Instruction

对于测序基因组进行 KEGG（Kyoto Encyclopedia of Genes and Genomes）和COG（clusters of orthologous groups，对直系同源基因进行聚类）功能注释，基本成为基因组注释的标配内容， 特别是微生物基因组基因注释，其基因功能注释逻辑基础是 直系同源基因具有相同的功能，最经典的鉴定直系同源基因策略是 BBH（bi-directional best hit）策略，但是通常最直接的直系同源基因很难鉴定，而对同源基因进行聚类并定义一个簇会是更好的策略：每一个簇会包含直系同源基因（伴随物种形成事件出现）和旁系同源基因（伴随拷贝事件出现），每一簇共享同一个功能， KO(KEGG Orholog), COG, eggNOG 等都是基于聚类的方式定义簇，并对簇进行注释。

区分直系同源非常适合对蛋白功能的推导。来自后代的Orthologous基因倾向于稳定地和始祖功能保持一致，而来自相同基因复制而产生的Paralogous基因则没那么保守。因此，可以将通过模式生物获得功能信息传递给其他难以获得的功能的生物个体。目前常用的软件工具有使用blast序列相似性比对的Blast2GO，RAST，或者通过序列profile搜索，根据序列同源性推导功能信息的。

Orthologs指来自不同物种的由垂直家系(物种进化)而来的蛋白，并且典型地保留与原始蛋白有相同的功能；Paralogs是那些在一定物种中的来源于基因复制的蛋白，可能会进化出新的与原来有关的功能。COG注释作用：1. 通过已知蛋白对未知序列进行功能注释； 2. 通过查看指定的COG编号对应的protein数目，存在及缺失，从而能推导特定的代谢途径是否存在； 3. 每个COG编号是一类蛋白，将query序列和比对上的COG编号的proteins进行多序列比对，能确定保守位点，分析其进化关系。

eggNOG(evolutionary genealogy of genes: Non-supervised Orthologous Groups(OGs))，从旁系同源gene中区分直系同源基因，用于注释大量序列，主要针对基因组/宏基因组转录后的基因编码区域，和转录的数据。

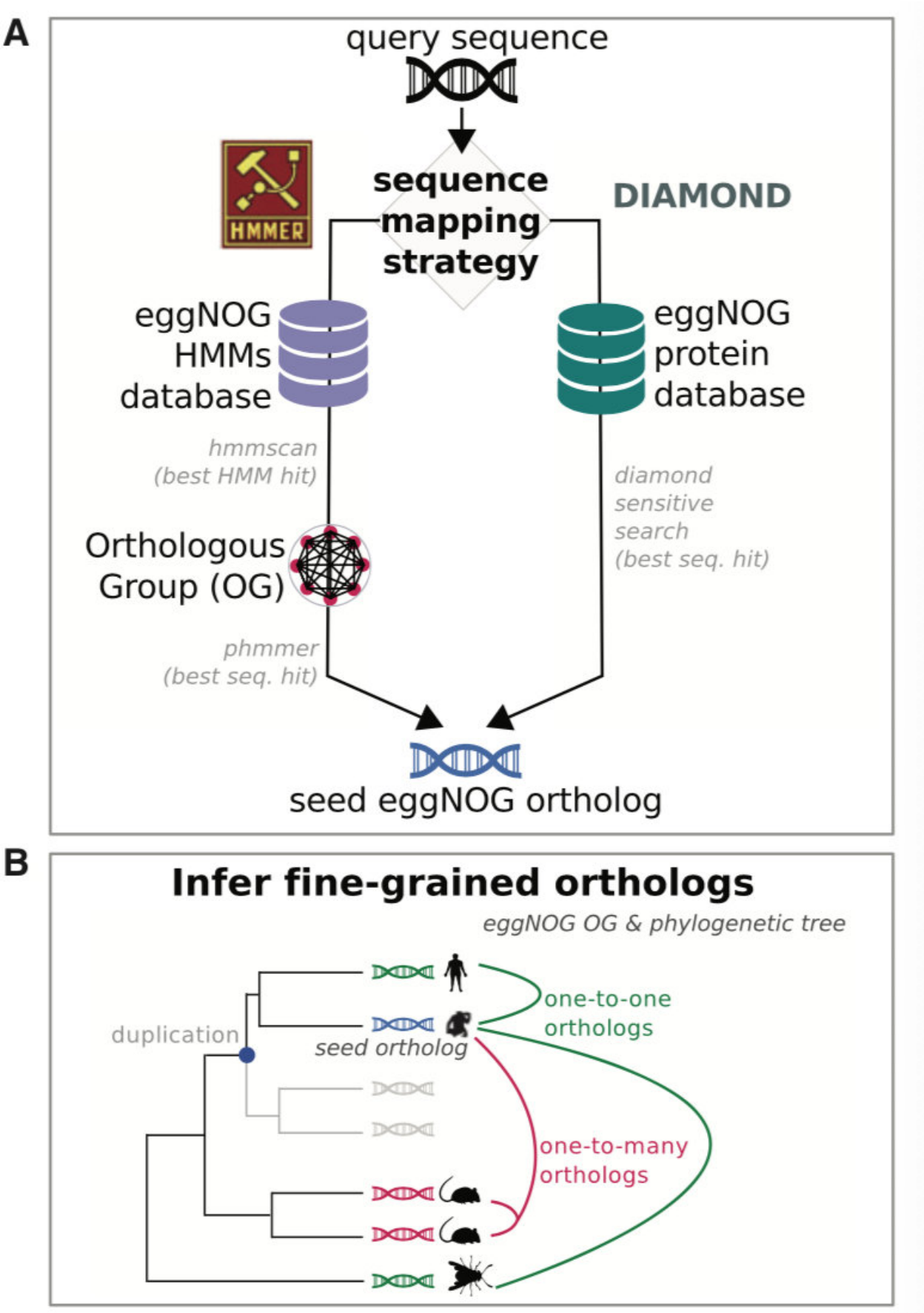
Approaches

1. Sequence Mapping

- HMMER3，使用序列profile的方式查询注释，每个HMM匹配都会注释一个eggNOG OG，提供第一层的功能注释；接着，使用phmmer对蛋白序列再次搜索，这次搜索范围为最佳匹配的HMM eggNOG蛋白，最终获得最佳匹配序列，保存为query's seed ortholog，用于查询orthologs。
- 更快的搜索方式为DIAMOND，直接在所有eggNOG蛋白序列中搜索best seed ortholog。推荐用于较大数据，例如宏基因组，基因组测序。但敏感度不如HMM方法

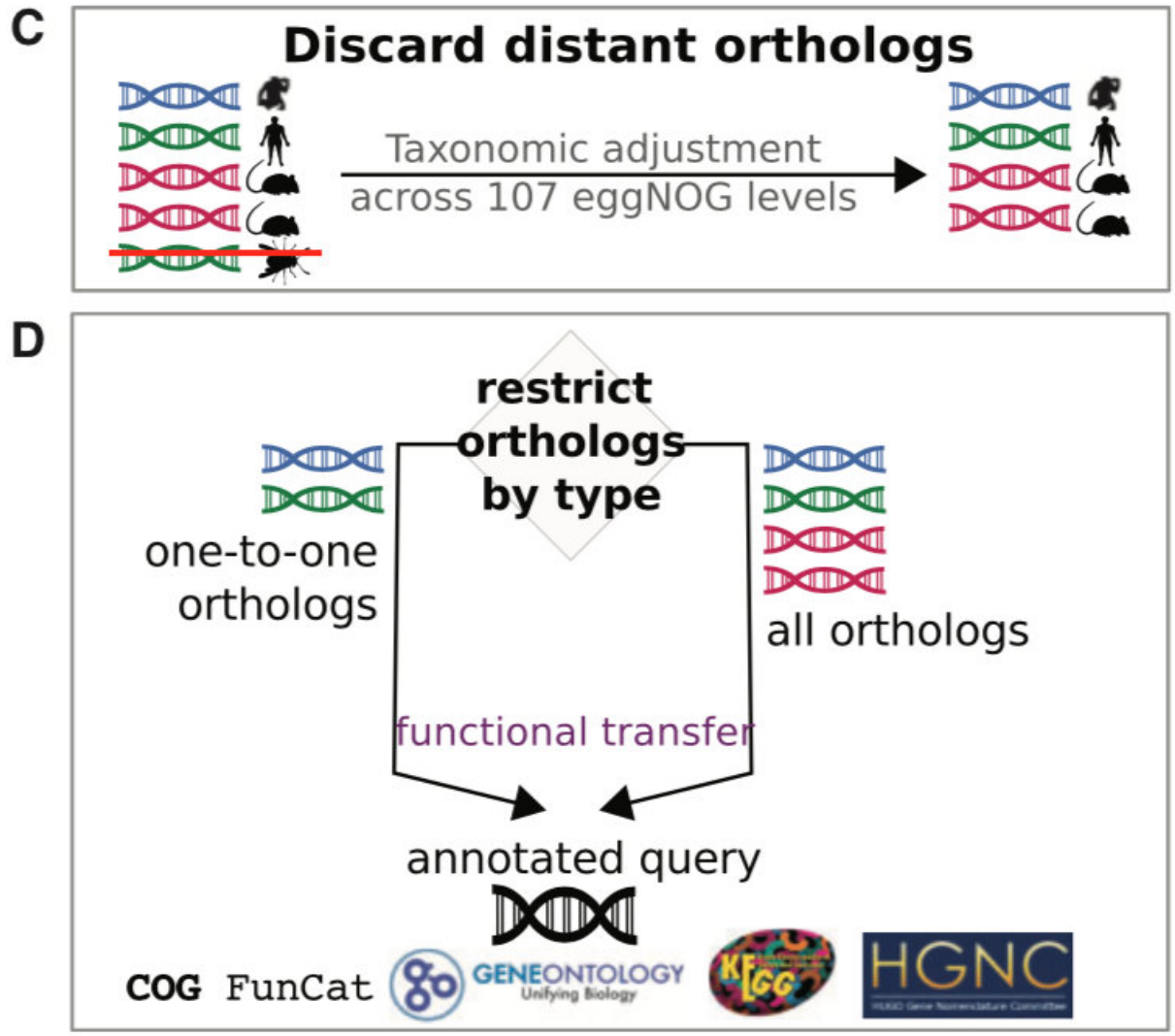
2. Otrhology Assignent

- 针对第一得到的seed ortholog，在eggNOG phylogenetic trees数据中查询获得一系列fine-grained orthology assignments，同时可以通过bit-score和E-value阈值避免获得非显著性结果



3. Functional Annotaion

- 获得query蛋白的功能信息，默认情况下，功能性描述会根据query蛋白序列来限制到最近的类别中。该信息包括最近的GO terms, KEGG和COG功能分类



Usage

-i： 输入fasta文件

--translate： 假设输入为gene序列

-m: hmmer/diamond，默认为hmmer

—output_dir：输出文件夹

-o：输出文件前缀

--seed_ortholog_evalue： 设置搜索E-value，默认0.01

--seed_ortholog_score：设置score，默认60

--target_orthologs：{one2one,many2one,one2many,many2many,all}

--cpu：指定cpu数目

```
emapper.py -m diamond --cpu 2 -i 36170_prodigal.faa -o 36170_prodigal_diamond -m diamond --target_orthologs all -d bact
```

diamond软件版本冲突问题:

- https://github.com/eggnogdb/eggnog-mapper/issues/48
- eggnog使用diamond构建大蛋白数据库为 v0.8.38，当前版本已经为v0.9xx，所以，搜索过程过程中无法调用diamond，解决方案，安装对应版本
- 或者直接下载构建diamond搜索数据库的蛋白序列：http://eggnogdb.embl.de/download/eggnog_4.5/eggnog-mapper-data/eggnog4.clustered_proteins.fa.gz，本地构建，更改名称即可