# 细菌样品建库测序分析报告

项目名称：

项目编号：

分析人员：

审核人员：

报告日期：

报告单位：西安浩瑞基因技术有限公司

# 1 技术背景

从早期的Sanger测序到第二代高通量测序、再到现在的单分子测序技术（第三代高通量测序），DNA测序技术一直推动着生命科学研究的发展。

当第二代高通量测序技术进入成熟阶段后，读长过短、PCR扩增带来的偏向性等问题开始日益凸显；作为基因组学上新的转折点，以PacBio单分子实时测序技术及纳米孔单分子测序技术为首的第三代高通量测序技术（Third-generation Sequencing，TGS）开始进入科研应用。

PacBio测序是以SMRT Cell为载体进行测序反应，通过在位于**零模波导孔**（Zero-mode Waveguides，**ZMW**）孔底部的荧光信号检测区锚定DNA聚合酶和一条DNA片段，并通过不同荧光标记的核苷酸及荧光激发的过程，将不同碱基的信号捕捉下来，从而得到DNA序列信息。

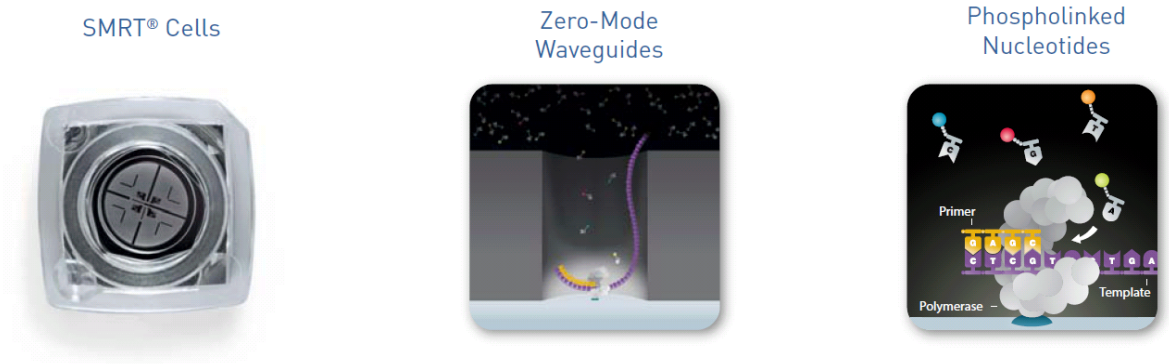



图1-1 PacBio SMRT 测序原理示意图

PacBio SMRT测序技术在单分子水平上对DNA分子的实时测序，成功解决了二代测序几大困扰：极端 GC含量区域覆盖度低、高度重复区域无法较好地拼装、大片段变异难以准确检测、不能直接检测碱基修饰等问题。

2015年10月PacBio公司发布的Sequel平台以其高通量的优势备受关注，与PacBio RS II相比，单个cell的ZMWs数目从150,000上升到1,000,000，单个cell的数据产出大约会是以前的7倍，但是体积只有PacBio RS II的三分之一。在2019年PacBio公司发布了Sequel II测序平台，在测序通量、准确度以及读长方面进行了进一步的技术革新。升级后的SMRT Cell 8M的ZMWs数目从1,000,000上升到了8,000,000，通量较之前提升约8倍。测序成本显著降低，项目周期也大大减少。目前均已广泛应用于科研动植物基因组、全长转录组和微生物等领域。

# 2 实验流程

Pacbio测序技术对样品DNA的要求非常高，浩瑞基因对客户提供的DNA进行严格的样品检测，从源头上保证质量。对检测合格的样品建库、上机测序，每个环节都严格把控，保证测序数据的准确性和PacBio测序长读长的特性。

首先用QIAGEN® Genomic试剂盒提取高质量的DNA，构建文库，利用Pacbio Sequel系列测序仪对DNA进行单分子实时荧光测序，获得原始测序数据。实验流程如下图所示：

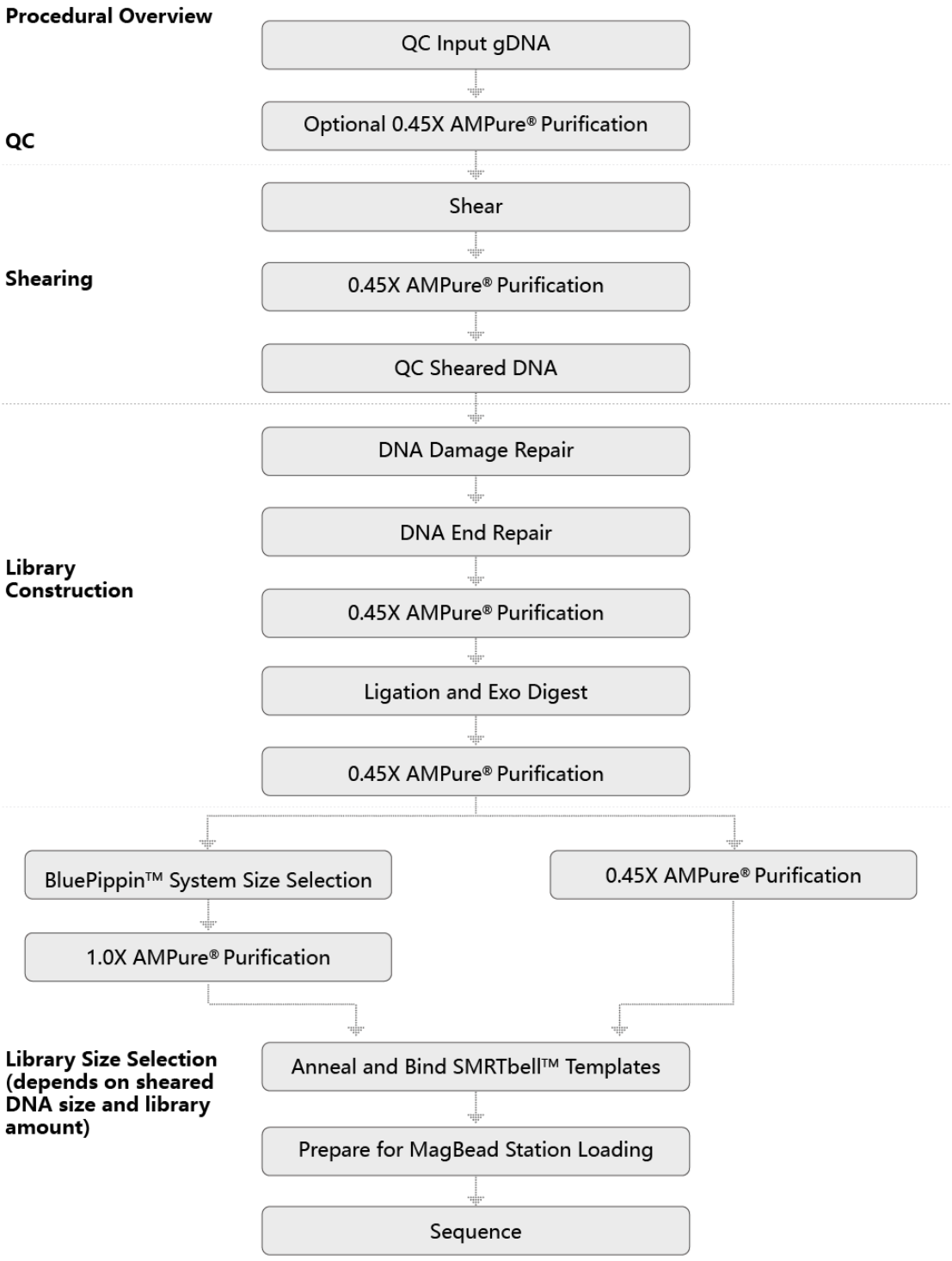



图2-1 实验流程图

## 2.1 DNA质检

采用以下四种方法检测DNA是否合格：

1. 样品的外观性状是否含有异物
2. 0.75%琼脂糖电泳：检测样品是否有降解以及DNA片段大小
3. Nanodrop：检测DNA纯度（OD260/280在1.8-2.0之间；OD 260/230在2.0-2.2之间）
4. Qubit：对DNA进行精确定量

## 2.2 文库构建及质量检测

样本质检合格后，根据建库的片段大小用g-TUBEs (Covaris, USA)对基因组DNA进行目的打断；利用磁珠富集、纯化目的片段DNA；将片段化的DNA进行损伤修复和末端修复；在DNA片段两端连接茎环状测序接头，并利用外切酶去除连接失败的片段。再用切胶仪BluePippin(Sage Science, USA)筛选目的片段，纯化后即文库。然后用Agilent 2100 Bioanalyzer (Agilent technologies, USA)检测文库片段大小。在文库片段两端添加发夹状接头构建SMRTbell结构测序文库。

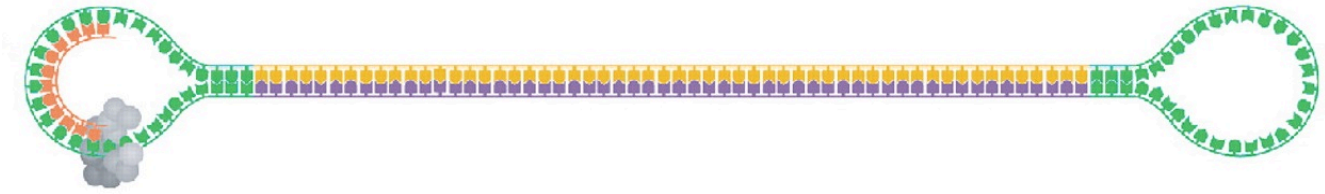


图2.2-1 SMRTbell模板结构示意图

> 注：发夹状接头（绿色）连接到双链DNA分子末端（黄色和紫色），构成闭环。锚定于ZMW纳米孔底部的聚合酶（灰色）与测序引物序列（橘黄色）结合，开启测序。

## 2.3 DNA测序

建库完成后将一定浓度和体积的DNA模板和酶复合物转移到Sequel系列测序仪纳米孔内开始实时单分子测序。


图2.3-1 Sequel II 单分子实时测序

# 3 分析流程

测序获得物种的组学数据后，对数据质量进行评估，利用高质量的数据进行组装得到基因组，进而对组装得到的基因组进行相关质量评估。具体分析流程如下图：

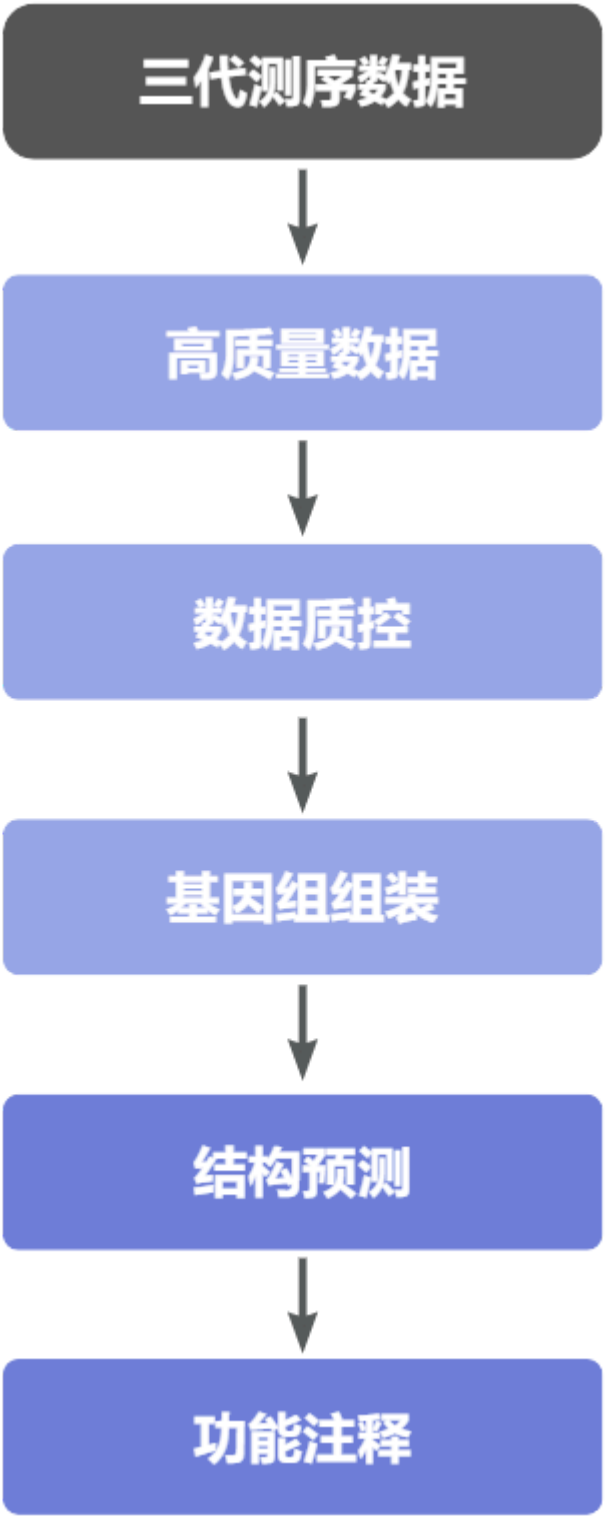



图3-1 数据分析流程

## 3.1 数据质控

### 3.1.1 质控方法

在PacBio的测序平台中，将通过零模波导孔的DNA产生的荧光信号记录成movie进而转化为相应的碱基序列的过程，称为basecalling。使用官方提供的工具SMRTLink进行basecalling获得含接头的测序序列，即酶读（polymerase reads），其长度由反应酶的活性和上机时间决定。酶读去除低质量序列和接头序列后得到subreads。环化共有序列（Circular Consensus Sequencing，CCS）测序模式获得subreads，通过校准同一序列模板多次测序的subreads的随机错误，可将测序准确率提升至99%以上，通过Q20阈值过滤获得高准确性的HiFi reads。

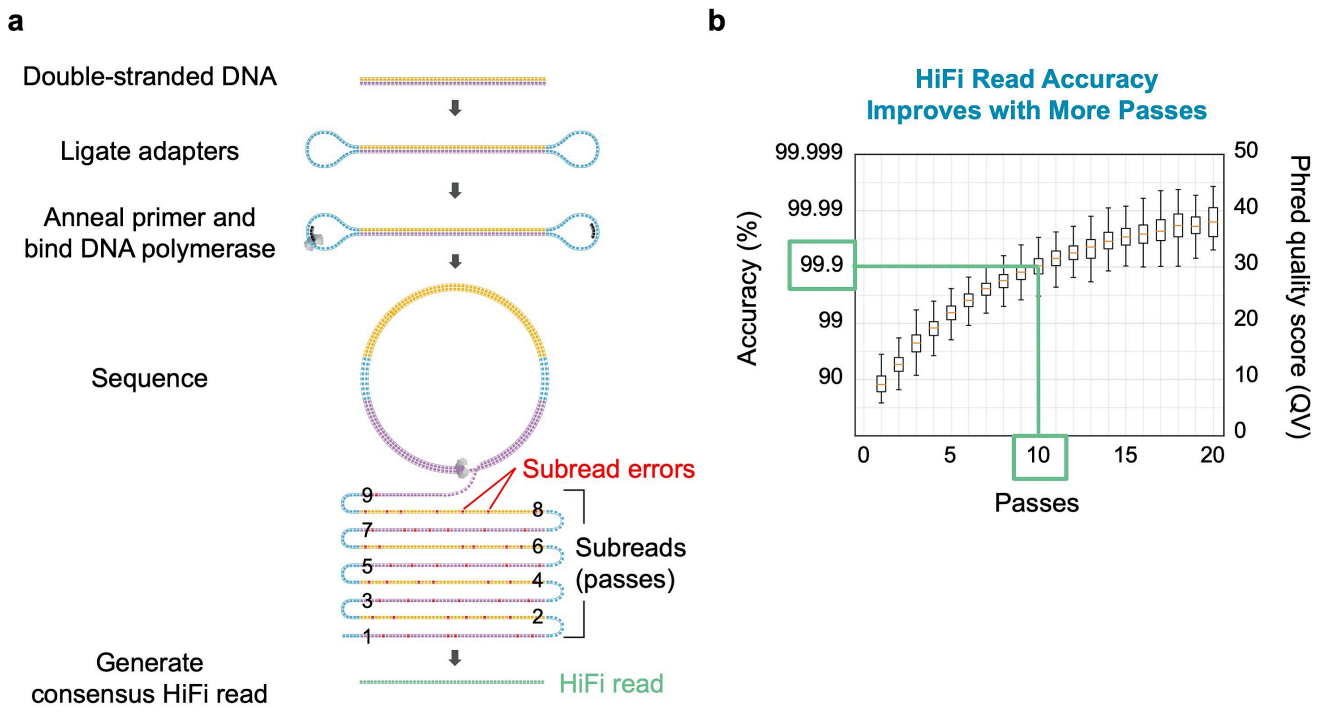



图3.1-1 HiFi read数据质控流程图

### 3.1.2 数据统计

对HiFi reads数据进行统计，总数据量为6.68Gb。各文库及总数据量统计信息如下表所示。

表3.1 三代下机数据统计

| Library id | Total bases(nt) | Total reads | Mean length(nt) | Max length(nt) | N50 length(nt) | >10kb rate(%) | >20kb rate(%) | >40kb rate(%) |
|---|---|---|---|---|---|---|---|---|
| | 6,682,447,063 | 513,499 | 13,014 | 45,974 | 13,445 | 76.94 | 6.09 | 0.01 |

> 注：下机数据的质量会根据前期实验提取建库质量的差别而有所不同。表格各列说明如下表：

| 列名 | 说明 |
|---|---|
| Library id | 测序文库编号； |
| Total bases(nt) | 有效数据的总碱基数； |
| Total reads | 有效数据的总reads数； |
| Mean length(nt) | 有效数据的平均长度； |

| 列名 | 说明 |
| --- | --- |
| Max length(nt) | 有效数据的最长reads长度； |
| N50 length(nt) | 有效数据的N50长度； |
| >10kb rate(%) | 有效数据中长度大于10kb的reads比例； |
| >20kb rate(%) | 有效数据中长度大于20kb的reads比例； |
| >40kb rate(%) | 有效数据中长度大于40kb的reads比例。 |

文库kp-1的read长度分布如下图所示：

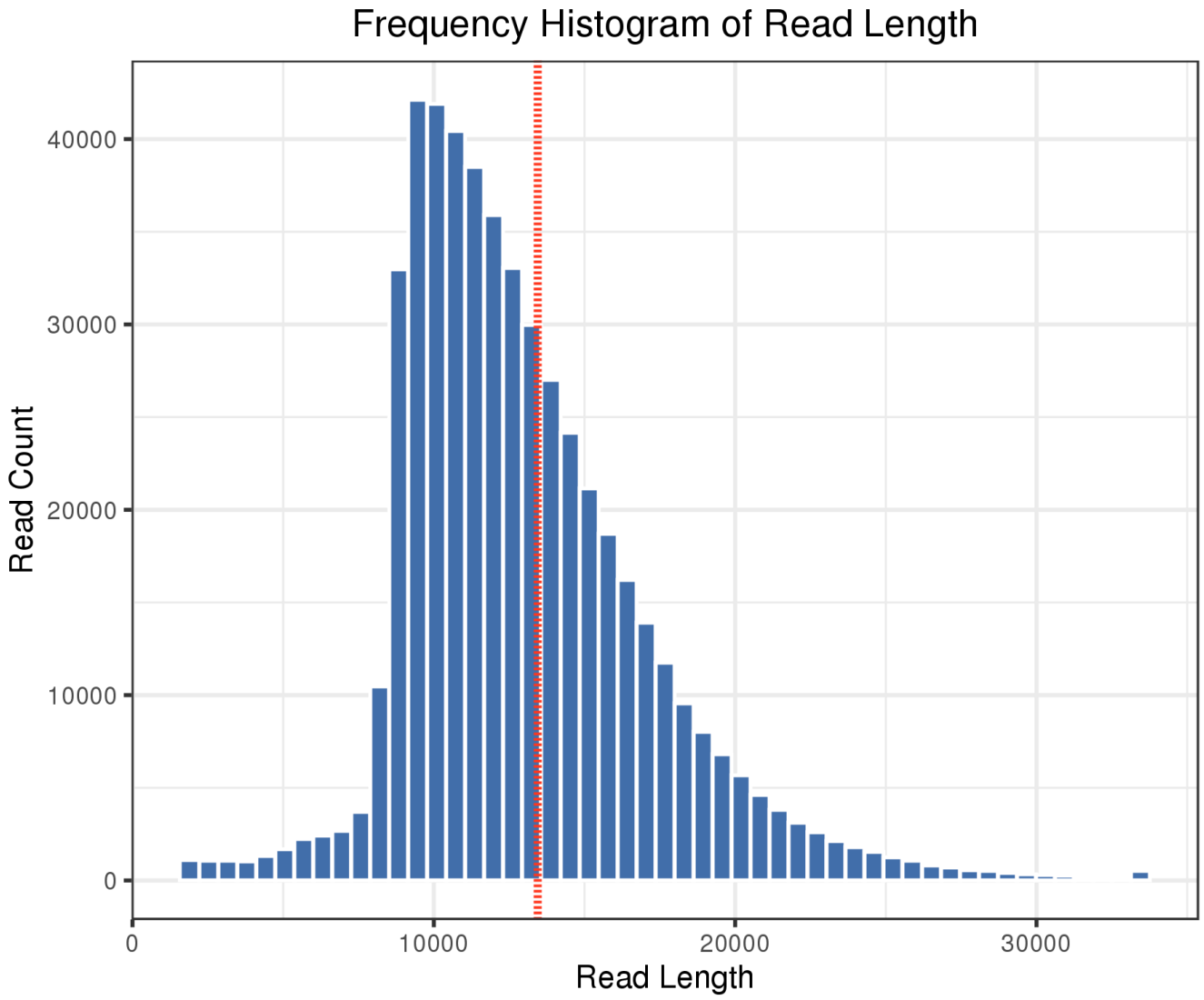



图3.1-2 文库kp-1的读长分布图

其他三代测序数据统计结果文件见目录：src/summary/1_data/TGS/。

根据下机数据过滤及质控的统计分析，该项目的HiFi reads总数据量为6.68Gb，reads数为513,499条，reads平均长度为13.01Kb，其中最长reads长度为45.97Kb。相对其他测序平台，HiFi reads兼顾长读长和高准确性，是目前推荐用于基因组组装的最佳选择。

## 3.2 组装结果

质控过后，利用高质量的HiFi reads进行纯三代组装，使用Flye[1]软件对基因组进行组装。最终的组装结果统计情况如下：

表3.2-1 样本kp-1组装结果统计

| seq_name | length | cov. | circ. |
|---|---|---|---|
| Chr | 5375743 | 1131 | Y |
| plasmid1 | 235980 | 1486 | Y |
| plasmid2 | 125421 | 859 | Y |
| plasmid3 | 65733 | 1004 | Y |

> 注：表格第一列为细菌序列名称，第二列为序列长度，第三列为序列对应的覆盖深度，第四列为是否为环化的序列。

## 3.3 基因组组分分析

组分分析主要包含编码基因预测、非编码 RNA 预测、重复序列分析、CRISPR（Clustered regularly interspaced short palindromic repeats, 规律成簇的间隔短回文重复）序列预测等分析。

### 3.3.1 基因结构预测

本次分析使用Bakta[2]进行基因结构的预测，该软件集合了多种基因结构预测软件，用于快速和标准化注释细菌基因组和质粒的工具。结果统计如下：

表3.3-1 样本kp-1结果统计

| Sample id | Length | Sequence Count | GC | coding density |
|---|---|---|---|---|
| kp-1 | 5802877 | 4 | 56.7 | 88.4 |

表3.3-2 样本kp-1基因结构预测结果

| Type | Number |
|---|---|
| CDSs | 5366 |
| tRNAs | 87 |
| tmRNAs | 1 |
| rRNAs | 25 |
| ncRNAs | 84 |
| ncRNA regions | 53 |
| pseudogenes | 16 |

| Type | Number |
| --- | --- |
| hypotheticals | 142 |
| sORFs | 19 |
| oriCs | 5 |
| oriVs | 0 |
| oriTs | 0 |

> 注：**CDSs**：编码蛋白产物的序列；**tRNAs**：转运RNA；**tmRNAs**：同时具有转运RNA和信使RNA功能的一类RNA；**rRNAs**：核糖体RNA；**ncRNAs**：非编码RNA；**ncRNA regions**：顺式调控区域 (*cis-regulatory regions*)；**pseudogenes**：假基因；**hypotheticals**：假设蛋白；**sORFs**：小开放阅读框；**oriCs**：复制起点；**oriVs**：质粒复制起点(*origin of vegetative replication*)；**oriTs**：转移起点(*origin of transfer*)

### 3.3.2 重复序列分析

重复序列根据组织形式可以分为两种，分别为串联重复序列和分散重复序列。前一种一般成簇的存在于染色体的特定区域，后一种分散于染色体的各个位置。TRF[3] (Tandem Repeat Finder) 软件是搜寻 DNA 序列中的串联重复序列（即相邻的重复两次或多次特定核酸序列模式的重复序列）准确性较高的软件，我们采用 TRF来进行串联重复序列的预测，参数为：2 7 7 80 10 50 500。

结果文件见目录：src/summary/2_assembly/Annotation。

文件名名称：*.2.7.7.80.10.50.500.1.html、*.2.7.7.80.10.50.500.1.txt.html

### 3.3.3 CRISPR 序列预测

CRISPR（Clustered regularly interspaced short palindromic repeats，规律成簇的间隔短回文重复）是由重复序列(Repeat)与长度相似的间隔序列(Spacer)排列而成的，是细菌免疫系统的中心组成部分，其负责序列识别。CRISPR 在细菌免疫系统中起着重要作用，它能够抵抗噬菌体、质粒等外来 DNA 的入侵。我们使用 CRT[4] 软件进行 CRISPR 序列的预测，结果不但会给出 CRISPR 在基因组上的位置，还包含重复序列与间隔序列。

表3.3-4 样本kp-1CRISPR结果统计

| Sequence ID | Number of CRISPR |
| --- | --- |
| kp-1 | 0 |

结果文件见目录：src/summary/2_assembly/Annotation。

文件名名称：*.CRISPR.tsv

## 3.4 基因功能注释分析

基因的功能注释主要是与各种功能数据库进行比对，了解基因的功能，掌握基因的产物及其在生命活动中的作用。目前微生物基因功能注释主要包含 NR/NT 注释、Swiss-Prot 功能注释、COG[5] 功能注释、GO[6]功能注释、KEGG[7] 功能注释等。编码基因的注释结果统计如下表所示：

表3.4-1 基因注释结果统计表

| Database | Number | Ratio(%) |
| --- | --- | --- |
| COG/KOG | 5,172 | 96.04 |
| KEGG | 3,704 | 68.78 |
| GO | 2,641 | 49.04 |
| Pfam | 4,948 | 91.88 |
| NR | 5,375 | 99.81 |
| Swiss-Prot | 4,391 | 81.54 |
| TrEMBL | 5,370 | 99.72 |
| Overall | 5,376 | 99.83 |
| Query | 5,385 | 100 |

> 第一列为注释数据库，第二列为注释到的数目，第三列为注释基因与总共预测基因集的占比。

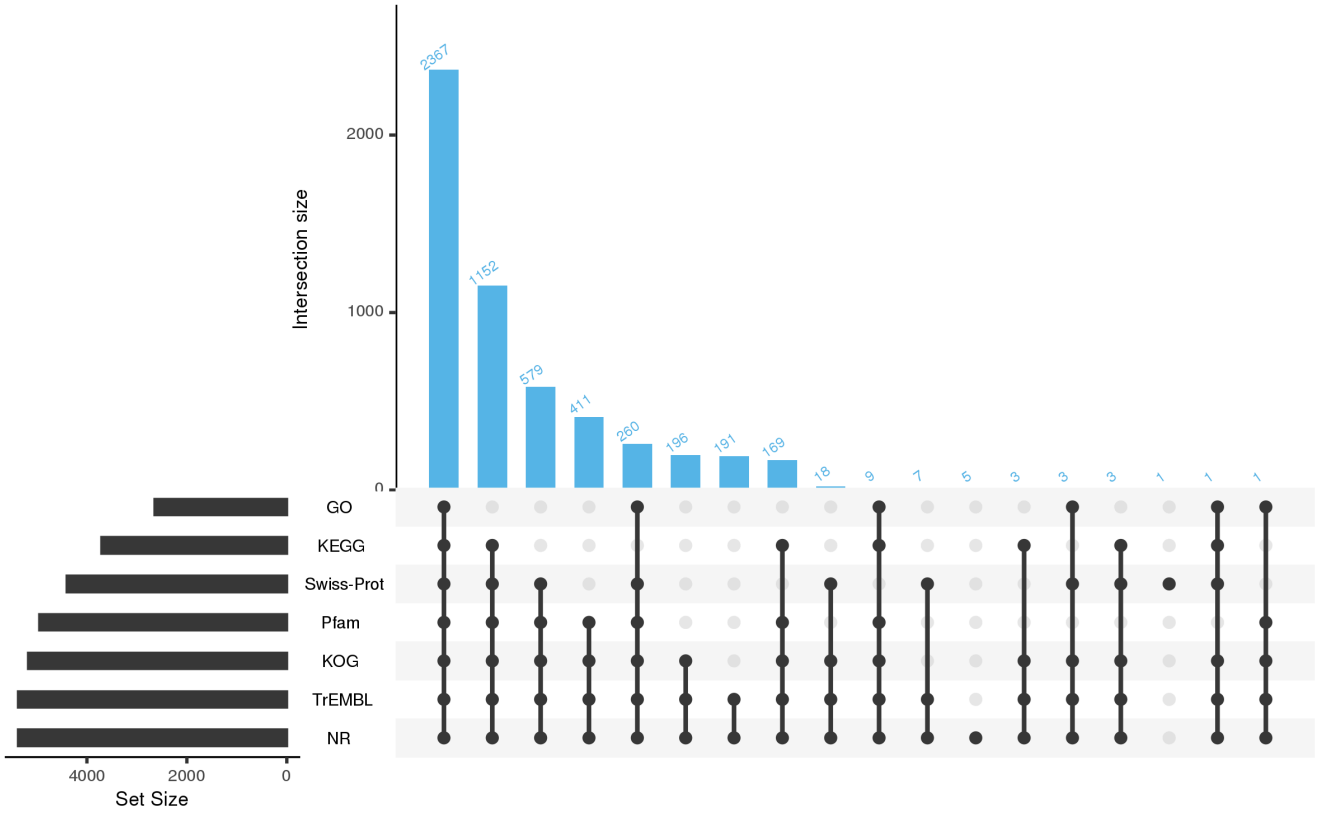



图3.4-1 数据库韦恩图

结果文件见目录：[src/summary/1_data/2_assembly/Annotation](src/summary/1_data/2_assembly/Annotation)。

### 3.4.1 GO数据库注释

GO的全称是Gene Ontology，1988年由基因本体联合会创立基因本体论数据库，其分为三大类：1）细胞组分（Cellular Component）：用于描述亚细胞结构、位置和大分子复合物，如核仁、端粒和识别起始的复合物等；2）分子功能（Molecular Function）：用于描述基因、基因产物个体的功能， 如与碳水化合物结合或ATP 水解酶活性等；3）生物过程（Biological Process）：用于描述分子功能的有序组合，达成更广的生物功能，如有丝分裂或嘌呤代谢

等。基因将依据产物性质归属到其中一类或者多类中。通过GO数据库注释，我们可以依据基因在不同大类中注释的情况，判断其可能的功能。样品GO数据库三大分类统计结果如下图：

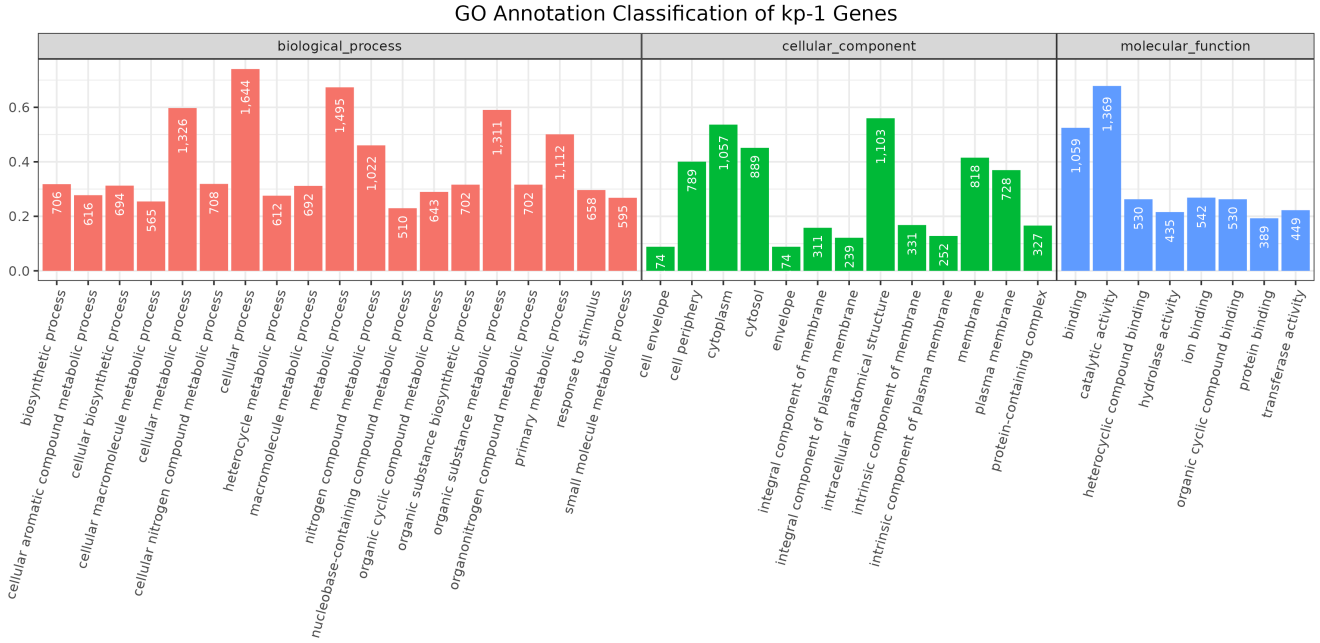



图3.4-2 基因功能注释GO分类图

## 3.4.2 KEGG数据库注释

KEGG全称为Kyoto Encyclopedia of Genes and Genomes，1995年由Kanehisa Laboratories推出0.1版，目前发展为一个综合性数据库，其中最核心的为KEGG PATHWAY数据库。该数据库将生物通路划分为八大类，每一大类下还有细分，每一类均标示上与之相关的基因，同时以图形的方式展示出来。通过该数据库注释，可以方便地寻找与行使某一类功能相关的所有注释上的基因。样品KEGG二级分类统计后获得的柱状图如下：

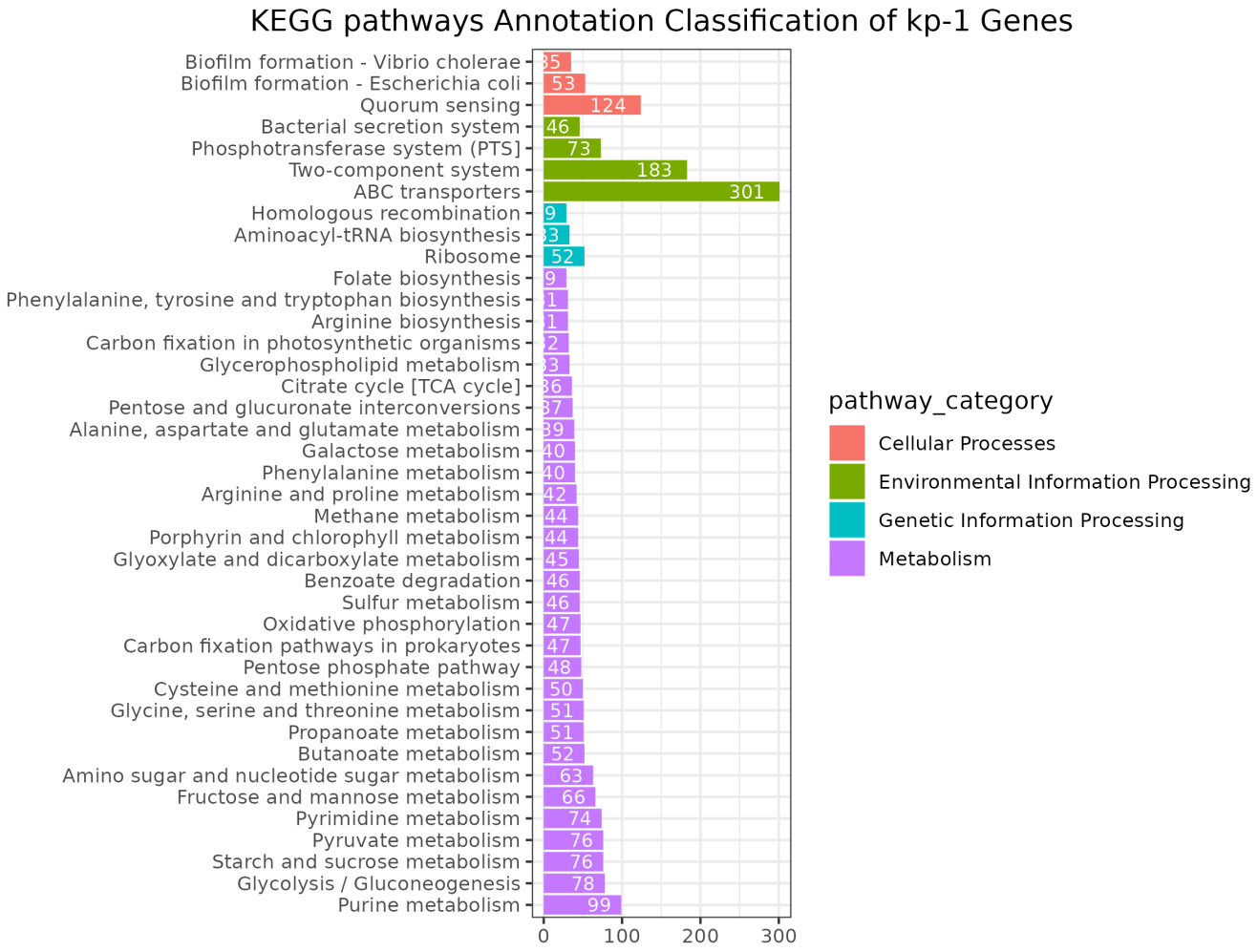



图3.4-3 基因功能注释KEGG分类图

### 3.4.3 COG/KOG数据库注释

COG，全称是Cluster of Orthologous Groups of proteins，由NCBI创建并维护的蛋白数据库，根据细菌、藻类和真核生物完整基因组的编码蛋白系统进化关系分类构建而成。通过比对可以将某个蛋白序列注释到某一个COG中，每一簇COG由直系同源序列构成，从而可以推测该序列的功能。样品的统计结果如下图：

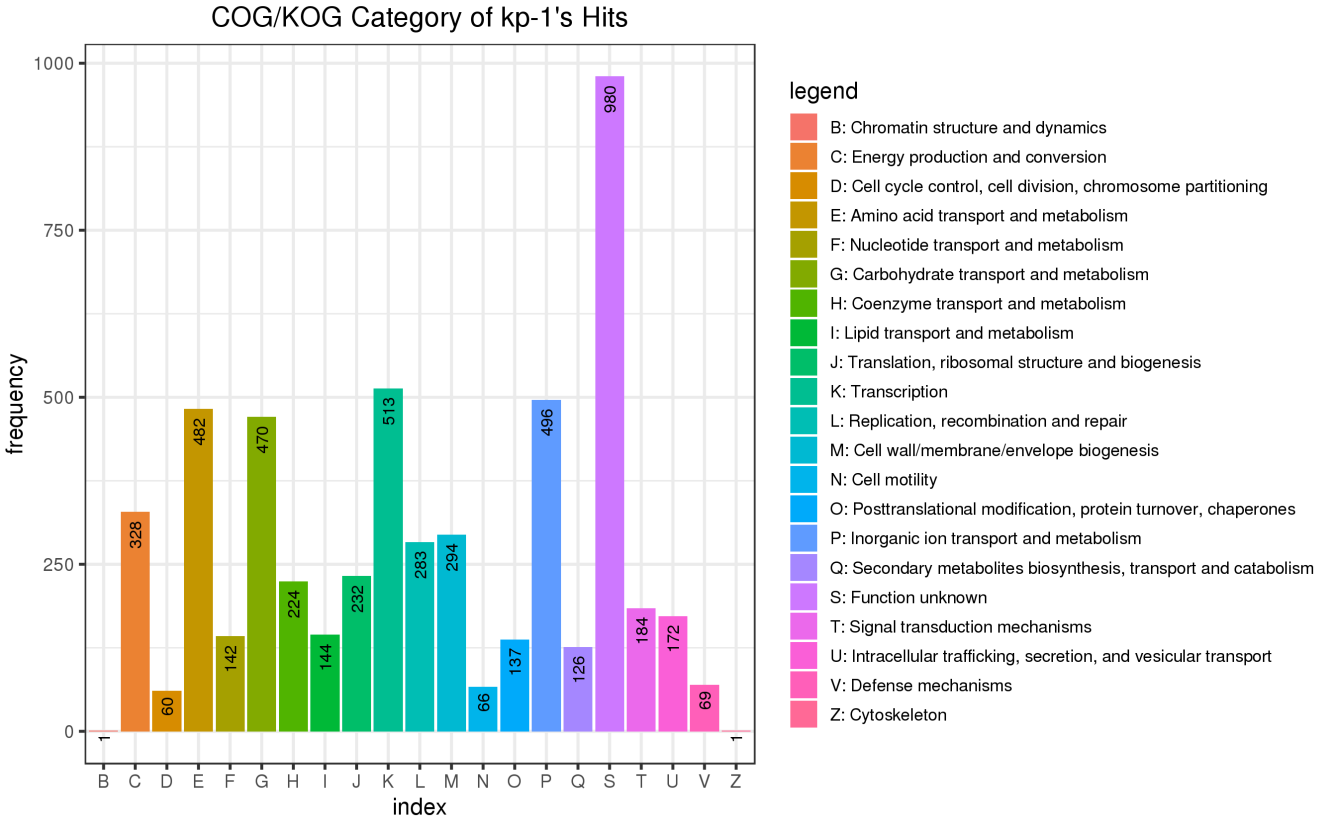



图3.4-4 基因功能注释COG/KOG分类图

### 3.4.4 NR数据库注释

NR全称为Non-Redundant Protein Database，是一个非冗余的蛋白质数据库，由NCBI创建并维护。NR数据库的优点在于内容比较全面，同时注释结果中包含有物种信息，可用作物种分类。缺点在于过多的数据未经过验证，可靠性尚有不足。其统计结果如下图：

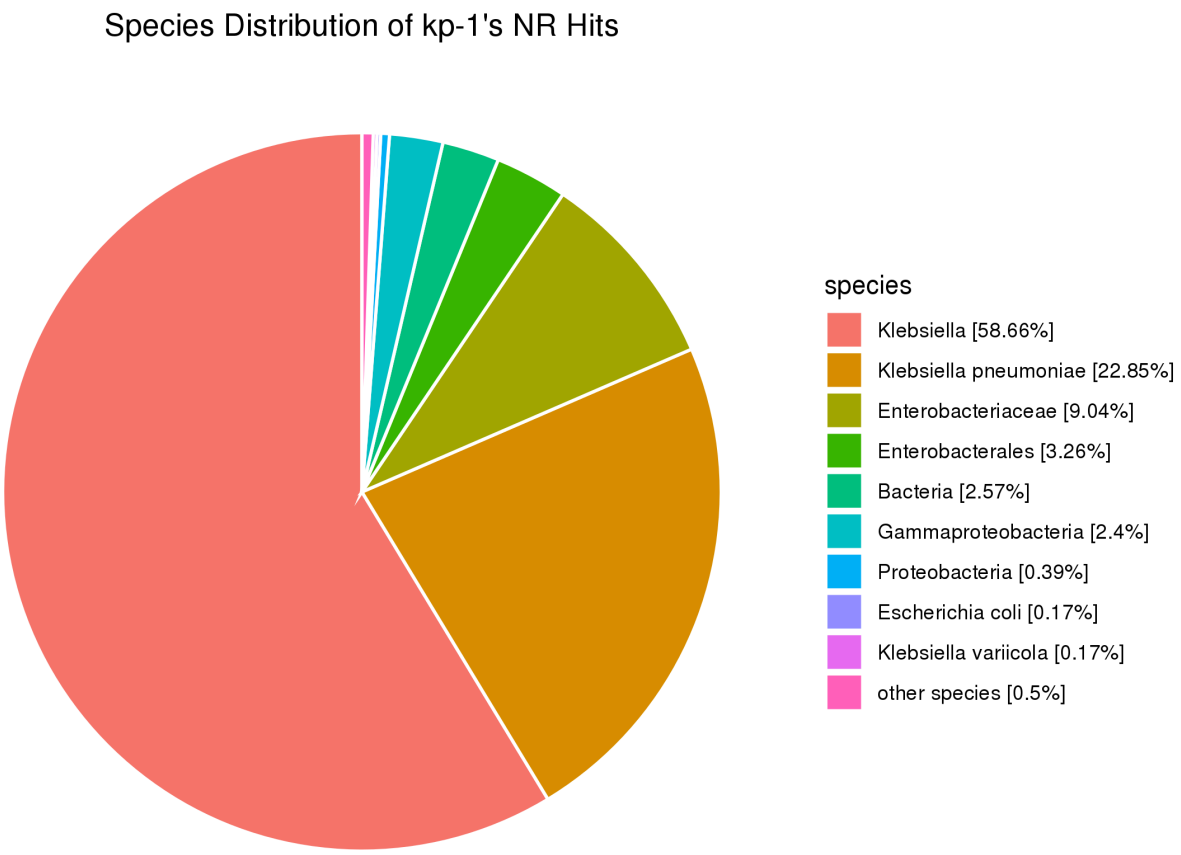



图3.4-5 基因功能注释NR分类图

### 3.4.5 Swiss-Prot 数据库注释

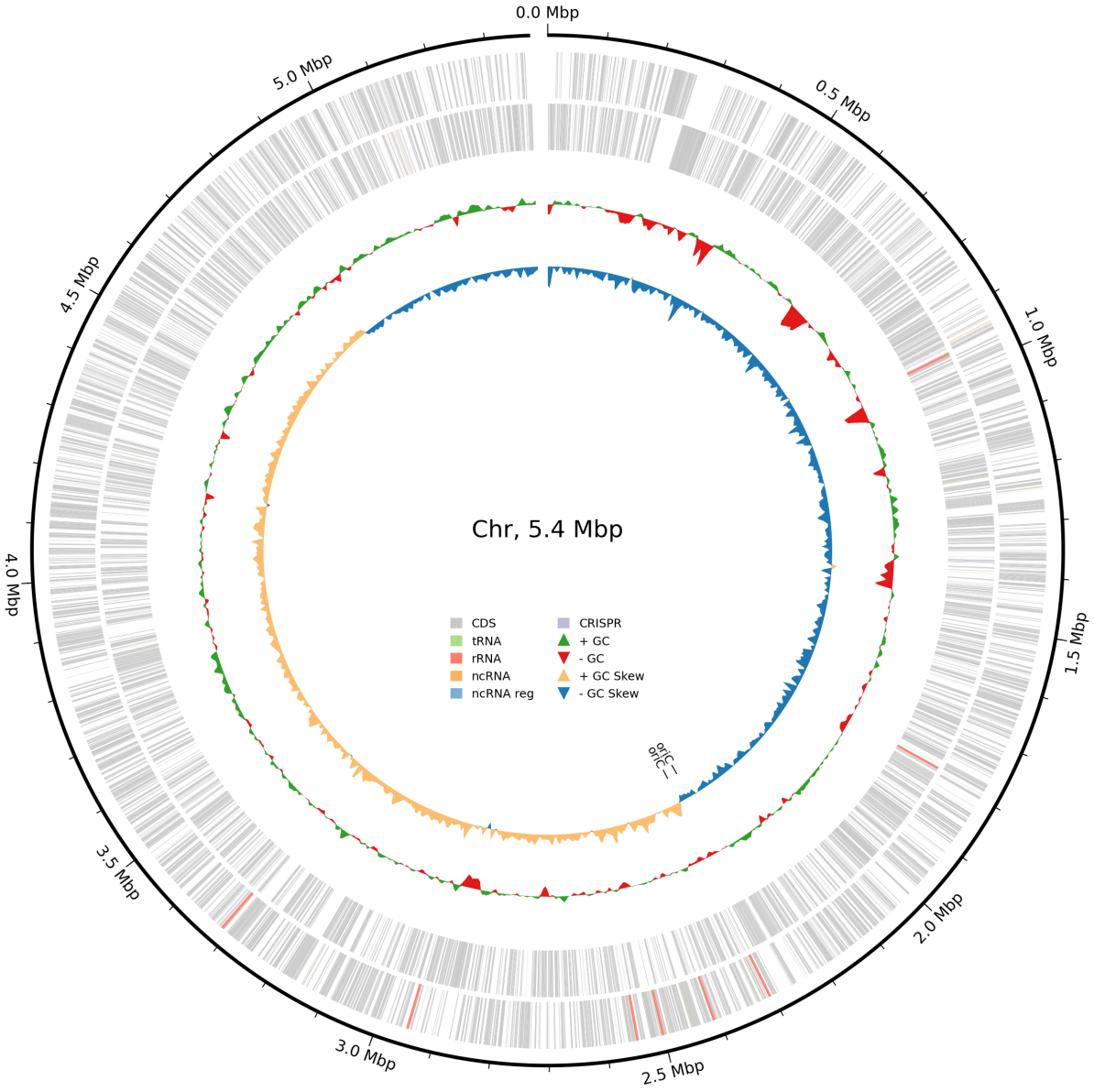
Swiss-Prot是 2002 年由 UniProt consortium 建立的基因数据库，其特点为注释结果经过实验验证，可靠性较高，可用作其他数据的参考。

### 3.4.6 TrEMBL 数据库注释

TrEMBL是从EMBL中的cDNA序列翻译得到的蛋白质序列数据库。TrEMBL数据库创建是于1996年，意为“Translation of EMBL”。该数据库采用SwissProt数据库格式，包含EMBL数据库中所有编码序列的翻译。TrEMBL数据库分两部分，SP-TrEMBL和 REM-TrEMBL。SP-TrEMBL中的条目最终将归并到SwissProt数据库中。而Rem-TrEMBL则包括其它剩余序列，包括免疫球蛋白、T细胞受体、少于8个氨基酸残基的小肽、合成序列、专利序列等。

### 3.4.7 细菌圈图

通过 pyCirclize 创建环形基因组图。分别使用以下特征颜色从外向内表示正向和反向链: 1.CDS: 灰色【#cccccc】 2.tRNA/tmRNA: 浅绿色【#b2df8a】 3.rRNA: 红色【#fb8072】 4.ncRNA: 橙色【#fdb462】 5.ncRNA-区域: 浅蓝色【#80b1d3】




# 4 分析软件

表4.1 分析所用软件信息表

| 分析模块 | 分析内容 | 工具名称 | 版本 |
| --- | --- | --- | --- |
| 数据质控 | 三代数据评估 | SMRTLink | v10.1.0 |
| 基因组组装 | 基因组组装 | Flye | v2.9.6 |
| 结构预测 | 基因结构预测 | Bakta | 1.11.0 |
| 结构预测 | CRISPR 预测 | CRT | v1.2 |
| 结构预测 | Tandem repeats 预测 | TRF | v4.09 |
| 基因注释 | 数据比对 | Diamond | v2.1.8 |

| 分析模块 | 分析内容 | 工具名称 | 版本 |
| --- | --- | --- | --- |
| 基因注释 | 数据比对 | eggnog-mapper | v2.11.1 |

# 5 参考文献


1. Logsdon GA, Vollger MR, Eichler EE. **Long-read human genome sequencing and its applications**. *Nat Rev Genet*. 2020 Oct;21(10):597-614. doi: 10.1038/s41576-020-0236-x. Epub 2020 Jun 5. PMID: 32504078; PMCID: PMC7877196.
2. Du H, Yu Y, .etc. **Sequencing and de novo assembly of a near complete indica rice genome**. *Nat Commun*. 2017 May 4;8:15324. doi: 10.1038/ncomms15324. PMID: 28469237; PMCID: PMC5418594.
3. Rhoads A, Au KF. **PacBio Sequencing and Its Applications**. *Genomics Proteomics Bioinformatics*. 2015 Oct;13(5):278-89. doi: 10.1016/j.gpb.2015.08.002. Epub 2015 Nov 2. PMID: 26542840; PMCID: PMC4678779. .
4. Mikhail Kolmogorov, Jeffrey Yuan, Yu Lin and Pavel Pevzner, **Assembly of Long Error-Prone Reads Using Repeat Graphs**, *Nature Biotechnology*, 2019 doi:10.1038/s41587-019-0072-8.
5. Schwengers O., Jelonek L., Dieckmann M. A., Beyvers S., Blom J., Goesmann A. (2021). **Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification**. *Microbial Genomics*, 7(11). https://doi.org/10.1099/mgen.0.000685.
6. **Sensitive protein alignments at tree-of-life scale using DIAMOND**.Buchfink B, Reuter K, Drost HG. 2021.*Nature Methods* 18, 366–368 (2021). https://doi.org/10.1038/s41592-021-01101-x.
7. **CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats.** Bland, C., Ramsey, T.L., Sabree, F. *et al. BMC Bioinformatics* **8**, 209 (2007). https://doi.org/10.1186/1471-2105-8-209
8. **Tandem repeats finder: a program to analyze DNA sequences.** G. Benson(1999). *Nucleic Acids Research*, Vol. 27,No. 2, pp. 573-580.
9. **eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale**. Carlos P. Cantalapiedra, Ana Hernandez-Plaza, Ivica Letunic, Peer Bork, Jaime Huerta-Cepas. 2021.*Molecular Biology and Evolution*, msab293, https://doi.org/10.1093/molbev/msab293.
10. **eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses**. Jaime Huerta-Cepas, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, SofiaK Forslund, Helen Cook, Daniel R Mende, Ivica Letunic, Thomas Rattei, LarsJ Jensen, Christian von Mering, *Peer Bork Nucleic Acids Res*. 2019 Jan 8;47(Database issue): D309–D314. doi: 10.1093/nar/gky1085.
11. **KEGG as a reference resource for gene and protein annotation.** Kanehisa, Minoru, et al. *Nucleic acids research* 44.D1 (2016): D457-D462.
12. **Expanded microbial genome coverage and improved protein family annotation in the COG database.**Galperin, Michael Y., et al. *Nucleic acids research* (2014): gku1223.
13. **Gene Ontology：tool for the unification of biology.** Bard J, Winter R (2000) *Nat Genet*. 25:25-29.


# 6 联系我们

西安浩瑞基因成立于2019年，公司引入了三代测序平台--3台PacBioRevio和7台Sequell设备，致力于深耕动植物基因组学、转录组和微生物组学研究的科研技术服务。2024年，与华大智造携手共建西北首家DCSLab组学前沿实验室，引入DNBSEO-T7测序平台，开展基于二代测序的单细胞转录组、时空转录组等前沿技术服务。凭借专业的一站式多组学技术，为广大科研客户提供专业、高效、可靠的组学科研技术服务。

## 联系方式

热线电话：+86 029-89303503

官方网站：www.xahorizon.cn

邮　箱：project@xahorizon.cn

地　址：陕西省西安市沣东新城中兴深蓝科技产业园A区2号楼3层