InterMine as a Science Gateway for Systems Biology

Aaron Golden
School of Mathematics, Statistics & Applied Mathematics
National University of Ireland, Galway

IWSG 2017 – Poznan, Poland

Talk Roadmap

(1) Desktop technologies & systems biology

(2) Problem – how to leverage resulting data space

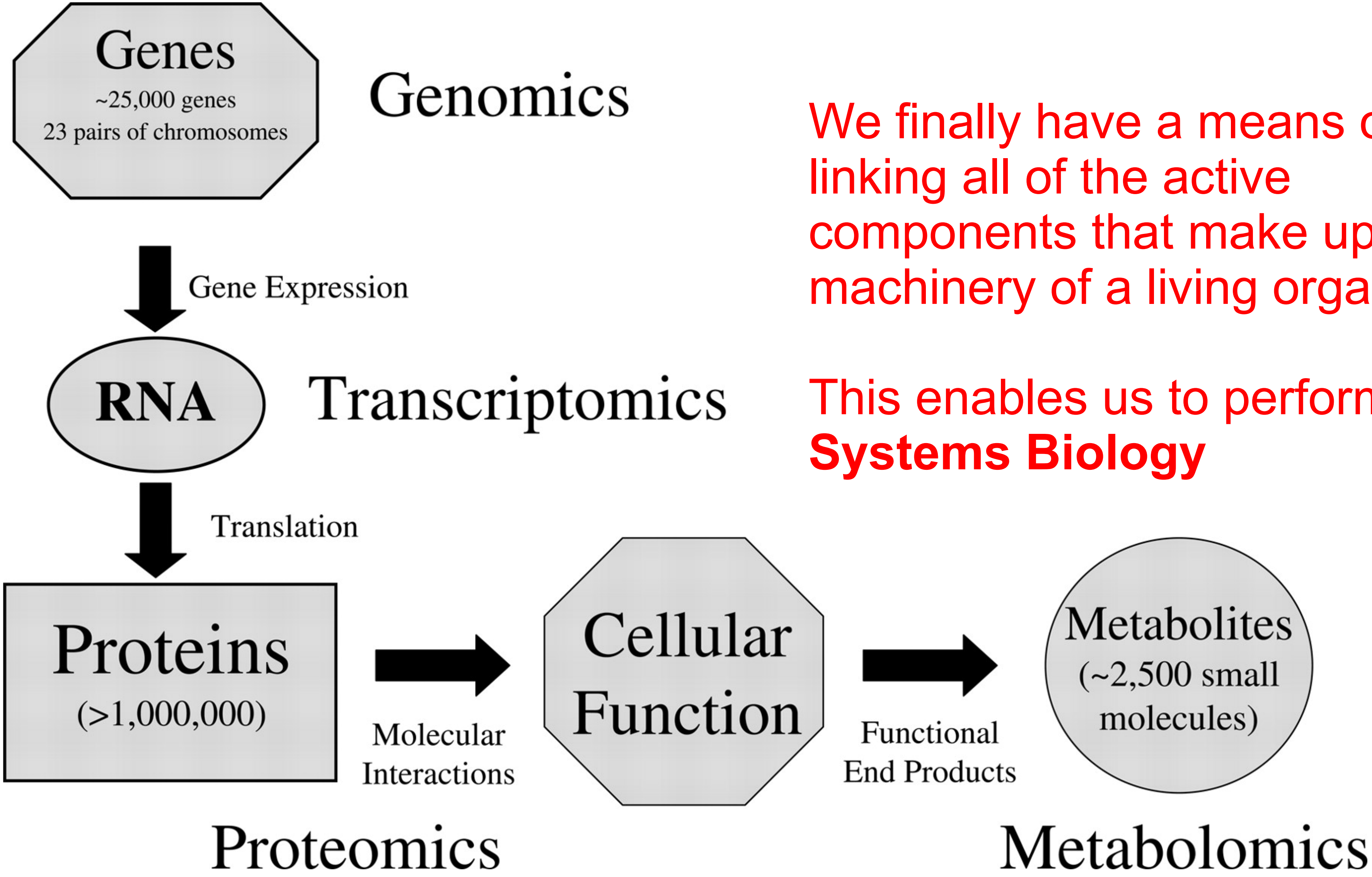
(3) Sequence Ontology & Data Schemas

(4) InterMine – how it works as a Science Gateway

NUI Galway
OÉ Gaillimh

InterMine

The Good News...



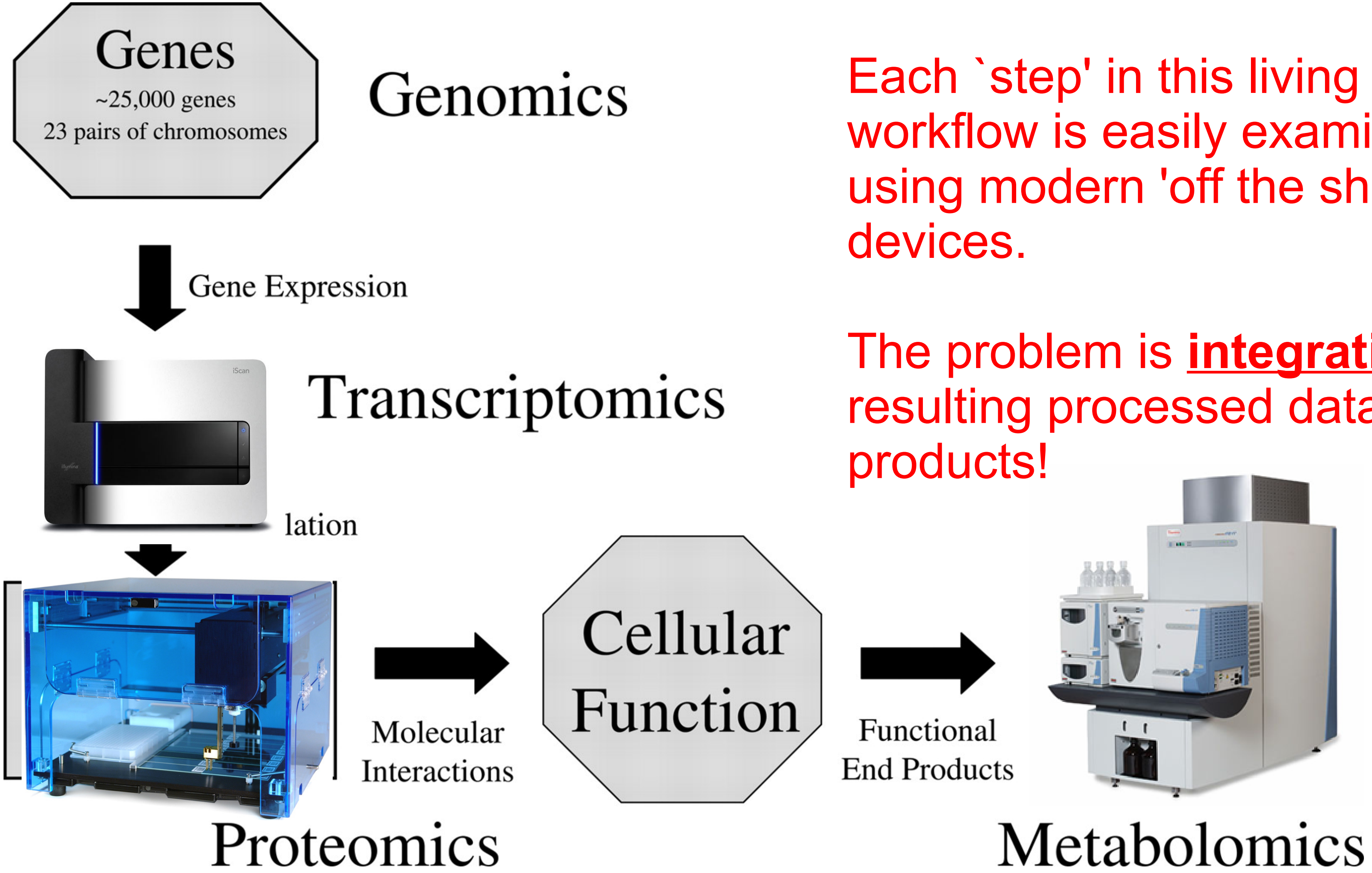
We finally have a means of fully linking all of the active components that make up the machinery of a living organism.

This enables us to perform **Systems Biology**

NUI Galway
OÉ Gaillimh

InterMine

The Problem...



Each `step' in this living workflow is easily examined using modern 'off the shelf' devices.

The problem is <u>**integrating**</u> the resulting processed data products!

NUI Galway
OÉ Gaillimh

InterMine

Ontologies to the Rescue...

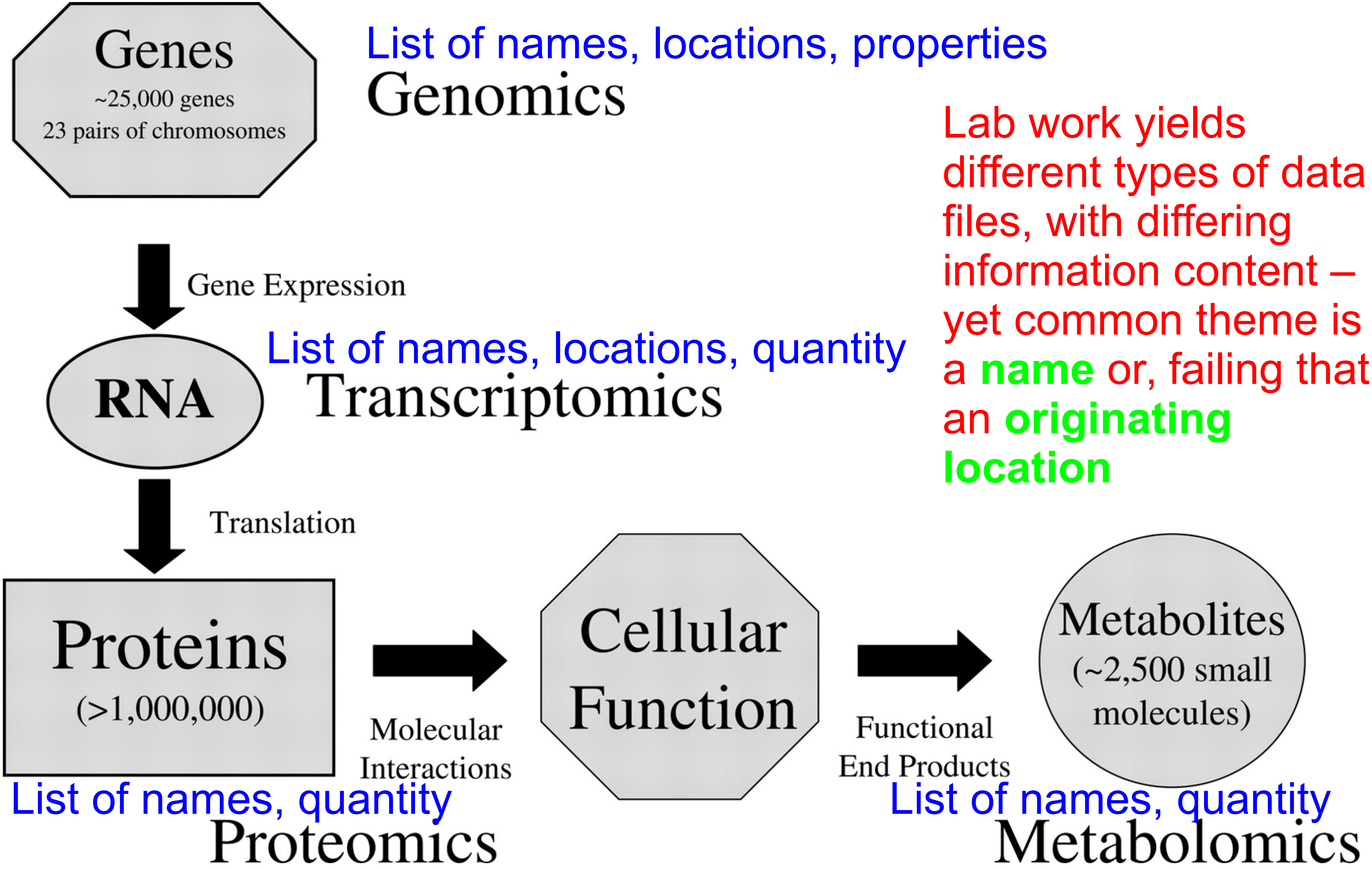
Solution lies in the SO The Sequence Ontology Project

Forms a schema based around the ultimate frame of reference
 - the genome

Genomic features described as parts of gene models, assembly components, experimental results generating annotations defined as being associated with a specific location...

Forms an excellent means ***to bind the various systems biology datasets together in the same data space***

The basis for building a schema...



Genes

~25,000 genes

23 pairs of chromosomes

List of names, locations, properties

Genomics

Gene Expression

RNA

List of names, locations, quantity

Transcriptomics

Translation

Proteins

(>1,000,000)

Molecular Interactions

Cellular Function

Functional End Products

Metabolites

(~2,500 small molecules)

List of names, quantity

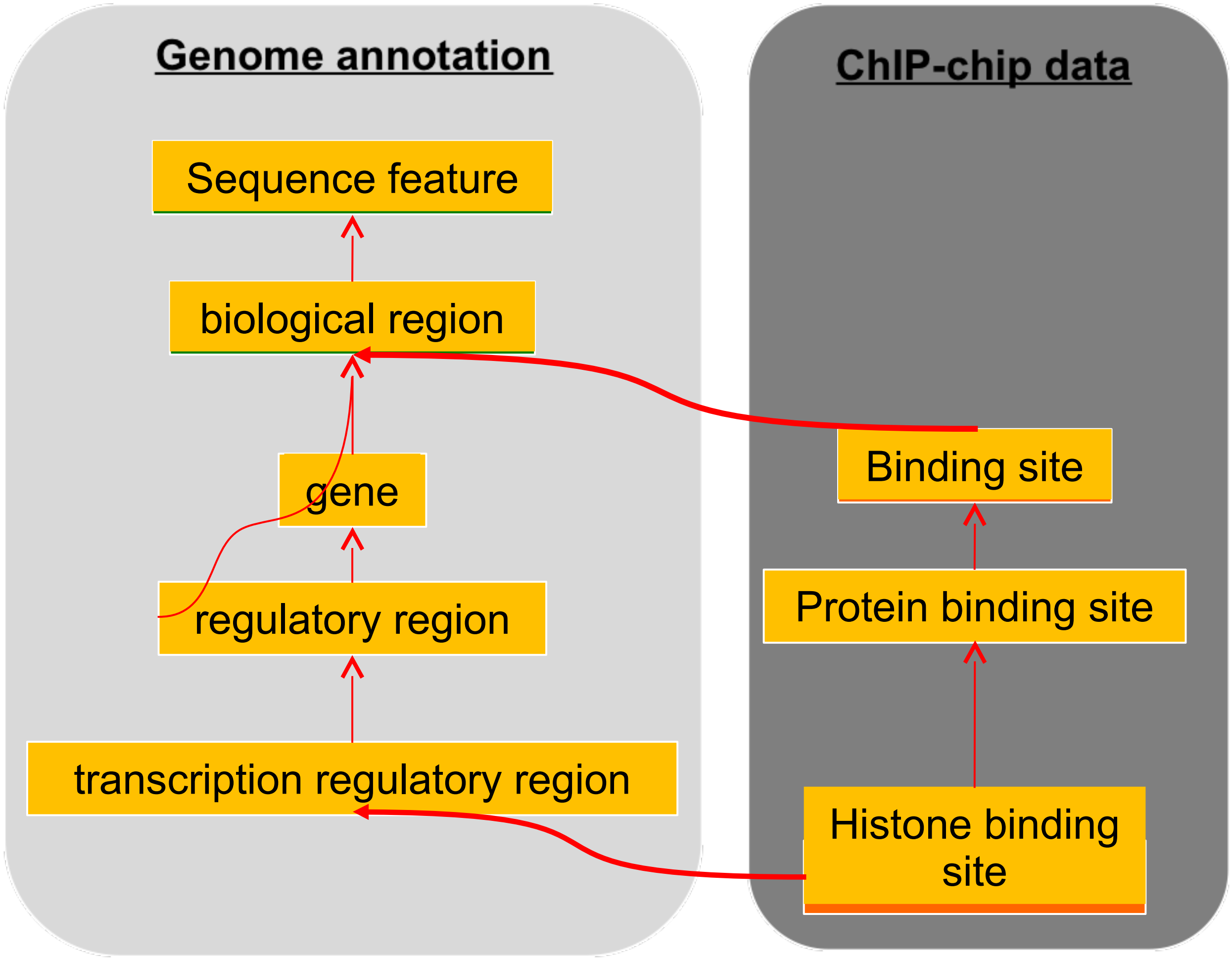
Proteomics

List of names, quantity

Metabolomics

Lab work yields different types of data files, with differing information content – yet common theme is a **name** or, failing that, an **originating location**

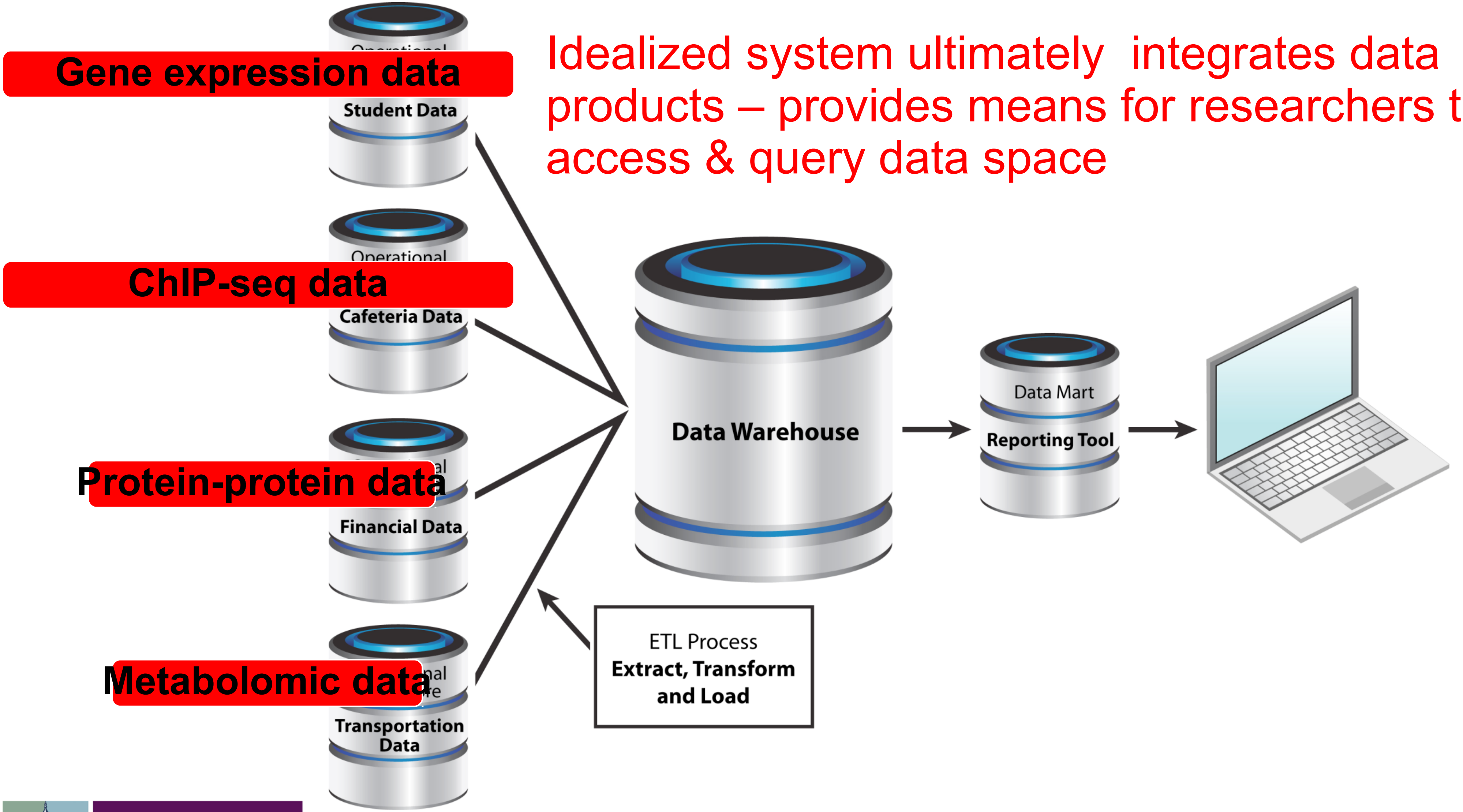
Sequence Ontology – e.g. mapping experimental results



NUI Galway
OÉ Gaillimh

InterMine

Conceptual Solution...



Idealized system ultimately integrates data products – provides means for researchers to access & query data space

NUI Galway
OÉ Gaillimh

InterMine

Introducing InterMine...

Open Source Data Warehouse

InterMine

github.com/intermine

InterMine has three parts:

1. Database
   a. Load data from different data sets into **a single database**
2. Webapp
   a. Mine the data
   b. Visualise!
3. Web services

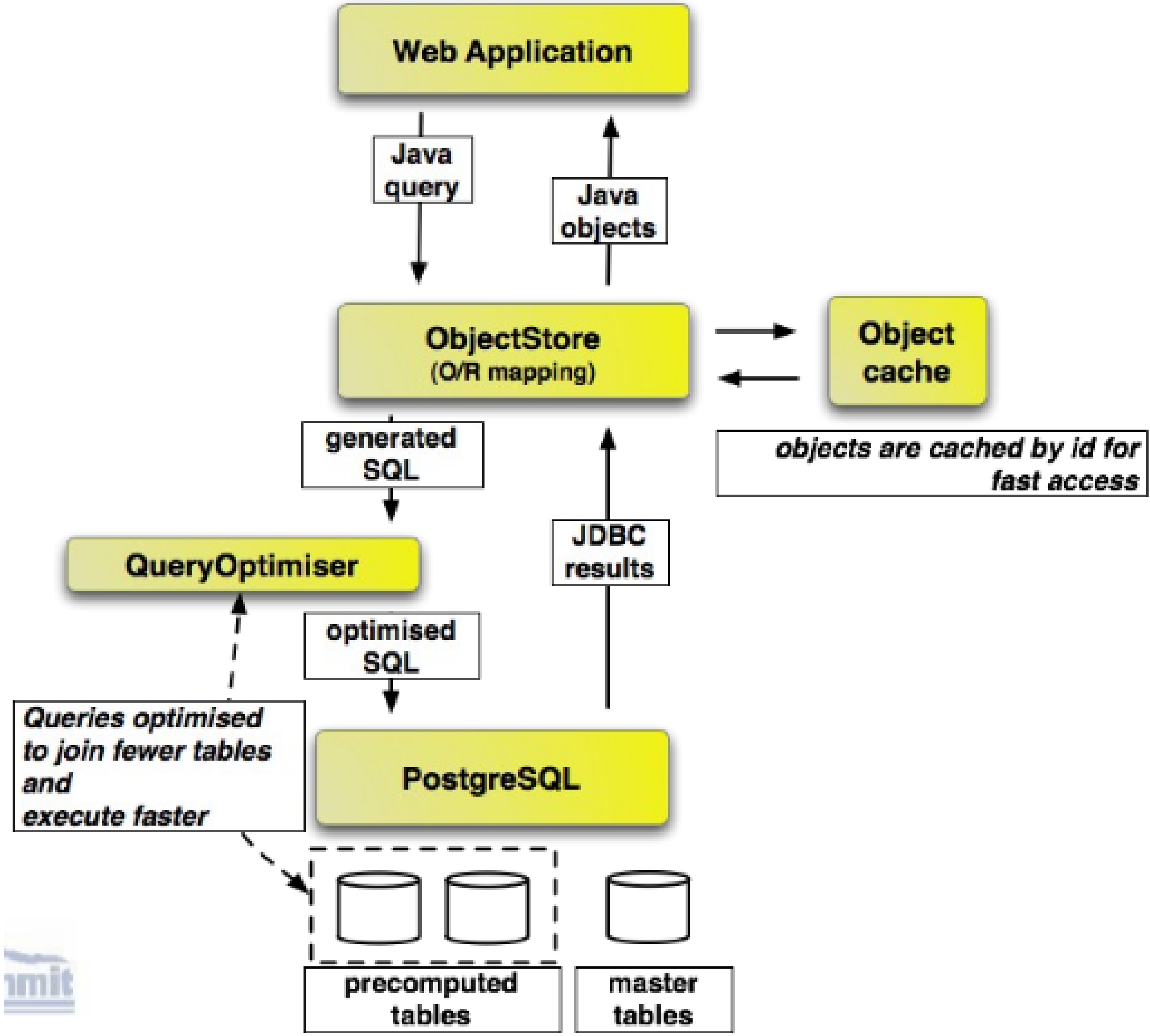
InterMine - Backend

Originally developed by Gos Micklem's group for FlyMine (2007)

*Uses read-only **ObjectStore** methodology to implement fast queries on linked dataspace using Seq. Ont. schema.*

Web-enabled access also via RESTful Perl, Python, Java, Ruby, R APIs

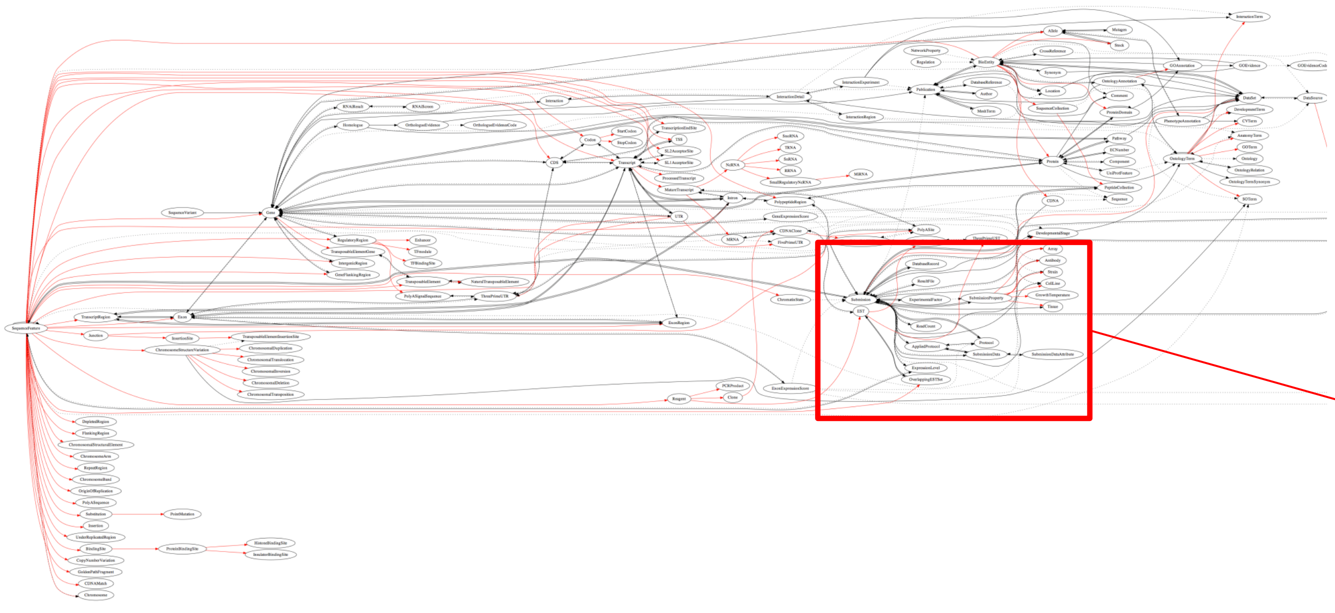
Java/PostgreSQL core code base, Tomcat for web services

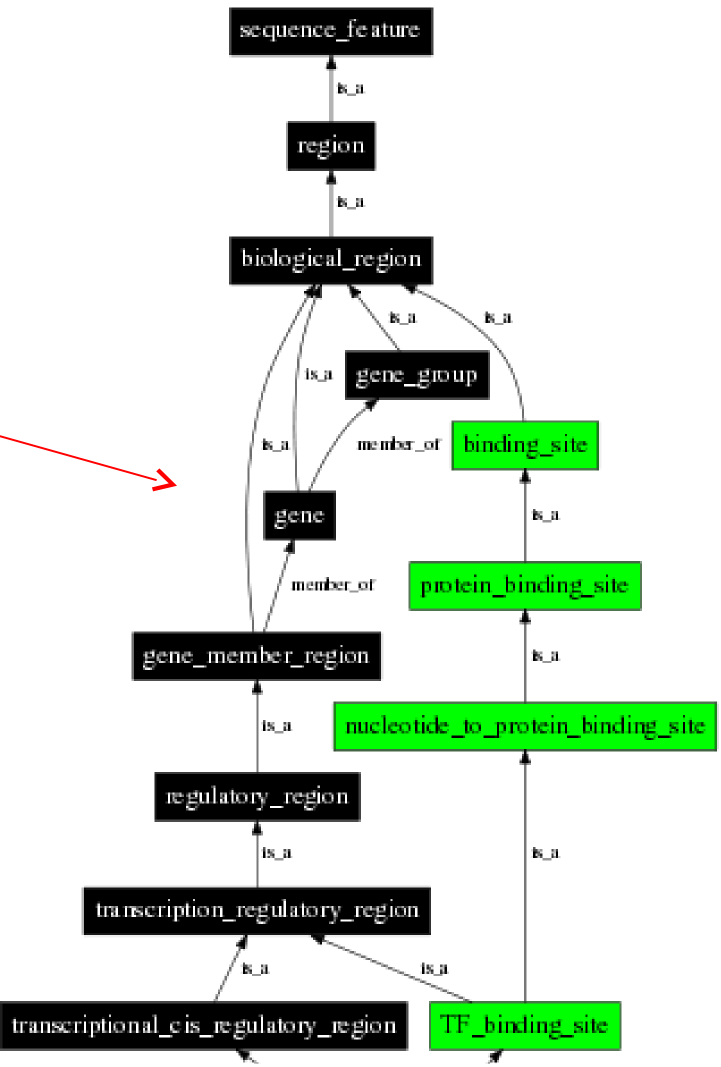


NUI Galway
OÉ Gaillimh

InterMine

Main step - Constructing & Loading a Data Model (using SO)

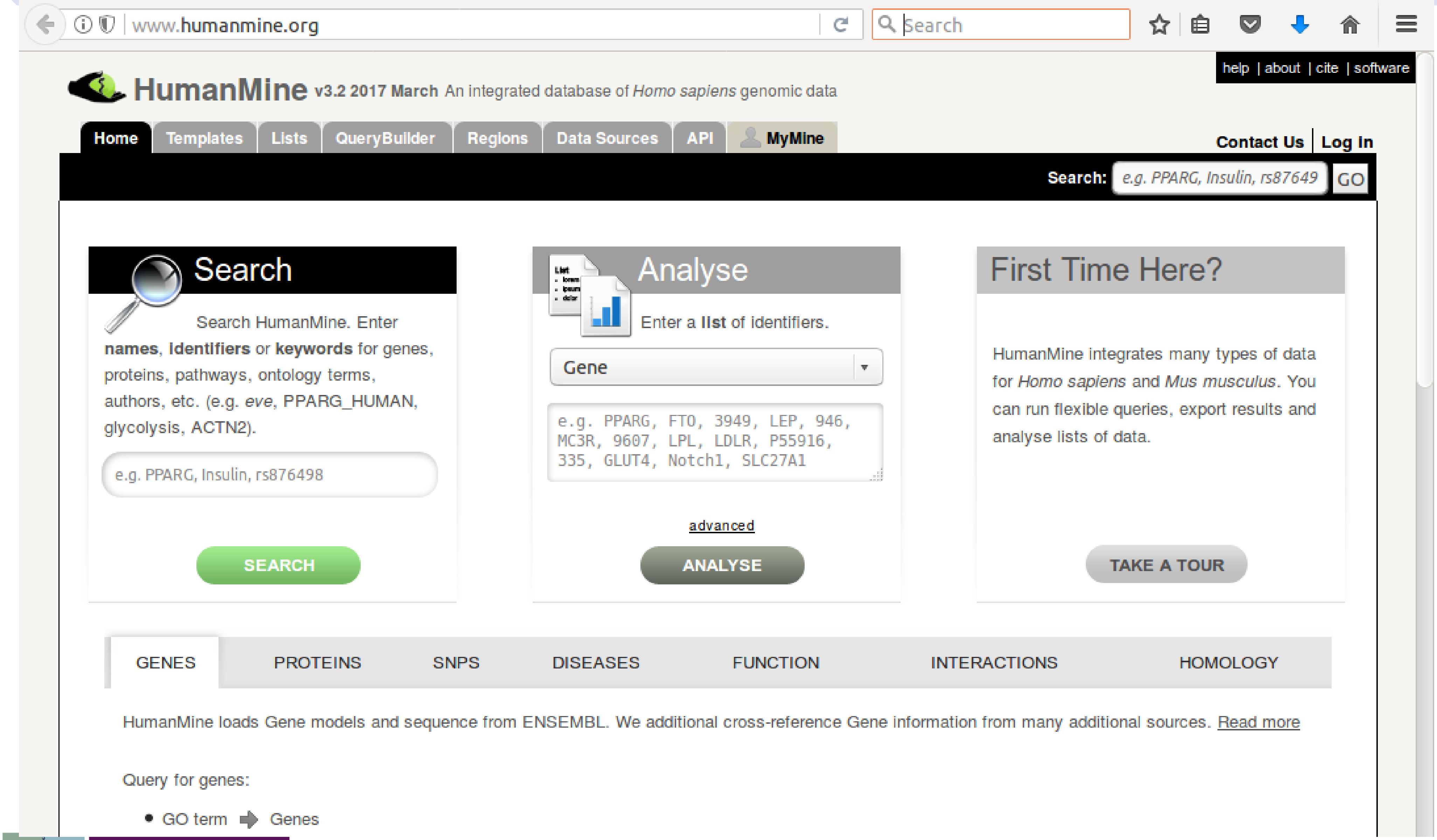




Data Model for ObjectStore based on

- experimental data space
- organism's genome(s)
- literature references
- published/archived 'omics data
- Data loaded from parsers to XML InterMine format

The InterMine `Science Gateway' web portal



www.humanmine.org
Search
help | about | cite | software

HumanMine v3.2 2017 March An integrated database of *Homo sapiens* genomic data

Home
Templates
Lists
QueryBuilder
Regions
Data Sources
API
MyMine
Contact Us
Log in
Search:
e.g. PPARG, Insulin, rs87649
GO

Search

Search HumanMine. Enter **names**, **identifiers** or **keywords** for genes, proteins, pathways, ontology terms, authors, etc. (e.g. *eve*, PPARG_HUMAN, glycolysis, ACTN2).

e.g. PPARG, Insulin, rs876498
SEARCH

Analyse

Enter a **list** of identifiers.

Gene
e.g. PPARG, FTO, 3949, LEP, 946, MC3R, 9607, LPL, LDLR, P55916, 335, GLUT4, Notch1, SLC27A1
advanced
ANALYSE

First Time Here?

HumanMine integrates many types of data for *Homo sapiens* and *Mus musculus*. You can run flexible queries, export results and analyse lists of data.

HumanMine loads Gene models and sequence from ENSEMBL. We additional cross-reference Gene information from many additional sources. Read more

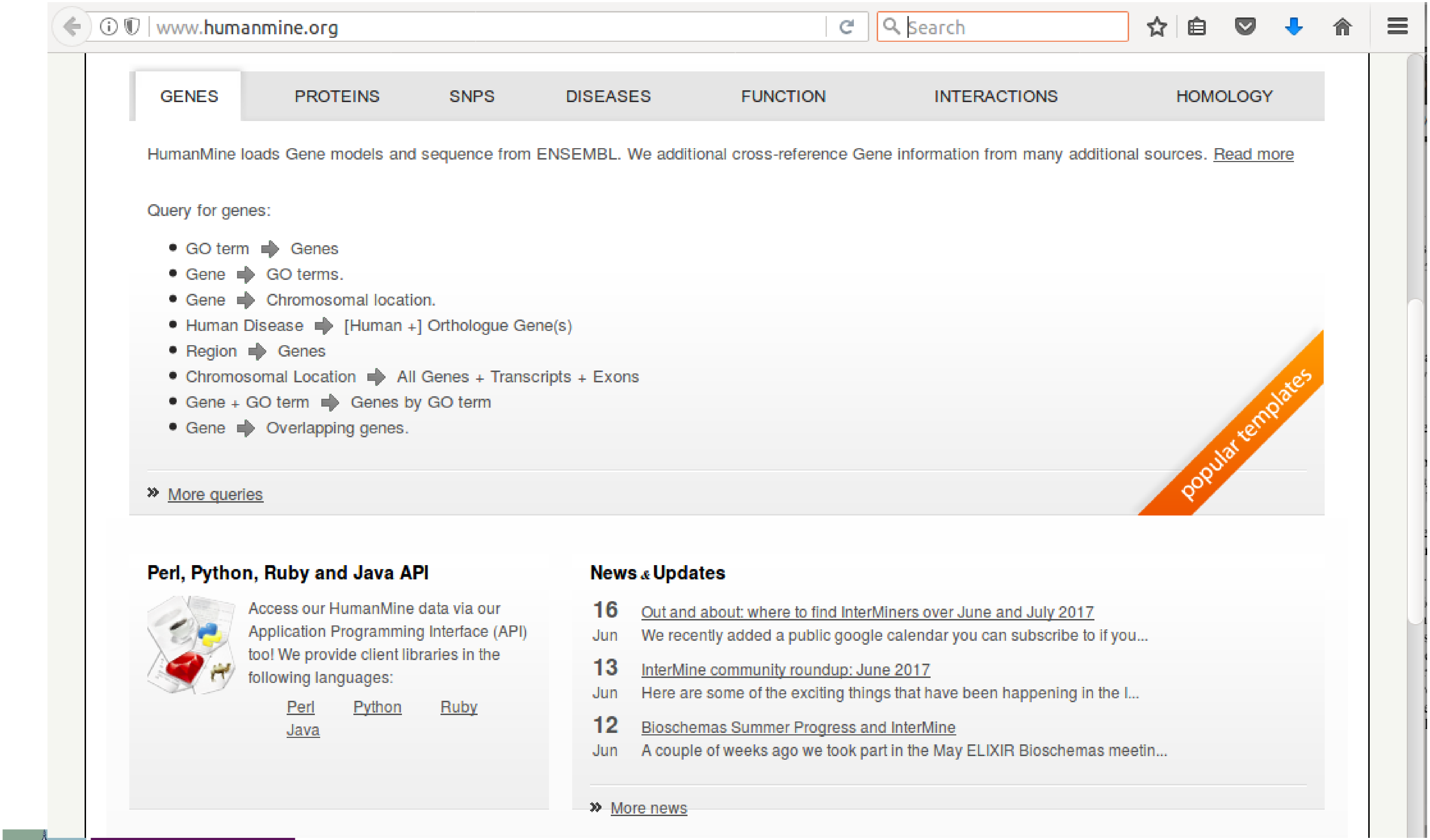
Query for genes:

- GO term ➡ Genes

NUI Galway
OÉ Gaillimh

InterMine

The InterMine `Science Gateway' web portal



HumanMine loads Gene models and sequence from ENSEMBL. We additional cross-reference Gene information from many additional sources. Read more

Query for genes:

- GO term ➡ Genes
- Gene ➡ GO terms.
- Gene ➡ Chromosomal location.
- Human Disease ➡ [Human +] Orthologue Gene(s)
- Region ➡ Genes
- Chromosomal Location ➡ All Genes + Transcripts + Exons
- Gene + GO term ➡ Genes by GO term
- Gene ➡ Overlapping genes.

» More queries

Perl, Python, Ruby and Java API

Access our HumanMine data via our Application Programming Interface (API) too! We provide client libraries in the following languages:

Perl Python Ruby
Java

News & Updates

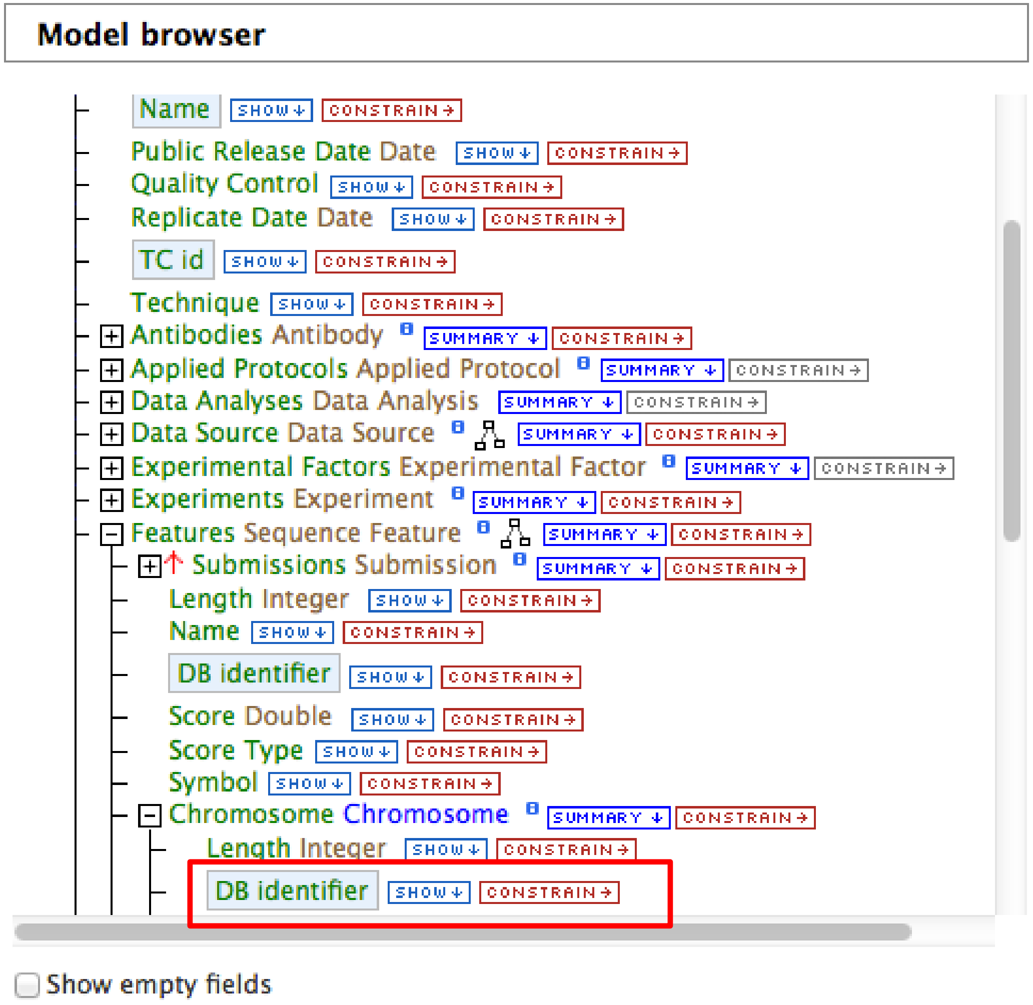
16 Jun Out and about: where to find InterMiners over June and July 2017
We recently added a public google calendar you can subscribe to if you...

13 Jun InterMine community roundup: June 2017
Here are some of the exciting things that have been happening in the I...

12 Jun Bioschemas Summer Progress and InterMine
A couple of weeks ago we took part in the May ELIXIR Bioschemas meetin...

» More news

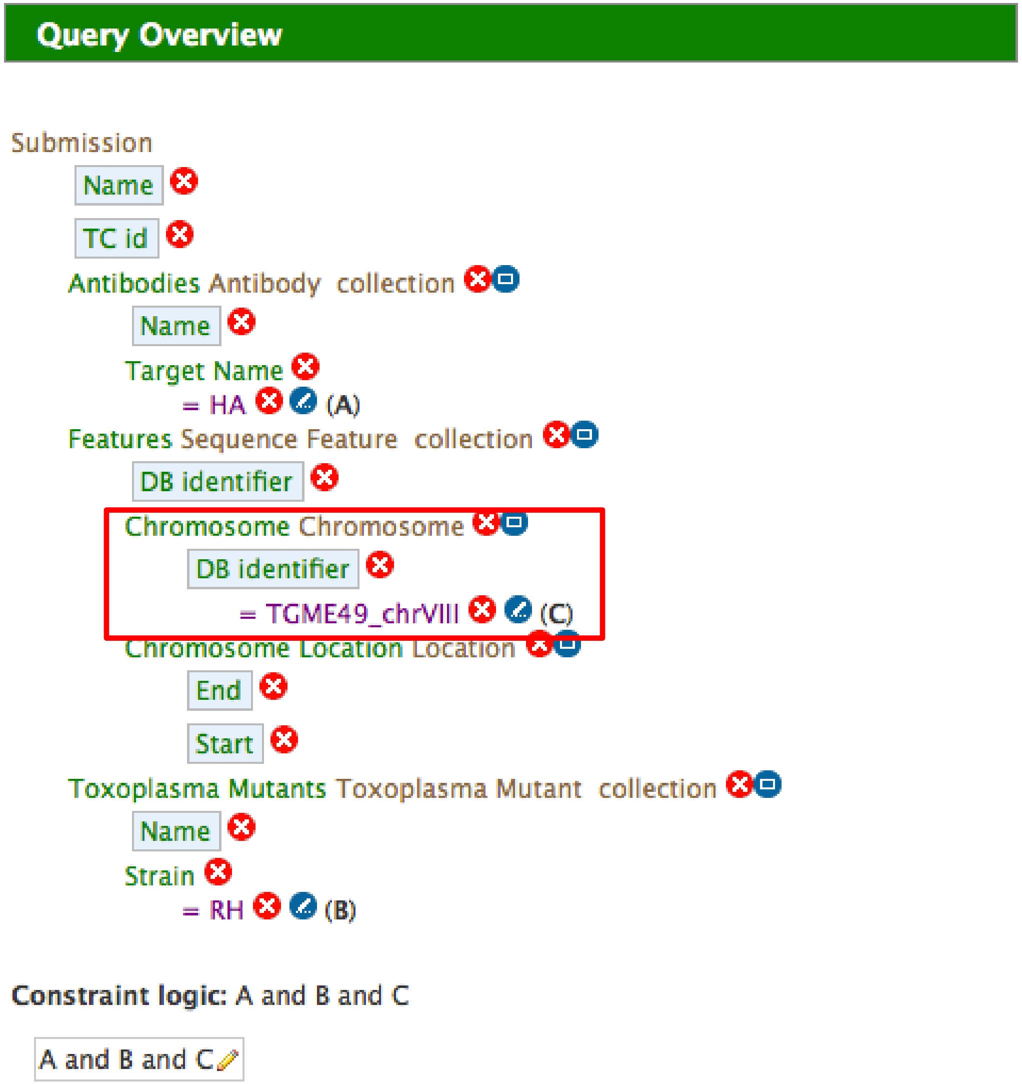
Using the QueryBuilder tool to explore dataspace



Model browser

- Name SHOW ↓ CONSTRAIN →
- Public Release Date Date SHOW ↓ CONSTRAIN →
- Quality Control SHOW ↓ CONSTRAIN →
- Replicate Date Date SHOW ↓ CONSTRAIN →
- TC id SHOW ↓ CONSTRAIN →
- Technique SHOW ↓ CONSTRAIN →
- ⊞ Antibodies Antibody SUMMARY ↓ CONSTRAIN →
- ⊞ Applied Protocols Applied Protocol SUMMARY ↓ CONSTRAIN →
- ⊞ Data Analyses Data Analysis SUMMARY ↓ CONSTRAIN →
- ⊞ Data Source Data Source SUMMARY ↓ CONSTRAIN →
- ⊞ Experimental Factors Experimental Factor SUMMARY ↓ CONSTRAIN →
- ⊞ Experiments Experiment SUMMARY ↓ CONSTRAIN →
- ⊟ Features Sequence Feature SUMMARY ↓ CONSTRAIN →
  - ⊞ ↑ Submissions Submission SUMMARY ↓ CONSTRAIN →
  - Length Integer SHOW ↓ CONSTRAIN →
  - Name SHOW ↓ CONSTRAIN →
  - DB identifier SHOW ↓ CONSTRAIN →
  - Score Double SHOW ↓ CONSTRAIN →
  - Score Type SHOW ↓ CONSTRAIN →
  - Symbol SHOW ↓ CONSTRAIN →
  - ⊟ Chromosome Chromosome SUMMARY ↓ CONSTRAIN →
    - Length Integer SHOW ↓ CONSTRAIN →
    - DB identifier SHOW ↓ CONSTRAIN →

☐ Show empty fields



Query Overview

- Submission
  - Name
  - TC id
  - Antibodies Antibody collection
    - Name
    - Target Name
      - = HA (A)
  - Features Sequence Feature collection
    - DB identifier
    - Chromosome Chromosome
      - DB identifier
        - = TGME49_chrVIII (C)
    - Chromosome Location Location
      - End
      - Start
  - Toxoplasma Mutants Toxoplasma Mutant collection
    - Name
    - Strain
      - = RH (B)

Constraint logic: A and B and C

A and B and C

Using Web Services remotely...

List of Identifiers → List of database records → Homologues in X → Their protein domains

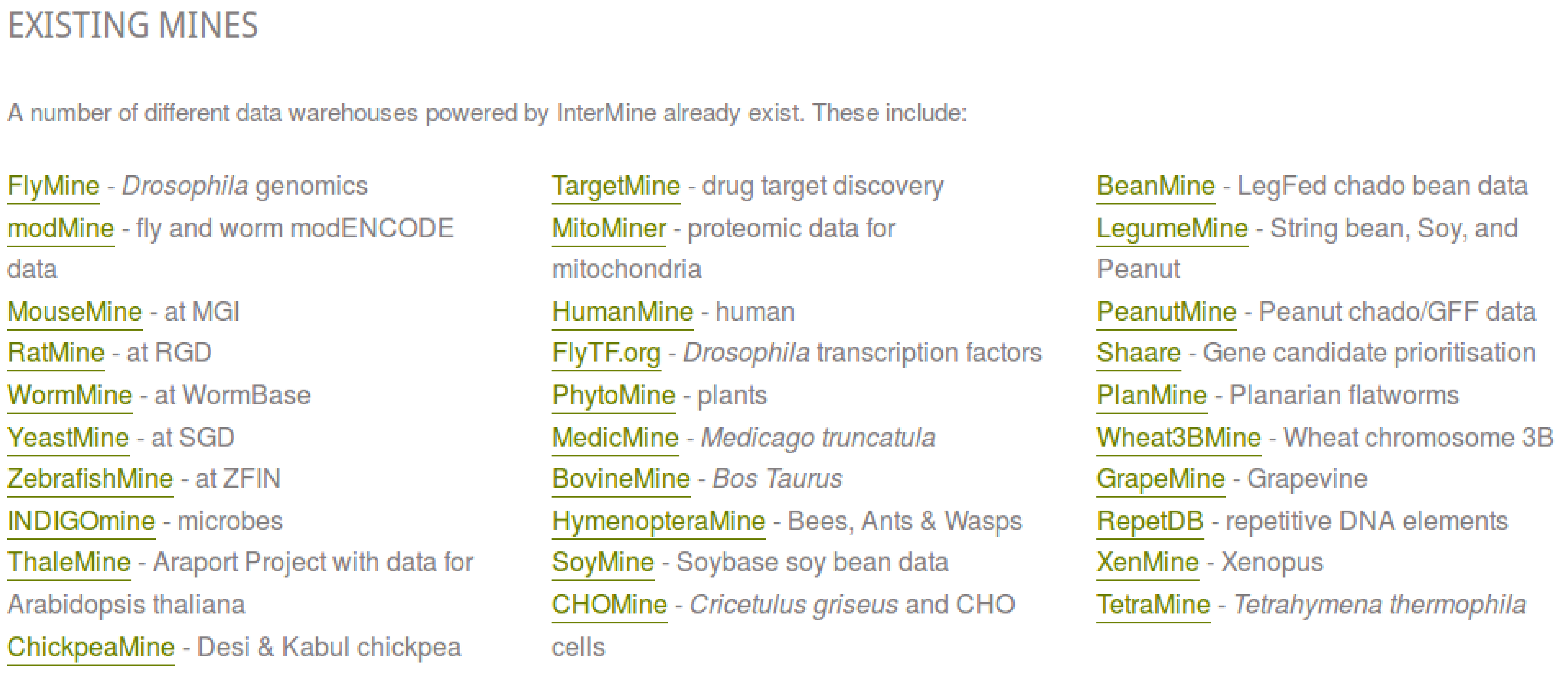
```
use InterMine::Webservice;

my $flymine = Webservice::InterMine->get_service('www.flymine.org/query', 'TOKEN');
my $list = $flymine->new_list(type => 'Gene', content => “some/file.txt”);

my $query = $list->to_query;
$query->add_views('homologues.homologue.symbol');
$query->add_constraint('homologues.homologue.organism.shortName', '=', 'R norvegicus');
my $rat_symbols = [map {$_->{homologues.homologue.symbol}} $query->results(''hashrefs”)];

my $ratmine = Webservice::InterMine->get_service('ratmine.mcw.edu/ratmine');
my $rat_query = $ratmine->new_query(class => 'Gene');
$rat_query->add_views(qw/symbol primaryIdentifier proteins.proteinDomains.name/);
$rat_query->add_constraint('symbol', 'ONE OF', $rat_symbols);
$rat_query->show;
```

InterMines, Intermines...



EXISTING MINES

A number of different data warehouses powered by InterMine already exist. These include:

FlyMine - *Drosophila* genomics
modMine - fly and worm modENCODE data
MouseMine - at MGI
RatMine - at RGD
WormMine - at WormBase
YeastMine - at SGD
ZebrafishMine - at ZFIN
INDIGOmine - microbes
ThaleMine - Araport Project with data for Arabidopsis thaliana
ChickpeaMine - Desi & Kabul chickpea

TargetMine - drug target discovery
MitoMiner - proteomic data for mitochondria
HumanMine - human
FlyTF.org - *Drosophila* transcription factors
PhytoMine - plants
MedicMine - *Medicago truncatula*
BovineMine - *Bos Taurus*
HymenopteraMine - Bees, Ants & Wasps
SoyMine - Soybase soy bean data
CHOMine - *Cricetulus griseus* and CHO cells

BeanMine - LegFed chado bean data
LegumeMine - String bean, Soy, and Peanut
PeanutMine - Peanut chado/GFF data
Shaare - Gene candidate prioritisation
PlanMine - Planarian flatworms
Wheat3BMine - Wheat chromosome 3B
GrapeMine - Grapevine
RepetDB - repetitive DNA elements
XenMine - Xenopus
TetraMine - *Tetrahymena thermophila*

NUI Galway
OÉ Gaillimh

InterMine

What about the DCI thing?

InterMines' tend to be standalone `silos' based one one model organism/biological process.

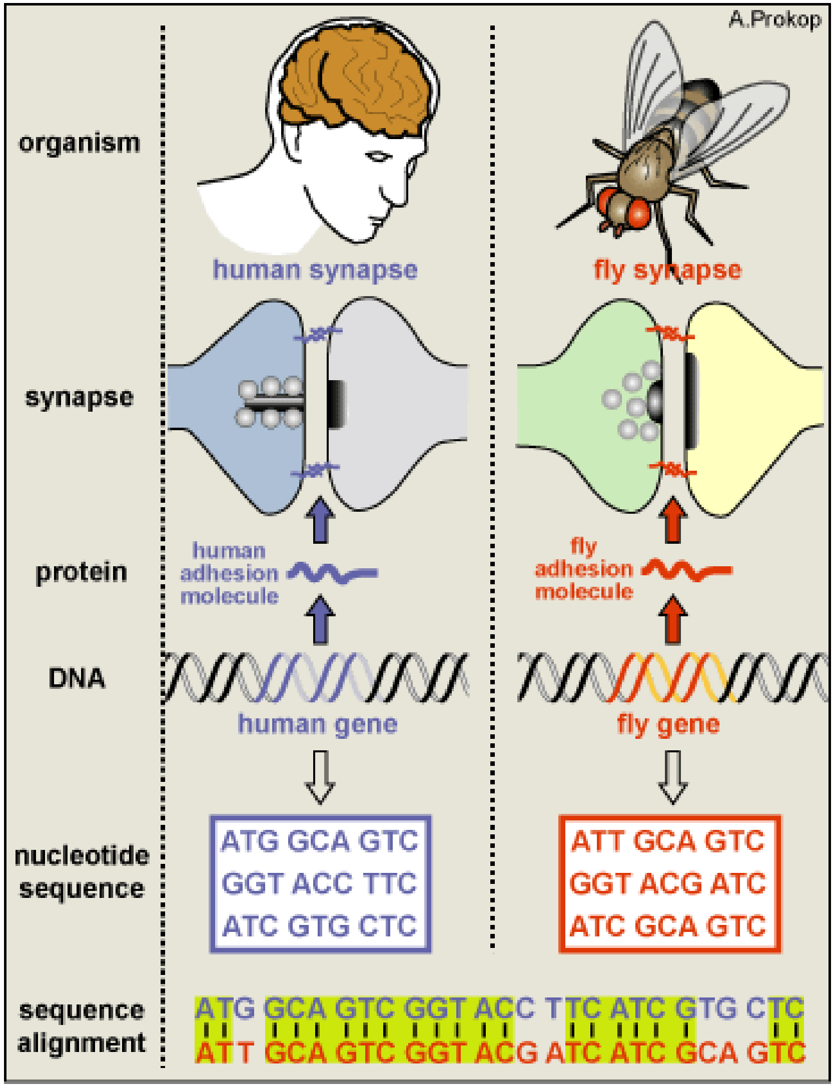
ToxoMine:

DB server (8 core, 32GB RAM, 2 TB)

Web servers (4 core, 16GB, 256 GB)

However – *fundamental homology between all living organisms enables interoperability 'for free' through the Sequence Ontology*.

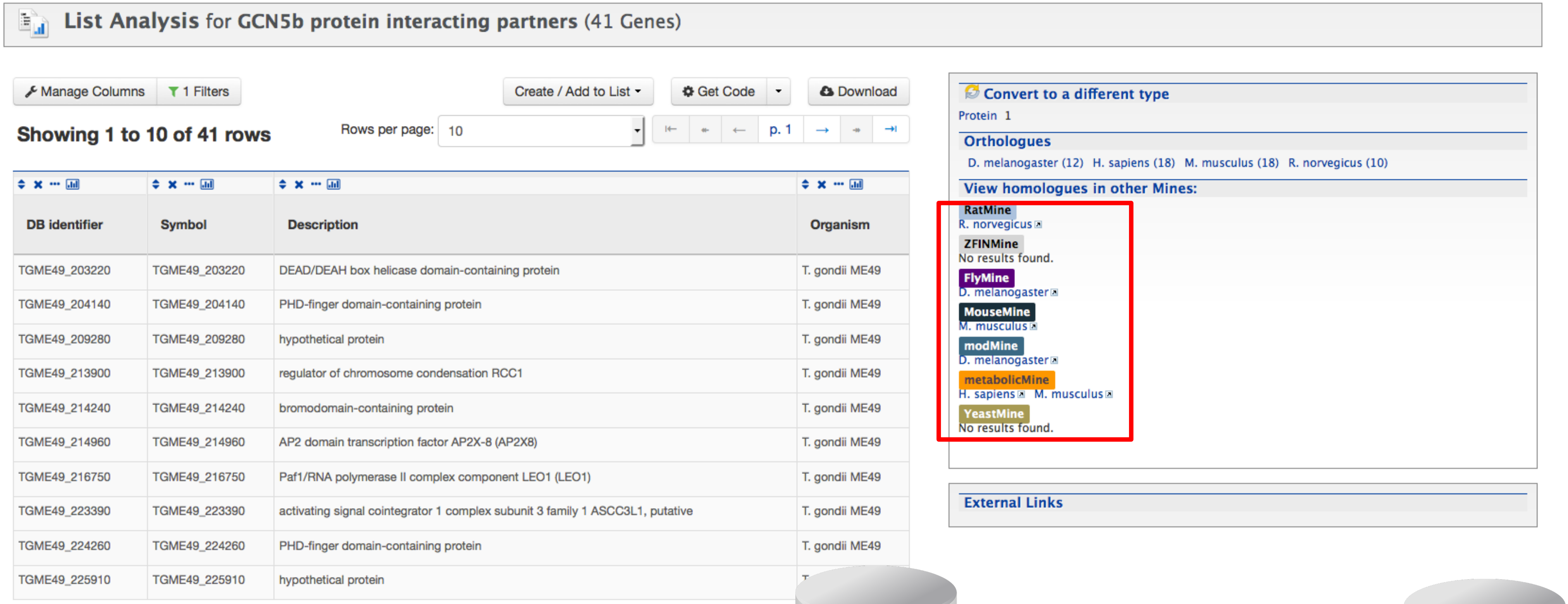
Offers a Virtual Observatory like experience across registered Mines.



NUI Galway
OÉ Gaillimh

InterMine

Cross-species query can be performed using the InterMOD infrastructure



List Analysis for **GCN5b protein interacting partners** (41 Genes)

Manage Columns | 1 Filters | Create / Add to List | Get Code | Download

Showing 1 to 10 of 41 rows

Rows per page: 10 | p. 1

DB identifier	Symbol	Description	Organism
TGME49_203220	TGME49_203220	DEAD/DEAH box helicase domain-containing protein	T. gondii ME49
TGME49_204140	TGME49_204140	PHD-finger domain-containing protein	T. gondii ME49
TGME49_209280	TGME49_209280	hypothetical protein	T. gondii ME49
TGME49_213900	TGME49_213900	regulator of chromosome condensation RCC1	T. gondii ME49
TGME49_214240	TGME49_214240	bromodomain-containing protein	T. gondii ME49
TGME49_214960	TGME49_214960	AP2 domain transcription factor AP2X-8 (AP2X8)	T. gondii ME49
TGME49_216750	TGME49_216750	Paf1/RNA polymerase II complex component LEO1 (LEO1)	T. gondii ME49
TGME49_223390	TGME49_223390	activating signal cointegrator 1 complex subunit 3 family 1 ASCC3L1, putative	T. gondii ME49
TGME49_224260	TGME49_224260	PHD-finger domain-containing protein	T. gondii ME49
TGME49_225910	TGME49_225910	hypothetical protein	T

Convert to a different type

Protein 1

Orthologues

D. melanogaster (12) H. sapiens (18) M. musculus (18) R. norvegicus (10)

View homologues in other Mines:

RatMine
R. norvegicus

ZFINMine
No results found.

FlyMine
D. melanogaster

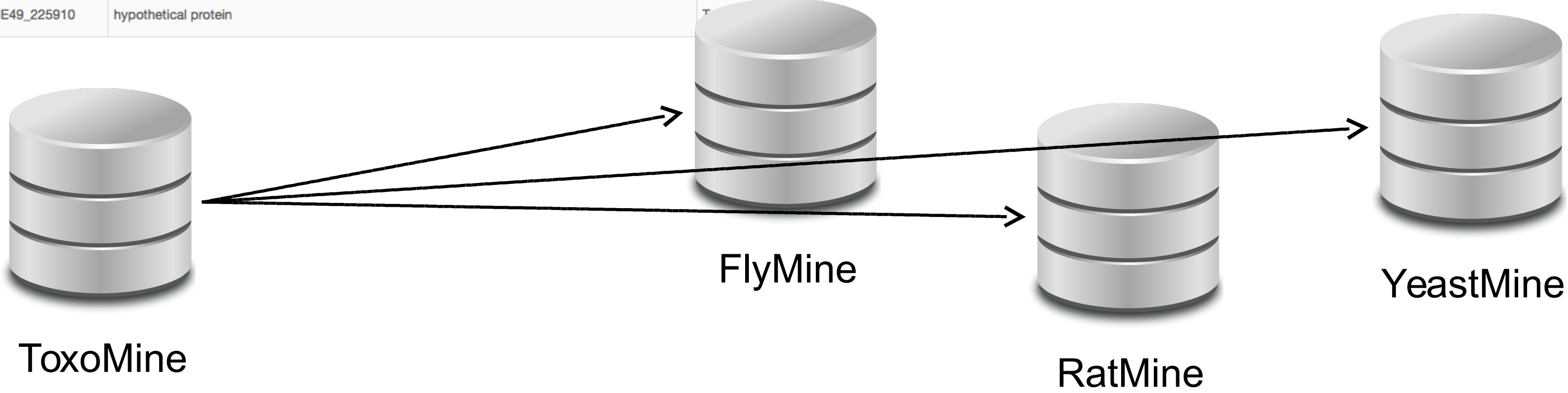
MouseMine
M. musculus

modMine
D. melanogaster

metabolicMine
H. sapiens M. musculus

YeastMine
No results found.

External Links



NUI Galway
OÉ Gaillimh

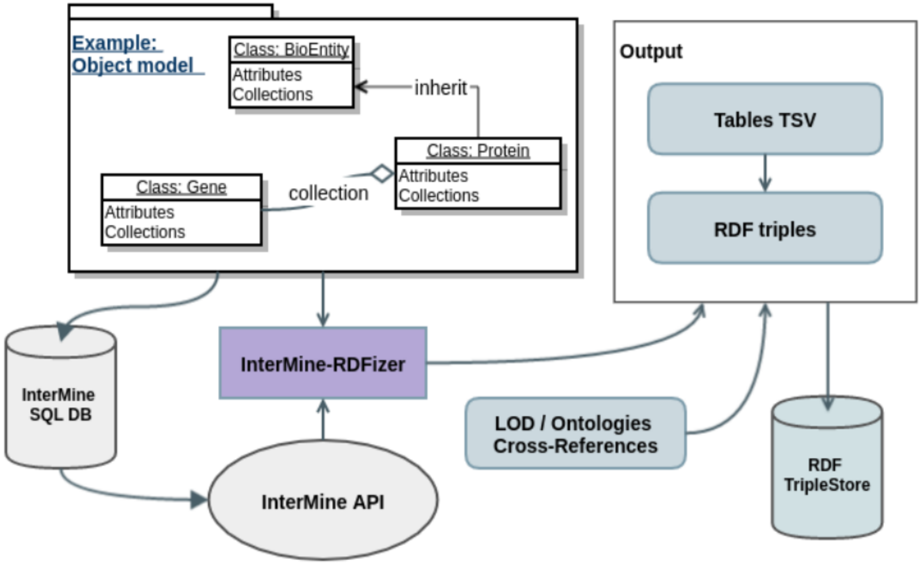
InterMine

Ongoing InterMine Community Activity

RDFization of Intermine

Create MO-LD (Model Organism Linked Data)

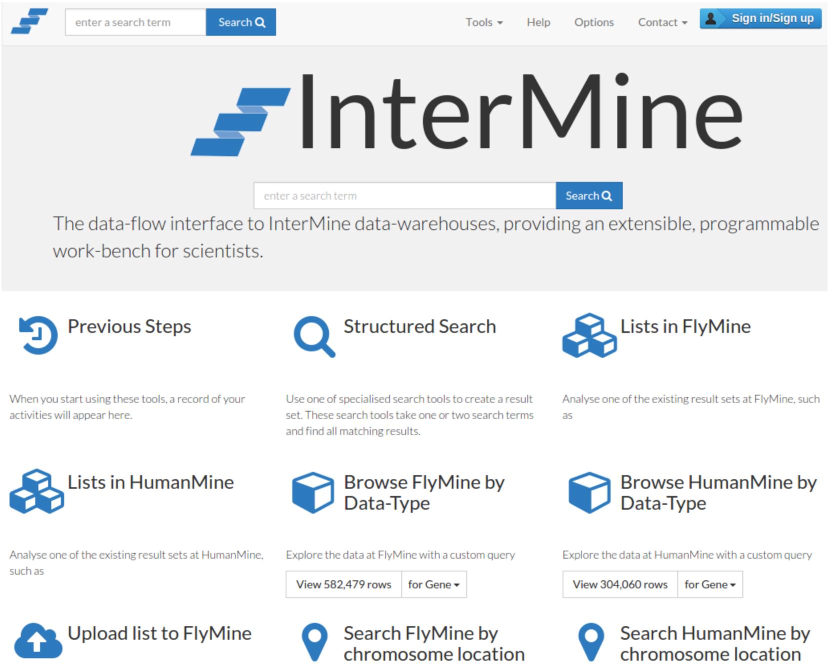
- improve interoperability
- enable federated queries across larger RDF space
- better packaging for local/cloud deployment



Enchanced Webservice functionality

Evolve infrastructure to current state-of-the-art

- move away from Java/Tomcat paradigm
- modularity to integrate third-party web apps
- Clojure, Angular, PostgreSQL, Web Services



The data-flow interface to InterMine data-warehouses, providing an extensible, programmable work-bench for scientists.

Previous Steps

When you start using these tools, a record of your activities will appear here.

Structured Search

Use one of specialised search tools to create a result set. These search tools take one or two search terms and find all matching results.

Lists in FlyMine

Analyse one of the existing result sets at FlyMine, such as

Lists in HumanMine

Analyse one of the existing result sets at HumanMine, such as

Browse FlyMine by Data-Type

Explore the data at FlyMine with a custom query

View 582,479 rows | for Gene

Browse HumanMine by Data-Type

Explore the data at HumanMine with a custom query

View 304,060 rows | for Gene

Upload list to FlyMine

Search FlyMine by chromosome location

Search HumanMine by chromosome location

NUI Galway
OÉ Gaillimh

InterMine

Acknowledgements

In Galway

David Rhee

Pilib Ó Broin

Kevin Shieh

Gos Micklem's Group (Cambridge)

Yo Yehuda

Julie Sullivan

Xiaonan Wang

Kami Kim's Lab (Albert Einstein College of Medicine)

Matthew Croken

Sheila C Nardelli

Ravikant Ranjan

Natalie Silmon De Monerri

Zoi Tampaki

Li-Min Ting

Netha Ulahannan

Funding

National Institute of Allergy and Infectious Diseases (#RC4AI092801)
Wellcome Trust (#099133)
National Human Genome Research Institute (#R01HG004834)

And the National Science Foundation!

NUI Galway
OÉ Gaillimh

InterMine