Microbiome

SOFTWARE ARTICLE

Open Access

MetaPop: a pipeline for macro- and microdiversity analyses and visualization of microbial and viral metagenome-derived populations

Ann C. Gregory[1,2†], Kenji Gerhardt[1,3†], Zhi-Ping Zhong[1,4], Benjamin Bolduc[1], Ben Temperton[5], Konstantinos T. Konstantinidis[3,6] and Matthew B. Sullivan[1,7,8*]

Abstract

Background: Microbes and their viruses are hidden engines driving Earth's ecosystems from the oceans and soils to humans and bioreactors. Though gene marker approaches can now be complemented by genome-resolved studies of inter-(macrodiversity) and intra-(microdiversity) population variation, analytical tools to do so remain scattered or under-developed.

Results: Here, we introduce MetaPop, an open-source bioinformatic pipeline that provides a single interface to analyze and visualize microbial and viral community metagenomes at both the macro- and microdiversity levels. Macrodiversity estimates include population abundances and α- and β-diversity. Microdiversity calculations include identification of single nucleotide polymorphisms, novel codon-constrained linkage of SNPs, nucleotide diversity (π and θ), and selective pressures (pN/pS and Tajima's D) within and fixation indices (F_{ST}) between populations. MetaPop will also identify genes with distinct codon usage. Following rigorous validation, we applied MetaPop to the gut viromes of autistic children that underwent fecal microbiota transfers and their neurotypical peers. The macrodiversity results confirmed our prior findings for viral populations (microbial shotgun metagenomes were not available) that diversity did not significantly differ between autistic and neurotypical children. However, by also quantifying microdiversity, MetaPop revealed lower average viral nucleotide diversity (π) in autistic children. Analysis of the percentage of genomes detected under positive selection was also lower among autistic children, suggesting that higher viral π in neurotypical children may be beneficial because it allows populations to better "bet hedge" in changing environments. Further, comparisons of microdiversity pre- and post-FMT in autistic children revealed that the delivery FMT method (oral versus rectal) may influence viral activity and engraftment of microdiverse viral populations, with children who received their FMT rectally having higher microdiversity post-FMT. Overall, these results show that analyses at the macro level alone can miss important biological differences.

Conclusions: These findings suggest that standardized population and genetic variation analyses will be invaluable for maximizing biological inference, and MetaPop provides a convenient tool package to explore the dual impact of macro- and microdiversity across microbial communities.

*Correspondence: sullivan.948@osu.edu
†Ann C. Gregory and Kenji Gerhardt contributed equally to this work.
[8] Department of Civil, Environmental and Geodetic Engineering, Ohio State University, Columbus, OH 43210, USA
Full list of author information is available at the end of the article

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Keywords: Metagenomes, Visualization, SNP profiling, Community ecology, Population genetics, Macrodiversity, Microdiversity, Phage, Microbes, Ecogenomics

Introduction

Microbiology has experienced a revolution as sequencing and computational advances have enabled the cultivation-independent study of microbial and viral communities across diverse ecosystems. These studies have revealed the importance of the "microbiome" and its viruses as critical drivers whose metabolisms and impacts alter nutrient, metabolite and energy flows that dictate human health and ecosystem outputs (e.g., [1–3]). Pragmatically, the sequence space exploration has helped rewrite foundational taxonomic rules even to the point of genomes alone being sufficient [4, 5]. Though early studies relied upon gene marker derived amplicons and could answer "who is there" questions (e.g., [6, 7]), sequencing and analytical advances have led to increasingly improved assemblies such that genome-resolved, population-level analyses now get beyond "who is there" to understand metabolism and even mechanism (e.g., [8–12]). This transformation has happened rapidly, with catalogs of tens to hundreds of thousands of microbial and viral metagenome-assembled genomes now emerging across diverse environments (e.g., [13–22]). Beyond such inter-population (macrodiversity) community questions, recent advances are now also providing a window into intra-population (microdiversity) variation. These latter observations provide complementary information by establishing niche-defining gene sets, as well as how genetic drift and selection shape populations and communities [20, 21, 23, 24].

A major challenge when assembling fragmented DNA from complex communities is assembling short-reads into biologically meaningful "genomes" that represent ecologically and evolutionarily relevant populations. At this point, however, there are several improved population definitions that account for ecological and evolutionary theory [25] and have been extrapolated and assessed to varying degrees community-wide [26–29]. Many of the remaining criticisms, e.g., chimeric "franken-genomes," are being increasingly addressed by the rapidly advancing capabilities enabled by long-read sequencing and hybrid assembly approaches (e.g., [30, 31]). Thus, researchers studying microbes and viruses in complex communities have or soon will have datasets that are ready for genome-resolved population-based studies where high-fidelity assemblies and base calls can be expected.

Once assembled, several obstacles remain to establish intra-population biological inferences. *First*, population genetic methods rely on defined genotypes of individuals within a population with equal coverage across each base within a sequence. These are conditions not satisfied in metagenomes as their populations have unequal coverage and are assembled from many individuals within a population. To date, researchers have developed many methods to overcome these issues. The most common ones try to resolve each individual's genotype within the population by linking single nucleotide polymorphisms (SNPs) into strain-level genotypes [32–43] or use strain proxies [44–46], but these are difficult to apply to or are insufficient for community-scale studies across bacteria and viruses. A *second* obstacle is that analyzing both macro- and microdiversity in these modern datasets has scaling and standardization issues, and user-friendly bioinformatic tools are not yet available. For the latter, while several bioinformatic tools have emerged, they require intensive data manipulation prior to use and few do more than one type of analysis ([29, 39, 42, 44], see Table 1 in "Implementation"). This creates a research barrier for microbiologists that are light on computational skills, which could be alleviated by a tool that provided a single interface to analyze and visualize macro- and microdiversity patterns in metagenomic data.

To fill this gap, we introduce the multi-functional bioinformatic pipeline MetaPop, written in python and R and bundled in a bioconda package, that analyzes and visualizes microbial and viral community metagenomic sequence data at both the inter-(macrodiversity) and intra-(microdiversity) population levels. MetaPop can be easily utilized by beginner microbiologists with little training. In this sense, the pipeline complements existing bioinformatic pipelines, such as Anvi'o [44] and metaSNV [42], which can require users to modify the code, train, consult detailed tutorials, and continually provide input while running the pipeline. MetaPop's distinctive features include (1) it combines both macro- and microdiversity analyses into a single easy-to-use pipeline, (2) all of MetaPop's functions and parameters are called in a single command-line that processes and analyzes the input data from start to finish (with the option to run steps independently), and (3) it improves adaptive selection (pN/pS) results by determining if SNPs are linked at the codon level. MetaPop is fully documented and maintained by the developers at https://github.com/metaGmetapop/metapop.

Table 1 Capabilities of MetaPop compared to existing complementary bioinformatic pipelines

	MetaPop	Anvi'o	MIDAS	metaSNV	InStrain
Input files:					
QC'd reads	No	No	Yes	No	No
BAM file(s)	Yes	Yes	No	Yes	Yes
Genome fasta file	Yes	Yes	No	Yes	Yes
Gene file	No	No	No	Yes	Yes
Read/bp numbers	Yes	No	No	No	No
Preprocessing:					
Sorts & indexes BAM files	Yes	Yes	Yes	No	No
Removes spuriously mapped reads from each genome	Yes	No	Yes	No	Yes
Removes genomes with low horizontal coverage in each BAM	Yes	Yes	Yes	Yes	Yes
Identifies genomes with high read depth coverage per BAM	Yes	Yes	Yes	Yes	Yes
Performs gene calls	Yes	Yes	Yes	No	No
Macrodiversity:					
Calculates raw population abundances	Yes	Yes	Yes	Yes	No
Normalizes population abundances across samples	Yes	No	No	No	No
Calculates alpha-diversity	Yes	No	No	No	No
Calculates beta-diversity	Yes	No	No	No	No
Microdiversity:					
Calls SNVs	Yes	Yes	Yes	Yes	Yes
Does consensus SNP calling	Yes	No	No	No	Yes
Identifies codon variants	Yes	Yes	No	No	No
Downsamples data prior to population genetics calculations	Yes	No	Yes	No	No
Calculates *intra*-population diversity	Yes	Yes	Yes	Yes	Yes
Calculates *inter*-population diversity	Yes	No	No	Yes	Yes
Calculates linkage disequilibrium	No	No	No	No	Yes
Additional analyses:					
Contig annotation	No	Yes	Yes	No	No
Codon bias analyses	Yes	No (but does calculate codon usage)	No	No	No
Dereplicates genomes	No	Yes	No	No	No
Contig binning	No	Yes	No	No	No
Pangenome analyses	No	Yes	Yes	No	No
Outputs:					
Text data files	Yes	Yes	Yes	Yes	Yes
Visualizations	Yes	Yes	No	No	Yes

Implementation

Technical overview of how MetaPop works: input, data processing, and output

MetaPop has three inputs: (1) a genome FASTA file, (2) a tab-delimited file of the number of reads or base pairs per metagenomic library, and (3) one BAM file per metagenomic library of read alignments (mappings) to the reference genomes. The reference genomes should be assembled microbial or viral contigs that represent populations. Currently, MetaPop only works for haploid organisms and works best for bacteria, archaea, and dsDNA phages. MetaPop is best applied when each BAM file is derived using reads from a single metagenomic community—as defined by the user—rather than reads pooled from multiple communities, in order to prevent the formation of hybrid populations that could violate underlying assumptions of population genetic inferences. The definition of a single community will need to be defined by the user and the question that they are trying to answer. Commonly, single communities in the context of population genetics are defined as samples taken from two different locations or different time points from the same site. With these inputs, MetaPop analyzes the data in three steps (Fig. 1):

1. Pre-processing
2. Macrodiversity and codon bias analyses
3. Microdiversity analyses

To clearly and explicitly lay out the capabilities of MetaPop compared to existing complementary bioinformatic pipelines [29, 39, 42, 44], we provide a summary (Table 1). With the exception of the filtered reads, which are in binary alignment format, data outputs are tab-delimited files and visual outputs vectorized images stored in PDFs. The final tab-delimited files include (*i*) percent alignment and read length for every read in each sample, (*ii*) percent of positions covered by reads and average depth of coverage for each genome, per BAM file, (*iii*) the raw and normalized genome abundances, calculated (*iv*) α- and (*v*) β-diversity values, (*vi*) genes with different codon biases, (*vii*) single nucleotide variant (SNV) calls and pileup files over SNV positions, (*viii*) called single nucleotide polymorphisms (SNPs), split into those which appear on genes and those which appear in intergenic regions, (*x*) linked SNP results, and (*xi*) intra- and (*xii*) inter-population genetic calculations. The visualization output include (*i*) overall summaries of read filtering, (*ii*) summary plots showing BAM file genome coverage and depth statistics, (*iii*) a heatmap of normalized genome abundances, (*iv*) scatterplots of α-diversity values, (*v*) ordination plots of β-diversity results, (*vi*) bar plots of codon position of detected SNPs, (*vii*) visualization of nucleotide diversity and codon bias per each gene for each genome, and identification of positively selected genes, and (*viii*) heatmaps of F_{ST} per genome across samples.

Step 1: Pre-processing

Though user-customizable, MetaPop defaults to a 95% nucleotide identity (ID) cut-off, but can be changed by the user, to define population boundaries, guided by studies exploring sequence space boundaries between microbial and dsDNA viral populations [5, 20, 21, 26–29, 47]. During pre-processing, input BAM files are first sorted and indexed and reads that map at < 95% ID to a reference genome or which are shorter than 30 base pairs are removed. Genomes that pass either a length and/or percentage coverage minimum after this initial filtering are considered present or "detected" in a given sample and move on to macrodiversity analyses [20, 21, 48]. If the user flags the genome fasta dataset as complete microbial genomes (-complete_bact), MetaPop will use a default detection cut-off of ≥ 20% genome length covered to consider the genome in further analyses [48, 49]. If the user flags the genome fasta dataset as viral (-viral) or fragmented microbial contigs (-frag_bact), MetaPop's default detection cut-offs require at least ≥ 5-kbp genome length covered in genomes > 5 kbp and ≥ 70% length for genomes < 5 kbp [20, 21]. All length cut-offs, however, can be adjusted by the user. We recommend using the same %ID cut-off and horizontal coverage cut-off as those used to de-replicate your genomes into populations. However, if detection of extremely rare taxa is important, consider lowering the horizontal coverage cut-off.

Once a population is detected by these above criteria, MetaPop calculates its relative abundance based on mean nucleotide coverage across the genome (see the "Macrodiversity analyses" section for more details). Loci with coverages below the 10th and above the 90th percentile are excluded from this assessment to prevent skewing of abundances from fast-evolving regions, such as genomic islands, and spurious recruitment of reads to highly conserved regions, respectively [50]. Importantly, users can customize any of these cut-offs for percent identity to define populations, horizontal coverage of the genome to "detect," and the quantiles for minimizing abundance data skew.

Step 2: Macrodiversity and codon bias analyses

Data processing and calculating alpha and beta diversity indices

Macrodiversity is the measure of population diversity within a community. While some diversity measurements rely strictly on the presence or absence of populations (such as richness and Jaccard distances), many rely on the relative abundances of populations between communities (such as Shannon's H, Simpson's, and Bray-Curtis distances). Importantly, metrics that rely on relative abundances have been shown to be more robust for metagenomic data because they are less susceptible to uneven sampling of rare taxa [51]. Thus, the raw abundances calculated during the pre-processing step must be transformed in order to allow for differential abundance testing. MetaPop proportionally normalizes per-sample abundances to those for the library with the highest number of either the number of reads or base pairs (selected by the user). For example, if library A has 1.5 million reads and library B has 2 million reads, all the raw population abundances in library A are multiplied by 1.33 to proportionally scale the abundances to the library with the highest number of reads. If more than one sequencing technology was used to create the different metagenomes and this resulted in vastly different

(See figure on next page.)

Fig. 1 MetaPop pipeline overview. MetaPop requires three primary inputs (a genome fasta file, file with the number of reads or bps per library, and unsorted BAM files). The BAM files are sorted and indexed and preprocessed (here showing the default setting for dsDNA viruses). The output of preprocessing goes through the macrodiversity or microdiversity arms of the pipelines. Codon usage bias is calculated as well and can be calculated independent of the whole MetaPop pipeline

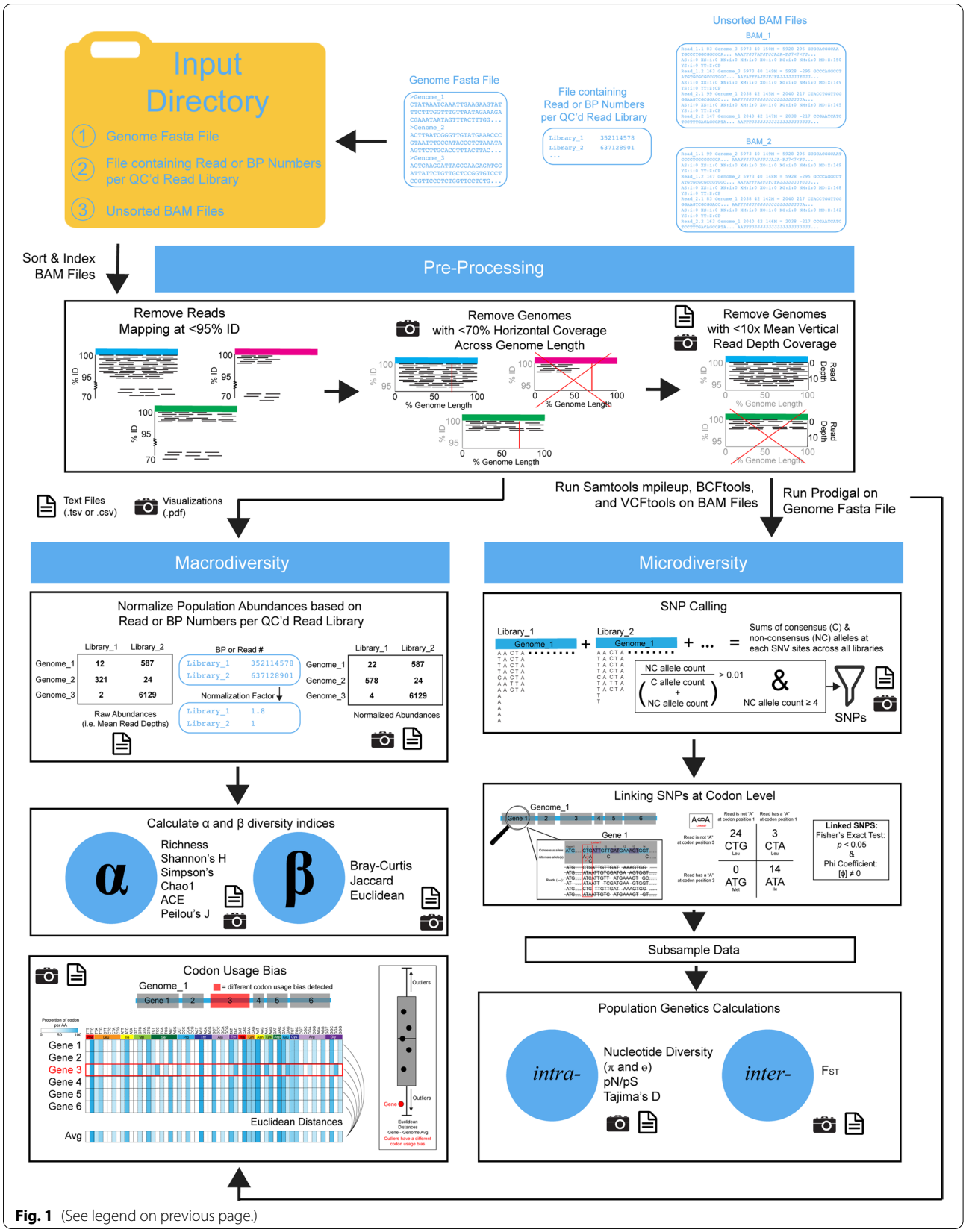

Fig. 1 (See legend on previous page.)

read lengths, we recommend using base pair counts for the normalization step. Normalized genome abundances for each metagenomic sample are used to calculate macrodiversity measurements with the "vegan" R package. If the input consists of a single BAM file, α-diversity (*within* community) indices—richness, Chao1, ACE, Shannon's H, Simpsons, inverse Simpsons, Fisher, and Pielou's J—for that community are calculated. With multiple BAM input files, β-diversity (*between* community) indices—Jaccard, Bray-Curtis, and centered log-ratio-transformed Euclidean distances—between all communities are also calculated. Importantly, MetaPop also outputs the raw abundances, so that the user can normalize their own data.

Codon bias analyses

Microbial and viral populations often have distinct codon biases for translational optimization [52]. Genes with codon usages different from the rest of the genome often have been recently horizontally transferred [53], have different temporal regulations [54], or are highly expressed [55]. MetaPop predicts putative genes using Prodigal [56] for all genomes in the reference genome FASTA file and identifies the codon usage for each gene within a genome and then calculates the codon bias for each gene. The bias for each codon per amino acid across every gene in the genome is then averaged to create the average codon bias. Each gene's codon usage is compared using Euclidean distances to the average codon bias. Genes with Euclidean distances greater than 1.5 times the interquartile range, the standard constant for discerning outliers (reviewed in [57]), of Euclidean distance for that genome are considered potential outliers and are marked as having aberrant codon usage for their respective genome.

Step 3: Microdiversity analyses

Data transformation

Microdiversity is the measure of genetic diversity within a population. In natural communities, where populations are represented at different abundance levels, only genomes with enough data can be evaluated. Thus, by default, genomes with < 70% length of their genome covered and < 10× average read depth coverage are excluded from these analyses to ensure that there is enough coverage to accurately call SNPs and to assess contig-level microdiversity. The 10× value was selected because prior work revealed that downsampling read depth to 10× did not statistically significantly impact downstream microdiversity calculations [23]. While deeper sequencing is now resulting in high coverage for many microbial and viral populations, it is also uncovering rare low-abundance species that remain with low coverage [20, 21, 58]. Thus, in order to compare these low and high coverage species, downsampling remains important. Users can set this parameter if they want to be more stringent or relaxed in the number of populations that pass to the microdiversity analyses step. Prior to SNP calling, MetaPop identifies SNVs within each population per BAM using the mpileup tool in samtools [59] and BCFtools [60] in order to obtain per-position variant information, followed by removal of low (PHRED < 20 by default) variant quality score calls. Importantly, decreasing or increasing the PHRED threshold for variant quality scores increases or decreases, respectively, the number of SNPs called and the downstream nucleotide diversity values. SNPs are identified using two methods, either a (1) global or (2) a local approach. For global SNPs calls, the base pair coverage for each SNV position per genome is pooled across all metagenomes and the consensus allele verified. Alternate alleles that make up ≥ 1% of the base pair coverage for that position [61] and represented by at least 4 reads are considered true SNPs [23]. For local SNP calls, the set of true positions identified in the global calls are reduced to the set of SNV positions identified in each BAM individually. SNV sites only observed in other BAM files are ignored.

Identified SNPs are cross-referenced with gene calls and assigned as either genic or non-genic. If genic, their position within each codon per gene is determined. Due to redundancy at the third position in codons that allows multiple codons to code for the same amino acid, most true SNPs should be at the 3rd position of the codon. MetaPop outputs all of the SNPs called and their codon positions if genic. MetaPop will issue a warning if there are more SNPs in the 1st and 2nd positions of a codon. Lastly, the global verified consensus allele per each SNP position is replaced as the consensus allele in each reference genome.

SNP linkages in codon variants and downsampling

SNPs are tested for local linkage at the codon level to identify codon variants by evaluating their co-presence within reads in each BAM file. Multiple programs try to link SNPs across the genome into strain haplotypes using the reads [39–41, 43]. However, given that shotgun sequencing read lengths are shorter than most gene and genome lengths, it makes it difficult to resolve genotype patterns that span across more than a single read length. Assessing linkage across small sequences that can be contained with a single read, nonetheless, provides the strongest evidence of linkage. MetaPop tests for linkage at the codon level due to its importance for studying protein evolution. The linkage of SNPs, for example, at positions 1 and 3 within a codon can code for a completely different amino acid than if each SNP independently arose. Further, codons are short enough to be contained with a single read, which allows us to accurately test the

linkage between or independence of SNPs shared within a codon. The resulting codon variants that code for different amino acids, most often those that have two SNPs or SNPs in the 1st or 2nd position of the codon, and their resulting impact on protein structure have recently become an active area of research in metagenomes [62]. Recent work, however, identified these codon variants by filtering for highly abundant codon variants already contained within 20 reads [62]. To the best of our knowledge, MetaPop is the first program to try to statistically link SNPs at the nucleotide level into these codon variants.

In order for MetaPop to link the SNPs at the codon level, SNPs that localize in the same codon on the same gene are selected as candidates for linked SNP identification. The original reads covering the positions of the candidates from their respective genomes are collected from each BAM file, and the codons relevant to the candidates are extracted from those reads. SNPs within the same codon are tested in pairs. If more than two SNPs occur within a codon, pairwise tests are performed between each combination of pairs. A contingency table of the frequencies of the extracted codons with both SNPs, the number of extracted codons with one SNP, the number of extracted codons with the other SNP, and those containing no SNPs is produced. Fisher's exact tests and phi coefficients are calculated. The linked SNP candidates are classified as either "linked" (Fisher's p-value < 0.05, ϕ > 0) meaning the SNPs occur together as a set disproportionately, "independent" (Fisher's p-value < 0.05, ϕ < 0), meaning the presence of one of the candidates excludes some of all of the others in that set disproportionately, or as "ambiguous," meaning that they occur together or separately at apparent random, or that there is insufficient data to classify them otherwise. In ambiguous cases, SNPs are treated as independent for downstream analyses.

SNP frequencies are subsampled down to 10× coverage proportionate to the frequency of different SNPs per site while maintaining SNP linkages. This stage normalizes the probability of a variant occurrence by chance across variant sites within a population genome. It also rarefies all the SNP frequencies across all genomes and samples allowing for differential SNP frequency testing. While the user can adjust the subsampling level, we recommend that you subsample down to the same average read depth cut-off.

Population genetic calculations: θ, π, F_{ST}, pN/pS, and Tajima's D

The subsampled SNPs are then used to assess population-level genetic diversity and explore protein and genome evolution. This will occur with both global and local SNP calls. If the population occurs in only one BAM file, only intra-population diversity (within-population microdiversity)—expected nucleotide diversity (θ; [63]) and the observed nucleotide diversity (π; [64])—are calculated. Both θ and π are calculated at the individual gene and whole-genome levels. Because we use a default minimum genome coverage of 70%, not every SNP position for a population will be covered within a BAM file. To correct for this, we use the following equation to estimate θ:

$$\theta = \frac{N}{\mid G \mid -(N-n)}$$

where N is the total number of SNP positions within a gene or genome, n is the number of SNPs covered within a metagenome, and $|G|$ is the total gene or genome length. To estimate π, we modified the Schloissnig et al. [23] equation:

$$\pi = \frac{1}{|G|-(N-n)} \sum_{i=1}^{|G|} \sum_{B_1 \in \{ACTG\}} \sum_{B_2 \in \{ACTG\} \setminus B_1} \frac{x_{i,B_1}}{c_1} \frac{x_{i,B_2}}{c_1 - 1}$$

where N is the total number of SNP positions within a gene or genome, n is the number of SNPs covered within a metagenome, $|G|$ is the total gene or genome length, $x_{i,Bj}$ is the number of nucleotide B_j seen at position i and c_i the coverage at position i in the gene or genome. If the population occurs in more than one BAM, F_{ST} ([65]; between population microdiversity) is calculated. Because F_{ST} requires comparing the nucleotide diversity per site across two metagenomes, we chose to keep the total genome length as the common denominator given that the SNP coverage may vary between both metagenomes. The implemented equation for F_{ST} is directly from [23].

To explore selective pressures on specific genes, MetaPop uses two methods: pN/pS [23] and Tajima's D [66]. The implemented equation pN/pS is directly from [23] except it factors in codon-constrained SNP linkages. Tajima's D is calculated using the original equation [66], but using the π value calculated above, the number of SNP positions within a gene as the number of segregating sites, and the ceiling mean read depth for the number of sequences.

Results and discussion

Biological evaluation of MetaPop

In order to test MetaPop, we ran the pipeline on three previously published datasets, a synthetic dataset representing mock bacterial communities and two biological virome dataset with natural variation (i.e., beyond that in the mock community). The synthetic dataset is composed of 30 mock, bacterial metagenomic communities of different known-proportions of *Staphylococcus aureus*, *Staphylococcus epidermidis*, and *Bacillus subtilis*

strains ([34]; see Table S1). The first biological virome dataset included 131 of the viromes deriving from the *Tara Oceans* expedition from the Global Oceans Virome 2 (GOV2; [20]) dataset. This dataset was the first dataset to assess microdiversity in metagenome-assembled viral genomes at a community-wide scale and provided the methodological backbone of MetaPop. The second biological virome dataset was composed of gut viromes from 12 autistic children that underwent fecal microbiota transfers and 6 neurotypical children ([67]; see Table S2). The default visualization outputs of MetaPop for the second biological virome dataset can be seen in Fig. 2.

MetaPop reproduces macrodiversity patterns in silico mock communities

We first tested MetaPop's default settings to accurately determine community composition and to calculate macrodiversity values across the 30 mock bacterial metagenomic communities. The communities are of varying known proportions of three distinct strains of *S. aureus*, three distinct strains of *S. epidermidis*, and a single strain of *B. subtilis* [34] and have varying numbers of reads, from ~ 2 million reads (communities 1–10), ~ 3 million reads (communities 11–20), ~ 6 million reads (communities 21–30). This dataset is practical to test macrodiversity calculations in MetaPop because the community is composed of two dominant closely related bacterial species-level populations (*S. aureus* and *S. epidermidis*) that share > 80% ANI and a more distantly related, rare, species-level population (*B. subtilis)* (Fig. S1). This taxonomic combination and different simulated sequencing depth enabled us to determine whether MetaPop could distinguish between closely related populations, and whether increased microdiversity within a population as well as sequencing depth impacted our ability to correctly assess macrodiversity.

Across the 30 communities, bacterial population relative abundances were almost identical to the simulated proportions with only a 0.98 and 1.08 mean fold change differences of the *S. aureus* and *S. epidermidis* species, respectively, and a 0.26-fold change average difference of the *B. subtilis* which is simulated to represent a rare taxon across the communities (Fig. S2A). This fold change difference is similar to or less than known quantitative biases, such as 10% divergence in alpha- and beta-diversity values seen in other metagenomic analyses for viruses [68] and ~ 3–9% divergence the microbes if genome length is accounted for [49]. The number of strains within each population and the number of reads did not impact detection of different populations (Fig. S2A). MetaPop estimates of α-diversity (Fig. S2B; all α-diversity indices: Wilcoxon $p > 0.05$) and β-diversity (Bray-Curtis dissimilarity) did not significantly differ from the actual values (Fig. S2C; Mantel's test $p > 0.05$). Thus, despite the minor fold change differences in community composition and difficulty in accurately detecting the abundances of rare taxa, MetaPop is generally able to accurately assess the community composition and macrodiversity biological trends.

MetaPop's codon usage bias analyses detect highly expressed and horizontally transferred genes in Staphylococcus aureus

MetaPop also looks for variation in codon biases among genes within each genome. Genes with different codon usage are often associated with horizontal gene transfer, high expression, or different temporal regulation of expression [53–55]. To determine the biological validity of MetaPop's codon's usage analysis, we evaluated codon bias across all 7 strains in the mock community. We choose to focus our analyses on the genome of an ST5 methicillin-resistant strain of *S. aureus* (see the full list of codon bias outliers in Table S3) because it is a well-studied human pathogen with known regions of

(See figure on next page.)

Fig. 2 MetaPop Visualization Outputs from our autism biological virome dataset. Pre-processing visualizations include **a** bar plot showing how many reads were kept and removed following removal of reads below a 95% ID cut-off across all samples, **b** scatter and bar plot composites (1 example shown) reported per sample showing how many genomes pass the horizontal and vertical coverage cut-offs, and (**b**—inset) donut plots summarizing the total number of genomes passing the different horizontal and vertical coverage cut-offs per sample. Macrodiversity visualizations include **a** heatmap summarizing the normalized abundances of covered genomes across the different samples (the max value on the color scale reported is the 75% quantile of all abundances to allow low abundance genomes to be better displayed; another heatmap not shown is also created showing a full range of abundances), **b** scatter plots per each alpha diversity index (4 examples shown) showing the alpha diversity value across all samples with horizontal lines showing the mean and median values, **c** ordination plots (PCA, PCoA, and NMDS) of all centered-log ratio transformed Euclidean distances, Bray-Curtis distances, and Jaccard distances, respectively (all distances are plotted using the 3 ordination methods by default). The color of each circle represents the species richness within each sample. The codon usage bias visualization is a circular bar plot per genome (1 example shown) showing the Euclidean distance of each gene from the average gene codon bias. Genes with outlier codon biases are displayed in red. Microdiversity visualizations include **a** Stacked bar plot (right) and standard bar plot (left) showing the distribution of SNPs across codon position and the total number of SNPs per sample. **b** F_{ST} heatmaps per genome (2 examples shown) showing the population differentiation per genome across all samples it has coverage within. **c** Genome plot composites for each genome in each sample where it has coverage (1 example shown) with four different tracks from top to bottom showing a line graph of the depth coverage of the genome, a genome plot of the genome with coloration of genes showing pN/pS results, a scatter plot showing π and θ values, and, lastly, a scatter plot showing Tajima's D values with the color background showing whether the value is indicative of selection

Fig. 2 (See legend on previous page.)

the genome that were horizontally transferred [69, 70] or highly expressed [69, 71, 72]. Importantly, MetaPop, in its default settings, is conservative because it compares each gene's codon bias to the average codon bias across the whole genome and will, thus, underestimate the number of genes with different codon usage. However, these settings are adjustable by the user. Given this conservative approach, we could only validate how many genes with a detected difference in codon bias were either mobile elements or highly expressed genes. Across the *S. aureus* ST5 strain, the vast majority (71%; 104 out 149) of the genes detected to have outlier codon usage have no known function. Of the remaining 45 annotated genes MetaPop detected with outlier codon usages, 20% are genes on known mobile elements (so prone to HGT) or are thought to be horizontally transferred and 47% are known to be highly expressed (Fig. S3). The identified known mobile elements include many toxin-antitoxin system genes [69, 70] and putatively transferred DltX and DltC proteins involved in wall teichoic acid, as well as poly(glycerol-phosphate) alpha-glucosyltransferase, a type of glycosyltransferase [73]. The highly expressed genes include ribosomal proteins [71], genes involved in transcription and translation such as elongation factor Tu [71, 74], chaperones [71, 75, 76], and all the phenol-soluble modulins (PSMα1-4 and PSMβ; [72]). Thus, MetaPop, in its default conservative settings, will not identify all horizontally transferred genes or highly expressed genes, but it provides an important first look at potential targets for further study.

MetaPop reproduces microdiversity patterns in the Global Oceans Virome 2 dataset

Using the GOV2 dataset, we next evaluated MetaPop's ability to assess microdiversity values and trends (Fig. S4). MetaPop calls SNPs using two methods, either a (1) global or (2) local approach. For global SNPs calls, the base pair coverage for each SNV position per genome is pooled across all metagenomes and the consensus allele verified. The original GOV2 paper explored microdiversity in the form of average nucleotide diversity (π) per sample by randomly subsampling the π values of different viral populations in each sample and averaging those values. We replicated these methods with the π values calculated using MetaPop where SNPs were called locally when they had a differential base with a PHRED score $\geq$ 30 (see "Materials and methods"). As a result, we ran MetaPop using PHRED $\geq$ 30 and its default of $\geq$ 20 on the GOV2 dataset.

Importantly, MetaPop calculates π slightly differently than the method used in the original analyses of the GOV2 dataset. The original method used the exact equation derived from Schloissnig et al. 2013 [23] which divides the calculated nucleotide diversity by the total genome length to obtain π. Because of unequal coverage across genomes in each sample, SNP positions are often not covered so it is impossible to assess the diversity at that site. As a result, MetaPop subtracts the number of SNP positions not covered from the genome length prior to dividing the nucleotide diversity in order to calculate π. Thus, π values from MetaPop will be slightly higher than values using the Schloissnig et al. 2013 [23] equation. As expected, MetaPop's average π using the same SNP calling thresholds (PHRED $\geq$ 30 and local SNP calls) were slightly higher (median 1.33 fold-change) than the original GOV2 average π (Fig. S4A, left). Due to the differences in random subsampling, there were also clear deviations between the original GOV2 average π values and the MetaPop derived values. Nonetheless, the original GOV2 average π and MetaPop's average π still strongly correlated (Fig. S4A, right; linear regression: $R^2 > 0.62$), indicating that despite higher average π and slight fluctuations in average π derived from the random subsampling process, the biological microdiversity patterns are still being captured due to a systematic adjustment consistent with calculations from fragmented genomes that derive from metagenomic datasets.

The SNP calling approach (global versus local) and PHRED score (i.e., a measure of the quality of the called nucleotide) can also impact downstream π values. Global SNP calling, for example, incorporates all SNP loci that were identified in any sample in the dataset into the π calculation for each sample (even if it was not called as an SNV for that exact sample), which will increase π. Using the GOV2 dataset, we see just that with PHRED $\geq$ 30 global SNP derived average π having a median 4.32-fold increase from the original GOV2 average π values calculated using PHRED $\geq$ 30 and local SNP calls (Fig. S4B, right) and a median 3.19-fold increase over MetaPops's average π using PHRED $\geq$ 30 and local SNP calls (Fig. S4E, right). Further, decreasing the minimum PHRED score requirements allows more potential SNVs and thus SNPs to be called per sample and, thus, the average π values should be higher. As expected, we see that using a PHRED $\geq$ 20 global SNP call approach increases average π values by a median 5.88 fold-increase from the original GOV2 average π values calculated using PHRED $\geq$ 30 and local SNP calls (Fig. S4C, left) and 1.32 fold change from MetaPop's PHRED $\geq$ 30 global SNP approach (Fig. S4D, left). Importantly, regardless of SNP calling approach or PHRED score cut-off, the π calculated using the original GOV2 approach or MetaPop's approaches are all strongly correlated (Fig. S4A-E, right; linear regression: $R^2 > 0.48$ to 0.68). Further, analyses of larger microdiversity trends across ecological zones in the ocean defined in the original GOV2 analyses were

also able to be replicated using both PHRED score cutoffs testing and a global and local approach (Fig. S4F). Taken together, MetaPop is able to accurately derive microdiversity values and biological trends.

MetaPop's codon-constrained linkage of SNPs improves detection of positively selected genes

Using the two biological datasets, we evaluated the impact of MetaPop's novel codon-constrained SNP linkages on pN/pS selection analyses. The original pN/pS equation [23] calculates the number of non-synonymous and synonymous codons without first evaluating if SNPs within the same codon are linked. If the two codon-constrained SNPs are linked, with the exception of two codon variants for leucine, the presence of two SNPs within the same codon will always lead to a non-synonymous codon. Thus, without codon-constrained SNP linkages, we hypothesized that we may be underestimating the number of genes detected under positive selection using pN/pS. MetaPop tries to resolve this issue by linking SNPs at the read level (as many tools do), but also at the codon level (see the methods in "Step 3: Microdiversity analyses" section above). We tested our hypothesis on the GOV2 and autism biological datasets using MetaPop and a global SNP calling approach to maximize the number of codons with putatively linked SNPs per sample.

Of the total genes in the autism dataset, ~ 1.4% of genes ($n = 248$) with enough coverage to evaluate selection had ≥ 1 codon with putatively linked SNP in at least one sample (Fig. 3A, top—larger circle). Of this subset, 16.97% contained at least one codon with potentially linked SNPs (Fig. 3A, top—smaller circle). The subset of genes containing a putatively linked codon had their pN/pS ratios calculated using both with and without linking SNPs, and their results were compared. When SNPs were not linked, we observed that 21.7% of the genes ($n = 54$) displayed positive selection, and when linking SNPs, we observed that 26.9% of the genes ($n = 63$) displayed positive selection (Fig. 3A, bottom). There were minimal differences using a PHRED ≥ 30 or ≥ 20, with PHRED ≥ 30 detecting 25.4% ($n = 258$) genes under positive selection using the codon-constrained SNP linkages, compared

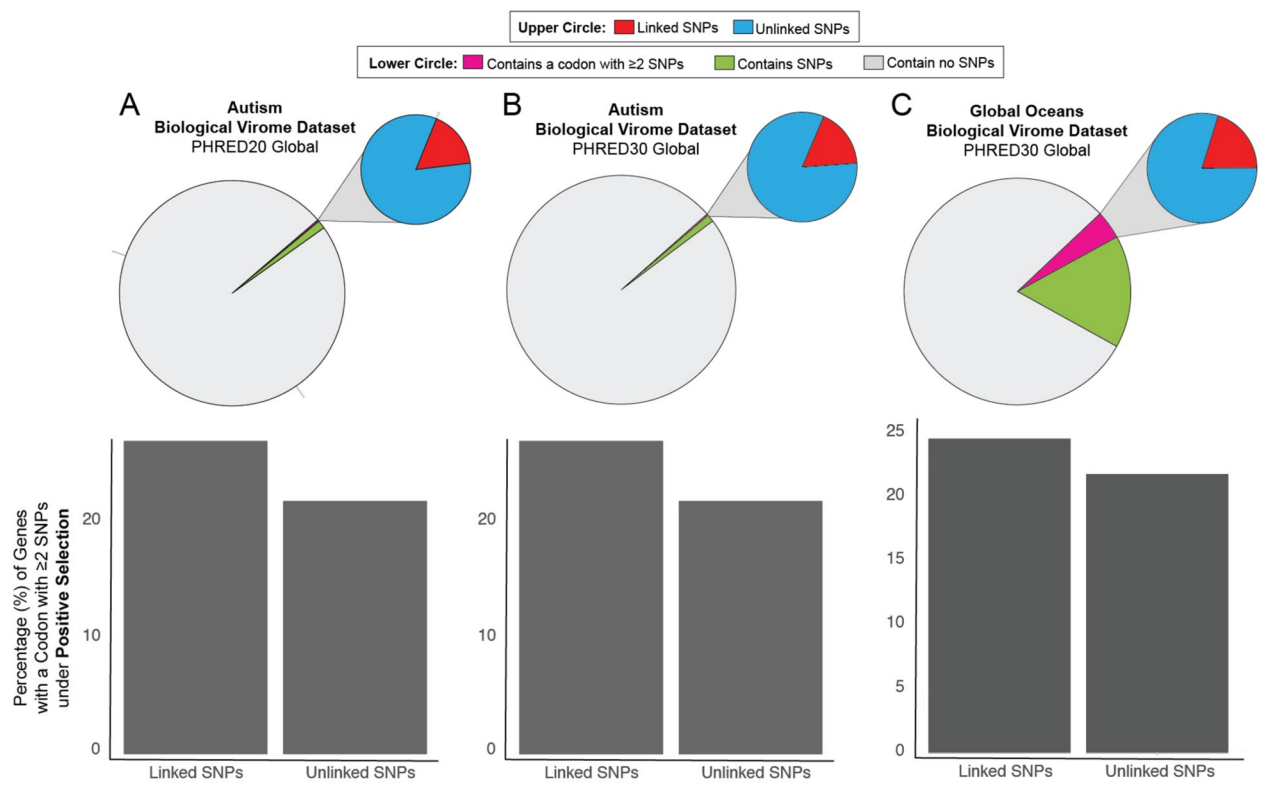

Fig. 3 MetaPop's codon-level SNP linkages increase the number of positively selected detected genes using pN/pS. (**A**–**C**, top) Pie charts displaying the predicted genes belonging to a contig which passed preprocessing coverage and depth cutoffs, genes with at least one SNP observed, and genes with at least one codon with putatively linked SNPs, and a pop-out pie chart showing the breakdown of genes with observed SNPs and those containing a codon with putatively linked SNPs. (**A**–**C**, bottom) Barplots comparing the percent of genes under selection when calculated after linking SNPs vs. not attempting to link SNPs. Data is shown for the autism biological virome dataset using **A** PHRED20 and **B** PHRED30 global SNP calls and on the Global Oceans' biological virome dataset using **C** PHRED30 global SNP calls

with 26.9% (n = 271) with PHRED ≥ 20 (Fig. 3B). In GOV2, we observed a similar pattern. In GOV2, 5% of genes (n = 229,058) with enough coverage to evaluate selection had ≥ 1 codon with putatively linked SNP (Fig. 3C, top—larger circle). Of this subset, 20.1% contained at least one codon with potentially linked SNPs (Fig. 3C, top—smaller circle). Similarly to the autism dataset, we saw that SNPs linkages increased the number genes detected under positive selection from 21.7% (n = 50,094) to 24.56% (n = 56,467) (Fig. 3C, bottom). Thus, MetaPop's codon-constrained SNP linkage shows that we are underestimating the number of positively selected genes that contain these putatively linked SNPs by an average of ~ 4%. Further, it shows that utilizing pN/pS to identify genes under selection without linking SNPs at the codon misses genes under positive selection.

Microdiversity: a case study in assessing intra-population variation reveals gut viruses may play a role in dysbiosis of autistic children's guts

To demonstrate the value of adding estimates of microdiversity to a researcher's toolkit, we used MetaPop to re-analyze the gut viromes of autistic children that underwent fecal microbiota transfer (FMT) and their neurotypical peers. Eighty percent of autistic children suffer gastrointestinal problems, so understanding how the gut microbiota differs between autistic and neurotypical children may be important for treating this symptom of autism [77]. Previously, we found bacterial macrodiversity was lower in autistic children compared to their neurotypical peers, but that viral macrodiversity (Shannon's H) was not significantly different (see Fig. 4A; Wilcoxon's test p = 0.89; [67]). Thus for this demonstration, we chose to specifically focus on the macro-α-diversity Shannon's H (as a positive control for whether MetaPop could recover our past observations), and the microdiversity average π (to assess what biological inferences can be gained by such measures).

Here, MetaPop revealed that average viral microdiversity (π) is significantly lower in autistic children than within their neurotypical peers (Fig. 4B; Wilcoxon's test p = 0.028), paralleling our previous bacterial macrodiversity findings [67]. High average π can indicate two biological outcomes: (*i*) more viruses from different populations are actively infecting host bacteria resulting in population expansion and more mutations, or (*ii*) more viral populations naturally maintain higher levels of microdiveristy in their standing populations. Either way, having a higher level of microdiveristy for a viral population can be an adaptive advantage because it better allows populations to "bet hedge" if their environment or hosts change [78]. We next looked to see if this increased microdiversity could be providing an adaptive advantage to neurotypical children by exploring the average number of genomes containing a gene found under positive selection using pN/pS. Indeed, neurotypical children had significantly more genomes with at least one gene under selection than the autistic children (Fig. 4C; Wilcoxon's test p = 0.026). Thus, we hypothesize that increased average π is beneficial in the gut virome because it allows viral populations to better adapt to their changing environments and hosts. Increased diversity at either the macro- and microdiversity level has consistently been shown to be important for maintaining ecosystem functions and services (e.g., [78, 79]), with the loss of diversity resulting in a loss of ecosystem resilience. Thus, the loss of viral microdiversity in autistic children's guts could potentially be indicative of a loss of the gut ecosystem's resilience.

Next, given that the autistic children underwent FMT, we were also curious how FMT impacted their viral microdiversity. We compared the pre-FMT (week 0) to the post-FMT (week 10) gut viral microdiversity (see [67] for full information about FMT design). Of the 12 autistic children with viromes available, 10 of the children responded positively to FMT treatment and 2 did not. Given the hypothesis that increased average π is beneficial, we expected to see that all of the responders would have increased viral microdiversity. Instead, we saw an interesting pattern (Fig. 4D). Six of the children with virome data were given the FMT orally, with the other six were given the FMT rectally. Across the responders that received FMT orally, only 2 of the 5 children had increased viral microdiversity, and only an average 1.36-fold increase at that. In contrast, 4 of the 5 children that received the FMT rectally responded with increased viral microdiversity, and they did so with a larger (11.58-fold) average increase. This suggests that rectal administration of the FMT may promote engraftment of more microdiverse viral populations, at least those surveyed in the feces, than the oral administration of FMT. This contrasts the findings at the clinical symptom level that found no significant difference in changes in children that received the oral or rectal FMT [67]. Taken together, though the study was pilot-scale and open-label, these microdiversity results suggest that viral population structure is associated with the autism disease phenotype and that the FMT delivery method may correlate to responder status. Again, however, a larger study is needed to better guide standard of care practices.

Computational evaluation of MetaPop

We next evaluated the processing time and computational resource consumption of MetaPop by running the synthetic mock bacterial and two biological datasets on the Ohio Supercomputer (OSC). To simulate different computational power, from a standard laptop to desktop

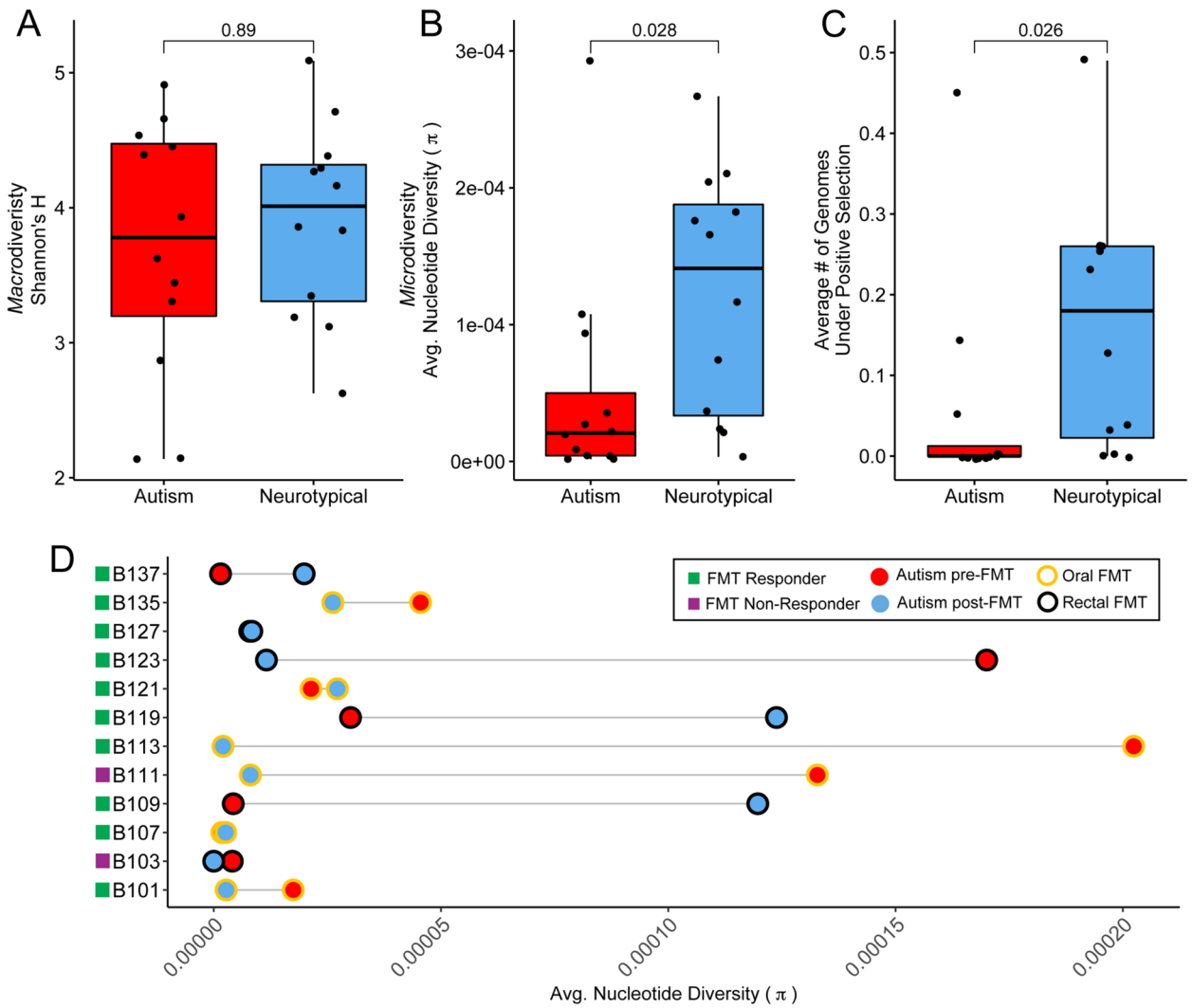

Fig. 4 MetaPop's microdiversity analyses reveal lower *intra*-population diversity in autistic children. Boxplots showing median and quartiles of **A** Shannon's H, **B** average nucleotide diversity (π), and **C** the average number of genomes under selection in pre-FMT autistic and neurotypical children. The Wilcoxon's test *p*-values above are the result from comparing pre-FMT autistic and neurotypical children values. **D** Connected dot plot showing the average π for the autistic children pre-FMT and post-FMT. The color filling the circle represents the pre- and post-FMT status and the outlier color of the circle represents the method in which the FMT was delivered (oral or rectal delivery). The colored squares next to the autistic child's ID corresponds to whether the child responded or did not respond to the FMT treatment

computer to a small-sized computer cluster, we ran MetaPop using 4, 6, and 14 cores, using 8, 16, and 64 GB of memory (RAM), respectively (see Fig. 5A). The much larger GOV2 dataset needed more cores and memory to run than the different settings tested, and was supplied with 48 cores and 128GB of memory. Importantly, MetaPop's code is parallelized, so increasing the number of supplied cores will increase the memory used because MetaPop will try to parallelize its steps as efficiently as possible given the computational resources supplied.

To assess both processing time and computational resource consumption, we first wanted to determine what steps in the MetaPop pipeline were rate-limiting. Across both the synthetic and biological datasets, the pre-processing and the SNP calling portion of the microdiversity section took the most processing time (Fig. 5B). These two steps operate on the entirety of data supplied by the user, and must perform multiple operations on every read in each BAM file, and for every contig supplied. This means that they must process large volumes of data and consume commensurate computational resources. Further, the resources consumed by these steps depend on the degree of parallelization, as each parallel process will operate on its own set of data, simultaneously. The latter

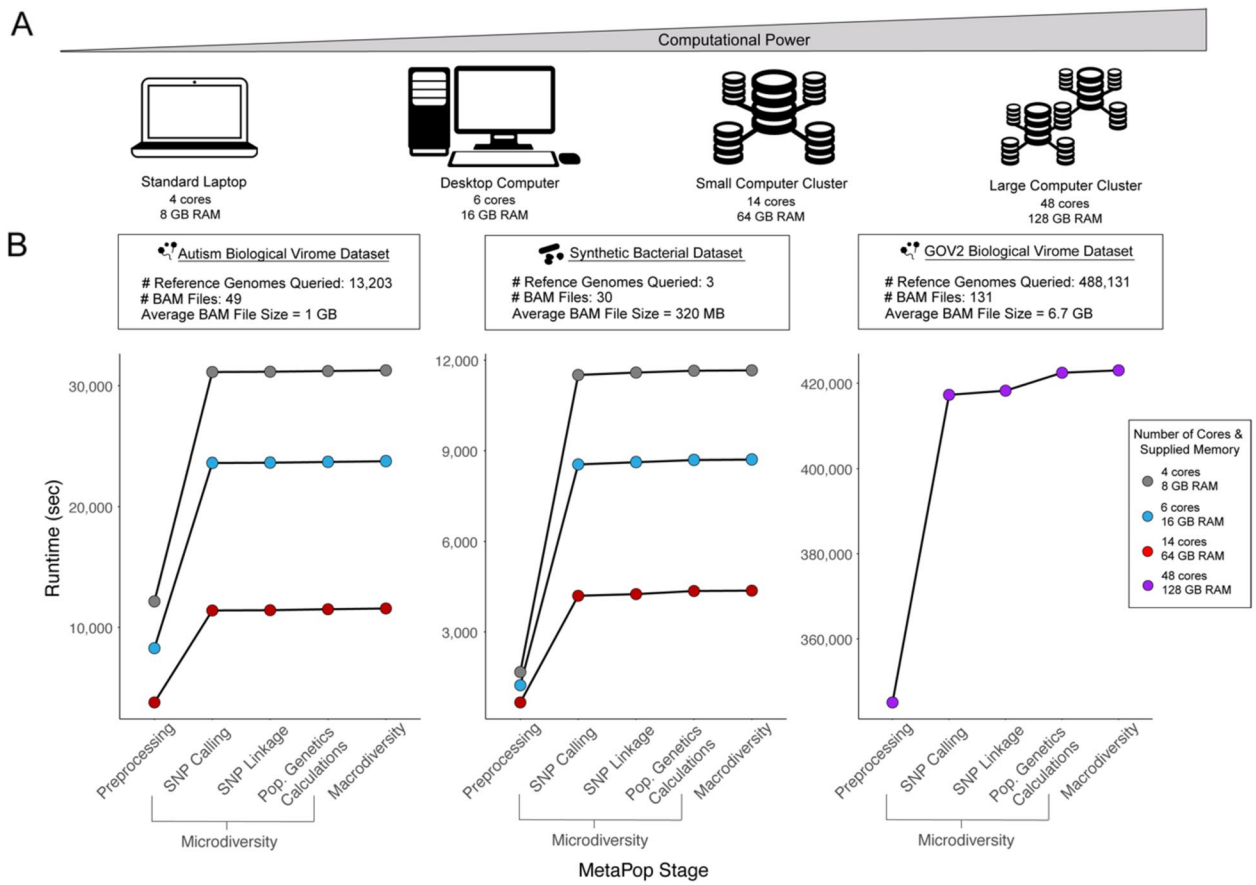

Fig. 5 Computational evaluation of MetaPop. **A** Schematogram showing the different computational power levels tested. **B** Line plots showing cumulative runtime of MetaPop on (left) the autism biological virome dataset, (center) the synthetic bacterial dataset with 4, 6, and 14 processors, and (right) the GOV2 biological virome dataset using 48 processors

portions of the pipeline work on summaries produced by pre-processing and SNP calling steps and are, correspondingly, faster because they do not have to operate on the entire input dataset, even though these later steps have substantially less parallelization.

With the knowledge of the rate-limiting steps, we then assessed the effect of BAM file size and the amount of computational power supplied on processing times. BAM files size is impacted by the number and length of reference genomes and the original number of reads. The synthetic and biological datasets differ substantially in these values resulting in vastly different BAM file sizes and allowing us to test MetaPop across a range of BAM file sizes, with the average synthetic dataset's BAM file size equal to 320 megabytes (MB), the average autistic biological dataset's BAM file size equal to 1 gigabyte (GB), and the average GOV2 biological dataset's BAM file size equal to 6.7 GB. The processing time of the rate-limiting steps were nearly linear functions of BAM file size across the biological datasets (linear regression: R^2 =0.985 (autism virome), R^2 = 0.9564 (GOV2); Fig. S5A&C). The synthetic dataset contained some samples with many more genomes present than others, which resulted in runtimes dividing into two distinct groups. This demonstrates an additional effect of community complexity on MetaPop's runtime. However, BAM file size was still linearly correlated with runtime within each group (linear regression: R^2 = 0.885 (synthetic group 1) and R^2 = 0.841(synthetic group 2); Fig. S5B).

We were next curious about how many of the computational resources were consumed of the memory supplied. The maximum memory consumption for the synthetic dataset was 5.75GB on 4 cores, 7.42 GB on 6 cores, and 10.82 GB on 14 cores. For the biological dataset, MetaPop utilized 3.94 GB of RAM on 4 cores, 5.39 GB on 6 cores, and 12.43 GB on 14 cores. Overall, these results show that MetaPop can be successfully run using low computational resources and will adjust the resources consumed based on the computational power supplied for datasets with average BAM sizes around 1GB, but will

need more computational resources for large datasets like GOV2.

Limitations and future directions

While MetaPop provides the sort of ease-of-use and scalability that we hope will open up microdiversity analyses to more researchers, our current implementation will benefit from future improvements. *First*, MetaPop was designed and optimized for single-contig genomes, so it does not work with binned-contig genomes and will treat each new contig as a different population. Nonetheless, it is possible to derive macro- and microdiversity calculations across binned contigs. The average read depth and nucleotide diversity per position for each contig per BAM file is output, so it is possible to derive the macrodiveristy abundance tables and microdiversity π values for binned contigs from MetaPop output. *Second*, MetaPop's default settings are optimized for short-read datasets. As more hybrid sequencing and assembly efforts are used to explore microbial and viral communities (e.g., [30, 31]), which capture more niche-defining hypervariable regions, adjustments for MetaPop's abundance calculations and SNP calling will need to be done. Though not prohibitive, this will require benchmarking studies that assess the nuances of new sequencing technology (e.g., homopolymers for nanopore sequencing) to correct for per base pair sequencing errors against the background of real biological mutations, many of which are now emerging (e.g., [80]). Finally, MetaPop is benchmarked for studying community and population-level diversity and selection, we have not optimized it for resolving strain-level genotypes. However, as other tools (e.g., InStrain; [29]) solve the problem of reconstructing strains from metagenomes, the resultant genotypes could be input to MetaPop.

Conclusions

MetaPop is a fast and scalable pipeline for the analyses of both macro- and microdiversity in metagenomic data. It combines both classical community ecology metrics with the full suite of population genetics parameters in a single integrated pipeline. While many of its functions are already available in existing pipelines [29, 39, 42, 44], MetaPop's easy user interface (i.e., single-line command) and ability to be run on a standard laptop for smaller datasets make it a practical choice for non-bioinformaticians and microbiology labs without access to large supercomputers. Further, MetaPop's default visualizations enable fast and easy interpretation of the results.

Molecular biology and sequencing technology advances have advanced the microbiologist's toolkit from 16S rRNA gene analyses to metagenomics and changed questions we could ask from "who is there" to also add "what could they be doing" and "with whom might they interact". Now, with further technological advances and by democratizing microdiversity analyses, we open a new window into the study of complex communities such that we can now ascertain "what populations have high microdiversity levels" and "which genes are under selection." While studying microdiversity has been hard due to lack of data and ease of toolkit, we are entering an era where such data and toolkits are available such that our understanding of these new biological windows will provide new insights into how complex systems work. MetaPop gives scientists an ideal toolkit to explore the dual impact of macro- and microdiversity across microbially impacted ecosystems.

Materials and methods

Preparing the mock and biological dataset input files for MetaPop

We chose three previously published datasets, a synthetic dataset representing mock bacterial communities [34] and two biological virome datasets including 131 viromes from the Global Oceans Virome 2 (GOV2) datasets [20] and gut viromes from autistic children that underwent FMT and their neurotypical peers [67], to evaluate MetaPop. The synthetic dataset was composed of known proportions of three distinct strains of *S. aureus* (ST5, ST8, and ST30), three distinct strains of *S. epidermidis* (TAW60, CV28, 1290N), and a single strain of *B. subtilis* [34]. Because MetPop explores inter- and intra-population-level analyses, we selected one strain of *S. aureus* (ST5 strain ECT-R2; GenBank: NC_017343.1), one strain of *S. epidermidis* (TAW60: binned assembly from [81]), and one strain of *B. subtilis* (strain 168; GenBank: NC_000964.3) as the population-level genome representatives. For the GOV2 dataset, we used the *Tara Oceans* 131 viromes from GOV2 and did not process the *Malaspina* viromes. The 488,130 viral populations identified in [20] were used as the reference genomes. For the biological gut virome dataset which comprised 49 gut viromes, we used the gut viral database from the *bioRxiv* version of [20] as the population-level reference genomes. Reads from both the mock and biological datasets were non-deterministically read mapped with a maximum fragment length of 2000 to their respective reference genomes using bowtie2 [82]. The resulting BAM files, the reference genomes, and read counts for each metagenome were used as input for MetaPop. MetaPop was run using the default settings if not otherwise noted.

Evaluating the processing time, computational resource consumption, and scalability of MetaPop

In order to computationally evaluate and benchmark MetaPop, we emulated the resource environment of

several likely computational platforms and attempted to process both the synthetic and biological datasets using these resources. The average BAM file size of the mock synthetic dataset BAM is 320 MB and has 30 samples, and of the biological dataset BAM is 1013 MB, with a total of 49 samples. The chosen computational scales reflect a fairly typical laptop computer, with 4 processing cores and 8 GB of RAM, a desktop computer with 6 cores and 16 GB or RAM, and a supercomputing environment using all 14 cores on one of the Ohio Super Owens (OSC; [84]) nodes and 64 GB of RAM.

The Owens nodes are each equipped with a Intel Xeon e5 2680 v4 Broadwell processor, which has 14 cores, and each node shares identical RAM and data storage characteristics. This provides parity between the differing scales of the computing environment, rendering maximum permissible memory and allocated cores the sole difference affecting runtime and memory usage. Finally, the OSC process manager terminates the execution of code which exceeds the supplied memory of any given job, meaning that exceeding specified RAM results in a failure of MetaPop to complete, just as it would in an environment actually limited by such resources. While the manager would also terminate a job which exceeded a specified runtime, we supplied each instance with excessive runtime so that this would not be a factor.

As it runs, MetaPop outputs timings for its five core components, namely preprocessing, SNP calling/refinement, linked SNP read mining and linkage calculations, calculation of microdiversity, and calculation of macrodiversity. In addition to the per-component timings, MetaPop outputs timings for each individual sample in sections of the code where each file is processed independently from the others. In both cases, these outputs include a date and time of start and finish for each step, and provide accurate timing for both the overall runtimes of each processing phase and for the time needed to run individual samples through preprocessing and SNP calling. MetaPop's overall runtime was calculated as a simple sum of the per-component runtimes.

In order to profile memory usage during the various phases of MetaPop, we relied on computational resource logs produced by the Ohio Supercomputer. These files report a variety of computational resource consumption statistics associated with a particular task, which includes the peak memory footprint for any collection of processes contained within a single job on the supercomputer. This approach answers the most pertinent question for users: what is the minimum RAM that is required to run a dataset of a given scale through MetaPop with a particular number of cores.

Mock community macrodiversity validation

In order to assess MetaPop's ability to resolve macrodiversity, we compared MetaPop's predicted relative abundances to the known relative abundances in the 30 mock communities [34]. Importantly, some of the reference genomes for the mock communities were not closed genomes. Thus, in order to calculate the raw abundances, the mean coverage across all base pairs in all fragments within the reference gene were calculated, excluding coverages below the 10th and above the 90th percentile per base pair. These values were then scaled as described in the "Macrodiversity analyses" above to create the normalized abundances. Prior to comparing the relative abundances, MetaPop's calculated normalized abundances were converted into relative abundances by dividing each population's normalized abundance in a community by the sum of all the population's normalized abundances. The fold change difference between MetaPop's calculated observed relative abundances and known relative abundances was then assessed using "foldchange" in the R package "gtools." Next, the calculated macrodiversity α (Richness, Shannon's *H*, and Peilou's *J*) between the observed MetaPop values and the expected actual values across the 30 mock communities were compared using Wilcoxon tests in the R package "ggpubr." β-(Bray-Curtis dissimilarity) diversity calculated distances calculated across all 30 communities then compared using a Mantel's test in the R package "vegan." Lastly, FastANI [83] using default settings was used to compare average nucleotide identity across the different strains and species within the mock community.

Mock community codon bias analyses

We chose to evaluate MetaPop's ability to pull out genes with different codon bias usage by examining the *Staphylococcus aureus* strains in the 30 mock communities [34]. *S. aureus* are well studied to their clinical relevance and there is a great deal of knowledge about the different genes and elements that have been horizontally acquired within their genomes [69, 70] and, to some extent, genes with increased expression (Malachow et al. 2011, [71, 74]). In order to assess the codon bias usage outlier part of MetaPop, we manually curated a list of known horizontally transferred and highly expressed genes and compared it to the list of genes with different codon bias usage predicted by MetaPop. We then calculated what proportion of the predicted genes with different codon bias usage were known to be horizontally acquired or highly expressed.

Global Oceans Virome 2 microdiversity validation

In order to assess MetaPop's ability to resolve microdiversity values and patterns, we computed the microdiveristy

(average π) per sample in the 131 GOV2 samples based on MetaPop's calculated nucleotide diversity (π) and compared it to the published microdiversity values [20]. To calculate microdiveristy for each sample, average π was calculated by randomly selecting the π values of 100 viral populations and then averaging their values (sensu [20]). This was repeated 1000× and the average of all 1000 subsamplings was used as the final average microdiversity value for each sample. The values were plotted using the line graph and scatter plot functions in Excel. The linear regressions were also run in Excel. The sample microdiversity values were then grouped by ecological zone as defined in [20] and, unlike the original analyses, the values were not subsampled from each ecological zone and then averaged in order to see the better spread of the values per zone. The ecological zone values were plotted and statistical differences assessed using the R package "ggboxplot." GOV2 SNPs were originally locally called using a PHRED $\geq$ 30. We ran MetaPop using a PHRED $\geq$ 20 (MetaPop's default) and $\geq$ 30 to filter the variants and then assessed SNP calls both globally (for PHRED $\geq$ 30 and $\geq$ 20) and locally (for PHRED $\geq$ 30). The fold change difference between MetaPop's average π values and the original GOV average π values were then assessed using "foldchange" in the R package "gtools."

Biological dataset microdiversity analyses

To compare macro- and microdiversity in the autism virome dataset [67], the predicted *macro*diversity α (Shannon's *H*) values, the microdiversity (average π), and the percentage of genomes under positive selection for all the autistic and neurotypical children prior to FMT treatment were compared using Wilcoxon tests in the R package "ggpubr." To calculate microdiveristy for each sample, average π was calculated by randomly selecting the π values of 10 viral populations and then averaging their values (*sensu* [20]). This was repeated 50× and the average of all 50 subsamplings was used as the final average microdiversity value for each sample. To calculate the percentage of genomes under positive selection per sample, viral populations with at least one gene detected under positive selection (pN/pS > 1) per sample were determined and pooled with the viral populations with enough coverage to analyze microdiversity. Similarly, to average π, the 10 viral populations per sample were randomly selected and then the percentage that were detected to be under positive selection assessed. This was repeated 50× and the average of all 50 subsamplings was used as the final average percentage of viral populations detected under positive selection for each sample. The pre- and post-FMT viromes of the autistic children were then plotted using the R package "ggplot2" and the fold change difference was assessed using "foldchange" in the R package "gtools."

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40168-022-01231-0.

Additional file 1: Figure S1. Heatmap showing % average nucleotide identities (ANI) similarities among the different strains and populations in the 30 mock communities. **Figure S2.** Validating MetaPop's *macro*diversity and codon bias analyses. (A) Tornado plot showing the relative abundances of *Staphylococcus aureus*, *Staphylococcus epidermidis*, and *Bacillus subtilis* across the 30 mock communities in the actual synthesized community and as determined by MetaPop. Bar charts contained within the gray bar to the left of the tornado plot reveal the number strains per each bacterial species, with three being the highest number of strains per species. (B) Boxplots showing median and quartiles of different α-diversity indices (richness, Shannon's H, and Peilou's J) compared between the actual and MetaPop derived abundances. The Wilcoxon test *p*-values above are the result from comparing actual and MetaPop derived α-diversity indices. (C) Heatmaps of β-diversity Bray-Curtis dissimilarity distances calculated using the actual and MetaPop derived abundances. **Figure S3.** Genome map of genes with outlier codon usage in ST5 *Staphylococcus aureus* ECT-R2. **Figure S4.** Validating MetaPops's microdiversity analyses using the Global Oceans Virome 2 dataset. (A-E, right) Line plots sorted by the original average nucleotide diversity (π) values from [20] and (A-E, left) scatter plots comparing the average π for the *Tara* Oceans stations in the GOV2 dataset derived from the (A) original GOV2 values versus MetaPops's PHRED≥30 local SNP calls, (B) original GOV2 values versus MetaPops's PHRED≥30 global SNP calls, (C) original GOV2 values versus MetaPops's PHRED≥20 global SNP calls, (D) MetaPops's PHRED≥20 global SNP calls versus MetaPops's PHRED≥30 global SNP calls, and (E) MetaPops's PHRED≥30 global SNP calls versus MetaPops's PHRED≥30 local SNP calls. The dashed line in the scatter plot represents the linear regression. (F, left to right) Bar plots showing the biological *micro*diversity trends across the ecological zones defined in [20] derived from the original GOV2 values, MetaPops's PHRED≥20 global SNP calls, MetaPops's PHRED≥30 global SNP calls, and PHRED≥30 local SNP calls. **Figure S5.** Scatterplots with Loess smoothing displaying runtime per sample for the rate-limiting part of MetaPop (i.e. pre-processing and the SNP calling section of microdiversity) as a factor of file size in megabytes on (left and right) the biological datasets and (center) the synthetic dataset.

Additional file 2: Table S1. Mock communities data actual and metapop derived abundances. **Table S2.** Biological virome dataset population abundances. **Table S3.** Full codon bias usage results for all genes in *Staphylococcus aureus* ECT-R2.

Acknowledgements

Pipeline design and discussion with Sergei Solonenko and Cesar J. Ignacio Espinoza is gratefully acknowledged. For help with digging into the *Staphylococcus* literature, we would like to thank Rodrigo Bacigalupe and Amy Richards. Lastly, we would like to thank the Sullivan lab and Konstantinidis lab members for testing the pipeline and providing input during the development of MetaPop.

Authors' contributions

A.C.G. and K.G. designed and coded the MetaPop pipeline. A.C.G., K.G., Z.Z., B.B., B.T., K.T.K, and M.B.S created the study design, analyzed the data, and wrote the manuscript. All authors approved the final manuscript.

Funding

Computational support was provided by an award from the Ohio Supercomputer Center (OSC) to MBS. Funding was provided by the Gordon and Betty Moore Foundation (#3790 to MBS), the U.S. Department of Energy (#DE-SC0020173 to MBS), the US National Science Foundation (OCE#1536989, OCE#1829831, and ABI#1759874 to MBS, and ABI#1759831 to KTK), and a National Institutes of Health T32 training grant fellowship (AI112542 to ACG).

Availability of data and materials

MetaPop is available for download at https://github.com/metaGmetapop/metapop. MetaPop can also be accessed and used as a GUI on iVirus on

Cyverse (https://de.cyverse.org/apps/agave/MetaPop-1.0.0). The synthetic datasets used for this study are available at: http://figshare.com/articles/Benchmarking_data_for_bacterial_strain_identification/1617539. The 131 GOV2 viromes used in this study can be downloaded from the European Nucleotide Archive (ENA) under the accession numbers found in Supplementary Table 3 of Gregory et al. 2019. The virome datasets used for this study are available in iVirus at the following link: http://mirrors.iplantcollaborative.org/browse/iplant/home/shared/iVirus/ABOR. Support for the pipeline is available on the issue tab of the github page.
Project name: MetaPop
Project home page: https://github.com/metaGmetapop/metapop
Operating system(s): Unix or Linux system
Programming language: Python and R
Other requirements: Samtools, BCFTools, Prodigal
License: GNU GPL
Any restrictions to use by non-academics: None

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Microbiology, Ohio State University, Columbus, OH 43210, USA. [2]Present Address: Department of Microbiology and Immunology, Rega Institute for Medical Research, VIB-KU Leuven Center for Microbiology, Leuven, Belgium. [3]School of Biological Sciences, Georgia Institute of Technology, Atlanta, Georgia, USA. [4]Byrd Polar and Climate Research Center, Ohio State University, Columbus, OH 43210, USA. [5]School of Biosciences, University of Exeter, Exeter, UK. [6]School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA. [7]Center of Microbiome Science, Ohio State University, Columbus, OH 43210, USA. [8]Department of Civil, Environmental and Geodetic Engineering, Ohio State University, Columbus, OH 43210, USA.

Received: 14 November 2021 Accepted: 29 November 2021
Published online: 15 March 2022

References
1. Falkowski PG, Fenchel T, Delong EF. The microbial engines that drive Earth's biogeochemical cycles. Science. 2008;320(5879):1034–9.
2. Kinross JM, Darzi AW, Nicholson JK. Gut microbiome-host interactions in health and disease. Genome Med. 2011;3(3):14.
3. Knight R, Callewaert C, Marotz C, Hyde ER, Debelius JW, McDonald D, et al. The microbiome and human biology. Annu Rev Genomics Hum Genet. 2017;18:65–86.
4. Parks DH, Chuvochina M, Chaumeil PA, Rinke C, Mussig AJ, Hugenholtz P. A complete domain-to-species taxonomy for Bacteria and Archaea. Nat Biotechnol. 2020;38(9):1079–86.
5. Roux S, Adriaenssens EM, Dutilh BE, Koonin EV, Kropinski AM, Krupovic M, Kuhn JH, Lavigne R, Brister JR, Varsani A, Amid C. Minimum information about an uncultivated virus genome (MIUViG). Nature biotechnology. 2019;37(1):29-37.
6. Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc Natl Acad Sci U S A. 1985;82(20):6955–9.
7. Woese CR. There must be a prokaryote somewhere: microbiology's search for itself. Microbiol Mol Biol Rev. 1994;58(1):1–9.
8. Koh A, De Vadder F, Kovatcheva-Datchary P, Bäckhed F. From dietary fiber to host physiology: short-chain fatty acids as key bacterial metabolites. Cell. 2016;165(6):1332–45.
9. Roager HM, Hansen LB, Bahl MI, Frandsen HL, Carvalho V, Gøbel RJ, et al. Colonic transit time is related to bacterial metabolism and mucosal turnover in the gut. Nat Microbiol. 2016;1(9):1–9.
10. Woodcroft BJ, Singleton CM, Boyd JA, Evans PN, Emerson JB, Zayed AA, et al. Genome-centric view of carbon processing in thawing permafrost. Nature. 2018;560(7716):49–54.
11. Solden LM, Naas AE, Roux S, Daly RA, Collins WB, Nicora CD, et al. Interspecies cross-feeding orchestrates carbon degradation in the rumen ecosystem. Nat Microbiol. 2018;3(11):1274–84.
12. Martinez-Guryn K, Hubert N, Frazier K, Urlass S, Musch MW, Ojeda P, et al. Small intestine microbiota regulate host digestive and absorptive adaptive responses to dietary lipids. Cell Host Microbe. 2018;23(4):458–69.
13. Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017;2(11):1533–42.
14. Delmont TO, Quince C, Shaiber A, Esen ÖC, Lee ST, Rappé MS, et al. Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. Nat Microbiol. 2018;3(7):804–13.
15. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell. 2019;176(3):649–62.
16. Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, et al. A new genomic blueprint of the human gut microbiota. Nature. 2019;568(7753):499.
17. Nayfach S, Shi ZJ, Seshadri R, Pollard KS, Kyrpides NC. New insights from uncultivated genomes of the global human gut microbiome. Nature. 2019;568(7753):505–10.
18. Emerson JB, Roux S, Brum JR, Bolduc B, Woodcroft BJ, Jang HB, et al. Host-linked soil viral ecology along a permafrost thaw gradient. Nat Microbiol. 2018;3(8):870–80.
19. Shkoporov AN, Clooney AG, Sutton TD, Ryan FJ, Daly KM, Nolan JA, et al. The human gut virome is highly diverse, stable, and individual specific. Cell Host Microbe. 2019;26(4):527–41.
20. Gregory AC, Zayed AA, Conceição-Neto N, Temperton B, Bolduc B, Alberti A, et al. Marine DNA viral macro-and microdiversity from pole to pole. Cell. 2019a;177(5):1109–23.
21. Gregory AC, Zablocki O, Zayed AA, Howell A, Bolduc B, Sullivan MB. The gut virome database reveals age-dependent patterns of virome diversity in the human gut. Cell host & microbe. 2020;28(5):724-40.
22. Camarillo-Guerrero LF, Almeida A, Rangel-Pineros G, Finn RD, Lawley TD. Massive expansion of human gut bacteriophage diversity. Cell. 2021;184(4):1098–109.
23. Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, Waller A, Mende DR, Kultima JR, Martin J, Kota K. Genomic variation landscape of the human gut microbiome. Nature. 2013;493(7430):45–50.
24. García-García N, Tamames J, Linz AM, Pedrós-Alió C, Puente-Sánchez F. Microdiversity ensures the maintenance of functional microbial communities under changing environmental conditions. ISME J. 2019;13(12):2969–83.
25. Cordero OX, Polz MF. Explaining microbial genomic diversity in light of evolutionary ecology. Nat Rev Microbiol. 2014;12(4):263–73.
26. Caro-Quintero A, Konstantinidis KT. Bacterial species may exist, metagenomics reveal. Environ Microbiol. 2012;14(2):347–55.
27. Gregory AC, Solonenko SA, Ignacio-Espinoza JC, LaButti K, Copeland A, Sudek S, et al. Genomic differentiation among wild cyanophages despite widespread horizontal gene transfer. BMC Genomics. 2016;17(1):1–3.
28. Bobay LM, Ochman H. Biological species in the viral world. Proc Natl Acad Sci U S A. 2018;115(23):6040–5.
29. Olm MR, Crits-Christoph A, Bouma-Gregson K, Firek BA, Morowitz MJ, Banfield JF. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. Nat Biotechnol. 2021:1–0.
30. Warwick-Dugdale J, Solonenko N, Moore K, Chittick L, Gregory AC, Allen MJ, et al. Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands. PeerJ. 2019;7:e6800.
31. Moss EL, Maghini DG, Bhatt AS. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. Nat Biotechnol. 2020:1–7.
32. Ahn T-H, Chai J, Pan C. Sigma: strain-level inference of genomes from metagenomic analysis for biosurveillance. Bioinformatics. 2015;31:170–7.
33. Hong C, Manimaran S, Shen Y, Perez-Rogers JF, Byrd AL, Castro-Nallar E, et al. PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. Microbiome. 2014;2:33.

34. Sankar A, Malone B, Bayliss SC, Pascoe B, Méric G, Hitchings MD, et al. Bayesian identification of bacterial strains from sequencing data. Microb Genom. 2016;2(8):e000075.
35. Albanese D, Donati C. Strain profiling and epidemiology of bacterial species from metagenomic sequencing. Nat Commun. 2017;8:2260.
36. Tamburini FB, Andermann TM, Tkachenko E, Senchyna F, Banaei N, Bhatt AS. Precision identification of diverse bloodstream pathogens in the gut microbiome. Nat Med. 2018;24(12):1809–14.
37. Luo C, Knight R, Siljander H, Knip M, Xavier RJ, Gevers D. ConStrains identifies microbial strains in metagenomic datasets. Nat Biotechnol. 2015;33:1045–52.
38. Sahl JW, Schupp JM, Rasko DA, Colman RE, Foster JT, Keim P. Phylogenetically typing bacterial strains from partial SNP genotypes observed from direct sequencing of clinical specimen metagenomic data. Genome Med. 2015;7:52.
39. Nayfach S, Rodriguez-Mueller B, Garud N, Pollard KS. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome Res. 2016;26(11):1612–25.
40. Quince C, Delmont TO, Raguideau S, Alneberg J, Darling AE, Collins G, et al. DESMAN: a new tool for de novo extraction of strains from metagenomes. Genome Biology. 2017;18(1):1–22.
41. Fischer M, Strauch B, Renard BY. Abundance estimation and differential testing on strain level in metagenomics data. Bioinformatics. 2017;33(14):i124–32.
42. Costea PI, Munch R, Coelho LP, Paoli L, Sunagawa S, Bork P. metaSNV: A tool for metagenomic strain level analysis. PLoS One. 2017;12(7):e0182392.
43. Smillie CS, Sauk J, Gevers D, Friedman J, Sung J, Youngster I, et al. Strain tracking reveals the determinants of bacterial engraftment in the human gut following fecal microbiota transplantation. Cell Host Microbe. 2018;23(2):229–40.
44. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ. 2015;3:e1319.
45. Scholz M, Ward DV, Pasolli E, Tolio T, Zolfo M, Asnicar F, et al. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nat Methods. 2016;13(5):435–8.
46. Greenblum S, Carr R, Borenstein E. Extensive strain-level copy-number variation across human gut microbiome species. Cell. 2015;160(4):583–94.
47. Konstantinidis KT, Ramette A, Tiedje JM. The bacterial species definition in the genomic era. Philos Trans R Soc Lond B Biol Sci. 2006;361(1475):1929–40.
48. Couto N, Schuele L, Raangs EC, Machado MP, Mendes CI, Jesus TF, et al. Critical steps in clinical shotgun metagenomics for the concomitant detection and typing of microbial pathogens. Sci Rep. 2018;8(1):1–3.
49. Nayfach S, Pollard KS. Toward accurate and quantitative comparative metagenomics. Cell. 2016;166(5):1103–16.
50. Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P. A bioinformatician's guide to metagenomics. Microbiol Mol Biol Rev. 2008;72(4):557–78.
51. Haegeman B, Hamelin J, Moriarty J, Neal P, Dushoff J, Weitz JS. Robust estimation of microbial diversity in theory and in practice. ISME J. 2013;7(6):1092–101.
52. Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol. 1985;2(1):13–34.
53. Lawrence JG, Ochman H. Molecular archaeology of the Escherichia coli genome. Proc Natl Acad Sci U S A. 1998a;95(16):9413–7.
54. Shin YC, Bischof GF, Lauer WA, Desrosiers RC. Importance of codon usage for the temporal regulation of viral gene expression. Proc Natl Acad Sci U S A. 2015a;112(45):14030–5.
55. Sharp PM, Li WH. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 1987a;15(3):1281–95.
56. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11(1):119.
57. Whitley E, Ball J. Statistics review 1: presenting and summarising data. Crit Care. 2001;6(1):66.
58. Knight R, Jansson J, Field D, Fierer N, Desai N, Fuhrman JA, et al. Unlocking the potential of metagenomics through replicated experimental design. Nat Biotechnol. 2012;30(6):513.
59. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
60. Danecek P, McCarthy SA. BCFtools/csq: haplotype-aware variant consequences. Bioinformatics. 2017;33(13):2037–9.
61. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061.
62. Delmont TO, Kiefl E, Kilinc O, Esen OC, Uysal I, Rappe MS, et al. Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. Elife. 2019;8:e46497.
63. Watterson GA. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 1975;7(2):256–76.
64. Nei M, Li WH. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci U S A. 1979;76(10):5269–73.
65. Wright S. The interpretation of population structure by F-statistics with special regard to systems of mating. Evolution. 1965;19(3):395–420.
66. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989;123(3):585–95.
67. Kang DW, Adams JB, Gregory AC, Borody T, Chittick L, Fasano A, et al. Microbiota transfer therapy alters gut ecosystem and improves gastrointestinal and autism symptoms: an open-label study. Microbiome. 2017;5(1):10.
68. Roux S, Emerson JB, Eloe-Fadrosh EA, Sullivan MB. Benchmarking viromics: an in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. PeerJ. 2017;5:e3817.
69. Malachowa N, DeLeo FR. Mobile genetic elements of Staphylococcus aureus. Cell Mol Life Sci. 2010;67(18):3057–71.
70. Alibayov B, Baba-Moussa L, Sina H, Zdeňková K, Demnerová K. Staphylococcus aureus mobile genetic elements. Mol Biol Rep. 2014;41(8):5005–18.
71. Karlin S, Mrázek J, Campbell A, Kaiser D. Characterizations of highly expressed genes of four fast-growing bacteria. J Bacteriol. 2001;183(17):5025–40.
72. Peschel A, Otto M. Phenol-soluble modulins and staphylococcal infection. Nat Rev Microbiol. 2013;11(10):667–73.
73. Li X, Gerlach D, Du X, Larsen J, Stegger M, Kühner P, Peschel A, Xia G, Winstel V. An accessory wall teichoic acid glycosyltransferase protects Staphylococcus aureus from the lytic activity of Podoviridae. Scientific reports. 2015;5(1):17219. https://doi.org/10.1038/srep17219.
74. Soufo HJ, Reimold C, Linne U, Knust T, Gescher J, Graumann PL. Bacterial translation elongation factor EF-Tu interacts and colocalizes with actin-like MreB protein. Proc Natl Acad Sci U S A. 2010;107(7):3163–8.
75. Bae W, Xia B, Inouye M, Severinov K. Escherichia coli CspA-family RNA chaperones are transcription antiterminators. Proc Natl Acad Sci U S A. 2000;97(14):7784–9.
76. Duval BD, Mathew A, Satola SW, Shafer WM. Altered growth, pigmentation, and antimicrobial susceptibility properties of Staphylococcus aureus due to loss of the major cold shock gene cspB. Antimicrob Agents Chemother. 2010;54(6):2283–90.
77. Hsiao EY. Gastrointestinal issues in autism spectrum disorder. Harv Rev Psychiatry. 2014;22(2):104–11.
78. Hughes AR, Inouye BD, Johnson MT, Underwood N, Vellend M. Ecological consequences of genetic diversity. Ecol Lett. 2008;11(6):609–23.
79. Tilman D, Isbell F, Cowles JM. Biodiversity and ecosystem functioning. Annu Rev Ecol Evol Syst. 2014;45:471–93.
80. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 2020;21(1):1–6.
81. Méric G, Miragaia M, de Been M, Yahara K, Pascoe B, Mageiros L, et al. Ecological overlap and horizontal gene transfer in Staphylococcus aureus and Staphylococcus epidermidis. Genome Biol Evol. 2015;7(5):1313–28.
82. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357.
83. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9(1):1–8.
84. Ohio Supercomputer Center. Ohio Supercomputer Center. Columbus: Ohio Supercomputer Center; 1987. http://osc.edu/ark:/19495/f5s1ph73

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.