ILLINOIS

Personalized Nutrition Initiative
Carle R. Woese Institute for Genomic Biology, Room 3002
University of Illinois at Urbana-Champaign
1206 West Gregory Drive
Urbana, IL 61801

December 15, 2020

Holly Nicastro, PhD, MPH and Christopher Lynch, PhD
Trans-NIH Nutrition for Precision Health Working Group
Office of Strategic Coordination
Bethesda, MD 20892
nutritionresearch@nih.gov

Dear Drs. Nicastro and Lynch,

On behalf of the Personalized Nutrition Initiative at Illinois and my colleagues (**Appendix 1**) at the University of Illinois at Urbana-Champaign (UIUC), I would like to thank you for the opportunity to respond to the ***Data Science Challenges and Opportunities in the Field of Precision Nutrition*** RFI (NOT-RM-21-005). The Personalized Nutrition Initiative was recently established in response to a call in the 2018-2023 University of Illinois Strategic Plan, which highlighted personalized nutrition as an area for strategic investment to enrich interdisciplinary connections and establish new resources and facilities to expand our campus's strength in food, nutrition, energy, health sciences, and cancer. We are thrilled that Precision Nutrition is the focus of the 2020-2030 NIH Nutrition Strategic Plan.

Pertinent to the RFI, the depiction of biological and imaging systems through the integration of 'omics and image data requires appropriate mathematical and statistical methodologies to infer and describe causal links between different subcomponents. Aside from the computational complexity of analyzing thousands of measurements, the extraction of correlations as true and meaningful biological interactions is not trivial.
Biomedical and omics datasets are complex and heterogeneous, and extracting meaningful knowledge from this vast amount of information is by far the most important challenge for bioinformatics and machine learning researchers (**1**). Biological systems include non-linear interactions and joint effects of multiple factors that make it difficult to distinguish signals from random errors (**2**).

A unique comprehensive strategy that automates data-driven analytical model building will need to be developed that incorporates the unique iterative integration of large-scale clinical record mining, 'omic analyses, hypothesis-based modeling, simulation, and advanced machine-learning approaches (**3**). This comprehensive, and daunting, approach will be required to make tangible progress toward personalized nutrition and precision medicine. For example, data-driven approaches that utilize multiparameter measures, such as the influence of the nutrients on gene expression, genetic variations, and interaction with environmental factors such as the influence of lifestyle measures on gut microbiome interaction could provide a comprehensive understanding of an individual's health.

Integration of systems science, data science, engineering, and computational analytics will be required to analyze complex data sets with variable formats and origins to achieve the NIH's vision of Nutrition for Precision Health (NPH). In this realm, we feel that the highest priority consideration is for ***NIH to provide leadership in creating standards for data and image collection, processing, storage, fusion, and integration.*** Establishing a ***NIH Tool Box for NPH*** will accelerate progress, increased validity, and reproducibility, and will facilitate the combination of data from other existing and future human studies with data from the *All of Us* Research Program to expand the evidence base and statistical power to identify factors important for NPH. For this purpose, NIH should ***build upon existing resources (see below) and expertise at NIH and consider collaborating with existing organizations*** (e.g., the International Society for Nutrigenetics and Nutrigenomics and the Nutrigenomics Organisation ([NuGo] as well as the numerous engineering and physics organizations that have significant life science expertise).

1. **Comments or caveats on disparate data type or format collection and needs as to how they could be made 'AI-ready'.**
   a. **Standardized Protocols:** We recommend that NIH take this opportunity to expand the *All of Us* Program Researcher Workbench to ***develop a comprehensive and detailed data curation and validation pipeline plan with detailed Standardized Protocols***. This could include recommended validated surveys and instruments to use as well as recommendations for how to approach problems related to discrepancies in data collection and inconsistencies across studies. The Standardized Protocols should be accompanied by a **Reference Manual** to support cross-disciplinary communication, common language, and collaboration.

      One example is microbiome analysis. A consistent strategy for data generation and analysis that supports the overarching research objectives should be implemented for all samples to mitigate biases in metagenomic and metatranscriptomic sequence analysis. These analyses are powerful tools for understanding the operation of microbial communities, but are subject to considerable error and bias due to differences in sample collection (**4**), extraction method (**5**), library preparation (**5,6**), sequencing platform and depth (**7**), as well as bioinformatic analysis (**8**). Also, internal microbial standards should be integrated into the experimental design to quantify extraction and sequencing biases (**9**). There are several commercially-available bacterial and viral community standards that could be used, ***but we suggest the non-commercial, widely available standards be developed*** (***see below***). External contamination should also be evaluated through the inclusion of appropriate negative controls throughout extraction, library preparation, and sequencing (**10**). These factors could be incorporated into a Reference Manual.

      Another example is quantitative ultrasound (QUS) imaging that yields liver fat content **(11-12**). Nonalcoholic fatty liver disease (NAFLD) is the most common chronic liver disease worldwide affecting about 25% of the human population and needs to be considered relative to metabolic disease. Real-time liver fat content assessment using QUS imaging uses various data analysis strategies (one being a deep learning approach) that, at UIUC, has merged with microbiome research activities.

   b. **Reference Standards:** Given the breadth of biological data that will be collected as part of the NPH, we also recommend that NIH work with the **National Institute for Biological Standards and Control (NIBSC)** and/or **The Center for Bioanalytic Metrology** (**13**) to develop reference standards for use in the *All of Us* Program and other NIH-funded, investigator-initiated grants. There are currently no accredited or certified reference materials available for the microbiome field. Amos and colleagues have observed that different bioinformatics tools introduced biases, with a trade-off occurring between sensitivity and the relative abundance of false positives in the final dataset (**14**). Going forward, the authors recommended the microbiome field use site-specific reagents of high complexity to ensure pipeline benchmarking is fit for purpose (**14**).

c. **Verification and Validation:** A multidisciplinary group of faculty at the University of Illinois is currently discussing how we can use already available data and biological samples from controlled studies to experimentally test different strategies, including ones that are being used for data collection and analysis in the *All of Us* program (currently and in the future). Gut microbiome DNA isolation techniques is one of the first protocols our group is discussing in regards to standardizing analytical approaches. Gaining a better understanding of the variation introduced by differences in sample collection, processing, and analysis would enable researchers to compare discrepancies in data outcomes from our controlled experimental approaches with the data obtained from *All of Us* cohort or publically available data sets. This would also allow the data to be more "AI-ready" as we could include estimated *'data corrections'* garnered from our experimental approaches.

d. **Longitudinal Data:** Based on our experience in microbiome data analysis, it is better to collect longitudinal data. The microbes composition relies on many factors, which bring a lot of challenges to removing confounding effects and extracting the true signal from the data. The longitudinal data is very helpful for detecting the real association. We also suggest using spike-in when the sequencing or other compositional data is collected. Sometimes data can reveal the false-positive signal because of its compositional nature (**15**). It would be very exciting to integrate different types of data sets, including imaging data that has point-of-care and AI/ML capabilities for quantitative liver fat content.

e. **Data Provenance:** Often, data are transformed and manipulated for analyses and, subsequently, it is not readily clear what has changed in the data. While transformations can be annotated/stored to provide provenance information, it is often labor-intensive to investigate and understand the series of transformations, especially when there is a time gap since the data were last analyzed, and especially if these data are shared. Ideally, it will be important to have visual/interactive ways to understand that provenance.

f. **Transdisciplinary Workshops:** We suggest that NIH organize workshops for leading researchers across disciplines to discuss various aspects of what data and biological samples should be collected, and how these will be handled, stored, and processed in preparation for analysis. The goal of these workshops would be to develop standardized protocols with a common language for all disciplines involved what will be highly encouraged to be used by researchers who receive NIH funding. We should as much as possible learn from what others have already developed, for example, the NIH Common Fund Molecular Transducers of Physical Activity Consortium (**MoTrPAC**) (**16**) and **PhenX** (**17**), which have developed standardized research protocols for data collections.

2. **Consideration of computational and modeling approach challenges, as well as important computational and technical parameters needed to develop algorithms for predicting precision diet recommendations**
   a. **Workflows**: It would be beneficial to develop a workflow that encapsulates the currently disparate and tedious processes required to study these data effectively - especially to facilitate rapid hypothesis generation and testing. An analysis platform could support: 1) the ability to easily create experimental cohorts on the fly from the investigators own private study or in combination with shared/public data; 2) saving/executing complex machine learning experiments effortlessly; and 3) engaging with sophisticated visualization tools to evaluate data and study and communicate results.

b. **Open-Source Data Visualization**: Integrative research, involving the modeling of living systems at different scales, is becoming more extensively used in the biomedical community. For this reason, an open-source library, called MSVTK, is being implemented to fill the gap in software visualization solutions handling multiscale data. The library adopts state-of-the-art visualization and interaction techniques to solve the various challenges. Thus, we encourage NIH to investigate whether existing open-source resources could be useful for precision nutrition and the *All of Us* program.

c. **Patient Metadata:** The microbiome varies across age, ethnicity, and geography (**18-21**). Emerging evidence suggests that phenotypic modeling using the microbiome is highly sensitive to these patient demographics as well as also for world-wide quantitative acceptance of image outcomes. For example, the extrapolation of highly accurate models failed to predict disease in individuals living in different geographic regions (**18**). Care should be taken to incorporate patient ethnicity, geography, and other relevant metadata into predictive models to properly account for the variance that these parameters may introduce.

3 Computational, analytical, system science or modeling resources or tools which NIH should consider adding to the *All of Us* researcher Workbench to leverage the data sets that will be generated by this study.

a. **Dietary Intake Data Collection:** Expanding the *All of Us* researcher Workbench to include protocols for accurate assessment of dietary intake and physical activity as these are vital components for quality research in public health, nutrition, and exercise science and to obtain accurate data to make "AI-ready". Currently, accurate and consistent methodology for the assessment of these components remains a major challenge. Therefore, we suggest that a first step should be for NIH to fund research to develop, optimize, and validate dietary data collection via Apps that include manual entry, selection entry (e.g. choose from a list), semi-automatic (scanning), voice-to-text, photo entry, digital receipts from restaurants or stores, and sensing of eating-related activities through wearables and non-wearables. (**22**). A desirable feature of Apps vs. more traditional dietary data collection will be a "push" feature, which can automatically prompt data entry. However, response to the push declines over time (**23**). Thus, enhancing technology acceptance (**24**), conducting comprehensive evaluation of app quality (**25**), and determining the reliability and validity of dietary apps as matched to the study purpose (e.g. individual data or population-based data, dietary change or monitoring) are important (**26**). Expanding dietary data collection to include several data collection features would enhance data validity.

In addition, NIH should fund studies focused on nutrient biomarker discovery through methods such as metabolomics and point-of-care image-based liver fat assessment as well the improvement of existing nutrient biomarkers. We also support the development of new and the strengthening of existing approaches that incorporate multiple methods of assessment (i.e. Food Frequency Questionnaires, diet records/24h and biomarkers) to accurately estimate dietary intake. Complete feeding studies that manipulate only the food item or food form under study should be used when appropriate (**27-29**). In addition, objective measures of physical activity, including actigraphy, are recommended. Precision Nutrition research should also leverage the resources available in **The PhenX toolkit**, which is a catalog of high-priority measures for consideration and inclusion in genome-wide association studies (GWAS) and other large-scale genomic research efforts (**17**).

b. In terms of the microbiome, there are hundreds of tools designed to analyze amplicon, metagenomic, and metatranscriptomic sequence data. It would be useful for *All of Us* Researcher Workbench to incorporate access to, or the output of a subset of these algorithms. For example:

   i. **16S tools**: There is still considerable discussion about the appropriate taxonomic unit to use in amplicon sequencing (**30**), thus users should have access to both amplicon sequence variant (ASV) and operational taxonomic unit (OTU) generating algorithms and or datasets. **DADA2** (**31**) or **QIIME2** (**32**) should be available for generating ASV based amplicon abundance tables and **Mothur** (**33**) for OTU abundances. If shotgun metagenomic data are not available for all samples **PICRUSt** (**34**) would be a useful tool that would allow users to impute metagenomic gene abundances from amplicon data. Tools that are agnostic to traditional taxonomic units, such as **ClaaTU** (**5**), would be useful in identifying phylogenetically linked traits.

   ii. **Metagenomic tools (reference based):** Reference based metagenomic and metatranscriptomic annotation provides a rapid assessment of microbiome operational diversity. This approach involves aligning reads to highly curated reference databases to generate taxonomic and or gene family abundance profiles. **HUMAnN2** (**36**) generates both microbial taxonomic and functional profiles and has been widely implemented for analysis of human fecal samples. **Kraken2** (**37**) also generates information on the compositional abundances of metagenomic communities and allows the user to provide a custom reference database. Regardless of the software, special care should be given to the selection of metagenomic databases, as this decision can substantially influence metagenomic profiles (**38**).

   iii. **Metagenomic tools (reference free):** This initiative has unprecedented potential to advance our knowledge of host-microbiota-nutrient interactions, and with noninvasive image-based liver fat. To realize this potential, we will need to move past referenced based annotation of metagenomes. Reference free methods have the potential to discover new genetic, taxonomic, and metabolic diversity. Because some of the genomic insights provided through reference free tools are novel, these methods can potentially uncover patterns of association that are missed using reference-based methods. This approach involves generating an assembly of the metagenomic or metatranscriptomic data and then aligning sequencing reads back to this assembly to generate coverage estimates. **Megahit** (**39**) is a flexible metagenome assembly algorithm that produces high quality assemblies. Novel taxonomic diversity can be identified in these assemblies using metagenomic binning algorithms (e.g., **DasTool** (**40**) which incorporate knowledge of sequence composition and differential coverage. Open reading frames are then located in these metagenome assembled genomes using **Prodigal** (**41**) and biosynthetic gene clusters identified using **antiSMASH** (**42**). These data can help discover potentially probiotic species that link with host health. These species can then be prioritized for downstream analyses to validate findings from clinical studies.

c. **Survey Validation and Access**: For social sciences, it would be useful for *All of Us* Researcher Workbench to incorporate access to agreed upon and validated surveys that incorporate social sciences and community factors important to use the data for predictions. For example, one researcher on campus currently conducts models from computational social science used to predict adherence to health guidelines, for instance, for diabetes, which can be adapted to represent nutrition decisions and utilized to predict the impact of interventions. Such models can represent nutrition decisions that depend on several multi-domain factors, including ecological, economic, and psychosocial. Moreover, the model can include interactions among the factors and the social actors themselves, including the influence of groupings and homophily. Approaches

from computational social science using Bayesian knowledge bases, as developed by UIUC Information School faculty can be validated and implemented to provide expected nutrition decisions of individuals based on a range of individual, group, and environmental attributes (**43**).

d. **Computer Infrastructure and Text Mining**: We also have a few suggestions to include in the *All of Us* Research Workbench:
   i. Building computational infrastructure that uses standardized data formats, vocabularies, and ontologies to represent, exchange, and reason over existing knowledge (on nutrition, diet, genetics, microbiome, electronic health records, image-based liver fat content, literature, as well as other modalities)
   II. Constructing new ontologies or enhancing existing ones to accommodate new types of nutrition-related knowledge
   III. Developing text mining methods and algorithms to extract and standardize relevant information from electronic health records, literature, and other resources, such as dietary guidelines
   IV. Developing AI/machine learning algorithms that combine personal data with literature knowledge and omics data to make personalized nutrition recommendations and explain these recommendations
   V. Developing methods (including text mining) to curate rigorous, transparent, and reproducible scientific publications on relevant topics

4. Opportunities for the NIH to partner in achieving the goals of the Nutrition for Precision Health program with dot.org-s, dot.com-s or dot.edu-s.

a. NIH should consider how to integrate data from the *All of Us Research* Project with the recently completed **PREDICT2** (Personalized Responses to Dietary Composition Trial 2) study, which was designed as a single-arm mechanistic intervention study and was registered in Clinical Trials.gov (**44**). The study included collaborators from King's College London, Massachusetts General Hospital, Stanford University and, Tufts University with ongoing NPH studies, such as the PREDICT trial in the UK. The study tested whether the gut microbiota affects metabolic responses to diet, weight, and health status in 18- to 70-year-old subjects. Participants were consumed standardized meals on up to 8 days while wearing glucose monitors (Abbott Freestyle Libre) to measure their blood sugar levels. Participants also self-monitored blood glucose at regular intervals and to record their appetite, food, physical activity, and sleep using apps and wearable devices. They collected a fecal and saliva sample before consuming the standardized meals and provided a fasted blood sample at the end of the study period.

b. The **American Gut Project** (**45**) is the world's largest crowd-sourced, citizen science microbiome research project. It is based in the laboratory of Dr. Robin Knight at the University of California San Diego, but is part of The Microsetta Initiative (TMI) (**46**). The mission of the TMI is to collect microbiome samples and rich phenotypic data spanning the world's populations and to couple these collections with educational outreach about microbiome science. Dr. Hannah Holscher at UIUC is working with the Knight lab on expanding and improving upon the quality of data being collected to assess dietary intake.

c. **Biosensors and Devices**: We recommend that NIH consider supporting transdisciplinary collaborations between computational, engineering, clinical, biological, and behavioral scientists to develop ***novel biosensors, quantitative image capabilities, and devices*** to collect physiological, dietary, and behavioral data in real-time with a limited burden to participants. This would allow transdisciplinary Precision Nutrition investigations to conduct longitudinal tracking of the

exposome and host biological fluids using biosensors. Ideally, the biosensors should incorporate sample collection in addition to monitoring, data fidelity, and reproducibility. While approaches such as genome sequencing, RNA-seq, qRT-PCR are powerful and sensitive, their protocols are time-intensive and complex. Precision Nutrition research would benefit from biomolecular analysis techniques that can be implemented in point-of-use settings, and in some cases integrated with personal mobile devices, to enable interfacing with cloud-based service systems. Nutrition biomarkers that can be easily and frequently quantified from noninvasively obtained bodily fluids and properties (blood finger stick, saliva, urine, perspiration, liver fat) can complement wearable sensors that monitor physiological status (heart rate, activity/motion, oximetry, sleep quality) to provide a holistic view of how an individual's environment, diet, sleep, and exercise all contribute to their well-being.

d. **Biomarker Identification:** We recommend NIH to support transdisciplinary collaborations to develop ***novel integrative analytical approaches that are robust and reproducible to ultimately support biomarker identification***. Advanced statistical machine learning algorithms have been developed to quantify the dynamics of microbiome composition (**47**), liver fat (**11,12**) and other outcomes. However, the effectiveness and robustness of such a strategy often rely on the availability of large data repositories. This effort could capitalize upon the NIH Common Fund investment in the **NIH Big Data to Knowledge (BD2K**) (**48**) initiative, which aims to enable biomedical scientists to capitalize more fully on the big data sets being generated by research communities. Precision Nutrition should also capitalize on the new NIH investment in the **Artificial Intelligence for BiomedicaL Excellence (AIBLE)** initiative on AI and machine learning (ML) in biomedicine (**49**). Lastly, continuous monitoring will also generate large amounts of data, which will produce challenges for data storage, analysis, and comparison with appropriate standards (for clinical outcomes). It would be ideal for NIH to support a centralized repository for Precision Nutrition data that could be accessed by researchers. This could be similar to **The Cancer Genome Atlas (TCGA)** cancer genomics program supported by the NCI (**50**).

e. **Industry Partnerships:** There is a broad range of companies, and even industries with a direct interest in precision nutrition, a field still new enough that the university and industrial research efforts are well aligned. We suggest the NIH develop a collaborative program modeled after the NSF's Industry-University Cooperative Research Centers (IUCRC) or NSF's Partnership for Accelerating Cancer Therapies (PACT), perhaps through the Foundation for the NIH (**51,52**). Relevant to this RFI, conversations with potential industry partners reveal common interests in big data analytics and artificial intelligence as well as in non-invasive data collection, monitoring, and validation. Even more important, the field of precision nutrition is still in need of the identification of relevant biomarkers for health and optimal nutrition. Fashioned after the pharmaceutical industry, we expect pre-competitive collaborations among industrial and university partners in precision nutrition would accelerate biomarker discovery, converge innovative efforts, and optimize economic investments. Other areas of interest to potential industrial partners, particularly relevant to this RFI, include more general interests in precision nutrition such as the democratization of precision nutrition and the promotion of food as medicine.

5. **Any other topic the respondent feels is relevant for the NIH to consider in developing this strategic plan.**

   We have identified several additional relevant areas that need strong considerations as NIH moves forward with its 10-year strategic plan in nutrition:

   a. **Ethical Issues**: Advances in precision nutrition can be realized only to the degree that diverse groups contribute to personal data and biological samples. Lack of diversity in genomic research contributes to ungeneralizable results and inequitable distribution of precisions medicine benefits derived from research (**53-55**). As NIH pursues the promise of the Precision Nutrition common fund, special attention should be paid to recruiting underrepresented groups for research participation. By extension then, issues regarding group privacy will also need to be addressed, ethically and technically. Genomic data about related people (Groups) can cause dignitary and tangible harm (**56**). One approach may be to consider re-consenting guidelines was part of the NIH Genomic Data Sharing (GDS) Policy (**57)** to ensure we are keeping to the highest ethical standard. While it is very hard for participants to give blanket consent not knowing what future research will be done, especially with large and comprehensive data and biological sample collections studies, dynamic and group consent technologies and processes may afford participants ethical assurances and control, beyond standard safe harbor practices.

   b. **Privacy:** Racial and ethnic minorities tend to have stronger genomic privacy concerns than caucasian respondents and that may contribute to the lack of diversity in genomic samples (**58,59**). Yet, studies examining whether genomic privacy attitudes influence decisions to participate in genomic research are inconclusive (**60**). People tend to be more willing to participate in genomic research if their data is anonymized (**61**). However, genomic data can be easily de-anonymized, which can lead to the re-identification of study participants (**62,63**) and inferring the participant's phenotypes (**64,65**). The availability of de-anonymized data raises concerns regarding discrimination by insurance companies, among others. While privacy tools may exist to address these issues, if the measures are too strenuous they might burden potential research participants and could hinder research. Overall, ***there is a need to balance usability and security/privacy.*** Research about how privacy attitudes predict participation behavior, in conjunction with the technical mechanisms to facilitate the secure transfer of data and rigorous standards for its privacy can help untangle these issues and improve participation in research and commercial databases such as GA4GH, dbGap, Thousand Genome Project, *All of Us*, etc. In addition, learning how well the consumer understands and trusts the technologies implemented to secure the privacy, anonymity, and control of data will be needed.

   c. **Limitations of Existing Cohort Studies:** In addition to the *All of Us* cohort, there are several other existing NIH cohorts that might be of use. These include the Environmental influences on Child Health Outcomes (ECHO) program (**66**), the Health Professionals Follow-Up Study (HPFS) (**67**), and the Nurses' Health Study (**68**). However, there are significant disadvantages to nesting Precision Nutrition research within existinglongitudinal cohort studies, since none were specifically designed to capture the complexity of the inputs needed to fully comprehend the contributions that shape Precision Nutrition outcomes. Therefore, ***we recommend that NIH consider funding new cohort(s) across the lifespan to investigate how individual variability affects responses to diet and health outcomes rather than trying to adapt existing cohorts***. These cohorts should include all ages as there are unanswered questions across the lifespan. The cohorts must reflect all aspects of ethnic diversity, socioeconomic status, and other social determinants of health. In addition, prospective cohort studies should also consider Mendelian randomization, which is a method of using measured variation in genes of known function to examine the causal effect of a

modifiable exposure on disease in observational studies (**69**). The design provides a method for obtaining unbiased estimates of the effects of a putative causal variable without conducting a traditional randomized trial. The design has a powerful control for reverse causation and confounding, which often impede or mislead epidemiological studies.

In closing, we appreciate the opportunity to assist NIH in identifying ***Data Science Challenges and Opportunities in the Field of Precision Nutrition***. We believe that this is the time for NIH to take leadership in establishing standards for data collection and carpentry to promote consistency and accuracy and, ultimately, data sharing. We also recommend that NIH consider how to leverage the work of several groups, including the International Society for Nutrigenetics and Nutrigenomics and the Nutrigenomics Organisation (NuGo) (**70**), which have been considering for nearly a decade how to conduct precision nutrition studies (**71,72**) and the needs for a "nutritional phenotype database (dbNP)" (**73**).

The University of Illinois has positioned itself as a world leader in tackling computational and modeling grand challenges as the home of four AI research institutes/centers on campus. The University has ~160 faculty from 30 different departments or institutes that are affiliated with at least one AI or computational institute, with about 17 of them having multiple affiliations. In addition, the university collaborates in an additional five AI research institutes at other universities. Thus, we have a breadth and strength of AI researchers on campus, we will look forward to joining with NIH and our colleagues across the country to contribute to this exciting initiative.

Thank you for considering our comments. If we can be of any further assistance, please contact me at sdonovan@illinois.edu or 217-333-2289.

Sincerely yours,

Sharon M. Donovan, PhD RD
Professor and Melissa M. Noel Endowed Chair, Department of Food Science & Human Nutrition
Member, Division of Nutritional Sciences
Director, Personalized Nutrition Initiative
Member, 2020-2025 Dietary Guidelines for American Advisory Committee
Member, National Academy of Medicine

Appendix 1. UIUC Faulty contributors to the RFI Response:

Name and Rank	Affiliations
Jacob Allen, PhD Assistant Professor	Department of Kinesiology and Community Health (**KCH**); Division of Nutritional Sciences (**DNS**)
Jaume Amengaul, PhD Assistant Professor	Department of Food Science & Human Nutrition (**FSHN**); DNS
Nicholas Burd, PhD Associate Professor	Department of KCH; DNS
Colleen Bushell, FAA Associate Director, Healthcare Innovation and Principal Research Scientist, Visual Analytics; Associate Research Professor	National Center for Supercomputing Applications (**NCSA**); Carle-Illinois College of Medicine (**CICOM**)
Brian Cunningham, PhD Professor; Director of Center for Genomic Diagnostics	Department of Electrical & Computer Engineering (**ECE**); Carl R. Woese Institute for Genomic Biology (**IGB**); CICOM
Elvira de Mejia, PhD Professor; Director of DNS	Department of FSHN; DNS
Jessica Dalhaus, PhD Associate Director for Research	College of Engineering – Office of Research
Sharon Donovan PhD RD Professor and Melissa M. Noel Endowed Chair; Director, Personalized Nutrition Initiative	Department of FSHN; DNS; IGB; CICOM
John Erdman, Jr, PhD Professor Emeritus; Deputy Director, Interdisciplinary Health Sciences Institute	Department of FSHN; DNS
Christopher Gaulke, PhD, Assistant Professor	Department of Pathobiology; School of Information Sciences
Cecilia Gentle, PhD IGB Fellow in Economic Development	IGB
Carl Gunter, PhD Professor	Department of Computer Science; IGB Genomic Security & Privacy theme
Hannah Holscher PhD, RD Assistant Professor	Department of FSHN; DNS; IGB
Aiguo Han, PhD Assistant Professor	Department of ECE
Matthew Hudson PhD Professor and Co-Director	Department of Crop Sciences; Center for Digital Agriculture
Eliu Huerta, PhD Senior Research Scientist; Head; Director	Senior Research Scientist; NCSA Gravity Group; NCSA Center for Artificial Intelligence Innovation
Rod Johnson, PhD Professor and Head	Department of Animal Sciences; DNS
Naiman Khan, PhD, RD Assistant Professor	Department of KCH; DNS
Halil Kilicoglu, PhD Associate Professor	School of Information Sciences
Aleks Ksiazkiewicz, PhD Assistant Professor	Department of Political Sciences; IGB Genomic Security & Privacy theme
Soo-Yeun Lee, PhD Professor	Department of FSHN; DNS
Ting Lu, PhD Associate Professor	Department of Bioengineering; IGB

Zeynep Madak-Erdogan PhD Assistant Professor	Department of FSHN; CICOM; DNS; IGB
William O'Brien, PhD Professor Emeritus, Donald Biggar Willet Professor of Engineering	Department of ECE; DNS
Shulei Wang, PhD Assistant Professor	Department of Statistics
Ken Wilund, PhD Professor	Department of KCH; DNS
M. Yanina Pepino PhD Assistant Professor	Department of FSHN; DNS
Lori Raetzman, PhD Associate Professor	Department of Molecular & Integrative Physiology
Christopher Rao, PhD Professor, Morris Professorial Scholar	Department of Chemical & Biomolecular Engineering; IGB
Eunice Santos, PhD Professor and Dean	School of Information Sciences
Saurabh Sinha, PhD Professor and Willett Faculty Scholar	Department of Computer Science; IGB
Stephan Schneider, PhD Fellow	IGB
Sebastian Souyris, PhD Post-Doctoral Research Associate	Gies College of Business
Don Takehara, PhD Associate Director for Research	College of Engineering – Office of Research
Margarita Teran-Garcia MD, PhD Assistant Professor	University of Illinois Extension, CICOM; DNS
Ruoqing Zhu, PhD Assistant Professor	Department of Statistics

Appendix 2: References Cited

1. Martorell-Marugán J, Tabik S, Benhammou Y, del Val C, Zwir I, Herrera F, Carmona-Sáez P. Deep Learning in Omics Data Analysis and Precision Medicine *In: Computational Biology.* Holger Husi (Editor), Codon Publications, Brisbane, Australia. doi: http://dx.doi.org/10.15586/computationalbiology.2019.
2. Alyass, Turcotte, Meyre D. From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Medical Genomics* 2015; 8:33. doi 10.1186/s12920-015-0108-y.
3. Verma M, Hontecillas R, Tubau-Juni N, Abedi V and Bassaganya-Riera J Challenges in Personalized Nutrition and Health. *Front Nutr* 2018; 5:117.doi: 10.3389/fnut.2018.00117.
4. Byrd DA, Sinha R, Hoffman KL, Chen J, Hua X, Shi J, et al. Comparison of methods to collect fecal samples for microbiome studies using whole-genome shotgun metagenomic sequencing. Rao K, editor. *mSphere.* 2020; 26;5(1):e00827-19. doi: 10.1128/mSphere.00827-19.
5. Sui H, Weil AA, Nuwagira E, Qadri F, Ryan ET, Mezzari MP, et al. Impact of DNA Extraction Method on Variation in Human and Built Environment Microbial Community and Functional Profiles Assessed by Shotgun Metagenomics Sequencing. *Front Microbiol*. 2020; 25;11:953. doi: 10.3389/fmicb.2020.00953
6. Bowers RM, Clum A, Tice H, Lim J, Singh K, Ciobanu D, et al. Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. *BMC Genomics*. 2015;16(1):856. doi: 10.1186/s12864-015-2063-6.
7. Nayfach S, Bradley PH, Wyman SK, Laurent TJ, Williams A, Eisen JA, et al. Automated and Accurate Estimation of Gene Family Abundance from Shotgun Metagenomes. Guigo R, editor. *PLOS Comput Biol*. 2015;11(11):e1004573. doi: 10.1371/journal.pcbi.1004573.
8. Shakya M, Lo C-C, Chain PSG. Advances and challenges in metatranscriptomic analysis. *Front Genet*. 2019;10. doi: 10.3389/fgene.2019.00904
9. Hardwick SA, Chen WY, Wong T, Kanakamedala BS, Deveson IW, Ongley SE, et al. Synthetic microbe communities provide internal reference standards for metagenome sequencing and analysis. *Nat Commun*. 2018;9(1):3096. doi: 10.1038/s41467-018-05555-0
10. Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, Gonzalez A, Kosciolek T, McCall LI, McDonald D, Melnik AV, Morton JT, Navas J, Quinn RA, Sanders JG, Swafford AD, Thompson LR, Tripathi A, Xu ZZ, Zaneveld JR, Zhu Q, Caporaso JG, Dorrestein PC. Best practices for analysing microbiomes. *Nat Rev Microbiol*. 2018;16(7):410-22.
11. Han A, Zhang YN, Boehringer AS, Montes V, Andre MP, Erdman JW, Jr, Loomba R, Valasek MA, Sirlin CB, and O'Brien WD, Jr. Assessment of Hepatic Steatosis in Nonalcoholic Fatty Liver Disease Using Quantitative Ultrasound. *Radiology* 2020: 295:106-13.
12. Han A, Byra Heba ME, Andre MP, Erdman JW, Jr, Loomba R, Sirlin CB, and O'Brien WD, Jr. Noninvasive Diagnosis of Nonalcoholic Fatty Liver Disease and Quantification of Liver Fat with Radiofrequency Ultrasound Data Using One-dimensional Convolutional Neural Networks. *Radiology* 2020: 295:342-50.
13. https://cbm.nd.edu/about/
14. Amos GCA, Logan A, Anwar S, Fritzsche M, Mate R, Bleazard T. Rijpkema S. Developing standards for the microbiome. Microbiome 2020; 8:898. /doi.org/10.1186/s40168-020-00856-3
15. Vandeputte D, Kathagen G, D'hoe K, Vieira-Silva S, Valles-Colomer M, Sabino J, Wang J, Tito RY, De Commer L, Darzi Y, Vermeire S, Falony G, Raes J. Quantitative microbiome profiling links gut community variation to microbial load. *Nature.* 2017;551(7681):507-511
16. https://www.motrpac.org
17. https://www.phenxtoolkit.org
18. Deschasaux M, Bouter KE, Prodan A, Levin E, Groen AK, Herrema H, et al. Depicting the composition of gut microbiota in a population with varied ethnic origins but shared geography. *Nat Med*. 2018;24(10):1526–31.
19. He Y, Wu W, Zheng H-M, Li P, McDonald D, Sheng H-F, et al. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. *Nat Med.* 2018; 24(10):1532–5.
20. Gaulke CA, Sharpton TJ. The influence of ethnicity and geography on human gut microbiome composition. *Nat Med*. 2018;24(10):1495–6.
21. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, et al. Human gut microbiome viewed across age and geography. *Nature*. 2012; 486(7402):222-7.
22. Fuchs KL, Haldimann M, Vuckovac D, Ilic A. Automation of data collection techniques for recording food intake: a review of publicly available and well-adopted diet apps. *International Conference on Information and Communication Technology Convergence (ICTC),* 2018, pp. 58-65, doi: 10.1109/ICTC.2018.8539468.

23. Freyne J, Yin J, Brindal E, Hendrie GS, Berkovsky S, Noakes M. Push notifications in diet apps: Influencing engagement times and tasks. *International Journal of Human–Computer Interaction* 2017; 33:833-845.
24. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: Toward a unified view. *MIS Quarterly* 2003*;* 27: 425-478.
25. DiFilippo KN, Huang W, Chapman-Novakofski KM. A new tool for nutrition App quality evaluation (AQEL): Development, validation, and reliability testing. *JMIR Mhealth Uhealth* 2017;5:e163. doi: 10.2196/mhealth.7441
26. Khazen W, Jeanne JF, Demaretz L, Schäfer F, Fagherazzi G. Rethinking the use of mobile Apps for dietary assessment in medical research. *J Med Internet Res* 2020;22(6):e15619. doi:10.2196/15619.
27. Baer, DJ, Gebauer SK, Novotny JA. Walnuts consumed by healthy adults provide less available energy than predicted by the Atwater factors. *J Nutrition* 2016; 146: 9-13.
28. Holscher HD, Guetterman HM, Swanson KS, An R, Matthan NR, Lichtenstein A, Novotny J, Baer DJ. Walnut consumption alters the gastrointestinal microbiota, microbially derived secondary bile acids, and health markers in healthy adults: a randomized controlled trial. *J Nutrition* 2018; 148: 861-867.
29. Tindall AM, McLimans CJ, Petersen KS, Kris-Etherton PM, Lamendella R. Walnuts and vegetable oils containing oleic acid differentially affect the gut microbiota and associations with cardiovascular risk factors: Follow-up of a randomized, controlled, feeding trial in adults at risk for cardiovascular disease. *J Nutrition* 2020; 150: 806-17.
30. Johnson JS, Spakowicz DJ, Hong B-Y, Petersen LM, Demkowicz P, Chen L, et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat Commun*. 2019; 10(1):5029. Available from: http://www.nature.com/articles/s41467-019-13036-1.
31. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods*. 2016;13(7):581–3.
32. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol*. 2019; 37(8):852–7.
33. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, et al. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol*. 2009.
34. Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol*. 2013;31(9):814–21.
35. Gaulke CA, Arnold HK, Humphreys IR, Kembel SW, O'dwyer JP, Sharpton TJ. Ecophylogenetics clarifies the evolutionary association between mammals and their gut microbiota. *MBio*. 2018/
36. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, et al. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods*. 2018;15(11):962–8.
37. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol*. 2019;20(1):257.
38. Tanca A, Palomba A, Fraumene C, Pagnozzi D, Manghina V, Deligios M, et al. The impact of sequence database choice on metaproteomic results in gut microbiota studies. *Microbiome*. 2016;4(1):51. doi: 10.1186/s40168-016-0196-8
39. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*. 2015: 1(10):1674-6. doi: 10.1093/bioinformatics/btv033
40. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol*. 2018;3(7):836-43.
41. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. 2010;11:119. doi: 10.1186/1471-2105-11-119.
42. Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, et al. AntiSMASH 3.0-A comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res*. 2015; 43(W1):W237-43. doi: 10.1093/nar/gkv437.
43. Santos, EE, Santos E, Korah J, Thompson JE, Zhao Y, Murugappan V, Russell JA. Modeling Social Resilience in Communities. *IEEE Transactions on Computational Social Systems* 2018; 5(1), 186-199
44. PREDICT 2: Personalized Responses to Dietary Composition Trial 2 - Full Text View - ClinicalTrials.gov
45. http://americangut.org/
46. https://microsetta.ucsd.edu/
47. Cullen CM, Aneja KK, Beyhan S, Cho CE, Woloszynek S, Convertino M, McCoy SJ, Zhang Y, Anderson MZ, Alvarez-Ponce D, Smirnova E, Karstens L, Dorrestein PC, Li H, Sen Gupta A, Cheung K, Powers JG, Zhao Z, Rosen

GL: Emerging Priorities for Microbiome Research. *Frontiers in Microbiology* 2020: 11:136. doi:10.3389/fmicb.2020.00136.
48. https://commonfund.nih.gov/bd2k
49. https://dpcpsi.nih.gov/sites/default/files/CoC_May_2020_1.05PM_Concept_Clearance_AIBLE_Brennan_508.pdf
50. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
51. Institute of Medicine (US). Building Public-Private Partnerships in Food and Nutrition: Workshop Summary. Washington (DC): National Academies Press (US); 2012. 2, Why Partner? Available from: https://www.ncbi.nlm.nih.gov/books/NBK109695/ https://www.ncbi.nlm.nih.gov/books/NBK97331/pdf/Bookshelf_NBK97331.pdf.
52. NIH News Release 10/12/2017: https://www.nih.gov/news-events/news-releases/nih-partners-11-leading-biopharmaceutical-companies-accelerate-development-new-cancer-immunotherapy-strategies-more-patients.
53. Hindorff LA, Bonham VL, Ohno-Machado L. Enhancing diversity to reduce health information disparities and build an evidence base for genomic medicine. *Personalized Medicine* 2018; 15(5): 403-412.
54. Hindorff LA, Bonham VL, Brody LC, Ginoza ME, Hutter CM, Manolio TA, Green ED. Prioritizing diversity in human genomics research. *Nature Reviews Genetics,* 2018; 19(3): 175-85.
55. Landry, L. G., Ali, N., Williams, D. R., Rehm, H. L., & Bonham, V. L. (2018). Lack of diversity in genomic databases is a barrier to translating precision medicine research into practice. *Health Affairs*, *37*(5), 780-785.
56. Fox, K. (2020). The Illusion of Inclusion—the "*All of Us*" research program and indigenous peoples' DNA. *New England Journal of Medicine 383*(5), 411-413.
57. NIH Genomic Data Sharing Policy, http://gds.nih.gov/03policy2.html.
58. Hull SC, Sharp RR, Botkin JR, Brown M, Hughes M, Sugarman J, et al. Patients' views on identifiability of samples and informed consent for genetic research. *American Journal of Bioethics* 2008; 8(10): 62-70.
59. Nwulia EA, Hipolito MM, Aamir S, Lawson WB, Nurnberger Jr JI, et al. Ethnic disparities in the perception of ethical risks from psychiatric genetic studies. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics,* 2011; 156(5):569-80.
60. Clayton EW, Halverson CM, Sathe NA, Malin BA. A systematic literature review of individuals' perspectives on privacy and genetic information in the United States. *PLoS One* 2018; 13(10): e0204417. doi: 10.1371/journal.pone.0204417.
61. Weidman J, Aurite W, Grossklags, J. On sharing intentions, and personal and interdependent privacy considerations for genetic data: A vignette study. *IEEEACM Transactions on Computational Biology and Bioinformatics* 2018; 16(4): 1349-61.
62. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. *Science,* 2013; 339(6117): 321-4.
63. Lippert C, Sabatini R, Maher MC, Kang EY, Lee S, Arikan O, et al. Identification of individuals by trait prediction using whole-genome sequencing data. *PNAS* 2017; 114(38): 10166-71.
64. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet,* 2008; 4(8): e1000167. doi: 10.1371/journal.pgen.1000167.
65. Wheelers DA, Srinivasan M, Egholm M, Shens Y, Chens L, McGuire A, et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 2008; 452(7189): 872-8.
66. https://www.nih.gov/research-training/environmental-influences-child-health-outcomes-echo-program
67. https://sites.sph.harvard.edu/hpfs/
68. https://www.nurseshealthstudy.org/
69. Goto A, Yamaji T, Sawada N, Momozawa Y, Kamatani Y, Kubo M, Shimazu T, Inoue M, Noda M, Tsugane S, Iwasaki M. Diabetes and cancer risk: A Mendelian randomization study. *Int J Cancer* 2020; 146(3): 712-9.
70. www.nugo.org
71. Ferguson LR, De Caterina R, Görman U, Allayee H, Kohlmeier M, Prasad C, Choi MS, et al. Guide and position of the International Society of Nutrigenetics/Nutrigenomics on personalized nutrition: Part 1 - Fields of precision nutrition. *J Nutrigenet Nutrigenomics* 2016, 9:12-27.
72. Kohlmeier M, De Caterina R, Ferguson LR, Görman U, Allayee H, Prasad C, Kang JX, Nicoletti CF, Martinez JA. Guide and position of the International Society of Nutrigenetics/Nutrigenomics on personalized nutrition: Part 2 - Ethics, challenges and endeavors of precision nutrition. *J Nutrigenet Nutrigenomics* 2016; 9:28-46.

73. van Ommen B, Bouwman J, Dragsted LO, Drevon CA, Elliott R, de Groot P, Kaput J, et al. Challenges of molecular nutrition research 6: the nutritional phenotype database to store, share and evaluate nutritional systems biology studies. *Genes Nutr*. 2010; 5:189-203.