Lpp Book

Lpp

目录

1 数据质量统计

使用 Pacbio 测序平台对 A16R 菌种进行测序。共得到 4 个 smrtcells 进行测序。

1.1 质量统计方法

Pacbio 使用单分子测序技术实现高读长测序。但由于其数据信号捕捉困难，原始数据的质量较低，并且具有高随机性。因此，在数据分析前需要进行严格的数据质量过滤和前处理。处理方法如下：

1. 由于 Pacbio 使用环状文库进行测序，并且其实际读长远远大于文库长度。因此，大部分的原始数据可能被测了多次。因此，首先需要对原始数据进行打断和开环，并根据开环结果对测序数据的单个碱基进行一致性检验，并给出单碱基的质量值。
2. 而后，使用严格的参数将残余的 adapter 进行去除。
3. 最后，根据单碱基质量值，对 reads 中 Q20<0.8 的区域进行过滤和筛除，得到相对高质量的 subreads 序列进入后续分析。

1.2 测序数据质量统计结果

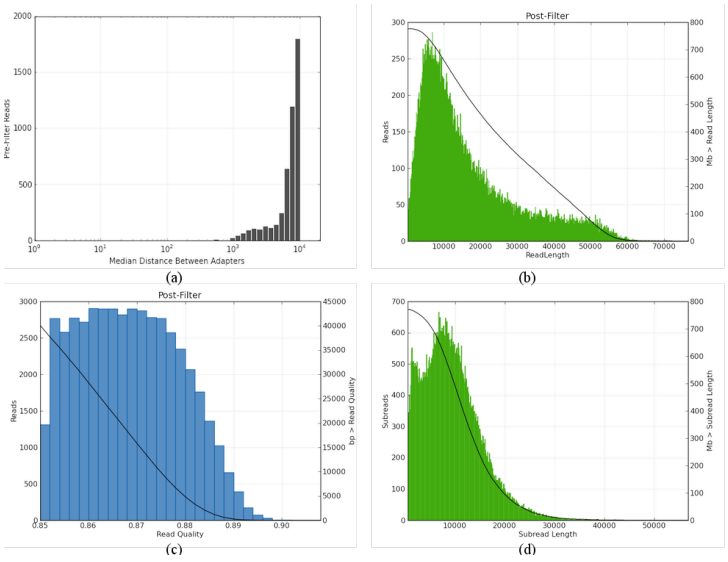

图 1. QC 分析结果

(a) 测序得到的 reads 读长的统计图

(b) 下机原始数据的序列读长统计

(c) 经过数据过滤处理后得到的序列质量和数据量

(d) 经过数据过滤处理后得到的序列读长

1.3 所有测序结果统计表

下表为测序结果的初步统计结果，其中，红色数据为经过数据质控后的可用数据的统计结果，后续分析全部使用这些质控后的数据进行分析。

表 1. 123

11	22	33
aa ee	bb ff	静静地 [1] 我将离开你，请将眼角的泪逝去。漫漫岁月里 dd

[1] 我的天啊啊哪!!!

2 基础分析

2.1 序列组装

使用 HGAP 对获得的测序数据进行组装

2.1.1 组装方法

由于 Pacbio 的测序数据质量较低，并且其错误呈现高随机性，因此，可以使用短的 reads 比对到长的 reads 上，并通过投票的方法对每一个碱基进行清洗，从而得到正确率能够满足分析需求的数据 [1]。

而后，使用经典的 OLC 算法架构的组装工具将经过算法处理后的 reads 进行组装，得到最终的组装结果。

2.1.2 重复序列预测

通过组装结果与已知的转座子序列库进行比对来查找转座子序列。具体方法是，通过 RepeatMasker 软件 [2]（使用 Repbase 数据库 [3]）和 RepeatProteinMasker 软件（使用 RepeatMasker 自带的转座子蛋白库）两种方法来预测转座子；通过 TRF（Tandem Repeat Finder）[4] 软件预测串联重复序列。

表 2. 重复序列分析结果

total.fna.masked	**重复序列 mask 后的序列**
total.fna.out	**重复序列比对后的初始结果**
total.fna.out.gff	**重复序列比对结果的 gff 文件**

3 基因预测和注释分析

使用 Prodigal[5] 对 cds 进行预测，使用 SignalP 对信号肽进行预测，使用 infernal[6]+Rnammer[7] 对 tRNA、rRNA 和 ncRNA 进行预测。使用 Phage Finder[8] 对可能的噬菌体前体区域进行预测，使用 PILE-CR[9] 预测可能存在的 CRISPR 序列。使用 RFam[10]、Nr[11]、KEGG[12]、Swissprot[13] 库对所有基因进行功能注释。使用 GIHunter（http://www5.esu.edu/cpsc/bioinfo/software/GIHunter/）预测可能存在的基因岛。

根据基因组序列绘出其 GC 图，使用 $G–C\ /\ G+C$ 的计算方法来进行 GC skew 分析，同时根据 COG [14] 的注释结果和基因的位置信息绘出 COG 注释的基因在基因组上的分布情况。结果如下所示：

由外至内分别是：tRNA 相关基因；正向基因的 COG 注释——以不同的颜色区分（可参见右上角说明）；正向基因的位置；rRNA 基因；反向基因坐标；反向基因 COG 注释；GC 含量，以平均 GC 为基准线，向外突出的表示高于均值，向内突出的表示低于均值；最内圈为 GC skew 值，紫色表示小于 0，绿色表示大于 0。

基因预测统计信息如下表所示：

3.1 COG 分析

COG，即 Clusters of Orthologous Groups of proteins 其中文释义即“同源蛋白簇”。构成每个 COG 的蛋白都是被假定为来自于一个祖先蛋白，并且因此或者是 orthologs 或者是 paralogs。Orthologs 是指来自于不同物种的由垂直家系（物种形成）进化而来的蛋白，并且典型的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 分为两类，一类是原核生物的，另一类是真核生物。原核生物的一般称为 COG 数据库；真核生物的一般称为 KOG 数据库。由于 COG 数据库早已停止更新，目前一般使用 EggNOG 数据库收录的 Ortholog Cluster 信息进行 COG 分析。我们将所有预测到的蛋白质与 EggNOG 4.0[14] 数据库进行比对后，取 e value 1e-35 为阈值，取 best hit one 作为映射依据进行 COG 注释。

3.2 KEGG 通路分析

KEGG[12][15] 数据库的主要特点是可以呈现经典的代谢通路。我们将所有预测到的基因比对到 KEGG 数据库中，并映射到通路上，结果入下所示：

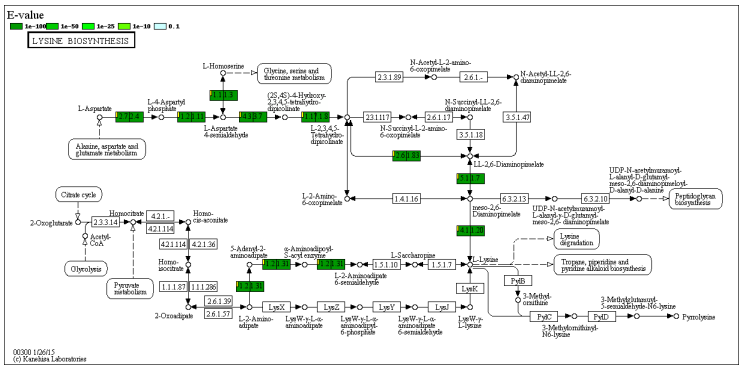

图 2. QC 分析结果

为便于查看差异基因在通路图中的分布情况，我们将差异基因标注到通路图中，查看方法如下：拿到全部分析结果后，打开 out 结果目录中文件夹下 04.OtherDatabase 文件夹子文件夹下的 KEGG 目录，解压 Pathway.tar.gz 压缩包，点击 index.html 进行浏览，您点击不同的通路名称会弹出相应的通路图。其中，蓝绿色标记代表以样本作为 query，以 KEGG 数据库作为 subject 进行序列比对的结果。红黄色标记代表以 KEGG 作为 query，以样本蛋白质序列作为 sujbect 进行比对的结果。比对结果的 e value 范围以不同颜色进行区分，如果两者 blast 结果形成双向最优，则在 KO 的节点左上角形成黄色标记。包含下鼠标悬停于标记的 KO 节点，弹出基因细节框，包括核酸序列，蛋白序列以及序列比对情况。以上步骤可脱机实现，如连接互联网，点击各个节点，可以连接到 KEGG 官方数据库进行查看。

3.3 CDS 序列多数据库比对注释

将所有的 CDS 序列利用 blast 比对到 KEGG、Swissprot、Nr、Nt、eggNOG 数据库，e-value 阈值统一为 1e-35。结果如下：

3.4 IS 序列分析

使用 ISFinder[16] 使用 blastn 进行预测，使用阈值为 e value$1e-5$，

3.5 CRISPR 分析

共发现多个 Crispr 功能单元。

共寻找到 0 个 CRISPR 单元。

3.6 噬菌体前体分析

使用 Prophage_Finder[8] 利用默认参数进行噬菌体前体预测，发现多个噬菌体前体。

表 4. 123

11	22	33
aa ee	bb ff	静静地 [1] 我将离开你，请将眼角的泪逝去。漫漫岁月里 dd

[1] 我的天啊啊哪!!!

表 3. 序列注释明细

	CDS	bases	contigs	gene	misc_RNA	rRNA	tRNA	tmRNA	Intergenetic Region%	GC%	Gene GC%	Intergenetic GC%	Gene Avg Length	Intergenetic Avg Length
LM1212_Genome1	5903	5705934	1	6185	132	42	107	1	14.11	35.56	36.41	30.39	792	130
LM1212_Plasmid1	269	248002	1	269	0	0	0	0	23.11	35.07	36.07	31.77	708	213
LM1212_Plasmid2	176	157839	1	177	1	0	0	0	19.77	35.92	36.49	33.57	715	176
LM1212_Plasmid3	127	112709	1	128	1	0	0	0	19.59	34.48	35.21	31.50	708	172
LM1212_Plasmid4	49	35121	1	49	0	0	0	0	25.40	33.84	34.81	30.99	534	182
LM1212_Plasmid5	41	28658	1	41	0	0	0	0	32.80	37.24	38.21	35.26	469	229
LM1212_Plasmid6	14	14193	1	16	2	0	0	0	40.11	32.95	33.34	32.36	531	355
LM1212_Plasmid7	7	13118	1	9	2	0	0	0	47.39	34.93	36.57	33.10	766	690
LM1212_Plasmid8	4	4975	1	5	1	0	0	0	34.75	34.85	34.90	34.76	649	345
total	6590	6320549	9	6879	139	42	107	1	14.99	35.52	36.38	30.69	781	137

4 Introduction2

我勒个去

表 5. 123

11	22	33
aa ee	bb ff	静静地 [1] 我将离开你，请将眼角的泪逝去。漫漫岁月里 dd

[1] 我的天哪!!!

5 Introduction2

我勒个去

表 6. 123

11	22	33
aa ee	bb ff	静静地 [1] 我将离开你，请将眼角的泪逝去。漫漫岁月里 dd

[1] 我的天哪!!!

6 附录和说明

6.1 名词解释

6.1.1 Polymerase Reads

Polymerase Reads 指测序下机产生的原始 reads 为 Polymerase Reads。由于 Pacbio 采用环状文库进行测序，如果读长长于文库长度的话，每一个碱基会被覆盖多次。因此，Polymerase Reads 的长度往往远高于文库的实际长度。

6.1.2 Subreads

Subreads 指对 Polymerase Reads 进行后续处理后得到的 reads。处理过程主要有两部：1. 侦测 Polymerase reads 序列中的 Adapter 序列并开环，并利用开环后的序列进行 Consensus Calling 以提升测序质量。2. 对 Consensus Calling 后的序列进行质量检验，将低质量区域进行滤除。Polymerase Reads 经过这两部分析后，得到了过滤后的数据叫做 Subreads。

6.1.3 Observe Insert length

Polymerase Reads 经过 Adapter Detection 开环后进行长度检验，其检验的方法是从一个 apadter 到下一个 adapter 序列之间的读长的平均值。

6.1.4 Overlap Graph

序列 Overlap Graph 也称为重叠关系图。序列组装是建立在序列末端具有重叠关系的基础上的。把每一个 reads 看成一个点，如果两个序列间有 overlap 关系，就有一条带有方向的变将两个序列相连。其中边的出发方向代表序列的 3‘端，边的进入方向，代表序列的 5‘端。

7 结果说明

7.1 总体结果展示

分析结果根目录下总共包含 6 个文件夹

表 7. 分析结果目录结构说明

	目录	用途
参考序列	reference/	存储参考序列
步骤 1.1	01-1.Circyled_contig/	进行环化处理的 Contig
步骤 1.2	01-2.Adjusted_contig/	将环化处理后的结果在 Oric 位置进行开环的结果
步骤 1.3	01-3.Assembly_END/	根据基因组大小对所有结果进行重新排序、命名
步骤 2	02.RepeatMask/	重复序列预测
步骤 3	03.Annotation/	序列基因预测,功能分析,生成 GBK 等文件
步骤 4	04.OtherDatabase/	基因与 GO，KEGG，Nr，eggNOG 进行比对
步骤 5	05.InsertionSequence/	IS 序列预测
步骤 6	06.ProPhage/	噬菌体前体预测
步骤 7	07.Crispr/	Crispr 单元预测
步骤 8	08.Genomic_Island/	基因岛预测
步骤 9	09.AllResult/	所有结果的汇总表
步骤 10	10.CircleGraph/	基因组分析可视化图

7.2 reference 文件夹

包含一系列的 gbk 和 fasta 文件，这些文件是拼接得到序列在 NCBI Blast NT 库的 best hit one 结果。我们将其 gbk 文件和序列 fasta 文件下载下来，方便后续的比较基因组分析。

7.3 01-1. Circyled_contig/文件夹

微生物基因组是环状序列，而我们拼接得到的是线性的字符串。在字符串两侧可能有过拼引起的序列重叠，我们将这些能够 overlap 在一起的重叠关系进行比对，并得到 consensus Sequence，从而实现序列的环化。内部文件的结果为 fasta 格式。

7.4 01-2.Adjusted_contig/文件夹

将所有的拼接结果预测复制起点，并从复制起点重新进行开环。对于有参考序列的物种，使用参考序列的开环基因作为开环结果，并将拼接结果重新开环，得到最终的分析结果。

7.5 Assembly_END/文件夹

我们以长度小于 $1MB$ 作为 Plasmid 和 Chromosome 区分标准，将所有的组装结果进行分类和重新编号。

7.6 02.RepeatMasker 文件夹

通过组装结果与已知的转座子序列库进行比对来查找转座子序列。具体方法是，通过 RepeatMasker 软件（使用 Repbase 数据库）和 RepeatProteinMasker 软件（使用 RepeatMasker 自带的转座子蛋白库）两种方法来预测转座子；通过 TRF（Tandem Repeat Finder）软件预测串联重复序列。

表 8. 重复序列分析结果

bfseries 文件名	说明
total.fna.masked	重复序列 mask 后的序列
total.fna.out	重复序列比对后的初始结果
total.fna.out.gff	重复序列比对结果的 gff 文件

7.7 03.Annotation/文件夹

对所有序列组装结果进行基因预测和 ncRNA 预测。并以基因预测结果的 swissprot 注释结果和 COG 注释结果作为标准，生成可提交 NCBI 的 tbl 文件。对于一些涉密菌株，我们直接生成 GBK，gff 等 NCBI FTP 下提供的文件格式，方便您日后的研究。

7.8 04.OtherDatabase 文件夹

每一个染色体或者质粒我们都进行了分别的单独注释。结果包含 Detail 和 Table 两个文件夹：Detail 文件夹是明细结果，包含：

表 9. 数据库注释结果

bfseries 文件夹	功能
Swiss 文件夹	SwissProt 比对结果，Gene Ontology 分析结果
eggNOG 文件夹	eggNOG 比对结果，COG 分析结果
KEGG 文件夹	KEGG Pathway 分析结果

其中，每一个文件夹下的都有 Readme 文件。Blast 的 e-value 阈值统一为 $1e-35$。Table 文件夹是将所有结果按照染色体和不同数据库进行了分门别类的整理，详情请见其下的 Readme 文件。

7.8.1 Detail 文件夹

7.8.1.1 Swiss 文件夹

将所有的基因序列比对到 swissprot 数据库，并使用 blast2go 进行 GO Mapping。所有的 GO 根据 GO 的有向无环图向上回溯，直到第三层。而后统计第三层 GO 的分析结果结果说明如下：

文件	说明
*.GO-mapping.detail	基因映射到 GO 的列表，可以提交到 WEGO 等网站自动生成可视化结果。EXCEL 打开
*.Genome1.GO-mapping.list	根据有向无环图自动回溯的 GO 过程，包含每一个第三级 GO 下所包含的基因和期 GO 映射, 用 Excel 打开。
*_GO.stats	每一个第三级 GO 所映射到的基因个数，由于 GO 是有向无环图结果，一个子 GO 可能具有多个 parent GO。所以该部分的基因总数要大于实际的基因总数。Excel 打开
*_GO.tsv	每一个基因的 GO 详细映射结果，用 Excel 打开
*_SwissAlignment.tsv	所有基因与 swissprot 数据库比对的详细结果，用 Excel 打开
*.xls	Swissprot 和 Gene Ontology 分析结果的整合结果。用 Excel 打开
stat*	GO 分析的可视化结果。
Draw.R	GO 分析可视化画图脚本，用 R 运行。

7.8.1.2 eggNOG 文件夹

为所有预测到的蛋白质与 Eggnog 4.2 数据库进行比对后，Mapping 到的 COG 结果。：文件夹结果如下：

文件	说明
*_AlignEggNOG.tsv	与 eggNOG 数据库的详细比对结果，Excel 打开
*_COG.tsv	每一个基因的 COG 映射结果,Excel 打开
*_COG.xls	每个基因与 eggNOG 比对和 COG 映射结果整合结果,Excel 打开
*_COG.stats	每一个 COG 功能分类的基因个数统计表,Excel 打开
stat*	GO 分析的可视化结果。
Draw.R	GO 分析可视化画图脚本，用 R 运行。

7.8.1.3 KEGG 文件夹

该文件夹放置 KEGG 通路分析结果，我们将所得的序列比对到 KEGG 数据库并进一步映射到 KO（KEGG Orthology），并通过 KO 映射到 Pathway。附件说明如下：

文件	说明
*_pathway.tsv	基因映射到 KEGG KO 和 Pathway 的明细，用 Excel 打开 *_AlignKEGG.tsv 基因序列与 KEGG 数据库比对结果，用 excel 打开
*_PathwayCategory_Stats.stat	KEGG 每一个功能模块的 Pathway 映射到的基因个数统计, 用 excel 打开 stat.* Pathway 分析可视化结果，提供 tiff 和 PDF 两个版本
*.tar.gz	Pathway 分析结果的可视化结果，请解压，有两个文件夹，其中 doctree 文件夹位搜索索引，无需打开。Pathway 文件夹下是一个网站，请点击 index.html 观看，每一个 Pathway 如果有基因被比对上，后面会出现 all 字样。
*.R	可视化画图脚本，用 R 语言运行，您可以根据需要自行调整。

7.8.1.4 Nr 文件夹

将所有的基因序列比对到 Nr 数据库, 结果说明如下：

文件	说明
**.xls	详细的比对结果，用 Excel 打开。

7.8.1.5 Nt 文件夹

将所有的基因序列比对到 Nt 数据库, 结果说明如下：

文件	说明
*.xls	详细的比对结果，用 Excel 打开。

7.8.2 Table 文件夹

该文件夹放置所有注释分析的表格，分成两类，第一类是按照不同的数据库进行分类，每一个子文件夹放置的 excel 文件包含该数据库下所有样本的注释结果。第二类是按照样本来源分类，每一个子文件夹下放置不同染色体的基因注释信息附件说明如下：

文件夹	说明
Database 文件夹	不同数据库注释的汇总文件
Choromsome 文件夹	按照不同染色体进行的分类汇总文件
All_HasAnnotation.xlsx	所有能够注释的基因的注释明细表
GeneFeature+Annotation.xlsx	注释的基因信息和基因序列等信息的总表

7.9 05.InsertionSequence 文件夹

存储所有的 IS 分析结果

使用在线网站 ISFINDER（https://www-is.biotoul.fr/）预测 IS 序列，使用 blastn+e value $1e-5$ 作为参数包含以下结果：

文件	说明
*.fa	预测到的 IS 序列
*.xls	IS 预测结果的表格
*.stat	IS 预测结果的统计结果，用 excel 打开!!

7.10 06.ProPhage 文件夹

使用 PhageFinder 进行前噬菌体寻找。结果如下：

文件	说明
*.con	前噬菌体序列
*.xls	前噬菌体序列的详细详细信息表格
*.pep	前噬菌体内包含的蛋白
*.seq	前噬菌体内包含的基因

7.11 07.Crispr 文件夹

使用 PILE-CR 预测 Crispr 序列。

文件	说明
*_DP.fa	Crispr 串联重复序列
*_Spacer.fa	Crispr 中 Spacer 序列
*_Spacer.xls	Crispr 中 Spacer 详细注释结果和基因组位置
*_NtAlign.tsv	Crispr 中 Spacer 的序列注释结果
*_RAW.txt	Pilercr 原始分析结果

7.12 08.Genomic_Island 文件夹

使用 GIHunter 预测基因岛的结果。

文件	说明
*_GIs.txt	GIHunter 预测的原始结果（如果没发现基因岛的话，该文件为空）
*_GI.xls	基因岛的明细信息
Genome1_GI.stat	基因岛的长度统计信息
*_GIProtein.faa	基因岛内包含的蛋白质序列
*_GIGene.ffn	基因岛内包含的基因序列
*_GI.fa	基因岛完整序列
*_GI_Component.xls	基因岛包含基因的明细信息

7.13 09.AllResult 文件夹

包含一个 excel 文件，该文件包含所有的分析结果。

8 分析用软件版本和数据库版本

除图片和网页外，所有的序列文件如 fasta 等推荐使用 notepad 的文本编辑器打开。其余的文件推荐使用 excel 打开

表 10. 分析所用软件和版本明细

用途	软件名	版本
序列组装	HGAP	v3.0
数据过滤	SMRTPipe	v2.3
序列比对	blasr	v1.3.1.140182
基因预测	Prodigal	v2.6
信号肽预测	Signalp	v4.1
RNA 预测,Crispr 预测	Infernal	v1.1.1
tRNA tmRNA 预测	Aragorn	v1.2.36
数据清洗	GATK	v1.9
二代数据比对	BWA	v0.9a
Phage 序列寻找	Phage Finder	v2.0
Crispr 序列寻找	PILE-CR	V1.0
KEGG 富集	KOBAS	v2.0
基因岛预测	GIHunter	v1.0 （http://www5.esu.edu/cpsc/bioinfo/software/GIHunter/）
rRNA 预测	RNAmmer	v 1.2

表 11. 分析所用数据库和版本明细

数据库	版本
KEGG	v74.0
Nr	2015/11/11
Rfam	v12.0
EggNOG	v4.2
Gene Ontology	2015/9/10

9 ok1

12344556 [12] 好好学习
123456

9.1 ok2

10 Conclusion

参考文献

[1] Chen-Shan Chin, David H. Alexander, Patrick Marks, Aaron A. Klammer, James Drake, Cheryl Heiner, Alicia Clum, Alex Copeland, John Huddleston, Evan E. Eichler, Stephen W. Turner, and Jonas Korlach. Nonhybrid, finished microbial genome assemblies from long-read smrt sequencing data. *Nature Methods*, 10(6):563–569, 2013. identifier: nmeth.2474.

[2] N. Chen. Using repeatmasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics*, Chapter 4:Unit 4.10, 2004.

[3] W. Bao, K. K. Kojima, and O. Kohany. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*, 6:11, 2015.

[4] G. Benson. Tandem repeats finder: a program to analyze dna sequences. *Nucleic Acids Res*, 27(2):573–80, 1999.

[5] Doug Hyatt, Gwo-Liang Chen, Philip F. LoCascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1):119, 2010. identifier: 1471-2105-11-119.

[6] E. P. Nawrocki and S. R. Eddy. Infernal 1.1: 100-fold faster rna homology searches. *Bioinformatics*, 29(22):2933–2935, 2013.

[7] K. Lagesen, P. Hallin, E. A. Rodland, H. H. Staerfeldt, T. Rognes, and D. W. Ussery. Rnammer: consistent and rapid annotation of ribosomal rna genes. *Nucleic Acids Research*, 35(9):3100–3108, 2007.

[8] D. E. Fouts. Phage_finder automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Res*, 34(20):5839–51, 2006.

[9] R. C. Edgar. Piler-cr: fast and accurate identification of crispr repeats. *BMC Bioinformatics*, 8:18, 2007.

[10] E. P. Nawrocki, S. W. Burge, A. Bateman, J. Daub, R. Y. Eberhardt, S. R. Eddy, E. W. Floden, P. P. Gardner, T. A. Jones, J. Tate, and R. D. Finn. Rfam 12.0: updates to the rna families database. *Nucleic Acids Research*, 43(D1):D130–D137, 2015. item_number: gku1063.

[11] K. D. Pruitt, T. Tatusova, G. R. Brown, and D. R. Maglott. Ncbi reference sequences (refseq): current status, new features and genome annotation policy. *Nucleic Acids Research*, 40(D1):D130–D135, 2011. item_number: gkr1079.

[12] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi. Kegg for linking genomes to life and the environment. *Nucleic Acids Research*, 36(Database):D480–D484, 2007.

[13] K. Watanabe and S. Harayama. [swiss-prot: the curated protein sequence database on internet]. *Tanpakushitsu Kakusan Koso*, 46(1):80–6, 2001.

[14] S. Powell, K. Forslund, D. Szklarczyk, K. Trachana, A. Roth, J. Huerta-Cepas, T. Gabaldon, T. Rattei, C. Creevey, M. Kuhn, L. J. Jensen, C. von Mering, and P. Bork. eggnog v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res*, 42(Database issue):D231–9, 2014.

[15] C. Xie, X. Mao, J. Huang, Y. Ding, J. Wu, S. Dong, L. Kong, G. Gao, C. Y. Li, and L. Wei. Kobas 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res*, 39(Web Server issue):W316–22, 2011.

[16] P. Siguier, J. Perochon, L. Lestrade, J. Mahillon, and M. Chandler. Isfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res*, 34(Database issue):D32–6, 2006.