

# *Assessment of the bimodality in the distribution of bacterial genome sizes*

Article

Accepted Version

Gweon, H. S., Bailey, M. J. and Read, D. S. (2017)  
Assessment of the bimodality in the distribution of bacterial  
genome sizes. *The ISME Journal*, 11 (3). pp. 821-824. ISSN  
1751-7370 doi: <https://doi.org/10.1038/ismej.2016.142>  
Available at <http://centaur.reading.ac.uk/75764/>

It is advisable to refer to the publisher's version if you intend to cite from the work. See [Guidance on citing](#).

Published version at: <https://doi.org/10.1038/ismej.2016.142>

To link to this article DOI: <http://dx.doi.org/10.1038/ismej.2016.142>

Publisher: Nature Publishing Group

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the [End User Agreement](#).

[www.reading.ac.uk/centaur](http://www.reading.ac.uk/centaur)

## **CentAUR**

Central Archive at the University of Reading

Reading's research outputs online

1 **Assessment of the bimodality in the distribution of bacterial genome sizes**

2

3 Hyun S. Gweon, Mark J. Bailey, Daniel S. Read

4

5 Centre for Ecology & Hydrology, Maclean Building, Benson Lane, Crowmarsh Gifford,

6 Wallingford, Oxfordshire, OX10 8BB, UK

7

8 Corresponding author full contact details: Dr Hyun Soon Gweon - Centre for Ecology and

9 Hydrology, Maclean Building, Benson Lane, Crowmarsh Gifford, Wallingford, Oxfordshire

10 OX10 8BB, UK. E-mail: [hyugwe@ceh.ac.uk](mailto:hyugwe@ceh.ac.uk).

11

12 **Abstract**

13 Bacterial genome sizes have previously been shown to exhibit a bimodal distribution. This  
14 phenomenon has prompted discussion regarding evolutionary forces driving genome size in  
15 bacteria and its ecological significance. We investigated the level of inherent redundancy in  
16 the public database and the effect it has on the shape of the apparent bimodal distribution. Our  
17 study reveals that there is a significant bias in the genome sequencing efforts towards a certain  
18 group of species, and that correcting the bias using species nomenclature and clustering of the  
19 16S rRNA gene, results in a unimodal rather than the previously published bimodal distribution.  
20 The true genome size distribution and its wider ecological implications will soon emerge as we  
21 are currently witnessing rapid growth in the number of sequenced genomes from diverse  
22 environmental niches across a range of habitats at an unprecedented rate.

23 **Short communication**

24 Significant progress has been made in understanding interactions between ecology and genome  
25 evolution in prokaryotes. A number of recent studies have focussed on the evolution of  
26 bacterial genome sizes (Kempes et al, 2016), indicating that the interaction between an  
27 organism and its ecological niche, for example resource availability and environmental stability,  
28 selects the genome size of the species (Konstantinidis & Tiedje, 2004; Bentkowski et al, 2015).  
29 The exact mechanisms driving the genome sizes are still not fully resolved (Sabath et al, 2013,  
30 Kempes et al, 2016). It has, however, been speculated that species living in invariant niches  
31 tend to have small genomes, as stability acts to reduce genome size due the metabolic burden  
32 of replicating DNA with no adaptive value (Giovannoni et al, 2005, 2014) such as in obligatory  
33 and intracellular pathogens or mutualists (Moya et al, 2009; Moran 2003; Klasson and  
34 Andersson 2004). Due to their metabolic diversity, species with large genomes are potentially  
35 able to tackle a wider range of environmental conditions (Schneiker et al, 2007) and tend to be  
36 more ecologically successful where resources are scarce but diverse and where there is little  
37 penalty for slow growth (Konstantinidis & Tiedje, 2004). The effect by which these two  
38 opposing evolutionary forces exert on the overall distribution of genome sizes was first  
39 observed by Koonin and Wolf in 2008, where it was reported that bacterial genome sizes show  
40 a bimodal distribution (Koonin and Wolf, 2008). The authors speculated that the observation  
41 of two distinct groups of bacteria, those with 'small' and those with 'large' genomes, directly  
42 reflects the balance between the opposing trends of genome expansion through gene  
43 duplication, horizontal gene transfer and replication, and genome contraction caused by  
44 genome streamlining and degradation (Koonin and Wolf, 2008). The observed bimodality in  
45 the database was the first empirical evidence to show the two forces at work in bacterial  
46 genomes, and the bimodalilty in the distribution has since attracted numerous citations in both  
47 peer-reviewed articles (Giovannoni et al, 2014; Moran et al, 2015; Mock et al, 2012; Lane et

48 a., 2011) and textbooks (Kirchman, 2012; Saitou, 2014; Seshasayee, 2015; Bergman, 2011;  
49 Koonin, 2011).

50

51 A substantial proportion of complete bacterial genomes in the public domain belong to human  
52 pathogens and very closely related genomes representing variations within the species  
53 (Tausova et al, 2014). As first reported by Graur and Zheng (2014), it has been suggested that  
54 this fact might introduce a bias to the bimodal distribution seen in the previous analyses. No  
55 formal treatment, however, has been carried out in the peer-reviewed literature to examine the  
56 extent of database bias and how it may affect bacterial genome size bimodality. The  
57 distribution of the bacterial genome size has broad and far-reaching implications in our  
58 understanding of prokaryotes and this in turn necessitates re-assessment of the distribution and  
59 the extent to which the bias distorts the apparent bimodality. Here, we present our finding that  
60 the bias in the database has profound influence in shaping the overall distribution of bacterial  
61 genome size.

62

63 Having obtained a total of 3923 complete bacterial genomes from Ensembl Bacteria database,  
64 which is the most comprehensive source of complete bacteria genomes (see Supplementary  
65 Information for detailed methods), the distribution of genome sizes was first evaluated and  
66 compared against the distribution from Koonin and Wolf (2007). Despite that almost six times  
67 more genomes have been archived since 2007, the current dataset exhibited a remarkably  
68 similar bimodal distribution with its distinctive bimodal peaks around 2Mbp and 5Mbp.  
69 Hartigan's dip test (Hartigan and Hartigan, 1985) was used to confirm that it features significant  
70 bimodality with a p-value of  $2.2e-16$  (Fig 1B), where p-values less than 0.05 indicate  
71 significant bimodality (or multimodality) and p-values greater than 0.10 indicate unimodality  
72 (Freeman and Dale, 2013).

73 The level of redundancy in the dataset was next assessed by counting the number of genomes  
74 which shared the same species classification. The entire dataset of 3923 genomes represented  
75 1,706 groups of species with a unique species classification based on names. As shown by Fig  
76 1C, there was a significant amount of bias in the genome sequencing efforts towards a certain  
77 group of species most of which belonged to well-characterised human pathogens. In fact,  
78 almost 25% of the entire genome dataset was composed of just 20 species (971 genomes). We  
79 also found that most of these highly redundant species belonged to the peaks in the bimodal  
80 distribution. Notably, the two most redundant species, namely *Salmonella enterica*,  
81 *Escherichia coli* belonged to peak  $\beta$  and *Helicobacter pylori*, *Staphylococcus aureus* belonged  
82 to peak  $\alpha$ .

83 Having observed the bias in the dataset, we assessed how much impact this has on the modality  
84 of the distribution by removing the redundant genomes from the dataset (Fig 2A). The resulting  
85 distribution exhibited much less pronounced peaks, and as confirmed by Hartigans' dip test,  
86 the distribution was non-significant for bimodality ( $p = 0.91$ ). The influence these redundant  
87 species has on the distribution became more apparent (Fig 2B) as we evaluated the modality of  
88 the distribution by progressively removing species from the dataset (from the most redundant  
89 to the least). There is a sharp incline towards unimodality as redundant species were gradually  
90 excluded (Fig 2B). In fact, the distribution became more or less unimodal after the top 60  
91 redundant species were removed from the dataset of 1,706 species.

92 One of the issues we faced with our approach was that a large number of genomes in the dataset  
93 had disorganised and inconsistent taxonomic classification. For instance, there were genomes  
94 using different naming convention such as ones with square brackets or strain identifier  
95 attached to their species name (e.g. "[Clostridium]-cellulolyticum", "*Francisella sp.*  
96 *TX077308*"). This meant that removing redundant genomes using a text based approach was  
97 only able to partially extirpate the bias. Also using this approach could not resolve the bias

98 arising from very closely related genomes representing variations within the species but with  
99 different species classification. A more suitable approach was to use a biomarker gene directly  
100 extracted from each genome to cluster dataset into units of redundant or very closely related  
101 species. For this purpose, we chose 16S rRNA gene as it had been demonstrated that 16S rRNA  
102 sequence on an individual strain with another exhibiting a similarity score of 97% or above  
103 represents the same species (Stackebrandt & Goebel, 1994; Tindall et al, 2010). The clustering  
104 resulted in 1081 groups of species or very closely related species, and as Fig 2C shows, the  
105 resulting distribution from the dataset indicated a unimodal distribution ( $p = 0.99$ , Hartigan's  
106 dip test).

107

108 Our results revealed that there is a significant amount of inherent redundancy in the public  
109 database with a strong bias towards certain groups of species, and they have strong influence  
110 in driving bacterial genome size distribution into bimodal. While it is plausible that bacterial  
111 genome size is heavily influenced by the specialist or generalist lifestyle, it is not immediately  
112 apparent whether or not this should lead to any particular distribution. To a great degree, it is  
113 still too early to make any conclusions as to whether the true distribution exhibits certain  
114 modality as the majority of genomes sequenced so far have only focussed on culturable species,  
115 in particular human pathogens and closely related species. Some interesting observations with  
116 a potential link to the nature of distribution have been emerging in recent years. For example,  
117 (i) the bimodality in flow cytometric analysis of bacterial DNA content has been implicated  
118 with the bimodal genome size distribution (Moran et al, 2015; Schattenhofer et al, 2011); (ii)  
119 there may be other factors such as physical cell space constraints playing a role in genome size  
120 selection (Kempes et al, 2016); and (iii) perhaps most intriguingly, numerous studies from  
121 metagenomics are indicating that species with small genomes are more common than  
122 previously thought (Giovannoni et al., 2014; Moran et al., 2015). With the rise of single-cell

123 genomics and improved bioinformatic assembly methods coupled with the continual reduction  
124 in genome sequencing, we are currently witnessing rapid growth in the number of sequenced  
125 genomes. Consequently, the true nature of the distribution together with its ecological  
126 implications will become more apparent as we gather more sequenced genomes from diverse  
127 niches across a wide range of habitats.



128 **Acknowledgements**

129 HSG acknowledges the support of NERC NBAF-W (NEC04916).

130

131 **Conflict of Interest**

132 The authors declare no conflict of interest.

133

134 **Supplementary Information**

135 Supplementary information is available at ISME Journal's website.

136

137 **References**

138 Bentkowski P, Van Oosterhout C & Mock T. (2015). A model of genome size evolution for  
139 prokaryotes in stable and fluctuating environments. *Genome Biology and Evolution* 7(8):  
140 2344–2351; e-pub ahead of print 4 August 2015, doi:10.1093/gbe/evv148.

141

142 Bergman NH. (2011). *Bacillus anthracis and Anthrax*. John Wiley & Sons.

143

144 Freeman JB, Dale R. (2013). Assessing bimodality to detect the presence of a dual cognitive  
145 process. *Behav Res Methods* 45:83–97.

146

147 Giovannoni SJ, Cameron Thrash J, Temperton B. (2014). Implications of streamlining theory  
148 for microbial ecology. *ISME J* 8:1–13.

149

150 Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, et al. (2005). Genome  
151 streamlining in a cosmopolitan oceanic bacterium. *Science* 309:1242–1245.

152

153 Graur D. (2014, April 27). “Take Another Good Look at the Data”: The Bimodal Distribution  
154 that Wasn’t, Retrieved from [http://judgestarling.tumblr.com/post/84095742522/take-another-](http://judgestarling.tumblr.com/post/84095742522/take-another-good-look-at-the-data-the-bimodal)  
155 [good-look-at-the-data-the-bimodal](http://judgestarling.tumblr.com/post/84095742522/take-another-good-look-at-the-data-the-bimodal)

156

157 Hartigan JA, Hartigan PM. (1985). The Dip Test of Unimodality. *Ann Stat* 13:70–84.

158

159 Kempes CP, Wang L, Amend JP, Doyle J, Hoehler T. (2016). Evolutionary tradeoffs in  
160 cellular composition across diverse bacteria. *ISME J*; e-pub ahead of print 5 April 2016, doi:  
161 10.1038/ismej.2016.21

162

163 Kirchman DL. (2012). *Processes in Microbial Ecology*. Oxford University Press: Oxford.

164

165 Klasson L, Andersson SGE. (2004). Evolution of minimal-gene-sets in host-dependent  
166 bacteria. *Trends Microbiol* 12:37–43.

167

168 Konstantinidis KT, Tiedje JM. (2004). Trends between gene content and genome size in  
169 prokaryotic species with larger genomes. *Proc Natl Acad Sci USA* 101:3160–3165.

170

171 Koonin EV, Wolf YI. (2008). Genomics of bacteria and archaea: The emerging dynamic  
172 view of the prokaryotic world. *Nucleic Acids Res* 36:6688–6719.

173

174 Koonin EV. (2011). *The Logic of Chance: The Nature and Origin of Biological Evolution*. FT  
175 Press: New Jersey.

176

177 Lane N. (2011). Energetics and genetics across the prokaryote-eukaryote divide. *Biol Direct*  
178 6:35.  
179  
180 Mock T, Kirkham A. (2012). What can we learn from genomics approaches in marine  
181 ecology? From sequences to eco-systems biology!. *Mar Ecol* 33:131–148.  
182  
183 Moran AG, Alonso-sa L, Nogueira E, Ducklow HW, Gonza N, Calvo-dí A, et al. (2015).  
184 More, smaller bacteria in response to ocean's warming? *Proc R Soc B*; e-pub ahead of print  
185 10 June 2015, doi: <http://dx.doi.org/10.1098/rspb.2015.0371>.  
186  
187 Moran NA. (2003). Tracing the evolution of gene loss in obligate bacterial symbionts. *Curr*  
188 *Opin Microbiol* 6:512–518.  
189  
190 Moya A, Gil R, Latorre A, Pereto J, Pilar Garcillan-Barcia M, De La Cruz F. (2009). Toward  
191 minimal bacterial cells: Evolution vs. design. *FEMS Microbiol Rev* 33:225–235.  
192  
193 Sabath N, Ferrada E, Barve A, Wagner A. (2013). Growth temperature and genome size in  
194 bacteria are negatively correlated, suggesting genomic streamlining during thermal  
195 adaptation. *Genome Biol Evol* 5:966–977.  
196  
197 Saitou N. (2014). *Introduction to Evolutionary Genomics*. Springer.  
198  
199 Schattenhofer M, Wulf J, Kostadinov I, Glöckner FO, Zubkov M V, Fuchs BM. (2011).  
200 Phylogenetic characterisation of picoplanktonic populations with high and low nucleic acid  
201 content in the North Atlantic Ocean. *Syst Appl Microbiol* 34: 470–5.  
202  
203 Schneiker S, Perlova O, Kaiser O, Gerth K, Alici A, Altmeyer MO et al. (2007). Complete  
204 genome sequence of the myxobacterium *Sorangium cellulosum*. *Nat Biotechnol* 25:1281–  
205 1289.  
206  
207 Seshasayee ASN. (2015). *Bacterial Genomics: Genome Organization and Gene Expression*  
208 *Tools*. Cambridge University Press.  
209  
210 Stackebrandt E, Goebel BM. (1994). Taxonomic Note: A Place for DNA-DNA Reassociation  
211 and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *Int J*  
212 *Syst Bacteriol* 44:846–849.  
213  
214 Tatusova T, Ciufu S, Federhen S, Fedorov B, McVeigh R, O'Neill K et al. (2014). Update on  
215 RefSeq microbial genomes resources. *Nucleic Acids Res* 43: D599–D605.  
216  
217 Tindall BJ, Rosselló-Móra R, Busse HJ, Ludwig W, Kämpfer P. (2010). Notes on the  
218 characterization of prokaryote strains for taxonomic purposes. *Int J Syst Evol Microbiol*  
219 60:249–266.

220 **Figure legends**

221

222 **Figure 1** (A) Distribution of genome sizes in bacteria and archaea: the curves were generated  
223 by Gaussian-kernel smoothing of the individual data points. The figure has a very similar  
224 pattern to the figure generated by Koonin and Wolf (2008). The distribution of archaea was  
225 included for comparison only. (B) Distribution of genome sizes in bacteria on a different scale:  
226 the distribution shows clear-cut bimodality. Hartigans' dip test for unimodality/multimodality  
227 with simulated p-value with 10000 Monte Carlo replicates:  $D = 0.02510$ ,  $p < 2.2e-16$  where  
228 values less than 0.05 indicate significant bi- or multimodality and values greater than 0.10  
229 indicate unimodality (Freeman and Dale, 2013). (C) Number of genomes from the top 20 most  
230 redundant species in the database with mean genome size and peak in which they belong. (Peak  
231  $\alpha$ : 1.5 Mbp - 3 Mbp, Peak  $\beta$ : 4 Mbp - 5.5 Mbp). The top 20 most redundant species belonged  
232 to 971 genomes representing almost 25% of the entire dataset. Most of them (18 species in  
233 total) formed part of the peaks ( $\alpha$  and  $\beta$ ) including the top 4 species, namely *Salmonella*  
234 *enterica*, *Escherichia coli*, *Helicobacter pylori* and *Staphylococcus aureus*.

235

236 **Figure 2** (A) Distribution of genome sizes in bacteria after removing redundant genomes. The  
237 grey area indicates 2217 redundant genomes (out of 3923 genomes in total). The distribution  
238 indicates unimodality (Hartigans' dip test:  $D = 0.0069289$ ,  $p = 0.908$ ). (B) Effect of removing  
239 500 most redundant species from the database on the modality of distribution measured by  
240 Hartigans' dip test. After removing around 60 most redundant species, the distribution becomes  
241 mostly unimodal. (C) Distribution of genome sizes in bacteria after removing redundant and  
242 very closely related genomes using 16S rRNA (2841 genomes). The distribution shows a clear-  
243 cut unimodal distribution (Hartigans' dip test:  $D = 0.0070418$ ,  $p = 0.996$ ).