Systematic overestimation of gene gain through false diagnosis of gene absence

Olga Zhaxybayeva, Camilla L. Nesbø and W. Ford Doolittle

Additional Data File 1. Materials and Methods

Genome data and group construction

We used 340 completed genomes available at NCBI FTP site on May 19, 2006. For each genome we extracted the sequence of one copy of 16S rRNA gene. We aligned the 16S rRNA genes using ClustalW 1.83 [23] and calculated p-distances using PAUP* 4.0b10 [24]. The genomes were split into 47 groups of >=96% 16S rRNA sequence identity between all members of the group using Markov Cluster (MCL) algorithm [25]. For each of 47 groups we calculated pairwise Average Nucleotide Identity (ANI) [26]. The genomes were clustered in 32 groups of from two to 11 genomes with at least 94% ANI between all members of the group. Clustering was performed using an MCL algorithm [25]) and 94% was chosen because Konstantindis and Tiedje [26] found it to correspond best to traditional species definitions. All further analyses were performed using these 32 groups (the list of organisms is available in Additional Data File 3).

Identification of gene families represented only in some genomes of a group

Within each group we identified patchily distributed gene families (i.e. gene families present only in some of the genomes of a group) using BLAST-based schemes. A variety of cutoffs for E-value and query match length were used (see Table S1 in Additional Data File 2). For synteny analysis we performed BLASTP and BLASTN searches with E-value cutoff of 10^{-4} and imposed an additional requirement, that there be conserved genes (A and B) present in all genomes of a group, flanking the patchily distributed genes X. That is, if gene X is present in two or more genomes, it has to be flanked by gene A on the 5' end in all genomes and by gene B on the 3' end in all genomes of a group, and if gene X appears as "absent" in a genome, genes A and B should be annotated as adjacent (5' to 3'). Nucleotide sequences for the genomic region between genes A and B, containing gene X or its remnants, if present, were extracted. In each set, one gene X was considered as a reference gene X. The regions with remnant of gene X were compared pairwise to the reference gene X using the PRSS program version 3.4 [27] with 1000 sequence shuffles and E-value cutoff of 10^{-3}.

Comparison of groups with three genomes each

For eight genome groups consisting of three genomes each, we performed a comparison of two selection criteria: BLASTP with E-value cutoff of 10^{-20} and BLASTN with E-value cutoff of 10^{-20} and 85% length requirement versus BLASTP with E-value cutoff of 10^{-20} and BLASTN with E-value cutoff of 10^{-20} and no match length requirement. The number of gene families where gene family type changes was calculated and classified into 12 scenarios (see Figure 2). Three-taxon phylogenetic trees for each group were calculated using ANI distance and mid-point rooting was used.