Alliance of Genome Resources

Genomic Resources and Sustainability Webinar

Presentation to NHGRI and External Advisors

October 25, 2019

Outline

- The Alliance
  - Goals and organization
  - Accomplishments and near term goals
- Alliance Knowledge Centers
  - Mission, approaches, impact
  - Organization: centralize or federate?
- *Gene Ontology Consortium*

Alliance of Genome Resources (the Alliance):

An experiment in “new ways of doing business” for Model Organism Databases (MODs) and the Gene Ontology Consortium (GOC)

Founding members of the Alliance

Alliance goals

- Common mechanisms for data access from MODs and GOC
  - Enhanced support for comparative genome biology

- Sustainable genome resource development
  - Shared modular infrastructure to reduce costs of resource development and maintenance

The Alliance is developing a "knowledge commons" approach to promote resource sustainability

The Alliance of Genome Resources has two components: Alliance Central and Alliance Knowledge Centers (MODs)

Alliance Central: Data and infrastructure

Data management

Programmatic and web data access

Shared user interface development

Platform for tool development

Alliance Knowledge Centers: Knowledgebases

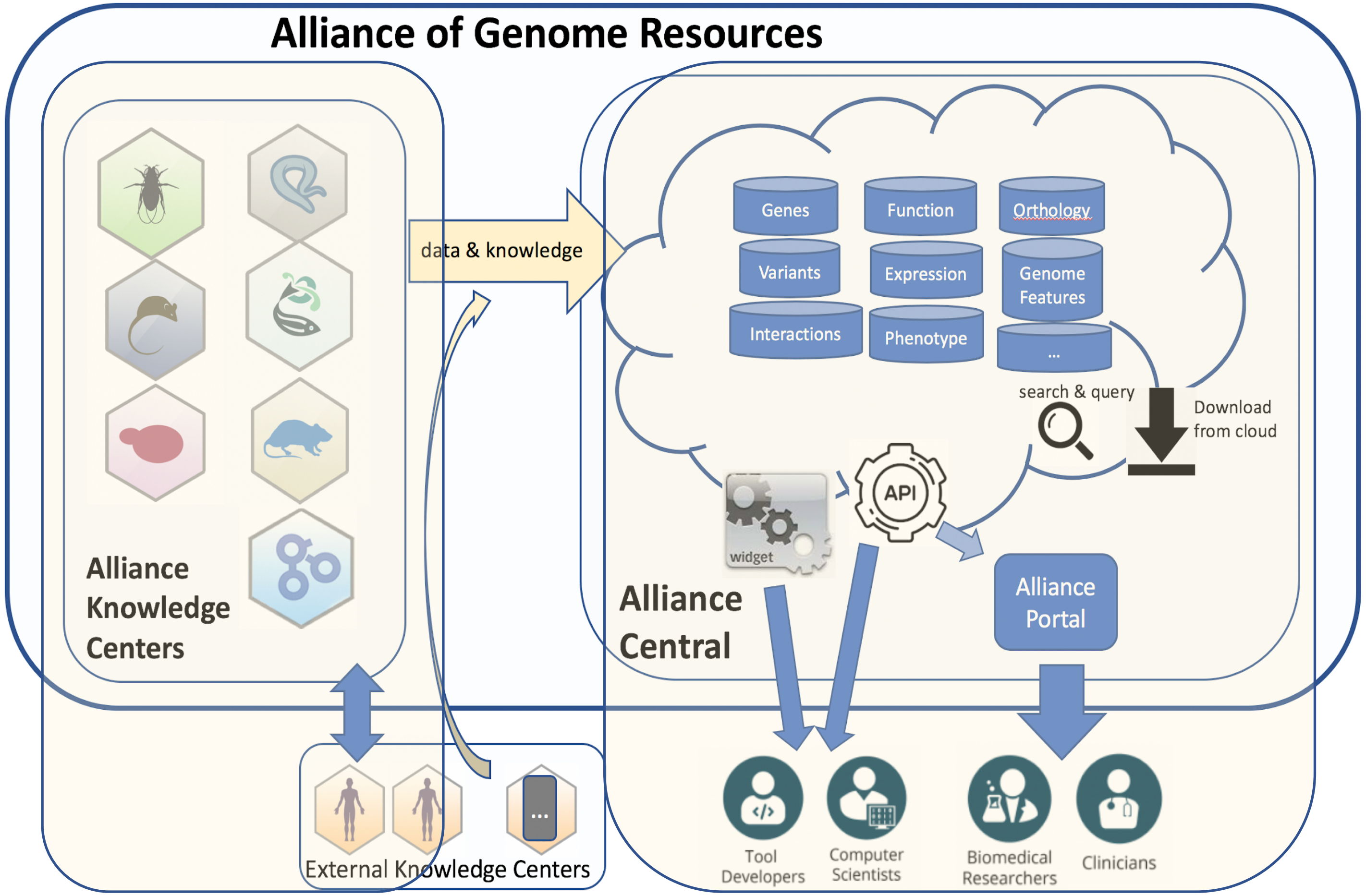
Data acquisition and expert curation

Nomenclature and knowledge representation standards

Data and concept harmonization

Organism- specific resources and reagents

Organism-specific research community engagement



Aims for Alliance Central

Alliance Central U24HG010859

- **Aim 1. Data In:** We will develop a shared platform for ingesting, storing, and harmonizing data from model organism databases and knowledgebases.
- **Aim 2. Data Out:** We will implement common methods for access to model organism data and annotations.
- **Aim 3. Website and Applications:** We will implement a framework to support the development and deployment of a multi-faceted web presence comprising software applications, workflows, and analysis tools that use model organism data and annotations.
- **Aim 4. Outreach and Management:** We will provide effective leadership and project management for the Alliance and a centralized user support helpdesk for community engagement.

The organization and operations of the Alliance align with principles in the NIH Data Science Strategic Plan

- **Modernizing the Data Ecosystem**
  - **Separate data-centric and knowledge-centric activities**
  - **Modular infrastructure**
    - Efficiency (reduction in duplication of effort)
    - Knowledge commons platform
    - Cloud based
  - **Adherence to FAIR principles**
    - Data standards
    - Data integration

High quality, “computation-ready” data

- Findable
  - Uniquely and persistently identifiable
- Accessible
  - Retrievable by machine or human
- Interoperable
  - Open, well-defined vocabulary
- Reusable
  - Machine process-able

Wilkinson et al. 2016, Sci Data 3:160018. doi: 10.1038/sdata.2016.18.

Examples of Alliance accomplishments

Starred items will be highlighted in the following slides.

Governance

- ★ *Centralized management/oversight*
- Working Groups focused on specific deliverables
- Scientific Advisory Board

User-centered Deliverables

- Search across all model organisms for harmonized data types: Gene, Allele, Function (GO), Disease
- Search and display for common data types with a shared “look and feel” across organisms
- ★ *Common ortholog gene set*
- ★ *Algorithm for generating short gene descriptions*
- ★ *Interactive ribbon graphic to summarize annotations for expression and disease*
- Common sequence visualization widget with links to common JBrowse instance
- Common molecular interaction data source for all Alliance organisms
- ★ *Common APIs for Phenotype, Disease, Orthology, Genes, Alleles, Expression*

Operations

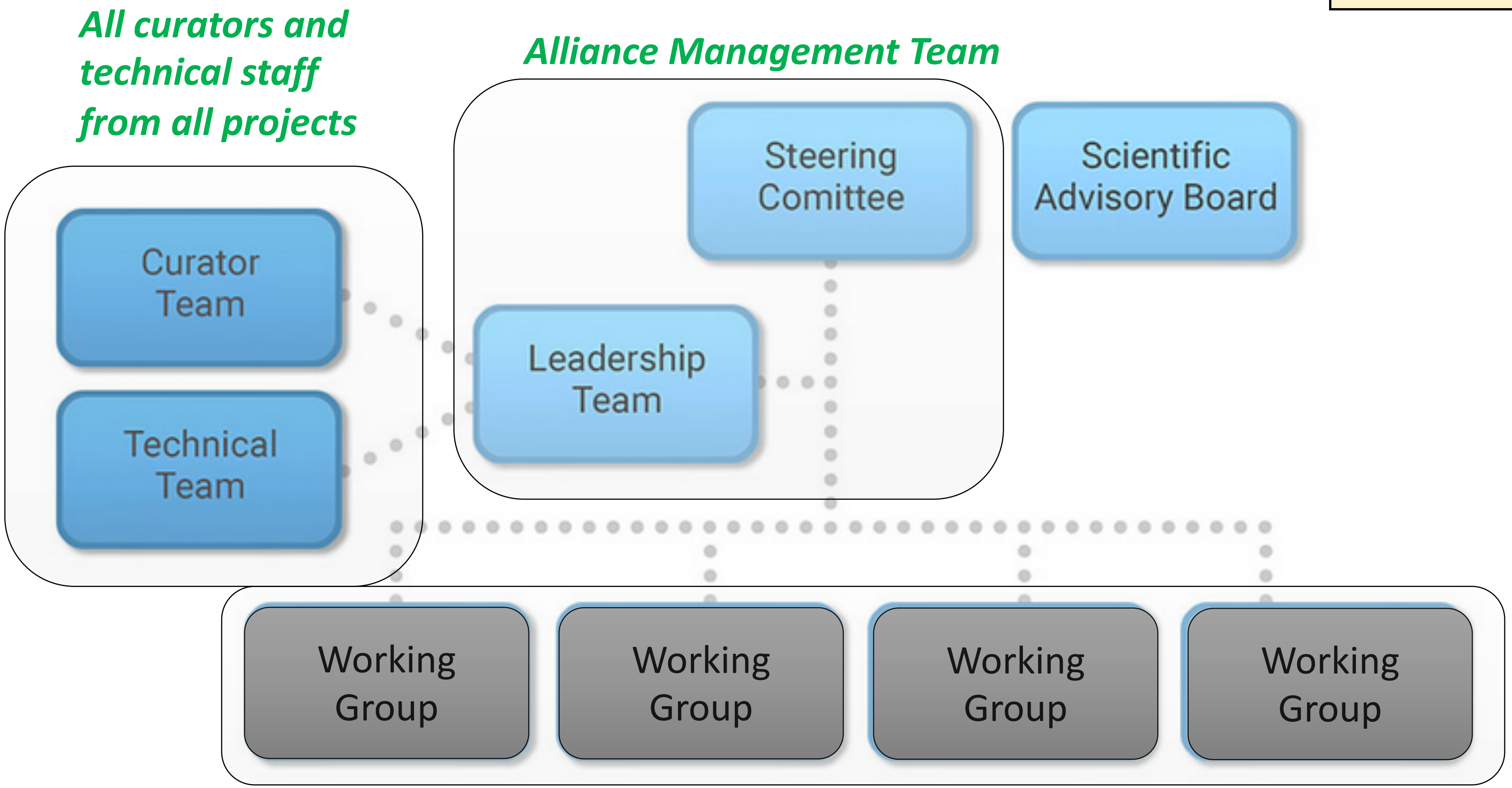
- Changed branching model to be more aligned with a Continuous Deployment model (should speed up release cycles)
- Common File Management System
- Updated front-end in accordance with API refactor changes
- Continual refinement to unified web page layout and navigation

Outreach

- Publications, Platform talks, Posters, Workshops, Booths
- Onboarding protocols for additional organisms in development

Alliance Governance

Steering Committee
includes PIs
from all the MODS and GO



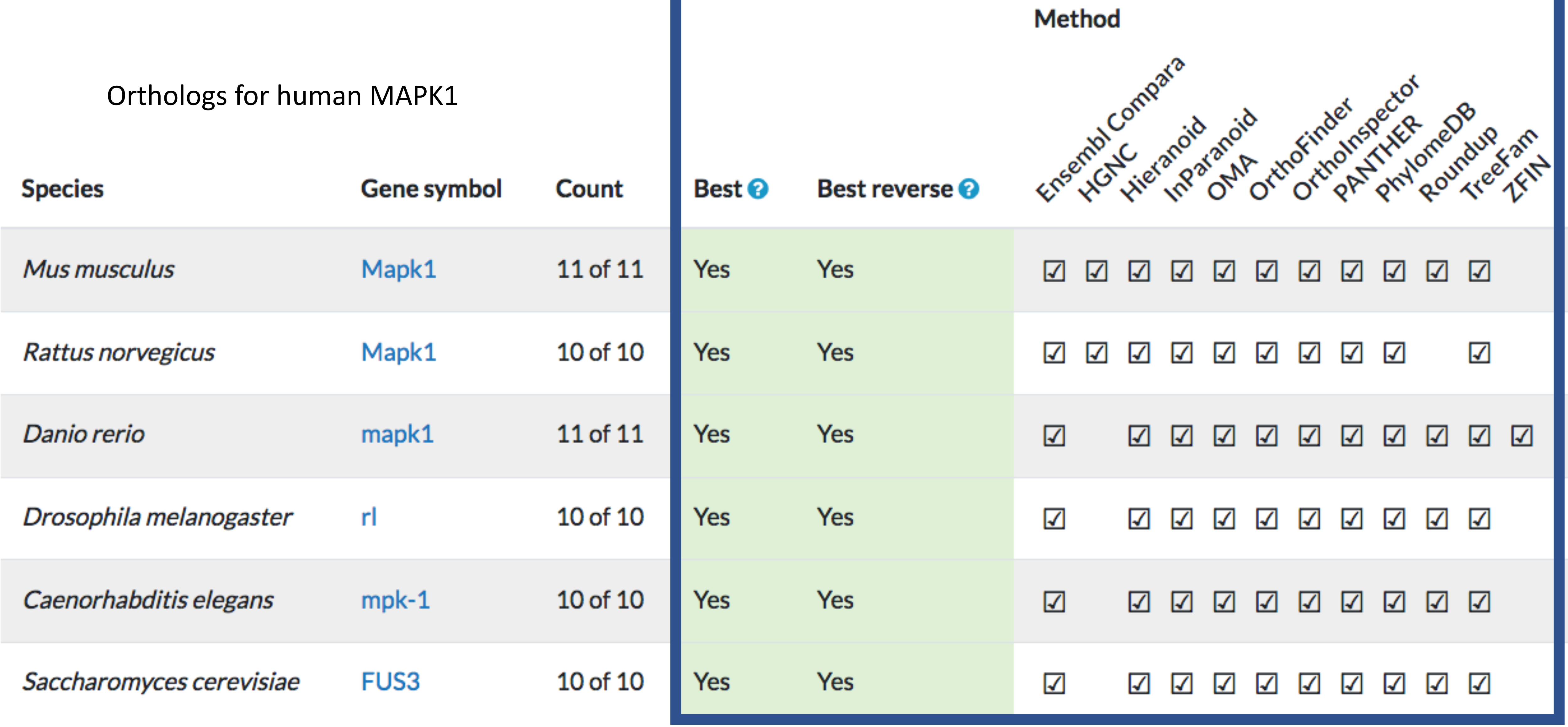
All curators and technical staff from all projects

Alliance Management Team

- *small, focused groups with specific deliverables*
- *often transient*
- *members from curator and technical teams from different MODs*

Common ortholog set

Orthologs for human MAPK1



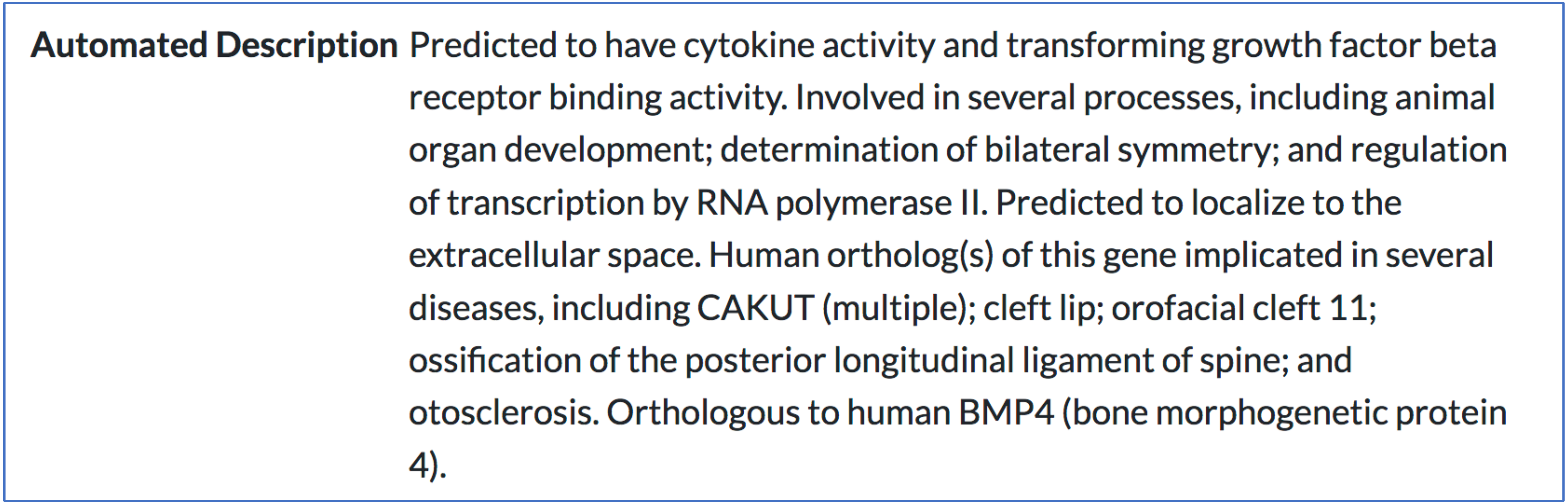
					Method											
Species	Gene symbol	Count	Best ?	Best reverse ?	Ensembl Compara	HGNC	Hieranoid	InParanoid	OMA	OrthoFinder	OrthoInspector	PANTHER	PhylomeDB	Roundup	TreeFam	ZFIN
Mus musculus	Mapk1	11 of 11	Yes	Yes	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑	
Rattus norvegicus	Mapk1	10 of 10	Yes	Yes	☑	☑	☑	☑	☑	☑	☑	☑	☑		☑	
Danio rerio	mapk1	11 of 11	Yes	Yes	☑		☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
Drosophila melanogaster	rl	10 of 10	Yes	Yes	☑		☑	☑	☑	☑	☑	☑	☑	☑	☑	
Caenorhabditis elegans	mpk-1	10 of 10	Yes	Yes	☑		☑	☑	☑	☑	☑	☑	☑	☑	☑	
Saccharomyces cerevisiae	FUS3	10 of 10	Yes	Yes	☑		☑	☑	☑	☑	☑	☑	☑	☑	☑	

Summary of orthology algorithms

Gene Descriptions

Automated. Rules-based. Generated from curated, structured annotations provided by Knowledge Centers.

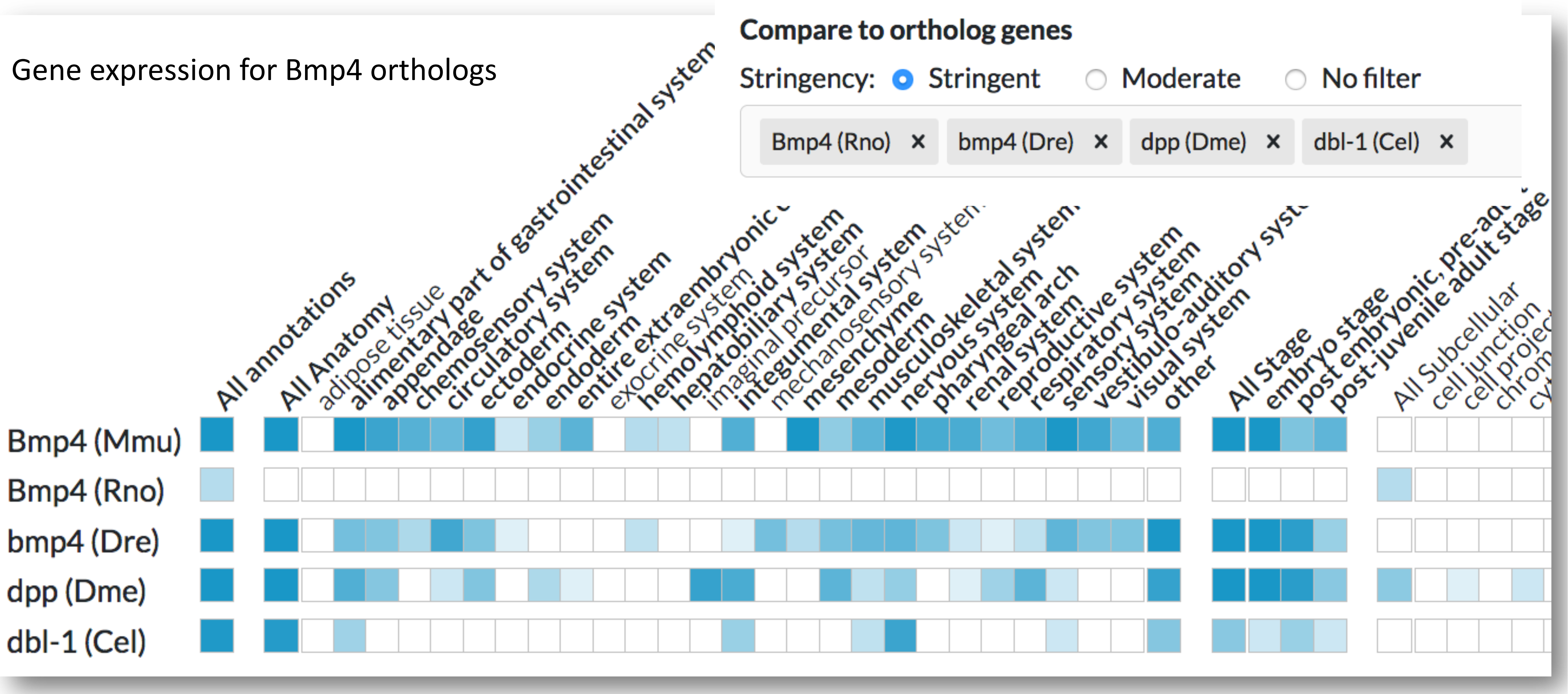
Gene description for bmp4 in zebrafish:



Automated Description Predicted to have cytokine activity and transforming growth factor beta receptor binding activity. Involved in several processes, including animal organ development; determination of bilateral symmetry; and regulation of transcription by RNA polymerase II. Predicted to localize to the extracellular space. Human ortholog(s) of this gene implicated in several diseases, including CAKUT (multiple); cleft lip; orofacial cleft 11; ossification of the posterior longitudinal ligament of spine; and otosclerosis. Orthologous to human BMP4 (bone morphogenetic protein 4).

https://www.alliancegenome.org/gene/ZFIN:ZDB-GENE-980528-2059

Comparative Gene Expression using Ribbon Annotation Summaries



Gene expression for Bmp4 orthologs

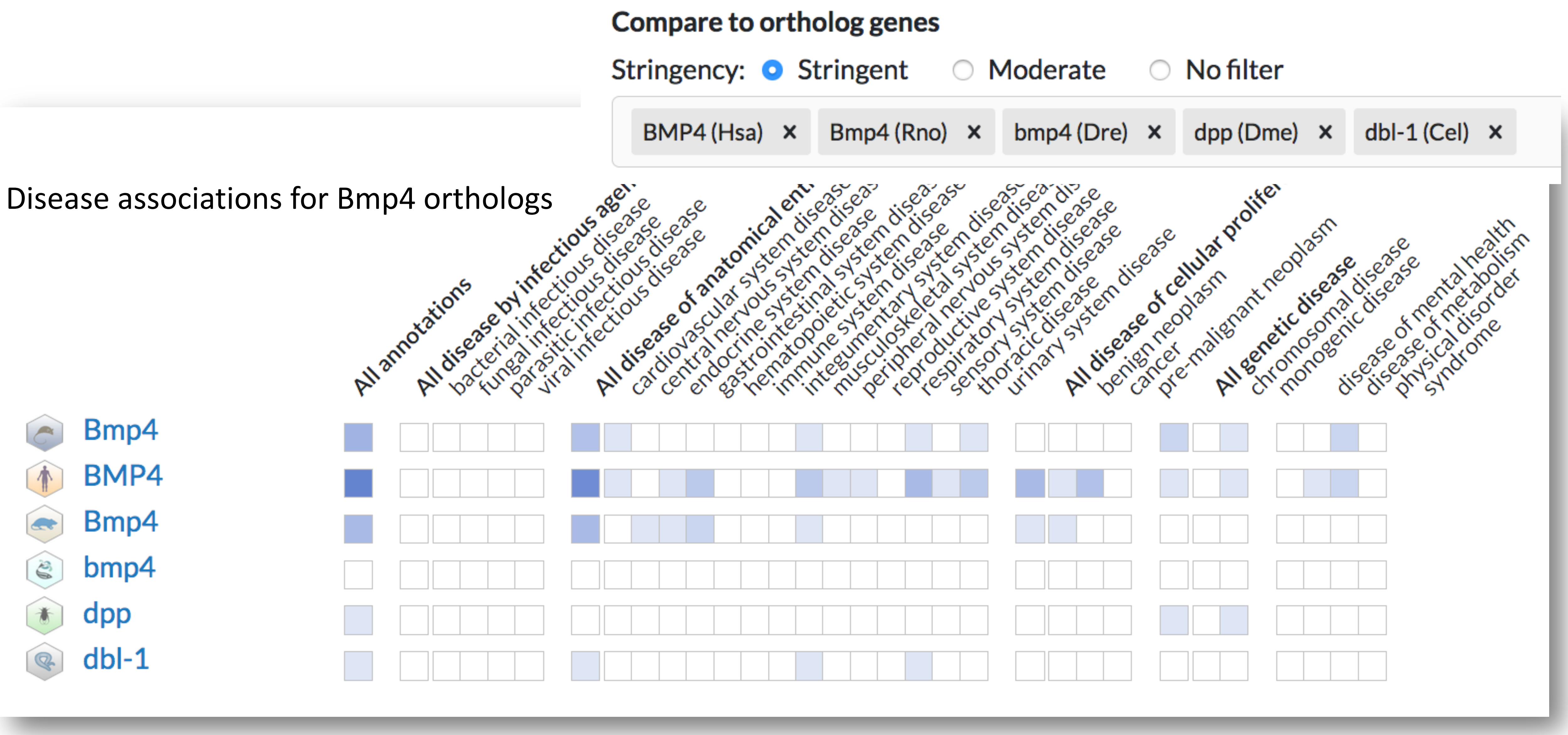
Compare to ortholog genes

Stringency: ● Stringent ○ Moderate ○ No filter

Bmp4 (Rno) × bmp4 (Dre) × dpp (Dme) × dbl-1 (Cel) ×

https://www.alliancegenome.org/gene/MGI:88180

Comparative Disease Annotation Using Ribbon Annotation Summaries



Compare to ortholog genes

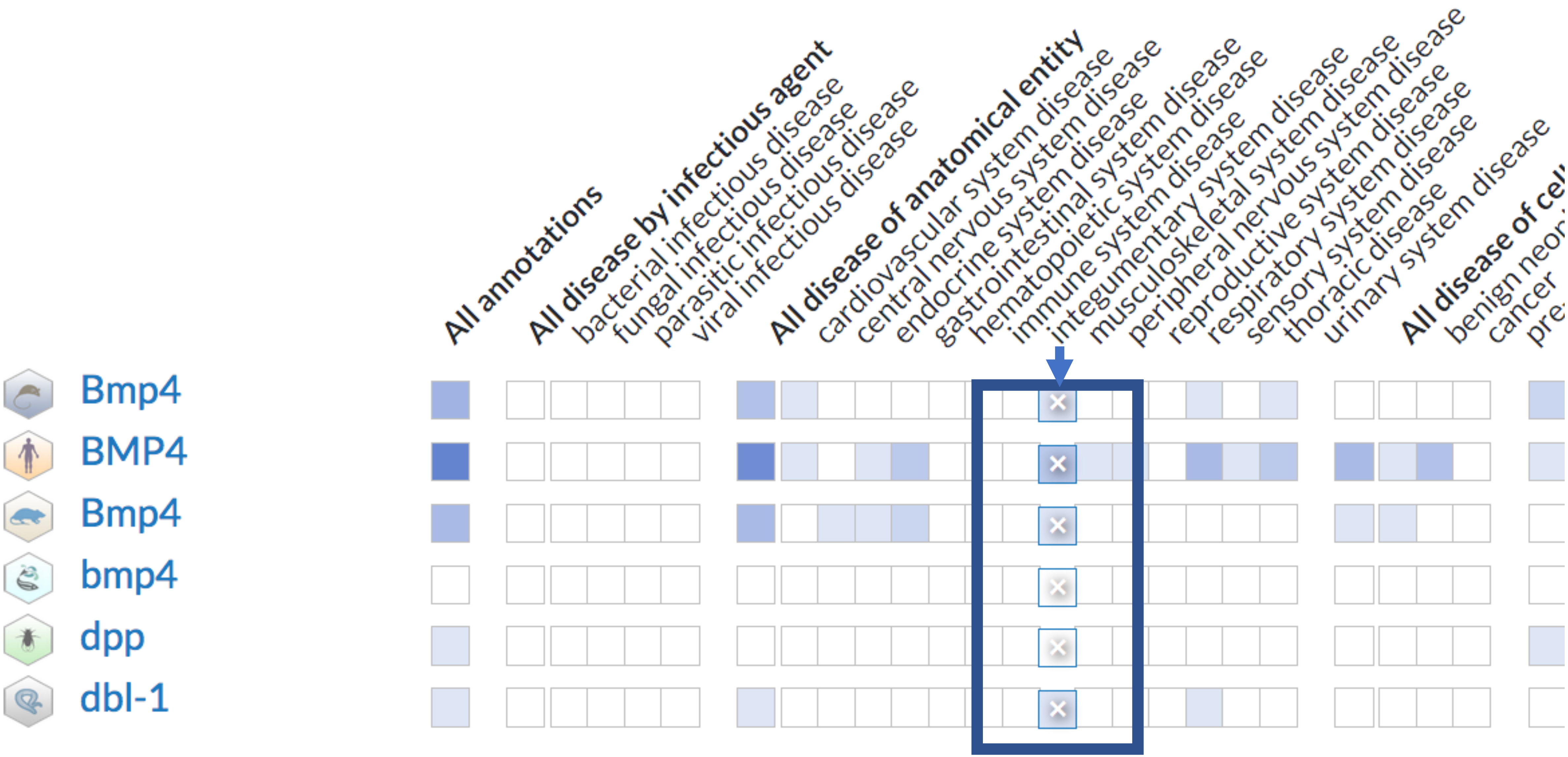
Stringency: ● Stringent ○ Moderate ○ No filter

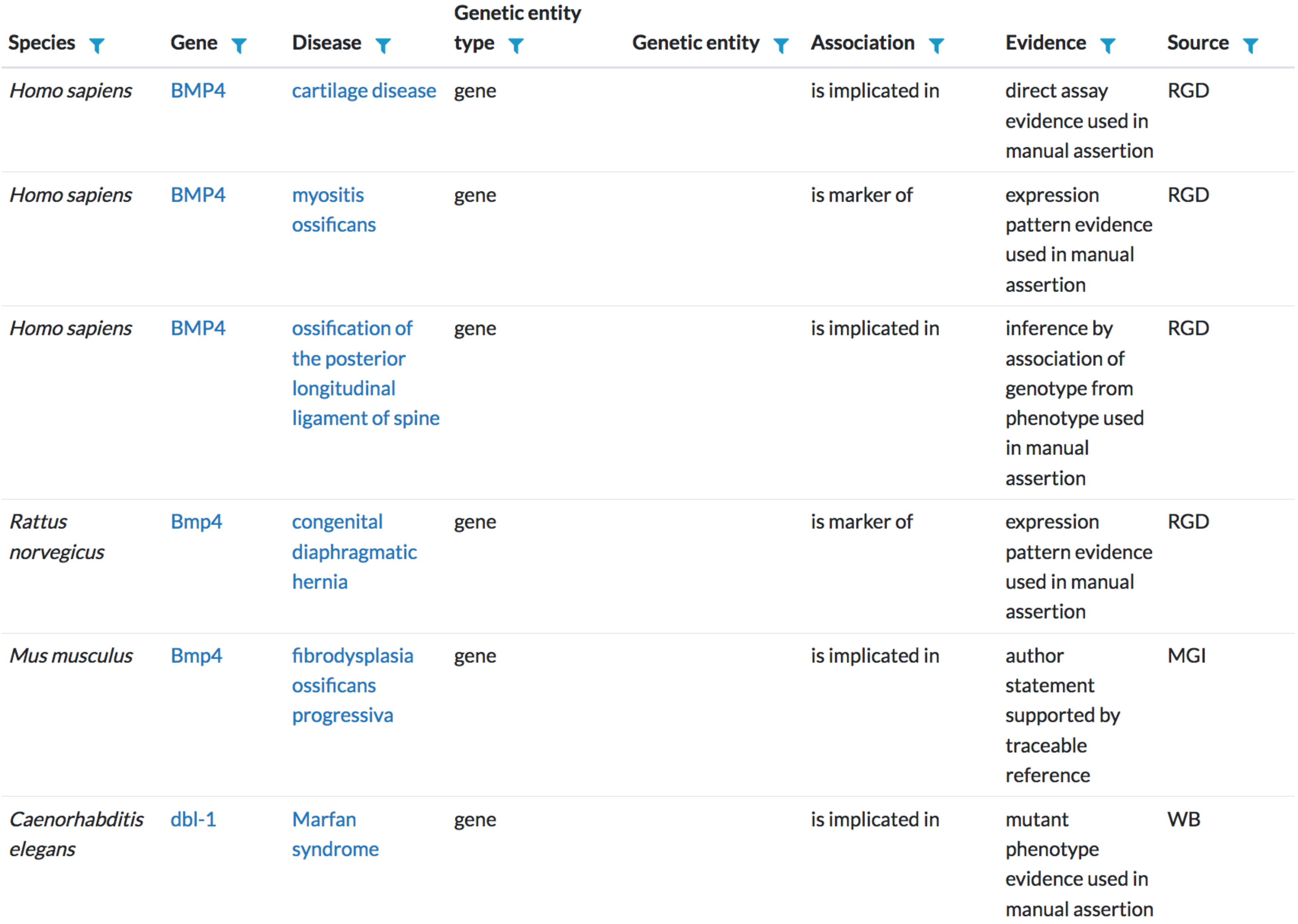
BMP4 (Hsa) × Bmp4 (Rno) × bmp4 (Dre) × dpp (Dme) × dbl-1 (Cel) ×

Disease associations for Bmp4 orthologs

All annotations
All disease by infectious agen
bacterial infectious disease
fungal infectious disease
parasitic infectious disease
viral infectious disease
All disease of anatomical ent
cardiovascular system diseas
central nervous system diseas
endocrine system disease
gastrointestinal system diseas
hematopoietic system disease
immune system disease
integumentary system diseas
musculoskeletal system disea
peripheral nervous system dis
reproductive system disease
respiratory system disease
sensory system disease
thoracic disease
urinary system disease
All disease of cellular prolifer
benign neoplasm
cancer
pre-malignant neoplasm
All genetic disease
chromosomal disease
monogenic disease
disease of mental health
disease of metabolism
physical disorder
syndrome
Bmp4
BMP4
Bmp4
bmp4
dpp
dbl-1

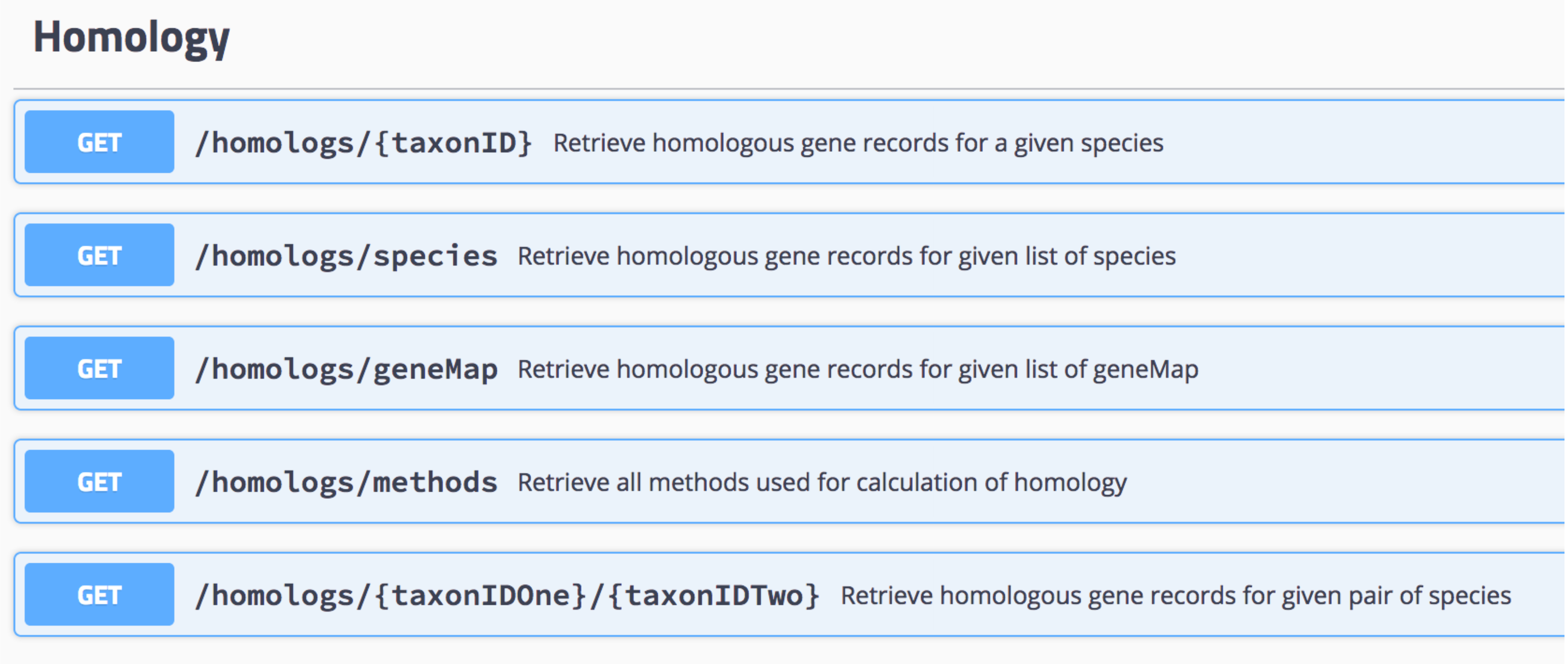
https://www.alliancegenome.org/gene/MGI:88180


All annotations
All disease by infectious agent
bacterial infectious disease
fungal infectious disease
parasitic infectious disease
viral infectious disease
All disease of anatomical entity
cardiovascular system disease
central nervous system disease
endocrine system disease
gastrointestinal system disease
hematopoietic system disease
immune system disease
integumentary system disease
musculoskeletal system disease
peripheral nervous system disease
reproductive system disease
respiratory system disease
sensory system disease
thoracic disease
urinary system disease
All disease of cel
benign neo
cancer
pre
Bmp4
BMP4
Bmp4
bmp4
dpp
dbl-1



Species	Gene	Disease	Genetic entity type	Genetic entity	Association	Evidence	Source
Homo sapiens	BMP4	cartilage disease	gene		is implicated in	direct assay evidence used in manual assertion	RGD
Homo sapiens	BMP4	myositis ossificans	gene		is marker of	expression pattern evidence used in manual assertion	RGD
Homo sapiens	BMP4	ossification of the posterior longitudinal ligament of spine	gene		is implicated in	inference by association of genotype from phenotype used in manual assertion	RGD
Rattus norvegicus	Bmp4	congenital diaphragmatic hernia	gene		is marker of	expression pattern evidence used in manual assertion	RGD
Mus musculus	Bmp4	fibrodysplasia ossificans progressiva	gene		is implicated in	author statement supported by traceable reference	MGI
Caenorhabditis elegans	dbl-1	Marfan syndrome	gene		is implicated in	mutant phenotype evidence used in manual assertion	WB

Common Application Programming Interfaces (APIs)



Homology

GET `/homologs/{taxonID}` Retrieve homologous gene records for a given species

GET `/homologs/species` Retrieve homologous gene records for given list of species

GET `/homologs/geneMap` Retrieve homologous gene records for given list of geneMap

GET `/homologs/methods` Retrieve all methods used for calculation of homology

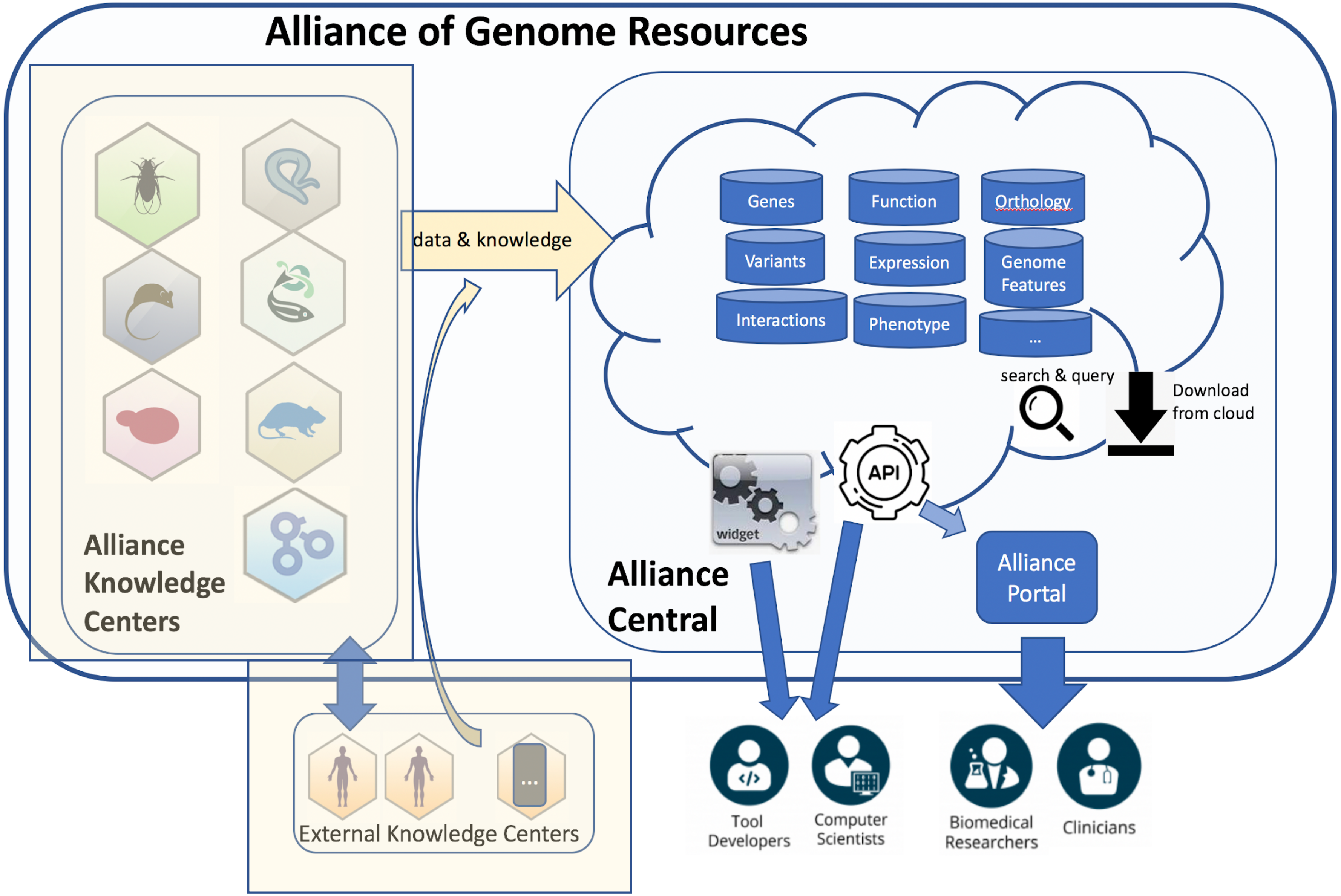
GET `/homologs/{taxonIDOne}/{taxonIDTwo}` Retrieve homologous gene records for given pair of species

https://www.alliancegenome.org/api/swagger-ui/

Near term focus areas

- Shared User Interfaces
    - Comparative ribbon display of harmonized phenotype data
    - Display genome variants associated with phenotypic alleles
    - Graphical representation of molecular interaction data
- Governance
    - Onboarding guidelines for new Alliance members/contributors
    - Core Seal Trust application to be certified as a “trustworthy data repository”
        - https://www.coretrustseal.org/

Alliance Knowledge Centers (e.g., MODs)



Alliance Knowledge Centers (MODs) will continue to focus on organism-centric curation

- Organism-centric knowledgebases
  - Develop and apply nomenclature and annotation standards
  - Expert curation of the scientific literature
  - Expert knowledge of the organism and organism-specific reagents and data
  - Centers for outreach to organism-specific research communities
  - Front lines for FAIR compliance

Knowledge Centers share a common mission: facilitate the use of model organisms and comparative biology to support investigations into the genetic and genomic basis of human health and disease.

NHGRI-funded components of Knowledge Centers focus on data relevant to genomics and genome biology: Genes, Variants, Expression, Phenotype, Disease

Expert curation is the core of any knowledgebase

Scalable expert curation

- Volume of genome data and scientific publications have grown exponentially. # of curators has not. Why?
  - More papers published does not equate to new knowledge in relevant data types
  - Innovations in tools and methods for expert curation have enhanced efficiency
  - Skill sets of biocurators evolving (data wranglers, data analysts) as the data landscape evolves

"I am tempted to conclude that a very large fraction of the alleged 35,000 journals now current must be reckoned as merely a distant background noise, and as very far from central or strategic in any of the knitted strips from which the cloth of science is woven".

D.J. de Solla Price. *Science*, 1965

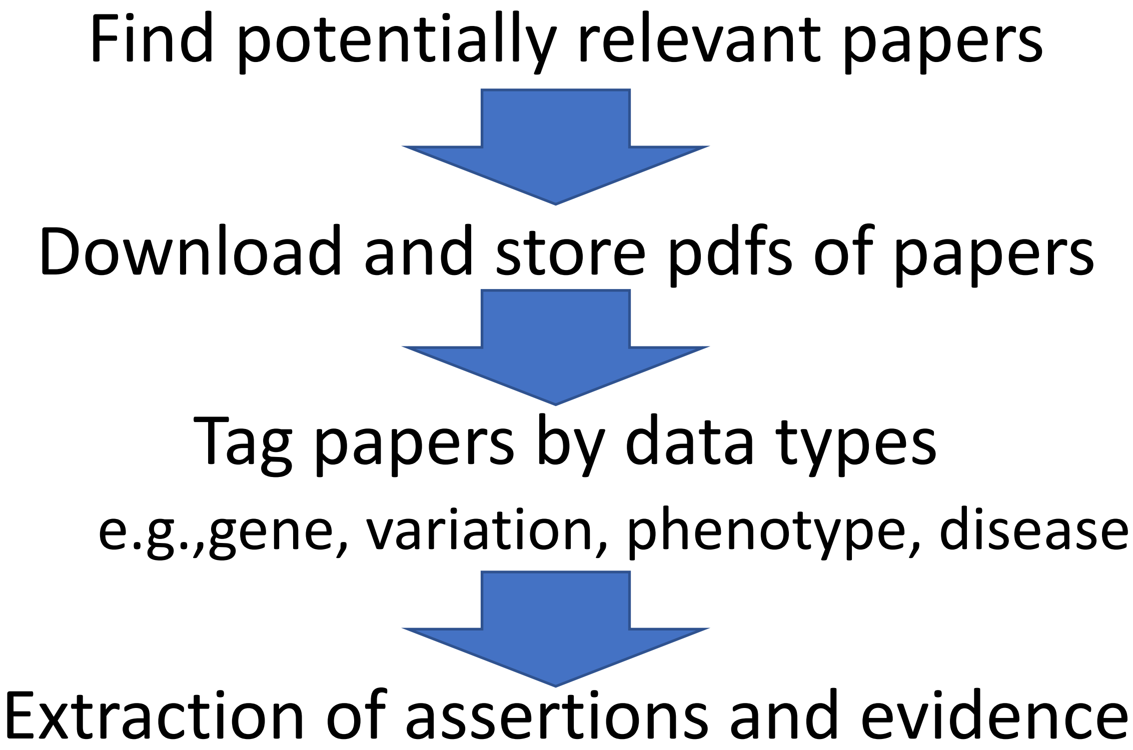
"We show that 90% of the papers in PubMed are out of the scope of UniProt, that a maximum of 2–3% of the papers indexed in PubMed each year are relevant for UniProt curation.."

Poux et al. *Bioinformatics*, 2017

"One striking change from the 2009 results is that, as of 2012, the seven databases that participated in 2012 track are using text mining in at least some parts of their workflow. This contrasts with the 2009 survey, where less than half of the biocurators (46%) reported that they were currently using text mining."

Lu and Hirschmann. *Database*, 2012

Literature curation: process overview (simplified)



Find potentially relevant papers

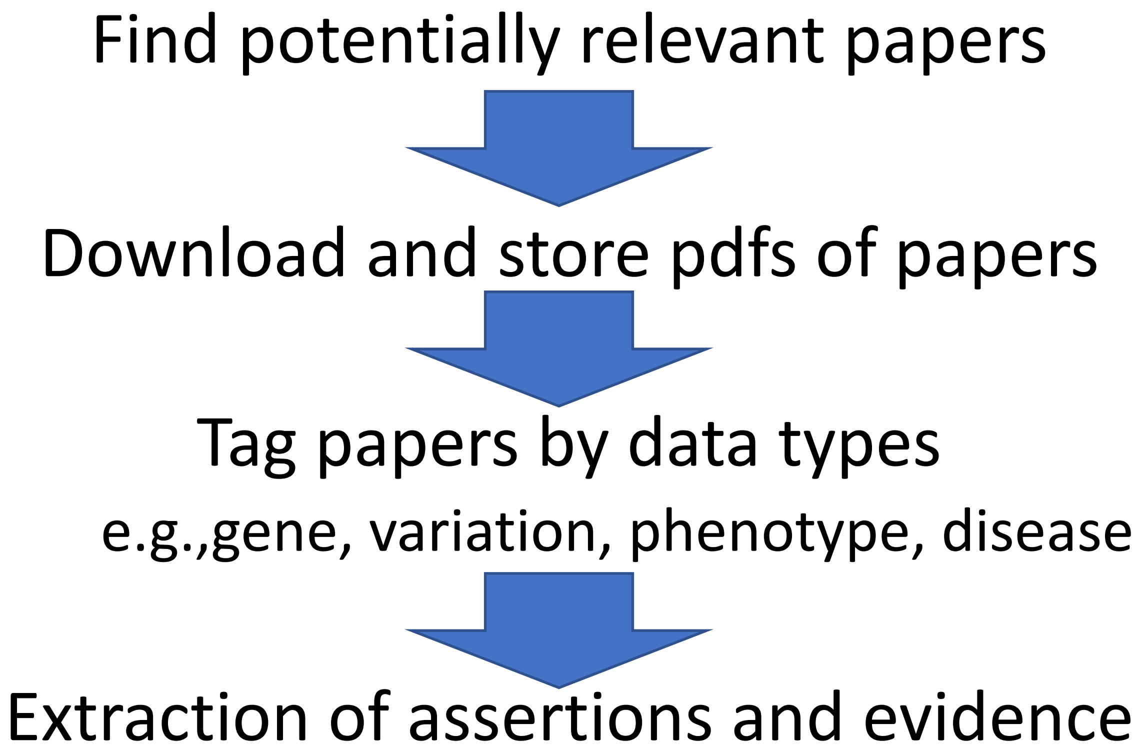
Download and store pdfs of papers

Tag papers by data types

e.g.,gene, variation, phenotype, disease

Extraction of assertions and evidence

Literature curation: automation



Find potentially relevant papers

Download and store pdfs of papers

Tag papers by data types
e.g.,gene, variation, phenotype, disease

Extraction of assertions and evidence

Automated…plus support for community input

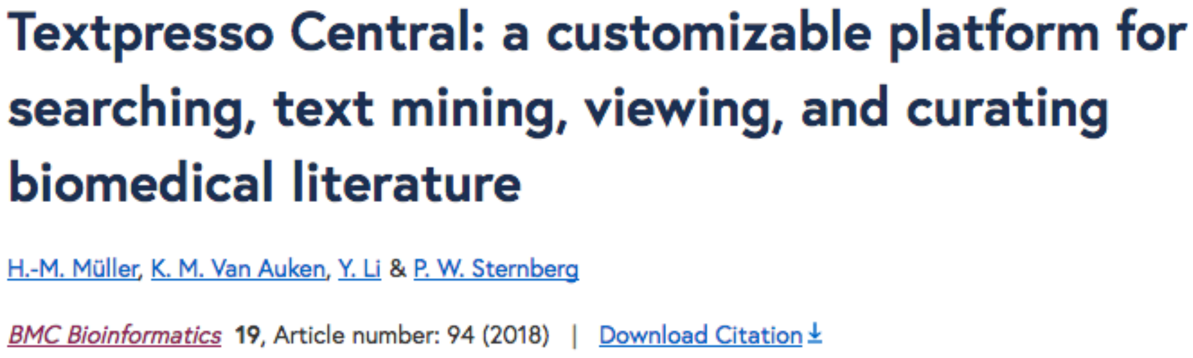
Automated when such access is supported

Semi-automated with human review

Custom annotation interfaces for curation efficiency

Tools to enhance efficiency, automation, and computability

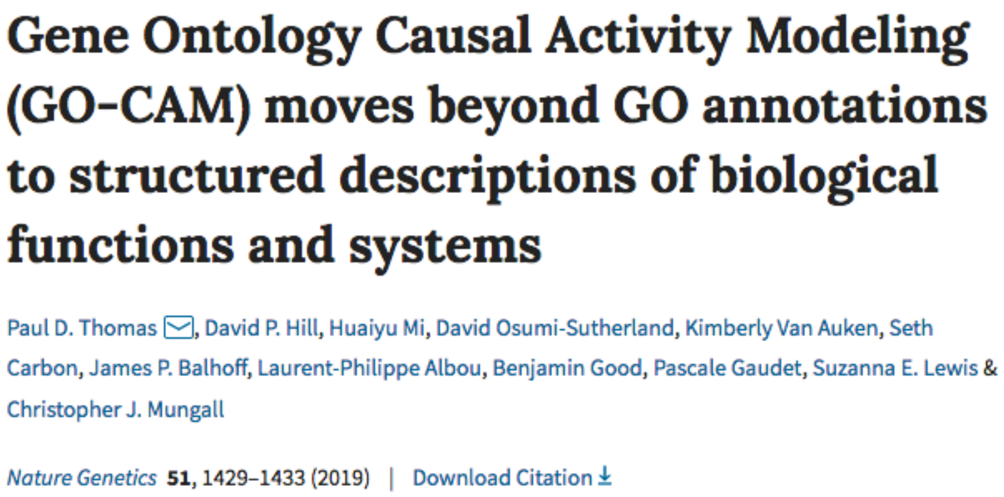
Examples of curation tools developed by Alliance members:



Textpresso Central: a customizable platform for searching, text mining, viewing, and curating biomedical literature

H.-M. Müller, K. M. Van Auken, Y. Li & P. W. Sternberg

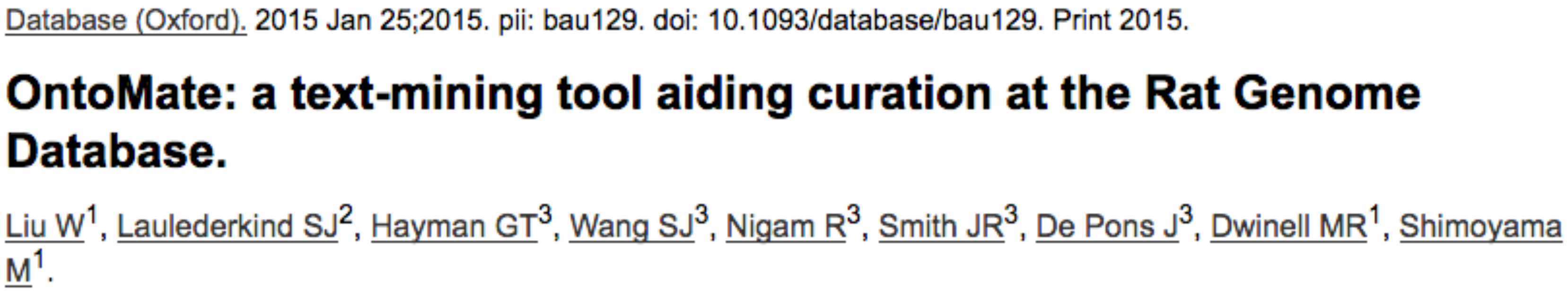
BMC Bioinformatics **19**, Article number: 94 (2018) | Download Citation ⤓



Gene Ontology Causal Activity Modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems

Paul D. Thomas ✉, David P. Hill, Huaiyu Mi, David Osumi-Sutherland, Kimberly Van Auken, Seth Carbon, James P. Balhoff, Laurent-Philippe Albou, Benjamin Good, Pascale Gaudet, Suzanna E. Lewis & Christopher J. Mungall

Nature Genetics **51**, 1429–1433 (2019) | Download Citation ⤓

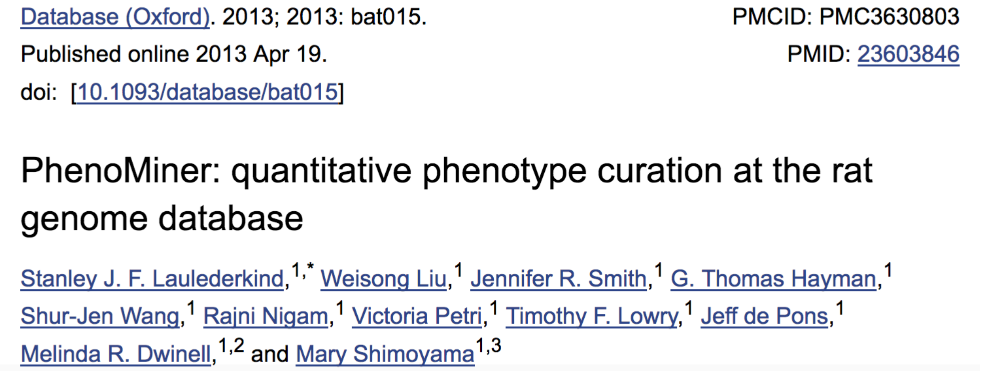


Database (Oxford). 2015 Jan 25;2015. pii: bau129. doi: 10.1093/database/bau129. Print 2015.

OntoMate: a text-mining tool aiding curation at the Rat Genome Database.

Liu W[1], Laulederkind SJ[2], Hayman GT[3], Wang SJ[3], Nigam R[3], Smith JR[3], De Pons J[3], Dwinell MR[1], Shimoyama M[1].

Examples of publications by Alliance members on curation methods/improvements:



Database (Oxford). 2013; 2013: bat015.
Published online 2013 Apr 19.
doi: [10.1093/database/bat015]

PMCID: PMC3630803
PMID: 23603846

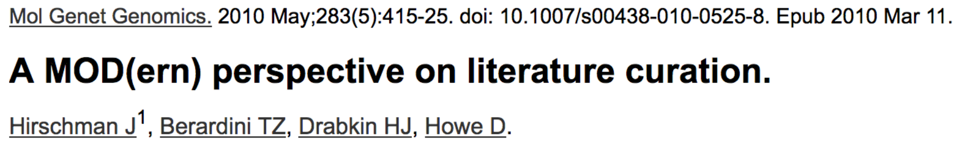
PhenoMiner: quantitative phenotype curation at the rat genome database

Stanley J. F. Laulederkind,[1,*] Weisong Liu,[1] Jennifer R. Smith,[1] G. Thomas Hayman,[1] Shur-Jen Wang,[1] Rajni Nigam,[1] Victoria Petri,[1] Timothy F. Lowry,[1] Jeff de Pons,[1] Melinda R. Dwinell,[1,2] and Mary Shimoyama[1,3]

Mamm Genome. 2015 Aug;26(7-8):295-304. doi: 10.1007/s00335-015-9571-1. Epub 2015 Jun 18.

A unified gene catalog for the laboratory mouse reference genome.

Zhu Y[1], Richardson JE, Hale P, Baldarelli RM, Reed DJ, Recla JM, Sinclair R, Reddy TB, Bult CJ.



Mol Genet Genomics. 2010 May;283(5):415-25. doi: 10.1007/s00438-010-0525-8. Epub 2010 Mar 11.

A MOD(ern) perspective on literature curation.

Hirschman J[1], Berardini TZ, Drabkin HJ, Howe D.

Disease Ontology: improving and unifying disease annotations across species

Susan M. Bello, Mary Shimoyama, Elvira Mitraka, Stanley J. F. Laulederkind, Cynthia L. Smith, Janan T. Eppig, Lynn M. Schriml

Disease Models & Mechanisms 2018 11: dmm032839 doi: 10.1242/dmm.032839 Published 12 March 2018

Methods Mol Biol. 2018;1757:307-347. doi: 10.1007/978-1-4939-7737-6_11.

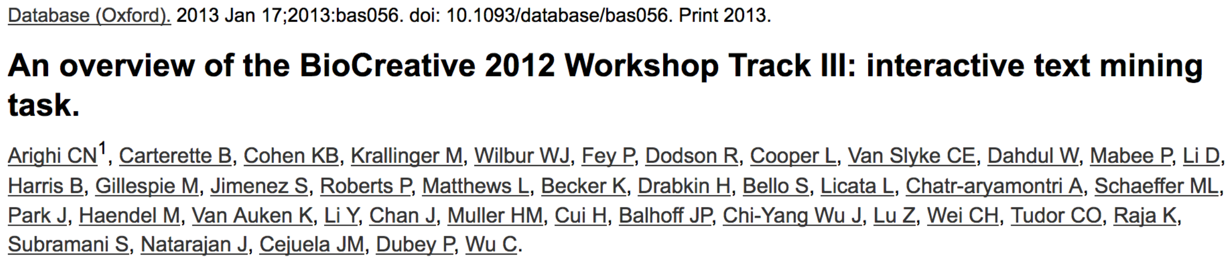
Using ZFIN: Data Types, Organization, and Retrieval.

Van Slyke CE[1], Bradford YM[2], Howe DG[2], Fashena DS[2], Ramachandran S[2], Ruzicka L[2]; ZFIN Staff*.

Nucleic Acids Res. 2013 Jan;41(Database issue):D751-7. doi: 10.1093/nar/gks1024. Epub 2012 Nov 3.

FlyBase: improvements to the bibliography.

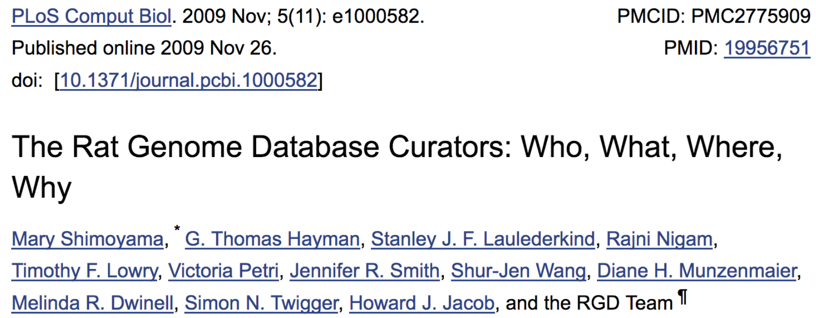
Marygold SJ[1], Leyland PC, Seal RL, Goodman JL, Thurmond J, Strelets VB, Wilson RJ; FlyBase consortium.



Database (Oxford). 2013 Jan 17;2013:bas056. doi: 10.1093/database/bas056. Print 2013.

An overview of the BioCreative 2012 Workshop Track III: interactive text mining task.

Arighi CN[1], Carterette B, Cohen KB, Krallinger M, Wilbur WJ, Fey P, Dodson R, Cooper L, Van Slyke CE, Dahdul W, Mabee P, Li D, Harris B, Gillespie M, Jimenez S, Roberts P, Matthews L, Becker K, Drabkin H, Bello S, Licata L, Chatr-aryamontri A, Schaeffer ML, Park J, Haendel M, Van Auken K, Li Y, Chan J, Muller HM, Cui H, Balhoff JP, Chi-Yang Wu J, Lu Z, Wei CH, Tudor CO, Raja K, Subramani S, Natarajan J, Cejuela JM, Dubey P, Wu C.

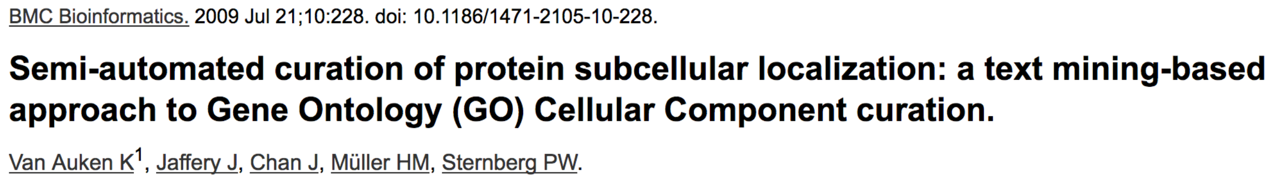


PLoS Comput Biol. 2009 Nov; 5(11): e1000582.
Published online 2009 Nov 26.
doi: [10.1371/journal.pcbi.1000582]

PMCID: PMC2775909
PMID: 19956751

The Rat Genome Database Curators: Who, What, Where, Why

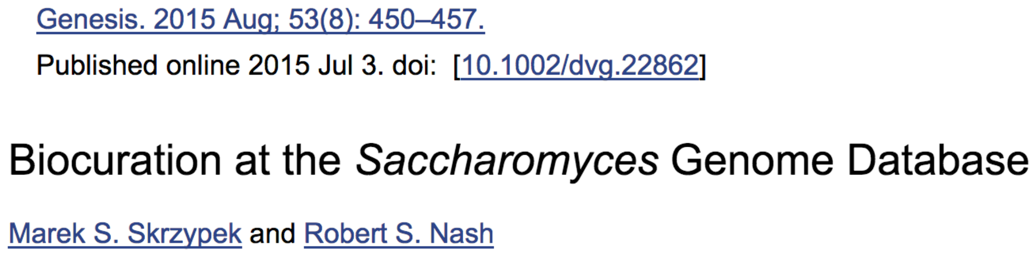
Mary Shimoyama,[*] G. Thomas Hayman, Stanley J. F. Laulederkind, Rajni Nigam, Timothy F. Lowry, Victoria Petri, Jennifer R. Smith, Shur-Jen Wang, Diane H. Munzenmaier, Melinda R. Dwinell, Simon N. Twigger, Howard J. Jacob, and the RGD Team[¶]



BMC Bioinformatics. 2009 Jul 21;10:228. doi: 10.1186/1471-2105-10-228.

Semi-automated curation of protein subcellular localization: a text mining-based approach to Gene Ontology (GO) Cellular Component curation.

Van Auken K[1], Jaffery J, Chan J, Müller HM, Sternberg PW.



Genesis. 2015 Aug; 53(8): 450–457.
Published online 2015 Jul 3. doi: [10.1002/dvg.22862]

Biocuration at the *Saccharomyces* Genome Database

Marek S. Skrzypek and Robert S. Nash

Barriers to automation of literature curation

- Only 1/3 of PubMed Central is open access for automatic download
- Full text of publications are not always available for data mining tools
- Publishers haven’t embraced ‘concept tagging’ of manuscripts prior to publication
- Errors, ambiguities, omissions in publications require human intervention
- Text mining for automated extraction of annotations and evidence don’t yet achieve levels of accuracy and quality compared to professional biocurators

“…54% of resources are not uniquely identifiable in publications, regardless of domain, journal impact factor, or reporting requirements. “

Vasilevsky et al. *PeerJ*, 2013

“…we conclude that the state of the art in automatically mining GO terms from literature has improved over the past decade while much progress is still needed for computer-assisted GO curation. “

Mao et al., *Database*, 2014

Literature curation: continual process improvements

Alliance Central and Alliance Knowledge Centers will….

- Continue to collaborate with AI/ML/NLP communities to improve automation at all steps
  - Alliance is a source of positive and negative training sets for advancing complex concept extraction methods
- Continue to work with publishers to require that authors include official nomenclature and annotation standards at time of publication
  - Work with tool developers to build software that simplifies this process for authors
- Continue to develop support for post publication community annotation and quality assurance
  - e.g., Author First pass (WormBase); FlyBase FastTrack Your Paper (FlyBase); Your Input Welcome (MGD), etc.

Impact: curated data saves researchers time

- Enhances data findability, aggregation, reproducibility, and re-use
  - Nomenclature and annotation standards
  - Evidence codes
  - Permanent and unique identifiers
  - Quality control
- Annotations and data from MODs are integrated into major genomic data resources and knowledge aggregators (both public and commercial)
  - e.g., NCBI, Ensembl, UCSC Genome Browser, GeneCards, Wiki Gene, Wiki Data, Ingenuity Pathway Analysis, UniProt, many others
- Makes data ‘hidden’ in supplemental data findable
- Makes it easy to find the most relevant publications for a specific annotation
- Advances timeframes for developing novel data mining methods
  - Reduces time needed for data ‘cleaning’ efforts
  - Computable annotations

Impact: curated data drives innovation in genomics and data science

- Gene Ontology is a foundation for *genomic data interpretation*
- MODs provide training sets for advances underlying innovations in *automated literature curation processes*
    - e.g., BioCreative
- MODs (and GOC) provide computation ready data sets and models that drive research in semantic reasoning for *predictive biology*
    - e.g., Monarch Initiative, Phenodigm, Exomizer, others
- Annotations and data from MODs essential for resources developed to aid in *functionalizing human genome variation*
    - e.g, MARRVEL

The Alliance is a hybrid of centralized and federated organization

- Centralization most effective when tight coordination of processes improves efficiency and reduces duplication of effort
  - Appropriate for infrastructure development goals of Alliance Central

- Federation most effective when context-specific adaptations and expertise are paramount to success and sustainability
  - A federated model is better suited for Knowledge Centers Coordination where coordination is achieved by cooperative development and deployment of data standards

What would be lost by centralizing Knowledge Centers?

- Effectiveness and efficiency of curation (time and money)
  - Tools and approaches for data acquisition and curation can differ among model organisms…even for common data types
    - e.g., genome feature annotation, phenotype and disease associations
- Quality
  - Assessing quality and accuracy of annotations and reagents in papers requires deep organism-specific knowledge and strong community connections
- Responsiveness to user organism-specific communities
  - Different model organisms have different experimental strengths and user communities…not “one-size fits all”
  - Many standards and innovations are developed in close partnership with organism-specific research communities
- Effectiveness of process management
  - Effectiveness of centralized management drops as “span of control” (i.e., number of different data resources) increases because of decision and resource bottlenecks
  - The number of model organisms interested in joining the Alliance is increasing!
  - Not all knowledgebases relevant to the Alliance mission will be members of the consortium
- Effective oversight
  - Meaningful evaluation of metrics for MODs requires appropriate context

Summary

- The Alliance of Genome Resources is modernizing the data ecosystem for model organism data and knowledge
  - Alliance Central
    - Centralized organization; <u>“develop once, use by all” </u>approach for data types in common across model organisms
  - Knowledge Centers
    - Federated organization; <u>“common mission, different paths” </u>with focus on standards to unify data and knowledge

Effective, scalable, adaptable, quality-focused, professional, user centered

Acknowledgements

Alliance Executive Steering Committee

- Judith A. Blake
- Carol J. Bult
- Brian Calvi
- J. Michael Cherry
- Kevin Howe
- Chris Mungall
- Norbert Perrimon
- Mary Shimoyama
- Paul Sternberg
- Paul Thomas
- Monte Westerfield

Alliance Scientific Advisory Board*

Helen Berman, Brian Oliver, *Anne Kwitek*, *Peter Robinson*, *Wayne Frankel*, Gary Bader, Shawn Burgess, Andrew Chisholm, Phil Hieter, Calum MacRae, Alex Bateman

* SAB members in italics rotating off in 2019

Alliance of Genome Resources All Hands meeting (Stanford University December 2018)

NIH ***U24HG010859***

Next All Hands/SAB meeting – Boston, MA December 2019