Pets pooled analysis specialized R scripts and their inputs/outputs

Date updated: 09/09/2024

metadata.R (preprocessing)

This script combines metadata from each cohort and format the input for all the following analysis. Since sample ids can be differed by sources, platforms and workflows, metaphlan files are brought to help match samples from metadata to the exact taxonomic output.
Input:

A. Metadata for public studies (animals) retrieved from ENA: **WGX_studylist.csv**
B. Metadata for private studies (animals): **new_Hills_Metadata.csv**
C. Metadata for HMP1-II (human): **hmp1-ii_public_metadata_tidy.tsv**
D. Metadata for Madagascar (human) retrieved from ENA: **filereport_read_run_PRJNA485056_tsv.txt**
E. Sample ids used in metaphlan after filtration and normalization: **merged_taxonomic_SGB_animal_profile_normalized.csv**

Output:

A. Metadata for all samples used in the study: **metadata.csv**

Pets_pooled_analysis_clean.R (Figure 1D-E, Figure 4B and Supplemental Figs)

This script brings in the metadata and metaphlan files and generates panels for Figure 1, Figure 3, and the Supplement. Filters SGBs based on relative abundance and prevalence to obtain the final subset of SGBs (total of 2,274) used to generate the phylogenetic tree in Figure 1 and subsequent figures.
Input:

A. Metadata (output from metadata.R script): **all_metadata_preformatted.csv**
B. List of duplicate samples for removal: **duplicated_samples.csv**
C. Pets metaphlan 4 table: **pets_meta/metaphlan_taxonomic_profiles.tsv**
D. HMP1-II metaphlan 4 table: **hmp1_II/metaphlan_taxonomic_profiles.tsv**
E. Madagascar metaphlan 4 table: **CM_madagascar__metaphlan-4.0.4_vOct22_CHOCOPhlAnSGB_202212.tsv**
F. MAGs checkm output from assembly pipeline: **checkm_qa_and_n50.tsv**

Outputs:

A. Formatted MAGs files from pets assembly to be used in the SGB_tree_metahlan4.R script:
    a. **pets_MAGs_formatted_for_SGBtree.csv**
B. Formatted metaphlan profiles for pets samples to be used in the SGB_tree_metahlan4.R script:
    a. **pets_metaphlan_v4_formatted_SGBtree.csv**
C. Formatted metaphlan profiles for human samples (HMP1-II and Madagascar) to be used in the SGB_tree_metahlan4.R script:
    a. **ALLhuman_metaphlan4_SGBs_formatted_SGBtree.csv**

Plots:

A. Fig 1D: venn diagram for overlap of SGBs across hosts
    a. **v4_VENN_0.00001_atLeast3samples.pdf**
B. Fig 1E: Proportion uSGBs/sample by study (abundance weighted)
    a. **proportion_uSGBs_alldata_abund_weighted.pdf**
C. Supplemental Fig@@@ MAG quality

        a. **MAG_quality_scatter.pdf**
        b. **MAG_quality_BAR.pdf**
        c. AI final figure: **checkm_figure_AI_version.ai**
    D. Fig 4A: microbes that are shared across all, CA only, or unique (prevalence when present)
        a. **figure4_top_panel.pdf**

SGB_tree_metaphlan4_clean.R (Figure 1C main tree and Figure 2)

The two main objectives for this script are to 1) prepare the annotation file and subset the Oct22 newik tree file (to include the SGBs identified by maaslin in our dataset) for visualization using graphlan (main tree in Figure 1), and 2) constructs the novel SGB tree in Figure 2.
Input:

A. Metadata (from metadata.R):
    a. **all_metadata_preformatted.csv**
B. Pets metaphlan file (formatted from Pets_pooled_analysis_clean.R):
    a. **pets_metaphlan_v4_formatted_SGBtree.csv**
C. Human metaphlan file (formatted from Pets_pooled_analysis_clean.R):
    a. **ALLhuman_metaphlan4_SGBs_formatted_SGBtree.csv**
D. Soil taxonomic profiles
    a. **NorthAmerican_soil/metaphlan_taxonomic_profiles.tsv**
E. MAGs (formatted from Pets_pooled_analysis_clean.R):
    a. **pets_MAGs_formatted_for_SGBtree.csv**
F. Oct 22 tree provided by Francesco:
    a. **Oct22.nwk**

Output:

A. Annotation file for graphlan
    a. **annotation_file_metaphlanv4Oct22_7.11.txt**
B. Subsetted Oct22 tree that includes the 2,274 SGBs identified in our dataset (for input into graphlan)
    a. **subsetted_tree_Oct22.txt**

Plots:

A. Novel tree (Figure 2)

Figure3_PcoA&Heatmap_renamed.R (Figure 3A,C)

This script prepares panels in Figure 3, showing the taxonomic patterns of samples across host species. The original metaphlan files are first transformed to the taxonomic profiles on SGBs level used in the analysis. Samples with unknown SGBs relative abundance = 100 are excluded. The relative abundance of each sample is then renormalized after removing the unknown SGBs (row).
For the visualization, this script includes the frequency-corrected (based on sample size) PCoA with SGB annotations selected and the union of the 8 most abundant SGBs found in each host SGBs heatmap. The script also generates major supplementary figures, including density plots, showing the distribution of relative abundance, and bar plots, showing the proportion of relative abundance = 0, of selected SGBs in each host species.
Input:

A. Formatted animal SGB profiles (metaphlan taxonomic profiles subsetted at the SGB level and contains all dog and cat profiles; output from Pets_pooled_analysis_clean.R): **pets_metaphlan_v4_formatted_SGBtree_Jun23tax.csv**

B. Combined human (HMP1-II and Madagascar) metaphlan table subsetted at the SGB level (output from Pets_pooled_analysis_clean.R): **ALLhuman_metaphlan4_SGBs_formatted_SGBtree_Jun23.csv**

C. Metadata (output from metadata.R script): **metadata.csv**

Output:

A. Combined MaAsLin2 results: **SGB_maaslin_FDRcorrected.csv**

Plots:

A. Figure 3A: Annotated frequency-corrected taxonomic PcoA colored by host species
    a. **weighted_pcoa_2xhuman_species_v4.pdf**

B. Figure 3C: Annotated taxonomic heatmap of top 8 union SGBs from each host species
    a. **heatmap_filtered_top27_2Xhuman_v4_naturalcluster.pdf**

C. Supplemental Fig@@@ Box plots of selected SGBs reflecting dog housing conditions (facility vs. private households)
    a. box_housing.pdf

D. Supplemental Fig@@@ Density plots of selected SGBs characterized by host species and SGBs identified from the union of top 10 most abundant SGBs in each host +miscellaneous SGBs
    a. density_all_supplement.pdf

strainphlan_tree_vis_automated.R (Figure 4A phylogenetic trees)

Generates a phylogenetic tree for each SGB from the strainphlan-generated bestTree files. Select trees are included in Figure 4.

Input:

A. Metadata (from metadata.R):
    a. **all_metadata_preformatted.csv**

B. Pets metaphlan file (formatted from Pets_pooled_analysis_clean.R):
    a. **pets_metaphlan_v4_formatted_SGBtree.csv**

C. Directory of bestTrees generated by Strainphlan 4

Output:

A. Directory containing a visualization of each SGB's phylogenetic tree

Plots:

A. Figure 4A phylogenetic trees for R. gnavus, P. vulgatus, B. wexlerae, and a Firmicutes uSGB

pets_coherence_score.R (Figure 4B coherence score)

This code is adapted from Dr. Ali Rahnavard's coherent_score.R written for his work in "Epidemiological associations with genomic variation in SARS-CoV-2". It adds a quantification component to the SGB-specific phylogenetic trees in Figure 4A by calculating a coherence score for each SGG (measures the host-specificity / niche-specificity of SGB subclades). Coherence score is shown in Figure 4B and Supplemental Figure@@@.

Input:

A. Metadata (from metadata.R):
    a. **all_metadata_preformatted.csv**

B. Pets metaphlan file (formatted from Pets_pooled_analysis_clean.R):
    a. **pets_metaphlan_v4_formatted_SGBtree.csv**

C. Directory of SGB-specific distance matrices: Kimura 2-parameter distance calculated on alignment files generated by Strainphlan 4

Output:

A. **all_silhouette_scores_per_bug.csv**: contains the raw data for the host-specific coherence scores

B. Figure 4B mean coherence score barplot (SGBs with highest 25 mean coherence score and lowest mean coherence score): **silhouette_score_barplot_truncated.pdf**

C. Supplemental Figure@@@, combines the following

   a. extended coherence score barplot: **silhouette_score_barplot_mean.pdf**

   b. heatmap showing host-specific coherence scores: **silhouette_score_heatmap_mean.pdf**

anpan_results_Figure_clean.R (Figure 5)

This script takes in the output from anpan's strain analysis, which was completed on the cluster. Specifically, we employed anpan's gene model which identified genes that are enriched in strains across different outcomes. In our case the outcomes are hosts and anpan was run pairwise for hosts i.e., cats vs. dogs, humans vs. dogs, humans vs. cats.

Input:

A. To generate **Figure 5A**: Species-specific **gene_terms.tsv** output file from anpan

B. To generate **Figure 5B**: three dataframes that contain pairwise "final gene" information, that is, the final genes that survived filtering in the anpan gene model (example: filtered_Ruminococcus_gnavus_catdog).

Output

A. A dataframe (called anpan_gene_diffs_df) that stores all of the species-specific values of enriched genes across hosts; this dataframe is the basis for Figure 5A.

B. Rgnavus_Anpan_heatmap_top20.pdf (**Figure 5B**), Rgnavus_Anpan_heatmap_top50.pdf (**Supplemental Fig.@@@**), Rgnavus_Anpan_heatmap_Glycosyl.pdf (**Supplemental Fig.@@@**)

Figure6_AMR_analysis.R (Figure 6)

Input:

A. Metadata (from metadata.R):

   a. **all_metadata_preformatted.csv**

B. Humann-generated table of gene families (unstratified, relative abundance, and for ease of use, subsetted to be include only the gene fams found in CARD)