KombOver: Efficient k-core and K-truss based characterization of perturbations within the human gut microbiome

Nicolae Sapoval†, Marko Tanevski and Todd J. Treangen

Department of Computer Science, Rice University,
Houston, TX 77005, USA
†*E-mail: nsapoval@rice.edu*

The microbes present in the human gastrointestinal tract are regularly linked to human health and disease outcomes. Thanks to technological and methodological advances in recent years, metagenomic sequencing data, and computational methods designed to analyze metagenomic data, have contributed to improved understanding of the link between the human gut microbiome and disease. However, while numerous methods have been recently developed to extract quantitative and qualitative results from host-associated microbiome data, improved computational tools are still needed to track microbiome dynamics with short-read sequencing data. Previously we have proposed KOMB as a *de novo* tool for identifying copy number variations in metagenomes for characterizing microbial genome dynamics in response to perturbations. In this work, we present KombOver (KO), which includes four key contributions with respect to our previous work: (i) it scales to large microbiome study cohorts, (ii) it includes both k-core and K-truss based analysis, (iii) we provide the foundation of a theoretical understanding of the relation between various graph-based metagenome representations, and (iv) we provide an improved user experience with easier-to-run code and more descriptive outputs/results. To highlight the aforementioned benefits, we applied KO to nearly 1000 human microbiome samples, requiring less than 10 minutes and 10 GB RAM per sample to process these data. Furthermore, we highlight how graph-based approaches such as k-core and K-truss can be informative for pinpointing microbial community dynamics within a myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) cohort. KO is open source and available for download/use at: https://github.com/treangenlab/komb

Keywords: metagenomics; graph based methods; anomaly detection.

1. Introduction

Metagenomics, the study of the genomes of microbes that inhabit a microbiome, offers an unprecedented and highly granular view into the interaction between host-associated microbiomes and host disease phenotypes. Numerous computational tools now exist to uncover the taxonomic composition and functional profiles of human host associated microbiomes [1–4]. Of particular relevance to this work, higher taxonomic and functional diversity of the microbiota is associated with healthy individuals, while lower diversity correlates with disease states [5–8]. Furthermore, with the growing number of metagenome assembled genomes (MAGs) [9, 10] the association between the genomic composition of microbial communities and the host health

has become better quantified and understood [11, 12]. However, metagenomic assembly from short reads remains a challenge in highly repetitive regions of bacterial genomes [13–15], and among closely related strains of a given bacterial species [16, 17]. Genomic repeats arising from horizontal gene transfer or duplication events have been associated with bacterial adaptation and evolution [18–20], functional diversification [20], and pathogenesis [18]. Recent advances in long-read sequencing offer a path to resolution of complex inter- and intra-genomic repeats in microbial communities [21, 22]. However, limitations in high molecular weight DNA extraction [23] and financial cost of hybrid or high-quality long-read approaches poses a roadblock for large scale studies involving long-read sequencing. Additionally, a large existing corpus of metagenomic sequencing data consists predominantly of short paired-end reads, thus warranting the development of novel methods that can better capture and quantify inter- and intra-genomic repeat dynamics and flux.

To address this challenge, we have previously proposed the software KOMB [24] to extract high copy number sequences of potential biological significance in the microbial communities from the short paired-end read metagenomic sequencing data, expanding on prior approaches [25–27]. As the genomic diversity of a bacterial community has been correlated with host health, we hypothesize that the corresponding inter- and intra-genomic repeat structures can act as a "biomarker" for host health. Our prior work highlighted the ability of KOMB to detect shifts in the microbial community associated with antibiotic treatment and bowel cleanse, as well as identify associations between observed shifts and key bacterial members of pre- and post-FMT bacterial communities. Additionally, similarly to *de novo* assembly methods, KOMB is a database independent tool, and hence it avoids database biases [28]. However, unlike the common assembly approaches [29, 30] KOMB does not simplify the compacted de Bruijn graph, thus retaining the diversity originally present in the sequencing data. Furthermore, in contrast to k-mer profiling methods [1, 2], KOMB offers a set of genomic sequences that can be annotated for downstream analyses. Thus, KOMB bridges the gap between fast profiling methods that either require a database or do not yield sequence units that can be readily annotated, and computationally expensive assembly-based approaches.

For the purpose of identifying key sequences in the graph, KOMB employs the graph mining concept of k-core decomposition, which iteratively determines densely connected graph components. Previously, we had not investigated the set of sequences contained in the core of the graph as a whole, only focusing on sequences with high Core-A anomaly score [31] which captures deviations in coreness/degree ratios of a vertex. In KombOver (KO), we introduce and implement analysis of the maximal K-truss subgraph. Similarly to the vertices of the maximal k-core, the vertices of the maximal K-truss have been shown to have strong spreading (i.e. centrality) [32] which can be relevant in certain biological contexts as an alternative to betweenness centrality measure [25, 33].

One of the limitations of our prior work was its scalability to large metagenomic studies. In particular, the construction of the main data structure employed by KOMB, the hybrid unitig graph (HUG) incurred a high computational cost. It resulted in run times ranging from over an hour per single metagenomic sample, resulting in overwhelming computational costs for thousands to tens of thousands of samples. To address this limitation, in this work we pro-

pose a set of improvements to the HUG construction and analysis in KO aimed at enabling large-scale processing of genomic data and characterization of phenotype-associated dynamics. Furthermore, in addition to computational improvements, we provide a more extensive characterization of HUGs within the context of bacterial pangenomics, and draw parallels between pangenome graphs constructed from MAGs and HUGs. In order, to assess our tool, we analyzed short-read metagenomic sequencing data from a cohort of controls and patients with myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS), previously published by Xiong *et al.* [34]. Additionally, we have benchmarked KOMB on integrative human microbiome project [35] inflammatory bowel disease (IBD) cohort samples, as well as human genome sequencing data from Genome in a Bottle project [36], to demonstrate KO's scalability to both large data volumes, and complex repeat architectures.

2. Results

2.1. *Hybrid unitig graphs*

Hybrid unitig graphs (HUGs) are an extension of compacted de Bruijn graphs [37–39] used as a primary data structure in *de novo* de Bruijn graph assembly approaches [29, 30, 40]. The key addition in HUGs is presence of paired-end edges coming from the sequencing read data. Hence, while during assembly, the de Bruijn graphs are iteratively simplified to construct MAGs [29, 30] in KO denser and more complex HUGs are analyzed directly to facilitate the capture of repeat dynamics within microbial communities. Conversely, pangenome graphs are typically constructed from annotated genome assemblies, and capture high-level variation in synteny and copy numbers of gene clusters across related microbial genomes. In this context, HUGs bridge the gap between exact compaction achieved in the compacted de Bruijn graphs and high-level genomic variation representation of pangenome graphs [41–43].

Thus, compared to compacted de Bruijn graphs (Figure 1c) HUGs offer additional connectivity information based on local similarity and inferred adjacency between unitigs. Compared to the pangenome graphs, HUGs do not require neither complete genome assembly nor identification of putative gene clusters (Figure 1d) and hence can be constructed more efficiently from short paired-end read data.

2.2. *Analysis of an ME/CFS cohort*

First, we compared the overall distributions of the number of unitigs reconstructed from control samples with the ones from patients with short and long-duration ME/CFS. We observe that all samples in the control cohort contain more than 50,000 unitigs per sample, with 85 out of 92 samples containing between 50,000 and 400,000 unitigs (Figure 2). In contrast, 3 samples derived from patients with short-term ME/CFS contain less than 50,000 unitigs, and 62 out of 73 samples in this category contain up to 250,000 unitigs (Figure 2). Similarly, the data for long-term ME/CFS contains 6 samples with less than 50,000 unitigs, and 68 out of 73 the samples fall into the 0 to 300,000 unitigs range (Figure 2).

Next, we have designated unitigs with Core-A anomaly scores three standard deviations (3σ) above their corresponding sample's mean (μ) as the anomalous unitigs for the corresponding samples. We have explored the distribution of degrees (Figure 3A) in the anomalous unitigs

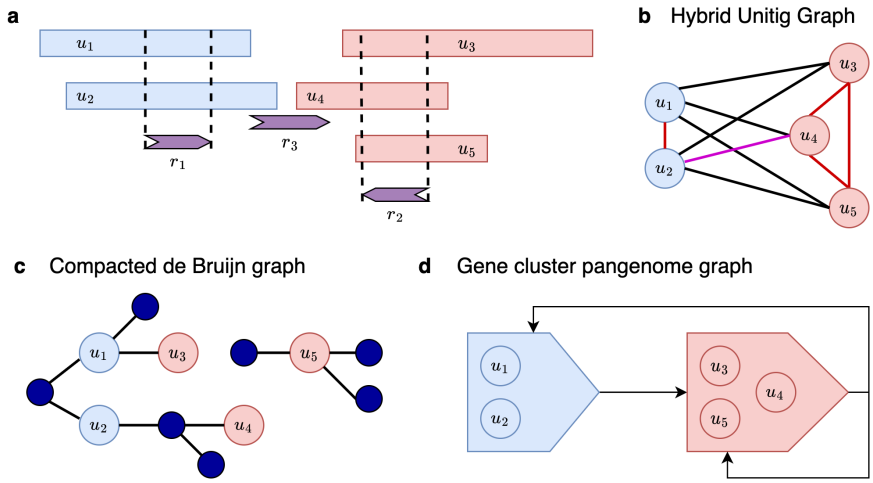

Fig. 1. **(a)** A set of 5 unitigs labeled u_1, u_2, u_3, u_4, and u_5 with the corresponding read mappings of r_1, r_2, and r_3. Note, that the read r_3 maps to both the end of unitig u_2 and the start of the unitig u_4. **(b)** A HUG corresponding to the unitigs and reads in (a). Edges marked in red are local similarity edges, and edges marked in black are adjacency edges arising from the paired reads (r_1, r_2). The magenta edge $\{u_2, u_4\}$ is an adjacency edge arising as a result of multi-mapping of a single read. **(c)** A schematic representation of a compacted de Bruijn graph. Dark blue nodes represent k-mers, while light blue and red nodes represent unitigs that have been compacted from unambiguous paths. **(d)** A schematic representation of a pangenome graph. Colored blocks represent gene clusters and arrows indicate possible paths through the gene sequences as indicated by corresponding genome assemblies.

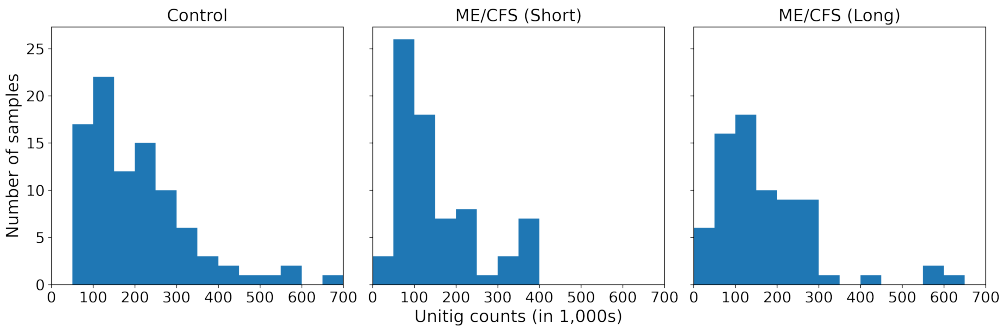

Fig. 2. Distribution of total unitig counts in HUGs constructed from control (left), ME/CFS short duration (center), and ME/CFS long duration (right) subjects's gut microbiome samples. Samples corresponding to short ME/CFS condition show lower absolute counts of unitigs, while those corresponding to the long ME/CFS are more similar to the controls. Both short and long ME/CFS associated samples have less high unitig count representatives.

based on the sample type and noted that the overall distributions are skewed to the left for all sample types. However, in the range of degrees from 250 to 1250, short-term ME/CFS samples exhibit sharper concentration towards the lower degree values than long-term ME/CFS and control samples. Additionally, in the 380-500 range of degrees long-term ME/CFS samples exhibit a more uniform distribution. The distributions in Figure 3A were tested for statistical

difference using Kolmogorov-Smirnov (KS) test. All three pairwise distribution comparisons were significant with p-value $< 10^{-9}$. Since the degree of a unitig in an HUG depends on the number of locally similar unitigs and potential genomic adjacencies of it, this indicates that long-term ME/CFS communities have more anomalous highly connected unitigs.

We also investigated the distribution of coreness values in the anomalous unitigs grouped by condition (Figure 3B). Similarly to the degree distributions for the low coreness values (0-100) all three sample types agree. Analogously, short-term ME/CFS samples also have the distribution of the coreness skewed towards lower values. This agrees with the observations in Figure 2 and Figure 3A, as the lower overall unitig count (and hence a smaller graph), and lower degrees (which provide an upper bound on coreness) would result in lower coreness values. Long-term ME/CFS samples have several peaks in the distribution (coreness 180-200, 280-320) that are not observed in the controls. The distributions in Figure 3B were tested for statistical difference using KS test. All three pairwise distribution comparisons were significant with p-value $< 10^{-9}$. Since coreness is a proxy for the level of interconnectedness in a group of unitigs, this can indicate the presence of clusters of unitigs corresponding to either a complex repeat architecture or a high abundance of closely related organisms.

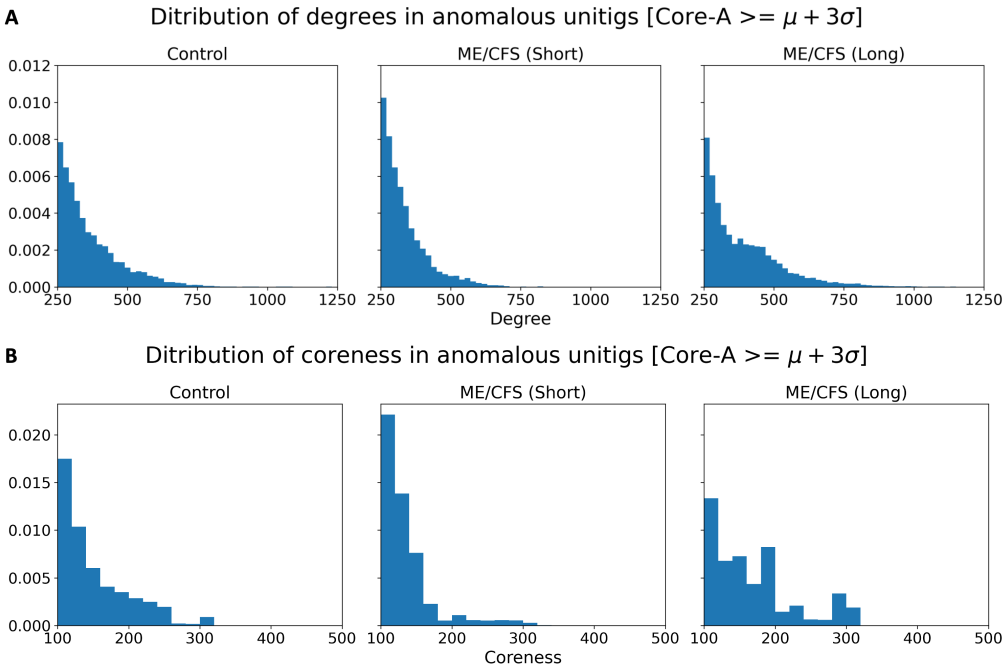

Fig. 3. **(A)** Distribution of the degrees of the unitigs that have Core-A anomaly score above $\mu + 3\sigma$ for their corresponding samples. Distributions for unitigs of degrees 0-250 are omitted for clarity. We note the samples corresponding to short-term ME/CFS have a distribution more skewed to the left. **(B)** Distribution of the coreness of the unitigs that have Core-A anomaly score above $\mu + 3\sigma$ for their corresponding samples. Distributions for unitigs of coreness 0-100 are omitted for clarity. Similarly to degree distribution, short-term ME/CFS samples show a skew towards lower coreness values. Additionally, long-term ME/CFS samples have more uniform distribution in the 100-200 range compared to controls and more unitigs in the 250-350 coreness range.

Next, we investigated the α-diversity at the species and genus level, as well as the overlaps between species and genus level classifications based on Kraken 2 [44] predictions for the anomalous unitigs in the three groups (Figure 4). We note that at both species and genus level the short-term ME/CFS samples exhibit a lower average α-diversity which can be indicative of dysbiosis. Additionally, at both species and genus levels, most taxonomic annotations are shared among the three cohorts. Still, the control group has consistently more unique taxa identified, further supporting the role of diverse microbial community composition in healthy individuals. Similarly, the long-term ME/CFS cohort has more unique taxa than the short-term ME/CFS cohort, indicating partial recovery from the dysbiosis.

We next compared the results for α-diversity and the overlaps obtained from anomalous unitigs, to the same information computed for the unitigs in the highest K-truss (Figure 5). We note that unlike in the case of general anomalous unitigs, those that belong to the highest K-truss show more similarity in the α-diversity between control and short-term ME/CFS samples, with long-term ME/CFS sample being the outlier (Figure 5A, B). Additionally, the total α-diversity in the trusses (Figure 5A, B) is noticeably lower than in general anomalous unitigs (Figure 4A, B). This is expected given trusses are densely connected subgraphs of a HUG, and hence have higher propensity to represent closely related genomic segments. Higher α-diversity in the long-term ME/CFS trusses can be a potential indicator for functional enrichment with multiple taxa coding for the same function in the long-term ME/CFS microbiota. Furthermore, we observed that, while the number of species and genera shared between all three cohorts makes up a smaller fraction of the total classifications. Namely, while species shared between all three categories make up 27.5% (1808 of 6567) of all species identified in the anomalous unitigs of the three cohorts (Figure 4C), they make up only 24.1% (177 of 735) of all species identified in the trusses of the samples (Figure 5C). Analogously, the shared genera make up 55.0% (1008 out of 1834) of all classifications for anomalous unitigs, and only 33.0% (156 out of 473) of all classifications for truss unitigs.

Additionally, when individual KO profiles are visualized for samples matched by age, gender, and race (Figure 6) we observe more compact profiles for the disease-associated samples. This matches the dysbiosis hypothesis, with the long-term ME/CFS sample showing a more complex profile than the short-term one. In all samples shown in Figure 6 unitigs with high anomaly scores are the ones for which the degree is larger than the expected coreness. This pattern occurs when a unitig is flanked by varying genomic contexts across the metagenome, and hence indicate unitigs with high inter- and intra-genomic copy numbers.

2.3. *Computational performance*

The k-core decomposition algorithm runs in $O(|V| + |E|)$ time [45] and hence scales linearly with the size of the graph. This scaling is particularly attractive in metagenomic communities, where the number of edges $|E|$ is proportional to the number of vertices $|V|$. In the case of more complex repeat architectures, such as Alu repeats in human genome, the number of edges is be proportional to the square of the number of vertices. Compared to Brandes's algorithm for betweenness centrality (a common algorithm for detecting influential nodes in a network) [46] k-core decomposition algorithm is significantly faster. The asymptotic time complexity of

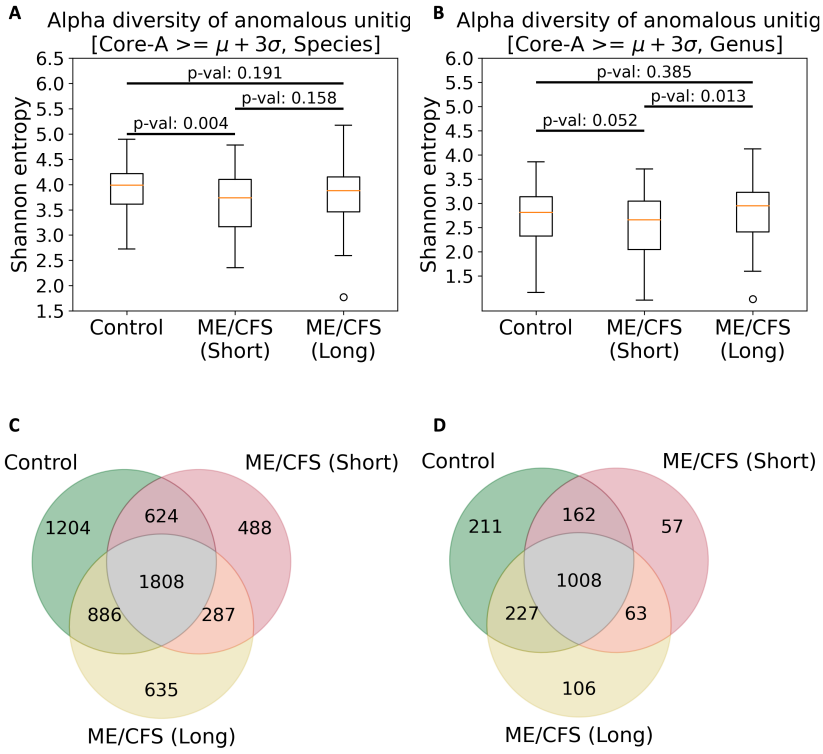

Fig. 4. **(A, B)** Distribution of alpha diversity (Shannon entropy) for anomalous unitigs that have anomaly score three standard deviations above the mean for the respective sample grouped by the condition and duration. p-values from Welch's t-test for equality of means are displayed above the boxplots. **(A)** Alpha diversity of samples associated with the ME/CFS condition is lower than that of control samples. Long ME/CFS samples on the other hand do not appear to be noticeably distinct from the control ones. Entropy was calculated based on the unitigs for which Kraken 2 provided species-level classification. **(B)** Alpha diversity of samples associated with the ME/CFS condition is lower than that of control samples and long ME/CFS samples. Long ME/CFS samples on the other hand do not appear to be noticeably distinct from the control ones. Entropy was calculated based on the unitigs for which Kraken 2 provided genus level classification (species annotations are rolled up into respective genus). **(C)** Venn diagram representing intersections between sets of species identified in the control, short ME/CFS, and long ME/CFS sample collections. **(D)** Venn diagram representing intersections between sets of genera identified in the control, short ME/CFS, and long ME/CFS sample collections.

Brandes's algorithm for unweighted graphs is $O(|E||V|)$, which even in the $|E| \sim \alpha|V|$ regime, leads to $O(|V|^2)$ complexity compared to $O(|V|)$ for the k-core decomposition.

The K-truss decomposition has an asymptotic time complexity of $O(|E|^{1.5})$ [47], making it slower than the k-core decomposition. Nevertheless, since we are only interested in the vertices contained in the maximal K-truss, we make the simplification of running the K-truss decomposition on only the maximal k-core of the graph, similar to the prior work [32].

Empirically, in addition to analyzing the ME/CFS data from Xiong *et al.* study [34], we have also benchmarked KO on the IBD data [35] from integrative HMP, as well as chromosome

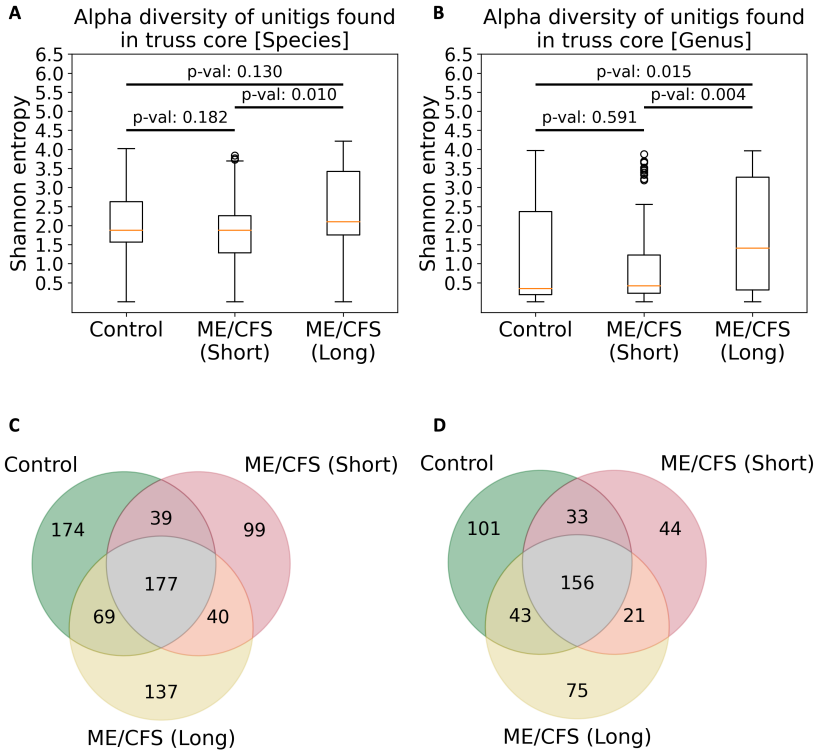

Fig. 5. **(A, B)** Distribution of alpha diversity (Shannon entropy) for anomalous unitigs that belong to the highest K-truss. p-values from Welch's t-test for equality of means are displayed above the boxplots. **(A)** Alpha diversity of samples associated with the ME/CFS condition is lower than that of control samples. Long ME/CFS samples on the other hand do not appear to be noticeably distinct from the control ones. Entropy was calculated based on the unitigs for which Kraken 2 provided species level classification. **(B)** Alpha diversity of samples associated with the ME/CFS condition is lower than that of control samples and long ME/CFS samples. Long ME/CFS samples on the other hand do not appear to be noticeably distinct from the control ones. Entropy was calculated based on the unitigs for which Kraken 2 provided genus level classification (species annotations are rolled up into respective genus). **(C)** Venn diagram representing intersections between sets of species identified in the control, short ME/CFS, and long ME/CFS sample collections. **(D)** Venn diagram representing intersections between sets of genera identified in the control, short ME/CFS, and long ME/CFS sample collections.

21 and chromosome 11 aligned reads from human genome HG002 from the Genome in a Bottle project [36]. The choice of human genome sequencing data is motivated by highly repetitive complex Alu regions present in the genome, hence the regime in which $|E| \sim \alpha|V|^2$ is in the HUG. All benchmarking was performed on a Ubuntu 18.04.6 LTS system with Intel(R) Xeon(R) Gold 5218 CPUs and 312GB of RAM and all runs used 60 threads. The results of benchmarking are summarized in Table 1.

The results in Table 1 showcase that resulting graph edge density is an important component of the overall computational performance, as indicated by a high run time value for the HG002 chromosome 11 experiment. Compared to the original KOMB implementation, we

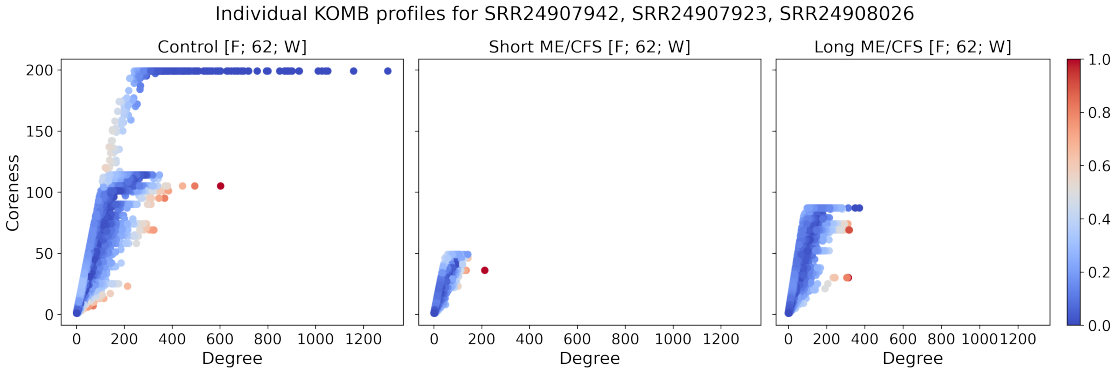

Fig. 6. Individual KO profiles (color: normalized Core-A anomaly score) for three samples from the ME/CFS cohort matched by age, gender and race. We observe a more complex profile in the control sample, while short- and long-term ME/CFS show compact profiles associated with lower bacterial genomic diversity. In all three samples, anomalous unitigs are predominantly high in degree compared to their coreness.

Table 1. Performance of KO on metagenomic and human genome datasets. Total dataset size refers to the cumulative size of all data processed, while average sample size describes the mean size of a single sample in a dataset. Analogously average runtime refers to mean time to process a single sample, while total runtime refers to cumulative time spent analyzing the dataset sequentially.

Dataset	# samples	Total dataset size (GB)	Average sample size (GB)	Average wall clock runtime (hrs)	Total wall clock runtime (hrs)
ME/CFS cohort	238	2,422	10.18	0.11	26.42
iHMP IBD	540	3,120	5.78	0.19	104.82
HG002 chr21 (300x)	1	26.17	-	-	0.71
HG002 chr21 (250bp)	1	5.90	-	-	0.14
HG002 chr11 (250bp)	1	20.43	-	-	1.70

achieve up to a 3-fold speed up for metagenomic samples containing an average of 16 million reads [24]. Additionally, we have performed a head-to-head comparison of KOMB and KO on a Zymo mock community sequenced by DOE Joint Genome Institute (BioProject Accession: PRJNA699918). We chose the Zymo mock community due to the large sample size (42 GB) and relatively simple genomic structure, allowing us to focus on the HUG construction performance, which we identified as a bottleneck, rather than the efficient k-core decomposition part of the analysis. On this data KOMB required a total of 7h16m of wall clock time (CPU time: 273h34m) and 48.28 GB of RAM to produce the final results, while KO required a total of 1h4m of wall clock time (CPU time: 14h34m) and 156.18 GB of RAM, note that both versions were ran with 40 threads for this experiment. The speedup is the result of three major changes in KO: (1) replacement of ABySS [48] with GGCAT [39] for the unitig construction, (2) change from BWA MEM to BWA MEM 2 for read mapping, and (3) improved parallelization in the KOMB codebase.

3. Discussion

In this work, we have provided a set of computational improvements to KOMB implemented in KO, and theoretical analysis of connections between HUGs and pangenome graphs. Addi-

tionally, we showcased the usage of KO on a ME/CFS patient cohort and identified disease-associated patterns. The dynamics inferred from KO profiles and taxa associated with important unitigs are concordant with the observations from a prior study [34]. Namely, we observe pronounced dysbiosis in short-term ME/CFS patients gut, and partial recovery from the dysbiosis in the long-term ME/CFS patients. We envision KO as an important tool to be integrated alongside existing approaches into clinically relevant microbiome studies. The key benefits of KO are: (a) selection of a small set of anomalous sequences without relying on taxonomy nor functional annotation, which can allow *de novo* analyses of these sequences and more sensitive detection of perturbations to host microbiome health, and (b) rapid profiling of a large number of samples, which can aid in the exploration of genotype-phenotype connections for large study cohorts.

An important next step is extending KO to an integrated approach that can annotate unitigs within the graph with associated transcriptomic or metabolomic information. Enriching the graph with multi-omic annotations can provide additional context for the nodes identified by KOMB as anomalous, enabling further functional associations to be extracted from the HUG structures. We believe that by adding -omics annotations KO can be further used to select genomic features relevant to the pathology, and hence enable better machine learning diagnostic tools. We also plan to add ability to distinguish between the edge types described in the Methods section and add the multi-omics annotations to the HUGs to directly extract hubs of functionally important genomic regions of a microbiome.

Additionally, it can be of interest to construct HUGs based on publicly available MAG catalogs as an annotation-rich reference for common community patterns identified in previous studies. We believe that this integrative large-scale approach can further illuminate mechanistic associations between microbiome and disease phenotypes.

4. Methods

4.1. *Hybrid unitig graph construction*

We begin construction of the HUG by constructing the underlying de Bruijn graph with a user-specified k-mer size parameter. The construction is done with the GGCAT [39], and the user can control the parameters exposed by the GGCAT command line interface. GGCAT produces a FASTA output file containing all maximal non-branching paths through the de Bruijn graph (unitigs). After unitigs are constructed the user has an option to specify an additional length based filtering step. Our recommended choice is setting this filter to be equal to the read length.

After construction and filtering, the final set of unitigs becomes the set of vertices in the HUG. Next, we perform read mapping of the input paired-end reads to the set of unitigs using BWA MEM 2 [49] v2.2.1 with the default parameters. We retain all read mappings for constructing the edges of the HUG. An edge is constructed between two unitigs u_1 and u_2 if either the same read maps to both of the unitigs, or, one read in the pair maps to u_1 and the other read in the pair maps to u_2. More precisely, let r_1 and r_2 be two paired end reads and let $M(r_1), M(r_2)$ be the sets of unitigs that r_1 and r_2 are mapped to. Then the initially empty set of edges (E) in the HUG is united with the set of newly created edges, i.e.

(a) $E \leftarrow E \cup \{\{u, v\} : u \in M(r_i) \text{ and } v \in M(r_i) \setminus u \text{ for } i = 1, 2\}$
(b) $E \leftarrow E \cup \{\{u, v\} : u \in M(r_1) \text{ and } v \in M(r_2) \setminus u\}$

Conceptually two kinds of edges arise from this construction: (a) local similarity edges, which capture subregions of unitigs that are similar as evidenced by the read mapping, and (b) adjacency edges which have potential proximity of two unitigs with a genome. While it is natural to expect that single read multi-mapping corresponds to similarity edges and paired-end information corresponds to the adjacency ones, it is worth noting that single read mapping also can contribute to the adjacency edge formation (see Figure 1a, b). We currently do not distinguish the two edge types (local similarity vs adjacency) in implementation.

4.2. *k-core decomposition and Core-A anomaly score*

The k-core of a graph is the maximal induced subgraph in which each node has a degree of at least k. If the vertices of the k-core of a graph are represented by V_k, then the coreness of vertex v is defined as $coreness(v) = \max\{k : v \in V_k\}$. Computing the coreness of each vertex is called k-core decomposition. Once the HUG is constructed, we perform a k-core decomposition of it using the igraph C library [50] implementation of the linear time Batagelj-Zaversnik [45] algorithm, which assigns a coreness to each vertex in the HUG. Subsequently, for each unitig a Core-A anomaly score is computed as specified in previous work on anomaly detection in networks [31]. In particular, for each vertex v we compute its rank based on the degree $\text{rank}_d(v)$, and its rank based on coreness $\text{rank}_c(v)$. The Core-A anomaly score is then defined as the absolute value of the difference of the log of the two ranks, i.e. $\text{Core-A}(v) = |\log \text{rank}_d(v) - \log \text{rank}_c(v)|$.

There are two key groups of unitigs with high anomaly scores: (a) individual anomalies and (b) anomalous clusters. In general, for any vertex v the shell number is upper bounded by the degree of that vertex. Thus, individual anomalies are nodes with a large discrepancy between their degree and coreness. In particular, this is can be described by the individual influence, ii value, defined as $ii = 1 - \text{coreness}(v)/\deg(v)$ that is equal to 0 if the degree and shell number are equal, and approaches 1 for values of degree significantly larger than that of coreness. Individual anomalies are unitigs likely to have varying genomic contexts in the metagenome. Thus, individual anomalies are good candidates for mobile genetic elements or duplicated genes. Anomalous clusters on the other hand are more likely to arise due to shared local similarities between a large group of unitigs. Those can be nearly identical repeats, such as Alu elements in human genomes, or hypervariable regions of ribosomal proteins in bacterial genomes.

4.3. *K-truss computation*

A K-truss of a graph is an induced subgraph in which every edge is present in at least $K - 2$ triangles. A method proposed by Malliaros *et al.* [32] computes the maximal K-truss by computing it for the k-core of the graph, since the K-truss of a graph is always a subgraph of its $K - 1$-core. Thus, as the k-core decomposition of the HUG is computed, we select the k-core subgraph of the HUG and then compute its K-truss decomposition using the igraph

C library's implementation of Wang and Cheng's algorithm [47]. The algorithm assigns a trussness value to each edge in the subgraph, representing the maximum value of K for which the edge is present in the K-truss.

Now, let V_k be the set of nodes and let E_k be the set of edges of the maximal k-core subgraph of the HUG, and define $\tau : E_k \to \mathbf{N}$ to be the mapping realized by the igraph algorithm, whose time complexity is $O(|E_k|^{1.5})$. We then set $K = \max\{\tau(e) : e \in E_k\}$ and select the vertices (i.e. unitigs) in the maximal K-truss to be those in $\{v \in V_k : \tau(e) = K \text{ for some } e \text{ incident to } v\}$.

4.4. *Taxonomic classification and α-diversity calculations*

Taxonomic classification of the unitigs was performed with Kraken 2 [2] with the standard parameters ($k = 35$, $\ell = 31$) and the standard Kraken 2 database consisting of RefSeq viral, bacterial, and archeal genomes, as well as human genome and known vector sequences from UniVec_Core. For α-diversity computations, the unclassified portion of unitigs was discarded, and the remaining fractions were re-normalized to add up to 1. The α-diversity was defined as the Shannon entropy of the classified unitig fractions $H = -\sum_{i \in T} f_i \log f_i$, where f_i is the fraction of unitigs classified as taxa i.

5. Data availability

This work has not produced any new sequencing data, and relied on publicly available datasets. Details for accessing these datasets are specified below.

ME/CFS metagenomic sequencing data. Illumina short paired-end sequences (150bp) from stool samples of 92 controls, 73 short-term ME/CFS, and 73 long-term ME/CFS patients were analyzed [34]. Original sequencing data was deposited into SRA under BioProject accession PRJNA878603.

IBD data from integrative HMP. Illumina short paired-end sequences from stool samples of patients with IDB were analyzed [35]. We analyzed a subset of 540 out of 1,613 available samples. Data is available from the HMP portal (https://portal.hmpdacc.org/) via study IBDMDB.

Human genome dataset. We have used Illumina short paired-end reads (150bp and 250bp) from Genome in a Bottle project. We used aligned reads for HG002 genome that can be accessed via the index hosted on GitHub: https://github.com/genome-in-a-bottle/giab_data_indexes.

6. Code availability

KOMB source code is publicly available on GitHub: https://github.com/treangenlab/komb.

7. Acknowledgements

The authors acknowledge helpful discussions and advice on statistical testing provided by Dr. Michael Nute. N.S. is supported by the Ken Kennedy Institute Andrew Ladd Memorial Excellence in Computer Science Fellowship. N.S. and T.J.T. were supported in part by the P01-AI152999 NIH award. T.J.T. was also supported by National Science Foundation grant EF-2126387, and NSF CAREER award (IIS-2239114).

References

[1] D. E. Wood and S. L. Salzberg, Kraken: ultrafast metagenomic sequence classification using exact alignments, *Genome biology* **15**, 1 (2014).

[2] D. E. Wood, J. Lu and B. Langmead, Improved metagenomic analysis with kraken 2, *Genome Biology* **20**, p. 257 (2019).

[3] K. D. Curry, Q. Wang, M. G. Nute, A. Tyshaieva, E. Reeves, S. Soriano, Q. Wu, E. Graeber, P. Finzer, W. Mendling *et al.*, Emu: species-level microbial community profiling of full-length 16s rrna oxford nanopore sequencing data, *Nature methods* **19**, 845 (2022).

[4] A. Blanco-Míguez, F. Beghini, F. Cumbo, L. J. McIver, K. N. Thompson, M. Zolfo, P. Manghi, L. Dubois, K. D. Huang, A. M. Thomas *et al.*, Extending and improving metagenomic taxonomic profiling with uncharacterized species using metaphlan 4, *Nature Biotechnology* , 1 (2023).

[5] O. Manor, C. L. Dai, S. A. Kornilov, B. Smith, N. D. Price, J. C. Lovejoy, S. M. Gibbons and A. T. Magis, Health and disease markers correlate with gut microbiome composition across thousands of people, *Nature communications* **11**, p. 5206 (2020).

[6] P. Scepanovic, F. Hodel, S. Mondot, V. Partula, A. Byrd, C. Hammer, C. Alanio, J. Bergstedt, E. Patin, M. Touvier *et al.*, A comprehensive assessment of demographic, environmental, and host genetic associations with gut microbiome diversity in healthy individuals, *Microbiome* **7**, 1 (2019).

[7] E. Castro-Nallar, M. L. Bendall, M. Pérez-Losada, S. Sabuncyan, E. G. Severance, F. B. Dickerson, J. R. Schroeder, R. H. Yolken and K. A. Crandall, Composition, taxonomy and functional diversity of the oropharynx microbiome in individuals with schizophrenia and controls, *PeerJ* **3**, p. e1140 (2015).

[8] K. Lange, M. Buerger, A. Stallmach and T. Bruns, Effects of antibiotics on gut microbiota, *Digestive Diseases* **34**, 260 (2016).

[9] A. Almeida, A. L. Mitchell, M. Boland, S. C. Forster, G. B. Gloor, A. Tarkowska, T. D. Lawley and R. D. Finn, A new genomic blueprint of the human gut microbiota, *Nature* **568**, 499 (2019).

[10] A. Almeida, S. Nayfach, M. Boland, F. Strozzi, M. Beracochea, Z. J. Shi, K. S. Pollard, E. Sakharova, D. H. Parks, P. Hugenholtz *et al.*, A unified catalog of 204,938 reference genomes from the human gut microbiome, *Nature biotechnology* **39**, 105 (2021).

[11] P. G. Wolf, E. S. Cowley, A. Breister, S. Matatov, L. Lucio, P. Polak, J. M. Ridlon, H. R. Gaskins and K. Anantharaman, Diversity and distribution of sulfur metabolic genes in the human gut microbiome and their association with colorectal cancer, *Microbiome* **10**, 1 (2022).

[12] K. Lee, S. Raguideau, K. Sirén, F. Asnicar, F. Cumbo, F. Hildebrand, N. Segata, C.-J. Cha and C. Quince, Population-level impacts of antibiotic usage on the human gut microbiome, *Nature Communications* **14**, p. 1191 (2023).

[13] A. M. Phillippy, M. C. Schatz and M. Pop, Genome assembly forensics: finding the elusive mis-assembly, *Genome biology* **9**, 1 (2008).

[14] M. Pop, Genome assembly reborn: recent computational challenges, *Briefings in bioinformatics* **10**, 354 (2009).

[15] N. D. Olson, T. J. Treangen, C. M. Hill, V. Cepeda-Espinoza, J. Ghurye, S. Koren and M. Pop, Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes, *Briefings in bioinformatics* **20**, 1140 (2019).

[16] C. Quince, S. Nurk, S. Raguideau, R. James, O. S. Soyer, J. K. Summers, A. Limasset, A. M. Eren, R. Chikhi and A. E. Darling, Strong: metagenomics strain resolution on assembly graphs, *Genome biology* **22**, 1 (2021).

[17] F. Meyer, A. Fritz, Z.-L. Deng, D. Koslicki, T. R. Lesker, A. Gurevich, G. Robertson, M. Alser, D. Antipov, F. Beghini *et al.*, Critical assessment of metagenome interpretation: the second

round of challenges, *Nature methods* **19**, 429 (2022).
[18] H. Ochman, J. G. Lawrence and E. A. Groisman, Lateral gene transfer and the nature of bacterial innovation, *nature* **405**, 299 (2000).
[19] F. A. Kondrashov, Gene duplication as a mechanism of genomic adaptation to a changing environment, *Proceedings of the Royal Society B: Biological Sciences* **279**, 5048 (2012).
[20] T. J. Treangen and E. P. Rocha, Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes, *PLoS genetics* **7**, p. e1001284 (2011).
[21] R. R. Wick and K. E. Holt, Polypolish: short-read polishing of long-read bacterial genome assemblies, *PLoS computational biology* **18**, p. e1009802 (2022).
[22] C. Y. Kim, J. Ma and I. Lee, Hifi metagenomic sequencing enables assembly of accurate and complete genomes from human gut microbiota, *Nature communications* **13**, p. 6367 (2022).
[23] F. Trigodet, K. Lolans, E. Fogarty, A. Shaiber, H. G. Morrison, L. Barreiro, B. Jabri and A. M. Eren, High molecular weight dna extraction strategies for long-read sequencing of complex metagenomes, *Molecular Ecology Resources* **22**, 1786 (2022).
[24] A. Balaji, N. Sapoval, C. Seto, R. L. Elworth, Y. Fu, M. G. Nute, T. Savidge, S. Segarra and T. J. Treangen, Komb: K-core based de novo characterization of copy number variation in microbiomes, *Computational and Structural Biotechnology Journal* **20**, 3208 (2022).
[25] S. Koren, T. J. Treangen and M. Pop, Bambus 2: scaffolding metagenomes, *Bioinformatics* **27**, 2964 (2011).
[26] G. Gautreau, A. Bazin, M. Gachet, R. Planel, L. Burlot, M. Dubois, A. Perrin, C. Médigue, A. Calteau, S. Cruveiller, C. Matias, C. Ambroise, E. P. C. Rocha and D. Vallenet, PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph, *PLoS Comput. Biol.* **16**, p. e1007732 (2020).
[27] J. Ghurye, T. Treangen, M. Fedarko, W. J. Hervey, 4th and M. Pop, MetaCarvel: linking assembly graph motifs to biological variants, *Genome Biol.* **20**, p. 174 (2019).
[28] D. J. Nasko, S. Koren, A. M. Phillippy and T. J. Treangen, Refseq database growth influences the accuracy of k-mer-based lowest common ancestor species identification, *Genome biology* **19**, 1 (2018).
[29] D. Li, C.-M. Liu, R. Luo, K. Sadakane and T.-W. Lam, Megahit: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph, *Bioinformatics* **31**, 1674 (2015).
[30] S. Nurk, D. Meleshko, A. Korobeynikov and P. A. Pevzner, metaspades: a new versatile metagenomic assembler, *Genome research* **27**, 824 (2017).
[31] K. Shin, T. Eliassi-Rad and C. Faloutsos, Corescope: Graph mining using k-core analysis—patterns, anomalies and algorithms, *2016 IEEE 16th international conference on data mining (ICDM)* , 469 (2016).
[32] F. D. Malliaros, M.-E. G. Rossi and M. Vazirgiannis, Locating influential nodes in complex networks, *Sci. Rep.* **6**, p. 19307 (January 2016).
[33] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley and H. A. Makse, Identification of influential spreaders in complex networks, *Nature Physics* **6**, 888 (Nov 2010).
[34] R. Xiong, C. Gunter, E. Fleming, S. D. Vernon, L. Bateman, D. Unutmaz and J. Oh, Multi-'omics of gut microbiome-host interactions in short-and long-term myalgic encephalomyelitis/chronic fatigue syndrome patients, *Cell Host & Microbe* **31**, 273 (2023).
[35] J. Lloyd-Price, C. Arze, A. N. Ananthakrishnan, M. Schirmer, J. Avila-Pacheco, T. W. Poon, E. Andrews, N. J. Ajami, K. S. Bonham, C. J. Brislawn *et al.*, Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases, *Nature* **569**, 655 (2019).
[36] E. D. Jarvis, G. Formenti, A. Rhie, A. Guarracino, C. Yang, J. Wood, A. Tracey, F. Thibaud-Nissen, M. R. Vollger, D. Porubsky *et al.*, Semi-automated assembly of high-quality diploid human reference genomes, *Nature* **611**, 519 (2022).

[37] J. Khan and R. Patro, Cuttlefish: fast, parallel and low-memory compaction of de bruijn graphs from large-scale genome collections, *Bioinformatics* **37**, i177 (2021).
[38] J. Khan, M. Kokot, S. Deorowicz and R. Patro, Scalable, ultra-fast, and low-memory construction of compacted de bruijn graphs with cuttlefish 2, *Genome biology* **23**, p. 190 (2022).
[39] A. Cracco and A. I. Tomescu, Extremely-fast construction and querying of compacted and colored de bruijn graphs with ggcat, *bioRxiv* (2022).
[40] P. E. Compeau, P. A. Pevzner and G. Tesler, How to apply de bruijn graphs to genome assembly, *Nature biotechnology* **29**, 987 (2011).
[41] Z. Iqbal, I. Turner and G. McVean, High-throughput microbial population genomics using the cortex variation assembler, *Bioinformatics* **29**, 275 (2013).
[42] R. M. Colquhoun, M. B. Hall, L. Lima, L. W. Roberts, K. M. Malone, M. Hunt, B. Letcher, J. Hawkey, S. George, L. Pankhurst *et al.*, Pandora: nucleotide-resolution bacterial pan-genomics with reference graphs, *Genome biology* **22**, 1 (2021).
[43] S. Martin, M. Ayling, L. Patrono, M. Caccamo, P. Murcia and R. M. Leggett, Capturing variation in metagenomic assembly graphs with metacortex, *Bioinformatics* **39**, p. btad020 (2023).
[44] D. E. Wood, J. Lu and B. Langmead, Improved metagenomic analysis with kraken 2, *Genome biology* **20**, 1 (2019).
[45] V. Batagelj and M. Zaversnik, An $O(m)$ algorithm for cores decomposition of networks, *arXiv preprint cs/0310049* (2003).
[46] U. Brandes, A faster algorithm for betweenness centrality, *Journal of mathematical sociology* **25**, 163 (2001).
[47] J. Wang and J. Cheng, Truss decomposition in massive networks, **5** (2012).
[48] S. D. Jackman, B. P. Vandervalk, H. Mohamadi, J. Chu, S. Yeo, S. A. Hammond, G. Jahesh, H. Khan, L. Coombe, R. L. Warren *et al.*, Abyss 2.0: resource-efficient assembly of large genomes using a bloom filter, *Genome research* **27**, 768 (2017).
[49] M. Vasimuddin, S. Misra, H. Li and S. Aluru, Efficient architecture-aware acceleration of bwa-mem for multicore systems, *2019 IEEE international parallel and distributed processing symposium (IPDPS)* , 314 (2019).
[50] G. Csardi, T. Nepusz *et al.*, The igraph software package for complex network research, *InterJournal, complex systems* **1695**, 1 (2006).