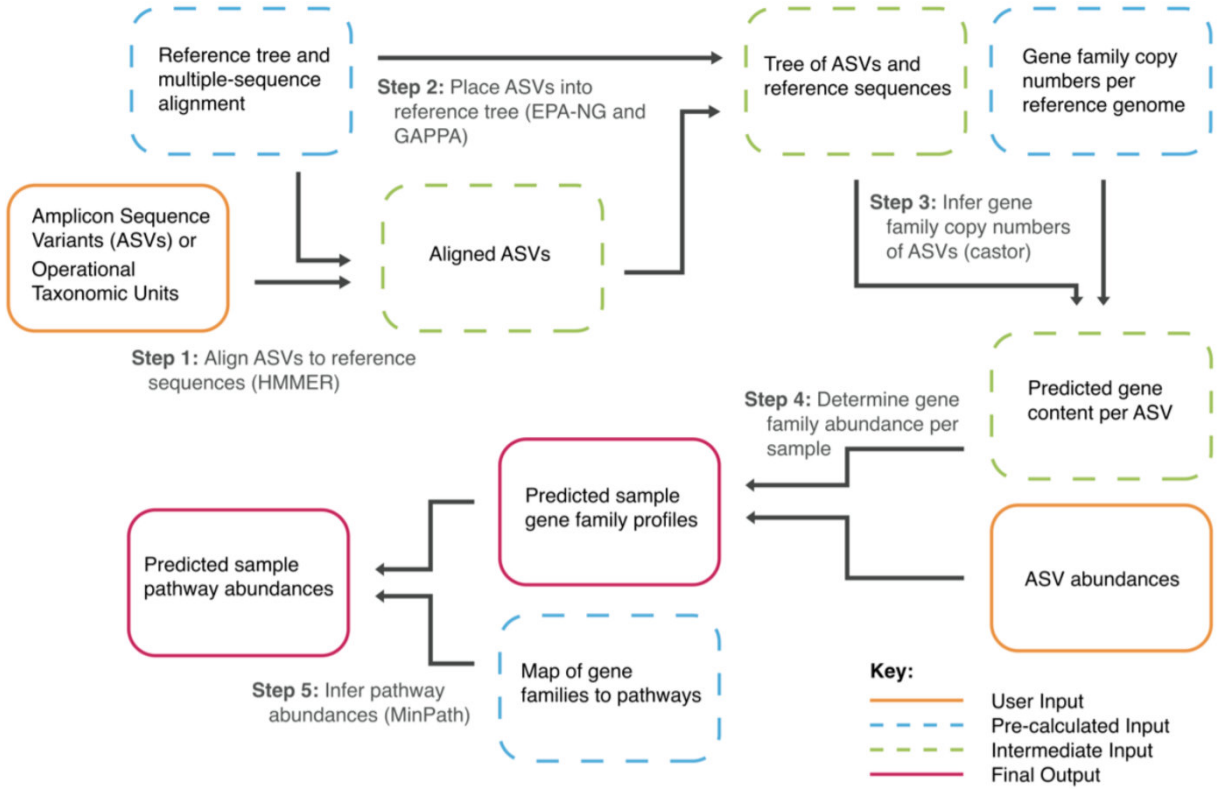
PICRUSt2

Phylogenetic Investigation of Communities by Reconstruction of Unobserved States, 仅基于marker基因序列(例如16S rRNA)预测功能性丰度.



The PICRUSt2 method now consists of phylogenetic placement, hidden-state-prediction, and sample-wise gene abundance tabulation. ASV sequences and abundances are taken as input, and stratified gene family and pathway abundances are output. All necessary reference tree and trait databases for the default workflow are included in the PICRUSt2 implementation.

'Function'一般指的是基因家族, 例如KEGG orthologs和Enzyme Classification numbers, 但是预测可以针对任何实验. 类似, 典型地基于16S rRNA基因序列数据的预测, 同时也可以根据其他标志基因.

Key Limitations

PICRUSt2局限性, 主要和预测相关, 受限于现存的参考基因组的基因组成. 评估该局限是否影响数据预测的方法是和 [functional profiles from metagenomes on a subset of samples] [https://github.com/picrust/picrust2/wiki/Validation-with-paired-metagenomes] 比较.

- **预测受限于被研究序列如何能放置到参考树状结构中(reference tree).** 默认采用`EPA-NG`来放置研究序列, 该过程会考虑多个因素. 最重要的, 基于其他同时被放置的序列, 该放置可能不会都能重复分析得到. 实际过程中, 每个样本的预测结果倾向于高度类似, 但是也会存在重大的差异, 尤其是当解释是根据单个ASV(amplicon sequence variation)或function.
- **给定样本的准确性将会严重依靠可获得的合适的参考基因组.** 可以通过计算per-ASV 和 sample - weighted nearest-sequenced taxon index(NSTI) values 来部分评估该问题, 将得到大概的评估关于ASVs能够被参考数据所代表的程度. 16S rRNA gene sequences do not typically enable resolution of strain variation within a species. 但是原核物种的基因组成可通过水平转移从关系

较远的物种获得外源基因, 因此应小心对待该预测结果.

- **一个相关问题是, 某一环境可被参考基因组很好的代表, 而其他环境不行.** 例如, PICRUSt2针对来自人体的16S 序列的执行情况要好于来自牛瘤胃的 16S 序列, 即使实际的16S序列非常相似. 原因是, 许多重要的瘤胃特异性的酶在默认的参考基因组中缺失. 其中一个可行的解决方法就是构建一个定制一个所感兴趣环境的参考基因组来预测.
- **默认, 若输入序列的NSTI值超过2, 将会从分析中舍去.** 这可能对一些样本的影响大于其他样本, 该影响应被评估. (i.e. you can determin what proportion of the community relative abundance was excluded per sample, which is typically extremely little)
- **PICURSt2仅能预测存在于输入功能表格中的基因.** 对应为KEGG orthologs 和 Enzyme Classification numbers. 尽管这些基因家族有用, 但是它们仅代表metagenomes中的一小部分遗传变化, 也可能被错误注释.
- **PICRUSt2 比对到高水平的功能, 例如pathways, 是完全根据使用的比对文件(entirely dependent on the mapping file used).** 因此, 任何pathway注释中的缺失或不准确性都将存在. 例如, 许多KEGG orhologs 列为参与'Human Diseases'相关的通路. 在许多情况中, 简单的因为细胞包含了(distant) 在哺乳动物pathways中具有重要作用的酶的同源物. 因此, KEGG pathway的注释需要仔细核查, 确保和当前生物系统相一致同时过滤掉不相关的pathways.
- **预测结果局限于使用引物扩增到的序列信息, 不能代表整个metagenome.** 若所使用引物扩增某taxa效率低于其他, 那么该taxa对function预测的功效将会低于其他taxa.

Tutorial (v2.3.0 beta)

chemerin_16S.zip: http://kronos.pharmacology.dal.ca/public_files/tutorial_datasets/picrust2_tutorial_files/chemerin_16S.zip