Illuminate dark matters in microbial communities by learning from massive genomic data

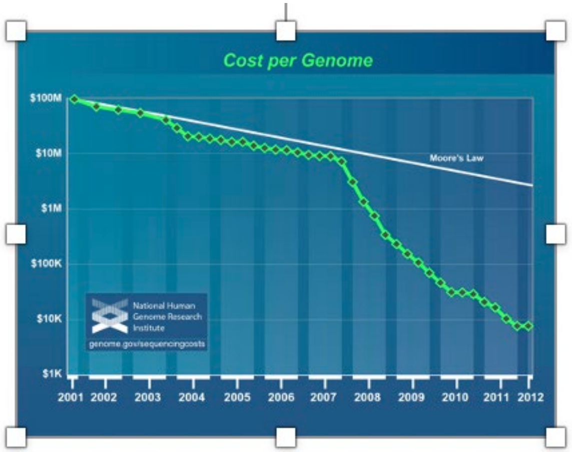
Dr. Yanni Sun, Electrical Engineering Department

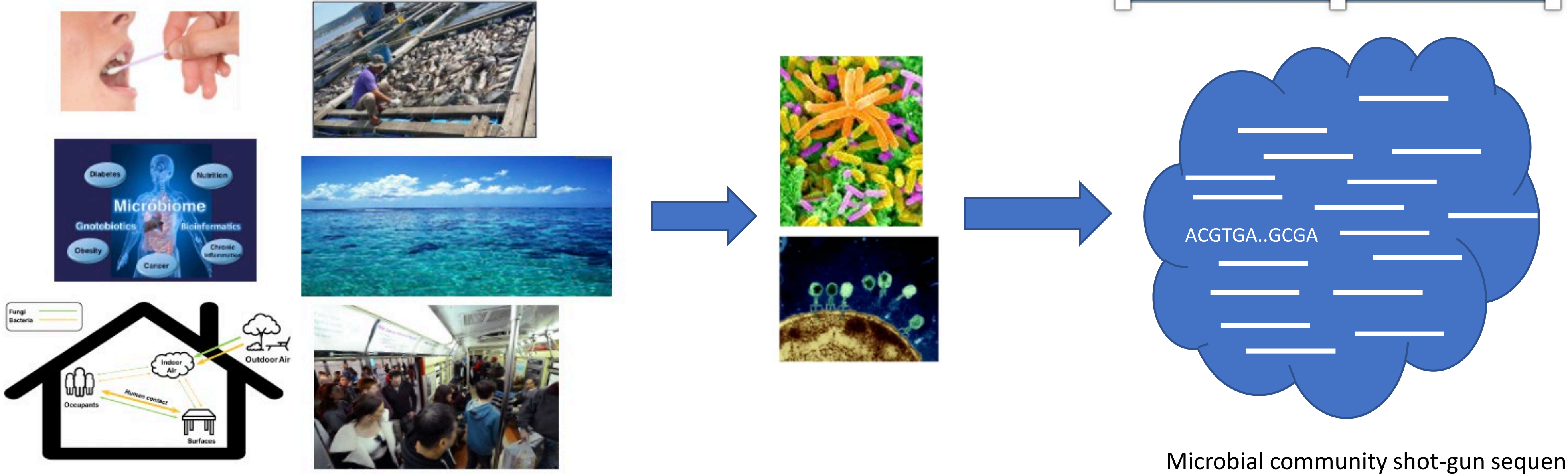
Dr. Patrick Lee, School of Energy and Environment

香港城市大學
City University of Hong Kong

Microbial communities and big genomic data

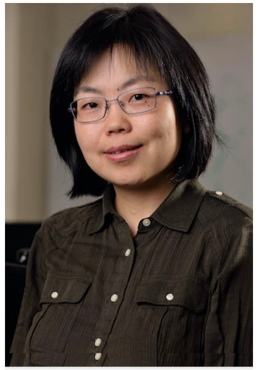
- Microbes: bacteria, viruses, etc.
  - Highly abundant and adaptive
  - Highly important to the earth and mankind
- Microbial communities: a group of microbes sharing a common space





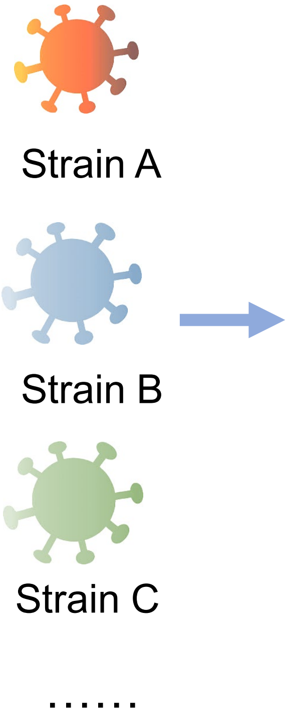
Microbial community shot-gun sequencing: metagenomic sequencing

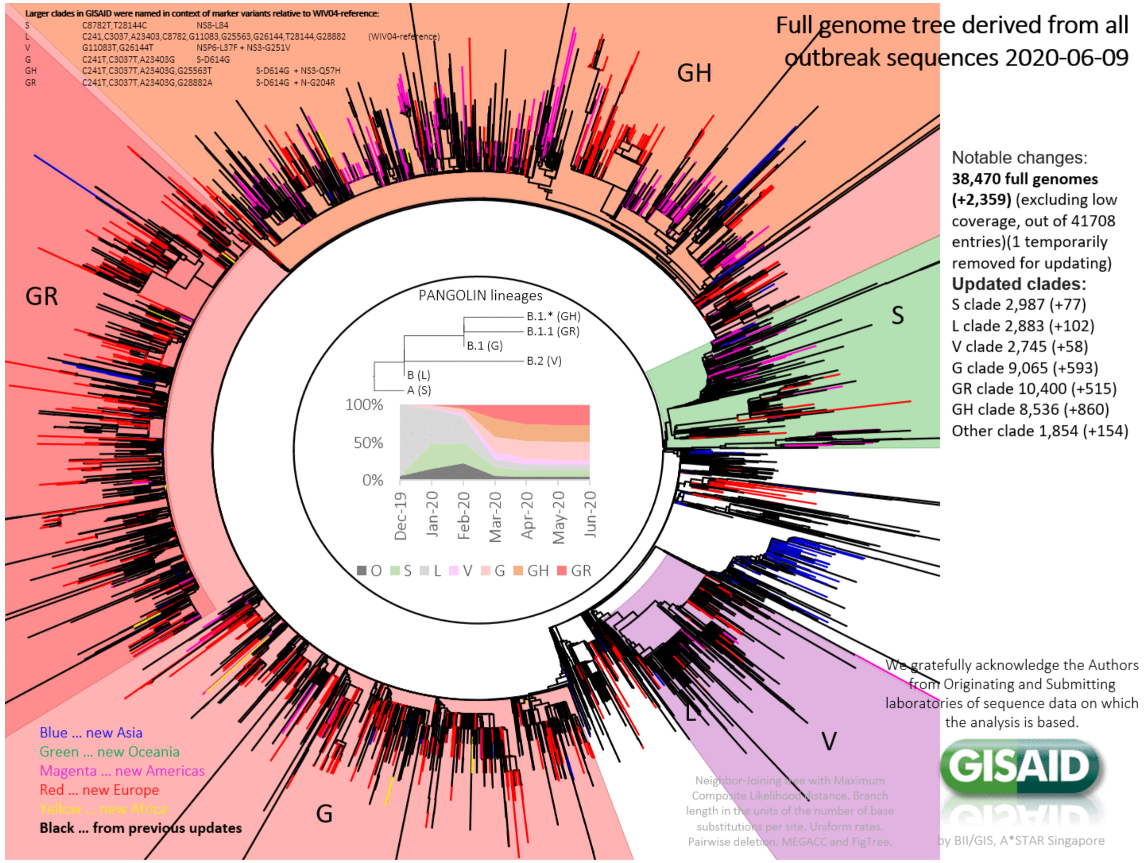
Dr. Yanni Sun, Electrical Engineering



Analyze viral metagenomic data using machine learning models

- **Who are there**?
  - Composition analysis: known and new viruses and their strain composition
- Answering these questions are important for characterizing viruses, diagnosis, vaccine design, and drug design





Full genome tree derived from all outbreak sequences 2020-06-09

Larger clades in GISAID were named in context of marker variants relative to WIV04-reference:

S	C8782T,T28144C	NS8-L84	
L	C241,C3037,A23403,C8782,G11083,G25563,G26144,T28144,G28882		(WIV04-reference)
V	G11083T,G26144T	NSP6-L37F + NS3-G251V	
G	C241T,C3037T,A23403G	S-D614G	
GH	C241T,C3037T,A23403G,G25563T	S-D614G + NS3-Q57H	
GR	C241T,C3037T,A23403G,G28882A	S-D614G + N-G204R	

Notable changes:
38,470 full genomes (+2,359) (excluding low coverage, out of 41708 entries)(1 temporarily removed for updating)
Updated clades:
S clade 2,987 (+77)
L clade 2,883 (+102)
V clade 2,745 (+58)
G clade 9,065 (+593)
GR clade 10,400 (+515)
GH clade 8,536 (+860)
Other clade 1,854 (+154)

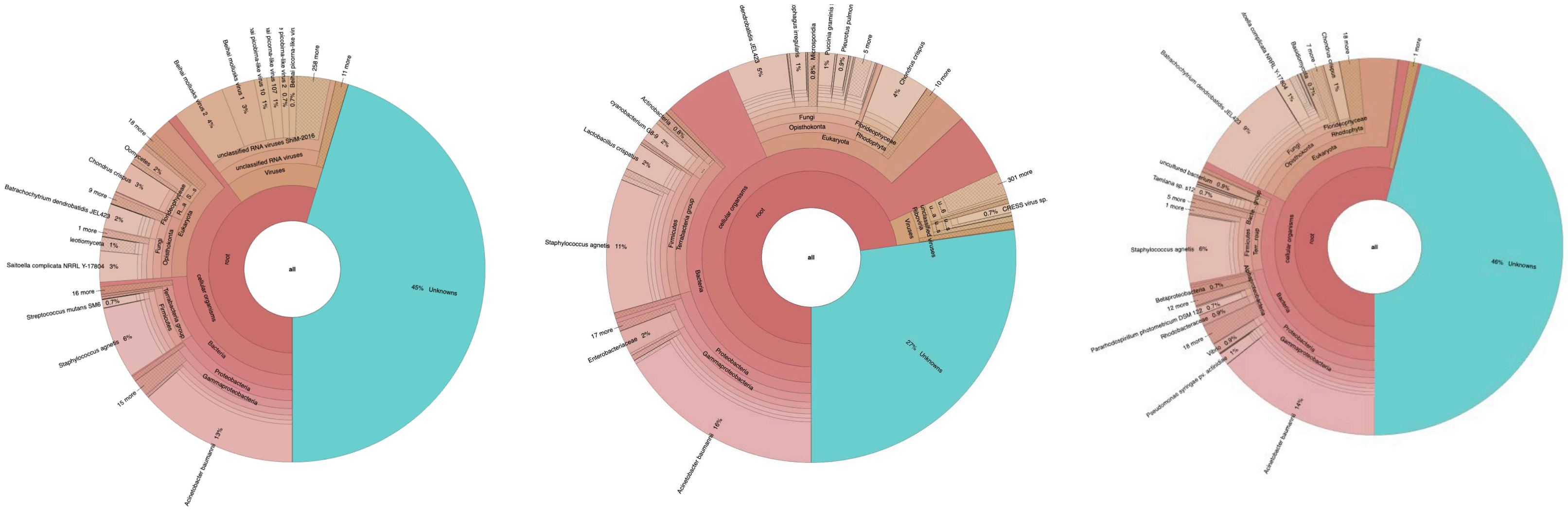
Blue ... new Asia
Green ... new Oceania
Magenta ... new Americas
Red ... new Europe
Yellow ... new Africa
Black ... from previous updates

Neighbor-Joining tree with Maximum Composite Likelihood distance. Branch length in the units of the number of base substitutions per site. Uniform rates. Pairwise deletion. MEGACC and FigTree.

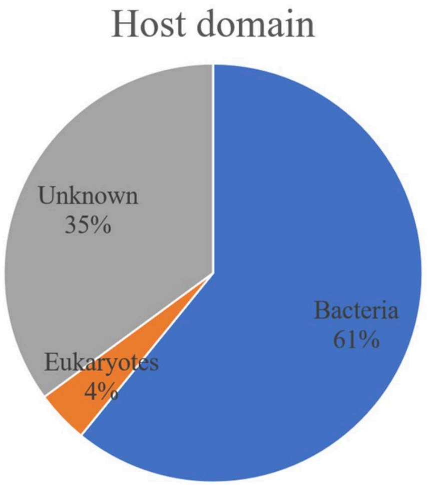
We gratefully acknowledge the Authors from Originating and Submitting laboratories of sequence data on which the analysis is based.

GISAID by BII/GIS, A*STAR Singapore

Knowledge gap and solution



Metagenomic Analysis of the Diversity of DNA Viruses in the Surface and Deep Sea of the South China Sea by Zhang et al. 2019



Limitations of existing tools

- Limited reference genomes (viral dark matter)
- Diverge viral genomes (marginal sequence similarity)
- Massive amount of reads (alignment-based tools are slow)

Solution: deep learning

- Automatic feature learning
- Mining degenerate patterns
- GPU for speedup

Challenges of applying deep learning to virus classification

- Highly imbalanced virus classes
- Need to reject non-viral reads (an open-set problem)

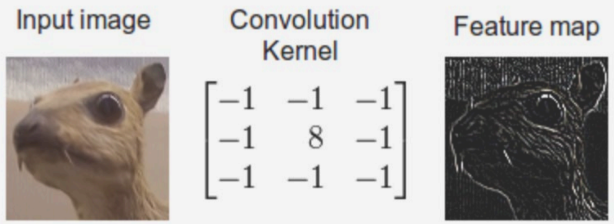
- Major goal: classify reads from new species
  - Species masking: the test species are not seen in the training data (i.e. new species)
  - Analogy: training using images of sedans, vans, SUVs, convertibles; testing with trucks

Motifs and convolution filters in CNN

- Conserved sequence patterns:
  - important features for genomic sequence classification
  - Can be represented by convolution filters

One hot encoding:

$$\text{AACGG} \rightarrow \begin{array}{c|ccccc} A & 1 & 1 & 0 & 0 & 0 \\ C & 0 & 0 & 1 & 0 & 0 \\ G & 0 & 0 & 0 & 1 & 1 \\ T & 0 & 0 & 0 & 0 & 0 \end{array}$$



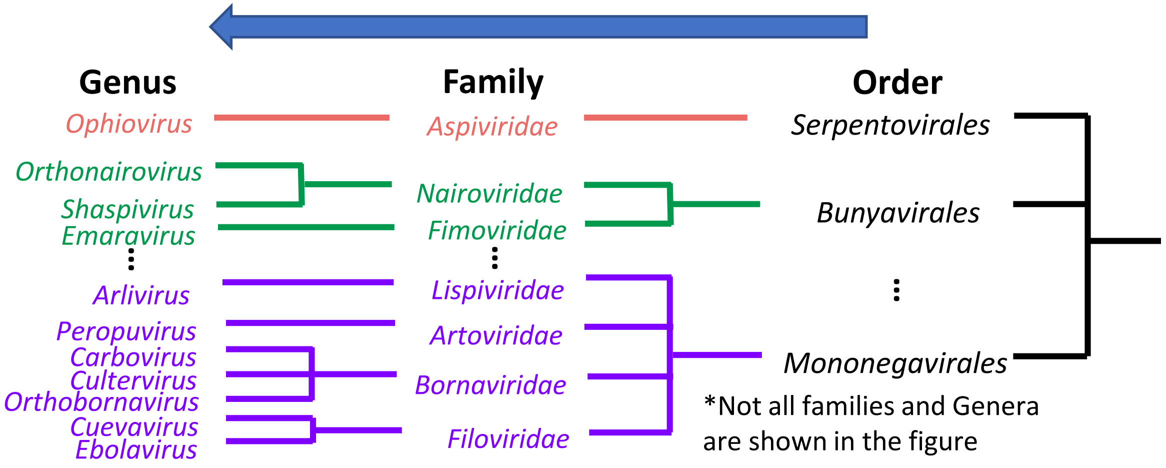
$$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

Edge detector kernel (https://developer.nvidia.com/discover/convolution)

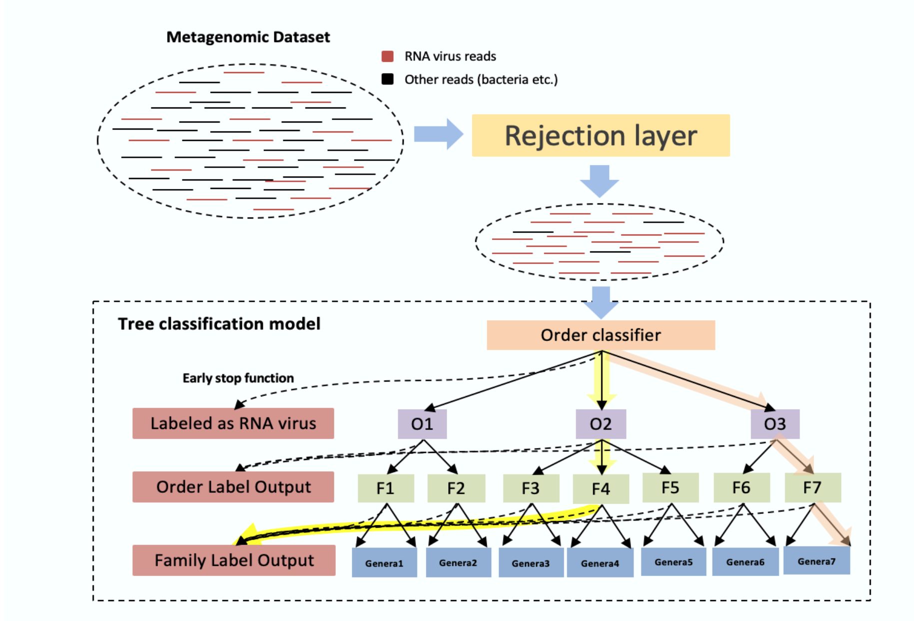
$$f = \frac{1}{5}\begin{bmatrix} 3&4&5&4&2&5&5&4&0&0 \\ 2&0&0&0&0&0&0&0&1&0 \\ 0&0&0&0&1&0&0&0&4&5 \\ 0&1&0&1&2&0&0&1&0&0 \end{bmatrix}\begin{matrix} A \\ C \\ G \\ T \end{matrix}$$

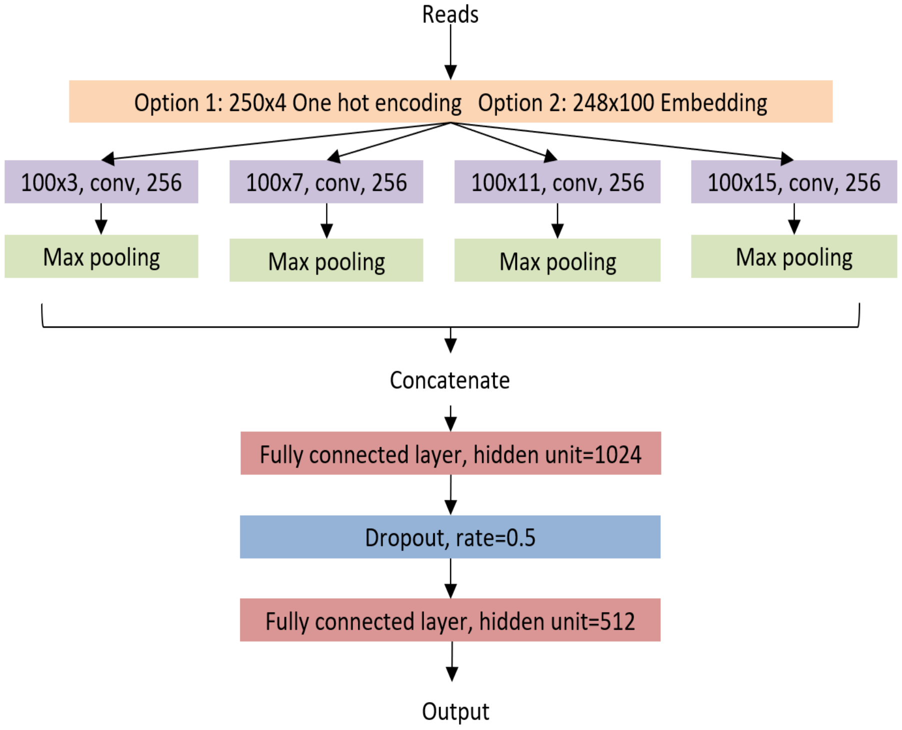
Zia, Amin & Moses, Alan. (2012). Towards a theoretical understanding of false positives in DNA motif finding. BMC bioinformatics

CHEER: hierarCHical taxonomic classification for viral mEtagEnomic data via deep learning



*Not all families and Genera are shown in the figure





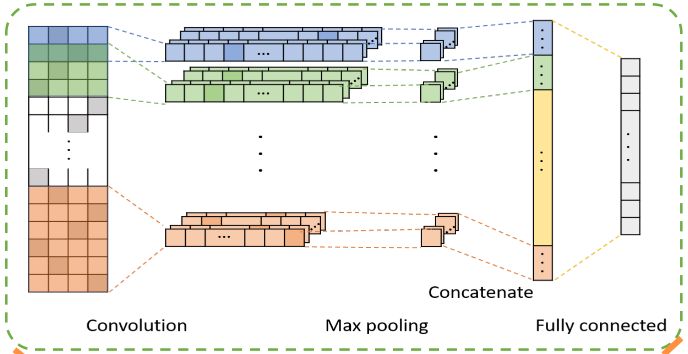
Architecture of a CNN model

Open set problem: outlier exposure

Labeled data from different orders

Highly similar protein sequences as **outlier exposure (OE)**

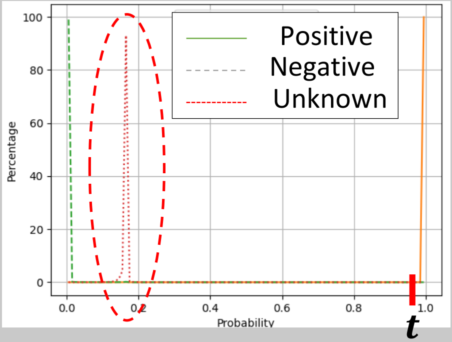
Training dataset



CNN classification model

Model

Softmax probability distribution



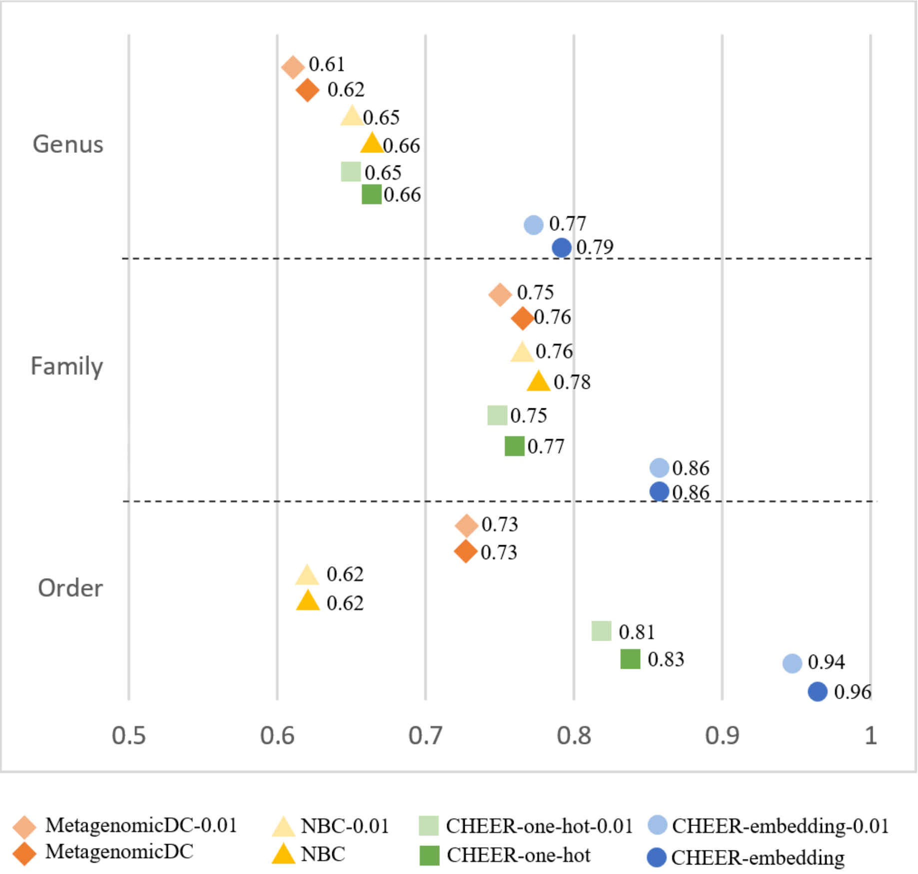
Set a softmax threshold $\boldsymbol{t}$ to reject non-RdRp reads

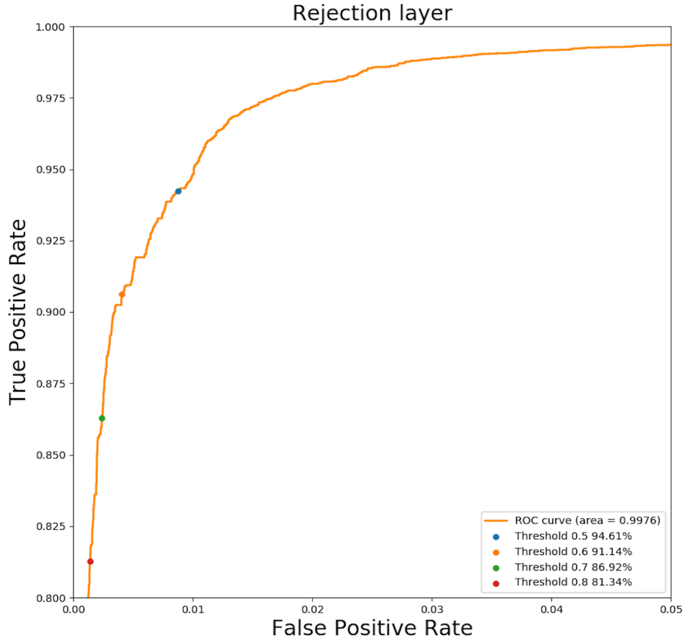
RdRp reads, if $p(i) \geq \boldsymbol{t}$

Non-RpRd reads, if $p(i) < \boldsymbol{t}$

Add $f_{loss}(\boldsymbol{oe})$ to the original loss function to enhance the ability of detecting non-RNA viruses. After training, the softmax probabilities of unknown samples tend to have uniform distributions.

Experimental results





Performance of rejection layer

Method	Time	Accuracy
CHEER	00:00:42	87.9%
BLASTx (alignment-based models)	00:46:39	74%

GPU acceleration

Prof. Dirk Pfeiffer, Department of Infectious Diseases and Public Health

Younjung Kim

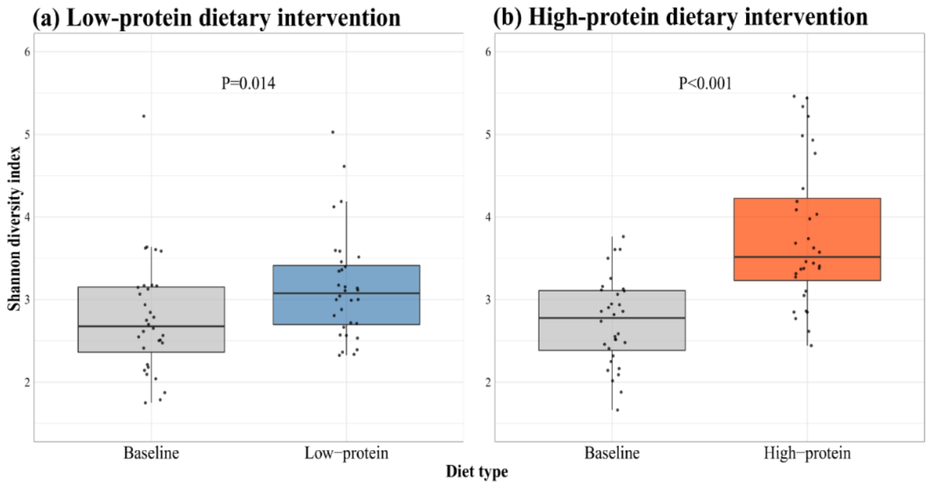
Antibiotic resistance genes in the canine gut microbiota

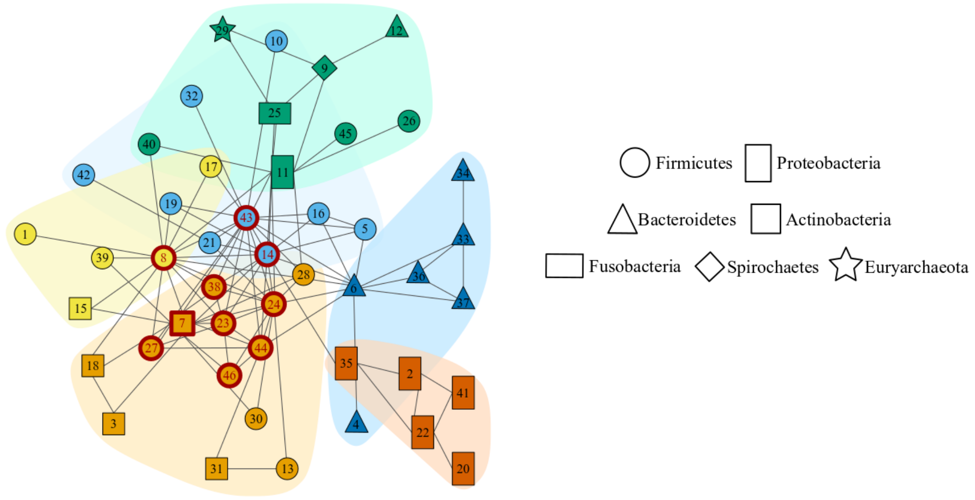
Objectives

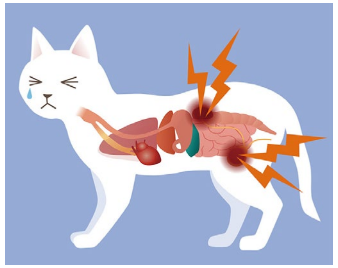
- Assess the **association between dietary protein content and diversity/composition of ARGs in the canine gut microbiota**
- Describe ARG-sharing relationships among canine gut bacteria using a novel network approach

Major findings

- Dietary protein content plays a significant role in determining the antibiotic resistance potential of the gut microbiota.
- The ARG-sharing network has general properties of social networks, with taxonomy being a barrier to the lateral transfer of ARGs.







Microbiome in a cat body site once believed to be sterile?

Objectives

- Characterise the diversity and composition of a **urinary microbial community (urobiome) in animals**
- Assess the association of the **urobiome** with different urinary tract diseases in animals

Pilot findings from cats

- The urinary tract indeed harbors a unique microbial community.
- Bacterial that were not cultured under standard conditions were identified by 16S rRNA sequencing.
- The main project is investigating the role of urinary microbiome with the **major urinary tract diseases** in cats.

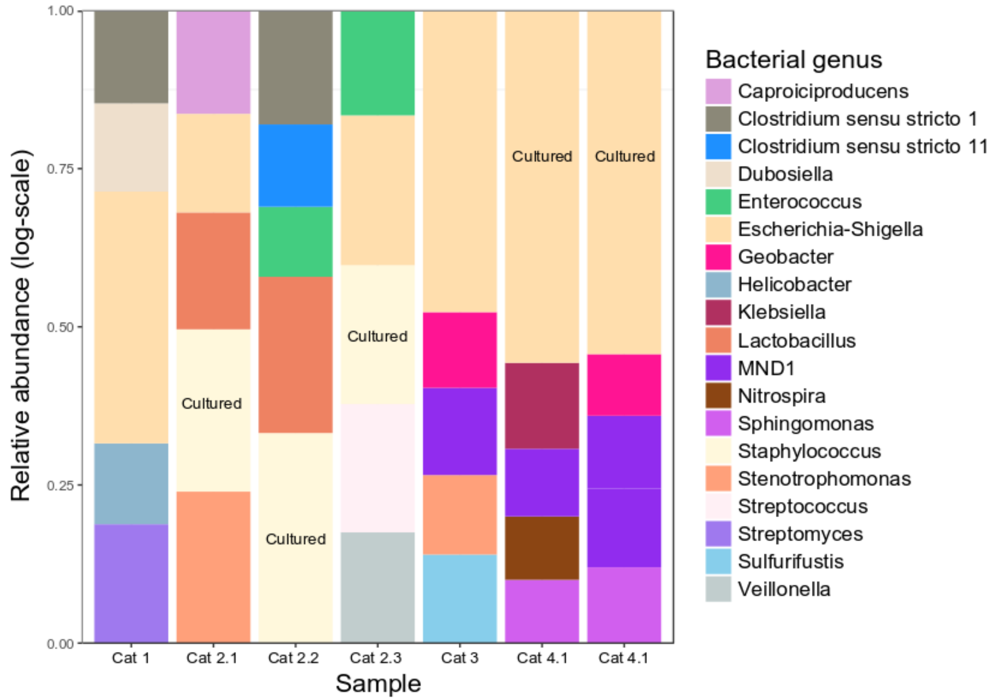
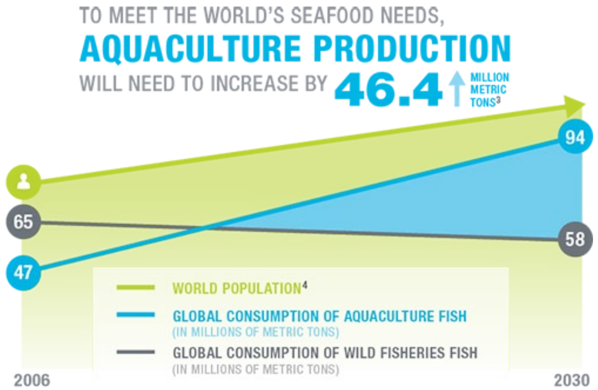


Figure 1. Bacterial genera in cat urine identified by culture-independent methods. Only the five most abundant genera are shown for each urine sample.

Dr. Giana Gomes, Department of Infectious Diseases and Public Health

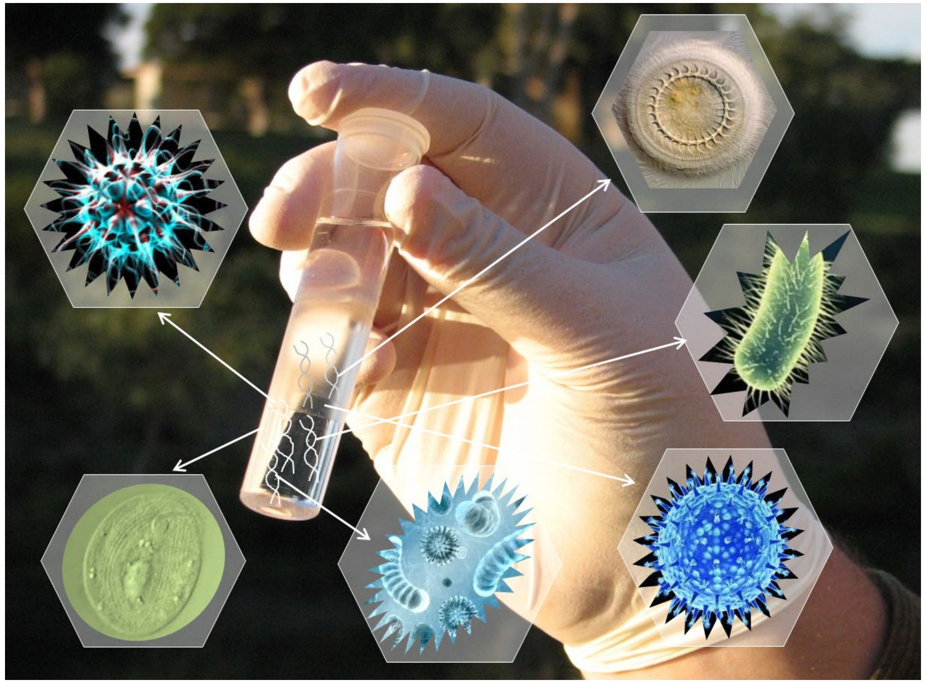
Underpinning the Use of Environmental DNA (eDNA) to Prevent Diseases in Aquaculture



But, up to 40% aquaculture production is lost due to diseases
(USD 6 billion/year)

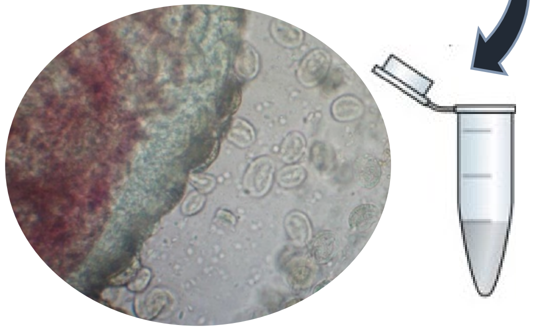
SOLUTION?

USE ENVIRONMENTAL DNA (eDNA) TO EARLY DETECT PATHOGENS IN FISH FARM WATER, BEFORE ANIMALS GET SICK!



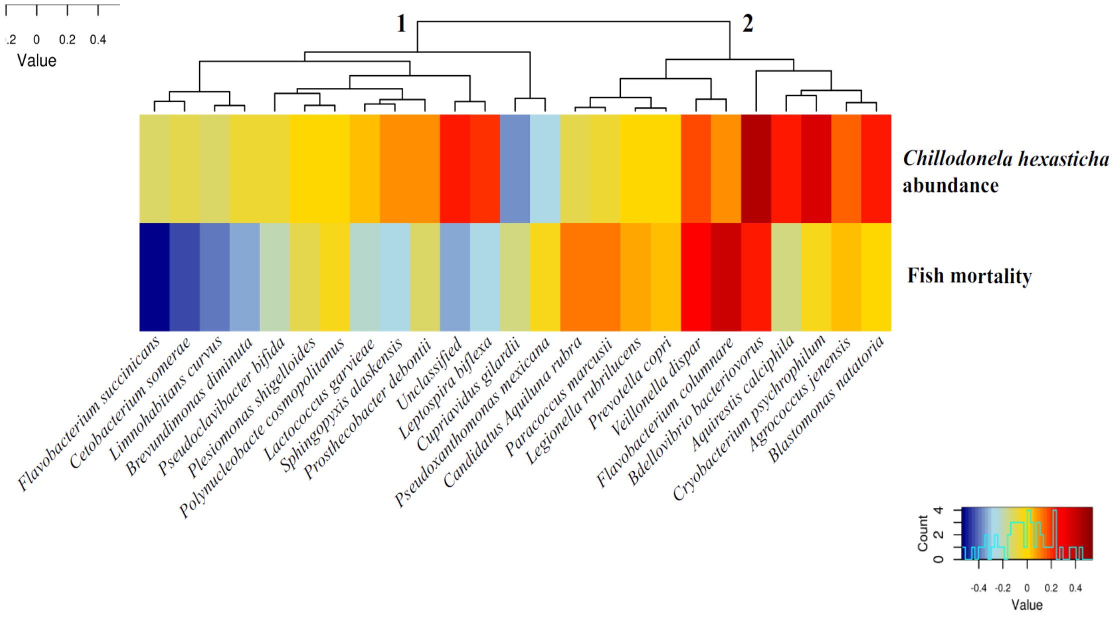
Investigation of Bacterial Microbiome and Parasite Interaction to Prevent Diseases in Aquaculture

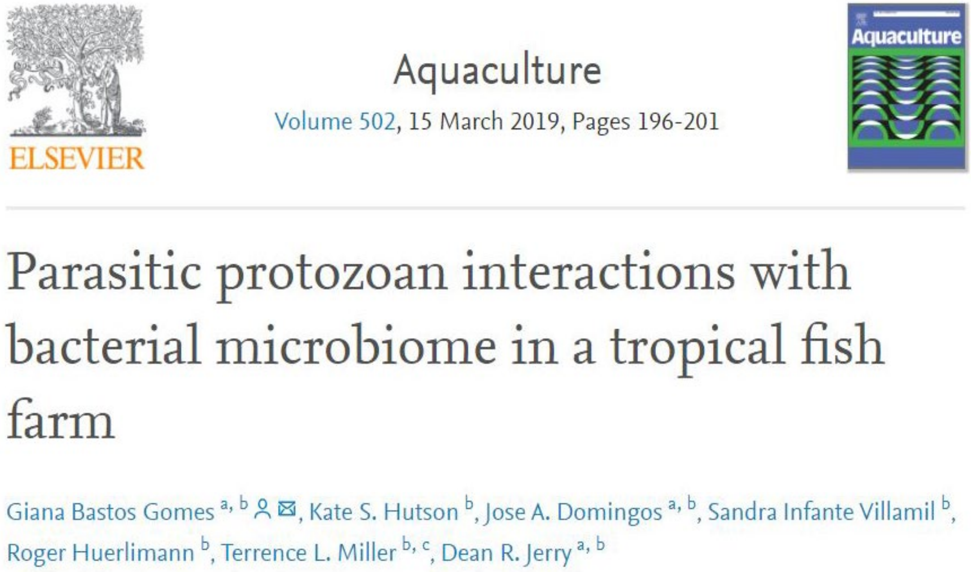
eDNA methodology



20 gills infected/ 18 non-infected by *Chilodonella*

Illumina Miseq Sequencing System





ELSEVIER

Aquaculture

Volume 502, 15 March 2019, Pages 196-201

Aquaculture

Parasitic protozoan interactions with bacterial microbiome in a tropical fish farm

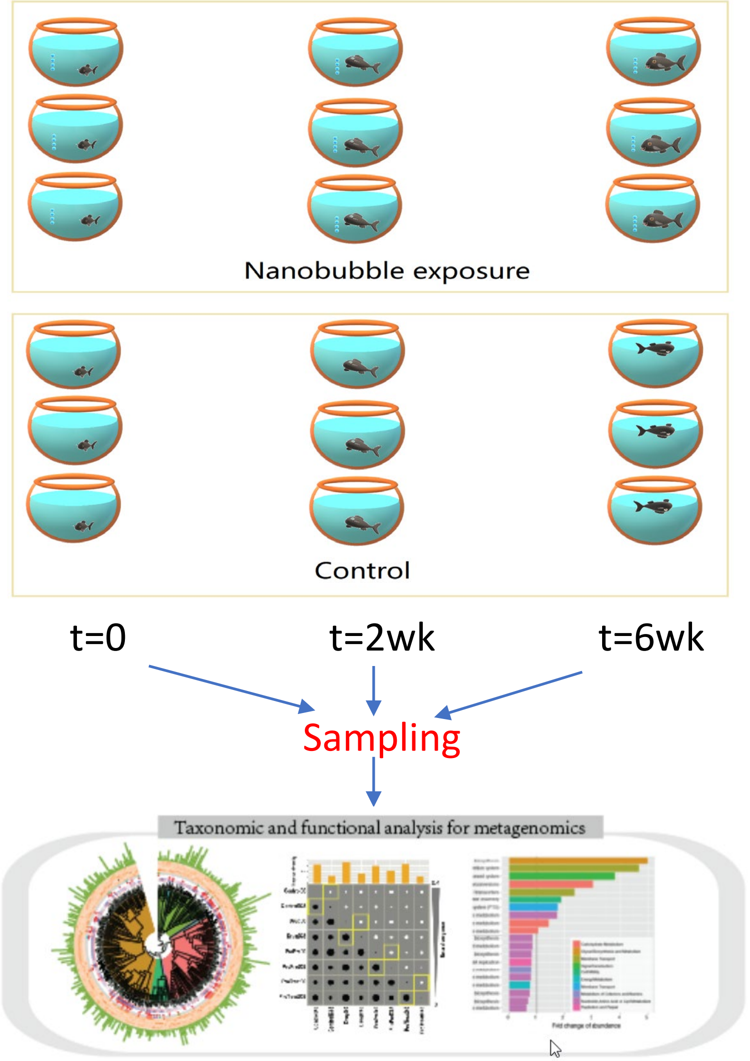
Giana Bastos Gomes [a, b], Kate S. Hutson [b], Jose A. Domingos [a, b], Sandra Infante Villamil [b], Roger Huerlimann [b], Terrence L. Miller [b, c], Dean R. Jerry [a, b]

Prof. Sophie St-Hilaire & Dr. Jun Li
Department of Infectious Diseases and Public Health

Innovative water treatment using ozone nanobubbles and copper carbon nanoparticles

Objective: Reduce waterborne bacterial pathogens in aquaculture

- Measure impact of disinfection strategies on bacterial gut microbiome and environmental bacterial community



Nanobubble exposure

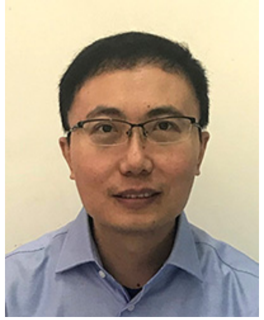
Control

t=0 t=2wk t=6wk

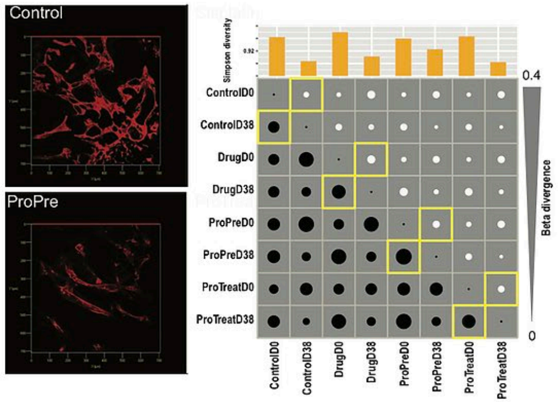
Sampling

Assess **gut microbiome of fish** and microbial environment across time for ozone nanobubble treated and untreated ponds

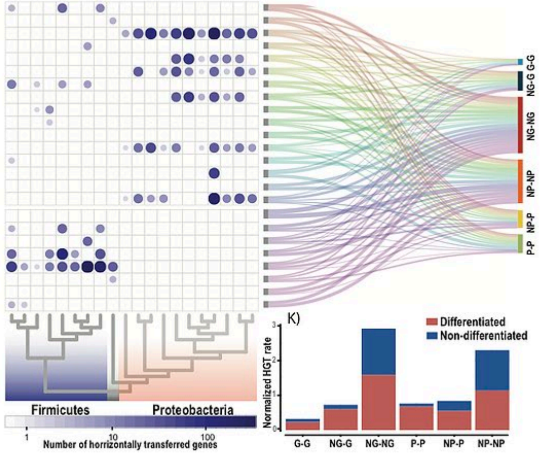
Dr. Jun Li
Department of Infectious Diseases and Public Health



- Antibiotic resistance and pathogenesis in human, animal and environments
- Microbial interactions and dynamics
- Disease-related microbiome and interventive studies
- Integrative data mining and Precision livestock farming



Health/Disease Related Microbiome



Integrative Biology for Pathogenesis and Antibiotic Resistance

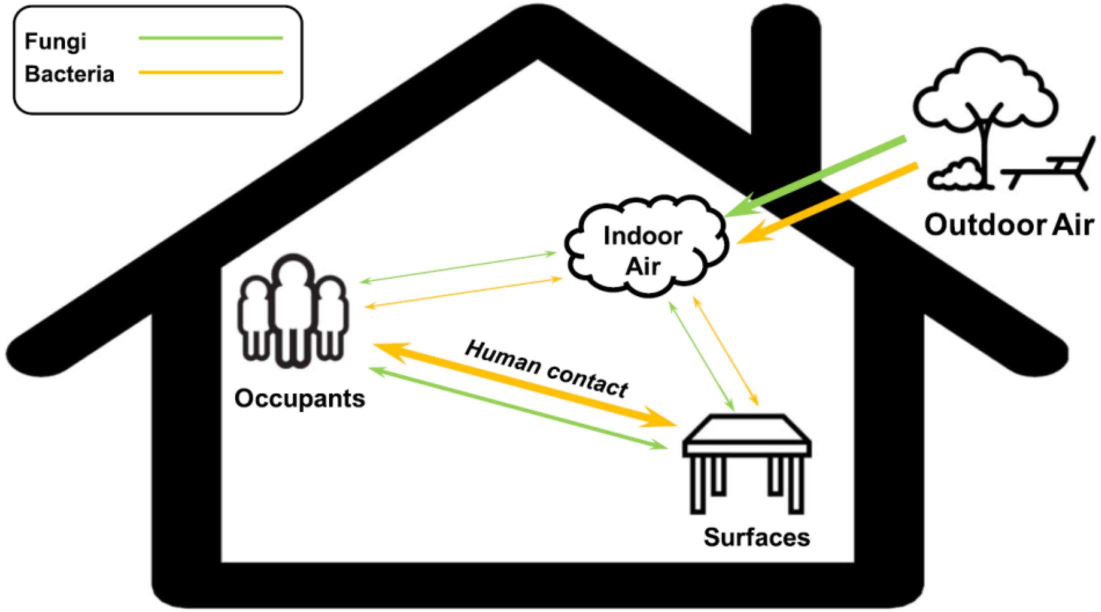
Dr. Patrick Lee, School of Energy and Environment

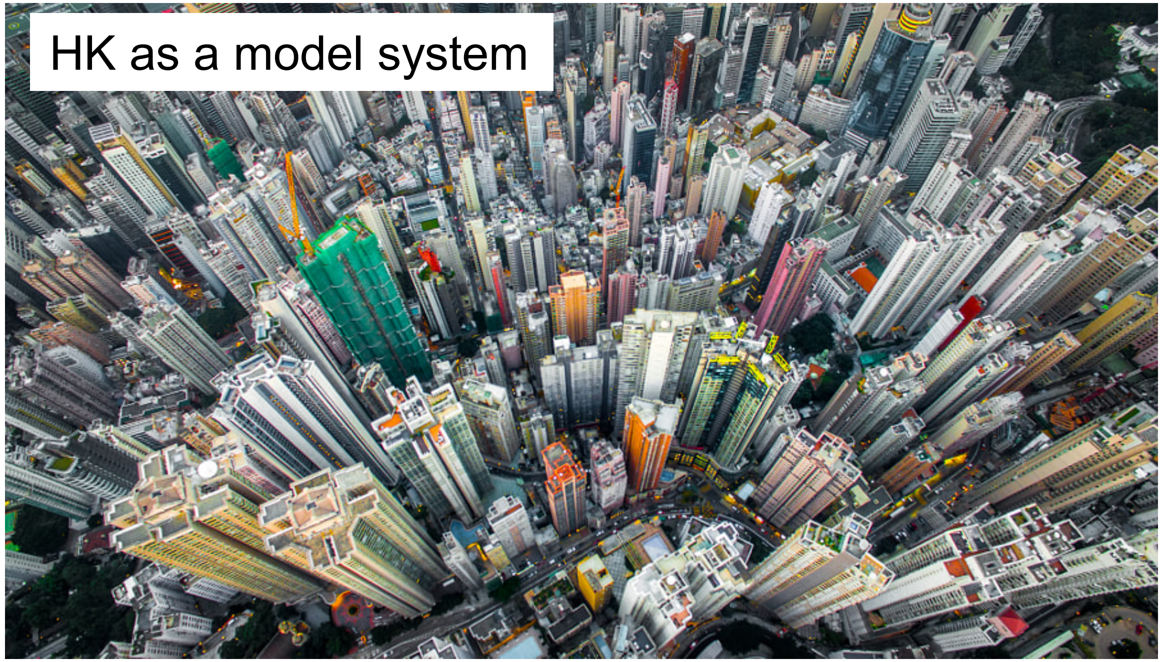
Urban & Built Environment Microbiome

- We spend ~90% of our time indoors and surrounded by a “cloud” of microbes
- In humans, ~1.3 bacterial cells for every 1 human cell (Nature Medicine 2018)

We are interested to understand:

(i) The composition of the built environment microbiome (Air & Surface)
(ii) The interactions and transfer between human and built environment
(iii) The novel biochemistry that we can harness from metagenomes (e.g., antibiotics)



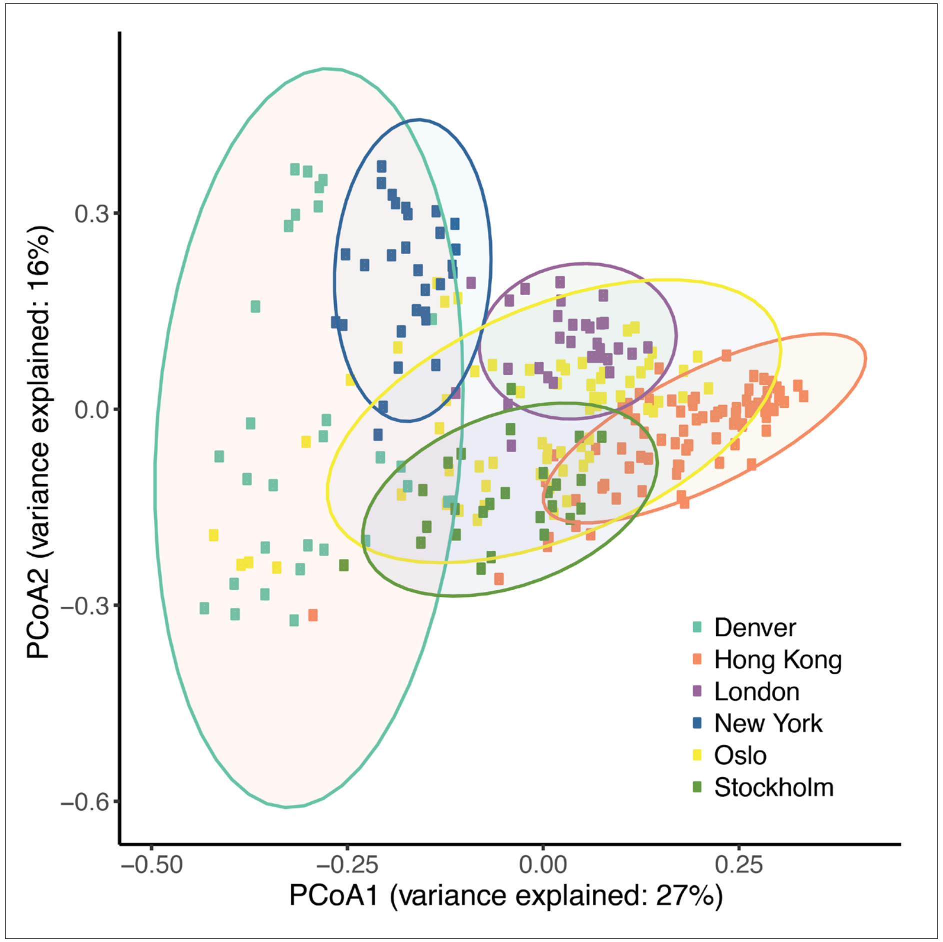


HK as a model system

Mapping the Global Urban Microbiome

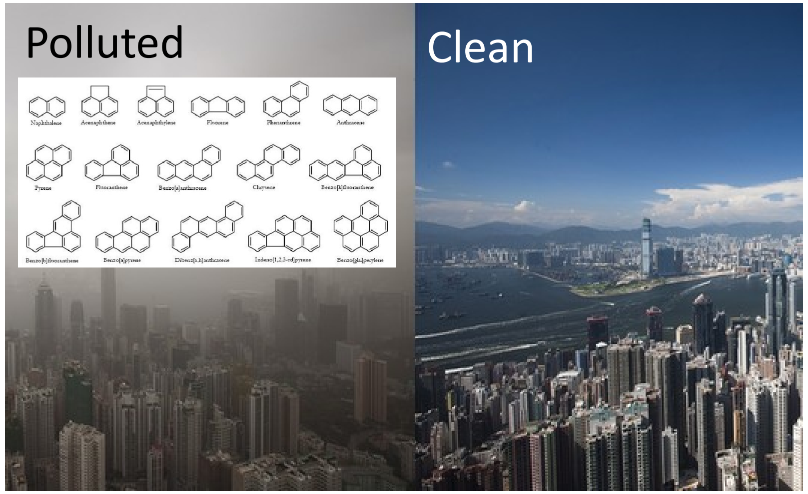
An international project striving to understand the metagenomes in the urban environment

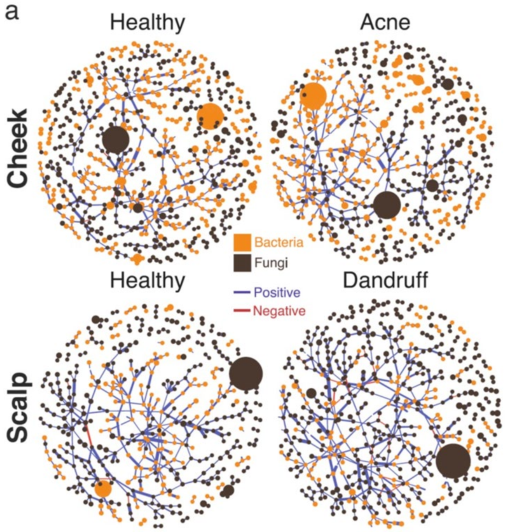
A 5-year longitudinal project

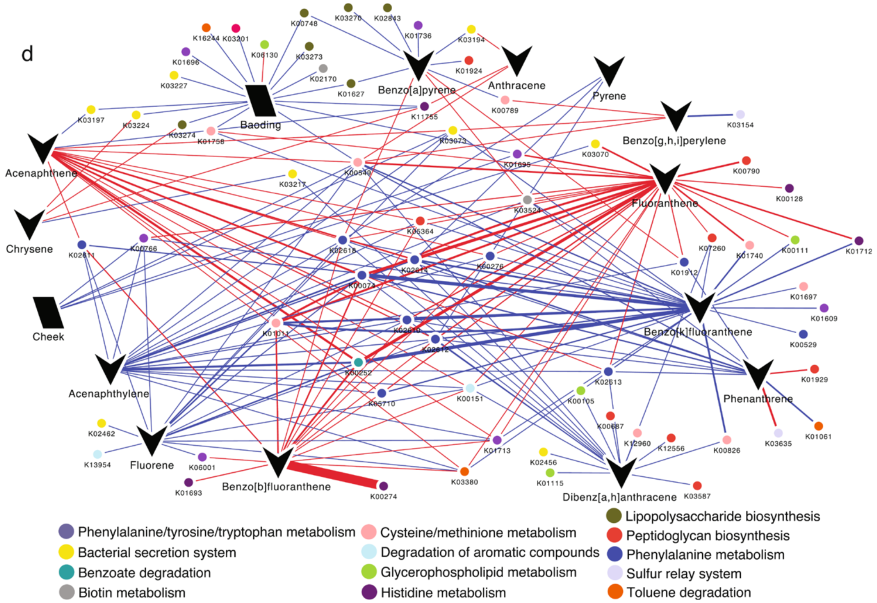


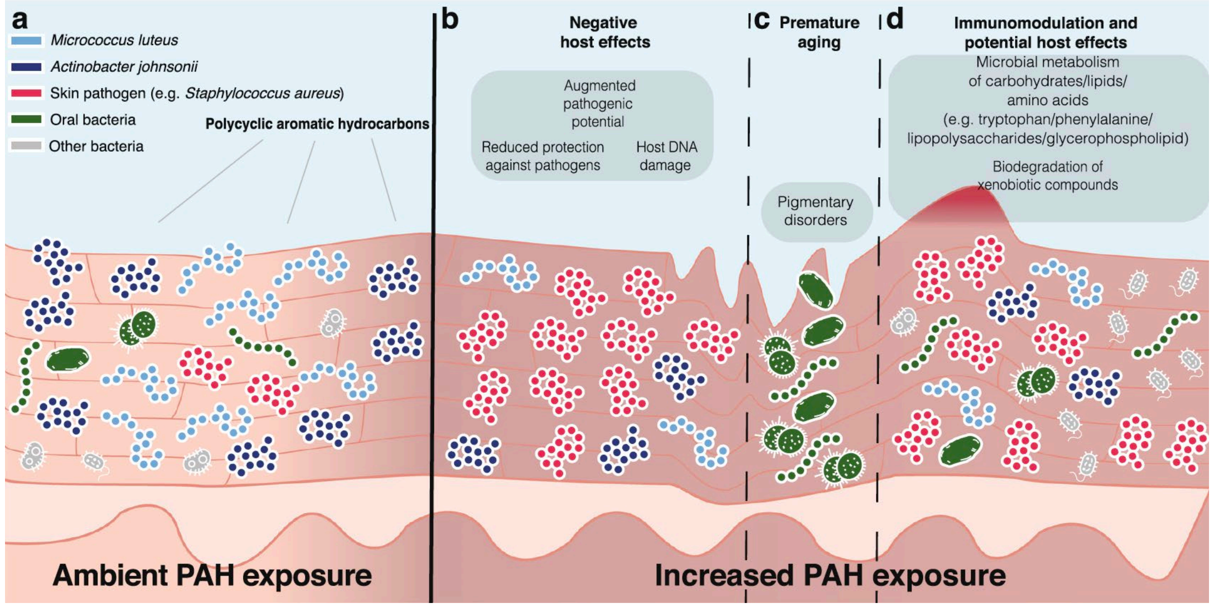
Air Pollution and Skin Microbiome

Skins (and the bacteria on it) are our external barrier against air pollutants









One Health with BIG genomic data…

CityU TODAY, Jan 2018

Acknowledgement

CityU

Hong Kong Institute for Data Science

香港城市大學
City University of Hong Kong