Workflow description for generating HMMs for a gene set of interest as used in “*Charting host-microbe co-metabolism in skin aging and application to metagenomics data*”

1. Literature study: domain experts generate a list of gene symbols, locus tags and organisms by reading scientific literature.
2. Genome collection: download genome sequences in genbank format from RefSeq (<https://www.ncbi.nlm.nih.gov/refseq/>).
3. Orthology: generate orthology information using one of the publicly available tools for this purpose (e.g., orthAgogue (<https://code.google.com/archive/p/orthagogue/>), OrthoFinder (<https://github.com/davidemms/OrthoFinder)>)
4. Protein sequences: extract orthologous protein sequences from the genbank files with the script provided (<https://github.com/andreiprodan/mask-publication>) (details on usage provided in the README file of the repository)
5. Alignment: for each fasta file with orthologous sequences, generate a multiple sequence alignment (for example with muscle (<http://www.drive5.com/muscle>): muscle -in example.aln -out example.aln)
6. HMMs: for each alignment, generate an HMM with hmmer (<http://hmmer.org/>):

hmmbuild example.hmm example.aln (note that the HMMs generated as described in the paper are provided in the file “S1\_Dataset.hmm” at <https://github.com/andreiprodan/mask-publication>)

1. Scoring: scan protein sequence fasta files with these HMMs with hmmer (“hmmsearch --tblout output\_table.txt example.hmm sequences.fasta")
2. Visualization: generate a heatmap of aggregated hmm output tables, for example with the R script provided in the git repository (<https://github.com/andreiprodan/mask-publication>) in the file “Correlation\_Heatmap.html” in the R\_script folder