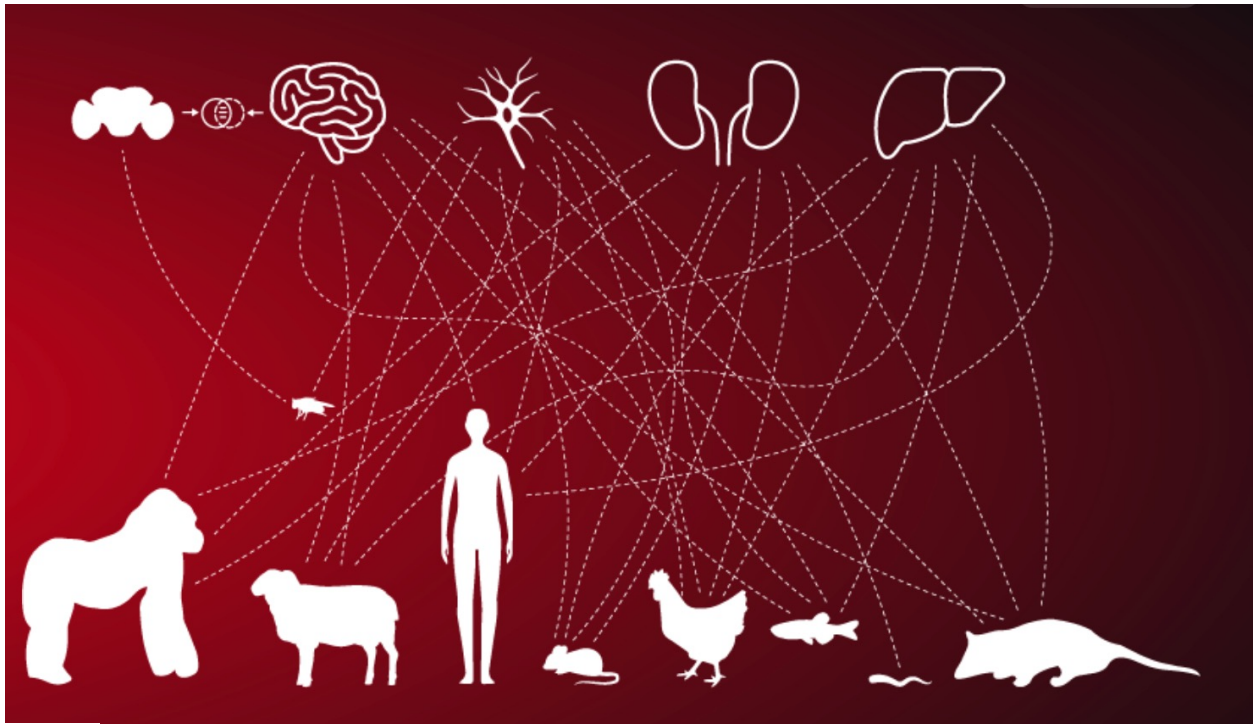


Basel Computational Biology Conference

Swiss Institute of Bioinformatics

Overview of Bgee

Marc Robinson-Rechavi

The Bgee suite: leveraging standardized and comparable transcriptomics data across animal

@bgeedb@genomic.social
@marcrr@ecoevo.social
@bgee.org
@marcrr.bsky.social

www.bgee.org

elixir SWITZERLAND

UNIL | Université de Lausanne Département d'écologie et évolution

Bgee Gene Expression Evolution

Bgee is the work of a team

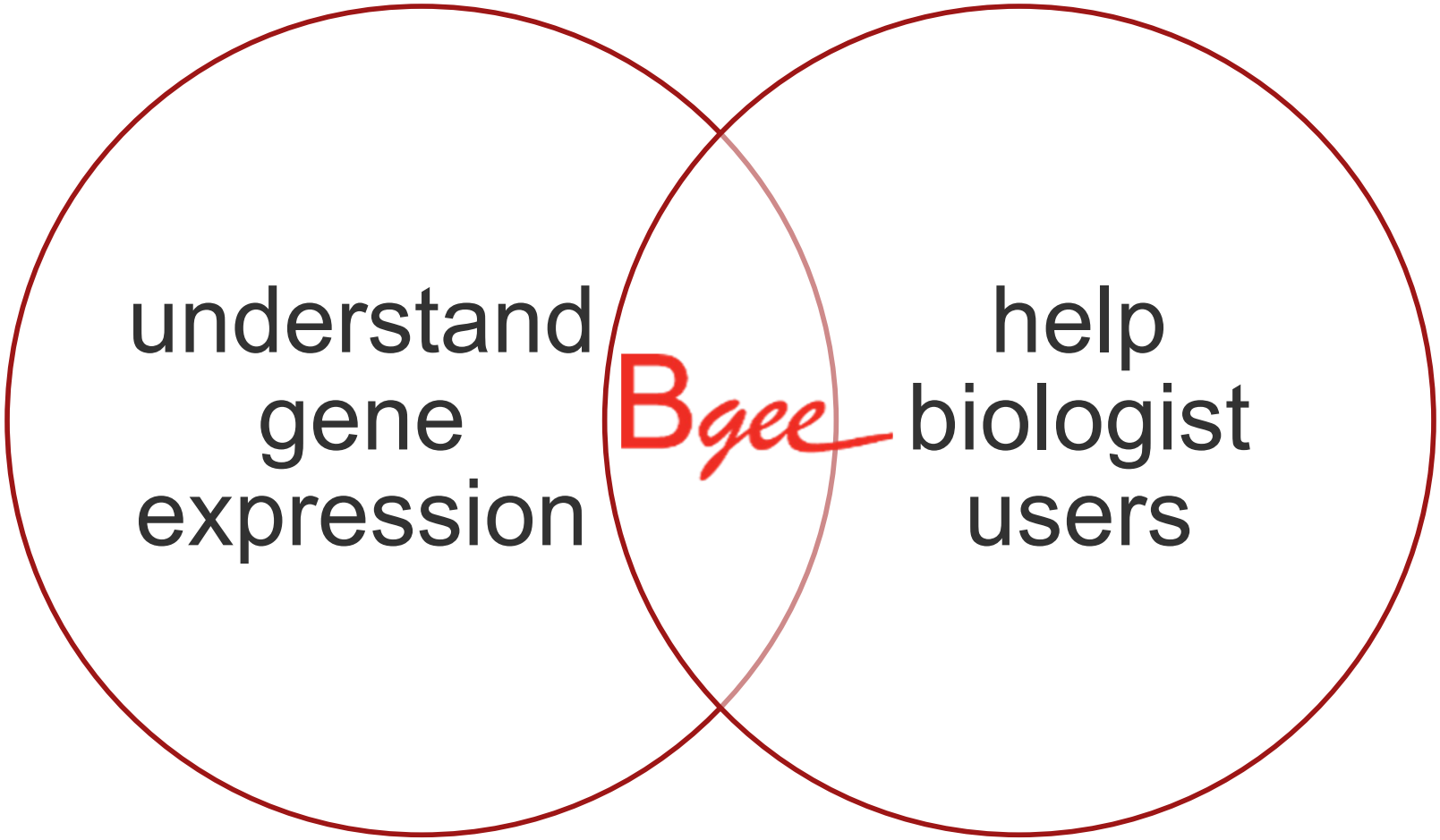
bgee@sib.swiss

@bgee.org

@bgeedb@genomic.social

Bgee

Bgee: Gene expression expertise



Demo

Do these species have data in Bgee?

- Platypus
- Human cancer
- Fly *D. melanogaster*
- Yeast *S. cerevisiae*
- Human healthy
- *Arabidopsis thaliana*

Data integration

ArrayExpress

Gene Expression Omnibus

SRA

ENA European Nucleotide Archive

DDBJ

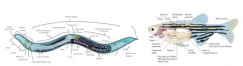
dbGaP

1. Data curation

- Identification of valid datasets, e.g.:
  - supported techonology
  - healthy wild-type samples
- Retrieval of all necessary files, e.g.:
  - barcode file
  - BAM files
  - association barcode -> cell population

2. Data annotation

- Information about experiment, e.g.:
  - protocol used
  - related paper
- Annotation to ontologies, e.g.:
  - anatomical entity + cell type
  - Developmental and life stage



  - sex

  - strain

3. Data processing

- Gene expression quantification
- Gene expression state determination

Bioconductor OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

BgeeCall

4. Data integration

- Integrate all data in Bgee

Biocuration

Uncurated databases

- Data redundancy
- Low organization of knowledge
- Main added value: **complete, up-to-date data**

Example: NCBI nucleotides (GenBank)

Curated databases

- Verified data
- Minimal redundancy
- Standardized annotations
- Main added value: **organized reliable knowledge**

Example: UniProtKB/Swiss-Prot

Biocuration involves the translation and integration of information relevant to biology into a database or resource that enables integration of the scientific literature as well as large data sets.

https://www.biocuration.org/dissemination/who-are-we/

Annotation

Associating a biological object to a feature, based on evidence.

Examples:

- associating a Gene Ontology term to a gene based on best Blast hit (uncurated annotation)
- associating a Gene Ontology term to a gene based on reading an experimental paper (curated annotation)

Bgee is a **curated** database

- All expression data in Bgee is verified
  - **data which doesn't fulfill criteria is excluded**
- Expression datasets are annotated by manual curation
  - **we read all the metadata, if needed the paper, if needed contact the authors**
- Annotations follow standards which are themselves curated

Some examples of curated databases?

Curation: wild-type healthy

Why wild-type healthy?

Informative on causal function of genes

Evolutionarily relevant (comparisons between species)

Reference for biomedical studies

Example: GTEx curation

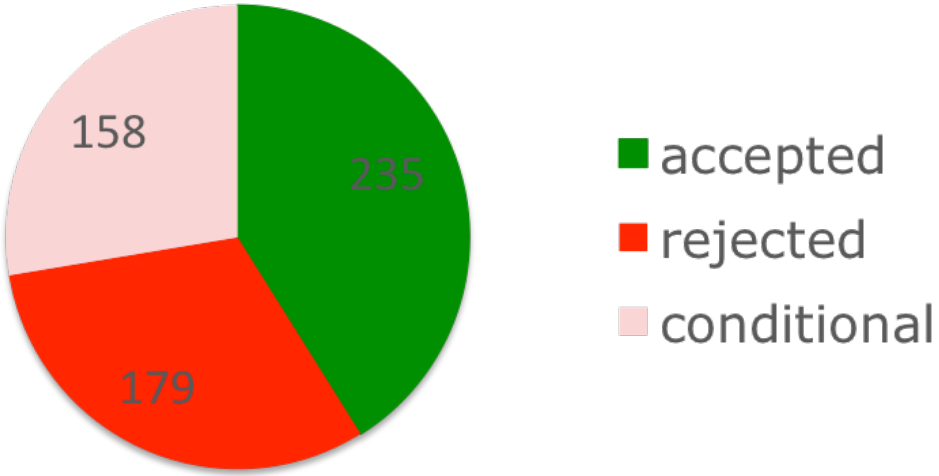
"The Genotype-Tissue Expression (GTEx) project is an ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation. Samples were collected from 54 non-diseased tissue sites across nearly 1000 individuals."

But "To date, we have not excluded specific donors from specific tissues based on their cause of death or medical history."

https://gtexportal.org/

https://www.gtexportal.org/home/faq#sampleExclusion

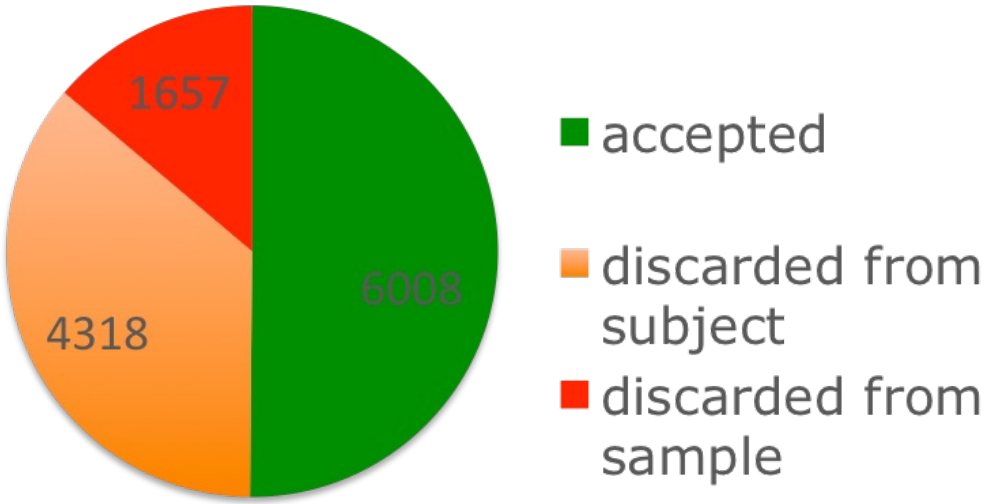
Curation GTEx data



e.g., drug abuse, cancer, BMI > 35

GTEx subjects

e.g., brain tissues from dementia, alzheimer; liver tissue from ascites, liver disease



GTEx samples

GTEx v6 in Bgee

11'983 libraries reviewed

- 4'860 libraries retained in Bgee

Private annotations:

- anatomical entity – exact age – sex – ethnicity

Public annotations in Bgee:

- anat. entity – broad age range – sex – no ethnicity
- 539 conditions, 75 anatomical entities

https://www.bgee.org/experiment/SRP012682

Fly Cell Atlas in Bgee

Annotation errors in FCA, corrected in Bgee:

- Inconsistent term names and IDs

“Male reproductive system” reported with 2 different IDs, FBbt:00004864 and FBbt:00004927

- Inconsistency between sex and cell type or tissue

“Ovary cell” in males; “Testis” in females

- Tissue annotation in cell type column

“adult hindgut” in cell type annotation column

Fly Cell Atlas in Bgee

61 libraries

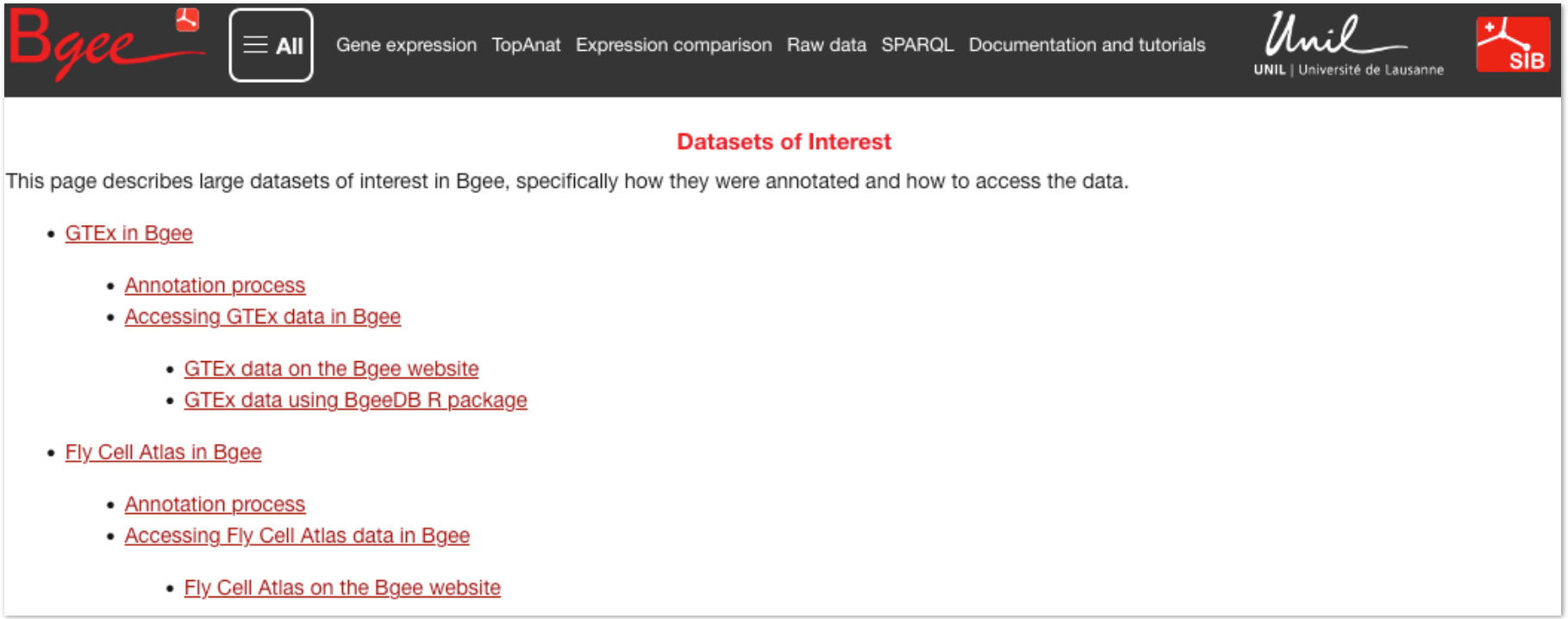
1518 conditions

- 17 anatomical entities
- 151 cell types
- 4 stages
- 3 sexes (mixed, male, female)
- 5 strains

27'031'026 processed expression values

https://www.bgee.org/experiment/ERP129698

Datasets of interest



To be continued in Bgee 16…

https://www.bgee.org/support/data-sets

Which terms in this experiment description fit "wild-type healthy"?

Circadian time course of liver mRNA profile of WT, Bmal1-liverKO, Rev-erbα/β-liverdoubleKO, Cry1Cry2 double KO after 12 weeks of high fat diet feeding ad libitum or time-restricted feeding

Curation: data standardization

Standardized data is reusable data

Standards define necessary metadata

when missing, biocurators contacts authors

➔ Complete files available from Bgee

Experiment ID: SRP222001

Technology: scRNA-Seq

Description: Here we perform massively parallel single-cell RNA sequencing (scRNA-seq) of human retinas using two independent platforms, and report the single-cell transcriptomic atlas of the human retina. Using a multi-resolution network-based analysis, we identify all major retinal cell types, and their corresponding gene expression signatures. Overall design: Three replicates of macula and peripheral retinas.

DOI: 10.1038/s41467-019-12780-8

Source: SRP222001 SRA

Download:

scRNA-Seq droplet-based processed expression values per cell population in Homo sapiens

scRNA-Seq droplet-based H5AD data per cell in Homo sapiens

Single-cell: H5AD standard

Bgee documentation on H5AD

Bgee's H5AD File Contents

In each H5AD file representing a single experiment, you will find:

1. **Main Matrix**: A central matrix that records the UMI (Unique Molecular Identifier) count data. Each row corresponds to a unique cell, while each column is aligned with a specific gene. The numeric value at a given matrix position reflects the UMI count, indicating the expression level of that gene in the corresponding cell.
2. **Metadata**: Accompanying the main matrix are annotations for each cell. These annotations offer insights into the cell's origin and characteristics.

Metadata Details

For every cell in the matrix, the following metadata is provided:

- **barcodes**: Unique sequences associated with individual cells or nuclei.
- **SampleId**: Sample identifier from which the cell was derived.
- **anatEntityId**: Unique identifier of the anatomical entity, from the Uberon ontology.
- **stageId**: Unique identifier of the developmental stage of the specimen, from the Uberon ontology.
- **cellTypeId**: Unique identifier of the cell type, from the Uberon ontology.
- **strain**: Genetic strain or variant information of the specimen.
- **sex**: Biological sex of the specimen ('not annotated', 'NA', 'mixed', 'male', 'female', 'hermaphrodite').
- **speciesId**: Code representing the species of the specimen.
- **anatEntityName**: Name of the anatomical structure/source.
- **stageName**: Name of the developmental stage of the specimen.
- **cellTypeName**: Name of the cell type.
- **libID**: Unique library identifier.
- **unique_cell_ids**: A distinct identifier ensuring every cell across the database is individually recognizable.

Usage

To access and manipulate the data in H5AD files, users can utilize the `scanpy` library in Python. This library offers a rich suite of methods for preprocessing, visualizing, and analyzing single-cell data.

Example:

```
import scanpy as sc

# Load H5AD file
adata = sc.read("path_to_your_file.h5ad")

# Access main matrix
matrix = adata.X

# Access metadata
metadata = adata.obs
```

https://www.bgee.org/support/tutorial-processed-expression-values-download-scRNA-seq-droplet-based

Annotation of protocols: single-cell RNA-seq

32 single-cell protocols classified in Bgee pipeline:

- plate-based, microfluidics, droplet-based, nanowells…
- same diversity of RNA targeted as for bulk
- full length, 3' or 5' RNA
- UMI and/or barcodes
- cells identified by FACS or a posteriori

4 protocols accepted at present in Bgee:

- Smart-seq and Smart-seq2, Full-length
- 10X Chromium V2 and V3, 3’ end
- single nuclei or single cell

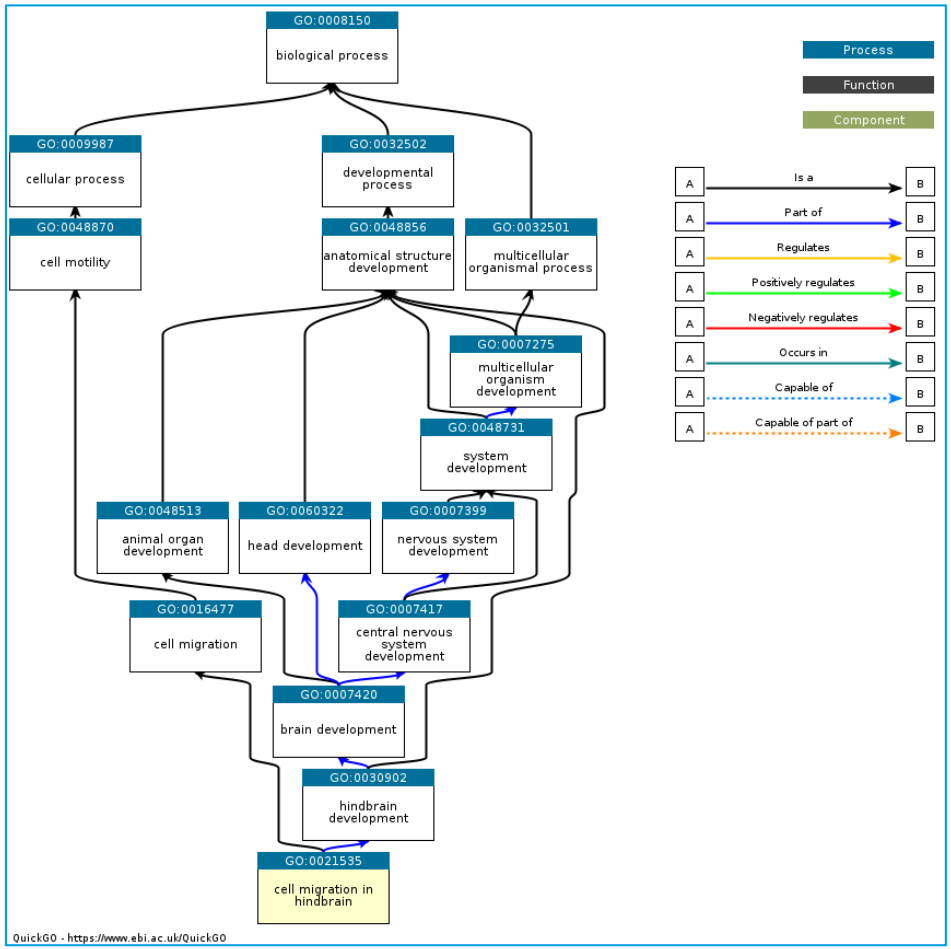
Annotation to ontologies

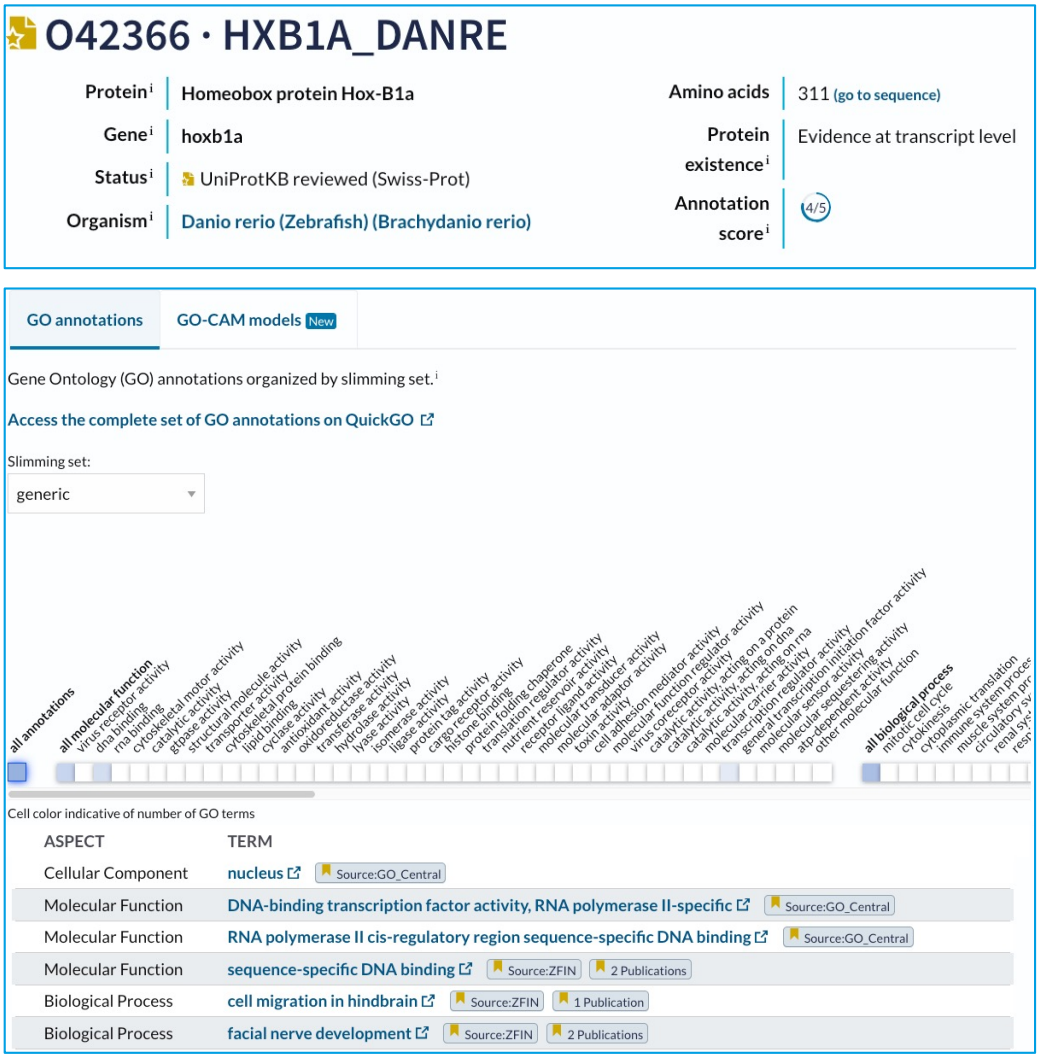
Ontologies in bioinformatics

An ontology is:

- a list of terms
  - **just a list of terms is a controlled vocabulary**
- definitions of the terms
  - **a list of terms with definitions is a dictionary**
- relations between the terms
  - **now we have an ontology!**
  - **various types of relations**
  - **allows automated reasoning**

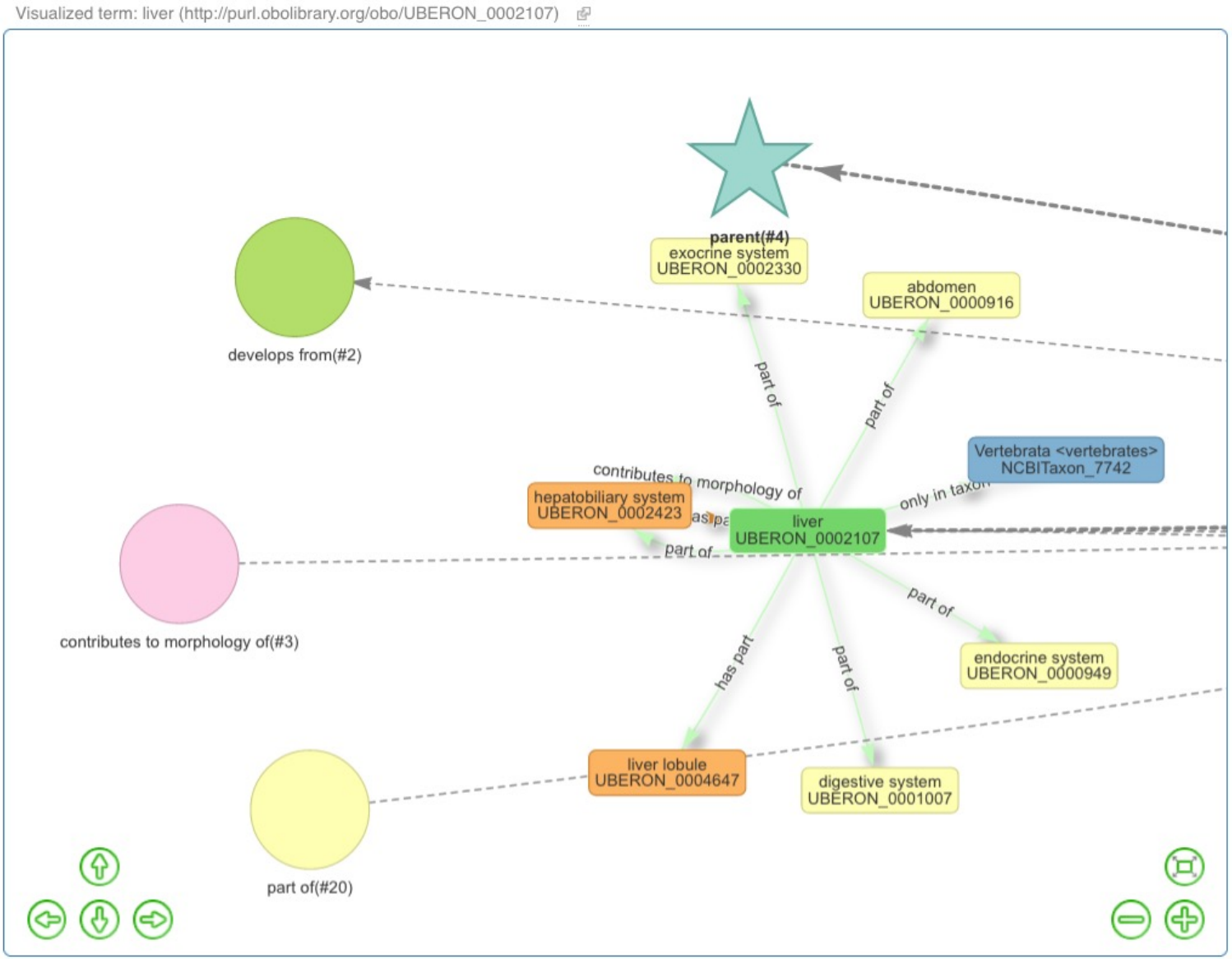
E.g., the Gene Ontology





https://www.ebi.ac.uk/QuickGO/term/GO:0021535
https://www.uniprot.org/uniprot/O42366

Uberon: the Uber-anatomy ontology



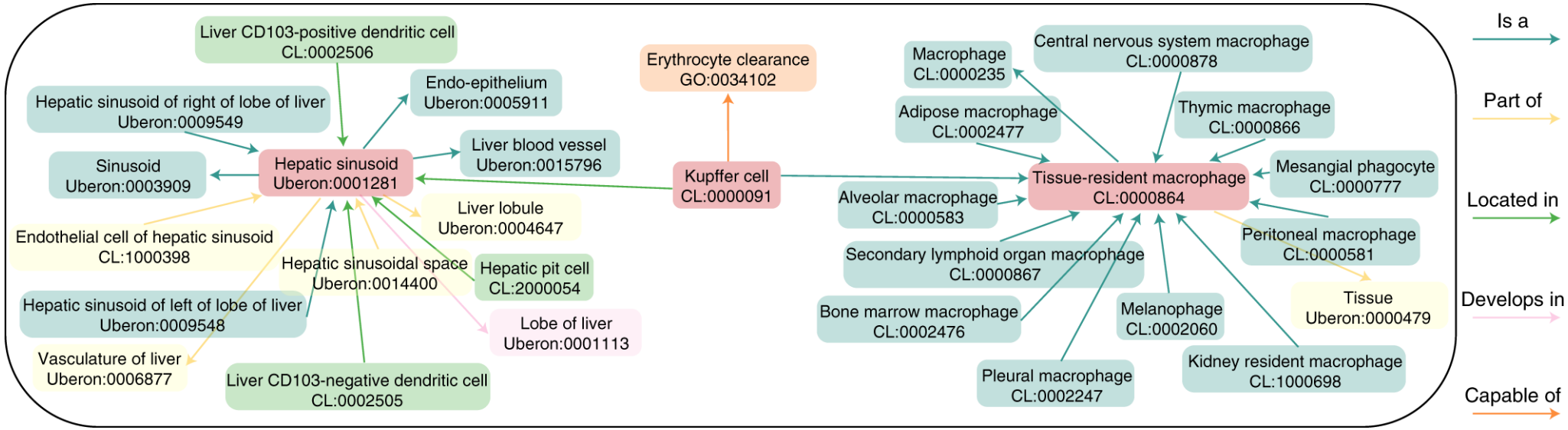
Legend

Relationship	Color	Visibility
Extended nodes (*)		-
is a		☑
part of		☑
produces		☐
site_of		☐
only in taxon		☑
has part		☑
develops from		☑
contributes to morphology of		☑
drains		☐
supplies		☐
attached to		☐

https://www.ebi.ac.uk/ols/ontologies/uberon/terms/graph?iri=http://purl.obolibrary.org/obo/UBERON_0002107

Mungall et al (2012) Genome Biology 13: R5

Cell Ontology



Osumi-Sutherland et al. (2021) Nature Cell Biol 23: 1129–1135

Challenges from single-cell RNA-seq:

- Unable to distinguish between cell types

T neuron T4a, T neuron T4b, T neuron T5a, T neuron T5b all annotated to FBbt:00058205 adult cholinergic neuron

- Constantly require new cell type terms

octopaminergic neuron / tyraminergic neuron

How many Uberon IDs for this experiment?

GSE30352 reporting RNA-seq from 6 organs across 10 species of mammals and birds

- 30352
- 6
- 10
- 60

What makes an ontology useful?

Used by many resources: one common standard

- Uberon common to all animal species
- Uberon used in large projects such as GTEx, Human Cell Atlas, Fly Cell Atlas

Covers a large domain of knowledge

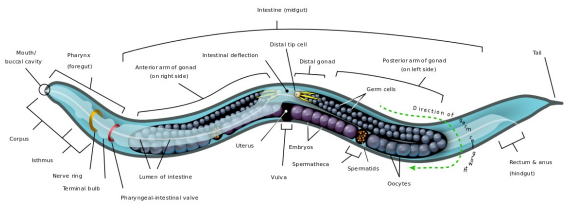
- Uberon covers animal anatomy

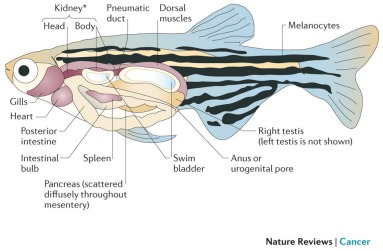
Many tools leveraging it, e.g. GO enrichments

- you'll see our tools today ☺

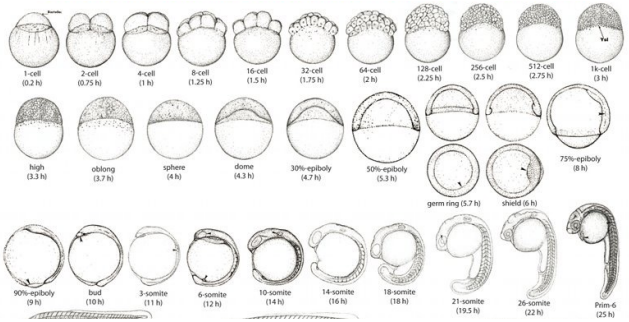
Bgee condition annotations

- anatomical entity + cell type

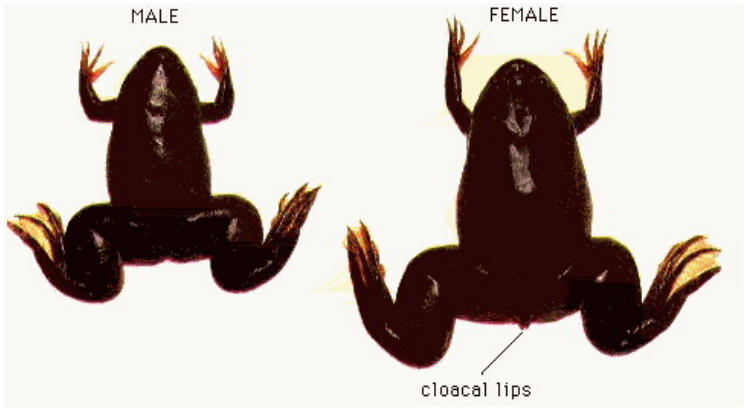




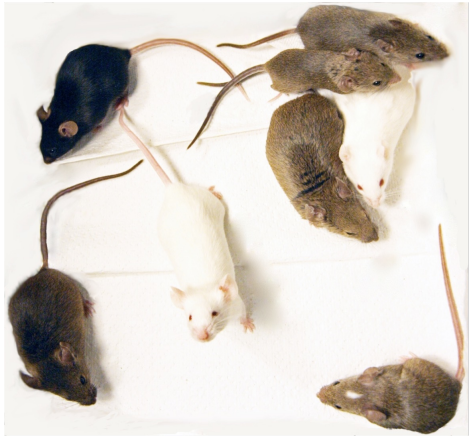
- Developmental and life stage



- sex



- strain



An example of annotation

Library ID	Cell type ID	Cell type name	Cell type author annotation	Anat. entity ID	Anat. entity name	Anat. entity author annotation	Stage ID	Stage name	Stage author annotation	Sex
SRX6859234	CL:0000748	retinal bipolar neuron	BPs	UBERON:0000053	macula lutea	macula in the central retina	HsapDv:0000162	68-year-old stage (human)	68 years	male
SRX6859234	CL:0000115	endothelial cell	Endo	UBERON:0000053	macula lutea	macula in the central retina	HsapDv:0000162	68-year-old stage (human)	68 years	male
SRX6859234	CL:0000745	retina horizontal cell	HCs	UBERON:0000053	macula lutea	macula in the central retina	HsapDv:0000162	68-year-old stage (human)	68 years	male
SRX6859234	CL:0000126	macroglial cell	Macroglia	UBERON:0000053	macula lutea	macula in the central retina	HsapDv:0000162	68-year-old stage (human)	68 years	male
SRX6859234	CL:0000129	microglial cell	Microglia	UBERON:0000053	macula lutea	macula in the central retina	HsapDv:0000162	68-year-old stage (human)	68 years	male
SRX6859234	CL:0000740	retinal ganglion cell	RGCs	UBERON:0000053	macula lutea	macula in the central retina	HsapDv:0000162	68-year-old stage (human)	68 years	male
SRX6859234	CL:0000604	retinal rod cell	Rods	UBERON:0000053	macula lutea	macula in the central retina	HsapDv:0000162	68-year-old stage (human)	68 years	male
SRX6859234	CL:0000561	amacrine cell	ACs	UBERON:0000053	macula lutea	macula in the central retina	HsapDv:0000162	68-year-old stage (human)	68 years	male
SRX6859235	CL:0000604	retinal rod cell	Rods	UBERON:0013682	peripheral region of retina	region of mid-peripheral retina	HsapDv:0000162	68-year-old stage (human)	68 years	male
SRX6859235	CL:0000740	retinal ganglion cell	RGCs	UBERON:0013682	peripheral region of retina	region of mid-peripheral retina	HsapDv:0000162	68-year-old stage (human)	68 years	male
SRX6859235	CL:0000129	microglial cell	Microglia	UBERON:0013682	peripheral region of retina	region of mid-peripheral retina	HsapDv:0000162	68-year-old stage (human)	68 years	male
SRX6859235	CL:0000573	retinal cone cell	Cones	UBERON:0013682	peripheral region of retina	region of mid-peripheral retina	HsapDv:0000162	68-year-old stage (human)	68 years	male
SRX6859235	CL:0000748	retinal bipolar neuron	BPs	UBERON:0013682	peripheral region of retina	region of mid-peripheral retina	HsapDv:0000162	68-year-old stage (human)	68 years	male
SRX6859235	CL:0000561	amacrine cell	ACs	UBERON:0013682	peripheral region of retina	region of mid-peripheral retina	HsapDv:0000162	68-year-old stage (human)	68 years	male

From Menon et al. (2019) *Single-cell transcriptomic atlas of the human retina identifies cell types associated with age-related macular degeneration*. Nat Commun 10: 4902

1. Data curation

SRA

ENA European Nucleotide Archive

ArrayExpress

DDBJ

GEO Gene Expression Omnibus

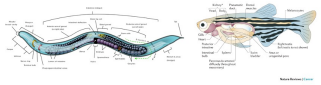
dbGaP

- Identification of valid datasets, e.g.:
  - supported techonology
  - healthy wild-type samples
- Retrieval of all necessary files, e.g.:
  - barcode file
  - BAM files
  - association barcode -> cell population

2. Data annotation

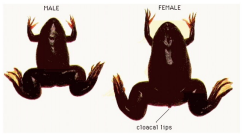
- Information about experiment, e.g.:
  - protocol used
  - related paper
- Annotation to ontologies, e.g.:

• anatomical entity + cell type

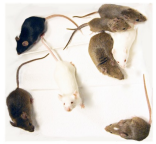


• Developmental and life stage

• sex



• strain



3. Data processing

- Gene expression quantification
- Gene expression state determination

Bioconductor OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

BgeeCall

4. Data integration

- Integrate all data in Bgee

End of part 1