The integration of fine-scale DNA-DNA associations by inclusion of Hi-C DNA cross-linking information into metagenomic community analysis

by

Matthew Zachariah DeMaere

A Thesis Submitted In
Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

University of Technology Sydney
Sydney, NSW, Australia

February, 2019

Nature uses only the longest threads to weave her patterns, so that each small piece of her fabric reveals the organisation of the entire tapestry.

Richard Feynman - The Character of Physical Law: Messenger Lectures

Certificate of Original Authorship

I, Matthew Zachariah DeMaere, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Science at the University of Technology Sydney. This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.

Production Note:

Signature removed
prior to publication.

Date: 2019-03-11

Matthew Z. DeMaere

Acknowledgments

I would like to thank my principal supervisor Associate Professor Aaron Darling for his time, advice and support during my candidature. In particular, for supporting the original initiative, a patient ear for long digressions and tolerance for random-walk problem solving. In addition, I would like to thank my co-supervisor Professor Steven Djordjevic for his time, advice and crucial support. I would like to thank the ithree institute for fostering an environment conducive to quality research and their continual pursuit of a positive postgraduate student experience. I would also like to thank Professor Ian Charles for his valuable assistance in initiating and arranging the candidature.

I would like to thank my mother Jennifer, father James and sister Kathleen for their enthusiasm and continued support in what has been a *long* journey.

Lastly, I would especially like to thank my wonderful wife, Klára, for her ceaseless support despite the increasing complication of our lives and her ever expanding contribution to this endeavor.

Matthew Z. DeMaere
Sydney, Australia, 2019.

Thesis Format

A thesis by publication.

List of Publications

Peer reviewed

1. **DeMaere, M. Z.**, & Darling, A. E. (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
2. **DeMaere, M. Z.**, & Darling, A. E. (2018). sim3C: simulation of Hi-C and Meta3C proximity ligation sequencing technologies. *GigaScience*, 7(2). https://doi.org/10.1093/gigascience/gix103
3. **DeMaere, M. Z.**, Darling, A. E. (2019). bin3C: exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes. *Genome Biology*, 20(1), 46. https://doi.org/10.1186/s13059-019-1643-1

Contents

List of Figures

List of Tables

Abstract

Much of our understanding of the microbial world has been obtained using culture-based methodologies, a paradigm that has stood since the 19th century. And yet it has long been known that most of the Earth's microbial species are resistant to laboratory culture. It is reasonable to expect; therefore, that applying equal scrutiny to all microbial life will lead to significant discoveries. Motivated by this, metagenomics eliminates the culturing dependency by directly sampling DNA from an environment; successfully shedding light on the once unseen majority.

The technical limitations of present-day sequencing technologies have meant, however, that in achieving culture-independence, traditional shotgun metagenomic sequencing experiments make a considerable sacrifice. That sacrifice comes in the form of information loss where, in preparing DNA for sequencing, much of the "same-cell" and "same-chromosome" information is destroyed; information which is essential when reconstructing the individual genomes. Purely computational solutions to overcoming this sacrifice have proved insufficient; surpassed instead by strategies which employ changes in the experimental design aimed at reducing the information loss.

A recent strategy is the inclusion of a new form of sequencing data, provided by the Hi-C sequencing technique. Originally conceived to study the three-dimensional structure of chromatin, the Hi-C sequencing technique captures *in vivo* proximity interactions between DNA loci in an all-vs-all manner. When applied to direct metagenomic sampling, the physical structure of the microbial community (chromosome, cell and community) strongly influences the probability of observing proximity interactions between loci, and this pronounced modulation can be exploited to recover the information lost during shotgun sequencing.

This thesis details the effective integration of Hi-C into metagenomic sequencing studies to accurately reconstruct individual genomes, thereby deconvoluting the metagenome. To accomplish this, first an *in silico* investigation of the effectiveness of graph clustering as a means of metagenome deconvolution was conducted; where Hi-C proximity interactions defined the edges and assembly contigs defined the nodes. A parametric sweep of experimental and community composition parameters was carried out, exploring how the degree of evolutionary divergence (from species to strains) affected the quality of deconvolution. For each iterate in the sweep, a ground-truth was constructed and quality assessed using a novel external validation measure supporting overlapping clusters and variable object weights.

This work led to the design and implementation of the first metagenomic Hi-C read-pair simulator, sim3C, capable of simulating complex community definitions and simple three-dimensional structural elements. While in pursuit of the final objective of metagenome deconvolution, sim3C enabled an externally validated development process.

Lastly, as the outcome of the final objective, bin3C is demonstrated; an open-source solution to Hi-C driven metagenome deconvolution. In an unsupervised manner, bin3C reconstructs individual genomes from metagenomic data. Using external validation of simulated data, bin3C is shown to have high precision and good recall. When a real human microbiome was analysed, bin3C achieved leading performance, resolving 20 more nearly-complete MAGs (57% gain) than its closest competitor.

Dissertation directed by Associate Professor Aaron E. Darling
ithree institute, Faculty of Science
University of Technology Sydney

CHAPTER 1

Introduction

1.1 Background

1.1.1 The uncultured majority

Microorganisms represent a large portion of the Earth's biodiversity and dominate many ecosystems in sheer biomass [1]. Their varied metabolisms play key functional roles within biogeochemical cycles across terrestrial and aquatic ecosystems [2], [3]. Within host-associated ecosystems, a growing body of evidence links microbiota composition to health and disease [4], [5]. Advancing human health, foreseeing and perhaps mitigating the impact of climate change and understanding the biosphere as a whole will require a deep understanding of the microbial world.

While the significance and diversity of microbial life is clear today, for much of the 20th-century microbiology focused its efforts on the study of species which were readily cultured in the laboratory. As a depth-first rather than breadth-first discovery process, it forsook a wider understanding of the microbial world for detailed knowledge about those organisms which accommodated the culture-based paradigm. A direct outcome of the decades of detailed work was the development of molecular tools which have made possible the study of all microbial life, whether culturable or not. Redirecting a portion of the accumulated inertia behind culture-based research to new approaches was perhaps the first task for proponents of the uncultured majority.

Although the so-called "uncultured majority" [6]–[8] had been remarked upon as much as nearly a century ago [9], addressing it did not begin to occur properly until the latter part of the 20th century [10]–[12]. It was at this time that the intractable morphological approach to bacterial phylogeny was finally overcome with the advent of DNA sequencing [13], [14] and the demonstration that ribosomal RNA (rRNA) genes were effective molecular clocks [12]. The development of polymerase chain reaction (PCR), and subsequently amplicon sequencing, both simplified and the enhanced the sensitivity of microbial marker surveys. This increasingly refined and targeted approach to diversity studies became the *de facto* standard and led to the founding of numerous publicly accessible phylogenetic databases [15]–[18]. Since their inception, these rRNA databases have grown exponentially in size, becoming significant information warehouses. As of release 132 the Silva SSURef database, housing only high-quality small-subunit rRNA sequences, contains more than 2 million records [19].

Phylogenetic marker surveys, however, do not suffice for all the questions science would like answered. Being targeted, they provide no information about an organism save for an indication of identity and its relative abundance. To infer function, studies must associate observed marker sequences with sufficiently similar examples from within a database of already well-characterised species, where characterisation might include genome sequencing or laboratory assays. This can be problematic in that, under a culture-based regime, species which resist cultivation are less likely to be well-characterised. Further, with as little as 39% of genes shared within a species [20], the conservation of phenotypic characteristics (or microbial traits) is not necessarily well reflected by phylogenetic relatedness [21]. Important aspects such as anti-microbial resistance, pathogenicity and niche exploitation can be acquired horizontally within an accessory genome, rather than existing within the vertically inherited core genome [22]. Thus, even with a fully characterised phylogenetic database, marker-based prediction of the realized phenotypic traits in an environment is unreliable. Instead, elucidating a species global and *in situ* pan-genomes is crucial to understanding both realized and potential behaviour.

1.1.2 Metagenomics

Metagenomic shotgun sequencing attempts to address these issues by sampling genetic information directly from an environment in a culture-independent manner. From its inception onward [23]–[25], as the technique has been brought to bear on various ecosystems, it has successfully shed light on the uncultured majority [4], [26], [27].

Current metagenomics relies on second-generation sequencing technologies [28] but makes a significant sacrifice in doing so. Relative to first-generation sequencing technologies, and even early second-generation sequencers such as the Roche 454, today's second-generation sequencers can produce huge yields measured both in the number of base-pairs (bp) and the number of reads. Although these massively-parallel machines have been steadily refined, a significant technological limitation remains in library preparation.

Input genomic DNA used as the template in library preparation must be first sheared into small fragments, on the order of 1000 bp or shorter. In addition, sequencing experiments are performed in an all-at-once shotgun style, where cells are lysed and the DNA extracted and purified.

In a clonal setting, the bulk DNA resulting from cell lysis is genomically homogeneous and simply an amplification process prior to sequencing, where all reads derive from the same genomic source. In a metagenomic setting, however, when the cells collected from an environment are lysed, all the DNA from all the species becomes intermixed and there are many potential genomic sources. Cell lysis, thus, destroys all connection between a cellular source and the genomic content it contained. DNA shearing further complicates the problem, destroying the long-range contiguity relationship between the resulting fragments. Together, the steps of cellular lysis and DNA shearing represent significant information loss in a metagenomic sequencing experiment.

Although current *de novo* metagenome assembly algorithms are capable of reconstructing a portion of the contiguity relationships from short-read sequencing, the problem is a significant challenge and the reconstructions are far from complete [29]. Even in a clonal setting, due to repetitive elements, the complete reconstruction *de novo* is unlikely without the assistance of more recent long-read sequencing technologies [30]–[32]. At present, long-read sequencing is not as well suited to metagenomics due to higher error rates and lower sequencing depth [33]. Even when long-read sequencing becomes more applicable to metagenomics, increased contiguity does nothing to address the lost cellular locality. Techniques which can capture or recover cellular locality would greatly benefit metagenome reconstruction, allowing at least the grouping of the fragmentary DNA sequences by microbial source. Beyond more thoroughly reconstructing the genomes of *in situ* species, such techniques could also be used to associate mobile elements such as plasmids, as well as study virus to host and cell to cell interactions.

1.1.3 Metagenome-assembled genomes

By current estimates, there are approximately 1 trillion (10^{12}) bacterial species on the Earth [34], and this number grows larger when consideration of strains within species is included. If a full reckoning of the planet's diversity is to be achieved in a practical amount of time, we must continually seek out and deploy high-throughput, low-bias knowledge gathering processes. Using direct sampling, metagenomics already eliminates much of the bias that exists in culture-base studies. In terms of genome sequencing, parallelism and thus increased throughput could be obtained by the refinement of metagenomics to the point that the extraction of individual genomes is systematic, routine and cost-efficient.

Recently, standards were proposed for reporting genomes isolated from metagenomic data [35], termed metagenome-assembled genomes (MAGs). Announced alongside a similar standard for single amplified genomes (SAGs), this publication by the Genome Standards Committee marks an increasing acceptance of genomes not determined through the traditional process of isolation and cultivation.

From early on, metagenomic sequencing has aimed to resolve or partition the contigs from assembly or the raw reads into genome bins. The first generation of such methods employed only intrinsic features such as GC content, kmer frequencies and depth of coverage within a single sequencing experiment [36], [37]. Refined implementations have managed to resolve high abundance species from real microbiomes, but struggle as abundance decreases [38] and possess undesirable median contamination rates (chapter 4).

Between 2004 and 2018 there has been a 10000-fold drop in the cost of DNA sequencing and proportionate increase in base-pair yield [39]. As a result, approaches which rely on deep sequencing have become much more accessible. Multi-sample metagenomic studies (timeseries and transects) have sought to leverage the correlated change in abundance across sample-points to improve the precision and accuracy of metagenome binning. These second generation methods have shown to be powerful, having resolved metagenomes down to at least species-level resolution [40]–[45]. Evidence of the success of these methods can be found in the public archive deposition of thousands of, so-called, metagenome-assembled genomes from studies where they have been employed [46]–[48].

Despite their power, however, the requirement of multiple samples (in some cases more than 200 [47]) is a cost burden that not all labs can handle. Beyond the question of cost, multiple samples can be a logistical barrier in cases such as clinical studies, where only a single time-point is available from a patient.

1.1.4 Validation of metagenome-assembled genomes

As the extraction of MAGs becomes increasingly common, a means of assessing the quality of these constructs has become a necessity. It does not suffice to report only the standard genomic statistics used in clonal genome sequencing studies. Isolated from a background of potentially hundreds of other species and strains, a seemingly small error can result in a highly contaminated genome bin. To be of value in downstream analyses, then, MAGs must be thoroughly validated.

Current tools which infer completeness and contamination use curated databases of taxonomically associated single copy marker genes [49]–[51]. As with any system reliant on an incomplete repository of reference data, these methods are imperfect. For instance, the CheckM reference database focuses on bacterial and archaeal marker genes, with the authors warning of its limitations with eukaryotes, phage and plasmids. BUSCO aims for wider phyletic coverage but does so using fewer marker genes, possibly resulting in less sensitivity. In all these approaches, the presence or absence of the non-marker gene content is not directly assessed. In an extreme pathological scenario, the possibility exists that an extracted MAG containing all the marker genes but only a small fraction of its entire genome would be assessed as complete. Therefore, the marker-gene approach to validation has its limitations, which must be kept in mind.

Reporting completeness and contamination are now enshrined in the standard most likely to be adopted by the public sequencing archives [35]. An alternative standard has been proposed and used in the literature concerning MAG extraction [50]. This standard is convenient due to its simplicity (Table 4.1). Here, a nearly-complete low-contamination genome would possess ≥ 90% of its expected marker genes and ≤ 5% with conflicting taxonomy.

1.1.5 Chromosome conformation capture

Chromosome conformation capture (3C) was originally conceived as a PCR based approach for observing *in vivo* the frequency of interaction between two targeted genomic loci, where these loci can be on the same or different chromosome [52]. Incorporating additional technologies, successive method iterations (4C, 5C, and Hi-C) aimed to improve on this original idea. Extending the one-vs-one of 3C to a one-vs-many, chromosome conformation capture on-chip (4C) compares a single loci's interactions against a genome-wide microarray [53]. The next advancement, chromosome conformation capture carbon copy (5C) extends the method to many-vs-many, comparing a wider genomic region (< 1 Mbp) against a genome-wide microarray [54]. Utilising high-throughput sequencing, Hi-C was the first method to perform an all-vs-all genome-wide interrogation of loci interactions [55].

The foundation of all 3C-based methods, the Hi-C protocol begins with the step of *in* *vivo* formalin fixation, crosslinking proteins bound to DNA while keeping cells intact. Next, the crosslinked DNA is extracted and purified. The purified DNA-protein complexes are then restriction digested to expose free-ends. These free-ends are then biotinylated and blunted. The DNA-protein complexes are placed under dilute conditions or immobilised on a solid substrate and the free-ends ligated. In this state, while held within the complexes, free-ends which were spatially close *in vivo* have a much greater probability of ligation than the random ligation of any two free-ends in solution. The crosslinking is then reversed, proteins digested and the DNA sheared in preparation for library construction. Biotin affinity purification is used to enrich for DNA fragments containing the proximity-ligation junctions (PL). Lastly, the purified PL fragments are used as the template DNA to create an Illumina paired-end sequencing library. After sequencing is completed, each end of a PL containing read-pair corresponds to different locations within the genome. The frequency of proximity (or interaction) between genomic regions can be established by mapping these read-pairs back onto a known reference sequence. This reference was originally the human genome but less well-characterised subjects of study, such as assembly scaffolds or contigs, can suffice (Figure 1.1).

While both 4C and 5C have subsequently been extended from relying on microarrays to using high-throughput sequencing [56], [57], Hi-C has come to dominate in many applications. In the seminal paper, the method was used successfully to confirm the existence of chromosome territories and determine that human chromatin is arranged as a fractal globule [55]. In clonal studies Hi-C has provided new insights on an array of topics such as chromosomal topologically-associated domains (TADs) [58], haplotype phasing [59], genome reassembly [60], supervised assembly clustering [61], centromere prediction [62], host-virus interactions [63], and epigenetics [64]. This growing list of applications which successfully leverage Hi-C data does well to highlight its inherent value in overcoming the many problems in genomics which stem from NGS information loss.

1.1.6 Addressing Metagenomics with Hi-C

The potential utility of Hi-C as an approach to metagenome deconvolution has been investigated independently by two groups [65]–[67]. By combining Hi-C read-pairs with contigs derived from a conventional shotgun metagenome assembly, it was hoped that proximity-based associations between contigs could be constructed and that these associations could then be used to deconvolute the metagenome. Importantly, if this

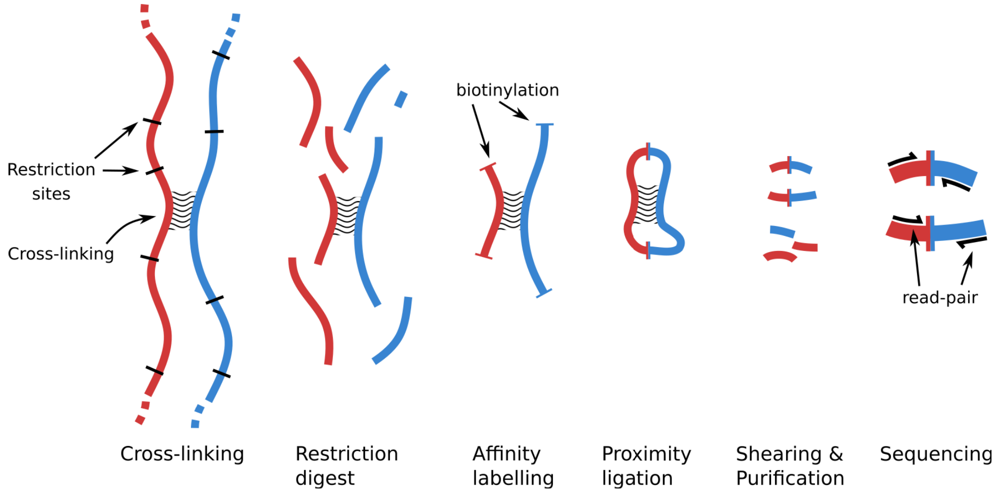

Figure 1.1: An outline of the steps involved in the Hi-C library protocol. From left to right, the protocol begins by cross-linking DNA *in situ*, after which cells are lysed and the DNA restriction digested. Next, free ends are affinity labelled with biotin and made blunt. Next, in dilute conditions or solid substrate immobilised, free ends are ligated, forming biotin labelled proximity ligation junctions. Next, cross-linking is reversed and bound proteins digested, followed by DNA shearing and purification. Lastly, sequenced as an Illumina paired-end library, each read in the pair samples one end of the junction fragment.

approach proved successful, deconvolution would require only a single time-point, a significant advantage over the correlated abundance approach which requires multiple time-points. In these first studies, what was important to learn was whether the physical structure of the community would be apparent in the Hi-C data.

As it turned out, the physical structure does, indeed, modulate the observed interaction rates; intra-chromosomal interactions are the most frequent, followed by inter-chromosomal and lastly inter-cellular [65]. Further, the degree of modulation is significant. On average, the three rates are separated by an order of magnitude and therefore offer a strong signal with which to assign contigs into “same cell” if not “same chromosome” bins.

With this in mind, it is no surprise perhaps that Hi-C mediated metagenome deconvolution was successfully demonstrated by both groups on synthetic microbial communities constructed in the laboratory [65], [66].

Beitel *et al.* [65] used the Markov clustering algorithm (MCL) [68] to partition the metagenome and employed reference genomes to quantitatively validate the binning solutions against a ground truth. At the species level, the authors found that their approach was able to obtain a near perfect solution (precision 0.96, recall 0.98) when ideal parameters were supplied to MCL. However, the authors also found that the resolution of closely related strains into separate bins was not achievable.

Since these initial experiments, Hi-C deconvolution has been employed on real microbial communities [69], [70]. On a low-complexity yeast dominated fermenter culture, Heil *et al.* [69] employed an *ad hoc* method to identify a novel inter-specific hybrid yeast genome. In a more complex human gut microbiome, Press *et al.* [70] extracted 35 nearly-complete metagenome-assembled genomes from a single sample. This result was encouraging when it was compared to a state of the art genome binner which does not make use of Hi-C, MaxBin [38], where Press *et al.* found that MaxBin could resolve only 20 nearly-complete genomes. In addition, the Hi-C genome bins were on average less contaminated than those of MaxBin. One issue of the Press *et al.* method, however, is that it is offered only as a proprietary service, which raises concerns in regard to the initiatives of open science [71] and the reproducibility of future studies which opt to use this closed service.

1.1.7 Graphical Model

Whether clonal or metagenomic in nature, the incorporation of Hi-C data into genome sequencing workflows has frequently been done in a graph-theoretic manner. Such representations are appealing as they can intuitively reflect the DNA-DNA proximity information captured by Hi-C and, importantly, allow researchers to take advantage of powerful analytical approaches from within network science, such as community detection [72].

The simplest such construct is a contig graph, where contigs become nodes and the strength of Hi-C interactions between contigs define edge weights. The contig graph is a weighted undirected graph, where edge weights in their raw form equate to the observed number of PL events between each pair of contigs. While more fine-scaled representations are easily constructed, such as the Hi-C linked variant graph used by the haplotype phaser HapCUT2 [59], the contig graph suffices for the purpose of metagenomic deconvolution (genome binning) to at least the level of species.

It is, in fact, particularly desirable to employ the simplest possible representation which elicits satisfactory solutions. The tendency for real-world problem sizes to grow substantially with technological advances in data collection means attention must be paid to algorithmic time and space complexity. Large datasets, full of scientific promise, can quickly make what was a viable but inefficient approach computationally intractable.

Computational efficiency is, therefore, an important consideration in metagenome deconvolution as a thoroughly sequenced microbial community of moderate complexity can possess more than half a million contigs when assembled [70], [73].

A metagenome, depicted as a contig graph, allows the deconvolution problem to be formulated as a community structure detection problem, where inferred communities correspond to genome bins. Here, ideally only the primary physical organisation of the community (cell envelopes and chromosomes) would be responsible for modulating the frequency of interaction between contigs. Other factors affecting the counting process include the relative abundance of species in the community and experimental factors such as differences in lysis efficiency, enzyme cut-site density and DNA accessibility. If left unaccounted for, these confounding factors can significantly influence algorithms searching for community structure. Deconvolution algorithms must, therefore, either account for these factors in their model or the contig graph must be normalised beforehand. Several articles devoted to Hi-C normalisation have been published [74]–[77]. ‘Iterative correction and eigendecomposition’ (ICE) is a popular and parameter-free approach. Initially regarded as a novel solution, it has since been shown [78] that ICE is equivalent to a conventional and long-established algorithm for determining doubly-balanced matrices [79].

In matrix form, a metagenomic Hi-C contig graph is highly sparse. That is to say, for most pairs of contigs in the assembly, no PL events will be observed between them. As the Sinkhorn-Knopp algorithm is known to have issues with sparsity [80], it is fortunate that a more recent and faster converging algorithm has resolved the issue [81]. Besides its use in bin3C (chapter 4) [73], the Knight-Ruiz algorithm has been incorporated into at least one other Hi-C analysis pipeline, Juicer [82].

1.1.8 Community Detection

With the advent of large social networks, a combination of financial backing and human fascination has driven progress in graph clustering. The developing field of community detection focuses on inferring the community structure of such complex networks with minimal reliance on *a priori* knowledge; this an essential feature when a researcher knows or wishes to assume little about the subject of study.

It is here, when little is known and a minimum of prejudice is desired, that frequently used traditional graph clustering methods are ill-suited to community detection. For example, graph partitioning algorithms, such as Kernighan-Lin and spectral bisection, require the number of clusters be known. Partitional clustering algorithms such as k-means, while also requiring the number of clusters further impose a potentially unnatural embedding space and metric on the graph. In hierarchical clustering, such as agglomerative, though the number of clusters is not required, a similarity measure must be chosen and the it is left to the researcher on how best to partition the resulting hierarchy [83].

Over the last two decades, a steady stream of community detection algorithms with increasing power and computational efficiency have been devised. They can be categorised by their major conceptual elements, such as: flow-based [68], [84], [85], modularity-based [86]–[88], label-propagation [89], and statistical [90]. The development of these algorithms has been supported by work on both external validation measures [91]–[93] and the introduction of the so-called LFR benchmarks [94], [95].

1.1.8.1 Validation Testing

External validation measures (indices) compare a given clustering solution against a ground truth, summarising its validity as a value on the unit interval. Though appearing to be a simple procedure, a range of subtleties exist on how different modes of disagreement are assessed. Additionally, not all measures support the notion of overlap among clusters and some require label matching between the ground truth and clustering (an $O(n^3)$ problem).

Work to enumerate and formalise the necessary constraints of an ideal measure has helped to understand why different popular measures disagree by identifying their individual deficiencies [91], [92], [96]. As no ideal measure has been devised, it is best practice to employ multiple well-behaving measures and take the time to thoroughly understand the strength and weakness of each. Currently, normalized information distance (NID) [97] , adjusted mutual information (AMI) [98], Extended BCubed (B^3) [92] and adjusted Rand index (ARI) [99] are some external measures with better behaviour.

The LFR benchmarks are graph generating algorithms which parametrize the major features of community structure (overlap, node degree and community size distributions) [94], [95]. LFR generated graphs are intended to more closely model real-world network characteristics and importantly the algorithms provide a ground truth of community membership. By providing this auxiliary information, they permit straightforward external validation testing of community detection algorithms.

1.1.8.2 Resolution Limits

One mathematical notion of community structure is termed Modularity [100] and maximising modularity as an objective function has become perhaps the most common basis used in community detection [86]–[88]. Although these methods tend to perform well, it has been shown that the modularity relation possesses a resolution limit [101]. The result is that, for larger and more complex graphs, modularity-based methods can struggle to detect smaller and fainter communities. Another serious issue with modularity maximisation is that, as an objective function, there exist many near-maximal degenerate solutions with fundamentally different small partitions [102].

An alternative community detection algorithm which does not suffer from a resolution limit [103] is the flow-based Infomap [85], [104]. As an information-theoretic approach, Infomap employs the notion of minimum description length (MDL) and attempts to describe the path taken by a random walker in the least bits. To do so, Huffman coding is used to assign codewords to nodes, where the frequency of visitation defines code length and an index codebook creates a 2-level system, allowing modules (communities) with the graph to use separate codings. The premise of the method is that the optimal community definition emerges simultaneously with the most efficient description of random walks. The authors refer to this as inference-compression duality.

In its default hard-clustering mode, Infomap possesses efficient time and space complexity. For example, the clustering of a contig graph with 29,653 nodes and 1,596,922 edges completed in 5 minutes and required 11.6 GB of memory on an Intel Xeon E5-2697 CPU. As Infomap does not suffer from a resolution limit, it is well suited to metagenomic contig graphs, which frequently possess a wide range of community size and weight due, in part, to the long-tailed abundance profiles in naturally occurring ecosystems [26], [105]. For these reasons, Infomap is bin3C's default community detection algorithm (chapter 4).

1.2 Research Aim

The overall aim of this research project was to investigate to what extent Hi-C DNA-DNA proximity information could be leveraged to deconvolve a metagenome accurately into the constituent species. As successful proofs-of-concept on simple synthetic communities had already been demonstrated, this project's focus was to advance and refine this basic premise.

Concerning advancement, the aim was that the final method should possess high precision while reconstructing as much of the community as possible with no *a priori* information on community structure. In terms of refinement, the final approach is to be encapsulated in a software tool, ready for easy use by the general research community and with sufficient computational efficiency as to permit the analysis of large and complex metagenomes.

1.3 Outline of Thesis

1.3.1 Objectives

The project has been divided into three primary objectives with each objective corresponding to one chapter of this thesis. In kind, the outcome of these objectives are detailed in chapters 2, 3 and 4; each of which represent a single peer-reviewed paper, with the exception that chapter 4 is presently under review following a favourable presubmission enquiry.

Objective 1 ⇔ Chapter 2 ⇔ Paper 1

An exploratory analysis of the problem space.

Objective 2 ⇔ Chapter 3 ⇔ Paper 2

Develop a refined Hi-C read simulator.

Objective 3 ⇔ Chapter 4 ⇔ Paper 3

Develop a refined solution for Hi-C metagenome deconvolution.

1.3.2 Chapter Summaries

Chapter 1

As an introduction to the topic, chapter 1 begins by outlining the major shortcomings of adhering to a culture-based paradigm in microbial genomics and ecology. This is followed by describing how as a culture-independent approach, metagenomics addresses these limitations while at the same time sacrificing information. Next, describing how the application of Hi-C to metagenomics aims to recover the lost information by means of a graph-theoretic approach. Lastly, a brief discussion on current community detection algorithms.

Chapter 2

Although proofs-of-concept on synthetic communities had been demonstrated [65], [66], it was not known how, as a system, the approach would respond to changes in depth of sequencing, community composition or the clustering algorithm. In chapter 2, I answer this question by way of an *in silico* parametric sweep, exploring how the quality of the deconvolution solutions respond to variation in evolutionary divergence and sequencing depth. As only non-overlapping clustering algorithms had been used in the proofs-of-concept, the value of overlapping soft-clustering algorithms is also assessed.

To quantitatively assess the sweep two subordinate objectives were identified: 1.1 and 1.2. The products of these suborbdinate objectives were subsequently employed in pursuit of the third primary objective in chapter 4.

Objective 1.1: A method to infer deconvolution ground truths from simulated metagenomes.

Objective 1.2: The definition of an external validation measure for overlapping clusters and, significantly, variable object value.

As we wished to quantitatively assess how the performance of the prototype deconvolution method responded to the sweep, this required that the simulation pipeline also produce an overlapping ground truth. Further, as longer contigs represent proportionately larger parts of any genome binning solution, we desired an external validation measure which could treat the notion of object value as well as overlap.

Chapter 3

In chapter 3, I detail the implementation and use of sim3C, the first open-source metagenomic Hi-C read simulator intended for use by the wider research community. A major refinement of the prototype implemented to achieve objective 1, sim3C includes support for Hi-C, meta3C and DNase Hi-C library protocols, the random simulation of topologically-associated domains and flexible community definitions.

Chapter 4

In chapter 4, I detail the implementation and performance of bin3C, the first open-source Hi-C driven genome binning tool. Bin3C's performance when deconvolving both real and simulated metagenomes into metagenome-assembled genomes (MAGs) is assessed. A performance comparison is made between bin3C, the commercial service ProxiMeta [70], and the non-Hi-C genome binner MaxBin [38]. The comparison demonstrates that bin3C currently represents the state of the art in Hi-C metagenome deconvolution.

1.4 Further Work

A question remaining in metagenomics is, for a given ecosystem, how does the degree of genomic microheterogeneity vary between species and within species in response to changes in environmental conditions. To answer this effectively will require deconvolution methods with a resolution at the strain level. This should be achievable with sufficient Hi-C depth of coverage and the progressive accumulation of variant site linkage within individual genotypes. A remaining significant contribution to this problem is the development of a tractable statistical model or graph construct with additional strain-level detail.

Assembly contigs are the current target of Hi-C deconvolution, however as a reduction of the much more information-rich assembly graph, using them as the basis of deconvolution represents another source of information loss. This is particularly the case with metagenomic assembly graphs which are complicated by the existence of closely related genotypes and shared sequence. Hi-C reads could, instead, be mapped directly to the assembly graph using such tools as vg [106], which may prove effective in improving resolving power and recall.

Rather than only assign assembly fragments to genome bins, reconstructing the chromosomal order of the fragments would be desirable in certain downstream analysis contexts. The strong power-law decay in Hi-C interaction strength with increasing loci separation (chapter 3) should provide sufficient signal to accomplish this task. While both Lachesis [61] and GRAAL [60] have addressed scaffolding, this was only in the context of clonal samples. A sscaffolder capable of functioning in a metagenomic context would be a useful contribution.

Community detection algorithms continue to improve, however, another alternative is consensus clustering [107]. Combining the clustering solutions of multiple algorithms, these meta-clustering algorithms attempt to reconcile the differences and agreements to produce improved solutions. Applying some form of consensus clustering when clustering the contig graph, or future more complex construct may prove fruitful.

Rather than clustering the same graph with multiple algorithms, multi-layer network clustering utilises multiple graphs over the same set of nodes [108]. For Hi-C metagenome deconvolution, one choice of multilayer graph would be the basic contig graph combined with the corresponding assembly graph. This would help reduce the bias against the large number of short contigs which suffer from mappability problems.

1.5 List of Abbreviations

- AMI - adjusted mutual information
- ARI - adjusted Rand index
- bp - base-pair
- CPU - central processing unit
- DNA - deoxyribonucleic acid
- GB - gigabyte
- ICE - iterative correction and eigendecomposition
- MAG - metagenome-assembled genome
- Mbp - megabase-pair
- MCL - Markov clustering
- MDL - minimum description length
- NGS - next-generation sequencing
- NID - normalized information distance
- PCR - polymerase chain reaction
- PL - proximity ligation
- rRNA - ribosomal ribonucleic acid
- SAG - single amplified genome
- TAD - topologically-associated domain

1.6 References

[1] L. Zinger, A. Gobet, and T. Pommier, “Two decades of describing the unseen majority of aquatic microbial diversity”, en, *Mol. Ecol.*, vol. 21, no. 8, pp. 1878–1896, Apr. 2012, ISSN: 0962-1083, 1365-294X. DOI: 10.1111/j.1365-294X.2011.05362.x. [Online]. Available: http://dx.doi.org/10.1111/j.1365-294X.2011.05362.x.

[2] M. G. A. van der Heijden, R. D. Bardgett, and N. M. van Straalen, “The unseen majority: Soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems”, en, *Ecol. Lett.*, vol. 11, no. 3, pp. 296–310, Mar. 2008, ISSN: 1461-023X, 1461-0248. DOI: 10.1111/j.1461-0248.2007.01139.x. [Online]. Available: http://dx.doi.org/10.1111/j.1461-0248.2007.01139.x.

[3] F. Azam and F. Malfatti, “Microbial structuring of marine ecosystems”, en, *Nat. Rev. Microbiol.*, vol. 5, no. 10, pp. 782–791, Oct. 2007, ISSN: 1740-1526, 1740-1534. DOI: 10.1038/nrmicro1747. [Online]. Available: http://dx.doi.org/10.1038/nrmicro1747.

[4] Human Microbiome Project Consortium, “Structure, function and diversity of the healthy human microbiome”, en, *Nature*, vol. 486, no. 7402, pp. 207–214, Jun. 2012, ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature11234. [Online]. Available: http://dx.doi.org/10.1038/nature11234.

[5] J. L. Round and S. K. Mazmanian, “The gut microbiota shapes intestinal immune responses during health and disease”, en, *Nat. Rev. Immunol.*, vol. 9, no. 5, pp. 313–323, May 2009, ISSN: 1474-1733, 1474-1741. DOI: 10.1038/nri2515. [Online]. Available: http://dx.doi.org/10.1038/nri2515.

[6] J. T. Staley and A. Konopka, “Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats”, en, *Annu. Rev. Microbiol.*, vol. 39, pp. 321–346, 1985, ISSN: 0066-4227. DOI: 10.1146/annurev.mi.39.100185.001541. [Online]. Available: http://dx.doi.org/10.1146/annurev.mi.39.100185.001541.

[7] M. S. Rappé and S. J. Giovannoni, “The uncultured microbial majority”, en, *Annu. Rev. Microbiol.*, vol. 57, no. 1, pp. 369–394, 2003, ISSN: 0066-4227. DOI: 10.1146/annurev.micro.57.030502.090759. eprint: http://dx.doi.org/10.1146/annurev.micro.57.030502.090759. [Online]. Available: http://dx.doi.org/10.1146/annurev.micro.57.030502.090759.

[8] P. Hugenholtz, “Exploring prokaryotic diversity in the genomic era”, *Genome Biol.*, vol. 3, no. 2, reviews0003.1, Jan. 2002, ISSN: 1465-6906, 1474-760X. DOI: 10.1186/

gb-2002-3-2-reviews0003. [Online]. Available: https://doi.org/10.1186/gb-2002-3-2-reviews0003.

[9] A. S. Razumov, “The direct method of calculation of bacteria in water: Comparison with the koch method”, Russian, *Mikrobiologija*, vol. 1, pp. 131–146, 1932. [Online]. Available: https://github.com/klloydbeaufort/great-plate-count-translated.

[10] W. B. Whitman, D. C. Coleman, and W. J. Wiebe, “Prokaryotes: The unseen majority”, en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 95, no. 12, pp. 6578–6583, Jun. 1998, ISSN: 0027-8424. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/9618454.

[11] N. R. Pace, D. A. Stahl, D. J. Lane, and G. J. Olsen, “The analysis of natural microbial populations by ribosomal RNA sequences”, in *Advances in Microbial Ecology*, K. C. Marshall, Ed., Boston, MA: Springer US, 1986, pp. 1–55, ISBN: 9781475706116. DOI: 10.1007/978-1-4757-0611-6_1. [Online]. Available: https://doi.org/10.1007/978-1-4757-0611-6_1.

[12] C. R. Woese, “Bacterial evolution”, en, *Microbiol. Rev.*, vol. 51, no. 2, pp. 221–271, Jun. 1987, ISSN: 0146-0749. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/2439888.

[13] F. Sanger, S. Nicklen, and A. R. Coulson, “DNA sequencing with chain-terminating inhibitors”, en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 74, no. 12, pp. 5463–5467, Dec. 1977, ISSN: 0027-8424. DOI: 10.1073/pnas.74.12.5463. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/271968.

[14] D. J. Lane, B. Pace, G. J. Olsen, D. A. Stahl, M. L. Sogin, and N. R. Pace, “Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses”, en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 82, no. 20, pp. 6955–6959, Oct. 1985, ISSN: 0027-8424. DOI: 10.1073/pnas.82.20.6955. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/2413450.

[15] G. J. Olsen, N. Larsen, and C. R. Woese, “The ribosomal RNA database project”, en, *Nucleic Acids Res.*, vol. 19 Suppl, pp. 2017–2021, Apr. 1991, ISSN: 0305-1048. DOI: 10.1093/nar/19.suppl.2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/2041798.

[16] T. Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen, “Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB”, en, *Appl. Environ. Microbiol.*, vol. 72, no. 7, pp. 5069–5072, Jul. 2006, ISSN: 0099-2240. DOI: 10.1128/AEM.03006-05. [Online]. Available: http://dx.doi.org/10.1128/AEM.03006-05.

[17] J. Wuyts, G. Perrière, and Y. Van De Peer, “The european ribosomal RNA database”, en, *Nucleic Acids Res.*, vol. 32, no. Database issue, pp. D101–3, Jan. 2004, ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkh065. [Online]. Available: http://dx.doi.org/10.1093/nar/gkh065.

[18] E. Pruesse, C. Quast, K. Knittel, B. M. Fuchs, W. Ludwig, J. Peplies, and F. O. Glöckner, “SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB”, en, *Nucleic Acids Res.*, vol. 35, no. 21, pp. 7188–7196, Oct. 2007, ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkm864. [Online]. Available: http://dx.doi.org/10.1093/nar/gkm864.

[19] *Silva release 132*, https://www.arb-silva.de/documentation/release-132/, Accessed: 2018-8-29, Dec. 2017. [Online]. Available: https://www.arb-silva.de/documentation/release-132/.

[20] R. A. Welch, V. Burland, G. Plunkett 3rd, P. Redford, P. Roesch, D. Rasko, E. L. Buckles, S.-R. Liou, A. Boutin, J. Hackett, D. Stroud, G. F. Mayhew, D. J. Rose, S. Zhou, D. C. Schwartz, N. T. Perna, H. L. T. Mobley, M. S. Donnenberg, and F. R. Blattner, “Extensive mosaic structure revealed by the complete genome sequence of uropathogenic escherichia coli”, en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 26, pp. 17 020–17 024, Dec. 2002, ISSN: 0027-8424. DOI: 10.1073/pnas.252529799. [Online]. Available: http://dx.doi.org/10.1073/pnas.252529799.

[21] J. B. H. Martiny, S. E. Jones, J. T. Lennon, and A. C. Martiny, “Microbiomes in light of traits: A phylogenetic perspective”, en, *Science*, vol. 350, no. 6261, aac9323, Nov. 2015, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aac9323. [Online]. Available: http://dx.doi.org/10.1126/science.aac9323.

[22] S. M. Soucy, J. Huang, and J. P. Gogarten, “Horizontal gene transfer: Building the web of life”, en, *Nat. Rev. Genet.*, vol. 16, no. 8, pp. 472–482, Aug. 2015, ISSN: 1471-0056, 1471-0064. DOI: 10.1038/nrg3962. [Online]. Available: http://dx.doi.org/10.1038/nrg3962.

[23] J. Handelsman, M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman, “Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products”, en, *Chem. Biol.*, vol. 5, no. 10, R245–9, Oct. 1998, ISSN: 1074-5521. DOI: 10.1016/S1074-5521(98)90108-9. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/9818143.

[24] J. L. Stein, T. L. Marsh, K. Y. Wu, H. Shizuya, and E. F. DeLong, “Characterization of uncultivated prokaryotes: Isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon”, en, *J. Bacteriol.*, vol. 178, no. 3,

pp. 591–599, Feb. 1996, ISSN: 0021-9193. DOI: 10.1128/jb.178.3.591-599.1996. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/8550487.

[25] F. G. Healy, R. M. Ray, H. C. Aldrich, A. C. Wilkie, L. O. Ingram, and K. T. Shanmugam, "Direct isolation of functional genes encoding cellulases from the microbial consortia in a thermophilic, anaerobic digester maintained on lignocellulose", en, *Appl. Microbiol. Biotechnol.*, vol. 43, no. 4, pp. 667–674, Aug. 1995, ISSN: 0175-7598. DOI: 10.1007/BF00164771. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/7546604.

[26] J. C. Venter, K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y.-H. Rogers, and H. O. Smith, "Environmental genome shotgun sequencing of the sargasso sea", *Science*, vol. 304, no. 5667, pp. 66–74, Apr. 2004, ISSN: 0036-8075. [Online]. Available: http://www.sciencemag.org/cgi/doi/10.1126/science.1093857.

[27] N. Fierer, J. W. Leff, B. J. Adams, U. N. Nielsen, S. T. Bates, C. L. Lauber, S. Owens, J. A. Gilbert, D. H. Wall, and J. G. Caporaso, "Cross-biome metagenomic analyses of soil microbial communities and their functional attributes", en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 52, pp. 21 390–21 395, Dec. 2012, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1215210110. [Online]. Available: http://dx.doi.org/10.1073/pnas.1215210110.

[28] J. M. Heather and B. Chain, "The sequence of sequencers: The history of sequencing DNA", en, *Genomics*, vol. 107, no. 1, pp. 1–8, Jan. 2016, ISSN: 0888-7543, 1089-8646. DOI: 10.1016/j.ygeno.2015.11.003. [Online]. Available: http://dx.doi.org/10.1016/j.ygeno.2015.11.003.

[29] A. Sczyrba, P. Hofmann, P. Belmann, D. Koslicki, S. Janssen, J. Dröge, I. Gregor, S. Majda, J. Fiedler, E. Dahms, A. Bremges, A. Fritz, R. Garrido-Oter, T. S. Jørgensen, N. Shapiro, P. D. Blood, A. Gurevich, Y. Bai, D. Turaev, M. Z. DeMaere, R. Chikhi, N. Nagarajan, C. Quince, F. Meyer, M. Balvočiūtė, L. H. Hansen, S. J. Sørensen, B. K. H. Chia, B. Denis, J. L. Froula, Z. Wang, R. Egan, D. Don Kang, J. J. Cook, C. Deltel, M. Beckstette, C. Lemaitre, P. Peterlongo, G. Rizk, D. Lavenier, Y.-W. Wu, S. W. Singer, C. Jain, M. Strous, H. Klingenberg, P. Meinicke, M. D. Barton, T. Lingner, H.-H. Lin, Y.-C. Liao, G. G. Z. Silva, D. A. Cuevas, R. A. Edwards, S. Saha, V. C. Piro, B. Y. Renard, M. Pop, H.-P. Klenk, M. Göker, N. C. Kyrpides, T. Woyke, J. A. Vorholt, P. Schulze-Lefert, E. M. Rubin, A. E. Darling, T. Rattei, and A. C. McHardy, "Critical assessment of metagenome interpretation-a benchmark of metagenomics software",

en, *Nat. Methods*, vol. 14, no. 11, pp. 1063–1071, Nov. 2017, ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.4458. [Online]. Available: http://dx.doi.org/10.1038/nmeth.4458.

[30] N. J. Loman, J. Quick, and J. T. Simpson, “A complete bacterial genome assembled de novo using only nanopore sequencing data”, en, *Nat. Methods*, vol. 12, no. 8, pp. 733–735, Aug. 2015, ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.3444. [Online]. Available: http://dx.doi.org/10.1038/nmeth.3444.

[31] R. R. Wick, L. M. Judd, C. L. Gorrie, and K. E. Holt, “Completing bacterial genome assemblies with multiplex MinION sequencing”, en, *Microb Genom*, vol. 3, no. 10, e000132, Oct. 2017, ISSN: 2057-5858. DOI: 10.1099/mgen.0.000132. [Online]. Available: http://dx.doi.org/10.1099/mgen.0.000132.

[32] S. Koren, B. P. Walenz, K. Berlin, J. R. Miller, N. H. Bergman, and A. M. Phillippy, “Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation”, en, *Genome Res.*, vol. 27, no. 5, pp. 722–736, May 2017, ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.215087.116. [Online]. Available: http://dx.doi.org/10.1101/gr.215087.116.

[33] J. A. Frank, Y. Pan, A. Tooming-Klunderud, V. G. H. Eijsink, A. C. McHardy, A. J. Nederbragt, and P. B. Pope, “Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data”, en, *Sci. Rep.*, vol. 6, p. 25 373, May 2016, ISSN: 2045-2322. DOI: 10.1038/srep25373. [Online]. Available: http://dx.doi.org/10.1038/srep25373.

[34] K. J. Locey and J. T. Lennon, “Scaling laws predict global microbial diversity”, en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 113, no. 21, pp. 5970–5975, May 2016, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1521291113. [Online]. Available: http://dx.doi.org/10.1073/pnas.1521291113.

[35] R. M. Bowers, N. C. Kyrpides, R. Stepanauskas, M. Harmon-Smith, D. Doud, T. B. K. Reddy, F. Schulz, J. Jarett, A. R. Rivers, E. A. Eloe-Fadrosh, S. G. Tringe, N. N. Ivanova, A. Copeland, A. Clum, E. D. Becraft, R. R. Malmstrom, B. Birren, M. Podar, P. Bork, G. M. Weinstock, G. M. Garrity, J. A. Dodsworth, S. Yooseph, G. Sutton, F. O. Glöckner, J. A. Gilbert, W. C. Nelson, S. J. Hallam, S. P. Jungbluth, T. J. G. Ettema, S. Tighe, K. T. Konstantinidis, W.-T. Liu, B. J. Baker, T. Rattei, J. A. Eisen, B. Hedlund, K. D. McMahon, N. Fierer, R. Knight, R. Finn, G. Cochrane, I. Karsch-Mizrachi, G. W. Tyson, C. Rinke, Genome Standards Consortium, A. Lapidus, F. Meyer, P. Yilmaz, D. H. Parks, A. M. Eren, L. Schriml, J. F. Banfield, P. Hugenholtz, and T. Woyke, “Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of

bacteria and archaea”, en, *Nat. Biotechnol.*, vol. 35, no. 8, pp. 725–731, Aug. 2017, ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.3893. [Online]. Available: http://dx.doi.org/10.1038/nbt.3893.

[36] M. Strous, B. Kraft, R. Bisdorf, and H. E. Tegetmeyer, “The binning of metagenomic contigs for microbial physiology of mixed cultures”, en, *Front. Microbiol.*, vol. 3, p. 410, Dec. 2012, ISSN: 1664-302X. DOI: 10.3389/fmicb.2012.00410. [Online]. Available: http://dx.doi.org/10.3389/fmicb.2012.00410.

[37] M. V. Brown, F. M. Lauro, M. Z. DeMaere, L. Muir, D. Wilkins, T. Thomas, M. J. Riddle, J. A. Fuhrman, C. Andrews-Pfannkoch, J. M. Hoffman, J. B. McQuaid, A. Allen, S. R. Rintoul, and R. Cavicchioli, “Global biogeography of SAR11 marine bacteria”, en, *Mol. Syst. Biol.*, vol. 8, p. 595, Jul. 2012, ISSN: 1744-4292. DOI: 10.1038/msb.2012.28. [Online]. Available: http://dx.doi.org/10.1038/msb.2012.28.

[38] Y.-W. Wu, Y.-H. Tang, S. G. Tringe, B. A. Simmons, and S. W. Singer, “MaxBin: An automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm”, en, *Microbiome*, vol. 2, p. 26, Aug. 2014, ISSN: 2049-2618. DOI: 10.1186/2049-2618-2-26. [Online]. Available: http://dx.doi.org/10.1186/2049-2618-2-26.

[39] K. A. Wetterstrand, *DNA sequencing costs: Data from the NHGRI genome sequencing program (GSP)*, https://www.genome.gov/sequencingcostsdata/, Accessed: 2018-8-28. [Online]. Available: https://www.genome.gov/sequencingcostsdata/.

[40] M. Imelfort, D. Parks, B. J. Woodcroft, P. Dennis, P. Hugenholtz, and G. W. Tyson, “GroopM: An automated tool for the recovery of population genomes from related metagenomes”, en, *PeerJ*, vol. 2, e603, Sep. 2014, ISSN: 2167-8359. DOI: 10.7717/peerj.603. [Online]. Available: http://dx.doi.org/10.7717/peerj.603.

[41] Y.-W. Wu, B. A. Simmons, and S. W. Singer, “MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets”, en, *Bioinformatics*, vol. 32, no. 4, pp. 605–607, Feb. 2016, ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btv638. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btv638.

[42] J. Alneberg, B. S. Bjarnason, I. de Bruijn, M. Schirmer, J. Quick, U. Z. Ijaz, N. J. Loman, A. F. Andersson, and C. Quince, “CONCOCT: Clustering cONtigs on COverage and ComposiTion”, Dec. 2013. arXiv: 1312.4038 [q-bio.GN]. [Online]. Available: http://arxiv.org/abs/1312.4038.

[43] Y. Y. Lu, T. Chen, J. A. Fuhrman, and F. Sun, “COCACOLA: Binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end

read LinkAge", en, *Bioinformatics*, vol. 33, no. 6, pp. 791–798, Mar. 2017, ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btw290. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btw290.

[44] B. Cleary, I. L. Brito, K. Huang, D. Gevers, T. Shea, S. Young, and E. J. Alm, "Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning", en, *Nat. Biotechnol.*, vol. 33, no. 10, pp. 1053–1060, Oct. 2015, ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.3329. [Online]. Available: http://dx.doi.org/10.1038/nbt.3329.

[45] D. D. Kang, J. Froula, R. Egan, and Z. Wang, "MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities", en, *PeerJ*, vol. 3, e1165, Aug. 2015, ISSN: 2167-8359. DOI: 10.7717/peerj.1165. [Online]. Available: http://dx.doi.org/10.7717/peerj.1165.

[46] D. H. Parks, C. Rinke, M. Chuvochina, P.-A. Chaumeil, B. J. Woodcroft, P. N. Evans, P. Hugenholtz, and G. W. Tyson, "Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life", en, *Nat Microbiol*, vol. 2, no. 11, pp. 1533–1542, Nov. 2017, ISSN: 2058-5276. DOI: 10.1038/s41564-017-0012-7. [Online]. Available: http://dx.doi.org/10.1038/s41564-017-0012-7.

[47] B. J. Tully, E. D. Graham, and J. F. Heidelberg, "The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans", en, *Sci Data*, vol. 5, p. 170 203, Jan. 2018, ISSN: 2052-4463. DOI: 10.1038/sdata.2017.203. [Online]. Available: http://dx.doi.org/10.1038/sdata.2017.203.

[48] B. J. Tully, R. Sachdeva, E. D. Graham, and J. F. Heidelberg, "290 metagenome-assembled genomes from the mediterranean sea: A resource for marine microbiology", en, *PeerJ*, vol. 5, e3558, Jul. 2017, ISSN: 2167-8359. DOI: 10.7717/peerj.3558. [Online]. Available: http://dx.doi.org/10.7717/peerj.3558.

[49] F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, "BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs", en, *Bioinformatics*, vol. 31, no. 19, pp. 3210–3212, Oct. 2015, ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btv351. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btv351.

[50] D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson, "CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes", en, *Genome Res.*, vol. 25, no. 7, pp. 1043–1055, Jul. 2015, ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.186072.114. [Online]. Available: http://dx.doi.org/10.1101/gr.186072.114.

[51] A. M. Eren, Ö. C. Esen, C. Quince, J. H. Vineis, H. G. Morrison, M. L. Sogin, and T. O. Delmont, “Anvi’o: An advanced analysis and visualization platform for ’omics data”, *PeerJ*, vol. 3, no. 358, e1319, Jan. 2015. [Online]. Available: https://peerj.com/articles/1319.

[52] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner, “Capturing chromosome conformation”, en, *Science*, vol. 295, no. 5558, pp. 1306–1311, Feb. 2002, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1067799. [Online]. Available: http://dx.doi.org/10.1126/science.1067799.

[53] M. Simonis, P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, B. van Steensel, and W. de Laat, “Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4c)”, en, *Nat. Genet.*, vol. 38, no. 11, pp. 1348–1354, Nov. 2006, ISSN: 1061-4036. DOI: 10.1038/ng1896. [Online]. Available: http://dx.doi.org/10.1038/ng1896.

[54] J. Dostie, T. A. Richmond, R. A. Arnaout, R. R. Selzer, W. L. Lee, T. A. Honan, E. D. Rubio, A. Krumm, J. Lamb, C. Nusbaum, R. D. Green, and J. Dekker, “Chromosome conformation capture carbon copy (5c): A massively parallel solution for mapping interactions between genomic elements”, en, *Genome Res.*, vol. 16, no. 10, pp. 1299–1309, Oct. 2006, ISSN: 1088-9051. DOI: 10.1101/gr.5571506. [Online]. Available: http://dx.doi.org/10.1101/gr.5571506.

[55] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker, “Comprehensive mapping of long-range interactions reveals folding principles of the human genome”, en, *Science*, vol. 326, no. 5950, pp. 289–293, Oct. 2009, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1181369. [Online]. Available: http://dx.doi.org/10.1126/science.1181369.

[56] E. Splinter, E. de Wit, E. P. Nora, P. Klous, H. J. G. van de Werken, Y. Zhu, L. J. T. Kaaij, W. van Ijcken, J. Gribnau, E. Heard, and W. de Laat, “The inactive X chromosome adopts a unique three-dimensional conformation that is dependent on xist RNA”, en, *Genes Dev.*, vol. 25, no. 13, pp. 1371–1383, Jul. 2011, ISSN: 0890-9369, 1549-5477. DOI: 10.1101/gad.633311. [Online]. Available: http://dx.doi.org/10.1101/gad.633311.

[57] M. A. Ferraiuolo, A. Sanyal, N. Naumova, J. Dekker, and J. Dostie, “From cells to chromatin: Capturing snapshots of genome organization with 5C technology”, en, *Methods*, vol. 58, no. 3, pp. 255–267, Nov. 2012, ISSN: 1046-2023, 1095-9130.

DOI: 10 . 1016 / j . ymeth . 2012 . 10 . 011. [Online]. Available: http://dx.doi.org/10.1016/j.ymeth.2012.10.011.

[58] B. D. Pope, T. Ryba, V. Dileep, F. Yue, W. Wu, O. Denas, D. L. Vera, Y. Wang, R. S. Hansen, T. K. Canfield, R. E. Thurman, Y. Cheng, G. Gülsoy, J. H. Dennis, M. P. Snyder, J. A. Stamatoyannopoulos, J. Taylor, R. C. Hardison, T. Kahveci, B. Ren, and D. M. Gilbert, "Topologically associating domains are stable units of replication-timing regulation", en, *Nature*, vol. 515, no. 7527, pp. 402–405, Nov. 2014, ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature13986. [Online]. Available: http://dx.doi.org/10.1038/nature13986.

[59] P. Edge, V. Bafna, and V. Bansal, "HapCUT2: Robust and accurate haplotype assembly for diverse sequencing technologies", en, *Genome Res.*, vol. 27, no. 5, pp. 801–812, May 2017, ISSN: 1088-9051, 1549-5469. DOI: 10 . 1101 / gr . 213462 . 116. [Online]. Available: http://dx.doi.org/10.1101/gr.213462.116.

[60] H. Marie-Nelly, M. Marbouty, A. Cournac, J.-F. Flot, G. Liti, D. P. Parodi, S. Syan, N. Guillén, A. Margeot, C. Zimmer, and R. Koszul, "High-quality genome (re)assembly using chromosomal contact data", en, *Nat. Commun.*, vol. 5, no. 5695, p. 5695, Dec. 2014, ISSN: 2041-1723. DOI: 10.1038/ncomms6695. [Online]. Available: http://dx.doi.org/10.1038/ncomms6695.

[61] J. N. Burton, A. Adey, R. P. Patwardhan, R. Qiu, J. O. Kitzman, and J. Shendure, "Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions", en, *Nat. Biotechnol.*, vol. 31, no. 12, pp. 1119–1125, Dec. 2013, ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.2727. [Online]. Available: http://dx.doi.org/10.1038/nbt.2727.

[62] N. Varoquaux, I. Liachko, F. Ay, J. N. Burton, J. Shendure, M. J. Dunham, J.-P. Vert, and W. S. Noble, "Accurate identification of centromere locations in yeast genomes using Hi-C", en, *Nucleic Acids Res.*, vol. 43, no. 11, pp. 5331–5339, Jun. 2015, ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkv424. [Online]. Available: http://dx.doi.org/10.1093/nar/gkv424.

[63] M. Marbouty, L. Baudry, A. Cournac, and R. Koszul, "Scaffolding bacterial genomes and probing host-virus interactions in gut microbiome by proximity ligation (chromosome capture) assay", en, *Sci Adv*, vol. 3, no. 2, e1602105, Feb. 2017, ISSN: 2375-2548. DOI: 10 . 1126 / sciadv . 1602105. [Online]. Available: http://dx.doi.org/10.1126/sciadv.1602105.

[64] W. W. Greenwald, H. Li, P. Benaglio, D. Jakubosky, H. Matsui, A. Schmitt, S. Selvaraj, M. D'Antonio, A. D'Antonio-Chrownowska, E. N. Smith, and

K. A. Frazer, “Integration of phased Hi-C and molecular phenotype data to study genetic and epigenetic effects on chromatin looping”, en, Jun. 2018, [Online]. Available: https://www.biorxiv.org/content/early/2018/06/20/352682.

[65] C. W. Beitel, L. Froenicke, J. M. Lang, I. F. Korf, R. W. Michelmore, J. A. Eisen, and A. E. Darling, “Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products”, en, *PeerJ*, vol. 2, no. 12, e415, May 2014, ISSN: 2167-8359. DOI: 10.7717/peerj.415. [Online]. Available: http://dx.doi.org/10.7717/peerj.415.

[66] J. N. Burton, I. Liachko, M. J. Dunham, and J. Shendure, “Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps”, en, *G3*, vol. 4, no. 7, pp. 1339–1346, May 2014, ISSN: 2160-1836. DOI: 10.1534/g3.114.011825. [Online]. Available: http://dx.doi.org/10.1534/g3.114.011825.

[67] M. Marbouty, A. Cournac, J.-F. Flot, H. Marie-Nelly, J. Mozziconacci, and R. Koszul, “Metagenomic chromosome conformation capture (meta3c) unveils the diversity of chromosome organization in microorganisms”, en, *Elife*, vol. 3, no. e03318, e03318, Dec. 2014, ISSN: 2050-084X. DOI: 10.7554/eLife.03318. [Online]. Available: http://dx.doi.org/10.7554/eLife.03318.

[68] S. Dongen, “A cluster algorithm for graphs”, Amsterdam, The Netherlands, The Netherlands, Tech. Rep., 2000. [Online]. Available: https://dl.acm.org/citation.cfm?id=868986.

[69] C. S. Heil, J. N. Burton, I. Liachko, A. Friedrich, N. A. Hanson, C. L. Morris, J. Schacherer, J. Shendure, J. H. Thomas, and M. J. Dunham, “Identification of a novel interspecific hybrid yeast from a metagenomic open fermentation sample using Hi-C”, en, Jun. 2017, [Online]. Available: http://biorxiv.org/content/early/2017/06/15/150722.

[70] M. O. Press, A. H. Wiser, Z. N. Kronenberg, K. W. Langford, M. Shakya, C.-C. Lo, K. A. Mueller, S. T. Sullivan, P. S. G. Chain, and I. Liachko, “Hi-C deconvolution of a human gut microbiome yields high-quality draft genomes and reveals plasmid-genome interactions”, en, Oct. 2017, [Online]. Available: https://www.biorxiv.org/content/early/2017/10/05/198713.

[71] M. Munafò, “Open science and research reproducibility”, en, *Ecancermedicalscience*, vol. 10, ed56, Jun. 2016, ISSN: 1754-6605. DOI: 10.3332/ecancer.2016.ed56. [Online]. Available: http://dx.doi.org/10.3332/ecancer.2016.ed56.

[72] M. E. J. Newman, “Detecting community structure in networks”, *Eur. Phys. J. B*, vol. 38, no. 2, pp. 321–330, Mar. 2004, ISSN: 1434-6028. DOI:

10.1140/epjb/e2004-00124-y. [Online]. Available: https://doi.org/10.1140/epjb/e2004-00124-y.

[73] M. Z. DeMaere and A. E. Darling, “bin3C: Exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes”, *Genome Biology*, vol. 20, no. 1, p. 46, Feb. 2019, ISSN: 1465-6906. DOI: 10.1186/s13059-019-1643-1. [Online]. Available: https://doi.org/10.1186/s13059-019-1643-1.

[74] E. Yaffe and A. Tanay, “Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture”, en, *Nat. Genet.*, vol. 43, no. 11, pp. 1059–1065, Oct. 2011, ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng.947. [Online]. Available: http://dx.doi.org/10.1038/ng.947.

[75] M. Imakaev, G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker, and L. A. Mirny, “Iterative correction of Hi-C data reveals hallmarks of chromosome organization”, en, *Nat. Methods*, vol. 9, no. 10, pp. 999–1003, Oct. 2012, ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.2148. [Online]. Available: http://dx.doi.org/10.1038/nmeth.2148.

[76] M. Hu, K. Deng, S. Selvaraj, Z. Qin, B. Ren, and J. S. Liu, “HiCNorm: Removing biases in Hi-C data via poisson regression”, en, *Bioinformatics*, vol. 28, no. 23, pp. 3131–3133, Dec. 2012, ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/bts570. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/bts570.

[77] W. Li, K. Gong, Q. Li, F. Alber, and X. J. Zhou, “Hi-Corrector: A fast, scalable and memory-efficient package for normalizing large-scale Hi-C data”, en, *Bioinformatics*, vol. 31, no. 6, pp. 960–962, Mar. 2015, ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btu747. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btu747.

[78] *Iterative [proportional fitting of a log-linear model for] correction of Hi-C data reveals hallmarks of chromosome organization*, https://liorpachter.wordpress.com/2013/11/17/imakaev_explained/, Accessed: 2018-9-4, Nov. 2013. [Online]. Available: https://liorpachter.wordpress.com/2013/11/17/imakaev_explained/.

[79] R. Sinkhorn and P. Knopp, “Concerning nonnegative matrices and doubly stochastic matrices”, *Pacific J. Math.*, vol. 21, no. 2, pp. 343–348, May 1967, ISSN: 0030-8730. [Online]. Available: https://msp.org/pjm/1967/21-2/p14.xhtml.

[80] P. Knight, “The Sinkhorn–Knopp algorithm: Convergence and applications”, *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 1, pp. 261–275, Jan. 2008, ISSN: 0895-4798. DOI: 10.1137/060659624. [Online]. Available: https://doi.org/10.1137/060659624.

[81] P. A. Knight and D. Ruiz, “A fast algorithm for matrix balancing”, *IMA J. Numer. Anal.*, vol. 33, no. 3, pp. 1029–1047, Jul. 2013, ISSN: 0272-4979. DOI: 10.1093/imanum/drs019. [Online]. Available: https://academic.oup.com/imajna/article-abstract/33/3/1029/659457.

[82] N. C. Durand, M. S. Shamim, I. Machol, S. S. P. Rao, M. H. Huntley, E. S. Lander, and E. L. Aiden, “Juicer provides a One-Click system for analyzing Loop-Resolution Hi-C experiments”, en, *Cell Syst*, vol. 3, no. 1, pp. 95–98, Jul. 2016, ISSN: 2405-4712. DOI: 10.1016/j.cels.2016.07.002. [Online]. Available: http://dx.doi.org/10.1016/j.cels.2016.07.002.

[83] S. Fortunato, “Community detection in graphs”, *Phys. Rep.*, vol. 486, no. 3, pp. 75–174, Feb. 2010, ISSN: 0370-1573. DOI: 10.1016/j.physrep.2009.11.002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0370157309002841.

[84] Y.-K. Shih and S. Parthasarathy, “Identifying functional modules in interaction networks through overlapping markov clustering”, en, *Bioinformatics*, vol. 28, no. 18, pp. i473–i479, Sep. 2012, ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/bts370. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/bts370.

[85] M. Rosvall and C. T. Bergstrom, “Maps of random walks on complex networks reveal community structure”, en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 4, pp. 1118–1123, Jan. 2008, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0706851105. [Online]. Available: http://dx.doi.org/10.1073/pnas.0706851105.

[86] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks”, en, *J. Stat. Mech.*, vol. 2008, no. 10, P10008, Oct. 2008, ISSN: 1742-5468. DOI: 10.1088/1742-5468/2008/10/P10008. [Online]. Available: http://iopscience.iop.org/article/10.1088/1742-5468/2008/10/P10008/meta.

[87] L. Waltman and N. J. van Eck, “A smart local moving algorithm for large-scale modularity-based community detection”, *Eur. Phys. J. B*, vol. 86, no. 11, p. 471, Nov. 2013, ISSN: 1434-6028, 1434-6036. DOI: 10.1140/epjb/e2013-40829-0. [Online]. Available: https://doi.org/10.1140/epjb/e2013-40829-0.

[88] P. Esmailian and M. Jalili, “Community detection in signed networks: The role of negative ties in different scales”, en, *Sci. Rep.*, vol. 5, p. 14 339, Sep. 2015, ISSN: 2045-2322. DOI: 10.1038/srep14339. [Online]. Available: http://dx.doi.org/10.1038/srep14339.

[89] J. Xie, B. K. Szymanski, and X. Liu, “SLPA: Uncovering overlapping communities in social networks via a Speaker-Listener interaction dynamic process”, in *2011 IEEE 11th International Conference on Data Mining Workshops*, Dec. 2011, pp. 344–349. DOI: 10.1109/ICDMW.2011.154. [Online]. Available: http://dx.doi.org/10.1109/ICDMW.2011.154.

[90] J. M. Hofman and C. H. Wiggins, “Bayesian approach to network modularity”, en, *Phys. Rev. Lett.*, vol. 100, no. 25, p. 258 701, Jun. 2008, ISSN: 0031-9007. DOI: 10.1103/PhysRevLett.100.258701. [Online]. Available: http://dx.doi.org/10.1103/PhysRevLett.100.258701.

[91] M. C. P. de Souto, A. L. V. Coelho, K. Faceli, T. C. Sakata, V. Bonadia, and I. G. Costa, “A comparison of external clustering evaluation indices in the context of imbalanced data sets”, in *2012 Brazilian Symposium on Neural Networks*, Oct. 2012, pp. 49–54. DOI: 10.1109/SBRN.2012.25. [Online]. Available: http://dx.doi.org/10.1109/SBRN.2012.25.

[92] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo, “A comparison of extrinsic clustering evaluation metrics based on formal constraints”, *Inf. Retr. Boston.*, vol. 12, no. 4, pp. 461–486, Aug. 2009, ISSN: 1386-4564, 1573-7659. DOI: 10.1007/s10791-008-9066-8. [Online]. Available: https://doi.org/10.1007/s10791-008-9066-8.

[93] S. Emmons, S. Kobourov, M. Gallant, and K. Börner, “Analysis of network clustering algorithms and cluster quality metrics at scale”, en, *PLoS One*, vol. 11, no. 7, e0159161, Jul. 2016, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0159161. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0159161.

[94] A. Lancichinetti and S. Fortunato, “Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities”, en, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, vol. 80, no. 1 Pt 2, p. 016 118, Jul. 2009, ISSN: 1539-3755. DOI: 10.1103/PhysRevE.80.016118. [Online]. Available: http://dx.doi.org/10.1103/PhysRevE.80.016118.

[95] A. Lancichinetti, S. Fortunato, and F. Radicchi, “Benchmark graphs for testing community detection algorithms”, en, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, vol. 78, no. 4 Pt 2, p. 046 110, Oct. 2008, ISSN: 1539-3755. DOI: 10.1103/PhysRevE.78.046110. [Online]. Available: http://dx.doi.org/10.1103/PhysRevE.78.046110.

[96] H. Rosales-Méndez and Y. Ramírez-Cruz, “Addressing the validation of overlapping clustering algorithms”, *Intelligent Data Analysis*, vol. 18, no. 6S,

S33–S45, 2014. DOI: 10.3233/IDA-140707. [Online]. Available: http://dx.doi.org/10.3233/IDA-140707.

[97] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance", *J. Mach. Learn. Res.*, vol. 11, no. Oct, pp. 2837–2854, 2010, ISSN: 1532-4435, 1533-7928. [Online]. Available: http://www.jmlr.org/papers/volume11/vinh10a/vinh10a.pdf.

[98] S. Romano, J. Bailey, V. Nguyen, and K. Verspoor, "Standardized mutual information for clustering comparisons: One step further in adjustment for chance", en, in *International Conference on Machine Learning*, jmlr.org, Jan. 2014, pp. 1143–1151. [Online]. Available: http://proceedings.mlr.press/v32/romano14.html.

[99] L. Hubert and P. Arabie, "Comparing partitions", *J. Classification*, vol. 2, no. 1, pp. 193–218, Dec. 1985, ISSN: 0176-4268, 1432-1343. DOI: 10.1007/BF01908075. [Online]. Available: https://doi.org/10.1007/BF01908075.

[100] M. E. J. Newman, "Modularity and community structure in networks", en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 23, pp. 8577–8582, Jun. 2006, ISSN: 0027-8424. DOI: 10.1073/pnas.0601602103. [Online]. Available: http://dx.doi.org/10.1073/pnas.0601602103.

[101] S. Fortunato and M. Barthélemy, "Resolution limit in community detection", en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 104, no. 1, pp. 36–41, Jan. 2007, ISSN: 0027-8424. DOI: 10.1073/pnas.0605965104. [Online]. Available: http://dx.doi.org/10.1073/pnas.0605965104.

[102] B. H. Good, Y.-A. de Montjoye, and A. Clauset, "Performance of modularity maximization in practical contexts", en, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, vol. 81, no. 4 Pt 2, p. 046 106, Apr. 2010, ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.81.046106. [Online]. Available: http://dx.doi.org/10.1103/PhysRevE.81.046106.

[103] T. Kawamoto and M. Rosvall, "Estimating the resolution limit of the map equation in community detection", en, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, vol. 91, no. 1, p. 012 809, Jan. 2015, ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.91.012809. [Online]. Available: http://dx.doi.org/10.1103/PhysRevE.91.012809.

[104] M. Rosvall, D. Axelsson, and C. T. Bergstrom, "The map equation", Jun. 2009. arXiv: 0906.1405 [physics.soc-ph]. [Online]. Available: http://arxiv.org/abs/0906.1405.

[105] M. L. Sogin, H. G. Morrison, J. A. Huber, D. Mark Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, and G. J. Herndl, "Microbial diversity in the deep sea and the underexplored "rare biosphere"", en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 32, pp. 12 115–12 120, Aug. 2006, ISSN: 0027-8424. DOI: `10.1073/pnas.0605127103`. [Online]. Available: `http://dx.doi.org/10.1073/pnas.0605127103`.

[106] E. Garrison, M. Lin, and R. Williams, *Vg*. [Online]. Available: `https://github.com/vgteam/vg`.

[107] L. G. S. Jeub, O. Sporns, and S. Fortunato, "Multiresolution consensus clustering in networks", en, *Sci. Rep.*, vol. 8, no. 1, p. 3259, Feb. 2018, ISSN: 2045-2322. DOI: `10.1038/s41598-018-21352-7`. [Online]. Available: `http://dx.doi.org/10.1038/s41598-018-21352-7`.

[108] M. De Domenico, A. Lancichinetti, A. Arenas, and M. Rosvall, "Identifying modular flows on multilayer networks reveals highly overlapping organization in interconnected systems", *Phys. Rev. X*, vol. 5, no. 1, p. 011 027, Mar. 2015. DOI: `10.1103/PhysRevX.5.011027`. [Online]. Available: `https://link.aps.org/doi/10.1103/PhysRevX.5.011027`.

[109] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg, "Genome sequencing in microfabricated high-density picolitre reactors", en, *Nature*, vol. 437, no. 7057, pp. 376–380, Sep. 2005, ISSN: 0028-0836, 1476-4687. DOI: `10.1038/nature03959`. [Online]. Available: `http://dx.doi.org/10.1038/nature03959`.

[110] S. Bennett, "Solexa ltd", en, *Pharmacogenomics*, vol. 5, no. 4, pp. 433–438, Jun. 2004, ISSN: 1462-2416. DOI: `10.1517/14622416.5.4.433`. [Online]. Available: `http://dx.doi.org/10.1517/14622416.5.4.433`.

[111] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham,

M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner, “Real-time DNA sequencing from single polymerase molecules”, en, *Science*, vol. 323, no. 5910, pp. 133–138, Jan. 2009, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1162986. [Online]. Available: http://dx.doi.org/10.1126/science.1162986.

[112] J. M. Rothberg, W. Hinz, T. M. Rearick, J. Schultz, W. Mileski, M. Davey, J. H. Leamon, K. Johnson, M. J. Milgrew, M. Edwards, J. Hoon, J. F. Simons, D. Marran, J. W. Myers, J. F. Davidson, A. Branting, J. R. Nobile, B. P. Puc, D. Light, T. A. Clark, M. Huber, J. T. Branciforte, I. B. Stoner, S. E. Cawley, M. Lyons, Y. Fu, N. Homer, M. Sedova, X. Miao, B. Reed, J. Sabina, E. Feierstein, M. Schorn, M. Alanjary, E. Dimalanta, D. Dressman, R. Kasinskas, T. Sokolsky, J. A. Fidanza, E. Namsaraev, K. J. McKernan, A. Williams, G. T. Roth, and J. Bustillo, “An integrated semiconductor device enabling non-optical genome sequencing”, en, *Nature*, vol. 475, no. 7356, pp. 348–352, Jul. 2011, ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature10242. [Online]. Available: http://dx.doi.org/10.1038/nature10242.

[113] M. Jain, H. E. Olsen, B. Paten, and M. Akeson, “The oxford nanopore MinION: Delivery of nanopore sequencing to the genomics community”, en, *Genome Biol.*, vol. 17, no. 1, p. 239, Nov. 2016, ISSN: 1465-6906. DOI: 10.1186/s13059-016-1103-0. [Online]. Available: http://dx.doi.org/10.1186/s13059-016-1103-0.

[114] E. W. Myers, G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. Reinert, K. A. Remington, E. L. Anson, R. A. Bolanos, H. H. Chou, C. M. Jordan, A. L. Halpern, S. Lonardi, E. M. Beasley, R. C. Brandon, L. Chen, P. J. Dunn, Z. Lai, Y. Liang, D. R. Nusskern, M. Zhan, Q. Zhang, X. Zheng, G. M. Rubin, M. D. Adams, and J. C. Venter, “A whole-genome assembly of drosophila”, en, *Science*, vol. 287, no. 5461, pp. 2196–2204, Mar. 2000, ISSN: 0036-8075. DOI: 10.1126/science.287.5461.2196. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/10731133.

[115] S. Batzoglou, D. B. Jaffe, K. Stanley, J. Butler, S. Gnerre, E. Mauceli, B. Berger, J. P. Mesirov, and E. S. Lander, “ARACHNE: A whole-genome shotgun assembler”, en, *Genome Res.*, vol. 12, no. 1, pp. 177–189, Jan. 2002, ISSN: 1088-9051. DOI: 10.1101/gr.208902. [Online]. Available: http://dx.doi.org/10.1101/gr.208902.

[116] J. R. Miller, S. Koren, and G. Sutton, “Assembly algorithms for next-generation sequencing data”, en, *Genomics*, vol. 95, no. 6, pp. 315–327, Jun. 2010, ISSN:

0888-7543, 1089-8646. DOI: 10.1016/j.ygeno.2010.03.001. [Online]. Available: http://dx.doi.org/10.1016/j.ygeno.2010.03.001.

[117] S. Nurk, A. Bankevich, D. Antipov, A. Gurevich, A. Korobeynikov, A. Lapidus, A. Prjibelsky, A. Pyshkin, A. Sirotkin, Y. Sirotkin, R. Stepanauskas, J. McLean, R. Lasken, S. R. Clingenpeel, T. Woyke, G. Tesler, M. A. Alekseyev, and P. A. Pevzner, "Assembling genomes and mini-metagenomes from highly chimeric reads", en, in *Research in Computational Molecular Biology*, ser. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, Apr. 2013, pp. 158–170, ISBN: 9783642371943. DOI: 10.1007/978-3-642-37195-0_13. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-37195-0_13.

[118] Y. Peng, H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin, "IDBA – a practical iterative de bruijn graph de novo assembler", in *Lecture Notes in Computer Science*, 2010, pp. 426–440. DOI: 10.1007/978-3-642-12683-3_28. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-12683-3_28.

[119] D. Coil, G. Jospin, and A. E. Darling, "A5-miseq: An updated pipeline to assemble microbial genomes from illumina MiSeq data", en, *Bioinformatics*, vol. 31, no. 4, pp. 587–589, Feb. 2015, ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btu661. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btu661.

[120] D. R. Zerbino and E. Birney, "Velvet: Algorithms for de novo short read assembly using de bruijn graphs", en, *Genome Res.*, vol. 18, no. 5, pp. 821–829, May 2008, ISSN: 1088-9051. DOI: 10.1101/gr.074492.107. [Online]. Available: http://dx.doi.org/10.1101/gr.074492.107.

[121] Y. Peng, H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin, "IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth", en, *Bioinformatics*, vol. 28, no. 11, pp. 1420–1428, Jun. 2012, ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/bts174. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/bts174.

1.7 Appendices

1.7.1 Definitions

Read

A read is a series of sequential base-pairs inferred – read, in a sense – from a fragment of DNA. This base-pair sequence is determined using one of a number of different DNA sequencing technologies [13], [109]–[113]. The choice of technology establishes both the expected accuracy of the predicted base-pairs and the overall length of each read.

Read-Set

A read-set is a collection of reads produced in a single sequencing experiment, prepared from a single genomic or metagenomic sample. The number of reads in a single read-set has increased dramatically since the advent of DNA sequencing. Present day high-throughput systems can produce read-sets on the order of 10^9 reads. In the common form of shotgun sequencing, the reads within a read-set are a random sampling over the source genome(s).

DNA Sequencing Technologies

Many different sequencing technologies have been devised since the dawn of the DNA sequencing era [13]. Technologies in use today can be partitioned effectively by the form of DNA sample used as input and by the characteristics of the resulting output sequencing information.

Sequencing technologies can be split into two classes by the intent of sample preparation method. These two intents being preparations that produce the efficient sequencing of bulk DNA and, alternatively, preparations whose precision is at the level of single-cells.

Bulk DNA systems begin from a sample of DNA ($1 - 5$ pg) which has been extracted from a large number cells ($\approx 10^6$ bacterial cells). These cells can be obtained from a growth culture, an environment, or tissue source. Contrastingly, and as the name implies, single-cell systems analyze the DNA of ideally just a single cell, however in practice might be somewhat more. The ultra-low input of single-cell systems lend these technologies to experimental situations where sample size is either extremely limited or desired (such as in studies of structural rearrangement) but are significantly hampered by high variation in depth of coverage.

In terms of output, sequencing technologies again fall into groups: short-read systems and long-read systems. Short-read systems ($50-500$ bp) have longer history than that of long-read systems ($5000-2,000,000$ bp), which, taken along with their traditionally higher throughput, represent the gross majority of publicly accessible DNA sequencing data. Present state of the art short-read systems can generate as many as 20×10^9 150 bp reads or 6×10^{12} bp per run, while current leading long-read systems are capable of yields on the order of 250×10^9 bp per run with varying but longer average read lengths. The accuracy of short- and long-read systems also differ, with short-read systems being much more accurate (short > 99%, long < 90%).

Depth of Coverage

Depth of coverage, or read-depth, refers to the number times a given genomic position was observed in a sequencing experiment – the number of generated reads which cover a given position. High read-depth is desirable as it represents increased experimental evidence. Consistent depth of coverage is an indication that the sequencing experiment did not suffer unduly from systematic errors or biases, while also suggesting such regions can be regarded as having been less difficult to reconstruct (and thereby more reliable) within an assembly.

Genome Assembly

A genome assembly is the algorithmic reconstruction of the genome from one or more read-sets produced in a shotgun (random sampling) format. With the exception of small entities such as virus genomes, reconstruction is necessary as the length of reads produced by current DNA sequencing technologies is much shorter than that of the individual chromosomes.

Current genome assembly algorithms fall into two primary categories: overlap-layout-consensus (OLC) [32], [114]–[116] and de Bruijn graph (DGB) methods [117]–[121].

In OLC, reads are first compared to find all pairwise overlaps with which to generate the overlap graph, where reads are nodes and overlaps are edges. Next the layout stage simplifies the overlap graph and merges unambiguous overlapping reads into longer sequences. Lastly, the consensus step walks the simplified graph, at each point choosing the base which is best supported by the overlapping reads. The most computationally expensive stage of OLC is overlap detection with a time complexity of $O(N^2)$ for N reads and $O(L^2)$ for alignments of length L. This exponential time complexity has become prohibitively expensive with the increasing yield of short-read sequencing technologies and motivated the adoption of DGB based assembly algorithms. It remains, however, that OLC is more robust in the presence of lower base-pair accuracy and so have renewed value in long-read sequencing experiments.

The central step of DBG assembly is the construction of the eponymous de Bruijn graph, which involves finding all overlapping sequences of some length k (a $k$mer), which less than the length of the reads. Two forms of DGB exist: Hamiltonian, where the $k$mers form nodes and the overlaps the edges; and Eulerian, where the nodes are instead the overlaps ($k-1$mers) and the edges the $k$mers. For either form, after the graph has been constructed, the next stage of the algorithm is to find long paths within the graph, which define contigs. For the Hamiltonian DBG the aim is a path with visits every node only once, whereas for the Eulerian DGB, paths may visit nodes multiple times, but can only traverse each edge once. Both forms have been successfully applied to genome assembly. As sequencing errors badly effect DGB methods by producing false $k$mers, an error-correction step is most often applied prior to the $k$mer analysis stage. A crucial benefit of DBG, in the face of the ever increasing yield of sequencing technologies, is their exceptional time and space-efficiency as the depth of coverage increases. Making DBG algorithms further suited to high-throughput short-read sequencing, is that the high depth of coverage can be exploited to improve algorithm heuristics and the impact false $k$mers is minimised by the higher base-pair accuracy.

Contig

A contig is a contiguous DNA sequence, without the suspicion of any gaps, inferred by a genome assembly algorithm. Gaps may be regions for which little or no sequence information has been generated or they may be due to unresolvable features within the assembly construct.

Scaffold

A scaffold is a series of one or more contigs whose relative order and orientation has been inferred by a genome assembly algorithm but for which gaps of indeterminate length and sequence composition exist between contigs.

CHAPTER 2

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

2.1 Authorship Declaration

By signing below I confirm that for the paper titled “Deconvoluting simulated metagenomes: the performance of hard- and soft- clustering algorithms applied to metagenomic chromosome conformation capture (3C)” and published by *PeerJ*, that:

Matthew Z. DeMaere conceived and designed the experiment, performed the experiments, analyzed the data, implemented analysis tools, wrote the paper, prepared figures and tables, reviewed drafts of the paper.

Aaron E. Darling conceived and designed the experiment and reviewed drafts of the paper.

Production Note: Signature removed prior to publication.	Production Note: Signature removed prior to publication.
Matthew Z. DeMaere	Associate Professor Aaron E. Darling

2.2 Abstract

Background Chromosome conformation capture, coupled with high throughput DNA sequencing in protocols like Hi-C and meta3C, has been proposed as viable means to generate data to resolve the genomes of microorganisms living in naturally occurring environments. Metagenomic Hi-C and meta3C 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 characterized.

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 they do not perform well when strain-level diversity exists in a metagenomic sample. While our simple extension of the Louvain method performed the best in these scenarios, its 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 sequencing depth must be carefully considered during experimental design. Finally, there appears to be significant scope to improve the accuracy of strain resolution through further algorithm development.

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

2.3 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 high-throughput 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 (during the process of DNA purification) and the subsequent shearing (during the sequencing library preparation step). What remains in the form of direct experimental observation is short-range contiguity information. From this beginning, the problem of reestablishing long-range contiguity and reconstructing the original genomic sources is handed over to genome assembly algorithms.

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 [1] thus results in a severely underdetermined inverse problem [2], [3], 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 [4], [5] and begun to work toward strain (genotype) resolution [6]. Inferring abundance per-sample from contig coverage [4], [5] or k-mer frequencies [6] respectively, the strength of this discriminating signal is a function of community diversity, environmental variation and sampling depth; and represents a significant computational task.

Chromosome conformation capture (3C), a technique first introduced to probe the three-dimensional structure of chromatin [7], has become the technological basis for a range of 3C-derived genomic strategies, all of which seek to detect the interaction of spatially proximate genomic loci. The fundamental goal in all cases is to in some way capture a snapshot of the 3D structure of a DNA target.

The methodology begins by fixation (cross-linking) of DNA within intact cells or nuclei, often by formaldehyde, to capture in-place native 3D conformational detail. The nuclei or cells are lysed and the protein-DNA complexes subjected to restriction digestion to produce free-ends. The resulting complex-bound free-ends are then religated under very low concentration, where conditions favour ligation between free-ends that were in close spatial proximity at the time of fixation. Originally, after this point, signal extraction involved known-primer locus-specific PCR amplification (3C), posing a significant experimental challenge [8] and limiting the scale of investigation. To extend its utility, subsequent advances (4C, 5C, Hi-C) have successively attempted to address the issue of scale by replacing PCR-mediated signal extraction with contemporaneous high-throughput technologies (microarrays, next-generation sequencing (NGS)) [8].

The genome-wide strategy of Hi-C [9] exploits NGS to extract interaction signal between all potential sites. To do so, before ligation the method inserts a step in which overhangs are filled with biotinylated nucleotides. Blunt-end ligation is then performed and the DNA purified and sheared. The junction-containing products are then selected for subsequent sequencing by biotin affinity pull-down.

Hi-C and the closely related meta3C (Hi-C/3C) have recently been applied to metagenomics [10]–[12], intended as an alternative to purely computational solutions to community deconvolution. Here conventional metagenomic sequencing is augmented with the information derived from Hi-C/3C read-pairs to provide strong experimental evidence of proximity between genomic loci. This map of interactions greatly increases the power of discrimination between community member genomes, by measuring which sequences were spatially nearby at the time of fixation.

Given sufficient sampling depth, Hi-C/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 [10], [13]. As with any real experimental process, the generation Hi-C/3C read-sets is imperfect. Three complications to downstream signal processing are: self-self religations which effectively produce local read-pairs, chimeric read-throughs which span the ligation junction and contain sequence from both ends, and spurious read-pairs involving non-proximity ligation products. Though not insurmountable when integrating Hi-C/3C data with that of conventional sequencing, these flawed products do at the very least represent a loss of efficiency in generating informative proximity ligation read-pairs.

Sequencing information generated in this way can recover a portion of the information lost in conventional whole genome shotgun (WGS) sequencing. It has been shown that the observational probability of intra-chromosomal read-pairs (*cis*) follows a long-tailed distribution decreasing exponentially with increasing genomic separation [10]. 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 [10]. This hierarchy in observational probability has the potential to be a precious source of information with which to deconvolute assembled sequence fragments derived from conventionally generated metagenomes into species and perhaps strains.

Previous work which leverages 3C data in assembly analysis has yielded algorithms focused on scaffolding [14], [15]. 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 each bin is scaffolded 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 been thoroughly investigated.

In this work, we describe the accuracy of various analysis algorithms applied to resolving the genomes of strains within metagenomic sequence data. The accuracy of these algorithms was measured over a range of simulated experimental conditions, including varying degrees of evolutionary divergence around our operationally defined species boundary (Figure 2.3), 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.

2.4 Materials and Methods

2.4.1 Representation

A contact map is formed by mapping proximity-ligation read-pairs to the available reference and counting occurrences between any two genomic regions [16]; where the definition of a genomic region is application dependent. Mathematically, the contact map is a square symmetric 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 [9], [17]–[19].

Noting that the contact map is equivalent to the weighted adjacency matrix A of an undirected graph G [20], an alternative graphical representation is obtained. Here, nodes n_i represent genomic regions and weighted edges $e(n_i, n_j, w_{ij})$ represent the observed frequency w_{ij} of 3C read-pairs linking regions n_i and n_j. Expressing the sequencing data as such, a host of graph-theoretic analysis methods can be brought to bear on domain-specific problems.

Possibly the simplest variation, the eponymous 3C-contig graph, defines the genomic regions (and thereby the nodes) to be the set contigs produced by WGS assembly. Fine details such as small indels or single nucleotide variants are not considered with this construction. Even so, the application of the 3C-contig graph to metagenomics [10]–[12] and multichromosomal genome scaffolding [14] has previously been studied.

The chosen granularity of any construct is a crucial factor in obtaining both sufficiently detailed answers and tractable problems. Though finer scale representations are possible when integrating Hi-C/3C data into conventional metagenomics, the 3C-contig graph is an effective means of controlling problem scale and can be regarded as a first step toward deeper Hi-C/3C metagenomic analyses.

2.4.2 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 the 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 an object o_i may only belong to a single cluster κ_k, it is termed hard-clustering (i.e. $\forall \kappa_k, \kappa_l \in K \mid k \neq l \rightarrow \kappa_k \cap \kappa_l = \emptyset$). When this condition on membership is relaxed and objects are allowed to belong to multiple clusters, it is termed soft-clustering. The outcome of this potential for multiple membership is cluster overlap, or more formally, that the intersection between clusters κ_k and κ_l is no longer strictly empty (i.e. $|\kappa_k \cap \kappa_l| \geq 0$).

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 metagenomes containing closely related species or strains, there is a tendency for the highly conserved core genome to co-assemble in single contigs while more distinct accessory regions do not. Co-assembly implies that uniquely placing (a 1-to-1 mapping of) contigs into source-genome bins (clusters) is not possible. Rather, an overlapping model is required, allowing co-assembled contigs to be placed multiple times in relation to their degree of source-heterogeneity.

From the aspect of prior knowledge, classification and clustering algorithms fall into three categories [21]. Supervised classification, where for a known set of classes, a set of class-labelled objects are used to determine a membership function; semi-supervised classification/clustering, which leverages additional unlabelled data as a means of improving the supervised membership function; and unsupervised clustering, where these prerequisites are not required. Unsupervised algorithms, in removing this *a priori* condition, are preferable if not necessary in situations where prior knowledge is unavailable (perhaps due to cost or accessibility) or the uncertainty in this information is high.

2.4.3 Appropriate Validation Measures

Simply put, clustering algorithms attempt to group together objects when they are similar (the same cluster) and separate those objects which differ (different clusters). Although algorithmic complexity can ultimately dictate applicability to a given problem domain, the quality of a clustering solution remains a primary concern in assessing an algorithm's value. 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 [22]: 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 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 (Equation 2.1) 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 [22] and further is sensitive to a preprocessing step where clusters and class labels must first be matched [23]. The entropy based V-measure [23] 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 B^3 [24] addresses all five aspects and building from this, extended B^3 [22] supports the notion of overlapping clusters and classes. Analogous to F_1-score and V-measure, extended B^3 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 B^3 (Equation 2.1) and propose using contig length (bp) as the measure of inherent value when clustering metagenomic contigs.

2.4.4 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 on both hard and soft clustering approaches.

Four graph clustering algorithms were considered: MCL, SR-MCL, the Louvain method and OClustR [25]–[28]. While MCL and Louvain have previously been applied to 3C-contig clustering [10], [12], to our knowledge SR-MCL and OClustR have not. We did not consider the clustering algorithm employed by [11] 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 2.2). The widely used MCL (markov clustering) algorithm [25] 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 [10] 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) [26] 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 (results not shown), and therefore only inflation was varied over the same range as MCL.

The Louvain modularity Q [29] 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 [27], following a greedy heuristic to determine the 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 [30] and Networkx [31]. We further extended this 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 [28] 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) [28]. The approach functions without the need for runtime parameters, thus avoiding their optimization, and aims to produce clusters of minimum overlap and maximal size.

2.4.5 Gold Standard

A crucial element of external validation is the gold-standard (ground truth). Particularly in the treatment of scientific data, what we call the gold-standard is frequently a "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 the genomic origin of each contig 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) [32] 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 matrix M of contig coverage indexed by both reference genome $g_k \in G$ and contig base position l. Rather than a binary representation, mask element M_{kl} was assigned a real value $[0, 1]$ proportional to the identity of the maximal covering alignment to reference genome g_k at site l. Lastly, the arithmetic mean μ_k was calculated over all base positions for each reference genome g_k (i.e. $\mu_k = L_k^{-1} \sum_l M_{kl}$, where L_k is the length of genome g_k) and an association between contig c_i and reference genome g_k was accepted if $\mu_k > 0.96$.

2.4.6 Graph Generation

Undirected 3C-contig graphs were generated by mapping simulated 3C read-pairs to WGS assembly contigs using BWA MEM (v0.7.9a-r786) [33]. 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 [11] effective in this case (results not shown). Assembly contigs defined the nodes n_i and inter-contig read-pairs the edges ((n_i, n_j) is an edge iff $i \neq j$), while intra-contig read-pairs ($(n_i, n_j) \iff i = j$) were ignored. Raw edge weights w_{ij} were defined as the observed number of read-pairs linking nodes n_i and n_j.

2.4.7 Validation

To assess the quality of clustering solutions a modification to the Extended B^3 external validation measure [22] was made, wherein each clustered object was given an explicit weight. We call the resulting measure “weighted B^3” (Equation 2.1). For a uniform weight distribution, this modification reduces to conventional Extended B^3. In our work, contig length (bp) was chosen as the weight when measuring the accuracy of clustered assembly contigs. Remaining the harmonic mean of B^3 precision and recall, the weights $w(o_i)$ are introduced here (Equation 2.2, 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).

$$F_{\mathrm{B}^3} = \frac{2 \left\langle P_{\mathrm{B}^3} \right\rangle \left\langle R_{\mathrm{B}^3} \right\rangle}{\left\langle P_{\mathrm{B}^3} \right\rangle + \left\langle R_{\mathrm{B}^3} \right\rangle} \tag{2.1}$$

where $\left\langle P_{\mathrm{B}^3} \right\rangle$ and $\left\langle R_{\mathrm{B}^3} \right\rangle$ are the weighted arithmetic means of $P_{\mathrm{B}^3}(o_i)$ and $R_{\mathrm{B}^3}(o_i)$ (Equation 2.2, 2.3) over all objects.

$$P_{\mathrm{B}^3}(o_i) = \frac{1}{\sum_{o_j \in D(o_i)} w(o_j)} \sum_{o_j \in D(o_i)} w(o_j) P^*(o_i, o_j) \tag{2.2}$$

$$R_{\mathrm{B}^3}(o_i) = \frac{1}{\sum_{o_j \in H(o_i)} w(o_j)} \sum_{o_j \in H(o_i)} w(o_j) R^*(o_i, o_j) \tag{2.3}$$

Unchanged from Extended B^3, the expressions for the Multiplicity B^3 precision $P^*(o_i, o_j)$ (Equation 2.4) and recall $R^*(o_i, o_j)$ (Equation 2.5) account for the non-binary relationship between any two items in the set when dealing with overlapping clustering.

$$P^*(o_i, o_j) = \frac{min\left(\left|K(o_i) \cap K(o_j)\right|, \left|\Theta(o_i) \cap \Theta(o_j)\right|\right)}{\left|K(o_i) \cap K(o_j)\right|} \tag{2.4}$$

$$R^*(o_i, o_j) = \frac{min\left(\left|K(o_i) \cap K(o_j)\right|, \left|\Theta(o_i) \cap \Theta(o_j)\right|\right)}{\left|\Theta(o_i) \cap \Theta(o_j)\right|} \tag{2.5}$$

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

2.4.8 Simulating Hi-C/3C read-pairs

A tool for simulating Hi-C/3C read-pairs was implemented in Python (`simForward.py`). Read-pairs were generated for a given community directly from its reference genomes, where the relative proportion of read-pairs from a given taxon adhered to the community's abundance profile. Inter-chromosomal (*trans*) pairs were modeled as uniformly distributed across the entire chromosomal extent of a given genome. For intra-chromosomal (*cis*) pairs, a linear combination of the geometric and uniform distributions was used to approximate a long-tailed probability distribution as a function of genomic separation and calibrated by fitting to real experimental data [10]. For these 3C reads, the modeling of experimental/sequencing error was not performed. Variation in intra-chomosomal probability attributable to 3D chromosomal structure was not included. In effect, chromosomes were treated as flat unfolded rings. The tool takes as input a seed, read length, number of read-pairs, abundance profile and inter-chromosomal probability and outputs reads in either interleaved FastA or FastQ format.

2.4.9 Pipeline Design

The chosen workflow (Figure 2.1) represents a simple and previously applied [10], [11] 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) [34] was used for assembly. BWA MEM was used for mapping 3C read-pairs to contigs [33]. 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. The Python implementation of our end-to-end pipeline is available at `https://github.com/koadman/proxigenomics`.

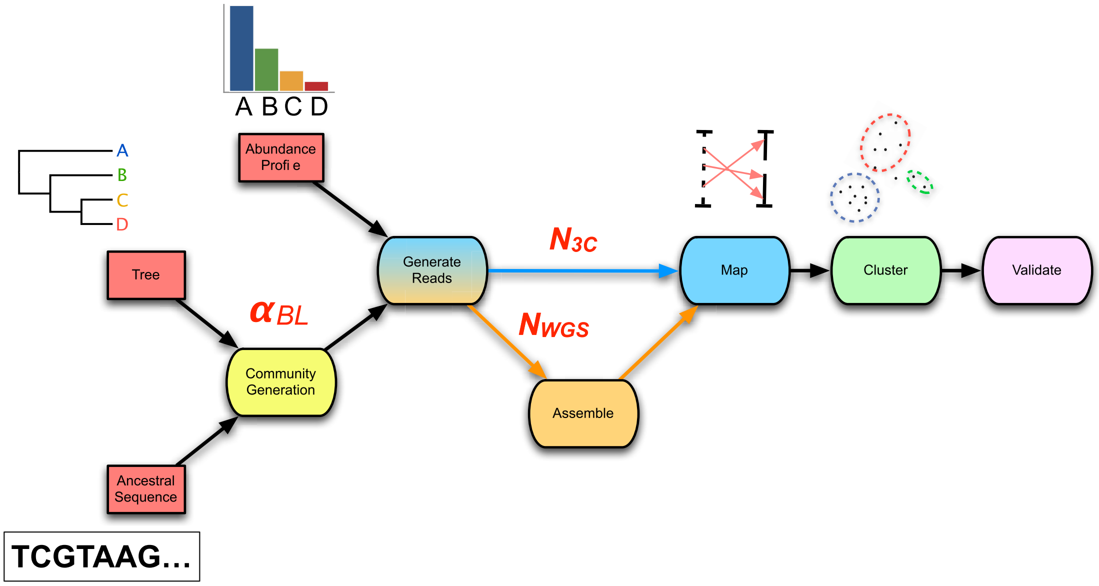

Figure 2.1: The 3C sequencing simulation pipeline used within the parameter sweep. An ancestral sequence and phylogenetic tree are used in simulating a process of genome evolution with varying divergence (α_{BL}). The resulting evolved genomes are subsequently subjected to *in silico* high-throughput 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 3C-contig graph. Finally, the graph is supplied to a clustering algorithm and the result validated against the relevant gold-standard.

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 [35], 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 2.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 2.1).

Following community generation, the read-simulation module takes as input the set of descendent genomes G and generates as output both simulated Illumina WGS paired-end reads and simulated Hi-C/3C read-pairs. For WGS reads, variation in relative abundance of descendent genomes G was produced by wrapping ART_illumina (v1.5.1) [36] within a Python script (`metaART.py`) with the added dependency of an abundance profile table as input. Hi-C/3C read-pairs were generated from community genomes as outlined above. Generation of the two forms of read-pairs introduces the following sweep parameters: WGS depth of coverage (xfold) and number of 3C read-pairs (n3c) (Table 2.1).

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 3C-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 [10], [11].

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 B^3 (Equation 2.1), while the set of contigs is supplied to QUAST (v3.1) [37] for the determination of conventional assembly statistics. The results from both clustering and assembly validation are then joined together to form a final output.

2.4.10 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 2.2), to develop insight into the challenges that face metagenomics researchers choosing to apply 3C to communities which contain closely related taxa.

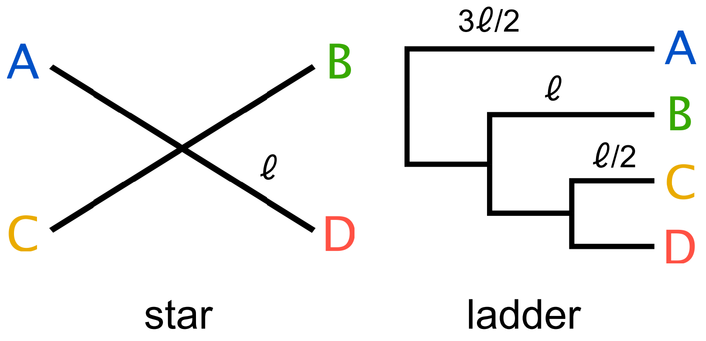

Figure 2.2: Two simple trees of four taxa (A,B,C,D) were used in the parameter sweep. The star; where all taxa have equal evolutionary distance ℓ to their ancestor and ladder; where the distance to the nearest ancestor decreases in incremental steps of $\ell/2$. For the ladder, the length of the internal 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.

2.4.11 Parameter Sweep

A single monochromosomal ancestral genome was used throughout (*Escherichia coli* K12 substr. MG1655 (acc: NC_000913)). Two simple ultrametric tree topologies of four taxa (tree: star, ladder) (Figure 2.2) 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 2.3). Two community abundance profiles were tested: uniform abundance and one of decreasing abundance by factors of $1/e$ (i.e. $1, 1/e, 1/e^2, 1/e^3$) (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 (3C sampling depth) was varied in five steps from ten thousand to one million pairs (n3c: 10k, 20k, 50k, 100k, 1M), while the remaining parameter variations can be found in Table 2.1 and Table 2.2.

From the 40 simulated microbial communities, the resulting 120 simulated metagenome read-sets were assembled and the assemblies evaluated using QUAST (v3.1) [37] 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

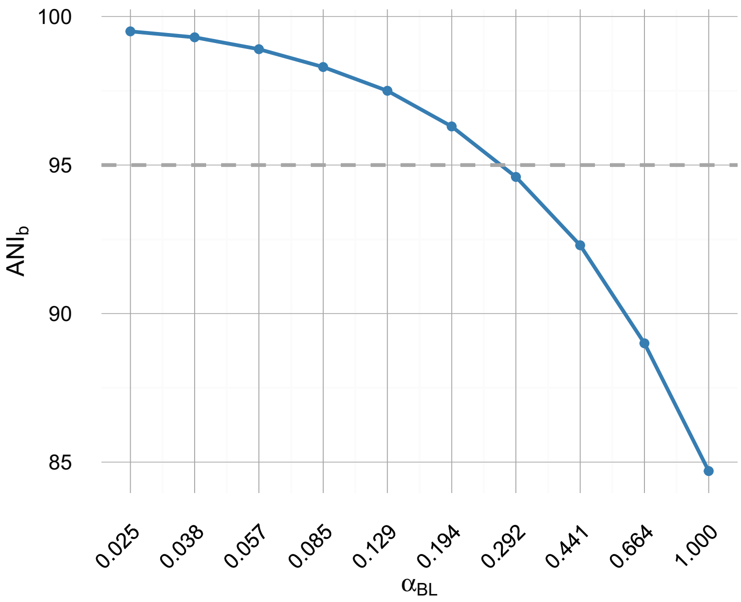

Figure 2.3: For sample points used in the sweep for the star topology, we depict the relationship between branch length scale factor α_{BL} and the resulting measure of average nucleotide identity from BLAST (ANIb). The 95% threshold indicated is used internally within IDBA-UD [38] to determine whether to merge highly similar contigs and has been proposed as a pragmatic definition of bacterial species [39], [40] akin to 97% 16S rRNA identity.

resulted in 600 distinct combinations of simulation parameters, forming the basis for input to the selected clustering algorithms. OClustR results in 600 clusterings; Louvain clustering was performed both as standard hard-clustering (Louvain-hard) and our naive soft-clustering modification (Louvain-soft) resulting in 600 clusterings each; lastly MCL and SR-MCL were both varied over one parameter (infl) in 5 steps resulting in 3000 clusterings each. Finally, the quality of the clustering solutions for all four algorithms was assessed using the weighted B^3 (Equation 2.1) external validation measure. Other parameters fixed throughout the sweep were: ancestral genome size (seq-len: 3 Mbp), indel/inversion/HT rate multiplier (sg_scale: 10^{-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); Hi-C/3C parameters (read-length: 150 bp, restriction enzyme: NlaIII [_CATGˆ]). As simulated genomes were monochromosomal, inter-chromosomal read-pair probability was not a factor.

2.4.12 Assembly Entropy

A normalized entropy based formulation S_{mixing} (Equation 2.6) 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 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 2.7), where mixing was regarded as a function of α_{BL} and the difference taken between successive sample points in the sweep.

$$\begin{gathered} S_{mixing} = -\frac{1}{L_{asm}\log_2(N_{eff})}\sum_{j=1}^{N_C} l_j \sum_{i=1}^{N_G} p_{ij}\log_2(p_{ij}) \\ L_{asm} = \sum_{j=1}^{N_C} l_j, \qquad N_{eff} = \sum_{i=1}^{N_G} q_i \end{gathered} \tag{2.6}$$

$$\Delta S_{mixing}(\alpha_{BL}) = \frac{1}{h}\left(S_{mixing}(\alpha_{BL}+h) - S_{mixing}(\alpha_{BL})\right) \tag{2.7}$$

2.4.13 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 2.8) 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

[41]–[43].

$$H_L = \sum_{\lambda_i \in \{\lambda : \lambda > 0\}} |\lambda_i| \log_2 |\lambda_i| \tag{2.8}$$

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

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	5	600	10000, 20000, 50000, 100000, 1000000
6	algo	Clustering algorithm	factor	5		MCL, SM-MCL, Louvain-hard, Louvain-soft, OClustR

Table 2.1: Primary parameters under control in the sweep. In total, each clustering algorithm is presented with 600 combinations which may further increase depending on whether a clustering algorithm also has runtime parameters under control.

2.5 Results

2.5.1 Experimental Design

We implemented a computational pipeline which is capable of simulating arbitrary metagenomic Hi-C/3C sequencing experiments (Figure 2.1). The pipeline exposes parameters governing both the process of sequencing and community composition for

Algorithm	Name	Description	Type	Number	Total	Values	Sampling
MCL	infl	Inflation parameter	numeric	5	3000	$1.1 - 2$	linear
SR-MCL	infl	Inflation parameter	numeric	5	3000	$1.1 - 2$	linear
Louvain-hard				1	600		
Louvain-soft				1	600		
OClustR				1	600		

Table 2.2: Clustering algorithm dependent parameters explored in the sweep, where the base set of combinations begins with the fundamental 600 combinations. Only MCL and SR-MCL were swept through additional runtime parameters.

control by the researcher and further, provides the facility for performing parametric sweeps on these parameters (Table 2.1).

The pipeline was used to vary community composition, in particular, the degree of within-community evolutionary divergence, and evaluate its impact on the accuracy of genomic reconstruction. Starting from an ancestral sequence, a phylogenetic tree and an abundance profile; 10 communities were generated with varying evolutionary divergence by scaling branch length (Figure 2.2). The range of evolutionary divergence was chosen so as to go from a region of easily separable species ($\approx 85\%$ ANI) to that of very closely related strains ($\approx 99.5\%$ ANI) (Figure 2.3). The sweep included variation of both WGS coverage (xfold: $10x$, $50x$, $100x$) and the number of Hi-C/3C read-pairs (n3c: 10^4 to 10^6) to assess the impact of increased sampling on reconstruction.

Genomic reconstruction was performed using five different graph clustering algorithms (Table 2.2) on the 600 3C-contig graphs resulting from the sweep. The quality of each solution was then evaluated using our weighted B^3 metric F_{B^3} (Equation 2.1), where the relevant gold-standard as also generated by the pipeline. The resulting dataset is publicly available at `http://doi.org/10.4225/59/57b0f832e013c`.

2.5.2 Assembly Complexity

Along with traditional assembly validation statistics (N50, L50) (Figure 2.4A, 2.4B), assembly entropy S_{mixing} and its approximate first order derivative ΔS_{mixing} (Equation 2.6, 2.7) (Figure 2.4C) were calculated for all 120 combinations resulting from the first four levels of the sweep (parameters: tree, profile, α_{BL}, xfold) (Table 2.1).

As community composition moved 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%), the degree of contig mixing increased more or less monotonically, and was delayed by increased read-depth. After α_{BL}, the only significant continuous variable influencing mixing was read-depth (Spearman's ρ=-0.26, $P < 4 \times 10^{-3}$), while abundance profile was the only significant categorical variable (one-factor ANOVA R^2=0.0774, $P < 3 \times 10^{-3}$) [44]. In all cases, as α_{BL} decreased mixing approached unity; implying that as genomic sources became more closely related, the resulting metagenomic assembly contigs were 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 2.4C 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 coincided with the region where assemblies were the most fragmented (max L50, min N50) (Figure 2.4A, 2.4B) and ΔS_{mixing} was well correlated with both N50 (Spearman's $\rho = 0.72$, $P < 1 \times 10^{-5}$) and L50 (Spearman's $\rho = -0.83$, $P < 1 \times 10^{-7}$), implying that as community divergence decreased through this critical point, traditional notions of assembly quality followed suit.

2.5.3 Graph Complexity

The introduction of 3C sampling depth (number of read-pairs) at the next level within the sweep (parameter: n3c) generated 480 3C-contig graphs (Table 2.1). To assess how assembly outcome affects the derived graph: order, size, density, and entropy H_L (Equation 2.8) were calculated and subsequently joined with the associated factors from assembly (Figure 2.4D).

Per the definition of the 3C-contig graph, there was a strong linear correlation between graph order $|n|$ and L50 (Pearson's $r = 0.96$, $P < 3 \times 10^{-16}$) and a weaker but still significant linear correlation between graph size $|e|$ and 3C sampling depth (parameter: n3c) (Pearson's $r = 0.66$, $P < 3 \times 10^{-16}$). Graphical density was moderately linearly correlated with graphical complexity H_L (Pearson's $r = -0.63$, $P < 3 \times 10^{-16}$), and strongly correlated with assembly statistics N50 (Spearman's $\rho = -0.97$, $P < 3 \times 10^{-16}$), L50 (Spearman's $\rho = 0.96$, $P < 3 \times 10^{-16}$) and ΔS_{mixing} (Spearman's $\rho = -0.73$, $P =< 1 \times 10^{-16}$).

The knock-on effect of evolutionary divergence on the 3C-contig graphs derived from metagenomic assemblies was clear; fragmented assemblies comprised of contigs of mixed heritage resulted in increased 3C-contig graph complexity. As 3C read-pairs are the direct observations used to infer an association between contigs, it could be expected that the correlation between 3C sampling depth and graphical size ($|e|$) would be high ($r \rightarrow 1$). In fact, we observed a more moderate correlation ($r = 0.66$) and, because spurious read-pairs were excluded in our simulations, what might be perceived as a shortfall in efficiency was simply the accumulation of repeated observations of read pairs linking the same contig pairs. Therefore by the nature of the experiment, increased 3C sampling depth did not lead to increased graphical complexity in the same way that a more fragmented assembly would. Instead, increased 3C sampling depth can significantly improve the quality of clustering solutions by increasing the probability of observing rare associations and repeat observations of existing associations.

2.5.4 Clustering Validation

The 300 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 2.1), were joined with their relevant sweep parameters and PCA performed in R (FactoMineR v1.32) [44]. The first three principal components explained 75% of the variation, where PC1 was primarily involved with factors describing graphical complexity (α_{BL}: $r = 0.91$, $P < 2 \times 10^{-118}$; density: $r = 0.67$, $P < 8 \times 10^{-41}$; order: $r = -0.75$, $P < 3 \times 10^{-56}$; ANIb: $r = -0.91$, $P < 2 \times 10^{-117}$; H_L: $r = -0.91$, $P < 3 \times 10^{-113}$), PC2 factors described the sampling of contig-contig associations and overall connectedness of the 3C-contig graph (size: $r = 0.84$, $P < 2 \times 10^{-79}$; n3c: $r = 0.84$, $P < 7 \times 10^{-82}$; modularity: $r = -0.40$, $P < 9 \times 10^{-13}$) and PC3 pertained to local community structure (modularity: $r = 0.73$, $P < 1 \times 10^{-49}$; and xfold: $r = 0.53$, $P < 3 \times 10^{-23}$) (Figure 2.5).

Of the five clustering algorithms, the performance of four (MCL, SR-MCL, Louvain-hard and OClustR) was strongly correlated with PC1 and so their solution quality was inversely governed by the degree of complexity in the input graph, which in turn was 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, possessed significant correlation with PC2 ($r = 0.53$, $P < 5 \times 10^{-23}$) and thus noticeably benefited from increasing the number of 3C read-pairs (Figure 2.5).

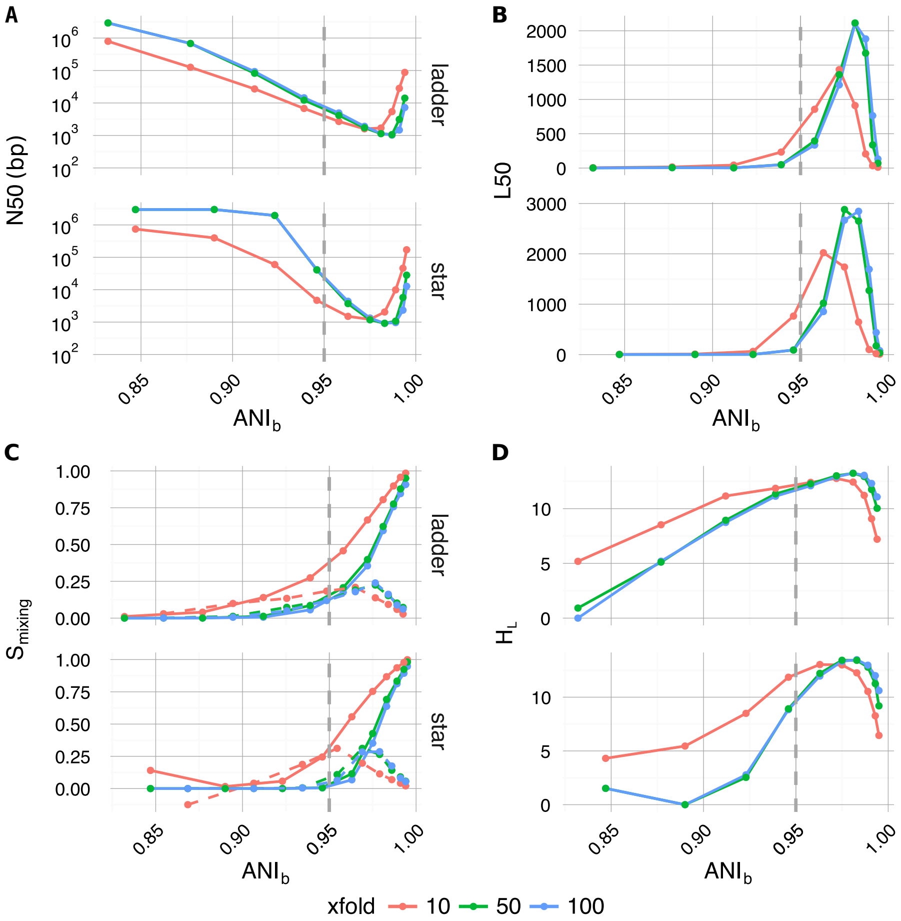

Figure 2.4: Plotted as a function of evolutionary divergence (measured by ANI_b) for the star and ladder communities at three depths of WGS coverage (10, 50 and 100x); assembly validation statistics N50 (**A**) and L50 (**B**), the degree of genome intermixing S_{mixing} and its approximate first order derivative ΔS_{mixing} (dashed lines) (**C**), lastly graphical complexity H_L (**D**). The vertical grey dashed line in each panel marks our operationally defined species boundary ($ANI_b = 95\%$). As evolutionary divergence decreased from easily separable species ($ANI_b \approx 85\%$) to very closely related strains ($ANI_b \rightarrow 1$), assemblies went through a transition from a state of high purity (S_{mixing} ≈0) to a highly degenerate state (S_{mixing} ≈1), where many contigs were composed of reads from all community members. A crisis point was observed for small evolutionary divergence ($\alpha_{BL} < 0.2924$, $ANI_b < 95\%$), where a sharp change in contiguity (implied by N50 and L50) occured. At very low divergence, N50 and L50 statistics implied that assemblies were recovering, while source degeneracy (S_{mixing}) monotonically increased. Graphical complexity (H_L) exhibited a similar turning point to L50 and was dominated by graph order $|n|$ (number of contigs/nodes).

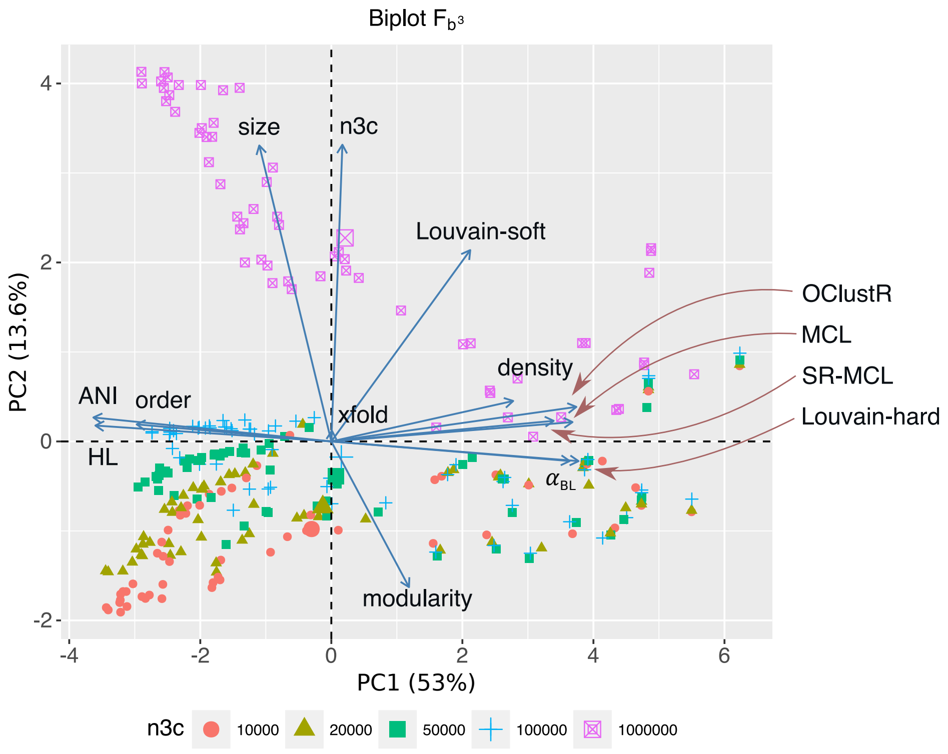

Figure 2.5: For the 300 3C-contig graphs pertaining to uniform abundance, a PCA biplot is shown for the two most significant components (PC1, PC2). Respectively, PC1 and PC2 explain 53% and 13.6% of the variation within the data-set. Here vectors represent sweep variables, while points represent individual 3C-contigs graphs and are coloured by 3C sequencing depth (n3c: 10k - 1M pairs). Double-sized points show mean values of these n3c groupings. Vectors labelled after the five clustering algorithms represent performance as measured by scoring metric F_{B^3} (Equation 2.1) PC1 and PC2) explained 53% and 13.6% of the variation within the data-set respectively. PC1 was most strongly correlated with graphical complexity (H_L) and the number of graph nodes ($order$), which come about with decreasing evolutionary divergence (ANI_b and α_{BL}) and explained the majority of variation in performance for four out of five clustering algorithms. The notable exception was Louvain-soft which had significant support on PC2. PC2 was related to Hi-C/3C sampling depth ($n3c$), which correlated strongly with the number of graph edges ($size$). The positive response Louvain-soft had to increasing the number of Hi-C/3C read-pairs ($n3c$) relative to the remaining four algorithms is evident.

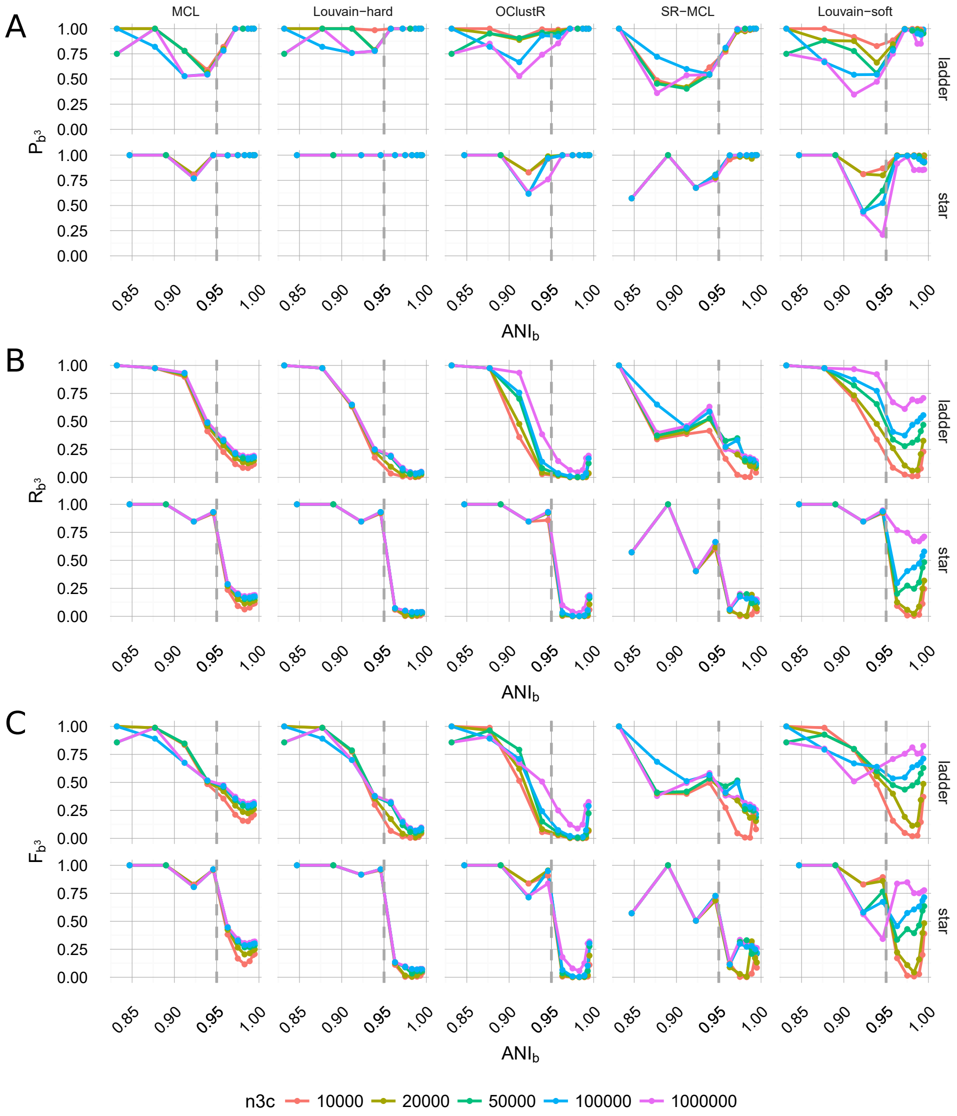

Figure 2.6: Performance of the five clustering algorithms (MCL, Louvain-hard, OClustR, SR-MCL, Louvain-soft), as measured by weighted extended B^3 precision P_{B^3} (**A**), recall R_{B^3} (**B**) and their harmonic mean F_{B^3} (**C**) (Equation 2.1 - 2.3). The slice from the sweep pertained to uniform abundance and 100x WGS coverage and the best performing runtime parameters specific to algorithms (i.e. for MCL and SR-MCL inflation=1.1). (**A**) Louvain-hard demonstrated high precision throughout, while our simple modification Louvain-soft lead to a drop, particular in the region of intermediate evolutionary divergence. (**B**) All algorithms struggled to recall the four individual genomes as evolutionary divergence decreased and cluster overlap grew. Within the region of overlap, Louvain-soft performed best and clearly benefited from increasing the number of Hi-C/3C read-pairs (n3c: $10^4 - 10^6$). (**C**) In terms of F_{B^3}, the harmonic mean of Recall and Precision, only Louvain-soft appeared to be an appropriate choice when it might be expected that strain-level diversity exists within a microbial community.

2.6 Discussion

By selecting a slice from within the sweep and the best-scoring runtime configuration for each algorithm, a qualitative per-algorithm comparison of clustering performance under ideal conditions can be made (Figure 2.6). For evolutionary divergence well above the level of strains and prior to the critical region of assembly ($\alpha_{BL} \gg 0.292$, $ANI_b \ll 95\%$), all algorithms achieved their best performance ($F_{\mathrm{B}^3} \rightarrow 1$) (Figure 2.6C). As evolutionary divergence decreased toward the level of strains and the assembly process approached the critical region, a fall-off in performance was evident for all algorithms and this performance drop is largely attributable to the loss of recall (Figure 2.6B). Hard-clustering algorithms (MCL, Louvain-hard) in general exhibited superior precision (Figure 2.6A) to that of soft-clustering algorithms (SR-MCL, OClustR, Louvain-soft) and the precision of soft-clustering algorithms was worst in the critical region where graphical complexity was highest.

A hundred-fold increase in the number of 3C read-pairs ($10^4 - 10^6$) had only a modest effect on clustering performance for four of the five algorithms, the exception being our naive Louvain-soft. Louvain-soft made substantial gains in recall from increased Hi-C/3C sampling depth at evolutionary divergences well below the level of strains ($\alpha_{BL} < 0.085$, $ANI_b < 98\%$), but sacrificed precision at large and intermediate evolutionary divergence. The soft-clustering SR-MCL also sacrificed precision but failed to make similar gains in recall as compared to Louvain-soft. Recall for all three hard-clustering algorithms (MCL, Louvain-hard, OClustR) decreased with decreasing evolutionary divergence as the prevalence of degenerate contigs grew. This drop in recall was particularly abrupt for the star topology where, within the assembly process, all taxa approached the transitional region simultaneously. Being primarily limited by their inability to infer overlap, increase in 3C sampling depth for the hard-clustering algorithms had 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 the 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 and 1 million Hi-C/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 [12], 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% [45]. We note that current Illumina MiSeq V3 kits are specified to produce up to $\approx 2 \times 10^7$ read-pairs, while HiSeq 2500 V4 lanes are specified to yield up to $\approx 5 \times 10^8$ 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 Hi-C protocol [10] may be advantageous to achieve higher proximity ligation read rates, with up to 50% of read pairs spanning over 1kbp.

2.6.1 Limitations and Future Work

Our simulation of 3C read-pairs did not include modeling of experimental noise in the form basic sequencing error nor spurious ligation products 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 [45]. As a first approximation, we feel it reasonable to assume that 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. Simulated communities were chosen to be particularly simple for the sake of downstream comprehension. Larger and more complex phylogenetic topologies are called for in fully assessing real-world performance. For the entire sweep, only a single ancestral genome (*Escherichia coli* K12 substr. MG1655) was used in generating the simulated communities and its particular characteristics represent will have biased genome assembly and sequence alignment tasks within the work-flow. As future work, a more thorough sampling of available microbial genomes and more complicated community structures could be investigated.

Only raw edge weights were used in our analysis because normalization procedures, such as have been previously employed [10]–[12], proved only weakly beneficial at higher 3C sampling depths and occasionally detrimental in situations of low sampling depth (results not shown). For higher sampling depth, the weak response can likely be attributed to a lack of complexity and the low noise environment inherent in the simulation. For low sampling depth, observation counts are biased to small values (mode $[w(n_i, n_j)] \rightarrow 1$) 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 ($f : \mathbb{N} \rightarrow \mathbb{R}$). 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 a growing number of tools intended explicitly for metagenomes, such as metaSPAdes [46], MEGAHIT [47], or populations of related genomes (Cortex) [48], while the modular MetAMOS suite [49] at once offers tantalising best-practice access to the majority of alternatives. For Hi-C/3C analysis, a desirable feature of read mapping tools is the capability to report split read alignments (e.g. BWA MEM) [33]. 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 [50] or re-expression of soft-clustering solutions into a non-overlapping context [51]. While a soft-clustering reformulation of normalized mutual information (NMI) [50] has become frequently relied on in clustering literature [52], alongside B^3 the two have been shown to be complementary measures [53]. Therefore, although the choice to rely on the single measure we proposed here (Equation 2.1) 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 3C-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 [54], [55]. 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, the parameter set we defined for the pipeline (Table 2.1) 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 for the purpose of deconvolving microbial communities via 3C read data. Not all algorithms are wholly unsupervised. Individually they 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 [51], [56], [57], 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 increasing 3C sampling depth (Figure 2.6) 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) [58] and shown to improve both Louvain and label propagation methods, is a clear candidate for future assessment.

2.7 Conclusion

For a microbial community, as evolutionary divergence decreases within the community, contigs derived from WGS metagenomic assembly increasingly become a mixture of source genomes. When combined with 3C information to form a 3C-contig graph, evolutionary divergence is directly reflected by the degree of community overlap. We tested the performance of both hard and soft clustering algorithms to deconvolute simulated metagenomic assemblies into their constituent genomes from this most simple 3C-augmented representation. Performance was assessed by our proposed weighted variation of extended B^3 validation measure (Equation 2.1), where here weights were set proportional to contig length. We have shown that soft-clustering algorithms can significantly outperform hard-clustering algorithms when intra-community evolutionary divergence approaches a level traditionally regarded as existing between microbial 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 of 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 focuses on the application of recent advances in soft-clustering methods.

2.8 Additional Information and Declarations

2.8.1 Competing Interests

The authors declare there are no competing interests.

2.8.2 Author Contributions

Matthew Z. DeMaere conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents, material, analysis tools, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.

Aaron E. Darling conceived and designed the experiments, contributed reagents, materials, analysis tools and reviewed drafts of the paper.

2.8.3 Data Availability

The following information was supplied regarding data availability:

Source code hosted on Github: `https://github.com/koadman/proxigenomics.git`

Data hosted by our institutions longterm archival service UTS Research Data. DOI:10.4225/59/57b0f832e013c.

2.8.4 Funding

This work was supported under Australian Research Council's Discovery Projects funding scheme (project number LP150100912). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

2.9 List of abbreviations

- 3C - chromosome conformation capture

- ANI - average nucleotide identity
- ANOVA - analysis of variance
- B^3 - an extrinsic validation measure
- bp - base-pair
- *cis* - intra-chromosomal
- DNA - deoxyribonucleic acid
- GMM - Gaussian mixture model
- Hi-C - high throughput sequencing 3C
- Mbp - mega base-pair
- MCL - Markov clustering
- meta3C - metagenomic 3C
- NGS - next generation sequencing
- NMF - non-negative matrix factorization
- PCA - principal component analysis
- SBM - stochastic block model
- SR-MCL - Soft regularised Markov clustering
- *trans* - inter-chromosomal
- WGS - whole genome shotgun

2.10 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 supported by the EIF and NCRIS.

2.11 References

[1] S. G. Tringe and E. M. Rubin, “Metagenomics: DNA sequencing of environmental samples”, en, *Nat. Rev. Genet.*, vol. 6, no. 11, pp. 805–814, Nov. 2005, ISSN: 1471-0056. DOI: `10.1038/nrg1709`. [Online]. Available: `http://dx.doi.org/10.1038/nrg1709`.

[2] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigó, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk,

Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu, “The sequence of the human genome”, en, *Science*, vol. 291, no. 5507, pp. 1304–1351, Feb. 2001, ISSN: 0036-8075. DOI: 10.1126/science.1058040. [Online]. Available: http://dx.doi.org/10.1126/science.1058040.

[3] E. W. Myers Jr, “A history of DNA sequence assembly”, *it - Information Technology*, vol. 58, no. 3, pp. 126–132, Jan. 2016, ISSN: 1611-2776, 2196-7032. DOI: 10.1515/itit-2015-0047. [Online]. Available: https://www.degruyter.com/view/j/itit.2016.58.issue-3/itit-2015-0047/itit-2015-0047.xml.

[4] J. Alneberg, B. S. Bjarnason, I. de Bruijn, M. Schirmer, J. Quick, U. Z. Ijaz, N. J. Loman, A. F. Andersson, and C. Quince, “CONCOCT: Clustering cONtigs on COverage and ComposiTion”, Dec. 2013. arXiv: 1312.4038 [q-bio.GN]. [Online]. Available: http://arxiv.org/abs/1312.4038.

[5] M. Imelfort, D. Parks, B. J. Woodcroft, P. Dennis, P. Hugenholtz, and G. W. Tyson, “GroopM: An automated tool for the recovery of population genomes from related metagenomes”, en, *PeerJ*, vol. 2, e603, Sep. 2014, ISSN: 2167-8359. DOI: 10.7717/peerj.603. [Online]. Available: http://dx.doi.org/10.7717/peerj.603.

[6] B. Cleary, I. L. Brito, K. Huang, D. Gevers, T. Shea, S. Young, and E. J. Alm, “Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning”, en, *Nat. Biotechnol.*, vol. 33, no. 10, pp. 1053–1060, Oct. 2015, ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.3329. [Online]. Available: http://dx.doi.org/10.1038/nbt.3329.

[7] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner, “Capturing chromosome conformation”, en, *Science*, vol. 295, no. 5558, pp. 1306–1311, Feb. 2002, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1067799. [Online]. Available: http://dx.doi.org/10.1126/science.1067799.

[8] E. de Wit and W. de Laat, “A decade of 3C technologies: Insights into nuclear organization”, en, *Genes Dev.*, vol. 26, no. 1, pp. 11–24, Jan. 2012, ISSN: 0890-9369, 1549-5477. DOI: 10.1101/gad.179804.111. [Online]. Available: http://dx.doi.org/10.1101/gad.179804.111.

[9] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker, "Comprehensive mapping of long-range interactions reveals folding principles of the human genome", en, *Science*, vol. 326, no. 5950, pp. 289–293, Oct. 2009, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1181369. [Online]. Available: http://dx.doi.org/10.1126/science.1181369.

[10] C. W. Beitel, L. Froenicke, J. M. Lang, I. F. Korf, R. W. Michelmore, J. A. Eisen, and A. E. Darling, "Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products", en, *PeerJ*, vol. 2, no. 12, e415, May 2014, ISSN: 2167-8359. DOI: 10.7717/peerj.415. [Online]. Available: http://dx.doi.org/10.7717/peerj.415.

[11] J. N. Burton, I. Liachko, M. J. Dunham, and J. Shendure, "Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps", en, *G3*, vol. 4, no. 7, pp. 1339–1346, May 2014, ISSN: 2160-1836. DOI: 10.1534/g3.114.011825. [Online]. Available: http://dx.doi.org/10.1534/g3.114.011825.

[12] M. Marbouty, A. Cournac, J.-F. Flot, H. Marie-Nelly, J. Mozziconacci, and R. Koszul, "Metagenomic chromosome conformation capture (meta3c) unveils the diversity of chromosome organization in microorganisms", en, *Elife*, vol. 3, no. e03318, e03318, Dec. 2014, ISSN: 2050-084X. DOI: 10.7554/eLife.03318. [Online]. Available: http://dx.doi.org/10.7554/eLife.03318.

[13] S. Selvaraj, J. R Dixon, V. Bansal, and B. Ren, "Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing", en, *Nat. Biotechnol.*, vol. 31, no. 12, pp. 1111–1118, Dec. 2013, ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.2728. [Online]. Available: http://dx.doi.org/10.1038/nbt.2728.

[14] J. N. Burton, A. Adey, R. P. Patwardhan, R. Qiu, J. O. Kitzman, and J. Shendure, "Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions", en, *Nat. Biotechnol.*, vol. 31, no. 12, pp. 1119–1125, Dec. 2013, ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.2727. [Online]. Available: http://dx.doi.org/10.1038/nbt.2727.

[15] H. Marie-Nelly, M. Marbouty, A. Cournac, J.-F. Flot, G. Liti, D. P. Parodi, S. Syan, N. Guillén, A. Margeot, C. Zimmer, and R. Koszul, "High-quality genome (re)assembly using chromosomal contact data", en, *Nat. Commun.*, vol. 5, no. 5695, p. 5695, Dec.

2014, ISSN: 2041-1723. DOI: 10.1038/ncomms6695. [Online]. Available: http://dx.doi.org/10.1038/ncomms6695.

[16] J.-M. Belton, R. P. McCord, J. H. Gibcus, N. Naumova, Y. Zhan, and J. Dekker, “Hi–C: A comprehensive technique to capture the conformation of genomes”, *Methods*, vol. 58, no. 3, pp. 268–276, Nov. 2012, ISSN: 1046-2023. DOI: 10.1016/j.ymeth.2012.05.001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1046202312001168.

[17] Z. Duan, M. Andronescu, K. Schutz, S. McIlwain, Y. J. Kim, C. Lee, J. Shendure, S. Fields, C. A. Blau, and W. S. Noble, “A three-dimensional model of the yeast genome”, en, *Nature*, vol. 465, no. 7296, pp. 363–367, May 2010, ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature08973. [Online]. Available: http://dx.doi.org/10.1038/nature08973.

[18] E. Yaffe and A. Tanay, “Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture”, en, *Nat. Genet.*, vol. 43, no. 11, pp. 1059–1065, Oct. 2011, ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng.947. [Online]. Available: http://dx.doi.org/10.1038/ng.947.

[19] M. Imakaev, G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker, and L. A. Mirny, “Iterative correction of Hi-C data reveals hallmarks of chromosome organization”, en, *Nat. Methods*, vol. 9, no. 10, pp. 999–1003, Oct. 2012, ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.2148. [Online]. Available: http://dx.doi.org/10.1038/nmeth.2148.

[20] R. E. Boulos, A. Arneodo, P. Jensen, and B. Audit, “Revealing long-range interconnected hubs in human chromatin interaction data using graph theory”, en, *Phys. Rev. Lett.*, vol. 111, no. 11, p. 118 102, Sep. 2013, ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.111.118102. [Online]. Available: http://dx.doi.org/10.1103/PhysRevLett.111.118102.

[21] K. Jajuga, A. Sokolowski, and H.-H. Bock, *Classification, Clustering, and Data Analysis: Recent Advances and Applications*, en, ser. Recent Advances and Applications. Springer Science & Business Media, Dec. 2012, ISBN: 9783642561818. [Online]. Available: https://market.android.com/details?id=book-0YrsCAAAQBAJ.

[22] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo, “A comparison of extrinsic clustering evaluation metrics based on formal constraints”, *Inf. Retr. Boston.*, vol. 12, no. 4, pp. 461–486, Aug. 2009, ISSN: 1386-4564, 1573-7659. DOI: 10.1007/s10791-008-9066-8. [Online]. Available: https://doi.org/10.1007/s10791-008-9066-8.

[23] J. B. Hirschberg and A. Rosenberg, “V-Measure: A conditional entropy-based external cluster evaluation”, in *Empirical Methods for Natural Language Processing*, Proceedings of EMNLP, Jun. 2007, pp. 410–420. DOI: 10.7916/D80V8N84. [Online]. Available: http://hdl.handle.net/10022/AC:P:21139.

[24] A. Bagga and B. Baldwin, “Algorithms for scoring coreference chains”, in *The first international conference on language resources and evaluation workshop on linguistics coreference*, vol. 1, 1998, pp. 563–566. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.47.5848&rep=rep1&type=pdf.

[25] S. M. Van Dongen, “Graph clustering by flow simulation”, PhD thesis, 2000. [Online]. Available: https://dspace.library.uu.nl/handle/1874/848.

[26] Y.-K. Shih and S. Parthasarathy, “Identifying functional modules in interaction networks through overlapping markov clustering”, en, *Bioinformatics*, vol. 28, no. 18, pp. i473–i479, Sep. 2012, ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/bts370. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/bts370.

[27] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks”, en, *J. Stat. Mech.*, vol. 2008, no. 10, P10008, Oct. 2008, ISSN: 1742-5468. DOI: 10.1088/1742-5468/2008/10/P10008. [Online]. Available: http://iopscience.iop.org/article/10.1088/1742-5468/2008/10/P10008/meta.

[28] A. Pérez-Suárez, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, and J. E. Medina-Pagola, “OClustR: A new graph-based algorithm for overlapping clustering”, *Neurocomputing*, vol. 121, pp. 234–247, Dec. 2013, ISSN: 0925-2312. DOI: 10.1016/j.neucom.2013.04.025. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231213005432.

[29] M. E. J. Newman and M. Girvan, “Finding and evaluating community structure in networks”, en, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, vol. 69, no. 2 Pt 2, p. 026 113, Feb. 2004, ISSN: 1539-3755. DOI: 10.1103/PhysRevE.69.026113. [Online]. Available: http://dx.doi.org/10.1103/PhysRevE.69.026113.

[30] T. Aynaud, *Python-louvain: Louvain community detection*, Dec. 2018. [Online]. Available: https://github.com/taynaud/python-louvain.

[31] A. A. Hagberg, D. A. Schult, and P. J. Swart, “Exploring network structure, dynamics, and function using NetworkX)”, in *Proceedings of the 7th Python in Science Conference.*, Jan. 2008, pp. 11–15. [Online]. Available:

http://permalink.lanl.gov/object/tr?what=info:lanl-repo/lareport/LA-UR-08-05495.

[32] S. M. Kiełbasa, R. Wan, K. Sato, P. Horton, and M. C. Frith, “Adaptive seeds tame genomic sequence comparison”, en, *Genome Res.*, vol. 21, no. 3, pp. 487–493, Mar. 2011, ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.113985.110. [Online]. Available: http://dx.doi.org/10.1101/gr.113985.110.

[33] H. Li, “Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM”, Mar. 2013. arXiv: 1303.3997 [q-bio.GN]. [Online]. Available: http://arxiv.org/abs/1303.3997.

[34] D. Coil, G. Jospin, and A. E. Darling, “A5-miseq: An updated pipeline to assemble microbial genomes from illumina MiSeq data”, en, *Bioinformatics*, vol. 31, no. 4, pp. 587–589, Feb. 2015, ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btu661. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btu661.

[35] A. Darling, M. Craven, B. Mau, and N. T. Perna, “Multiple alignment of rearranged genomes”, in *Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004.*, IEEE, Aug. 2004, pp. 738–739, ISBN: 9780769521947. DOI: 10.1109/CSB.2004.1332564. [Online]. Available: http://dx.doi.org/10.1109/CSB.2004.1332564.

[36] W. Huang, L. Li, J. R. Myers, and G. T. Marth, “ART: A next-generation sequencing read simulator”, en, *Bioinformatics*, vol. 28, no. 4, pp. 593–594, Feb. 2012, ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btr708. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btr708.

[37] A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler, “QUAST: Quality assessment tool for genome assemblies”, en, *Bioinformatics*, vol. 29, no. 8, pp. 1072–1075, Apr. 2013, ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btt086. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btt086.

[38] Y. Peng, H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin, “IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth”, en, *Bioinformatics*, vol. 28, no. 11, pp. 1420–1428, Jun. 2012, ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/bts174. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/bts174.

[39] K. T. Konstantinidis, A. Ramette, and J. M. Tiedje, “The bacterial species definition in the genomic era”, en, *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, vol. 361, no. 1475, pp. 1929–1940, Nov. 2006, ISSN: 0962-8436. DOI: 10.1098/rstb.2006.1920. [Online]. Available: http://dx.doi.org/10.1098/rstb.2006.1920.

[40] M. Richter and R. Rosselló-Móra, “Shifting the genomic gold standard for the prokaryotic species definition”, en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 106, no. 45, pp. 19 126–19 131, Nov. 2009, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0906412106. [Online]. Available: http://dx.doi.org/10.1073/pnas.0906412106.

[41] M. Dehmer and A. Mowshowitz, “A history of graph entropy measures”, *Inf. Sci.*, vol. 181, no. 1, pp. 57–78, Jan. 2011, ISSN: 0020-0255. DOI: 10.1016/j.ins.2010.08.041. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020025510004147.

[42] A. Mowshowitz and M. Dehmer, “Entropy and the complexity of graphs revisited”, en, *Entropy*, vol. 14, no. 3, pp. 559–570, Mar. 2012. DOI: 10.3390/e14030559. [Online]. Available: http://www.mdpi.com/1099-4300/14/3/559.

[43] M. Dehmer, L. Sivakumar, and K. Varmuza, “Uniquely discriminating molecular structures using novel eigenvalue—based descriptors”, *Match-Communications in Mathematical and Computer Chemistry*, vol. 67, no. 1, p. 147, 2012. [Online]. Available: http://match.pmf.kg.ac.rs/electronic_versions/Match67/n1/match67n1_147-172.pdf.

[44] S. Lê, J. Josse, and F. Husson, “FactoMineR: An R package for multivariate analysis”, *J. Stat. Softw.*, vol. 25, no. 1, pp. 1–18, Mar. 2008. DOI: 10.18637/jss.v025.i01. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.422.7829&rep=rep1&type=pdf.

[45] M. Liu and A. Darling, “Metagenomic chromosome conformation capture (3c): Techniques, applications, and challenges”, en, *F1000Res.*, vol. 4, no. 1377, p. 1377, Nov. 2015, ISSN: 2046-1402. DOI: 10.12688/f1000research.7281.1. [Online]. Available: http://dx.doi.org/10.12688/f1000research.7281.1.

[46] A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev, and P. A. Pevzner, “SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing”, en, *J. Comput. Biol.*, vol. 19, no. 5, pp. 455–477, May 2012, ISSN: 1066-5277, 1557-8666. DOI: 10.1089/cmb.2012.0021. [Online]. Available: http://dx.doi.org/10.1089/cmb.2012.0021.

[47] 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”, en, *Bioinformatics*, vol. 31, no. 10, pp. 1674–1676, May 2015, ISSN:

1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btv033. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btv033.

[48] Z. Iqbal, I. Turner, and G. McVean, “High-throughput microbial population genomics using the cortex variation assembler”, en, *Bioinformatics*, vol. 29, no. 2, pp. 275–276, Jan. 2013, ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/bts673. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/bts673.

[49] T. J. Treangen, S. Koren, D. D. Sommer, B. Liu, I. Astrovskaya, B. Ondov, A. E. Darling, A. M. Phillippy, and M. Pop, “MetAMOS: A modular and open source metagenomic assembly and analysis pipeline”, en, *Genome Biol.*, vol. 14, no. 1, R2, Jan. 2013, ISSN: 1465-6906. DOI: 10.1186/gb-2013-14-1-r2. [Online]. Available: http://dx.doi.org/10.1186/gb-2013-14-1-r2.

[50] A. Lancichinetti, S. Fortunato, and J. Kertész, “Detecting the overlapping and hierarchical community structure in complex networks”, en, *New J. Phys.*, vol. 11, no. 3, p. 033 015, Mar. 2009, ISSN: 1367-2630. DOI: 10.1088/1367-2630/11/3/033015. [Online]. Available: http://iopscience.iop.org/article/10.1088/1367-2630/11/3/033015/meta.

[51] J. Xie, B. K. Szymanski, and X. Liu, “SLPA: Uncovering overlapping communities in social networks via a Speaker-Listener interaction dynamic process”, in *2011 IEEE 11th International Conference on Data Mining Workshops*, Dec. 2011, pp. 344–349. DOI: 10.1109/ICDMW.2011.154. [Online]. Available: http://dx.doi.org/10.1109/ICDMW.2011.154.

[52] J. Xie, S. Kelley, and B. K. Szymanski, “Overlapping community detection in networks: The state-of-the-art and comparative study”, *ACM Computing Surveys (CSUR)*, vol. 45, no. 4, p. 43, Aug. 2013, ISSN: 0360-0300. DOI: 10.1145/2501654.2501657. [Online]. Available: https://dl.acm.org/citation.cfm?doid=2501654.2501657.

[53] D. Jurgens and I. Klapaftis, “Semeval-2013 task 13: Word sense induction for graded and non-graded senses”, in *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, vol. 2, 2013, pp. 290–299. [Online]. Available: http://www.aclweb.org/website/old_anthology/S/S13/S13-2.pdf#page=326.

[54] A. Lancichinetti, S. Fortunato, and F. Radicchi, “Benchmark graphs for testing community detection algorithms”, en, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, vol. 78, no. 4 Pt 2, p. 046 110, Oct. 2008, ISSN: 1539-3755. DOI: 10.1103/PhysRevE.78.046110. [Online]. Available: http://dx.doi.org/10.1103/PhysRevE.78.046110.

[55] A. Lancichinetti and S. Fortunato, “Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities”, en, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, vol. 80, no. 1 Pt 2, p. 016 118, Jul. 2009, ISSN: 1539-3755. DOI: 10.1103/PhysRevE.80.016118. [Online]. Available: http://dx.doi.org/10.1103/PhysRevE.80.016118.

[56] W. Chen, Z. Liu, X. Sun, and Y. Wang, “A game-theoretic framework to identify overlapping communities in social networks”, *Data Min. Knowl. Discov.*, vol. 21, no. 2, pp. 224–240, Sep. 2010, ISSN: 1384-5810, 1573-756X. DOI: 10.1007/s10618-010-0186-6. [Online]. Available: https://doi.org/10.1007/s10618-010-0186-6.

[57] C. Gaiteri, M. Chen, B. Szymanski, K. Kuzmin, J. Xie, C. Lee, T. Blanche, E. Chaibub Neto, S.-C. Huang, T. Grabowski, T. Madhyastha, and V. Komashko, “Identifying robust communities and multi-community nodes by combining top-down and bottom-up approaches to clustering”, en, *Sci. Rep.*, vol. 5, p. 16 361, Nov. 2015, ISSN: 2045-2322. DOI: 10.1038/srep16361. [Online]. Available: http://dx.doi.org/10.1038/srep16361.

[58] M. Burgess, E. Adar, and M. Cafarella, “Link-Prediction enhanced consensus clustering for complex networks”, en, *PLoS One*, vol. 11, no. 5, e0153384, May 2016, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0153384. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0153384.

2.12 Appendices

			Star	**Ladder**		
$\boldsymbol{\alpha}_{\mathbf{BL}}$	**ANI**	**1 - ANI**	$\mathbf{d}_{*,*}$	$\mathbf{d}_{\mathbf{A},[\mathbf{B}\vert\mathbf{C}\vert\mathbf{D}]}$	$\mathbf{d}_{\mathbf{B},[\mathbf{C}\vert\mathbf{D}]}$	$\mathbf{d}_{\mathbf{C},\mathbf{D}}$
0.0250	0.9950	0.0050	0.0050	0.0075	0.0050	0.0025
0.0377	0.9930	0.0070	0.0075	0.0113	0.0075	0.0038
0.0568	0.9890	0.0110	0.0114	0.0171	0.0114	0.0057
0.0855	0.9830	0.0170	0.0171	0.0257	0.0171	0.0086
0.1288	0.9750	0.0250	0.0258	0.0387	0.0258	0.0129
0.1941	0.9630	0.0370	0.0388	0.0582	0.0388	0.0194
0.2924	0.9460	0.0540	0.0585	0.0877	0.0585	0.0292
0.4405	0.9230	0.0770	0.0881	0.1322	0.0881	0.0441
0.6637	0.8900	0.1100	0.1327	0.1991	0.1327	0.0664
1.0000	0.8470	0.1530	0.2000	0.3000	0.2000	0.1000

Table 2.S1: The relationship between the scale factor (α_{BL}) and the phylogenetic distance (d) between taxa in the resulting tree. Additionally for the star we report ANI as calculated using BLAST alignments. As the Star is isotropic, only a single distance is reported; while for the Ladder all three distinct distance combinations are reported. The two underlined columns indicate the taxon pairs for which the Star and Ladder have equivalent evolutionary divergence

CHAPTER 3

sim3C: simulation of Hi-C and Meta3C proximity ligation sequencing technologies

3.1 Authorship Declaration

By signing below I confirm that for the paper titled “sim3C: simulation of Hi-C and Meta3C proximity ligation sequencing technologies” and published by *GigaScience*, that:

Matthew Z. DeMaere designed and implemented sim3C and wrote the manuscript and prepared figures.

Aaron E. Darling assisted in the design and contributed to the manuscript.

Production Note:
Signature removed
prior to publication.

Matthew Z. DeMaere

Production Note:
Signature removed
prior to publication.

Associate Professor Aaron E. Darling

3.2 Abstract

Background Chromosome conformation capture (3C) and Hi-C DNA sequencing methods have rapidly advanced our understanding of the spatial organization of genomes and metagenomes. Many variants of these protocols have been developed, each with their own strengths. Currently there is no systematic means for simulating sequence data from this family of sequencing protocols, potentially hindering the advancement of algorithms to exploit this new datatype.

Findings We describe a computational simulator that, given simple parameters and reference genome sequences, will simulate Hi-C sequencing on those sequences. The simulator models the basic spatial structure in genomes that is commonly observed in Hi-C and 3C datasets, including the distance-decay relationship in proximity ligation, differences in the frequency of interaction within and across chromosomes, and the structure imposed by cells. A means to model the 3D structure of randomly generated topologically associating domains (TADs) is provided. The simulator considers several sources of error common to 3C and Hi-C library preparation and sequencing methods, including spurious proximity ligation events and sequencing error.

Conclusions We have introduced the first comprehensive simulator for 3C and Hi-C sequencing protocols. We expect the simulator to have use in testing of Hi-C data analysis algorithms, as well as more general value for experimental design, where questions such as the required depth of sequencing, enzyme choice, and other decisions can be made in advance in order to ensure adequate statistical power with respect to experimental hypothesis testing.

Keywords Hi-C, Meta3C, 3C, DNA sequencing, simulation, metagenomics

3.3 Findings

3.3.1 Software testing

To the casual observer, formal software testing is often thought to begin and end with the validation of fine-grained behavioural (functional) aspects; such as the correct execution of individual methods. In day to day use however, what can matter most to end-users are broader system attributes such as speed, scalability, reproducibility and ease of use. To ensure a project offers maximum value, a thorough testing process would collectively examine all aspects.

For inferential software within scientific fields, the system-level attributes of precision and accuracy are of primary interest, and their quantification is best accomplished by comparison to a known truth (gold standard). Therefore, any testing methodology capable of providing an *a priori* gold standard, particularly without estimation, improves this facet of testing significantly.

Purpose-built bioinformatics software ultimately acts on experimentally collected observations. The inherent noise and variation that comes with experimental data means achieving testing thoroughness is a great challenge. Ready access to sufficient data sources is a fundamental necessity for adequate software testing.

For established experimental methods, public data archives are a first choice for the necessary testing data. When high quality metadata is available, testing driven by real data becomes possible. However, even when sufficient depth and description of data is available, difficulty can remain in matching desired test data characteristics to what actually exists in one or several public dataset(s). Further, fine-grained whole-corpus querying of metadata on remote data archives is not always possible, frequently making the up-front job of data selection a difficult task. Once selected, obtaining said real data can be time-consuming or even infeasible in locations with lower network speeds and/or high bandwidth costs. In advancing fields such as DNA sequencing, new experimental datatypes can appear for which the public data archives contain only a handful of examples and few researchers would have the time and financial resources to commit to experimental generation of new data purely for software testing.

Though performance on real data is the ultimate arbiter of analytical value, advantaged by explicit control over its characteristics, a faithful simulation of real data can act as a valuable proxy. Simulation-driven development and testing has proven to be a highly cost effective and time efficient approach. It offers the possibility to explore a near continuum of data characteristics, subjecting software to an otherwise unavailable degree of testing thoroughness. Certainty and control makes attaining the twin objectives of rigorous testing and an *a priori* gold standard straightforward. This enables us not only to be more certain about when we have failed, but also to extrapolate this process to infer the limits of success within the experimental parameter space.

Tools for simulating DNA sequencing reads have existed from the very early days of genomics, beginning with the many anonymous implementations of simple DNA shearing algorithms, up to the most recent highly detailed empirical model simulators [1]–[4]. From read simulation in isolation, field advancements such as metagenomics have been accompanied soon after by simulators reflecting their specific data characteristics and evolving experimental methodology [5]–[7].

We introduce sim3C, a software package designed to simulate data generated by Hi-C and other 3C-based proximity ligation (PL) sequencing protocols. The software includes flexible support for a range of sequencing project scenarios and choice of three 3C methods Hi-C, Meta3C, DNase Hi-C). The resulting output (paired-end FastQ) is easily assimilated into existing analysis workflows. It is our intention that sim3C provide the Hi-C/3C research community with means to further validate existing software projects, to support new experimental or analysis development initiatives and as a platform for exploration, such as the comparative analysis of clustering algorithms [8].

3.3.2 3C sequencing

3C-based sequencing protocols, including Hi-C, 4C-seq, and Meta3C, have great potential to address questions directed at the spatial organization of DNA in samples ranging from eukaryotic tissue, to single cells, to microbial communities. The growing use of these protocols creates a legitimate need for a simulator capable of generating data with relevant characteristics.

Chromosome conformation capture (3C) was originally designed as a PCR-based assay to measure interactions among a small number of defined regions of eukaryotic chromosomes [9]. In 2009 Lieberman-Aiden [10] reported an extension of the protocol to high throughput sequencing, enabling the global spatial arrangement of chromosomes to be reconstructed at unprecedented resolution. All 3C protocols depend on an initial formalin fixation step, which crosslinks proteins bound to DNA in vivo. Subsequently cells are lysed and the DNA:protein complexes are sheared enzymatically and/or physically to create free ends in the bound DNA strands. These free ends are then subjected to a proximity ligation reaction, in which ligation of free ends preferentially occurs among DNA strands cobound in a protein complex. The DNA:protein crosslinks are then reversed, the DNA is purified, and an Illumina-compatible sequencing library is constructed. In Hi-C protocols, the proximity ligation junctions can then be further purified in the sequencing library.

3C-derived methods have found several applications beyond their initial use to reconstruct 3D chromosome structure. For example, it has been shown that 3C-derived data provide a valuable signal for genome scaffolding [11], [12], as well as a signal that can support genome-wide haplotype phasing [13], [14]. 3C-derived data has also proven valuable for metagenomics, where initial studies on mock communities demonstrated that highly accurate genome reconstruction in mixed microbial communities could be facilitated by proximity ligation sequence data [15]–[17]. Subsequent application to naturally occurring microbial communities has also suggested that bacteriophage can be linked to their hosts with this data type [18].

In the remainder of this manuscript we describe the sim3C software and demonstrate how it can be used to simulate data for various 3C-derived experiments.

3.3.3 Experiment scenarios

Beyond simple monochromosomal genome sequencing experiments, sim3C offers support for the more complex scenarios of multi-chromosomal genomes and metagenomes. A scenario is defined by way of a community profile; assigning a copy-number and containing genome to each chromosome and a relative abundance to each genome. The profile and supporting reference sequences form a skeleton definition with which to initialize the weighted random sampling process within a simulation. The user can elect to supply a profile either as an explicit table (Listing 3.1, 3.2) or allow sim3C to draw abundances at runtime from one of three distributions (equal abundance, uniformly random, log-normal distribution) for communities made up of strictly mono-chromosomal genomes.

```
#chrom   cell   abund   copynum
chr1     bac1   0.4     1
plas1    bac1   0.4     1
chr2     bac2   0.6     1
```

Listing 3.1: **A mock two genome community.** For demonstration purposes, we assume that the plasmid (plas1) is present in four copies and that there is a 0.4/0.6 relative abundance split between the two organisms (bac1, bac2) in the community

```
#chrom   cell   abund   copynum
chr1     euk1   1       1
chr2     euk1   1       1
chr3     euk1   1       1
chr4     euk1   1       2
```

Listing 3.2: **A mock four chromosome genome**. Cellular abundance is a constant across the profile, while chr4 exists in two copies. Note that relative abundances specified in a profile are not required to sum to 1, but are normalised internally.

3.3.4 Error Modelling

Sim3C models three forms of experimental noise: machine-based sequencing error, the formation of spurious ligation products and the contamination of PL libraries with WGS read-pairs.

To simulate machine-based sequencing error, the paired-end mode from art_illumina [2] has been reimplemented as a Python module (`Art.py`). This approach was taken as delegating read-pair generation to native invocations of art_illumina proved cumbersome. More explicitly, a loosely coupled solution (via subprocess calls but without an IPC mechanism) lacked sufficient control to generate PL read-pairs in an efficient and robust manner. On the other hand, tightly coupling sim3C to the ART C/C++ source code (i.e. implementing hooks) would have left sim3C vulnerable to changes in a non-public external API (i.e. a codebase without formal definition or guarantee of stability). Reimplementation also meant Art's many empirically derived machine profiles are available for use by sim3C, allowing equivalent treatment of machine-error when experiments involve both PL (sim3C) and pure WGS (art_illumina) libraries.

The production of spurious ligation products is an inherent source of noise in PL library construction [19]. Sim3C models spurious pairs as the uniformly random ligation of any two cut-sites across all source genomes. While this process disregards cellular organisation, it respects the relative abundance of chromosomes. Spurious pairs, and to a lesser extent sequencing error, represent an important confounding signal to downstream analyses that attempt to infer the cellular or chromosomal organisation of DNA sequences.

Lastly, conventional WGS read-pairs represent a source of contamination within a PL library, which even after Hi-C enrichment steps, are not completely eliminated. The rates at which spurious and WGS read-pairs are injected into a simulation run are controllable by the end-user.

3.3.5 Simulation modes

Since Hi-C was first introduced [10], the development of variants and extensions has been continual [17], [20]–[22]. Variants have often strived to further enhance the discriminatory power of the original experiment, while seemingly adding yet more complexity to an already challenging protocol (*in-situ* DNase Hi-C, sciHi-C) [22]. Others instead have sought compromise, with the aim of lessening the burden on the laboratory (Meta3C). While not considering more recent and complex extensions, sim3C offers three simulation modes: traditional Hi-C, Meta3C and DNase Hi-C. The first two of these modes were chosen as representing the fundamental basis (traditional Hi-C) and an attractive and pragmatic simplification of the original (Meta3C). The third mode (DNase Hi-C) replaces the restriction endonuclease driven production of the free-ends, used to form PL products, with an ideally-free process of DNA fragmentation. In the laboratory, this ideally-free process could be carried out by DNase digestion or mechanical shearing via sonication.

The most notable difference between the methods of Hi-C and the more recent Meta3C, is that after restriction digest,Hi-C employs additional steps leading to the incorporation of biotin tags at each PL junction. This biotinylation permits Hi-C libraries to be subsequently enriched for fragments containing PL junctions by streptavidin-mediated affinity purification. Without enrichment, the simpler Meta3C protocol results in a gross mixture of both WGS and PL read-pairs, where only a small percentage of the total read-pair yield (approx. 1%) will possess

PL junctions [23]. The enrichment process within Hi-C, however, is not perfectly efficient and WGS read-pairs are still observed (approx. 10-50% of reads contain a PL product) [23]. DNase Hi-C replaces restriction digest with a non-specific endonuclease (e.g. DNase I) [24] or mechanical DNA shearing process (e.g. sonication) [20]. In this operational mode, sim3C treats DNA cleavage as a completely unbiased (free) process and as such all genomic positions have equal probability of participating in proximity ligation events.

Within sim3C, each of the three methodological variations is conceptualised as a sequencing strategy (Figure 3.1) and each iteration of a strategy produces one read-pair (PL or WGS in origin). For all strategies, an iteration begins by drawing a 3-tuple of insert parameters: length, direction and junction point (L_{ins}, dir, x_{junc}).

After obtaining insert parameters, the Hi-C strategy (Figure 3.1a) first tests if the insert will represent a WGS or PL read-pair $(\sim Bern(p_{eff}))$, where efficiency p_{eff} is defined in the sense of enrichment. When $p_{eff} = 1$, there is perfect filtering and all WGS read-pairs are eliminated from the experiment. In the case of WGS, the iteration reaches an end-point and the simulation emits a conventional read-pair drawn from the community definition. In the case of PL, a cut-site 3-tuple is drawn (gen_1, chr_1, x_1), where the categorical distribution over chromosomes is weighted by relative abundances (A) and chromosomal copy-numbers (n_{cpy}); genomic position is sampled uniformly from the set of restriction sites $(sites(chr_1))$; and parent genome (gen_1) is implicit from the chromosome. Next, a spurious ligation test is performed $(\sim Bern(p_{spur}))$. If a spurious event has occurred, the 3-tuple defining the second cut-site (gen_2, chr_2, x_2) is drawn i.i.d. as the first. If not spurious, next a test for inter-chromosomal (*trans*) ligation is performed. Only source chromosome and position (chr_2, x_2) need be drawn as the second genome is implicitly the same as the first $(gen_2 = gen_1)$. Here, chr_2 is selected without replacement from the set of chromosomes of genome (gen_1), where the categorical distribution is adjusted by removal of chr_1. Finally, an intra-chromosomal (*cis*) ligation must have occurred. As now both genome and chromosome are implicit $(gen_2 = gen_1, chr_2 = chr_1)$, all that is left is to draw genomic position x_2. The pair of positions (x_1, x_2) are constrained by their separation $(s = |x_2 - x_1|)$, which is represented by a mixture model of the geometric and uniform distributions (Equation 3.1). This relation possesses rapid falloff with increasing separation and non-zero probability for all chromosomal positions, as has been commonly observed in real experimental data [10], [25].

$$Pr(X = s|\alpha, \beta, l) = \beta(1-\alpha)^s\alpha + (1-\beta)/l \tag{3.1}$$

where β is a mixing parameter, α the geometric distribution shape parameter and l chromosome length.

For Meta3C (Figure 3.1b) after insert parameters are determined, in the same fashion as a regular WGS read, an initial free genomic position is drawn (chr_1, x_1^*), uniformly distributed over the extent of chr_1 rather than only over its cut-sites. In real datasets, it has been observed that neither the restriction digestion nor the re-ligation of free ends are perfectly efficient. Taken as independent probabilities, in our model we conceptualise their joint occurrence as an efficiency factor, p_{eff} and a Bernoulli trial $(Bern(p_{eff}))$ determines whether a sequence read is successful in containing an observable proximity ligation event. Failing this coverage test relegates the iteration and end-point and emit a WGS read-pair. Successful candidates instead continue akin to the Hi-C decision tree, beginning with the test for spurious ligation.

For both Hi-C and Meta3C, PL read-pairs are produced by joining the free-ends drawn above as defined by the fragment parameters (Figure 3.2a). Here the location of the PL junction within the insert is determined by x_{junc}. At the junction,Hi-C differs from Meta3C as the process of biotinylation results in the duplication of the restriction cut-site overhang sequence. The overhang duplication in Hi-C is included in the simulation.

DNase Hi-C is handled similarly to traditional Hi-C, with the exception that, as *in-silico* digestion trivially leads to all sites, the simulated digestion is unnecessary to perform and positions can be drawn directly from the uniform distribution over the interval $[0..L_{chr})$. Site duplication, attributable to the likely production of random overhangs in this scenario, is not presently simulated.

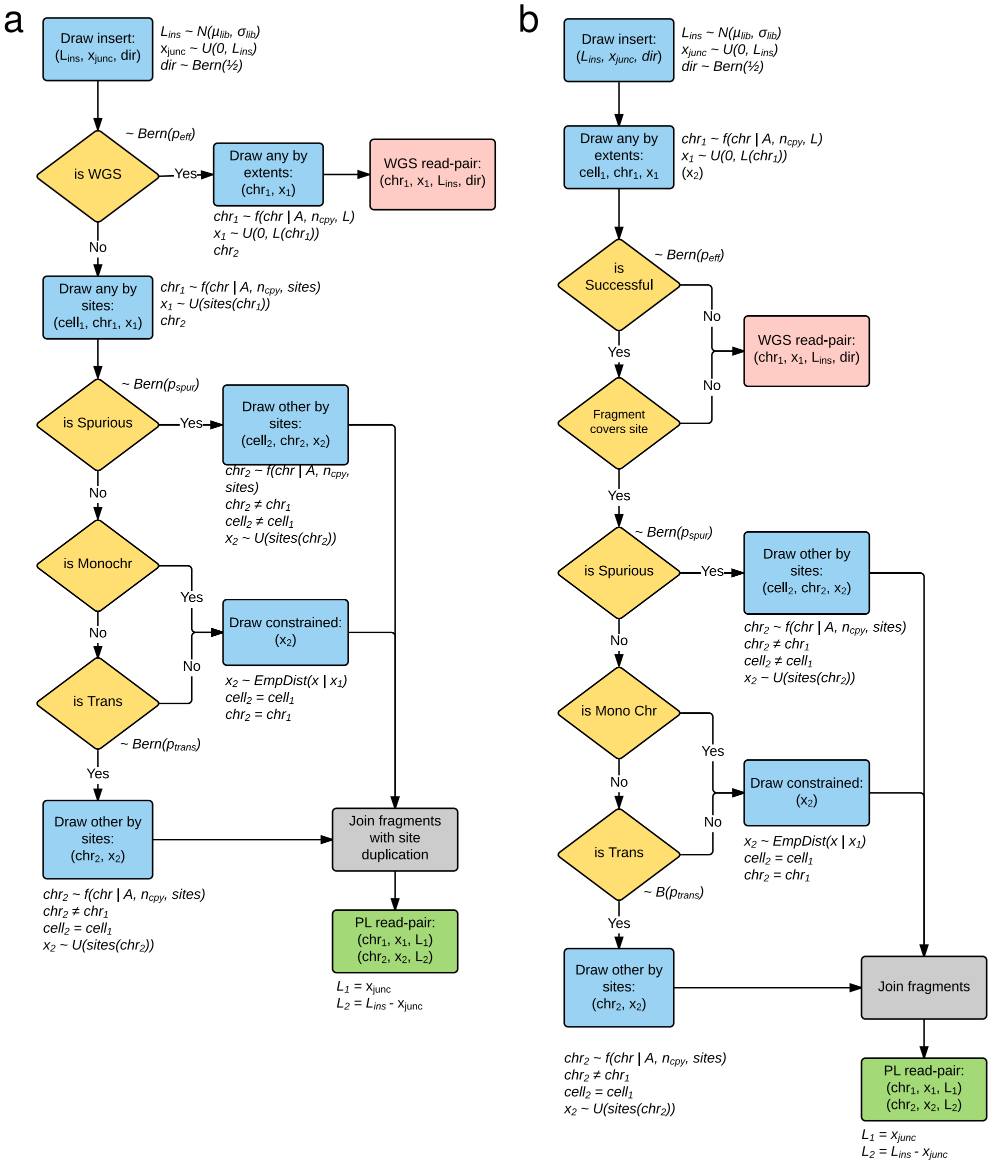
a
Draw insert: (L_{ins}, x_{junc}, dir)
$L_{ins} \sim N(\mu_{lib}, \sigma_{lib})$
$x_{junc} \sim U(0, L_{ins})$
$dir \sim Bern(½)$
$\sim Bern(p_{eff})$
is WGS
Yes
Draw any by extents: (chr_1, x_1)
WGS read-pair: (chr_1, x_1, L_{ins}, dir)
$chr_1 \sim f(chr \mid A, n_{cpy}, L)$
$x_1 \sim U(0, L(chr_1))$
chr_2
No
Draw any by sites: ($cell_1$, chr_1, x_1)
$chr_1 \sim f(chr \mid A, n_{cpy}, sites)$
$x_1 \sim U(sites(chr_1))$
chr_2
$\sim Bern(p_{spur})$
is Spurious
Yes
Draw other by sites: ($cell_2$, chr_2, x_2)
$chr_2 \sim f(chr \mid A, n_{cpy}, sites)$
$chr_2 \neq chr_1$
$cell_2 \neq cell_1$
$x_2 \sim U(sites(chr_2))$
No
is Monochr
Yes
Draw constrained: (x_2)
$x_2 \sim EmpDist(x \mid x_1)$
$cell_2 = cell_1$
$chr_2 = chr_1$
No
No
is Trans
$\sim Bern(p_{trans})$
Yes
Draw other by sites: (chr_2, x_2)
$chr_2 \sim f(chr \mid A, n_{cpy}, sites)$
$chr_2 \neq chr_1$
$cell_2 = cell_1$
$x_2 \sim U(sites(chr_2))$
Join fragments with site duplication
PL read-pair: (chr_1, x_1, L_1) (chr_2, x_2, L_2)
$L_1 = x_{junc}$
$L_2 = L_{ins} - x_{junc}$
b
Draw insert: (L_{ins}, x_{junc}, dir)
$L_{ins} \sim N(\mu_{lib}, \sigma_{lib})$
$x_{junc} \sim U(0, L_{ins})$
$dir \sim Bern(½)$
Draw any by extents: $cell_1$, chr_1, x_1
$chr_1 \sim f(chr \mid A, n_{cpy}, L)$
$x_1 \sim U(0, L(chr_1))$
(x_2)
$\sim Bern(p_{eff})$
is Successful
No
WGS read-pair: (chr_1, x_1, L_{ins}, dir)
Yes
No
Fragment covers site
Yes
$\sim Bern(p_{spur})$
is Spurious
Yes
Draw other by sites: ($cell_2$, chr_2, x_2)
$chr_2 \sim f(chr \mid A, n_{cpy}, sites)$
$chr_2 \neq chr_1$
$cell_2 \neq cell_1$
$x_2 \sim U(sites(chr_2))$
No
is Mono Chr
Yes
Draw constrained: (x_2)
$x_2 \sim EmpDist(x \mid x_1)$
$cell_2 = cell_1$
$chr_2 = chr_1$
No
No
is Trans
$\sim B(p_{trans})$
Yes
Draw other by sites: (chr_2, x_2)
$chr_2 \sim f(chr \mid A, n_{cpy}, sites)$
$chr_2 \neq chr_1$
$cell_2 = cell_1$
$x_2 \sim U(sites(chr_2))$
Join fragments
PL read-pair: (chr_1, x_1, L_1) (chr_2, x_2, L_2)
$L_1 = x_{junc}$
$L_2 = L_{ins} - x_{junc}$

Figure 3.1: **Logical schema used within sim3C**. (**a**)Hi-C and (**b**) Meta3C simulation strategies. Gold diamonds represent simple Bernoulli trials. Blue boxes represent sampling distributions defined by runtime input data (community profile, genomic sequences, enzyme) and the empirically derived distribution for intra-chromosome (*cis*) interaction probability (equation 1). Logical end-points to a single iteration of either algorithm are represented as red (producing a WGS read-pair) and green boxes (producing a PL read-pair). Due to the elimination of the biotinylation step, Meta3C does not produce a duplication of the restriction cut-site overhang (grey boxes).

3.3.6 Structurally related interactions

Independent of any 3D structure that might exist, the primary and most frequently observed interactions are those which occur along a chromosome (intra-arm) (Figure 3.2b), seen as the primary $(y \simeq x)$ diagonal in the contact map. sim3C can approximate the less frequent interactions occurring between chromosomal arms (inter-arm) [26], which are visible as anti-diagonal $(y \simeq L - x)$ in the contact map.

At progressively smaller scales, the hierarchical 3D folding of DNA into topologically associated domains (TADs) produces overlapping regions of interaction visible in the contact map as block-like intensity modulations. Though the agents responsible for their formation vary [27], [28], the characteristic patterns evident in real-data derived 3C contact maps have been observed across all three domains [25], [26], [29]. Sim3C can optionally approximate the sense of TAD related modulation by means of a recursive stochastic process.

Our approximation of hierarchical folding begins from the full extent L of a chromosome (Figure 3.2c). Folding is portrayed by the division of the interval $[0..L)$ into a set of non-overlapping sub-intervals $\{[0, x_1), [x_1, x_2), \cdots, [x_{n-1}, x_n)\}$, the number and widths of which are drawn at random $(U(l_{min}, l_{max}), U(n_{min}, n_{max}))$. The procedure is then recursively applied to each sub-interval until a depth d, producing a nested set of coverings of the full interval $[0..L)$ at progressively finer scales. Across this hierarchical collection each interval is assigned a uniformly distributed random probability p_i and empirical distribution $f_i(s|\theta_i)$ (equation 1) for separation s parameterised by shape parameter α_{TAD} and interval length $l_{inv} = x_{i+1} - x_i$, where $\theta = (\alpha_{TAD}, \beta, l_{inv})$.

The process of drawing samples of separation begins by determining the set of intervals $\{l_{inv}\}$ which contain an initial point x_0. The intervals, as tuples $(p_i, f_i(s|\theta_i))$, then form a categorical distribution (equation (3.7)), from which a governing distribution $f_i(s|\theta_i)$ is drawn and finally a sample of separation is taken, $s \sim f_i(s|\theta_i)$. To efficiently sample from the full collection, an interval-tree data structure is employed. When queried, an interval-tree returns the set of intervals $\{l\}$ overlapping a position x in order $O(\log n + m)$, where n is number of intervals and m is number of intervals returned by the query.

$$\mathbf{f} = \{f_0(s|\theta_0), f_1(s|\theta_1), \cdots, f_i(s|\theta_i)\} \tag{3.2}$$

$$N = \text{number of distributions} = |\mathbf{f}| \tag{3.3}$$

$$\mathbf{p} = \{p_0, p_1, \cdots, p_i\} \tag{3.4}$$

$$p_i \sim U(0,1) \text{ and } \sum p_i = 1 \tag{3.5}$$

$$n \sim Cat(N, \mathbf{p}) \tag{3.6}$$

$$f(s|n) = \prod_{i=0}^{N-1} f_i(s|\theta_i)^{[i=n]} \tag{3.7}$$

where $[i = n]$ is the Iverson bracket.

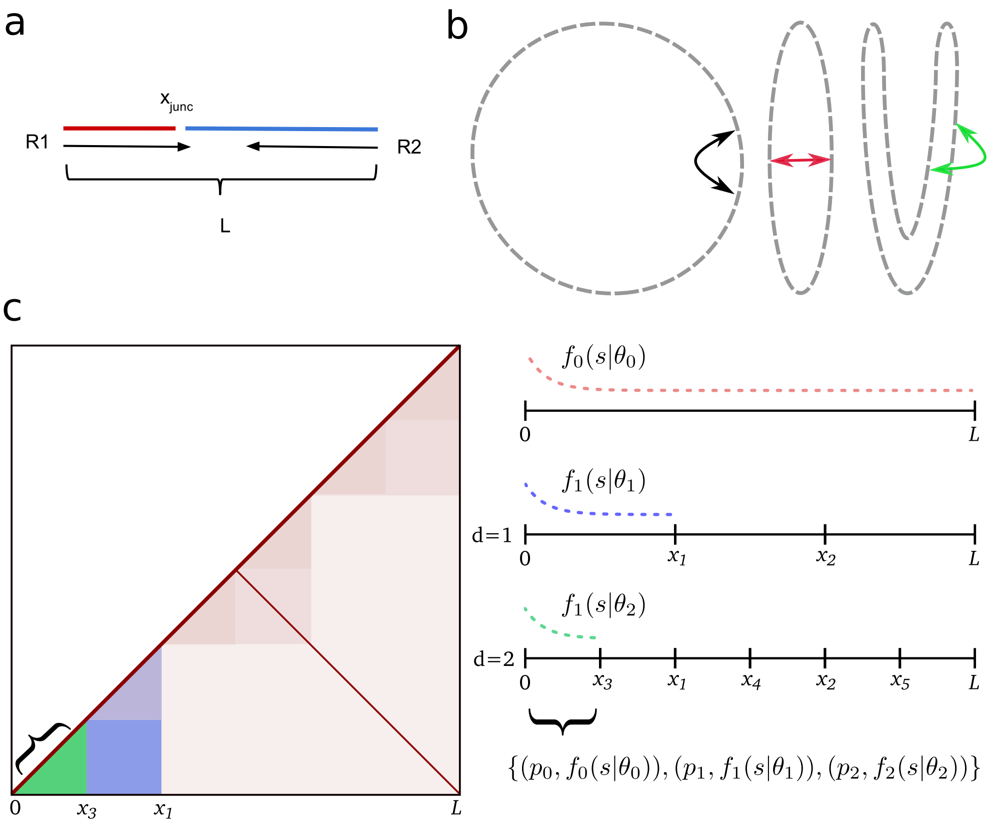

Figure 3.2: **Model details**. Generation of proximity ligation inserts (**a**) involves joining two randomly drawn parts (red and blue), from which the read-pair (R1, R2) is then simulated. The junction point (x_{junc}) varies over the interval $[0..L)$ and reproduction of read-through events is possible. For an unbounded chromosome (**b**) (circular here), besides strictly primary separation (black arrow) spatial proximity can be induced from successive folding (red, green arrows). When the spatial arrangement is consistent across the population of cells, this will be observable as modulations in the contact frequencies. Sim3C models simple structurally related modulation of observed contact frequencies (**c**). Beyond primary interactions forming the main diagonal, users can reproduce inter-arm mediated anti-diagonals. Finer scale modulations attributed to topologically associated domains (TADs) can optionally be randomly simulated. Primary interactions $f_0(s|\theta_0)$ (Equation 3.1) cover the full interval $[0, L)$. Each level of recursion $(d = 1, 2 \cdots n)$ generates a finer set of intervals, to which a distribution $f_i(s|\theta_i)$ and probability p_i is assigned. The final covering of intervals each define a range (green, curly braces) over which a set of probabilities and empirical distribution pairs govern interaction separation s.

3.3.7 Example scenarios

In the following, three use-cases are presented to demonstrate aspects of the resulting simulation output: bacterial genome, multi-chromosomal eukaryotic (yeast) genome, and metagenome. For each use-case, 3C contact maps have been used to pit simulation output against the corresponding real experimental data (Table 3.1).

3.3.8 Bacterial

A monochromosomal bacterial genome is perhaps the simplest scenario to which proximity ligation methods have been applied, making for a sensible entry point from which to make comparison. Due to the smaller extent, a bright and high resolution contact map (10 kbp bin size) is possible for a practical volume of sequencing data, potentially revealing fine detail not easily discerned with larger bin sizes (50-100 kbp bin size).

The genome of *Caulobacter crescentus* NA1000, a model organism in the study of cellular differentiation and regulation of the cell cycle, is comprised of a single 4 Mbp circular chromosome [30]. Deep Hi-C sequencing of *C. crescentus* has been used to explore the degree to which bacterial chromosomes can be regarded as organised and provided evidence for the existence of so called chromosomal interaction domains (CIDs) [26]. As a prokaryotic analog of topologically associated domains (TADs) from eukaryotic literature [28], [31], [32], these regions are believed to promote intra-domain loci interactions and thereby act to functionally compartmentalize the genome. The chromosomal structure was observed to have boundaries defined, at least in part, by highly expressed genes and these boundaries were found to disruptable through rifampicin mediated inhibition of transcription [26].

For the raw contact map of *C. crescentus*, prominent rectilinear features are apparent for both real and simulated traditional Hi-C sequencing data (Figure 3.3a,b), while notably for simulated unrestricted Hi-C the field is much smoother (Figure 3.3c). Within the sim3C model, a single distribution governs both intra- and inter-arm interactions. Inspection of the real-data contact map (Figure 3.3a) suggests that the true relationship governing inter-arm interactions is more dispersed. This perhaps is not surprising, where different arms associating spatially possess a greater number of potential configurations than can be taken on by the primary chromosome backbone. Additionally for the real contact map, long-range interactions away from either diagonal can be seen to drop to a lower threshold than that produced from simulation.

Within the unrestricted Hi-C map, the fine zero-intensity rectilinear features are a direct result of poor mappability (non-unique sequence), where their small size reflects the extent of the non-unique regions (example: rRNA genes) and the single base-pair resolution of the less constrained read generation process. The process of enzymatic digestion is the only difference between the unrestricted and traditional Hi-C simulation models. The clear contrast in their contact maps is thus a combination of factors either directly inherent to digestion (cut-site density) or a byproduct of downstream bioinformatics analysis (e.g. filtering heuristics). Though the problem of mappability exists for any reference based representation, for real and simulated traditional Hi-C, zero-intensity rectilinear features mark regions devoid of cut-sites over at least 10 kbp.

Enabling TAD approximation in simulated traditional Hi-C (Figure 3.3d) has the effect of modulating map intensity in a manner not particularly distinct from that produced purely from experimental/workflow bias. Discriminating between these two feature sources; one representing experimental signal, the other representing noise; demands attention when developing solutions to problems such as normalisation. Contact map normalisation methods, whether based upon explicit or implicit bias models [33], may leave behind remnants of noise-related features from either a lack of convergence or model limitations. Downstream inferencing should therefore not be made under an assumption of bias-free signal.

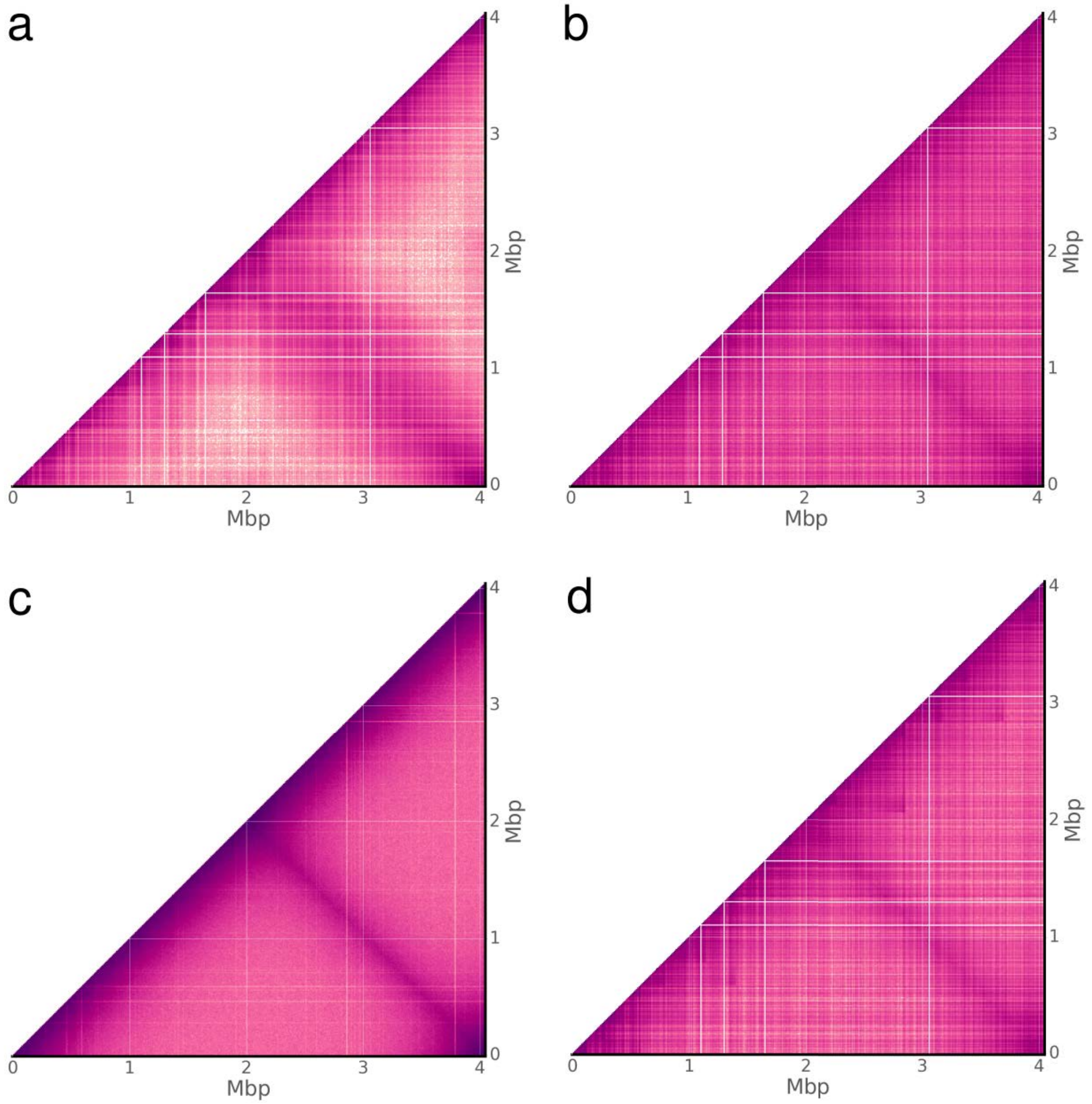

Figure 3.3: **Bacterial contact maps**. Observed Hi-C interactions for the monochromosomal genome of *Caulobacter crescentus* NA1000. Comparing (**a**) real experimental data [26], to the three simulation choices (**b**) traditional Hi-C, (**c**) DNase Hi-C and (**d**) traditional Hi-C with TADs enabled. Sharp rectilinear modulations of the intensity within (**a**) and (**b**) indicate a reduction in PL observations within a given bin. Not due to 3D chromosome structure, rather such features can be attributed largely to mappability and low cut-site density. (**c**) Without an enzymatic constraint a significantly smoother field is apparent, yet still susceptible to mappability. (**d**) Enabling topologically associated domains (TADs) highlights the similarity between features produced merely from biases and what could be truly associated with 3D structure.

3.3.9 Eukaryotic

The eight chromosomes of the 15.4 Mbp genome of the native xylose-fermenting yeast *Scheffersomyces stipitis* CBS 6054 [34] range in size from 970 kbp to 3.5 Mbp. The organism was one of 16 yeasts included in a synthetic community to explore the application of Hi-C sequencing to deconvolving metagenomic assemblies [16] and is divergent enough from other synthetic community members to permit unambiguous read mapping, and thus act as a proxy for a clonal experiment.

From the contact map of real Hi-C data (Figure 3.4a), it can be seen that the rates of intra-chromosomal and inter-chromosomal interactions are roughly equivalent in magnitude. Across the eight chromosomes of *S. stipitis*, there is significant uniformity in the degree of physical intimacy within and between all chromosomes. The subtleties of this chromosomal organisation reveals a self-similar “fuzzy-x” pattern repeated between all chromosomes across the contact map. The convergence point within the pattern is attributed to centromere-SPB binding and has been used to predict centromere locations [35]. It has been shown that the physical constraints generated from the interaction of centromeres to the spindle pole body (SPB) and telomeres to the nuclear envelope are sufficient to explain a number of experimental observations in real data [36], [37]. As sim3C was derived from study of bacterial datasets, our simulation model does not currently include a notion of these higher organism physical constraints. Consequently, the contact map derived from simulated traditional Hi-C sequencing elicits a flat field (Figure 3.4b), where the intensity variation that does exist is a byproduct of aforementioned factors such as mappability and cut-site density. For the runtime parameters employed, the rate of intra-chromosomal contact is higher than that of inter-chromosomal, making clear the boundaries between the eight chromosomes (Figure 3.4b). Though our model is presently incomplete for higher organisms, there remains a potential utility as an analytical or simply observational prior.

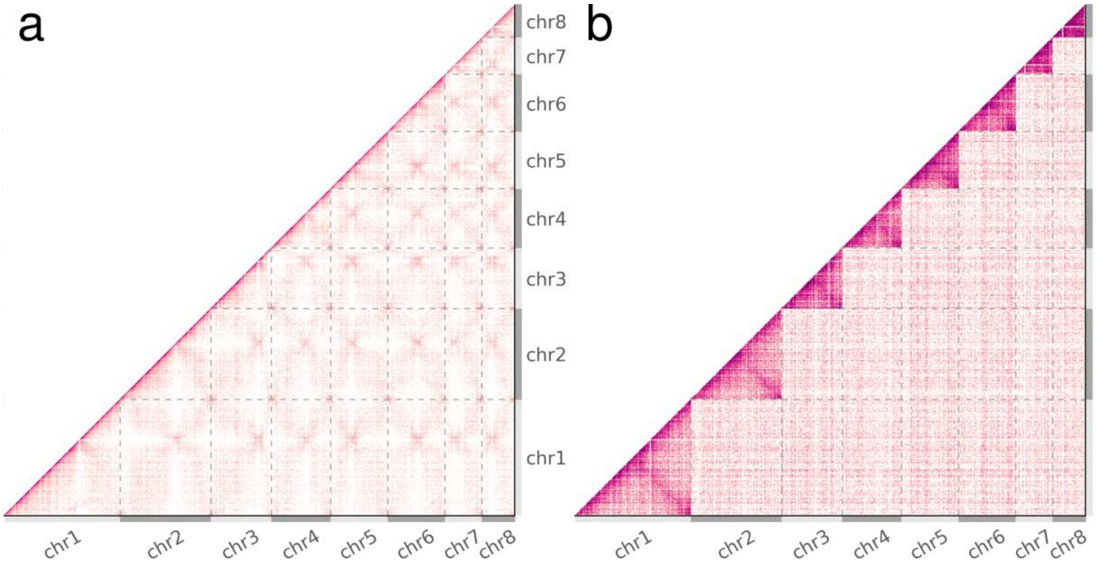

Figure 3.4: **Eukaryotic contact maps**. Observed Hi-C interactions (**a**) real and (**b**) simulated data from the eight chromosome genome of the budding yeast *Scheffersomyces stipitis* CBS 6054 [16]. Grey dashed lines and alternating light and dark grey axes demarcate the boundaries between chromosomes. (**b**) Simulated data elicits a flat field and the clearly evident higher rate of intra- to inter- interactions makes for easily observable chromosomal boundaries within the map. (**a**) Contrastingly for real data, the similar rates of intra-chr and inter-chr interactions reveals the physical constraints imposed by centromere-SPB tethering on all eight chromosomes [35].

3.3.10 Metagenomic

In the deconvolution of metagenomes, proximity ligation methods hold great potential as new sources of information and have been investigated by the construction and sequencing of synthetic communities [15]–[17]. We selected two previously constructed synthetic bacterial communities, one employing traditional Hi-C and the other Meta3C (Table 3.1). Intended as "proof of concept" experiments, neither community reflects a real environment, but rather were intended to be easily interpreted and include interesting features, such as: range of GC, single and multi- chromosomal genomes and strain-level divergence. The Hi-C community involved five genotypes from four species, one genome of two chromosomes (*B. thailandensis*), *E. coli* strains BL21 and K12 (Average Nucleotide Identity, ANI 99%) and a wide overall GC range of 37-68% (Table 3.2). Of lower complexity, the Meta3C community involved three genomes from three species, included one genome of two chromosomes (*V. cholerae*) and had a narrower GC range of 44-51% (Table 3.3).

Relative to the single genome experiments above, a lower depth of sequencing resulted in a lower overall contact map intensity (Figure 3.5). This is particularly the case for Meta3C, where, by the nature of the method, a large proportion (approx. 99%) of the sequencing yield is in reality conventional WGS read-pair data [17]. As a direct result, in binning the Meta3C dataset, there were insufficient counts to fully establish finer detail within the contact maps, leaving a smoother appearance.

As with single-genome experiments, metagenomic contact maps are locally modulated by factors such as mappability and cut-site density. Importantly now for metagenomes, the factors of relative abundance and GC content interact to alter the observed intensity of each chromosome within the contact map.

As a first approximation and assuming agreement in nucleotide sampling frequency, we expect $n_0 = L/4^{\lambda}$ recognition sites for an enzyme of site length λ and DNA sequence length L. The degree to which an enzyme and DNA sequence deviate from this estimate could be described as how well they match, $m = n_x/n_0$. Poorer quality matches $(m < 1)$ occur when an enzyme's recognition site is underrepresented, while conversely, better quality matches $(m > 1)$ describe a situation of more recognition sites than expected.

When multiple chromosomes are taken as a community, the relative proportion of sites from each represents an observational bias when conducting 3C-based experiments. For community C, the number of sites n_x from chromosome x determines the number of potential PL pairings N_x within C which involve x (Equation 3.8). The number of intra-chromosomal and inter-chromosomal potential pairs thus respectively vary quadratically and linearly with n_x. Regarding the process of observing a PL event (read-pair) from the community as a random draw with replacement, and the selection pool as comprised of all potential events from all chromosomes, then variation in match quality constitutes a per-chromosome bias. In real laboratory experiments, the composition of the selection pool is further modified by variation in other factors, such as cellular lysis efficiency, unintended DNA fragmentation and relative abundance. In particular, when relative abundances A are introduced, the odds of observing a PL event involving chromosome x is then proportional the product $p_x \propto A_x N_x / N_C$. Although the processes of intra-chromosomal, inter-chromosomal, and inter-cellular (spurious) ligation are treated independently in our simulation model, in this manner, per-chromosome intensity (observation rate of chromosome x) can vary significantly within a metagenome.

$$N_x = n_x^2 + n_x \sum_{n_y \in C \setminus n_x} n_y \tag{3.8}$$

Though the original laboratory experiments reported by Beitel et al. 2014 and Marbouty et al. 2014 intended to create synthetic communities with uniform relative abundances, in practice each possesses a non-uniform profile. The variation in GC content is largest for the Hi-C experiment and together with non-uniform relative abundances produces a wide range of chromosome intensity for both real and simulated data (Figure 3.5a,b). For both the real and simulated Hi-C maps, the frequent observation of PL events involving *P. pentosaceus* (Pp) and *L. brevis* (Lb), suggests the possibility that inter-cellular interaction is significant. Within the simulated map at least, inter-cellular pairs are produced exclusively through the process of spurious ligation (noise) and are observed at a higher rate than in the real data, indicating that as expected, spurious ligation rates across species are correlated with their relative abundances.

Further for the Hi-C data, the two-chromosome genome of *B. thailandensis* (Bt1, Bt2) (Figure 3.5a) has a greater rate of inter-chromosomal interaction than expected from comparing it to simulation (Figure 3.5b). Meanwhile, the clear delineation of *E. coli* strains BL21 and K12 ($ANI > 99$%), with little inter-cellular signal, helps to support the notion that the inter-chromosomal interactions observed between *B. thailandensis* chromosomes ($ANI \simeq 83$%) are real and not a by-product of inadequate filtering.

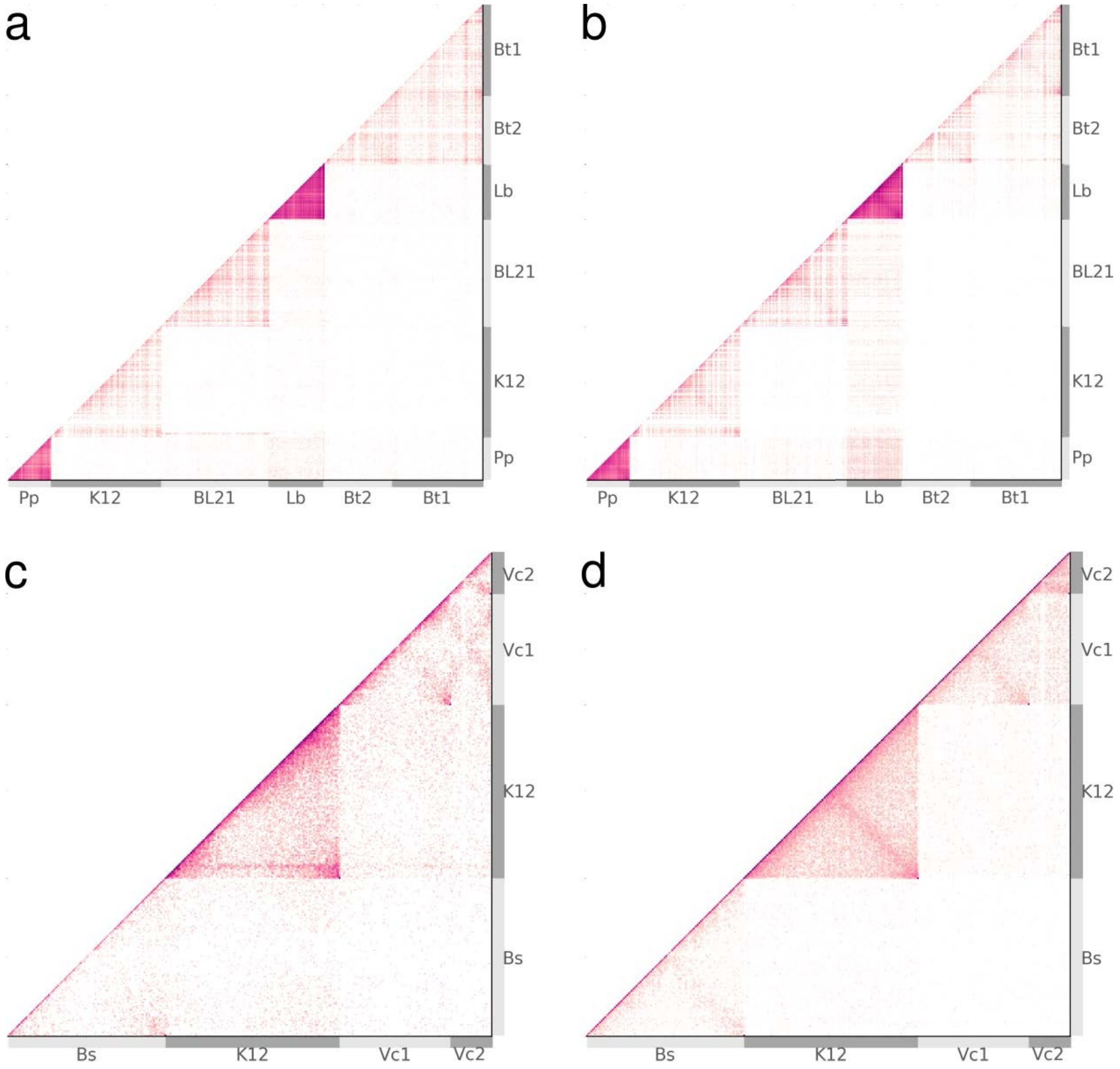

Figure 3.5: **Metagenomic contact maps**. From synthetic microbial communities, raw contact maps from real (**a**) and simulated (**b**) traditional Hi-C, and real (**c**) and simulated (**d**) Meta3C. Chromosome boundaries are demarcated by alternating light and dark grey bands (Table 3.2, 3.3), while the small plasmids of *L. brevis* are omitted for clarity. Although the original works [15], [17] intended uniform abundance, the results exhibit significant variation in abundance. Lysis efficiency (not modelled) and enzyme suitability are significant factors contributing to the overall intensity of a given chromosome. For more abundant members of the Hi-C community (*P. pentosaceus* and *L. brevis*), signal due only to spurious ligation can appear to suggest inter-cellular interactions when none are present (**b**).

3.3.11 Limitations and future work

Sim3C in its current form has several limitations, some of which present opportunities for future work. Sim3C's repertoire of structural features is currently limited to those found in microbes - circular and linear chromosomes with randomly generated approximations of self-associating domains (CIDs/TADs). Sim3C does not model structural features observed in larger, more complex genomes (CTCF/cohesin loops, A/B compartments, chromosome territories) [10], [38]. Such features are becoming increasingly well characterised [39] and a simulator capable of modelling these features would surely be valuable. Mammalian genomes are much larger than microbial genomes however, and additional work to improve scalability of sim3C will likely be required.

Some features of microbial eukaryotes, such as the point centromeres found in budding yeast genomes [40] are computationally simpler [35], [36] yet remain unmodelled in sim3C. The addition of these sorts of model details would be best supported by introducing model initialisation via external data (experimental observations, motif detection, cell phase), which subsequently would require extension of the community profile definition. Careful design would be required to ensure these features could be added without compromising ease-of-use.

3.4 Methods

3.4.1 Reference Data

To compare sim3C against real experiments, we obtained previously published experimental read-pair datasets (Table 3.1) and their accompanying reference genomes (Table 3.2, 3.3) from public archives. In the case of the single genome project of *Caulobacter crescentus* CB15 [26], sequencing data derived from untreated swarmer cells was chosen and the laboratory strain *C. crescentus* NA1000 (acc: NC_011916) was used as the reference genome. For the yeast genome, the completed eight chromosome genome of *Scheffersomyces stipitis* CBS 6054 was used as a reference (acc: PRJNA18881) and the respective reads were extracted from the MY16 yeast synthetic metagenome [16] by direct mapping with BWA MEM. Extraction by mapping in isolation was employed as *S. stipitis* was the second furthest phylogenetically removed yeast in the synthetic community and was the most contiguous (N50: 60kbp) from the whole synthetic community de novo metagenomic WGS assembly.

Authors	Type	Method	Accession	Sequencing details	Mapped reads
Beitel et al[15]	Synthetic bacterial metagenome	Hi-C	SRX377733	MiSeq 160bp PE insert range: 280-420bp enzyme: HindIII	20552775
Burton et al[16]	Synthetic yeast metagenome	Hi-C	SRX527868	HiSeq2500 100bp PE insert range: 450-550bp enzyme: HindIII	9704944
Le et al[26]	Single bacterial genome	Hi-C	SRX263925	HiSeq2000 40bp PE insert range: 200-600bp enzyme: NcoI	22324360
Marbouty et al[41]	Synthetic bacterial metagenome	Meta3C	doi:10.5061/dryad.gv595	HiSeq2000 100bp PE insert range: 400-800bp enzyme: HpaII	7975740

Table 3.1: **Real Hi-C and Meta3C data-sets used within this work.** The total off-diagonal weight of the contact map was used to calibrate the amount of simulated sequencing required to approximately match the outcome of the real experiments.

3.4.2 Read Generation

Experimental parameters used in read simulation were set to agree as closely as reasonably possible to the respective real experiments, employing the same read length and restriction enzyme (Table 3.1). In each experiment, the published fragment size range was approximated by a normal distribution (Table 3.4). For ease of reproducibility, a single random seed (1234) was used in all simulations. As our intent was primarily to demonstrate functionality, rates of inter-chromosomal and spurious events were adjusted per-experiment only through a qualitative process. For simulation of metagenomic datasets, relative abundances were estimated by mapping real experimental reads to the respective reference genomes. From each real experiment, the off-diagonal weight of the resulting contact map was used to calibrate the amount of simulated sequencing required to achieve roughly equivalent intensity (Table 3.4). Both real and simulated read-pair datasets were mapped to their respective reference genomes using BWA MEM (v0.7.15-r1140, RRID:SCR_010910) [42]

Name	**Replicons**	**Accession**	**Chr abbr.**	A	n_{cpy}	**%GC**	n_x	m
Burkholderia thailandensis E264	2	NC_007651	Bt1	0.054	1	67.29	225	0.24
		NC_007650	Bt2			68.07	144	0.20
Escherichia coli BL21	1	NC_012892	BL21	0.242	1	50.83	508	0.46
Escherichia coli K12 DH10B	1	NC_010473	K12	0.166	1	50.78	568	0.50
Lactobacillus brevis ATCC 367	3	NC_008497	Lb	0.436	1	46.22	629	1.12
		NC_008498	-			38.64	3	0.92
		NC_008499	-			38.51	16	1.84
Pediococcus pentosaceus ATCC 25745	1	NC_008525	Pp	0.102	1	37.36	863	1.93

Table 3.2: **Synthetic Hi-C community.** A synthetic community used to demonstrate the utility of Hi-C sequencing data in resolving a microbial metagenome [15]. It is composed of 5 bacteria, including two closely related strains (*E. coli* K12 and BL21), a genome with two plasmids (*L. brevis*) and a two-chromosome genome (*B. thailandensis*). A is relative abundance, n_{cpy} is copy number, n_x is number of restriction sites, and $m = n_x/n_0$ is match quality between chromosome and enzyme choice: $m < 1$ is worse, $m > 1$ is better.

Name	**Replicons**	**Accession**	**Chr abbr.**	A	n_{cpy}	**%GC**	n_x	m
Bacillus subtilis subsp. subtilis str. 168	1	NC_000964	Bs	0.123	1	43.51	14529	0.88
Escherichia coli str. K-12 substr. MG1655	1	NC_000913	K12	0.562	1	50.79	24311	1.34
Vibrio cholerae O1 biovar El Tor str. N16961	2	NC_002505	Vc1	0.332	1	47.70	5909	0.51
		NC_002506	Vc2			46.91	1802	0.43

Table 3.3: **Synthetic Meta3C community.** A synthetic community used to demonstrate the utility of Meta3C sequencing data in resolving a microbial metagenome [17], [41]. It is composed of three bacteria with one possessing two chromosomes. A is relative abundance, n_{cpy} is copy number, n_x is number of restriction sites, and $m = n_x/n_0$ is match quality between chromosome and enzyme choice: $m < 1$ is worse, $m > 1$ is better.

Experiment	**Insert μ** (bp)	**Insert σ** (bp)	**Anti** **rate**	**Spurious** **rate**	**Trans** **rate**	**Reads** ($\times 10^6$)
Beitel et al	300	50	0.2	0.05	0.1	7
Burton et al	400	50	0.2	0.5	0.15	1.5
Le et al	400	100	0.2	0.2	0.1	22
Marbouty et al	600	100	0.2	0.2	0.2	7.5

Table 3.4: **Runtime simulation.** Parameters supplied to sim3C during read generation.

3.4.3 Contact Maps

Contact maps were produced using our own tool (`contact_map.py`), where heatmap intensity was plotted as log-scaled observational frequency. To reduce the potential for spurious assignment, aligned reads were subject to the same basic filtering criteria: BWA MEM mapq > 5 and alignment length $\geq 50\%$ of read length, with the added restriction that read alignments must have begun with a match. For methods which employed a restriction enzyme (traditional Hi-C, Meta3C), we constrained the maximum allowable distance from an aligned read to the nearest upstream cut-site. Calculated per chromosome, this distance constraint could not exceed two-fold the median cut-site spacing. Rather than simply delete the primary diagonal for the sake of reducing the displayed dynamic range in figures, we instead to reduced its intensity by categorizing properly paired reads with an estimated fragment size of less than 2 of the reported mean as being conventional WGS (non-PL) reads and ignored them. The resolution of contact maps was adjusted between experiments so as to present a sufficiently bright image without undue loss of resolution. The contact map bin sizes employed were: 10000 bp for the single bacterial genome, 25000 bp for the yeast genome and 40000 bp for the Hi-C and Meta3C metagenomes (Table 3.2, 3.3).

3.5 Availability of data and materials

Snapshots of the supporting code are available from the GigaScience repository, GigaDB [43].

3.6 Availability of supporting source code and requirements

- Project name: sim3C
- Release version: 0.1
- Project homepage: `https://github.com/cerebis/sim3C`
- RRID: SCR_015772
- DOI: `https://doi.org/10.5281/zenodo.1030812`
- Operating system: Platform independent
- Programming languages: Python 2.7
- License: GNU GPL v3

3.7 List of abbreviations

- IPC - interprocess communication
- PL - proximity ligation
- WGS - whole genome shotgun
- CID - chromosomal interaction domain
- TAD - topologically associated domain
- $Bern(x)$ - Bernoulli distribution
- $U(x)$ - uniform distribution
- $N(\mu, \sigma)$ - normal distribution
- *cis* - intra-chromosomal
- *trans* - inter-chromosomal

3.8 Declarations

3.8.1 Funding

This work was supported under Australian Research Council's Discovery Projects funding scheme (project number: LP150100912, CI: Djordjevic, SP). The NeCTAR Research Cloud is an Australian Government project conducted as part of the Super Science initiative and financed by the and the Education Investment Fund (EIF) and National Collaborative Research Infrastructure Strategy (NCRIS).

3.8.2 Authors contributions

MD designed and implemented sim3C and wrote the manuscript and prepared figures. AD assisted in the design and contributed to the manuscript.

3.9 Acknowledgements

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, the QCIF and the UTS eResearch Group.

3.10 References

[1] H. Li, *Lh3/wgsim*, https://github.com/lh3/wgsim, Accessed: 2017-3-21, Oct. 2011. [Online]. Available: https://github.com/lh3/wgsim.

[2] W. Huang, L. Li, J. R. Myers, and G. T. Marth, “ART: A next-generation sequencing read simulator”, en, *Bioinformatics*, vol. 28, no. 4, pp. 593–594, Feb. 2012, ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btr708. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btr708.

[3] Y. Ono, K. Asai, and M. Hamada, “PBSIM: PacBio reads simulator–toward accurate genome assembly”, en, *Bioinformatics*, vol. 29, no. 1, pp. 119–121, Jan. 2013, ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/bts649. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/bts649.

[4] X. Hu, J. Yuan, Y. Shi, J. Lu, B. Liu, Z. Li, Y. Chen, D. Mu, H. Zhang, N. Li, Z. Yue, F. Bai, H. Li, and W. Fan, “pIRS: Profile-based illumina pair-end reads simulator”, en, *Bioinformatics*, vol. 28, no. 11, pp. 1533–1535, Jun. 2012, ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/bts187. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/bts187.

[5] B. Jia, L. Xuan, K. Cai, Z. Hu, L. Ma, and C. Wei, “NeSSM: A next-generation sequencing simulator for metagenomics”, en, *PLoS One*, vol. 8, no. 10, e75448, Oct. 2013, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0075448. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0075448.

[6] F. E. Angly, D. Willner, F. Rohwer, P. Hugenholtz, and G. W. Tyson, “Grinder: A versatile amplicon and shotgun sequence simulator”, en, *Nucleic Acids Res.*, vol. 40, no. 12, e94, Jul. 2012, ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gks251. [Online]. Available: http://dx.doi.org/10.1093/nar/gks251.

[7] D. C. Richter, F. Ott, A. F. Auch, R. Schmid, and D. H. Huson, “MetaSim: A sequencing simulator for genomics and metagenomics”, en, *PLoS One*, vol. 3, no. 10, e3373, Oct. 2008, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0003373. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0003373.

[8] M. Z. DeMaere and A. E. Darling, “Deconvoluting simulated metagenomes: The performance of hard- and soft- clustering algorithms applied to metagenomic chromosome conformation capture (3c)”, en, *PeerJ*, vol. 4, e2676, Nov. 2016, ISSN: 2167-8359. DOI: 10.7717/peerj.2676. [Online]. Available: http://dx.doi.org/10.7717/peerj.2676.

[9] J. Dekker, K. Rippe, M. Dekker, and N. Kleckner, “Capturing chromosome conformation”, en, *Science*, vol. 295, no. 5558, pp. 1306–1311, Feb. 2002, ISSN:

0036-8075, 1095-9203. DOI: 10.1126/science.1067799. [Online]. Available: http://dx.doi.org/10.1126/science.1067799.

[10] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker, "Comprehensive mapping of long-range interactions reveals folding principles of the human genome", en, *Science*, vol. 326, no. 5950, pp. 289–293, Oct. 2009, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1181369. [Online]. Available: http://dx.doi.org/10.1126/science.1181369.

[11] J. N. Burton, A. Adey, R. P. Patwardhan, R. Qiu, J. O. Kitzman, and J. Shendure, "Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions", en, *Nat. Biotechnol.*, vol. 31, no. 12, pp. 1119–1125, Dec. 2013, ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.2727. [Online]. Available: http://dx.doi.org/10.1038/nbt.2727.

[12] O. Dudchenko, S. S. Batra, A. D. Omer, S. K. Nyquist, M. Hoeger, N. C. Durand, M. S. Shamim, I. Machol, E. S. Lander, A. P. Aiden, and E. L. Aiden, "De novo assembly of the aedes aegypti genome using Hi-C yields chromosome-length scaffolds", en, *Science*, vol. 356, no. 6333, pp. 92–95, Apr. 2017, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aal3327. [Online]. Available: http://dx.doi.org/10.1126/science.aal3327.

[13] S. Selvaraj, J. R Dixon, V. Bansal, and B. Ren, "Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing", en, *Nat. Biotechnol.*, vol. 31, no. 12, pp. 1111–1118, Dec. 2013, ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.2728. [Online]. Available: http://dx.doi.org/10.1038/nbt.2728.

[14] J. O. Korbel and C. Lee, "Genome assembly and haplotyping with Hi-C", en, *Nat. Biotechnol.*, vol. 31, no. 12, pp. 1099–1101, Dec. 2013, ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.2764. [Online]. Available: http://dx.doi.org/10.1038/nbt.2764.

[15] C. W. Beitel, L. Froenicke, J. M. Lang, I. F. Korf, R. W. Michelmore, J. A. Eisen, and A. E. Darling, "Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products", en, *PeerJ*, vol. 2, no. 12, e415, May 2014, ISSN: 2167-8359. DOI: 10.7717/peerj.415. [Online]. Available: http://dx.doi.org/10.7717/peerj.415.

[16] J. N. Burton, I. Liachko, M. J. Dunham, and J. Shendure, "Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability

maps”, en, *G3*, vol. 4, no. 7, pp. 1339–1346, May 2014, ISSN: 2160-1836. DOI: 10.1534/g3.114.011825. [Online]. Available: http://dx.doi.org/10.1534/g3.114.011825.

[17] M. Marbouty, A. Cournac, J.-F. Flot, H. Marie-Nelly, J. Mozziconacci, and R. Koszul, “Metagenomic chromosome conformation capture (meta3c) unveils the diversity of chromosome organization in microorganisms”, en, *Elife*, vol. 3, no. e03318, e03318, Dec. 2014, ISSN: 2050-084X. DOI: 10.7554/eLife.03318. [Online]. Available: http://dx.doi.org/10.7554/eLife.03318.

[18] M. Marbouty, L. Baudry, A. Cournac, and R. Koszul, “Scaffolding bacterial genomes and probing host-virus interactions in gut microbiome by proximity ligation (chromosome capture) assay”, en, *Sci Adv*, vol. 3, no. 2, e1602105, Feb. 2017, ISSN: 2375-2548. DOI: 10.1126/sciadv.1602105. [Online]. Available: http://dx.doi.org/10.1126/sciadv.1602105.

[19] T. Nagano, C. Várnai, S. Schoenfelder, B.-M. Javierre, S. W. Wingett, and P. Fraser, “Comparison of Hi-C results using in-solution versus in-nucleus ligation”, en, *Genome Biol.*, vol. 16, p. 175, Aug. 2015, ISSN: 1465-6906. DOI: 10.1186/s13059-015-0753-7. [Online]. Available: http://dx.doi.org/10.1186/s13059-015-0753-7.

[20] P. Y. H. Huang, Y. Han, L. Handoko, S. Velkov, E. Wong, E. Cheung, X. Ruan, C.-L. Wei, M. J. Fullwood, and Y. Ruan, “Protocol: Sonication-based circular chromosome conformation capture with next-generation sequencing analysis for the detection of chromatin interactions”, *Protocol Exchange*, Dec. 2010. DOI: 10.1038/protex.2010.207. [Online]. Available: http://dx.doi.org/10.1038/protex.2010.207.

[21] V. Ramani, D. A. Cusanovich, R. J. Hause, W. Ma, R. Qiu, X. Deng, C. A. Blau, C. M. Disteche, W. S. Noble, J. Shendure, and Z. Duan, “Mapping 3D genome architecture through in situ DNase Hi-C”, en, *Nat. Protoc.*, vol. 11, no. 11, pp. 2104–2121, Nov. 2016, ISSN: 1754-2189, 1750-2799. DOI: 10.1038/nprot.2016.126. [Online]. Available: http://dx.doi.org/10.1038/nprot.2016.126.

[22] V. Ramani, X. Deng, R. Qiu, K. L. Gunderson, F. J. Steemers, C. M. Disteche, W. S. Noble, Z. Duan, and J. Shendure, “Massively multiplex single-cell Hi-C”, en, *Nat. Methods*, vol. 14, no. 3, pp. 263–266, Mar. 2017, ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.4155. [Online]. Available: http://dx.doi.org/10.1038/nmeth.4155.

[23] M. Liu and A. Darling, “Metagenomic chromosome conformation capture (3c): Techniques, applications, and challenges”, en, *F1000Res.*, vol. 4, no. 1377, p. 1377,

Nov. 2015, ISSN: 2046-1402. DOI: 10.12688/f1000research.7281.1. [Online]. Available: http://dx.doi.org/10.12688/f1000research.7281.1.

[24] W. Ma, F. Ay, C. Lee, G. Gulsoy, X. Deng, S. Cook, J. Hesson, C. Cavanaugh, C. B. Ware, A. Krumm, J. Shendure, C. A. Blau, C. M. Disteche, W. S. Noble, and Z. Duan, "Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes", en, *Nat. Methods*, vol. 12, no. 1, pp. 71–78, Jan. 2015, ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.3205. [Online]. Available: http://dx.doi.org/10.1038/nmeth.3205.

[25] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren, "Topological domains in mammalian genomes identified by analysis of chromatin interactions", en, *Nature*, vol. 485, no. 7398, pp. 376–380, Apr. 2012, ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature11082. [Online]. Available: http://dx.doi.org/10.1038/nature11082.

[26] T. B. K. Le, M. V. Imakaev, L. A. Mirny, and M. T. Laub, "High-resolution mapping of the spatial organization of a bacterial chromosome", en, *Science*, vol. 342, no. 6159, pp. 731–734, Nov. 2013, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1242059. [Online]. Available: http://dx.doi.org/10.1126/science.1242059.

[27] A. Badrinarayanan, T. B. K. Le, and M. T. Laub, "Bacterial chromosome organization and segregation", en, *Annu. Rev. Cell Dev. Biol.*, vol. 31, no. 1, pp. 171–199, 2015, ISSN: 1081-0706, 1530-8995. DOI: 10.1146/annurev-cellbio-100814-125211. [Online]. Available: http://dx.doi.org/10.1146/annurev-cellbio-100814-125211.

[28] R. D. Acemel, I. Maeso, and J. L. Gómez-Skarmeta, "Topologically associated domains: A successful scaffold for the evolution of gene regulation in animals", en, *Wiley Interdiscip. Rev. Dev. Biol.*, vol. 6, no. 3, May 2017, ISSN: 1759-7684, 1759-7692. DOI: 10.1002/wdev.265. [Online]. Available: http://dx.doi.org/10.1002/wdev.265.

[29] T. Sexton, E. Yaffe, E. Kenigsberg, F. Bantignies, B. Leblanc, M. Hoichman, H. Parrinello, A. Tanay, and G. Cavalli, "Three-dimensional folding and functional organization principles of the drosophila genome", en, *Cell*, vol. 148, no. 3, pp. 458–472, Feb. 2012, ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2012.01.010. [Online]. Available: http://dx.doi.org/10.1016/j.cell.2012.01.010.

[30] M. E. Marks, C. M. Castro-Rojas, C. Teiling, L. Du, V. Kapatral, T. L. Walunas, and S. Crosson, "The genetic basis of laboratory adaptation in caulobacter crescentus", en, *J. Bacteriol.*, vol. 192, no. 14, pp. 3678–3688, Jul. 2010, ISSN: 0021-9193, 1098-5530.

DOI: 10.1128/JB.00255-10. [Online]. Available: http://dx.doi.org/10.1128/JB.00255-10.

[31] E. P. Nora, B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N. L. van Berkum, J. Meisig, J. Sedat, J. Gribnau, E. Barillot, N. Blüthgen, J. Dekker, and E. Heard, "Spatial partitioning of the regulatory landscape of the x-inactivation centre", en, *Nature*, vol. 485, no. 7398, pp. 381–385, Apr. 2012, ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature11049. [Online]. Available: http://dx.doi.org/10.1038/nature11049.

[32] B. D. Pope, T. Ryba, V. Dileep, F. Yue, W. Wu, O. Denas, D. L. Vera, Y. Wang, R. S. Hansen, T. K. Canfield, R. E. Thurman, Y. Cheng, G. Gülsoy, J. H. Dennis, M. P. Snyder, J. A. Stamatoyannopoulos, J. Taylor, R. C. Hardison, T. Kahveci, B. Ren, and D. M. Gilbert, "Topologically associating domains are stable units of replication-timing regulation", en, *Nature*, vol. 515, no. 7527, pp. 402–405, Nov. 2014, ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature13986. [Online]. Available: http://dx.doi.org/10.1038/nature13986.

[33] A. D. Schmitt, M. Hu, and B. Ren, "Genome-wide mapping and analysis of chromosome architecture", en, *Nat. Rev. Mol. Cell Biol.*, vol. 17, no. 12, pp. 743–755, Dec. 2016, ISSN: 1471-0072, 1471-0080. DOI: 10.1038/nrm.2016.104. [Online]. Available: http://dx.doi.org/10.1038/nrm.2016.104.

[34] T. W. Jeffries, I. V. Grigoriev, J. Grimwood, J. M. Laplaza, A. Aerts, A. Salamov, J. Schmutz, E. Lindquist, P. Dehal, H. Shapiro, Y.-S. Jin, V. Passoth, and P. M. Richardson, "Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast pichia stipitis", en, *Nat. Biotechnol.*, vol. 25, no. 3, pp. 319–326, Mar. 2007, ISSN: 1087-0156. DOI: 10.1038/nbt1290. [Online]. Available: http://dx.doi.org/10.1038/nbt1290.

[35] N. Varoquaux, I. Liachko, F. Ay, J. N. Burton, J. Shendure, M. J. Dunham, J.-P. Vert, and W. S. Noble, "Accurate identification of centromere locations in yeast genomes using Hi-C", en, *Nucleic Acids Res.*, vol. 43, no. 11, pp. 5331–5339, Jun. 2015, ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkv424. [Online]. Available: http://dx.doi.org/10.1093/nar/gkv424.

[36] K. Gong, H. Tjong, X. J. Zhou, and F. Alber, "Comparative 3D genome structure analysis of the fission and the budding yeast", en, *PLoS One*, vol. 10, no. 3, e0119672, Mar. 2015, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0119672. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0119672.

[37] H. Wong, H. Marie-Nelly, S. Herbert, P. Carrivain, H. Blanc, R. Koszul, E. Fabre, and C. Zimmer, “A predictive computational model of the dynamic 3D interphase yeast nucleus”, en, *Curr. Biol.*, vol. 22, no. 20, pp. 1881–1890, Oct. 2012, ISSN: 0960-9822, 1879-0445. DOI: 10.1016/j.cub.2012.07.069. [Online]. Available: http://dx.doi.org/10.1016/j.cub.2012.07.069.

[38] S. S. P. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden, “A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping”, en, *Cell*, vol. 159, no. 7, pp. 1665–1680, Dec. 2014, ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2014.11.021. [Online]. Available: http://dx.doi.org/10.1016/j.cell.2014.11.021.

[39] T. J. Stevens, D. Lando, S. Basu, L. P. Atkinson, Y. Cao, S. F. Lee, M. Leeb, K. J. Wohlfahrt, W. Boucher, A. O’Shaughnessy-Kirwan, J. Cramard, A. J. Faure, M. Ralser, E. Blanco, L. Morey, M. Sansó, M. G. S. Palayret, B. Lehner, L. Di Croce, A. Wutz, B. Hendrich, D. Klenerman, and E. D. Laue, “3D structures of individual mammalian genomes studied by single-cell Hi-C”, en, *Nature*, vol. 544, no. 7648, pp. 59–64, Apr. 2017, ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature21429. [Online]. Available: http://dx.doi.org/10.1038/nature21429.

[40] G. Cottarel, J. H. Shero, P. Hieter, and J. H. Hegemann, “A 125-base-pair CEN6 DNA fragment is sufficient for complete meiotic and mitotic centromere functions in saccharomyces cerevisiae”, en, *Mol. Cell. Biol.*, vol. 9, no. 8, pp. 3342–3349, Aug. 1989, ISSN: 0270-7306. DOI: 10.1128/MCB.9.8.3342. eprint: https://mcb.asm.org/content/9/8/3342.full.pdf. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/2552293.

[41] M. Marbouty, A. Cournac, J.-F. Flot, H. Marie-Nelly, J. Mozziconacci, and R. Koszul, “Data from: Metagenomic chromosome conformation capture (meta3c) unveils the diversity of chromosome organization in microorganisms”, 2014. [Online]. Available: https://dryad2.lib.ncsu.edu/resource/doi:10.5061/dryad.gv595.

[42] H. Li, “Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM”, Mar. 2013. arXiv: 1303.3997 [q-bio.GN]. [Online]. Available: http://arxiv.org/abs/1303.3997.

[43] T. P. Sneddon, P. Li, and S. C. Edmunds, “GigaDB: Announcing the GigaScience database”, en, *Gigascience*, vol. 1, no. 1, p. 11, Jul. 2012, ISSN: 2047-217X. DOI: 10.1186/2047-217X-1-11. [Online]. Available: http://dx.doi.org/10.1186/2047-217X-1-11.

CHAPTER 4

bin3C : exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes (MAGs)

4.1 Authorship Declaration

By signing below I confirm that for the paper titled "bin3C : exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes (MAGs)" and under review with *BMC Genome Biology*, that:

Matthew Z. DeMaere developed the methods, implemented the software, performed and analyzed the experiments, and drafted the manuscript.

Aaron E. Darling revised and edited the manuscript, and conceived and supervised the project.

Production Note:
Signature removed
prior to publication.

Matthew Z. DeMaere

Production Note:
Signature removed
prior to publication.

Associate Professor Aaron E. Darling

4.2 Abstract

Most microbes inhabiting the planet cannot be easily grown in the lab. Metagenomic techniques provide a means to study these organisms, and recent advances in the field have enabled the resolution of individual genomes from metagenomes, so-called Metagenome Assembled Genomes (MAGs). In addition to expanding the catalog of known microbial diversity, the systematic retrieval of MAGs stands as a tenable divide and conquer reduction of metagenome analysis to the simpler problem of single genome analysis. Many leading approaches to MAG retrieval depend upon time-series or transect data, whose effectiveness is a function of community complexity, target abundance and depth of sequencing. Without the need for time-series data, promising alternative methods are based upon the high-throughput sequencing technique called Hi-C.

The Hi-C technique produces read-pairs which capture in-vivo DNA-DNA proximity interactions (contacts). The physical structure of the community modulates the signal derived from these interactions and a hierarchy of interaction rates exists (Intra-chromosomal > Inter-chromosomal > Inter-cellular).

We describe an unsupervised method that exploits the hierarchical nature of Hi-C interaction rates to resolve MAGs from a single time-point. As a quantitative demonstration, next, we validate the method against the ground truth of a simulated human faecal microbiome. Lastly, we directly compare our method against a recently announced proprietary service ProxiMeta, which also performs MAG retrieval using Hi-C data.

Bin3C has been implemented as a simple open-source pipeline and makes use of the unsupervised community detection algorithm Infomap (`https://github.com/cerebis/bin3C`).

Keywords Metagenomics, Hi-C, clustering, next generation sequencing, metagenome-assembled genome

4.3 Background

The number of microbial organisms which can be readily investigated using culture-based techniques is relatively small in proportion to the Earth's apparent total diversity [1], [2]. Although concerted efforts have found the individual conditions necessary to cultivate a relatively small number of species in the laboratory [3]–[5], scaling-up this discovery process to the remaining majority is daunting, if not intractable.

Beyond the issue of cultivation, an environmental population can possess at once phenotypic microdiversity and within that group large differences in gene content. With as little as 40% of genes shared within a species [6], this accessory genome is thought to contribute significantly to the dynamics of microbial adaptation in the environment [7]–[9]. Phylogenetic marker surveys (16S amplicon sequencing), while still informative, stand essentially as a proxy for broader discovery processes of the genomic landscape, should they exist. The systematic extraction of entire genomes from an environment will enable a more thorough determination of the constituent species core and accessory gene content (pangenome). The extracted pangenome and community profile will enable investigation of the functional basis of species fitness and niche partitioning within an environment, and further longitudinal experiments will permit studying the dynamics.

Metagenomics offers a direct culture-independent sampling approach as a means to study the unculturable majority. Recent advances in this field have begun to make possible the systematic resolution of genomes from metagenomes; so-called Metagenome Assembled Genomes (MAGs). Tools designed to assess the quality of retrieved MAGs [10], [11] have brought with them suggestions for categorical quality rankings (Table 4.1). Marking an increasing acceptance, the Genomic Standards Consortium (GSC) recently introduced standardised reporting criteria (Table 4.2) for the submission of MAGs to public archives [12], and as of mid-2018 there are more than 5200 MAGs registered in the Genomes Online (GOLD) database [13]. As retrieval methodologies improve and new complex environments are studied, the registration rate of new MAGs is expected to eventually exceed that of culture-based studies [12].

Most current approaches to the accurate retrieval of MAGs (also called genome binning or clustering) depend on longitudinal or transect data series, operating either directly on WGS sequencing reads (LSA) [14] or on assembly contigs (CONCOCT, GroopM, metaBAT, MaxBin2, Cocacola) [15]–[19]. The need for multiple samples can, however, pose a barrier both in terms of cost of sequencing and the logistics of obtaining multiple

Rank	Completeness (%)	Rank	Contamination (%)
Near	≥ 90	Low	≤ 5
Substantial	≥ 70 to < 90	Medium	> 5 to ≤ 10
Moderate	≥ 50 to < 70	High	> 10 to ≤ 15
Partial	< 50	Very high	> 15

Table 4.1: A proposed standard for reporting the quality of retrieved MAGs which uses only estimates of completeness and contamination [10]. Completeness and contamination are independently ranked and are intended to be used in conjunction, e.g. "nearly complete and low contamination."

Rank	Assembly Quality Criteria	
Finished	Single, validated contiguous sequence per replicon without gaps or ambiguities, with consensus error rate or equivalent $>$ Q50.	
	Completeness and Contamination (%)	**Additionally**
High-quality draft	$> 90, < 5$	Presence of 23S, 16S and 5S and ≥ 18 tRNAs.
Medium-quality draft	$\geq 50, < 10$	
Low-quality draft	$< 50, < 10$	

Table 4.2: A small component of the reporting details for MAGs as proposed by the Genomic Standards Consortium include ranks of quality [12]. The "finished" rank is left to future advances, while lower ranks are achievable now by Hi-C based genome binning methods. The additional criterion of rRNA genes makes the "high-quality" rank challenging to achieve with current methods.

samples as, for instance, with clinical studies. As an alternative single-sample approach, Hi-C (a high throughput sequencing technique which captures in-vivo DNA-DNA proximity) can provide significant resolving power from a single time-point when combined with conventional shotgun sequencing.

The first step of the Hi-C library preparation protocol is to crosslink proteins bound to DNA in vivo using formalin fixation. Next, cells are lysed and the DNA-protein complexes are digested with a restriction enzyme to create free ends in the bound DNA strands. The free ends are then biotin labelled and filled to make blunt ends. Next is the important proximity-ligation step, where blunt ends are ligated under dilute conditions. This situation permits ligation to occur preferentially among DNA strands bound in the same protein complex, that is to say, DNA fragments which were in close proximity in vivo at the time of crosslinking. Crosslinking is then reversed, the DNA is purified and a biotin pull-down step employed to enrich for proximity junction containing products. Lastly, an Illumina-compatible paired-end sequencing library is constructed. After sequencing, each end of a proximity-ligation containing read-pair is composed of DNA from two potentially different intra-chromosomal, inter-chromosomal or even inter-cellular loci.

As a high-throughput sequencing adaptation of the original 3C (chromosome conformation capture) protocol, Hi-C was originally conceived as a means to determine, at once, the 3-dimensional structure of the whole human genome [20]. The richness of information captured in Hi-C experiments is such that the technique has subsequently been applied to a wide range of problems in genomics, such as: genome reassembly [21], haplotype reconstruction [22], [23], assembly clustering [24], centromere prediction [25]. The potential of Hi-C (and other 3C methods) as a means to cluster or deconvolute metagenomes into genome bins has been demonstrated on simulated communities [26]–[28] and real microbiomes [29], [30].

Most recently, commercial Hi-C products ranging from library preparation kits through to analysis services [30], [31] have been announced. These products aim to lessen the experimental challenge in library preparation for non-specialist laboratories, while also raising the quality of data produced. In particular, one recently introduced commercial offering is a proprietary metagenome genome binning service called ProxiMeta, which was demonstrated on a real human gut microbiome, yielding state of the art results [30].

Here we describe a new open software tool bin3C which can retrieve MAGs from metagenomes, by combining conventional metagenome shotgun and Hi-C sequencing data. Using a simulated human faecal microbiome, we externally validate the binning performance of bin3C in terms of adjusted mutual information, and B^3 Precision and Recall against a ground truth. Finally, for a real microbiome from human faeces, we compare the retrieval performance of bin3C against that published for the ProxiMeta service [30].

4.4 Method

4.4.1 Simulated Community

To test the performance of our tool on the task of genome binning, we designed a simulated human gut microbiome from 63 high-quality draft or better bacterial genomes randomly chosen from the Genome Taxonomy Database (GTDB) [32]. Candidate genomes were required to possess an isolation source of faeces or feces, while not specifying a host other than human. To include only higher quality drafts, the associated metadata of each was used to impose the following criteria: contig count ≤ 200, CheckM completeness $> 98\%$, MIMAG quality rank of “High” or better and lastly a total gap length < 500 bp. For these metadata based criteria, there were 223 candidate genomes.

In addition to the metadata based criteria, FastANI (v1.0) [33] was used to calculate pairwise average nucleotide identity (ANI) between the 223 candidate genome sequences. As we desired a diversity of species and mostly unambiguous ground truth, a maximum pairwise ANI of 96% was imposed on the final set of genomes. This constraint controlled for the over-representation of some species within the GTDB. Additionally, when two or more genomes have high sequence identity, the assignment process becomes more difficult and error-prone as it challenges both the assembler [34] and creates ambiguity when assigning assembly contigs back to source genomes.

The resulting 63 selected genomes had an ANI range of 74.8% to 95.8% (median: 77.1%) and GC content range of 28.3% to 73.8% (median: 44.1%) (Figure 4.1) (Table 4.S1). A long-tailed community abundance profile was modelled using a Generalized Pareto distribution (parameters: `shape=20, scale=31, location=0`) (Figure 4.S1), where there was approximately a 50:1 reduction in abundance from most to least abundant. Lastly, before read simulation, genomes in multiple contigs were converted to a closed circular form by concatenation, thereby simplifying downstream interpretation.

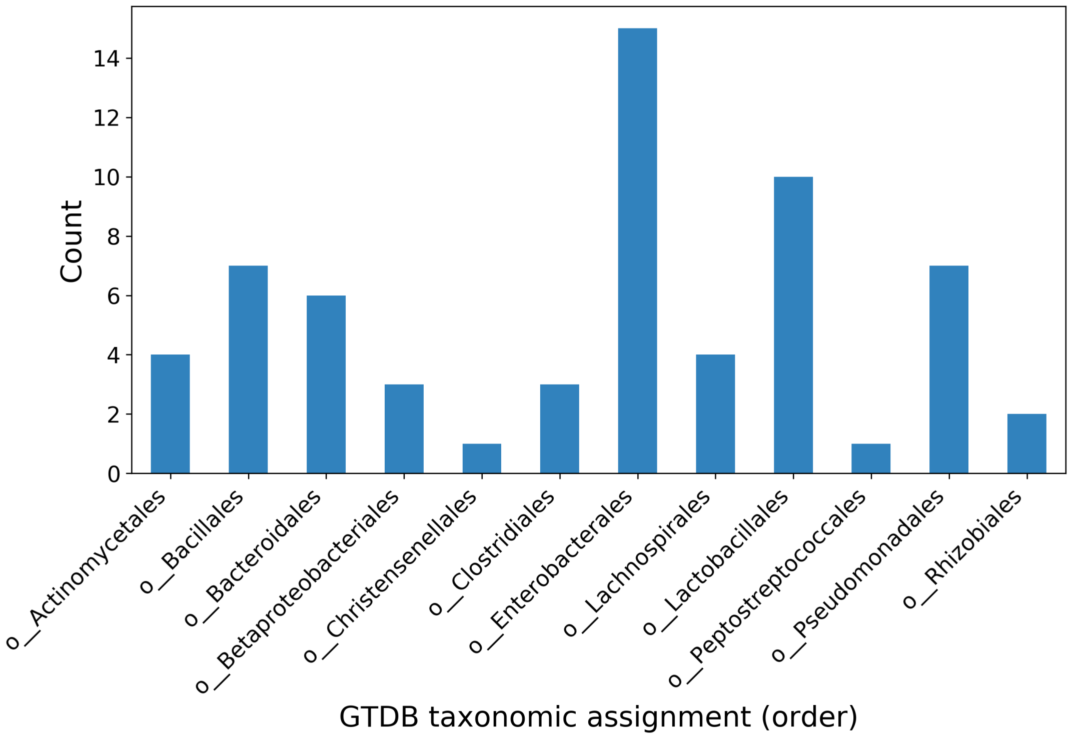

Figure 4.1: Taxonomic distribution at the order rank of 63 selected bacterial genomes used in the simulated community. The number of each order is a product of the taxonomic distribution of genomes existing in the GTDB, while the constraint that no two genomes be more similar than 96% ANI restricts the over-representation of deeply sequenced species.

4.4.2 Read-set generation

To explore how increasing depth of coverage affects bin3C's ability to correctly retrieve MAGs, Hi-C read-sets were generated over a range of depths while keeping shotgun coverage constant. Hi-C depth was parameterised simply by the total number of pairs generated, while shotgun depth was parameterised by the depth of the most abundant community member.

From this definition, an initial read-set with high depth of coverage was produced with 250x shotgun and 200 million Hi-C pairs. The shotgun dataset at this depth constituted 18.2M pairs.

Shotgun reads were generated using the metagenomic shotgun simulator MetaART which wraps the short-read simulator art_illumina (v2.5.1) [35], [36] (options: `-M 100 -S 12345 -l 150 -m 350 -s 20 -z 1`).

Hi-C reads were generated in two equal parts from two different 4-cutter restriction enzymes (NEB names: MluCI and Sau3AI) using Sim3C [36] (options: `-e ${enzyme} -m hic -r 12345 -l 150 -insert-sd 20 -insert-mean 350 -insert-min 150 -linear -simple-reads`). Two enzymes were used to mimic the library construction of the real data-set we also analyzed.

From the initial read-set, a parameter sweep was produced by serially downsampling the initial read-set by factors of 2 using BBTools (v37.25) [37]. The initial Hi-C read-set was reduced 4 times for a total of 5 different depths or 200M, 100M, 50M, 25M, 12.5M pairs (command: `reformat.sh sampleseed=12345 samplerate=${d}`). In terms of the community genomes, depth of coverage for the subsampling with the greatest reduction factor ranged from 3.5x to 171x for Hi-C.

4.4.3 Ground Truth Inference

For the task of the whole-community genome binning, a ground truth was constructed by aligning scaffolds resulting from the SPAdes assembly to the 'closed' reference genomes using LAST (v941) [38]. From the LAST alignments, overlapping source assignment was determined using a methodology we have described previously [34] and implemented as the program `alignmentToTruth.py` (see availability section). An overlapping (soft) ground truth better reflects the possibility of co-assembly of sufficiently similar regions among reference genomes and the tendency that these regions cause breakpoints in assembly algorithms, leading to highly connected assembly fragments which belong equally well to more than one source.

4.4.4 Performance Metrics

To validate genome binning, we employed two extrinsic measures; adjusted mutual information (AMI) (sklearn v0.19.2) and weighted Bcubed (B^3). AMI is a normalized variant of mutual information which corrects for the tendency that the number of agreements between clusters by random chance tends to increase with increasing problem size [39]. Weighted B^3 is a soft extrinsic metric which, analogous to the F-measure, is the harmonic mean of the B^3 formulation of Precision and Recall. Here, precision is a measure of cluster homogeneity (like with like), while recall is a measure of the cluster completeness. The B^3 measure handles overlapping (soft) clusters and better satisfies the constraints that an ideal metric should possess; i.e. homogeneity, completeness, rag-bag and size vs quantity when compared to other metrics. Weighted B^3 extends the definition to allow the objects under study to have variable values, for which contig length is a natural choice with genome binning problems [34], [40], [41].

In employing two measures, we seek to gain confidence in their agreement while also obtaining the additional insight afforded by the separate facets B^3 Precision and Recall.

4.4.5 Real Microbiome

To demonstrate bin3C on real data and make a direct comparison to the proprietary Hi-C based genome binning service (ProxiMeta), we obtained the publicly available high-quality combined whole-metagenome shotgun and Hi-C sequencing data-set used in the previous study [30]. The data-set derives from the microbiome of a human gut (BioProject: PRJNA413092, Acc: SRR6131122, SRR6131123 and SRR6131124).

For this data-set, two separate Hi-C libraries (SRR6131122, SRR6131124) were created using two different 4-cutter restriction enzymes (MluCI and Sau3AI respectively). In using two enzymes, the recognition sites were chosen to be complementary in terms of GC content. When the libraries were subsequently combined during the generation of the contact map, site complementarity provided a higher and more uniform site density over a wider range of target sequence. We conjecture that for metagenome deconvolution, site complementarity is particularly helpful in obtaining a consistent signal from all community members, while higher site density improves recovery of smaller assembly fragments.

All read-sets were obtained from an Illumina HiSeq X Ten at 150 bp. After clean-up (described below), the shotgun read-set (SRR6131123) consisted of 248.8 million paired-end reads, while the two Hi-C libraries consisted of 43.7 million (SRR6131122) and 40.8 million (SRR6131124) paired-end reads.

4.4.6 Initial Processing

Read clean-up is occasionally overlooked in the pursuit of completing the early stages of genomic analysis. This initial processing step is however essential for optimal shotgun assembly and particularly for Hi-C read mapping where remnants of adapter sequence, PhiX or other contaminants can be a significant noise source.

Data Set	N50	L50	Contigs $\geq$ 1kbp	All contigs	Scaffolds $\geq$ 1kbp	All scaffolds	Total extent (bp)
Real human gut	56,282	1277	97,760	670,379	95,521	652,723	719,550,669
Simulated human gut	29,009	1170	24,324	116,696	23,364	41,704	240,133,820

Table 4.3: Assembly statistics for real and simulated human gut microbiomes.

A standard cleaning procedure was applied to all WGS and Hi-C read-sets using bbduk from the BBTools suite (v37.25) [37], where each was screened for PhiX and Illumina adapter remnants by reference and by kmer (options: `k=23 hdist=1 mink=11 ktrim=r tpe tbo`), quality trimmed (options: `ftm=5 qtrim=r trimq=10`). For Hi-C read-sets, only paired reads are kept to expedite later stages of analysis. Shotgun assemblies for both simulated and read read-sets (Table 4.3) were produced using SPAdes (v.3.11.1) [42] in metagenomic mode with a maximum kmer size of 61 (options: `-meta -k 21,33,55,61`).

4.4.7 Hi-C Read Mapping

As bin3C is not aimed at assembly correction, we opted to use assembly scaffolds rather than contigs as the target for genome binning, electing to trust any groupings of contigs into scaffolds done by SPAdes.

Both simulated and real Hi-C reads were mapped to their respective scaffolds using BWA MEM (v0.7.17-r1188) [43]. During mapping with BWA MEM, read pairing and mate-pair rescue functions were disabled and primary alignments forced to be the alignment with lowest read coordinate (5' end) (options: `-5SP`). This latter option is a recent introduction to BWA at the request of the Hi-C bioinformatics community. The resulting BAM files were subsequently processed using samtools (v1.9) [44] to remove unmapped reads, supplementary and secondary alignments (option: `-F 0x904`), then sorted by name and merged.

4.4.8 Contact Map Generation

The large number of contigs ($>$ 500,000) typically returned from metagenomic shotgun assemblies for non-trivial communities is a potential algorithmic scaling problem. At the same time, biologically important contigs can be on the order of 1000 bp or smaller, challenging the effective analysis of metagenomic datasets from both sides.

A Hi-C analysis, when conducted in the presence of experimental biases, involves the observation of proximity-ligation events, which in turn rely on the occurrence of restriction sites. The signal we desire to exploit is therefore not smoothly and uniformly distributed between and across all contigs. As a counting experiment, the shortest contigs can be problematic as they tend to possess a weaker signal with higher variance; as a result, they can have a deleterious effect on normalisation and clustering if included. Therefore, bin3C imposes constraints on minimum acceptable length (default: 1000 bp) and minimum acceptable raw signal (default: 5 non-self observations) for contig inclusion. Any contig which fails to meet these criteria is excluded from the clustering analysis.

With this in mind, bin3C constructs a contact map from the Hi-C read-pairs. As in previous work [26], the bins pertain to whole contigs and capture global interactions, which work effectively to cluster a metagenome into genome bins. In doing so, we make the implicit assumption that assembly contigs contain few misassemblies that would confound or otherwise invalidate the process of partitioning a metagenome into genome bins.

Bin3C can also optionally construct a contact map binned on windows of genomic extent. These maps are not used in the analysis per se but can be used to plot visual representation of the result in the form of a heatmap (Figure 4.S2).

4.4.9 Bias Removal

The observed interaction counts within raw Hi-C contact maps contain experimental biases, due in part to factors such as mappability of reads, enzyme digestion efficiency, in vivo conformational constraints on accessibility, and restriction site density. In order to apply Hi-C data to genome binning, a uniform signal over all DNA molecules would be ideal, free of any bias introduced by the factors mentioned above. Correcting for these biases is an important step in our analysis, which is done using a two-stage process. First, for each enzyme used in library preparation, the number of enzymatic cut sites are tallied for each contig. Next, each pairwise raw Hi-C interaction count c_{ij} between contigs i and j is divided by the product of the number of cut sites found for each contig n_i, n_j. This first correction is then followed by general bistochastic matrix balancing using the Knight-Ruiz algorithm [45].

4.4.10 Genome binning

After bias removal, the wc-contact map (whole contig) is transformed to a graph where nodes are contigs and edge weights are normalized interaction strength between contigs i and j. It has been shown that DNA-DNA interactions between loci within a single physical cell (intra-cellular proximity interactions) occur an order of magnitude more frequently than interactions between cells (inter-cellular) [26] and, in practice, the signal from inter-cellular interactions is on par with experimental noise. The wc-graph derived from a microbial metagenome is then of low density (far from fully connected), being composed of tightly interacting groups (highly modular) representing intra-cellular interactions and against a much weaker background of experimental noise. Graphs with these characteristics are particularly well suited to unsupervised cluster analysis, also known as community detection.

Unsupervised clustering of the wc-graph has previously been demonstrated using Markov clustering [26], [46] and the Louvain method [28], [47]. In a thorough investigation using ground truth validation, we previously found neither method to be sufficiently efficacious in general practice [34]. Despite the high signal to noise from recent advances in library preparation methods, accurate and precise clustering of the wc-graph remains a challenge. This is because resolving all of the structural detail (all of the communities) becomes an increasingly fine-grained task as graphs grow in size and number of communities. Clustering algorithms can, in turn, possess a resolution limit if a scale exists below which they cannot recover finer detail. As it happens, modularity-based methods such as Louvain have been identified as possessing such a limit [48]. For Hi-C based microbiome studies, the complexity of the community and the experiment are sufficient to introduce significant structural variance within the wc-graph. A wide variation such aspects as in the size of clusters and weight of intra-cluster edges relative to the whole graph make a complete reconstruction difficult for algorithms with limited resolution.

The state of unsupervised clustering algorithms has however been advancing. Benchmarking standards have made thorough extrinsic validation of new methods commonplace [49], and comparative studies have demonstrated the capability of available methods [50]. Infomap is another clustering algorithm, which like Markov clustering is based upon flow [51], [52]. Rather than considering the connectivity of groups of nodes versus the whole, flow models consider the tendency for random walks to persist in some regions of the graph longer than others. Considering the dynamics rather than the structure of a graph, flow models can be less susceptible to resolution limits as graph size increases

[53]. Additionally, the reasonable time-complexity and the ability to accurately resolve clusters without parameter tuning makes Infomap well suited to a discovery science where unsupervised learning is required.

We have therefore employed Infomap (v0.19.25) to cluster the wc-graph into genome bins (options: `-u -z -i link-list -N 10`). Genome bins greater than a user-controlled minimum extent (measured in base-pairs) are subsequently written out as multi-FASTA in descending cluster size. A per-bin statistics report is generated detailing bin extent, size, GC content, N50, and read depth statistics. By default, a whole sample contact map plot is produced for qualitative assessment.

In the following analyses, we have imposed a 50 kbp minimum extent on genome bins, partly for the sake of figure clarity and as a practical working limit for prokaryotic MAG retrieval. That is to say, being less than half the minimum length of the shortest known bacterial genome [54], it is unlikely that this threshold would exclude a candidate of moderate or better completeness. If a user is in doubt or has another objective in mind, the constraint can be removed.

4.5 Results

4.5.1 Simulated Community Analysis

We validated the quality of bin3C solutions as Hi-C depth of coverage was swept from 12.5M to 200M pairs on an assembly (Figure 4.2). A sharp gain in AMI, B^3 Recall and B^3 F-score was evident as Hi-C coverage rose from 12.5M to 100M pairs, while the gain between 100M and 200M pairs was less pronounced. Accompanying the upward trend for these first three measures was an inverse but relatively small change in B^3 Precision. In terms of AMI, the highest scoring solution of 0.848 was at the greatest simulated depth of 200M pairs. Concomitantly this solution had B^3 Precision, Recall and F-scores of 0.909, 0.839 and 0.873 respectively. For this highest depth sample, 22,279 contigs passed the bin3C filtering criteria and represented 95.4% of all assembly contigs over 1000 bp. There were 62 genome bins with an extent greater than 50 kbp, with total extent 229,473,556 bp. This was 95.6% of the extent of the entire shotgun assembly, which itself was 91.1% of the extent of the set of reference genomes. The remaining small clusters of less than 50 kb extent totalled 1,413,596 bp or 0.6% of the assembly extent (Table 4.3), while unanalyzed contigs below 1000 bp represented 8,103,486 bp or 3.4%.

As a soft clustering measure, B^3 can consider overlaps both within predicted clusters and the ground truth. Regions of shared sequence within our simulated community meant that for 4.4% of assembly contigs the assignment in the ground truth was ambiguous, being shared by two or more source genomes. Meanwhile, bin3C solutions are hard clusters placing contigs in only one genome bin. Even without mistakes, this leaves a small but unbridgeable gap between the ground truth and the best possible bin3C solution. Due to this, when overlap exists in the ground truth, the maximum achievable B^3 Precision and Recall will be less than unity. Conversely, AMI is a hard clustering measure that requires assigning each of these shared contigs in the ground truth to a single source genome through a coin-toss process. It remains, however, that when bin3C selects a bin for such contigs, either source would be equally valid. For this reason, AMI scores are also unlikely to achieve unity in the presence of overlapping genomes.

Despite these technicalities, a quantitative assessment of overall completeness and contamination is robustly inferred using B^3 Recall and Precision, as they consider contig assignments for the entirety of the metagenomic assembly. This is in contrast to marker gene based measures of completeness and contamination, where only those contigs containing marker genes contribute to the score. The overall completeness of bin3C solutions, as inferred using B^3 Recall, rose monotonically from 0.189 to 0.839 as Hi-C depth of coverage was increased from 12.5M to 200M pairs. At the same time, the overall contamination, as inferred using B^3 Precision, dropped slightly from 0.977 to 0.909. Thus bin3C responded positively to increased depth of Hi-C coverage while maintaining an overall low degree of contamination.

We validated our simulation sweep using the marker gene tool CheckM [10]. CheckM estimated that bin3C retrieved 33 nearly complete MAGs using 12.5M Hi-C pairs, while 39 nearly complete were retrieved using 200M pairs (Figure 4.3). For the deepest run with the most retrieved MAGs, genome bins deemed nearly complete had a total extent which ranged from 1.56 Mbp to 6.97 Mbp, shotgun depth of coverage from 3.34x to 161.2x, N50 from 5797 bp to 2.24 Mbp, GC content from 28.0% to 73.9% and number of contigs from 4 to 787 (Figure 4.S3) (Table 4.S2).

Broadening the count to include MAGs of all three ranks: moderate, substantial and nearly (Table 4.1); 37 were retrieved at 12.5M Hi-C pairs, which increased to 48 when using 200M Hi-C pairs. The small increase in the number of retrieved MAGs for the relatively large increase in Hi-C depth of coverage may seem perplexing, particularly in the face of a large change in the extrinsic validation measures AMI, B^3 Recall and F-score. To explain this, we referred to the cluster reports provided by bin3C, where we found that the average number of contigs in nearly complete MAGs increased from 94 at 12.5M pairs to 179 at 200M pairs. Thus, although marker gene associated contigs are efficiently found at lower Hi-C depth of coverage, obtaining a more complete representation of each MAG can require significantly more depth.

With respect to contamination as inferred by marker genes, CheckM estimated a low median contamination rate of 1.08% across all genome bins with completeness greater than 70%. CheckM, however, also identified four bins where contamination was estimated to be higher than 10% and for which marker gene counting suggested that two genomes had merged into a single bin. We interrogated the ground truth to determine the heritage of these bins and found that each was a composite of two source genomes, whose pairwise ANI values ranged from 93.1% to 95.8%. Each pair shared an average of 131 contigs within the ground truth with an average Jaccard index of 0.19, which was significant when compared against the community-wide average Jaccard of 6.5×10^{-4}. Thus, a few members of the simulated community possessed sufficiently similar or shared sequence to produce co-assembled contigs. Although the co-assembled contigs were short, with a median length of 2011 bp, the degree of overlap within each pair was enough to produce single clusters for sufficiently deep Hi-C coverage. Reference genomes corresponding to two of these merged bins fall within the definition of intraspecies, with pairwise ANI values of 95.80% and 95.85% respectively. The reference genomes involved with remaining two bins are close to this threshold, with ANI values of 93.1% and 93.5%. From this, we would concede that although bin3C is precise, it is not capable of resolving strains.

4.5.2 Library Recommendations

The time, effort and cost of producing a combined shotgun and Hi-C metagenomic dataset should be rewarded with good results. As bin3C is reliant on both the quality and quantity of data supplied, we felt it important to highlight two factors beyond Hi-C depth of coverage which can influence results.

Shotgun sequencing data forms the basis on which Hi-C associations are made and therefore, the more thoroughly a community is sampled, the better. To demonstrate how this affects bin3C, we reduced the shotgun depth of coverage of our simulated community by half (to 125x) and reassembled the metagenome. Basic assembly statistics for this half-depth assembly were N50 6289 bp and L50 4353. There were 43,712 contigs longer

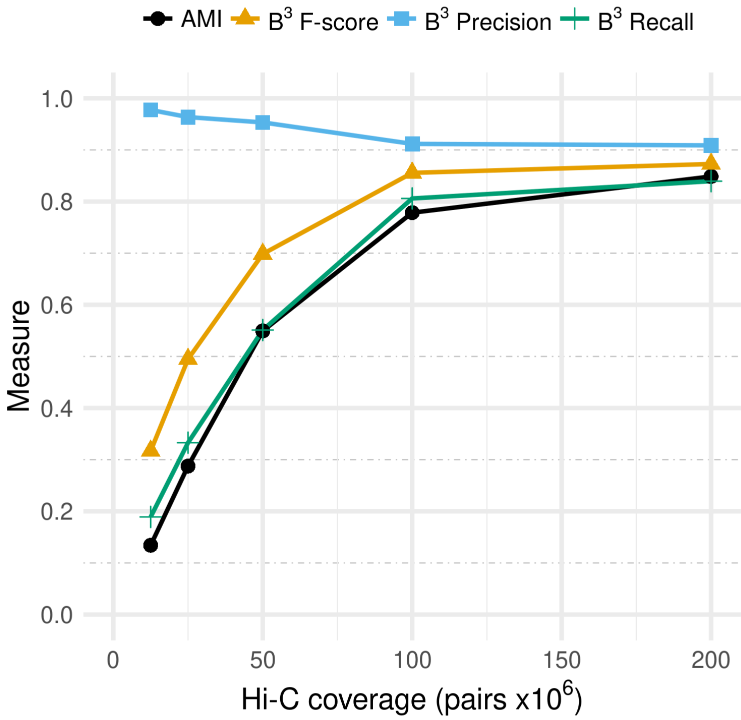

Figure 4.2: Validation of bin3C solutions using extrinsic measures and a ground truth. Bin3C was run against five simulated experiments, with increasing Hi-C depth of coverage while keeping shotgun coverage fixed. With diminishing returns from 100M to 200M pairs, the highest depth of coverage produced the best scoring genome binning solution, with an AMI 0.849 and B^3 Precision, Recall and F-score of 0.909, 0.839 and 0.873 respectively.

than 1000 bp with an extent of 187,388,993 bp and overall, there were 113,754 contigs with the total extent of 222,522,774 bp. This contrasts to the full-depth (250x) assembly, which had N50 30,402 bp and L50 1105, with 23,364 contigs over 1000 bp with an extent of 232,030,334 bp, and 41,704 total contigs with an extent of 240,133,820 bp. Clearly, the reduction in shotgun depth has resulted in a more fragmented assembly. In particular, the decrease in depth has lead to a 45 Mbp drop in total extent for contigs longer than 1000 bp. This large proportional shift of assembly extent to fragments smaller than 1000 bp is significant as we have found that this length is an effective working limit within bin3C.

We then analysed the resulting contigs with bin3C over the same range of Hi-C depth of coverage as before. Comparison of the AMI validation scores using the half and full depth assemblies (Figure 4.4) shows that, for the more deeply sampled community, bin3C's reconstruction of the community greatly improved. CheckM estimation of completeness and contamination followed a similar trend (Figure 4.S4), where the best result at half depth

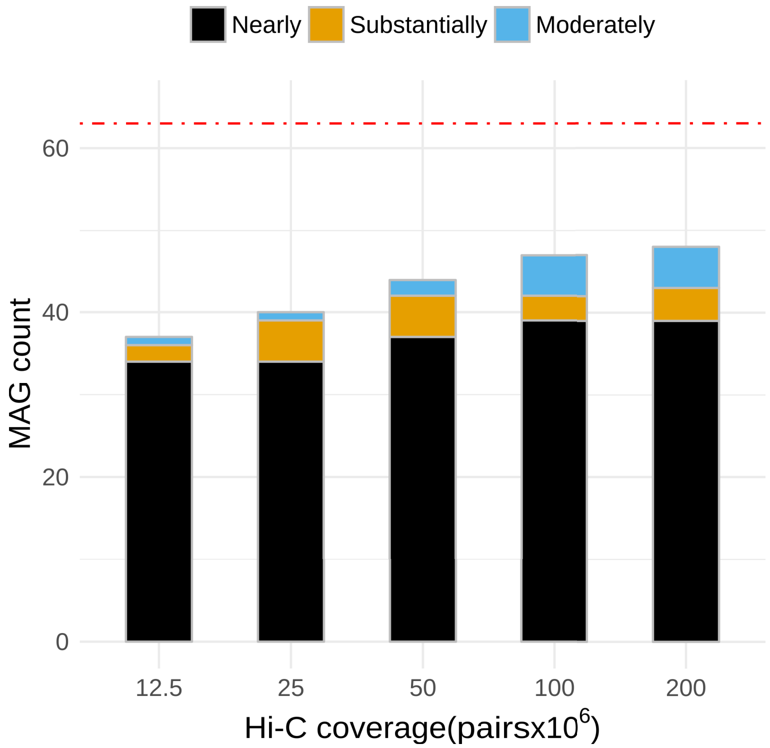

Figure 4.3: For the simulated community, CheckM was used to validate MAGs retrieved using bin3C for increasing depth of Hi-C coverage. The red dashed line indicates the total number of reference genomes used in constructing the simulated community. The step with the highest depth and consequently highest B^3 Recall retrieved 39 nearly, 4 substantially and 5 moderately complete MAGs. Nearly complete MAG retrieval at 100M pairs was equal to that of 200M, with 3 substantially and 5 moderately complete MAGs.

produced 25 nearly, 4 substantially and 6 moderately complete MAGs, compared against 39 nearly, 4 substantially and 5 moderately complete at full depth.

A recent trend in the preparation of metagenomic Hi-C libraries involves employing two different restriction enzymes during the digestion step [30]. The enzymes are chosen to have different GC biases at their restriction sites. For a microbial community with a diversity of species and consequently a wide range of GC content, the intent of this strategy is more uniform digestion of the extracted DNA, and therefore coverage of Hi-C reads across the metagenome. With wider and more uniform coverage, so the logic goes, should come improved results when performing Hi-C based genome binning.

As our work already involved simulating a two-enzyme library, as used in recent real experiments [30], we elected to repurpose this data to ascertain what gain was had in using two enzymes rather than one alone. The two enzymes used in our simulated libraries are

Sau3AI and MluCI. While the Sau3AI restriction site ^GATC is GC balanced, the ^AATT restriction site of MluCI is AT-rich. For our simulated community, source genomes ranged in GC content from 28.3% to 73.8% and their abundances were randomly distributed. For Sau3AI, these extremes of GC content translated to expected cut-site frequencies of 1 in every 338 bp at 28.3% and 1 in every 427 bp at 73.8%. For the less balanced MluCI, the expected cut-site frequencies were instead 1 in every 61 bp at 28.3% and 1 in every 3396 bp at 73.8%. Thus, relative to a naive 4-cutter frequency of 1 in every 256 bp, while the predicted density of sites from Sau3AI is not ideal at either extreme, the site density of MluCI will be very high in the low GC range but very sparse at the high GC range.

For the simulated community full depth assembly, we used bin3C to analyze three Hi-C scenarios: two single enzyme libraries generated using either Sau3AI or MluCI, and a two-enzyme library using Sau3AI and MluCI together. The performance of bin3C was then assessed against the libraries at equal Hi-C depth of coverage using our ground truth. In terms of AMI, the performance of bin3C for the single enzyme libraries was less than that of the combined Sau3AI+MluCI library (Figure 4.5). Although the gain was small at lower depth, the advantage of a two enzyme model grew as depth increased, where at 100M Hi-C pairs the AMI scores were MluCI: 0.63, Sau3AI: 0.71 and Sau3AI+MluCI: 0.78.

4.5.3 Real Microbiome Analysis

We analyzed the real human gut microbiome (Table 4.3) with bin3C using the same parameters as with the simulated community along with a randomly generated seed (options: `–min-map 60 –min-len 1000 –min-signal 5 -e Sau3AI -e MluCI –seed 9878132`). Executed on a 2.6GHz Intel Xeon E5-2697, contact map generation required 586 MB of memory and 15m26s of CPU time, while the clustering stage required 11.6 GB of memory and 9m06s of CPU time. Of the 95,521 contigs longer than 1000 bp, 29,653 had sufficient signal to be included in clustering. The total extent of contigs greater than 1000 bp was 517,309,710 bp for the whole assembly, while those with sufficient Hi-C observations totalled 339,181,288 bp or 65.6% of all those in the assembly.

Clustering the contact map into genome bins, bin3C identified 296 genome bins with extents longer than 50 kbp and 2013 longer than 10 kbp. The 296 clusters longer than 50 kbp had a total extent of 290,643,239 bp, representing 40.4% of the total extent of the assembly, while clusters longer than 10 kbp totalled 324,223,887 bp in extent or 45.1% of the assembly.

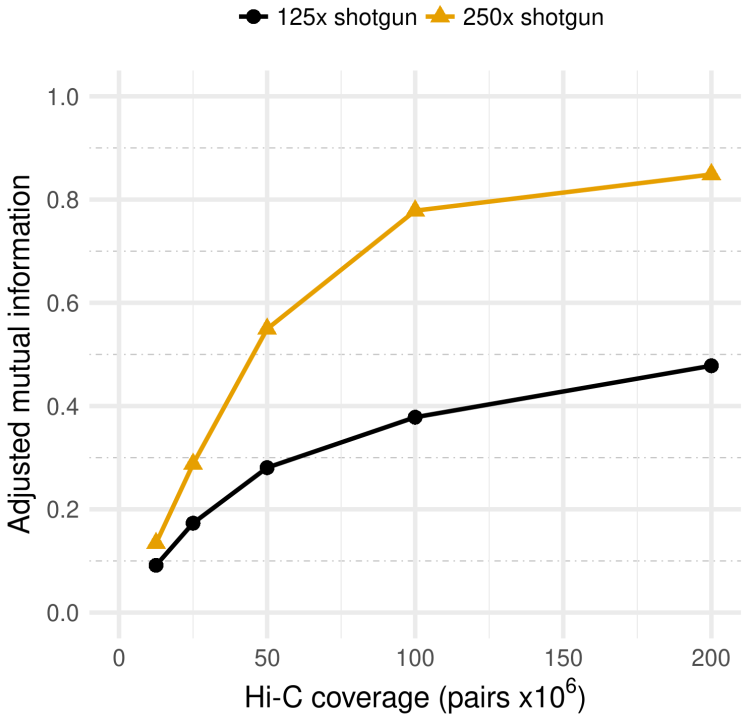

Figure 4.4: Adjusted mutual information (AMI) scores for bin3C solutions at two different shotgun depths of coverage. For our simulated community, shotgun libraries generated at 125x and 250x coverage demonstrate that although the depth of Hi-C coverage is crucial, so too is the depth of shotgun sequencing.

For clusters greater than 50 kb, shotgun depth of coverage ranged from 3.4x to 498x, N50 ranged from 3119 bp to 297,079 bp, GC content from 28.2% to 65.0%, total extent from 50,315 bp to 5,460,325 bp and number of contigs from 1 to 495 (Table 4.S3).

We analyzed these 296 genome bins using CheckM (Figure 4.6) [10]. For the proposed MAG ranking standard based on only measures of completeness and contamination (Table 4.1), bin3C retrieved 55 nearly, 29 substantially and 12 moderately complete MAGs. In terms of total extent, MAGs ranked as nearly complete ranged from 1.68 Mbp to 4.97 Mbp, while for the substantially complete ranged from 1.56 Mbp to 5.46 Mbp and moderately complete ranged from 1.22 Mbp to 3.40 Mbp (Table 4.S4). In terms of shotgun coverage, MAGs ranked as nearly complete ranged from 5.9x to 447.5x, substantially from 4.3x to 416.4x and moderately from 3.7x to 83.4x.

Using the more detailed ranking instead from the recently proposed extension to MIxS (Table 4.2) [12], the bin3C solution represented 17 high quality, 78 medium quality and 105 low-quality MAGs. For the high-quality MAGs, shotgun coverage ranged from 10.7x to

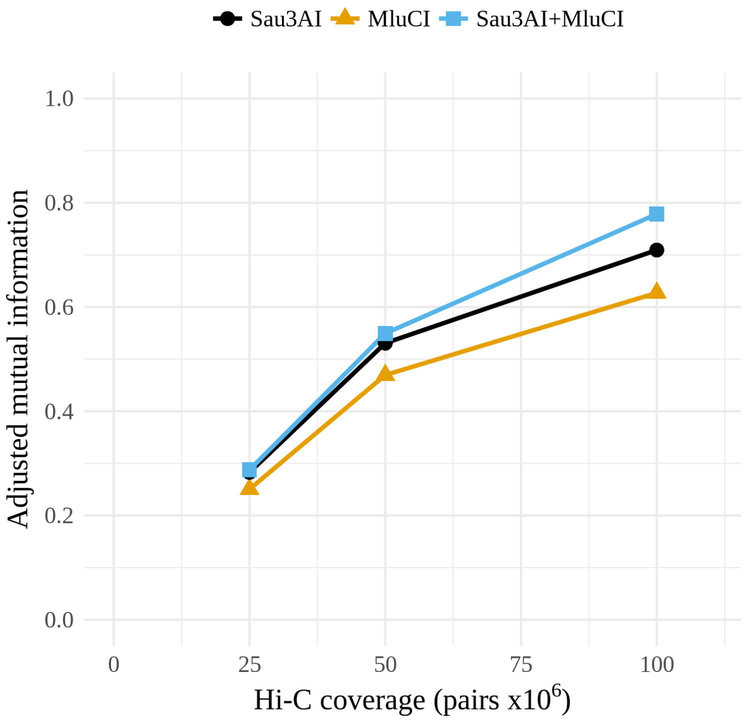

Figure 4.5: For a simulated community whose GC content varied between 28.3% to 73.8%, bin3C retrieval performance improved when simulated reads were generated as if from a library prepared using a two enzyme digestion model (Sau3AI+MluCI), rather than if the library was prepared using either enzyme in isolation.

447.5x, extent from 1.86 Mbp to 4.10 Mbp (Table 4.S5).

4.5.4 Comparison to previous work

The real microbiome we analyzed with bin3C was first described in a previous study to demonstrate a metagenomic Hi-C analysis service called ProxiMeta [30]. ProxiMeta is the only other complete solution for Hi-C based metagenome deconvolution with which to compare bin3C. As ProxiMeta is a proprietary service rather than open source software, the comparison was made by reanalysis of the same dataset as used in their work (Bioproject: PRJNA413092). As their study included a comparison to the conventional metagenomic binner MaxBin (v2.2.4) [55], which was one of the best performing MAG retrieval tools evaluated in the first CAMI challenge [56], we have included those results here as well. It should be noted that although MaxBin 2 is capable of multi-sample analysis, all software

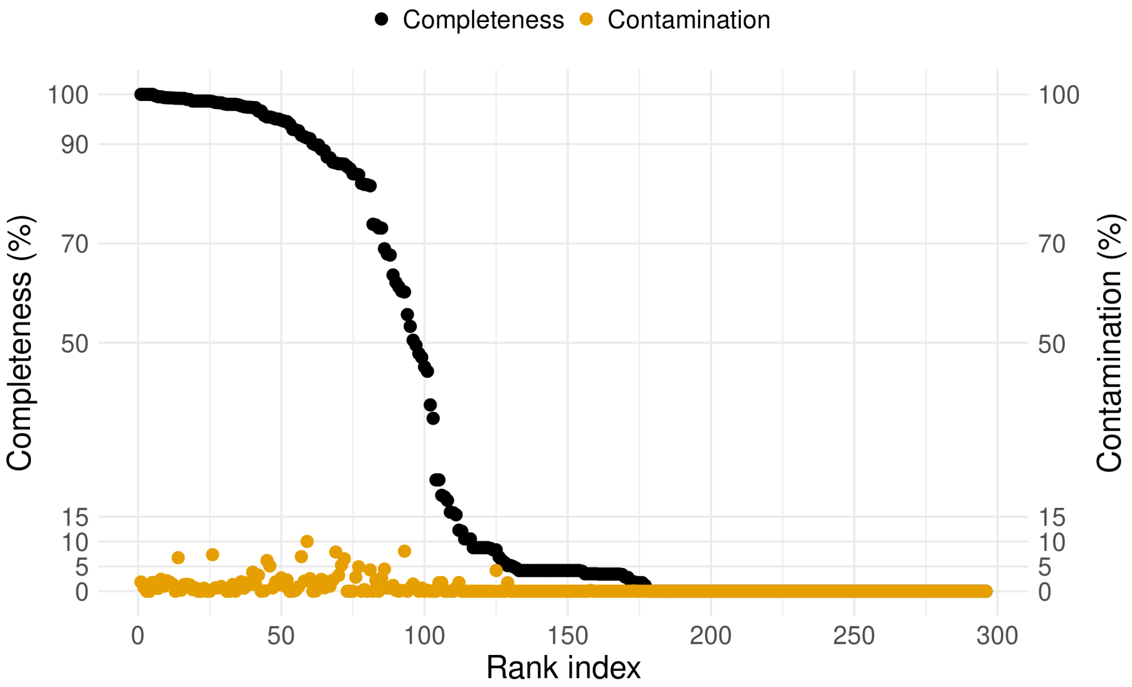

Figure 4.6: Bin3C retrieved MAGs from a real human gut microbiome, ordered by descending estimate of completeness (black circles). Plotted along with completeness is estimated contamination (gold circles). The y-axis grid lines pertain to thresholds used in quality assessment standards: completeness of 50%, 70% and 90% and contamination of 5%, 10% and 15%. Although there is a sharp fall-off in completeness after roughly 75 MAGs, estimated contamination remains consistently low.

was run against a single shotgun sequencing sample. We have compared the CheckM validation of bin3C results to the CheckM validation of ProxiMeta and MaxBin as provided in their supplementary data [57].

Regarding the simple ranking standard (Table 4.1), it was reported that ProxiMeta retrieved 35 nearly, 29 substantially and 13 moderately complete MAGs, while MaxBin retrieved 20 nearly, 22 substantially and 17 moderately complete MAGs. On the same metagenomic Hi-C dataset, we found that bin3C retrieved 55 nearly, 29 substantially and 12 moderately complete MAGs (Figure 4.7A). Against MaxBin, bin3C retrieved fewer moderately complete MAGs but otherwise bettered its performance. Against ProxiMeta, bin3C had equivalent performance for the substantially and moderately complete ranks, while retrieving 20 additional nearly complete genomes, representing an improvement of 57%.

In terms of the more complex MIMAG standard (Table 4.2), it was reported that

ProxiMeta retrieved 10 high and 65 medium quality MAGs, while MaxBin retrieved 5 high and 44 medium quality MAGs. The bin3C solution retrieved 17 high and 78 medium quality MAGs, which against ProxiMeta represents 70% improvement in high-quality MAG retrieval from the same sample (Figure 4.7B).

It was demonstrated previously that ProxiMeta possessed a higher binning precision than MaxBin and resulted in a much lower rate of contamination [30]. We have found that the precision of bin3C improves on the mark set by ProxiMeta. bin3C's gains, when retrieving MAGs in the highest quality ranks, are mainly due to the rejection of fewer bins for excessive contamination. For all genome bins over 1 Mbp in extent, bin3C had a median contamination rate of 0.8%, while for ProxiMeta median contamination was 3.5% and MaxBin this was 9.5%.

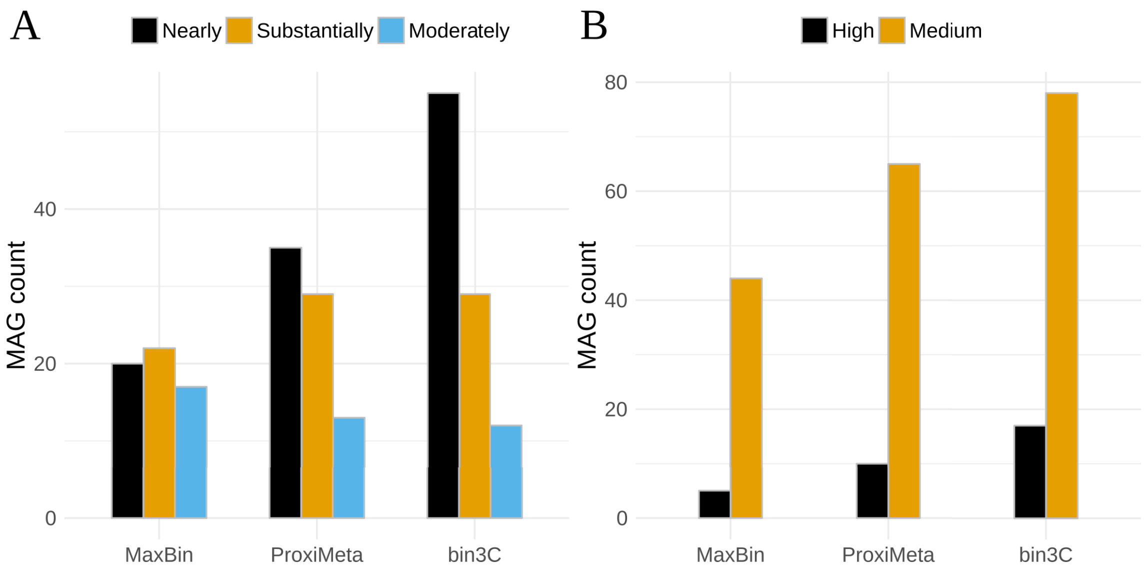

Figure 4.7: In comparison to existing conventional and Hi-C based single-sample metagenome binning tools, bin3C performs well. When compared by ranking standards, based either on measures of completeness and contamination only (A) [10] or the recent GSC MIMAG reporting standard (B) [12], bin3C retrieves a higher or equivalent number of MAGs in each category. The apparent stringency of the MIMAG high quality is primarily due to the requirement that 5S, 16S and 23S rRNA genes be present.

4.6 Discussion

We have introduced bin3C, an openly implemented and generic algorithm which reproducibly and effectively retrieves MAGs on both simulated and real metagenomic data.

To demonstrate this, we assessed bin3C's retrieval performance on a simulated human gut microbiome, by way of a ground truth and the extrinsic validation measures of AMI, as well as B^3 Precision, Recall and F-score (Figure 4.2). Bin3C proved to be consistently precise over a wide range of Hi-C depth of coverage, while recall and the overall quality of solutions improved substantially as more Hi-C data was included. Although a high shotgun depth of coverage is not necessary to obtain low contamination MAGs, greater depth of shotgun sequencing has a strongly positive influence on the recall and overall completeness of MAG retrieval (Figure 4.4).

Hi-C MAGs have a characteristically low rate of contamination by foreign genomic content [30]. On a real human gut microbiome, we have shown that bin3C achieves a lower estimated rate of contamination than both the conventional metagenome binner MaxBin [55] and the recently introduced commercial Hi-C analysis service ProxiMeta [30]. For all bins over 1 Mbp as determined by each approach, bin3C's median contamination rate was 0.8%, while MaxBin was 9.5% and ProxiMeta was 3.5%.

This low contamination rate is a primary reason why bin3C attained the most complete retrieval of MAGs from the real human gut dataset when compared to MaxBin and ProxiMeta (Figure 4.6). Retrieving 20 more nearly complete MAGs than ProxiMeta, bin3C achieved a gain of 57% on this previous best result (Figure 4.7A). For the stringent GSC MIMAG high-quality ranking, bin3C retrieved 17 MAGs from the gut microbiome, a gain of 70% against the previous best result (Figure 4.7B).

For best results, we recommend that Hi-C metagenomic libraries be constructed using a two enzyme digestion model.

4.6.1 Limitations and future work

The ground truth as determined in our work is imperfect, notably when a simulated community possesses multiple strains of a single species. The plethora of extrinsic validation measures from which to choose also have their limitations and differences [40], [41], [50]. Though we chose measures which we felt best suited our problem space, these are not in widespread use. Different measures can have significantly different opinions on the agreement between a ground truth and a given solution. Those with the lowest scoring results are not always the most readily chosen for publication.

The use of non-trivial simulated microbial communities makes determining ground truth and measuring accuracy difficult, and yet these are a crucial element of the development process if the resulting methods are to be robust in real experimental use. Under such circumstances, we work from the premise that achieving close to unity on strong validation measures is unlikely to be possible. In our work here, bin3C demonstrated a B^3 Precision varying between 0.909 and 0.977, while in work pertaining to metagenome binning with multiple samples, precision values as high as 0.998 were reported using a different formulation of the measure [17]. In practical terms by using CheckM as an operational measure of precision, bin3C achieved a much lower rate of MAG contamination on real data than has previously been reported.

Though marker gene based validation with tools such as CheckM or BUSCO [10], [11] are of great value and easily applied to our work, as validators, their perception is limited only to those sequences which contain marker genes. Ideally, metagenome binning approaches should aim to gather together all the sequence fragments pertaining to a given genome and not only those which contained marker genes. The generalizability of an approach is not assured when the validation measure used in development is systematically insensitive to some aspect of the problem. Therefore, we believe refining the ground truth determination process, to be independent of community complexity, is warranted and would be a useful contribution.

Although bin3C can analyze sequences shorter than 1000 bp, it is our experience that allowing them into the analysis does not lead to improvements in MAG retrieval. We believe the weaker signal and higher variance in the raw observations for Hi-C contacts involving shorter sequences is to blame. A weakness here is relying on the final assembly contigs or scaffolds as the subject of read mapping, where the ends of sequences interrupt alignment. In future work, we believe aligning Hi-C reads to an assembly graph has the potential to achieve better results.

Against the simulated community, the performance of bin3C as indicated by the validation scores AMI and B^3 Recall, suggests that further gains in retrieval completeness are possible (Figure 4.2). In particular, strains of the same species can fail to be resolved into separate bins. Improving the resolving power of bin3C or the addition of a post hoc reconciliation process to separate these merged bins would be worthwhile.

4.7 List of abbreviations

- AMI - adjusted mutual information
- ANI - average nucleotide identity
- bp - base-pairs
- CPU - central processing unit
- DNA - deoxyribonucleic acid
- GOLD - Genomes Online Database
- GSC - Genomic Standards Consortium
- GTDB - Genome Taxonomy Database
- M - million
- Mbp - mega base-pairs
- kbp - kilo base-pairs
- MAG - metagenome-assembled genome
- MIMAG - Minimum information about a metagenome-assembled genome
- MIxS - Minimum information about “some” sequence
- 3C - chromosome conformation capture

4.8 Declarations

4.8.1 Author contributions

MZD developed the methods, implemented the software, performed and analyzed the experiments, and drafted the manuscript. AED revised and edited the manuscript, and conceived and supervised the project. All authors read and approved the final manuscript.

4.8.2 Competing interests

The authors declare that they have no competing interests.

4.8.3 Consent for publication

Not applicable

4.8.4 Ethics approval and consent to participate

Not applicable

4.8.5 Funding

This research was supported partially by the Australian Government through the Australian Research Council Discovery Projects funding scheme (project DP180101506, URL `http://purl.org/au-research/grants/arc/DP180101506` and project LP150100912, CI: S.P. Djordjevic, `http://purl.org/au-research/grants/arc/LP150100912`)

4.8.6 Availability of data and materials

- Project name: bin3C
- Repository: `https://github.com/cerebis/bin3C`
- O/S: Linux
- Language: Python 2.7, C/C++
- License: GNU Affero General Public License v3.
- Manuscript DOI: `https://doi.org/10.5281/zenodo.1341423`

4.8.7 Supporting tools

- sim3C metagenomic Hi-C reads simulator
  - Repository: `https://github.com/cerebis/sim3C`
  - Manuscript DOI: `https://doi.org/10.5281/zenodo.1035049`
- MetaART metagenomic shotgun reads simulator and `alignmentToTruth.py`
  - Repository: `https://github.com/cerebis/meta-sweeper`
  - Manuscript DOI: `https://doi.org/10.5281/zenodo.1341441`

Simulated datasets used in this study are available at `https://doi.org/10.5281/zenodo.1342169`. The real human gut microbiome used in this study was downloaded from the NCBI Sequence Read Archive (`http://www.ncbi.nlm.nih.gov/sra`) under the accession numbers: shotgun read-set SRR6131123, Hi-C libraries SRR6131122 and SRR6131124 [30]. Supporting material from a previous study used in comparison is available at `https://doi.org/10.1101/198713`.

4.9 Acknowledgements

We thank Professor Steven P. Djordjevic for his gracious 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, the QCIF and the UTS eResearch Group.

4.10 References

[1] J. T. Staley and A. Konopka, “Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats”, en, *Annu. Rev. Microbiol.*, vol. 39, pp. 321–346, 1985, ISSN: 0066-4227. DOI: 10.1146/annurev.mi.39.100185.001541. [Online]. Available: http://dx.doi.org/10.1146/annurev.mi.39.100185.001541.

[2] M. S. Rappé and S. J. Giovannoni, “The uncultured microbial majority”, en, *Annu. Rev. Microbiol.*, vol. 57, no. 1, pp. 369–394, 2003, ISSN: 0066-4227. DOI: 10.1146/annurev.micro.57.030502.090759. eprint: http://dx.doi.org/10.1146/annurev.micro.57.030502.090759. [Online]. Available: http://dx.doi.org/10.1146/annurev.micro.57.030502.090759.

[3] P. H. Janssen, P. S. Yates, B. E. Grinton, P. M. Taylor, and M. Sait, “Improved culturability of soil bacteria and isolation in pure culture of novel members of the divisions acidobacteria, actinobacteria, proteobacteria, and verrucomicrobia”, en, *Appl. Environ. Microbiol.*, vol. 68, no. 5, pp. 2391–2396, May 2002, ISSN: 0099-2240. DOI: 10.1128/AEM.68.5.2391-2396.2002. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/11976113.

[4] M. Sait, P. Hugenholtz, and P. H. Janssen, “Cultivation of globally distributed soil bacteria from phylogenetic lineages previously only detected in cultivation-independent surveys”, en, *Environ. Microbiol.*, vol. 4, no. 11, pp. 654–666, Nov. 2002, ISSN: 1462-2912. DOI: 10.1046/j.1462-2920.2002.00352.x. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/12460273.

[5] B. S. Stevenson, S. A. Eichorst, J. T. Wertz, T. M. Schmidt, and J. A. Breznak, “New strategies for cultivation and detection of previously uncultured microbes”, en, *Appl. Environ. Microbiol.*, vol. 70, no. 8, pp. 4748–4755, Aug. 2004, ISSN: 0099-2240. DOI: 10.1128/AEM.70.8.4748-4755.2004. [Online]. Available: http://dx.doi.org/10.1128/AEM.70.8.4748-4755.2004.

[6] R. A. Welch, V. Burland, G. Plunkett 3rd, P. Redford, P. Roesch, D. Rasko, E. L. Buckles, S.-R. Liou, A. Boutin, J. Hackett, D. Stroud, G. F. Mayhew, D. J. Rose, S. Zhou, D. C. Schwartz, N. T. Perna, H. L. T. Mobley, M. S. Donnenberg, and F. R. Blattner, “Extensive mosaic structure revealed by the complete genome sequence of uropathogenic escherichia coli”, en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 26, pp. 17 020–17 024, Dec. 2002, ISSN: 0027-8424. DOI: 10.1073/pnas.252529799. [Online]. Available: http://dx.doi.org/10.1073/pnas.252529799.

[7] T. O. Delmont and A. M. Eren, “Linking pangenomes and metagenomes: The prochlorococcus metapangenome”, en, *PeerJ*, vol. 6, e4320, Jan. 2018, ISSN: 2167-8359. DOI: 10.7717/peerj.4320. [Online]. Available: http://dx.doi.org/10.7717/peerj.4320.

[8] S. J. Biller, P. M. Berube, D. Lindell, and S. W. Chisholm, “Prochlorococcus: The structure and function of collective diversity”, en, *Nat. Rev. Microbiol.*, vol. 13, no. 1, pp. 13–27, Jan. 2015, ISSN: 1740-1526, 1740-1534. DOI: 10.1038/nrmicro3378. [Online]. Available: http://dx.doi.org/10.1038/nrmicro3378.

[9] J. Wiedenbeck and F. M. Cohan, “Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches”, en, *FEMS Microbiol. Rev.*, vol. 35, no. 5, pp. 957–976, Sep. 2011, ISSN: 0168-6445, 1574-6976. DOI: 10.1111/j.1574-6976.2011.00292.x. [Online]. Available: http://dx.doi.org/10.1111/j.1574-6976.2011.00292.x.

[10] D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson, “CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes”, en, *Genome Res.*, vol. 25, no. 7, pp. 1043–1055, Jul. 2015, ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.186072.114. [Online]. Available: http://dx.doi.org/10.1101/gr.186072.114.

[11] F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, and E. M. Zdobnov, “BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs”, en, *Bioinformatics*, vol. 31, no. 19, pp. 3210–3212, Oct. 2015, ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btv351. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btv351.

[12] R. M. Bowers, N. C. Kyrpides, R. Stepanauskas, M. Harmon-Smith, D. Doud, T. B. K. Reddy, F. Schulz, J. Jarett, A. R. Rivers, E. A. Eloe-Fadrosh, S. G. Tringe, N. N. Ivanova, A. Copeland, A. Clum, E. D. Becraft, R. R. Malmstrom, B. Birren, M. Podar, P. Bork, G. M. Weinstock, G. M. Garrity, J. A. Dodsworth, S. Yooseph, G. Sutton, F. O. Glöckner, J. A. Gilbert, W. C. Nelson, S. J. Hallam, S. P. Jungbluth, T. J. G. Ettema, S. Tighe, K. T. Konstantinidis, W.-T. Liu, B. J. Baker, T. Rattei, J. A. Eisen, B. Hedlund, K. D. McMahon, N. Fierer, R. Knight, R. Finn, G. Cochrane, I. Karsch-Mizrachi, G. W. Tyson, C. Rinke, Genome Standards Consortium, A. Lapidus, F. Meyer, P. Yilmaz, D. H. Parks, A. M. Eren, L. Schriml, J. F. Banfield, P. Hugenholtz, and T. Woyke, “Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea”, en, *Nat. Biotechnol.*, vol. 35, no. 8, pp. 725–731, Aug. 2017,

ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.3893. [Online]. Available: http://dx.doi.org/10.1038/nbt.3893.

[13] S. Mukherjee, D. Stamatis, J. Bertsch, G. Ovchinnikova, O. Verezemska, M. Isbandi, A. D. Thomas, R. Ali, K. Sharma, N. C. Kyrpides, and T. B. K. Reddy, “Genomes OnLine database (GOLD) v.6: Data updates and feature enhancements”, en, *Nucleic Acids Res.*, vol. 45, no. D1, pp. D446–D456, Jan. 2017, ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkw992. [Online]. Available: http://dx.doi.org/10.1093/nar/gkw992.

[14] B. Cleary, I. L. Brito, K. Huang, D. Gevers, T. Shea, S. Young, and E. J. Alm, “Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning”, en, *Nat. Biotechnol.*, vol. 33, no. 10, pp. 1053–1060, Oct. 2015, ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.3329. [Online]. Available: http://dx.doi.org/10.1038/nbt.3329.

[15] J. Alneberg, B. S. Bjarnason, I. de Bruijn, M. Schirmer, J. Quick, U. Z. Ijaz, N. J. Loman, A. F. Andersson, and C. Quince, “CONCOCT: Clustering cONtigs on COverage and ComposiTion”, Dec. 2013. arXiv: 1312.4038 [q-bio.GN]. [Online]. Available: http://arxiv.org/abs/1312.4038.

[16] M. Imelfort, D. Parks, B. J. Woodcroft, P. Dennis, P. Hugenholtz, and G. W. Tyson, “GroopM: An automated tool for the recovery of population genomes from related metagenomes”, en, *PeerJ*, vol. 2, e603, Sep. 2014, ISSN: 2167-8359. DOI: 10.7717/peerj.603. [Online]. Available: http://dx.doi.org/10.7717/peerj.603.

[17] Y. Y. Lu, T. Chen, J. A. Fuhrman, and F. Sun, “COCACOLA: Binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge”, en, *Bioinformatics*, vol. 33, no. 6, pp. 791–798, Mar. 2017, ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btw290. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btw290.

[18] Y.-W. Wu, B. A. Simmons, and S. W. Singer, “MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets”, en, *Bioinformatics*, vol. 32, no. 4, pp. 605–607, Feb. 2016, ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btv638. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btv638.

[19] D. D. Kang, J. Froula, R. Egan, and Z. Wang, “MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities”, en, *PeerJ*, vol. 3, e1165, Aug. 2015, ISSN: 2167-8359. DOI: 10.7717/peerj.1165. [Online]. Available: http://dx.doi.org/10.7717/peerj.1165.

[20] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker, "Comprehensive mapping of long-range interactions reveals folding principles of the human genome", en, *Science*, vol. 326, no. 5950, pp. 289–293, Oct. 2009, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1181369. [Online]. Available: http://dx.doi.org/10.1126/science.1181369.

[21] H. Marie-Nelly, M. Marbouty, A. Cournac, J.-F. Flot, G. Liti, D. P. Parodi, S. Syan, N. Guillén, A. Margeot, C. Zimmer, and R. Koszul, "High-quality genome (re)assembly using chromosomal contact data", en, *Nat. Commun.*, vol. 5, no. 5695, p. 5695, Dec. 2014, ISSN: 2041-1723. DOI: 10.1038/ncomms6695. [Online]. Available: http://dx.doi.org/10.1038/ncomms6695.

[22] P. Edge, V. Bafna, and V. Bansal, "HapCUT2: Robust and accurate haplotype assembly for diverse sequencing technologies", en, *Genome Res.*, vol. 27, no. 5, pp. 801–812, May 2017, ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.213462.116. [Online]. Available: http://dx.doi.org/10.1101/gr.213462.116.

[23] S. Selvaraj, J. R Dixon, V. Bansal, and B. Ren, "Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing", en, *Nat. Biotechnol.*, vol. 31, no. 12, pp. 1111–1118, Dec. 2013, ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.2728. [Online]. Available: http://dx.doi.org/10.1038/nbt.2728.

[24] J. N. Burton, A. Adey, R. P. Patwardhan, R. Qiu, J. O. Kitzman, and J. Shendure, "Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions", en, *Nat. Biotechnol.*, vol. 31, no. 12, pp. 1119–1125, Dec. 2013, ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.2727. [Online]. Available: http://dx.doi.org/10.1038/nbt.2727.

[25] N. Varoquaux, I. Liachko, F. Ay, J. N. Burton, J. Shendure, M. J. Dunham, J.-P. Vert, and W. S. Noble, "Accurate identification of centromere locations in yeast genomes using Hi-C", en, *Nucleic Acids Res.*, vol. 43, no. 11, pp. 5331–5339, Jun. 2015, ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkv424. [Online]. Available: http://dx.doi.org/10.1093/nar/gkv424.

[26] C. W. Beitel, L. Froenicke, J. M. Lang, I. F. Korf, R. W. Michelmore, J. A. Eisen, and A. E. Darling, "Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products", en, *PeerJ*, vol. 2, no. 12, e415, May 2014,

ISSN: 2167-8359. DOI: 10.7717/peerj.415. [Online]. Available: http://dx.doi.org/10.7717/peerj.415.

[27] J. N. Burton, I. Liachko, M. J. Dunham, and J. Shendure, “Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps”, en, *G3*, vol. 4, no. 7, pp. 1339–1346, May 2014, ISSN: 2160-1836. DOI: 10.1534/g3.114.011825. [Online]. Available: http://dx.doi.org/10.1534/g3.114.011825.

[28] M. Marbouty and R. Koszul, “Metagenome analysis exploiting High-Throughput chromosome conformation capture (3c) data”, en, *Trends Genet.*, vol. 31, no. 12, pp. 673–682, Dec. 2015, ISSN: 0168-9525. DOI: 10.1016/j.tig.2015.10.003. [Online]. Available: http://dx.doi.org/10.1016/j.tig.2015.10.003.

[29] M. Marbouty, L. Baudry, A. Cournac, and R. Koszul, “Meta3C analysis of a mouse gut microbiome”, en, Dec. 2015, [Online]. Available: https://www.biorxiv.org/content/early/2015/12/17/034793.

[30] M. O. Press, A. H. Wiser, Z. N. Kronenberg, K. W. Langford, M. Shakya, C.-C. Lo, K. A. Mueller, S. T. Sullivan, P. S. G. Chain, and I. Liachko, “Hi-C deconvolution of a human gut microbiome yields high-quality draft genomes and reveals plasmid-genome interactions”, en, Oct. 2017, [Online]. Available: https://www.biorxiv.org/content/early/2017/10/05/198713.

[31] J. Ghurye, A. Rhie, B. P. Walenz, A. Schmitt, S. Selvaraj, M. Pop, A. M. Phillippy, and S. Koren, “Integrating Hi-C links with assembly graphs for chromosome-scale assembly”, en, Feb. 2018, [Online]. Available: https://www.biorxiv.org/content/early/2018/02/07/261149.

[32] D. H. Parks, M. Chuvochina, D. W. Waite, C. Rinke, A. Skarshewski, P.-A. Chaumeil, and P. Hugenholtz, “A proposal for a standardized bacterial taxonomy based on genome phylogeny”, en, Jan. 2018, [Online]. Available: https://www.biorxiv.org/content/early/2018/01/31/256800.

[33] C. Jain, L. M. Rodriguez-R, A. M. Phillippy, K. T. Konstantinidis, and S. Aluru, “High-throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries”, en, Nov. 2017, [Online]. Available: https://www.biorxiv.org/content/early/2017/11/27/225342.

[34] M. Z. DeMaere and A. E. Darling, “Deconvoluting simulated metagenomes: The performance of hard- and soft- clustering algorithms applied to metagenomic chromosome conformation capture (3c)”, en, *PeerJ*, vol. 4, e2676, Nov. 2016, ISSN: 2167-8359. DOI: 10.7717/peerj.2676. [Online]. Available: http://dx.doi.org/10.7717/peerj.2676.

[35] W. Huang, L. Li, J. R. Myers, and G. T. Marth, “ART: A next-generation sequencing read simulator”, en, *Bioinformatics*, vol. 28, no. 4, pp. 593–594, Feb. 2012, ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btr708. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btr708.

[36] M. Z. DeMaere and A. E. Darling, “sim3C: Simulation of Hi-C and Meta3C proximity ligation sequencing technologies”, en, *Gigascience*, vol. 7, no. 2, Feb. 2018, ISSN: 2047-217X. DOI: 10.1093/gigascience/gix103. [Online]. Available: http://dx.doi.org/10.1093/gigascience/gix103.

[37] B. Bushnell, *BBTools*, https://www.sourceforge.net/projects/bbmap/, Accessed: 2018-5-1, Feb. 2014. [Online]. Available: https://www.sourceforge.net/projects/bbmap/.

[38] S. M. Kiełbasa, R. Wan, K. Sato, P. Horton, and M. C. Frith, “Adaptive seeds tame genomic sequence comparison”, en, *Genome Res.*, vol. 21, no. 3, pp. 487–493, Mar. 2011, ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.113985.110. [Online]. Available: http://dx.doi.org/10.1101/gr.113985.110.

[39] N. X. Vinh, J. Epps, and J. Bailey, “Information theoretic measures for clusterings comparison: Is a correction for chance necessary?”, in *Proceedings of the 26th Annual International Conference on Machine Learning*, ACM, Jun. 2009, pp. 1073–1080, ISBN: 9781605585161. DOI: 10.1145/1553374.1553511. [Online]. Available: https://dl.acm.org/citation.cfm?doid=1553374.1553511.

[40] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo, “A comparison of extrinsic clustering evaluation metrics based on formal constraints”, *Inf. Retr. Boston.*, vol. 12, no. 4, pp. 461–486, Aug. 2009, ISSN: 1386-4564, 1573-7659. DOI: 10.1007/s10791-008-9066-8. [Online]. Available: https://doi.org/10.1007/s10791-008-9066-8.

[41] M. C. P. de Souto, A. L. V. Coelho, K. Faceli, T. C. Sakata, V. Bonadia, and I. G. Costa, “A comparison of external clustering evaluation indices in the context of imbalanced data sets”, in *2012 Brazilian Symposium on Neural Networks*, Oct. 2012, pp. 49–54. DOI: 10.1109/SBRN.2012.25. [Online]. Available: http://dx.doi.org/10.1109/SBRN.2012.25.

[42] S. Nurk, D. Meleshko, A. Korobeynikov, and P. A. Pevzner, “metaSPAdes: A new versatile metagenomic assembler”, en, *Genome Res.*, vol. 27, no. 5, pp. 824–834, May 2017, ISSN: 1088-9051, 1549-5469. DOI: 10.1101/gr.213959.116. [Online]. Available: http://dx.doi.org/10.1101/gr.213959.116.

[43] H. Li, “Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM”, Mar. 2013. arXiv: 1303.3997 [q-bio.GN]. [Online]. Available: http://arxiv.org/abs/1303.3997.

[44] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup, “The sequence Alignment/Map format and SAMtools”, en, *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009, ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btp352. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btp352.

[45] P. A. Knight and D. Ruiz, “A fast algorithm for matrix balancing”, *IMA J. Numer. Anal.*, vol. 33, no. 3, pp. 1029–1047, Jul. 2013, ISSN: 0272-4979. DOI: 10.1093/imanum/drs019. [Online]. Available: https://academic.oup.com/imajna/article-abstract/33/3/1029/659457.

[46] S. Dongen, “A cluster algorithm for graphs”, Amsterdam, The Netherlands, The Netherlands, Tech. Rep., 2000. [Online]. Available: https://dl.acm.org/citation.cfm?id=868986.

[47] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks”, en, *J. Stat. Mech.*, vol. 2008, no. 10, P10008, Oct. 2008, ISSN: 1742-5468. DOI: 10.1088/1742-5468/2008/10/P10008. [Online]. Available: http://iopscience.iop.org/article/10.1088/1742-5468/2008/10/P10008/meta.

[48] S. Fortunato and M. Barthélemy, “Resolution limit in community detection”, en, *Proc. Natl. Acad. Sci. U. S. A.*, vol. 104, no. 1, pp. 36–41, Jan. 2007, ISSN: 0027-8424. DOI: 10.1073/pnas.0605965104. [Online]. Available: http://dx.doi.org/10.1073/pnas.0605965104.

[49] A. Lancichinetti and S. Fortunato, “Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities”, en, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, vol. 80, no. 1 Pt 2, p. 016 118, Jul. 2009, ISSN: 1539-3755. DOI: 10.1103/PhysRevE.80.016118. [Online]. Available: http://dx.doi.org/10.1103/PhysRevE.80.016118.

[50] S. Emmons, S. Kobourov, M. Gallant, and K. Börner, “Analysis of network clustering algorithms and cluster quality metrics at scale”, en, *PLoS One*, vol. 11, no. 7, e0159161, Jul. 2016, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0159161. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0159161.

[51] M. Rosvall, D. Axelsson, and C. T. Bergstrom, “The map equation”, Jun. 2009. arXiv: 0906.1405 [physics.soc-ph]. [Online]. Available: http://arxiv.org/abs/0906.1405.

[52] M. De Domenico, A. Lancichinetti, A. Arenas, and M. Rosvall, “Identifying modular flows on multilayer networks reveals highly overlapping organization in interconnected systems”, *Phys. Rev. X*, vol. 5, no. 1, p. 011 027, Mar. 2015. DOI: 10.1103/PhysRevX.5.011027. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevX.5.011027.

[53] T. Kawamoto and M. Rosvall, “Estimating the resolution limit of the map equation in community detection”, en, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, vol. 91, no. 1, p. 012 809, Jan. 2015, ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.91.012809. [Online]. Available: http://dx.doi.org/10.1103/PhysRevE.91.012809.

[54] A. Nakabachi, A. Yamashita, H. Toh, H. Ishikawa, H. E. Dunbar, N. A. Moran, and M. Hattori, “The 160-kilobase genome of the bacterial endosymbiont carsonella”, en, *Science*, vol. 314, no. 5797, p. 267, Oct. 2006, ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1134196. [Online]. Available: http://dx.doi.org/10.1126/science.1134196.

[55] Y.-W. Wu, Y.-H. Tang, S. G. Tringe, B. A. Simmons, and S. W. Singer, “MaxBin: An automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm”, en, *Microbiome*, vol. 2, p. 26, Aug. 2014, ISSN: 2049-2618. DOI: 10.1186/2049-2618-2-26. [Online]. Available: http://dx.doi.org/10.1186/2049-2618-2-26.

[56] A. Sczyrba, P. Hofmann, P. Belmann, D. Koslicki, S. Janssen, J. Dröge, I. Gregor, S. Majda, J. Fiedler, E. Dahms, A. Bremges, A. Fritz, R. Garrido-Oter, T. S. Jørgensen, N. Shapiro, P. D. Blood, A. Gurevich, Y. Bai, D. Turaev, M. Z. DeMaere, R. Chikhi, N. Nagarajan, C. Quince, F. Meyer, M. Balvočiūtė, L. H. Hansen, S. J. Sørensen, B. K. H. Chia, B. Denis, J. L. Froula, Z. Wang, R. Egan, D. Don Kang, J. J. Cook, C. Deltel, M. Beckstette, C. Lemaitre, P. Peterlongo, G. Rizk, D. Lavenier, Y.-W. Wu, S. W. Singer, C. Jain, M. Strous, H. Klingenberg, P. Meinicke, M. D. Barton, T. Lingner, H.-H. Lin, Y.-C. Liao, G. G. Z. Silva, D. A. Cuevas, R. A. Edwards, S. Saha, V. C. Piro, B. Y. Renard, M. Pop, H.-P. Klenk, M. Göker, N. C. Kyrpides, T. Woyke, J. A. Vorholt, P. Schulze-Lefert, E. M. Rubin, A. E. Darling, T. Rattei, and A. C. McHardy, “Critical assessment of metagenome interpretation-a benchmark of metagenomics software”, en, *Nat. Methods*, vol. 14, no. 11, pp. 1063–1071, Nov. 2017, ISSN: 1548-7091, 1548-7105. DOI: 10.1038/nmeth.4458. [Online]. Available: http://dx.doi.org/10.1038/nmeth.4458.

[57] Press, Maximilian O, A. H. Wiser, Z. N. Kronenberg, K. W. Langford, M. Shakya, C.-C. Lo, K. A. Mueller, S. T. Sullivan, P. S. G. Chain, and I. Liachko, *Supporting tables from CheckM*, Title of the publication associated with this dataset: Hi-C deconvolution of a human gut microbiome yields high-quality draft genomes and reveals plasmid-genome interactions., Oct. 2017. [Online]. Available: https://www.biorxiv.org/highwire/filestream/60380/field_highwire_adjunct_files/1/198713-2.xlsx.

4.11 Appendices

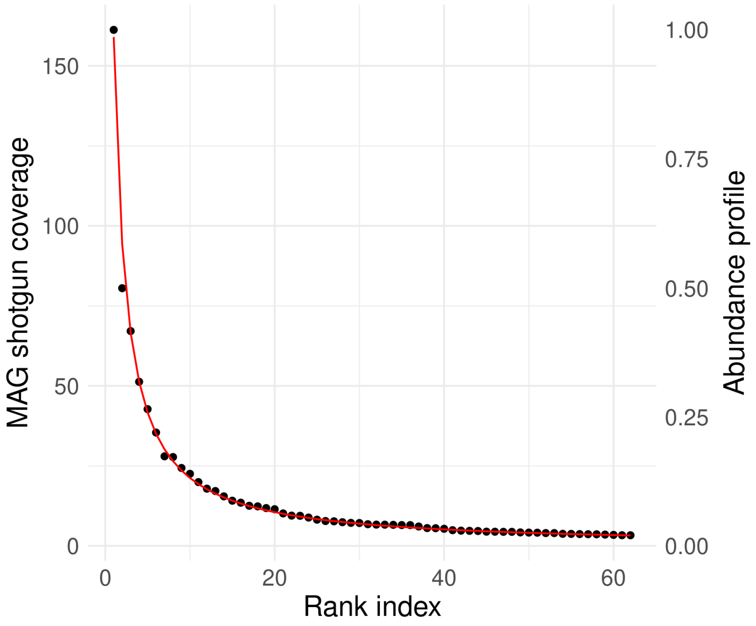

Figure 4.S1: Relative abundance of the simulated community was modelled as a Generalized Pareto distribution (red curve). After genome binning was completed, the estimated coverage of MAGs (black circles) agrees closely with the input abundances. Here we have defined the most abundant member as equal to unity.

Table 4.S1: Download GTDB metadata associated with genomes selected for the simulated community.

Table 4.S2: Download CheckM validation result for the simulated community.

Table 4.S3: Download cluster report and CheckM validation result for the real human gut microbiome.

					Simple MAG Rank				
		Nearly			Substantially			Moderately	
Statistic	**min**	**max**	**median**	**min**	**max**	**median**	**min**	**max**	**median**
Contigs	14	294	83	35	495	109	41	416	151
Coverage	5.9	447.5	34.4	4.3	416.4	22.1	3.7	83.4	21.6
N50	13,103	297,079	73,278	6,203	169,501	38,287	5,009	74,774	17,627
Extent	1,681,638	4,967,006	2,810,566	1,555,274	5,460,325	2,480,584	1,224,207	3,402,418	1,836,393
gc_expect	34.40	62.82	52.88	28.18	64.99	44.54	35.92	60.07	49.60

Table 4.S4: Summary statistics for MAGs retrieved using bin3C from a real human gut microbiome, divided into ranks as defined by Parks et al based only on completeness and contamination

			GSC MIMAG Rank			
		High quality			Medium quality	
Statistic	**min**	**max**	**median**	**min**	**max**	**median**
Contigs	27	275	85	14	495	92
Coverage	10.7	447.5	68.38	3.7	416.4	25.6
N50	31,316	221,523	75159	5,009	297,079	52,246
Extent	1,863,635	4,099,346	2549586	1,224,207	5,460,325	2,623,866

Table 4.S5: Summary statistics for MAGs retrieved using bin3C from a real human gut microbiome, divided into ranks defined by the GSC MIMAG standard.

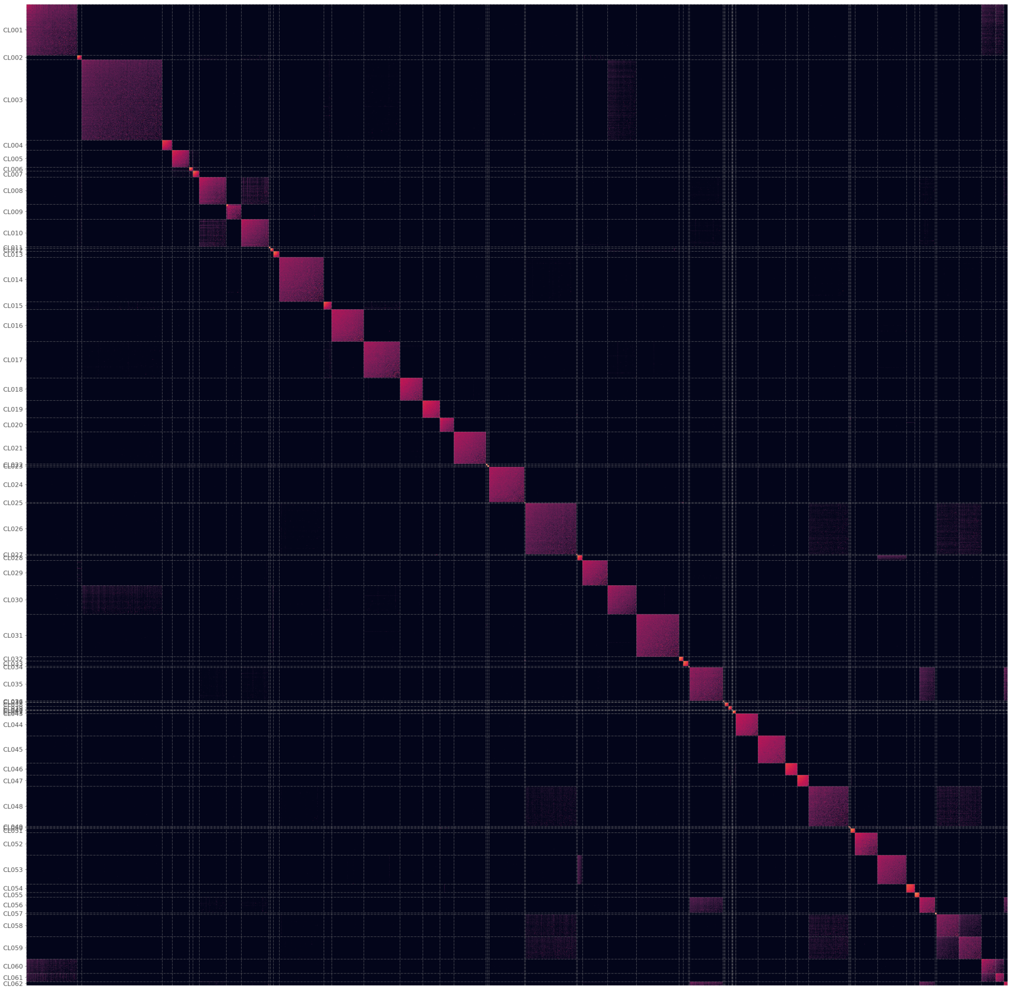

Figure 4.S2: The contact map of the simulated community at 250x shotgun and 200M Hi-C pairs. Here, intensity of a pixel is equal to the natural log of the normalized interaction strength between two contigs. When clustered, the the heatmap appears in block diagonal form, where each block represents a cluster. Each cluster is sorted largest to smallest contig, giving the impression of a gradient which is only an artefact. Blocks are proportional to the number of contigs.

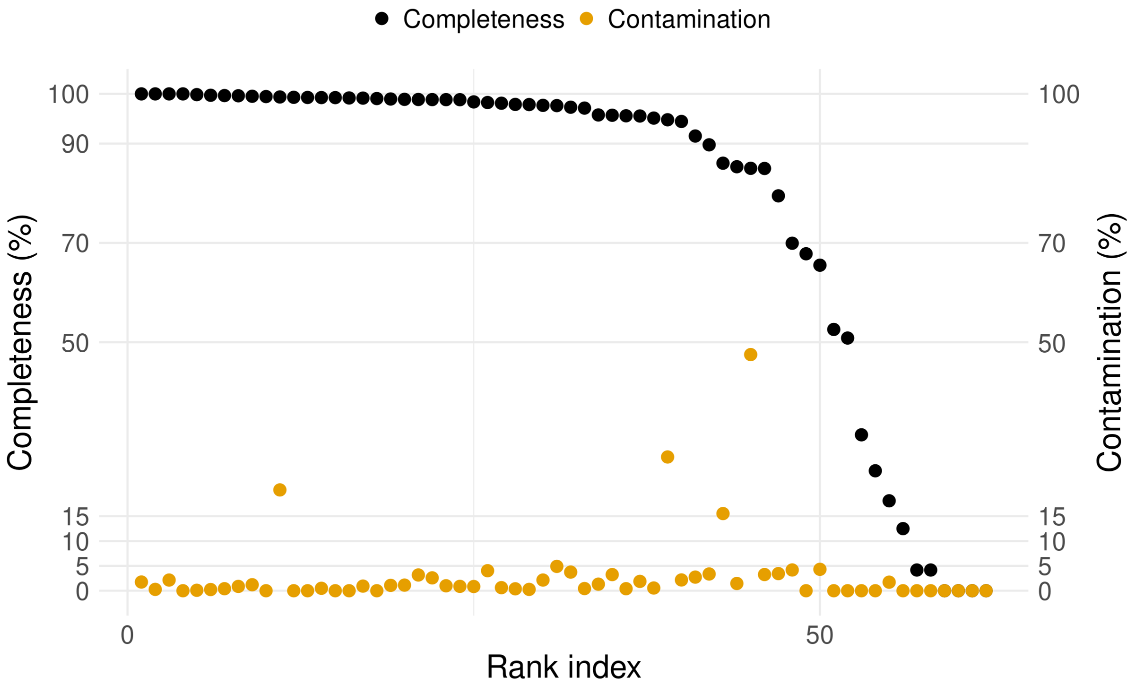

Figure 4.S3: Completeness and contamination plot of the full depth 250x/200M pair run of the simulated community. There were 62 clusters in the solution from an initial 63 genomes. Ticks along the y-axis mark thresholds used in the simple CheckM standard for MAG quality. Completeness (>90, >70, >50) and Contamination (>5, >10, >15).

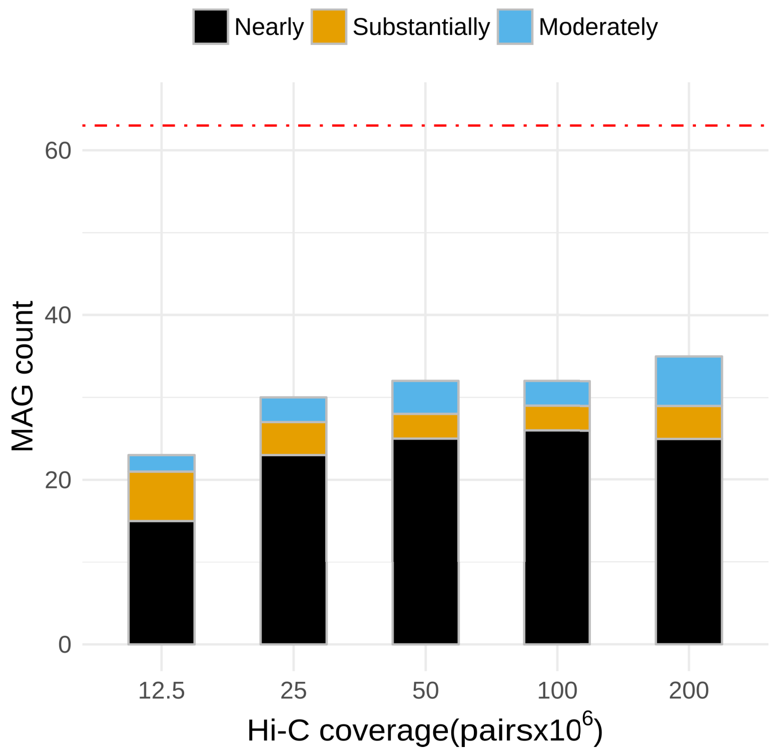

Figure 4.S4: MAGs retrieved from the simulated community when shotgun coverage is reduced by half (125x most abundant genome). At the maximum Hi-C depth of coverage, CheckM estimated that there were 25 nearly, 4 substantially and 6 moderately complete genomes.