百易汇能
微生物个性化分析手册 v1.0

撰写人：张兴国

2022 年 01 月 22 日

武汉百易汇能生物科技有限公司

目　录

一 分析方法与结果.....1
1.1 cgMLST 分析.....1
1.2 K-locus 分析.....2
1.3 基因和基因组的突变分析.....3
1.4 wzi 基因分析.....4
1.5 全基因组 ANI 分析.....5
1.6 基因进化分析.....7
二 分析方法.....8
2.1 软件及数据库.....8
2.2 数据库的链接.....8
三 帮助文档.....9
3.1 结果文件及数据说明.....9
3.2 常见文件格式.....9
3.2.1 fastq 格式文件.....9
3.2.2 fasta 格式文件.....9

一 分析方法与结果

1.1 cgMLST 分析

从 Ridom cgMLST 数据库下载 Klebsiella pneumoniae/variicola/quasipneumoniae cgMLST（ https://www.cgmlst.org/ncs/schema/2187931 ）数据集，使用 chewBBACA 对数据库进行处理和转换获得 2358 等位基因图谱。使用 chewBBACA 软件结合处理好的数据库，对测序获的基因组进行 cgMLST 分析，相关结果如下（其中样本 158 与 234 的汉明距离为 0；样本 188 与 190 的汉明距离为 0；样本 206 与 207 的汉明距离为 0）。

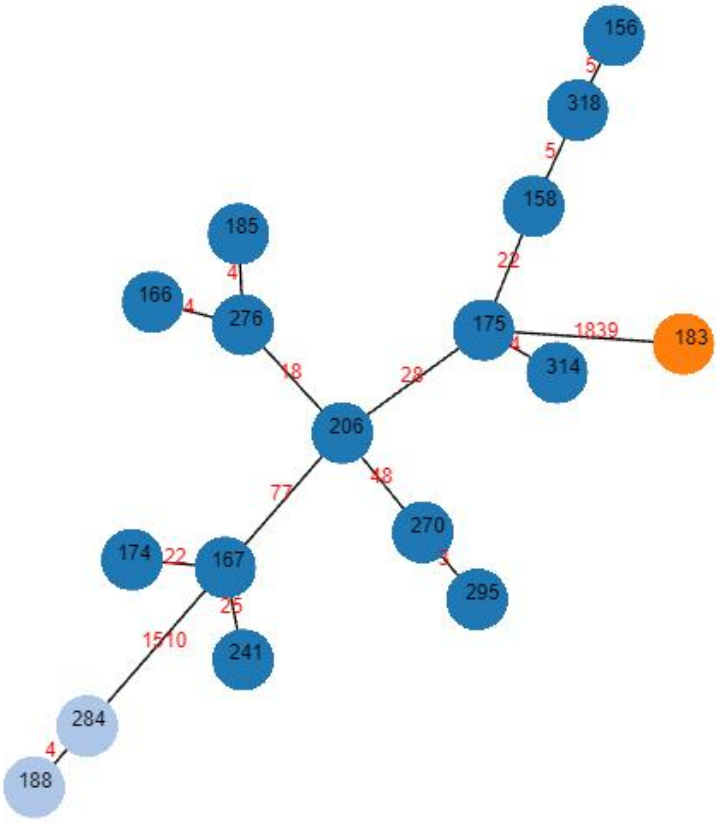

图 1.1 cgMLST 分析构建的最小生成树

结果说明：图中每个圈代表样本的名字，两个圈之间连线上的数字表示两个菌株之间的等位基因差异大小（汉明距离）。不同的多位点序列类型以圆圈颜色区分。等位基因的功能介绍见（https://www.cgmlst.org/ncs/schema/2187931/locus/）。图的美化和调整可以用结果中的 cgMLST.tsv 文件导入 PHYLOViZ（https://online.phyloviz.net/index）即可。（如果两个样本之间的汉明距离为 0，则只会显示其中一个样本）

分析结果文件说明：

cgMLST.tsv：具有 cgMLST 等位基因图谱的矩阵；

distances.tab：各样本之间的汉明距离矩阵；

mdata_stats.tsv：每个基因组缺失的等位基因总数和百分比；

Presence_Absence.tsv：等位基因存在和缺失矩阵（分别为 1 或 0 ）

cgMLSTFinder：https://cge.cbs.dtu.dk/services/cgMLSTFinder/ （分析的物种有限，为在线版本）

1.2 K-locus 分析

使用 Kaptive（参数： --k_refs Klebsiella_k_locus_primary_reference.gbk --allelic_typing wzi_wzc_db.fasta ）对基因组进行分析，具体几个见附件 Klocus.results.tsv，部分结果如下。

表 1.1 各样本的分型

Sample	Locus	Type
156	KL64	K64
158	KL64	K64
166	KL64	K64
167	KL47	K47
174	KL47	K47
175	KL64	K64
183	KL102	unknown (KL102)
185	KL64	K64
188	KL16	K16
190	KL16	K16
206	KL64	K64
207	KL64	K64
234	KL64	K64
241	KL47	K47
270	KL64	K64
276	KL64	K64
284	KL16	K16
295	KL64	K64
314	KL64	K64
318	KL64	K64

1.3 基因和基因组的突变分析

以 MGH 78578 的基因组作为参考（链接：https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/016/305/GCF_000016305.1_ASM1630v1/），使用 bwa 将各样本质控后的 read 与参考基因组进行比对。使用 snippy 流程进行 snp 分析和 snp 注释。根据注释结果找到对应关注的基因，其中基因 mgrB、PhoP、PmrB 在 20 个样本中均未发现突变（基因 CrrA 在参考基因组中未发现，是否有其他名字），突变基因及突变位点数目统计见下表。

表 1.2 各样本对应基因的突变数目统计

Sample	CrrB	PhoQ	PmrA
156	0	3	4
158	2	3	4
166	2	3	4
167	1	3	5
174	2	4	4
175	2	3	4
183	4	3	7
185	1	3	5
188	5	1	4
190	5	1	4
206	2	3	4
207	2	3	4
234	2	3	4
241	2	3	4
270	2	3	5
276	1	3	4
284	5	3	4
295	2	3	5
314	2	3	4
318	2	3	4

结果说明：结果文件中有单个样本基因组的突变结果文件*.snps.gff、*.snps.html 和*.snps.vcf（三种不同的格式）。建议看*.snps.gff 文件，里面含有突变的基因、突变的位点、突变的类型和突变位点的深度等信息。

*.gc_depth.png：为各样本 reads 在参考基因组上的深度图；

*.length_gc.xls：为样本 reads 在参考基因组上的深度统计；

core_snp.vcf：各样本 snp 汇总结果；

core_snp.aln：为各样本 snp 对齐后的序列文件，可以用于绘制 snp 进化树；

merge.vcf：为样本 snp 和其他突变类型的汇总结果；

stat_map_coverage.tsv：为个样本的覆盖情况统计。

1.4 wzi 基因分析

从组装的基因组中提取 wzi 基因（见：all.wzi.fasta），使用 MAFFT 比对对齐，使用 Gblocks 提取保守位点和展示对齐结果（wzi.aln.fasta.gb.htm），使用 raxml 构建进化树。使用 pyani 进行 ani 分析。相关结果如下：

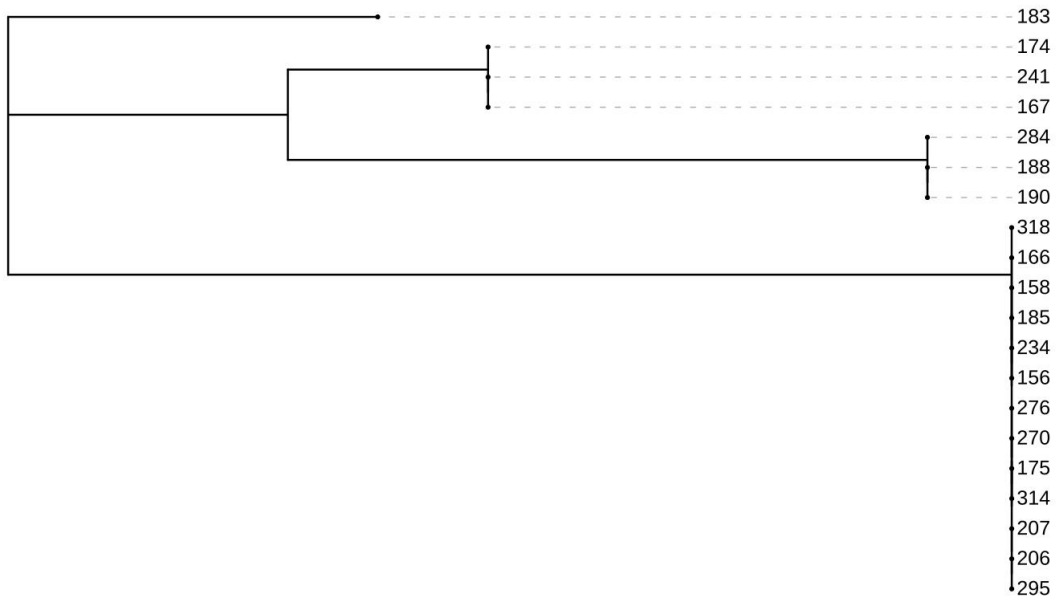

图 1.2 wzi 基因进化树

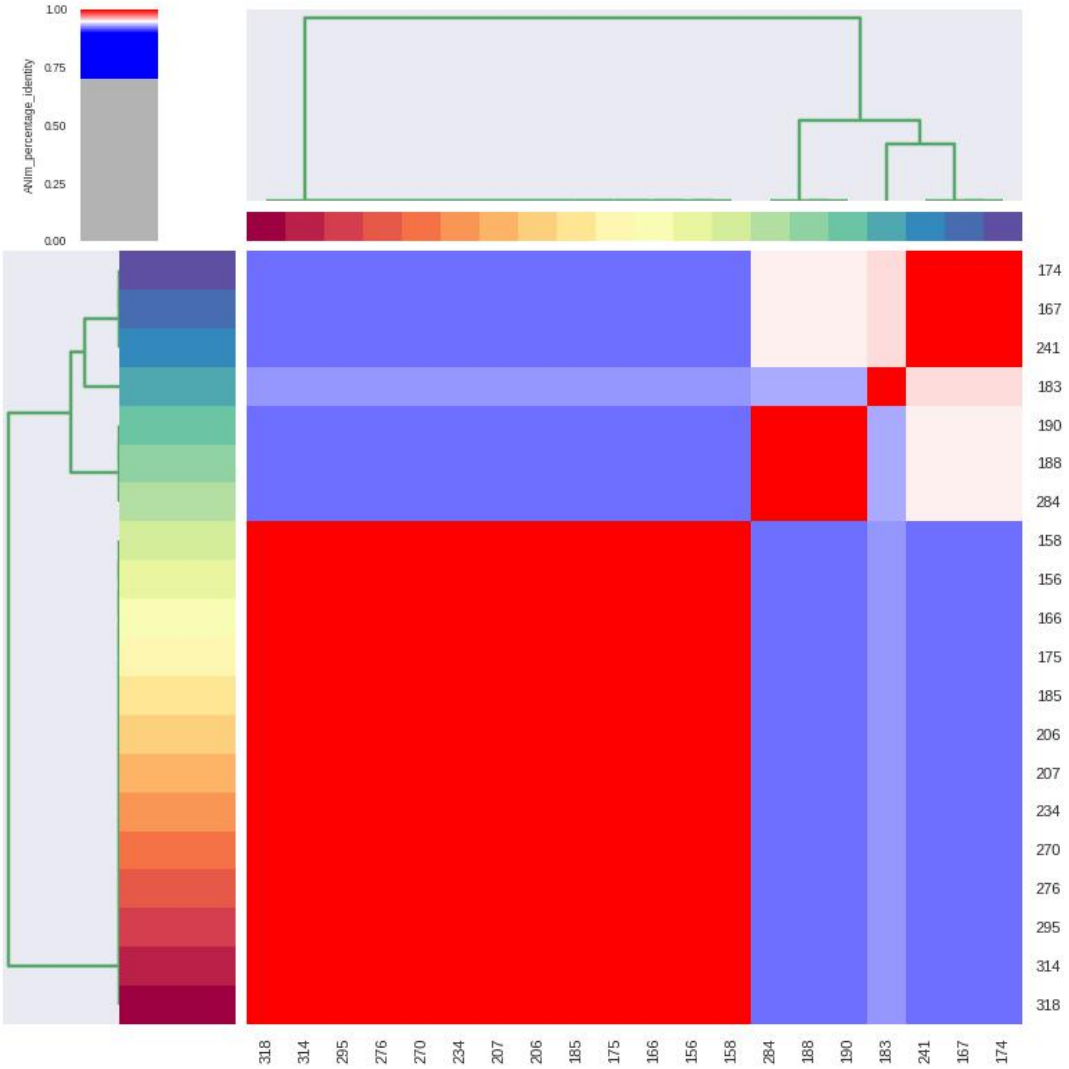

图 1.3 wzi 基因 ani 分析结果

结果说明：wzi 基因相似度矩阵见附件 ANIm_percentage_identity.tab；进化树的美化和调整可以使用 Itol（https://itol.embl.de/），将文件 wzi.tree 导入即可。

1.5 全基因组 ANI 分析

从 NCBI 上下载近源物种的基因组序列，使用 nucmer 将基因进行比较，使用 pyani 绘制各基因组之间的相似度热图,见下图。

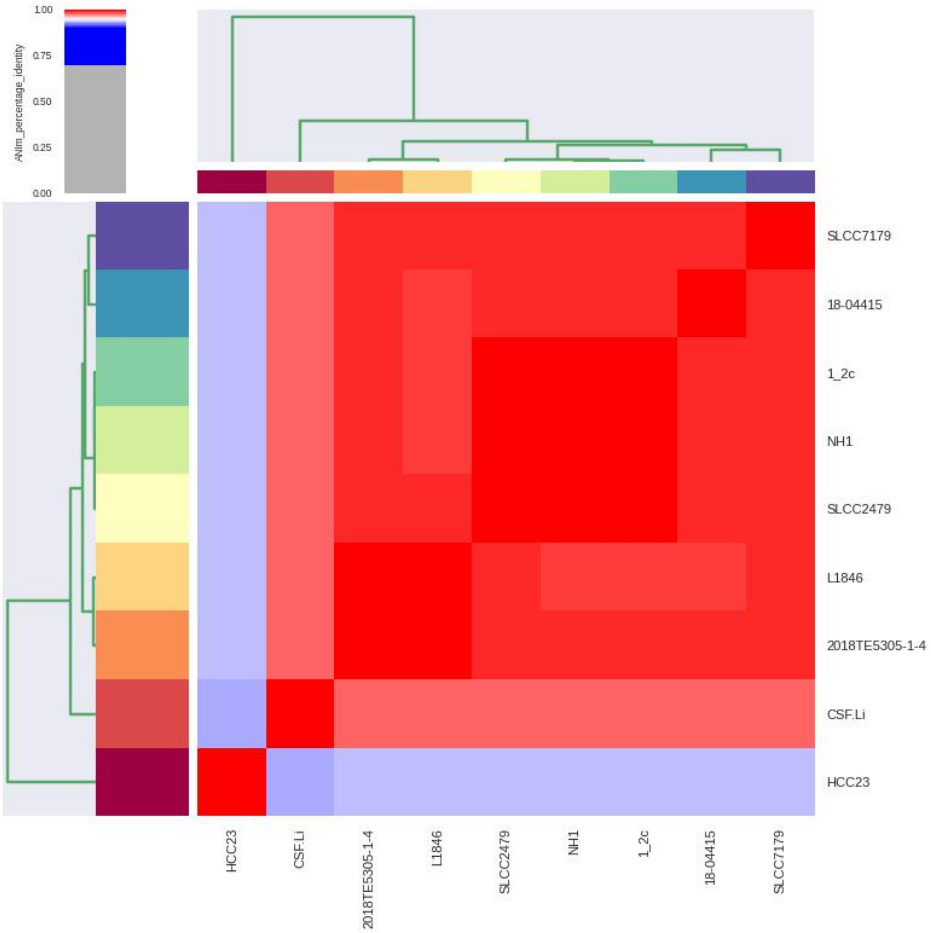

图 1.1 各样本全基因组相似度热图

1.6 基因进化分析

从 NCBI 上下载近源物种的基因组序列，使用 OrthoFinder 获得共有的单拷贝蛋白。使用 MAFFT 对蛋白的序列进行比对，使用 Gblocks 提取共有的保守序列，使用 RAxML 构建进化树。进化树的美化可以使用 Itol 和 FigTree，进化树的绘图结果见下图。

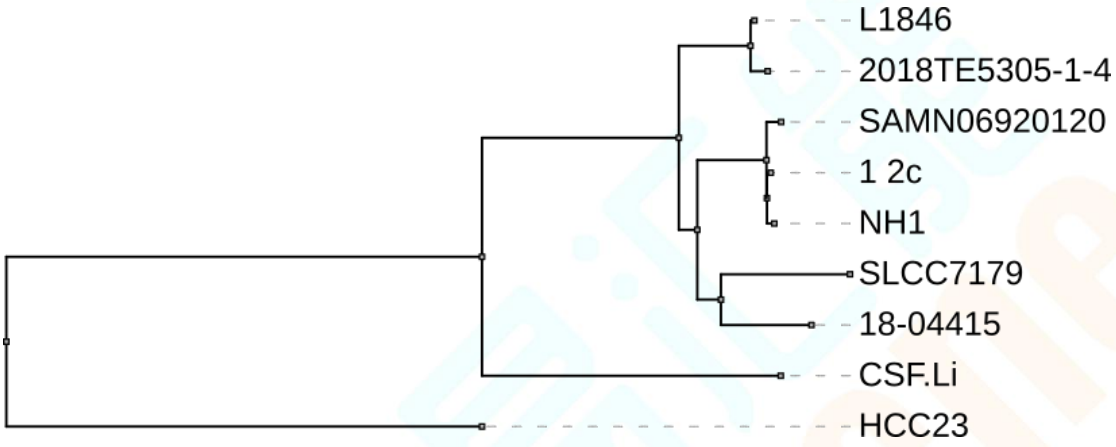

图 1.2 进化树图

二 分析方法

2.1 软件及数据库

生物信息分析所用的软件及数据库信息如下。

表 2.1 分析所用软件信息

Software	Version	Website
chewBBACA	v2.8.5	https://github.com/B-UMMI/chewBBACA
PHYLOViZ	-	https://online.phyloviz.net/index
cgmlst-dists	v0.4.0	https://github.com/tseemann/cgmlst-dists
Kaptive	v2.0.0	https://github.com/katholt/Kaptive
bwa	v0.7.17	https://github.com/lh3/bwa
snippy	v4.5.9	https://github.com/tseemann/snippy
OrthoFinder	v2.4.0	https://github.com/davidemms/OrthoFinder
OrthoMCL	-	http://orthomcl.org/orthomcl/
MAFFT	v7.313	https://www.ebi.ac.uk/Tools/msa/mafft/
Gblocks	0.91b	http://www.phylogeny.fr/one_task.cgi?task_type=gblocks
RAxML	8.2.10	https://github.com/stamatak/standard-RAxML
blast+	2.7.1+	https://blast.ncbi.nlm.nih.gov/Blast.cgi
Itol	-	https://itol.embl.de/
FigTree	v1.4.3	https://github.com/rambaut/figtree
pyani	v0.2.11	https://github.com/widdowquinn/pyani
nucmer	v3.1	http://mummer.sourceforge.net/

2.2 数据库的链接

生物信息分析所用的数据库信息如下。

表 2.2 分析使用的数据库信息

Database	Version	Website
Ridom cgMLST	-	https://www.cgmlst.org/ncs

三 帮助文档

3.1 结果文件及数据说明

(1) 上传目录中有 Readme.txt 说明，详细介绍了每个文件所代表的内容。上传的结果数据文件多以文本格式为主(fa 文件、txt 文件、detail 文件、xls 文件等)。在 Windows 系统下查看文件，推荐使用 Editplus 或 UltraEdit 作为文本浏览程序，或者使用 PilotEdit Lite 打开超大文本文件。在 Unix 或 Linux 系统下可以浏览较大的文本文件，用 Less 等操作命令可以顺利地查看。

(2) 报告文件含有 SVG 格式的图片文件，SVG 是矢量化的图片文件，可以随意放大而不失真。要查看 SVG 格式的文件，请先安装 SVG 插件。

3.2 常见文件格式

3.2.1 fastq 格式文件

fastq 文件中包含有测序数据的序列信息以及对应的质量值。每条 read 包含 4 行信息，如下：

```
@HWI-ST531R：144：D11RDACXX：4：1101：1212：1946 1：N：0：ATTCCT
ATNATGACTCAAGCGCTTCCTCAGTTTAATGAAGCTAACTTCAATGCTGAGATCGTTGA
CGACATCGAATGGG
+ HWI-ST531R：144：D11RDACXX：4：1101：1212：1946 1：N：0：ATTCCT
?A#AFFDFFHGFFHJJGIJJJIICHIIIIJJGGHIIJJIIJIIJIHGI@FEHIIJBFFHGJJIIHHHDFFFFDCC
CCEDDCDDCDEACC
```

其中第一行以“@”开头，为序列名称；第二行是碱基序列，由大写“ACGTN”组成，N 表示测序未知的碱基；第三行以“+”开头，为序列 ID；第四行是对应序列的测序质量，以 ASCII 值表示，每个字母对应第 2 行每个碱基。

3.2.2 fasta 格式文件

fasta 格式是一种基于文本用于表示核苷酸序列或氨基酸序列的格式。在这种格式中碱基对或氨基酸用单个字母来编码。如下：

```
>tig00001 [topology=circular] [completeness=complete] [location=chromosome]
ATGAAAAAAGAAACAAGTAAGTTTACGAAAAAAGTCATCCTTGGTCAGCTTCTCATCT
GT
ACCAGTTTTTTGGTATGGGCAGGGCAGACTGTTGCAGCAGAGGAGGTATCTCAGGGAG
```

AA
CCAGTAGCAATCAACAATGAAACTCCACCTGTTATTGAAGCAAAACCACTTGTAGAAGCT
ACTCCATCAACTGAAACAGAACCTTCAACTCCAGAAGAAAAACAGGAGGAACCACAAGTA
GCTGGTAAAAATGTGGTAGAAGGGAAAAAGCCAGTAACCAACAGCCAGCATCAGATGGTT
TCTGAATTATCAACGGTCCCTTCTGTTTCTATAAAAAATCCAGATGCAGCTACCGATGGG
AGTGCTTACGGTAGTAGCGACGATGCCAATAGCAATACCAAAATCATTGCTGGTAAAGAA

序列文件的第一行是由大于号">"或分号";"打头的任意文字说明（习惯常用">"作为起始），用于序列标记。从第二行开始为序列本身，只允许使用既定的核苷酸或氨基酸编码符号。