Gweon, Hyun S.; Bailey, Mark J.; Read, Daniel S. 2017. **Assessment of the bimodality in the distribution of bacterial genome sizes**. *ISME Journal*, 11 (3). 821-824. 10.1038/ismej.2016.142

Assessment of the bimodality in the distribution of bacterial genome sizes

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

Centre for Ecology & Hydrology, Maclean Building, Benson Lane, Crowmarsh Gifford, Wallingford, Oxfordshire, OX10 8BB, UK

Corresponding author full contact details: Dr Hyun Soon Gweon - Centre for Ecology and Hydrology, Maclean Building, Benson Lane, Crowmarsh Gifford, Wallingford, Oxfordshire OX10 8BB, UK. E-mail: hyugwe@ceh.ac.uk.

Abstract

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

Short communication

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

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

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

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

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

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

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

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

Conflict of Interest
The authors declare no conflict of interest.

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

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

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

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

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

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

Graur D. (2014, April 27). "Take Another Good Look at the Data": The Bimodal Distribution that Wasn't, Retrieved from http://judgestarling.tumblr.com/post/84095742522/take-another-good-look-at-the-data-the-bimodal

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

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

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

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

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

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

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

Lane N. (2011). Energetics and genetics across the prokaryote-eukaryote divide. Biol Direct 6:35.

Mock T, Kirkham A. (2012). What can we learn from genomics approaches in marine ecology? From sequences to eco-systems biology!. Mar Ecol 33:131–148.

Moran AG, Alonso-sa L, Nogueira E, Ducklow HW, Gonza N, Calvo-dı A, et al. (2015). More, smaller bacteria in response to ocean's warming? Proc R Soc B; e-pub ahead of print 10 June 2015, doi: http://dx.doi.org/10.1098/rspb.2015.0371.

Moran NA. (2003). Tracing the evolution of gene loss in obligate bacterial symbionts. Curr Opin Microbiol 6:512–518.

Moya A, Gil R, Latorre A, Pereto J, Pilar Garcillan-Barcia M, De La Cruz F. (2009). Toward minimal bacterial cells: Evolution vs. design. FEMS Microbiol Rev 33:225–235.

Sabath N, Ferrada E, Barve A, Wagner A. (2013). Growth temperature and genome size in bacteria are negatively correlated, suggesting genomic streamlining during thermal adaptation. Genome Biol Evol 5:966–977.

Saitou N. (2014). Introduction to Evolutionary Genomics. Springer.

Schattenhofer M, Wulf J, Kostadinov I, Glöckner FO, Zubkov M V, Fuchs BM. (2011). Phylogenetic characterisation of picoplanktonic populations with high and low nucleic acid content in the North Atlantic Ocean. Syst Appl Microbiol 34: 470–5.

Schneiker S, Perlova O, Kaiser O, Gerth K, Alici A, Altmeyer MO et al. (2007). Complete genome sequence of the myxobacterium Sorangium cellulosum. Nat Biotechnol 25:1281–1289.

Seshasayee ASN. (2015). Bacterial Genomics: Genome Organization and Gene Expression Tools. Cambridge University Press.

Stackebrandt E, Goebel BM. (1994). Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. Int J Syst Bacteriol 44:846–849.

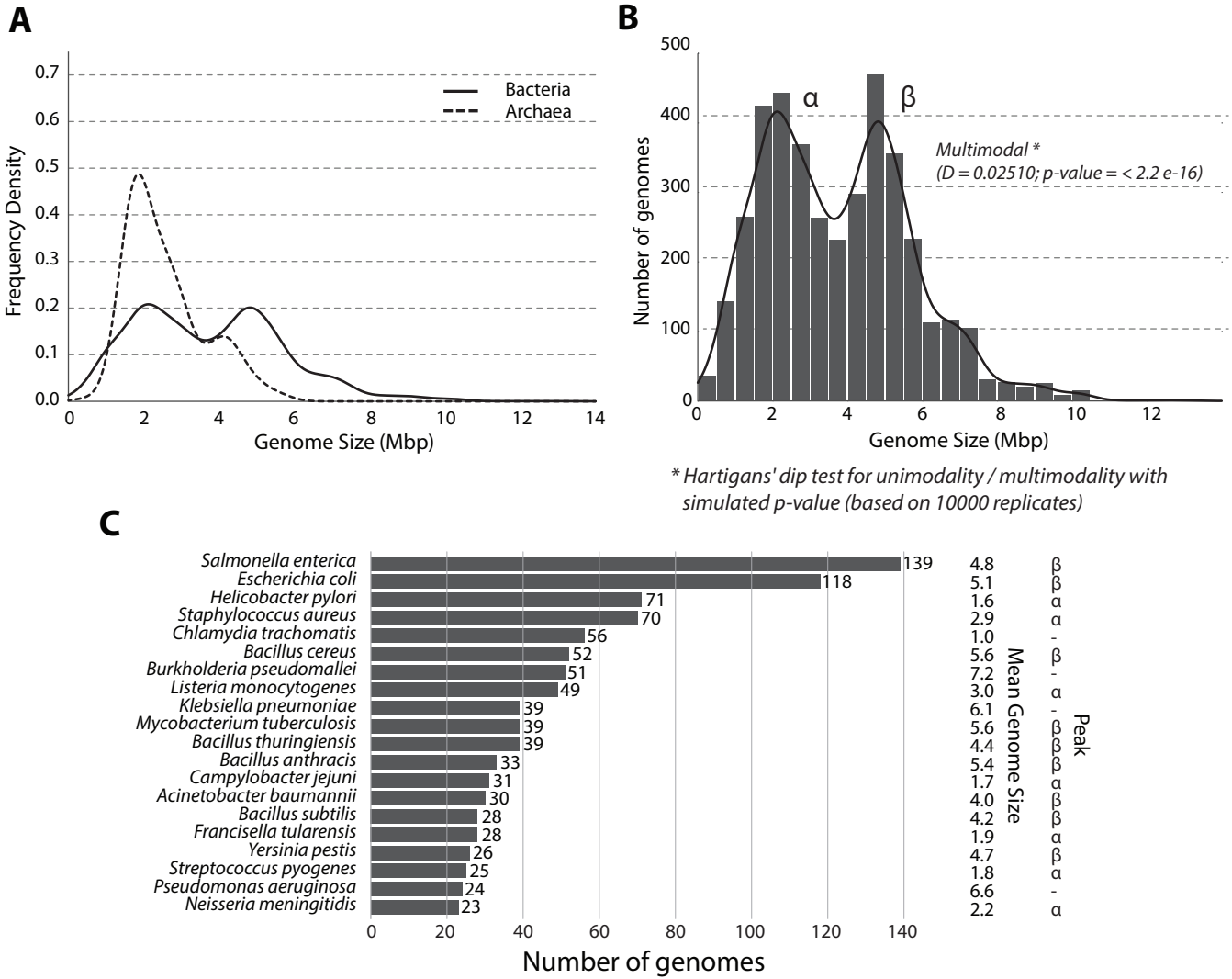
Tatusova T, Ciufo S, Federhen S, Fedorov B, McVeigh R, O'Neill K et al. (2014). Update on RefSeq microbial genomes resources. Nucleic Acids Res 43: D599–D605.

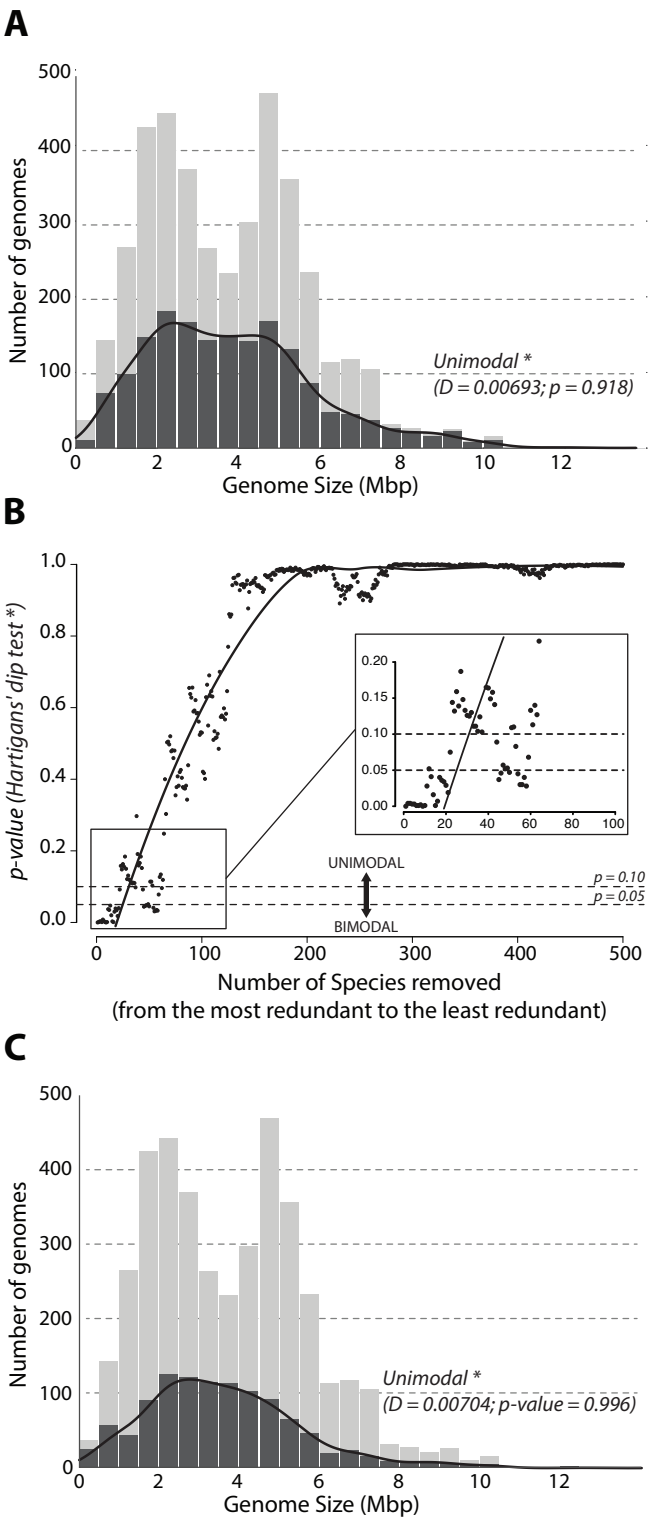
Tindall BJ, Rosselló-Móra R, Busse HJ, Ludwig W, Kämpfer P. (2010). Notes on the characterization of prokaryote strains for taxonomic purposes. Int J Syst Evol Microbiol 60:249–266.

Figure legends

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

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


A
Frequency Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
2
4
6
8
10
12
14
Genome Size (Mbp)
Bacteria
Archaea
B
Number of genomes
0
100
200
300
400
500
α
β
Multimodal *
(D = 0.02510; p-value = < 2.2 e-16)
Genome Size (Mbp)
* Hartigans' dip test for unimodality / multimodality with simulated p-value (based on 10000 replicates)
C
Salmonella enterica 139
Escherichia coli 118
Helicobacter pylori 71
Staphylococcus aureus 70
Chlamydia trachomatis 56
Bacillus cereus 52
Burkholderia pseudomallei 51
Listeria monocytogenes 49
Klebsiella pneumoniae 39
Mycobacterium tuberculosis 39
Bacillus thuringiensis 39
Bacillus anthracis 33
Campylobacter jejuni 31
Acinetobacter baumannii 30
Bacillus subtilis 28
Francisella tularensis 28
Yersinia pestis 26
Streptococcus pyogenes 25
Pseudomonas aeruginosa 24
Neisseria meningitidis 23
0
20
40
60
80
100
120
140
Number of genomes
Mean Genome Size
4.8
5.1
1.6
2.9
1.0
5.6
7.2
3.0
6.1
5.6
4.4
5.4
1.7
4.0
4.2
1.9
4.7
1.8
6.6
2.2
Peak
β
β
α
α
-
β
-
α
-
β
β
β
α
β
β
α
β
α
-
α



* *Hartigans' dip test for unimodality / multimodality with simulated p-value (based on 10000 replicates)*