StrainPanDA: Linked reconstruction of strain composition and gene content profiles via pangenome-based decomposition of metagenomic data

Han Hu^{1,2}, Yuxiang Tan¹, Chenhao Li³, Junyu Chen¹, Yan Kou², Zhenjiang Zech Xu⁴, Yang-Yu Liu⁵, Yan Tan^{2,*}, Lei Dai^{1,*}

 ¹CAS Key Laboratory of Quantitative Engineering Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
 ²Xbiome, Scientific Research Building, Tsinghua High-Tech Park, Shenzhen, China
 ³Center for Computational and Integrative Biology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA
 ⁴State Key Laboratory of Food Science and Technology, Nanchang University, Nanchang, China
 ⁵Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA



Hu, Han, Yuxiang Tan, Chenhao Li, Junyu Chen, Yan Kou, Zhenjiang Zech Xu, Yang-Yu Liu, Yan Tan, and Lei Dai. 2022. "StrainPanDA: Linked reconstruction of strain composition and gene content profiles via pangenome-based decomposition of metagenomic data." *iMeta.* e41. <u>https://doi.org/10.1002/imt2.41</u>

Motivation

- Multiple within-species variants coexist in microbiomes, which can have substantial variations in their gene contents.
- Within-species variations can lead to substantial phenotypic differences, and play important role in microbial adaptation across environments and host-microbiome interaction.

Current approaches

- Most strain-level analysis tools focus on identifying strain composition based on single nucleotide variants (SNVs).
- Current pangenome-based tool such as PanPhlAn only infers the gene content of dominant strain in a metagenomics sample.

Solution

• A method to simultaneously reconstruct the composition and gene contents of coexisting strains from metagenomic data.

StrainPanDA (<u>Strain</u>-level <u>Pangenome</u> <u>D</u>ecomposition <u>A</u>nalysis)

Gene family abundance data matrix (<i>D</i>)								
	Sample1	Sample2	Sample3	SampleS				
GF 1	5	1	3	10				
GF 2	1	2	1	7				
GF 3	1	1	0	2				
GF N	5	1	1	1				

Gene content profile matrix (P)



Strain composition matrix (S)

		Sample1	Sample2	Sample3	SampleS
	Strain 1	0.1	0.2	0	0.1
	Strain 2	0.1	0.1	0.3	0.4
	Strain 3	0.3	0.3	0.3	0
	Strain K	0.2	0	0.1	0

Decomposition of the gene family abundance data matrix enables linked reconstruction of strain composition and gene content profile





StrainPanDA workflow



Study design

Validation: synthetic mixtures of multiple strains

- *E. coli* strains: sequencing errors, sequencing depths, and background metagenomes
- Other species: B. longum, C. difficile, E. faecalis, F. prausnitzii and P. copri.

Benchmarking: strain analysis tools

• StrainEst, PStrain, PanPhlAn

Application: longitudinal metagenomic datasets

- Infant gut microbiome (Bäckhed *et al.* 2015)
- Post-FMT gut microbiome (Kong *et al.* 2020)

StrainPanDA allows accurate inference of strain composition



JSD: Jensen-Shannon divergence

StrainPanDA allows accurate inference of gene content profile



Genome not used for constructing pangenome database

AUPRC: area under the Precision-Recall Curve

Case study #1: succession of *B. longum* subspecies in infant gut microbiome

А Mother Infant: newborn Infant: 12-month Infant: 4-month 3 subspecies3 subspecies3 subspecies3 subspecies3 в Breastfeeding
Discontinued
Continued **** 1.0 1.0 1.0 atimepoints c.0 0.5 0.5 differ ≥ 0.0 0.0 0.0 ap 2.0-the -0.5 -0.5 Relativ -1.0 -1.0 -1.0 subspecies1 subspecies2 subspecies3

B. longum subspecies



• Breastfeeding status change (discontinued or continued) was associated with the shift in *B. longum* subspecies.

43 kbp HMO cluster Sialidase cluster Fucosidase cluster LNT related gene GNB-LNB cluster

GluNAc-GalNAc related gene alpha mannose related gene urease related gene CRISPR related gene

Case study #1: succession of B. longum subspecies in infant gut microbiome

Gene content profiles reconstructed by StrainPanDA

class	subspecies1	subspecies2	subspecies3	<i>Red:</i> unique gene families of subspecies 1	Ure
		152 150 150 150 150 150 150 150 150 150 150		Family 1 extracellular solute-binding protein Galactoside symporter Binding-protein-dependent transport system Inner membrane protein Sugar ABC transporter permease Alpha-galactosidase Dihydrodipicolinate synthase Beta-galactosidase Oligopeptide-binding protein oppA Oligopeptide transport system permease protein tusion between oligopeptide transport permease oppC and ATP-binding protein oppD Lacl-type transcriptional regulator inner-membrane translocator N-Acetylmannosamine-6-phosphate 2-epimerase,NanE Sialidase A ROK family transcriptional regulator formate C-acetyltransferase L-fucose permease lactaldehyde reductase L-fuconolactonase Alpha-L-fucosidase Mandelate racemase/muconate lactonizing domain-containing protein LNT beta-1,3-galactosidase Solute-binding protein of ABC transporter system for sugars Permease protein of ABC transporter system for sugars Lacto-N-biose phorylase Phosphocarrier protein HPr Phosphoenolpyruvate-protein phosphotransferase Endo-alpha-N-acetylgalactosaminidase N-Acetylglucosamine-6-phosphate deacetylase PTS system, N-acetylglaccosamine-specific IIBC component PTS system, N-acetylglaccosamine-specific IIBC component Alpha-mannosidase high-affinity nickel-transporter urea inner-membrane translocator Urease accessory protein type I-E type I-C	class 43 Sia Fu UN GN GN GN
				Beta−lactamase N−Acetylneuraminate lyase	

- Distinct nutrient utilization genes among the subspecies
- Subspecies 1 had unique gene families (marked in red) that are key enzymes related to human milk oligosaccharides (HMOs)

Case study #2: Crohn's disease patients treated with FMT



B. ovatus subspecies

- Strain composition was individualized;
- Two subspecies had opposite correlation trends with the species and distinct enrichment patterns with FMT outcome;
- Subspecies 2 had more CAZy gene families and strain-specific virulence factor genes, which may contribute to its competitive advantage and association with post-FMT relapse.

Summary

- StrainPanDA is a novel method that reconstructs the strain composition and gene contents with high accuracy and robustness, compared to state-of-the-art methods.
- Linked reconstruction of strain composition and gene contents is crucial for understanding the relationship between microbial adaptation and strain-specific function.
- StrainPanDA is accessible from https://github.com/xbiome/StrainPanDA



Strain composition



Strain gene content profile

Hu, Han, Yuxiang Tan, Chenhao Li, Junyu Chen, Yan Kou, Zhenjiang Zech Xu, Yang-Yu Liu, Yan Tan, and Lei Dai. 2022. "StrainPanDA: Linked reconstruction of strain composition and gene content profiles via pangenome-based decomposition of metagenomic data." *iMeta.* e41. <u>https://doi.org/10.1002/imt2.41</u>

iMeta: Integrated meta-omics to change the understanding of the biology and environment





"*iMeta*" is an open-access Wiley partner journal and launched by scientists of the Chinese Academy of Sciences. iMeta aims to promote metagenomics, microbiome and bioinformatics development by publishing original researches, methods or protocols, and reviews. The goal is to publish highly quality papers (Top 10%, IF > 15) targeting broad audience. Unique features including video submission, reproducible analysis, figure polishing, APC waiver, and promotion by social media with 500,000 followers. The first issue released in March 2022.



Society: <u>http://www.imeta.science</u>

Publisher: https://onlinelibrary.wiley.com/journal/2770596x



Submission: https://mc.manuscriptcentral.com/imeta

