

1 **dRep: A tool for fast and accurate genome de-replication that enables tracking of microbial**
2 **genotypes and improved genome recovery from metagenomes**

3

4 Matthew R. Olm¹, Christopher T. Brown¹, Brandon Brooks¹, and Jillian F. Banfield^{2,3*}

5

6 ¹Department of Plant and Microbial Biology, University of California, Berkeley, CA, USA.

7 ²Department of Environmental Science, Policy, and Management, University of California,

8 Berkeley, CA, USA

9 ³Department of Earth and Planetary Science, University of California, Berkeley, CA, USA.

10

11 *Corresponding author

12

13 **Running title**

14 dRep: De-replication of microbial genomes.

15

16 **Corresponding author contact info**

17 Jillian Banfield

18 Department of Environmental Science, Policy, & Management

19 UC Berkeley

20 369 McCone Hall

21 Berkeley, CA 94720

22 Jbanfield@berkeley.edu

23 **The number of microbial genomes sequenced each year is expanding rapidly, in part due to**
24 **genome-resolved metagenomic studies that routinely recover hundreds of draft-quality**
25 **genomes. Rapid algorithms have been developed to comprehensively compare large**
26 **genome sets, but they are not accurate with draft-quality genomes. Here we present dRep,**
27 **a program that sequentially applies a fast, inaccurate estimation of genome distance and a**
28 **slow but accurate measure of average nucleotide identity to reduce the computational time**
29 **for pair-wise genome set comparisons by orders of magnitude. We demonstrate its use in a**
30 **study where we separately assembled each metagenome from time series datasets. Groups**
31 **of essentially identical genomes were identified with dRep, and the best genome from each**
32 **set was selected. This resulted in recovery of significantly more and higher-quality genomes**
33 **compared to the set recovered using the typical co-assembly method. Documentation is**
34 **available at <http://drep.readthedocs.io/en/master/> and source code is available at**
35 **<https://github.com/MrOlm/drep> .**

36
37 Genome-resolved metagenomics involves the recovery of genomes directly from environmental
38 shotgun DNA sequence datasets (Tyson *et al.*, 2004). Metagenomic analysis of related samples
39 from the same ecosystem is often employed to investigate compositional stability and spatial or
40 temporal variation. The approach can also reveal microbial co-occurrence patterns and identify
41 factors or processes that control organism abundances. Analysis of sample series data is also
42 important technically, as different abundance patterns across the sample series for different
43 organisms provide valuable constraints for binning of assembled fragments into genomes
44 (Sharon *et al.*, 2013). In this process, reads from individual samples are mapped back to a
45 collection of genomes that is often obtained by combining the reads from all samples and

46 assembling them together (co-assembly) (Vineis *et al.*, 2016; Bendall *et al.*, 2016; Lee *et al.*,
47 2016). However, co-assembly dramatically increases the dataset size and complexity, especially
48 when multiple different strains of the same species are present across the sample series, and can
49 result in fragmented assemblies (Sczyrba *et al.*, 2017).

50 An alternative process is to map reads to a collection of genomes independently
51 assembled from the individual samples (**Supp. Figure S1**). Independent assembly should
52 generate more and higher quality genomes than the co-assembly based approach because the
53 complexity of individual samples is lower than that of the combination of samples. The
54 challenge that arises from independent assembly is that de-replication of the resulting genome set
55 is required (Olm *et al.*, 2017; Raveh-Sadka *et al.*, 2015; Probst *et al.*, 2016). In addition to
56 identifying genomes that are the “same,” another important aspect of de-replication is identifying
57 the highest quality genome in each replicate set. This is important to maximize the accuracy of
58 metabolic predictions and other downstream analyses.

59 De-replication requires pair-wise genome comparisons, and thus the time required scales
60 exponentially with an increasing number of genomes. Hundreds of thousands of CPU hours may
61 be needed to de-replicate larger genome sets with robust algorithms (gANI) (Varghese *et al.*,
62 2015). Mash, a recently developed algorithm that utilizes MinHash distance to estimate
63 similarity between genomes, is an attractive alternative due to its incredibly fast speed (Ondov *et*
64 *al.*, 2016). However, we found that the accuracy of MASH decreases as the completeness of the
65 compared genome bins decreases (**Figure 1A**). Thus, it cannot be used to de-replicate collections
66 of partial genomes.

67 Here we present dRep, a program that utilizes both gANI and Mash in a bi-phasic
68 approach to dramatically reduce the computational time required for genome de-replication,

69 while ensuring high accuracy. The genome set is first divided into primary clusters using Mash,
70 and then each primary cluster is compared in a pair-wise manner using gANI, forming secondary
71 clusters of near-identical genomes that can be de-replicated. Using published information about
72 time required for genome comparisons, we performed an *in silico* simulation of de-replication
73 time for Mash, gANI, and dRep (**Figure 1B**). The results indicate that dRep affords a multiple
74 orders of magnitude increase in computation efficiency compared to naïve gANI. To verify this
75 prediction, we ran dRep on 1,125 genomes assembled from 195 fecal metagenomes collected
76 from 21 premature infants during the first months of life (Raveh-Sadka *et al.*, 2016), and found
77 the actual run time to be very close to that predicted by our simulation: 92 versus 93 CPU hours,
78 respectively. This is compared to 2,784 hours required for naïve gANI. As the run-time of dRep
79 depends on the diversity of the genome set, and pre-term infant gut communities are especially
80 non-diverse (Gibson *et al.*, 2016), even greater increases in computational efficiency are
81 expected from most other environments than predicted by our simulation.

82 We analyzed the same 195 metagenomes to test the prediction that, for each infant,
83 individual assembly and de-replication would generate more and higher quality genomes than
84 co-assembly of the read datasets. We de-replicated genomes obtained from assemblies generated
85 from each sample individually as well as from a co-assembly (to recover low-abundance
86 genomes), and recovered a genome set with 34% more bins ($\geq 50\%$ complete, $\leq 25\%$
87 contaminated) than were obtained from co-assembly alone (**Figure 1C**). In cases where genomes
88 were recovered using both methods, genomes assembled from individual samples were
89 significantly less fragmented (N50; $p = 9.6e-13$) and more complete ($p = 2.3e-8$) (Wilcoxon
90 signed-rank test). We also tested the ability of dRep to track strains present in multiple infants.
91 We defined strains as the “same” if at least 50% of the genomes aligned with 99.9% average

92 nucleotide identity, and identified 10 strains present in at least three of the infants (**Figure 1D**).
93 Taken together, dRep enabled recovery of more and better genomes than co-assembly alone, and
94 served as an effective tool for strain tracking.

95 To explore the effect of strain heterogeneity on assembly and genome recovery, we
96 performed co-assemblies and individual dataset assemblies followed by dRep on a sample series
97 from a single infant. The infant dataset has known strain heterogeneity (Sharon *et al.*, 2013) and
98 has been used for benchmarking many new tools (Luo *et al.*, 2015; Eren *et al.*, 2015; Graham *et*
99 *al.*, 2016). In the case of *Staphylococcus hominis*, co-assembly generated a contaminated bin (i.e.,
100 many duplicate and triplicate single copy genes) (**Figure 2A**). In contrast, a near-complete,
101 uncontaminated genome was recovered from several individual time-points. Previous work on
102 the same dataset (Eren *et al.*, 2015) has shown manual bin curation of the co-assembled bin with
103 anvio can increase the *S. hominis* bin quality (73% complete; 6.6% redundant), but still not to
104 the level of the un-curated bin from the individual assembly (98% complete; 0% redundant).

105 For *Staphylococcus aureus*, both co-assembly and individual assembly resulted in near-
106 complete and uncontaminated genomes. However, alignment of the scaffolds from both *S.*
107 *aureus* assemblies to a complete *S. aureus* reference genome showed that the genome from the
108 co-assembly was more fragmented than that from the single sample assembly. (**Figure 2BC**).
109 Fragmentation was also concentrated in areas of extensive population variation, as evident based
110 on SNPs between metagenome reads and the genome sequence (**Figure 2D**). Genome
111 fragmentation in sites of elevated strain variation could systematically decrease measures of
112 within-population heterogeneity that rely on mapping reads to reconstructed genome sequences
113 (Bendall *et al.*, 2016; Quince *et al.*, 2016).

114 It is both logical, based on the well known effects of sample complexity, and clear from
115 the analysis of human microbiome samples presented here, that assembly of data from individual
116 samples followed by de-replication has major advantages over co-assembly (especially as co-
117 assembled genomes can be included in the de-replication process). Because it relies on Mash,
118 dRep can only be used if the genomes in the comparison set are >50% complete. dRep combines
119 checkM for completeness-based genome filtering (Parks *et al.*, 2015), Mash (Ondov *et al.*, 2016)
120 for fast grouping of similar genomes, gANI (Varghese *et al.*, 2015) or ANIm (Richter and
121 Rosselló-Móra, 2009) for accurate genomic comparisons, and Scipy (Jones *et al.*, 2001) for
122 hierarchical clustering. In the case of viruses and plasmids, dRep requires use of an independent
123 method to estimate genome completeness because there are no established metrics for this in
124 checkM.

125 dRep is easy to use, highly customizable, and parallelizable. The algorithm has the
126 sensitivity of gANI but scales more efficiently with the size of the genome set. If desired, dRep
127 can perform rapid pairwise genomic comparisons (without de-replication) to enable visualization
128 of the degree of similarity among groups of similar genomes (**Supp. Figure S2**). This may be
129 particularly valuable for classifying strains as indistinguishable vs. different, a task that is
130 important, for example, for strain tracking. For the full source-code, installation instructions, and
131 manual, see <https://github.com/MrOlm/drep>.

132

133 **Conflict of Interest**

134 The authors declare no conflict of interest.

135

136 **Acknowledgements**

137 Funding was provided by the Sloan Foundation (<http://www.sloan.org/>, grant number: G 2012-
138 10-05, PI: JFB) and the National Institutes of Health (NIH; award reference number 5R01-AI-
139 092531). This material is based upon work supported by the National Science Foundation
140 Graduate Research Fellowship under Grant No. DGE 1106400.

141

142 **Figure Legends**

143 **Figure 1. Assembly and de-replication with dRep results in more and higher-quality**
144 **genome bins as compared to co-assembly. (A)** A complete *Escherichia coli* genome was subset
145 ten times in increments of 10% (10%, 20%, 30%, etc.). Subsets were compared to each other in
146 a pair-wise manner (100 total comparisons) using three algorithms- ANIm, MASH, and gANI.
147 For each pair of subsets, the percent overlap between the two genomes is shown on the x-axis
148 (alignment coverage measured using mummer as implemented in ANIm), and the ANI reported
149 from each algorithm is shown on the y-axis. ANIm and gANI are accurate when genomes
150 overlap by $\geq 50\%$, but MASH is only accurate when genomes are essentially complete. **(B)** Using
151 previously reported algorithm run-times, we estimated the time required to de-replicate genome
152 sets of various sizes. gANI exhibits a sharp exponential climb, limiting its use on larger genome
153 sets; MASH and dRep do not. **(C)** De-replication of bins from individual assemblies and a co-
154 assembly (dRep assembly method) resulted in more bins ($\geq 50\%$ complete, $\leq 25\%$ contaminated)
155 than co-assembly alone. **(D)** Defining strains as having $>99.9\%$ ANI over 50% of their genomes,
156 10 strains are present in at least 3 of the 21 infants analyzed.

157

158 **Figure 2: Strain heterogeneity reduces genome assembly quality and causes fragmentation**
159 **in areas of extensive population-level variation. (A)** The 104 universal bacterial single copy

160 genes from checkM reported for a *Staphylococcus hominis* bin are shown on the x-axis, and their
161 copy number is reported on the y-axis. Co-assembly resulted in many duplicate and triplicate
162 single copy genes (compare top and bottom panels). **(B-D)** The *Staphylococcus aureus* bin
163 attained from co-assembly is more fragmented than that from an individual assembly. **(B)**
164 Scaffolds from both bins are aligned to a complete reference genome (2.7 Mbp). **(C)** Scaffolds
165 from the co-assembly are aligned to a single scaffold (shown in grey in **B**) from the individual
166 assembly. **(D)** Reads from all samples aligned to a gap in the alignment in **(C)**. Reads mapped to
167 the area where co-assembly failed to recover a genome sequence (highlighted in blue) show
168 signs of population-level strain variation **(D)**.

169

170 **Supplemental Information**

171 **Supplemental Figure S1: De-replication of individual assemblies vs. co-assembly.** The
172 methods of individual assembly and de-replication vs. co-assembly are shown for a sample series
173 of metagenomic shotgun datasets. If there are two closely related strains changing abundance
174 over the sample series, co-assembly will likely result in a fragmented genome. If there are some
175 samples in which one strain or the other is dominant, individual assemblies may result in high-
176 quality genomes. During the de-replication process the best genome from all individual
177 assemblies is chosen, leading to better genomes compared with co-assembly alone.

178

179 **Supplemental Figure S2: Visualization of dRep primary and secondary clusters.** The first
180 step of dRep is comparing genomes in a pair-wise manner using Mash. The resulting
181 dendrogram **(A)** is part of the output of the program by default, and allows the user to visualize
182 the global Mash relationship of all bins, as well as the cutoff for primary clusters (the black

183 dotted line at 90% Mash ANI). A dendrogram for each primary cluster is also produced (**B, C**).
184 This allows the user to visualize the secondary clustering relationship (using ANIm or gANI), as
185 well as the cutoff for secondary clusters of “same” genomes (black line at 99% ANI). The red
186 dotted line is the value of the lowest ANI resulting from a self-vs-self alignment of each genome
187 in the primary cluster and represents a “limit of detection” of sorts.

188
189 **Supplemental Data 1: dRep.** A clone of the dRep program (version 0.3.3) that is available at
190 <https://github.com/MrOlm/drep>.

191
192 **Supplemental Data 2: dRep manual.** A PDF reconstruction of the dRep program manual that is
193 available at drep.readthedocs.io/en/latest/.

194

195 References

196 Bendall ML, Stevens SL, Chan L-K, Malfatti S, Schwientek P, Tremblay J, *et al.* (2016).
197 Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME*
198 *J* **10**: 1589–1601.

199 Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, *et al.* (2015). Anvi'o: an
200 advanced analysis and visualization platform for 'omics data. *PeerJ* **3**: e1319.

201 Gibson MK, Wang B, Ahmadi S, Burnham C-AD, Tarr PI, Warner BB, *et al.* (2016).
202 Developmental dynamics of the preterm infant gut microbiota and antibiotic resistome. *Nat*
203 *Microbiol* 16024.

204 Graham E, Heidelberg J, Tully B. (2016). BinSanity: Unsupervised Clustering of Environmental
205 Microbial Assemblies Using Coverage and Affinity Propagation. *bioRxiv*. e-pub ahead of print,
206 doi: 10.1101/069567.

207 Jones E, Oliphant T, Peterson P. (2001). SciPy: Open source scientific tools for Python. *URL*
208 *Httpscipy Org*.

209 Lee ST, Kahn SA, Delmont TO, Hubert NA, Morrison HG, Antonopoulos DA, *et al.* (2016).
210 High-resolution tracking of microbial colonization in Fecal Microbiota Transplantation

- 211 experiments via metagenome-assembled genomes. *bioRxiv*. e-pub ahead of print, doi:
212 10.1101/090993.
- 213 Luo C, Knight R, Siljander H, Knip M, Xavier RJ, Gevers D. (2015). ConStrains identifies
214 microbial strains in metagenomic datasets. *Nat Biotechnol*. e-pub ahead of print, doi:
215 10.1038/nbt.3319.
- 216 Olm MR, Brown CT, Brooks B, Firek B, Baker R, Burstein D, *et al.* (2017). Identical bacterial
217 populations colonize premature infant gut, skin, and oral microbiomes and exhibit different in
218 situ growth rates. *Genome Res* gr-213256.
- 219 Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, *et al.* (2016). Mash:
220 fast genome and metagenome distance estimation using MinHash. *Genome Biol* **17**. e-pub ahead
221 of print, doi: 10.1186/s13059-016-0997-x.
- 222 Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. (2015). CheckM: assessing
223 the quality of microbial genomes recovered from isolates, single cells, and metagenomes.
224 *Genome Res* **25**: 1043–1055.
- 225 Probst AJ, Castelle CJ, Singh A, Brown CT, Anantharaman K, Sharon I, *et al.* (2016). Genomic
226 resolution of a cold subsurface aquifer community provides metabolic insights for novel
227 microbes adapted to high CO₂ concentrations. *Environ Microbiol* n/a-n/a.
- 228 Quince C, Connelly S, Raguideau S, Alneberg J, Shin SG, Collins G, *et al.* (2016). De novo
229 extraction of microbial strains from metagenomes reveals intra-species niche partitioning.
230 <http://biorxiv.org/lookup/doi/10.1101/073825> (Accessed September 12, 2016).
- 231 Raveh-Sadka T, Firek B, Sharon I, Baker R, Brown CT, Thomas BC, *et al.* (2016). Evidence for
232 persistent and shared bacterial strains against a background of largely unique gut colonization in
233 hospitalized premature infants. *ISME J*. e-pub ahead of print, doi: 10.1038/ismej.2016.83.
- 234 Raveh-Sadka T, Thomas BC, Singh A, Firek B, Brooks B, Castelle CJ, *et al.* (2015). Gut bacteria
235 are rarely shared by co-hospitalized premature infants, regardless of necrotizing enterocolitis
236 development Kolter R (ed). *eLife* **4**: e05477.
- 237 Richter M, Rosselló-Móra R. (2009). Shifting the genomic gold standard for the prokaryotic
238 species definition. *Proc Natl Acad Sci* **106**: 19126–19131.
- 239 Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Droege J, *et al.* (2017). Critical
240 Assessment of Metagenome Interpretation – a benchmark of computational metagenomics
241 software. *bioRxiv*. e-pub ahead of print, doi: 10.1101/099127.
- 242 Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman DA, Banfield JF. (2013). Time series
243 community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during
244 infant gut colonization. *Genome Res* **23**: 111–120.

- 245 Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, *et al.* (2004).
246 Community structure and metabolism through reconstruction of microbial genomes from the
247 environment. *Nature* **428**: 37–43.
- 248 Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides NC, *et al.*
249 (2015). Microbial species delineation using whole genome sequences. *Nucleic Acids Res* **43**:
250 6761–6771.
- 251 Vineis JH, Ringus DL, Morrison HG, Delmont TO, Dalal S, Raffals LH, *et al.* (2016). Patient-
252 Specific *Bacteroides* Genome Variants in Pouchitis. *mBio* **7**: e01713-16.
- 253



