枯草芽孢杆菌群比较基因组分析

1 前言

2 材料与方法

2.1 数据收集

2.2 基因组注释

2.3 功能基因注释

2.4 泛基因组分析

2.5 表面素合成基因与调控的识别与比对

2.6 基因频率分布热图的构建

3 结果与分析

3.1 基因组注释

3.2 功能基因注释

3.3 泛基因组结果分析

3.3.1 Genome gerneral infotmation

3.4 表面素合成基因 *srfA* 操纵子及调控基因分布分析

4 讨论

摘要

枯草芽孢杆菌（*Bacillus subtilis*）是一类在工业和环境科学中具有广泛应用的细菌。为了深入理解其适应性和生物技术潜力，我们对从 NCBI（https://www.ncbi.nlm.nih.gov/，访问于 2024 年 2 月 24 日）获取的 338 个枯草芽孢杆菌基因组进行了全面的比较基因组分析。

本研究旨在通过结构和功能注释、泛基因组分析、比较基因组分析，探究枯草芽孢杆菌的遗传多样性，并特别关注表面素合成基因簇的分布情况。

利用高通量基因组测序数据，我们对选定的枯草芽孢杆菌群进行了基因组组装、注释和比较分析。通过基因簇分析，我们鉴定了表面素合成相关的 *srfA* 操纵子基因，并与参考序列进行了比对。

我们的分析揭示了枯草芽孢杆菌群中表面素合成基因簇的多样性和分布特征。发现不同菌株间 srfA 操纵子的序列变异可能影响表面素的合成效率。此外，泛基因组分析扩展了我们对枯草芽孢杆菌功能性基因内容的认识。

本研究为枯草芽孢杆菌的生物技术应用提供了新的遗传资源，特别是在表面素的生物合成方面。通过比较基因组分析，本文为理解枯草芽孢杆菌的生态适应性和工业应用潜力提供了分子层面的见解。

关键词：枯草芽孢杆菌、比较基因组、表面素、*srfA* 操纵子、泛基因组

Abstract

Bacillus subtilis, a group of bacteria with broad applications in industry and environmental science, has been a subject of interest for its adaptability and biotechnological potential.We conducted a comprehensive comparative genomic analysis of 338 Bacillus subtilis genomes obtained from NCBI.

The study aimed to explore the genetic diversity of Bacillus subtilis through structural and functional annotation, pan-genome analysis, comparative genomic analysis, with a particular focus on the distribution of surfactin synthesis gene clusters.

Utilizing high-throughput genomic sequencing data, we performed genome assembly, annotation, and comparative analysis on the selected Bacillus subtilis group. Through gene cluster analysis, we identified the *srfA* operon genes associated with surfactin synthesis and compared them with reference sequences.

Our analysis revealed the diversity and distribution characteristics of surfactin synthesis gene clusters within the Bacillus subtilis group. Sequence variations in the *srfA* operon among different strains were found to potentially affect surfactin synthesis efficiency. Additionally, pan-genome analysis expanded our understanding of the functional gene content in Bacillus subtilis.

This study provides new genetic resources for the biotechnological application of Bacillus subtilis, especially in the biosynthesis of surfactin. Comparative genomic analysis offers insights into the ecological adaptability and industrial potential of Bacillus subtilis at the molecular level.

Keywords: *Bacillus subtilis*, Comparative Genomics, Surfactin, *srfA* Operon, Pan-Genome

1 前言

枯草芽孢杆菌（*Bacillus subtilis*）作为一类具有高度适应性和遗传多样性的细菌，在全球生态系统中扮演着重要角色。它们不仅在土壤生态循环中发挥着关键作用，而且在工业生物技术领域，尤其是在酶的生产、生物修复和作为生物控制剂方面展现出广泛的应用潜力。随着高通量测序技术的进步，我们现在能够对大量枯草芽孢杆菌基因组进行深入分析，以揭示其生物学特性和进化历史。 本研究旨在通过对从 NCBI 下载的 338 个枯草芽孢杆菌基因组进行全面的结构和功能注释、泛基因组分析以及比较基因组分析，来探索这一细菌群的遗传多样性和功能复杂性。特别地，我们关注于表面素（surfactin）合成基因簇的分布情况，表面素作为一种具有强大表面活性的脂肽，已被证明在生物降解、医学和增强石油回收等领域具有重要价值。 通过对参考序列与特定菌株的表面素合成 *srfA* 操纵子基因进行比对分析，我们期望揭示不同枯草芽孢杆菌株之间在表面素合成能力上的差异，并理解这些差异背后的分子机制。

2 材料与方法

2.1 数据收集

我们从国家生物技术信息中心（NCBI）收集了 338 个枯草芽孢杆菌的基因组数据，这些基因组的组装水平均为完整基因组。

2.2 基因组注释

收集到的基因组数据首先使用 Prokka 软件进行基因注释。Prokka 是一款广泛使用的开源工具，能够预测基因组中的蛋白质编码基因，并提供基因的功能注释。注

```
注释结果包括：faa，gff 等文件，做为下游分析的输入文件
```

2.3 功能基因注释

为了进一步了解基因组中的功能基因，我们使用了本地化的 eggNOG-mapper（基因组归类和功能注释器）对 Prokka 注释的基因组进行更深入的功能注释。eggNOG-mapper 通过将基因序列与 eggNOG 数据库进行比较，为每个基因分配一个功能类别。

2.4 泛基因组分析

Table x: pan-genome content

Gene Category	Strain Range (% of strains)	Number of Genes
Core genes	$99\% \leq$ strains $\leq 100\%$	303
Soft core genes	$95\% \leq$ strains $< 99\%$	2284
Shell genes	$15\% \leq$ strains $< 95\%$	2390
Cloud genes	$0\% \leq$ strains $< 15\%$	26584
Total genes	$0\% \leq$ strains $\leq 100\%$	31561

2.5 表面素合成基因与调控的识别与比对

通过 DIAMOND 比对从 NCBI 下载的 *srfA* 操纵子相关基因和调控基因序列，我们进行了功能基因分析，以确定这些基因在枯草芽孢杆菌群中的分布情况。

Table x Surfactin gene and regulate gene info

Gene symbol	Specie
srfA-A	bs
srfA-B	bs
srfA-C	ba
srfA-D	bs
abrB	bs
codY	bs
comA	bs
degU	bs
sinR	bs

1 表格需要后续优化

2.6 基因频率分布热图的构建

我们利用 DIAMOND 软件分析得到的表面素合成基因和 5 种调控基因的分布数据，构建了基因频率分布热图。这一步骤帮助我们直观地展示了不同基因在枯草芽孢杆菌群中的分布模式。

3 结果与分析

3.1 基因组注释

我们从 NCBI 数据库中收集了 338 个枯草芽孢杆菌 (Bacillus subtilis) 的基因组数据，这些基因组的组装水平均达到了基因组级别。利用 Prokka 软件对这些基因组进行了详细的注释，包括基因的位置、功能、以及可能的代谢途径。注释结果包括了基因的预测、基因产物的功能分类以及基因组中的 RNA 基因。

……

3.2 功能基因注释

通过 eggNOG-mapper，我们进一步对基因组中的蛋白质编码基因进行了功能注释。这一步骤帮助我们理解了枯草芽孢杆菌在不同生物学过程中的潜在角色。

Fig. x: Top 3 go enrich bar? dot? emap? cnet? tree? plot

```
338 个菌株基因组已经完成 eggnogMapper 注释
准备使用 clusterProfiler 做 GO KEGG 富集
如何展示结果?
    top3 Heatmap?
        选取所有基因组功能前 3 作热图?
    top10 bar? dot? emap? cnet? tree?
        考虑到基因组相似性高，选取所有物种功能富集前 10 作图?
可视化脚本位置:
/public/pipeline/besalttools/latest/public/gokegg_heatmap_clusterpro
file.sh
```

3.3 泛基因组结果分析

3.3.1 Genome gerneral infotmation

```
结果来自 roary github 仓库可视化脚本
```

```
Fig. x genome gerneral information
A Number of new genes bar plot
B Number of conserved genes bar plot
C Number of genes in the pan-genome
D Number of unique genes
E Number of blastp hits with different percentage identity
使用 BLASTP 搜索时，找到的与查询序列具有不同百分比相似性的命中数。BLASTP 是一种用于比较蛋白质序列的工具，它基于局部序列相似性来识别相似的蛋白质序列。具体来说，这个指标通常在 Roary 的输出中以一个表格形式出现，表格中会列出不同百分比的相似性阈值，以及在每个阈值下找到的命中数。例如，它可能会显示在 30%、50%、70% 和 90% 的相似性阈值下分别找到的命中数。这些信息对于理解不同基因组之间的相似性和差异性非常有用。较低的相似性阈值（如 30%）可能会产生更多的命中，但这些命中可能包含许多不相关的序列。而较高的相似性阈值（如 90%）则会产生较少的命中，但这些命中更可能是生物学上相关的序列。
```

```
如果在更高的相似性阈值下找到的命中数更多，这可能由几个原因导致：

基因组之间的高度相似性：如果比较的基因组之间存在高度的相似性，那么在较高的相似性阈值下找到的命中数自然会增加。这可能是因为这些基因组来自相近的物种或者具有共同的进化历史。

基因组的保守性：某些基因或蛋白质在进化过程中高度保守，即使在远缘物种之间也保持了较高的序列相似性。因此，这些保守基因在较高的相似性阈值下仍然能够被 BLASTP 识别出来。

基因组的完整性：如果被比较的基因组较为完整，那么在进行 BLASTP 搜索时更容易找到匹配的序列，尤其是在较高的相似性阈值下。
```

Table x: pan-genome content

Gene Category	Strain Range (% of strains)	Number of Genes
Core genes	$99\% \leq \text{strains} \leq 100\%$	303
Soft core genes	$95\% \leq \text{strains} < 99\%$	2284
Shell genes	$15\% \leq \text{strains} < 95\%$	2390
Cloud genes	$0\% \leq \text{strains} < 15\%$	26584
Total genes	$0\% \leq \text{strains} \leq 100\%$	31561

在对 338 个枯草芽孢杆菌（*Bacillus subtilis*）基因组进行深入的泛基因组分析后，我们鉴定了不同基因集的组成和它们在种群中的分布。泛基因组是指一个物种所有成员的基因总和，它包括核心基因组（存在于所有菌株中）和可变基因组（存在于部分菌株中）。这种分析有助于我们理解物种的遗传多样性和适应性。 我们的研究结果揭示了枯草芽孢杆菌基因组的复杂性，其中核心基因（存在于 99% 至 100% 的菌株中）共有 303 个，这些基因对维持基本生物学功能至关重要。软核心基因（存在于 95% 至 99% 的菌株中）数量较多，达到 2 284 个，它们可能涉及一些特定的生物学过程或环境适应性。壳层基因（存在于 15% 至 95% 的菌株中）有 2 390 个，而云层基因（存在于 0% 至 15% 的菌株中）数量最多，达到 26 584 个，这些基因可能与特定的生态位或菌株特异性功能相关。总体而言，我们鉴定了 31 561 个基因，这反映了枯草芽孢杆菌基因组的高度可变性和种群的遗传多样性。与先前的研究相比，我们观察到的核心基因数量较为保守，而云层基因的数量则显著增加，这可能与我们分析的菌株数量增加和使用的高通量测序技术有关。这些发现不仅加深了我们对枯草芽孢杆菌基因组结构的理解，而且为进一步探索其生物学特性和生态功能提供了新的视角。

```
图、表和文字表述会出现内容重叠
考虑使用图还是表展示数据
```

Fig. x Gene presence absence upset plot

```
数据量较大，不使用花瓣图
图右下角表示交集
左下角表示每个菌株基因量
右上柱状图表示该交集的数量
```

```
泛基因组分析跑了 roary 和 OrthoFinder，两个软件的结果有重叠
考虑可视化的难度选择不同的结果分析
```

……

3.4 表面素合成基因 *srfA* 操纵子及调控基因分布分析

为了研究枯草芽孢杆菌表面素合成的关键基因 *srfA* 操纵子在不同菌株中的分布情况，使用 DIAMOND 软件，我们针对枯草芽孢杆菌群中的表面素合成关键基因 *srfA* 操纵子进行了同源性搜索。通过这种方法，我们能够确定 *srfA* 操纵子在枯草芽孢杆菌群中的分布模式。

……

通过对比对结果的分析，我们发现 *srfA* 操纵子在枯草芽孢杆菌群中的分布呈现出显著的多样性。一些菌株中 *srfA* 操纵子的存在表明它们可能具有合成表面素的潜力，这是一种重要的生物膜形成因子，对细菌的粘附和生物膜形成具有重要作用。

……

利用 DIAMOND 得到的比对数据，我们进一步分析了 *srfA* 操纵子基因的频率分布，并生成了基因频率分布热图。热图显示了不同基因组中 *srfA* 操纵子基因出现的频率，颜色的深浅代表了基因出现的频率高低。通过热图，我们可以直观地观察到 *srfA* 操纵子在枯草芽孢杆菌群中的分布情况，以及不同菌株间基因频率的差异。

……

Fig. Gene frequency distribution heatmap

```
Y 轴聚类，X 轴不聚类
```

4 讨论

本研究的比较基因组分析揭示了枯草芽孢杆菌群中表面素合成基因的多样性，为理解其生物学功能和生态角色提供了新的见解。这些结果对于未来的微生物基因组研究和应用开发具有重要的指导意义。

我们的研究结果表明，枯草芽孢杆菌群中 srfA 操纵子的分布具有高度的变异性，这可能与菌株的生态适应性和进化历史有关。基因组功能基因分析揭示了枯草芽孢杆菌在生物膜形成和环境适应性方面的潜在机制。此外，基因频率分布热图为我们提供了一个直观的工具，用以比较和分析不同菌株间 srfA 操纵子的分布差异。

我们的分析结果与现有文献中关于枯草芽孢杆菌基因组多样性的研究相吻合。特别是，*srfA* 操纵子的变异与菌株的环境适应性之间的联系，已经在先前的研究中被提出。我们的研究通过大规模基因组分析，为这一领域提供了新的视角。

这些发现不仅增进了我们对枯草芽孢杆菌生物学特性的理解，而且为开发新的生物技术应用提供了可能性，如通过基因编辑技术改造枯草芽孢杆菌，以增强其在工业发酵或生物修复中的应用潜力。

……

参考文献

- Prokka v1.14.5:
  Seemann T. , Prokka: rapid prokaryotic genome annotation, *Bioinformatics* 2014 Jul 15;30(14):2068-9. PMID:24642063
- DIAMOND v0.9.25:
  Buchfink B, Xie C, Huson DH, Fast and sensitive protein alignment using DIAMOND, *Nature Methods* **12**, 59-60 (2015). doi:10.1038/nmeth.3176

附录

致谢