PennCHOP
MICROBIOME PROGRAM

Introduction to Shotgun metagenomics analysis using Sunbeam pipeline

Ceylan Tanes, PhD

Principal Bioinformatics Scientist

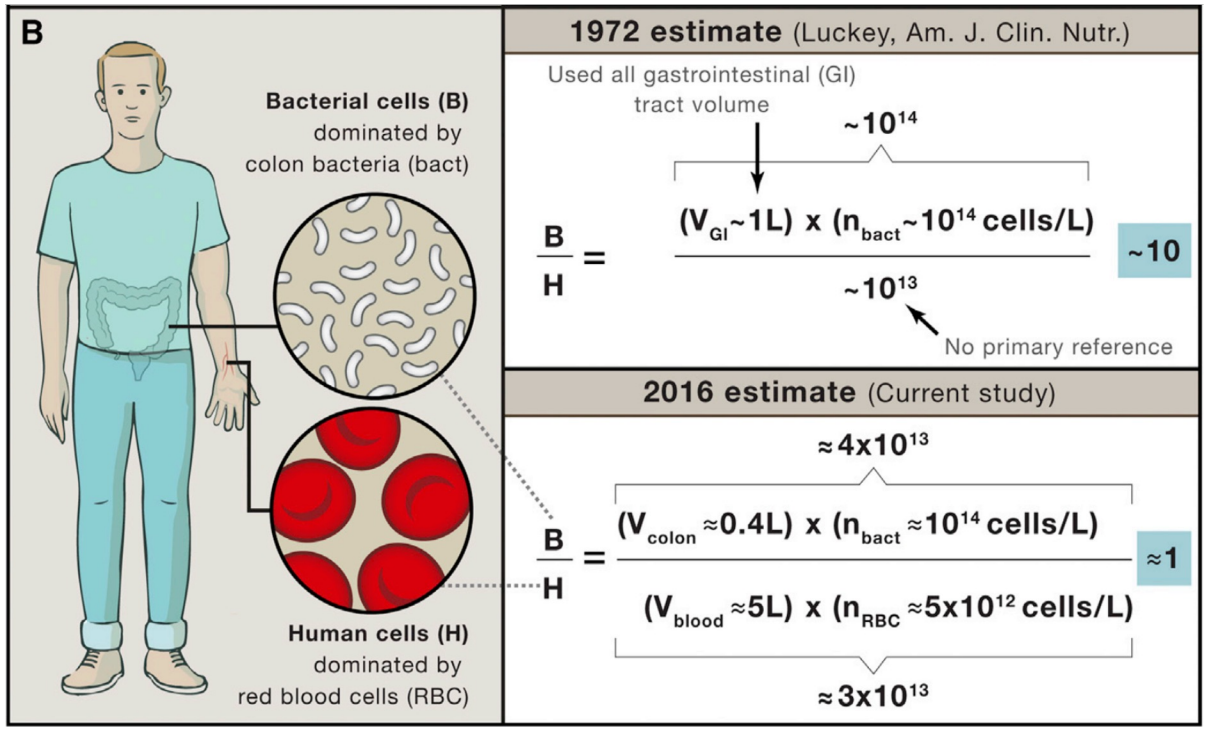
CHOP Microbiome Program

Division of Gastroenterology, Hepatology and Nutrition

Children's Hospital of Philadelphia

Why should we study the microbiome?

- There are as many bacteria cells in our body as human cells
- Bacteria metabolize the foods we consume
- Bacteria interact with the host immune system



Sender R. *Cell* **164**, 337 (2016).

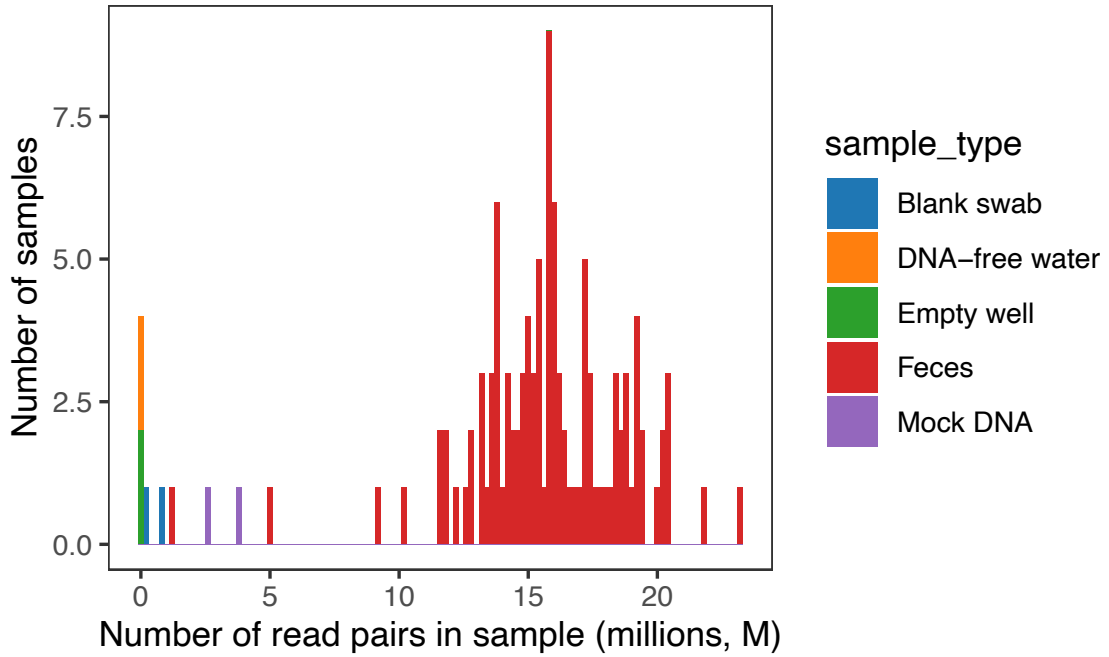
Methods to study the microbiome

- **Marker gene sequencing**
  - **16S rRNA gene**
  - Internal transcribed spacer (ITS)
- **Shotgun sequencing**
- Metabolome - small molecules
- Transcriptome – RNA
- Proteome - proteins

Shotgun sequencing overview

- Extract DNA
- Fragment the DNA randomly
- Amplify
- Size selection (Illumina sequences short reads)
- Pool and sequence

- Number of reads you need per sample: Order of millions



Is Shotgun sequencing right for your project?

Pros

- No primer bias – get information on viruses and eukaryotes as well
- With enough sequencing depth (and processing power), sky is the limit on what kind of analyses you can do
- Information on gene content of bacteria
- Strain level analysis

Cons

- More complicated to analyze
- Host reads are also sequenced so not a good choice for low biomass samples
- You are limited by how complete your reference genomes are

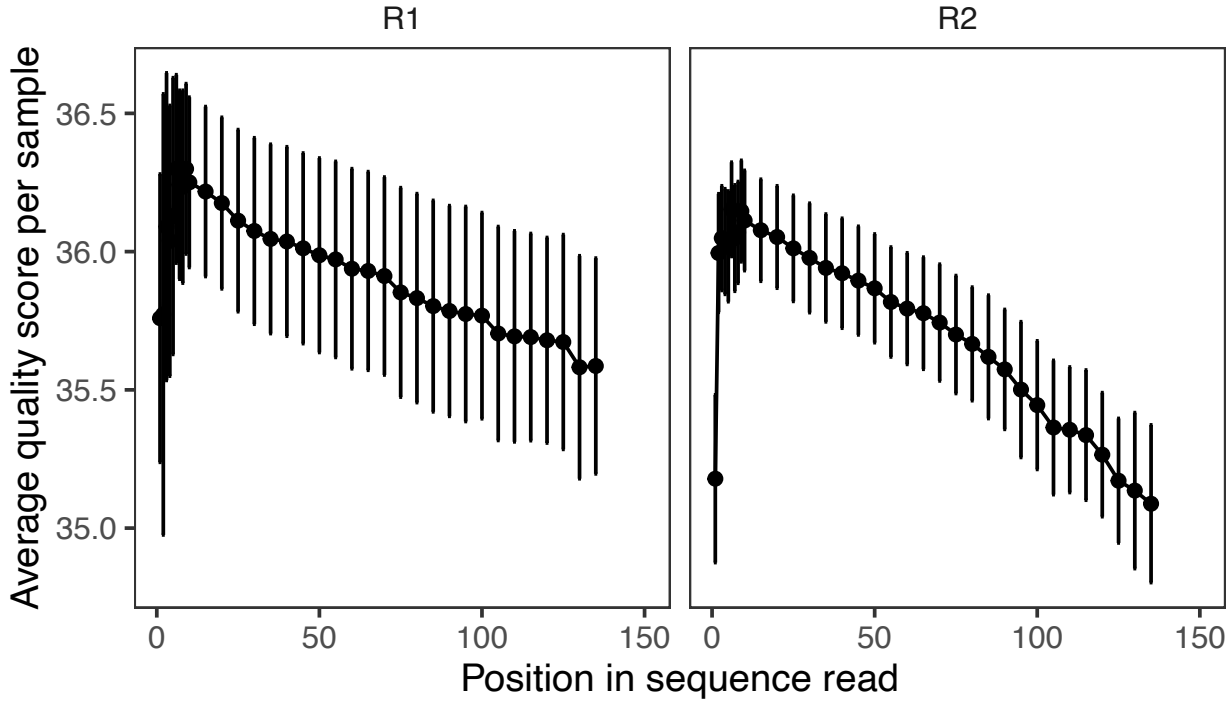
Steps to analyze Shotgun data

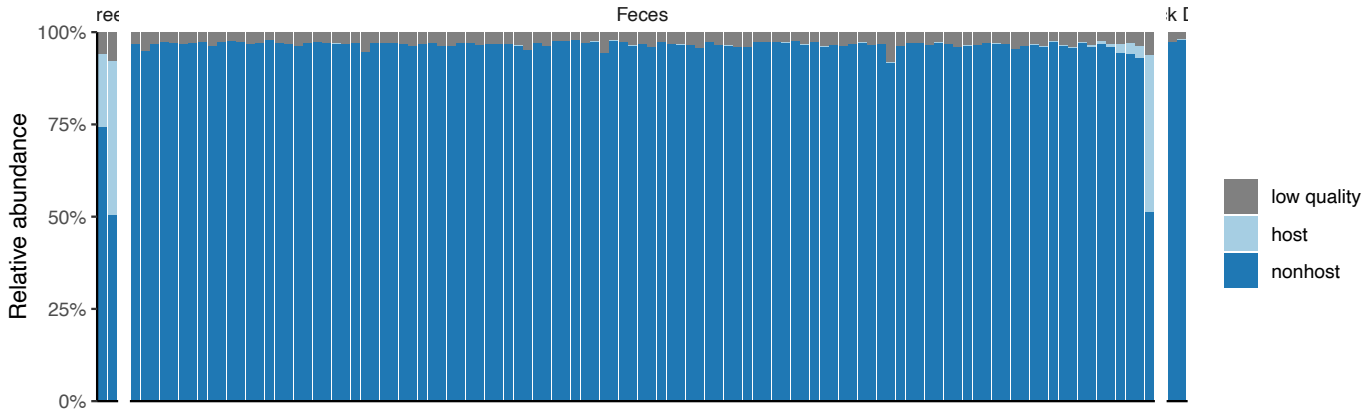
- Quality control
  - Remove/trim low quality reads
  - Remove host reads
- Assign taxonomy to reads
  - Kraken, Metaphlan
- Build contigs / assemble genomes
- Align to protein databases

Quality control

- Trim adapters
- Check sequence quality
- Remove low complexity reads

- Align sequences to genomes and remove:
  - Host
  - phiX (spiked in during library prep)





Metaphlan

- Database of clade specific markers
- Small database so it's fast and not memory intensive
  - Uses only ~4% of sequenced microbial genes
- It requires a certain amount of the clade specific marker to make a call.
- Normalize the total number of reads in each clade by nucleotide length

Metagenomic microbial community profiling using unique clade-specific marker genes

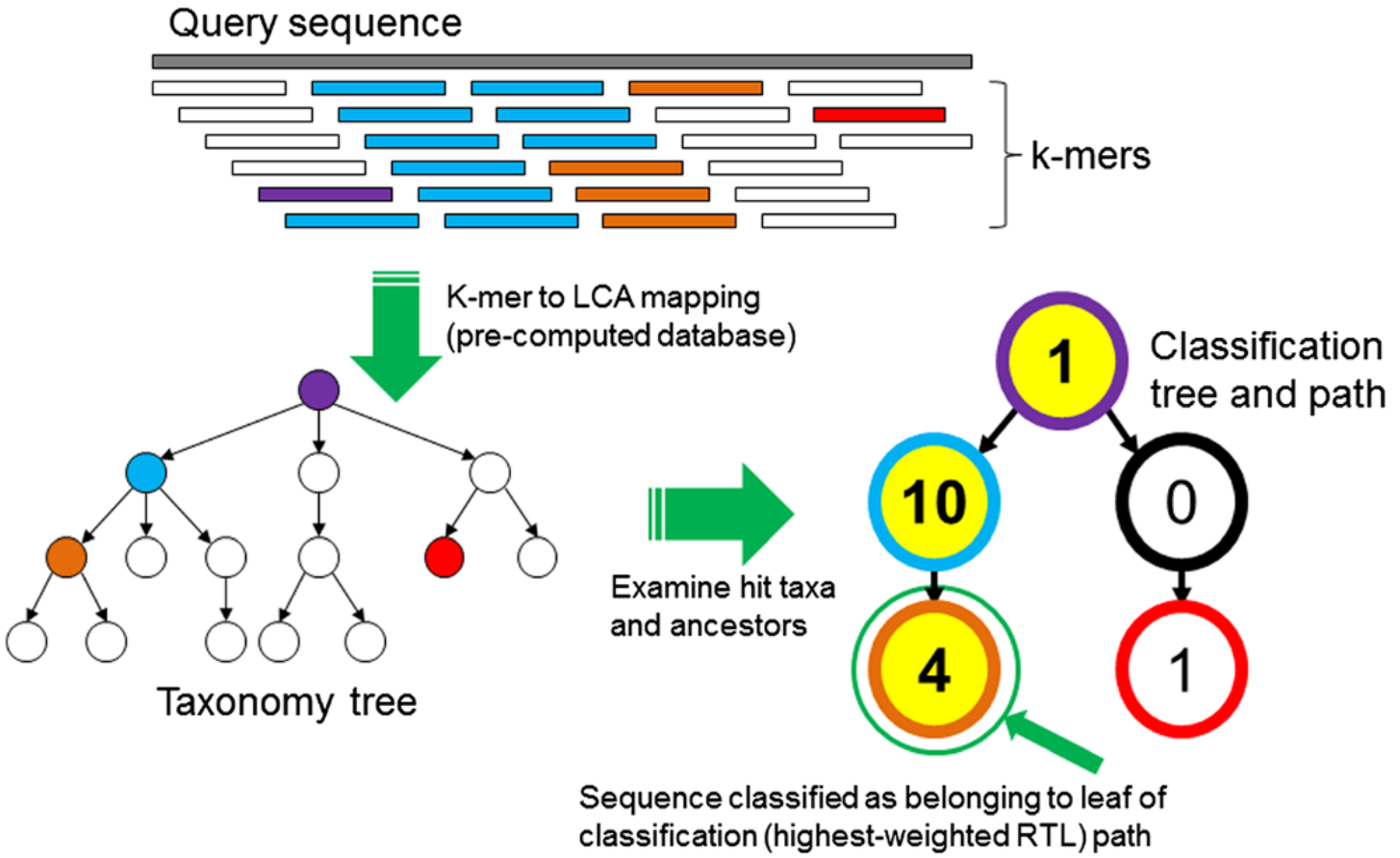
Nicola Segata[1], Levi Waldron[1], Annalisa Ballarini[2], Vagheesh Narasimhan[1], Olivier Jousson[2] & Curtis Huttenhower[1]

Kraken

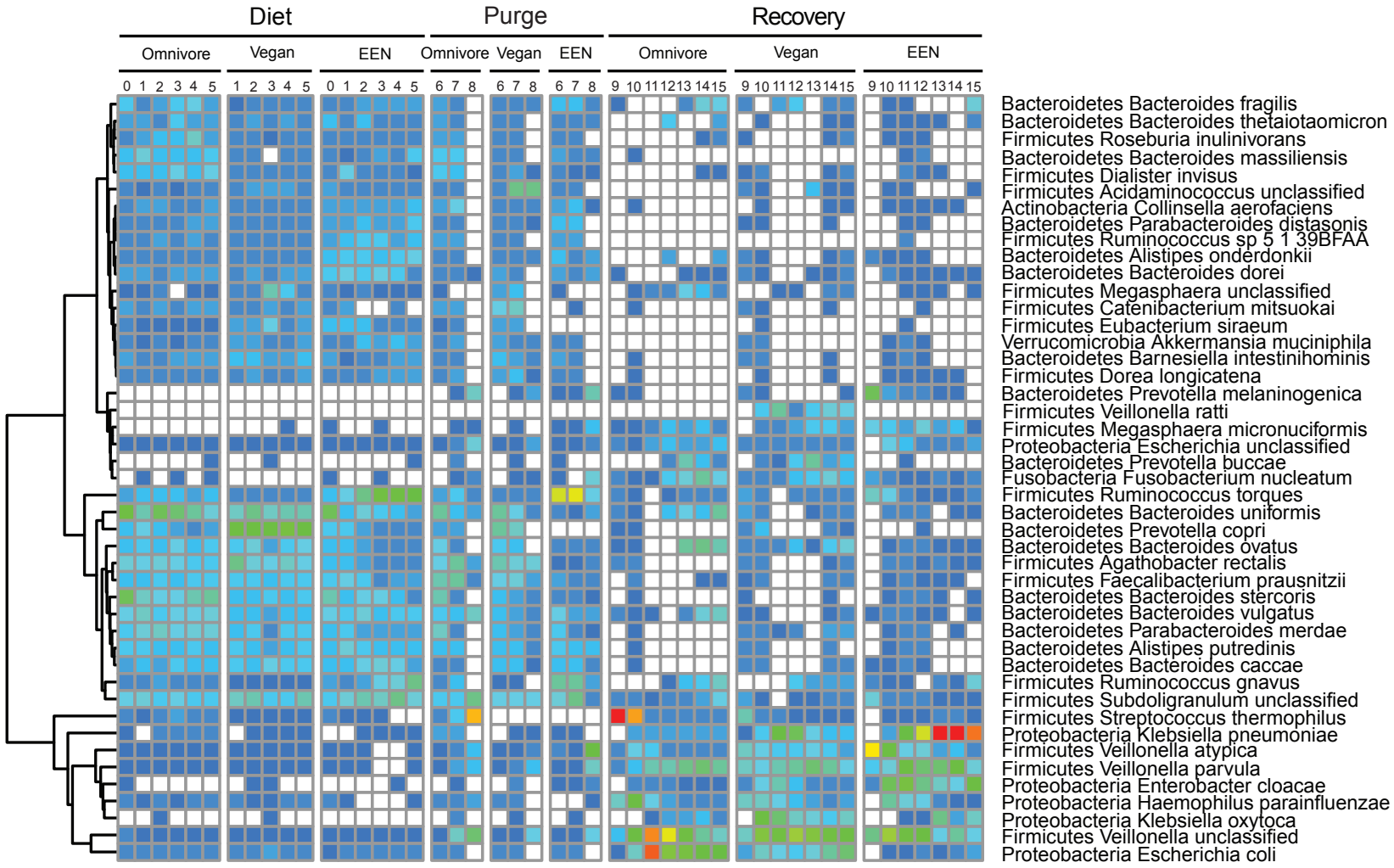
- Build a database of k-mers and least common ancestors of all organisms whose genomes contain that k-mer.
- Default k=31

Kraken: ultrafast metagenomic sequence classification using exact alignments

Derrick E Wood[1,2*] and Steven L Salzberg[2,3]



How to visualize the taxonomic assignment results

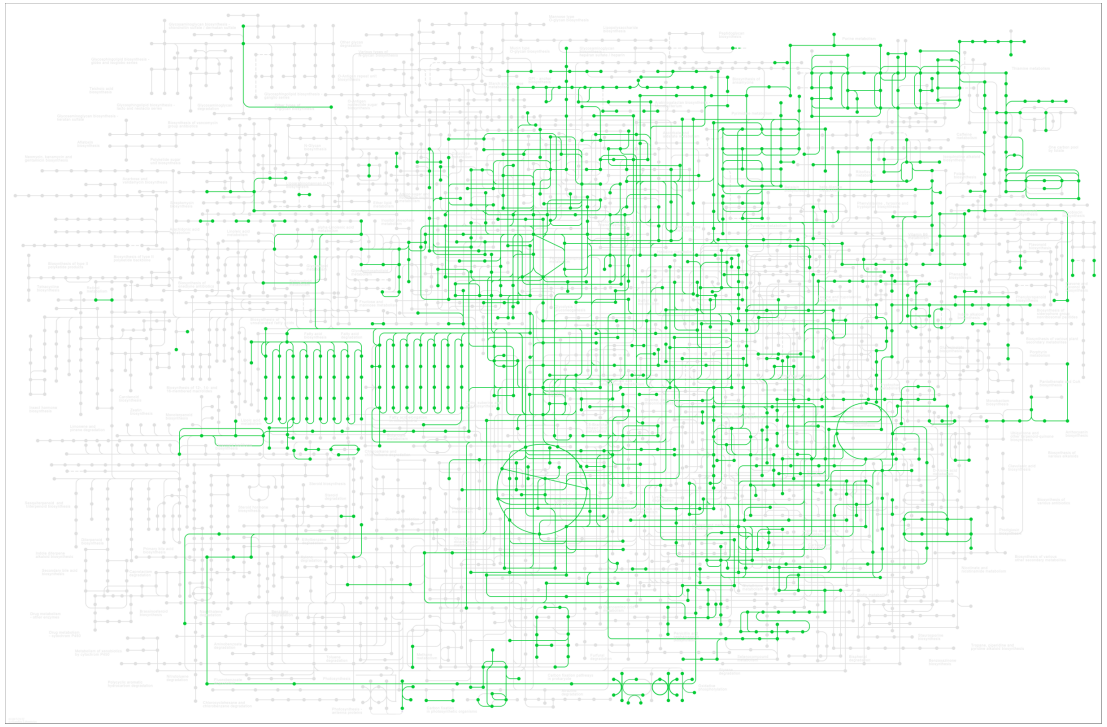


Tanes et al.2021 https://doi.org/10.1016/j.chom.2020.12.012

Considering the microbiome as a gene pool

- Bacteria has a unique set of genes that interact with the metabolites available to them in their environment.

- Some examples include but are not limited to:
  - Production of butyrate through amino acids or fermentation of complex carbohydrates
  - Broader set of glycoside hydrolases to ferment the fiber that reaches the large intestine
  - Production of indole-propionic acid from tryptophan
  - Modifying bile acids
  - Production of TMAO

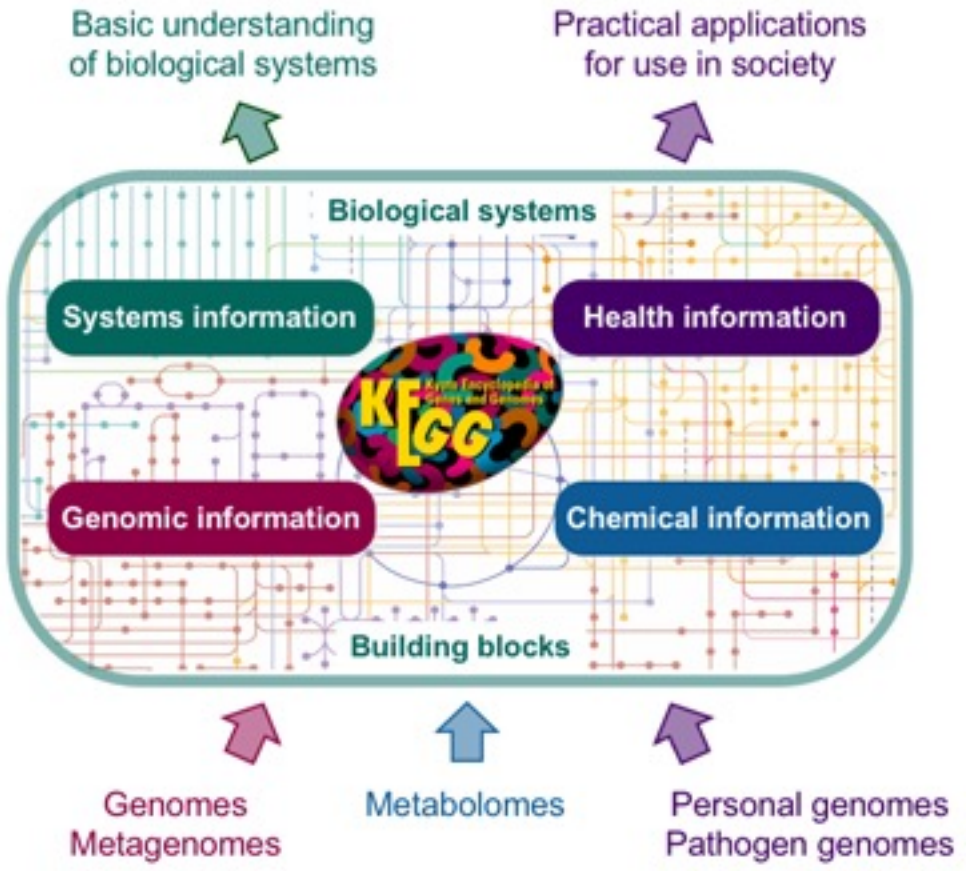


Metabolic pathway for E. coli according to KEGG database

Kyoto Encyclopedia of Genes and Genomes is a great resource

KEGG Database as of 2022/6/9

Systems information		
KEGG PATHWAY	Pathway maps, reference (total)	551 (931,487)
KEGG BRITE	Functional hierarchies, reference (total)	185 (316,073)
KEGG MODULE	KEGG modules	457
	Reaction modules	46
Genomic information		
KEGG ORTHOLOGY	KEGG Orthology (KO) groups	25,198
KEGG GENES	Genes in KEGG organisms	41,397,244
	Viral genes	595,312
	Viral mature peptides	256
	Addendum proteins	4,106
KEGG GENOME	KEGG organisms	8,137
	(753 eukaryotes, 6995 bacteria, 389 archaea)	
	KEGG selected viruses	359
Chemical information		
KEGG COMPOUND	Metabolites and other chemical substances	18,918
KEGG GLYCAN	Glycans	11,084
KEGG REACTION	Biochemical reactions	11,774
	Reaction class	3,177
KEGG ENZYME	Enzyme nomenclature	7,962
Health information		
KEGG NETWORK	Disease-related network elements	1,233
	Network variation maps	133
KEGG VARIANT	Human gene variants	456
KEGG DISEASE	Human diseases	2,563
KEGG DRUG	Drugs	11,910
	Drug groups	2,392
Drug labels		
KEGG MEDICUS	Japanese prescription drug labels from JAPIC	13,947
	Japanese OTC drug labels from JAPIC	10,692
KEGG MEDICUS	FDA prescription drug labels linked to DailyMed	33,522



https://www.kegg.jp/kegg/kegg1a.html

Cell Host & Microbe

CellPress

Article

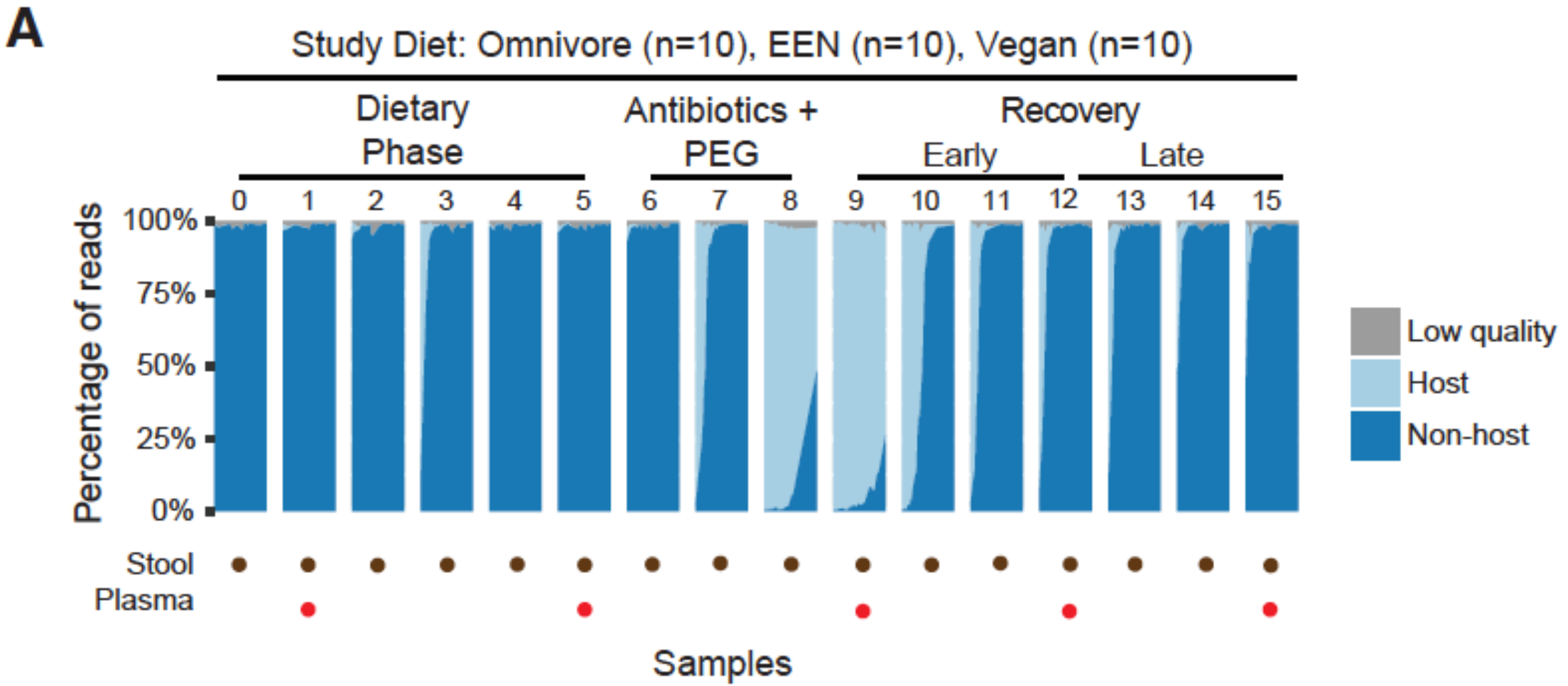
Role of dietary fiber in the recovery of the human gut microbiome and its metabolome

Ceylan Tanes,[1] Kyle Bittinger,[1] Yuan Gao,[2] Elliot S. Friedman,[3] Lisa Nessel,[2] Unmesha Roy Paladhi,[2] Lillian Chau,[3] Erika Panfen,[3] Michael A. Fischbach,[4] Jonathan Braun,[5] Ramnik J. Xavier,[6] Clary B. Clish,[7] Hongzhe Li,[2] Frederic D. Bushman,[8] James D. Lewis,[2,3,9,*] and Gary D. Wu[3,9,10,*]

- Microbiome utilizes the nutrients we consume that reach the large intestine
- What is the effect of diet on microbiome and metabolome, before and after a gut purge
- 3 divergent diets
  - Omnivore
  - Vegan
  - Enteral nutrition diet

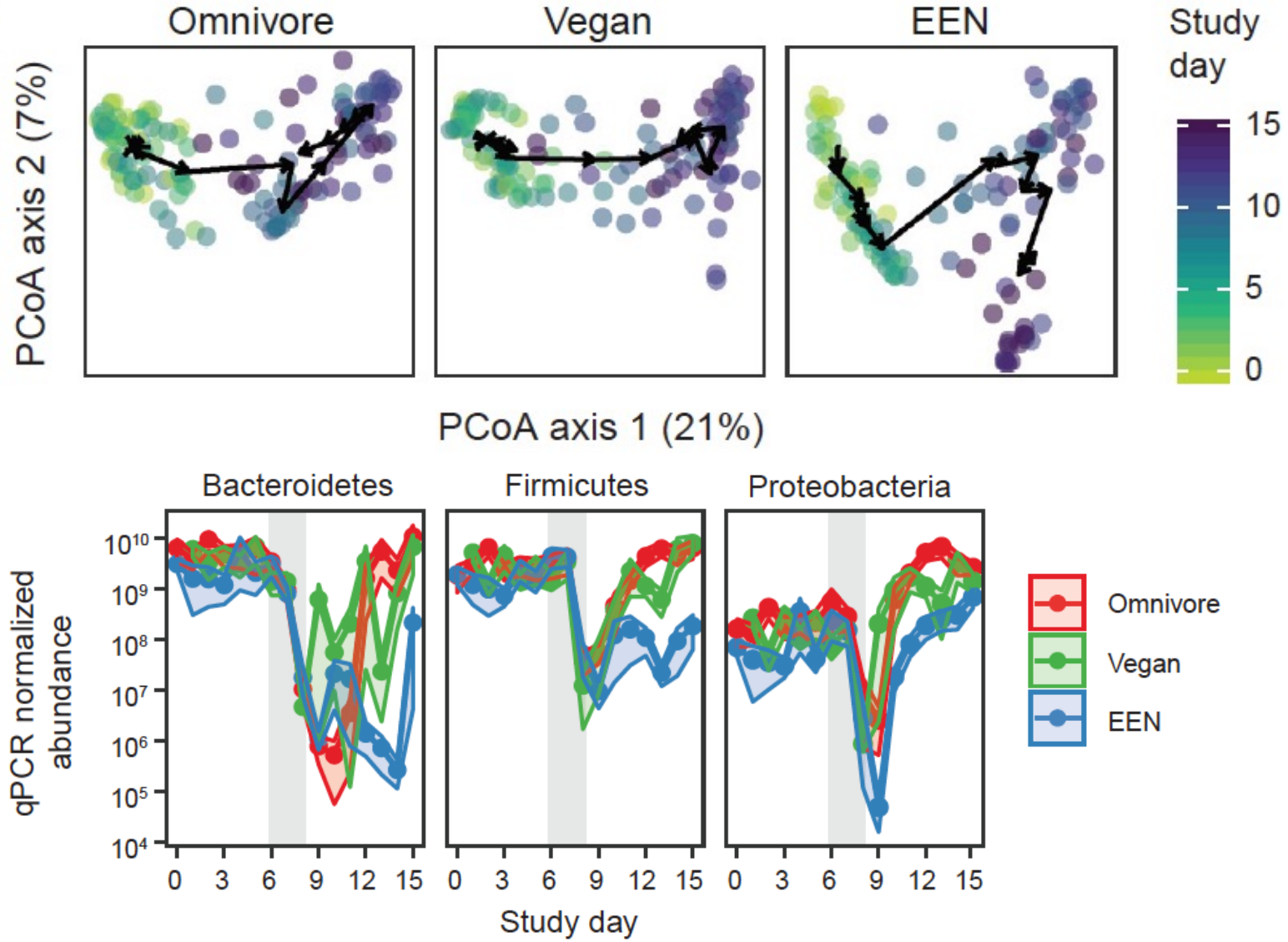
Tanes et al.2021 https://doi.org/10.1016/j.chom.2020.12.012

Study design



Tanes et al.2021 https://doi.org/10.1016/j.chom.2020.12.012

Community doesn't recover in EEN as well as other diets



Tanes et al.2021 https://doi.org/10.1016/j.chom.2020.12.012

Bacterial degradation of complex carbohydrates

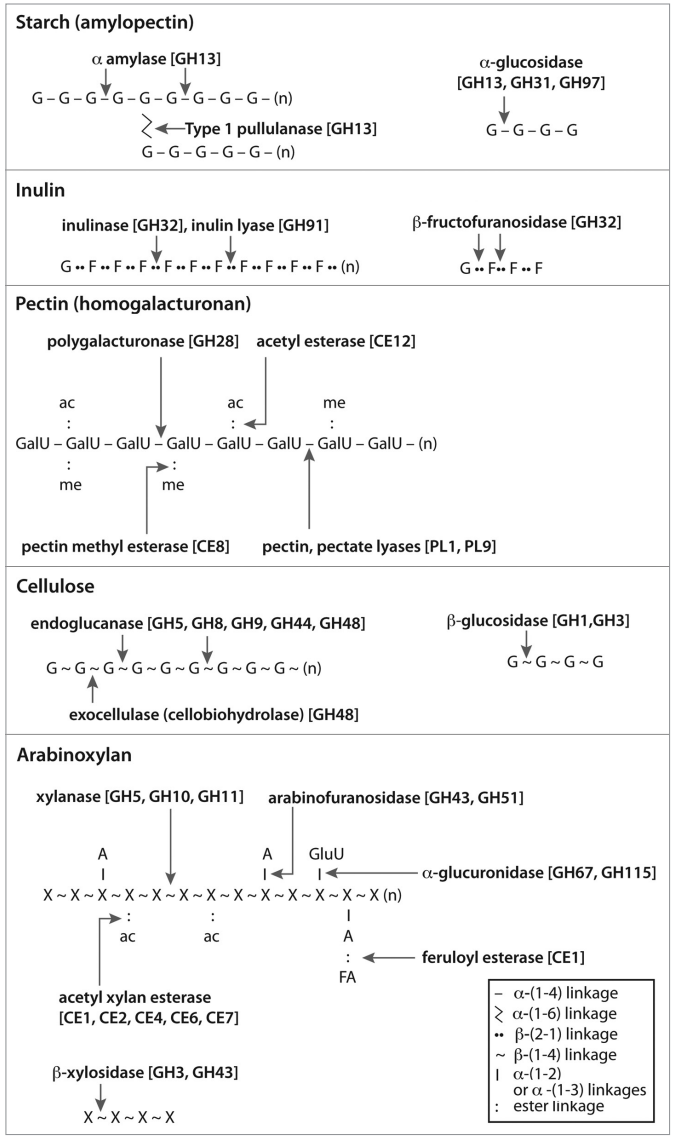


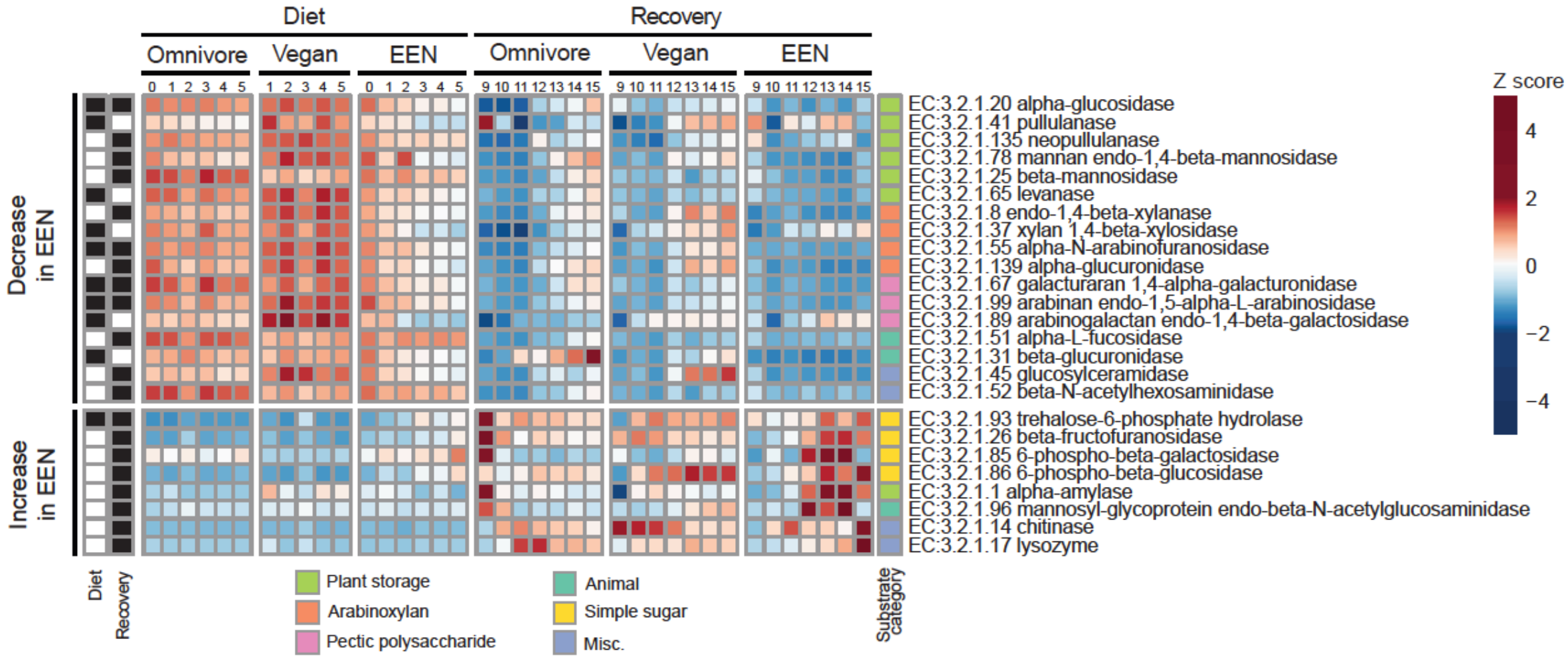
Table 1. Predicted CAZymes encoded by the genomes of selected fibrolytic gut bacteria

Ecosystem	Phylum (family)	Bacterium	Total CAZymes	GH	GT	PL	CE	Total CBMs
Human colon	Bacteroidetes	*Bacteroides thetaiotaomicron* VPI-5482	386	263	87	16	20	31
		B. xylanisolvens XB1A*	349	224	81	22	22	26
		B. vulgatus ATCC-8482	279	177	78	7	17	18
		B. fragilis 638R	223	138	78	1	6	26
	Firmicutes: (Lachnospiraceae) (Ruminococcaceae)	*Roseburia intestinalis* XB6B4*	175	115	46	0	14	11
		Butyrivibrio fibrisolvens 16/4*	115	75	37	0	3	31
		Ruminococcus champanellensis 18P13*	87	54	12	9	12	34
	Actinobacteria	*Bifidobacterium adolescentis* ATCC15703	94	54	37	0	3	6
Rumen	Fibrobacteres/ Acidobacteria	*Fibrobacter succinogenes* S85	183	100	54	12	17	73
	Bacteroidetes	*Prevotella ruminicola* 23	215	133	60	3	19	16
		P. bryantii	203	107	53	14	19	un
	Firmicutes: (Ruminococcaceae)	*Ruminococcus albus* 7	145	96	24	7	18	128
		R. flavefaciens FD1	140+	101	un	13	26	68

(GH, glycoside hydrolases; GT, glycosyl transferases; PL, polysaccharide lyases; CE, carbohydrate esterases; CBM, carbohydrate binding modules)
*For these strains, data were provided by the Pathogen Genomics group at the Wellcome Trust Sanger Institute and can be obtained from http://www.sanger.ac.uk/resources/downloads/bacteria/metahit/; The information presented is available from the CAZY website, except in the case of *R. flavefaciens* FD1[27,30] and *P. bryantii*.[44] Un, information not available.

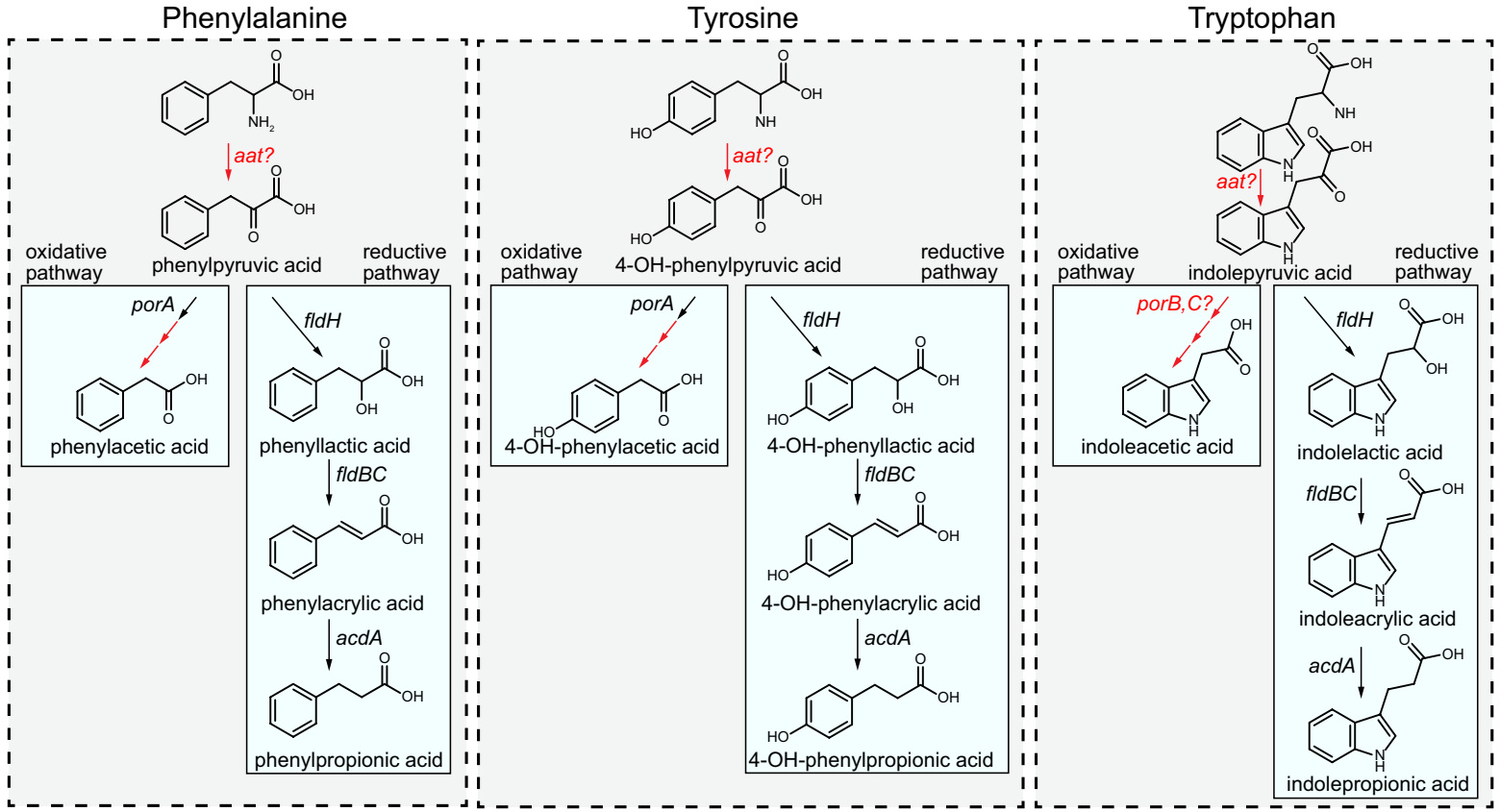
Flint et al. Gut Microbes. Jul-Aug 2012;3(4):289-306. doi: 10.4161/gmic.19897.

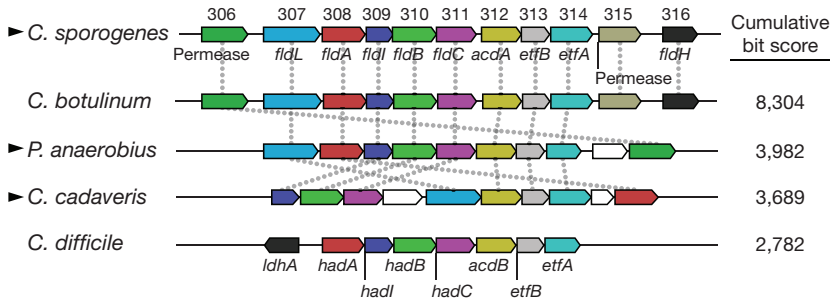
Low levels of glycoside hydrolases that break down complex carbohydrates in the EEN group



Tanes et al.2021 https://doi.org/10.1016/j.chom.2020.12.012

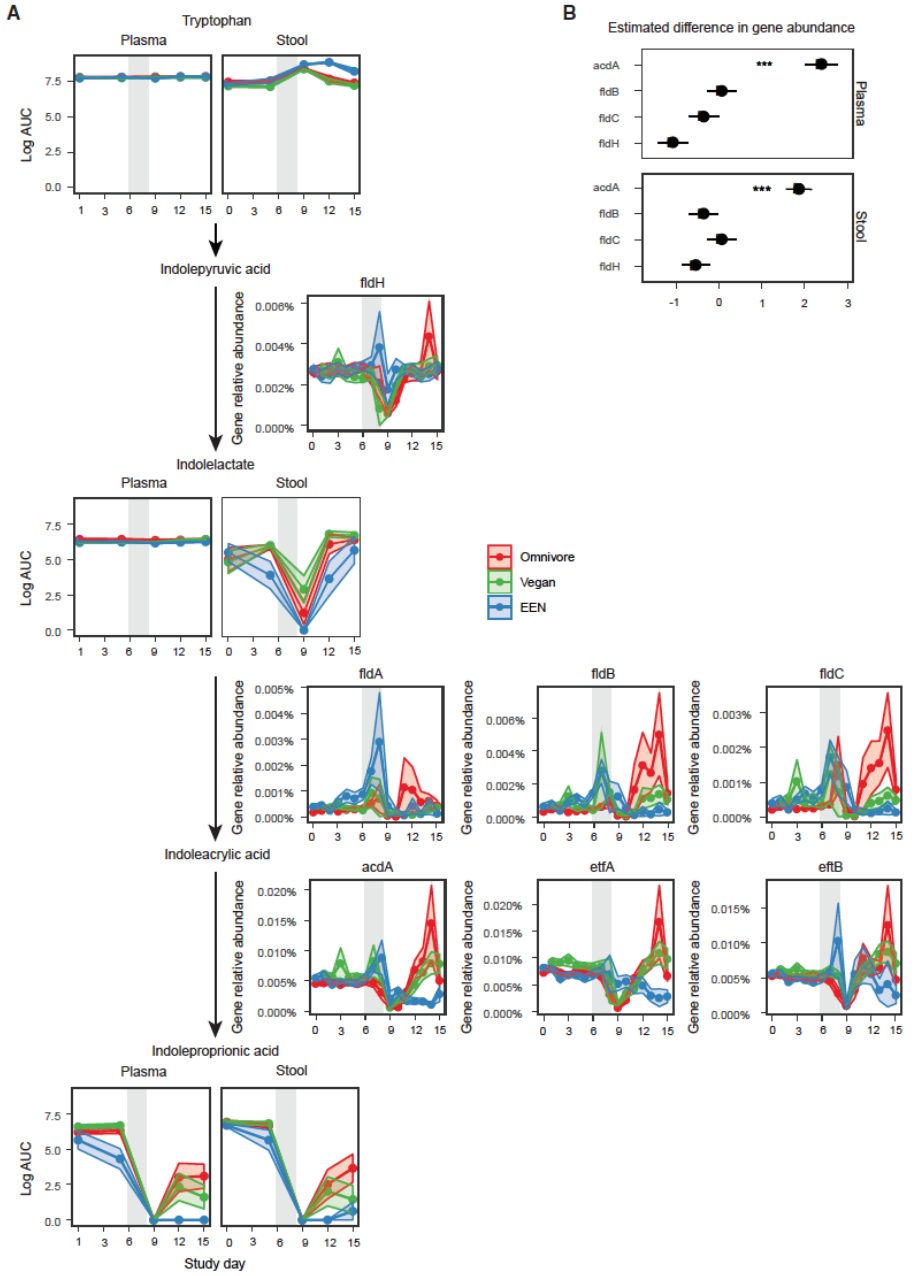
Bacterial metabolites – Indole propionic acid





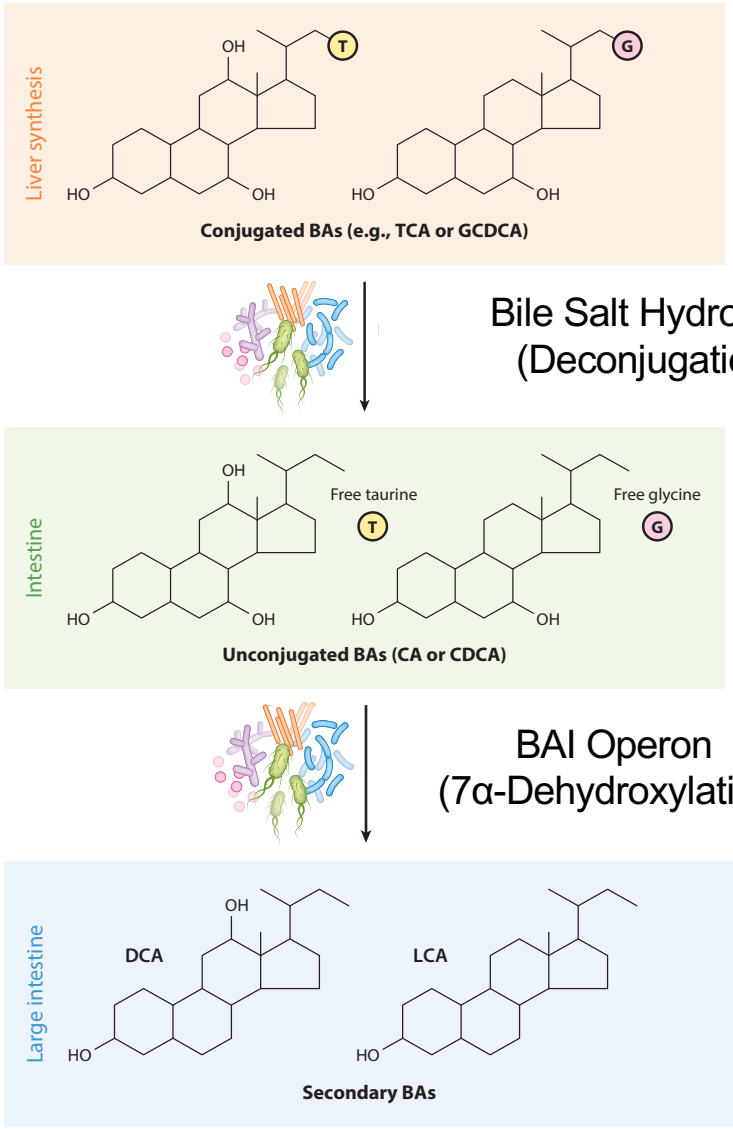
Dodd et al. *Nature* **volume 551**, pages 648–652 (2017)

Indole propionic acid levels don't recover in the EEN group



Tanes et al.2021 https://doi.org/10.1016/j.chom.2020.12.012

Bile acid modification of bacteria



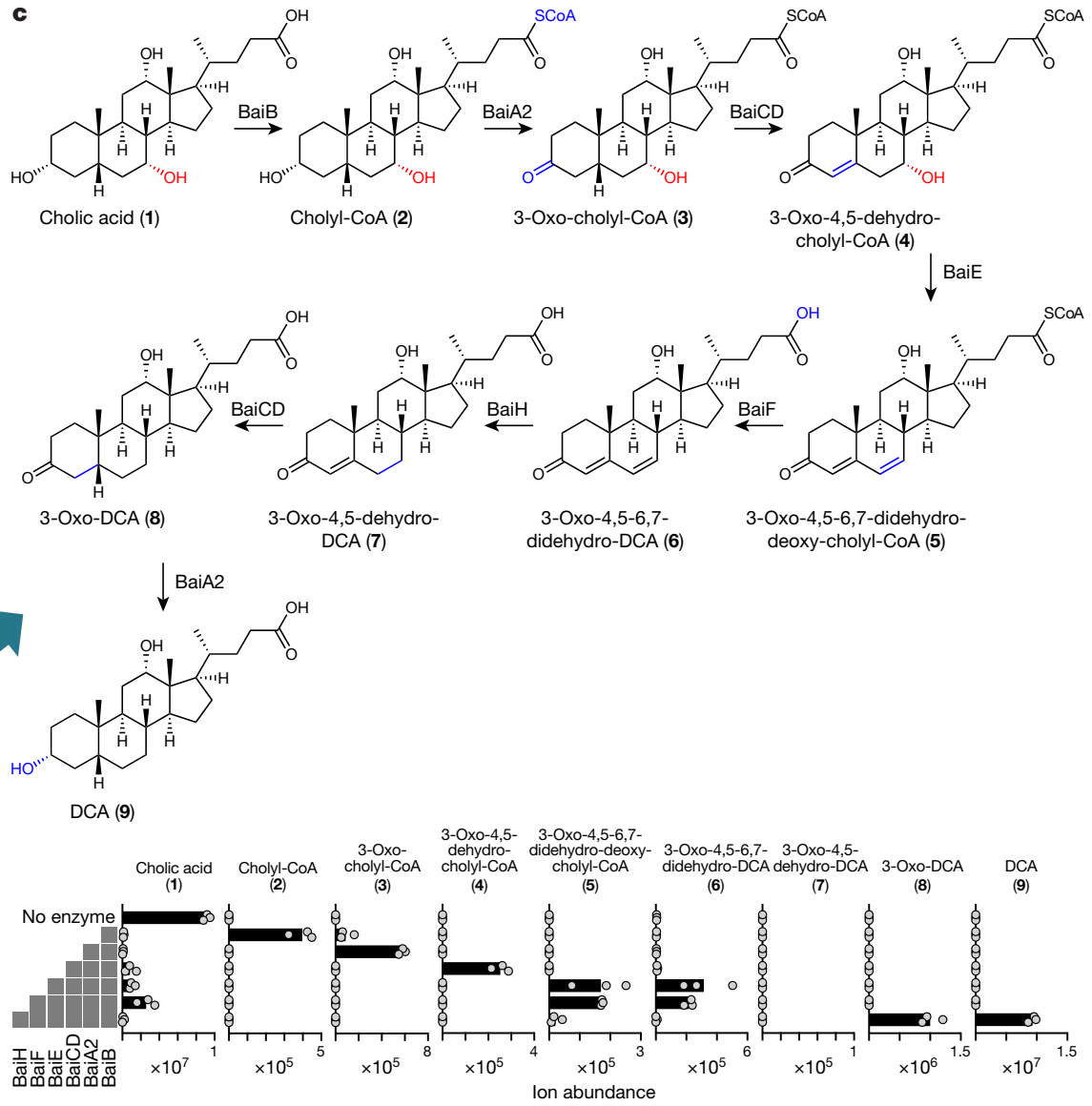
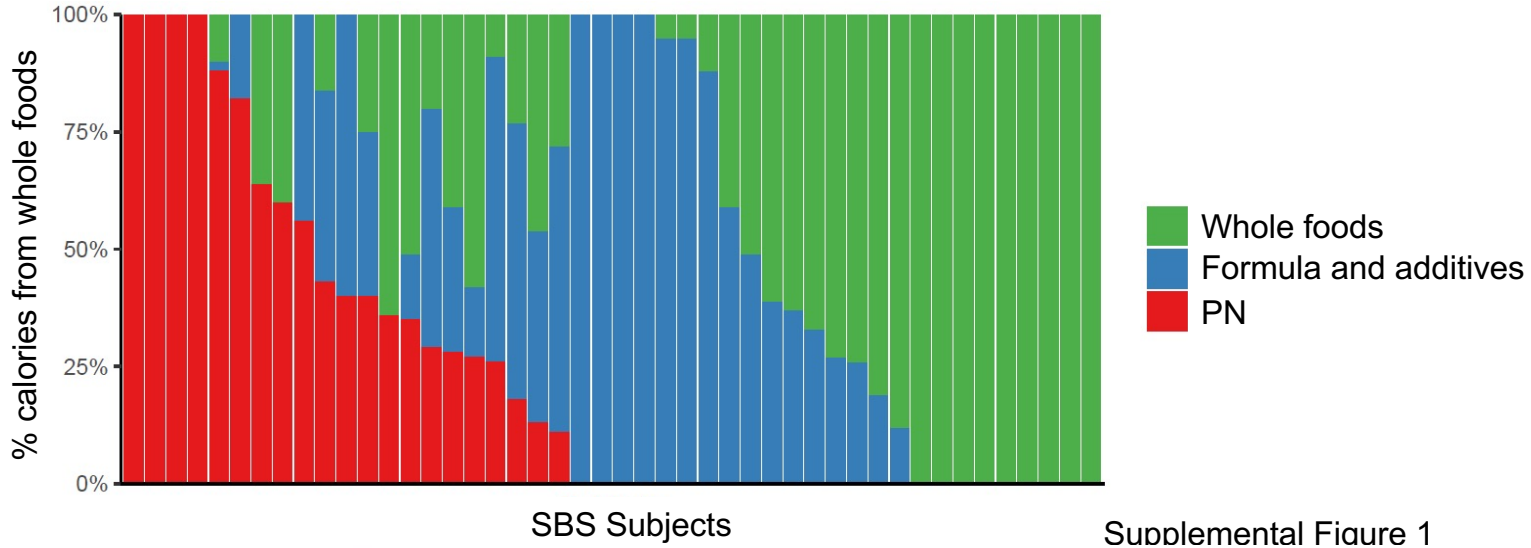
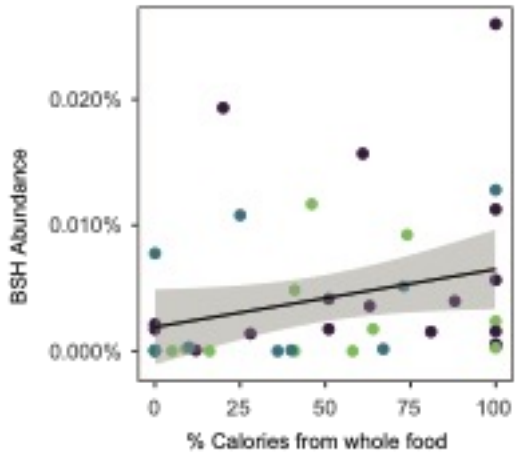


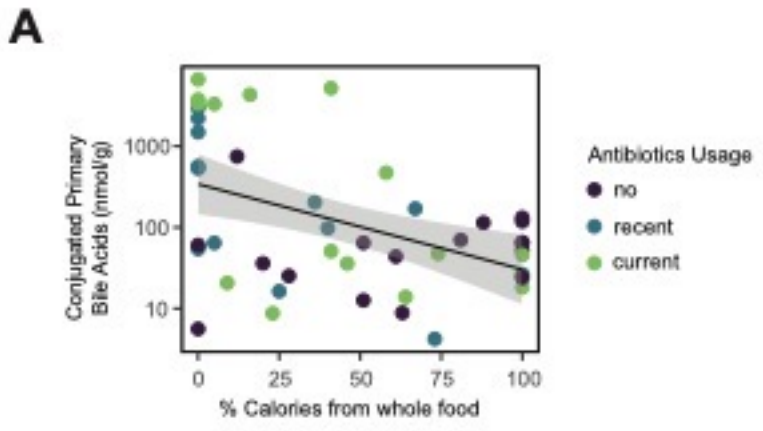
Figure on the left is from Elliot Friedman
Funabashi et al. (2020) https://doi.org/10.1038/s41586-020-2396-4

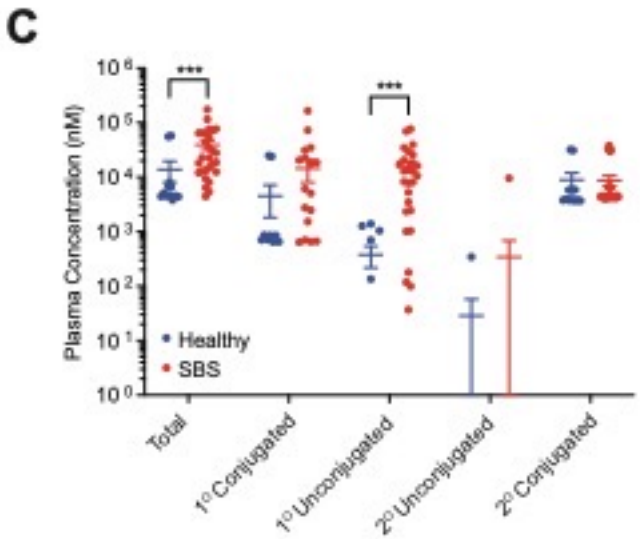
Subjects with Short bowel syndrome have altered bile acid levels



Supplemental Figure 1



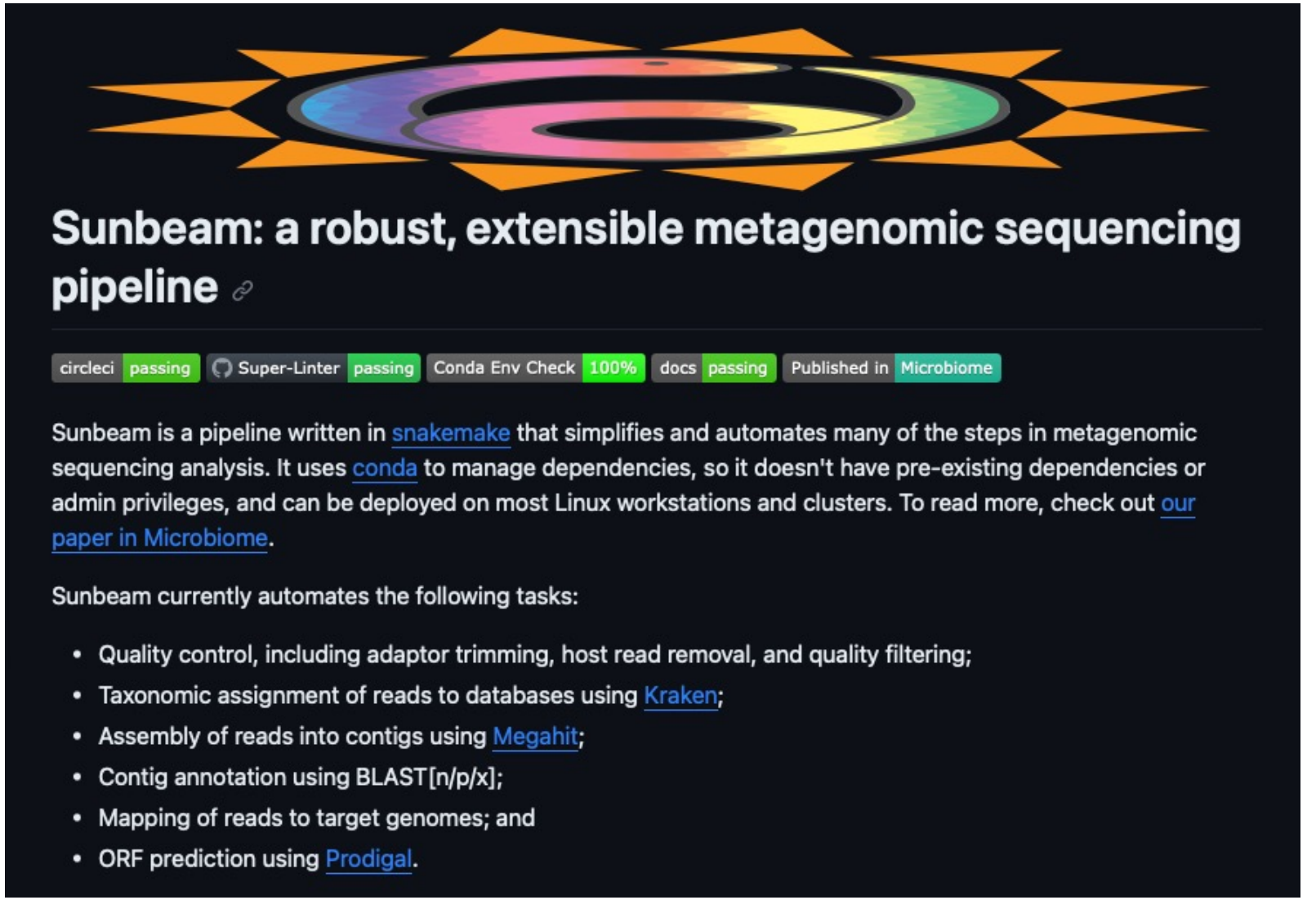




Kastl et al. 2022. Gut Microbes. Jan-Dec 2022;14(1):2083417

Sunbeam pipeline

Sunbeam software - https://github.com/sunbeam-labs/sunbeam



Sunbeam: a robust, extensible metagenomic sequencing pipeline

circleci passing | Super-Linter passing | Conda Env Check 100% | docs passing | Published in Microbiome

Sunbeam is a pipeline written in snakemake that simplifies and automates many of the steps in metagenomic sequencing analysis. It uses conda to manage dependencies, so it doesn't have pre-existing dependencies or admin privileges, and can be deployed on most Linux workstations and clusters. To read more, check out our paper in Microbiome.

Sunbeam currently automates the following tasks:

- Quality control, including adaptor trimming, host read removal, and quality filtering;
- Taxonomic assignment of reads to databases using Kraken;
- Assembly of reads into contigs using Megahit;
- Contig annotation using BLAST[n/p/x];
- Mapping of reads to target genomes; and
- ORF prediction using Prodigal.

Installing Sunbeam

- You can use the Docker image instead!

- If you would like to install from scratch, here is what to run:
  - https://github.com/sunbeam-labs/sunbeam
  - Click “To get started, see our documentation!” and then ”Quickstart Guide” for instructions

- Make sure you activate the sunbeam environment once it’s installed
  - conda activate sunbeam4.1.0

- Install the kraken extension: https://github.com/sunbeam-labs/sbx_kraken
  - sunbeam extend https://github.com/sunbeam-labs/sbx_kraken.git

Getting started – What you need

- A built sunbeam environment – make sure you activate the environment while you are executing commands
  - conda activate sunbeam4.1.0

- Your dataset of fastq.gz files in a folder

- Any databases you would need to use.
  - In this case the Kraken2 database downloaded from here: https://benlangmead.github.io/aws-indexes/k2
  - We want the "Standard-8" database for this demonstration
  - Download the .tar.gz file
  - Go to the directory where you downloaded and extract:
    - mkdir `k2_standard_08gb_20231009`
    - mv `k2_standard_08gb_20231009.tar.gz k2_standard_08gb_20231009`
    - `Cd k2_standard_08gb_20231009`
    - `tar zxvf k2_standard_08gb_20231009.tar.gz`
    - `Remember the location of this file!!`

Initiate the project

- sunbeam init sunbeam_demonstration --data_fp /sequencing/project/reads
  - This creates 3 files needed to run sunbeam
  - Samples.csv: location of all the files and the predicted sample names
  - Config.yaml: allocates resources
  - Sunbeam_config.yml: specifies parameters for the rules and file paths to data/databases

- We would ideally change the host_fp under qc, however for this demonstration, we will skip this.
- We need to change the `kraken_db_fp` to the path of the folder that contains the database you downloaded.

Run the pipeline!

- sunbeam run --profile sunbeam_demonstration/

- The first time you run the pipeline, it will create the necessary environments.

- Some useful additional commands:
  - If you would like to do a "dry-run": sunbeam run --profile sunbeam_demonstration/ -n
  - Sometimes pipelines will exit before the commands are successfully completed. Snakemake keeps track of the files who are completed vs failed mid way. It's prudent to add sunbeam run --profile sunbeam_demonstration/ --rerun-incomplete
  - If for any reason Sunbeam doesn't exit properly, you will have to "unlock" the directory: sunbeam run --profile sunbeam_demonstration/ --unlock

Output from the pipeline

- Everything will be under sunbeam_output/qc so far

- What the pipeline did:
  - Removed Illumina adapters
  - Trimmed low quality reads
  - Removed low complexity reads: These are harder to align and the assumption is that they are coming from the human genome and should be removed
  - Aligned against the host genome: we didn't specify any genomes in this demonstration, so technically we skipped this step

- Decontam sub-folder under qc contains the final reads we can use for other steps

Run Kraken!

- sunbeam run --profile sunbeam_demonstration/ all_classify
- The results will be under sunbeam_output/classify/kraken/all_samples.tsv
- You may have to modify the config.yaml to increase resources if it's failing. The database is 8Gb so you should have at least 8Gb available.
  - If your computer doesn't have 8Gb memory, don't worry, we will provide you with the output.

Let's look at the results and visualize them using R!