The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals

Frederic B. Bastian[1,2,*], Julien Roux[1,2,3], Anne Niknejad[1,2], Aurélie Comte[1,2], Sara S. Fonseca Costa[1,2], Tarcisio Mendes de Farias[1,2], Sébastien Moretti[1,2], Gilles Parmentier[1,2], Valentine Rech de Laval[1,2], Marta Rosikiewicz[1,2,4], Julien Wollbrett[1,2], Amina Echchiki[1,2], Angélique Escoriza[1,2,5], Walid Gharib[1,2,4], Mar Gonzales-Porta[1,2,6], Yohan Jarosz[1,2,7], Balazs Laurenczy[1,2,4], Philippe Moret[1,2], Emilie Person[1,2,8], Patrick Roelli[1,2,9], Komal Sanjeev[1,2], Mathieu Seppey[1,2,10], Marc Robinson-Rechavi[1,2,*]

[1] Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland

[2] SIB Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

[3] Current affiliation: Bioinformatics Core Facility, Department of Biomedicine, University of Basel, Switzerland

[4] Current affiliation: SOPHiA GENETICS, Switzerland

[5] Current affiliation: Medion Grifols AG, Switzerland

[6] Current affiliation: Genome Institute of Singapore, Singapore

[7] Current affiliation: Luxembourg Centre for System Biomedecine, Luxembourg

[8] Current affiliation: Office fédéral de l'environnement OFEV, Switzerland

[9] Current affiliation: 10x Genomics, Sweden

[10] Current affiliation: University of Geneva, Switzerland

[*] Authors for Correspondence: Frederic B. Bastian and Marc Robinson-Rechavi, Department of Ecology and Evolution, University of Lausanne, Switzerland, bgee@sib.swiss

Abstract

Bgee is a database to retrieve and compare gene expression patterns in multiple animal species, produced by integrating multiple data types (RNA-Seq, Affymetrix, in situ hybridization, and EST data). It is based exclusively on curated healthy wild-type expression data (e.g., no gene knock-out, no treatment, no disease), to provide a comparable reference of normal gene expression. Curation includes very large datasets such as GTEx (re-annotation of samples as "healthy" or not) as well as many small ones. Data are integrated and made comparable between species thanks to consistent data annotation and processing, and to calls of presence/absence of expression, along with expression scores. As a result, Bgee is capable of detecting the conditions of expression of any single gene, accommodating any data type and species. Bgee provides several tools for analyses, allowing, e.g., automated comparisons of gene expression patterns within and between species, retrieval of the prefered conditions of expression of any gene, or enrichment analyses of conditions with expression of sets of genes. Bgee release 14.1 includes 29 animal species, and is available at https://bgee.org/ and through its Bioconductor R package BgeeDB.

Introduction

Gene expression is central to the relation between genes or genomes and their function. It mediates or indicates the involvement of genes in organ functions, pathways, pathologies, reactions to the environment, and differences between individuals, species, or genes. Many of the features which make expression so interesting also complicate its bioinformatics analysis, especially for multicellular organisms such as animals. While databases of reference genomes, providing standardized annotations and comparative frameworks within and between species, are well established[123], this is not the case for expression data. There is not any concept of a "reference expression" for a species, and comparisons are complicated by the need to define comparable samples and conditions.

The reusability and ease of access and interpretation of genome data have been a major part of the success of genomics. Despite rare somatic mutations, a genome sequence can be characterized in a stable manner for one individual. And although individuals differ at some small proportion of sites, most of the genome is stable for a given species, allowing the use of a reference genome. On the other hand, the transcriptome varies between cell types and conditions. Thus many samples representing this diversity need to be integrated to provide an accurate picture. Moreover, like for genomes, different technologies (e.g. in situ hybridization vs. RNA-Seq) provide different resolutions and levels of information. So while there is a concept of a reference transcriptome in the sense of a collection of transcript sequences, there is not a clear concept of a reference set of expression patterns and expression levels. Yet such a reference is needed for many applications, such as comparing species, characterizing gene function, or comparing expression in disease to the healthy expectation.

This integration and availability of genome data has allowed the widespread development of comparative genomics, shedding light on the evolution of species and on our own human history. A variety of tools has been developed for the automatic comparison of genomics data, since the creation of the widely-used BLAST software in 1990[4]. And indeed, the best predictor of phenotypic importance of a genomic feature is evolutionary conservation[5,6]. An early example of evolutionary conservation of expression patterns was the study of contribution of hox genes to vertebrate axial morphology[7]. Such discoveries were instrumental in the emergence of the field of evo-devo[8], the study of the evolution of development, allowing to understand the mechanisms responsible for the evolution of species morphology. But we are limited in our capacity to perform systematic comparative transcriptomics analyses, because of factors such as heterogeneous data, non trivial anatomical homology, or transient gene expression[9].

There are a few datasets which can play a role close to that of a reference expression pattern. Such datasets cover typically many anatomical structures (tissues or organs), and are often used to present expression in gene-centric databases. We call such experiments covering many anatomical structures “atlases”. For example, as of May 2020, NCBI Genes[10] presents by default expression from the HPA set[11] for human genes, with the option of switching to another of three atlases. Each of these is a RNA-seq atlas covering a different set of 6 to 27 tissues. Because they are presented separately, it can be difficult to come to a conclusion concerning the expression pattern of a gene. For example, human insulin is expected from its biology to be most expressed in pancreas, and more specifically in the Langerhans islets B cells. In NCBI Genes, the default HPA experiment shows indeed very pancreas-specific expression of human insulin (NCBI Gene ID: 3630), but without finer detail on the specificity at the cell type level. But the other proposed atlases, lacking pancreas samples, show top expression in the prostate, the spleen, or the stomach. Mouse orthologs Ins1 and Ins2 (Gene ID: 16333, 16334) are reported expressed in the duodenum from the mouse ENCODE RNA-seq atlas[12], again without any pancreas samples in this experiment. There are other such “reference” atlases. One of the oldest, still used, e.g., in GeneCards[13] or Wikipedia gene pages, is the GNF microarray atlas of human and mouse tissues[14]. One of the most recent, and probably the most complete so far, is the GTEx human transcriptome atlas[15]. While each of these atlases are valuable, they are dispersed in different databases, do not have common standards of annotation or quality-control, are processed in different ways, and provide differently processed values. Moreover, the concept of a reference transcriptome is obviously dependent on health, genetic background, and conditions of sampling. For example, GTEx does not guarantee that samples come from healthy individuals[16]. In another large dataset, ENCODE, many samples are from immortalized cell lines which may not reflect in vivo conditions. And as the insulin example shows, each of these individual datasets only covers a limited subset of anatomical structures, of developmental stages, or of other conditions relevant to characterizing expression patterns.

In response to these challenges, there exist a few databases which specialize in providing expression pattern descriptions[17]. The Expression Atlas[18] presents curated and processed expression data from a diversity of species and sources. Curation notably includes annotation to standard ontologies for anatomy, experimental conditions, etc. As of May 2020, it includes 3,942 studies from 65 species, mostly mammals (including human), and plants. From the perspective discussed here, an important feature is that 214 of these studies are annotated as “baseline”, meaning that they *“report transcript or protein abundance typically in constitutive conditions, such as healthy tissues, cell types, developmental stages or cell lines”*, and are used to present an expression pattern of each gene. The other studies are used only for differential expression, e.g., between treated and untreated samples, or between healthy and pathological. In the Expression Atlas, expression is presented per experiment. There is no integration over all experiments to give one reference description of a gene expression pattern. Users can sort anatomical entities by the highest expression of the gene in any experiment. For example, for both human and mouse insulin, the top organ is the pancreas.

Another database of expression, Tissues[19] presents expression patterns derived from UniProtKB annotations, analysis of selected reference atlases, and text mining of PubMed abstracts, for human, rat, mouse and pig. Notably, Tissues favors a different approach, focusing on confidence in expression presence rather than levels of expression. The reference atlases cover diverse techniques, from microarrays and RNA-seq to immunochemistry and proteomics. While text mining results are combined to provide an integrated score per tissue per gene, other experiments are presented separately, as in Expression Atlas. An interesting feature is that tissues are ranked within each evidence type, allowing the most important features of a complex expression pattern to emerge. For human insulin, the pancreas is one of two tissues from UniProtKB and Pancreatic beta cell is the top tissue from text mining, as it is for mouse insulin 1 and 2 (as of May 2020).

Finally, Model Organism Databases (MODs)[2021] provide integration of many datasets for a given species (or a targeted set of species). They were the first to standardize the capture of experimental conditions and to use ontologies for anatomy and development, allowing integration of many small scale experiments such as in situ hybridization. A first, obvious, limitation of this approach is that it does not cover species outside of a small core of model organisms. This approach does not scale to covering more biodiversity, nor does it aim to, unlike a genome browser such as Ensembl[3]. A second limitation is that much of the expression data in each MOD comes from mutant strains (e.g., Knock-out or Knock-down), and it can be difficult to identify only “normal” expression data, produced from healthy wild-type individuals.

We have developed the Bgee database to answer questions about gene expression while avoiding limitations due to the separation of data sets and species. We notably provide an integrated reference of gene expression patterns in animals. With Bgee we aim to provide a framework enabling the comparison of expression patterns between genes and between animal species, with a focus on healthy wild-type data. This is why we have developed methods to integrate different types of expression data to make them comparable, and tools to automatically compare expression patterns between genes, within or between species. Bgee integrates curated expression data from different sources in diverse animal species. The data produced by Bgee allow: i) to answer the questions “where and when this gene is expressed?”, and “what are the genes expressed in this condition?”; ii) to identify the conditions most relevant to the expression of each gene; iii) to study gene expression evolution, by comparing expression patterns between species. Bgee provides a consistent vision over all species, and integrates data over a large number of atlases and datasets.

We describe in this paper the Bgee dataset of curated and processed data available to users, and the gene expression calls integrated both among data types and among species. We present the tools developed on top of these data to discover biological insights into normal gene functions and expression evolution. We also detail the complete procedure to produce these data, and the resources we provide to reproduce these analyses.

Results

We describe here: i) the data available in Bgee that we have produced; ii) the tools to leverage these data for biological insights; iii) the tools to reproduce Bgee analyses; iv) other resources that we have developed for Bgee.

Overview of Bgee

Bgee currently integrates RNA-Seq, Affymetrix, *in situ* hybridization, and EST data. One of the advantages of such an integrative approach is to obtain data for as many conditions as possible, with as much anatomical and developmental detail as possible. Moreover, new technologies do not require a new database, but allow integration with historical data.

For developing Bgee, we have curated all expression data to the same standard, of high quality, non-redundant, healthy wild-type data. We have annotated or mapped all data to the same anatomical and developmental ontologies, with also information of sex and strain. We have consistently reprocessed all RNA-Seq data, and Affymetrix data where raw data were available. This allows us to propose to users a dataset of consistently curated, annotated, and processed expression data, in various animal species, available for downstream analyses. This is a first level of data integration.

From all these data we have then produced qualitative calls of presence/absence of expression, and quantitative expression ranks, integrated from all data types, and comparable between conditions and species. Bgee provides a single answer to the question “where and when is this gene expressed?”. For instance, for the insulin gene, both human and mouse insulin have in Bgee top expression in islets of Langerhans and more generally pancreas; in human the information is even more fine-grained, with top expression in the Beta cells. Users of this second level of integration (calls) are free to ignore the complexity of the underlying data, to concentrate on the biological signal of interest, while we also provide high quality processed data for downstream studies.

Data provided by Bgee

In this section we describe the Bgee release 14.1 datasets of: i) annotated and processed expression values; ii) calls of presence/absence of expression and expression ranks; iii) anatomical similarity annotations; iv) developmental stage similarity annotations. All original data sources used to build Bgee release 14.1 are in Table 1.

We have published all these data under the Creative Commons Zero license (CC0), with the aim of facilitating downstream integration into other resources, and avoiding the problem of license stacking. Although CC0 doesn’t legally require users of the data to cite the source, we would still appreciate being cited if Bgee data or tools were used.

Data sources	**URL**	**reference**	**Description**
Genomics databases			
Ensembl	http://www.ensembl.org/index.html	3	Source for gene annotations, mappings to the Gene Ontology to Affymetrix probeset IDs, and cross-references to other databases
miRBase	http://www.mirbase.org/	22	Source for miRNA families
OMA	https://omabrowser.org/oma/home/	23	Source of gene orthology information
EnsemblMetazoa	http://metazoa.ensembl.org/index.html	2	Source for gene annotations, mappings to the Gene Ontology to Affymetrix probeset IDs, and cross-references to other databases
RNA-Seq			
GEO	https://www.ncbi.nlm.nih.gov/geo/	24	RNA-Seq data source for various species
GTEx - dbGAP	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v8.p2	25	GTEx RNA-Seq data
SRA	https://www.ncbi.nlm.nih.gov/sra	26	RNA-Seq data source for various species
Affymetrix			
ArrayExpress	https://www.ebi.ac.uk/arrayexpress/	27	Affymetrix data source for various species
GEO	https://www.ncbi.nlm.nih.gov/geo/	24	Affymetrix data source for various species
In Situ			
BDGP	https://insitu.fruitfly.org/cgi-bin/ex/insitu.pl	28	Drosophila in situ data source
FlyBase	http://flybase.org/	29	Drosophila in situ data source
MGI	http://www.informatics.jax.org/expression.shtml	30	Mouse in situ data source

WormBase	https://wormbase.org/#%23012-34-5	31	Nematode Information Resource
Xenbase	http://www.xenbase.org/entry/	32	Xenopus in situ data source
ZFIN	http://zfin.org/	33	Zebrafish in situ data source
EST			
smiRNAdb	http://www.clipz.unibas.ch/cloningprofiles/	34	EST data for miRNAs
UniGene	https://www.ncbi.nlm.nih.gov/UniGene/	35	EST data source for various species
Ontologies			
CIO	http://obofoundry.org/ontology/cio.html	36	Confidence Information Ontology
Developmental stage ontologies	https://github.com/obophenotype/developmental-stage-ontologies/		Collection of developmental and life stage ontologies in animals. Integrated into Uberon
GO	http://geneontology.org/GO.downloads.ontology.shtml	37	Filtered Gene Ontology
Uberon	https://uberon.github.io/	38	Integrated cross-species ontology covering anatomical structures in animals. Use of the subset "composite-metazoan"
Other sources			
Anatomical similarity annotations	https://github.com/BgeeDB/anatomical-similarity-annotations/		Define evolutionary relations between anatomical entities described in the Uberon ontology
NCBI Taxonomy	https://www.ncbi.nlm.nih.gov/taxonomy	39	Source taxonomy used in Bgee

Table 1: data sources of Bgee, with URLs, references, and descriptions.

Annotated and processed expression values

To date, we have annotated conditions in which expression data were produced with 5 types of information: anatomical localization, developmental and life stage, sex, strain/ethnicity, and species. Bgee 14 includes 29 animal species. We have annotated anatomical localization using the multi-species anatomical ontology Uberon[38]. We have annotated developmental stages by using a multi-species ontology developed in collaboration with Uberon development[40], integrating existing ontologies for some species, and ontologies that we have created for other species. We have annotated sex and strain with a basic controlled vocabulary. We have annotated species using the NCBI taxonomy[39]. We have curated all data to retain only healthy, wild-type data (no treatment, no gene knockout, etc). Since we have annotated and remapped all these data in a consistent framework using ontologies, and reprocessed all of them, it allows a complete integration of all data. Bgee thus allows users to retrieve a dataset of expression data consistently curated, annotated, and processed, usable in their own downstream analyses. Statistics about the data integrated per species are presented in Table 2.

Species	**EST**	**In Situ**		**Affymetrix**		**RNA-Seq**	
	libraries	evidence	experiments	chips	experiments	libraries	experiments
Anolis carolinensis	-	-	-	-	-	31	5
Bos taurus	-	-	-	-	-	121	8
Caenorhabditi selegans	-	360	141	177	34	41	6
Canis lupus familiaris	-	-	-	-	-	141	21
Cavia porcellus	-	-	-	-	-	28	4
Danio rerio	108	42153	4674	186	33	147	16
Drosophila ananassae	-	-	-	-	-	4	1
Drosophila melanogaster	62	92227	5468	965	92	253	11
Drosophila mojavensis	-	-	-	-	-	8	1
Drosophila pseudoobscura	-	-	-	-	-	10	1
Drosophila simulans	-	-	-	-	-	15	2
Drosophila virilis	-	-	-	-	-	4	1
Drosophila yakuba	-	-	-	-	-	4	1
Equus caballus	-	-	-	-	-	232	22
Erinaceus europaeus	-	-	-	-	-	6	1
Felis catus	-	-	-	-	-	32	5
Gallus gallus	-	-	-	-	-	48	6
Gorilla gorilla	-	-	-	-	-	13	2
Homo sapiens	2393	-	-	5452	323	5676	36
Macaca mulatta	-	-	-	14	1	238	18
Monodelphis domestica	-	-	-	-	-	108	15
Mus musculus	706	223513	37654	6095	698	330	30
Ornithorhynchus anatinus	-	-	-	-	-	21	4
Oryctolagus cuniculus	-	-	-	-	-	55	13

Pan paniscus	-	-	-	-	-	13	2
Pan troglodytes	-	-	-	-	-	250	18
Rattus norvegicus	-	-	-	107	2	106	8
Sus scrofa	-	-	-	-	-	169	14
Xenopus tropicalis	66	2400	1304	-	-	259	5

Table 1: data statistics for release Bgee 14.1 per species for the 29 species included in this version. This table provides number of EST libraries integrated, number of *in situ* hybridization experiments and evidence (e.g., the image of a staining) counts, number of Affymetrix chips and experiments, number of RNA-Seq libraries and experiments.

RNA-Seq data

For RNA-Seq data, we have remapped all raw data to transcriptome to produce counts at the transcript level, then aggregated them per gene. These aggregated counts are used to compute FPKMs and TPMs per gene. We have notably re-curated the GTEx human dataset phs000424.v6.p1[15]. Only 50% of samples were kept, to discard unhealthy or contaminated samples, representing a high quality subset of GTEx. For each RNA-Seq library, we provide detailed information, including annotations, library information, and relevant statistics (see online documentation). For each gene in each library, we provide several measures of expression level, and calls of presence/absence of expression.

Affymetrix data

For Affymetrix data, we have processed raw CEL files when available, or used MAS5 processed files when raw CEL files were not available. Similar to RNA-Seq libraries, for each Affymetrix chip, we provide annotations, chip information, and relevant statistics. In Affymetrix, expression is measured per probeset rather than per gene; several probesets can map to one gene, and provide different measures. For each probeset in each chip, we provide the gene the probeset maps to, the signal intensity, and the call of presence/absence of expression (see online documentation, and Supplementary materials).

In situ hybridization data

We have retrieved *in situ* hybridization data from relevant Model Organism Databases (MODs; Table 1). For each evidence, we do not store the original image, or paper figure, but we provide a link to the original data. For each evidence and each spot, meaning the report of an area with staining from expression of a gene, or lack of staining from absence of expression of a gene, we provide annotations, mapping to gene, call of presence/absence of expression and quality of the call for this spot. For details of what Bgee re-uses or generates by new annotation, see *Materials and Methods, Integration of in situ hybridization data from Model Organism Databases*. At present, these data are not available for direct download, but are used in integrated calls and scores.

Expressed Sequence Tags

Both EST databases which we used as data sources (Table 1) are now retired, thus we are no longer updating these data. For each EST library, we provide annotations. We provide the mapping between genes and ESTs per library. For each gene in each library, we provide the number of ESTs mapped to it, and the call of presence of expression. Note that calls of absence of expression are not produced from EST data. At present, these data are not available for direct download, but are used in integrated calls and scores.

Calls of presence/absence of expression and expression ranks

While many applications focus on expression levels in different conditions, it is also important to determine which genes are actively transcribed, and which are not. This is the information provided by calls of presence/absence of expression. Such calls are very similar to the data that can be reported using *in situ* hybridization methods, and can be compared between conditions and between species. By making such calls from each data type, we can combine their strengths: *in situ* hybridization provides very fine anatomical resolution; Affymetrix provide transcriptome-wide data on a large variety of samples, including a wealth of heritage data (e.g., aging); and RNA-seq has the highest sensitivity and specificity, and provides high quality transcriptomics for non model species.

Expression calls

A call corresponds to a unique combination of a gene in a condition, with reported presence or absence of expression. A condition is defined according to anatomical entity, life stage, sex, strain or ethnicity, and species. For each call, a confidence score is computed. At time of writing, users can retrieve calls considering the anatomical entity, life stage, and species, but we have annotated the expression data to capture sex and strain information as well. This other information will be publicly available for calls in a future release of Bgee.

For each evidence integrated into Bgee (a RNA-Seq library, a probeset on an Affymetrix chip, an *in situ* hybridization spot, or an EST library), we have produced a qualitative call of presence/absence of expression in the condition studied, for each gene analyzed. For *in situ* hybridization, it is the only form that the data takes. For continuous data types, we apply a threshold which depends on the data type: for Affymetrix a Wilcoxon test on the signal of the probesets against a subset of weakly expressed probesets[41]; for RNA-seq, a library-specific threshold depending on the distribution of TPMs on intergenic sequences (Julien Roux, Marta Rosikiewicz, Julien Wollbrett, Marc Robinson-Rechavi, Frederic B. Bastian; in preparation); and for ESTs a threshold based on the number of tags[42]. These calls can then be integrated over experiments and over data types.

We have propagated these individual calls produced from multiple techniques and experiments along a graph of conditions (see Materials and Methods, Call Propagation). Then, we have integrated these individual calls from multiple experiments, propagated along the condition graph, to produce one global call of presence/absence of expression per gene - condition, associated to a confidence level[36]. As a result, for each gene in each condition, Bgee provides one global call, the confidence level of the call (gold, silver, bronze), the number of experiments from each data type supporting the presence and absence of expression, and whether the call has been propagated or observed directly in the condition.

Expression ranks and scores

While the expression calls are useful to make data comparable between experiments, data types, and species, they lack quantitative information. To overcome this limitation, we have developed a method to rank genes in a condition based on their expression level. We have also integrated these ranks over all data types and experiments. The lower the rank score of a gene in a condition, the higher its expression. Because this is not very intuitive to users, we also compute an "expression score" for visualisation, which is simply 100*(Max(ranks) +1 - rank)/Max(ranks), Max computed over all genes in a species. Thus the top expressed gene has a rank score of 1 and an expression score of 100, and the lowest expressed has the maximum rank and an expression score of 100/Max(ranks). E.g., for human maximum rank is 55,062 and the corresponding expression score is 0.0018. This is for instance the case for gene MT4 (ENSG00000102891) in "lung" (UBERON:0002048) at stage "young adult" (HsapDv:0000089). This gene in this condition is not actively expressed (absent expression call).

Expression ranks can be retrieved for different combinations of the information used in annotations. To date, these are anatomical entity and developmental stage. To produce a rank considering only the anatomical entity, we have retained the minimum rank over all developmental stages with data for this anatomical entity.

For instance, the two first ranked anatomical entities of the human insulin gene *ins* are "type B pancreatic cell" (rank score 2.84, expression score 99.99660) and "islet of Langerhans" (rank score 5.60, expression score 99.99151), with information produced from Affymetrix and RNA-seq data. The first anatomical entity of the mouse insulin gene *ins1* is "islet of Langerhans" (rank score 4.44, expression score 99.99161), with information produced from RNA-Seq, Affymetrix, and *in situ* hybridization data; the first two entities of the mouse insulin gene *ins2* are "dorsal pancreas" (rank score 3.18, expression score 99.99469) and "islet of Langerhans" (rank score 3.51, expression score 99.99388), with information produced from Affymetrix and *in situ* hybridization data.

Anatomical and developmental similarity

To allow the study of the evolution of gene expression patterns, it is necessary to determine what are the conditions that can be compared between species. In an evolutionary context, the most relevant criterion for making comparisons is historical homology, a similarity that results from common evolutionary origin. For anatomical entities, that means entities in different species that derive from a common ancestral structure in a common ancestor. For developmental stages, heterochrony prevents stage-wide homology calls (organs do not develop at the same speed and in the same sequence in different species[43]), but comparable stages can still be defined.

We use the ontology of homology and related concepts[44] to capture the type of similarity considered between anatomical entities.These similarity annotations can be retrieved from GitHub[45], and will be described in detail elsewhere (Anne Niknejad, Marc Robinson-Rechavi, Frederic B. Bastian; unpublished). It should be emphasized that they are derived from primary literature (paleontology, Evo-Devo, etc), not from the expression data in Bgee; thus the anatomical homology annotations and the gene expression calls are independent. As of Bgee 14.1, we have integrated 2,328 relations of homology, involving 1,845 anatomical entities. Which means that Bgee can allow the comparison of expression patterns between species in these 1,845 entities.

For development stages, we have merged all the developmental stage ontologies of species integrated in Bgee into one common multi-species ontology, within Uberon. Common general stages have been merged between species, and more precise species-specific stages are children of these general terms. For instance, the human-specific precise stage HsapDv:0000016 “Carnegie stage 09” is a child of the more general, multi-species term UBERON:0000111 “organogenesis stage”. While we could not map data to the exact equivalent of Carnegie stage 09 in all species, we could compare them at the more general organogenesis stage, thanks to propagation of calls to parent terms. This makes comparison with a developmental aspect possible.

The merged multi-species ontology used in Bgee is available at https://github.com/obophenotype/developmental-stage-ontologies/blob/master/external/bgee/dev_stage_ontology.obo (the archived version used in Bgee 14 releases is available at https://github.com/obophenotype/developmental-stage-ontologies/blob/97fa5b7e176f8c7f72cd5c7115d2fc36f35e8bb3/external/bgee/dev_stage_ontology.obo).

Web interface of Bgee

In this section, we present the tools which allow to interact with Bgee data through the web interface: i) gene expression patterns, allowing to sort the conditions where a gene is expressed to study its normal healthy function; ii) TopAnat, to perform expression enrichment analyses; iii) expression comparison, allowing to compare expression patterns between genes, even in different species; and iv) homology retrieval, allowing to retrieve homologous anatomical entities between different species. An overview of the interfaces can be found in Figure 1.

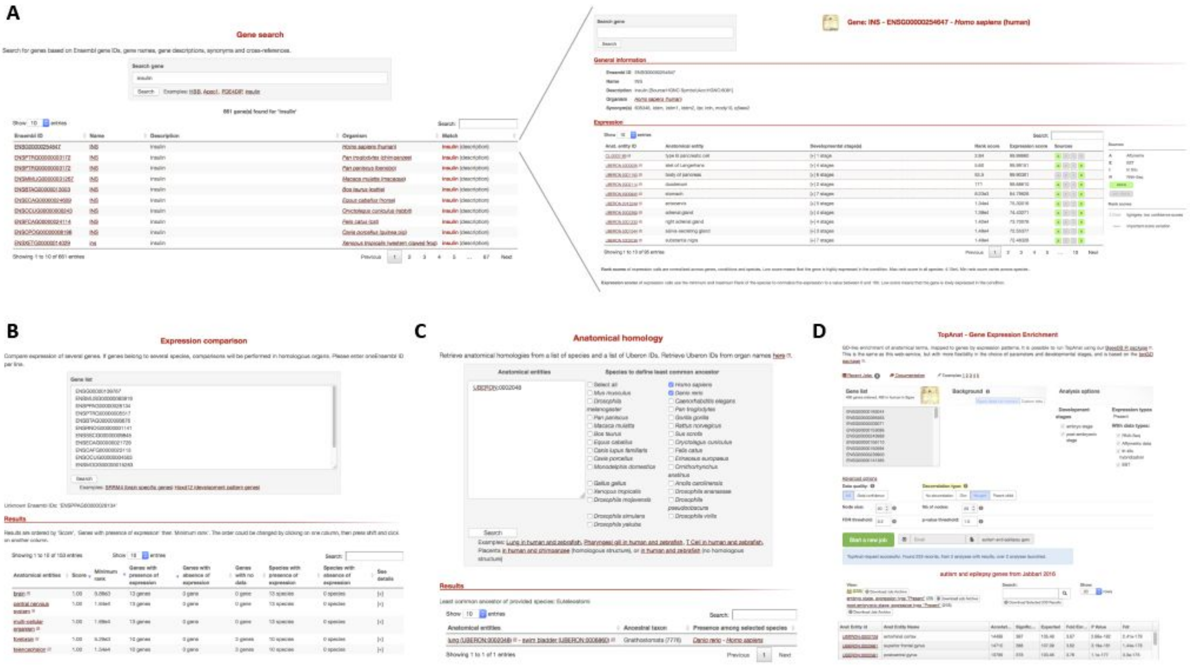

A
Gene search
Gene: INS - ENSG00000254647 - Homo sapiens (human)
B
Expression comparison
C
Anatomical homology
D
TopAnat - Gene Expression Enrichment

Figure 1: screenshots of the Bgee web interfaces. A: example of gene search (top left) for the term “insulin” (https://bgee.org/?page=gene&query=insulin), allowing to go to the gene page (top right) displaying ranked conditions with expression for the human gene INS (https://bgee.org/?page=gene&gene_id=ENSG00000254647). B: example of comparison of expression patterns for the SRRM4 genes (brain-related genes) in 13 species (https://bgee.org/?page=expression_comparison&data=34beddfc93bb7fbb440e757e6de24d91fc0ce177). C: Anatomical homology retrieval tool, with here an example query allowing to identify swim bladder as the anatomical structure in zebrafish homlogous to the human lung (https://bgee.org/?page=anat_similarities&species_list=9606&species_list=7955&ae_list=UBERON%3A0002048). D: example of TopAnat analysis on a set of human genes associated to autism and epilepsy, identifying the enriched conditions with expression of these genes as specific brain regions (https://bgee.org/?page=top_anat#/result/8fce889da7b4519c5792573ed3933032c8122819/).

Gene expression patterns

The Bgee website allows to retrieve, for each gene, the conditions where it is expressed, sorted by their expression rank mentioned above. It provides information about the normal healthy function of a gene. As of Bgee 14.1, it is possible to sort anatomical entities where a gene is expressed, by its expression level, integrated from all data types and experiments; and for each anatomical entity, to rank the developmental stages where the gene is expressed. The Gene Page can be accessed at https://bgee.org/?page=gene.

TopAnat

A powerful use of the calls of presence of expression is to associate gene lists to specific anatomical structures. For example spermatogenesis genes are more often expressed in testis than expected by chance. We provide the anatomical expression enrichment test “TopAnat” to do this. TopAnat uses a similar approach to Gene Ontology enrichment tests[4647], but genes are associated to the anatomical structures from Uberon by their expression calls, instead of to their functional classification (Roux J., Seppey M., Sanjeev K., Rech de Laval V., Moret P., Artimo P., Duvaud S., Ioannidis V., Stockinger H., Robinson-Rechavi M., Bastian F.B.; in preparation). The algorithms from the package topGO[4849] are available in TopAnat to account for the non-independence of anatomical structures, and avoid the over-representation of lowly-informative top-level terms. TopAnat is available as a web-tool at https://bgee.org/?page=top_anat, and in the BgeeDB R package (see below).

As an example, we have used TopAnat to analyze a list of genes associated with autism and epilepsy in human[50]. TopAnat returns a list of anatomical structures where expression of these genes is enriched relative to the background of all human genes with expression. These structures are almost all specific brain regions, including all brain regions known to be affected by autism[51]: frontotemporal lobe (examples of TopAnat results part of this structure: UBERON:0002771 “middle temporal gyrus” and UBERON:0002810 “right frontal lobe”); frontoparietal cortex (UBERON:0001872 “parietal lobe” and UBERON:0001870 “frontal cortex”); amygdala (UBERON:0001876 “amygdala”); hippocampus (UBERON:0003881 “CA1 field of hippocampus”); basal ganglia (UBERON:0002038 “substantia nigra”, part of UBERON:0002420 “basal ganglion”); and anterior cingulate cortex (UBERON:0009835 “anterior cingulate cortex”).

Expression comparison

The Bgee website makes it possible to compare expression patterns between genes, within and between species. When genes in a single species are compared, Expression Comparison considers present/absent expression calls in anatomical entities (e.g., “lung” in human). When genes from several species are compared, it considers calls in homologous anatomical entities (e.g., “lung - swim bladder” for comparing data in human and zebrafish[525354]).

For comparing expression patterns, a user enters a list of gene IDs. Expression Comparison then retrieves all anatomical entities (for single-species comparison), or homologous anatomical entities (for multi-species comparison), where at least one of the genes is expressed. For each of these entities, it then displays the genes from the user list that are either expressed, or not expressed, or have no data. It proposes different sorting methods of these entities. For instance, when studying the conservation of expression of the brain specific genes SRRM4 in mammals, the top homologous anatomical entities, when sorted by maximum expression conservation and minimum expression rank, are “forebrain”, “telencephalon”, and “cerebral hemisphere”[55]. The Expression Comparison can be accessed at https://bgee.org/?page=expression_comparison.

Anatomical homology retrieval

Relations of historical homology are useful not only to gene expression, but for comparisons of any anatomy-related data, such as phenotypes. Bgee thus provides a way to easily query them through its web interface.

For instance, for comparing such anatomy-related data, a user may ask: “what is, in human, the comparable organ to the zebrafish ‘pharyngeal gill’?”. The homology retrieval tool will return the anatomical entity in human “parathyroid gland”[56]. Indeed, the parathyroid gland and the pharyngeal gill likely derive from a common ancestral structure, present in the ancestor of *Vertebrata*, as they both regulate extracellular calcium levels, and the parathyroid gland is positioned within the pharynx in *Tetrapoda*[57]. We have captured this information in our annotations of similarities between anatomical entities, that the homology retrieval webtool uses to answer the user query. The homology retrieval tool can be accessed at https://bgee.org/?page=anat_similarities.

Resources to access data

We provide several ways to access the data in Bgee, through: i) a Bioconductor R package; ii) a SPARQL endpoint; and iii) direct download of TSV files.

BgeeDB R package

We provide access to the annotated and processed expression values (see “Data” section) through the Bioconductor[58] R[59] package BgeeDB[60]. BgeeDB is a collection of functions to import into R these data, facilitating downstream analyses, such as gene expression analyses with other Bioconductor packages. BgeeDB also allows to run TopAnat analyses (see section “Tools leveraging data for biological insights”), the R package offering more flexibility in the choice of input data and analysis parameters than the web-interface, and possibilities of inclusion within programs or pipelines. The BgeeDB Bioconductor page can be found at https://bioconductor.org/packages/BgeeDB/.

SPARQL endpoint

Bgee also provides a SPARQL[61] endpoint which is based on the EasyBgee database. EasyBgee is a lighter version of the Bgee database, that contains the most useful information, made explicit. For instance, it contains only calls of presence of expression, and does not contain calls of absence of expression; the conditions used reference only the anatomical entity, developmental stage, and species information, and do not reference sex and strain information.

The endpoint is accessible at the address https://bgee.org/sparql/. In addition to SPARQL endpoint querying through any programmatic language, a web-interface also offers the possibility to run more user-friendly queries, developed as part of the BioSODA project[62]. This web-interface is available at http://biosoda.expasy.org/. Bgee specific queries are present under the category “Bgee database queries”.

Download files

We provide TSV files to retrieve, for each species: i) annotated and processed expression values for RNA-Seq and Affymetrix data; and ii) calls of presence/absence of expression with confidence and rank scores, generated from all data types.

The files containing expression values for RNA-Seq and Affymetrix data can be browsed at https://bgee.org/?page=download&action=proc_values. For each species and data type, one file lists all the experiments included in Bgee; another file lists all the evidence annotated (Affymetrix chips, or RNA-Seq libraries), with the information described in the “Data” section above. Finally, one file per experiment contains all the processed expression values from all evidence in this experiment.

The files containing expression calls can be browsed at https://bgee.org/?page=download&action=expr_calls. For each species, the user can specify whether the calls should be retrieved using only the anatomical entity, or both the anatomical entity and the developmental stage. More advanced information can also be selected, such as information about the calls for each data type independently.

Discussion

We provide a reference of gene expression patterns and data in diverse animal species, from humans and established model organisms, to the “new model army”[63] emerging with genomics. This reference can be used to study the function of genes in healthy wild-type conditions, to provide a control for mutant or disease studies, and to study expression conservation and evolution. We have built a robust framework which we can expand to new animal species and to new transcriptomics technologies.

The philosophy of Bgee is to provide information and data which users can trust, so that they can be built on to do further work. Thus Bgee shares not only expression data, but also expertise on the curation and analysis of these data. Because gene expression is complex, we need to be careful at all steps of this sharing: how the data is curated, how it is analysed, and how it is presented to users. Expert manual curation is at the core of providing trustworthy information[64–67]. It is key to certify that expression data is from healthy wild-type samples, and to annotate precisely and accurately to anatomy, stage, sex and strain. Manual curation is also the only way to generate reliable annotations of anatomical homology from a diverse literature of Evo-Devo, paleontology, zoology, etc. While expression data analysis is not specific to Bgee[17], our approach is to focus on delivering predefined clear biological signals, and choosing the best methodology to do so. Whether it is in calling genes present or absent from each source of data, ranking anatomical structure expression per gene integrated over experiments and data types, or computing enriched anatomical structures for gene lists, we strive to provide solutions which serve life scientists. To present a balance between trust and ability to verify, it it is easy to use our calls, relying on Bgee’s expertise, yet we also provide the code and data access to reproduce our analyses or modify them. It is thus critical that all our pipeline is public and documented (https://github.com/BgeeDB/bgee_pipeline). Finally, for users to have an easy and intuitive access to information from expression data, despite its complexity, we have chosen to present different views which answer different questions. These views must be consistent with the analysis and curation choices. While each view is necessarily a simplification, we combine very synthetic views, such as the Gene Page or Expression Comparison, with more advanced access, such as the R packages or SPARQL endpoint, to serve different use cases and ensure flexibility.

A key aspect to allow a meaningful integration of a variety of datasets is to standardize the capture of metadata information available for each sample, notably by using ontologies and controlled vocabularies. This information is often provided as free text information in primary databases, and the nomenclature and lexicon used can vary greatly between different depositors, and in the different species studied. Standardization is enforced only inside specific projects, using project-specific guidelines (see, e.g., 189 variables associated with the GTEx dataset[68]). This hampers the community's capability of integrating and comparing the data already available. Thus by annotating and curating data, we make it more findable, interoperable, and re-usable, in accordance with FAIR principles[69].

Allowing to link these gene expression data to other types of data in a necessity of utility in biology. This is why Bgee is part of the Bio-SODA project[62]. This project allows semantic queries across federated bioinformatics databases, by allowing the conversion of any repository into a semantic representation, that can be queried through a SPARQL endpoint. It allows Bgee data to be retrieved along with gene orthology information from OMA[70], and protein and functional annotations from UniProtKB/Swiss-Prot[71]. And this SPARQL endpoint allows to federate these data with any other bioinformatics resource providing a semantic representation as well. This integration of gene expression data along with other information is possible thanks to our approach of providing one clear answer to the question "where and when a gene is expressed". Otherwise, the amount of data to retrieve would be overwhelming (e.g., TPMs from tens of RNA-seq libraries), and would be difficult to link to other resources without well defined gene-level information.

Providing this clear gene-level expression information allows comparison between species. Comparative transcriptomics is essential to understand the molecular basis for phenotypes, such as, for instance, evolution of animal morphology[72], embryonic cell differentiation[73], species lifespan[74,75], or cancer evolution[76]. Most studies are based on the generation of a new dataset, tailored to answer a specific question. The integration of all available datasets in a meaningful way should allow new discoveries[77]. Bgee, by generating comparable reference sets of expression patterns in multiple species, and homology relations to link them, is the first resource to allow the systematic and automated comparison of gene expression patterns between species.

These comparisons are based on the present/absent calls of expression produced by Bgee. It is thus essential that these calls capture the relevant functional aspects of gene expression. This relevance is demonstrated, for instance, by the results provided by our webtools "gene expression page", and by TopAnat. On the gene expression page, the top ranked conditions of genes are relevant to their known biology (e.g., as of Bgee 14.1: several muscle regions for human *PDE4DIP* gene, "liver" for mouse *Apoc1* gene, "pancreas" for Xenopus *ins* gene). In TopAnat, results for list of genes are highly representative of their known function (e.g., as of Bgee 14.1: top ranked condition is "spermatocyte" for a list of mouse genes associated with spermatogenesis, "musculature of body" for a list of cow genes with a relation to muscle in their description). Bgee thus provides reference sets of expression patterns that are accurate and predictive of gene functions.

A challenge when providing a reference set of expression patterns is how to capture their variability, both between conditions and individuals. Capturing the variability between conditions is highly dependent on the depth of annotation when capturing condition information. A transcriptomics assay could be annotated by capturing only the anatomical localization aspect, or also capturing developmental stage aspect, sex, strain, genetic background, environmental conditions, etc. The less precise the definition of a condition is, the more variability of gene expression in this condition we will artificially observe. For instance, we observe sex difference in gene expression in mammals[78], as in, e.g., adult brain in human[79]. If we were to capture expression patterns of genes in brain of mixed sexes, without capturing the sex information, we would believe that many genes have a high variability of expression. But if we captured the sex information when analyzing these samples, we would observe less variability. If we could capture in the annotations all conditions impacting gene expression, what would be left to observe would only be gene expression noise and stochasticity[80], but we are far from that perspective.

Providing information about normal gene functions, that is possible to prioritize by expression conservation between species, has many applications in biology. One example is the use of the Bgee data in the OncoMX consortium[56]. This consortium has the aim of producing a resource integrating all available information about cancer biomarkers, in order to enhance new discoveries. In this context, Bgee provides its reference sets of healthy gene expression, in human and mouse, to be able to relate the changes occuring in cancer to the healthy normal expectation. And it is because Bgee can give a clear answer about what is this healthy normal expectation that it can be useful to the OncoMX project.

Here we have presented the latest release of Bgee, which integrates expression information from four well established data types. An important feature is that the model of Bgee allows seamless integration of new data types into the same framework. From each new data type, we need to define quality control criteria, conditions for calling gene expression present or absent, and rule of inference of expression ranks. Once this is done, our model will allow views and tools such as the Gene Page, the Expression Comparison, or TopAnat to make use of the new data together with the previously available data. Notably, single-cell RNA-Seq presents an important perspective of combining the anatomical precision of in situ hybridization (and even more) with the broad coverage of microarrays or RNA-Seq. Unlike dedicated single-cell databases, Bgee offers the perspective of a unified view of gene expression from the cellular to the organism level, which we believe will be increasingly relevant.

Materials and methods

Pipeline overview

We retrieve information about species taxonomy and genomes. We collect expression data from different sources, depending on their data type. We curate these data to retain only samples from healthy wild-type individuals. We annotate them to anatomical and life stage ontologies, along with information of population/strain and gender/sex. We perform quality controls to remove low-quality and duplicated samples. We process these data to produce present/absent expression calls, along with expression level information, represented by our expression ranks and expression scores. We propagate these calls along the anatomical and life stage ontologies, to allow the integration of data generated with various granularities (e.g., “cerebellum” vs. “brain, “human adult” vs. “human late adulthood”). See figure 2 for an overview, see sub-sections below for precise descriptions.

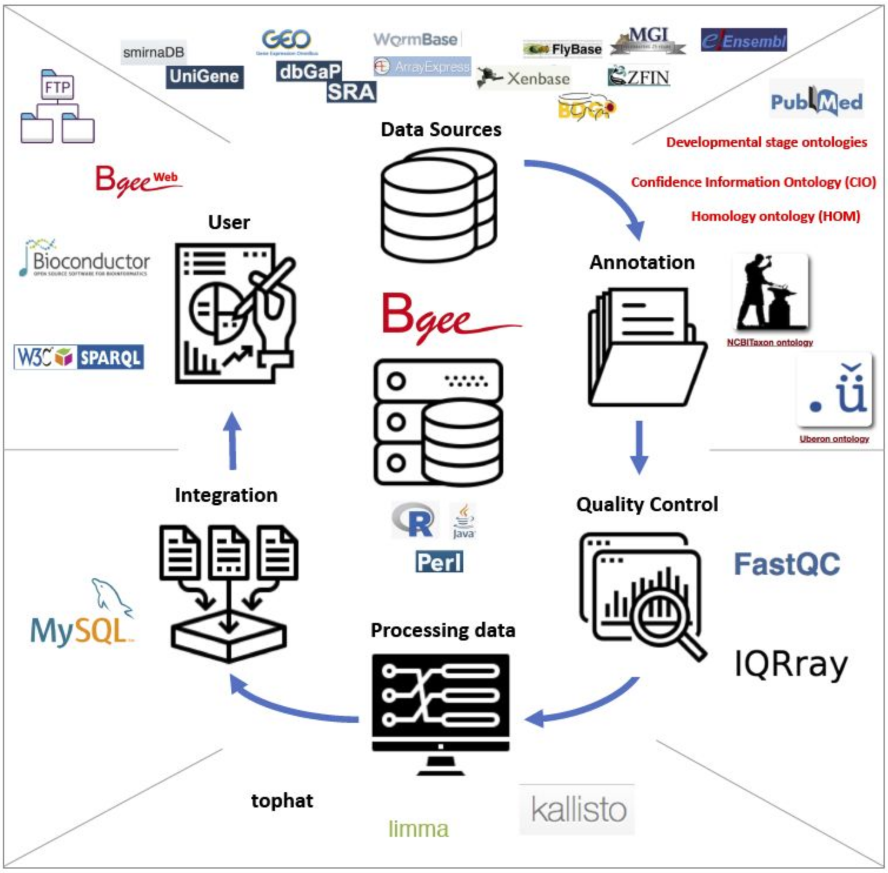


Figure 2: Bgee pipeline overview. Expression data are retrieved from various databases; they are annotated by the Bgee team, or annotations from Model Organism Databases are remapped by the Bgee team, to ontologies describing developmental stages, anatomy, taxa; quality controls are performed, using for instance FastQC for RNA-Seq data, IQRray for Affymetrix data; data are then analyzed using specific tools, such as kallisto to produce TPM values from RNA-Seq data, or limma to compute TMM normalization factors, and

presence/absence expression calls are then produced; all the expression data and analysis results are integrated into the MySQL Bgee database; these data are then leverated by the different tools offered by Bgee: Bgee web-interface, Bioconductor packages, SPARQL endpoint, FTP server.
Icons for tools and databases retrieved from their respective website. Other icons made by smashicons, Becris, monkik, Flat Icons, Eucalyp, phatplus, from www.flaticon.com.

Ontologies and genome annotations are only updated for each major release (e.g., Bgee 13 to 14), while expression data can be updated at minor releases (e.g. Bgee 14.0 to 14.1). Thus all expression data presented here concern Bgee 14.1 specifically, while ontology and genome annotations concern Bgee 14, both 14.0 and 14.1.

Species information: taxonomy, genomes, gene orthology

We annotate species information using the NCBI taxonomy database[39], and integrate the taxonomy for species present in Bgee. We retrieve gene models for each species in Bgee from Ensembl[3] (Ensembl 84[81] for Bgee 14) and Ensembl Metazoa[2] (release 30[82] for Bgee 14) using the Perl Ensembl API. We also retrieve cross-references to other databases, gene biotypes, and annotations to Gene Ontology[8337] terms, and gene orthology information from OMA[70], although not all of these are used in tools accessible to users as of Bgee 14.1.

Uberon integration

To describe the anatomy of diverse animals, we use the Uberon ontology[38]. Its integration into Bgee requires processing, described below. The relevant part of the Bgee pipeline documentation is presented in Sup. Material. The corresponding Bgee pipeline source code and instructions can be found at https://github.com/BgeeDB/bgee_pipeline/tree/master/pipeline/uberon (archived version for Bgee 14 releases: https://github.com/BgeeDB/bgee_pipeline/tree/v14.0/pipeline/uberon). The custom version of Uberon generated for Bgee can be found at https://github.com/BgeeDB/bgee_pipeline/blob/master/generated_files/uberon/ (OBO format: file custom_composite.obo; OWL format: file custom_composite.owl.zip; archived version for Bgee 14 releases: https://github.com/BgeeDB/bgee_pipeline/tree/v14.0/generated_files/uberon).

Taxon constraints

In order for automatic reasoners to determine in which taxa an anatomical entity exists, Uberon uses taxon constraints[8485]. Bgee needs to override some of these taxon constraints. First, because there can be errors in the ontology, which we need to correct. And second, because we need to define taxon constraints for species-specific terms. Indeed, Uberon comes notably in two flavors. One is the ext.owl ontology[86], which is the core version of Uberon, containing the taxon constraint axioms. Another one is the composite-metazoan.owl ontology[87], which merges species ontologies into the structure of Uberon, merging in taxonomic equivalents, relabeling species-specific classes, and also merging in all of the cell ontology[8889]. Bgee integrates the composite-metazoan version, in order to be able to describe precisely species-specific anatomies, with taxon constraints for generic terms from the core ontology. But the species-specific terms (e.g., ZFA terms for zebrafish) lack taxon constraints. We thus override these taxon constraints to define them for species-specific terms.

Uberon simplification

This simplification aims at keeping only the knowledge needed for Bgee, while simplifying it for an easier browsing. We keep in the ontology only some specific relation types (object properties in OWL), and their child relation types (sub-properties): *is_a*, *part_of*, *develops_from*, *transformation_of*. We remove some terms (e.g., CARO:0000006 “material anatomical entity”), while merging their incoming edges with their outgoing edges. We remove some relations that we feel make the ontology more difficult to browse, e.g., UBERON:0000463 “organism substance” *part_of* UBERON:0000468 “multi-cellular organism” (as of Uberon ext release 2019-06-27, this relation is annotated as “"this relationship may be too strong and may be weakened in future"). And we remove some branches we are not interested in, e.g., BFO:0000141 “immaterial anatomical entity”.

Cycle removal

We remove cycles between terms in Uberon. All the cycles were caused by a single issue, see https://github.com/obophenotype/uberon/issues/651. See also documentation at https://github.com/BgeeDB/bgee_pipeline/tree/master/pipeline/uberon for a description of the procedure.

Improvement of mappings to species-specific ontologies

Many terms from species-specific ontologies, used in MODs providing *in situ* hybridization data, lack a mapping to Uberon; we can thus not integrate them into Bgee without corrections. We correct as many as possible of these missing or incorrect mappings (see https://github.com/obophenotype/uberon/issues/664 and https://github.com/obophenotype/uberon/issues/1288 for a report of the issues discovered for Bgee 14).

Developmental stage integration

Each species in Bgee has a developmental stage ontology, available in the obophenotype developmental stage ontology repository[40]. We have developed most of the species-specific ontologies in this repository, i.e. all except those from Model Organism Databases (EMAPA for mouse[59]; FBdv for fly[90]; ZFA for zebrafish[60]; XAO for Xenopus[62]; WBls[63] for worm). The aim of this step is to produce a composite stage ontology merging all of these species-specific ontologies, that we can insert into Bgee. Having one common multi-species ontology, with high level terms merged between species (e.g., “gastrula”), allows us to propagate expression data to comparable stages in different species, and thus take development into account when automatically comparing expression patterns. The merge is performed by the Uberon mainteners. For insertion into Bgee, we use a nested set model, which impacts the representation needed: all terms must have exactly one parent by *part_of* relations; all terms must be uniquely ordered relative to another term. We describe here our strategy for insertion into the Bgee database. The motivation for design choices are described at https://github.com/obophenotype/developmental-stage-ontologies/blob/master/external/bgee/known_issues.md (archived version for Bgee 14 releases: https://github.com/obophenotype/developmental-stage-ontologies/blob/97fa5b7e176f8c7f72cd5c7115d2fc36f35e8bb3/external/bgee/known_issues.md). The resulting merged developmental stage ontology used in Bgee can be found at https://github.com/obophenotype/developmental-stage-ontologies/blob/master/external/bgee/dev_stage_ontology.obo (archived version for Bgee 14 releases: https://github.com/obophenotype/developmental-stage-ontologies/blob/97fa5b7e176f8c7f72cd5c7115d2fc36f35e8bb3/external/bgee/dev_stage_ontology.obo).

Term merging

To merge each species-specific ontology into the structure of Uberon, we add cross-references to associate species-specific terms to Uberon terms they can be merged with. For instance, the term from the human ontology HsapDv:0000010 “gastrula stage” has a cross-reference to the Uberon term UBERON:0000109 “gastrula stage”. HsapDv:0000010 is merged with UBERON:0000109, and all its children are considered as children of UBERON:0000109 (e.g., HsapDv:0000011 “Carnegie stage 06”).

Term ordering

Stages are ordered using the relations “preceded_by” or “immediately_preceded_by”. For integration into Bgee, all terms need to be strictly ordered related to each other, as if all terms possessed a relation “immediately_preceded_by” to another term. This is not always the case. For instance, if the relations in the human ontology provided the information that “Carnegie stage 06” is immediately preceded by “Carnegie stage 05”, and that “Carnegie stage 06a” is the first term in temporal ordering among its siblings, and “Carnegie stage 05c” the last; then the Bgee pipeline can infer that “Carnegie stage 06a” is immediately preceded by “Carnegie stage 05c”.

Of note, for some species, stages can be immediately preceded by more than one term. This is the case for instance in worm, where the stage “L4” can be immediately preceded by either “Dauer”, or “L3”. As of Bgee 14, Bgee cannot accommodate such relations, and we conserve only one of them for insertion into the database.

Also, some developmental stages can be asynchronous in an individual, so that the ontology cannot explicitly state what is their temporal ordering. This is for instance the case in fly for the stages embryonic cycle 15 and 16. As of Bgee 14, Bgee needs an absolute ordering of terms for insertion into the database. We thus add “preceded_by” relations to these terms.

part_of relations

For insertion into Bgee, a stage must have one parent at most by a *part_of* relation. But the merging of species-specific ontologies into Uberon can cause inconsistencies. For instance, in the ontology for *Drosophila melanogaster* FBdv, FBdv:00005349 “pupal stage”, mapped to UBERON:0000070 “pupal stage”, is a child of FBdv:00007001 “P-stage”. But in Uberon, UBERON:0000070 “pupal stage” is a child of UBERON:0000092 “post-embryonic stage”. The merging of FBdv and Uberon results in having UBERON:0000070 “pupal stage” child of both UBERON:0000092 “post-embryonic stage” and FBdv:00007001 “P-stage”. This is not possible for insertion as of Bgee 14. In this case, we discard the term FBdv:00007001 “P-stage”, and we map all its children to UBERON:0000092 “post-embryonic stage” (FBdv:00005342 “prepupal stage”, FBdv:00005349 “pupal stage”, FBdv:00006011 “pharate adult stage”).

Insertion into Bgee

For Bgee 14, we then transform this merged developmental stage ontology into a nested set model (we will likely abandon this approach in future releases, to avoid the issues described above). One of the main difficulties at this step is to order consistently all terms from all species-specific ontologies, related to each other. This cannot be done with a classical sort algorithm, such as merge sort (see [91]). The reason is that some developmental stages in different species cannot be ordered relative to each other. While the classical sorting algorithm considers these stages as equal, it only means that they could not be ordered, not that they truly have the same position in the stage ordering. To overcome this problem, we have implemented a bubble sort algorithm (see [92]), to compare all stages to each other (archived source code for Bgee 14 releases of the sort algorithm: https://github.com/BgeeDB/bgee_apps/blob/Bgee_v14.0/bgee-pipeline/src/main/java/org/bgee/pipeline/ontologycommon/OntologyUtils.java; archived source code for Bgee 14 releases of the sorting algorithm implementation and the generation of the nested set model: https://github.com/BgeeDB/bgee_apps/blob/Bgee_v14.0/bgee-pipeline/src/main/java/org/bgee/pipeline/uberon/UberonDevStage.java).

Curation of expression data

Our curation steps have the aim of selecting data: i) produced from healthy wild-type animals in normal conditions; ii) that are of high quality; iii) that are non-redundant. Because the amount of data available in public repositories is large, we can be stringent regarding our criteria to accept data for inclusion.

Data sources

All data sources are detailed in Table 1. For SRA or GEO studies for which the corresponding publication is missing at time of curation, we determine the publication, and report missing PMIDs to GEO. We use the information from these publications, as well as those which are already linked, for precise annotation of samples to anatomy, stage, sex and/or strain. For example the SRP075519 study contains six RNA-seq libraries, with tissue information reported as 'muscle', while the related paper[93] reports that “RNA-Seq analysis was performed on six piglets representing two breeds: Duroc and Ronghcang (three animals of each breed). RNA was extracted from the longissimus dorsi of each individual.” We can thus annotate the libraries to the relevant term UBERON:0001401 “longissimus thoracis muscle", and we can also annotate each library to the corresponding breed (here 'Duroc' or 'Ronghcang').

Healthy wild-type samples in normal conditions

For inclusion into Bgee, we reject samples from animals with abnormal genetic backgrounds (e.g., project SRX2751118, nono$^{-/-}$ mice), or subject to diseases, to gene knockouts, or to treatments not expected in the wild (e.g., irradiated sample DRX012402). We manually review the information about each sample before inclusion (see Supplementary Material for our criteria for inclusion). It is a difficult task to consistently define what is a normal healthy wild-type sample: while we aim at being stringent regarding the quality of the data we integrate, we also need our data to reflect the genetic diversity of wild-type animals, or the various expected conditions encountered in the wild.

For instance, we have defined whether we could integrate samples used to study the effect of fasting on gene expression: a reasonable fasting time can correspond to conditions that animals encounter in the wild; but an abnormally long fasting time can have adverse effects on the health of the individual, so that it cannot be considered as "healthy" anymore. As a result, we have defined 5-6 hours as being a reasonable fasting time in *Mus musculus*. Another example is the use of lab strain animals: they do not exhibit the exact same phenotypes or gene expression patterns than animals harvested in the wild[94], and some lab or domesticated strains are selected to maximize a given phenotype (e.g., DBA/2J mouse model for experimental glaucoma[95]). But they also represent a part of the genetic diversity of the species studied. For this reason, and because lab strain animals represent the vast majority of the data available, we include them into Bgee, while annotating the strain used whenever possible (see below).

This annotation effort allows us to gather data from a diversity of datasets. For instance, we can use an experiment designed for cancer research, by selecting only the control samples. We can leverage experiments of any size, from those including very few samples, to big data projects. Because we act as external experts when reviewing a dataset, we can discard samples not fitting our criteria objectively.

For *in situ* hybridization data that we retrieve from MODs, we coordinate with developers of these resources to find the appropriate parameters to retrieve only wild-type healthy samples in normal conditions.

Removal of hidden redundancies

It is important for our statistical analyses and generation of our quality scores to avoid the use of redundant data points. During our quality control procedure, we have identified duplicated content in the GEO and ArrayExpress databases, affecting about 14% of our Affymetrix data[96]: fully or partially duplicated experiments from independent data submissions, Affymetrix chips reused in several experiments, or reused within an experiment. We have thus created a method to discard these duplicated samples, as well as to avoid this problem appearing in our RNA-Seq dataset. This allows Bgee, despite the use of potentially overlapping multiple datasets, to provide and use a clean reference of unique samples for downstream analyses.

Curation of GTEx dataset

We have curated the GTEx human dataset phs000424.v6.p1[15]. We apply a stringent re-annotation process to the GTEx data to retain only healthy and non-contaminated samples, using the information available under restricted-access. For instance, we reject all samples for 31% of subjects, deemed globally unhealthy from the pathology report (e.g., drug abuse, diabetes, BMI > 35), as well as specific samples from another 28% of subjects who had local pathologies (e.g., brain from Alzheimer patients). We also reject samples with contamination from other tissues according to the pathologist report. In total, we have kept only 6008 of 11983 samples (50%); these represent a high quality subset of GTEx, restricted to only healthy and non-contaminated samples. We re-annotate some GTEx samples to Uberon ontology, according to the original sampling sites. For instance, we map GTEx samples ‘Minor Salivary Gland - Inner surface of lower lip’ to UBERON:0001830 minor salivary gland, while GTEx reports on UBERON:0006330 anterior lingual gland. Note that both lip inner surface and tongue have minor salivary glands. The precise criteria for our curation of GTEx can be found in Sup. Material. More information to use these data is available at https://bgee.org/?page=doc&action=data_sets, and is also provided in Sup. Material.

Annotation of expression data

We capture information about the anatomical localization of samples, their developmental and life stage, their sex, and their strain or ethnicity. We either manually capture this information using ontologies and controlled vocabularies (for Affymetrix, RNA-Seq, and EST data), or we map existing annotations provided by MODs to these ontologies and vocabularies (for all *in situ* hybridization data, and for some Affymetrix and RNA-Seq data annotations provided by Wormbase for *C. elegans*). For Affymetrix and RNA-Seq data, we also capture whether replicates are technical or biological.

Anatomy is annotated to Uberon[22], a multi-species anatomical ontology allowing to capture anatomical information in any animal species. The use of Uberon is fundamental for the development of Bgee, and we contribute to its development. Developmental and aging stages are annotated to the developmental stage ontologies detailed in “Developmental stage integration”. Sex is annotated to a naive controlled vocabulary including the following terms: “male”, “female”, “hermaphrodite”, “mixed”, and “NA” (information not available at time of annotation).

We maintain a naive controlled vocabulary of strains or breeds, based on the UniProt 'strains.txt' file (https://www.uniprot.org/docs/strains) and completed with strain/breed names found in the literature, or from species specific resources (e.g. https://www.gov.uk/guidance/official-cattle-breeds-and-codes). We standardize strain information provided in free-text format, notably to remove most duplicates, resulting in 341 different strain and ethnicity terms used in Bgee 14.1.

Processing of expression data

RNA-Seq data

We parse Bgee annotations in order to retrieve the relevant information about SRA files to download, using the NCBI e-utils[97] (for instance, SRR IDs of runs part of a library). We download data from SRA[26] using the NCBI SRA Toolkit[98] and the Aspera software[99], and then convert them to FASTQ files[100]. For GTEx data, we download FASTQ files through the dbGaP[25] Authorized Access System[101] using Aspera. GTF annotation files and genome sequence fasta files from Ensembl[3] and Ensembl metazoa[2] are used to identify sequences of genic regions, exonic regions, and intergenic regions. We also use them to generate indexed transcriptome files for all species in Bgee, using TopHat[102] and Kallisto[103]. We use Kallisto to generate pseudo-counts for each transcript, which are then summed for each gene. All further analysis and reporting is done at the gene level as of Bgee 14. TPM and FPKM (or RPKM) values are computed from pseudo-counts, and the results between libraries for each experiment are normalized independently using TMM[104]. The source code and documentation for RNA-Seq data integration can be found at https://github.com/BgeeDB/bgee_pipeline/tree/master/pipeline/RNA_Seq (archived version for release Bgee 14.1: https://github.com/BgeeDB/bgee_pipeline/tree/v14.1/pipeline/RNA_Seq).

Affymetrix data

We retrieve Affymetrix data preferentially as CEL files, or as MAS5[105] processed files when the CEL files are not available. For each chip type, we retrieve the mapping of probesets to genes from Ensembl. We remove hidden redundancy as explained above. We remove low quality chips by using either the IQRray[106] quality score for CEL file data, or the percentage of genes considered as expressed for MAS5 file data. When CEL files are available for an experiment, we normalize the probeset signal intensities using gcRMA[107]. Otherwise, we store the MAS5 flags of expression “present”, “marginal”, “absent”. The source code and documentation for Affymetrix data integration can be found at https://github.com/BgeeDB/bgee_pipeline/tree/master/pipeline/Affymetrix (archived version for release Bgee 14.1: https://github.com/BgeeDB/bgee_pipeline/tree/v14.1/pipeline/Affymetrix).

EST data

For ESTs, we retrieve data from UniGene[35] and smiRNAdb[34] (both now retired; Table 1). We retrieve mappings of UniGene data to genomes, for human, mouse, zebrafish and Xenopus, from Ensembl. For fly, as a mapping is not available from Ensembl, we retrieve cDNA information in FASTA format[108] from Ensembl, and use BLAST[4] to map UniGene clusters to genes. In each library, we simply count the number of ESTs mapped to each gene. The source code and documentation for EST data integration can be found at https://github.com/BgeeDB/bgee_pipeline/tree/master/pipeline/ESTs (archived version for release Bgee 14.1: https://github.com/BgeeDB/bgee_pipeline/tree/v14.1/pipeline/ESTs).

Integration of *in situ* hybridization data from Model Organism Databases

We retrieve *in situ* hybridization data from the following MODs: for *Mus musculus*, from GXD[30]; for *Drosophila melanogaster*, BDGP[109] and FlyBase[29]; for *Danio rerio*, ZFIN[33]; for *Xenopus tropicalis*, Xenbase[32]; for *Caenorhabditis elegans*, WormBase[31]. When possible, our pipeline relies on the use of InterMine[110]. The source code and documentation for *in situ* hybridization data integration can be found at https://github.com/BgeeDB/bgee_pipeline/tree/master/pipeline/In_situ (archived version for release Bgee 14.1: https://github.com/BgeeDB/bgee_pipeline/tree/v14.1/pipeline/In_situ).

Healthy wild type data

For each MOD, we filter out the data from animals with abnormal genetic backgrounds, e.g., animals with transgenes or knockout. We discuss with the developers of these resources to determine how to accurately filter out these data. For instance, for data in *C. elegans*, WormBase has provided us a file containing only healthy wild-type data (Daniela Raciti, personal communication).

Data quality

When no data quality information is provided, we assume by default that *in situ* hybridization data are high quality. When the information is provided, we remap the quality levels from each MOD to Bgee quality levels, and discard samples of low quality (see Sup. Material for correspondences and filtering criteria).

Remapping of annotations

Each MOD uses their own controlled vocabulary or ontology to annotate data: EMAPA and EMAPS for GXD in mouse[111]; FBbt and FBdv for BDGP and FlyBase in fly[90]; ZFA for ZFIN in zebrafish[112]; XAO for Xenbase in Xenopus[113,114]; WBbt[115] and WBls[116] for WormBase in worm. For integration into Bgee, we remap these species-specific terms to Uberon, both for the anatomical localizations and the developmental stages. This is achieved by using cross-references or OWL equivalentClass axioms present in Uberon, mapping to these species-specific ontologies. We have also corrected and added many cross-references as part of our integration of Uberon, and of developmental stage ontologies (see above). Other species-specific terms, not merged to Uberon terms, are also present in the composite-metazoan version of Uberon, allowing us to integrate them when needed.

For some MODs, both the anatomy and developmental stage information are entailed in a single term annotation. It is for instance the case in GXD, that uses the EMAPS ontology. The EMAPS ontology maps to anatomical terms from EMAPA, but targets them at a specific developmental stage (see [111] for details). For integration into Bgee, we transform these annotations to retrieve the anatomical localization on the one hand, and the developmental stage on the other hand.

In some other MODs, expression patterns are captured without keeping a stage-specific identity. In WormBase, a gene could be reported as expressed, for instance, in the pharynx at L1 stage and in neurons from embryo to adult; the resulting annotation would be pharynx and neuron, in embryo, larva and adult. The stage specific information is lost. Even when there is only one stage, a gene could be reported as expressed in the pharynx, vulva, and intestine, and, at stage L1, in the nerve ring. Wormbase would capture L1 for nerve ring, but along with the other anatomical parts (Daniela Raciti, personal communication). For this reason, we keep stage information from WomBase only when there is only one anatomical entity annotated for a gene, so that we can be certain that the annotated stages should be associated to it. For other annotations, we map the stage information to the root of our stage ontology, UBERON:0000104 “life cycle”, meaning that no precise information about stage is captured.

The species-specific ontologies used by MODs also differ in their structure, as compared to Uberon, making the mapping task challenging. For instance, the EMAPA ontology makes no distinction between “myocardium” and “cardiac muscles”: “myocardium” is a synonym of the term EMAPA:32688 “cardiac muscle tissues”. EMAPA:32688 is mapped to 3 different terms in Uberon: UBERON:0002349 “myocardium”, UBERON:0001133 “cardiac muscle tissue”, and UBERON:0004493 “cardiac muscle tissue of myocardium”. In this case, Bgee retains the mapping to the most specific term, which is here UBERON:0004493 “cardiac muscle tissue of myocardium”. In other cases, an arbitrary choice has to be made, and only one mapping is kept in the ontology.

Verification of annotations and sex inference

Several sanity checks are automatically performed on all annotations, from Bgee and from MODs. We verify that the anatomical entity or developmental stage used in annotation is expected to exist in the species annotated. This checks both our annotations and the Uberon taxon constraints. We verify that the anatomical entity is consistent with the sex annotated, to avoid, for instance, an annotation of “ovary” in “male”. This information is inferred from Uberon, and by using information captured by Bgee about the sexes existing in species (for instance, to know whether hermaphrodite individuals can exist in the species). The inference of the sexes where anatomical entities can exist is reported in our pipeline source code at https://github.com/BgeeDB/bgee_pipeline/tree/master/generated_files/uberon/uberon_sex_info.tsv (archived version for release Bgee 14 releases: https://github.com/BgeeDB/bgee_pipeline/tree/v14.0/generated_files/uberon/uberon_sex_info.tsv). When the sex information is not available for a sample, or we did not capture it at time of annotation, we use the same information to infer the sex automatically. For instance, in species with no hermaphrodite individuals, an “ovary” annotation allows to infer that the sex is “female”. We store whether the sex is inferred, or originally annotated.

Calls of presence/absence of expression

Bgee provides calls of presence or absence of expression for unique combinations of a gene in a condition. As of Bgee 14.1, a condition is defined by 5 information: an anatomical entity, a life stage, a sex, a strain or ethnicity, and a species.

Calls produced from RNA-Seq data

We use a new method to estimate for each RNA-Seq library independently the TPM threshold to consider a gene as actively transcribed (Julien Roux, Marta Rosikiewicz, Julien Wollbrett, Marc Robinson-Rechavi, Frederic B. Bastian; unpublished). While this method will be described elsewhere, the documentation of our pipeline describing its use in Bgee is available at https://github.com/BgeeDB/bgee_pipeline/tree/master/pipeline/RNA_Seq (archived version for release Bgee 14.1: https://github.com/BgeeDB/bgee_pipeline/tree/v14.1/pipeline/RNA_Seq). Briefly, we use all RNA-seq data from a species to identify a stringent set of intergenic regions, then use these regions to define the background level of read mapping per library. Then, we call genes as expressed when their level of mapped reads is significantly higher than this background, and as not expressed otherwise. As of Bgee 14.1, RNA-Seq calls are all considered of high quality.

Calls produced from Affymetrix data

When only the MAS5 files of an analysis are available, we use the flags provided by the MAS5 software[117]. Although MAS5 classification is efficient[118], the estimation of the background signal can be biased depending on probe sequence affinity[41]. For this reason, we use preferentially CEL files when available to produce present/absent calls, and, when not available, we consider all calls produced from MAS5 files as low quality. MAS5 “present” and “marginal” flags correspond in Bgee to a low quality present call; “absent” flag to a low quality absent call. For CEL data, we use the gcRMA algorithm[107] to normalize the signal taking into account probe sequences, and use a subset of weakly expressed probesets for estimating the background, as described in [41]. We then apply a Wilcoxon test to compare the normalized signal of the probesets with the background signal, as implemented in the ‘mas5calls’ function of the bioconductor package ‘affy’[118]: High quality present call corresponds to a p-value threshold of < 0.03 (corresponding to a FDR of ≈5%, see Schuster *et al.*[119]), low quality present call to a p-value >= 0.03 and <= 0.12, and high quality absent call to a p-value > 0.12 (< 0.12 corresponds to a FDR ≈ 10%).
We then exclude all probesets that are never seen as MAS5 “present” or CEL high quality present over the whole dataset. Because a same gene can be covered by several probesets, we reconcile this information to produce one call per gene and per chip, by retaining the best probeset signal, in this order: high quality present, low quality present, high quality absent, low quality absent.

Calls produced from EST data

Based on the number of ESTs mapped to a gene for which the 95% confidence interval of the EST count excludes 0 [42], we call of presence of expression of high quality when an experiment has at least 7 ESTs mapped to a gene, and of low quality from 1 to 6 ESTs. Of note, we do not produce calls of absence of expression from EST data because of the low sampling.

Calls produced from *in situ* hybridization data

In situ hybridization data are already provided as calls of presence/absence of expression. Bgee retrieves these data directly from MODs. Some MODs also provide a quality level for the calls, that we have mapped to either high or low qualities (see above).

Remapping to more generalized conditions

We annotate and retrieve expression data with as much granularity as possible. For instance, in human, when the information was available, we annotate expression data with the exact age (e.g., HsapDv:0000150 “56-year-old human stage”). But when it comes to integrating all expression data to produce expression calls, such a granularity can hamper call integration and comparison. For this reason, to produce expression calls, we remap the granular annotations of the data to more generalized conditions.

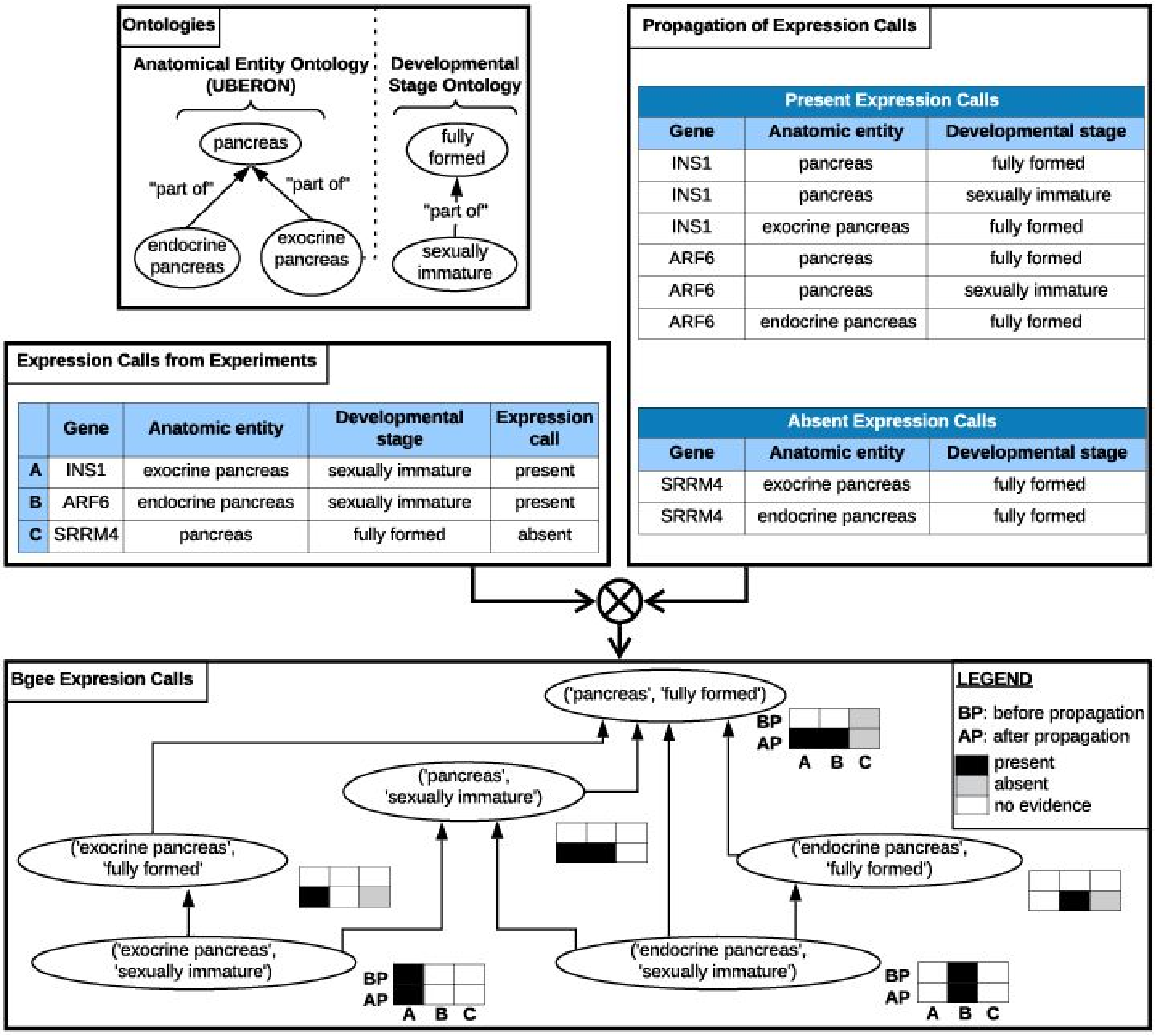
For developmental stages, we have created in developmental stage ontologies a subset named “granular_stage”. When some expression data are annotated to a stage part of this subset, we remap the stage to its closest parent not part of this subset, for expression call generation. For instance, the term HsapDv:0000150 “56-year-old human stage” in the previous example would be automatically remapped to the term HsapDv:0000092 “human middle aged stage”. All calls are mapped to the relevant sex, whether it is inferred or manually annotated.

In this way, users can still benefit from the highest granularity possible when retrieving our annotated and processed expression values. But the data are more generalized for call generation, allowing more powerful integration and comparisons.

Call propagation

After producing these calls from multiple techniques and experiments, we propagate them along the graph of conditions (see Figure 3 for an overview). We produce the graph of conditions by: i) using the graph of anatomical entities (Uberon ontology) from *is_a* (subClassOf in OWL) and *part_of* relations (an object property in OWL); ii) using the graph of developmental stages (developmental stage ontology) from *part_of* relations; iii) adding a root parent to all sex terms by *is_a* relation; iv) adding a root parent to all strain terms by *is_a* relation. For instance, for the condition [UBERON:0001891 “midbrain”, HsapDv:0000084 “2-5 year-old child stage”, “female”, “caucasian”, “Homo sapiens”], a parent condition would be [UBERON:0000955 “brain”, HsapDv:0000081 “child stage”, “any sex”, “wild-type”, “Homo sapiens"]; a direct child condition would be [UBERON:0019267 “gray matter of midbrain”, HsapDv:0000096 “2-year-old human stage”, “female”, “caucasian”, “Homo sapiens”].

We propagate calls of presence of expression along the graph of conditions to all parent conditions. The idea is that if a gene is expressed, e.g., in the midbrain, it is expressed in the brain, the parent structure. We propagate calls of absence of expression to direct anatomical sub-structures, keeping the same developmental stage, sex, and strain. We propagate only to direct anatomical sub-structures, and not all sub-structures, because we consider that an experiment could miss the expression of a gene in a very small part of a larger anatomical entity. We do not propagate to child terms for stages, sexes, or strains, as most of the time an experiment is not performed by mixing samples from all stages of a time period, or all sexes or strains of a species.


Ontologies
Anatomical Entity Ontology (UBERON)
Developmental Stage Ontology
pancreas
"part of"
"part of"
endocrine pancreas
exocrine pancreas
fully formed
"part of"
sexually immature
Propagation of Expression Calls
Expression Calls from Experiments
Bgee Expresion Calls
('pancreas', 'fully formed')
BP
AP
A B C
('pancreas', 'sexually immature')
('exocrine pancreas', 'fully formed'
('endocrine pancreas', 'fully formed')
('exocrine pancreas', 'sexually immature')
BP
AP
A B C
('endocrine pancreas', 'sexually immature')
BP
AP
A B C
LEGEND
BP: before propagation
AP: after propagation
present
absent
no evidence

Present Expression Calls

Gene	Anatomic entity	Developmental stage
INS1	pancreas	fully formed
INS1	pancreas	sexually immature
INS1	exocrine pancreas	fully formed
ARF6	pancreas	fully formed
ARF6	pancreas	sexually immature
ARF6	endocrine pancreas	fully formed

Absent Expression Calls

Gene	Anatomic entity	Developmental stage
SRRM4	exocrine pancreas	fully formed
SRRM4	endocrine pancreas	fully formed

	Gene	Anatomic entity	Developmental stage	Expression call
A	INS1	exocrine pancreas	sexually immature	present
B	ARF6	endocrine pancreas	sexually immature	present
C	SRRM4	pancreas	fully formed	absent

Figure 3: propagation of calls of presence/absence of expression. Calls of presence/absence of expression are produced from the raw data (left table), for instance: call of presence of expression for gene INS1 in exocrine pancreas at sexually immature developmental stage; call of presence of expression for gene ARF6 in endocrine pancreas at sexually immature developmental stage; call of absence of expression for gene SRRM4 in pancreas at fully formed developmental stage. A graph of conditions is generated by using the anatomical

ontology and the developmental stage ontology, to allow propagation of expression calls (top left box): for instance, the condition “endocrine pancreas - sexually immature” is a child of the condition “pancreas - fully formed”; the condition “endocrine pancreas - fully formed” is a parent of the condition “endocrine pancreas - sexually immature”. Calls of presence of expression are propagated to all parent conditions; calls of absence of expression are propagated to direct child conditions (top right box). This propagation of calls allow the integration of data that were produced and annotated with different granularity: for instance, while before propagation there was information in “pancreas - fully formed” only for the gene SRRM4, after propagation the expression of the three genes can be compared in this condition (bottom box).

Of note, call propagation varies between species, as the localization of some anatomical entities can vary between species. For instance, the brain structure islands of Calleja is part of the olfactory tubercle in rodents, but part of the nucleus accumbens in primates. Expression calls produced in islands of Calleja are then propagated to different parent structures depending on the species. This information is captured in Uberon thanks to OWL General Class Inclusion (GCI) axioms. For the previous example, it notably includes the axiom: 'islands of Calleja' (UBERON:0001881) and *part_of* some Primates (NCBITaxon:9443) SubClassOf *part_of* some 'nucleus accumbens' (UBERON:0001882).

Integration to produce one global call and confidence level per gene - condition combination

Multiple calls from multiple data types and assays can be produced for a given gene in a given condition, from direct observation or from propagation along the condition graph. We integrate all these individual calls to produce one global call of presence/absence of expression per gene - condition, along with a confidence level (gold, silver, or bronze).

Global calls of absence of expression are reported when all experiments consistently report an absence of expression for a gene in a condition, with no conflicting presence of expression reported in the condition itself, or any sub-condition. Otherwise, a global call of presence of expression is reported: presence of expression always “wins” over absence of expression, whatever the quality level and number of experiments producing the call of presence of expression. This means that Bgee is very stringent when reporting absence of expression.

We follow the principles of the Confidence Information Ontology (CIO)[36]. These basic principles are that an assertion supported by several experiments is more reliable than an assertion supported by a single experiment, and that an assertion supported by experiments of several data types is more reliable than by a single data type. We translate this principle in Bgee into three confidence levels for calls of presence/absence of expression: gold, silver, bronze. We assign these confidence levels depending on the number of experiments supporting a call, and the level of support (low or high) from each experiment to produce the call.

After call propagation and reconciliation of presence/absence calls, we count for each call the number of experiments supporting the call with high quality evidence, and the number of experiments supporting the call with low quality evidence. We assign gold confidence level to global calls supported by at least two different experiments with a high quality call; silver to global calls supported by either only one experiment with a high quality call, or at least two different experiments with a low quality call; and bronze level to global calls supported by only one experiment with a low quality call. Because of call propagation, we also consider in support of a global call the experiments which contribute to calls in sub-conditions (for calls of presence of expression) or in direct parent conditions (for calls of absence of expression).

Expression ranks and expression scores

In addition to calls of presence/absence, Bgee provides a more quantitative ranking of the conditions where a gene is expressed, also integrated over experiments and data types. To compute expression ranks: 1) we compute ranks using different methods per data type; 2) we normalize ranks over all genes, conditions and data types for each species; 3) we compute a global weighted mean rank for each gene in each condition over all data types considered. We also transform these ranks into expression scores, more easily understandable by users: higher gene expression translates into lower rank but higher expression score, from 0 to 100.

As of Bgee 14.1, we compute ranks using only data annotated to a condition itself (no propagation), and only considering anatomical entity, developmental stage, and species (i.e., over all sexes and strains indifferently). The rank of a gene in an anatomical entity for a species is the minimum of its individual ranks at each developmental stage with data for it. The pipeline source code and documentation can be found at https://github.com/BgeeDB/bgee_pipeline/tree/master/pipeline/post_processing (archived version for Bgee release 14.1: https://github.com/BgeeDB/bgee_pipeline/tree/v14.1/pipeline/post_processing).
Of note, this pipeline also offers the possibility to compute ranks by pooling all data in a condition itself and its sub-conditions. But we have not yet evaluated and released these ranks, although they would allow to quantify expression levels in many more conditions.

Rank computation for each data type

RNA-Seq

The Perl script to generate ranks for RNA-Seq data can be found at https://github.com/BgeeDB/bgee_pipeline/blob/master/pipeline/post_processing/ranks_rnaseq.pl (archived version for Bgee release 14.1: https://github.com/BgeeDB/bgee_pipeline/tree/v14.1/pipeline/post_processing/ranks_rnaseq.pl).

To compute ranks for RNA-Seq data, we first identify the set of valid genes that should be considered, and that is always the same for computing ranks in all libraries: the set of all genes that received at least one read over all libraries in Bgee for this species. Then, in each library, we compute fractional ranks of genes, based on their TPM value in this library. For each gene and each condition with RNA-Seq data in the condition itself, we compute a weighted mean of the gene ranks, using all libraries annotated to this condition itself in the species. We weigh the mean by the number of distinct ranks in each library, under the assumption that libraries with a higher number of distinct ranks have a higher power for ranking genes.

To be able to normalize ranks between conditions and data types, we store the maximum fractional rank over all libraries in each condition and species (and not the maximum weighted mean). To be able to compute the global weighted mean rank of a gene over all data types considered in a condition, we store the sum of the number of distinct ranks in each library, for each condition and species. For example, if we have a set of 3 genes to rank, and 2 libraries annotated to the same condition; in library 1, gene 1 has rank 1, gene 2 has rank 2, and gene 3 has rank 3; in library 2, gene 1 has rank 1, gene 2 has rank 2.5, and gene 3 has rank 2.5; the maximum rank in this condition is 3, and the sum of the number of distinct ranks is 3 + 2 = 5.

Affymetrix

The Perl script to generate ranks for Affymetrix data as of Bgee 14.1 is available at https://github.com/BgeeDB/bgee_pipeline/blob/master/pipeline/post_processing/ranks_affymetrix.pl (archived version for Bgee release 14.1: https://github.com/BgeeDB/bgee_pipeline/tree/v14.1/pipeline/post_processing/ranks_affymetrix.pl).

To compute ranks for Affymetrix data, we first compute fractional ranks for each gene, for each Affymetrix chip, based on the highest signal intensity from all the probesets mapped to this gene. Then, we normalize ranks between chips annotated to the condition itself in this species. Indeed, different chip types do not have the same probeset design, and thus the same number of genes represented on it. The ranks are then inconsistent between different chip types.

For each chip type, we compute its max rank over all conditions. We normalize ranks of each chip based on the max rank of its corresponding chip type, compared to the max of the max ranks of all chip types used in the same condition. The idea is to correct for the different genomic coverage of different chip types. We do not normalize simply based on the max rank in a given condition, but based on the max rank of the chip types represented in that condition, to not penalize conditions with a lower number of expressed genes, or higher number of ex-aequo genes and lower fractional max ranks. For example, if two chip types are represented in a condition, and the max rank of chipType 1 over all conditions is 1,000, and the max rank of chipType 2 over all conditions is 2,000; the putative rank of a gene represented on chipType 1 could then be shifted, as compared to genes on chipType 2, by a factor of (2000/1000). We thus normalize fractional ranks of genes on chipType 1 by considering the average of their actual rank and their maximal shifted rank: (fractional_rank + fractional_rank * 2000/1000)/2.

We then compute the weighted mean of the normalized ranks per gene and condition, weighting by the number of distinct ranks in each chip, under a similar assumption to RNA-Seq that chips with higher number of distinct ranks have a higher power for ranking genes. Similar to RNA-Seq, to be able to normalize ranks between conditions and data types, we store the max of max ranks of chip types represented in each condition and species. To be able to compute the global weighted mean rank of a gene over all data types considered in a condition, we store the sum of the number of distinct ranks in each chip, for each condition and species.

Of note, for RNA-Seq data, we do not normalize ranks between samples in the same condition before computing the weighted mean, as for Affymetrix data: we use all libraries to produce ranking over always the same set of genes in a given species, so the genomic coverage is always the same, and no normalization is required. The higher power at ranking genes of a library (for instance, thanks to a higher number of mapped reads) is taken into account by weighting the mean by the number of distinct ranks in the library, not by normalizing libraries with lower max ranks; this would penalize conditions with a lower number of expressed genes, and thus with more ex-aequo ranked genes, corresponding to genes receiving 0 read.

In situ hybridization

Since *in situ* hybridization data are not quantitative, we use an approach based on the assumption that the more often an expression information is reported, the more biologically important this expression is likely to be. The Perl script to generate ranks for *in situ* hybridization data is available at https://github.com/BgeeDB/bgee_pipeline/blob/master/pipeline/post_processing/ranks_in_situ.pl (archived version for Bgee release 14.1: https://github.com/BgeeDB/bgee_pipeline/tree/v14.1/pipeline/post_processing/ranks_in_situ.pl).

To compute ranks for *in situ* hybridization data, we compute a score for each gene in each condition, based on the detection flag of *in situ* evidence (present, absent), and the quality level (high quality, low quality). Each spot is given the following score: [present, high quality], 1; [present, low quality], 0.5; [absent, low quality], -0.5; [absent, high quality], -1. Then we sum these scores for each gene in each condition. We consider all experiments pooled together, because each *in situ* experiment usually studies a very limited number of genes. We compute a dense ranking of the genes in each condition, based on the scores computed. Again, we have considered all experiments pulled together. A fractional ranking is not appropriate, because there are many ex-aequo, leading to an artificially high max rank. This gives too much weight to *in situ* hybridization data when computing a global weighted mean rank over all data types. To be able to normalize ranks between conditions and data types, and to compute the global weighted mean rank of a gene over all data types considered in a condition, we store the max rank in each condition and species.

ESTs

While EST data are in principle quantitative, their very low coverage leads us to treat them similarly to *in situ* hybridization data. The Perl script used is available at https://github.com/BgeeDB/bgee_pipeline/blob/master/pipeline/post_processing/ranks_est.pl (archived version for Bgee release 14.1: https://github.com/BgeeDB/bgee_pipeline/tree/v14.1/pipeline/post_processing/ranks_est.pl).

To compute ranks for EST data, we sum for each gene in each condition the count of ESTs mapped to this gene, from all EST libraries in this condition pooled. We compute a dense ranking of the genes in each condition, based on the EST count sums, considering all libraries in the condition all together. A fractional ranking was not appropriate for the same reason as for *in situ* hybridization data. To be able to normalize ranks between conditions and data types, and to compute the global weighted mean rank of a gene over all data types considered in a condition, we store the max rank in each condition and species.

Integration over all data types and experiments

Normalization over all data types and conditions

Different experiments and techniques might have different power and resolution at ranking genes in a same condition. For instance, some *in situ* hybridization data might allow to rank 5 genes in a condition; while RNA-Seq data in the same condition might allow to rank 30,000 genes, with a max rank of 25,000, and 20,000 distinct ranks. The ranks from different experiments and data types can thus be highly heterogeneous, and difficult to compare as such. To overcome this limitation, we normalize ranks by taking into account the max rank for each data type, condition and species, as compared to the max rank over all data in this species. We thus normalize ranks by considering the average of their actual rank and their maximal shifted rank: (rank + rank * (species_max_rank/data_type_condition_max_rank))/2. The Perl script performing this normalization step is available at https://github.com/BgeeDB/bgee_pipeline/blob/master/pipeline/post_processing/normalize_ranks.pl (archived version for Bgee release 14.1: https://github.com/BgeeDB/bgee_pipeline/tree/v14.1/pipeline/post_processing/normalize_ranks.pl).

We normalize the mean ranks for each gene in each condition and for each data type, based on the max rank for this data type in this condition and species, as compared to the max rank over all conditions and data types in this species. For instance, if the max rank in mouse over all data types and conditions is 30,000, and the dense ranking of a gene in a condition by *in situ* hybridization data is 10, and the max rank by *in situ* hybridization data in this condition in mouse is 30, the dense ranking of the gene will become (10 + 10 * 30,000/30)/2 = 5,005. While this will receive a low weight if there is more informative data (see below), it allows to compute a rank in the absence of other data type in this condition. It also avoids always ranking conditions studied only by *in situ* hybridization data at the top.

Computation of global weighted mean rank for each gene and condition over all data types

This computation is done in the Java Bgee API directly, to be able to choose the data types to consider to compute the ranks, based on the user request. The Java source code performing this computation is available at https://github.com/BgeeDB/bgee_apps/blob/master/bgee-dao-sql/src/main/java/org/bgee/model/dao/mysql/expressiondata/MySQLGlobalExpressionCallDAO.java (archived version for Bgee release 14.1: https://github.com/BgeeDB/bgee_apps/blob/Bgee_v14.1/bgee-dao-sql/src/main/java/org/bgee/model/dao/mysql/expressiondata/MySQLGlobalExpressionCallDAO.java).

With the information computed per data type we compute a global weighted mean rank for each gene and condition. All types can be used, or only some. This mean is computed by using the normalized mean ranks or normalized dense ranks for a gene in a condition, from each data type considered. For RNA-Seq and Affymetrix data, if they are considered and data exist in this condition for this gene, the mean rank is weighted by the sum of the number of distinct ranks in this condition and species for the corresponding data type. For *in situ* hybridization and EST data, if they are considered and data exist in this condition for this gene, the dense rank is weighted by the max rank in this condition and species for the corresponding data type.

We transform global weighted mean ranks into expression scores. To compute the expression score of a gene in a condition, we retrieve the max rank over all conditions and data types considered, for this species; and the global weighted mean rank for a gene in a condition, computed over all data types considered and data available, as described above. The expression score is equal to (MaxRank + 1 - meanRank) * (100/maxRank).

Funding

We acknowledge support by the SIB Swiss Institute of Bioinformatics, the Canton de Vaud, Swiss National Science Foundation grants 31003A_173048, 31003A_153341, 31003A_133011, and CRSII3_160723, SystemsX project AgingX, PNR grant 407540_167149 (BioSODA), NIH Award Number U01CA215010 (OncoMX). This project has also received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 863410.

Acknowledgment

Chris Mungall, Melissa Haendel, Suzanna Lewis, David Osumi-Sutherland, Daniela Raciti, and Torsten Henrich helped and coordinated with us concerning ontologies and annotation. Vincent Laudet was instrumental in helping the project get started. Bgee was also improved by very helpful discussions with Amos Bairoch, Pascale Gaudet, Lydie Lane, Christophe Dessimoz, Adrian Altenhoff, Natasha Glover, and many members of the SIB Swiss Institute of Bioinformatics and SIB UX team; as well as many members of the biocuration and OBO Foundry communities, notably the teams of the Uberon project, the Model Organisms Databases BDGP, FlyBase, GXD, WormBase, Xenbase, and ZFIN, as well as the Alliance for Genome Resources (AGR), the ArrayExpress team, and the owlcollab and obophenotype communities. The NCBI Taxonomy rapidly added new taxa when needed for homology annotations.

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data integrated in Bgee described in this manuscript were obtained from dbGaP accession number phs000424.v6.p1.
Finally, we thank all members and alumni of the Robinson-Rechavi lab; special thanks to Nadezda Kryuchkova for discussions on tissue-specificity, and to Andrea Komljenovic and Tina Begum for suggestions of data to annotate and for use-cases.

References

1. Haeussler, M. *et al.* The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res.* **47**, D853–D858 (2018).
2. Kersey, P. J. *et al.* Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Res.* **46**, D802–D808 (2018).
3. Yates, A. D. *et al.* Ensembl 2020. *Nucleic Acids Res.* (2019) doi:10.1093/nar/gkz966.
4. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
5. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
6. Ritchie, G. R. S., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nat. Methods* **11**, 294–296 (2014).
7. Burke, A. C., Nelson, C. E., Morgan, B. A. & Tabin, C. Hox genes and the evolution of vertebrate axial morphology. *Development* **121**, 333 (1995).
8. Carroll, S. B. *Endless forms most beautiful: the new science of evo devo and the making of the animal kingdom*. (Norton, 2005).
9. Roux, J., Rosikiewicz, M. & Robinson-Rechavi, M. What to compare and how: Comparative transcriptomics for Evo-Devo. *J. Exp. Zoolog. B Mol. Dev. Evol.* **324**, 372–382 (2015).
10. Brown, G. R. *et al.* Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.* **43**, D36–D42 (2014).

11. Fagerberg, L. *et al.* Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell. Proteomics MCP* **13**, 397–406 (2014).

12. Yue, F. *et al.* A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355 (2014).

13. Stelzer, G. *et al.* The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses: The GeneCards Suite. in *Current Protocols in Bioinformatics* (eds. Bateman, A., Pearson, W. R., Stein, L. D., Stormo, G. D. & Yates, J. R.) 1.30.1-1.30.33 (John Wiley & Sons, Inc., 2016). doi:10.1002/cpbi.5.

14. Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 6062–6067 (2004).

15. Carithers, L. J. *et al.* A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreservation Biobanking* **13**, 311–319 (2015).

16. *Sample exclusion in GTEx dataset*. https://gtexportal.org/home/faq#sampleExclusion

17. Rung, J. & Brazma, A. Reuse of public genome-wide gene expression data. *Nat. Rev. Genet.* **14**, 89 (2012).

18. Papatheodorou, I. *et al.* Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Res.* **46**, D246–D251 (2017).

19. Palasca, O., Santos, A., Stolte, C., Gorodkin, J. & Jensen, L. J. TISSUES 2.0: an integrative web resource on mammalian tissue expression. *Database* **2018**, (2018).

20. Vize, P. D. & Westerfield, M. Model organism databases. *genesis* **53**, 449–449 (2015).

21. Howe, D. G. *et al.* Model organism data evolving in support of translational medicine. *Lab Anim.* **47**, 277–289 (2018).

22. Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. miRBase: from microRNA sequences to function. *Nucleic Acids Res.* **47**, D155–D162 (2019).

23. Altenhoff, A. M. *et al.* The OMA orthology database in 2018: retrieving evolutionary

relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Res.* **46**, D477–D485 (2018).

24. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* **41**, D991–D995 (2012).
25. Tryka, K. A. *et al.* NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res.* **42**, D975–D979 (2013).
26. Kodama, Y. *et al.* The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.* **40**, D54–D56 (2011).
27. Athar, A. *et al.* ArrayExpress update - from bulk to single-cell expression data. *Nucleic Acids Res.* **47**, D711–D715 (2019).
28. Hammonds, A. S. *et al.* Spatial expression of transcription factors in Drosophila embryonic organ development. *Genome Biol.* **14**, R140 (2013).
29. Thurmond, J. *et al.* FlyBase 2.0: the next generation. *Nucleic Acids Res.* **47**, D759–D765 (2018).
30. Smith, C. M. *et al.* The mouse Gene Expression Database (GXD): 2019 update. *Nucleic Acids Res.* **47**, D774–D779 (2018).
31. Harris, T. W. *et al.* WormBase: a modern Model Organism Information Resource. *Nucleic Acids Res.* (2019) doi:10.1093/nar/gkz920.
32. Karimi, K. *et al.* Xenbase: a genomic, epigenomic and transcriptomic model organism database. *Nucleic Acids Res.* **46**, D861–D868 (2017).
33. Howe, D. G. *et al.* ZFIN, the Zebrafish Model Organism Database: increased support for mutants and transgenics. *Nucleic Acids Res.* **41**, D854–D860 (2012).
34. Landgraf, P. *et al.* A Mammalian microRNA Expression Atlas Based on Small RNA Library Sequencing. *Cell* **129**, 1401–1414 (2007).
35. Pontius, J. U., Wagner, L. & Schuler, G. D. *UniGene: A unified view of the transcriptome. In: The NCBI Handbook*. (2004).

36. Bastian, F. B. *et al.* The Confidence Information Ontology: a step towards a standard for asserting confidence in annotations. *Database* **2015**, (2015).

37. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).

38. Haendel, M. A. *et al.* Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon. *J. Biomed. Semant.* **5**, 21 (2014).

39. Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Res.* **40**, D136–D143 (2011).

40. *Developmental stage ontology collection*. https://github.com/obophenotype/developmental-stage-ontologies

41. Schuster, E. F., Blanc, E., Partridge, L. & Thornton, J. M. Correcting for sequence biases in present/absent calls. *Genome Biol.* **8**, R125 (2007).

42. Audic, S. & Claverie, J.-M. The Significance of Digital Gene Expression Profiles. *Genome Res.* **7**, 986–995 (1997).

43. Jeffery, J. E., Bininda-Emonds, O. R. P., Coates, M. I. & Richardson, M. K. A New Technique for Identifying Sequence Heterochrony. *Syst. Biol.* **54**, 230–240 (2005).

44. Roux, J. & Robinson-Rechavi, M. An ontology to clarify homology-related concepts. *Trends Genet.* **26**, 99–102 (2010).

45. *Anatomical similarity annotations*. https://github.com/BgeeDB/anatomical-similarity-annotations

46. Yon Rhee, S., Wood, V., Dolinski, K. & Draghici, S. Use and misuse of the gene ontology annotations. *Nat. Rev. Genet.* **9**, 509–515 (2008).

47. Dessimoz, C. & Škunca, N. *The Gene Ontology Handbook*. (Humana Press New York, NY, USA:, 2017).

48. Alexa, A., Rahnenfuhrer, J. & Lengauer, T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**,

1600–1607 (2006).

49. Adrian Alexa, J. R. *topGO*. (Bioconductor, 2017). doi:10.18129/b9.bioc.topgo.

50. Jabbari, K. & Nürnberg, P. A genomic view on epilepsy and autism candidate genes. *Genomics* **108**, 31–36 (2016).

51. Ha, S., Sohn, I.-J., Kim, N., Sim, H. J. & Cheon, K.-A. Characteristics of Brains in Autism Spectrum Disorder: Structure, Function and Connectivity across the Lifespan. *Exp. Neurobiol.* **24**, 273–284 (2015).

52. Schmidt-Rhaesa, A. *The evolution of organ systems*. (Oxford University Press, 2007).

53. Zheng, W. *et al.* Comparative Transcriptome Analyses Indicate Molecular Homology of Zebrafish Swimbladder and Mammalian Lung. *PLoS ONE* **6**, e24019 (2011).

54. Zaccone, D. *et al.* Morphology and innervation of the teleost physostome swim bladders and their functional evolution in non-teleostean lineages. *Acta Histochem.* **114**, 763–772 (2012).

55. *SRRM4 expression comparison in mammals*. https://bgee.org/bgee14_1/?page=expression_comparison&data=34beddfc93bb7fbb440e757e6de24d91fc0ce177

56. *Pharyngeal gill homology*. https://bgee.org/bgee14_1/?page=anat_similarities&species_list=9606&species_list=7955&ae_list=UBERON%3A0000206

57. Graham, A., Okabe, M. & Quinlan, R. The role of the endoderm in the development and evolution of the pharyngeal arches: Endoderm in the development and evolution of the pharyngeal arches, A. Graham et al. *J. Anat.* **207**, 479–487 (2005).

58. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **12**, 115–121 (2015).

59. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2018).

60. Komljenovic, A., Roux, J., Wollbrett, J., Robinson-Rechavi, M. & Bastian, F. B. BgeeDB, an R package for retrieval of curated expression datasets and for gene list expression localization enrichment tests. *F1000Research* **5**, 2748 (2018).

61. Segaran, T., Taylor, J. & Evans, C. *Programming the Semantic Web*. (O'Reilly, 2009).

62. Sima, A. C. *et al. Enabling Semantic Queries Across Federated Bioinformatics Databases*. http://biorxiv.org/lookup/doi/10.1101/686600 (2019) doi:10.1101/686600.

63. Braasch, I. *et al.* A new model army: Emerging fish models to study the genomics of vertebrate Evo-Devo: NEW FISH MODELS FOR VERTEBRATE EVO-DEVO-GENO. *J. Exp. Zoolog. B Mol. Dev. Evol.* **324**, 316–341 (2015).

64. Howe, D. *et al.* The future of biocuration. *Nature* **455**, 47–50 (2008).

65. International Society for Biocuration. Biocuration: Distilling data into knowledge. *PLOS Biol.* **16**, e2002846 (2018).

66. Tang, Y. A. *et al.* Ten quick tips for biocuration. *PLOS Comput. Biol.* **15**, e1006906 (2019).

67. The SIB Swiss Institute of Bioinformatics' resources: focus on curated databases. *Nucleic Acids Res.* **44**, D27–D37 (2016).

68. Common Fund (CF) Genotype-Tissue Expression Project (GTEx) - Phenotype Datasets phs000424.v8.p2.

69. the FAIRsharing Community *et al.* FAIRsharing as a community approach to standards, repositories and policies. *Nat. Biotechnol.* **37**, 358–367 (2019).

70. Altenhoff, A. M. *et al.* OMA standalone: orthology inference among public and custom genomes and transcriptomes. *Genome Res.* **29**, 1152–1163 (2019).

71. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2018).

72. Ahi, E. P., Richter, F., Lecaudey, L. A. & Sefc, K. M. Gene expression profiling suggests differences in molecular mechanisms of fin elongation between cichlid

species. *Sci. Rep.* **9**, 9052 (2019).

73. Briggs, J. A. *et al.* The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360**, eaar5780 (2018).

74. Fushan, A. A. *et al.* Gene expression defines natural changes in mammalian lifespan. *Aging Cell* **14**, 352–365 (2015).

75. Holland, P. W., Holland, L. Z., Williams, N. A. & Holland, N. D. An amphioxus homeobox gene: sequence conservation, spatial expression during development and insights into vertebrate evolution. *Dev. Camb. Engl.* **116**, 653–661 (1992).

76. Lam, S. H. *et al.* Conservation of gene expression signatures between zebrafish and human liver tumors and tumor progression. *Nat. Biotechnol.* **24**, 73–75 (2006).

77. Sudmant, P. H., Alexis, M. S. & Burge, C. B. Meta-analysis of RNA-seq expression data across species, tissues and studies. *Genome Biol.* **16**, 287 (2015).

78. Naqvi, S. *et al.* Conservation, acquisition, and functional impact of sex-biased gene expression in mammals. *Science* **365**, eaaw7317 (2019).

79. Trabzuni, D. *et al.* Widespread sex differences in gene expression and splicing in the adult human brain. *Nat. Commun.* **4**, 2771 (2013).

80. Eling, N., Morgan, M. D. & Marioni, J. C. Challenges in measuring and understanding biological noise. *Nat. Rev. Genet.* **20**, 536–548 (2019).

81. *Ensembl 84*. http://mar2016.archive.ensembl.org/

82. *Ensembl Metazoa 30*. http://ensemblgenomes.org/info/release-notes/30

83. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).

84. Deegan (née Clark), J. I., Dimmer, E. C. & Mungall, C. J. Formalization of taxon-based constraints to detect inconsistencies in annotation and ontology development. *BMC Bioinformatics* **11**, 530 (2010).

85. Chris Mungall, J. D. Taxon constraints in Uberon.

https://github.com/obophenotype/uberon/wiki/Taxon-constraints

86. *Uberon ext.owl*. http://purl.obolibrary.org/obo/uberon/ext.owl

87. *Uberon composite-metazoan.owl*. http://purl.obolibrary.org/obo/uberon/composite-metazoan.owl

88. Meehan, T. F. *et al.* Logical Development of the Cell Ontology. *BMC Bioinformatics* **12**, 6 (2011).

89. *Cell Ontology*. http://obofoundry.org/ontology/cl.html

90. Costa, M., Reeve, S., Grumbling, G. & Osumi-Sutherland, D. The Drosophila anatomy ontology. *J. Biomed. Semant.* **4**, 32 (2013).

91. *Algorithms and complexity: third Italian conference, CIAC '97, Rome, Italy, March 12-14, 1997: proceedings*. (Springer, 1997).

92. Astrachan, O. Bubble sort: an archaeological algorithmic analysis. *ACM SIGCSE Bull.* **35**, 1 (2003).

93. Wang, J. *et al.* Convergent and divergent genetic changes in the genome of Chinese and European pigs. *Sci. Rep.* **7**, 8662 (2017).

94. Yoshiki, A. & Moriwaki, K. Mouse phenome research: implications of genetic background. *ILAR J.* **47**, 94–102 (2006).

95. Turner, A. J., Vander Wall, R., Gupta, V., Klistorner, A. & Graham, S. L. DBA/2J mouse model for experimental glaucoma: pitfalls and problems. *Clin. Experiment. Ophthalmol.* **45**, 911–922 (2017).

96. Rosikiewicz, M., Comte, A., Niknejad, A., Robinson-Rechavi, M. & Bastian, F. B. Uncovering hidden duplicated content in public transcriptomics data. *Database* **2013**, (2013).

97. Entrez Programming Utilities Help. (2010). https://www.ncbi.nlm.nih.gov/books/NBK25501/

98. *SRA Toolkit download*. https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft/

99. *Aspera software*. https://asperasoft.com/software/clients/

100. Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* **38**, 1767–1771 (2009).

101. GaP FAQ Archive - Authorized Access System Login. (2009). https://www.ncbi.nlm.nih.gov/books/NBK63573/

102. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).

103. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525 (2016).

104. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).

105. Hubbell, E., Liu, W.-M. & Mei, R. Robust estimators for expression analysis. *Bioinformatics* **18**, 1585–1592 (2002).

106. Rosikiewicz, M. & Robinson-Rechavi, M. IQRray, a new method for Affymetrix microarray quality control, and the homologous organ conservation score, a new benchmark method for quality control metrics. *Bioinformatics* **30**, 1392–1399 (2014).

107. Wu, Z., Irizarry, R. A., Gentleman, R., Martinez-Murillo, F. & Spencer, F. A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *J. Am. Stat. Assoc.* **99**, 909–917 (2004).

108. Lipman, D. & Pearson, W. Rapid and sensitive protein similarity searches. *Science* **227**, 1435–1441 (1985).

109. *BDGP Berkeley Drosophila Genome Project*. https://insitu.fruitfly.org

110. Kalderimis, A. *et al.* InterMine: extensive web services for modern biology. *Nucleic Acids Res.* **42**, W468–W472 (2014).

111. Hayamizu, T. F., Baldock, R. A. & Ringwald, M. Mouse anatomy ontologies:

enhancements and tools for exploring and integrating biomedical data. *Mamm. Genome* **26**, 422–430 (2015).

112. Van Slyke, C. E., Bradford, Y. M., Westerfield, M. & Haendel, M. A. The zebrafish anatomy and stage ontologies: representing the anatomy and development of Danio rerio. *J. Biomed. Semant.* **5**, 12 (2014).

113. Segerdell, E., Bowes, J. B., Pollet, N. & Vize, P. D. An ontology for Xenopus anatomy and development. *BMC Dev. Biol.* **8**, 92 (2008).

114. Segerdell, E. *et al.* Enhanced XAO: the ontology of Xenopus anatomy and development underpins more accurate annotation of gene expression and queries on Xenbase. *J. Biomed. Semant.* **4**, 31 (2013).

115. *C. elegans Gross Anatomy Ontology*. http://www.obofoundry.org/ontology/wbbt.html

116. *C. elegans Development Ontology*. http://www.obofoundry.org/ontology/wbls.html

117. Liu, W. -m *et al.* Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics* **18**, 1593–1599 (2002).

118. Choe, S. E., Boutros, M., Michelson, A. M., Church, G. M. & Halfon, M. S. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biol.* **6**, R16 (2005).

119. Schuster, E. F., Blanc, E., Partridge, L. & Thornton, J. M. Correcting for sequence biases in present/absent calls. *Genome Biol.* **8**, R125 (2007).

120. *Zenodo Bgee intergenic community*. https://zenodo.org/communities/bgee_intergenic

121. Kuśnierczyk, W. Taxonomy-based partitioning of the Gene Ontology. *J. Biomed. Inform.* **41**, 282–292 (2008).

122. Smith, B. *et al.* The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* **25**, 1251 (2007).

123. Chris Mungall. Taxon constraints in OWL. *Monkeying around with OWL* https://douroucouli.wordpress.com/2012/04/24/taxon-constraints-in-owl/ (2012).

124. Feldman, S. I. Make—A program for maintaining computer programs. *Softw. Pract. Exp.* **9**, 255–265 (1979).

125. Gosling, J., Joy, B., Steele, G. L., Bracha, G. & Buckley, A. *The Java Language Specification, Java SE 8 Edition. 1st*. (Addison-Wesley Professional, 2014).

126. Reenskaug, T. A note on DynaBook requirements. *Xerox PARC Available Httpheim Ifi Uio No~ Trygverthemesmvcmvc-Index Html Accesed 21 March 2013* (1979).

127. Rosikiewicz, M. & Robinson-Rechavi, M. IQRray, a new method for Affymetrix microarray quality control, and the homologous organ conservation score, a new benchmark method for quality control metrics. *Bioinformatics* **32**, 2565–2565 (2016).

Supplementary material

Additional resources

In order to serve the development of Bgee, we have contributed to the development of additional resources, listed here: i) a collection of developmental stage ontologies; and ii) an ontology capturing the information present in the NCBI Taxonomy database.

BgeeCall R package

The BgeeCall R package allows to reproduce a specific part of the Bgee pipeline: the generation of present/absent expression calls from RNA-Seq data (see “Material and methods”). Our method avoids the use of a fixed threshold to consider a gene as expressed (such as 1 RPKM, or 2 TPM). It is based on the estimation of background experimental and transcriptional noise from expression level of intergenic regions (Julien Roux, Marta Rosikiewicz, Julien Wollbrett, Marc Robinson-Rechavi, Frederic B. Bastian; in preparation). The BgeeCall R package allows to generate present/absent calls for any RNA-Seq library, as long as the species studied is present in Bgee. It is also possible to use it for other species, if a set of validated intergenic regions is provided (see the Zenodo ‘bgee_intergenic’ community[120] for more details). The BgeeCall Bioconductor page can be accessed at https://bioconductor.org/packages/BgeeCall/.

NCBITaxon ontology

In order to compare expression patterns between species, and to identify homologous anatomical entities to do so, we need to be able to retrieve the common ancestor of any set of species in Bgee. Moreover, to define taxon constraints in a multi-species ontology, a taxonomy ontology is needed (see, e.g., [121] and [84]). To this aim, the NCBI taxonomy database[39] has been translated into an ontology (and more specifically, a tree, where each term has only one parent, except the root taxon “cellular organisms”, which has no parent). Each term in the ontology represents a taxon; and each taxon is linked to its direct ancestral taxon by the OBO *is_a* relation (equivalent to *SubClassOf* in OWL). The ontology is part of the OBOFoundry[122], and is available at http://obofoundry.org/ontology/ncbitaxon.html.

But in order to compute taxon constraints in Uberon (see Material and methods), we needed to use a version of the taxon ontology that includes disjoint classes axioms between sibling taxa[123]. These axioms are not included in the public version of the ontology. We have thus produced a custom version, including these axioms, for the species present in Bgee. This version can be found at https://github.com/BgeeDB/bgee_pipeline/blob/master/generated_files/species/bgee_ncbitaxon.owl (archived version for Bgee 14 releases: https://github.com/BgeeDB/bgee_pipeline/blob/v14.0/generated_files/species/bgee_ncbitaxon.owl).

Source code

Bgee pipeline source code

The source code of the entire Bgee pipeline is available at https://github.com/BgeeDB/bgee_pipeline/. The pipeline is decomposed in different steps, focused on a specific part of the pipeline (e.g., genome integration, RNA-Seq analyses). For each step, we have created a Makefile[124], allowing to list all the computations necessary, and the dependencies between the computations. And for each step, a README file documents it directly in the step folder.

Bgee application source code

The source code of the Bgee application is available at https://github.com/BgeeDB/bgee_apps. The Bgee application is developed in Java[125]. It is structured in different modules: *bgee-dao-api* defines the interfaces to access to the data source; *bgee-dao-sql* is an implementation of *bgee-dao-api* to access the Bgee MySQL database; *bgee-core* is the business layer of the application, or model layer in a model-view-controller architecture[126], making use of *bgee-dao-api*; *bgee-pipeline* contains the Java components used in the Bgee pipeline, making use of *bgee-core* and *bgee-dao-sql*; *bgee-webapp* contains the code to build the Bgee website, using the *bgee-core* module, representing the view and controller layers in a model-view-controller architecture.

IQRray source code

To perform quality controls of Affymetrix data where raw CEL files are available, we have developed the IQRray[127] method. IQRray outperforms other methods in identification of poor quality arrays in datasets composed of arrays from many independent experiments. IQRray is implemented in R, and the source code is available at https://github.com/BgeeDB/IQRray.

GTEx data into Bgee

In addition to the continuous growth of transcriptomics datasets, some specific projects produce large amounts of data, generated and accessible in a consistent manner, as, notably, the GTEx project. The GTEx project aims at building a comprehensive resource for tissue-specific gene expression in human. Here we describe how this dataset was integrated into Bgee.

Annotation process

We applied a stringent re-annotation process to the GTEx data to retain only healthy tissues and non-contaminated samples, using the information available under restricted-access (see Supplementary Material for more details). For instance, we rejected all samples for 31% of subjects, deemed globally unhealthy from the pathology report (e.g., drug abuse, diabetes, BMI > 35), as well as specific samples from another 28% of subjects who had local pathologies (e.g., brain from Alzheimer patients). We also rejected samples with contamination from other tissues.

In total, only 50% of samples were kept; these represent a high quality subset of GTEx. All these samples were re-annotated manually to specific Uberon anatomy and aging terms.

GTEx data into Bgee

All corresponding RNA-seq were reanalyzed in the Bgee pipeline, consistently with all other healthy RNA-seq from human and other species. These data are being made available both through the website, and through BgeeDB R package (with sensitive information hidden).

GTEx data into our website

- Annotations can be retrieved from RNA-Seq human experiments/libraries info. Experiment ID of GTEx is 'SRP012682'.
- Processed expression values, from GTEx only, are available on our FTP (download file).
- Gene expression calls are included into human files
- Each human gene page includes GTEx data if there is any (search a gene here).
- TopAnat analyses can be performs here, which leverage the power of the abundant GTEx data integrated with many smaller datasets to provide biological insight into gene lists.

GTEx data using BgeeDB R package

More information and examples can be found on the BgeeDB R package page.

Annotations can be retrieved from RNA-Seq human experiments/libraries information. Experiment ID of GTEx is 'SRP012682'.

```
  source("https://bioconductor.org/biocLite.R")
  biocLite("BgeeDB")
  library(BgeeDB)
  bgee <- Bgee$new(species = "Homo_sapiens", dataType = "rna_seq")
  myAnnotation <- getAnnotation(bgee)
```

Quantitative expression data and presence calls for GTEx can be loaded.

```
  bgee <- Bgee$new(species = "Homo_sapiens", dataType = "rna_seq")
  # This step can take a lot of time as all Bgee GTEx data have to be downloaded and
uncompressed.
  dataGTEx <- getData(bgee, experimentId = "SRP012682")
```

TopAnat analyses can be performs, which leverage the power of the abundant GTEx data integrated with many smaller datasets to provide biological insight into gene lists.

```
  bgee <- Bgee$new(species = "Homo_sapiens")
  myTopAnatData <- loadTopAnatData(bgee)
  # Retrieve all genes with data in Bgee
  allGenes <- unique(row.names(myTopAnatData$gene2anatomy))
  # List of genes related to autism and epilepsy from Jabbari 2016
  genesOfInterest <- c("ENSG00000183044", "ENSG00000085563", "ENSG00000006071",
"ENSG00000153086",
  "ENSG00000243989", "ENSG00000156110", "ENSG00000150594",
"ENSG00000239900",
  "ENSG00000141385", "ENSG00000038002", "ENSG00000142208",
"ENSG00000275199",
  "ENSG00000117020", "ENSG00000163631", "ENSG00000159423",
"ENSG00000112294",
  "ENSG00000164904", "ENSG00000033011", "ENSG00000182858",
"ENSG00000101901",
  "ENSG00000119523", "ENSG00000214160", "ENSG00000088035",
"ENSG00000159063",
  "ENSG00000086848", "ENSG00000242110", "ENSG00000137074",
"ENSG00000124198",
  "ENSG00000118520", "ENSG00000198844", "ENSG00000131089",
"ENSG00000100299",
  "ENSG00000113273", "ENSG00000004848", "ENSG00000104763",
"ENSG00000108381",
  "ENSG00000066279", "ENSG00000138363", "ENSG00000159363",
"ENSG00000018625",
  "ENSG00000174437", "ENSG00000182220", "ENSG00000185344",
"ENSG00000171953",
  "ENSG00000175054", "ENSG00000158321", "ENSG00000086062",
"ENSG00000103507",
  "ENSG00000074582", "ENSG00000176697", "ENSG00000157764",
"ENSG00000106009",
  "ENSG00000164061", "ENSG00000169814", "ENSG00000111678",
"ENSG00000130921",
  "ENSG00000131943", "ENSG00000197603", "ENSG00000141837",
"ENSG00000007402",
  "ENSG00000182389", "ENSG00000198668", "ENSG00000143933",
"ENSG00000147044",
  "ENSG00000036828", "ENSG00000110395", "ENSG00000015133",
"ENSG00000108691",
  "ENSG00000136861", "ENSG00000008086", "ENSG00000064309",
"ENSG00000151849",
  "ENSG00000103995", "ENSG00000173575", "ENSG00000100888",
"ENSG00000168539",
  "ENSG00000181072", "ENSG00000120903", "ENSG00000101204",
"ENSG00000175344",
  "ENSG00000274542", "ENSG00000160716", "ENSG00000114859",
"ENSG00000073464",
  "ENSG00000186510", "ENSG00000184908", "ENSG00000188603",
"ENSG00000102805",
  "ENSG00000128973", "ENSG00000182372", "ENSG00000278220",
"ENSG00000184144",
  "ENSG00000278728", "ENSG00000174469", "ENSG00000166685",
"ENSG00000168434",
  "ENSG00000213380", "ENSG00000142173", "ENSG00000173085",
"ENSG00000088682",
  "ENSG00000006695", "ENSG00000014919", "ENSG00000047457",
"ENSG00000165078",
  "ENSG00000157184", "ENSG00000169372", "ENSG00000147571",
"ENSG00000160213",
  "ENSG00000064601", "ENSG00000117984", "ENSG00000174080",
"ENSG00000115827",
  "ENSG00000077279", "ENSG00000100150", "ENSG00000181192",
"ENSG00000091140",
  "ENSG00000101152", "ENSG00000116675", "ENSG00000116641",
"ENSG00000172269",
  "ENSG00000000419", "ENSG00000136908", "ENSG00000179085",
"ENSG00000188641",
  "ENSG00000197102", "ENSG00000157540", "ENSG00000101210",
"ENSG00000096093",
  "ENSG00000111361", "ENSG00000119718", "ENSG00000070785",
"ENSG00000115211",
  "ENSG00000145191", "ENSG00000170370", "ENSG00000133216",
"ENSG00000112425",
  "ENSG00000178607", "ENSG00000140374", "ENSG00000105379",
"ENSG00000171503",
   "ENSG00000103089", "ENSG00000122591", "ENSG00000145982",
"ENSG00000091483",
   "ENSG00000112367", "ENSG00000196924", "ENSG00000162769",
"ENSG00000119686",
   "ENSG00000110195", "ENSG00000170345", "ENSG00000125740",
"ENSG00000176165",
   "ENSG00000160973", "ENSG00000087086", "ENSG00000179163",
"ENSG00000022355",
   "ENSG00000166206", "ENSG00000187730", "ENSG00000113327",
"ENSG00000054983",
   "ENSG00000141012", "ENSG00000130005", "ENSG00000171766",
"ENSG00000105607",
   "ENSG00000140905", "ENSG00000131095", "ENSG00000170266",
"ENSG00000178445",
   "ENSG00000074047", "ENSG00000145888", "ENSG00000109738",
"ENSG00000173540",
   "ENSG00000087258", "ENSG00000159921", "ENSG00000111670",
"ENSG00000090581",
   "ENSG00000135677", "ENSG00000108433", "ENSG00000171723",
"ENSG00000233276",
   "ENSG00000176884", "ENSG00000183454", "ENSG00000273079",
"ENSG00000152822",
   "ENSG00000169919", "ENSG00000138796", "ENSG00000170445",
"ENSG00000172534",
   "ENSG00000164588", "ENSG00000138622", "ENSG00000213614",
"ENSG00000049860",
   "ENSG00000165102", "ENSG00000153187", "ENSG00000158104",
"ENSG00000174775",
   "ENSG00000276536", "ENSG00000072506", "ENSG00000114378",
"ENSG00000181873",
   "ENSG00000010404", "ENSG00000127415", "ENSG00000134049",
"ENSG00000166333",
   "ENSG00000124313", "ENSG00000150995", "ENSG00000120071",
"ENSG00000278458",
   "ENSG00000275867", "ENSG00000111262", "ENSG00000169282",
"ENSG00000069424",
   "ENSG00000184408", "ENSG00000140015", "ENSG00000151704",
"ENSG00000177807",
   "ENSG00000187486", "ENSG00000156113", "ENSG00000281151",
"ENSG00000075043",
   "ENSG00000184156", "ENSG00000107147", "ENSG00000243335",
"ENSG00000068796",
   "ENSG00000276734", "ENSG00000168280", "ENSG00000185467",
"ENSG00000118162",
   "ENSG00000133703", "ENSG00000087299", "ENSG00000196569",
"ENSG00000143815",
   "ENSG00000108231", "ENSG00000121897", "ENSG00000138095",
"ENSG00000187391",
   "ENSG00000169032", "ENSG00000126934", "ENSG00000109339",
"ENSG00000204406",
   "ENSG00000090674", "ENSG00000147316", "ENSG00000169057",
"ENSG00000081189",
   "ENSG00000164073", "ENSG00000168282", "ENSG00000100427",
"ENSG00000124615",
   "ENSG00000164172", "ENSG00000129255", "ENSG00000178802",
"ENSG00000177000",
   "ENSG00000198793", "ENSG00000196091", "ENSG00000108784",
"ENSG00000072864",
   "ENSG00000275911", "ENSG00000125356", "ENSG00000131495",
"ENSG00000023228",
   "ENSG00000213619", "ENSG00000164258", "ENSG00000115286",
"ENSG00000110717",
   "ENSG00000167792", "ENSG00000049759", "ENSG00000223957",
"ENSG00000204386",
   "ENSG00000234343", "ENSG00000228691", "ENSG00000227129",
"ENSG00000184494",
   "ENSG00000227315", "ENSG00000234846", "ENSG00000196712",
"ENSG00000151092",
   "ENSG00000187566", "ENSG00000087303", "ENSG00000164190",
"ENSG00000156574",
   "ENSG00000074181", "ENSG00000141458", "ENSG00000119655",
"ENSG00000122585",
   "ENSG00000185149", "ENSG00000213281", "ENSG00000179915",
"ENSG00000079482",
   "ENSG00000116329", "ENSG00000112038", "ENSG00000187848",
"ENSG00000135124",
   "ENSG00000007168", "ENSG00000125779", "ENSG00000173599",
"ENSG00000165194",
   "ENSG00000160299", "ENSG00000131828", "ENSG00000148459",
"ENSG00000164494",
   "ENSG00000127980", "ENSG00000108733", "ENSG00000142655",
"ENSG00000121680",
   "ENSG00000164751", "ENSG00000215193", "ENSG00000034693",
"ENSG00000139197",
   "ENSG00000124587", "ENSG00000112357", "ENSG00000102144",
"ENSG00000156531",
   "ENSG00000092621", "ENSG00000165195", "ENSG00000108474",
"ENSG00000165282",
   "ENSG00000124155", "ENSG00000060642", "ENSG00000121879",
"ENSG00000184381",
   "ENSG00000182621", "ENSG00000123560", "ENSG00000140650",
"ENSG00000039650",
  "ENSG00000108439", "ENSG00000140521", "ENSG00000115138",
"ENSG00000131238",
  "ENSG00000102103", "ENSG00000139174", "ENSG00000163637",
"ENSG00000100033",
  "ENSG00000167371", "ENSG00000197746", "ENSG00000185920",
"ENSG00000179295",
  "ENSG00000172053", "ENSG00000151552", "ENSG00000155961",
"ENSG00000132155",
  "ENSG00000108557", "ENSG00000146282", "ENSG00000078328",
"ENSG00000167281",
  "ENSG00000189056", "ENSG00000163933", "ENSG00000155906",
"ENSG00000104889",
  "ENSG00000136104", "ENSG00000172922", "ENSG00000067836",
"ENSG00000151835",
  "ENSG00000101347", "ENSG00000138760", "ENSG00000144285",
"ENSG00000105711",
  "ENSG00000136531", "ENSG00000153253", "ENSG00000183873",
"ENSG00000196876",
  "ENSG00000169432", "ENSG00000130489", "ENSG00000073578",
"ENSG00000178980",
  "ENSG00000152217", "ENSG00000127990", "ENSG00000181523",
"ENSG00000164690",
  "ENSG00000108061", "ENSG00000138083", "ENSG00000064651",
"ENSG00000124140",
  "ENSG00000119899", "ENSG00000135917", "ENSG00000106688",
"ENSG00000110436",
  "ENSG00000079215", "ENSG00000102743", "ENSG00000125454",
"ENSG00000177542",
  "ENSG00000117394", "ENSG00000164414", "ENSG00000117620",
"ENSG00000181830",
  "ENSG00000076351", "ENSG00000144290", "ENSG00000142319",
"ENSG00000276996",
  "ENSG00000165970", "ENSG00000130821", "ENSG00000198689",
"ENSG00000072501",
  "ENSG00000108055", "ENSG00000166311", "ENSG00000102172",
"ENSG00000163877",
  "ENSG00000115904", "ENSG00000104450", "ENSG00000152583",
"ENSG00000166068",
  "ENSG00000197694", "ENSG00000102359", "ENSG00000126091",
"ENSG00000115525",
  "ENSG00000124356", "ENSG00000123473", "ENSG00000136854",
"ENSG00000144455",
  "ENSG00000139531", "ENSG00000148290", "ENSG00000008056",
"ENSG00000197283",
  "ENSG00000227460", "ENSG00000102003", "ENSG00000198198",
"ENSG00000164458",
  "ENSG00000136463", "ENSG00000143374", "ENSG00000054611",
"ENSG00000162065",
  "ENSG00000145979", "ENSG00000184058", "ENSG00000143178",
"ENSG00000196628",
  "ENSG00000177426", "ENSG00000175606", "ENSG00000061938",
"ENSG00000166340",
  "ENSG00000213689", "ENSG00000165699", "ENSG00000103197",
"ENSG00000154743",
  "ENSG00000274672", "ENSG00000275165", "ENSG00000274796",
"ENSG00000274078",
  "ENSG00000278712", "ENSG00000273896", "ENSG00000170892",
"ENSG00000274129",
  "ENSG00000278605", "ENSG00000278622", "ENSG00000182173",
"ENSG00000175894",
  "ENSG00000104833", "ENSG00000131462", "ENSG00000128159",
"ENSG00000198431",
  "ENSG00000114062", "ENSG00000104517", "ENSG00000173218",
"ENSG00000137411",
  "ENSG00000236178", "ENSG00000234032", "ENSG00000206476",
"ENSG00000230985",
  "ENSG00000223494", "ENSG00000213585", "ENSG00000165637",
"ENSG00000197969",
  "ENSG00000141252", "ENSG00000196998", "ENSG00000075702",
"ENSG00000186153",
  "ENSG00000169554", "ENSG00000043355")
  # Build the gene vector for the analysis
  geneList <- factor(as.integer(unique(allGenes) %in% genesOfInterest))
  names(geneList) <- unique(allGenes)
  # Run the test
  myTopAnatObject <- topAnat(myTopAnatData, geneList)
  resFis <- runTest(myTopAnatObject, algorithm ="elim", statistic ="fisher")
  # Format results
  tableOver <- makeTable(myTopAnatData, myTopAnatObject, resFis, 0.1)
  tableOver
```

GTEx filtering criteria

Overall summary

1. We selected subjects with biological characteristics matched by Bgee requirements. We created a set of “accepted subjects” by filtering via a subset of variables reported in "phs000424.v6.pht002742.v6.p1.c1.GTEx_Subject_Phenotypes.GRU.txt". We came up with three category tags:
    1.1. “YES”: subjects for which we had no evidence for unmatched Bgee requirements
    1.2. “NO”: subjects for which we had evidence for large set of unmatched Bgee requirements (eg. systemic disease)
    1.3. “PENDING”: subjects for which we had evidence for restricted set of unmatched Bgee requirements (eg. localized disease)
2. We selected samples from “YES” and “PENDING” subjects. The samples were selected from "phs000424.v6.pht002743.v6.p1.c1.GTEx_Sample_Attributes.GRU.txt" file, based on the first part of the GTEx_SampleID (exact string match with GTEx_SubjectID) and created a set of “accepted samples”:
    2.1. “YES” subjects: select samples for which we had no evidence for unmatched Bgee requirements, aka select all the samples from this subject
    2.2. “PENDING” subjects: select samples for which we had no evidence for unmatched Bgee requirements
3. We filtered accepted samples:
    3.1. We checked assay type for accepted samples, based on the SMGEBTCHT [‘Type of genotype or expression batch’] variable in “phs000424.v6.pht002743.v6.p1.c1.GTEx_Sample_Attributes.GRU.txt" file. We only accepted TrueSeq.v1(RNA-seq).
    3.2. We checked tissue quality for accepted samples, based on free-text SMPTHNTS [‘Pathology Notes’] variable in “phs000424.v6.pht002743.v6.p1.c1.GTEx_Sample_Attributes.GRU.txt" file. We predefined a set of quality requirements. We defined a “tissue quality” variable. We came up with two category tags for this variable:
        3.2.1. “GOOD”: samples for which we had no evidence for unmatched Bgee quality requirements
        3.2.2. “BAD”: samples for which we had evidence for unmatched Bgee quality requirements
    3.3. For all the samples retained so far by our Bgee filters, we then checked the status of the anatomical mapping provided by GTEx (SMUBRID/SMUBRTRM, Uberon ID and Uberon name) by analysing three variables describing tissue location (SMSMPSTE/SMTS/SMTSD) in “phs000424.v6.pht002743.v6.p1.c1.GTEx_Sample_Attributes.GRU.txt" file. We defined a “mapping quality” variable. We came up with two category tags for this variable:

3.3.1. “GOOD”: samples for which the Uberon mapping proposed by GTEx was consistent with the tissue location information.

3.3.2. “BAD”: samples for which the Uberon mapping proposed by GTEx was not fully consistent with the tissue location information.

4. we then perform a review of Uberon mapping:
    4.1. We proposed a new Uberon mapping for filtered samples of “BAD” mapping quality
    4.2. We kept as it is the GTEx Uberon mapping for those samples filtered as “GOOD” mapping quality
5. We built the final annotation file, with the samples to be processed into the Bgee pipeline

How did we select GTEx subjects whose samples could be included in Bgee?

Only "normal" expression is considered in Bgee (i.e., no treatment, no disease, no gene knock-out...). Since we are dealing with data from human source, it would be too conservative to only keep perfectly healthy subjects, thus we had to find a compromise. We annotated in Bgee only samples from subjects for which we had no evidence for abnormal status (eg. diseases, non-ordinary physiological parameters...).

We excluded from Bgee all samples from subjects suffering from potentially systemic diseases (eg. fungal, bacterial, viral infections), drug/medicine abuse (cocaine, heroin, prescription pills...), abnormally high BMI (above an arbitrarily set threshold of BMI=35), and other "selected" diseases following discussions with annotators (eg. gonorrhea, lupus, sex-transmitted diseases...); we also excluded subjects for which we had no details on reason of death or if the subject was reported to be ineligible for GTEx. Samples coming from cultured cell lines were also excluded (including EBV-transformed cells). We selectively excluded tissues sampled from unhealthy organs while the subject was otherwise judged to be healthy or affected by a disease not reported to lead to systemic infections in literature: for example, if a subject was reported to be suffering from "ascites", then we excluded liver samples from this subject, but we kept all samples from other organ source for the analyzed subject. We considered as includable in Bgee all samples from individuals which could not be excluded following the given parameters:

We defined a "subject status" variable, harbouring three levels: "YES", "PENDING", O". We assigned one of the levels in variable "status" to each subject provided in GTEx v6.0, depending on the subject's features. We looked up the subject features in the provided file "phs000424.v6.pht002742.v6.p1.c1.GTEx_Subject_Phenotypes.GRU.txt".

In particular, for status assignment of each subject, we were interested in the following columns:

- BMI : Autocalculated field of BMI: general indicator of the body fat an individual is carrying based upon the ratio of weight to height. We excluded from Bgee all samples from subjects having a BMI value greater than 35.

- INCEXC : A verification field of whether the Donor has met the overall eligibility criteria for GTEx collection based on answers to eligibility questions. We excluded from Bgee all samples from subjects that were reported "false" for this field.
- DTHFUCOD / DTHCOD: First Underlying Cause Of Death / Immediate Cause of Death. We excluded from Bgee all samples from subjects that were reported to died from: drug/medicine abuse/intoxication, cancer, end-stage diseases (eg. kidney, liver...)
- DTHHRDY: Death Classification: 4-point Hardy Scale. We excluded from Bgee all samples from subjects that were tagged with a Hardy Scale level of 4 (aka: Slow death Death after a long illness, with a terminal phase longer than 1 day -commonly cancer or chronic pulmonary disease-; deaths that are not unexpected) only if DTHFUCOD / DTHCOD and DTHHRDY are consistent (eg: a subject had pneumonia and died of pneumonia, DTHHRDY 4).
- LBHBCABM / LBHBCABT / LBHBSAB / LBHBSAG: Hepatitis B virus infection serology analysis. We excluded from Bgee all samples from subjects that were tagged as "positive" (=1) in at least one of these columns.
- LBHCV1NT / LBHBHCVAB: Flaviviridae infection serology analysis. We excluded from Bgee all samples from subjects that were tagged as "positive" (=1) in at least one of these columns.
- LBHBHCVAB / LBHIV1NT / LBHIVAB / LBHIVO: HIV serology analysis. We excluded from Bgee all samples from subjects that were tagged as "positive" (=1) in at least one of these columns.
- LBPRRVDRL / LBRPR: Syphilis or syphilis agent serology analysis. We excluded from Bgee all samples from subjects that were tagged as "positive" (=1) in at least one of these columns.
- MHALS: Amytrophis Lateral Sclerosis. We excluded from Bgee all samples from subjects that were tagged as "positive" (=1) in this column.
- MHALZDMT / MHALZHMR: Alzheimer. We excluded from Bgee brain samples coming from subjects that were tagged as "positive" (=1) in at least one of these columns.
- MHARTHTS: Arthritis. We excluded from Bgee all samples from subjects that were tagged as "positive" (=1) in this column.
- MHASCITES: Ascites. We excluded from Bgee liver samples coming from subjects that were tagged as "positive" (=1) in this column.
- MHBCTINF: Bacterial infection. We excluded from Bgee all samples from subjects that were tagged as "positive" (=1) in this column.
- MHCANCER5 / MHCANCERC / MHCANCERNM: Cancer detection or cancer history. We excluded from Bgee all samples from subjects that were tagged as "positive" (=1) in this column.
- MHCLLULTS: Cellulites. We excluded from Bgee skin samples coming from subjects that were tagged as "positive" (=1) in this column.
- MHCOCAINE5: Cocaine use/abuse in the 5 years before death. We excluded from Bgee all samples from subjects that were tagged as "positive" (=1) in this column.
- MHCOPD: Chronic low-respiratory disease. We excluded from Bgee lung samples coming from subjects that were tagged as "positive" (=1) in this column.
- MHCOUGHU: Chronic respiratory disease. We excluded from Bgee lung samples coming from subjects that were tagged as "positive" (=1) in this column.
- MHCVD: Cerebrovascular disease/stroke/encephalitis. We excluded from Bgee brain samples coming from subjects that were tagged as "positive" (=1) in this column.

- MHDLYSIS: Dialysis. We excluded from Bgee all samples from subjects that were tagged as "positive" (=1) in this column.
- MHDMNTIA: Dementia. We excluded from Bgee brain samples coming from subjects that were tagged as "positive" (=1) in this column.
- MHENCEPHA: Active encephalitis. We excluded from Bgee brain samples coming from subjects that were tagged as "positive" (=1) in this column.
- MHFNGINF: Fungal infections. We excluded from Bgee all samples from subjects that were tagged as "positive" (=1) in this column.
- MHGNRR12M: Gonhorrea. We excluded from Bgee all samples from subjects that were tagged as "positive" (=1) in this column.
- MHHEPBCT: Hepatitis B infection. We excluded from Bgee all samples from subjects that were tagged as "positive" (=1) in this column.
- MHHEPCCT: We excluded from Bgee all samples from subjects that were tagged as "positive" (=1) in this column.
- MHHEROIN: Heroin users. We excluded from Bgee all samples from subjects that were tagged as "positive" (=1) in this column.
- MHHMPHLIA / MHHMPHLIAB: Hemophilia. We excluded from Bgee blood sampless from subjects that were tagged as "positive" (=1) in this column.
- MHHRTDIS: Ischemic heart disease. We excluded from Bgee heart samples coming from subjects that were tagged as "positive" (=1) in this column.
- MHHRTDISB: Heart disease history. We excluded from Bgee heart samples coming from subjects that were tagged as "positive" (=1) in this column.
- MHIVDRG5: Intravenous drug abuse. We excluded from Bgee all samples from subjects that were tagged as "positive" (=1) in this column.
- MHLUPUS: Lupus. We excluded from Bgee all samples from subjects that were tagged as "positive" (=1) in this column.
- MHLVRDIS: Liver chronic disease. We excluded from Bgee liver samples coming from subjects that were tagged as "positive" (=1) in this column.
- MHMENINA: Meningitis. We excluded from Bgee brain samples coming from subjects that were tagged as "positive" (=1) in this column.
- MHOPPINF: Opportunistic infections. We excluded from Bgee all samples from subjects that were tagged as "positive" (=1) in this column.
- MHOSTMYLTS: Osteomyelitis. We excluded from Bgee bone samples coming from subjects that were tagged as "positive" (=1) in this column.
- MHPLLABS: Prescription pills abuse. We excluded from Bgee all samples from subjects that were tagged as "positive" (=1) in this column.
- MHPNMIAB / MHPNMNIA: Penumonia. We excluded from Bgee lung samples coming from subjects that were tagged as "positive" (=1) in this column.
- MHPRKNSN: Parkinson. We excluded from Bgee all samples from subjects that were tagged as "positive" (=1) in this column.
- MHPSBLDCLT: Blood culture shown positive to bacteria. We excluded from Bgee all samples from subjects that were tagged as "positive" (=1) in this column.
- MHRA: Arthritis rheumatoides. We excluded from Bgee all samples from subjects that were tagged as "positive" (=1) in this column.
- MHRNLFLR: Renal failure. We excluded from Bgee kindney samples coming from subjects that were tagged as "positive" (=1) in this column.

- MHSCHZ: Schizophrenia. We excluded from Bgee brain samples coming from subjects that were tagged as "positive" (=1) in this column.
- MHSCLRDRM: Scleroderma, aka systemic sclerosis. We excluded from Bgee all samples from subjects that were tagged as "positive" (=1) in this column.
- MHSDRGABS: Signs of drug abuse. We excluded from Bgee all samples from subjects that were tagged as "positive" (=1) in this column.
- MHSEPSIS: Sepsis. We excluded from Bgee all samples from subjects that were tagged as "positive" (=1) in this column.
- MHSTD: Sex-transmitted diseases. We excluded from Bgee all samples from subjects that were tagged as "positive" (=1) in this column.
- MHSTRDLT: Long-term steroid use. We excluded from Bgee all samples from subjects that were tagged as "positive" (=1) in this column.
- MHSUBABSA / MHSUBABSB: Drug use for non-medical reasons. We excluded from Bgee all samples from subjects that were tagged as "positive" (=1) in this column.
- MHT1D: Diabete type I "genetic". We excluded from Bgee all samples from subjects that were tagged as "positive" (=1) in this column.
- MHTBHX: Tubercolosis. We excluded from Bgee all samples from subjects that were tagged as "positive" (=1) in this column.

In addition, we decided to:
- exclude brain samples from subjects who died from CerebroVascular Accident (CVA) and whose death was reported on Hardy Scale 3 or 4

According to information reported in the described columns: if samples from a given subject had to be completely excluded from Bgee, then the subject was assigned status "NO"; if some but not all the samples from a given subject had to be excluded from Bgee, then the subject was assigned status "PENDING"; if samples from a subject had no reason to be excluded from Bgee, then the subject was assigned status "YES".

How did we retrieve status/organ information for subjects from GTEx whose samples could be included in Bgee?

For the final annotation file, we needed to retrieve information such as organ source or the subject's attributes such as age, sex, ethnicity...etc.

Useful columns reported in file "phs000424.v6.pht002742.v6.p1.c1.GTEx_Subject_Phenotypes.GRU.txt" were the following:
- SUBJID : Subject ID, GTEx Public Donor ID
- GENDER : The Donor's Identification of gender based upon self-report, family/next of kin, or medical record abstraction. Note: values from this variable were used to obtain the "Sex" value for the final RNAseqLibrary.tsv file.
- AGE : Elapsed time since birth in years. Note: values from this variable were used to obtain the "InfoStage" value for the final RNAseqLibrary.tsv file.
- RACE : Report of Donor's race (geographically based) as reported by Donor, family/next of kin, or medical record abstraction. Note: values from this variable were used to obtain the "Strain" value for the final RNAseqLibrary.tsv file.

Useful columns reported in file "phs000424.v6.pht002743.v6.p1.c1.GTEx_Sample_Attributes.GRU.txt" were the following:
- SAMPID: Sample ID, GTEx Public Sample ID
- SMSMPSTE: Tissue Location
- SMTS: Tissue Type, area from which the tissue sample was taken.
- SMTSD: Tissue Type, more specific detail of tissue type

note: values from these three variables were merged together to obtain the "InfoOrgan" value for the final RNAseqLibrary.tsv file.

How did we check quality for each sample that could be included in Bgee?

Only good quality samples have to be included in Bgee. Quality level for each sample was assessed for both biological quality, inferred from microscopy sample subset observations by MDs, and annotation mapping quality.

We defined a "Tissue quality" variable, harbouring two levels: "GOOD" or "BAD", to define biological quality of each sample. We defined a "Mapping quality" variable, harbouring two levels: "GOOD", "BAD", to define annotation mapping quality of each sample.
We assigned one of the levels of both "Tissue quality" and "Mapping quality" variables to each sample to be included in Bgee, depending on the sample's features. We looked up the sample features in the provided file "phs000424.v6.pht002743.v6.p1.c1.GTEx_Sample_Attributes.GRU.txt".

We selected only RNA-seq data from GTEx, because they’re the ones that the current Bgee pipeline is able to deal with. To select only samples processed with RNA-seq, we selected:
- SMGEBTCHT (Type of genotype or expression batch) = ”TrueSeq.v1”

In particular, for "Tissue quality", we were interested in columns:
- SMSTYP: Indicates whether sample is a tumor or a normal. We excluded from Bgee all tumoral samples.
- SMPTHNTS: Pathologist notes, free-text field. We considered as "BAD" samples that contained more than 10% of foreign tissue, as well as samples displaying diseases or early disease stage.

For "Mapping quality", we were interested in columns:
- SMSMPSTE: The location on the organ or tissue from which the specimen was collected.
- SMTS: The type of Tissue that was collected
- SMTSC: Any comments the prosector wishes to make concerning this case with respect to tissue collection (this field contains for example information about the organ sub-anatomy, eg. left/rigt)
- SMTSD: The type of Tissue that was collected on a Donor by Donor basis
- SMUBRID/SMUBRTRM: Uberon term and Uberon ID proposed by annotators.

We considered that a sample had "GOOD" mapping quality if the columns SMSMPSTE/SMTS/SMTSD were consistent with the suggested SMUBRTRM value and if the eventual detail reported in SMTSC was of no biological relevance. Otherwise, we considered that a sample had "BAD" mapping quality if the annotation was considered not precise enough, or if SMUBRTRM was inconsitent with what reported in SMSMPSTE/SMTS/SMTSD, or if there was a biological difference, with possible precision found in Uberon, that was not accurately reported. For all samples with "BAD" mapping quality, we proposed a new Uberon term (and identifier) only if the sample was reported to be of "GOOD" quality.

Criteria of inclusion regarding healthy wild-type data

Factor	Include into Bgee?	Comment	Example
BMI (Body Mass Index) from 18.5 to 25 ~~(or less than 30)~~ (or less than 35)	Yes	In the normal range or overweight (overweight is common), see https://en.wikipedia.org/wiki/Body_mass_index to exclude other ranges (Sept, 2019, Update: see comment here above about GTEx samples, we have to consider BMI value is continuously rising inside human population). Actually, weight difference between individuals may be part of the natural variability	SRP046752
Cell lines (3T3-L1, Hela, MCF-7) and cell cultures	No		
Fasted animals	Yes	If the fastening time is reasonable: a mouse fasting duration of 5–6 h might offer a better comparison to humans overnight(16–18 h), see PMID:24025567	E-GEOD-7137
Dark/light circadian rhythms and temperature variation	yes/no	Depends on the animal, if reasonable for its physiology, yes as in the example	GSE23528
Low or high fat diet for short time	Yes	If the diet time is short (3 days), can be considered as part of the wild life variability for animals	E-GEOD-8524
Mammary glands from virgin, pregnant and lactating females	Yes	From all types of females	E-TABM-199
Oocytes at different stages of maturation	Yes	Without including the info on the stage (except for Drosophila where the different maturation stages are present in the ontology)	E-GEOD-3351

Placenta and extraembryonic components during development	Yes	Be careful whether to put it in the adult or the embryo!	E-GEOD-7674
Injury	No		
Animals selected for their behaviour (e.g. fear)	Yes	Part of the natural variability	E-GEOD-4035
Animals from different strains (e.g. C57BL/6, BALB/c...)	Yes	Part of the natural variability	
Intestinal germ free animals	No	Normal animals have an intestinal flora	E-GEOD-5156
Removal of the eye (monocular enucleation) or cochlea on one side; the eye, visual cortex or cochlea on the other side were analysed	No		E-GEOD-4265
Cell types (T-cells, stem cells...)	Yes	Should be included only if enough precision in the ontologies (e.g. T-cell in zebrafish). If not enough precision in the ontology, store the experiment ID in the "not_included_for_now" file	
Polysomal RNA only hybridized	No	Method that pellets the polyribosomes while leaving the mono and non-polysomal mRNA fractions in the supernatant.	E-GEOD-3962
Anesthesia	No/Yes	Similar to drug treatment (but sometimes anesthesia is simply not described, and anyway has to be accepted for human tissue samples)	
Human post-mortem tissues	Yes	Mainly the simple way to get human tissues	
Light impulse to stress the animals	Yes	Stress is probably usual for lab animals	
Killed by cervical dislocation or decapitation	Yes	Common for Mouse	
Killed by inhalants (CO2)	Yes	Common for Mouse	

Killed by exsanguination under CO2 anesthesia	No	Used for mouse lung retrieval	
Killed by intravenous anesthetic	No	Method of killing is sometimes simply not described	
Killed by intracardiac or intraperitoneal injection	No	Method of killing is sometimes simply not described	
Mock-treated (plasmid, surgery...)	No/Yes	Typical 'control' that is not normal condition, to check. Can be acceptable when this is for example 'mock inoculated in a similar manner with Minimal Essential Medium' (SRP061418), but rejected when involving experimental surgery (SRP039511)	SRP061418, SRP039511
Normal adjacent tissues from tumor	No/Yes	Sometimes difficult to get the information, be careful if both tumor and normal samples come from the same patient (paired samples). Also sometimes we consider the data because no other data are available for these tissues (described in comments)	
Treatment with Alkaline Hypochlorite Solution ("Bleaching") of C.elegans worm: larva are collected	Yes	Use for synchronizing C. elegans culture, see here	