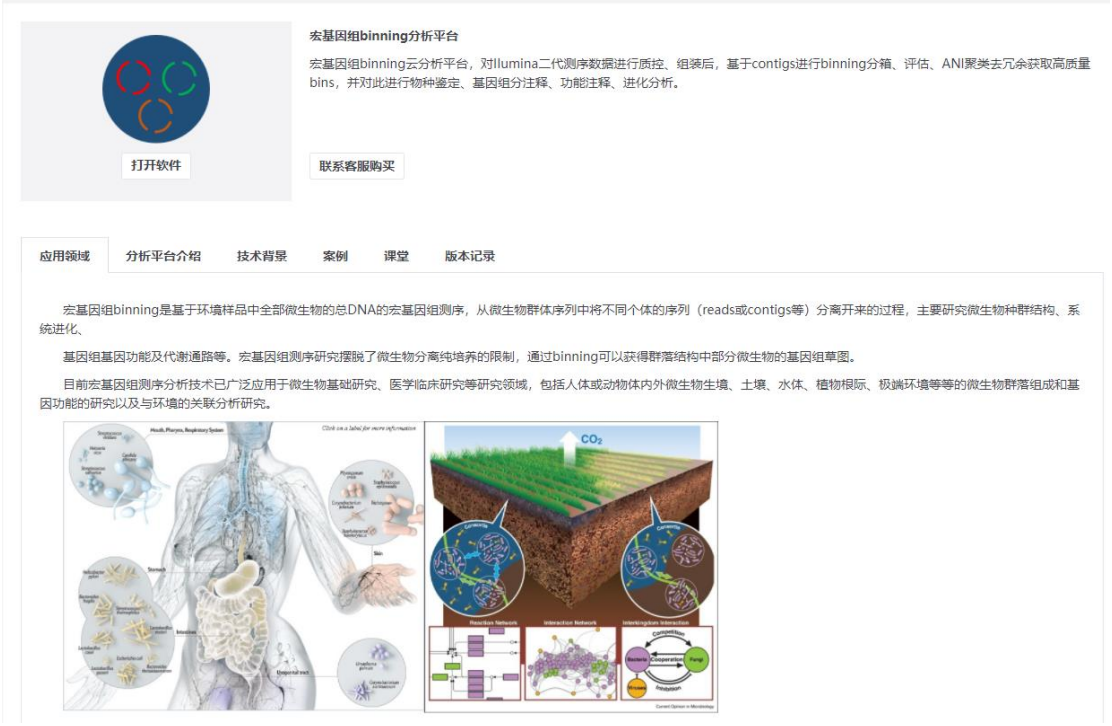
宏基因组 binning 分析流程参数



1 背景介绍

宏基因组 binning：宏基因组是环境中全部微小生物遗传物质的总和，包含了可培养的和不可培养的微生物的基因。宏基因组研究可以获得群落中全部的物种信息和功能信息，宏基因组 Binning 就是对序列进行聚类、分装，是根据基因组特征以及组装信息等将属于不同基因组的序列分离开来的过程。通过 binning 得到的 bins，很可能是实验室无法纯培养的未知微生物，对其进行组学分析很有意义。

2 基本分析结果

2.1 宏基因组分箱

使用软件 MetaBat2[1]、Maxbin2[2]或 CONOCOCT[3] 进行宏基因组分箱。

默认 MetaBat2，老师可自行选择

MetaBat2：v2.12，默认参数

Maxbin2：v2.2.6，默认参数

Conococt：v1.0.0

使用软件 DAS_Tool[4]（v1.1.2 --search_engine diamond --write_bins 1 --score_threshold 0）进行不同宏基因组分箱软件结果整合。

采用 checkM[5] （v1.1.3，默认参数）软件对整合后 bins 进行评估。高质量 bins 评估（默认完整度≥80　污染度≤10 ）。

采用 drep[6] （v3.0.0，　-sa 0.95　-comp 50　-nc 10，　-comp 与用户选择一致）软件对高质量 bins 进行去冗余。

4.2 高质量 bins 组分分析

4.2.1 编码基因预测

采用 MetaGeneMark[7]软件（http://exon.gatech.edu/meta_gmhmmp.cgi，Version 3.26），使用默认参数（参数-A -D -f G）来识别基因组中的编码区域。

4.2.2 重复序列预测

相比于真核生物，原核生物具有更紧凑的基因组。编码蛋白或者 RNA 分子的区域一般会占到基因组的 85-90%。而剩余的基因组区域中的绝大多数被各种调控性区域所占据，所以原核生物基因组中重复序列含量极少。

采用 RepeatMasker[8] （v4.0.5 -engine wublast ）软件，将基因组与已知重复序列数据库(Repbase)进行比对（默认参数）来搜索基因组中的重复序列。

4.3 功能注释分析

4.3.1 Nr 数据库注释

Nr[9]数据库的全称是 Non-Redundant Protein Database,是一个非冗余的蛋白质数据库,由 NCBI 创建并维护,该数据库含有全面的蛋白序列和注释信息，而且在注释信息中存在相应的物种信息。与使用其他数据库相比，使用 Nr 数据库一般能使基因组中更多的基因具有注释

信息，即具有最高的注释比例。但该数据库中很多蛋白序列和注释信息未经过验证,可靠性有待提高。

具体注释时，通过将非冗余基因的蛋白序列和 Nr 数据库进行 BLAST 比对（diamond v0.9.29.130 比对筛选阈值 E-value 1e-5），找到在 Nr 数据库中最相似的序列，该序列对应的注释信息即为对应测序基因组基因的注释信息。

4.3.2 kegg 功能注释

kegg[10]（Kyoto Encyclopedia of Genes and Genomes）是收集了生物的基因组、通路和化合物信息的综合性的数据库。该数据库对收录的序列进行分析聚类分析，形成直系同源蛋白群，对于不同的蛋白群指定不同的 KO 序列。根据已发表文献，kegg 人工绘制了大量的生物过程图（如代谢途径、信号传导途径等），在图中的特殊的矩形框或线条对应发挥相应功能的某种 KO 序号的蛋白群，通过这些生物过程图更能直观地体现生物的生命过程。

kegg（Kyoto Encyclopedia of Genes and Genomes）是收集基因组、生物通路、疾病、药物和化学物质的数据库。在 kegg 中有一个“专有名词”KO(kegg Orthology)，它是蛋白质(酶)的一个分类体系，将序列高度相似并且在同一条通路上有相似功能的蛋白质序列归为一组，然后打上 KO 标签，对应每个 KO 标签添加相应的功能注释信息。而这些 KO 标签对应到基因组在 kegg 中对于许多代谢通路或者细胞过程。

具体注释时,通过将非冗余基因的蛋白序列和 kegg 数据库中收录的蛋白序列进行 BLAST 比对（diamond v0.9.29 比对筛选阈值 E-value 1e-5），找到在 kegg 数据库中最相似的序列，该序列的注释信息、对应的 KO 号、KO 对应的通路中的位置即为测序基因组中对应的基因的注释信息、KO 号以及在生物过程通路中的位置该序列对应具有 KO 序号。

4.3.3 eggNOG 数据库注释

eggNOG[11]是收录生物直系同源基因簇的数据库，是在 COG 数据库的基础上的持续更新。构成每个直系同源基因簇(Cluster of Orthologous Groups)的蛋白都是被假定为来自于一个祖先蛋白，具有相同的功能。该数据库是通过比较完整基因组的蛋白序列而产生的。该数据库常被用来对新测序基因组的基因进行分类和注释。

具体注释时,通过将非冗余基因的蛋白序列和 eggNOG 数据库进行 BLAST 比对(diamond v0.9.29 比对筛选阈值 E-value 1e-5），找到在 eggNOG 数据库中最相似的序列，该序列对应的注释信息和分类信息即为对应测序基因组基因的注释信息和分类信息。

4.3.4 CAZy 数据库注释

CAZy[12]（Carbohydrate-active enzymes database）是收录碳水化合物活性酶的数据库，该数据库由已发表文献和相关蛋白的收集和分类形成，并由专家精心维护。该数据库中蛋白分为五大类功能类别:糖苷水解酶(glycoside hydrolases,GHs)、糖基转移酶(glycosyltransferases,GTs）、多糖裂解酶（polysaccharide lyases，PLs）、碳水化合物酯酶（carbohydrate esterases，CEs)和非催化的结合碳水化合物的功能域（CBMs）。上述每一种大的类别中都含有许多不同的家族。不过每个家族中都是由蛋白序列组成的，不能突出该家族共有的特征性的功能域，

而细胞内的碳水化合物活性酶一般具有许多的结构域。dbCAN 具体分析 CAZy 数据库收集的每一个家族，并对每一个家族构建该家族特征性结构域的隐马尔可夫模型。软件 hmmer 使用这些隐马尔可夫模型可以识别出属于特定家族的保守功能域。

在具体注释时，使用 hmmer 软件（版本 3.0）对于通过将非冗余基因的蛋白序列分别与 CAZy 数据库的每一个家族的隐马尔可夫模型比对（比对参数 默认，默认筛选阈值"if alignment > 80aa, use E-value < 1e-5, otherwise use E-value < 1e-3; covered fraction of HMM > 0.3"），找出所有满足过滤阈值的家族，这样可以找出基因组中的碳水化合物活性酶，也可以分析出它们的含有几个保守碳水化合物相关功能域。

4.3.5 CARD 数据库注释

CARD[13](Comprehensive Antibiotic Research Database)是一个持续更新的抗生素抗性相关信息的数据库。CARD 包含描述抗生素和它们的靶标，涉及抗生素抗性基因、相关蛋白、抗生素抗性机制信息。在 CARD 的核心是一个高度开发的抗生素抗性本体（Antibiotic Resistance Ontology (ARO)），用于抗生素抗性基因数据的分类。

在注释时，使用 CARD 数据库中的工具软件 rgi（版本 4.2.2 默认 Perfect, Strict 算法）将非冗余基因的蛋白序列分别数据库进行比对，找出比对上的数据库中对应序列，并得到对应的抗性基因和抗性相关信息。

5 物种鉴定分析

5.1 物种鉴定

利用 GTDB-Tk[14] （v1.2.0）软件根据基因组数据库分类（release 95）GTDB 为 bins 进行物种分类注释，GTDB 基于大量基因组的系统发育分析来构建基因组分类学研究对微生物进行分类。基于保守的单拷贝基因构建的进化树数据库，将 binning 数据和数据库进行比对，看定位到哪个物种进化区间内。

5.2 物种进化分析

进化树在生物学中用来表示物种之间的进化关系。根据各类生物间的亲缘关系的远近，把各类生物安置在有分枝的树状的图表上，简明地表示生物的进化历程和亲缘关系（进化树绘图包：R 3.6.1 ggtree）。进化树由结点（node）和进化分支（branch）组成，每一结点表示一个分类学单元（属、种群、个体等），进化分支定义了分类单元（祖先与后代）之间的关系。

通过鉴定单拷贝标记基因、多序列比对，构建进化树。样品层次聚类树图：样品越靠近，枝长越短，说明两个样品的物种组成越相似。

参考文献

1. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome

assemblies. PeerJ. 2019 Jul 26;7:e7359. doi: 10.7717/peerj.7359. PMID: 31388474; PMCID: PMC6662567.

2. Wu Y W , Tang Y H , Tringe S G , et al. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm[J]. Microbiome, 2014, 2.
3. Johannes Alneberg, Brynjar Smári Bjarnason, Ino de Bruijn, Melanie Schirmer, Joshua Quick, Umer Z Ijaz, Leo Lahti, Nicholas J Loman, Anders F Andersson & Christopher Quince. 2014. Binning metagenomic contigs by coverage and composition. Nature Methods, doi: 10.1038/nmeth.3103
4. Sieber C M K , Probst A J , Sharrar A , et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy[J]. Nature Microbiology, 2018, 3(7).
5. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Research, 25: 1043–1055.
6. Olm M R , Brown C T , Brooks B , et al. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication[J]. Isme Journal, 2017.
7. Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. Nucleic acids research. 2010;38(12):132–132.
8. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences[J]. Current Protocols in Bioinformatics, 2009: 4.10. 1-4.10. 14.
9. Deng Y Y, Li J Q, Wu S F, et al. Integrated nr database in protein annotation system and its localization[J]. Comput Eng, 2006, 32(5): 71-74.
10. Kanehisa M, Goto S, Kawashima S, et al. The kegg resource for deciphering the genome[J]. Nucleic acids research, 2004, 32(suppl 1): D277-D280.
11. Powell S , Forslund K , Szklarczyk D , et al. eggNOG v4.0: nested orthology inference across 3686 organisms[J]. Nucleic Acids Research, 2014, 42(Database issue):231-9.
12. Cantarel B L, Coutinho P M, Rancurel C, et al. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics[J]. Nucleic acids research, 2009, 37(suppl 1): D233-D238.
13. Jia B, Raphenya A R, Alcock B, et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database[J]. Nucleic acids research, 2016: gkw1004.
14. Chaumeil PA, et al. 2019. GTDB-Tk: A toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics, btz848.