Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/6/18/eaba0111/DC1

Supplementary Materials for

Phytoplankton pangenome reveals extensive prokaryotic horizontal gene transfer of diverse functions

Xiao Fan, Huan Qiu, Wentao Han, Yitao Wang, Dong Xu, Xiaowen Zhang, Debashish Bhattacharya*, Naihao Ye*

*Corresponding author. Email: d.bhattacharya@rutgers.edu (D.B.); yenh@ysfri.ac.cn (N.Y.)

Published 29 April 2020, *Sci. Adv.* **6**, eaba0111 (2020)
DOI: 10.1126/sciadv.aba0111

The PDF file includes:

Figs. S1 to S10

Other Supplementary Material for this manuscript includes the following:

(available at advances.sciencemag.org/cgi/content/full/6/18/eaba0111/DC1)

Data files S1 to S12

Supplementary figures:

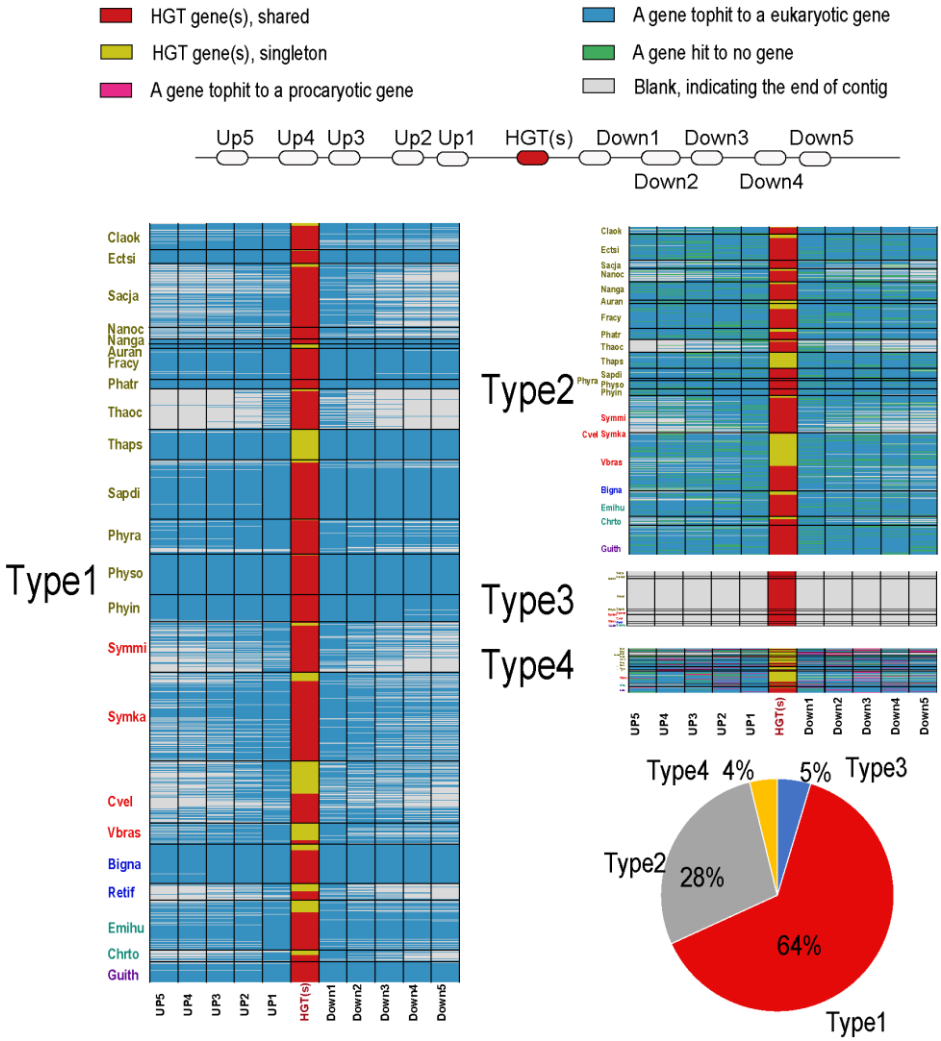

Fig. S1. Analysis of HGT flanking genes. Heatmaps indicate the HGT (red (shared) and yellow (singleton) blocks) and their flanking genes (up to 5). Colors of flanking genes indicate their origins including blue (eukaryotic genes), purple (prokaryotic genes), green (no hit except query themselves). Grey color indicate contig ends. Type1 indicates HGT gene(s) that are flanked by eukaryotic genes on both sides. Type2 indicates HGT gene(s) that are flanked by eukaryotic genes on one side and species-specific gene on the other side. Type3 indicates HGT gene(s) sitting at the contig with no flanking genes. Type4 indicates HGT gene(s) that were flanked by a mixture of eukaryotic genes, species-specific genes and prokaryotic genes. Pie chart indicates the percentage of each type.

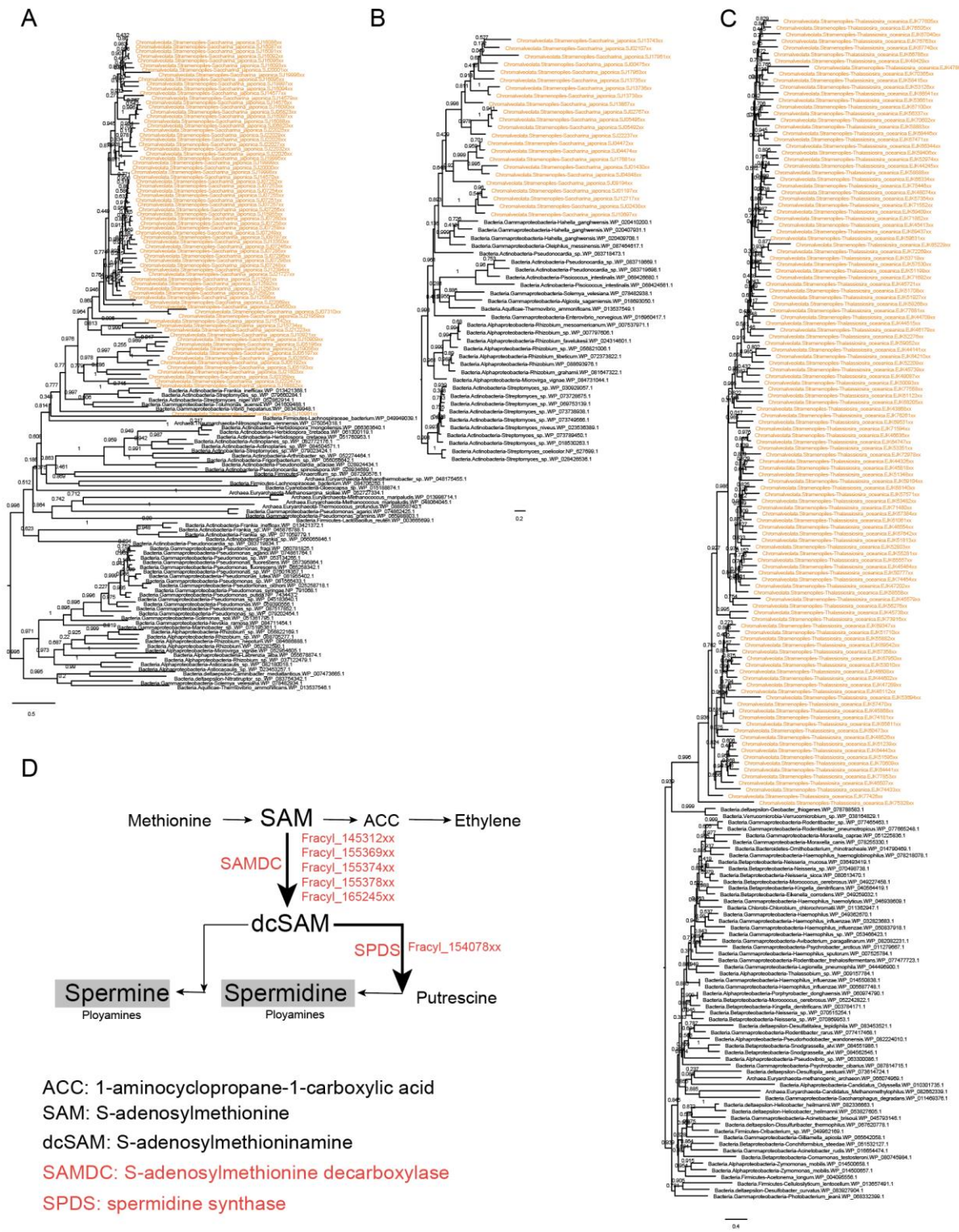

Fig. S2. Phylogenetic analysis of HGT-derived genes. (**A**) Tree of prokaryotic-derived mannuronan C-5-epimerases in brown algae (76 copies). (**B**) Tree of prokaryotic-derived polysaccharide lyases (23 copies) in brown algae. (**C**) Tree of prokaryotic-derived polysaccharide lyases 120 self-repeat family proteins in *Thalassiosira oceanica.* (**D**) Two important HGT events identified in the polyamine metabolic pathway in *F. clindrus*.

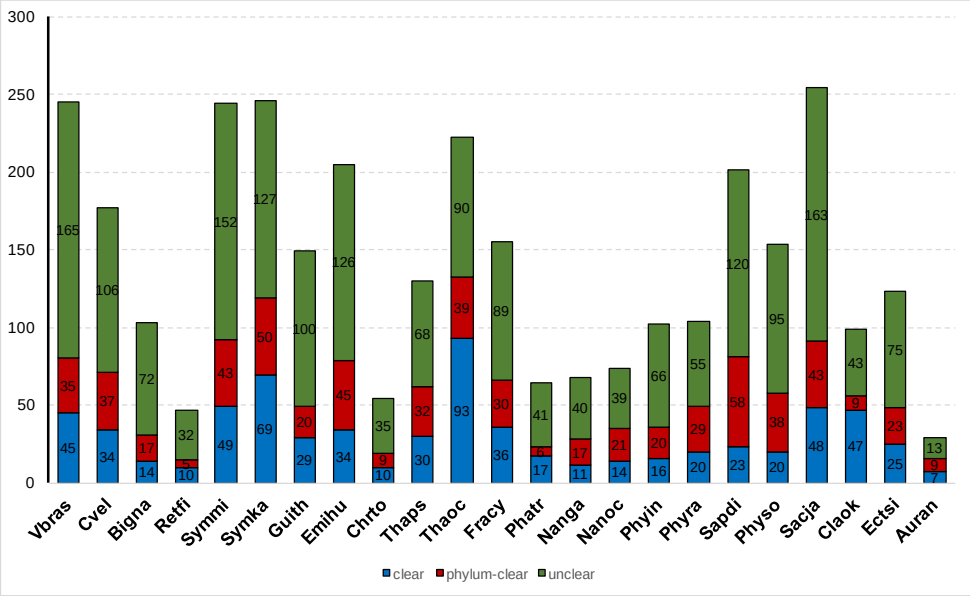

Fig. S3. Composition of CRASH HGTs with respect to how their origins can be identified. Blue color indicates the traceable origins at species level, red color indicates traceable origins at phylum level, and green color indicates no clear origin to be traced back to.

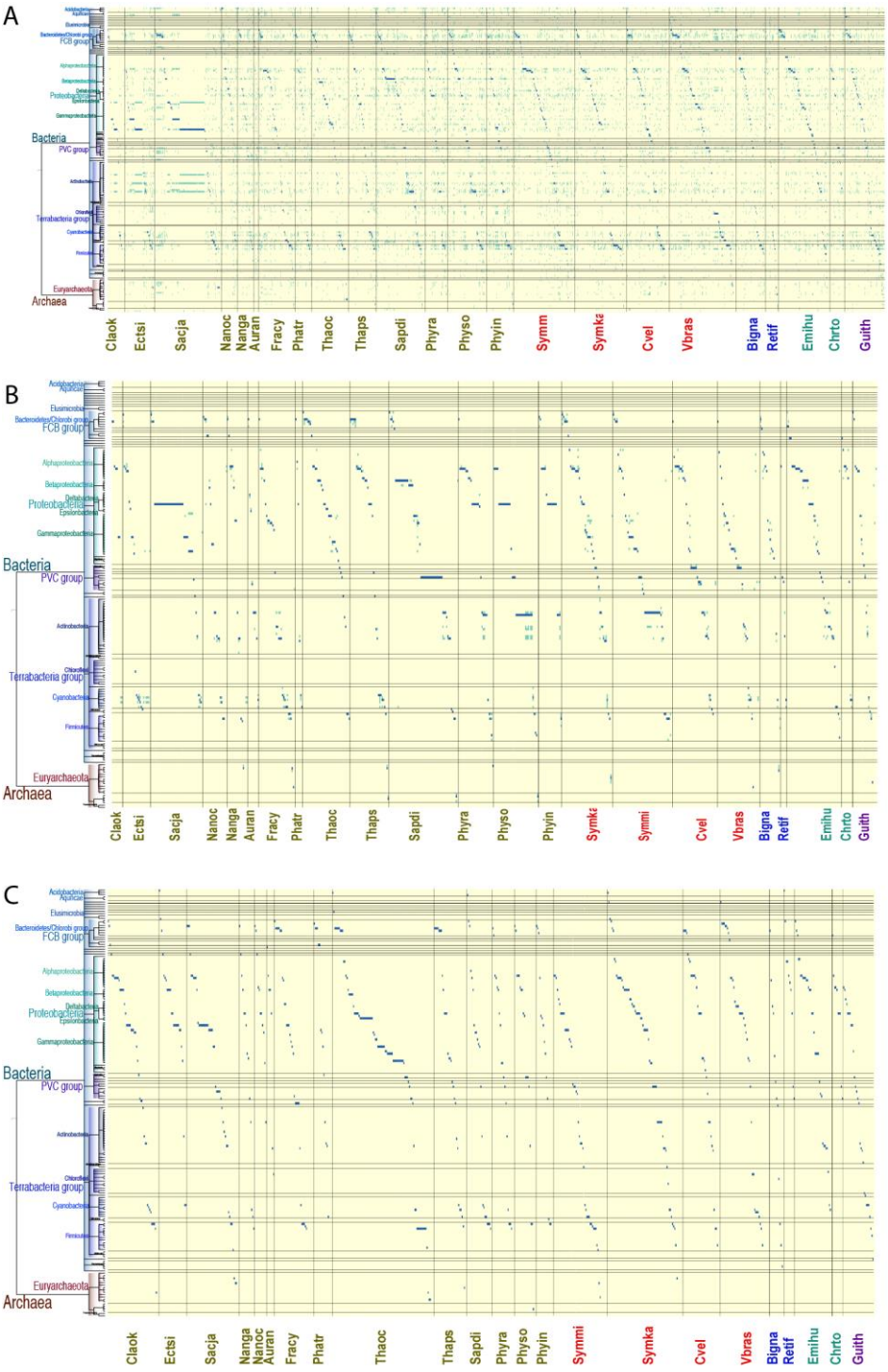

Fig. S4. Taxonomic distribution of prokaryotes within CRASH lineages. (**A**) Distribution that indicates ambiguous HGT origins. These monophyletic groups (UFboot >85%) comprise CRASH HGTs and two or more different prokaryotic phyla. The exact prokaryotic donors are unclear at the phylum level. The dark blue sticks indicate the prokaryotic genes that are the most similar to the HGTs whereas the green sticks indicate the remaining prokaryote genes in the monophyletic groups. The vertical lines separate data for different CRASH taxa and the horizontal lines for different prokaryotic phyla. Each column indicates a monophyletic group (or an HGT). For better visual appreciation, within each CRASH taxa, the monophyletic groups are ordered following the prokaryotic phylum associated with the genes the most similar the HGTs. The prokaryotic phylogeny (left panel) is retrieved from the NCBI taxonomy database. (**B**) Distribution that indicates unambiguous HGT origins at phylum level. These monophyletic groups (UFboot >85%) comprise CRASH HGTs and two or more genes from the same prokaryotic phylum. The figure is

annotated the same as Fig. S5. (**C**) Distribution that indicates unambiguous HGT origins at species level. These monophyletic groups (UFboot >85%) comprise CRASH HGTs and genes from the same prokaryotic species. The figure is annotated the same as Fig. S5.

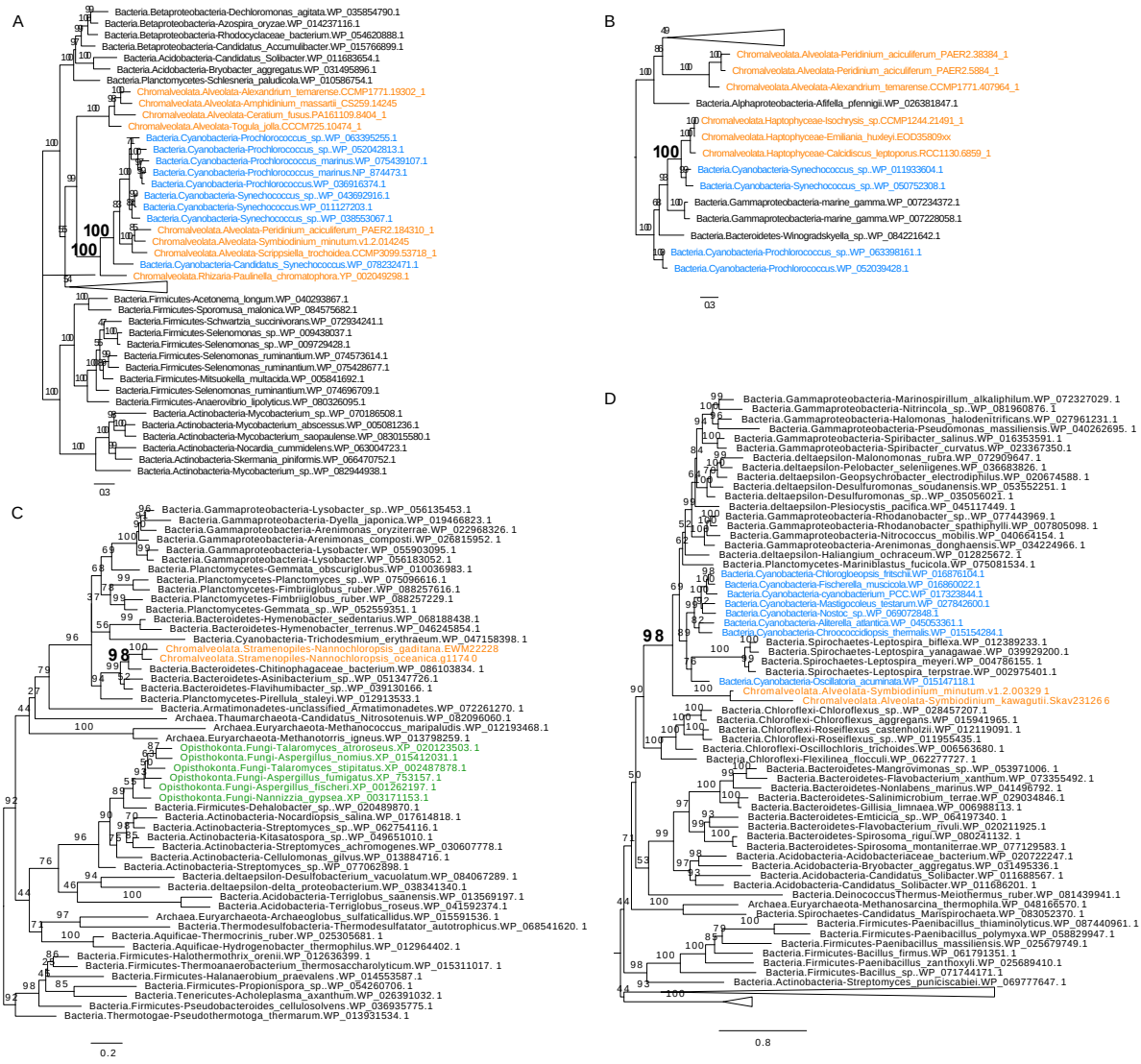

Fig. S5. Example of HGTs in CRASH species. (A) Maximum likelihood tree of an ABC transporter ATP-binding protein encoding genes transferred from *Prochlorococcus* into *Symbiodinium*. (B) Maximum likelihood tree of a DUF1254 domain-containing protein encoding gene transferred from *Synechococcus* into several haptophytes species. (C) An example of the genus-specific HGTs that are shared between two *Nannochloropsis* species (*N. oceanica* and *N. gaditana*). (D) An example of genus-specific HGT that is shared between two *Symbiodinium* species (*S. kawagutii* and *S. minutum*). Cyanobacteria are shown in blue and other prokaryotic sequences are in black. CRASH sequences are shown in orange and other eukaryotic sequences are in green.

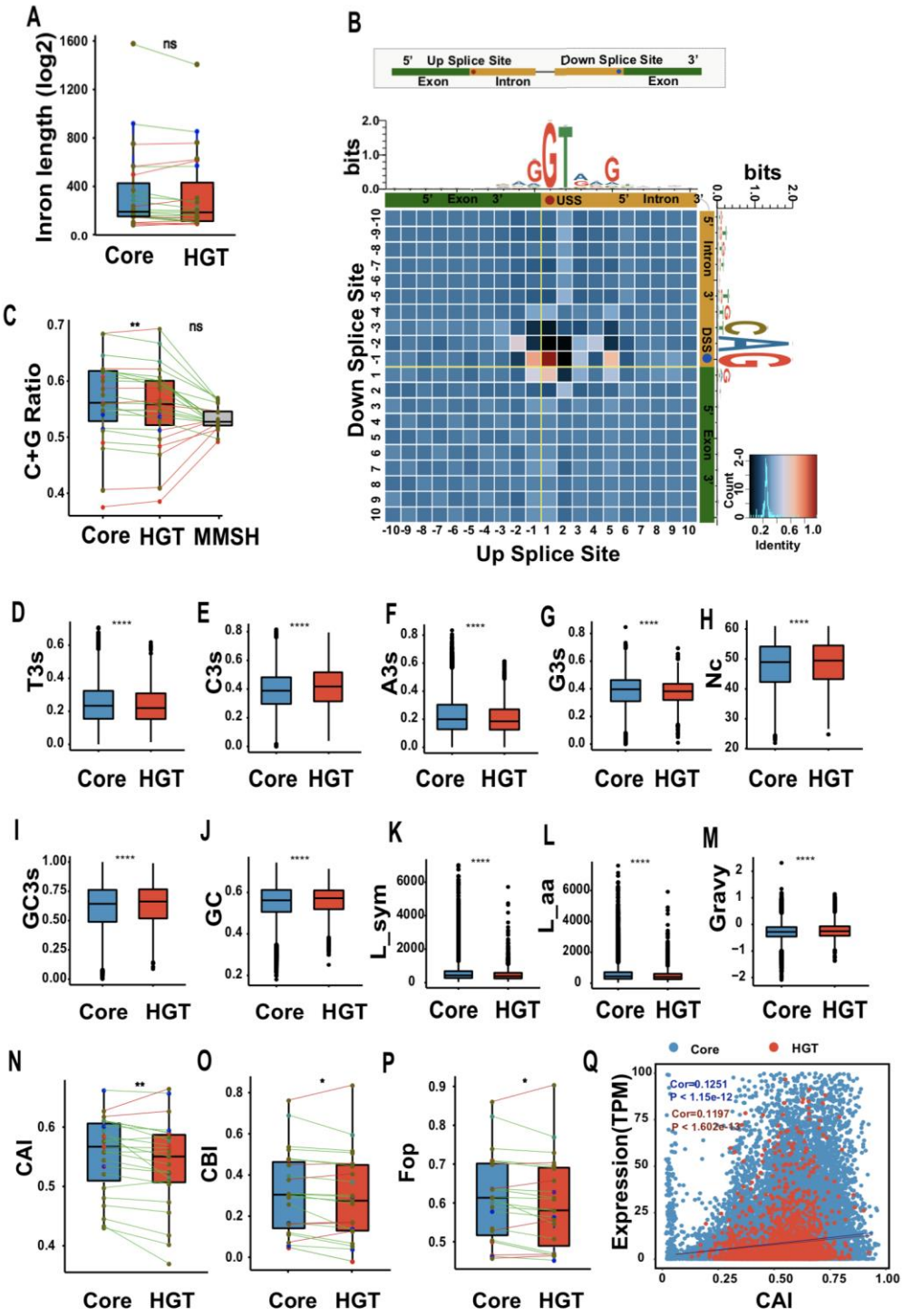

Fig. S6. Comparison of gene features between Core and HGT genes. (A) Non-significant difference in intron length between HGTs and core genes. (B) Conserved sequence motifs at HGT intron splice sites. (C) Higher exon GC-content in HGTs than in core genes. (D-M) Sequence statistics in HGTs and Core genes using the combined 23 CRASH-taxa data. A/T/G/C3s indicate the A/T/G/C content 3rd position of synonymous codons, respectively. Nc indicates the effective number of codons, a simple measure of overall codon bias analogous to the effective number of alleles measure used in population genetics. GC3s indicate G+C content 3rd position of synonymous codons. L_aa indicate the Length of amino acids. **** indicate *P*-value<1e-5. (N) Significant weaker codon adaptation index (CAI) in HGTs than in core genes. (O) Stronger CBI in core genes than in HGTs. (P) Higher Fop in core genes than in HGTs. (Q) Significant correlation between CAI and gene expression levels in Chromalveolate. The data from all ~600 Chromalveolate transcriptome data

were combined. HGT genes and core genes are distinguished by red and blue color respectively, both show the significant correlation between CAI and gene expression.

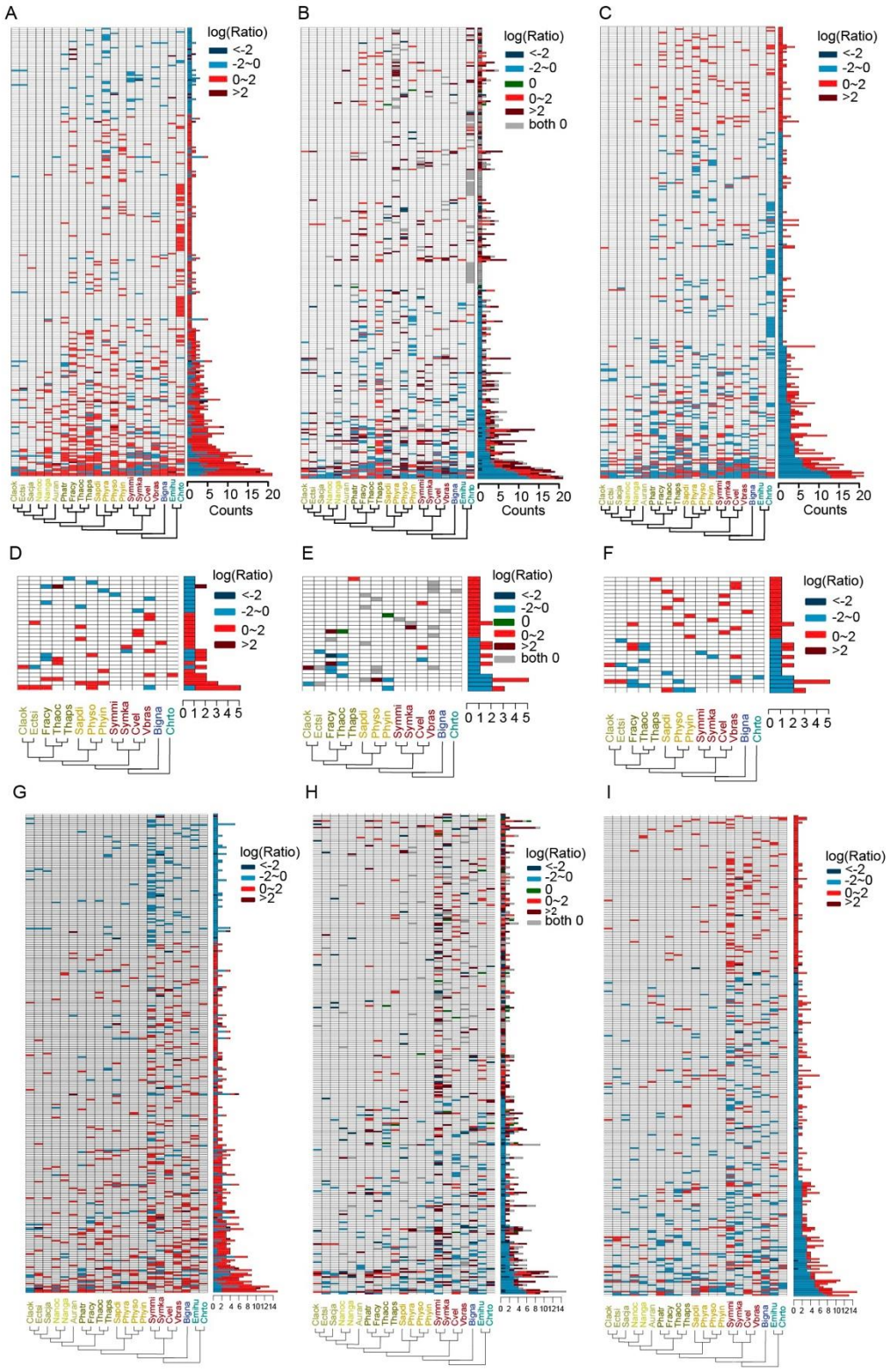

Fig. S7. Comparison of gene expression between Core and HGT genes. (A) Comparison of average gene expression levels between core genes and HGT genes among ~200 GO terms. Each row indicates a GO term. Red colors indicate higher mean expression levels in core gene than in HGTs (mean value of core genes / mean value of HGT genes >1), and blue colors indicate the opposite. For each species, all transcriptome data regardless of experimental conditions were combined. (B) Comparison of the coefficient of variation (CV) of gene expression across different conditions between core genes and HGT genes among ~200 GO terms pathways. Each row indicates a GO term. Red color indicates higher gene expression CV in core gene than in HGTs (CV_{core}/ CV_{HGT} > 1), and blue colors indicate the

opposite. (C) Comparison of gene expression specificity (for a given GO term, the ratio of the genes that unexpressed (defined as TPM <1) in any condition) between core and HGT genes. Red color indicates a greater portion of non-transcribed gene in core genes than in HGTs (specificity of core / HGT >1), and blue color indicates the opposite. (D) Comparison of average gene expression levels between core genes and HGT genes among 28 KEGG pathways. Each row indicates a KEGG pathway. Red colors indicate higher mean expression levels in core gene than in HGTs (mean value of core genes / mean value of HGT genes >1), and blue colors indicate the opposite. (E) Comparison of the Coefficient of Variation (CV: standard deviation / mean value) for gene expression across different conditions between core genes and HGT genes among 28 KEGG pathways. Each row indicates a KEGG pathway. Red colors indicate higher gene expression CV in core gene than in HGTs (CV_{core}/ CV_{HGT} > 1), and blue colors indicate the opposite. (F) Comparison of gene expression specificity (for a given KEGG pathway, the ratio the genes that unexpressed in any condition, TPM <1) between core and HGT genes. Red colors indicate a greater portion of genes is not expressed among core genes than among HGTs (specificity of core / HGT >1), and blue color indicates the opposite. (G) Comparison of average gene expression levels between core genes and HGT genes among Pfam terms. Each row indicates a Pfam term. Red colors indicate higher mean expression levels in core gene than HGTs (mean value of core genes / mean value of HGT genes >1), and blue colors indicate the opposite. (H) Comparison of the Coefficient of Variation (CV: standard deviation / mean value) for gene expression across different conditions between core genes and HGT genes among Pfam terms. Each row indicates a Pfam term. Red colors indicate higher gene expression CV in core gene than HGTs (CV_{core}/ CV_{HGT} > 1), and blue colors indicate the opposite. (I) Comparison of gene expression specificity (for a given Pfam term, the ratio the genes that unexpressed in any condition, TPM <1) between core and HGT genes. Red colors indicate a greater portion of genes is not expressed among core genes than among HGTs (specificity of core / HGT >1), and blue color indicates the opposite.

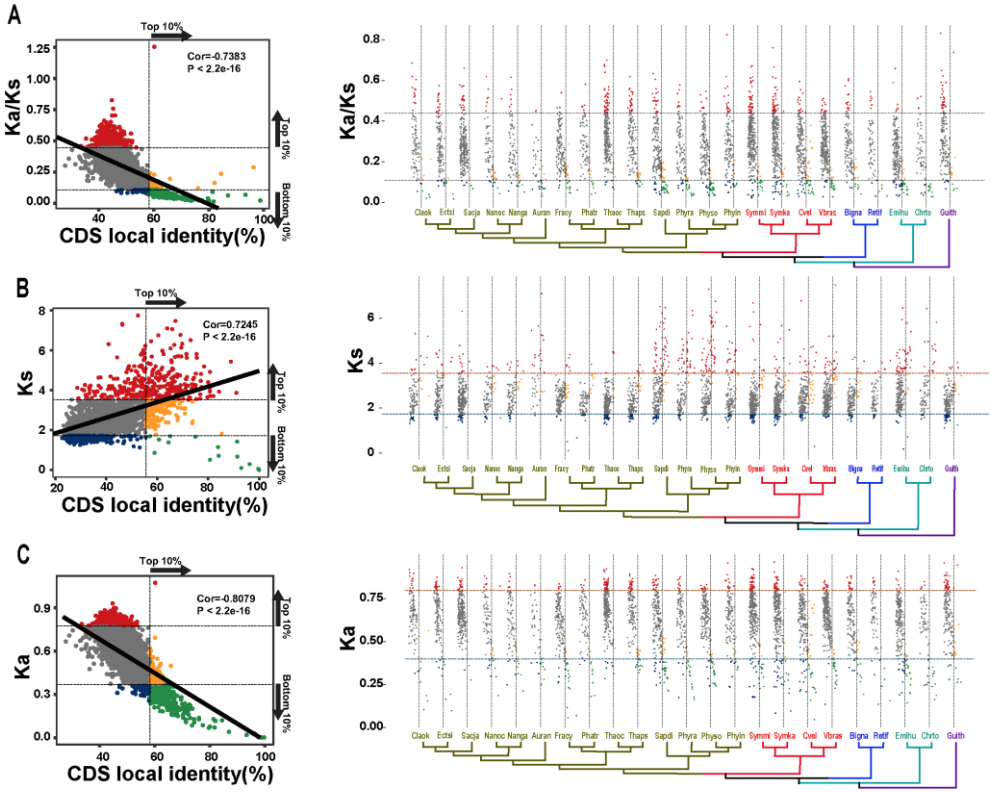

Fig. S8. Correlation between Ka/Ks, Ks, Ka and CDS local identity. (A) Significant correlation between Ka/Ks and CDS local identity. Sub-figure in left stand for the combined dataset of CRASH and sub-figure in right stand for separate dataset for each species. Y axis mark the Ka/Ks value where the dash lines stands for top 10% (up line) and bottom 10% (down line) respectively, X axis mark the CDS local identity where dash line stand for the top 10%. The same meaning for axis and dash lines for figure of left hand. (B) Significant correlation between Ks and CDS local identity. (C) Significant correlation between Ka and CDS local identity.

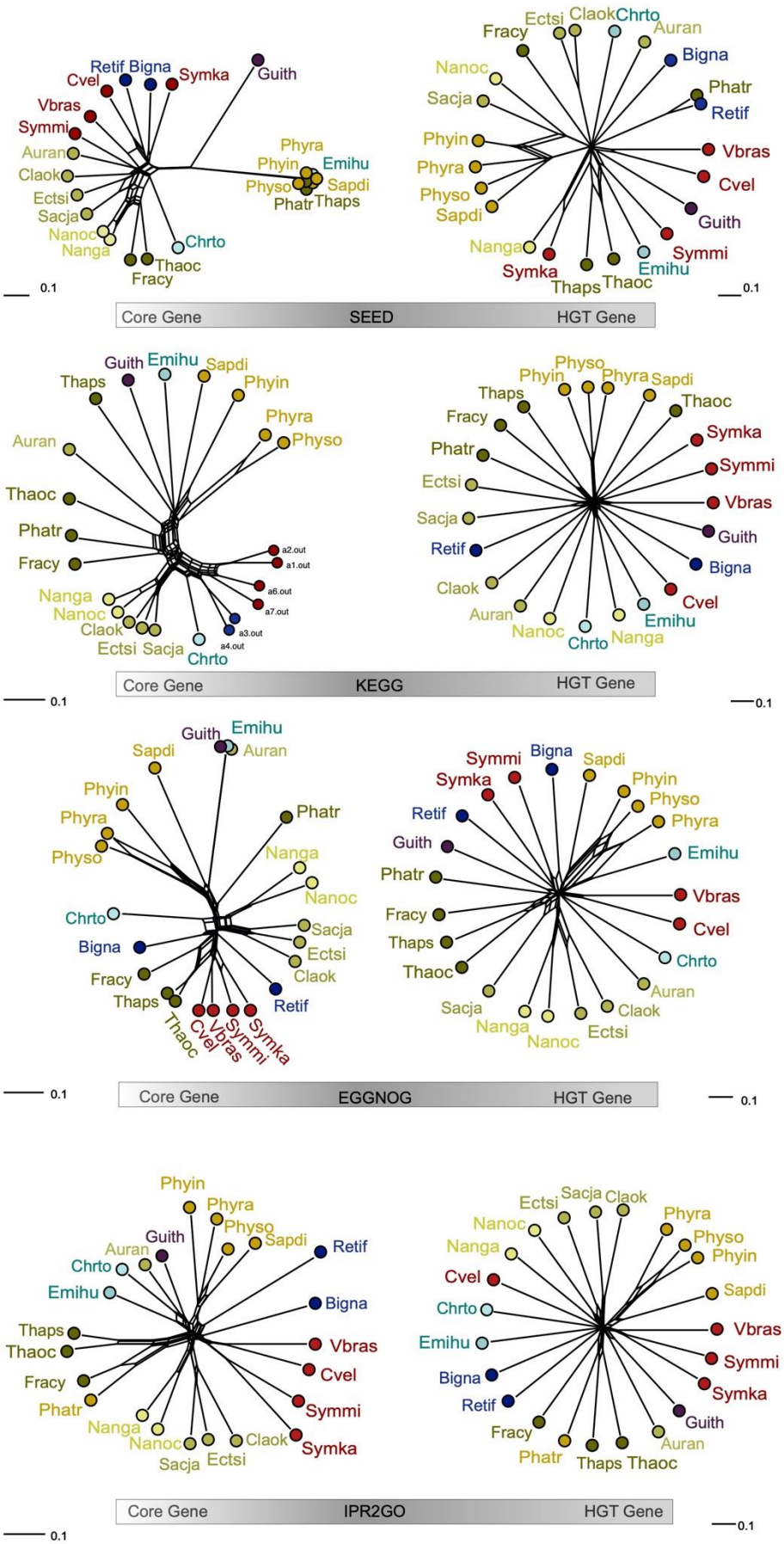

Fig. S9. Split neighbor-nets of CARSH lineages built on functions encoded in host core genes (left panel) and HGT genes (right panel). Four datasets are presented including those

based on annotation according to SEED database, KEGG database, EGGNOG database, and IPR2GO database from top to bottom.

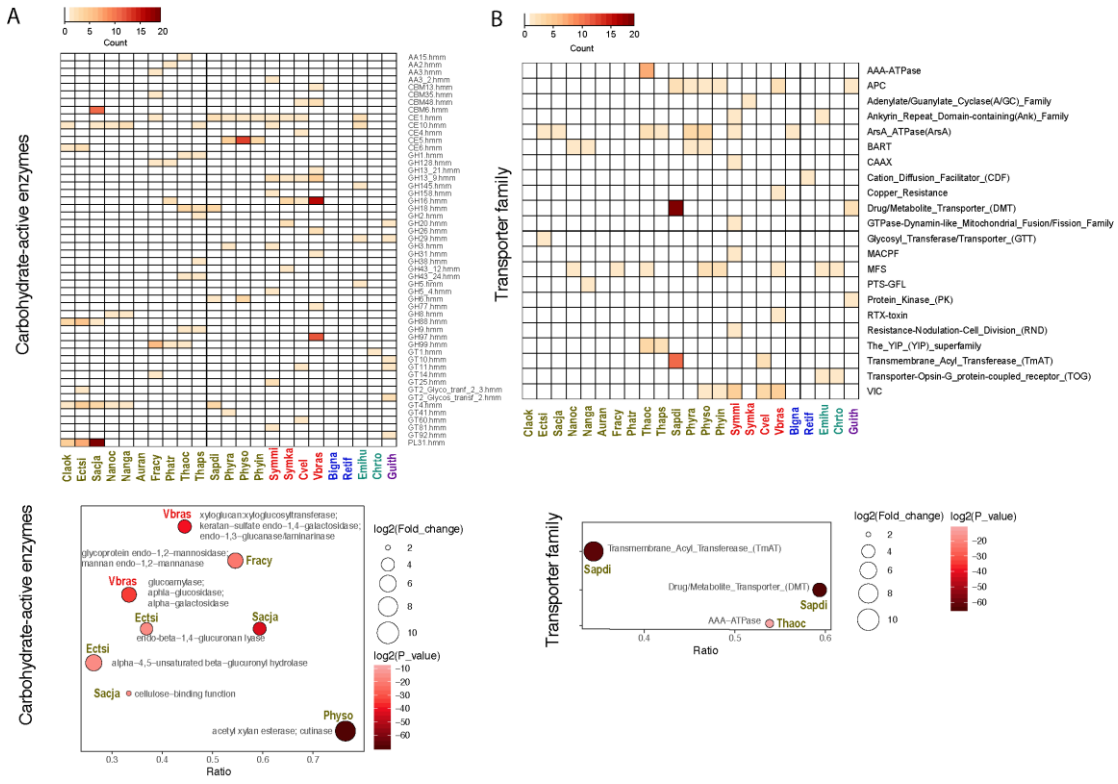

Fig. S10. CAZyme (A) and Transporter (B) distribution and enrichment in HGTs. Note that the enrichments were due to post-HGT gene duplications and not by preferential transfer of CAZyme genes.

#Data files S1 to S12.