# The population genetics of pangenomes

To the Editor — Two short papers recently attempted to explain prokaryotic pangenomes in light of population genetic theory, reaching surprisingly different conclusions. McInerney, McNally and O'Connell[1] define pangenomes as "the entire collection of gene families that are found in a given species" and conclude that "pangenomes are the result of adaptive, not neutral, evolution", whereas Andreani, Hesse and Vos[2] conclude the opposite, that "accessory gene turnover is for a large part dictated by neutral evolution". How can these views be reconciled? Going forward, how can we best use population genetic theory to understand pangenomes?

Both papers rely on neutral or nearly neutral population genetic theory to make their arguments. Using a dataset of 90 bacterial and archaeal species, Andreani et al. used a correlation analysis to demonstrate that species with large effective population sizes ($N_e$) also tend to have highly diverse pangenomes. The mechanism underlying this correlation is not known, but Andreani et al. suggest that larger effective population sizes harbour more genetic diversity, which is true (by definition) for nucleotide diversity[3], and is plausible (by extension, although not formally proven) for flexible genome diversity.

McInerney et al. focus on another prediction of population genetic theory: that large populations (with large $N_e$) will respond more efficiently to natural selection. They therefore argue that large populations have large pangenomes as gene acquisitions are on average adaptive, as suggested by recent modelling work[4]. Specifically, genes in the pangenome are thought to provide adaptation to one of many micro-niches inhabited by the species at large, as suggested by another model[5].

At face value, the two views can be reconciled by accepting that large populations contain high levels of neutral genetic diversity, and also respond more efficiently to natural selection. Thus, some of the correlation between $N_e$ and pangenome fluidity observed by Andreani et al. can be explained by neutral evolution, and some by niche adaptation.

While I applaud the use of population genetic theory to explain pangenomes, the theory relies on clearly defined populations — a concept that does not apply well to pangenomes as they are shaped by horizontal gene transfer (HGT), which can occur across population (or species) boundaries. McInerney et al. state that "HGT is a form of mutation and can be treated as such in models of pangenome evolution". This is a very strong assumption, which does not necessarily work. They go on to explain some basic population theory: "A truly neutral allele in a population of size $N$ will have an initial frequency of $1/N$. If the underlying acquisition rate of new alleles is $\mu$, then the rate of fixation of new alleles purely by drift is $N\mu \times 1/N = \mu$".

However, the rate of new alleles arriving by HGT will not be $N\mu$ (where $N$ is the size of the recipient population). Rather, $N$ will depend on the donor gene pool, which is potentially much larger than the recipient population. As a result, different genes will have different $N_e$, with more frequently recombined genes known to have higher $N_e$ (refs. [6,7]). Therefore, population genetic theory is not so easily ported to pangenomes, because the genes in pangenomes inhabit multiple populations. This is certainly true for the estimated ~12% of any given microbial genome that contains mobile genes with effectively infinite turnover rates[8], and is probably true for many other genes as well.

Should we give up on population genetic theory to explain pangenomes? Absolutely not. Both McInerney et al. and Andreani et al. have made some important headway, but I suggest that the next step will be to apply the theory on a gene-by-gene basis rather than a species-by-species basis, as different mobile genes will have different values of $N_e$. This will allow us to answer questions such as: do genes with high $N_e$ tend to occur in very many niches? And which genes are associated with which ecological niches? The fact that Andreani et al. do observe a strong correlation between $N_e$ and pangenome fluidity using species rather than genes as units suggests that most named species are indeed coherent units. If HGT is more frequent within than between species[9,10], most genes within a species would tend to have a similar $N_e$. However, distinguishing between neutral and adaptive interpretations of pangenomes will require a gene-focused and ecological approach. ❐

B. Jesse Shapiro
*Département de sciences biologiques, Université de Montréal, Montréal, Québec, Canada*
*e-mail: jesse.shapiro@umontreal.ca*

Published online: 24 November 2017
https://doi.org/10.1038/s41564-017-0066-6

## References

1. McInerney, J. O., McNally, A. & O'Connell, M. J. *Nat. Microbiol.* **2**, 17040 (2017).
2. Andreani, N. A., Hesse, E. & Vos, M. *ISME J.* **11**, 1719–1721 (2017).
3. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, 1984).
4. Sela, I., Wolf, Y. I. & Koonin, E. V. *Proc. Natl Acad. Sci. USA* **113**, 11399–11407 (2016).
5. Niehus, R., Mitri, S., Fletcher, A. G. & Foster, K. R. *Nat. Commun.* **6**, 8924 (2015).
6. Comeron, J. M., Kreitman, M. & Aguadé, M. *Genetics* **151**, 239–249 (1999).
7. Yahara, K. et al. *Mol. Biol. Evol.* **33**, 456–471 (2016).
8. Wolf, Y. I., Makarova, K. S., Lobkovsky, A. E. & Koonin, E. V. *Nat. Microbiol.* **2**, 16208 (2016).
9. Bobay, L.-M. & Ochman, H. *Genome Biol. Evol.* **9**, 491–501 (2017).
10. Shapiro, B. J., Leducq, J.-B. & Mallet, J. *PLoS Genet.* **12**, e1005860 (2016).

## Competing interests

The author declares no competing financial interests.