“Interrogating the Gut Microbiome: Estimation of Growth Dynamics and Prediction of Biosynthetic Gene Clusters”

Hongzhe Li

Perelman Professor of Biostatistics, Epidemiology and Informatics

Professor of Biostatistics and Statistics

Vice Chair of Integrative Research

Director, Center for Statistics in Big Data

Perelman School of Medicine

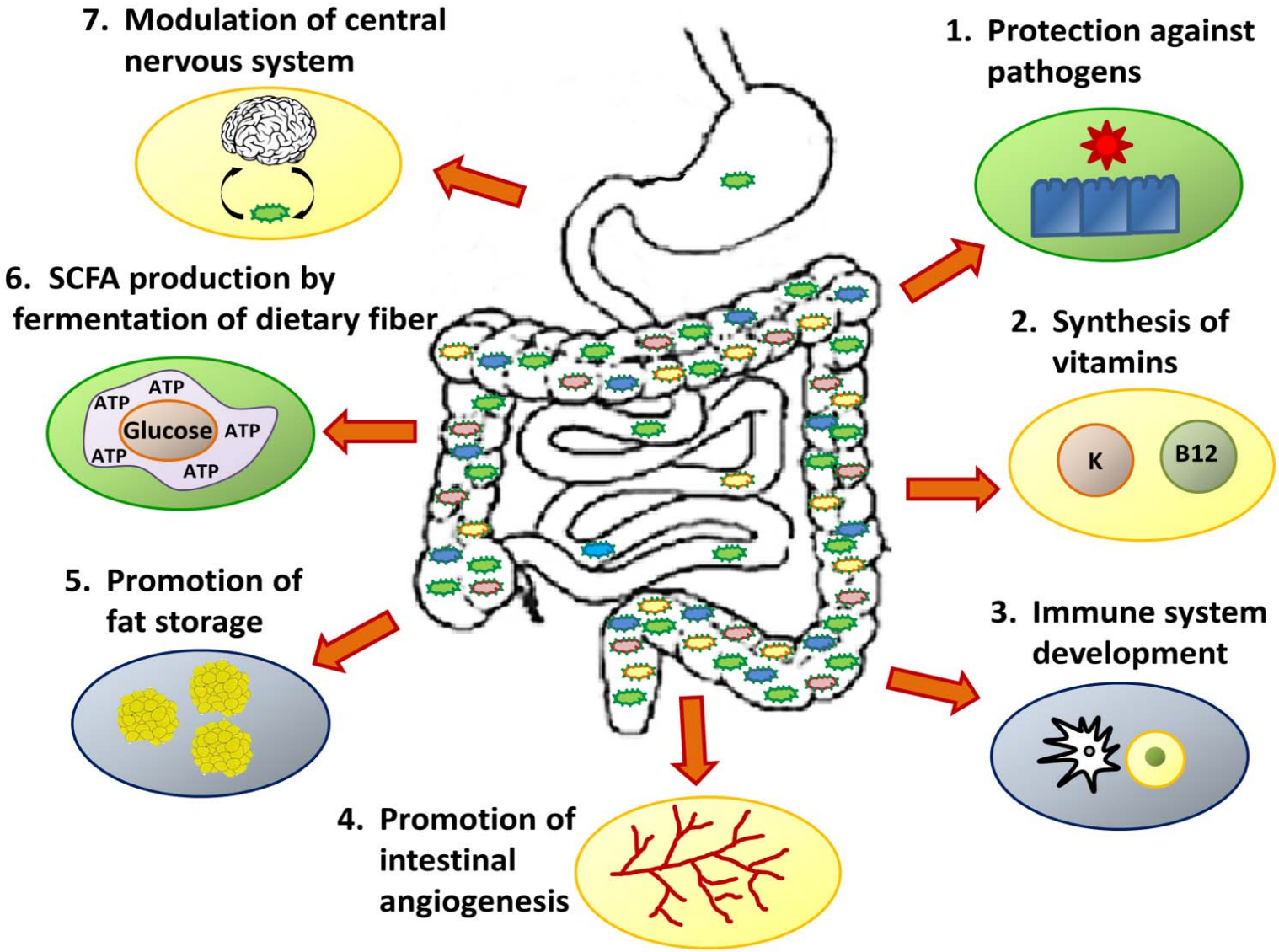
University of Pennsylvania

Interrogating the Gut Microbiome: Estimation of Growth Dynamics and Prediction of Biosynthetic Gene Clusters

Hongzhe Li
University of Pennsylvania

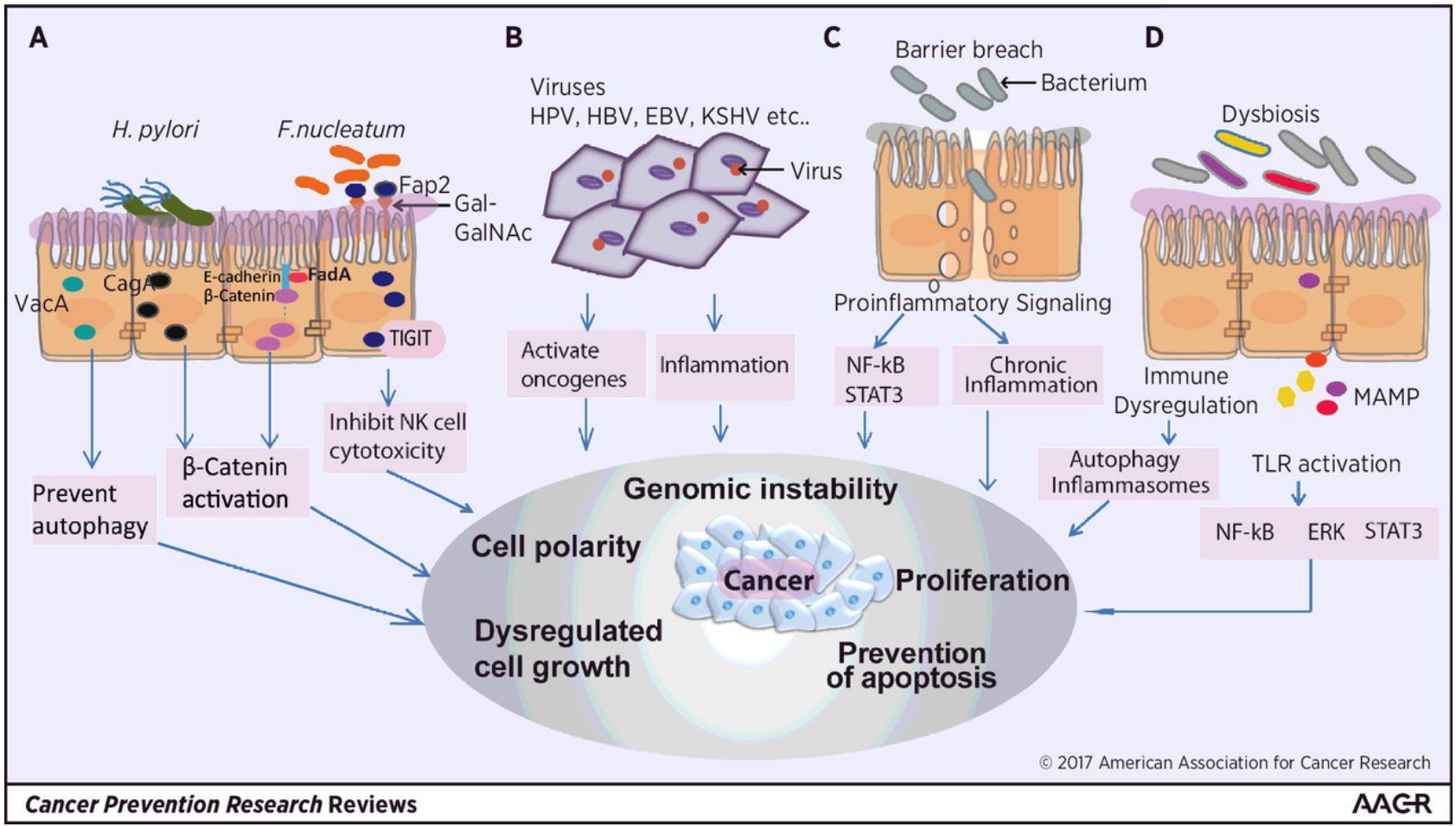
05/01/2020

Microbiome and its Function



https://ep.bmj.com/content/102/5/257 (Amon and Sanderson, 2016)

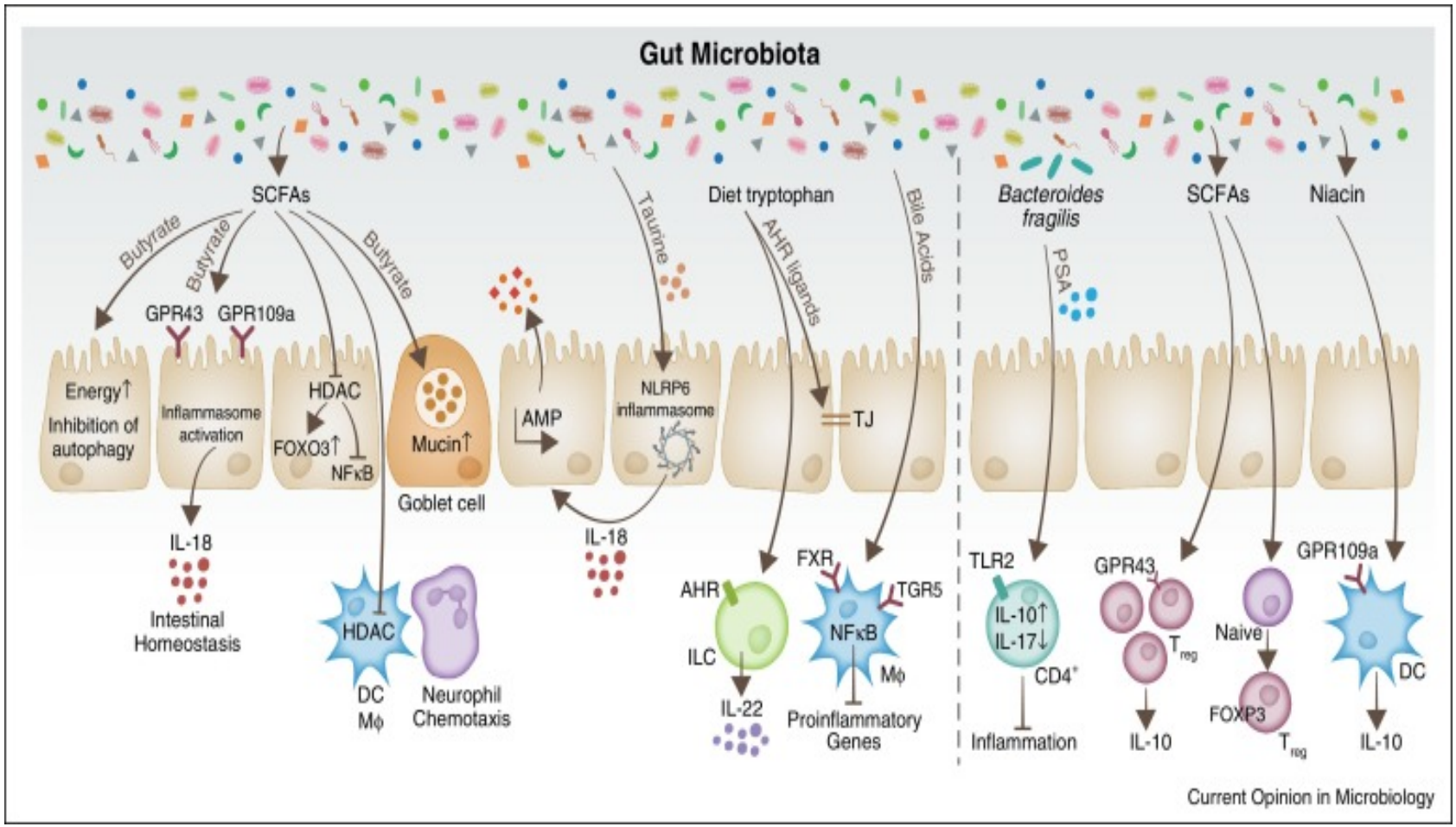
The Human Microbiome and Cancer



Rajagopala (2017 Cancer Prevention Research).

Question - microbiome-based individual treatment assignment?

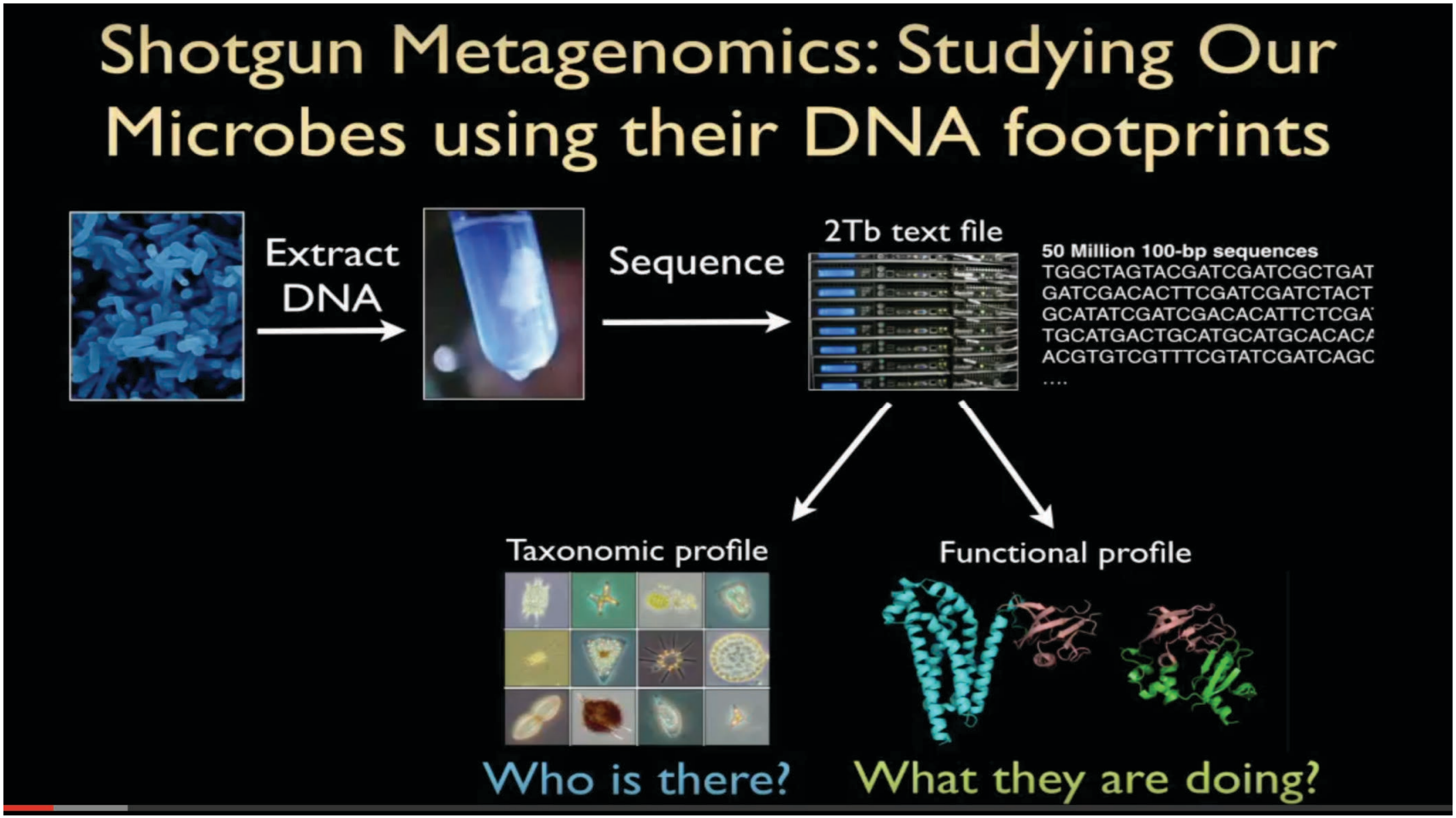
Microbiome, metabolites and immunology



Levy, Blacher and Elinav (2017, Current Opinion in Microbiology)

Question: how microbiome produces different metabolites?

Shotgun Metagenomics



Slide from Katie Pollard

Question: can we understand the growth dynamics?

Microbiome configurations/features in shotgun metagenomic data

Static Features

- Composition of taxa.
- Microbial genes/gene set or pathway abundance.
- Diversity of microbes.
- Metagenomic SNPs/structural variants.

Dynamic Features

- Bacterial growth rates
- Dynamic interactions

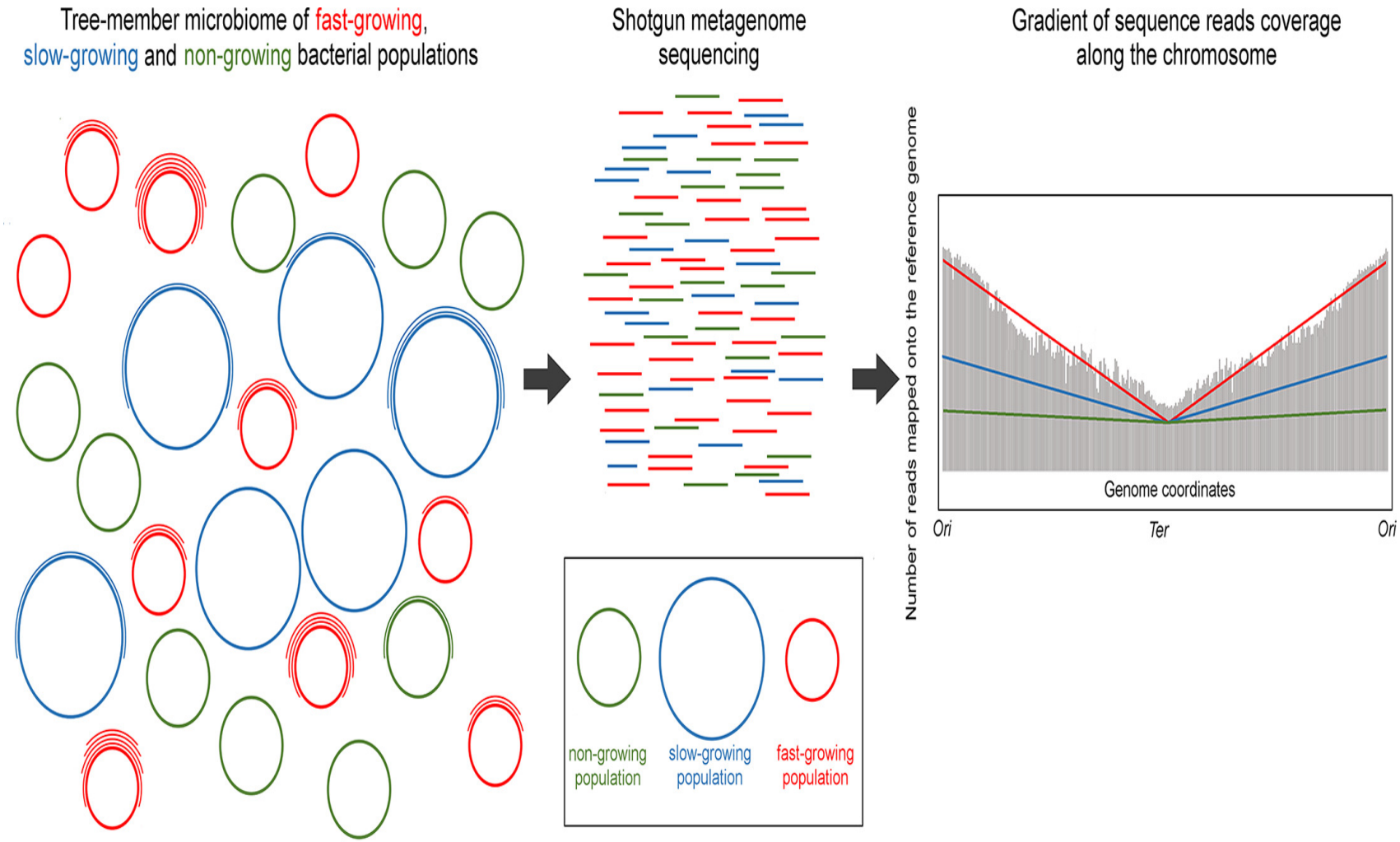
Statistical questions - how to quantify and model these features?

Topics to be discussed

- Basic microbiology science
  Estimation of bacterial growth dynamics based on genome assemblies.
- Functional microbiome
  Deep learning approach for predicting biosythetic gene clusters.

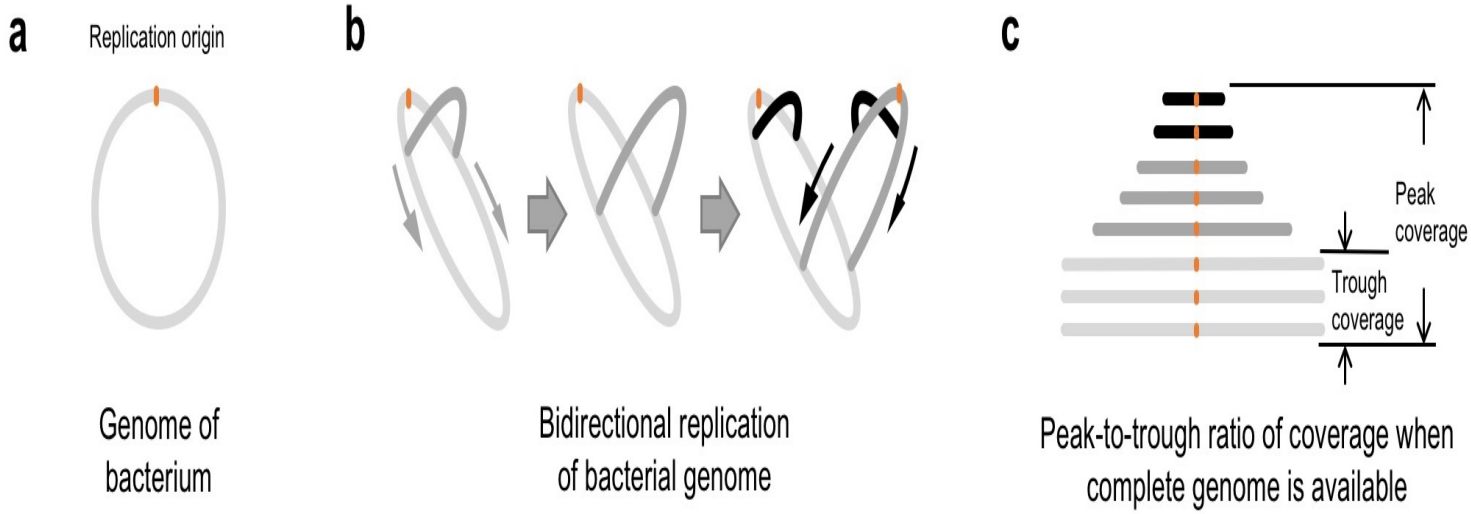
Bacterial Growth Dynamics in Metagenomics

Pienkowska et al., 2019.



Bacterial DNA Replication and Growth Dynamics

Uneven coverage of read counts reveals bacterial growth rates.



- growth dynamics for species with complete genome sequences
  Korem et al. 2015 Science.
- growth dynamics for genome assemblies - new species
  Brown et al. 2016 Nature Biotechnology
  Gao and Li, 2018 Nature Methods

Genome assemblies from shotgun data

Sangwan et al (2016): Microbiome

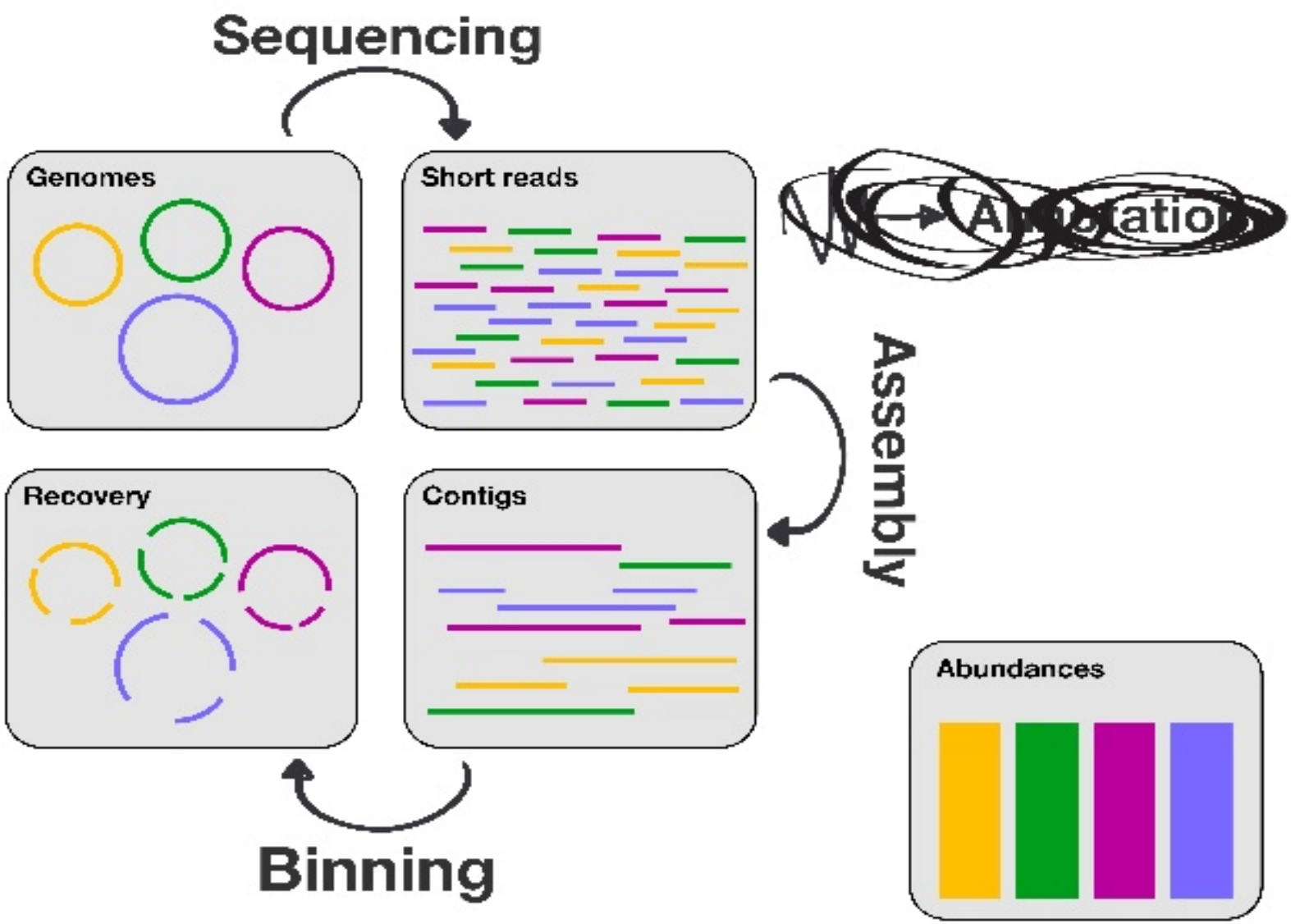


Illustration of the Statistical/Computational Problem

For a given bacteria:

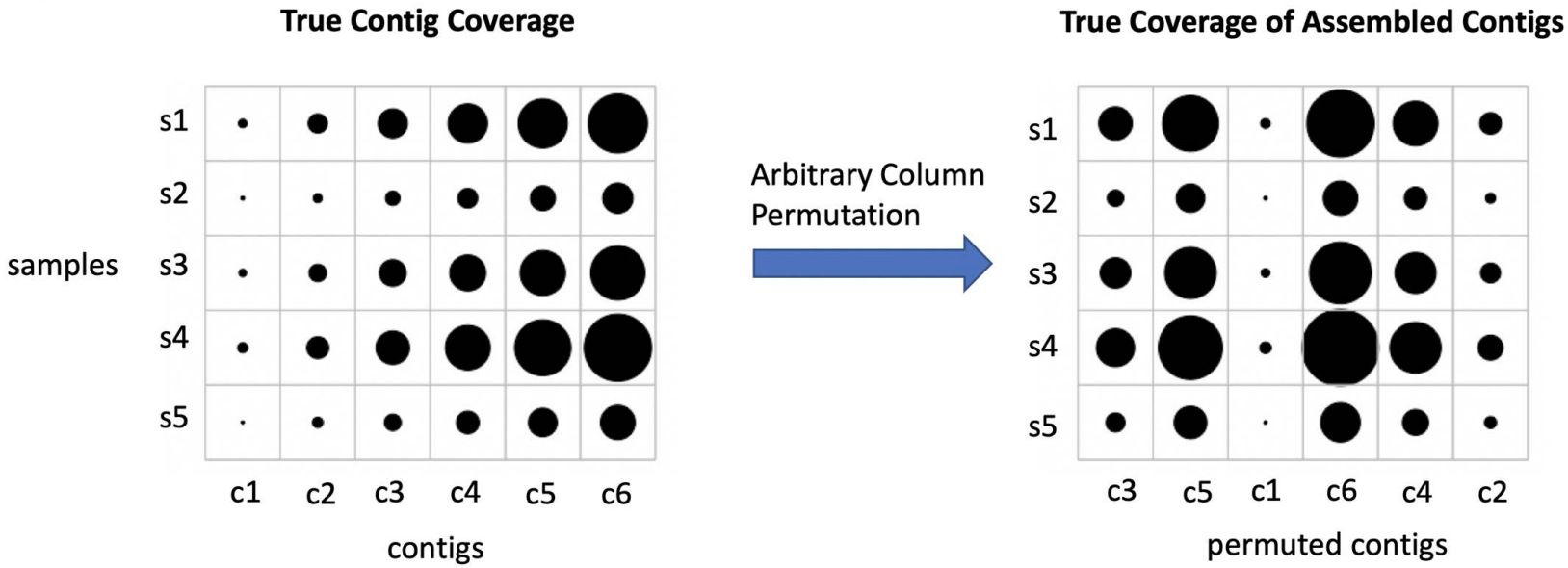


Illustration of the Statistical/Computational Problem

For a given bacteria:

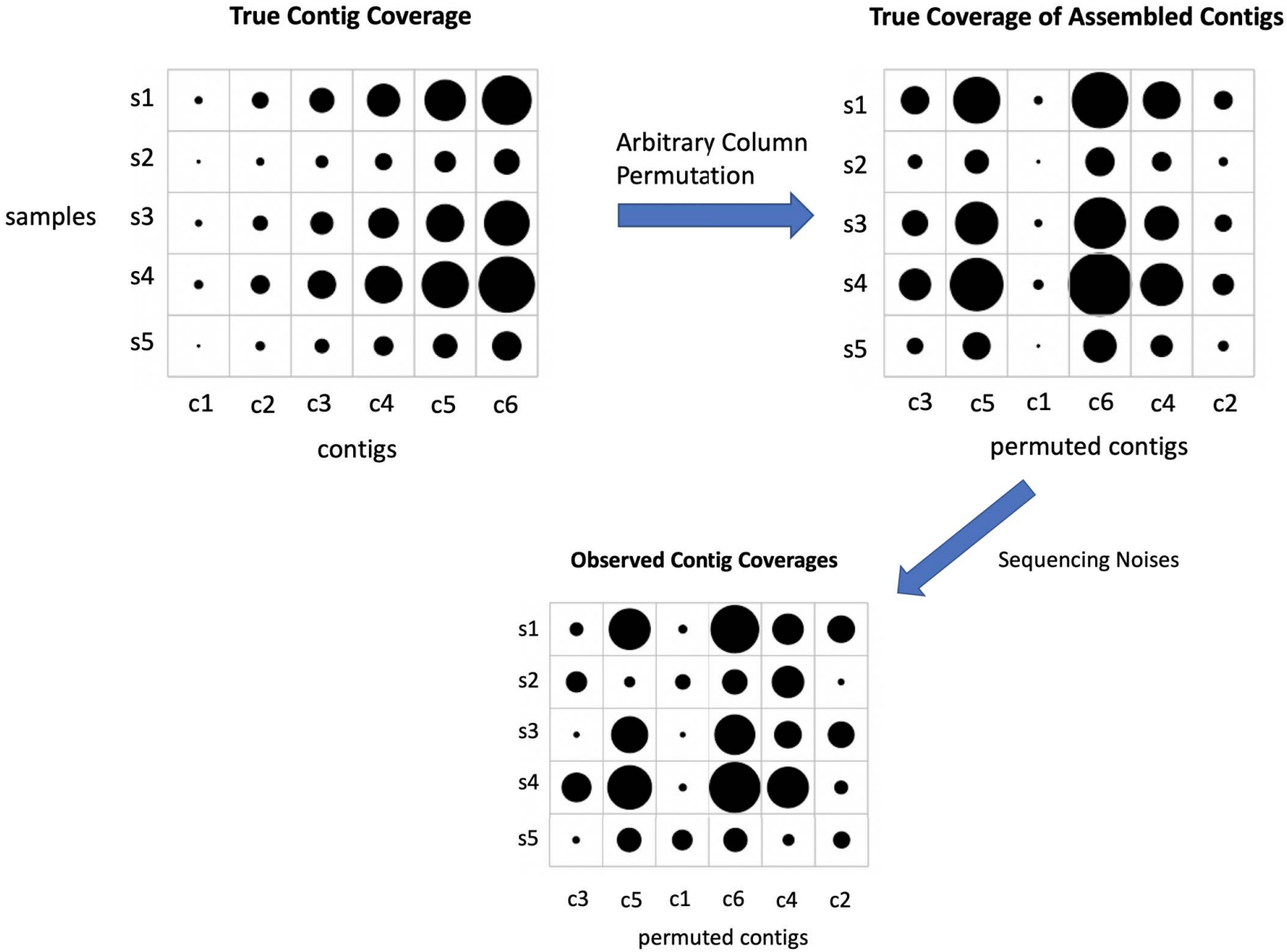
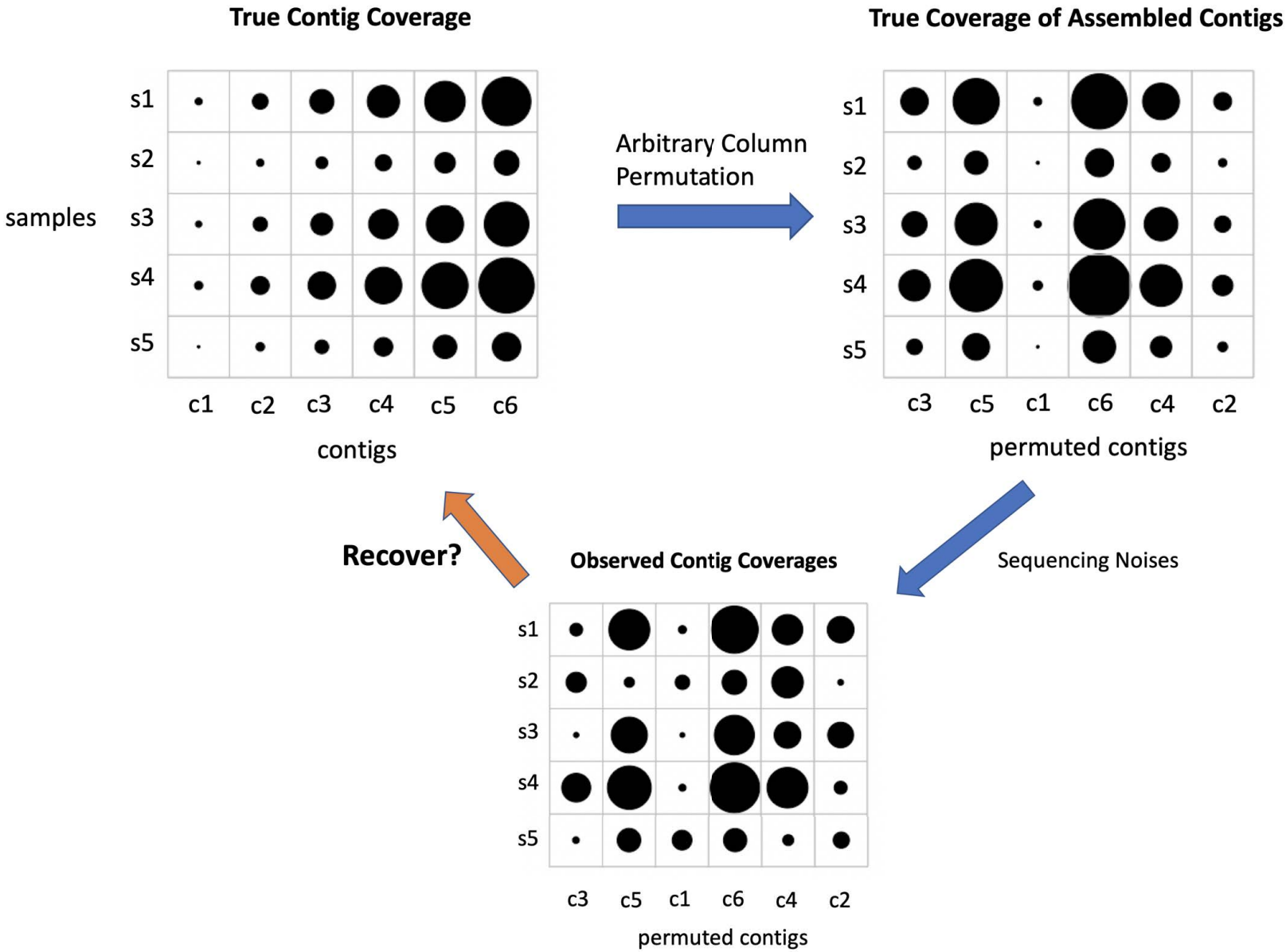


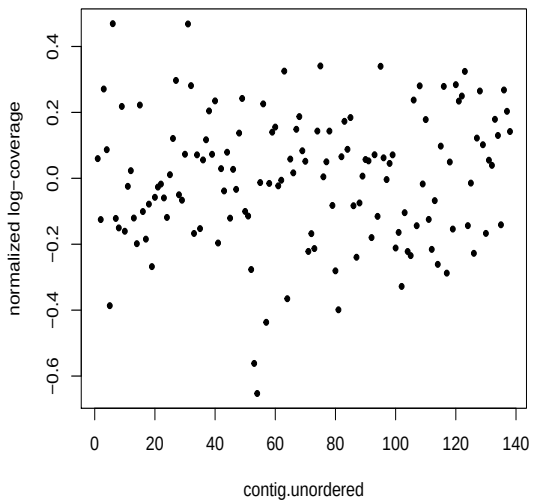
Illustration of the Statistical/Computational Problem

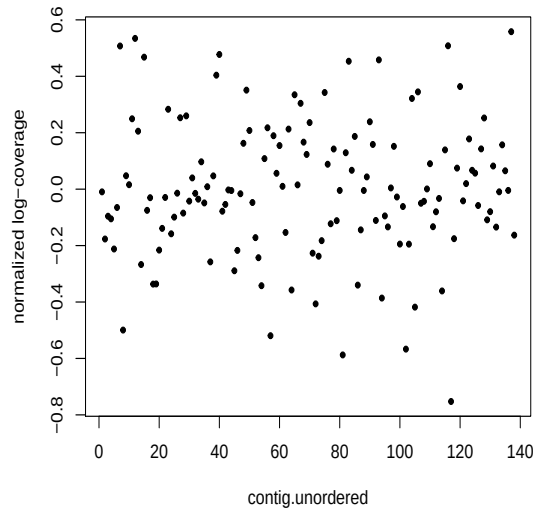
For a given bacteria:

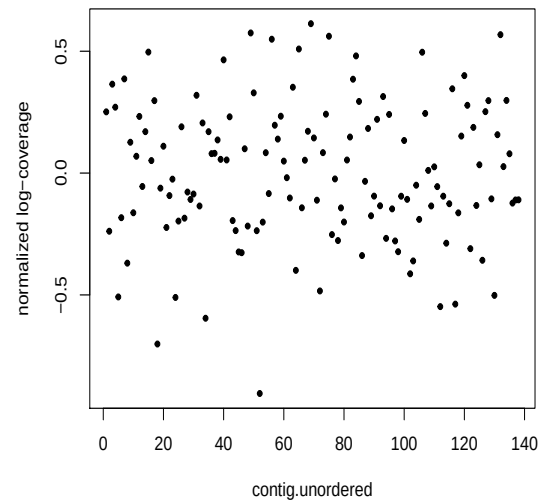


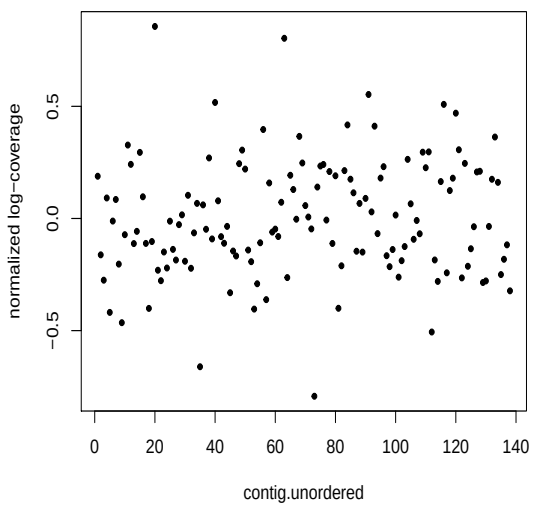
Coverages of contigs - 6 PLEASE samples

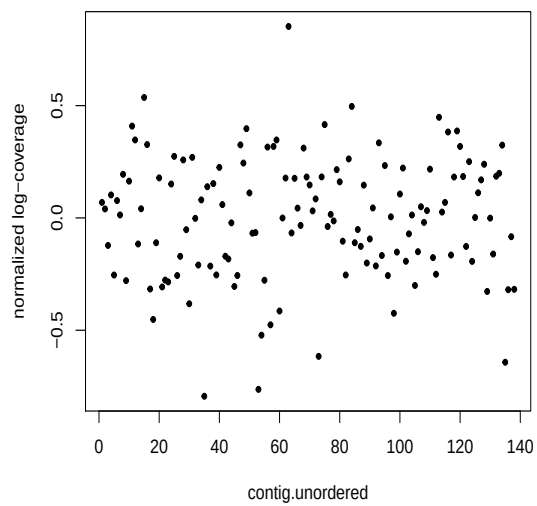
Top 3: normal. Bottom 3: IBD patients.

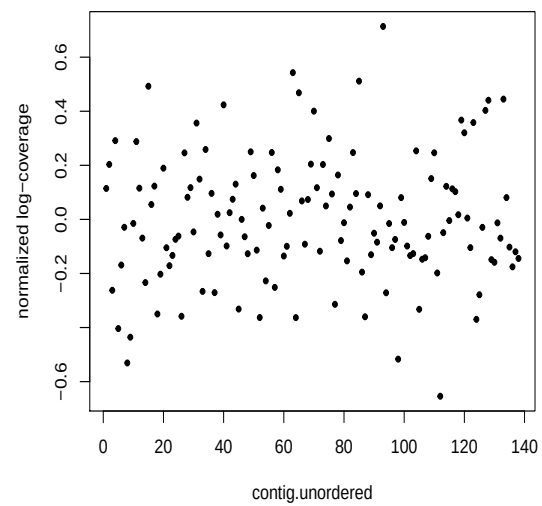






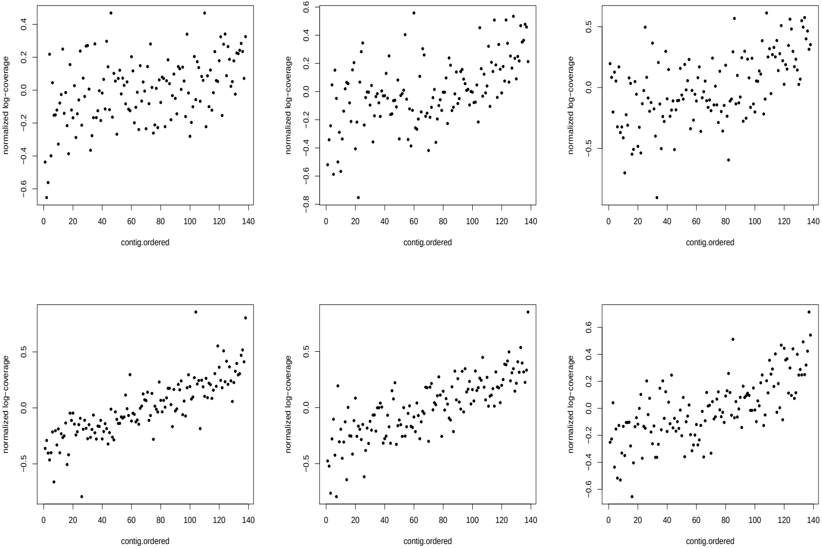






PCA vs Coverages - 6 PLEASE samples

Top 3: normal. Bottom 3: IBD patients.



Optimal permutation recovery

For a given assembly bin (species)

- Permuted Monotone Matrix Model: X is GC-adjusted log-read counts along the genome - n samples and p contigs,

$$Y_{n\times p} = \pi(X_{n\times p}), \quad X_{n\times p} = \Theta_{n\times p} + Z_{n\times p}$$

where $X, \Theta, Z \in \mathbb{R}^{n\times p}$, π is a column-permutation operator, and

$$\Theta \in \mathcal{D} = \left\{ \Theta = (\theta_{ij}) : \quad 0 < \theta_{i,j} \leq \theta_{i,j+1} < \infty, \forall i, j \right\}.$$

Z: some additive noise (i.i.d. Gaussian, $N(0, \sigma^2)$).

- The goal is to recover π based on observed Y.
- Solution: 1st PC, $\hat{\pi} = \mathfrak{r}(\hat{w}_1^\top Y)$ as an estimate of π, $\hat{w}_1$ is loading coefficients of the 1st PC.

Theoretical Properties (Ma, Cai and Li 2020 JASA)

Linear growth model - the parameter space for Θ:

$$\mathcal{D}_L = \left\{ \Theta \in \mathbb{R}^{n\times p} : \begin{array}{l} \theta_{ij} = a_i \eta_j + b_i, \text{ where } a_i, b_i \geq 0 \text{ for } 1 \leq i \leq n, \\ 0 \leq \eta_j \leq \eta_{j+1} \text{ for } 1 \leq j \leq p-1 \end{array} \right\},$$

A key quantity:

$$\Gamma(\Theta) = \left(n^{-1} \sum_{i=1}^{n} a_i^2 \right)^{1/2} \cdot \min_{1\leq i<j\leq p} |\eta_i - \eta_j|.$$

Theorem (Exact Recovery)

Suppose the noise Z are i.i.d. $N(0, \sigma^2)$. Then under some mild conditions, whenever

$$\Gamma \gtrsim \sigma \sqrt{\frac{\log p}{n}},$$

we have $\hat{\pi} = \pi$ with probability at least $1 - p^{-c}$.

Estimation of PTR

Proposed estimators of peak/trough coverage: $\hat{\Theta}_{\max}$/ $\hat{\Theta}_{\min}$:

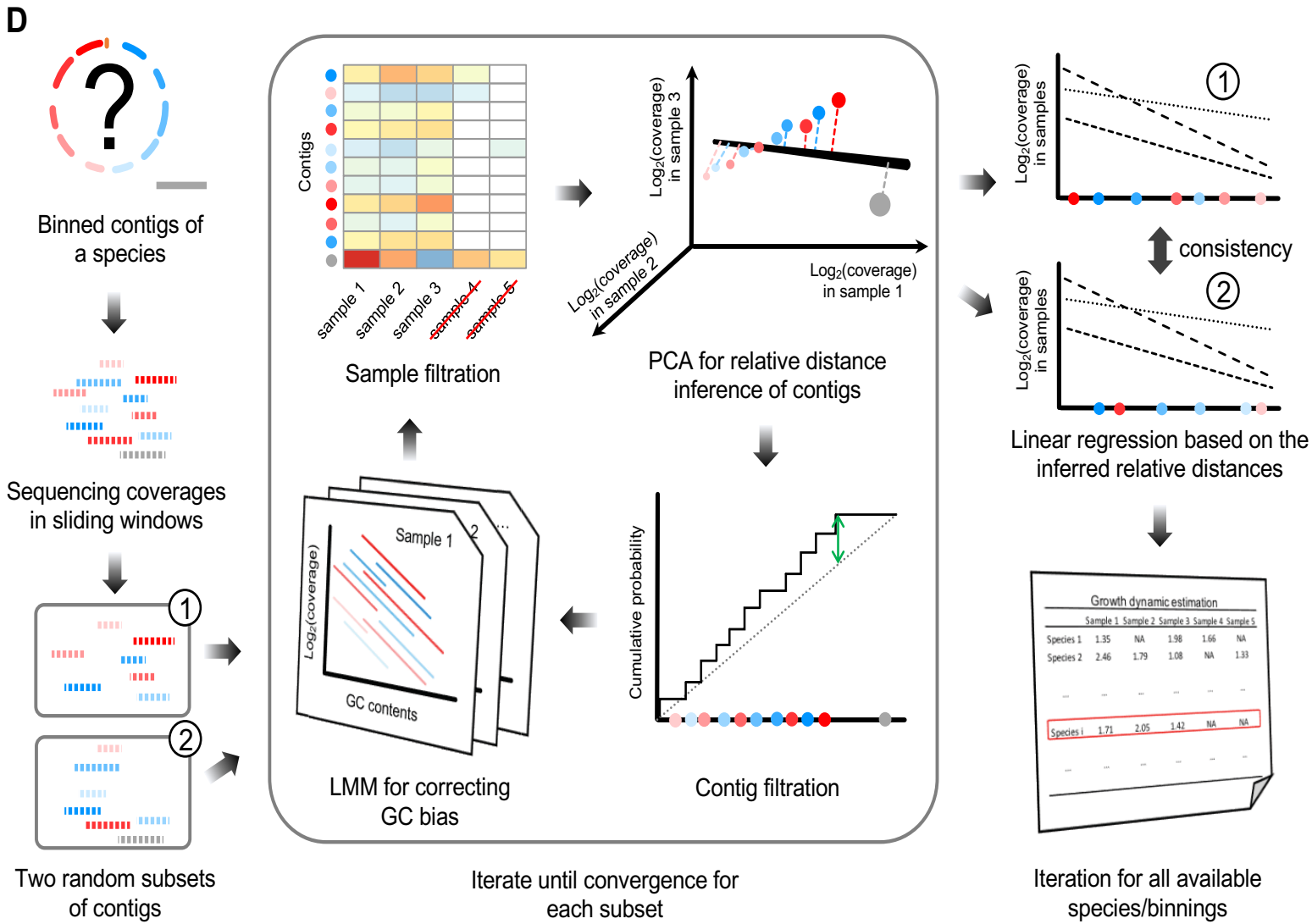
1. Obtain the optimal permutation estimator $\hat{\pi}$ to reorder the columns (contigs);
2. Fit simple linear regression for each row (sample);
3. Define $\hat{\Theta}_{\max}$ and $\hat{\Theta}_{\min}$ as the **fitted maximum and minimum values**.

$\Longrightarrow$ DEMIC algorithm.

Optimal and adaptive estimation of PTR and the two extreme values (peak and trough) for general growth model.
Ma, Cai and Li: 2020 submitted

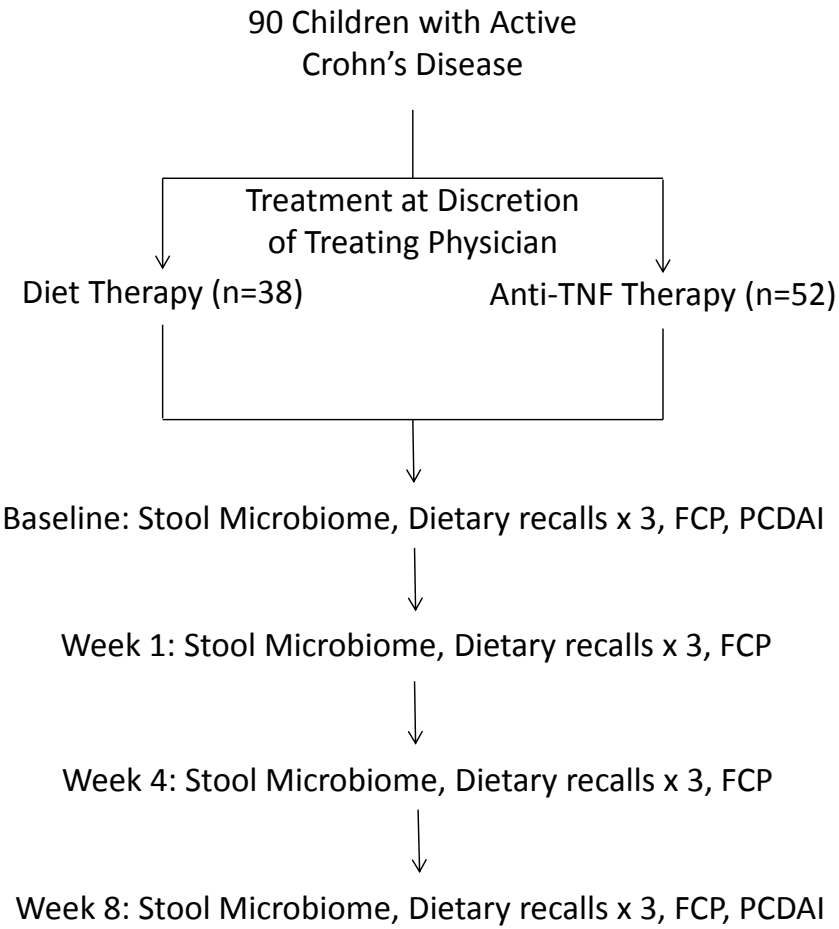
DEMIC Software

Dynamics Estimator of Microbial Communities (DEMIC)
https://github.com/scottdaniel/sbx_demic (Scott Daniel)



Penn PLEASE Study (Lewis et al. (2015): Cell Host & Microbe)

PLEASE (Pediatric Crohn's Disease) study at Penn: 90×4 shotgun metagenomic samples and 26 normal children (ave 11×10^6 paired-end reads).
Outcome: Fecal calprotection (FCP) (reduction below 250mcg/g).
Metabolomics: fecal metabolites.



Anti-TNF: 26 (50%) a reduction in FCP below 250 mcg/g.

Enteral Diet: 12 (32%) a reduction in FCP below 250 mcg/g.

Lewis, Chen et al. (2015): Cell Host & Microbe.

Species with differential growth dynamics

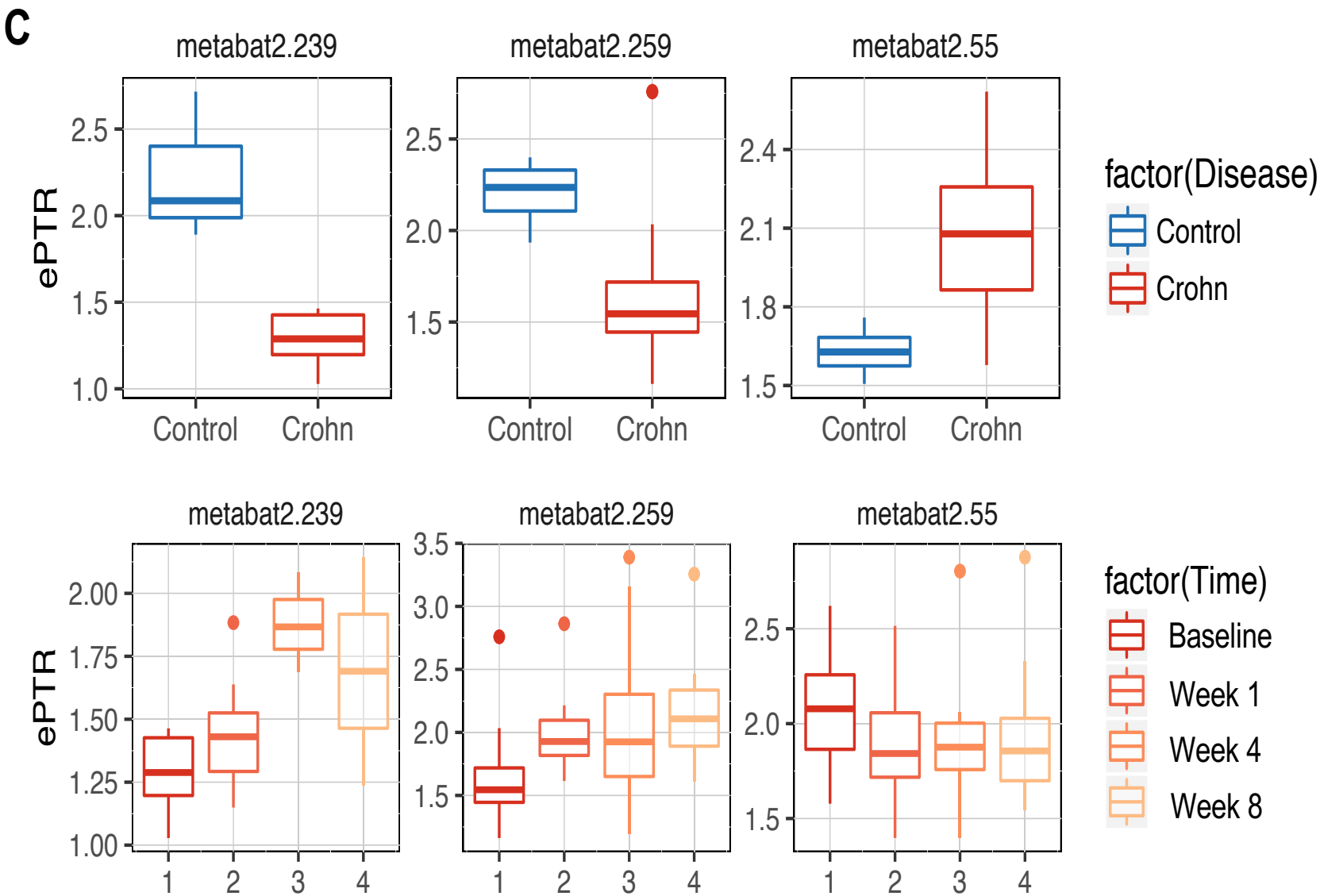
DEMIC estimated growth dynamics for 278 species, 20% in 50 or more samples.

The assembly quality and marker lineage of seven contig clusters with different growth rates in healthy and Crohn's disease samples of PLEASE data set (FDR<0.05)

Contig cluster	Completeness	Contamination	Control vs Crohn's	Marker lineage
metabat2.187	61.7%	0	High	kBacteria
metabat2.239	58.5%	1.8%	High	oClostridiales
metabat2.250	66.6%	0.8%	High	pProteobacteria
metabat2.259	79.3%	2.1%	High	kBacteria
metabat2.270	72.0%	2.0%	High	fLachnospiraceae
metabat2.369	68.8%	2.8%	High	fLachnospiraceae
metabat2.55	55.2%	1.9%	Low	oClostridiales

Shift of growth dynamics after treatment

oClostridiales, oClostridiales, kbacteria (uncharacterized)



Summary and software

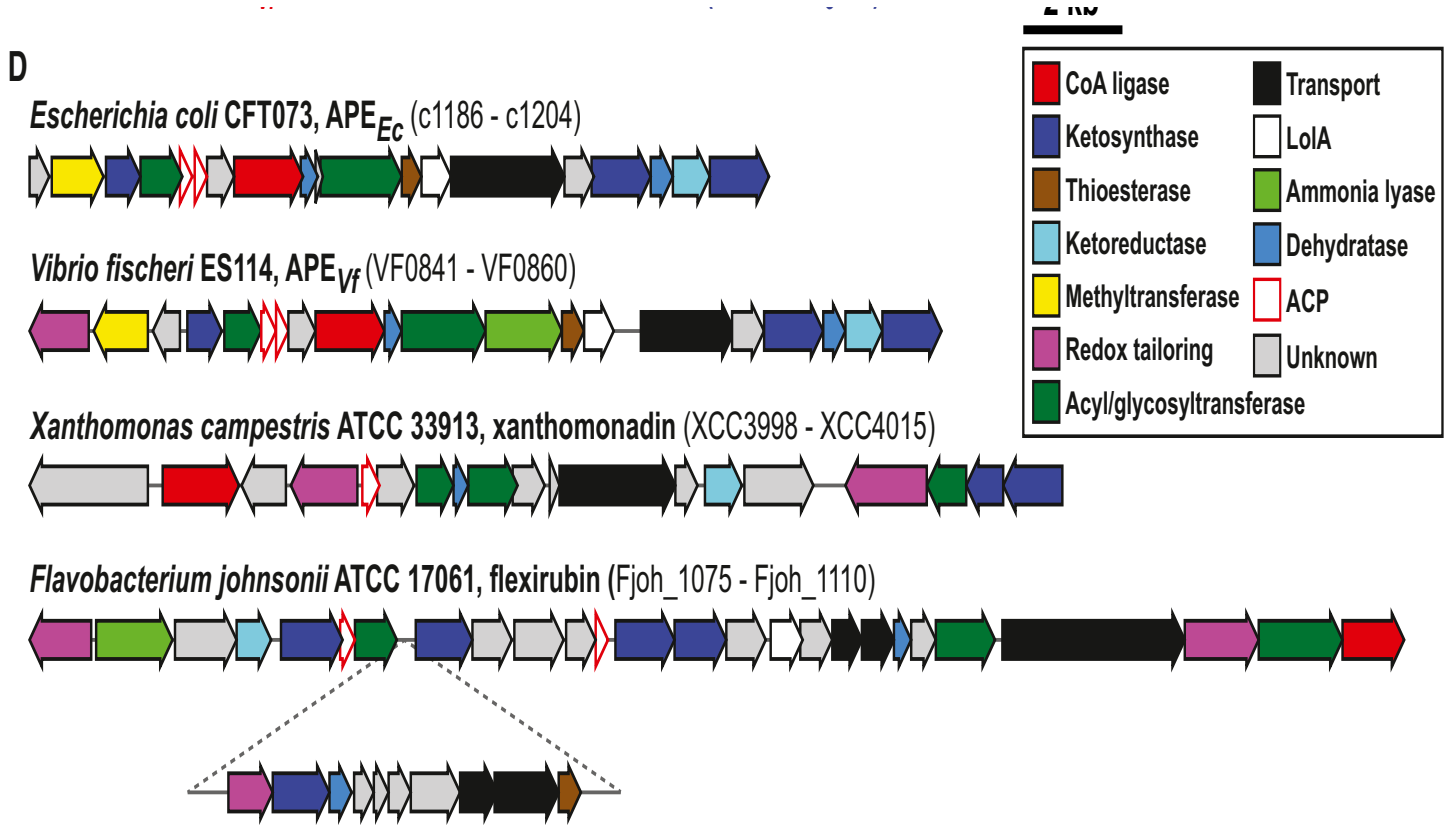
Dynamics Estimator of Microbial Communities (DEMIC)
https://github.com/scottdaniel/sbx_demic (Scott Daniel)
(Gao and Li, 2018 Nature Methods)

Optimal permutation recovery for monotone permuted matrix.
(Ma, Cai and Li, 2020 JASA)

Biosynthetic gene clusters (BGCs)

Bioactive secondary metabolites (SMs) - antibiotics, anticancer reagents, etc

SMs - encoded by genes that cluster together in a genetic package, referred to as a biosynthetic gene cluster (BGC).



Cimermancic et al (2014, Cell)

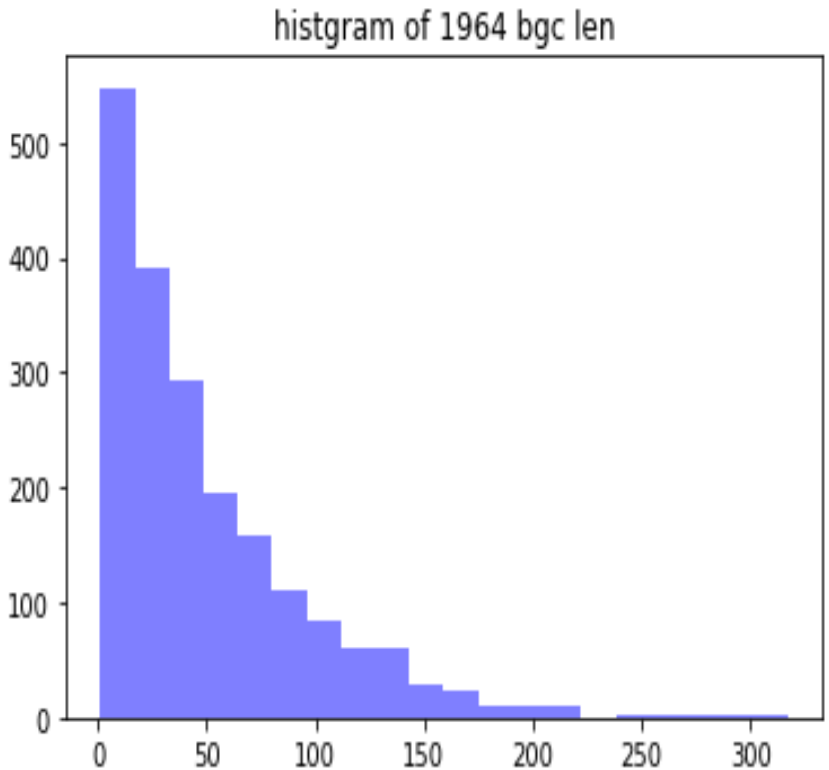
Identification of all BGCs in bacterial genomes

Training Data set:

1,984 BGC gene sequences from MIBiG v1.4 database, ORF/gene prediction, Pfam domains. 3,685 Pfam domains.

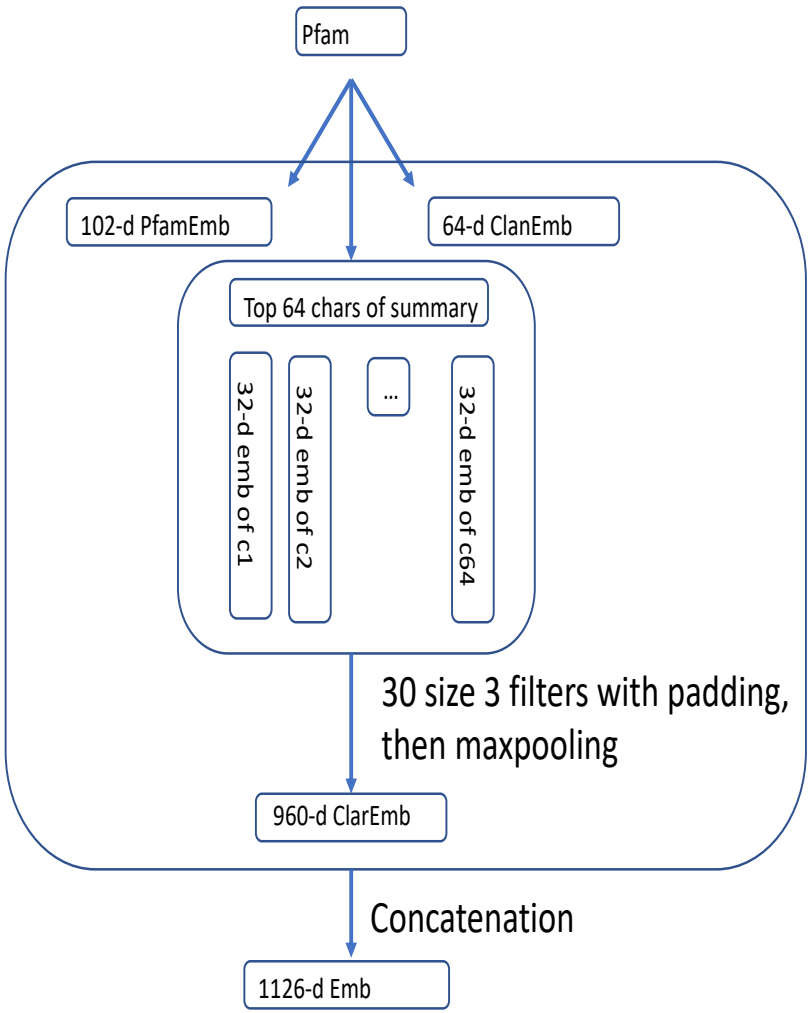
1,868 BGCs with 3-250 Pfam domains, 1094 species

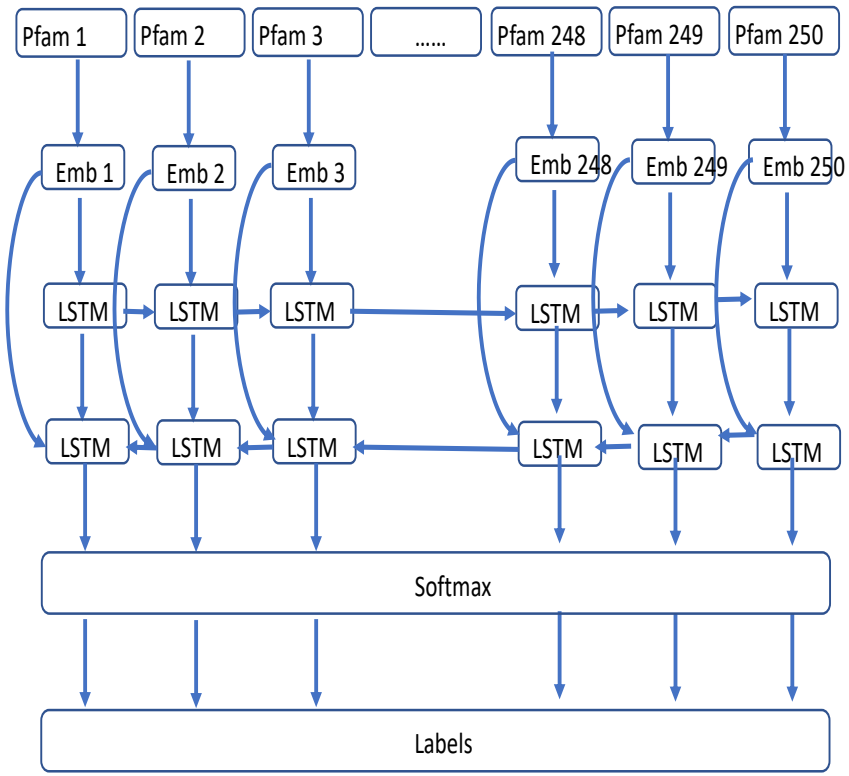
Background: 5,666 reference genomes from NCBI database, 11,427 unique Pfam domains. $n_{non-BGC} = 10,128$ controls.



DeepMBGC - deep learning and embedding

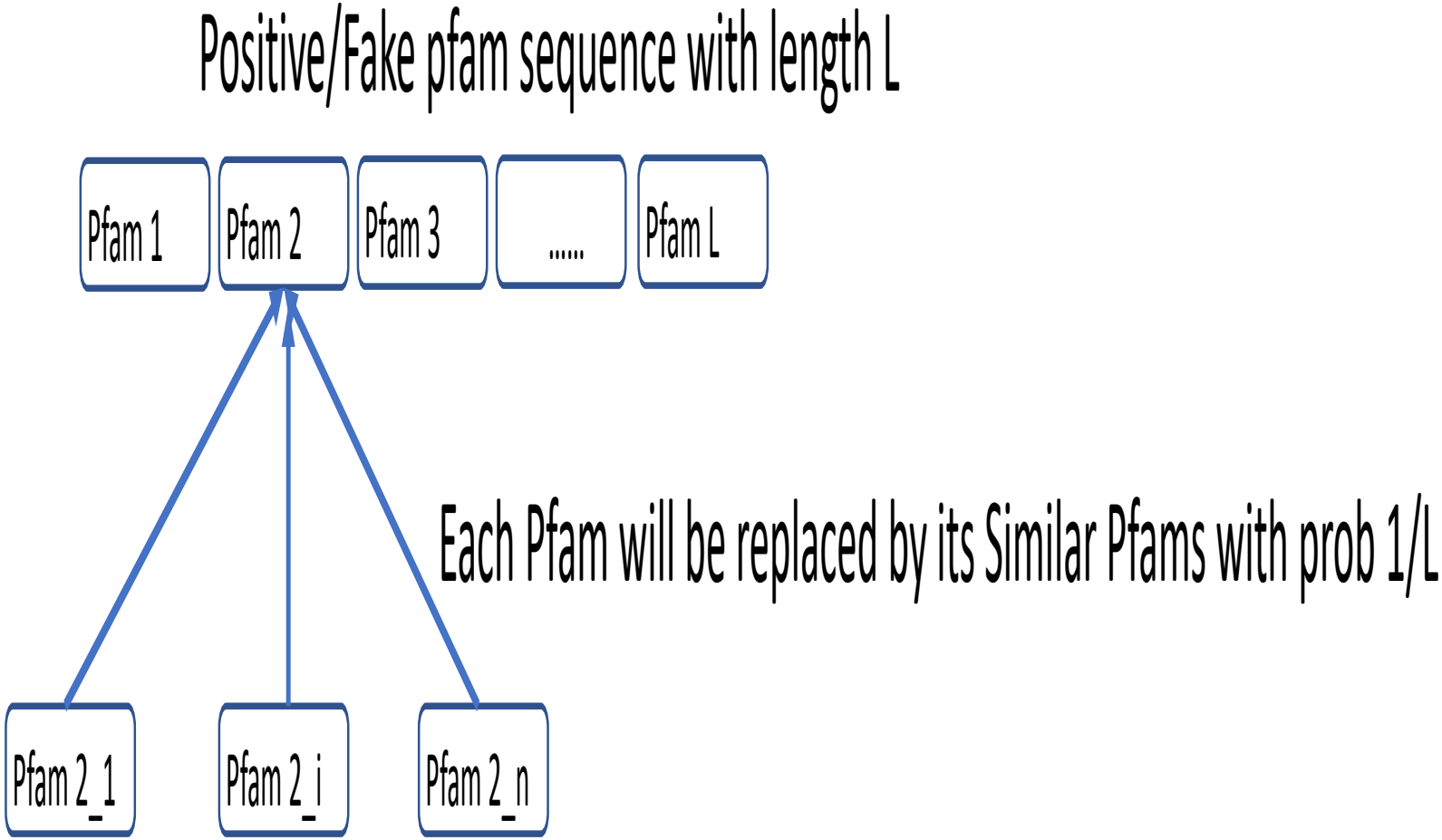
Embedding: Pfam domain names, Pfam clans, Pfam function descriptions (Liu, Li and Li, in preparation) $\Rightarrow$ LSTM RNN



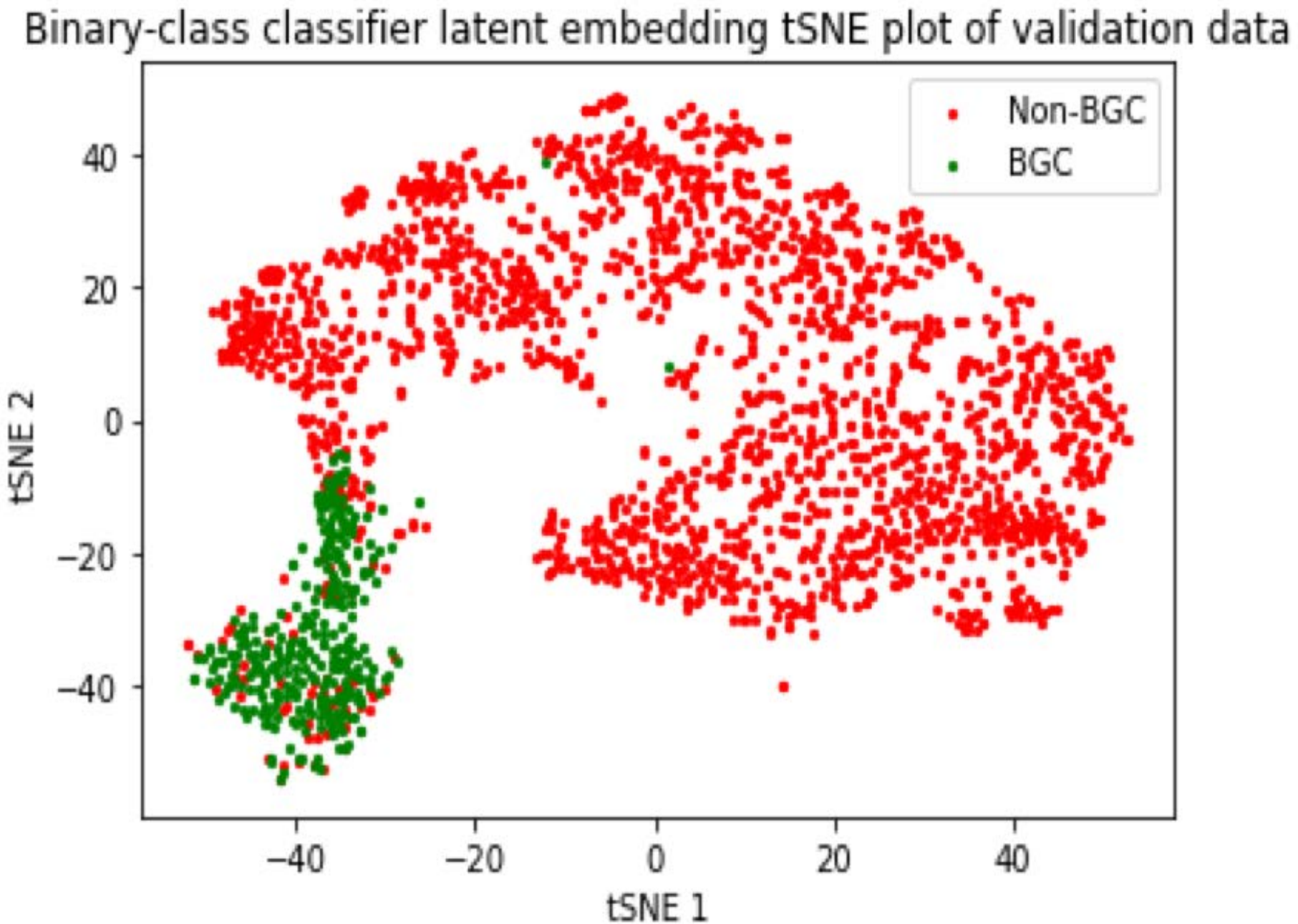


DeepMBGC - Data Augmentation

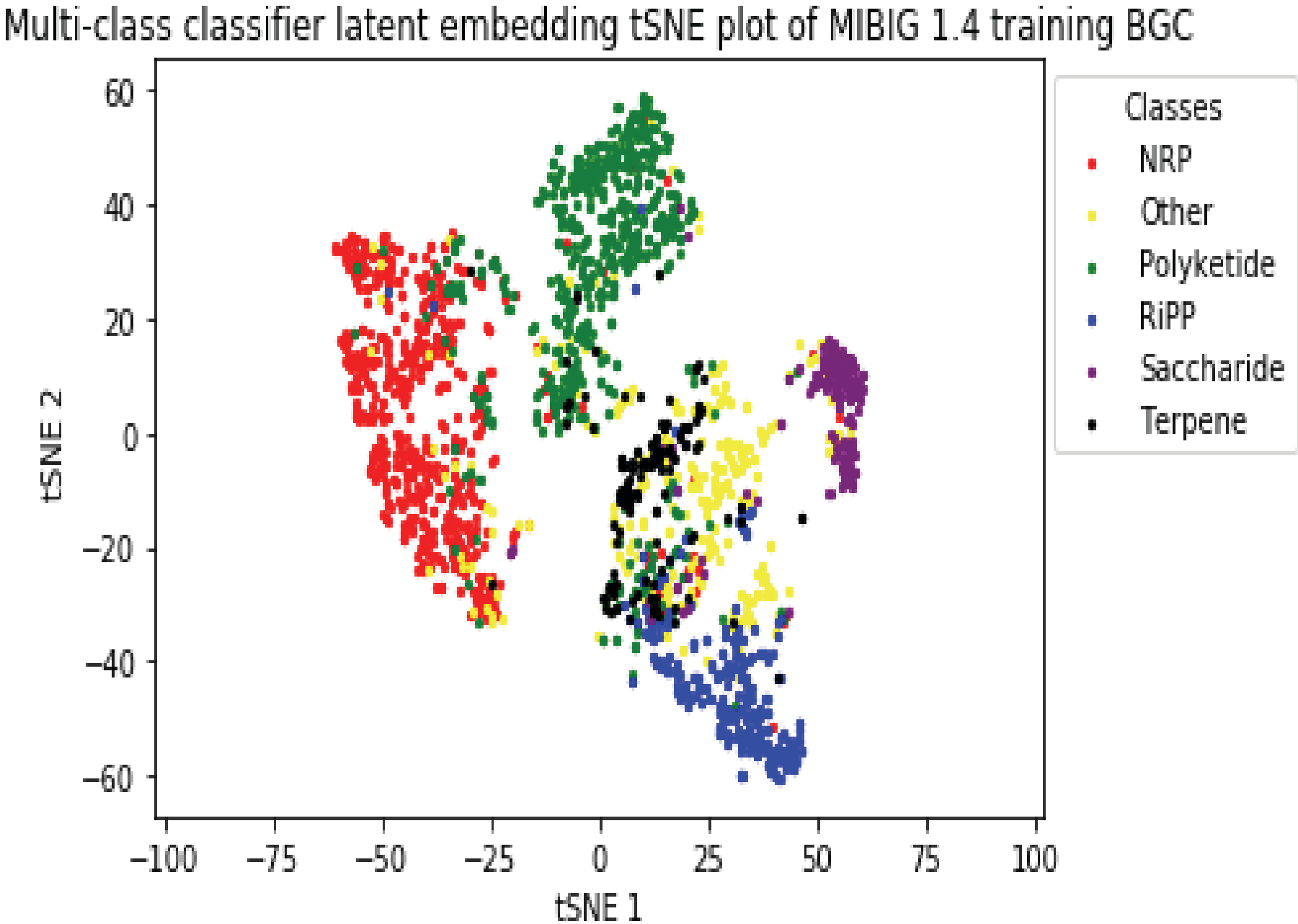
On expectation, a sequence has one Pfam domain being replaced, each epoch with new perturbed data.



DeepMBGC - Embedding, binary case



DeepMBGC - Embedding, multi-class case



DeepMBGC Prediction Results - Pfam level

Testing set: 13 genomes with 291 known BGCs never used in training, 10x13=130 artificial genomes with 291 known BGCs fixed in original genomes, other replaced with non-BGCs.

Table: Prediction performance at the Pfam level

	DeepBGC	DeepMBGC	DeepMBGC+ Data Argumentation
precision	0.831(0.0069)	0.774(0.0053)	**0.833(0.0042)**
recall	0.748(0.0025)	**0.883(0.0018)**	0.852(0.0016)
f1	0.788(0.0029)	0.825(0.0026)	**0.842(0.0024)**
roc	0.984(0.0002)	0.989(0.0003)	**0.989(0.0002)**
pr	0.881(0.0023)	0.919(0.0017)	**0.921(0.0016)**

DeepBGC: Hannigan et al., 2019 NAR.

DeepMBGC Prediction Results - BGC level

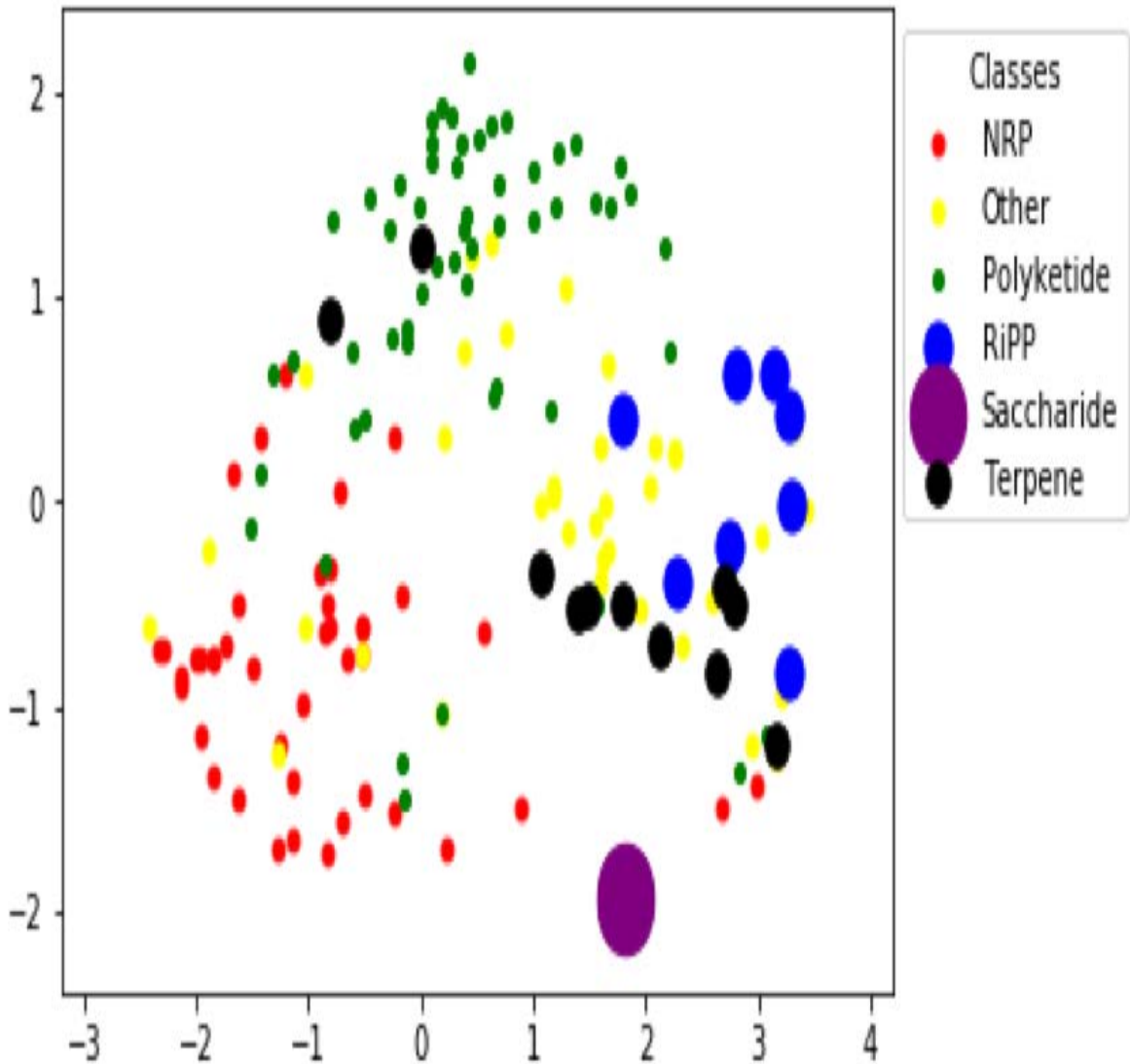
BGCs - infered based on estimated max Pfam probabilties, length between 3 and 250 Pfams.

Table: Prediction performance at the BGC level, F1 score

	DeepBGC	DeepMBGC	DeepMBGC+ Data Argumentation
overlap>0.0	0.74(0.0026)	0.808(0.0030)	**0.817(0.0029)**
overlap≥0.2	0.736(0.0023)	0.805(0.0028)	**0.815(0.0029)**
overlap≥0.4	0.711(0.0029)	0.784(0.0028)	**0.799(0.0030)**
overlap≥0.6	0.661(0.0037)	0.733(0.0052)	**0.753(0.0041)**
overlap≥0.8	0.556(0.0051)	0.609(0.0051)	**0.645(0.0044)**
overlap= 1	0.268(0.0048)	0.218(0.0065)	**0.286(0.0062)**

DeepMBGC multiclass prediction

Testing set: 160 new BGC were deposited to MiBIG v1.5



Multi-class accuracy: 74.8%

Recall rate: 77.5%

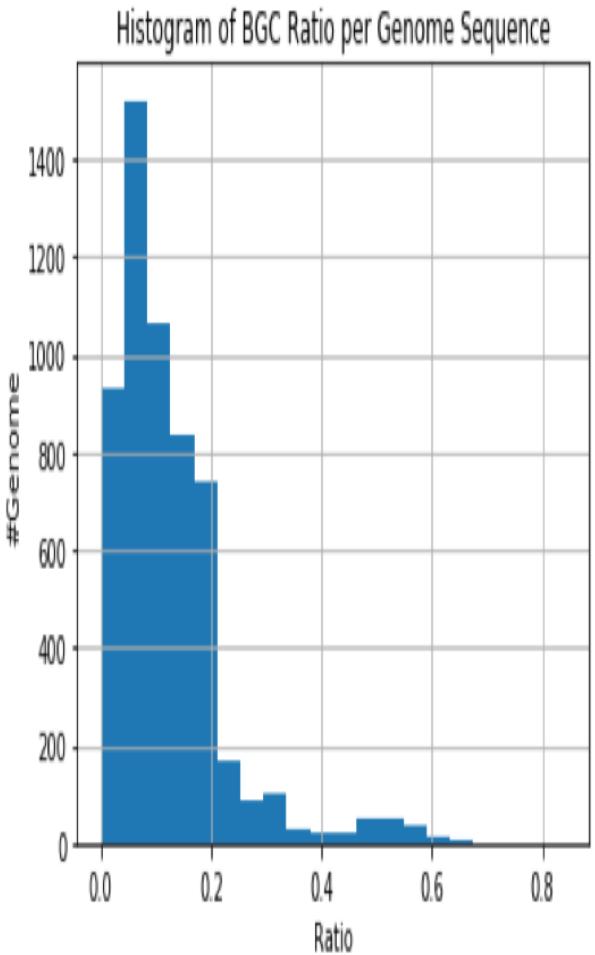
All BGCs predicted by DeepMBGC

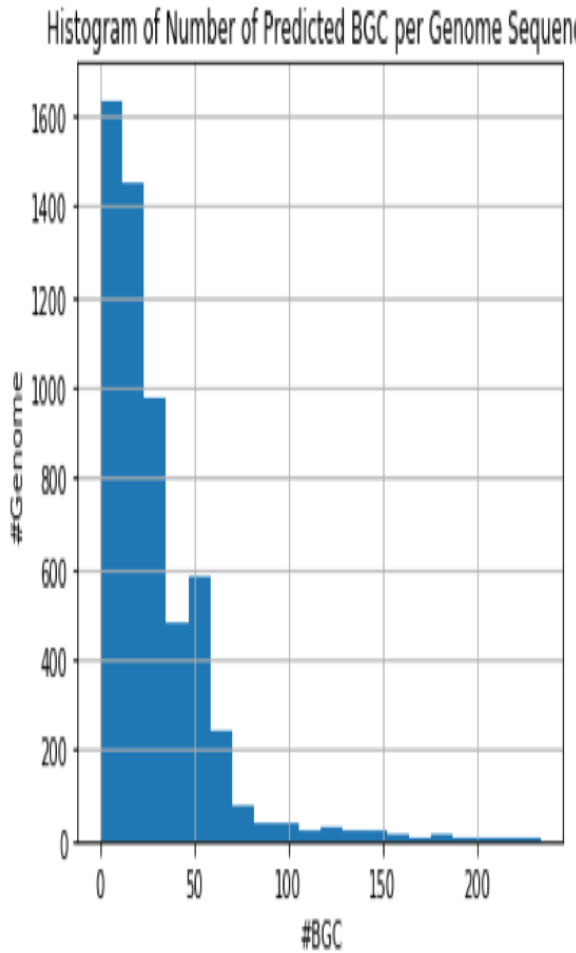
There are 161,026 predicted BGCs in all 5666 bacteria genomes.

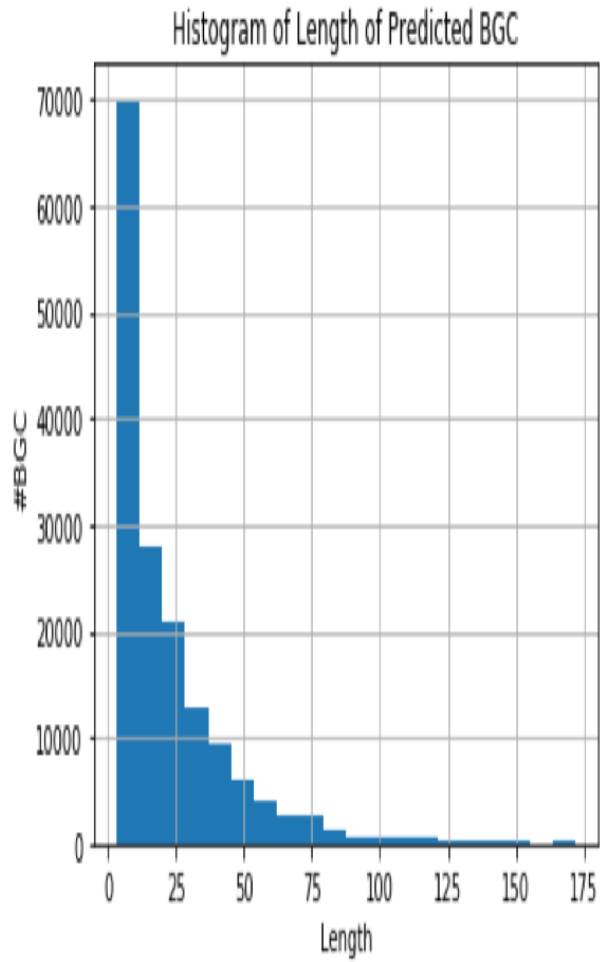
RiPP	41%
Non-ribosomal peptides (NRPs)	12.5%
Polyketide (PKS)	9.8%
Saccharide	9.7%,
Terpene	4.8%
other	21.6%

RiPP: Ribosomally synthesized and post-translationally modified peptides. Conserved genomic arrangement of many genes.

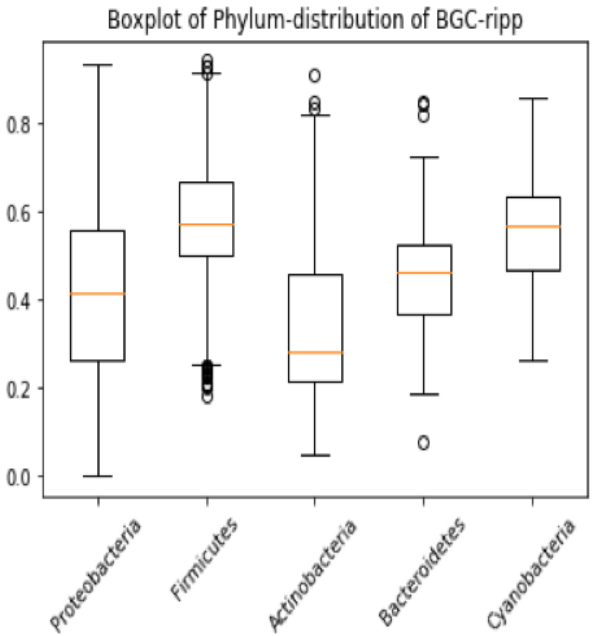
All BGCs predicted by DeepMBGC

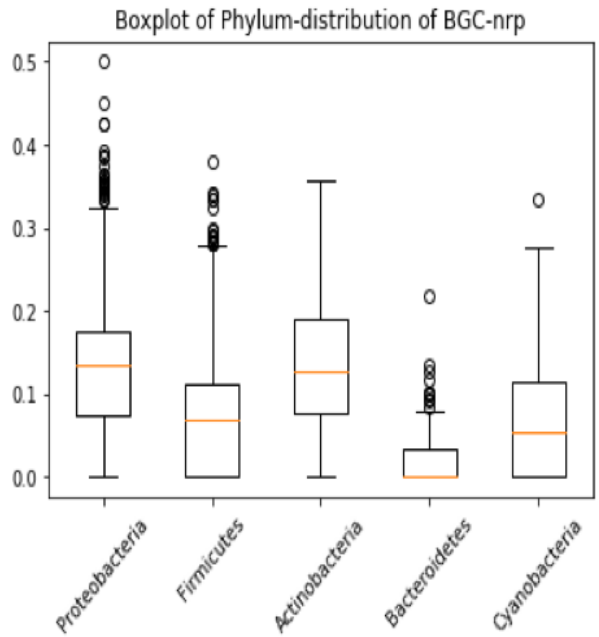


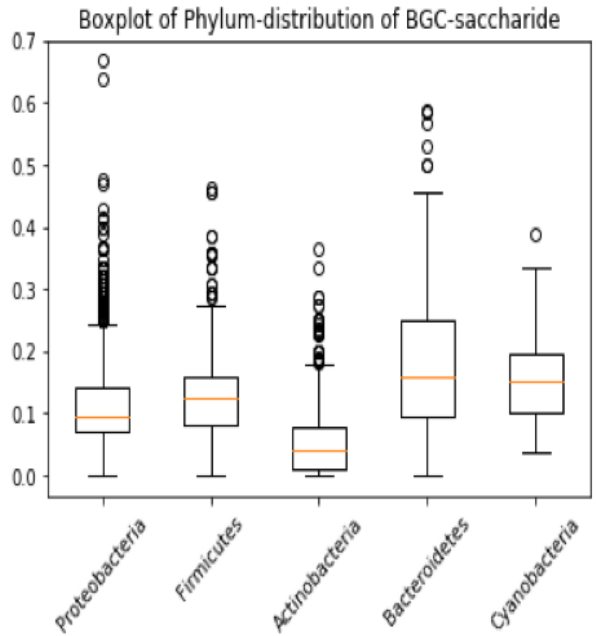


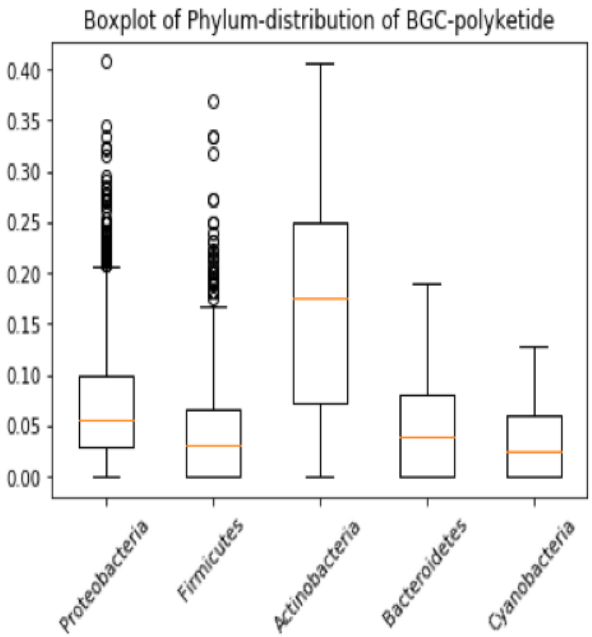


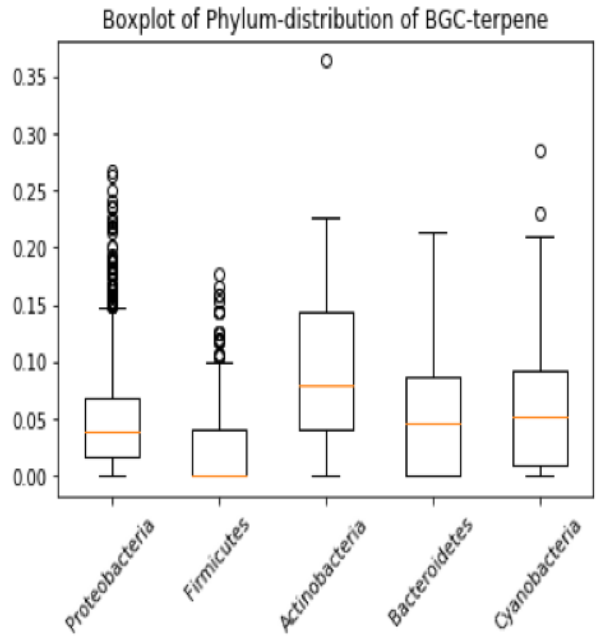
BGCs in Species Stratified by Phylum

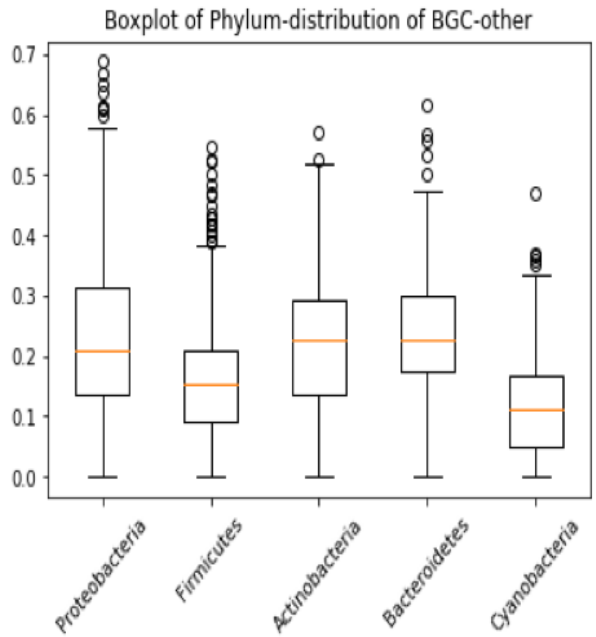












Summary of DeepMBGC

DeepMBGC

- deep learning for multi-class BGC discovery, better performance than DeepBGC (Hannigan et al., 2019 NAR)

- can make multi-class prediction
- database for BGCs coded by each species
- discovery of novel natural products

Acknowledgments

Many thanks to:

Li lab (NIH grants: R01GM129781; R01GM123056)

- Yuan Gao, Rong Ma
- Mingyang Liu and Yun Li

Tony Cai, PhD (The Wharton School)

Biology collaborators

- Gary Wu, MD (Gastroenterology)
- Rick Bushman, PhD (Microbiology)
- James Lewis, MD (Gastroenterology and DBEI)
- People in their labs