**Methods**

Our method relies on a phylogenetic scan of the entire exomes (protein coding region of the genome) of the species under consideration using maximum likelihood methods as in Parker et al. (2013). Therefore, it is necessary to first assemble an appropriate sample of exomes. These exomes can be from cell lines or species but the selection of exomes for comparison is important for appropriately specifying the null and alternative hypotheses. For detecting convergence among species it is important to include exomes from species which are not expected to be convergent but are closely related to the convergent species under consideration. In particular, the species for which convergence is hypothesized should not have the most recent common ancestor as in figure 1. For cell lines it suffices to use cell lines from more than one individual with a common disease or, if the disease affects multiple tissues within an individual, it is sufficient to include both healthy and diseased cells from each affected tissue. The null hypothesis for the phylogenetic model then naturally takes the form of the accepted species or gene tree. The alternative hypothesis re-orders the species tree so that the convergent species become artificially sister as in figure 2. We will use species as an example for the remainder of this section but cell lines could be substituted without any changes to the analysis.

Once the exomes have been assembled and the null and alternative hypotheses are appropriately determined orthologous genes must be assembled. Parker et al. and others have used 1:1 orthologs in their respective methodologies in an effort to reduce the false discovery rate. However, this only indirectly reduces the false discovery rate while severely limiting the power to detect convergence since duplicated genes will be excluded from the analysis. Gene duplication events are a well-established mechanism for evolution since a duplicated gene is relieved from biological and functional constraints associated with the gene’s function and is therefore more easily re-purposed for new functionality. Instead of using 1:1 orthologs we conduct hierarchical gene clustering using a community detection algorithm (Kalinka and Tomancak 2011). We first determine the relatedness of each gene to every other gene in the set using an accelerated profile hidden markov model search algorithm (HMMER; Eddy 2011). We then weight the edge between each pair of genes with the inverse of the E-value for that pair as an approximation of the multidimensional distance between the two genes. For computational efficiency we then remove any edges with a weight less than 0.05. We then assemble gene clusters using the R package “LinkComm” (Kalinka and Tomancak 2011).

We then use a maximum likelihood based method to assess the likelihood for both the null and alternative hypotheses for each gene cluster by cycling through each possible combination of individual species. We use the program Genetic Algorithm for Rapid Likelihood Inference (GARLI; Zwickl 2006) to determine the likelihood of the null and alternative phylogenies and take the maximum of the likelihood ratio for all comparisons within a cluster. Large likelihood ratios indicate support for the alternative hypothesis and evidence for convergence.

Although the convergence detected in the likelihood ratios is real in the molecular sense it may have occurred by chance and may not be related to the phenotype of interest. Therefore, we test several alternative hypotheses that could explain the existence of random molecular convergence: 1) convergence due to genetic drift, 2) convergence due to poor model fit, and 3) convergence due to rapidly evolving sites. To assess the likelihood of the convergent changes occurring due solely to genetic drift we conduct a markov-chain mote carlo (MCMC) simulation using the Bayesian MCMC sampler Phylobayes (Lartillot and Philippe 2004). We then calculate the likelihood ratio of 10,000 simulated gene alignments (sampled from 1 million iterations of the MCMC chain) to determine the probability of a likelihood ratio equal to or larger than what we observe in our data. To assess whether the convergence signal is due to a poorly fitting model or a random selection of species we permute the alternative hypothesis tree to generate every possible alternative hypothesis in which the species of interest are not convergent. We then calculate the likelihood ratios for these random hypotheses. If any random alternative hypothesis exceeds the designated alternative hypothesis we infer chance convergence. Finally, to assess whether convergence could be due solely to rapidly evolving genes we conduct a branch-sites test using the software PAML (Yang 1997). Although fast evolving genes are more likely to produce chance convergence we nonetheless retain these genes since this property does not preclude convergence. However, these genes are appropriately flagged since they should be interpreted with additional care.

Eddy SR. 2011. Accelerated Profile HMM Searches. PLoS Comput Biol 7(10): e1002195. doi: 10.1371/journal.pcbi.1002195

Kalinka, A.T. and Tomancak, P. (2011). linkcomm: an R package for the generation, visualization, and analysis of link communities in networks of arbitrary size and type. Bioinformatics 27 (14), 2011-2012.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Computer Applications in BioSciences 13:555-556

Zwickl, D. J., 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Ph.D. dissertation, The University of Texas at Austin.

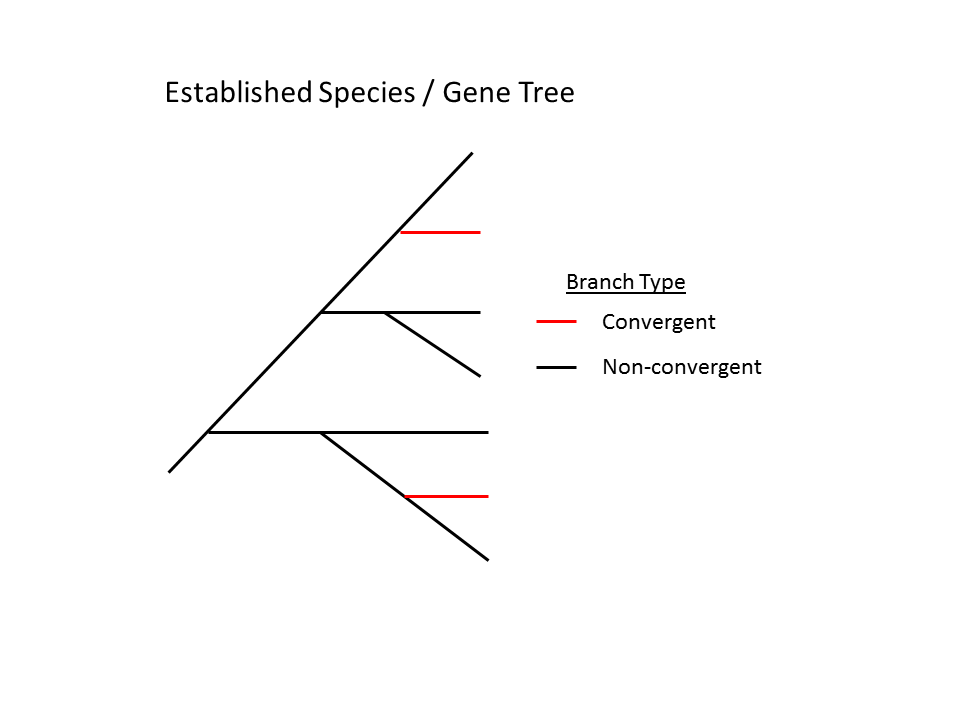


Figure 1. Appropriate species or gene tree with adequate separation of nodes which are expected to be convergent. The power of the method increases with increasing separation between the convergent nodes.

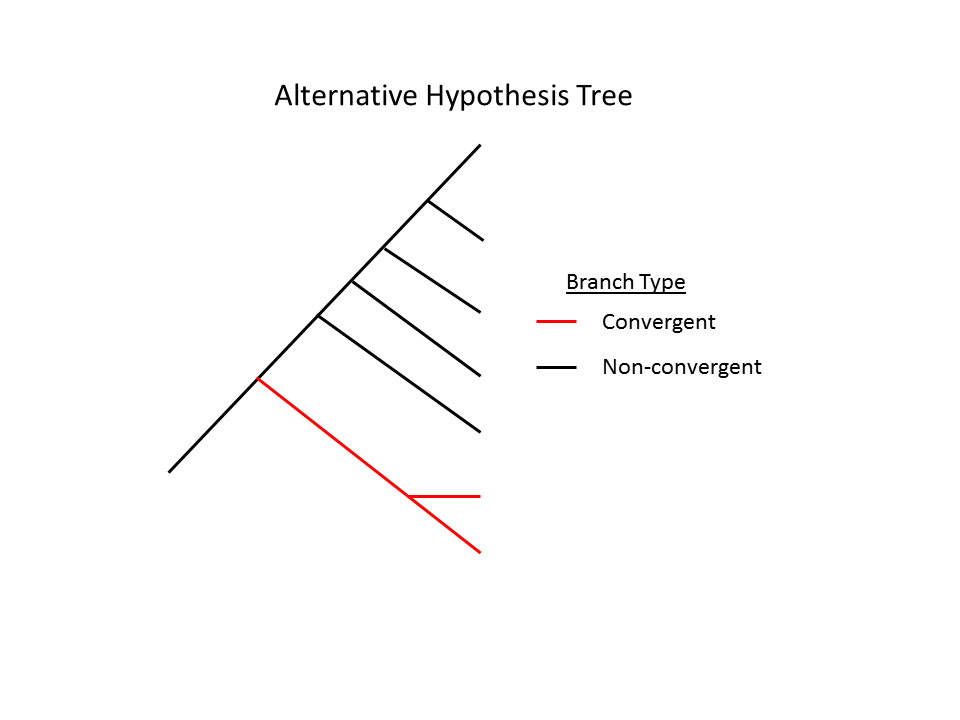


Figure 2: The alternative hypothesis should represent the expected convergence and simplify the species tree as much as possible.