OPEN

The ISME Journal (2017) 11, 1719–1721
www.nature.com/ismej

SHORT COMMUNICATION

# Prokaryote genome fluidity is dependent on effective population size

Nadia Andrea Andreani[1,2,4], Elze Hesse[3] and Michiel Vos[2]
[1]Department of Comparative Biomedicine and Food Science (BCA), University of Padova, Padua, Italy; [2]European Centre for Environment and Human Health, University of Exeter Medical School, University of Exeter, Penryn, UK and [3]Department of Biosciences, University of Exeter, Penryn, UK

**Many prokaryote species are known to have fluid genomes, with different strains varying markedly in accessory gene content through the combined action of gene loss, gene gain via lateral transfer, as well as gene duplication. However, the evolutionary forces determining genome fluidity are not yet well understood. We here for the first time systematically analyse the degree to which this distinctive genomic feature differs between bacterial species. We find that genome fluidity is positively correlated with synonymous nucleotide diversity of the core genome, a measure of effective population size $N_e$. No effects of genome size, phylogeny or homologous recombination rate on genome fluidity were found. Our findings are consistent with a scenario where accessory gene content turnover is for a large part dictated by neutral evolution.**
*The ISME Journal* (2017) **11,** 1719–1721; doi:10.1038/ismej.2017.36; published online 14 April 2017

## Results and discussion

Many bacterial species have been shown to exhibit extensive variation in gene repertoires, where a set of core genes shared by all strains are supplemented with a set of accessory genes that are only present in a subset of strains (Ochman *et al.*, 2000; Gogarten *et al.*, 2002; Tettelin *et al.*, 2005). Although accessory genome analyses are routinely performed in prokaryote genomics studies, whether certain genome characteristics are associated with particularly low or high genome fluidity has not been systematically tested. We here make use of the increasing availability of whole-genome sequences to, for the first time, perform a meta-analysis to (1) gauge the extent to which genome fluidity varies among different species and (2) test which genome characteristics best explain genome fluidity.

Methods to quantify pan-genome diversity are generally sensitive to the absence of rare accessory genes from genome samples. We therefore use the $\varphi$ measure of genome fluidity that has been shown to be robust to sample size (Kislyuk *et al.*, 2011) (Supplementary Methods). This measure of genomic fluidity is defined as the ratio of unique gene families to the sum of gene families in pairs of genomes averaged over randomly chosen genome pairs from within a group of sampled genomes. Because it is vital to reliably score gene presence/absence and most available genomes are not sequenced to completion, we first verified that good quality ($<150$ contigs) non-closed genomes resulted in fluidity estimates comparable to those based on closed genomes (linear regression, $R^2=0.70$, $P<0.001$; Supplementary Figure S1). Genome fluidity could be calculated for 90 free-living species for which five or more genomic data sets were available (3 archaea and 87 bacteria belonging to 15 major taxonomic groups, Supplementary Table S1). Only a single species was selected per genus to minimize phylogenetic bias. As estimates for individual species are dependent on genome selection and to a degree on the specifics of bioinformatics processing, they are not to be taken as absolutes and we will refrain from highlighting individual species, analysing broad patterns only.

Genome fluidity $\varphi$ was plotted against synonymous nucleotide diversity of the core genome ($\pi_{syn}$) on a natural log scale for all species (Figure 1), which showed a significant positive relationship (linear regression: $\ln(\varphi) = -1.39(0.12)+0.27(0.03)\times\ln(\pi)$; a: $t=-11.61$*** and b: $t=8.59$***, adjusted $R^2=0.45$). No genetically monomorphic species with high gene content variation or species with diverse core genomes but limited variation in accessory gene content were found. The same analysis was performed for the genera *Pseudomonas* and *Streptococcus* for which multiple species genome sets are available (Supplementary Tables S2 and S3).

Correspondence: M Vos, European Centre for Environment and Human Health, University of Exeter Medical School, University of Exeter, ESI Building, Penryn Campus, Exeter TR10 9EZ, UK.
E-mail: m.vos@exeter.ac.uk
[4]Current address: Istituto Zooprofilattico della Lombardia e dell'Emilia Romagna, reparto Substrati Cellulari e Immunologia Cellulare, via Bianchi 8, Brescia 25124, Italy.
Received 23 August 2016; revised 2 February 2017; accepted 9 February 2017; published online 14 April 2017

All estimates of $\varphi$ for these two genera were found to lie inside the 95% prediction interval of the relationship depicted in Figure 1 (Supplementary Figure S2), adding to the generality of our finding. A linear mixed-effects model was used with phylogenetic grouping included (group-dependent random intercepts) to test for the effect of genome size in addition to $\pi_{syn}$ (fixed effects) (Table 1). This analysis was limited to the 77 species belonging to the broad Proteobacteria and Terrabacteria classifications. No effect of phylogeny or genome size (ranging from 0.9 to 10.2 Mb) on genome fluidity was found, but the positive relation with evolutionary divergence of the core genome remained highly significant (Table 1).

Interestingly, the intercept of the relationship of $\varphi$ with $\pi_{syn}$ is significantly different from zero (Table 1), indicating that accessory genomes diverge before single-nucleotide polymorphisms appear in the core genome. This finding supports the emerging view that changes in gene content occur at high rates relative to mutation in bacteria (Touchon *et al*., 2009; Nowell *et al*., 2014; Vos *et al*., 2015; Wielgoss *et al*., 2016). The uptake and loss of accessory genes is in part mediated via recombination of flanking homologous sequences (Polz *et al*., 2013). To test whether the flexibility of the accessory genome is dependent on the rate of homologous recombination in the core genome, we compared $\varphi$ estimates and $r/m$ estimates (the probability that a nucleotide is changed as the result of recombination relative to point mutation) for 26 species that also featured in a meta-analysis of homologous recombination rate (Vos and Didelot, 2009). No significant relationship was detected (linear regression: $\varphi = 0.13(0.01)+0.01(0.01) \times \ln(r/m)$, a: $t = 9.78^{***}$ and b: $t = 0.54^{NS}$, adjusted $R^2 = -0.03$; Supplementary Table S4), confirming results of a previous analysis (Narra and Ochman, 2006).

The $\varphi$ estimate only provides a general indication of genome fluidity as it ignores genome rearrangements or plasmids, and we cannot exclude the fact that elevated or decreased levels of genome fluidity are associated with some of the many phyla that could not be included in this analysis due to a lack of data. These caveats aside, the positive relationship of genome fluidity with synonymous diversity is highly significant. The synonymous nucleotide diversity equals two times the product of the mutation rate $\mu$ and effective population size $N_e$ for haploid species. As variation in prokaryote mutation rate is believed to be relatively small (Lynch, 2010), $\pi_{syn}$ can be taken as a proxy for $N_e$. Large effective population size is expected to result in generally higher levels of genetic diversity due to neutral evolution (Kimura, 1984). The result of our cross-species meta-analysis is therefore consistent with the expectation that large $N_e$ species exhibit greater accessory genome variation. A variety of studies have suggested that many gene content changes have only minor effects on fitness and are effectively neutral (Gogarten and Townsend, 2005; Baumdicker *et al*., 2012; Haegeman and Weitz, 2012; Knöppel *et al*., 2014), although it is clear that a proportion of gene gains and losses will be significantly deleterious or beneficial. To gain a full understanding of selection on the accessory genome, it will be vital to collect data on the distribution of fitness effects of gene content changes (Vos *et al*., 2015).

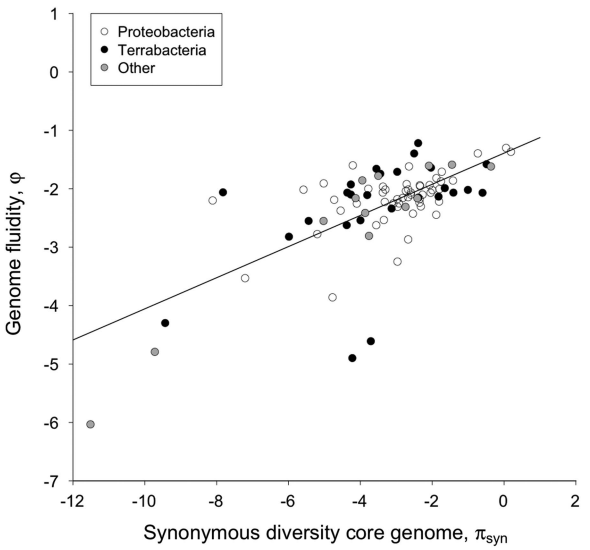


**Figure 1** The genome fluidity statistic $\varphi$ as a function of synonymous core genome nucleotide variation $\pi$ for 90 free-living prokaryote species on a ln-ln scale. White dots: Proteobacteria, black dots: Terrabacteria (Actinobacteria, Firmicutes and Cyanobacteria), grey dots: other taxa.

**Table 1** Results of the linear mixed-effects model testing the additive effects of genome size and synonymous core genome diversity ($\pi_{syn}$, ln-transformed) on accessory genome fluidity ($\varphi$, ln-transformed) with random intercepts fitted for each broad phylogenetic group (that is, Proteobacteria and Terrabacteria)

| | *Parameter estimate ± s.e.*[a] | F-*test* |
|---|---|---|
| Intercept | $-1.64 \pm 0.18^{***}$, $t = -8.87$ | |
| Genome size | $-0.02 \pm 0.04^{NS}$, $t = -0.42$ | $F_{1,4} = 0.18$, $P = 0.67$ |
| $\pi_{syn}$ | $0.17 \pm 0.04^{***}$, $t = 4.05$ | $F_{1,4} = 15.42$, $P < 0.001$ |
| Phylogenetic group | <1% of total variance | |

Abbreviation: NS, not significant.
[a]Note: significance of parameter estimates are based on Wald's *t*-test, $^{***}P < 0.001$.
The most parsimonious model was arrived at by sequentially deleting terms and comparing model fits using *F*-tests of likelihood ratios.

## Conflict of Interest

The authors declare no conflict of interest.

## Acknowledgements

This work was supported by NERC grant NE/L013177/1 to MV, and the Fondazione Ing. Aldo Gini and the PhD school of Veterinary Science of the University of Padova to NAA. We thank Adam Eyre-Walker, Haiwei Luo and Joshua Weitz for helpful discussion.

## References

Baumdicker F, Hess WR, Pfaffelhuber P. (2012). The infinitely many genes model for the distributed genome of bacteria. *Genome Biol Evol* **4**: 443–456.

Gogarten JP, Doolittle WF, Lawrence JG. (2002). Prokaryotic evolution in light of gene transfer. *Mol Biol Evol* **19**: 2226–2238.

Gogarten JP, Townsend JP. (2005). Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol* **3**: 679–687.

Haegeman B, Weitz JS. (2012). A neutral theory of genome evolution and the frequency distribution of genes. *BMC Genomics* **13**: 196.

Kimura M (1984). *The Neutral Theory of Molecular Evolution*. Cambridge University Press: Cambridge, UK.

Kislyuk AO, Haegeman B, Bergman NH, Weitz JS. (2011). Genomic fluidity: an integrative view of gene diversity within microbial populations. *BMC Genomics* **12**: 32.

Knöppel A, Lind PA, Lustig U, Näsvall J, Andersson DI (2014). Minor fitness costs in an experimental model of horizontal gene transfer in bacteria. *Mol Biol Evol* **31**: 1220–1227.

Lynch M. (2010). Evolution of the mutation rate. *Trends Genet* **26**: 345–352.

Narra HP, Ochman H. (2006). Of what use is sex to bacteria? *Curr Biol* **16**: 705–710.

Nowell RW, Green S, Laue BE, Sharp PM. (2014). The extent of genome flux and its role in the differentiation of bacterial lineages. *Genome Biol Evol* **6**: 1514–1529.

Ochman H, Lawrence JG, Groisman EA. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**: 299–304.

Polz MF, Alm EJ, Hanage WP. (2013). Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends Genet* **29**: 170–175.

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL *et al.* (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc Natl Acad Sci* **102**: 13950–13955.

Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P *et al.* (2009). Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet* **5**: e1000344.

Vos M, Didelot X. (2009). A comparison of homologous recombination rates in bacteria and archaea. *ISME J* **3**: 199–208.

Vos M, Hesselman MC, te Beek TA, van Passel MW, Eyre-Walker A. (2015). Rates of lateral gene transfer in prokaryotes: high but why? *Trends Microbiol* **23**: 598–605.

Wielgoss S, Didelot X, Chaudhuri RR, Liu X, Weedall GD, Velicer GJ *et al.* (2016). A barrier to homologous recombination between sympatric strains of the cooperative soil bacterium *Myxococcus xanthus*. *ISME J* **10**: 2468–2477.



Supplementary Information accompanies this paper on *The ISME Journal* website (http://www.nature.com/ismej)