UC RIVERSIDE | Institute for Integrative Genome Biology

Proteomics Core

# Robust, scalable virus-host proteomics with AI-generative library of life kingdoms and iterative search of zoological space

Mingqi Liu[1], Duo Xu[1], **Quanqing Zhang**[2]*, Rong Hai[1]*

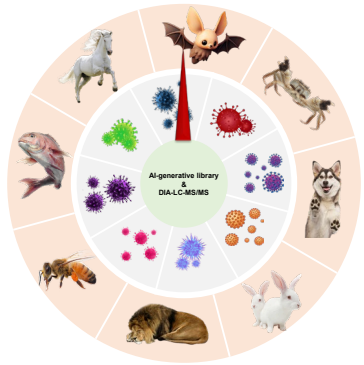
1. Department of Microbiology and Plant-pathology, University of California, Riverside, Riverside, CA, United States
2. Institute for Integrative Genome Biology, Proteomics Core, University of California, Riverside, Riverside, CA, United States

## Abstract

The surveillance of environmental viral pathogens is crucial for public health.[1,2] Ideally, the detection information should include both virus's identity and the host. However, current nucleic acid-based detection is better suited for identifying the former. An innovative approach that combines advanced liquid chromatography-mass spectrometry (MS) proteomics with artificial intelligence (AI) and iterative search was developed. AI/MS-proteomics identified diverse viruses in environmental samples and revealed the correct host information for exotic samples such as bat-origin virions. This framework of proteome-based surveillance presents a potent tool for studying viral dynamics, informing public health responses, enhancing ecological understanding, and mitigating emerging viral threats.

## Introduction

The presence of viruses in waste water has been shown to precede their prevalence in humans. The surveillance of viral pathogens in waste water, therefore, becomes a critical need in public health, particularly for proactively responding to emerging zoonotic viruses. Monitoring viral pathogens in waste water allows for early detection and identification of potentially harmful viruses before they can cause widespread outbreaks. Ideally, waste water surveillance requires monitoring a population's qualitative and quantitative health status within a specific area by detecting the concentration of target chemical or biological markers in waste water, in conjunction with human metabolism, water inflow volume, and population information. We propose to develop a tandem hydrophilic size-exclusion chromatography system for concentrating viruses in residential environment waste water and a protein-based detection that combines advanced liquid chromatography-mass spectrometry (LC-MS) proteomics with artificial intelligence (AI).




## Results

**Robust, scalable AI-generative library of proteome of life kingdoms**

We developed a scalable AI-generative library pipeline capable of handling proteomes of any size, achieving over a 20-fold performance boost across ultra-large search spaces (Figure 1a-1e). This pipeline leverages open-source deep learning models to generate in-silico LC-MS libraries from proteome databases, enabling efficient identification and quantification of DIA-LC-MS data. To enhance adaptability, we applied transfer learning to fine-tune models for various LC-MS signal types. Notably, the pipeline maintained performance across different samples and instruments, with no signs of overfitting. DIA-based transfer learning consistently outperformed DDA-based approaches in our datasets. We benchmarked our pipeline against conventional sample-specific methods using public datasets.

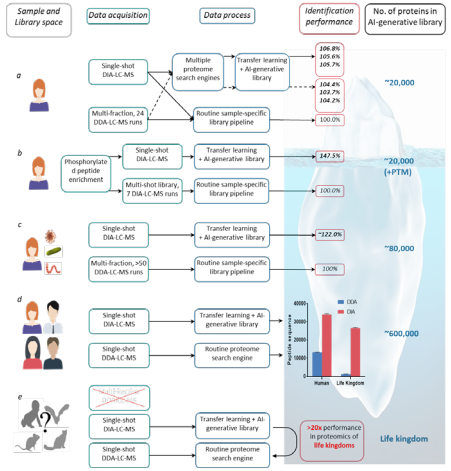



Figure 1. Robust, scalable AI-generative library of proteome of life kingdoms. (A) Performance evaluation of AI-generative library with routine sample-specific library pipeline on public proteomic dataset. (B) Performance evaluation of AI-generative library with routine sample-specific library pipeline on public phosphoproteomic dataset. (C) Performance evaluation of AI-generative library with routine sample-specific library pipeline on public proteomic dataset with multiple species. (D) Performance comparison of DDA/DIA-LC-MS in different search spaces. (E) Necessity of AI-generative library in analysis of small sample volume with unknown species information.

**Framework of fast, sensitive LC-MS based virus-host proteomics**

We developed a fast and sensitive virus-host proteomics workflow that completes within half a day, combining optimized virion sample preparation, DIA-LC-MS, and an AI-generative library (Figure 2a-1c). Removing detergents improved digestion efficiency, and complex DIA samples enhanced transfer learning. Using a serial dilution of influenza A PR8 virions (1-512 ng), our pipeline identified strain-specific viral peptides and over 3,600 host proteins-exceeding DDA-based methods by over 10×. This scalable approach enables high-depth, accurate analysis across diverse proteomes with minimal sample input.

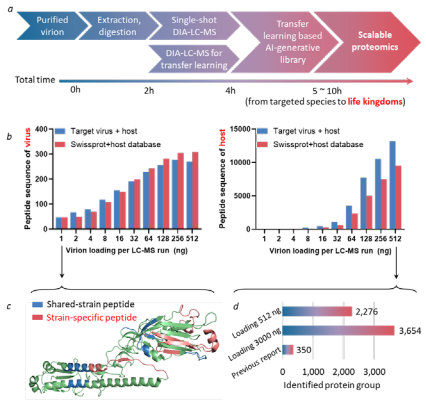



Figure 2. Fast, sensitive LC-MS based virus-host proteomics. (A) Workflow of virus-host proteomics with LC-MS and AI-generative library. (B) Peptide sequence identification of virus and host under different conditions. (C) HA protein coverage under 1 ng virion loading. (D) Performance comparison of existing virion proteomic study.

**Framework of fast, sensitive LC-MS based virus-host proteomics**

Building on the strength of our AI-generative library with large proteomes (Figure 2B), we developed a three-step iterative search framework for unbiased virus-host identification from unknown sources (Figure 3a). A direct one-step search across all known and unknown proteomes (>250 million proteins) is currently infeasible, so we implemented a decision-tree approach using DIA-LC-MS data. Step 1 screens all Swiss-Prot proteomes to narrow down top virus and host candidates. Step 2 refines identification using UniProt databases of these candidates. Step 3 confirms identities using focused virus-host proteomes.

We demonstrated this approach on PR8 virions and extended it to additional viruses (WSN, ZIKA) from various sources, successfully identifying virus-host pairs in each case (Figure 3b). Functional analysis of host proteomes revealed consistent biological themes-such as enrichment in protein synthesis and degradation pathways-regardless of the culturing system (Figure 3c), suggesting selective incorporation of host proteins into virions. In summary, our iterative, AI-assisted workflow enables accurate, scalable virus-host profiling from unknown samples, validated across multiple viruses and host systems.

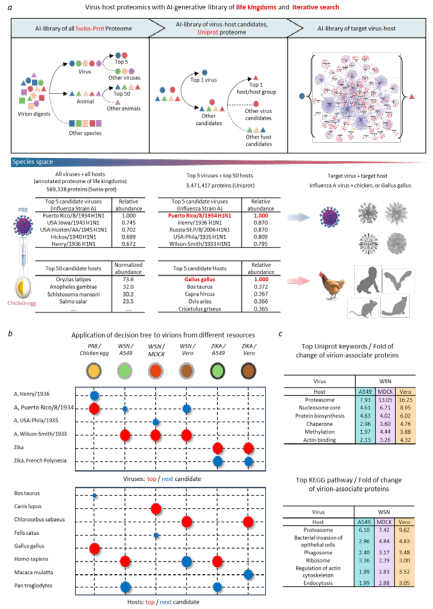



Figure 3. dedicated iterative search and AI-generative library reveals virus-host resources. (A) Workflow and example of iterative search for virus-host proteomics, in the form of decision-tree. (B) Application of proposed iterative search to virions from different resources. (C) Bioinformatic analysis of virion-associated proteins from different resources.

## Conclusions

In summary, we designed a dedicated experiment with comparison between double-spike in mixtures before and after virion enrichment processes. The mixture consists of a known quantity of virions and environmental bacteria. We observed dozens of folds of increase in host protein intensity after virion enrichment, confirming the source of the identified host proteins is the viral particle. Additionally, consistent host protein signals were observed in different virus-host combinations.

## References (if necessary)

1. Bruderer R, Bernhardt O M, Gandhi T, et al. Optimization of experimental parameters in data-independent mass spectrometry significantly increases depth and reproducibility of results[J]. Molecular & Cellular Proteomics, 2017, 16(12): 2296-2309.

2. Chi H, Liu C, Yang H, et al. Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine[J]. Nature biotechnology, 2018, 36(11): 1059-1061.

## UCR Proteomics Core

**At the Proteomics Core, state-of-the-art instruments provide exceptional sensitivity and accuracy in protein identification and quantification. These advanced technologies empower researchers to explore a wide range of proteomics applications. We provide professional consultation, sample preparation, data acquisition, data search, and comprehensive data analysis and interpretation services.**

https://proteomics.iigb.ucr.edu/contact

# IIGB Genomics Core

Let us help your genomics research!

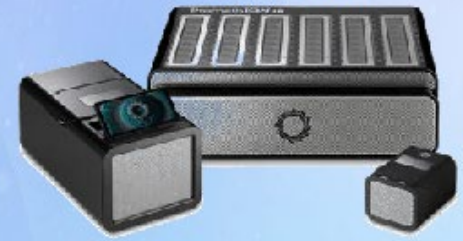
Wei Zhang
Academic Coordinator

Holly Clark
Research Staff

Clay Clark
Research Staff

**Sanger Sequencing**
ABI 3730XL DNA Analyzer

**Long-read Sequencing**
Oxford Nanopore
PromethION



**Next-Generation Sequencing**

Illumina NextSeq 2000

Illumina NextSeq 500

Illumina MiSeq

Keen Hall
gencore@ucr.edu
https://genomics.iigb.ucr.edu/

UC RIVERSIDE | Institute for Integrative Genome Biology

# IIGB Bioinformatics Core

1207G Genomics | facility.bioinformatics.ucr.edu |
bioinformatics@ucr.edu | 951-827-7990

## IIGB Bioinformatics Core

The bioinformatics facility provides convenient access to the NGS data sequenced at UCR, data analysis, and programming expertise. The resources serve scientists at UC Riverside to master the informatics needs of their research in a proficient and cost-effective manner. With over years of data analysis experiences, the core can assist with your bioinformatics need to advance your research.


SERVICES

### DATA ANALYSIS

The core provides data analysis services for major common datasets including: **Genomics and Transcriptomics** (e.g., RNA-seq, small RNA-seq, scRNA-seq), **Epigenomics** (ChIP-seq, Methyl-Seq, ATAC-seq), **Variant-seq**, and **Metagenomics**. Additionally, the core can work with users to develop custom/novel analysis pipeline. Please contact the core for a free consultation and quote regarding your projects.

TRAINING

### WORKSHOPS

The core provides hands-on training and workshops on basic and advance bioinformatic topics from introduction to basic command line tools to NGS data analysis workflows (e.g., RNA-seq). These training sessions and workshops are open to undergraduate and graduate students, post-docs, staff scientists and faculties.

Check out the Github repository for workshop info: https://github.com/orgs/bioinformatics-workshop/repositories

HELP

### OFFICE HOUR

The core provides daily office hours (2-3PM) to assist with bioinformatics-related questions/ issues. Make an appointment via the Calendly link: https://calendly.com/ucr_iigb_bioinfocore or stop by the office in 1207G Genomics building.

PEOPLE

### BIOINFORMATICS COORDINATOR

Brandon Le, Ph.D.

Scan Me

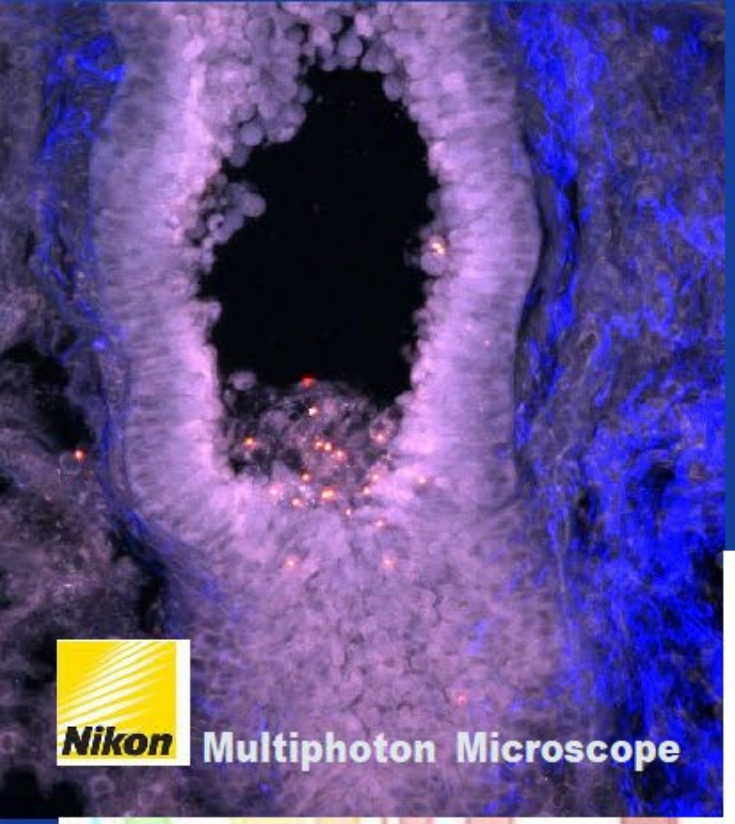
# School of Medicine Research Core

UC RIVERSIDE School of Medicine

## Request a tour of the SOM Research Core

Learn more about the instruments we offer to *all* researchers

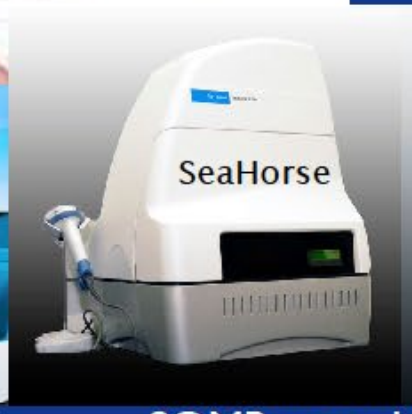
**including:** 10X: ChromiumX, Visium
Novocyte Quanteon
Nanostring: nCounter, GeoMX, CosMX
Astrios Cell Sorter, SeaHorse,
Multiphoton Microscope, Histology Core






10X GENOMICS

Contact Mary Hamer
mary.hamer@medsch.ucr.edu



SOMResearchCore@medsch.ucr.edu

UCR Library

RESEARCH SERVICES

# MAKER SERVICES

Maker Services consists of three labs: Creat'R, Robotics and 3DXP Lab. Each lab is equipped with emerging technology tailored for the innovation and creative prototpying for our patrons. For any questions regarding our services, please contact creatrlab@ucr.edu for more information.

1. CREAT'R LAB
2. ROBOTICS LAB
3. 3DXP LAB

### 3D PRINTING

Use 3D printing to rapidly prototype research projects for a sustainable future.

### PROGRAMMING

Learn how to program using various sensors and microcontrollers to create your own robotics system.

### MULTIMEDIA DEVELOPMENT

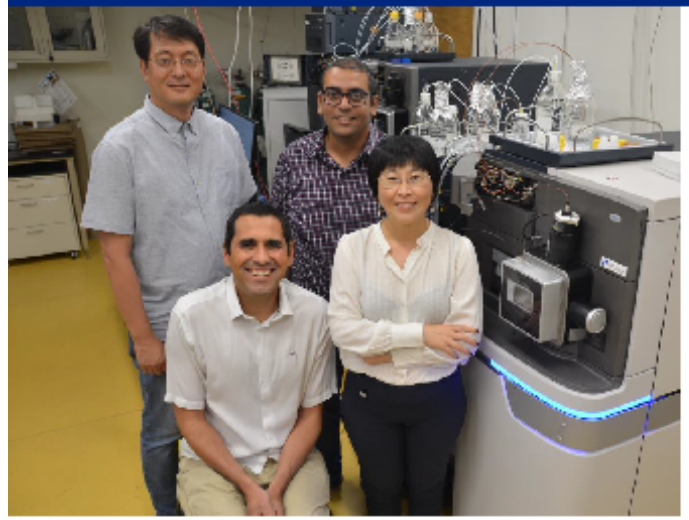
Use multimedia tools to create fun or academic projects for you or your organization.

Innovation and Making

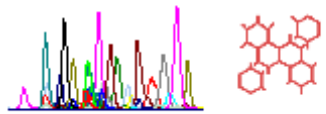
creatrlab@ucr.edu

# UC RIVERSIDE | Metabolomics Core Facility

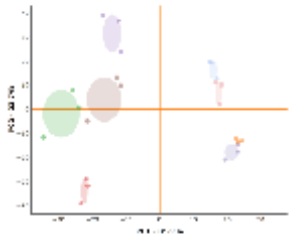
INSTITUTE FOR INTEGRATIVE GENOME BIOLOGY (IIGB)

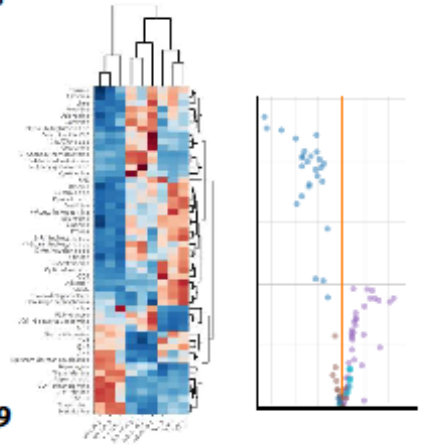


Left to right: Manhoi, Amancio, Anil, Haiyan



**Are you ready for a small molecule adventure?**







## We have full-service platforms for:

- Secondary metabolites / Natural products
- Lipids
- Central carbon metabolism
- Phytohormones
- Oxylipins
- Tailored methods and more

Instruments: Synapt G2-Si (QTof with ion mobility), Xevo TQ-XS (QQQ), Xevo G2-XS (QTof)

**Schedule a meeting today to discuss you project!**

**metabolomics@ucr.edu; Phone: (951) 5631119**
**https://metabolomics.iigb.ucr.edu/**

# Who we are

- Xiaoping Hu, PhD
  - Director
- Chelsea Evelyn
  - MRI Technologist
- Xu (Jerry) Chen, PhD
  - Research Scientist
- Jason Langley
  - MRI Physicist

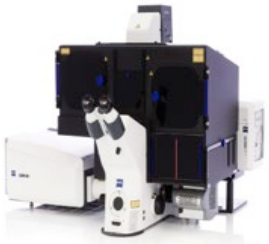

- We provide expertise for protocol development, task design for functional MRI scans, data acquisition, creation of data processing pipelines, and grant preparation

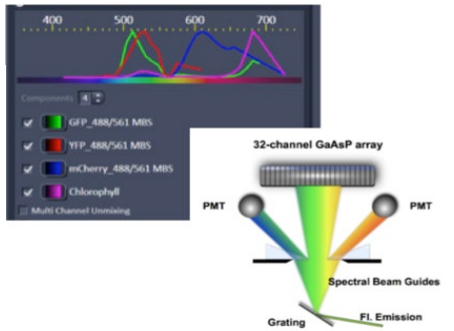
Center for Advanced Neuroimaging

# Microscopy and Imaging Core Facility:

David Carter, M.A., Ph.D. *(Cantab.)*
2025 Keen Hall. Dept. MCSB
o.951 827 2694;  c.951 850 2559
Microscopycore.ucr.edu  faces.ccrc.uga.edu Keenscopes

## Zeiss 880 Airyscan Fast Confocals

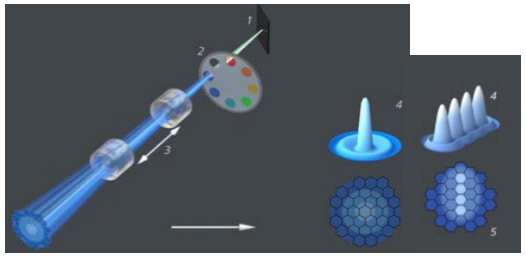




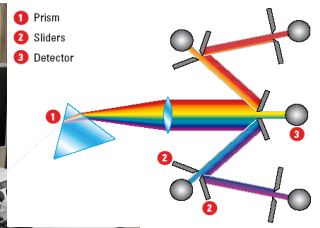

Spectral imaging
- Collects 32 colors simultaneously
- Auto-unmixes overlapping dyes

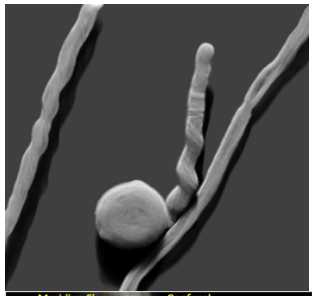
Airyscan 70% higher resolution
- Collects 32 facets simultaneously

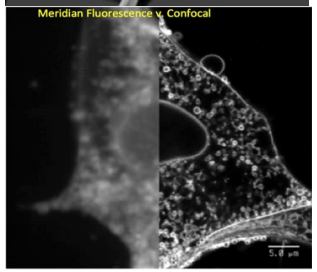
AiryFast 50% sharper ~10x faster



## Leica SP5 Confocal



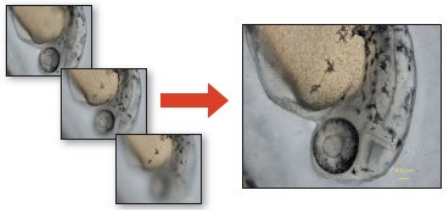



Any laser line - any emission range.
Easy and ergonomic

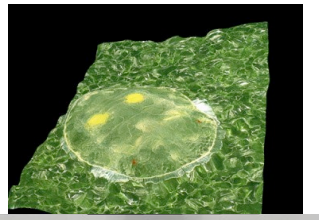
## Keyence BZ-X710 All-In-One

Inverted fluorescence/bf microscope
- Climate controlled stage incubator
- Birefringence cube and Analyzer added for Collagen
- Dry and (for us) water lenses; 2x-60x
- built-in color or b/w camera
- Tiling, kinetics, video capable

Extended Depth Imaging



## Keyence VHX-7000 Inspection Scope



Camera-Based 20x-2500x
- Topographic Projection
- Extreme Pixel Tiling
- Nightsea Fluorescence

Proteomics Core Facility IIGB

# A Glimpse in PROTEOMICS CORE


At the Proteomics Core, state-of-the-art instruments provide exceptional sensitivity and accuracy in protein identification and quantification. These advanced technologies empower researchers to explore a wide range of proteomics applications.

## OUR MISSION

We provide professional consultation, sample preparation, data acquisition, data search, and comprehensive data analysis and interpretation services.

## SERVICES:

- **Discovery Proteomics**
  (Data-Dependent Acquisition(DDA)/Data-Independent Acquisition(DIA), TMT labeling quantification/label free quantification)
- **Post-translation Modification Proteomics**
  (Phosphorylation/Glycosylation/SUMOylation, and others)
- **Liquid Chromatography**
  (Reversed Phase Chromatography, Affinity chromatography, HILIC Chromatography, size exclusion chromatography)
- **Proteomics data analysis**
  (GO analysis, String analysis, and others)
- **Ultra Centrifuge**
  (Various types of rotors are available)


**ACADEMIC COORDINATOR:**
Dr. Quanqing Zhang

**LOCATION:**
Office: Keen Hall 1019

**EMAIL:**
quanqing.zhang@ucr.edu

**PHONE:**
951-827-7114

# Plant Transformation Research Center
# Services, Training and Research


**Presented by**

**William Hsu, Assistant Specialist and**
**Martha L. Orozco-Cárdenas, Ph.D.**
**Director - PTRC**


**October 13, 2025**

UC RIVERSIDE