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

**Olexiy Kyrgyzov**
CEA, DRF, Institut Jacob, Genoscope
CNRS-UMR8030, Université Paris-Saclay

**Vincent Prost**
CEA, DRF, Institut Jacob, Genoscope
CNRS-UMR8030, Université Paris-Saclay

**Stéphane Gazut**
CEA, LIST, LADIS

**Bruno Farcy**
Bull Atos Technologies

**Thomas Brüls***
CEA, DRF, Institut Jacob, Genoscope
CNRS-UMR8030, Université Paris-Saclay

July 24, 2019

## ABSTRACT

Sequence binning techniques enable the recovery of a growing 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.

We present here a pre-assembly binning scheme (i.e. operating on unassembled short reads) enabling latent genomes recovery by leveraging sparse dictionary learning and elastic-net regularization, and demonstrate its efficiency and scalability by recovering hundreds of metagenome-assembled genomes, including very low-abundance genomes, from a joint analysis of microbiomes from the LifeLines-Deep population cohort (n=1135, $> 10^{10}$ reads, 10 terabytes of sequence data).

***Keywords*** Metagenomics · Human microbiome · Sequence binning · Sparse coding

*To whom correspondence should be addressed.

## 1 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, requiring 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, 3, 4, 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. Though highly successful, such methods are nevertheless "inherently biased towards the most abundant organisms, meaning consistently less abundant organisms may still be missed" (quoted from ref[4]). For example, although thousands of MAGs were reconstructed from more than 1500 public metagenomes in the remarkable study [2], over 93% of these MAGs had an average coverage of more than 10x in at least one of the samples analyzed. The high proportions of phylogenetically unassigned reads typical in medium to high complexity metagenomes is another consequence of this limitation[6].

## 2 Rationale

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 to make the assembly problem more tractable by targeting lower complexity sequence subsets (bins). Binning unassembled reads is however more computationally demanding, as the number of raw sequences is typically orders of magnitude higher than the number of assembled contig sequences. A pioneering pre-assembly binning scheme[7] 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 is 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 and the size of which is kept tractable by indexing its columns using locality sensitive hashing (ref[7] and Methods). Matrix decomposition techniques are then used to define two sets of orthogonal latent vectors analogous to principal components of sample and sequence space. The large memory requirements incurred by the factorisation of large abundance matrices naturally drove ref[7] toward a rank-reduced singular value decomposition (SVD), for which efficient streaming libraries[8] 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

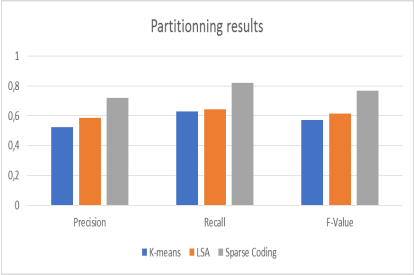



Figure 1: Binning accuracy estimates: LSA refers to the original algorithm of ref[7], with a cosine similarity threshold of 0.7 as recommanded by the authors, k-means refers to a direct clustering of the columns of the abundance matrix, with the number of clusters set to 1000 (equal to the number of components for the sparse decomposition), see main text and Methods.

from the right singular vectors matrix until about 0.4% of the latter has been covered. This heuristic is however acknowledged as a significant hindrance, the authors calling for "more sophisticated methods [are needed] to computationally discover a natural clustering" (quoted from ref[7]).

## 3 Results

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[9] 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 recover consistently less abundant genomes that are typically missed by assembly-first approaches.

### 3.1 Enhanced binning accuracy

To validate the new method, we first compared its read clustering accuracy (measured in terms of precision, recall and F-value metrics) with that of the original LSA method by using previously described benchmark datasets[10] (downloadable from http://www.genoscope.cns.fr/SCdata/vc50/), for which read to genome assignments were known (ref[10] and Methods). The results from these experiments are summarized in Figure 1, and show improved accuracy over both the canonical LSA and k-means based algorithms.

### 3.2 Partial strain separation

By design, locality sensitive hash functions increase the probability of collision for related items[11], which provides a convenient way to handle sequencing errors. However, the occurrence in natural environments of multiple strains from the same species

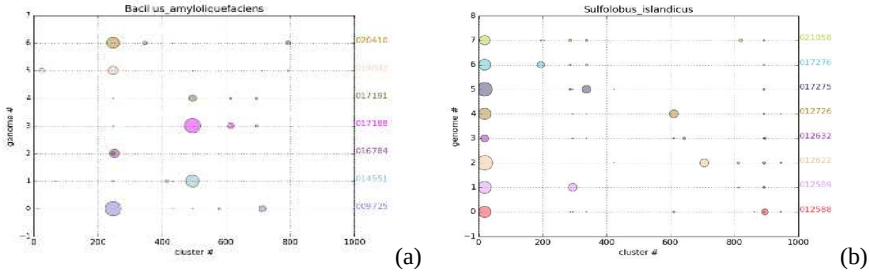



Figure 2: Partial resolution of species pangenomes. x-axis: partition identifier, y-axis: horizontal axes correspond to different strains from the same species (left: *B. amyloliquefaciens* starins, right: *S. islandicus* strains). Circle area is proportional to the number of reads from a given strain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

(the species pangenome) could lead to artefactual k-mer merging and potential overlap between the genomic partitions generated.

To assess the behavior of the method in the presence of extensive pangenomic variation, we quantified its ability to separate closely related (up to 98% average nucleotide identity) strains that were deliberately included in the genome mixtures from the virtual cohort used in the test experiments.

Results from these analyses are shown in Figure 2, which makes apparent a partial separation of the closely related strains (left panel), 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 (right panel).

### 3.3 Sensitivity and scalability on real-life data

To assess the sensitivity and scalability of the sparse coding method, we applied it to a real world dataset of over $10^{10}$ reads (about 10 terabytes of raw sequence data) derived from 1135 gut microbiomes of healthy Dutch individuals from the LifeLinesDeep cohort[12]. The pre-assembly binning of the cohort's reads resulted in 983 partitions, which were then assembled individually using the Spades engine[13] (Methods). The distribution of assembly sizes is shown in figure 2, making apparent that the vast majority of partitions are bacterial-genome sized (i.e. in the 2-5 Mbp range). A handful of coarse-grained partitions harboring unresolved genomes makes up the right tail of the distribution. As 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 toolbox[14]. A summary of completion and contamination statistics of the genome-resolved partitions is presented in Table 1, while another facet of the homogeneity of reconstructed genomes is displayed in the left panels of figure 5.

| Classification | Completeness | Genomes | Contamination |
|---|---|---|---|
| Nearly Complete | >90% | 14 | $\leq$ 5% |
| Substantial | >70% to $\leq$ 90% | 53 | $\leq$ 5% |
| Moderate | > 50% to $\leq$70% | 97 | $\leq$ 5% |
| Partial | $\leq$ 50% | 724 | $\leq$ 5% |

Table 1: Genome completion and contamination statistics of assembled partitions/bins, see main text and Methods.

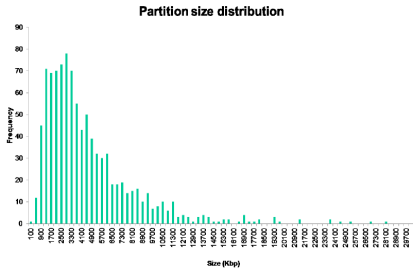



Figure 3: Distribution of assembled bin sizes. x-axis: assembled partition size (in kbp), y-axis: partition frequency

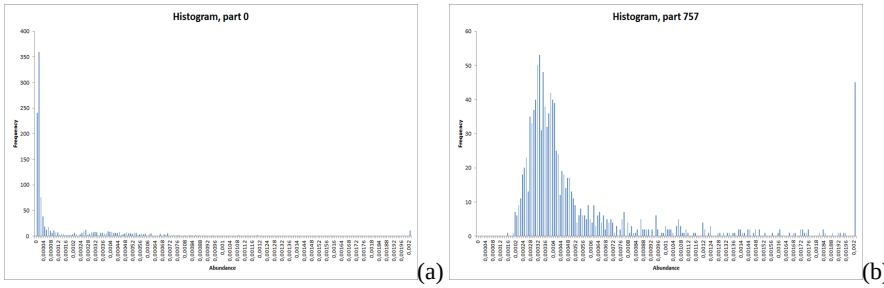
### 3.4 Recovery of very low-abundance genomes

A key motivation for the pre-assembly processing of reads was the theoretical possibility to aggregate reads from low abundance organisms. We therefore evaluated relative enrichment levels of a subset of partitions corresponding to $> 70\%$ complete genomes by measuring the fraction of raw reads contributed by each sample to these genome-resolved partitions (Methods). Given the large number of microbiomes analyzed, we quite frequently observed situations where a given genome reaches medium to high relative abundance in at least one sample (as illustrated in the right panel of Figure 4), but we could also detect instances of genomes that consistently segregated at low abundance levels across the whole cohort (left panel of Figure 4 and right panels of Figure 5).

The recovery of these genomes was made possible by aggregating a few thousands reads per sample across a large number of samples, thus demonstrating the ability of the method to recover substantial genome portions from rare organisms that would typically be missed by current sample by sample assembly-first approaches. Overall, the high proportion of homogeneous partitions corresponding to partial genomes (Table 1) is consistent with the recovery of sequences from lower abundance organisms, whose cumulated coverage across the cohort is not sufficient to allow complete genome reconstruction.

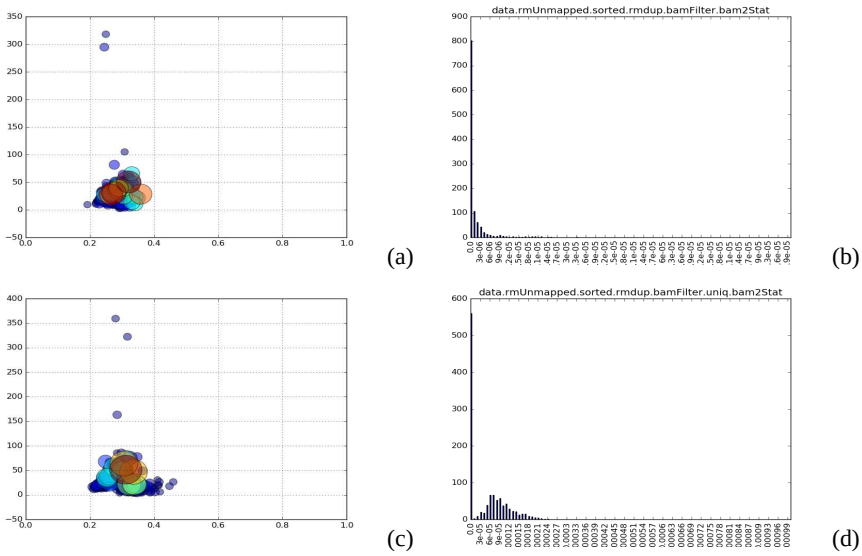
### 3.5 Assessing novelty against reference genome compendia

To investigate the extent to which the recovered genomes could correspond to novel organisms, we screened a subset of 164 of them (more than 50% complete with less than 5% contamination, that can be accessed at http://www.genoscope.cns.fr/SCdata/MAGs/) against several reference genome li-




(a) (b)

Figure 4: Enrichment histograms displaying the fraction of raw reads contributed by each sample to two distinct genome-resolved bins. x-axis: read abundance of partition 0 (left) and partition 757 (right); y-axis: sample frequency (among the 1135 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 rightmostest peak for the genome-resolved bin shown in panel b), while panels a illustrates that substantial (i.e. $\geq$70% complete) genomes of low-abundance organisms can also be recovered by aggregating only a few thousands reads per sample across the full cohort.




(a) (b) (c) (d)

Figure 5: Left panels (a,c): GC-coverage plots (x-axis: contig GC%; y-axis: contig coverage) illustrating the homogeneity of two assembled bins (top: bin 470, bottom: bin 766) corresponding to two 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 1135 samples to the two genomes whose GC-coverage plots are displayed in the corresponding left panel. x-axis: read abundance of partition 470 (panel b) and 766 (panel d); y-axis: sample frequency (among the 1135 samples).

braries. We first compared the genomes against the Kraken 2[15] database built from NCBI's Refseq bacteria, archaea and viral libraries (on October 2018). Only 21 out of the 164 genomes compared had at least one fragment classified against this reference database (Methods). We also compared the genomes against the "Global Human Gastrointestinal Bacteria Genome Collection" (HGG, ref[6]), that represents one of the most comprehensive resources of gastrointestinal bacterial reference sequences currently available. Only less than half (72/164) of the genomes displayed convincing similarity to the HGG genome catalogue (Methods).

## 4 Discussion

Covariance-based binning has the power to identify biologically meaningful associations between metagenomic sequences that could go unnoticed by analyses based on assembly or nucleotide signatures. This is illustrated in the present study by the recovery of hundreds metagenome-derived genomes, including from consistently less abundant organisms whose overall low coverage would prevent their reconstruction through metagenome assembly. As both the sensitivity and resolution of covariance-based methods increase with the number of samples, it is reasonable to anticipate further gains in the application of such methods in relation to future increases in the scale of sequence data generated (e.g. increased cohort sizes).

We already mentioned a limitation of the method arising from the natural occurrence of strain-level variation at the origin of differential segregation of core and variable fractions of species pangenomes (subsection 3.2 and Figure 2). Assembly-based approaches are neither 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 for example ref[16]).

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 distribution shown in Figure 3). This problem is examplified in the large cohort analysis, that yielded 983 (non empty) partitions, 888 of which displayed low-levels (<5%) of contamination (Table 1), but also produced a handful large clusters holding dozens of microbial genomes. The generation of such unresolved partitions is probably inevitable as the extent to which genomes differ to each other is variable across phylogenetic groups. As a result, there is no single setting (e.g. k-mer length, hash functions, number of hyperplanes) that could achieve perfect separation of genomes from highly diverse genome mixtures.

Overall, 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 for example that close to 5,000 species-level genome bins were recovered from the analysis of nearly 10,000 metagenomes in ref[3]. On one hand, the occurrence of coarse-grained partitions discussed above contributes to limit the resolution of the method. On the other hand, when analyzing a large number of related samples, it appears to be relatively common that distinct organisms are able to reach a sufficiently high (to be assembled) relative abundance level in at least one sample (a situation examplified in the right panel of Figure 4). When following a sample by sample assembly-based approach, a high coverage reached in a

single sample might be sufficient to assemble significant portions of a genome, even if it segregates at low levels in the vast majority of the individuals of the cohort.

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 illustrated in the left panel of Figure 4. 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.

Most of the recent development efforts in the field of metagenomic sequence binning are directed toward assembly-first approaches, which have delivered a vast array of performant and user-friendly software[17, 18, 19], some of which have shown capabilities to recover genomes as low as 0.6 % ($10^{-3}$)in relative abundance[18]. However, we have shown that the method presented here is able to reconstruct genomes by sequence enrichments of the order of up to $10^{-6}$ ($10^{-7}$ for some plasmid sequences), and therefore we believe that it could complement existing more mainstream approaches, especially for targeting more rare organisms.

**Potential implications** As global metagenome assembly (and even more coassembly) 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 post-binning assembly tractable while allowing to gain access to genomes from the rare biosphere.

# 5 Methods

## 5.1 Control datasets

The dataset described in ref[10] (and accessible from http://www.genoscope.cns.fr/SCdata/vc50/) 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 10x (see ref[10] for details), was used in the control experiments. The assignment of reads to genomes is exhaustively known for these metagenomes, that we call semi-synthetic because they are made of real bacterial genomes assembled into artificial mixtures.

## 5.2 Real dataset: LifeLines-DEEP metagenomes

The LifeLines-DEEP cohort features 1135 individuals (474 males and 661 females) from the general Dutch population whose gut microbiomes were analyzed using metagenomic shotgun sequencing with the Illumina short read technology, generating an average of 32 million reads per sample, see [12] and EBI dataset accession number EGAD00001001991.

## 5.3 Locality Sensitive Hashing (LSH)

We used the SimHash[11] scheme described in ref[7] 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[7]. 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. The rows of the k-mer abundance matrix were scaled to unit $\ell_2$ norm. The LSH scheme is sequence sensitive, increasing the probability of collision for more similar k-mers[11], and allows the representation of k-mer abundance matrices of arbitrary dimensions in fixed memory.

### 5.4 Sparse coding

We want to learn sparse and non negative factors from the sample by (hashed) k-mer abundance matrix $\mathbf{X}$. 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[9]. We used the spams library (http://spams-devel.gforge.inria.fr/), that implements the learning algorithm of ref[9]: given a training set $\mathbf{x}^1, ..., \mathbf{x}^n$ it tries to solve

$$\min_{\mathbf{D} \in C} \lim_{n \to +\infty} \frac{1}{n} \sum_{i=1}^{n} \min_{\alpha_i} \left( \frac{1}{2} \parallel \mathbf{x_i} - \mathbf{D}\alpha_\mathbf{i} \parallel_\mathbf{2}^\mathbf{2} + \psi(\alpha_\mathbf{i}) \right)$$

where $\psi$ is a sparsity-inducing regularizer (e.g. the $\ell_1$ norm) and *C* is a constraint set for the dictionary (positivity constraints can be added to $\alpha$ as well). The following optimization scheme was used (FL stands for fused lasso):

$$\min_{\mathbf{D} \in C} \frac{1}{n} \sum_{i=1}^{n} (1/2) \parallel \mathbf{x_i} - \mathbf{D}\alpha_\mathbf{i} \parallel_\mathbf{2}^\mathbf{2} + \lambda_\mathbf{1} \parallel \alpha_\mathbf{i} \parallel_\mathbf{1} + \lambda_\mathbf{2} \parallel \alpha_\mathbf{i} \parallel_\mathbf{2}^\mathbf{2}$$

with *C* a convex set verifying

$$C = \mathbf{D} \in \mathbf{R}^{\mathbf{m} \times \mathbf{p}} \mathbf{s.t.} \forall \mathbf{j}, ||\mathbf{d_j}||_\mathbf{2}^\mathbf{2} + \gamma_\mathbf{1} ||\mathbf{d_j}||_\mathbf{1} + \gamma_\mathbf{2} \mathbf{FL}(\mathbf{d_j}) <= \mathbf{1}$$

Once the dictionary has been learned, the spams library offers an efficient implementation of the LARS algorithm[20] for solving the Lasso or Elastic-Net problem: given the data matrix $\mathbf{X}$ in $\mathbf{R}^{m \times n}$ and a dictionary $\mathbf{D}$ in $\mathbf{R}^{m \times p}$, this algorithm returns a matrix of coefficients $\mathbf{A} = [\alpha^\mathbf{1}, ..., \alpha^\mathbf{n}]$ in $\mathbf{R}^{p \times n}$ such that for every column $\mathbf{x}$ of $\mathbf{X}$, the corresponding column $\alpha$ of $\mathbf{A}$ is the solution of

$$\min_{\alpha} (1/2) ||\mathbf{x} - \mathbf{D}\alpha||_\mathbf{2}^\mathbf{2} + \lambda_\mathbf{1} ||\alpha||_\mathbf{1} + (\mathbf{1/2}) \lambda_\mathbf{2} ||\alpha||_\mathbf{2}^\mathbf{2}$$

The spams implementation of this algorithm allows to add positivity constraints on the solutions $\alpha$, 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 $c = \mathrm{argmax}_j \, A_{i,j}$.

### 5.5 Read classification

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 co-varying k-mers sets ("eigengenomes"), a post-processing step is thus necessary to assign reads to their cognate k-mer partitions in order to achieve a read-level clustering. We sticked to the LSA procedure[7] 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, iii) an inverse document frequency (IDF)-style weight expressing the rarity of each of the overlapping k-mers (see ref[7] for details). After read assignment, the partitions were assembled with the Spades engine[13] using default settings.

### 5.6 Comparison of pre-assembly binning algorithms

The virtual cohort dataset described in section 5.1 was used to compare the clustering accuracies of the original LSA[7] and sparse coding methods, as well as the performance of a k-means clustering of the columns of the abundance matrix as a baseline. The reads to genome memberships being comprehensively known in these controlled genome mixtures, clustering accuracy metrics (precision, recall and F-measure) could be straightforwardly quantified (Figure 1). 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.

### 5.7 Initial estimate of genome richness in the LifeLines-DEEP microbiomes

A meaningful number of components for the sparse decomposition of the k-mer abundance matrix was initially estimated based on the number of distinct rpS3 ribosomal protein sequences (clustered at 98% identity) in the analyzed metagenomes.

### 5.8 Evaluation of read enrichment levels

Relative enrichment levels were estimated by mapping the original reads (after removal of duplicated reads) to the genome-resolved partitions using bwa-mem[21]. Uniquely and consistently (i.e. paired) mapped reads were scored to compute the enrichment ratios as the number of mapped reads divided by the number of raw reads analyzed (as displayed on the x-axis of Figure 4 and the right panels of Figure 5).

### 5.9 Comparison of genome-resolved partitions to reference genomes

To assess the novelty of the genomes assembled from individual partitions recovered from the analysis of the LifeLines-Deep cohort, we screened them against two reference libraries. First, the genomes were compared to the Kraken2 (v1) database (https://ccb.jhu.edu/software/kraken2/) built from NCBI's refseq bacteria, archaea, and viral libraries (in August 2018), using the Kraken2 classifier[15] and a confidence score threshold of 0.2. Second, the same genomes were compared against the Human Gastrointestinal Bacteria Genome Collection[6] (HGG, encompassing more than 100 GB of sequence data) using the nucmer aligner[22] with default parameters. A genome was marked as already known if it shared at least ten distinct 99% identity alignments of length $\geq$5 kb to any HGG entry.

### 5.10 Binning implementation

Code for the pipeline used in the analysis of the LifeLinesDeep cohort can be cloned from https://gitlab.com/kyrgyzov/lsa_slurm, while a more lightweight implementation of key algorithms (including sparse NMF) can be downloaded from https://github.com/vincentprost/LSA_NMF; they draw on the code base of the LSA method (ref[7] https://github.com/brian-cleary/LatentStrainAnalysis) and on the SPAMS (SPArse Modeling Software) library, that can be downloaded from http://spams-devel.gforge.inria.fr/. 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 about three weeks wall time, with the k-mer abundance matrix decomposition taking less than one day. The bulk of the execution time was spent in pre and post-processing tasks: pre-processing of the 10 terabytes of raw reads to improve load balancing ($\sim$5 days), k-mer hashing and counting for abundance matrix construction ($\sim$4.5 days), assignments of reads to eigengenomes following the sparse decomposition ($\sim$ 6 days), and assembly of individual read partitions using the spades assembly engine[13] ($\sim$2.5 days).

A desirable feature for scalable computational pipelines is that resource requirements, especially memory, scale in a way independent of the sheer data volume. This is the case for the analytical method presented here, that 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 relevant for a dataset and consistent with the available resource budget. The use of efficient online matrix factorization techniques[9] can lead to limited memory footprints. Even though we leveraged here a powerful computer infrastructure to carry out the analysis of the large cohort dataset (10 terabytes of data), our pipeline is routinely executed on commodity hardware for smaller projects.

### 5.11 Data access

Assembled sequences of the genome-resolved bins (more than 50% complete and with less than 5% contamination) recovered from the analysis of the LifeLines-DEEP cohort are available at http://www.genoscope.cns.fr/SCdata/MAGs/. The dataset corresponding to the virtual cohort used in the test experiments can be downloaded from http://www.genoscope.cns.fr/SCdata/vc50/.

**Acknowledgments** This research was funded by the French Investments for the Future ("Investissements d'Avenir") program FSN-CISN2 (ADAMme project). The authors wish to thank Brian Cleary and Eric Alm for pioneering pre-assembly binning, Julien Mairal and Ghislain Durif for developing the spams library and for useful discussions, and Alexandre d'Aspremont for helpful insights.

**Competing Interests** The authors declare no competing financial interests.

**Author's 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.

## References

[1] Cindy J Castelle and Jillian F Banfield. Major new microbial groups expand diversity and alter our understanding of the tree of life. *Cell*, 172(6):1181–1197, 2018.

[2] Donovan H Parks, Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J Woodcroft, Paul N Evans, Philip Hugenholtz, and Gene W Tyson. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature microbiology*, 2(11):1533, 2017.

[3] Edoardo Pasolli, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica Armanini, Francesco Beghini, Paolo Manghi, Adrian Tett, Paolo Ghensi, et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell*, 2019.

[4] Alexandre Almeida, Alex L Mitchell, Miguel Boland, Samuel C Forster, Gregory B Gloor, Aleksandra Tarkowska, Trevor D Lawley, and Robert D Finn. A new genomic blueprint of the human gut microbiota. *Nature*, page 1, 2019.

[5] Stephen Nayfach, Zhou Jason Shi, Rekha Seshadri, Katherine S Pollard, and Nikos Kyrpides. Novel insights from uncultivated genomes of the global human gut microbiome. *Nature*, page 1, 2019.

[6] Samuel C Forster, Nitin Kumar, Blessing O Anonye, Alexandre Almeida, Elisa Viciani, Mark D Stares, Matthew Dunn, Tapoka T Mkandawire, Ana Zhu, Yan Shao, et al. A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nature biotechnology*, 37(2):186, 2019.

[7] Brian Cleary, Ilana Lauren Brito, Katherine Huang, Dirk Gevers, Terrance Shea, Sarah Young, and Eric J Alm. Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nature biotechnology*, 33(10):1053, 2015.

[8] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

[9] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11(Jan):19–60, 2010.

[10] Anestis Gkanogiannis, Stéphane Gazut, Marcel Salanoubat, Sawsan Kanj, and Thomas Brüls. A scalable assembly-free variable selection algorithm for biomarker discovery from metagenomes. *BMC bioinformatics*, 17(1):311, 2016.

[11] Moses S Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 380–388. ACM, 2002.

[12] Alexandra Zhernakova, Alexander Kurilshikov, Marc Jan Bonder, Ettje F Tigchelaar, Melanie Schirmer, Tommi Vatanen, Zlatan Mujagic, Arnau Vich Vila,

Gwen Falony, Sara Vieira-Silva, et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science*, 352(6285):565–569, 2016.

[13] Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey A Gurevich, Mikhail Dvorkin, Alexander S Kulikov, Valery M Lesin, Sergey I Nikolenko, Son Pham, Andrey D Prjibelski, et al. Spades: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*, 19(5):455–477, 2012.

[14] Donovan H Parks, Michael Imelfort, Connor T Skennerton, Philip Hugenholtz, and Gene W Tyson. Checkm: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research*, 25(7):1043–1055, 2015.

[15] Derrick E Wood and Steven L Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, 15(3):R46, 2014.

[16] Chengwei Luo, Rob Knight, Heli Siljander, Mikael Knip, Ramnik J Xavier, and Dirk Gevers. Constrains identifies microbial strains in metagenomic datasets. *Nature biotechnology*, 33(10):1045, 2015.

[17] Johannes Alneberg, Brynjar Smári Bjarnason, Ino De Bruijn, Melanie Schirmer, Joshua Quick, Umer Z Ijaz, Leo Lahti, Nicholas J Loman, Anders F Andersson, and Christopher Quince. Binning metagenomic contigs by coverage and composition. *Nature methods*, 11(11):1144, 2014.

[18] Yu-Wei Wu, Yung-Hsu Tang, Susannah G Tringe, Blake A Simmons, and Steven W Singer. Maxbin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome*, 2(1):26, 2014.

[19] Yang Young Lu, Ting Chen, Jed A Fuhrman, and Fengzhu Sun. Cocacola: binning metagenomic contigs using sequence composition, read coverage, coalignment and paired-end read linkage. *Bioinformatics*, 33(6):791–798, 2017.

[20] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

[21] Heng Li and Richard Durbin. Fast and accurate long-read alignment with burrows–wheeler transform. *Bioinformatics*, 26(5):589–595, 2010.

[22] Guillaume Marçais, Arthur L Delcher, Adam M Phillippy, Rachel Coston, Steven L Salzberg, and Aleksey Zimin. Mummer4: a fast and versatile genome alignment system. *PLoS computational biology*, 14(1):e1005944, 2018.