

An introduction of Integrated Prokaryotes Genome  
and pan-genome Analysis (IPGA) platform

## 泛基因组分析平台IPGA的应用



中国科学院微生物研究所  
Institute of Microbiology, Chinese Academy of Sciences



国家微生物科学数据中心  
National Microbiology Data Center



Dongmei Liu, Yifei Zhang, Guomei Fan, *et al.* 2022. IPGA: a handy integrated prokaryotes genome and pan-genome analysis web service. *iMeta* 1: e55. <https://doi.org/10.1002/imt2.55>

# Content 报告目录

1. Background 背景介绍
2. Pangenome research 泛基因组相关研究
3. Introduction 平台介绍
4. Methods 方法
5. Results 结果

6.  
Dongmei Liu, Yifei Zhang, Guomei Fan, *et al.* 2022. IPGA: a handy integrated prokaryotes genome and pan-genome analysis web service. *iMeta* 1: e55. <https://doi.org/10.1002/imt2.55>

# Background 背景介绍

**Pan-genome**是用于描述一个类群基因组信息的总和，比单一基因组更能展示出一个类群的基因多样性。

In the fields of molecular biology and genetics, a pan-genome is the entire set of genes from all strains within a clade. More generally, it is the union of all the genomes of a clade.

Perspective | [Published: 20 April 2022](#)

## The Human Pangenome Project: a global resource to map genomic diversity

[Ting Wang](#) , [Lucinda Antonacci-Fulton](#), [Kerstin Howe](#), [Heather A. Lawson](#), [Julian K. Lucas](#), [Adam M. Phillippy](#), [Alice B. Popejoy](#), [Mobin Asri](#), [Caryn Carson](#), [Mark J. P. Chaisson](#), [Xian Chang](#), [Robert Cook-Deegan](#), [Adam L. Felsenfeld](#), [Robert S. Fulton](#), [Erik P. Garrison](#), [Nanibaa' A. Garrison](#), [Tina A. Graves-Lindsay](#), [Hanlee Ji](#), [Eimear E. Kenny](#), [Barbara A. Koenig](#), [Daofeng Li](#), [Tobias Marschall](#), [Joshua F. McMichael](#), [Adam M. Novak](#), [the Human Pangenome Reference Consortium](#)  Show authors

[Nature](#) **604**, 437–446 (2022) | [Cite this article](#)

Article | [Open Access](#) | [Published: 15 January 2018](#)

## Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice

[Qiang Zhao](#), [Qi Feng](#), [Hengyun Lu](#), [Yan Li](#), [Ahong Wang](#), [Qilin Tian](#), [Qilin Zhan](#), [Yiqi Lu](#), [Lei Zhang](#), [Tao Huang](#), [Yongchun Wang](#), [Danlin Fan](#), [Yan Zhao](#), [Ziqun Wang](#), [Congcong Zhou](#), [Jiaying Chen](#), [Chuanran Zhu](#), [Wenjun Li](#), [Qijun Weng](#), [Qun Xu](#), [Zi-Xuan Wang](#), [Xinghua Wei](#), [Bin Han](#) & [Xuehui Huang](#) 

[Nature Genetics](#) **50**, 278–284 (2018) | [Cite this article](#)



Article | [Published: 13 May 2019](#)

## The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor

[Lei Gao](#), [Itay Gonda](#), [Honghe Sun](#), [Qiyue Ma](#), [Kan Bao](#), [Denise M. Tieman](#), [Elizabeth A. Burzynski-Chang](#), [Tara L. Fish](#), [Kaitlin A. Stromberg](#), [Gavin L. Sacks](#), [Theodore W. Thannhauser](#), [Majid R. Foolad](#), [Maria Jose Diez](#), [Jose Blanca](#), [Joaquin Canizares](#), [Yimin Xu](#), [Esther van der Knaap](#), [Sanwen Huang](#), [Harry J. Klee](#), [James J. Giovannoni](#)  & [Zhangjun Fei](#) 

[Nature Genetics](#) **51**, 1044–1051 (2019) | [Cite this article](#)



# Background 背景介绍

RESEARCH ARTICLE | BIOLOGICAL SCIENCES |



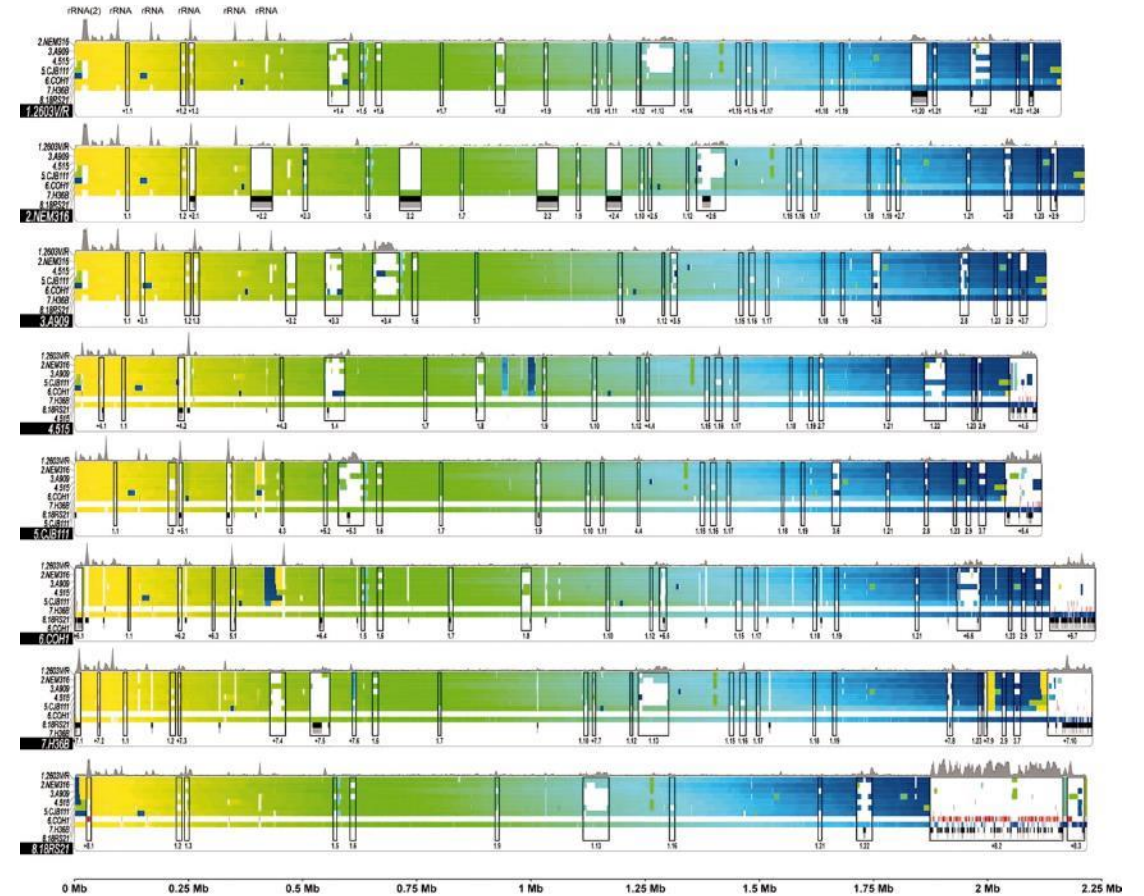
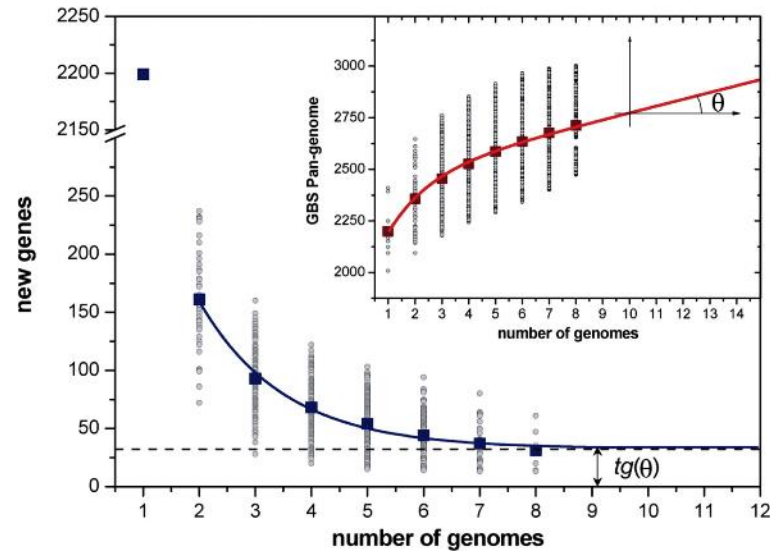
## Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial “pan-genome”

Hervé Tettelin, Vega Masignani, Michael J. Cieslewicz, +42, and Claire M. Fraser [Authors Info & Affiliations](#)

September 19, 2005 | 102 (39) 13950-13955 | <https://doi.org/10.1073/pnas.0506758102>

Tettelin H et al. PNAS, 2005

GBS pan-genome.

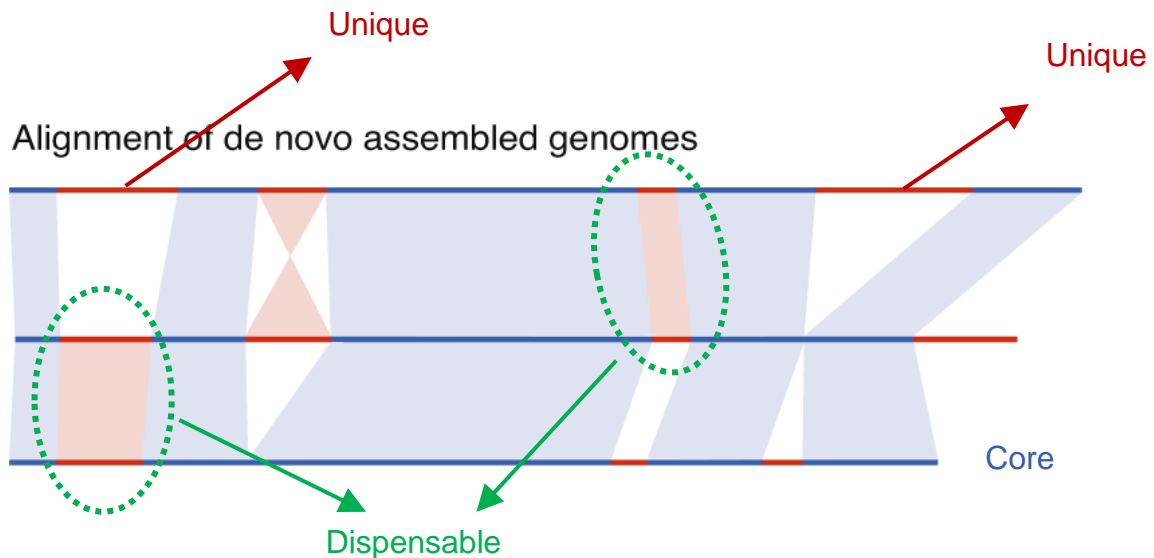


Synteny analysis between eight group B *Streptococcus* (GBS) genomes.

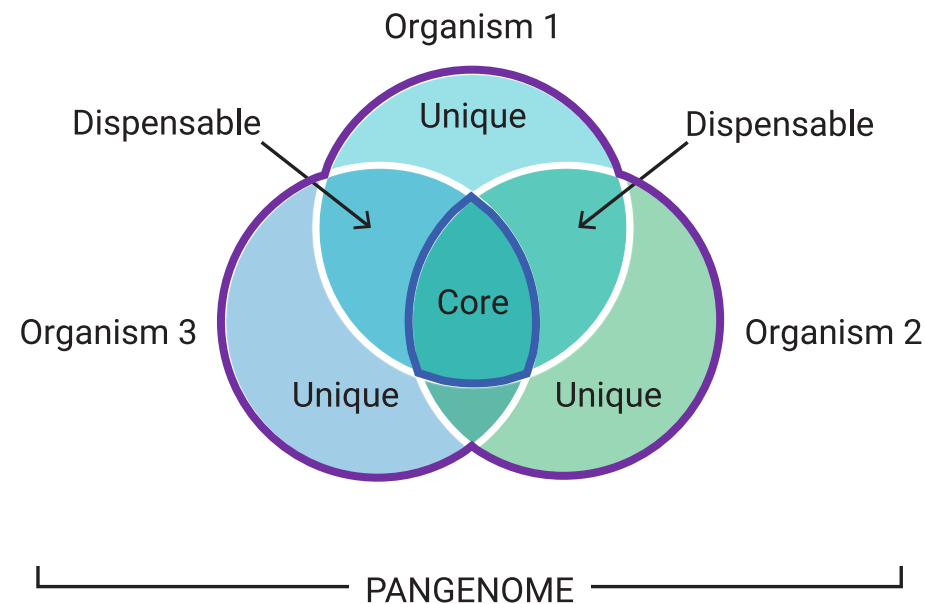
# Background 背景介绍

Pan-genome可以被拆分为核心泛基因组 (**core pangenome**)，由核心基因 (**core genes**) 组成；**shell pangenome**，由 **dispensable genes** 组成；**cloud pangenome**，由独有基因 (**unique gene**) 组成。

The pan-genome can be broken down into a "core pangenome" that contains genes present in all individuals (core genes), a "shell pangenome" that contains genes present in two or more strains (dispensable genes), and a "cloud pangenome" that contains genes only found in a single strain (unique gene).



Bayer, P.E. et al. Nat. Plants, 2020



# Background 背景介绍

## 描述Pangenome的方式

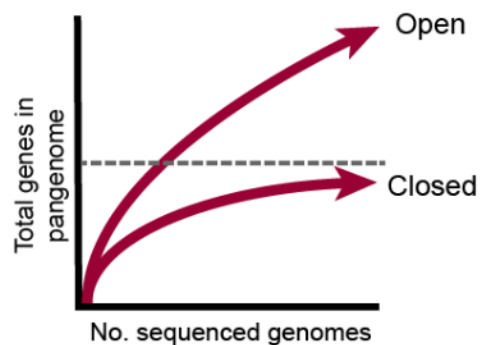
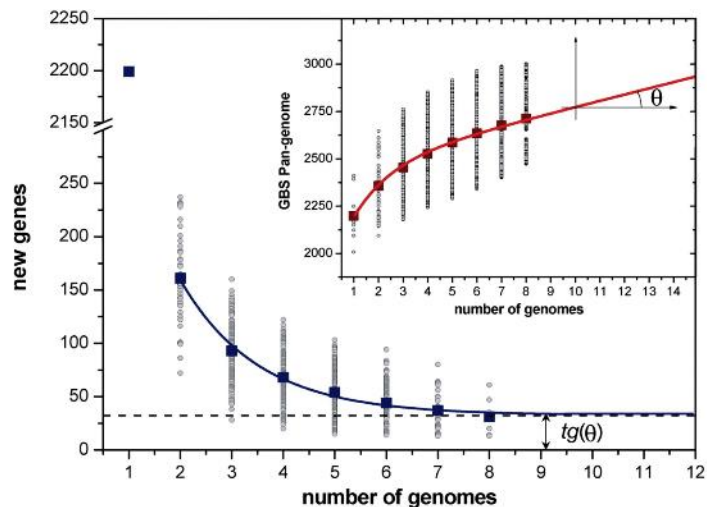


Table 1. Main software used for pan-genomic analysis and their respective algorithms.

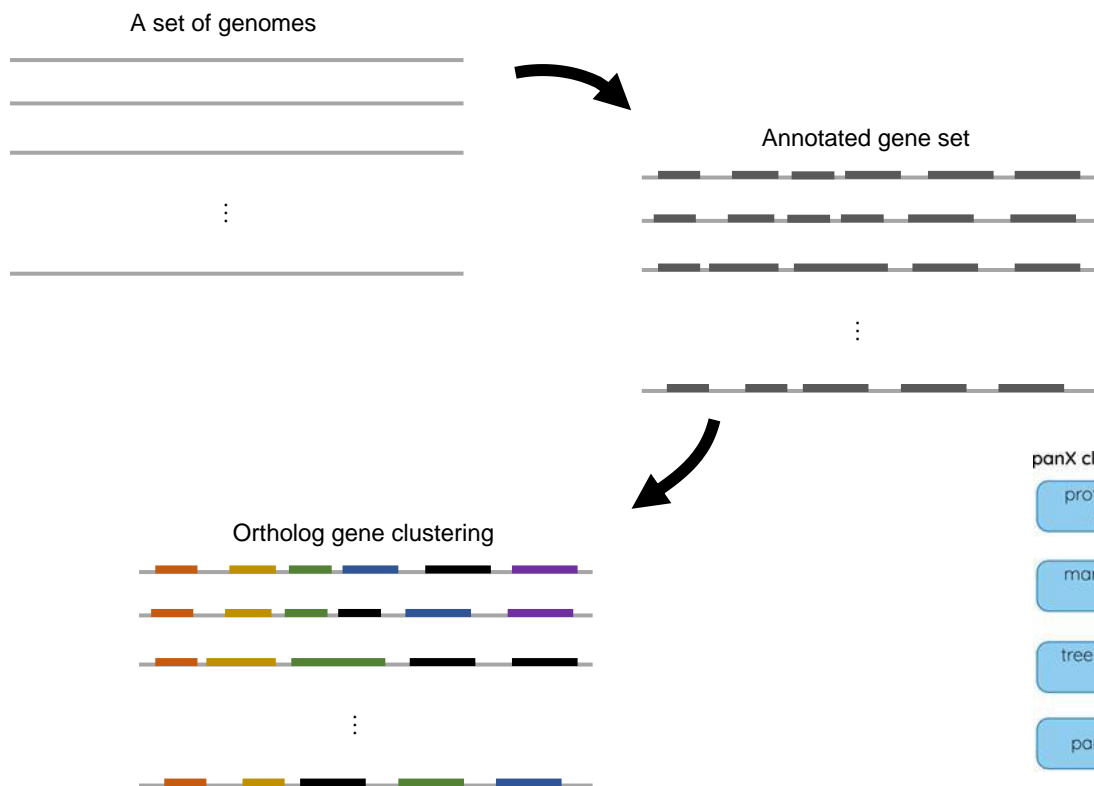
SOFTWARE	ORTHOLOGY ANALYSIS	PAN-GENOME DEVELOPMENT	CORE GENOME DEVELOPMENT	REFERENCES
BPGA	USEARCH, CD-HIT and OrthoMCL	Power-law regression	Exponential curve fit	Chaudhari et al <sup>48</sup>
EDGAR and EDGAR 2.0	Score ratio values	Heaps' law	Exponential curve fit	Blom et al <sup>53</sup>
GET_HOMOLOGUES	Bidirectional best-hit, COGtriangles, or OrthoMCL	<i>plot_pancore_matrix.pl</i>	<i>plot_pancore_matrix.pl</i>	Contreras-Moreira and Vinuesa <sup>26</sup>
PanDelos	Dictionary-based method	— <sup>a</sup>	— <sup>a</sup>	Bonnici et al <sup>54</sup>
Panseq	MUMmer and BLASTn	— <sup>a</sup>	— <sup>a</sup>	Laing et al <sup>55</sup>
PanWeb	PGAP	PGAP	PGAP	Pantoja et al <sup>62</sup>
PanX	Diamond and MCL	— <sup>a</sup>	— <sup>a</sup>	Ding et al <sup>56</sup>
PGAP	Inparanoid, MultiParanoid and Gene Family	Heaps' law	Exponential curve fit	Zhao et al <sup>61</sup>
PGAT	BLASTp (sequence alignment of >80% and sequence identity >91%-92%)	— <sup>a</sup>	— <sup>a</sup>	Brittnacher et al <sup>64</sup>
PGAweb	PGAP and PGAP-x modules	PGAP and PGAP-x modules	PGAP and PGAP-x modules	Chen et al <sup>63</sup>
Piggy <sup>b</sup>	Roary	— <sup>a</sup>	— <sup>a</sup>	Thorpe et al <sup>66</sup>
Roary	CD-HIT, BLAST and MCL	— <sup>a</sup>	— <sup>a</sup>	Page et al <sup>65</sup>

<sup>a</sup>Not mentioned in manuscript.

<sup>b</sup>Pan-genome analysis of intergenic regions.

# Background 背景介绍

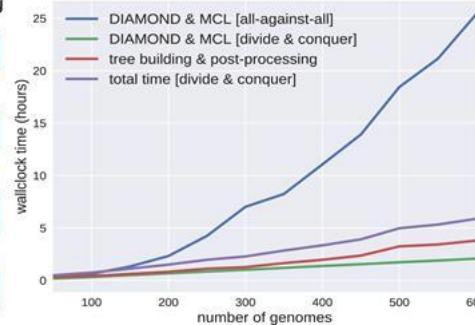
构建Pangenome的两个重要过过程和参数



如何找同源基因？

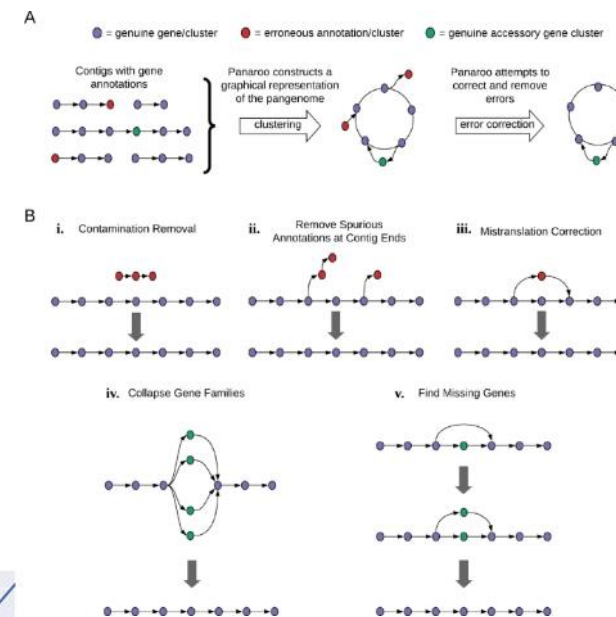
BLAST/DIAMOND  
CD-HIT  
Markov Clustering

panX clustering strategy



PanX analysis pipeline.

Ding W et al. Nucleic acids research, 2018

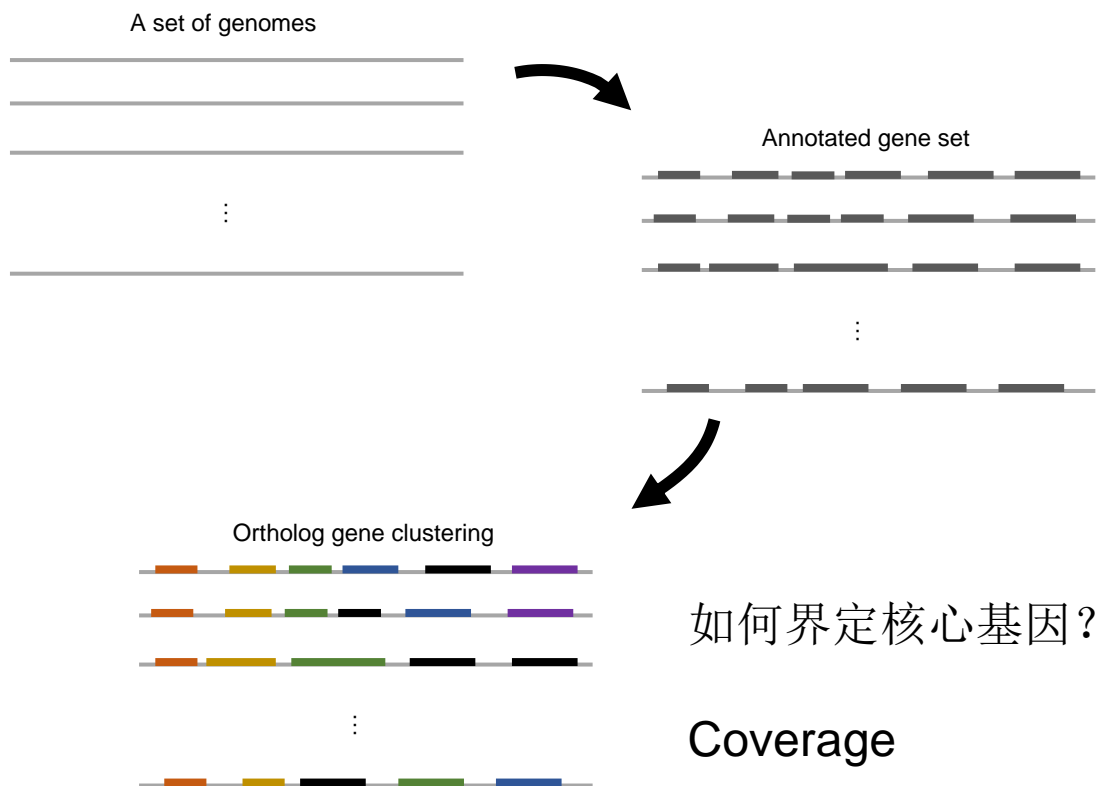


Panaroo gene annotation pipeline.

Tonkin-Hill G et al. Genome biology, 2020

# Background 背景介绍

构建Pangenome的两个重要过过程和参数



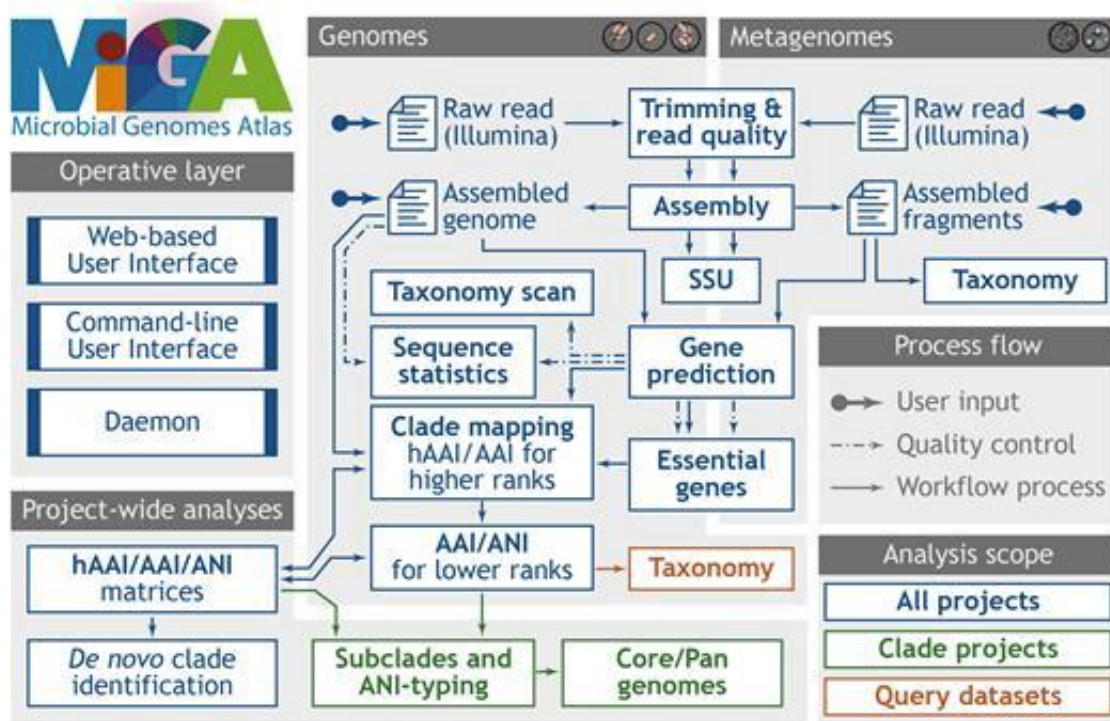
SOFTWARE	ORTHOLOGY ANALYSIS	REFERENCES
BPGA	USEARCH, CD-HIT and OrthoMCL	Chaudhari et al <sup>48</sup>
EDGAR and EDGAR 2.0	Score ratio values	Blom et al <sup>53</sup>
GET_HOMOLOGUES	Bidirectional best-hit, COGtriangles, or OrthoMCL	Contreras-Moreira and Vinuesa <sup>26</sup>
PanDelos	Dictionary-based method	Bonnici et al <sup>54</sup>
Panseq	MUMmer and BLASTn	Laing et al <sup>55</sup>
PanWeb	PGAP	Pantoja et al <sup>62</sup>
PanX	Diamond and MCL	Ding et al <sup>58</sup>
PGAP	Inparanoid, MultiParanoid and Gene Family	Zhao et al <sup>61</sup>
PGAT	BLASTp (sequence alignment of >80% and sequence identity >91%-92%)	Brittnacher et al <sup>64</sup>
PGAweb	PGAP and PGAP-x modules	Chen et al <sup>63</sup>
Piggy <sup>b</sup>	Roary	Thorpe et al <sup>66</sup>
Roary	CD-HIT, BLAST and MCL	Page et al <sup>65</sup>

<sup>a</sup>Not mentioned in manuscript.

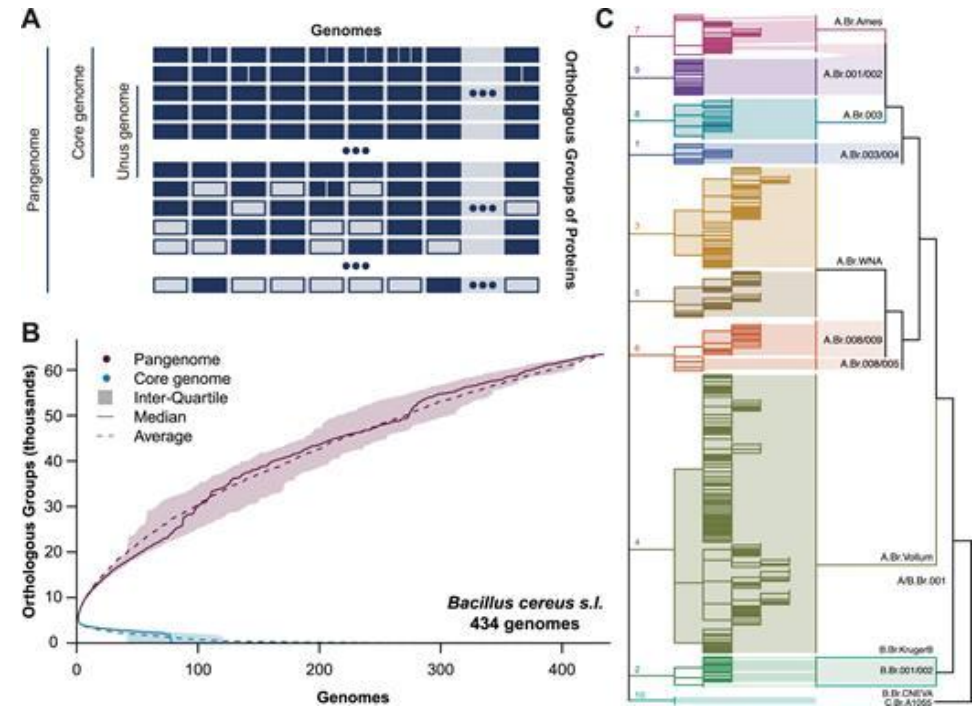
<sup>b</sup>Pan-genome analysis of intergenic regions.

# Pangenome research 相关研究

Microbial Genomes Atlas, 可以做给定基因组的分类阶元的确定、基因多样性评估和进化分支属性推断。



MiGA Workflow.



Pangenome profile of *Bacillus cereus* sensu lato (s.l.).

Rodriguez-R L M, et al. Nucleic acids research, 2018

# Pangenome research 相关研究

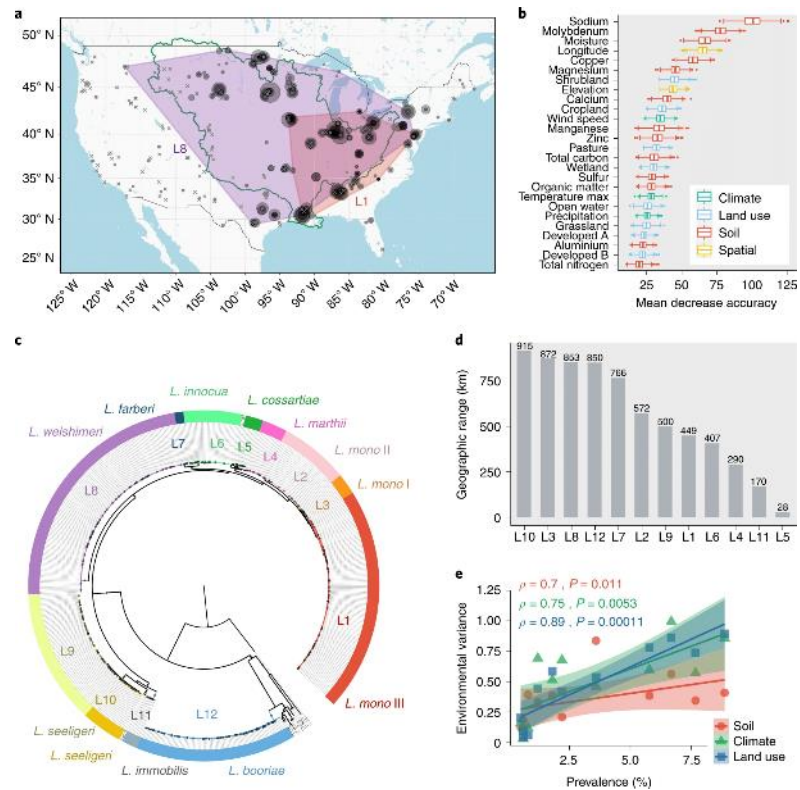
Article | Published: 15 July 2021

## Nationwide genomic atlas of soil-dwelling *Listeria* reveals effects of selection and population ecology on pangenome evolution

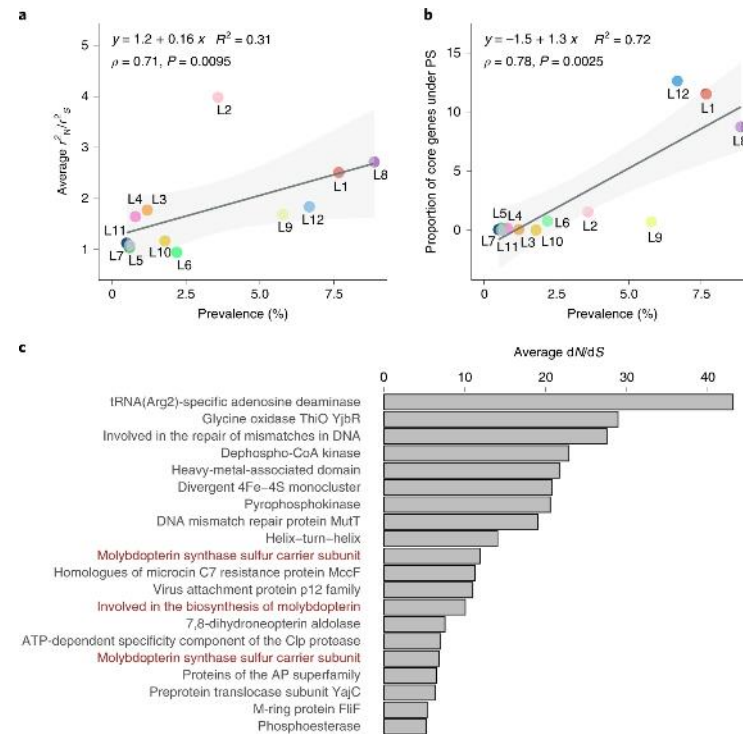
Jingqiu Liao, Xiaodong Guo, Daniel L. Weller, Shaul Pollak, Daniel H. Buckley, Martin Wiedmann & Otto X. Cordero

*Nature Microbiology* 6, 1021–1030 (2021) | Cite this article

- 作者通过全美范围的采样，获得了1854株李斯特菌。发现李斯特菌的存在严重受到环境因子的影响，尤其是土壤湿度、钼含量和盐离子浓度。
- 594个代表性菌株的全基因组分析可以把所有样本分到12个系统类群(phylogroups)，而且每个系统类群在生态地域的宽度和类型上都有很大差异。
- 普遍存在的“流行性”类群有更开放的泛基因组，并表现出较弱的连锁不平衡特性，表明这样的类群有更强的基因获得、丢失、以及交换能力。
- “流行性”类群也有相当大比例的基因受到正向选择，正向选择效应在核心基因组中表现更为明显，表明系别特异性的核心基因是环境适应的重要驱动因素。

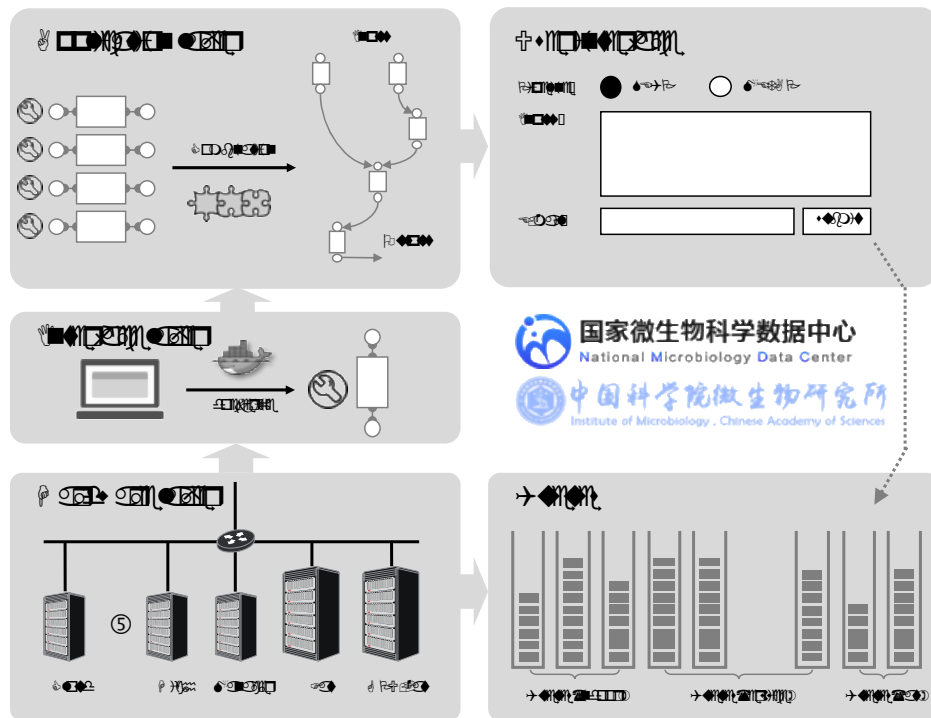


Nationwide distribution and ecological drivers of *Listeria*.



Stronger positive selection in more cosmopolitan phylogroup.

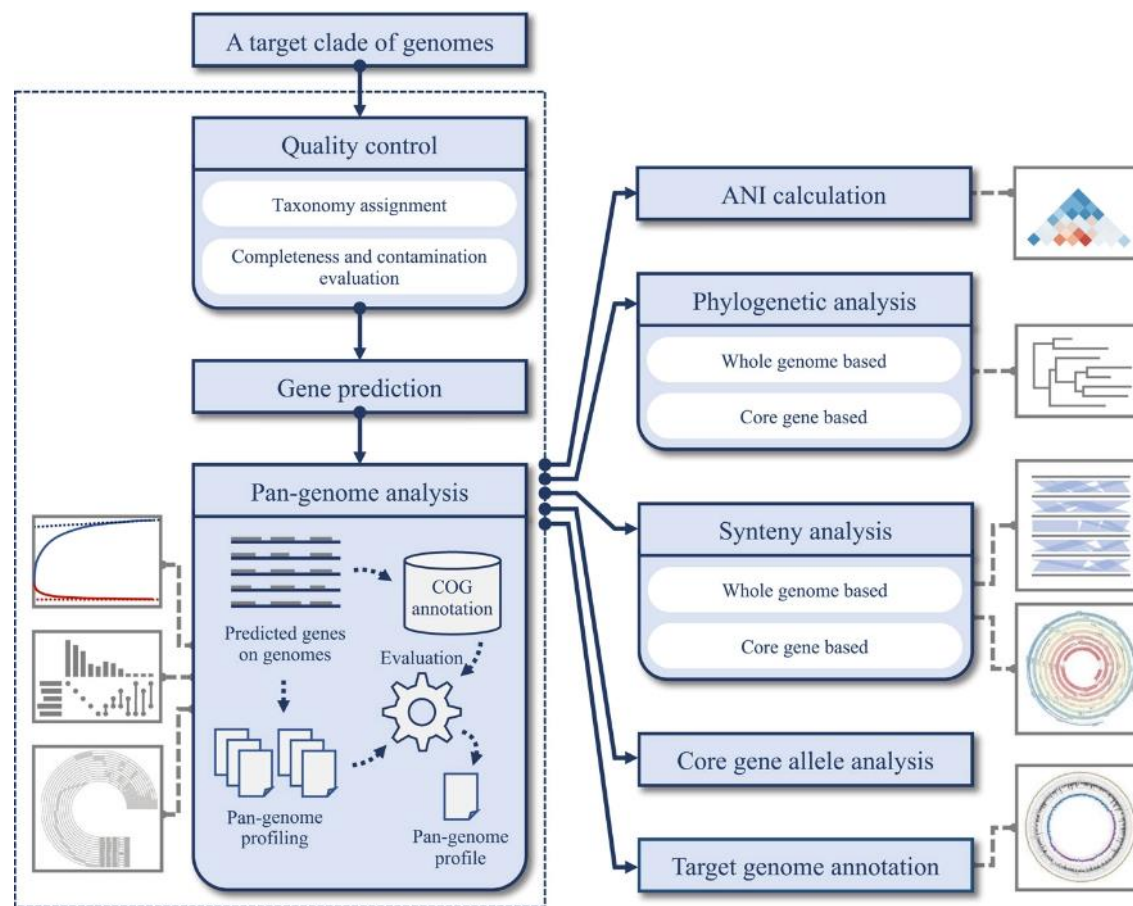
# Introduction 平台介绍



国家微生物数据中心的高性能计算云服务框架

Service status: **RUNNING** Running tasks: 0 Total tasks: 295

服务状态 运行中的作业



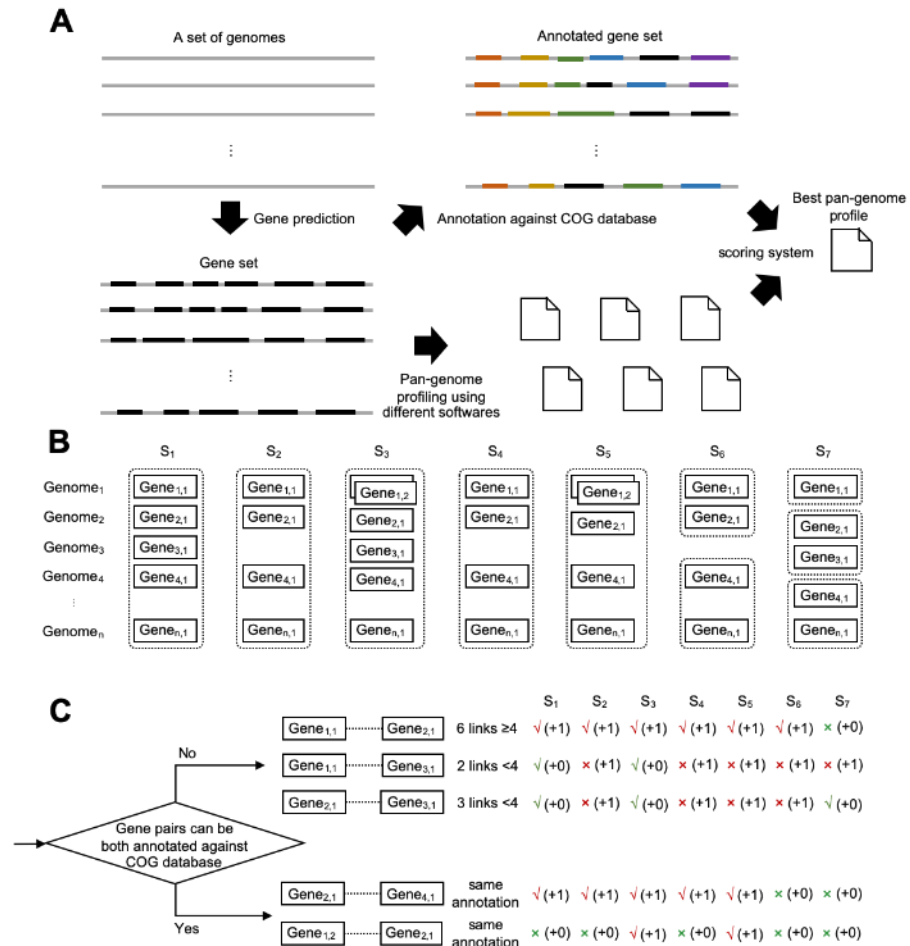
IPGA服务流程

# Methods 方法

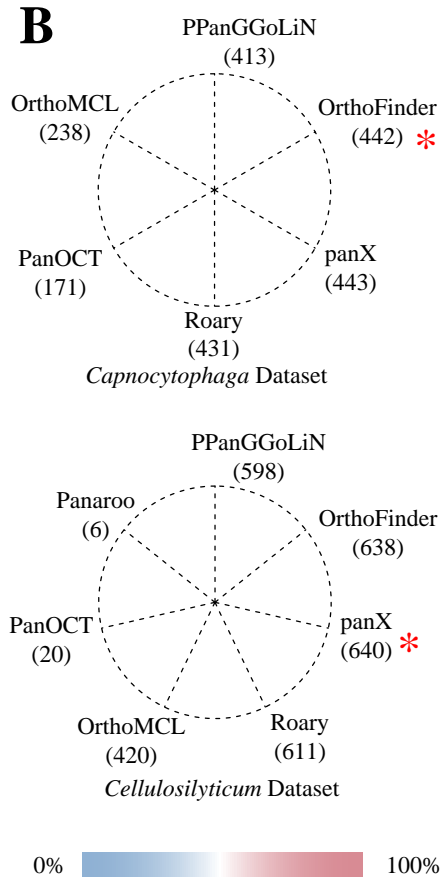
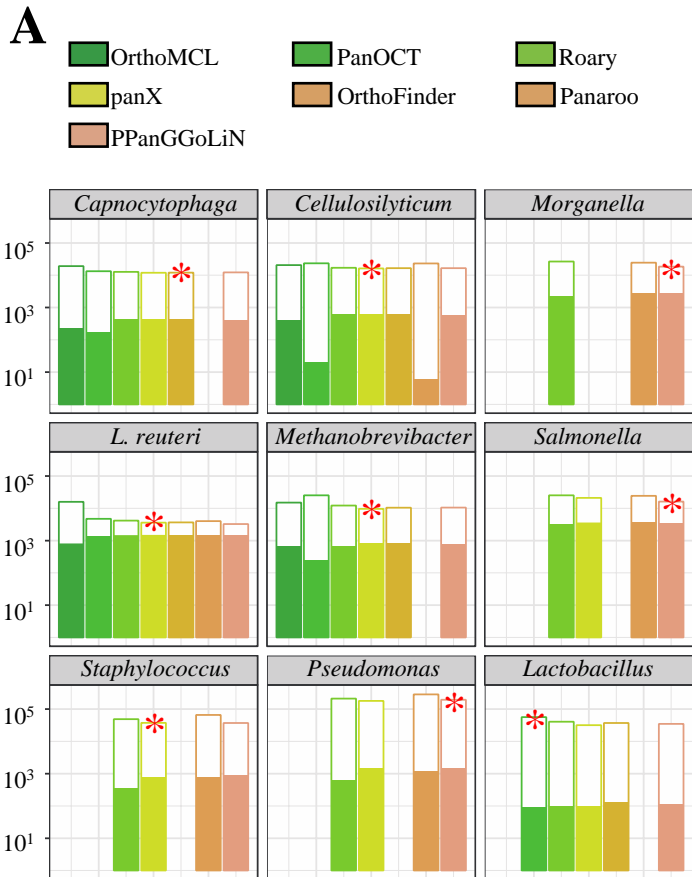
SOFTWARE	ORTHOLOGY ANALYSIS	REFERENCES
BPGA	USEARCH, CD-HIT and OrthoMCL	Chaudhari et al <sup>48</sup>
EDGAR and EDGAR 2.0	Score ratio values	Blom et al <sup>53</sup>
GET_HOMOLOGUES	Bidirectional best-hit, COGtriangles, or OrthoMCL	Contreras-Moreira and Vinuesa <sup>26</sup>
PanDelos	Dictionary-based method	Bonnici et al <sup>54</sup>
Panseq	MUMmer and BLASTn	Laing et al <sup>55</sup>
PanWeb	PGAP	Pantoja et al <sup>62</sup>
PanX	Diamond and MCL	Ding et al <sup>58</sup>
PGAP	Inparanoid, MultiParanoid and Gene Family	Zhao et al <sup>61</sup>
PGAT	BLASTp (sequence alignment of >80% and sequence identity >91%-92%)	Brittnacher et al <sup>64</sup>
PGAweb	PGAP and PGAP-x modules	Chen et al <sup>63</sup>
Piggy <sup>b</sup>	Roary	Thorpe et al <sup>66</sup>
Roary	CD-HIT, BLAST and MCL	Page et al <sup>65</sup>

<sup>a</sup>Not mentioned in manuscript.

<sup>b</sup>Pan-genome analysis of intergenic regions.

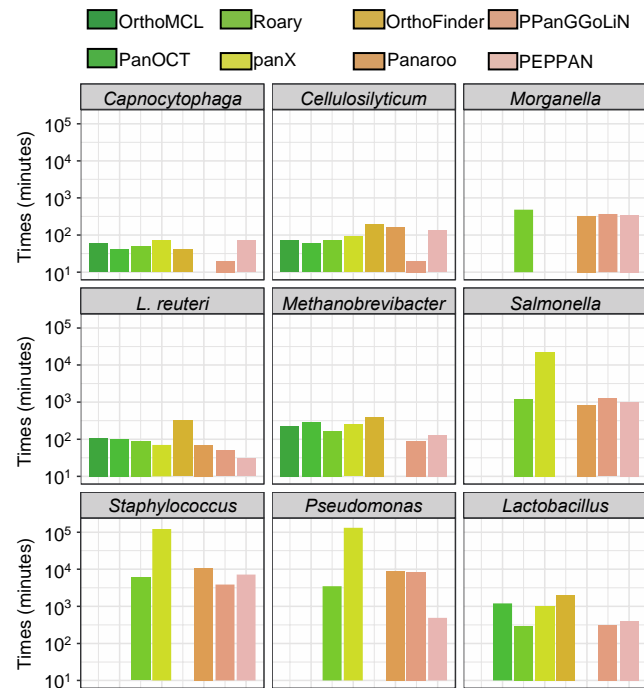


# Results 结果



orthoMCL	6793
PanOCT	9599
Roary	11046
panX	11365
orthoFinder	11524
PPanGGGoLiN	11134

orthoMCL	5338
PanOCT	1532
Roary	7237
panX	7497
orthoFinder	7432
Panaroo	1663
PPanGGGoLiN	7451



Pan-genome analysis procedure

PanOCT

OrthoFinder

OrthoMCL

PanX

Roary

(-ap)

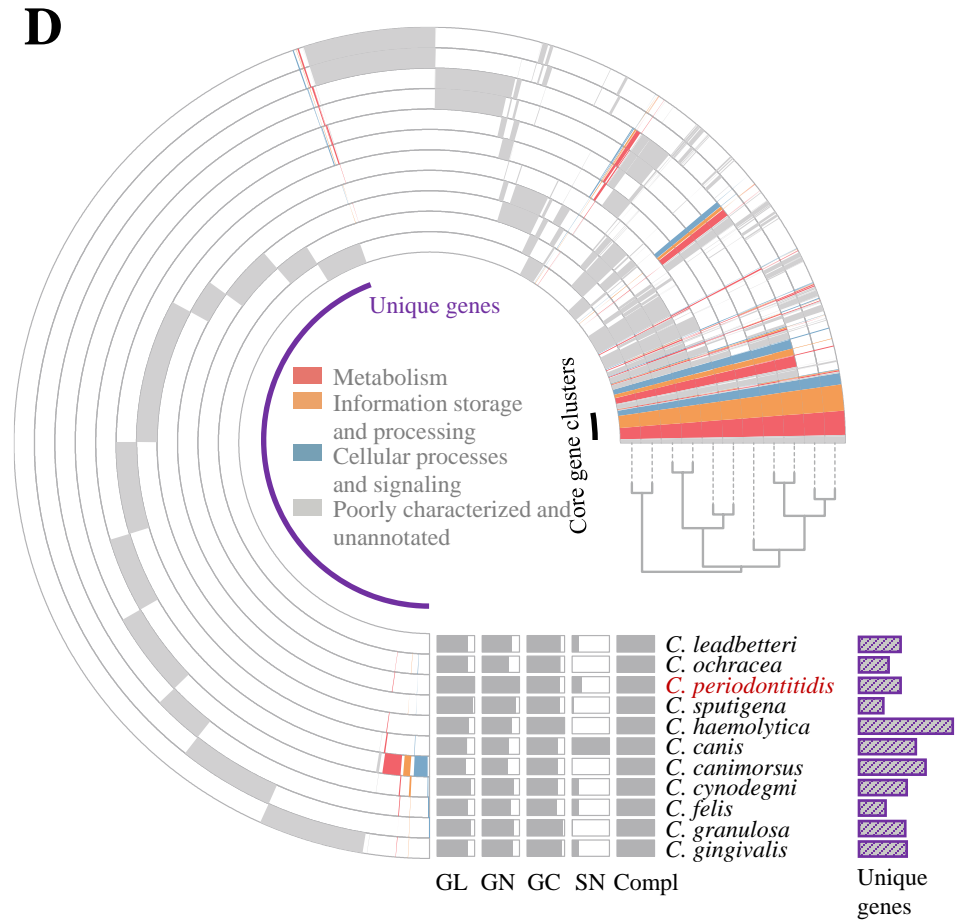
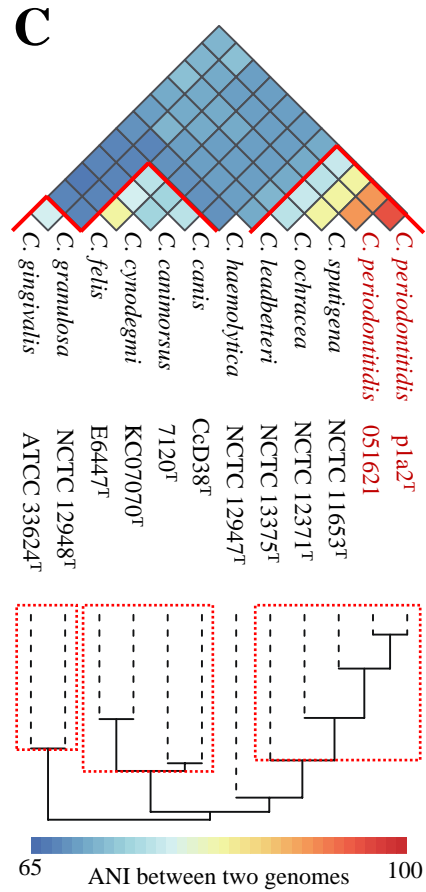
PPanGGOLIN

Panaroo

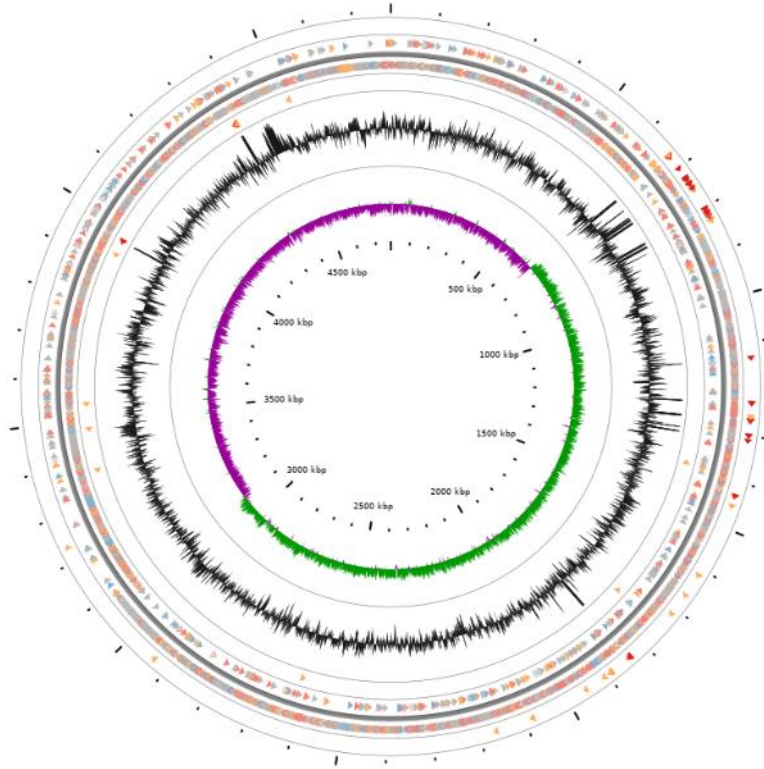
PEPPAN

execution speed

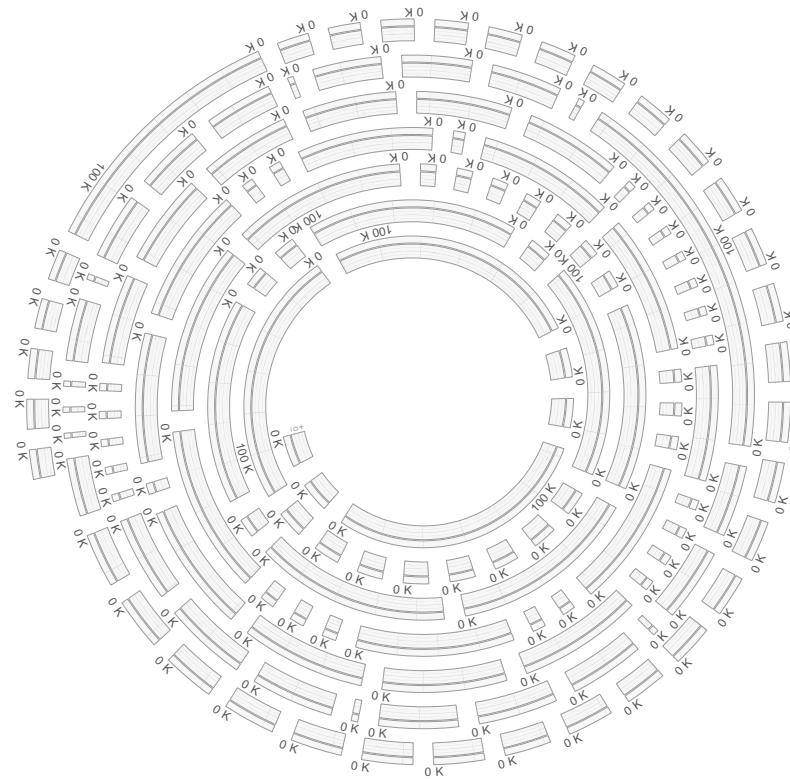
# Results 结果



# Results 结果

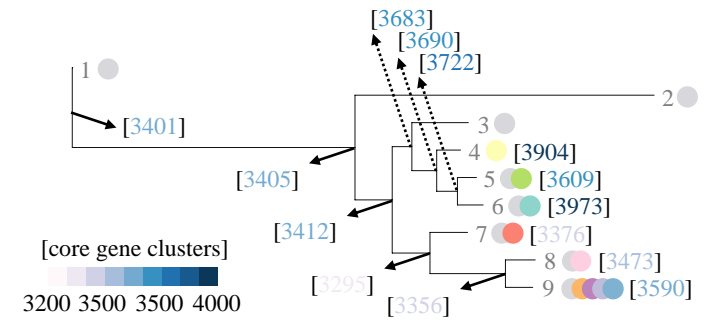
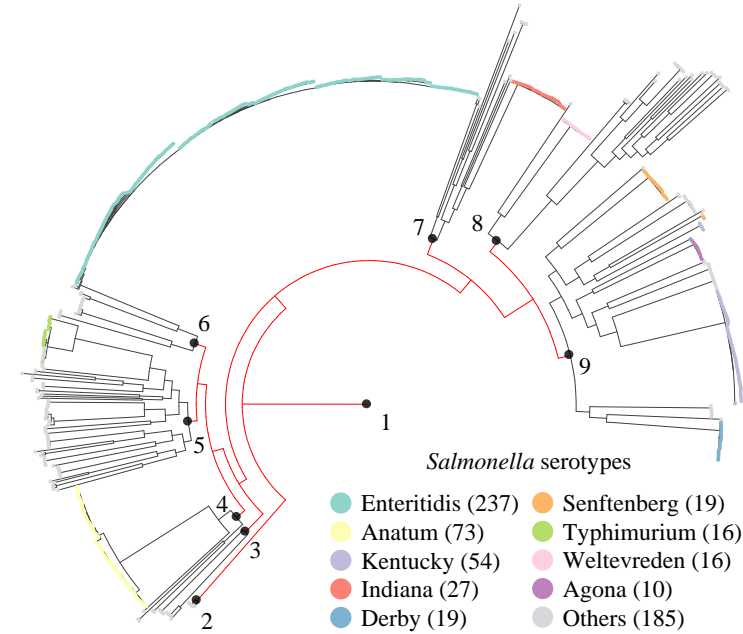
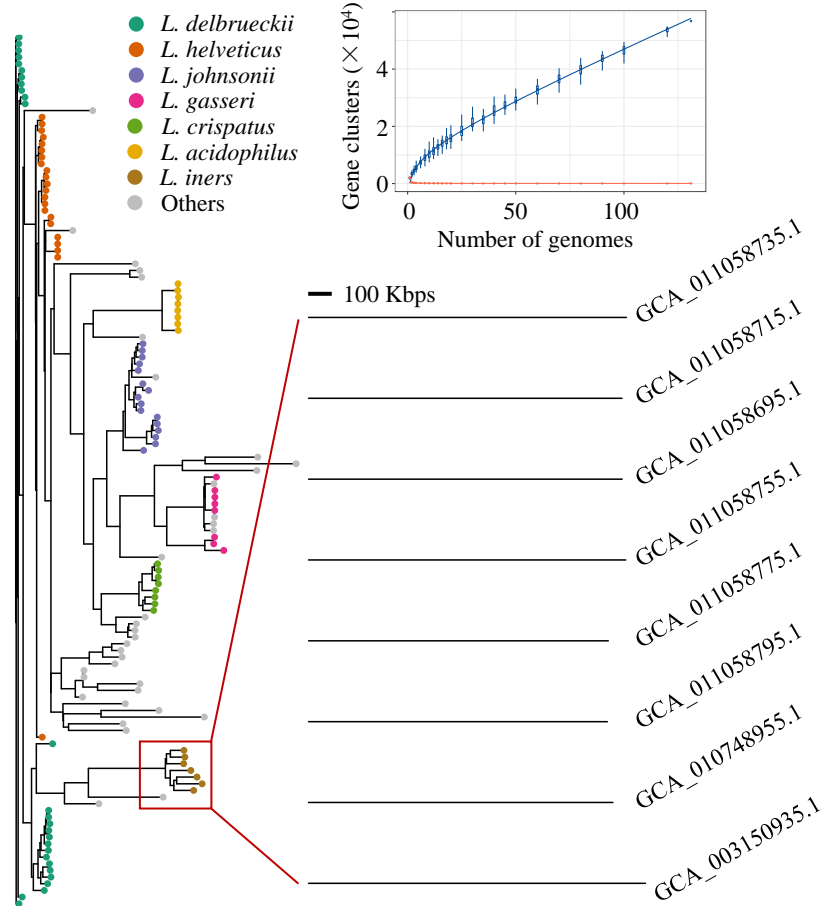


*Cellulosilyticum* sp. (strain WCF-2) isolated from cow feces  
(GCA\_003990395.1)



*Cellulosilyticum* sp. (MAG SIG270) isolated from goat feces  
(GCA\_015058045.1)

# Results 结果



# 结语

IPGA V1.09  
integrated prokaryotes genome  
and pan-genome analysis service

国家微生物科学数据中心  
National Microbiology Data Center

About Proceed Download Updates Help

Service status: **RUNNING** Running tasks: 1 Total tasks: 299

Home page Analysis Output query Download Manual Video Q&A

### IPGA

IPGA (<https://nmdc.cn/ipga/>) is a one-stop web service to analyze, compare, and visualize pan-genome as well as individual genomes, which avoid users to install any tools. IPGA features a score system that helps users to evaluate the reliability of pan-genome profiles generated by different packages. Thus, the users can choose the profiling method that is most suitable for their dataset for the following analysis. In addition, IPGA introduces several downstream comparative analysis module and genome analysis module to make users achieve diverse targets. All the tables and figures can be viewed and downloaded though result page which will be sent to your e-mail.

Detail ...

### WORKFLOW

A target clade of genomes

Contact shiwy@im.ac.cn, ma@im.ac.cn National Microbiology Data Center, Beijing 100101, China

<https://nmdc.cn/ipga/>

通讯方式:  
[shiwy@cau.edu.cn](mailto:shiwy@cau.edu.cn)  
[ma@im.ac.cn](mailto:ma@im.ac.cn)

中国科学院微生物科学数据中心  
Chinese academy of sciences  
Microbial resources and big data center

所有数据库 请输入您要搜索的内容 搜索

16,573万 访问次数 794个 数据库 52亿 数据条数 4TB 储存容量

### 重点数据库 / CHARACTERISTIC RESOURCES

- VarEPS 细菌基因组变异位点...
- gcType 全球模式菌株测序计划
- GCM 全球微生物物种目录
- CCINFO 全球保藏中心名录
- Refs 参考菌株数据库
- gcPathogen 全球病原菌株目录
- PQFUNGUS 植物内生菌数据库
- CASBRC 中国科学院战略生物资源库
- FOOD 食物链微生物全库...
- hGMB 人类肠道微生物资源
- FN 真菌微生物名录
- 4Gdb 全球基因组数据库

### 微生物数据资源 / MICROBIAL DATA RESOURCES

生物项目数据 159,556	生物样本数据 14,245,114	核糖序列数据 26,278,851	原基因组数据 1,054,615	基因组数据 81,373
宏基因组数据 81,373	多元组学数据 81,373	蛋白质序列数据 12,373	晶体结构数据 27	期刊附件数据 1,411