# Binning unassembled short reads based on k-mer covariance in a sparse coding framework

Olexiy Kyrgyzov[1,2], Vincent Prost[1,3], Stéphane Gazut[3], Bruno Farcy[4] & Thomas Brüls[1,2]

[1]*CEA, DRF, Institut Jacob, Genoscope, 91057 Evry, France*

[2]*CNRS-UMR8030, Université Paris-Saclay, UEVE, 91057 Evry, France*

[3]*CEA, LIST, Laboratoire d'Analyse de Données et Intelligence des Systèmes, 91191 Gif-sur-Yvette, France*

[4]*Bull Atos Technologies, 78340 Les Clayes sous Bois, France*

**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. We present here a pre-assembly binning scheme enabling latent genomes recovery by leveraging sparsity and non-negativity constraints, and demonstrate its efficiency by recovering low-abundance genomes from a joint analysis of microbiomes from a large population cohort (LifeLines-Deep, n=1135).**

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 meaning-

ful 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. operated on assembled sequence contigs. 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 a further consequence of this limitation[6].

Considering that global metagenome assembly is currently unpractical to recover low abundance genomes or complex microbial consortia from terabytes of data, one may contemplate the possibility of resorting to 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 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 is to view metagenomes as linear mixtures of genomic variables, which 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 the authors toward a rank-reduced singular value decomposition (SVD), which also powers LSA in NLP applications and 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 were then recovered by an iterative sampling and merging heuristic that samples blocks of eigen k-mers from the right singular vectors matrix until about 0.4% of the latter has been covered. This heuristic is acknowledged as a significant hindrance, the authors calling for "more sophisticated methods [are needed] to computationally discover a natural clustering" (quoted from ref[7]).

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 matri-

ces; it also bypasses the aforementioned k-mer clustering heuristic and removes interpretability issues associated with the SVD (e.g. the physical meaning of negative contributions).

To validate the approach, we first compared its read clustering accuracy with that of the original LSA method by using previously described benchmark datasets for which read to genome assignments were known (ref [10] and Methods). Clustering accuracy metrics from these experiments are shown in panel a) of figure 1.

To assess the scalability and sensitivity of the sparse coding method, we applied it to a real dataset of over $10^{10}$ reads (about 10 TB raw sequence data) derived from 1135 gut microbiomes of healthy dutch individuals from the LifeLinesDeep cohort[11]. The pre-assembly binning of the LifeLinesDeep cohort's reads resulted in 983 partitions, which were then assembled individually (using the spades engine[12]). The distribution of assembly sizes is shown in panel b) of figure 1, making apparent that the vast majority of partitions are bacterial-genome sized (i.e. in the 2-5 Mbp range). As a direct read to genome mapping is not available for real life data, we assessed clustering performance by quantifying the genomic homogeneity and completeness of the resulting partitions based on the presence pattern of universal single-copy markers using the checkm toolbox[13]. A summary of completion and contamination estimates of the genome-resolved partitions is presented in Table 1, while another facet of the homogeneity of partitions is illustrated in the panels e) and g) of figure 1.

A key motivation for a pre-assembly processing of reads being the theoretical possibility to aggregate reads from low abundance organisms, we 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 panel c) of figure 1), but we could also detect instances of genomes that consistently segregated at low abundance levels across the whole cohort (as in panels d), f) and h)). 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.

[Figure 1]

(a) (b)

(c) (d)

(e) (f)

(g) (h)

To investigate the extent to which such genomic partitions could correspond to novel organisms, we screened a subset of 164 of them (more than 50% complete with less than 5% contamination) against several reference genome libraries. We first compared the genomes against

the Kraken 2[14] 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 contig 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. Less than half (72/164) of the genomes displayed convincing similarity to the HGG genome catalogue (Methods).

As global metagenome assembly (or coassembly) remains unpractical for multi terabase-sized datasets, methods like the one described here -for which memory requirements are independent of sequence depth- could prove valuable by making post-binning assembly tractable while allowing to access low-abundance genomes from the rare biosphere.

## Methods

**Control datasets** The dataset described in ref[10] 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, was used in the control experiments.

**LifeLines-DEEP cohort data** The LifeLines-DEEP cohort consists in 1135 individuals (474 males and 661 females) from the general Dutch population whose gut microbiomes were analyzed using metagenomic shotgun Illumina sequencing, generating an average of 32 million reads

per sample, see [11] and EBI dataset accession number EGAD00001001991.

**Locality Sensitive Hashing (LSH)** We used the same SimHash[15] scheme as in ref[7] to obtain a proxy for k-mer abundance. Briefly, raw reads are parsed into k-mers of fixed size, the bases of which are individually mapped to a complex simplex by also incorporating quality scores. k-mers are thus represented in k-dimensional space in which $n$ hyperplanes are randomly drawn, creating $2^n$ subspaces indexing the columns of the sample by k-mer abundance matrix. The LSH scheme is sequence sensitive, increasing the probability of collision for more similar k-mers[15], and allows the representation of k-mer abundance matrices of arbitrary dimensions in fixed memory. The k-mer abundance matrix is conditioned by scaling input row (samples) vectors to unit L2 norm.

**Sparse coding** We want to learn sparse and non negative factors from the k-mer abundance matrix. The sparsity assumption has biological roots in the fact that every individual only harbors a small subset of all the genomes forming 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 $l_1$-norm) and $C$ is a constraint set for the

dictionary (positivity constraints can be added to $\alpha$ as well). We used the following optimization scheme (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[16] 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 hashed k-mers to the latent genomes. Clusters were created by assigning k-mers to the latent component corresponding to each $\alpha$ vector's maximum value.

**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 calculation of 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) the cluster size (in terms of k-mer numbers), ii) the overlap between k-mers in reads and those in the clusters, iii) an IDF weight expressing the rarity of each of the overlapping k-mers.

**Comparison of pre-assembly binning algorithms** The virtual cohort dataset mentioned above was used to compare clustering accuracies of the original LSA[7] and sparse coding methods, as well as the performance of a direct k-means clustering of the columns of the abundance matrix. The reads to genome memberships being comprehensively known in these genome mixtures, clustering accuracy metrics (precision, recall and F-measure) could be straightforwardly quantified (panel a) of figure 1). In order to focus the comparison of the binning methods on their matrix decomposition core, the same k-mer abundance matrices (built using a k-mer size of 30 and a number of hash bits (hyperplanes) equal to 30) were used as input for all the methods.

**Initial estimate of the number of distinct genomes** A meaningful number of components was initially estimated based on the number of distinct rpS3 ribosomal protein sequences (clustered at 98% identity) in the analyzed microbiomes.

**Comparison of genome-resolved partitions to reference genomes** To assess the novelty of the genome-resolved partitions outputted by the LifeLines Deep cohort analysis, we screened the genomes assembled from individual partitions 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[14]

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[17] with default parameters. A genome was considered as already known if it shared at least ten 99% identity alignments of length $\geq$5 kb to a given HGG entry.

**Evaluating 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[18]. Uniquely and consistently (i.e. paired) mapped reads were scored to compute the enrichment ratios (number of mapped reads divided by number of analyzed raw reads) displayed in figure 1.

**Binning implementation** Code for the analysis pipeline is available at https://gitlab.com/kyrgyzov/xxx_slurm/; it draws on the code base of the original LSA method (ref[7] https://github.com/brian-cleary/LatentStrainAnalysis) and uses the SPAMS (SPArse Modeling Software) library that can be downloaded from http://spams-devel.gforge.inria.fr/. The analysis of the LifeLines-DEEP metagenomes 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 scheduling ($\sim$5 days), k-mer hashing and counting for abundance matrix construction ($\sim$4.5 days), assignments of reads to eigengenomes following the sparse matrix factorization step ($\sim$ 6 days), and assembly of individual read partitions ($\sim$2.5 days).

**Data access** Assembled sequences of the genome-resolved bins (more than 50% complete and less than 5% contamination) recovered from the analysis of the LifeLines-DEEP cohort are available at https://www.genoscope.cns.fr/xxx/lldeep_mags.tar.gz

**Acknowledgements** 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 Alex d'Aspremont for helpful insights.

**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 inputs from O.K. and V.P. All authors approved the final version of the manuscript.

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

**Correspondence** Correspondence should be addressed to T.B. (email: bruls@genoscope.cns.fr).

1. Castelle, C. J. & Banfield, J. F. Major new microbial groups expand diversity and alter our understanding of the tree of life. *Cell* **172**, 1181–1197 (2018).

2. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature microbiology* **2**, 1533 (2017).

3. Pasolli, E. *et al.* Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* (2019).

4. Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* 1 (2019).

5. Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S. & Kyrpides, N. Novel insights from uncultivated genomes of the global human gut microbiome. *Nature* 1 (2019).

6. Forster, S. C. *et al.* A human gut bacterial genome and culture collection for improved metagenomic analyses. *Nature biotechnology* **37**, 186 (2019).

7. Cleary, B. *et al.* Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. *Nature biotechnology* **33**, 1053 (2015).

8. 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*, 45–50 (ELRA, Valletta, Malta, 2010). `http://is.muni.cz/publication/884893/en`.

9. Mairal, J., Bach, F., Ponce, J. & Sapiro, G. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research* **11**, 19–60 (2010).

10. Gkanogiannis, A., Gazut, S., Salanoubat, M., Kanj, S. & Brüls, T. A scalable assembly-free variable selection algorithm for biomarker discovery from metagenomes. *BMC bioinformatics* **17**, 311 (2016).

11. Zhernakova, A. *et al.* Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565–569 (2016).

12. Bankevich, A. *et al.* Spades: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology* **19**, 455–477 (2012).

13. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. Checkm: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research* **25**, 1043–1055 (2015).

14. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology* **15**, R46 (2014).

15. Charikar, M. S. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, 380–388 (ACM, 2002).

16. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. *et al.* Least angle regression. *The Annals of statistics* **32**, 407–499 (2004).

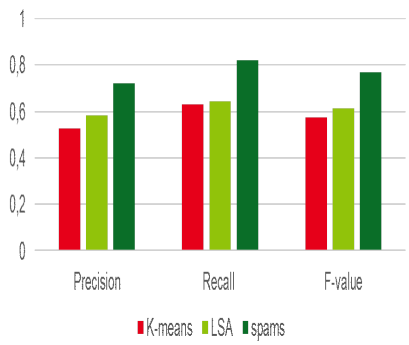
17. Marçais, G. *et al.* Mummer4: a fast and versatile genome alignment system. *PLoS computational biology* **14**, e1005944 (2018).

18. Li, H. & Durbin, R. Fast and accurate long-read alignment with burrows–wheeler transform. *Bioinformatics* **26**, 589–595 (2010).

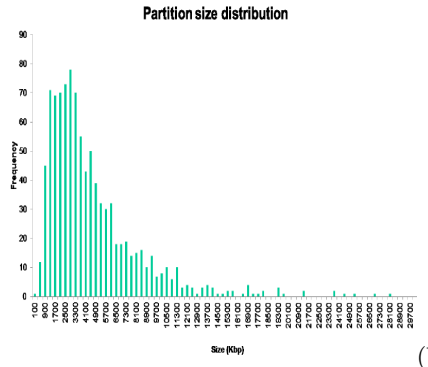
**Figure 1** (a): Binning accuracy estimates; LSA refers to the original algorithm of ref[7], with a cosine similarity threshold of 0.7 as recommanded by the authors, k-means refers to a direct clustering of the columns of the abundance matrix, with the number of clusters set to 1000 (equal to the number of components for the sparse factorization), see main text and Methods. (b): Distribution of assembled bin sizes. (c, d, f, h): Enrichment histograms, displaying the fraction of raw reads contributed by each sample to distinct genome-resolved bins. 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 c), while panels d, f and h illustrate that substantial ($\sim$70% complete) genomes of low-abundance organisms can be recovered by aggregating only a few thousands reads per sample across the cohort. (e, g): GC-coverage plots (x-axis: contig GC%; y-axis: contig coverage) illustrating the homogeneity of bins corresponding to the two low-abundance genomes whose enrichment histograms are shown in panels f and h respectively.

| 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% |

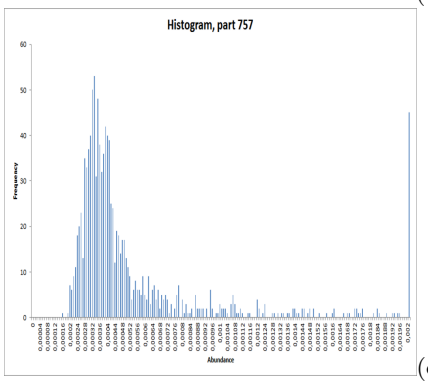
Table 1: Genome completion statistics of assembled partitions, see main text.


1
0,8
0,6
0,4
0,2
0
Precision
Recall
F-value
K-means
LSA
spams

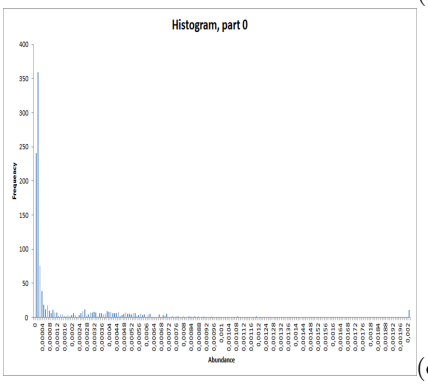
(a)


Partition size distribution
Frequency
Size (Kbp)

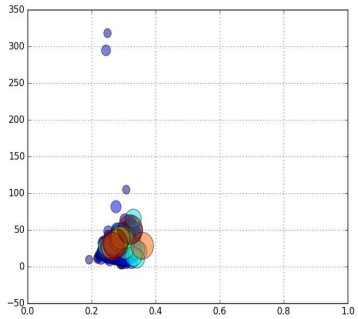
(b)


Histogram, part 757
Frequency
Abundance

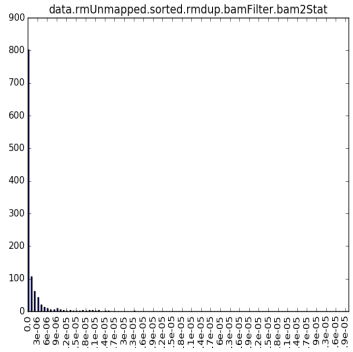
(c)


Histogram, part 0
Frequency
Abundance

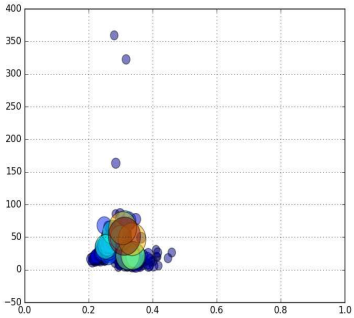
(d)



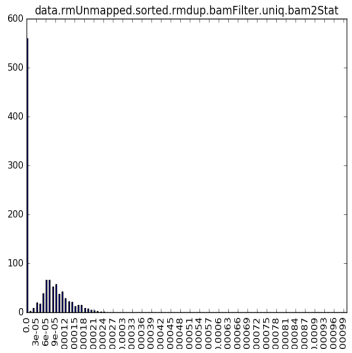
(e)


data.rmUnmapped.sorted.rmdup.bamFilter.bam2Stat

(f)



(g)


data.rmUnmapped.sorted.rmdup.bamFilter.uniq.bam2Stat

(h)