A peer-reviewed version of this preprint was published in PeerJ on 8 November 2016.

DeMaere MZ, Darling AE. 2016. Deconvoluting simulated metagenomes: the performance of hard- and soft- clustering algorithms applied to metagenomic chromosome conformation capture (3C) PeerJ 4:e2676 https://doi.org/10.7717/peerj.2676

Deconvoluting simulated metagenomes: the performance of hard- and soft-clustering algorithms applied to metagenomic chromosome conformation capture (3C)

Matthew Z. DeMaere[1] **and Aaron E. Darling**[1]

[1]**ithree institute, University of Technology Sydney, Sydney, NSW, Australia**

ABSTRACT

Background. Chromosome conformation capture, coupled with high throughput DNA sequencing in protocols like Hi-C and 3C-seq, has been proposed as viable means to generate data to resolve the genomes of microorganisms living in naturally occuring environments. Metagenomic Hi-C and 3C-seq datasets have begun to emerge, but the feasibility of resolving genomes when closely related organisms (strain-level diversity) are present in the sample has not yet been systematically characterised.

Methods. We developed a computational simulation pipeline for metagenomic 3C and Hi-C sequencing to evaluate the accuracy of genomic reconstructions at, above, and below an operationally defined species boundary. We simulated datasets and measured accuracy over a wide range of parameters. Five clustering algorithms were evaluated (2 hard, 3 soft) using an adaptation of the extended B-cubed validation measure.

Results. When all genomes in a sample are below 95% sequence identity, all of the tested clustering algorithms performed well. When sequence data contains genomes above 95% identity (our operational definition of strain-level diversity), a naive soft-clustering extension of the Louvain method achieves the highest performance.

Discussion. Previously, only hard-clustering algorithms have been applied to metagenomic 3C and Hi-C data, yet none of these perform well when strain-level diversity exists in a metagenomic sample. Our simple extension of the Louvain method performed the best in these scenarios, however, accuracy remained well below the levels observed for samples without strain-level diversity. Strain resolution is also highly dependent on the amount of available 3C sequence data, suggesting that depth of sequencing must be carefully considered during experimental design. Finally, there appears to be great scope to improve the accuracy of strain resolution through further algorithm development.

Keywords: 3C, HiC, chromosome conformation capture, microbial ecology, metagenomics, synthetic microbial communities, simulation pipeline, metagenome assembly, read mapping, clustering, soft clustering, external index

INTRODUCTION

The explicit and complete determination of the genomes present in an environmental sample is a highly prized goal in microbial community analysis. When combined with their relative abundances, this detailed knowledge affords a great deal of power to downstream investigations in such things as: community metabolism inference, functional ecology, genetic exchange and temporal or inter-community comparison. Unfortunately, the current standard methodology in DNA sequencing is incapable of generating data of such exquisite detail and although raw base-pair yield has increased dramatically with technological progress, a significant methodological source of information loss remains.

The organization of DNA into chromosomes (long-range contiguity) and cells (localization) is almost

completely lost as a direct result of two requirements of high-throughput library based sequencing; cell lysis (the DNA purification step) and the subsequent shearing (the DNA fragmentation step). What remains in the form of direct experimental observation is short-range contiguity information. Beginning from this observational evidence, the problem of determining long-range contiguity and thus reconstruction the original genomic sources is handed over to computational algorithms, in particular genome assembly algorithms. Computational approaches to genome assembly would have little success if not for the known physical constraints (the sequential structure of DNA), a high degree of oversampling in what remains (read depth) and the assumption that a very high degree of sequence identity (>95% ANI) implies a common chromosomal origin.

Though the damage done in the steps of purification and fragmentation amounts to a tractable problem in single-genome studies, in metagenomics the whole-sample intermingling of free chromosomes of varying genotypic abundance is an enormous blow to assembly algorithms. Conventional whole-sample metagenome sequencing (Tringe and Rubin, 2005) thus results in a severely underdetermined system, where the number of unknowns exceeds the number of observations and the degree to which a given metagenome is underdetermined depends variously on community complexity. Accurately and precisely inferring cellular co-locality for this highly fragmented set of sequences, particularly in an unsupervised *de novo* setting, and thereby achieving genotype resolution, remains an unsolved problem.

Recent techniques which repeatedly sample an environment, extracting a signal based on correlated changes in abundance to identify genomic content that is likely to belong to individual strains or populations of cells, have confidently obtained species resolution (Alneberg et al., 2013; Imelfort et al., 2014) and begun to work toward strain (genotype) resolution (Cleary et al., 2015). Inferring abundance per-sample from contig coverage (Alneberg et al., 2013; Imelfort et al., 2014) or k-mer frequencies (Cleary et al., 2015) respectively, the strength of this discriminating signal is a function of community diversity, environmental variation and sampling depth; and represents a significant computational task.

HiC (Lieberman-Aiden et al., 2009) is an extension of 3C (Dekker et al., 2002) exploiting high-throughput sequencing, as is 3C-seq (Stadhouders et al., 2013) and related techniques such as tethered conformation capture (Kalhor et al., 2012). Recently introduced to metagenomics (Beitel et al., 2014; Burton et al., 2014; Marbouty et al., 2014) as an alternative to purely computational solutions, this family of techniques attempts to capture the native conformational state of DNA prior to cell lysis. Thus bound, cellular lysis is followed by DNA extraction and restriction digestion. In dilute conditions, the religation of free ends within cross-linked DNA-protein complexes favors ligation of free ends that were in close physical proximity prior to cross-linking. The cross-linking is then reversed and a series of steps applied to prepare the ligation products for high-throughput sequencing. Post sequencing, the resulting proximity ligation read-pairs provide direct experimental evidence of the cellular co-locality of the respective short DNA sequences (read-pairs). Given sufficient sampling, proximity ligation (3C) read-pairs have the potential to link points of genomic variation at the genotype level at much longer ranges than has previously been possible (Selvaraj et al., 2013; Beitel et al., 2014).

Sequencing information generated in this way thus acts to recover a portion of the information lost in conventional whole genome shotgun (WGS) sequencing. Expressed in terms of observation probability, it has been shown that intra-chromosomal read-pairs (*cis*) follow a long-tailed distribution decreasing exponentially with increasing genomic separation (Beitel et al., 2014). Inter-chromosomal read-pairs (*trans*), modeled as uniformly distributed across chromosome pairs, typically occur an order of magnitude less frequently than *cis* pairs, and inter-cellular read-pairs are an order of magnitude less frequently again (Beitel et al., 2014). This hierarchy in observational probability has potential to be an extremely valuable source of information with which to deconvolute conventionally generated metagenomes into species and perhaps strains.

Previous work which leverages 3C data in assembly analysis has yielded algorithms focused on scaffolding (Burton et al., 2013; Marie-Nelly et al., 2014). In the context of clonal genome sequencing, 3C directed scaffolding can be applied directly to the entire draft assembly with reasonable success. Beyond monochromosomal genomes, it has been necessary to first cluster (group) assembly contigs into chromosome (plasmid) bins, after which scaffolding is applied to each bin in turn. A move to metagenomics generally entails increased sample complexity and less explicit knowledge about composition. Effectively clustering metagenomic assemblies, containing a potentially unknown degree of both species and strain diversity, represents a challenge that to date has not be thoroughly investigated.

In this work, we describe the accuracy of various analysis algorithms applied to resolving the genomes

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1974v1 | CC-BY 4.0 Open Access | rec: 20 Apr 2016, publ: 20 Apr 2016

of strains in metagenomic sequence data. The accuracy of these algorithms was measured over a range of simulated experimental conditions, including varying degrees of evolutionary divergence around the species boundary, and varying depths of generated sequence data. Finally, we discuss implications for the design of metagenomic 3C experiments on systems containing strain-level diversity, and describe the limitations of the present work.

MATERIALS AND METHODS

Representation

A contact map is formed by mapping proximity ligation read-pairs to the available reference and counting occurrences between any two genomic regions (Belton et al., 2012); where the definition of a genomic region is application dependent. Mathematically, the contact map is a symmetric square matrix $\mathscr{M}$, whose raw elements m_{ij} represent the set of observational frequencies between all genomic regions. The removal of experimental bias by normalization, inference of spatial proximity and finally prediction of chromosome conformation represents the majority of published work in the field to date (Lieberman-Aiden et al., 2009; Noble et al., 2011; Yaffe and Tanay, 2011; Imakaev et al., 2012).

Defining the genomic regions as the set of contigs produced by WGS assembly and noting that the contact map is equivalent to the adjacency matrix A of a weighted undirected graph G (Boulos et al., 2013), an alternative graphical representation can be obtained. The eponymous contig graph expresses the combined 3C and WGS assembly data as an undirected graph, where nodes n_i represent contigs and weighted edges $e(n_i, n_j, w_{ij})$ represent the observed frequency w_{ij} of 3C read-pairs linking contigs n_i and n_j. Expressed as a graph, a host of graph theoretic analysis methods can be brought to bear on problems in metagenomics. In particular, the utility of graph clustering in the reconstruction of community member genomes has been successfully demonstrated for both synthetic and real communities (Beitel et al., 2014; Burton et al., 2014; Marbouty et al., 2014).

Clustering

Placing entities into groups by some measure of relatedness is often used to reduce a set of objects O into a set of clusters K and ideally where number of clusters is much less than the number of objects (i.e. $|K| \ll |O|$). When object membership within the set of clusters K is mutually exclusive and discrete, so that no object o_i belongs to more than a single cluster κ_k, it is termed hard-clustering (i.e. $\forall o_i, o_j \in O \mid (o_i \subset \kappa_k) \wedge (o_j \subset \kappa_l) \wedge (\kappa_k \cap \kappa_l = \emptyset)$). Termed soft-clustering when the constraint is removed and objects allowed to belong to multiple clusters, the potentially non-empty intersection between clusters ($\kappa_k \cap \kappa_l \supseteq \emptyset$) is known as the overlap between κ_k, κ_l.

Possibly motivated by a desire to obtain the plainest answer with maximal contrast, and for the sake of relative mathematical simplicity, hard-clustering is the more widely applied approach. Despite this, many problem domains exist in which cluster overlap reflects real phenomena. For instance, in the metagenomic analysis of closely related species or strains, the tendency of the highly conserved core genome to co-assemble into single contigs while the more distinct accessory genomes tend not to, implies that a 1-to-1 correspondence of cluster-to-genome is not possible and an overlapping model is required.

From the aspect of prior knowledge, clustering algorithms fall into two categories: supervised, where important details with respect to cluster definition are presented to the algorithm at the outset (e.g. number of clusters, archetype objects); and unsupervised, where this is not required. Unsupervised algorithms, in removing this *a priori* condition, would be preferable if not necessary in situations where such information is unavailable (perhaps due to cost or accessibility) or the uncertainty in this information is high.

Appropriate Validation Measures

Although algorithmic complexity can ultimately dictate applicability to a given problem domain, the quality of the resulting solution remains an overriding concern in assessing an algorithm's value. Simply put, clustering algorithms group objects together when they are similar (the same cluster) and separates those objects which differ (different clusters). To fully assess the quality of a given clustering solution, multiple aspects must be considered. Measures that fail to account for one aspect or another may incorrectly rank solutions. Five important yet often incompletely addressed aspects of clustering quality have been proposed (Amigó et al., 2009): homogeneity, completeness, size, number and lastly the notion of a ragbag. Here, a ragbag is when preference is given to placing uncertain assignments in a single

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1974v1 | CC-BY 4.0 Open Access | rec: 20 Apr 2016, publ: 20 Apr 2016

catch-all cluster, rather than spreading them across otherwise potentially homogeneous clusters or leaving them as isolated nodes.

External measures, which compare a given solution to a gold-standard are a powerful means of assessing quality and they themselves vary in effectiveness. F_1-score, the harmonic mean of the traditional measures precision and recall, is frequently employed in the assessment of bioinformatics algorithms. For clustering algorithms, it is perhaps not well known that F_1-score fails to properly consider the aspect of completeness (Amigó et al., 2009) and further is sensitive to a preprocessing step where clusters and class labels must first be matched (Hirschberg and Rosenberg, 2007). The entropy based V-measure (Hirschberg and Rosenberg, 2007) was conceived to address these shortcomings, but does not consider the ragbag notion nor the possibility of overlapping clusters and classes. The external validation measure Bcubed (Bagga and Baldwin, 1998) addresses all five aspects and building from this, extended Bcubed (Amigó et al., 2009) supports the notion of overlapping clusters and classes. Analogous to F_1-score and V-measure, extended Bcubed is also the harmonic mean of a form of precision and recall.

Still, all of these measures treat the objects involved in clustering as being equal in value when assessing correct and incorrect placements. For some problem domains, it could be argued that correctly classifying object *A* may be more important than correctly classifying object *B*. Conversely, that incorrectly classifying object *A* may represent a larger error than incorrectly classifying object *B*. To this end, we introduce per-object weighting to extended Bcubed (Equation 1) and propose using contig length (bp) as the measure of inherent value when clustering metagenomic contigs.

Clustering Algorithm Selection

Supervised algorithms require *a priori* descriptive detail about the subject of study prior to analysis, while unsupervised algorithms make no such demand. This *a priori* knowledge can be of crucial importance scientifically, such as informing a clustering algorithm how many clusters exist within a dataset under study. For the genome of a single organism, where cluster count corresponds to chromosome count, independent estimation may be tenable. Extracting such descriptive information from an uncultured microbial community in the face of ecological, environmental and historical variation is an onerous requirement. For this reason, we only consider unsupervised algorithms, and focus attention to both hard and soft clustering approaches.

Four graph clustering algorithms were considered: MCL, SR-MCL, the Louvain method and OClustR (van Dongen, 2001; Shih and Parthasarathy, 2012; Blondel et al., 2008; Pérez-Suárez et al., 2013). While MCL and Louvain have previously been applied to 3C contig clustering (Beitel et al., 2014; Marbouty et al., 2014), to our knowledge SR-MCL and OClustR have not. We did not consider the clustering algorithm employed by (Burton et al., 2014) as it requires the number of clusters to be specified *a priori*.

Runtime parameters particular to each algorithm were controlled in the sweep as necessary (Table 2). The widely used MCL (markov clustering) algorithm (van Dongen, 2001) uses stochastic flow analysis to produce hard-clustering solutions, where cluster granularity is controlled via a single parameter ("inflation"). For this parameter, a range of 1.1 to 2.0 was chosen based on prior work (Beitel et al., 2014) and the interval sampled uniformly in five steps (inflation: 1.1 - 2.0). A soft-clustering extension of MCL, SR-MCL (soft, regularized Markov clustering) (Shih and Parthasarathy, 2012) attempts to sample multiple clustering solutions by iterative re-execution of MCL, penalizing node stochastic flows between iterations depending on the previous run state. Beyond MCL's inflation parameter, SR-MCL introduces four additional runtime parameters (balance, quality, redundancy and penalty ratio). It was determined that default settings were apparently optimal for these additional parameters (data not shown) and therefore only inflation was varied over the same range as MCL.

The Louvain modularity Q (Newman and Girvan, 2004) quantifies the degree to which a graph is composed of pockets of more densely interconnected subgraphs. Density is uniform across a graph when $Q = 0$ and there is essentially no community structure, while as $Q \to 1$ it indicates significant community structure with a strong contrast in the degree to which nodes are linked within and between communities. Louvain clustering builds upon this modularity score (Blondel et al., 2008), following a greedy heuristic to determine a best partitioning of a graph by the measure of local modularity, identifying sets of nodes more tightly interconnected with each other than with the remainder of the graph. Although a hierarchical solution by recursive application of the Louvain method on the subsequent subgraphs can be obtained, at each step the result is a hard-clustering. We implemented a one-step Louvain clustering algorithm in Python making use of the modules python-louvain and Networkx. We further extended this

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1974v1 | CC-BY 4.0 Open Access | rec: 20 Apr 2016, publ: 20 Apr 2016

hard-clustering method (Louvain-hard) to optionally elicit a naive soft-clustering solution (Louvain-soft), where after producing the hard-clustering, any two nodes in different clusters that are connected by an edge in the original graph are made members in both clusters.

We implemented the OClustR algorithm (Pérez-Suárez et al., 2013) in Python. The algorithm employs a graph covering strategy applied to a thresholded similarity graph using the notion of node relevance (the average of relative node compactness and density) (Pérez-Suárez et al., 2013). The approach functions without the need for runtime parameters, thus avoiding their optimization, and aims to produce clusters of minimum overlap and maximal size.

Gold Standard

The gold standard (ground truth) is a crucial element of external validation and often overlooked is the reality that it itself is only an approximation to the absolute truth. Particularly in the treatment of scientific data, what we call the gold standard is frequently the "best we can do". Despite the powerful *a priori* advantages gained by the explicit nature of simulation based studies, practical limitations can introduce uncertainty. In particular, the loss of read placement information in de Bruijn graph assembly means we must infer contig origin rather than obtain it explicitly from assembly output metadata.

In this study, the gold standard must accurately map the set of assembly contigs C to the set of community source genomes G, while supporting the notion of one-to-many associations from contig c_i to some or all genomes $g_i \in G$. It is this one-to-many association that represents the overlap between genomes at low evolutionary divergence. The mapping must also contend with spurious overlap signal from significant local alignments due to such factors as conserved gene content and try to minimize false positive associations.

We used LAST (v712) (Kiełbasa et al., 2011) to align the set of assembly contigs C onto the respective set of community reference genomes G. For each contig $c_i \in C$, LAST alignments were traversed in order of descending bitscore and used to generate a mask $M(g_j,x)$ of contig coverage indexed by both reference genome $g_j \in G$ and contig base position x. Rather than a binary representation, mask elements were assigned real values $[0,1]$ in proportion to the identity of the maximally scoring alignment to reference genome g_j at the given site x. Lastly the arithmetic mean was calculated over all base positions for each reference genome g_j (i.e. $\mu(g_j) = |x|^{-1} \sum_x M(g_j,x)$) and an association was recognized between contig c_i and reference genome g_j when $\mu(g_j) > 0.96$.

Graph Generation

Undirected contig graphs were generated by mapping simulated 3C read-pairs to WGS assembly contigs using BWA MEM (v0.7.9a-r786) (Li, 2013). Read alignments were accepted only in the case of matches with 100% coverage of each read and zero mismatches. In general, this restriction to 100% coverage and identity should be relaxed when working with real data and we found the iterative strategy employed by (Burton et al., 2014) effective in this case (data not shown). Assembly contigs defined the nodes n_i and inter-contig (*trans*) read-pairs the edges ($(n_i,n_j) \iff i \neq j$), while intra-contig (*cis*) read-pairs ($(n_i,n_j) \iff i = j$) were ignored. Raw edge weights $w(n_i,n_j)$ were defined as the observed number of read-pairs linking nodes n_i and n_j.

Validation

To assess the quality of clustering solutions a modification to the Extended Bcubed external validation measure (Amigó et al., 2009) was made, wherein each clustered object was attributed with an explicit weight. We call the resulting measure "weighted Bcubed" (Equation 1). For a uniform weight distribution this modification reduces to conventional Extended Bcubed. In our work, contig length (bp) was chosen as the weight when measuring the accuracy of clustered assembly contigs. Remaining the harmonic mean of Bcubed precision and recall, the weights $w(o_i)$ are introduced here (Equation 2, 3) and the result normalized. For an object o_i, the sum is carried out over all members of the set of objects who share at least one class $H(o_i)$ or cluster $D(o_i)$ with object o_i (Equation 3).

$$F_{b^3} = \frac{2 \langle P_{b^3} \rangle \langle R_{b^3} \rangle}{\langle P_{b^3} \rangle + \langle R_{b^3} \rangle} \tag{1}$$

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1974v1 | CC-BY 4.0 Open Access | rec: 20 Apr 2016, publ: 20 Apr 2016

where $\langle P_{b^3} \rangle$ and $\langle R_{b^3} \rangle$ are the weighted arithmetic means of $P_{b^3}(o_i)$ and $R_{b^3}(o_i)$ (Equation 2, 3) over all objects.

$$P_{b^3}(o_i) = \frac{1}{\sum_{o_j \in D(o_i)} w(o_j)} \sum_{o_j \in D(o_i)} w(o_j) P^*(o_i, o_j) \tag{2}$$

$$R_{b^3}(o_i) = \frac{1}{\sum_{o_j \in H(o_i)} w(o_j)} \sum_{o_j \in H(o_i)} w(o_j) R^*(o_i, o_j) \tag{3}$$

Unchanged from Extended Bcubed, the expressions for the Multiplicity Bcubed precision $P^*(o_i, o_j)$ (Equation 4) and recall $R^*(o_i, o_j)$ (Equation 5) account for the non-binary relationship between any two items in the set when dealing with overlapping clustering.

$$P^*(o_i, o_j) = \frac{min\left(\left|K(o_i) \cap K(o_j)\right|, \left|\Theta(o_i) \cap \Theta(o_j)\right|\right)}{\left|K(o_i) \cap K(o_j)\right|} \tag{4}$$

$$R^*(o_i, o_j) = \frac{min\left(\left|K(o_i) \cap K(o_j)\right|, \left|\Theta(o_i) \cap \Theta(o_j)\right|\right)}{\left|\Theta(o_i) \cap \Theta(o_j)\right|} \tag{5}$$

where $K(o_i)$ the set of clusters and $\Theta(o_i)$ the set of classes either of which contain object o_i.

Pipeline Design

The selected workflow (Figure 1) represents a simple and previously applied (Beitel et al., 2014; Burton et al., 2014) means of incorporating 3C read data into traditional metagenomics, via *de novo* WGS assembly and subsequent mapping of 3C read-pairs to assembled contigs. Inputs to this core process are 3C read-pairs and WGS sequencing reads. Outputs are the set of assembled contigs C and the set of "3C read-pairs to contig" mappings M_{3C}. Although tool choices vary between researchers, we chose to keep the assembly and mapping algorithms fixed and focus instead on how other parameters influence the quality of metagenomic reconstructions with 3C read data. The A5-miseq pipeline (incorporating IDBA-UD, but skipping error correction and scaffolding via the –metagenome flag) (Coil et al., 2015) was used for assembly. BWA MEM was used for mapping 3C read-pairs to contigs (Li, 2013). Parameters placed under control were: WGS coverage (xfold), the number of 3C read-pairs (n3c) and a random seed (S). Prepended to this core process are two preceding modules: community generation and read simulation.

From a given phylogenetic tree and an ancestral sequence, the community generation module produces a set of descendent taxa with an evolutionary divergence defined by the phylogeny and evolutionary model. The simulated evolutionary process is implemented by sgEvolver (Darling et al., 2004), which models both local changes (e.g. single nucleotide substitutions and indels) and larger genomic changes (e.g. gene gain, loss, and rearrangement). The degree of divergence is controlled through a single scale factor α_{BL} (Table S1) that uniformly scales tree branch lengths prior to simulated evolution. As data inputs, the module takes a phylogenetic tree and an ancestral genome. As data outputs, the module generates a set of descendent genomes G and an accompanying gold-standard. Overall, community generation introduces the following two sweep parameters: branch length scale factor α_{BL} and random seed (S) (Table 1).

Following community generation, the read-simulation module takes the set of descendent genomes G as input and generates as output simulated Illumina WGS paired-end reads and 3C read-pairs. Variation in relative abundance of the descendent genomes G in simulated metagenomes was produced by wrapping ART_illumina (v1.5.1) (Huang et al., 2012) within a Python script (metaART.py) with the added dependency of an abundance profile table as input. A 3C read-pairs simulator was implemented in Python (simForward.py), capable of simulating both inter- and intra- chromosomal pairs from whole communities when supplied a set of reference genomes and a per-genome abundance profile. Here, a linear combination of the geometric and uniform distributions was used to model a long-tailed probability distribution of intra-chromosomal (*cis*) read-pairs as a function of genomic separation and the distribution was calibrated

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1974v1 | CC-BY 4.0 Open Access | rec: 20 Apr 2016, publ: 20 Apr 2016

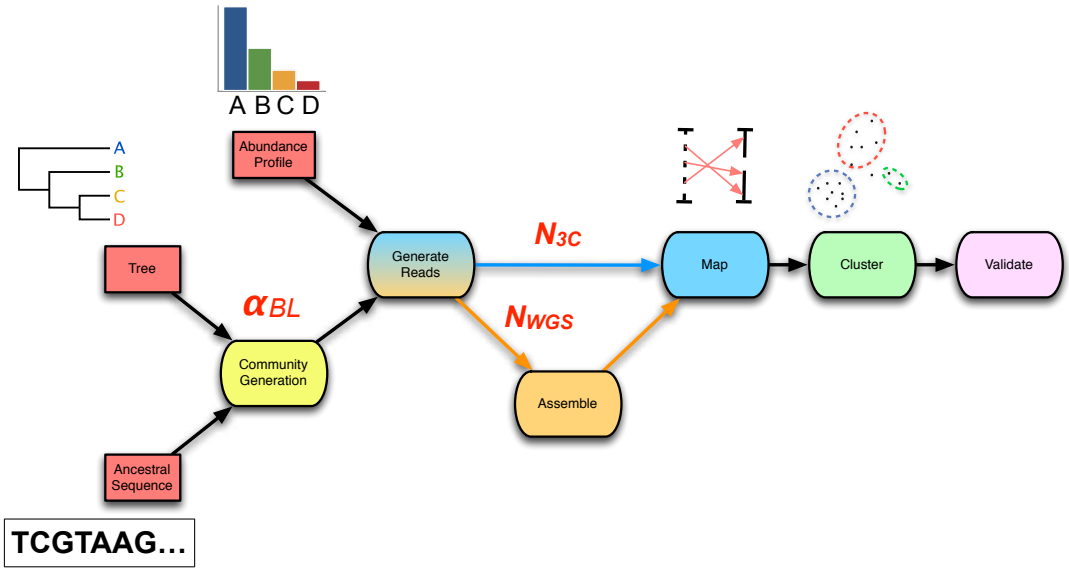

Figure 1. The 3C sequencing simulation pipeline used within the parameter sweep. An ancestral sequence and phylogenetic tree are used in simulating a process of genome evolution with varying divergence (α_{BL}). The resulting evolved genomes are subsequently subjected to *in silico* high-throughput sequencing, producing both WGS and 3C read-sets of chosen depth (N_{WGS}, N_{3C}). WGS reads are assembled and 3C read-pairs are mapped to the resulting contigs to generate a contig graph. Finally, the graph is supplied to a clustering algorithm and the result validated against the relevant gold standard.

by fitting it to the real experimental data of (Beitel et al., 2014). No constraints that would come about by modeling 3D chromosome structure were imposed on the simulation. Read-generation introduces the following sweep parameters: WGS depth of coverage (xfold) and number of 3C read-pairs (n3c) (Table 1).

After the assembly and mapping module comes the community deconvolution module, taking as input the set of 3C read mappings M_{3C}. Internally, the first step of the module generates the contig graph $G(n, e, w(e))$. Deconvolution is achieved by application of graph clustering algorithms, where the set of output clusters K reflect predicted genomes of individual community members (Beitel et al., 2014; Burton et al., 2014).

Lastly, the validation module takes as inputs: a clustering solution, a gold-standard and a set of assembly contigs. The first two inputs are compared by way of weighted Bcubed (Equation 1), while the set of contigs is supplied to QUAST (v3.1) (Gurevich et al., 2013) for the determination of conventional assembly statistics. The results from both clustering and assembly validation are then joined together to form a final output.

Simulation

Variational studies require careful attention to the number of parameters under control and their sampling granularity, so as to strike a balance between potential value to observational insight and computational effort. Even so, the combinatorial explosion in the total number of variations makes a seemingly small number of parameters and steps quickly exceed available computational resources. Further, an overly ambitious simulation can itself present significant challenges to the interpretation of fundamental system behaviour under the induced changes.

End-to-end, the simulation pipeline makes a large number of variables available for manipulation, and the size and dimensionality of the resulting space is much larger than can be explored with available computational resources. Therefore we decided to focus our initial exploration on a small part of this space. We used two simple phylogenetic tree topologies (a four taxon ladder and a four taxon star) (Figure 2), to develop insight into the challenges that face metagenomics researchers choosing to apply 3C to communities which contain closely related taxa.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1974v1 | CC-BY 4.0 Open Access | rec: 20 Apr 2016, publ: 20 Apr 2016

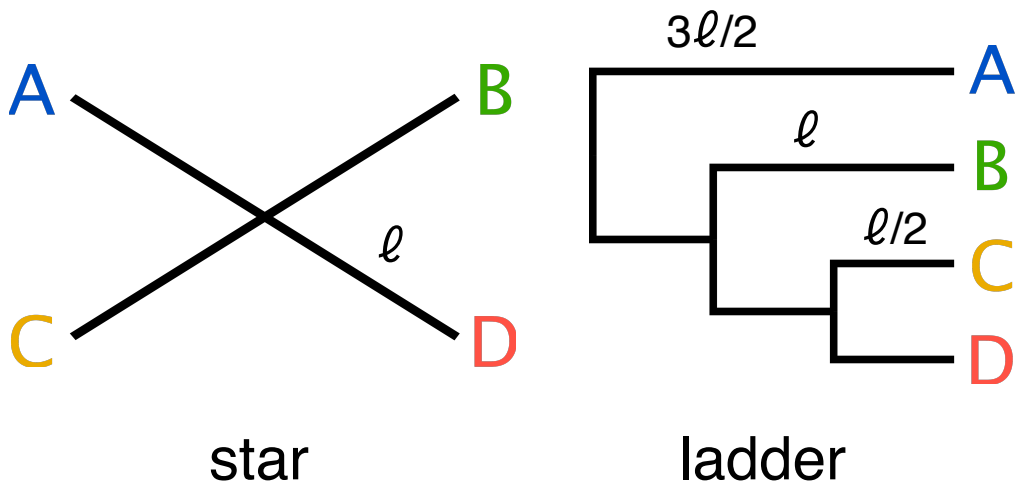

Figure 2. Two simple trees of four taxa (A,B,C,D) were used in the parameter sweep. The star; where all taxa are of equal evolutionary distance ℓ and ladder; where evolutionary distance decreases in incremental steps of $\ell/2$. For the ladder, the length of the internal branch for taxon B was set equal to the branch length of the star and therefore possesses both more closely and more distantly related community members for any value of the scale factor α_{BL} relative to the star topology.

Parameter Sweep

A single monochromosomal ancestral genome was used throughout (Escherichia coli K-12 substr. MG1655 (acc: NC_000913)). Two simple ultrametric tree topologies of four taxa (tree: star, ladder) (Fig. 2) were included and evolutionary divergence was varied over ten values on a log-scale (α_{BL}: 1 – 0.025; mean taxa ANIb 85 – 99.5%) (Figure 3). Two community abundance profiles were tested (profile: uniform and $1/e$). WGS coverage was limited to three depths (xfold: 10, 50, 100), which for uniform abundance represents 0.12, 0.60 and 1.2 Gbp of sequencing data respectively. Being a simple simulated community, greater depths did not appreciably improve the assembly result. The number of 3C read-pairs was varied from 10 to 100 thousand pairs (n3c: 10k, 20k, 50k, 100k), while the remaining parameter variations can be found in Tables 1 and 2.

From the 40 simulated microbial communities, the resulting 120 simulated metagenome read datasets were assembled and the assemblies evaluated using QUAST (v3.1) (Gurevich et al., 2013) against the 20 respective reference genome sets. Both external reference based (E.g. rates of mismatches, Ns, indels) and internal (E.g. N50, L50) statistics were collected and later joined with the results from the downstream cluster validation measures. Data generation resulted in 480 distinct combinations of simulation parameters, forming the basis for input to the selected clustering algorithms. OClustR results in 480 clusterings; Louvain clustering was performed both as standard hard-clustering (Louvain-hard) and our naive soft-clustering modification (Louvain-soft) resulting in 480 clusterings each; lastly MCL and SR-MCL were both varied over one parameter (infl) in 5 steps resulting in 2400 clusterings each. Finally, the quality of the clustering solutions for all four algorithms was assessed using the weighted extended Bcubed (Equation 1) external validation measure. Other parameters fixed throughout the sweep were: ancestral genome size (seq-len: 3 Mbp), indel/inversion/HT rate multiplier (sg_scale: 1e-4), small HT size (P̃oisson(200 bp)), large HT size range (Ũniform(10-60 kbp)), inversion size (G̃eometric(50 kbp)), WGS read generation parameters (read-length: 150 bp, insert size: 450 bp, standard deviation: 100 bp); HiC/3C parameters (read-length: 150 bp, restriction enzyme: NlaIII [_CATGˆ]). As simulated genomes were monochromosomal, inter-chromosomal read-pair probability was not a factor.

Assembly Entropy

A normalized entropy based formulation S_{mixing} (Equation 6) was used to quantify the degree to which a contig within an assembly is a mixture of source genomes, averaged over the assembly with terms

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1974v1 | CC-BY 4.0 Open Access | rec: 20 Apr 2016, publ: 20 Apr 2016

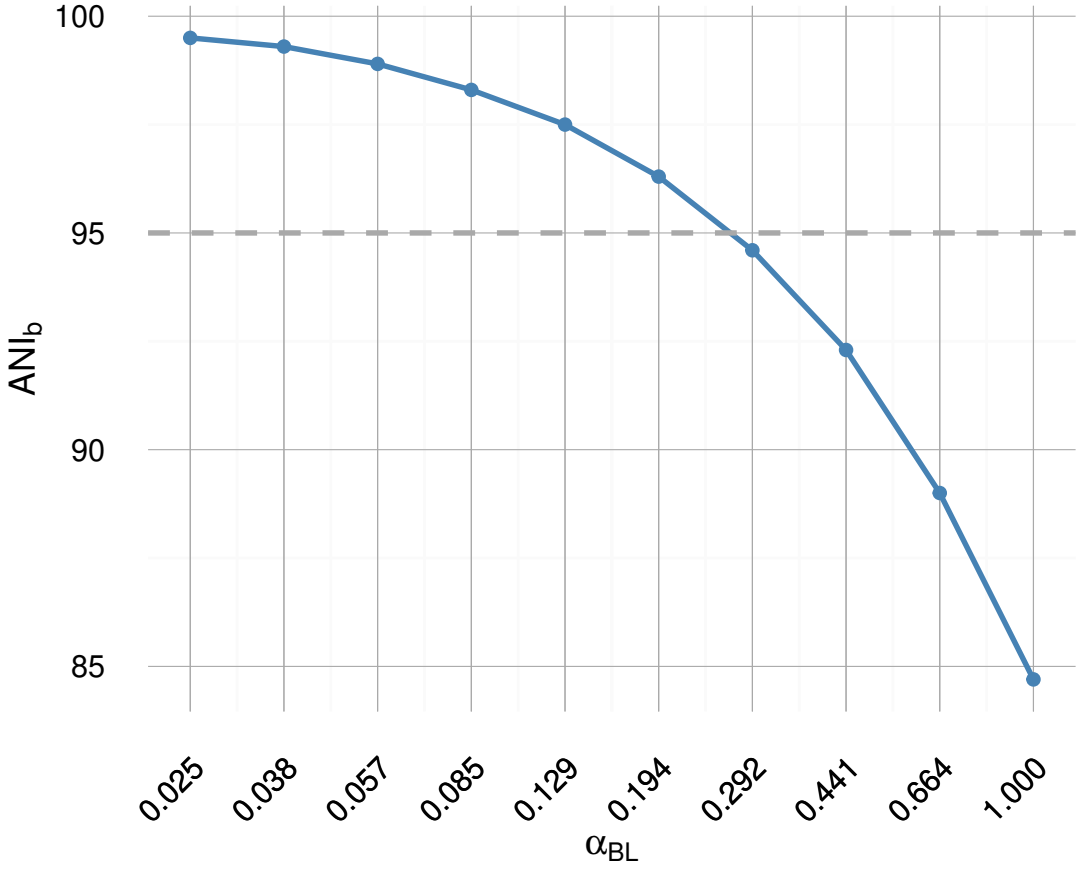

Figure 3. For sample points used in the sweep for the star topology, we depict the relationship between BL and the resulting measure of average nucleotide identity from BLAST (ANIb). The 95% threshold indicated is used internally within IDBA-UD (Peng et al., 2012) to determine whether to merge highly similar contigs and has been proposed as a pragmatic definition of bacterial species (Konstantinidis et al., 2006) akin to 98% 16S rRNA identity.

weighted in proportion to contig length. For simulated communities, the maximum attainable value is equal to the logarithm of the sum of the relative abundances q_i, the effective number of genomes N_{eff} (uniform profile $N_{eff} = 4$, $1/e$ profile $N_{eff} \approx 1.37$). Here N_C is the number of contigs within an assembly, N_G the number of genomes within a community and L_{asm} simply the total extent of an assembly, p_{ij} is the proportion of reads belonging to i^{th} genome mapping to the j^{th} contig, l_j the length of the j^{th} contig, and h the step size in α_{BL}.

When each contig in an assembly is derived purely from a single genomic source $S_{mixing} = 0$, conversely when all contigs possess a proportion of reads equal to the relative abundance the respective source genome $S_{mixing} = 1$. A forward finite difference was used to approximate the first order derivative ΔS_{mixing} (Equation 7), where mixing was regarded as a function of α_{BL} and the difference taken between successive sample points in the sweep.

$$S_{mixing} = -\frac{1}{L_{asm}\log_2(N_{eff})}\sum_{j=1}^{N_C} l_j \sum_{i=1}^{N_G} p_{ij}\log_2(p_{ij}) \qquad L_{asm} = \sum_{j=1}^{N_C} l_j, \qquad N_{eff} = \sum_{i=1}^{N_G} q_i \tag{6}$$

$$\Delta S_{mixing}(\alpha_{BL}) = \frac{1}{h}\left(S_{mixing}(\alpha_{BL}+h) - S_{mixing}(\alpha_{BL})\right) \tag{7}$$

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1974v1 | CC-BY 4.0 Open Access | rec: 20 Apr 2016, publ: 20 Apr 2016

Graph Complexity

Although simple intrinsic graph properties such as order, size and density can provide a sense of complexity, they do not consider the internal structure or information content present in a graph. One information-theoretic formulation with acceptable computational complexity is the non-parametric entropy H_L (Equation 8) associated with the non-zero eigenvalue spectrum of the normalized Laplacian matrix $N = D^{-1/2}LD^{-1/2}$, where $L = D - A$ is the regular Laplacian matrix, D is the degree matrix and A the adjacency matrix of a graph (Dehmer and Mowshowitz, 2011; Mowshowitz and Dehmer, 2012).

$$H_L = \sum_{\lambda_i \in \{\lambda : \lambda > 0\}} |\lambda_i| \log_2 |\lambda_i| \tag{8}$$

where $\{\lambda : \lambda > 0\}$ is set the non-zero eigenvalues of the normalized Laplacian N.

Table 1. Primary parameters under control in the sweep. In total, each clustering algorithm is presented with 480 combinations which may further increase depending on whether a clustering algorithm also has runtime parameters under control.

Level	Name	Description	Type	Number	Total	Values
1	tree	Phylogenetic tree topology	factor	2	2	star, ladder
2	profile	Relative abundance profile	factor	2	4	uniform, $1/e$
3	α_{BL}	Branch length scale factor	numeric	10	40	0.025–1 (log scale)
4	xfold	WGS paired-end depth of coverage	numeric	3	120	10, 50, 100
5	n3c	Number of 3C read-pairs	numeric	4	480	10000, 20000, 50000, 100000
6	algo	Clustering algorithm	factor	5		MCL, SM-MCL, Louvain-hard, Louvain-soft, OClustR

Table 2. Clustering algorithm dependent parameters explored in the sweep, where the base set of combinations begins with the fundamental 480 combinations. Only MCL and SR-MCL were swept through additional runtime parameters.

Algorithm	Name	Description	Type	Number	Total	Values	Sampling
MCL	infl	Inflation parameter	numeric	5	2400	1.1–2	linear
SR-MCL	infl	Inflation parameter	numeric	5	2400	1.1–2	linear
Louvain-hard				1	480		
Louvain-soft				1	480		
OClustR				1	480		

RESULTS

Assembly Complexity

Along with traditional assembly validation statistics (N50, L50) (Figure 4a, 4b), assembly entropy S_{mixing} and its approximate first order derivative ΔS_{mixing} (Equations 6, 7) (Figure 4c) were calculated for all 120 combinations resulting from the first four levels of the sweep (parameters: tree, profile, α_{BL}, xfold) (Table 1).

As community composition moves from the realm of distinct species (α_{BL}=1.0, ANI≈85%) to well below the conventional definition of strains (α_{BL}=0.025, ANI≈99.5%), and though increased read-depth

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1974v1 | CC-BY 4.0 Open Access | rec: 20 Apr 2016, publ: 20 Apr 2016

assists in delaying the onset, the degree of contig mixing increases more or less monotonically. After α_{BL}, the only significant continuous variable influencing mixing is read-depth (Spearman's ρ=-0.26, P=$3.83x10^{-3}$), while abundance profile is the only significant categorical variable (one-factor ANOVA R^2=0.0774, P=$2.09x10^{-3}$) (Lê et al., 2008). In all cases, as α_{BL} decreases mixing approaches unity; implying that as genomic sources become more closely related, the resulting metagenomic assembly contigs are of increasingly mixed origin.

Regarding the assembly process as a dynamic system in terms of evolutionary divergence, the turning point evident in ΔS_{mixing} (Figure 4c dashed lines) could be regarded as the critical point in a continuous phase transition from a state of high purity ($S_{mixing} \approx 0$) to a state dominated by completely mixed contigs ($S_{mixing} \rightarrow 1$). This point in evolutionary divergence coincides with the region where assemblies are the most fragmented (max L50, min N50) (Figure 4a, 4b) and ΔS_{mixing} is well correlated with both N50 (Spearman's ρ=0.72, $P < 1x10^{-5}$) and L50 (Spearman's ρ=-0.83, $P < 1x10^{-7}$), implying that as community divergence decreases through this critical point, traditional notions of assembly quality follow suit.

Graph Complexity

Introduction of 3C sampling depth at the next level within the sweep (parameter: n3c) generated 480 contig graphs (Table 1). To assess how assembly outcome affects the derived graph: order, size, density, and entropy H_L (Equation 8) were calculated and subsequently joined with the associated factors from assembly (Figure 4d).

Per the definition of the contig graph, there is a strong linear correlation between graph order $|n|$ and L50 (Pearson's r=0.96, $P < 1x10^{-16}$) and a weaker but still significant linear correlation between graph size $|e|$ and 3C sampling depth (parameter: n3c) (Pearson's r=0.61, $P < 1x10^{-16}$). Graphical density was strongly linearly correlated with graphical complexity (Pearson's r=-0.87, $P < 1x10^{-16}$). Graph entropy H_L is strongly correlated with assembly statistics N50 (Spearman's ρ=-0.97, $P < 1x10^{-16}$), L50 (Spearman's ρ=0.96, $P < 1x10^{-16}$) and ΔS_{mixing} (Spearman's ρ=-0.73, $P =< 1x10^{-16}$).

The knock-on effect of evolutionary divergence on the contig graphs derived from metagenomic assemblies is clear; fragmented assemblies comprised of contigs of mixed heritage result in increased contig graph complexity. As 3C read-pairs are the direct observations used to infer association between contigs, it could be expected that the correlation between 3C sampling depth and graphical size ($|e|$) is high ($\rho \rightarrow 1$) and so too the rate at which new edges are formed as read-pair data is added. As we observe a more moderate correlation (ρ=0.61) and with the absence of unhelpful spurious read-pairs in our simulation model, the perceived efficiency shortfall is in fact the necessary and expected repeat observation of contig-to-contig associations. Therefore by the nature of the experiment, increased 3C sampling depth does not confer increased graphical complexity in the same way that a more fragmented assembly would and increased 3C sampling depth can significantly improve the quality of clustering solutions.

Clustering Validation

The 240 contig graphs resulting from the sweep at uniform abundance were used to assess the influence of the various parameters on the performance of five clustering algorithms. For each clustering algorithm, overall performance scores, using F_{b^3} (Equation 1), were joined with their relevant sweep parameters and PCA performed in R (FactoMineR v1.32) (Lê et al., 2008). The first three principal components explain 78% of the variation, where PC1 is primarily involved with factors describing graphical complexity (α_{BL}: r=0.91, $P = 2.77x10^{-93}$; density: r=0.70, $P = 4.59x10^{-}36$; order: r=-0.76, $P = 2.81x10^{-47}$; ANIb: r=-0.91, $P = 5.25x10^{-92}$; H_L: r=-0.92, $P = 4.76x10^{-96}$), PC2 factors describing the sampling of contig-contig associations and overall connectedness of the contig graph (size: r=0.87, $P = 2.58x10^{-16}$; n3c: r=0.72, $P = 2.65x10^{-39}$; modularity: r=-0.48, $P = 5.41x10^{-15}$) and PC3 pertaining to local community structure (modularity: r=0.72, $P = 5.19x10^{-40}$; xfold: r=0.52, $P = 2.72x10^{-18}$) (Figure 5).

Of the five clustering algorithms, the performance of four (MCL, SR-MCL, Louvain-hard and OClustR) is strongly correlated with PC1 and so their solution quality is inversely governed by the the degree of complexity in the input graph, which in turn is largely influenced by within-community evolutionary divergence. The fifth algorithm, our naive Louvain-soft, though also correlated with PC1 and so negatively affected by graphical complexity, possesses significant correlation with PC2 (r=0.53, $P = 1.27x10^{-18}$) and thus noticeably benefits from increased 3C sampling depth (Figure 5).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1974v1 | CC-BY 4.0 Open Access | rec: 20 Apr 2016, publ: 20 Apr 2016

DISCUSSION

Selecting a deeply sequenced slice from within the sweep (profile: uniform, xfold: 100) and ideal algorithm-specific runtime parameters (MCL, SR-MCL inflation: 1.1), we can visually compare clustering performance under best tested conditions for a given algorithm (Figure 6). For evolutionary divergence well above the level of strains and prior to the critical region of assembly ($\alpha_{BL} \gg 0.292$, ANIb $\ll 95\%$), all algorithms achieve their best performance ($F_{b^3} \rightarrow 1$) (Figure 6c). As evolutionary divergence decreases toward the level of strains and the assembly process approaches the critical region, a fall-off in performance is evident for all algorithms and this performance drop is largely attributable to loss of recall (Figure 6b). Hard-clustering algorithms (MCL, Louvain-hard) in general exhibit superior precision (Figure 6a) to that of soft-clustering algorithms (SR-MCL, OClustR, Lovain-soft) and the precision of soft-clustering algorithms is worst in the critical region where graphical complexity is highest.

A ten-fold increase of 3C sampling depth (10k to 100k) has only a modest effect on clustering performance for 4 of the 5 algorithms, the exception being our naive Louvain-soft. Louvain-soft makes substantial gains in recall from increased 3C sampling depth at evolutionary divergences well below the level of strains ($\alpha_{BL} < 0.085$, ANIb $< 98\%$), but sacrifices precision at larger evolutionary divergences. The soft-clustering SR-MCL also sacrifices precision but fails to make similar gains in recall as compared to Louvain-soft. Recall for all three hard-clustering algorithms (MCL, Louvain-hard, OClustR) decreases with decreasing evolutionary divergence and the growing prevalence of degenerate contigs. This drop in recall is particularly abrupt for the star topology where, within the assembly process, all taxa approach the transitional region simultaneously. Being primarily limited by their inability to infer overlap, increase in 3C sampling depth for the hard-clustering algorithms has little effect on recall.

Our results have implications for the design of metagenomic 3C sequencing experiments. When genomes with >95% ANI exist in the sample, the power to resolve differences among those genomes can benefit greatly from generation of additional sequence data beyond what would be required to resolve genomes below 95% ANI. In our experiments the best results were achieved with 100x WGS coverage in addition to 100,000 3C read-pairs. For the simple communities of four genomes each of roughly 3Mbp considered here, 100x coverage corresponds to generating approximately 1.2Gbp of Illumina shotgun data. In a metagenomic 3C protocol (Marbouty et al., 2014), obtaining 100,000 proximity ligation read-pairs would require approximately 10^7 read-pairs in total; when we assume a proximity ligation read-pair rate of 1% (Liu and Darling, 2015). We note that current Illumina MiSeq V3 kits are specified to produce up to $\approx 2x10^7$ read-pairs, while HiSeq 2500 V4 lanes are specified to yield up to $\approx 5x10^8$ read-pairs per lane. Therefore, while it may be possible to resolve closely related genomes in very simple microbial communities with the capacity of a MiSeq, the scale of the HiSeq is likely to be required in many cases. Alternatively, the more technically complicated HiC protocol may be advantageous to achieve higher proximity ligation read rates, with up to 50% of read pairs spanning over 1kbp.

Limitations and Future Work

Our simulation of 3C read-pairs did not include modeling of experimental noise in the form of spurious read-pairs that do not reflect true DNA:DNA interactions. Such aberrant products have been estimated to occur in real experiments at levels up to 10% of total yield in 3C read-pairs (Liu and Darling, 2015). As a first approximation, we feel it reasonable to assume these erroneous read-pairs are a result of uniformly random ligation events between any two DNA strands present in the sample. The sampling of any such spurious read-pair will be sparse in comparison to the spatially constrained true 3C read-pairs and therefore amount to weak background noise. As currently implemented, the Louvain-soft clustering method would be prone to creating false cluster joins in the presence of such noise, but a simple low frequency threshold removal (e.g. requiring some constant number N links to join communities instead of 1) could in principle resolve the problem.

Only 3C read-pairs were used when inferring the associations between contigs, while conventional WGS read-pairs were used exclusively in assembly. It could be argued that also including WGS read-pairs during edge inference would have had positive benefits, particularly when assemblies were highly fragmented in the critical region.

Only raw edge weights were used in our analysis as normalization procedures, such as have been previously employed (Beitel et al., 2014; Marbouty et al., 2014; Burton et al., 2014), proved only weakly beneficial when at higher 3C sampling depths and occasionally detrimental in situations of low sampling depth. For higher sampling depth, the weak response can likely be attributed to a lack of complexity

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1974v1 | CC-BY 4.0 Open Access | rec: 20 Apr 2016, publ: 20 Apr 2016

and the low noise environment inherent in simulation. For low sampling depth, observation counts are biased to small values (mode $[w(n_i,n_j)] \to 1$) and simple counting statistics would suggest there is high uncertainty ($\pm\sqrt{w(n_i,n_j)}$) in these values. As such, this uncertainty is propagated via any normalization function $f(w(n_i,n_j))$ that attempts to map observation counts to the real numbers ($f : \mathbb{N} \to \mathbb{R}$). Even under conditions for high sampling depth, pruning very infrequently observed low-weight edges can prove beneficial to clustering performance as, beyond this source of uncertainty, some clustering algorithms appear to unduly regard the mere existence of an edge even when its weight is vanishingly small relative to the mean.

For the sake of standardization and to focus efforts on measuring clustering algorithm performance we elected to use a single assembly and mapping algorithm. However, many alternative methods for assembly and mapping exist. In the case of assembly, there are an increasing number of tools intended explicitly for metagenomes, such as metaSPAdes (Bankevich et al., 2012), MEGAHIT (Li et al., 2015), or populations of related genomes (Cortex) (Iqbal et al., 2013), while the modular MetAMOS suite (Treangen et al., 2013) at once offers tantalising best-practice access to the majority of alternatives. For HiC/3C analysis, a desirable feature of read mapping tools is the capability to report split read alignments (E.g. BWA MEM) (Li, 2013). Because of the potential for 3C reads to span the ligation junction, mappers reporting such alignments permit the experimenter the choice to discard or otherwise handle such events. Though we explored the effects of substituting alternative methods to a limited extent (not shown), both in terms of result quality and practical runtime considerations, a thorough investigation remains to be made.

The present implementation state of the simulation pipeline does not meet our desired goal for ease of configuration and broader utility. Of the numerous high-throughput execution environments (SLURM, PBS, SGE, Condor) in use, the pipeline is at present tightly coupled to PBS and SGE. It is our intention to introduce a grid-agnostic layer so that redeployment in varying environments is only a configuration detail. Although a single global seed is used in all random sampling processes, the possibility for irreproducibility remains due to side-effects brought on by variance in a deployment target's operating system and codebase. Additionally, though the pipeline and its ancillary tools are under version control, numerous deployment specific configuration settings are required post checkout. Preparation of a pre-configured instance within a software container such as Docker would permit the elimination of many such sources of variance and greatly lower the configurational barrier to carrying out or reproducing an experiment.

Many commonly used external validation measures (E.g. F-measure, V-measure) have traditionally not handled cluster overlap and were inappropriate for this study. Ongoing development within the field of soft-clustering (also known as community detection in networks) has, however, led to the reformulation of some measures to support overlap (Lancichinetti et al., 2009) or re-expression of soft-clustering solutions into a non-overlapping context (Xie et al., 2011). While a soft-clustering reformulation of normalized mutual information (NMI) (Lancichinetti et al., 2009) has become frequently relied on in clustering literature (Xie et al., 2013), alongside Bcubed the two have been shown to be complementary measures (Jurgens and Klapaftis, 2013). Therefore, although the choice to rely on the single measure we proposed here (Equation 1) is a possible limitation, it simultaneously avoids doubling the number of results to collate and interpret.

We chose to limit the representation of the combined WGS and 3C read data to a contig graph. While other representations built around smaller genomic features, such as SNVs, could in principle offer greater power to resolve strains, they bring with them a significant increase in graphical complexity. How more detailed representations might impact downstream algorithmic scaling, or simply increase the difficulty in accurately estimating a gold standard remains to be investigated.

Benchmark graph generators (so called LFR benchmarks) have been developed that execute in the realm of seconds (Lancichinetti et al., 2008; Lancichinetti and Fortunato, 2009). Parameterizing the mesoscopic structure of the resulting graph, their introduction is intended to address the inadequate evaluation of soft-clustering algorithms, which too often relied on unrepresentative generative models or *ad hoc* testing against real networks. Our pipeline may suffice as a pragmatic, albeit much more computationally intensive means of generating a domain specific benchmark on which to test clustering algorithms. Whether it is feasible to calibrate the LFR benchmarks so as to resemble 3C graphs emitted by our pipeline could be explored. Ultimately however, the parameter set we defined for the pipeline (Table 1) has the benefit of being domain-specific and therefore meaningful to experimental researchers.

Detection of overlapping communities in networks is a developing field and much recent work has left the performance of many clustering algorithms untested in deconvolving microbial communities via

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1974v1 | CC-BY 4.0 Open Access | rec: 20 Apr 2016, publ: 20 Apr 2016

3C read data. Not all algorithms are wholly unsupervised and individually fall into various algorithmic classes (i.e. clique percolation, link partitioning, local expansion, fuzzy detection and label propagation). Label propagation methods have shown promise with respect to highly overlapped communities (Xie et al., 2011; Chen et al., 2010; Gaiteri et al., 2015), which we might reasonably expect to confront when resolving microbial strains. Empirically determined probability distributions, such as those governing the production of intra-chromosomal (*cis*) read-pairs as a function of genomic separation, might naturally lend themselves to methods from within the fuzzy-detection class. With a generative community model in hand, exploring the performance of gaussian mixture models (GMM), mixed-membership stochastic block models (SBM) or non-negative matrix factorization (NMF) could be pursued.

The incomplete nature of graphs derived from experimental data can result in edge absence or edge weight uncertainty for rare interactions, with the knock-on effect that clustering algorithms can then suffer. We have shown that increase in 3C sampling depth (Figure 6) can significantly improve the quality of the resulting clustering solutions. A computational approach, which could potentially alleviate some of the demand for increased depth has been proposed (EdgeBoost) (Burgess et al., 2015) and shown to improve both Louvain and label propagation methods, is a clear candidate for future assessment.

CONCLUSION

For a microbial community, as intra-community evolutionary divergence decreases, contigs derived from WGS metagenomic assembly increasingly become a mixture of source genomes. When combined with 3C information to form a contig graph, evolutionary divergence is directly reflected by the degree of community overlap. In an effort to deconvolute simulated metagenomic assemblies into their constituent genomes, we tested the performance of both hard and soft clustering algorithms when applied to the contig graph. Performance was assessed by our proposed object-weighted variation of the extended Bcubed validation measure (Equation 1). We have shown that soft-clustering algorithms can significantly outperform hard-clustering algorithms when intra-community evolutionary divergence approaches that of bacterial strains. In addition, although increasing sampling depth of 3C read-pairs does little to improve the quality of hard-clustering solutions, it can noticeably improve the quality soft-clustering solutions. Of the tested algorithms, the precision of the hard-clustering algorithms often equalled or exceeded that of the soft-clustering algorithms across a wide range of evolutionary divergence. However, the poor recall of hard-clustering algorithms at low divergence greatly reduces their value in genomic reconstruction. We recommend that future work focus on the application of recent advances soft-clustering methods.

ACKNOWLEDGMENTS

We thank Steven P. Djordjevic for his support and helpful discussions. This work was supported by the AusGEM initiative, a collaboration between the NSW Department of Primary Industries and the ithree institute. We acknowledge the use of computing resources from the NeCTAR Research Cloud, an Australian Government project conducted as part of the Super Science initiative and financed by the EIF and NCRIS.

REFERENCES

Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., Loman, N. J., Andersson, A. F., and Quince, C. (2013). CONCOCT: Clustering cONtigs on COverage and ComposiTion.

Amigó, E., Gonzalo, J., Artiles, J., and Verdejo, F. (2009). A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information retrieval*.

Bagga, A. and Baldwin, B. (1998). Algorithms for scoring coreference chains. *The first international conference on language ...*.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology*, 19(5):455–477.

Beitel, C. W., Lang, J. M., Korf, I. F., Michelmore, R. W., Eisen, J. A., and Darling, A. E. (2014). Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ*, 2(12):e415.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1974v1 | CC-BY 4.0 Open Access | rec: 20 Apr 2016, publ: 20 Apr 2016

Belton, J.-M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods*, 58(3):268–276.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.

Boulos, R. E., Arneodo, A., Jensen, P., and Audit, B. (2013). Revealing long-range interconnected hubs in human chromatin interaction data using graph theory. *Physical Review Letters*, 111(11):118102.

Burgess, M., Adar, E., and Cafarella, M. (2015). Link-prediction enhanced consensus clustering for complex networks. *arXiv.org*.

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., and Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. 31(12):1119–1125.

Burton, J. N., Liachko, I., Dunham, M. J., and Shendure, J. (2014). Species-Level Deconvolution of Metagenome Assemblies with Hi-C-Based Contact Probability Maps. *G3 (Bethesda, Md.)*, 4(7):1339–1346.

Chen, W., Liu, Z., Sun, X., and Wang, Y. (2010). A game-theoretic framework to identify overlapping communities in social networks. *Data Mining and Knowledge Discovery*, 21(2):224–240.

Cleary, B., Brito, I. L., Huang, K., Gevers, D., Shea, T., Young, S., and Alm, E. J. (2015). Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nature Biotechnology*, 33(10):1053–1060.

Coil, D., Jospin, G., and Darling, A. E. (2015). A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data. *Bioinformatics (Oxford, England)*, 31(4):587–589.

Darling, A. E., Craven, M., Mau, B., and Perna, N. T. (2004). *Multiple alignment of rearranged genomes*. IEEE.

Dehmer, M. and Mowshowitz, A. (2011). A history of graph entropy measures. *Information Sciences*, 181(1):57–78.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. *Science*, 295(5558):1306–1311.

Gaiteri, C., Chen, M., Szymanski, B., Kuzmin, K., Xie, J., Lee, C., Blanche, T., Neto, E. C., Huang, S.-C., Grabowski, T., Madhyastha, T., and Komashko, V. (2015). Identifying robust communities and multi-community nodes by combining top-down and bottom-up approaches to clustering. *Scientific Reports*, 5:16361.

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics (Oxford, England)*, 29(8):1072–1075.

Hirschberg, J. B. and Rosenberg, A. (2007). V-Measure: A conditional entropy-based external cluster evaluation.

Huang, W., Li, L., Myers, J. R., and Marth, G. T. (2012). ART: a next-generation sequencing read simulator. *Bioinformatics (Oxford, England)*, 28(4):593–594.

Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., Dekker, J., and Mirny, L. A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods*, 9(10):999–1003.

Imelfort, M., Parks, D., Woodcroft, B. J., Dennis, P., Hugenholtz, P., and Tyson, G. W. (2014). GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ*, 2:e603.

Iqbal, Z., Turner, I., and McVean, G. (2013). High-throughput microbial population genomics using the Cortex variation assembler. *Bioinformatics (Oxford, England)*, 29(2):275–276.

Jurgens, D. and Klapaftis, I. (2013). Semeval-2013 task 13: Word sense induction for graded and non-graded senses. *Second joint conference on lexical and . . . .*

Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., and Chen, L. (2012). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature biotechnology*, 30(1):90–98.

Kiełbasa, S. M., Wan, R., Sato, K., Horton, P., and Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Res*, 21(3):487–493.

Konstantinidis, K. T., Ramette, A., and Tiedje, J. M. (2006). The bacterial species definition in the genomic era. 361(1475):1929–1940.

Lancichinetti, A. and Fortunato, S. (2009). Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 80(1 Pt 2):016118.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1974v1 | CC-BY 4.0 Open Access | rec: 20 Apr 2016, publ: 20 Apr 2016

Lancichinetti, A., Fortunato, S., and Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms. *Physical review E*.

Lancichinetti, A., Fortunato, S., and sz, J. n. K. (2009). Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015.

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: an R package for multivariate analysis. *Journal of statistical software*.

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics (Oxford, England)*, 31(10):1674–1676.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv.org*.

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., and Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293.

Liu, M. and Darling, A. (2015). Metagenomic Chromosome Conformation Capture (3C): techniques, applications, and challenges. *F1000Research*, 4:1377.

Marbouty, M., Cournac, A., Flot, J.-F., Marie-Nelly, H., Mozziconacci, J., and Koszul, R. (2014). Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. *eLife*, 3.

Marie-Nelly, H., Marbouty, M., Cournac, A., Flot, J.-F., Liti, G., Parodi, D. P., Syan, S., Guillén, N., Margeot, A., Zimmer, C., and Koszul, R. (2014). High-quality genome (re)assembly using chromosomal contact data. *Nature communications*, 5:5695.

Mowshowitz, A. and Dehmer, M. (2012). Entropy and the complexity of graphs revisited. *Entropy*.

Newman, M. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*.

Noble, W., Duan, Z.-j., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y. J., Lee, C., Shendure, J., Fields, S., and Blau, C. A. (2011). A Three-Dimensional Model of the Yeast Genome. In *Algorithms in Bioinformatics*, pages 320–320. Springer Berlin Heidelberg, Berlin, Heidelberg.

Peng, Y., Leung, H. C. M., Yiu, S. M., and Chin, F. Y. L. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics (Oxford, England)*, 28(11):1420–1428.

Pérez-Suárez, A., Martínez-Trinidad, J. F., Carrasco-Ochoa, J. A., and Medina-Pagola, J. E. (2013). OClustR: A new graph-based algorithm for overlapping clustering. *Neurocomputing*, 121:234–247.

Selvaraj, S., R Dixon, J., Bansal, V., and Ren, B. (2013). Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. 31(12):1111–1118.

Shih, Y.-K. and Parthasarathy, S. (2012). Identifying functional modules in interaction networks through overlapping Markov clustering. *Bioinformatics (Oxford, England)*, 28(18):i473–i479.

Stadhouders, R., Kolovos, P., Brouwer, R., Zuin, J., van den Heuvel, A., Kockx, C., Palstra, R.-J., Wendt, K. S., Grosveld, F., van Ijcken, W., and Soler, E. (2013). Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. *Nature protocols*, 8(3):509–524.

Treangen, T. J., Koren, S., Sommer, D. D., Liu, B., Astrovskaya, I., Ondov, B., Darling, A. E., Phillippy, A. M., and Pop, M. (2013). MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome Biol*, 14(1):R2.

Tringe, S. G. and Rubin, E. M. (2005). Metagenomics: DNA sequencing of environmental samples. *Nat Rev Genet*, 6(11):805–814.

van Dongen, s. (2001). *Graph Clustering by Flow Simulation*. PhD thesis, Utrecht University Repository, Utrecht University.

Xie, J., Kelley, S., and Szymanski, B. K. (2013). Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM Computing Surveys (csur)*, 45(4):43–35.

Xie, J., Szymanski, B. K., and Liu, X. (2011). *SLPA: Uncovering Overlapping Communities in Social Networks via a Speaker-Listener Interaction Dynamic Process*. IEEE.

Yaffe, E. and Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1974v1 | CC-BY 4.0 Open Access | rec: 20 Apr 2016, publ: 20 Apr 2016

to characterize global chromosomal architecture. *Nature Genetics*, 43(11):1059–1065.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1974v1 | CC-BY 4.0 Open Access | rec: 20 Apr 2016, publ: 20 Apr 2016

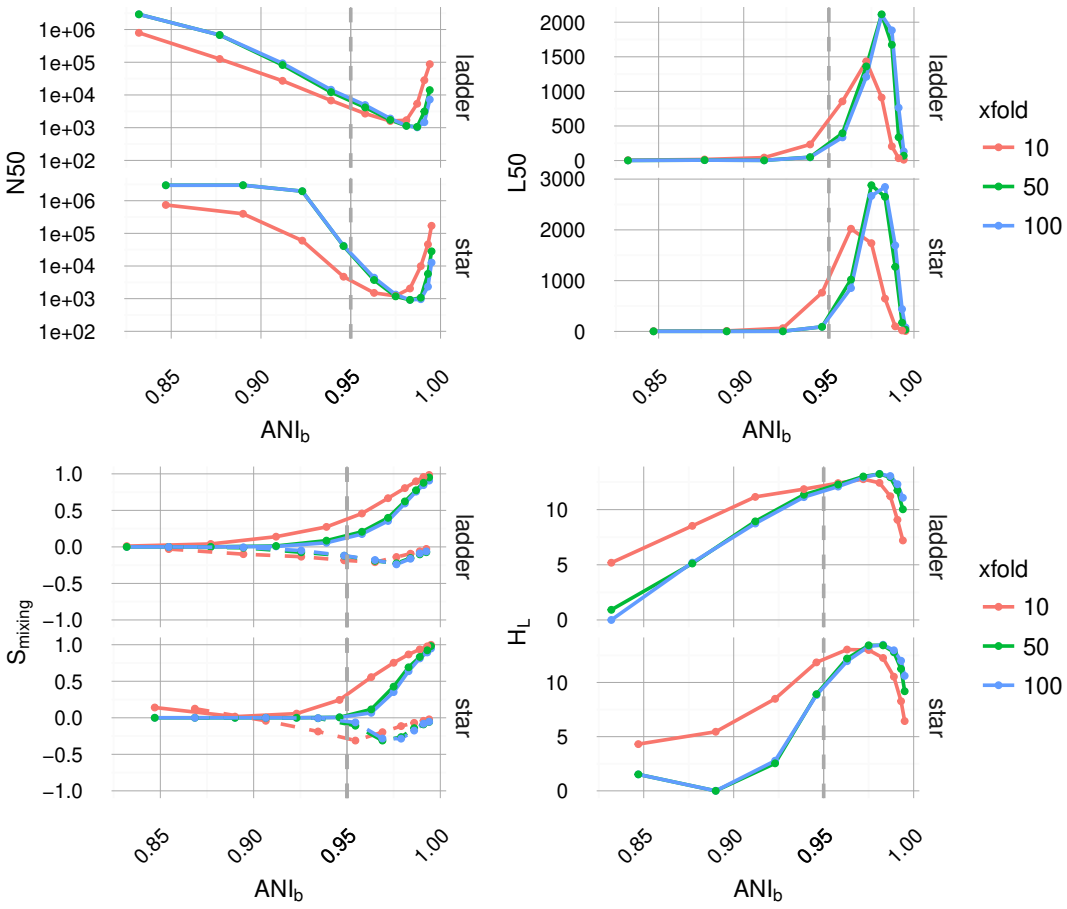

Figure 4. Plotted as a function of evolutionary divergence (measured by ANIb) for the star and ladder communities at three depths of WGS coverage (10, 50 and 100x); assembly validation statistics N50 (*top left*) and L50 (*top right*), the degree of genome intermixing Smixing and its approximate first order derivative ΔS_{mixing} (dashed lines) (*bottom left*), lastly graphical complexity H_L (*bottom right*). As community member similarity increases (ANIb $\rightarrow$ 1), assemblies go through a transition from a state of high purity ($S_{mixing} \approx 0$) to a highly degenerate state ($S_{mixing} \approx 1$), where many contigs are composed of reads from all community members. A crisis point is observed for small evolutionary divergence ($\alpha_{BL} < 0.2924$, ANIb $< 95\%$), where a sharp change in contiguity (implied by N50 and L50) occurs. At very low divergence, N50 and L50 statistics imply that assemblies are recovering, while source degeneracy (S_{mixing}) monotonically increases. Graphical complexity (H_L) exhibits a similar turning point to L50 and for the 3C contig graph is dominated by graphical order.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1974v1 | CC-BY 4.0 Open Access | rec: 20 Apr 2016, publ: 20 Apr 2016

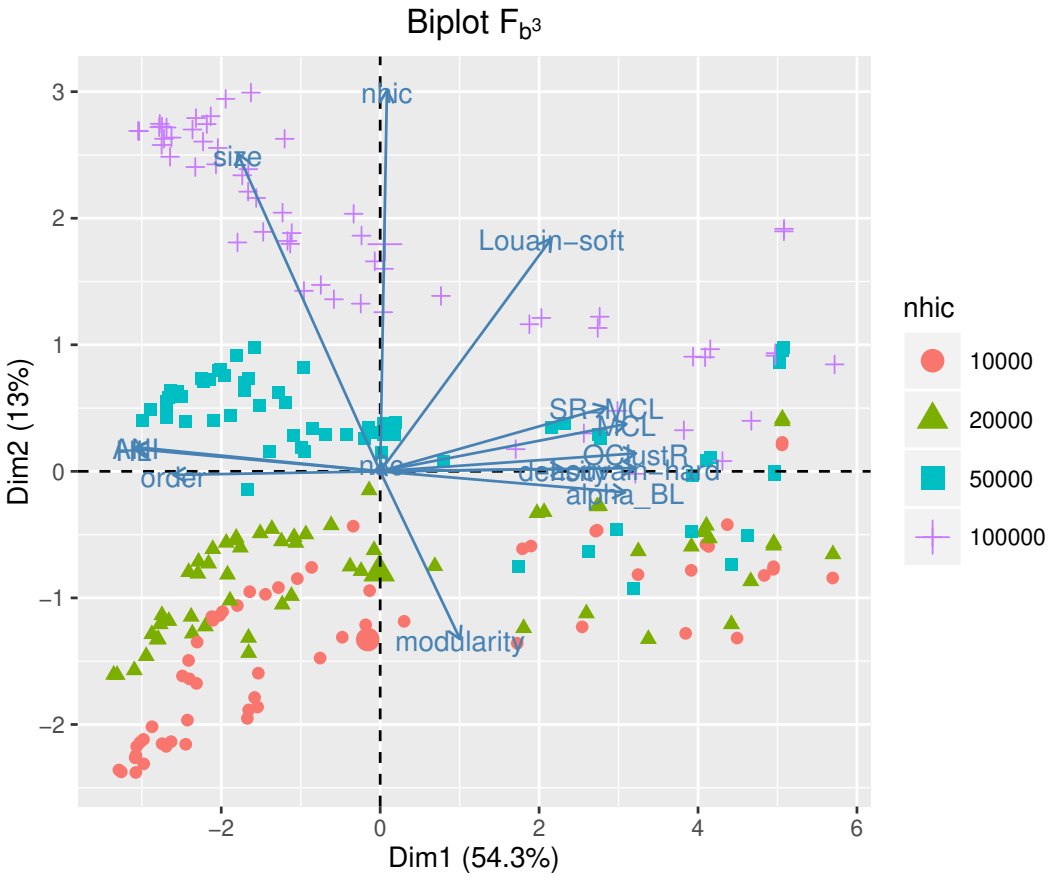

Figure 5. A PCA biplot of individuals and variables (first two components) for the 240 contig graphs pertaining to the uniform abundance profile. PC1 is most strongly correlated with graphical complexity (parameters: order, H_L), which comes about with decreasing evolutionary divergence (parameters: ANI, α_{BL}) and explains the majority of variation in performance for 4 of 5 clustering algorithms, with the notable exception being Louvain-soft. PC2 is related sampling depth and overall connectedness of contig graphs (parameters: size, n3c) with which Louvain-soft has significant positive correlation.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1974v1 | CC-BY 4.0 Open Access | rec: 20 Apr 2016, publ: 20 Apr 2016

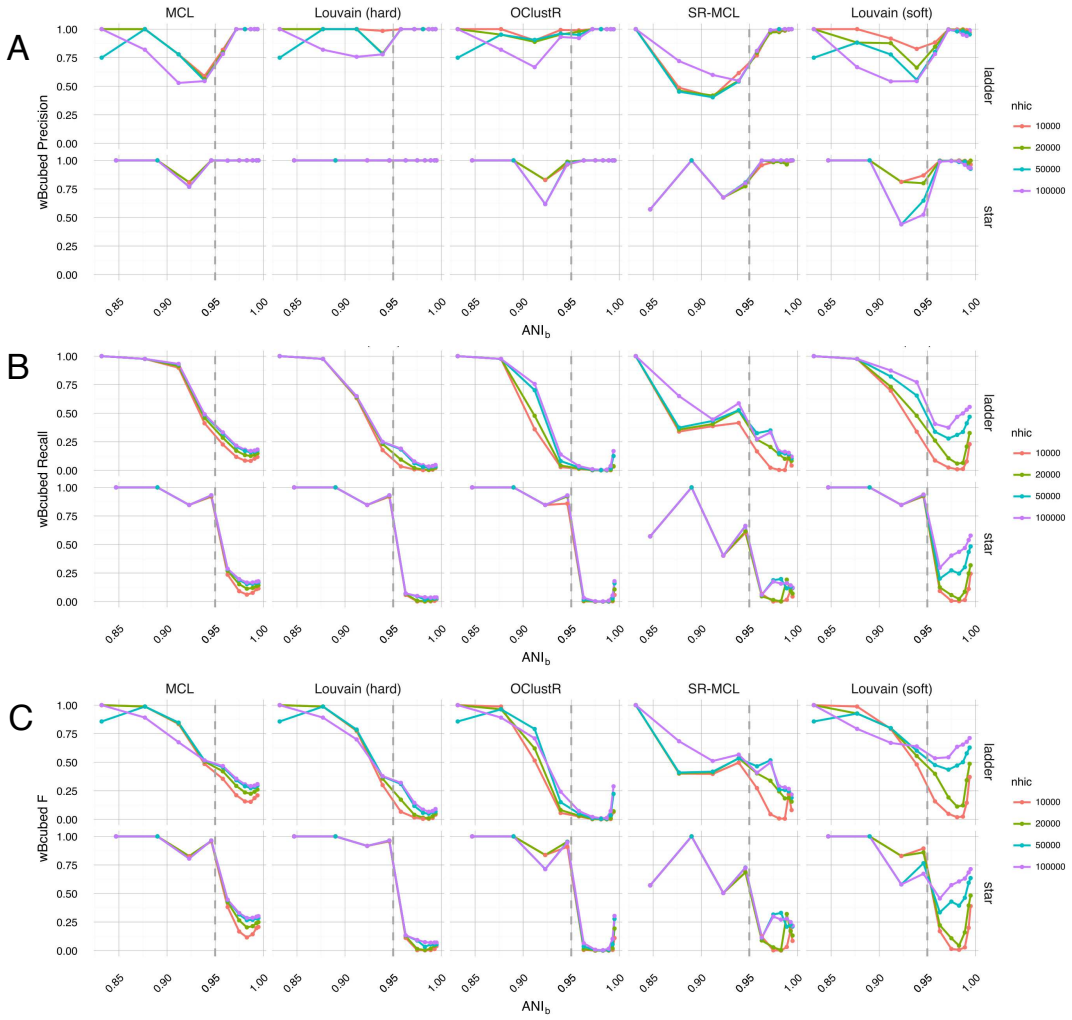

Figure 6. Performance of the five clustering algorithms (MCL, Louvain-hard, OClustR, SR-MCL, Louvain-soft), as measured by the weighted extended Bcubed F_{b^3} (Equation 1). The slice from the sweep pertains to uniform abundance and 100x WGS coverage and the best performing runtime parameters specific to algorithms (MCL, SR-MCL). (**A**) Louvain-hard demonstrates high precision throughout, while our simple modification Louvain-soft leads to a drop. (**B**) All algorithms struggle to recall the four individual genomes as evolutionary divergence decreases and cluster overlap grows. Within the region of overlap, Louvain-soft performs best and clearly gains from deeper 3C coverage. (**C**) In terms of F_{b^3}, the harmonic mean of Recall and Precision, only Louvain-soft appears to be an appropriate choice when it is expected that strain-level diversity exists within a microbial community.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1974v1 | CC-BY 4.0 Open Access | rec: 20 Apr 2016, publ: 20 Apr 2016