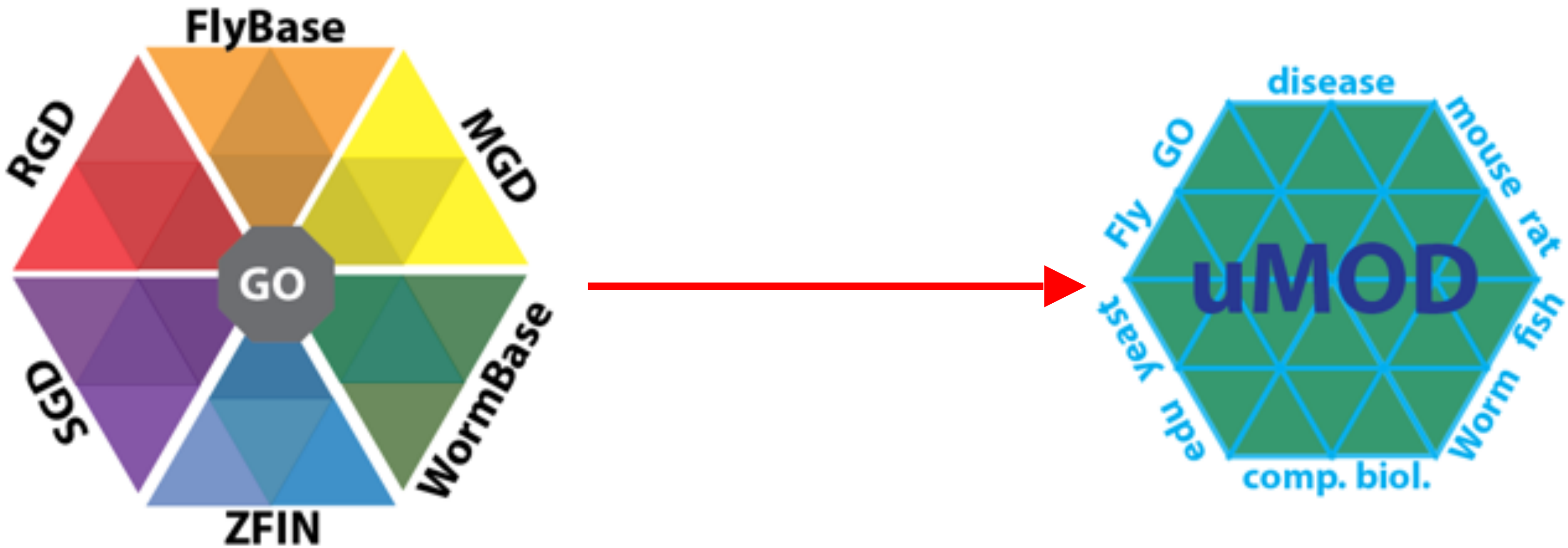
Bringing the MODs together to create the *Alliance of Genome Resources*



1. Who?
2. Why?
3. How? Long and short term goals
4. Funding implications
5. Should smaller mods to work towards this sort of integration too?

Impetus for unifying the MODs (from NHGRI)

User confusion for lack of homogeneity

- User access interfaces
  - need different navigation skills and data access approaches for each resource
  - Semantic inconsistencies and different data structures for the same genomic entities
- Analyses
  - human/model organism association for disease and phenotypes
  - functional annotation
  - Homology representation

Redundancy of operations at 6 resources

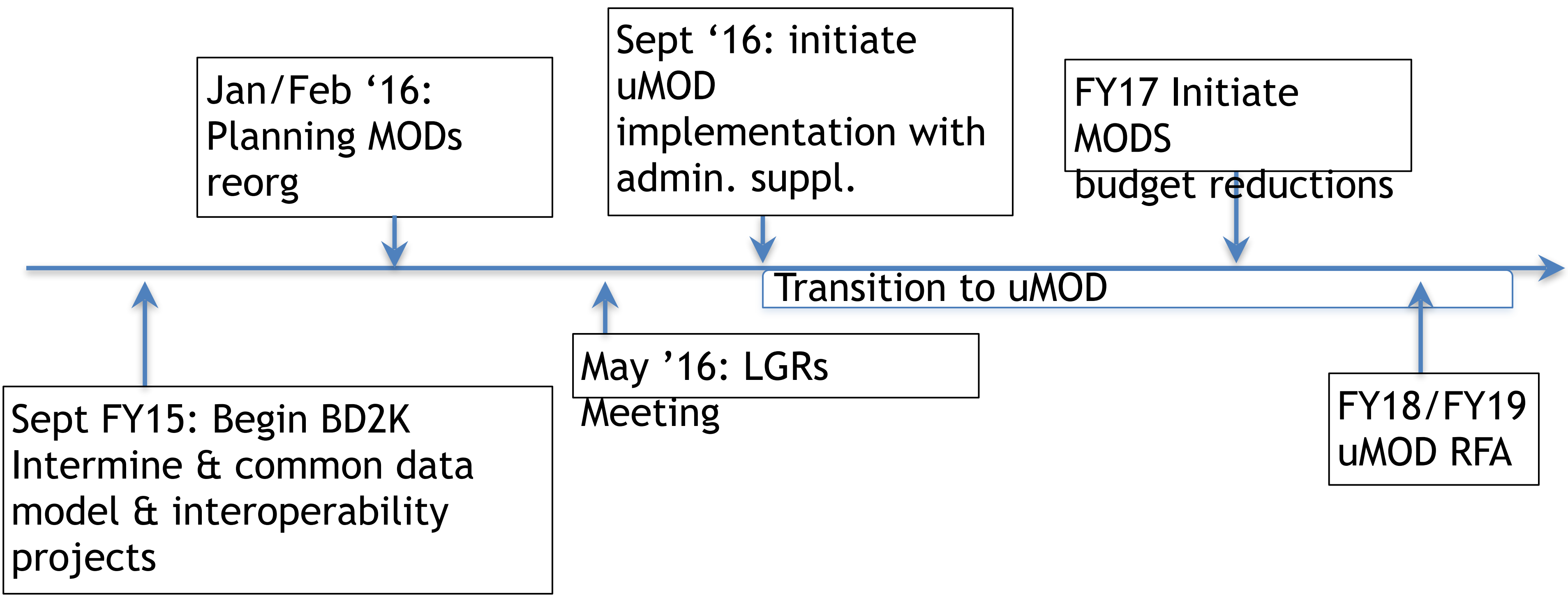
- Data management systems for related data structures and types
- System administration and IT support
- Technical user support
- Links to the same public resources which need updates and maintenance

From the NHGRI Large Genomic Resources Meeting - Valentina Di Francesco, NHGRI Computational Genomics and Data Science Program, May 23, 2016

Goals of the Reorganization

- Facilitate access to these resources
- Continue to support the value and services provided by the MODs resources
- Transition the resources to a more effective and sustainable funding model
- Gain flexibility for new informatics program activities at NHGRI

Timelines


Jan/Feb '16: Planning MODs reorg
Sept '16: initiate uMOD implementation with admin. suppl.
FY17 Initiate MODS budget reductions
Transition to uMOD
Sept FY15: Begin BD2K Intermine & common data model & interoperability projects
May '16: LGRs Meeting
FY18/FY19 uMOD RFA

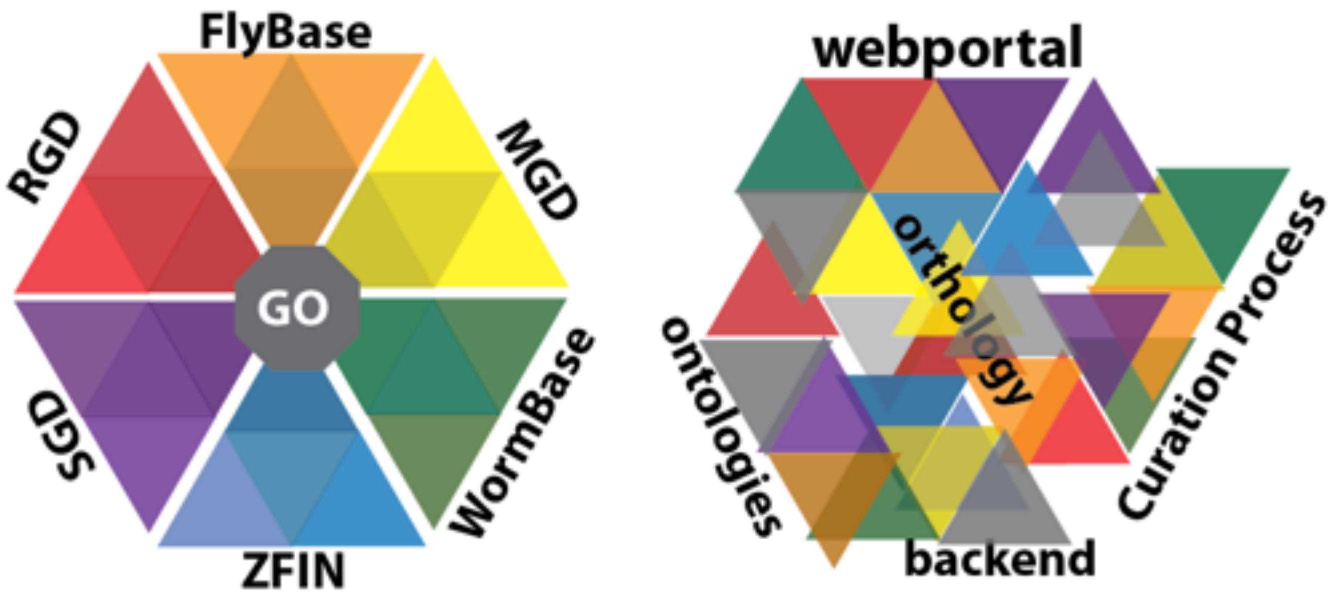
Unifying the MODs?

1. Who?
2. Why?
3. How? Long and short term goals
4. Funding implications
5. Should smaller mods work towards this sort of integration too?

1. Who are the MODs?

1. Who?
2. Why?
3. How? Long and short term goals
4. Funding implications
5. Should smaller mods to work towards this sort of integration too?

2. How



1. Identify and integrate the common tools

HOW: Priority Goals for MOD harmonization

- Understand existing components of all resources
- Understand the strong models/tools within the consortium
- Map a clear path to integration and common views to allow federated access to all MOD data

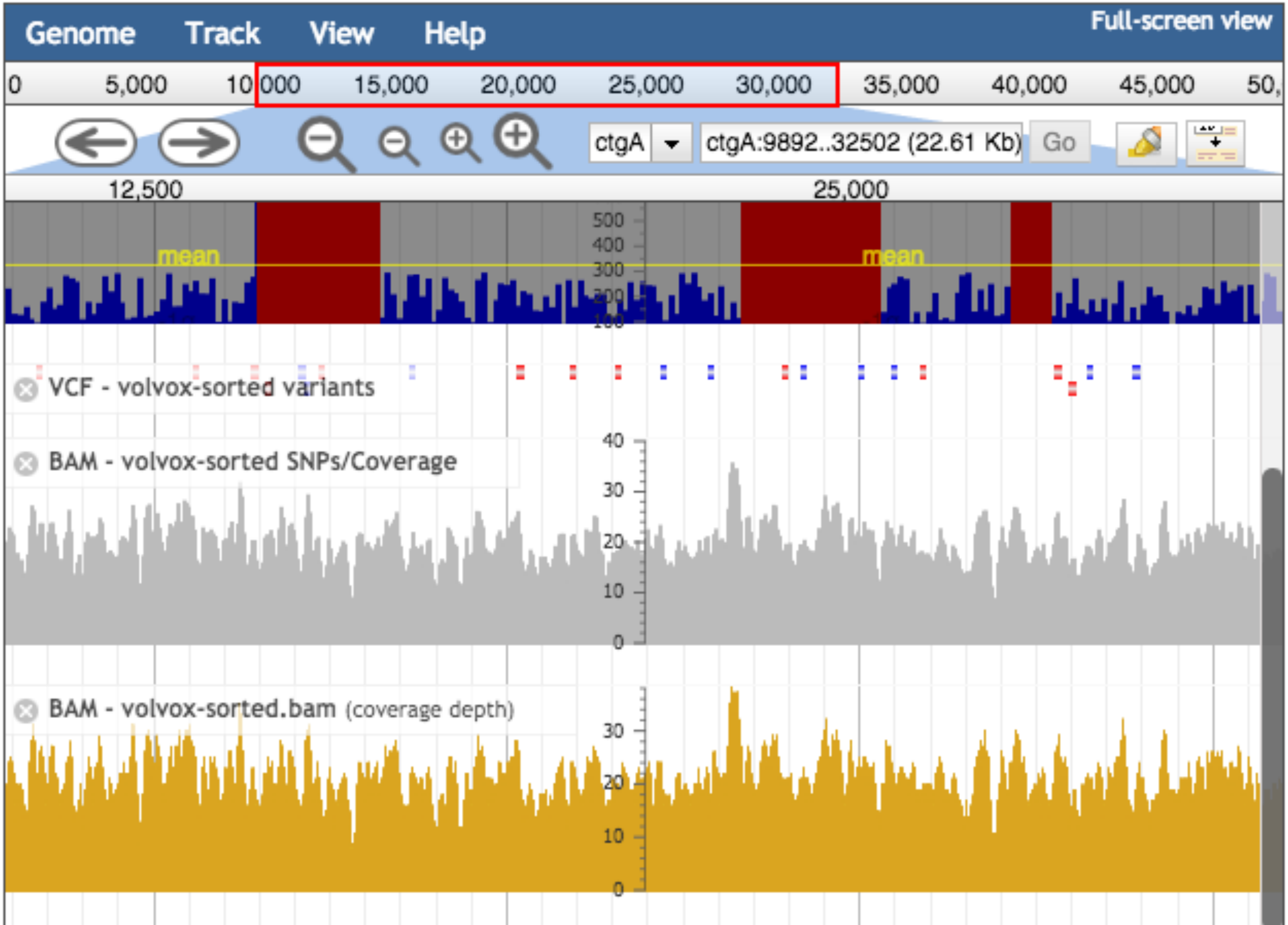
Low Hanging Fruit for harmonization

- Genome Features / Gene Models
- Orthology / Homology
- Human Disease / Phenotypes
- Biomedical Ontologies

Genomes and Genome Features: Two Initial Objectives

- **Common Genome Browser**
  - JBrowse
  - Already in wide use
- **Genome Features / Gene Detail Pages**
  - Initially protein coding genes
  - Build for all genome features

Latest Release – JBrowse 1.12.1



Initial Steps for Gene Pages

- Support query on any official gene symbol or synonym
- Summation page with possible matches and why
- Links to current gene detail pages at MOD sites

Gene Model Standardization

<u>Core Identifiers</u>

- primaryIdentifier
- secondaryIdentifier
- Symbol
- Name
- Cross references

<u>Descriptions</u>

- Limited to two fields: description and briefDescription

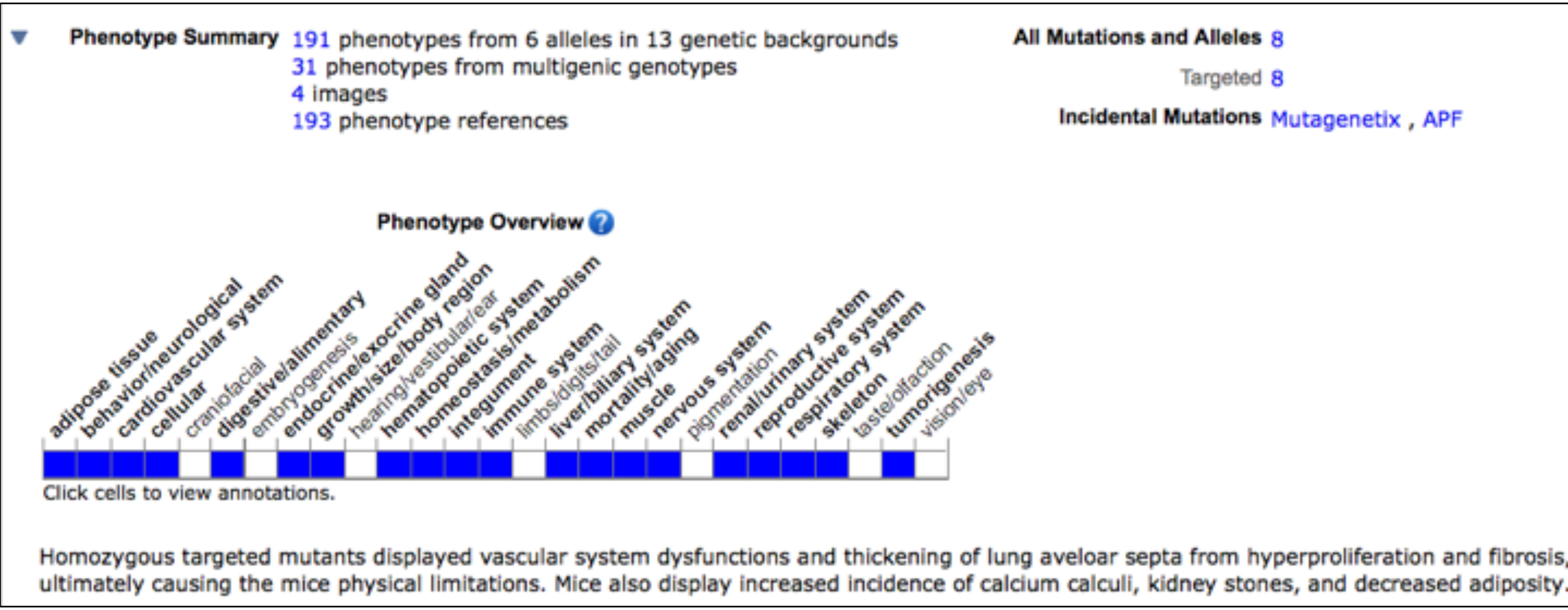
<u>Gene Types</u>

- originally represented by several field names, featureType, geneType etc
- Will be limited to Sequence Ontology terms

Gene Detail Pages

- **Protein-coding Genes**
  - Isoforms and modified protein forms
- **RNA genes**
  - Regulatory interactions
- **Two aspects to capture**
  - 'what does this gene do' summations
  - Drill down to data detail

Summations Ex: Phenotype Ribbon



- Click cells to view annotations

Orthology /Homology: Two Initial Objectives

- **Define uMOD orthology set(s) for comparative tools and representations**
- **Incorporate comparative orthology tool to interrogate and visualize orthology data.**

Define uMOD orthology set(s) for comparative tools

- **InterMOD**
  - Each MOD's orthology used with search on human gene
  - But what to do when no human gene in set?
- **Panther / GO *****
  - Already in deep use in AmiGO; extends to analysis
  - Positive: Quest 4 Orthologs provides community input into definition of reference proteomes to drive algorithm
  - Negative: Updates and completion for some MODs needed
- **Other? InParanoid?**
- **ZFIN - custom set, special purpose**

Incorporate comparative orthology tool to interrogate and visualize orthology data.

- **Model 1: HCOP**
  - Nice visualizations
  - Only from Human, to 17 model organisms
- **Model 2: DIOPT (****)**
  - Human, fly, mouse, worm, yeast, zebrafish
  - Confusing stats
  - 'in house' so further improvements quick
- **Comparative data graphs**
  - http://www.informatics.jax.org/homology/GOGraph/7247

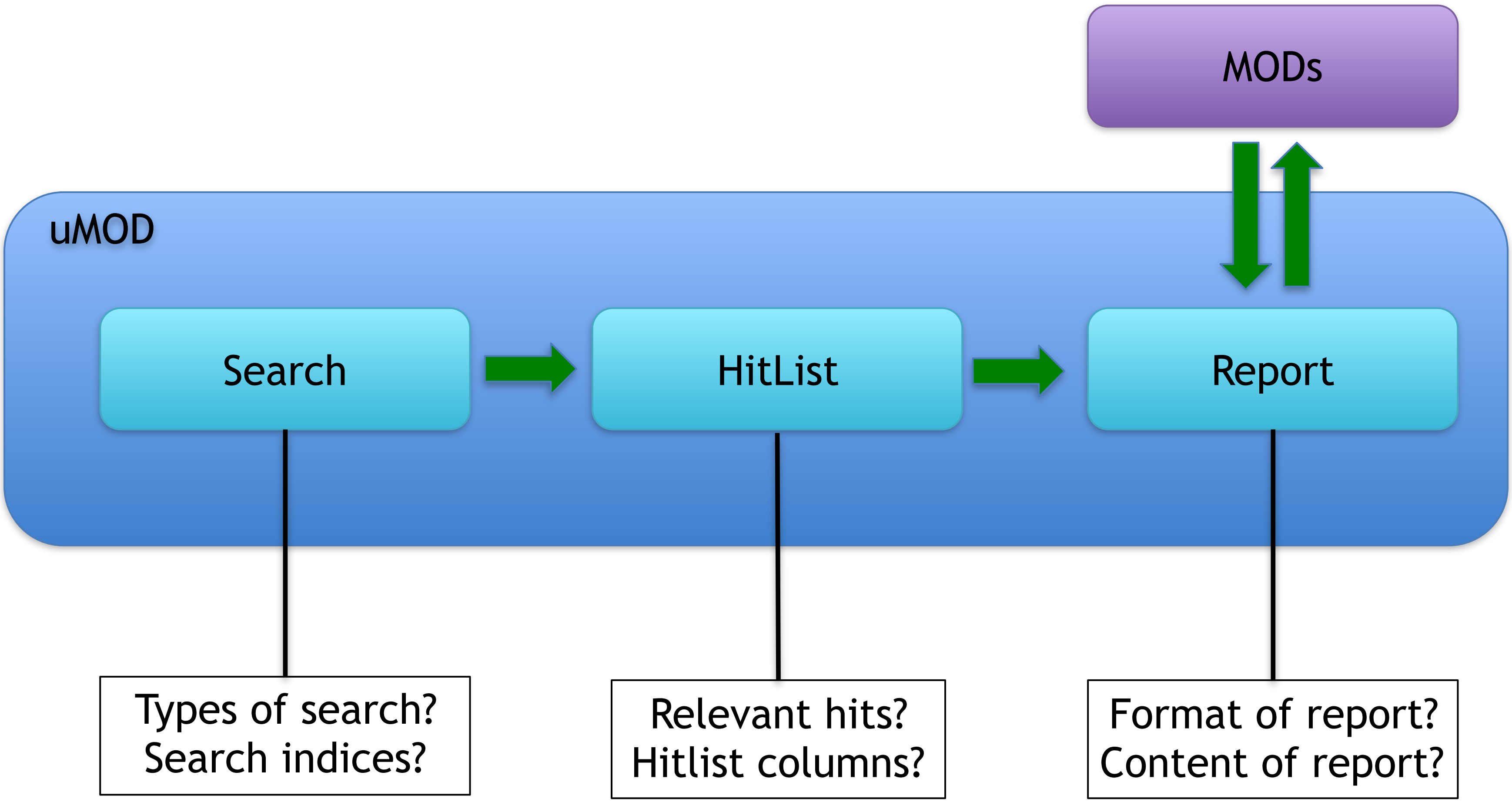
Phenotype and Disease

Cindy Smith and Carol Bult (MGD)
Suzanna Lewis (GO)
Mary Shimoyama(RGD)
Monte Westerfield (ZFIN)
Thom Kaufman(FlyBase)
H-MOD Disease Working Group

Two Initial Objectives

- **Support query by human disease term**
  - OMIM
  - Disease Ontology (DO)
- **Support query by phenotype term**
  - Continued work on alignment of vocabularies

Objective


MODs
uMOD
Search
HitList
Report
Types of search?
Search indices?
Relevant hits?
Hitlist columns?
Format of report?
Content of report?

Use cases

1. Search with a human gene
a) Which **diseases** are associated with this human gene?
b) What are the **orthologs** for this human gene in MOs, and are there **experimental disease models**?
c) Which **diseases** are associated with the region containing this human gene (GWAS)?
2. Search with a Model Organism gene
a) Which **diseases** are associated with human ortholog(s) of this gene?
b) Are there existing **experimental disease models** in this MO using this gene, or a human transgene?
c) What are the **orthologs** for this gene in other MOs, and are there **experimental disease models**?
d) Which **diseases** are associated with the region containing the human ortholog(s) of this gene (GWAS)?
3. Search with a human disease
a) Which **human gene(s)** are associated with this disease?
b) Are there **experimental disease models** in MOs for this disease?
c) Which human **genomic region(s)** (GWAS) are associated with this disease, what human genes are contained therein? (What are there MO orthologs, and are there experimental disease models?)
4. Search with a human genomic region?
a) Which **diseases** are associated with this region, either direct gene-disease associations or via GWAS?
b) Are there **experimental disease models** in MOs for these diseases?
5. Search with a 'phenotype'?

H-MOD Working Group

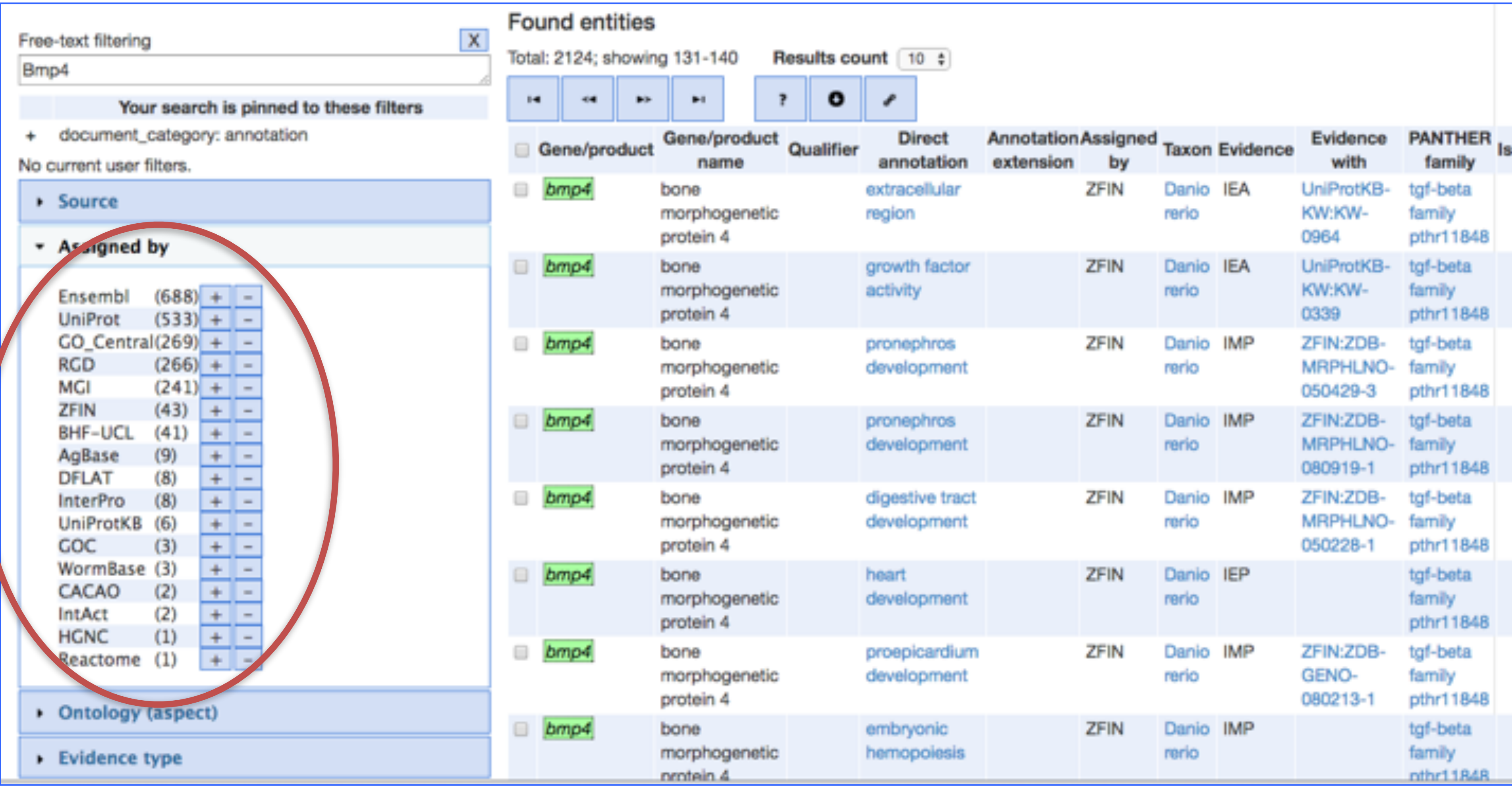
Disease Model annotation table?

Disease		Human gene		Model Organism gene			
Name (DO)	**Subtype (OMIM pheno)**	**Symbol**	**Genomic location**	**Species**	**Symbol**	**Ortholog score**	**Exp. models ?**
Alzheimer's disease	ALZHEIMER DISEASE; AD	APP	21q21.3	Mouse	App	10	108
				Rat	App	10	343
				Frog	app	6	-
				Fish	appb	6	-
				Fly	Appl	7	4
				Fly	Hsap\APP	n/a	36
				Worm	apl-1	7	?
		APBB2	4p14-p13	Mouse	Apbb2	8	-
				Frog	apbb2	5	-
				Fish	apbb2b	3	-
				Worm	feh-1	6	-
	ALZHEIMER DISEASE 2	APOE	9q13.32	Mouse	Apoe	8	-
				etc…	*etc…*	*etc…*	*etc…*

Biomedical Ontologies

GOC group

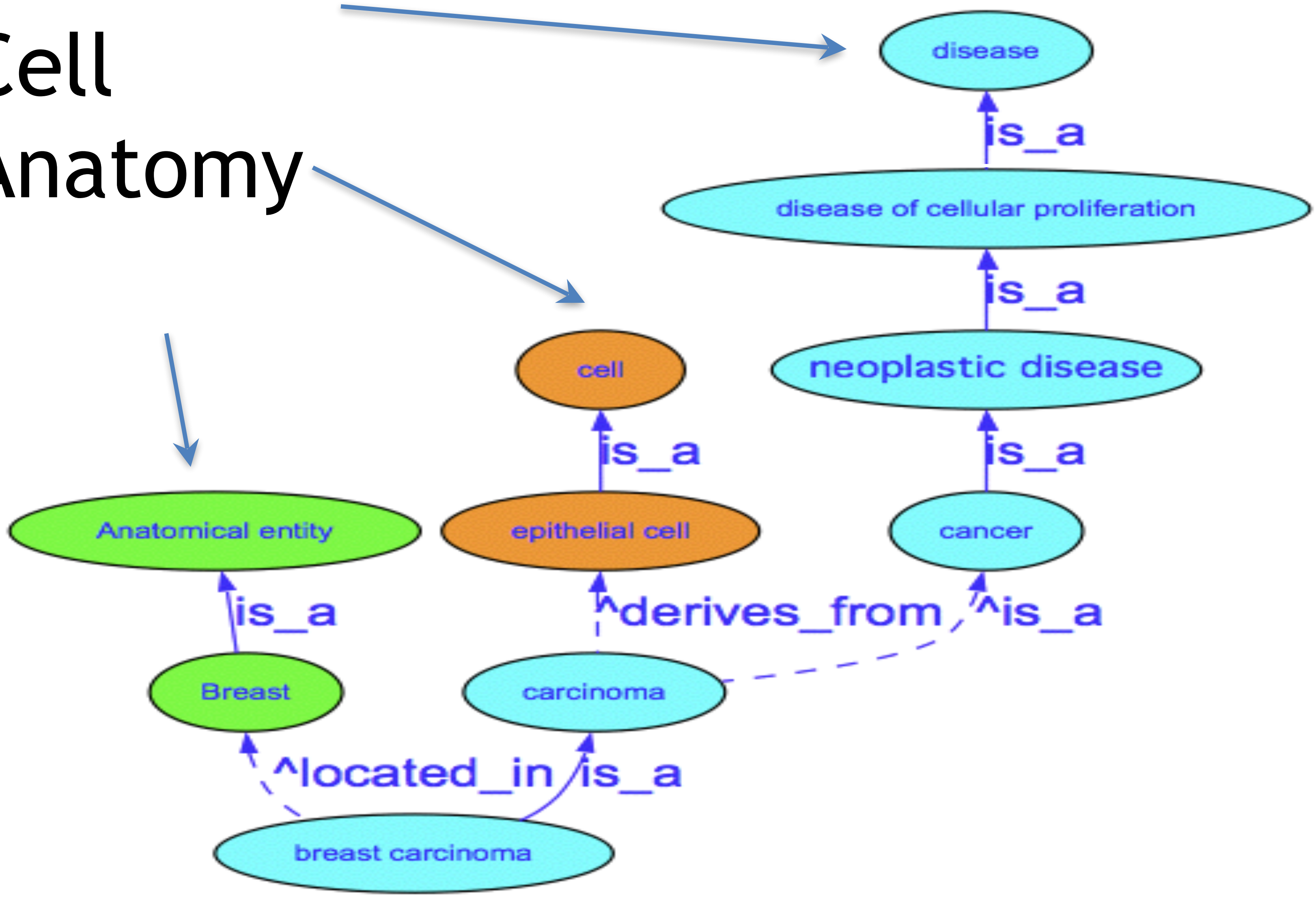
GENEONTOLOGY
Unifying Biology



Biomedical Ontologies

- GO: (function, process, cellular location)
- SO: (sequence features)
- PRO: (specific proteins by species/strain)
- MP, HPO, others: (phenotypes)
- Anatomies / Homologies (morphology)
- DO: (diseases, not phenotypes; definitions not diagnoses)
- CL: (cells and their lineages)

Disease
Cell
Anatomy



Adapted from Schriml and Kibbe: ICBO submission 2013

Summary

- Start with compelling use cases that join models and build collaborative environment
- Small work groups and teams define requirements and bring to steering committee for refinement and agreement
- Groups exist now, focus and interactions strong

Acknowledgements

- Karen Yook and the other curators at WormBase for helping me sort some of this out.
- Lincoln Stein, Paul Sternberg, Paul Kersey and Matt Berriman , co-PIs of WormBase.

A Few Questions

1) Specifically, for each of the 5 MODs and GO, what activities and data would be supported in common, and which would be unique to each site?

2) What is the likely platform(s) for the common components? For the individual sites?

3) Which common elements/activities might be phased in by what order?

4) How will issues of scalability and variety of data types be addressed?

5) Would there be incremental deliverables along the way. What might these specific milestones be?

6) Are there any components or steps for which there is already agreement among the groups?

7) What are the components or steps which are not yet resolved?

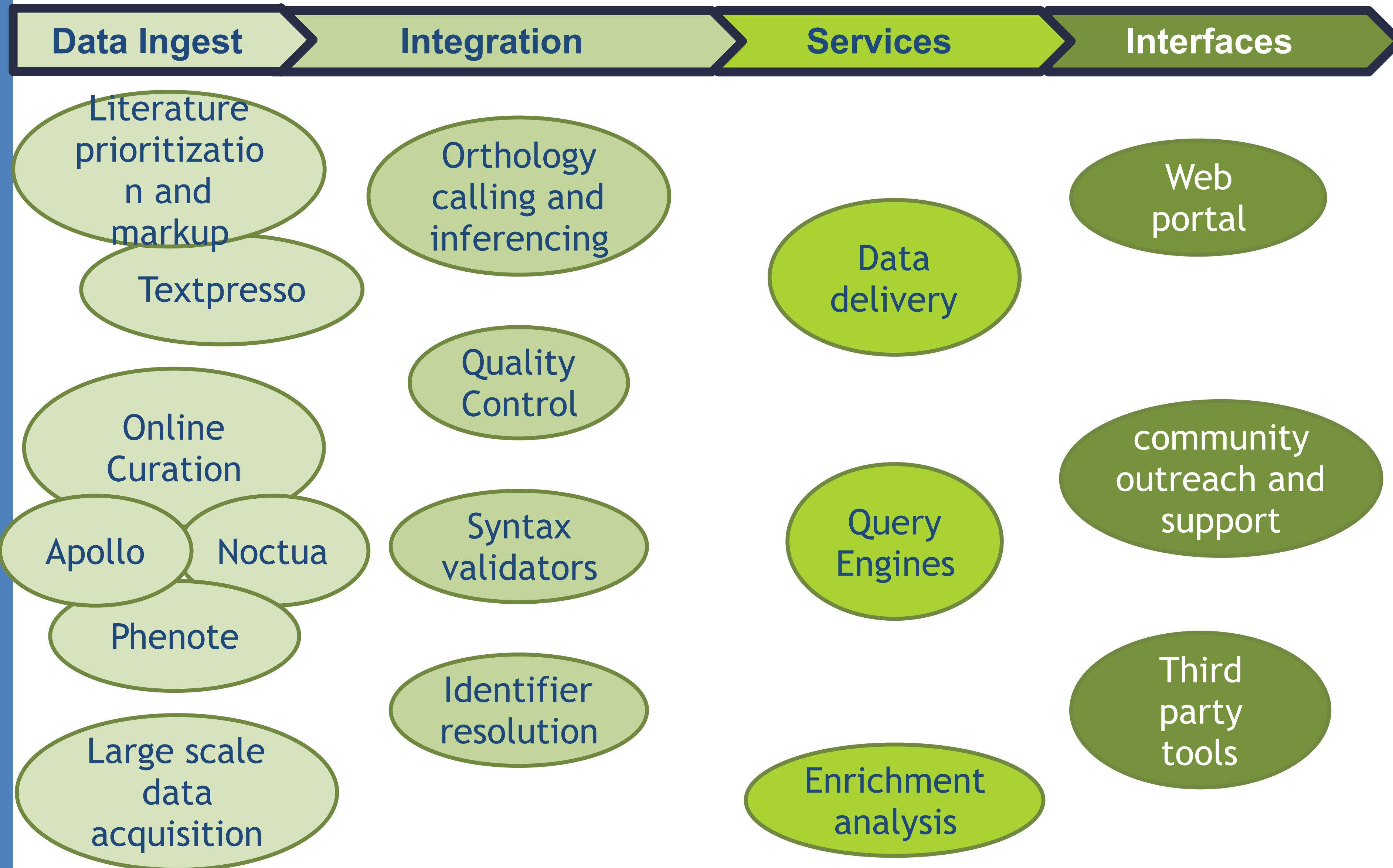
8) Are there any components or elements where there are clear conflicts, or differences of opinion among the sites now?

9) In the proposed new model, where exactly will the savings be realized? Staff reductions? Which staff?

10) Will any current activities by any of the participants stop? Which ones?

11) Will any new activities not currently done by any participant be started?

uMOD Functions



SHARED SEMANTIC FRAMEWORK

SHARED FRAMEWORK

UBERON + Cell Type

- ✓ Human
- ✓ Worm anatomy
- ✓ Mouse anatomy
- ✓ Fly anatomy
- ✓ Etc.

Assays

NCBI taxonomy

Evidence

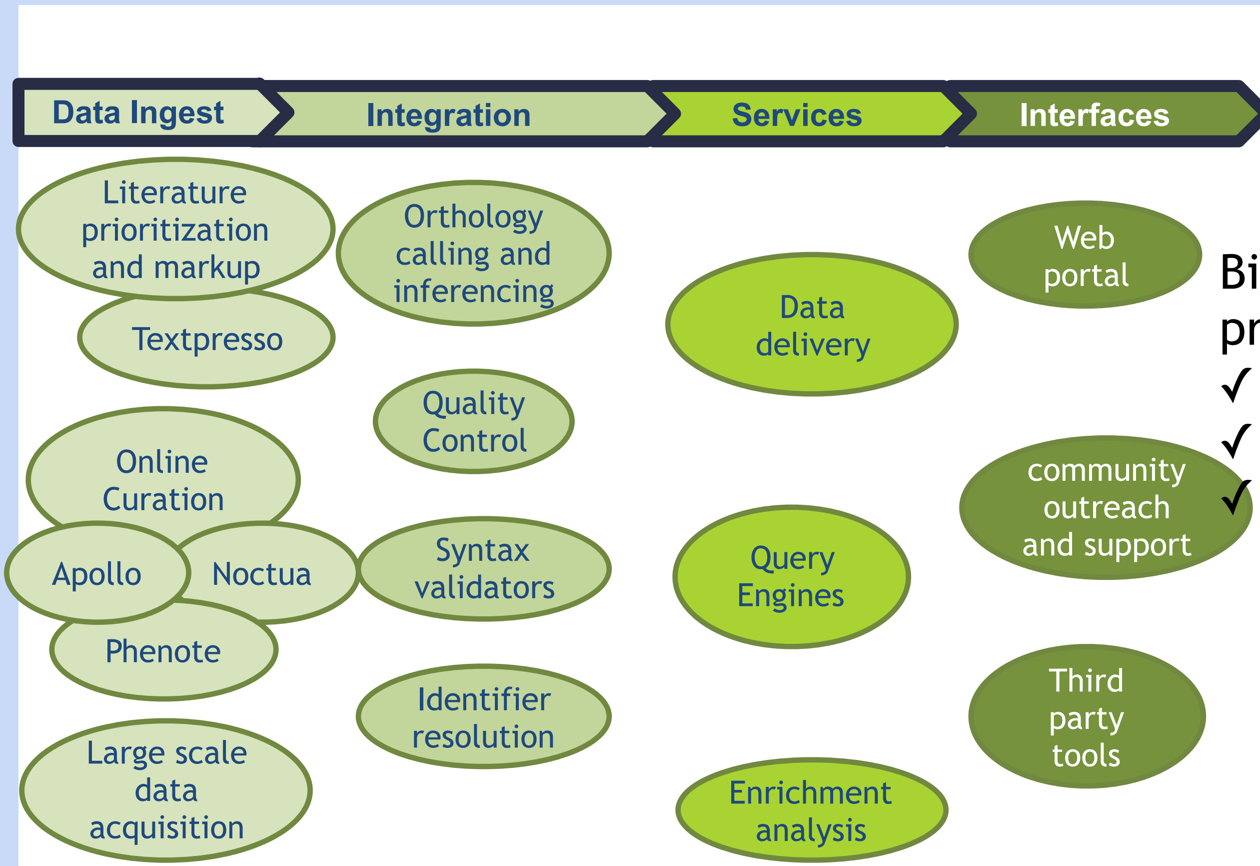
Environment

Phenotype

- ✓ HPO
- ✓ WP
- ✓ MP
- ✓ Yeast
- ✓ Etc.

Cellular component

Molecular function



Biological process

- ✓ Physiological
- ✓ Behavior
- ✓ Population

Sequence features

Molecular process pathway

Recommendation: **Standard APIs**

- **Shared Web APIs for MODs and related resources**
  - Standard **URL structure** to programmatically fetch database objects
  - Standard **JSON format** for returning data
- **Wins**
  - Loose coupling, lightweight, no/few backend changes required
  - Shared UI components
    - Steps toward more consistent front ends
  - (Power) User efficiencies
  - Steps towards tighter integration

Standardized exchange formats

- **Pre-requisite for Web API** (previous slide)
  - Can be implemented independently
  - Replace or extend current ad-hoc database dumps available from most MODs FTP sites
  - Enables aggregators and bioinformatics power users
    - Aggregators: Intermine, Monarch, ClinGen
- **Recommendations**:
  - JSON or JSON-LD
    - Not XML, TSV
  - Leverage semantics of ontologies rather than reinvent
- **Track with existing efforts**
  - GA4GH: G2P, metadata, …
  - FHIR, Bio2RDF
  - PhenoPackets
  - Domain-specific efforts
    - Pathways, interactions, pathways, …

Best behavior for Identifiers

1. Use established identifiers
2. Design unique identifiers for use by other groups
3. Help local identifiers travel well: document Prefix and Namespace
4. Opt for simple durable web resolution
5. Avoid embedding meaning
6. Make URIs clear and findable
7. Implement a version management policy
8. Do not re-assign or delete identifiers
9. Document the identifiers you issue and use
10. Reference responsibly

Shared UI components

- Current shared components
  - Industry-standard 3rd party frameworks (e.g. jquery)
  - Occasional use of shared bioinformatics components
    - E.g. JBrowse genome browser
  - Sometimes: intermine, biomart

Standardized Curation (Tools)

- Begin with curator developed curation standards
- Move to shared curation tools, for example:
  - Apollo for genome features
  - Noctua GO curation tool
    - Being extended to model phenotypes
  - TextPressoCentral (integrated with Noctua now!)
- Wins
  - Curator efficiencies
  - Developer efficiencies (less duplication of effort)
  - Community contribution

A new way of developing software: Commons based Peer Production

Distributed Version Control

GitHub

Social Coding

Graph-based audit trail of every contribution

Issue tracking

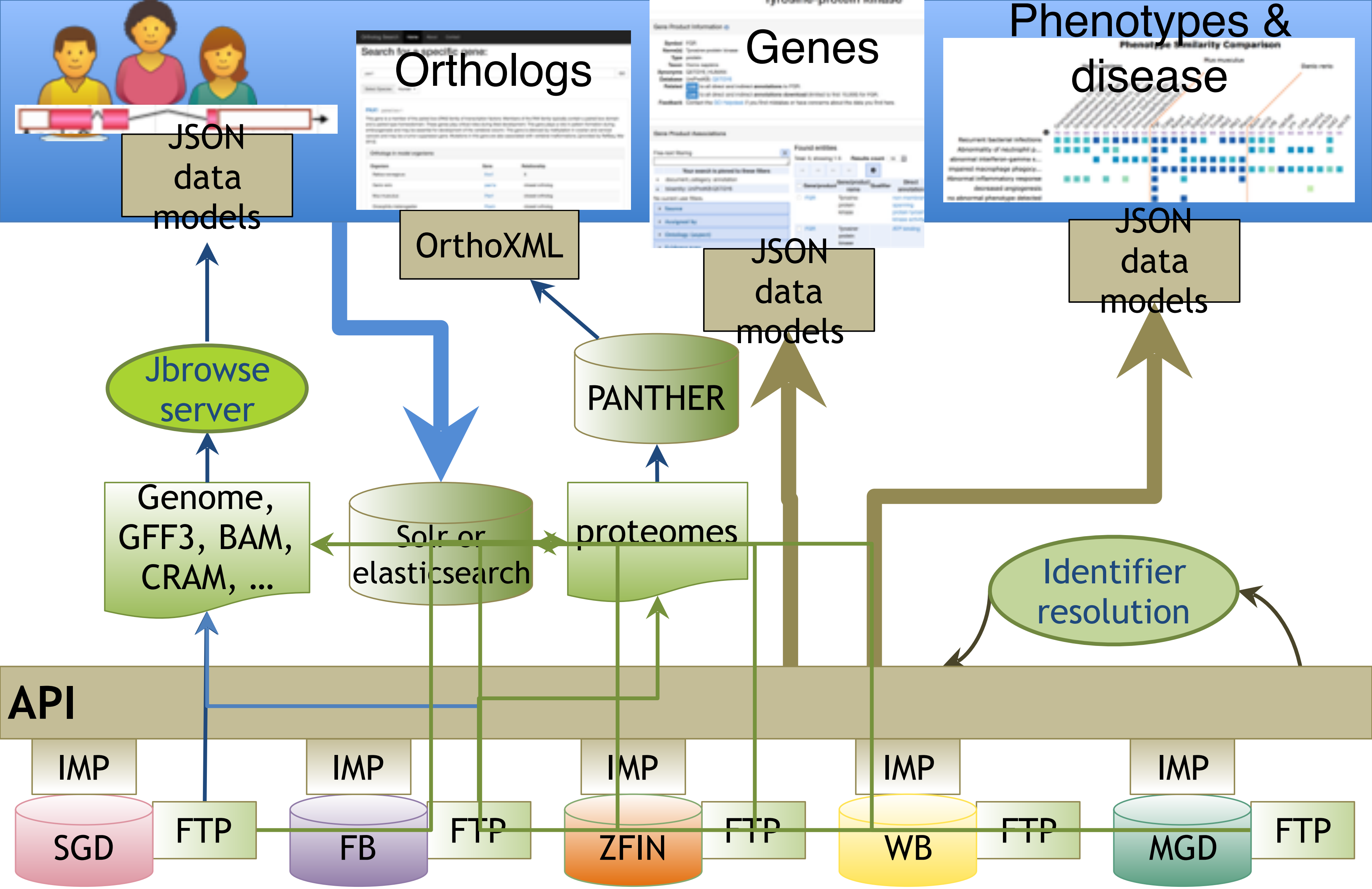
Radical transparency

Commons based Peer Development for community standards

- Does the GitHub model make sense for community standards development?
  - Yes!
  - Watch this space: Global Alliance for Genomic Health (GA4GH)
    - forks, branches, +1s, pull requests, attribution, immediacy, dynamism, experiments and crazy ideas

Shared Hygiene

- Use of version control systems
  - Open hosted repos (GitHub, GitLab)
  - Trackers for features and bugs
- Unit tests and Continuous Integration
- Common language-level APIs



Shared API and formats

Potential end goal: Shared full stack

- A unified database/schema and architecture
  - Long term developer efficiencies
  - Reuse of core backend tools, e.g. curation tools
  - Structural analysis of the relationships between different entities organized in huge networks of graph-like structures enables data science
- Recommendations:
  - Incremental steps: Start with common exchange format (bootstrap), more shared tooling and APIs
  - Exploratory task force to evaluate Cost/benefit tradeoffs

MODs have similar user

human geneticists who want access to all model organism data which are the main source of experimental annotation of human genes

basic science researchers who use specific model organisms to understand fundamental biology

computational biologists and data scientists who need access to standardized, well-structured data, both big and small

educators and students who want to teach and learn

uMOD Consortium Mission

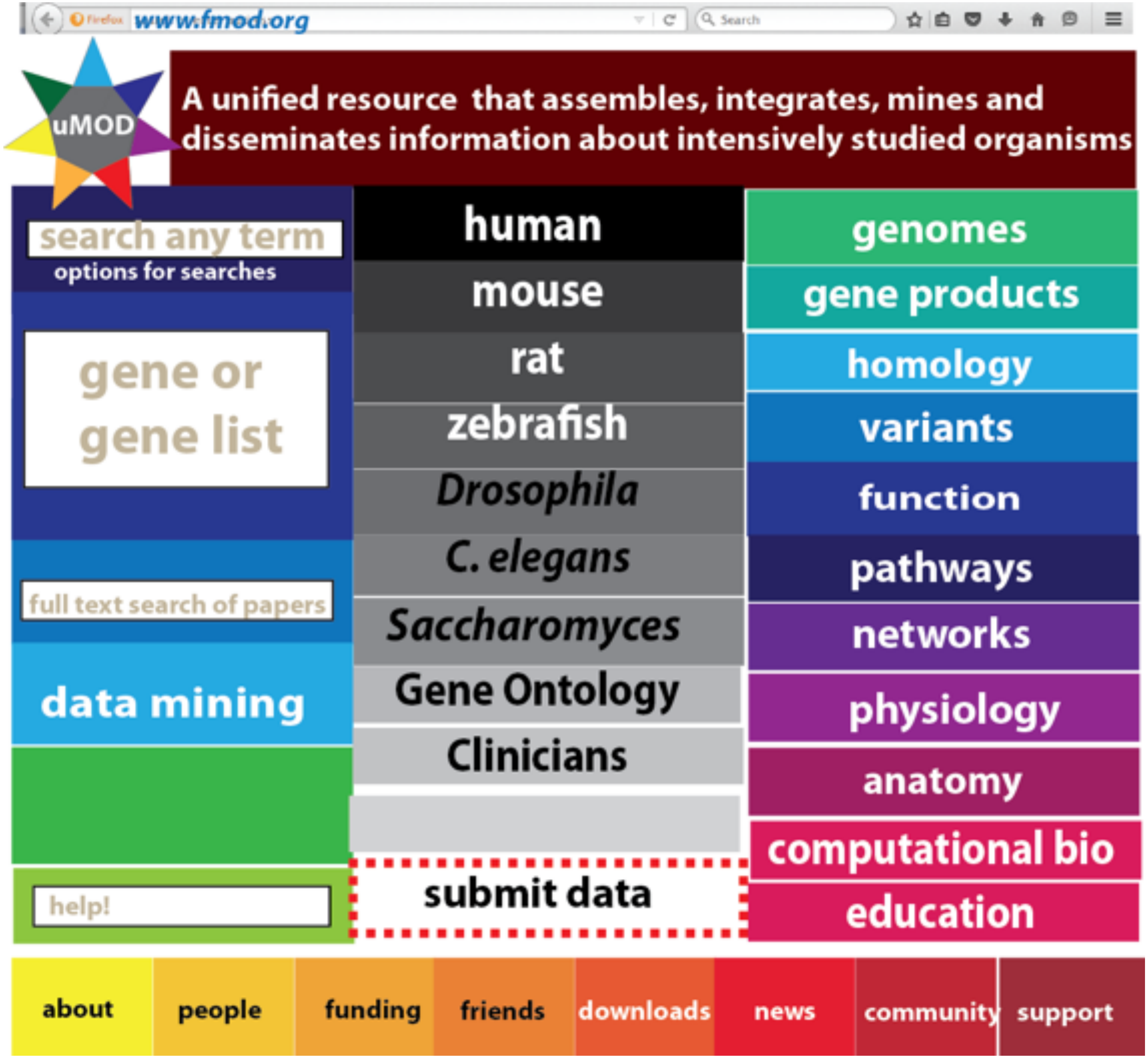
Develop and maintain a unified information resource that facilitates the use of diverse model organisms in understanding the genetic and genomic basis of human biology and health. This understanding is fundamental for advancing genome biology and for translating genome data into clinical utility

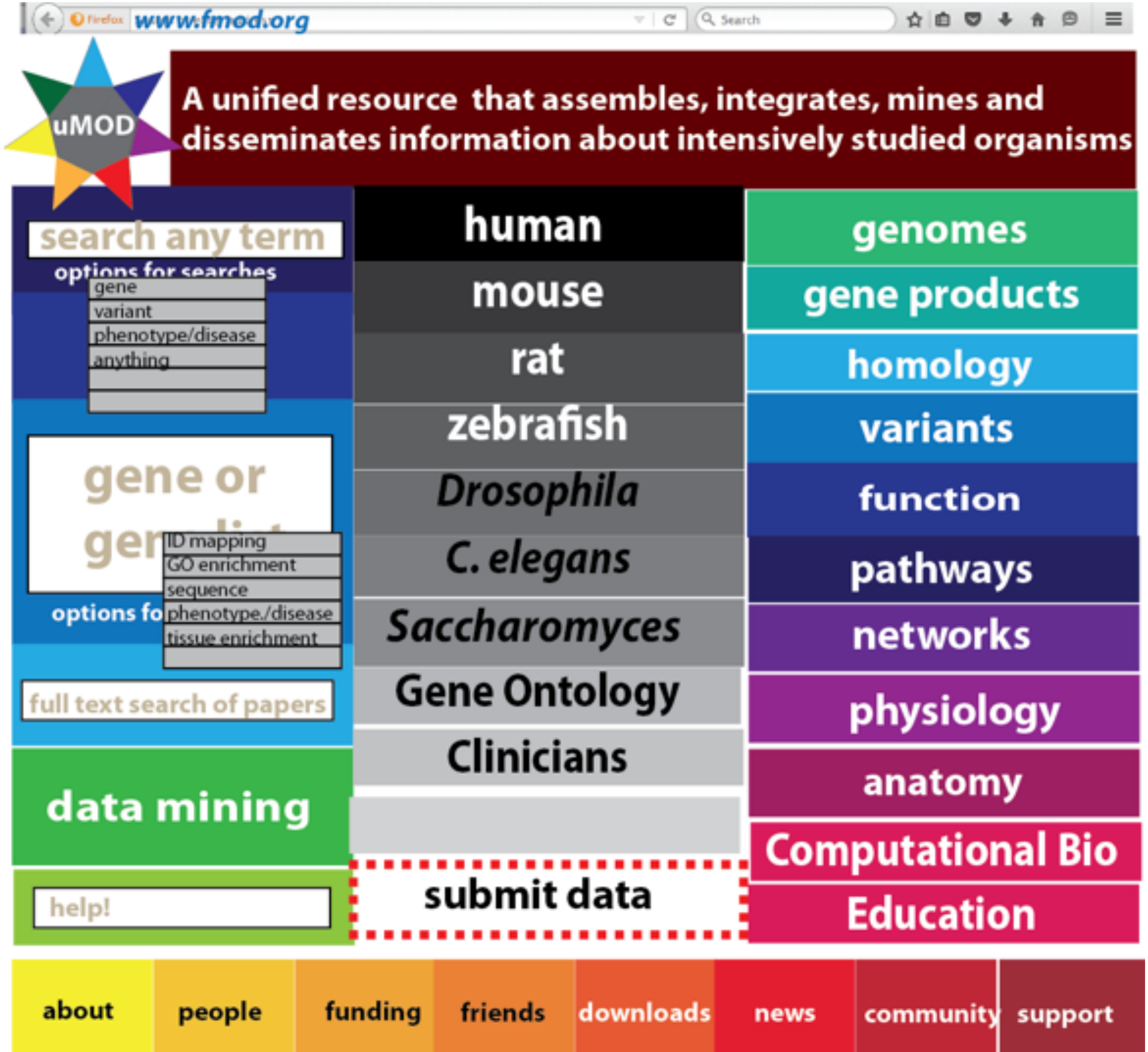
The MATRIX

Matrix of shared infrastructure modules comprising
outwardly facing components such as common web portal, APIs, hosted online curation and data analysis tools, web services, and community outreach;
internal data management applications in concert with biological expertise to enable comprehensive integration across biological domains including, orthology analysis, a shared ontology framework, and data consistency verification.

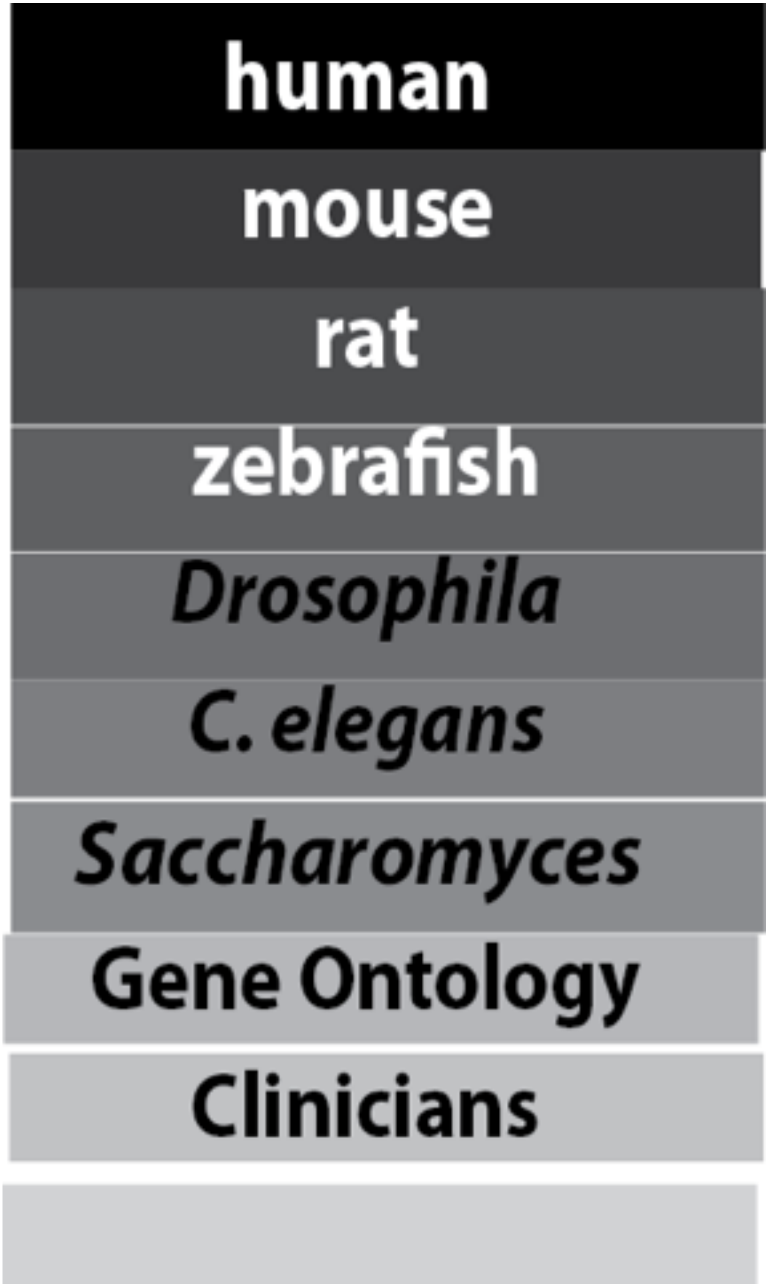
Expert teams that select or develop modules from sources within and outside the consortium.

Organism-specific working groups that use shared


www.fmod.org
Search
uMOD
A unified resource that assembles, integrates, mines and disseminates information about intensively studied organisms
search any term
options for searches
gene or gene list
full text search of papers
data mining
help!
human
mouse
rat
zebrafish
Drosophila
C. elegans
Saccharomyces
Gene Ontology
Clinicians
submit data
genomes
gene products
homology
variants
function
pathways
networks
physiology
anatomy
computational bio
education
about
people
funding
friends
downloads
news
community
support

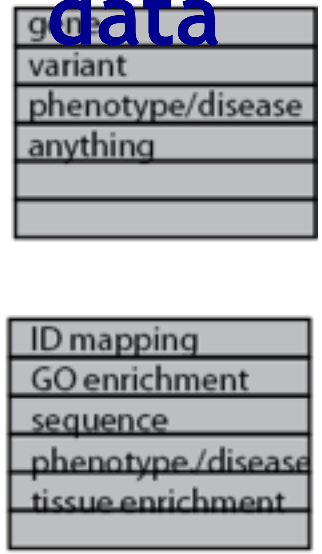


www.fmod.org
Search
uMOD
A unified resource that assembles, integrates, mines and disseminates information about intensively studied organisms
search any term
options for searches
gene
variant
phenotype/disease
anything
gene or gene list
ID mapping
GO enrichment
sequence
phenotype./disease
tissue enrichment
options fo
full text search of papers
data mining
help!
human
mouse
rat
zebrafish
Drosophila
C. elegans
Saccharomyces
Gene Ontology
Clinicians
submit data
genomes
gene products
homology
variants
function
pathways
networks
physiology
anatomy
Computational Bio
Education
about
people
funding
friends
downloads
news
community
support



Links to community-specific portals that replace and enhance existing MODs or new views

Perform simple searches on data



Perform queries and analyses of gene lists

Textpresso searches of PMC

interMOD interMine

InterMine platform from InterMOD

Will enable federated queries between MOD data and the network of linked data

- InterMOD standardizes access to MOD data with InterMine software.
- InterMine creates data warehouses that integrate data and enable bioinformatics analyses.
- 28 InterMines are available for different community databases:
  - **YeastMine** (SGD), **WormMine** (WB), **FlyBaseMine** (FB), **ZebrafishMine** (ZFIN), **MouseMine** (MGI), ThaleMine (AraPort), HumanMine (Micklem), **RatMine** (RGD), **XenMine** (Cherry), BovineMine (USDA), SoyMine (USDA), modMine (modENCODE), GrapeMine (EU), LegumeMine (NSF), PlanMine (planarian, EU), …
- Core data object model (commonly used) + specific MOD customizations for flexibility

- genomes
- gene products
- homology
- variants
- function
- pathways
- networks
- physiology
- anatomy
- Computational Bio
- Education

Links to topic-specific portals that provide global entry into the combined data.

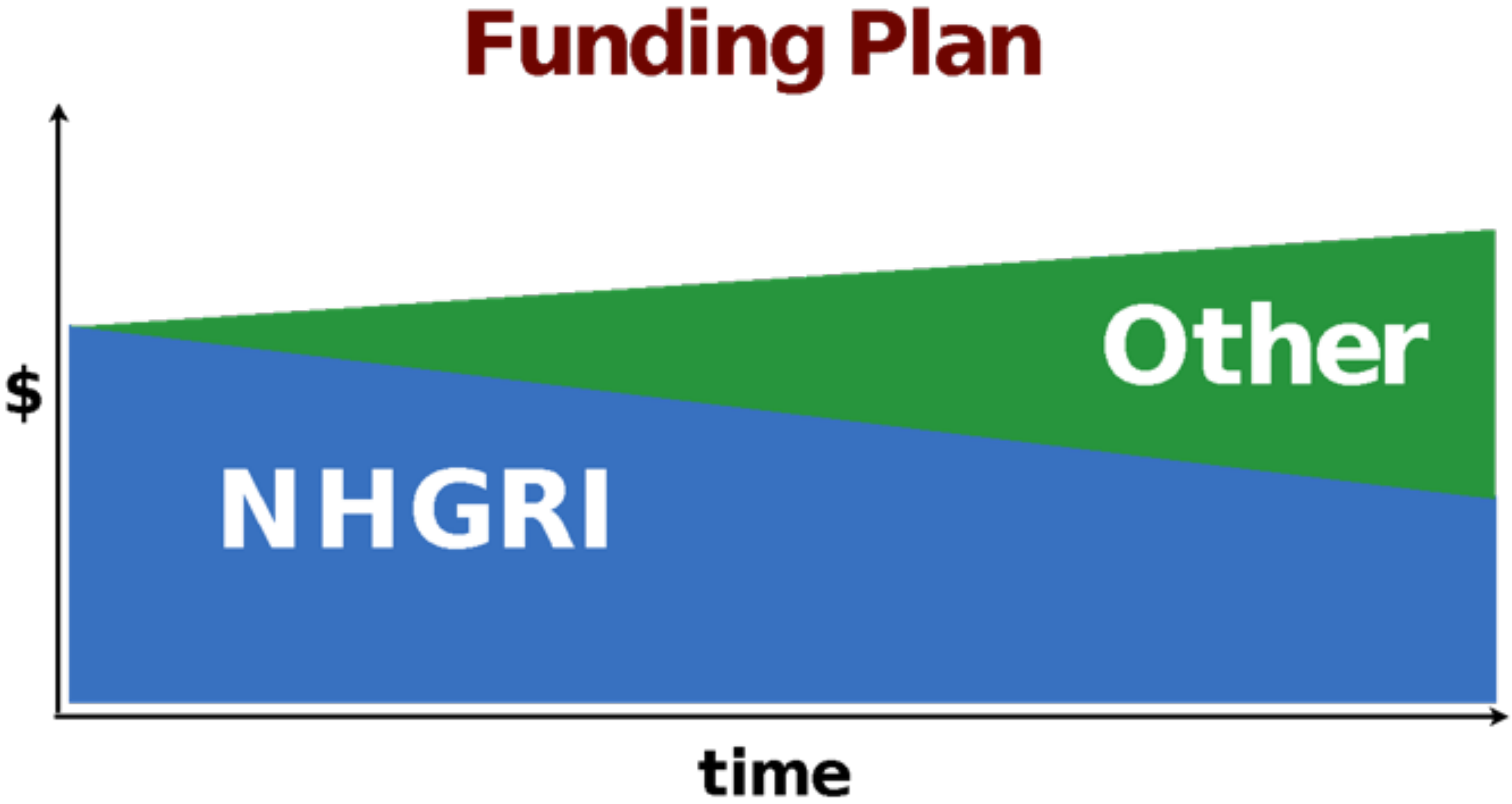
New pathway viewer + reactome + etc.

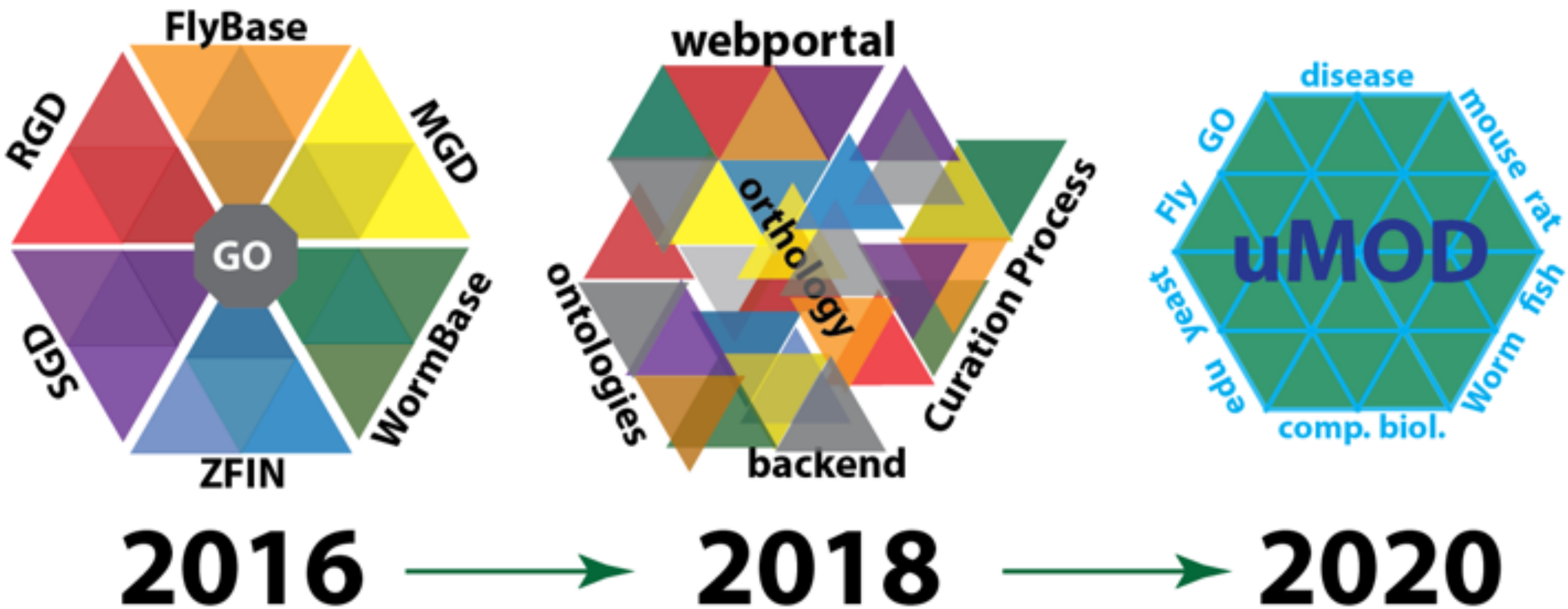
Anatomy ontology browsers and links to atlas sites

Link to community annotation systems

Link to Standard website pages, ftp sites, help desk, etc.

The Finances





Sept 2017 Web Portal sufficient for use and as target of further development

Sept 2018 Backend sufficient for use and as target of further development