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

Frederic B. Bastian[1,2,*], Julien Roux[1,2], 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], Julien Wollbrett[1,2], Amina Echchiki[1,2], Angélique Escoriza[1,2], Walid H Gharib[1,2], Mar Gonzales-Porta[1,2], Yohan Jarosz[1,2], Balazs Laurenczy[1,2], Philippe Moret[1,2], Emilie Person[1,2], Patrick Roelli[1,2], Komal Sanjeev[1,2], Mathieu Seppey[1,2], 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

* To whom correspondence should be addressed. Tel: +41 692 4220; Fax: +41 21 692 4165; Email: bgee@sib.swiss

Present Address:

Julien Roux: Bioinformatics Core Facility, Department of Biomedicine, University of Basel, Switzerland

Marta Rosikiewicz, Walid H Gharib, Balazs Laurenczy: SOPHiA GENETICS, Switzerland

Angélique Escoriza: Medion Grifols AG, Switzerland

Mar Gonzales-Porta: Genome Institute of Singapore, Singapore

Yohan Jarosz: Luxembourg Centre for System Biomedecine, Luxembourg

Emilie Person: Office fédéral de l'environnement OFEV, Switzerland

Patrick Roelli: 10x Genomics, Sweden

Mathieu Seppey: University of Geneva, Switzerland

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 functions, pathways, pathologies, 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(1–3), this is not the case for expression data. Both the concept of a "reference expression" for a species, and comparisons, are complicated by the need to define comparable samples and conditions (4).

There are a few datasets which can play a role close to that of a reference expression pattern. Such “atlases” cover typically many anatomical structures (tissues or organs), and are often used to present expression in gene-centric databases, such as NCBI Genes(5). Genes presents by default expression from the HPA set(6) for human genes, with the option of switching to another of three RNA-seq atlases, covering 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 (NCBI Gene ID: 3630) shows very pancreas-specific expression from HPA, but the other proposed atlases lack pancreas samples, and show top expression in the prostate, the spleen, or the stomach. Probably the most complete atlas so far is GTEx (7). While atlases are valuable, they are dispersed in different databases, do not have common standards of annotation or quality-control, provide differently processed values, and cover different limited subsets of anatomy and conditions.

There exist databases which specialize in providing expression pattern descriptions(8). The Expression Atlas(9) presents curated and processed expression data from a diversity of species and sources. As of May 2020, it includes 3,942 studies from 65 species, of which 214 are annotated as “baseline”, and are used to present an expression pattern of each gene. Expression is presented per experiment. Tissues(10) presents expression patterns derived from UniProtKB annotations, selected atlases, and text mining. The focus is on confidence in expression presence rather than levels of expression. Data from different techniques is presented separately. An interesting feature is that tissues are ranked within each evidence type, allowing the most important features of a complex expression pattern to emerge. Thus, for human insulin, the pancreas is one of two tissues from UniProtKB and Pancreatic beta cell is the top tissue from text mining. Finally, Model Organism Databases (MODs)(11)(12) provide integration of many datasets for a given species (or a targeted set of species). They integrate many small scale experiments such as in situ hybridization, thanks to curation and the pioneering use of ontologies. They do not cover species outside of a small core of

model organisms. A limitation, relative to “reference expression”, is that much of the expression data comes from mutant strains.

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. 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 small-scale datasets.

MATERIALS AND METHODS

Pipeline overview

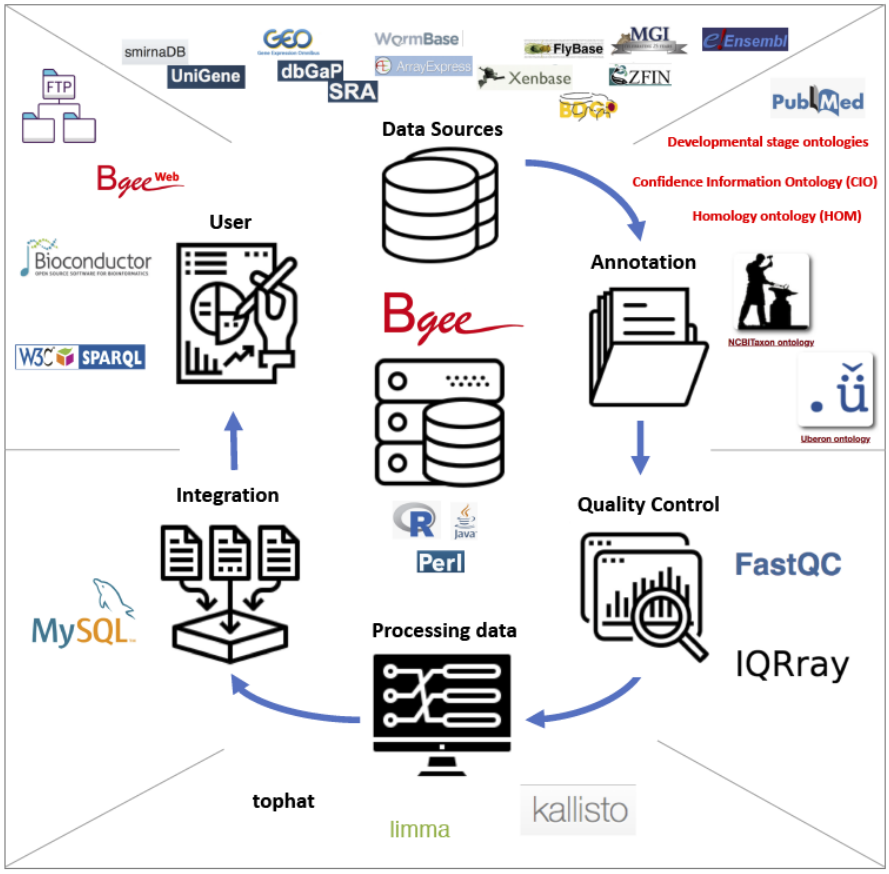

Figure 1: Bgee pipeline overview. Expression data are retrieved from various databases; they are annotated by the Bgee team, or annotations from Model Organism Databases are remapped by the Bgee team, to ontologies describing developmental stages, anatomy, taxa; quality controls are performed, using 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.

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. We provide an overview of this pipeline in Figure 1, and detail some aspects in the following subsections; see the Supplementary Methods and the online documentation for further details.

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(13), and integrate the taxonomy for species present in Bgee. We retrieve gene models for each species in Bgee from Ensembl(3) (Ensembl 84 for Bgee 14) and Ensembl Metazoa(2) (release 30 for Bgee 14) using the Perl Ensembl API. We also retrieve cross-references to other databases, gene biotypes, and annotations to Gene Ontology(14)(15) terms, and gene orthology information from OMA(16), 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(17). Its integration into Bgee requires processing, described in Supplementary Methods. Briefly, we refine taxon constraints; we simplify Uberon to keep only the knowledge needed for Bgee, and simplify it for an easier browsing; we remove any cycles; and we correct mappings between model organism databases (MODs) and Uberon.

Developmental stage integration

Each species in Bgee has a developmental stage ontology, available in the obophenotype developmental stage ontology repository. We have developed all of the species-specific ontologies in this repository, except those from Model Organism Databases (EMAPA for mouse; FBdv for fly(18); ZFA for zebrafish; XAO for Xenopus; WBls for worm). We 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.

We describe our strategy for insertion into Bgee in detail in the Supplementary Methods. Briefly, we merge terms from species-specific ontologies into the structure of Uberon; we order stages using the relations “preceded_by” or “immediately_preceded_by”; and we ensure that each stage has at most one parent by a *part_of* relation. Finally, we transform the merged developmental stage ontology into a nested set model for integration into Bgee.

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 S1. 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(19) 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 Methods for our criteria for inclusion). 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 *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

During our quality control procedure, we have identified duplicated content in the GEO and ArrayExpress databases, affecting about 14% of our Affymetrix data(20). 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(7). 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.

In addition to Uberon and developmental stage ontologies described above, sex is annotated to a simple controlled vocabulary, and 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

We present key processing steps and software very briefly here. Details of expression data download and processing are in Supplementary Methods, and in the online code documentation. For RNA-seq, we use Ensembl and Ensembl metazoa for sequences and annotations of coding and non coding genes. We use Kallisto(21) to generate pseudo-counts per transcript, which are then summed for each gene. All further analysis and reporting is done at the gene level as of Bgee 14. For Affymetrix microarray data, we use CEL files when they are available, using IQRray(22)for quality control and gcRMA(23) for normalization. Otherwise we use MAS5(24) processed files. For ESTs, we simply mapped by BLAST(25) to UniGene clusters, counted the number of ESTs per gene.

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

We retrieve *in situ* hybridization data from the relevant MODs (Table S1), using InterMine(26) whenever possible. We filter out the data from animals with abnormal genetic backgrounds, e.g., animals with transgenes or knockout. We remap species-specific terms from the MOD ontology or controlled vocabulary to Uberon. When needed, we transform annotations to separate anatomical localization and developmental stage. We provide details of this integration, and of specific issues and their solutions, in Supplementary Methods.

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, Sara S. Fonseca Costa, 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(27). Although MAS5 classification is efficient(28), the estimation of the background signal can be biased depending on probe sequence affinity(29). 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(23) to normalize the signal taking into account probe sequences, and use a subset of weakly expressed probesets for estimating the background, as described in (29). 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’(28): High quality present call corresponds to a p-value threshold of < 0.03 (corresponding to a FDR of ≈5%, see Schuster *et al.*(29)), 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 (30), we call 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. We do not produce calls of absence of expression from EST data because of the low sampling.

Remapping to more generalized conditions

We annotate and retrieve expression data with as much granularity as possible. But such a granularity can hamper call integration and comparison. Thus, 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”. For instance, HsapDv:0000150 “56-year-old human stage” would be automatically remapped to HsapDv:0000092 “human middle aged stage”. The highest granularity possible is still available when retrieving our annotated and processed expression values.

Call propagation

After producing these calls from multiple techniques and experiments, we propagate them along the graph of conditions (see Figure 2 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.

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.

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.

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, as of Bgee 14. This means that Bgee is very stringent when reporting absence of expression.

We follow the principles of the Confidence Information Ontology (CIO)(31), and translate them into three confidence levels. After call propagation and reconciliation of presence/absence calls, 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 anatomical entities (for calls of absence of expression).

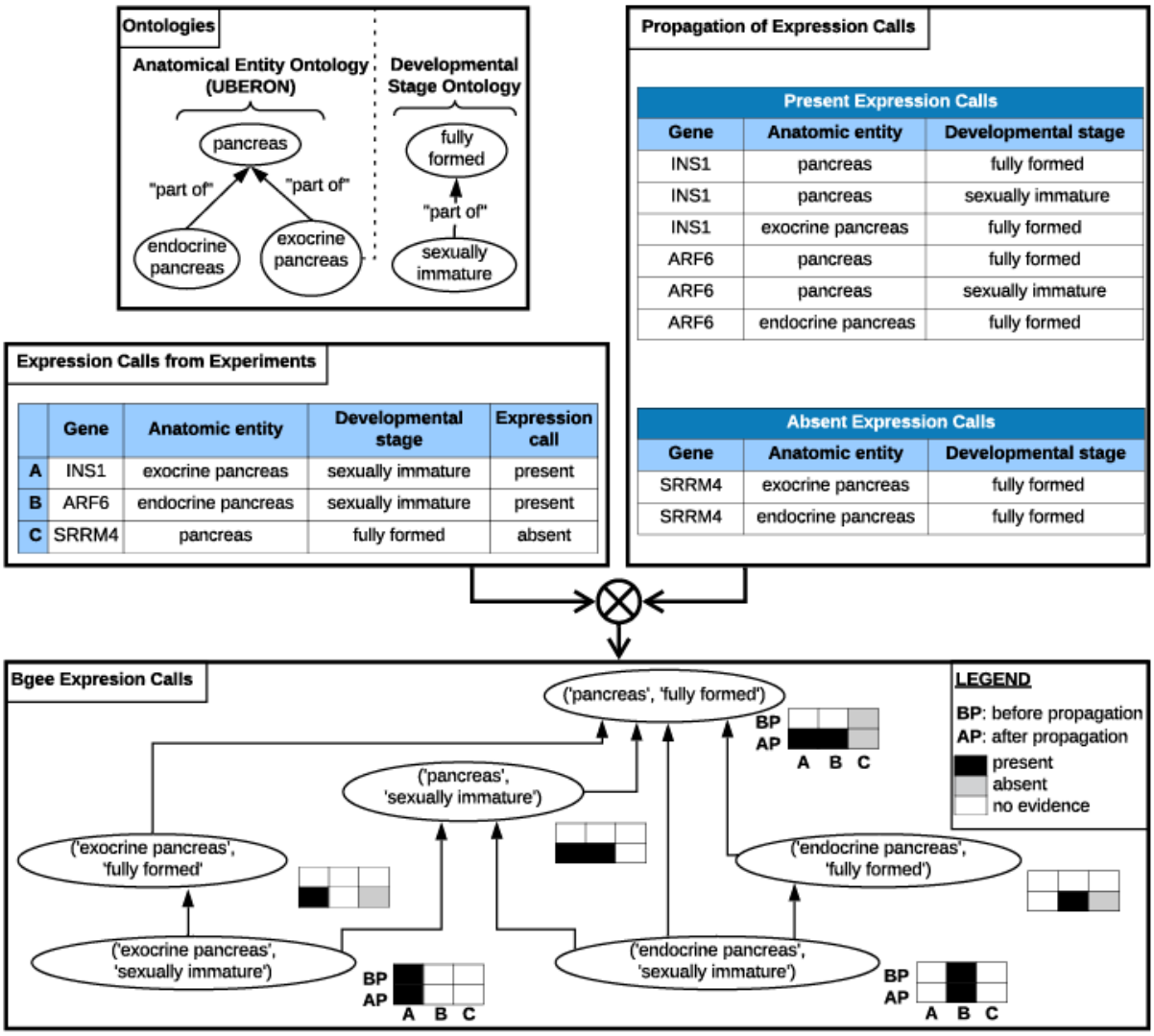

Figure 2: propagation of calls of presence/absence of expression. Calls of presence/absence of expression are produced from the raw data (left table), for instance: call of presence of expression for gene INS1 in exocrine pancreas at sexually immature developmental stage; call of presence of expression 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).

Expression ranks and expression scores

We also provide a 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.

We present briefly the principle of our rank computation. Detailed methods are available in Supplementary Methods and online code documentation.

Rank computation for each data type

For RNA-Seq, we compute fractional ranks of genes in each library, based on their TPM value. Then for each gene and each condition with RNA-Seq data in the condition itself, we compute a weighted mean of the gene ranks over relevant libraries. 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.

Similarly, for Affymetrix we compute fractional ranks based on signal intensity, considering the highest signal intensity for each gene. We normalize the ranks between chip types, because they do not have the same probeset design. e then compute the weighted mean of the normalized ranks per gene and condition, weighting by the number of distinct ranks in each chip.

Since *in situ* hybridization data are not quantitative, we make the assumption that the more often an expression information is reported, the more biologically important this expression is likely to be. Thus the rank is computed on the number of calls and their confidence pulling together all data available in each condition.

While EST data are in principle quantitative, their very low coverage leads us to treat them similarly to *in situ* hybridization data, summing EST counts from all libraries in each condition.

Integration over all data types and experiments

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. To correct for this, 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. 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.

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).

RESULTS

We describe here: i) the data available in Bgee that we have produced; ii) the web interfaces to leverage these data for biological insights; iii) the resources to access Bgee data.

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 and developmental stage similarity information. All original data sources used to build Bgee release 14.1 are in Table S1. All data are published under the Creative Commons Zero license (CC0).

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. Anatomical localization is annotated to the multi-species anatomical ontology Uberon (17), and developmental stages to a multi-species ontology (32), which integrates existing ontologies when available (URLs of the reference ontologies and other resources are in Table S3). Sex and strain are annotated with a basic controlled vocabulary, and species with the NCBI taxonomy(13). 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 S2.

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(7). Only 50% of samples were kept, to discard unhealthy or contaminated samples (see Table S3 for link to GTEx criteria), 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).

In situ hybridization data

We have retrieved *in situ* hybridization data from relevant Model Organism Databases (MODs; Table S1). 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. 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 S1) 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

Expression calls

A call corresponds to a unique combination of a gene in a condition, with reported presence or absence of expression. We produce calls for each evidence integrated into Bgee. 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(29); 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(30). These calls can then be integrated over experiments and over data types.

We have propagated these individual calls along a graph of conditions (see *Materials and Methods*). 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(31). 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

To compare expression between experiments in a quantitative manner, we 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: 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). Expression ranks can be retrieved per anatomical entity and developmental stage.

Anatomical and developmental similarity

We use the ontology of homology and related concepts(33) to capture the type of similarity between anatomical entities. These similarity annotations can be retrieved from GitHub(34), 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 (e.g., paleontology, Evo-Devo), 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.

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.

Web interface of Bgee

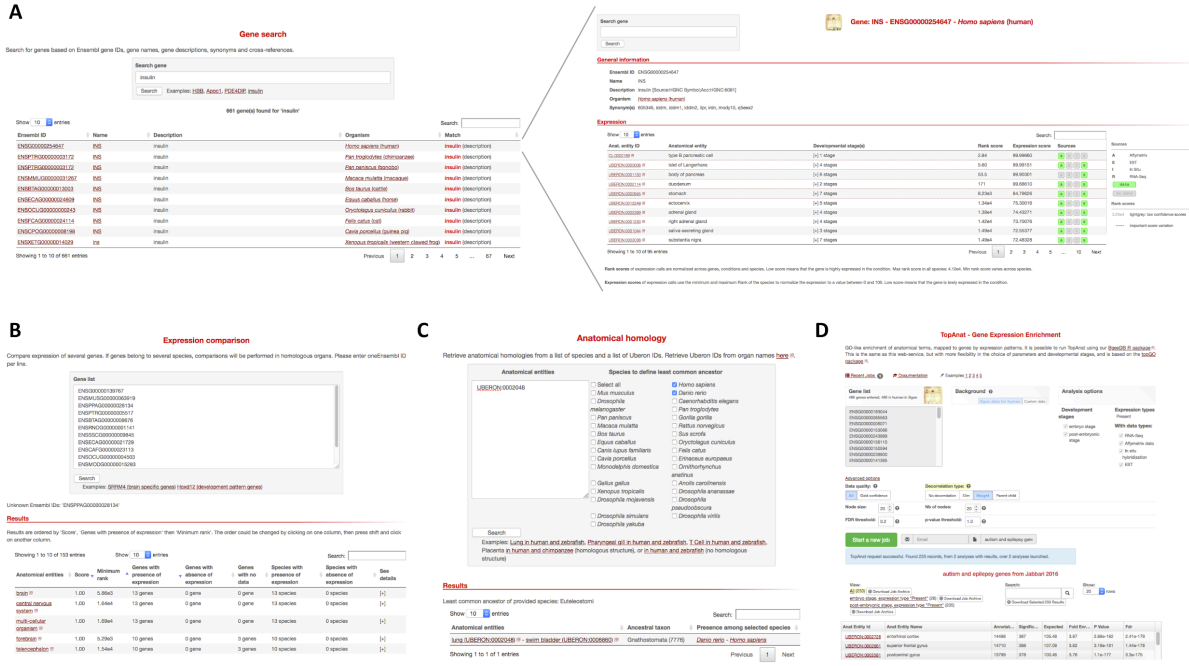

Figure 3: screenshots of the Bgee web interfaces. A: example of gene search (top left) for the term “insulin” (https://bgee.org/?page=gene&query=insulin), allowing to go to the gene page (top right) displaying ranked conditions with expression for the human gene INS (https://bgee.org/?page=gene&gene_id=ENSG00000254647). 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/).

An overview of the Bgee web interfaces is presented in Figure 3. The Bgee Gene page provides for each gene the conditions where it is expressed, sorted by their expression rank. Primary sorting is on anatomical entities.

The anatomical expression enrichment test TopAnat uses a similar approach to Gene Ontology enrichment tests(35)(36), 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.; unpublished). The algorithms from the package topGO(37) are available in TopAnat to account for the non-independence of anatomical structures, and avoid the over-representation of lowly-informative top-level terms.

To compare expression between genes in a list, Expression Comparison (38) 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, by expression rank and by conservation of calls.

Relations of historical homology can also be queried directly through the homology retrieval tool. Thus they can be used for other applications than expression in Bgee, such as phenotypes.

Resources to access data

We provide access to the annotated and processed expression values through the Bioconductor(39) R(40) package BgeeDB(41). BgeeDB is a collection of functions to import into R these data, facilitating downstream analyses. BgeeDB also allows to run TopAnat analyses, offering more flexibility in the choice of input data and analysis parameters than the web-interface.

Bgee also provides a SPARQL(42) 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. 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(43). This web-interface is available at http://biosoda.expasy.org/. Bgee specific queries are present under the category “Bgee database queries”.

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

DISCUSSION

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. Expert manual curation is at the core of providing trustworthy information (44–47). It is key to certify that expression data is from healthy wild-type samples, and to annotate precisely and accurately. It is also the only way to generate reliable annotations of anatomical homology. While expression data analysis is not specific to Bgee(8), our approach is to focus on delivering clear biological signals, and choosing the best methodology to do so. 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 include 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.

Comparative transcriptomics is essential to understand the molecular basis for phenotypes, such as, for instance, evolution of animal morphology(48), embryonic cell differentiation(49), species lifespan(50, 51), or cancer evolution(52). 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.

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 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 rules for 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 anatomical precision beyond that of in situ hybridization 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.

FUNDING

This work was supported 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), and from the European Union's Horizon 2020 research and innovation program under grant agreement No 863410.

AVAILABILITY

Bgee, documentation, and links to resources used. are available at https://bgee.org/. Source code and annotations are available in the GitHub repository (https://github.com/BgeeDB). R packages are available in Bioconductor (https://bioconductor.org/packages/BgeeDB/ and https://bioconductor.org/packages/BgeeCall/).

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.

CONFLICT OF INTEREST

No conflicts of interest to declare.

REFERENCES

1. Haeussler,M., Zweig,A.S., Tyner,C., Speir,M.L., Rosenbloom,K.R., Raney,B.J., Lee,C.M., Lee,B.T., Hinrichs,A.S., Gonzalez,J.N., *et al.* (2018) The UCSC Genome Browser database: 2019 update. *Nucleic Acids Res.*, **47**, D853–D858.
2. Howe,K.L., Contreras-Moreira,B., De Silva,N., Maslen,G., Akanni,W., Allen,J., Alvarez-Jarreta,J., Barba,M., Bolser,D.M., Cambell,L., *et al.* (2019) Ensembl Genomes 2020-enabling non-vertebrate genomic research. *Nucleic Acids Res.*, 10.1093/nar/gkz890.
3. Yates,A.D., Achuthan,P., Akanni,W., Allen,J., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Azov,A.G., Bennett,R., *et al.* (2019) Ensembl 2020. *Nucleic Acids Res.*, 10.1093/nar/gkz966.
4. Roux,J., Rosikiewicz,M. and Robinson-Rechavi,M. (2015) What to compare and how: Comparative transcriptomics for Evo-Devo: COMPARATIVE TRANSCRIPTOMICS FOR Evo-Devo. *J. Exp. Zoolog. B Mol. Dev. Evol.*, **324**, 372–382.
5. Brown,G.R., Hem,V., Katz,K.S., Ovetsky,M., Wallin,C., Ermolaeva,O., Tolstoy,I., Tatusova,T., Pruitt,K.D., Maglott,D.R., *et al.* (2014) Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.*, **43**, D36–D42.
6. Fagerberg,L., Hallström,B.M., Oksvold,P., Kampf,C., Djureinovic,D., Odeberg,J., Habuka,M., Tahmasebpoor,S., Danielsson,A., Edlund,K., *et al.* (2014) Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell. Proteomics MCP*, **13**, 397–406.
7. Carithers,L.J., Ardlie,K., Barcus,M., Branton,P.A., Britton,A., Buia,S.A., Compton,C.C., DeLuca,D.S., Peter-Demchok,J., Gelfand,E.T., *et al.* (2015) A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreservation Biobanking*, **13**, 311–319.
8. Rung,J. and Brazma,A. (2012) Reuse of public genome-wide gene expression data. *Nat. Rev. Genet.*, **14**, 89.
9. Papatheodorou,I., Fonseca,N.A., Keays,M., Tang,Y.A., Barrera,E., Bazant,W., Burke,M., Füllgrabe,A., Fuentes,A.M.-P., George,N., *et al.* (2017) Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Res.*, **46**, D246–D251.
10. Palasca,O., Santos,A., Stolte,C., Gorodkin,J. and Jensen,L.J. (2018) TISSUES 2.0: an integrative web resource on mammalian tissue expression. *Database*, **2018**.
11. Vize,P.D. and Westerfield,M. (2015) Model organism databases. *genesis*, **53**, 449–449.
12. Howe,D.G., Blake,J.A., Bradford,Y.M., Bult,C.J., Calvi,B.R., Engel,S.R., Kadin,J.A., Kaufman,T.C., Kishore,R., Laulederkind,S.J.F., *et al.* (2018) Model organism data evolving in support of translational medicine. *Lab Anim.*, **47**, 277–289.
13. Federhen,S. (2011) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.
14. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T., *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
15. The Gene Ontology Consortium (2019) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.
16. Altenhoff,A.M., Glover,N.M., Train,C.-M., Kaleb,K., Warwick Vesztrocy,A., Dylus,D., de Farias,T.M., Zile,K., Stevenson,C., Long,J., *et al.* (2018) 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.
17. Haendel,M.A., Balhoff,J.P., Bastian,F.B., Blackburn,D.C., Blake,J.A., Bradford,Y., Comte,A., Dahdul,W.M., Dececchi,T.A., Druzinsky,R.E., *et al.* (2014) Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon. *J. Biomed. Semant.*, **5**, 21.
18. Costa,M., Reeve,S., Grumbling,G. and Osumi-Sutherland,D. (2013) The Drosophila anatomy ontology. *J. Biomed. Semant.*, **4**, 32.
19. Wang,J., Zou,H., Chen,L., Long,X., Lan,J., Liu,W., Ma,L., Wang,C., Xu,X., Ren,L., *et al.* (2017) Convergent and divergent genetic changes in the genome of Chinese and European pigs. *Sci. Rep.*, **7**, 8662.
20. Rosikiewicz,M., Comte,A., Niknejad,A., Robinson-Rechavi,M. and Bastian,F.B. (2013) Uncovering hidden duplicated content in public transcriptomics data. *Database*, **2013**.

21. Bray,N.L., Pimentel,H., Melsted,P. and Pachter,L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**, 525.
22. Rosikiewicz,M. and Robinson-Rechavi,M. (2016) 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.
23. Wu,Z., Irizarry,R.A., Gentleman,R., Martinez-Murillo,F. and Spencer,F. (2004) A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *J. Am. Stat. Assoc.*, **99**, 909–917.
24. Hubbell,E., Liu,W.-M. and Mei,R. (2002) Robust estimators for expression analysis. *Bioinformatics*, **18**, 1585–1592.
25. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
26. Kalderimis,A., Lyne,R., Butano,D., Contrino,S., Lyne,M., Heimbach,J., Hu,F., Smith,R., Štěpán,R., Sullivan,J., *et al.* (2014) InterMine: extensive web services for modern biology. *Nucleic Acids Res.*, **42**, W468–W472.
27. Liu,W. -m, Mei,R., Di,X., Ryder,T.B., Hubbell,E., Dee,S., Webster,T.A., Harrington,C.A., Ho,M. -h., Baid,J., *et al.* (2002) Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics*, **18**, 1593–1599.
28. Choe,S.E., Boutros,M., Michelson,A.M., Church,G.M. and Halfon,M.S. (2005) Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biol.*, **6**, R16.
29. Schuster,E.F., Blanc,E., Partridge,L. and Thornton,J.M. (2007) Correcting for sequence biases in present/absent calls. *Genome Biol.*, **8**, R125.
30. Audic,S. and Claverie,J.-M. (1997) The Significance of Digital Gene Expression Profiles. *Genome Res.*, **7**, 986–995.
31. Bastian,F.B., Chibucos,M.C., Gaudet,P., Giglio,M., Holliday,G.L., Huang,H., Lewis,S.E., Niknejad,A., Orchard,S., Poux,S., *et al.* (2015) The Confidence Information Ontology: a step towards a standard for asserting confidence in annotations. *Database*, **2015**.
32. Developmental stage ontology collection.
33. Roux,J. and Robinson-Rechavi,M. (2010) An ontology to clarify homology-related concepts. *Trends Genet.*, **26**, 99–102.
34. Anatomical similarity annotations.
35. Yon Rhee,S., Wood,V., Dolinski,K. and Draghici,S. (2008) Use and misuse of the gene ontology annotations. *Nat. Rev. Genet.*, **9**, 509–515.
36. Dessimoz,C. and Škunca,N. (2017) The Gene Ontology Handbook Humana Press New York, NY, USA:
37. Alexa,A., Rahnenfuhrer,J. and Lengauer,T. (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, **22**, 1600–1607.
38. Robinson-Rechavi,M., Laval,V.R. de, Bastian,F.B., Wollbrett,J. and Bgee,T. (2020) The Expression Comparison Tool in Bgee. In *Phylogenetics in the Genomic Era*. No commercial publisher | Authors open access book, p. 4.3:1--4.3:4.
39. Huber,W., Carey,V.J., Gentleman,R., Anders,S., Carlson,M., Carvalho,B.S., Bravo,H.C., Davis,S., Gatto,L., Girke,T., *et al.* (2015) Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods*, **12**, 115–121.
40. R Core Team (2018) R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.
41. Komljenovic,A., Roux,J., Wollbrett,J., Robinson-Rechavi,M. and Bastian,F.B. (2018) BgeeDB, an R package for retrieval of curated expression datasets and for gene list expression localization enrichment tests. *F1000Research*, **5**, 2748.
42. Segaran,T., Taylor,J. and Evans,C. (2009) Programming the Semantic Web 1st ed. O'Reilly, Beijing ; Sebastopol, CA.
43. Sima,A.C., de Farias,T.M., Zbinden,E., Anisimova,M., Gil,M., Stockinger,H., Stockinger,K., Robinson-Rechavi,M. and Dessimoz,C. (2019) Enabling Semantic Queries Across Federated Bioinformatics Databases Bioinformatics.
44. Howe,D., Costanzo,M., Fey,P., Gojobori,T., Hannick,L., Hide,W., Hill,D.P., Kania,R., Schaeffer,M., St Pierre,S., *et al.* (2008) The future of biocuration. *Nature*, **455**, 47–50.
45. International Society for Biocuration (2018) Biocuration: Distilling data into knowledge. *PLOS Biol.*, **16**, e2002846.
46. Tang,Y.A., Pichler,K., Füllgrabe,A., Lomax,J., Malone,J., Munoz-Torres,M.C., Vasant,D.V., Williams,E. and Haendel,M. (2019) Ten quick tips for biocuration. *PLOS Comput. Biol.*, **15**, e1006906.
47. The SIB Swiss Institute of Bioinformatics' resources: focus on curated databases (2016)

Nucleic Acids Res., **44**, D27–D37.
48. Ahi,E.P., Richter,F., Lecaudey,L.A. and Sefc,K.M. (2019) Gene expression profiling suggests differences in molecular mechanisms of fin elongation between cichlid species. *Sci. Rep.*, **9**, 9052.
49. Briggs,J.A., Weinreb,C., Wagner,D.E., Megason,S., Peshkin,L., Kirschner,M.W. and Klein,A.M. (2018) The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science*, **360**, eaar5780.
50. Fushan,A.A., Turanov,A.A., Lee,S.-G., Kim,E.B., Lobanov,A.V., Yim,S.H., Buffenstein,R., Lee,S.-R., Chang,K.-T., Rhee,H., *et al.* (2015) Gene expression defines natural changes in mammalian lifespan. *Aging Cell*, **14**, 352–365.
51. Holland,P.W., Holland,L.Z., Williams,N.A. and Holland,N.D. (1992) An amphioxus homeobox gene: sequence conservation, spatial expression during development and insights into vertebrate evolution. *Dev. Camb. Engl.*, **116**, 653–661.
52. Lam,S.H., Wu,Y.L., Vega,V.B., Miller,L.D., Spitsbergen,J., Tong,Y., Zhan,H., Govindarajan,K.R., Lee,S., Mathavan,S., *et al.* (2006) Conservation of gene expression signatures between zebrafish and human liver tumors and tumor progression. *Nat. Biotechnol.*, **24**, 73–75.