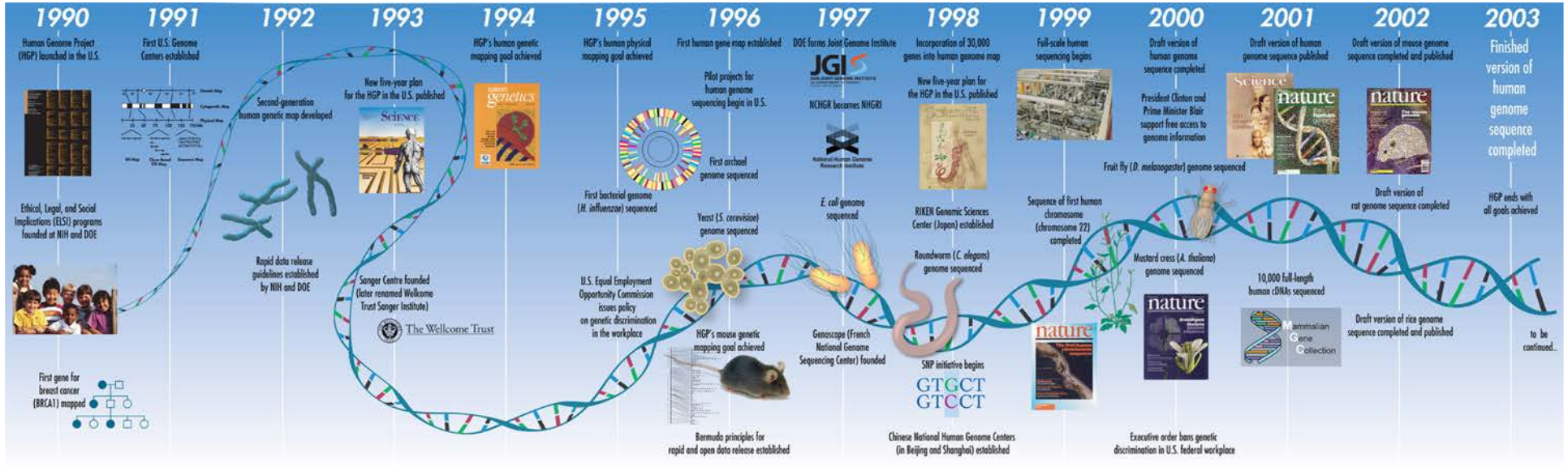
Networks of Networks: Sequence, Genomes and People

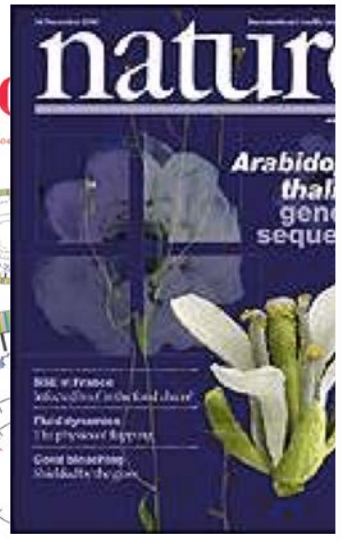
Owen White
Director of Bioinformatics
Institute for Genome Sciences
University of Maryland Baltimore
School of Medicine

NHGRI Genomic Timeline



A Series of Consortia

1995



2000

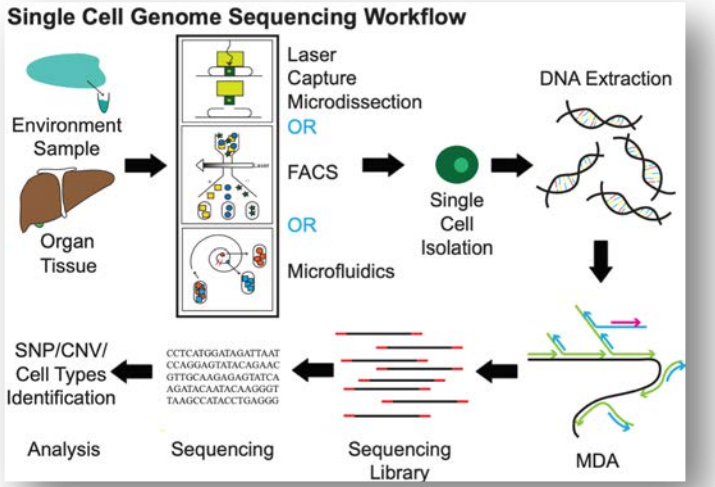
2002

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

A.

	4D Nucleome	GTEx	HMP / iHMP	HuBMAP	Kids First	LINCS	Metabolomics	MoTrPAC	SPARC
Clinical Data		X	X		X	X			
Whole Genome/Exome Sequence		X	X		X			P	
Transcriptomics	X	X	X	P	X	X		P	P
Histology Images					X				
Radiology Images					X				
Metatranscriptomics			X					P	
Metaproteomics			X						
Marker Sequence Metagenomics			X					P	
Microbial Reference Genomes			X					P	
ChIPseq	X					X			
FISH	X			P					
ATACseq	X			P		X			
Hi-C	X								
ChIA-PET	X								
Proteomics			X	P		X		P	P
KINOMEscan						X			
Metabolomics			X	P			X	P	
Lipidomics				P					
scDNAseq				P					
Epigenomics			X	P		X		P	

B.

	Systems	Organs	Cells	Molecules
MoTrPAC	X	X		
SPARC	X	X		
HubMap		X	X	
LINCS			X	X
4D Nucleome			X	X
GTeX				X
KidsFirst				X
HMP/iHMP				X
Metabolomics				X

A.

	4D Nucleome	GTEx	HMP / iHMP	HuBMAP	Kids First	LINCS	Metabolomics	MoTrPAC	SPARC
Clinical Data		X	X		X	X			
Whole Genome/Exome Sequence		X	X		X			P	
Transcriptomics	X	X	X	P	X	X		P	P
Histology Images					X				
Radiology Images					X				
Metatranscriptomics			X					P	
Metaproteomics			X						
Marker Sequence Metagenomics			X					P	
Microbial Reference Genomes			X					P	
ChIPseq	X					X			
FISH	X			P					
ATACseq	X			P		X			
Hi-C	X								
ChIA-PET	X								
Proteomics			X	P		X		P	P
KINOMEscan						X			
Metabolomics			X	P			X	P	
Lipidomics				P					
scDNAseq				P					
Epigenomics			X	P		X		P	

B.

	Systems	Organs	Cells	Molecules
MoTrPAC	X	X		
SPARC	X	X		
HubMap		X	X	
LINCS			X	X
4D Nucleome			X	X
GTeX				X
KidsFirst				X
HMP/iHMP				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)

A.

	4D Nucleome	GTEx	HMP / iHMP	HuBMAP	Kids First	LINCS	Metabolomics	MoTrPAC	SPARC
Clinical Data		X	X		X	X			
Whole Genome/Exome Sequence		X	X		X			P	
Transcriptomics	X	X	X	P	X	X		P	P
Histology Images					X				
Radiology Images					X				

B.

	Systems	Organs	Cells	Molecules
MoTrPAC	X	X		
SPARC	X	X		
HubMap		X	X	
LINCS			X	X
4D Nucleome			X	X

Problem Statement:

No common electronic specification for assets

No common specification for asset inventories

No common transport system, “commerce”

Complement

- 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)

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

The Jackson Laboratory
Leading the search for tomorrow's cures

EMBL-EBI

Samples, Phenotypes, Ontology Team

cognitive atlas

NeuroDevNet

MGI

Serving Our Community

- Term requests & review
- Integrating rare diseases
- Coordinating development with clinicians
- Providing support for disease curation & annotation

Gene Wiki

orphanet

PomBase

Sifem Inner Ear disease

REACTOME

NIF Neuroscience Information Framework

OMIM Online Mendelian Inheritance in Man — Johns Hopkins University

WormBase

efo Experimental Factor Ontology

PRINCETON UNIVERSITY
Lewis-Sigler Institute for Integrative Genomics

FlyBase

Neurocarta

HPO
Human Phenotype Ontology

HIVE

gsc genomic STANDARDS consortium

Cross-mapping disease concepts (UMLS), disparate representation of disease across vocabularies (37, 988 xref mappings)

disease of cellular proliferation
cancer
benign neoplasm

NCI

genetic disease

Orphanet: Rare Diseases

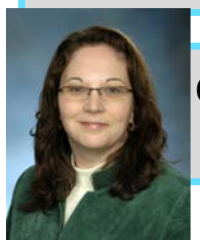
disease of metabolism
inherited metabolic disorder

disease of mental health

DSM - IV

disease of infectious agent

disease of anatomical entity



ICD9 (ICD-10)

neoplasms

congenital anomalies

mental disorders

infectious and parasitic diseases

disease of … system

MeSH

[C] Diseases

[C04] neoplasms

[F03] mental disorders

[C01] bacterial infections and mycoses
[C02] virus diseases
[C03] parasitic diseases

[C19] endocrine system diseases
[C20] immune system diseases

OMIM

phenotypes

phenotypes

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 → frequent demos
- Github software registries

Training

- Empowerment
- Builds social networks
- Test early and often
- Understand usage patterns

Thanks