Binning unassembled short reads based on k-mer abundance covariance using sparse coding

Olexiy Kyrgyzov, Vincent Prost, Stéphane Gazut, Bruno Farcy, Thomas Brüls

▶ **To cite this version:**

Olexiy Kyrgyzov, Vincent Prost, Stéphane Gazut, Bruno Farcy, Thomas Brüls. Binning unassembled short reads based on k-mer abundance covariance using sparse coding. GigaScience, 2020, 9 (4), pp.1-13. 10.1093/gigascience/giaa028 . cea-04252039

HAL Id: cea-04252039

https://cea.hal.science/cea-04252039

Submitted on 20 Oct 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

GigaScience, 9, 2020, 1–13

doi: 10.1093/gigascience/giaa028
Technical Note

TECHNICAL NOTE

Binning unassembled short reads based on *k*-mer abundance covariance using sparse coding

Olexiy Kyrgyzov[1], Vincent Prost[1,2], Stéphane Gazut[2], Bruno Farcy[3] and Thomas Brüls[1,*]

[1]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Université Paris-Saclay, 2 rue Gaston Crémieux, 91057 Evry, France; [2]Laboratoire Sciences des Données et de la Décision, LIST, CEA, Bâtiment 565, 91191 Gif-sur-Yvette, France and [3]Atos Bull Technologies, 68 avenue Jean Jaurès, 78340 Les Clayes-sous-Bois, France

[*]**Correspondence address.** Thomas Brüls, Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Université Paris-Saclay, 2 rue Gaston Crémieux, 91057 Evry, France. E-mail: bruls@genoscope.cns.fr http://orcid.org/0000-0003-3525-8979

Abstract

Background: Sequence-binning techniques enable the recovery of an increasing number of genomes from complex microbial metagenomes and typically require prior metagenome assembly, incurring the computational cost and drawbacks of the latter, e.g., biases against low-abundance genomes and inability to conveniently assemble multi-terabyte datasets. **Results:** We present here a scalable pre-assembly binning scheme (i.e., operating on unassembled short reads) enabling latent genome recovery by leveraging sparse dictionary learning and elastic-net regularization, and its use to recover hundreds of metagenome-assembled genomes, including very low-abundance genomes, from a joint analysis of microbiomes from the LifeLines DEEP population cohort (n = 1,135, $>10^{10}$ reads). **Conclusion:** We showed that sparse coding techniques can be leveraged to carry out read-level binning at large scale and that, despite lower genome reconstruction yields compared to assembly-based approaches, bin-first strategies can complement the more widely used assembly-first protocols by targeting distinct genome segregation profiles. Read enrichment levels across 6 orders of magnitude in relative abundance were observed, indicating that the method has the power to recover genomes consistently segregating at low levels.

Keywords: metagenomics; human microbiome; sequence binning; sparse coding

Background

Metagenomic shotgun sequencing has dramatically increased our appreciation of the intricacies of microbial systems, whether sustaining biogeochemical processes or underlying health status of their hosts. Several limitations, including sequencing errors, strain-level polymorphism, repeat elements, and inequal coverage, among others, concur however to yield fragmented metagenome assemblies, which require post-processing in order to cluster (bin) assembled fragments into meaningful biological entities, ideally strain-resolved genomes.

The advent of reasonably efficient sequence-binning techniques, often exploiting a coverage covariance signal across multiple samples, allowed the field of metagenomics to move toward more genome-centric analyses [1], and recently thousands of so-called metagenome-assembled genomes (MAGs) have been reported, both from environmental sources and human surfaces or cavities [2–5]. The vast majority of these MAGs have been produced by post-assembly binning approaches, i.e., operating on sequence contigs assembled on a sample-by-sample basis. Although highly successful, such methods are nevertheless "inherently biased towards the most abundant or-

Received: 27 November 2019; **Revised:** 6 March 2020; **Accepted:** 10 March 2020

© The Author(s) 2020. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/gigascience/article/9/4/giaa028/5812700 by CEA user on 20 October 2023

ganisms, meaning consistently less abundant organisms may still be missed" [4]. For example, although thousands of MAGs were reconstructed from >1,500 public metagenomes in the remarkable study by Parks et al. [2], over 93% of these MAGs had an average coverage of >10× (5th percentile, 9.2x, 95th percentile, 268x). The high proportions of phylogenetically unassigned reads typical in medium- to high-complexity metagenomes is another consequence of this limitation [6].

Even though the ecological or community-level importance of rare species is a matter of debate, there are both theoretical and empirical observations supporting the notion that rare organisms can substantially contribute to community-level behavior and resilience and hence represent valuable targets for genome recovery.

Theoretical modeling of microbial trade of diffusible goods [7] has, for example, highlighted an apparent paradox (called "curse of increased efficiency" by Kallus et al [7]), where 1 bacterial species becomes rarer in the population despite becoming fitter and more efficient at producing a key metabolic resource. This situation is provoked by metabolic interdependencies that can evolve via trade in microbial consortia and that can lead to low-abundance organisms becoming essential for a faster growth rate of the community. On the other hand, several empirical studies have documented the ecosystem-level relevance of rare bacteria (see [8] for a review); e.g., Kalenitchenko et al. [9] make a case for the role of "ultrarare" bacteria in ecosystem-level productivity, and Benjamino et al. [10] highlight the role of some low-abundance bacteria in driving termite hindgut bacterial community composition.

Considering that global metagenome assembly (or cross-assembly) is currently unpractical to recover low-abundance genomes or complex microbial consortia from terabytes of data, we decided to investigate a "bin first and assemble second" paradigm that could make the assembly problem more tractable by targeting lower-complexity sequence subsets (bins). Binning unassembled reads is however more computationally demanding because the number of raw sequences is typically orders of magnitude higher than the number of assembled contig sequences.

Even though the dominating paradigm nowadays is assembly-first binning, it is worth noting that the first sequence-binning methods reported, such as AbundanceBin [11] and MetaCluster [12], operated at the read level. This shift towards contig binning was mainly driven by the increase in data throughput, as the first read-level binning methods were designed at the time of 454 (Roche) and even Sanger sequencing (both providing longer reads) to process individual samples. They were thus not designed to scale to large multi-sample terabase-sized short-read datasets. In this perspective, assembly can be viewed as a pre-processor to reduce the computational burden of binning.

A pioneering pre-assembly binning scheme [13] was proposed a couple of years ago, with the read partitioning problem formulated by analogy to the latent semantic analysis (LSA) technique widely used in natural langage processing (NLP). The core idea to view metagenomes as linear mixtures of genomic variables can lead to read clustering formulations based on the deconvolution of latent variables ("eigengenomes") driving the k-mer (subsequences of length k) abundance covariance across samples. The raw sequence data are first summarized in a sample by k-mer occurrence matrix (analogous to term-document matrices in NLP), approximating the abundance of k-mers across samples. Matrix decomposition techniques can then be used to define 2 sets of orthogonal latent vectors analogous to principal components of sample and sequence space. The large memory requirements incurred by the factorization of large abundance matrices naturally drove Cleary et al. [13] toward a rank-reduced singular value decomposition (SVD), for which efficient streaming libraries [14] enable a parallel processing of blocks of the abundance matrix by updating the decomposition iteratively. Clusters of k-mers can then be recovered by an iterative sampling and merging heuristic that samples blocks of eigen k-mers from the right singular vectors matrix until an arbitrary portion (∼0.4% in [13]) of the latter has been covered. This heuristic is however acknowledged as a significant hindrance, the authors stating that "more sophisticated methods are needed to computationally discover a 'natural' clustering" [13].

We describe here a pre-assembly binning method based on sparse dictionary learning and elastic-net regularization that exploits sparsity and non-negativity constraints inherent to k-mer count data. This sparse coding formulation of the binning problem can leverage efficient online matrix factorization techniques [15] and scales to very large (terabyte-sized) k-mer abundance matrices; it also bypasses the aforementioned problematic k-mer clustering heuristic, removes interpretability issues associated with the SVD (e.g., the physical meaning of negative contributions), and is able to enrich sequences from a given genome across 6 orders of magnitude in relative abundance (see section "Recovery of very low-abundance genomes").

Analyses

We describe in the following section some analyses and results of the proposed binning scheme based on the modeling of data vectors as sparse linear combinations of basis elements (sparse coding [15]).

We start with a preliminary experiment illustrating the ability of read binning to recover a target genome whose sequences segregate at levels too low to yield any kilobase-sized fragment by assembly in any single sample and hence would not be recoverable by assembly-first approaches. We then describe results from a direct comparison of assembly-first versus bin-first methods that illustrate the complementarity of the 2 approaches in terms of the profiles of genomes recovered. The next subsection describes a comparison of the sparse coding–based bin-first approach with a state of the art read-binning method. The next subsections describe strain separation results obtained with the new method and document its scalable behavior and its ability to enrich rare sequences, thereby enabling the recovery of low-abundance genomes. We conclude with a discussion of some important limitations of the method and consider some of its potential applications.

Read-level binning can recover low-abundance genomes that escape assembly-first protocols

We devised an experiment to illustrate a situation where assembly-first approaches are not able to recover a target genome—because target genome sequences are too low in number in any single sample—whereas a bin-first approach is successful at it. The experimental set-up involved distributing a very low number of short reads (100 paired reads) randomly sampled from a target genome (a 10-kb plasmid) into 14 samples containing each a background of 20,000 unrelated bacterial sequences (4 further samples contained only background sequences with no read from the target genome at all). Because no single kilobase-sized fragment could be recovered by assembling the sequences from each sample individually, this precluded

Downloaded from https://academic.oup.com/gigascience/article/9/4/giaa028/5812700 by CEA user on 20 October 2023

Table 1: Binning accuracy estimates

Parameter	*k*-means	LSA	Sparse coding
Precision	0.52	0.58	0.72
Recall	0.63	0.64	0.82
F-value	0.57	0.61	0.77

LSA refers to the original algorithm of [13], with a cosine similarity threshold of 0.7 as recommended by the authors; *k*-means refers to a direct clustering of the columns of the abundance matrix, with the number of clusters set to 1,000 (equal to the number of components for the sparse decomposition); see main text and Methods.

the application of assembly-first methods (e.g., contig binning methods like metabat [16,17] require ≥1,500 bp sequences as input). On the other hand, ∼90% of the reads originating from the target genome could be segregated in a single cluster/bin using our read binning pipeline (Supplementary Table 1), leading to the complete recovery of the target genome in a single contig after assembly (Methods).

Bin-first and assembly-first strategies recover distinct and complementary genome sets

A second experiment aimed at directly comparing the genome recovery yield of assembly-first versus bin-first strategies on a real-life dataset. We selected the raw sequence data from 18 (randomly chosen) individuals of the LifeLines DEEP cohort [18] and either assembled these individually (i.e., on a sample-by-sample basis) with metaSPAdes (v3.13.0) followed by contig binning across samples with the MetaBat2 adaptive algorithm [17] or clustered the raw reads using our read-level binning pipeline, followed by metaSPAdes assembly of the resulting partitions/bins.

Fourteen nearly (>90%) complete and uncontaminated (<5%) genomes were recovered using the assembly-first approach, versus 7 using the bin-first method. Interestingly, the 2 genome sets were disjoint, with no complete genome recovered by both approaches. Among the 14 genomes recovered by the assembly-first approach, 3 were not represented in the set of 164 MAGs recovered from the analysis of the entire cohort using the bin-first protocol. More surprisingly, only 3 of the 7 complete genomes retrieved by our bin-first pipeline from the analysis of 18 samples were represented among the complete or nearly complete MAGs identified from the full cohort analysis, indicative of a lack of stability of the algorithm that we relate to bin fragmentation provoked by extensive strain-level variation across the samples (see Discussion).

The surprising lack of overlap between the 2 genome sets in this experiment is not attributable to fundamental differences in abundance levels between the genomes recovered by the 2 approaches because in both cases the genome bins could be directly aligned to individual sample assemblies; i.e., the genomes recovered using both approaches were of sufficiently high coverage to yield relatively large contigs in the assemblies of individual samples. We assessed potential differences between the distributions of binned genome sequences across the samples, which highlighted distinct patterns for the 2 approaches, with the genomes identified by the bin-first approach aggregating sequences from a larger number of samples (and harboring a higher number of contigs per genome bin on average) (Fig. 1).

Thus, in the present experiment, the assembly-first approach targeted genomes reaching high abundance in a limited number of samples, for which the weaker abundance covariation signal probably hampered the bin-first approach. Consistent with this view, sequences from genomes produced through the assembly-first approach were frequently located in large (dozens of megabase pairs in size) and unresolved partitions computed by read-level binning (see Discussion).

On the other hand, we should keep in mind that the number of samples (18) used in this experiment is relatively low. Related approaches based on abundance-covariance, like Concoct [19] or LSA [13] among others, require a higher number of samples to achieve best performance (∼50 samples for the former and 30–50 for the latter).

Despite these limitations, the fact that the bin-first approach was able to recover a significant number of complete genomes not identified by the assembly-first approach illustrates the complementarity of the 2 strategies.

Enhanced accuracy of sparse coding–based read binning versus state of the art read binning

Besides the 2 pioneering read-binning methods already mentioned (AbundanceBin [11] and MetaCluster [12]), we could also mention CompostBin [20], which is a principal component analysis–based read-level–binning algorithm that was designed and tested on Sanger reads. BiMeta [21] and MetaProb [22] are other tools that operate at the read level but describe themselves as "assembly-assisted," meaning that they rely on the detection of read overlaps. BiMeta was tested on 454 reads simulating bacterial communities of a dozen different genomes at most and on the Acid Mine Drainage dataset [23], which is of low complexity and consists of Sanger reads. MetaProb shares some principles with BiMeta: it is also "assembly-assisted" and was tested on the same low-complexity synthetic datasets as the latter. The authors also tested their method on a real microbiome sample consisting of 43 million reads, but only after filtering the latter down to 2 million reads.

Thus, all the above methods were designed to operate on individual samples, at a time when scalability issues were less acute. Moreover, with the exception of AbundanceBin, which exploits a coverage signal extracted from unique *k*-mers, the other methods are better described as composition-based, using a nucleotide composition signal measured from short *k*-mers (typically of length 4 or 5).

We developed our method with scalability in mind because we wanted it to be able to process on the order of 10^{10} short reads and to be able to process increasingly larger multi-sample datasets by simply stacking additional computing resources. In this respect, there is only 1 competing method left, Latent Strain Analysis [13], that is both scalable and designed to operate on unassembled short reads from a large number of samples.

To evaluate our method, we first compared its read clustering accuracy (measured in terms of precision, recall, and F-value metrics; see Methods) with that of the original LSA method by using previously described benchmark datasets [24] (downloadable from the GigaScience database [25]), for which read to genome assignments were known ([24] and Methods). The results from these experiments are summarized in Table 1 and show improved accuracy of the sparse-coding framework over both the original LSA and a naive *k*-means algorithm.

Partial strain separation

The counting and indexing of *k*-mers in fixed memory is achieved by locality sensitive hashing (Methods). By design, locality sensitive hash functions increase the probability of colli-

Downloaded from https://academic.oup.com/gigascience/article/9/4/giaa028/5812700 by CEA user on 20 October 2023

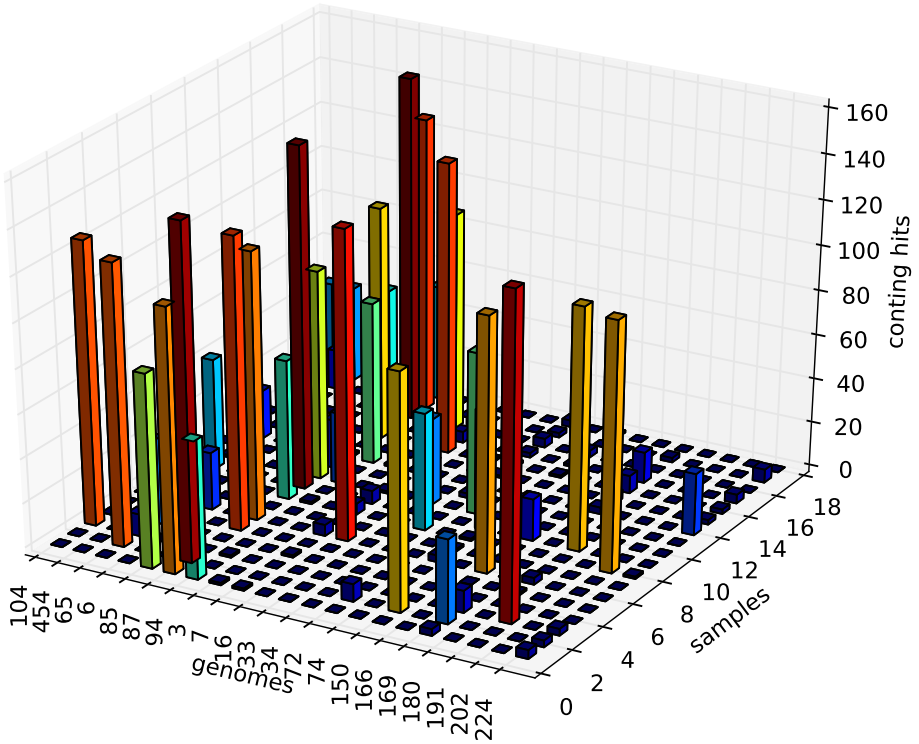

Figure 1: Sample origins of the sequences aggregated into genome bins (displayed by their genome identifier on the x-axis) using our bin-first method (first 7 genomes [104–94] on the left) vs assembly-first binning using metabat2 (14 rightmost [3–224] genomes). Genomes retrieved by the bin-first method aggregate sequences from a larger number of samples.

Table 2: Average nucleotide identity (ANI) between the *Bacillus amyloliquefaciens* strains used in the strain separation experiment illustrated in Fig. 2A

1						
94.20	1					
94.30	99.44	1				
94.21	99.96	99.44	1			
98.74	94.20	94.27	94.20	1		
99.02	94.25	94.33	94.25	98.72	1	
97.75	94.09	94.20	94.09	97.79	97.77	1

See main text and Methods.

sion for related items [26]. On one hand, this provides a convenient way to handle sequencing errors. On the other hand, the occurrence in natural environments of multiple strains from the same species (the so-called species pangenome) could lead to artifactual *k*-mer merging and potential overlap between distinct genomic partitions. This represents an issue potentially exacerbated by the inter-sample read aggregation process.

To assess the behavior of the method in the presence of extensive pangenomic (i.e., strain-level) variation, we quantified its ability to separate closely related (up to 99.96% average nucleotide identity [ANI]; Table 2) strains that were deliberately included in the genome mixtures of the virtual cohort used in the test experiments.

Fig. 2 illustrates 2 practical examples of partial strain separation achieved with the method. Fig. 2A illustrates a partial separation of 7 strains of the species *Bacillus amyloliquefaciens* (whose ANI ranged from 94.18 to 99.96; Table 2), while Fig. 2B shows similar results for 8 strains of *Sulfolobus islandicus* (whose ANI ranged from 97.84 to 99.59). Because the genomic origin of each read is known in the virtual cohort dataset, these plots show, for each strain (represented by a horizontal line), the distribution of its reads among the full set of clusters/bins generated by the pipeline (and arbitrarily ordered along the x-axis). Fig. 2A illustrates that the 7 strains of *B. amyloliquefaciens* are mostly separated into 2 groups according to whether their main cluster is located near x-coordinate 220 or x-coordinate 500. Fig. 2AB on the other hand shows that the 8 strains of *S. islandicus* share a common "core" cluster (located near the origin), while a variable portion of their genomes are segregated into distinct "variable" clusters.

Overall, this analysis makes apparent a partial separation of closely related strains (Fig. 2A), as well as the differential segregation of the core (i.e., the genome fraction that is shared between all the strains of a species) and variable portions of the species pangenomes (Fig. 2B).

In practice, some level of strain mix-up is probably inherent to the inter-sample read aggregation process, and approaches based on sample-by-sample assembly limit the risk of strain mixing, but at the expense of focusing on those genomes that reach high coverage (~10×). Our approach aimed at relaxing the latter constraint, but by doing so through the aggregation of lower-abundance reads across samples, it becomes vulnerable to extensive strain-level variation. Dealing with this problem is the focus of future research; e.g., a possible workaround could be to carry out a "soft-clustering" by allowing "core" sequences to belong to >1 "variable" cluster.

Sensitivity and scalability on real-life data

By scalability, we refer to the ability of the method to adapt to order-of-magnitude change in the input (raw reads) and its abil-

Downloaded from https://academic.oup.com/gigascience/article/9/4/giaa028/5812700 by CEA user on 20 October 2023

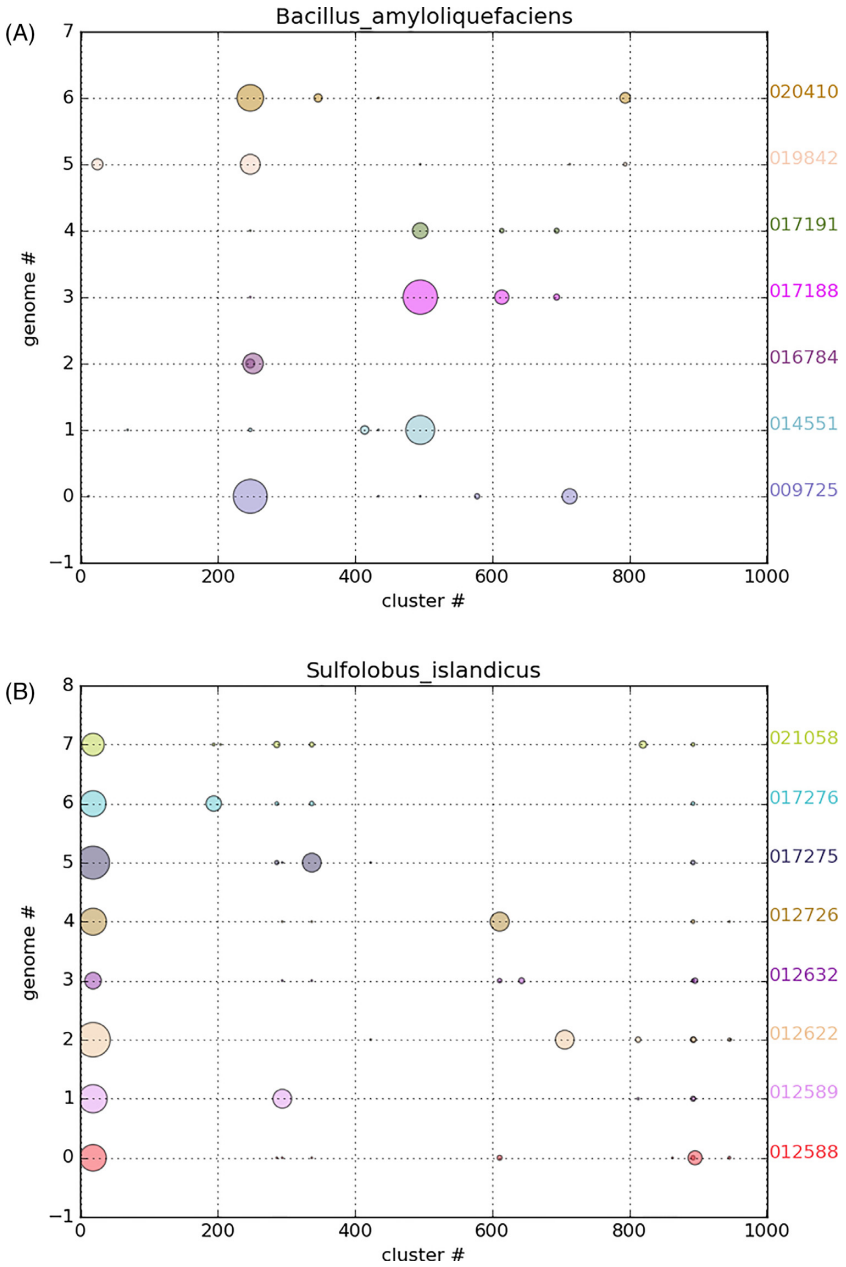

Figure 2: Partial resolution of species pangenomes. x-axis: partition identifier; y-axis: horizontal axes correspond to different strains from the same species (left: *B. amyloliquefaciens* strains; right: *S. islandicus* strains). Circle area is proportional to the number of reads from a given strain assigned to the given partition. A illustrates the partial separation of 7 strains into 2 distinct partitions. B illustrates the differential segregation of the core (at the left of the figure) and variable portions of the species pangenome.

ity to maintain its functionality and performance under high demand (i.e., increasingly higher data volumes).

To assess the sensitivity and scalability of the sparse coding method, we applied it to a real-world dataset of $>10^{10}$ reads (~10 TB of raw sequence data) derived from 1,135 gut microbiomes of healthy Dutch individuals from the LifeLines DEEP cohort [18]. The pre-assembly binning of the cohort's reads resulted in 983 partitions, which were then assembled individually using the SPAdes engine [27] (Methods). The distribution of assembly sizes is shown in Fig. 3, making apparent that almost all partitions are bacterial-genome sized (i.e., in the 2–5 Mb range). A few dozens of coarse-grained partitions harboring unresolved genomes make up the right tail of the distribution. Because a direct read to genome mapping is not available for real-life metagenomes, we assessed clustering performance by quantifying the genomic homogeneity and completeness of the resulting partitions based on the occurrence pattern of universal single-copy markers using the checkm toolkit [28]. A summary of completion and contamination statistics of the genome-resolved partitions is presented in Table 3 .

The fact that many of the partitions display low contamination is somehow balanced by the concomitant generation of large and unresolved partitions. The production of these unresolved partitions arises from the fact that the extent of genome divergence is not uniform across the range of taxa occurring in the samples. As discussed above, strain-level ("pangenomic")

Downloaded from https://academic.oup.com/gigascience/article/9/4/giaa028/5812700 by CEA user on 20 October 2023

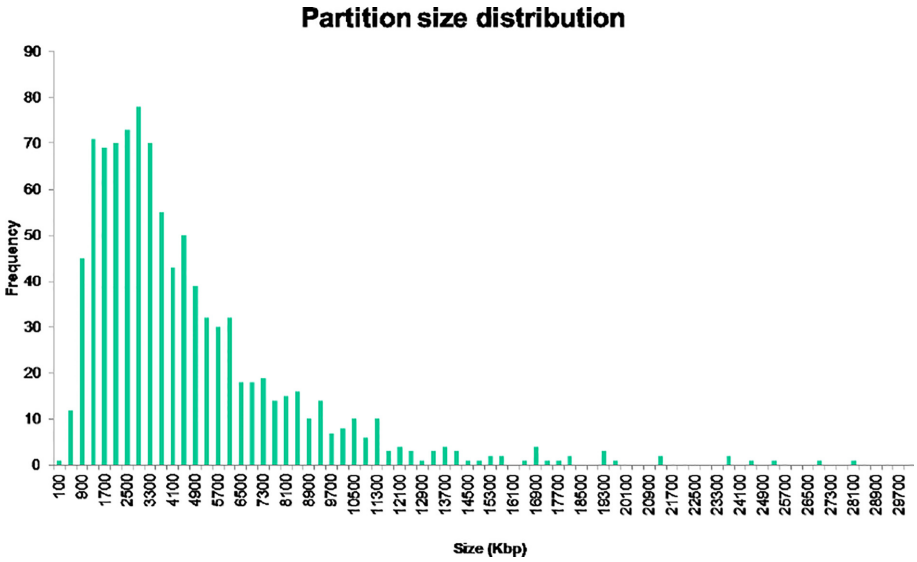

Figure 3: Distribution of assembled bin sizes. x-axis: assembled partition size (in kilobase pairs); y-axis: partition frequency.

Table 3: Genome completion and contamination statistics of assembled partitions/bins

Classification	Completeness (%)	Genomes (bins)	Contamination (%)
Nearly complete	>90	14	≤5
Substantial	>70 to ≤90	53	≤5
Moderate	>50 to ≤70	97	≤5
Partial	≤50	724	≤5
Unresolved	>100	95	>5

See main text and Methods.

variation is another factor contributing to cluster fragmentation, by inducing a differential segregation of the core and variable portions of genomes, and is exacerbated by the inter-sample read aggregation process.

Recovery of very low-abundance genomes

A key motivation for the pre-assembly processing of reads was the theoretical possibility to aggregate reads from low-abundance organisms across samples.

To assess whether we could indeed identify such consistently low-abundance genomes in real-life datasets, we characterized the abundance of a subset of > 70% complete genomes from the LifeLines DEEP cohort analysis by directly mapping the raw reads of the original samples against them. Given the large size of the cohort, this analysis was not performed on the full set of MAGs but on a limited number of genomes, the aim being to validate the ability of the method to retrieve such low-abundance genomes by exhibiting some of them.

The relative enrichment levels of these genomes was measured as the fraction of raw reads contributed by each sample to them (Methods) and is illustrated in Fig. 4 for 2 genomes, with panel A showing an example of a consistently low-abundance genome (i.e., with nearly all the samples contributing no more than 10^{-5} to 10^{-4} of their reads to the given genome), while panel B shows a genome of overall moderate abundance (10^{-4}) but reaching higher abundance (10^{-3}) in a few dozen samples (represented by the rightmost peak in the histogram).

Given the large number of microbiomes analyzed, we quite frequently observed situations where a given genome reaches medium to high relative abundance in ≥1 sample (as illustrated in Fig. 4B). However and importantly, we could also detect instances of genomes that consistently segregated at low abundance levels across the whole cohort (Figs 4A and 5B and D).

The recovery of these genomes was made possible by aggregating a few thousand reads per sample across a large number of samples, thus demonstrating the ability of the method to isolate rarer genomes. Overall, the high proportion of homogeneous partitions corresponding to partial genomes (Table 3) is consistent with the recovery of sequences from lower-abundance organisms, whose cumulative coverage across the cohort is not sufficient to allow complete genome reconstruction.

Assessing novelty against reference genome compendia

To investigate the extent to which the recovered genomes could correspond to novel organisms, we screened a subset of 164 of them (>50% complete with <5% contamination, accessible via GigaDB [25]) against several reference genome libraries. We first compared the genomes against the Kraken 2 [29] database built from NCBI's Refseq bacteria, archaea, and viral libraries (accessed October 2018). Only 21 of the 164 genomes compared had ≥1 fragment classified against this reference database (Methods). We also compared the genomes against the "Global Human Gastrointestinal Bacteria Genome Collection" (HGG [6]), which represents one of the most comprehensive resources of gastrointestinal bacterial reference sequences currently available. Less than half (72 of 164) of the genomes displayed convincing similarity to the HGG genome catalogue (Methods).

Discussion

Abundance covariance-based binning has the power to identify biologically meaningful associations between metagenomic sequences that could go unnoticed by analyses based on sequence overlap (assembly) or nucleotide signatures. This is illustrated

Downloaded from https://academic.oup.com/gigascience/article/9/4/giaa028/5812700 by CEA user on 20 October 2023

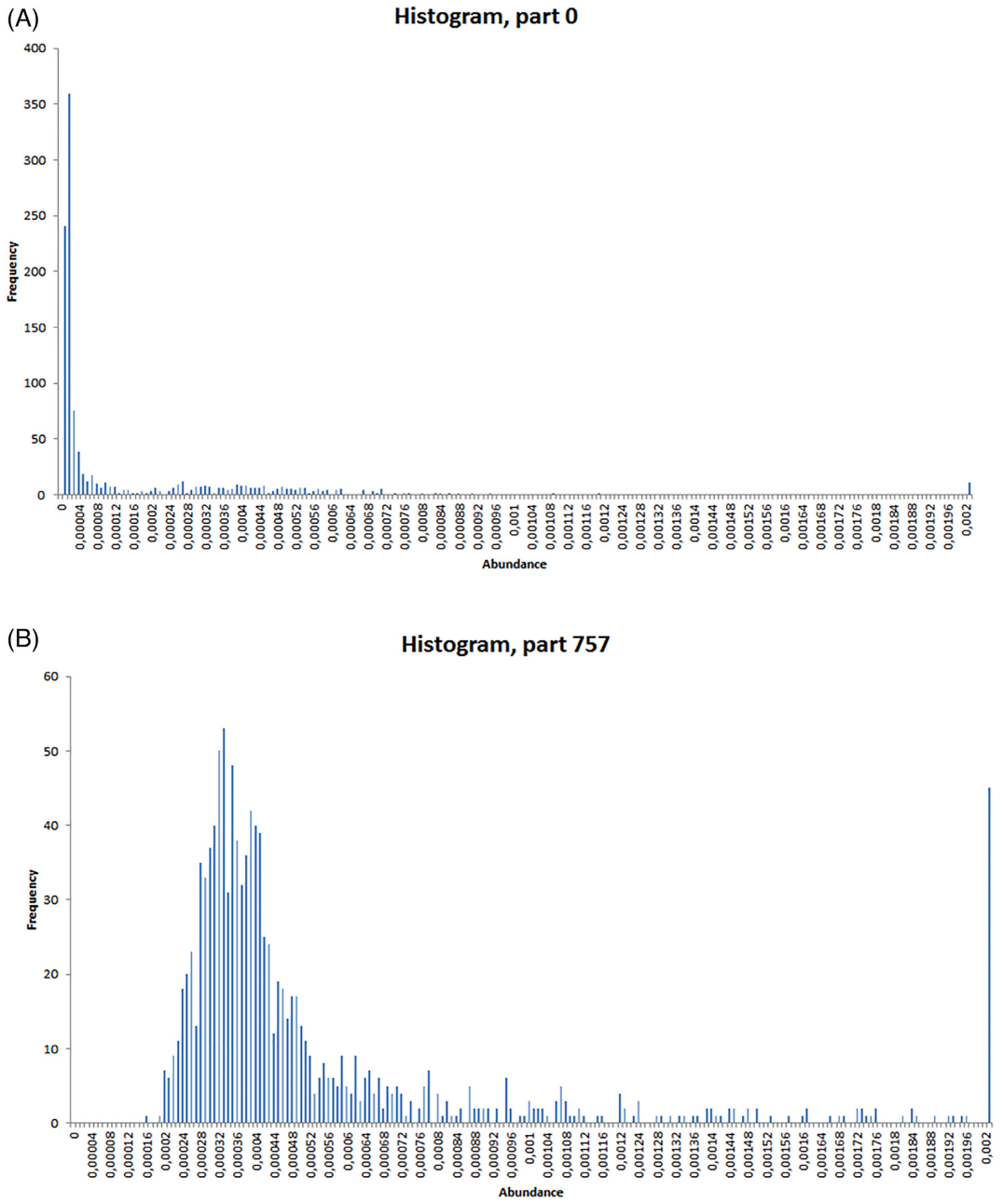

Figure 4: Enrichment histograms displaying the fraction of raw reads contributed by each sample to 2 distinct genome-resolved bins. x-axis: read abundance of partition 0 (left) and partition 757 (right); y-axis: sample frequency (among the 1,135 samples). Different situations are illustrated: a relatively high proportion of reads can be contributed by a small subset of individuals (a few dozens, corresponding to the rightmost peak for the genome-resolved bin shown in panel B), while panel A illustrates that substantial (i.e., ≥70% complete) genomes of low-abundance organisms can also be recovered by aggregating only a few thousand reads per sample across the full cohort.

in the present study by a preliminary experiment using a synthetic dataset spiked with low-abundance sequences from a target genome that does not reach a sufficient coverage to yield kilobase-sized fragments after assembly in any individual sample (thus precluding the application of contig binning) but that is successfully recovered via read-level binning (Supplementary Table). When applied to the $>10^{10}$ reads from the LifeLines DEEP cohort's metagenomes, our bin-first protocol recovers hundreds of metagenome-derived genomes, including from consistently less abundant organisms (Figs 4 and 5B and D). By increasing the number of distinct abundance profiles that can be generated, larger sample numbers increase both the sensitivity and resolution of covariance-based methods; one may therefore anticipate further gains in the application of such methods in relation to future increases in the scale of sequence data generated (i.e., increased cohort sizes).

We need however to acknowledge several important limitations that impede the overall performance and applicability of our bin-first framework. First, we already mentioned a limitation arising from the natural occurrence of strain-level variation at the origin of differential segregation of core and variable fractions of species pangenomes (Fig. 2). The large number

Downloaded from https://academic.oup.com/gigascience/article/9/4/giaa028/5812700 by CEA user on 20 October 2023

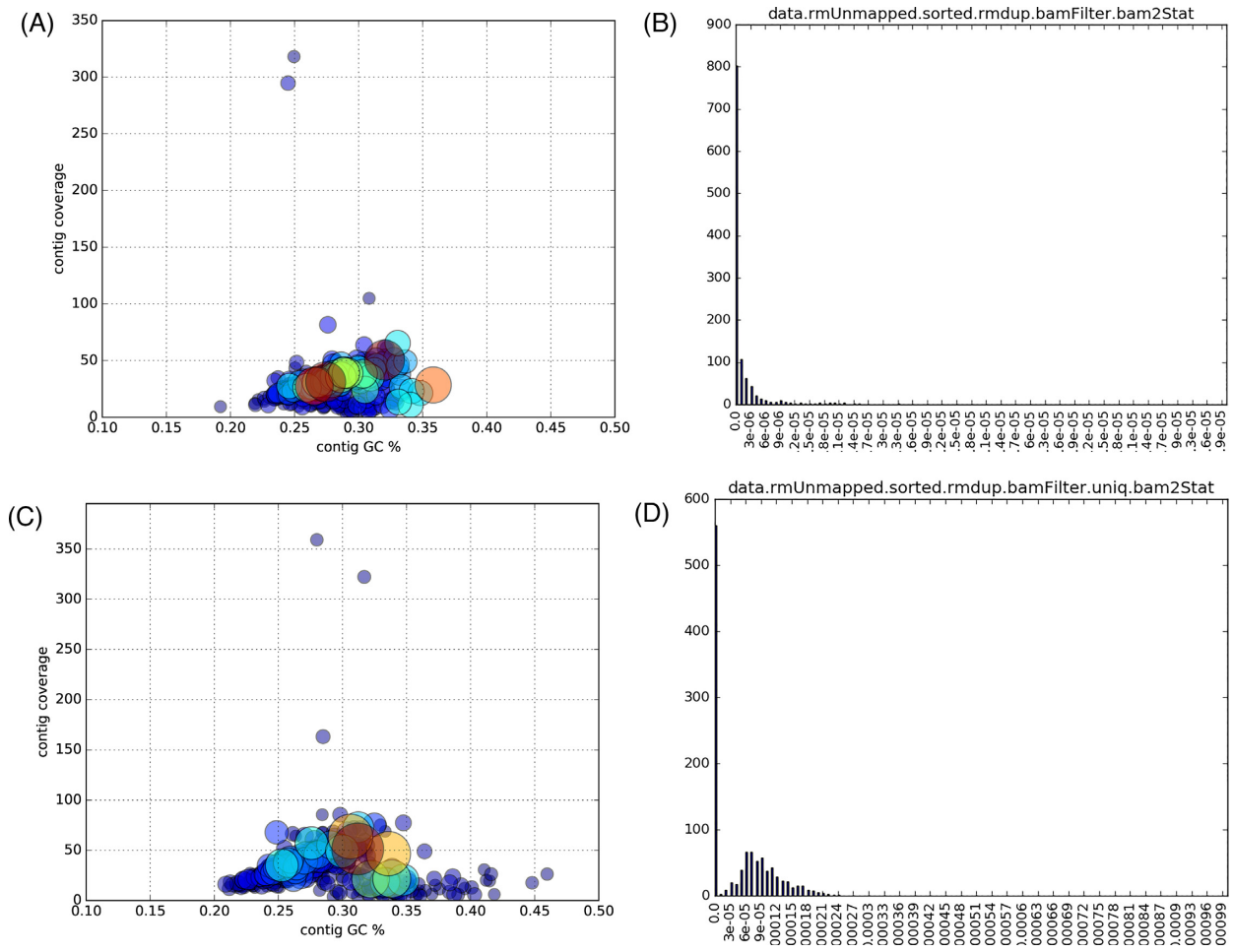

Figure 5: Left panels (A, C): GC-coverage plots (x-axis: contig GC%; y-axis: contig coverage) illustrating the homogeneity of 2 assembled bins (A, bin 470 [70% complete, 4.8% contamination]; C, bin 766 [70% complete, 3.5% contamination]) corresponding to 2 unclassified Firmicutes genomes of low abundance, whose enrichment histograms are shown in the corresponding right panel (B, D). Right panels (B, D): Enrichment histograms showing the fraction of raw reads contributed by each of the 1,135 samples to the 2 genomes whose GC-coverage plots are displayed in the corresponding left panel. x-axis: read abundance of genome bin 470 (B) and 766 (D); y-axis: sample frequency (among the 1,135 samples).

of incomplete but otherwise uncontaminated partitions/bins in the LifeLines DEEP analysis partly reflects the widespread occurrence of this type of variation in natural habitats. However, it should be noted that neither are assembly-based approaches immune to this type of variation, frequently discarding it when building "flattened" consensus contigs. This type of polymorphism is difficult to handle in a *de novo* way, and current methods for strain-level surveys of metagenomes typically rely on reference databases of strain-specific nucleotide polymorphisms (see, e.g., [30]). Sample-by-sample assembly limits the risk of strain mix-up, but at the expense of focusing on those genomes reaching high coverage (~10×). Our approach aimed at relaxing the latter constraint, but by doing so through the aggregation of lower-abundance reads across samples, it becomes vulnerable to extensive strain-level variation.

To the best of our knowledge, a method that could target—in an unsupervised way—low-coverage genomes in a strain-resolved manner is not available today, and working towards this goal is clearly a promising research area. It should be noted however that, to some extent, the degree of similarity that one wishes to distinguish can be tuned through the choice of the *k*-mer length and hash size. Increasing the *k*-mer size would increase the separation of closely related sequences, but only to some extent because the locality-sensitive hashing (LSH) scheme will inherently increase the probability of collision for similar sequences. Thus, we face here another trade-off: besides efficient in-memory indexing, the same LSH trick that allows convenient handling of sequencing errors (noise) can also put a limit on the power to separate very similar sequences (e.g., strains).

The observation that 4 of 7 genomes retrieved in the preliminary experiment based on 18 samples were not among the set of MAGs identified by analyzing the full dataset is indicative of a lack of stability of the algorithm. This effect of the sample number is most likely mediated by the increasing presence of strain variation when aggregating reads across increasing numbers of samples, leading to more fragmented partitions, and suggests that, above a certain level, increases in sample number can lead to diminishing returns in terms of complete genome recovery. We probably underestimated the extent of strain-level variation in real-world data, and the high level of genome fragmentation in the LifeLines DEEP partitions can be partly attributed to this problem, with low sequence coverage able to contribute as well.

Another limitation of the method is the generation of coarse-grained partitions harboring a large number of unresolved genomes (corresponding to the tail of the partition size distribu-

Downloaded from https://academic.oup.com/gigascience/article/9/4/giaa028/5812700 by CEA user on 20 October 2023

tion shown in Fig. 3). This problem is already manifest in the preliminary experiment comparing assembly-first versus bin-first approaches, and further exemplified in the large cohort analysis that yielded 983 partitions, 888 of which displayed low levels (<5%) of contamination (Table 3), but also produced several large clusters holding dozens of microbial genomes. The generation of such unresolved partitions seems difficult to avoid because the extent to which genomes differ from each other is variable across phylogenetic groups. As a result, it is unlikely that a single setting (e.g., k-mer length and hash size) could achieve perfect separation of genomes from highly diverse genome mixtures.

These 2 limitations probably concur to explain that the number of moderate to nearly complete genomes recovered from the population cohort analysis appears much lower than the number of "species genomes" recoverable via assembly-first approaches (remember, e.g., that close to 5,000 species-level genome bins were recovered from the analysis of nearly 10,000 metagenomes in [3]; one should however note that an average of 5.3 Gb per sample after quality control was generated in the latter study, vs 3.0 Gb before quality control in the LifeLines DEEP [18] data analyzed here).

When analyzing a large number of related samples, we noticed quite commonly that distinct organisms are able to reach a sufficiently high (to be assembled) relative abundance level in ≥1 sample (a situation exemplified in Fig. 4B). When following a sample-by-sample assembly-based strategy, a high coverage reached in a single sample (the likelihood of which increases with the number of samples analyzed) might be sufficient to assemble significant portions of a genome, even if it segregates at much lower levels in the remaining part of the cohort. This probably contributes to explain the high genome recovery yields of assembly-based approaches.

However, a key feature of the presented method is its ability to recover genomes of organisms consistently segregating at low levels across the entire cohort, as verified in a test experiment and on real-world data (cf. Fig. 4A). The observation that more than half of the genomes recovered here were not detected in a very large compendium of human gut genomes assembled from thousands of samples [3] is consistent with this view.

Metagenomic sequence binning is still a very active research field, and there are many interesting ongoing efforts, including some attempts to cast binning as an assembly graph partitioning problem [31]. Recent efforts include Brown et al. [32], which exploits the structural sparsity of compact de Bruijn assembly graphs to compute succinct indexes in linear time, allowing neighborhood queries to be performed on large assembly graphs in an "assembly-free" manner. One should note however that, even though this can leverage developments in efficient k-mer counting and graph compaction (e.g., [33]), assembling large multi-terabyte datasets can remain problematic in the first place. Nevertheless, most of the recent development efforts in the field of metagenomic sequence binning remain directed toward assembly-first approaches, which have already delivered a vast array of high-performing and user-friendly software [19,34–36], some of which have shown capabilities to recover genomes as low as 0.6% (10^{-3}) in relative abundance [35]. However, we have shown that the method presented here is able to recover genomes by sequence enrichments of the order of up to 10^{-6} (10^{-7} for some plasmid sequences) and therefore believe that it could be a useful adjunct to existing more mainstream approaches, especially for targeting more rare organisms. On the other hand, benefits of read-binning for comparative metagenomics have also been recently reported [37].

Potential Implications

As global metagenome assembly (and even more co-assembly) remains unpractical for multi-terabase–sized datasets, methods like the one described here—for which computer memory requirements remain independent of sequence depth—could prove valuable by making pre-assembly binning tractable while allowing researchers to gain access to genomes from the rare biosphere.

Methods

Control datasets

The control experiments used the dataset described in Gkanogiannis et al. [24] (and accessible from GigaScience's GigaDB [25]) corresponding to a virtual cohort of 50 individuals each harboring a microbiome of 100 distinct bacterial genomes sampled under a power-law abundance distribution (with power parameter $\alpha = 1.0$) from a pool of 750 fully sequenced genomes at an average depth of 10× (see [24] for details). We call these datasets semi-synthetic because they are made of real bacterial genome sequences assembled into artificial mixtures. The read to genome assignments (ground truth) being known in advance for all the reads, the precision and recall metrics were computed from the read clustering output as in equations (10) and (11) of Meyer et al. [38] (see section "Comparison of read binning algorithms"), with precision corresponding to what the authors refer to as purity and recall corresponding to completeness.

Real dataset: LifeLines-DEEP metagenomes

The LifeLines-DEEP cohort features 1,135 individuals (474 men and 661 women) from the general Dutch population, whose gut microbiomes were shotgun sequenced using the Illumina short-read technology, generating an average of 32 million reads per sample (see [18] and EBI dataset accession No. EGAD00001001991).

Locality-Sensitive Hashing

We used the SimHash [26] scheme described by Cleary et al. [13] to obtain a proxy for k-mer abundance. Briefly, raw reads are parsed into k-mers of fixed size ($k = 31$ was used in our experiments), the bases of which are individually mapped to a complex simplex via a mapping of the form $A = 1, C = i, G = -i, T = -1$ that can also incorporate base-call confidence scores [13]. k-mers are thus represented in k-dimensional space in which n hyperplanes (we used $n = 30$ in our experiments) are randomly drawn, creating 2^n subspaces, or buckets, indexing the columns of the sample by k-mer abundance matrix whose rows were scaled to unit ℓ_2 norm. The LSH scheme is sequence sensitive, increasing the probability of collision for more similar k-mers [26], and allows the representation of k-mer abundance matrices of arbitrary dimensions in fixed memory.

Regarding the choice of a k-mer length, the key requirement is that k-mers should be sufficiently long so that most of them will be specific to each genome, thereby capturing genuine abundance patterns of individual genomes. In our experiments, the k-mer length (31) was chosen to be close to the value used by Cleary et al. [13] to analyze their largest (terabase-sized) dataset. Some limited experiments with varying k-mer length values were carried out on smaller subsets of the data to check that

Downloaded from https://academic.oup.com/gigascience/article/9/4/giaa028/5812700 by CEA user on 20 October 2023

small variations in k-mer size did not result in disproportionate differences in clustering outputs.

In choosing the k-mer length, we were also guided by the observations in [39] that k-mer similarity between genomes at different k approximates various degrees of taxonomic similarity, with $k = 31$ appearing to correspond to species-level similarity. We also noticed that $k = 31$ is the default setting in the popular sequence classification engine kraken [29].

Sparse coding

Our aim is to learn sparse and non-negative factors from the sample by (hashed) k-mer abundance matrix **X**. The sparsity assumption has biological roots in the fact that every individual only harbors a small subset of all the genomes that constitute the global microbiome, while each genome only contains a very small subset of the k-mers encountered across all the samples. Sparse coding aims at modeling data vectors as sparse linear combinations of elements of a basis set (aka dictionary) that can be learned from the data by solving an optimization problem [15]. We used the SPAMS library [40], which implements the learning algorithm of [15]: given a training set $\mathbf{x}^1, \ldots, \mathbf{x}^n$ it tries to solve

$$\min_{\mathbf{D}\in C} \lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^{n} \min_{\alpha_i} \left[\frac{1}{2} \parallel \mathbf{x}_i - \mathbf{D}\alpha_i \parallel_2^2 + \psi(\alpha_i)\right],$$

where ψ is a sparsity-inducing regularizer (e.g., the ℓ_1 norm) and C is a constraint set for the dictionary (positivity constraints can be added to α as well). The following optimization scheme was used (FL stands for fused LASSO):

$$\min_{\mathbf{D}\in C} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2} \parallel \mathbf{x}_i - \mathbf{D}\alpha_i \parallel_2^2 + \lambda_1 \parallel \alpha_i \parallel_1 + \lambda_2 \parallel \alpha_i \parallel_2^2$$

with C a convex set verifying

$$C = \mathbf{D} \in R^{m\times p} \text{s.t.} \forall j, ||d_j||_2^2 + \gamma_1||\mathrm{d_j}||_1 + \gamma_2 \mathrm{FL}(d_j) \leq 1.$$

Once the dictionary has been learned, the SPAMS library offers an efficient implementation of the least-angle regression algorithm [41] for solving the LASSO or elastic net problem: given the data matrix **X** in $\mathbf{R}^{m\times n}$ and a dictionary **D** in $\mathbf{R}^{m\times p}$, this algorithm returns a matrix of coefficients $\mathbf{A} = [\alpha^1, \ldots, \alpha^n]$ in $\mathbf{R}^{p\times n}$ such that for every column **x** of **X**, the corresponding column α of **A** is the solution of

$$\min_{\alpha} \frac{1}{2}||\mathbf{x} - \mathbf{D}\alpha||_2^2 + \lambda_1||\alpha||_1 + \frac{1}{2}\lambda_2||\alpha||_2^2.$$

The SPAMS implementation of this algorithm allows the addition of positivity constraints on the solutions α, which have a natural interpretation as weighing the contribution of the different hashed k-mers to the latent genomes. In practice, we defined clusters by assigning hashed k-mers from bucket *i* to component *c* if $c = \mathrm{argmax}_j A_{i,j}$.

Read classification and assembly

Starting with the raw reads and their decomposition into k-mers, the bulk of the binning algorithm thus operates in k-mer space. After computing covarying k-mer sets ("eigengenomes"), a post-processing step is thus necessary to assign reads to their cognate k-mer clusters to achieve a read-level clustering. We stuck to the LSA procedure [13] for this step, with the original reads being assigned to k-mer clusters based on a log-likelihood score aggregating (i) cluster sizes (measured in terms of k-mer numbers), (ii) the overlap between k-mers in reads and those in clusters, and (iii) an inverse document frequency–style weight expressing the rarity of each of the overlapping k-mers. After read assignment, the partitions were assembled with the SPAdes (v3.13.0) engine [27] using default settings.

First experiment for comparing assembly-first versus bin-first protocols

An experimental set-up was designed to illustrate the ability of read binning to cluster rare reads from a target genome across samples, while assembly-first protocols are inoperable because the low coverage of the target genome prevents the generation of any kilobase-sized contig from the assembly of the individual samples.

The dataset consisted of 18 samples each containing a subset of 20,000 reads sampled from the 18 metagenomic libraries analyzed in Sharon et al. [42] and randomly spiked with mock reads from a *Bacillus thuringiensis* plasmid (NG_035027.1) as in the test data used in the original LSA paper [13]. However, as the number of spiked reads (up to 4,000) distributed among the samples in LSA's test dataset was sufficient to yield contigs covering a large fraction of the plasmid genome upon assembly, we derived a new dataset only containing 0–100 paired reads (14 samples contained 100 paired reads while 4 were entirely devoid of plasmid reads) and used the latter for this experiment. This dataset is available on the repository associated with this publication [25].

After checking that no kilobase-sized contig could be assembled in any of the samples—thus precluding the application of contig binning—the dataset was processed by our pre-assembly pipeline using the following settings: a k-mer length of 30 and a hash size of 22 were used to build the k-mer abundance matrix; the latter was decomposed by SVD and the columns of the eigen–k-mer matrix were clustered using a cosine similarity threshold of 0.25, followed by read assignment and assembly (using SPAdes) of the partitions. More than 99% (2,782 of 2,800) of the plasmid-derived reads ended up in a single partition (Supplementary Table), leading to the recovery of the complete target genome sequence upon assembly.

Second experiment for comparing assembly-first versus bin-first strategies

The raw sequence data from 18 (randomly chosen) individuals of the LifeLines DEEP cohort were either assembled individually (i.e., on a sample-by-sample basis) with metaSPAdes (v3.13.0) followed by contig binning across samples with the MetaBat2 adaptive algorithm [17], or the raw reads were clustered using our read-level binning pipeline, followed by metaSPAdes assembly of the resulting partitions/bins.

The raw reads were first mapped to the assembled contigs using bwa-mem [43] using default parameters. MetaBat2 was then invoked in the following way: first, the jgi_summarize_bam_contig_depths script was called to compute contig abundance statistics from the read mapping bam files, with the default options (minimum percent identity for a mapped read: 0.97; minimum contig length: 1,000; minimum contig depth: 1). The metabat2 program was then called using the default parameters (minCV 1.0, minCVSum 1.0, maxP 95%,

Downloaded from https://academic.oup.com/gigascience/article/9/4/giaa028/5812700 by CEA user on 20 October 2023

minS 60, and maxEdges 200) and the previously generated coverage statistics file, leading to the generation of 225 bins covering 694,000,907 bases.

For the comparison, our sparse coding pipeline was then executed under the same settings as in the full cohort analysis (hash size and k-mer size equal to 30 and 31, respectively, and default parameters for the dictionary learning and sparse decomposition of the abundance matrix), with the exception of the number of components that was matched to the number of bins (225) generated by MetaBat2. To generate Fig. 1, the complete genomes retrieved using both approaches were aligned (using nucmer [44] with default parameters) to individual assemblies from all the samples, and the number of distinct contig hits (≥99% identity and ≥2,500 bp) was recorded.

Comparison of read-binning algorithms

The virtual cohort dataset described above and in Gkanogiannis et al. [24] was used to compare the clustering accuracies of the original LSA [13] and sparse coding methods, as well as the performance of directly clustering the columns of the abundance matrix using a k-means algorithm as a baseline.

The read to genome memberships being comprehensively known in these controlled genome mixtures, clustering accuracy metrics (precision, recall, and F-measure) could be quantified as in Meyer et al. [38] (Table 1). Briefly, each bin is first mapped to its most abundant (in terms of number of reads) genome (note that if each bin is mapped to a single genome, a given genome can be mapped to multiple bins). Precision is defined as the ratio of reads originating from the mapped genome to all the bin's reads. Recall on the other hand reflects how complete a bin is with respect to the sequence of its cognate (mapped) genome. Average precision is the fraction of correctly assigned reads for all assignments to a given cluster averaged over all clusters, while average completeness is averaged over all genomes (including those possibly not assigned to any cluster). We follow Meyer et al. [38] to give larger bins higher weight in performance determinations. Specifically, if X is the set of clusters and Y the set of underlying genomes, precision and recall are defined, respectively, as:

$$p = \frac{\sum_{x \in X} TP_x}{\sum_{x \in X} TP_x + FP_x} = \frac{\sum_{x \in X} \max_y |x \cap y|}{\sum_{x \in X} |x|}$$

and

$$r = \frac{\sum_{y \in Y} \max_x |x \cap y|}{\sum_{y \in Y} |y|}.$$

The same k-mer abundance matrices (built using a k-mer size of 31 and a number of hash bits [hyperplanes] equal to 30) were used as input for all the methods.

Initial estimate of genome richness and number of components

For the test experiments based on synthetic microbiomes of controlled complexity (e.g., the virtual cohort of 50 individuals, where each microbiome consisted of 100 genomes drawn from a pool of 750 genomes under a given abundance distribution), the number of clusters was set to match the (known) number of distinct genomes segregating in the complete set of samples.

For the analysis of real-world data (the LifeLines DEEP cohort), where the total number of genotypes is unknown, a meaningful number of components for the sparse decomposition was estimated on the basis of the number of distinct rpS3 ribosomal protein sequences in the analyzed metagenomes, clustered at 98% identity, which roughly corresponds to species-level delineations according to Sharon et al. [45].

Evaluation of read enrichment levels

To assess whether we could identify genomes segregating at consistently low abundance levels in real-life datasets, we characterized the abundance of a dozen MAGs reconstructed from the LifeLines DEEP cohort analysis by directly mapping the raw reads from the original samples against them. Given the large size of the cohort (and the significant amount of computer resources associated with this analysis), and given that our objective was to establish whether consistently rare genomes can be identified by the method, this analysis was performed on a limited number of genomes.

Relative enrichment levels were estimated by mapping the original reads (after removal of duplicated reads) to the genome-resolved partitions using bwa-mem [43] with default parameters. Uniquely and consistently (i.e., paired) mapped reads were scored to compute enrichment ratios as the number of mapped reads divided by the number of raw reads analyzed, as displayed, e.g., on the x-axes of Figs 4 and 5B and D.

Comparison of genome-resolved partitions to reference genomes

To assess the novelty of the genomes assembled from individual partitions produced by our pipeline through the analysis of the LifeLines DEEP cohort, we screened them against 2 reference libraries. First, the genomes were compared to the Kraken2 (v1) database [46] built from NCBI's Refseq bacteria, archaea, and viral libraries (accessed October 2018), using the Kraken2 classifier [29] and a confidence score threshold of 0.2. Second, the same genomes were compared against the Human Gastrointestinal Bacteria Genome Collection [6] (HGG, encompassing >100 GB of sequence data) using the nucmer aligner [44] with default parameters. A genome was marked as already known if it shared ≥10 distinct 99% identity alignments of length ≥5 kb to any HGG entry.

Binning implementation

Code for the pipeline used to perform the analysis of the LifeLines DEEP cohort can be cloned from https://gitlab.com/kyrgyzov/lsa_slurm, while a more lightweight implementation of key algorithms (including sparse non-negative matrix factorization [NMF]) is available from [47]; they draw on the code base of the LSA tool ([13] and [48]) and on the SPAMS library that can be downloaded from [40]. The analysis of the metagenomes from the LifeLines DEEP cohort was carried out on a Bullion S6130 octo module server equipped with 2 Intel Xeon Haswell E7-4890 v3 CPU (18 cores) per module, 8 TB of RAM, and 35 TB storage. Most of the tasks being embarassingly parallelizable, they were run through a Slurm workload manager. The analysis took ∼3 weeks wall time, with the sparse decomposition of the k-mer abundance matrix taking <1 day. The bulk of the execution time was spent in pre- and post-processing tasks: pre-processing of the 10 TB of raw reads to improve load balancing (∼5 days), k-mer hashing and counting for constructing the k-mer abundance matrix (∼4.5 days), assignments of reads to eigengenomes following the sparse decomposition step (∼6 days), and assembly of in-

Downloaded from https://academic.oup.com/gigascience/article/9/4/giaa028/5812700 by CEA user on 20 October 2023

dividual read partitions using the SPAdes assembly engine [27] (~2.5 days).

A desirable feature when designing computational pipelines is to have resource requirements, especially memory, scale in a way independent of the sheer data volume. This is the case for the analytical method presented here because it can be executed "in memory" with the dimensionality of the empirical abundance matrix tailored via the LSH scheme to capture the desired amount of sequence diversity while remaining consistent with the available resource budget. The use of efficient online matrix factorization techniques [15] leads to limited memory footprints. Even though we leveraged here a powerful computer infrastructure to carry out the analysis of the large cohort dataset (10 TB of data), our pipeline is routinely executed on commodity hardware for smaller projects.

Availability of Source Code and Requirements

- Project name: Metagenomic read binning using sparse coding
- Project home page: https://gitlab.com/kyrgyzov/lsa_slurm
- Operating system(s): Linux
- Programming language: Python
- Other requirements: NumPy, SciPy, Gensim, SPAMS (https://gitlab.inria.fr/thoth/spams-devel)
- License: MIT License
- RRID:SCR_018134
- biotoolsID:Metagenomic_read_binning_using_sparse_coding

A lightweight implementation of key algorithms (including sparse NMF) is available from [47].

Availability of Supporting Data and Materials

Assembled sequences of the genome-resolved bins (>50% complete and with <5% contamination) recovered from the analysis of the LifeLines DEEP cohort are available via the *GigaScience* database [25]. The datasets used in the test experiments (virtual cohort and spiked datasets), as well as supporting data and an archival copy of the code, are also available via GigaDB [25].

Additional Files

Supplementary Table 1: Cluster assignments of reads from a target genome vs background (unrelated) reads. Nearly all the 2,800 reads from the target genome segregating at low levels in the samples (100 paired reads per sample in 14 samples; none in the remaining samples) are binned in a single partition using our bin-first pipeline, leading to the complete genome after assembly. No kilobase-sized contig could be assembled from any individual sample, making the assembly-first protocol inoperable.

Abbreviations

ANI: average nucleotide identity; bp: base pairs; CPU: central processing unit; Gb: gigabase pairs; GC: guanine-cytosine; kb: kilobase pairs; LASSO: least absolute shrinkage and selection operator; LSA: latent semantic analysis; LSH: locality sensitive hashing; MAG: metagenome-assembled genome; Mb: megabase pairs; NCBI: National Center for Biotechnology Information; NLP: natural langage processing; NMF: non-negative matrix factorization; RAM: random access memory; SPAdes: St. Petersburg genome assembler; SPAMS: Sparse Modeling Software; SVD: singular value decomposition.

Competing Interests

The authors declare that they have no competing financial interests.

Funding

This research was funded by the French Investments for the Future ("Investissements d'Avenir") program FSN-CISN2 (ADAMme project).

Authors' Contributions

T.B. conceived the project. O.K., V.P., and T.B. performed the analyses. B.F., S.G., and T.B. supervised the project. T.B. wrote the manuscript, with contributions from O.K. and V.P. All authors approved the final version of the manuscript.

Acknowledgments

We would like to express our sincere thanks to Brian Cleary and Eric Alm for their work on the LSA method, Julien Mairal and Ghislain Durif for developing the SPAMS library and for useful discussions, and Alexandre d'Aspremont for helpful insights. Finally, we thank both reviewers for valuable feedback that improved the manuscript.

References

1. Castelle CJ, Banfield JF. Major new microbial groups expand diversity and alter our understanding of the tree of life. Cell 2018;**172**(6):1181–97.
2. Parks DH, Rinke C, Chuvochina M, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol 2017;**2**(11):1533.
3. Pasolli E, Asnicar F, Manara S, et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell 2019;**176**:649–62.e20.
4. Almeida A, Mitchell AL, Boland M, et al. A new genomic blueprint of the human gut microbiota. Nature 2019;**568**(7753):499–504.
5. Nayfach S, Shi ZJ, Seshadri R, et al. Novel insights from uncultivated genomes of the global human gut microbiome. Nature 2019;**568**:505–10.
6. Forster SC, Kumar N, Anonye BO, et al. A human gut bacterial genome and culture collection for improved metagenomic analyses. Nat Biotechnol 2019;**37**(2):186.
7. Kallus Y, Miller JH, Libby E. Paradoxes in leaky microbial trade. Nat Commun 2017;**8**(1):1361.
8. Jousset A, Bienhold C, Chatzinotas A, et al. Where less may be more: how the rare biosphere pulls ecosystems strings. ISME J 2017;**11**(4):853.
9. Kalenitchenko D, Le Bris N, Peru E, et al. Ultrarare marine microbes contribute to key sulphur-related ecosystem functions. Mol Ecol 2018;**27**(6):1494–504.
10. Benjamino J, Lincoln S, Srivastava R, et al. Low-abundant bacteria drive compositional changes in the gut microbiota after dietary alteration. Microbiome 2018;**6**(1):86.
11. Wu YW, Ye Y. A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. J Comput Biol 2011;**18**(3):523–34.
12. Yang B, Peng Y, Leung HCM, et al. Unsupervised binning of environmental genomic fragments based on an error robust

Downloaded from https://academic.oup.com/gigascience/article/9/4/giaa028/5812700 by CEA user on 20 October 2023

selection of l-mers. BMC Bioinformatics 2010;**11**(2):S5.
13. Cleary B, Brito IL, Huang K, et al. Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. Nat Biotechnol 2015;**33**(10):1053.
14. Řehůřek R, Sojka P. Software framework for topic modelling with large corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta: ELRA. 2010:45–50.
15. Mairal J, Bach F, Ponce J, et al. Online learning for matrix factorization and sparse coding. J Mach Learn Res 2010;**11**:19–60.
16. Kang DD, Froula J, Egan R, et al. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ 2015;**3**:e1165.
17. Kang D, Li F, Kirton ES, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ 2019;**7**:e7359.
18. Zhernakova A, Kurilshikov A, Bonder MJ, et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science 2016;**352**(6285):565–69.
19. Alneberg J, Bjarnason BS, De Bruijn I, et al. Binning metagenomic contigs by coverage and composition. Nat Methods 2014;**11**(11):1144.
20. Chatterji S, Yamazaki I, Bai Z, et al. CompostBin: A DNA composition-based algorithm for binning environmental shotgun reads. In: Vingron M, Wong L , eds. Research in Computational Molecular Biology. Berlin, Heidelberg: Springer; 2008, doi:10.1007/978-3-540-78839-3˙3.
21. Van Lang T, Van Hoai T, et al. A two-phase binning algorithm using l-mer frequency on groups of non-overlapping reads. Algorithms Mol Biol 2015;**10**(1):2.
22. Girotto S, Pizzi C, Comin M. MetaProb: accurate metagenomic reads binning based on probabilistic sequence signatures. Bioinformatics 2016;**32**(17):i567–i575.
23. Tringe SG, Von Mering C, Kobayashi A, et al. Comparative metagenomics of microbial communities. Science 2005;**308**(5721):554–7.
24. Gkanogiannis A, Gazut S, Salanoubat M, et al. A scalable assembly-free variable selection algorithm for biomarker discovery from metagenomes. BMC Bioinformatics 2016;**17**(1):311.
25. Kyrgyzov O, Prost V, Gazut S, et al. Supporting data for "Binning unassembled short reads on the basis of *k*-mer covariance using sparse coding." GigaScience Database 2020; http://dx.doi.org/10.5524/100719.
26. Charikar MS. Similarity estimation techniques from rounding algorithms. Proceedings of the Thirty-Fourth Annual ACM Symposium on Theory of Computing. 2002:380–8.
27. Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 2012;**19**(5):455–77.
28. Parks DH, Imelfort M, Skennerton CT, et al. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 2015;**25**(7):1043–55.
29. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 2014;**15**(3):R46.
30. Luo C, Knight R, Siljander H, et al. ConStrains identifies microbial strains in metagenomic datasets. Nat Biotechnol 2015;**33**(10):1045.
31. Pell J, Hintze A, Canino-Koning R, et al. Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. Proc Natl Acad Sci U S A 2012;**109**(33):13272–77.
32. Brown CT, Moritz D, O'Brien M, et al. Exploring neighborhoods in large metagenome assembly graphs reveals hidden sequence diversity. BioRxiv 2019;462788.
33. Chikhi R, Limasset A, Medvedev P. Compacting de Bruijn graphs from sequencing data quickly and in low memory. Bioinformatics 2016;**32**(12):i201–8.
34. Albertsen M, Hugenholtz P, Skarshewski A, et al. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol 2013;**31**(6):533.
35. Wu YW, Tang YH, Tringe SG, et al. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome 2014;**2**(1):26.
36. Lu YY, Chen T, Fuhrman JA, et al. COCACOLA: binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge. Bioinformatics 2017;**33**(6):791–98.
37. Song K, Ren J, Sun F. Reads binning improves alignment-free metagenome comparison. Front Genet 2019;**10**:1156.
38. Meyer F, Hofmann P, Belmann P, et al. AMBER: assessment of metagenome binners. Gigascience 2018;**7**(6):giy069.
39. Koslicki D, Falush D. MetaPalette: A k-mer painting approach for metagenomic taxonomic profiling and quantification of novel strain variation. MSystems 2016;**1**(3):e00020–16.
40. Online resource for the SPArse Modeling Software (SPAMS). http://spams-devel.gforge.inria.fr/.
41. Efron B, Hastie T, Johnstone I, et al. Least angle regression. Ann Stat 2004;**32**(2):407–99.
42. Sharon I, Morowitz MJ, Thomas BC, et al. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. Genome Res 2013;**23**(1):111–20.
43. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics 2010;**26**(5):589–95.
44. Marçais G, Delcher AL, Phillippy AM, et al. MUMmer4: a fast and versatile genome alignment system. PLoS Comput Biol 2018;**14**(1):e1005944.
45. Sharon I, Morowitz MJ, Thomas BC, et al. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. Genome Res 2013;**23**(1):111–20.
46. Online resource for the Kraken 2 taxonomic classification software. https://ccb.jhu.edu/software/kraken2/. Accessed on, 2019/10/03.
47. Code repository for a lightweight implementation of sparse NMF based read binning. https://github.com/vincentprost/LSA_NMF.
48. Code repository for the "Latent Strain Analysis" (LSA) toolkit. https://github.com/brian-cleary/LatentStrainAnalysis.

Downloaded from https://academic.oup.com/gigascience/article/9/4/giaa028/5812700 by CEA user on 20 October 2023