Seven Years of Biomedical Ontology in Production Service in Ten Minutes

Simon Jupp, Tony Burdett, James Malone

malone@ebi.ac.uk

@jamesmalone

EMBL-EBI

Data we work with: experimental variables

ArrayExpress

Gene Expression atlas

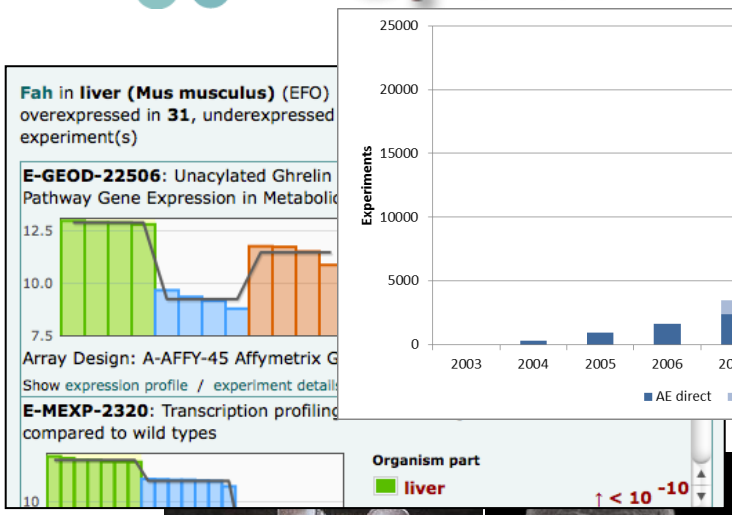
UniProt

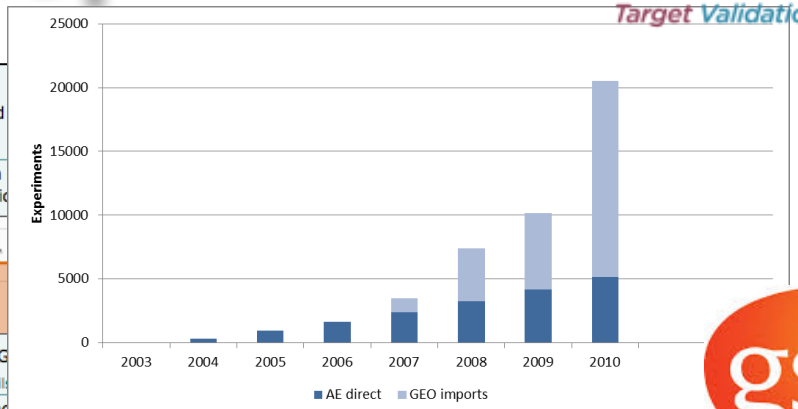
Ensembl

CTTV

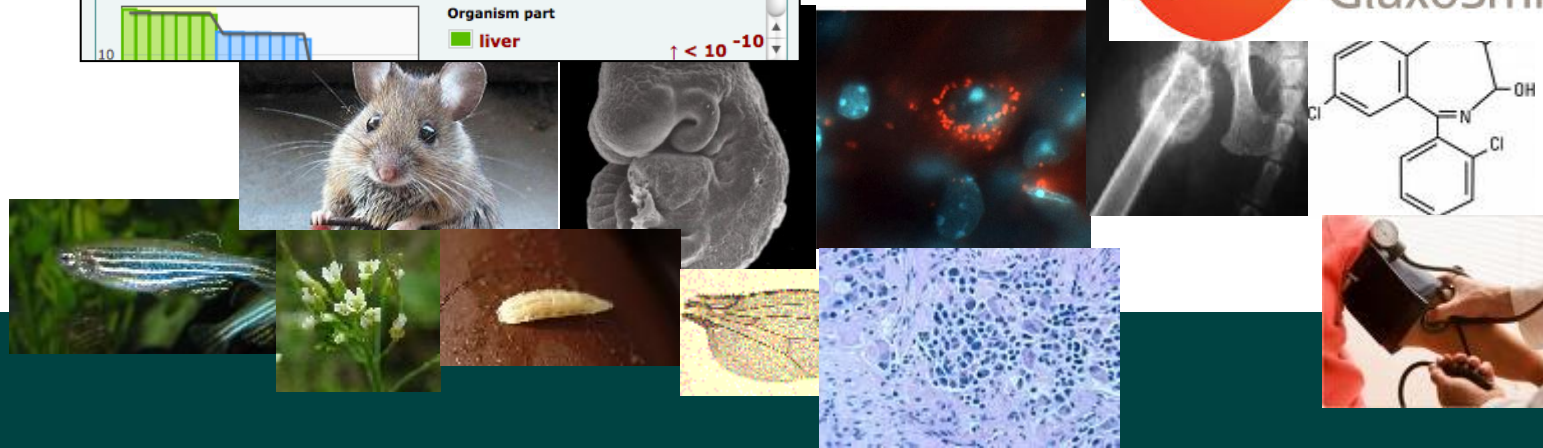
Centre for Therapeutic Target Validation

ENCODE





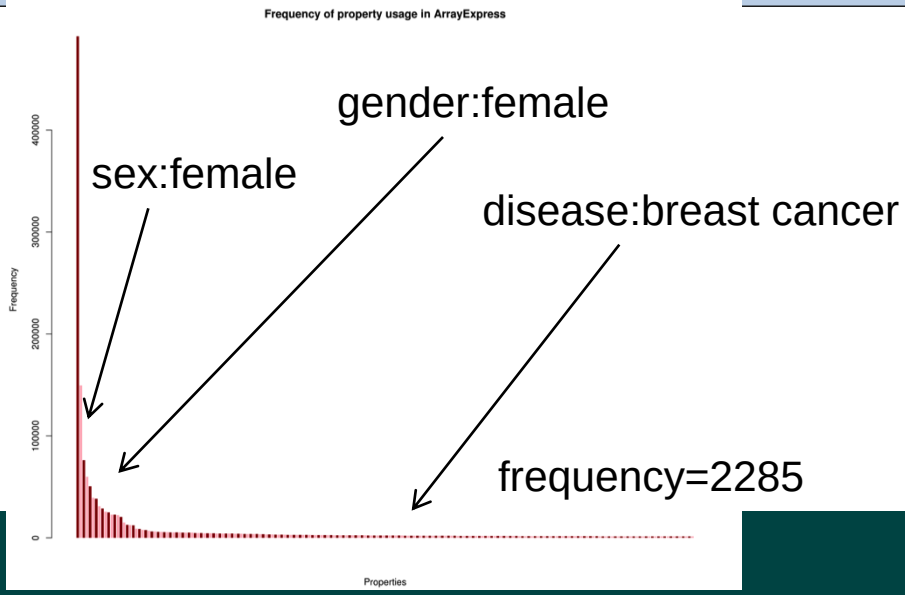
gsk GlaxoSmithKline



-EBI

Lots of stuff, lots of types of stuff

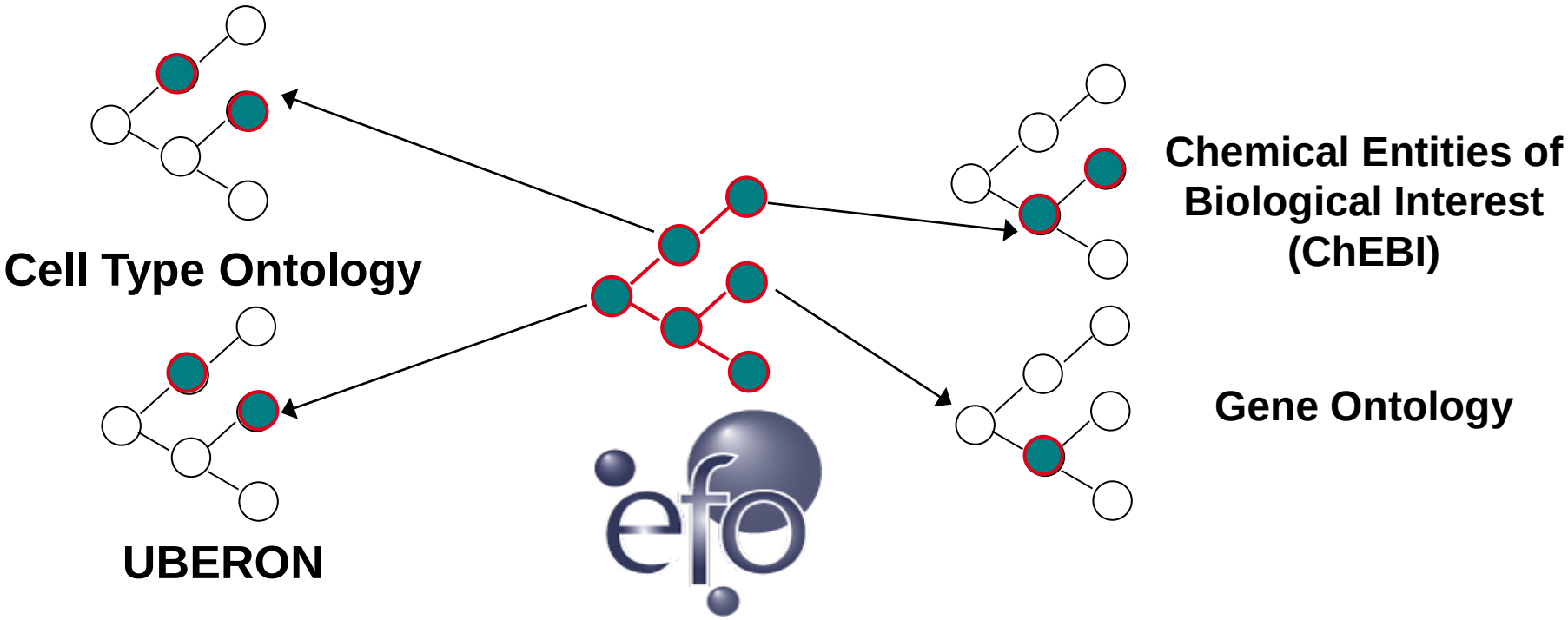
Annotations	Gene Expression
Species	~1821
Samples	1,126,457
Annotations on samples	7,672,825
Unique sample annotations	**243,650**
Assays (Hybridizations)	965,638
Annotations on assays	3,248,298
Unique assay annotations	**189,381**



This is top 200. There are nearly 250,000 in total and half are used only once

Ontology www.ebi.ac.uk/efo

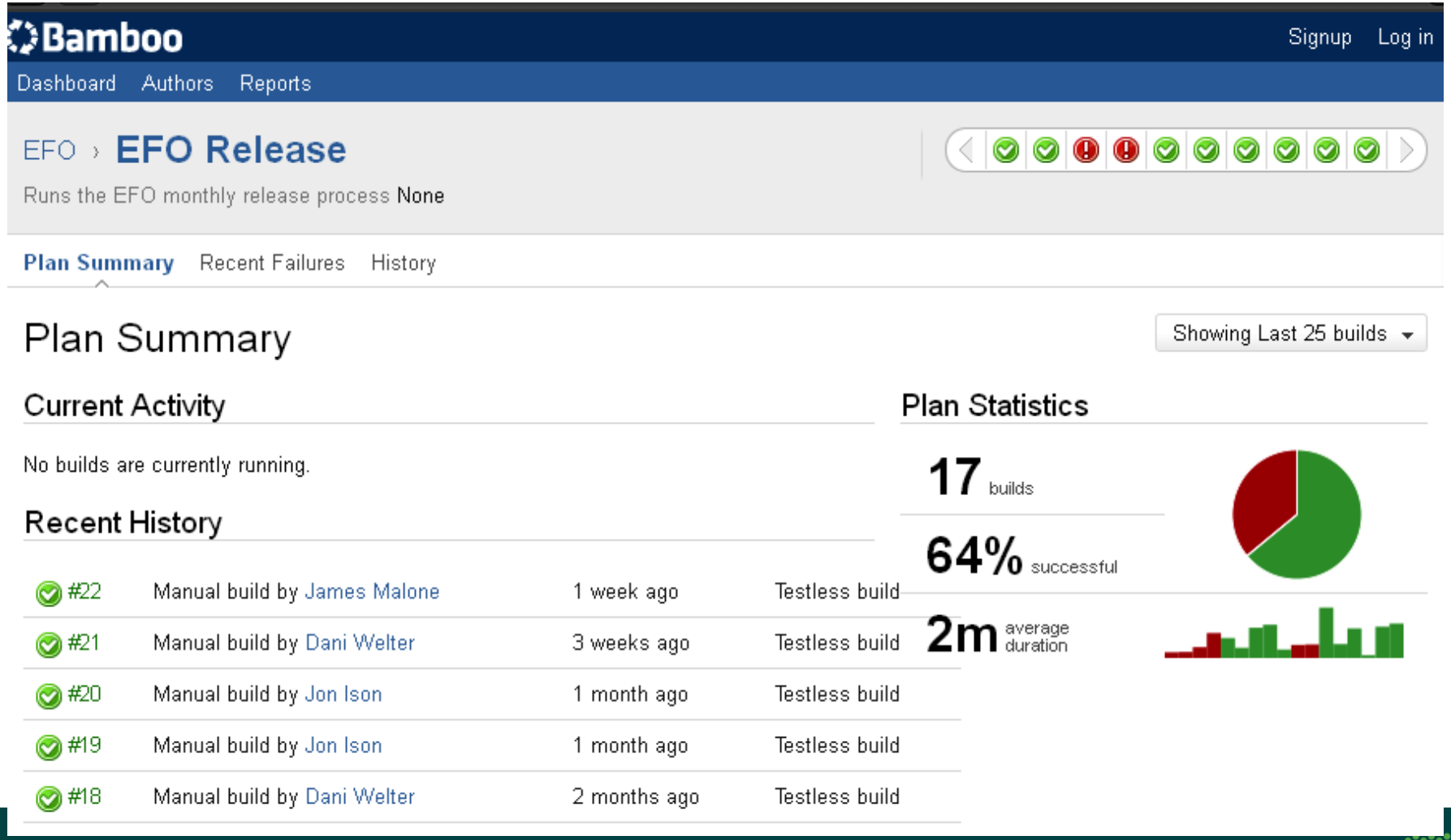
- We build the Experimental Factor Ontology in OWL
- EFO is an application ontology, built for use in production services
- Lots of external dependencies - reuses where possible and appropriate other references



Ontology challenges

- Managing external dependencies
- Expressivity trade-offs
  - Working in different OWL profiles
- Managing ontology views
  - Application specific slices
- Engaging our curators and domain experts
  - Aligning with production release cycles
  - Agile development
- Continuous integration and testing environments

Ontology as Software – continuous integration



Testing in Bamboo build

EFO › **EFO Check**

Checks and validates each EFO commit None

Plan Summary | Recent Failures | History

Plan Summary

Showing Last 25 builds

Current Activity

No builds are currently running.

Recent History

#34	Updated by James Malone	6 days ago	Testless build
#33	Updated by James Malone	6 days ago	Testless build
#32	Updated by James Malone	1 week ago	Testless build
#31	Updated by James Malone	1 week ago	Testless build
#30	Updated by James Malone	1 week ago	Testless build
#29	Updated by Dani Welter	3 weeks ago	Testless build
#28	Updated by Dani Welter	3 weeks ago	Testless build
#27	Updated by James Malone	1 month ago	Testless build
#26	Updated by James Malone	1 month ago	Testless build

Plan Statistics

25 builds

72% successful

30s average duration

Testing in Bamboo build

EFO › EFO Check › **#26**

Checks and validates each EFO commit

#26 failed

Build Summary | Tests | Changes | Artifacts | **Logs** | Metadata

Logs

The following logs have been generated by the Jobs in this Plan.

Expand All | Collaps

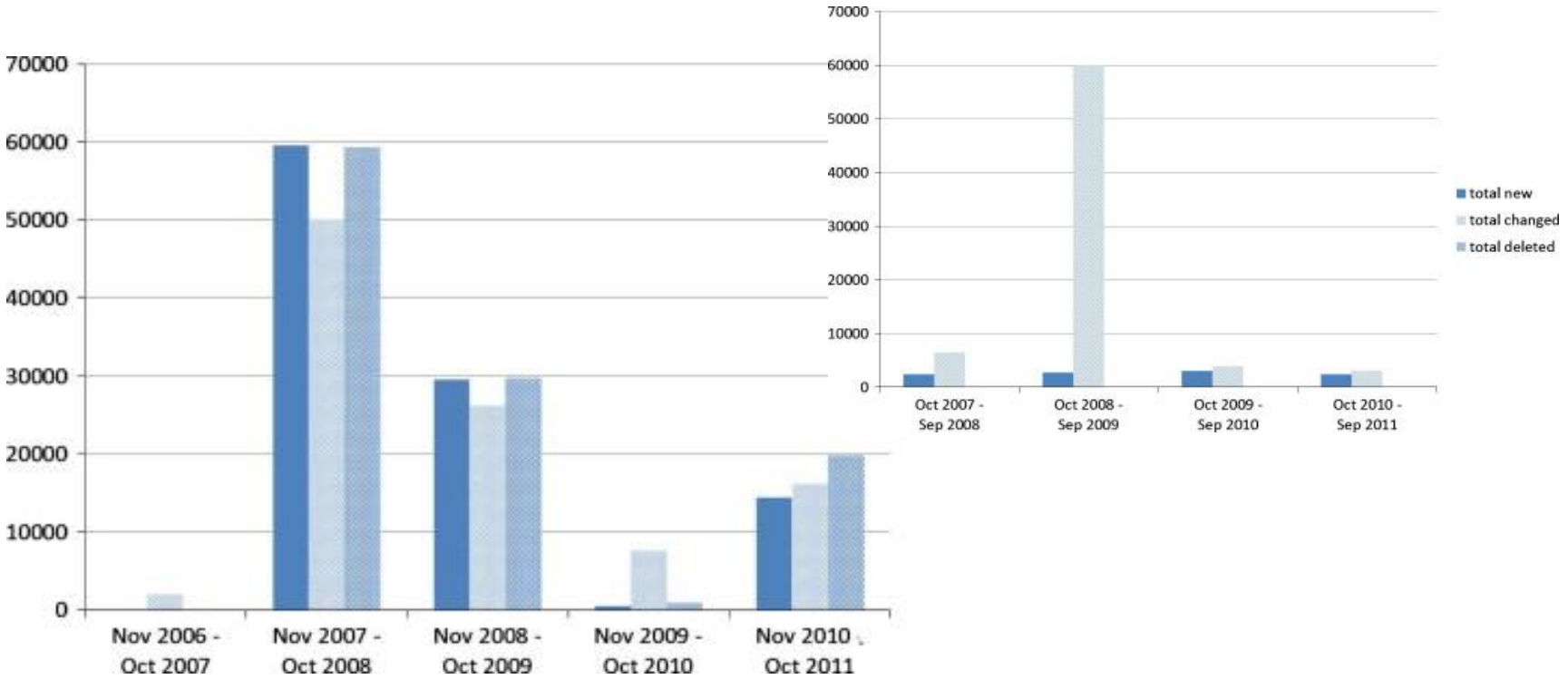
Job	Logs
Check EFO Default Stage	Download

```
20-Sep-2012 17:21:56 # Report of validation of EFO generated by efovalidator
20-Sep-2012 17:21:56 #   log file  : /ebi/microarray/home/fgpt/sw/efovalidator/validation.log
20-Sep-2012 17:21:56 #   EFO input : efo_release_candidate.owl
20-Sep-2012 17:21:56 #   generated : 20 September 2012 @ 05:21:56 PM
20-Sep-2012 17:21:56 #
20-Sep-2012 17:21:56 # For more information, see:
20-Sep-2012 17:21:56 #   http://www.ebi.ac.uk/fgpt/sw/efoimporter/index.html
20-Sep-2012 17:21:56 #
20-Sep-2012 17:22:01 # Summary
20-Sep-2012 17:22:01 # Critical errors
20-Sep-2012 17:22:01 # 1)    Check input is in valid OWL format
20-Sep-2012 17:22:01 # 2)    Check version number
20-Sep-2012 17:22:01 # 3)    Check date
20-Sep-2012 17:22:01 # 4)    Check class IDs are URIs formed from a valid namespace name
20-Sep-2012 17:22:01 # 5)    Check EFO IDs used in class IRIs are valid
20-Sep-2012 17:22:01 # 6)    Check every class has a single, unique label
```

Managing external dependencies

Activity Study PMID:22554701

- Lots of activity, yes, but lots of deletions
- Question: is this a stable resource?



Ontology diffs

Ontology 1: "http://efo.svn.sourceforge.net/viewvc/efo/trunk/src/efoinowl/efo.owl?revision=207"
Ontology 2: "http://efo.svn.sourceforge.net/viewvc/efo/trunk/src/efoinowl/efo.owl?revision=214"
Number of classes that have changed: 120
Number of classes that have been added: 258
Number of classes that have been deleted: 111

Perform another diff | Export results as XML <xml/> | Export results as text

Classes modified:

Class: http://purl.obolibrary.org/obo/GO_0007608
Label: sensory perception of smell
- SubClassOf('sensory perception of smell' 'sensation')
+ SubClassOf('sensory perception of smell' http://purl.obolibrary.org/obo/GO_0007600)

Class: http://www.ebi.ac.uk/efo/EFO_0004525
Label: bitter taste sensitivity
- SubClassOf('bitter taste sensitivity' 'phenotype')
+ SubClassOf('bitter taste sensitivity' http://purl.obolibrary.org/obo/GO_0050909)

Class: http://www.ebi.ac.uk/efo/EFO_0004356
Label: taste
- SubClassOf('taste' 'sensation')

parents

disease

term history

The panel below shows changes to EFO that have affected this term over recent versions.

Changes in version: 2.46

ADD SUPERCLASS
ordo:Orphanet_98056 (Rare genetic renal disease)

ADD SUPERCLASS
ordo:Orphanet_68336 (Rare genetic tumor)

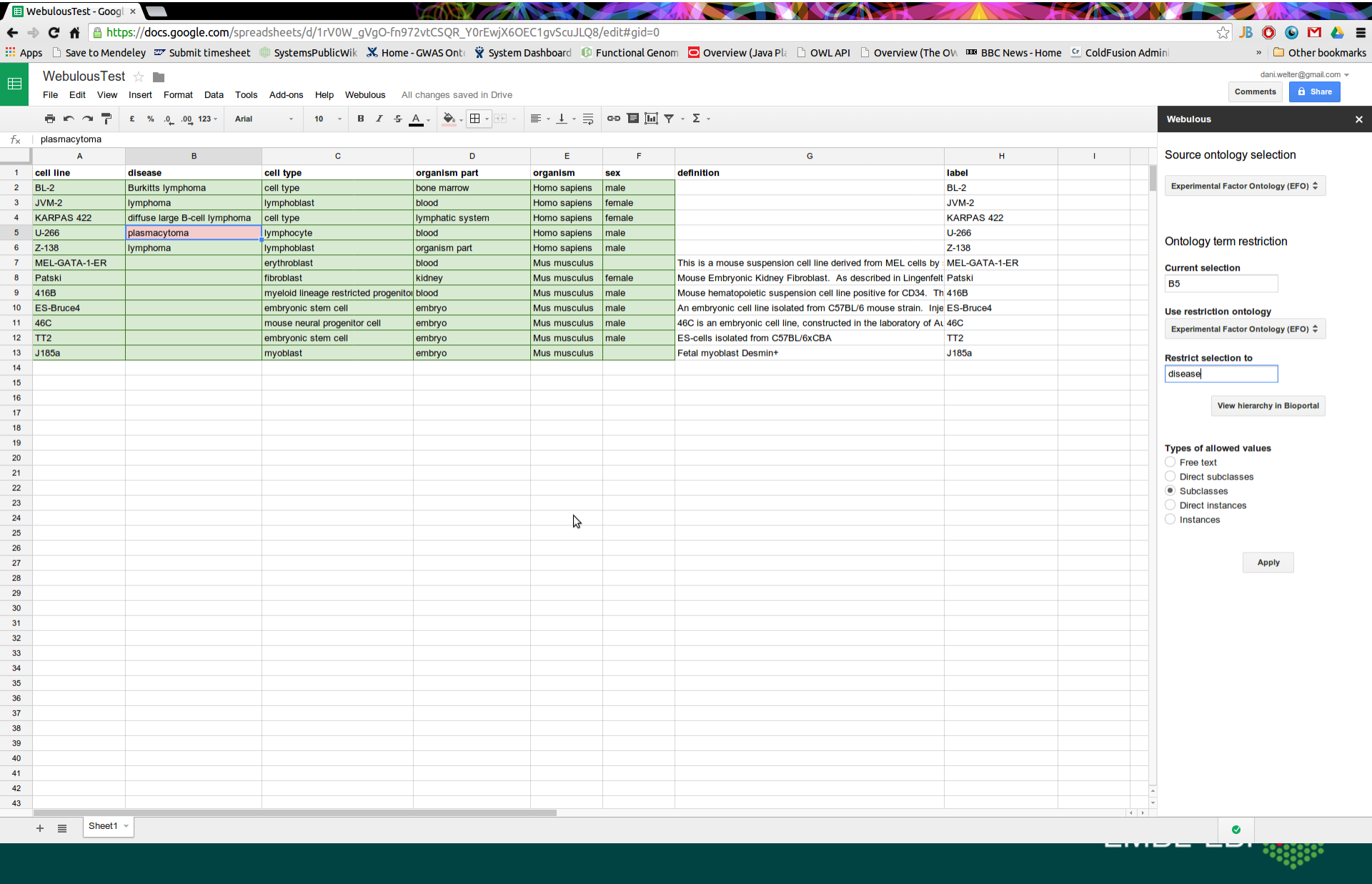
ADD SUPERCLASS
ordo:Orphanet_168615 (Hereditary persistence of alpha-fetoprotein)

ADD SUPERCLASS
ordo:Orphanet_68335 (Chromosomal anomaly)

ADD SUPERCLASS
ordo:Orphanet_140162 (Inherited cancer-predisposing

- aphthous ulcer
- carcinoid tumor and carcinoid syndrome
- cardiovascular disease
- chronic fatigue syndrome
- Complete hydatidiform mole
- cryptorchidism
- delayed encephalopathy after acute carbon monoxide poisoning
- dementia
- developmental disability
- digestive system disease
- endocrine system disease
- erectile dysfunction
- eye disease
- Familial sinus histiocytosis with massive lymphadenopathy
- **genetic disorder**
- head disease
- Hepatoblastoma
- hippocampal atrophy
- hyperplasia
- hypersomnia
- hypertrophy
- hypothyroidism
- Idiopathic copper-associated cirrhosis
- immune system disease

Webulous



	A	B	C	D	E	F	G	H	I
1	**cell line**	**disease**	**cell type**	**organism part**	**organism**	**sex**	**definition**	**label**	
2	BL-2	Burkitts lymphoma	cell type	bone marrow	Homo sapiens	male		BL-2	
3	JVM-2	lymphoma	lymphoblast	blood	Homo sapiens	female		JVM-2	
4	KARPAS 422	diffuse large B-cell lymphoma	cell type	lymphatic system	Homo sapiens	female		KARPAS 422	
5	U-266	plasmacytoma	lymphocyte	blood	Homo sapiens	male		U-266	
6	Z-138	lymphoma	lymphoblast	organism part	Homo sapiens	male		Z-138	
7	MEL-GATA-1-ER		erythroblast	blood	Mus musculus		This is a mouse suspension cell line derived from MEL cells by	MEL-GATA-1-ER	
8	Patski		fibroblast	kidney	Mus musculus	female	Mouse Embryonic Kidney Fibroblast. As described in Lingenfelt	Patski	
9	416B		myeloid lineage restricted progenitor	blood	Mus musculus	male	Mouse hematopoietic suspension cell line positive for CD34. Th	416B	
10	ES-Bruce4		embryonic stem cell	embryo	Mus musculus	male	An embryonic cell line isolated from C57BL/6 mouse strain. Inje	ES-Bruce4	
11	46C		mouse neural progenitor cell	embryo	Mus musculus	male	46C is an embryonic cell line, constructed in the laboratory of Au	46C	
12	TT2		embryonic stem cell	embryo	Mus musculus	male	ES-cells isolated from C57BL/6xCBA	TT2	
13	J185a		myoblast	embryo	Mus musculus		Fetal myoblast Desmin+	J185a	
14									
15									
16									
17									
18									
19									
20									
21									
22									
23									
24									
25									
26									
27									
28									
29									
30									
31									
32									
33									
34									
35									
36									
37									
38									
39									
40									
41									
42									
43									

Different views – tooling gaps

External ontologies

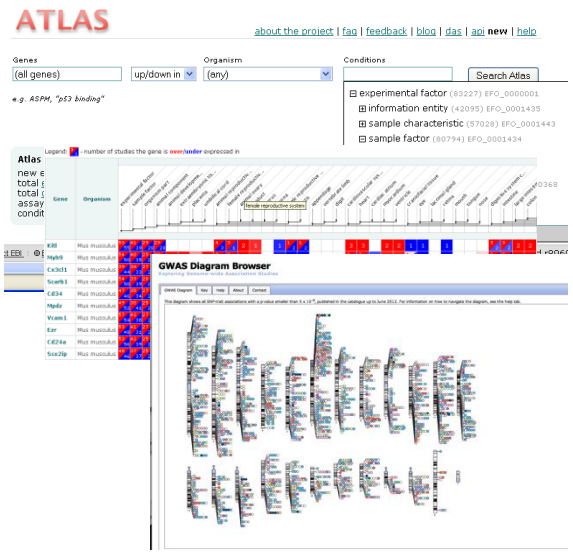
Ontology development

protégé

efo

Multiple Views

User & Analysis



Database curators

Data submission

GWAS View

GWAS Diagram Browser

Exploring Genome-wide Association Studies

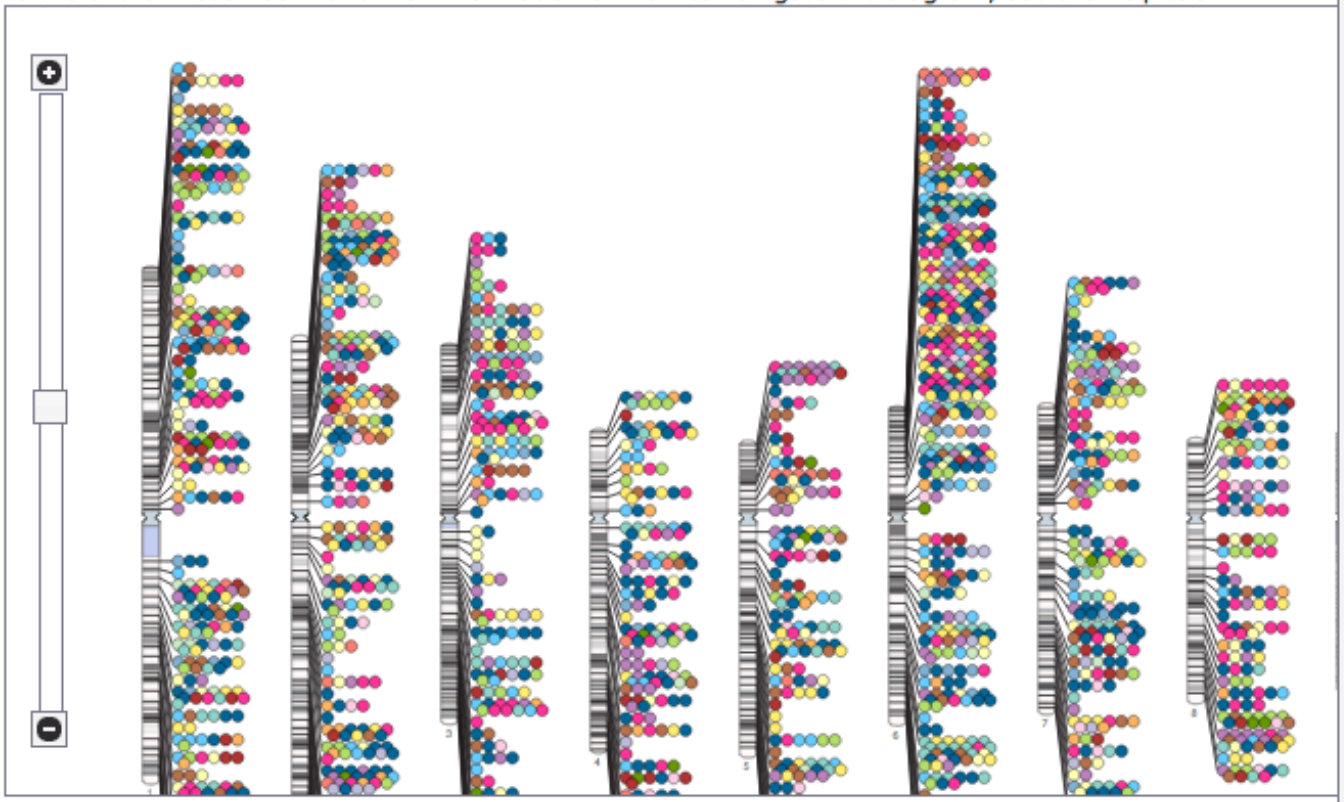
National Human Genome Research Institute | EMBL-EBI

Query by trait | Clear

To show only one trait, e.g. "breast cancer" or "schizophrenia", type the trait in the box on the left and hit "Query by trait"

Interactive GWAS Diagram | Trait-specific Views | Time Series Views | Downloads | Help | Curation process | Hide Legend

This diagram shows all SNP-trait associations with p-value ≤ 5.0×10^{-8}, published in the GWAS catalogue to the end of December 2013. For information on how to navigate the diagram, see the help tab.



SNP-associated trait categories

- Digestive system disease
- Cardiovascular disease
- Metabolic disease
- Immune system disease
- Nervous system disease
- Liver enzyme measurement
- Lipid or lipoprotein measurement
- Inflammatory marker measurement
- Hematological measurement
- Body weights and measures
- Cardiovascular measurement
- Other measurement
- Response to drug
- Biological process
- Cancer
- Other disease

GWAS View

GWAS Diagram Browser

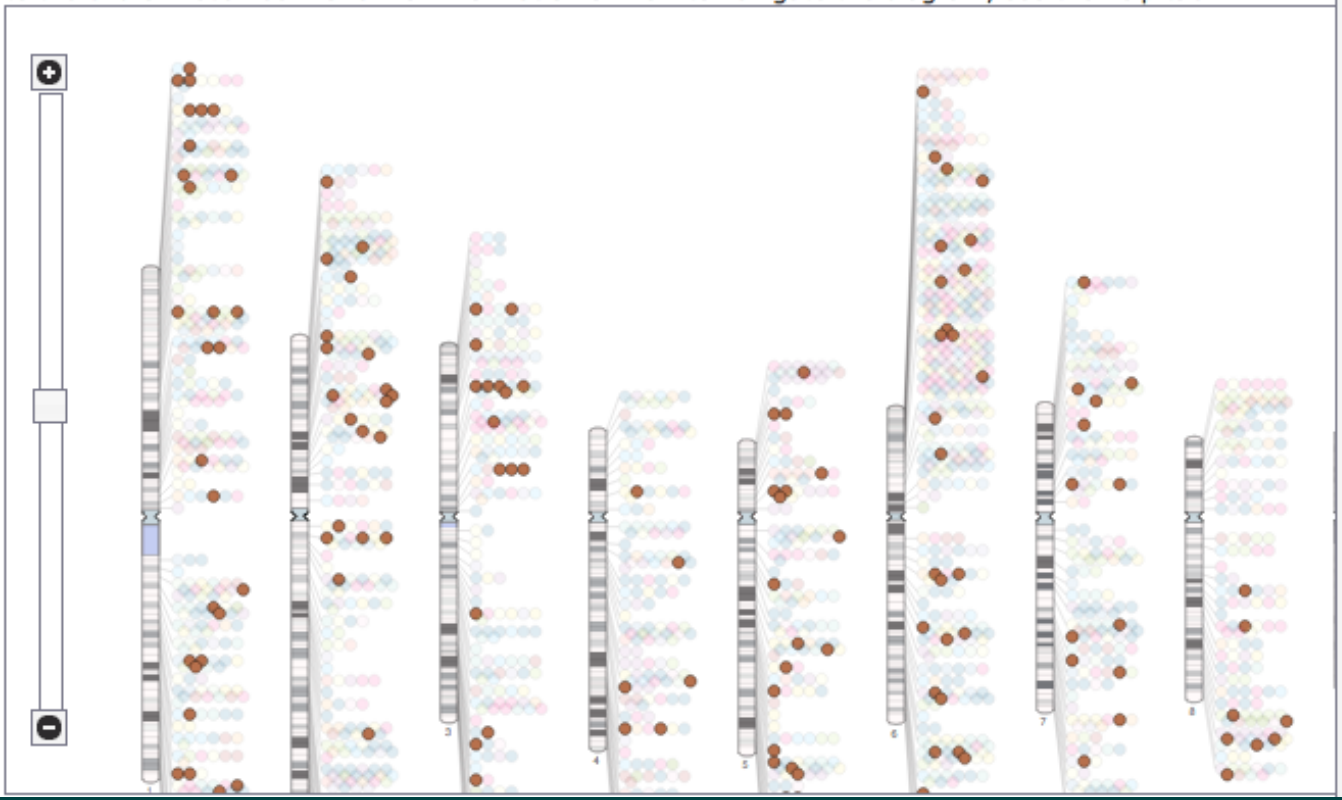
Exploring Genome-wide Association Studies

National Human Genome Research Institute | EMBL-EBI

Query by trait | Clear | *To show only one trait, e.g. "breast cancer" or "schizophrenia", type the trait in the box on the left and hit "Query by trait"*

Interactive GWAS Diagram | Trait-specific Views | Time Series Views | Downloads | Help | Curation process | Hide Legend

This diagram shows all SNP-trait associations with p-value ≤ 5.0×10^{-8}, published in the GWAS catalogue to the end of December 2013. For information on how to navigate the diagram, see the help tab.

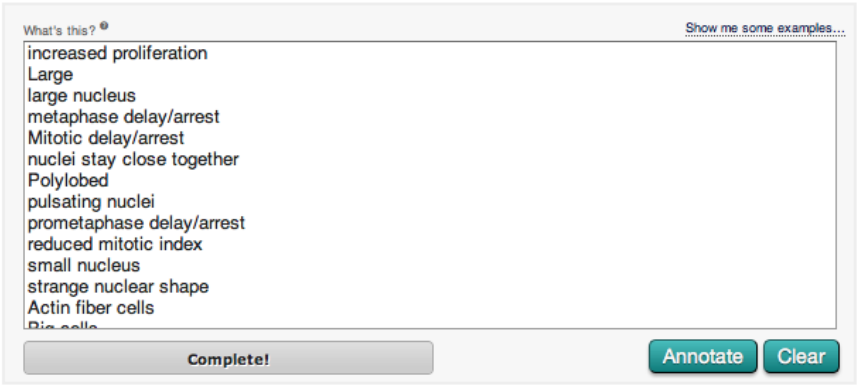


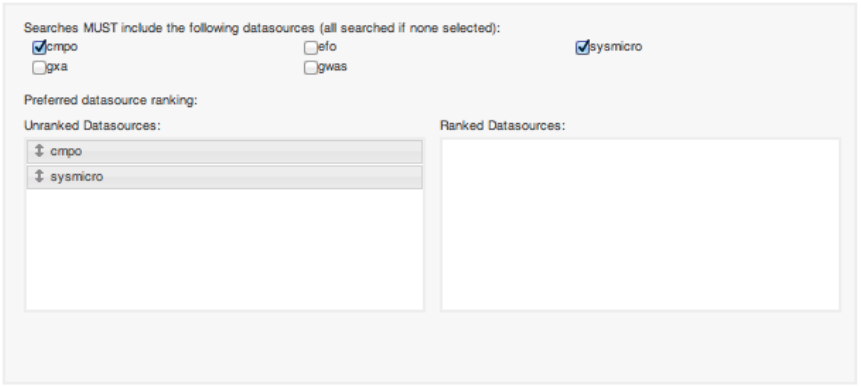
SNP-associated trait categories

- Digestive system disease
- Cardiovascular disease
- Metabolic disease
- Immune system disease
- Nervous system disease
- Liver enzyme measurement
- Lipid or lipoprotein measurement
- Inflammatory marker measurement
- Hematological measurement
- Body weights and measures
- Cardiovascular measurement
- Other measurement
- Response to drug
- Biological process
- Cancer

Annotating data with ontologies – Zooma

www.ebi.ac.uk/fgpt/zooma





Results

The table below shows a report describing how ZOOMA annotates text terms supplied above.

Hide results that did not map? ☑

Download my results

Term Type	Term Value	Ontology Class Label	Mapping Type	Ontology Class ID	Source
[NO TYPE]	Binuclear	binuclear cell phenotype	Automatic	CMPO_0000213	SysMicro
[NO TYPE]	cell death	cell death phenotype	Automatic	CMPO_0000030	SysMicro
[NO TYPE]	cell migration	cell migration phenotype	Automatic	CMPO_0000033	SysMicro
[NO TYPE]	Dynamic changes	increased variability of nuclear shape in population	Automatic	CMPO_0000345	SysMicro
[NO TYPE]	failure in decondensation	absence of mitotic chromosome decondensation phenotype	Requires curation	CMPO_0000216	http://www.ebi.ac.uk/cmpo/cmpo.owl
[NO TYPE]	Grape	graped micronucleus phenotype	Automatic	CMPO_0000156	SysMicro
[NO TYPE]	increased proliferation	proliferating cells	Automatic	CMPO_0000241	SysMicro
[NO TYPE]	Large	increased nucleus size phenotype	Automatic	CMPO_0000140	SysMicro

Conclusion

- Ontology as software adds QC
  - Borrowing from Software Engineering
  - Tooling is still major bottleneck (e.g. views)
- Data curation still costs too much
  - We have a good number of bio-ontologies now but applying them ex post facto is difficult and expensive
  - We need to make this part of everyday process
- Capturing expert knowledge requires familiar modes
  - Viva spreadsheets!

Acknowledgements

- EBI: Simon Jupp, Tony Burdett, Helen Parkinson, Eleanor Williams, Dani Welter, Jon Ison, Emma Hastings, Gabriella Rustici
- External collaborators, especially Chris Mungall, Michel Dumontier

Funding:

- NCBO, one of the National Centers for Biomedical Computing supported by the NHGRI, the NHLBI, and the NIH Common Fund under grant U54-HG004028
- EMBL funding
- BioMedBridges EC Grant number: 284209,
- Systems Microscopy EC Grant number: 258068

Links

EFO: www.ebi.ac.uk/efo

Zooma: www.ebi.ac.uk/fgpt/zooma

GWAS viz: www.ebi.ac.uk/fgpt/gwas

Bubastis: http://www.ebi.ac.uk/efo/bubastis

My blog: http://drjamesmalone.blogspot.co.uk