saul.sarria@fda.hhs.gov

# Scalable Workflow for Genome Binning on HPC Clusters

Saul Sarria[1], Mike Mikailov[2], Caitlin A. Welsh[1], Fu-Jyh Luo[2], Kenny Cha[2], Jiuzhou Song[3], Daniel A. Tadesse[1]

[1]Food and Drug Administration, Center for Veterinary Medicine, Office of Applied Science, Laurel, MD, USA [2]Center for Devices and Radiological Health Office of Science and Engineering Laboratories. Silver Spring, MD, USA [3]College of Computer, Mathematical, and Natural Sciences. Biological Science Graduate Program, University of Maryland, College Park, MD, USA

## Abstract

**Introduction:** Metagenomics is a culture-independent method for studying microbes directly in their natural environments. It provides a snapshot of the microbial community composition, its genomic contents including resistome, virulome and mobilome, and the functional profiles of the microbial community. However, the assembly of high-quality genomes and linking antimicrobial resistance genes to host bacteria and plasmids are major challenges in metagenomics. High-throughput proximity ligation (Hi-C) has the capacity to link DNA fragments within cells, and when coupled with metagenomic shotgun sequencing, shows remarkable genome binning and high-quality metagenome-assembled genomes (MAG) reconstruction from microbiome samples. The bin3C pipeline applies an unsupervised method that exploits the hierarchical nature of Hi-C proximity ligation interaction rates to resolve MAGs.

**Methods**: We randomly selected 21 cecal samples from food animal sources cattle, swine, chicken, and turkey. Shotgun and Hi-C metagenomics libraries were sequenced on the Illumina HiSeq 2500 platform, generating >75 million reads/sample for Hi-C and >100 million reads/sample for shotgun libraries. Hi-C based proximity-guided assembly was used to deconvolute the shotgun assembly into genomes using ProxiMeta software. To increase efficiency, we developed a scalable bin3C workflow that employs a data-driven parallelization technique for efficient and distributed processing of Hi-C and deep shotgun metagenomics data across massively parallel High-Performance Computing (HPC) clusters. The presence and abundance of antimicrobial resistance (AMR) genes were determined using the Short, Better Representative Extract Dataset (ShortBRED).

**Results:** We developed a scalable workflow leveraging the job array functionality of job schedulers for dynamic and efficient resource allocation. Each job array task processes a single input dataset, with two independent job arrays of M and N tasks running concurrently to handle M Hi-C and N deep shotgun metagenomics datasets, respectively. The job scheduler's job dependency mechanism is used to synchronize these independent task streams before Hi-C read mapping. The overall speedup scales proportionally to O(M+N). Unlike traditional methods, such as the widely used Message Passing Interface (MPI), this approach is more fault-tolerant (only requiring failed tasks to be rerun), utilizes system resources dynamically - automatically queuing and starting tasks as resources become available, and avoids the MPI "job starvation" phenomenon. The developed workflow was used to analyze shotgun and Hi-C data generated from 21 randomly selected samples. The bioinformatics analysis revealed over 200 bins containing metagenome assembled genomes (MAGs). We reconstructed 130 near-complete MAGs with high level of confidence (>90% Completeness and < 10% contamination). Resistome analysis revealed wide ranging AMR genes representing 18 antimicrobial classes including: aminoglycoside, beta-lactams, colistin, macrolide, quinolone and tetracycline.

**Conclusions**: Parallel processing of 21 samples on CDRH HPC cluster Betsy (5,000 CPU cores) yielded > 10x speedup at initial steps; > 426x speedup at annotation step. A scalable bin3C innovative approach harnesses the power of the CDRH High-Performance Computing (HPC) Clusters to parallelize, scale, and accelerate large-scale computations, enabling tasks that were once impractical due to time limitations. Metagenomics provides valuable insights into the diversity and identity of AMR genes AMR genes present in samples.

## Background

Metagenomics is a cultivation-independent method for studying microbes directly in their natural environments. It provides a comprehensive means of deciphering the antimicrobial resistance potential for entire microbial communities. Despite recent advances in metagenomics, reconstruction of microbial genomes and linking resistance genes to plasmid and source bacteria from metagenomics dataset remains a challenge. High-throughput chromosome conformation capture (Hi-C) metagenome deconvolution combines proximity ligation with deep shotgun sequencing data to assemble high-quality metagenomes and associate mobile genetic elements with their hosts. We employed deep shotgun metagenomic and proximity ligation (Hi-C) chemistry sequencing to assemble genomes from metagenomic samples. These dataset requires large-scale computations on HPC clusters. Message Passing Interface (MPI), the most widely used HPC parallelization method, can lead to job "starvation," is unable to utilize dynamically available resources, and has limited scalability. To address these limitations, we leverage the job array feature of HPC job schedulers.

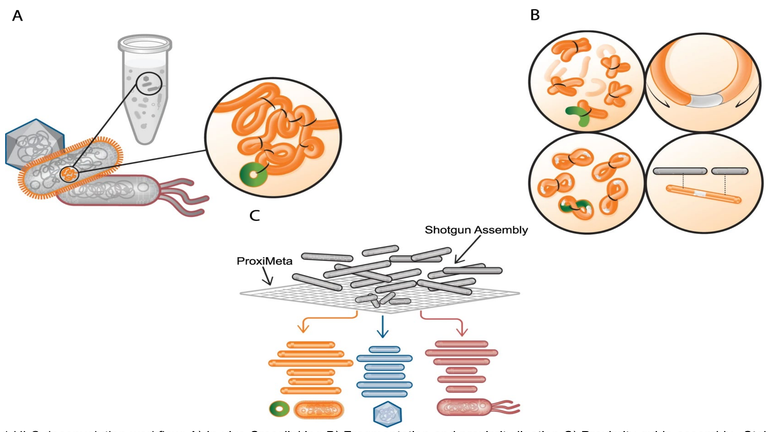


Figure 1 Hi-C deconvolution workflow: A) In-vivo Crosslinking B) Fragmentation and proximity ligation C) Proximity guide assembly. Stalder et al, 2019.

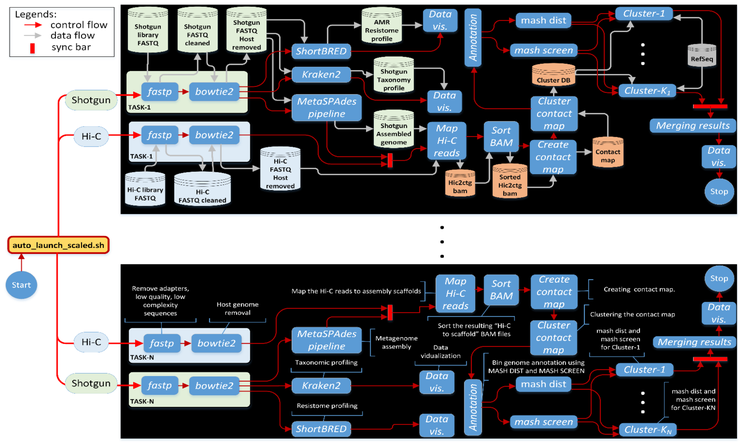


Figure 2 The scalable workflow for genome binning on HPC environment

## Results

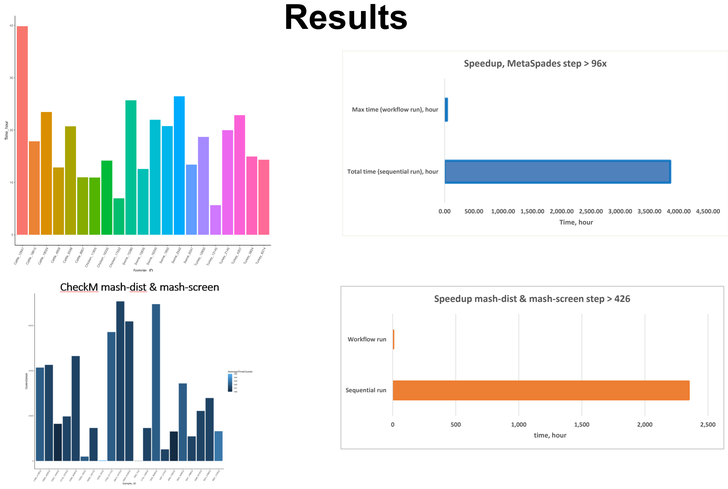


Figure 3. Metrics showing improved efficiency gained by scalable workflow

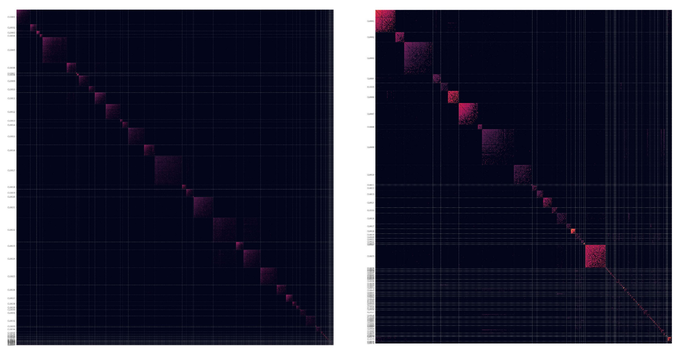

Figure 4. Hi-C contact map that shows pairs of genomic positions adjacent to each other in three-dimensional space

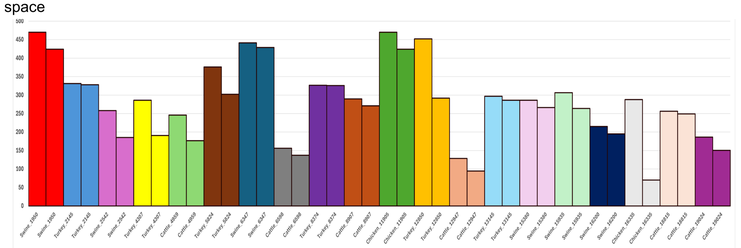

Figure 5. Average contigs length per sample reconstructed using Hi-C guided shotgun assembly

| Sample ID | Bin ID | Genomes | Completeness | Contamination | Genome size (bp) | # scaffolds | N50 (scaffolds) | Longest scaffold (bp) | GC |
|---|---|---|---|---|---|---|---|---|---|
| 2542 | CL0010 | o__Chlamydiales (UID2994) | 99.1 | 0.0 | 1084648 | 18 | 164610 | 258002 | 42.0 |
| 2542 | CL0001 | o__Clostridiales (UID1120) | 96.2 | 0.0 | 1816456 | 139 | 27660 | 131922 | 42.0 |
| 4959 | CL0003 | o__Actinomycetales (UID2012 | 99.3 | 0.9 | 4006330 | 108 | 102107 | 245919 | 68.0 |
| 4959 | CL0001 | f__Streptomycetaceae (UID204 | 95.5 | 0.4 | 5762750 | 555 | 17823 | 66406 | 72.0 |
| 5824 | CL0009 | o__Clostridiales (UID1212) | 97.3 | 0.0 | 2273391 | 37 | 110310 | 232866 | 45.1 |
| 5824 | CL0001 | o__Bacteroidales (UID2716) | 96.8 | 1.9 | 3684059 | 117 | 68663 | 191159 | 48.0 |
| 5824 | CL0004 | p__Bacteroidetes (UID2605) | 96.4 | 1.0 | 2590648 | 43 | 119087 | 376165 | 60.0 |
| 6347 | CL0002 | p__Bacteroidetes (UID2605) | 98.5 | 0.7 | 3389572 | 42 | 140501 | 428692 | 57.0 |
| 6347 | CL0006 | f__Lachnospiraceae (UID1256 | 98.2 | 0.8 | 2671474 | 330 | 14060 | 102147 | 40.0 |
| 6347 | CL0004 | k__Bacteria (UID2982) | 98.0 | 0.0 | 3019513 | 54 | 163957 | 441298 | 54.0 |
| 6347 | CL0013 | o__Clostridiales (UID1212) | 97.3 | 0.0 | 1919961 | 37 | 106873 | 319229 | 53.0 |
| 6598 | CL0012 | p__Actinobacteria (UID2112) | 95.0 | 3.4 | 2738177 | 588 | 6772 | 42057 | 58.0 |
| 6598 | CL0020 | o__Clostridiales (UID1120) | 90.1 | 0.7 | 2098332 | 442 | 7702 | 35007 | 48.0 |
| 8347 | CL0004 | o__Clostridiales (UID1212) | 96.9 | 1.4 | 2325707 | 214 | 21385 | 113432 | 55.5 |
| 8347 | CL0016 | o__Clostridiales (UID1212) | 96.6 | 0.2 | 1729906 | 145 | 23019 | 78725 | 51.7 |
| 8347 | CL0010 | o__Clostridiales (UID1212) | 94.0 | 0.0 | 2012053 | 92 | 29613 | 137129 | 62.7 |
| 8347 | CL0008 | o__Bacteroidales (UID2617) | 93.7 | 0.1 | 2050186 | 53 | 55878 | 203566 | 46.6 |
| 8907 | CL0004 | c__Deltaproteobacteria (UID32 | 99.3 | 0.9 | 2837866 | 89 | 50723 | 251671 | 53.9 |
| 8907 | CL0001 | c__Clostridia (UID1118) | 99.2 | 1.3 | 3298764 | 136 | 46802 | 174214 | 56.8 |
| 8907 | CL0005 | o__Clostridiales (UID1212) | 98.4 | 6.3 | 2798880 | 96 | 68539 | 219819 | 53.9 |
| 8907 | CL0002 | o__Clostridiales (UID1212) | 98.1 | 1.0 | 3246994 | 131 | 47975 | 155355 | 53.1 |
| 8907 | CL0003 | o__Clostridiales (UID1212) | 96.6 | 1.3 | 2911787 | 47 | 93012 | 242781 | 51.6 |
| 11905 | CL0020 | o__Selenomonadales (UID1024 | 99.4 | 1.2 | 1945558 | 18 | 287869 | 417009 | 45.0 |
| 11905 | CL0027 | g__Lactobacillus (UID380) | 96.6 | 1.3 | 1523223 | 156 | 17999 | 74443 | 34.0 |
| 11905 | CL0004 | o__Bacteroidales (UID2657) | 94.4 | 0.7 | 3559935 | 64 | 96070 | 196133 | 45.0 |
| 12850 | CL0002 | p__Bacteroidetes (UID2605) | 98.3 | 0.0 | 2722336 | 34 | 113749 | 247913 | 57.6 |
| 13145 | CL0007 | p__Actinobacteria (UID2112) | 98.4 | 1.1 | 1795608 | 76 | 80331 | 296956 | 51.0 |
| 13145 | CL0004 | p__Bacteroidetes (UID2605) | 95.7 | 1.4 | 2481390 | 35 | 133253 | 263356 | 57.9 |
| 15380 | CL0001 | f__Enterobacteriaceae (UID516 | 99.5 | 2.2 | 6438707 | 682 | 22372 | 109263 | 49.8 |
| 15380 | CL0017 | f__Bifidobacteriaceae (UID145 | 99.1 | 0.8 | 2163878 | 54 | 86190 | 286097 | 59.4 |
| 15380 | CL0052 | c__Clostridia (UID1118) | 98.4 | 1.8 | 1607157 | 233 | 10041 | 60014 | 53.8 |
| 15380 | CL0024 | p__Actinobacteria (UID2112) | 98.4 | 0.0 | 2057301 | 84 | 55981 | 266247 | 56.7 |
| 15380 | CL0056 | o__Lactobacillales (UID355) | 95.1 | 1.1 | 1539374 | 179 | 17856 | 65035 | 49.8 |
| 15935 | CL003 | o__Pseudomonadales (UID4679) | 99.7 | 0.1 | 3153942 | 133 | 49214 | 131559 | 41.4 |
| 15935 | CL009 | c__Bacilli (UID284) | 92.4 | 0.7 | 2440825 | 241 | 18485 | 96918 | 36.3 |
| 16200 | CL0001 | o__Bacteroidales (UID2621) | 98.9 | 0.4 | 3325418 | 173 | 32026 | 129965 | 44.6 |
| 16200 | CL0007 | p__Actinobacteria (UID2112) | 98.8 | 0.8 | 1994938 | 328 | 9866 | 38404 | 59.2 |
| 16200 | CL0005 | f__Lachnospiraceae (UID1256 | 98.0 | 2.1 | 2432660 | 304 | 14867 | 49282 | 54.0 |
| 16200 | CL0020 | c__Clostridia (UID1118) | 92.3 | 0.6 | 1620412 | 171 | 14539 | 47671 | 51.9 |
| 17552 | CL0003 | k__Bacteria (UID2372) | 90.5 | 1.1 | 1817629 | 151 | 21869 | 88483 | 55.5 |

Table 1. Selected list of genomes reconstructed using MAGs with high level of confidence (>90% Completeness, < 10% contamination).

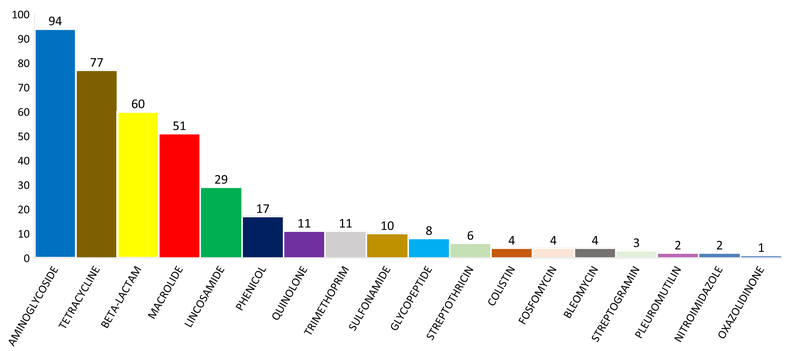


Figure 6. Number of antimicrobial resistance gene variants by drug class

## Conclusions

- Parallel processing of 21 samples on CDRH HPC cluster Betsy (5,000 CPU cores) yielded > 10x speedup at initial steps; > 426x speedup at annotation step.
- A scalable bin3C innovative approach harnesses the power of the CDRH High-Performance Computing (HPC) Clusters to parallelize, scale, and accelerate large-scale computations, enabling tasks that were once impractical due to time limitations.
- Metagenomics provides valuable insights into the diversity and identity of AMR genes present in samples. When coupled with Hi-C sequencing, it enables linking AMR genes to plasmids and host bacterial cells and facilitates public health decisions aimed at reducing the source and exposure routes of antimicrobial resistance to humans.

## Disclaimer

The views expressed in this poster are those of the author(s) and may not reflect the official policy of the Department of Health and Human Services, the U.S. Food and Drug Administration or the U.S. Government.