Bioinformatics, 36(7), 2020, 1994–2000
doi: 10.1093/bioinformatics/btz866
Advance Access Publication Date: 25 November 2019
Original Paper

Genome analysis

SCGid: a consensus approach to contig filtering and genome prediction from single-cell sequencing libraries of uncultured eukaryotes

Kevin R. Amses*, William J. Davis and Timothy Y. James

Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, USA

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on May 3, 2019; revised on October 9, 2019; editorial decision on November 15, 2019; accepted on November 22, 2019

Abstract

Motivation: Whole-genome sequencing of uncultured eukaryotic genomes is complicated by difficulties in acquiring sufficient amounts of tissue. Single-cell genomics (SCG) by multiple displacement amplification provides a technical workaround, yielding whole-genome libraries which can be assembled *de novo*. Downsides of multiple displacement amplification include coverage biases and exacerbation of contamination. These factors affect assembly continuity and fidelity, complicating discrimination of genomes from contamination and noise by available tools. Uncultured eukaryotes and their relatives are often underrepresented in large sequence data repositories, further impairing identification and separation.

Results: We compare the ability of filtering approaches to remove contamination and resolve eukaryotic draft genomes from SCG metagenomes, finding significant variation in outcomes. To address these inconsistencies, we introduce a consensus approach that is codified in the SCGid software package. SCGid parallelly filters assemblies using different approaches, yielding three intermediate drafts from which consensus is drawn. Using genuine and mock SCG metagenomes, we show that our approach corrects for variation among draft genomes predicted by individual approaches and outperforms them in recapitulating published drafts in a fast and repeatable way, providing a useful alternative to available methods and manual curation.

Availability and implementation: The SCGid package is implemented in python and R. Source code is available at http://www.github.com/amsesk/SCGid under the GNU GPL 3.0 license.

Contact: amsesk@umich.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Contamination is an ever-present concern in the preparation of high-throughput sequencing libraries. Certain methods of sample preparation are more susceptible to contamination, whether it is from the laboratory or from the environment. Approaches that involve a non-specific amplification step, such as the multiple displacement amplification (MDA) associated with single-cell genomics (SCG), are especially prone to contamination from these sources. As DNA is amplified non-specifically, even small amounts of contamination, including that derived from the MDA reagents themselves, can lead to significant dilution of target molecules (Gawad *et al.*, 2016; Rinke *et al.*, 2014). Perhaps best known for its applications in model systems where it can capture cell-to-cell heterogeneity in molecular processes, SCG has also been leveraged toward generating genome-scale data for groups of uncultured bacteria, archaea, fungi and protozoans (Ahrendt *et al.*, 2018; Davis *et al.*, 2019; Gawryluk *et al.*, 2016; Mikhailov *et al.*, 2016; Rinke *et al.*, 2013; Roy *et al.*, 2014).

Uncultured microbes are those that cannot or have not been successfully grown axenically in pure laboratory cultures. Their study necessitates that tissues be collected directly from the environment or from highly mixed *in vitro* microcosms. Collecting ample material that is reasonably pure and yields sufficient quantities of DNA to serve as input for whole-genome sequencing is often a near insurmountable obstacle. While SCG techniques circumvent this obstacle through non-specific DNA amplification, the data they yield poses a unique set of bioinformatic challenges: (i) SCG is highly subject to contamination, making most, if not all, SCG-derived genomes of uncultured microbes mildly to moderately metagenomic (Davis *et al.*, 2019; Gawryluk *et al.*, 2016; Mikhailov *et al.*, 2016; Roy *et al.*, 2014); (ii) despite the capacity for fully factorial priming, the replication enzymes involved in MDA introduce amplification biases that eventually manifest as read libraries that do not accurately

represent the starting population of template molecules, making coverage statistics less reliable (Gawad *et al.*, 2016; Pinard *et al.*, 2006) and, (iii) uncultured organisms are often underrepresented in sequence databases, complicating taxonomic delineation from contaminants. Taken together, all of these factors make identification of the target genome from noise a major obstacle.

2 Approaches to isolating genomes from metagenomes

Methods for extracting individual genomes from metagenomic data are diverse. Utilizing features inherent to or derivative of nucleotide sequences, these approaches cluster contigs independent of any taxonomy assigned by BLAST searches of large sequence repositories (i.e. taxonomy-independent binning) (Sedlar *et al.*, 2017). Common features include the relationship between GC-content and coverage, *k*mer frequencies and relative synonymous codon usage (RSCU) (Dick *et al.*, 2009; Kumar *et al.*, 2013; Laczny *et al.*, 2015; McInerney, 1998; Mikhailov *et al.*, 2016; Sedlar *et al.*, 2017; Wu *et al.*, 2016). Despite being clustered independent of taxonomy, identification, selection and verification of clusters is almost always informed by assigned taxonomy (Dick *et al.*, 2009; Kumar *et al.*, 2013; Laetsch *et al.*, 2017; Mikhailov *et al.*, 2016).

2.1 The relationship between GC-content and coverage

GC-coverage-taxonomy (GCT) plots graph contigs as points in two dimensions, allowing visualization and separation of metagenomic assemblies into clusters based on the GC-content and sequencing depth (i.e. coverage) of their constituent contigs (e.g. Fig. 1) (Kumar *et al.*, 2013; Laetsch *et al.*, 2017). Since GC-content varies in between organisms and per-organism genome coverage is correlated with the relative abundance of fragments of its DNA in the sequencing library, these clusters can correspond to the genomes of individual organisms. Points are annotated with taxonomic information from nucleotide BLAST searches of large sequencing databases to determine the taxonomic affinities of clusters. Resolution depends on the complexity of the metagenome, the quality of the annotations and the phylogenetic distances between constituent genomes. GCT plots quickly visualize the 'metagenomic-ness' of assemblies and can be used to determine GC and coverage cutoffs for extracting particular clusters for independent processing and analysis (Kumar *et al.*, 2013; Laetsch *et al.*, 2017).

2.2 *k*mer frequencies

Separation of individual genomes from metagenomic backgrounds by *k*mer frequencies hinges on the assumption that the frequencies of specific oligonucleotide sequences of length *k* are internally consistent across each genome. Under this assumption, the frequencies of any particular *k*mer on assembled contigs that originate from the same genome will be similar, distributed around the frequency of that *k*mer in the entire genome. While *k*mers cannot be homogeneously distributed within genomes, *k*mer frequencies can be used to cluster metagenome assemblies and separate sets of contigs belonging to individual genomes (Dick *et al.*, 2009). One approach applies unsupervised machine learning to cluster a matrix of the relative frequencies of all informative *k*mers across a contig (its *k*mer profile) to generate emergent self-organizing maps (ESOMs) that visualize this *n*-dimensional data (Dick *et al.*, 2009; Ultsch and Moerchen, 2005). This approach yields a 2D topology that visualizes the boundaries between clusters, and theoretically, individual genomes (e.g. Fig. 1). Taxonomic annotations can be overlaid the final topology to predict the identity of clusters, which can then be carved out of the larger map by eye and analyzed independently (Dick *et al.*, 2009).

2.3 Relative synonymous codon usage

RSCU measurements are numerical representations of codon bias, describing the preferential use of different codons coding for the same amino acid (i.e. synonymous codons) in protein coding nucleotide sequences (CDS) (McInerney, 1998; Mikhailov *et al.*, 2016). As

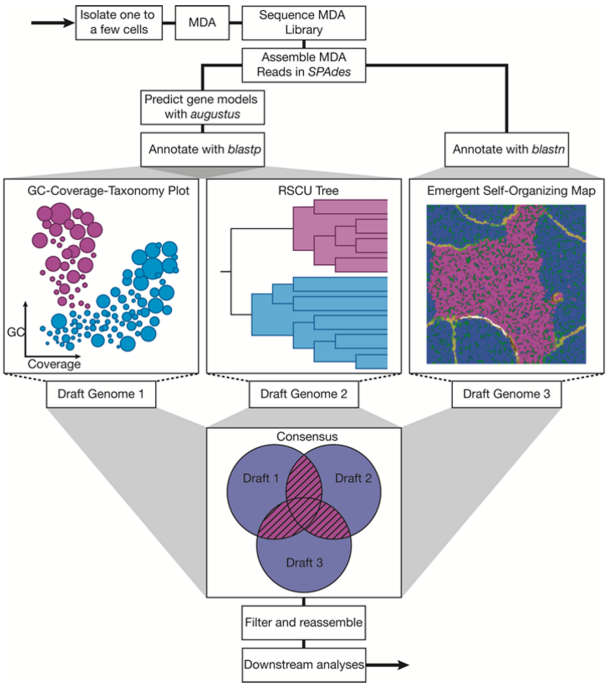

Fig. 1. Flow chart showing overview of the automated SCGid workflow, from isolation of one to a few cells of an uncultured eukaryote to a consensus-filtered assembly. The initial SCG assembly is annotated with predicted protein models and taxonomy based on BLAST searches. Three draft genomes are independently predicted by separate binning methods, representing each method's inference of whether each contig belongs in the target genome (purple) or not (blue). Consensus takes these three draft genomes and identifies their overlaps, generating a final filtered draft assembly by majority rule that is at the interstices of the three independent methods, averaging over variation and reinforcing confidence in the final position of contigs. Parameters affecting filtering decisions at each step are highly customizable and the SCGid workflow is built to be run iteratively (Color version of this figure is available at *Bioinformatics* online.)

codon bias is often species-specific, RSCU profiles represent another feature by which assembled contigs can be clustered and separated. Following protein annotation, coding portions of contigs are concatenated into a single joint CDS sequence for each contig, upon which whole-contig RSCU profiles are calculated. RSCU values for each of the 59 codons with alternative synonymous codons that are not STOP codons are calculated across the entire concatenate according to the generalized expression considering codon *i*...

$$RSCU_i = \frac{X_i}{\frac{1}{n}\sum_{i=1}^{n} X_i}$$

...where n is the number of codons synonymous to i and X_i is the number of occurrences of i in the concatenate (McInerney, 1998). These profiles are subsequently used to generate an RSCU distance matrix based on the generalized distance measure...

$$D_{jk} = \sum_{i=1}^{n} \frac{|RSCU_{ji} - RSCU_{ki}|}{n}$$

...where $RSCU_{ji}$ is the RSCU of codon i on CDS concatenate j, $RSCU_{ki}$ is the RSCU of codon i on concatenate k and n is the total number of synonymous codons in the concatenate (McInerney, 1998). Hierarchical clustering of RSCU matrices exposes clusters of contigs with similar profiles that can be assigned taxonomy by BLAST searches (Fig. 1). Clusters of contigs with known or inferred origin can be used as training sets in subsequent rounds of clustering by different features to retrieve the short, protein-less contigs that could not be included in the initial clustering (Mikhailov *et al.*, 2016).

3 Obstacles to filtering single-cell eukaryotic metagenomes

Methods currently available for separating metagenomes address some of the issues associated with filtering SCG assemblies, but there remain gaps in their ability to do so. While useful features for clustering are necessarily present in sequences from across the tree of life, the collection of tools that use them is generally skewed toward prokaryotes (Sieber *et al.*, 2018; Wu *et al.*, 2016). This limits the pool of available options when filtering SCG assemblies of uncultured eukaryotes. Tools that lean on contig coverage for clustering (Kumar *et al.*, 2013; Wu *et al.*, 2016) have considerably less utility with SCG because of the biased sequencing depth that characterizes *de novo* assemblies (Davis *et al.*, 2019; Pinard *et al.*, 2006). This bias tends to lead to *de novo* assemblies that are highly fragmented, introducing significant variance in contig-level sequence features (e.g. *k*mer frequencies) used for clustering and negatively affecting filtering outcomes (Davis *et al.*, 2019). Moreover, as the target organisms of SCG and their relatives are usually uncultured, contigs belonging to their genomes rarely share sufficient sequence similarity with those contained in public sequence repositories. This impairs the ability of BLAST searches to assign taxonomy for the vast majority of contigs, making annotation difficult.

Despite these obstacles, filtering SCG metagenomes of uncultured eukaryotes with available tools can yield draft genomes predicted to be nearly complete (Davis *et al.*, 2019; Mikhailov *et al.*, 2016). However, there often remains uncertainty in the fidelity of filtered drafts because verification by unified taxonomy or coverage information is difficult or impossible. Downstream analyses of these drafts are imbued with similar uncertainty when the inclusion or exclusion of a contig could arbitrarily introduce false negatives for genome functionality or attribute functionality that is derived from a contaminant.

4 SCGid: a consensus-based filtering tool for SCG of uncultured eukaryotes

To address this uncertainty and fully investigate the efficacy of different approaches in filtering *de novo* assemblies of single-cell sequencing libraries, we implemented three in SCGid, an automated filtering tool for SCG assemblies. SCGid filters assemblies separately using each approach described above, generating three intermediate drafts. A final consensus draft is generated by majority rule at the overlaps of different approaches, where inclusion of a contig is dependent on its inclusion in two of the three intermediate drafts (pipeline summarized in Fig. 1). Each filtering approach, including consensus, is invoked separately from the CLI. Module-specific implementations, discussed in the coming sections, introduce novel code automating all but a single step of the tripartite pipeline. Automation is enabled by SCGid's requirement of *a priori* specifications of 'target' taxa. This duality of 'target' and 'nontarget' taxonomic annotations is hereafter referred to as such. To reduce computational time spent assigning taxonomy, SCGid uses the Uniprot swissprot database (SPDB) for protein sequences and the full NCBI nt database for nucleotide sequences (NCBI Resource Coordinators, 2017; The UniProt Consortium, 2017). To increase coverage of non-model lineages, utilities are included to supplement the SPDB with additional protein sequences.

4.1 GCT plots (SCGid gc-cov)

SCGid plots BLAST-annotated AUGUSTUS-predicted (Stanke and Morgenstern, 2005) proteins as points in GC-coverage space and draws a total of 13 separate flexible selection windows (FSWs) around them (Fig. 2A). The 2D bounds of windows are calculated with respect to proteins that had a significant hit (*e*-value $\leq$ 1*e*-5, by default) in the SPDB. These bounds are used downstream to make inclusion decisions on contigs that either contain no proteins or contain proteins with no significant hit (i.e. unclassified contigs). All contigs identified as target by virtue of the sum and strength of their protein hits are *ad hoc* included in the GC-coverage-based filtered draft, by default. The flexibility of FSWs provides a unique SCG optimization as it allows for wide GC and coverage distributions, artifacts of highly fragmented assemblies and MDA amplification, respectively.

The bounds of FSWs are calculated through two sequential rounds of 1D expansion, one along each axis (e.g. round 1 along GC, round 2 along coverage). Beginning at the mean value of target points on that axis, expansion outward is incremental, proceeding

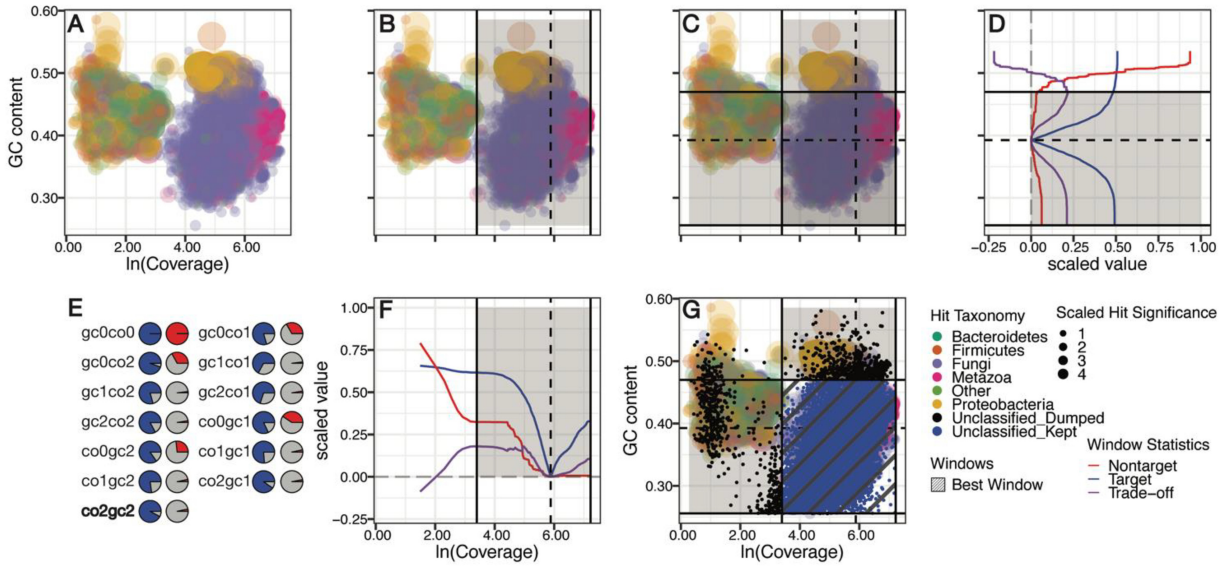

Fig. 2. Plots visualizing the process of 2D GC-coverage window expansion over SCG data for the zoopagalean fungus *Stylopage hadra* (Sh). (**A**) GCT plot generated by SCGid, points are AUGUSTUS-predicted proteins plotted by GC and ln(coverage) of the containing contig, colors represent phylum-level taxonomic classification and size represents strength of the hit. (**B, F**) First round of window expansion along the coverage axis using method 'co2', window arms (grey box in B, F) originate from the target mean (dashed line in B, F). The final window arms are defined by maximization of a trade-off value (purple line in F) which balances the proportions of target (blue line in F) and nontarget (red line in F) in the window as it expands. (**C, D**) Second round of window expansion, along the GC axis using method 'gc2'. (**E**) All 13 window expansion methods and associated P_{tar} (blue) and P_{ntar} (red) values for final windows; note that only method 'co2gc2' is shown in (A–D), (F) and (G). The optimal window (crosshatched box in G) is defined by maximization of P_{tar} below set P_{ntar} stringency threshold. (**G**) GCT plot (from A) now overlaid with all unclassified contigs (black and blue points) showing optimal final window (crosshatched box). Unclassified contigs falling within the optimal final window (blue) are included in the final genome while the rest (black) are discarded

to the limits of annotated points (Fig. 2B). The proportions of target and nontarget proteins inside versus outside the bounds are computed at each step, $P_{tar} = tar_{inside}/tar_{total}$ and $P_{ntar} = ntar_{inside}/ntar_{total}$, and used to calculate a trade-off value defined as $D_{tradeoff} = P_{tar}(P_{tar} - P_{ntar})$ (Fig. 2B and F). At the end of each round of expansion, bounds are set where D_{tradeoff} is maximized (Fig. 2F). The second round of expansion is identical to the first except that all points outside the bounds set in round 1 are ignored (Fig. 2C and D). The end product is a 2D FSW with cutoffs on both the GC and coverage axes (Fig. 2G, crosshatched region).

Accounting for all thirteen FSWs, to cope with dataset-specific distributional differences in GC-Coverage space, SCGid draws FSWs for all factorial combinations of first axis analyzed (GC or coverage) and three expansion types (unbounded =0, coupled =1 or uncoupled =2) (Fig. 2E). Unbounded (0) expansion rounds do not compute P_{tar}, P_{ntar} or D_{tradeoff} at all, merely setting the bounds at the limits of annotated points along that axis (gc0 or co0) (Fig. 2E). Coupled (1) expansion rounds compute P_{tar}, P_{ntar} or D_{tradeoff} once per step for the positive and negative directions taken together (gc1 or co1) (Fig. 2E). Uncoupled (2) expansion rounds compute P_{tar}, P_{ntar} or D_{tradeoff} twice per step for the positive and negative directions separately, allowing for unequal bound divergence from the mean (gc2 or co2) (Fig. 2A–G). From this set of FSWs (Fig. 2E), an optimal window is chosen that maximizes P_{tar}, but minimizes P_{ntar} at or below a set stringency level, s (i.e. $P_{\text{ntar}} \leq s$). As stated above, cutoffs defined by the optimal window determine the inclusion or exclusion of unclassified contigs (Fig. 2G, blue points in crosshatched region). All contigs identified as target are included, by default.

4.2 Emergent self-organizing maps (SCGid kmers)

SCGid provides automated preparation of all the files required to train and generate an ESOM topology using outside scripts and Databionics ESOM Tools (Dick *et al.*, 2009; Ultsch and Moerchen, 2005). SCGid introduces an automated annotation pipeline that links contigs with their best BLAST hit in the NCBI nt database, coloring them according to user-defined taxonomic levels. The task of sectioning-out a target cluster from the topology (using Databionics ESOM tools) relies on the user. An automated algorithm has not yet been implemented in SCGid and mouse-sectioning by human eye is standard practice (Dick *et al.*, 2009; Ultsch and Moerchen, 2005). Following sectioning and export, SCGid pulls the contigs belonging to the target class, yielding the ESOM-filtered draft assembly.

4.3 Relative synonymous codon usage (SCGid codons)

SCGid implements RSCU-based metagenome filtering in line with the concepts and applications described above (McInerney, 1998; Mikhailov *et al.*, 2016). CDS sequences are pulled from AUGUSTUS (Stanke and Morgenstern, 2005) models and joined into a single CDS concatenate for each contig. Short concatenates are discarded (<3000 bp, by default). RSCU profiles are calculated for large concatenates and used to compute an RSCU distance matrix (McInerney, 1998). A neighbor-joining tree is computed from this matrix, the tips of which are assigned taxonomy. The tree is iteratively searched, and all sufficiently sized clades (≥30 tips, by default) are binned by shared node architecture to avoid duplication. Clades in bins are ranked by the target-nontarget ratio of their descendant tips; ties are resolved by maximizing clade size. The highest-ranking clades from each bin are compared and the best clade, presumed to originate from the target genome, is nominated as a training set to collect small protein-less contigs from the rest of the metagenome. Clustering is done in ClaMs (Pati *et al.*, 2011), a *k*mer-based ($k = 2$, by default) binning algorithm that assesses contig similarity to the trainset (Pearson's distance ≤0.1) and bins them accordingly (Mikhailov *et al.*, 2016).

5 Validation

5.1 Methods

To assess the performance of our filtering implementations and the ability of consensus to resolve inconsistencies between them, we ran SCGid on two mock and three elsewhere-published SCG datasets (Davis *et al.*, 2019; Mikhailov *et al.*, 2016; Roy *et al.*, 2014).

5.1.1 Dataset selection

We generated two mock-MDA read libraries from the *Saccharomyces cerevisiae* S288C reference genome. We selected three studies where MDA was used to prepare sequencing libraries and that had the goal of generating draft genome sequences for one or more uncultured eukaryotes. To sample from a range of eukaryotic lineages, we selected two studies targeting fungi, one microsporidian (PRJNA321520) and five zoopagalean fungi (PRJNA451036) and one targeting a stramenopile (PRJNA244411) (Davis *et al.*, 2019; Mikhailov *et al.*, 2016; Roy *et al.*, 2014). The studies' filtering methods involved various tools and levels of scrutiny, sometimes implementing similar approaches to those implemented in SCGid, other times relying solely on taxonomy-dependent approaches.

5.1.2 Mock dataset preparation

We artificially contaminated the *S.cerevisiae* S288C reference genome with the green alga *Chlamydomonas reinhardtii* CC-503 cw92 mt+, and the bacteria *Bacillus cereus* ATCC 14579, *Cellulomonas* sp. FA1 GY42 and *Pseudomonas putida* KT2440 (Belda *et al.*, 2016; Cohen *et al.*, 2015; Fisk *et al.*, 2006; Ivanova *et al.*, 2003; Merchant *et al.*, 2007). To test SCGid on both eukaryotic and prokaryotic contamination, we generated two mock-MDA read libraries *in silico* that were either contaminated with just the bacteria (mockB) or the bacteria and *C.reinhardtii* (mockBE). To simulate biased and unequal coverage across the metagenome, we used bounded Brownian motion to generate unique discrete probability mass functions for each chromosome or contig that modulated the likelihood of each nucleotide being sampled as a start point for a 500 bp fragment (e.g. Supplementary Fig. S1). We sampled fragment start locations from these distributions and read 150 bp from both ends (i.e. paired-end), sampling to a mean expected coverage of 80× without simulating sequencing errors. In this way, we simulated the output of sequencing an MDA-derived library from three or four cells on the Illumina NextSeq platform. The mock metagenomes were assembled using SPAdes v3.9.0 (Bankevich *et al.*, 2012), yielding initial assemblies of 58.48 Mbp on 3102 contigs (coverage range: 2.45–17 369.63×, mean =60.04×) and 127.80 Mbp on 31 781 contigs (coverage range: 1.172–10 261, mean =134.26×) for mockB and mockBE, respectively, confirming that our fabricated SCG metagenomes were MDA-like (i.e. fragmented with wide coverage distributions). All contigs <200 bp were trimmed from initial assemblies prior to filtering. To simulate under-representation of *S.cerevisiae* during filtering, we manually purged the SPDB of all entries corresponding to the Saccharomycotina.

5.1.3 Genuine SCG dataset preparation

Since initial unfiltered assemblies are not usually made publicly available upon publication, we independently processed and assembled libraries of raw paired-end reads deposited in NCBI SRA according to the methods and parameters outlined by the authors (Mikhailov *et al.*, 2016; Roy *et al.*, 2014). Since we authored the study for the five zoopagalean fungi featured here, we worked directly with our initial assemblies (Davis *et al.*, 2019).

5.1.4 Analysis of filtering outcomes

We compared filtering outcomes to each other, to their corresponding consensus draft, and to the published assembly. Comparisons were made on the basis of cumulative assembly size, number of contigs and CEGMA/BUSCO completeness (Parra *et al.*, 2007; Waterhouse *et al.*, 2017). Where informative, we made whole-genome alignments in MUMmer v3.23 (Kurtz *et al.*, 2004) to quantify and visualize the proportion of the published assembly that was recapitulated in the SCGid consensus draft. For the mock datasets, we split the read libraries based on origin and mapped them to each

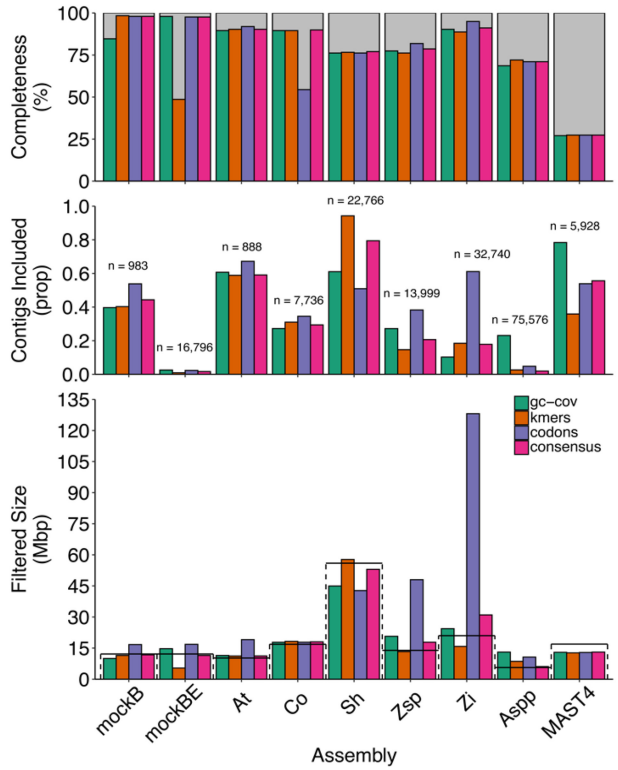

Fig. 3. Set of grouped bar charts showing variation in filtering outcomes of the three different filtering approaches implemented in SCGid (green, orange and purple bars) and the averaging effect of consensus (pink bars). Filtered assemblies were often different in terms of cumulative filtered assembly size (bottom), proportion of initial assembly contigs persisting into the filtered draft (middle) and predicted genome completeness (top). Cumulative assembly sizes of the references are shown as dashed bars in the lower pane. The total number of contigs in each initial assembly is shown above bars in middle pane. Abbreviations are as follows: mockB, mock with bacterial contamination only; mockBE, mock with bacterial and eukaryotic contamination; At, *Acaulopage tetraceros*; Co, *Cochlonema odontosperma*; Sh, *Stylopage hadra*; Zsp, *Zoopage* sp.; Zi, *Zoophagus insidians*; Aspp, *Amphiamblys* sp. and MAST4, MAST4-like strameopile (Color version of this figure is available at *Bioinformatics* online.)

assembly with BWA-MEM (Li, 2013) to quantify the respective contribution of yeast or contamination to filtered draft size.

5.2 Results

Automated filtering with SCGid yielded three filtered drafts and one consensus-filtered draft for each organism. In total, we generated 36 filtered assemblies for 9 target organisms. The filtered drafts predicted by separate approaches were often different, distinct in number of contigs, cumulative sequence length and predicted completeness (Fig. 3). Filtering with the consensus method applied by SCGid averaged sometimes dramatic variation where present, yielding conservative filtered drafts at the overlaps of different approaches and sometimes improving completeness (Fig. 3, pink bars; Supplementary Table S4). In general, SCGid consensus recapitulated the sequence content and genome size of reference genomes and published drafts (Fig. 3 bottom, dashed bars; Supplementary Table S4).

5.2.1 Mock *Saccharomyces cerevisiae* SCG metagenomes

Automated filtering of the two mock SCG metagenomes yielded *S.cerevisiae* consensus drafts that nearly recapitulated the size of the 12.16 Mbp S288C reference genome: 11.76 Mbp on 436 contigs and 11.47 Mbp on 280 contigs for mockB and mockBE, respectively. Individual filtering approaches commonly yielded different drafts compared to the reference or even the draft produced by that approach on the other mock (Fig. 3; Supplementary Table S4). GC-coverage (i.e. SCGid gc-cov) either over- or under-filtered (mockB: 10.01 Mbp on 390 contigs, mockBE: 14.7 Mbp on 432 contigs), *k*mer frequencies (i.e. SCGid kmers) over-filtered in both cases, dramatically so for mockBE (mockB: 11.47 Mbp on 396 contigs; mockBE: 5.4 Mbp on 158 contigs) and RSCU (i.e. SCGid codons) under-filtered both metagenomes, generating similarly sized drafts (mockB: 16.72 Mbp on 529 contigs; mockBE: 16.83 Mbp on 398 contigs). Consensus outperformed all three on the basis of closest cumulative sequence length.

SCGid consensus drafts were mostly composed of yeast sequence data with relatively small fractions of contamination. With only bacterial contamination included (i.e. mockB), SCGid kmers produced the best draft, with a 99.21–0.15% ratio of mapped reads originating from yeast versus contamination, compared to 98.69–2.62% for consensus (SCGid gc-cov: 82.78–2.62%; SCGid codons: 98.68–34.46%). With bacterial and eukaryotic contamination included (mockBE), consensus outperformed individual approaches with a 98.04–1.38% ratio (SCGid gc-cov: 98.32–9.08%; SCGid kmers: 48.04–2.8%; SCGid codons: 98.94–5.70%). Taken together, these results underpin the uncertainty in filtering SCG metagenomes using any one approach and demonstrate the benefits of consensus.

5.2.2 Five zoopagalean fungi

As we noted in the original publication, manually-applied consensus averaged variation among separate filtering approaches and reduced uncertainty in the final drafts (Davis *et al.*, 2019). Compared to those consensus drafts, automated SCGid filtering tended to increase assembly size and predicted completeness (Fig. 3; Supplementary Table S4). The filtered assemblies of *Zoopage* sp. (Zsp) and *Zoophagus insidians* (Zi) were significantly increased in size from 13.92 Mbp on 1958 contigs to 17.84 Mbp on 2892 contigs (SCGid gc-cov: 20.71 Mbp, 3809 contigs; SCGid kmers: 13.29 Mbp, 2056 contigs; SCGid codons: 48.01 Mbp, 5358 contigs) and from 21.01 Mbp on 2432 contigs to 31.01 Mbp on 5839 contigs (SCGid gc-cov: 24.37 Mbp, 3360 contigs; SCGid kmers: 15.83 Mbp, 6055 contigs; SCGid codons: 128.10 Mbp, 20 013 contigs), respectively. Those of *Acaulopage tetraceros* (At) and *Cochlonema odontosperma* (Co) were only marginally increased from 10.20 Mbp on 472 contigs to 11.20 Mbp on 525 contigs (SCGid gc-cov: 11.45 Mbp, 539 contigs; SCGid kmers: 11.20 Mbp, 523 contigs; SCGid codons: 19.10 Mbp, 597 contigs) and 16.84 Mbp on 1819 contigs to 18.05 Mbp on 2274 contigs (SCGid gc-cov: 17.81 Mbp, 2108 contigs; SCGid kmers: 18.26 Mbp, 2399 contigs; SCGid codons: 17.84 Mbp, 2670 contigs), respectively. Finally, the *Stylopage hadra* (Sh) assembly decreased in size from 55.96 Mbp on 20 112 contigs to 53.01 Mbp on 18 082 contigs (SCGid gc-cov: 44.96 Mbp, 13 902 contigs; SCGid kmers: 57.76 Mbp, 21 459 contigs; SCGid codons: 42.77 Mbp, 11 592 contigs).

Increases in assembly size were often accompanied by boosts in predicted completeness. Predicted completeness of At and Zsp were greatly increased from 83.06 to 90.32% and 71.77 to 78.63%, respectively. Co and Zi only saw marginal boosts from 89.52 to 89.92% and 90.73 to 91.13%, respectively. Consistent with a decrease in assembly size, predicted completeness of Sh was marginally decreased from 77.42 to 77.02% (Fig. 3; Supplementary Table S4).

5.2.3 *Amphiamblys* sp.

SCGid yielded a consensus draft of 6.09 Mbp on 1464 contigs, compared to the published assembly of 5.62 Mbp on 1727 contigs (Mikhailov *et al.*, 2016). The SCGid consensus draft was more similar in size to the published draft than those of separate approaches (SCGid gc-cov: 13.08 Mbp, 17 469 contigs; SCGid kmers: 8.60 Mbp, 1987 contigs; SCGid codons: 10.69 Mbp, 3628 contigs; Fig. 3, Supplementary Table S4).

In the original publication, completeness was estimated at ~90% with a custom microsporidian database of core eukaryotic genes in BUSCO v1.1b (Mikhailov *et al.*, 2016; Sima *et al.*, 2015). Unable to directly replicate the unpublished custom database, we instead compared completeness of both assemblies using the fungi_odb9 database in BUSCO v3.0.2 (Waterhouse *et al.*, 2017). Of the 290 core fungal genes in fungi_odb9, the SCGid assembly

contained 205 complete copies (70.69%) while the original published assembly contained only 193 complete copies (66.55%), equating to a 4.14% completeness advantage in favor of the SCGid assembly (Fig. 3, Supplementary Table S4).

Whole-genome alignment detected ~740 contigs with cumulative length 0.508 Mbp in the published assembly that was unaccounted for in the SCGid-filtered assembly and ~880 contigs with cumulative length 1.59 Mbp in the SCGid-filtered assembly that was unaccounted for in the published draft (Supplementary Fig. S2). These values indicate that the unaligned contigs were generally quite short. To confirm that alignments were not being made too liberally, we measured sequence similarity between the two drafts (Supplementary Fig. S3). When ordered by decreasing contig size, there is a general trend of decreased sequence identity toward the end of the published draft that we explain as variability in initial assemblies.

5.2.4 MAST-4-type stramenopile

The automated SCGid run yielded a final consensus draft of 13.08 Mbp on 3298 contigs compared to the published draft of 16.93 Mbp on 4611 contigs (Roy *et al.*, 2014). The SCGid consensus draft was most similar in size to the published draft (SCGid gc-cov: 12.98 Mbp, 4647 contigs; SCGid kmers: 12.71 Mbp, 2128 contigs; SCGid codons: 12.88 Mbp, 3195 contigs; Fig. 3; Supplementary Table S4). Predictions of genome completeness using the eukaryota_odb9 database (303 core eukaryotic genes) favored the published draft with 102 complete copies (33.66%) compared to 83 complete copies (27.39%) in the SCGid consensus draft. Whole-genome alignment with MUMmer identified 1803 contigs with a cumulative sequence length of 1.61 Mbp in the published draft that were unaccounted for in the SCGid consensus draft. The SCGid-predicted genome draft contained 172 contigs with a cumulative sequence length of 0.081 Mbp that were unaccounted for in the published draft.

6 Discussion

We demonstrate that the outcomes of filtering SCG metagenomes can vary dramatically with the particular approach taken. SCGid is a consensus filtering tool designed to address this problem. It brings automation to the process of filtering SCG metagenomes, offering an alternative to the time-consuming manual curation or strict BLAST-based filtering that are typical of most SCG projects to date. It is a fast and informative tool that quickly characterizes the landscape of SCG metagenomes and produces filtered drafts at the interstices of three different approaches.

We go on to show that SCGid successfully filters both genuine and fabricated SCG metagenomes. We demonstrate SCGid's ability to recover the well-known *S.cerevisiae* S288C reference genome from a significantly muddled background using databases simulating its novelty. We benchmark SCGid against filtering approaches used in the literature, where it recapitulates final genome size, content and completeness. For five zoopagalean fungi, SCGid generally predicted larger filtered drafts than those we previously published (Davis *et al.*, 2019). Compared to the published *Amphiamblys* sp. assembly, SCGid yielded a similarly sized draft that corresponds well to the published draft (Mikhailov *et al.*, 2016). While SCGid generated a smaller draft for the MAST-4-like stramenopile, it is not evident that any filtering was conducted in the original publication, indicating that perhaps SCGid filtered out previously-overlooked contamination (Roy *et al.*, 2014). In terms of predicted completeness, SCGid-filtered drafts landed on both sides of the line, overall tending to increase completeness: an average +2.91% for five zoopagalean fungi, +4.14% for microsporidian Am*phiamblys* sp. and -6.27% for a MAST-4-like stramenopile.

SCGid's consensus approach blends the outcomes of the filtering approaches it employs, leading to conservatism in contig inclusion decisions. We view this is as a beneficial trait as it protects against the over-inclusion of sequence data, the converse of which can lead to misrepresentations of biology as inferred from genome annotation and pollute public repositories with misidentified sequence data. While there is the potential for contigs that belong to be excluded by consensus, the majority of contigs that are selected against are either non-coding or of unknown function and do not usually contribute to predicted genome function or completeness. Further, consensus offers protection against the unstable behavior of individual approaches confronted with different metagenomic backgrounds. Given the fundamental reliance of these filtering approaches on sequence data, it is not surprising that decreasing phylogenetic distance between contaminants and target can obscure filtering outcomes. In filtering mock *S.cerevisiae* S288C SCG metagenomes, two of the three filtering approaches (SCGid gc-cov and SCGid kmers) yielded very different outcomes dependent on the inclusion of algal contamination. Encouragingly, despite over- or under-filtered intermediate drafts, the consensus outcome was similar to that reached from a solely bacterial background. We noted similar successful removal of rotifer contamination from the genuine *Zoophagus insidians* (Zi) SCG metagenome (Davis *et al.*, 2019). Taken together, these examples demonstrate moderate resilience of SCGid's consensus approach to both bacterial and eukaryotic contamination.

SCGid can yield draft genomes ready for downstream analyses or partial solutions in need of further manual curation (Davis *et al.*, 2019). This depends on the robustness of at least two of its integrated filtering approaches and the nature of planned downstream analyses. SCGid was conceived with these outcomes in mind. As such, it comes with a highly customizable set of options and utilities to augment the ways in which filtering decisions are made. SCGid can be iteratively rerun with different settings fast as it recycles the results of long-running steps. While the first run on an assembly can take 1–2 days, alternative filtered drafts can be produced by additional runs within minutes. An iterative SCGid workflow combined with tweaks to module and database configurations leads to increasingly refined filtering outcomes. By virtue of its consensus approach, SCGid has the potential to grow through the addition of novel filters leveraging variation in intergenic distance, intron length, etc.

SCG, despite its biases, weaknesses to contamination and inherent noise, generates genome-level sequence data for microbes that are inaccessible via standard approaches. This data contains fewer constituent genomes at higher coverage than analogous high-complexity metagenomes, but their identities are shrouded by unique biases. Where the goal of metagenomic binning may be the separation of many genomes, the goal of SCG is the separation of one or a few genomes from background contamination, endosymbionts and noise. This sets SCG apart from metagenomics and in turn sets SCGid apart from other tools. SCGid takes prior expectations of taxonomy into account, using it as a central driver of filtering outcomes. SCGid is not intended for use in determining community composition or isolating hundreds of genomes from soil samples, but for filtering the genomes of the uncultured targets of sequencing efforts where whole community sequencing and brute-force metagenomics is unfeasible or extraneous. SCGid is made for SCG and is capable of mitigating its downsides in a fast, automated, and repeatable way. As such, it wields potential to unlock genome-enabled biology for the innumerable uncultured eukaryotes that depend on SCG for the acquisition of genome-scale data.

Funding

This work was supported by the National Science Foundation [DEB1441677, DEB1441604]. K.R.A. was supported by the following National Institutes of Health training grant: "Michigan Predoctoral Training in Genetics [T32GM007544].

Conflict of Interest: none declared.

References

Ahrendt,S.R. *et al.* (2018) Leveraging single-cell genomics to expand the fungal tree of life. *Nat. Microbiol.*, **3**, 1417–1428.

Bankevich,A. *et al*. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol*., **19**, 455–477.
Belda,E. *et al*. (2016) The revisited genome of *Pseudomonas putida* KT2440 enlightens its value as a robust metabolic chassis. *Environ. Microbiol*., **18**, 3403–3424.
Cohen,M.F. *et al*. (2015) Genome sequence of the alkaline-tolerant Cellulomonas sp. strain FA1. *Genome Announc*., **3**, e00646–15.
Davis,W.J. *et al*. (2019) Genome-scale phylogenetics reveals a monophyletic Zoopagales (Zoopagomycota, Fungi). *Mol. Phylogenet. Evol*., **133**, 152–163.
Dick,G.J. *et al*. (2009) Community-wide analysis of microbial genome sequence signatures. *Genome Biol*., **10**, R85.
Fisk,D.G. *et al*. (2006) *Saccharomyces cerevisiae* S288C genome annotation: a working hypothesis. *Yeast*, **23**, 857–865.
Gawad,C. *et al*. (2016) Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet*., **17**, 175–188.
Gawryluk,R.M.R. *et al*. (2016) Morphological identification and single-cell genomics of marine diplonemids. *Curr. Biol*., **26**, 3053–3059.
Ivanova,N. *et al*. (2003) Genome sequence of *Bacillus cereus* and comparative analysis with *Bacillus anthracis*. *Nature*, **423**, 87.
Kumar,S. *et al*. (2013) Blobology: exploring raw genome data for contaminants, symbionts, and parasites using taxon-annotated GC-coverage plots. *Front. Genet*., **4**, 1–12.
Kurtz,S. *et al*. (2004) Versatile and open software for comparing large genomes. *Genome Biol*., **5**, R12.
Laczny,C.C. *et al*. (2015) VizBin - an application for reference-independent visualization and human-augmented binning of metagenomic data. *Microbiome*, **3**, 1–7.
Laetsch,D.R. *et al*. (2017) BlobTools: interrogation of genome assemblies. *F1000Res*., **6**, 1287–1216.
Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arxiv:1303.3997, 1–3.
McInerney,J.O. (1998) GCUA: general codon usage analysis. *Bioinformatics*, **14**, 372–373.
Merchant,S.S. *et al*. (2007) The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science*, **318**, 245–251.
Mikhailov,K.V. *et al*. (2016) Genomic survey of a hyperparasitic microsporidian *Amphiamblys* sp. (Metchnikovellidae). *Genome Biol. Evol*., **9**, 454–467.
NCBI Resource Coordinators (2017) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **45**, 12–17.
Parra,G. *et al*. (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, **23**, 1061–1067.
Pati,A. *et al*. (2011) ClaMS: a classifier for metagenomic sequences. *Stand. Genomic Sci*., **5**, 248–253.
Pinard,R. *et al*. (2006) Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics*, **7**, 216.
Rinke,C. *et al*. (2014) Obtaining genomes from uncultivated environmental microorganisms using FACS – based single-cell genomics. *Nat. Protoc*., **9**, 1038–1048.
Rinke,C. *et al*. (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, **499**, 431–437.
Roy,R.S. *et al*. (2014) Single cell genome analysis of an uncultured heterotrophic stramenopile. *Sci. Rep*., **4**, 1–8.
Sedlar,K. *et al*. (2017) Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. *Comput. Struct. Biotechnol. J*., **15**, 48–55.
Sieber,C.M.K. *et al*. (2018) Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol*., **3**, 836–843.
Sima,F.A. *et al*. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.
Stanke,M. and Morgenstern,B. (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res*., **33** (**Suppl. 2**), 465–467.
The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res*., **45**, D158–D169.
Ultsch,A. and Moerchen,F. (2005) ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM. *Technical Report*. Department of Mathematics and Computer Science, University of Marburg, Germany, 46.
Waterhouse,R.M. *et al*. (2017) BUSCO applications from quality assessments to gene prediction and Phylogenomics letter fast track. *Mol. Biol. Evol*., **35**, 543–548.
Wu,Y.-W. *et al*. (2016) MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, **32**, 605–607.