NHGRI Genomic Data Science Virtual Town Halls

Summary of Process & Participant Feedback
Work done by Contractors (KnowInnovation)

I. Summaries of Town Hall discussions

The questions/challenges submitted by participants were grouped. Each group was assigned to a breakout session. These original questions/challenges are organized as bulleted items under the phrase "How would we…?" During the town halls these questions were discussed, refined and expanded upon. The results of these discussion are itemized under each breakout session summary below. The participants were asked to identify the most important recommendations, which are provided in the box in each breakout session's summary.

Challenges submitted by a subset of participants (Breakout sessions: 1 & 13 in Town Halls 1 & 2) – What is stopping us from overcoming these challenges? What are the gaps in knowledge?

How would we:

- develop an open data culture
- accelerate and improve access to data sharing in a way that enables collaboration between researchers and facilitates efficient development of computational and statistical methods
- improve data storage and data/method sharing
- improve access to results/data/algorithms/methods from industry laboratories
- create and manage more standardized, publicly available, and controlled data
- reduce administrative approvals to allow international data exchange
- access large human data sets that avoid HIPAA requirements
- create incentives to produce 'sharing-ready' tools/data or to cite those tools/data -- we do not value methodology contributions as much as we value the findings enabled by these methods
- eliminate privacy concerns that limit data sharing
- obtain consent such that the associated samples and data are usable in further analyses

Box: Summary of the "most important" challenges/recommendations from the breakout session.

- Standards: data formats that interoperate with EHR data, database, models, technology
- Standard guidance/policies from NIH needed for data sharing and consent structure (across institutes/agencies)
- Catalog of datasets and resources for easy discovery
- Scalable visualization platform
- Designing broad/standard consent forms to be used across institutes
- Encouraging and enabling data sharing
- Centralized data analysis service from NIH

Notes

1. **Standard data models and formats, as well as database and technology (e.g., programming languages, lab tests)**
    - There is a communication gap between computer scientists and geneticists regarding data storage and formatting.
    - The process of data sharing is an obstacle – time is spent finding the right database to deposit data (e.g., metabolomics data - dbGaP or SRA?) and the right cloud platform (so that data are findable).
    - Lack of dbGaP for other omics data types.
    - Standardize the data processing and submission – raw data, processed data, ready-to-use data. Sometimes data is more granular or less specific than the standard.
    - Lab tests and their parameters change, so it is difficult to standardize that data.
    - <u>Recommendation</u>: Need to standardize data formats and models that communicate with EHR data.
    - <u>Recommendation</u>: Discussions to standardize data format, database, and technology (e.g., programming language) for data sharing before data capture.
    - <u>Recommendation</u>: Comparing and merging metadata results, and sharing the interfaces to share the data.
2. **Standard data sharing policy and administration across jurisdictions**
    - NIH expectations for standards of data sharing with collaborators and with the general research community are unclear.
    - Reducing administrative approvals to use data/resources from one jurisdiction to another: It is unclear whether there is an organizational body that already serves this purpose, and whether any clear standards exist.
    - Inclusion of data reuse or data sharing in consent forms as standard practice.
    - <u>Recommendation</u>: Develop clear legal requirements and policies for data sharing, standardize legal expectations across different jurisdictions (countries, states, institutions, labs).
    - <u>Recommendation</u>: Researchers need to be made aware of such information, if it exists.
3. **Central directory of resources**
    - It is challenging for a researcher to locate all resources that are relevant to their interest.
    - <u>Recommendation</u>: The research community needs a central repository of resources that directs users to specific resources relevant to their scientific interest (both NIH-funded and other resources).
4. **How to incentivize de-identifying and uploading data for sharing, if this is not explicitly required?**
5. **ELSI implications of subject consent**
    - For example, can a single instance of consent cover all possible uses of the subject's data, even if technology continues to evolve, and might permit in the future some uses of the data to which the subject might object?
    - The current model of a single consenting contact may be a barrier to encouraging sharing.
    - <u>Recommendation</u>: standardize definition of subject consent. Would consent include all possible future uses of their data, or should it be restricted to specific use cases defined in a consent form agreement?
6. **Reluctance to share data and defining data ownership**

   - Intellectual property of data (particularly human data) is protected even after publication because of the fear of others using their data to publish something.
   - Recommendation: Need to clarify data ownership and emphasize the value of data sharing.
   - Lack of incentives for data sharing.
7. **Standard data sharing (format, database, and technology)**
8. **Data egress and data analysis in the cloud**
   - Difficult to get data out of dbGaP
   - Uncertainty of what data are computable on the cloud
9. **Dynamic and long-term data sharing plan needed to prevent data silos**
10. **Lack of deep analysis and re-analysis of data, with well-curated data and documentation**
11. **Fear of non-HIPAA-compliance if investigators share their data**
12. **Difference between clinical hierarchy and research hierarchy for data**
    - Clinical data is designed to not be shared, but it would be very valuable.
13. **Challenges of sharing tools, analysis workflows, etc.**

Challenges submitted by a subset of participants (Breakout sessions: 2 & 14 in Town Halls 1 & 2) – What is stopping us from overcoming these challenges? What are the gaps in knowledge?

How would we:

- find datasets that are good exemplars of a problem we want to illuminate
- access main data modalities in the public and controlled domains – for instance, there are not many single-cell, cell-free DNA, methylation, etc. data available.
- develop detailed and standardized metadata
- produce standardized benchmark datasets that are updated regularly (similar to ImageNet for image recognition)
- develop a shared resource for genome/exome analysis that overcomes consent and privacy concerns
- build/create high-quality genotype-phenotype datasets
- improve management of unstructured research data (i.e. publications)
- develop datasets with spiked-in genes, known quantities of molecular features, or dilution studies to be tested against reference datasets
- develop reliable and accurate reference data or sequencing methods to verify epigenetic and epitranscriptomic data
- create a "harness" for problems with defined inputs and outputs where new methods/tools can be submitted and slotted into a publicly supported pipeline for users to access (i.e., an extreme form of benchmarking)

Box: Summary of the "most important" challenges/recommendations from the breakout session.

- Creating and communicating new/existing standards to the community
- Bioinformatics curricula for students at all levels, especially for the rapidly evolving field of genomics and data science.
- Creating protocols for data aggregation, appending, merging, and representation
- Need for novel analytical models and longitudinal data analysis
- Need to develop more tools

Notes

1. **Interoperable standards and ontologies, benchmarking, and communication of standards/ontologies to research community**
    - Interoperable standards with minimal requirements.
    - Better phenotype/metadata standards are needed.
    - There is a lack of community-supported ontologies for relevant entities such as traits, phenotypes, and diseases, especially among journals ("gatekeepers").
    - More sophisticated skill sets for more sophisticated datasets are needed (e.g. computer programming, ontology expertise) required to make use of those sophisticated structures for representing metadata, accessing data, etc.
    - Provenance and identifiers are important metadata, and yet it seems that we are faced with having to create NEW data fields, new standards in order to describe the data that is being captured.
    - Creation and integration of standards without broad adoption by and communication with the rest of the community.
    - Difference between data unification v. aggregation is an issue in standardization.
    - Standardization across data modalities would be ideal, though maybe not possible.
    - Standardization/normalization challenges start from a representation problem.
    - Recommendation: There should be incentives from journals, funders, academic institutions, regulations, for standards cooperation and adoption.
    - Recommendation: Requirements by journals to submit data to repositories.
2. **Lack of understanding of consent and privacy concerns for data sharing**
    - What can and cannot be shared without explicit consent? What can be shared in certain contexts?
3. **Managing unstructured research data using ontologies and metadata standards**
    - The community needs better management of unstructured research data and better ways to describe the fields in biology. Publications should be tagged with relevant terms.
    - Ontologies are needed, and formal education on ontologies may be required.
    - Requirements and guidelines for metadata capture are not clear.
    - Recommendation: Minimum requirements for datasets for journal publications.
4. **Centralized repositories needed with firm requirements to share data to those repositories**
    - Data submission to such repositories should be made easier.
    - Training for data submission process would be helpful.
    - Investment in staff focused on acquiring data submissions may be useful.
5. **Demand for clinical data access and knowledge, database search, tool development, and standards**
    - Clinical standards, terminologies, integration, and interoperability are needed.
6. **Training, education, and career**
    - There is a lack of verified, tested approaches to integrate data science training within standard bioinformatics curricula.
    - Few bioinformatics PhD programs exist.
    - Computer science PhDs specializing in bioinformatics end up in biology, not computer science, departments, which creates isolation between the two fields.
7. **Genotype-phenotype data and infrastructure needed for the public**

  - Human pangenome, high quality reference genomes, central repository for phenotype data.

Challenges submitted by a subset of participants (Breakout session 3 in Town Hall 1) – What is stopping us from overcoming these challenges? What are the gaps in knowledge?

How would we:

- develop accurate phenotype data for outcomes that matter?
- develop universal standards in phenotyping?
- encourage phenotypic data to be shared in the same way as genomic data?
- generate more and better genomics information?
- better coordinate standards?

Box: Summary of the "most important" challenges/recommendations from the breakout session.

- FAIR datasets, including harmonizing data
- Programmatic access to phenotypic data
- Anonymizing phenotypic data

Notes

1. **Making (raw) data available**
   a. There is no common source for raw data. Raw data often obtained via e-mail.
2. **FAIR data requirements are not yet widely adopted**
3. **Definition before standardization and anonymization of phenotype data**
4. **Machine learning models to map genotypes to phenotypes are needed**

Challenges submitted by a subset of participants (Breakout sessions 4 & 16 in Town Halls 1 & 2) – What is stopping us from overcoming these challenges? What are the gaps in knowledge?

How would we:

- unify and integrate statistical data, topological (2D), Hi-C (2D), and 3D/4D structural (computational/imaging) data, whether the data is obtained from experiment or in theory
- improve computational and functional methods for combining different types and fields of data (e.g., microbiome data, genomic sequence, expression data, and proteomics)
- manage data collected from different sequencing platforms, which may require different computational and statistical methods
- effectively integrate different modalities of genomic data
- identify and integrate all genomic data sets that might be relevant for a particular analysis
- standardize the entire process—from consenting, consent/data sharing, to dataset metadata and data dictionaries, to accessing existing data set processes

Box: Summary of the "most important" challenges/recommendations from the breakout session.

- NHGRI collaboration with other institutions and agencies on data resources and sharing
- Data integration from individuals with the same phenotype and ability to document
- Time resolved phenotype crossed with community predisposition haplotypes
- Standardization of all elements in the process, from consenting, consent/data sharing, dataset/metadata/data dictionaries, to data access.
- Creating and making accessible standardized datasets with standardized metadata that can be shared internationally

Notes

1. **Difficult to find and access available datasets across separate forms and platforms, both domestic and international**
    - Recommendation: People are not putting data on EHRs for research, so it would be helpful to identify structured elements that are useful and develop standardized phenotypes.
    - Simplifying the dbGaP process would streamline the data access requests.
    - Can we stream data out of buckets to enable access to public data?
    - No consistent way of advertising that datasets are available, and advertised datasets are not where they are supposed to be.
    - Proper metadata and data dictionaries would help correlate datasets – automated methods and tools to generate metadata would help.
    - Need detailed documentation of various platforms.
    - Recommendation: NHGRI should work with other institutes and agencies (including international efforts) to get the data out there.
2. **Need data integration methods that are agnostic of models/formats/general representation**
3. **Lack of machine learning methods for data governance of SNGS data**
4. **Inconsistent consents for data sharing, and patient consent needs to be streamlined**
5. **Data access requests are not standardized or streamlined**
6. **Dataset quantity and quality are not equal across diseases (e.g., more chronic diseases are more "manageable" and have less access to resources)**
7. **Gene-environment data is lacking and important**
8. **Lack of data from underserved/underrepresented populations to diversity research datasets**
9. **Expertise for data analysis and standardization**
    a. This is especially important in under-sourced regions.
10. **Inability to ship samples across borders, need data sharing agreements in place**
11. **Need to align scientific/financial/business incentives (e.g., Bill and Melinda Gates, WHO, GA4GH, HHS, NGOs)**

Challenges submitted by a subset of participants (Breakout sessions: 5 & 17 in Town Halls 1 & 2) – What is stopping us from overcoming these challenges? What are the gaps in knowledge?

How would we:

- develop new cost-effective technologies to measure meaningful biology that cannot be easily modeled?
- develop functional analysis that follows the pace of discovery of new disease-related genes?

- develop mature methods for identifying causal relationships from genomic variants to phenotypes at various scales?
- develop systems that can learn and apply the specific information relevant at a variant-gene-phenotype level without becoming an automated pipeline where everything ends up classified as unknown significance?
- simplify the mathematical modelling of biological events? The relationships between cellular demands and genome-supplies for phenome progression remain largely unclear.
- develop more sophisticated models that consider interactions or complex inheritance

Box: Summary of the "most important" challenges/recommendations from the breakout session.

- Lack of diversity in research and clinical populations
- Sophisticated models that consider interactions and complex inheritance
- Improving communication between disciplines and providing more training programs for computer science

Notes

1. **Biological mechanisms and interactions are complex – There is much to be discovered in basic science.**
2. **Data standards**
    - Need standard/consensus for gene names or other basic variables, phenotypes, exposures, etc.
    - Centralizing and standardizing reported results, though that requires manual curation to ensure that aggregated information is accurate.
3. **Collaboration culture**
    - Lack of large-scale, meaningful collaborations among labs working on similar problems.
4. **Patient/research subject privacy and confidentiality and the consent process**
    - Difficulty in balancing data access, sharing with patient / research subject privacy and confidentiality.
    - Potential for loss of confidentiality- with implications for consent process. Do we include these risks as part of consent, or develop new technologies to protect confidentiality?
    - Extent of consent is unclear; communicating the level of data sharing in consent process.
5. **Lacking in diverse populations and statistical power**
    - Data need to be rich and granular for a very large number of individuals in order to find signals, including phenotype data, genetic data, and environmental variables.
    - Data in current databases are not representative of diverse populations; models we are developing cannot be utilized for everyone at this point.
    - None of this can be done without adequate representation from diverse populations, standardization of data collection, categorize populations, etc.
        - Mathematical modeling will be biased toward European ancestry (racially white) patient populations.
        - Gene-disease discovery missed opportunities.
        - Causal variant-phenotype relationships at large scale cannot be ascertained without accounting for population-level and family-level confounding.
6. **Missing data, e.g. hospital data points, events**
7. **Need more sophisticated machine learning models**

    - Models that integrate human judgement.
    - Models that consider interactions and complex inheritance.

8. **Lack of funding and expertise for methods in statistical analysis of networks**
9. **“For advancing the integration of different scales of analysis, we are currently lacking standards of truth”**
10. **Difficult to find a publishing home for interdisciplinary work**
    - Inconsistent ontologies between fields and their journals, different standards for publishing research work.
11. **Resources and best pipelines are not “findable”**
12. **Investigators are budgeting only for production of data but not computation**
13. **Lack of training in programming skills for medical/graduate students**
14. **Lack of infrastructure at some institutes, need some guidelines on the computation resources needed for certain research**

Challenges submitted by a subset of participants (Breakout sessions 6 & 18 in Town Halls 1 & 2) – What is stopping us from overcoming these challenges? What are the gaps in knowledge?

How would we:

- combine data generation AND development of new methods that can interpret genome sequence using that data?
- ensure that method-development is not treated as an add-on and is considered a process driving data science innovation?
- develop guidelines that encourage uniformity in developing methods and applying them in clinical and research genomics?
- experimentally ascertain sensitivity and specificity for each new method developed?
- validate the performance of developed new methods?
- help researchers understand computational methods and how to develop them? We need clear resources on how current methods work and their limitations.
- develop mature methods for selecting which experiments are likely to be most informative for elucidating a biological system of interest?
- develop incentives, funding, recognition of work/effort?
- show proper recognition of alternative career paths (data scientists) via K awards?

Box: Summary of the “most important” challenges/recommendations from the breakout session.

- Incentives, funding, recognition of work/effort
- Recognition of alternative career paths via K awards
- Benchmarks and gold standards, as well as metrics for evaluation

Notes

1. **Siloed work between different communities**
    - Communication gap between data generators and methods developers, clinical and research labs, etc.
    - <u>Recommendation</u>: Funding for joint work.
    - Academic software development is also siloed; should encourage the development of reusable components like SDK, APIs.
    - Lack of reproducibility and validation, as well as standardization and utility of tools.
2. **Publishing venues for computational methods**

  - High impact journals are not as friendly to computational articles (pure methods), so there is a lack of publishing venues for methods.
  - Issues with the review process to carefully look at the datasets.
  - Recommendation: code review upon publication.
3. **Issues with data access**
   - Limited publicly available data that is easy to access.
   - It is too time intensive to access data through dbGaP, and the process needs to be redesigned.
4. **Standard datasets, tools, and analysis methods and benchmarking**
   - Lack of a gold standard datasets and analysis methods. Lack of developed and user-friendly suite of tools for testing sensitivity and specificity of analyses.
   - Uniform testing and analyses across the US.
   - Validate performance with outputs that give true physical quantities, rather than scores or rankings.
   - Need better benchmarking datasets and standardized benchmarks for methods.
   - Baseline implementations are hard to develop and make portable.
   - Need correct, large, uniform, and unbiased database development to train machine learning algorithms.
   - Recommendation: tools documentation that shows what software is available for use, how well it is maintained, etc. Essentially, tracking software usage.
   - Recommendation: Incentives, funding, credit for tool development.
5. **Training for post-grad researchers**
   - Computational courses and resources should be available to post-grad researchers.
   - Need platform for sharing user experiences with tools/workflows.
   - Training on how to use tools/workflows.
   - Training in methods development (incl. data science, statistics, machine learning, algorithms, programming, etc.).
6. **Overcoming logistical and organizational work to pull together resources**

Challenges submitted by a subset of participants (grouped into breakout sessions: 7 & 19 in Town Halls 1 & 2) – What is stopping us from overcoming these challenges? What are the gaps in knowledge?

How would we:

- develop the capability to sequence multiple chemical modifications?
- understand molecular biology in its spatial environment, that is, how does a gene's spatial context affect its regulation?
- develop a standardized language and associated metrics for chromatin folding and structural descriptors?
- fill in the gaps in knowledge in chemical modification in DNA, RNA and protein?

Box: Summary of the "most important" challenges/recommendations from the breakout session.

- Can a standardized language and associated metrics for chromatin folding and structural descriptors be developed/devised?
- Gap in knowledge in chemical modification in DNA, RNA, protein
- Continued NHGRI support of large grants to bring collaborations together and disseminating research findings to greater community through training opportunities
- Organization around large-scale data generation, announcing new datasets, data accessibility
- Diversification of funding across all institutions. Increased participation in areas of expertise, with more narrow funding scope to allow this, from more institutes.

Notes

1. **Standardize spatial context in molecular biology and sequencing technology**
   - There is a need for methods to record the spatial context and representing it in standardized, faithful, accessible manner.
   - Standard sequencing technologies will allow us to fill in the gaps for how to improve current technologies.
   - Barriers include organizing people to develop the standards, getting producers and users involved, methods for broad sharing and computing.
2. **Cultural difference between traditional genomics and new methods**
   - Veterans in traditional genomics may tend to ignore recent work in epigenetics.
3. **3D/4D genomics technology development.**
   - It is challenging to represent genomics in 3D+ contexts. The current schema doesn't cover epigenetics sufficiently.
   - Development and dissemination of technology are essential.
   - Each molecule will require specialized, new, and advanced technology.
4. **Modest understanding of underlying biological basis of spatial transcriptomics**
   - For example, how a single molecule interacts with multiple molecules in the genome.
5. **Need to establish a mechanism to create the economies of scale needed**
   - This includes infrastructure and developing efficient logistics for sample sharing.
6. **Informing clinicians of new technologies**
7. **Community task forces led by PIs to address questions/knowledge gaps**
   - Community can know what questions to pick up and work on.
   - Sequencing providers can support efforts to data production and data scientists can support analysis efforts.

Challenges submitted by a subset of participants (grouped into breakout sessions: 8 & 20 in Town Halls 1 & 2) – What is stopping us from overcoming these challenges? What are the gaps in knowledge?

How would we:

- address the problem that current computational methods do not work well for high dimensional low sample size gene expression data sets?
- deal with small data sets?
- standardize transcriptome profiling? Single cell and bulk profiling of alternative transcripts is yet to be perfect.
- describe the nature of gene expression level when inputting transcriptome data?
- reduce the uncertainty associated with genomic testing results?
- reproduce analyses done in publications?
- address the problem that current computational methods do not work well for high dimensional low sample size gene expression data sets?

Box: Summary of the "most important" challenges/recommendations from the breakout session.

- Reproducible analyses in publications
- Reducing uncertainty associated with genomic testing results
- Develop novel statistical approaches for high-dimensional, low sample size gene expression datasets
- Innovative incentives to encourage collaboration around data sharing
- Addressing metadata challenges, including dataset annotation

1. **Establishing standards (best practices, tools, computing environment) for reproducible analyses**
   - Create standards, a possible committee/community that comes up with standards and best practices.
   - Educate scientists on using these standards and best practices which should be adhered to in order to publish.
   - Making everyone use the same tools and tools sets for analyses of data - changed if computational languages.
   - Provide an environment for the people who are doing the experiment to replicate it to reproduce it.
   - More discussion for current computational tools that are available for analysis.
   - Data generation approach should be well-described and shared with the public.
2. **Metadata challenges**
   - Issues of protocol used, study design, technical aspects.
3. **Unannotated datasets**
4. **Other factors in ensuring reproducible science**
   - For the highly impactful studies, do a parallel replication study to determine if reproducible.
   - Scientists are busy rushing from deadline to deadline, once an analysis is about to be published or published, have teams of "professionals" that go and assist the scientist to get their analyses factored so as to be reproducible rather than rely on a grad student about to leave.
   - Recommendation: require publications to have accompanying code on GitHub or another repository to share scripts/data.
5. **New computational methods needed for high-dimensional, low sample size gene expression datasets**
   - Additional types of information can be used (integromics).
6. **Single cell technology should be more developed**
   - Needed to have standardized transcriptome profiling.
   - Single cell and bulk profiling are subject to total available budget and research design/purposes.
7. **Need policies for reassessing genomic test results, including delegating the reassessment and communication of results of reassessment**
8. **Access to supercomputer systems and cost-effective IT systems**
9. **Education for users of genomic tests to understand uncertainty around risk prediction and variant interpretation**

10. **Need an online website/tool that includes models developed by each group and the corresponding datasets to be shared with the whole research community**

Challenges submitted by a subset of participants (Breakout sessions 9 & 19 in Town Halls 1 & 2) – What is stopping us from overcoming these challenges? What are the gaps in knowledge?

How would we:

- develop computational and statistical methods that are more generalize-able rather than being developed in a lab - for a lab?
- provide software solutions that are transferrable, updatable, and easy to implement across multiple platforms and architectures?
- create resources that everyone can access and use (rather than downloading the data and analyze it separately on their own computer)?
- make the methods very interactive for the biologist with limited computational and statistics skills?
- create a multi-omics platform that will bring things together in one place?

Box: Summary of the "most important" challenges/recommendations from the breakout session.

- How to make the methods very interactive for the biologist with limited computational and statistics skills?
- Note: We think that emphasizing the collaborative element between the life sciences and computational sciences is the most important point.
- Multi-omics platform to bring things together in one place

Notes

1. **Interoperability between cloud environments**
    - Having different cloud environments makes transferring across systems hard.
    - Don't have many datasets connected to the cloud environment.
    - Connecting disparate data sources, especially for non-experts in one omics field connecting with other data.
    - Recommendation: Multi-omics platform to bring everything together in one place, where all software and data can plug into.
2. **Generalizable tools and standardization**
    - The field is moving very quickly, standardization would be hard.
    - Computational methods need to be made more generalizable.
    - Requirements to include script in publications. Is there a way to make codes transferable and generalizable (i.e. may run on one Java but not on another version of Java)?
    - Lack of standardization regarding data sets.
    - Need a standard in data analysis and data generation.
    - Standardization is hard because different labs have different equipment.
    - Individual standards are not effective, need consortiums. Collaboration rather than individual algorithm development.
    - Take into account different countries, populations across the globe.
3. **Competitive culture obstructing collaboration**

   - Collaboration is underappreciated, due to desire to be first author etc. Issue of balancing individual credit and progress.
   - People are rewarded for creating new models and tools. Misalignment of scientific incentives.
4. **"Multi-omics" is not well defined**
5. **Lack of usable interfaces for biologists and clinicians**
   - Need reliability and reproducibility in the clinical setting, which academic models sometimes lack.
6. **Cross-training between wet lab/clinicians and data scientists**
   - Biologists and clinicians should learn more about data science, and data scientists should learn more on the basic biology.

Challenges submitted by a subset of participants (grouped into breakout sessions: 10 in Town Halls) – What is stopping us from overcoming these challenges? What are the gaps in knowledge?

How would we:

- develop better productive collaboration between the computation/statistics community and the genomics/geneticist community?
- reduce the fragmentation of effort across the scientific community?
- make connections to people in genetics fields to foster interdisciplinary collaboration?
- make it easier for statistician and computer scientist, especially non-native speakers, to understand the complex genomic language?
- find/develop reviewers with expertise in the area of proposed projects, who can provide insightful comments and suggestions?
- recruit reviewers who will benefit from the methods, e.g., epidemiologists focusing on genetic or epigenetic studies on certain health outcomes?
- develop better productive collaboration between the computation/statistics community and the genomics/geneticist community?

Box: Summary of the "most important" challenges/recommendations from the breakout session.

- Education and training of different stakeholders on what are an individual researcher's expertise, knowledge, and skill sets
- Developing better and productive collaboration between the computation/statistics community and the genomics/geneticist community

Notes

1. **Fostering collaboration between computational and genomics/genetics communities**
   - Challenge in interpretation of data types - interpreting experimental techniques/technologies (caveats, strengths and weaknesses).
   - Challenge in educating users on resource: richness of content and resource usability/querying.
   - Challenge in educating biologists, geneticists, etc. how data analysis techniques work and what are important considerations for interpreting analysis results.
   - <u>Recommendation</u>: HG-organized effort to bring issues to broader scientific community and bring researchers together from different domains.

    - Recommendation: A coordination platform for cross-disciplinary investigators to talk about potential problems and solutions. Can match between problems/solutions across fields to identify reviewers with expertise in each area.
    - Recommendation: 'Genomic Salon'—informal talks/seminars within institutions.

2. **Reducing siloed work across the scientific community**
    - Gaps in knowledge: Lack of a comprehensive database/resource that catalogs all scientists/researchers and their topics of focus.
    - Requires uniform format/guidelines to capture and report researchers and their focus (lack of uniform language to communicate this information).
    - Need better discoverability of common efforts.
    - Good to share data, but also equally important is sharing of methods (and parameters) to produce or analyze such data.
    - Short projects do not leave time to develop terminologies and connections; need longer-term collaborations.
3. **Making connections in the genomics community to foster interdisciplinary collaboration**
    - Challenge to recruit community to participate in a workshop (like a UniProt workshop).
    - NIH program officers should communicate significance to community members.
    - Need good ways to educate how resources are of relevance to the research community.
4. **Cross-education between statisticians and computer scientists, and genomicists**
    - Need long-term interdisciplinary cross-education between genomicists and data scientists (important to establish collaborations with knowledgeable people).
    - Lack of resources for cross-education.
5. **Finding reviewers with the appropriate expertise to provide comments and suggestion**
    - Again, lacking a comprehensive catalog of researchers and their focus/expertise.
    - Maybe would benefit from an ontology of research topics.
    - SRO may select reviewers by suggestion.
    - Keywords are too few and generic for applications, but do not want narrow expertise representation on review panels.
6. **Recruiting reviewers who will benefit from the methods, e.g., epidemiologists focusing on genetic or epigenetic studies on certain health outcomes?**
    - Conversely, we have a need for a database of needs for researchers; if researchers could express and disseminate their technological needs, that would help.
    - Challenge embedded in the language issue.

Challenges submitted by a subset of participants (Breakout session 11 in Town Hall 1) – What is stopping us from overcoming these challenges? What are the gaps in knowledge?

How would we:

- develop competencies across computing, statistics and genomics?
- find and/or develop a workforce with the right skill set (computer science AND genomics)?
- address the large computational burden of data science and the high bar of entry (supercomputer needed) to increase diversity among computational and statistical genomics researchers?
- attract bright data scientists away from lucrative careers in industry?
- develop more professionals with computational and statistical skills?
- develop domain knowledge in genetics for quantitatively trained researchers?
- support the training of instructors, or helping them develop modules for their classes?

- train the trainers so that they can train students by letting them choose research topics of interest?
- make sure that researchers on statistical genomics or computational genomics are considered mainstream in their own discipline?
- provide education resources to train people, including courses and instructions for trainees (high school students, students with computer science background) to get to know the fundamentals of biology and genomics?

Box: Summary of the "most important" challenges/recommendations from the breakout session.

- Appropriate incentives for data scientists to continue working in the field of health care
- Need more education resources for training, including courses and instructions for trainees (high school students, students with computer science background, etc.) to learn the fundamentals of biology

Notes

1. **Lack of incentives for training**
    - In most cases, for service units, teaching is not part of the mandate or JD, and graduate students may not get credit for teaching.
    - Recommendation: Holding workshops and courses that help develop skills, incentivizing alternate career paths.
2. **Funding, grantsmanship, and budgeting**
    - Easy to get NIH funding to generate data, not so easy to get funding for analysis of data.
    - Need help at the institutional level to develop proper skills and budget appropriately for work.
3. **K Grant for Data scientists? -- should acknowledge Jason Williams for this one :)**
4. **Data science not seen as a "fundamental" or valuable field**
    - One possible knowledge gap: the field of data science is not yet as "fundamental" as other fields for which it is easier to attract/fund scientists. We do not yet have a common universal "model" for what data science is that we could study in ways that are traditional for experimental approaches.
    - Is a 'genomic data scientist' useful if the science doesn't support it?
    - Genomic data management - is there a field for this?
    - Pushing to create a PhD to specialize in genomic data science. Keep this as part of the broader 'data science' field to take advantage of improvements in non-genomic data science fields.
    - Genomics data science is not as valuable as data science in other fields (finance, targeted marketing). What is the value that genomics data science will provide?
    - Is there a clear path for developing a career in genomic data science that is financially incentivized and stable?
5. **Lacking education resources to train people with various backgrounds in statistics/computer science and in fundamental biology**
6. **Low salary compared to industry jobs**
7. **Lack of data/script sharing and standards for dataset annotation**

Challenges submitted by a subset of participants (grouped into breakout sessions: 12 in Town Halls 1) – What is stopping us from overcoming these challenges? What are the gaps in knowledge?

How would we:

- identify a "Grand Theory" or a "Grand Challenge" of genomics that seeks physical laws/principles that govern the field?
- develop good hypotheses that generate better models to depict the biological questions?
- build multidisciplinary teams to formulate hypotheses and tackle questions using a variety of different approaches?
- encourage cross-domain expertise across biological/genomic subfields?
- encourage a multi-dimensional and holistic understanding that fully comprehends the multiple levels of "interactions"?
- enable cross-discipline collaboration and communication?

- Computational representation of the human phenome for maximizing our genomic interpretation in order to realize the promise of precision medicine.
- Must have cross-disciplinary collaboration with researchers in basic research and across the globe. Genomic health is a global endeavor.
- Building multidisciplinary teams to formulate hypotheses and tackle questions using a variety of different approaches

Box: Summary of the "most important" challenges/recommendations from the breakout session.

1. **Variant catalog, standards, analysis, and methods**
    - We need to expand the catalogue of variation – CNVs, etc.
    - Underserved/underrepresented populations are poorly represented in variant catalogs
    - No standardized variant representation across systems, lack of standardization in pipelines and their comparison, and lack of transparency/benchmarking for comparison. This is particularly true for variant interpretation!
    - More sophisticated algorithms for structural variant analysis are needed.
    - Machine learning could be used to help triage variants into functional/impactful classes.
    - Experimental methods for assessing function is not clear for all.
2. **Lack of currency for tools in accessing current database knowledge**
3. **Resource integration, shared model across diseases across knowledge sources needed for standard genomic interpretation**
    - Integration of various resources to one focused question (no set basis/standards between them).
    - Essentially, we need a better shared model (standards) of disease across knowledge sources in order to standardize genomic interpretation (while still being dynamic).
4. **Computational representation of human phenome to maximize genomic interpretation in a way that is useful for precision medicine.**
5. **Patient data privacy**
    - Challenges in identifiability and consent when you bring more data types together, which is really required for maximizing interoperation.
    - Ethical challenges in labelling of pops/groups and protecting your samples (and in creating trust with these populations).

   - Patients need rights (and education) about what, where, and when to share their own (multimodal) data (sometimes we take their rights away, sometimes we consent without adequate education).

6. **Science/research language needs to be "interoperable"**
7. **Funding scope and exclusivity**
   - It is hard to break into the "club" of NHGRI collaborations without being funded by them first. There needs to be an "it takes a village" mentality to be instantiated at HG. Other programs funded by NIH or globally are relevant to HG, but they are not always leveraged if they are not HG-funded.
   - Genomic/cross-disciplinary/cross-disease research often falls to the side when others say, "HG will do that", but then HG doesn't address it because it is too cross-disciplinary or leverages non-HG resources.
8. **Cross-domain expertise across biological/genomic subfields.**
   - Genomics is truly multidimensional, and the lack of complete understanding about interactions results in difficulties formulating a holistic view.
9. **Cross-discipline collaboration**
   - There is a divide between basic and translational science.
   - Need cross-domain expertise across biological/genomic subfields.
   - We need computable access to the basic research data, not just searching for functional validation evidence. Conversely, we also need to provide computable and accessible human data to basic scientists.
   - Investigators do not know *who* to reach out to, or even what to ask when thinking about a very different field of interest. There needs to be more funded opportunities to get together across disciplines BEFORE funding calls and BEFORE proposal writing happens. Most conference grants and programs either don't fund enough people/gatherings or don't include enough cross-disciplinary participants.
      i. Example: Why don't we collaborate with folks in geography and their GIS systems to use in combination with our genetics work?
   - Recommendation: Collaboration between people to explore different hypotheses and different aspects of data analysis through the formation of WGs.