

A peer-reviewed version of this preprint was published in PeerJ on 27 May 2014.

[View the peer-reviewed version](https://peerj.com/articles/415) (peerj.com/articles/415), which is the preferred citable publication unless you specifically need to cite this preprint.

Beitel CW, Froenicke L, Lang JM, Korf IF, Michelmore RW, Eisen JA, Darling AE. 2014. Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. PeerJ 2:e415 <https://doi.org/10.7717/peerj.415>

1 **Strain- and plasmid-level deconvolution of a synthetic metagenome by** 2 **sequencing proximity ligation products**

3
4 Christopher W. Beitel [1,2], Lutz Froenicke [1], Jenna M. Lang [1], Ian F. Korf [1,2], Richard W.
5 Micheltmore [1, 2, 3], Jonathan A. Eisen [1,4,5], and Aaron E. Darling [6]

6
7 1 The University of California, Davis Genome Center, Davis, California, United States of
8 America

9 2 Department of Molecular & Cellular Biology, University of California, Davis, California, United
10 States of America

11 3 Department of Plant Sciences, University of California, Davis, California, United States of
12 America

13 4 Department of Medical Microbiology and Immunology, University of California Davis,
14 California, United States of America

15 5 Department of Evolution and Ecology, University of California Davis, Davis, California, United
16 States of America

17 6 ithree institute, University of Technology Sydney, Sydney, NSW, Australia
18

19 20 **ABSTRACT**

21
22 Metagenomics is a valuable tool for the study of microbial communities but has been
23 limited by the difficulty of “binning” the resulting sequences into groups corresponding to
24 the individual species and strains that constitute the community. Moreover, there are
25 presently no methods to track the flow of mobile DNA elements such as plasmids
26 through communities or to determine which of these are co-localized within the same
27 cell. We address these limitations by applying Hi-C, a technology originally designed for
28 the study of three-dimensional genome structure in eukaryotes, to measure the cellular
29 co-localization of DNA sequences. We leveraged Hi-C data generated from a synthetic
30 metagenome sample to accurately cluster metagenome assembly contigs into groups
31 that contain nearly complete genomes of each species. The Hi-C data also reliably
32 associated plasmids with the chromosomes of their host and with each other. We
33 further demonstrated that Hi-C data provides a long-range signal of strain-specific
34 genotypes, indicating such data may be useful for high-resolution genotyping of
35 microbial populations. Our work demonstrates that Hi-C sequencing data provide
36 valuable information for metagenome analyses that are not currently obtainable by other
37 methods. This metagenomic Hi-C method could facilitate future studies of the fine-scale
38 population structure of microbes, as well as studies of how antibiotic resistance
39 plasmids (or other genetic elements) mobilize in microbial communities. The method is
40 not limited to microbiology; the genetic architecture of other heterogeneous populations
41 of cells could also be studied with this technique.
42

43 **INTRODUCTION**

44
45 Microbial ecology is the study of microbial communities in terms of their composition,
46 functional diversity, interactions, stability, and emergent properties (Handelsman, 2004;
47 Konopka, 2009). Knowledge of the roles microbes play in ecosystems is essential for
48 understanding how these ecosystems function (Konopka, 2009). Readily-cultivated

49 organisms are estimated to constitute less than 1% of all microbial species, leading to
50 the development of culture-independent methods for studying microbial communities
51 (Gilbert & Dupont, 2011; Hugenholtz, 2002; Staley, 1985). These culture-independent
52 methods allow communities to be characterized directly.

53 Current sequencing-based metagenomic methods do not capture some of the
54 most informative genetic information in microbial communities, in particular the long-
55 range sequence contiguity and associations of genetic material in individual cells. In
56 nearly all metagenomic methods, cells from the microbial community are lysed *en*
57 *masse* to obtain a bulk DNA sample. This results in DNA from many different cells being
58 mixed together, so that the genotype and species identity of individual cells are lost.
59 Chromosomal DNA is then fragmented into pieces (~500 bp to 40 kbp, depending on
60 the sequencing strategy), further reducing contiguity.

61 Improved sample-processing workflows might preserve this information and
62 thereby yield greater insight into the genetic structure of microbial communities. High
63 throughput single-cell genomics (e.g. applied to thousands of cells) offers an alternative
64 to shotgun metagenomics that preserves information about cellular
65 compartmentalization of genetic material. However, single-cell genomics currently has
66 its own challenges associated with isolating individual cells from certain types of
67 samples (requirement for specialized equipment, rarity of some cells within the sample)
68 and amplifying DNA from single cells (which can introduce significant amplification bias
69 and artifacts) to attain sufficient amounts of DNA for sequencing. This approach is also
70 highly sensitive to contamination by foreign DNA. Long-read technologies, such as
71 Pacific Biosciences (Eid et al., 2009) and nanopore (Maitra, Kim, & Dunbar, 2012)
72 sequencing, may help address this challenge but are still constrained by the difficulty of
73 preparing adequate amounts of very long DNA fragments.

74 Computational methods have been developed to infer genomic contiguity from
75 metagenomic data by binning metagenome assembly contigs by species. These binning
76 procedures pose a significant analytical challenge. Several methods have been
77 developed that can be divided into comparative, compositional, and assembly
78 approaches. Comparative approaches use alignments to reference sequences to assign
79 contigs to species within existing taxonomies (Droge & Mchardy, 2012). Comparative
80 approaches are limited by their reliance on existing taxonomies. Compositional
81 approaches form clusters of contigs that share similar oligomer (usually 4 bp to 8 bp)
82 composition (Droge & Mchardy, 2012). Compositional approaches tend to be limited as
83 well due to their underlying assumption that contigs with similar sequence composition
84 belong together. Horizontal gene transfer complicates both of these analysis methods
85 because it can introduce gene content from a taxonomically distant relative with unusual
86 nucleotide composition. Thirdly, assembly approaches seek to infer correct groupings of
87 assembly contigs by using paired- and long-read technologies to scaffold contigs
88 according to their original genome order, consequentially placing these sequences
89 within the same bin. These methods are limited by sequencing read lengths, which span
90 only a small fraction of a bacterial genome in a single read or read pair.

91 Binning seeks to address the challenge of determining which sequences were
92 present within cells of the same species prior to DNA extraction. We reasoned that such
93 co-localizations could be inferred from Hi-C data, a method originally developed for the
94 study of three-dimensional genome structure in eukaryotes (Lieberman-Aiden et al.,

2009). This method relies on cross-linking molecules in close physical proximity and consequently identifies both intra- and inter-chromosomal associations, reflecting the spatial arrangement of DNA at the time of cross-linking within intact nuclei or non-nucleated cells (Umbarger et al., 2011). We predicted that sequences of DNA not present in the same cell at the time of cross-linking would not be cross-linked together and should not be associated by Hi-C reads (Supplementary Figure 1). Herein we demonstrate the utility of Hi-C as a tool for addressing metagenomic binning and related problems in microbial ecology. To do so we first constructed a synthetic microbial community by culturing and mixing five organisms with available reference genomes. We then performed a metagenome assembly on sequences that were simulated *in silico* from the genomes of these organisms. Our first objective was to group these metagenome assembly contigs according to species using Hi-C reads that were generated from the synthetic microbial community. We then sought to differentiate two closely related *E. coli* strains included within this mixture. To do so we constructed contig and variant graphs and analyzed those graphs to characterize the extent to which Hi-C data might resolve the genotypes of species and strains present in our synthetic community.

MATERIALS AND METHODS

Construction of a synthetic microbial community

Pediococcus pentosaceus and *Lactobacillus brevis* were provided by the UC Davis Enology Culture Collection (<http://wineserver.ucdavis.edu>). Single colonies were used to start cultures in 5 ml liquid MRS broth. *Escherichia coli* BL21 (ATCC# PTA-5073), *E. coli* K12 DH10B (ATCC# 207214), and *Burkholderia thailandensis* (ATCC# 700388) were obtained as freeze-dried stocks from the American Type Culture Collection (ATCC). The *E. coli* strains were re-suspended in 5 ml of LB liquid medium (10g/L Tryptone, 10 g/L NaCl, 5 g/L Yeast Extract) and the *B. thailandensis* was re-suspended in 5 ml of Nutrient Broth (Peptone 15.0g/L, yeast extract 3g/L, sodium chloride 6g/L, D(+)-glucose 1g/L). All were incubated, with shaking, overnight at 37°C to produce starter cultures.

A separate 50 ml culture for each organism was created by inoculation with 10 ul of the appropriate starter culture and grown, with shaking, at 37°C, for 24 hours. The cell density of each culture was estimated by measuring the OD600. The cultures were then mixed in quantities proportional to their optical density, seeking to have equal representation of each organism in the synthetic community. Glycerol was added to a final concentration of 7% and the final mixture was divided into 2 ml tubes and frozen at -80°C.

Simulated metagenome assemblies

We simulated Illumina paired-end sequencing of the synthetic microbial community to obtain metagenomic assemblies that we could subsequently attempt to bin using experimentally derived Hi-C reads. Reads were simulated using Grinder (Angly, Willner, Rohwer, Hugenholtz, & Tyson, 2012) v0.4.5, a tool for simulating metagenomic shotgun

141 sequence reads. Replicons were sampled assuming uniform abundance of species. A
142 total of 61,063,000 million reads were simulated to cover the genomes at 500x. From
143 this set, paired-end read datasets of varying levels of coverage (100, 50, and 5x) were
144 formed. Read length was simulated to 165 bp and fragment size was simulated with a
145 normal distribution around 550 bp with a standard deviation of 50 bp (“grinder -am
146 uniform -cf \$cov -rd 165 -id 550 normal 50 -rf \$ref -fq 1 -ql 30 10 -bn
147 grinder.dp\$cov.\$tag”).

148 Assembly from each of the simulated metagenomic read sets was performed
149 using SOAPdenovo (Luo et al., 2012) with a k-mer length of 23, yielding assemblies of
150 varying quality (Supplementary Table 1). Reads were aligned to the resulting assembly
151 contigs using BWA MEM (Li, 2013). The rate of misassembly was determined by
152 mapping contigs back to the reference assemblies and counting the number of contigs
153 that joined sequences from different species.

154 **Application of Hi-C to the synthetic microbial community**

156 We performed Hi-C on the synthetic microbial community for the purpose of obtaining
157 information that could be used to group (by species) the simulation-derived contigs
158 described above, as well as to differentiate closely related strains present in our
159 synthetic community. Hi-C was carried out by combining the cross-linking and cell wall
160 digestion procedures described by (Umbarger et al., 2011) for bacterial 3-C experiments
161 and the Hi-C protocol developed for mammalian cells (Lieberman-Aiden et al., 2009)
162 with minor modifications. Cells were transferred into a 50 ml centrifuge tube and
163 washed three times in 25 ml of TE buffer (pH= 8.0) by centrifugation for 5 minutes at
164 4000 rpm at 4°C. Cells were re-suspended at an OD600 of 0.2 in TE and 37%
165 formaldehyde was added to a final concentration of 1% to cross-link proteins in the cell.
166 Cells were incubated at room temperature for 30 min and subsequently for another 30
167 min on ice (Umbarger et al., 2011). The formaldehyde was quenched by adding glycine
168 to a final concentration of 0.125 M and incubated on ice for 10 minutes. After
169 centrifugation, cells were re-suspended in TE and lysozyme digestion was carried out
170 as described to release the protein-DNA complexes (Umbarger et al., 2011). The
171 samples were centrifuged and re-suspended in Hi-C lysis buffer and incubated on ice
172 for 15 minutes (Lieberman-Aiden et al., 2009). From this step on the original Hi-C
173 protocol employing *HindIII* (Lieberman-Aiden et al., 2009) was applied with some
174 modifications. To summarize the Lieberman-Aiden et al. (2009) protocol, DNA in the
175 cross-linked protein complexes is digested with *HindIII* endonuclease following cell lysis
176 and free DNA ends are tagged with biotin. Blunt-ended DNA fragments are ligated
177 under highly dilute conditions, resulting in preferential ligation of fragments that are
178 within the same cross-linked DNA/protein complex. Next, crosslinks are removed, DNA
179 is purified, biotin is eliminated from un-ligated ends, DNA is size-selected, and ligation
180 products are selected for through a biotin pull-down. One modification we made to the
181 published procedure was to reduce the concentration of biotin-14-dCTP (Life
182 Technologies) by half. Also, ligation to Illumina-compatible sequencing adapters
183 (Biooscientific) was carried out in solution before capture with streptavidin beads. The
184 DNA sample was size selected after end-repair and before adapter ligation by gel
185 extraction for fragment sizes ranging from 280 to 420 bp. The bead-captured Hi-C
186

187 library was amplified by 10 cycles of PCR before a final cleanup with Ampure XP beads
188 (Agencourt). The library was sequenced in a single run on an Illumina Miseq machine
189 using 160 bp paired-end reads.

190

191 **Sequence alignment and quality filtering**

192

193 Reference assembly sequences were obtained from the NCBI RefSeq database (Pruitt,
194 Tatusova, Brown, & Maglott, 2012) with the following accession numbers for each of *P.*
195 *pentosaceus* (NC_008525), *L. brevis* (NC_008497, NC_008498, NC_008499), *E. coli*
196 BL21 (NC_012892), *E. coli* K12 DH10B (NC_010473), and *B. thailandensis*
197 (NC_007651, NC_007650). These sequences were pooled into a single reference
198 database for sequence alignment. No quality filtering was performed on raw reads,
199 leaving this to be performed later using alignment quality scores. Split-read sequence
200 alignment was performed (independently for each read in a pair) using BWA MEM (Li,
201 2013; default parameters) against the pooled reference assemblies as well as
202 (separately) against the metagenome assembly described above. Heat map
203 visualizations and insert distribution plots were generated from unfiltered alignments
204 using custom R (The R Core Development Team, 2010) scripts.

205 We investigated the effect of various alignment filtering parameters on the
206 subsequent variant graph analysis (illustration of the concept of a variant graph can be
207 found in Supplementary Figure 7). To this end, alignments of Hi-C reads to the
208 reference genomes were filtered according to 24 parameter combinations, with three
209 minimum mapping quality (0, 20, 60), two CIGAR filtering (none, CIGAR=160M), and
210 four minimum insert filtering (0, 1kb, 10kb, 40kb) conditions. In the latter, Hi-C read
211 pairs were excluded when their alignments within the reference assembly had an insert
212 size below the specified minimum, including cases of alignments spanning the
213 linearization points of these assemblies (e.g. for *E. coli*, near coordinates 0 and
214 4686137). An open-source graph visualization tool, Gephi (0.8.2-beta), was used to
215 visualize the Hi-C contig association network (Bastian et al., 2009).

216

217 **Contig clustering**

218 We inferred grouping of metagenome assembly contigs by applying the Markov
219 Clustering Algorithm (MCL) to a matrix of contig association data (van Dongen, 2000).
220 MCL is an unsupervised clustering algorithm which simulates flow and accumulation of
221 edge weights within a given weighted graph structure. It has a computational complexity
222 of $O(Nk^2)$ given an implementation designed for sparse matrices. The matrix of edge
223 weights provided to this algorithm was computed from contig association counts by
224 normalizing edge weights according to the following formula, which corrects for the
225 expected inflation of association between large contigs. Formally, given contig length
226 set L , the raw contact count c_{ij} between each pair of contigs $\{i, j\}$ was normalized by the
227 ratio of the square of the maximum contig length and the lengths l_i and l_j of contigs i and
228 j , respectively.

$$c'_{ij} = \frac{\max(L)^2 c_{i,j}}{l_i l_j}$$

229

230 Prior to normalization, we filtered the contig association data for (1) contig associations
231 greater than some minimum k , and (2) associations between contigs of size greater

232 than L. We explored the (k,L) parameter space by performing normalization and MCL
233 clustering for 205 parameter combinations, with 41 contig size minimums chosen evenly
234 across [0,40000] and five contact minimums, {0,3,5,7,9}. For each of the 205 filtering
235 parameter combinations, clustering was performed using 100 different MCL inflation
236 values chosen to span the interval [1,2] in increments of 0.01.

237

238 **Assessment of clustering quality**

239

240 Each metagenome assembly contig was aligned to the reference assemblies to
241 determine its species or strain of origin, allowing us to determine which strains were
242 present in each cluster of contigs. A measure of clustering quality was computed by
243 sampling random pairs of contigs (N=100,000 pairs) and comparing their cluster
244 assignments to their species of origin. This random sampling was weighted according to
245 the size of each contig such that the probability of sampling any contig from the set was
246 equal to the size of that contig divided by the sum total of all contig sizes. Any contig not
247 present in the clustering solution (but which was present as input for the clustering run
248 that generated that solution) was added to the solution in a singleton cluster. For each
249 sampled contig pair, if those contigs belonged to the same species and had been
250 placed in the same cluster, they were counted as a true positive (“TP”). If they
251 originated from the same species but had been placed in different clusters, they were
252 counted as a false negative (“FN”). Likewise, contigs originating from different species
253 that were placed in the same or different clusters were counted as false positives (“FP”)
254 and true negatives (“TN”), respectively. The true positive rate (a.k.a. sensitivity, recall),
255 false positive rate, positive predictive value (a.k.a. precision), and negative predictive
256 value were calculated from these counts according to standard formulae. All clustering
257 quality measures were computed in two ways, one treating the two *E. coli* strains as
258 independent classes (strain-level) and another treating them as the same (species-
259 level).

260

261 **Analysis of SNP graph connectivity**

262

263 A SNP graph is an undirected graph wherein SNP sites are nodes and edges link pairs
264 of SNP sites that were observed together in a sequence read pair. We expected SNP
265 graphs that were constructed using Hi-C data to be more densely connected than SNP
266 graphs that were constructed using mate-pair data because Hi-C read pairs can span
267 entire chromosomes while reads from mate-pair libraries span no more than 40 kb +/- 5
268 kb. In this way, Hi-C provides global information while that provided by mate pairs is
269 locally constrained. To quantify this, SNP graphs were constructed for the Hi-C data and
270 simulated mate-pair data and the shortest path between randomly chosen SNP pairs
271 was plotted relative to the distance between the those variants within the reference
272 assembly. These graphs were constructed from alignments of reads to the *E. coli* K12
273 reference assembly after it had been masked at variant positions identified through
274 pairwise sequence alignment of *E. coli* K12 and BL21 using progressiveMauve (Darling,
275 Mau, & Perna, 2010). In this way, the masked K12 reference assembly was used as a
276 scaffold for our analysis of Hi-C and mate-pair variant graphs.

277 Hi-C read datasets typically contain a mixture of reads derived from ligation and
278 non-ligation products, the latter having short inserts. In our analysis of SNP graph
279 connectivity we sought to understand the contribution of these non-ligation products on
280 the connectivity gains seen with Hi-C reads over mate-pairs. For comparison to our Hi-C
281 reads, mate-pair read sets were computationally simulated for a range of sizes (5 kb, 10
282 kb, 20 kb, 40 kb). These were compared to three Hi-C read sets: the entire Hi-C
283 dataset, Hi-C reads with inserts below 1kb, and Hi-C reads with inserts above 1 kb.

284 SNP graph connectivity was analyzed using the simulated mate-pair read sets
285 described above combined with Hi-C reads aligned and filtered for alignment qualities
286 above or equal to 60 and for CIGAR encodings of 160M Read pairs with both ends
287 aligning to SNP positions were identified and for each corresponding SNP pair an edge
288 was formed. Shortest path lengths between sampled SNP positions were computed
289 using a custom breadth-first search program, relying on the Boost Graph Library
290 (<http://boost.org>). The program constructs a graph from a user-specified SNP edge list
291 and performs a breadth-first search to identify the shortest path length between a user-
292 specified number of randomly selected SNP pairs. We calculated path lengths between
293 10,000 randomly chosen SNP pairs. To aid in visualization, the full range of variant
294 separation distances was divided into 20 kb segments and the average path length was
295 computed for each segment. These data were smoothed using locally-weighted
296 scatterplot smoothing (LOWESS).

297 298 **RESULTS**

299 **The synthetic microbial community for metagenomic Hi-C**

300
301
302 Five microorganisms were chosen to test the metagenomic Hi-C approach:
303 *Lactobacillus brevis*, *Pediococcus pentosaceus*, *Burkholderia thailandensis*, *Escherichia*
304 *coli* K12 DH10B and *E. coli* BL21 (DE3). These were selected because high quality
305 reference genomes are available. In addition, the multiple replicons of *B. thailandensis*
306 and plasmids present in *L. brevis* allowed us to explore whether Hi-C might link
307 separate replicons present in the same cell. We selected two strains of *E. coli* (K12 and
308 BL21) to evaluate whether Hi-C sequence data could be used to resolve inter-strain
309 differences. Genome alignment of these two *E. coli* shows that 87.9% of their genomes
310 can be aligned and that the average nucleotide identity across aligned regions is 99.5%.
311 Finally, because differences in cell membrane structure and GC content could
312 potentially lead to bias in DNA extraction and/or cross-linking efficiency, we selected
313 two lactic acid bacteria (*P. pentosaceus* and *L. brevis*), which are low-GC, Gram-
314 positive organisms for which only 39.5% of their genomes can be aligned with 84.3%
315 average nucleotide identity across aligned regions.
316

317 318 **Metagenome assembly**

319
320 We generated a metagenome assembly of the synthetic microbial community that we
321 could use as input for our analysis of the utility of Hi-C for species clustering. Hi-C
322 sequencing data is biased by the distribution of restriction sites for the restriction

323 enzyme used to construct the library as well as by other factors including GC content,
324 restriction fragment length, and “mappability” (Yaffe & Tanay, 2011). Hi-C data contain
325 numerous chimeric sequences and thus are not suitable for *de novo* contig assembly.
326 Therefore, we simulated and assembled Illumina metagenomic sequence data at
327 varying coverage levels to yield the assemblies summarized in (Supplementary Table
328 1). The size (bp) of each of these assemblies was approximately 77% of the sum of the
329 synthetic community reference genome sizes and this fraction did not change when
330 increasing the quantity of input reads from 5x to 100x. Alignment of assembled contigs
331 to the collection of reference genomes indicates that similar regions of the two *E. coli*
332 genomes were co-assembled into single contigs (data not shown). The assembly on the
333 lowest amount of input sequence (5x coverage) contained two misassembled contigs
334 and three misassembled scaffolds. Assemblies at 50x and 100x coverage were free
335 from misassembled contigs and scaffolds and were similar in terms of their contig
336 counts, N50s, and total amounts of sequence assembled. We used the 100x coverage
337 assembly (SOAP-3) for all further analysis.
338

339 Hi-C Library Statistics

340

341 A total of 20,623,187 read pairs were obtained from a single MiSeq run to yield ~6.4 Gb
342 of raw sequence data. Of these, 98.25% could be aligned back to the reference
343 genomes by BWA MEM. A total of 21,260,753 (51.55% of original and 52.46% of raw
344 aligned reads) read pairs were retained after filtering for both reads in the pair aligning
345 at high quality (MapQ \geq 60) and in full (CIGAR=160M). Due to either self-ligation or
346 imperfect enrichment for ligation junctions, some of the reads present in the dataset
347 represent local genomic DNA fragments. Therefore we classify read pairs mapping
348 within 1,000 nt as fragment reads, while all other reads are considered to be Hi-C reads
349 (3% of reads). The abundance of each replicon was estimated using filtered alignments
350 and unfiltered alignments (Table 1), as well as by normalizing each unfiltered alignment
351 count with the restriction site counts for each replicon (Supplementary Table 2). These
352 figures can only be used as approximate measures of abundance because these values
353 are affected by the frequency of restriction sites in each organism and a multitude of
354 other confounding factors (Morgan, Darling, & Eisen, 2010). Insert distances derived
355 from the alignment of Hi-C reads to the *E. coli* K12 genome were distributed in a similar
356 manner as previously reported (Figure 1; Lieberman-Aiden et al., 2009). We observed a
357 minor depletion of alignments spanning the linearization point of the *E. coli* K12
358 assembly (e.g. near coordinates 0 and 4686137) due to edge effects induced by BWA
359 treating the sequence as a linear chromosome rather than circular.
360

361 Clustering contigs by species with Hi-C

362

363 The experimentally derived Hi-C read pairs have a long tail in their insert
364 distribution (Figure 1; Supplementary Figures 3-6) indicating that they provide
365 information that can be used to link metagenome assembly contigs originating from
366 distant parts of the same chromosome. We evaluated whether Hi-C reads could be
367 used to group the simulated assembly contigs described above into clusters that

368 correspond with each species' genome. We tested this process on the SOAP-3
369 assembly, using only contigs with a length of at least 5 kbp. This threshold was applied
370 to exclude short contigs that may not have a *HindIII* restriction site. As *HindIII*
371 recognizes a 6bp motif, it cuts on average every 4,096bp. We note that >25% of a 4Mbp
372 genome is expected to have inter-site distances > 10Kbp in simulations that treat 6-
373 cutter restriction sites as uniformly distributed (data not shown) suggesting that many of
374 the contigs < 10Kbp may cluster poorly due to lack of restriction sites. The dataset was
375 further reduced to exclude links among contig pairs which are associated by 5 or fewer
376 read pairs.

377 We tested Markov clustering (van Dongen, 2000) on these data over a range of
378 inflation (affecting cluster solution granularity) parameters (Table 2). In the best case,
379 Markov clustering produced four clusters, each of which correspond to the nearly
380 complete genome of a species in our synthetic community. In this clustering, the two
381 strains of *E. coli* appear in the same group. When using the default inflation parameter
382 of 2.0 we find that the data is under-clustered, but there are no false positive
383 associations among contigs for this choice of input.

384 To further understand the sensitivity of MCL to choices of filtering and inflation
385 parameters, we performed clustering across the 204 filtering and 100 inflation (total of
386 20,400) conditions (see Methods: Contig Clustering) using MCL. A representative
387 subset of all parameter combinations tested is shown in Supplementary Figure 2. These
388 data suggest that once a sufficient contact and contig size minimums have been
389 applied, cluster solutions vary primarily in terms of their granularity (as the inflation
390 parameter varies), not their PPV (remaining close to 1) or FPR (remaining close to 0).
391 Low inflation values, close to 1, give clustering solutions with the highest TPR's, but this
392 does not hold true without sufficient filtering.

394 **Association of species with metagenomic Hi-C data**

396 We next sought to quantify the cellular co-localization signal underlying the above-
397 described species clustering. For this analysis we studied Hi-C reads aligned directly to
398 the reference assemblies of the members of our synthetic microbial community with the
399 same alignment parameters as were used in the top ranked clustering (described
400 above). We first counted the number of Hi-C reads associating each reference
401 assembly replicon (Figure 2; Supplementary Table 3), observing that Hi-C data
402 associated replicons within the same species (cell) orders of magnitude more frequently
403 than it associated replicons from different species. The rate of within-species
404 association was 98.8% when ignoring read pairs mapping less than 1,000 bp apart.
405 Including read pairs < 1,000 bp inflated this figure to 99.97%. Figure 3 illustrates this by
406 visualizing the graph of contigs and their associations. Similarly, for the two *E. coli*
407 strains (K12, BL21) we observed the rate of within-strain association to be 96.36%.
408 When evaluated on genes unique to each strain (where read mapping to each strain
409 would be unambiguous), the self-association rate was observed to be >99%.

410 We observed that the rate of association of *L. brevis* plasmids 1 and 2 with each
411 other and with the *L. brevis* chromosome was at least 100-fold higher than with the
412 other constituents of the synthetic community (Figure 2). Chromosome and plasmid Hi-
413 C contact maps show that the plasmids associate with sequences throughout the *L.*

414 *brevis* chromosome (Figure 4; Supplementary Figures 3,4,5) and exhibit the expected
415 enrichment near restriction sites. This demonstrates that metagenomic Hi-C can be
416 used to associate plasmids to specific strains in microbial communities as well as to
417 determine cell co-localization of plasmids with one another.

418 419 **Variant graph connectedness**

420
421 Algorithms that reconstruct single-molecule genotypes from samples containing two or
422 more closely-related strains or chromosomal haplotypes depend on reads or read pairs
423 that indicate whether pairs of variants coexist in the same DNA molecule. Such
424 algorithms typically represent the reads and variant sites as a variant graph wherein
425 variant sites are represented as nodes, and sequence reads define edges between
426 variant sites observed in the same read (or read pair). We reasoned that variant graphs
427 constructed from Hi-C data would have much greater connectivity (where connectivity is
428 defined as the mean path length between randomly sampled variant positions) than
429 graphs constructed from mate-pair sequencing data, simply because Hi-C inserts span
430 megabase distances. Such connectivity should, in theory, enable more accurate
431 reconstruction of single-molecule genotypes from smaller amounts of data.
432 Furthermore, by linking distant sites with fewer intermediate nodes in the graph,
433 estimates of linkage disequilibrium at distant sites (from a mixed population) are likely to
434 have greater precision.

435 To evaluate whether Hi-C produces more connected variant graphs we
436 compared the connectivity of variant graphs constructed from Hi-C data to those
437 constructed from simulated mate-pair data (with average inserts of 5 kb, 10 kb, 20 kb,
438 and 40 kb). To exclude paired-end products from the analysis, Hi-C reads with inserts
439 under 1 kb were excluded from the analysis. For each variant graph constructed from
440 these inputs, 10,000 variant position pairs were sampled at random, with 94.75% and
441 100% of these pairs belonging to the same connected graph component of the Hi-C and
442 40 kb variant graphs, respectively. These rates fell to 6.21%, 16.6%, and 32.38% for the
443 5 kb, 10 kb, and 20 kb mate-pair variant graphs, respectively (Table 3).

444 Across conditions, variant graphs differed in terms of their connectivity, with Hi-C
445 graphs showing the greatest connectivity. Despite having simulated an equal number of
446 reads for each mate-pair distance, the numbers of variant positions linked by such
447 reads was different across conditions. We observed that the variant graph derived from
448 Hi-C data (>1 kb inserts, no alignment filtering), despite having the lowest number of
449 variant links, had the lowest mean and maximum path length (5.47, 11; Table 3). Path
450 length was not correlated with distance within Hi-C variant graphs, in contrast to the
451 mate-pair conditions (Figure 5). The lengths of paths between variant pairs in the mate-
452 pair graphs did increase with distance, reaching maximums of 71, 96, 94, and 111 in the
453 5 kb, 10 kb, 20 kb, and 40 kb cases, respectively. We further examined the effect of
454 alignment quality and completeness filtering and observed that in the latter case such
455 filtering vastly reduced the rate at which variant positions occur within the same
456 connected graph component.

457 458 **DISCUSSION**

459

460 This study demonstrates that Hi-C sequencing data provide valuable information
461 for metagenome analyses that are not currently obtainable by other methods. By
462 applying Hi-C to a synthetic microbial community we showed that genomic DNA was
463 associated by Hi-C read pairs within strains orders of magnitude more frequently than
464 between strains. Hi-C reads associated genomic regions at distances not achievable
465 with mate-pair or long-read sequencing technologies. The long-range contiguity
466 information provided by Hi-C reads enabled us to perform species-level clustering of
467 metagenome assembly contigs with perfect precision and recall scores when the input
468 had been filtered sufficiently. We performed an exploration of the clustering parameter
469 space to understand the factors affecting clustering quality and identified a number of
470 key filtering parameters. Optimal filtering involved retention of only contigs that are large
471 enough to contain (or occur near) a *HindIII* restriction site and furthermore to remove
472 low-frequency contig associations that constitute a form of “noise”. Additional work is
473 needed to develop methods to identify and remove “noise” from Hi-C datasets. Lastly,
474 we compared the connectivity of variant graphs constructed from mate-pair and Hi-C
475 read datasets, observing much greater connectivity in the latter case, illustrating the
476 global nature of the Hi-C signal.

477 We also observed orders of magnitude greater rates of association between
478 plasmids and chromosomes of their hosts than between plasmids and the genomes of
479 other species. This indicates Hi-C can be used to study horizontal gene transfer. Given
480 a metagenome assembly, Hi-C provides a means to link plasmid sequences to
481 chromosomes of the host strain, and may provide the means to detect cases where
482 plasmids have been transferred among co-existing species of bacteria. We have thus
483 far demonstrated that Hi-C provides a signal of cell co-localization for the two plasmids
484 present within the *L. brevis* genome. Alternative methods do not allow identification of
485 which cells in a microbial community harbor such mobile DNA elements. Hi-C data has
486 the potential to help quantify the dynamics of horizontal gene transfer and help
487 characterize the spread of antibiotic resistance and virulence factors. It remains to be
488 determined whether this signal will be sufficient to localize small, low-copy, or highly
489 variable mobile elements within the species that contain them.

490 The resolving power of Hi-C and related methods such as ChIA-PET (Fullwood
491 et al., 2009) when applied to complex natural microbial communities is as of yet
492 undetermined. In principle, as the number of species and genotypes in a community
493 grows the amount of sequence data required to resolve species and strains also grows.
494 This challenge is common to all metagenomic approaches and is not specific to the Hi-
495 C method described. Improvements on metagenomic analysis of complex communities
496 may require integration of Hi-C data with other information sources such as sequence
497 composition, phylogeny, and measurements of abundance.

498 The problems of differentiating contigs originating from different species is similar
499 to that of differentiating contigs originating from different chromosomes of the same
500 species. Recently a study reported the use of Hi-C to perform genome scaffolding of
501 several individual eukaryotic genomes, first by inferring chromosomal groupings of
502 contigs and then ordering sequences along the chromosome (Burton et al., 2013).
503 Markov Clustering of Hi-C association data may be used to cluster contigs into
504 chromosomal groups without specifying the number of chromosomes *a priori*. This may
505 be important for samples where the number of chromosomes (e.g. tumor samples),

506 species, or species abundances (e.g. environmentally-isolated microbial communities)
507 are not known.

508 Hi-C analysis can be applied to communities other than environmentally-isolated
509 microbial communities, such as pools of BAC clones. Heterogeneous tumor populations
510 are analogous in some ways to microbial communities and Hi-C may be applied to
511 identify sub-populations therein. The problem of resolving the membership of variants in
512 closely related strains (between different cells) shares some common features with the
513 problem of differentiating closely related haplotypes within polyploid eukaryotic
514 genomes (within the same cells). Recent work has demonstrated that Hi-C data can be
515 used to phase haplotypes in a diploid organism (Selvaraj, R Dixon, Bansal, & Ren,
516 2013). Our analysis indicates that the average degree of separation between variants
517 within a Hi-C variant graph is dramatically lower than that in mate-pair variant graphs.
518 This is significant because as the degree of separation between distant graph regions
519 grows, error is compounded and the reliability of inferences regarding the phase of
520 these regions declines. Our analysis thus indicates that Hi-C data provide an
521 informative signal for the analysis of haplotype and strain mixtures.

522 **AVAILABILITY**

524 Raw sequence data for this experiment has been deposited to the NCBI Short Read
525 Archive, accession SRX377733.

526 **AUTHOR CONTRIBUTIONS**

528 Prepared the mixture: J.M.L. Prepared Hi-C Libraries: L.F. and C.B. Analyzed the data
529 C.B. and A.D. Wrote the paper: All. Conceived the method: C.B. and A.D. Designed the
530 experiment: All.

531 **ACKNOWLEDGEMENTS**

532 We wish to acknowledge the substantial efforts of Michael Lewis, administrator of the
533 UC Davis Genome Center Cluster Computing Resource, to provide and maintain the
534 computing resources on which these analyses were performed. We would also like to
535 acknowledge Matthew DeMaere (UTS) for his contribution to Figure 3.

537 **FUNDING STATEMENT**

538 This work was supported by a gift from MARS, Inc. and by Department of Homeland
539 Security contract #HSHQDC-11-C-00091.

541 **REFERENCES**

- 543 Angly, F. E., Willner, D., Rohwer, F., Hugenholtz, P., & Tyson, G. W. (2012). Grinder: a
544 versatile amplicon and shotgun sequence simulator. *Nucleic acids research*,
545 *40*(12), e94. doi:10.1093/nar/gks251
- 546 Avis, D. (1978). Two Greedy Heuristics for the Weighted Matching Problem.
547 *Proceedings of the 9th S-E Conference on Combinatorics, Graph Theory, and*
548 *Computing* (pp. 55–76).
- 549 Bastian, M., Heymann, S. & Jacomy, M. (2009). Gephi: An Open Source Software for

- 550 Exploring and Manipulating Networks.. In E. Adar, M. Hurst, T. Finin, N. S. Glance,
551 N. Nicolov & B. L. Tseng (eds.), *ICWSM*, : The AAAI Press. ISBN: 978-1-57735-
552 421-5
- 553 Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., & Shendure, J.
554 (2013). Chromosome-scale scaffolding of de novo genome assemblies based on
555 chromatin interactions. *Nature biotechnology*, (November). doi:10.1038/nbt.2727
- 556 Darling, A. E., Mau, B., & Perna, N. T. (2010). progressiveMauve: multiple genome
557 alignment with gene gain, loss and rearrangement. *PloS one*, 5(6), e11147.
558 doi:10.1371/journal.pone.0011147
- 559 Droge, J., & Mchardy, A. C. (2012). Taxonomic binning of metagenome samples
560 generated by next-generation sequencing technologies, 13(6), 646–655.
561 doi:10.1093/bib/bbs031
- 562 Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., et al. (2009). Real-
563 Time DNA Sequencing from Single Polymerase Molecules, (January), 133–138.
- 564 Ellson, J., Gansner, E. R., Koutsofios, E., North, S. C., & Woodhull, G. (2003). Graphviz
565 and Dynagraph – Static and Dynamic Graph Drawing Tools.
- 566 Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. Bin, Orlov, Y. L., et
567 al. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome.
568 *Nature*, 462(7269), 58–64. doi:10.1038/nature08497
- 569 Gilbert, J. a., & Dupont, C. L. (2011). Microbial Metagenomics: Beyond the Genome.
570 *Annual Review of Marine Science*, 3(1), 347–371. doi:10.1146/annurev-marine-
571 120709-142811
- 572 Handelsman, J. (2004). Metagenomics : Application of Genomics to Uncultured
573 Microorganisms Metagenomics : Application of Genomics to Uncultured
574 Microorganisms, 68(4). doi:10.1128/MBR.68.4.669
- 575 Hugenholtz, P. (2002). Exploring prokaryotic diversity in the genomic era. *Genome*
576 *biology*, 3(2), REVIEWS0003.
- 577 Konopka, A. (2009). What is microbial community ecology? *The ISME journal*, 3(11),
578 1223–30. doi:10.1038/ismej.2009.88
- 579 Li, H. (2013). Aligning sequence reads , clone sequences and assembly contigs with
580 BWA-MEM, 00(00), 1–3.
- 581 Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T.,
582 Telling, A., Amit, I., et al. (2009). Comprehensive mapping of long-range
583 interactions reveals folding principles of the human genome. *Science (New York,*
584 *N.Y.)*, 326(5950), 289–93. doi:10.1126/science.1181369
- 585 Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., et al. (2012).
586 SOAPdenovo2: an empirically improved memory-efficient short-read de novo
587 assembler. *GigaScience*, 1(1), 18. doi:10.1186/2047-217X-1-18
- 588 Maitra, R. D., Kim, J., & Dunbar, W. B. (2012). Recent advances in nanopore
589 sequencing. *Electrophoresis*, 33(23), 3418–28. doi:10.1002/elps.201200272

- 590 Morgan, J. L., Darling, A. E., & Eisen, J. a. (2010). Metagenomic sequencing of an in
591 vitro-simulated microbial community. *PloS one*, 5(4), e10209.
592 doi:10.1371/journal.pone.0010209
- 593 Pruitt, K. D., Tatusova, T., Brown, G. R., & Maglott, D. R. (2012). NCBI Reference
594 Sequences (RefSeq): current status, new features and genome annotation policy.
595 *Nucleic acids research*, 40(Database issue), D130–5. doi:10.1093/nar/gkr1079
- 596 Selvaraj, S., R Dixon, J., Bansal, V., & Ren, B. (2013). Whole-genome haplotype
597 reconstruction using proximity-ligation and shotgun sequencing. *Nature*
598 *biotechnology*, (November). doi:10.1038/nbt.2728
- 599 Staley, J. T. (1985). Microorganisms in aquatic and terrestrial habitats. *Annual Review*
600 *of Microbiology*. 10.1146/annurev.mi.39.100185.001541
- 601 van Dongen, S. (2000), Graph Clustering by Flow Simulation. *PhD thesis, University of*
602 *Utrecht* (May).
- 603 The R Core Development Team. (2010). *R : A Language and Environment for Statistical*
604 *Computing* (Vol. 1). Vienna, Austria.: R Foundation for Statistical Computing.
605 Retrieved from <http://www.r-project.org/>
- 606 Umbarger, M. a, Toro, E., Wright, M. a, Porreca, G. J., Baù, D., Hong, S.-H., Fero, M.
607 J., et al. (2011). The three-dimensional architecture of a bacterial genome and its
608 alteration by genetic perturbation. *Molecular cell*, 44(2), 252–64.
609 doi:10.1016/j.molcel.2011.09.010
- 610 Yaffe, E., & Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates
611 systematic biases to characterize global chromosomal architecture. *Nature*
612 *genetics*, 43(11), 1059–65. doi:10.1038/ng.947
- 613
- 614

615
616
617
618
619**TABLES**

Sequence	Alignment	% of Total	Filtered	% of Aligned	Length	GC	#R.S.
<i>Lac0</i>	10603204	26.17%	10269562	96.85%	2291220	0.462	629
<i>Lac1</i>	145718	0.36%	145478	99.84%	13413	0.386	3
<i>Lac2</i>	691723	1.71%	665825	96.26%	35595	0.385	16
<i>Lac</i>	11440645	28.23%	11080865	96.86%	2340228	0.46	648
<i>Ped</i>	2084595	5.14%	2022870	97.04%	1832387	0.373	863
<i>BL21</i>	12882177	31.79%	2676458	20.78%	4558953	0.508	508
<i>K12</i>	9693726	23.92%	1218281	12.57%	4686137	0.507	568
<i>E. coli</i>	22575903	55.71%	3894739	17.25%	9245090	0.51	1076
<i>Bur1</i>	1886054	4.65%	1797745	95.32%	2914771	0.68	144
<i>Bur2</i>	2536569	6.26%	2464534	97.16%	3809201	0.672	225
<i>Bur</i>	4422623	10.91%	4262279	96.37%	6723972	0.68	369

620

621 Table 1. Species alignment fractions. The number of reads aligning to each replicon
 622 present in the synthetic microbial community are shown before and after filtering, along
 623 with the percent of total constituted by each species. The GC content (“GC”) and
 624 restriction site counts (“#R.S.”) of each replicon, species, and strain are shown. Bur1: *B.*
 625 *thailandensis* chromosome 1. Bur2: *B. thailandensis* chromosome 2. Lac0: *L. brevis*
 626 chromosome, Lac1: *L. brevis* plasmid 1, Lac2: *L. brevis* plasmid 2, Ped: *P.*
 627 *pentosaceus*, K12: *E. coli* K12 DH10B, BL21: *E. coli* BL21. An expanded version of this
 628 table can be found in Supplementary Table 2.

629

630

Inflation	Precision	Recall	# clusters
2.0	1	0.19	33
1.3	1	0.33	25
1.125	1	0.98	5
1.1	0.96	0.98	4

631

632 Table 2. Markov clustering of metagenome assembly contigs using Hi-C data. A range
633 of inflation parameters were applied, and the precision and recall for the resulting
634 clusters was calculated as described in the text. An inflation parameter of 1.1 produced
635 a near perfect clustering of contigs by species.

636

637

638

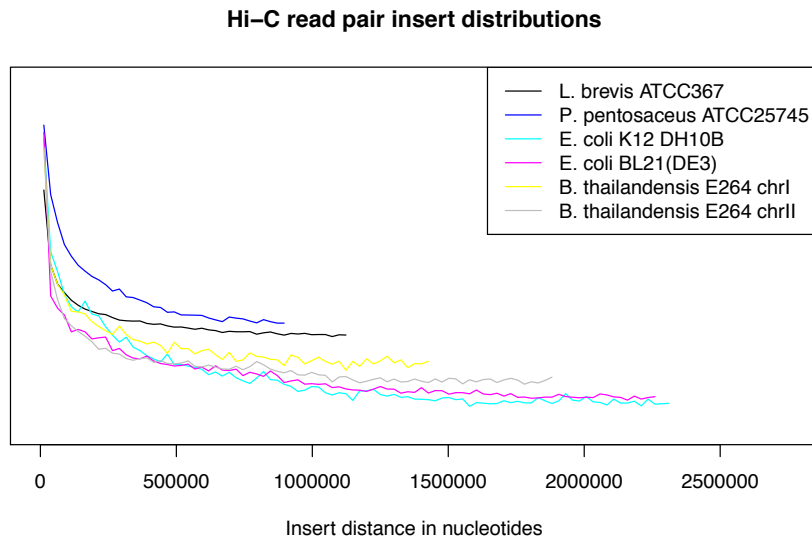
	Num. reads	Max	Avg.	% Same c.c
5 kb, MP	10287315	71	14.81	6.21
10 kb, MP	7681515	96	24.45	16.6
20 kb, MP	4871227	94	27.58	32.38
40 kb, MP	4257896	111	37.19	100
Hi-C (all)	16429505	10	5.11	97.77
Hi-C (>1 kb)	111525	11	5.47	94.75

639

640 Table 3. Variant graph statistics. Connectivity statistics are shown for variant graphs
641 constructed from various simulated mate-pair (# kb, MP) and Hi-C read datasets. Graph
642 constructed from all Hi-C data are compared to those constructed using only Hi-C read
643 pairs with inserts over 1 kb. The Hi-C variant graphs are highly connected in contrast to
644 the mate-pair graphs that have both lower connectedness and lower rates of variants
645 occurring in the same connected components.

646 **FIGURES**

647



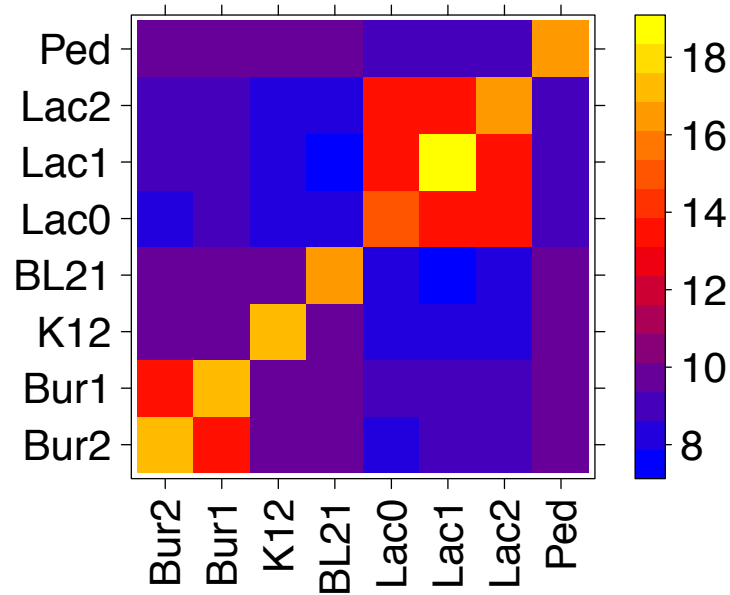
648

649

650 Figure 1. Hi-C insert distribution. The distribution of genomic distances between Hi-C
651 read pairs is shown for read pairs mapping to each chromosome. For each read pair the
652 minimum path length on the circular chromosome was calculated and read pairs
653 separated by less than 1000 bp were discarded. The 2.5 Mb range was divided into 100
654 bins of equal size and the number of read pairs in each bin was recorded for each
655 chromosome. Bin values for each chromosome were normalized to sum to 1 and
656 plotted.

657

658

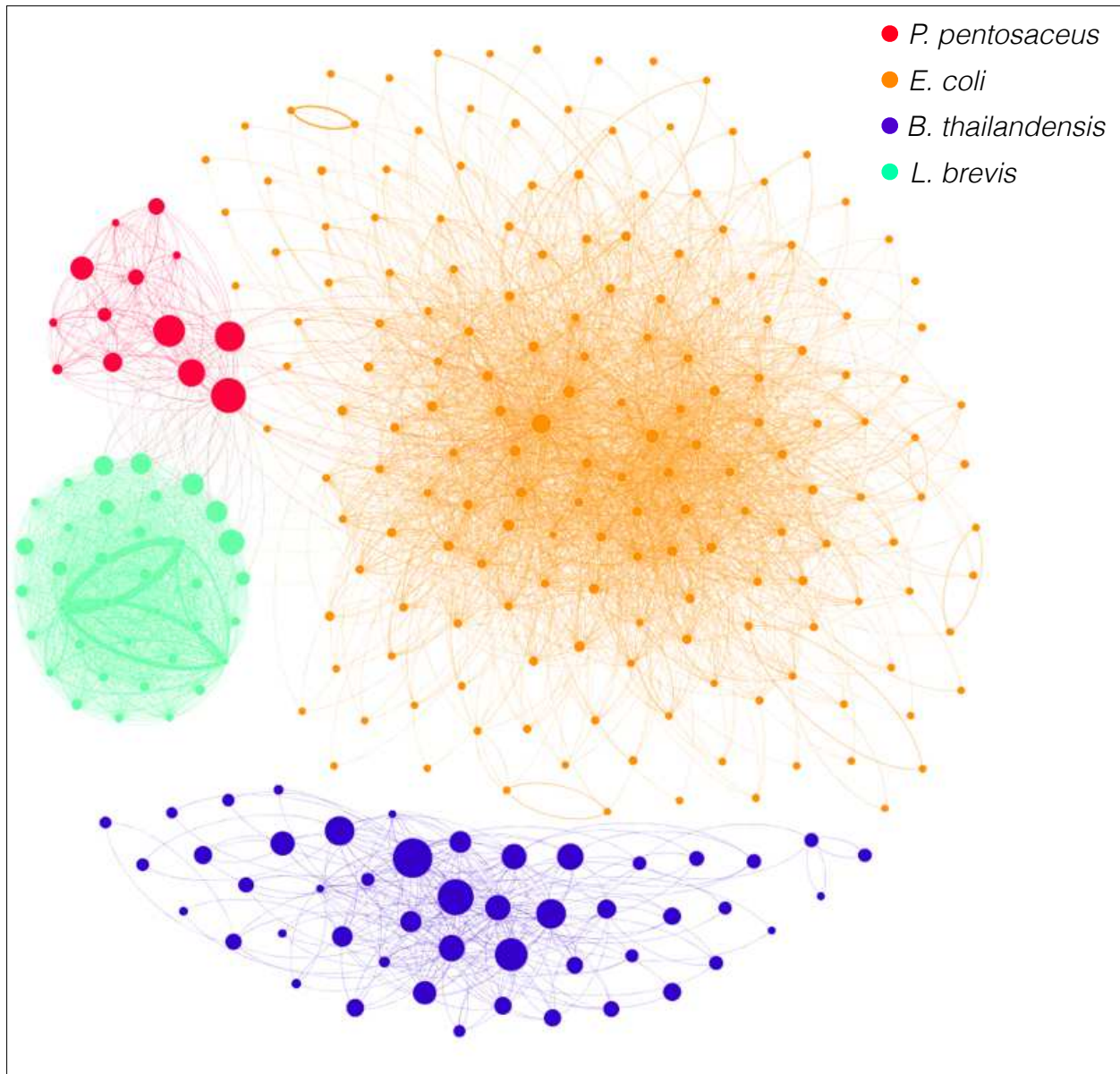


659

660 Figure 2. Metagenomic Hi-C associations. The log-scaled, normalized number of Hi-C
661 read pairs associating each genomic replicon in the synthetic community is shown as a
662 heat map (see color scale, blue to yellow: low to high normalized, log scaled association
663 rates). Bur1: *B. thailandensis* chromosome 1. Bur2: *B. thailandensis* chromosome 2.
664 Lac0: *L. brevis* chromosome, Lac1: *L. brevis* plasmid 1, Lac2: *L. brevis* plasmid 2, Ped:
665 *P. pentosaceus*, K12: *E. coli* K12 DH10B, BL21: *E. coli* BL21.

666

667

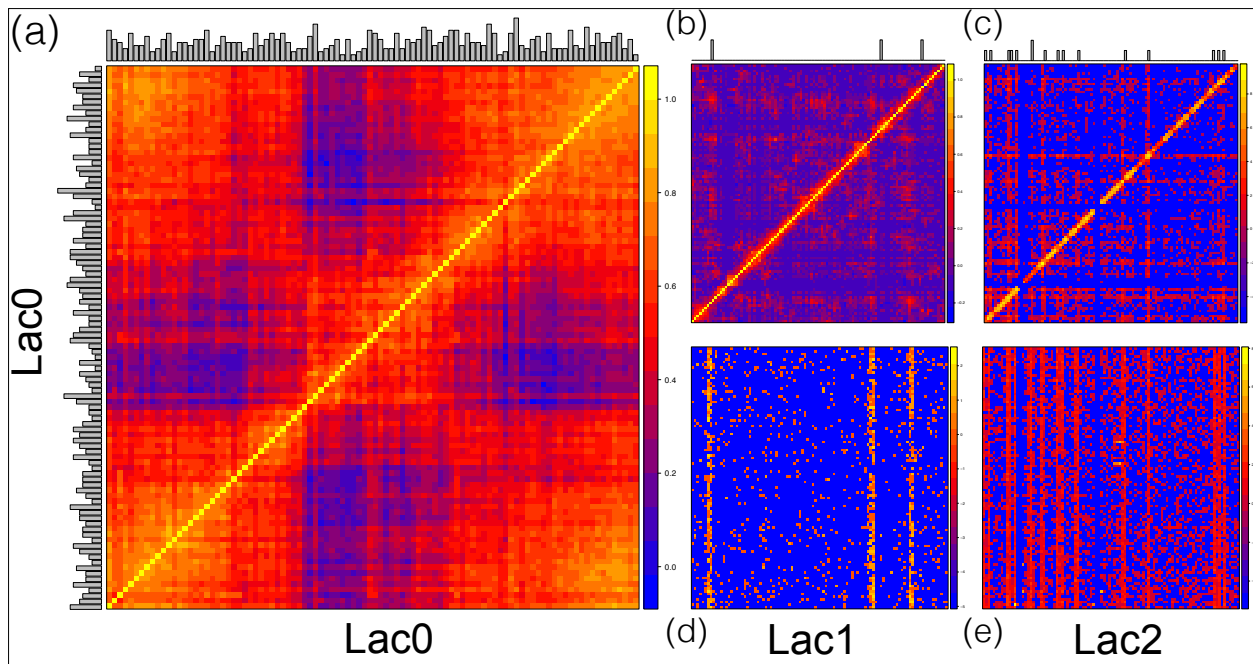


668

669 Figure 3. Contigs associated by Hi-C reads. A graph is drawn with nodes depicting
670 contigs and edges depicting associations between contigs as indicated by aligned Hi-C
671 read pairs, with the count thereof depicted by the weight of edges. Nodes are colored to
672 reflect the species to which they belong (see legend) with node size reflecting contig
673 size. Contigs below 5kb and edges with weights less than 5 were excluded. Contig
674 associations were normalized for variation in contig size.

675

676

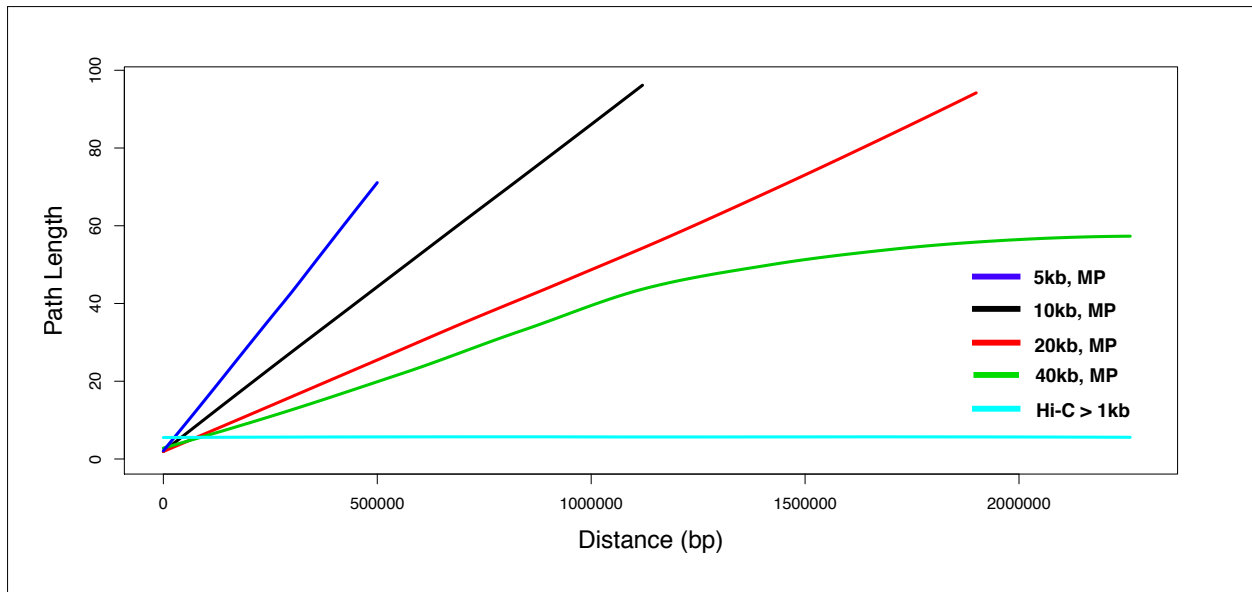


677

678 Figure 4. Hi-C contact maps for replicons of *Lactobacillus brevis*. Contact maps show
 679 the number of Hi-C read pairs associating each region of the *L. brevis* genome. The *L.*
 680 *brevis* chromosome (Lac0, a, Spearman rank correlation) and plasmids (Lac1, b; Lac2,
 681 c) show enrichment for local associations (bright diagonal band). Interactions between
 682 Lac1 and Lac0 (d) and Lac2 and Lac0 (e) are shown. All except Lac0 are log-scaled.
 683 Circularity of Lac0 became apparent after transforming data with the Spearman rank
 684 correlation (computed for each matrix element between the row and column sharing that
 685 element) in place of log transformation (a) indicated by the high number of contacts
 686 between the ends of the sequence. In all plots, pixels are sized to represent interactions
 687 between blocks sized at 1% of the interacting genomes. The number of *HindIII*
 688 restriction sites in each region of sequence is shown as a histogram on the left and top
 689 of each panel.

690

691



692

693

694

695

696

Figure 5. Relationship of distance to degree of separation in Hi-C and mate-pair variant graphs. The length of paths between random pairs of SNP sites in a SNP graph constructed from both Hi-C and mate-pair libraries of varying sizes (left; 5 kb, 10 kb, 20 kb, 40 kb), smoothed using locally-weighted regression.

697 **SUPPLEMENT**

698

699 **Supplementary Tables**

700

	Cov.	Contigs (scaff.)	Bp assem.	N50 (scaff.)	N90 (scaf)	Max. len.	Contig. error	Scaff. error	Pct. assem.
SOAP-1	5	11609 (1594)	15,512,802	14136	3427	160684	2	3	77.01
SOAP-2	50	7687 (561)	15,606,703	87595	9083	379938	0	0	77.48
SOAP-3	100	7687 (557)	15,577,164	87634	8552	379623	0	0	77.37

701

702 Supplementary Table 1. SOAPdenovo assembly results. Statistics are shown for three
703 assemblies, including the simulated coverage and the number of contigs (and scaffolds)
704 present in the assembly. Assembly quality is reflected in the count of misassembled
705 contigs and scaffolds (“contig error” and “scaffold error”). The percent of the total
706 reference sequence size constituted by each assembly is also shown.

707

708

709

	Align.	% of Total	Filtered	% of Align.	Length	GC	#R.S.	RE site % of tot.	Align. / RE count	A/R %
Lac0	10603204	26.17%	10269562	96.85%	2291220	0.462	629	21.28%	16857.24	9.48%
Lac1	145718	0.36%	145478	99.84%	13413	0.386	3	0.10%	48572.67	27.31%
Lac2	691723	1.71%	665825	96.26%	35595	0.385	16	0.54%	43232.69	24.31%
Lac	11440645	28.23%	11080865	96.86%	2340228	0.46	648	21.92%	17655.32	61.09%
Ped	2084595	5.14%	2022870	97.04%	1832387	0.373	863	29.19%	2415.52	1.36%
BL21	12882177	31.79%	2676458	20.78%	4558953	0.508	508	17.19%	25358.62	14.26%
K12	9693726	23.92%	1218281	12.57%	4686137	0.507	568	19.22%	17066.42	9.59%
E. coli	22575903	55.71%	3894739	17.25%	9245090	0.51	1076	36.40%	20981.32	23.85%
Bur1	1886054	4.65%	1797745	95.32%	2914771	0.68	144	4.87%	13097.60	7.36%
Bur2	2536569	6.26%	2464534	97.16%	3809201	0.672	225	7.61%	11273.64	6.34%
Bur	4422623	10.91%	4262279	96.37%	6723972	0.68	369	12.48%	11985.43	13.70%

710

711 Supplementary Table 2. Species alignment fractions (expanded table). The number of
712 reads aligning to each replicon present in the synthetic microbial community are shown
713 before and after alignment filtering, along with the percent of total constituted by each
714 species. The GC content and restriction site (R.S.) counts of each replicon, species,
715 and strain are shown. Total and fractional raw alignment counts adjusted by R.S. counts
716 are also shown, constituting our best approximation of relative abundances of synthetic
717 community members. Bur1: *B. thailandensis* chromosome 1. Bur2: *B. thailandensis*
718 chromosome 2. Lac0: *L. brevis* chromosome, Lac1: *L. brevis* plasmid 1, Lac2: *L. brevis*
719 plasmid 2, Ped: *P. pentosaceus*, K12: *E. coli* K12 DH10B, BL21: *E. coli* BL21.

720

721

	Bur2	Bur1	K12	BL21	Lac0	Lac1	Lac2	Ped
Bur2	239305	4908	111	199	517	16	84	385
Bur1	4908	161432	74	127	396	14	49	264
K12	111	74	160137	244	252	5	26	222
BL21	199	127	244	348021	497	7	61	458
Lac0	517	396	252	497	2302389	10816	57786	1501
Lac1	16	14	5	7	10816	31659	1085	28
Lac2	84	49	26	61	57786	1085	138921	111
Ped	385	264	222	458	1501	28	111	547694

722

723 Supplementary Table 3. Raw metagenomic Hi-C association counts. The number of Hi-
724 C read pairs associating each genomic replicon in the mock community is shown
725 without normalization. These data with normalization applied can be found in
726 Supplementary Table 4.

727

728

	Bur2	Bur1	K12	BL21	Lac0	Lac1	Lac2	Ped
Bur2	22375775	673634	15820	13067	4999	8418	9733	16050
Bur1	673634	32523886	15482	12241	5620	10813	8334	16155
K12	15820	15482	34791627	24423	3714	4010	4592	14108
BL21	13067	12241	24423	16048911	3375	2586	4963	13409
Lac0	4999	5620	3714	3375	2302389	588559	692463	6471
Lac1	8418	10813	4010	2586	588559	93744228	707501	6569
Lac2	9733	8334	4592	4963	692463	707501	19948758	5734
Ped	16050	16155	14108	13409	6471	6569	5734	10180848

729

730 Supplementary Table 4. Normalized association counts. Shown are the counts of Hi-C
731 read pairs associating each pair of replicons included in the synthetic community,
732 normalized as described in the methods. These data without normalization applied can
733 be found in Supplementary Table 3.

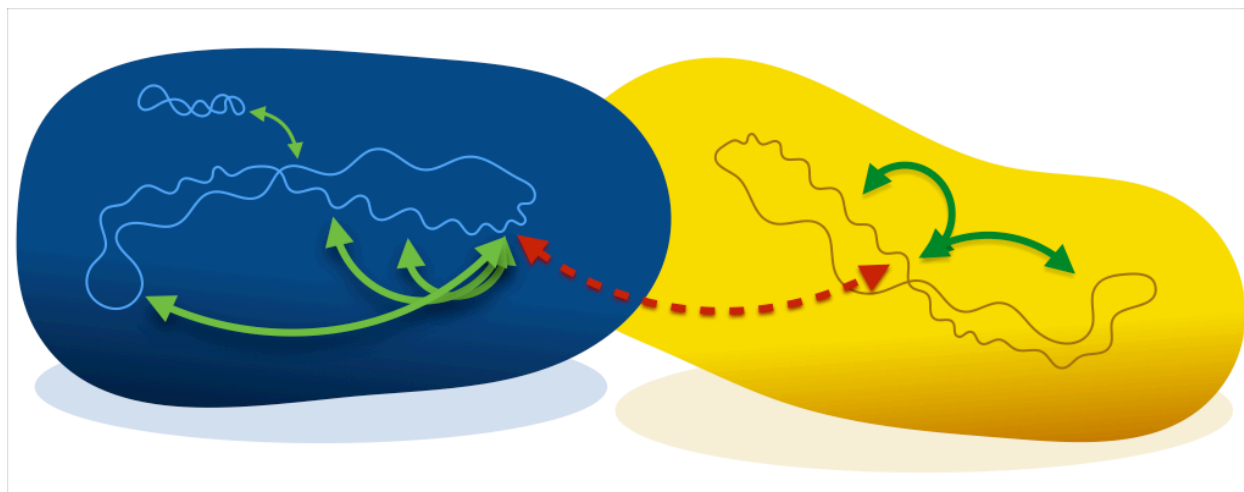
734

735

736

737 **Supplementary Figures**

738



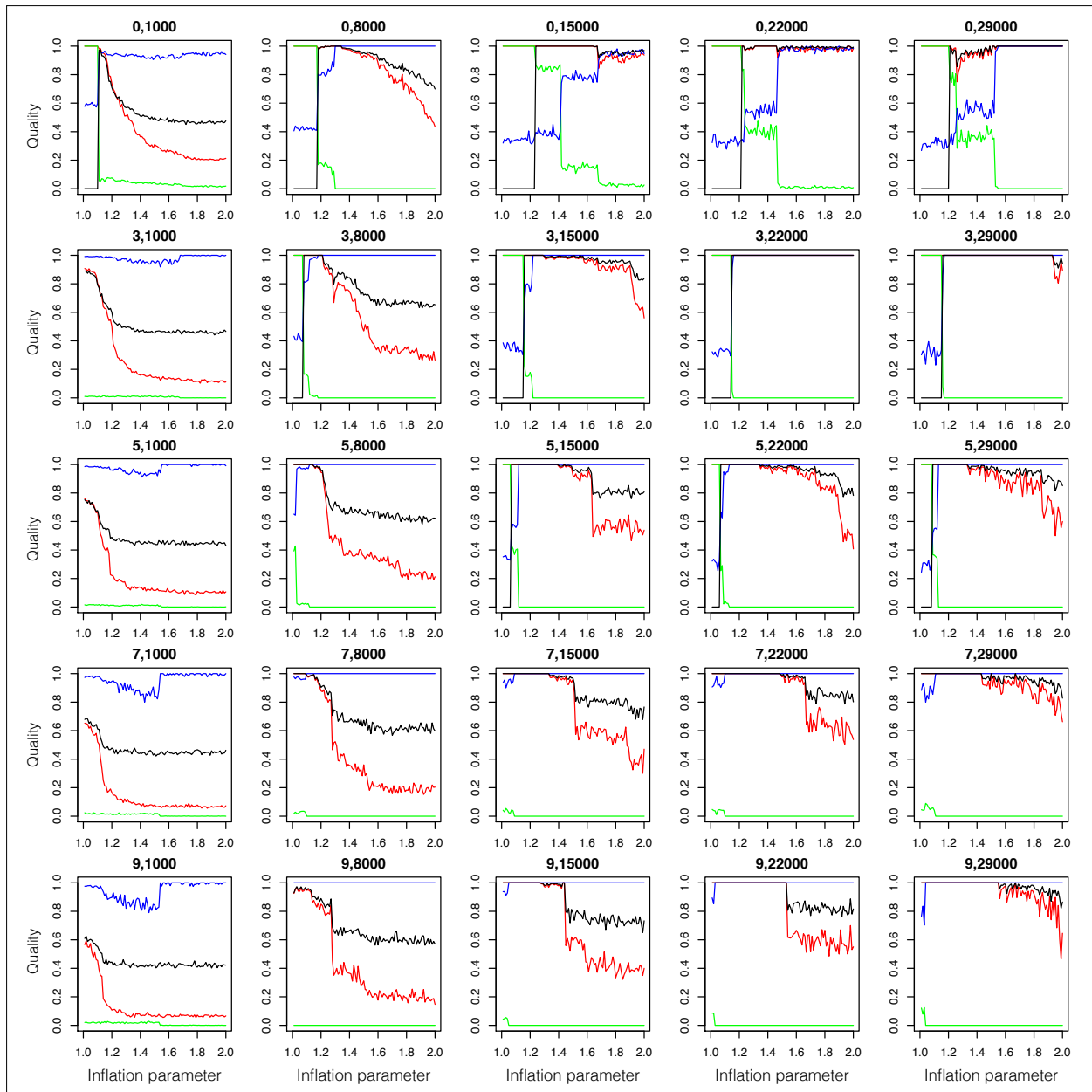
739

740

741 Supplementary Figure 1. Illustration of the metagenome binning signal provided by Hi-
742 C. Two bacterial cells are illustrated, each containing a single circular chromosome. For
743 one genomic region in each of the two species, examples of associations that are likely
744 (green; red is “not likely”) to be derived from Hi-C are illustrated.

745

746

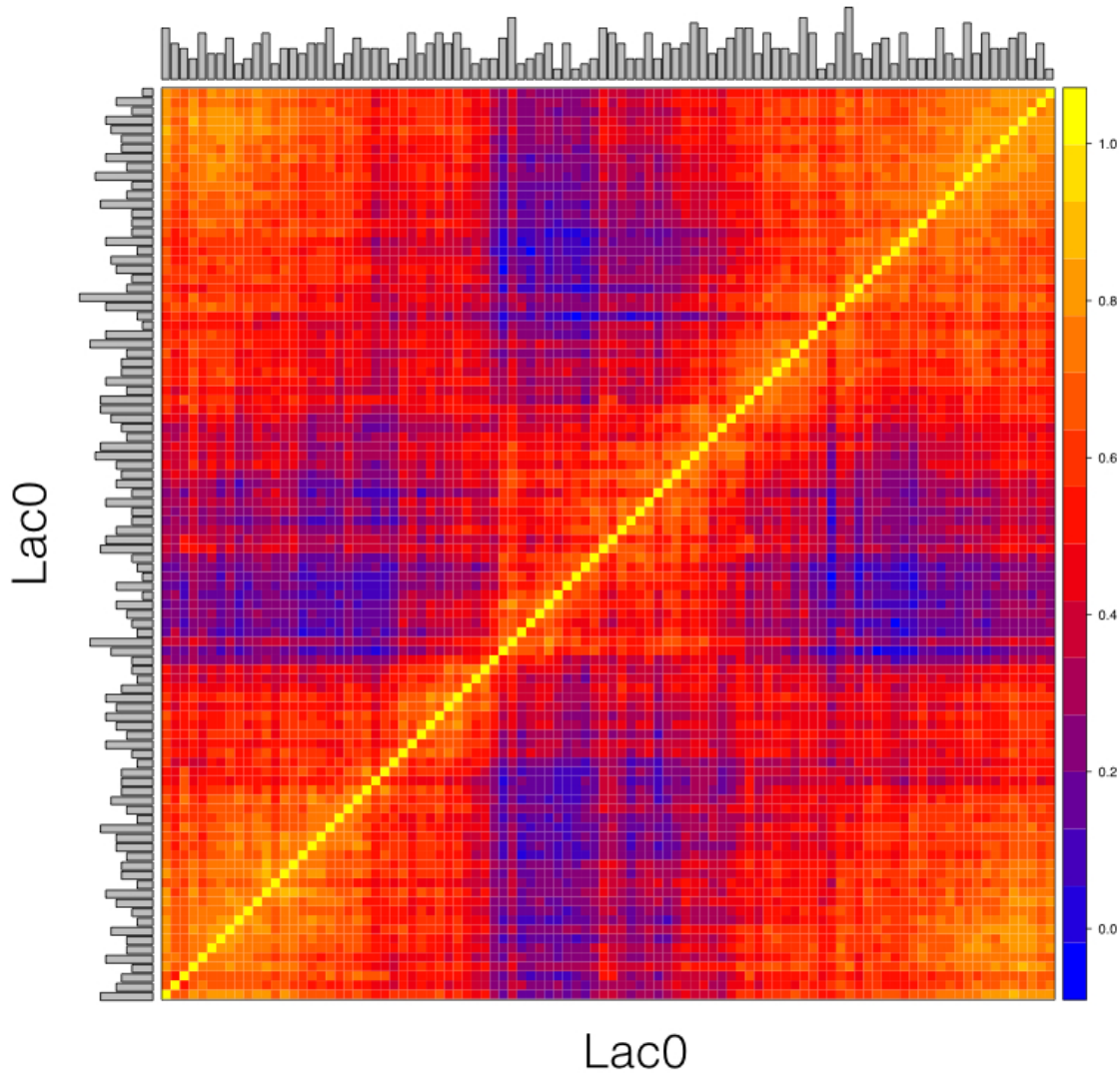


747

748

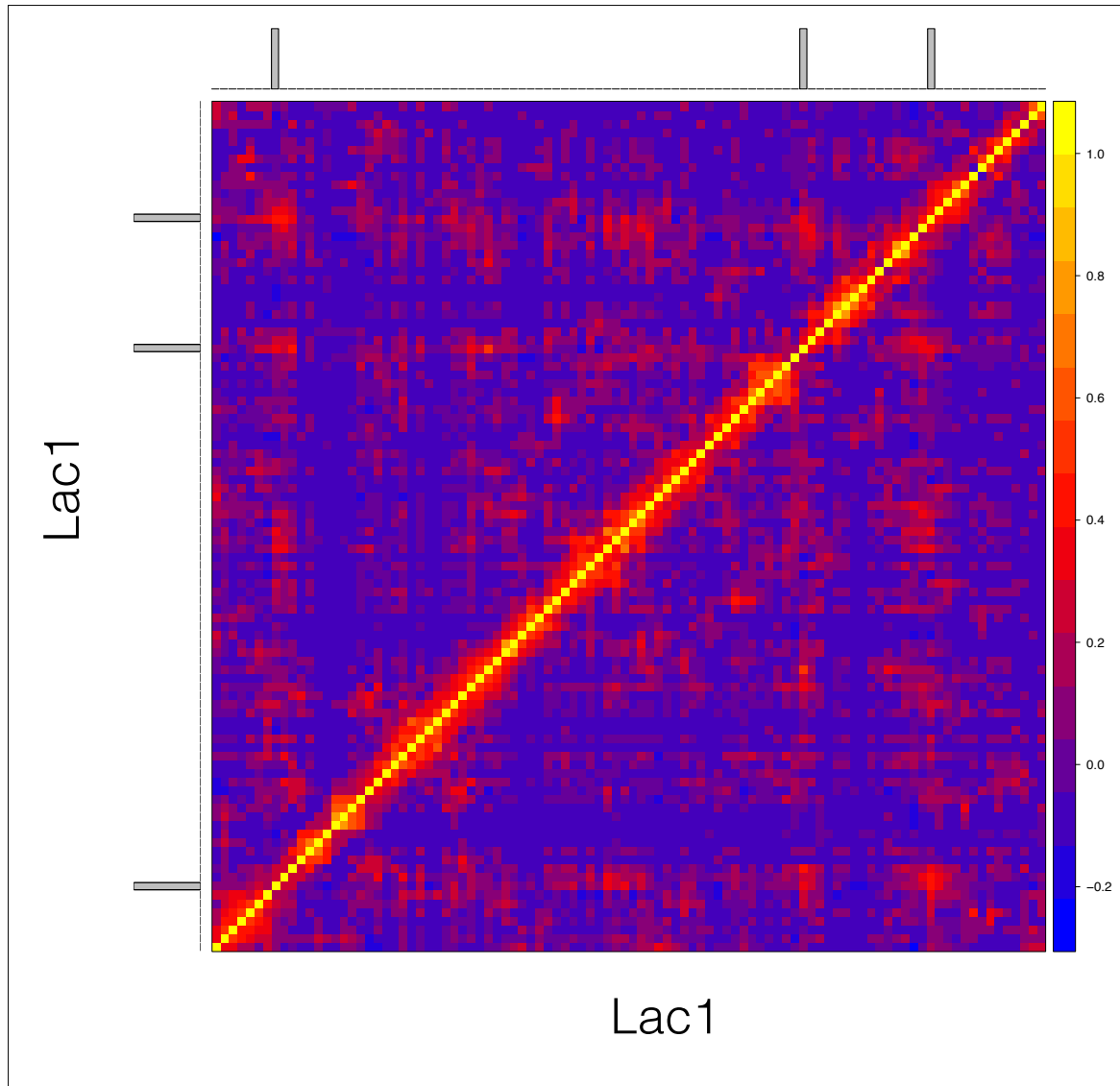
749 Supplementary Figure 2. Visualization of the impact of parameter choice on the quality
750 of clustering solutions. A small-multiples plot is showing 5x5 combinations of contact
751 minimum (top to bottom; 0, 3, 5, 7, 9) and contig size minimum (left to right; 1,000,
752 8,000, 15,000, 22,000, 29,000) thresholds. For each parameter combination, line plots
753 show the quality (y-axis) of clustering solutions performed for inflation values in the
754 interval [1,2]. The quality of clustering solutions is measured in terms their true-positive
755 rate (red), false-positive rate (green), positive predictive value (blue), and negative
756 predictive value (black) are shown.

757



758
759 Supplementary Figure 3. Hi-C contact frequency within *L. brevis* genome. Contact
760 frequency is visualized as a heat map, after normalization and application of the
761 spearman rank correlation (matrix elements are the spearman correlation of the row and
762 column of which they are the intersection). Circularity is apparent in the elevated contact
763 between either end of the reference assembly sequence.
764

765

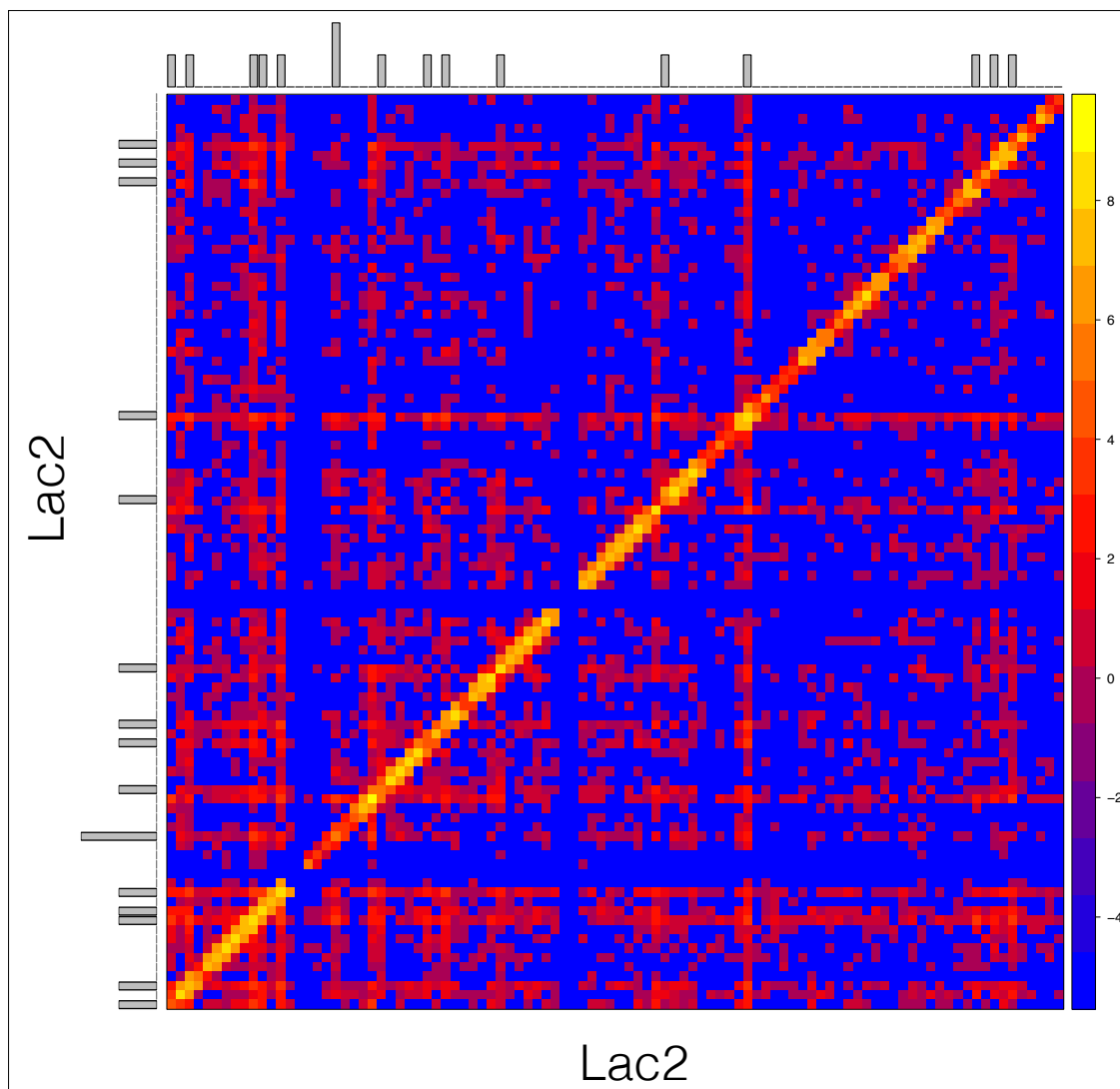


766

767 Supplementary Figure 4. Hi-C contact map for *Lactobacillus brevis* plasmid 1. Contact
768 maps show the number of Hi-C read pairs associating each region of the *L. brevis*
769 plasmid 1. Contact values are Spearman rank correlation transformed following
770 normalization. Pixels are sized to represent interactions between blocks sized at 1% of
771 the interacting sequence. A minimal signal of circularity is apparent with enrichment for
772 contact between the minimum and maximum positions within the reference assembly.

773

774

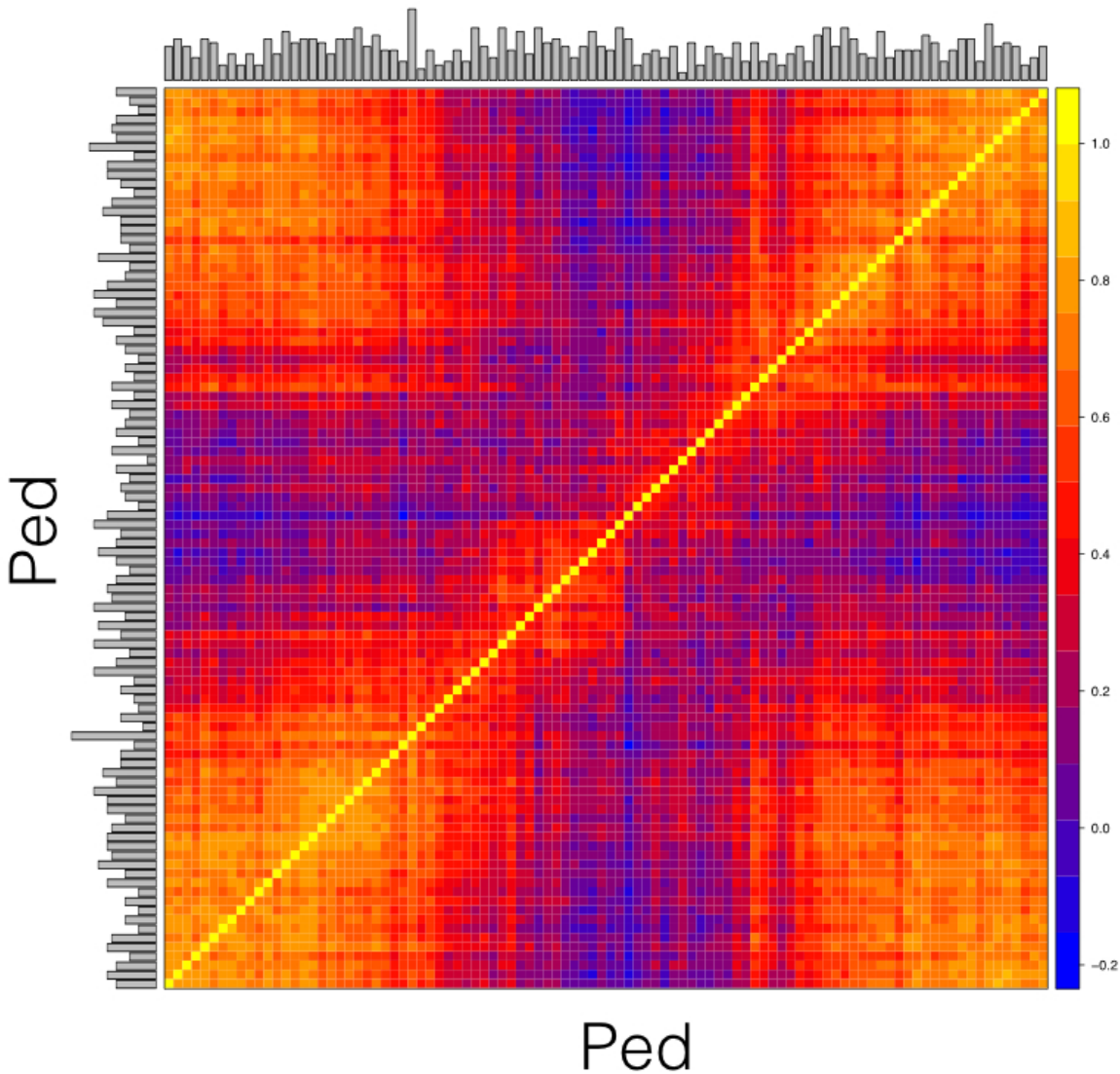


775

776 Supplementary Figure 5. Hi-C contact map for *Lactobacillus brevis* plasmid 2. Contact
777 maps show the number of Hi-C read pairs associating each region of the *L. brevis*
778 plasmid 2. Contact values are Spearman rank correlation transformed following
779 normalization. Pixels are sized to represent interactions between blocks sized at 1% of
780 the interacting sequence. A signal indicative of circularity is not apparent.

781

782

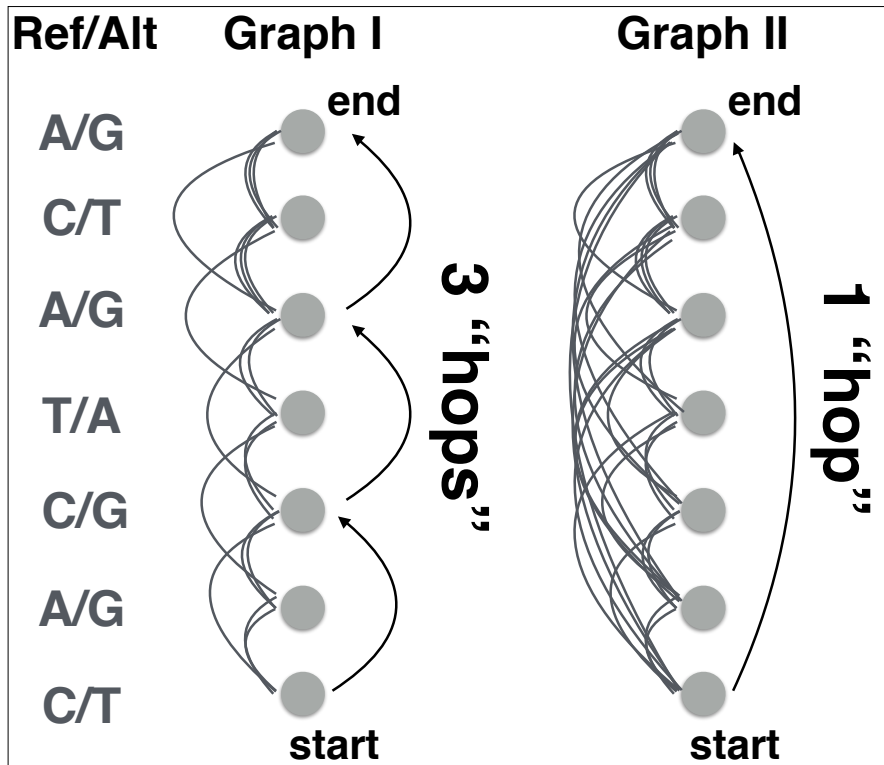


783

784 Supplementary Figure 6. Hi-C contact frequency within *P. pentosaceus* genome.
785 Contact frequency is visualized as a heat map, after normalization and application of the
786 spearman rank correlation (matrix elements are the spearman correlation of the row and
787 column of which they are the intersection). Circularity is apparent in the elevated contact
788 between either end of the reference assembly sequence.

789

790



791

792

793

794

795

796

797

798

Supplementary Figure 7. Variant graph illustration. Two examples of variant graphs (non-data illustration). Variant nodes (circles) are linked by edges (light grey lines) derived from read pair data with small and medium (Graph I) or small, medium, and large (Graph 2) inserts. A path between two nodes (start, end) is illustrated and this path is shorter in the graph representing the dataset that includes larger-insert reads.