QUANTIFYING THE CONTRIBUTION OF HORIZONTAL TRANSFER TO GENOME EVOLUTION IN BACTERIA AND ARCHAEA Paul G Higgs. McMaster University, Hamilton, Ontario. higgsp@mcmaster.ca

Introduction: Comparison of the many complete genomes of prokaryotes now available makes it clear that horizontal gene transfer is important in allowing genes to spread between unrelated genomes. However, opinions differ on the quantitative extent of horizontal transfer. Some authors argue that it is so frequent that there can be no tree of life for prokaryotes, while others argue that there is a clear underlying tree visible in genome data.

It is sometimes argued that horizontal transfer was much more frequent in the early stages of evolution, and that separate lineages of organisms emerged only when the horizontal transfer rate became lower and the Darwinian threshold was crossed. Our own earlier work [1] supports this view by showing that there would be substantial benefits to accepting horizontally transferred genes in the early stages of life when vertical transmission of genes was very inaccurate, but that high levels of horizontal transfer became unfavourable in later stages when large genomes could be accurately passed down vertically.

When genomes are compared across groups of related species, it is usually found that the pan-genome (i.e. the set of genes present in the group as a whole) is much larger than the size of any one of the genomes. The distribution of numbers of genes as a function of there frequency in the group of genomes has a characteristic U shape, with a certain number of core genes present in almost all genomes, a large number of rare genes present in only a small number of genomes, and a relative small number of genes present in intermediate frequency [2].

Phylogenetics with Infinitely Many Genes: The presence and absence patterns of genes across species can be represented as simple 1/0 phylogenetic characters. Here we use data of this type to investigate the role that horizontal gene transfer plays in genome evolution. Presence/absence characters can be analyzed in a maximum likelihood framework using two-state model with rate parameters for gain and loss.

The standard way of doing this allows independent transitions from 0 to 1 or 1 to 0 in different parts of a tree, *i.e.* a gene can be gained or lost more than once. While independent losses of the same gene are likely to be frequent, independent gains of the same gene need to be considered more carefully. A gene gain could represent a horizontal transfer or the evolution of a new gene within the lineage being studied, *e.g.* by duplication and sequence divergence, or by scrambling of protein domains. If a gene is gained more than once, then at least one of these gains must be a horizontal transfer.

A key parameter is the ratio of insertion rate to deletion rate, a/v. Here we consider the limiting case known as the infinitely many genes (IMG) model [2]. This corresponds to a situation where the number of possible gene types in the pool available for horizontal transfer is infinite, but a/v tends to zero, in such a way that the overall genome size remains finite. In the IMG model a gene cannot be gained more than once. The standard recursion algorithm for calculating the likelihood of a phylogenetic pattern fails in this limit; therefore we develop a new method of calculation of the likelihood of presence and absence patterns with the IMG model. The IMG can be used as a null model in comparison to standard models with finite a/v that allow independent gains.

Applications: As a test case, we applied these methods to study the evolution of Cyanobacteria. It was found that a model with a small positive a/v ratio is a significant improvement over the IMG limit. This demonstrates that horizontal transfer has a significant effect that is visible in the presence/absence patterns. However, only about 15% of genes are best explained by evolutionary scenarios that involve more than one gain. The rest are either present from the root or inserted only once within the tree. Thus the majority of genes have a presence and absence pattern that is consistent with a treelike picture of evolution. Horizontal transfer does not erase the signal of this underlying evolutionary tree.

Currently we are using these methods to consider the relationships between the major groups of archaea and eukaryotes. It is commonly believed that eukaryotes contain genes derived from both bacteria and archaea. The bacterial partner is thought to be an alpha proteobacterium that was the ancestor of mitochondria. The nature of the archaeal partiner is less clear. We are using gene presence/absence data to determine if there is a particular subgroup of archaea that is most closely related to eukaryotes, or if the eukaryotes appear as a sister group to the whole of archaea, as in the traditional three-domain picture of life.

References:

[1] Vogan, A.A., Higgs, P.G. (2011) Biol. Direct. 6:1.
[2] Collins, R.E., Higgs, P.G. (2012) Mol. Biol. Evol. 29: 3413-3425.