NIH Symposium

Summary

Linking Disease Model Phenotypes to Human Conditions
September 10–11, 2015

Fishers Lane Conference Center
Rockville, MD

Table of Contents

Summary

Day 1: Thursday, September 10, 2015

Introductions and Welcome

Symposium Introductions

Oleg Mirochnitchenko, Ph.D., Office of Research Infrastructure Programs (ORIP), NIH
Harold Watson, Ph.D., ORIP, NIH

Dr. Oleg Mirochnitchenko welcomed all of the participants to the workshop. He thanked the members of the organizing committee for their efforts and the speakers for their participation. He encouraged active participation in the workshop discussions.

In his introduction Dr. Oleg Mirochnitchenko mentioned significant advances in biomedical science have been achieved, including phenotyping of model organisms. Precision medicine is becoming a focus of biomedical research, including phenotyping of research cohorts of human subjects. The need for new interpretive language to describe these phenotypes provided the impetus for organizing this meeting. The Division of Comparative Medicine (DCM) recognizes the need for new resources for cross-species comparisons of phenotypes.

Welcome

Franziska Grieder, D.V.M., Ph.D., ORIP, NIH
Stephanie Murphy, V.M.D., Ph.D., DCM, NIH

Dr. Franziska Grieder, Director of ORIP, also welcomed the participants and thanked them for their attendance. She indicated that Dr. James M. Anderson, Director of the Division of Program Coordination, Planning, and Strategic Initiatives (DPCPSI), is very supportive of this workshop and would have liked to attend, but he had a schedule conflict and sends his regrets. Dr. Grieder stressed how appropriate it is that ORIP and the DPCPSI organized this workshop. The DPCPSI mission is identifying scientific opportunities or knowledge gaps that merit further research, as well as assisting the NIH in addressing such emerging scientific opportunities, supporting cross-cutting, trans-NIH programs that would benefit from strategic planning and organization. ORIP focuses on supporting and enhancing research and research resources that benefit basic, translational, and clinical science and that are of interest to any of the disease entities that the NIH supports. ORIP also is interested in supporting data integration in collaboration with other NIH ICs and the Office of Data Science, which is led by Dr. Philip Bourne. Dr. Grieder also thanked the organizing committee, especially DCM staff; Dr. Mirochnitchenko; Dr. Harold Watson; and Dr. Stephanie Murphy, Director of DCM.

Dr. Murphy indicated that a main focus of DCM's mission is supporting and advancing models of diseases. Disease models involving animals, as well as cell lines and tissues, are some of the most important tools being used in biomedical research. The degree to which disease mechanisms and phenotypes are conserved among animal model species and with humans will dictate the efficacy of the use of these models. One of the biggest obstacles in using animal models to translate biomedical research to clinical practice is the lack of alignment of disease phenotypes, especially at the molecular level, among different species, including humans. Increasing the availability of more precise molecular phenotypes for different diseases and enhancing uniformity in how phenotypes are described and organized will help researchers overcome challenges in translating results from animal models to help human patients. Dr. Murphy thanked the participants in advance for their willingness to engage in the

discussion topics that would be introduced by the meeting speakers and panelists and their contributions to this meeting.

Dr. Watson introduced Dr. Peter Robinson, the keynote speaker. In his research, Dr. Robinson uses mathematical models to understand the biology and genetics of disease. Biologists are realizing that computational approaches are needed to translate biological data into knowledge. One of the important purposes of this meeting is to apply these approaches to phenotyping data from animal models to provide answers about human disease and treatments.

Keynote Presentation
Deep Phenotyping for Translational Research and Precision Medicine
Peter Robinson, M.D., M.Sc., Max Planck Institute for Molecular Genetics

Dr. Peter Robinson described the broader context of why the Human Phenotype Ontology (HPO) is needed. Since its inception, a central theme of bioinformatics has been comparing biological entities to produce information that indicates how similar they are, leading to ranked lists. A similar approach can be used for clinical phenotypes (e.g., influenza, a broken arm) to determine how similar they are. An ontology can be thought of as a hierarchy of terms from general to specific. The HPO contains approximately 11,000 terms with 110,000 annotations for 7,000 mainly monogenic diseases. The HPO is widely used in the rare disease community, in databases, and in the bioinformatics community because it offers substantially better coverage of phenotype concepts than any other terminology.

Ontology algorithms address basic problems of phenotypic descriptions, including difficulty interpreting such descriptions by computers and production of different results when searching on terms that are synonymous. An ontology has been defined as conceptualization in which the interrelation between concepts (e.g., subclasses, instances, or properties) is specified. The average similarity between terms can be used to compare diseases. The human phenome can be depicted as a network of human diseases and disease genes; for each disease phenotype, searching for the most closely matched phenotype will provide an indication of how similar two diseases are. Semantic Web technology allows the researcher to start with a certain amount of knowledge and create new knowledge; analogous algorithms are the basis for human to model organism comparisons. Phenotype ontology is defined as using precise language, interoperability, and database models to reliably capture and interpret phenotype information; medical phenotype ontology describes deviations from normal health through the individual manifestations of disease.

Ontological diagnoses are used by physicians to make diagnoses by searching for semantically similar diseases; an exact match is not needed. The Phenomizer, a freely available web tool for clinical genetics allows a user to enter a phenotype and produces a ranked list of potential diagnoses.

The HPO can be used for translational research, either from basic science to clinical research or clinical research to clinical practice, when it is used to link phenotypic data to other types of data, such as gene function and data from model organisms. In the traditional view of copy number variation (CNV) pathogenesis, for example, the scientific and medical research problem is determining which genes are responsible for phenotypic features. A recent study of Liebenberg syndrome, a rare disease in which the elbow develops like a knee, supports a different model in which a deletion alters the long-range control of gene expression, removing a topological domain barrier (TDB) and leading to ectopic gene expression. Dr. Robinson and his colleagues decided to explore the prevalence of haplo-insufficiency as opposed to TDB disruption in explaining CNVs associated with congenital disease. They assigned tissue-specific enhancers to phenotypic categories. The researchers found that approximately 7 percent of Database of Chromosomal Imbalance and Phenotype in Humans using Ensemble Resources (DECIPHER, a web-based resource for clinical community to share and compare phenotypic and genotypic data) deletion

cases were potentially related to TDB disruptions, with almost 12 percent predicted from an analysis that included model organism phenotype data.

Because obtaining a precise diagnosis for individuals with rare diseases can be difficult, phenotypic analysis was used to try to improve diagnoses. Using patient exome and phenotypic data, the Phenotypic Interpretation of eXomes (PhenIX) identifies predicted pathogenic mutations in the exome and ranks the corresponding genes according to phenotypic relevance. In 10,000 simulations with mutations from the Human Gene Mutation Database® (HGMD®), PhenIX was successful 86 percent of the time in identifying the correct disease cause variant. Cross-species phenotypic analyses on a large scale have the potential to provide insight into the genetic basis of human disease.

The community is beginning to appreciate using phenotypic analysis in diagnosing rare diseases; the next frontier in phenotypic analysis is to explore the genetic origins of common human disease. Genome-wide association studies (GWAS) have identified more than 6,000 strong associations to common complex diseases, and some GWAS hits have been associated with multiple diseases. Often mutations are associated with increased risk, not Mendelian disease. In developing phenotypic networks of common disease, the researchers identified a substantial amount of phenotypic overlap among diseases. Common disease annotations are available for browsing: a disease name can be entered and abstracts associated with the HPO term will be retrieved.

Genomiser, a new web bionformatics tool for prioritization of non-coding variants was developed to aid in understanding noncoding Mendelian mutations. The Synthetic Minority Oversampling Technique (SMOTE) screened for positive associations among a much larger number of negative ones.

In conclusion, Dr. Robinson emphasized the pressing needs and goals for deep phenotyping. In cross-species phenotype analysis, many projects are underfunded, including Monarch (Initiative to integrate, align, and re-distribute cross-species gene, genotype, variant, disease, and phenotype data) and the HPO (Human Phenotype Ontology), and some areas in human and mammalian phenotyping will require extension (e.g., behavior, metabolism). For precision medicine, the current algorithms perform well for rare diseases, but more sophisticated phenotypes will be needed for common diseases; integrated algorithms for matching phenotype to molecular pathophysiology also are needed as well as connecting to a molecular taxonomy of disease; and animal models of common diseases need to be developed. Regarding gene sequencing and noncoding variations, researchers are beginning to explore the role of the entire genome in human disease. Additional opportunities include studies of how phenotype differs from that of coding mutations in rare disease, the phenotypic spectrum of common disease, and annotation of animal models of gene regulation to integrate them in medical analysis.

Discussion

Dr. Warren Kibbe, NCI/NIH, asked how deep phenotyping can impact oncology, particularly how Dr. Robinson's model integrates with small molecule datasets. Dr. Robinson replied that PhenIX could be used to conduct exome analysis. Algorithms assume that there are multiple, intersecting lines of evidence. Better cancer resources may be needed, and phenotype ontologies may have a role to play.

In determining the connection between complex disease and regulatory variants related to Mendelian diseases, different methods of identifying associations between regions and genes are needed. Data are not well annotated regarding enhancers, and this provides an opportunity for an ontology project. In addition, the bio-creation of regulatory elements should be improved.

A participant asked how well this method could predict a non-specific condition (e.g., autism). Dr. Robinson replied that phenotypic tools can narrow down disease candidates but that some cases (e.g., skeletal dysplasia) will not be solvable using these tools.

Session 1: The Current Status of the Human Clinical Phenotype Ontology and Terminology, and Associated Data Annotation and Use
Chair: *Olivier Bodenreider, M.D., Ph.D., National Library of Medicine*

Chair Overview

HPO is being integrated into the Unified Medical Language System® (UMLS®). Integration is bringing together resources that previously were separate, providing new opportunities for use of phenotypic data. Phenotypic data are produced from two major sources: (1) basic research and clinical trials, and (2) clinical care and medical claims data. Many resources are available to help analyze these data sets, and several are described in this session. The round table will include a discussion of how resources can be can reconciled to meaningfully and efficiently integrate data sets collected for research and clinical care purposes.

The PhenX Toolkit: Standard Measures for Collaborative Research
Carol Hamilton, Ph.D., RTI International

Dr. Carol Hamilton introduced the Web-based PhenX Toolkit developed for the biomedical community. The PhenX Toolkit contains measures for phenotypes and exposures that were established using a consensus-based process and driven by the scientific community. Initially, PhenX was focused on measures for GWAS, but the scope has broadened to include translational and clinical research. A common framework (i.e., ontology) is needed that connects PhenX with other resources. In PhenX terminology, a measure is a certain characteristic of, or related to, a study subject; and a protocol is a standard procedure describing how a measure is collected. The criteria for selecting PhenX measures include that they be clearly defined, well established, broadly applicable, and reproducible, as well as have standard measurement protocols. The scope of PhenX is broad - ranging from cancer to nutrition and dietary supplements to physical activity and fitness, but shallow, with only 15 measures per working group. Examples of PhenX measures include height and weight (anthropometrics), blood pressure (cardiovascular), stroke (neurology), ultraviolet light exposure (environmental exposures), exposure to violence (psychosocial), liver function (gastrointestinal), and pain type and intensity (gastrointestinal). New measures are being added that are related to rare genetic conditions, but the PhenX Toolkit already contains measures used in common, complex diseases that also can be used to assess rare genetic conditions. PhenX supplements -including for sickle cell disease, mental health, tobacco regulatory science, and substance abuse and addiction - are adding depth to the toolkit.

As explained on the PhenX Toolkit home page (https://www.phenxtoolkit.org/), the PhenX Toolkit allows researchers to add measures to expand their study design beyond their primary research focus. The PhenX Toolkit provides information about each protocol, including information needed to collect a measure. Researchers are encouraged to register their studies to facilitate collaborations and provide access to information about ongoing studies in advance of publication of results.

The PhenX Toolkit is being linked to a variety of vocabulary standards, ontologies, and resources, including the NCI's Common Data Element (CDE) Browser and the National Library of Medicine's CDE Resource Portal. PhenX is collaborating with Research Electronic Data Capture™ (REDCap™) to make PhenX protocols available as REDCap™ Zip files. Pilot studies have been completed to map PhenX measures to the database of Genotypes and Phenotypes (dbGaP). There is no standard set of rules for

names of variables in dbGaP studies; therefore, manual curation is needed. The resulting ontology will be useful for dbGaP and other resources. PhenX variables can be cross-referenced to other measures, such as Logical Observation Identifiers Names and Codes (LOINC) and P3G variables.

Clinical Phenotyping from Electronic Health Records (EHRs): Opportunities and Challenges
Rachel Richesson, M.P.H., M.S., FACMI, Duke University

EHRs provide opportunities and challenges for clinical phenotyping. Opportunities include access to clinical evaluations of diagnoses and problems and clinical notes; treatments, procedures, and medications; laboratory results; and patient-reported outcomes and biometric uploads. The Office of the National Coordinator for Health Information Technology (ONC) has established several programs and infrastructure to enhance the capture and use of EHR data. "Meaningful Use" of EHR data is being implemented in stages: establishing basic EHR functionality and data structure, coordinating care and informing patients, using data to improve delivery and outcomes, and improving population health. The Centers for Medicare and Medicaid Services (CMS) is using financial incentives to encourage Meaningful Use. Between 91 and 95 percent of hospitals are demonstrating some degree of Meaningful Use. Uptake by office-based providers is lower, at approximately 54 percent, but is growing. A majority of patients in large health plans still do not use patient portals, but the potential for use exists.

The outstanding challenges for clinical phenotyping are that EHRs are designed to support clinical care, not research, and completeness and accuracy vary significantly. EHRs still largely are not standardized, more than 100 EHR vendor products exist, as well as standardized coding systems, but they are used differently, and researchers do not control EHR design or coding practices.

A clinical phenotype, in the context of EHRs, is defined as specifications for identifying patients or populations with a given characteristic or condition from EHRs using data that are routinely collected in EHRs or ancillary data sources. Phenotypes are based on widely adopted coding systems. For example, diabetes can be defined as a set of inpatient ICD-9-CM (International Classification of Diseases, Ninth Revision, Clinical Modification) codes or a combination of laboratory results, outpatient diagnosis codes, and medications. The Electronic Medical Records and Genomics (eMERGE) Network is a vanguard effort from the informatics community to combine DNA biorepositories with EHRs, changing the approach to GWAS. Resources include the PheKB (The Phenotype Knowledgebase) database, which contains 30 public phenotypes and more than 90 phenotypes that are in development. Other sources for clinical phenotypes are clinical classifications software; the CMS Chronic Conditions Warehouse; Quality Net; and research networks, such as Mini-Sentinel. Comparing groups of people using different phenotypes can be problematic, as shown in a comparison of phenotype definitions for diabetes. In another example, a multicenter study of definitions used to identify patients with chronic obstructive pulmonary disease (COPD) showed variations between patients who did and did not meet the clinical trial reference standard, including differences in rates of comorbidities, race, and education. The NIH collaboration activity and National Patient-Centered Clinical Research Network (PCORnet) are examining how clinical phenotypes are defined, and PCORnet is focusing on how rare diseases might fit in the clinical phenotype infrastructure. Comparing ICD-9-CM, ICD-10-CM (International Classification of Diseases, Tenth Revision, Clinical Modification), and Systematized Nomenclature of Medicine—Clinical Terms (SNOWMED CT), the three coding systems differ in coverage and precision of rare disease names. SNOWMED CT is considered best suited for clinical data capture because it has better content coverage, is clinically oriented, and has flexible data entry and retrieval. EHRs are one component of the various subdomains integrated in the UMLS®, which connect the subdomains by linking terminologies. There is a need for a robust clinical interface terminology, and SNOMED CT can perform that function.

Progress toward Precision Medicine and the Challenges of Integrating Genomics into EHRs

Rex Chisholm, Ph.D., Northwestern University

Critical factors for defining genetic contributions to disease, a key to the success of precision medicine, includes methods to measure genetic variation, which are available, and large numbers of well-phenotyped human genomes. Northwestern's biobank, NUgene, addresses the second factor, and includes a collection of biological specimens (blood for DNA analysis), a one-time questionnaire, retrospective and prospective longitudinal data from EHRs, and a re-contact option for additional deep phenotyping. Currently, approximately 11,000 participants are enrolled, 58 percent of whom are female, and with an ethnic distribution that reflects local census data. By providing demographic, environmental exposure, and self-reported family and medical history information, the questionnaire supplements the EHR data, which includes the free text physician notes, and electronic billing record data. The prevalence of diagnoses such as hypertension and diabetes by ICD-9 code matches known population prevalence. The magnitude of the data is exemplified by the glucose laboratory test data, which is available for 88 percent of the participants, representing more than 250,000 tests.

Northwestern University is participating in the eMERGE Network, which in Phase 1 tested the ability to leverage EHRs and biobanks for genomic research. The approach for electronic phenotyping was to identify the phenotype of interest, develop and refine a case and control algorithm, manually review data, deploy the phenotyping algorithm, and conduct and replicate genetic association tests. The type 2 diabetes case algorithm had a 98 positive predictive value. The genomic analysis identified the same genes as purpose-built cohorts. Most of the Phase 1 phenotypes developed by different eMERGE Network sites now are available. The researchers also tested whether the merged genotype data set, with samples collected for various studies (e.g., dementia, cataracts, diabetes), could be used in a different experiment. An algorithm incorporating ICD-9 codes, laboratory values, medications, and known secondary causes was used to develop a case/control algorithm for hypothyroidism. A genomic analysis discovered that *FoxE1*, which is involved in thyroid development, is associated with hypothyroidism. Such a new type of study, a phenome-wide association study (PheWAS), requires a large cohort of patients with genotype data and many diagnoses.

The goal of eMERGE phenotyping is to discover sharable, high-throughput phenotypes. Phase 2 phenotyping is ongoing, requiring natural language processing. For example, development of a phenotype for colon polyps required the analysis of the EHR for the colonoscopy and the pathology report if polyps were discovered.

Returning genomic results to the EHR requires addressing questions of how to store the data, what to store, and how to provide clinical decision support. The system architecture for returning results involves inputting data to a secure data receiver, formatting and storing results, filtering data with a knowledge engine, and producing data for the physician (e.g., best practice alerts, laboratory results) and patient (e.g., MyChart). The resources compiled for physicians and patients to explain genomic results are available on MyResults.org on the eMERGE website.

Round Table Discussion

Dr. Olivier Bodenreider asked presenters to delve further into the areas where phenotyping could be applied. He also asked about the different perspectives and tools, particularly tools used primarily for research purposes versus documentation of clinical care. How can these resources be reconciled to effectively integrate data on a larger scale to help advance the era of "Big Data"?

Dr. Rachel Richesson observed that although patterns in a lot of data can provide a surrogate to assist clinicians with diagnoses (e.g., medications as a surrogate for symptoms, patient-reported histories may indicate paths that those with rare diseases may have followed before a final diagnosis), many different

types of data are used because of the different patterns of practice. Regarding tools, the research field makes a number of assumptions about data (e.g., coding, data quality procedures) that present challenges in data integration of clinical care data. The research and clinical care environments are two different worlds that historically and sociologically have not come together; terminology remains distinct.

Dr. Rex Chisholm agreed with the difficulties but pointed out that electronic phenotyping of EHR data has worked fairly well given the intense pressure on clinicians (e.g., their time, how they can code phenotypes). He said that driving the research use of clinical data is an important step. Mapping and other tools should be put into the infrastructure to better integrate multiple terminologies and capture quality data, at the entry point, that are usable to researchers.

Dr. Carol Hamilton concurred that a sharable, portable framework would be helpful. She added that studies of the response to treatment and environmental exposure are important areas to consider.

Dr. Bodenreider asked the panelists whether the volume of clinical data would compensate for quality. The panelists agreed that big data approaches can work for clinical data, and that scale has some compensatory value: quantity can make up for some quality. Dr. Chisholm said that the database which underlies his data mining activities has more than 5 million people who account for more than 1 billion data elements; this scale definitely compensates for some of the noisy background. Dr. Hamilton commented that a wider use of PheKB would be an important step in aggregating and sharing data across multiple sites. Dr. Richesson stated that concerns still will remain, including potential biases and barriers to collection (e.g., the capture of some elements but not others).

Dr. Hamilton reflected on the challenge to get data input into the EHRs. Dr. Chisholm pointed out that because EHRs are legal documents, the type of data included and its use must be controlled. Ancillary genomic and other information desired by researchers might need to be obtained through other data sources.

Dr. Bodenreider mentioned that the Observational Health Data Sciences and Informatics (OHDSI) Consortium has a distributed system of records covering 250 million patients, with a goal of reaching 1 billion patients. The effort has been successful because OHDSI requires mapping of the data in the original systems to a few standards (e.g., SNOMED).

Participants asked about the possibility of mandated collection, the involvement of data experts more globally in the collection process, and the boundaries between models and terms. Dr. Chisholm agreed that someone (e.g., insurance provider) is paying for the collection of phenotyping data and added that Meaningful Use is driving the field into the direction of mandated collection. He noted an age-effect among clinicians; younger physicians are keener on coded data, whereas older physicians are more willing to tell the story in notes. Dr. Richesson observed that her group has spent a significant time mapping to specific systems, and manual handling of data can be costly. The solution is to bring in data experts and have a more guided approach, including identifying the structure of data elements. Dr. Hamilton added that challenges arise in terms of work flows, but building in a bioinformatics component through such tools as natural language processing or HPO mapping should help.

Dr. Bodenreider asked about the best way to incorporate a bioinformatics component. He also noted that data often must be curated after collection before being useful for research. The panelists agreed that data collection for research is complicated, and good online tools are difficult to find. Dr. Chisholm suggested that the first step is to use natural language processing to distill instances of particular terms; it would be interesting to use this approach to search for 25,000 terms in GWAS space and compare the mapped results with HPO, which has 25,000 terms mapped.

Dr. Richesson stated that Meaningful Use has provided some levels of incentives regarding online capture. As improvements in collecting patient information are realized, gaps in knowledge about patients who are high users or are not functioning well could be identified, with therapies enhanced for those patients. Dr. Chisholm pointed out that models could inform on how researchers think about genotypes and phenotypes; the clinical flow to model space is an area for future studies.

Session 2: Cross-Species Phenotype Analysis and Ontology
Chair: *Melissa Haendel, Ph.D., Oregon Health and Science University*

Chair Overview

The ultimate aim of cross-species phenotype analysis and ontology is to create better models of human disease and use model data to inform patient care and disease discovery. The standard ways to associate models to disease involve approaches based on direct assertion, homology, semantic similarity, and enrichment or statistical association. Clinicians and basic researchers, however, do not speak the same language. Unfortunately, a standardized vocabulary is not sufficient to cross the translational divide. Integrating disease sources through mapping also is inadequate as a single approach to this complex problem. The goal is to facilitate collaboration and discovery across diverse groups, using phenotype ontologies as a bridge between them. This session described models and technologies used to connect diverse groups of scientists.

Crossing the Species Divide
Chris Mungall, Ph.D., Lawrence Berkeley National Laboratory

Dr. Mungall described the Monarch Initiative's ontology project, which addresses analogous problems in animal systems and aims to build better models. Animal models have experienced a large growth of phenomic data. With the increase of high-throughput systems, more animal and human phenotype data are expected from less classic models. A knowledge base of animal phenotypes is useful to inform human phenome research, which in turn can help to build better animal models. Animals and humans have some shared genes and biology: two analogous genes will share some similar function at the molecular level. Model mutants recapitulate disease phenotypes. This is evidenced by a gene that is capitulated in human disease, when, if knocked out in the mouse, will show similar phenotypes that can inform the human disease. This similarity can be seen in cell, neuroanatomy, and behavior phenotypes. Researchers need a way to quantify what models make a good match.

Advances in sequencing provide a model for precision comparative phenomics. Sequencing data increased because of next-generation sequencing technologies. Phenotype data will increase as researchers move from knockout models to CRISPR technologies. Whereas sequencing has enabled the bioinformatics revolution, phenomics has the potential to revolutionize biomedical research (e.g., phenome databases). Challenges for phenomics are that quantifying distance is complex, and a single unifying model is elusive.

Dr. Chris Mungall showed a graph tree used by Monarch Ontology to illustrate how phenotypes can be classified based on anatomy, mechanism, or function, with some phenotypes having "dual parents." Likewise, genes can be classified to describe genomic entities (e.g., Pten Atxn1). The graph tree can measure some similarity of entities; terms may be similar but not completely match at the subcellular or atomic levels, or two genes can be seen to function in a similar way at specific or broader levels. Other approaches include the comparison of multiple phenotypes at once, and probabilistic methods to explicitly model unobserved data as a way to measure uncertainty.

Multiple phenotype ontologies and terminologies exist among species, and researchers are challenged with how to map between these ontologies. To address this, the Monarch Initiative used "helping machines" to understand phenotype terms; decomposed complex concepts into simpler, underlying concepts; and moved into species-neutral areas and homologous components to allow mapping between mouse and human models. The researchers found anatomical homology between structures in the ontology and developed network approaches for deep phenotype matching of cellular ontologies with complex phenotypes. Monarch Ontology integrates and links these ontologies into a graph showing phenotypes and genotypes (e.g., phenotype of abnormality of cerebellum), and is delving into other approaches for matching non-obvious phenotypes to determine whether enrichment is possible for orthologs to infer non-obvious connections.

Next steps include joint genomic-phenome comparison for use in the clinic; probabilistic modeling, with a focus on evolutionary models of phenotypic change and the incorporation of negation uncertainty; and the use of ontologies for causal models, including genotype to phenotype associations as well as toward the development of temporal and causal models of phenotype progression. For the next generation of phenomic curation, researchers need a pan-species approach, including a comparative phenomics perspective, enabled collaboration, and easy-to-integrate new organisms. In addition, support for expert curators, new tools to enhance efficiency, and widely used, open standards would help to increase knowledge curation. Finally, an integrated phenotype curation that encompasses ontology, genotype-to-phenotype curation, and causal networks is needed.

Disease Variant Prioritization and Model Discovery through Cross-Species Phenotype Analysis
Damian Smedley, Ph.D., Wellcome Trust Sanger Institute

Dr. Damian Smedley presented a group that includes members of the NIH-funded MONARCH initiative and KOMP2 (Knockout Mouse Phenotyping Program) projects that have applied semantic similarity methods to match phenotypes within and across species spread in the field of rare disease diagnostics and gene discovery. Standard rare disease exome analysis pipelines work to identify the rarest variants; they eliminate hundreds of candidates but rarely obtain a single, striking candidate. An alternative approach called Exomiser takes advantage of the multi-species phenotype comparisons made possible through Monarch. For every gene in the exome, existing phenotype knowledge is compared with the patient's HPO phenotype profile and provides a score for how similar its known phenotypes are to the patient. The end result is a single variant candidate that is rare, predicted to be highly pathogenic, and affecting a gene where previous disruption has been shown to cause a similar phenotypic effect to that seen in the patient. Exomiser is a software suite that includes PHIVE (Phenotypic Interpretation of Variants in Exomes), hiPHIVE, PhenIX, and ExomeWalker tools. Dr. Smedley described studies using the PHIVE tool.

Dr. Smedley's group tested Exomiser extensively by determining how well it could be used to identify known disease variants from the Human Gene Mutation Database; variants were randomly added to unaffected exomes produced by the 1000 Genomes Project. Human disease phenotype data are critical for detecting known disease-gene associations, and the Exomiser achieved almost 100 percent performance for known associations. Other exome analysis tools that use HPO annotations of the patient also have been developed, including PhenIX, Phen-Gen, eXtasy, and Phevor; each of these tools bring various strengths to phenotype analyses, and the Exomiser was found to perform comparatively with them.

Monarch researchers also tested Exomiser on the NIH Undiagnosed Diseases Program (UDP), running 11 cases that had been solved through Exomiser, which also found the same diagnosed variant. Exomiser is now being used in the UDP pipeline, and it reported a novel disease-gene discovery recently (York Platelet syndrome and *STIM1*). Exomiser is being expanded to do whole genome analysis, through a process of filtering by phenotype similarity, regulatory features, and frequency, followed by prioritization

by pathogenicity and phenotype similarity. A similar strategy is being applied to 414 Mendelian regulatory mutations, and preliminary results suggest 80 percent performance if all phenotypes are used.

Monarch phenotype comparison methods also have been used for animal model discovery in the context of the KOMP2 project at the International Mouse Phenotype Consortium (IMPC) portal. There are 530 genes associated with a Mendelian disease that now have a phenotyped IMPC line. Of these, 85 percent have never had a mouse mutant generated or one that was described at Mouse Genome Informatics (MGI) as a disease model, 24 show phenotype similarity from partial results on the IMPC broad screen, and opportunities exist regarding 75 novel disease gene candidates from phenotypic similarity where the human ortholog lies in correct linkage locus. Dr. Smedley described an example of this approach in the first Bernard-Soulier mouse model in a study of bone mineral density, which identified a novel candidate for isolated microphthalmia, with cataract, 1.

In conclusion, this work demonstrated that semantic phenotype comparisons greatly improve diagnosis and candidate gene identification, and highlighted good disease models; mouse and fish phenotypes should be included to advance novel disease gene discovery; and the collection of deep clinical phenotype data has value for transformational bioinformatics. Challenges include the inclusion of phenotype frequency data and negative phenotype data. In addition, some phenotypes (e.g., behavior) are not well covered by mouse or fish models, and focus is needed on common diseases.

Exploiting Mouse Genotype-Phenotypic Associations for Disease Genomics
Caleb Webber, Ph.D., Oxford University

Dr. Caleb Webber described how mouse knock-out genotype and phenotype associations can be used to advance the study of disease genomics. Patients whose variants disrupt the same “pathway” share a broad range of phenotypic similarities. Knowing why dispersed loci are involved in similar phenotypes is useful because they can provide multiple drug targets and help with the tracking of a larger population group. However, difficulties arise when comparing patient phenotypes as there can be absence of sufficient details and description can present issues unless there is an underlying ontology to relate phenotypes. A study of 4,000 patients who have been systematically phenotyped and annotated using HPO allowed researchers to examine copy number variants for patients and group them non-exclusively. Patients had many of the same phenotypes, and standard functional enrichment tests were used to ascertain molecular commonalities. The study considered whether there was a subset of patients whose genes contributed to a particular pathway phenotype that was similar. The study found that the same pathway equaled a similar phenotype.

Dr. Webber’s group found that mouse phenotypic data can be used to evaluate other functional data, especially for particular phenotypes of interest. The value of mouse phenotype data compared to gene ontology annotations is in the ability to predict whether two genes are involved in the same disorder. Researchers integrated function-linkage networks to identify human disease genes and to predict human phenotypic associations by analyzing the similarity and dissimilarity of human disease phenotypes associated with genes as well as the functionality between genes. Mouse data can now be used to measure other types of functional genomics data. A comparison of functional data sources also shows the value of a phenotypic linkage network. In disorder-specific networks, a disease-relevant phenotype can be selected to see how those relate to functional genetics and ultimately to a disorder, such as Type 2 diabetes; this can be seen in a study that clustered exome variants from nearly 13,000 ethnic patients.

Phenotyping genes that are less studied will help researchers to estimate variant deleteriousness. Dr. Webber described study bias in haplo-insufficiency prediction as an example. Using a method built on unbiased genomic information, results improved when a mouse model was used in addition to human data.

Round Table Discussion

Dr. Haendel noted that in the clinic, researchers complete a full clinical work up, particularly for rare diseases, and every system is at least partially phenotyped. This is not the case in animal modeling, where many researchers are experts in one system (e.g., skeleton, neuron development). She asked the panelists to consider ways to help model organism researchers populate those kinds of phenotype gaps. Dr. Webber responded that from an informatics perspective, the most helpful thing is knowing what is most informative to associate a gene with a disease. Dr. Mungall noted the amount of bias that is often present, and Dr. Smedley said that the greatest issues surround tools and training in the use of tools.

A participant asked about the use of analyses from the methods described by the panelists to predict gaps in knowledge, inform the creation of better models, and prioritize them. Dr. Haendel replied that existing data should be aggregated and used to inform, look for knowledge gaps, and help prioritize research in both human and animal models. Dr. Mungall referred to his group's ongoing work on the annotation sufficiency score to help determine where more phenotyping is needed. Dr. Webb encouraged realism regarding what model organism (e.g., a knockout mouse) can deliver and its relevance to human conditions. He added that deep comparison may not be possible.

A participant referred to Russ Allman's ranking of the PhenIX paper as a major step forward in genomic to phenotype associations, but wondered how model organism phenotype data could be made more relevant as most clinicians do not consider it. Dr. Smedley agreed that clinicians addressing complex diseases mostly look at data for their specific research fields. Dr. Robinson provided an example of how his group used Exomiser to conduct exome sequencing and has supported initial gene discoveries in which clinicians, in hindsight, acknowledged similar findings in model organism studies.

Dr. Chessler reflected on her experience with knockout mice. Some databases (e.g., phenome database, genenetwork.org) contain more than 1,500 phenotypes from the mice, and every phenotype that is added can be correlated with every transcript abundance. The fully saturated data provides quantitative associations of multiple genes, phenotypes, and variants.

Another participant remarked on the types of polymorphisms seen in human versus mouse, noting that GWAS SNPs would not be useful to see in knockout mice; some literature shows that the loss of function and percent of variants in humans is important for novel drug discovery. Dr. Webb agreed; loss of function data are informative but not for specific phenotypes.

A participant raised the question of how many phenotypes are missing in the phenotype "universe" and how many of those missed are due to lethality. Other participants responded that for approximately 20 percent of coding genes in humans and up to 30 percent in mice, no phenotype information is available.

A participant asked Dr. Webb how, using his methodology, he would inform which combinations to try for more complex genotypes (e.g., multi-gene knockouts). Dr. Webb recognized the idea that emergent effects from a combination of genes is the "elephant in the room" in human genetics. He recently studied copy number variants from autism using the network approach and hit 4 genes out of 30 that were thought to be involved in the phenotype. He knocked out those genes alone or in combination and found that only in combination was a phenotype seen. Asking the mouse community to knock out combinations may be too much "to ask" at this time. A participant noted that antagonism indeed exists, although many researchers are focused on synergies. Dr. Haendel commented that much of the available data have not been curated well and should be re-mined.

Session 3: Large Scale High Throughput Analysis of Disease Model Phenotyping Data and Annotation of Gene Function
Chair: *Janan Eppig, Ph.D., The Jackson Laboratory*

Chair Overview

Dr. Janan Eppig highlighted several important projects involved in disease model phenotyping. She noted that there is a need for generating a comprehensive system for mutations. To conduct rigorous phenotyping, researchers will need shared standard operating procedures, standardized phenotype pipelines, and integration of data within large-scale efforts.

The International Gene Trap Consortium has used different trapping strategies to create more than 300,000 gene traps. Scientists continue to obtain mutants of interest from these resources.

N-Ethyl-N-Nitrosourea (ENU) mutagenesis (forward genetics) involves phenotype-driven screens based on specific systems. It has produced more than 3,400 important mutations for specific phenotypes. All mutants are incorporated into MGI. ENU mutagenesis is in a revival, aided by sequence-based rapid mapping of new mutations. Sequencing also identifies "incidental mutations" that are important for modifier identification and the discovery of new point mutations in other genes.

The Knockout Mouse Project (KOMP) provides a systematic approach to knocking-out all protein coding genes in the mouse. The overarching goal of the project was that mutations would be made in embryonic stem cell lines to knockout genes systematically. It is a multi-national effort, with all stakeholders working on a single genetic background. KOMP2, which is the U.S.-funded portion of the project, is a phenotype pilot phase that aimed to phenotype 5,000 mouse lines. It is now in a completion stage.

Other important data are found in individual laboratory research fields focused on specific systems and with a variety of mutation types. Because large-scale projects (mammalian scale) are limited in scope, focused laboratories will need to provide granular deep phenotyping. The integration of these data with large-scale efforts will maximize knowledge.

Several lessons from these first large-scale phenotype/mutant screens in mice include that data should be organized with a long-term plan; community standards for nomenclature, strains, and identification of objects screened should be adopted; and phenotype data should be annotated or assigned using standard terms and descriptive metadata.

Functional Exploration of Human Cancer Genomes Using Flies
Erdem Bangi, Ph.D., Mount Sinai Hospital

Dr. Erdem Bangi described tumor phenotype studies of human colorectal cancer using *Drosophila* as the animal model. The Cancer Genome Atlas (TCGA) has identified 30 drivers of colorectal cancer with recurrent mutations found in five pathways (Wnt, Ras/MAPK, PI3K, TGF beta, and TP53). Researchers selected pathways in flies to represent these recurrent pathways and created 33 multigenic *Drosophila* models to illustrate the complexity of the human tumors. Tumor phenotypes that were observed included proliferation, multilayering, evasion of apoptosis and senescence, migration, and others. Having a diverse number of models allowed the correlation of tumor genotypes with cancer phenotypes. Many phenotypes are emerging properties and require complex interactions between individual mutations.

The researchers also tested drug response for 16 agents using genetically complex models that considered dissemination to distant sites as a readout for response. In two models, 12 of 16 drugs were effective against Ras alone, but none of the 16 were effective against the other pathways, indicating that intrinsic

drug resistance is an emergent property of genetically complex models. A study of PI3K pathway inhibitors found molecular biomarkers of resistance and response, and resistance mechanisms were investigated leading to designing a drug combination that overcame the resistance. This approach was validated in mammalian models.

The studies showed that complexity matters when comes to modeling cancer. A large number of models are needed when conducting genotype-phenotype studies. Questions for the future include: How much complexity is needed for accurate drug response? Can fly models be used for personalized drug discovery tools? Can they be used to help patients?

Dr. Bangi's group will be shifting to next-generation studies that focus on specific genes rather than pathways, as well as including patient-specific variants. The Center for Personalized Cancer Therapeutics pipeline will begin with the generation of high-quality tumor genomic profiles, build patient-specific fly models (both base and personalized), and conduct drug screenings using FDA's cancer set (62 drugs) and full FDA set (1,200 drugs), with the aim for a multidisciplinary tumor board to provide personalized treatment recommendations. He noted the advantages of using flies as a model organism, including sophisticated genetics; conserved epithelia, pathways, and drug activity; speed, scale, and low cost; and facilitation of possible *in vivo* drug screens.

The Zebrafish Mutation Project
Derek Stemple, Ph.D., Welcome Trust Sanger Institute

Dr. Derek Stemple described an animal model project focused on creating a knockout allele in every protein-coding gene in the zebrafish genome using exome and next-generation technologies. He shared several examples of the Zebrafish Mutation Project's activities related to morphological and molecular phenotyping.

Functional annotation of a vertebrate genome (zebrafish) involved the identification of a disruptive mutation in every protein-coding gene, morphological phenotype description of 8,000 genes, and mRNA expression profiles of alleles producing abnormal phenotypes. Steps in mutation detection in genetics are sequencing frozen sperm (whole exome) and analyzing all mutations. The steps are reversed for phenomics. When researchers identified one mutation, they were able to find it in half of the population; however, they found 15 disruptive mutations when they looked at all the mutations at once. Because the majority of mutations did not lead to any visible phenotype, the research collected and phenotyped wild-type fish. Dr. Stemple shared an example of this conducted with *lamc1* and referred attendees to the European Nucleic Acid Archive (ENA), where the Project has placed all of its detected mutations. The Project has sequenced 3,500 F1 males and identified more than 32,000 nonsense and disruptive splice-site alleles in more than 14,000 genes.

The Project also is using Cas9/CRISPR technologies to generate highly multiplexed targeted mutations. An example is the application of the differential expression transcript counting technique (DeTCT) to identify genes displaying alterations in transcript levels and aid researchers in better determining drug response. A total of 39 experiments have been completed, with 78 sets of samples collected. Dr. Stemple showed examples of work related to transcriptional profiles, including expression enrichment, of *slc2a11b* and *sox10*; heterozygous mutations of *sox10* in humans cause Waardenburg syndrome through haplo-insufficiency. He observed that CRISPR is an ideal technology to analyze complex diseases.

Finding Cross-Species Phenomic Similarity through Integration of Heterogeneous Functional Genomic Data
Elissa Chessler, Ph.D., The Jackson Laboratory

Dr. Elissa Chessler described a successful approach to address the diversity in models for complex disease. Data-driven classification of traits and models are based on the underlying biology. Objective phenotypes can assist in a better alignment of disease and models. Phenotype ontologies have to balance competing priorities and enable harmonization, including among various approaches and resources (e.g., mammalian phenotype, vertebrate trait ontology, neuro behavioral ontology, and animal behavior ontology).

Many mouse genetic strategies associate genes to traits and phenotype terms, including mutant characterization, genetic loci, and differential expression. Systems genetic analysis holistically connects traits to sets of genes and variants, and online data resources such as GeneWeaver, which can facilitate the identification of extremes from advanced mouse populations as disease models, enable cross-species and cross-population integration in a single database.

Dr. Chessler described research questions that were enabled by GeneWeaver for integrative functional genomics, including the identification of a new mouse model for alcohol preference. Promising new models were identified by characterizing the "ignorome." Using a model from the KOMP repository, researchers completed an aggregate analysis of many studies of alcohol preference using genes commonly associated with alcohol and found five intersections, the highest ones of which were not annotated for alcohol and many were not formally associated. She showed her laboratory's approach to find models for related facets of alcohol use disorder and provided convergent evidence (genetic and Encyclopedia of DNA Elements, ENCODE data) that yielded a single causal variant and can abet the design of precision mouse models. A latent ontology from empirical genomic evidence also was constructed.

Dr. Chessler highlighted the main points of her presentation. Linking animal models to human disease through phenotypes often exploits face validity. The desired characteristic is "construct validity." Underlying construct similarity can be obtained through genome-wide comparison of assays and models. A wealth of data sources from mouse and other organisms exist. Cross-species integrative functional genomics enables global comparison of animal models, assays, and diseases based on the underlying biology.

Round Table Discussion

Dr. Eppig asked Dr. Bangi how he chooses the target gene/s and knows he has enough candidates when developing models. Dr. Bangi responded that his project's aim is to bring as many cancer drivers into the model as possible, but what comprises an adequate number is unknown. He considers potential deleteriousness, gene function, and other factors; ranks the genes; and selects the top 10. He also described work with a separate pipeline in which each variant is tested against the base model before a selection is made, thus having a functional test to determine ahead of time whether a patient variant was a driver or passenger. The goal of the study is to determine the differences of drug effects in recurrent lines for patient-specific models; the current phase is focused on base drug screening.

A participant asked about the use of imaging in the panelists' studies. Dr. Chessler's work usually begins with phenotypic assays, but others use broader monitoring techniques to extract from image and video analyses frequently occurring features or events by passively monitoring the activity of a model organism. Dr. Bangi's group starts with detailed phenotype characterization for a small number of models, and works to identify surrogate readouts for high-throughput screening. Secondary assays that are more detailed are conducted to compensate for loss of resolution.

Participants lauded the studies described as a good representation of how reduced models can have an important impact, and asked about the models' limits for studying mental health conditions with a genetic basis such as schizophrenia or autism. Dr. Chessler noted that if the underlying biology networks

associated with the disorders can be found, then organism characteristics that reflect the activity of the relative pathway can be identified; detection of biological pathways in the early stages of disease is important. Dr. Stemple described a study of depression among women that found two loci of interest in a first analysis, and then found a third signal after asking about environmental conditions that revealed the women had experienced childhood sexual abuse; it raises the question of what is the phenotype that should be studied. Movement between biological components (e.g., synaptic functions, mitochondrial functions) in a meaningful way may yield more associations with behaviors. Dr. Bangi noted the advantages in breaking a complex disease into parts to study in more relevant, smaller, or more cost-effective organism models. Participants noted the challenges in working in a multi-directional manner.

A participant noted a point raised in Dr. Bangi's presentation that "It is really important to do representative complexity" and wondered how much complexity is needed. The same slide noted the need for "lots of cell lines," and the participant requested clarity on how much is "lots." Dr. Bangi clarified that his team is working on as many personalized models as possible (up to 200) based on TCGA data and then conducting drug screening. He is using transgenic flies and is not focused on cell lines. He noted that as more genomes are sequenced, the landscape of the mutated genes changes.

Participants reflected on the move in genetics from genes to specific areas and asked about pathogenic prediction for a disease such as congenital heart disease. Dr. Semple said that he would first want to understand the loss of function in the zebrafish.

A participant stated that in the toxicology world, where there often are well-defined animal model organism phenotypes, the question is how to query or even identify if there is a human phenotype. Dr. Chessler said that her group has integrated the Comparative Toxicogenomics Database that compiles drug-interacting genes; the dataset can be probed for traits in any organism that are associated with drug-interacting genes. The phenotypes could suggest which diseases might be relevant.

A participant noted that each of the panelists represent a different species and wondered if there was a logical, iterative process to move from one species to another to inform human disease. Dr. Chessler responded that it depends on the disease, the phenotype(s), and which species are most amenable to the specific phenotype(s). Dr. Bangi said that for cancer, it is best to try as many models as possible based on the resources available.

Session 4: Linking Disease-Relevant Phenotypes with Physiologically Relevant Molecular Pathways and Networks

Chair: *Olga Troyanskaya, Ph.D., Princeton University & Simons Center for Data Analysis*

Chair Overview

Biological scientists want to map phenotypes across organisms to understand the molecular bases that underlay those phenotypes, determine the best model organisms to use for specific diseases and conditions, and understand how these phenotypes relate to each other within organisms. Another area of importance relates to the phenotypes that look semantically interesting but have no annotation overlap. This session focuses on three questions: Can gene and phenotype-gene associations be improved within human models? What ways can relationships between phenotypes (and phenotype networks) within an organism be studied? How can phenotypes be mapped across organisms?

Using Networks to Re-Examine the Genome-Phenome Connection
John Quackenbush, Ph.D., Dana-Farber Cancer Institute

Dr. John Quackenbush described work funded by the National Heart, Lung, and Blood Institute (NHLBI) to examine the connection between genome and phenome through network structures. As GWAS has not been fully successful in finding a genetic variance that influences complex traits, researchers have employed a technique called eQTL analysis, which examines *trans*-acting SNPs instead of the more commonly focused *cis*-acting SNPs. eQTL networks are based on the idea that eQTLs should group into communities with core SNPs regulating particular cellular functions. Researchers noted that many strong eQTLs are found near the target gene but wondered about multiple SNPs that are correlated with multiple genes. They studied degrees of SNP and gene distribution in COPD, and found almost no SNPs in the "hub" (target gene area); the hubs are a GWAS desert. Network structure matters because the collection of highest degree SNPs is devoid of disease-related SNPs. Highly deleterious SNPs that affect many processes likely are removed by strong negative selection.

Researchers focused on using the network to identify groups of SNPs and genes that have functional roles in the cell by clustering the nodes into communities. Dr. Quackenbush used 31 communities in COPD eQTL networks to demonstrate that community structure algorithms group nodes in such way that the number of links within a community are higher than expected by chance. His group found that disease SNPs were skewered higher on a SNP list when ranked by a community core score, which calculates local connectivity. The median core score for GWAS SNPs was 1.7 times higher than the median for the non-GWAS SNPs.

This research showed that the property of the hubs to be devoid of GWAS hits and is consistent with strong selection against highly deleterious SNPs/survival bias. The study of communities indicate that a family of SNPs are associated with regulation of a process consistent with complex traits. Many communities are apparently preserved across disease states, reflecting processes common to many cell types. The Core SNPs are highly enriched for disease associations.

Human Phenotype Networks
Jason Moore, Ph.D., M.S., The Perelman School of Medicine, University of Pennsylvania

Common diseases have complex systems that affect individual trajectories. These complexities have been seen even at the level of a single gene. Because the univariate approach (e.g., one SNP at a time) has had limited success, a multivariate-bioinformatics approach is needed to tease apart factors and understand complexity. Tools and other resources are needed to handle this complexity as many current tools look at one factor at a time under the assumption that an individual SNP or factor has an impact separate from other SNPs and factors.

Dr. Jason Moore described studies of bladder cancer responses to benzene pyrene using an epistasis network approach. Epistasis concerns a phenomenon that consists of the effect of one gene being dependent on the presence of one or more "modifier genes." His group considered epistasis and pleiotropy, which occurs when one gene influences two or more seemingly unrelated phenotypic traits. Their research yielded "shadows" of complexity from which they postulated new gene-gene interactions. They extended this network approach to phenotypes based on risk factors and used SNPs to build similarity among phenotypes, which was shown in clusters. They used the Reactome Pathway database and found that focusing on shared pathways instead of SNPs to map genes and pathways showed a different clustering of human conditions; for example, on the map that used shared pathways, coronary heart disease clustered in a different site on the map than where it clustered when SNPs were used. A different location also was seen when examining exposure-based human phenotype networks; environmental exposure provides different information than SNPs. Dr. Moore further illustrated this with another example: genome-wide genetic interaction analysis of glaucoma using expert knowledge derived from human phenotype networks. The challenge going forward is how to harness this information to improve genetic analyses.

Understanding the Molecular Basis of Human Disease by Mapping across Tissues and Organisms
Olga Troyanskaya, Ph.D., Princeton University & Simons Center for Data Analysis

Dr. Olga Troyanskaya presented ways to use big data in biology to map animal models to human disease, starting with tissue complexity. More than 200 cell types are present in the human body, encompassing a wide variety of tissues and organ systems. Tissue specificity is critical in human disease: each cell type performs a specialized function, and pathways and processes need to be understood in a cell-and tissue-specific context. Although there is a significant amount of high-throughput genomic data available, including gene expression, the data are not resolved to cell types and tissues. In addition, many datasets are not annotated to cell-type or tissues of origin, are only partly annotated, or are not annotated adequately for phenotyping.

To address this issue, Dr. Troyanskaya's group built integrated networks for 144 human tissues and cell types. Their studies on inflammation in blood vessels showed that tissue networks can predict disease-relevant, lineage-specific molecular responses.

A tool called NetWAS, which focuses on functional genomics, has been developed to help reprioritize GWAS results to identify disease genes and potential drug targets. For example, a case study involving the Women's Genome Health Study successfully applied the NetWAS approach to reprioritize GWAS results by considering phenotypic, functional, and therapeutic values for several endpoints such as hypertension. NetWAS is discovery driven, retains the unbiased nature of GWAS, does not depend on known disease-gene associations, and can be used to re-analyze GWAS in which no associations reached genome-wide significance.

Most human diseases are under-characterized at the molecular level, and model systems can help improve the understanding of genes and biological mechanisms as well as facilitate the conduct of some genetic experiments. Mapping diseases and phenotypes is challenging, and researchers are faced with the question of which model system to use. One approach is to link human disease to model phenotypes at the molecular level. It relies on a rich curated resource of genes annotated to biological processes in Gene Ontology that is available for human and all major models. Examples of this approach included studies of macular degeneration in zebrafish and humans, and of candidate genes for Parkinson's disease in *Caenorhabditis elegans* and human GWAS.

Round Table Discussion

A participant noted that GWAS was not intended to diagnose individual patients and asked the panelists whether any of the methods that they described could be used to diagnose a patient as opposed to the population setting. Dr. Troyanskaya stated that genes identified through GWAS could be used as a diagnostic tool applied to individuals. The research that Dr. Quackenbush presented is not focused on individual patients, but in other research he used the Predict Gene Regulatory Network (PANDA), which centers on prediction of downstream factors, and found that omission of even a single sample resulted in a slightly different network; he found druggable targets through analysis and network estimation. Dr. Moore referred to an editorial that he wrote in 2009 on "genome type" that wondered if the human health state is based on the entire genome, a subset of the genome, or an SNP, and where various common disease genes can be mapped on the genome; it is the individual genome-type that confirms the clinical phenotype.

A participant asked about using tissue-specific networks to make sense of gene expression data of individual tumor samples and used the example of tumor and normal colon tissue. Dr. Troyanskaya recognized the difficulties of moving from prediction via networks to single samples. For tissue-specific

networks, her group has a method to subset the networks (i.e., make the networks specific to a given sample) and works with networks for most biological processes; she can reconstruct networks by regressing samples (i.e., taking out what is most relevant).

A participant asked Dr. Quackenbush how he deals with tiny effect sizes in his process model (the "500 SNPs for 20%" problem). Dr. Quackenbush said that is how it is, and that completely predictive biology probably cannot be built from genetics: gene, environment, and chance all play a part. He added that one might say that "these SNPs perturb this function" as a reason for sensitivity. He shared a thought from Eric Sondheimer that small effects will always exist. He added that one approach that he has taken is to reprioritize genes, select the top 1,000 genes, and retest them in GWAS. Weak effects are a fact of nature that researchers have to accept.

Another participant asked if enough whole genome sequences for people with particular conditions existed such that researchers could say yes or no that there is a reasonable probability that what is being seen is an additive effect of multiple low-effect-size genes. Dr. Quackenbush dissected the question into (1) Are the data available? and (2) Are analyses done? Environmental data are not included in most datasets. He also commented that analysis has barely scratched the surface in this area and stressed that in focusing on analysis, context is important to anchor genetic information, particularly when phenotypic data do not exist for studies.

A participant asked Dr. Moore if he thought there might be a "tipping point" where small additive effects become sufficiently distilled or if it will be a point of reaching thousands of polymorphisms to elucidate the heritability of diabetes. Dr. Moore confirmed the latter: it will be thousands, and some will be small independent effects. Dr. Troyanskaya added that personalized medicine may not be fully achieved in our lifetime. Dr. Quackenbush affirmed the importance of understanding the continuum of healthy to disease conditions; for example, a liver cell should be considered as a continuum and not just as a single cell.

Day 2: Friday, September 11, 2015

Session 5: Clinical and Experimental Biology Data Integration Emerging Field of Precision Medicine

Chair: *Yves Lussier, M.D., FAMCI, University of Arizona*

Chair Overview

Dr. Lussier introduced the subject of the session and provided examples of his group's work in moving assays into clinical practice thus illustrating the integration of data to support precision medicine. Using an N-of-1 approach, researchers examined dynamic changes *in vivo* in the transcriptome of patients whose conditions were hypothesized to be exacerbated with asthma. In another assay, his group used one patient sample to analyze the approaches of four papers on the pathways of Mahalanobis distance to better determine which could predict the outcome of patients, specifically hospitalization within the coming year.

Precision Medicine and the Reclassification of Cancer: Divide and Conquer

Razelle Kurzrock, M.D., University of California, San Diego

Dr. Razelle Kurzrock shared her opinion on the terms precision medicine and personalized medicine, which are often used interchangeably. Both are correct: precise alterations in an individual's cancer or disease and how to target it is being learned. To be *precise*, the treatment must be attenuated or *personalized* to the individual. Lessons already garnered from precision medicine activities are to use combinations of matched drugs for metastatic or complex tumors, treat newly diagnosed patients, forcing

precision medicine technologies to retrofit into traditional paradigms is suboptimal, harness the immune system, and develop new models for the clinic to support transformative changes.

Cancers are difficult to treat; they are complex diseases and traditional therapies often have low survival gains. A master protocol called Profile-Related Evidence Determining Individualized Cancer Therapy (PREDICT) is a histology-independent targeted approach that can assess multiple molecular aberrations, and match patients with targeted agents to increase response rates. Dr. Razelle Kuzrock's group has partnered with the University of California, San Diego (UCSD) Super Computer Center in precision medicine studies. Research focused on whether every patient with metastatic disease is different, and molecular analysis of 75 patients confirmed that although some have an element in common, no two have the same genomic portfolio. Hence, drugs should be customized for the individual patient.

Beyond genomics, much more is to be learned from studying the transcriptome. Strategies in the clinic generally involve the treatment of newly diagnosed disease and customized combinations and immunotherapy for advanced disease. The example of chronic myelogenous leukemia (CML), in which the median survival changed from 4 to 20 years, shows how a fatal disease can be transformed. End-stage CML does not have a good response rate, but outcomes changed dramatically when matched targeted treatment was moved to early diagnosed disease. The same success story is possible for solid tumors. Key factors in cancer treatment success are knowing the target, using a targeted agent, and treating early diagnosed disease.

Success also is being seen in tumor micro-heterogeneity, which can harness the immune system. Immunology is revolutionizing melanoma cancer treatment, and markers for immunotherapy are emerging. These advances have been helped by new technologies and approaches, such as the Liquid Biopsy Program, which extracts DNA from serum (blood, urine, or ascites). For example, researchers can detect resistant mutations related to lung cancer months before progression is visible on a CAT scan. In other studies, 37 percent of brain tumors were found to be shed into the blood, and EGFR amplifications in ascites were detected in lung cancer. Liquid biopsy can facilitate customized combination therapies and support early diagnosis of disease.

Lessons from a meta-analyses of 70,000 patients showed that the personalization of therapy was one of the most important variables that correlated with improved response rate. Non-personalized targeted arms led to poorer outcomes than cytotoxic arms (e.g., chemotherapy).

Linking Disease Model and Human Phenotypes: The Clinical Geneticist Perspective
Gail Herman, M.D., Ph.D., Nationwide Children's Hospital

Precision medicine has been made possible through the disruptive technology of next-generation sequencing and advances in computation biology. Clinical utility currently encompasses cancer and the diagnosis of rare Mendelian disorders. Its future use will be in pharmacogenomics and multifactorial disorders. The past few years have witnessed a small explosion of gene identification through whole exome or whole genome sequencing. Strategies for exome sequencing in pediatrics disorders include: (1) original diagnoses techniques; (2) trios, which compare variants of the single patient affected with his/her/its parents and elucidate *de novo* inheritance; and (3) recessive analysis.

A case study of the Central Ohio Registry for Autism (CORA) enrolled 300 families and focused on exome sequencing of families with a single affected patient without other significant neuro/psych disease. Work has been completed on 75 families with simplex, 11 with multiplex, and 16 with *PTEN* tumor suppressor. Strict criteria have been used in the simplex families to narrow variants to allow data to be shown on a simple spreadsheet. Results have included 65 patients confirmed *de novo*, of which 30 were damaging, and 5 were considered clinically significant.

Dr. Gail Herman's group followed the recommendations of the American College of Medical Genetics and Genomics (ACMG) and Presidential Commission on Bioethics. These encompassed a minimum list of 56 actionable genes and specific mutations. Specific recommendations included that pathogenic variants in the list should be reported regardless of indication for clinical exome sequencing; laboratories should report only the variants listed; and the clinician should provide appropriate counseling. In addition, the list should be refined and updated annually.

Clinical exome sequencing has a high diagnostic yield and is important in studying trios. Dr. Herman referred attendees to the Clinical Exome Report for useful information for phenotype-genotype research. Although full exome sequencing is valuable, it is not always the best option for cases in which an obvious phenotype exists or cost is a consideration. Genetic diagnosis is helpful in that it can prevent additional unnecessary testing, could predict future medical complications, and provides genetic counseling and guidance.

Dr. Herman described the experience at her hospital. Samples are sent out for exome sequencing, and only clinical geneticists can order them. In total, 131 exomes have been completed, with positive results for 52 percent. The first 100 cases had 46 percent with positive result and lead to a change in management. One half were found to be *de novo*, and three novel genes have been identified.

Trends in clinical sequencing include the expansion to carrier and population screening; a shift from gene identification to the validation of variant pathogenicity that warrants the development of rapid, robust tools to validate potential disease-causing variants, particularly missense variants; a movement toward whole genome sequencing with the assessment of chromosome rearrangements included in the analysis; and increased complexity of assessing non-coding variants.

"Vertical Integration" Around Clinical Problems
Calum MacRae, M.D., Ph.D., Harvard Medical School and Brigham and Women's Hospital

Dr. Calum MacRae presented several family cases from his practice indicating challenges with clinical phenotyping of the patients. In the case of a family with a single gene disorder but whose members had multiple phenotypes, most wanted to know if they will live longer or feel better. Asymptomatic EKG findings, 12 different lamin syndromes, and enlarged dilated cardiomyopathy (DCM) chamber were determined to be pleiotropic manifestations of DCM causing genes. In another example, a patient suffered from extreme anxiety attacks based on fear of sudden death; a pathogenicity assessment was conducted and implicated *KCNQ1*. For some clinical studies, diagnosis is only possible by provoked phenotypes, most commonly attained through observation of posture, exercise, and recovery time. The most predictive of these is the change seen in Q-T during recovery.

Phenotyping has limitations and issues in multiple arenas, including clinical care, genetics, and personalized medicine because of the domination of morphology, legacy phenotyping, binary technologies, diagnosis of diseases at later stages, high costs, and other factors. Advances in phenotype information can be stymied by silent alleles; the dependence of genetic architecture or phenotype architecture (resolution, selection pressures, and environmental contribution); and limitations of genetic studies to date. Dr. MacRae shared examples of areas where model organisms have helped, such as: saturation screens, reverse genetics, empiric predictions of genotype-phenotype correlation at scale, environmental modeling, and gaps in genetic or phenotypic architecture. These included initial predictions for congestive heart failure and modeling the chronic disease arrhythmogenic right ventricular cardiomyopathy in 5 days.

New, translatable human phenotypes are needed. Existing data types should be reappraised, and new analytic approaches need to be developed. In addition, functional genomics has to be brought both into the clinic and then back for the animal model adjustment. New technologies should be implemented, and phenotype narrative should be collected in the clinic.

Current external pressures demand revolution in multiple elements of the translational cycle. These include comprehensive approaches to phenotyping to maximize yield from genomics; clinical investment in research and development infrastructure; and new translational teams addressing curation, biology, and clinical care. Clinical utility of phenotyping also should be proven. Model organisms offer scalable *in vivo* genetics, biology, and chemical genomics, and fundamental biology should be built around clinical problems. Phenotypic innovation aligns discovery, clinical care, and cost, and work should be conducted using a shared lexicon; it should focus on advancing discoveries about genomes, phenomes, perturbations, and networks, while avoiding unaffordable duplication. Finally, a new minimal clinical data set for the 21st century needs to be established that is rooted in fundamental biology rather than technology; complements the current clinical care, genomics, and eHealth; accelerates translation; and optimally is both portable and affordable.

Round Table Discussion

Dr. Lussier summarized the panelists' presentations. Dr. Lussier described hypotheses that consider the transcriptome as an iteration of the intergenic and genetic polymorphisms observed in GWAS, integrators, and *ex vivo* assays of the transcriptome. Dr. Kurzrock presented the advancements in oncology, particularly precision medicine. Dr. Herman shared her experience in genetics, including that 40 percent of the coding genome has been characterized, as well as recommendations from the ACMG. Dr. MacRae presented on increasing the capability of phenotyping within clinical care.

A participant asked whether and how immunologic phenotyping can be done in model organisms. Dr. MacRae replied affirmatively and said that it has been gathered from model organisms, such as in lineage markers in fish. He added that emphasis should be on building collective animal model communities to run the problems that impact society by having vital phenotypes for many model organisms. He also noted that genomics has been successful because of its comprehensiveness, and said that to match the investment in genomics, phenomics needs a similar level of investment that facilitates a more cohesive, systematic approach.

A participant referred to Dr. Herman's talk and asked about the possibility of making full genome sequencing a clinical standard and asked whether mosaicism was noticed in her studies. Dr. Herman stated that families can request that samples be de-identified if they want, but many prefer panels either because of insurance coverage or their desire not to have secondary findings. Her preference is to conduct full exome sequencing and indicated that research would benefit from access to complete genome information. She added that researchers are picking up some mosaicism.

Participants wondered what could be done reasonably and efficiently to improve the data on clinical records and to incorporate animal model organisms into the clinical cycle; that the EHR needs a major revision to support phenotype research was affirmed as a given. Dr. Lussier observed that the EHR was designed in the 1970s and 1980s based on the IT systems that were available; IT needs to better support physicians. Dr. MacRae agreed that the phenotype landscape should be recast to allow ambient collection of phenotyping, and it is an ideal time to change a key component in building a phenotype data community that includes the model organism community. Dr. Kurzrock commented that the current EHR reduces productivity 30 to 50 percent but can be fixed.

A participant noted that semantics will need to be supported to move phenotype focus from rare to common diseases. Dr. Herman said that standards need to be developed with an aim to impact clinical care. She added that resources also are needed to develop technologies that can help researchers obtain more quantitative phenotypes. Dr. MacRae agreed that attention should be focused on innovative technologies (e.g., combination of risk model, medical, and quantified cell movements) rather than on traditional phenotyping tools.

Session 6: Informatics Tools for Phenotypic Analysis and Data Sharing
Chair: *Philip Bourne, Ph.D., Office of the Director, NIH*

Connecting the Pieces: How to Make a Biomedical Information Ecosystem Run
Maryann Martone, Ph.D., University of California, San Diego

In this session, Dr. Haendel substituted for both the Chair (Phil Bourne, NIH) and first scheduled speaker (Maryann Martone, UCSD).

Enabling a Cross-Species Disease Research Ecosystem
Dr. Melissa Haendel

To find data and use it, researchers have to identify what to talk about and who to talk to. Experts are relatively easy to identify, but the model organisms are not. Dr. Haendel described an experiment in reproducibility to determine how identifiable the models are in published literature. Her team studied the domains, impact factors, and reporting guidelines from 248 papers in 84 journals and found that only 50 percent of resources were identifiable, that is, which organisms or cell lines were being discussed in the article/journal. This finding had no correlation with journal impact factors or the level of strictness in journal reporting guidelines.

To begin to alleviate the problem, the Resource Identification Initiative supported a pilot project to help authors find the resource identifiers. In the pilot workflow, the author goes to the Research Identification Portal to locate the Research Resource Identifier (RRID) and includes the RRID in the methods section and as keywords. Only resources actually used in the research are included in the RRID. A post-pilot identification activity found improvement in the identification of antibodies, organisms and tools.

Dr. Haendel shared work to enable genotype-phenotype data capture and interoperability, particularly through the development of PheNote, an online collaborative genotype-to-phenotype curation tool that works both with and across any species. Other activities include joint efforts with journals to develop phenotype records at the time of publication, the development of a PubMed browser to allow users to find co-occurrences of a phenotype profile of interest, collaboration with the Global Alliance for Genomics and Health on such tools as a “Matchmaker Exchange,” and the development of global data sharing practices. Finding collaborators for functional validation, such as connecting phenotyping experts with patient phenotype profiles, also is key to enabling a cross-species disease research ecosystem.

Evolutionary Relationships as a Paradigm for Integrating Biological Knowledge: The Gene Ontology Phylogenetic Annotation Project
Paul Thomas, Ph.D., University of Southern California

Darwin’s species tree is useful to identify characters that extant species have in common due to inheritance. This approach allows inferences about common ancestors, visualizes specific changes of biology as seen through branches, and facilitates inferences about the position of uncharacterized sequences in the tree. This approach can be used at the level of genes to describe gene function as an evolutionary character in a biology tree. The evolutionary framework exists and allows inferences about

integration and unknown characteristics to be made, recorded, and traced; these inferences could be improved by additional curation.

The GO Phylogenetic Annotation Project (GO PAP) has developed a database containing gene function information ("annotations") extracted from 120,000 science papers whose authors come primarily from 12 organizations. The goal is to integrate gene phylogeny (i.e., how genes are related) with experimental information about gene function, including the relationship between genes among different model organisms, and which functions are conserved among which homologs. The project uses gene trees to integrate multiple types of knowledge in a software model called PAINT that provides graphic representation of the relationships. GO PAP has completed a total of 1,914 annotated families, contributing to the doubling of the information about human genes.

Large models can bring together information about as many genes as possible, compare each model organism to specific human diseases or conditions, and query the similarities. Dr. Thomas demonstrated this through work done on *ALDH1L1*, in which a new gene and a new function appeared due to gene duplication. The evolution of a one-carbon folate pathway can be traced by building on gene presence and absence over time.

Many GO annotations are derived from experimentally observed phenotypes. However, phylogenetic annotation to date does not usually build phenotyping into its model. New tools should be developed to make computational representation of these causal connections.

Evolutionary information can help identify similarities and differences between a model and a human system at the level of biological pathways/processes. The GO Consortium has developed a general infrastructure for inferring and annotating the evolution of any biological "character." This encompasses integrating information at points of common ancestry and inferring unknown character states of living organisms. It could be extended to or integrated with phenotype information, such as through capturing information from a computational model of normal biology and how perturbations can result in particular phenotypes.

Of Worms and Men: A Data Journey
Nicole Washington, Ph.D., Lawrence Berkeley National Laboratory

Dr. Nicole Washington described the Monarch Initiative's activities to elucidate the role and utility of phenotyping to determine the genetic causes of diseases. She highlighted important strategies to improve the use of the widest array of animal models possible. To answer "big" health science research questions, integration of data related to phenotypes, genotypes, evolutionary conservation, and knowledge networks are needed. Concepts should be mapped to a common language using ontologies. This can be for a number of biomedical health aspects, such as genotypes, phenotypic features, diseases, drugs, and pathways. A common model for integrating genotypes should be developed to provide the scaffold to integrate historical genotype-to-phenotype data, as various species (e.g., human, worm, fly, fish, frog, and mouse) makes different genotype-to-phenotype associations, which may need to be curated differently. Understanding the relationships of each of the genotypes is instrumental to understanding how to relate phenotypes.

Model organisms alone supply 50 percent of phenotypic knowledge about human genes, and all the models are needed to help fill the gap of what is not known about genotype-to-phenotype associations. Many and varied sources of genotype-to-phenotype association data should be found, acquired, and integrated with the above strategies. Software should be provided to enable others to map their data to a common model, and visual tools should be created to aid in the interpretation of phenotype data. In

addition, the provenance of the assertions linking genotype-to-phenotype should be tracked, and data and standards should be shared with the community.

Round Table Discussion

Participants reflected on the use of the RRID project to potentially impact reproducibility and noted that many animal species (e.g., cat, dog, horse, and others shown on Dr. Washington's slides) are seen clinically, and their data could be captured from existing databases and included. Dr. Haendel said that the hope is that the RRID project will help facilitate author recognition to identify resources. Some resources are further along and the project is working with resources that are not as advanced in this (e.g., cell line vendors to uniquely identify cell lines). Much of this has to do with the provenance and types of data gathered in animal clinics; most of the data are attached to the resource. Although identification is the first step, the data attached and how researchers go about getting it are what matter. Dr. Washington added that owners often have their animals genotyped, and veterinary hospitals could be a potential partner in data sharing. The diversity of possible animal models is an important aspect.

Participants discussed non-human primate (NHP) models. One participant commented that standard animal models may not be useful in NHP studies and noted that the National Primate Research Centers program, which is supported by the NIH, is developing genotype and phenotype databases, which could be useful resources for the model organism community. The National Primate Research Centers have conducted an extreme phenotype survey to identify animals (or groups of animals) in the Centers that have phenotypes relevant to some diseases that do not yet have good models. One of the challenges of conducting research in a non-standard animal model is obtaining funding; a significant portion of a grant application is devoted to a rationale of why it is best to do a study in the proposed animal model. It would be helpful for those submitting applications to more easily justify certain animal models if clinician scientists indicate the most important phenotypes for the disease(s) in which they are interested. The participant asked if other such gaps (i.e., where standard animal models are falling short) could be identified as well. Dr. Haendel agreed that some types of studies (e.g., behavioral) are best conducted in NHP models, and that similarities of physiological and molecular functions make NHPs appropriate for comparison with human conditions. To help justify a model system, researchers must consider a variety of data types that might be informative; it is not just about the right anatomy, but also about having the right tools to support the assays to be done and determine that models have the same molecular function for comparison. Dr. Thomas agreed that a framework that coalesces a wide variety of information (molecular systems, assay ability) and requires the community at large to work together is ideal. Dr. Washington mentioned preliminary scoring work underway to assess how to understand if the particular phenotypes of a given system are the right complement to create a model for specific disease or condition.

A participant noted the importance of evolutionary data in NHPs, which include a group of genes that cannot be modeled (i.e., the genes are duplicated only in chimpanzees and humans and specialized only in humans), and asked whether the GO project will consider incorporating anatomical databases. Dr. Thomas replied that the GO project has started integrating information about cellular and subcellular levels from anatomical databases via low-throughput assays; it is a significant effort, however, challenges remain to capture, integrate, and ensure the quality level of data. Dr. Haendel added that the goal is to incorporate anatomical information from the cellular and subcellular levels.

Closing Remarks and Recommendations
Mary Mullins, Ph.D., University of Pennsylvania

Dr. Mary Mullins synthesized the conference presentation and discussions over the past day and a half, highlighting modeling progress and data needs to advance research on human phenotypes and cross-species analysis.

Human Genomic Information and Phenomics

A wealth of information about the human genome has become available with the advent of high-throughput sequencing methods in recent years. Human genome studies, patient cohorts, and studies of human conditions have had as their goal the identification of effector genes of human diseases. GWAS and single SNP variants can account for new disease gene identification, and although these approaches have been valuable, they also have been limiting regarding many human phenotypes. This workshop highlighted that integrating these data with phenotype information (environment conditions) can provide better candidates for disease gene effectors than are currently available. The goal is to use this information for precision medicine.

Data from the human genome, exome, or SNP needs to be associated with thorough (deep) patient phenotype data that are accessible broadly using human phenotype ontology and software programs (e.g., PhenIX) and that have demonstrated the value of this approach. To cross-analyze studies and integrate data information from multiple studies, research studies need to use common standards and measures in analysis. Dr. Hamilton described the PhenX toolkit, initially developed to help investigators with GWAS studies and now being expanded for the broader human condition and other research domains using a common ontology and standard terminology that is important for phenotype analysis. Multiple ontologies and terminologies exist for human phenotypes and are partially integrated with each other and used in a variety of communities.

The challenge with EHRs is that no one system covers all known diseases and conditions. Currently, there are no standards for developing clinical phenotype definitions. Although the majority (90%) of hospitals have demonstrated Meaningful Use to support clinical care, the completeness and accuracy in EHRs varies. In addition, the information is not standardized in a way that facilitates evaluation. More than 100 EHR vendor products are available that further stymy access to information. Researchers currently do not control the design of EHR documentation and coding practices.

Common terminology to report patient data and increase the population base of phenotype data for human conditions is needed. As seen in studies, a use of HPO in clinics to report patient data by physicians and database use can provide more efficient diagnostics and reduce health care costs. eMERGE has demonstrated that data collected for the purpose of clinical documentation of EHRs or billing claims data also can support research. The large volume of data in EHRs partly compensates for limited quality and resolution. Standardization currently happens behind the scenes. Dr. Chisholm described Northwestern's BioBank as an example of how EHR data can be integrated through eMERGE, as well as the value of flipping the normal GWAS method by starting with a phenome-wide association study and examining target genotypes. The PheWAS approach requires a large cohort of patients with genotype data and many diagnoses.

Free text remains an important source of phenotypic information in EHRs. Natural language processing techniques are required for information extraction and standardization.

Multiple inconsistent phenotype definitions have a negative impact on clinical research from comparability and reproducibility. A key to defining genetic contributions to precision medicine requires large numbers of well-phenotyped human genomes.

The following points were made in discussion:

- It might be helpful to consider use cases and implementation to obtain a better understanding of how many resources would be needed going forward.
- Much of the computational phenotype work that has been accomplished has been focused on rare disease phenotypes; opportunities exist in the common diseases arena, and new data structures or algorithms should be developed and tested to see what works.

Cross-Species Phenotype Analyses and Ontology

Ontology is important for cross-species data integration to aid disease diagnostics and to understand the correlation between genotypes and phenotypes.

Semantic similarity approaches, that is, non-exact phenotype approaches, are effective. Because orthologous genes generally conserve their function across species, model organisms can inform human disease genes. Evolutionary information on gene function can identify similarities and differences between model organisms and human systems. This could be extended to an integrated approach for phenotype information.

GeneWeaver, described by Dr. Chessler, provides a way to integrate data across species, including the mouse. Cross-species integrative genomics enables global comparison of animal models, assays, and diseases based on underlying biology. Different models need to be used for different diseases and for different phenotypes of one disease.

Species-neutral ontologies are needed. Dr. Mungall discussed the Monarch ontology for this purpose. An increase in knowledge curation by expert curators also is needed, using cross-species, integrative phenotype ontology approaches to allow comparative phenomics. Model organism and human phenomics can be integrated into whole exome human disease analysis to prioritize candidate variants that are identified. There can be hundreds of variants, and Dr. Smedley showed how Exomiser can be used successfully in such analyses. Expanding this to include nonhuman genome analysis is being done in a Genomiser program. Deep clinical phenotype data can greatly inform translational bioinformatics. A challenge is to include negative phenotype data as well as phenotype frequency data. In addition, behavioral information has not been well covered in models or ontology terms.

Patients whose variants disrupt the same pathway share a broad range of phenotypic similarities. Functionally linking genes through orthogonal data sources, protein interaction studies, co-expression, and similar model organism phenotypes is valuable in connecting human disease with animal models and in identifying these important exome variants as disease effectors.

Phenotyping of the less-studied gene is needed to allow model organism phenomics to inform candidate disease analysis; there is no information about this for approximately 20 percent of the human genome.

Large-scale resources in model organisms must have a well-planned foundation, including standard operating procedures (SOPs), standardized pipelines, appropriate analysis, and use of standards to enable data integration and reuse. The analysis, whether it relies on existing or new algorithms, is a key component of the success of the use of these data. Resource availability will ensure large-scale data usable to the broader scientific community. A careful curation process must be part of the annotation of data results and is key to the value of the data over time.

There is no one right model organism for a given study. Different aspects of an investigation may best be performed in different species, depending on the questions being addressed and the aspect of the disease being studied.

A general theme is that both the breadth and the depth of data are needed to compare and link phenotypes across species and to develop multiple phenotype profiles to more precisely define how models map to disease.

In addition to semantic mappings, it is important to consider the molecular phases of phenotypes when mapping across organisms. This requires data-driven algorithms that are capable of data-driven association of phenotypes across organisms based on molecular similarity that is not restricted to existing annotations and biological knowledge. Complexities of genetic-based phenotypes should be taken into account by considering network-based approaches that integrate functional information. Characterizing genetic bases of human disease requires sophisticated computational algorithms that can integrate functional and genomic data in both human and model organisms.

Dr. Kurzrock described how precision medicine works in the clinic with demonstrated improved outcomes. She suggested that the patients should be treated earlier and that the approach should be used at earlier stages of the disease. To determine the best drug treatment, the specific characteristics of the mutations that are identified should be studied, as well as how protein function is affected.

Other presenters discussed techniques and technologies of interest. A new program that uses liquid biopsy allows targeting of multiple metastatic sites at once. In addition, transcriptome analysis of clinical samples add valuable phenotyping and precision information. Dr. Herman presented clinical exome sequence data in pediatrics that also showed high diagnostic yield.

Dr. McRae highlighted the need to change diagnostic tests in the clinic, noting that many tests exist for historical reasons, and that they could be better informed for phenotype analyses and the use of orthogonal phenotyping. Broadened clinical phenotyping can lead to quicker diagnosis, more quantitative metrics, and more phenotype data for exome variant analysis. More model organism phenotype expansion can be done for quantitative metrics in the environmental modeling arena as well as more imaging in organisms such as the zebrafish for quantitative analysis. He described the successful use of model organisms to generate disease models with specific human alleles and homologous phenotype results, which were found in zebrafish.

Areas for specific action include:

- Funding for comprehensive, computational phenotype resourcing is needed for the current large phenotype projects (e.g., IMPC, UDP).
- Consider funding collaborative consortia integrating human phenotype ontology along with model organism ontologies and functional validation of human disease causes.
- Monarch and HPO, which are comprehensive interlinked databases of human phenome and “disease-ome” with relevant animal model data, are underfunded and need to grow.
- Some areas in human and mouse phenotype will require extension (e.g., behavior, metabolism, and craniofacial data).
- The great majority of GWAS hits are noncoding, and there may be more regulatory mutations in disease than has been thought. Although many animal models of gene regulation exist, their data are not well integrated or annotated for analysis.
- Although current algorithms are working well for rare diseases, more sophisticated representations of phenotype will be required for common, complex diseases, including cancer.
- Integrated algorithms for matching phenotypes to molecular pathophysiology are needed to shorten the time to enable diagnosis.
- The terms used in the EHR should be standardized to improve consistency between different clinical sites.

- Standardization of phenotypes, extracted from health record data in reference to ontologies and better integration between the ontologies using basic and clinical research for health care should be implemented.
- Standardization and validation of phenotype definitions are needed across research studies.
- Data scientists should be involved in the collection of electronic health data.
- Governmental requirements could help to change the culture of recording data, such as through standardized terms and greater accessibility to EHRs for analysis.
- Phenotype data being generated in clinics could be better leveraged for disease diagnosis and disease gene identification. They are currently underutilized and difficult to mine. The NIH could have researchers develop EHR programs that are accessible or mandated to the clinics.
- Deep phenotyping in the clinics should be improved, such as through better assays, integration of model organism phenomics and genetics, and phenotyping of understudied genes, so that phenotype information from variants can be better ascertained as candidates.
- Curation of the literature with currently non-identifiable data can increase the phenotype information base.
- Support for model organism research to guide Gene-Environment, Gene-Gene-Environment (microbiome) studies. Consider implementing of the standard measures for environmental exposures. Define and recommend a standard panel of environmental exposures for inclusion in a "minimal clinical data set".
- Consider metabolomics phenotypes as biomarkers for the metabolic profiles.
- The goal is to use all of the information to improve the ability to do precision medicine, be more effective in the clinics, and improve human health.

Final Discussion

Dr. Mirochnitchenko asked the moderators to reflect on the Symposium discussions and share what they might change. Participants also shared their thoughts on the future directions and opportunities in model organism phenotyping.

Moderators recognized that all the various experts participating in the Symposium have a piece of the puzzle and commended the NIH for thinking innovatively about collaborative activities. People involved in these fields need to talk with each, recognizing that each field brings a separate way of thinking.

One area not raised in the meeting is the sociology of phenotyping. An important driver will be when patients tell hospitals that they want a personalized approach to their condition; in such a situation, "ambient" phenotyping will follow genotyping.

An opportunity exists regarding a comprehensive database. Phenomics does not have MGI or a similar database to use.

The phenotype information must be gathered and curated before semantic similarities can be identified. There is an increasing criticism of expert curation, which could be ameliorated by having investigators use standard terms, provide clearer resource information in their publications, and so forth. It was noted that sequencing data was generated electronically from the onset and provided a breakthrough for genomics.

The person who designs instruments has a role to ensure its adaptability to the database. Participants were encouraged to attend the upcoming FORCE11 Workshop.

Participants drew a mental picture of the physician's office in the future. It will involve collaboration between physician and patient, with direct digital input of phenotypic information as part of a broader EHR. Physicians should be able to press another button on a tablet to conduct a real-time, basic search of research regarding symptoms the patient has mentioned right then. The physician will review the survey results and select appropriate tests. The EHR should be perceived as providing this type of assistance to clinical care.

Phenotype information will never be fully captured with consideration of environmental factors. However, the environmental arena is far behind in this area of science; an environmental exposure ontology has been started but needs much more expansion.

It was noted that induced pluripotent stem cells (IPCs) offer a range of possibilities for biomedical research and personalized medicine. Moderators agreed but noted that IPC data are in some systems (Eagle Eye) but not yet in others (Monarch).

For the past 14 years, NHP researchers have gathered environmental health on 4,000 animals since age 6 months old and will have aged animals soon. The influence of the microbiome on immune profiles in NHP models was noted.

Dr. Kibbe, NCI, referred to NIH's genomic data sharing policy, which was effective January 2015, and asked how it might be useful from the phenotype perspective and what core elements should be shared.

Participants also noted that it would be helpful to have use cases from UDP or other programs. In addition, the pharmaceutical industry has worked extensively with animal models and may be a potential resource.

Adjournment

Dr. Mirochnitchenko thanked the moderators, presenters, and attendees for their contributions and adjourned the Symposium at 1:12 p.m.

Appendix A. Symposium Agenda

Agenda

Purpose of the Meeting: The purpose of the Symposium is to convene a colloquium on the current status of phenomics and its role in closing the gap that exists between biomedical research and clinical medical practice. The wealth of whole organism, cellular, and molecular data generated in the research laboratory must be translated into clinically relevant knowledge that enables the physician to make the best possible treatment decisions. Phenomics is gaining momentum due to the availability of the complete genomes for many organisms as well as higher throughput methods to genetically modify model organism genomes and observe and record phenotypes. Disease models comprise some of the most important tools of biomedical research. The efficacy of the use of disease models is based upon the principles of evolutionary conservation between species, including conservation of pathogenic disease mechanisms. The lack of alignment of phenotypes between model species and humans has been a historic impediment to understanding disease processes. Further progress depends upon integration of clinical, biological, and genomic data, and development of the tools for identification and analysis of specific and amendable disease-causing molecular phenotypes of various diseases. Determining the molecular "fingerprints" that define similarities, and differences, between disease models and the actual human conditions will ultimately lead to more predictive models. The availability of precise molecular phenotypes for diseases, increasing the uniformity with which these phenotypes are described, and better systems for organization and retrieval of this information, will allow the "historic impediment" to be circumvented.

Meeting participants will provide insight to the Division of Comparative Medicine and other NIH units for the development of potential initiatives in this rapidly evolving area of research and development.

Organizing Committee: Olivier Bodenreider (National Library of Medicine, MD), Philip Bourne (DS/NIH, MD), Janan Eppig (The Jackson Laboratory, ME), Melissa Haendel (Oregon Health & Science University, OR), Yves Lussier (University of Arizona, AZ), Oleg Mirochnitchenko (ORIP/NIH, MD), Mary Mullins (University of Pennsylvania, PA), Olga Troyanskaya (Princeton University, NJ), and Harold Watson (ORIP/NIH, MD)

Day 1 – Thursday, September 10, 2015

7:30 a.m. – 8:30 a.m.	**Registration**
8:30 a.m. – 9:00 a.m.	**Introductions and Welcome**
	Symposium Introductions: *Oleg Mirochnitchenko, Ph.D., Office of Research Infrastructure Programs (ORIP), NIH (Speaker)* *Harold Watson, Ph.D., ORIP, NIH (Speaker)*
	Welcome: *Franziska Grieder, D.V.M., Ph.D., ORIP, NIH* *Stephanie Murphy, V.M.D., Ph.D., DCM, NIH*
9:00 a.m. – 9:45 a.m.	**Keynote Presentation**
	Deep Phenotyping for Translational Research and Precision Medicine *Peter Robinson, M.D., M.Sc., Max Planck Institute for Molecular Genetics*
Session 1:	**The Current Status of the Human Clinical Phenotype Ontology and Terminology, and Associated Data Annotation and Use** **Chair:** *Olivier Bodenreider, M.D., Ph.D., National Library of Medicine*
9:45 a.m. – 9:55 a.m.	**Chair Overview**
9:55 a.m. – 10:15 a.m.	**The PhenX Toolkit: Standard Measures for Collaborative Research** *Carol Hamilton, Ph.D., RTI International*
10:15 a.m. – 10:35 a.m.	**Clinical Phenotyping from Electronic Health Records: Opportunities and Challenges** *Rachel Richesson, M.P.H., M.S., FACMI, Duke University*
10:35 a.m. – 10:55 a.m.	**Progress Toward Precision Medicine and the Challenges of Integrating Genomics into Electronic Health Records** *Rex Chisholm, Ph.D., Northwestern University*
10:55 a.m. – 11:25 a.m.	**Round Table Discussion**
11:25 a.m. – 12:25 p.m.	**LUNCH**

Meals and light refreshments are at the expense of attendees.
(Attendees will be responsible for meals/light refreshments on their own, at their own cost. The government and/or government contractors are not involved in facilitating the provision of food and/or light refreshments.)

Session 2:	**Cross-Species Phenotype Analysis and Ontology** **Chair:** *Melissa Haendel, Ph.D., Oregon Health and Science University*

12:25 p.m. – 12:35 p.m.	**Chair Overview**
12:35 p.m. – 12:55 p.m.	**Crossing the Species Divide** *Chris Mungall, Ph.D., Lawrence Berkeley National Laboratory*
12:55 p.m. – 1:15 p.m.	**Disease Variant Prioritization and Model Discovery Through Cross-Species Phenotype Analysis** *Damian Smedley, Ph.D., Wellcome Trust Sanger Institute*
1:15 p.m. – 1:35 p.m.	**Exploiting Mouse Genotype-Phenotypic Associations for Disease Genomics** *Caleb Webber, Ph.D., Oxford University*
1:35 p.m. – 2:05 p.m.	**Round Table Discussion**
2:05 p.m. – 2:25 p.m.	**Break**
Session 3:	**Large Scale High Throughput Analysis of Disease Model Phenotyping Data and Annotation of Gene Function** **Chair:** *Janan Eppig, Ph.D., The Jackson Laboratory*
2:25 p.m. – 2:35 p.m.	**Chair Overview**
2:35 p.m. – 2:55 p.m.	**Functional Exploration of Human Cancer Genomes Using Flies** *Erdem Bangi, Ph.D., Mount Sinai Hospital*
2:55 p.m. – 3:15 p.m.	**The Zebrafish Mutation Project** *Derek Stemple, Ph.D., Welcome Trust Sanger Institute*
3:15 p.m. – 3:35 p.m.	**Finding Cross-Species Phenomic Similarity through Integration of Heterogeneous Functional Genomic Data** *Elissa Chesler, Ph.D., The Jackson Laboratory*
3:35 p.m. – 4:05 p.m.	**Round Table Discussion**
Session 4:	**Linking Disease-Relevant Phenotypes with Physiologically Relevant Molecular Pathways and Networks** **Chair:** *Olga Troyanskaya, Ph.D., Princeton University & Simons Center for Data Analysis*
4:05 p.m. – 4:15 p.m.	**Chair Overview**
4:15 p.m. – 4:35 p.m.	**Using Networks to Re-Examine the Genome-Phenome Connection** *John Quackenbush, Ph.D., Dana-Farber Cancer Institute*
4:35 p.m. – 4:55 p.m.	**Human Phenotype Networks** *Jason Moore, Ph.D., M.S., The Perelman School of Medicine, University of Pennsylvania*

4:55 p.m. – 5:15 p.m.	**Understanding the Molecular Basis of Human Disease by Mapping Across Tissues and Organisms** *Olga Troyanskaya, Ph.D., Princeton University & Simons Center for Data Analysis*
5:15 p.m. – 5:45 p.m.	**Round Table Discussion**
5:45 p.m.	**Adjournment**

Day 2 – Friday, September 11, 2015

Session 5:	**Clinical and Experimental Biology Data Integration Emerging Field of Precision Medicine** **Chair:** *Yves Lussier, M.D., FAMCI, University of Arizona*
8:30 a.m. – 8:40 a.m.	**Chair Overview**
8:40 a.m. – 9:00 a.m.	**Precision Medicine and the Reclassification of Cancer: Divide and Conquer** *Razelle Kurzrock, M.D., University of California, San Diego*
9:00 a.m. – 9:20 a.m.	**Linking Disease Model and Human Phenotypes: The Clinical Geneticist Perspective** *Gail Herman, M.D., Ph.D., Nationwide Children's Hospital*
9:20 a.m. – 9:40 a.m.	**"Vertical Integration" Around Clinical Problems** *Calum MacRae, M.D., Ph.D., Harvard Medical School and Brigham and Women's Hospital*
9:40 a.m. – 10:10 a.m.	**Round Table Discussion**
10:10 a.m. – 10:30 a.m.	**Break**
Session 6:	**Informatics Tools for Phenotypic Analysis and Data Sharing** **Chair:** *Philip Bourne, Ph.D., Office of the Director, NIH*
10:30 a.m. – 10:40 a.m.	**Chair Overview**
10:40 a.m. – 11:00 a.m.	**Connecting the Pieces: How to Make a Biomedical Information Ecosystem Run** *Maryann Martone, Ph.D., University of California, San Diego*
11:00 a.m. – 11:20 a.m.	**Evolutionary Relationships as a Paradigm for Integrating Biological Knowledge: The Gene Ontology Phylogenetic Annotation Project** *Paul Thomas, Ph.D., University of Southern California*
11:20 a.m. – 11:40 a.m.	**Of Worms and Men: A Data Journey** *Nicole Washington, Ph.D., Lawrence Berkeley National Laboratory*

11:40 a.m. – 12:10 p.m.	**Round Table Discussion**
12:10 p.m. – 1:00 p.m.	**Closing Remarks and Recommendations** *Mary Mullins, Ph.D., University of Pennsylvania*
1:00 p.m.	**Adjournment**

Appendix B. Abstracts of presentations

Abstracts of the Presentations

Keynote: Deep Phenotyping for Translational Research and Precision Medicine
Dr. Peter N. Robinson, Institute for Medical and Human Genetics, Charité-Universitätsmedizin, Max Planck Institute for Molecular Genetics, Germany

Phenotype assessment plays a key role in clinical practice and medical research, and yet phenotypic descriptions in clinical notes and medical publications are often imprecise. Deep phenotyping can be defined as the precise and comprehensive analysis of phenotypic abnormalities in which the individual components of the phenotype are observed and described in a way that allows computational analysis. The Human Phenotype Ontology (HPO) is being developed to enable phenotypic information to be described in an unambiguous, standardized fashion in medical publications and databases. The HPO has been adopted by a number of groups in rare disease research, including the DECIPHER and DDD databases of the Sanger Institute, the NIH Undiagnosed Diseases Program/Network, ECARUCA, the rare disease part of the UK 100,000 Genomes project, and many others, enabling them to exchange next-generation sequencing data to assist in disease diagnosis. This is realized using Phenotype-driven bioinformatics algorithms that leverage HPO for prioritizing genes in exome studies. I will explain how such HPO-driven algorithms work and demonstrate how we have used them to integrate clinical and basic research data for translational research. I will conclude by presenting current projects aimed at developing HPO-based resources for common (complex) disease and for the understanding of mutations in the non-coding portion of the human genome.

Session 1. The Current Status of the Human Clinical Phenotype Ontology and Terminology, and Associated Data Annotation and Use
Chair: ***Dr. Olivier Bodenreider, NIH, MD***

In parallel to the development of methods for studying genetic variation, the past decade has seen the collaborative development of many ontologies for biomedical research (e.g., Gene Ontology, Human Phenotype Ontology), as well as increased use of standard terminologies in clinical institutions (e.g., LOINC, SNOMED CT). Yet the representation of clinical phenotypes in

standard terminologies remains insufficient, and limited interoperability has been achieved between datasets collected for research and clinical care. Two main approaches to phenotyping have emerged. On the one hand, detailed phenotype ontologies and clinical data elements support the precise annotation of datasets (prospectively). On the other hand, pragmatic methods for identifying phenotypes in EHR data (retrospectively) enable the secondary use of observational data. These two approaches are complementary and equally important.

The PhenX Toolkit: Standard Measures for Collaborative Research
Dr. Carol M. Hamilton, RTI International, NC

To help investigators identify opportunities for collaborative biomedical research and to improve the consistency of data-collection, the Web-based PhenX Toolkit (consensus measures for Phenotypes and eXposures, https://www.phenxtoolkit.org/) provides standard measures, protocols and bioinformatics support for assessing human phenotypes and exposures. In PhenX Phase I, the emphasis was on identifying recommended, well established measures and protocols that were suitable for genome-wide association studies (GWAS) with an emphasis on common complex diseases. PhenX Phase I established a consensus process, a bioinformatics pipeline, and addressed 21 broad research domains, as well as adding depth in substance abuse and addiction. In PhenX Phase II, with continued funding as a Genomic Resource, the scope has expanded to include measures relevant for clinical and translational studies and rare genetic conditions. PhenX Phase II will address four additional research domains, and will assemble Expert Review Panels to review and update Phase I Toolkit content. In Phase II, the PhenX Steering Committee prioritized expansion of the Toolkit to include measures relevant to rare genetic conditions, including a crowdsourcing effort to annotate measures already in the Toolkit for use in specific rare genetic conditions. To support investigators who want to collect data via the Web, PhenX protocols are being made available as REDCap instrument zip files that can be directly uploaded to REDCap studies (http://project-redcap.org/(. In addition, Phase II supplemental efforts are adding depth to the PhenX Toolkit in mental health, tobacco regulatory, and sickle cell disease research. PhenX is managing change (updating measures and protocols) and extending the scope to meet the evolving needs of the scientific community. These efforts will ensure that the PhenX Toolkit will continue to provide the biomedical research community with easy access to standard measures and the potential to increase the overall impact of individual studies by facilitating cross-study analysis. Funding provided by NHGRI, co-funded by NIDA: Genomic Resource award U41 HG007050.

Clinical Phenotyping from Electronic Health Records: Opportunities and Challenges
Dr. Rachel Richesson, Duke University, NC

Wide-spread adoption of electronic health records (EHRs) containing rich longitudinal clinical data has led to expanded opportunities to repurpose these data for clinical and genomic research. Standardized EHR-based condition definitions (also called “clinical phenotypes”) can support the rapid development of new biomedical investigations. However, the development of standard EHR-based clinical phenotype definitions for various diseases and conditions is challenging due to the heterogeneity of EHR systems and valid concerns about the

completeness and accuracy of these data. A number of different coding systems are used in EHRs but no single system provides complete coverage for all known diseases and conditions, with rare and genetic disorders being most unrepresented. There is no standard methodology for developing clinical phenotype definitions, nor is there a single recognized authority to endorse or host them. This talk will describe the strengths and limitations of EHR data for research, relevant coding systems and tools for linking between coding systems and specialized ontologies, and strategies for the standardization and dissemination of clinical phenotype definitions, particularly in large research networks.

Progress Toward Precision Medicine and the Challenges of Integrating Genomics into Electronic Health Records
Dr. Rex L. Chisholm, Northwestern University, IL

Biobanks linked to electronic health records (EHR) provide a unique opportunity to study the association between genetic variation and phenotypes. In 2002, Northwestern University established a EHR-linked biobank called NUgene (http://nugene.org). Participants in NUgene have consented to mining of their EHR and provide a DNA sample for studies of genomic variation. This has allowed us to develop electronic phenotyping algorithms to identify cases and controls for GWAS studies, demonstrating the value of this approach. The NUgene biobank has enabled Northwestern to become a site for the eMERGE network. Funded by the National Human Genome Research Institute, eMERGE has developed over 40 high throughput phenotyping algorithms and enabled multiple gene-disease association studies. In addition, the eMERGE network has begun to develop methods for associating specific genomic variants with EHRs and providing clinical decision support to assist care providers in the use of this information.

Session 2. Cross-Species Phenotype Analysis and Ontology
Chair: Dr. Melissa Haendel, Oregon Health & Science University, OR

Disease modeling has traditionally been focused on a few key model organisms that have been especially useful for certain kinds of analyses. While we have been successful in developing technologies and in discoveries thus far in the context of these organisms, we can do more to take advantage of the fact that we can learn different things about the phenotypic consequences of mutation in different organisms. This session aims to highlight the use of ontologies for cross-species data integration to aid disease diagnostics and understand fundamental correlations of genotype with phenotype.

Crossing the Species Divide
Dr. Chris Mungall, Lawrence Berkeley Laboratory, CA

Humans exhibit fundamental similarities with all forms of life, and this forms the basis for use of model organisms to study human diseases and advance health. However, each organismal community uses their own experimental designs and vocabularies for assaying and recording phenotypes. Bridging these vocabularies is a challenge, especially as the phylogenetic distance

increases. In the Monarch Initiative, we have applied a systematic ontology-based approach to describing the attributes of humans and model organisms, allowing automated cross-species phenotype matching to aid disease diagnosis and mechanism discovery. Such approaches also leverage gene knowledge resources such as the Gene Ontology. We have also developed an online collaborative curation tool for performing ontology-based genotype-phenotype curation according to community standards and in support of maximal interoperability across species. While we have come far in providing semantic interoperability across species, we must include more experts across the translational spectrum in the end-to-end process of generating such ontologies, gene or disease annotation using these ontologies, and the downstream use and discovery that invariably leads to overall and incremental improvement.

Disease Variant Prioritization and Model Discovery through Cross-Species Phenotype Analysis
Dr. Damian Smedley, Wellcome Trust, UK

Whilst whole-exome sequencing has revolutionized rare disease research, many cases still go unsolved and result in extended diagnostic odysseys. This is due in part because prioritizing the ~100-1000 loss of function candidate variants that remain after removing those deemed as common or non-pathogenic is still very difficult. Our Exomiser software suite tackles this problem by leveraging the Monarch infrastructure to semantically compare patient phenotypes to existing phenotypic knowledge from disease and model organism databases. Our successes in diagnosis and gene discovery in partnership with the NIH Undiagnosed Disease Program will be presented. Finding quality candidate variants is just the first step towards a diagnosis and potential therapeutics. The identification or creation of appropriate animal models is a vital part of subsequent functional validation and mechanistic studies. Use of semantic phenotype comparison techniques is also extremely useful for identifying such models in the context of the NIH KOMP2 mouse knockout project.

Exploiting Mouse Genotype-Phenotypic Associations for Disease Genomics
Dr. Caleb Webber, Oxford University, UK

The use of the large and growing collection of mouse knock-out genotype-phenotype associations to identify unusually frequent phenotypic-associations amongst, and therefore prioritize, human candidate disease genes is now common practice. Indeed, mouse genotype-phenotype associations provide an excellent benchmark for functional genomics data, enabling integrating tools such as phenotypic-linkage networks, while the alignment of mouse and human phenotype ontologies enables such tools to be more sensitively focused towards specific disease phenotypes of interest. However, ascertainment biases in the mouse data can lead to significant issues when developing gene disease-association prediction methods. Given the need for such methods, addressing these biases through a targeted exploration of functionally-unannotated gene space would significantly aid in their development.

<u>Session 3. Large Scale High Throughput Analysis of Disease Model Phenotyping Data and Annotation of Gene Function</u>
Chair: ***Dr. Janan Eppig, The Jackson Laboratory, ME***

Over the last two decades our ability to design, collect, integrate, analyze, and disseminate large amounts of experimental biological data (and results) has increased dramatically. Biotechnology has revolutionized what and how we address biological questions, and computer technology advances have enabled rapid communication, large-scale data storage, and large collaborative programs, many carried out at great physical distances. As an exemplar of the challenges and potential of large-scale high throughput systematic phenotyping systems, how data can provide valuable biological insights, and why these data are critical, but not yet sufficient, I will briefly describe the mouse KOMP2/IMPC (Knockout Mouse Project2 /International Mouse Phenotyping Consortium) project. I also will briefly describe how the Mouse Genome Informatics (MGI, www.informatics.jax.org) resource is incorporating IMPC phenotyping data and presenting these data in the context of phenotyping data for all other mutant alleles for the same genes, providing a comprehensive and comparative view of the manifestation of gene mutations, their functions, and potential contributions as human disease models. This session includes talks on three projects generating and using high throughput data on mutants and genome variation in Drosophila, zebrafish, and mouse, and varying approaches designed to traverse mutant phenotypes-to-models of human diseases. The advantages and disadvantages of these approaches will be examined, including the lessons learned along the way, and the new insights that many large-scale high throughput projects might yield in the near future.

Functional Exploration of Human Cancer Genomes Using Flies
Dr. Erdem Bangi, Icahn School of Medicine at Mount Sinai, NY

Personalized cancer genomics is providing unprecedented access into the genetic complexity and diversity of human tumors. The next challenge is to utilize this information to establish effective therapeutics. Functional interrogation of cancer genomes using genetic model systems provides a powerful step towards realizing this goal. We have used publicly available human tumor genomes from the Cancer Genome Atlas (TCGA) to generate a set of fly models that capture the genetic complexity and diversity of human tumors. These genetically diverse set of models provide an excellent opportunity to study tumorigenesis and metastasis and explore mechanisms of drug response and resistance in the context of the whole animal. We are now leveraging a platform we established to generate and screen large numbers of personalized fly models in a rapid and cost effective manner to treat individual patients in a clinical study. We start by generating high quality genomic profiles for our patients and use this information to build a personalized fly model for each patient. These models are then screened against an FDA approved drug library to identify drug combinations specifically tailored to each patient. This approach to personalized cancer therapeutics takes advantage of sophisticated genetic tools and high throughput drug screening methods in Drosophila to address tumor and whole body complexities and to identify treatment options for individual patients based on functional exploration of their tumor genomes.

The Zebrafish Mutation Project
Dr. Derek Stemple, Welcome Trust Sanger Institute, UK

In the Zebrafish Mutation Project (www.sanger.ac.uk/Projects/D_rerio/zmp/), we are generating and phenotyping disruptive mutations in protein-coding genes on a genome-wide scale. While we have identified and archived disruptive mutations in more than half of all zebrafish protein-coding genes using chemical mutagenesis and whole-exome sequencing, we have reached a significant point of diminishing returns and are radically shifting our mutagenesis approach. We are adopting Cas9/CRISPR technology to generate highly multiplexed targeted mutations, which will allow for continued high-throughput phenotype analysis. With more than 10 disruptive mutations in each F2 family, we screen in detail for mutant morphological phenotypes arising in F3 progeny within the first five days post-fertilisation. For any specific mutation, by comparing mutants with wild-type siblings using our differential expression transcript counting technique (DeTCT), we find a wealth of genes displaying alterations in transcript levels, broadly reflecting observed morphological changes. Ontology term enrichment analysis using the gene ontology (GO) annotations combined with the zebrafish anatomical and development (ZFA) ontology has led to surprisingly detailed insights into phenotypes. Thus far two general trends have emerged. Firstly, transcript profiles for previously uncharacterised mutants confirm predicted cellular function and show tissue specific effects on transcript abundance, thus providing mechanistic evidence. Secondly, we are beginning to build pathway specific gene networks. Transcript counting analysis of mutants has revealed novel candidate genes, which lead to a phenotype affecting the same developmental pathway when mutated. Our approach and the results will be discussed in the context of human disease related genes.

Finding Cross-Species Phenomic Similarity through Integration of Heterogeneous Functional Genomic Data

Dr. Elissa Chessler, The Jackson Laboratory, ME

Methods that use phenotype similarity across species to match animal models to human disease rely heavily on face validity of the assays used to define the phenotypes. The resemblance of objectively measured phenotypic characteristics across species is limited by the extent to which the phenotypic inferences supported by these assays are relevant to the disease under investigation and reflect similar characteristics across species. 'Construct validity' is a more important criterion for the matching of phenotypes across species, and to the matching of phenotypes to disease. Construct-valid assays are expected to be associated with similar molecular and other biological characteristics across species, even when the external manifestation of the disease related phenotypes is quite different in humans and model organisms. There is a wealth of relevant data consisting of gene-phenotype associations obtained through high throughput, whole genome experimentation, including genetic mapping, expression correlation, differential expression, systems genetics, mutant screens, proteomic assays and curated functional genomics experiments. To integrate data from diverse studies, gene identifiers are harmonized across various experimental platforms, and through gene homology data are harmonized across diverse species. A variety of statistical and combinatorial approaches may then be applied to match data from various experiments and known gene-disease or gene phenotype associations. This is may be done either agnostically or based on

meta-content describing the disease or phenotypic assay represented by the gene associations. This approach to data driven inference of the relationships among the biological characteristics of animal models, assays and disease features has been implemented in the GeneWeaver.org system, a web service consisting of a database and analytic tools for collaborative integration of functional genomic experiments. This work is supported by AA18776 jointly funded by NIDA and NIAAA.

Session 4. Linking Disease-Relevant Phenotypes with Physiologically Relevant Molecular Pathways and Networks

Chair: ***Dr. Olga Troyanskaya, Princeton University & Simons Center for Data Analysis, NJ***

To discover the molecular basis of human disease, we must effectively and accurately leverage the wealth of data and knowledge from experimental systems and humans to identify genes, pathways, and networks whose malfunction promotes the emergence of clinical disease. This requires sophisticated experimental and computational approaches in human and model organisms, and relies on development of cross-species modeling frameworks to identify genes, pathways, phenotypes, and model organisms that are the most informative for studying each human disease and the drugs that can treat them. This session will provide examples of such approaches and discuss next challenges and exciting directions in using human and model organism data to identify molecular-level models of human disease.

Using Networks to Re-Examine the Genome-Phenome Connection

Dr. John Quackenbush, Dana-Farber Cancer Institute and Harvard TH Chan School of Public Health, MA

The problem with genome-wide association studies (GWAS) is dramatically illustrated in two recent publications. The first analyzed data from 253,288 individuals and found that 697 single nucleotide polymorphisms (SNPs) could explain about 20% of human height variability, but approximately 9,500 SNPs were needed to raise that to 29% . The second surveyed 339,224 individuals and identified 97 loci that can account for 2.7% of body mass index (BMI) variation. These and other similar results leave little hope that using standard GWAS studies, surveying millions of genetic variants across ever larger populations, will lead us to identify the genetic factors driving complex traits. As an alternative, we have developed a revolutionary new way of exploring and exploiting the structure of expression Quantitative Trait Loci (eQTL) networks to explain how weak effect SNPs can combine to drive biological processes and to identify those SNPs most likely to perturb cellular function. As a way of bridging the gap between SNPs and phenotype, we will also explore modeling of gene regulatory networks and methods that can help us model regulation in individuals as well as transitions between phenotypic states.

Human Phenotype Networks

Dr. Jason Moore, Institute for Biomedical Informatics, The Perelman School of Medicine, University of Pennsylvania, PA

Networks are commonly used to represent and analyze large and complex systems of interacting elements. In systems biology, human disease networks show interactions between disorders sharing common genetic background. We present pathway-based human phenotype networks (PHPN) of over 800 physical attributes, diseases, and behavioral traits; based on about 2,300 genes and 1,200 biological pathways. Using GWAS phenotype-to-genes associations, and pathway data from Reactome, we connect human traits based on the common patterns of human biological pathways, detecting more pleiotropic effects, and expanding previous studies from a gene-centric approach to that of shared cell-processes. We also show how these phenotype networks can serve as expert knowledge for subsequent studies of gene-gene interactions.

Understanding the Molecular Basis of Human Disease by Mapping Across Tissues and Organisms
Dr. Olga Troyanskaya, Princeton University & Simons Center for Data Analysis, NJ

A key challenge in biomedical research is to effectively and accurately leverage the wealth of knowledge from research in experimental systems and humans in order to identify what genes, pathways, phenotypes, and model organisms are the most informative for studying each human disease and the drugs that can treat them. Addressing this challenge requires leveraging both the limited and biased, but high-confidence, existing knowledge about molecular processes underlying human disease and the untapped wealth of molecular-level information in genomic data available across multiple species. This relies on sophisticated data integration and analysis methods that can extract functional signals to identify genes and pathways underlying human disease and map genes, processes, phenotypes and diseases across organisms in an unbiased, large-scale way. I will discuss our work in this area, including development of tissue-specific networks to study pathways in specific cell lineage contexts, leveraging them for accurate functional mapping across organisms, and using these approaches to accurately identify genes and pathways underlying complex human disease.

Session 5. Clinical and Experimental Biology Data Integration in the Emerging Field of Precision Medicine
Chair: ***Dr. Yves Lussier, University of Arizona, AZ***

The current approaches to precision therapy proceed incrementally from established genetics when interpreting the genome. Indeed, mutation are straightforwardly identified in protein-coding areas of the genome. The majority of the inherited polymorphisms or acquired mutations are thus uninterpretable as they occur in non-protein codeine areas that cover 97% of the genome. Additionally, each human genome harbors ˜10,000 private polymorphism variants not found in the 1000 genomes or hapmap. Few studies address how biologic systems properties affect human diseases (e.g. protein domain interactions, pathways, etc.). Connecting a personal phenotype to a personal genotype or epigenotype remains highly challenging. Yet, in his seminal Nature paper entitled “Human disease genes” of 2001, David Valle demonstrated that systems properties emerge from the combined analyses of animal genetic models and single-gene diseases. I will briefly present two tables summarizing the

milestones pertaining to combining clinical and experimental biology data for systems and precision therapeutic. The gaps of knowledge and methods will be highlighted in a third slide. I will then introduce three speakers that are addressing these challenges and advancing precision therapeutics via clinical trials, human genetics and experimental biology.

Precision Medicine and the Reclassification of Cancer: Divide and Conquer
Dr. Razelle Kurzrock, University of California San Diego, CA

Cancer remains one of the leading causes of mortality in the world. Precision medicine is however at the threshold of transforming outcomes. Precision therapy implies treatments that "precisely" target the tumor and not the normal elements. It is now increasingly apparent that therapy that is "precise," must also be "personalized," as each patient has a distinct genomic and immune landscape. Hence, the aim of precision/personalized oncology is to customize treatment to the unique molecular and biologic characteristics of each individual and their cancer. Selecting optimal therapy relies first and foremost on a correct diagnosis. Historically, the diagnostic process has been based on light microscopy, which examines the surface of the cell, and identifies the organ of tumor origin. Yet, we now know that cancer is driven by genomic processes, in the setting of a permissive immune system, and that the underlying defects do not necessarily segregate by organ of origin. Fortunately, remarkable technological advances in genomic sequencing and understanding immune cognition, as well as the increasing availability of targeted and immunotherapeutic drugs, are now facilitating precision/personalized therapy. It is now apparent that neoplasms classified uniformly (e.g., non-small cell lung cancer or breast cancer) are actually comprised of a multitude of distinct molecular entities. For instance, tumors bearing ALK alterations make up about 4% of non-small cell lung cancers, and tumors bearing epidermal growth factor receptor (EGFR) mutations, approximately 10%. Importantly, matching patients to therapies targeted against their driver molecular aberrations, or specifically enhancing their immune system, has resulted in remarkable response rates. There is now a wealth of evidence supporting a divide-and-conquer strategy. Indeed, it is evident that advanced tumors have heterogeneous molecular and immune landscapes that mostly differ between patients, and that these tumors may be analogous to "malignant snowflakes." Traditional models of clinical research/practice are drug centered, with a strategy of finding commonalities between patients so that they can be grouped together and treated similarly. However, if each patient with metastatic cancer has a unique molecular and immunological portfolio, a new patient-centered, N-of-one approach that utilizes individually tailored treatment is needed.

Linking Disease Model and Human Phenotypes: The Clinical Geneticist Perspective
Dr. Gail Herman, Nationwide Children's Hospital, Columbus, OH

Improved sequencing technology and bioinformatics capabilities have resulted in dramatic advances in the ability to diagnose genetic diseases. While new gene discovery for rare and common diseases will continue, the domain of the clinical and laboratory geneticist will rapidly move toward understanding the pathogenicity of specific variants in disease, as well as in

healthy individuals via population and carrier screening. Experimental and in silico modeling of specific variants for disease genes of interest will be essential. Current and future efforts in the area of human gene identification and variant classification will be discussed.

"Vertical Integration" around Clinical Problems
Dr. Calum MacRae, Harvard Medical School and Brigham and Women's Hospital, MA

The full realization of Precision medicine will require transformative change in multiple components of the biomedical enterprise including the elucidation of disease mechanisms, ongoing clinical management, drug discovery and translation. The fundamental need for a new and translatable phenotypic lexicon to facilitate this transformation will be discussed and the major hurdles to be overcome in disease modeling and in the clinical arena will be outlined along with some potential solutions.

Session 6. Informatics Tools for Phenotypic Analysis and Data Sharing
Chair: ***Dr. Phil Bourne, NIH, MD***

Genotype-phenotype data is generated, analyzed, and disseminated in a variety of contexts amongst our scientific landscape and throughout the research data life cycle. Tools that work together synergistically across this landscape are essential to support interoperability and downstream data analysis. This session explores a variety of tools aimed to support unique identification of organismal resources, propagation of gene function across taxa, and community standards and tools for assessing genotype-phenotype data quality.

Connecting the Pieces: How to Make a Biomedical Information Ecosystem Run
Dr. Maryann Martone, University of California, San Diego, CA

The last decade has seen major investments into digital infrastructure designed to expand the power and diversity of available research by providing digital access to data, code and knowledge. As we continue to transition slowly from our paper-based system to a fully digital enterprise, we are starting to converge on some sets of basic principles that allow these individual investments to work not in isolation but as part of a connected and networked information ecosystem. We have not solved all of the technical hurdles, nor are all the standards in place, but I believe that a set of best practices and new types of tools are emerging for effective publishing of research objects, including narrative, data, code and workflows. First and foremost, the digital world runs on unique and persistent identifiers. We have different types of entities, e.g., people, concepts, tools and research objects, and different identifier systems for them, but the principles are the same. Anything that enters or is born in the digital world should come with a persistent handle that provides a primary key for finding and aggregating information. In this presentation, I will provide an overview of the issue of identifiers and the Resource Identification Initiative, a partnership between the researchers, informaticians and publishers to develop and implement a simple system for identifying key type of research resources, including genetically modified animals, antibodies and software tools/databases, within the published literature. I will then demonstrate how such identifiers

can anchor new types of information channels across silos of information through a unique technology for annotating web-based objects.

Evolutionary Relationships as a Paradigm for Integrating Biological Knowledge: The Gene Ontology Phylogenetic Annotation Project
Dr. Paul Thomas, University of Southern California, CA

The common ancestry of all living organisms means that discoveries in one biological model organism may shed light on the biology of even its distant relatives. Thus, the evolutionary relationships, or "natural classification" can be used to integrate the knowledge obtained across disparate model organisms. I will discuss how the Gene Ontology Consortium's Phylogenetic Annotation project is integrating information from multiple model organisms, in order to make inferences about the functions of human genes. This integration is done in the context of gene trees and reconstructions of ancestral genomes, and has required development of substantial infrastructure to handle the constant revision of experimental knowledge on the one hand, and of genomes on the other. I will also give examples of how phenotype-based GO annotations are helping to infer shared, as well as diverged, biology among different organisms.

Of Worms and Men: A Data Journey
Dr. Nicole Washington, Lawrence Berkeley National Laboratory, CA

Despite the fact that all living organisms use the same DNA alphabet and that they exhibit both conservation and convergent evolution of biological mechanisms, the way in which we exchange information is not uniform across research specialties or taxa. This fragmented landscape makes it difficult to grasp the collective knowledge that animal models provide in understanding the genetics that underlie disease mechanisms. While we want to allow flexibility in syntactic expression for each domain or taxa to describe the variation in phenotyping and other data, we also need standards to enable cross-cutting analyses and improve integration with clinical data. By standardizing the representation of genotypes and phenotypes across taxa, we are now able uniformly survey the collective knowledge across all organisms and disease models. Furthermore, these standards enable global data sharing to facilitate research, diagnosis, and treatment. Online tools can help support improved specificity of genotypes and their associations with phenotypes. Contextual comparison of phenotype data against this pan-organism knowledgebase can provide an assessment of sufficiency of the phenotyping, as well as reveal coverage and/or gaps in our knowledge and understanding of disease models. Community efforts for data sharing and standardizing genotype-phenotype representation will facilitate discovery, and enable mechanisms to better track data (re)use.

Appendix C. Speaker's Bios

Speakers and Session Chairs

Keynote Speaker

Peter N. Robinson, M.D., M.Sc.
Institute for Medical Genetics, Universitätsklinikum Charité
Berlin, Germany
Email: peter.robinson@charite.de

Peter N. Robinson is a Research Scientist and leader of the Computational Biology Group in the Institute of Medical Genetics and Human Genetics at Charité-Universitätsmedizin Berlin. Dr. Robinson completed his medical education at the University of Pennsylvania, followed by an internship at Yale University. He also studied Mathematics and Computer Science at Columbia University. His research interests involve the use of mathematical and bioinformatics models to understand biology and hereditary disease. In addition to computational biology, his group also does 'wetlab' molecular genetics research in hereditary disease as well as in the molecular mechanisms of fracture healing. His group's output in recent years has included the development of a novel treatment strategy for Marfan Syndrome in mice based on antagonism of a class of bioactive motives that are common in fragments of elastin and fibrillin-1, and the identification of novel disease genes for a form of ataxia (CA8) and hyperphosphatasia with mental retardation syndrome (PIGV).

Dr. Robinson's computational group has developed the Human Phenotype Ontology (HPO), as well as a number of algorithms for disease gene prediction and next-generation sequencing data. A major current focus lies in the development of algorithms for using phenotype and genotype information for diagnostics and computational biology. Dr. Robinson is a member of the Scientific Advisory Board of RD-Connect.

Olivier Bodenreider, M.D., Ph.D.
Branch Chief, Cognitive Science Branch
Lister Hill National Center for Biomedical Communications
U.S. National Library of Medicine, National Institutes of Health (NIH)
Email: obodenreider@mail.nih.gov

Dr. Bodenreider is a Fellow of the American College of Medical Informatics. He received an M.D. degree from the University of Strasbourg, France, and a Ph.D. in Medical Informatics from the University of Nancy, France. Before joining NLM, he was Assistant Professor for Biostatistics and Medical Informatics at the University of Nancy, France, Medical School.

His research focuses on terminology and ontology in the biomedical domain, both from a theoretical perspective (quality assurance, interoperability) and in their application to natural language processing, knowledge discovery, and information integration. He has investigated the representation of phenotypes in standard terminologies (e.g., SNOMED CT), which have paved the way for integrating HPO into the Unified Medical Language System (UMLS). He has also developed methods for extending the coverage of phenotype terms in standard terminologies through post-coordination and partial mappings. Dr. Bodenreider is a Fellow of the American College of Medical Informatics.

Philip E. Bourne, Ph.D.
Associate Director for Data Science, NIH
Email: philip.bourne@nih.gov

Dr. Bourne was trained as a physical chemist and obtained his Ph.D. at The Flinders University of South Australia. He then moved to the University of Sheffield to do postdoctoral research, followed by a move to Columbia University and eventually to the University of California, San Diego, where he was a Professor in the Department of Pharmacology. In 2014, he moved to the National Institutes of Health to become its Associate Director for Data Science. Dr. Bourne is co-developer of the Combinatorial Extension algorithm for the three-dimensional alignment of protein structures and co-director of the Protein Data Bank. Dr. Bourne's professional interests focus on relevant biological and educational outcomes derived from computation and scholarly communication. This work involves the use of algorithms, text mining, machine learning, metalanguages, biological databases, and visualization applied to problems in systems pharmacology, evolution, cell signaling, apoptosis, immunology, and scientific dissemination. He has published over 300 papers and five books. One area to which he is extremely committed is to furthering the free dissemination of science through new models of publishing and better integration and subsequent dissemination of data and results. He is the co-founder and founding Editor-in-Chief of the *open access* journal *PLOS Computational Biology*. Dr. Bourne is a Past President of the International Society for Computational Biology, an elected fellow of the American Association for the Advancement of Science (AAAS), the International Society for Computational Biology (ISCB) and the American Medical Informatics Association (AMIA).

Erdem Bangi, Ph.D.
Senior Scientist, Center for Personalized Cancer Therapeutics
Icahn School of Medicine at Mount Sinai
Email: erdem.bangi@mssm.edu

Dr. Erdem Bangi received his Ph.D. from Brown University, where he took a classical developmental genetics approach to study cell-cell signaling and epithelial patterning using the fruit fly Drosophila. He then moved to Novartis Institutes for Biomedical Research as a postdoctoral fellow to explore new applications for Drosophila in drug discovery. He continued developing this approach further in Dr. Ross Cagan's laboratory, where he took advantage of a century of powerful genetic tools to build a collection of fly models that reflect the genetic complexity and diversity of human colorectal cancer genomes from the Cancer Genome Atlas (TCGA). This collection proved to be an invaluable resource to explore mechanisms of drug response and resistance in a diverse set of genetically complex models in a whole animal setting. He is currently the lead scientist at the Center for Personalized Cancer Therapeutics (CPCT), whose goal is to identify cancer treatments based on a patient's personal cancer genome. The CPCT's approach is to build a personalized fly model for each patient based on the genomic profile of their tumor and use these patient specific models to screen FDA approved drug libraries and identify drug combinations specifically tailored to each patient.

Elissa Chesler, Ph.D.
Associate Professor, The Jackson Laboratory
Bar Harbor, ME
Email: Elissa.Chesler@jax.org

Dr. Chesler began her scientific career in Psychology and Behavioral Neuroscience. She was an M.D./Ph.D. student at the University of Illinois and completed her Ph.D. work in the Psychology Department on neuroendocrine effects on brain and behavior. Dr. Chesler did her post-doctoral work at the University of Tennessee, where she made use of mouse strains to find genes that influence behavior and gene expression in the brain. This work was an early example of a new synthesis of systems biology to genetics, now called "systems genetics". She also became involved early on with the Collaborative Cross (CC) mouse project, which sought to develop better strains of mice for more precise complex trait research. Dr. Chesler's current work involves analyzing massive data sets to try to tease apart what genes and gene networks are associated with behaviors and to effectively correlate mouse behavior in labs with human behaviors. To this end, she has developed GeneWeaver.org in close collaboration with computer scientists at the University of Tennessee and Baylor University. This system allows users to integrate heterogeneous phenotype-centered gene sets across species, tissues, and experimental platforms. Her laboratory integrates quantitative genetics, bioinformatics and behavioral science to understand and identify the biological basis for the relationships among behavioral traits. Dr. Chesler develops and applies cross-species genomic data integration, advanced computing methods, and novel high-precision mouse populations to find genes associated with a constellation of behavioral disorders. This integrative strategy enables to relate mouse behavior to specific aspects of human

disorders, to test the validity of behavioral classification schemes, and to find genes and genetic variants that influence behavior.

Rex L Chisholm, Ph.D.
Vice Dean Scientific Affairs and Graduate Education
Adam and Richard T. Lind Professor of Medical Genetics
Professor in Cell and Molecular Biology,Center for Genetic Medicine and Surgery
Northwestern University Feinberg School of Medicine, Chicago, IL
Email: r-chisholm@northwestern.edu

Dr. Chisholm graduated from the Department of Microbiology and Immunology at the University of Michigan Medical School. He did postdoctoral work in the Department of Biology at MIT and the Whitehead Institute, focusing on the molecular genetics of the cellular slime mold Dictyostelium discoideum. As a faculty member at Northwestern University in the Department of Cell Biology and Anatomy, he studied the role of molecular motors in cell migration. In 2000, Dr. Chisholm became the founding Director of the Center for Genetic Medicine. At that time he had begun to focus on the areas of bioinformatics and understanding the contribution of genetic variation to human disease, especially common, complex disease. Dr. Chisholm currently serves as PI of dictyBase, the model organism database for Dictyostelium. He is also PI of the NUgene project, a biobank linked to the electronic health records of consented participants who receive their health care through the Feinberg School's clinical affiliates: Northwestern Memorial Hospital, Northwestern Medical Faculty Foundation and Children's Memorial Hospital. NUgene is also a member of the NHGRI (National Human Genome Research Institute - one of the institutes of the National Institutes of Health) eMERGE network. The eMERGE network is a consortium of biobanks that are linked to electronic health records. He is also an active participant in the Gene Ontology Consortium.

Janan T. Eppig, Ph.D.
Professor, The Jackson Laboratory, Bar Harbor, ME
Email: jte@informatics.jax.org

Janan Eppig graduated from the University of Washington in Seattle, Washington, and received her Ph.D. from the University of Maine in Orono, ME. She is currently a Professor at The Jackson Laboratory, Bar Harbor, ME. Dr. Eppig's interests include comparative genomics, genome organization, model systems for human diseases and cancers, bioinformatics, and the development of database resources and semantic standards for annotation and data sharing. She plays a very active role in developing integration of databases of mouse genetic, genomic and biological data and currently is a PI of the Mouse Genome Informatics Database (MGI). MGI is used by the international scientific community as its primary resource for mouse information and as a tool for new biological discovery. The database contains a wide variety of data pertaining to genes, their DNA and protein sequences, and the phenotypes that result from mutations in different genes. The three central components of MGI are the Mouse Genome Database (MGD), an internationally recognized database for the laboratory mouse; the Mouse Tumor Biology (MTB) database, which facilitates the selection of experimental models for cancer research; and the International Mouse Strain Resource (IMSR), a searchable online database cataloging mouse stocks available worldwide. The database continues to expand to keep abreast of new

technologies and to grow with our expanding knowledge of how the genetic blueprint of DNA manifests in traits of a living individual. Dr. Eppig chairs the International Committee for Standardized Genetic Nomenclature for Mice that develops guidelines for the naming of genes, mutations, strains, and genome features and is a member of the HGNC (Human Genome Nomenclature Committee) International Advisory Committee. She is an initiating member of the Gene Ontology (GO) Consortium. Dr. Eppig has served on many Advisory Boards and Review Panels for bioinformatics and database resources in North America and Europe. Current Advisory Board appointments include the Drosophilia Genome Database (FlyBase), the European Mouse Mutagenesis Consortium tools for functional annotation of the mouse genome (EUCOMMtools), and the European infrastructure for phenotyping and archiving of model mammalian genomes (Infrafrontier-I3).

Melissa Haendel, Ph.D.
Associate Professor
Oregon Health & Science University Library
Department of Medical Informatics and Clinical Epidemiology
Oregon Health & Science University
Email: haendel@ohsu.edu

Dr. Haendel has a B.A. from Reed College in Chemistry and a Ph.D. in Neuroscience from the University of Wisconsin, Madison. She was trained in Molecular and Developmental Biology, using chick, mouse, and zebrafish model systems. She is currently the basic research PI of the Monarch Initiative, with the aim of providing integrated access to human and model systems genotype-phenotype data for the purposes of disease hypothesis exploration. Dr. Haendel led zebrafish genome nomenclature and ontology interoperability efforts for the Zebrafish Model Organism Database (ZFIN). More recently, she has been leading efforts to assess reproducibility relating to specification of model systems in the literature. She also participates in development of eagle-i and VIVO, designed to collect and disseminate information about biomedical resources and enable research profiling, and to promote collaboration across translational boundaries. Her research interests are in using ontologies to promote synthetic science though connections within biomedical data, to utilize information science during the course of research and its publication, to promote team science, and to enable scientific reproducibility.

Carol M. Hamilton, Ph.D.
PhenX Principal Investigator, Director of Bioinformatics
Research Computing Division, RTI International, Research Triangle Park, NC
Email: chamilton@rti.org

Dr. Hamilton earned a B.S. in Botany at the University of California, Davis, and a Ph.D. in Genetics from the University of Georgia. She has experience in biochemistry, molecular biology, technology development, data management, and analysis of genomic and clinical data. She has been involved in the development of bioinformatics and analytic systems for a variety of technologies, including DNA sequencing, RNA expression profiling, metabolic profiling, and phenotype profiling. Research areas have included biomarker discovery for monitoring drug safety, efficacy, and target identification. Key areas of interest are the visualization and analysis of complex data sets that include genotype, phenotype (including "-omics") and environmental exposures data. Dr. Hamilton is currently the Principal Investigator of PhenX, a cooperative agreement funded by the National Human Genome Research Institute (NHGRI) of

the National Institutes of Health (NIH). Led by RTI International, and driven by the scientific community, PhenX has established a toolkit of standard measures for use in genome-wide association studies (GWAS) and other studies involving human subjects.

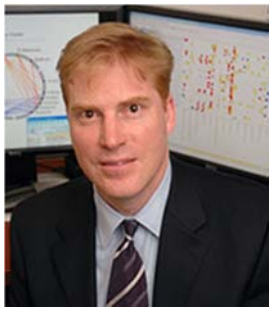
Gail E. Herman, M.D., Ph.D.
King Fahd Professor of Molecular Medicine
Department of Medicine, and Department of Oncology and the Department of Molecular Biology and Genetics
The Research Institute at Nationwide Children's Hospital, Columbus, OH
Email: Gail.Herman@nationwidechildrens.org

Dr. Herman received her medical degree and a Ph.D. in Biochemistry from Duke University and completed a residency in Pediatrics and a fellowship in Genetics at the Baylor College of Medicine. She is board-certified in Pediatrics and Clinical and Biochemical Genetics. Her research focuses on X-linked developmental disorders. Current NIH-funded research focuses on mouse models for cholesterol biosynthesis disorders. A new research focus involves translational work with the development of the Central Ohio Registry for Autism (CORA) and genetics evaluation of children with autism funded through the Department of Defense. Through the registry and clinical genetics evaluations of children with autism, Dr. Herman and colleagues have developed guidelines for genetic testing for newly diagnosed patients, including the sequencing of the PTEN tumor suppressor gene in patients with autism or developmental delay and macrocephaly. They are performing whole exome sequencing on selected registry families to identify de novo pathogenic variants, as well as second site modifiers in children with an identifiable primary mutation. Association studies in trios in Nationwide Children's Hospital registry using polymorphic functional variants in genes involved in relevant neurotransmitter pathways are under the way.

Razelle Kurzrock, M.D.
Chief, Division of Hematology and Oncology
Murray Professor of Medicine, Senior Deputy Director, Clinical Science Director, University of California, San Diego
Moores Cancer Center, La Jolla, CA
Email: rkurzrock@ucsd.edu

Dr. Kurzrock received her M.D. from the University of Toronto, Canada. Dr. Kurzrock joined the University of California, San Diego Moores Cancer Center in 2012, as Senior Deputy Center Director for Clinical Science. She is also the Murray Professor of Medicine, Director of the Clinical Trials Office and, in 2014, became the Chief of the Division of Hematology-Oncology Division (in the University of California, San Diego School of Medicine). Dr. Kurzrock's charge includes growing and innovating the clinical trials program, and heading the newly established Center for Personalized Cancer Therapy and the UCSD Moores Cancer Center Clinical Trials Office. As a Physician-Scientist, she brings extraordinary expertise and experience in clinical research, business operations, regulatory operations, financial and budget planning, and administrative oversight, in addition to her world-recognized work in translational science. Dr. Kurzrock is best known for successfully creating and chairing the largest Phase I clinical trials department in the world while at the University of Texas M.D. Anderson Cancer Center. A central theme of that program was the personalized medicine strategy, embodied in a protocol called PREDICT (Profile-Related Evidence Determining Individualized Cancer Therapy). Dr. Kurzrock's unique approach

emphasizes using cutting-edge molecular profiling technologies to match patients with novel targeted therapies, reflecting a personalized strategy to optimize cancer treatment.



Yves A. Lussier, M.D., FAMCI
Professor of Medicine, Associate Vice President for Health Sciences Associate Director, BIO5 Informatics
Cancer Center at The University of Arizona
yves@email.arizona.edu

Dr. Lussier received a Bachelor of Engineering and his medical degree from the University of Sherbrooke, Quebec, Canada. He performed predoctoral research in the Departments of Medicine and Human Physiology at the University of Sherbrooke. After medical school, Dr. Lussier completed an internship in Ophthalmology at Laval University Hospital in Quebec City, and a residency in Family Medicine at the University of Sherbrooke Medical Center. He was a post-doctoral residential fellow in the Department of Biomedical Informatics in the College of Surgeons & Physicians at Columbia University. Dr. Lussier is an international expert in translational bioinformatics and a pioneer in research informatics techniques including systems biology, data representation through ontologies and high-throughput methods in personalized medicine. At the University of Arizona, he will lead efforts to fully develop novel programs in biomedical informatics, computational genomics and precision health. Dr. Lussier will provide critical leadership in efforts to advance precision health approaches to health outcomes and healthcare delivery and in the development of big data analytical tools and resource services in support of the University's clinical research and service missions.

Dr. Lussier's research interests focus on the use of ontologies, knowledge technologies, and genomic network models to accurately individualize the treatment of disease and to repurpose therapies. He has National Institutes of Health funding for a clinical trial that repositioned a combination therapy. He also bioinformatically predicted and obtained biological confirmation of several novel tumor suppressor microRNAs, including the first one underpinning the oligo- vs poly- metastasis development of cancer. A Fellow of the American College of Medical Informatics, Dr. Lussier is a member of numerous governance, technology transfer, scientific and editorial boards, including the American Medical Informatics Association, International Society for Computational Biology, Society for Clinical and Translational Science, American Society for Cancer Research, Healthcare Information and Management Systems Society, American Association of Pharmaceutical Scientists, American Association for the Advancement of Science and American Society for Human Genetics.

Calum A. MacRae, M.D., Ph.D.
Chief, Cardiovascular Medicine, Brigham and Women's Hospital
Associate Professor of Medicine, Harvard Medical School
Email: cmacrae@partners.org

Dr. Calum A. MacRae is a Geneticist, Developmental Biologist and a Cardiologist at Brigham and Women's Hospital and Harvard Medical School. He is an Associate Member at the Broad Institute and a Principal Faculty Member at the Harvard Stem Cell Institute. His research is focused on understanding the fundamental mechanisms of cardiovascular disease using human studies and complementary efforts in systems modeling with empiric high-throughput biology in the zebrafish. His lab is using automated screens in fish to define the

genetic architecture of disease and to explore gene-drug (environment) interactions through the interrogation of large-scale chemical libraries. His clinical interests include the management of inherited heart disease and cardiac involvement in systemic diseases. Dr. MacRae also is the Director of the Cardiology Fellowship Program and is responsible for Physician Scientist training initiatives at the Cardiovascular Research Center.

Maryann E. Martone, Ph.D.
Co-Director
National Center for Microscopy and Imaging Research (NCMIR)
Professor in Residence, Department of Neurosciences
University of California, San Diego, La Jolla, CA
Email: maryann@ncmir.ucsd.edu

Dr. Martone received her B.A. from Wellesley College in Biological Psychology and her Ph.D. in Neuroscience from the University of California, San Diego (UCSD). She joined the Department of Neurosciences at UCSD in 1993, where she is currently a Professor in residence. She is the Principal Investigator of the Neuroscience Information Framework project (NIF), a national project to establish a uniform resource description framework for neuroscience. Her recent work has focused on building ontologies for neuroscience for data integration. She recently finished her tenure as the U.S. Scientific Representative to the International Neuroinformatics Coordinating Facility (INCF), an international organization dedicated to developing tools and standards for neuroscience data exchange. Dr. Martone is a practicing neuroscientist, with expertise in neuroanatomy, and light and electron microscopy. For the past decade, she has been working in the area of neuroinformatics to increase access to and utilization of neuroscience data. To further develop the framework, she heads the Ontology Development Program for the INCF and the Data Standards Workstream for the newly launched One Mind for Research campaign. Through NIF and her neuroscience background, Dr. Martone has a unique global perspective on issues in data sharing and utilization in the neurosciences and has gained considerable insight and expertise in working with diverse biomedical data. She has also continued to explore how these knowledge frameworks can be used to solve difficult problems in neurodegenerative disease through modeling of structural phenotypes in animal models of human neurodegenerative conditions.

Jason H. Moore, Ph.D.
Edward Rose Professor of Informatics
Director, Institute for Biomedical Informatics, Division of Informatics
Department of Biostatistics and Epidemiology
The Perelman School of Medicine, University of Pennsylvania, PA
Email: jhmoore@exchange.upenn.edu

Dr. Moore, is a Translational Bioinformatics Scientist and Human Geneticist. He graduated from Florida State University (B.S.) and University of Michigan (M.S., M.A.). He finished his Ph.D. at the University of Michigan. Dr. Moore was an Ingram Associate Professor of Cancer Research and a member of the Center for Human Genetics Research at Vanderbilt University before joining the faculty at The Geisel School of Medicine at Dartmouth in 2004. He was elected a Fellow of the American Association for the Advancement of Science (AAAS) in 2011. Moore's research focuses on the development and application of informatics methods for identifying

combinations of DNA sequence variations and environmental factors that are predictive of human health and complex disease. Along the way, he pioneered the theory and formalisms of the multifactor dimensionality reduction (MDR) machine learning method for detecting and characterizing combinations of attributes or independent variables that interact to influence a dependent or class variable. He then applied MDR for improved understanding of the interplay of multiple genetic polymorphisms of complex traits in genome-wide association studies. Of note, his seminal MDR work has been cited > 900 times and this led the scientific community to publish over 300 manuscripts in the nascent MDR field. He is the Edward Rose Professor of Informatics and the Director of the Institute for Biomedical Informatics, Division of Informatics, Department of Biostatistics and Epidemiology, and Senior Associate Dean for Informatics at the Perelman School of Medicine, University of Pennsylvania. Dr. Moore is the founding Editor-in-Chief of the journal, *BioData Mining*.

Mary C. Mullins, Ph.D.
Professor of Cell and Developmental Biology, Perelman School of Medicine
University of Pennsylvania, Philadelphia, PA
Email: mullins@mail.med.upenn.edu

Dr. Mullins received her B.S. from the University of Wisconsin, Madison and her Ph. D. from the University of California, Berkeley. She did her postdoctoral training at the Max Planck Institute, Tübingen, Germany with Christiane Nüsslein. Dr. Mullins is an expert in the field of Embryonic Development and Oogenesis in the zebrafish and her laboratory at the Perelman School of Medicine is studying the molecular mechanisms by which a BMP (Bone Morphogenetic Protein) signal transduction pathway establishes different aspects of the vertebrate body plan. Various zebrafish mutants of BMP pathway components, as well as antisense knockdown approaches, are used to dissect the molecular mechanisms by which this pathway establishes different cell types. The interest is in the formation, function, and temporal regulation of a BMP activity gradient, which is implicated in specification of diverse cell types along the dorsal-ventral axis. She has shown that this gradient is essential in neural crest specification and is linked to dorsal-ventral patterning of neural tissue. Moreover, a subset of defined components is also function in post-embryonic heart development. Misregulation of BMP signaling leads to a debilitating disease in humans and laboratory is trying to establish a model for this disease in the zebrafish. In addition, to study maternally-controlled processes Dr. Mullins performed a large-scale maternal-effect mutant screen, not previously performed in a vertebrate, to identify mutants of key genes specifically required in the mother for oocyte development, egg activation, fertilization, the midblastula transition, and establishment of the axes of the vertebrate embryo. Numerous mutants in these processes were obtained and currently the molecular and cellular basis for the defects is under investigation.

Chris Mungall, Ph.D.
Bioinformatics Scientist, Berkeley Bioinformatics Open-source Projects
Lawrence Berkeley National Laboratory, Berkeley, CA
Email: cjmungall@lbl.gov

Dr. Mungall is currently a PI of the Monarch Initiative. He has a B.Sc. in Computer Science from the University of Edinburgh and a Ph.D. from the University of Edinburgh in Biological Sciences. His current research focuses on the use of informatics techniques and tools to integrate and interpret life science data across a variety of domains. In particular, he is interested in

connecting biological models at multiple scales: the biochemical, molecular, cellular, organismal and ecological. He is the creator of the OWLSim algorithm, which allows for the computation of the similarity of two organisms in phenotype space. He developed the OBD database and reasoning system which is geared towards phenotype-based search. The methods are described in the paper Linking Human Diseases to Animal Models using Ontology-based Phenotype Annotation, and OWLSim analyses are used in tools such as mousefinder. In a collaboration with the Neurosciences Information Framework (NIF) project, he developed a Phenotype Knowledge Base (PKB) system that allows neurodegenerative diseases to be automatically matched to model systems based on phenotypes in common. The system integrates phenotypes that are manifest at different scales, from the molecular up through the cellular to the level of gross neuroanatomy. The Gene Ontology Project was established to systematize this knowledge to allow for automated computational inference - for example, predicting the function of human genes based on phylogeny, or interpreting the expression patterns of genes regulated in diseases. Dr. Mungall currently manages the Gene Ontology software group, which produces software, resources and standards such as AmiGO, the GO database, OBO-Edit, TermGenie, the GO Galaxy Environment, and the obo-format specification. One of the long term research goals of Dr. Mungall is to render the bulk of biological and medical knowledge computable, allowing for a new generation of intelligent bioinformatics tools. He is a Co-founder and Coordinating Editor of the Open Bio-Ontologies Foundry, which was initiated to further this goal.

John Quackenbush, Ph.D.
Professor, Department of Biostatistics, Harvard School Of Public Health
Biostatistics and Computational Biology, Dana-Farber Cancer Institute
Boston, MA
Email: johnq@jimmy.harvard.edu

Dr. Quackenbush attended the California Institute of Technology, where he earned a Bachelor's degree in Physics. He went on to earn a doctorate in Theoretical Particle Physics from the University of California, Los Angeles. After working two years as a postdoctoral fellow in Physics, Dr. Quackenbush was awarded a Special Emphasis Research Career Award from the National Center for Human Genome, and subsequently spent the next two years at the Salk Institute working on physical maps of human chromosome 11. He was a faculty member of The Institute for Genomic Research (TIGR) in Rockville, Maryland, developing analytical methods based on the integration of data across domains to derive biological meaning from high-dimensional data. In 2005, Dr. Quackenbush was appointed to his current positions at the Dana-Farber Cancer Institute (DFFI) and the Harvard School of Public Health. Four years later, he launched the DFCI's Center for Cancer Computational Biology (CCCB), which he directs and which provides broad-based bioinformatics and computational biology support to the research community through a collaborative consulting model, and which also performs and analyzes large-scale second-generation DNA sequencing. A leader in the fields of Genomics and Computational Biology, Dr. Quackenbush's current research focuses on the analysis of human cancer using systems biology-based approaches to understanding and modeling the biological networks that underlie disease.

Rachel Richesson, Ph.D., M.P.H., FACMI
Associate Professor, Duke University School of Nursing
Durham, NC
Email: rachel.richesson@dm.duke.edu

Dr. Richesson is an Associate Professor of Informatics at the Duke University of Nursing. She is particularly interested in applications and standards specifications that increase the efficiency of clinical research and enable interoperability between clinical research and healthcare information systems. She co-leads the Phenotyping, Data Standards, and Data Quality Core for the NIH Health Care Systems Research Collaboratory, a demonstration program for the transformation of clinical trials based upon use of Electronic Health Records (EHRs) and healthcare systems partnerships. In this role, she is developing standard approaches and guidance for using computable phenotypes in the extraction of clinical data to support research and learning healthcare. She is also the co-lead of the Rare Diseases Task Force for the nationally distributed Patient Centered Outcomes Research Network (PCORnet), specifically promoting standardized computable phenotype definitions for rare diseases, and helping to develop a national research infrastructure that can support observational and interventional research for various types of conditions. Prior to joining Duke University in 2011, she provided informatics and data standards support for the NIH Rare Diseases Clinical Research Network as part of the Data Coordinating Center at the University of South Florida. She edited the first textbook on the topic of Clinical Research Informatics (Springer, 2012) and was inducted as a Fellow of the College of Medical Informatics in 2014.

Damian Smedley, Ph.D.
Senior Scientific Manager, Mouse Informatics, Wellcome Trust Sanger Institute
Cambridge, United Kingdom
Email: ds5@sanger.ac.uk

Dr. Smedley leads development of clinical software tools for the Monarch Initiative, utilizing phenotypic comparison of patient and model organism phenotypes for diagnosis and disease gene discovery as part of projects such as the National Institutes of Health's (NIH) Undiagnosed Disease Program. As part of the NIH KOMP2 Program, his team identifies new disease models and gene candidates arising from the effort of the International Mouse Phenotyping Consortium. His current research focuses on how best we can use model organism phenotype data to better understand the role of gene mutation in human disease. He has a B.Sc. in Biochemistry from the University of Bristol and a Ph.D. from the University of Cambridge, also in Biochemistry, studying the effect of mutation on the structure and function of a family of bacterial proteins. His postdoctoral research at the Institute for Cancer Research, London, studied the role of large-scale chromosomal changes in cancer patients and led to the identification and characterisation of a novel FGFR1-ZNF198 fusion protein associated with a mixed leukaemia/lymphoma syndrome. He switched to a purely computational approach after completing an M.Sc. in Bioinformatics from Birbeck, University of London and, after working on candidate gene selection in type 2 Diabetes GWAS regions at Imperial College London, joined the Ensembl Group at the European Bioinformatics Institute (EBI) as the main developer of the BioMart distributed data integration solution. Following this, he set up the Mouse Informatics Group at the EBI that plays a key

role in international mouse distribution and phenotyping programs, before joining the Wellcome Trust Sanger Institute in his current position.

Derek Stemple, Ph.D.
Head of Mouse and Zebrafish Genetics
The Wellcome Trust Sanger Institute, Cambridge, United Kingdom
Email: ds4@sanger.ac.uk

Dr. Stemple obtained his first degrees from the University of Colorado Boulder, in Applied Mathematics (B.S.) and Molecular, Cellular and Developmental Biology (B.A.). As a Postgraduate he worked for several years studying microtubule dynamics and mitosis with Professor J. Richard McIntosh. He began his Ph.D. studies in Neurobiology at the California Institute of Technology under the supervision of Professor David Anderson, which concluded with his discovery of the mammalian neural crest stem cell. As a Helen Hay Whitney postdoctoral fellow working at the Massachusetts General Hospital, Dr. Stemple participated in a large-scale systematic screen for mutations affecting embryogenesis in. At the National Institute for Medical Research in London, Dr. Stemple's group identified several genes important for development of the notochord. Currently, Dr. Stemple is a Senior Investigator at the Wellcome Trust Sanger Institute, where his group studies the genetics of vertebrate early development, skeletal sarcomere formation and muscle integrity as well as zebrafish genomics. His group is also involved with two major Sanger Institute projects, the Zebrafish Mutation Resource and the Zebrafish Genome Sequencing Project.

Paul D. Thomas, Ph.D.
Associate Professor in the Preventive Medicine Department and Molecular and Computational Biology Program
Director of Bioinformatics Division
University of Southern California, Los Angeles
Email: pdthomas@usc.edu

Dr. Thomas has a B.A. degree from the University of California, Los Angeles and a Ph.D. from the University of California, San Francisco. While at Celera Genomics, Dr. Thomas contributed to the first analysis of the human genome (published in *Science Magazine*), co-writing (with Mani Subramanian) the 10-page overview of the function and evolution of human genes. Dr. Thomas is also known for the development of the PANTHER (Protein Analysis through Evolutionary Relationships) web server, which is used by a broad community of researchers seeking information about gene function and evolution. Dr. Thomas's own research lab focuses on the development and application of computational methods for reconstructing gene evolution, and using these techniques to understand the function of human genes, and how genetic factors may impact disease risk. Dr. Thomas is a Principal Investigator for the Gene Ontology Project, which is among the world's largest bioinformatics projects. It is now possible to perform "Genomics" experiments in which, for example, variations in all 20,000 human genes are determined for thousands of different individuals, and compared between those affected by a particular disease, and those unaffected. The Project is exploring how ontologies can be used in an informed fashion, to help interpret the results of genome-wide association studies (GWAS) in humans. The Project is also developing a genomics data and analysis resource for large-scale experiments with the

well-studied lab strains of *E. coli*. Dr. Thomas is also a PI for the PortEco project, designed to help the research community get the most out of E. coli as a biological "model organism."

Olga Troyanskaya, Ph.D.
Professor, Department of Computer Science and Lewis-Sigler Institute of Integrative Genomics, Princeton University, Deputy Director for Genomics, Simons Center for Data Analysis, Simons Foundation, NY
Email: ogt@genomics.princeton.edu

Olga Troyanskaya has a B.S. from the University of Richmond, Richmond, Virginia and a Ph.D. in Biomedical Informatics from Stanford University, Stanford, California. The goal of her research is to bring the capabilities of computer science and statistics to the study of gene function and regulation in the biological networks through integrated analysis of biological data from diverse data sources--both existing and yet to come (e.g. from diverse gene expression data sets and proteomic studies). She is designing systematic and accurate computational and statistical algorithms for biological signal detection in high-throughput data sets. More specifically, Dr. Troyanskaya is interested in developing methods for better gene expression data processing and algorithms for integrated analysis of biological data from multiple genomic data sets and different types of data sources (e.g. genomic sequences, gene expression, and proteomics data).

Dr. Troyanskaya's laboratory combines computational methods with an experimental component in a unified effort to develop comprehensive descriptions of genetic systems of cellular controls, including those whose malfunctioning becomes the basis of genetic disorders, such as cancer, and others whose failure might produce developmental defects in model systems. The experimental component the lab focuses on is S. cerevisiae (baker's yeast).

Nicole Washington, Ph.D.
Research Scientist at Genomics Division
Lawrence Berkeley National Laboratory, Berkeley, CA
Email: NLWashington@lbl.gov

Dr. Washington has a B.S. in Biology from Harvey Mudd College and a Ph.D. in Molecular and Cellular Biology from The University of Arizona, where she focused on the study of the *C. elegans* as a model for human disease. She is focusing on the intersection of computers with biological research, in particular making tools to realize hidden connections in diverse types of biological data and integrating large-scale experiments. She participates in the following projects: Monarch Initiative, The modENCODE Data Coordination Center, Phenotype Annotation using Ontologies. The Monarch Initiative aims to provide easy-to-use tools and services to enable navigation of model systems data by semantically aggregating this data and making it queryable based on a number of facets, such as phenotypic similarity, network analysis, gene expression and function, and genomics.

Caleb Webber, Ph.D.
Program Leader, Neurological Disease Genomics
MRC Functional Genomics Unit, Department of Physiology, Anatomy and Genetics, Oxford University, Oxford, United Kingdom
Email: caleb.webber@dpag.ox.ac.uk

Dr. Weber obtained a Ph.D. from the European Bioinformatics Institute, The Wellcome Trust Genome Campus, Hinxton Cambridge, and from the Department of Genetics, Cambridge University. Afterwards, he returned to Oxford University to work with Professor Chris Ponting on several major large-scale genome projects of the last decade. His interest in Synteny breaks from those projects led to an interest in copy number variation, and in turn to the role of genetic variation in disease. His group is gaining insights in to complex neurodevelopmental and neuropsychiatric disorders using functional and integrative genomics. He exploits recent large-scale genetic data sets to identify significant molecular features that can help identify which genes contribute to these complex disorders. In particular, his group has applied the phenotypic-associations made by disrupting genes in the mouse ("mouse knockouts") as a novel large-scale functional genomics resource. In a proof-of-concept publication, they showed how significant biases could be detected among the set of mouse phenotypes associated with those human genes affected by mutations in patients with intellectual disability. This approach is now extended using integrative genomics to incorporate annotations from gene expression, protein-protein interactions and other resources thereby creating functional linkage networks of relevance to particular disorders. Dr. Weber's group collaborates with many European and International partnerships, especially through the Genetics of Cognitive Dysfunction (Gencodys) Consortium and through the IMI StemBANCC consortium, where he leads the Data Interpretation package.

Appendix D. Participant List

Darrell Abernethy, Ph.D.
Associate Director for Drug Safety
Office of Clinical Pharmacology
Food and Drug Administration
10903 New Hampshire Avenue
Silver Spring, MD 20993
Telephone: (301) 796-3719
Email: darrell.abernethy@fda.hhs.gov

Robert Adelstein, M.D.
Principal Investigator
Genetics and Developmental Biology Center
National Heart, Lung, and Blood Institute
National Institutes of Health
Building 10, Room 6C-103B
10 Center Drive, MSC 1583
Bethesda, MD 20892-1583
Telephone: (301) 496-1865
Fax: (301) 402-1542
Email: robert.adelstein@nih.gov

Beena Akolkar, Ph.D.
Senior Advisor, Research Programs
National Institute of Diabetes and Digestive and Kidney Diseases
National Institutes of Health
Two Democracy Plaza, Room 6105
6707 Democracy Boulevard, MSC 5460
Bethesda, MD 20892-5460
Telephone: (301) 594-8812
Email: akolkarb@mail.nih.gov

Joanna Amberger
Program Manager
Online Mendelian Inheritance in Man
The Johns Hopkins University School of Nursing
242 Garland Hall
3400 N. Charles Street
Baltimore, MD 21218
Telephone: (410) 955-0313
Email: joanna@peas-welch.jhu.edu

Carl Baker, Ph.D., M.D.
Program Director
Division of Skin and Rheumatic Diseases
National Institute of Arthritis and Musculoskeletal and Skin Diseases
National Institutes of Health
One Democracy Plaza, Room 892
6701 Democracy Boulevard, MSC 4872
Bethesda, MD 20892-4872
Telephone: (301) 435-1240
Email: bakerc@mail.nih.gov

Erdem Bangi, Ph.D., B.Sc.
Senior Associate Dean for the Graduate School of Biomedical Sciences
Professor, Department of Developmental and Regenerative Biology
The Icahn School of Medicine at Mount Sinai
Annenberg Building, 25th Floor, Rooms 25-40
1468 Madison Avenue
New York, NY 10029
Telephone: (212) 241-0135
Fax: (212) 860-9279
Email: erdem.bangi@mssm.edu

Zhirong Bao, Ph.D.
Associate Member
Developmental Biology Program
Memorial Sloan Kettering Cancer Center
1275 York Avenue
New York, NY 10065
Telephone: (646) 636-6027
Email: baoz@mskcc.org

Judith Blake, Ph.D.
Professor
Bioinformatics and Computational Biology
The Jackson Laboratory
600 Main Street
Bar Harbor, ME 04609
Telephone: (207) 288-6248
Email: judith.blake@jax.org

Olivier Bodenreider, M.D., Ph.D.
Branch Chief, Cognitive Science Branch
Lister Hill National Center for Biomedical Communications
U.S. National Library of Medicine
National Institutes of Health
Building 38A, Room 09S904
8600 Rockville Pike
Bethesda, MD 20892
Telephone: (301) 435-3246

Email: obodenreider@mail.nih.gov

Philip Bourne, Ph.D.
Associate Director for Data Science
Office of the Director
National Institutes of Health
Building 1, Room 228
1 Center Drive
Bethesda, MD 20892
Telephone: (301) 496-0786
Email: philip.bourne@nih.gov

Rebecca Boyles, M.S.P.H.
Data Scientist
National Institute of Environmental Health Sciences
National Institutes of Health
Keystone Building, Room 2047
530 Davis Drive
Durham, NC 27713
Telephone: (919) 541-7853
Email: rebecca.boyles@nih.gov

Catherine Brownstein, Ph.D.
Instructor
Departments of Genetics and Genomics
Boston Children's Hospital/Harvard Medical School
300 Longwood Avenue
Boston, MA 02115
Telephone: (203) 901-6451
Email: catherine.brownstein@childrens.harvard.edu

Shawn Burgess, Ph.D.
Senior Investigator
Translational and Functional Genomics Branch
National Human Genome Research Institute
National Institutes of Health
Building 50, Room 5537
50 South Drive, MSC 8004
Bethesda, MD 20892-8004
Telephone: (301) 594-8224
Email: burgess@mail.nih.gov

Aditi Chandra, B.A.
Post-Baccalaureate Intramural Research Training Award Fellow
National Institute of Diabetes and Digestive and Kidney Diseases
National Institutes of Health
Building 8, Room 323
8 Center Drive
Bethesda, MD 20892
Telephone: (408) 391-6797
Email: aditic93@gmail.com

Faye Chen, Ph.D.
Program Director
Division of Musculoskeletal Diseases
National Institute of Arthritis and Musculoskeletal and Skin Diseases
National Institutes of Health
One Democracy Plaza, Room 852
6701 Democracy Boulevard, MSC 4872
Bethesda, MD 20892-4872
Telephone: (301) 594-5055
Email: chenf1@mail.nih.gov

Quan Chen, Ph.D.
Program Officer
Division of Allergy, Immunology and Transplantation
National Institute of Allergy and Infectious Diseases
National Institutes of Health
Building 5601FL, Room 7A76
5601 Fishers Lane, MSC 9828
Rockville, MD 20852
Telephone: (240) 627-3489
Email: quan.chen@nih.gov

Elissa Chesler, Ph.D.
Associate Professor
The Jackson Laboratory
600 Main Street
Bar Harbor, ME 04609
Telephone: (207) 288-6453
Email: elissa.chesler@jax.org

Rex Chisholm, Ph.D.
Vice Dean for Scientific Affairs and Graduate Education
Center for Genetic Medicine
Robert H. Lurie Medical Research Center
Feinberg School of Medicine
Northwestern University
303 E. Superior Street
Chicago, IL 60611

Telephone: (312) 503-3209
Fax: (312) 503-5603
Email: r-chisholm@northwestern.edu

Joyce Cohen, D.V.M.
Associate Director of Animal Resources
Yerkes National Primate Research Center
Emory University
954 Gatewood Road N.E.
Atlanta, GA 30329
Telephone: (404) 712-8103
Email: joyce.cohen@emory.edu

Daniel Colón-Ramos, Ph.D.
Associate Professor of Cell Biology
Department of Cell Biology
Boyer Center for Molecular Medicine
Yale University
295 Congress Avenue, Suite 436B
New Haven, CT 06511
Telephone: (203) 737-3438
Email: daniel.colon-ramos@yale.edu

Miguel Contreras, Ph.D.
Program Officer
Office of Research Infrastructure Programs
Division of Program Coordination, Planning, and Strategic Initiatives
Office of the Director
National Institutes of Health
One Democracy Plaza, Room 945
6701 Democracy Boulevard, MSC 4877
Bethesda, MD 20892-4877
Telephone: (301) 594-9410
Fax: (301) 480-3819
Email: contre1@mail.nih.gov

Laura Cox, Ph.D.
Scientist
Department of Genetics
Texas Biomedical Research Institute
7620 N.W., Loop 410
San Antonio, TX 78227
Telephone: (210) 258-9687
Email: lcox@txbiomed.org

Kara Dolinski, Ph.D.
Assistant Director
Lewis-Sigler Institute for Integrative Genomics
142 Carl Icahn Laboratory
Princeton University
Washington Road
Princeton, NJ 08544
Telephone: (609) 258-1895
Fax: (609) 258-7070
Email: kara@genomics.princeton.edu

Janan Eppig, Ph.D.
Professor
The Jackson Laboratory
600 Main Street
Bar Harbor, ME 04609
Telephone: (207) 288-6422
Email: jte@informatics.jax.org

Anna Fernandez, Ph.D.
Lead Associate, Health Informatics
BoozAllenHamilton
One Preserve Parkway, Suite 200
Rockville, MD 20852
Telephone: (301) 838-3866
Email: fernandez_anna@bah.com

Colin Fletcher, Ph.D.
Program Director
Department of Genome Sciences
National Human Genome Research Institute
National Institutes of Health
Building 5635FL, Room 4076
5635 Fishers Lane, MSC 9305
Rockville, MD 20850-9305
Telephone: (301) 451-1340
Email: fletcherc2@mail.nih.gov

Elise Flynn, B.A.
Technical Consultant
Appistry
1538 17th Street, N.W.
Washington, DC 20036
Telephone: (240) 462-8786
Email: elise@appistry.com

Gilberto Fragoso, Ph.D.
Biomedical Informatics Program Manager
Center for Biomedical Informatics and Information Technology
National Cancer Institute
National Institutes of Health
Building 9609, Room 1W102
9609 Medical Center Drive
Rockville, MD 20850
Telephone: (240) 276-5129
Email: fragosog@mail.nih.gov

Zorina Galis, Ph.D.
Chief
Vascular Biology and Hypertension Branch
National Heart, Lung, and Blood Institute
National Institutes of Health
Two Rockledge Center, Room 8116
6701 Rockledge Drive
Bethesda, MD 20892
Telephone: (301) 435-0560
Email: zorina.galis@nih.gov

Tina Gatlin, Ph.D.
Program Director
Division of Genome Sciences
National Human Genome Research Institute
National Institutes of Health
Building 5635FL, Room 4076
5635 Fishers Lane
Rockville, MD 20850
Telephone: (301) 402-2851
Email: christine.gatlin@nih.gov

Marc Gillespie, Ph.D.
Professor/Biocurator
Department of Pharmaceutical Sciences
St. John's University
8000 Utopia Parkway
Jamaica, NY 11439
Telephone: (718) 990-5249
Email: gillespm@stjohns.edu

Andy Golden, Ph.D.
Senior Investigator
Genetics of Simple Eukaryotes Section
National Institute of Diabetes and Digestive and Kidney Diseases
National Institutes of Health
Building 8, Room 323
8 Center Drive, MSC 0840
Bethesda, MD 20892-0840
Telephone: (301) 594-4367
Email: andyg@mail.nih.gov

Benjamin Good, Ph.D.
Assistant Professor of the Department of Molecular and Experimental Medicine
The Scripps Research Institute
10550 N. Torrey Pines Road, MEM-216
La Jolla, CA 92037
Telephone: (619) 261-2046
Email: ben.mcgee.good@gmail.com

Jo Anne Goodnight, B.S.
Director, Project Development and Analysis
Department of External Affairs
The Jackson Laboratory
600 Main Street
Bar Harbor, ME 04609
Telephone: (207) 664-8427
Email: joanne.goodnight@jax.org

Franziska Grieder, Ph.D., D.V.M.
Director
Office of Research Infrastructure Programs
Division of Program Coordination, Planning, and Strategic Initiatives
Office of the Director
National Institutes of Health
One Democracy Plaza, Room 962
6701 Democracy Boulevard, MSC 4877
Bethesda, MD 20892-4877
Telephone: (301) 435-0744
Email: griederf@mail.nih.gov

Xiang Guo, Ph.D.
Senior Scientist
Division of Translational Science
MedImmune
One Medimmune Way
Gaithersburg, MD 20878
Telephone: (301) 398-6808
Email: guox@medimmune.com

Ada Hamosh, M.D.
Professor and Scientific Director
Online Mendelian Inheritance in Man
Institute of Genetic Medicine
Blalock 1007
600 N. Wolfe Street
Baltimore, MD 21286

Telephone: (410) 614-3313
Fax: (410) 614-9246
Email: ahamosh@jhmi.edu

Melissa Haendel, Ph.D.
Associate Professor
Department of Medical Informatics and Clinical Epidemiology
Oregon Health and Science University Library
3181 S.W. Sam Jackson Park Road – LIB
Portland, OR 97239-3098
Telephone: (503) 494-3460
Fax: (503) 494-3322
Email: haendel@ohsu.edu

Carol Hamilton, Ph.D.
Director, Bioinformatics Program
RTI International
3040 Cornwallis Road
Research Triangle Park, NC 27709
Telephone: (919) 485-2706
Email: chamilton@rti.org

Jack Harding, Ph.D.
Health Scientist Administrator
Division of Comparative Medicine
Office of Research Infrastructure Programs
Division of Program Coordination, Planning, and Strategic Initiatives
Office of the Director
National Institutes of Health
One Democracy Plaza, Room 950
6701 Democracy Boulevard, MSC 4877
Bethesda, MD 20892-4877
Telephone: (301) 435-0776
Fax: (301) 480-3819
Email: hardingj@mail.nih.gov

Sharie Haugabook, Ph.D.
Project Manager/Drug Development Team Lead
Division of Preclinical Innovation
National Center for Advancing Translational Sciences
National Institutes of Health
Building 9800, Room 325
9800 Medical Center Drive, MSC 3370
Rockville, MD 20850
Telephone: (301) 217-6038
Fax: (301) 217-5736
Email: sharie.haugabook@nih.gov

Mervi Heiskanen, Ph.D.
Program Manager
Center for Biomedical Informatics and Information Technology
National Cancer Institute
National Institutes of Health
Building 9609, Room 1W378
9609 Medical Center Drive, MSC 9719
Rockville, MD 20850
Telephone: (240) 276-5175
Fax: (240) 276-7886
Email: heiskame@mail.nih.gov

Gail Herman, M.D., Ph.D.
Principal Investigator
Center for Molecular and Human Genetics
Nationwide Children's Hospital
700 Children's Drive
Columbus, OH 43205
Telephone: (614) 722-3535
Fax: (614) 722-3546
Email: gail.herman@nationwidechildrens.org

Jiarong Hong, Ph.D.
Assistant Professor
Department of Mechanical Engineering
University of Minnesota
111 Church Street, S.E.
Minneapolis, MN 55455
Telephone: (612) 626-4562
Email: jhong@umn.edu

Lorette Javois, Ph.D.
Program Director, Organogenesis
Developmental Biology and Structural Variation Branch
National Institute of Child Health and Human Development
National Institutes of Health
Building 6100, Room 4B01D
6100 Executive Boulevard, MSC 7510
Rockville, MD 20852-7510
Telephone: (301) 435-6890
Email: javoisl@mail.nih.gov

Warren Kibbe, Ph.D.
Director

Center for Biomedical Informatics and Information Technology
National Cancer Institute
National Institutes of Health
Building 9609, Room 1W412
9609 Medical Center Drive, MSC 9719
Rockville, MD 20850-9719
Telephone: (240) 276-7300
Email: warren.kibbe@nih.gov

Teresa Krakauer, Ph.D.
Principal Investigator
Department of Immunology
U.S. Army Medical Research and Materiel Command
1425 Porter Street
Frederick, MD 21702
Telephone: (301) 619-4733
Email: teresa.krakauer.civ@mail.mil

Razelle Kurzrock, M.D.
Senior Deputy Director for Clinical Science
Chief, Division of Hematology and Oncology
University of California, San Diego Moores Cancer Center
3855 Health Sciences Drive
La Jolla, CA 92083
Telephone: (858) 657-7000
Email: rkurzrock@ucsd.edu

Stan Laulederkind, Ph.D.
Research Scientist
Human and Molecular Genetics Center
Medical College of Wisconsin
8701 Watertown Plank Road
Milwaukee, WI 53226
Telephone: (414) 955-7513
Email: slaulede@mcw.edu

Janine Lewis, M.S.
Program Manager
Biomedical Information Services
ICF International
530 Gaither Road
Rockville, MD 20850
Telephone: (301) 407-6668
Email: janine.lewis@icfi.com

Huiqing Li, Ph.D.
Postdoc Fellow
Cell Biology and Physiology Center
National Heart, Lung, and Blood Institute
National Institutes of Health
Building 50, Room 3535
50 South Drive
Bethesda, MD 20892
Telephone: (301) 594-0864
Fax: (301) 402-1915
Email: huiqing.li@nih.gov

Yan Li, Ph.D.
Health Specialist
Division of Diabetes, Endocrinology, and Metabolic Diseases
National Institute of Diabetes and Digestive and Kidney Diseases
National Institutes of Health
Two Democracy Plaza, Room 608
6707 Democracy Boulevard, MSC 5464
Bethesda, MD 20892-5464
Telephone: (301) 435-3721
Email: liy7@mail.nih.gov

Ti Lin, Ph.D.
Program Analyst
Division of Construction and Instruments
Office of Research Infrastructure Programs
Division of Program Coordination, Planning, and Strategic Initiatives
Office of the Director
National Institutes of Health
One Democracy Plaza, Room 949
6701 Democracy Boulevard
Bethesda, MD 20892
Telephone: (301) 594-5367
Fax: (301) 480-3658
Email: linti@mail.nih.gov

Yu Lin, M.D., Ph.D., M.S.
Oak Ridge Institute for Science and Education Fellow
Office of Translational Sciences
Center for Drug Evaluation and Research
Food and Drug Administration
10903 New Hampshire Avenue
Silver Spring, MD 20993
Telephone: (612) 806-2032
Email: yu.lin@fda.hhs.gov

Kent Lloyd, Ph.D., D.V.M.
Professor of Surgery

Director, Mouse Biology Program
University of California, Davis
2795 Second Street, Suite 400
Davis, CA 95618
Telephone: (530) 902-1699
Email: kclloyd@ucdavis.edu

Harvey Luksenburg, M.D.
Special Advisor to the Director
Division of Blood Diseases and Resources
National Heart, Lung, and Blood Institute
National Institutes of Health
Two Rockledge Center, Room 9164
6701 Rockledge Drive, MSC 7950
Bethesda, MD 20892-7950
Telephone: (301) 435-0050
Fax: (301) 480-1046
Email: luksenburgh@mail.nih.gov

Nadya Lumelsky, Ph.D.
Program Director
Integrative Biology and Infectious Diseases Branch
National Institute of Dental and Craniofacial Research
National Institutes of Health
One Democracy Plaza, Room 618
6701 Democracy Boulevard, MSC 4878
Bethesda, MD 20892-4878
Telephone: (301) 594-7703
Fax: (301) 480-8319
Email: nadyal@nidcr.nih.gov

Yves Lussier, M.D., FAMCI
Associate Vice President for Health Sciences
Associate Director, BIO5 Informatics
The University of Arizona Cancer Center
1515 N. Campbell Avenue
Tucson, AZ 85724
Telephone: (520) 626-3968
Email: yves@email.arizona.edu

Xuefei Ma, Ph.D.
Staff Scientist
Laboratory of Molecular Cardiology
Genetics and Development Biology Center
National Heart, Lung, and Blood Institute
National Institutes of Health
Building 10, Room 6C104
10 Center Drive, MSC 1762
Bethesda, MD 20892-1762
Telephone: (301) 402-1993
Email: max@nhlbi.nih.gov

Calum MacRae, M.D., Ph.D.
Chief, Cardiovascular Medicine
Associate Professor
Brigham and Women's Hospital/Harvard Medical School
75 Francis Street, PBB-1
Boston, MA 02115
Telephone: (857) 307-4000
Fax: (617) 732-7134
Email: cmacrae@partners.org

Ann-Marie Mallon, Ph.D.
Head of BioComputing
MRC Harwell
Harwell Campus
Oxfordshire
OX11 0RD
United Kingdom
Telephone: +44 1235 841077
Email: a.mallon@har.mrc.ac.uk

Prashanti Manda, Ph.D.
Postdoctoral Research Associate
University of North Carolina at Chapel Hill
120 South Road
Chapel Hill, NC 27705
Telephone: (850) 766-2775
Email: manda.prashanti@gmail.com

Willie McCullough, Ph.D.
Program Director
Division of Construction and Instruments
Office of Research Infrastructure Programs
Division of Program Coordination, Planning, and Strategic Initiatives
Office of the Director
National Institutes of Health
One Democracy Plaza, Room 958
6701 Democracy Boulevard
Bethesda, MD 20892
Telephone: (301) 435-0783
Email: mccullow@mail.nih.gov

Stephen McFadden, M.S.
Researcher
International Society for Research Activity

5717 Beech Avenue
Bethesda, MD 20817
Telephone: (301) 897-9614
Email: isra01@hush.com

Terrence Meehan, Ph.D.
Coordinator of Mouse Informatics
Samples Phenotype Ontologies Team
Wellcome Trust Genome Campus
Hinxton, Cambridgeshire
CB1 3PJ
United Kingdom
Telephone: +44(0)7544191516
Email: tmeehan@ebi.ac.uk

Melissa Mendez, Ph.D.
Postdoctoral Researcher
Assay Development and Screening Technology
Division of Pre-Clinical Innovation
National Center for Advancing Translational Sciences
National Institutes of Health
Building 9800, Room 347C
9800 Medical Center Drive, MSC 3375
Rockville, MD 20892-3375
Telephone: (301) 217-5475
Email: melissa.mendez@nih.gov

Peter Midford, Ph.D.
Independent Contractor
6003C Willow Oaks Drive
Richmond, VA 23225
Telephone: (785) 218-0662
Email: peter.midford@gmail.com

Oleg Mirochnitchenko, Ph.D.
Health Scientist Administrator
Office of Research Infrastructure Programs
Division of Program Coordination, Planning, and Strategic Initiatives
Office of the Director
National Institutes of Health
One Democracy Plaza, Room 942
6701 Democracy Boulevard, MSC 4877
Bethesda, MD 20892-4877
Telephone: (301) 435-0748
Email: oleg.mirochnitchenko@nih.gov

Elvira Mitraka, Ph.D.
Postdoctoral Fellow
Institute for Genome Sciences
University of Maryland
Biopark II
801 W. Baltimore Street
Baltimore, MD 21201
Telephone: (443) 240-1006
Email: emitraka@som.umaryland.edu

Andrew Mitz, Ph.D.
Staff Scientist
Office of the Chief
Laboratory of Neuropsychology
National Institute of Mental Health
National Institutes of Health
Building 49, Room 1B80
49 Convent Drive, MSC 4401
Bethesda, MD 20892-4401
Telephone: (301) 402-5573
Fax: (301) 402-5444
Email: arm@nih.gov

Stephanie Mohr, Ph.D.
Director
Drosophila RNAi Screening Center
Harvard Medical School
New Research Building, Room 336
77 Avenue Louis Pasteur
Boston, MA 02115
Telephone: (617) 432-5626
Email: stephanie_mohr@hms.harvard.edu

Jason Moore, Ph.D.
Professor of Genetics
Geisel School of Medicine
Dartmouth-Hitchcock Medical Center
Dartmouth University
706 Rubin Building
One Medical Center Drive, HB 7937
Lebanon, NH 03756
Telephone: (603) 653-9939
Fax: (603) 653-9952
Email: jason.h.moore@dartmouth.edu

Manuel Moro, Ph.D., D.V.M.
Program Official
Office of Research Infrastructure Programs
Division of Program Coordination, Planning, and Strategic Initiatives
Office of the Director
National Institutes of Health
One Democracy Plaza, Room 941

6701 Democracy Boulevard, MSC 4877
Bethesda, MD 20892-4877
Telephone: (301) 435-0960
Fax: (301) 480-3819
Email: manuel.moro@nih.gov

Mary Mullins, Ph.D.
Professor of Cell and Development Biology Department
Perelman School of Medicine at the University of Pennsylvania
1152 BRB II/III
421 Curie Boulevard
Philadelphia, PA 19104-6058
Telephone: (215) 898-2644
Fax: (215) 898-9871
Email: mullins@mail.med.upenn.edu

Chris Mungall, Ph.D.
Bioinformatics Scientist
Berkeley Bioinformatics Open-Source Projects
Lawrence Berkeley National Laboratory
University of California, Berkeley
1 Cyclotron Road, Mailstop 100PGF100
Berkeley, CA 94720
Telephone: (510) 486-7508
Fax: (510) 486-4229
Email: cjmungall@lbl.gov

Stephanie Murphy, V.M.D, Ph.D.
Director
Division of Comparative Medicine
Office of Research Infrastructure Programs
Division of Program Coordination, Planning, and Strategic Initiatives
Office of the Director
National Institutes of Health
One Democracy Plaza, Room 954
6701 Democracy Boulevard, MSC 4877
Bethesda, MD 20892-4877
Telephone: (301) 451-7818
Fax: (301) 480-3819
Email: stephanie.murphy@nih.gov

Nancy Nadon, Ph.D.
Chief
Biological Resources Branch
Division of Aging Biology
National Institute on Aging
National Institutes of Health
Gateway Building, Room 2C231
7201 Wisconsin Avenue, MSC 9205
Bethesda, MD 20892-9205
Telephone: (301) 402-7744
Fax: (301) 402-0010
Email: nadonn@nia.nih.gov

Bindu Nanduri, Ph.D.
Assistant Professor
Department of Basic Sciences
College of Veterinary Medicine
Mississippi State University
240 Wise Center Drive
Mississippi State, MS 39762
Telephone: (662) 325-5859
Fax: (662) 325-1031
Email: nanduribindu@gmail.com

Tristan Nelson, B.A.
Technical Analyst
Autism and Developmental Medicine
Geisinger Medical Center
100 N. Academy Avenue
Danville, PA 17822
Telephone: (570) 271-6211
Email: tnelson@geisinger.edu

Claire O'Donovan
UniProt Team Leader (Content)
Wellcome Trust Genome Campus
Hinxton, Cambridge
CB10 1SD
United Kingdom
Telephone: +44 1223 494460
Email: odonovan@ebi.ac.uk

Andrew Oler, Ph.D.
Clinical Genomics Program Lead
Bioinformatics and Computational Bioscience Branch
National Institute of Allergy and Infectious Diseases
National Institutes of Health
Building 31, Room 3B62
31 Campus Drive, MSC 2135
Bethesda, MD 20892-2135
Telephone: (240) 507-3791
Fax: (301) 480-4743
Email: andrew.oler@nih.gov

Andras Orosz, Ph.D.
Program Director
Division of Metabolism and Health Effects
National Institute on Alcohol Abuse and Alcoholism
National Institutes of Health
Building 5635FL, Room 2119
5635 Fishers Lane
Rockville, MD 20852
Telephone: (301) 443-2193
Email: orosza@mail.nih.gov

Elizabeth Ottinger, Ph.D.
Senior Project Manager/Drug Development Team Lead
Division of Pre-Clinical Innovation
National Center for Advancing Translational Sciences
National Institutes of Health
Building 9800, Room 326
9800 Medical Center Drive, MSC 3370
Rockville, MD 20850-3370
Telephone: (301) 217-5478
Email: ottingerea@mail.nih.gov

Anil Panackal, M.D., Sc.M.
Staff Clinician
Division of Intramural Research
Laboratory of Clinical Infectious Diseases
National Institute of Allergy and Infectious Diseases
National Institutes of Health
Building 10, Room 11N222
10 Center Drive
Bethesda, MD 20892
Telephone: (301) 496-2775
Email: anil.panackal@nih.gov

Sarah Pendergrass, Ph.D., M.S.
Assistant Professor
Biomedical and Translational Informatics Program
100 N. Academy Avenue
Danville, PA 17822
Telephone: (603) 369-7709
Fax: (570) 214-9451
Email: spendergrass@geisinger.edu

Jonathan Pollock, Ph.D.
Chief
Genetics and Molecular Neurobiology Research Branch
Division of Neuroscience and Behavior
National Institute on Drug Abuse
National Institutes of Health
Neuroscience Building, Room 4255
6001 Executive Boulevard, MSC 9555
Bethesda, MD 20892-9555
Telephone: (301) 435-1309
Fax: (301) 594-6043
Email: jpollock@mail.nih.gov

John Quackenbush, Ph.D.
Professor of Computational Biology and Bioinformatics
Department of Biostatistics
Dana-Farber Cancer Institute
450 Brookline Avenue, Mailstop SM822
Boston, MA 02215
Telephone: (617) 582-8163
Fax: (617) 582-7760
Email: johnq@jimmy.harvard.edu

Armin Raznahan, M.D., Ph.D.
Clinical Investigator, Child Psychiatry Branch
Section on Childhood Neuropsychiatric Disorders
National Institute of Mental Health
National Institutes of Health
Building 10, Room 4C108
10 Center Drive
Bethesda, MD 20892
Telephone: (301) 453-7927
Email: raznahana@mail.nih.gov

Kirk Reardon, B.A.
Graduate Student
National Institute of Mental Health
National Institutes of Health
10 Center Drive, Building 60
Bethesda, MD 20892
Telephone: (617) 641-6834
Email: kirk.reardon@nih.gov

Matthew Reilly, Ph.D.
Program Director
Division of Neuroscience and Behavior
National Institute on Alcohol Abuse and Alcoholism
National Institutes of Health

Building 5635FL, Room 2065
5635 Fishers Lane, MSC 9304
Rockville, MD 20852-9304
Telephone: (301) 594-6228
Email: reillymt@mail.nih.gov

Rachel Richesson, Ph.D., M.P.H., FACMI
Associate Professor
Duke University School of Nursing
2007 Pearson Building
307 Trent Drive, DUMC 3322
Durham, NC 27710
Telephone: (919) 681-0825
Email: rachel.richesson@duke.edu

Jeffrey Roberts, D.V.M.
Clinical Director
Pathogen Detection Laboratory
California National Primate Research Center
University of California, Davis
One Shields Avenue
Davis, CA 95616
Telephone: (530) 752-6490
Email: jaroberts@ucdavis.edu

Peter Robinson, M.D., M.Sc.
Research Scientist
Institute for Medical Genetics of the Charité Berlin
Universitätsklinikum Charité
Augustenburger Platz 1
13353 Berlin
Germany
Telephone: +49 (0) 3045 0566006
Email: peter.robinson@charité.de

Steven Scholnick, Ph.D.
Program Director
Developmental Biology and Genetics Program
Translational Genomics Research Branch
National Institute of Dental and Craniofacial Research
National Institutes of Health
One Democracy Plaza, Room 606
6701 Democracy Boulevard, MSC 4878
Bethesda, MD 20892-4878
Telephone: (301) 594-3977
Fax: (301) 480-8319
Email: scholnis@nidcr.nih.gov

Charlene Schramm, Ph.D.
Program Officer
Division of Cardiovascular Sciences
Heart Development and Structural Diseases Branch
National Heart, Lung, and Blood Institute
National Institutes of Health
Two Rockledge Center, Room 8222
6701 Rockledge Drive, MSC 7940
Bethesda, MD 20892-7940
Telephone: (301) 402-3793
Email: schrammc@nhlbi.nih.gov

Mark Simpson, Ph.D., D.V.M.
Senior Scientist
Laboratory of Cancer Biology and Genetics
National Cancer Institute
National Institutes of Health
Building 37, Room 2000
37 Convent Drive, MSC 4256
Bethesda, MD 20892-4256
Telephone: (301) 435-6276
Email: mark_simpson@nih.gov

Damian Smedley, Ph.D.
Senior Scientific Manager
Wellcome Trust Genome Campus
Wellcome Trust Sanger Institute
Hinxton, Cambridge CB10 1SA
United Kingdom
Telephone: +44 (0) 1223 834244
Fax: +44 (0) 1223 494919
Email: ds5@sanger.ac.uk

Cynthia Smith, Ph.D.
Research Scientist
Mouse Genome Informatics
The Jackson Laboratory
600 Main Street, Box 45
Bar Harbor, ME 04609
Telephone: (207) 288-6663
Email: cynthia.smith@jax.org

Lisa Spain, Ph.D.
Program Director
Division of Diabetes, Endocrinology, and Metabolic Diseases
National Institute of Diabetes and Digestive and Kidney Diseases

National Institutes of Health
Two Democracy Plaza, Room 695
6707 Democracy Boulevard, MSC 5343
Bethesda, MD 20892-5343
Telephone: (301) 451-9871
Email: spainl@niddk.nih.gov

Derek Stemple, Ph.D.
Head of Mouse and Zebrafish Genetics
Wellcome Trust Genome Campus
Wellcome Trust Sanger Institute
Hinxton, Cambridge CB10 1SA
United Kingdom
Telephone: +44 (0) 1223 834244
Fax: +44 (0) 1223 194919
Email: ds4@sanger.ac.uk

Juilee Thakar, Ph.D.
Assistant Professor
Microbiology and Immunology and Biostatistics and Computational Biology
University of Rochester Medical Center
601 Elmwood Avenue
Rochester, NY 14618
Telephone: (585) 276-6925
Email: juilee_thakar@urmc.rochester.edu

Rajesh Thangapazham, Ph.D.
Research Assistant Professor
Uniformed Services University of the Health Sciences
4301 Jones Bridge Road, C2082
Bethesda, MD 20814
Telephone: (301) 295-3820
Email: rajesh.thangapazham.ctr@usuhs.edu

Barbara Thomas, Ph.D.
Scientific Review Officer
Genes, Genomes and Genetics Integrated Review Group
Center for Scientific Review
National Institutes of Health
Two Rockledge Center, Room 22101
6701 Rockledge Drive, MSC 7890
Bethesda, MD 20892-7890
Telephone: (301) 435-0603
Email: bthomas@csr.nih.gov

Paul Thomas, Ph.D.
Director, Division of Bioinformatics
Associate Professor, Department of Preventive Medicine
University of Southern California
1450 Biggy Street, NRT 2502
Los Angeles, CA 90089-9601
Telephone: (323) 442-7799
Fax: (323) 442-7995
Email: pdthomas@usc.edu

Olga Troyanskaya, Ph.D.
Primary Investigator
Laboratory for Bioinformatics and Functional Genomics
Lewis-Sigler Institute of Integrative Genomics
Princeton University & Simons Center for Data Analysis
242 Carl Icahn Laboratory
Princeton, NJ 08544
Telephone: (609) 258-1749
Fax: (609) 258-1771
Email: ogt@princeton.edu

Terry Van Dyke, Ph.D.
Director
Center for Advanced Preclinical Research
Basic Research Laboratory
National Cancer Institute
National Institutes of Health
Building 560, Room 32-32
1050 Boyles Street
Frederick, MD 21702
Telephone: (301) 846-6680
Email: vandyket@mail.nih.gov

Ceri Van Slyke, Ph.D.
Researcher
The Zebrafish Model Organism Database
University of Oregon
5291 University of Oregon
Eugene, OR 97403-5291
Telephone: (541) 346-2116
Email: van_slyke@zfin.org

Amanda Vinson, Ph.D.
Assistant Scientist/Assistant Professor
Primate Genetics Section
Division of Neuroscience
Oregon Health & Science University
505 N.W. 185th Avenue, Mail Code L584

Beaverton, OR 97006
Telephone: (503) 533-2421
Email: vinsona@ohsu.edu

Lan-Hsiang Wang, Ph.D.
Program Director
Division of Cardiovascular Sciences
Heart Failure and Arrhythmias Branch
National Heart, Lung, and Blood Institute
National Institutes of Health
Two Rockledge Center, Room 8184
6701 Rockledge Drive, MSC 7956
Bethesda, MD 20892-7956
Telephone: (301) 435-0504
Fax: (301) 480-7404
Email: lw72f@nih.gov

Nicole Washington, Ph.D.
Research Scientist
Berkeley Bioinformatics Open-Source Projects (BBOP)
Lawrence Berkeley National Laboratory
University of California, Berkeley
1 Cyclotron Road, Mailstop 64-121
Berkeley, CA 94720
Telephone: (510) 486-6217
Fax: (510) 486-6798
Email: nlwashington@lbl.gov

Harold Watson, Ph.D.
Program Director
Division of Comparative Medicine
Office of Research Infrastructure Programs
Division of Program Coordination, Planning, and Strategic Initiatives
Office of the Director
National Institutes of Health
One Democracy Plaza, Room 944
6701 Democracy Boulevard
Bethesda, MD 20892
Telephone: (301) 435-0884
Email: watsonh@mail.nih.gov

Caleb Webber, Ph.D.
MRC Programme Leader
Medical Sciences Division
Department of Physiology, Anatomy, and Genetics
University of Oxford
LeGros Clark Building
South Parks Road
Oxford OX1 3QX
United Kingdom
Telephone: +44 (0) 1865 272169
Fax: +44 (0) 1865 272420
Email: caleb.webber@dpag.ox.ac.uk

Mark Williams, Ph.D.
Project Officer
Research Resources Section
National Institute of Allergy and Infectious Diseases
National Institutes of Health
Building 5601FL, Room 8G54
5601 Fishers Lane, MSC 9825
Rockville, MD 20892-9825
Telephone: (240) 627-3327
Email: mark.williams4@nih.gov

Peter Williamson, M.D., Ph.D.
Chief
Translational Mycology Unit
Laboratory of Clinical Infectious Diseases
National Institute of Allergy and Infectious Diseases
National Institutes of Health
Building 101, Room 11N22
10 Center Drive, MSC 1888
Bethesda, MD 20892-1888
Telephone: (301)443-8339
Email: peter.williamson2@nih.gov

Keenan Withers, B.A.
Post-Baccalaureate Intramural Research Training Award Fellow
Division of AIDS Research
Office of the Clinical Director
National Institute of Mental Health
National Institutes of Health
Building 10, Room 6-5340
10 Center Drive
Bethesda, MD 20892
Telephone: (301) 496-1338
Fax: (301) 402-2588
Email: keenan.withers@nih.gov

Lawrence Wright, M.A.
Program Manager
Center for Biomedical Informatics and Information Technology
National Cancer Institute
National Institutes of Health
Building 9609, Room 1W106
9609 Medical Center Drive, MSC 9719
Rockville, MD 20850-9719
Telephone: (240) 276-5131
Email: larry.wright@nih.gov

Ashley Xia, Ph.D., M.D.
Program Officer
Division of Allergy, Immunology, and Transplantation
National Institute of Allergy and Infectious Diseases
National Institutes of Health
Building 5601FL, Room 7A78
5601 Fishers Lane, MSC 9828
Rockville, MD 20852-9828
Telephone: (240) 627-3479
Email: axia@niaid.nih.gov

Sandhya Xirasagar, Ph.D.
Project Manager
Bioinformatics and Computational Bioscience Branch
National Institute of Mental Health
National Institutes of Health
Building 5601FL, Room 4A40
5601 Fishers Lane, MSC 9815
Rockville, MD 20852-9815
Telephone: (301) 594-0715
Email: xirasasa@mail.nih.gov

Wenxing Zha, Ph.D.
Project Officer
Division of Epidemiology and Prevention Research
National Institute on Alcohol Abuse and Alcoholism
National Institutes of Health
Building 5635FL, Room 2091
5635 Fishers Lane, MSC 9304
Bethesda, MD 20852-9304
Telephone: (301) 443-0633
Email: wenxing.zha@nih.gov

Yantian Zhang, Ph.D.
Program Director
Imaging Technology Development Branch
National Cancer Institute
National Institutes of Health
Building 9609, Room 4W330
9609 Medical Center Drive, MSC 9729
Rockville, MD 20850-9729
Telephone: (240) 276-5980
Email: yantian.zhang@nih.gov

Sige Zou, Ph.D.
Health Scientist Administrator
Division of Comparative Medicine
Office of Research Infrastructure Programs
Division of Program Coordination, Planning, and Strategic Initiatives
Office of the Director
National Institutes of Health
One Democracy Plaza, Room 943
6701 Democracy Boulevard, MSC 4877
Bethesda, MD 20892-4877
Telephone: (301) 435-0749
Email: zous@mail.nih.gov