Networks of Networks: Sequence, Genomes and People

Owen White
Director of Bioinformatics
Institute for Genome Sciences
University of Maryland Baltimore
School of Medicine
NHGRI Genomic Timeline
A Series of Consortia

1995
2000
2002
2005
2007
2012-2018
Present
Biomedical Research is Large

• Millions of genome-equivalents

• 1,000s of centers

• Gargantuan cloud-based systems

• Abundant resources, e.g.:
  - HMP: $120M
  - BRAIN Initiative $180M
Data Topology is Distributed

• There is no one “genome repository”
  – Imagine: PubMed → 100s of libraries

• National Institutes of Health
  – 100s of Data Coordination Centers, $10^5$ labs, $10^7$ samples, # of files?

• Consider: 1,000s of hospitals
  – human sequencing as an assay
Distributed Data Implications

Puts a high premium on:
• Open access / data release

But this is very hard:
• Discoverability
• Combining datasets
• Reproducibility
# NIH Common Fund Assets

## Table

<table>
<thead>
<tr>
<th>A.</th>
<th>4D Nucleome</th>
<th>GTEx</th>
<th>HMP/HMP</th>
<th>HubMap/P</th>
<th>KidsFirst</th>
<th>LINCS</th>
<th>Metabolomics</th>
<th>MoTrPAC</th>
<th>SPARC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical Data</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whole Genome/Exome Sequence</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transcriptomics</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>P</td>
<td>X</td>
<td>P</td>
<td>P</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histology Images</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Radiology Images</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metatranscriptomics</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metaproteomics</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marker Sequence Metagenomics</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Microbial Reference Genomes</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ChIPseq</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FISH</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATACseq</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hi-C</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ChIA-PET</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proteomics</td>
<td>X</td>
<td>P</td>
<td>X</td>
<td>P</td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KINOMEscan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Metabolomics</td>
<td>X</td>
<td>P</td>
<td>X</td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lipidomics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>P</td>
</tr>
<tr>
<td>scRNAseq</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>P</td>
</tr>
<tr>
<td>Epigenomics</td>
<td>X</td>
<td>P</td>
<td>X</td>
<td>P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Diagram

```
A.   B.  System  Organ  Cell  Molecule
     |     MoTrPAC X X
     |     SPARC  X X
     |     HubMap X X
     |     LINCS  X X
     |     4D Nucleome X X
     |     GTEx   X
     |     KidsFirst X
     |     HMP/HMP X
     |     Metabolomics X
```
Complementary Assets

- Same assets across sites
- Assets useful in combination across sites
- Sites host data associated with core entities:
  - Human genes - link between expression, epigenetic, and variant
- Data linked to concepts
  - Part of the body (e.g. "liver")
  - Patient information (e.g. body mass index, blood pressure)
Complementary Assets

- Same assets across sites
- Assets useful in combination across sites
- Sites host data associated with core entities:
  - human genes - link between expression, epigenetic, and variant
  - Data linked to concepts
    - Part of the body (e.g. "liver")
    - Patient information (e.g. body mass index, blood pressure)

Problem Statement:

- No common electronic specification for assets
- No common specification for asset inventories
- No common transport system, “commerce”
The Challenge: Distributed Data is a Fact of Life

Puts a high premium on:
• Open access / data release
But this is very hard:
• Discoverability
• Combining datasets
• Reproducibility

Unexpected surprise:
These are significant social issues – technical agreement is nearly trivial
Genome Standards Have Always Been Built On Community Engagement

- **Community Members**
  - identify initial set of key stakeholders
  - develop plans to grow the community
  - define contributor and leader roles

- **Communication**
  - project goals, solicit community input
  - match goal to meet community needs,
  - set up mechanism to field community requests

- **Collaborative - Iterative - Development**
  - reuse – recycle – repurpose Existing Ontologies
  - evaluate ontology utility to data needs
  - refine the ontology & establish update process
DO Community
Connecting Disease to Gene, Protein, Variation

Serving Our Community
• Term requests & review
• Integrating rare diseases
• Coordinating development with clinicians
• Providing support for disease curation & annotation
Cross-mapping disease concepts (UMLS), disparate representation of disease across vocabularies (37,988 xref mappings)

- Disease of cellular proliferation
  - Cancer
  - Benign neoplasm

- Genetic disease
  - Orphanet: Rare Diseases

- Disease of metabolism
  - Inherited metabolic disorder

- Disease of mental health
  - DSM - IV

- Disease of infectious agent

- Disease of anatomical entity

ICD9
- Neoplasms
  - (ICD-10)
  - Congenital anomalies

MeSH
- Diseases
  - Mental disorders
  - Infectious and parasitic diseases
  - Endocrine system diseases
  - Immune system diseases

OMIM
- Phenotypes

NCI
- [C] Diseases
  - Neoplasms

DSM
- [F03] Mental disorders

MeSH
- [C01] Bacterial infections and mycoses
- [C02] Virus diseases
- [C03] Parasitic diseases

MeSH
- [C19] Endocrine system diseases
- [C20] Immune system diseases

Cross-mapping disease concepts (UMLS), disparate representation of disease across vocabularies (37,988 xref mappings)
Challenges: Fairness and Trust

• Stakeholders have vested interest in the implementation (read: continued funding)
• Across consortia, no incentives to get in the room
• Prisoner’s dilemma: no one group member can get buy-in from the rest of the group
• Not everyone needs to agree with a decision, but everyone does need to agree with the process for how to make decisions
ORGANIZATIONAL AND COMMUNICATION STRATEGIES
Elements of Success: Open Communication Tools

- Google drive
- Github
- Slack
- Groups.io
- Zoom
- Figshare

Goal: raise openness
Drivers of Success in a Consortium
(and drivers of primate behavior)

Fairness, trust, and “seeing” each other
Elements of Success: Communication Team

• Listening missions (physical travel)
• Do not talk about implementation, listen, take notes
• See what their life is like
• Determine incentives for participation
• Disseminate info
• Buffer between funder

Goal: raise trust, “see” each contributor, promote buy-in
Elements of Success: Working groups

• Vertical and horizontal communication (everyone is seen)

• Decisions should not be based on who is in the room, take notes, disseminate openly
Elements of Success: RFCs

Note: academics are notorious for NOT wanting standards
Requests for Comments are:
• Open
• Iterative
• Binding
• Triangulates on consensus/community agreement
• Incremental engagement --> routine dissemination
• Basis of standards formation
Other elements of success

Increase accessibility
• Use open communication tools
• Record everything
• Disseminate everything
• Publish release cycles
• Instant messaging

Think: football coach
• Personalize contacts
• Liaison with mothership / let people do what they’re good at

Promote: Everyone is seen, everyone contributes
• Examples: consortium-wide meetings, pairwise interactions, recording institutional memory, newsletters, social media
Other elements of success

Promote fairness, open methods devel
• Bake-offs, objective validation of methods
• Agile development $\rightarrow$ frequent demos
• Github software registries

Training
• Empowerment
• Builds social networks
• Test early and often
• Understand usage patterns
Thanks