Southern Chinese populations harbour non-nucleatum Fusobacteria possessing homologues of the colorectal cancer-associated FadA virulence factor

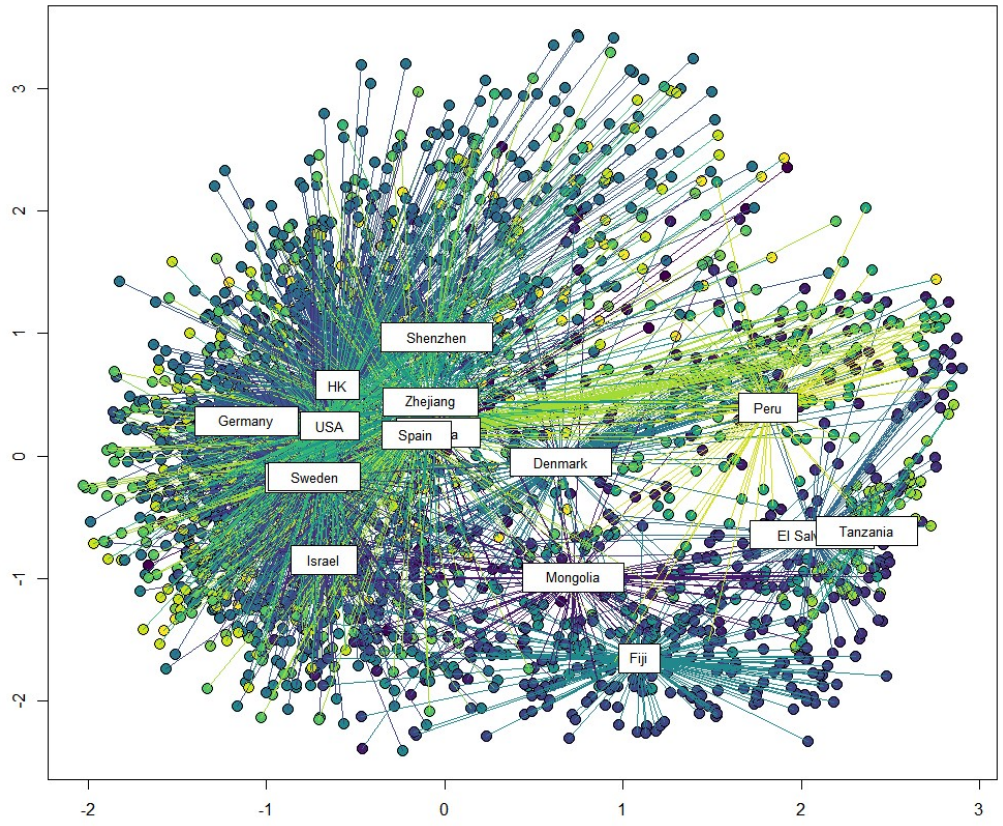

Figure S1: Principal component analysis of gut microbiota composition in non-colorectal cancer (CRC) individuals from Hong Kong,[2,29] China,[30,31] USA,[3,27] Austria,[4] Denmark,[32] France, Germany,[5] Spain,[32], Israel,[33] Sweden,[34,35] El Salvador, Peru,[36] Fiji,[37] Mongolia,[38] and Tanzania [37,40]. Read mapping counts generated by MetaPhlAn2 were centered log ratio-transformed. Each circle represents community composition of one sample; the closer two circles are the more compositionally similar their microbial communities. The spokes connect circles to their respective centroids according to cohort (centroids indicated by position of cohort label).

Southern Chinese populations harbour non-nucleatum Fusobacteria possessing homologues of the colorectal cancer-associated FadA virulence factor

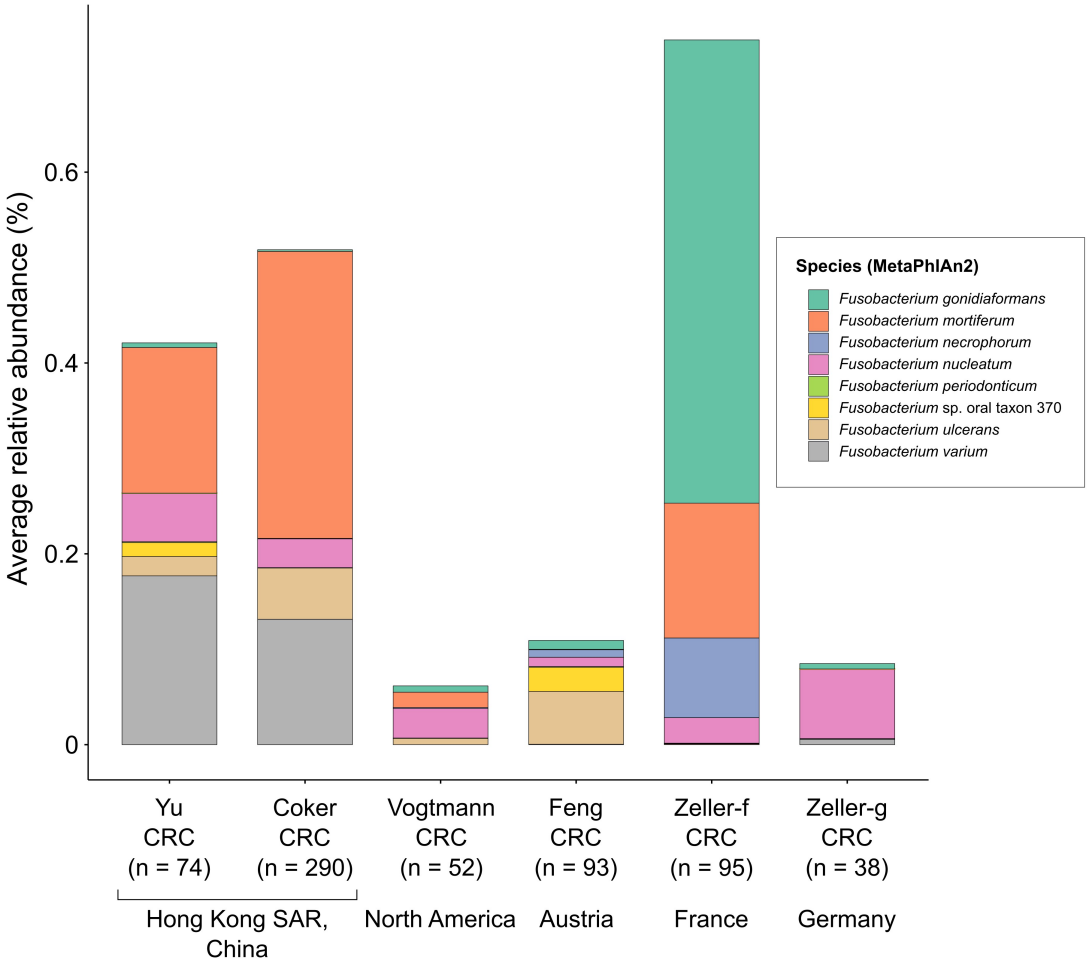

Figure S2: Average relative abundances of fusobacterial species detected in stool metagenomes of previously described CRC patients from various geographical backgrounds. Average relative abundances were calculated using MetaPhlAn2 on quality-filtered metagenome sequences, and values shown here for fusobacterial species are percentages of the total community. Each of the six stacked bars represent a study cohort: (i) Yu (Hong Kong),[2] (ii) Coker (Hong Kong),[29] (iii) Vogtmann (USA),[3] (iv) Feng (Salzburg, Austria),[4] (v) Zeller-f (France),[5] and (vi) Zeller-g (Germany).[5] Colours represent fusobacterial species detected in all metagenomes. Number of CRC subjects in each cohort are indicated in the cohort labels.

Southern Chinese populations harbour non-nucleatum Fusobacteria possessing homologues of the colorectal cancer-associated FadA virulence factor

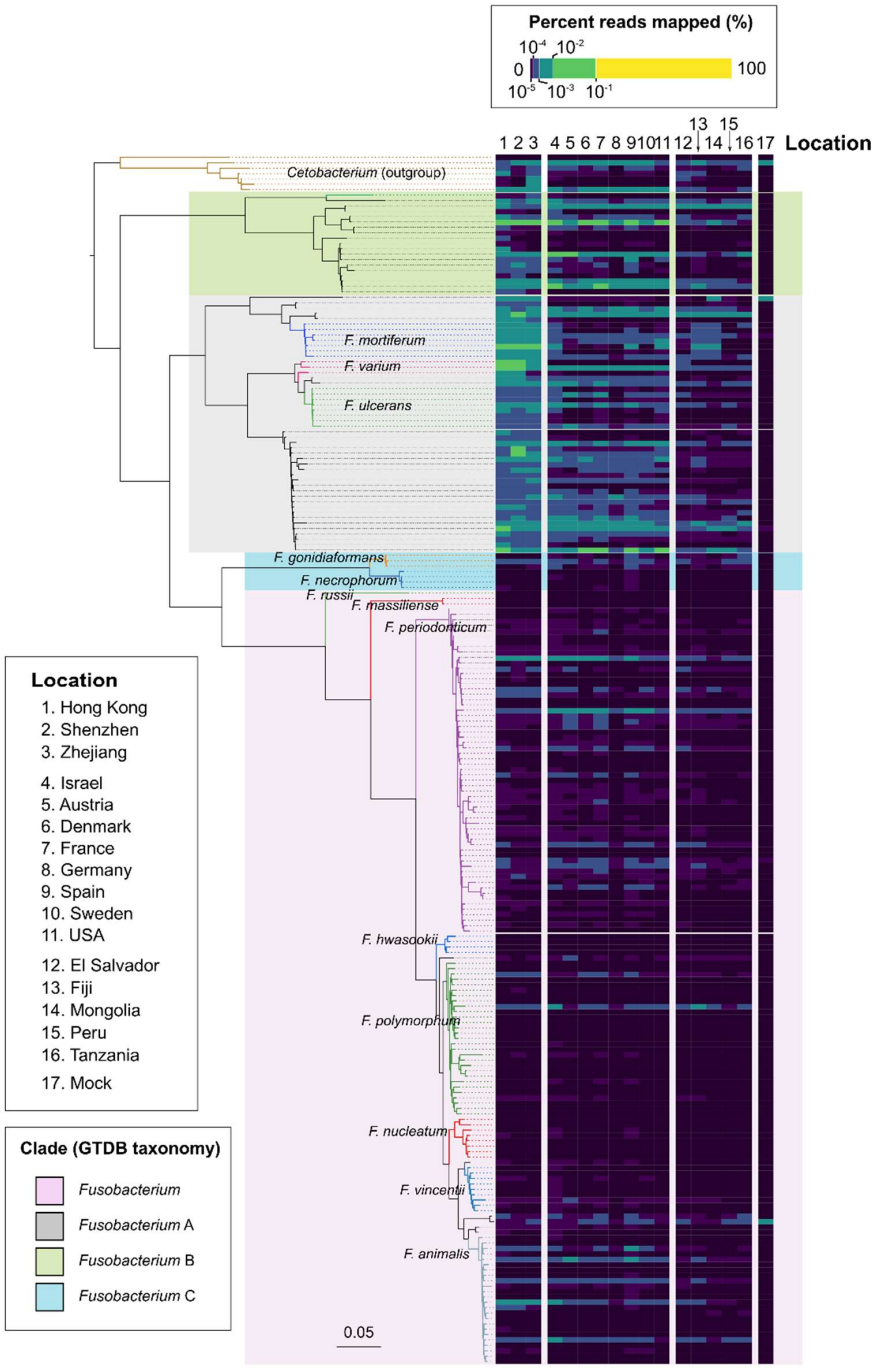

Figure S3: Proportion of sequence reads in non-colorectal cancer gut metagenomes mapped to dereplicated fusobacterial genomes. Increasing proportions of mapped reads are represented by lighter colours. Phylogenetic tree on left of figure shows inferred evolutionary relationships between genomes. Branch colours delineate species boundaries and do not represent any taxa in particular.

Southern Chinese populations harbour non-nucleatum Fusobacteria possessing homologues of the colorectal cancer-associated FadA virulence factor

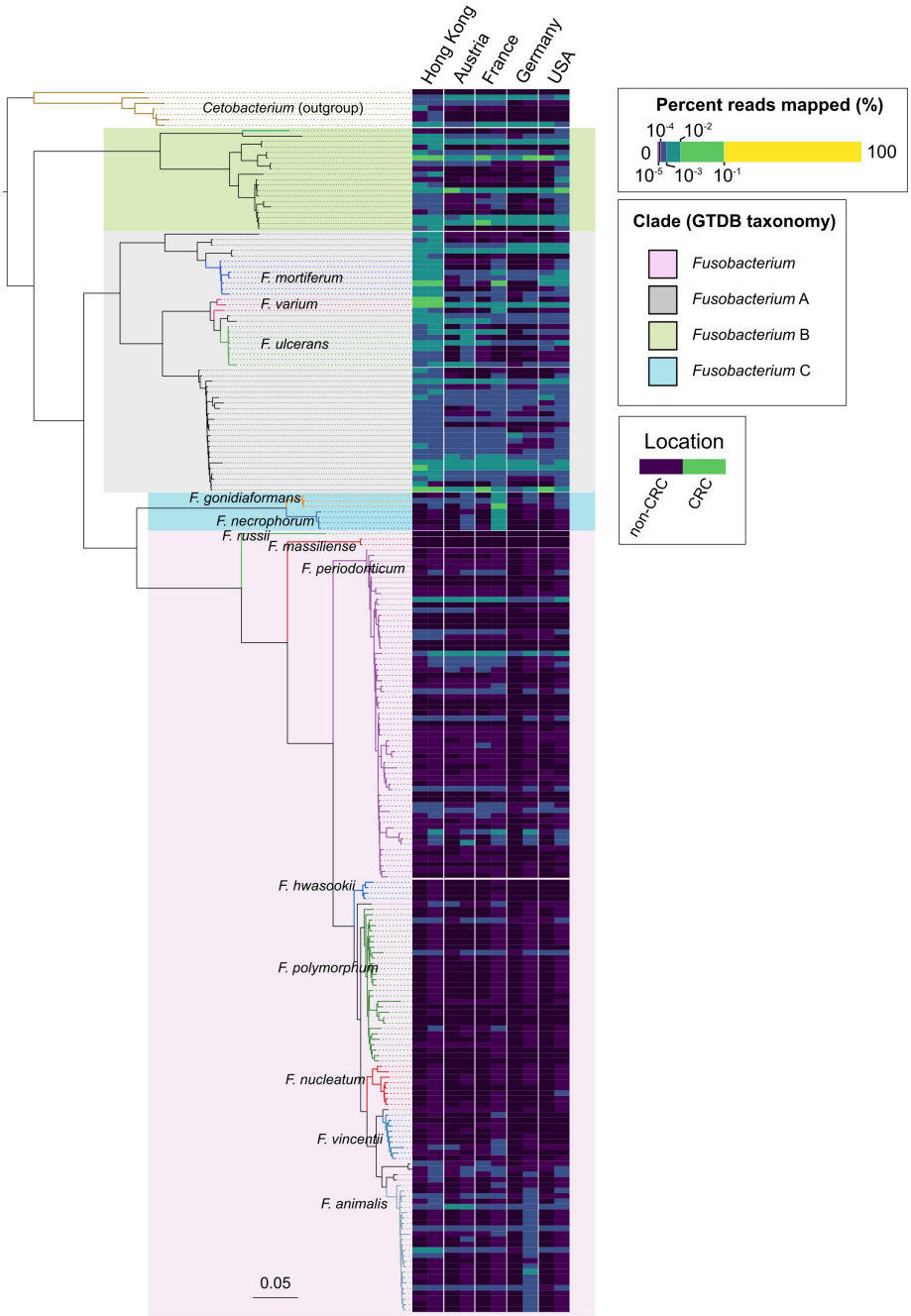

Figure S4: Proportions of sequence reads in colorectal cancer gut metagenomes mapped to dereplicated fusobacterial genomes. For each study location, the first and second columns indicate proportions from non-CRC controls and CRC cases, respectively. Increasing proportions of mapped reads are represented by lighter colours. Phylogenetic tree on left of figure shows inferred evolutionary relationships between genomes. Branch colours delineate species boundaries and do not represent any taxa in particular.

Southern Chinese populations harbour non-nucleatum Fusobacteria possessing homologues of the colorectal cancer-associated FadA virulence factor

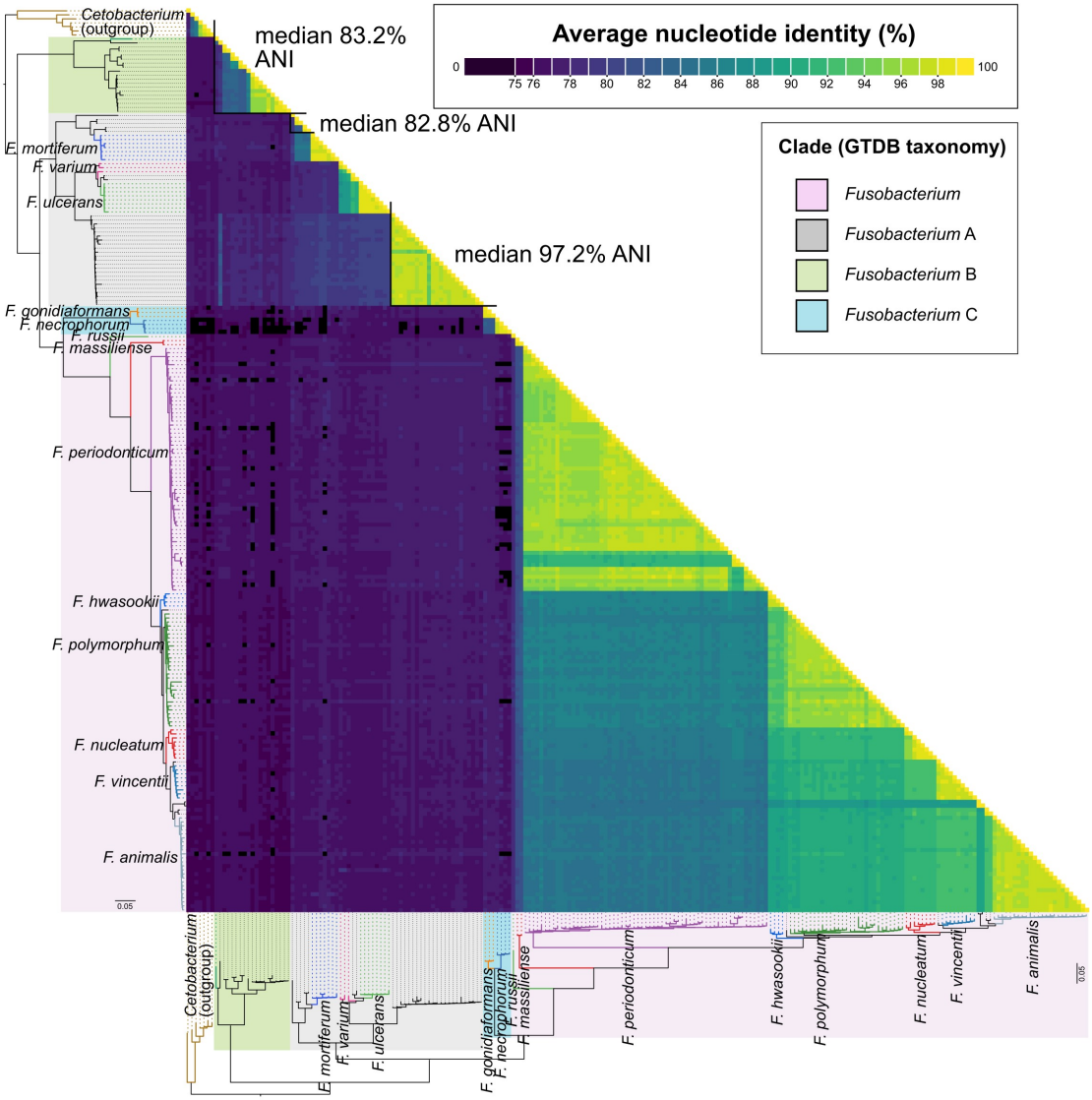

Figure S5: Pairwise average nucleotide identity (ANI) comparisons among dereplicated fusobacterial genomes. See **Table S10** for ANI values and species boundaries of putatively novel genomes identified in this study. Increasing percentage ANI is represented by lighter colours. Phylogenetic tree on left and bottom of figure shows inferred evolutionary relationships between genomes. Branch colours delineate species boundaries and do not represent any taxa in particular. ANI comparisons were performed using FastANI.[53]

Southern Chinese populations harbour non-nucleatum Fusobacteria possessing homologues of the colorectal cancer-associated FadA virulence factor

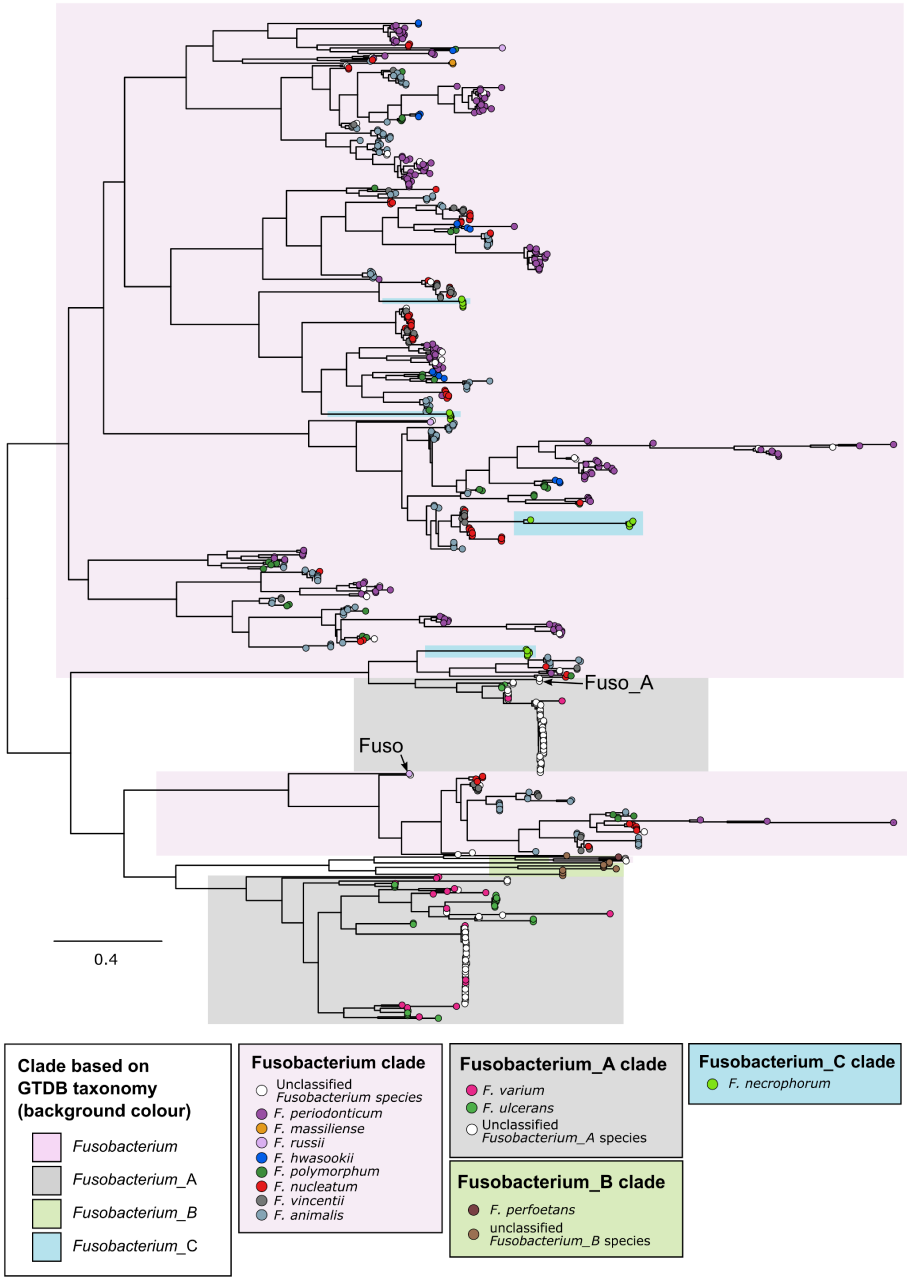

Figure S6: Phylogenetic relationships Fap2 protein homologues identified in fusobacterial genomes. Figure shows a maximum-likelihood tree of aligned amino acid sequences of Fap2 homologues rooted at the midpoint. Homologues were identified using eggNOG-mapper.[58] Each tip represents a homologue and is coloured according to species of the corresponding genomes they were found in. Background shading is according to the four major monophyletic clades identified in the genome-based phylogenetic tree in figure 2. Scale bars indicate amino acid substitutions per site. See **Table S14** for output from eggNOG-mapper.