Perspectives in Comparative Genomics and Evolution Meeting Report

On August 15-16, 2019, the National Human Genome Research Institute (NHGRI), the National Science Foundation (NSF) and the United States Department of Agriculture (USDA) sponsored a meeting, *Perspectives in Comparative Genomics and Evolution,* at the Bethesda North Marriott Hotel & Conference Center in Rockville, Maryland. This meeting's objective was to obtain a big picture look at the current state of the field of comparative genomics with a focus on commonalities across genomic investigations into humans, model organisms (both traditional and non-traditional), agricultural species, wildlife species and microbes.

Objectives were threefold:

1. Identify areas of synergy;
2. Identify gaps in knowledge and resources that need attention and development;
3. Recommend areas of focus for comparative genomics as a whole, and specifically to NHGRI, with relevance to understanding human health and disease.

In addition to the goal of all three agencies to work together to advance the field of comparative genomics, there was the NHGRI strategic planning goal of identifying what part of this field (if any) belongs with NHGRI and would be at the "forefront" of comparative genomics.

Opening and Keynote Speakers: Representatives from NIH (Jennifer Troyer, Program Director NHGRI), NSF (Donal Manahan, Director, Division of Integrative Organismal Systems) and USDA (Adele Turzillo, Director, Division of Animal Systems) provided welcoming remarks as well as a mandate for the meeting to focus on big picture ideas for the future. The first keynote speaker, ***Harris Lewin*** (Chair of Earth BioGenome as well as member of the National Academy and Endowed Chair and Distinguished Professor of Evolution and Ecology at UC Davis) set the stage in this regard by providing a perspective on the history of genomics and a vision for a National sequencing program to sequence all the species in the US. He argued that the return on investment would far exceed that of the Human Genome Project and need the same audacious effort and commitment. ***Beth Shapiro*** (Professor of ecology and evolution at UCSC and HHMI Investigator) broadened this view by addressing the utility of ancient DNA and comprehensive DNA sampling to provide evolutionary, environmental, and ecological contexts.

Each of the following sessions consisted of several short talks that address a particular common problem in genomics followed by a longer facilitated discussion where participants were asked to consider unmet needs and long-term vision for this area of investigation.

Session 1: **What Is A Reference Genome and What Other Data Types Are Needed?**

Speakers: ***Erich Jarvis, Huaijun Zhou, Robin Buell, Carlos Bustamante***
Facilitators: Elinor Karlsson, Sofie Salama

This session addressed the current state-of-art in reference genomes and the balance between what is ideal and what is "good enough". The reality is that circumstances will be different depending on the organism because of availability of samples, funding, the research community, and the complexity of the genome. Speakers emphasized the scientific value of a high-quality reference genome and the costly mistakes imposed by inadequate references; the increased capacity for including genome annotation and regulatory information; the utility of adopting universal standards alongside the need to avoid a "one-size fits all approach" given the variability of genomes and the research communities who rely on them; and the vast potential for comparative genomics to inform biology and biomedical research.

Unmet needs:

- High quality reference genomes across the tree of life
- Significant investment in infrastructure and resources, including appropriate review mechanisms
- Adoption of standards, especially for functional annotation and representation of alternative structures; confidence measures for annotation (gene models for example)
- High quality samples with HMW DNA, preferably with associated phenotypic data
- Reliable and affordable methods for chromosome level telomere to telomere assemblies
- Timely, accurate data released with appropriate data structures and associated metadata

Long-term vision:

- A national sequencing program
- High quality reference genomes with common protocols, QC, and data structures for many (or all) species with:
    - Representation of population sequence and structural variation
    - Associated phenotypic data from sequenced individuals
    - RNAseq data, epigenomic marks for common tissues, life stages, etc.
      Coordinated data across species and ways to perform cross-species analyses

Session 2: **Integration and Usability of Comparative Genomics Data**

Speakers: ***Bonnie Hurwitz , Melissa Haendel, Beth Dumont***
Facilitators: Lucia Carbone, Adam Arkin

This session explored the multiple genomic data types, how to increase the interoperability of data across species as well as new tools and resources needed to facilitate data use. Speakers emphasized the needs for data to be FAIR and how to accomplish that with frictionless data packages, standardized ontologies, appropriate associated metadata, and dockerized containers; the need to integrate phenotypic data so that we can relate form, function, and dysfunction in order to interpret the genome; and a caution that genomes are dynamic and that individual and population level variation can impact many traits and phenotypes that we might think of as fixed, including mutation rate, in ways that we do not yet understand.

Unmet needs:

- Interoperability of data
    - Metadata curation
    - Standards
    - Vocabulary/ontologies with translation
    - Quality scores
- Better ways to characterize and harmonize phenotypic data, especially for conditional phenotypes
- Appropriate repositories for multiple data layers
- Universal use of dockerized code and pipelines, protocols.io, digital object identifiers (DOI), etc.
- Cultural incentives for data (and tool) curation and sharing
    - Career advancing recognition for providing useful data
    - Common expectations from journals, funders, and reviewers
    - Meaningful, reviewed, data management plans with certified data repositories

Long-term vision:

- "Trait database" for all species
- Federated methods to combine data and find "touchpoints" for deep phenotypic homology across species

Session 3: **Mechanistic Function, Variation and Phenotype**

Speakers: ***Emma Farley, Yoav Gilad, Tom Spencer, Joanna Kelley***
Facilitators: Claudio Mello / Charles Danko

This session discussed ways to untangle genomic function, inform our understanding of biological mechanisms, and connect genomic variation to phenotypes. Speakers described methods to determine universal regulatory grammar, the utility of comparative iPSC lines to understand developmental pathways, the use of natural systems and the utility of comparing natural systems and model organisms in order to disentangle complex phenotypes.

Unmet needs:

- Untapped resource of companion and zoo animals
- Standard quantitative environmental measures
- General-use genome manipulation tools
- High throughput phenotyping
- More studies that leverage both natural and manipulated systems

Long-term vision:

- Trans-agency funding of "One Health" research
- Connected ecosystem that allows flexibility in models to address specific biological questions

- Moon Shot: One system with a concerted trans-agency effort to have ALL the data and understand everything

Breakout sessions: The breakout sessions were designed to be an opportunity to discuss areas of overlap in our different communities (medical, agricultural, and wildlife) and think about ways that these communities might increase interactions and synergies to advance this area of science.

- Breakout Session 1: **Population Genomics**
  Facilitators: María Ávila-Arcos, Clare Gill, Elaine Ostrander, Pam Soltis

  Recommendations for Areas of Synergy:
  - Uniform methods for sample collection, sample archiving, and standards
  - The extended specimen - the importance of collecting the phenotypic and ecological information on the samples that are used
  - Leveraging data science to connect what we want to do with the data - everything needs an identifier to connect things
  - One is not enough – high coverage vs. low coverage and careful sampling strategies
  - Putting things in an ethical international context: sampling, funding, training, and work with/return to local populations

Breakout Session 2: **Developmental and Reproduction Genomics**
Facilitators: John Liu, Nipam Patel, Ollie Ryder, Tom Spencer

Recommendations for Areas of Synergy:
- Methods to go beyond genomes and comparative genomics to comparative phenomics and transcriptomics
- AI for understanding enhancer function and timing to move towards predictive understanding of the relationship between genome sequence and body form and function
- Distributed data repositories at multiple scales

Breakout Session 3: **Host/Microbial/Vector Evolution and Interaction**
Facilitators: Hans Cheng, Elinor Karlsson, Lynn Schriml, Sue VandeWoude

Recommendations for Areas of Synergy:
- Fostering interdisciplinary interaction & data sharing, with training for (creative) data scientists
- Focus on intersecting/connecting research projects currently focused on different scales (host to pathogen to microbiome to environment)
- Added value opportunities and projects that support the higher risk, cross disciplinary projects and can be done at minimal additional cost in conjunction with existing funded project (ex: cross-disciplinary postdoctoral opportunities)

- Capture all the information - so that we can capture change over time; response to new pathogens; capture the background state as well as when something exciting happens

Breakout Session 4: **Systems Biology**
Facilitators: Charles Danko, Erich Jarvis, Peter Johnson, James Koltes

Recommendations for Areas of Synergy:

- Concerted effort to identify broadly conserved networks across domains of life
- Programs to foster interdisciplinary relationships and/or training
- Linked trait-genome database
- Programs linking systems biology to synthetic biology

Big Ideas, NHGRI Strategic Planning, and Closing Discussion

The meeting ended with several discussions directed towards the future of the field. Participants voted on priorities, provided ideas, and advised NHGRI on community needs for better telomere to telomere sequencing and analysis methods; higher quality reference genomes for more species across the tree of life; methods for cross-species comparison of genomes, epigenomes, transcriptomes, and traits; better genomic visualization methods; technology development support for synthetic genomics; and NHGRI leadership in trans-NIH support for non-traditional models of health and disease.

Post-meeting, participants were asked to rank the collated big ideas that came out of the discussions. Highest priority recommendations were:

1. Provide trans-agency funding opportunities that encourage interdisciplinary, synergistic research that will take advantage of multiple systems, which might normally be supported by different agencies, to bring new understanding to functional genomics

2. Support new computational methods that leverage comparative genomic information across a diversity of species to relate form, function, and dysfunction in order to interpret the genome (for diagnostics or otherwise)

3. Support a concerted effort to generate comprehensive genomic resources and annotation (genomes, transcriptomes, metabolomes, proteomes, ATAC-Seq, methylomes, in multiple tissues) in a targeted smaller sampling of species across the tree of life.

4. Support a browser (or browser-deployable method packages) to display systematic collections of omic datasets together with computational methods to analyze them across multiple species including:
    a. better, more tractable methods for aligning divergent genomes and visualizing the comparisons

b. comparison of non-coding regions
c. species-level diversity representation
d. functional annotation and ways to meaningfully compare expression and functional elements across species

5. Normalize the use of any organism as a “model organisms” with tools, workstreams, and resources that can be deployed as well as standardized sample collection protocols including collection of phenotype information, data analysis - including sequencing, annotation and alignments, and proper avenues for data storage

Appendix

1. Meeting Agenda
2. Meeting Minutes
3. Participant Roster
4. Dotstormming