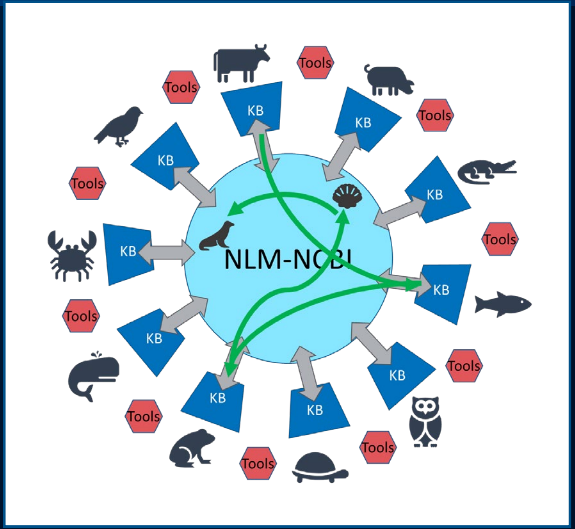
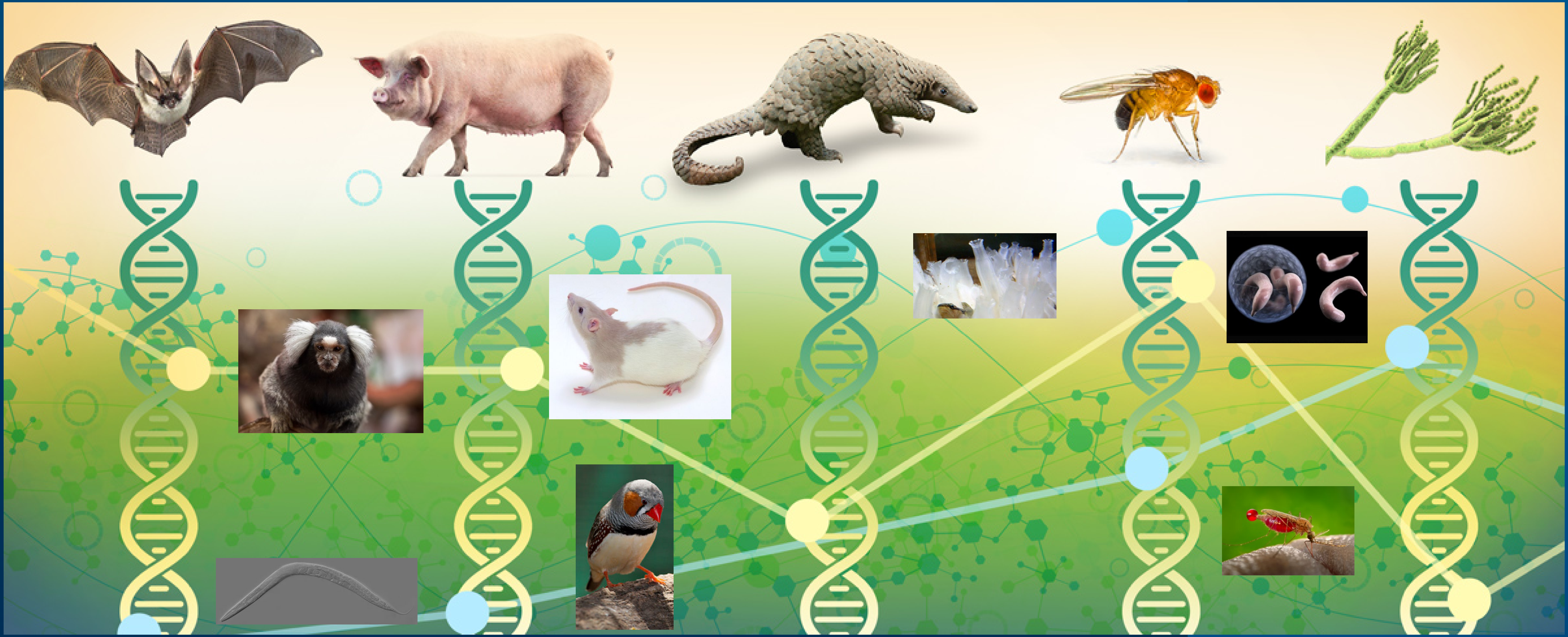
Introduction to the NIH Comparative Genomics Resource (CGR)

Valerie Schneider, Ph.D. Jan 6, 2022



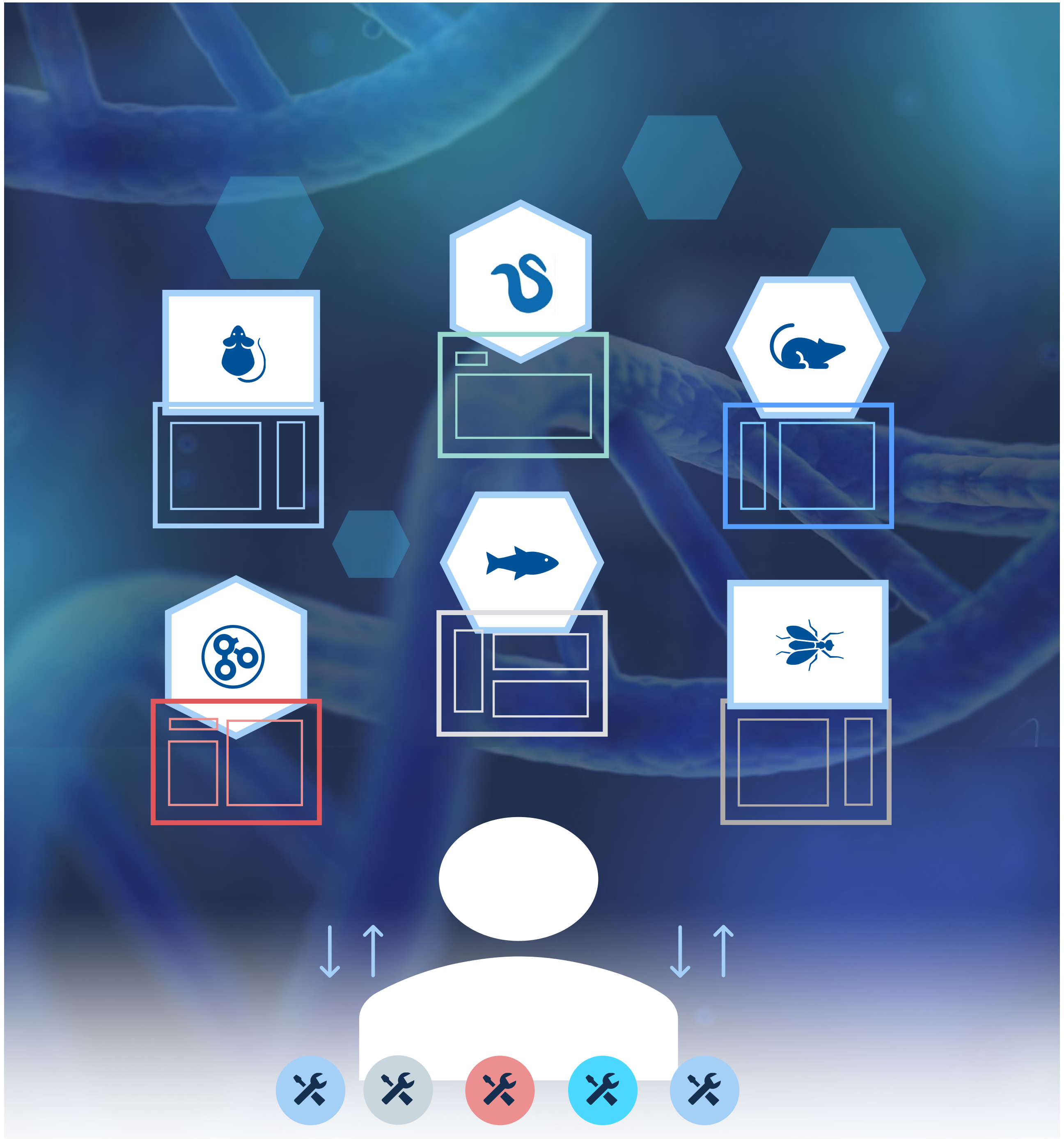
National Library of Medicine
National Center for Biotechnology Information

The Promise of Research Organisms



Understand biological processes

Understand disease



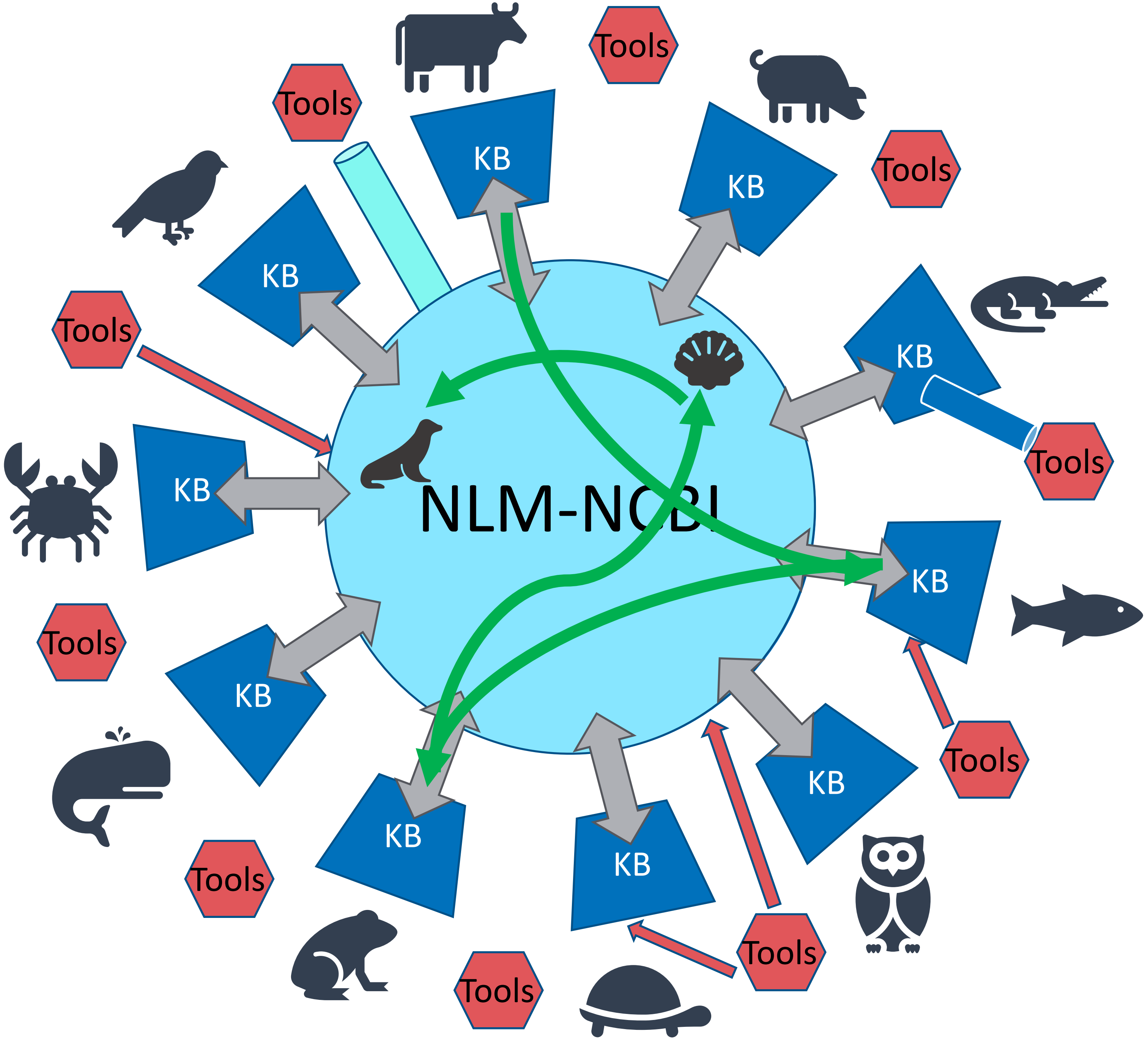
Current State

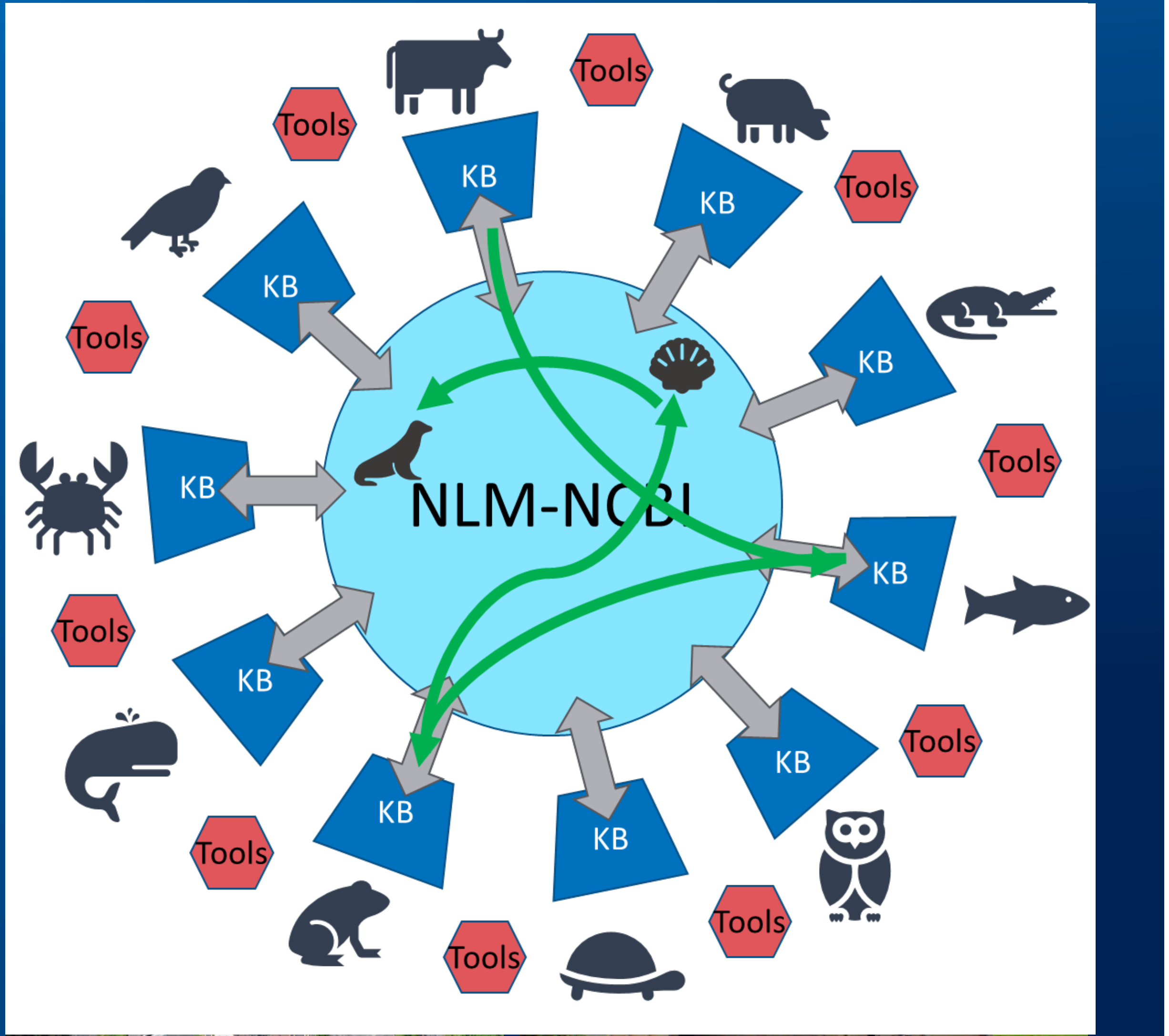
Limitations and Challenges

- Multiple different user interfaces
- Limited number of organisms supported
- Siloed data and applications
- Must download data to apply tools
- Limited scalability

CGR Vision

A consistently annotated comparative genomics cloud-based data resource for all eukaryotic research organisms that integrates gene and organism knowledge and provides a foundation for reliable comparative analysis.





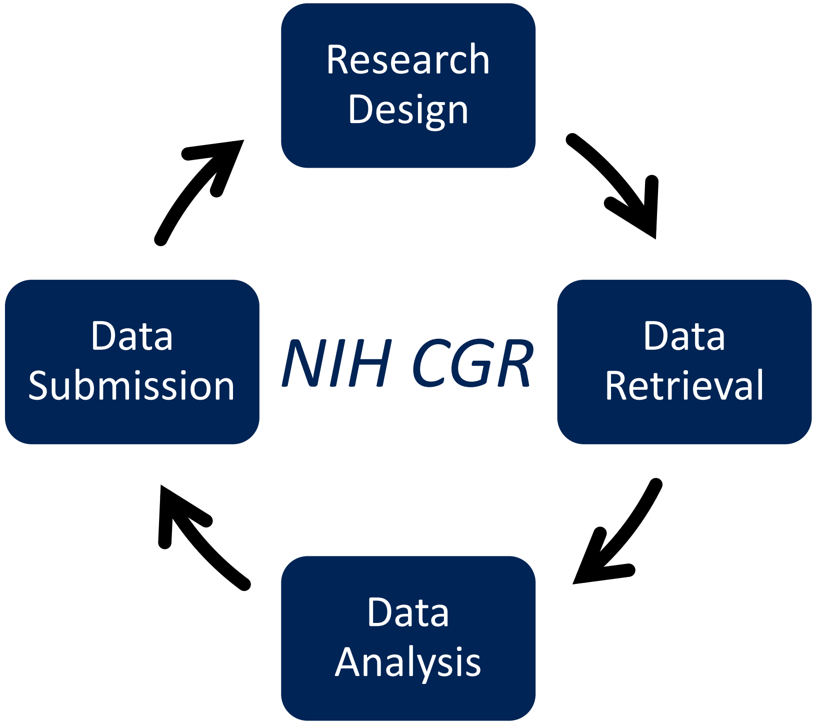
CGR: Infrastructure at the Core

Strategic Goals

Promote high quality data submission and re-use

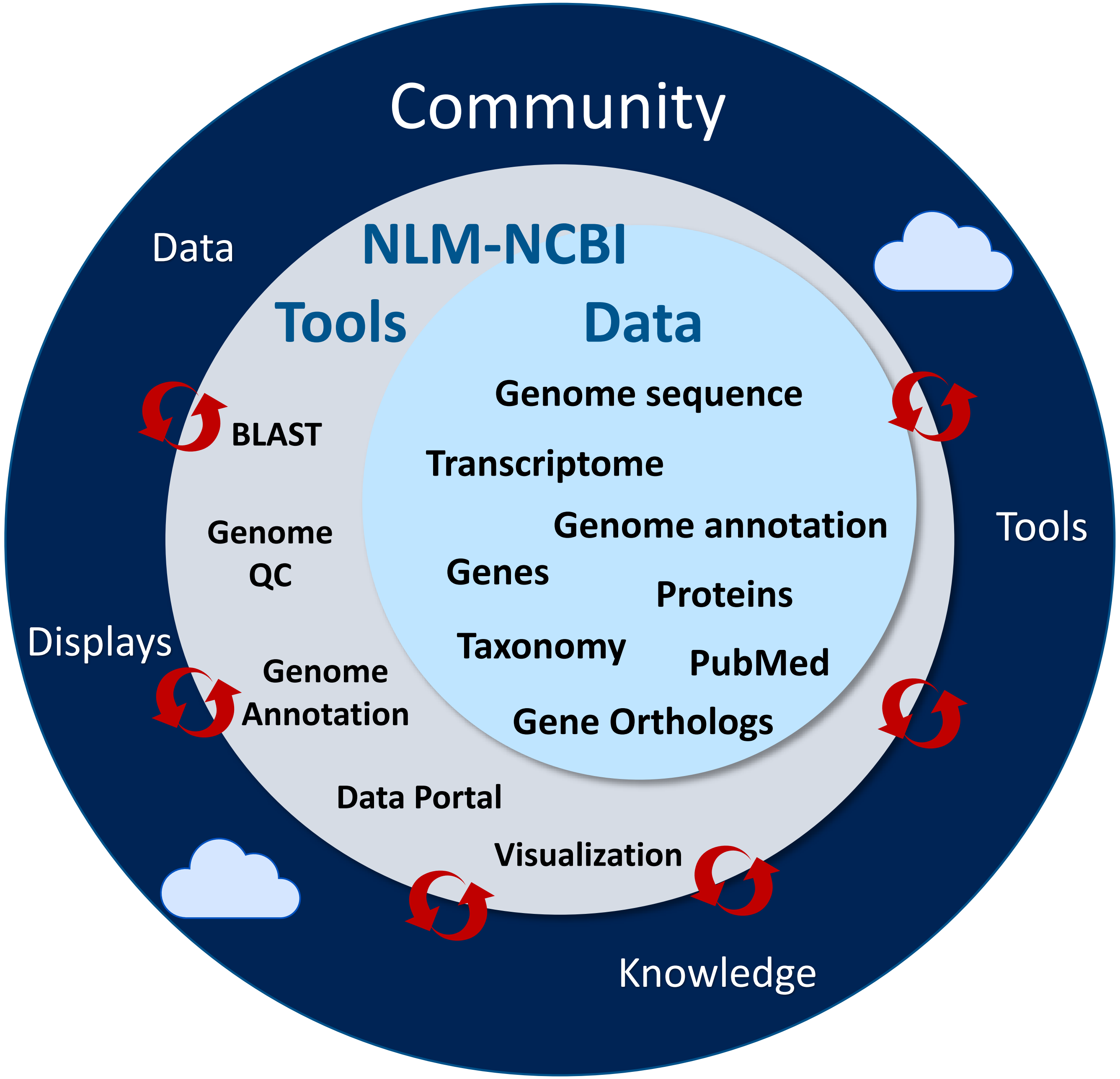
Offer best and most complete content

Support efficient and effective scientific discovery at the NLM/NCBI website



CGR Structure/ Benefits

- ✓ **Central portal** — support *all* research organisms; integrate data, metadata, links
- ✓ **Scalable analysis –** efficiency and economy at scale
- ✓ **Shared public tools** — accelerate research
- ✓ **Work in the cloud —** no need to download data to apply tools.
- ✓ **Meet new research needs** – create AI ready data sets
- ✓ **Community engagement** – FAIR data sharing



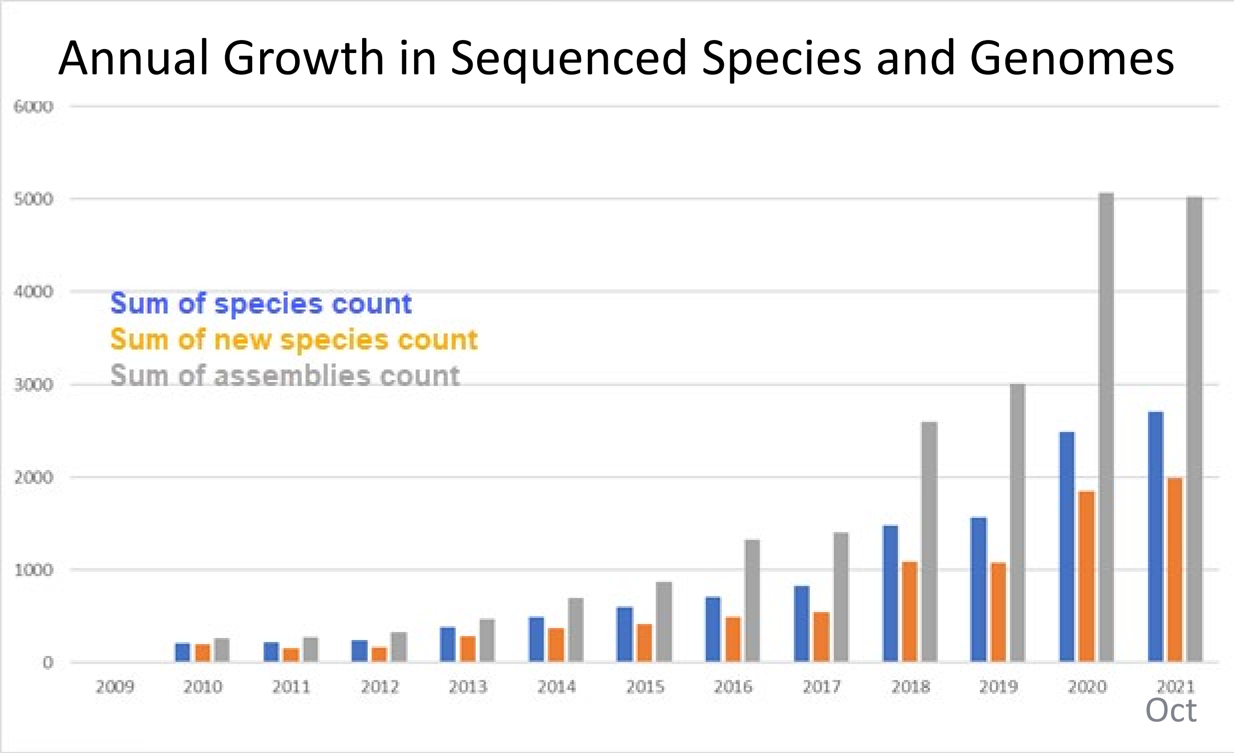
GenBank eukaryotic genome submissions:

- 64% are contaminated
- 80% lack annotation
- 20% have annotation
  - 58% have >50% proteins annotated as "*HYPOTHETICAL*"

Genome issues reveal the need for CGR

ALL EUKARYOTIC GENOMES (Cumulative: December 2021):

GenBank genomes (all):	20,927 (8,807 species)
GenBank (with annotation):	4,518 (2,612 species)
NCBI RefSeq annotated genomes (all):	1,357 (1,340 species)



National Library of Medicine
National Center for Biotechnology Information

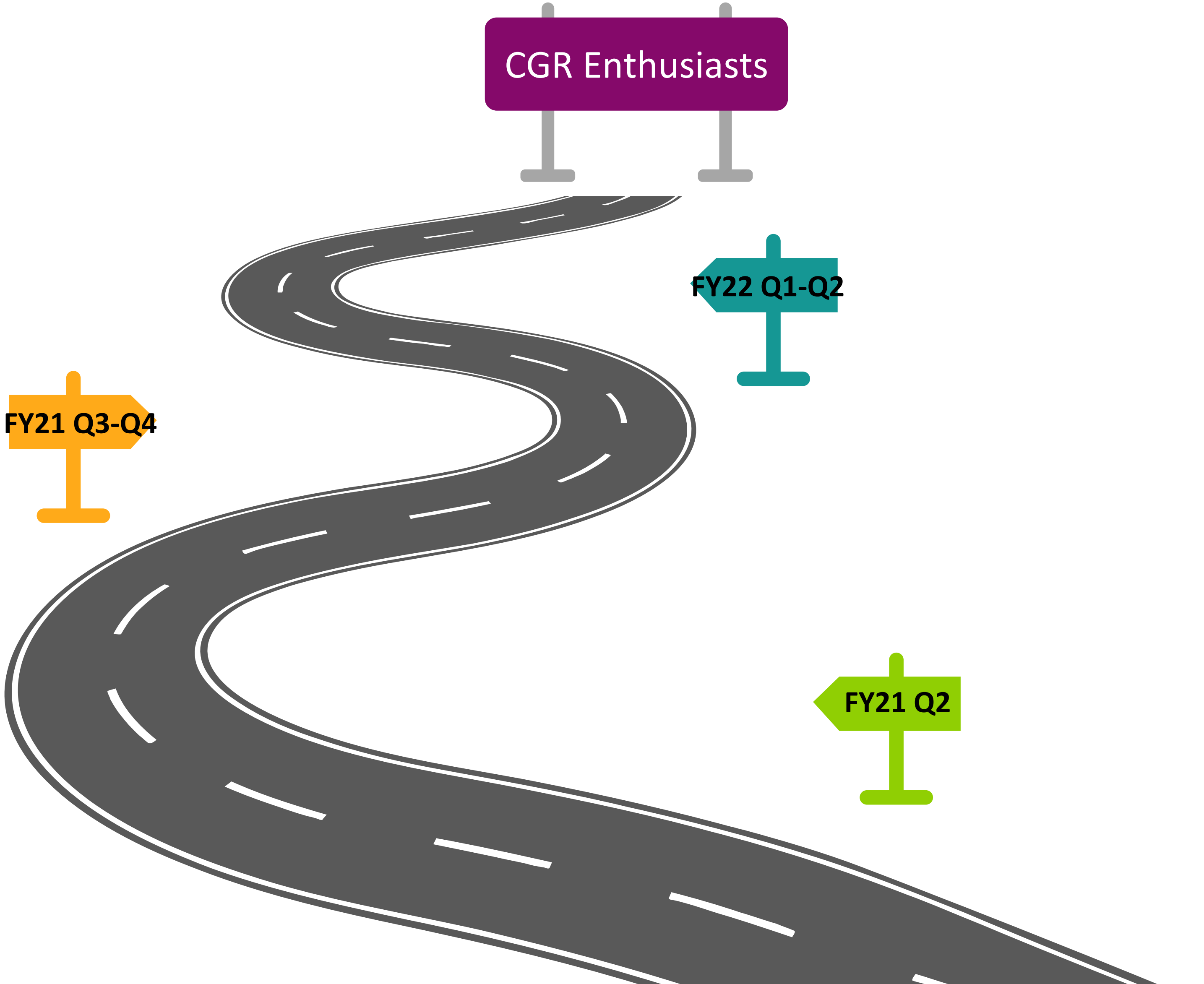
Paving the Way for CGR in Communities

Preparatory Activities

Outreach/communication team and plan
1st BoR WG meeting
2nd NLM blog, PAG Conference

Meetings w/ NIH CGR steering committee
Invite BoR WG members
Evaluate NIH organism “landscape”

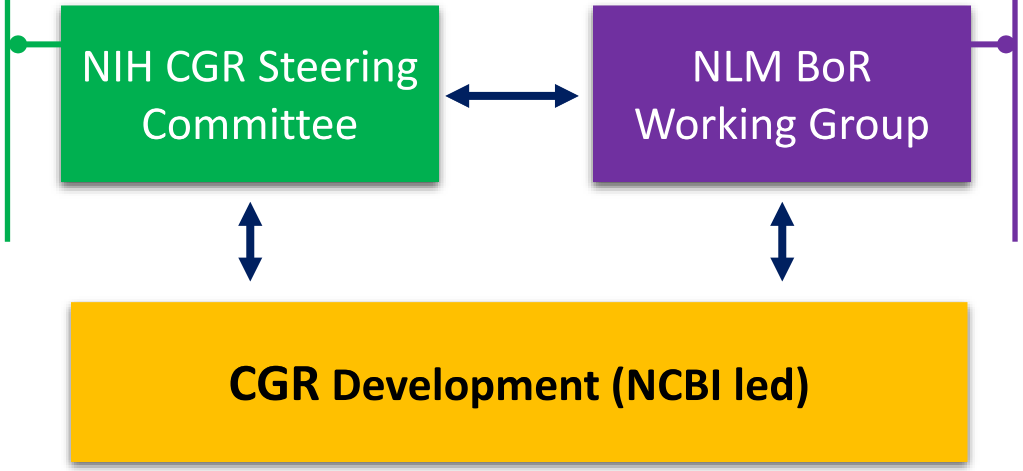
Refine vision & alignment with NLM goals
Governance structure organized
Initial public blog released



Supporting CGR Stakeholders

NIH CGR Steering Committee

- Report progress to SDC
- Monitor budget, milestones, progress, success metrics
- Ensure project remains within approved and funded scope.
- Amplify communications about value of initiative deliverables to NIH stakeholders



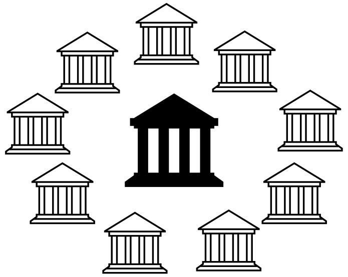
NIH CGR Steering Committee

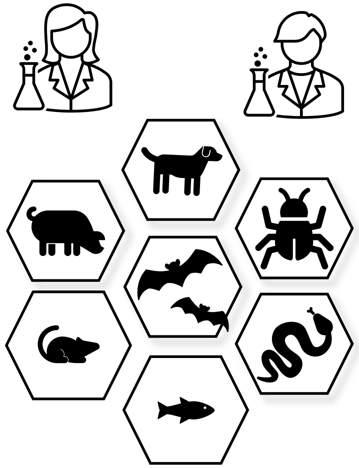
NLM BoR Working Group

CGR Development (NCBI led)

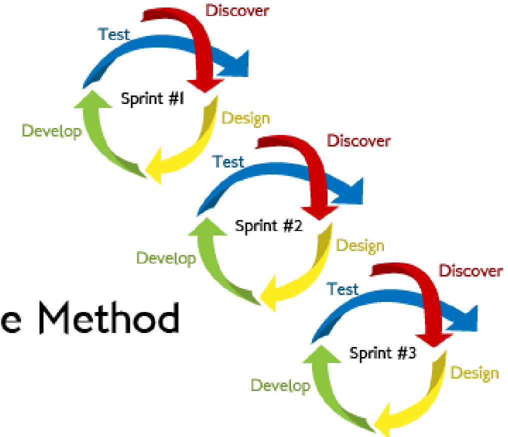
NLM BoR WG

- Help engage the scientific community
- Help NLM set priorities
- Guide the development of a new approach to scientific discovery
- Liaise with NIH CGR steering committee





Agile Method

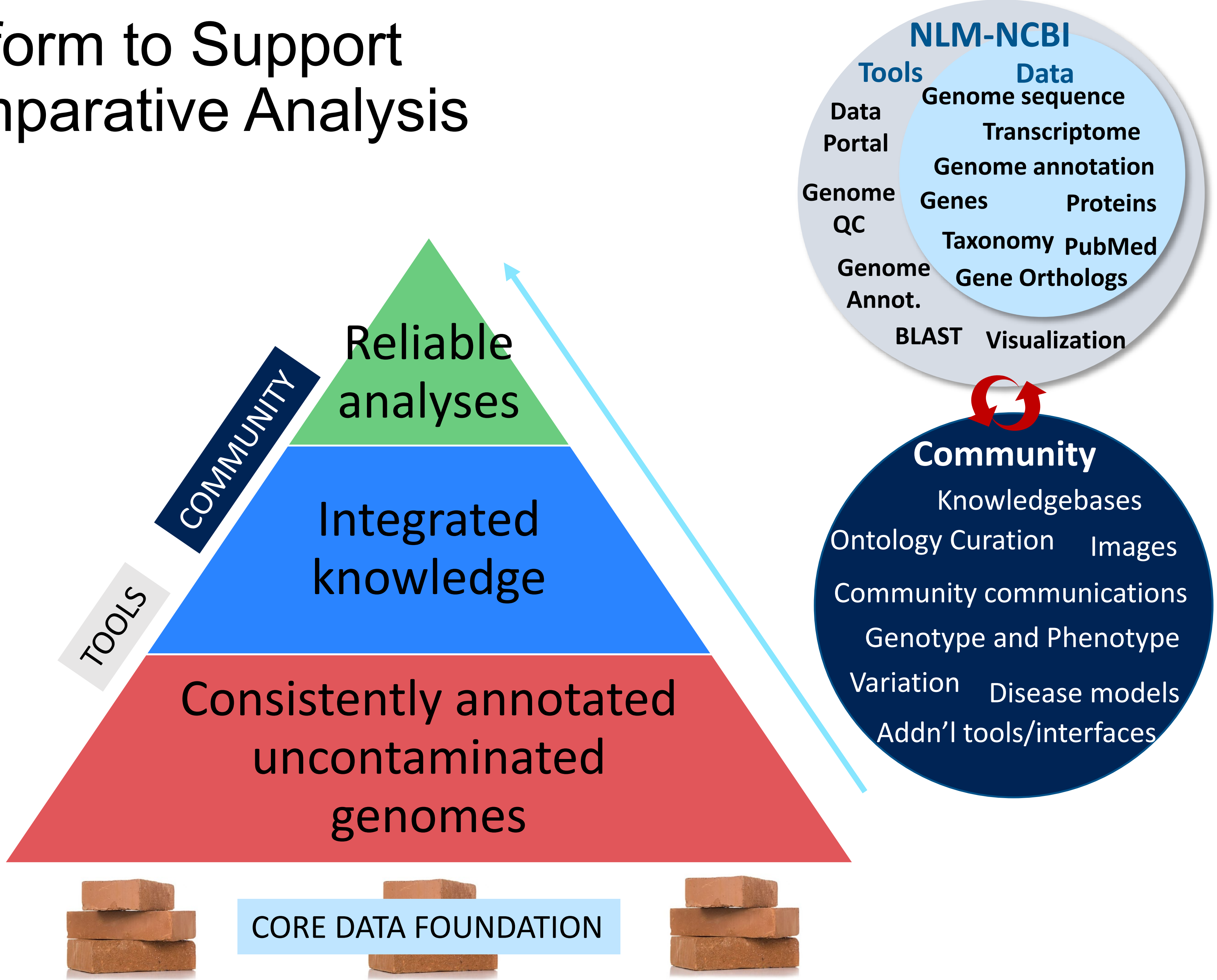


Beck et al.'s Agile Methodology

NIH National Library of Medicine
National Center for Biotechnology Information

Building a New Platform to Support Pan-Eukaryotic Comparative Analysis

- Organism-focused web portals and APIs
- Community engagement
- Public cloud-ready tools to screen and annotate all eukaryotic genomes



NLM-NCBI

Tools

Data Portal

Genome QC

Genome Annot.

BLAST

Visualization

Data

Genome sequence

Transcriptome

Genome annotation

Genes

Proteins

Taxonomy

PubMed

Gene Orthologs

Community

Knowledgebases

Ontology Curation

Images

Community communications

Genotype and Phenotype

Variation

Disease models

Addn'l tools/interfaces

COMMUNITY

TOOLS

Reliable analyses

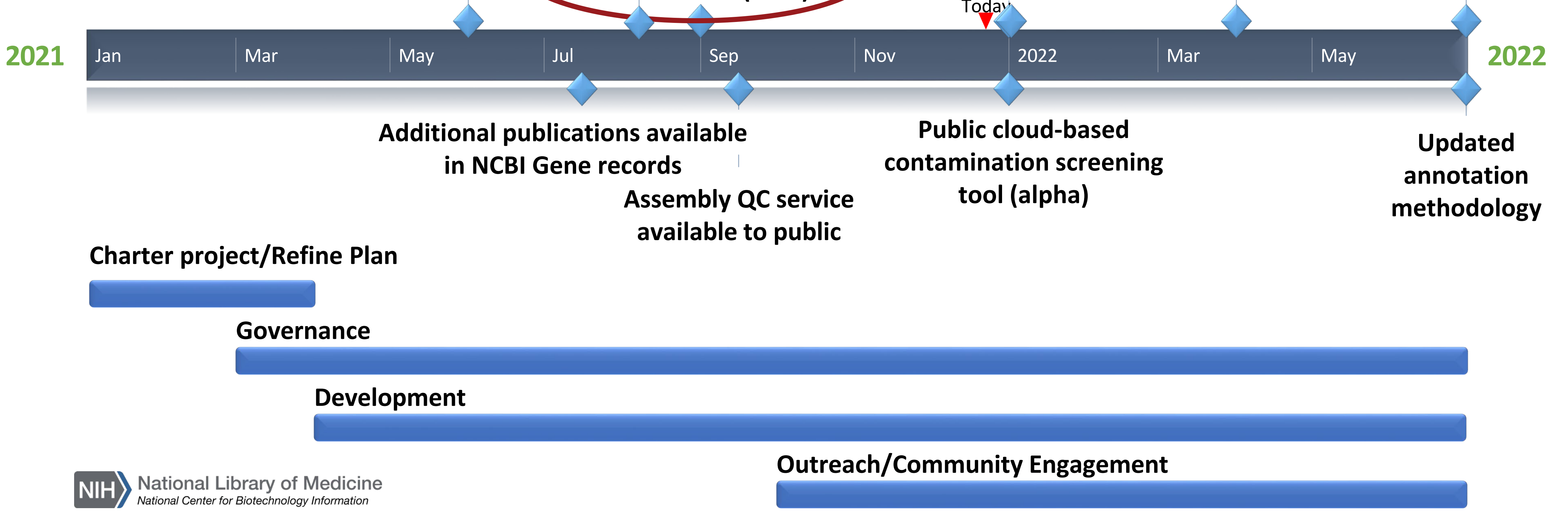
Integrated knowledge

Consistently annotated uncontaminated genomes

CORE DATA FOUNDATION

NIH National Library of Medicine
National Center for Biotechnology Information

Recent and Upcoming Highlights



Submitter-annotated assemblies displayed in genome browser

Portfolio and PubMed organism research

New web portal for genome-related data (beta)

CGR website

New BLAST databases

Comparative genome viewer (alpha release)

Expanded Gene Ortholog content

Additional publications available in NCBI Gene records

Assembly QC service available to public

Public cloud-based contamination screening tool (alpha)

Updated annotation methodology

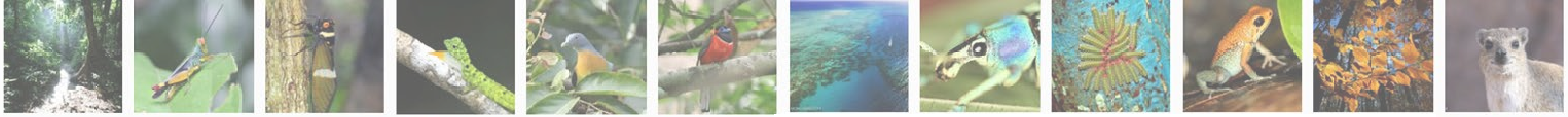
Charter project/Refine Plan

Governance

Development

Outreach/Community Engagement

NIH National Library of Medicine
National Center for Biotechnology Information



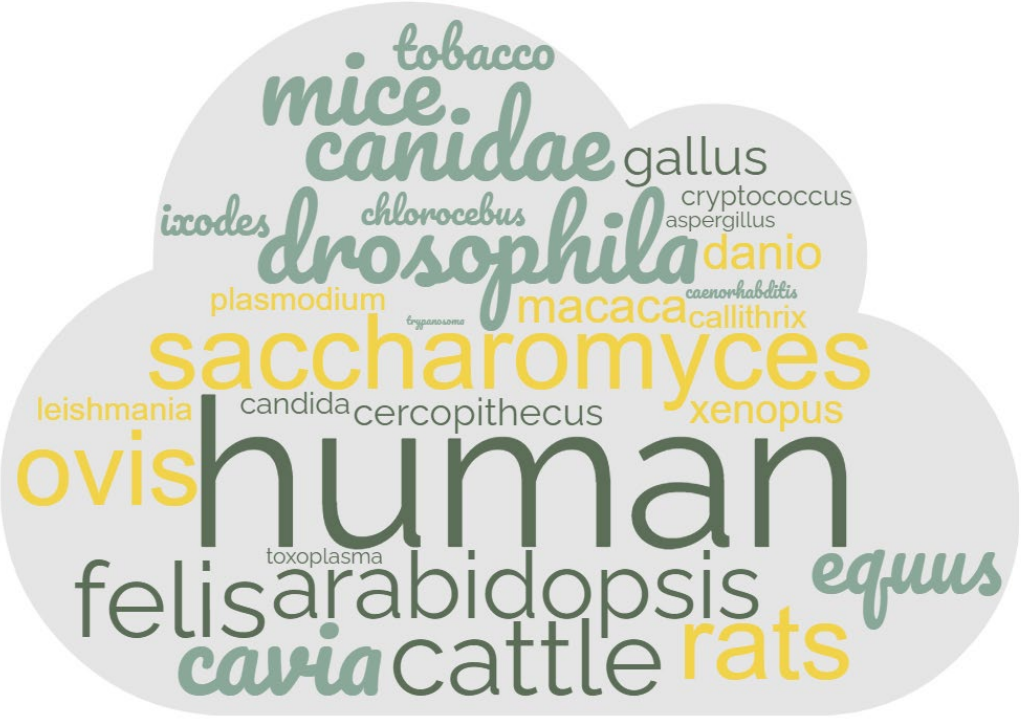
Organism Representation: Portfolio and PubMed

FY17-FY21

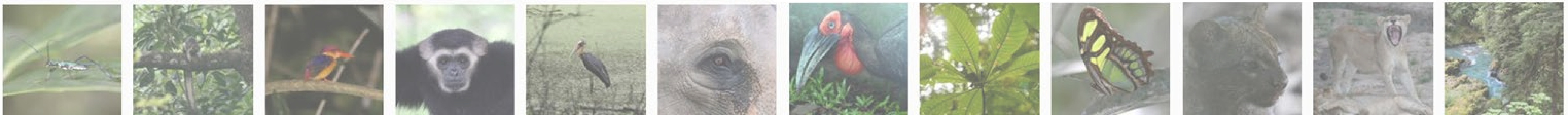
2015-2020

PubMed.gov

- Inputs
  - NIH Funded grants
  - PubMed publications
  - PubMed article views
- "Unweighted" analysis
- At Genus level:
  - Top 50/year
  - Total number/year
  - Most changed/year

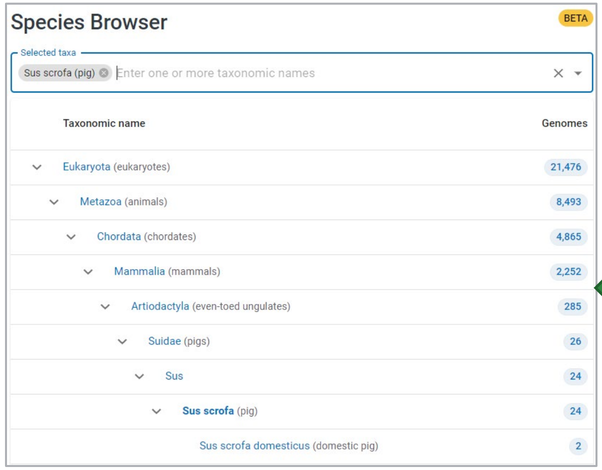


"Top 30"



https://www.ncbi.nlm.nih.gov/data-hub/taxonomy/9823/

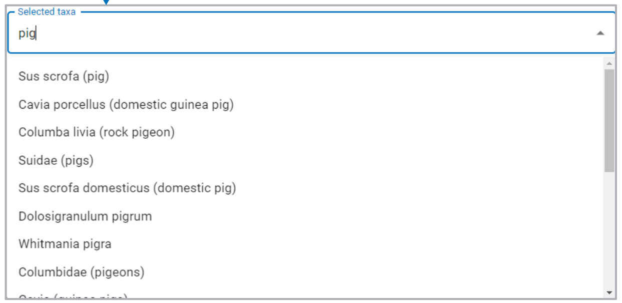
Browsable Taxonomy Tree



Species Browser BETA

Selected taxa: Sus scrofa (pig) ⊗ Enter one or more taxonomic names

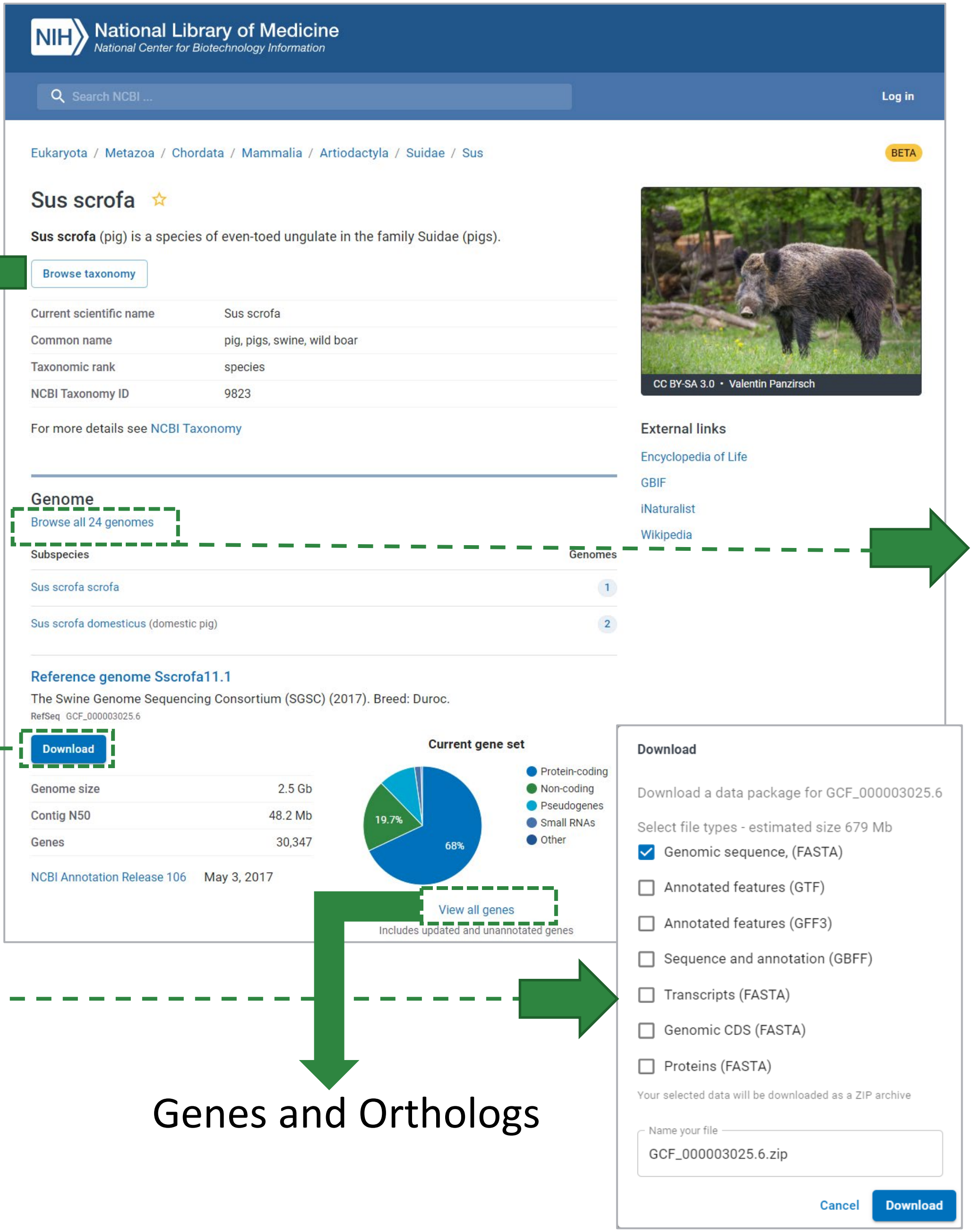
Taxonomic name	Genomes
Eukaryota (eukaryotes)	21,476
Metazoa (animals)	8,493
Chordata (chordates)	4,865
Mammalia (mammals)	2,252
Artiodactyla (even-toed ungulates)	285
Suidae (pigs)	26
Sus	24
Sus scrofa (pig)	24
Sus scrofa domesticus (domestic pig)	2



Selected taxa: pig

- Sus scrofa (pig)
- Cavia porcellus (domestic guinea pig)
- Columba livia (rock pigeon)
- Suidae (pigs)
- Sus scrofa domesticus (domestic pig)
- Dolosigranulum pigrum
- Whitmania pigra
- Columbidae (pigeons)

Taxonomic Autocomplete



NIH National Library of Medicine
National Center for Biotechnology Information
Search NCBI ...
Log in

Eukaryota / Metazoa / Chordata / Mammalia / Artiodactyla / Suidae / Sus BETA

Sus scrofa ☆

Sus scrofa (pig) is a species of even-toed ungulate in the family Suidae (pigs).

Browse taxonomy

Current scientific name	Sus scrofa
Common name	pig, pigs, swine, wild boar
Taxonomic rank	species
NCBI Taxonomy ID	9823

For more details see NCBI Taxonomy

CC BY-SA 3.0 • Valentin Panzirsch

External links

- Encyclopedia of Life
- GBIF
- iNaturalist
- Wikipedia

Genome

Browse all 24 genomes

Subspecies	Genomes
Sus scrofa scrofa	1
Sus scrofa domesticus (domestic pig)	2

Reference genome Sscrofa11.1

The Swine Genome Sequencing Consortium (SGSC) (2017). Breed: Duroc.

RefSeq GCF_000003025.6

Download

Genome size	2.5 Gb
Contig N50	48.2 Mb
Genes	30,347

NCBI Annotation Release 106 May 3, 2017

Current gene set

View all genes

Includes updated and unannotated genes

Download

Download a data package for GCF_000003025.6

Select file types - estimated size 679 Mb

- ☑ Genomic sequence, (FASTA)
- ☐ Annotated features (GTF)
- ☐ Annotated features (GFF3)
- ☐ Sequence and annotation (GBFF)
- ☐ Transcripts (FASTA)
- ☐ Genomic CDS (FASTA)
- ☐ Proteins (FASTA)

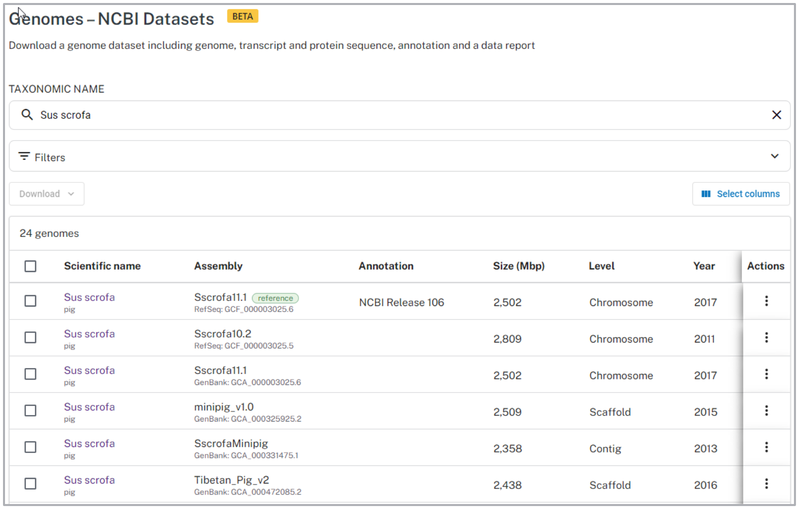
Your selected data will be downloaded as a ZIP archive

Name your file: GCF_000003025.6.zip

Cancel Download

Command-Line Tools API

Genomes



Genomes – NCBI Datasets BETA

Download a genome dataset including genome, transcript and protein sequence, annotation and a data report

TAXONOMIC NAME: Sus scrofa

24 genomes

☐	Scientific name	Assembly	Annotation	Size (Mbp)	Level	Year	Actions
☐	Sus scrofa pig	Sscrofa11.1 reference RefSeq: GCF_000003025.6	NCBI Release 106	2,502	Chromosome	2017	⋮
☐	Sus scrofa pig	Sscrofa10.2 RefSeq: GCF_000003025.5		2,809	Chromosome	2011	⋮
☐	Sus scrofa pig	Sscrofa11.1 GenBank: GCA_000003025.6		2,502	Chromosome	2017	⋮
☐	Sus scrofa pig	minipig_v1.0 GenBank: GCA_000325925.2		2,509	Scaffold	2015	⋮
☐	Sus scrofa pig	SscrofaMinipig GenBank: GCA_000331475.1		2,358	Contig	2013	⋮
☐	Sus scrofa pig	Tibetan_Pig_v2 GenBank: GCA_000472085.2		2,438	Scaffold	2016	⋮

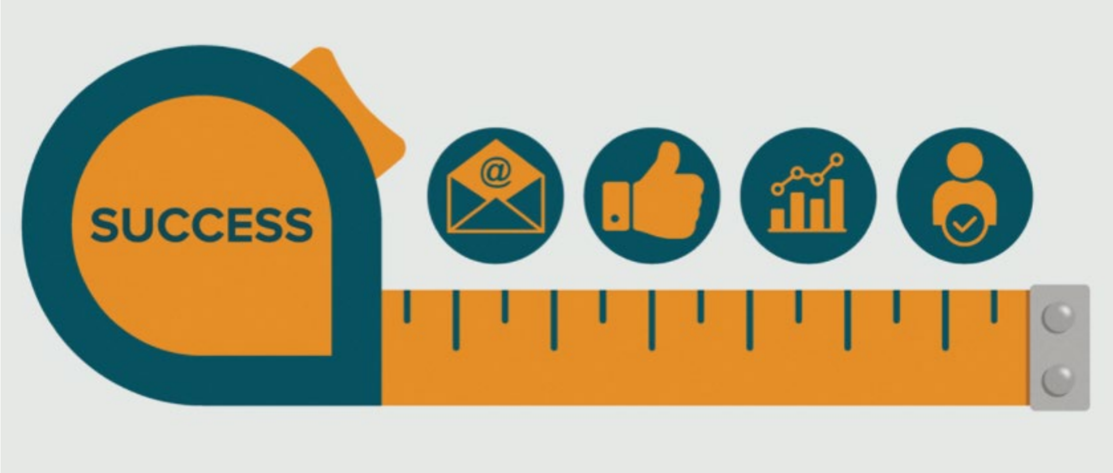
User friendly tables

Genes and Orthologs

Data Package Downloads

NIH National Library of Medicine
National Center for Biotechnology Information

CGR: Measuring Success



- Net Promoter Score
  - On a scale from 0-10, how likely are you to recommend our site to a friend, family member, or colleague?
- Quantification of tool usage, data/content submission, and data/content use
- Decrease in contaminated genome submissions
- Increase in genomes submitted with high quality annotations
- TBD: Measure scientific impact

We want your input!

How can the working group help NCBI connect with communities and share the CGR vision?

What criteria should NCBI use to prioritize organisms within CGR?

What measures will reveal the scientific impact of CGR?

Thank You

Kim Pruitt

Steve Sherry

Anatoly Mnev

Anne Ketter

Paul Ciprich

Kawaldeep Chadha

Jim Ostell

Terence Murphy

Françoise Thibaud-Nissen

Paul Kitts

Nuala O'Leary

Sanjida Rangwala

Tom Madden

Aron Marchler-Bauer

Wratko Hlavina

Peter Meric

Patti Brennan

Janet Coleman

Jodi Nurik

Diane Tuncer

Susan Gregurick

Rick Woychik

NIH CGR Steering Committee

Kristi Holmes