Incorporating Systems Biology to Vector-Borne Pathogen Research: Current Landscape, Challenges, and Opportunities

A Virtual Workshop Sponsored by the
Division of Microbiology and Infectious Diseases
National Institute of Allergy and Infectious Diseases
National Institutes of Health
May 17–18, 2022

Meeting Report

NIH National Institute of Allergy and Infectious Diseases

Table of Contents

Executive Summary

Organizers from the Division of Microbiology and Infectious Diseases (DMID) of the National Institute of Allergy and Infectious Diseases (NIAID) at the National Institutes of Health (NIH) brought together experts in arthropod vectors/vector-borne diseases and computation/modeling to explore systems biology (SysBio) approaches in both the vertebrate and invertebrate host. In accordance with NIAID's goals, the virtual workshop promoted communication and multidisciplinary collaboration among researchers, highlighted existing and needed resources, and sparked ideas about accelerating research.

Prior to the workshop, a Microsoft Teams communication platform was established to encourage participants to exchange ideas, information, and resources. The tool will be maintained to encourage continued collaboration among experts from the many different disciplines involved in applying systems biology approaches to vector-borne diseases.

Meeting Summary

Introduction and Purpose

Organizers from the Division of Microbiology and Infectious Diseases (DMID) of the National Institute of Allergy and Infectious Diseases (NIAID) at the National Institutes of Health (NIH) brought together experts in arthropod vectors/vector-borne diseases and computation/modeling to explore systems biology approaches in both the vertebrate and invertebrate host. The goals were to:

- Highlight examples of large dataset generation, integration, and analyses.
- Highlight research at both the vertebrate-pathogen interface and the invertebrate-pathogen interface.
- Facilitate experts' assessment of the availability of systems data and computational tools for the study of vector-borne disease systems (vector, pathogen, host).

Keynote speakers discussed vector-borne disease (VBD) research from both a systems biology and technology perspective. Day 1 of the workshop consisted of oral presentations on Flaviviruses, Borrelia, and Plasmodium as use cases to discuss VBD systems biology and modeling in both the vertebrate and invertebrate pathogen interface. Day 1 ended with optional informal networking in breakout rooms. Day 2 focused on technology advances and systems analysis, with specific examples of their implications for the VBD community. Next came a moderated panel that set the stage for discussion with the recent work of VBD and computational biology/modeling researchers. The final session was a panel discussion on roadblocks, opportunities, and next steps in VBD/systems biology research.

Day One, May 17

Remarks from Cristina Cassetti, Deputy Director, Division of Microbiology and Infectious Diseases (DMID), NIAID

Dr. Cassetti framed the workshop in terms of the missions of NIAID and DMID. Besides NIAID's mandate to produce a robust research portfolio on infectious diseases, the agency also develops nimble mechanisms to respond rapidly to new epidemics and pandemics.

Two of DMID's scientific priorities directly involve vector-borne pathogens: develop countermeasures that mitigate the impact of emerging infectious diseases and biodefense threats; and improve strategies for the prevention and treatment of malaria.

VBDs are responsible for 17 percent of infectious diseases worldwide, Dr. Cassetti noted. Many emerging and re-emerging diseases are vector-borne. Environmental and climate changes are driving the emergence of diseases in areas where they have not been seen before. VBDs are complex, involving the pathogen, arthropod vector, and host. Understanding VBDs requires the most sophisticated tools available, including systems biology. The outbreak of Zika, a mosquito-borne viral infection, prompted one example of NIAID's response to a global VBD epidemic. The agency spent more than $150 million in two years to build a program from scratch, from basic viral biology to developing vaccines, diagnostics, and therapeutics.

Dr. Cassetti emphasized the workshop's purpose of gathering experts in arthropod vectors/vector-borne diseases and computation/modeling to discuss how to generate and analyze large datasets to improve understanding of these important pathogens.

Session 1: Keynotes: An Overview of Systems Biology (SysBio) Approaches & Their Use in VBD Research

Modulation of Human Innate Immune Responses by Arboviruses
Ana Fernandez-Sesma, Icahn School of Medicine at Mount Sinai

Dr. Fernandez-Sesma noted that her focus is the dengue virus (DENV). She said that about 3.5 billion people are at risk of infection globally, and this number may expand with increased travel and environmental shifts due to climate change. Dengue is spread by mosquitoes, and a good animal model does not currently exist. Dr. Fernandez-Sesma and her team used human monocyte-derived dendritic cells (MDDCs) to study the biology and effects of the dengue virus.

Dengue virus 2 (DENV-2) and dengue virus 4 (DENV-4) infection kinetics are different in human dendric cells. DENV-2 secondary infection is associated with more severe disease, whereas DENV-4 infection produces a milder clinical presentation. DENV-4 infection also induces greater cytokine secretion compared with DENV-2.

Dr. Fernandez-Sesma served as co-Principal investigator (PI) for the NIAID-funded Dengue Human Immunology Project Consortium (DHIPC) that included studies on natural DENV human infections, vaccines, and *ex-vivo* infections. Core techniques employed included genomics, proteomics, immune phenotyping, data analysis and modeling, and data dissemination. Results include:

- Bystander cells produce different cytokines than DENV-infected cells.
- The study of human tonsil cells *ex vivo* using aural spectral flow cytometry revealed again that DENV-2 and DEN-4 show different infection kinetics, with DENV-4 replicating rapidly and DENV-2 starting slowly and picking up.
- DENV-4 infection induces greater type 1 interferon (IFN)-related cytokine secretion than DENV-2 infection. DENV-4 induces T helper type 1 (TH1) secretion compared with DENV-2 infection.

The recognition and examination of the complexity of DENV infections has led to a proposed model for viral fitness strategies:

- DENV-4: Exhibits early viral replication that results in early viremia and early transmission to the mosquito. This induces a higher human innate immunity, which may result in a higher human adaptive immunity (TH1 response), early clearance of virus, and less severe disease.

- DENV-2: Exhibits late viral replication, late viremia, and late transmission to the mosquito. This results in a more muted human innate immunity, late clearance of the virus, and a more severe disease.

Discussion

In response to an audience query, Dr. Fernandez-Sesma discussed reproducibility and the challenges with tissue in terms of the tonsil model. Her team did not see the infectivity of dengue as pronounced in all tonsil tissue but did see the consequence of exposure to the virus. Her team is working now to divide tonsils into high responders and low responders. This difference supports the observation that with dengue, 80 percent of people are asymptomatic or have a mild version of the disease. Its remarkable reproducibility allows a focus on key elements that define the profile, such as the interferon signature.

A meeting participant inquired about emerging pictures of defective virus genomes and what role they might play in dengue disease. Dr. Fernandez-Sesma said the tetravalent live attenuated vaccine is being used to observe if there is any role for the sub-genomic remains that are generated in an infection as compared with wild type viruses.

In response to other audience queries, Dr. Fernandez-Sesma noted:

- The strain of DENV-4 used in her study was 1668 from Indonesia, with primary isolates from Nicaragua.
- Both DENV2 and DENV4 are sensitive to IFN and can inhibit IFN production similarly. Her team believes the bystander cells contribute to the IFN signature in DENV4.
- Her team is using proteomics to look at the metabolomic output of cells infected with the two viruses and its correlation with the infection dynamics.

Session 1: Keynotes: Cutting-Edge SysBio Technologies & Potential Applications for VBDs

Systems Immunology of Human Immune Variations: Do Differences Make a Difference?

John Tsang, NIAID Laboratory of Immune System Biology, Multiscale SysBio Section

Dr. Tsang outlined three areas that his research team has been exploring at the human population (systems human immunology) level:

- The factors that determine why individuals mount variable immune responses to perturbations such as infection, vaccination, and disease and whether those outcomes can be predicted.
- Comparative analysis of the human immune system across various health and disease states with a goal of discovering the shared commonalities across different types of diseases and whether those can help researchers learn more about health.
- Building tools to enable this kind of systems approach.

Dr. Tsang noted that there is also fascinating biology at the single cell level. When researchers look at homogenous populations of cells, there is still a lot of variation in gene expression. This raises the question of whether the immune system leverages this kind of extensive variation to allow it to respond to different perturbations.

Dr. Tsang highlighted his team's research efforts in these areas and the technologies used, including Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-Seq). CITE-Seq performs multi-modal single-cell profiling by measuring hundreds of proteins together with transcriptome messenger RNA (mRNA) expression within single cells (Stoekius *et al*., *Nature Methods* 2017). Dr. Tsang's team has been scaling up CITE-Seq analysis to study individuals over multiple time points. Individual donor cells are identified by single nucleotide polymorphism (SNP) calling, enabling batches of 45-50 samples to be processed together. As automation increases, capacity will rise to hundreds of samples. This pooling allows processing of the samples together, followed by sequencing of single cells in the pool.

The data produced by these technologies presents new challenges in understanding where the noise is coming from. For example, proteins have distinct noise structures compared to mRNAs. The denoised and scaled by background (DSB) method is used to normalize and denoise droplet-based protein expression data (Mule *et al*., *Nature Communications* 2022). DSB is an open-source package available to all researchers.

Dr. Tsang's team clustered the single cells based on the protein expression on the cell surface. Within each cell cluster, researchers looked at the transcriptional profiles and how they differ across groups of individuals in a study. Dr. Tsang highlighted other types of variations that exist in this data, including variations over time and in reaction to perturbations such as an infection or vaccine. Researchers are attempting to predict responses, uncover potential determinants, and discover shared predictors and mechanisms among determinants.

These determinants could include genetics, intrinsic factors (age, sex), and pre-existing immunity (antibody levels), but these do not explain all phenotypic variability. A study conducted by Dr. Tsang's team found the baseline immune state of an individual to be a major determinant, and the informative markers of that immune status to be temporally stable (Tsang, JS, *Trends in Immunology* 2015). Further study using various cohorts across location, season, and perturbations (influenza and yellow fever vaccines) suggests that this baseline immune state may predict the responsiveness potential of an individual (Kotliarov, Sparks *et al*., *Nature Medicine* 2020).

Dr. Tsang's team applied CITE-Seq to the whole blood transcriptional signature of a cohort of both high and low responders. Researchers found that in high responders, a whole circuitry of cells (such as B cells) were in a more poised, temporally stable state before a perturbation (vaccine).

Dr. Tsang's colleagues conducted a study in 2020 to explore how such antigen-agnostic baseline setpoint states are established. Researchers enlisted a healthy, non-obese cohort that experienced mild (non-hospitalized) COVID-19. Researchers compared this cohort's reaction to a flu vaccine with the reaction of a cohort of people who had never had COVID-19 to examine whether COVID-19 may have reshaped the baseline of some cells (). Researchers found that cluster of differentiation 8 (CD8) + T cells at the baseline—which are an imprint of mild COVID-19—are also the cells that responded to the flu antigens.

Discussion

In answer to the question of whether baseline signatures are specifically associated with immune responsiveness or broader indicators of general physiological response, Dr. Tsang replied that physiology may indeed play a role in immune responsiveness. Subclinical inflammation and hormones are examples. Females tend to have higher levels of the baseline signature score, potentially explaining why they generally mount more potent responses to vaccines. But within each sex alone, the score still delineates high vs. low responders. This indicates that while sex and hormones play a role, other factors contribute as well.

A workshop participant asked how many proteins are generally used to differentiate a tertiary cell type and to what Dr. Tsang attributes the variability baseline tertiary cell types. Dr. Tsang said that his team typically clustered the cells using all the proteins, then refined further if needed. The annotation of the cell clusters has been done both manually based on known markers and using approaches in Seurat to map cell clusters to reference populations.

When asked whether analysis of yellow fever vaccine datasets can be used to identify the differentiation pathway of CD8T cells responding to a vaccine, Dr. Tsang said that tracking the kinetics of T cell phenotypes over time might provide some information about the sequence of events. Right now, his team is only looking at blood.

Session 2: Vector-Borne Disease Systems Biology & Modeling in the Vertebrate & Invertebrate Host: Flavivirus as a Use Case

The Vertebrate Host-Flavivirus Interface

Let's Get Physical: A Zika Virus-Host Protein Interaction in Replication and Pathogenesis

Priya Shah, University of California, Davis

Dr. Shah focused on a single Zika virus-protein interaction identified using proteomics and the consequences of that protein interaction.

She explained that for a virus to be successful at a molecular level, it must get into the cell, replicate its gnome, get out of the cell, and accomplish these actions while counteracting host defenses. These activities are often accomplished by co-opting host machinery through physical interactions between a viral protein and a host protein. Identifying these interactions can give researchers insight into the mechanisms of replication and how viruses potentially cause disease.

A focus of Dr. Shah's team is using comparative approaches to highlight biologically relevant protein interactions—comparing and contrasting virus-host networks across different viruses or different hosts, human and arthropod. The idea is to identify protein interactions that are highly conserved and reveal conserved mechanisms of replication.

Dr. Shah's team generated three different flavivirus-host protein interaction networks—dengue-human interactions, dengue-mosquito interactions, and Zika-human interactions (Shah *et al*., *Cell* 2018). The team identified individual proteins that are conserved across all three networks as well as those that are unique to one or two. A systems-level view of the data clearly showed categories and processes that were enriched in multiple networks and those that were unique to one virus or host.

Of particular interest was the Zika virus enrichment of protein interactions related to human development. The protein interaction that stood out to Dr. Shah's team was that between Zika virus NS4A and human protein ANKLE2. ANKLE2 variants are implicated in a congenital hereditary form of microcephaly. Dr. Shah's team coordinated with Dr. Nichole Link at the University of Utah to test whether NS4A inhibits ANKLE2 to force it to act like a loss of function mutant. Dr. Link found that NS4A inhibits fruit fly brain development in an ANKLE2-dependent manner. Dr. Shah emphasized that in the process of discovering this interaction, researchers were able to take a large systems-level data set, leverage the comparative proteomics approach to identify an interaction specific to the Zika virus, and quickly resolve molecular mechanisms of viral disease using the fruit fly model.

Dr. Shah's team also explored whether ANKLE2 is involved in Zika virus replication and whether the NS4A-ANKLE2 interaction can be broken. Researchers were able to establish an ANKLE2 involvement in Zika replication (Fishburn *et al.,* bioRxiv 2022). Across a range of multiplicities of infection with Zika virus, researchers were able to identify a three- to eight-fold decrease in Zika virus infection virion production approximately 24 hours post infection. The team is now working on a more complete depletion of ANKLE 2 using clustered regularly interspaced short palindromic repeats (CRISPR) knockout.

Dr. Shah's team dissected the NS4A-ANKLE2 interaction at a molecular level using AlphaFold, a new tool that predicts protein structure. The process not only identified ANKLE2 established domains, but three other uncharacterized structured regions that the team could use to design the deletion mutants. NS4A failed at co-immunoprecipitation (co-IP) with a deletion mutant that is missing both the transmembrane and LEM domains. During infection, this mutant also failed to co-localize with NS4A during infection compared to the full-length ANKLE2. This process illustrates the movement from systems to molecular mechanisms using both computation and experimentation. The team is now including proteomics on ANKLE2 itself to understand how it is coordinating virus replication, how specific mutations in ANKLE2 might lead to pathogenesis or potentially protection, and whether the viral replication and disease aspects of this interaction can be uncoupled.

Dr. Shah concluded by discussing how SysBio can be made more efficient. Given that hundreds of interactions are generated in a protein interaction network, it is important to be able to choose protein interactions more confidently. She highlighted the potential of high throughput animal models, leveraging human variants in systems level analysis, and using predictions of protein structures and interactions themselves by leveraging AlphaFold, RosettaFold, or experimental approaches like deep mutational scanning.

Discussion
Dr. Shah was asked how the tissue or cell type and where those data and interactions come from factor into her analyses. She responded that with proteomics, one cannot amplify input for analysis, so proteomics research tends to be done in easy-to-use cell types. In terms of expression patterns, her team is interested in whether ANKLE2 is expressed in relevant cell lines or tissues. Her team focuses on protein interactions that are present in many different tissue types. There has been some work on these virus-host protein interaction networks in neuro progenitor cells for Zika virus and Dr. Shah said she would love to directly compare those two approaches.

A participant asked whether Dr. Shah cleans up the initial interaction data computationally or also uses logical reasoning to exclude interactions. Dr. Shah replied that her team does it entirely computationally to remove any element of human bias based on a mathematical formula of how reproducible the interaction is and how abundant the protein is that has been detected. Both interactions can be relevant. The key is to focus on specificity, which is where doing this repeatedly for every single viral protein can help discern whether the protein interaction is truly relevant.

The Invertebrate Host-Flavivirus Interface
Integrating Different 'Omics Approaches with Molecular Tools to Study Flavivirus-Vector Interactions That Drive Transmission and Emergence
Kevin Maringer, The Pirbright Institute

Dr. Maringer focused on tools developed to support molecular SysBio approaches to study flavivirus transmission and pathogenesis. His laboratory takes a comparative virology approach to examining how flaviviruses emerge. Most work focuses on *Aedes aegypti*, a major arboviral vector because it transmits arboviruses that effect humans, is distributed across the tropics and subtropics, and can sustain explosive outbreaks in urbanized areas.

Dr. Maringer explained that to unlock the potential of SysBio in the study of mosquitoes, it is important that the molecular models that generate initial data are tractable, standardized, accessible to people across the field, and safeguarded for the future. This is also key during molecular follow-up of the targets identified and will be increasingly important as researchers apply artificial intelligence to their work.

Among the limitations of mosquito cell lines are the fact that they are derived from embryos and larvae, not adults; lack standardized, defined, line cultures; and can be immunocompromised. Steps taken by Dr. Maringer and his team to address challenges include:

- Generated Aag-AF5, a clonal, standardized *A. aegypti* cell line that is immunocompetent and supports replication with a number of arboreal cell lines (Fredericks *et al.*, *PLOS Negl Trop Dis*. 2019).

- Identified several Aag 2-derived single-cell clones with low or undetectable levels of insect viruses in culture.
- Worked towards a tractable molecular toolkit for *A. aegypti* cell culture, including significantly boosting protein expression in individual cloned cells and developing the first CRISPR knockout cell line for any mosquito species (Varjak *et al*., *mSphere* 2017; Scherer *et al*., *Viruses* 2021).

Dr. Maringer outlined work being done by his research team to integrate proteomics with transcriptomics ("proteomics informed by transcriptomics," or PIT) for non-model organisms (Aag2 cells):

- Proteomics experiment followed by RNA sequencing, assembly of transcripts *de novo* without the need for a reference genome. Used these to predict the proteins that would be present in the actual sample and compare the predicted spectra with the detected spectra for identification of additional proteins (Maringer *et al*., *BMC Genomics* 2017).
- Conducted global proteomics to verify 4,760 *A. aegypti* proteins at the protein level for the first time. Conducted refined genome annotation using PIT to examine the proteins identified as non-insect proteins. Most are new annotations that previously have been missed. PIT was used to identify any problem hotspots in the reference genome to target efforts to improve it. Although none were identified, Dr. Maringer called PIT a powerful technique to determine which reference genomes from which vector species might benefit from attention to improve the reference.
- Conducted the first proteomics analysis of the entire proteome derived from transposable elements (Maringeret al., *BMC Genomics* 2017). PIT helps identify specific peptides that can be mapped onto transcripts, allowing researchers to catalogue transcripts actively translated.

Dr. Maringer concluded by discussing research gaps and opportunities:

- Molecular models – Develop better cell lines that are standardized, defined-lineage, karyotyped, immunocompetent, and derived from adult tissue. Leverage repositories for reagent safeguarding and sharing. Develop antibodies.

- Bioinformatic approaches – Improve PIT using long mRNA sequencing. Maximize community benefit of integrated 'omics data through data sharing and machine learning.

Discussion

When asked about his strategy for ensuring a consistent improvement to annotation that will benefit all researchers, Dr. Maringer said that his team put the PIT reference genome in VectorBase and would like to engage more with this resource to map the most recent assembly. In the meantime, Dr. Maringer will share that database with interested individuals.

A discussion ensued about what can be done to catalyze antibody development. One suggestion was bringing groups of people together to generate some of the most sought-after antibodies. Funding such an endeavor is a challenge. Some of the commercial antibody companies have been open to developing new antibodies if the companies are convinced there is a market. There were a few examples from the early/mid 2000s in mice where new clones against key polymorphic markers were developed after community-wide guarantees for purchase.

Session 3: Vector Borne Disease Systems Biology & Modeling in the Vertebrate & Invertebrate Host: *Borrelia* as a Use Case

The Vertebrate Host-Borrelia Interface

Robert Moritz, Institute for Systems Biology

Dr. Moritz discussed the work being done at his laboratory on Borrelia and Lyme disease in a three-step strategy to better understand Borrelia on a systems level to define new diagnostic pathways and understand the Borrelia-human cell interface.

1. Deciphering the Borrelia genome/proteome.

Sequencing performed in the past has been done at a low level that misses a lot of components, Dr. Moritz explained. His team cultured Borrelia laboratory strains and Borrelia clinical isolates from around the United States and also drew from public datasets to build a PeptideAtlas with the aim of quantifying every protein in Borrelia. This process identified 1240 proteins. Dr. Moritz's team also conducted a two-phase Borrelia genome annotation correction process on curated databases with their own proteomic data. The team identified common peptides—as well as different annotated proteins and pseudogenes—and validated them. This process has allowed the recovery of missing genes in Borrelia that are being expressed and proteins that are being induced. The process also recovered peptides unique to pseudogenes in Borrelia and showed that these are uniquely expressed proteins as well.

2. Building comprehensive proteomes across Borrelia isolates.

Dr. Moritz explained that when one examines the gene bank, many Borrelia have differing amounts of plasmids. Most lab strains lose a fair number of plasmids, something that researchers need to be aware of. Dr. Moritz's lab has observed a full complement of plasmids in infected strains from throughout the United States, except in isolates that have been grown mainly in the lab or those that are not infective.

3. Quantitative approaches to understand the vertebrate host and Borrelia interactome.

Dr. Moritz next discussed how to take the information built through improved databases and other sources and take a quantitative approach to understanding the protein interactions between the Borrelia and the human host. Steps include identifying all the membrane proteins in Borrelia strains, quantitatively measuring these proteins, and defining the true outer membrane proteins (Borrelia have proteins on both the inner and outer surface of the outer membrane). The idea is to target the proteins that change in expression when they go from the tick phase to the mammalian phase (Kurokawa *et al*., *Nat Rev Microbiol* 2020).

Dr. Moritz discussed two strategies for defining the vertebrate host-Borrelia interface—a simple affinity purification with mass spectrometry, which produces hundreds of thousands of interactions that require bioinformatic sorting into true interactions and false positives, and TurboID proximity labeling mass spectrometry, a more defined approach that expresses each one of the proteins with the enzyme attached.

Discussion
A workshop participant asked whether the spirochetes for mass spectrum analysis by Dr. Moritz's lab are grown in culture or obtained directly from humans. Dr. Moritz replied that they are membrane proteins grown in culture, with the clinical strains obtained from skin biopsies and grown out into defined isolates.

The Invertebrate Host-Borrelia Interface: Tick Immunology and Microbial Interactions
Joao Pedra, University of Maryland

The focus of Dr. Pedra's laboratory is understanding the tick immune system—in particular, how lipids trigger the immune deficiency (IMD) signaling pathway. The research team discovered that certain lipids can bind to a molecule called croquemort and croquemort can lead to activation of NF-kappaB, translocation to the nucleus, and production of antimicrobial peptides (Shaw *et al*., *Nature Comms*. 2017; McClure Carroll *et al*., *PNAS* 2019; O'Neal *et al*., bioRxiv 2022).

Ticks have three hemocyte (immune cell) populations—prohemocytes (stem-like cells), plasmatocytes (involved primarily with phagocytosis), and granulocytes (can secrete antimicrobial peptides). The team took a SysBio approach to studying how the tick immune system responds to bacterial infection:

- Single cell RNA sequencing (scRNAseq) of hemocytes, followed by computational analysis – As the team navigated computational challenges, they developed their own pipeline, apparently for the first time in the tick research community. The team also used new genome sequencing data from the University of Maryland and the J. Craig Venter Institute.

- Tick hemocyte cluster analysis data with non-engorged nymphs – In addition to the three hemocyte populations already described, Dr. Pedra's team found a transitional population and an undefined population believed to be important in lineage differentiation.

- Gene annotation and expression – Researchers used both the unbiased and the manual approach.

- Pathway analysis and protein networks – In the context of granulocytes, there was an overrepresentation of protein synthesis; in plasmatocytes, an overrepresentation with protein networks associated with actin-dependent cell motility and protein processing and transport; and with prohemocytes, an overrepresentation of cell networks associated with energy metabolism, electrolyte homeostasis, and protein synthesis and assembly.

Dr. Pedra and colleagues would ultimately like to perform a single cell analysis to observe what happens with hemocytes not only during infection but also during feeding. His team would also like to develop a monoclonal library available to all researchers for their studies.

Dr. Pedra highlighted the work of Dr. Monika Gulia-Nuss (University of Nevada) showing that CRISPR is possible in organismal-based genetics (Sharma *et al*., *iScience* 2022) and Dr. Nisha

Singh's work with cell-based genetics. He outlined the cell-based CRISPR methodology used in his lab on tick cells to isolate genomic DNA and do polymerase chain reaction (PCR) quality control. Dr. Pedra's lab is also developing CRISPR activation technology, noting that both the loss of function assay and gain of function assay will be important for the research community. He added that his team would ultimately like to do high throughput screening analyses looking at the higher genome in tick cells.

Discussion
When asked about the model organisms that Dr. Pedra uses for Enrichment analysis, he said his team uses flies, humans, or basically anything available so that researchers can annotate. For the unbiased analysis, the team uses an *Ixodes scapularis* genome.

A workshop participant inquired whether Dr. Pedra's team has done any functional validation for hemocyte subtypes examining phagocytosis in only the plasmatocytes. Dr. Pedra replied that they have not done any of that work yet. Researchers do not want to make assumptions that whatever fundamental principle is happening in mosquitoes and flies is happening in ticks. He said he has learned earlier that ticks can teach his team new principles in immunology. His team is taking an unbiased approach and will let the biology guide them before undertaking the reductionist approach.

In response to the problem of accurate annotation of translational start sites in many species. Dr. Pedra said his team is trying to design more single guide RNA and considering tiling them.

Session 4: Vector Borne Disease Systems Biology & Modeling in the Vertebrate & Invertebrate Host: *Plasmodium* as a Use Case

The Invertebrate Host-Plasmodium Interface: Using scRNAseq to Understand P. falciparum Development and Evolution

Virginia Howick, University of Glasgow

Dr. Hardwick described the work she has done with colleagues to build an atlas of malaria parasite transcription at the single cell level and how colleagues are using these tools and other genetic and genomic techniques to understand how *P. falciparum* has adapted to different vector species. Using single cell methods, researchers can get a transcriptome for each individual parasite from a population of heterogenous parasites to place individuals along a developmental trajectory and understand how genotypic differences are influencing gene expression at a particular stage.

The initial goal of the Malaria Cell Atlas project was to provide a resource of single cell transcriptomic data across the full parasite life cycle that includes diverse morphological forms. This was accomplished using the *Plasmodium berghei* mouse model. The project also covered the anthracitic development cycle of three parasite species using the *in vitro* culture system, then demonstrated how this data could be used as a reference by mapping on clinical isolate samples (Howick, Russell *et al.*, *Science* 2019).

Dr. Howick expanded on that work to thoroughly study the transmission stages using *P. falciparum*, the human malaria parasite. The transmission through the mosquito holds the greatest bottlenecks in the parasite lifecycle. Understanding the biology of these stages will help researchers target vulnerabilities in the life cycle to inform malaria control efforts. Dr. Howick's team optimized cell sorting to better capture these stages and produce data combined with data from asexual blood stages to get a more life cycle-wide view of transcription in *P. falciparum*. These data allow researchers to understand global patterns of expression as well as zoom into a particular stage of interest (Real, Howick, Dahalan *et al.*, *Nature Comms* 2021). A heat map of global patterns of gene expression revealed mostly stage-specific expression for each gene cluster. When a gene graph is overlayed with any gene level statistic of interest, researchers can link gene usage to stage-specific functions.

Dr. Howick concluded that when building a single-cell atlas of *P. falciparum* transmission through the mosquito, researchers can use single-cell data to reveal fine-scale developmental patterns and connect transcript expression to protein localization or show life cycle-wide patterns of co-expression to infer gene function and patterns of selection. She noted that all the data for her presentation are from a single genotype and remarked that researchers do not yet know how different parasites collected from different geographic regions would look if profiled in a similar manner.

The process of transmission in the wild is made much more complicated by not knowing which vector species will end up the host. The combination of a bottleneck in the parasite lifecycle and variation in vector species make this a strong selective pressure on the parasite. Dr. Howick's team hypothesizes that the patterns of differentiation seen in the population genomic data are driven by local adaptation to different vector species. There is also evidence for this selection at the phenotypic level.

Dr. Howick concluded by describing her team's current work to better understand patterns of local adaptation:

- Using single-cell transcriptomics of geographically diverse strains and vector species to understand how variation and developmental process at super-fine scale varies across different parasite-vector combinations.
- Taking a more direct approach using CRSPR-Cas9 based on population genomic data for allelic-replacement of candidate non-synonymous single nucleotide polymorphisms (SNPs) to observe how this alters infection in African and South American vector species.
- Optimizing low-input transcriptomic profile to understand how the physiological aspects of the vector may be exerting a strong selective pressure on the parasite.

Dr. Howick looked forward to expanding the SysBio approach to non-model vectors and to better understanding the geographic variation in *P. falciparum* and how these different combinations influence the pattern of transmission that is seen globally.

Discussion

A workshop participant asked about the level of resolution Dr. Howick sees at the single cell level and how many genes are detected and speculated that the resolution in field samples may not be as strong as that seen in lab isolates. The participant also asked if this issue affects clustering and if so, how Dr. Howick's team corrects for it. Dr. Howick said that her team sees the greatest difference in terms of genes per cell detection across the different stages. Not as many genes per cell are detected in smaller cells—just a couple hundred on average—whereas at the replicative stages, many thousands are detected. Dr. Howick's team can control for this in the clustering analysis using different normalization techniques and a sample of the most highly variable genes. The field samples shown are from early in the development of the methods being used and the hit rate was low. The few cells from which researchers got good transcriptomes matched in terms of genes per cell what would be expected based on lab data.

Dr. Howick was asked to comment on approaches to explore discordance of transcript and protein expression levels (transcriptional regulation) using data generated for known regulated genes, e.g., Pfs25 in the transmission stages of Plasmodium. Dr. Howick commented that there is much work to be done in the transmission stages where more translational repression is seen. Her lab has not worked on these issues yet but said it will be interesting to watch as the proteomic methods improve with detection.

The Invertebrate Host-Plasmodium Interface: Systems Analysis of Plasmodium falciparum in Relation to Parasite Density and Severity of Disease

Karen Day and Michael Duffy, University of Melbourne

Life cycle transitions of Plasmodium are dependent on parasite density. Dr. Day focused on the density-dependent death pathway with apoptotic-like cells and what researchers are doing with transcriptomic, metabolomic, etc., and the relationship to cell density, including how it can cause significant changes in transcriptome and metabolome.

To investigate this pathway, Dr. Day and her research team set up a high density/low density parasite culture system. In a high density culture, the predominant phenotype begins to exhibit aberrant morphology in the later asexual lifecycle. Features of density-dependent death include failure of schizont maturation, failure of merozoite formation, appearance of small size and blebbing, release from erythrocytes, and loss of mitochondrial membrane potential. Further investigation revealed that exposure to conditioned medium induces cell death in *P. falciparum* (Chou *et al., FEBS J* 2018). Researchers collected a high density conditioned medium from a high-density culture, a low density conditioned medium from a low density culture, and uninfected condition medium with no parasite. Researchers then put these media onto low density trophozoite culture. The death phenotype is generated only in the high-density culture. Transcriptional and biochemical analysis of high density conditioned medium revealed that transcriptional profiles are discriminated by density conditions.

When looking at results of transcriptional experiments, the team could see significant changes in the transcriptome of high-density cultures. Sixteen percent of the more than 4000 genes tested were differentially expressed. This is significantly more than what would be seen, for example, when parasites are treated with anti-malarial drugs. This includes a dysregulation of the genes involved in antigenic variation and ribosome biogenesis.

An analysis of the metabolite and lipid composition of the conditioned media using liquid chromatography–mass spectrometry (LC-MS) showed again that there were significant differences in the metabolites found in high density, low density, and unaffected media. Because metabolite profiles are discriminated by density conditions, researchers are starting to conduct compound testing on metabolites to generate the phenotypes. Conditions for density sensing are:

1. Occurs during specific growth stages or physiological/environmental conditions.
2. Threshold concentration of signal generates concerted response.
3. Cellular response beyond metabolism or detoxification of signal.

Researchers are seeking to identify a factor that would be recognized by a specific receptor.

Dr. Day concluded by describing the complexities faced when thinking about SysBio in a human host. In endemic areas where she has worked, she has seen at least four species co-infecting hosts and multiple genotypes of these species.

Dr. Duffy described a systems analysis of *P. falciparum* in relation to parasite density and severity of disease, accomplished through transcriptomics of *P. falciparum* clinical samples. The presentation outlined how parasite transcriptomes differ between severe and uncomplicated, or mild malaria. Moreover, integration of serology and genomic data provides further confidence to identify mild versus severe cases. Dr. Duffy concluded that transcriptomics of clinical malaria parasite samples can identify pathogenesis associated proteins and processes. However, there are limitations to their approach. For example, samples should be controlled for variations in parasite development between disease states, which can be partially resolved through single cell genomics/transcriptomics. Moreover, hypotheses derived from transcriptome data should be tested with orthogonal approaches ideally using the same samples and from longitudinal sampling. Finally, additional insights can be gained by integrating analyses of other large, published datasets which points to the promise of 'secondary' systems dataset analyses for uncovering novel insights into *P. falciparum* pathogenesis drivers and disease biomarkers.

Day Two, May 18

Recap and Day 2 Preview by Reed Shabman, Meeting Organizer, DMID, NIAID
Dr. Shabman briefly recapped the presentations from Day 1 and previewed Day 2 topics covering technological advances in systems analyses followed by a moderated discussion and a panel discussion in which session presenters and workshop attendees shared perspectives and feedback.

Dr. Shabman proceeded to set the stage by posing three high-level survey questions to attendees, then providing key results:

1. *What is your biggest challenge in systems biology approaches*?

Data analysis and data integration (e.g., multi-omics analyses): 40 percent
Computational expertise: 29 percent
Access to relevant 'systems' data sets: 14 percent
Data generation (e.g., access to technologies): 9 percent
My biggest challenge is not listed: 8 percent

2. *What technological advance is most exciting to you?*

Advances in genomics (e.g., vector genomes and annotations): 20 percent
Advances in transcriptomics (e.g., spatial transcriptomics): 25 percent

Advances in ’less mature’ technologies (e.g., metabolomics, lipidomics, glycomics): 27 percent
Advances in informatics applications (e.g., graphical interfaces for omics analyses): 20 percent
My most exciting advance is not listed: 9 percent

3. *Opportunities to collaborate on systems biology projects are available to me.*

Strongly agree: 17 percent
Agree: 48 percent
Disagree or strongly disagree: 35 percent

Session 5: Technological Advances in Systems Analyses: Implications for the Vector-Borne Disease Community

Advances in "Omics" as It Relates to the Vector, Pathogen, Vertebrate Host: Unmasking Ammonia Metabolism in A. aegypti Mosquitoes with Mass Spectrometry-Based Metabolomics

Patricia Y. Scaraffia, Tulane University

Dr. Scaraffia began by emphasizing the diversity of ‘omics techniques in systems biology, and their corresponding datasets. In her research, she has been unraveling mosquito metabolism with mass spectrometry (MS)-based metabolomics with a variety of state-of-the-art MS instruments (Horvath, Dagan and Scaraffia, *Trends Parasitol*, 2021).

Dr. Scaraffia discussed her team’s work to uncover ammonia metabolism in *A. aegypti* with MS-based metabolomics and RNA interference (RNAi). Given that a huge amount of ammonia is released during digestion, her team worked to understand how female mosquitoes detoxify such a concentration of ammonia in the absence of the urea cycle. Researchers observed a decrease in concentration of certain amino acids in the mosquitoes’ hemolymph 24 hours after feeding, leading them to wonder whether certain amino acids act as temporary sinks of nitrogen.

Through ^{15}N, ^{13}C isotope tracing and RNAi techniques, Dr. Scaraffia’s team discovered multiple unexpected metabolic networks and cross-talk mechanisms for ammonia detoxification. The team mapped the interconnectedness between glucose and ammonia pathways in mosquitoes to determine whether the C skeleton of glucose can be utilized to support ammonia detoxification pathways (Ma, Dagan, Somogyi, Wysocki, and Scaraffia, *J Am Soc Mass Spectrom* 2013; Horvath, Dagan, Lorenzi , Hawke and Scaraffia, *FASEB J* 2018). This mapping of metabolic interactions at the C atomic level in *A. aegypti* revealed a metabolic link between glucose and ammonia metabolism and revealed that glucose supports ammonia detoxification and nitrogen disposal.

The application of MS-based metabolomics, isotope tracing, and RNAi empowered researchers to perform high-precision dynamic studies, monitor multiple isotopologs simultaneously, and reveal pathways at the atomic level. MS-based metabolomics has emerged as a powerful approach to study mosquito metabolism and host-pathogen-vector interplay. Dr. Scaraffia said she anticipates that the integration of metabolomics with other ‘omics and modern genome editing techniques will be expanded to the study of additional vectors to assist in the development of novel strategies to reduce vector populations and/or to disrupt parasite development or virus replication in both human and vector. To this end, cross-collaborations between experts in their fields will continue to be essential.

Dr. Scaraffia outlined remaining challenges and knowledge gaps:

- The need to better understand the biology of the vector as a whole organism rather than the linear sum of its parts in response to internal or external perturbations.
- The need to be able to combine MS-based metabolomics, stable-label isotope tracers, genetic techniques, and other ‘omics to define the metabolic networks in non-infected and pathogen-infected vectors.
- Analytical challenges, such as data accuracy and validation and data interpretation and integration, which present the opportunity for collaboration among experts in different fields.

Discussion

One participant asked Dr. Scaraffia to discuss how conserved the pathways are in different mosquito species and within the Diptera vectors. She replied that the range can be 60 to70 percent, depending on the genes being investigated. Some of the metabolic pathways do not work in humans and vertebrates. Mosquito biology brings out a unique pathway to study—what happens after the female takes the blood in? Researchers have been fascinated by the regulation, which is different when compared to other organisms.

Another attendee asked whether Dr. Scaraffia has looked at whether triglycerides from the blood meal may also feed into ammonia detoxification. She replied that her team did not monitor that aspect, as they were doing target metabolomics. She noted, however, that 20 percent of the amino acids released are used to build reserves, including carbohydrates, proteins, and lipids. Her team was focusing on discovering a link between the glucose metabolism and the nitrogen metabolism.

A critical question raised was if Dr. Scaraffia had examined mosquito metabolomics in the presence and absence of pathogen infection. The answer was no, due to cost and experimental challenges. This represents a gap in systems data for the pathogen and its vector host.

Advances in Data Integration and Analytics: The VEuPathDB Family of Knowledge Bases

Mary Ann McDowell, University of Notre Dame

Dr. McDowell explained that VEuPathDB is a result of the merger of all the data, tools, and teams of VectorBase and EuPathDB. She and Dr. David Roos are the co-PIs. VEuPathDB is one of two NIAID-funded Bioinformatics Resources and supports data on more than 230 different species, including vector and vertebrate hosts and pathogens. All leverage the OrthoMCL Database so researchers can ask questions about orthology.

Data types supported include genome sequence and annotation; genetic and epigenetic variation; transcriptomes; proteomes; protein structure; metabolism; signaling networks; subcellular localization; interaction data when available; phenotypes from field, clinical, and mutant studies; and curated metadata so users can understand what data types are available and compare across data types.

Analyses provided include a genome browser where users can look at the genomic structure, put in tracks to look at the evidence underlying that structure, and examine synteny by comparing genomic structure across organisms. Also included are functional enrichment tools, transcriptomics, a variant calling tool, and orthology inference. VEuPathDB supports vertebrate and invertebrate host data and provides pathway analysis where users can look at a metabolic pathway and map transcriptomic data to discover what genes in that pathway are regulated. VEuPathDB also allows users to analyze their own data.

Other VEuPathDB tools highlighted by Dr. McDowell include the Apollo web-based genome annotation editing platform that allows researchers to curate genomes, and MapVEu which leverages GPS data to map where samples are collected throughout the world and allows users to identify and download the data that interest them. All VEuPathDB data is downloadable. Operators welcome researchers to submit their own data so that it can be made available to others.

Dr. McDowell told participants that the VEuPathDB currently has approximately 2,000 datasets, with a bimonthly release cycle to incorporate new datasets and data types. The datasets are searchable and users can set their organism preferences. She demonstrated through various screenshots how VEuPathDB's extensive and specific search functions and tools allow researchers to seek complex information such as the differentially regulated secreted proteins during dengue infection with orthologs in all arthropods.

Tools to enhance information include:

- The ability for users to save complex search strategies and share them with lab members and collaborators.
- The ability to form word clouds to help determine types of gene functions.
- The ability to create a word cloud based on gene ontology (GO) enrichment analysis.
- Graphic depictions of metabolic pathways onto which users can map RNASeq data to see how well the pathway is regulated.
- A community chat function for user interaction.
- Tutorials and exercises to learn how to use specific tools.

Dr. McDowell encouraged participants to alert VEuPathDB staff to new datasets. She then described tools available on ClinEpiDB, an open-access data analysis platform with datasets from a variety of epidemiological studies. ClinEpiDB includes sophisticated analysis tools such as a geolocation map, a drop-down menu for variables, overlays, and faceting of data. Each dataset has a summary page that includes data characteristics. The site also links to other relevant dataset sites.

Dr. McDowell concluded by encouraging participants to access a third platform, MicrobiomeDB, to explore more relevant datasets and tools.

Discussion

Dr. McDowell was asked whether vector-based and other based sites would be interested in hosting data beyond transcriptomic and genomic datasets, such as metabolomics or more diverse datasets—such as micro computed tomography volumes—from associated organisms. She answered that VEuPathDB has proteome data and a variety of metabolomic data, depending on a particular dataset. She said the platform is constantly trying to keep up with the type of datasets that come out and find the best ways to provide that information to users.

In answer to another inquiry, Dr. McDowell said that the latest *I. scapularis* genome is scheduled to be added to the database during the next release cycle.

When asked whether there are plans to integrate host microbiome data into VEuPathDB, she replied that there already is a robust ontology—a data dictionary with a hierarchy that allows researchers to talk between their databases and apply ontologies that other databases use. There are already vertebrate host microbiome data. Dr. Roos added that data from several tens of thousands of studies loaded onto on MicrobiomeDB also include some vector data. That platform's staff has been working with NIAID to ensure the ability for researchers to analyze their own data from microbiome studies and feed directly into MicrobiomeDB. He added that NIAID also funds the Bacterial and Viral Bioinformatics Resource Center (BV-BRC). Staff from these BRCs are eager to discuss how to best integrate with VEuPathDB across the silos more easily.

Advances in Modeling Development and Application: Advances in Multi-Scale Modeling for Vector-Borne Diseases: Integration and Data Fusion to Enable Prediction

Carrie Manore, Los Alamos National Laboratory

Dr. Manore introduced data analytics and modeling as one way to meet what she called the grand challenge—the need for rapid conversion of observations to knowledge in order to gain decision support and understanding about vector-borne disease systems. The large amount of data available is often noisy and gathered from many different sources. Mathematical and statistical models put that data together and pick the right combination to provide forecasts and predictions, quantify uncertainty in those predictions, and drive forward science and decision support. The idea is to combine data-driven and model-driven approaches.

Dr. Manore said that modeling can take the form of forecasting based on the data that researchers already have. This is a statistical approach that can provide decisionmakers with a forecast of expected cases. Modeling can also take the form of prediction, where the simulation of different what-if scenarios (such as what if officials spray for mosquitoes, make a treatment available, etc.) are studied for their impact.

For an epidemiological overview, Dr. Manore discussed the question of what scale and resolution are needed. When thinking of scale during the spread of an epidemic, for example, are researchers looking at a community, city, state, or nation? There is also fidelity (resolutions)—are researchers considering the mean, homogeneous mixing, or each individual (cell, person, mosquito)? Models will range from course to fine and require very different computational resources.

Statistical and Machine Learning Models for Forecasting (specific outcomes based on current or past data) – Dr. Manore used the example of a dengue outbreak in Brazil, but modeling also applies across scale. She listed the data sources that could help researchers predict a sengue outbreak (satellite imagery, weather stations, demographics, and non-traditional sources such as Google queries about health trends). She discussed the issue of data fusion—a major issue for forecasting. The data must be fused to spatial and time scale. Dr. Manore said alignment of space and time across organisms and other conditions is tricky and a hidden cost of modeling. She showed results of the modeling, including spatial maps and number of cases in a season.

Challenges of these models include finding data at multiple spatial scales and across species and pathogens, finding data across time and other variables such as weather, and piecing together data sets collected under different conditions, making data fusion an ever-present challenge.

Mechanistic Models – These are used to put together rules about how things should be working to create a model of what the outcome should be. Dr. Manore used a mosquito-borne disease model as an example. In this case, humans can be put into predictable stages of susceptible, exposed, incubating, infectious, and recovered. Humans move between these stages at different rates. Those rates were determined through extensive fieldwork and lab work to help parameterize the rates and how they change based on the heterogeneity of the people involved. Notably, mosquitoes also have similar stages, and this is an area of future investigation in the modeling space.

Dr. Manore showed workshop participants the progression from theoretical framework and model development, data integration and parameterization, refining simulations, and estimates and uncertainty quantification (UQ) for quantities of interest based on fitting the mechanistic model to the reported data. Examples of information that can be provided to decisionmakers and researchers with this model are basic reproduction numbers and human feeding rates.

Dr. Manore also presented work on a new tool to produce a climate-integrated model of mosquito-borne infectious diseases. This work produced tools for:

- A habitat scaling method to capture vector population dynamics coupled with human population.
- Continental-scale short- and long-term forecasting of mosquito-borne diseases with a mechanistic model that can inform mitigations while incorporating heterogeneous data.
- Mosquito/vector/virus occurrence and species distribution maps; vector population density through time.
- A data-model fusion product across multiple political and grid scales and through time (daily/weekly).

Challenges include finding data that can be used specifically to parameterize mechanistic models, connecting scales with models (e.g., within-cell to within-host to population scale interactions), and incorporating sufficient heterogeneity.

Incorporating Biology/Ecology – Dr. Manore used the example of mosquito modeling to demonstrate the importance of incorporating factors that impact mosquito development at various stages, including temperature, water, and daylight.

Discussion

A participant inquired whether Dr. McDowell’s team has given any thought to incorporating data from the molecular and cellular scales. She replied that this is already being worked on, but there are no results to show publicly at the moment. She said this will be key to understanding how to put together what is going on within the cells, the host, and the vector at the molecular scale and how that translates to outcomes at the public health population level.

Another participant asked how Dr. McDowell’s team is making their models available to others. She said the team tries to provide code and documentation to make models as available as possible.

Researchers must determine whether the need is for graphical user interface (GUI) features for tweaks or information that will allow people to adapt the model at a deeper level. She said her team is creating modules so that people who have the skills and interest can slide modules in and out for various input. She concluded that the way forward will be to make sure models are usable for the wider community.

Session 6: Moderated Discussion: Crosstalk and Collaboration Among Vector-Borne Disease and Computational Biology/Modeling Researchers

For Session 6, the organizers transitioned from scientific updates from experts in the field to a more interactive format with the virtual audience. The session was divided into two segments, one with a 'vertebrate' and one with an 'invertebrate' focus. Session 6 proceeded as follows:

- Presenters gave a five-minute 'flash talk'. providing their perspective on comparing systems approaches in the vertebrate and invertebrate.
- These flash talks were followed by a 20-minute discussion.

Comparing and Contrasting Systems Approaches in the Vertebrate and Invertebrate – The Vertebrate Perspective
Rhoel Dinglasan, University of Florida

Dr. Dinglasan began by discussing how systems biology leverages biological insight at different scales—from cell to tissue to organism to ecology—and also from data to assay to intellectual property that can lead to interventions for public health. He said that in the end, systems biology is not just for biology's sake, but potentially for the development of public health interventions.

Dr. Dinglasan labeled the 'omics cascade (genome transcriptome → proteome → metabolome → phenotype) the "molecular network of networks" (Sexton *et al.*, *ACS Inf Dis* 2019). He asked attendees to think about how well they can visualize, interpret, and validate the data. Once models are developed to generate data, how do researchers go back and revalidate it, refine it, develop new models, and test those? He provided two examples of successful large interdisciplinary collaborations on malaria, then discussed the potential pitfalls of a collaborative systems biology approach:

- Trash in, trash out – If there is an incorrect experimental design for generation of the data type, the result is "trash out"—lots of data, but not necessarily good data.
- Non-tailored workflow – The core facility on which a lab must rely may have a workflow designed for cancer or some other work. Dr. Dinglasan said that his lab works seven to eight months to develop a workflow that fits with its system.
- Systems biology does not necessarily have to be 'omics-based, with capital-intensive instrumentation requirements. There are other data types that form part of systems biology.
- Glycomics – NIH has invested quite a bit in this field.
- Animal models – A lot of the basic systems (baseline information) have not been generated, so it becomes an expensive process. He suggested that NIH could develop resources so that more labs can take advantage of systems biology.
- Embrace variation – Variation can be an issue, especially for those who have moved from

reductionist science to systems biology. Replication time and instrumentation time are two examples.

Dr. Dinglasan also discussed opportunities:

- Multi-datatype integration is really the issue, not the fact that data analytics are different in respective disciplines. The challenge is how you bring them together.
- Researchers should bridge the biology-computational divide. If researchers do not have a good collaborator, they will ask for data that cannot be generated, causing friction. New opportunities are needed to enhance the dialogue between different disciplines.

Jishnu Das, University of Pittsburgh

Dr. Das noted that he spoke from a systems immunology lab perspective. He emphasized that systems immunology is bioinformatics. It is embedded between two communities—the statistical and informatics community and the immunology community. He said his team can speak both languages. He described his lab's work in using systems biology to study malaria vaccine-induced humoral responses using predictive machine learning to identify correlates, then observe how the correlates change with the vaccine dose. The team used prior biological knowledge to get from correlates to mechanisms. Researchers had to do mechanistic experiments to show that antibodies were functional.

Dr. Das then challenged participants with two concepts:

1. There are many multi-omic technologies, but how do researchers integrate them (transcriptomics, proteomics, and metabolomics) with demographic, genetic and clinical features—prediction vs. inference?
2. When researchers look at the generated data, how can they integrate these with prior-knowledge biological networks—data-driven vs. prior knowledge-driven approaches?

Multi-omic integration has become the way to go in precision medicine, according to Dr. Das, including for biomarker identification and study of disease processes. He named three broad areas for the types of input data and the types of methods used—based on data focused on prediction—identifying predictive correlates; based on inference beyond predictive correlates; and the mechanistic model.

Dr. Das expanded on the concept of interpretable machine learning. Predictive modeling simply generates correlates, he said. If researchers want to know the "how" and the "why" beyond the "what," inference is needed. The number of interpretable machine learning approaches that provide inference beyond prediction is limited. He concluded that in human systems immunology, researchers need inference beyond prediction. His own lab uses multi-omics integration to uncover both correlates and mechanisms using predictive and interpretable machine learning. These need to be integrated with prior knowledge biological networks, not just depend on the data generated in the context of a specific study.

Discussion

- One attendee commented that when it comes to machine learning and systems biology, researchers are not so much interested in predictions, but want to understand the reasons behind the predictions. If a transcriptomics study is done, for example, researchers want to know which genes at what levels of expression contributed to discern between cases and controls.

- Dr. Das commented on physics-based machine learning techniques. Although his lab does not deal much with biophysical approaches and sticks with data-driven as well as prior knowledge-based predictive and interpretable machine learning, his team works with collaborators who bring out these biophysical perspectives. He said his lab uses a technique called network propagation and has been quite interested in the insights provided by propagating signals derived from genomic, epigenomic, and transcriptomic datasets on prior knowledge networks using this physics-based approach.

- Researchers need to design experiments together when collaborating on a multi-investigator project. It is easier for data analysts and computer scientists to analyze results if they were involved in the design. Each discipline could learn a lot from the other by working together to figure out how things should be done from the beginning as a team. This is especially true when working from field samples that cannot be collected again easily.

- Participants discussed the integration of different ‘omics modalities/machine learning in non-model organisms, such as arthropod vectors. They commented that this is not much of an issue right now when compared to human systems immunology, but it is an emerging problem. The biggest challenge raised was the existence of reference and prior knowledge. Some of the interpretable machine learning approaches could still be brought to bear with the right data modalities, but careful consideration needs to be given to underlying power calculations and how noisy the data is.
- Participants discussed the dependency on core facilities and the use of workflows that are forced onto non-model organisms and possible solutions.
- Regarding the lack of protein-protein interaction networks in vector species, participants discussed how these data are needed to jump from transcriptomics to function. The problem is that there are not existing networks. There are correlation data based on transcriptomics, but those are not true networks that can be validated independently. The real protein-protein networks are needed, not information inferred from correlations.

Comparing and Contrasting Systems Approaches in the Vertebrate and Invertebrate: The Invertebrate Perspective

Utpal Pal, University of Maryland

Ticks are globally prevalent, with 900 species; 50 species can transmit many discrete infections. *I. scapularis* are the predominant disease vectors that are genetically distinct and yet are the most diverse and distributed globally.

Dr. Pal said that in systems biology approaches to address tick biology, researchers are fortunate to have eight reference genomes for five tick species. There are also various transcriptome and proteome databases of tick organs and cell lines. His team has started to work on improved *I. scapularis* genome, transcriptome, and proteome. Results include assembly of a complete genome, the first complete catalog of isoforms of any tick species with no transcript assembly required, and

14 chromosome level scaffolds. His team also developed a high-resolution whole tick proteome representing various feeding and developmental stages. Dr. Pal concluded by describing efforts to analyze new National Center for Biotechnology Information (NCBI) 103 (PalLab Reference Sequence) to identify loci that are the most different between northern and southern populations of *I. scapularis* and may be driving genetic differences in vectorial capacity.

Rushika Perera, Colorado State University
Dr. Perera provided a perspective on the state of metabolomics data acquisition, especially for the mosquito and the metabolic changes of pathogen infection. The metabolic environment of a mosquito changes upon exposure to any stimulus. The metabolome is impacting vector control strategies because of this, so if researchers want to understand how to reduce vectorial capacity or vector competence, they must understand how the metabolome changes. Dr. Perera said that many metabolomic studies have been done on whole mosquitoes, isolated organs, and cells that provide perspective on these changes.

- Whole mosquitoes – Results showed that three arboviruses had different metabolic requirements and profiles, that this was time-dependent, and that co-infection by multiple viruses might be possible because they have different metabolic requirements.

- Isolated organs – Infection of the midgut and salivary glands (sites of virus replication) of *A. aegypti* mosquitoes with dengue virus resulted in a change in 15 percent of the detected metabolome in the midgut, including a large change in the lipid metabolome (Chotiwan *et al., PLoS Pathogen* 2018). This provides a metabolic basis and support for other studies showing that mitochondria elongate during dengue infection.

- Cells – Dr. Perera noted that if one wants to understand how an enzyme or a pathway is altered, one must go back to the genome. She gave an example of fatty acid synthase, which has seven possible transcripts in the mosquito, so the redundancy of the *A. aegypti* genome is a problem. Researchers had to understand how many transcripts there are that are close to the human transcript, during what stage of lifecycle these genes are expressed, and what happens during a blood meal. Researchers were able to show that the first blood meal reduces the expression of the fatty acid synthase gene and a second blood meal enhances this process. This provides a possible metabolic basis for why a second blood meal can increase viral infection.

Dr. Perera highlighted other studies done on the mosquito that infer the metabolome is important. She concluded that researchers need to focus on the metabolome and finds ways to integrate metabolomic, transcriptomic, and proteomics.

Discussion
- Participants discussed the effect of genetic drifts between *I. scapularis* in different geographic regions and whether using ticks from lab colonies or from facilities such as the Centers for Disease Control and Prevention (CDC) may confound experimental data. Dr. Pal noted that transmission differences have been observed between ticks from different regions and researchers are addressing the high chance that there is a genetic reason for the difference.

- Dr. Perera was asked whether she is considering using vector strains that have midgut infection and escape barriers. Dr. Perera replied that her team has collected field mosquito lines and will be interesting to look at escape barriers and the metabolic basis for that escape barrier. She added that experimental design is critical because the effort put into design on the front end produces clean data, including in metabolomics.

- The discussion turned to the fact that there seem to be many elements that condition whether a population or individuals are going to be susceptible to or be able to transmit a pathogen. Being able to integrate the metabolomics, genetic, proteomic and other datasets provide a clearer picture as to what is contributing to the variation in populations. Researchers will be interested to see if the differences are observed not only in the presence or absence of dengue virus, but what are the metabolic differences between a mosquito in Thailand versus. a mosquito in Senegal versus a mosquito in Miami-Dade County. Computational biologists can help with integrating these datasets.

- Participants discussed Dr. Pal's single nucleotide polymorphisms data for tick clades in the U.S. Southeast and South-Central coasts. Dr. Pal said that his team is interested in pursuing whether the genetic differences have any link to the vectorial capacity. Dr. Pal added that the microbiome is one of the huge factors for vectorial competence or capacity.

Session 7: Roadblocks, Opportunities, and Next Steps

NIAID continued the interactive portion of the meeting by engaging with seven panelists. Dr. Shabman asked pre-prepared questions to each panelist and asked them to provide a 1-2 minute response per question. He then opened the floor for a general conversation before moving to the next question.

Session 7, Block 1 questions provided in advance:

Question 1 (roadblock): Can you describe your experience with developing and/or accessing large-scale datasets to conduct vector-borne disease research, and can you highlight **one** challenge in applying systems biology in VBD research?

Question 2 (opportunity): Are you aware of technological advances (either wet lab or informatics) that will facilitate systems analyses for VBDs?

Session 7, Block 2 questions provided in advance:

Question 1 (opportunity): How should scientists collaborate on VBD research, especially in the context of collaborations between pathogen-vertebrate and pathogen-invertebrate systems?

Question 2 (next steps): From your experiences, and from the meeting discussions, what would excite/incentivize you to establish new collaborations in the VBD space?

What is Needed (Data, Technologies, Tools) to Apply SysBio to Vector-Borne Diseases of Public Health Importance?

Carrie Manore, Los Alamos National Laboratory; Sukanya Narasimhan, Yale University; David Roos, (VEuPathDB); John Adams, University of South Florida

Dr. Shabman opened the panel discussion by posing the following question:

Can you describe your experience with developing and/or accessing large-scale datasets to conduct vector-borne research and can you highlight one challenge in applying SysBio to vector-borne disease research?

Dr. Adams – A current challenge is the annotation for genomes for things that are not called *falciparum*. His team is using the reference genome A1H1 and is finding a lot of pseudogenes that are not really pseudogenes, so the annotation is not as thorough as was done for the reference genome. Some focus on the genomes that are most frequently used would be helpful.

Dr. Manore – One challenge is getting appropriate metadata. Her lab is using data from different experiments, labs, and field locations, and putting it all together is always a challenge. It is also a challenge to get a time series of what is actually happening.

Dr. Roos – Data analysis and data integration are two different things. It was a shock to discover the difference between production database management, for example, and the kinds of things done in the research lab. The high value data that people generate need to be broadly accessible. That issue is a real challenge in all aspects of systems biology.

Dr. Narasimhan –The biggest challenge coming from the bench perspective is data management. There is so much 'omics data available to be integrated. It requires both accessing and mining the data. Her team relies on core facilities to do the analysis. Dr. Narasimhan recommended a hands-on workshop or have a designated group of bioinformatics experts who would be willing to collaborate.

> Dr. Roos responded that while there is no one-size-fits-all comprehensive course on data science, the VEuPathDB team is committed to try to make sure researchers have the resources they need. He referenced the organization's recent webinar series available on YouTube that highlighted resources for tick biology. Dr. Adams suggested the strategy of finding a partner to team with. Dr. Narasimhan noted that it can be difficult to contact experts on an individual basis who have the time for a collaboration, prompting her suggestion of a designated core group of experts. Dr. Roos noted that his organization runs a production contract to support researchers in the field and encouraged Dr. Narasimhan and others to contact the VEuPathDB outreach team.

Workshop participants weighed in on the subject with their own comments, including:

- Proteomics core facilities are not set up to do higher bioinformatics. There is a need for a systems biology core. Once you get to that higher level of data, it is a fairly flat field. Systems biologists need education too. Most of the data they are used to is genomic data. They need to learn how to use proteomics and metabolomics data.

- One participant commented that he finds it overwhelming to get started on accessing the supercomputers and getting up to speed with command lines, writing code and scripts, and choosing the best software. It is a challenge if one is not constantly on top of things.

- About project funding, NIH is often rightfully focused on hypothesis-driven research. There also needs to be space for exploratory research (not through R21 grants).

Dr. Shabman posed a last flash question to the panelists, asking them what technology they are most excited about:

Dr. Adams – Spatial transcriptomics. Applying the technology to whole genome screens. One worry is whether what is done in the laboratory has relevance to the field. Having good partners and being able to carry research in the lab to the field would be a great opportunity.

Dr. Manore – It is interesting to feed data from all the different ‘omics into the models and translate that as researchers try to understand what is happening on a population scale and make new connections.

Dr. Roos – What are most exciting are technologies to cope with the cross-silo problem of integrating data across databases—being able to use field-based detailed metadata to ask questions of genomic-type data. The opportunity is that systems biology is an inherently intellect-driven discipline rather than a capital-intensive discipline, which is a global leveling opportunity for junior researchers all over the world.

Dr. Narasimhan – Spatial transcriptomics and spatial proteomics are very exciting technologies.

Collaboration Opportunities in SysBio/Data Science for Invertebrate and Vertebrate Researchers

Martin Ferris, University of North Carolina-Chapel Hill; Karine Le Roch, University of California Riverside; Kevin Maringer, Pirbright Institute

Dr. Shabman began the discussion by posing the following question:

From your experience and from discussions at this meeting and others, what would excite or incentivize you to establish new collaborations in the vector-borne disease space?

Dr. Le Roch: Meet with people and find a common interest. For the past two years, there has been an issue with traveling and meeting with colleagues and the amount of collaboration significantly decreased. It is hoped with the end of COVID-19 restrictions, researchers can go back to meeting in person. Symposiums and workshops are critical.

Dr. Ferris: These meetings are important where researchers get to see what others are doing and how individual expertise might be able to help somebody answer a specific question, even if it is not necessarily high-dimensional data. He has experience in challenges integrating scientists from different fields, such as virologists and geneticists and data scientists, so he understands the challenges of getting integrative programs going.

Dr. Maringer: The biggest barrier is money is necessary to make collaborative efforts happen. There is an important role for funders. The research field is both small and global. For many people, it is challenging to find all the expertise they need for a project within one geographic area. What is needed are funders to do is work across borders and involve endemic countries. It is the people in the affected countries that know what the important and biological questions are.

Dr. Ferris added that the R01 mechanism often does not suffice. Larger programmatic mechanisms are needed to facilitate these interactions. Dr. Le Roch said that moving from large datasets into specific hypotheses is essential,an unbiased approach, and merits broader support.

Dr. Shabman posed another question to the panel:

In your opinion, do you think that the technology advances over the past decade in systems biology are happening fast enough where you can do an unbiased approach and get to a hypothesis because you have some results, as opposed to doing 'omics on samples. Do you see progress there?

Dr. Le Roch: Sequencing technology continues to improve. It is time to start thinking about generating as many genomes as possible to get back to what is happening in the field.

Dr. Maringer: All of his grants include proteomics and transcriptomics. It is usually used in a hypothesis-driven way. It is not completely unbiased. His grants are reviewed on an international level. The datasets that generated are maybe not as grand as mapping every action in the mosquito genome, but they do generate data that is still valuable to the community.

Dr. Ferris: Even with technological advances, it would be nice to have a recognition among reviewers about timelines—if it is a new type of approach or if researchers are working with a non-referenced genome organism there is a lag, but that lag is actually valuable in leading to those hypotheses. It is setting up the fundamental frameworks that will then be used fieldwide.

Participants added their own comments:

- A lot of genome annotations can be hit or miss in terms of quality. It can be difficult to put together a comprehensive annotation of a genome. It would be helpful if there were a mechanism by which a large structure effort to do large-scale genome annotations was available.

- There are some interesting challenges with improving annotations because there is no one-size-fits-all solution. Discussions have been ongoing among bioinformatic resource center investigators and others that are intended to try to provide more effective automated approaches for improved annotations that would also be consistent across sites.

- Those who assist researchers with data for no compensation need to be given credit in some way for their efforts, since they are not published.

Meeting Summary and Next Steps
Adriana Costero-Saint Denis, NIAID/DMID, and Reed Shabman, NIAID/DMID

Meeting Summary

Day 1: A major theme was multi-disciplinary, trans-disciplinary research (across viruses, bacteria, and eukaryotic VBD) and pathogen research on invertebrate and vertebrate hosts.

Day 2: Technology highlights in VBD research included challenges with data analysis and integration, networks that work on non-model organisms, core facilities that can provide services for non-model organisms, high-quality modeling across scales, and enhanced communication for informatics and experiments, especially at the planning stages.

Next Steps

- Meeting summary report will be shared on the Teams Channel and available upon request in no later than six months.
- Considering possible publication about bringing systems biology approaches to vector-borne diseases.
- NIAID consideration of how to facilitate multi-disciplinary collaborations between scientists, including future workshops that focus on specific topics raised at the current workshop. Dr. Costero-Saint Denis encouraged participants to email her or Dr. Shabman or contact them through the Teams account with ideas.

Adjourn

Appendix A: Agenda

Meeting Agenda
Information website link: https://cvent.me/18mbN9

Time (EST)	Session Title	Speaker/Institution
Day 1, May 17, 2022 11:00am-3:00pm		
11:00am	Welcome and Introduction	Adriana Costero-Saint Denis (NIAID/DMID)
11:05am	The importance of systems biology approaches to vector-borne disease research to NIAID	Cristina Cassetti (Deputy Director, NIAID/DMID)
11:12am	Meeting Agenda Rationale	Reed Shabman (NIAID/DMID)
11:15am-12:05pm	1. Keynotes	Moderator: Organizers (NIAID/DMID)
11:15am-11:40am	An overview of systems biology approaches and their use in vector-borne disease research - Modulation of human innate immune responses by arboviruses	Ana Fernandez-Sesma (Icahn School of Medicine at Mount Sinai)
11:40am-12:05pm	Cutting-edge SysBio technologies and potential applications for vector-borne diseases - Systems immunology of human immune variations: do differences make a difference?	John Tsang (NIAID, Multiscale SysBio Center)
12:05pm-12:55pm	2. Vector Borne Disease Systems Biology & Modeling in the Vertebrate & Invertebrate Host: Flavivirus as a Use Case	Moderator: Ashley St. John (Duke-Nat University Singapore)
12:05pm-12:25pm	The Vertebrate Host-Flavivirus Interface- Let's get physical: a Zika virus-host protein interaction in replication and pathogenesis	Priya Shah (University of California, Davis)
12:25pm-12:45pm	The Invertebrate Host-Flavivirus Interface - Integrating different 'omics approaches with molecular tools to study flavivirus-vector interactions that drive transmission and emergence	Kevin Maringer (The Pirbright Institute, UK)
12:45pm-12:55pm	Discussion	
12:55pm-1:10pm	Break (Optional Networking)	
1:10pm-1:55pm	3. Vector-borne Disease Systems Biology & Modeling in the Vertebrate and Invertebrate Host: *Borrelia* as a use case	Moderator: Sukanya Narasimhan (Yale University)
1:10pm-1:30pm	The Vertebrate Host-Borrelia Interface	Robert Moritz (Institute for Systems Biology)

Time (EST)	Session Title	Speaker/Institution
1:30pm-1:50pm	The Invertebrate Host-Borrelia Interface – Tick Immunobiology and Microbial Interactions	Joao Pedra (University of Maryland)
1:50pm-2:00pm	Discussion	
2:00pm-3:00pm	4. Vector-borne Disease Systems Biology & Modeling in the Vertebrate & Invertebrate Host: Plasmodium as a Use Case	Moderator: Karine Le Roch (University of California, Riverside)
2:00pm-2:20pm	The Invertebrate Host-Plasmodium Interface – Using scRNAseq to understand *P. falciparum* development and evolution	Virginia Howick (University of Glasgow)
2:20pm-2:40pm	The Vertebrate Host-Plasmodium Interface – Systems analysis of *P. falciparum* in relation to parasite density and severity of disease	Karen Day & Michael Duffy (University of Melbourne)
2:40pm-3:00pm	Discussion & End of Day 1	Informal networking in breakout rooms until 3:30pm
Day 2, May 18, 2022 11:00am-2:30pm		
11:00am-11:10am	Day 1 Recap	Organizers (NIAID/DMID)
11:10am-12:10pm	5. Technological Advances in Systems Analyses: Implications for the Vector Borne Disease Community	Moderator: Priya Shah (University of California, Davis)
11:10am-11:30am	Advances in "Omics" as it relates to the vector, pathogen, vertebrate host – Unmasking ammonia metabolism in A. aegypti mosquitoes with mass spectrometry-based metabolomics	Patricia Y. Scaraffia (Tulane University)
11:30am-11:50am	Advances in data integration and analytics – the VEuPathDB Family of Knowledge Bases	Mary Ann McDowell (University of Notre Dame)
11:50am-12:10pm	Advances in modeling development and application – Advances in Multi-scale Modeling for Vector-borne Diseases: Integration and Data Fusion to Enable Prediction	Carrie Manore (Los Alamos National Laboratory)
12:10pm-12:20pm	Discussion	
12:20pm-1:20pm	6. Moderated Discussion: Crosstalk and collaboration among vector-borne disease and computational biology/modeling resaerchers	

Time (EST)	Session Title	Speaker/Institution
12:20pm-12:50pm	Comparing and contracting systems approaches in the vertebrate and invertebrate – the vertebrate perspective *Dr. Dinglasan and Dr. Das will give two five-minute flash talks followed by a 20-minute discussion*	Rhoel Dinglasan (University of Florida) Jishnu Das (University of Pittsburgh)
12:50pm-1:20pm	Comparing and contrasting systems approaches in the vertebrate and invertebrate – the invertebrate perspective *Dr. Pal and Dr. Perera will give two five-minute flash talks followed by a 20-minute discussion*	Utpal Pal (University of Maryland) Rushika Perera (Colorado State University)
1:20pm-1:30pm	Break	
1:30pm-2:30pm	7. Panel discussion: Roadblocks, opportunities, and next steps	
1:30pm-2:00pm	What is needed (data, technologies, tools) to apply SysBio to vector-borne diseases of public health importance?	Carrie Manore (Los Alamos National Laboratory) Sukanya Narasimhan (Yale Unviersity) Davis Roos (VEuPathDB) John Adams (University of South Florida)
2:00pm-2:30pm	Collaboration opportunities in SysBio/Data Science for invertebrate and vertebrate researchers	Martin Ferris (University of North Carolina, Chapel Hill) Karine Le Roch (University of California, Riverside) Kevin Maringer (Pirbright Inst)
2:30pm-2:45pm	Meeting Summary and next steps	Organizers (NIAID/DMID)
2:45pm	Adjourn	