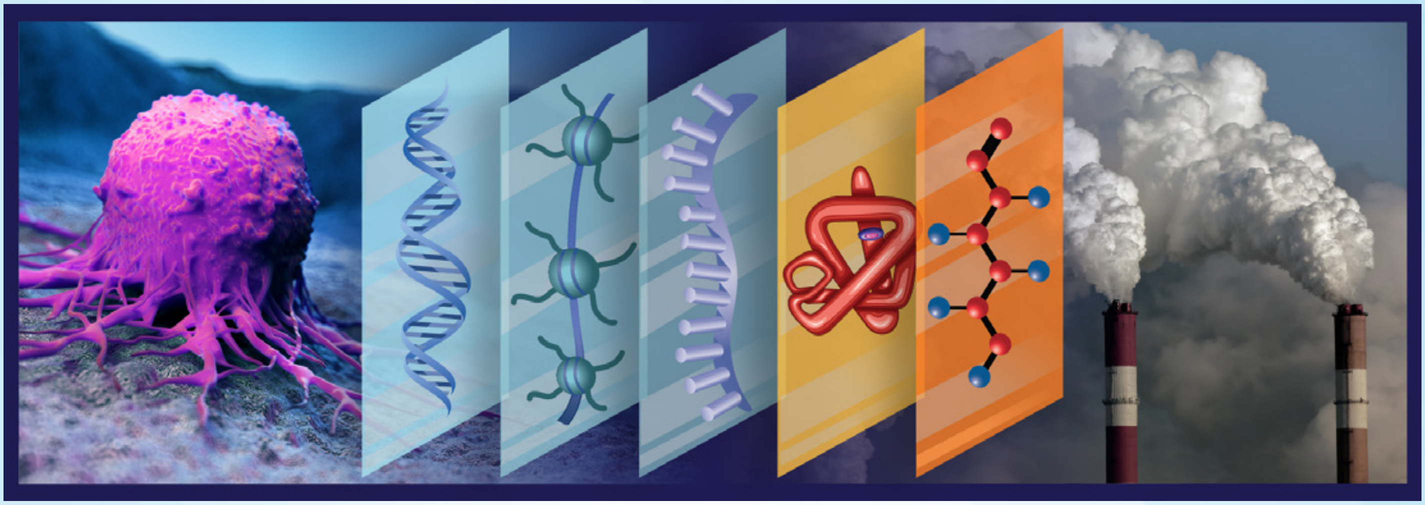
NIH National Institute of Environmental Health Sciences

Integrating Environmental Data With Other Omics for Cancer Epidemiology



Workshop Report

National Institutes of Health • U.S. Department of Health and Human Services

This workshop report was prepared by Lauren Sprouse, MDB, Inc. This work has been funded with federal funds from the National Institute of Environmental Health Sciences, National Institutes of Health, under Contract No. HHSN273201800006I. The statements, conclusions, and recommendations contained in this document reflect opinions of the workshop participants and are not intended to represent the official position of the National Institute of Environmental Health Sciences, the National Cancer Institute, the National Institutes of Health, or the U.S. Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

The author thanks Kimberly McAllister, Leah Mechanic, Julie Leibach, and workshop participants for their review of earlier versions of this workshop summary.

The workshop planners gratefully acknowledge the contributions of all workshop participants, listed in Appendix 1, for their contributions to workshop deliberations and outcomes.

Suggested citation: Integrating Environmental Data with Other Omics for Cancer Epidemiology. National Institute of Environmental Health Sciences & National Cancer Institute, February 14-15, 2023. https://www.niehs.nih.gov/news/events/pastmtg/2023/environmental_data/index.cfm

Table of Contents

List of Abbreviations and Acronyms

BMI	Body mass index
CEDCD	Cancer Epidemiology Descriptive Cohort Database
CEECR	Cohorts for Environmental Exposures and Cancer Risk
DCCPS	Division of Cancer Control and Population Sciences
EHR	Electronic health record
FAIR	Findable, Accessible, Interoperable, Reusable
GWAS	Genome-wide association studies
GxE	Gene–environment
MR	Mendelian randomization
NCI	National Cancer Institute
NHGRI	National Human Genome Research Institute
NIEHS	National Institute of Environmental Health Sciences
NIH	National Institutes of Health
PFAS	Per- and polyfluoroalkyl substances
PM	Particulate matter
PRS	Polygenic risk score
RCT	Randomized controlled trial
SNP	Single nucleotide polymorphism
WGCNA	Weighted gene co-expression network analysis

Glossary

Term	Definition
Central dogma	Developed by Francis Crick in 1958, the central dogma of molecular biology is a theory stating that genetic information flows only in one direction: from DNA, to RNA, to protein, or from RNA directly to protein. Scientists have since discovered several exceptions to the theory — prions, for example, are infectious proteins that replicate without going through DNA or RNA intermediates.[1]
Data harmonization	Data harmonization refers to all efforts to combine data from different sources and provide users with a comparable view of data from different studies.[2] Data harmonization and curation improves overall data quality through careful review of processes and procedures related to data collection, documentation, and interim analyses that can identify potential differences in administration, collection, procedural differences, scoring and measurement, and missing data.

[1] Central Dogma. NHGRI Talking Glossary of Genomic and Genetic Terms. Available at https://www.genome.gov/genetics-glossary/Central-Dogma

[2] Inter-university Consortium for Political and Social Research, Data Sharing for Demographic Research. Data Harmonization. Available at https://www.icpsr.umich.edu/web/pages/DSDR/harmonization.html

Term	Definition
Epigenomics	Derived from Greek, “epigenome” means “above” the genome. The epigenome consists of chemical compounds that modify, or mark, the genome in ways that tell it what to do, where to do it, and when to do it. The marks, which are not part of the DNA itself, can be passed on from cell to cell as cells divide, and from one generation to the next.[3] Epigenomics is seen as the interface between genes and environment.
Exposome	The exposome can be defined as the measure of all the exposures of an individual in a lifetime and how those exposures relate to health. Exposures can begin before birth and include insults from environmental and occupational sources. Understanding how exposures from our environment, diet, lifestyle, etc. interact with our own unique characteristics such as genetics, physiology, and epigenetics to impact our health is how the exposome will be articulated.[4]
Gene–environment interaction (GxE)	Gene–environment interaction refers to the interplay of genes (and, more broadly, genome function) and the physical and social environment. These interactions influence the expression of phenotypes. For example, most human traits and diseases are influenced by how one or more genes interact in complex ways with environmental factors, such as chemicals in the air or water, nutrition, ultraviolet radiation from the sun, and social context.[5]
Gene module	A gene module is a group of genes showing a concordant change in expression under a given set of circumstances.[6] Gene modules can also be defined based on existing knowledge, such as known metabolic or signaling pathways or protein–protein interactions. Once defined, these modules can be used as the basis for analyzing any biological question.[7]
Genetics	Genetics is a term that refers to the study of genes and their roles in inheritance and explores how specific traits or conditions are biologically passed down from one generation to another. Genes (units of heredity) carry the instructions for making proteins, which direct the activities of cells and the functions of the body. Certain medical conditions, such as cystic fibrosis, Huntington's disease, and phenylketonuria (PKU), are caused by mutations, or alterations, in a single gene.[8]

[3] According to the Fact Sheet on Epigenomics published by the National Human Genome Research Institute: http://www.genome.gov/27532724

[4] National Institute for Occupational Safety and Health (NIOSH). Exposome and Exposomics. Available at https://www.cdc.gov/niosh/topics/exposome/

[5] Gene Environment Interaction. NHGRI Talking Glossary of Genomic and Genetic Terms. Available at https://www.genome.gov/genetics-glossary/Gene-Environment-Interaction

[6] DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science. 1997;278(5338):680-6. doi: 10.1126/science.278.5338.680.

[7] Keidar Haran T, Keren L. From genes to modules, from cells to ecosystems. Mol Syst Biol. 2022;18(5):e10726. doi: 10.15252/msb.202110726.

[8] Frequently Asked Questions About Genetic and Genomic Science. NHGRI. Available at http://www.genome.gov/19016904

Term	Definition
Genome-wide association studies (GWAS)	Genome-wide association studies involve scanning the genomes of many people to find genetic variations associated with a particular disease. Once new genetic associations are identified, researchers can use the information to develop better strategies to detect, treat, and prevent the disease. Such studies are particularly useful in finding genetic variations that contribute to common, complex diseases, such as asthma, cancer, diabetes, heart disease, and mental illnesses.[9]
Genomics	Genomics, a newer term than genetics, describes the study of all of a person's genes (the genome), including interactions of those genes with each other and with the person's environment. Genomics includes the scientific study of complex diseases such as heart disease, asthma, diabetes, and cancer, because these diseases are typically caused more by a combination of genetic and environmental factors than by individual genes. Genomics is offering new possibilities for more targeted therapies and treatments for complex diseases, as well as new diagnostic methods.[10]
Heterogeneity	Heterogeneity in data broadly refers to differences within individual samples, between samples, and between experimental results. Environmental data poses challenges due to its heterogeneity, meaning it can be highly variable in location, measure, and analysis, as well as across the life course and time of day. Human populations are also heterogeneous in their genomes and exposomes, requiring large data sets to account for diversity among populations.
Mendelian randomization (MR)	Mendelian randomization is a method of using genetic data to infer causal relationships between exposure and outcome in conventional epidemiological (observational) studies. Mendelian randomization studies examine how genetic differences affect the way people's bodies react to certain behaviors, environments, or other factors, thus leading to specific health outcomes. This is achieved through the properties of genetic variants that render them not susceptible to reverse causation and confounding, which otherwise pose issues in epidemiological studies.[11]
Metabolomics	Metabolomics is the study of the biological metabolic profile of a cellular specimen in a specific environment at an isolated timepoint. This discipline depicts the physiological states of cells and organisms by focusing on carbohydrates, lipids, and other metabolites. Several analytical techniques are utilized to quantify the metabolic content of specimens such as mass spectrometry and electrophoretic applications.[12]

[9] According to the Fact Sheet on Genome-Wide Association Studies published by NHGRI: https://www.genome.gov/about-genomics/fact-sheets/Genome-Wide-Association-Studies-Fact-Sheet

[10] Frequently Asked Questions About Genetic and Genomic Science. NHGRI. Available at http://www.genome.gov/19016904

[11] CDC Genomics & Precision Health. Mendelian Randomization: Using Genetics to Study Behaviors and Environments that Cause Disease. Available at https://www.cdc.gov/genomics/disease/mendelian_randomization.htm

[12] Derived from the NCI Thesaurus found at https://ncit.nci.nih.gov/ncitbrowser/pages/multiple_search.jsf?nav_type=terminologies

Term	Definition
Omics	Omics has been defined as the study of related sets of biological molecules in a comprehensive fashion.[13] Examples of omics disciplines include genomics, transcriptomics, proteomics, metabolomics, and epigenomics. For the purposes of this workshop, a broad definition of omics includes the integration of various environmental exposures measures (encompassing exposomics) and resulting phenotypes with these other omics disciplines.
Organoid	A tiny, three-dimensional mass of tissue that is made by growing stem cells (cells from which other types of cells develop) in the laboratory. Organoids that are similar to human tissues and organs, or to a specific type of tumor, can be grown. Organoids are used in the laboratory to study how normal tissues or cancers form and to test new drugs and other types of treatment before they are given to people.[14]
Phenotype	The set of observable characteristics of an individual resulting from the interaction of the individual's genes with the environment. Phenotype can refer to anything from a common trait, such as height or hair color, to the presence or absence of a disease.[15]
Polygenic risk score (PRS)	Polygenic risk scores provide a measure of an individual's disease risk due to their genes. The score considers each version of a gene — related to a specific disease — that the individual has. Combining polygenic risk scores with other factors that affect disease risk, such as environmental exposures, can give a better idea of how likely an individual is to get a specific disease than considering either alone.[16]
Precision Environmental Health	Analogous to Precision Medicine, where treatments are tailored to the individual and specific disease, Precision Environmental Health has the goal of individualized risk assessment and interventions to prevent disease. Precision Medicine and Precision Environmental Health are similar in the use of omics data but differ in that the former has the goal of disease treatment, while the latter focuses on disease prevention. The ultimate goal of Precision Environmental Health is to reduce the adverse health effects of exposures through the air we breathe, the water we drink, and the food we eat by identifying individuals who are specifically susceptible to environmental threats and enabling precise, targeted, and effective prevention.[17]
Proteomics	The study of the structure and function of proteins, including the way they work and interact with each other inside cells.

[13] IOM (Institute of Medicine). 2012. Evolution of Translational Omics: Lessons Learned and the Path Forward. Washington, DC: The National Academies Press.

[14] NCI Dictionary of Cancer Terms. Available at https://www.cancer.gov/publications/dictionaries/cancer-terms/def/organoid

[15] Phenotype. NHGRI Talking Glossary of Genomic and Genetic Terms. Available at https://www.genome.gov/genetics-glossary/Phenotype

[16] Polygenic Risk Scores. CDC, Genomics & Precision Health. Available at https://www.cdc.gov/genomics/disease/polygenic.htm

[17] Walker CL, Dolinoy D, Baccarelli A. Perspectives on Precision Environmental Health. National Advisory Health Sciences Council, February 2021.

Term	Definition
Single nucleotide polymorphism (SNP)	A single nucleotide polymorphism (abbreviated SNP, pronounced “snip”) is a type of genetic variation where one nucleotide — a component of DNA — is replaced with a different nucleotide. These changes may cause disease, and may affect how a person reacts to bacteria, viruses, drugs, and other substances.[18]
Social determinants of health	Social determinants of health are the conditions in the environments where people are born, live, learn, work, play, worship, and age that affect a wide range of health, functioning, and quality-of-life outcomes and risks. Examples include housing, transportation, education, income, discrimination, access to nutritious foods, and environmental pollution.[19]
Transcriptome	A transcriptome is the full range of messenger RNA, or mRNA, molecules expressed by an organism. The term “transcriptome” can also be used to describe the array of mRNA transcripts produced in a particular cell or tissue type.[20]
Weighted gene co-expression network analysis (WGCNA)	Weighted gene co-expression network analysis (WGCNA) is a bioinformatics application for exploring relationships between different gene sets (modules), or between gene sets and physical health attributes. WGCNA uses hierarchical clustering to identify modules and then relates those modules to phenotypes. WGCNA provides straightforward, biologically functional interpretations of gene network modules.[21]

[18] NCI Dictionary of Cancer Terms. Available at https://www.cancer.gov/publications/dictionaries/cancer-terms/def/snp

[19] U.S. Department of Health and Human Services. Healthy People 2030. Available at https://health.gov/healthypeople/priority-areas/social-determinants-health

[20] According to Nature’s “Scitable” at: http://www.nature.com/scitable/definition/transcriptome296

[21] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559. doi: 10.1186/1471-2105-9-559.

Executive Summary

On February 14 and 15, 2023, the National Institute of Environmental Health Sciences (NIEHS) and the National Cancer Institute (NCI) convened a virtual workshop called Integrating Environmental Data with Other Omics for Cancer Epidemiology. Co-chaired by Kimberly McAllister, Ph.D., of NIEHS, and Leah Mechanic, Ph.D., MPH, of NCI, the workshop brought together a multidisciplinary group of experts to explore challenges and opportunities related to the integration of environmental exposure data with other omics data for cancer studies in human populations. The workshop sought to inform future research directions for NIEHS and NCI, as well as address key questions related to developing new methods and computational approaches for advancing knowledge in the field of multi-omics, using appropriate study designs for cancer molecular epidemiology studies that incorporate measures of environmental exposures, and enhancing experimental models of environmental exposures to use in conjunction with human multi-omics studies to enhance our understanding of complex human diseases.

There are many known environmental factors that can lead to cancer — for example, arsenic, ultraviolet radiation, and tobacco smoke — and some of the best described gene–environment (GxE) interaction examples are for cancer outcomes. Though many GxE relationships are well established, integrating environmental data with other omics, including metabolomics, proteomics, and transcriptomics, poses challenges related to the heterogeneity of environmental data, measurements with different instruments at different scales, and temporality of exposures. There is a need for analytical tools, techniques, and established best practices for integrating various environmental data with other omics data and for evaluating methods for integration, which is particularly challenging in longitudinal cancer epidemiology studies.

Cancer consists of over 100 diseases that can arise from nearly every cell type in the body. Once an individual is exposed to a cancer-causing environmental agent, cancer can take many years to develop, increasing the complexity of studying the role of environmental factors in cancer development. While there are challenges in collecting, managing, and integrating environmental, phenotypic, and genomic data, there are also opportunities to involve multiple disciplines and team science.

Day 1 of the workshop began with introductory presentations from Trevor Archer, Ph.D., Deputy Director of NIEHS, and Gary Ellison, Ph.D., MPH, Deputy Director of the Division of Cancer Control and Population Sciences (DCCPS) at NCI, who highlighted challenges with environmental data and cancer epidemiology, respectively. Drs. McAllister and Mechanic then discussed the workshop's purpose and intended outcomes: to identify the challenges and opportunities related to the integration of environmental exposure data with other omics data for human cancer populations, to establish best practices and develop methods focused on the integration of omics with other environmental exposure data, and to inform future supported research directions for NIEHS and NCI.

The workshop was organized into four main sessions over two days, with an additional closing session on Day 2 to integrate themes from the prior sessions.

Session I, moderated by Ulrike Peters, Ph.D., MPH, of the University of Washington Fred Hutchinson Cancer Center, covered the role of tumor tissue, long latency periods, and exposures during susceptibility windows in cancer initiation and promotion.

Mary Beth Terry, Ph.D., of the Columbia University Mailman School of Public Health, presented on the role of study design in research on environmental exposures and cancer susceptibility. According to Dr. Terry, all samples are not equal, and simply having samples for omics research without the context of when they were collected and in whom will lead to biased and underpowered research results. Additionally, because most cancers are identified later in life, population-based research studies tend to be insufficient to detect GxE interactions. Potential solutions include establishing new cohorts of younger, more diverse individuals that are enriched for cancer risk. The session also indicated that study designs should capture windows of susceptibility, and that oversampling groups underrepresented in omics research is key for reducing cancer burdens.

Genevieve Leyden, Ph.D., of the University of Bristol Medical School, presented on understanding pathways between body mass index (BMI) and tissue-specific cancer risk using a Mendelian randomization (MR) approach, which uses genetic data to infer causal relationships in observational epidemiological studies. Dr. Leyden studied the effects of BMI on different cancer outcomes by obtaining over 900 genetic variants identified for BMI genome-wide and relating them to genetic variants for diseases identified in other genome-wide association studies (GWAS) from similar populations. Results showed that variants related to BMI found in brain tissue samples were predominantly implicated in lung cancer, while BMI-related variants from adipose tissue samples were implicated in endometrial cancer.[22] According to Dr. Leyden, the results of this study provide an example of how multi-omics datasets can be used to help us get closer to understanding the relationship between a complex lifestyle risk factor, such as obesity or high BMI, on site-specific cancer risk.

Session II was moderated by Andrea Baccarelli, M.D., Ph.D., of the Columbia University Mailman School of Public Health. The session focused on methods to deal with heterogeneity and high dimensionality of environmental data, as well as differences in data scales and collection methods.

James Gauderman, Ph.D., of the University of Southern California Keck School of Medicine, discussed various approaches for assessing GxE interactions. Using data from the Functionally Informed GxE Interaction (FIGI) study, Dr. Gauderman and team conducted a genome-wide scan to determine how genes and environment, separately and in combination, impacted colorectal cancer outcomes, using a two-step approach.[23] Another approach for assessing GxE interactions is to use polygenic risk scores (PRS), which provide a measure of an individual's disease risk due to their genes. According to Dr. Gauderman, discovering GxE interactions will provide insight into the underlying carcinogenic mechanisms affected by established environmental risk factors, and in turn help identify potential targets for future interventions.

Cristian Coarfa, Ph.D., of the Baylor College of Medicine, presented on the Houston Hurricane Harvey Health (Houston3H) study to assess health outcomes, chemical exposures, and the microbiome in a cohort of households exposed to Hurricane Harvey. The researchers used a variety of data collection methods, including questionnaires, geographic information, and wristband samplers, as well as household swabs and nasal, saliva, and stool biospecimens. Dr. Coarfa and team assessed correlations among data on Harvey exposure, diet, demographic, health, and pollution burden, and found that most correlations with chemical exposures were found in Black study participants. Exposure to particulate matter showed the strongest correlation among participants in East Houston who experienced flooding. According to Dr. Coarfa, the Houston3H study showed

[22] Leyden GM, Greenwood MP, Gaborieau V, Han Y, Amos CI, Brennan P, Murphy D, Davey Smith G, Richardson TG. Disentangling the aetiological pathways between body mass index and site-specific cancer risk using tissue-partitioned Mendelian randomisation. Br J Cancer. 2023;128(4):618-625. doi: 10.1038/s41416-022-02060-6.

[23] Kawaguchi ES, Kim AE, Lewinger JP, Gauderman WJ. Improved two-step testing of genome-wide gene-environment interactions. Genet Epidemiol. 2023;47(2):152-166. doi: 10.1002/gepi.22509.

the utility of multi-omics environmental profiling, and future work should plan for the use of more powerful technologies to assess other omics, including metabolomics and proteomics.

Session III, moderated by Gary Miller, Ph.D., of the Columbia University Mailman School of Public Health, sought to inform our understanding of environmentally relevant cancer outcomes, particularly timing, dose, and mixtures challenges.

Dean Jones, Ph.D., of the Emory University Winship Cancer Institute, presented ways to apply integrative omics using model systems to environmental epidemiology research on cancer. Improvements in high-resolution metabolomics allow us to measure up to a million chemical signals in a small biological sample. Meanwhile, omics-scale exposome analyses provide an opportunity to explore the multiplicity of exposures and complexities of their interactions. We have genomics, proteomics, and metabolomics capabilities that must be aligned with immunologic, microbiome, and exposome measures to move the field forward. Environmental health research, together with cell and rodent studies, has established the utility of integrating untargeted mass spectrometry with other omics data. Model system studies with nontargeted exposomics methods can reveal biologically relevant chemicals that have escaped discovery and what their functions are. Experimental models can then be integrated with observational studies to link chemical exposures to health outcomes.

Carmen Marsit, Ph.D., of the Emory University Rollins School of Public Health, discussed several examples of omics integration in studies of the environment on human health, including studies identifying potential mechanisms linking exposure to per- and polyfluoroalkyl substances (PFAS) and metal mixtures to poor birth outcomes. While studies integrating environmental data with other omics have challenges and limitations, including temporality of exposures and ability to access target tissue, there is a real opportunity to demonstrate downstream biological effects to prove how social determinants of health are causal agents in disease development. Integrating omics with environmental data can inform mechanisms underlying links between the environment and health, provide direction for public health interventions, inform exposure histories, and help with individualized risk assessment to prevent adverse health effects — goals aligned with Precision Environmental Health.

Session IV was moderated by Stephen Montgomery, Ph.D., of Stanford University. The session covered how advances in in vitro functional genomics and model organisms might be used to inform multi-omics human studies related to environmentally sensitive cancers.

William Hill, Ph.D., of The Francis Crick Institute, proposed a mechanism of action for non-small cell lung cancer in people who have never smoked, suggesting that carcinogens act via non-mutagenic mechanisms, perhaps on preexisting mutations present in normal tissues. Given that non-small cell lung cancer cases among those who have never smoked are increasing, there is a clinical need to understand the biological underpinnings of this patient group. Using mouse models, Dr. Hill explored whether air pollution acted as a tumor promoter in lung tissue and found that particulate matter likely induces inflammation and tumor initiation among cells with preexisting mutations. According to Dr. Hill, this initial study to develop mouse models will allow them to dive deeper to explore which chemicals induce cancer.

Matthew Devall, Ph.D., of the University of Virginia, discussed colon organoids as a useful model for studying environmental factors that increase risk of colorectal cancers. In a study assessing the risk of colorectal cancer from smoking and eating red or processed meat, Dr. Devall and team treated independent organoid lines with a cocktail of carcinogens commonly found in cigarette smoke and red meat. They identified over 700 differentially expressed genes in response to carcinogen exposure. According to Dr. Devall, genes rarely act in isolation, and single-gene analysis was associated with a high false discovery rate. They identified consistent modules associated with smoking carcinogens, highlighting the impact of carcinogens on epithelial cells of the colon directly. Though there are challenges in enhancing representation in colon organoid studies, Dr. Devall and team are prioritizing sample collection from younger age groups and African American individuals, as they have been underrepresented in colon organoid studies.

In the workshops closing sessions, Chirag Patel, Ph.D., of Harvard University Medical School, summarized the prior workshop sessions. For an exposure to be considered as cancer-causing, it must induce a biological change and response, according to Dr. Patel. Multi-omics can be used as indicators of this exposure-cancer biology. The final workshop discussion was moderated by Kari Nadeau, M.D., Ph.D., of the Harvard University T.H. Chan School of Public Health. Drs. Patel and Nadeau highlighted key discussion themes from throughout the workshop.

Workshop Discussion Themes

Several themes arose during workshop discussions:

- **Study design challenges**. Participants cited timing of exposure, long latency periods, prioritizing sample quality, and optimal sample sizes as key study design challenges. Although case-control study designs will likely remain an important part of epidemiological studies, measuring exposure at the time of disease diagnosis provides little information in terms of causality. Participants noted difficulty in identifying optimal sample sizes, particularly for rarer cancers. In some cases, prioritizing sample quality over quantity — for instance, capturing very detailed omics data on fewer participants — is preferable. Existing cohorts for cancer studies are also skewed toward older individuals. Establishing younger and more diverse cohorts, capturing windows of susceptibility, and oversampling groups underrepresented in omics research are key steps for reducing cancer burdens.
- **Scientific goals should control choice of analysis**. Biological hypotheses need to be clearly defined prior to initiating studies. Researchers need to take their time to understand the steps involved in analysis and what might be gained or lost by prioritizing some dimensions over others.
- **Environment drives cancer risk**. Given the results of twin studies, it is clear that the environment, rather than genes, drives cancer risk. There are challenges in measuring other types of environmental exposures, such as social stressors, as well as in combining different measurement scales for different types of exposures.
- **Data harmonization**. Measuring multiple omics in the same individuals, same tissues, at the same time, and correlating different omics signatures across samples is one step to harmonizing datasets. Resources should be dedicated to the weakest link in the model because that is where most power loss will occur. Questionnaires should also be standardized so other researchers can use them to collect data in the same way, making it easier to harmonize data.
- **Developing a cancer exposome atlas**. Creating a cancer exposome atlas — a map of how specific exposures affect different tissues and developmental stages — is essential, especially when incorporating multi-omics to understand cancer risk. However, no coordinated effort to do so exists, and consensus is lacking regarding which methods should be used for analyzing the exposome, and what the definition of "exposome" is. Collecting samples for a variety of omics analysis and uploading them to a platform to look at concurrently could spark the beginning of an atlas.
- **Assessing social determinants of health**. There is a need to establish causal biological pathways underlying the influence of social structures on health. Larger and higher-quality life history data are needed to link social stressors to cancer outcomes but scaling and logistics present challenges. An

increasing wealth of geospatial data can provide information on social determinants of health, and those factors can then be linked to cancer disparities.

- **Coupling experimental models with population-level studies**. Experimental models can be coupled with human population studies by elucidating mechanisms of action, though challenges persist. Colon organoid studies, for example, pose challenges with representation in terms of age and race and ethnicity. There are also challenges in connecting mouse models to human studies, where timing, dose, and the effects of cumulative exposures are difficult to replicate and validate. Next steps to understanding the impacts of exposure mixtures on health will take a lot of coordination.
- **Current and future technologies**. With where we are in omics research, it may be worth taking a step back to consider what technologies are needed to reach our goals. Analytical technology will evolve, so it is important to understand the other side of the equation — study populations and samples — and characterize them correctly in order to develop better analytical approaches. Conceptual frameworks can help us understand which environmental factors impact each other and downstream omics.
- **Coordination among research communities**. Participants emphasized the importance of collaboration across disciplines to identify environmental exposure factors of interest and improve experimental models. Leveraging findings from existing studies can also benefit data harmonization.
- **Reliability, repeatability, and reproducibility.** Studies should begin with the end in mind (e.g., policy change, health outcomes). As with all scientific research, omics studies demand reliability, repeatability, and reproducibility — reproducible results inform approval and regulatory processes. However, there are challenges with exposures that only affect a small fragment of the population, as results may be difficult to reproduce.

Suggested Next Steps

Workshop participants also suggested future research directions for NIEHS and NCI:

- Decide on key exposures to prioritize and develop a series of cross-sectional or short-term longitudinal studies that are optimized, diverse, and fully characterized. Create standardized protocols for exposure assessments and epidemiological data collection. Then, with standardized sample collection and analysis protocols, it may be possible to probe how exposures affect biological function and put together a cancer exposome atlas.
- Bring together researchers working in multi-omics to communicate across study designs and exposures. Develop consensus on how to move the field forward.
- Develop guidance on optimal sample sizes for studies of different cancer types, as well as consensus on measurement tools to improve data harmonization. Emphasize the need for clear hypotheses and developing the best models possible to improve statistical power.
- Focus on providing omics data to researchers. The Multi-Omics for Health and Disease consortium — an upcoming joint initiative among NIH institutes — will establish best practices and develop methods focused on the integration of omics with other environmental exposure data.
- Create a specific NIH study section on multi-omics and exposome integration. The field of omics has an important role to play as agencies start to use multi-omics data for risk characterization. This may be especially important because multi-omics applications are not always "hypothesis-driven."

Introduction and Overview

Many environmental factors can induce biological responses at the genome, epigenome, transcriptome, proteome, and metabolome levels, altering gene expression and function. Interactions between the environment and these omics layers drive many complex disease outcomes, making it critical to incorporate environmental exposures into multi-omics studies to enhance our understanding of human diseases. Multi-omics data integration fits within the Precision Environmental Health Framework, which seeks to reduce the adverse health effects of exposures that make their way to us through the air we breathe, the water we drink, and the food we eat by identifying individuals who are specifically susceptible to environmental threats and enabling precise, targeted, and effective prevention.[24]

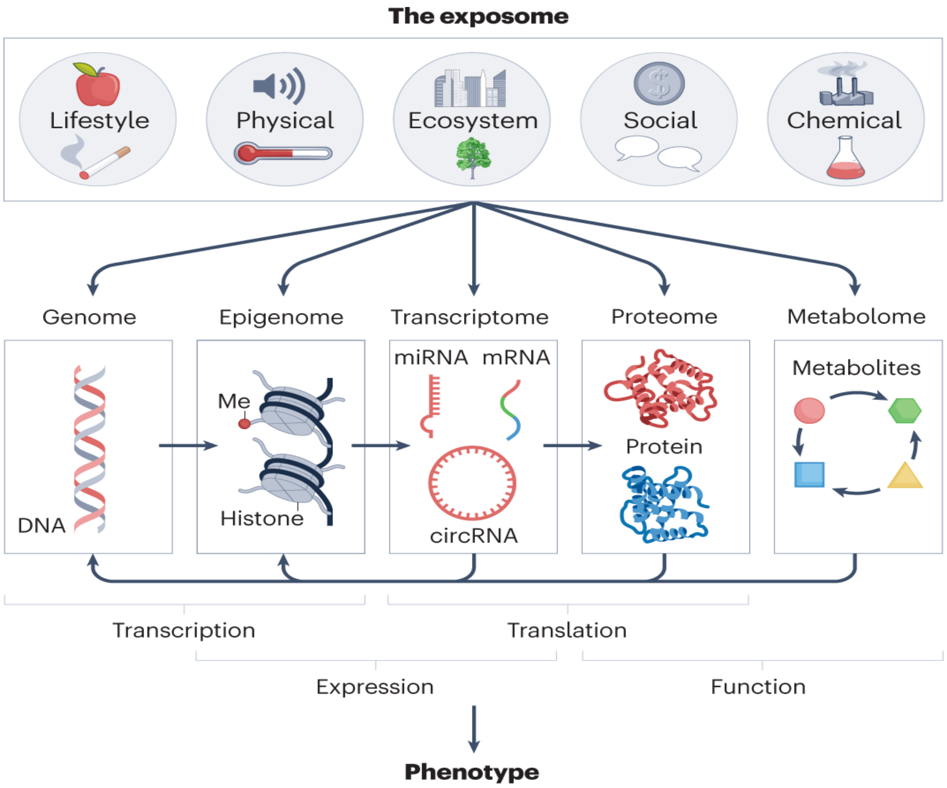


Figure 1. Environmental factors impact gene expression and function through a variety of pathways.
(From Wu H, Eckhardt CM, Baccarelli AA. Molecular mechanisms of environmental exposures and human disease. Nat Rev Genet. 2023 May;24(5):332-344. doi: 10.1038/s41576-022-00569-3.)

On February 14 and 15, 2023, the National Institute of Environmental Health Sciences (NIEHS) and the National Cancer Institute (NCI) convened a virtual workshop on Integrating Environmental Data with Other Omics for Cancer Epidemiology. The workshop was co-chaired by Kimberly McAllister, Ph.D., of NIEHS, and Leah Mechanic, Ph.D., MPH, of NCI. As part of the NCI/NIEHS Cancer and the Environment Working Group's effort to promote research into the effects of environmental exposures on cancer risk and etiology, the workshop brought together a multidisciplinary group, including environmental scientists, epidemiologists, toxicologists,

[24] Walker CL, Dolinoy D, Baccarelli A. Perspectives on Precision Environmental Health. National Advisory Health Sciences Council, February 2021.

physicians, and bioinformatics researchers (See Appendix 1 for the full list of participants and biographies). Presentations and moderated discussions highlighted challenges and opportunities related to the integration of environmental exposure data with other omics data for human cancer population studies. The full workshop agenda is included in Appendix 2. Key publications and other resources are available in Appendices 3 and 4.

The workshop was organized into four main sessions over two days, with introductory presentations at the beginning of Day 1 and a closing session on Day 2, designed to integrate themes from the prior sessions. Each session included a *State of the Science* speaker to give a broad overview of the session theme and an *Applications* speaker to apply the session theme to a particular research project, with an audience question-and-answer forum following each presentation. Sessions also included moderated panel discussions with the speakers and other experts to address key questions posed by the moderator and audience.

Day 1:

- Introduction and Overview
- Session I: Specific Cancer Considerations
- Session II: Computational Approaches

Day 2:

- Session III: Integration of Environmental Data with Other Data Types
- Session IV: Experimental Models and Functional Approaches
- Closing Session

Trevor Archer, Ph.D., Deputy Director of NIEHS, and Gary Ellison, Ph.D., M.P.H., Deputy Director of the Division of Cancer Control and Population Sciences (DCCPS) at NCI, set the stage for the workshop, with opening presentations highlighting challenges facing environmental exposure studies and cancer epidemiological studies.

Dr. Archer explained that there are many known environmental exposures that can lead to cancer — for example, arsenic, ultraviolet radiation, and tobacco smoke — and some of the best described gene–environment (GxE) interaction examples are for cancer outcomes. An example of a well-established GxE interaction involves the NAT2 gene and smoking as the environmental factor, where smokers with one variant in NAT2 have a much higher risk of bladder cancer compared to smokers with a different variant.[25] Though many GxE relationships are well established, integrating environmental data with other omics, including metabolomics, proteomics, and transcriptomics, poses challenges related to the heterogeneity of environmental data, measurements with different instruments at different scales, and temporality of exposures. As we continue to increase our ability to gather large amounts of omics data, there is a need for analytical tools, techniques, and established best practices for integrating various environmental data with other omics data and for evaluating methods for integration, which is particularly challenging in longitudinal cancer epidemiology studies.

Dr. Ellison discussed the complexity of cancer, which is not just one disease but consists of over 100 diseases that can arise from nearly every cell type in the body. Lifestyle, social, psychological, and environmental factors interact with intrinsic factors across the life course to affect health. Once an individual is exposed to a cancer-causing environmental agent, cancer can take many years to develop, which makes studying the role of

[25] Gene Environment Interaction. NHGRI Talking Glossary of Genomic and Genetic Terms. Available at https://www.genome.gov/genetics-glossary/Gene-Environment-Interaction

environmental factors a difficult task. Although there are challenges in collecting, managing, and integrating environmental, phenotypic, and genomic data, there are also opportunities to involve multiple disciplines and team science, as expertise is needed at all steps in the Precision Environmental Health pipeline. Dr. Ellison also discussed NCI's cancer epidemiology cohort resources, including the Cancer Epidemiology Descriptive Cohort Database (CEDCD) and Cohorts for Environmental Exposures and Cancer Risks (CEECR). More information on these key resources is available in Appendix 4.

Following opening remarks from Drs. Archer and Ellison, Drs. McAllister and Mechanic discussed the workshop's goals and intended outcomes. For the purposes of this workshop, "omics" refers to the comprehensive quantification and characterization of separate layers of biological regulation, such as genomics, transcriptomics, proteomics, epigenomics, metabolomics, and others. The workshop's goal was to identify the challenges and opportunities related to the integration of environmental exposure data with other omics data for human cancer population studies and to inform future supported research directions for NIEHS and NCI. An upcoming joint initiative between NCI, NIEHS, and the National Human Genome Research Institute (NHGRI) — Multi-Omics for Health and Disease — will establish best practices and develop methods focused on the integration of omics with other environmental exposure data. The workshop will inform this consortium as well as many other multi-omics related efforts in years to come.

Throughout the sessions, the workshop sought to answer these key questions:

- What omics layers or platforms best inform environmentally relevant biomarkers for cancer outcomes?
- What new methods and computational approaches are needed to move the field forward?
- What study designs and approaches should be considered for cancer molecular epidemiology studies incorporating environmental exposure measures and other omics data?
- How can experimental models of environmental exposures be used in conjunction with human multi-omics studies to enhance our understanding of complex human diseases?

The remainder of this report summarizes speaker presentations and accompanying discussion, as well as key takeaways and suggested future directions for NCI and NIEHS.

Session I: Specific Cancer Considerations

The first session, moderated by Ulrike Peters, Ph.D., MPH, of the University of Washington Fred Hutchinson Cancer Center, covered the role of tumor compared to normal tissue, long latency periods, and exposures during susceptibility windows in cancer initiation and promotion.

Cancer Susceptibility and Environmental Exposures: Why Study Designs Matter

Mary Beth Terry, Ph.D., Columbia University, Mailman School of Public Health

Study design is essential to consider when making inferences about results, particularly in studies of diseases with long induction times, like cancer. Importantly, knowing when samples were collected and from whom is vital to avoiding biased and underpowered research results.

Study design considerations include:

- Moving beyond the evidence-based hierarchy to consider context and the most appropriate design given the study question.

- Incorporating omics into the research hypothesis.
- Understanding why studies using population-based sampling may have insufficient power to detect GxE interactions.
- Life course, acquired mutations, and windows of susceptibility.
- Designing studies across the cancer continuum, from etiology to prognosis.

Study designs for public health impact, policy, and evaluation often prioritize generalizability and, as a result, favor population-based sampling, while randomized controlled trials (RCTs) incorporate more targeted sampling. Almost all cancer causes have been identified from observational research rather than RCTs. Because most cancers are identified later in life, population-based research studies are underpowered to detect GxE interactions, particularly for alterations in the germline. By using more targeted sampling methods — for example, studying a high-risk group or individuals with a family history of cancer — we can ensure that there is enough power to draw valid inferences in omics studies.

The traditional paradigm for cancer risk separated the majority of cancers that occurred in the population as environmentally caused, while the remaining cancers that clustered within families were thought to be driven primarily by germline genetics. The identification of many of the hundreds of cancer genes started with these family-based studies. We now know that no cancer gene is 100% penetrant, meaning all cancers involve GxE interactions. Cancer genes are relevant to everyone — through both germline and acquired mutations throughout life — not just people with a family history, but many required an enriched study, such as a family study, for identification. Roughly half of all mutations and epigenetic changes occur before the body has matured,[26] indicating that studies on germline variants should be conducted earlier in life. Studies involving older adults must consider the burden of acquired mutations, as conducting studies of germline variants in older individuals without factoring in other susceptibility factors could lead to biased results.

Epidemiological studies of cancer have mainly focused primarily on genomics, and to a lesser extent epigenomics, but we are starting to collect samples in a way that can allow for measurement of other omics. As technology improves to allow for inclusion of different omics platforms, it is essential to think about the timing of when they are applied. Genomics and other omics platforms, if measured before the exposure occurs, can be used to understand underlying susceptibility. If measured after the exposure, all of the omics except genomics may be used to evaluate potential mediating pathways or to measure outcomes of an intervention study — for example, measuring the metabolome after an intervention to reduce exposure to a carcinogen.

Beyond study design, participant recruitment and timing of sampling are crucial. However, most genomic research studies that lead to cancer drug discoveries and clinical trials are conducted with samples that do not represent the global population.

Most environmental studies also do not factor in susceptibility. For example, many breast cancer studies do not account for pregnancy or lactation periods when tissue is undergoing rapid changes and cell division that may increase cancer susceptibility. Studies that have factored in susceptibility, whether based on windows of susceptibility of genetic susceptibility, showed much more consistent evidence for the relationship between environmental chemical exposures and risk of breast cancer.

Many existing cohorts are skewed toward older individuals, which will lower statistical power for assessing GxE interactions. These cohorts also potentially introduce survivorship bias, if they include participants who are

[26] Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14:R115.

healthier and have a lower cancer risk than the population they are intended to represent. Solutions include establishing cohorts of younger, more diverse individuals that are enriched for cancer risk. When using existing cohorts of older individuals, researchers should consider measuring susceptibility beyond germline genetics. Moving forward, studies need to consider windows of susceptibility as well as the influence of underlying susceptibility, particularly for common environmental exposures outside of occupational settings. Population-based cohorts or cohorts that enroll most participants at midlife or later need to consider survivorship bias and acquired mutations. Additionally, oversampling groups underrepresented in omics research — to better estimate attributes of those groups and then use sampling weights in analyses to avoid unintended biases associated with oversampling[27] — is key for reducing cancer burdens.

Disentangling the Etiological Pathways Between Body Mass Index and Site-Specific Cancer Risk Using Tissue-Partitioned Mendelian Randomization

Genevieve Leyden, Ph.D., University of Bristol, Bristol Medical School

Obesity is an established risk factor for many types of cancer. In epidemiological studies, body mass index (BMI) is often used to define obesity. BMI is also a highly heritable trait — large-scale, genome-wide studies have been important in identifying regions of the genome that are reliably associated with BMI, with approximately 900 BMI loci uncovered to date. However, there are multiple causal pathways other than genetics by which an individual might be predisposed to having a higher BMI, not all of which have the same effect on disease risk.

Using a Mendelian randomization (MR) approach, which uses genetic data to infer causal relationships in observational epidemiological studies, Dr. Leyden studied the effects of BMI on different cancer outcomes. By obtaining all 900 genetic variants identified for BMI genome-wide — and relating them to genetic variants for diseases identified in other genome-wide association studies (GWAS) from similar populations — it is possible to estimate the total effect of BMI on the disease outcome. This methodology is increasingly playing an important role in genetic epidemiology studies.

Insight on genetic mechanisms brings new understanding to how people are predisposed to developing obesity and how that predisposition relates to disease. The contribution of individual tissue types is a growing area of research. Dr. Leyden's study aimed to identify and distinguish the contribution of BMI single nucleotide polymorphisms (SNPs) on disease risk based on tissue-derived gene expression Using subcutaneous adipose and neural tissue, Dr. Leyden aimed to understand the contribution of individual BMI-related SNPs to disease.

Results showed that variants related to BMI found in brain tissue samples were predominantly implicated in lung cancer, while BMI-related variants from adipose tissue samples were implicated in endometrial cancer..[28] According to Dr. Leyden, the results of this study provide an example of how multi-omics datasets can be used to help us get closer to understanding the relationship between a complex lifestyle risk factor, such as obesity or high BMI, on site-specific cancer risk. Other omics data, such as methylation or protein data, may be used to provide additional context for studying genetic effects on complex disease risk.

[27] Vaughan R. Oversampling in Health Surveys: Why, When, and How? Am J Public Health. 2017;107(8):1214-1215. doi: 10.2105/AJPH.2017.303895.

[28] Leyden GM, Greenwood MP, Gaborieau V, Han Y, Amos CI, Brennan P, Murphy D, Davey Smith G, Richardson TG. Disentangling the aetiological pathways between body mass index and site-specific cancer risk using tissue-partitioned Mendelian randomisation. Br J Cancer. 2023;128(4):618-625. doi: 10.1038/s41416-022-02060-6.

Session II: Computational Approaches

Session II was moderated by Andrea Baccarelli, M.D., Ph.D., of the Columbia University Mailman School of Public Health. The session focused on methods to deal with heterogeneity and high dimensionality of environmental data, as well as differences in data scales and collection methods.

Statistical Approaches for Integrating Environmental and Omics Data in Cancer Epidemiology Studies

James Gauderman, Ph.D., University of Southern California, Keck School of Medicine

Limited statistical power remains a primary concern in GxE analyses, making it essential to have the largest possible sample size by pooling resources across studies. Bringing in various levels of omics data can enhance our understanding of the connection between environmental exposures and cancer outcomes. There are many techniques for identifying features within each omics layer; however, there are challenges in deciding which features to include as well as in data integration among different studies. Incorporating environmental exposures into multi-omics studies adds another level of complexity due to differences in the resolution and precision of environmental data.

For example, there are a variety of methods for collecting air pollution data. Questionnaires that ask about smoking, gas stove use, or time spent outdoors can provide some crude data on air pollution exposure, whereas regional monitors, exhaled biomarkers, wearable sampling devices, or satellite imaging could provide a higher level of resolution.

Omics data consist of a single omics layer measured on a set of individuals. These data may include millions of SNPs from thousands of individuals, for example, where it may not be possible to preserve all complexity. High dimensional data reduction is a statistical technique that can be used to reduce the number of attributes in a dataset while keeping as much of the variation in the original dataset as possible.

Using data from the Functionally Informed GxE Interaction (FIGI) study, Dr. Gauderman and team conducted a genome-wide scan to determine how genes and environment, separately and in combination, impacted colorectal cancer outcomes, using a two-step approach:[29]

- Step 1: Initial screening to allocate SNPs into bins based on priority, or likelihood that they are involved in a GxE interaction.
- Step 2: Test SNPs for an interaction under a modified significance threshold, where higher priority bins are tested at a more liberal significance threshold, and lower priority bins are tested much more stringently. SNPs that are more likely to have an interaction based on the screening statistic in Step 1 will have a higher chance of being discovered.

Though environmental factors were considered as a single variable, consideration of multiple environmental exposures as separate variables may help identify underlying lifestyle profiles. When incomplete or no environmental data are available, a separate data resource can be used to build an exposure model, with omics layers then applied for mediation or interaction analysis.

[29] Kawaguchi ES, Kim AE, Lewinger JP, Gauderman WJ. Improved two-step testing of genome-wide gene-environment interactions. Genet Epidemiol. 2023;47(2):152-166. doi: 10.1002/gepi.22509.

Another approach for assessing GxE interactions is to use polygenic risk scores (PRS), which provide a measure of an individual's disease risk due to their genes. The score considers each version of a gene — related to a specific disease — that the individual has. PRS, however, condenses tens to thousands of SNPs to a single measure, assuming the same interaction effects across all SNPs within the PRS. Dr. Gauderman posed the solution of breaking PRS into subgroups. Rather than modeling a single PRS, it is possible to model multiple, which is still much simpler than attempting to model the thousands of interactions among each of the millions of SNPs.

According to Dr. Gauderman, incorporating environmental data with omics holds promise for determining cancer risk and prognostic factors. A large and well-characterized study population, combined with integration of functional genomics data into novel statistical methods, provide opportunities to better understand how genetic and environmental risk factors contribute to individual disease risk. Discovering GxE interactions will provide insight into the underlying carcinogenic mechanisms impacted by established environmental risk factors, and in turn help identify potential targets for future interventions.

Longitudinal Multi-Omic Characterization of a Community Cohort After Chemical Exposures From Hurricane Harvey

Cristian Coarfa, Ph.D., Baylor University, Baylor College of Medicine

Hurricane Harvey hit Houston, Texas, in August 2017. Beyond the disaster's immediate threat, flooding led to major public health issues — drinking water contamination, chemical and explosive hazards, vector-borne and infectious diseases, and mental health concerns, among others. The crisis sparked creation of the Houston Hurricane Harvey Health (Houston3H) study to assess health outcomes, chemical exposures, and the microbiome in a cohort of Harvey-exposed households.

Using questionnaires and geographic information, the Houston3H study learned about individuals' experiences and neighborhood characteristics, and assessed allergic symptoms, injuries, and mental health. Household swabs, nasal, saliva, and stool biospecimens were collected to examine the fungal mycobiome and bacterial microbiome. Wristband samplers were used to assess exposures to over a thousand chemicals. The study enrolled a total of 347 unique participants at two times: within approximately one month of Harvey (n = 206) and approximately 12 months after Harvey (n = 266), including 125 individuals who participated at both time points.[30]

Dr. Coarfa and team assessed correlations between Harvey exposure, diet, demographic, health, and pollution burden data, and found that most correlations with chemical exposures were found in Black study participants. Exposure to particulate matter showed the strongest correlation with participants in East Houston who experienced flooding.

The team then used a weighted gene co-expression network analysis (WGCNA) to identify similarities in gene expression among different clinical traits. They found the strongest module-trait correlations among nasal mycobiome samples collected from participants in East Houston. Using machine learning, they developed a

[30] Oluyomi AO, Panthagani K, Sotelo J, Gu X, Armstrong G, Luo DN, Hoffman KL, Rohlman D, Tidwell L, Hamilton WJ, Symanski E, Anderson K, Petrosino JF, Walker CL, Bondy M. Houston hurricane Harvey health (Houston-3H) study: assessment of allergic symptoms and stress after hurricane Harvey flooding. Environ Health. 2021;20(1):9. doi: 10.1186/s12940-021-00694-2.

model to predict both numerical and binary categorical traits based on metagenomic and chemical data. While nasal mycobiomes were similar among East Houston participants, their chemical exposures tended to differ.

Dr. Coarfa concluded that questionnaire-based data revealed a comprehensive interplay between questionnaire variables, including diet, demographics, hurricane exposure, pollution burden, and self-reported health. Robust omics profiling of the microbiome, mycobiome, and exposome was allowed through stool, nasal, saliva, and home environment samples as well was through chemical wristband sampling. While the study showed strong associations among demographic and pollution variables, there were fewer associations with health, which may have been due to the limited set of questions and no use of electronic health records (EHRs). Additionally, comprehensive omics profiling — which is crucial for identifying rich multi-omics features and modules — within the same individuals was not available due to logistical challenges. Overall, the Houston3H study showed the power of multi-omics environmental profiling, and future work should plan for the use of more powerful technologies to assess other omics, including metabolomics and proteomics.

Day 1 Discussion Points

The Session I panel discussion, moderated by Ulrike Peters, Ph.D., MPH, featured the following panelists, in addition to Drs. Terry and Leyden:

- Catherine Metayer, M.D., Ph.D., University of California, Berkeley, School of Public Health
- Sophia Wang, Ph.D., City of Hope Comprehensive Cancer Center
- Peter Kraft, Ph.D., Harvard University, T.H. Chan School of Public Health
- Cathrine Hoyo, Ph.D., North Carolina State University

The Session II panel discussion, moderated by Andrea Baccarelli, M.D., Ph.D., featured the following panelists, in addition to Drs. Gauderman and Coarfa:

- Marylyn Ritchie, Ph.D., University of Pennsylvania, Perelman School of Medicine
- Nilanjan Chatterjee, Ph.D., Johns Hopkins University, Bloomberg School of Public Health
- Marina Sirota, Ph.D., University of California, San Francisco, Baka Computational Health Sciences Institute

Each panelist provided brief comments relating the session topic to their research area. Panelist biographies are available in Appendix 1.

Key discussion points included:

Study design issues. Timing of exposure, long latency periods, prioritizing sample quality, and optimal sample sizes were key themes discussed in Sessions I and II. Although case-control study designs will likely remain an important part of epidemiological studies, measuring exposure at the time of disease diagnosis provides little information in terms of causality. There is a need to identify stable intermediate markers, such as DNA methylation markers, that can be used as surrogates for past exposures.

Participants noted the difficulty in identifying optimal sample sizes, particularly for rarer cancers. Assessing GxE interactions often requires very large sample sizes, which is not possible for rare cancers like childhood leukemia. We are much farther ahead in understanding some types of cancers compared to others. It is important to prioritize the quality of samples over the quantity — in some cases, it may be better to have fewer participants but capture very detailed omics data.

Existing cohorts for cancer studies are also skewed toward older individuals. Establishing younger and more diverse cohorts, capturing windows of susceptibility, and oversampling groups underrepresented in omics research are key steps for reducing cancer burdens.

Scientific goals should control choice of analysis. Biological hypotheses should be clearly defined prior to initiating studies and vague terminology should be avoided. Researchers need to take their time to understand the steps involved in analysis and what might be gained or lost by prioritizing some dimensions over others.

Environment drives cancer risk. Given the results of twin studies, it is clear that the environment, rather than genes, drives cancer risk. There is a need for robust exposure assessment with minimal measurement error to identify which exposures influence risk. However, there are challenges in measuring other types of environmental exposures, such as social stressors, that interact with environmental contaminants in ways that are still not fully understood. There are also challenges in combining different measurement scales for different types of exposures.

Data harmonization. Leveraging multi-omics datasets and pulling them together would be a useful start to combining different assays. In other words, measure multiple omics in the same individuals, same tissues, and at the same time point and correlate different omics signatures across samples to combine datasets. Resources should be put into the weakest link in the model because that is where most power loss will occur. Questionnaires should also be standardized so other researchers can use them to collect data in the same way, making it easier to harmonize data.

Developing a cancer exposome atlas. Creating a cancer exposome atlas is essential, especially when incorporating multi-omics to understand cancer risk. However, there has not been a coordinated effort to do so, and there is a lack of consensus about methods for analyzing the exposome as well as what the definition of "exposome" is. Collecting samples in a way that they can be analyzed for a variety of omics and uploaded to a platform to look at concurrently could spark the beginning of an atlas, starting with classes of chemicals and looking at their metabolomic effects and transcriptome impacts downstream.

Session III: Integration of Environmental Data with Other Data Types

Session III, moderated by Gary Miller, Ph.D., of the Columbia University Mailman School of Public Health, sought to inform our understanding of environmentally relevant cancer outcomes, particularly timing, dose, and mixtures challenges.

Dose Versus Burden in Exposome Research: Translation of Integrative Omics of Model Systems to Environmental Epidemiology in Cancer Research

Dean Jones, Ph.D., Emory University, Winship Cancer Institute

In 2002, then-NIH director Elias Zerhouni, M.D., introduced the concept of a roadmap, a far-reaching plan to transform key areas of biomedical research.[31] This novel approach — "Data of the Future," as Dr. Jones called it — could take metabolomics to an omics scale, with simple, high-throughput, reproducible, standardizable, and affordable studies that provide extensive coverage. Over the next 15 years, high-resolution metabolomics continued to improve and can now measure up to a million chemical signals in a small biological sample. In

[31] NIH Office of Strategic Coordination – The Common Fund. A Decade of Discovery: The NIH Roadmap and Common Fund. Available at https://commonfund.nih.gov/commemoration

2016, the Findable, Accessible, Interoperable, and Reproducible (FAIR) Principles were published, adding additional guidelines for human exposome research to the roadmap established by Dr. Zerhouni.[32]

Improved analytic tools and databases for targeted and nontargeted metabolic profiling, along with bioinformatics, pathway mapping, and computational modeling, are now used for nutrition research on diet, metabolism, microbiome, and their associations with health. Metabolomics studies measure components of a broad range of the exposome, including diet-derived metabolites, post-infection products, drugs and ethnobotanicals, personal care products, and environmental chemicals.[33] Environmental epidemiology has generally worked with targeted measures of small numbers of environmental exposures. Advances in metabolomics now allow for exposome epidemiology at an omics scale.

Individuals are not exposed to one chemical at a time, and lifelong exposures are cumulative. Omics-scale exposome analyses provide an opportunity to explore the multiplicity of exposures and complexities of their interactions. Dr. Jones discussed several needs for the field of exposome epidemiology:[34]

- Extension of environmental epidemiology to analysis of omics-scale exposure data.
- Improved environmental chemical and xenobiotic metabolite identification.
- Improved quantification procedures for low-abundance environmental chemicals.
- Coupling exposome data with epigenetic data to develop systematic knowledge of the best characterized mechanism for exposure memory.
- Maintenance of cohorts with coverage of critical exposure windows and consequences.

The central dogma of molecular biology — the theory that genetic information flows from DNA, to RNA, to protein, to metabolite — directs the use of omics technologies to gain insight through the integration of detailed measures of these omics layers. We have genomics, proteomics, and metabolomics capabilities that need to be aligned with immunologic, microbiome, and exposome measures to move the field forward.

Environmental health research with cell and rodent models has established the utility of integrating untargeted mass spectrometry with other omics data. These model systems have critical value because findings can be tested and validated through genetic, pharmacologic, flux analyses and other mechanistic tests. Model system studies with nontargeted exposomics methods can reveal biologically relevant chemicals that have escaped discovery and what their functions are. Experimental models can then be integrated with observational studies to link chemical exposures to health outcomes. The gap between mechanistic research in model systems and epidemiology research using environmental exposure data and readily accessible human samples can be bridged by integrative exposomics and metabolomics of normal and diseased human tissues.

Examples of Omics Integration in Studies of the Environment on Human Health

Carmen Marsit, Ph.D., Emory University, Rollins School of Public Health

Various omics measures can be examined with environmental exposures. Approaches for integrating environmental data with omics include meet-in-the-middle analyses, polygenic and epigenetic risk scores, and

[32] Wilkinson MD, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018. doi: 10.1038/sdata.2016.18.

[33] Jones DP, Park Y, Ziegler TR. Nutritional metabolomics: progress in addressing complexity in diet and health. Annu Rev Nutr. 2012;32:183-202. doi: 10.1146/annurev-nutr-072610-145159.

[34] Jones DP, Cohn BA. A vision for exposome epidemiology: The pregnancy exposome in relation to breast cancer in the Child Health and Development Studies. Reprod Toxicol. 2020;92:4-10. doi: 10.1016/j.reprotox.2020.03.006.

central dogma-driven quantitative trait analyses, as well as new machine learning approaches to bring data together. In meet-in-the-middle analyses, overlapping pathways and omics features among genome-wide association studies (GWAS) are considered intermediate mechanisms and biomarkers, respectively. Identifying those overlapping pathways and features then allows for exploration of the potential biological mechanisms linking an environmental exposure to a health outcome.

In a study examining per- and polyfluoroalkyl substance (PFAS) exposure and reduced fetal growth, Dr. Marsit and collaborators applied a meet-in-the-middle approach with an advanced untargeted metabolomics workflow to investigate associations between PFAS levels, metabolome perturbations, and adverse birth outcomes in a cohort of 313 African American mother and newborn pairs.[35] They characterized PFAS exposure among pregnant women and newborns in Atlanta, Georgia, investigated the association between serum PFAS and metabolome using a nontargeted metabolomics approach, and then identified the potential biological pathways for adverse birth outcomes. They found perturbations in amino acid, lipids and fatty acid, bile acid, and androgenic hormone metabolism associated with PFAS exposure, suggesting that oxidative stress, inflammation, and placental transport function may be mechanisms through which PFAS affect fetal growth.

Dr. Marsit and team also studied pathways through which in utero multi-metal exposures impact fetal growth.[36] Higher concentrations of metals like arsenic, cadmium, and tin are related to an increased risk of being small for gestational age. However, individuals are not exposed to a single metal — rather, mixtures of metals interact to impact risk. Using a weighted quintile sum regression — an approach that estimates the effects of a mixture of correlated chemicals to identify the individual chemicals most strongly associated with a health outcome — the team found that metal mixtures predominated by arsenic and cadmium drive low birthweight risk. Then, using WGCNA to define specific co-regulated gene clusters, they found associations between exposure to metal mixtures and changes in placental gene expression, suggesting that the placenta may be mediating the effects of these environmental contaminants on birthweight.

While studies integrating environmental data with other omics have challenges and limitations, including temporality of exposures and ability to access target tissue, there is an opportunity to demonstrate downstream biological effects to prove how social determinants of health are causal agents in disease development. Integrating omics with environmental data can uncover mechanisms underlying links between the environment and health, provide direction for public health interventions, reveal exposure histories, and help with individualized risk assessment to prevent adverse health effects, aligned with the goals of Precision Environmental Health.

Session IV: Experimental Models and Functional Approaches

Stephen Montgomery, Ph.D., of Stanford University, introduced Session IV with a presentation on challenges and opportunities for modeling environmental effects at an omics scale. When measuring the impact of environment on disease, challenges include imperfect measures of exposure, variability in natural populations,

[35] Chang CJ, Barr DB, Ryan PB, Panuwet P, Smarr MM, Liu K, Kannan K, Yakimavets V, Tan Y, Ly V, Marsit CJ, Jones DP, Corwin EJ, Dunlop AL, Liang D. Per- and polyfluoroalkyl substance (PFAS) exposure, maternal metabolomic perturbation, and fetal growth in African American women: A meet-in-the-middle approach. Environ Int. 2022;158:106964. doi: 10.1016/j.envint.2021.106964.

[36] Deyssenroth MA, Gennings C, Liu SH, Peng S, Hao K, Lambertini L, Jackson BP, Karagas MR, Marsit CJ, Chen J. Intrauterine multi-metal exposure is associated with reduced fetal growth through modulation of the placental gene network. Environ Int. 2018;120:373-381. doi: 10.1016/j.envint.2018.08.010.

limited molecular understanding, and difficulty in proving causality. Model systems can help to reduce confounders that might exist in natural populations, converting a noisy signal to a more robust signal.

Experimental models of environmental exposures include cell and organoid in vitro models, as well as animal models. Cellular models allow for differential response testing of GxE interactions, though there are challenges posed by the multitude of environmental conditions to test as well as a limited number of cell types available. Organoid models allow for testing a range of cell types and cellular phenotypes, though they make it difficult to study GxE interactions. Animal models allow for prospective studies of complex exposures and provide access to hard-to-acquire cell and tissue types, though few studies translate well to humans.

The session covered how the expansion of in vitro functional genomics and model organisms might be used to inform multi-omics human studies related to environmentally sensitive cancers.

Mechanism of Action and Inflammatory Axis for Air Pollution-Induced Non-Small Cell Lung Cancer

William Hill, Ph.D., The Francis Crick Institute

Lung cancer in never smokers (LCINS) is distinct from lung cancer in smokers in that there is a relatively low mutational burden, it is more common in females, and there are geographic disparities. Risk factors for LCINS include germline susceptibility, radon exposure, secondhand smoke, environmental pollution, and diet. As the proportion of non-small cell lung cancer cases that are never smokers is increasing, there is an unmet clinical need to understand the biological underpinnings of this patient group.

The classical mutation model of cancer indicates that a carcinogen causes a DNA mutation, leading to tumor growth. However, the vast majority of mutations in LCINS cases do not appear to be in the context of an exogenous mutational signature, but rather display an endogenous mutational behavior, suggesting that carcinogens are behaving in a non-mutagenic fashion. The tumor promotion model of cancer is a two-step process: an endogenous process, or initiator, causes a mutation; then, that mutation is exposed to a carcinogen, or promoter, leading to clonal expansion and eventually cancer. Recent research has shown that normal healthy tissue can harbor mutant clones with cancer-driver mutations with no evidence of cancer.[37] Because somatic mutations may be found in normal lung tissue prior to a lung cancer diagnosis, Dr. Hill and team hypothesized that carcinogens act via non-mutagenic mechanisms, perhaps on nascent mutant clones present in normal tissues.

Previous studies have shown geographic associations between particulate matter (PM) exposures and lung cancer incidence, so the team sought to establish a functional link using mouse models. They found that air pollution promoted cancer in mice with preexisting mutations on the epidermal growth factor receptor (EGFR) gene, which has been associated with lung cancer in humans. However, pollution did not lead to an increase in clonal expansions in immunodeficient mice, suggesting that a competent immune system is required for pollution to promote tumorigenesis.

Because tumor promoters act on preexisting mutations in latent cells, the team sought to determine whether EGFR mutations exist in normal human lung tissue. They found that the mutation is present in normal tissue, and EGFR mutations in non-smoking adults increase with age. However, presence of the mutation alone does

[37] Martincorena I. Somatic mutation and clonal expansions in human tissues. Genome Med. 2019;11(1):35. doi: 10.1186/s13073-019-0648-4.

not indicate that an individual will develop cancer. PM is likely a tumor promoter, inducing inflammation and tumor initiation among cells with EGFR mutations.

Though mouse models poorly replicate decades of exposure that humans experience, they can elicit questions that can later be addressed in clinical cohorts. PM is also a broad type of exposure that may encompass other chemicals that initiate carcinogenesis. According to Dr. Hill, this initial study to develop mouse models will allow the team to further explore which chemicals induce cancer.

Deciphering Mechanisms Through Which Environmental Risk Factors Mediate Colorectal Cancer Risk Through Weighted Gene Co-Expression Networks

Matthew Devall, Ph.D., University of Virginia

Colon organoids, which contain many cell markers present in the colon crypt and can be grown seemingly indefinitely, may be a useful tool for modeling environmental factors that increase risk of colorectal cancers. The University of Virginia established a biorepository to house colon biopsies from mostly healthy individuals (n > 140), with the goal of generating 500 lines over the next two years. This system has several advantages, allowing researchers to account for racial disparities, observe differences in left and right colons, determine the role of BMI changes over time in colorectal cancer risk, differentiate cell compositions, and use it as a tool to model environmental exposures.

Dr. Devall and colleagues identified mechanisms driving the colorectal cancer effects following environmental exposures like those in observational and epidemiological studies. Across studies, they found that organoids derived from right and left colon responded differently to drug treatments, that exposures altered colon epithelial cell composition, and that some of the genes identified in exposure studies were the same as those associated with colorectal cancer through GWAS.

In a study assessing the risk of colorectal cancer from smoking and from eating red or processed meat, Dr. Devall and team treated independent organoid lines with a cocktail of carcinogens commonly found in cigarette smoke and red meat.[38] They then used WGCNA to identify and prioritize genes that represent plausible targets through which environmental factors exert their effects. Networks were defined based on gene-gene correlation structures, which were not limited by an arbitrary p-value, as many genes in the module may have small but important effects.

The team identified over 700 differentially expressed genes in response to carcinogen exposure. Network analysis revealed that those genes rarely act in isolation, and single-gene analysis was associated with a high false discovery rate. They identified consistent modules associated with smoking-related carcinogens, highlighting the impact of carcinogens directly on epithelial cells of colon crypts. Advances in organoid co-culturing methods may help clarify effects on other cells of the colon.

Improving integration of epidemiology and molecular biology can help researchers to better assess the interplay between genetics and environment in colon epithelial cells by incorporating extensive phenotypic data, as well as increasing cohort power and diversity. Though there are challenges in enhancing representation

[38] Devall M, Dampier CH, Eaton S, Ali MW, Díez-Obrero V, Moratalla-Navarro F, Bryant J, Jennelle LT, Moreno V, Powell SM, Peters U, Casey G. Novel insights into the molecular mechanisms underlying risk of colorectal cancer from smoking and red/processed meat carcinogens by modeling exposure in normal colon organoids. Oncotarget. 2021;12(19):1863-1877. doi: 10.18632/oncotarget.28058.

in colon organoid studies, Dr. Devall and team are prioritizing sample collection from younger age groups and African Americans, who have been underrepresented in colon organoid studies.

Day 2 Discussion Points

The Session III panel discussion, moderated by Gary Miller, Ph.D., featured the following panelists, in addition to Drs. Jones and Marsit:

- Scarlett Gomez, Ph.D., M.P.H., University of California, San Francisco, Helen Diller Family Comprehensive Cancer Center
- Thomas Metz, Ph.D., Pacific Northwest National Laboratory
- Douglas Walker, Ph.D., Emory University, Rollins School of Public Health
- Ivana Yang, Ph.D., University of Colorado, Anschutz Medical Campus
- Nathaniel Rothman, M.D., M.P.H., M.H.S., NCI

The Session IV panel discussion, moderated by Stephen Montgomery, Ph.D., featured the following panelists, in addition to Drs. Hill and Devall:

- David Reif, Ph.D., NIEHS
- Rebecca Fry, Ph.D., University of North Carolina at Chapel Hill, Gillings School of Global Public Health
- Francesca Luca, Ph.D., Wayne State University
- Justin Colacino, Ph.D., University of Michigan, School of Public Health

Each panelist provided brief comments relating the session topic to their research area. Panelist biographies are available in Appendix 1.

Key discussion points included:

Assessing social determinants of health. Social structures influence health, and there is a need to establish causal biological pathways of that health impact. Community partners embrace the concept of the exposome, because they feel their concerns about multiple potential exposures are being heard, rather than only hearing about the chemicals that scientists are interested in studying.

A major challenge in using EHR data is the ability to link and understand relevant outcomes. Larger sets of higher-quality life history data are needed to link social stressors to cancer outcomes, but there are challenges in scaling and logistics. An increasing wealth of geospatial data can provide information on social determinants of health, and those factors can then be linked to cancer disparities.

Coupling experimental models with population-level studies. Experimental models can be coupled with human population studies to elucidate mechanisms of action. Experimental models help to refine and dive deeper into questions that arise from human studies, and the impact of confounding is lessened in experimental models. Traditional epidemiological approaches (e.g., questionnaires) can be used for cross-referencing to test for accuracy in omics studies.

Colon organoid studies pose challenges in terms of representation. For example, colonoscopies are typically only performed at age 45 and onward, leading to a lack of data from younger individuals. Black individuals have also been historically underrepresented in colon organoid studies, though that trend is changing. There are also challenges in connecting mouse models to human studies. Timing, dose, and the effects of cumulative exposures are difficult to replicate and validate.

Next steps to understanding the impacts of exposure mixtures on health will require extensive coordination. Models can be trained to predict chemical effects, and in silico approaches will continue to evolve, moving the field forward.

Current and future technologies. Different methodologies pose different biases. With where we are in omics research, it may be worth taking a step back to consider what technologies are needed to reach our goals. Analytical technology will evolve, so it is important to understand the other side of the equation — study populations and samples — and characterize them correctly in order to develop better analytical approaches. Conceptual frameworks can help us understand which environmental factors impact each other and downstream omics.

When the Human Genome Project launched, technology had to catch up. Similarly, this effort requires innovative technologies and dedicated coordination. The cancer exposome atlas is a useful step in moving toward this goal.

We do not want to truncate the omics we are able to measure. There is a need for frameworks to determine which data are most important to the omics we are focusing on. Conceptual frameworks can help us understand which environmental factors influence each other as well as affect downstream omics. However, massive datasets hinder efforts to make connections among many measurements.

Coordination among research community. Participants noted the importance of connecting and collaborating with geographic information scientists and disease experts to assess relationships between exposures and disease outcomes. Working with epidemiological experts can also help narrow down factors of interest. Leveraging findings from existing studies, rather than having siloed research, will be important in data harmonization. Collaborations and team science are important since no model is perfect.

Closing Session: Integration of All Four Themes and Future Directions

Closing Summary

Chirag Patel, Ph.D., of Harvard University Medical School, summarized the prior workshop sessions. Dr. Patel noted that if exposures are causal for cancer, they must induce a biological change and response — multi-omics can be used as indicators of this exposure-cancer biology.

Cancer etiology is complex, and there are opportunities to formalize the role of the environment in etiological models. The environment, genetics, and clinical outcomes are strongly connected to demographic structure, including ancestry, race, and geographic location. Cancers show much phenotypic variation that has yet to be explained. There are also disparities in clinical and biological cancer outcomes due to confluence of screening, medical decisions, and therapeutics.

Environmental and multi-omics studies measure change throughout the life course; considering windows of susceptibility provides a big opportunity for moving the field forward. Sampling at key windows of susceptibility is a critical gap in studies of environmental impacts on cancer. For example, many breast cancer studies do not account for pregnancy or lactation periods when tissue is undergoing rapid changes that may increase cancer susceptibility.

Though several new approaches for measuring the exposome exist, there are bottlenecks in capturing the exposure of interest in an etiological-specific manner. Data and tissue analytic resources need to be integrated with existing cohorts — one step in creating the exposome atlas. Data harmonization of heterogeneous sources, including individual-level and geographic social determinants of health as well as genotypes and polygenic risk scores, is a necessary step, but it must take into account hypotheses and overarching scientific goals.

Final Discussion Points and Suggested Future Directions

The final workshop discussion was moderated by Kari Nadeau, M.D., Ph.D., of the Harvard University T.H. Chan School of Public Health. Key discussion points and suggested future directions were highlighted:

- **Reliability, repeatability, and reproducibility.** Studies should begin with the end in mind (e.g., policy change, health outcomes). As with all scientific research, omics studies demand reliability, repeatability, and reproducibility — reproducible results inform approval and regulatory processes. However, there are challenges with exposures that only affect a small fragment of the population, as results may be difficult to reproduce.
- **Cancer exposome atlas**. Conduct comprehensive cross-sectional or short-term longitudinal studies of healthy human populations that were selected for an exposure of interest with a wide exposure range. Create standardized protocols for study population selection, exposure assessment for current and historical exposures alike, questionnaire collection, and biological sample collection, processing, and storage. Exposures selected for study could be both known as well as suspected carcinogens. Then, with standardized laboratory and data analysis protocols, it should be possible to probe into biological perturbations of those exposures and put together a cancer exposome atlas. The cancer exposome atlas would allow one to see what pathways specific exposures perturb in different tissues. The cancer exposome atlas could also be complemented by or linked to experimental studies of the same exposures analyzed on the same platforms, where feasible.
- **Need for multidisciplinary collaboration**. Bring together researchers working in multi-omics to communicate across study designs and exposures. Develop a consensus to move the field forward.
- **Develop guidance and provide data to researchers**. Develop guidance on optimal sample sizes for studies of different cancer types, as well as consensus on measurement tools to improve data harmonization. Emphasize the need for clear hypotheses and for developing the best models possible to maximize statistical power. Focus on providing omics data to researchers. There is no single path forward; rather, many innovative efforts exist to harmonize large-scale data. The Multi-Omics for Health and Disease consortium — an upcoming joint initiative among NIH institutes — will establish best practices and develop methods focused on the integration of omics with other environmental exposure data.

Appendix 1: Participant Biographies

Trevor Archer, Ph.D.

National Institute of Environmental Health Sciences

Dr. Trevor Archer received a Ph.D. in Biochemistry in 1987 at Queen's University, Kingston, Ontario, Canada, after which he completed postdoctoral training on chromatin gene transcription and steroid receptors at the National Cancer Institute in Bethesda, Maryland. In 1992, Dr. Archer joined the University of Western Ontario in Canada, as a National Cancer Institute of Canada Scientist. Dr. Archer was recruited to the NIEHS in 1999 as head of the Chromatin Structure and Gene Expression Group and was later appointed as Chief, Laboratory of Molecular Carcinogenesis in February 2003. In 2014, Dr. Archer became the founding chief of the new Epigenetics and Stem Cell Biology laboratory at NIEHS. Dr. Archer has made numerous original and important contributions to the study of chromatin structure/function, epigenetics, and gene transcriptional regulation in breast cancer cells while publishing ~120 peer reviewed manuscripts.

Andrea Baccarelli, M.D., Ph.D.

Columbia University, Mailman School of Public Health

Dr. Andrea Baccarelli is the Leon Hess Professor and Chair of the Department of Environmental Health Sciences at the Mailman School of Public Health at Columbia University and serves as the Director of the NIH/NIEHS P30 Center for Environmental Health and Justice in Northern Manhattan. Dr. Baccarelli's work has supported international best practices for air pollution control developed by multiple agencies worldwide, and his findings have served as the basis for the Environmental Protection Agency's decision to enforce stricter guidelines for human exposure. Dr. Baccarelli's research investigates molecular mechanisms as pathways linking environmental exposures to human disease, and current projects investigate a range of mechanisms, including epigenomics, epitranscriptomics, extracellular vesicles and small non-coding RNAs, mitochondrial DNA, and the microbiome.

Nilanjan Chatterjee, Ph.D.

Johns Hopkins University, Bloomberg School of Public Health

Dr. Nilanjan Chatterjee is a Bloomberg Distinguished Professor of Biostatistics and Oncology at Johns Hopkins University. He leads research on statistical and computational methods for the analysis of genome-wide association studies, gene-environment interactions, integrative -omics studies, polygenic scores, and risk prediction. His applied research has led to the discovery of novel genetic susceptibility loci for cancers, broad characterization of underlying genetic architecture, understanding the nature of and public health implications for gene-environment interactions, and development of risk prediction models.

Cristian Coarfa, Ph.D.

Baylor University, Baylor College of Medicine

Dr. Cristian Coarfa is an Associate Professor and Director of Multi-omic Bioinformatics for the Dan L Duncan Comprehensive Cancer Center at Baylor College of Medicine. His research interests are in multi-omics data integration, with a focus on epigenetics and the microbiome, and detection of molecular biomarkers and novel disease endotypes.

Justin Colacino, Ph.D.

University of Michigan, School of Public Health

Dr. Justin Colacino is an Associate Professor of Environmental Health Sciences at the University of Michigan School of Public Health whose research focuses on understanding environmental and dietary factors in the development of chronic diseases like cancer. Specifically, the goal of his research is to characterize the susceptibility of normal stem cell populations to environmental stress to understand the link between dysregulated development and disease. Of particular interest are understanding the changes that occur at the epigenetic and transcriptional level, changes which affect not only gene expression but also how progenitor cells differentiate and divide. His research group combines wet lab bench work and bioinformatic and statistical analysis of large scale genomic and epidemiologic data sets to translate findings from *in vitro* and *in vivo* models to the population level.

Matthew Devall, Ph.D.

University of Virginia

Dr. Matthew Devall is a Senior Scientist in the Department of Family Medicine at the University of Virginia. He had four years of research experience almost exclusively centered on the evaluation of colorectal cancer genetic and environmental risk factors using the colon organoid model. His work has expanded the net of statistical and computational models employed within the field of colon organoids. By coupling large-scale exposure studies with machine-learning and network-based approaches, Dr. Devall's work aims to identify novel and robust mechanisms through which common lifestyle factors influence colorectal cancer risk.

Gary Ellison, Ph.D., M.P.H.

National Cancer Institute

Dr. Gary L. Ellison is the deputy director of the Division of Cancer Control and Population Sciences at the National Cancer Institute. Prior to his current position, Dr. Ellison was the chief of the Environmental Epidemiology Branch (EEB) in the division's Epidemiology and Genomics Research Program (EGRP), where he oversaw extramural research focused on modifiable factors and cancer risk. Dr. Ellison led program directors within EEB with expertise spanning all domains of the exposome, including the general external (e.g., broader social context), specific external (e.g., lifestyle factors; environmental pollutants; chemical, physical, and infectious agents), and internal environments (e.g., biomarkers of effect, early damage). He joined EGRP as an epidemiologist and program director in 2008 and became chief of EEB in 2016.

Rebecca Fry, Ph.D.

University of North Carolina at Chapel Hill, Gillings School of Global Public Health

Dr. Rebecca Fry is the Carol Remmer Angle Distinguished Professor and Associate Chair of the Department of Environmental Sciences and Engineering at the Gillings School of Global Public Health at UNC-Chapel Hill. She uses multi-omic techniques with a particular emphasis on the epigenome to identify mechanisms of environmentally mediated disease.

James Gauderman, Ph.D.

University of Southern California, Keck School of Medicine

Dr. Jim Gauderman is Director of the Biostatistics Division in the USC Department of Population and Public Health Sciences. He has conducted methodological research related to the design and analysis of studies of complex traits for over 30 years, with particular focus on gene-environment interactions. He is currently Co-PI of a program project aimed at developing statistical methods for integrative genomics of cancer, and Co-PI of a separate project to identify novel risk factors for colorectal cancer using multi-omics data.

Scarlett Lin Gomez, Ph.D., M.P.H.

University of California, San Francisco, Helen Diller Family Comprehensive Cancer Center

Dr. Scarlett Lin Gomez is Professor and Vice Chair for Faculty Development in the Department of Epidemiology and Biostatistics and Co-Leader of the Cancer Control Program of the Helen Diller Family Comprehensive Cancer Center, at the University of California, San Francisco. She is Director of the Greater Bay Area Cancer Registry, a participant in the NCI SEER (Surveillance, Epidemiology, End Results) program and the California

Cancer Registry. Her research focuses primarily on cancer health disparities/inequities and aims to understand the multilevel drivers of those disparities.

William Hill, Ph.D.

Francis Crick Institute

Dr. William Hill is a postdoctoral fellow in the Francis Crick Institute, Cancer Evolution and Genomic Instability lab. He completed his doctoral research at Cardiff University in pancreatic cancer cell biology using mouse models to investigate cell competition. Since joining the Francis Crick Institute, Dr. Hill has applied these learnings to understand how risk factors, such as air pollution, contribute to lung cancer.

Cathrine Hoyo, Ph.D.

North Carolina State University

Dr. Cathrine Hoyo is the Goodnight Distinguished Innovation Chair and Professor in Biological Sciences and directs the Epidemiology and Environmental Epigenomics Laboratory at North Carolina State University. Her group's research program aims to improve our understanding of how early development influences risk of common chronic diseases, especially those that exhibit racial/ethnic differences in outcomes, including liver-cancer and metabolic diseases. Her group assembled and is following a cohort of new-borns from the first trimester to identify epigenetic targets that are acquired early and contribute to these chronic diseases. In adults, her team is assembling a cohort of adults to identify epigenetic targets contributing to liver cancer disparities that can be harnessed for susceptibility biomarkers.

Dean Jones, Ph.D.

Emory University

Dr. Dean Jones is a Professor of Medicine and Director of the Clinical Biomarkers Laboratory at Emory University. He has formal training in chemistry, biochemistry, nutrition and molecular toxicology, and more than 40 years directing an academic research program on redox biology and environmental health. Over the past decade, he advanced the use of ultra-high resolution mass spectrometry for high-throughput metabolomics. This research established methods to measure metabolites in most metabolic pathways, as well as thousands of environmental chemicals, dietary, microbiome and related metabolites, in human plasma and urine. Through development of advanced computational methods, this approach now provides an affordable platform for precision medicine, including biomonitoring of dietary supplements, personal use products and environmental chemical exposures. He has authored or co-authored more than 600 research articles and these papers have been cited more than 65,000 times.

Peter Kraft, Ph.D.

Harvard University, T.H. Chan School of Public Health

Dr. Peter Kraft's research concentrates on devising and applying statistical techniques to large-scale observational studies of genetic and circulating biomarkers, with particular emphasis on studies understanding the joint contribution of germline DNA variation, environmental exposures, and biomarkers to risk of complex disease. He has participated in multiple international consortia studying genetics and other exposures in relation to breast, prostate, and pancreatic cancer risk. Dr. Kraft currently co-leads a project developing risk models for breast cancer that combine questionnaire, imaging, genetic and biomarker data, and a project examining the relationship between pre-diagnostic exposures and tumor signatures in triple-negative tumors.

Anshul Kundaje, Ph.D.

Stanford University

Dr. Anshul Kundaje is an Associate Professor of Genetics and Computer Science at Stanford University. The Kundaje lab develops interpretable machine learning and deep learning models to decipher gene regulation and the genetic and molecular basis of disease from multi-modal genomic and molecular profiling experiments. Dr. Kundaje completed his Ph.D. in Computer Science in 2008 from Columbia University. During his postdoctoral research at Stanford University from 2008-2012 and MIT/Broad Institute from 2012-2014, he led the integrative analysis efforts of The Encyclopedia of DNA Elements (ENCODE) and The Roadmap Epigenomics Consortia.

Genevieve Leyden, Ph.D.

University of Bristol

Dr. Genevieve Leyden is an early career researcher working as a Research Associate at the MRC Integrative Epidemiology Unit at the University of Bristol. In 2018, Dr. Leyden was awarded a 4-year studentship from the British Heart Foundation and pursued a cross disciplinary Ph.D. specializing in genetic epidemiology and molecular genetics at the University of Bristol. Her current focus is on the development of methodology for the integration of multi-omics datasets into Mendelian randomization frameworks. Prior to this, after completing an undergraduate degree in Human Genetics in 2016 (Trinity College Dublin), she worked as a research assistant at the Wellcome Sanger Institute (Cambridge, UK) in a role focused on the generation of patient derived cancer organoid models, contributing to the Human Cancer Models Initiative (HCMI).

Francesca Luca, Ph.D.

Wayne State University

Dr. Francesca Luca's research focuses on understanding the genetic and environmental components of human variation in molecular and complex phenotypes of biomedical interest. Dr. Luca performed her graduate studies in Population Genetics at the University of Calabria in Italy, while collaborating with the Musee de l'Homme and the Pasteur Institute in Paris and continued her training as a postdoctoral fellow at the University of Chicago, Department of Human Genetics. She develops and applies high-throughput experimental approaches followed by genomic analysis to identify functional regulatory variants associated with complex traits and their underlying mechanisms. For her research, she has received funding from NIGMS, the American Heart Association, NHLBI, and NIEHS.

Carmen Marsit, Ph.D.

Emory University, Rollins School of Public Health

Dr. Carmen J. Marsit is Executive Associate Dean for Faculty Affairs and Research Strategy, Rollins Distinguished Professor of Research, and Professor in the Gangarosa Department of Environmental Health and Department of Epidemiology at the Rollins School of Public Health of Emory University. He leads a multi-disciplinary research program focused on understanding the impacts of the pre- and perinatal environments on maternal and child health, utilizing genomics, epigenomics, and bioinformatics to uncover mechanisms underlying the impact of the environment on health within epidemiologic studies. Dr. Marsit was the recipient of an NIMH Biobehavioral Research Award for Innovative New Scientists. His current projects are examining the impacts of maternal structural, psychosocial, and chemical exposures on the transcriptomes and epigenomes of the placenta in populations in the United States and in Thailand. He also has an extensive record of research in the utilization of epigenetic biomarkers to understand the etiology and outcomes of human exposure related cancers. Dr. Marsit serves as Director of the NIEHS-funded Emory HERCULES Exposome Research Center and Training Program in the Environmental Health Sciences and Toxicology and was the founding Director of the Emory-Georgia Clean Air Research and Education Program in the Republic of Georgia. Dr. Marsit received his B.S. in Biochemistry from Lafayette College and his Ph.D. in the Biological Sciences in Public Health from the Graduate School of Arts and Sciences at Harvard University.

Kimberly McAllister, Ph.D.

National Institute of Environmental Health Sciences

Dr. Kimberly McAllister received a B.S. in honors biology at the University of Illinois and a Ph.D. in human genetics at the University of Michigan. Her Ph.D. dissertation involved identifying the first gene known to cause the disease Hereditary Hemorrhagic Telangiectasia. She completed postdoctoral training on a Department of Defense breast cancer grant in the Division of Intramural Research at NIEHS with research focusing on the development of BRCA2-deficient mice as a model for breast cancer and Fanconi Anemia. Dr. McAllister is presently a program administrator in the extramural division of NIEHS in the Genes and Environment Health Branch. She manages a portfolio of grants in genetic epidemiology and gene-environment interaction studies, human genetics, GxE statistical and bioinformatics methods, basic genetics, DNA repair, animal models of human disease, and comparative biology and population-based model approaches. She represents NIEHS on multiple trans-NIH committees and large trans-NIH consortium efforts.

Leah Mechanic, Ph.D., M.P.H.

National Cancer Institute

Dr. Leah Mechanic is a Program Director in the Genomic Epidemiology Branch (GEB) of the Epidemiology and Genomics Research Program (EGRP) in NCI's Division of Cancer Control and Population Sciences (DCCPS). Her responsibilities include managing a portfolio of grants related to genetic factors modulating susceptibility to cancer.

Catherine Metayer, M.D., Ph.D.

University of California, Berkeley, School of Public Health

Dr. Catherine Metayer received her medical degree from the University of Bordeaux II in France, and her Ph.D. in Epidemiology from Tulane University, School of Public Health in New Orleans. She is currently an Adjunct Professor of Epidemiology and Biostatistics at the UC Berkeley School of Public Health. Prior to joining UC Berkeley, Dr. Metayer was a scientist at the U.S. National Cancer Institute, Division of Cancer Epidemiology and Genetics. Her work primarily focuses on environmental, dietary, and genetic risk factors of childhood leukemia and testicular cancer, which disproportionately affect the Latinx communities in California. She is the Director of the Center for Integrative Research on Childhood Leukemia and the Environment (CIRCLE), and the immediate past Chair of the Childhood and Cancer Leukemia International Consortium (CLIC). She collaborates with researchers at the intersection of various omics including genetics, epigenetics, metabolomics, and adductomics.

Thomas Metz, Ph.D.

Pacific Northwest National Laboratory

Dr. Thomas (Tom) Metz received a Ph.D. in Chemistry from the University of South Carolina, then joined Pacific Northwest National Laboratory in 2003 for postdoctoral work in mass spectrometry with Dr. Richard D. Smith, where he focused on metabolomics. He became Staff Scientist and a Principal Investigator in the Integrative Omics Group in 2005 and is the Metabolomics Team Lead for a group of scientists that focuses on development and applications of high-throughput metabolomics and lipidomics methods to various biological questions. His research has focused primarily on applying mass spectrometry-based omics approaches, including proteomics, in studies of diabetes mellitus and infectious diseases, resulting in over 180 publications to date. Currently, he is PI of the Pacific Northwest Advanced Compound Identification Core within the NIH Common Fund Metabolomics Program, PI of the Proteomics Laboratory for The Environmental Determinants of Diabetes in the Young consortium, Lead of the PNNL-funded m/q Initiative, and President of the Metabolomics Association of North America.

Gary Miller, Ph.D.

Columbia University, Mailman School of Public Health

Dr. Gary Miller serves as Vice Dean for Research Strategy and Innovation and Professor of Environmental Health Sciences in the Mailman School of Public Health, and Professor of Molecular Pharmacology and Therapeutics in the Vagelos College of Physicians and Surgeons at Columbia University in New York. His laboratory studies the role of environmental factors in neurodegenerative diseases, including Parkinson's disease and Alzheimer's disease. Dr. Miller founded the first exposome center in the U.S. and wrote the first book on the topic. He is a member of the NIH All of Us Research Program Advisory Panel and the National Institute of Environmental Health Sciences Advisory Council. Dr. Miller is the founding editor of the new journal *Exposome*, published by Oxford University Press.

Stephen Montgomery, Ph.D.

Stanford University

Dr. Stephen B. Montgomery is an Associate Professor of Pathology, Genetics, Biomedical Data Science and, by courtesy, Computer Science at Stanford University. His laboratory focuses on both functional genomics and bioinformatics approaches to understanding the molecular origins of genetic diseases. Dr. Montgomery has been a member of multiple large-scale consortia including GREGoR, GTEx and MoTrPAC. His laboratory has developed novel transcriptome-based approaches to identify impactful rare variants in individuals, families, and populations and for measuring gene-by-environment effects.

Kari Nadeau, M.D., Ph.D.

Harvard University, T.H. Chan School of Public Health

Dr. Kari Nadeau is the Chair of the Department of Environmental Health at Harvard School of Public Health and John Rock Professor of Climate and Population Studies. She holds appointments at Harvard Medical School and Stanford University as an adjunct faculty member as of January 2023. She is a pediatrician and practices Allergy, Asthma, Immunology and occupational health in children and adults. She has published over 400 papers, many in the field of climate change and health. Dr. Nadeau is a member of the National Academy of Medicine, the U.S. EPA Children's Health Protection Committee, the CA Governor's Science and Technology Committee, and was appointed as a member of the U.S. Federal Wildfire Commission in 2022. Dr. Nadeau earned her M.D./Ph.D. from Harvard Medical School, completing her doctoral work in biochemistry and immunology, followed by a pediatric internship and residency at Boston Children's Hospital.

Chirag Patel, Ph.D.

Harvard University, Harvard Medical School

Dr. Chirag Patel is an associate professor in the Department of Biomedical Informatics at Harvard Medical School. His long-term research goal is to address problems in human health and disease by developing computational and bioinformatics methods to reproducibly and efficiently reason over high-throughput data streams spanning molecules to populations. Patel's group aims to dissect inter-individual differences in human phenomes through strategies that integrate data sources that capture the comprehensive clinical experience (e.g., through the electronic medical record), the complex phenomena of environmental exposure (e.g., high-throughput measures of the exposome), and inherited genomic variation. He received his Ph.D. in biomedical informatics from Stanford University.

Ulrike Peters, Ph.D., M.P.H.

University of Washington, Fred Hutchinson Cancer Center

Dr. Ulrike (Riki) Peters is Associate Director for Public Health Sciences and Professor at the Fred Hutchinson Cancer Center and Research Professor at the School of Public Health University of Washington. Her research centers on understanding the underlying risk factors of colorectal cancer that will lead to evidence-based targeted interventions and treatments with a specific focus on the impact of race and ethnicity.

David Reif, Ph.D.
National Institute of Environmental Health Sciences
Dr. David Reif joined NIEHS in 2022 as Chief of the Predictive Toxicology Branch in the Division of Translational Toxicology. In this role, he will leverage expertise of the branch in data science, toxicogenomics, spatiotemporal health analytics, computational methods development, and new approach methods (NAMs) to advance predictive toxicology applications with partners across NIEHS, the interagency Tox21 Program, and the Interagency Coordinating Committee on the Validation of Alternative Methods. Prior to joining NIEHS, Dr. Reif was a professor of bioinformatics at North Carolina State University, in the Department of Biological Sciences. His lab focused on integrated analysis of high-dimensional data from diverse sources to understand the complex interactions between human health and the environment.

Marylyn Ritchie, Ph.D.
University of Pennsylvania, School of Medicine
Dr. Marylyn D. Ritchie is a Professor of Genetics, Director of the Center for Translational Bioinformatics, Associate Director for Bioinformatics in the Institute for Biomedical Informatics, and Associate Director of the Center for Precision Medicine at the University of Pennsylvania School of Medicine. Dr. Ritchie is a statistical and computational geneticist with a focus on understanding genetic architecture of complex human disease. She has expertise in developing novel bioinformatics tools for complex analysis of big data in genetics, genomics, and clinical databases, in particular in the area of Pharmacogenomics. Dr. Ritchie has over 15 years of experience in the analysis of complex data and has authored over 250 publications. She has received several awards and honors, including selection as a Genome Technology Rising Young Investigator in 2006, an Alfred P. Sloan Research Fellow in 2010, a KAVLI Frontiers of Science fellow by the National Academy of Science from 2011-2014, and she was named one of the most highly cited researchers in her field by Thomson Reuters in 2014. Dr. Ritchie has extensive experience in all aspects of genetic epidemiology and translational bioinformatics as it relates to human genomics.

Nathaniel Rothman, M.D., M.P.H., M.H.S.
National Cancer Institute
Dr. Nathaniel Rothman received an A.B. in biochemistry and molecular biology at Harvard College and an M.D. at Northwestern University. At the Johns Hopkins University, he received training in internal medicine, occupational and environmental medicine, and occupational and environmental epidemiology. He was Chief Resident in the Division of Occupational Medicine and received an MPH and MHS from the Johns Hopkins Bloomberg School of Public Health. Dr. Rothman joined the intramural research program at the U.S. National Cancer Institute in 1990. He is currently a senior investigator and Head, Molecular Epidemiology Studies in the Occupational and Environmental Epidemiology Branch in the Division of Cancer Epidemiology and Genetics, and an Adjunct Professor at Johns Hopkins, Yale, and Georgetown Universities. His research focuses on using occupational and environmental exposure data paired with biological markers of exposure, early biologic effect, genetic susceptibility, and disease in cross-sectional, case-control and prospective epidemiologic studies of cancer. He is the author of more than 700 publications and was the lead editor of the text Molecular Epidemiology: Principles and Practices. Dr. Rothman conducts research on populations exposed to known or suspected carcinogens including benzene, formaldehyde, chlorinated solvents, aromatic amines, organochlorines, PAHs, and indoor and outdoor air pollution and studies the etiology of lymphoma, leukemia, bladder cancer, and lung cancer.

Marina Sirota, Ph.D.

University of California, San Francisco, Bakar Computational Health Sciences Institute

Dr. Marina Sirota is an Associate Professor at the Bakar Computational Health Sciences Institute at UCSF. Prior to that she worked as a Senior Research Scientist at Pfizer where she focused on developing Precision Medicine strategies in drug discovery. She completed her Ph.D. in Biomedical Informatics at Stanford University. Dr. Sirota's research experience in translational bioinformatics spans over 10 years during which she has co-authored over 100 scientific publications. Her research interests lie in developing computational integrative methods and applying these approaches in the context of disease diagnostics and therapeutics with a special focus on studying the role of the immune system in disease. The Sirota laboratory is funded by NIA, NLM, NIAMS, Pfizer, March of Dimes and the Burroughs Wellcome Fund. As a young leader in the field, she has been awarded the AMIA Young Investigator Award in 2017. Dr. Sirota also is the director of the AI4ALL program at UCSF, with the goal of introducing high school girls to applications of AI and machine learning in biomedicine and serves as the director of outreach and advocacy at the Bakar Computational Health Sciences Institute.

Mary Beth Terry, Ph.D.

Columbia University, Mailman School of Public Health

Dr. Mary Beth Terry is a Professor of Epidemiology and Environmental Health at Columbia University's Mailman School of Public Health. She is a cancer researcher with a primary focus on understanding how cancer susceptibility is modified by environmental and lifestyle factors across the life course. She currently co-leads multi-institutional efforts in New York City to reduce health disparities in multiple chronic diseases and to increase diversity in cancer clinical trials. She also serves as Associate Director of Population Science and Community Outreach for Columbia's Herbert Irving Comprehensive Cancer Center. She teaches epidemiological methods and data science to public health students, medical students, and undergraduate students. Dr. Terry serves on the Board of Scientific Counselors and the PDQ Genetics Board for the National Cancer Institute.

Douglas Walker, Ph.D.

Emory University, Rollins School of Public Health

Douglas Walker, Ph.D., is an Associate Professor in the Gangarosa Department of Environmental Health at Emory University. Dr. Walker's research focuses on continued development and application of advanced analytical strategies for measuring the occurrence, distribution and magnitude of previously unidentified environmental exposures and assisting in delineating the mechanisms underlying environment-related diseases in humans. The approaches he developed show it is possible to measure over 100,000 chemical signals that include exposure biomarkers, nutrients, dietary chemicals and associated biological response in a high-throughput and cost-effective manner, establishing a foundation for operationalizing the exposome framework for precision medicine. Ongoing research projects are now focused on using high-throughput exposome methods to establish disease-exposome atlases, and development of methods for measuring biomarkers of complex exposures of emerging concern, including microplastics, e-waste and polyfluorinated chemicals. Dr. Walker leads the Comprehensive Laboratory for Untargeted Exposome Science (CLUES), which was established to provide high-quality, untargeted screening of biological samples for nutrition, precision medicine and environmental health research.

Sophia Wang, Ph.D.

City of Hope Comprehensive Cancer Center

Dr. Sophia Wang is a Professor at the City of Hope Comprehensive Cancer Center within the Department of Health Analytics. Throughout her career, she has integrated genetics, molecular characteristics, and exposures to gain an understanding of gene-molecular-environment interactions in disease and cancer etiology. She leads the evaluation of gene-environment interactions within the International Lymphoma (InterLymph) Consortium. As part of her research, she also aims to identify intermediate biomarkers associated with environmental exposures such as air pollution and pesticides, and further identify exposures associated with specific tumor molecular characteristics. Dr. Wang completed her training at the Massachusetts Institute of Technology (B.S.) and The Johns Hopkins Bloomberg School of Public Health (Ph.D.).

Ivana Yang, Ph.D.

University of Colorado, Anshutz Medical Campus

Dr. Ivana V. Yang is a tenured Professor and Vice Chair in the Department of Biomedical Informatics at the University of Colorado Anschutz Medical Campus. Dr. Yang's research program uses genetics, transcriptomics, epigenomics and animal/cell models of disease to enhance early detection, predict outcomes, develop biomarkers, and design personalized therapeutic strategies in lung disease. Specific current disease areas of interest include pulmonary fibrosis, chronic beryllium disease, sarcoidosis, and asthma and allergy in underrepresented minority populations.

Appendix 2: Workshop Agenda

All times listed are Eastern Standard Time

Day 1: Tuesday, February 14, 2023

11:00 a.m. **Setting the Stage**

Trevor Archer, Ph.D., Deputy Director, National Institute of Environmental Health Sciences (NIEHS); National Institutes of Health (NIH) Distinguished Investigator

Gary Ellison, Ph.D., M.P.H., Deputy Director, Division of Cancer Control and Population Sciences, National Cancer Institute (NCI)

11:10 a.m. **Purpose and Outcomes**

Kimberly McAllister, Ph.D., NIEHS

Leah Mechanic, Ph.D., MPH, NCI

11:20 a.m. **Session I: Specific Cancer Considerations**

Moderator: Ulrike Peters, Ph.D., M.P.H., University of Washington, Fred Hutchinson Cancer Center

11:20 a.m. **Cancer Susceptibility and Environmental Exposures: Why Study Designs Matter**

Mary Beth Terry, Ph.D., Columbia University, Mailman School of Public Health

11:50 a.m. **Disentangling the Etiological Pathways Between Body Mass Index and Site-Specific Cancer Risk Using Tissue-Partitioned Mendelian Randomization**

Genevieve Leyden, Ph.D., University of Bristol, Bristol Medical School

12:20 p.m. **Session I Panel Discussion and Audience Questions**

Panelists:

- Catherine Metayer, M.D., Ph.D., University of California, Berkeley, School of Public Health
- Sophia Wang, Ph.D., City of Hope Comprehensive Cancer Center
- Peter Kraft, Ph.D., Harvard University, T.H. Chan School of Public Health
- Cathrine Hoyo, Ph.D., North Carolina State University

1:05 p.m. **Lunch Break**

1:35 p.m. **Session II: Computational Approaches**

Moderator: Andrea Baccarelli, M.D., Ph.D., Columbia University, Mailman School of Public Health

1:35 p.m. **Statistical Approaches for Integrating Environmental and Omics Data in Cancer- Epidemiology Studies**

James Gauderman, Ph.D., University of Southern California, Keck School of Medicine

2:05 p.m. **Longitudinal Multi-Omic Characterization of a Community Cohort After Chemical Exposures From Hurricane Harvey**

Cristian Coarfa, Ph.D., Baylor University, Baylor College of Medicine

2:35 p.m. **Break**

2:50 p.m. **Session II Panel Discussion and Audience Questions**

Panelists:

- Marylyn Ritchie, Ph.D., University of Pennsylvania, Perelman School of Medicine
- Nilanjan Chatterjee, Ph.D., Johns Hopkins University, Bloomberg School of Public Health
- Marina Sirota, Ph.D., University of California, San Francisco, Bakar Computational Health Sciences Institute

3:35 p.m. **Open Discussion**

4:00 p.m. **Day 1 Closing**

Leah Mechanic, Ph.D., M.P.H., NCI

Day 2: Wednesday, February 15, 2023

11:00 a.m. **Day 2 Introduction**

Kimberly McAllister, Ph.D., NIEHS

11:05 a.m. **Session III: Integration of Environmental Data with Other Data Types**

Moderator: Gary Miller, Ph.D., Columbia University, Mailman School of Public Health

11:05 a.m. **Dose Versus Burden in Exposome Research: Translation of Integrative Omics of Model Systems to Environmental Epidemiology in Cancer Research**

Dean Jones, Ph.D., Emory University, Winship Cancer Institute

11:35 a.m. **Examples of Omics Integration in Studies of the Environment on Human Health**

Carmen Marsit, Ph.D., Emory University, Rollins School of Public Health

12:05 p.m. **Session III Panel Discussion and Audience Questions**

Panelists:

- Scarlett Gomez, Ph.D., M.P.H., University of California, San Francisco, Helen Diller Family Comprehensive Cancer Center
- Thomas Metz, Ph.D., Pacific Northwest National Laboratory
- Douglas Walker, Ph.D., Emory University, Rollins School of Public Health
- Ivana Yang, Ph.D., University of Colorado, Anschutz Medical Campus
- Nathaniel Rothman, M.D., M.P.H., M.H.S., NCI

12:50 p.m. **Lunch Break**

1:15 p.m. **Session IV: Experimental Models and Functional Approaches**

Moderator: Stephen Montgomery, Ph.D., Stanford University

1:15 p.m. **Introductory Remarks**

Stephen Montgomery, Ph.D., Stanford University

1:25 p.m. **Mechanism of Action and Inflammatory Axis for Air Pollution-Induced Non-Small Cell Lung Cancer**

William Hill, Ph.D., The Francis Crick Institute

1:55 p.m. **Deciphering Mechanisms Through Which Environmental Risk Factors Mediate Colorectal Cancer Risk Through Weighted Gene Co-Expression Networks**

Matthew Devall, Ph.D., University of Virginia

2:25 p.m. **Session IV Panel Discussion and Audience Questions**

Panelists:

- David Reif, Ph.D., NIEHS
- Rebecca Fry, Ph.D., University of North Carolina at Chapel Hill, Gillings School of Global Public Health
- Francesca Luca, Ph.D., Wayne State University
- Justin Colacino, Ph.D., University of Michigan, School of Public Health

3:10 p.m. **Break**

3:30 p.m. **Closing Session: Integration of All Four Themes and Future Directions**

3:30 p.m. **Summary**

Chirag Patel, Ph.D., Harvard University, Harvard Medical School

3:50 p.m. **Final Open Discussion**

Moderator: Kari Nadeau, M.D., Ph.D., Harvard University, T.H. Chan School of Public Health

4:30 p.m. **Closing Remarks**

Kimberly McAllister, Ph.D., NIEHS

Leah Mechanic, Ph.D., M.P.H., NCI

Appendix 3: Key Publications

Reviews

1. Adkins-Jackson, P.B., et al., *Measuring Structural Racism: A Guide for Epidemiologists and Other Health Researchers*. Am J Epidemiol, 2022. 191(4): p. 539-547.
2. Everson, T.M. and C.J. Marsit, *Integrating -Omics Approaches into Human Population-Based Studies of Prenatal and Early-Life Exposures*. Curr Environ Health Rep, 2018. 5(3): p. 328-337.
3. Ghosh, D., et al., *Leveraging Multilayered "Omics" Data for Atopic Dermatitis: A Road Map to Precision Medicine*. Front Immunol, 2018. 9: p. 2727.
4. Gillette, M.A., et al., *Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma*. Cell, 2020. 182(1): p. 200-225.e35.
5. Graw, S., et al., *Multi-omics data integration considerations and study design for biological systems and disease*. Mol Omics, 2021. 17(2): p. 170-185.
6. Gruszecka-Kosowska, A., A. Ampatzoglou, and M. Aguilera, *Integration of Omics Approaches Enhances the Impact of Scientific Research in Environmental Applications*. Int J Environ Res Public Health, 2022. 19(14): p. 8758.
7. López de Maturana, E., et al., *Challenges in the Integration of Omics and Non-Omics Data*. Genes (Basel), 2019. 10(3).
8. Noble, A.J., et al., *A Final Frontier in Environment-Genome Interactions? Integrated, Multi-Omic Approaches to Predictions of Non-Communicable Disease Risk*. Front Genet, 2022. 13: p. 831866.
9. Price, E.J., et al., *Merging the exposome into an integrated framework for "omics" sciences*. iScience, 2022. 25(3): p. 103976.
10. Ritchie, M.D., et al., *Methods of integrating data to uncover genotype–phenotype interactions*. Nature Reviews Genetics, 2015. 16(2): p. 85-97.
11. Tarazona, S., A. Arzalluz-Luque, and A. Conesa, *Undisclosed, unmet and neglected challenges in multi-omics studies*. Nature Computational Science, 2021. 1(6): p. 395-402.
12. Xiao, Y., et al., *Multi-omics approaches for biomarker discovery in early ovarian cancer diagnosis*. eBioMedicine, 2022. 79: 104001.
13. Shah, R.V., et al., *Dietary metabolic signatures and cardiometabolic risk*. Eur Heart J, 2022. ehac446.
14. Davies, N.M., M.V. Holmes, and G.D. Smith, *Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians*. BMJ, 2018. 362: k601.
15. Fang, Z., et al., *The Role of Mendelian Randomization Studies in Deciphering the Effect of Obesity on Cancer*. J Natl Cancer Inst, 2022. 114(3): p. 361-371.
16. Zeinomar, N., et al., *Environmental exposures and breast cancer risk in the context of underlying susceptibility: A systematic review of the epidemiological literature*. Environ Res, 2020. 187: 109346.
17. Terry, M.B., et al., *Environmental exposures during windows of susceptibility for breast cancer: a framework for prevention research*. Breast Cancer Res, 2019. 21(1): 96.
18. Huls, A. and D. Czamara, *Methodological challenges in constructing DNA methylation risk scores*. Epigenetics, 2020. 15(1-2): p. 1-11.

19. Yalcin, G.D., et al., *Systems Biology and Experimental Model Systems of Cancer*. J Pers Med, 2020. 10(4): 180.

20. Pfohl, U., et al., *Precision Oncology Beyond Genomics: The Future Is Here—It Is Just Not Evenly Distributed*. Cells, 2021. 10(4): 928.

Applications:

21. Kehm, R.D., et al., *Associations of prenatal exposure to polycyclic aromatic hydrocarbons with pubertal timing and body composition in adolescent girls: Implications for breast cancer risk*. Environ Res, 2021. 196: 110369.

22. Shen, J., et al., *Dependence of cancer risk from environmental exposures on underlying genetic susceptibility: an illustration with polycyclic aromatic hydrocarbons and breast cancer*. Br J Cancer, 2017. 116(9): p. 1229-1233.

23. Peng, C., et al., *A latent unknown clustering integrating multi-omics data (LUCID) with phenotypic traits*. Bioinformatics, 2020. 36(3): p. 842-850.

24. Dixon, H.M., et al., *Discovery of common chemical exposures across three continents using silicone wristbands*. R Soc Open Sci, 2019. 6(2): 181836.

25. Samon, S.M., et al., *Associating Increased Chemical Exposure to Hurricane Harvey in a Longitudinal Panel Using Silicone Wristbands*. Int J Environ Res Public Health, 2022. 19(11): 6670.

26. Oluyomi, A.O., et al., *Houston hurricane Harvey health (Houston-3H) study: assessment of allergic symptoms and stress after hurricane Harvey flooding*. Environ Health, 2021. 20(1): 9.

27. Dutta, D., et al., *Aggregative trans-eQTL analysis detects trait-specific target gene sets in whole blood*. Nat Commun, 2022. 13(1): 4323.

28. Boye, C., et al., *Characterization of caffeine response regulatory variants in vascular endothelial cells*. bioRxiv, 2022. doi: 10.1101/2022.11.22.517533.

29. Balmain, A., *The critical roles of somatic mutations and environmental tumor-promoting agents in cancer risk*. Nat Genet, 2020. 52(11): p. 1139-1143.

30. Bissell, M.J. and W.C. Hines, *Why don't we get more cancer? A proposed role of the microenvironment in restraining cancer progression*. Nat Med, 2011. 17(3): p. 320-329.

31. Freedman, A.N., et al., *The placenta epigenome-brain axis: placental epigenomic and transcriptomic responses that preprogram cognitive impairment*. Epigenomics, 2022. 14(15), p. 897-911.

32. Santos Jr., H.P., et al., *Evidence for the placenta-brain axis: multi-omic kernel aggregation predicts intellectual and social impairment in children born extremely preterm*. Mol Autism, 2020. 11(1): 97.

33. Garcia, A.L.C., V.M. Arlt, and D.H. Phillips, *Organoids for toxicology and genetic toxicology: applications with drugs and prospects for environmental carcinogenesis*. Mutagenesis, 2022. 37(2): p. 143-154.

34. Cao, Z.J. and G. Gao, *Multi-omics single-cell data integration and regulatory inference with graph-linked embedding*. Nat Biotechnol, 2022. 40(10): p. 1458-1466.

Appendix 4: Other Key Resources

Title	Description	Web Link
Cancer Epidemiology Descriptive Cohort Database (CEDCD)	The NCI Cancer Epidemiology Descriptive Cohort Database (CEDCD) contains descriptive information about cohort studies that follow groups of persons over time for cancer incidence, mortality, and other health outcomes. The CEDCD is a searchable database that contains general study information (e.g., eligibility criteria and size), the type of data collected at baseline, cancer sites, number of participants diagnosed with cancer, and biospecimen information. All data included in this database are aggregated for each cohort; there are no individual level data.	https://cedcd.nci.nih.gov/
Cohorts for Environmental Exposures and Cancer Risk (CEECR)	The New Cohorts for Environmental Exposures and Cancer Risk (CEECR) is a collaborative effort co-funded by NIEHS and NCI that aims to support innovative scientific research in new prospective cohorts that address knowledge gaps in cancer etiology and carcinogenesis processes with a focus on environmental exposures. The CEECR Consortium is made up of five new prospective cohorts that are interested in diverse study populations representative of various regions throughout the Unites States.	https://ceecr.org/
Environmental Health Language Collaborative	The NIEHS Environmental Health Language Collaborative is a new initiative to advance community development and application of a harmonized language for describing Environmental Health Science (EHS) research. Applying a harmonized language to environmental health data enhances its value by enabling consistent classification of data, increasing interoperability of databases, and promoting data sharing, reuse, and reanalysis; thereby, accelerating the pace of scientific discovery in EHS research.	https://www.niehs.nih.gov/research/programs/ehlc/index.cfm

Title	Description	Web Link
GWAS Catalog	The Catalog was founded by NHGRI in 2008, in response to the rapid increase in the number of published genome-wide association studies (GWAS). The GWAS Catalog provides a consistent, searchable, visualizable, and freely available database of SNP-trait associations, which can be easily integrated with other resources, and is accessed by scientists, clinicians and other users worldwide.	https://www.ebi.ac.uk/gwas/home
MR-Base	MR-base is a database and analytical platform for Mendelian randomization being developed by the Integrative Epidemiology Unit at the University of Bristol.	https://www.mrbase.org/
NCI Cohort Consortium	The NCI Cohort Consortium is an extramural-intramural partnership formed to address the need for large-scale collaborations to pool the large quantity of data and biospecimens necessary to conduct a wide range of cancer studies. Through its collaborative network of investigators, the Consortium provides a coordinated, interdisciplinary approach to tackling important scientific questions, economies of scale, and opportunities to quicken the pace of research.	https://epi.grants.cancer.gov/cohort-consortium/
PhenX Toolkit	The PhenX Toolkit (consensus measures for **Phen**otypes and e**X**posures) provides recommended standard data collection protocols for conducting biomedical research. The Toolkit provides detailed protocols for collecting data and tools to help investigators incorporate these protocols into their studies. Using protocols from the PhenX Toolkit facilitates cross-study analysis, potentially increasing the scientific impact of individual studies.	https://www.phenxtoolkit.org/
Surveillance, Epidemiology, and End Results (SEER) Data & Software	NIH SEER research data include SEER incidence and population data associated by age, sex, race, year of diagnosis, and geographic areas (including SEER registry and county). SEER releases new research data every Spring based on the previous November's submission of data.	https://seer.cancer.gov/data-software/

Title	Description	Web Link
UK Biobank	UK Biobank is a large-scale biomedical database and research resource, containing in-depth genetic and health information from half a million UK participants. The database is regularly augmented with additional data and is globally accessible to approved researchers undertaking vital research into the most common and life-threatening diseases. It is a major contributor to the advancement of modern medicine and treatment and has enabled several scientific discoveries that improve human health.	https://www.ukbiobank.ac.uk/