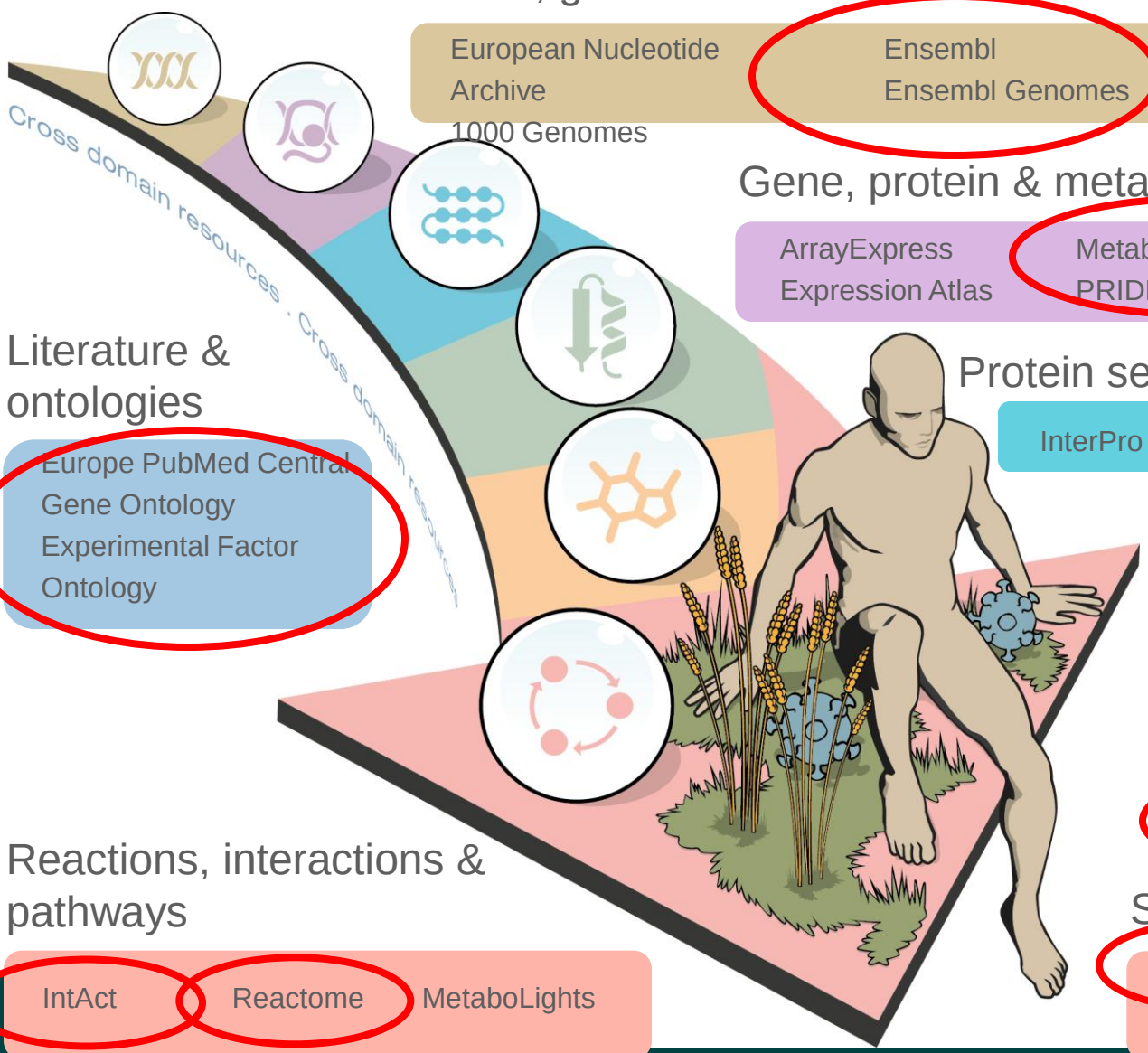
Capturing Provenance for a Linkset of Convenience

Simon Jupp 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

Literature & ontologies

Europe PubMed Central
Gene Ontology
Experimental Factor Ontology

Protein sequences, families & motifs

InterPro
Pfam
UniProt

Molecular structures

Protein Data Bank in Europe
Electron Microscopy Data Bank

Chemical biology

ChEMBL
ChEBI

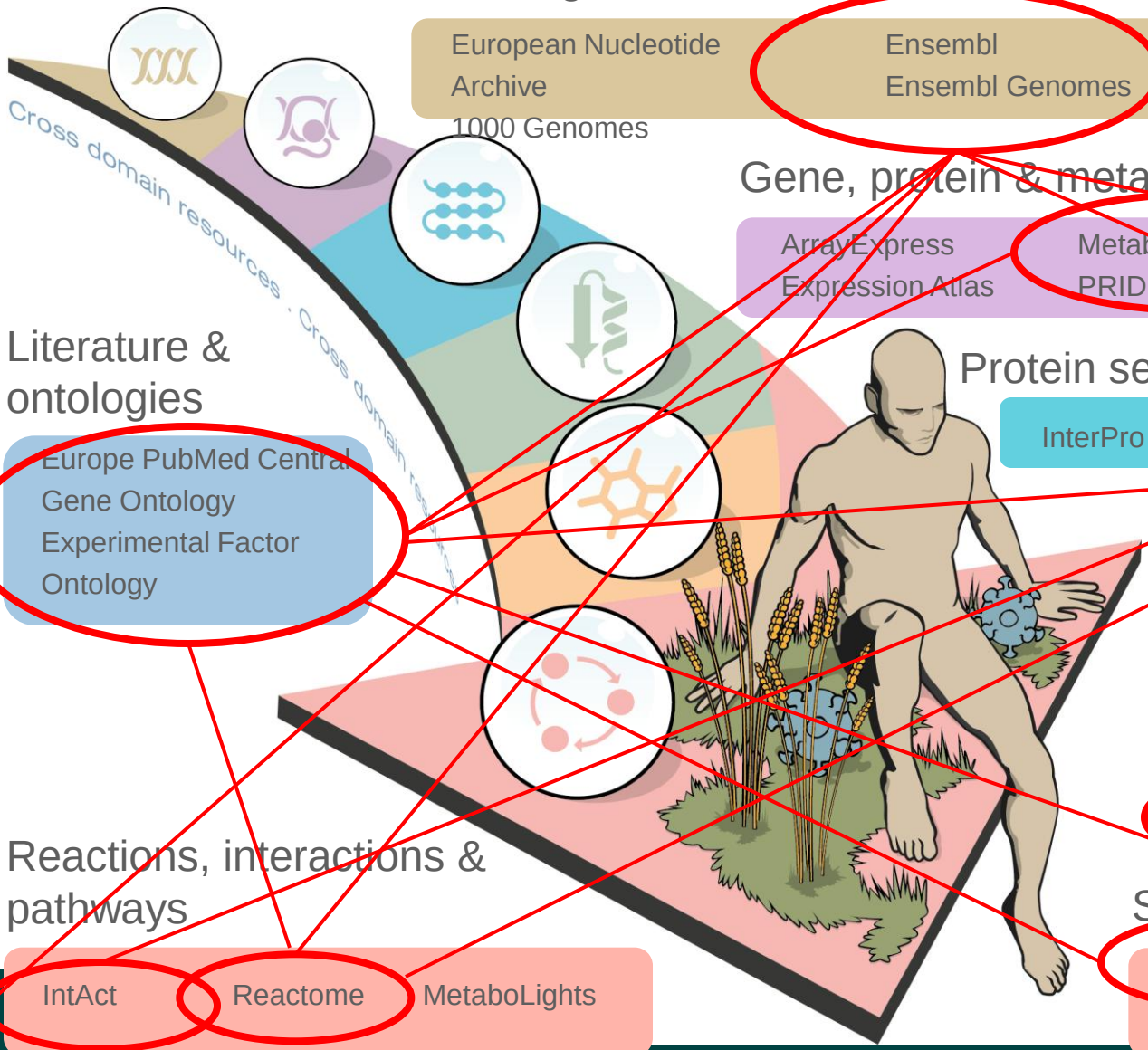
Reactions, interactions & pathways

IntAct
Reactome
MetaboLights

Systems

BioModels
Enzyme Portal
BioSamples

Lots of links, but what are the integration points?



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

Literature & ontologies

Europe PubMed Central
Gene Ontology
Experimental Factor Ontology

Protein sequences, families & motifs

InterPro
Pfam
UniProt

Molecular structures

Protein Data Bank in Europe
Electron Microscopy Data Bank

Chemical biology

ChEMBL
ChEBI

Reactions, interactions & pathways

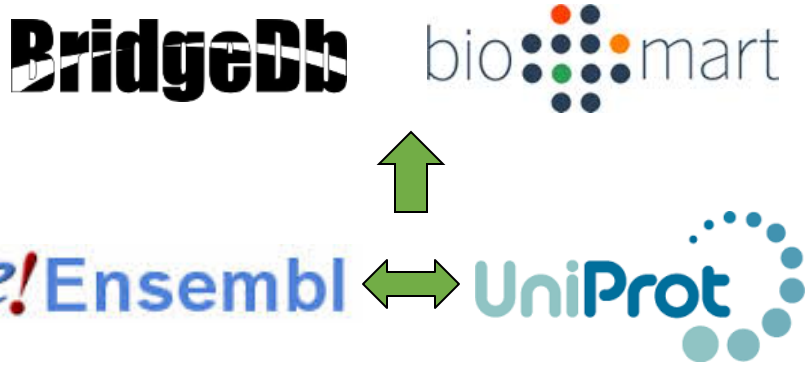
IntAct
Reactome
MetaboLights

Systems

BioModels
Enzyme Portal
BioSamples

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      47060116
Q6GZX4 GI      49237298
Q6GZX4 UniRef100      UniRef100_Q6GZX4
Q6GZX4 UniRef90       UniRef90_Q6GZX4
Q6GZX4 UniRef50       UniRef50_Q6GZX4
Q6GZX4 UniParc        UPI00003B0FD4
Q6GZX4 EMBL    AY548484
Q6GZX4 EMBL-CDS       AAT09660.1
Q6GZX4 NCBI_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      47060117
Q6GZX3 GI      81941548
Q6GZX3 UniRef100      UniRef100_Q6GZX3
Q6GZX3 UniRef90       UniRef90_Q6GZX3
Q6GZX3 UniRef50       UniRef50_Q6GZX3
Q6GZX3 UniParc        UPI00003B0FD5
Q6GZX3 EMBL    AY548484
Q6GZX3 EMBL-CDS       AAT09661.1
Q6GZX3 NCBI_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      123808694
Q197F8 UniRef100      UniRef100_Q197F8
Q197F8 UniRef90       UniRef90_Q197F8
Q197F8 UniRef50       UniRef50_Q197F8
Q197F8 UniParc        UPI0000D83464
Q197F8 EMBL    DQ643392
Q197F8 EMBL-CDS       ABF82032.1
Q197F8 NCBI_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      123808693
Q197F7 UniRef100      UniRef100_Q197F7
Q197F7 UniRef90       UniRef90_Q197F7
```

Ensembl 2 UniProt example – BRAF gene

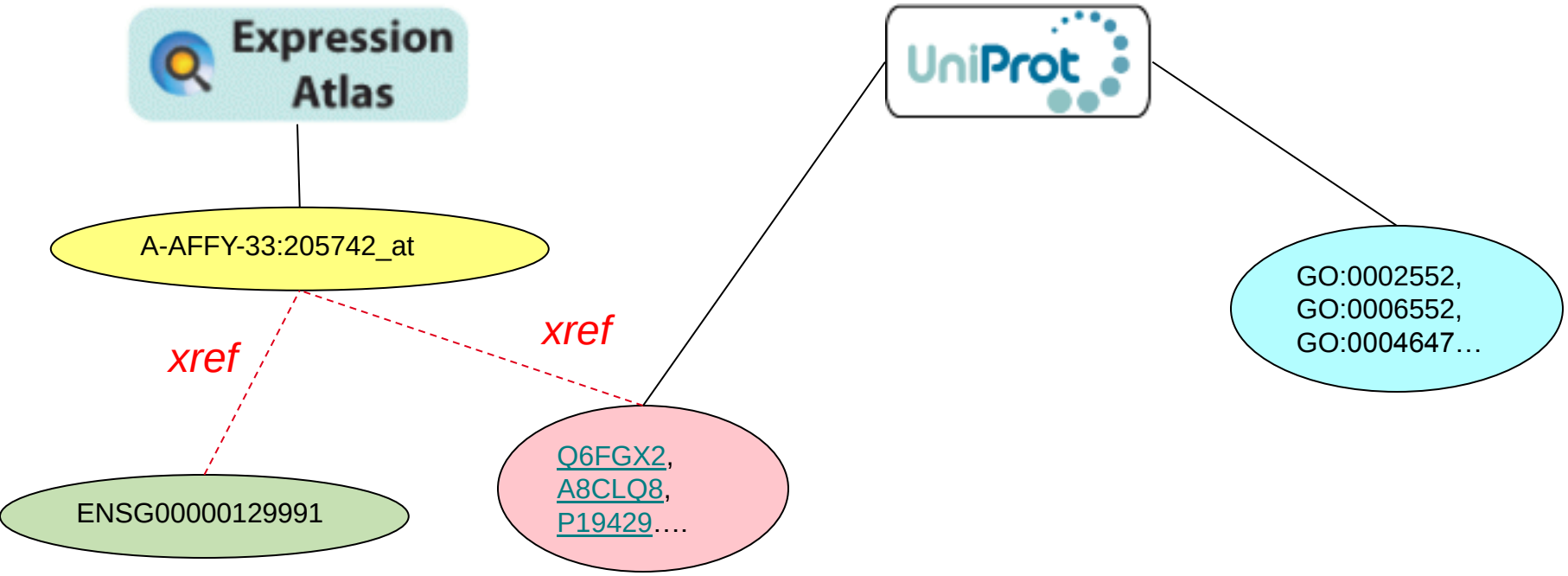
Ensembl gene ENSG00000157764

transcriptid	peptide	uniprotacc
BRAF-001	ensemblprotein:ENSP00000288602	E5FF37
BRAF-001	ensemblprotein:ENSP00000288602	Q75MQ8
BRAF-001	ensemblprotein:ENSP00000288602	D7PBN4
BRAF-001	ensemblprotein:ENSP00000288602	P15056
BRAF-002	ensemblprotein:ENSP00000420119	H7C5K3
BRAF-003	ensemblprotein:ENSP00000419060	E5FF37
BRAF-005	ensemblprotein:ENSP00000418033	H7C4S5
BRAF-005	ensemblprotein:ENSP00000418033	E5FF37

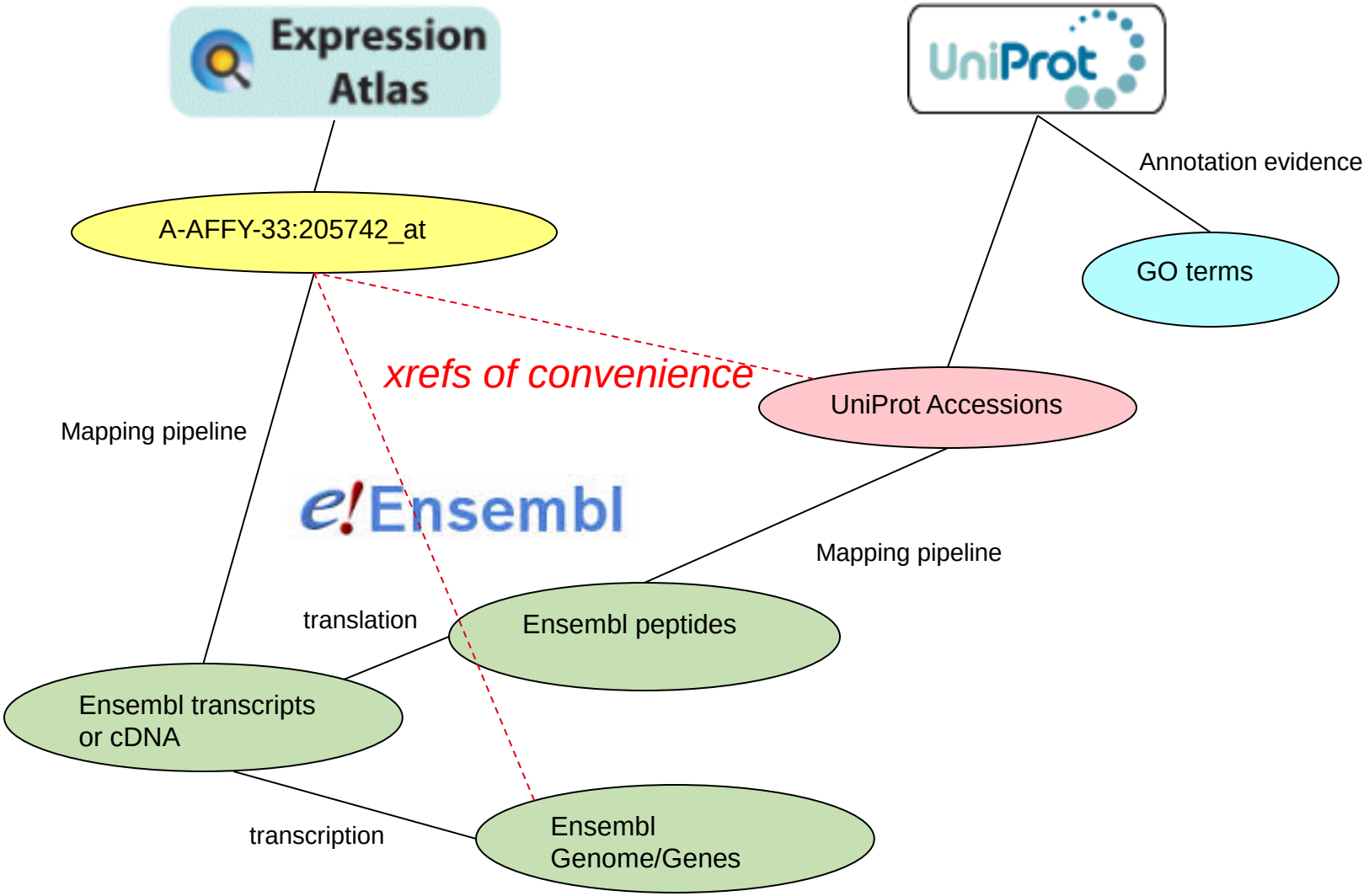
Entry	Entry name	Protein names	Gene names	Organism	Length
H7C560	H7C560_HUMAN	**Serine/threonine-protein kinase B-r...**	**BRAF**	Homo sapiens (Human)	375
P15056	BRAF_HUMAN	**Serine/threonine-protein kinase B-r...**	**BRAF**, BRAF1, RAFB1	Homo sapiens (Human)	766
H7C4S5	H7C4S5_HUMAN	**Serine/threonine-protein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

Querying across resources

"Show expression for ENSG00000129991 (TNNI3) with its GO annotations from Uniprot"

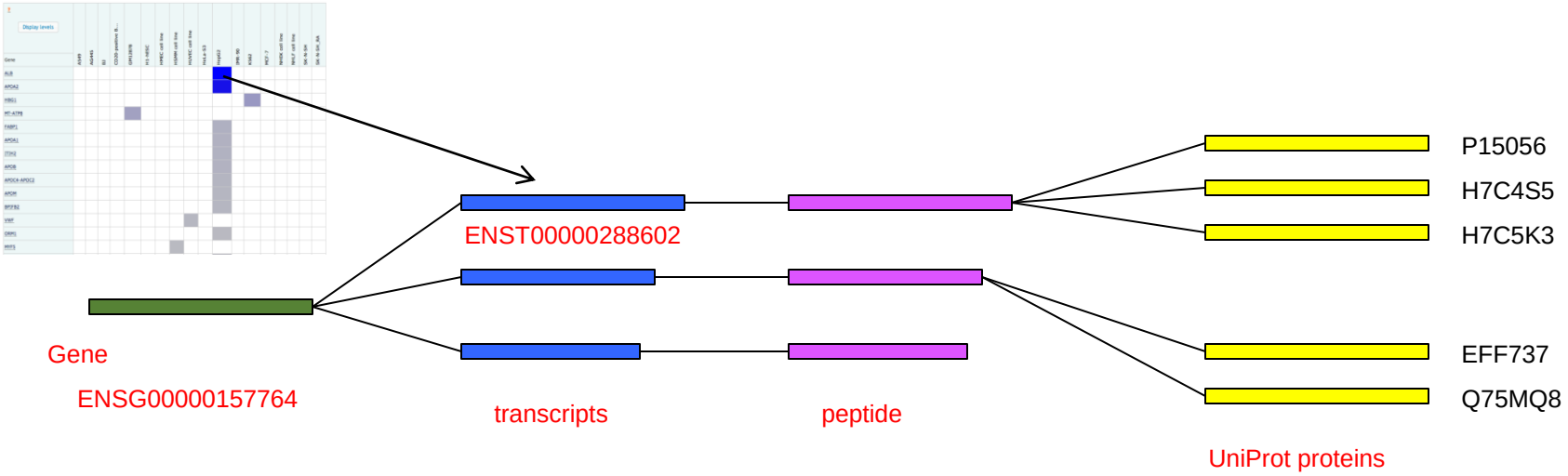


Convenience links derived from stronger links



Need to be careful with Xrefs

RNASeq experiment shows expression of transcript ENST00000288602



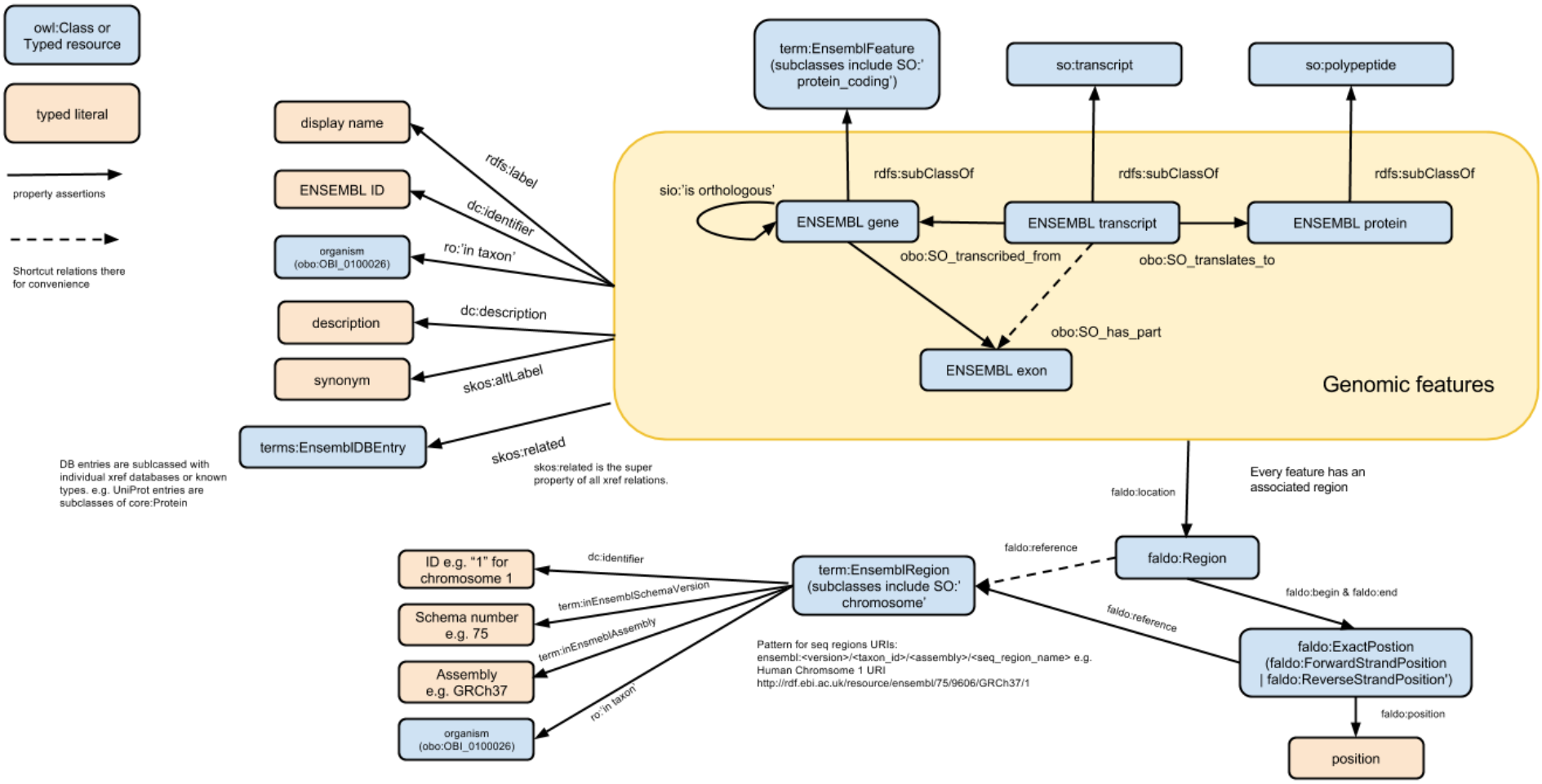
Would you expect this experiments to be returned in a search for Q75MQ8?

- probably yes

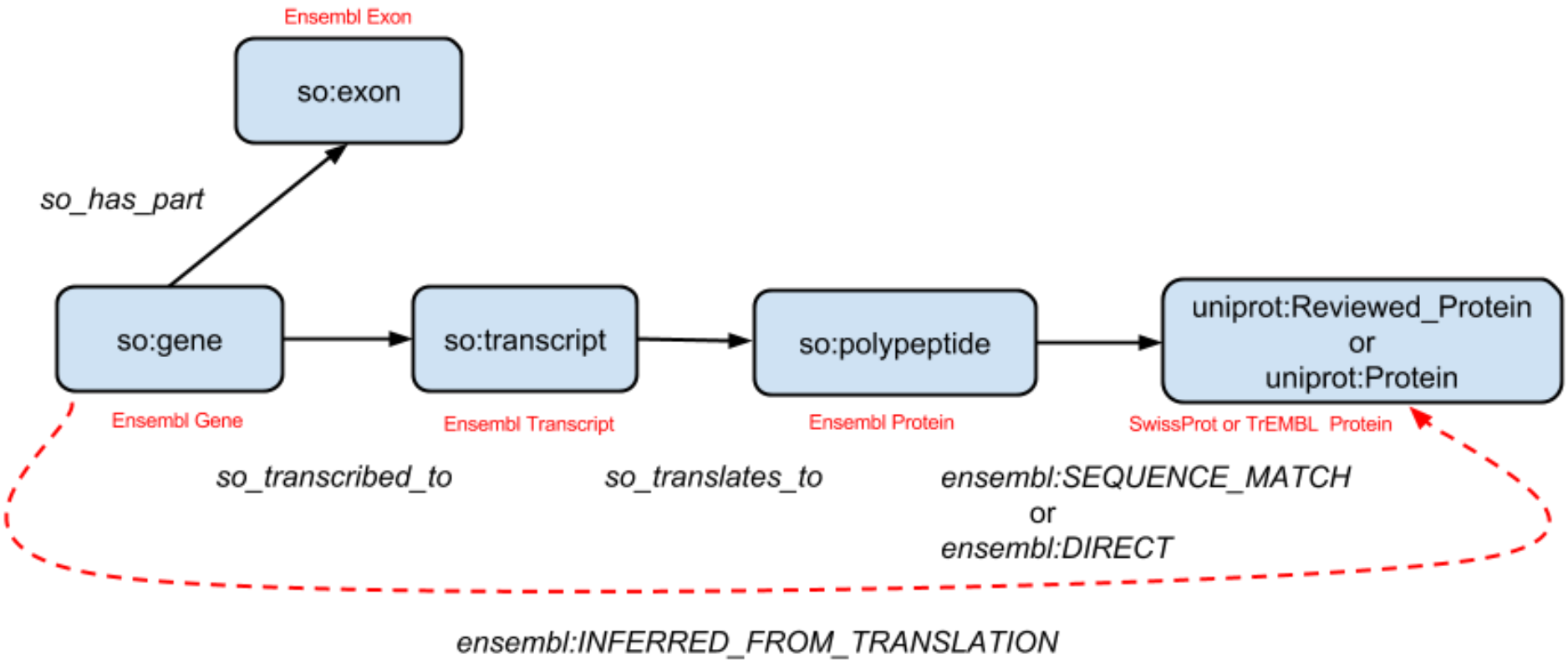
Is there evidence for the expression on this protein?

- possibly not

Ensembl core RDF schema



Ensembl Gene to UniProt protein

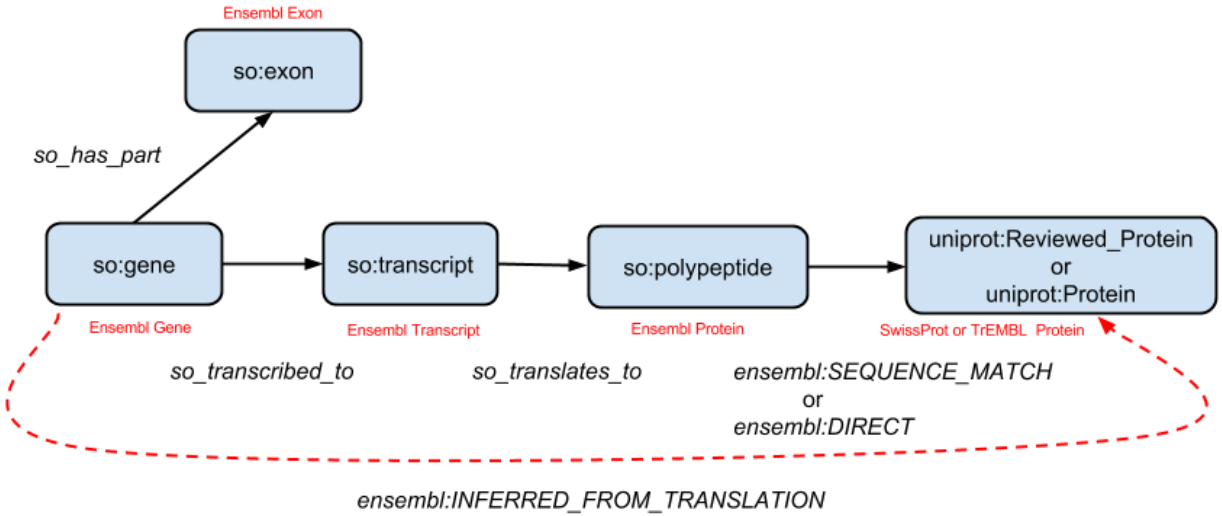


How can we capture that the link of convenience is derived from the longer chain of links above?

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 releate 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 ;*

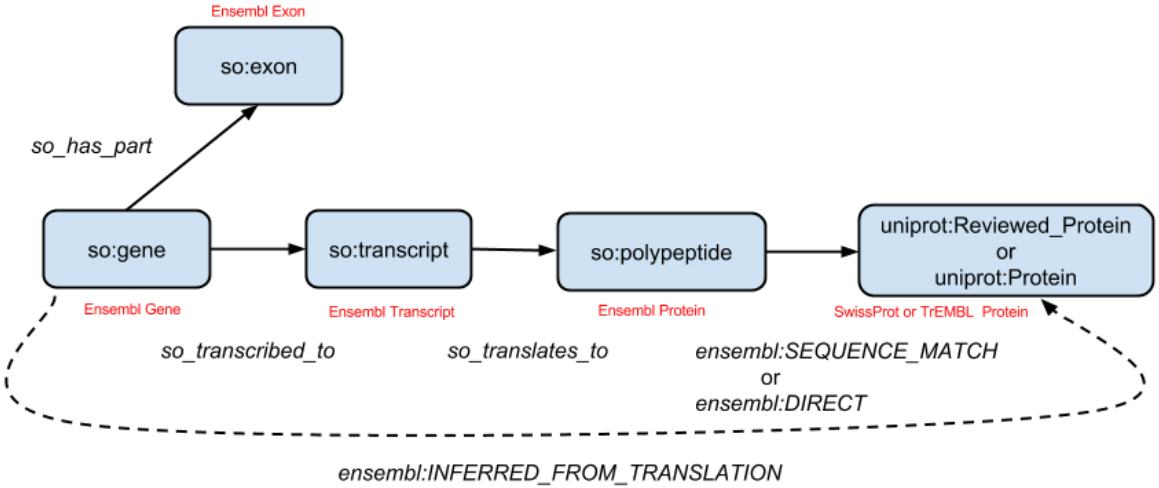
  void:linkPredicate so:transcribed_to;

  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_linkest *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-3d94016fef4d6626d4540a5eaeb1d5e8

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

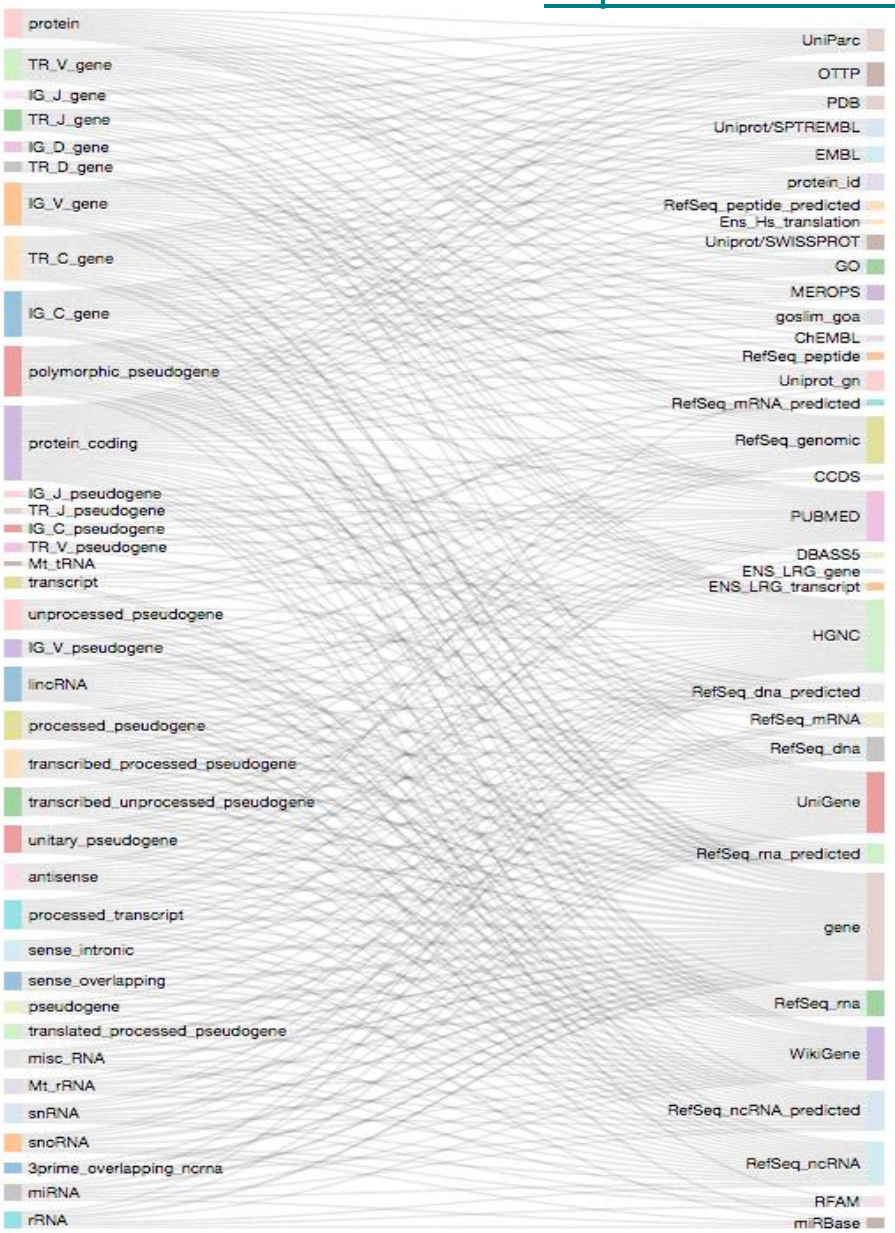
```
PREFIX void:<http://rdfs.org/ns/void#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>

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:ENSACAG00000000004	ensembleterms:INFERRED_FROM_TRANSLATION	<http://identifiers.org/uniprot/H9G367>	ensembleterms:SEQUENCE_MATCH
ensembl:ENSACAG00000001545	ensembleterms:INFERRED_FROM_TRANSLATION	<http://purl.uniprot.org/uniprot/H9G5C3>	ensembleterms:SEQUENCE_MATCH
ensembl:ENSACAG00000001545	ensembleterms:INFERRED_FROM_TRANSLATION	<http://identifiers.org/uniprot/H9G5C3>	ensembleterms:SEQUENCE_MATCH
ensembl:ENSACAG00000001305	ensembleterms:INFERRED_FROM_TRANSLATION	<http://purl.uniprot.org/uniprot/H9G4X0>	ensembleterms:SEQUENCE_MATCH
ensembl:ENSACAG00000001305	ensembleterms:INFERRED_FROM_TRANSLATION	<http://identifiers.org/uniprot/H9G4X0>	ensembleterms:SEQUENCE_MATCH
ensembl:ENSACAG00000002395	ensembleterms:INFERRED_FROM_TRANSLATION	<http://identifiers.org/uniprot/H9G672>	ensembleterms:SEQUENCE_MATCH
ensembl:ENSACAG00000001272	ensembleterms:INFERRED_FROM_TRANSLATION	<http://purl.uniprot.org/uniprot/H9GVG0>	ensembleterms:SEQUENCE_MATCH
ensembl:ENSACAG00000001005	ensembleterms:INFERRED_FROM_TRANSLATION	<http://purl.uniprot.org/uniprot/H9G4J3>	ensembleterms:SEQUENCE_MATCH
ensembl:ENSACAG00000001005	ensembleterms:INFERRED_FROM_TRANSLATION	<http://identifiers.org/uniprot/H9G4J3>	ensembleterms:SEQUENCE_MATCH
ensembl:ENSACAG00000001908	ensembleterms:INFERRED_FROM_TRANSLATION	<http://identifiers.org/uniprot/H9G5P3>	ensembleterms:SEQUENCE_MATCH
ensembl:ENSACAG00000000004	ensembleterms:INFERRED_FROM_TRANSLATION	<http://identifiers.org/uniprot/H9G367>	ensembleterms:DIRECT
ensembl:ENSACAG00000001545	ensembleterms:INFERRED_FROM_TRANSLATION	<http://purl.uniprot.org/uniprot/H9G5C3>	ensembleterms:DIRECT
ensembl:ENSACAG00000001545	ensembleterms:INFERRED_FROM_TRANSLATION	<http://identifiers.org/uniprot/H9G5C3>	ensembleterms:DIRECT
ensembl:ENSACAG00000001305	ensembleterms:INFERRED_FROM_TRANSLATION	<http://purl.uniprot.org/uniprot/H9G4X0>	ensembleterms:DIRECT
ensembl:ENSACAG00000001305	ensembleterms:INFERRED_FROM_TRANSLATION	<http://identifiers.org/uniprot/H9G4X0>	ensembleterms:DIRECT
ensembl:ENSACAG00000002395	ensembleterms:INFERRED_FROM_TRANSLATION	<http://identifiers.org/uniprot/H9G672>	ensembleterms: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