

Landscape of mobile genetic elements and their antibiotic resistance cargo in prokaryotic genomes

Supriya Khedkar¹, Georgy Smyshlyaev^{1,2}, Ivica Letunic³, Oleksandr M. Maistrenko¹, Luis Pedro Coelho⁴, Askarbek Orakov¹, Sofia K. Forslund^{1,5,6,7}, Falk Hildebrand^{1,8,9}, Mechthild Luetge^{1,10}, Thomas S. B. Schmidt¹, Orsolya Barabas^{1,2}, Peer Bork^{1,5,11,12*}

¹European Molecular Biology Laboratory, Structural and Computational Biology Unit, 69117 Heidelberg, Germany

²Department of Molecular Biology, University of Geneva, 1211 Geneva, Switzerland

³Biobyte solutions GmbH, Bothestr 142, 69117 Heidelberg, Germany

⁴Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China

⁵Max Delbrück Centre for Molecular Medicine, Berlin, Germany

⁶Experimental and Clinical Research Center, Charité-Universitätsmedizin and Max-Delbrück Center, Berlin, Germany

⁷Charité – Universitätsmedizin Berlin, Berlin, Germany

⁸Present address: Gut Microbes and Health, Quadram Institute Bioscience, Norwich, Norfolk, UK

⁹Present address: Digital Biology, Earlham Institute, Norwich, Norfolk, UK

¹⁰Present address: Institute of Immunobiology, Kantonsspital St. Gallen, 9007 St. Gallen, Switzerland

¹¹Department of Bioinformatics, Biocenter, University of Würzburg, Würzburg, Germany

¹²Yonsei Frontier Lab (YFL), Yonsei University, Seoul 03722, South Korea

*Correspondence: peer.bork@embl.org

Supplementary figure legends

Supplementary Figure 1. HMM building and calibration workflow

Workflow for building recombinase subfamily profile HMMs (see Methods)

Supplementary Figure 2. Evaluation and abundance of recombinase subfamily profile HMMs

A. Evaluation statistic: recall, specificity, f.p.r (false positive rate), accuracy and m.c.c (mathews correlation coefficient) of the performance of the built recombinase subfamily profile HMMs on the training dataset; B. Abundance of DDE recombinase based on recombinase family specific counts of the five major recombinase families; C. Recombinase subfamily counts for each of the five major families represented in panel B.

Supplementary Figure 3. Validation of recombinase annotations

A. Comparable counts of recombinase predictions by EggNOG and this study; B and C. Panels corresponding to five major recombinase families representing large repertoire of diverse domains significantly associated with recombinases; D. Majority of recombinase subfamilies showing the presence of the putative active site in the known and novel fraction either in a single recombinase domain of a protein or in one of the recombinase domains in a multi-domain protein; E. Major proportion of proteins across all recombinase subfamilies showing presence of putative active site residue.

Supplementary Figure 4. Evaluation of pangenome features

A. Box plots comparing the percentage of accessory genes within a species pangenome as a function of number of strains within a species; B. Average length (in base pairs) of MGE categories coloured according to the number of conspecific strains within a species (numbers in boxes indicate the total number of MGEs in each category).

Supplementary Figure 5. MGE assignment and quantification

A. Prokaryotic MGE disentanglement workflow representing MGE category determination strategy for recombinase subfamilies shared by different MGE types; B. Black bar showing the small contribution of putative IS element counts (transposable elements with a single transposase gene) to total transposable element numbers (same as Figure 2A with separate IS elements counts); C. Venn diagram showing the substantial overlap in transposable element predictions from ISEScan and this study, and expansion of phage predictions by this study compared to phage predictions by PHASTER; D. High average counts of transposable

elements per genome compared to other MGE categories. The whiskers span from the 10th to the 90th percentile; E. Positive correlation between number of MGEs in a genome and proteome size (Spearman's $\rho = 0.31$, $p < 2.2e-16$).

Supplementary Figure 6. MGE-mediated Horizontal Gene Transfer (HGT)

A. Marker gene based prokaryotic phylogenetic tree with coloured arcs representing MGE mediated HGT events across taxonomic classes (same as Figure 4A with arcs coloured according to MGE categories); B. High transposable elements mediated distant HGT events compared to other MGE categories; C. Enrichment of horizontally transferred IS_Tn with phage_like elements (HGT calls at 95% sequence identity) and conjugative elements (HGT calls at 100% sequence identity see methods) implicating mechanism of inter-cellular transfer of identical sequences IS_Tn via hitchhiking; D. Heatmaps illustrating the promiscuous horizontal transfer of transposable elements across habitats compared to other MGE categories. All MGE categories show high within habitat MGE dynamics (diagonals) compared to between habitat (same as Figure 4D but with absolute counts).

Supplementary Figure 7. Potential of MGEs to carry ARG across habitats

A. Distribution of normalised Antibiotic Resistance Genes (ARG) counts per 1000 genes for each MGE category compared to the (non-mobile) Genome. On average 29% of all MGEs carry at least one ARG. The fraction for each MGE category is indicated on the boxplot; B. Analysis of MGEs (per category) that carry ARGs in non-host associated habitats (soil, aquatic and food habitat) shows enrichment of ARG in transposable elements compared with other genomics regions, indicating transposable elements as major MGE associated ARG carriers; C. Distribution of ARGs per 1000 genes by MGE category and separated by habitat shows high fractions of ARGs in transposable elements and integrons across all habitats.

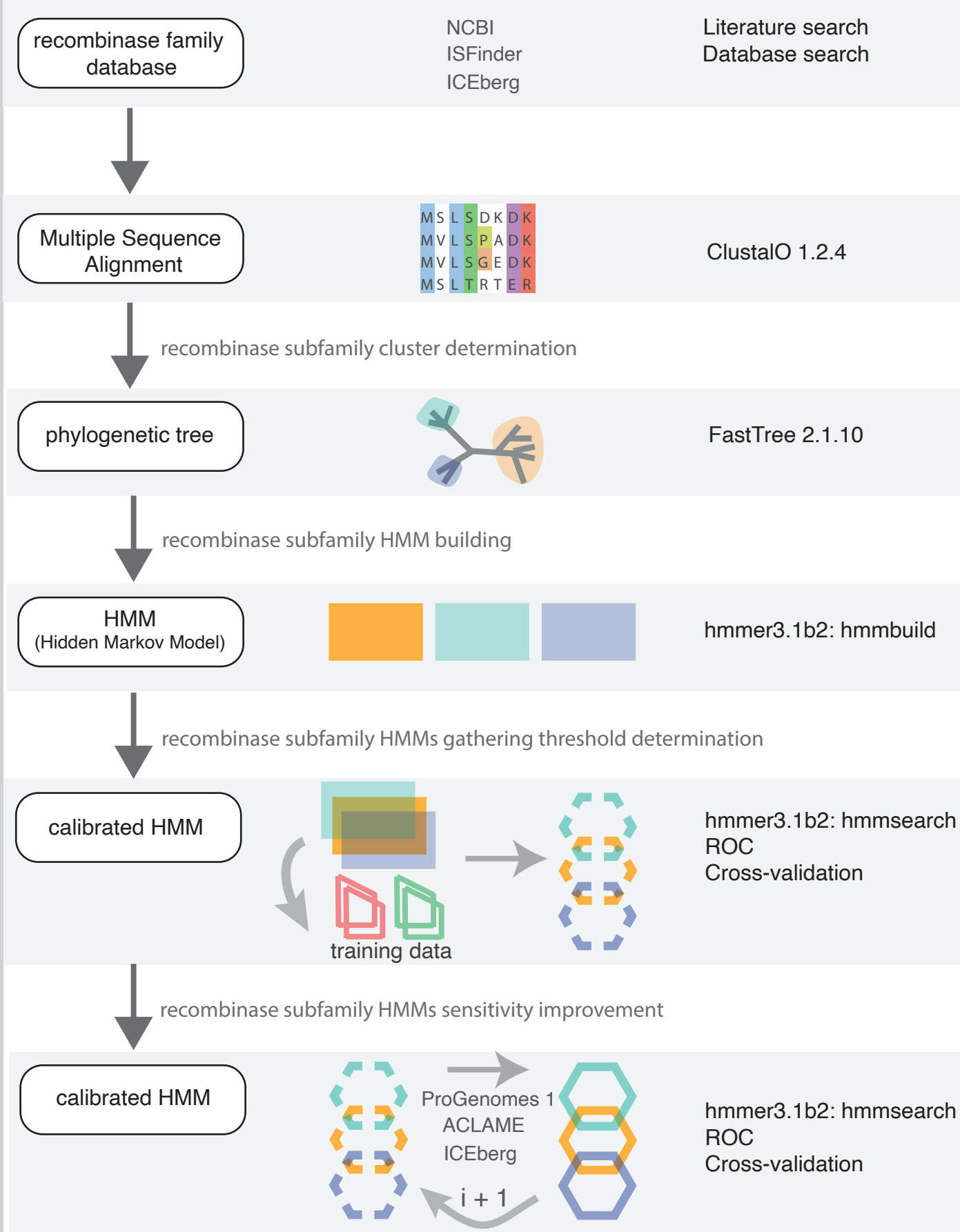
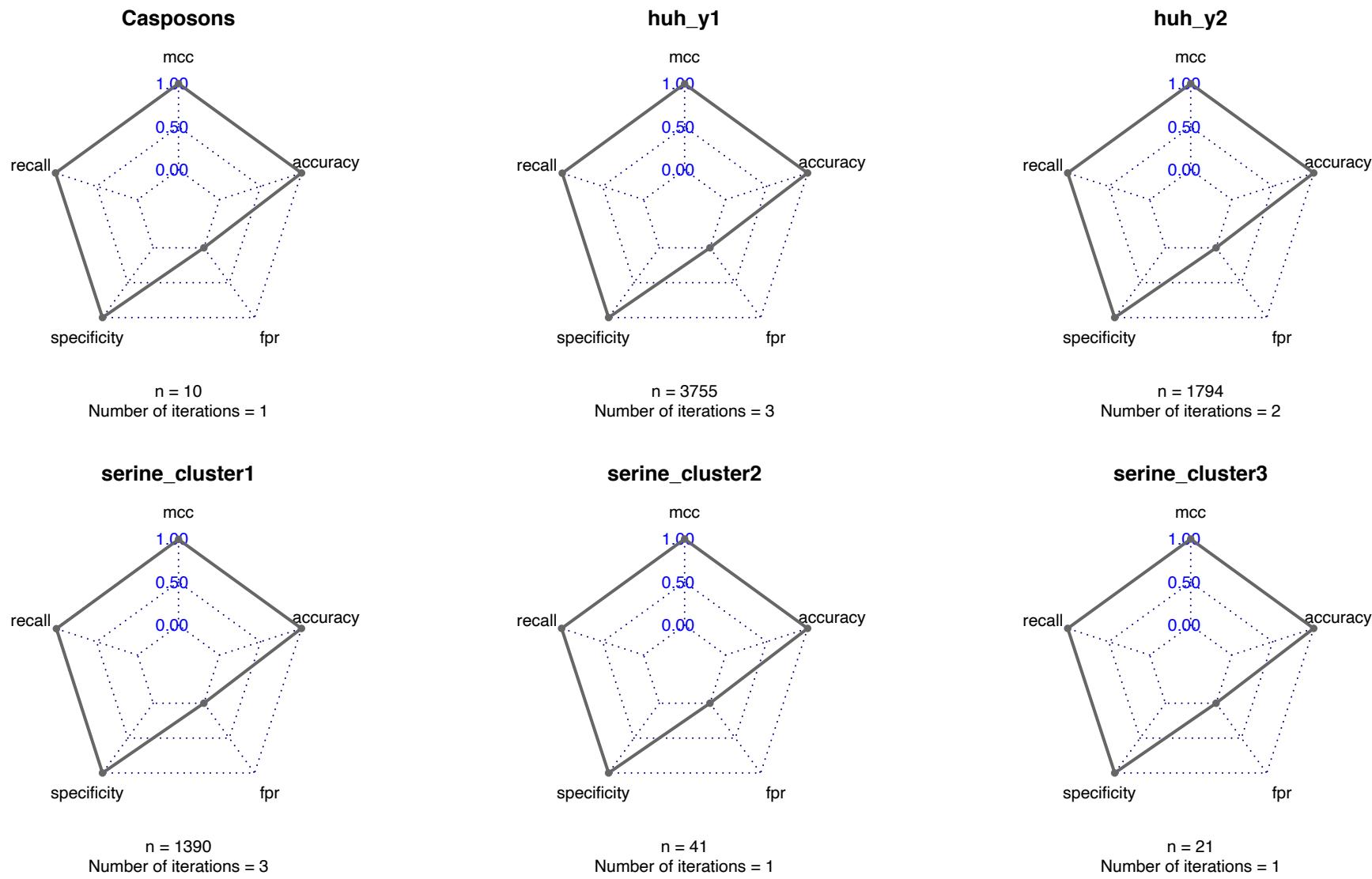
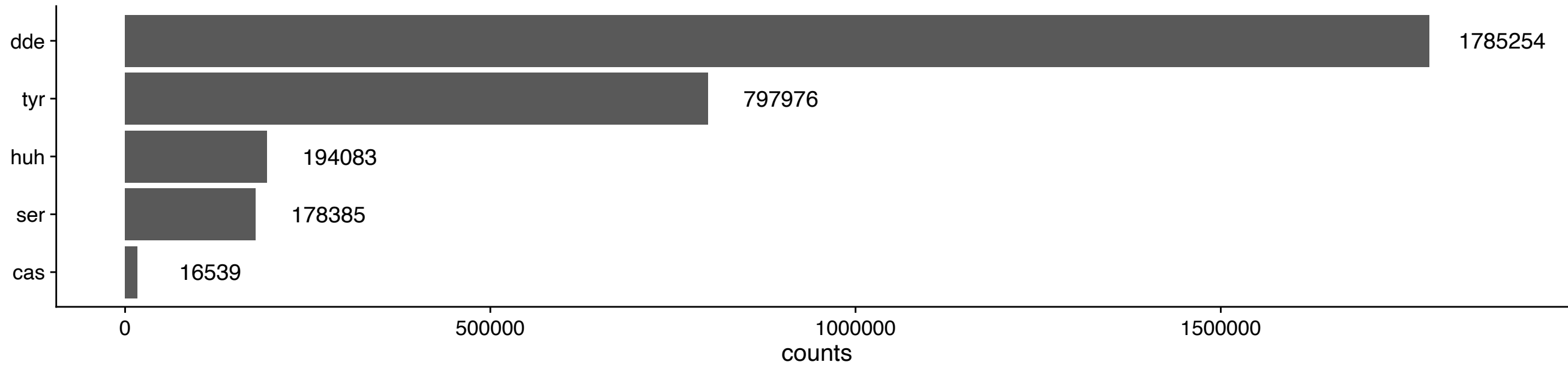


Figure S1

A



B



C

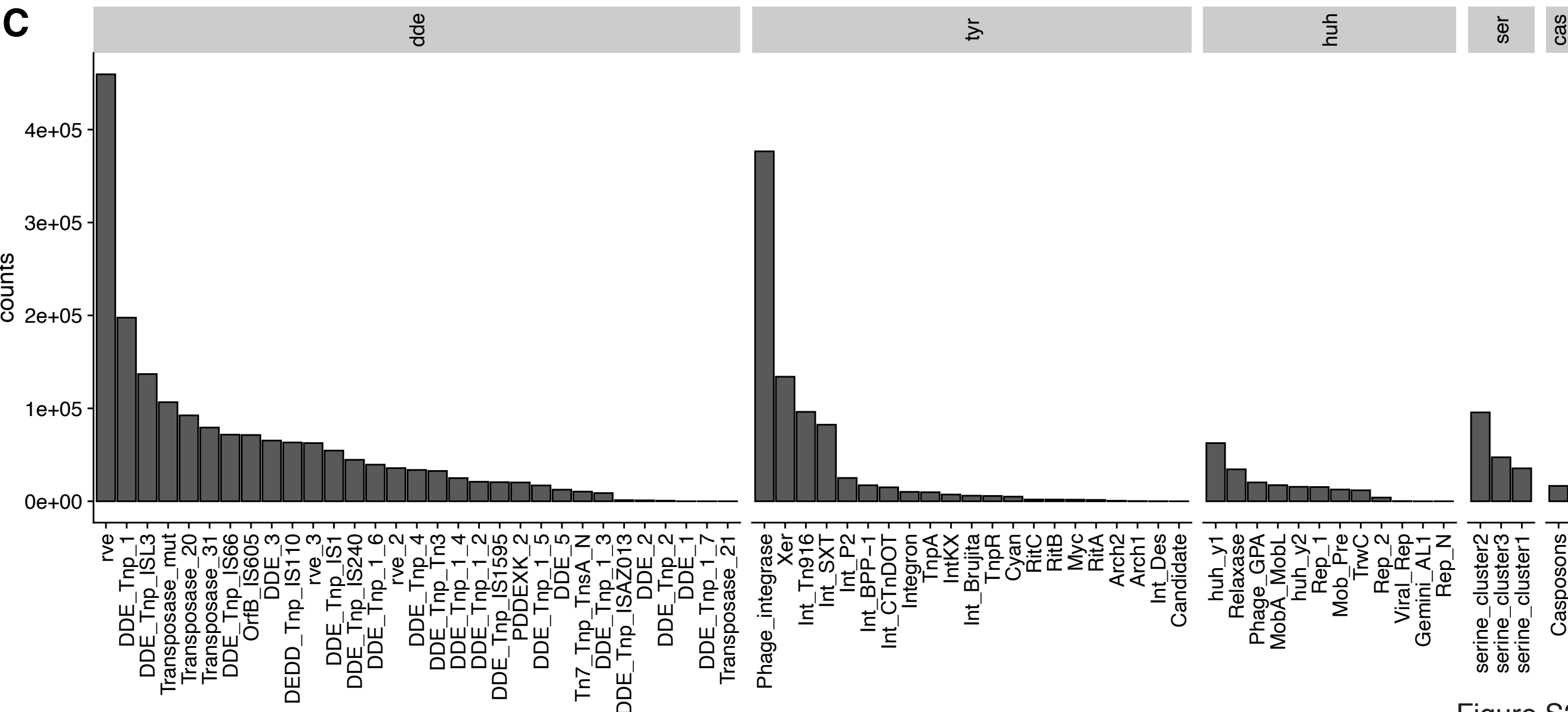


Figure S2

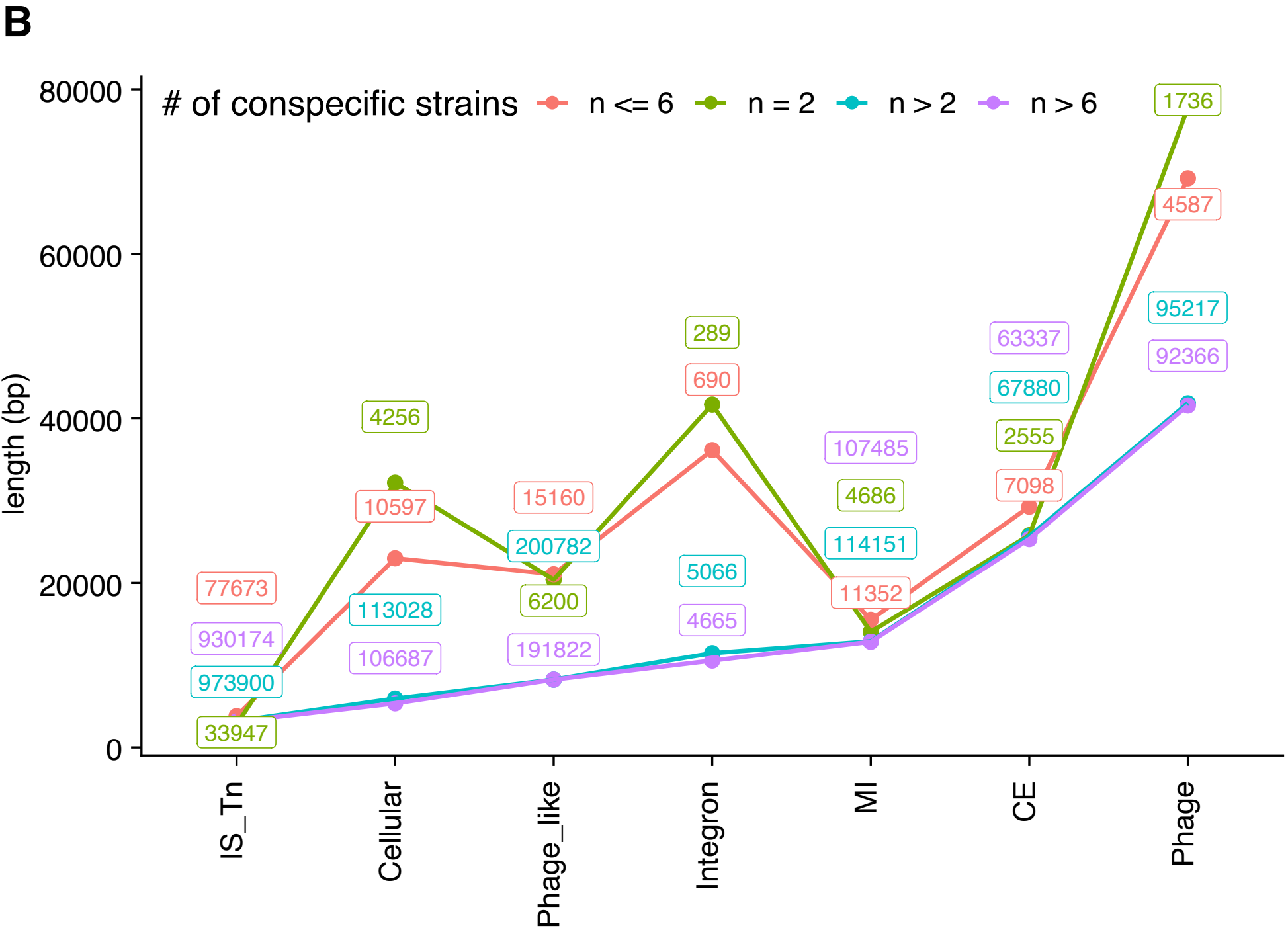
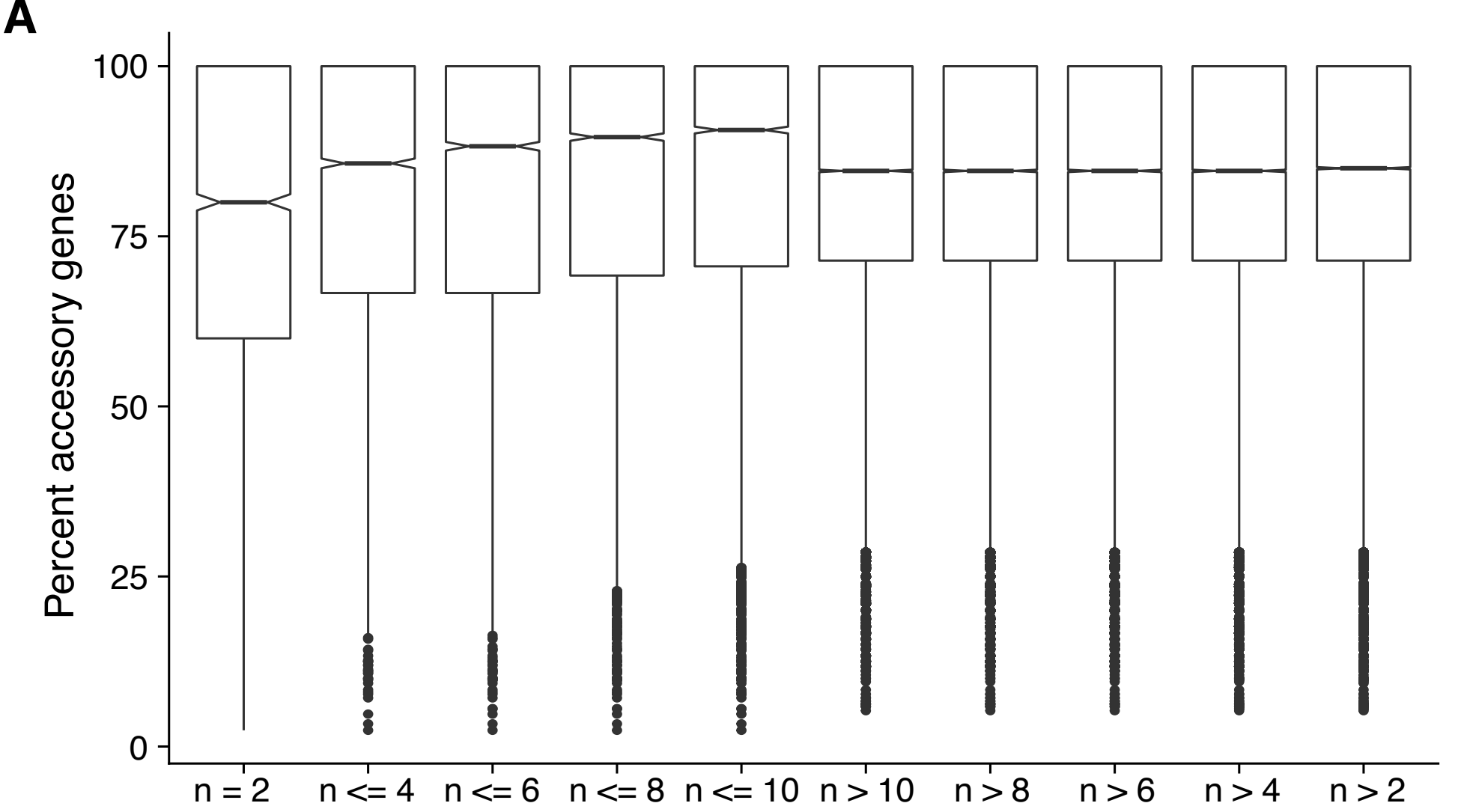


Figure S4

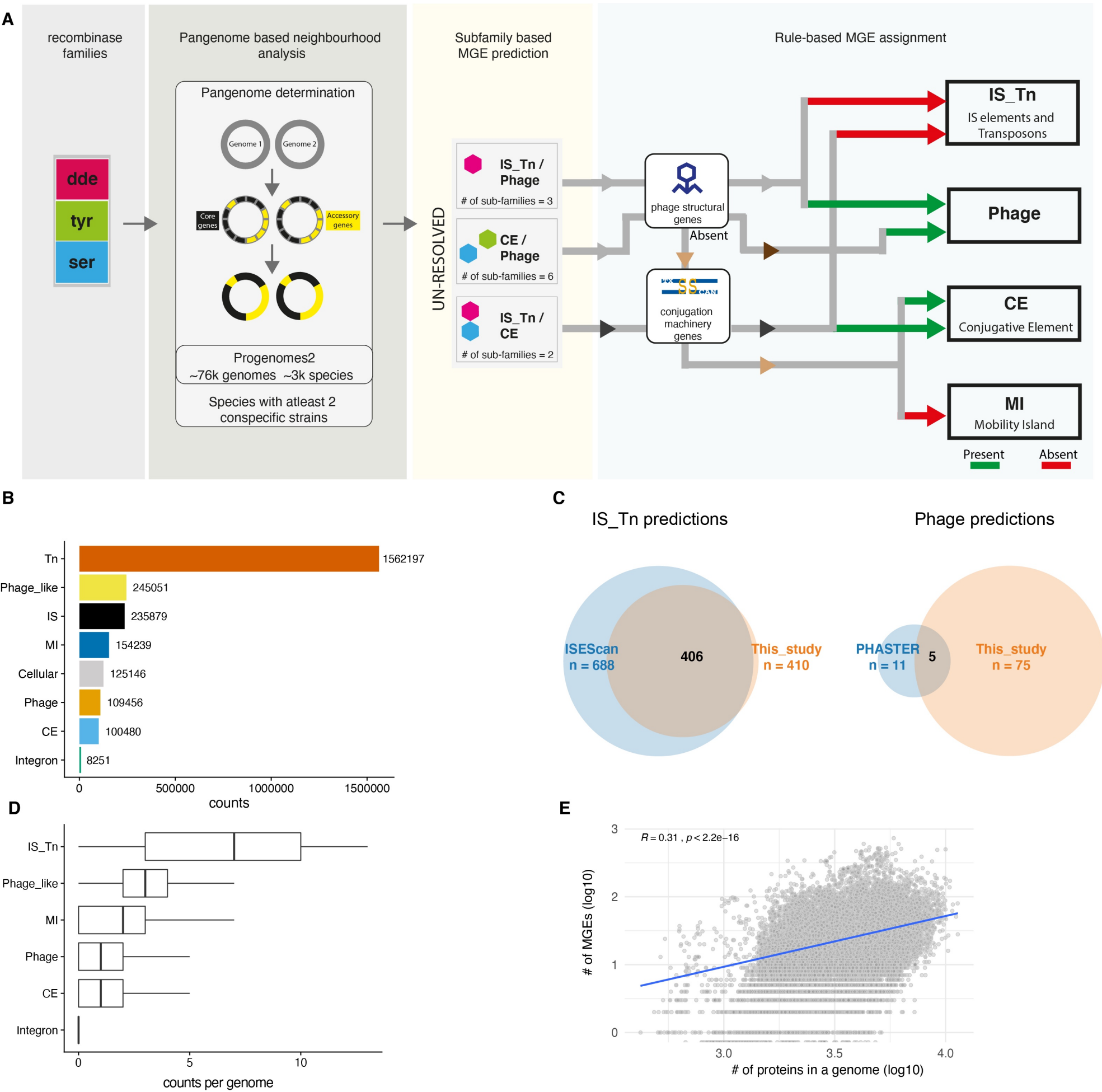
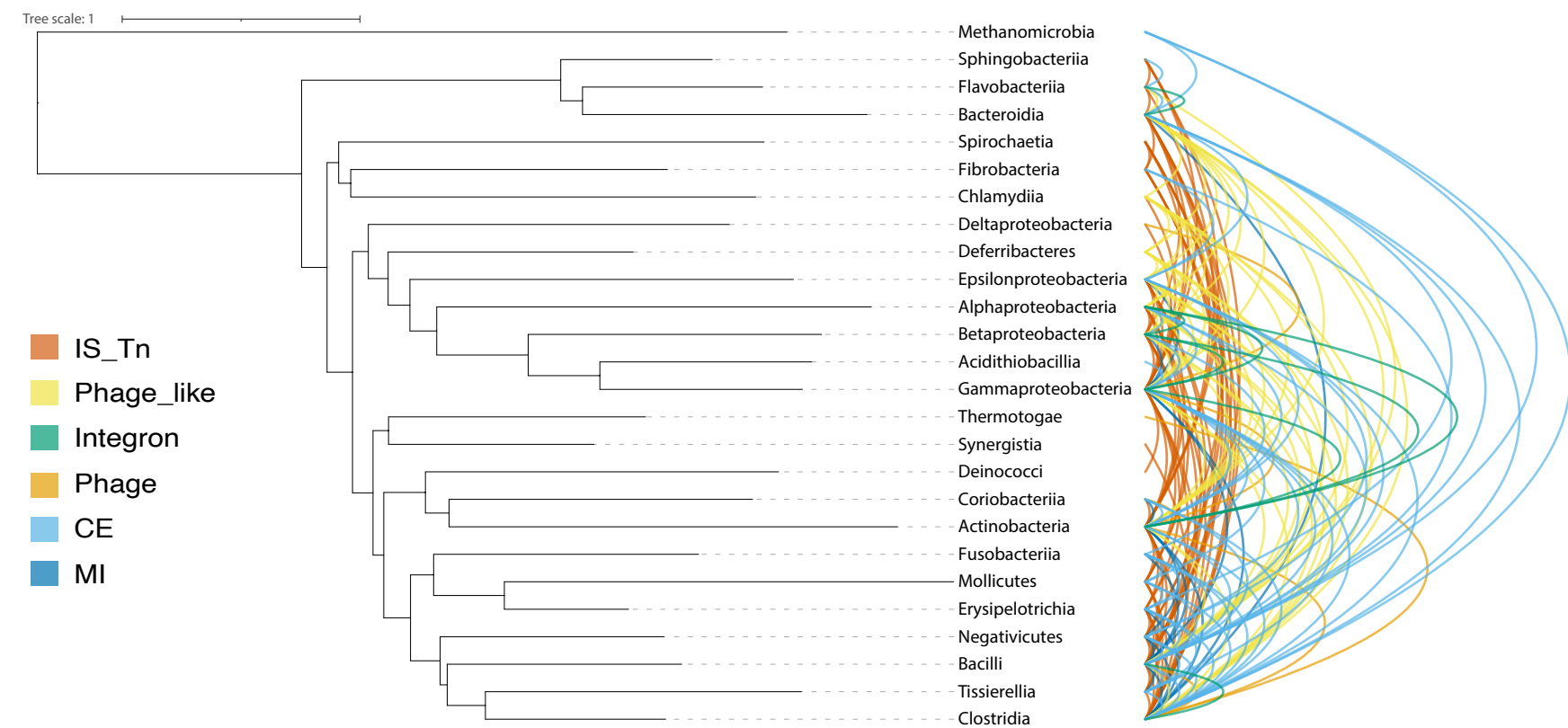
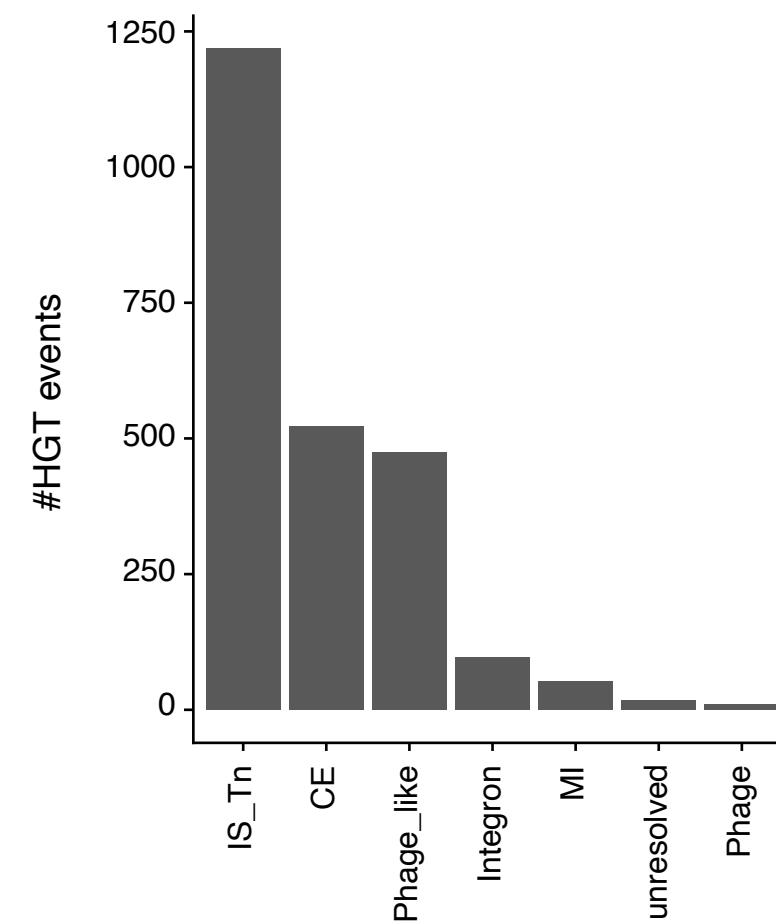


Figure S5

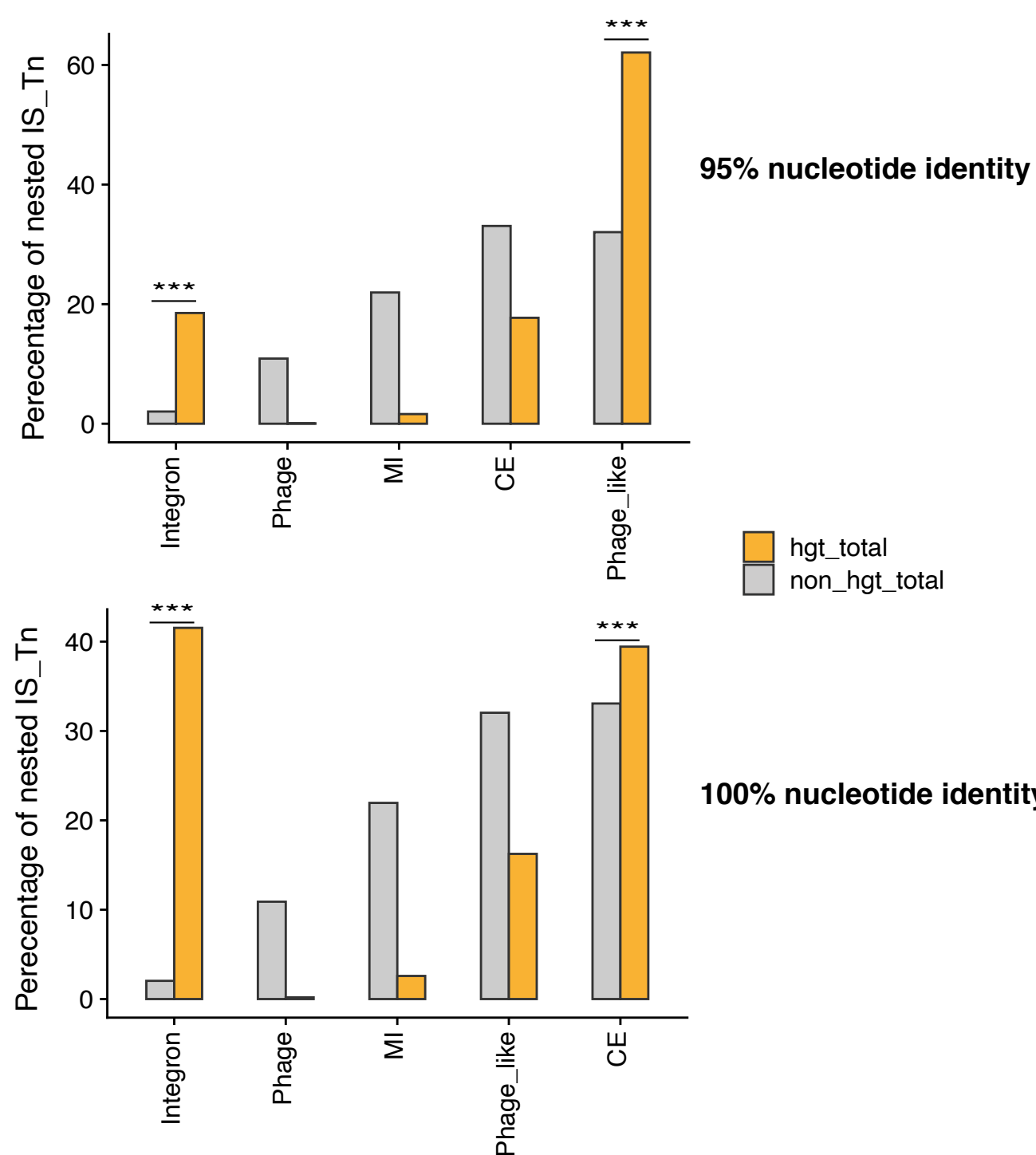
A



B



C



D

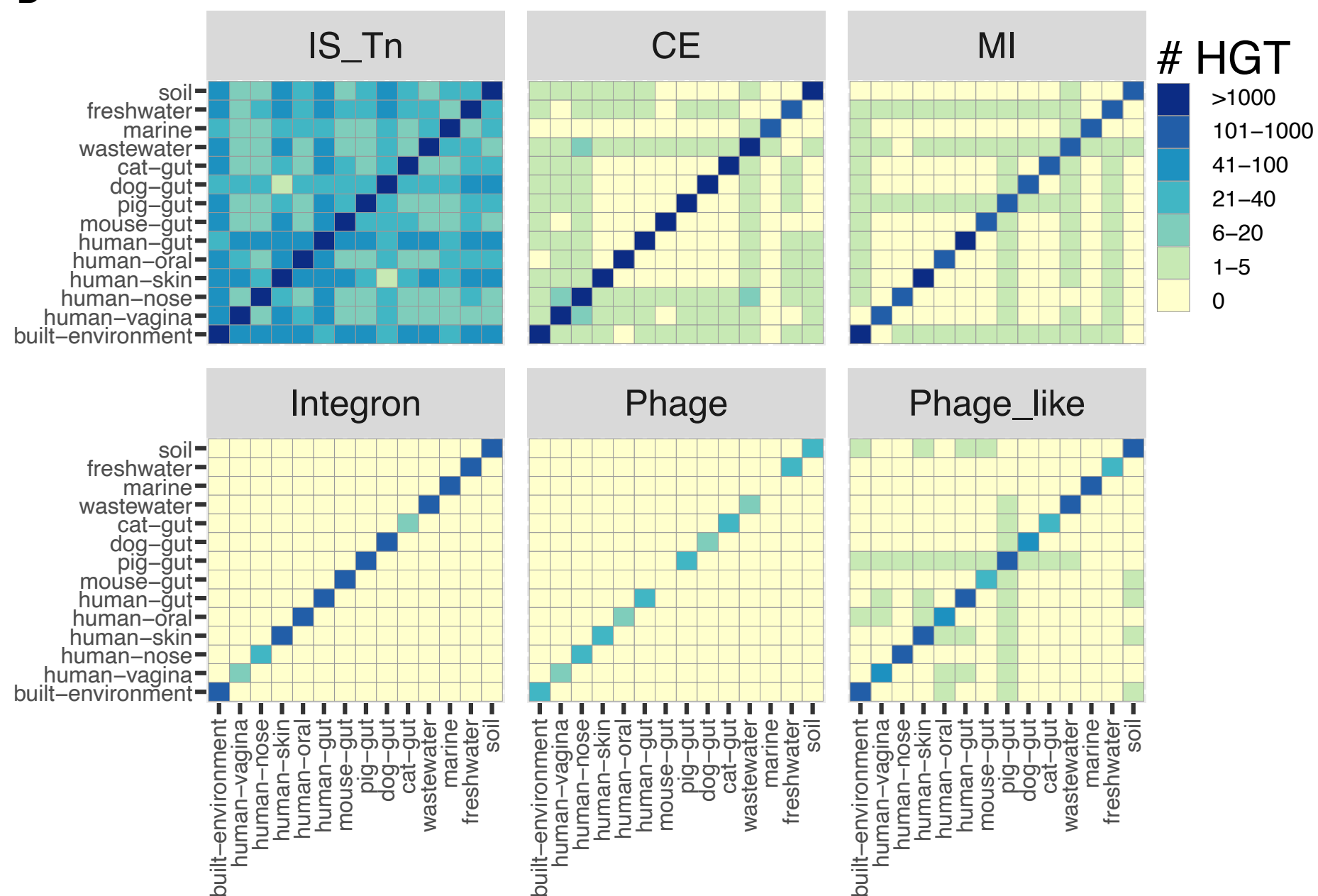


Figure S6

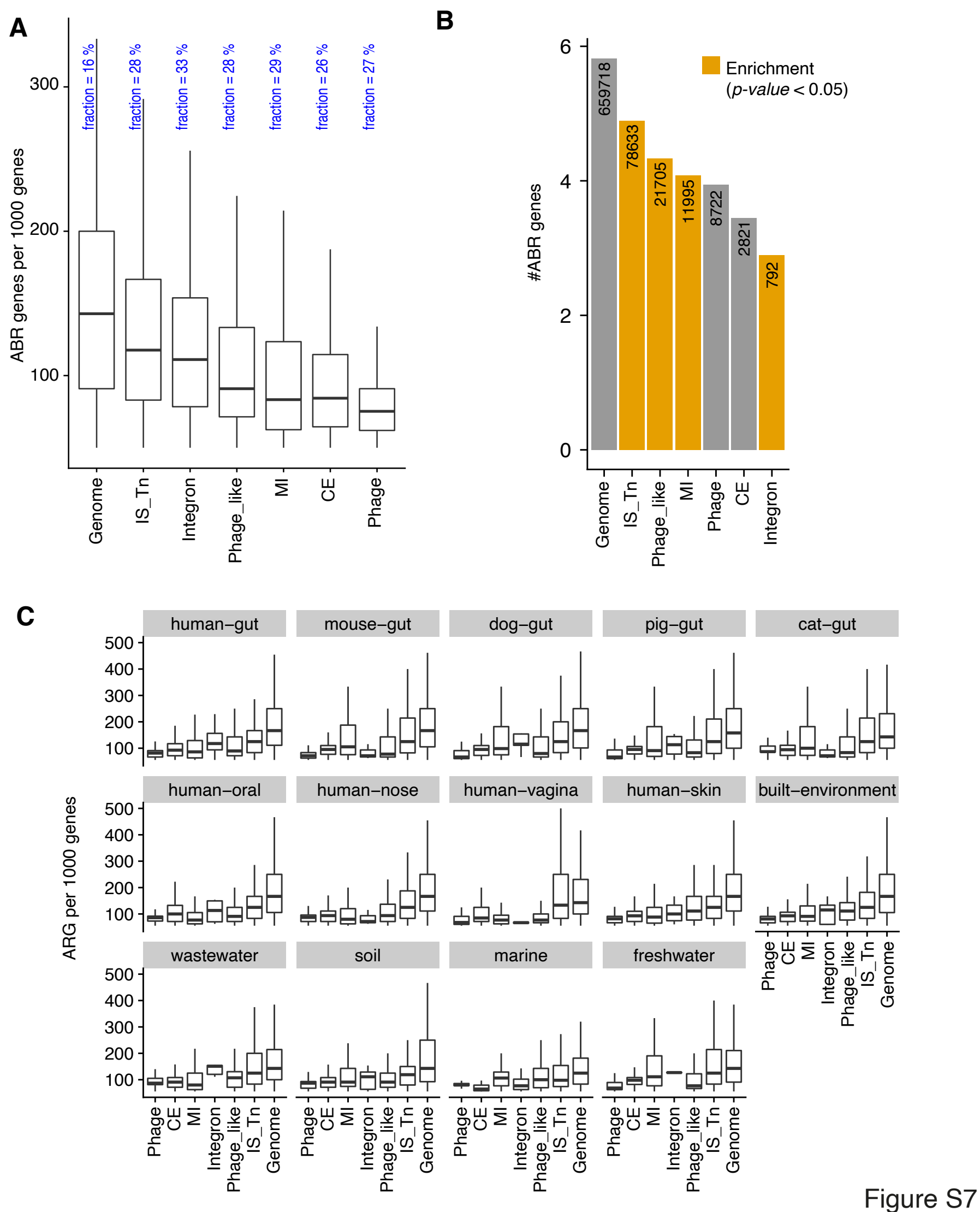


Figure S7