WormBase

Todd Harris, PhD

todd@wormbase.org @tharris

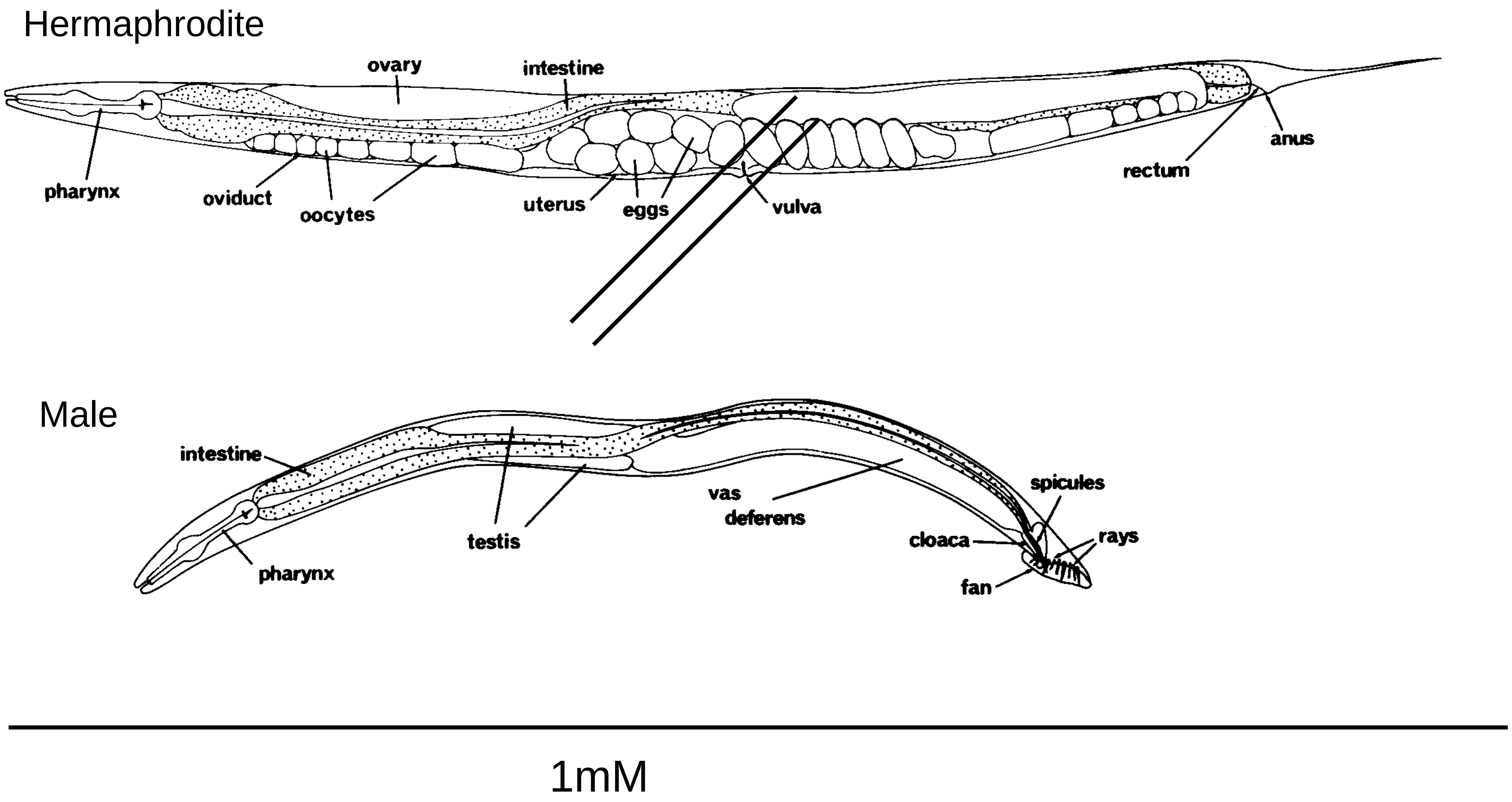
CBPSS Mini Symposium

Mission

Provide the biomedical research community with **accurate**, **current**, and **accessible** information on the genetics, genomics, and biology of the model system *Caenorhabditis elegans* and related nematodes.

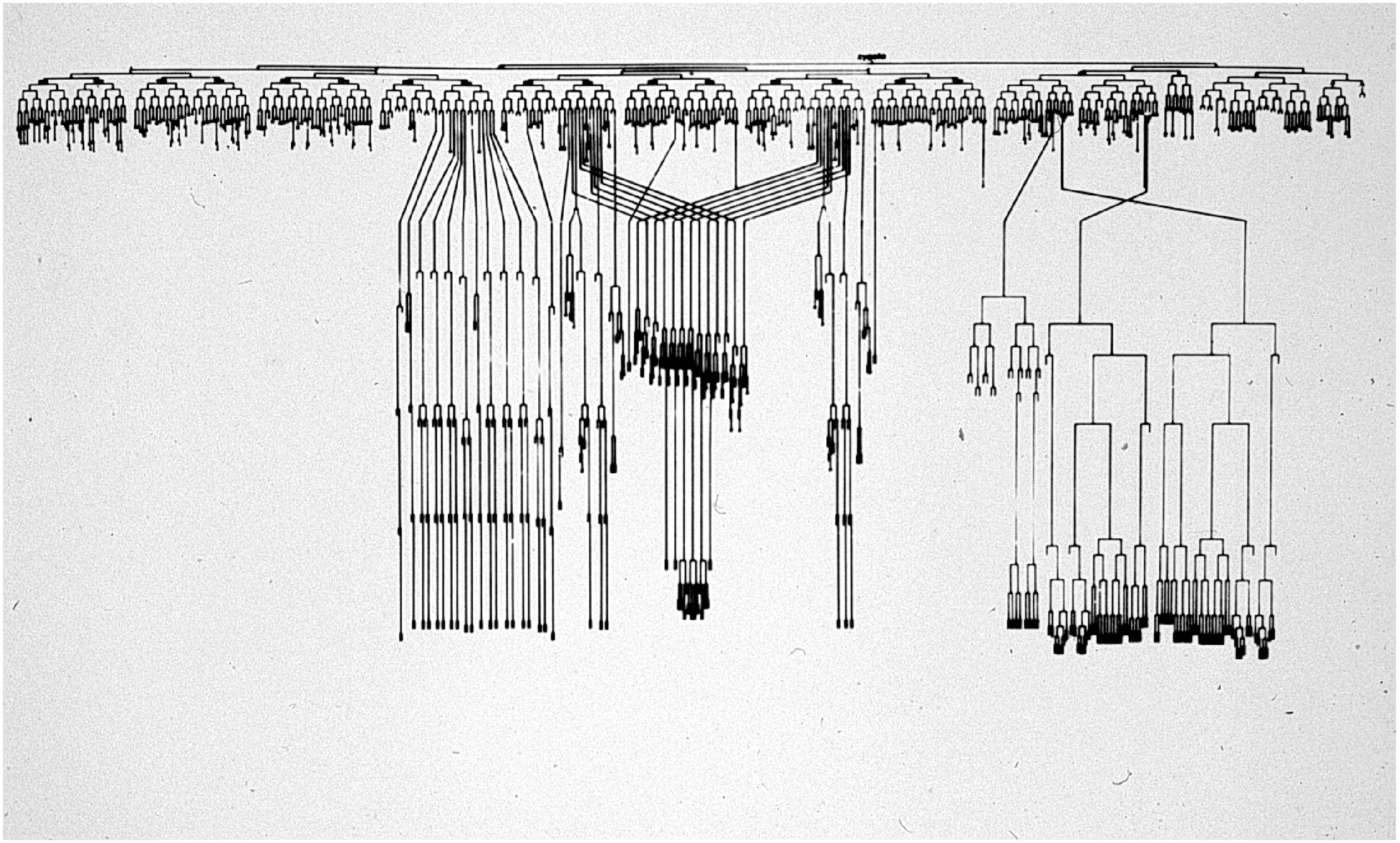
C. elegans in 30 seconds

Relatively simple organism, advanced genetic system.



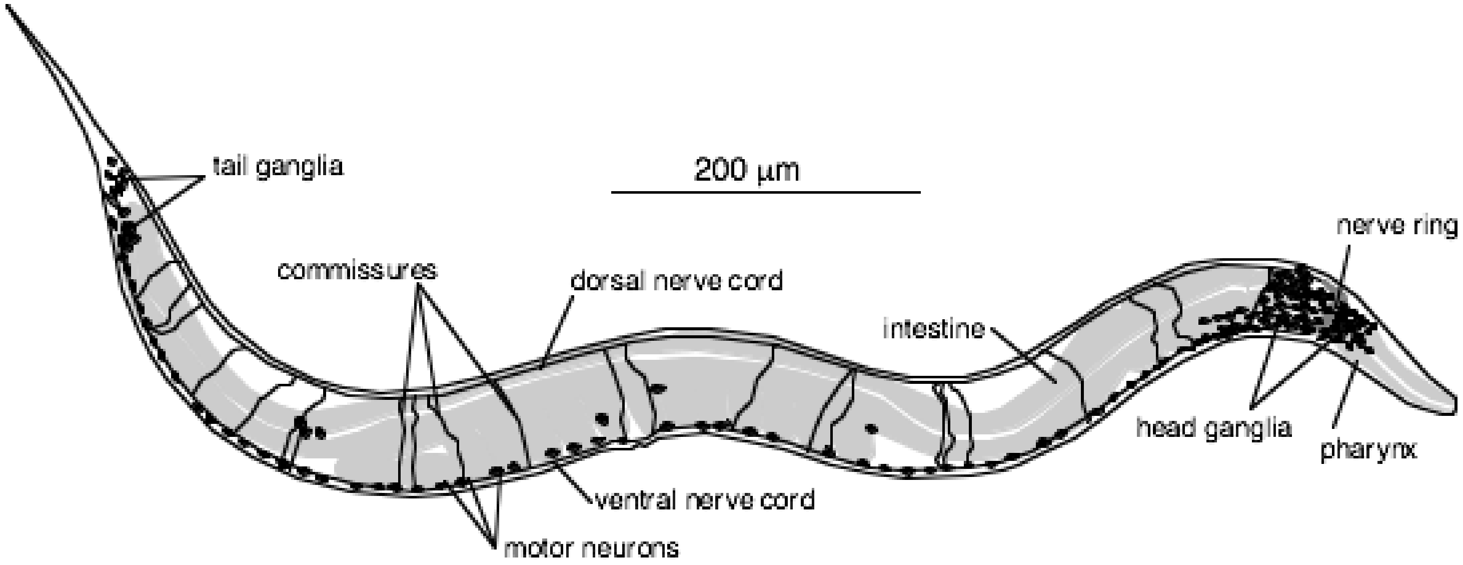
C. elegans in 30 seconds

Invariant lineage



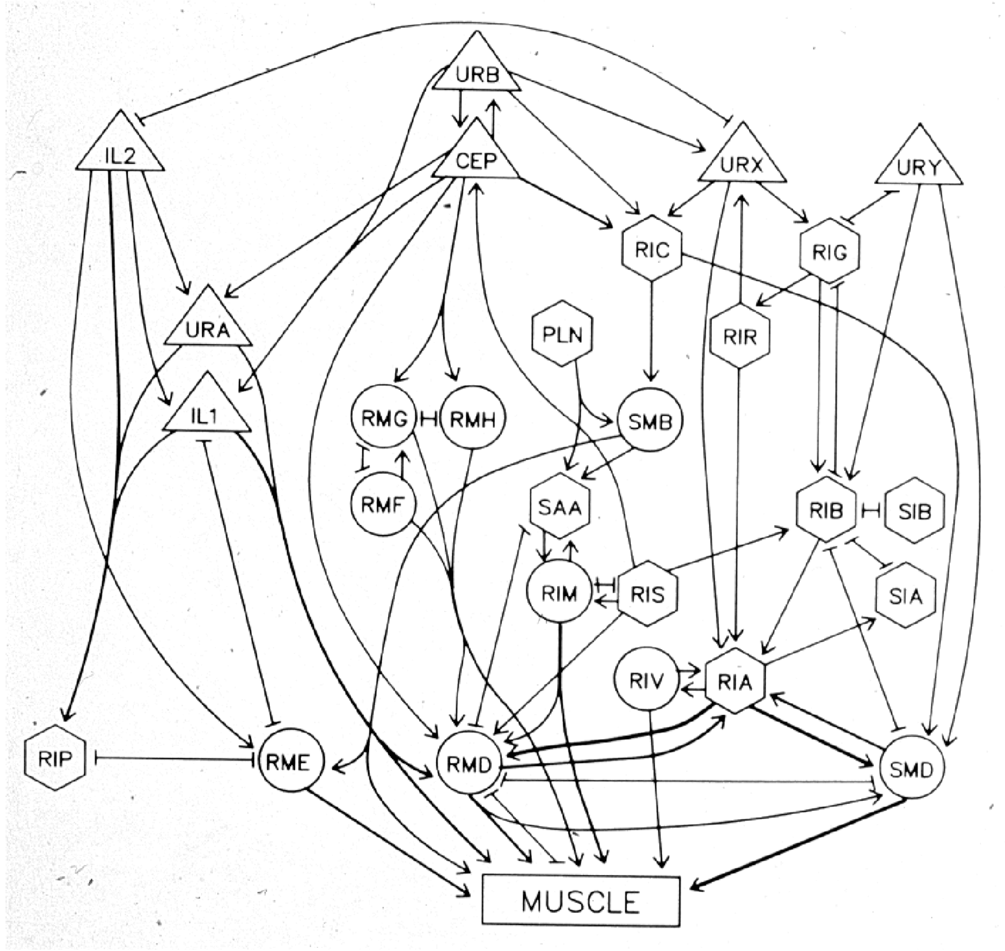
C. elegans in 30 seconds

Simple nervous system



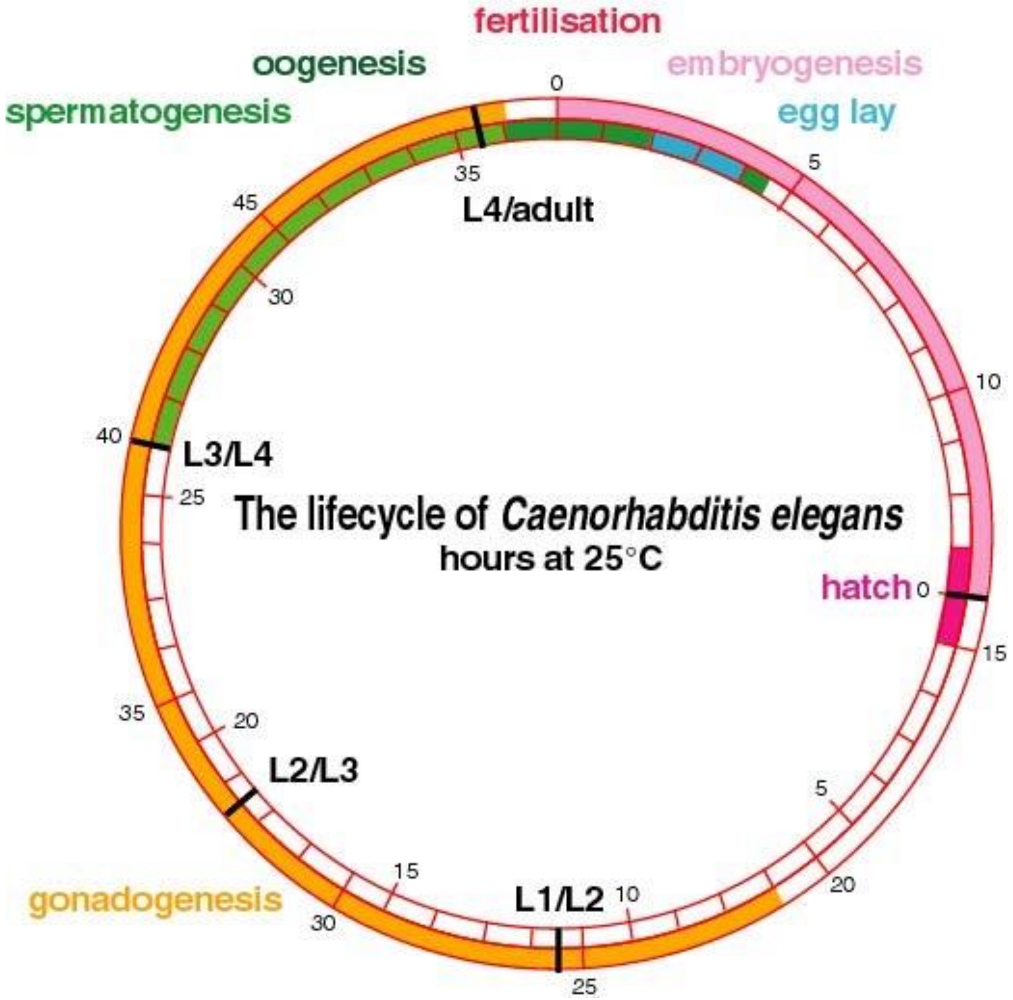
302 neurons

Described connectivity



C. elegans in 30 seconds

Rapid generation time



A frozen *C. elegans* library

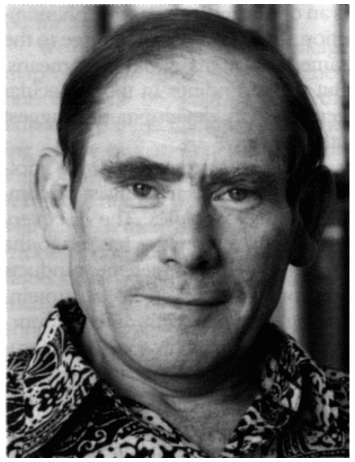
C. elegans in 30 seconds

100 MBp Genome

~20K genes

1998 (!)

A tradition of Open Science



Brenner's
Letters

1963

1st genetic screen
published

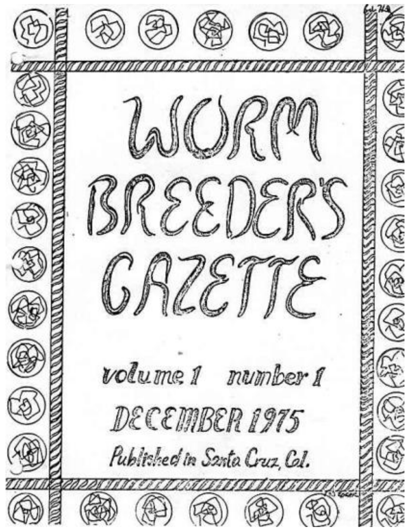
1974

AceDB
development
begins

1989

BioNet
www
gopher

1994



WORM
BREEDER'S
GAZETTE

volume 1 number 1
DECEMBER 1975
Published in Santa Cruz, Cal.

Gazette

1995

WormBase

2000

WormBook
THE ONLINE REVIEW OF *C. elegans* BIOLOGY

2003

The WormBase Consortium

CALIFORNIA INSTITUTE OF TECHNOLOGY
1891

EMBL-EBI

OICR
Ontario Institute
for Cancer Research

CSH

WASHINGTON UNIVERSITY IN ST. LOUIS
Per
Veritatem
Vis
· 1853 ·

User Community

Registered *C. elegans* laboratories

Country	Labs
United States	594
Canada	62
United Kingdom	60
Japan	58
Germany	48
France	31
China	28
Spain	20
Switzerland	20
The Netherlands	16

1106 laboratories

53 countries

3000 researchers

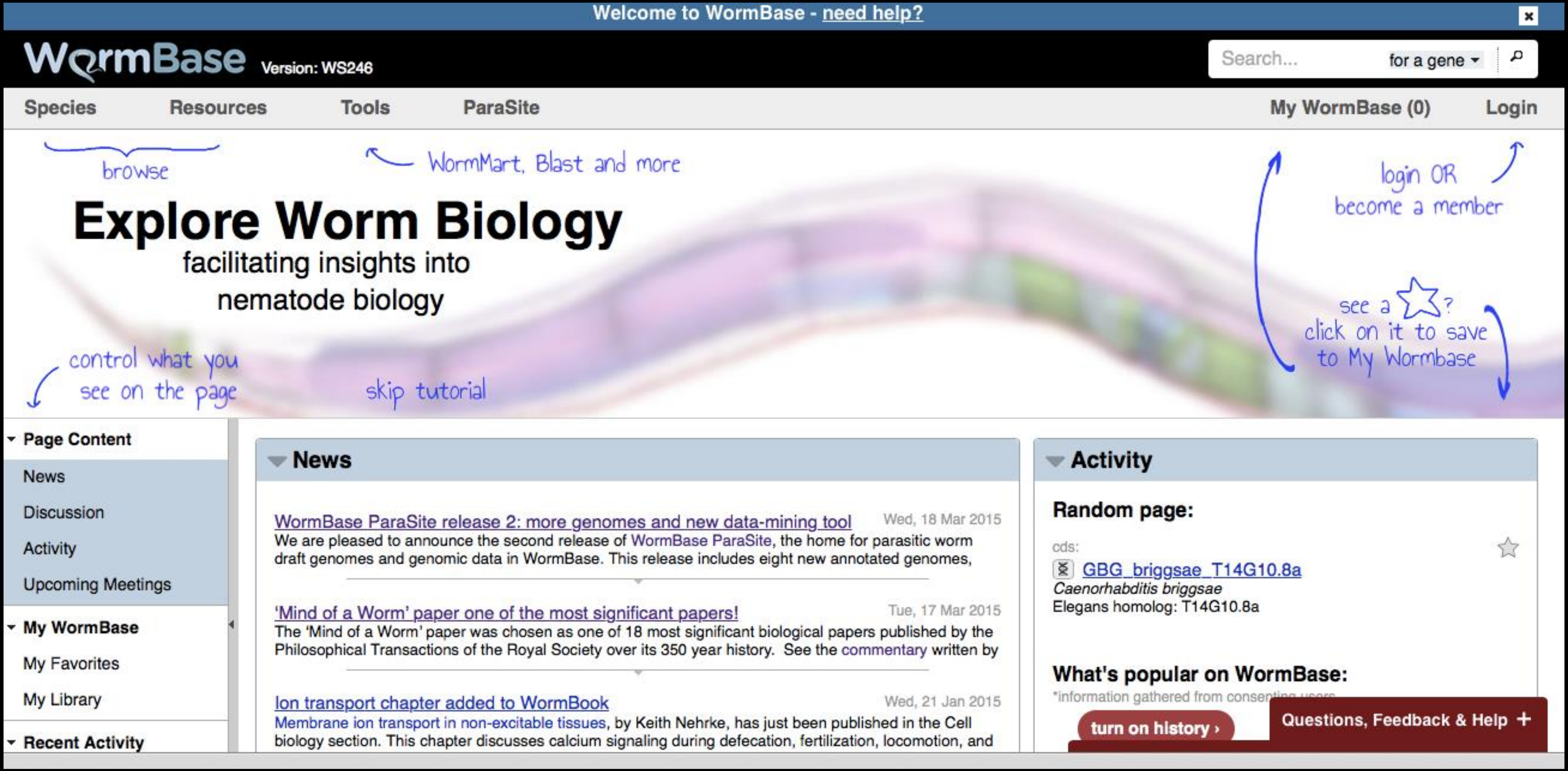
User Community

Biomedical researchers studying aging, neurobiology, cancer, etc.

185 **countries**

37K **unique users/month**

5.5M **page views / month**



Welcome to WormBase - need help?

WormBase Version: WS246

Search... for a gene

Species | Resources | Tools | ParaSite | My WormBase (0) | Login

browse
WormMart, Blast and more
login OR become a member
see a ☆? click on it to save to My Wormbase
control what you see on the page
skip tutorial

Explore Worm Biology

facilitating insights into nematode biology

Page Content
- News
- Discussion
- Activity
- Upcoming Meetings

My WormBase
- My Favorites
- My Library

Recent Activity

News

WormBase ParaSite release 2: more genomes and new data-mining tool — Wed, 18 Mar 2015

We are pleased to announce the second release of WormBase ParaSite, the home for parasitic worm draft genomes and genomic data in WormBase. This release includes eight new annotated genomes,

'Mind of a Worm' paper one of the most significant papers! — Tue, 17 Mar 2015

The 'Mind of a Worm' paper was chosen as one of 18 most significant biological papers published by the Philosophical Transactions of the Royal Society over its 350 year history. See the commentary written by

Ion transport chapter added to WormBook — Wed, 21 Jan 2015

Membrane ion transport in non-excitable tissues, by Keith Nehrke, has just been published in the Cell biology section. This chapter discusses calcium signaling during defecation, fertilization, locomotion, and

Activity

Random page:

cds:
GBG_briggsae_T14G10.8a
Caenorhabditis briggsae
Elegans homolog: T14G10.8a

What's popular on WormBase:

*information gathered from consenting users

turn on history ›

Questions, Feedback & Help +

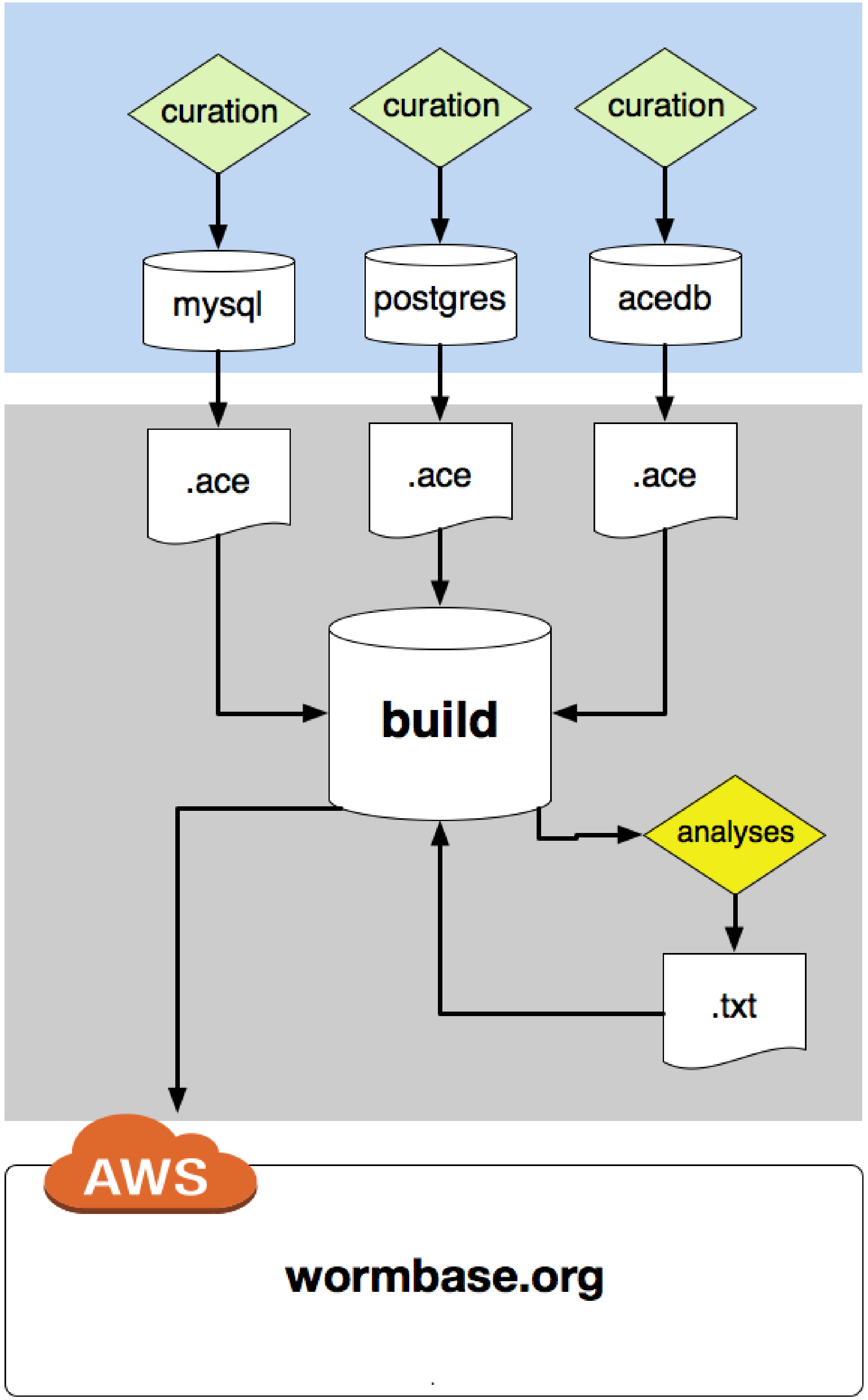
wormbase.org

Contents & Features

28 Species
Genomes
Genes
Orthology / Homology / Paralogy
Comparative Genomics
Strains / Antibodies / Oligos
Expression
Lineage & Connectivity
Authors & Publications
Labs

Reports
Genome Browsers
Alignment Tools
Query Tools
APIs
Data Mining Platforms
Social Features
FTP
Forums, Wikis, Blogs

Workflow



1. Curation

2. Integration & analysis

3. Presentation

Curation Goals

1. Extract data from the scientific literature.

2. Develop standards to structure data.

3. Facilitate new insights by making prose observations computable.

Curated Sources

Scientific literature (~30K papers)

User submissions

Genomic sequences (gene models)

3rd party datasets

Early Realizations

Curation is hard and time-consuming!

Requires automation.

Need tools to facilitate.

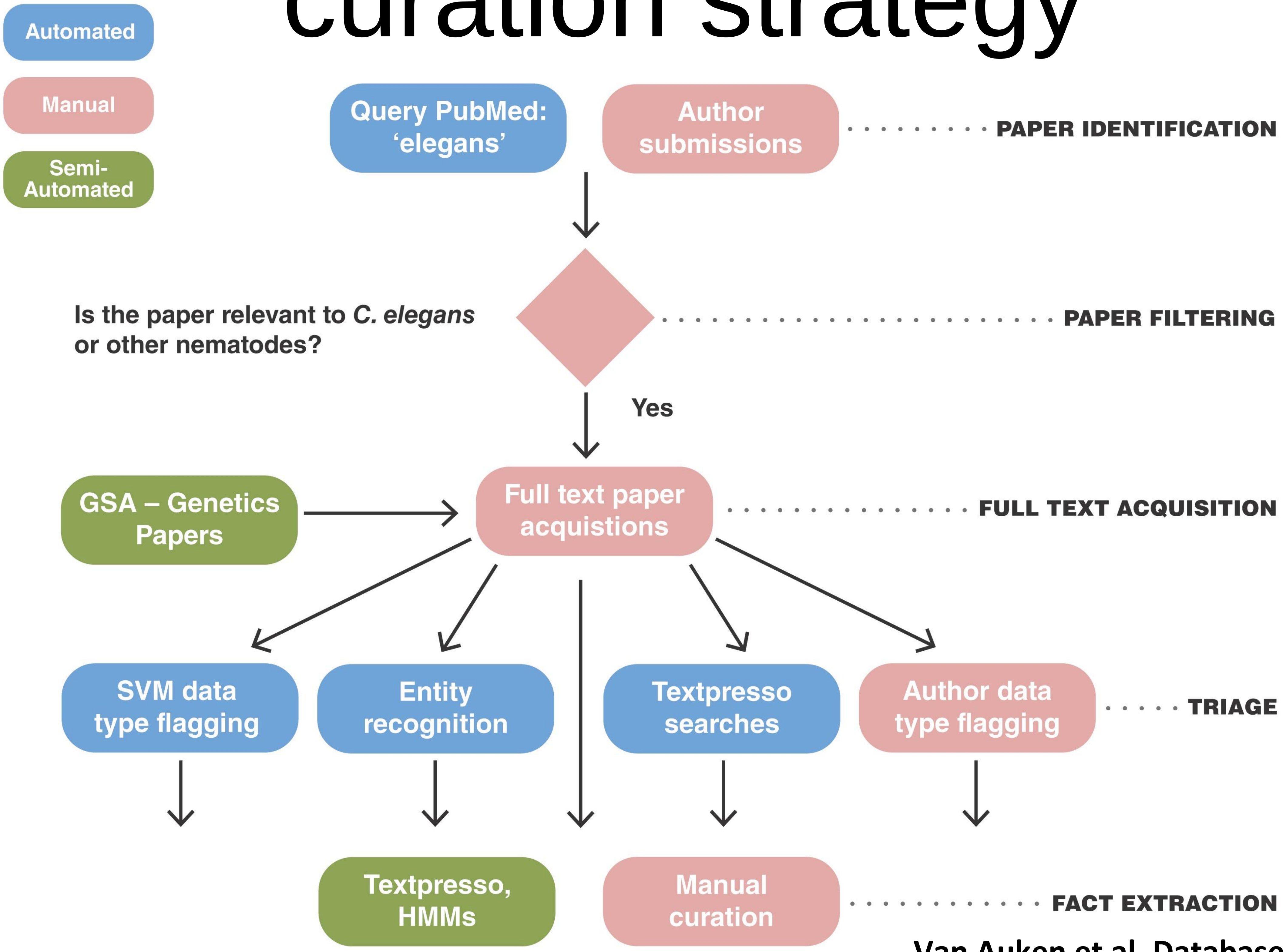
Many data types.

Prioritization is key.

Work procedurally through data types.

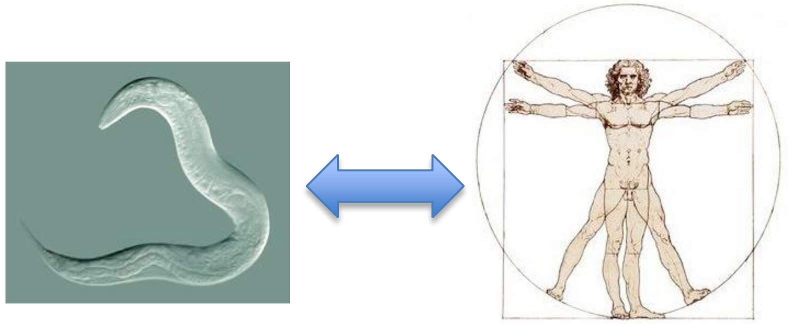
Balance of breadth **and** depth critical for making useful community resource.

Hybrid automated/manual curation strategy

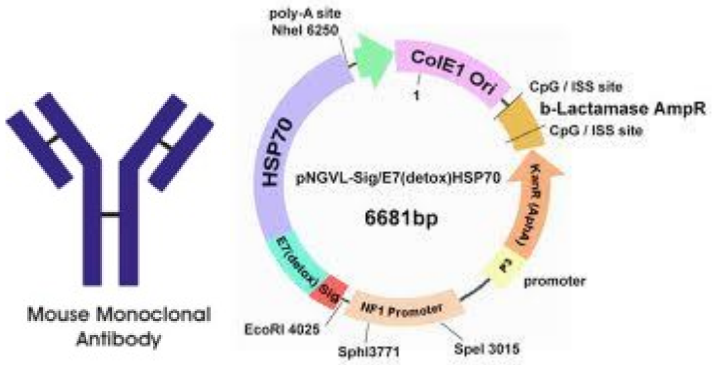


Van Auken et al, Database, 2012

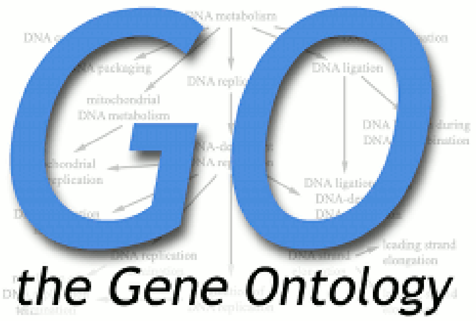
Curated data types

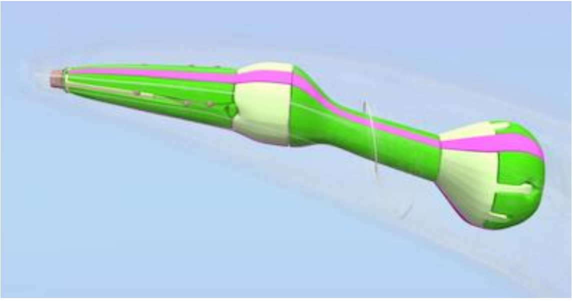


Human Disease Relevance

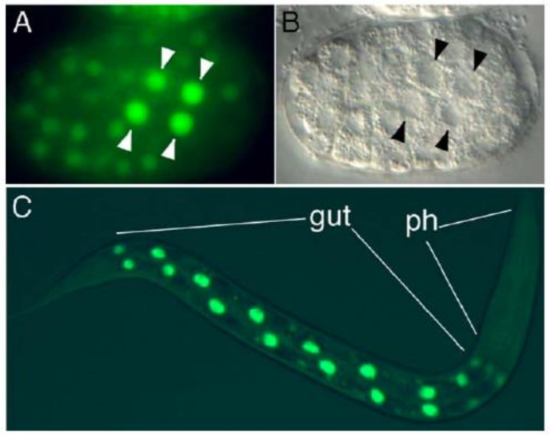


Reagents

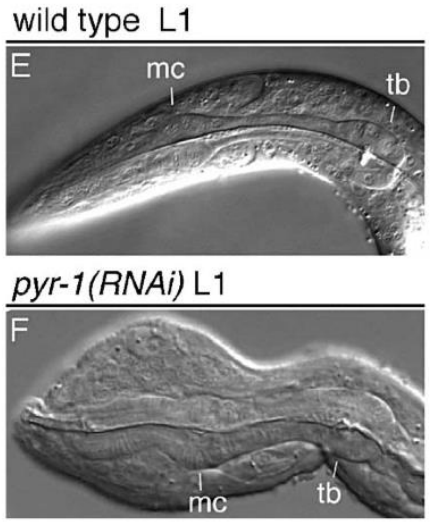




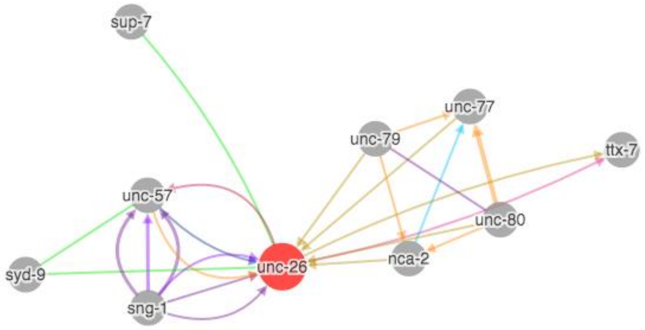
Anatomy Function



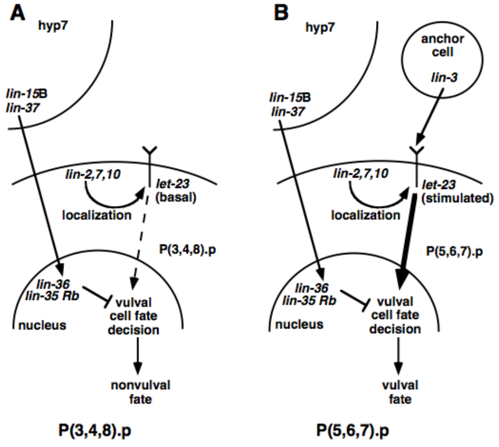
Expression



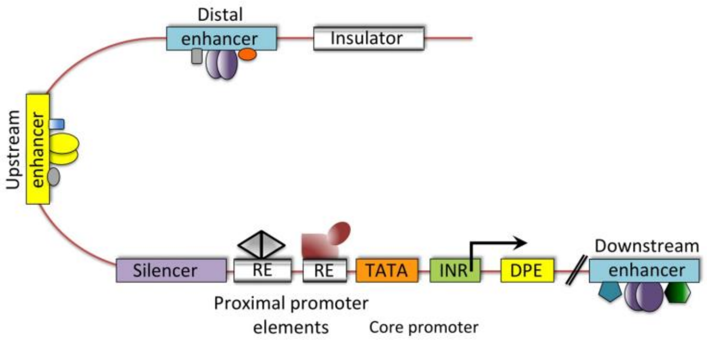
Phenotypes



Gene Interactions



Pathways



Sequence Features

Reference datasets

Large scale data at WormBase

- Proteomics (mass spec)
- Transcriptomics (splicing, UTRs)
- Expression (microarray, *in vivo* imaging)
- Interactions (physical, genetic)
- Perturbation: RNAi, systematic mutation
- Lineage and connectivity

Reference datasets

Broad reference data sets can fill knowledge gaps

- Verification can be difficult
- Relevance?
- Utilization varies greatly. Confidence?

Do we assess the quality of…

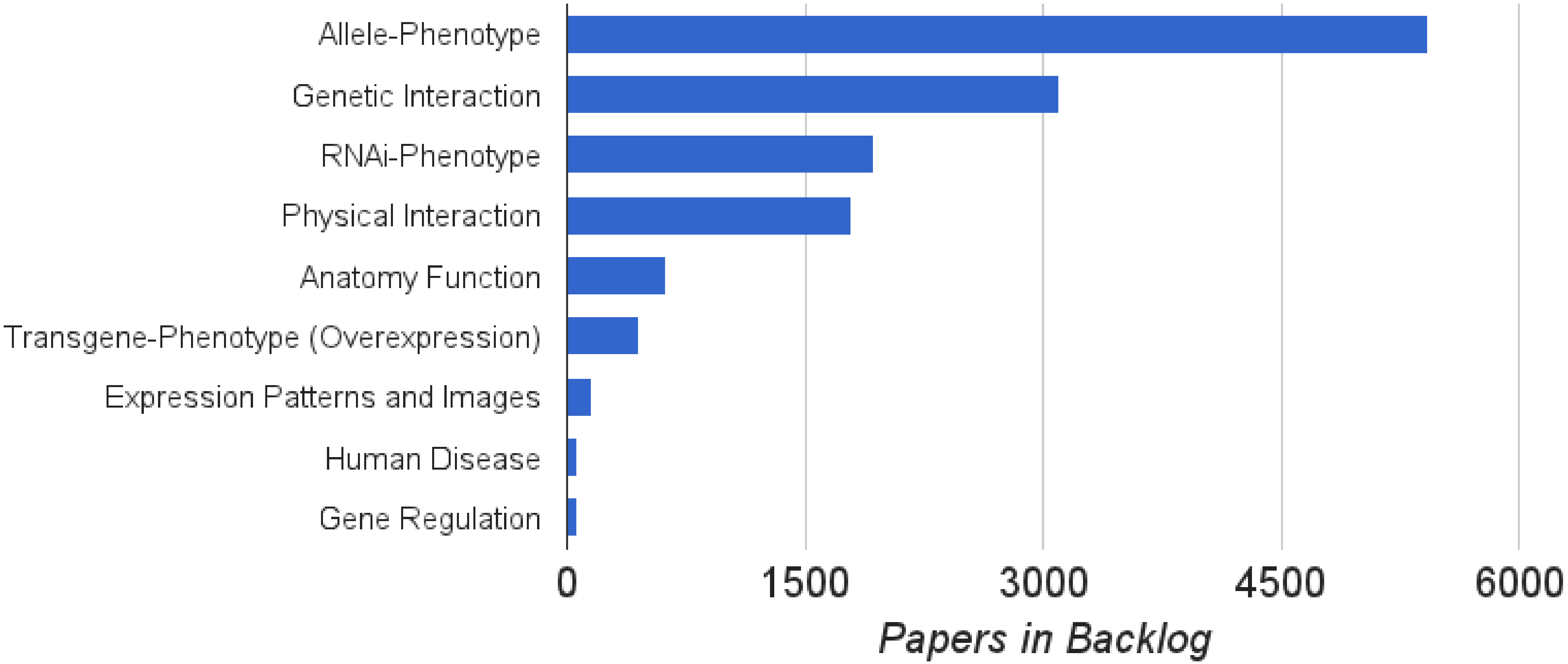
experimental design? external data?

Publication is the gold standard.

Revisit: erroneous data

Request corrections or clarifications when warranted

Remaining backlog



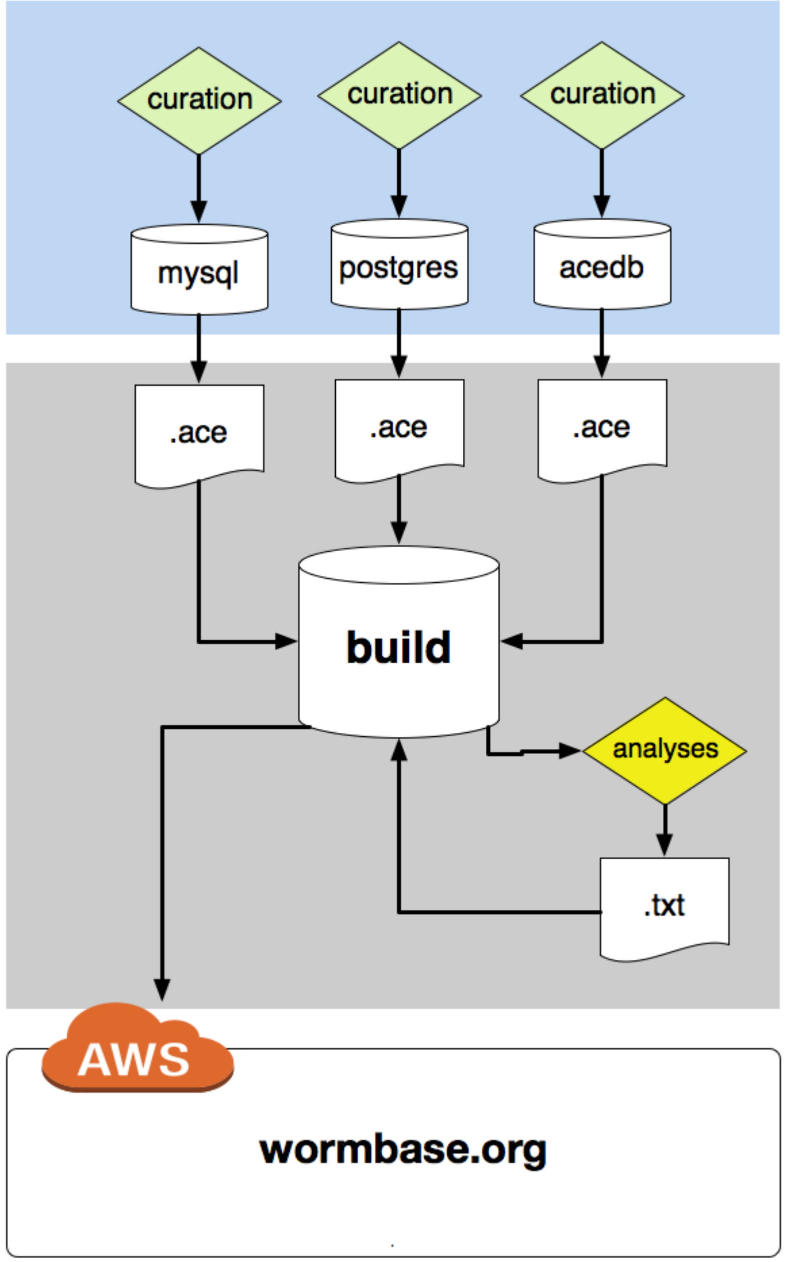
Curation: Lessons Learned

- **harder** and **consumes more time** than expected
- more **enriching** to the final product than expected
- curation ensures data integrity and builds **trust** in the resource

Curation: Suggestions

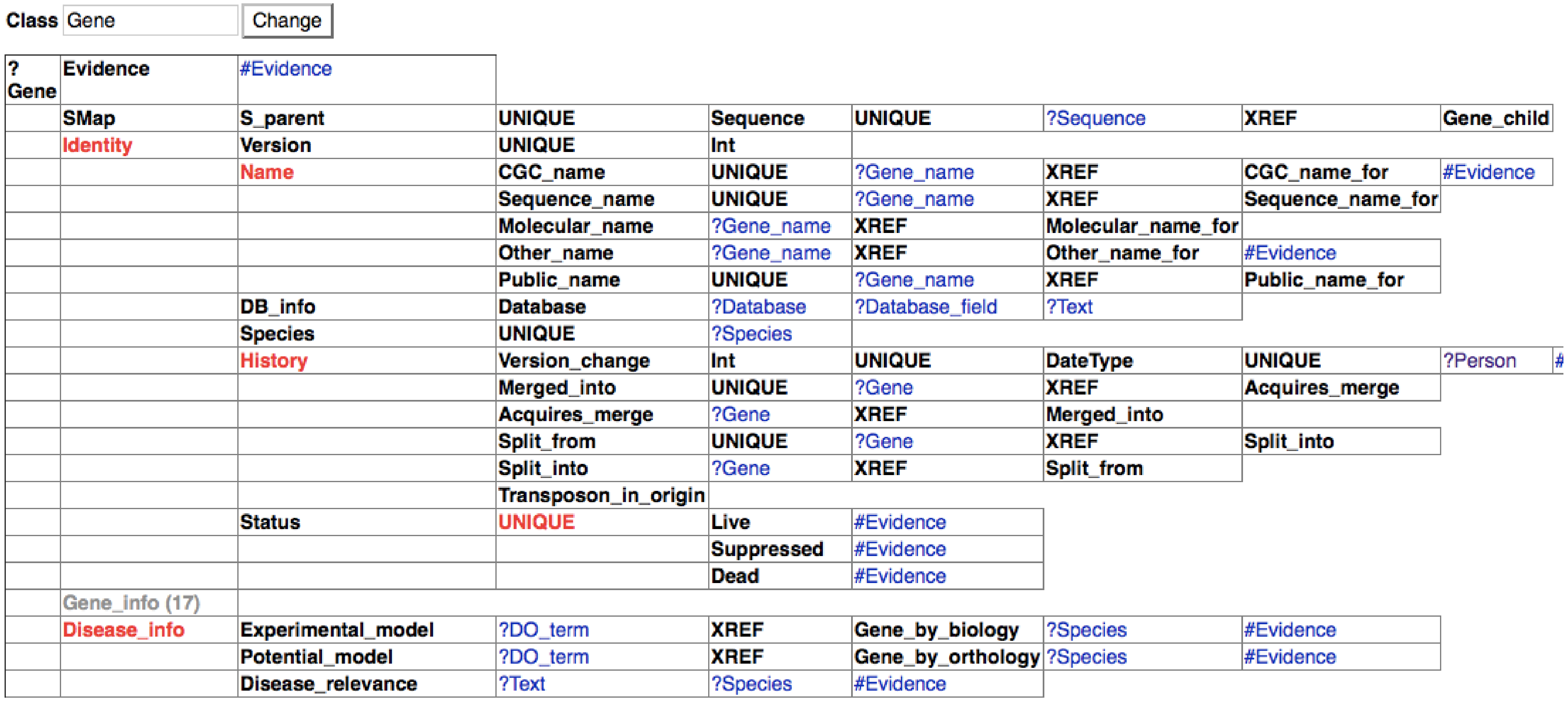
- Start early to develop best practices.
- Automate as much as possible.
- Employ domain experts for high value **manual curation** and to confirm **precision** of automated curation.
- **Expect** publication rate and new data types to **exceed** manual curation capacity (10% Y-o-Y).
- **Refining** curation will be an ongoing enterprise.

What fundamentals have driven our workflow design?



What fundamentals have driven our design?

1. Ease of data modeling and loading



Class Gene Change

? Gene	Evidence	#Evidence							
	SMap	S_parent	UNIQUE	Sequence	UNIQUE	?Sequence	XREF	Gene_child	
	Identity	Version	UNIQUE	Int					
		Name	CGC_name	UNIQUE	?Gene_name	XREF	CGC_name_for	#Evidence	
			Sequence_name	UNIQUE	?Gene_name	XREF	Sequence_name_for		
			Molecular_name	?Gene_name	XREF	Molecular_name_for			
			Other_name	?Gene_name	XREF	Other_name_for	#Evidence		
			Public_name	UNIQUE	?Gene_name	XREF	Public_name_for		
		DB_info	Database	?Database	?Database_field	?Text			
		Species	UNIQUE	?Species					
		History	Version_change	Int	UNIQUE	DateType	UNIQUE	?Person	#
			Merged_into	UNIQUE	?Gene	XREF	Acquires_merge		
			Acquires_merge	?Gene	XREF	Merged_into			
			Split_from	UNIQUE	?Gene	XREF	Split_into		
			Split_into	?Gene	XREF	Split_from			
			Transposon_in_origin						
		Status	UNIQUE	Live	#Evidence				
				Suppressed	#Evidence				
				Dead	#Evidence				
	Gene_info (17)								
	Disease_info	Experimental_model	?DO_term	XREF	Gene_by_biology	?Species	#Evidence		
		Potential_model	?DO_term	XREF	Gene_by_orthology	?Species	#Evidence		
		Disease_relevance	?Text	?Species	#Evidence				

Emphasis on collecting and sharing data.

What fundamentals have driven our design?

2. Handling unknown unknowns

Yet-to-be-discovered …

- datatypes

- data relationships

Data model must be able to evolve.

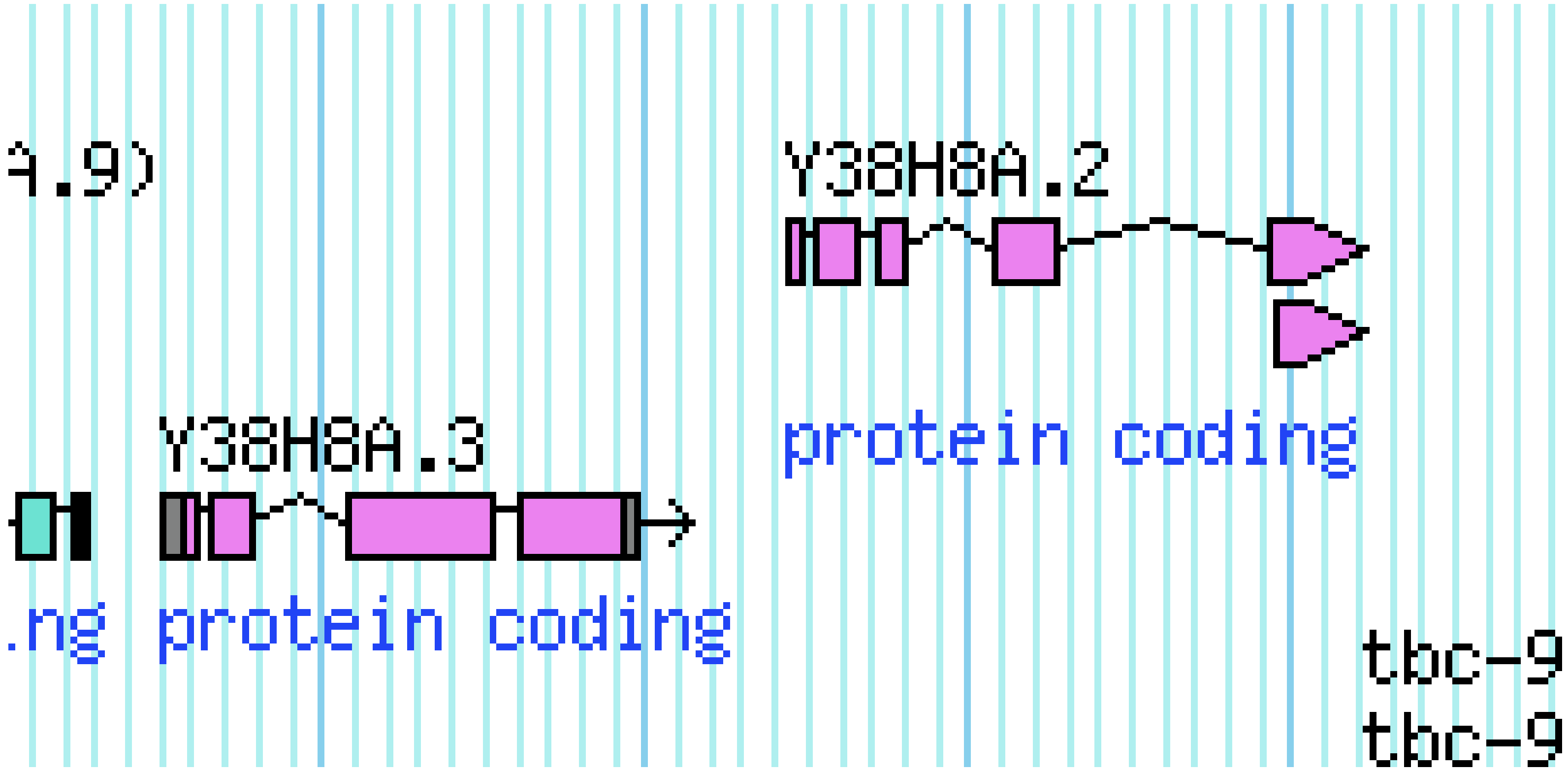
What fundamentals have driven our design?

3. **Ability to track supporting evidence, metadata, and provenance**

Reproducibility and accountability.

What fundamentals have driven our design?

4. Coping with high-connectivity data



eg: What happens to downstream annotations if gene merge? Orthology, proteomics, expression, etc…

What fundamentals have driven our design?

5. Finding a suitable refresh rate

Datasets evolve. New data becomes available. Analyses need to be updated.

How often will you update analyses?

How tolerant will your community be of ***stale data****?*

What fundamentals have driven our design?

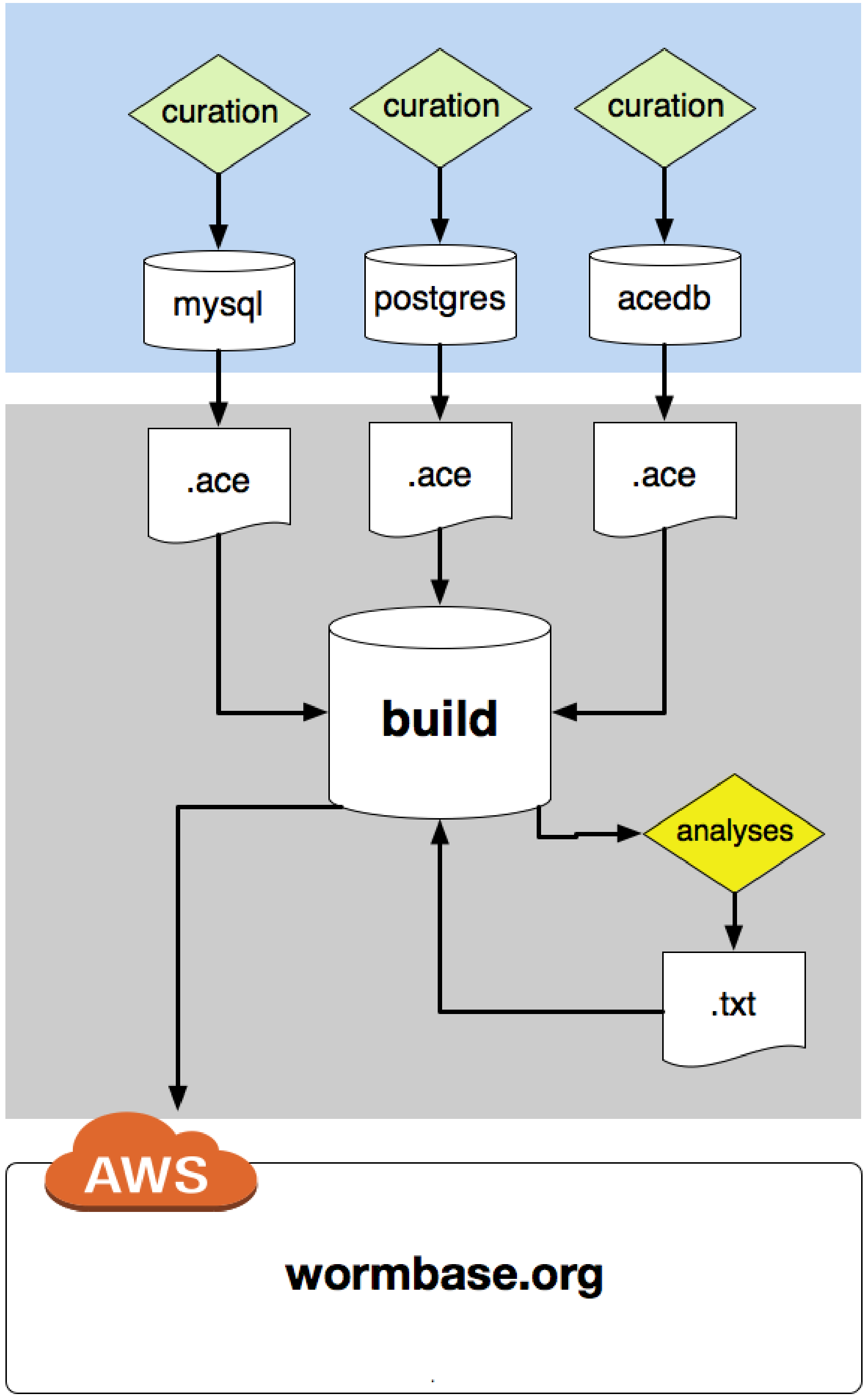
5. Finding a suitable refresh rate

1 week -> 2 weeks -> 3 weeks -> 1 month -> 2 months

2001 2002 2005 2008 2011

Balance of stability, rate of new data, cost/time of analysis, churn.

Design: Lessons Learned



1. **A flexible model/workflow is essential.**
2. **Evidence and metdata collection needs to be central to process.**
3. **High connectivity data presents unique challenges.**
4. **Needed to adjust release frequency.**

Design: Suggestions

1. Build flexibility into both the data model and workflow.

2. Be aware of consequences of changing high connectivity data.

3. Refresh frequency is a balance of user needs, resources, and rate of change.

Integration & Interoperability

Suggestions for integrating with organismal databases (easy)

- Liaise with organismal databases early and often!

- Use **stable identifiers**! Most organism databases have them. Please?

Suggestions for integrating with organismal databases (harder)

Reciprocal data exchange and cross links

Crosslinks alone are boring and do not engage users.

Without some supporting context, crosslinks do not increase interoperability.

Suggestions for integrating with organismal databases (hardest)

Avoid direct data import

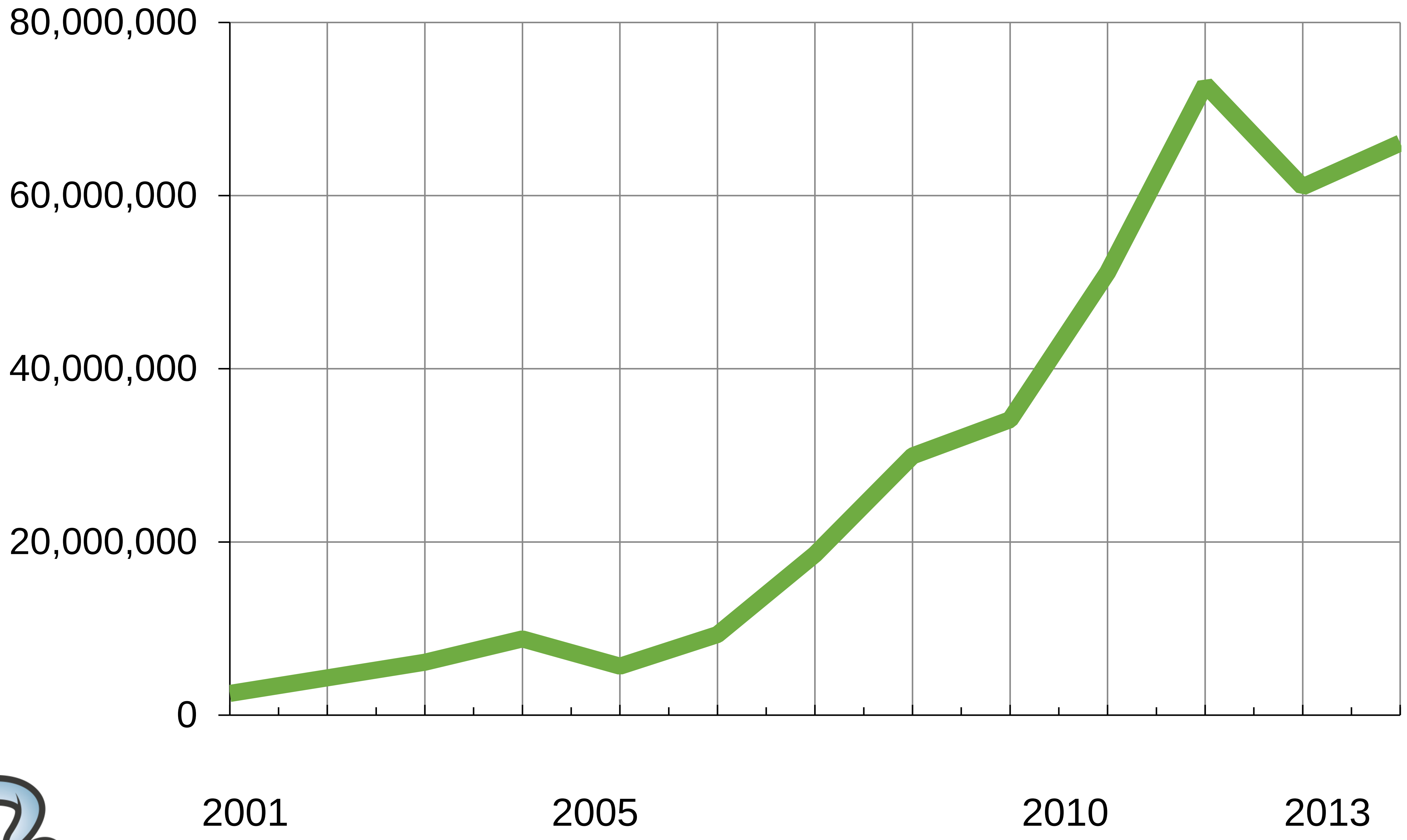
*Except for core scaffolding features (genomes, genes, eg), use **APIs** to fetch and embed functional data.*

Interoperability Suggestions

1. Provide data in (multiple) common formats
2. API (RESTful) with JSON and XML delivery
3. Data files programmatically accessible — simple is better (FTP), no registration barrier or fancy web-based download scheme.
4. Consistent, shared identifiers

If you build it, will they come?

Pageviews vs time



Nurture Your Community

Collect feedback

Chat, Twitter, Google Alerts, mailing lists, conferences, webinars, surveys.

Measure

Web logs, CloudWatch, Google Analytics

Set standards

Data quality, curation, submission, help desk response times.

Metrics of success

Not easy to measure.

Small user communities, niche domains.

Providing annotation or feedback is a low priority for busy scientists.

Positive feedback rare, but you’ll **know** when users don’t like something!

Suggested Metrics

- Page Views
- Citation Rate
- Downloads
- Queries & Resolutions
- Rate / precision of curation
- Database size / objects / submissions

Performance Metrics

Acknowledgments

@tharris

@wormbase

CALIFORNIA INSTITUTE OF TECHNOLOGY 1891

Paul Sternberg
Juancarlos Chan
Wen Chen
Chris Grove
Raymond Lee
Ranjana Kishore
Cecilia Nakamura
Daniela Raciti
Gary Schindelman
Mary Ann Tuli
Kimberly Van Auken
Xiaodong Wang
Karen Yook
Hans-Michael Muller
Yuling Li
James Done

OICR
Ontario Institute
for Cancer Research

Lincoln Stein
Sibyl Gao
Todd Harris

EMBL-EBI

Matt Berriman
Paul Kersey
Paul Davis
Thomas Done
Kevin Howe
Michael Paulini
Gary Williams

National Human
Genome Research
Institute

MRC
Medical
Research
Council