

# *Capturing Provenance for a Linkset of Convenience*

Simon Jupp [jupp@ebi.ac.uk](mailto:jupp@ebi.ac.uk)

Samples, Phenotypes and Ontologies

European Bioinformatics Institute

# RDF 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

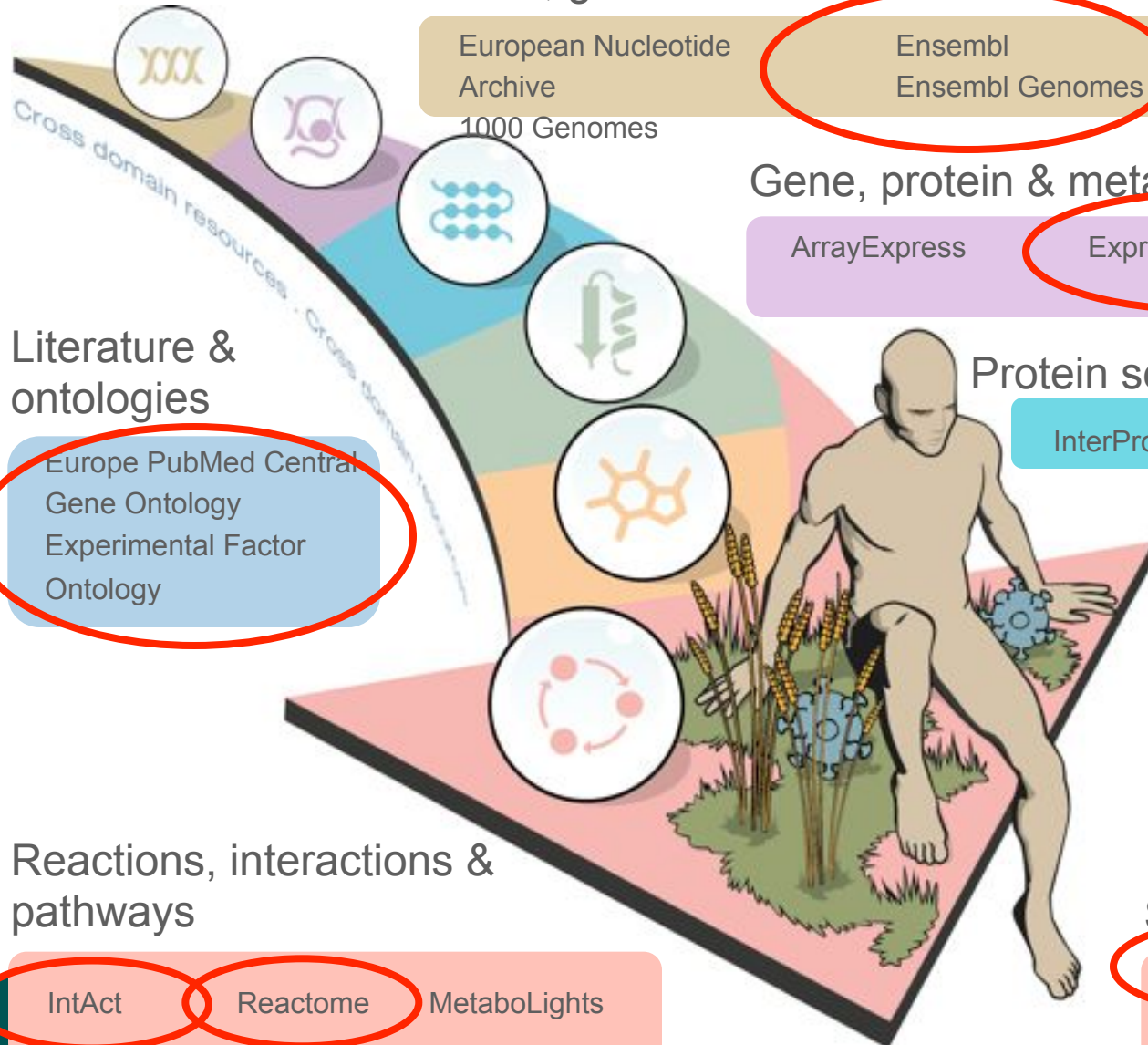
Europe PubMed Central  
Gene Ontology  
Experimental Factor Ontology

## Reactions, interactions & pathways

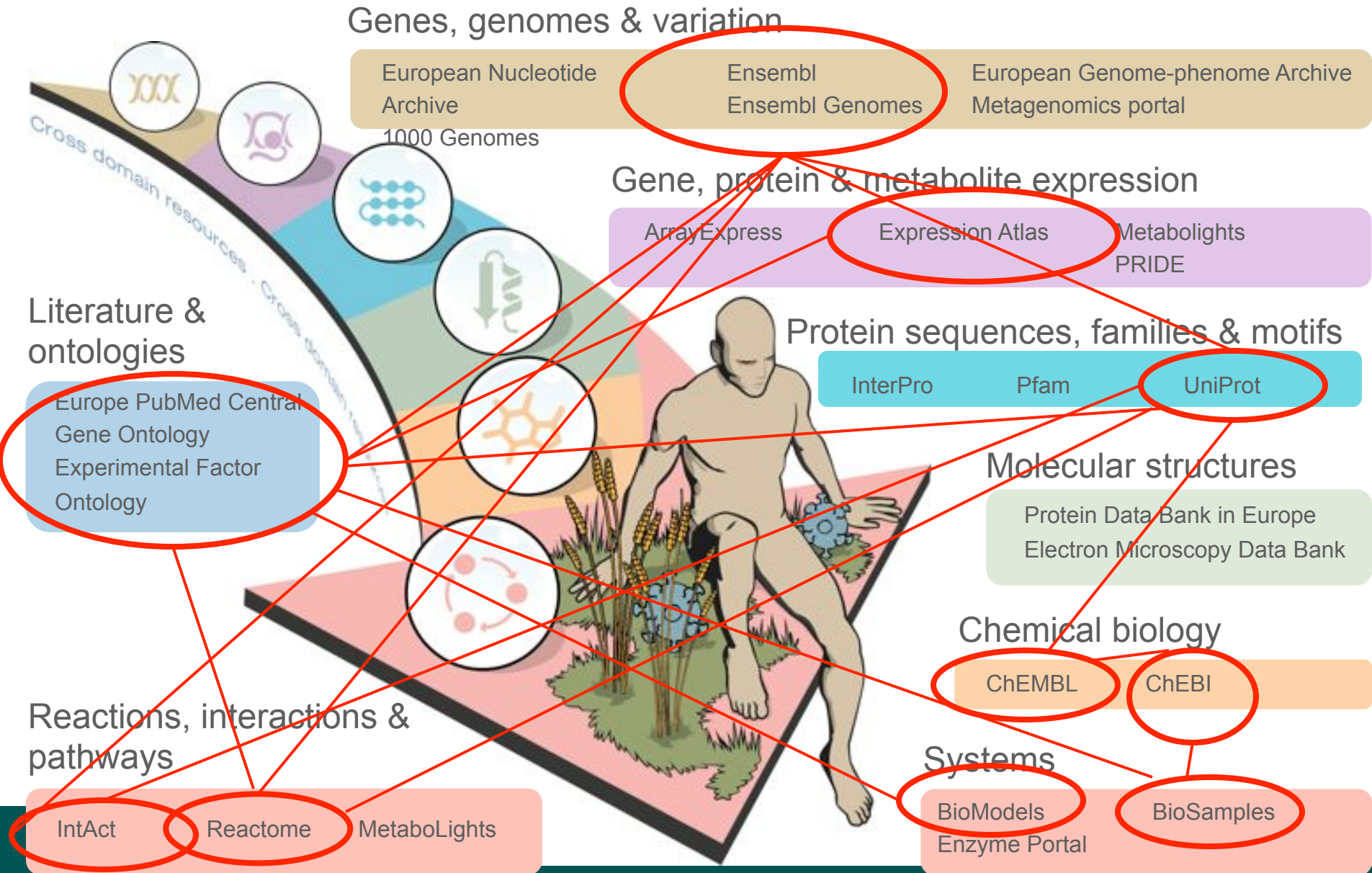
IntAct

Reactome

MetaboLights



# Lots of links, but what are the integration points?



# Integration points

- Most common integration point gene or protein id
  - UniProt or Ensembl id
  - Ontology terms (GO, EFO)
- Most databases include db cross references to support hyperlinking and search
  - Mappings come from multiple source
  - Mappings get imported from other resources

```
Q6GZX4 UniProtKB-ID 001R_FRG3G
Q6GZX4 GI 81941549
Q6GZX4 GI 47960116
Q6GZX4 GI 49237298
Q6GZX4 UniRef100 UniRef100_Q6GZX4
Q6GZX4 UniRef90 UniRef90_Q6GZX4
Q6GZX4 UniRef50 UniRef50_Q6GZX4
Q6GZX4 UniParc UP10000308FD4
Q6GZX4 EMBL AY548484
Q6GZX4 EMBL-CDS AAT09660.1
Q6GZX4 NCBJ_TaxID 654924
Q6GZX4 RefSeq YP_031579.1
Q6GZX4 RefSeq_NT NC_005946.1
Q6GZX4 GeneID 2947773
Q6GZX3 UniProtKB-ID 002L_FRG3G
Q6GZX3 GI 49237299
Q6GZX3 GI 47960117
Q6GZX3 GI 81941548
Q6GZX3 UniRef100 UniRef100_Q6GZX3
Q6GZX3 UniRef90 UniRef90_Q6GZX3
Q6GZX3 UniRef50 UniRef50_Q6GZX3
Q6GZX3 UniParc UP10000308FD5
Q6GZX3 EMBL AY548484
Q6GZX3 EMBL-CDS AAT09661.1
Q6GZX3 NCBJ_TaxID 654924
Q6GZX3 RefSeq YP_031580.1
Q6GZX3 RefSeq_NT NC_005946.1
Q6GZX3 GeneID 2947774
Q197F8 UniProtKB-ID 002R_IIV3
Q197F8 GI 106073503
Q197F8 GI 109287880
Q197F8 GI 123880694
Q197F8 UniRef100 UniRef100_Q197F8
Q197F8 UniRef90 UniRef90_Q197F8
Q197F8 UniRef50 UniRef50_Q197F8
Q197F8 UniParc UP10000083464
Q197F8 EMBL DQ643392
Q197F8 EMBL-CDS ABF82032.1
Q197F8 NCBJ_TaxID 345201
Q197F8 RefSeq YP_654574.1
Q197F8 RefSeq_NT NC_008187.1
Q197F8 GeneID 4156251
Q197F7 UniProtKB-ID 003L_IIV3
Q197F7 GI 106073504
Q197F7 GI 109287881
Q197F7 GI 123880693
Q197F7 UniRef100 UniRef100_Q197F7
Q197F7 UniRef90 UniRef90_Q197F7
```



# Ensembl 2 UniProt example – BRAF gene



Ensembl gene **ENSG00000157764**

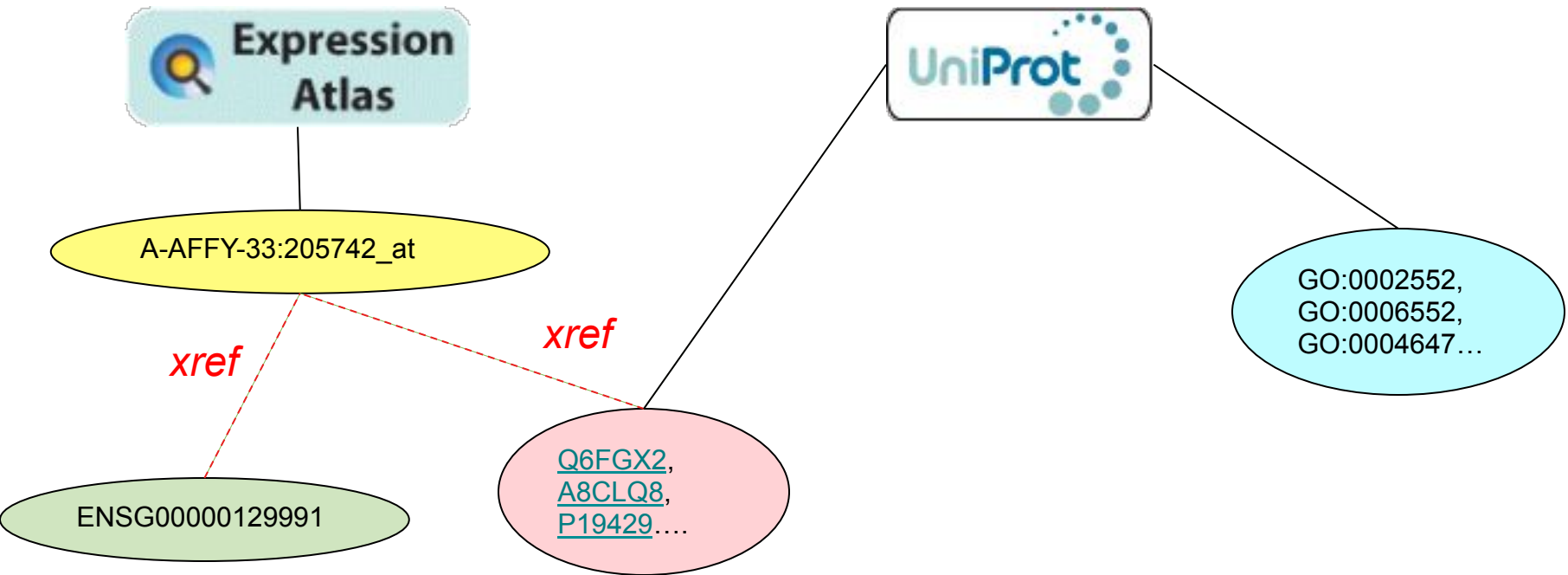
transcriptid	peptide	uniprotacc
BRAF-001	<a href="#">ensemblprotein: ENSP00000288602</a>	ESPF37
BRAF-001	<a href="#">ensemblprotein: ENSP00000288602</a>	Q75MQ8
BRAF-001	<a href="#">ensemblprotein: ENSP00000288602</a>	D7PBN4
BRAF-001	<a href="#">ensemblprotein: ENSP00000288602</a>	<b>P15056</b>
BRAF-002	<a href="#">ensemblprotein: ENSP00000420119</a>	<b>H7C5K3</b>
BRAF-003	<a href="#">ensemblprotein: ENSP00000419060</a>	ESPF37
BRAF-005	<a href="#">ensemblprotein: ENSP00000419033</a>	<b>H7C4S5</b>
BRAF-005	<a href="#">ensemblprotein: ENSP00000419033</a>	ESPF37



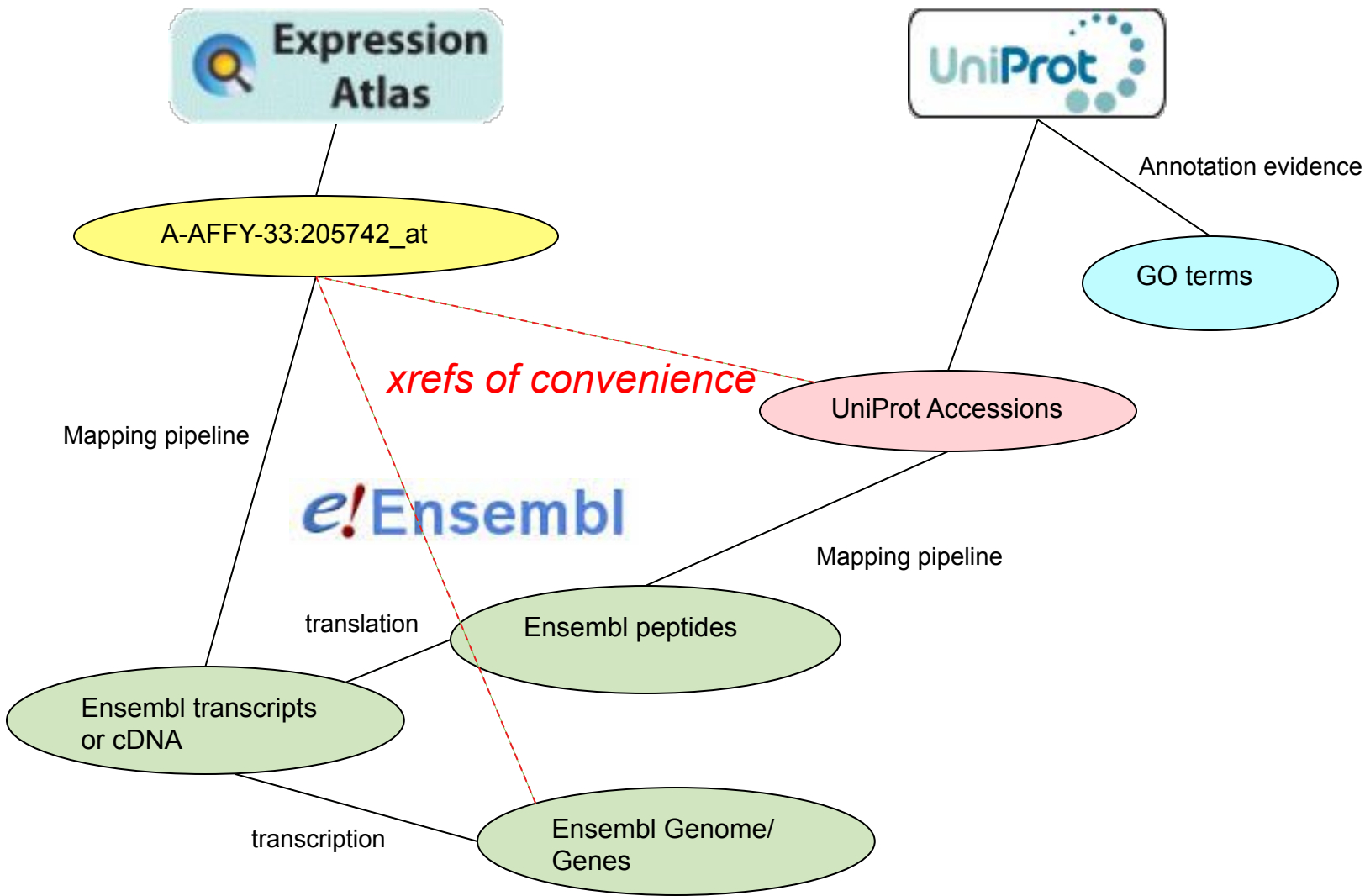
Entry	Entry name	Protein names	Gene names	Organism	Length
<input type="checkbox"/> H7C560	H7C560_HUMAN	Serine/threonine-protein kinase B-r...	<b>BRAF</b>	Homo sapiens (Human)	375
<input type="checkbox"/> <b>P15056</b>	BRAF_HUMAN	Serine/threonine-protein kinase B-r...	<b>BRAF, BRAF1, RAFB1</b>	Homo sapiens (Human)	766
<input type="checkbox"/> <b>H7C4S5</b>	H7C4S5_HUMAN	Serine/threonine-protein kinase B-r...	<b>BRAF</b>	Homo sapiens (Human)	102
<input type="checkbox"/> <b>H7C5K3</b>	H7C5K3_HUMAN	Serine/threonine-protein kinase B-r...	<b>BRAF</b>	Homo sapiens (Human)	194

# Querying across resources

*“Show expression for ENSG00000129991 (TNNI3) with its GO annotations from Uniprot”*

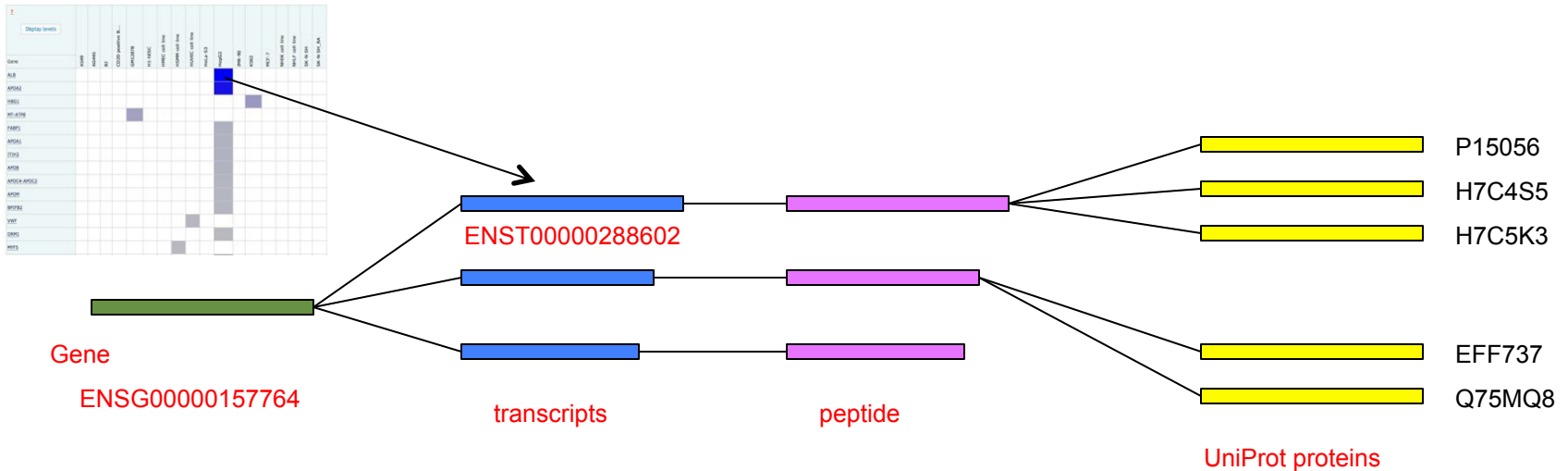


# Convenience links derived from stronger links



# Need to be careful with Xrefs

RNASeq experiment shows expression of transcript **ENST00000288602**



Would you expect this experiments to be returned in a search for Q75MQ8?

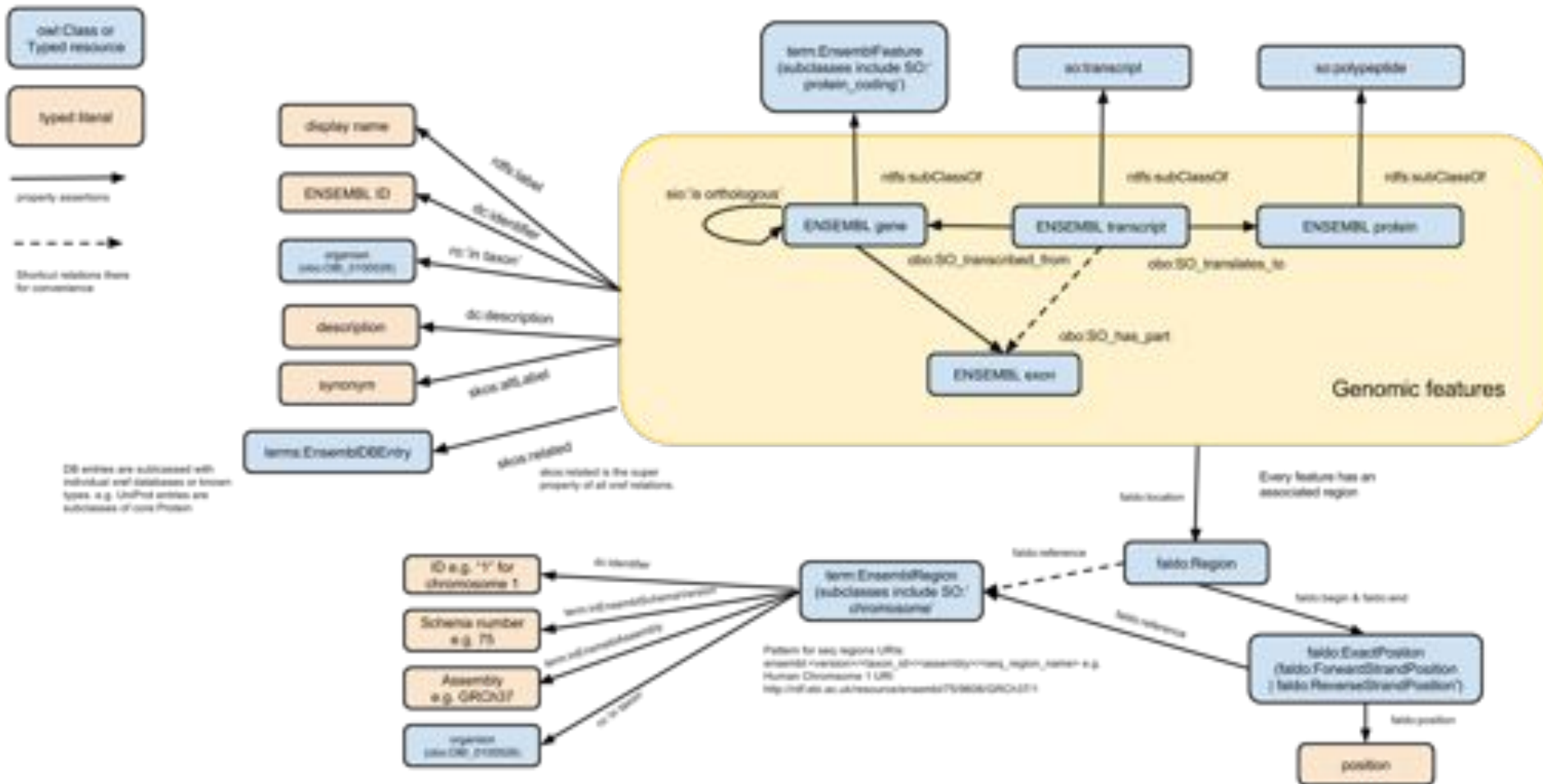
- probably yes

Is there evidence for the expression on this protein?

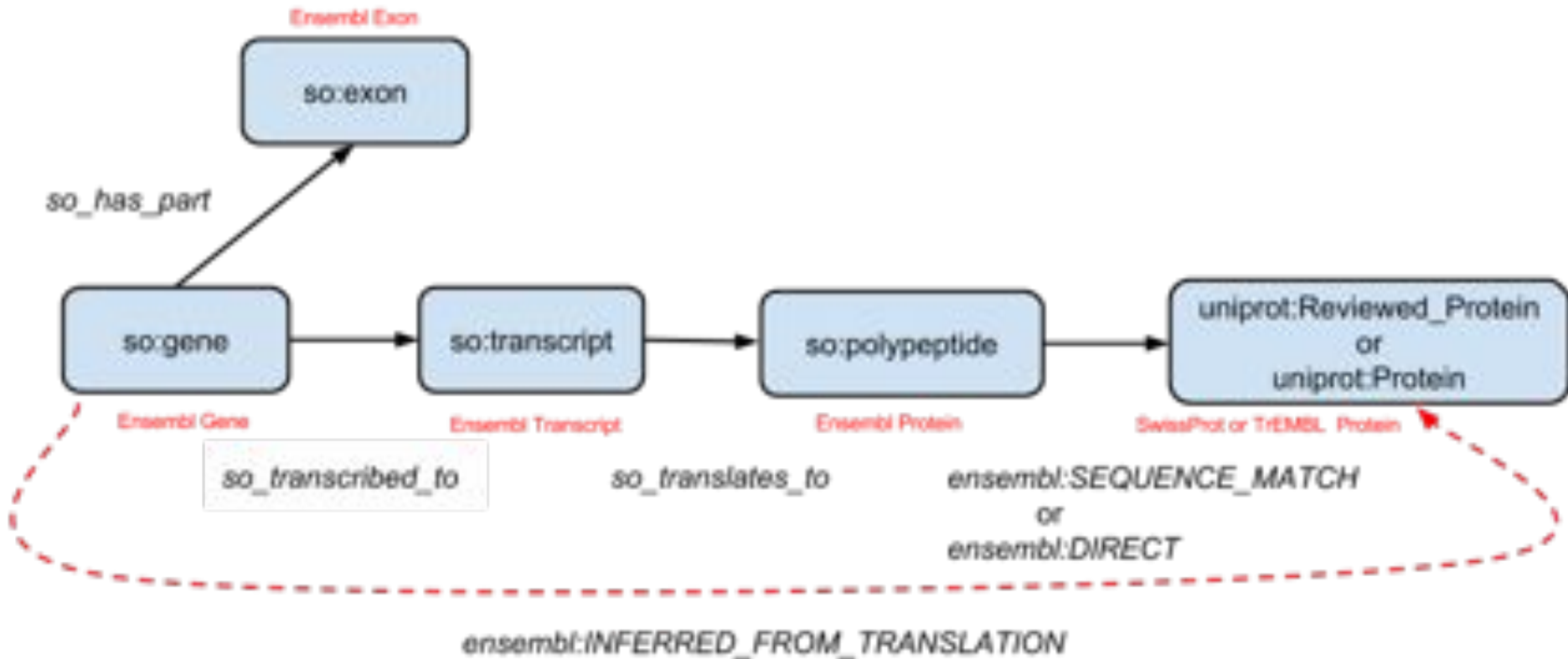
- possibly not



# Ensembl core RDF schema



# Ensembl Gene to UniProt protein

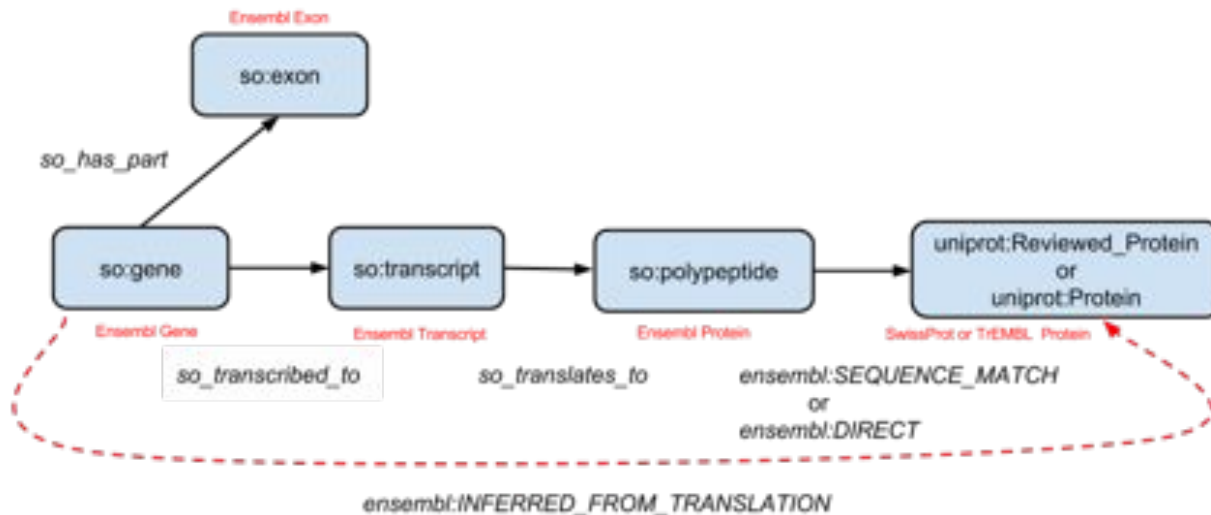


*How can we capture that the link of convenience is derived from the longer chain of links above?*

# VoID

- Vocabulary for describing linked datasets
- Describe partitions/subsets in a dataset
  - Class partitions e.g. all triples where subject of a given type
  - Property partitions e.g. all triples where a given property is used
- VoID linksets
  - Description of links between datasets (or partitions)
  - We can use VoID to cut a large dataset like ensembl into smaller datasets based on partitions
  - We can use linksets to describe how these subsets relate to each within Ensembl, and to external datasets like UniProt

# Describing subsets and linksets with VoID



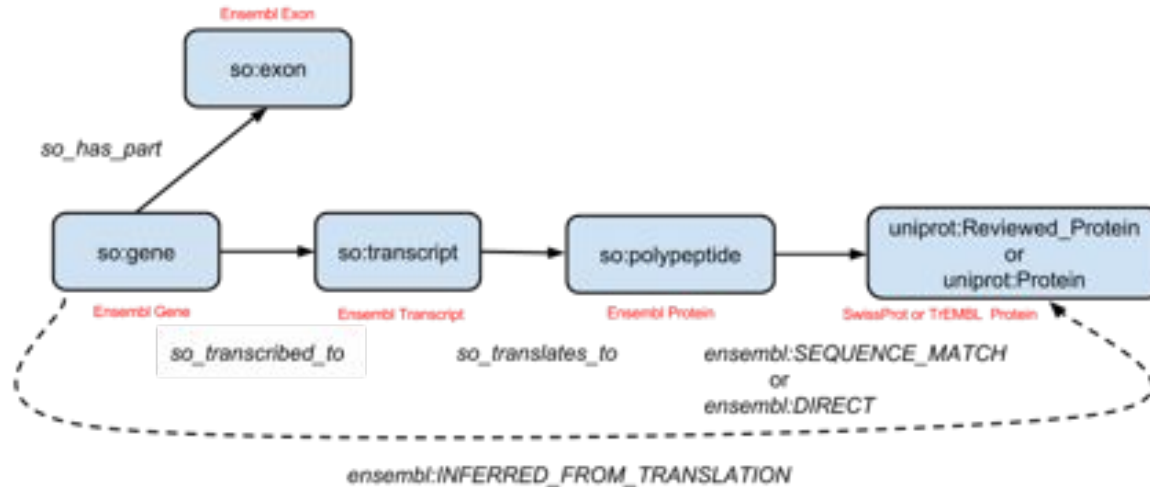
## Subsets

- **All Ensembl gene subset:** :gene\_partition *void:class* ensembl:Gene
- **All Ensembl transcript subset:** :transcript\_partition *void:class* ensembl:Transcript
- **All Ensembl protein subset:** :protein\_partition *void:class* ensembl:Protein
- **All UniProt reviewed protein subset:** :uniprot\_reviewed\_partition *void:class* uniprot:Reviewed\_Protein

## Example Linkset

- **Gene Transcript linkset:**  
:gene\_transcript\_partition a *void:Linkset* ;  
*void:linkPredicate* so:transcribed\_to;  
*void:subjectTarget* :gene\_partition ;  
*void:objectTarget* :transcript\_partition
- Repeat for transcript-protein, protein-uniprot, gene-uniprot

# Derived Linksets



## Linkset of convenience

- A linkset derived from other linksets

:gene\_to\_uniprot\_protein\_linkset *prov:wasDerivedFrom*

:gene\_to\_transcript, :transcript\_to\_peptide, :peptide\_to\_uniprot ;

*void:subjectTarget* :gene\_partition ;

*void:objectTarget* :reviewed\_uniprot\_partition

***prov:wasDerivedFrom*** - “A derivation is a transformation of an entity into another, an update of an entity resulting in a new one, or the **construction of a new entity based on a pre-existing entity.**”

# Implementation – 1013 linksets defined

<http://tinyurl.com/pswgo8l>

**About:** protein\_coding INFERRED\_FROM\_TRANSLATION Uniprot/SWISSPROT linkset



<http://rdf.ebi.ac.uk/dataset/ensembl/76/linkset-3d94016fef4d6626d4540a5eeeb1d5e8>

**Type:** Linkset

more types...

## Related to

### **wasDerivedFrom** (Linkset)

- protein NONE Uniprot/SWISSPROT linkset
- protein CHECKSUM Uniprot/SWISSPROT linkset
- protein DEPENDENT Uniprot/SWISSPROT linkset
- protein DIRECT Uniprot/SWISSPROT linkset
- protein SEQUENCE\_MATCH Uniprot/SWISSPROT linkset
- linkset-transcript-SO\_translates\_to-protein
- linkset-transcript-SO\_transcribed\_from-protein\_coding

### **linkPredicate** (ObjectProperty)

- inferred mapping from translation

### **objectTarget**

- SWISSPROT-dataset-partition

### **subjectTarget**

- protein\_coding-dataset-partition

# Generic query to explore a dataset

```
PREFIX void:<http://rdfs.org/ns/void#>
```

```
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
```

```
SELECT DISTINCT ?subject ?subject_name ?object ?object_name
```

```
WHERE {
```

```
    ?linkset a void:Linkset ;
```

```
    void:objectTarget [void:class ?object ] ;
```

```
    void:subjectTarget [void:class ?subject ] .
```

```
    ?subject rdfs:label ?subject_name .
```

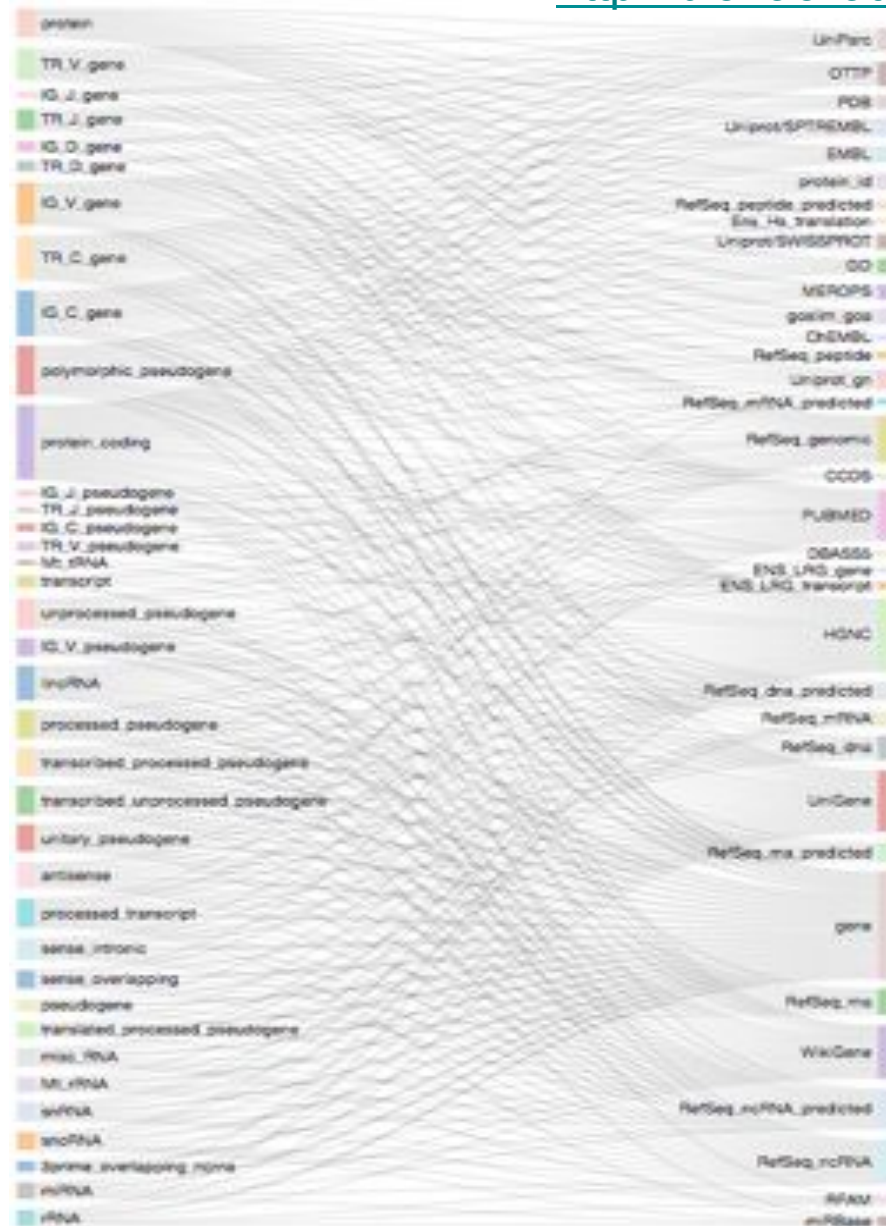
```
    ?object rdfs:label ?object_name .
```

```
}
```

# Visualising Linksets

<http://tinyurl.com/p3gw7of>

<http://biohackathon.org/d3sparql/>





# Querying convenience links

```
PREFIX void:<http://rdfs.org/ns/void#>  
PREFIX prov:<http://www.w3.org/ns/prov#>  
PREFIX ensemblterms: <http://rdf.ebi.ac.uk/terms/ensembl/>  
PREFIX core: <http://purl.uniprot.org/core/>
```

```
SELECT ?subject ?rel ?object ?derivedLinksetRelation WHERE  
{  
  ?subject ?rel ?object .  
  ?subject a ?subjectType .  
  ?object a ?objectType .  
  ?linkset void:linkPredicate ?rel .  
  ?linkset prov:wasDerivedFrom ?derivedLinkset .  
  ?derivedLinkset void:linkPredicate ?derivedLinksetRelation .
```

```
VALUES ?subjectType {ensemblterms:protein_coding}
```

```
VALUES ?objectType {core:Reviewed_Protein core:Protein}
```

← Ensembl gene

← UniProt Protein

```
}
```

# Querying convenience links

subject	rel	object	derivedLinksetRelation
ensembl:ENSACAG0000000004	ensemblterms:INFERRED_FROM_TRANSLATION	<http://identifiers.org/uniprot/H9G367>	ensemblterms:SEQUENCE_MATCH
ensembl:ENSACAG000000001545	ensemblterms:INFERRED_FROM_TRANSLATION	<http://purl.uniprot.org/uniprot/H9G5C3>	ensemblterms:SEQUENCE_MATCH
ensembl:ENSACAG000000001545	ensemblterms:INFERRED_FROM_TRANSLATION	<http://identifiers.org/uniprot/H9G5C3>	ensemblterms:SEQUENCE_MATCH
ensembl:ENSACAG000000001305	ensemblterms:INFERRED_FROM_TRANSLATION	<http://purl.uniprot.org/uniprot/H9G4X0>	ensemblterms:SEQUENCE_MATCH
ensembl:ENSACAG000000001305	ensemblterms:INFERRED_FROM_TRANSLATION	<http://identifiers.org/uniprot/H9G4X0>	ensemblterms:SEQUENCE_MATCH
ensembl:ENSACAG000000002395	ensemblterms:INFERRED_FROM_TRANSLATION	<http://identifiers.org/uniprot/H9G672>	ensemblterms:SEQUENCE_MATCH
ensembl:ENSACAG000000001272	ensemblterms:INFERRED_FROM_TRANSLATION	<http://purl.uniprot.org/uniprot/H9GVS0>	ensemblterms:SEQUENCE_MATCH
ensembl:ENSACAG000000001005	ensemblterms:INFERRED_FROM_TRANSLATION	<http://purl.uniprot.org/uniprot/H9G413>	ensemblterms:SEQUENCE_MATCH
ensembl:ENSACAG000000001005	ensemblterms:INFERRED_FROM_TRANSLATION	<http://identifiers.org/uniprot/H9G413>	ensemblterms:SEQUENCE_MATCH
ensembl:ENSACAG000000001608	ensemblterms:INFERRED_FROM_TRANSLATION	<http://identifiers.org/uniprot/H9G5P3>	ensemblterms:SEQUENCE_MATCH
ensembl:ENSACAG000000000004	ensemblterms:INFERRED_FROM_TRANSLATION	<http://identifiers.org/uniprot/H9G367>	ensemblterms:DIRECT
ensembl:ENSACAG000000001545	ensemblterms:INFERRED_FROM_TRANSLATION	<http://purl.uniprot.org/uniprot/H9G5C3>	ensemblterms:DIRECT
ensembl:ENSACAG000000001545	ensemblterms:INFERRED_FROM_TRANSLATION	<http://identifiers.org/uniprot/H9G5C3>	ensemblterms:DIRECT
ensembl:ENSACAG000000001305	ensemblterms:INFERRED_FROM_TRANSLATION	<http://purl.uniprot.org/uniprot/H9G4X0>	ensemblterms:DIRECT
ensembl:ENSACAG000000001305	ensemblterms:INFERRED_FROM_TRANSLATION	<http://identifiers.org/uniprot/H9G4X0>	ensemblterms:DIRECT
ensembl:ENSACAG000000002395	ensemblterms:INFERRED_FROM_TRANSLATION	<http://identifiers.org/uniprot/H9G672>	ensemblterms:DIRECT

# Utility of the linksets

- Provenance based schema for publishing and sharing linksets across resources
- Tune query precision (scientific lenses)
- Used to assist SPARQL autocomplete widgets
- Potential for use by query optimizers to improve query plans for federated querying

# Is VoID+PROV the right vocabulary for this?

- It's possible that linksets are still over generalisations
  - Semantics aren't well defined
  - Some cases require more fine grained solution (consider % sequence similarity)
- Alternative approaches
  - Nanopublications, Open Biological Associations (OBAN)
  - Property singleton pattern
  - OWL modeling
- Open to other ideas???

# Acknowledgement

- Alasdair Gray and James Malone
- Kieron Taylor and Ensembl development team
- EBI RDF platform
  - Andy Jenkinson, Mark Davies, Marco Brandizi, Sarala Wimalaratne, Leyla Garcia, Jerven Bolleman