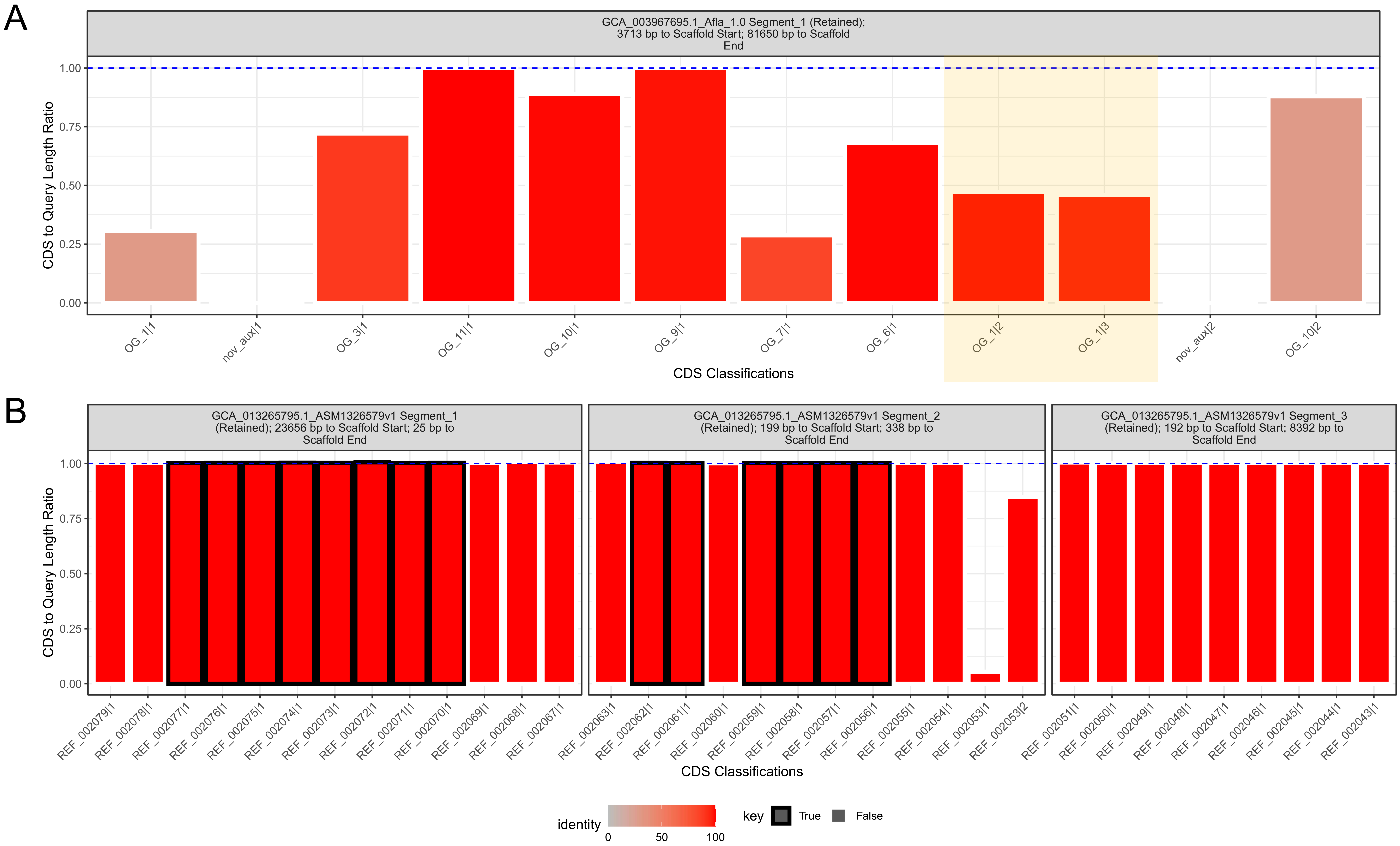

**Figure S1: Example illustrations for assessing quality of homologous gene-clusters produced by fai. A**) Gene calling or frame-shift differences between the query gene-cluster and coding-sequence predictions in the target genome have resulted in a discrepancy for OG_1 (highlighted) from the query being regarded as two separate coding-sequences in the target genome. **B**) Three candidate gene-cluster segments located near scaffold edges which match the query gene-cluster and meet the thresholds needed for detection as requested in fai in aggregate.

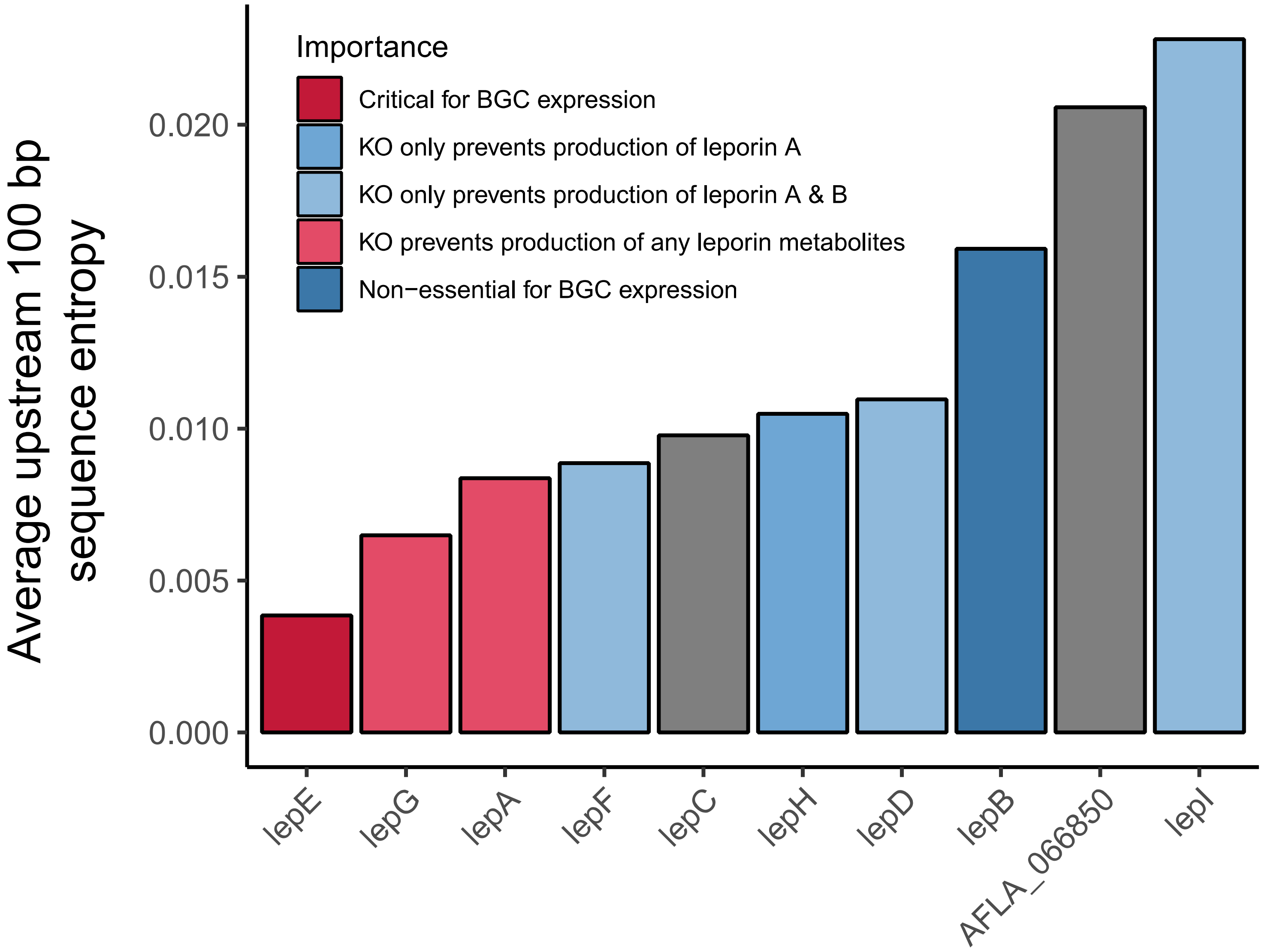

**Figure S2: Conservation in the upstream regions of coding sequences of genes in the leporin BGC.** The average entropy of the 100 bp upstream regions is shown for each of the genes from the leporin BGC. Coloring of the bars corresponds to effects on BGC expression (for *lepE* and *lepB*) or metabolite production (using a mutant with overexpression of *lepE*) when genes were knocked out as determined by Cary *et al.* 2015.

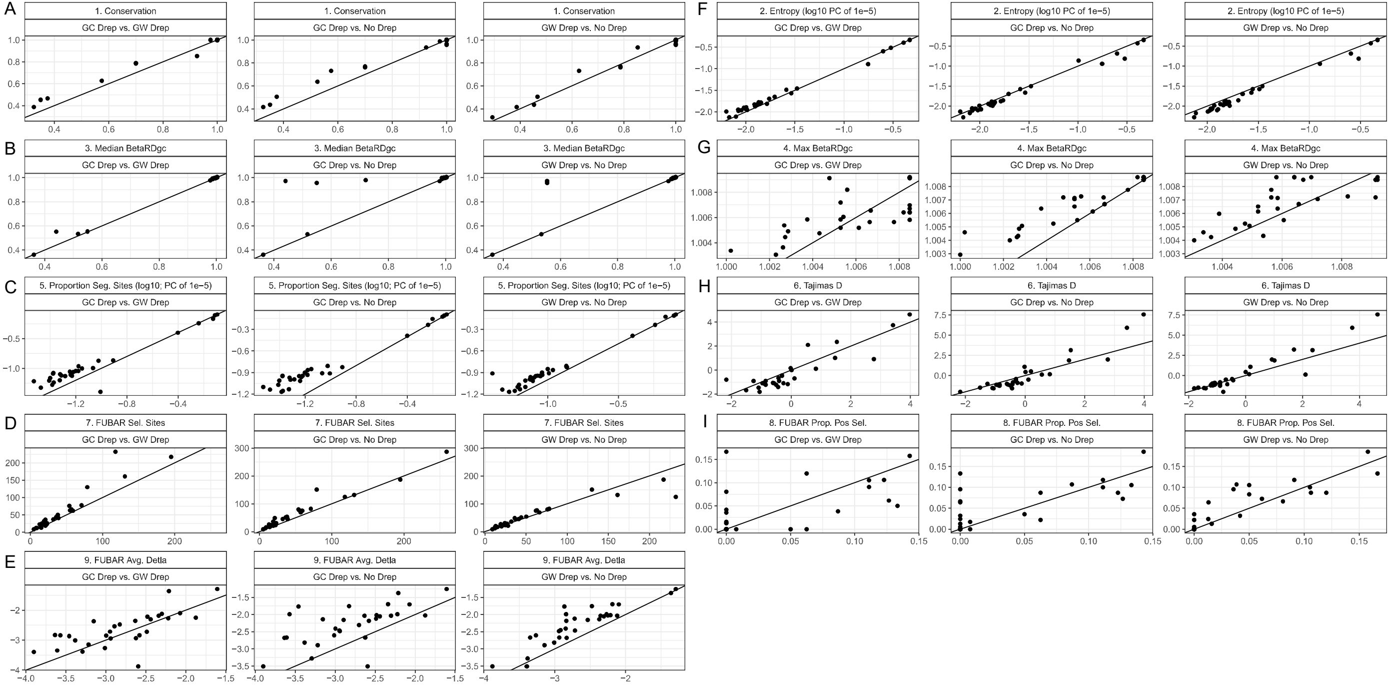

**Figure S3: Influence of dereplication on evolutionary statistics computed by zol**. The relationship in values for analogous ortholog groups which map to query proteins from *E. faecalis* V583 for different evolutionary statistics (**A-I**) when different sets of gene-clusters corresponding to different approaches in dereplication are shown. Only ortholog groups which lacked any paralogous proteins are shown and accounted for. A line is shown in each plot corresponding to a 1:1 ratio.

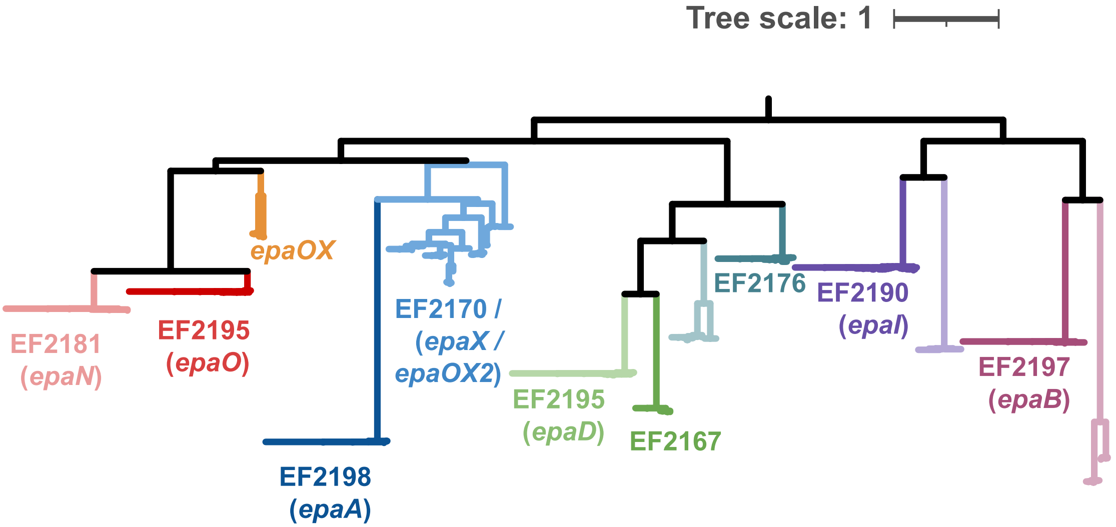

**Figure S4: The ortholog group with *epa*X features greater diversity relative to other glycosyl transferase related ortholog groups from the *epa* locus in *E. faecalis*.** An approximate maximum-likelihood phylogeny based on gap-filtered protein alignments of ortholog groups with domains featuring “glycosyl” and “transferase” as key words. Ortholog groupings (coloring of phylogeny branches) by zol were largely consistent with phylogenetic clades. Association of clades to genes from *E. faecalis* V583 based on sequence alignment are noted.