WORKSHOP REPORT: NHGRI EXTRAMURAL INFORMATICS & DATA SCIENCE WORKSHOP, SEPT 29-30, 2016

EXECUTIVE SUMMARY

The National Human Genome Research Institute (NHGRI) held an Informatics and Data Science workshop on Sept 29-30, 2016. The goal of the workshop was to identify and prioritize opportunities of significance to the NHGRI extramural computational genomics and data science over the next 3-5 years.

The workshop started with focused discussions around new and emerging opportunities in bioinformatics for genomics in both basic biology and clinical sciences. Following this, consideration shifted to data and compute resources, algorithms to undertake computations at scale, and NHGRI's role in enhancing genome informatics within the broader biomedical community. The participants emphasized that good data science requires both strong statistics, and robust computer science. NHGRI is expected to significantly support both of these scientific areas in order to adequately grow genomic data science.

Several high-priority opportunities were highlighted during the workshop. Given the scale at which genomic data are being generated, one of the top priorities identified was to support the development of statistical and computational tools that enable interactive analysis and visualization to address causality, imputation, handling multi-scale data, and networks. A second recommendation was to invest in methods that enhance our understanding of how genotypes translate to phenotypes from the level of the individual nucleotide all the way to the patient. Other important priorities identified include (a) development of tools, technologies, and policies that support enhanced genomic data sharing; (b) investment in research that integrates patients more fully into genomic medicine (e.g., patients as data providers and consumers); (c) identification and functional annotation of causal variants; (d) development of scalable methods to collaboratively develop with the community multi-scale phenotype-focused ontologies and standards; (e) development of algorithms that scale well (linearly or better with sample size); (f) support for all aspects of single-cell studies; (g) continuation of support of protein-based data resources that support and enhance other complementary resources such as model organism and disease process databases; (h) development of methods to make metadata for genomic and phenotypic data, software, and services that are Findable, Accessible, Interoperable, and Resuable (FAIR); (i) promotion of the development of the cloud environment while working with the larger NIH-wide Data Commons efforts; (j) support of rigorous benchmarking, and where possible/necessary, development of gold standards to improve analytical methods and enhance reproducibility; and (k) investment in improvements needed to integrate genomic medicine into clinical decision systems (CDS).

Finally, it was emphasized that training needs for the community are and should be a cross-cutting focus for NHGRI. These include training experimental scientists in computation and statistics, computational scientists in experimental design, and clinicians in genomics and visualization.

WORKSHOP REPORT

NHGRI Extramural held an Informatics and Data Science-focused workshop on Sept 29-30, 2016, in Bethesda, MD. The goal of the workshop was to identify and prioritize opportunities of significance to the NHGRI Extramural Computational Genomics and Data Science (Informatics) Program over the next 3-5 years. The workshop presentations and discussions were designed to elicit and explore in depth new opportunities and continuing challenges that face the genomic computational and data science community.

WORKSHOP OVERVIEW, INCLUDING IDENTIFIED PROIRITIES

DESIGN AND DEVELOPMENT OF THE WORKSHOP

An external Organizing Committee (Drs. Mike Boehnke, Carol Bult, Trey Ideker, Aviv Regev, and Lincoln Stein) was recruited to work with NHGRI Informatics staff members to structure the meeting agenda (see Appendix 1). The meeting was structured around five broad topic areas that define the state of the field: 1) Challenges in enabling new biology in basic sciences; 2) Challenges in enabling new clinical insights; 3) Data and compute resources; 4) Algorithms for computations at scale; 5) Interfacing NHGRI data and resources with other efforts. A subcommittee was formed for each topic area, and was tasked with addressing one topic in depth, summarizing the state of the art, identifying gaps and brainstorming about future opportunities (see Appendix 2). In order to include a diverse array of viewpoints and opinions, workshop attendees were appointed to each succommitte, with expertise from different areas of genomics, bioinformatics, and computational and data science included in each subcommittee. They were tasked with addressing their assigned topic area in depth, summarizing the state of the art, identifying gaps, and brainstorming about the future (Appendix 3 shows attendee participation in breakout sessions).

Prior to the workshop, NHGRI staff provided all participants baseline information about the current informatics and data science portfolio, along with the goals of the workshop. The subcommittees were involved in planning individual breakout sessions. NHGRI informatics staff also served as an effective glue between the different discussion streams. Important contributions during the workshop from colleagues at the National Cancer Institute (NCI) and the Associate Director for Data Science (ADDS) office were crucial to this outcome.

As outlined in the annotated agenda (Appendix 2), the first day began with a framing presentation by Dr. Phil Bourne covering the activities of the NIH-wide BD2K program and a summary by staff of the current NHGRI portfolio of direct relevance to the workshop. This was followed by genomic sciences and genomic medicine focused breakout sessions that aimed to identify new questions and rethink old ones based upon recent developments in both experimental and computational methods. The summary session focued on identifying bottlenecks and reflecting on needs for genomics and biology to flourish. The afternoon of the first day was devoted to consideration of methods to address the challenges identified. This was accomplished through three breakout groups: one focused on enabling data and compute resources; a second on algorithms for computations at scale; and a third on the challenges of interfacing among data providers, integrating diverse data types, and presenting data and metadata to the broader community so that they can be effectively used for new analysis and discovery. Through presentations and vigorous discussions, each session developed a set of separate priorities.

The second day was devoted to developing a final combined list of priorities (Table 1). Breakout session leaders presented their group's priorities to the entire workshop. Further discussions refined

these priorities by sharpening the description, removing redundancy, narrowing or broadening the focus, and clarifying potential areas of confusion. Everyone was given a limited number of votes to select their top priorities. They were then reviewed, combined, reordered, and a second round of voting occurred. A final discussion of the prioritized list then consolidated the recommendations.

PRIORITIZATED RECOMMENDATIONS

Meeting participants identified thirteen recommendations (spelt out in Table 1 below) as the highest priority for NHGRI. The details in the table are intended to provide context and to more precisely define the meaning and scope of the different recommendations. These details were brought up during the workshop.

TABLE 1

List of Priorities Identified at Workshop[1]	
Recommendation	**Additional Details & Specifics on the Recommendation**
NHGRI should support development of statistical and computational tools that enable interactive analysis and visualization of large datasets to enable thoughtful balancing of human expert input with machine automation.	• Visualization tools that combine exploration, history record & story as a well-defined goal. • Tools that address latency of complex data sets. • Invest in data generation for easier data integration. • Avoid standardizing on pipelines/workflows prematurely.
NHGRI should invest in methods and data that enhance our understanding of how genotypes translate to phenotypes at all "scales" from nucleotide to patient.	• Translation of raw molecular profiles and interactions into multi-scale models of biological systems. • Inference of causality by genetic perturbations and drugs. • Improve imputation for a broad array of data types, both theoretical methods to approach imputation and applied methods to use in practice. • Investment in data generation, analysis methods, and integration tools.
NHGRI should participate in the development of tools and technologies, and policies to ensure that genomic data can be shared for research and clinical applications.	• Need common standards and policies for consent and for data access agreements. • Linking private and public data. • Linking individual and population level data. • Secure cloud computing. • Quantification of privacy risk. • Common workflow language to standardize operations behind APIs.

[1] These recommendations are not in any specific priority order. These have been edited very lightly by NHGRI staff to improve clarity.

	• Advocate for licensing language that allows for data reuse and redistribution; tackle social and legal impediments to data reuse.
NHGRI should support development of statistical and computational methods and tools to identify causal variants, taking advantage of large genotype data (across multiple studies), multi-scale phenotypes and functional annotations.	• Understanding of how individual genetic variation coalesces/integrates within pathways, interactions, and networks. • Development of innovative methods for filtering variants, e.g. based on genealogy.
NHGRI should support development of scalable methods to coherently and collaboratively develop multi-scale phenotype-focused ontologies and standards.	• Support curation of ontologies and standards. • Support annotation of publicly available data.

NHGRI should support development of efficient statistical and computational methods, libraries, algorithms, and tools that scale well (ideally linearly or better with sample size) for sequence and association analysis and other compute-intensive applications.	• Examples include the following:
	  ○ Methods for complex sequence analysuis, such as structural variant detection.
	  ○ Scalable methods for rare variant association analysis.
	  ○ Approaches for multifactorial disease associations, such as gene-gene and gene-environment interactions.

NHGRI should promote the development of vertically integrated data resources that are designed to support the needs of horizontally-organized knowledgebases.	• Resources that focus on a specific data type, such as proteins, across species, and promote the transfer of knowledge across species and disease process boundaries and semantic infrastructures. • Connections are required between vertical pillars of deeply-characterized data types and horizontal community-organized activities organized broadly around specific activities such as a disease process or species. • This is an area where coordination and collaboration should be emphasized to avoid redundancy, recognizing the challenges of doing this across funding agencies and the need for meaningful metrics addressing the entire lifecycle of these resources.
NHGRI should support the development of methods to enable the scalable, intelligent and cost-effective curation of FAIR	• Resources focused on curated metadata and computable phenotypes. • Standard phenotype descriptions to aid diagnosis and diagnostic categories.

metadata for genomic and phenotypic data, software and services.	• Standard phenotype metadata to aid data analysis and re-use. • Methods and tools that can facilitate large-scale phenotyping. • FAIR representation of software and services need to be developed and supported that enable re-use of these products by the larger community.
NHGRI should promote the development of a cloud environment that can be used by NHGRI-funded projects to share their data and tools, and which will participate in the NIH "Data Commons."	• Develop as a "tributary" to the "data lake" of NIH Commons. • A substantial software stack is required to implement this, including the creation of indexing and search services, and integration with the authentication and authorization services. • Use more than one vendor to promote choice and competition. Different solutions should interoperate.
NHGRI should support rigorous benchmarking and, where possible/necessary, development of "gold standards."	• Strengthen reproducibility or research results from informatics and data science perspective. • Mechanisms for head-to-head comparison of methods. • Applies to both computational and experimental benchmarking.
NHGRI should invest in improvements needed to integrate genomic medicine into Clinical Decision Systems (CDS).	• Improve interfaces/visualization tools for presenting genomic data to clinical audiences. • Representi uncertainty/risk in CDS. • Methods to validate outcomes based on actions that results from CDS recommendations. • Improve CDS algorithms, for example to accommodate polygenic models. • Develop methods for communicating changes as data and models are updated. • Quantification of improved patient health as prime metric for retrun on investment. • Engage health systems, including community clinics, in implementing genomic medicine.
NHGRI should invest in research that integrates patients more fully into genomic medicine research and clinical practice.	• Patients as data providers, consumers of genomic information, and collaborators. • Consent and privacy should both be addressed with patients. There are informatics challenges in this space (not investigated deeply within the workshop).
NHGRI should support single cell studies.	• Support methods development to improve the utilization of single cell genomic data being generated. • Look for ways to support the community driven Human Cell Atlas project.

	• Dealing with heterogeneity. • Understanding cell types and interpreting single cell data. • Development of statistical models and computational methods that account for stochastic behavior in biology. • User access to well-processed, large scale single cell data via simple portals (federated) for access; QC/normalization and application on portals.

SUMMARY OF WORKSHOP DISCUSSION BY TOPIC AREA

The meeting was organized around five areas (see the sessions in Appendix 1). Each of the areas was highlighted in presentations and discussions during the breakout sessions. Below we provide an overview of the topics the group found to be important, and outline specific details discussed in relation to the topic.

Area 1:

Challenges in Enabling New Biology in Basic Sciences: The broad topics that were discussed in this area included how genotypes translate to phenotypes, single cell studies, challenges to building useful models, and focusing on rigorous benchmarking and gold standards.

- Enhance our understanding of how genotypes translate to phenotypes at all "scales" from nucleotide to patient
    - Translation of raw molecular profiles and interactions into multi-scale models of biological systems
    - Inference of causality by genetic perturbations and drugs
    - Imputation (theory and practice)
    - Investment in both data generation and methods will be required
    - Think about open and dynamic consent models
- Support single cell studies since they are essential for understanding cell type and eventually relating that back to the genome
    - Methods development, comparison and standardization
    - Developing a cell atlas
    - Dealing with heterogeneity
    - Understanding cell types and interpreting single cell data
    - Support development of statistical models and computational methods that account for stochastic behavior in biology
- Support useful models
    - Interactive
    - Inference of causality by perturbation experiments
    - Networks
    - Multi-scale
- Ensure rigorous benchmarking and, where possible, enable development of "gold standards" (both computational and experimental)
    - Reproducibility
    - Head-to-head comparison of methods

- Data collection can be informed by analysis needs
- Develop ways to merge data at different scales, and across time and space
- Develop better imputation methods for broader data types
- Provide standardized methods, data types, and methods for naming and indexing elements
- Incentivize "good behavior"

Challenges in Enabling New Clinical Insights: Important topics discussed here included barriers to sharing clinical data, getting to curated metadata and computable phenotypes, integrating genomics into clinical decision systems, and approaches to integrating patients more fully into genomic research and clinical practice.

- Need tools, technologies, and policies that remove barriers to sharing clinical data
  - Consents/data access agreements – need common standards and policies; new models for sharing clinical data
  - Secure cloud computing
  - Quantification of privacy
  - Linking private and public data
  - Linking individual and population level data
  - Common workflow language (standardize operations behind APIs)
- Developing curated metadata and computable phenotypes are important bottlenecks
  - Standard phenotype descriptions to aid diagnosis and diagnostic categories
  - Standard phenotype metadata to aid data analysis and re-use
  - Methods and tools that can facilitate large-scale phenotyping
- Despite the myriad challenges in validating and utilizing clinical decision systems, it is important to consider the specific challenges of integrating genomic medicine into them:
  - Improve interfaces/visualization tools for presenting genomic data to clinical audiences
  - Representing uncertainty/risk
  - Methods to validate outcomes based on actions that results from CDS recommendations
  - Improve CDS algorithms, for example to accommodate polygenic models
  - Develop methods for communicating changes as data and models are updated
  - Quantification of the prime ROI – improved patient health
  - Engage health systems, including community clinics, in implementing genomic medicine
- Integrating patients fully into genomic medicine research and clinical practice
  - Patients as data providers, consumers of genomic information, and collaborators
  - Consent and privacy should both be addressed with patients

Data & Compute Resources: This group focused on topics including importance of developing interoperability standards, ontologies and controlled vocabularies, sustaining both 'horizontal' (e.g. species or disease or model organism) and 'vertical' (e.g. proteins, pathways) data resources, reducing costs of maintaining data resources, developing a cloud 'sandbox' in coordination with the NIH Commons, and maintaining strict guidelines for developing FAIR metrics for data and tools.

- Development and use of data and service interoperability standards
  - This includes: common data models, controlled vocabularies, adherence to common terminology standards (e.g., gene models), common APIs
  - Semantic interoperability and misaligned data models: Invest in better informatics, software, and coordination to align and integrate across diverse downstream data re-users and data providers

- Standards (develop standard/uniform terms for data contribution, use, and licensing)
    - For example, standard language for data sharing in patient consent forms, to reduce the workload on data access committees and to allow for machine-readable data access rules
- Continue development of vertically integrated data resources that are designed to support the needs of horizontally-organized knowledgebases
    - These are resources that focus on a specific data type, such as proteins, across species, and which promote the transfer of knowledge across species and disease process boundaries. It also covers such semantic infrastructures as GO
- Incentivize the information resources to consolidate their technical infrastructures, and to create a framework on which new community resources can be built
    - The combination of this point and the previous one will together reduce the cost of creating and maintaining a new MOD or similar database. It also opens a potential path to "HumanBase"
    - Downstream data users should play a role in defining these requirements
- A cloud "sandbox" environment that can be used by NHGRI-funded projects to share data and tools,
    - This should participate in the NIH "Data Commons" as a tributary to the "data lake" that Vivien Bonazzi described
    - There are lots of corollaries to this, including the creation of indexing and search services, and integration with the authentication & authorization services
    - Use more than one vendor to promote choice and competition. Different solutions should interoperate
- NHGRI should promote the FAIR use of software tools & data
    - Packaging in Docker-style packages
    - Registry in findable repositories
    - Focusing on published analytic methods
    - Creation of methods repositories where use and citations are tracked
    - Promote citation of software use
    - Better recognize contributions to existing software, including proper citations

Algorthims for Computations at Scale: This group discussed the following topics: critical need to develop efficient statistical and computational methods, and tools that promote interactive analysis and visualization.

- Development of efficient statistical and computational methods, libraries, algorithms, and tools that scale well (ideally linearly or better with sample size) for sequence analysis and other compute-intensive applications
- Development of statistical and computational tools that enable interactive analysis and visualization of large datasets to enable thoughtful balancing of human expert input with machine automation. Avoid building pipelines/workflows prematurely.
- Development of statistical models and computational methods that account for stochastic behavior in biology, for example, for single cell data.
- Development of statistical and computational methods and tools to identify causal variants, taking advantage of large genotype data (across multiple studies), multi-scale phenotypes and functional annotations
    - Understanding how individual genetic variation coalesces/integrates within pathways, interactions, and networks
    - Development of innovative new methods for filtering variants, e.g. based on genealogy

- Development of methods to enable the scalable, intelligent and cost-effective curation of FAIR metadata for genomic and phenotypic data, software and services
- Scalable and reliable computational methods for the prioritization of variant association validation using evidence gathered across scales and multiple sources
- Scalable methods to coherently and collaboratively develop phenotype-focused ontologies and standards

Interfacing HG Data and Resources for Others: This group focused on interoperability questions and sharing, generating data and interfaces and exploration tools to interpret genetic variations, visualization tools, enabling user access to well-processed and large-scale single cell data, and scalable methods to collaboratively develop phenotype-focused ontologies and standards.

- Improving the social, technical, and legal interoperability of data sources and in their sharing
    - Licensing/data use restrictions: Tackle social/legal impediments, by considering best practices, common licenses, policies, and incentives to promote the permission to re-use and re-distribute data (write white papers)
    - Poor data identification: Build on recent advances to harmonize identifiers and educate users on their management, to improve evidence metadata from sources, and build tools for data resolution, in more efficient, automated ways
- NHGRI should invest in data, interfaces and exploration tools to help users interpret the function of genetic variants in cancer, common and rare disease:
    - In partnership with informatics effort, perform data collection that drives functional, causal interpretation (e.g., functional assays of variants, at scale; epigenomics of tumors; large scale perturbation experiments of single and multiple genes)
    - Develop educational tools for clinical users
    - Develop visualizations and tools to relate the "called function" of genetic variants to the underlying evidence
    - Streamline large scale data processing and sharing in international consortia
    - Provide a role for citizen scientists
- NHGRI should invest heavily in visualization tools, as a well defined goal
    - Build visualization tools ("Vistory") that combine exploration, history record, and story (outcome)
    - Invest in tools that address latency for complex data sets
    - Invest in tools that allow graceful integration
- NHGRI should invest in users' access to well-processed, large scale single cell data:
    - Develop simple portal for access; perhaps consider a federated model
    - Invest in tools for QC/normalization, their application to single-cell data publicly available to the community.
    - Articulate the biomedical use cases: Where is a GWAS gene expressed?;Signatures for a next-generation "complete blood count"?;, etc.

ACKNOWLEDGEMENTS

First and foremost, we thank the five Organizing Committee members: Drs. Mike Boehnke, Carol Bult, Trey Ideker, Aviv Regev and Lincoln Stein. Their critical insights helped guide the process through articulation, design and chairing of the different sessions during the workshop. The Informatics group within NHGRI Extramural consists of Drs. Ajay Pillai (Lead), Lisa Brooks, Valentina di Francesco, Dan Gilchrist, Peter Good, Mike Pazin, Heidi Sofia, Jennifer Troyer, Chris Wellington, Ken Wiley. Dr. Jeff

Schloss helped in multiple ways to steer the design of the workshop. NHGRI Director Dr. Eric Green commented on the agenda and provided invaluable support and participation during the prioritization segment of the meeting. Kevin Lee organized dozens of planning meetings prior to the workshop, facilitated planning for the day of the workshop, took notes, and contributed to this workshop report; his contributions were indispensable. Last but not the least, we thank all the participants (Appendix 3), both external scientists, and colleagues from NCI and ADDS who worked prior to the meeting in thinking through and designing the sessions, in enthusiastic participation during the session and being great sports through an intense day and half of the workshop, especially the first day which lasted 10+ hours.

FINAL

APPENDIX 1: AGENDA

NHGRI Computational Genomics and Data Science Workshop AGENDA OUTLINE

Sept 29, 2016 8am – 6:30pm (Day 1)

SESSION 1: INTRODUCTION AND SETTING THE STAGE **8:00-10:30 AM**

(5635 Fishers Lane, Terrace Room)

BREAK **10:30-11:00AM**

SESSION 2: SCIENTIFIC QUESTIONS (Concurrent) **10:45-1:00PM**

- **Session 2A: Challenges in Enabling New Biology in Basic Sciences** (5635 Fishers Lane, Terrace Room)
- **Session 2B: Challenges in Enabling New Clinical Insights** (5625 Fishers Lane, 5th Floor Conference Room)

LUNCH **1:00-2:00 PM**

SESSION 3: JOINT REVIEW OF SESSION 2 **2:00-3:00 PM**

(5635 Fishers Lane, Terrace Room)

SESSION 4: HOW TO BEST EXECUTE? (Concurrent) **3:00-5:00 PM**

- **Session 4A: Data and Compute Resources** (5635 Fishers Lane, Terrace Room)
- **Session 4B: Algorithms for Computations At Scale** (5625 Fishers Lane, 5th Floor Conference Room)
- **Session 4C: Interfacing HG Data and Resources for Others** (5635 Fishers Lane, 4th Floor Conference Room)

BREAK **5:00-5:30 PM**

SESSION 5: DESIGNING DAY 2 **5:30-6:30 PM**

(5635 Fishers Lane, Terrace Room)

ADJOURN DAY 1 **6:30 PM**

EXECUTIVE SESSION: Organizing Committee & NHGRI Staff: Preparing for Day 2 **7-8PM**

Sept 30, 2016 8am – 12:30pm (Day 2)

SESSION 6: PRIORITIZE **8:00-12:30 PM**

FINAL

(5635 Fishers Lane, Terrace Room)

BREAK **9:45-10:15 AM**

MEETING ADJOURNS

APPENDIX 2: ANNOTATED AGENDA

Session Details

SESSION 1: INTRODUCTION AND SETTING THE STAGE **8:00-10:30 AM**

(5635 Fishers Lane, Terrace)

- Welcome 8:00-8:05AM
- NIH-wide efforts in data science 8:05-8:35 AM

Phil Bourne

- Workshop Background and Goals 8:35-9:30 AM

Ajay Pillai

- Summary and Charge for Each Session 9:30-10:05 AM

Organizing Committee

- General Discussion of Scope 10:05-10:30AM

All

SESSION 2: SCIENTIFIC QUESTIONS **10:45AM-1:00PM**

Parallel Session 2A: Challenges in Enabling New Biology in Basic Sciences

(5635 Fishers Lane, Terrace Room)

Co-Leads: Trey Ideker, Lucia Ohno-Machado, Mike Pazin

Rapporteurs: *Anne Kwitek, Nils Gehlenborg, Jen Troyer*

- New integrative methods to address the genotype-phenotype problem 11:00-11:30AM *Trey Ideker (presenter) and leads discussion*
- Relating single-cell information to tissues and phenotypes 11:30-2:00PM *Dana Pe'er (presenter) and leads discussion*
- The epigenetic layer: Machine learning for understanding regulation 12:00-12:30PM *Anshul Kundaje (presenter) and leads discussion*

- Standard practices for validation of bioinformatic methods 12:30-12:45PM *Hector Corrada-Bravo (presenter) and leads discussion*

- Data Sharing 12:45-1:00 PM *Lucila Ohno-Machado (presenter) and leads discussion*

Parallel Session 2B: Challenges in Enabling New Clinical Insights

(5625 Fishers Lane, 5th Floor Conference Room)

Co-Leads: Carol Bult, Mark Gerstein, Ken Wiley

Rapporteurs: *Adam Arkin, David Glazer, Ajay Pillai*

- Introduction 11:00-11:05AM *Carol Bult*

Integration of genomic information with clinical phenotypes and issues related to data privacy 11:05-11:15AM *Mark Gerstein*

- Deep Phenotyping for Precision Medicine 11:15-11:25AM *Peter Robinson*

- Title: 11:25-11:30AM *David Glazer*

- From Precision to Clinician – What can we learn from previous bioinformatics efforts to impact the point-of-care 11:30-11:35AM *Sachin Kheterpal*

- Informatics tools and infrastructure for distributed and variably protected data 11:35-11:45AM *James Taylor*

- Effectively communicating evidence from omics data analyses for clinical applications 11:45-11:55AM *Casey Overby*

- Heterogeneous data, conditional information, and communicating uncertainty in biomedical predictions 11:55-12:05PM *Adam Arkin*

- Discussion 12:05-1:00PM *All*

SESSION 3: BRINGING IT TOGETHER **2:00-3:00 PM**

(5635 Fishers Lane, Terrace Room)

Co-Leads: Valentina di Francesco, Ajay Pillai

- Summary from Basic Biology Session 2:00-2:15 PM

Review notes with assistance from Session 2A Rapporteurs

- Summary from Clinical Session 2:15-2:30 PM

Review notes with assistance from Session 2B Rapporteurs

- Discussion (setting the stage for afternoon sessions) 2:30-2:45 PM

All

SESSION 4: HOW TO BEST EXECUTE? **3:00-5:00 PM**

Parallel Session 4A: Data and Compute Resources

(5635 Fishers Lane, Terrace Room)

Co-Leads: Lincoln Stein, Mike Lin, Valentina di Francesco

Rapporteurs: Anthony Philippakis, Bonnie Berger, Erin Ramos

- Motivation for why we are here 3:00-3:05PM Valentina di Francesco
- Overview and survey of major data resources 3:05-3:10PM *Mike Lin*
- What are impediments to using the current data resources? 3:10-3:15PM *Paul Flicek*
- Discussion 3:15-4:00PM *All*
- Overview and survey of compute resources 4:00-4:05PM *Owen White*

  What are impediments to using the current compute resources? 4:05-4:10PM *Bob Grossman*
- Discussion 4:10-5:00PM

All

Parallel Session 4B: Algorithms for Computations At Scale

(5625 Fishers Lane, 5th Floor Conference Room)

Co-Leads: Mike Boehnke, Rafael Irizarry, Lisa Brooks

Rapporteurs: Chiara Sabatti, Jyotishman Pathak, Dan Gilchrist

- Introduction 3:00-3:10PM

Michael Boehnke

- Variant calling and association analysis on millions of genomes—how

do we scale up? 3:10-3:25PM

Hyun Min Kang

- Single-cell genomics data: dealing with stochasticity 3:25-3:40PM

Rafael Irizarry

- Understanding the role of genetic variants: where are we in terms of establishing causality? 3:40-3:55PM

Chiara Sabatti

- Data integration and normalization of information across multiple data types to understand molecular processes 3:55-4:10PM

Michel Dumontier

- Discussion 4:10-5:00PM

All

Parallel Session 4C: Interfacing HG Data and Resources for Others

(5635 Fishers Lane, 4th Floor)

Co-Lead: Aviv Regev, Warren Kibbe, Heidi Sofia

Rapporteurs: Corrie Painter, Nicola Mulder, Chris Wellington

- Introduction and charge 3:00-3:05PM

Heidi Sofia

- A (future) user's guide to the human cell atlas 3:05-3:20PM

Aviv Regev

- Interpretation of variant pathogenicity in human disease 3:20-3:35PM

Dan Rader/Ekta Khurana

- Interoperability of organism data (technical, semantic, legal) 3:35-3:50PM

Melissa Haendel

- Frontiers of Data Visualization 3:50-4:05PM

Nils Gehlenborg

FINAL

- Brief comments 4:05-4:15PM

Any participant

- Discussion & analysis 4:15-5:00 pm

All

SESSION 5: DESIGNING DAY 2 **5:30-6:30 PM**

(5635 Fishers Lane, Terrace Room)

Co-Leads: Aviv Regev, Jeff Schloss

- Discussion: What are the major themes from sessions 4A, 4B, and 4C? (10-15 mins each)

Review notes from each group's rapporteurs. Are we missing something important?

- Discussion: Review of outcome of sessions 3 & 5 together (15 mins)

  Review notes from each group's rapporteurs. Are we missing something important?

SESSION 6: PRIORITIZE **8:00-12:30AM**

(5635 Fishers Lane Terrace Room)

Lead: Jeff Schloss

- Informatics and Data Science at NHGRI 8:00-8:10AM

Eric Green

- Revisit the major themes — any new insights? 8:10-8:40 AM
- What are the best methods to prioritize? 8:40-9:10 AM
- Prioritize for Basic Sciences 9:10-9:45 AM

Facilitators: Aviv Regev and Trey Ideker

Rapporteurs: Mike Pazin & Dan Gilchrist

FINAL

BREAK **9:45-10:15 AM**

- Prioritize for Clinical Sciences 10:15-11:00 AM

Facilitators: Mike Boehnke and Carol Bult

Rapporteurs: Lisa Brooks & Ken Wiley

- Establishing a virtuous cycle between informatics for basic science and clinical practice 11:00-11:45 AM

Facilitator: Jeff Schloss

Rapporteurs: Heidi Sofia & Chris Wellington

- Prioritize on Infrastructure Questions 11:45-12:30 PM

Facilitator: Lincoln Stein

Rapporteurs: Valentina di Francesco & Jen Troyer

APPENDIX 3: LIST OF ATTENDEES

Name	Morning Breakout Session	Afternoon Breakout Session	Sector	Website/Bio URL
Adam Arkin	B	A	Academic; Government	http://genomics.lbl.gov
Bonnie Berger	A	A	Academic	http://people.csail.mit.edu/bab/
Mike Boehnke	B	B	Academic	https://sph.umich.edu/faculty-profiles/boehnke-michael.html
Angela Brooks	A	B	Academic	https://brookslab.soe.ucsc.edu
Hector Corrada Bravo	A	A	Academic	http://www.hcbravo.org
Carol Bult	B	C	Academic	https://www.jax.org/research-and-faculty/faculty/carol-bult
Michel Dumontier	B	B	Academic	http://dumontierlab.stanford.edu
Paul Flicek	A	A		http://www.ebi.ac.uk/research/flicek
Nils Gehlenborg	A	C	Academic	http://gehlenborg.com
Mark Gerstein	B	C	Academic	http://www.gersteinlab.org
David Glazer	B	B	Industry	
Bob Grossman	A	A	Academic	https://www.ci.uchicago.edu/profile/202
Melissa Gymrek	B	B	Academic	http://melissagymrek.com
Melissa Haendel	B	C	Academic	http://www.ohsu.edu/xd/education/library/about/staff-directory/melissa-haendel.cfm
Trey Ideker	A	C	Academic	http://healthsciences.ucsd.edu/som/medicine/research/labs/ideker/Pages/default.aspx
Rafael Irizarry	A	B	Academic	http://rafalab.dfci.harvard.edu
Hyun Min Kang	A	B	Academic	https://sph.umich.edu/faculty-profiles/kang-hyunmin.html

Sachin Kheterpal	B	C	Academic	https://umchop.org/faculty/sachin.html
Ekta Khurana	A	C	Academic	http://khuranalab.med.cornell.edu
Anshul Kundaje	A	B	Academic	https://med.stanford.edu/profiles/anshul-kundaje
Anne Kwitek	A	A	Academic	http://www.medicine.uiowa.edu/dept_primary_apr.aspx?appointment=Pharmacology&id=kwiteka
Mike Lin	B	A	Industry	http://www.mlin.net
Lucila Ohno-Machado	A	B	Academic	http://healthsciences.ucsd.edu/som/dbmi/people/faculty/Pages/lucila-ohno-machado.aspx
Nicola Mulder	B	C	Academic	http://www.idm.uct.ac.za/nmulder/
Casey Overby	B	B	Academic	https://scholar.google.com/citations?user=qeLrD0sAAAAJ&hl=en
Corrie Painter	B	C	Academic	http://bcsm.org/all-hands-on-deck-the-metastatic-breast-cancer-project/
Jyotishman Pathak	A	B	Academic	http://vivo.med.cornell.edu/display/cwid-jyp2001
Dana Pe'er	A	C	Academic	http://www.c2b2.columbia.edu/danapeerlab/html/
Anthony Philippakis	B	A	Academic	https://www.broadinstitute.org/bios/anthony-philippakis-0
Angel Pizarro	B	A	Industry	https://www.linkedin.com/in/angelpizarro
Dan Rader	B	C	Academic	http://www.med.upenn.edu/apps/faculty/index.php/g5165284/p17778
Aviv Regev	B	C	Academic	https://www.broadinstitute.org/scientific-community/science/core-faculty-labs/regev-lab/regev-lab-home
Chiara Sabatti	B	B	Academic	http://statweb.stanford.edu/~sabatti/
Lincoln Stein	A	A	Academic	https://oicr.on.ca/person/oicr-investigator/lincoln-stein
Jessica Tenebaum	B	C	Academic	https://www.dtmi.duke.edu/who-we-are/tenenbaum-jessica
James Taylor	B	C	Academic	http://jamestaylor.org

Owen White	A	A	Academic	http://medschool.umaryland.edu/FACULTYRESEARCHPROFILE/viewprofile.aspx?id=20313
Thomas James	A	A	NISC	http://www.nisc.nih.gov
Jim Mullikin	B	B	NISC	https://www.genome.gov/11007681/mullikin--group/
Susan Gregurick	A	A	NIGMS	
Vivien Bonazzi	A	A	ADDS	
Steve Sherry	A	A	NCBI	
Warren Kibbe	A	C	NCI	
Sean Davis	B	C	NCI	
Tony Kerlavage	A	B	NCI	
Cashell Jaquish	B	A	NHLBI	
Lisa Brooks	B	B	NHGRI	
Valentina Di Francesco	A	A	NHGRI	
Peter Good	A	A	NHGRI	
Mike Pazin	A	A	NHGRI	
Ajay Pillai	B	C	NHGRI	
Jeff Schloss	A	A	NHGRI	
Heidi Sofia	B	C	NHGRI	
Jennifer Troyer	A	C	NHGRI	
Ken Wiley	B	B	NHGRI	
Erin Ramos	B	A	NHGRI	
Dan Gilchrist	A	B	NHGRI	
Tina Gatlin	A	A	NHGRI	
Mike Smith	A	B	NHGRI	