NCBI Data types, management, monitoring, and outreach activities

Kim D. Pruitt

Biomedical Information Science and Technology Initiative

(BISTI)

November 3, 2016

- Data types, archives and the role of curation
- Data privacy
- Data management
- Metrics of success
- Outreach activities

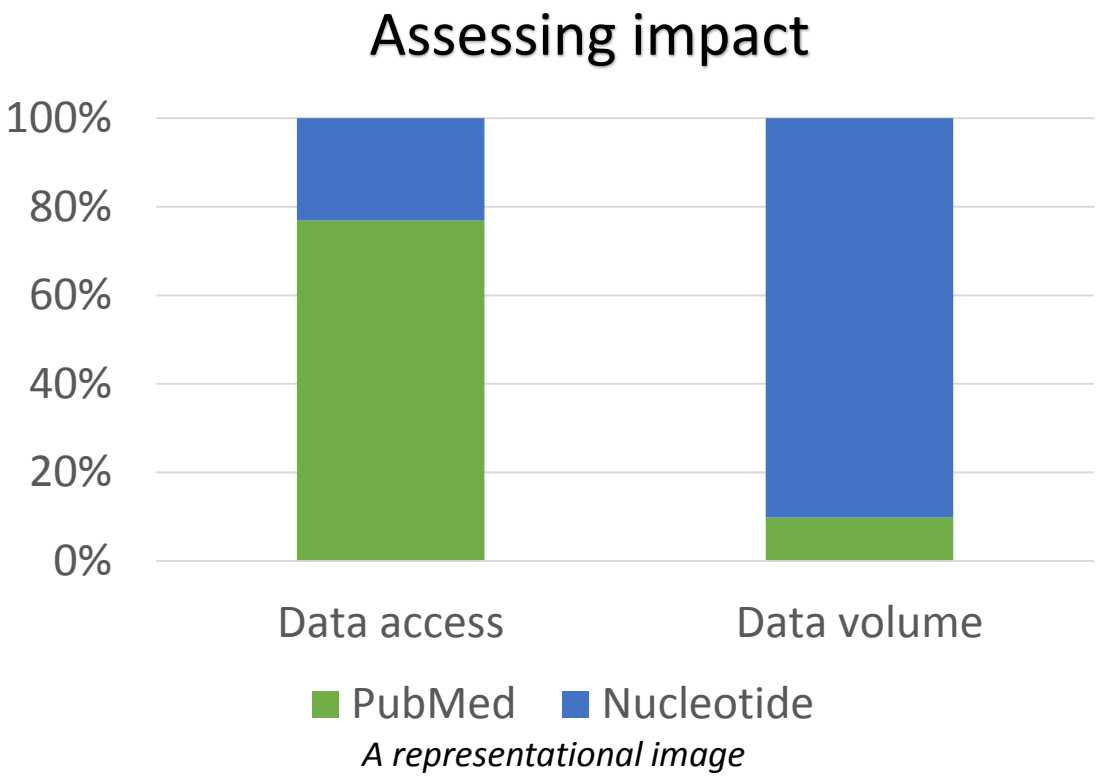
Primary data types

Literature

- PubMed
- PubMed Central
- Bookshelf

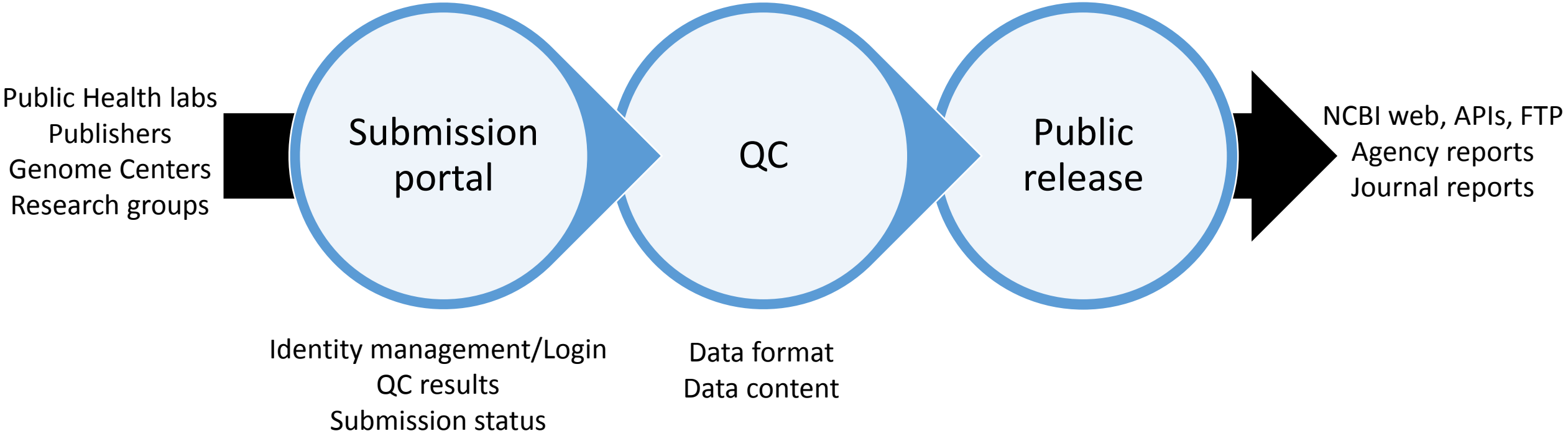
Sequence

- Sequencing results
- Meta-data
- Derivative databases
- Medical Genetics / Variation

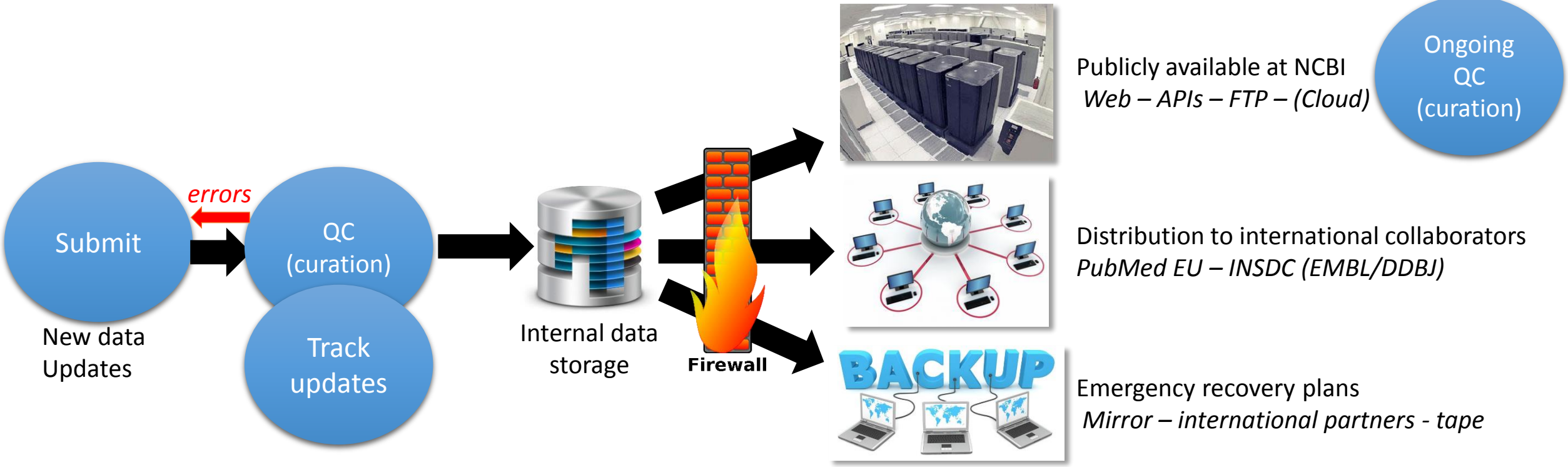


A representational image

Archival data submission – common elements



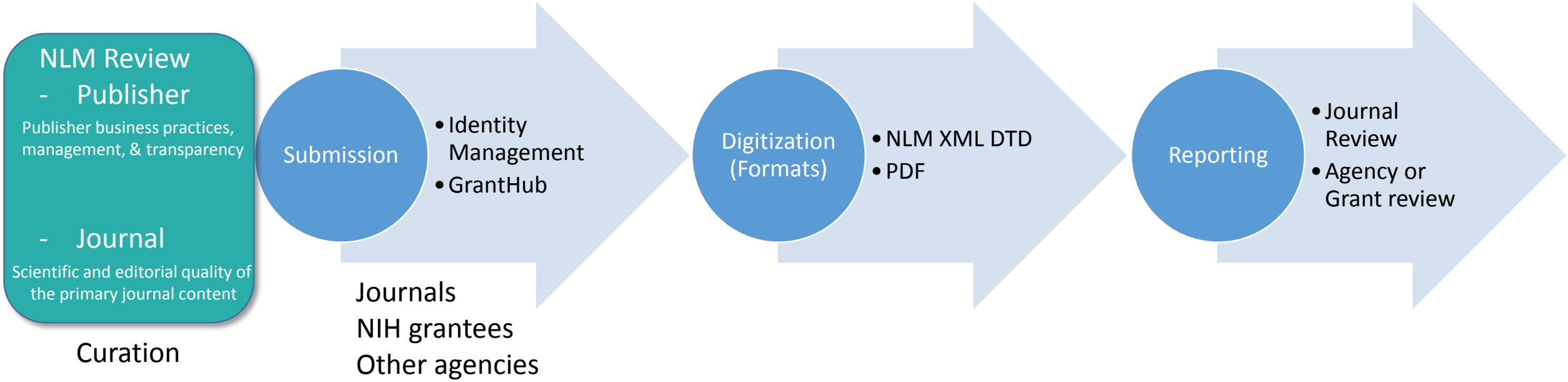
Archival data lifecycle



Not all archival data is the same – some are backed by political agreements, others reflect community needs at a given time.
Archival data continues to be publicly available as long as delivery resources are available.
Some archival data types could in theory be discontinued due to budgetary constraints, or revised mission priorities.
In practice, content would likely continue to be available in some mode (ftp, backup) for a considerable time.

Literature

- PubMed - Abstracts
- PubMed Central (PMC) - full text
- NCBI Bookshelf: Books, review articles



PubMed Central - a full text article archive

Foundation Principles

Two roles of PMC

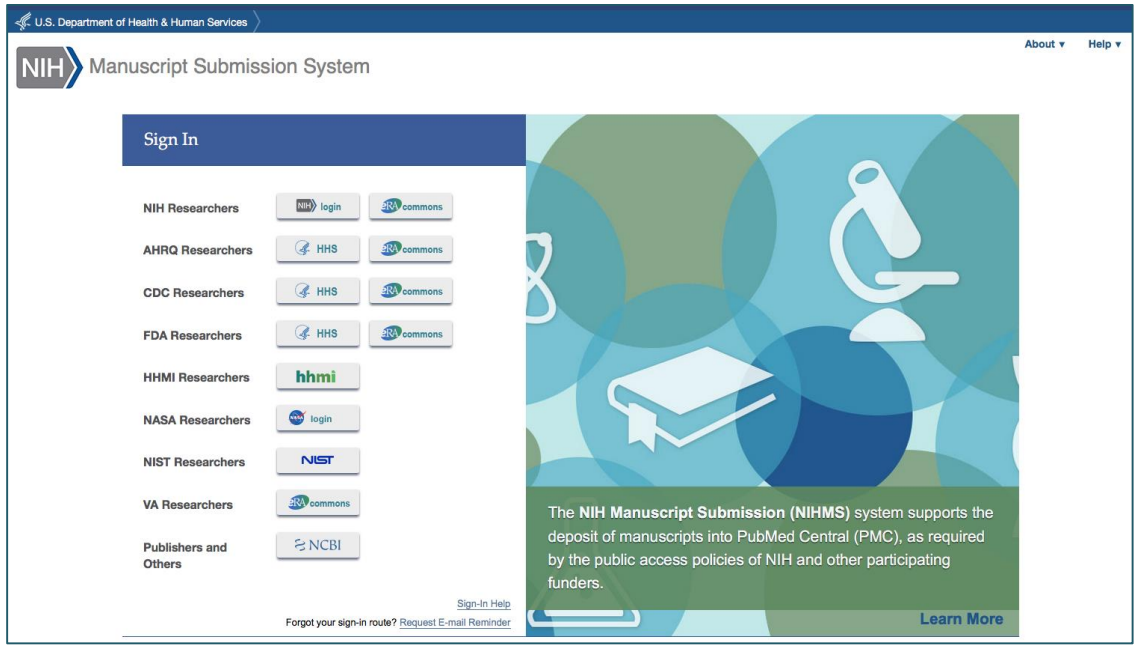
Journal Literature Archive

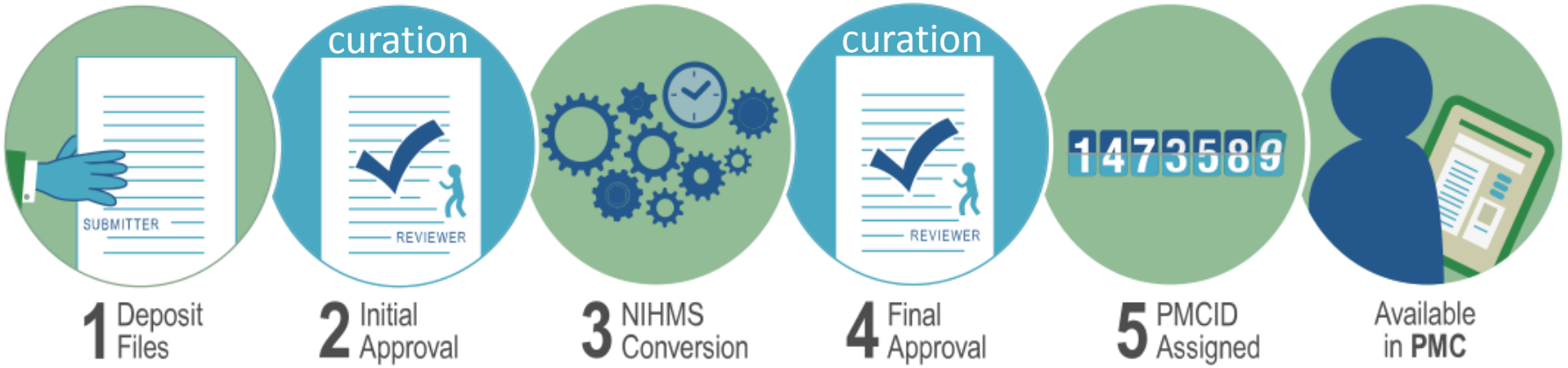
For a journal to be archived in PMC, it must meet certain scientific and technical criteria and demonstrate adherence to best practices.

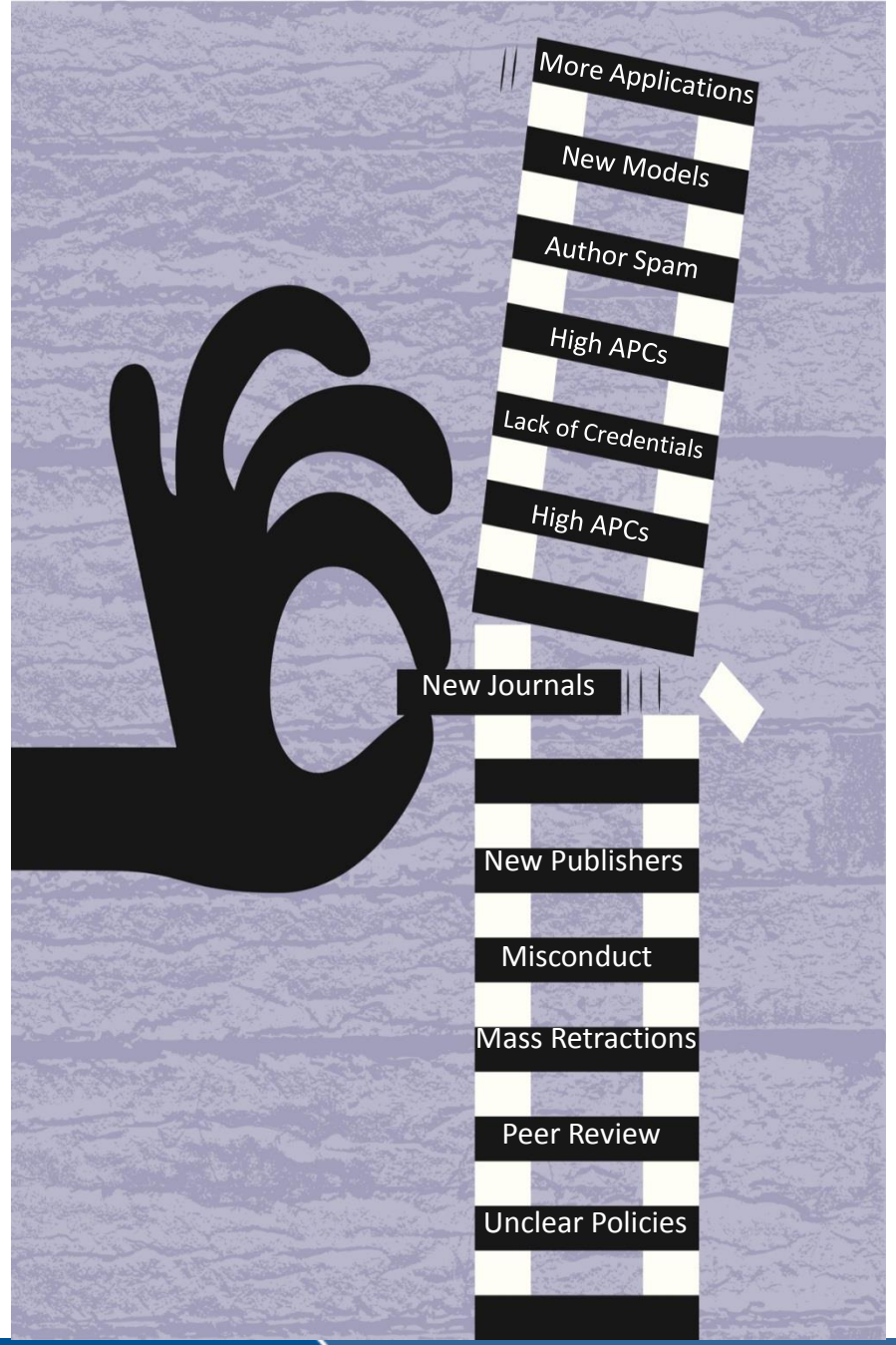
Funded Article Repository

For an article to be archived in PMC outside of a journal or publisher agreement it must be funded by a PMC-participating funder.

Funded Author Manuscript /NIHMS Submission Process







Outreach Activities

"In 2014, with the approval of the PMC National Advisory Committee, PMC implemented a scientific and editorial quality review procedure whereby **expert consultants from outside NLM conduct an independent review of journals seeking to participate in PMC**….

The independent review follows an assessment by NLM that the journal meets NLM's criteria for its collection, as outlined in the Collection Development Manual. NLM's Library Operations Division takes the independent reviewers' opinions into account in making the final decision on a journal's suitability for inclusion in PMC."

http://www.ncbi.nlm.nih.gov/pmc/about/faq/#q14

PMC as A Funded- and Historical-Content Repository

U.S. and international cooperation efforts

U.S. Federal Gov't Funders

NIH
AHRQ
CDC Centers for Disease Control and Prevention
FDA
ASPR Assistant Secretary for Preparedness and Response
ACL Administration for Community Living
NASA
Department of Veterans Affairs
United States Environmental Protection Agency
NIST

U.S. NGO Funders

hhmi Howard Hughes Medical Institute
Bill & Melinda Gates foundation
Health Research Alliance

International Funders

Europe PubMed Central

PMC Europe supports the policies of 27 European research funders.

CIHR IRSC Canadian Institutes of Health Research / Instituts de recherche en santé du Canada

CIHR-funded papers are deposited in PMC Canada.

Funders added to
http://www.ncbi.nlm.nih.gov/pmc/about/public-access/
as policies are implemented

PMC (etc.) - measuring success

- Data volume
- Data access
- More submitters
- Adoption of format standards

- Submission process time
- QC issues
- Email volume

Sequence

Archival

- Nucleotide – GenBank, SRA
- Meta-data – Sample, Project, Taxonomy
- Medical genetics – dbGaP, ClinVar, GTR, dbSNP

Services, value ads, curation

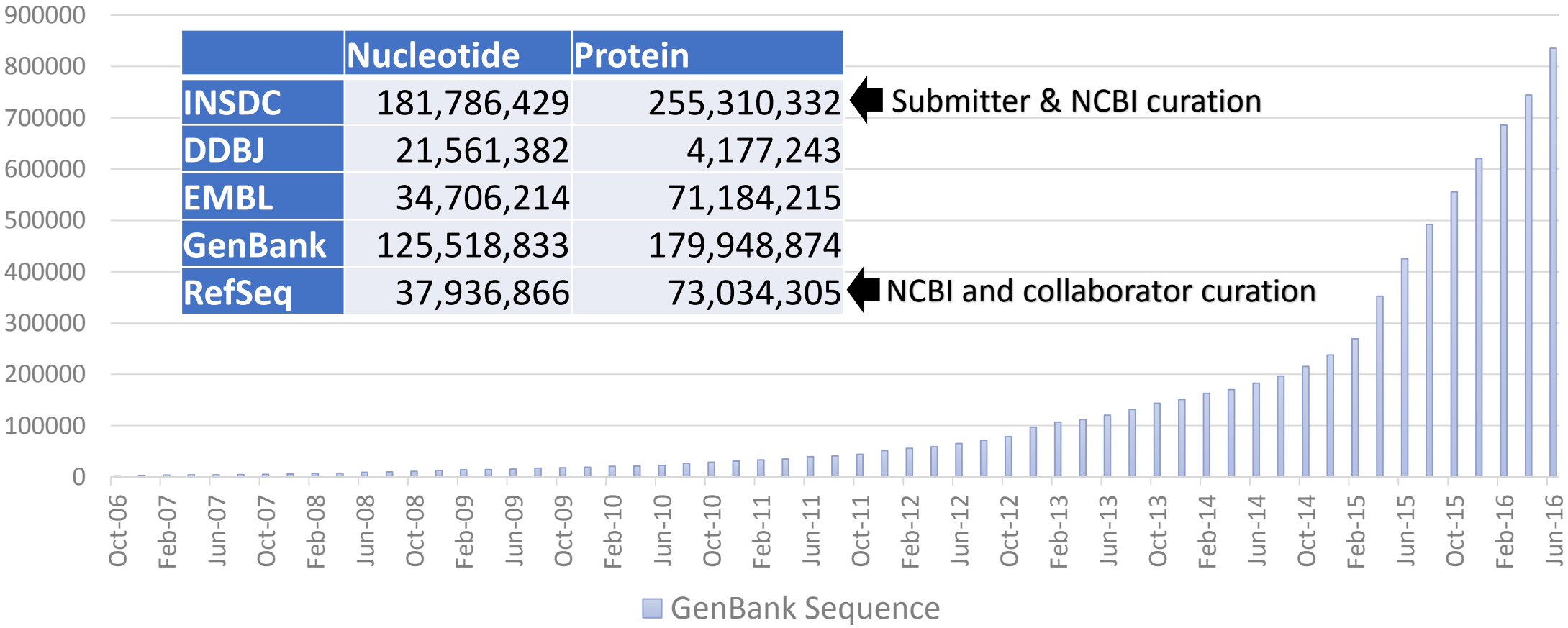
- Reference sequences (RefSeq) and annotation pipelines
- Pathogen detection (FDA collaboration)
- Gene, Genomics resources

Nucleotide data (***volume***) & curation

	Nucleotide	Protein
INSDC	181,786,429	255,310,332
DDBJ	21,561,382	4,177,243
EMBL	34,706,214	71,184,215
GenBank	125,518,833	179,948,874
RefSeq	37,936,866	73,034,305

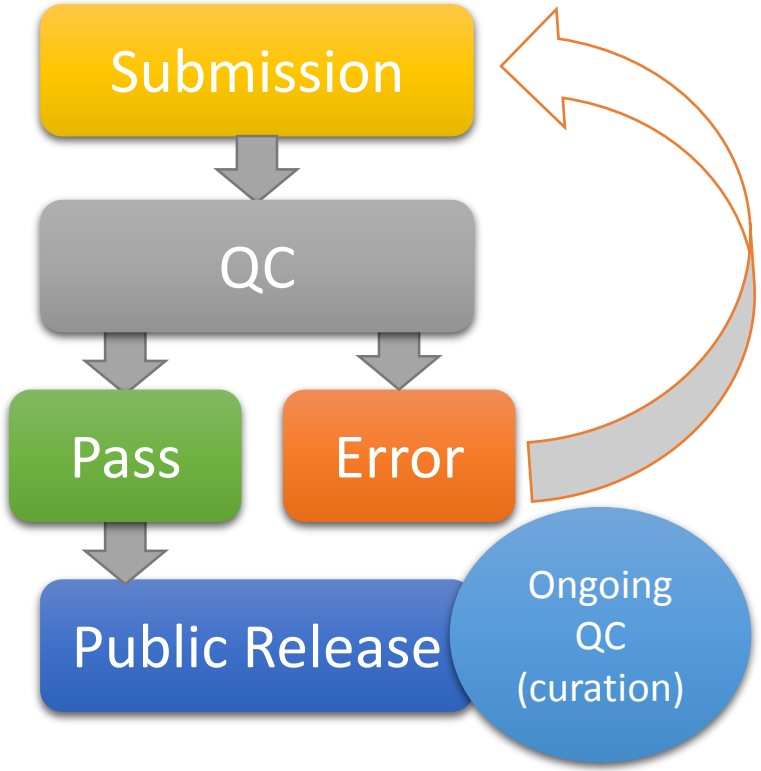
Submitter & NCBI curation

NCBI and collaborator curation



Automation: the key to keeping up with volume

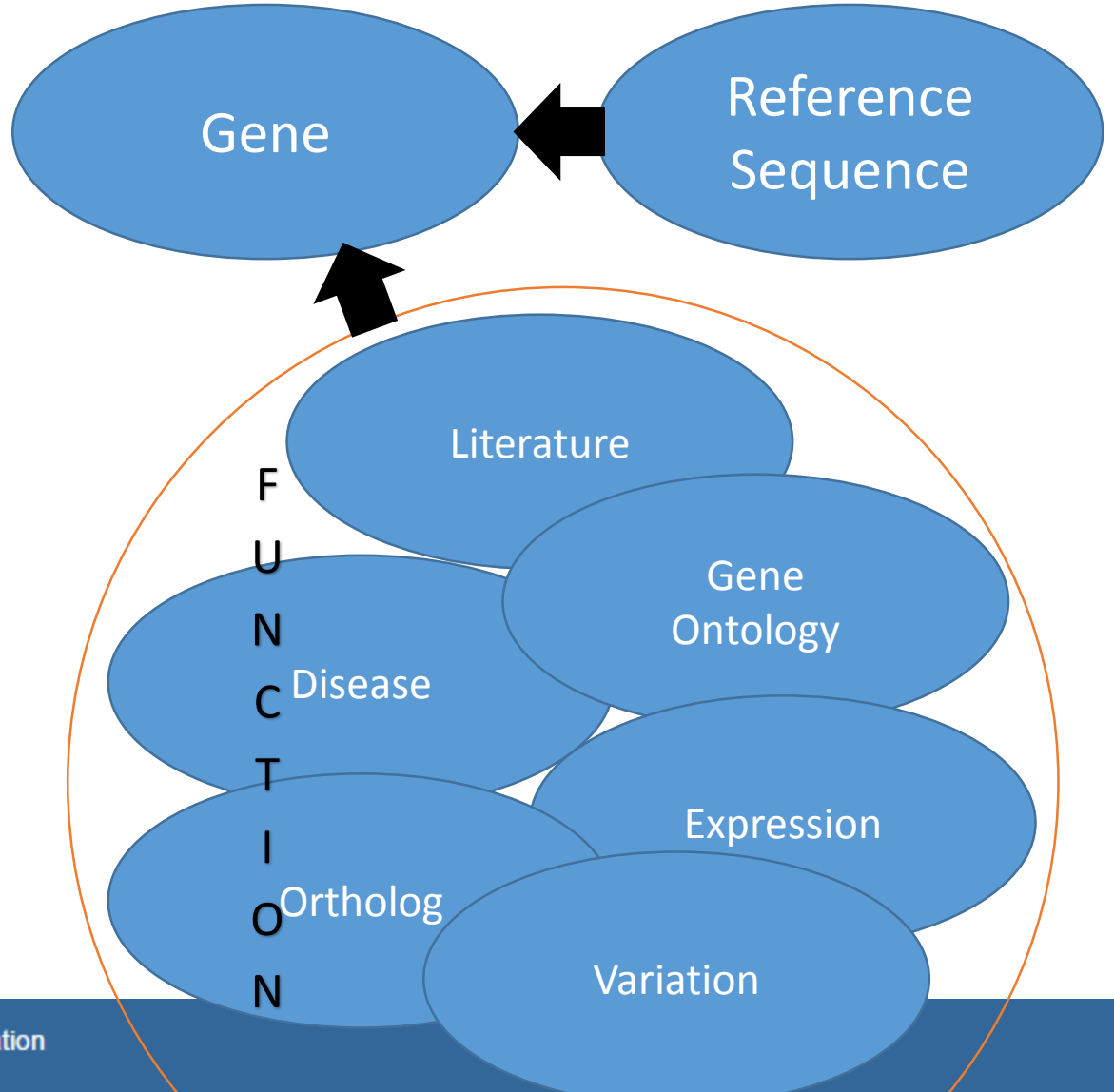
- Submission portal
  - Customer can monitor status
  - Customer can access QC results
- Automating QC checks
  - Taxonomy
  - Meta-data
  - Mark-up errors
  - More focused curation
- Certain errors get pushed to GenBank curators for follow-up with submitter



Extensive Curation – RefSeq & Gene Projects

Multiple Value Additions:

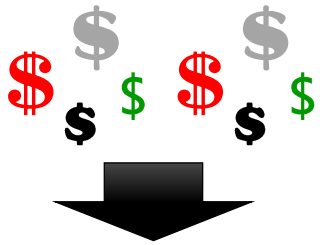
- Genome annotation
- Curated data
- Connect SEQUENCE to FUNCTION
- Collaborations
  - Gene & protein names
  - Correct taxonomy
  - Annotation methods
  - GMODs
  - Clinical support resources



More information:
https://www.ncbi.nlm.nih.gov/refseq/
https://www.ncbi.nlm.nih.gov/gene/

Connecting sequence data to knowledge

Billions of dollars have been invested by federal agencies, international governments, and charitable foundations to obtain sequence data for thousands of organisms.



Sequence Data

Knowledge

Genes & function

Food security

Health & disease

Treatment

```
AATACTGAAAAACACTGGTAGGTGCTCAGTAAGTGATTTTCACTAATGGCAAAATGATTGAGGACAGTAC
TGGAAATAAAAGCAGCCATTGAACACACATTTGTCTAAAAGTTACAAGAAATTTATAAATGAGGCTGTGA
TAAAGTCATTCAAAAGCGGACAGTCCAGGATAAAATAAAAAGTACAATGTATTAAAATGTAGTAAGGTGT
TTCTTAATGTATTTCACTGGTGTTACTATGTTTCTAAGGAATTTGGAGAGAAACTGTAGTAGCCTGATGT
TAAACTTAAGGTTCTGGAAGTCCAGACTCCACAGTGGTGGGTTCCAGGTGTCAAGTCGGAGTTGTGTCAG
CGGGAGGATTAGGGTCAGCTGCCCGCTAATGGCAGCCAGCAAGCCTCCGCTGCAGGAAATGCTGCGGGAT
GGGGAAGGTCTGAGCTTGCTTCTTTACTGGCCAAATCCCATGTGACATTGAGGGGCAGACCTTAGTAGGT
ATG                                                                    G
GCA                                                                    A
TTA                                                                    T
ATT                                                                    C
AAA                                                                    A
GGACAATTGTTTCGTGCTTTCTTAAGTCCCTGACCAAGATCAGTAGGGAAATCCCATTAGCTCTATTCAT
CGCATTTCTCATAAATCTAAAAGGAAAGCAAGTATCCAAAGCTTTATTCTTTAATAGTTCTTATTTGTAA
AACCTTGACAGGGCTTTATTCTAAATTGCTTCCAAATTACATAAAGAATAGTTAGAAATATGATGGGCAG
TGAATCGTGATAGTTTCAGTCATCGAATAAATAAGTAGAACAGTAAAAGGGGACTAATCTGAAAGCAAAC
AGAGTATGATGGGCTACGATGAGGCAAAAAGAACAGAACTGTGGAATTCTTAAACTACCATCTGGCTCAA
TCCCCTTTAAGTGGCCAAGTTGAAGAGCCTTTTATTCAGTGCCCTAAAACTTATTCAAAAGGCGCAGTAT
TTGAATATTAATATTTCCATACAGTAAAAAAAAAAGTCTTTTGTAACCAAAAAGGAATGTTTAACAAAAT
AATGGGTAATCCTCTGGCTAAACATGAGGGCGTGGGGAGCACCCTGTGGGTGGGAGGTCGGCTGGAACCC
```

Where are the genes, transcripts, proteins?

What is the function?

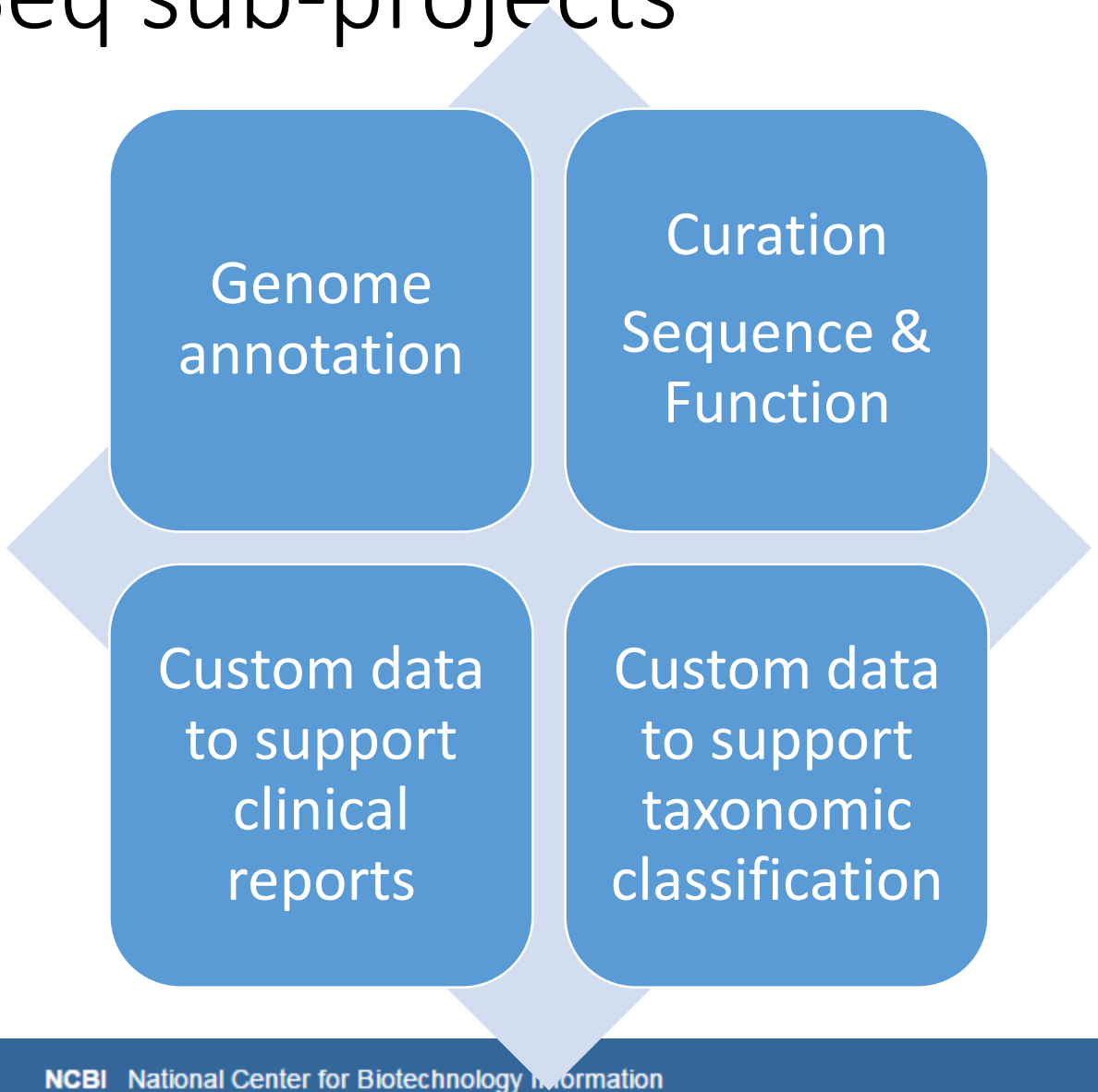
Is a disease associated with changes to this sequence?

How is RefSeq different from GenBank?

- GenBank is an archival sequence database
  - ‘the research article’
- RefSeq is a reference, vetted, sequence data set
  - ‘a review article’
- RefSeq records are updated to stay current
- RefSeq records are supported by curation and collaboration

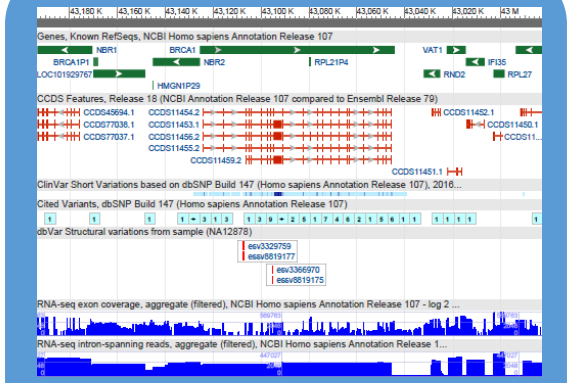
RefSeq provides genome annotation, transcripts and proteins that may not be available in GenBank.

Many RefSeq sub-projects

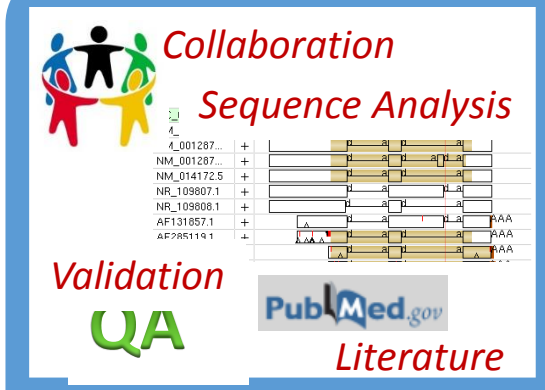


Eukaryotes example (we have prokaryotes too)

Sequence Data



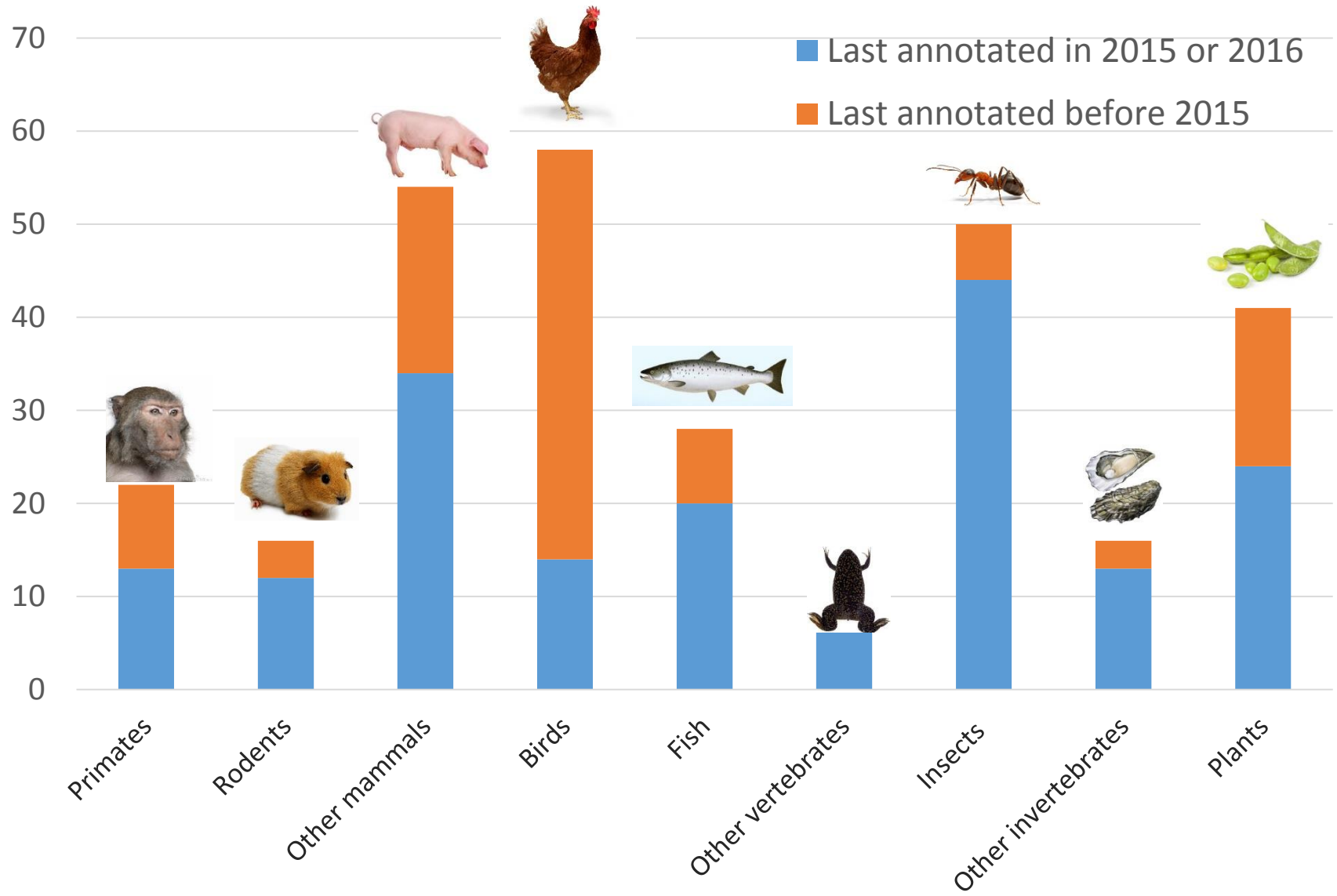
Annotation Pipeline



Curation

	Curated
Nucleotide	272,162
Protein	269,638

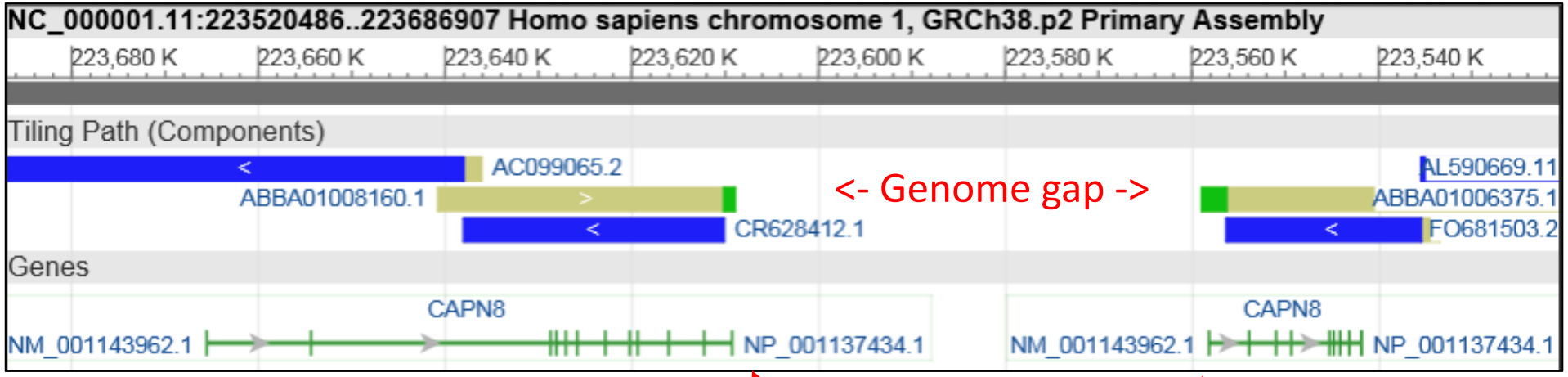
RefSeq annotated genomes



Nearly 350 plant & animal organisms annotated.
Most of these do not have GenBank annotation in the GenBank version.

Providing complete transcripts & proteins even when genome is incomplete

A region of human chromosome 1



This RefSeq transcript is complete (extends across the gap)

Collaborations supporting human RefSeq data (a few examples)

Genome reference consortium

- Maintaining the human reference genome sequence

Consensus CDS collaboration

- Harmonizing international annotation of the human genome

HUGO gene nomenclature committee

Immunogenetics

Supporting clinical reporting needs

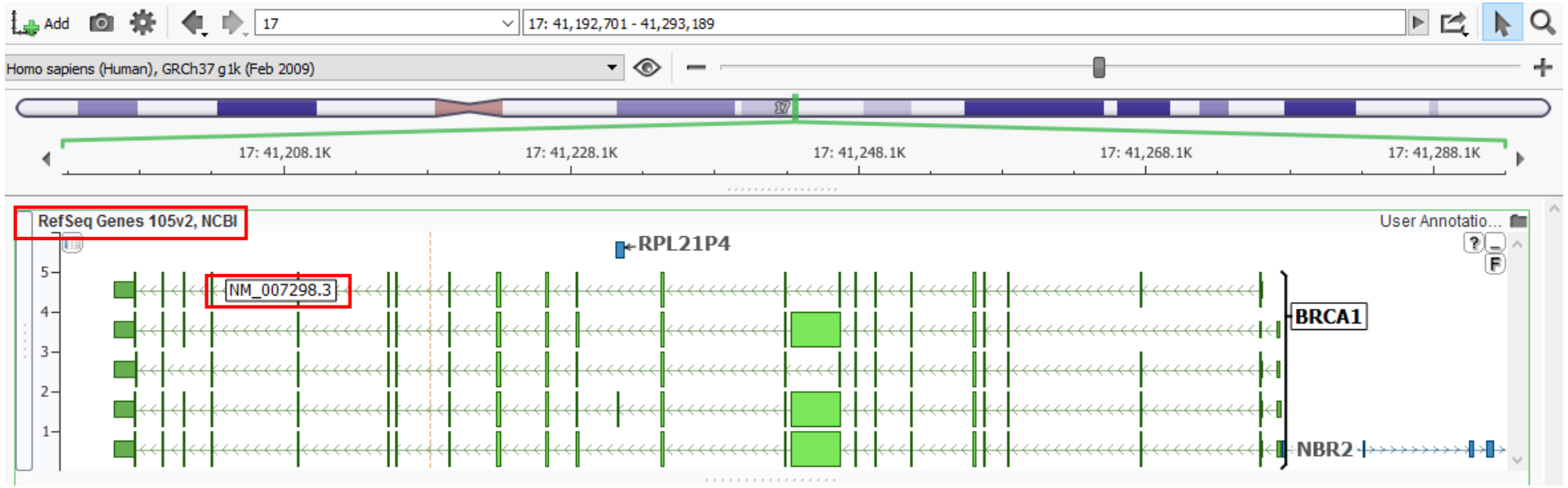
RefSeqGene Locus·Reference·Genomic

RefSeq use in clinical genomics

Commercial products

National Library of Medicine

SolveBio | GOLDEN HELIX | ClinVar | RefSeqGene | PheGenI Phenotype-Genotype Integrator



VARIANT NAMES

CODING DNA

NM_000492.3:c.1521_1523delCTT

GENOMIC

NC_000007.13:g.117199646delCTT

OTHER NAMES

CFTR NM_000492.3:c.1521_1523delCTT rs199826652

A standard framework to report sequence locations

AJHG

Cell PRESS

Volume 98, Issue 3, 3 March 2016, Pages 541–552

Article

Disruption of *POGZ* Is Associated with Intellectual Disability and Autism Spectrum Disorders

Table 1. Clinical Features of Individuals Harboring *POGZ* Mutations from ID and/or DD Cohorts

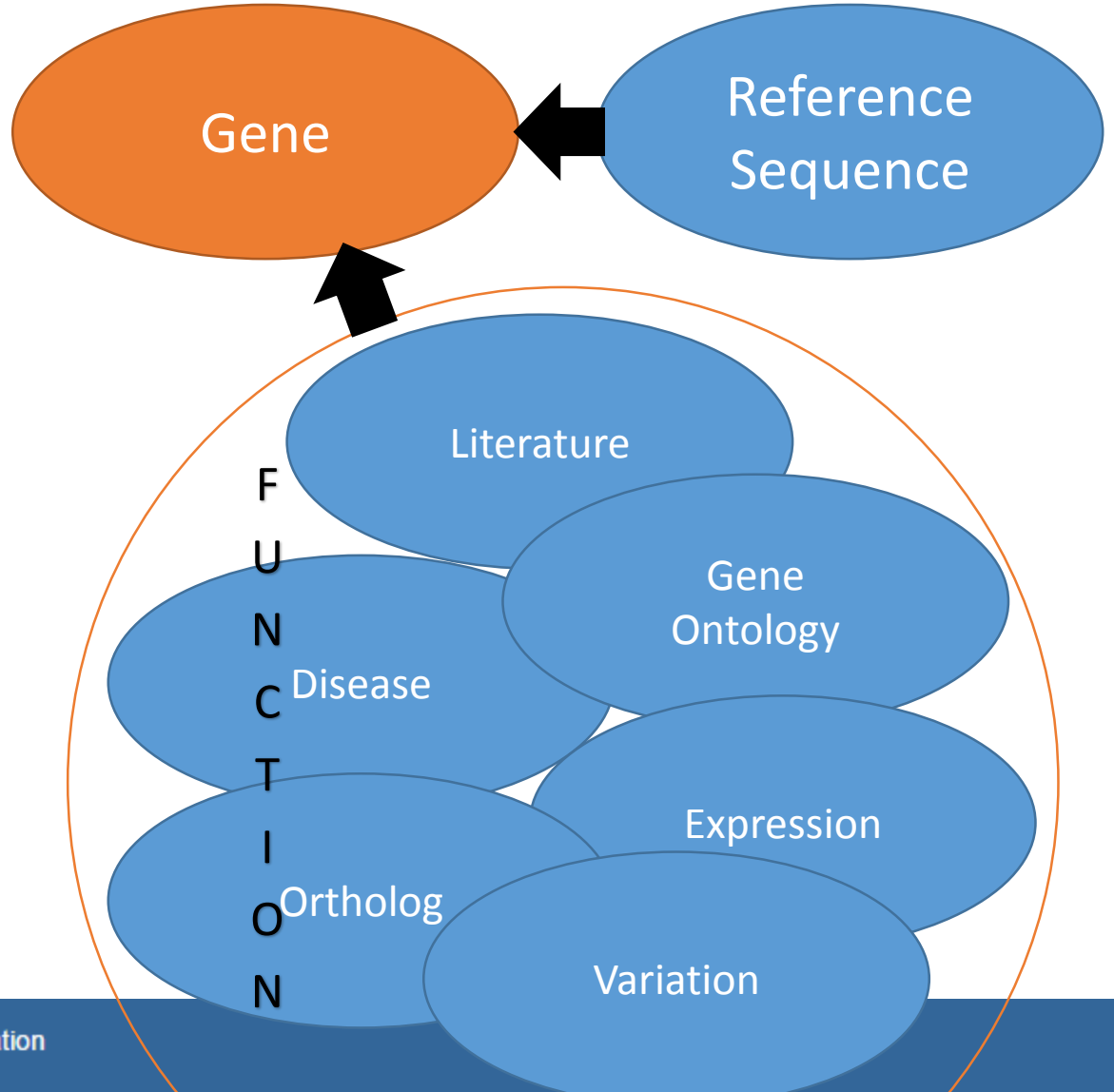
	UMCN1	UMCN2	UMCN3	UMCN4	UMCN5	UMCN6	UMCN7	UMCN8	UMCN9
Mutation	c.2590C>T (p.Arg864*)	c.3001C>T (p.Arg1001*)	c.3456_3457del (p.Glu1154 Thrfs*4)	c.2263del (p.Glu755 Serfs*36)	c.1152dup (p.Arg385 Serfs*4)	c.2432+ 1G>A (p.?)	c.2020del (p.Arg674 Valfs*9)	c.3847C>T (p.Gln1283*)	c.3456_3457del (p.Glu1154 Thrfs*4)
Age (years)	5	13	9	6	2	12	5	26	8
Gender	F	M	M	M	M	M	F	M	M
[illegible]	N.	[illegible]	[illegible]	[illegible]	[illegible]	–	–	+	[illegible]
Vision problems	–	+	+	–	+	+	–	–	+
Obesity tendency	ND	–	–	–	+	+	–	+	+

All HGVS annotations were annotated on RefSeq transcript (GenBank: NM_015100.3). A full clinical description for each individual can be found in Table S1. Abbreviations are as follows: mo., months; ID, intellectual disability; DD, developmental delay; ASD, autism spectrum disorder; M, male; F, female; +, formal diagnosis (mild or moderate); ++, severe presentation; +/−, possessing some features and/or mild presentation; −, not present; ND, no data.
[a]Inheritance unknown

Extensive Curation – RefSeq & Gene Projects

Multiple Value Additions:

- Genome annotation
- Curated data
- Connect SEQUENCE to FUNCTION
- Collaborations
  - Gene & protein names
  - Correct taxonomy
  - Annotation methods
  - GMODs
  - Clinical support resources



NCBI Gene resource – a ‘central hub’ of information

Nomenclature, aliases, chromosome, homology

Genome annotation

Bibliography, GeneRIF

Gene Ontology

Sequence

NCBI Resources How To pruitt My NCBI Sign Out

Gene Gene Search

Advanced Help

Full Report Send to: Hide sidebar >>

BRCA1 breast cancer 1 [*Homo sapiens* (human)]

Gene ID: 672, updated on 13-Mar-2016

- Summary
- Genomic context
- Genomic regions, transcripts, and products
- Bibliography
- Phenotypes
- Variation
- HIV-1 interactions
- Pathways from BioSystems
- Interactions
- General gene information
- General protein information
- NCBI Reference Sequences (RefSeq)
- Related sequences
- Additional links

Table of contents

Related information

Links to other resources

General information

Related sites

Feedback

Subscription

Recent activity

Human BRCA1:
http://www.ncbi.nlm.nih.gov/gene/672

Human BRCA1

http://www.ncbi.nlm.nih.gov/gene/672

Summary

Official Symbol BRCA1 provided by HGNC

Official Full Name breast cancer 1 provided by HGNC

Primary source HGNC:HGNC:1100

See related Ensembl:ENSG00000012048; HPRD:00218; MIM:113705; Vega:OTTHUMG00000157

Gene type protein coding

RefSeq status REVIEWED

Organism Homo sapiens

Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Euther
Primates; Haplorrhini; Catarrhini; Hominidae; Homo

Also known as IRIS; PSCP; BRCAI; BRCC1; FANCS; PNCA4; RNF53; BROVCA1; PPP1R53

Summary This gene encodes a nuclear phosphoprotein that plays a role in maintaining genomic
acts as a tumor suppressor. The encoded protein combines with other tumor suppres
sensors, and signal transducers to form a large multi-subunit protein complex known as the BRCA1-
associated genome surveillance complex (BASC). This gene product associates with R
and through the C-terminal domain, also interacts with histone deacetylase complexes.
plays a role in transcription, DNA repair of double-stranded breaks, and recombination.
gene are responsible for approximately 40% of inherited breast cancers and more than
breast and ovarian cancers. Alternative splicing plays a role in modulating the subcellu
physiological function of this gene. Many alternatively spliced transcript variants, some
disease-associated mutations, have been described for this gene, but the full-length na
of these variants has been described. A related pseudogene, which is also located on
been identified. [provided by RefSeq, May 2009]

Orthologs mouse all

Homology

Homologs of the BRCA1 gene: The BRCA1 gene is conserved in chimpanzee, Rhesus monkey, dog, cow, mouse, rat, and chicken.

Orthologs from Annotation Pipeline: 177 organisms have orthologs with human gene BRCA1

Map Viewer (Mouse, Rat)

OrthoDB: The Hierarchical Catalog of Eukaryotic Orthologs

Gene Ontology Provided by GOA

Function	Evidence Code	Pubs
DNA binding	TAS	PubMed
RNA binding	IDA	PubMed
androgen receptor binding	NAS	PubMed
chromatin binding	IEA	

HIV-1 interactions

Replication interactions

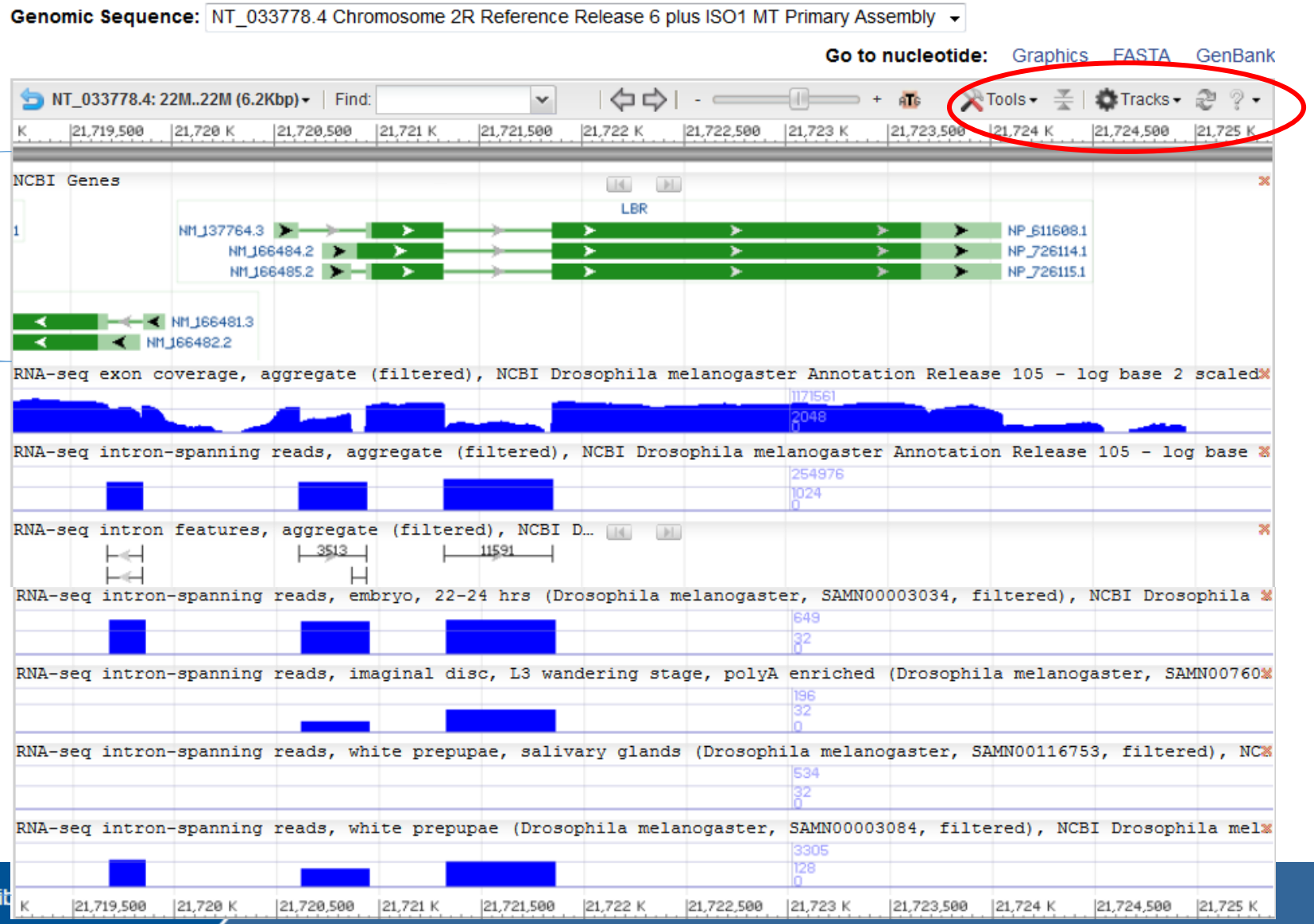
Interaction	Pubs
Knockdown of breast cancer 1, early onset (BRCA1) by siRNA inhibits HIV-1 replication in HeLa P4/R5 cells	PubMed

Protein interactions

Protein	Gene	Interaction	Pubs
Tat	tat	HIV-1 Tat associates with BRCA1 in cells and the amino-acid 504-802 region of BRCA1 physically interacts with HIV-1 Tat	PubMed
	tat	BRCA1 enhances HIV-1 Tat-dependent transcription in cells, and BRCA1 phosphorylation at positions S1387, S1423, S1457, and S1524 is important for the	PubMed

NCBI Gene: integrated genome browser

Example: Drosophila melanogaster Twist http://www.ncbi.nlm.nih.gov/gene/37655



RefSeq Annotation

(Ensembl also available)

RNA-seq tracks

Add tracks
Change track style
Zoom/Pan
Click/Drag to move tracks
Search
Zoom
Import data
Add BLAST RID
Export high res PDF

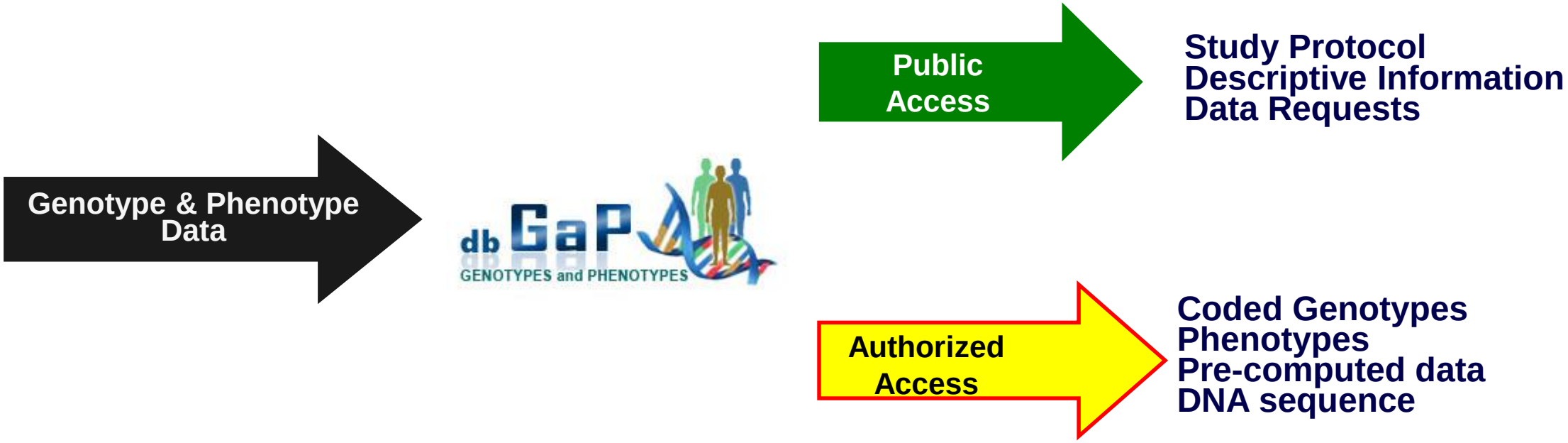
Privacy concerns

- The database of Genotypes and Phenotypes (dbGaP) archives and distributes the data and results from studies that have investigated the interaction of genotype and phenotype in Humans.

- The data
  - Genotypes – genetic markers in individuals
  - Phenotype traits – measured traits that differ in individuals
  - Genotype:phenotype association – personally identifiable information

NIH GWAS Policy: http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html

Privacy concerns: dbGaP public & private access



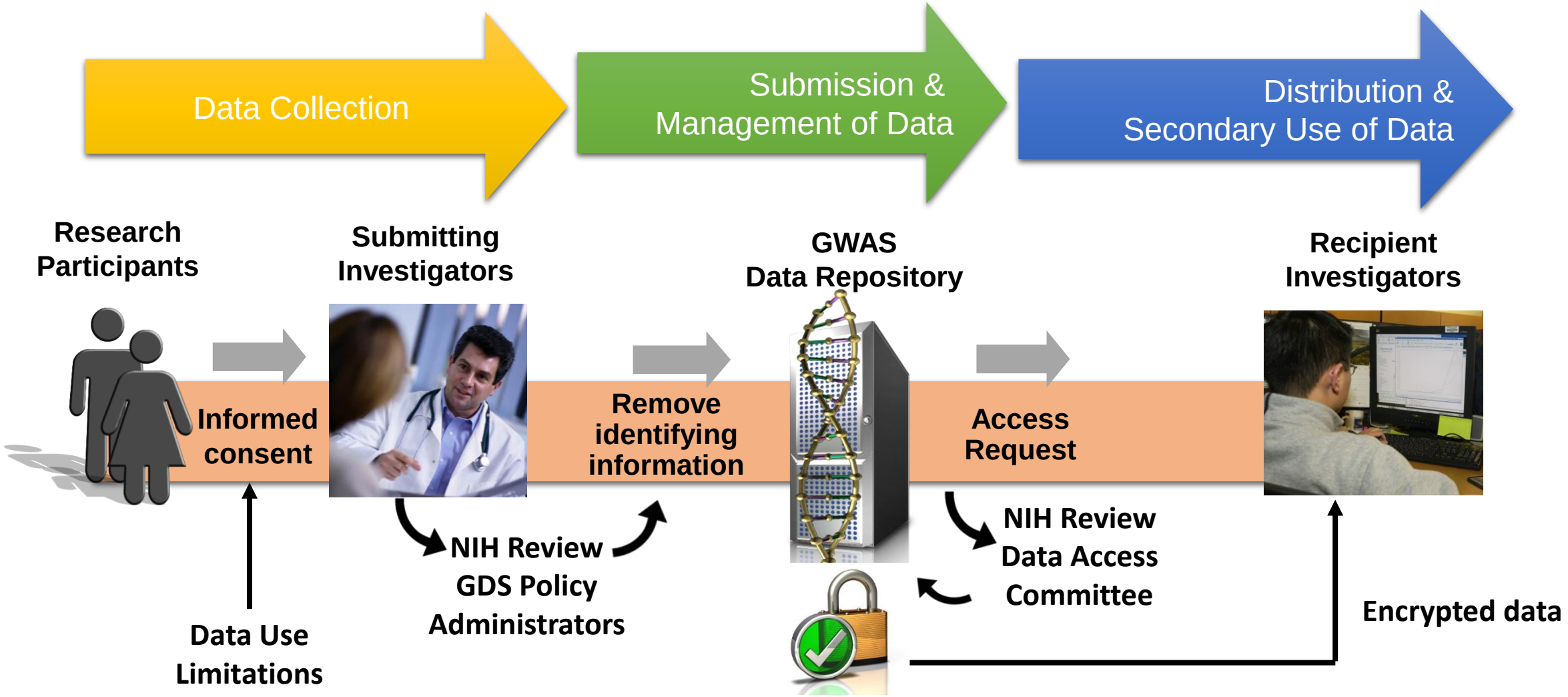
Important aspects:

- Accurate identity management system
- Secure/encrypted data delivery system

Guiding Principle for Data Sharing

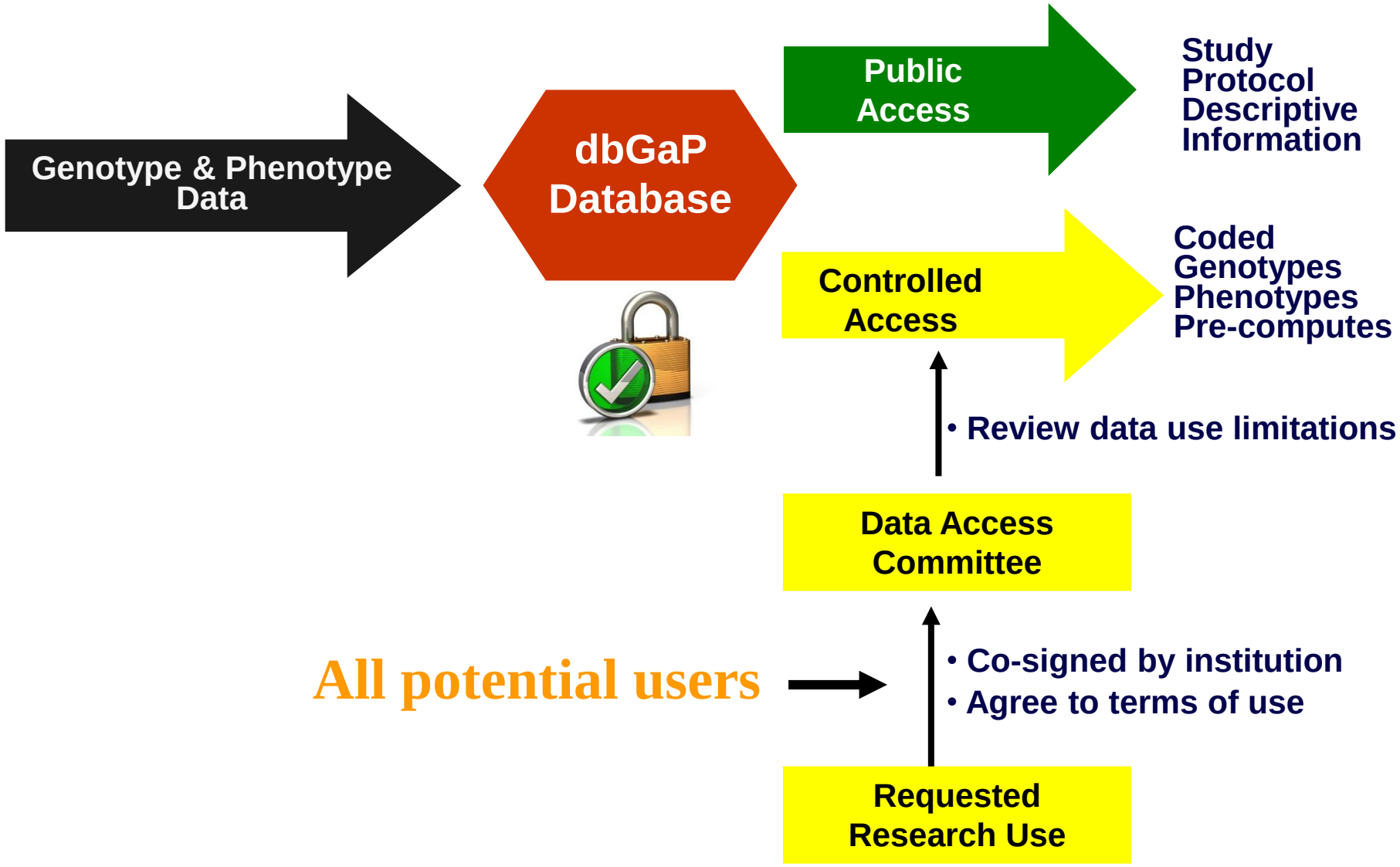
The greatest public benefit will be realized if [data] are made available, under terms and conditions consistent with the informed consent provided by individual participants, in a timely manner to the largest possible number of investigators.

NIH GWAS Policy Overview

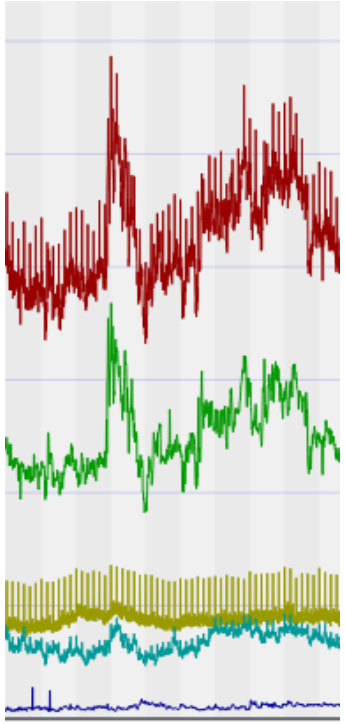


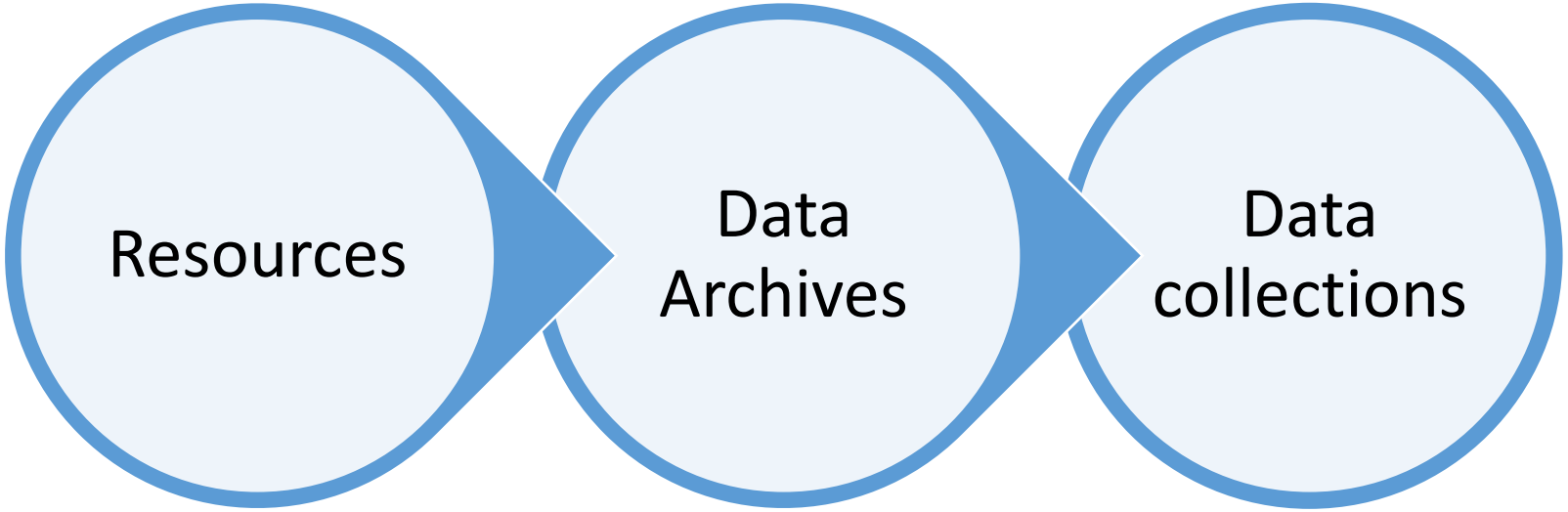
Effective as of January 25, 2008

Controlled Access for Specific Questions



Data management considerations





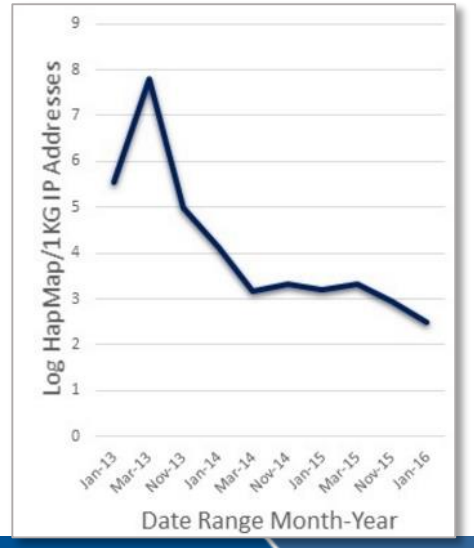
Usage and network logs

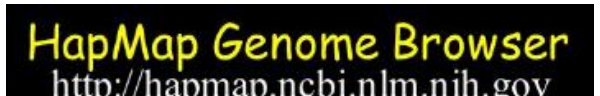
Data storage and access costs

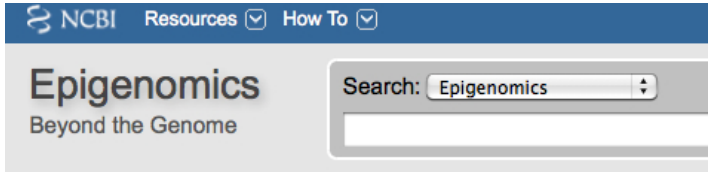
Use cases, enterprise priorities

Data lifecycles

- Archives – keep the data forever (but not necessarily online)
- NCBI Hosted, but non-archival, data collections may come and go
  - HapMap - Example decommissioned resource
    - haplotype map of the human genome to describe patterns of human variation
    - relevance (and usage) declined over time (replaced by 1000Genomes project)







Decommissioning versus finding new solutions

- Usage analysis
- Data volume and activity
- Increased storage/compute costs -> different solutions
  - Cloud services – let customer pay to compute
- Cost:benefit analysis for maintaining or upgrading hosted data sets
- Data relevance for current research
- NIH, NLM, NCBI Mission/Policy changes

For instance
- As sequencing becomes very inexpensive, how much redundant data should be archived?
- Is there a more economical way to store sequence data and meta-data?

- Data types, archives and the role of curation
- Data privacy
- Data management
- Metrics of success
- Outreach activities

- Slide contributions –
  - Katherine Funk (PMC)
  - Steve Sherry (dbGaP)

- NCBI leadership
  - David Lipman
  - Jim Ostell