

Interoperability of large scale image data sets with ontologies

CrEDIBLE wokshop
10 October 2014, Sophia Antipolis

Simon Jupp
Samples Phenotypes and Ontologies Team
EMBL-European Bioinformatics Institute

Data resources at EMBL-EBI

Genes, genomes & variation

European Nucleotide Archive
1000 Genomes

Ensembl
Ensembl Genomes

European Genome-phenome Archive
Metagenomics portal

Gene, protein & metabolite expression

ArrayExpress

Expression Atlas

Metabolights
PRIDE

Protein sequences, families & motifs

InterPro

Pfam

UniProt

Molecular structures

Protein Data Bank in Europe
Electron Microscopy Data Bank

Chemical biology

ChEMBL

ChEBI

Systems

BioModels
Enzyme Portal
BioSamples

Literature & ontologies

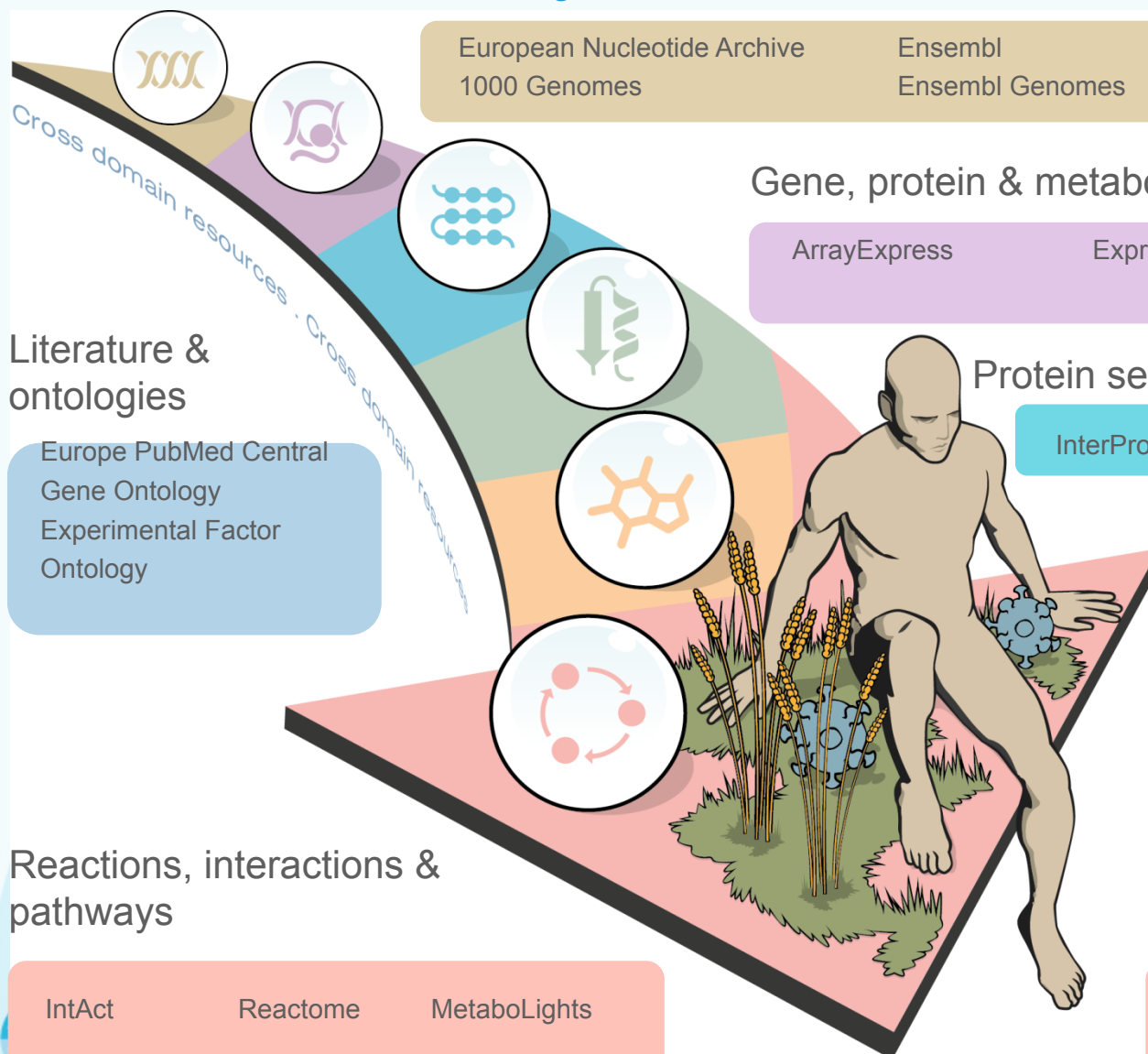
Europe PubMed Central
Gene Ontology
Experimental Factor
Ontology

Reactions, interactions & pathways

IntAct

Reactome

MetaboLights



Sample description with semantic markup

Export Experiment Design as Tab-Delimited file

Assay	Array	cell type	Compound treatment	organism
GSM589566	A-AFFY-141	human umbilical vein endothelial cell	valproic acid	Homo sapiens
GSM589557	A-AFFY-141	human umbilical vein endothelial cell	none	Homo sapiens
GSM589564	A-AFFY-141	human umbilical vein endothelial cell	valproic acid	Homo sapiens
GSM589558	A-AFFY-141	human umbilical vein endothelial cell	valproic acid	Homo sapiens
GSM589559	A-AFFY-141	human umbilical vein endothelial cell	none	Homo sapiens
GSM589563	A-AFFY-141	human umbilical vein endothelial cell	none	Homo sapiens
GSM589561	A-AFFY-141	human umbilical vein endothelial cell	none	Homo sapiens
GSM589565	A-AFFY-141	human umbilical vein endothelial cell	none	Homo sapiens
GSM589560	A-AFFY-141	human umbilical vein endothelial cell	valproic acid	Homo sapiens
GSM589562	A-AFFY-141	human umbilical vein endothelial cell	valproic acid	Homo sapiens

CL:CL_0000071
(blood vessel endothelial cell)

obo:CHEBI_39867
(valproic acid)

NCBITaxon:NCBITaxon_9606
(Homo Sapiens)

Ontologies add value

Smarter searching

Experiment, citation, sample and factor annotations [clear]

leuk

- leukaemia
 - leukemia
 - acute lymphoblastic leukemia
 - B-cell acute lymphoblastic leukemia
 - T-cell acute lymphoblastic leukemia
 - acute myeloid leukemia
 - chronic lymphocytic leukemia
 - chronic myelogenous leukemia
- leukemias
- leukemic
- leukemogenesis
- leukocyte
- leukocytes

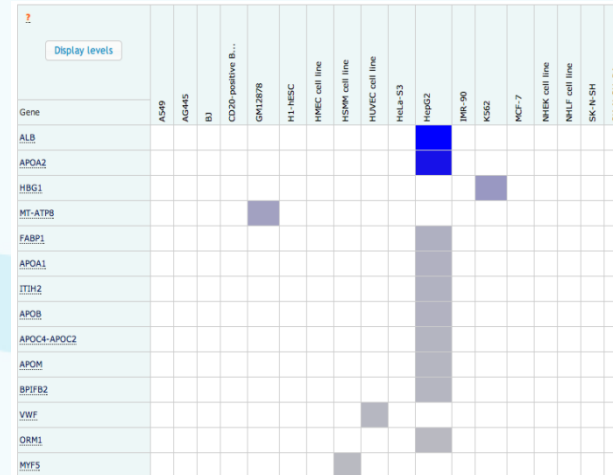
Filter on [reset] Display options [reset]
Any species 25 experiments
Any array
Any experiment type

Submitter/reviewer login ArrayExpress Browser Help

ID	Title	Assays	Species	Date
E-TABM-763	MicroRNA profiling of peripheral blood mononuclear cells from CLL patients to L...	61	Homo sapiens	2009-08-14
E-TABM-762	MicroRNA profiling of peripheral blood mononuclear cells from CLL patients to L...	97	Homo sapiens	2009-08-14
E-TABM-726	Transcription profiling of liver from wild type, Rev knock out and Rev-erb-alpha...	6	Mus musculus	2009-09-03
E-TABM-696	Transcription profiling of human chronic lymphocytic leukemia cells with muta...	24	Homo sapiens	2009-04-30
E-TABM-695	Transcription profiling of mouse LSK hematopoietic stem cells from wild type a...	12	Mus musculus	2009-04-28
E-TABM-694	Transcription profiling of mouse LT-HSC hematopoietic stem cells from wild typ...	5	Mus musculus	2009-04-28
E-TABM-670	Transcription profiling of mouse embryonic stem cell line CGR8 grown in prese...	9	Mus musculus	2009-03-18
E-TABM-667	Transcription profiling of mouse embryonic stem cell line CGR8 treated with L...	30	Mus musculus	2009-03-18
E-TABM-632	Transcription profiling of human acute myeloid leukemia cells before and early ...	24	Homo sapiens	2009-02-03
E-TABM-628	MicroRNA profiling of human chronic lymphocytic leukemia cells in response to L...	12	Homo sapiens	2009-01-27
E-TABM-431	Chromatin immunoprecipitation of mouse hematopoietic cell lines and tissue ty...	25	Mus musculus	2008-03-26
E-TABM-429	MicroRNA profiling of human acute myeloid leukemia samples from patients cha...	85	Homo sapiens	2008-03-13
E-TABM-405	MicroRNA profiling of patients with acute myeloid leukemia to identify microRN...	176	Homo sapiens	2008-01-09
E-TABM-391	Transcription profiling of human CD4+ leukemia Jurkat T-cells in which the PI3...	4	Homo sapiens	2007-12-18
E-TABM-346	Transcription profiling of human patients with diffuse large B-cell lymphoma tre...	53	Homo sapiens	2007-10-16
E-TABM-293	Comparative genomic hybridization of human chronic lymphocytic leukemia sa...	2	Homo sapiens	2007-07-17
E-SMDB-2850	Transcription profiling of acute myeloid leukemia FLT3 wild type and mutants in childhood AML samples from the Pediatric Oncology Group Study 9421.	87	Homo sapiens	2005-12-08

Description This data set was used to study FLT3 wild type and mutants in childhood AML samples from the Pediatric Oncology Group Study 9421, and published in the Journal Blood in 2002 by Lacayo NJ et al.

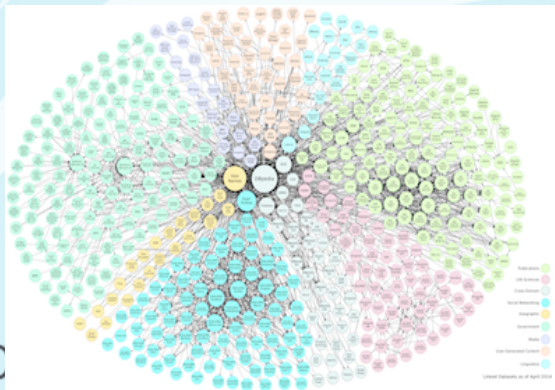
Data analysis



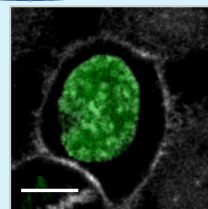
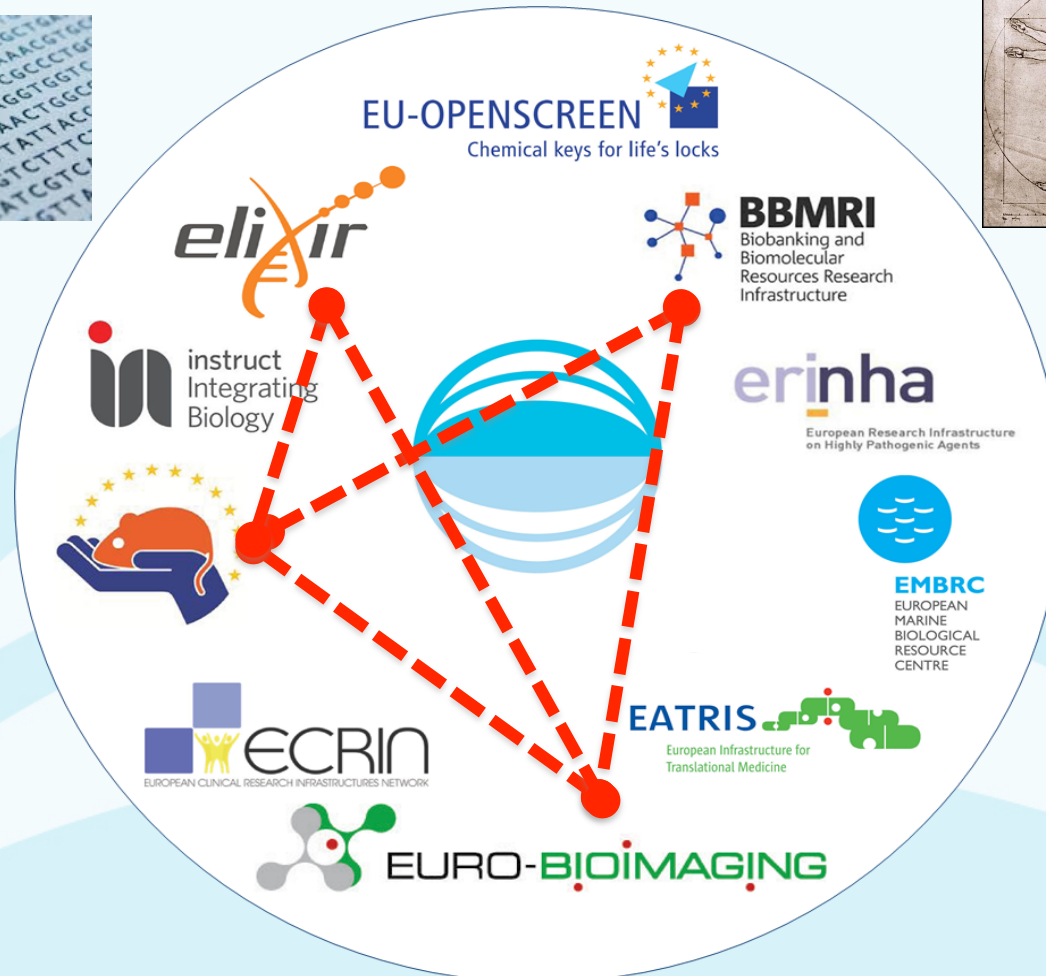
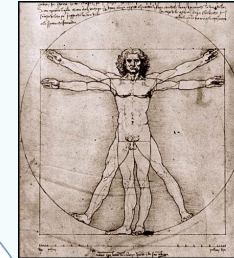
Data visualisation



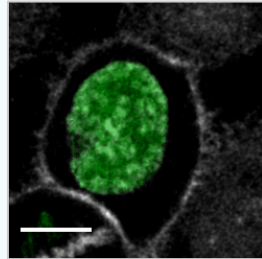
Data integration



BioMedBridges Project



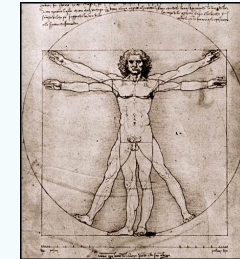
Scientific problem



Cell



Mouse



Human

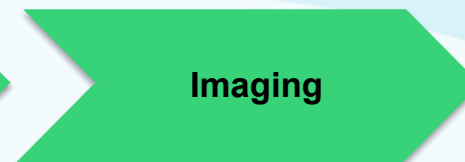
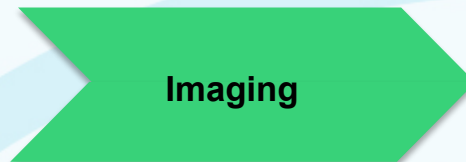


Genome

- ✓ Cellular **phenotype**
- ✓ Genetic information
- ✓ Molecular mechanism

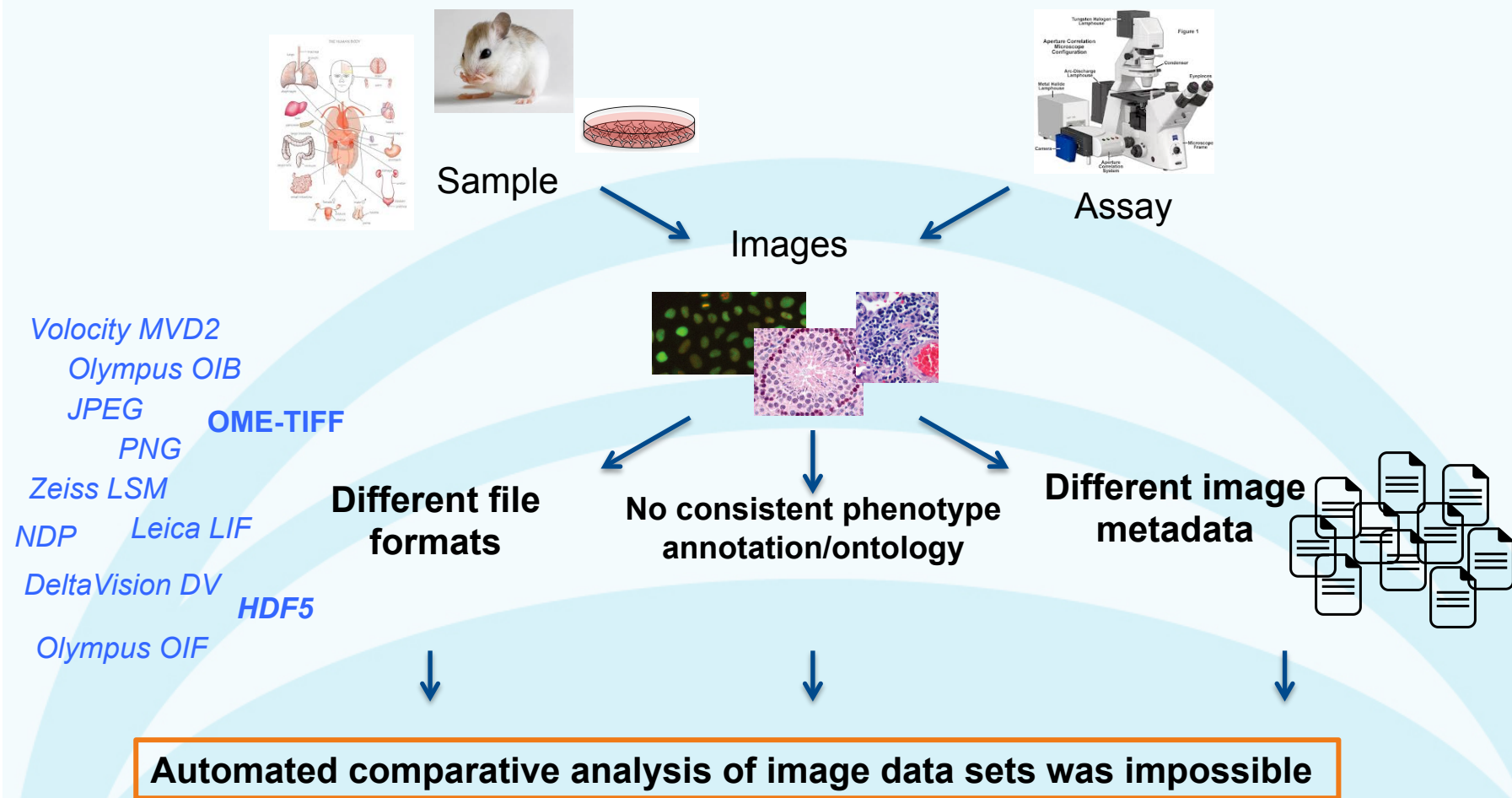
- ✓ Tissue **phenotype**
- ✓ Genetic information

- ✓ Tissue **phenotype**



By linking these three different types of data sets, we can better understand diseases, predict novel drug targets and biomarkers

To compare and integrate image data we need interoperable standards

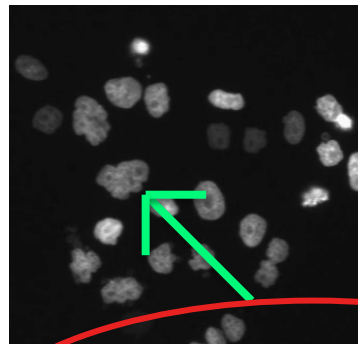


Correlative analysis and biomarker prediction

Promising gene candidates from cellular screens

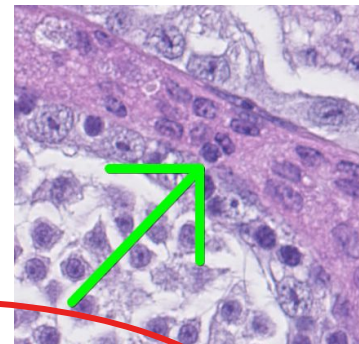
- MLL3
- PAPP A
- SF3B1
- PRPF8
- CENPE
- CIT
- **ASPM**
- ESPL1
- DYNC1H1
- ASCC3
- KIF4A

Cell line
ASPM Knockdown



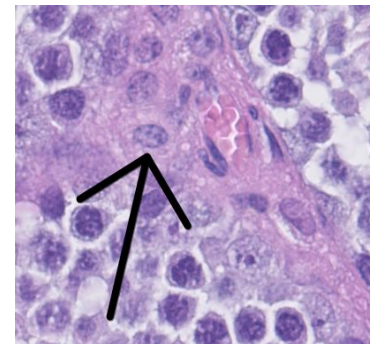
Polylobed nucleus

Mouse
ASPM Mut



Polylobed
nucleus

Mouse
ASPM WT



Mouse and human tissue WP6 partners are looking for and/or generating the data relevant to these genes to be used for analysis.

Matching phenotypes at different scales

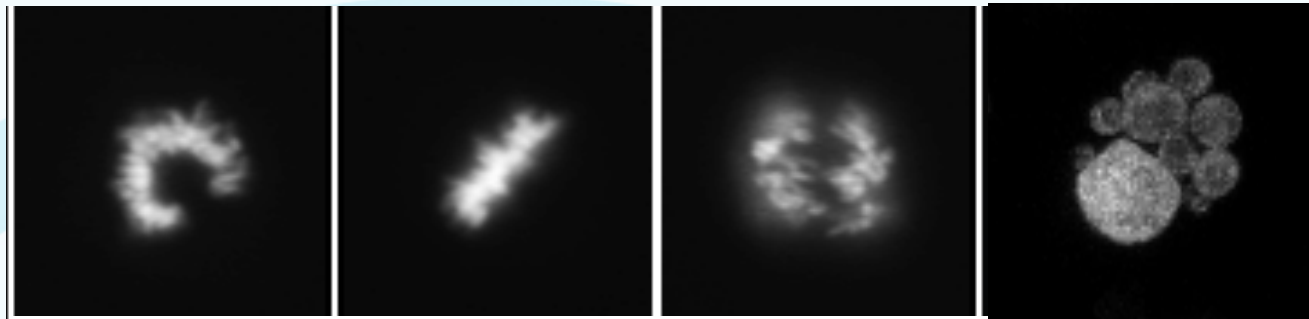
Prometaphase

Metaphase

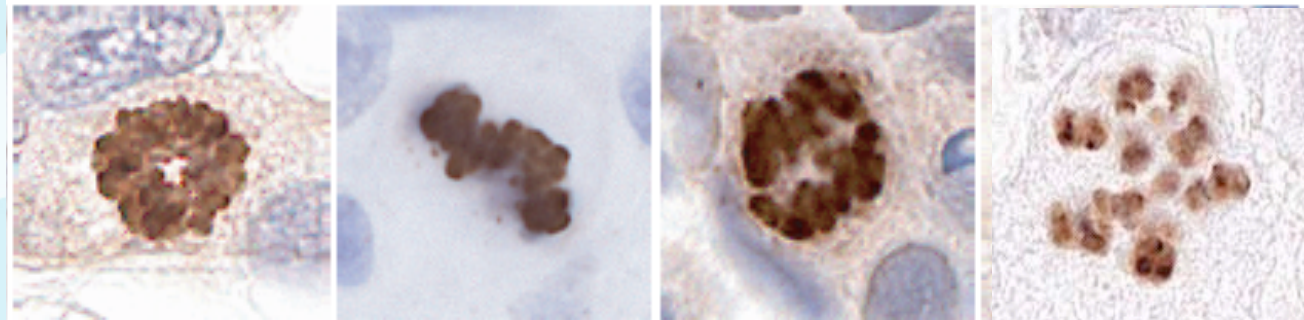
Anaphase

Graped
micronucleus

Cell line – gene
knockdown



Human cancer
tissue



State of the art: finding a match by chance

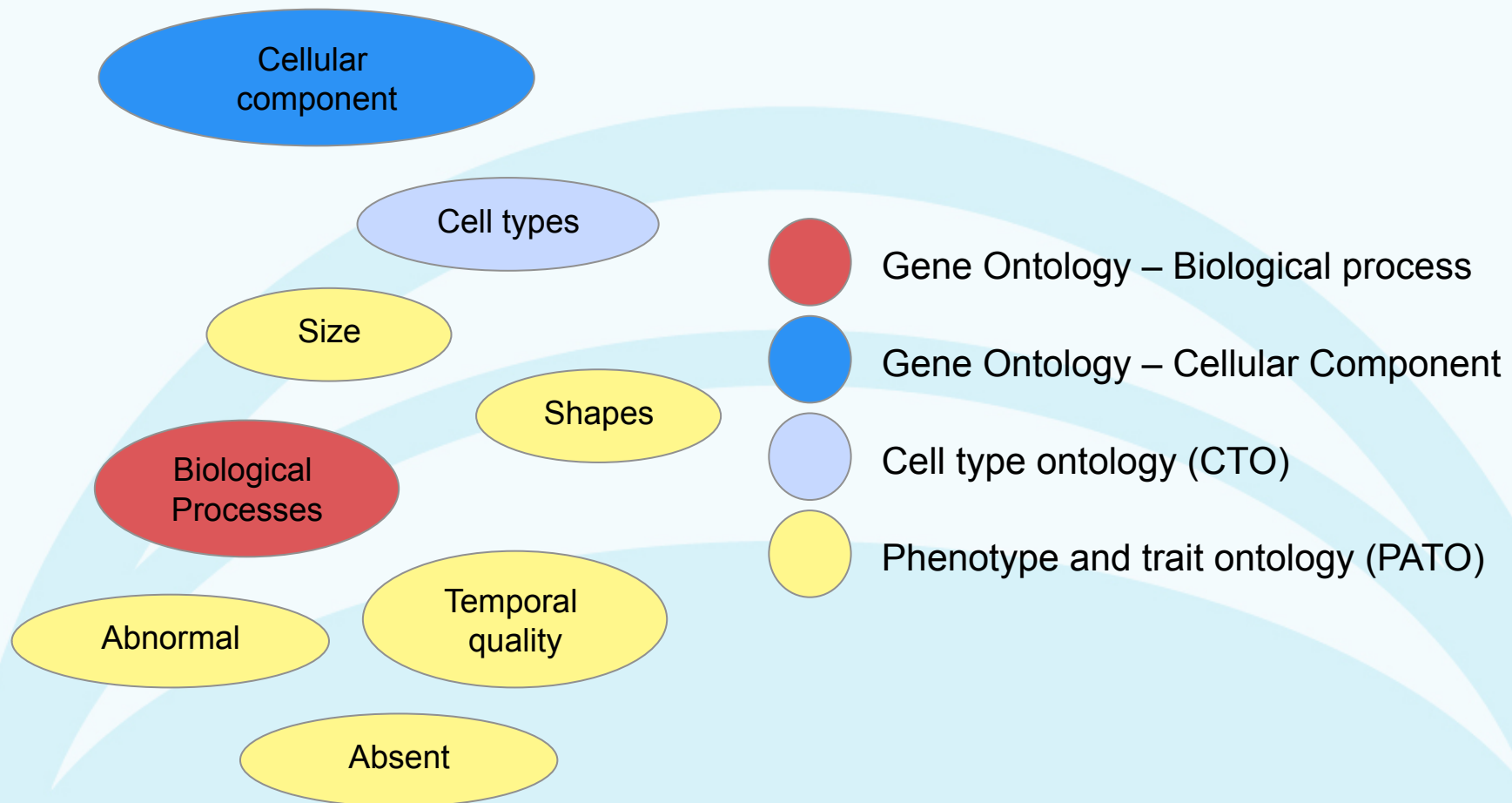
Use cases

- Defined classes: “*Abnormal cellular membrane phenotypes*” includes all parts of the cellular membrane
- Real queries
 - Genes / siRNAs / Images having a **mitotic phenotype** in all screens or in one particular screen.
 - Genes / siRNAs / Images having a **cytokinesis phenotype** in all screens or in one particular screen.
 - Genes / siRNAs / Images having a **mitotic phenotype**, but **no cell death**.
 - Genes / siRNAs / Images having a **mitotic phenotype**, but **no** problem in **Prophase**.
 - Genes / siRNAs / Images having a **mitotic phenotype followed** by **cytokinesis defects**.
 - Genes / siRNAs / Images having a **mitotic phenotype before** 30 hours in all screens or in one particular screen.
 - Genes / siRNAs / Images having a **mitotic phenotype** restricted to **human** and **Drosophila**.

Existing ontologies are not enough

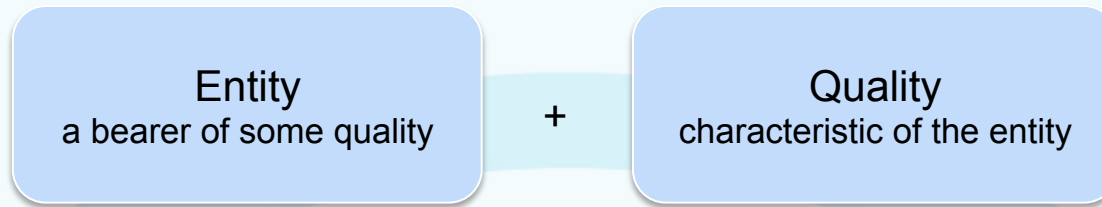
- Existing ontologies either lack coverage or are incomplete to describe cellular scale phenotypes
- No species neutral ontology for cellular phenotypes
- Such ontology is needed for data interoperability
- **WP6 developed the**
Cellular Microscopy Phenotype Ontology (CMPO)

Cellular phenotypes: entities, processes and qualities



Building a phenotype ontology

Composing a phenotype description



Examples:

- Phenotype: "*Large nucleus*"
 - Entity: nucleus (GO_000xxxx)
 - Quality: large (PATO_000xxxx)
- Phenotype: "*Cells stuck in metaphase due to metaphase arrest*"
 - Entity: mitotic metaphase (GO_0000089)
 - Quality: arrested (PATO_0000297)

Cellular Microscopy Phenotype Ontology (CMPO)

- Annotate the available data using EQ based ontology annotations (post composition)
- Translate EQs into OWL axioms following OBO style “part-of some” pattern
- Create new terms via post composition and import into the CMPO ontology (assign labels, definitions, provenance)

Composing ontology terms from annotations in OWL

- More expressive, faster reasoners (ELK)
- Basic modeling pattern (*has_part some (<Quality> and inheres_in some <Entity>)*)

Original phenotype: metaphase arrest

EQ annotation :

mitotic metaphase (GO_0000089)
arrested (PATO_0000297)

Translation to OWL class description:

CMPO_0000305 **equivalentTo**

has_part **some** ('arrested'

and (*inheres_in* **some** 'mitotic metaphase'))

Enabling standardised data generation

Phenotator: user-friendly ontology annotation of image data

Edit Entity

Accession:

Phenotype:

segregation problems/chromatin bridges/lagging chromosomes/multiple dna masses

Comment:

Annotations:

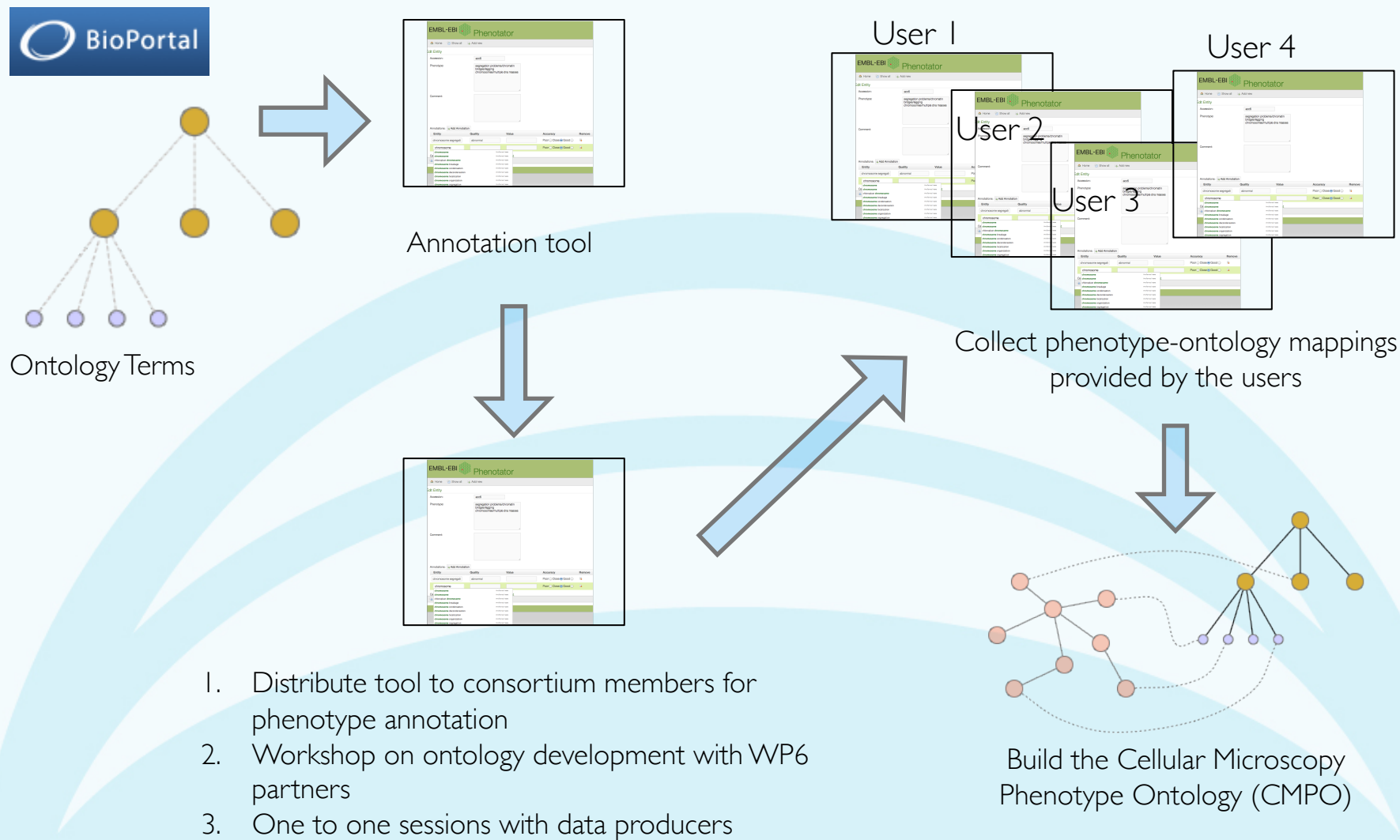
Entity	Quality	Value	Accuracy	Remove
chromosome segregati	abnormal		Poor <input type="radio"/> Close <input checked="" type="radio"/> Good <input type="radio"/>	<input type="button" value="Remove"/>
chromosome			Poor <input type="radio"/> Close <input checked="" type="radio"/> Good <input type="radio"/>	<input type="button" value="Remove"/>
chromosome		Preferred Name		
Ca chromosome		Preferred Name		
chloroplast chromosome		Preferred Name		
chromosome breakage		Preferred Name		
chromosome condensation		Preferred Name		
chromosome decondensation		Preferred Name		
chromosome localization		Preferred Name		
chromosome organization		Preferred Name		
chromosome segregation		Preferred Name		

Original
phenotypic
description

Ontology based
annotations

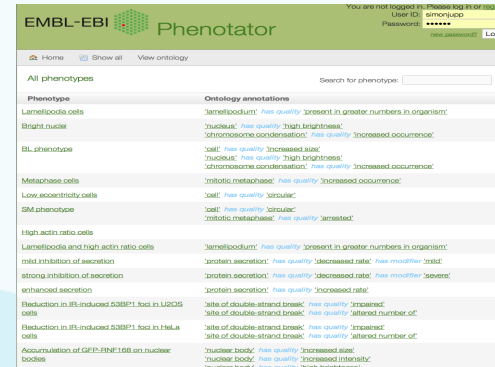
<http://wwwdev.ebi.ac.uk/fgpt/phenotator/>

Ontology building using Phenotator



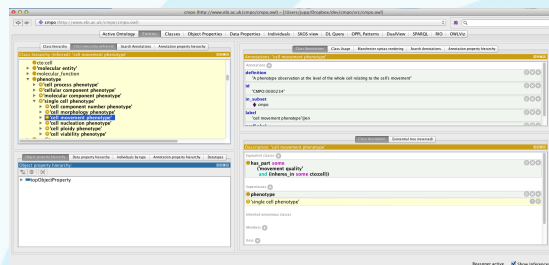
Building the ontology

CMPO upper level species neutral ontology
(cell process phenotypes,
cellular component phenotypes,
whole cell phenotypes,
cell population)

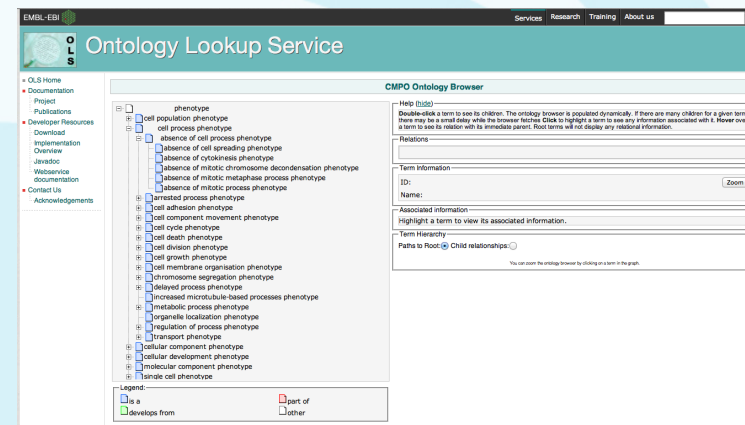


Phenotype	Ontology annotations
Lamellipodia cells	'lamellipodium' <i>has_quality</i> 'present in greater numbers in organism'
Bright nuclei	'nucleus' <i>has_quality</i> 'high brightness'
BL phenotype	'chromosome condensation' <i>has_quality</i> 'increased occurrence'
Metaphase cells	'cell' <i>has_quality</i> 'increased size'
Low acentrinrichy cells	'nucleus' <i>has_quality</i> 'high brightness'
SM phenotype	'chromosome condensation' <i>has_quality</i> 'increased occurrence'
High actin ratio cells	'mitotic metaphase' <i>has_quality</i> 'increased occurrence'
Lamellipodia and high actin ratio cells	'cell' <i>has_quality</i> 'scruled'
mild inhibition of secretion	'cell' <i>has_quality</i> 'scruled'
strong inhibition of secretion	'mitotic metaphase' <i>has_quality</i> 'scruled'
enhanced secretion	'lamellipodium' <i>has_quality</i> 'present in greater numbers in organism'
Reduction in IFN induced G3BP1 foci in L2COS cells	'protein secretion' <i>has_quality</i> 'decreased rate' <i>has_modifier</i> 'mild'
Reduction in IFN induced G3BP1 foci in HeLa cells	'protein secretion' <i>has_quality</i> 'decreased rate' <i>has_modifier</i> 'severe'
Accumulation of GFP-RNF168 on nuclear foci	'protein secretion' <i>has_quality</i> 'increased rate'
	'sites of double strand break' <i>has_quality</i> 'increased number of'
	'sites of double strand break' <i>has_quality</i> 'increased number of'
	'sites of double strand break' <i>has_quality</i> 'increased number of'
	'nuclear body' <i>has_quality</i> 'increased size'
	'nuclear body' <i>has_quality</i> 'increased intensity'
	'nuclear body' <i>has_quality</i> 'high brightness'

Convert annotation to OWL axioms
and owl:import and classify under
upper level CMPO

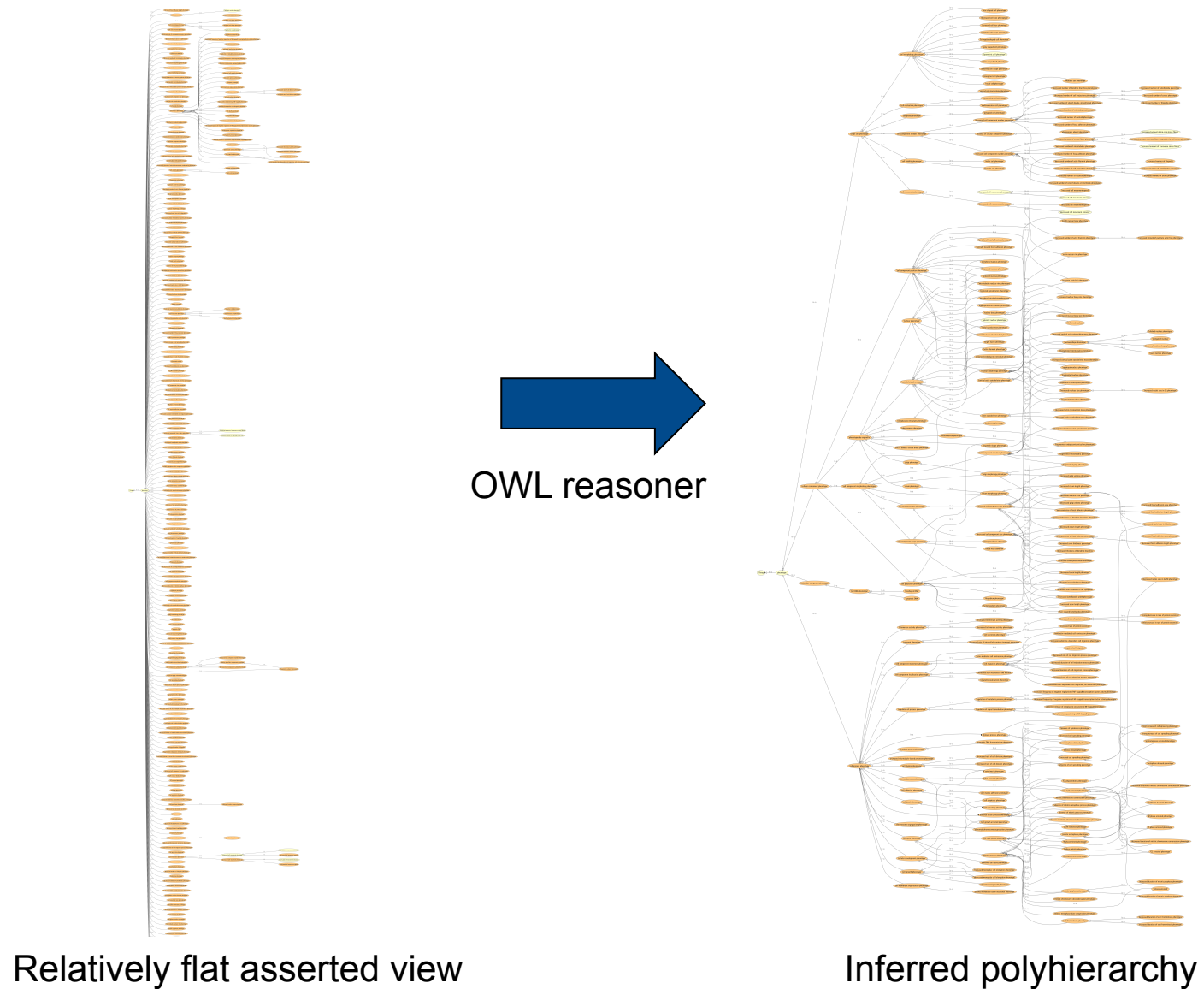


Ontology editor merges terms,
adds definitions,



Biologists verify new classification

Exploit GO and PATO to automatically construct CMPO hierarchy



Inferring equivalence across species (MP, FYPO and CMPO)

- ⊖ 'decreased axon thickness phenotype'
 - ▶ ⊖ 'short cochlear hair cell stereocilia'
 - ⊖ 'short photoreceptor inner segment'
 - ⊖ 'short vestibular hair cell stereocilia' http://purl.obolibrary.org/obo/MP_0004524
 - ⊖ 'thin cochlear hair cell stereocilia'
 - ⊖ 'thin sperm flagellum'
 - ⊖ 'thin vestibular hair cell stereocilia'
- ▶ ⊖ 'increased cell component size phenotype'
- ▶ ⊖ 'cell component structure phenotype'
- ▶ ⊖ 'cell morphology phenotype'
- ▼ ⊖ 'cilium morphology phenotype'
 - ▶ ⊖ 'abnormal cilium morphology'
 - ▶ ⊖ 'decreased cilium length phenotype'
 - ⊖ 'disorganized photoreceptor outer segment'
 - ▶ ⊖ 'increased cilium length phenotype'
 - ⊖ 'photoreceptor outer segment degeneration'
- ▶ ⊖ 'golgi morphology phenotype'
- ▶ ⊖ 'nuclear morphology phenotype'
- ▶ ⊖ 'cell component position phenotype'
- ▼ ⊖ 'cell projection phenotype'
 - ▶ ⊖ 'abnormal cilium physiology'
 - ▶ ⊖ 'abnormal neurite morphology'
 - ▶ ⊖ 'abnormal photoreceptor outer segment morphology'
 - ▶ ⊖ 'abnormal small intestinal microvillus morphology'
 - ▶ ⊖ 'absent cochlear hair cell stereocilia'
 - ▶ ⊖ 'absent vestibular hair cell stereocilia'
 - ▶ ⊖ 'axon degeneration'
 - ▼ ⊖ 'cilium morphology phenotype'
 - ▼ ⊖ 'abnormal cilium morphology'
 - ▶ ⊖ 'abnormal motile cilium morphology'
 - ▶ ⊖ 'abnormal primary cilium morphology'
 - ▼ ⊖ 'decreased cilium length phenotype'
 - ⊖ 'short photoreceptor outer segment'

Annotations +

definition

"A phenotype observation at the level of a cilium relating to the components shape, size or structure"

id

"CMPO:0000252"

in_subset

◆ cmpo

label

"cilium morphology phenotype"@en

prefLabel

"cilium morphology"@en

Description: 'cilium morphology phenotype'

Equivalent classes +

● has_part some
(morphology
and (inheres_in some cilium))

Superclasses +

● phenotype













⊖ 'cell component morphology phenotype'

⊖ 'cell projection phenotype'

⊖ 'cilium phenotype'

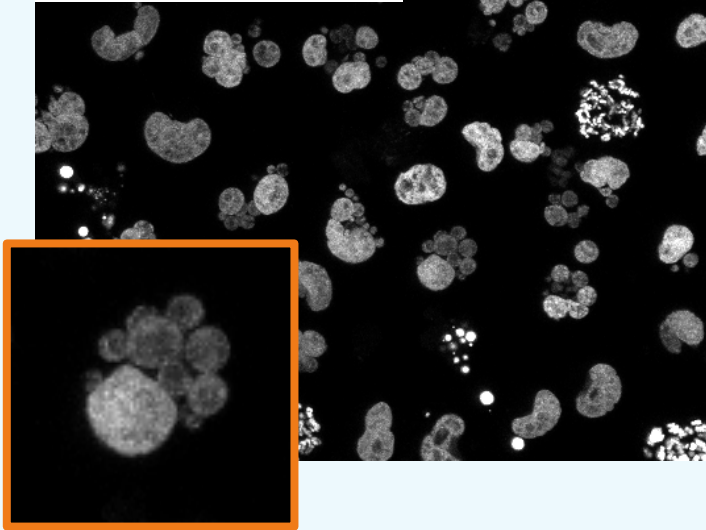
Inherited anonymous classes

Automated ontology mapping with

elongated cells	elongated cell phenotype	Automatic	CMPO_0000077	 SysMicro
lamellipodia + high actin ratio cells	more lamellipodia cells, increased number of actin filament phenotype	Automatic	CMPO_0000083 , CMPO_0000105	 SysMicro
lamellipodia cells	more lamellipodia cells	Automatic	CMPO_0000083	 SysMicro
sm phenotype	metaphase arrested phenotype, round cell phenotype	Automatic	CMPO_0000305 , CMPO_0000118	 SysMicro
reduction in ir-induced 53bp1 foci in u2os cells	site of double-strand break phenotype, decreased number of site of double-strand break phenotype	Automatic	CMPO_0000180 , CMPO_0000181	 SysMicro
accumulation of gfp-rnf168 on nuclear bodies	bright nuclear body phenotype, increased nuclear body size phenotype	Automatic	CMPO_0000335 , CMPO_0000126	 SysMicro
intracellular retention of sh4(yes)-mcherry	decreased rate of intracellular protein transport phenotype	Automatic	CMPO_0000346	 SysMicro
no nf-kb oscillation	cytoplasmic sequestering of NF-kappaB phenotype	Automatic	CMPO_0000332	 SysMicro
decreased nf-kb oscillation	decreased frequency of negative regulation of NF-kappaB transcription factor activity phenotype	Automatic	CMPO_0000330	 SysMicro
cell shape processes or spiky or stretchy	star shaped cell phenotype	Automatic	CMPO_0000267	 SysMicro
increased number of actin puncta or dots	increased amount of punctate actin foci phenotype	Automatic	CMPO_0000291	 SysMicro
increased number of zigzag actin stress fibers	increased amount of zig-zag stress fibers	Automatic	CMPO_0000299	 SysMicro

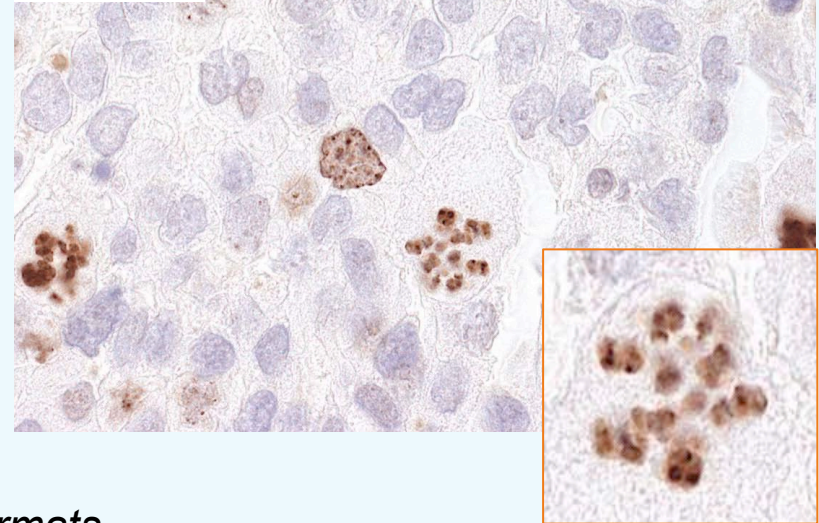
Cell

Gene knockdown



Human

Disease



Integrate file formats

Integrate metadata

Apply phenotype ontology

CMPO term:

graped micronucleus

CMPO_0000156

CMPO term:

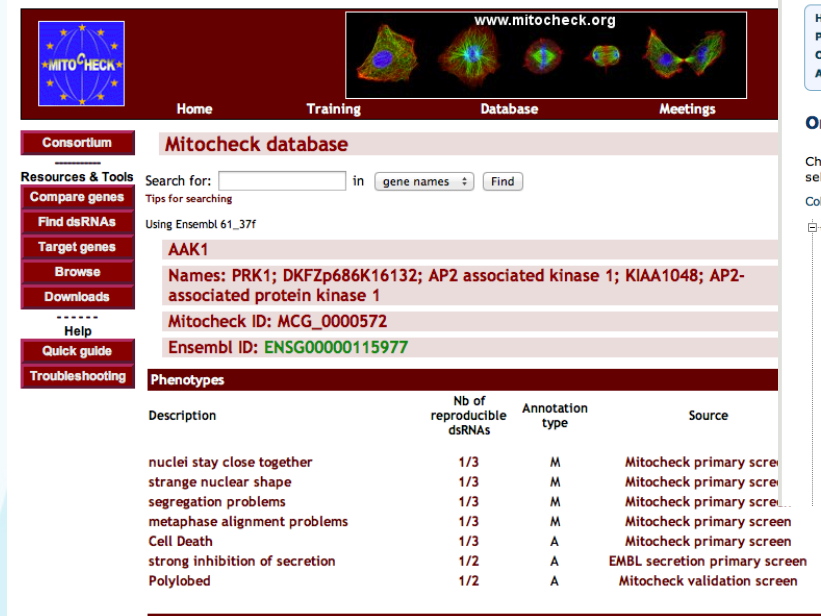
graped micronucleus

CMPO_0000156

Predict disease gene/biomarkers

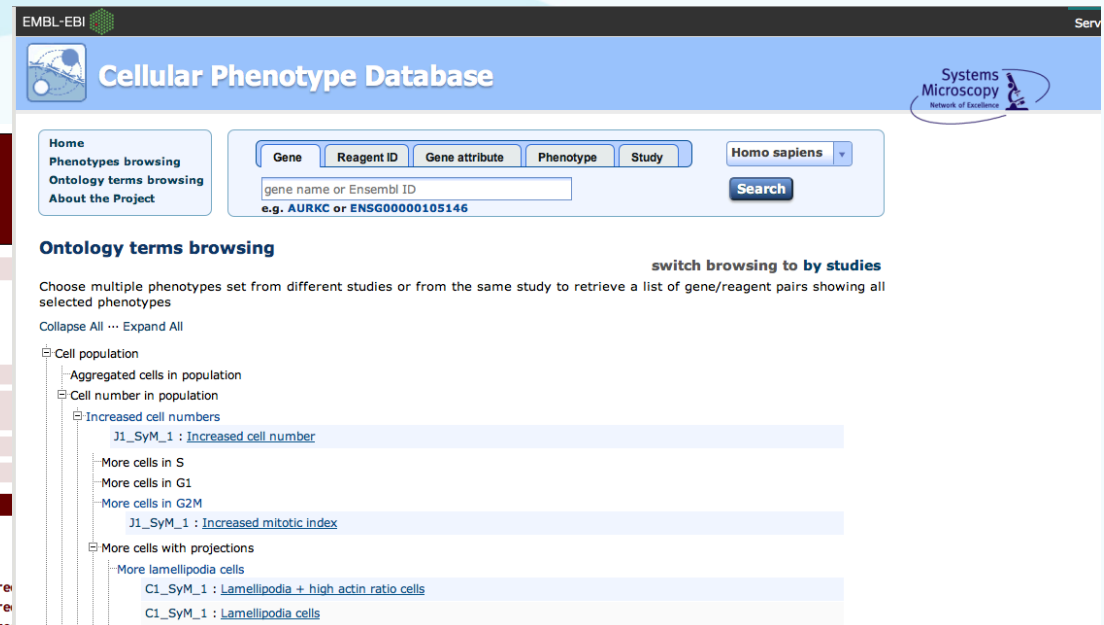
Integrating CMPO

- Cellular phenotype database (EMBL-EBI), Mitosys.org (EMBL), Webmicroscope (coming soon)
- OMERO, Phenoimage share, CellCognition, CellProfiler, Knime, ImageJ, Bioconductor, GenomeRNAi



The screenshot shows the Mitosys.org website. At the top, there's a navigation bar with links: Home, Training, Database, and Meetings. Below this is a search bar with the text "Search for:" and a dropdown menu set to "gene names". To the left of the search bar is a sidebar with links: Consortium, Resources & Tools, Compare genes, Find dsRNAs, Target genes, Browse, Downloads, Help, Quick guide, and Troubleshooting. The main content area displays the "Mitochondrial database" for the gene "AAK1". It lists various phenotypes with their descriptions, the number of reproducible dsRNAs, the annotation type, and the source. For example, "nuclei stay close together" has 1/3 reproducible dsRNAs, is annotated as "M", and is from a "Mitochondrial primary screen".

Description	Nb of reproducible dsRNAs	Annotation type	Source
nuclei stay close together	1/3	M	Mitochondrial primary screen
strange nuclear shape	1/3	M	Mitochondrial primary screen
segregation problems	1/3	M	Mitochondrial primary screen
metaphase alignment problems	1/3	M	Mitochondrial primary screen
Cell Death	1/3	A	Mitochondrial primary screen
strong inhibition of secretion	1/2	A	EMBL secretion primary screen
Polylobed	1/2	A	Mitochondrial validation screen



The screenshot shows the EMBL-EBI Cellular Phenotype Database website. At the top, there's a navigation bar with links: Home, Phenotypes browsing, Ontology terms browsing, About the Project, and a search bar. The search bar has tabs for Gene, Reagent ID, Gene attribute, Phenotype, and Study. Below the search bar is a section titled "Ontology terms browsing" with a link to "switch browsing to by studies". It lists various ontology terms such as "Cell population", "Aggregated cells in population", "Cell number in population", "Increased cell numbers", "More cells in S", "More cells in G1", "More cells in G2M", "More cells with projections", and "More lamellipodia cells".

<http://www.ebi.ac.uk/fg/sym>

<http://www.mitosys.org>

EBI Collaborations with Image Generation Projects

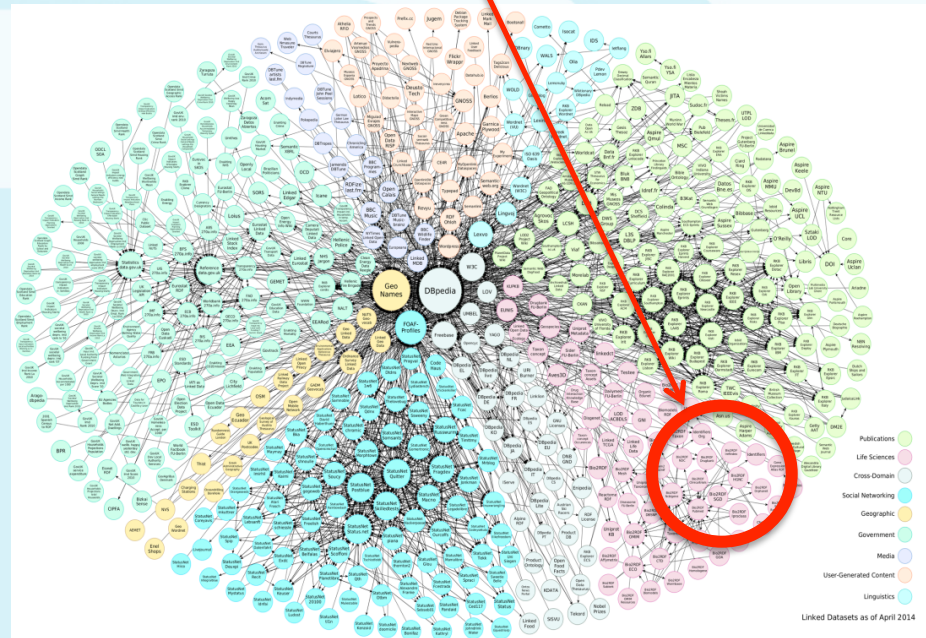
- KOMP2 – adult and mouse embryonic phenotyping (WTSI, MRC Harwell)
- HIPSCI – Cellular Imaging Watt Lab of iPSCs (WTSI, EBI, Lamond)
- BioMedBridges – Cross scale image integration – EuroBioImaging, Jan Ellenberg
- EMDB – Electron Micrographs (Gerard Kleywegt)
- Virtual FlyBrain (O’Kane, Jefferis, Armstrong, David Osumi-Sutherland)
- BioStudies – Unstructured Data (Alvis Brazma)
- EBiSC – iPSC – via CellFinder – Andreas Kurtz, Charite

Publishing biological data as Linked Open Data

- The EBI RDF platform
 - Released Nov 2013
 - Currently over 16 billion RDF triples
- Datasets updated ~ quarterly

Jupp et al (2013). **The EBI RDF Platform: Linked Open Data for the Life Sciences**. Bioinformatics.

LOD diagram August 2014



Bridging the semantic gap

- SPARQL extensions to support OWL queries
 - OWLET, Aber-OWL
- Dedicated SPARQL endpoints that can answer OWL queries

```
SERVICE <http://www.ebi.ac.uk/ols-owl/sparql> {  
  GRAPH <http://www.ebi.ac.uk/cmpo> {  
    ?classes ebi-ols:subclasses  
      "phenotype and inheres_in some 'mitotic metaphase'"^^mosi .  
  }  
}
```

- Bit of a hack, need triple stores capable of EL plus nicer SPARQL syntax for writing OWL class expressions

Summary

- We have a pipeline for building an application ontology for cellular microscopy data
 - The biologists build define the ontology terms by annotating data
- Need for templates to guide the annotators e.g. “Increased cytoplasmic actin”
 - EQ (‘actin filament’, ‘present in greater number in organism’)
 - EQE2 (‘actin filament’, ‘localised’, ‘cytoplasm’)
 - EQE2 (‘cytosol’, ‘has extra parts of type’, ‘actin filament’),
 - EQ (‘localisation of actin to cytosol’, ‘increased rate’)
- NCBO Driving biological project grant underway to generify the phenotator approach to OWL ontologies
 - Building ontologies from Google spreadsheets

BMS RI partners

Euro-BioImaging



Jan Ellenberg



Tanja Ninkovic



*Jean-Karim
Heriche*



*Wolfgang
Huber*

Elixir



Gabriella Rustici



Simon Jupp

Infrafrontier



Frauke Neff



*Philipp
Gormanns*

BBMRI



Johan Lundin



Mikael Lundin

Acknowledgments

- WP6 partners
- James Malone, Tony Burdett and Helen Parkinson, EMBL-EBI
- In particular, we wish to thank:
 - Anna Melidoni, Ruth Lovering and Jennifer Rohn (UCL)
 - Beate Neumann and Jean Karim Heriche (EMBL)
 - Bob Van De Water (U. Leiden)
 - Bram Herpers (Ocello)
 - Claudia Lukas (U. Copenhagen)
 - Greg Pau (Genentech)
 - Sylvia Le Dévédec (LUMC)
 - Thomas Walter (Institut Curie)
 - Wies Roosmalen (U. Twente)
 - Zvi Kam (Weizmann Institute)

Thank you for your attention.

