Counting the unseen

Sarah Christofides

University of Cardiff

DNA barcoding (amplicon sequencing) has revolutionised microbial ecology, allowing researchers to detect organisms invisible to traditional culture-based methods. Despite its enormous popularity, amplicon data is a statistician's nightmare: huge, multivariate, sparse, overdispersed, heteroscedastic, compositional and – worst of all – with observations per sample that differ by orders of magnitude due to technical artefacts. The latter is usually combatted with rarefaction: random subsampling of observations down to a constant level. Rarefaction was made controversial by McMurdie and Holmes' 2014 paper Waste not, want not: why rarefying amplicon data is inadmissable, but has nonetheless remained common practice due a lack of viable alternatives. Nine years on from Waste not, want not, I look at how the field has changed since then, the progress that has been made, and the outstanding challenges.