



中国科学院北京基因组研究所（国家生物信息中心）

BEIJING INSTITUTE OF GENOMICS CHINESE ACADEMY OF SCIENCES / CHINA NATIONAL CENTER FOR BIOINFORMATION

基因组学 - Genomics

第12章 泛基因组与宏基因组

章张 博士



中国科学院大学

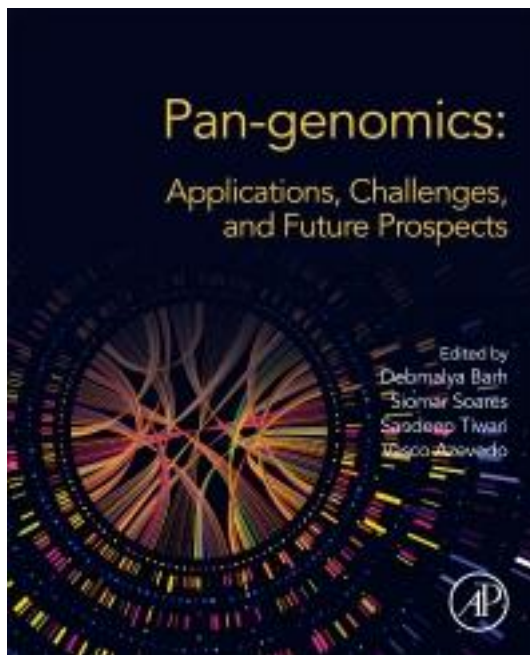
University of Chinese Academy of Sciences



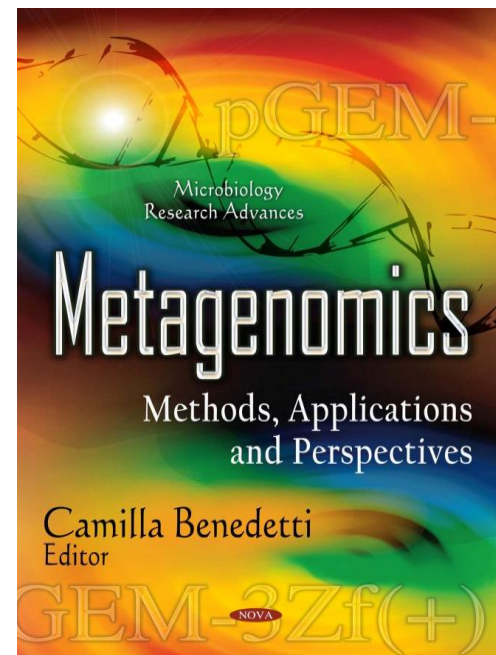
国家基因组科学数据中心
National Genomics Data Center

内容提纲

泛基因组Pangenome



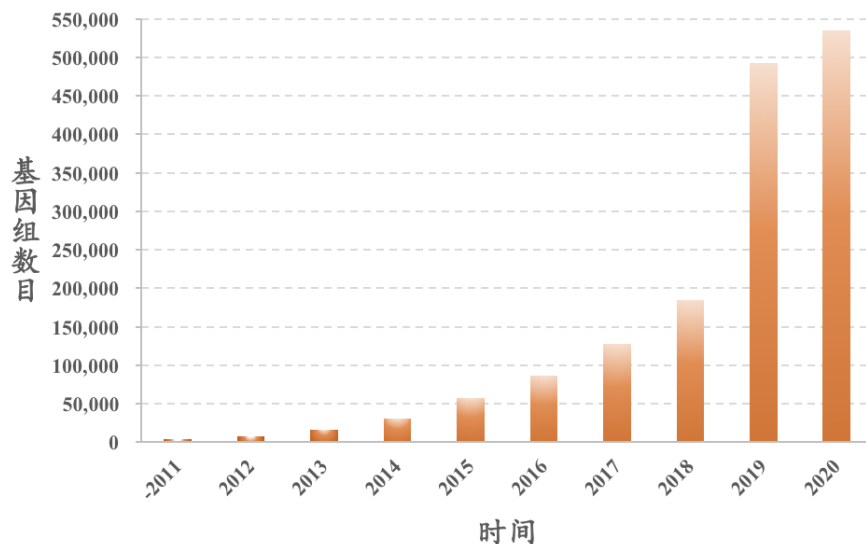
宏基因组Metagenome



Pangenome 泛基因组

同一物种的基因组序列

NCBI原核生物基因组数目统计



菌株数超过100株的菌种有195种

菌株数超过5株的菌种有2075种

Top10菌种及菌株数

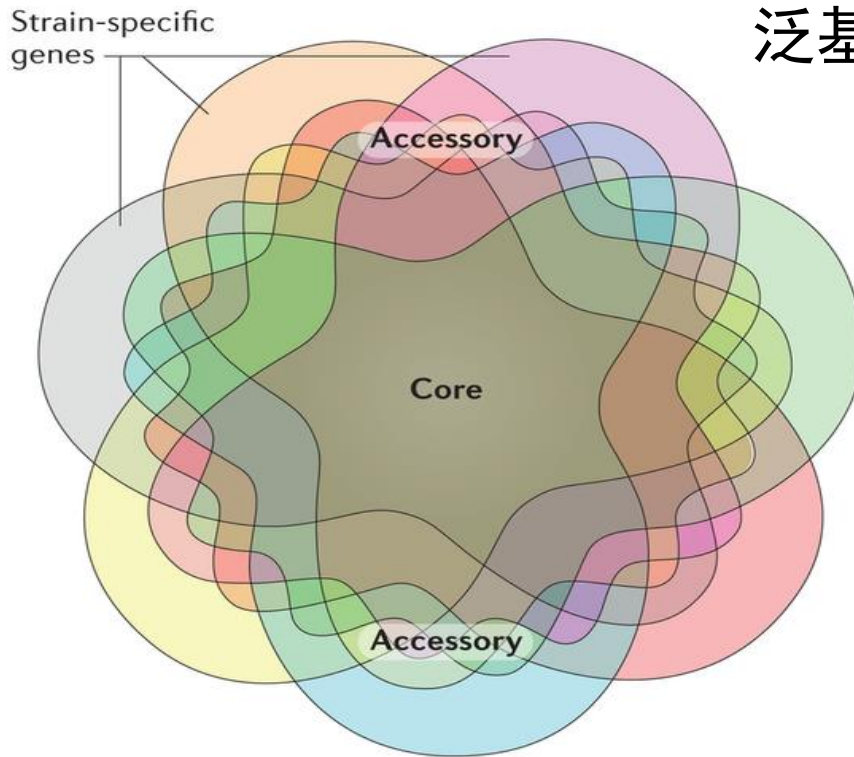
种名	菌株数
<i>Salmonella enterica</i>	207612
<i>Campylobacter jejuni</i>	33956
<i>Listeria monocytogenes</i>	31684
<i>Escherichia coli</i>	22839
<i>Streptococcus pneumoniae</i>	21571
<i>Campylobacter coli</i>	13234
<i>Staphylococcus aureus</i>	13147
<i>Klebsiella pneumoniae</i>	9866
<i>Mycobacterium tuberculosis</i>	6722
<i>Pseudomonas aeruginosa</i>	5395

为研究原核生物的基因功能、环境适应性及致病菌的致病耐药机制等问题提供了新的视角

Pangenome (泛基因组)

A **pangenome** refers to a collection of genomic sequences in the entire species or population rather than in a single individual.

泛基因组：某一物种全部基因的集合



- **Core genome:** containing genes present in all strains.
- **Accessory genome:** containing '**dispensable**' genes present in a subset of the strains.
- **Strain-specific genes:** present in only one strain.

Nature Reviews | **Genetics**

Escherichia coli genomes

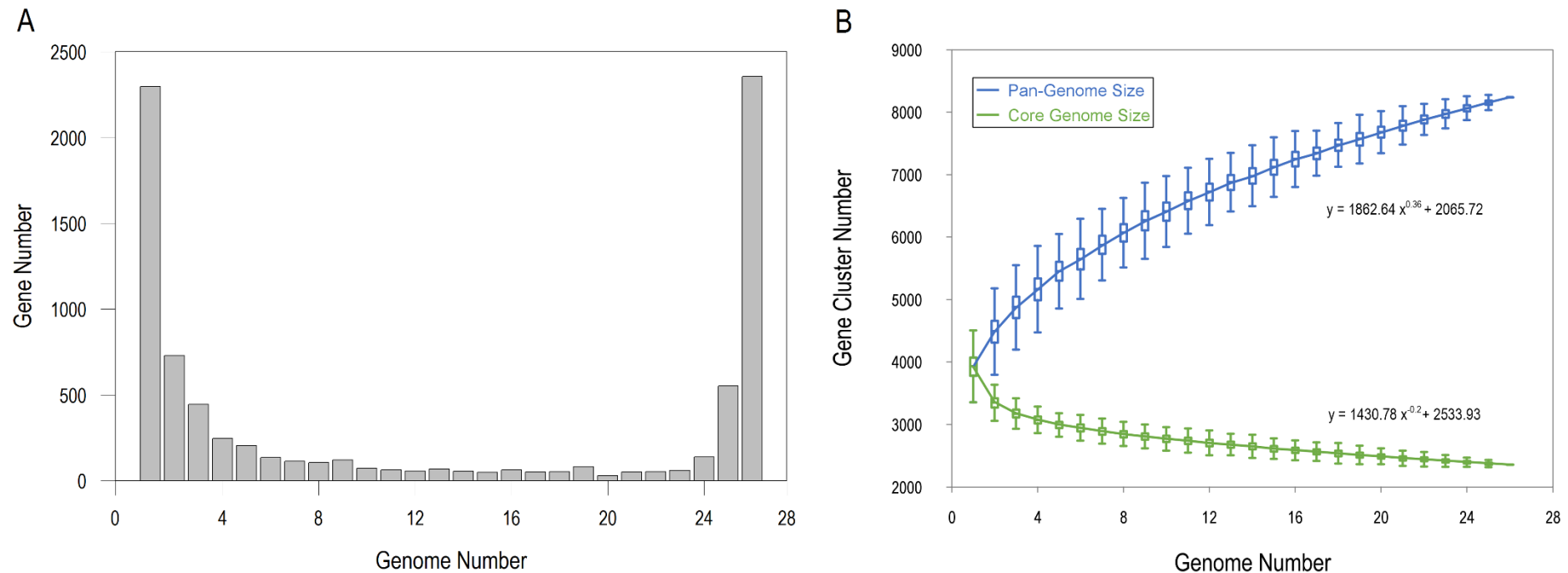
Overview (1); **Prokaryotes (19466)**; Plasmids (2759)

Filters 3

View 1 - 50 of 1,229										
#	Organism Name	Organism Groups	Strain	Assembly	Lev	Size	GC%	Genes	CDS	Release Da
1	Escherichia coli str. K-12 substr. MG1655	Bacteria;Proteobacteria;Ga	K-12 substr. MG1655	GCA_000005845.2	●	4.64	50.80	4,566	4,242	13-Oct-1998
2	Escherichia coli O157:H7 str. EDL933	Bacteria;Proteobacteria;Ga	EDL933	GCA_000006665.1	●	5.62	50.35	5,567	5,151	25-Jan-2001
3	Escherichia coli O157:H7 str. Sakai	Bacteria;Proteobacteria;Ga	Sakai substr. RIMD 0509952	GCA_000008865.2	●	5.59	50.45	5,417	5,155	07-Mar-2001
4	Escherichia coli CFT073	Bacteria;Proteobacteria;Ga	CFT073	GCA_000007445.1	●	5.23	50.50	5,023	4,681	06-Dec-2002
5	Escherichia coli str. K-12 substr. W3110	Bacteria;Proteobacteria;Ga	K-12 substr. W3110	GCA_000010245.1	●	4.65	50.80	4,485	4,236	21-Jan-2006
6	Escherichia coli UTI89	Bacteria;Proteobacteria;Ga	UTI89	GCA_000013265.1	●	5.18	50.61	5,147	4,863	05-Apr-2006
7	Escherichia coli 536	Bacteria;Proteobacteria;Ga	536	GCA_000013305.1	●	4.94	50.50	4,771	4,436	13-Jul-2006
8	Escherichia coli APEC O1	Bacteria;Proteobacteria;Ga	APEC O1	GCA_000014845.1	●	5.50	50.29	5,543	5,191	04-Oct-2006
9	Escherichia coli HS	Bacteria;Proteobacteria;Ga	HS	GCA_000017765.1	●	4.64	50.80	4,507	4,213	10-Sep-2007
10	Escherichia coli O139:H28 str. E24377A	Bacteria;Proteobacteria;Ga	E24377A	GCA_000017745.1	●	5.25	50.54	5,120	4,771	11-Sep-2007
11	Escherichia coli O157:H7 str. TW14588	Bacteria;Proteobacteria;Ga	TW14588	GCA_000155125.1	●	5.67	50.48	11,842	11,195	03-Mar-2008
12	Escherichia coli ATCC 8739	Bacteria;Proteobacteria;Ga	ATCC 8739	GCA_000019385.1	●	4.75	50.90	4,568	4,301	11-Mar-2008
13	Escherichia coli str. K-12 substr. DH10B	Bacteria;Proteobacteria;Ga	K-12 substr. DH10B	GCA_000019425.1	●	4.69	50.80	4,536	4,256	14-Mar-2008
14	Escherichia coli SMS-3-5	Bacteria;Proteobacteria;Ga	SMS-3-5	GCA_000019645.1	●	5.22	50.50	5,088	4,754	20-Mar-2008
15	Escherichia coli O157:H7 str. EC4115	Bacteria;Proteobacteria;Ga	EC4115	GCA_000021125.1	●	5.70	50.39	5,709	5,359	08-Oct-2008
16	Escherichia coli SE11	Bacteria;Proteobacteria;Ga	SE11	GCA_000010385.1	●	5.16	50.75	5,031	4,794	22-Oct-2008
17	Escherichia coli O127:H6 str. E2348/69	Bacteria;Proteobacteria;Ga	E2348/69	GCA_000026545.1	●	5.07	50.55	5,052	4,728	26-Nov-2008
18	Escherichia coli 55989	Bacteria;Proteobacteria;Ga	55989	GCA_000026245.1	●	5.15	50.70	5,110	4,767	16-Dec-2008
19	Escherichia coli IAI39	Bacteria;Proteobacteria;Ga	IAI39	GCA_000026345.1	●	5.13	50.60	4,950	4,608	16-Dec-2008
20	Escherichia coli S88	Bacteria;Proteobacteria;Ga	S88	GCA_000026285.2	●	5.17	50.66	5,052	4,772	16-Dec-2008
21	Escherichia coli IAI1	Bacteria;Proteobacteria;Ga	IAI1	GCA_000026265.1	●	4.70	50.80	4,524	4,311	16-Dec-2008
22	Escherichia coli UMN026	Bacteria;Proteobacteria;Ga	UMN026	GCA_000026325.2	●	5.36	50.64	5,185	4,949	17-Dec-2008
23	Escherichia coli ED1a	Bacteria;Proteobacteria;Ga	ED1a	GCA_000026305.1	●	5.21	50.70	5,276	4,864	17-Dec-2008
24	Escherichia coli LF82	Bacteria;Proteobacteria;Ga	LF82	GCA_000284495.1	●	4.77	50.70	4,651	4,434	23-Jan-2009
25	Escherichia coli BW2952	Bacteria;Proteobacteria;Ga	K-12 substr. BW2952	GCA_000022345.1	●	4.58	50.80	4,456	4,180	04-Jun-2009

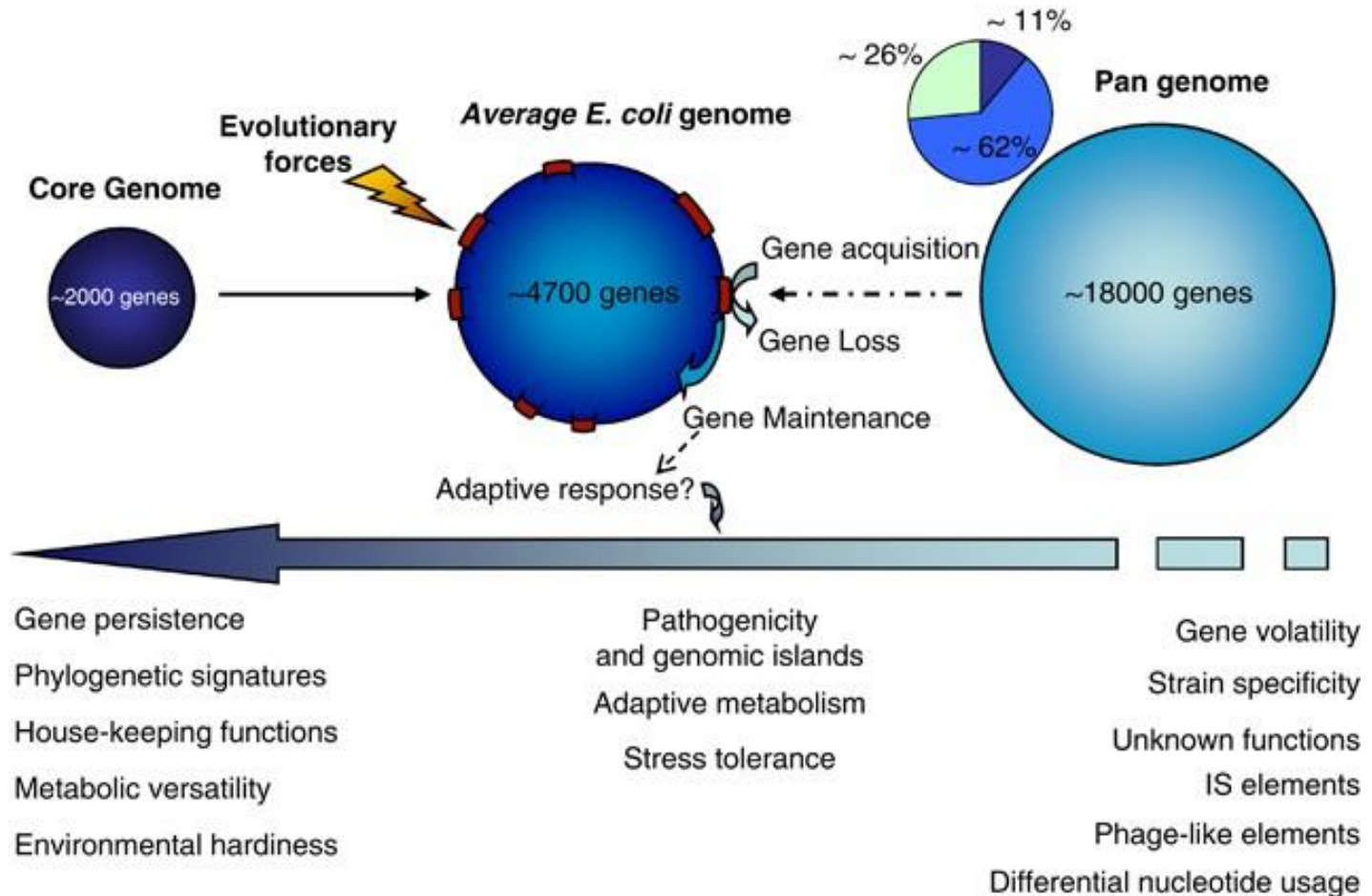
<https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/Escherichia%20coli>

E. coli Pangenome



Pangenome evidence for higher codon usage bias and stronger translational selection in core genes of *Escherichia coli*. *Frontiers in Microbiology* 2016

E. coli Pangenome



Survival of *Escherichia coli* in the environment: fundamental and public health aspects. *The ISME Journal* 2011

Core vs. dispensable genome

- 核心基因组：
 - 高度保守，在同一种细菌所有菌株中都存在
 - 保持生命体基本功能，代表物种的基本表型特征
- 非必须基因组：
 - 在细菌整个进化史上，为适应环境条件而不断发生水平转移与基因缺失
 - 使不同菌株间在表型、生活方式、代谢等方面发生明显的分化，最终表现为丰富的遗传多样性

Pangenomes arise due to **gene gain** by genomes from other species through **horizontal gene transfer** and differential **gene loss** among genomes.

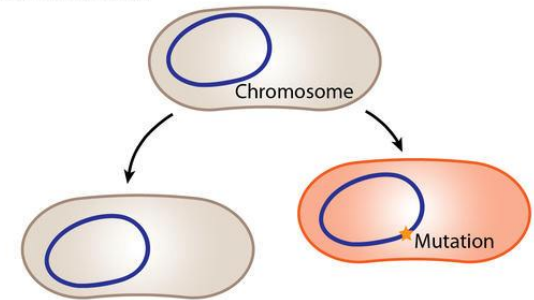
Horizontal Gene Transfer

水平基因转移：遗传物质传递给非其子代

- Horizontal gene transfer (HGT), first described in microorganisms in the late 1940s, is the **movement of genetic material** between organisms that are **not in a parent-offspring relationship**.
- HGT is an important factor in the **evolution** of many organisms and a widely recognized mechanism for **adaptation**.
- The **pan-genome** size of many microbial species illustrates the importance of **additive gene transfer**, which is the process of the **integration of novel genetic material** into a genome.
- Microbial **antibiotic resistance** and **pathogenicity** are associated with HGT.

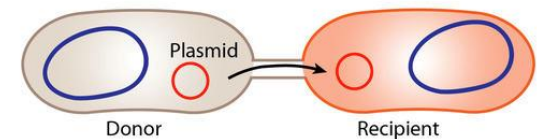
Horizontal gene transfer: building the web of life.
Nature Reviews Genetics 2015

A. Vertical evolution

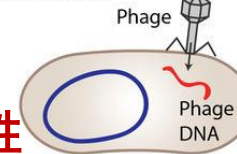


B. Horizontal evolution

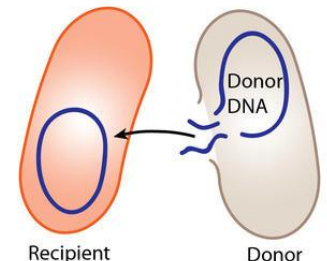
Conjugation from another bacterium



Transduction



Transformation

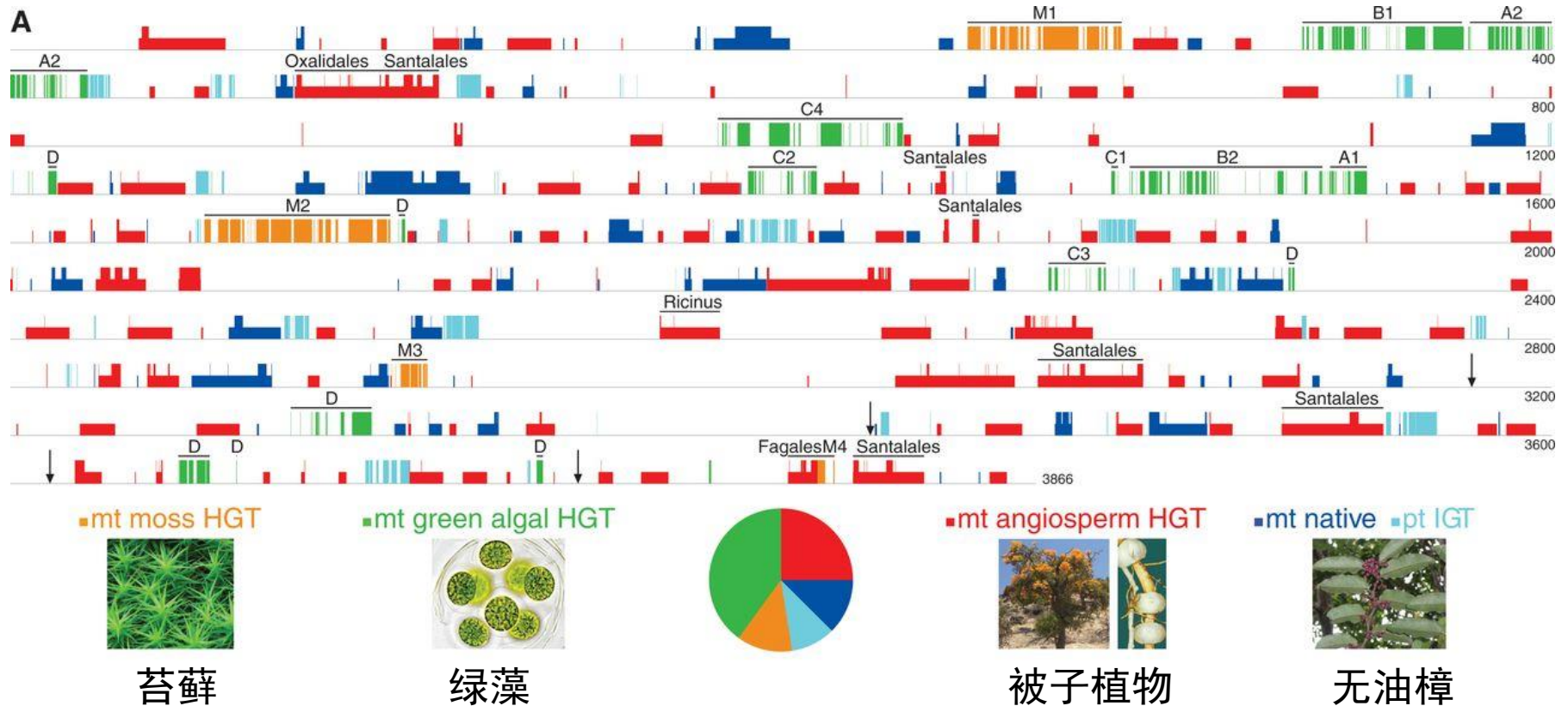


from a bacterial virus

from the environment

<https://www.futurelearn.com/courses/introduction-to-bacterial-genomics/0/steps/45331>

Horizontal Transfer of Entire Genomes via Mitochondrial Fusion in the Angiosperm Amborella



mt, mitochondrial; pt, plastid; Labeled black lines: horizontally transferred mitochondrial genomes.

M1 to M4: moss-derived genome; A1, A2, B1, B2, and C1 to C4: three green algal-derived

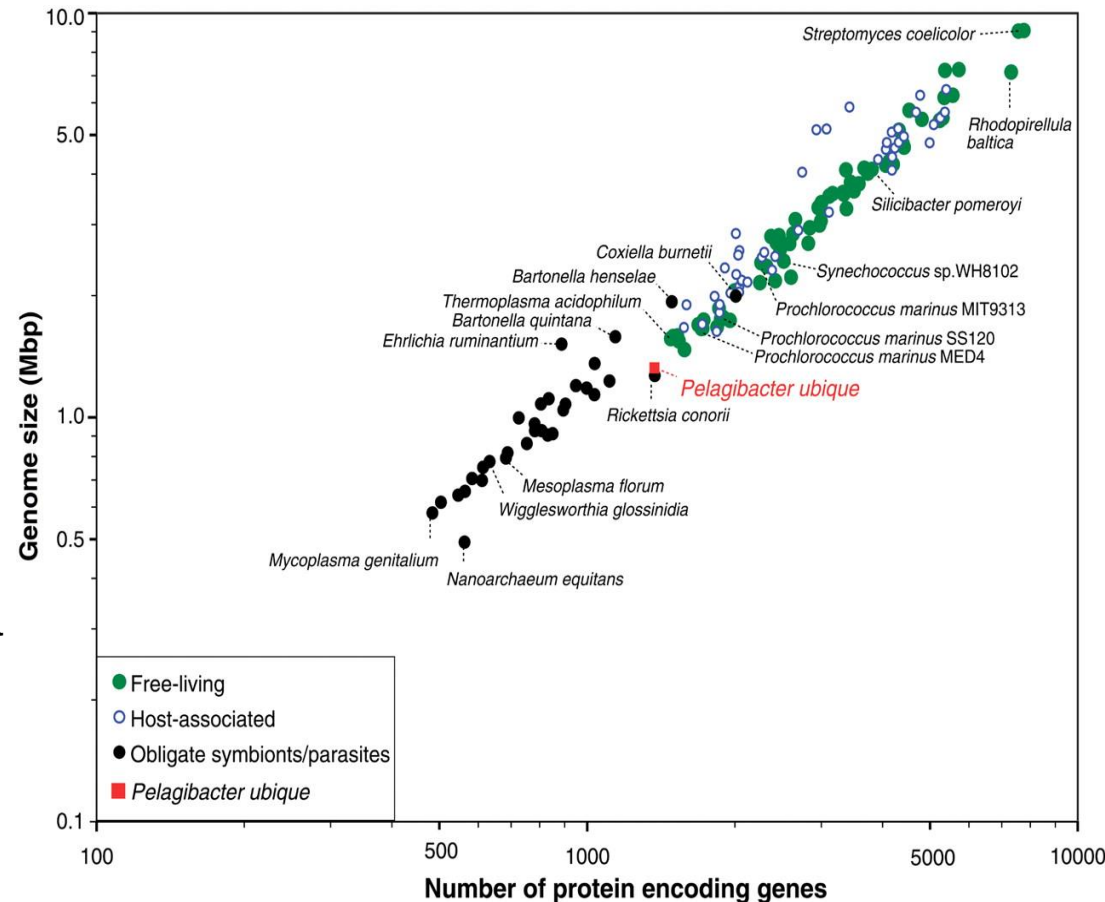
genomes. D: seven fragments of a partial genome from a fourth green algal donor.

Science 342, 1468 (2013)

完美基因组

海洋浮游菌 *Pelagibacter ubique*

- 基因组~1.3Mb, 含1354个蛋白质基因
- 基因组含有最简单的独立生存细胞所需基因
- *Pelagibacter ubique* has the **smallest genome** and encodes the **smallest number of predicted open reading frames** known for a free-living microorganism.
- *P. ubique* has **no pseudogenes, introns, transposons, extrachromosomal elements, or inteins**; few paralogs; and the shortest intergenic spacers yet observed for any cell.

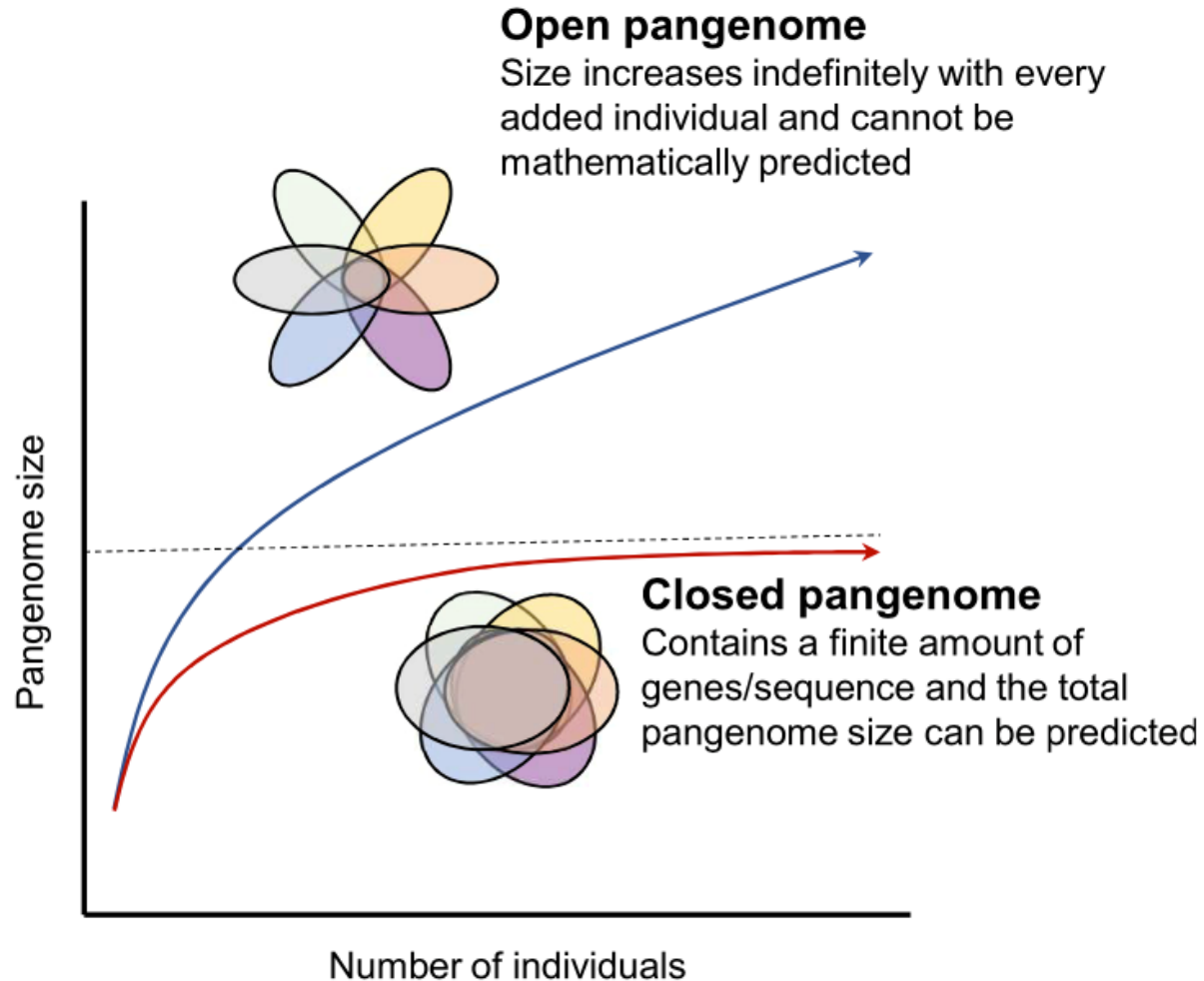


Genome Streamlining in a Cosmopolitan Oceanic Bacterium. *Science* 2005

<http://www.doi.org/10.1126/science.1114057>

Pangenome: Open or Closed

- **Open pangenome:** sequencing a new individual adds to the pangenome content.
- **Closed pangenome:** sequencing a sufficient number of individuals would capture almost the entire gene space and the theoretical size of the pangenome can be predicted.



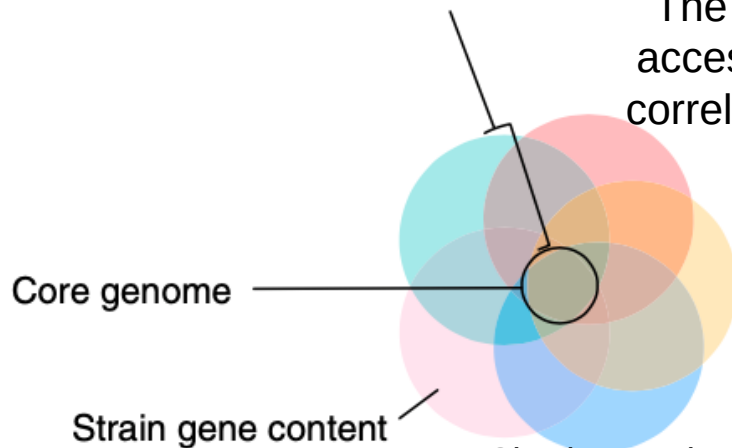
Pangenomics Comes of Age: From Bacteria to Plant and Animal Applications. *Trends in Genetics* 2020

Open vs Closed

Open pangenomes

Species with sympatric lifestyles tend to have open pangenomes.

Large accessory genome



Sharing environment with other organisms facilitates gene transfer, leading to larger, open pangenomes.

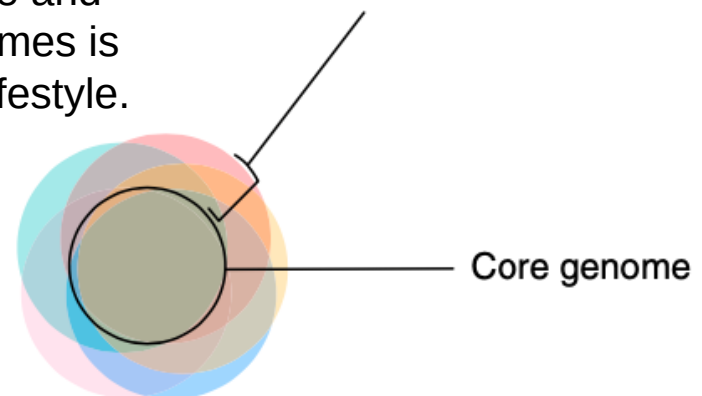
Common among....

Niche generalists
Diverse community interactions
Large population size

Closed pangenomes

Allopatric bacterial species tend to have closed pangenomes.

Small accessory genome



Niche specialists
Limited community interactions
Small population size

The Ecology and Evolution of Pangenomes. *Current Biology* 2019

First pangenome paper

RESEARCH ARTICLE

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

Hervé Tettelin, Vega Massignani, Michael J. Cieslewicz, Claudio Donati, Duccio Medini, Naomi L. Ward, Samuel V. Angiuoli, Jonathan Crabtree, Amanda L. Jones, A. S. Robert T. DeBoy, Tanja M. Davidsen, Marirosa Mora, Maria Scarselli, Immaculada Margarit y Ros, Jeremy D. Peterson, Christopher R. Hauser, Jaide William C. Nelson, Ramana Madupu, Lauren M. Brinkac, Robert J. Dodson, M. Steven A. Sullivan, Sean C. Daugherty, Daniel H. Haft, Jeremy Selengut, M. Liwei Zhou, Nikhat Zafar, Hoda Khouri, Diana Radune, George Dimitrov, Kisha ' Kevin J. B. O'Connor, Shannon Smith, Teresa R. Utterback, Owen White, Craig Guido Grandi, Lawrence C. Madoff, Dennis L. Kasper, John L. Telford, Michael Rino Rappuoli, and Claire M. Fraser

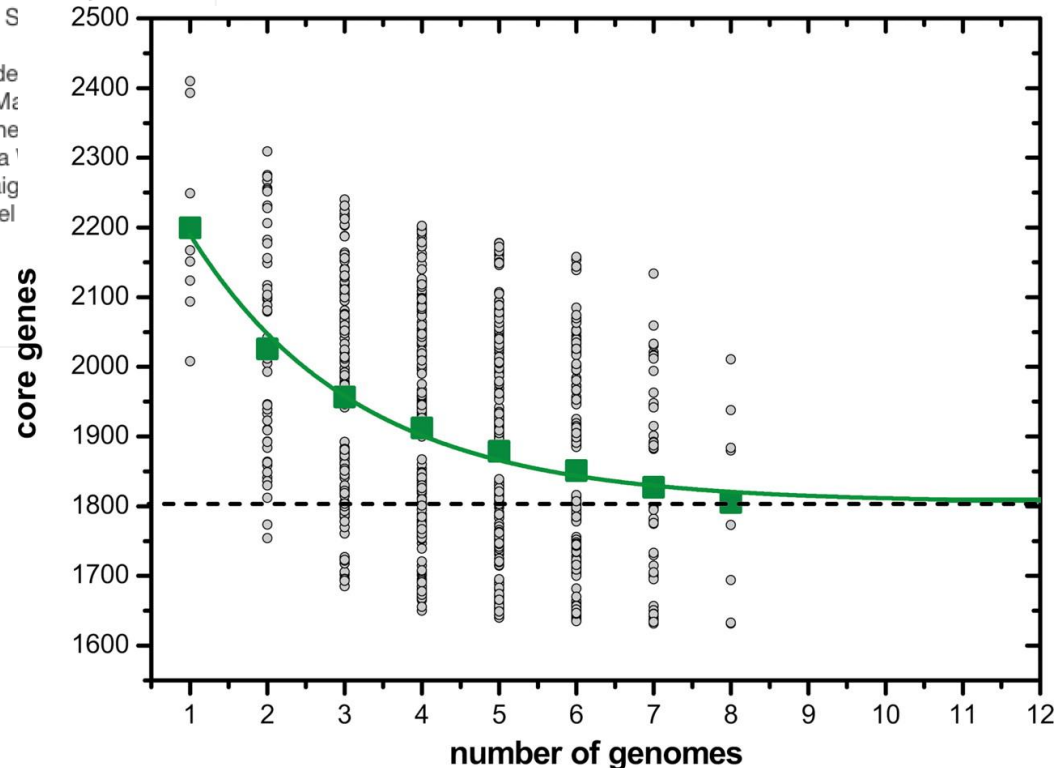
PNAS September 27, 2005 102 (39) 13950-13955; <https://doi.org/10.1073/pnas.0506758102>

Contributed by Rino Rappuoli, August 5, 2005

The term ‘pan-genome’ was defined with its current meaning by Tettelin *et al* in PNAS 2005.

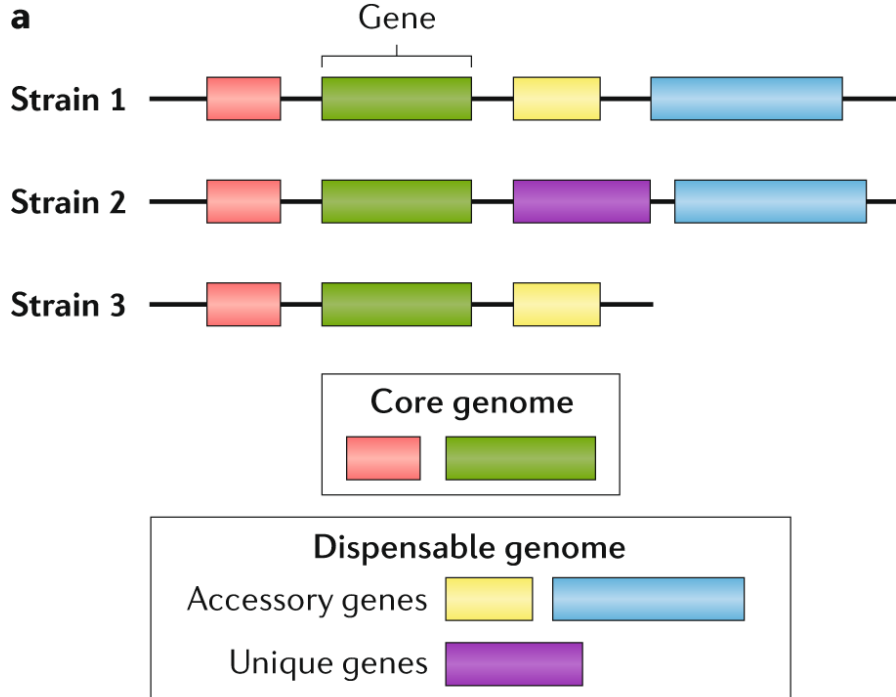


无乳链球菌是一种革兰氏阳性链球菌，可引起新生儿早发和晚发型感染，也被发现于各种畜牧及水产鱼类的流行病中，导致动物体产生败血症和脑膜炎，具有较高的致死率

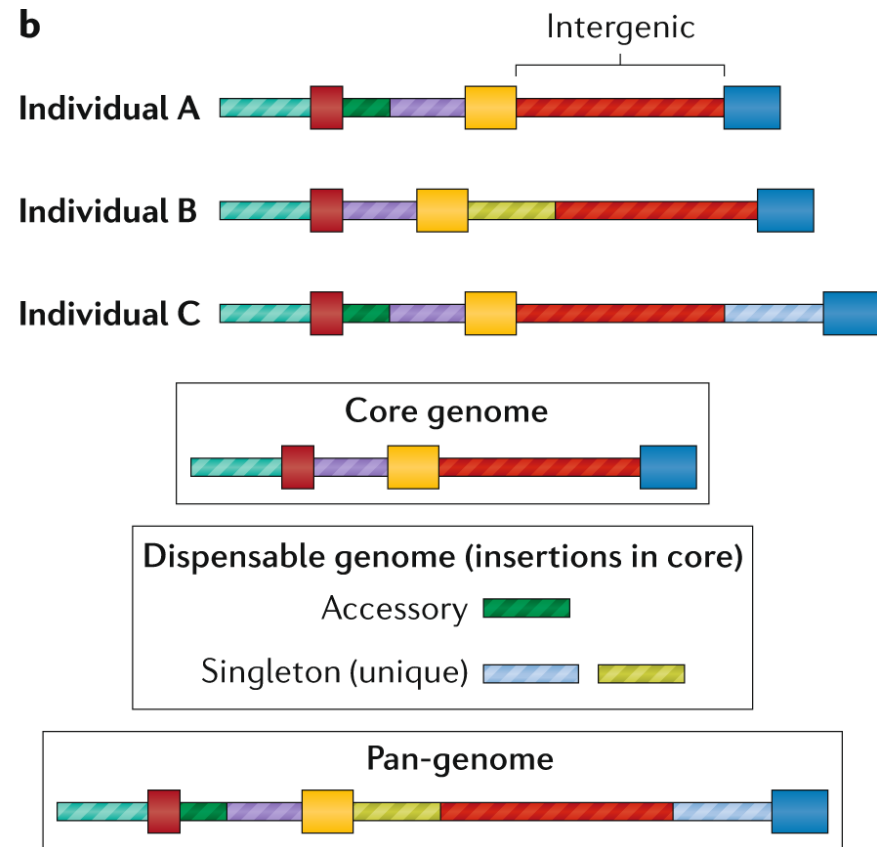


Core and dispensable genomes

Prokaryotes

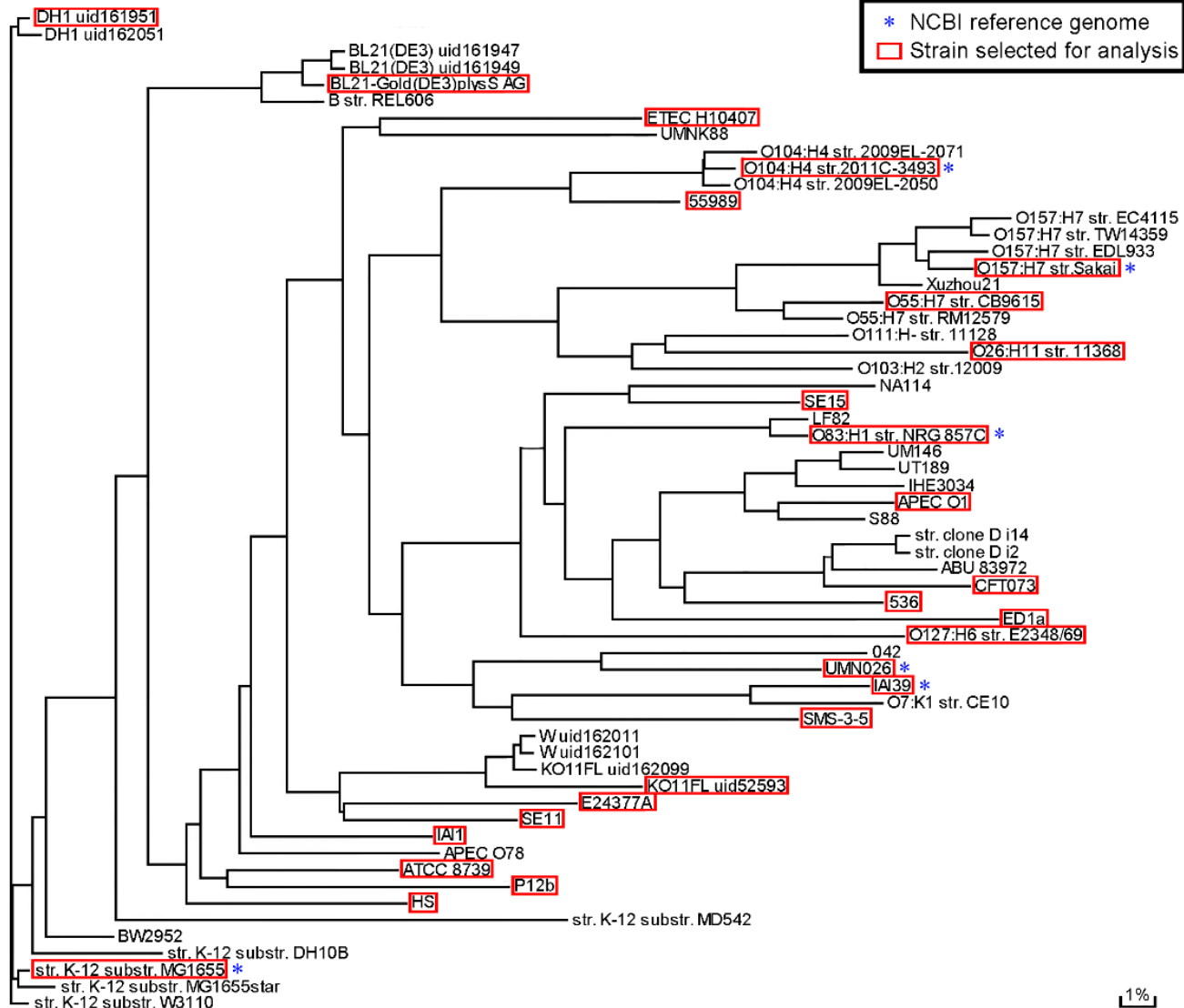


Eukaryotes



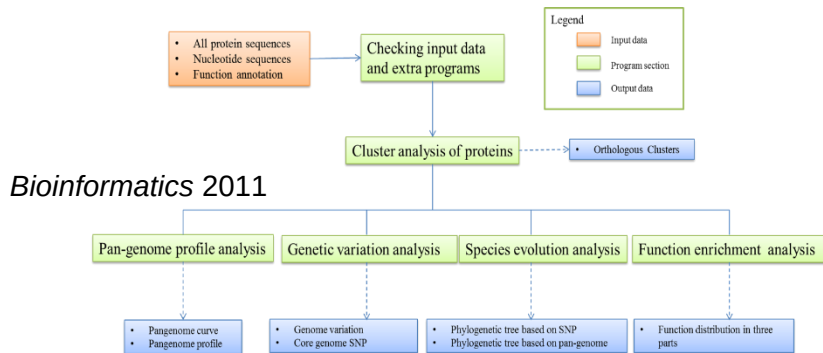
Pan-genomics in the human genome era. *Nature Reviews Genetics* 2020

E. coli strains



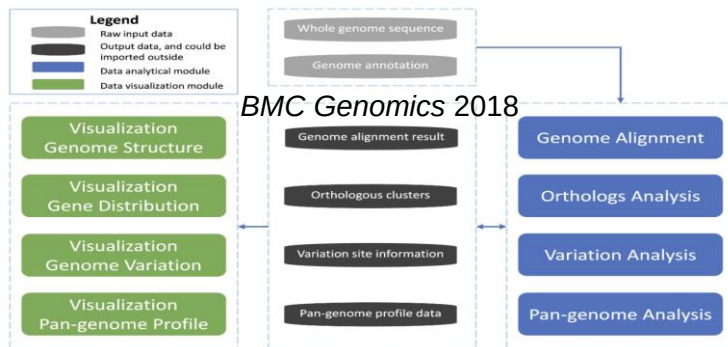
原核生物泛基因组分析工具

PGAP (Linux)



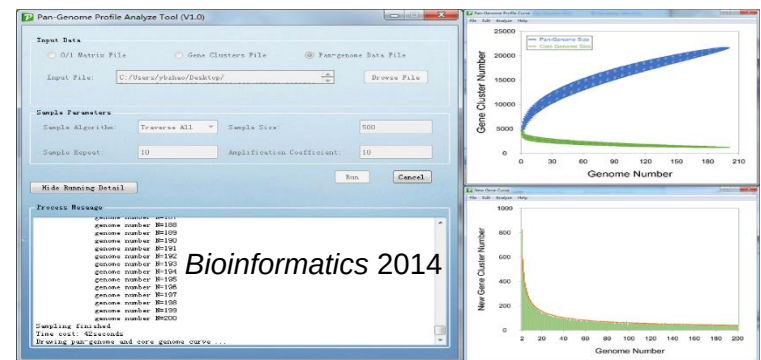
同源基因家族聚类、泛基因组特征计算、单核苷酸变异分析、菌株演化分析、功能富集分析

PGAP-X (Linux, windows)



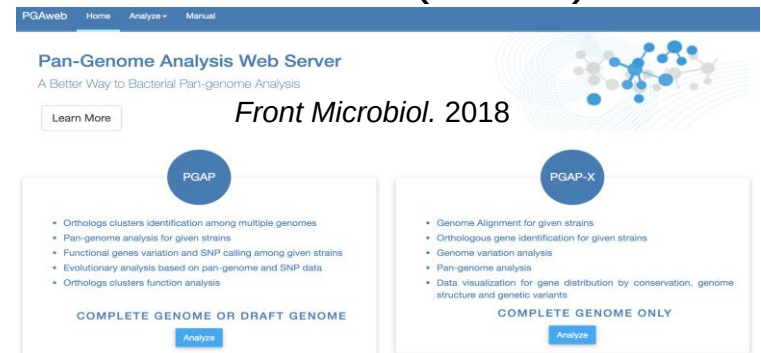
基因组比对、直系同源基因家族聚类、遗传变异分析、泛基因组特征计算、分析结果可视化

PanGP (Linux, Windows)



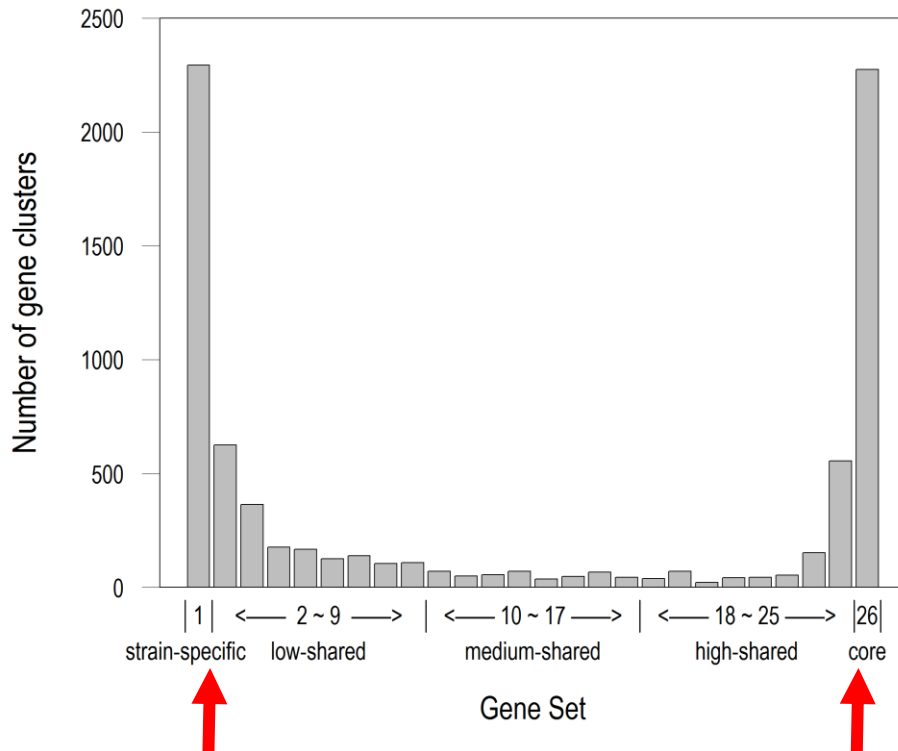
融合多种抽样算法用于泛基因组特征曲线和新基因曲线计算

PGAweb (online)



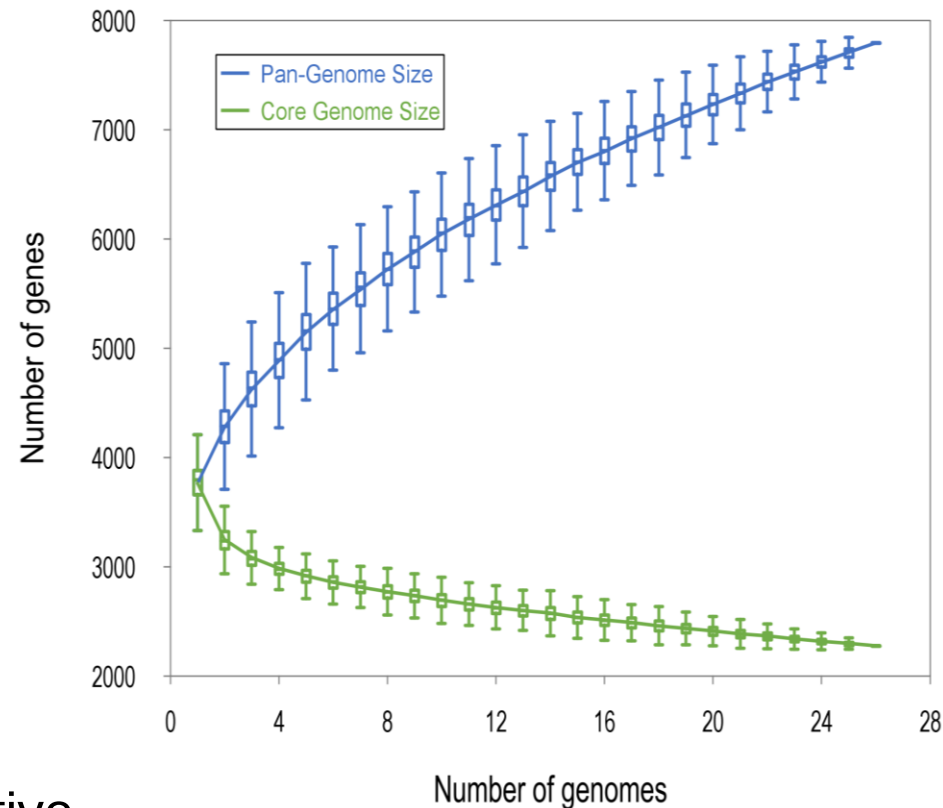
原核生物泛基因在线分析、点击式操作、交互可视化 (<http://pgaweb.vlcc.cn>)

E. coli pangenome



active in gene birth/gain
for adaptation to new
environments

conservative



E. coli is an open genome.

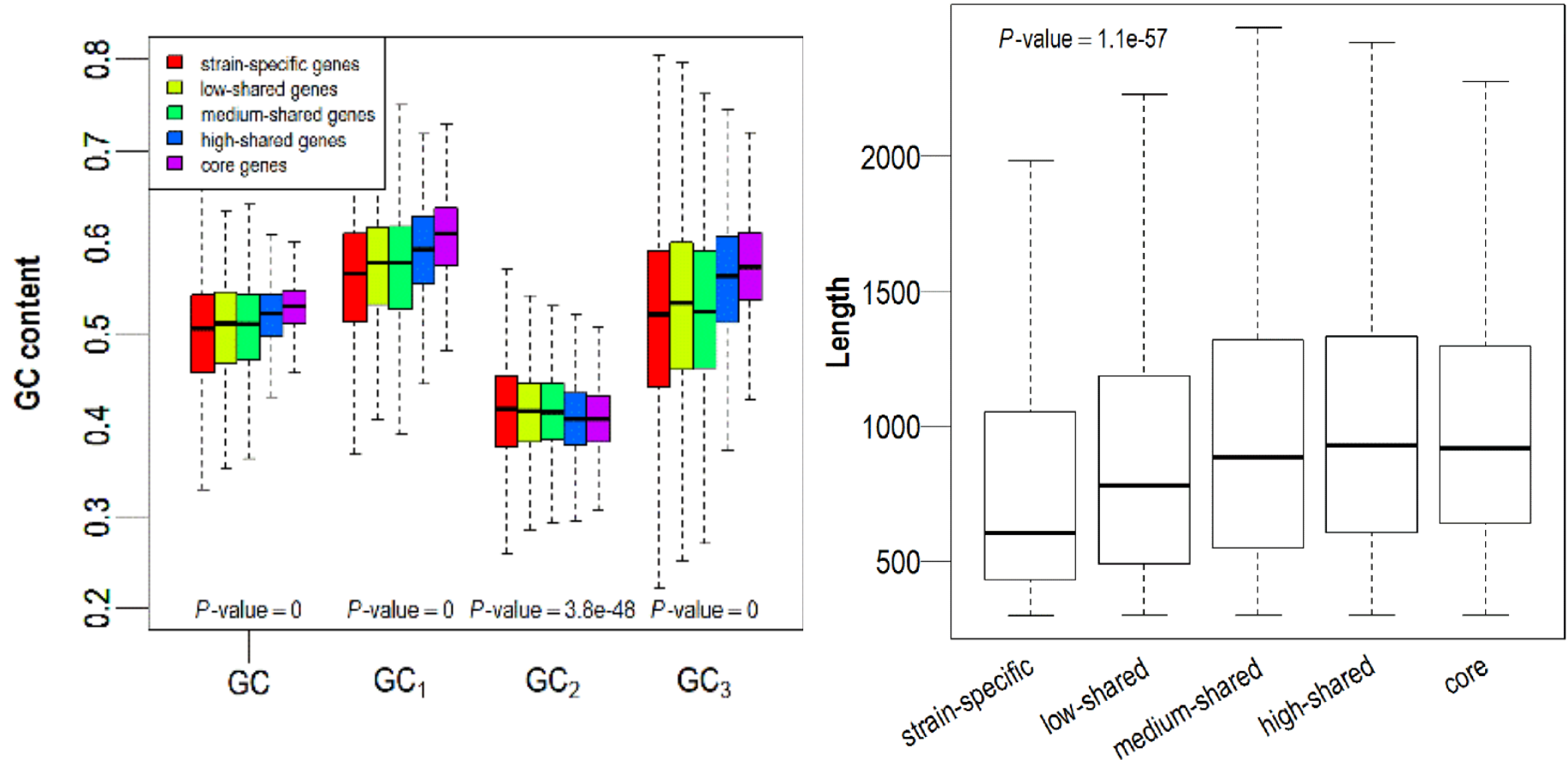
Core genes and strain-specific genes

Core genes	50S ribosomal protein L20
	30S ribosomal protein S9
	membrane protein, GPR1/FUN34/yaaH family / hypothetical protein
	DoxX family protein / quinol oxidase subunit / inner membrane protein YqjF / hypothetical protein
	50S ribosomal protein L15
	50S ribosomal protein L16
	30S ribosomal protein S12
	fumarate reductase subunit D / fumarate reductase (anaerobic), membrane anchor subunit
	30S ribosomal protein S14
	periplasmic ATP-independent protein refolding chaperone, stress-induced / spheroplast protein Y / hypothetical protein

Strain-specific genes	hypothetical protein
	hypothetical protein
	hypothetical protein
	hypothetical protein
	hypothetical protein
	hypothetical protein
	hypothetical protein
	hypothetical protein
	hypothetical protein
	putative transposase ORF2, IS66 family

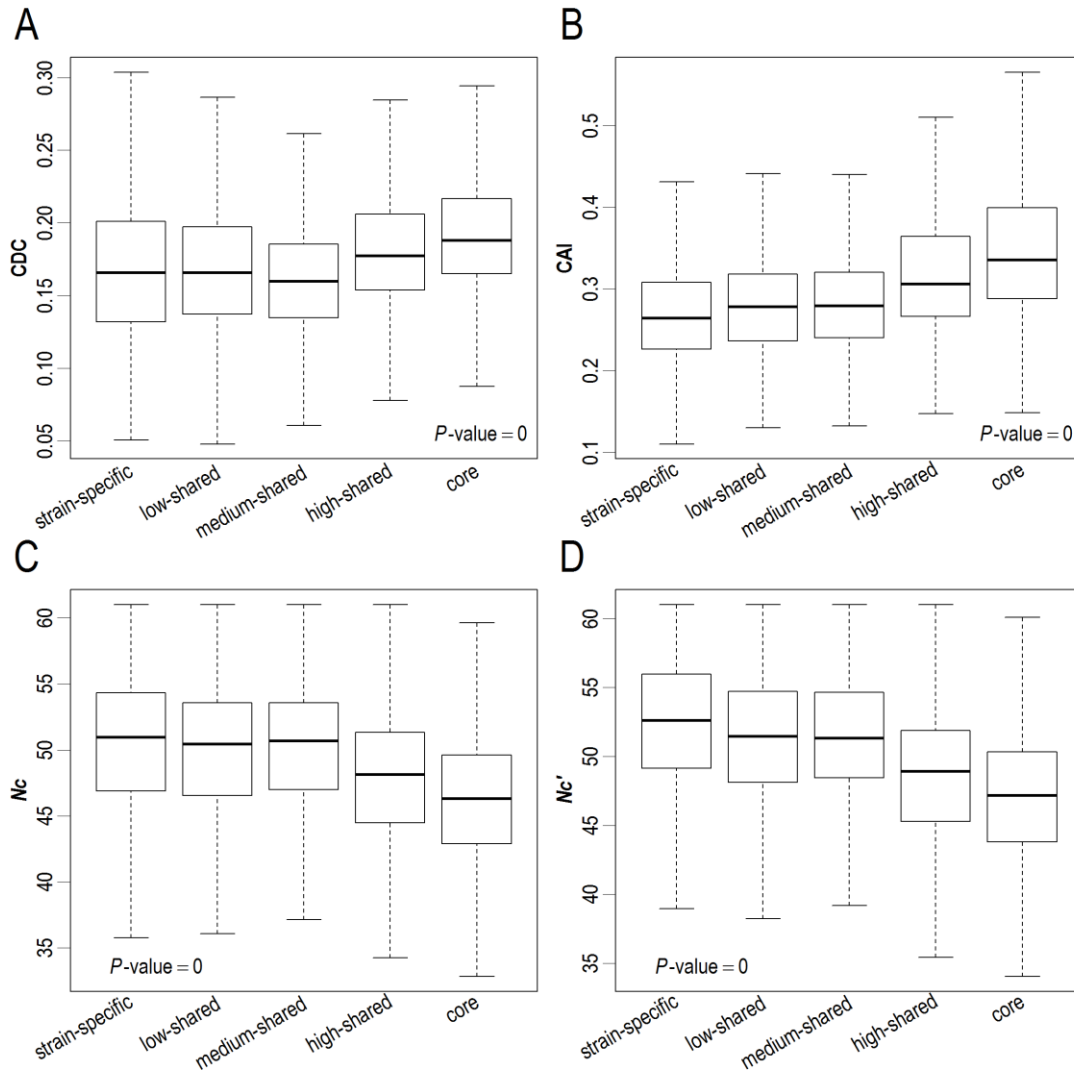
GC content and gene length in pangenome

A



With an increased presence in more isolates, genes tend to have higher GC content and longer sequence.

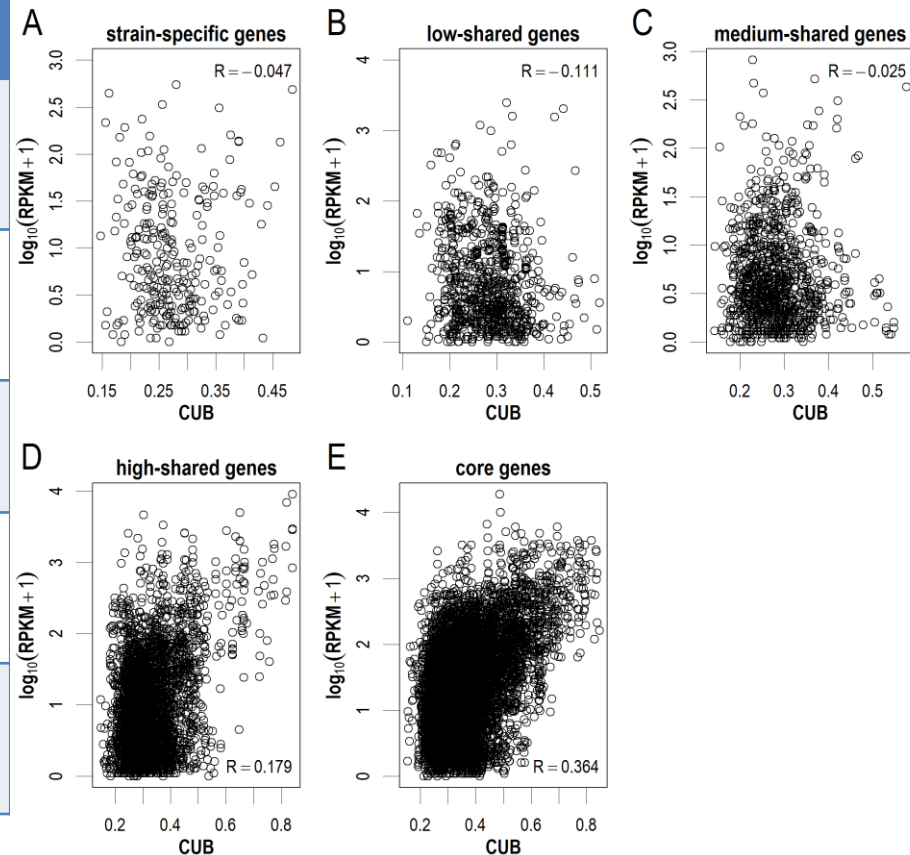
Codon usage bias in *E. coli* pangenome



Core genes possess more biased codon usage than strain-specific genes.

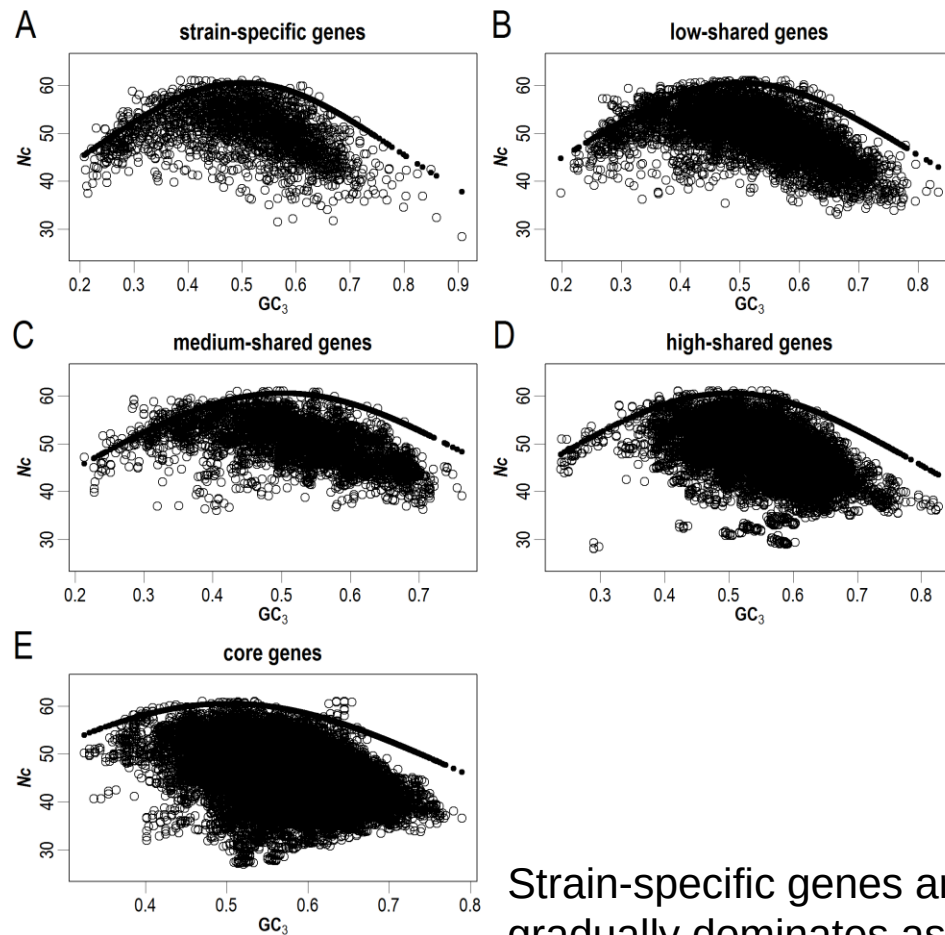
Translational selection in pangenome

Gene sets	CDC	CAI	Nc	Nc'
Strain-specific	-0.223*	-0.047	-0.015	-0.036
Low-shared	-0.128*	-0.111*	0.053	-0.025
Medium-shared	0.026	-0.025	0.071*	0.007
High-shared	0.193*	0.179*	-0.095*	-0.138*
Core	0.249*	0.364*	-0.193*	-0.258*



Codon usage bias correlates positively with gene presence and core genes tend to have higher CUBs and experience stronger translational selection.

ENC-plot: investigate the influence of mutation and selection acting on codon usage

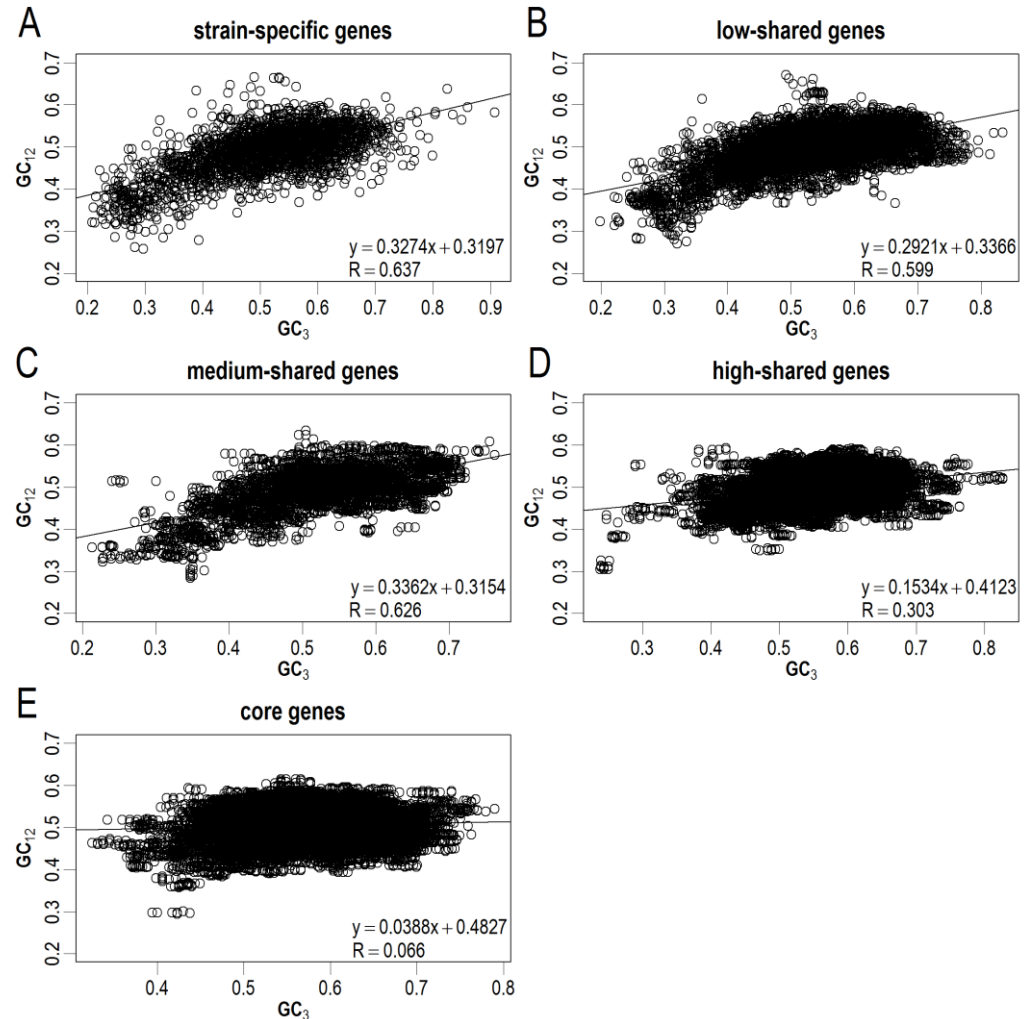


Gene set	Percentage (Deviation Value>5%)	Percentage (Deviation Value>10%)
Strain-specific genes	82%	38%
Low-shared genes	87%	46%
Medium-shared genes	88%	45%
High-shared genes	96%	69%
Core genes	99%	80%

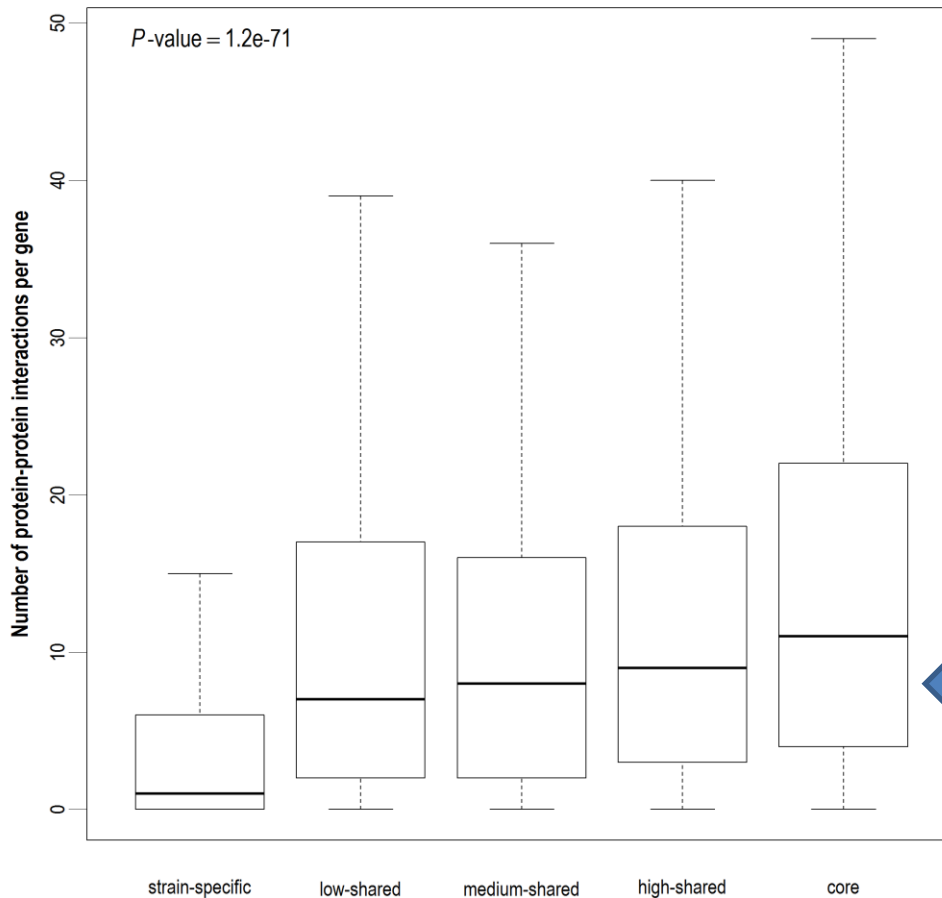
Strain-specific genes are primarily affected by mutation, and selection gradually dominates as an important factor to influence codon usage as genes are present in an increased number of isolates.

Neutrality plot: quantify the relative ratio between mutation and selection

- The correlation between GC_3 and GC_{12} the slope of GC_3 - GC_{12} regression function decreases from strain-specific genes to core genes.
- Mutation governs strain-specific genes and selection dominates core genes.



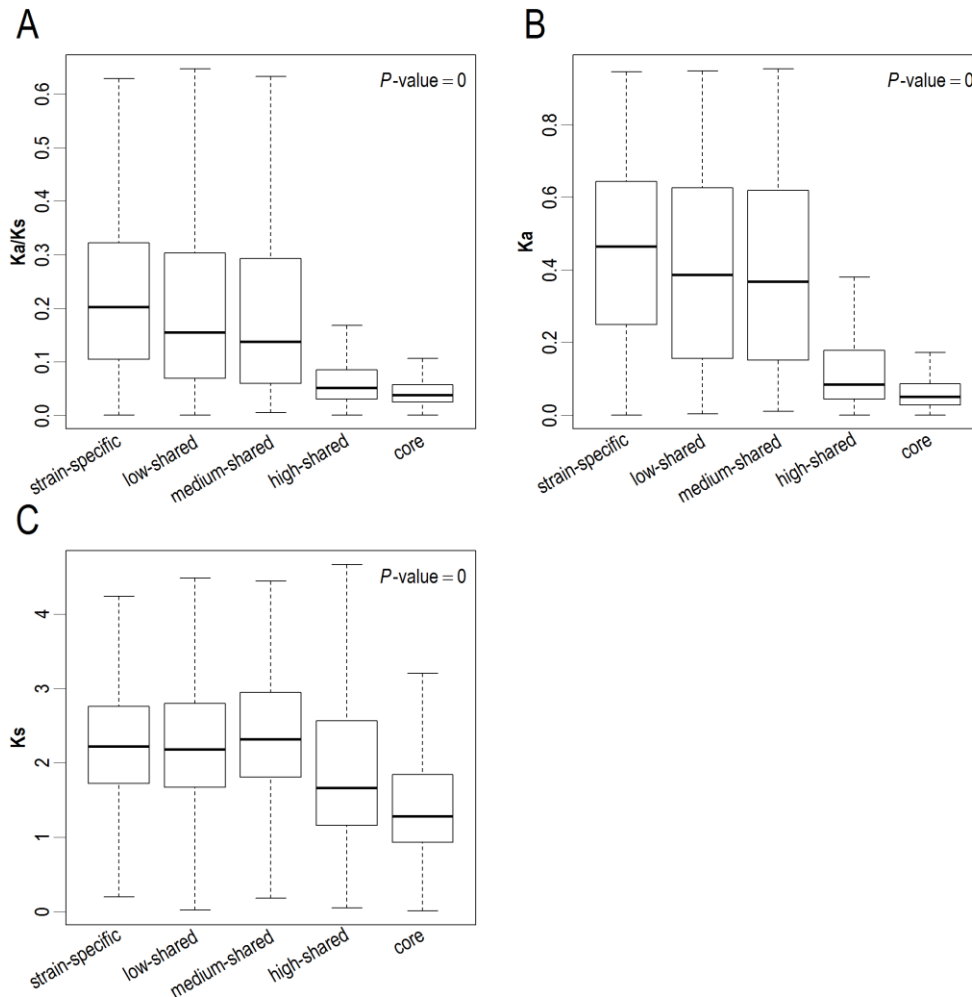
Protein-protein interaction (PPI)



Translational selection caters for high expression and highly expressed genes are more enriched in PPI network.

Core genes are hubs in protein interaction network.

Evolutionary rates: synonymous (Ks) & nonsynonymous (Ka) substitution rates



In contrast to strain-specific genes, core genes are more conservative and ancient, highly connected in protein-protein interaction network and thereby undergoing stronger selection constraints.

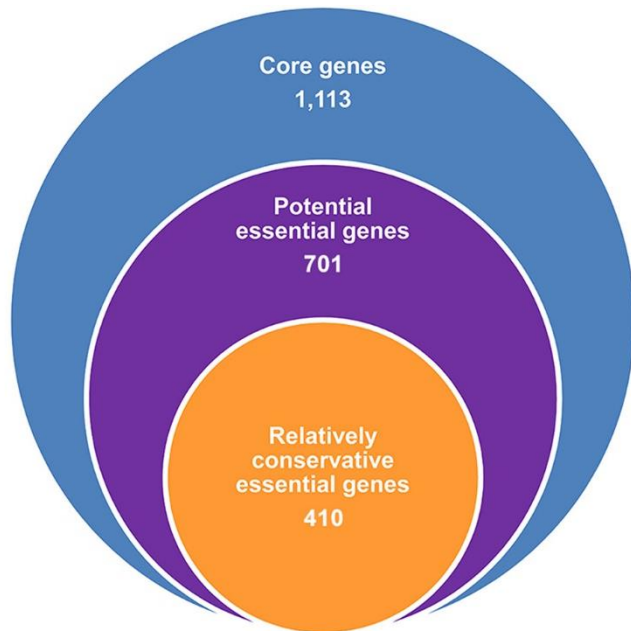
Pangenome

Pangenome is the global gene repertoire of a bacterial species.

- Core genes --- shared by all individuals
 - basic aspects of the biology
 - major phenotypic traits
- Dispensable genes --- shared by some individuals
 - species diversity supplementary biochemical pathways
- Strain-specific genes --- unique to an individual

Core genes vs. essential genes

- Essential genes (必需基因) are those genes of an organism that are thought to be **critical for its survival**.
- Essential genes encode proteins to **maintain a central metabolism, replicate DNA, translate genes into proteins, maintain a basic cellular structure, and mediate transport processes** into and out of the cell.

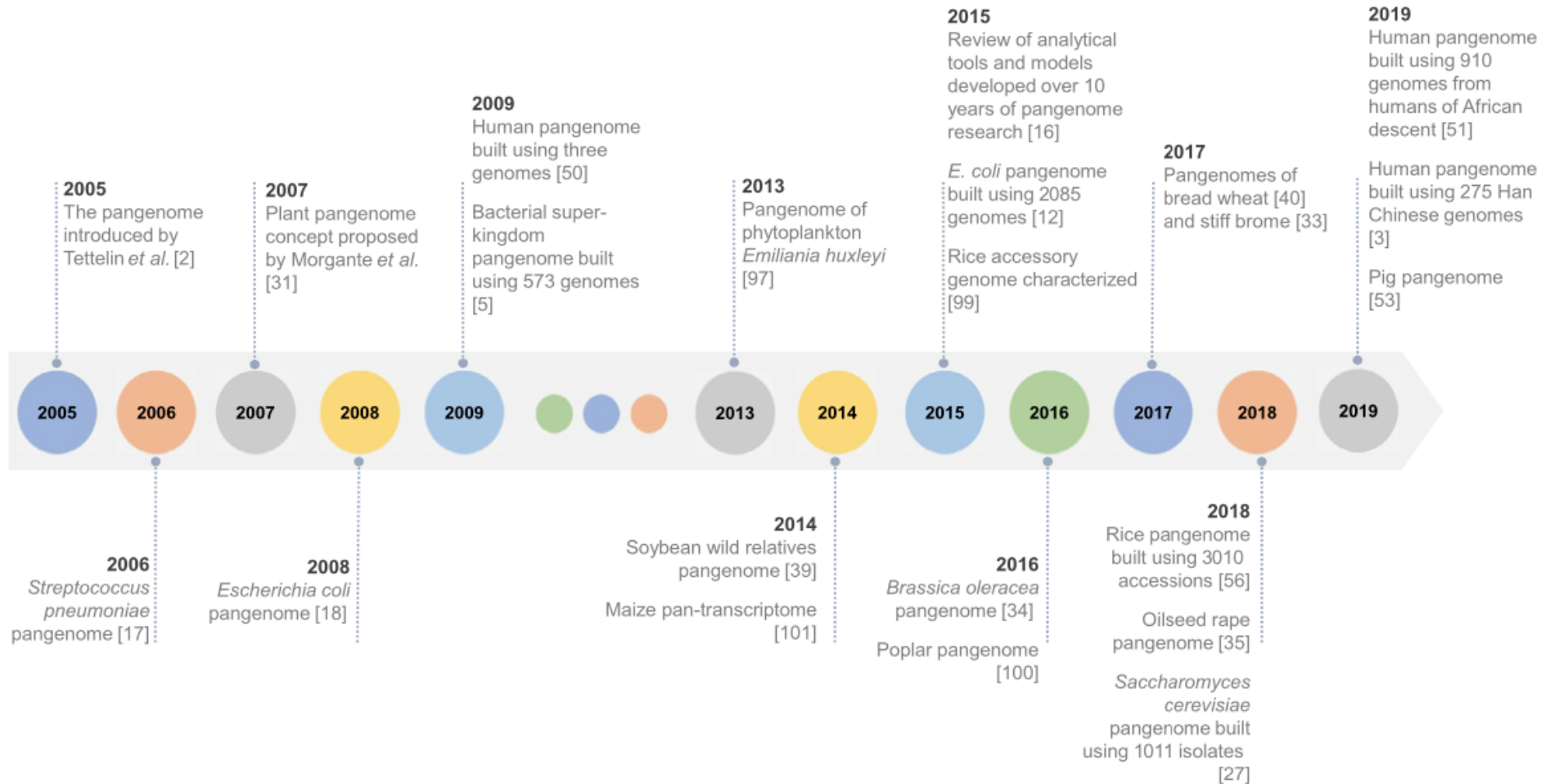


Organism	Essential Genes
<i>Escherichia coli</i>	1617
<i>Bacillus subtilis</i>	271
<i>Haemophilus influenzae</i>	642
<i>Streptococcus pneumoniae</i>	244
<i>Mycoplasma genitalium</i>	381
<i>Vibrio cholerae</i>	779
<i>Staphylococcus aureus</i>	653
<i>Saccharomyces cerevisiae</i>	1110

<https://doi.org/10.1111/1462-2920.14635>

Database of Essential Genes: <http://tubic.tju.edu.cn/deg/>

Timeline of Pangenomic Research



Pangenomics Comes of Age: From Bacteria to Plant and Animal Applications. *Trends in Genetics* 2020

Virus pangenome

Giant viruses open Pandora's box

Genome of largest viruses yet discovered hints at 'fourth domain' of life.

Ed Yong

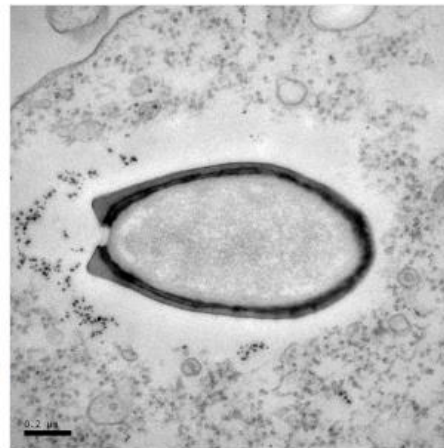
<https://doi.org/10.1038/nature.2013.13410>

18 July 2013

 Rights & Permissions

The organism was initially called NLF, for “new life form”. Jean-Michel Claverie and Chantal Abergel, evolutionary biologists at Aix-Marseille University in France, found it in a water sample collected off the coast of Chile, where it seemed to be infecting and killing amoebae. Under a microscope, it appeared as a large, dark spot, about the size of a small bacterial cell.

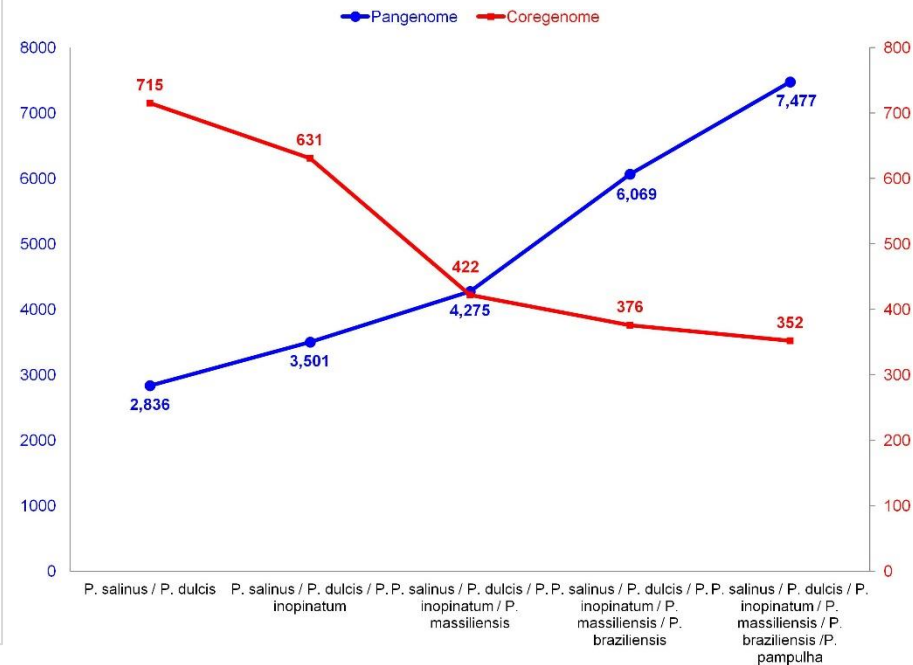
Later, after the researchers discovered a similar organism in a pond in Australia, they realized that both are viruses — the largest yet found. Each is around 1 micrometre long and 0.5 micrometres across, and their respective genomes top out at 1.9 million and 2.5 million bases — making the viruses larger than many bacteria and even some eukaryotic cells.



Chantal Abergel/Jean-Michel Claverie

Pandoraviruses infect amoebae and are larger than some bacteria.

The core genome of **8 pandoraviruses** comprises **352 genes**, and the ratio core genome/pangenome is less than **0.05**.



***Pandoravirus salinus*, 2473870 bp, 1647 genes**

<https://www.ncbi.nlm.nih.gov/genome/22431>

A Large Open Pangenome and a Small Core Genome for Giant Pandoraviruses. *Front Microbiol* 2018

Soybean pangenome

Cell

Volume 182, Issue 1, 9 July 2020, Pages 162-176.e13

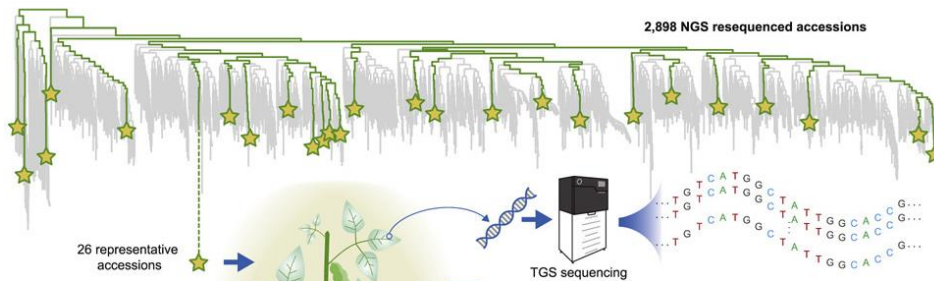


Resource

Pan-Genome of Wild and Cultivated Soybeans

Yucheng Liu^{1,7,8}, Huilong Du^{2,7,8}, Pengcheng Li³, Yanting Shen⁴, Hua Peng^{2,7}, Shulin Liu¹, Guo-An Zhou¹, Haikuan Zhang³, Zhi Liu^{1,7}, Miao Shi³, Xuehui Huang⁵, Yan Li⁶, Min Zhang¹, Zheng Wang¹, Baoge Zhu¹, Bin Han⁶, Chengzhi Liang^{2,7}✉, Zhixi Tian^{1,7,9}✉

de novo genome assemblies for **26 representative soybeans** that were selected from **2,898** deeply sequenced accessions.



Summary of the Assembly and Annotation of 26 Soybean Genomes

Sample	Contig N50 (Mb)	Scaffold N50 (Mb)	Assembly Size (Mb)	Chr Size (Mb)	Chr Loading Ratio (%)	Repetitive Content (%)	Gene Number (#)
SoyW01	23.9	52.1	1,021.3	1,010.7	99.0	53.9	57,920
SoyW02	20.5	51.9	1,007.0	983.9	97.7	54.2	55,516
SoyW03	26.8	52.3	1,014.5	1,008.1	99.4	55.3	57,884
SoyL01	23.3	50.6	999.2	991.8	99.3	54.8	54,502
SoyL02	23.1	52.3	1,011.7	1,003.6	99.2	54.7	54,803
SoyL03	21.3	51.3	1,039.5	1,002.7	96.5	54.2	55,769
SoyL04	23.0	51.4	1,005.2	995.5	99.0	54.5	58,881
SoyL05	22.1	51.2	1,059.8	993.8	93.8	54.5	56,086
SoyL06	21.5	50.7	997.8	997.2	99.9	54.5	56,573
SoyL07	23.7	50.9	1,004.7	991.0	98.6	54.8	58,259
SoyL08	22.6	51.1	999.7	993.6	99.4	54.8	58,496
SoyL09	23.1	51.0	1,028.2	997.1	97.0	54.6	59,588
SoyC01	23.5	51.5	1,003.8	999.7	99.6	54.7	54,405
SoyC02	23.8	51.2	1,011.1	997.1	98.6	54.4	55,191
SoyC03	23.9	51.3	1,007.8	994.6	98.7	54.0	57,792
SoyC04	22.7	50.9	1,001.5	998.7	99.7	54.4	57,777
SoyC05	22.7	50.3	992.3	989.3	99.7	54.3	58,040
SoyC06	23.0	52.0	1,007.9	997.4	99.0	54.1	55,619
SoyC07	21.8	51.2	1,008.8	1,000.9	99.2	54.7	54,792
SoyC08	22.4	50.9	1,002.3	990.0	98.8	54.0	54,747
SoyC09	22.6	50.4	1,004.1	995.3	99.1	53.6	55,926
SoyC10	19.8	51.0	1,004.9	995.4	99.1	54.1	56,011
SoyC11	18.8	51.4	1,024.5	1,000.1	97.6	53.8	56,750
SoyC12	20.0	50.8	1,025.1	995.6	97.1	53.8	55,000
SoyC13	23.8	51.1	1,010.1	991.4	98.2	54.3	57,976
SoyC14	23.0	51.6	1,007.4	988.5	98.1	54.1	55,267

Cell 2020 <https://doi.org/10.1016/j.cell.2020.05.023>

High-quality genome sequencing

Cell

Volume 182, Issue 1, 9 July 2020, Pages 162-176.e13



Resource

Pan-Genome of Wild and Cultivated Soybeans

Yucheng Liu^{1,7,8}, Huilong Du^{2,7,8}, Pengcheng Li³, Yanting Shen⁴, Hua Peng^{2,7}, Shulin Liu¹, Guo-An Zhou¹, Haikuan Zhang³, Zhi Liu^{1,7}, Miao Shi³, Xuehui Huang⁵, Yan Li⁶, Min Zhang¹, Zheng Wang¹, Baoge Zhu¹, Bin Han⁶, Chengzhi Liang^{2,7} ㉟, Zhixi Tian^{1,7,9} ㉟

...using single-molecule real-time (SMRT) sequencing with an average coverage depth of 96 $\times$, optical mapping with an average coverage depth of 277 $\times$, chromosome conformation capture (Hi-C) sequencing with an average coverage depth of 136 $\times$, and Illumina sequencing (HiSeq) with an average coverage depth of 68 $\times$. **The contig N50 sizes of the 26 whole-genome assemblies ranged from 18.8 to 26.8 Mb pairs with a mean of 22.6 Mb.**

Article | [Open Access](#) | Published: 14 March 2019

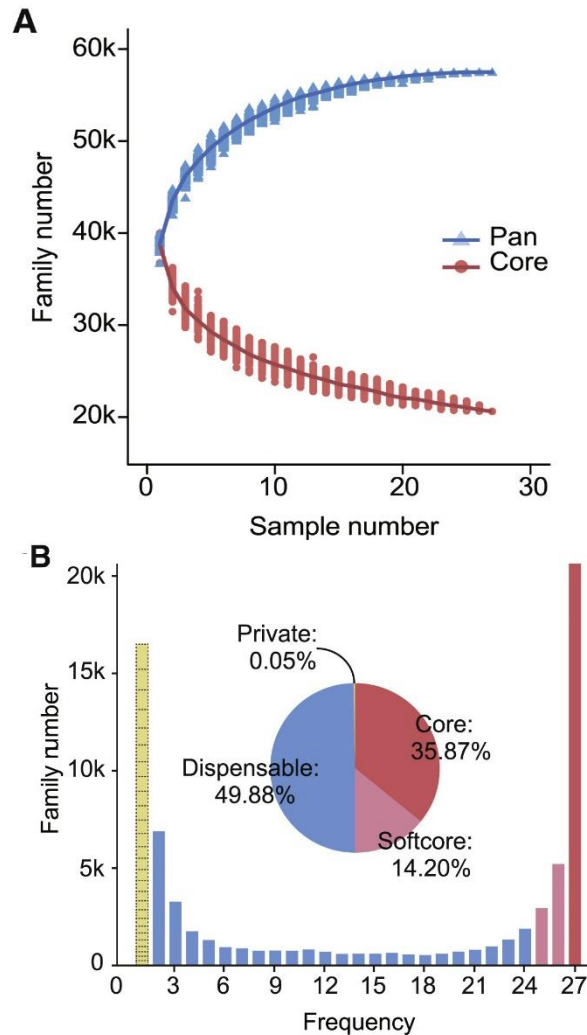
A reference-grade wild soybean genome

Min Xie, Claire Yik-Lok Chung, Man-Wah Li, Fuk-Ling Wong, Xin Wang, Ailin Liu, Zhili Wang, Alden King-Yung Leung, Tin-Hang Wong, Suk-Wah Tong, Zhixia Xiao, Kejing Fan, Ming-Sin Ng, Xinpeng Qi, Linfeng Yang, Tianquan Deng, Lijuan He, Lu Chen, Aisi Fu, Qiong Ding, Junxian He, Gyuhwa Chung, Sachiko Isobe, Takanari Tanabata, Babu Valliyodan, Henry T. Nguyen, Steven B. Cannon, Christine H. Foyer, Ting-Fung Chan ㉟ & Hon-Ming Lam ㉟ -Show fewer authors

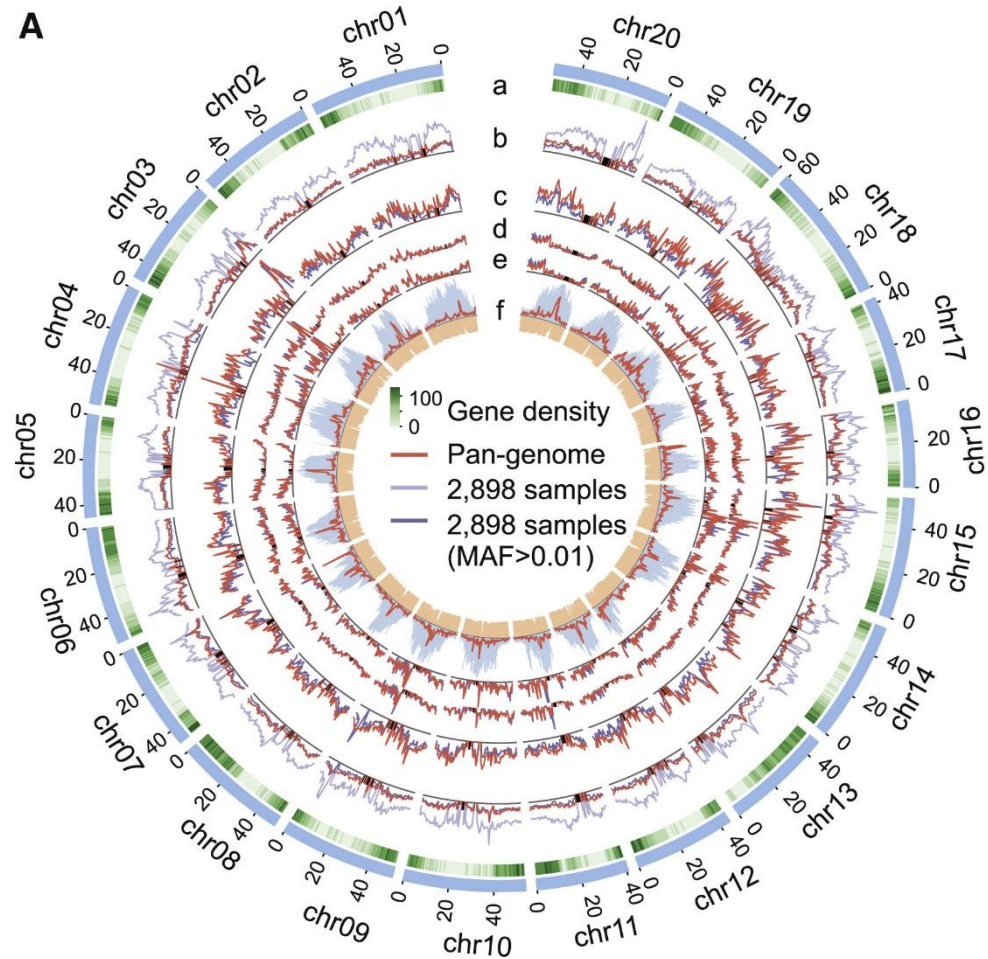
Nature Communications **10**, Article number: 1216 (2019) | [Cite this article](#)

After gap filling and polishing, the final assembly for W05 is 1013.2 Mb in length, with 988.6 Mb unambiguous bases (Supplementary Figure [1g](#) and Table [1](#)). In total, 95.7% of sequences are anchored to 20 superscaffolds, corresponding to 20 chromosomes, whereas 43.6 Mb in 1098 contigs remain unplaced. **The contig N50 of the final assembly is 3.3 Mb.**

Soybean Pangenome



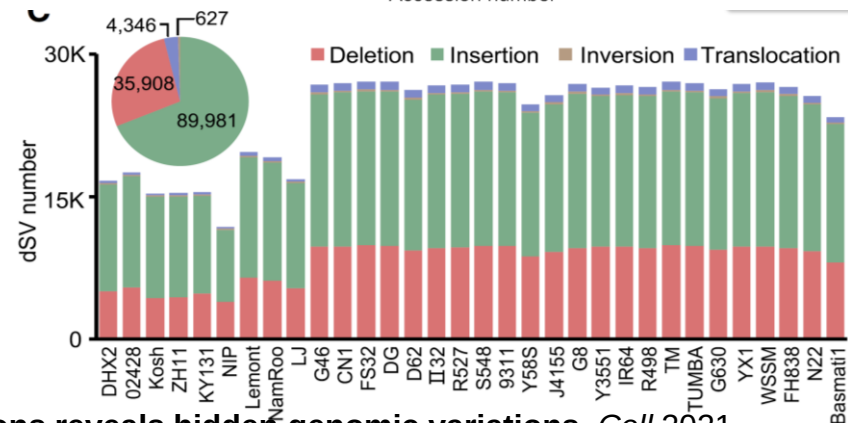
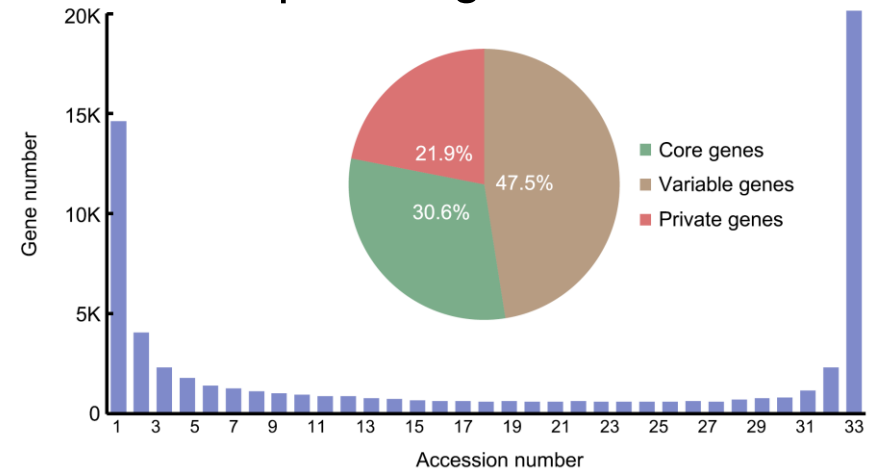
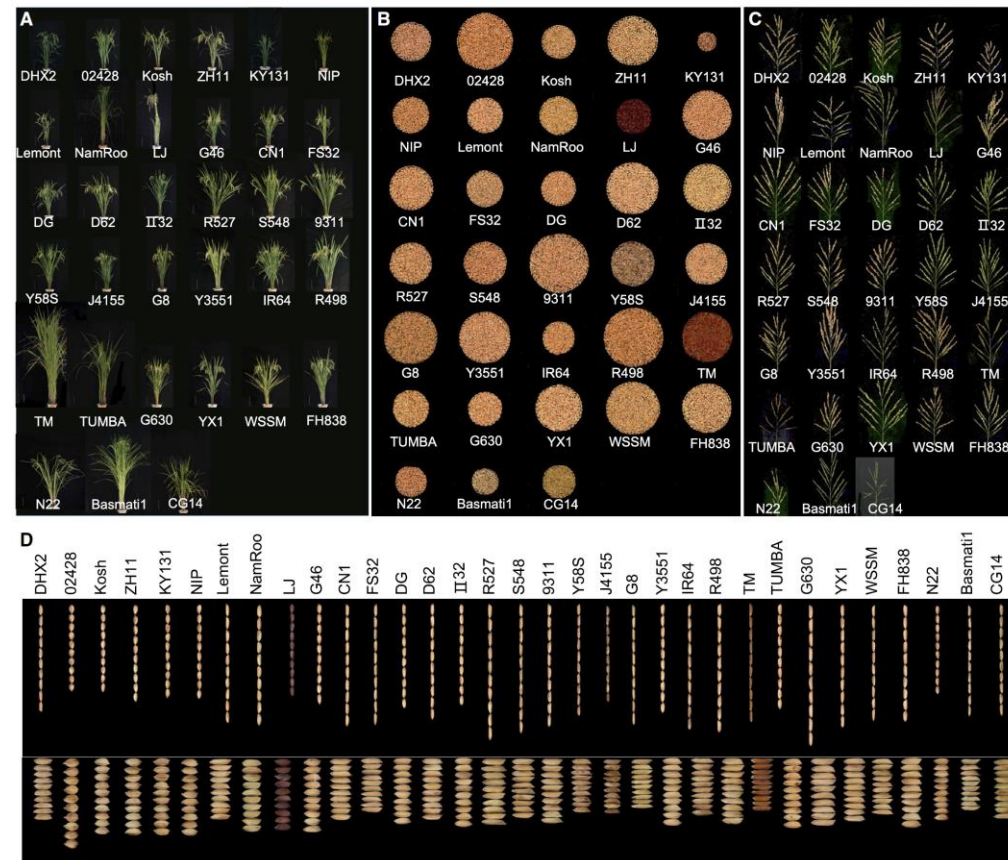
Cell 2020 <https://doi.org/10.1016/j.cell.2020.05.023>



(a) Gene density. (b–e) SNP density, π , d_N , and d_S . (f) Distribution of larger structure variations and repeat sequences.

Rice Pangenome

a pan-genome of rice comprising **66,636** protein-coding genes: **20,374** genes categorized as “core genes”; **46,262** genes categorized as “variable genes” which included **14,609** accession-private genes



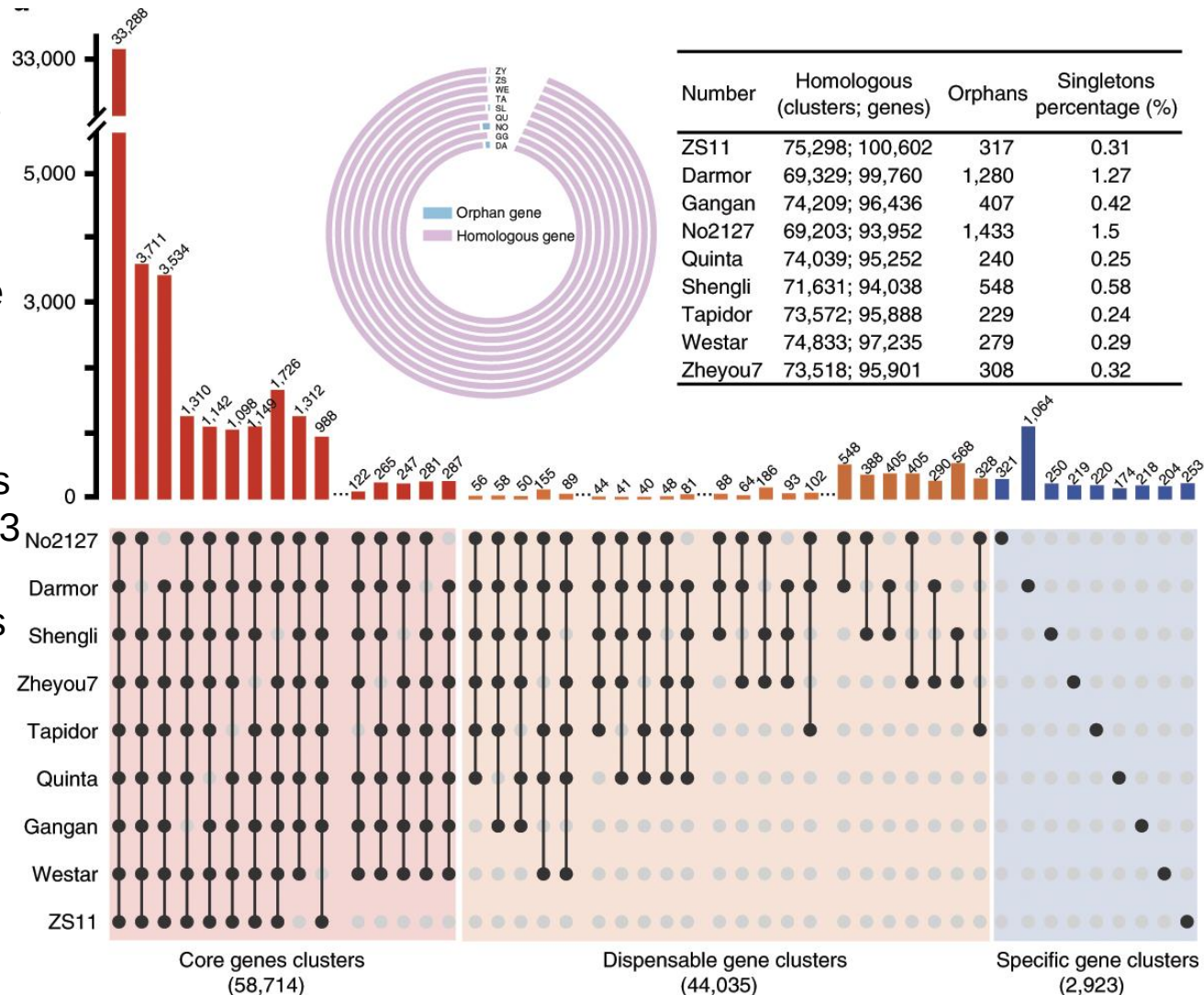
Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* 2021

Pangenome of *Brassica napus*(油菜)

All 874,105 annotated genes in the **9 *B. napus* genomes** were assigned to 105,672 gene clusters.

- 58,714 (~56%) core-gene clusters that existed in at least seven genomes
- 44,035 (~42%) dispensable gene clusters
- The remaining ~2% (2,923 gene clusters and 5,041 singletons) specific genes with no orthologues in other genomes

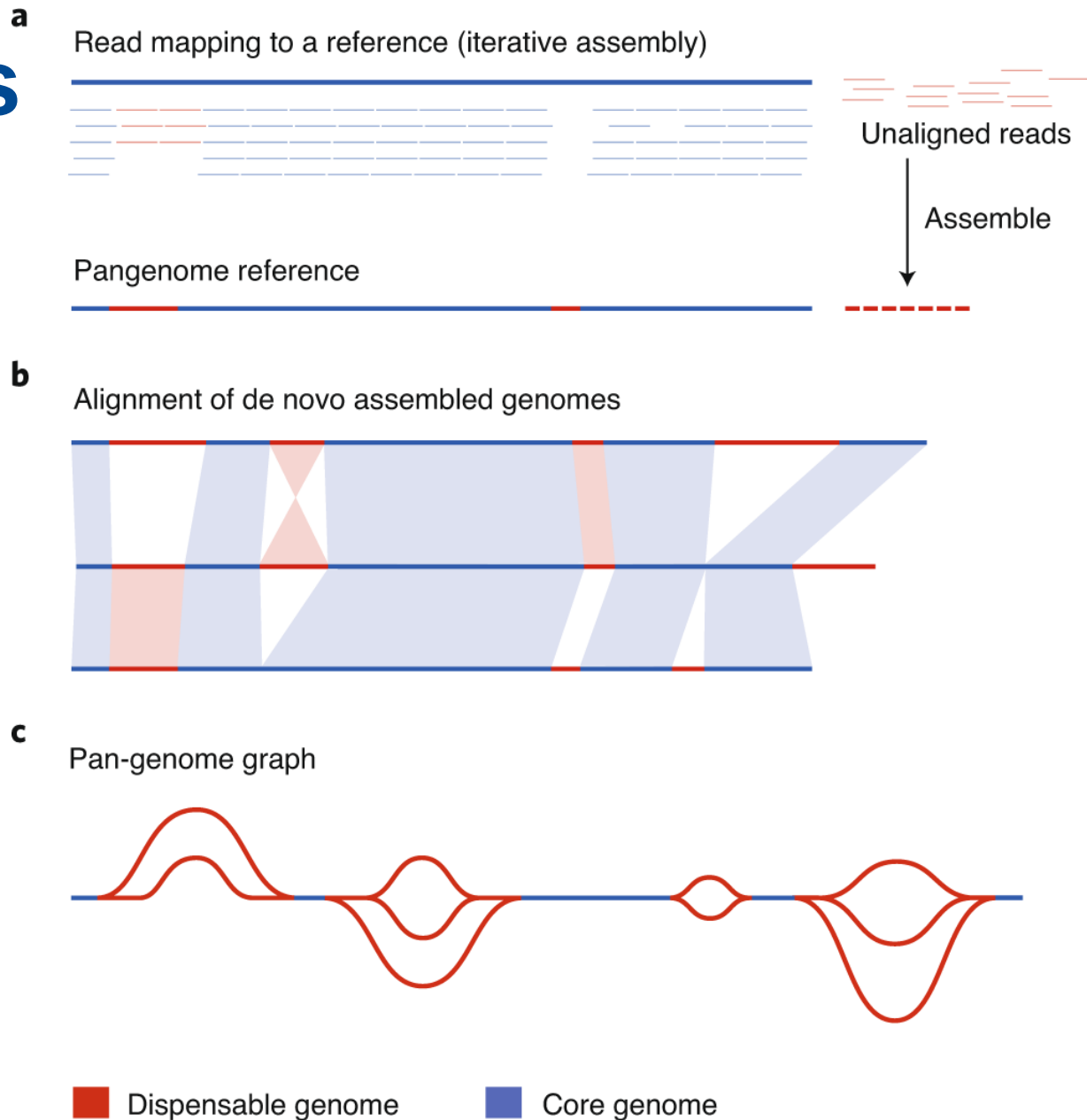
Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. *Nature Plants* 2020



Pangenomes are the new reference

A **pangenome graph** can be constructed from whole genome alignments or by de novo graph assembly, and efficiently stores variant information of dispensable regions as unique paths through the graph.

Plant pan-genomes are the new reference. *Nature Plants* 2020

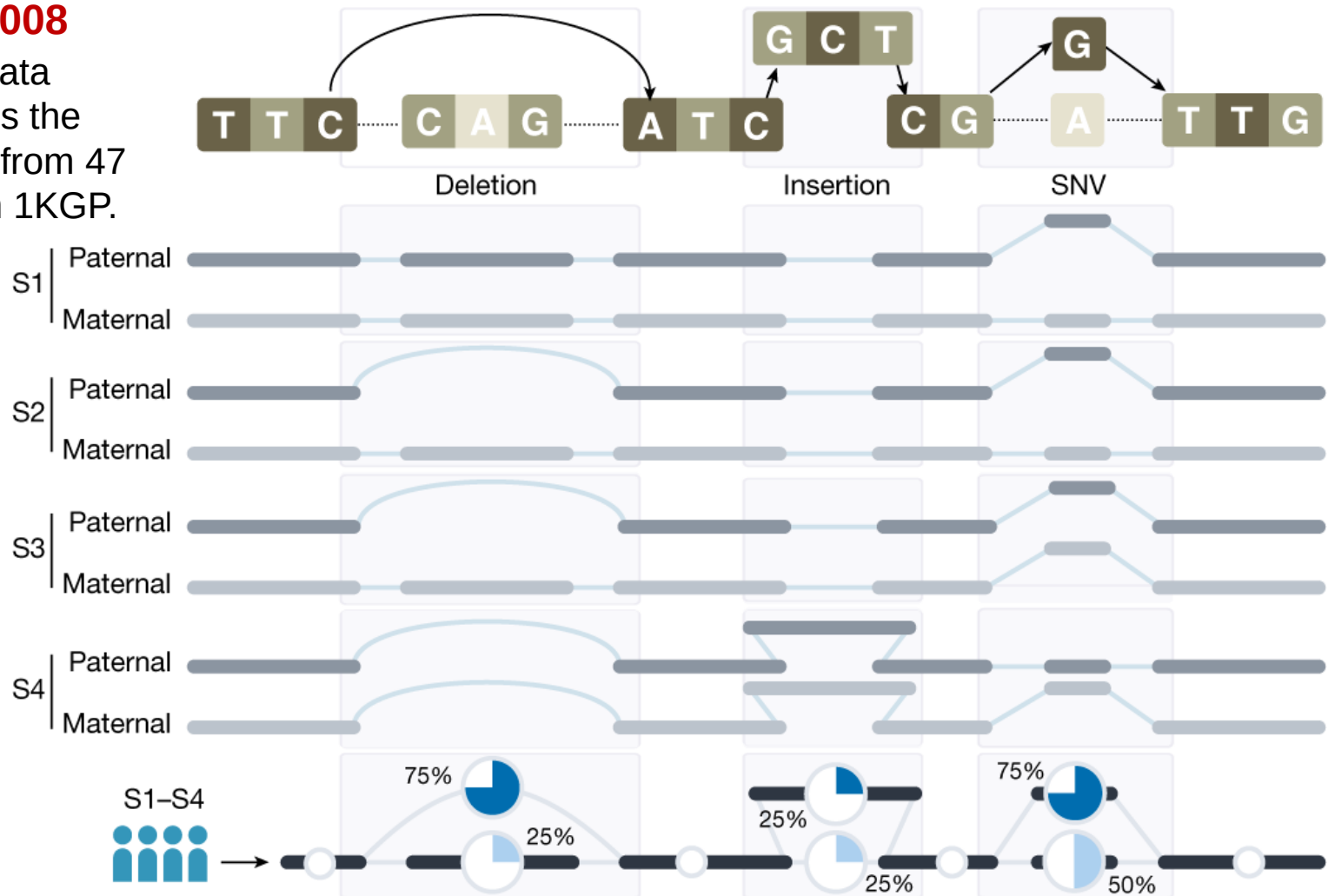


Human Pangenome Reference

Launched in 2008

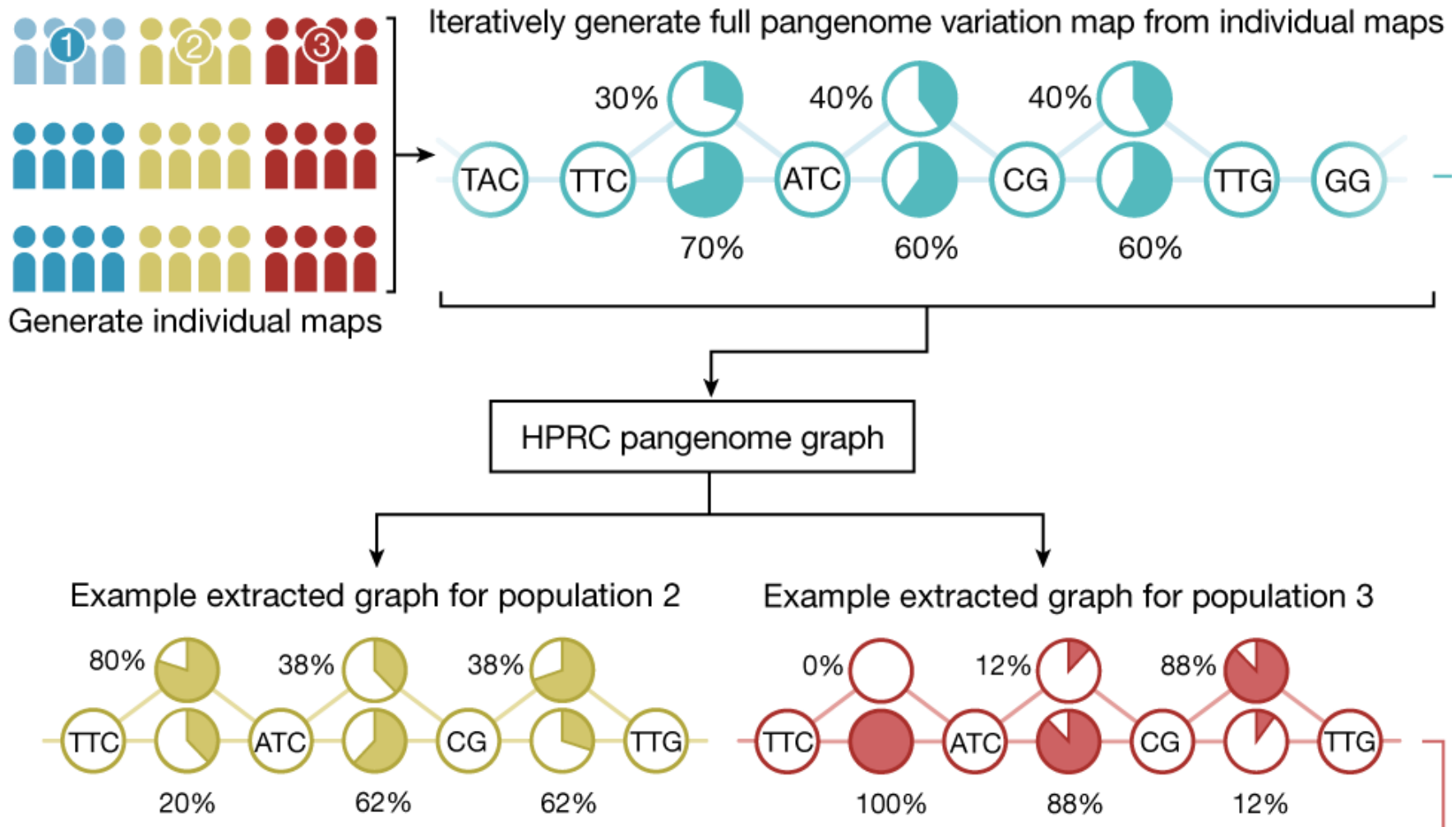
The first HPRC data release comprises the sequencing data from 47 participants, from 1KGP.

350 people from 26 populations whose genomes will be analysed in the Human Pangenome Project.



The Human Pangenome Project: a global resource to map genomic diversity. *Nature* 2022

Human Pangenome Reference

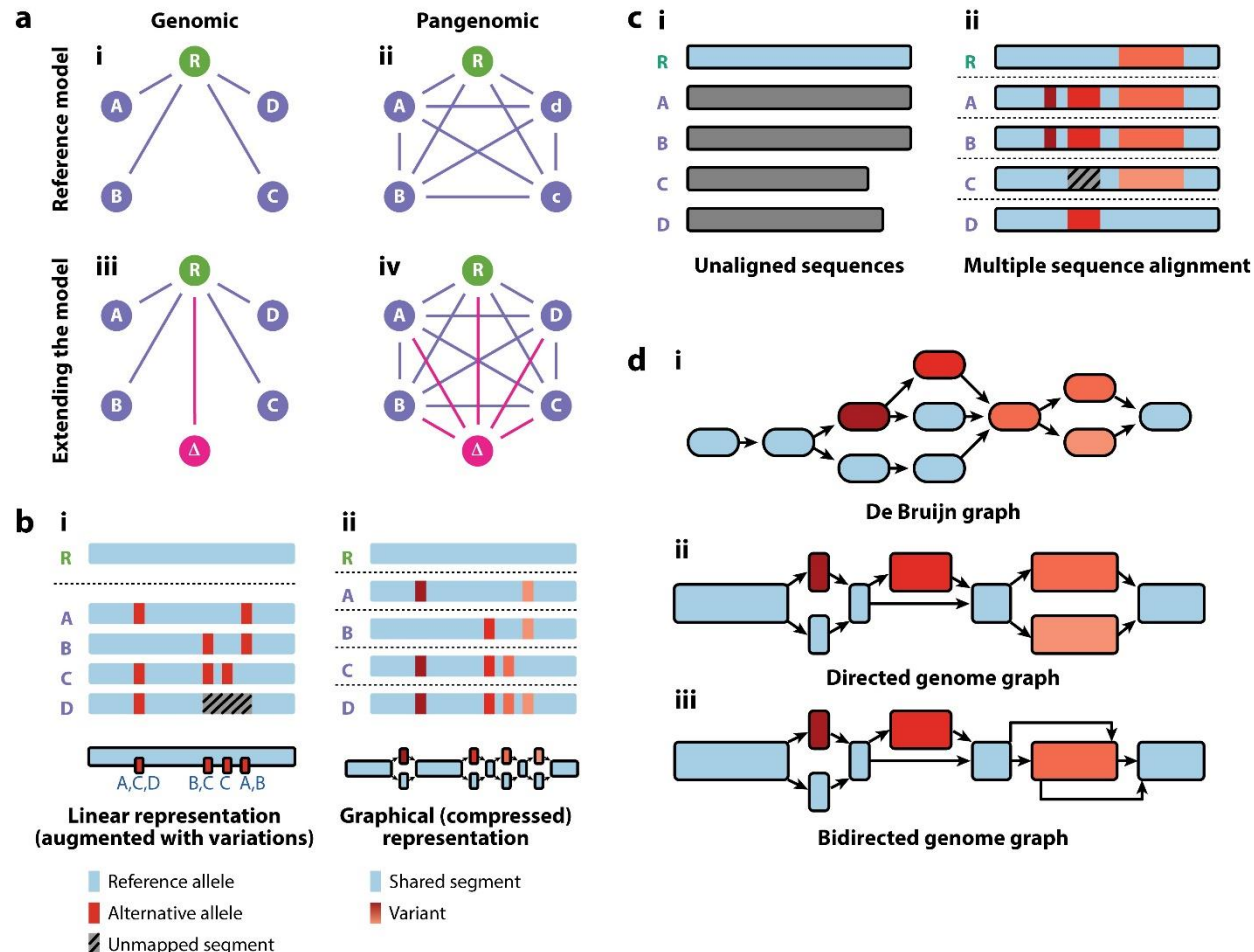


The Human Pangenome Project: a global resource to map genomic diversity. *Nature* 2022

Pangenome graph

Pangenome graphs are emerging data structures designed to represent pangenomes.

Genome graphs are sequence graphs used to represent whole genome relationships, in which *sequence graph* serves to compress many redundant input sequences into a smaller data structure that is still representative of the full set.

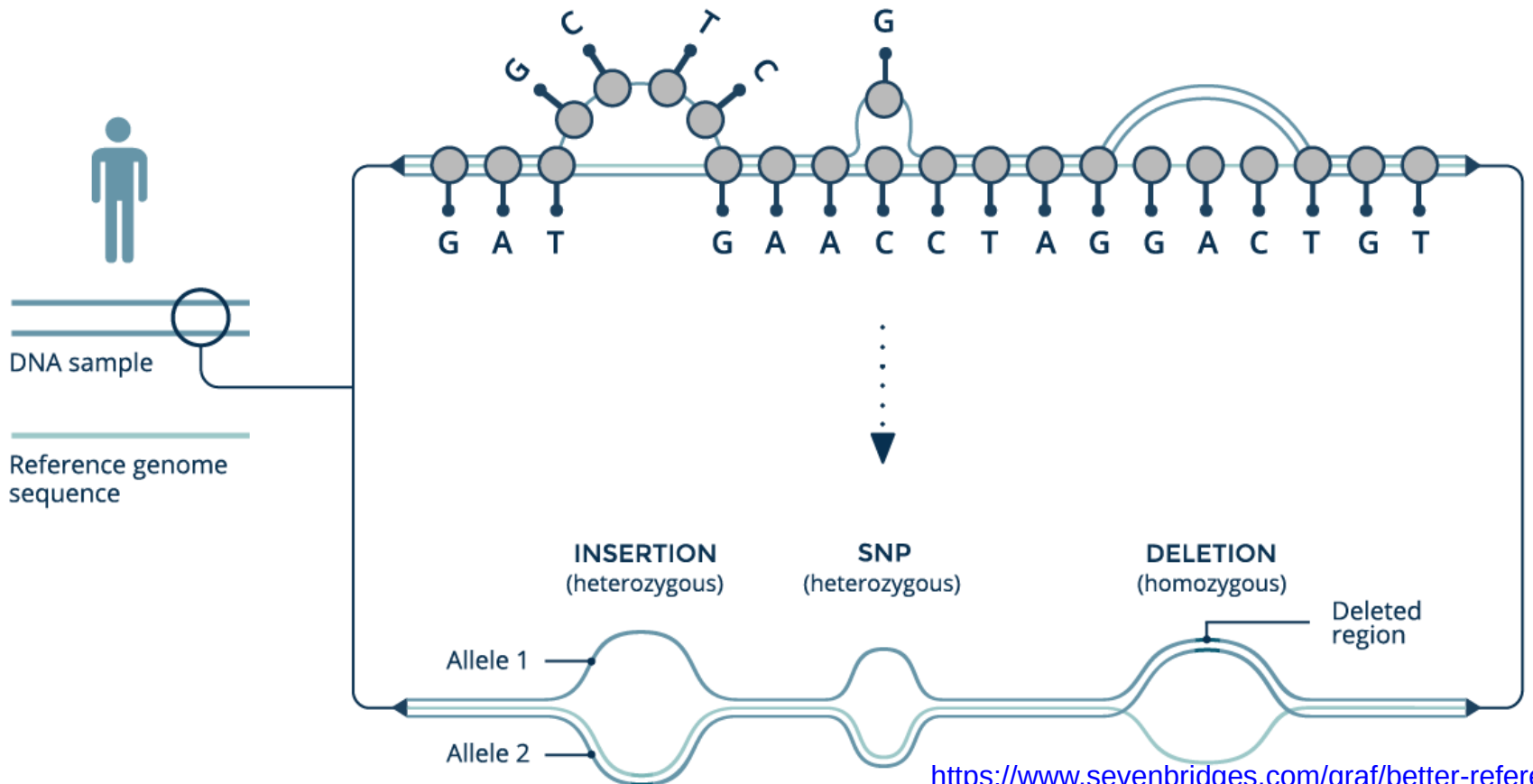


Eizenga JM, et al. 2020. Annu. Rev. Genom. Hum. Genet. 21:139–62

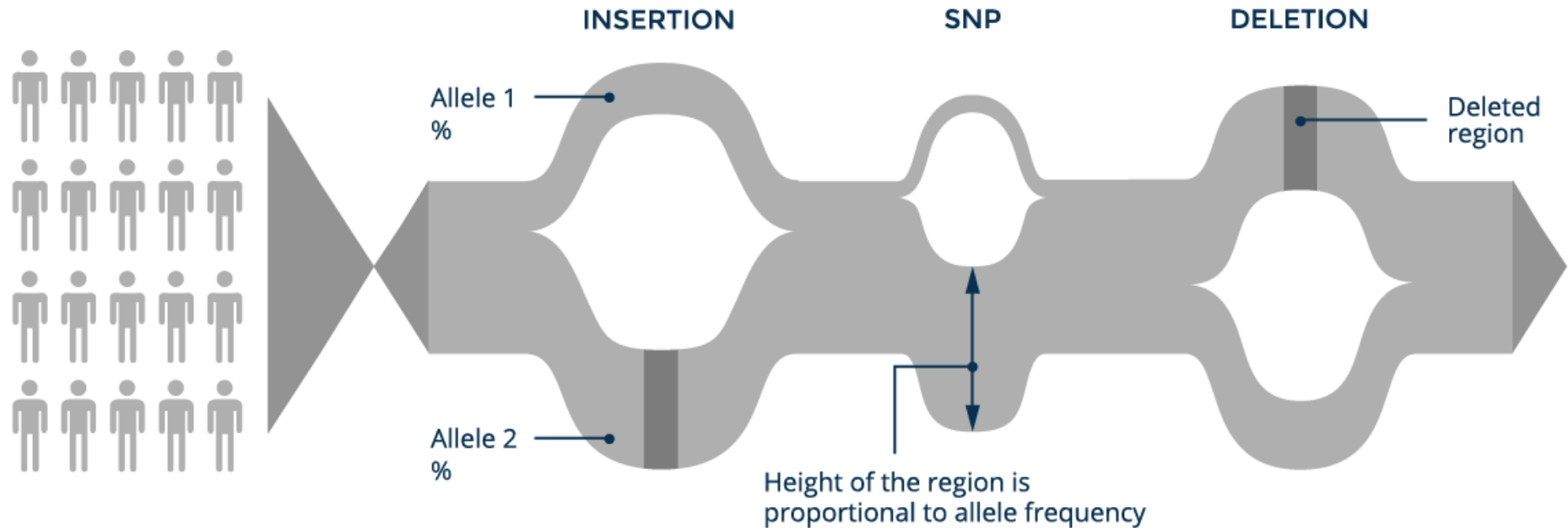
Pangenome graphs. *Annu Rev Genomics Hum Genet.* 2020 <https://dx.doi.org/10.1146%2Fannurev-genom-120219-080406>

Pangenome graph

A genome graph is built by starting with a conventional genomic reference sequence and adding new variants as additional branches. The power of the graph structure is that it can be updated as new haplotypes are discovered.



Pangenome graph: world population

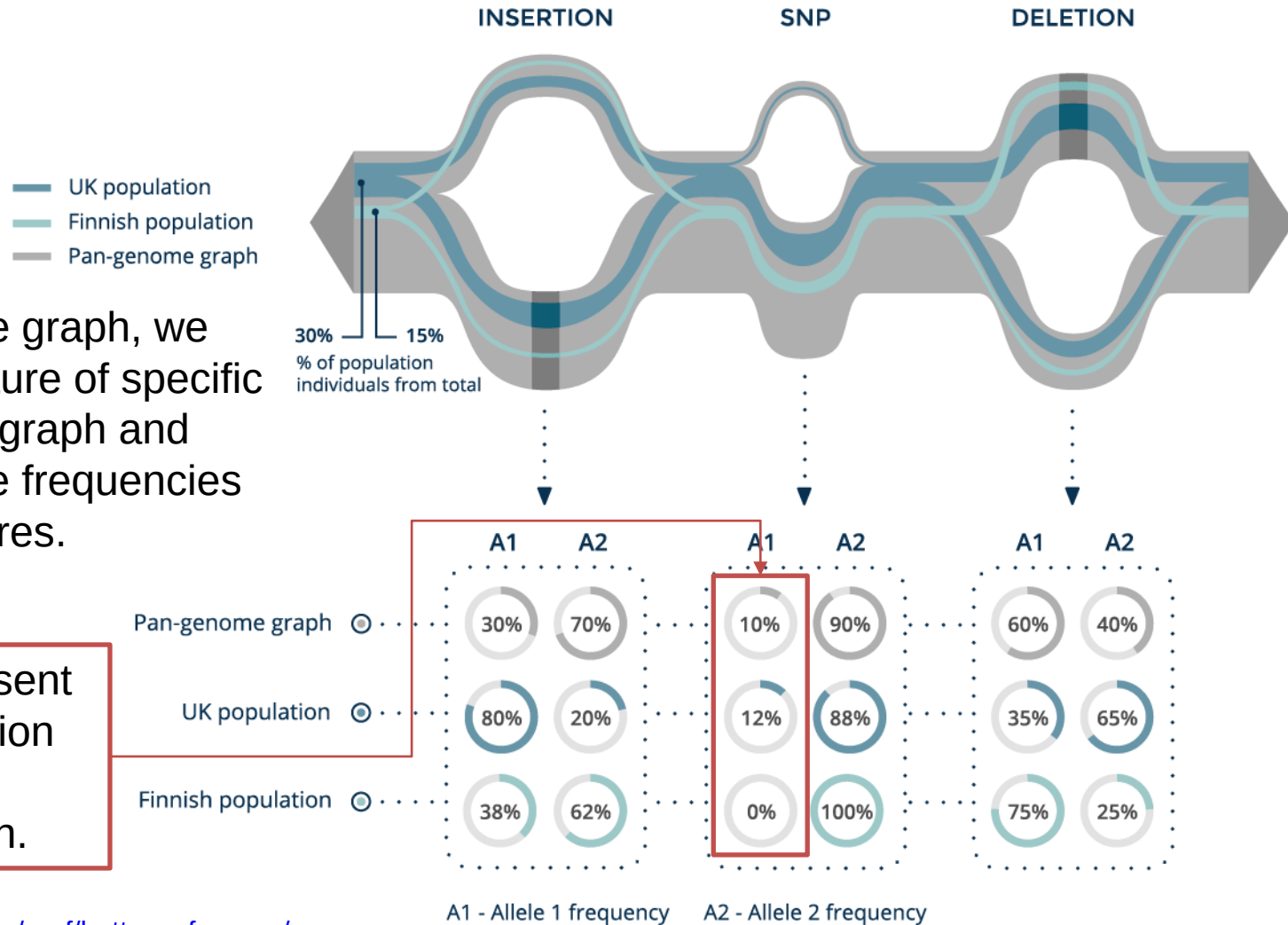


If we take all the genomic information available worldwide and put it into a single graph structure, we can build a world population or pan-genome graph. **Each branch has an allele frequency associated with it**, which measures how often it appears in the individuals included in the graph.

<https://www.sevenbridges.com/graf/better-reference/>

Pangenome graph: population tracing

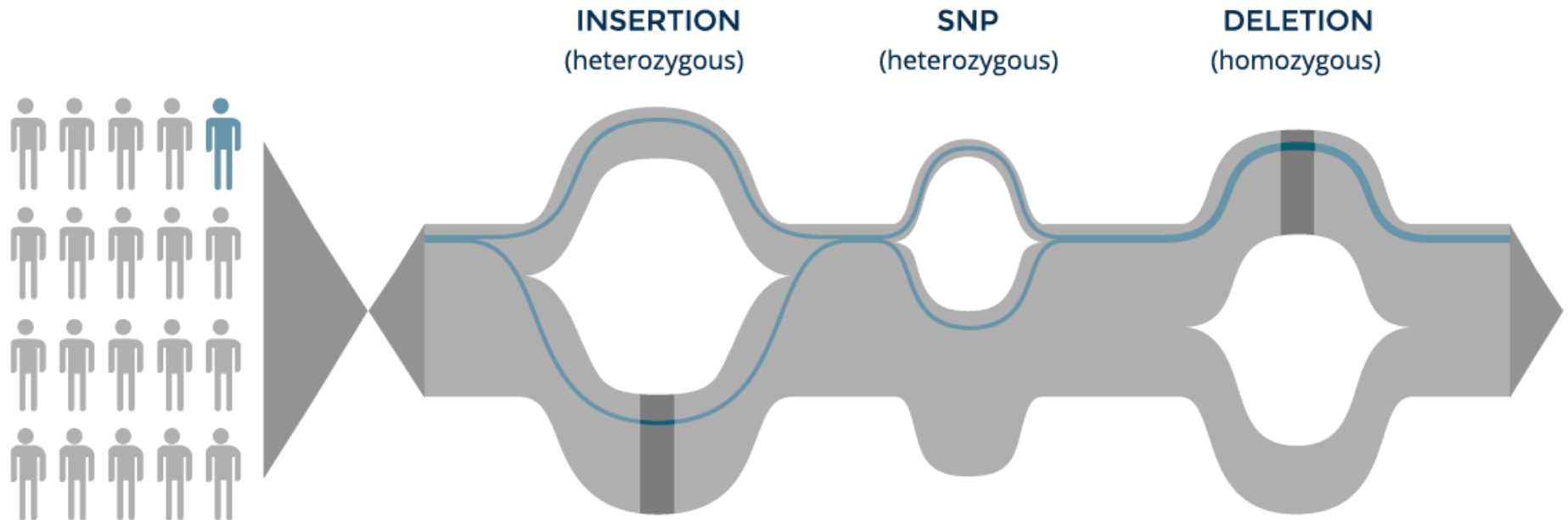
With a pan-genome graph, we can trace the structure of specific populations on the graph and compare their allele frequencies and linkage structures.



a SNP that is present in the UK population but absent in the Finnish population.

<https://www.sevenbridges.com/graf/better-reference/>

Pangenome graph: individual tracing

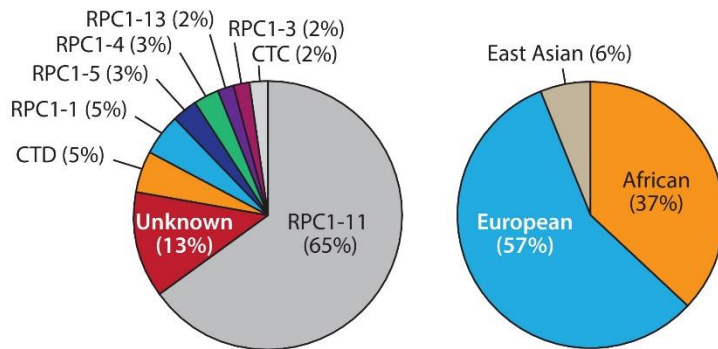


The genome of each individual is a set of paths along the population graph and can be tracked to compare his or her variants and their allele frequencies to the population frequencies. On the other hand, each individual can be anonymized by not storing path information, whereby his or her variants are aggregated to the population graph with no link to the individual.

<https://www.sevenbridges.com/graf/better-reference/>

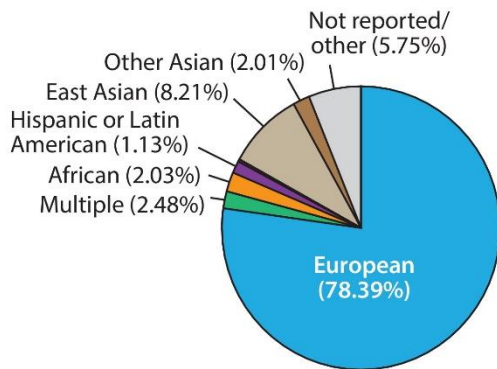
The Human Pangenome Reference Initiative

a



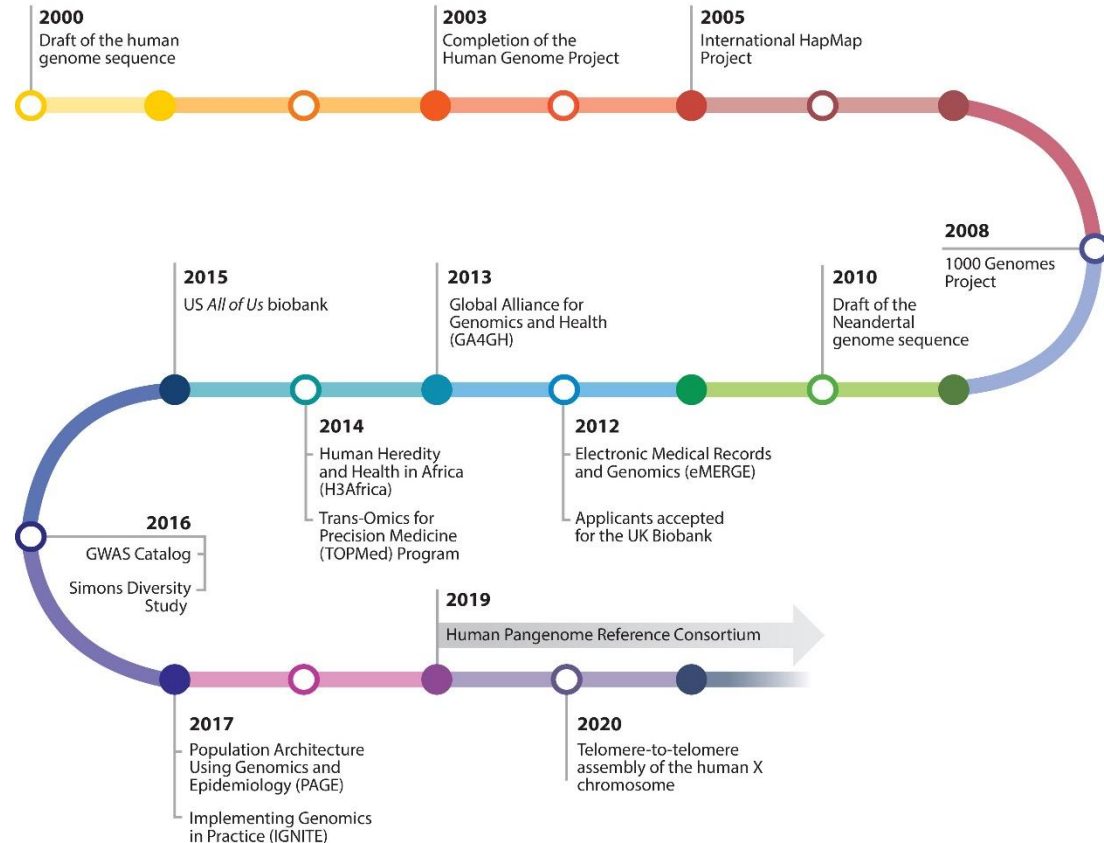
The majority of the reference human genome sequence (65%) is derived from a single bacterial artificial chromosome library (RPC1-11).

b



Also reported in GWAS databases, where the vast majority of data are from people of European ancestry (78%).

c



The Need for a Human Pangenome Reference Sequence. *Annual Review of Genomics and Human Genetics* 2021

<https://doi.org/10.1146/annurev-genom-120120-081921>

The Human Pangenome Reference Initiative

a Population representation and sampling



- Population genetics
- Ethics, policy, and oversight
- Production of cell line resources

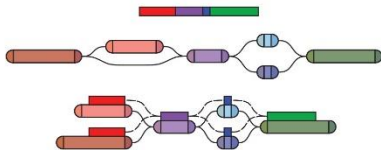
b Multicenter sequencing technology and production



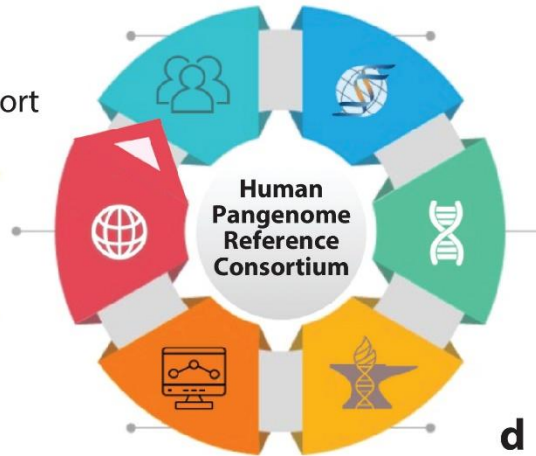
f Outreach and education to support implementation



e Human pangenomics: genome reference representations

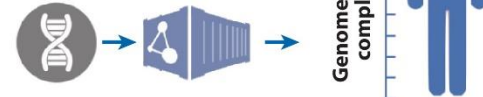


- Development of pangenome reference data structure
- Support for standard analyses



c Reproducible workflows to reach finished telomere-to-telomere assemblies

- Automated and scalable workflow
- Goal: telomere-to-telomere diploid genome sequence



d Open data sharing and cloud-based data management



The Need for a Human Pangenome Reference Sequence. *Annual Review of Genomics and Human Genetics* 2021

<https://doi.org/10.1146/annurev-genom-120120-081921>

A draft human pangenome reference

Article | [Open Access](#) | Published: 10 May 2023

A draft human pangenome reference

[Wen-Wei Liao](#), [Mobin Asri](#), [Jana Ebler](#), [Daniel Doerr](#), [Marina Haukness](#), [Glenn Hickey](#), [Shuangjia Lu](#), [Julian K. Lucas](#), [Jean Monlong](#), [Haley J. Abel](#), [Silvia Buonaiuto](#), [Xian H. Chang](#), [Haoyu Cheng](#), [Justin Chu](#), [Vincenza Colonna](#), [Jordan M. Eizenga](#), [Xiaowen Feng](#), [Christian Fischer](#), [Robert S. Fulton](#), [Shilpa Garg](#), [Cristian Groza](#), [Andrea Guarracino](#), [William T. Harvey](#), [Simon Heumos](#), ... [Benedict Paten](#) 

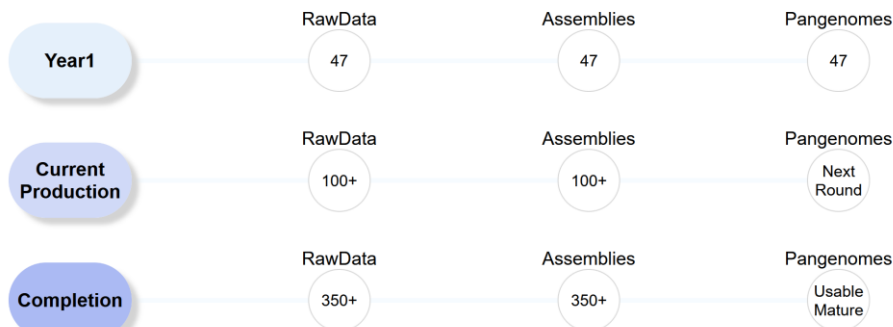
[+ Show authors](#)

[Nature](#) **617**, 312–324 (2023) | [Cite this article](#)

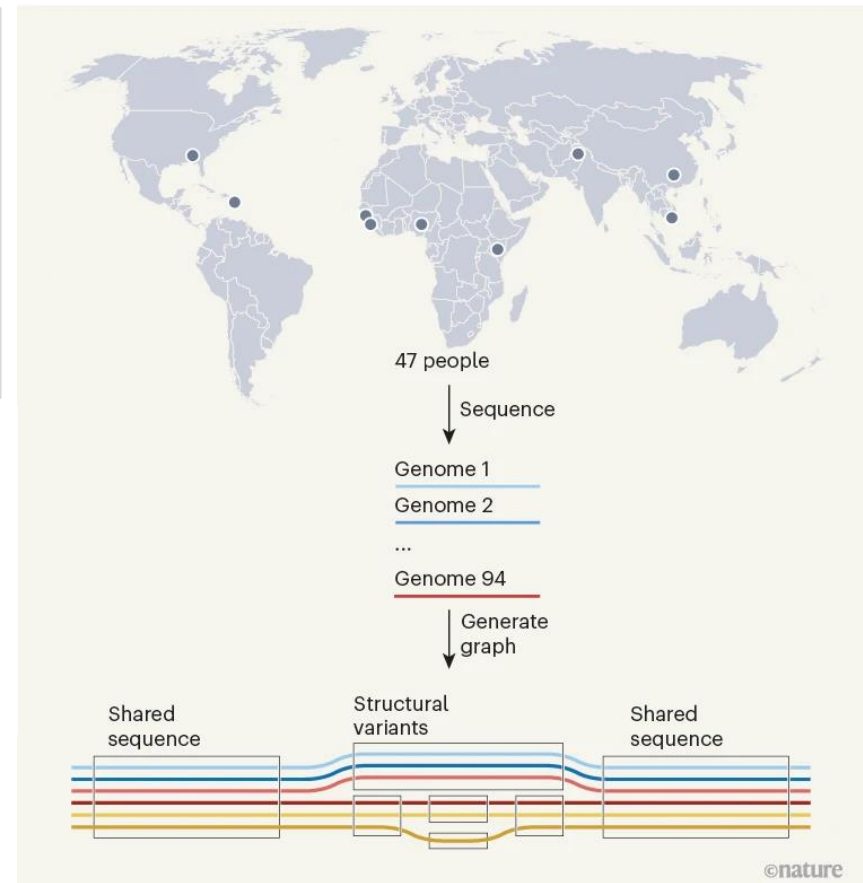


Human Pangenome Reference Consortium

The Human Pangenome Reference Consortium (HPRC) is a project funded by the National Human Genome Research Institute to sequence and assemble genomes from individuals from diverse populations in order to better represent genomic landscape of diverse human populations.



<https://humanpangenome.org>



<https://www.nature.com/articles/d41586-023-01490-3>

47个二倍体基因组，与GRCh38相比，增加了1.19 亿个碱基对和1115个基因重复，结构变异数量增加了104%

Metagenome

宏基因组

Metagenome and Metagenomics

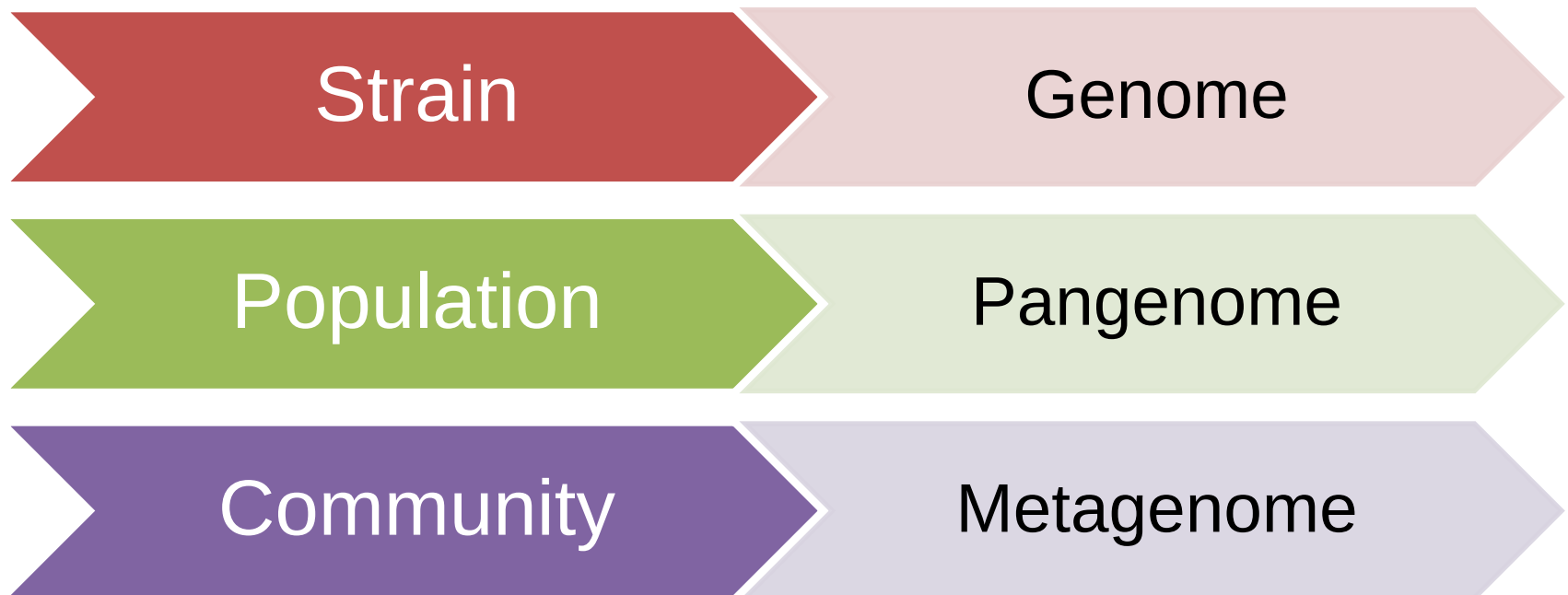
- 宏基因组 (Metagenome, 也称环境基因组或元基因组), 是由Handelsman等1998年提出的新名词, 其定义为
“the genomes of the total microbiota found in nature”,
即**生境中全部微小生物遗传物质的总和**, 包含了可培养的和未可培养的微生物的基因。
- 宏基因组学 (metagenomics) 就是一种以环境样品中的微生物群体基因组为研究对象, 以功能基因筛选和/或测序分析为研究手段, 以微生物多样性、种群结构、进化关系、功能活性、相互协作关系及与环境之间的关系为研究目的的新的微生物研究方法。

Metagenomics is the study of genetic material recovered directly from environmental samples. The broad field may also be referred to as environmental genomics, ecogenomics or community genomics.

Metagenome vs. Pangenome

- Pangenome is the global gene repertoire of a bacterial species.
- Metagenome is the genetic material present in an environmental sample that consists genomes of many individual organisms.

- **culturable / non-culturable** (~99% of microbial species cannot be cultivated)
- **known / unknown**



Metagenomics Applications

- 人类健康 (疾病、机体免疫、食物消化)
- 农业 (1克的土中包含约 10^9 - 10^{10} 个微生物细胞)
- 生物多样性
- 生态环境
- 生化燃料
- 生物技术
- 人工生物环境 (如废水处理、食品发酵、堆肥、沼气池)



取样来源：土壤、水体（湖泊、河流、海洋等）、空气、肠道、口腔、皮肤等

宏基因组学发展

1991年首次提出环境基因组学的概念

1998年提出生命研究对象应是生物环境中全部微小生物的基因组，针对样品中细菌和真菌的基因组总和进行研究

广义宏基因组是指特定环境下所有生物遗传物质的总和，它决定了生物群体的生命现象

人类宏基因组学研究人体宏基因组结构和功能、相互之间关系、作用规律和与疾病的关系

环境
基因组学

微生物
基因组学

宏
基因组学

人类宏
基因组学

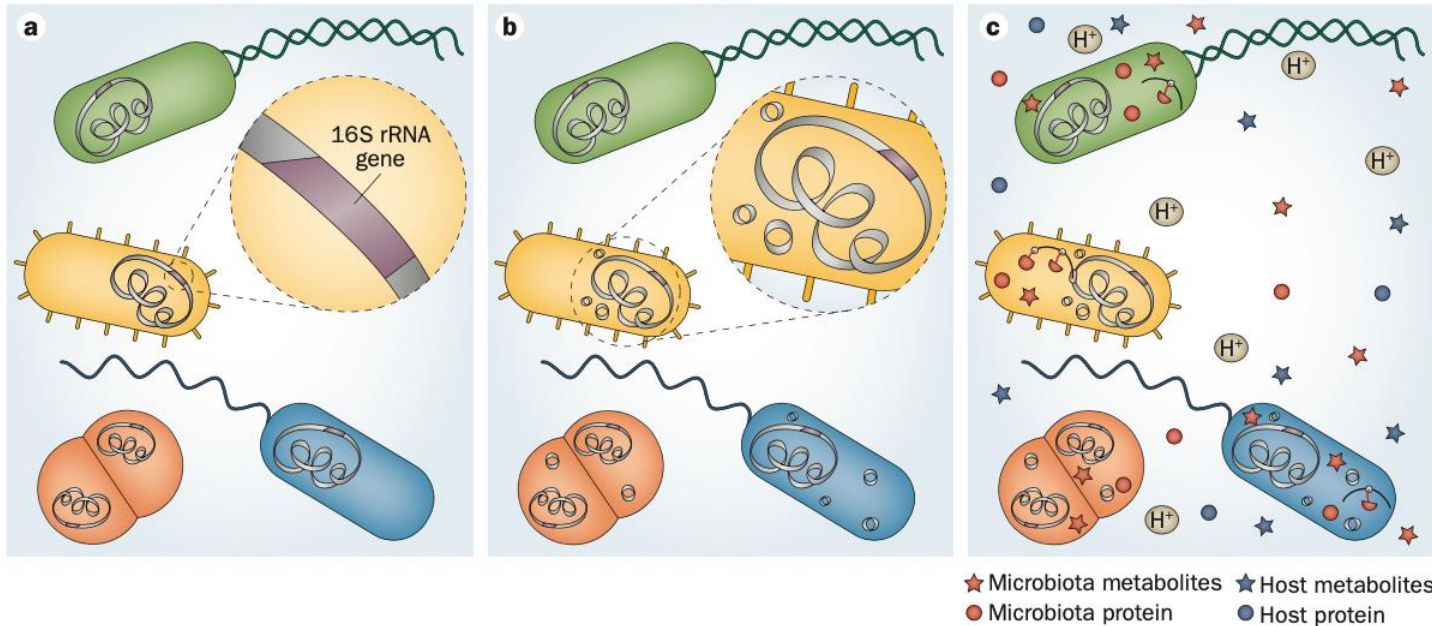
1998年美国启动环境基因组计划

2007年3月，美国国家科学院以“环境基因组学新科学——揭示微生物世界的奥秘”为题发表咨询报告

把人体内所有微生物菌群基因组的总和称为“人类宏基因组” (human metagenome)

目标是：把人体内共生菌群的基因组序列信息都测定出来，研究与人体发育和健康有关的基因功能

Microbiota, Metagenome & Microbiome



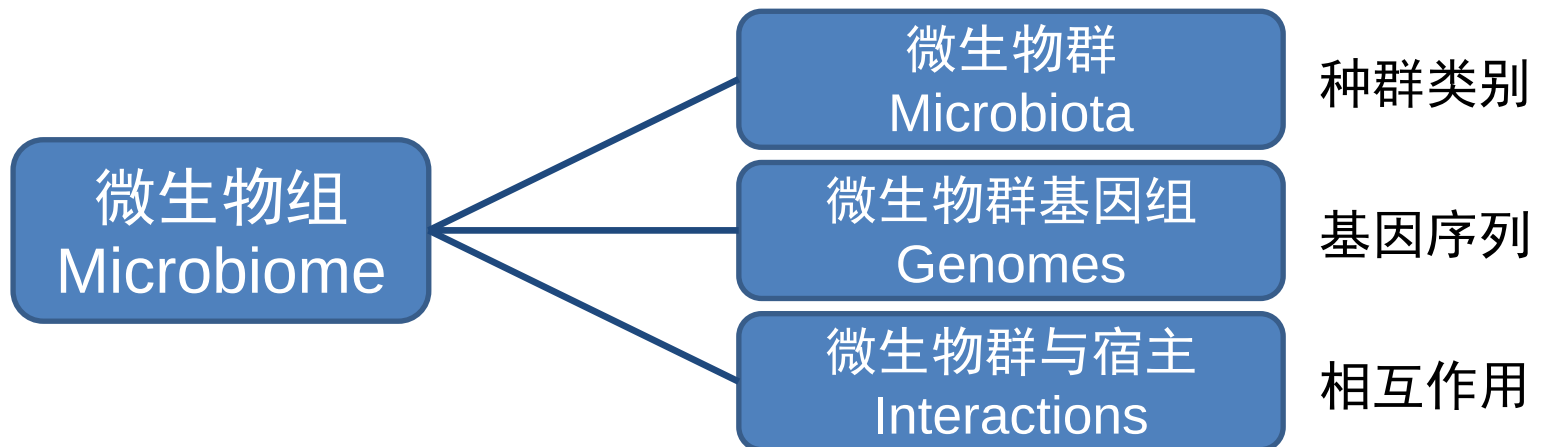
Microbiota 微生物群	Metagenome 宏基因组	Microbiome 微生物组
Microorganisms (by type) living in a specific environment	The genes of microorganisms in a specific environment	Microorganisms (and their genes) living in a specific environment
Refers to the taxonomy (name) of microorganisms in the environment identified based on 16S rRNA.	Refers to the microbial genes and genomes, highlighting the genetic potential of the population.	Refers to the microbial genes and genomes, as well as the products of the microbiota and the host environment.

The microbiome of the urinary tract—a role beyond infection. *Nat Rev Urol* 2015

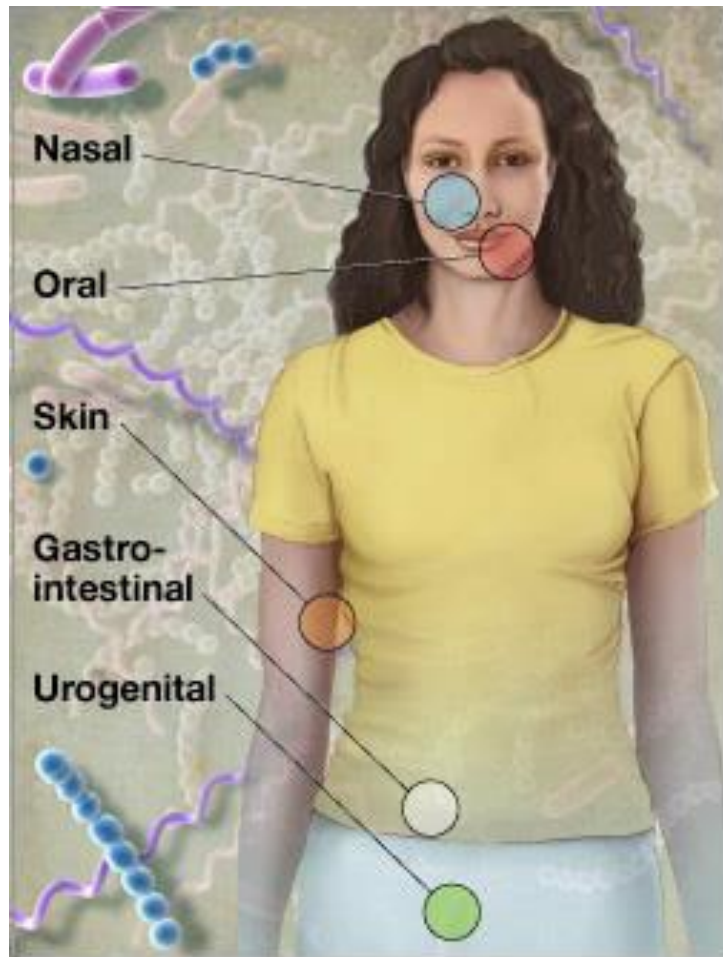
Microbiome – 微生物组

A microbiome is the collection of all microbes, such as bacteria, fungi, viruses, and their genes and genomes that naturally live on our bodies and inside us.

- **Diversity:** human microbiome is diverse and each body site – the gut, skin, and oral and nasal cavities – is home to a unique community of microbes.
- **Role:** microbes contribute to human health and wellness in many ways.
 - Microbes protect us against pathogens, help our immune system develop, and enable us to digest food to produce energy.
 - Some microbes alter environmental chemicals in ways that make them more toxic, while others act as a buffer and make environmental chemicals less toxic.



人类微生物组

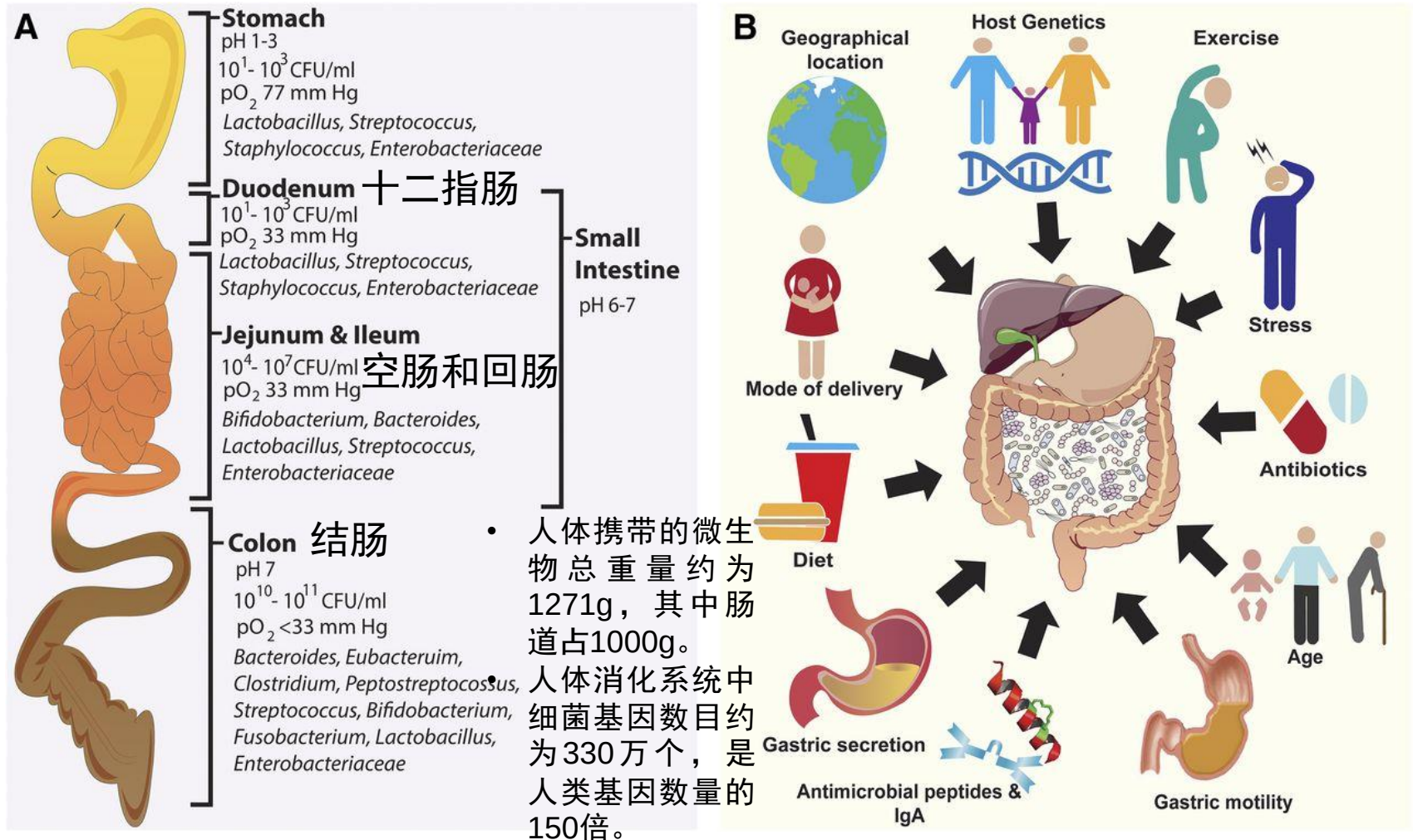


人类微生物组是指与人类共生的全部微生物的总和

- 人体内的微生物种类多达3万种，数量约100兆至1000兆，特别是胃肠道内的微生物最为丰富，又被称为**人类第二基因组**。
- **肠道菌群**（Gut或gastrointestinal microbiota）：指在消化道内生存的微生物复杂群落，宏基因组在狭义上指的是肠道源基因组。
- **肠-脑轴**（gut-brain axis）：是大脑和肠消化道两个器官间的沟通桥梁，而其中肠道中的菌群也对此路径贡献匪浅，三者相互影响并调控全身各种生理作用，从脑部早期发育到晚期老年的神经疾病的皆与此连结轴线有着密切的关系。

<https://www.genome.gov/27528490/2008-release--researchers-establish-international-human-microbiome-consortium>

Gut Microbiome



Gut Reactions: Breaking Down Xenobiotic–Microbiome Interactions. *Pharmacological Reviews* 2019

宏基因组学是研究人体微生物组的主要方法

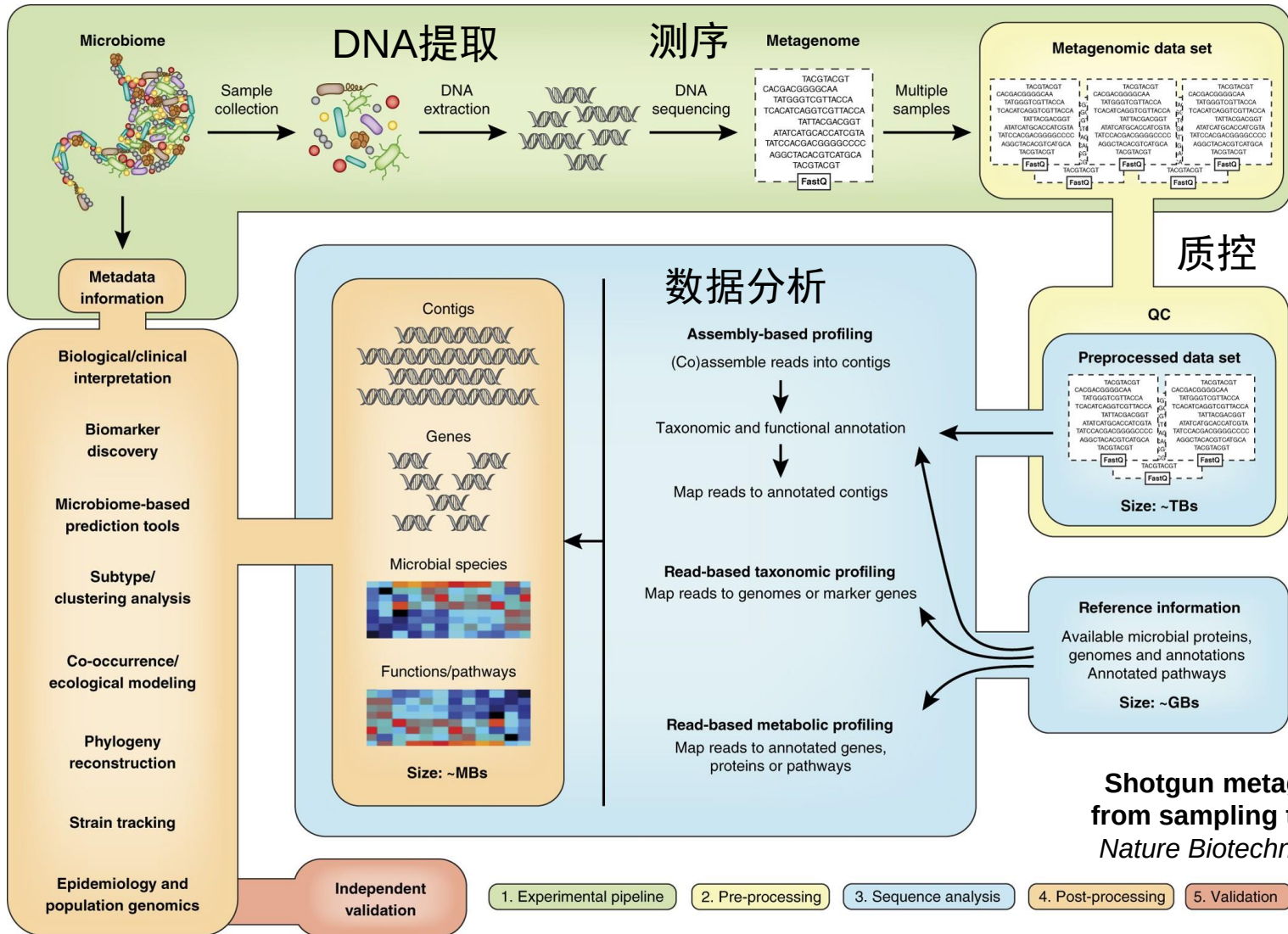
- 宏基因组学是在微生物基因组学的基础上发展起来的一种研究微生物多样性、开发新的生理活性物质（或获得新基因）的新理念和新方法。
- 通过直接从环境样品中提取全部微生物的DNA，构建宏基因组文库，利用基因组学的研究策略，研究环境样品所包含的全部微生物的遗传组成及群落功能。

人体“微生物组”是人体微生物及微生物基因组的总和。

——诺贝尔奖得主Joshua Lederberg于2001年提出

宏基因组学研究方法

多样性功能分析, 包括物种组成、丰度及其演化等



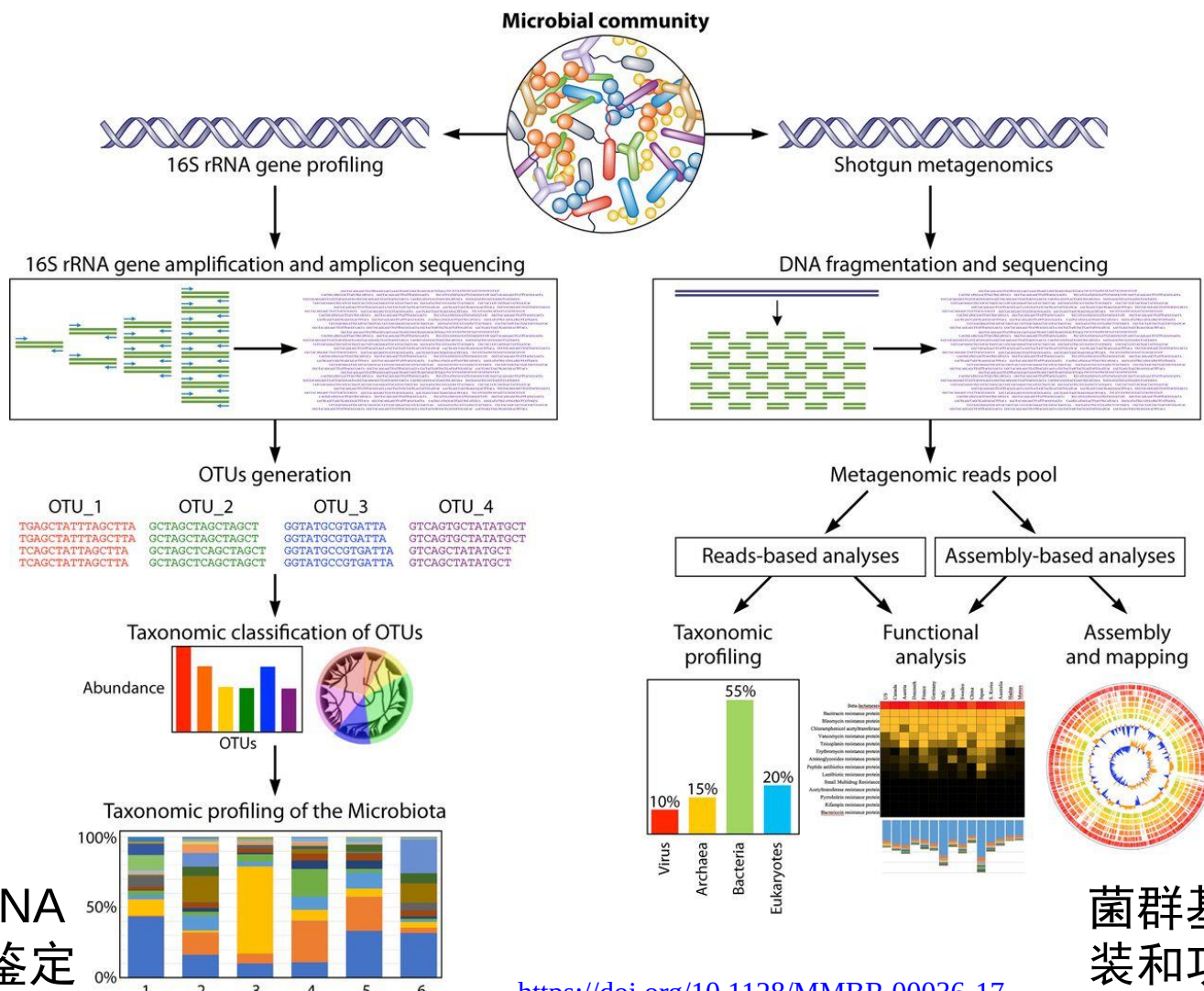
Operational Taxonomic Unit (OTU)

可操作分类单元 (Operational Taxonomic Unit, OTU)

- 为便于进行分析，人为给某一个分类单元（品系、种、属、分组等）设置的同一标志
- 每一个OTU通常被视为一个微生物物种
- 用于估计样本中的物种组成及丰度
- OTU代表生物多样性程度，在典型情况下：
 - 原核生物的OTU使用16S rDNA来衡量
 - 真核生物的OTU使用18S rDNA来衡量

- An **operational taxonomic unit (OTU)** is an operational definition used to classify groups of closely related individuals.
- OTU also refers to clusters of (uncultivated or unknown) organisms, grouped by DNA sequence similarity of a specific taxonomic marker gene (originally coined as mOTU; molecular OTU).

宏基因组学生物信息分析流程



基于16S rRNA
的菌群丰度鉴定

菌群基因组组
装和功能分析

<https://doi.org/10.1128/MMBR.00036-17>



人类微生物组计划

Human Microbiome Project



Terabases

美国NIH2008年发起，旨在鉴定与阐明和人类健康与疾病相关的微生物功能的计划

<https://commonfund.nih.gov/hmp>

Gigabases

600 Mb Open ocean

150 Mb Lean mouse

80 Mb Deep sea

Megabases

10 Mb Diarrheal illness

250 Mb Human microbiome

4 Gb Coastal ocean

2 Gb Lean/obese gut

MetaHit 600Gb

130 Gb Soil

14Gb Cow rumen

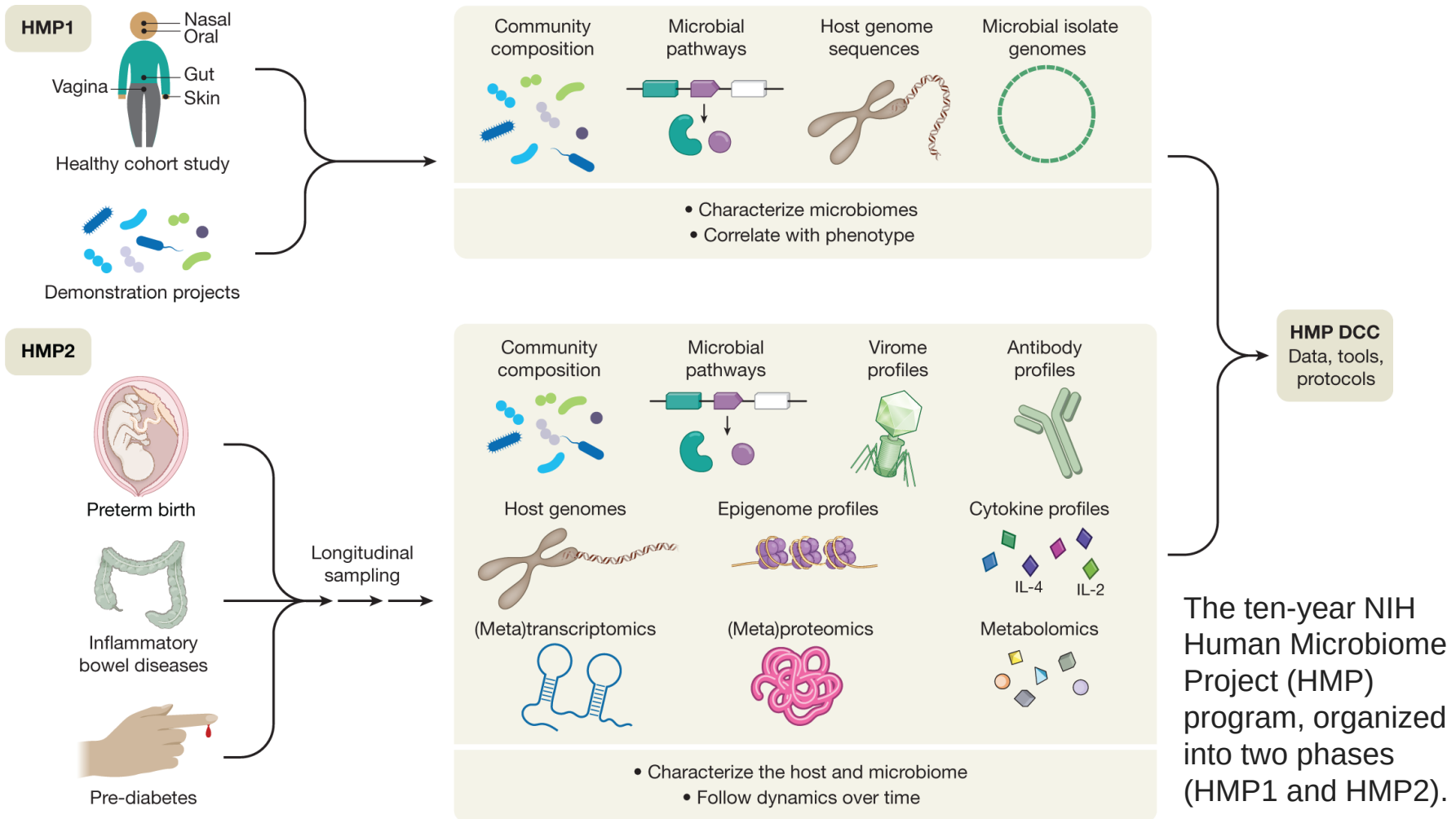
HMP WGS 3.5Tb

100 Gb human gut

30 Gb HMP 16S

The Human Microbiome Project: A Community Resource for the Healthy Human Microbiome. *PLoS Biology* 2012

人类微生物组计划HMP



The Integrative Human Microbiome Project. *Nature* 2019

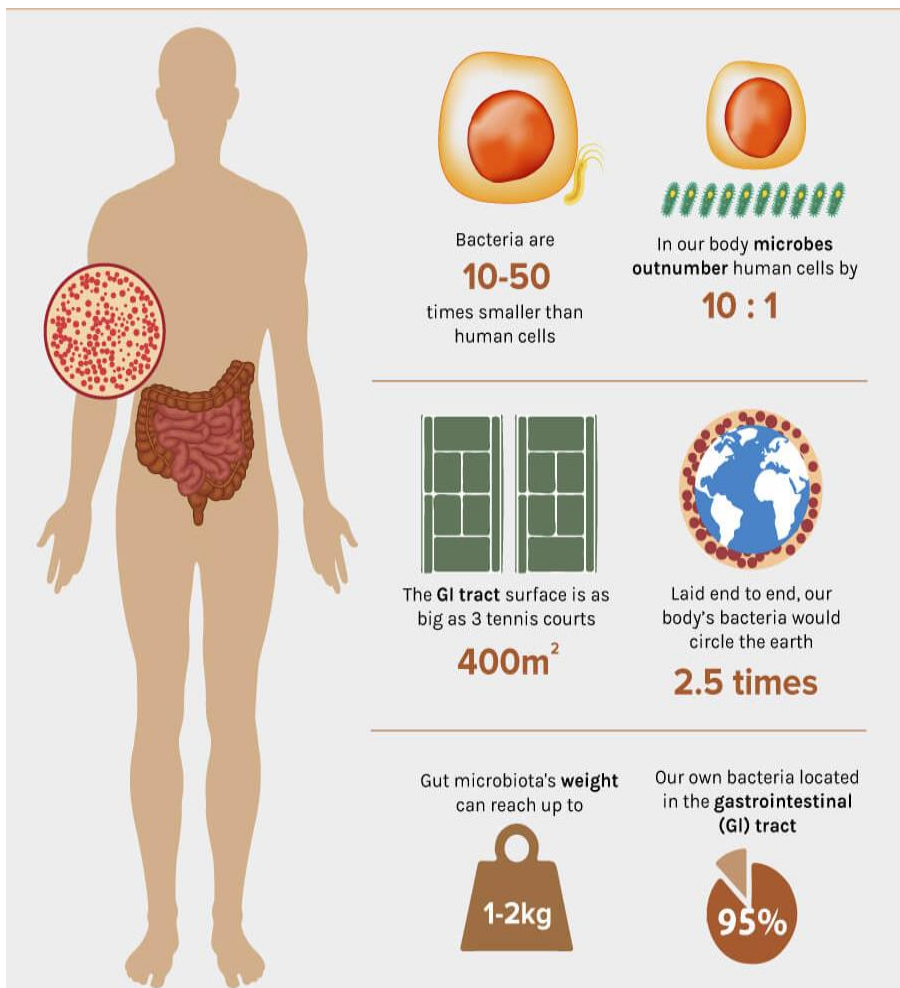
人体肠道宏基因组研究计划(MetaHIT)

2008年，欧盟倡议并启动了“人体肠道宏基因组计划”（Metagenomics of the Human Intestinal Tract, MetaHIT），由来自英国、法国、德国、丹麦、意大利、中国等九个国家的研发团队共同承担。



<https://cordis.europa.eu/project/id/201052>

正常微生物组生理作用



1. **生物拮抗 (antagonism)**: 抵抗外来致病菌，维持正常菌群内部的平衡，保护机体免受感染
2. **营养作用**: 参与机体代谢，产生维生素等
3. **免疫作用**: 促进免疫器官发育，刺激免疫系统应答
4. **抗衰老作用**: 产生过氧化物歧化酶(SOD)，清除自由基($O_2^{\cdot-}$)
5. **抗肿瘤作用**:
 - 产生自身抑癌产物
 - 激活巨噬细胞等发挥免疫功能抑制肿瘤

<https://drjockers.com/10-ways-improve-gut-microbiome/>

肠道菌群影响人类健康

Intestinal Flora Affects Your Health

The microbes that live inside your intestines influence your health in beneficial and harmful ways

Immunity

Providing a physical barrier to invasive microbes, our gut flora enhances the functionality of the immune system.



Vitamins

Bacteria in the gut plays a direct role in the synthesis of vitamins B and K as well as the absorption of calcium and iron.



Metabolism

Metabolic activity of the gut flora allows our body to utilize food that would otherwise not be digested.



Obesity

In 2009, Dr. Krajmalnic-Brown discovered gut bacteria of obese patients differs significantly from normal individuals.



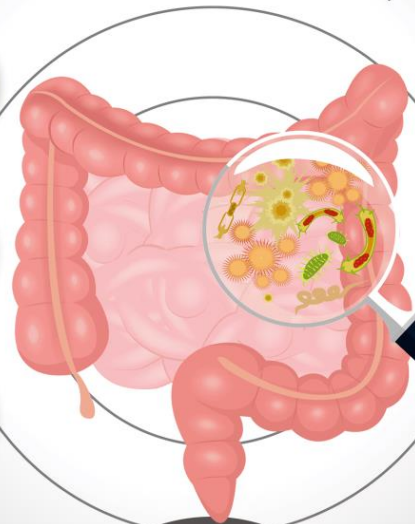
Inflammation

Gut flora likely plays a major role in the development of various inflammatory diseases including IBD and colitis.



Autism

New research by Dr. Krajmalnic-Brown suggests a link between autism and decreased gut bacterial diversity.



Bad Gut Microbes



Increased Inflammation



Poor Quality of Life

DRJOCKERS.COM
SUPERCHARGE YOUR HEALTH

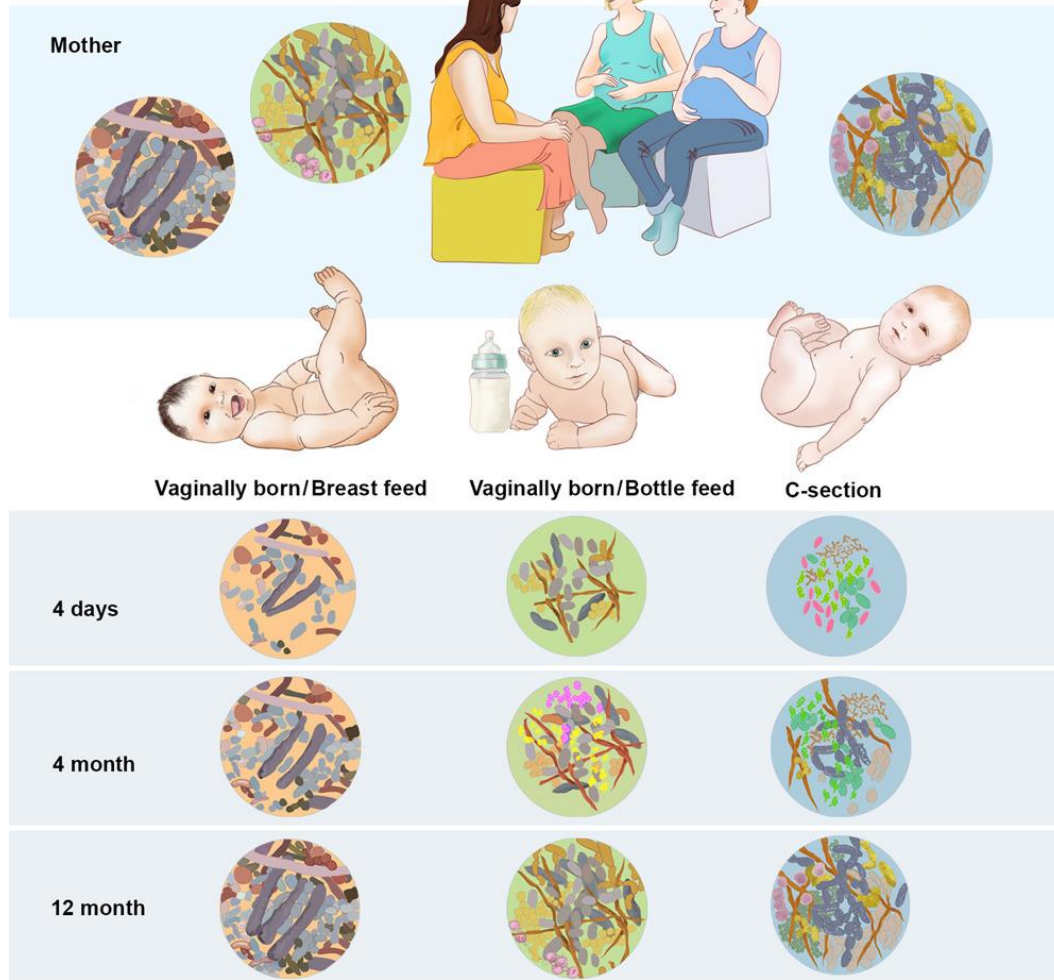
<https://drjockers.com/10-ways-improve-gut-microbiome/>



中国科学院大学
University of Chinese Academy of Sciences

基因组学 - Genomics

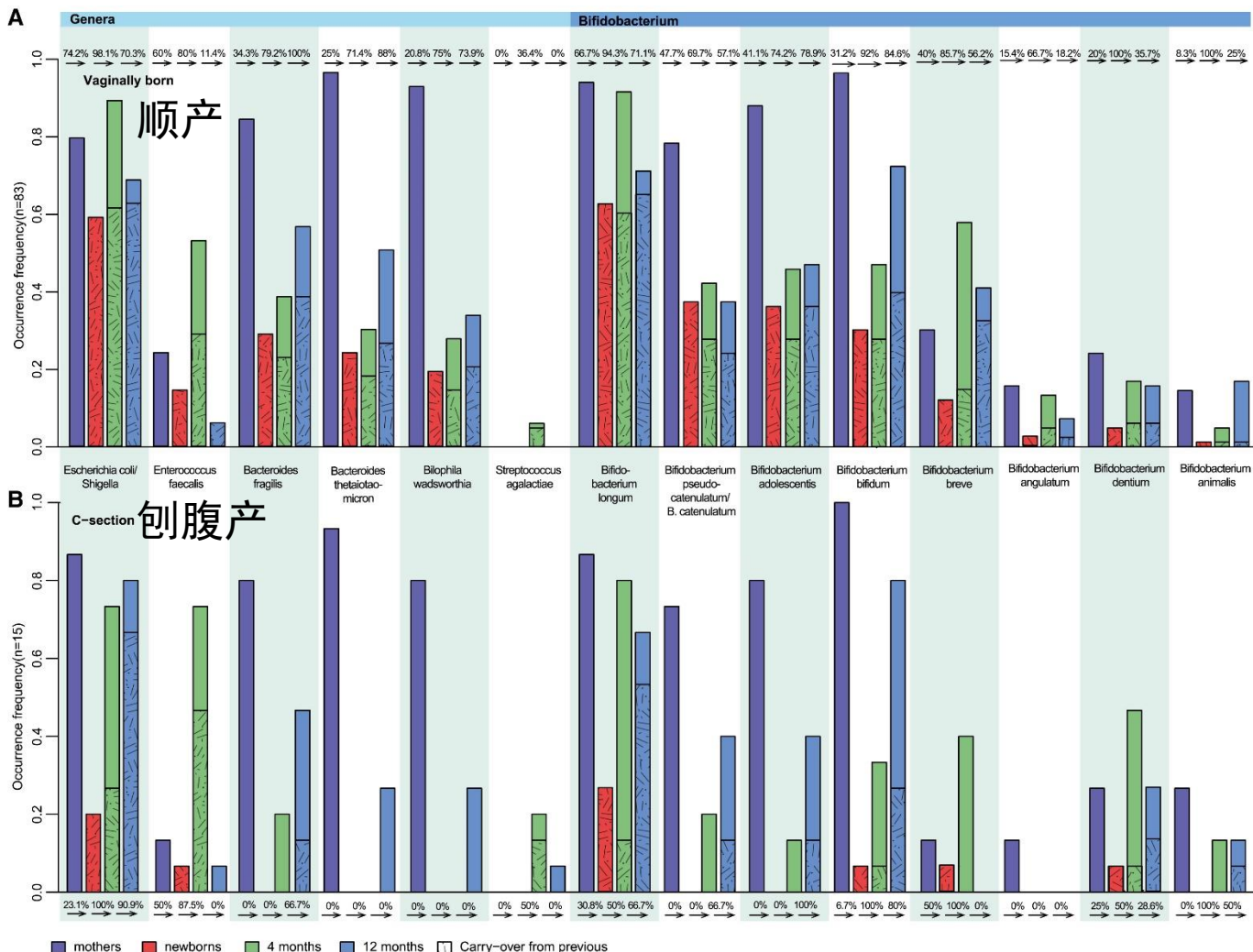
肠道菌群的起源



- 肠道菌群是“外来户”
- 胎儿肠道内是无菌的
- 当胎儿出生之后的几天内，细菌通过分娩时阴道物质摄入、哺乳时的口腔摄入以及空气吸入等途径进入新生儿体内，并在肠道内定植，形成新生儿最初的肠道菌群
- 随着婴儿的成长，肠道菌群的种类结构逐渐趋于稳定，最终形成成熟的肠道菌群

Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host & Microbe* 2015

顺产婴儿肠道菌群更类似成人

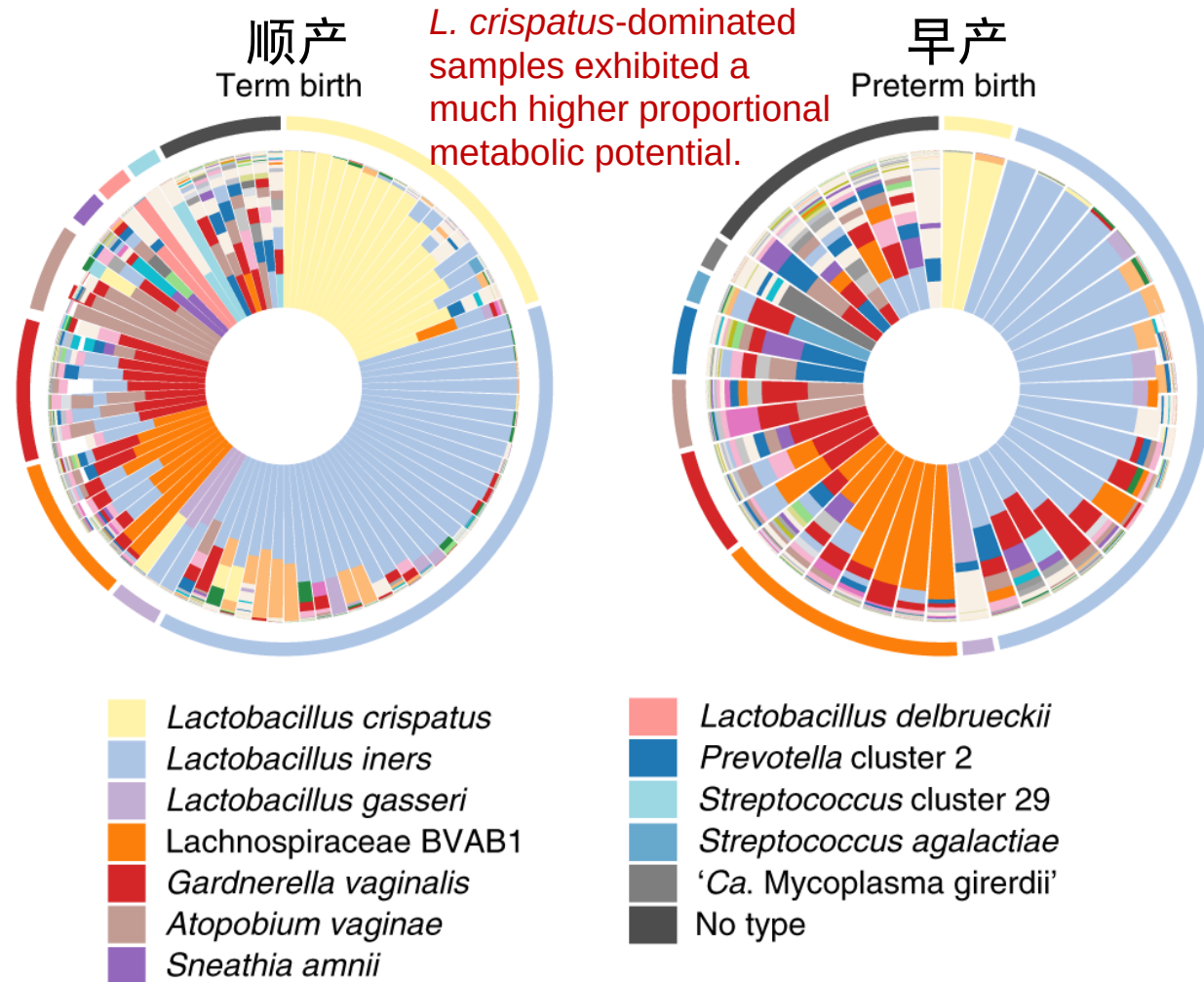


- 样本：98名母亲及其新生儿的粪便样本及婴儿4个月和12个月的粪便样本，共鉴定了四千多个新细菌基因组。
- 剖腹产婴儿的肠道菌群含有许多口腔和皮肤菌，而顺产婴儿则与母亲共享更多的肠道菌。
- 喂母乳的婴儿受到感染的比例比较低

Dynamics and Stabilization of the Human Gut Microbiome during the First Year of Life. *Cell Host & Microbe* 2015

Vaginal microbiome: TB vs PTB

- ~15 million preterm births at less than 37 weeks of gestation occur annually worldwide.
- Preterm birth (PTB) remains the second most common cause of neonatal death across the globe, and the most common cause of infant mortality in middle- and high-income economies.
- Despite these statistics, there remains a paucity of effective strategies for predicting and preventing PTB.

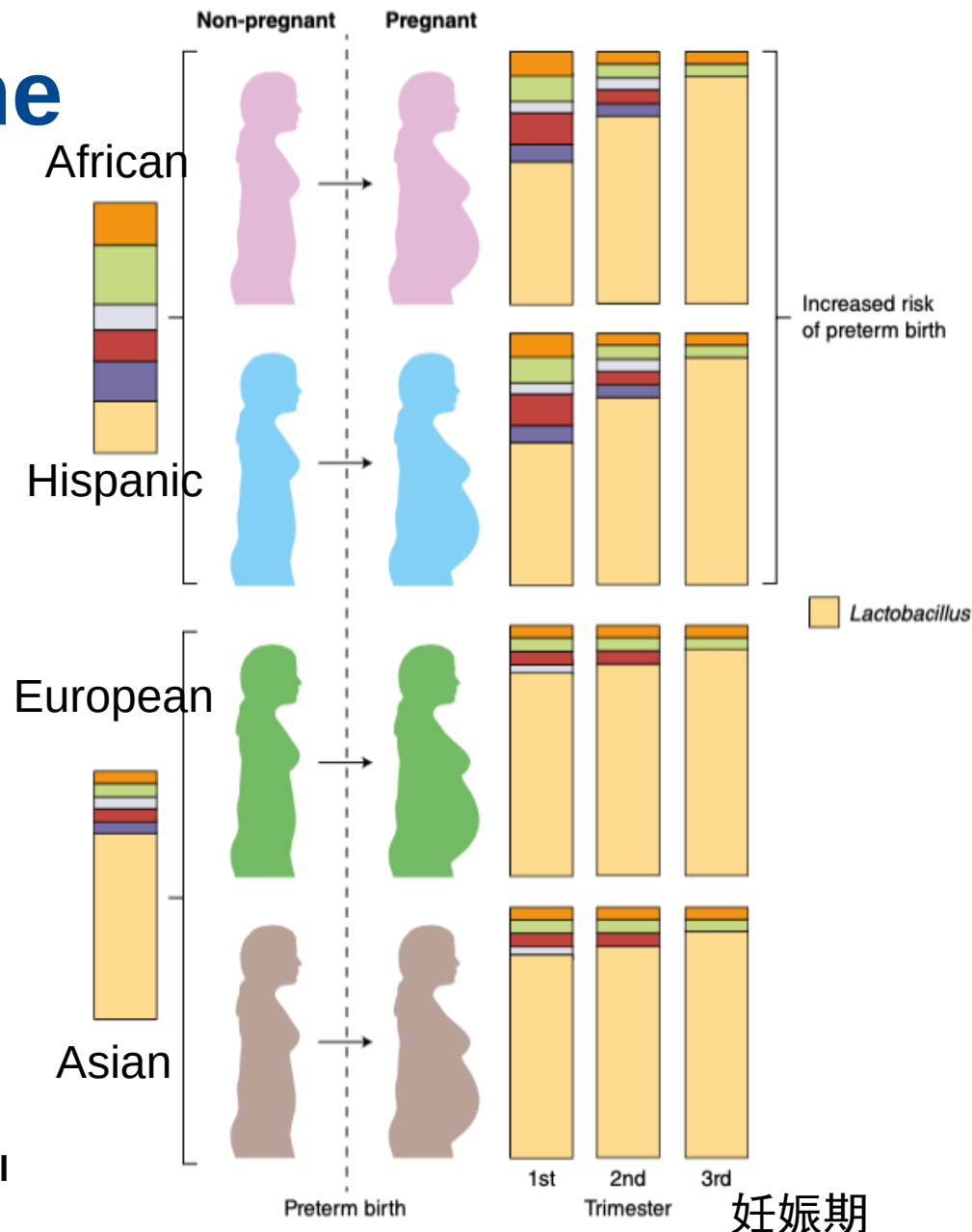


The vaginal microbiome and preterm birth. *Nature Medicine* 2019

Vaginal Microbiome

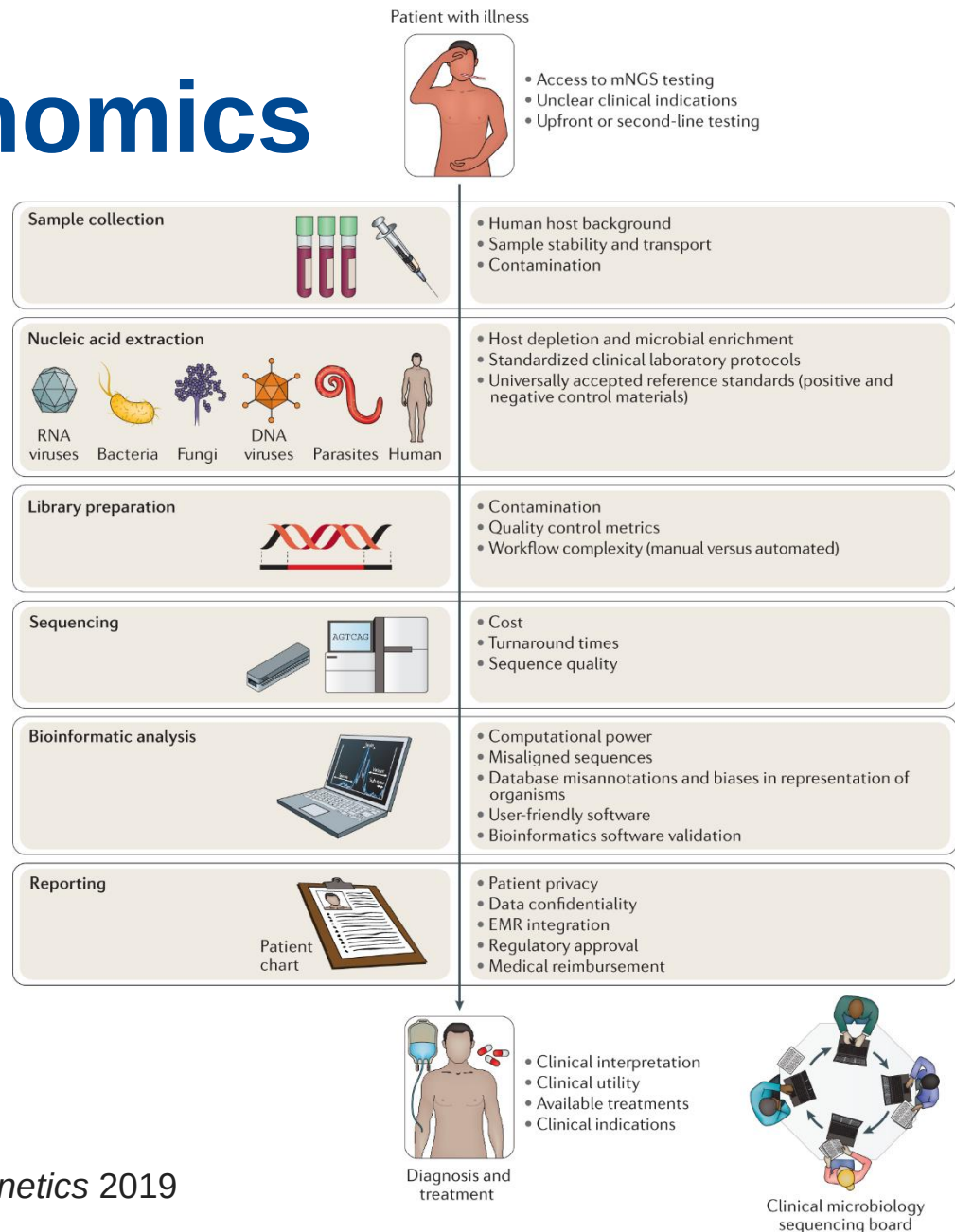
- There is variation in vaginal microbiome **diversity in healthy women according to ethnic group**.
- Women of **African** ancestry show **higher diversity** and women of **European and Asian** ancestry show **lower diversity**.
- High vaginal diversity is associated with bacterial **vaginosis**.
- **Pregnancy leads to the convergence of the maternal microbiome into a low-diversity, Lactobacillus-dominated microbiota.**
- **High vaginal diversity during pregnancy is associated with increased risk of preterm birth.**

Gestational shaping of the maternal vaginal microbiome. *Nature Medicine* 2019



Clinical metagenomics

- Clinical metagenomic next-generation sequencing (mNGS), the comprehensive analysis of microbial and host genetic material (DNA and RNA) in samples from patients, is rapidly moving from research to clinical laboratories.
- The capacity to detect all potential pathogens — bacteria, viruses, fungi and parasites — in a sample and simultaneously interrogate host responses has great potential utility in the diagnosis of infectious disease.

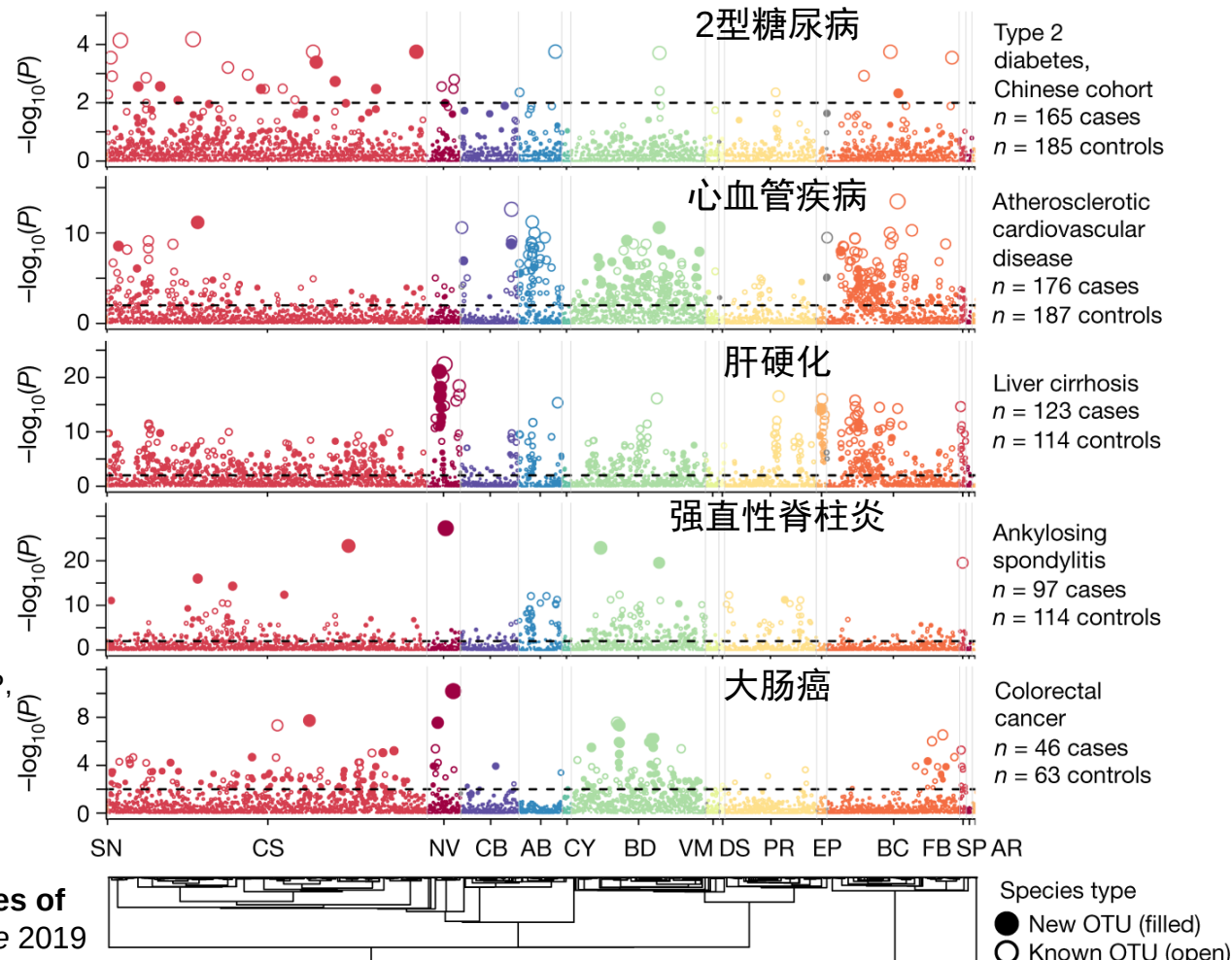


Clinical metagenomics. *Nature Reviews Genetics* 2019

Metagenome-wide association of gut OTUs with human diseases

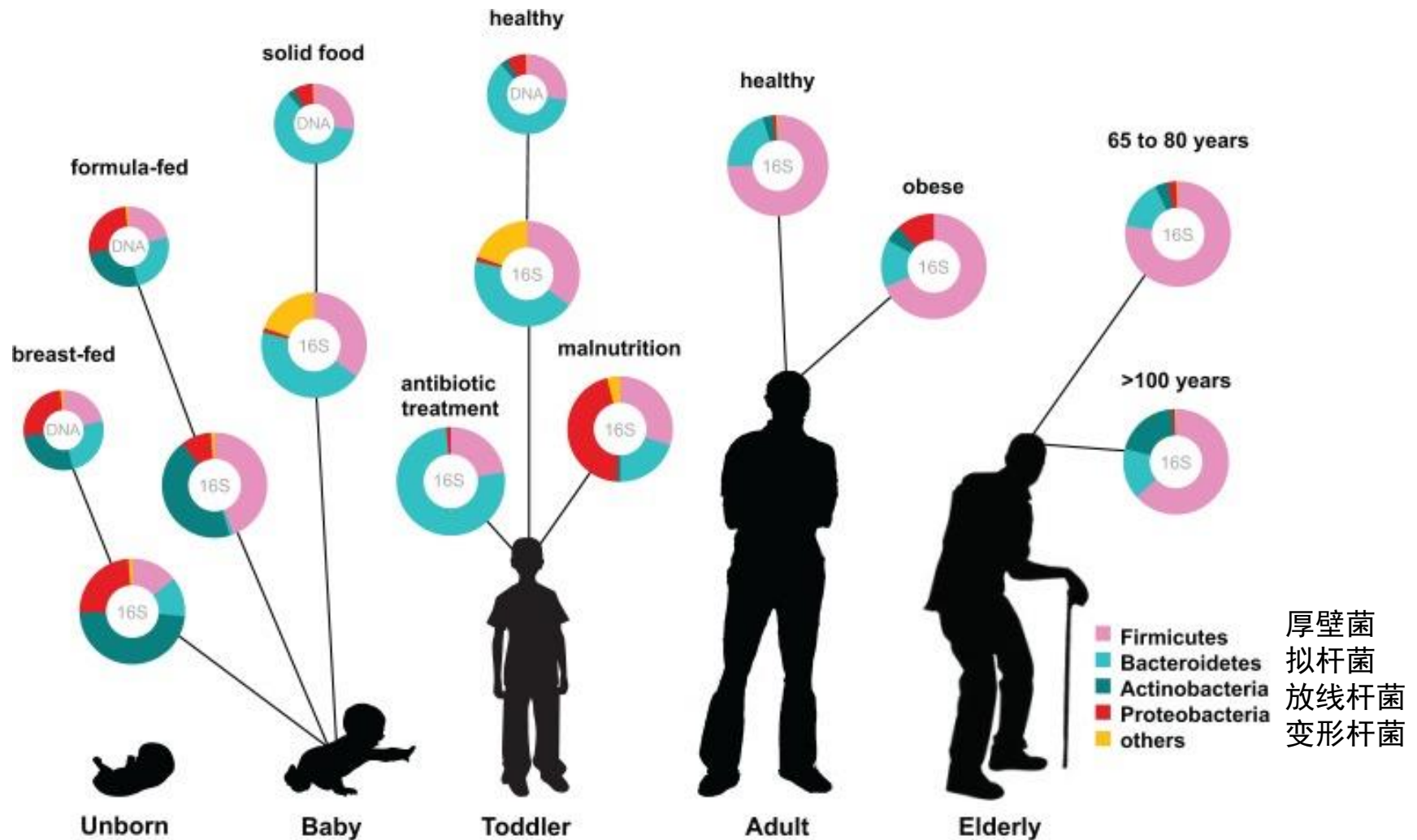
The Manhattan plot shows the phylogenetic distribution of species–disease associations for different metagenomic studies. **Each point is one species-level OTU and point height indicates the P value** from a two-sided Wilcoxon rank-sum test of estimated species abundance between diseased and healthy individuals after correction for multiple hypothesis tests.

AR, Archaea; AB, Actinobacteria; BC, Bacilli; BD, Bacteroidetes; CB, Coriobacteriia; CS, Clostridia; CY, Cyanobacteria; DS, Desulfobacteraeota; EP, Epsilonbacteraeota; FB, Fusobacteria; NV, Negativicutes; PR, Proteobacteria; SN, Synergistetes; SP, Spirochaetes; VM, Verrucomicrobia.



New insights from uncultivated genomes of the global human gut microbiome. *Nature* 2019

肠道微生物在人类不同成长阶段的变化

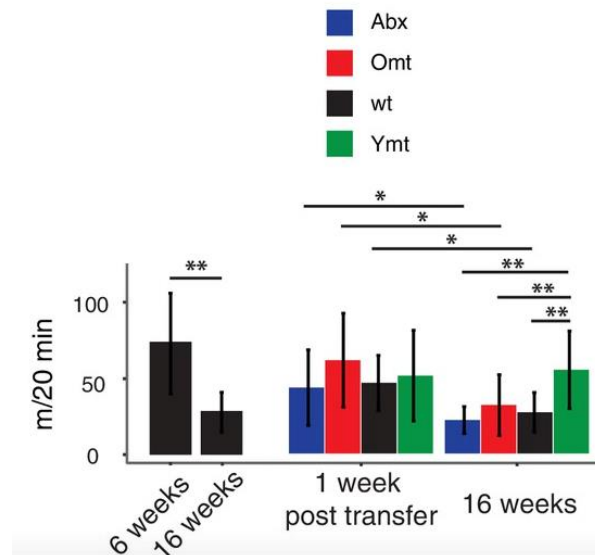
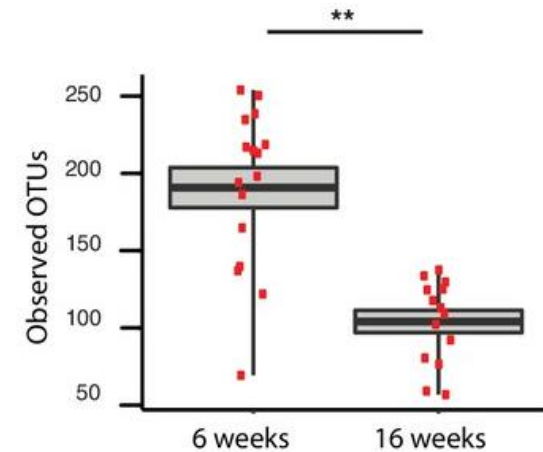
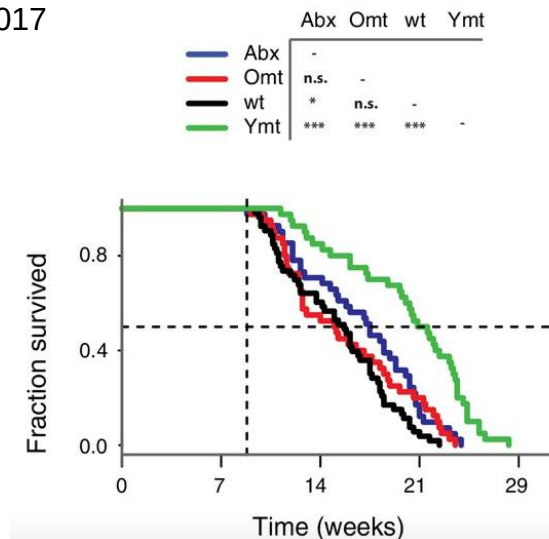
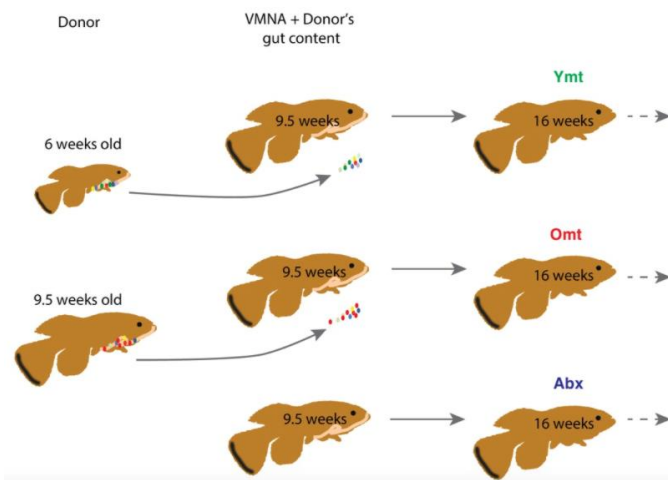


The function of our microbiota: who is out there and what do they do? *Front Cell Infect Microbiology* 2012

Metagenomics and Longevity



Regulation of life span by the gut microbiota in the short-lived African turquoise killifish. *eLife* 2017



粪菌移植FMT

粪菌移植（Fecal Microbiota Transplantation, FMT），是通过重建肠道菌群来治疗疾病的方法，即把经过处理的健康人的粪便液，灌到患者肠道内，操作过程需要通过严格的筛选和实验方案。

粪便银行：开放生物群（OpenBiome）是美国一家非营利性“粪便银行”，初建于2012年，由麻省理工学院博士后Mark Smith创立，向33个州的122家医院提供粪便样本和**粪菌胶囊**用于肠道菌群移植治疗。

<https://www.openbiome.org>

- 东晋时期(公元300 ~ 400年)的葛洪，在其所著《肘后备急方》中，记载了当时用粪清治疗食物中毒和严重腹泻，“绞粪汁，饮数合至一二升，谓之黄龙汤，陈久者佳”。
- 16世纪，中国明代李时珍在《本草纲目》中记载了口服粪水治疗严重腹泻、发热、呕吐和便秘等疾病。



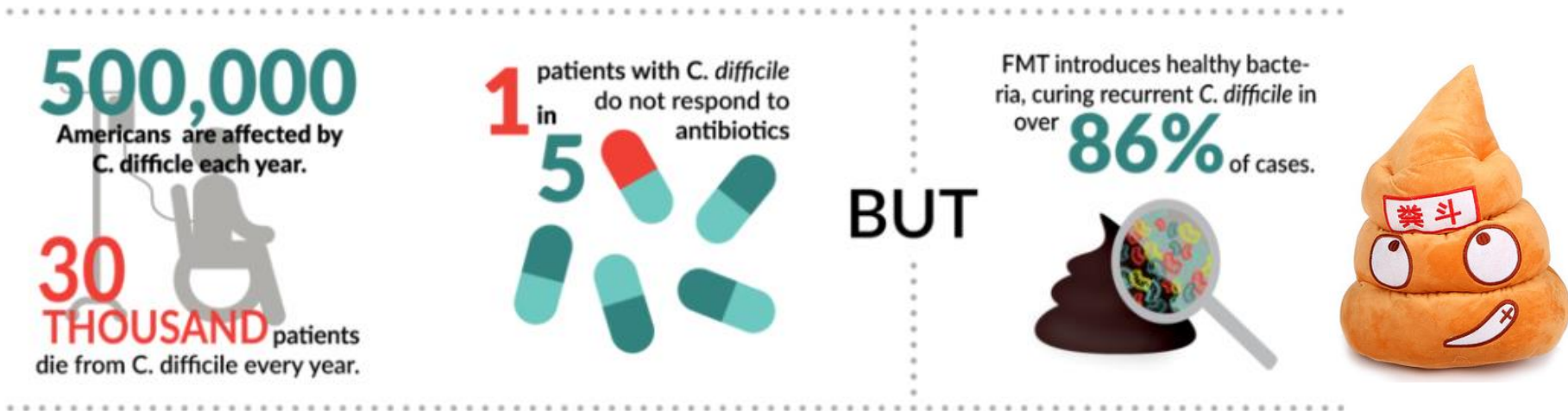
粪菌移植FMT案例

- 1958年，一种名叫伪膜性肠炎的消化道疾病死亡率高达75%。美国科罗拉多大学医学院外科医生beneiseman，对4名患有严重伪膜性肠炎的患者，采用了抗生素、氢化可的松、益生菌等治疗手段，但病人仍腹泻严重，甚至出现休克。无奈之下，医生最终和患者及家属商议，用患者家属的大便制成粪水对患者进行灌肠。结果，其中3名患者在几天之内奇迹般地康复出院，另一名患者在住院期间则死于与肠道感染无关的其他疾病。
- 2013年，美国一名32岁女子此前因被“艰难梭菌”感染，多年来经常呕吐及腹泻。于是她接受了一项手术，从自己体型肥胖的女儿胃部抽取粪便微生物，移植至自己的肠中，期望根治疾病。但该女子接受手术后虽然痊愈，体重却无故暴增，从手术前的62公斤长到了16个月后的77公斤，与自己的女儿一样，也变成了一个肥胖的人。

改善肠道菌群

粪菌胶囊：粪便库收集的粪便样本主要用于治疗一种特殊的肠道感染——艰难梭菌感染，导致肠道菌群失调。艰难梭菌感染难治、易复发，患病人群每年50万人，每年可导致**3万**美国患者死亡，移植健康人粪便中的肠道菌群是目前最理想的治疗方法。

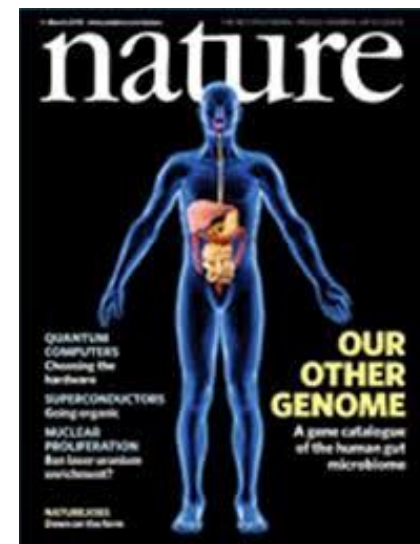
It was hard for doctors to treat patients with the most common hospital-acquired infection, *C. difficile*.



<https://www.openbiome.org/impact>

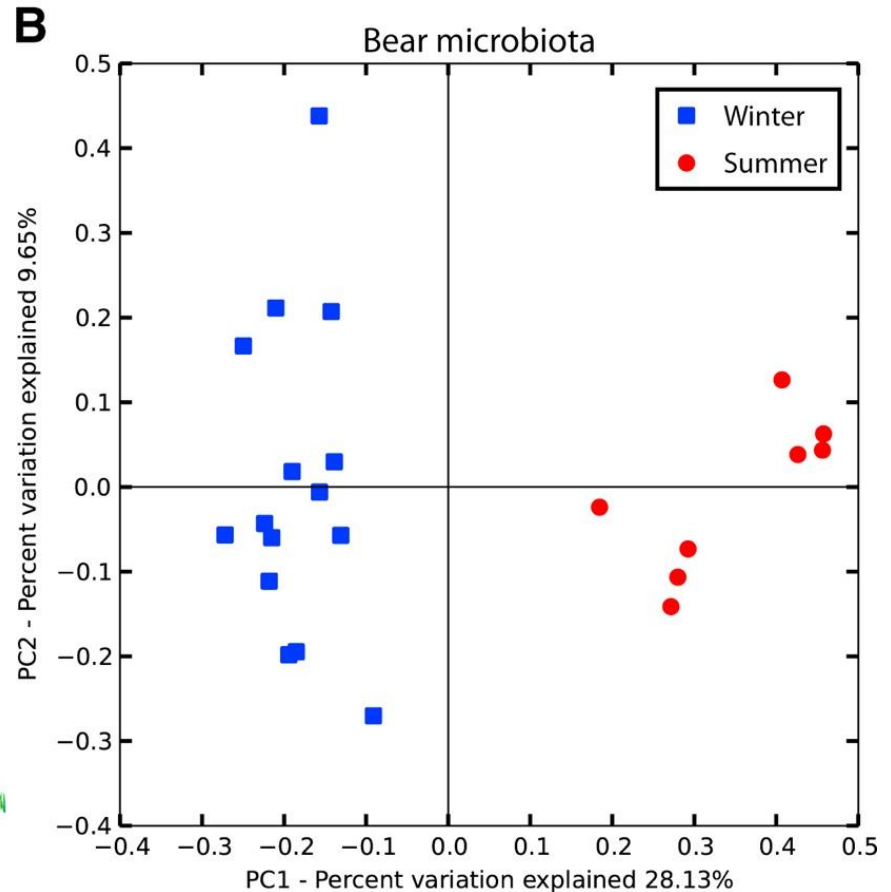
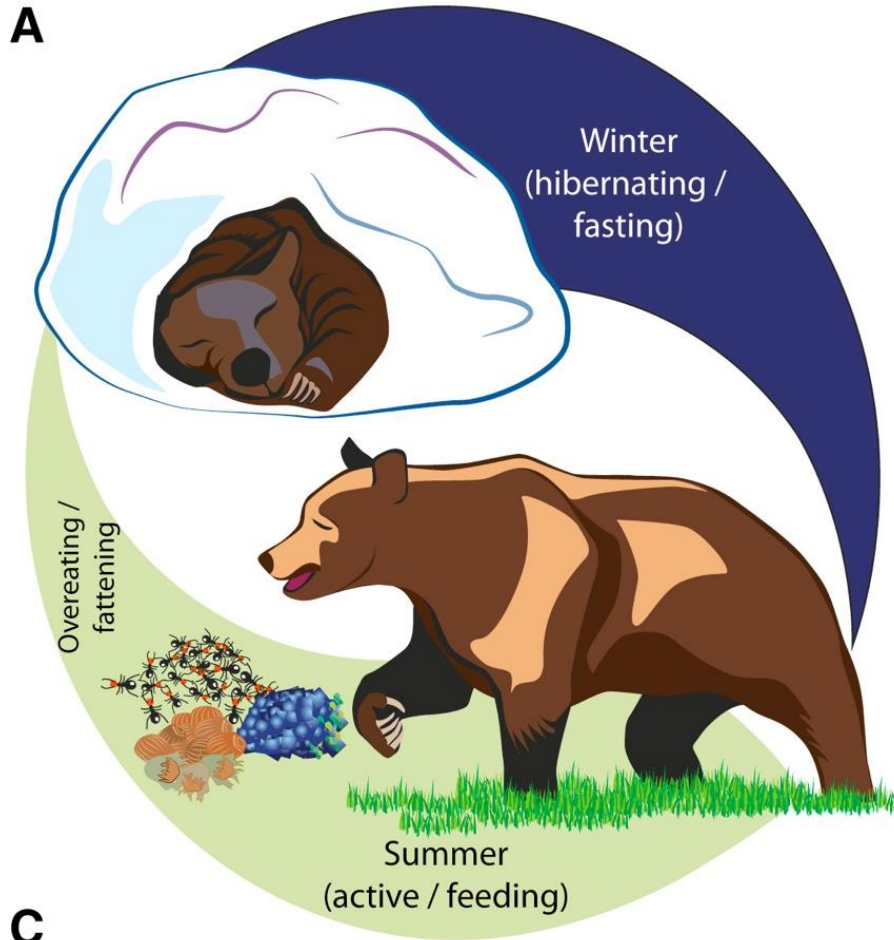
肠道宏基因组：人类第二基因组

肠道菌群数量庞大、种类繁多，参与机体许多重要的生理功能，包括食物的消化、新陈代谢、免疫反应和炎症，以及与大脑相互作用，影响情绪、食欲甚至生物节律



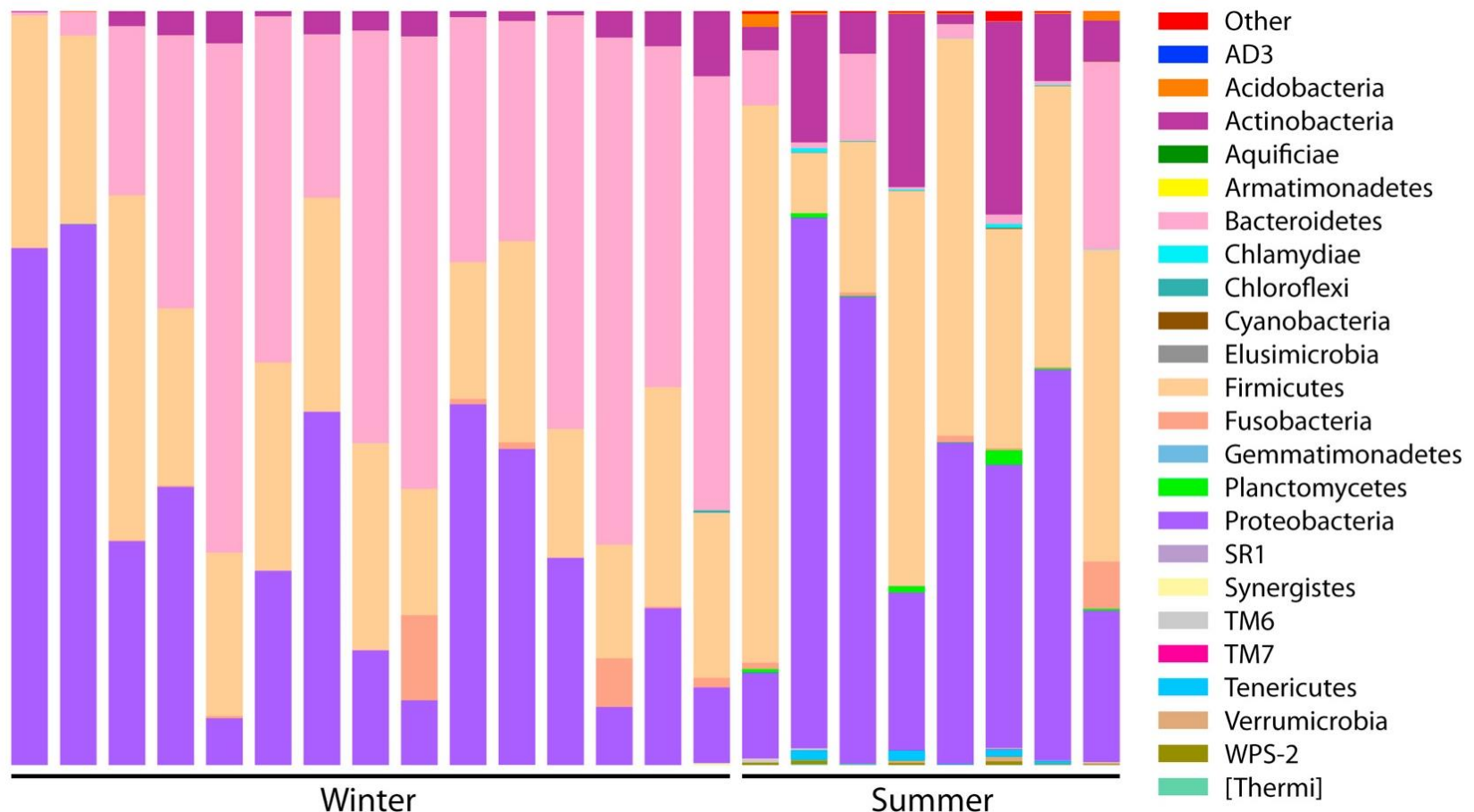
- 哮喘：四种关键肠道细菌的低水平将更容易使婴儿在1岁时表现出哮喘
- 肥胖：高脂饮食的导致啮齿类动物肠道菌群发生改变，从而在体内产生更多乙酸，同时也刺激胃泌素和胃饥饿素释放，进一步导致暴饮暴食
- 精神疾病：与移植来自健康个体粪便样品的无菌小鼠相比，被移植帕金森病患者肠道菌群样本的无菌小鼠表现出更强的症状

棕熊的肠道菌群：增肥、储存能量



The Gut Microbiota Modulates Energy Metabolism in the Hibernating Brown Bear *Ursus arctos*. *Cell Reports* 2016

棕熊夏冬季的肠道菌群变化

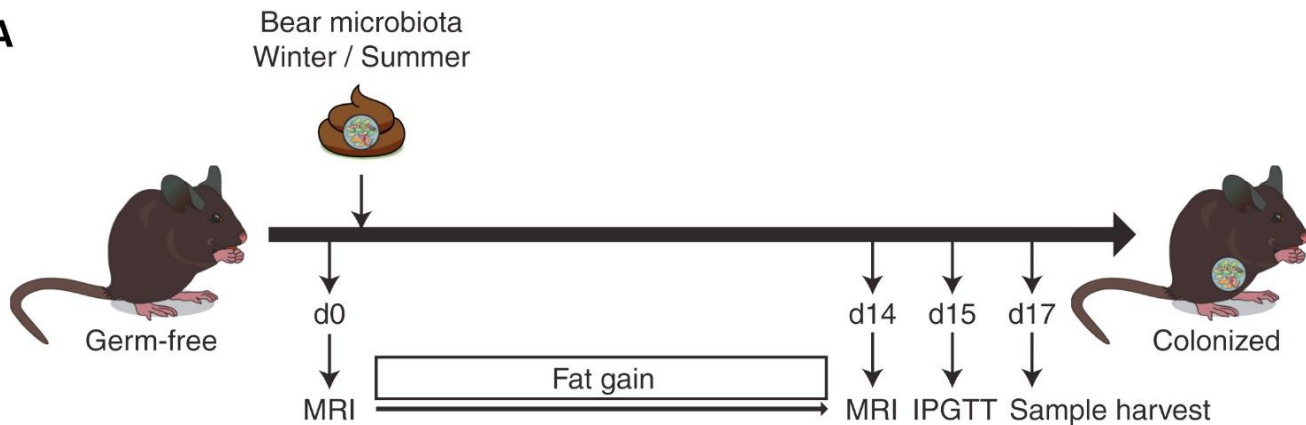


比较夏天活动期和冬眠期的肠道微生物样本，发现棕熊的肠道微生物组成和丰度均有明显变化

The Gut Microbiota Modulates Energy Metabolism in the Hibernating Brown Bear *Ursus arctos*. *Cell Reports* 2016

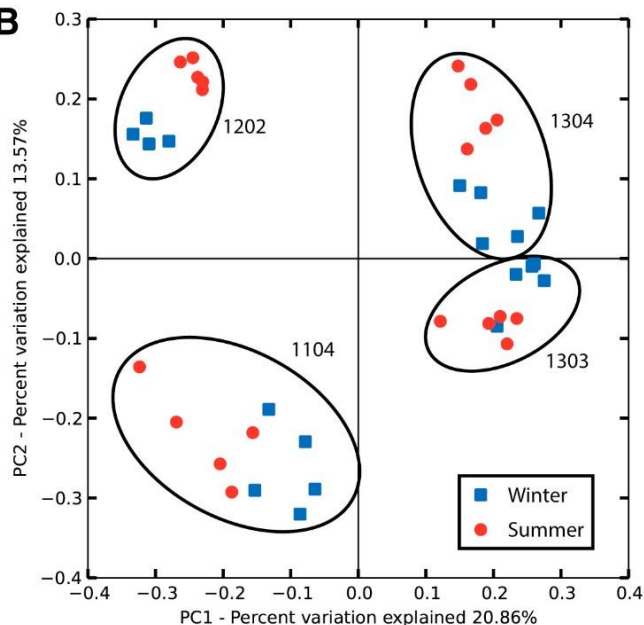
棕熊的肠道菌群：增肥并储存能量

A

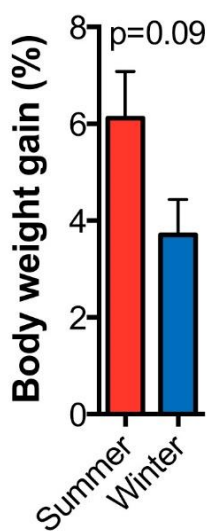


为验证棕熊肠道菌群对于其能量代谢的影响，研究人员分别将冬天和夏天的棕熊肠道微生物群移植给无菌小鼠。结果显示，夏天棕熊的肠道微生物群会促进脂肪的储存，且不会引起血糖和血脂的增加。

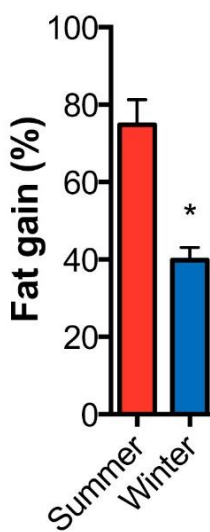
B



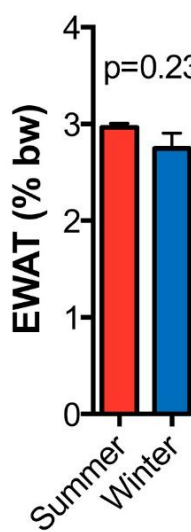
C



D



E



The Gut Microbiota Modulates Energy Metabolism in the Hibernating Brown Bear *Ursus arctos*. *Cell Reports* 2016

MGnify: microbiome analysis resource



Human
(129201)



Digestive
system
(85564)



Aquatic
(43822)



Marine
(32069)



Digestive
system
(25668)



Plants
(24135)



Soil
(20553)



Skin
(9564)



Wastewater
(3479)



Food
production
(2668)

The Earth Microbiome Project



[Read more](#)

The Earth Microbiome Project is now available in MGnify. The Earth Microbiome Project (EMP)¹ is a wide ranging collaborative effort that attempts to characterise the taxonomic and functional diversity of microbial life on the planet. Founded in 2010, the project includes 96 different studies, comprising approximately 26k individual sequencing runs from a diverse range of biomes.

[Project Home](#)

<https://www.ebi.ac.uk/metagenomics/>

<https://www.earthmicrobiome.org>

Summary

- 泛基因组
 - 泛基因组和泛基因组学
 - 核心基因组、非必需基因组、菌株特异性基因
 - 泛基因组：Open vs Closed, 水平基因转移HGT
 - 核心基因 vs 必需基因 (essential genes)
 - 泛基因组图 (pangenome/genome graph)
- 宏基因组
 - 宏基因组和宏基因组学的定义
 - 宏基因组、微生物组、微生物群
 - 宏基因组研究方法和分析流程, OTU
 - 人类第二基因组：肠道菌群、人类健康(粪菌移植)

Further Reading

- **A draft human pangenome reference.** *Nature* 2023
- **The Human Pangenome Project: a global resource to map genomic diversity.** *Nature* 2022
- **Pangenome graphs.** *Annu Rev Genomics Hum Genet.* 2020
- **Pan-genomics in the human genome era.** *Nature Reviews Genetics* 2020
- **Pan-Genome of Wild and Cultivated Soybeans.** *Cell* 2020
- **The Ecology and Evolution of Pangenomes.** *Current Biology* 2019
- **Computational pan-genomics: status, promises and challenges.** *Briefings in Bioinformatics* 2018
- **Why prokaryotes have pangenomes.** *Nature Microbiology* 2017
- **Clinical metagenomics.** *Nature Reviews Genetics* 2019
- **New insights from uncultivated genomes of the global human gut microbiome.** *Nature* 2019
- **The Integrative Human Microbiome Project.** *Nature* 2019
- **Shotgun metagenomics, from sampling to analysis.** *Nature Biotechnology* 2017