

Données ouvertes :

L'expérience de la base de données de référence d'expression de gènes Bgee

Marc Robinson-Rechavi
Département d'écologie et évolution
FBM



@bgeedb@genomic.social



@bgee.org

Unil.



<https://www.bgee.org/>

L'expression des gènes

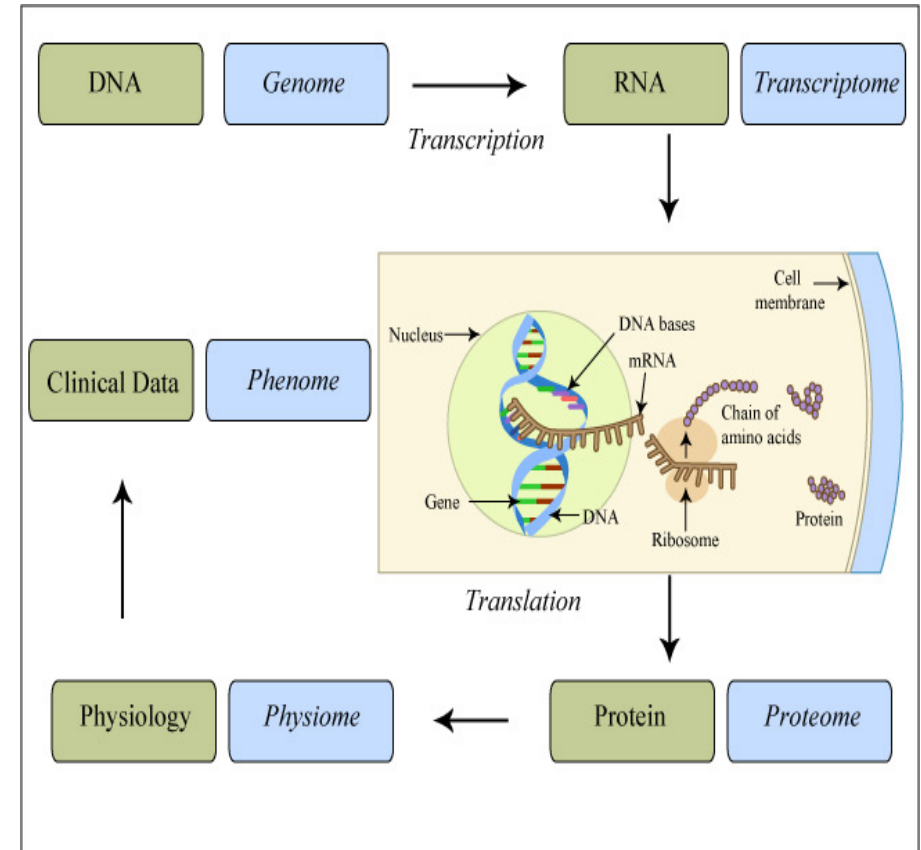
Quoi ?

Pourquoi ?

ARN ça vous dit quelque chose ?

ARN ça vous dit quelque chose ?

- Toutes les cellules ont le même ADN
- Elles diffèrent par l'expression des gènes
= leur transcription en ARN
- ADN -> hérédité
- ARN -> fonction



Pour comprendre le fonctionnement des gènes et des cellules, il faut l'expression des gènes

L'expression est différente

- pour chaque gène
- pour chaque cellule
- pour chaque organe
- selon l'âge, le sexe, ...
- entre espèces

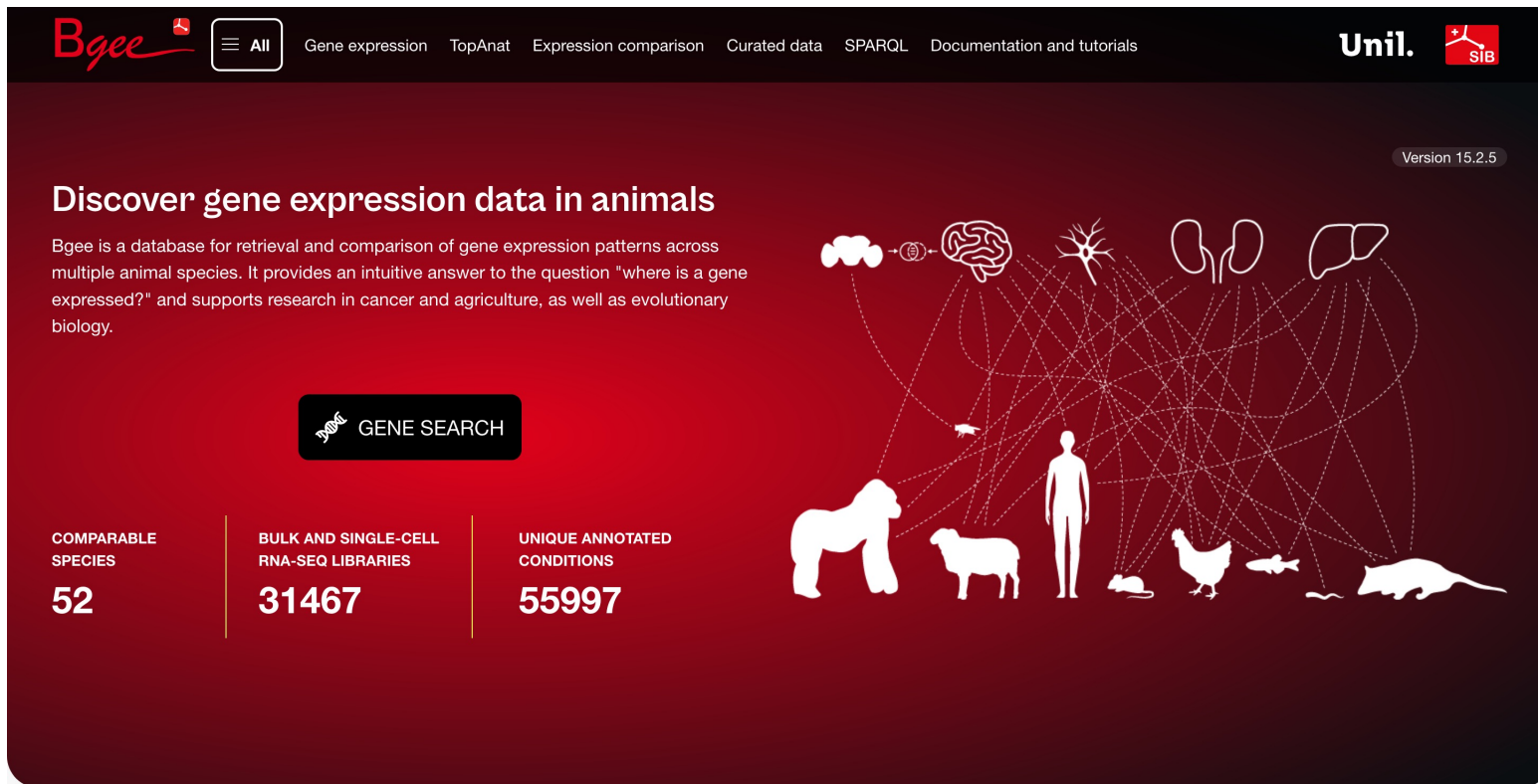
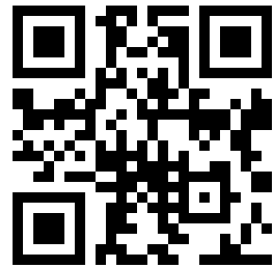


Le défi :

Organiser les données d'expression des gènes pour les rendre utiles

La base de données Bgee à l'UNIL

<https://www.bgee.org/>



The screenshot shows the Bgee database homepage. The header includes the Bgee logo, a menu icon, and navigation links: Gene expression, TopAnat, Expression comparison, Curated data, SPARQL, and Documentation and tutorials. The Unil. SIB logo is in the top right corner. The main content area has a dark red background. On the left, it says 'Discover gene expression data in animals' and provides a brief description of the database. In the center, there is a 'GENE SEARCH' button with a magnifying glass icon. Below this, three statistics are listed: 'COMPARABLE SPECIES' (52), 'BULK AND SINGLE-CELL RNA-SEQ LIBRARIES' (31467), and 'UNIQUE ANNOTATED CONDITIONS' (55997). On the right, there is a diagram showing various animal silhouettes (dog, sheep, human, mouse, chicken, fish, pig) connected by lines to icons representing different organs (brain, heart, liver, etc.), illustrating the database's focus on gene expression across species and tissues. A 'Version 15.2.5' badge is visible in the top right corner of the main content area.

Bgee   All [Gene expression](#) [TopAnat](#) [Expression comparison](#) [Curated data](#) [SPARQL](#) [Documentation and tutorials](#) **Unil.** 

Version 15.2.5

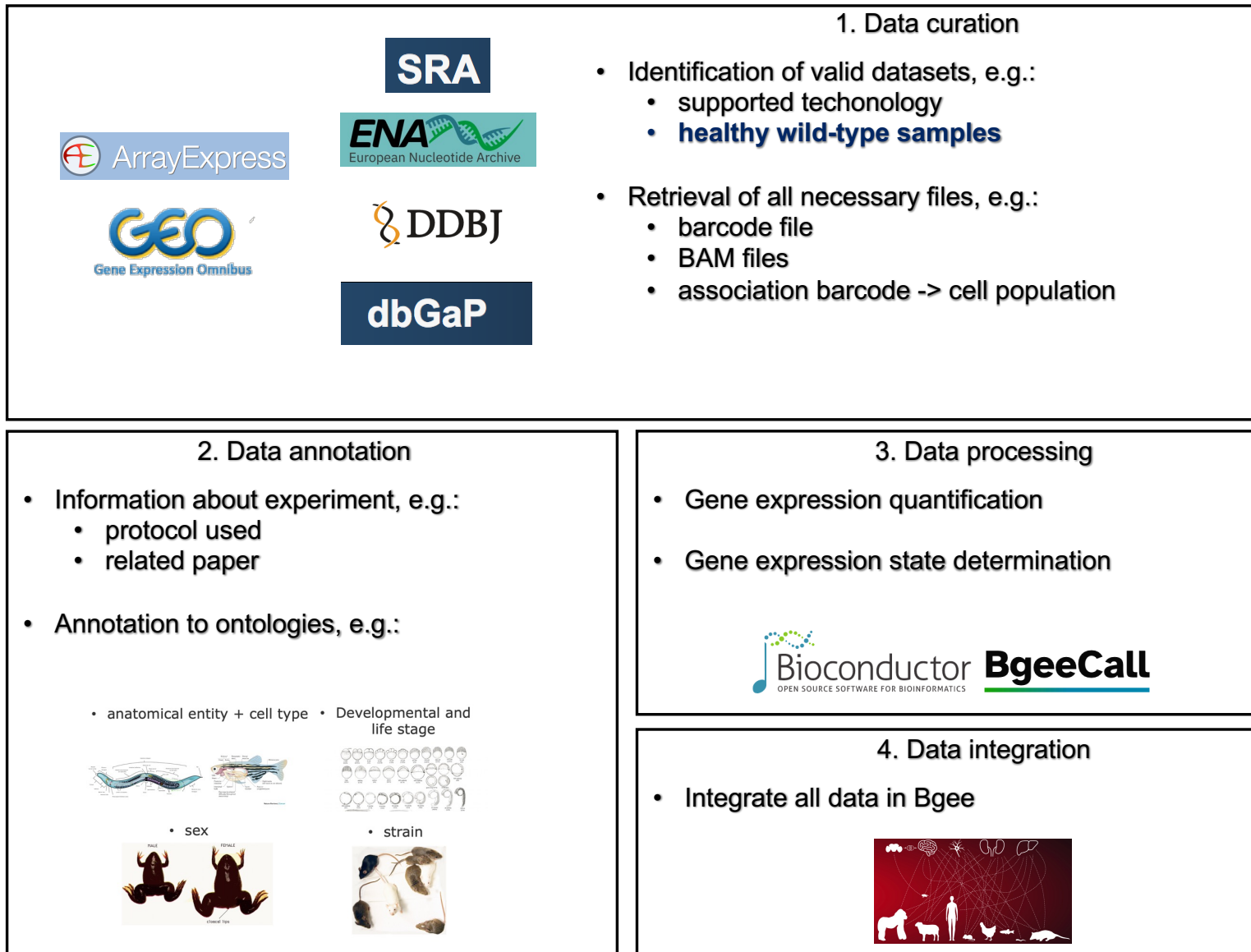
Discover gene expression data in animals

Bgee is a database for retrieval and comparison of gene expression patterns across multiple animal species. It provides an intuitive answer to the question "where is a gene expressed?" and supports research in cancer and agriculture, as well as evolutionary biology.

 **GENE SEARCH**

COMPARABLE SPECIES	BULK AND SINGLE-CELL RNA-SEQ LIBRARIES	UNIQUE ANNOTATED CONDITIONS
52	31467	55997





Curation de jeux de données

[expression-annotations-documents](#) / [scRNA](#) / [ERP132587](#) / [ERP132587.ipynb](#)

```
In [24]: cell_df_complete = pd.read_csv(info_celltype_complete_path, sep='\t', index_col=False, keep_default_na=False, na_display_df(cell_df_complete))
```

	cell_type	notes_cell	cellTypeId_	cellTypeName_	cellTypeAnnotationStatus_	cell_ty
0		should be NA/ none	CL:0000255	eukaryotic cell	other	
1	Undifferentiated_Spermatogonia		CL:0000020	spermatogonium	missing child term	Undifferentiated_Sperma
2	Differentiated_Spermatogonia		CL:0000020	spermatogonium	missing child term	Differentiated_Sperma
		early spermatid aka round spermatid				

```
In [25]: barcode_file = barcode_file.merge(cell_df_complete, left_on='cell_type', right_on='cell_type')
```

```
In [26]: barcode_file['cellTypeId'] = barcode_file['cellTypeId_'].values
barcode_file['cellTypeName'] = barcode_file['cellTypeName_'].values
barcode_file['cellTypeAnnotationStatus'] = barcode_file['cellTypeAnnotationStatus_'].values
```

```
In [27]: # optional - update cell type name and add comments
barcode_file['cell_type'] = barcode_file['cell_type_new'].values
#barcode_file['comment'] = barcode_file[['comment', 'comment_cell']].agg(' '.join, axis=1)
```

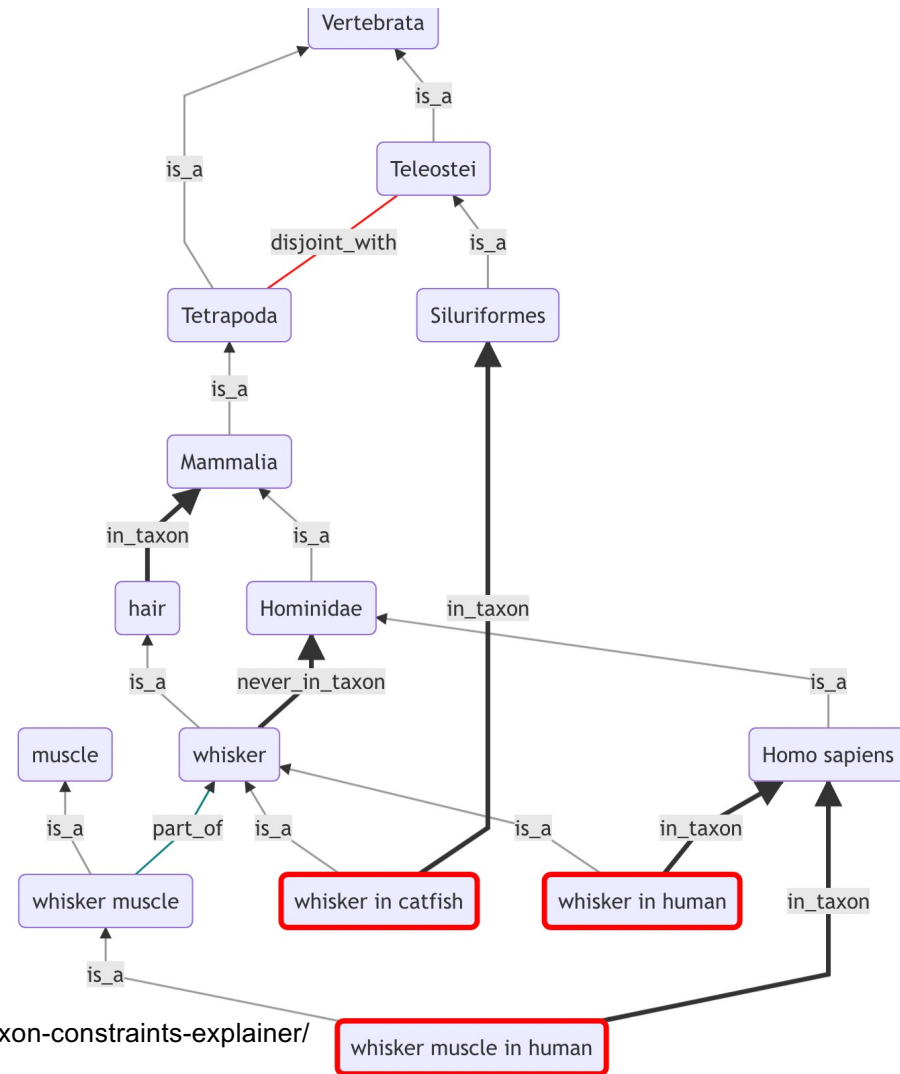
```
In [28]: barcode_file_complete = barcode_file[barcode_cols]
display_df(barcode_file_complete)
barcode_file_complete.to_csv(barcode_to_add_path, sep="\t", index=False, quoting=csv.QUOTE_ALL)
```

1221	TGCCAAATCAGCGGT	ERX6700458	ERP132587	testis	Differentiated_Spermatogonia	nan
1222	CCTCTGACAGTAACT	ERX6700458	ERP132587	testis	Differentiated_Spermatogonia	nan
1223	CGTTCTGTCCGATATG	ERX6700457	ERP132587	testis	Leptotene_Spermatocytes	nan
1224	TTGACTTGCTCTTTA	ERX6700458	ERP132587	testis	Leptotene_Spermatocytes	nan
1225	CAGGTGCGTATAGGTA	ERX6700458	ERP132587	testis	Leptotene_Spermatocytes	nan



2 biocuratrices, 140% temps

Annotation à des ontologies de référence



<https://oboacademy.github.io/oobook/explanation/taxon-constraints-explainer/>

L'annotation permet l'intégration et la comparaison

1) Identifying scRNA-Seq experiments for inclusion

SRA experiment ID: SRP343046
SRA library ID: SRX13121072

SRA experiment ID: SRP167086
SRA library ID: SRX4947727

SRA experiment ID: SRP125768
SRA library ID: SRX3426759

2) Retrieving sample and cell population information as provided by authors

Species: "Homo sapiens"
Tissue: "CA1"
Life stage: "51y"
Cell population: "Oligo OPALIN LAMA2"

Species: "Mus musculus"
Tissue: "Olfactory bulb"
Life stage: "14-20 weeks"
Cell population: "Astro3"

Species: "Drosophila melanogaster"
Tissue: "brain"
Life stage: "less than 8 Hours"
Cell population: "Ensheathing glia"

3) Mapping free text information to ontologies

Tissue:
UBERON:0003881
"CA1 field of hippocampus"
Life stage:
HsapDv:0000145
"51-year-old stage"
Cell population:
CL:0000128
"oligodendrocyte"

Tissue:
UBERON:0002264
"olfactory bulb"
Life stage:
MmusDv:0000137
"3-6 month-old mature stage"
Cell population:
CL:0000127
"astrocyte"

Tissue:
UBERON:0000955
"brain"
Life stage:
FBdv:00007075
"day 0 of adulthood"
Cell population:
FBbt:00100504
"ensheathing neuropil associated glial cell"

 multi-species terms
 species-specific terms

4) Ontology inference allows to match information across experiments and species

Tissue:
UBERON:0000955
"brain"
Life stage:
UBERON:0000066
"fully formed stage"
Cell population:
CL:0000125
"glial cell"

Données ouvertes



Sign in

Sign up



BgeeDB

11 followers

<https://www.bgee.org/>

[@bgee.org](#)

[@bgeedb@genomic.social](#)

[company/bgee](#)

[@bgeedatabase](#)

[Overview](#) [Repositories 19](#) [Projects](#) [Packages](#) [People 4](#)

Pinned

[BgeeDB_R](#) Public

Source code of the R package BgeeDB to use data from the Bgee database

R 16 7

[BgeeCall](#) Public

Source code of the BgeeCall R package, for automatic RNA-Seq present/absent gene expression calls generation

R 5 2

[bgee_pipeline](#) Public

Source code of the Bgee pipeline used to build the Bgee database

Perl 12 3

[confidence-information-ontology](#) Public

Repository of the Confidence Information Ontology (CIO)

Perl 5 2

[anatomical-similarity-annotations](#) Public

Project hosting resources used for annotating relations of similarity between anatomical structures

Makefile 2 2

[homology-ontology](#) Public

Home for the ontology of homology and related concepts in biology

Makefile 1

People



Top languages

R HTML Perl Makefile
 VBA

Most used topics

[biology](#) [gene-expression](#) [rna-seq](#)
[obofoundry](#) [ontology](#)

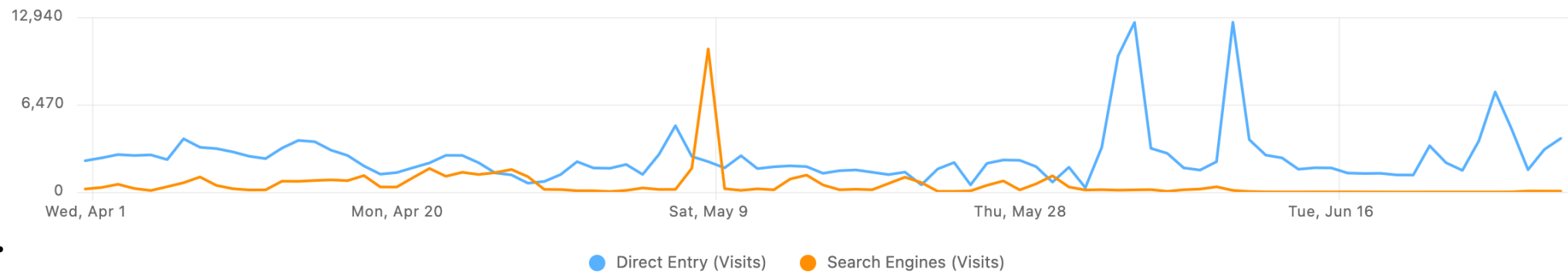
Des licences claires, un accès facile

Toutes les données sont CC0

Accès :

- Téléchargement web
- Package R (licence GPL)
- SPARQL

Unil.



Pourquoi CC0?

- Des conditions très claires
- Pas de coûts de suivi
- Confiance dans les normes académiques

Alternatives :

- CC-BY
- CC-BY-NC

Données organisées
Données ouvertes
Données utiles



Unil.

Unil.

<https://www.bgee.org/>



@bgeedb@genomic.social

@bgee.org



Swiss National
Science Foundation

swissuniversities¹⁷



Swiss Institute of
Bioinformatics

08.09.26