Robust inference of expression state in bulk and single-cell RNA-Seq using curated intergenic regions

Sara S. Fonseca Costa[1,2*], Marta Rosikiewicz[3*], Julien Roux[2,4*], Julien Wollbrett[1,2], Frederic B. Bastian[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
[3] SOPHiA GENETICS, Switzerland.
[4] Bioinformatics Core Facility, Department of Biomedicine, University of Basel, Switzerland
* Equal contribution
+ Corresponding author: marc.robinson-rechavi@unil.ch

Abstract

RNA-Seq is a powerful technique to provide quantitative information on gene expression. While many applications focus on estimated expression levels, it is also important to determine which genes are actively transcribed, and which are not. The problem can be viewed as simply setting a biologically meaningful threshold for calling a gene expressed. We propose to define this threshold per sample relative to the background level for non-expressed genomic features, inferred by the amount of reads mapped to intergenic regions of the genome. To this aim, we first define a stringent set of reference intergenic regions, based on available bulk RNA-Seq libraries for each species. We provide predefined regions selected for different animal species with varying genome annotation quality through the Bgee database. We then call genes expressed if their level of expression is significantly higher than the background noise. This approach can be applied to bulk as well as single-cell RNA-Seq, on a single library as well as on a combination of libraries over one condition. We show that the estimated proportion of expressed genes is biologically meaningful and stable between libraries originating from the same tissue, in both model and non-model organisms.

Introduction

A typical biological sample contains genes which are actively transcribed, above technical and biological noise, and others which are inactive. Thus a qualitative description of the transcriptome as genes being "present" or "absent" is of wide interest in many fields. Indeed, it is at the core of the link between the genome and the phenotype. Such gene expression calls are also routinely used in genomics studies as a filter before downstream analyses [1] [2], to allow the annotation of genes to sampling conditions, and to facilitate integration across datasets or even data types [3].

Expression is typically measured by the amount of RNA molecules present in samples. A difficulty in identifying active and inactive genes in a biological sample is that the latter still potentially produce RNA molecules through various processes [4]. Indeed transcription is a noisy phenomenon [5]–[7], with contributions from active gene regulation but also border effects from active genes, stochastic binding of RNA polymerase [8]–[10] or permissive chromatin states allowing leaky transcription [11]–[14]. In addition, different technical factors introduce additional noise in gene expression measurements, e.g., sample contamination, RNA isolation, library amplification, or allelic variants.

RNA-Seq is generally accepted as one of the most accurate technologies to quantify expression as a relative measurement. The number of RNA-seq reads mapped to a gene is often normalized by library size, for example in expression level units such as CPM (counts per million mapped reads). CPM does not normalize for gene length, with the important consequence that within-sample expression levels comparison across genes is not possible, except for 3' or 5' end biased protocols (e.g., CAGE). Additional gene length normalization is performed when using units such as RPKM (reads per kilobase of transcript per million reads mapped), FPKM (fragments per kilobase of transcript per million fragments mapped), or TPM (transcripts per million) [15], allowing the comparison of expression levels between genes. Arbitrary thresholds of these units are often used in transcriptomics studies to call genes active or inactive, yet there is little consensus on the exact value to use. Values as low as 0.1 RPKM [16] or as high as 3 TPM [17] (based on [18]) are used. The Expression Atlas Baseline [19] uses a threshold of 0.5 TPM or 0.5 FPKM to report expression (https://www.ebi.ac.uk/gxa/FAQ.html#blResults). Many RNA-seq studies consider a gene expressed if at least one read is mapped to a gene [1], [20]. Yet given the relative nature of expression levels measured by RNA-Seq, one fixed threshold cannot be expected to fit a variety of transcriptomic samples from different tissues and conditions. Moreover, the sensitivity or specificity of calls at different thresholds has rarely been evaluated.

In an effort to define sample-specific thresholds, Hebenstreit et al. [4] showed that cultured murine Th2 cells displayed two classes of genes, active and inactive, the latter contributing a clear left shoulder on the distribution of log-transformed RPKMs across genes. The genes classes were identified by deconvoluting this distribution into two Gaussians, and the status of genes was validated by RT-PCR and measurements of histone modifications. Wagner and Lynch [18] proposed to fit a model to the raw TPM values, aiming at deconvoluting a mixture of an exponential distribution for inactive genes, and a negative binomial distribution for active genes. Hart et al. [21] proposed an approach based on the distribution of RPKMs in genes to model the most expressed genes as a simple Gaussian, and measure the distance to this distribution with a Z-score (zFPKM [21]).

Recently, Thompson et al [22] proposed a Bayesian mixture model (implemented in the method "zigzag") to infer active expression, formalizing the logic of Hebenstreit et al. [4]. Briefly, their model fits to the log-transformed TPM values a mixture of a Gaussian for inactive expression, one or more Gaussians for active expression (e.g., low and high expression), and a compartment of genes with no detected reads. They then infer posterior probabilities for genes to belong to an inactive or an active component. This represents to our knowledge the most advanced method available to call genes actively transcribed or not from RNA-Seq. However this approach presents some practical limitations, such as the requirement of at least two libraries to perform the inferences, as well as convergence problems when there is high discrepancy in variance between samples (see Results).

All these methods rely on the reads mapped to annotated genes only. We propose that other genomic regions could also be informative, notably to define a background expectation for the class of inactive genes. Intergenic regions are genomic features rarely actively transcribed, yet libraries contain reads mapped to them, because of the technical and biological noise sources described above.

We thus propose a new approach for objective expression calls using the level of spurious intergenic expression – due to other factors than active expression – to estimate per sample the transcriptomic noise, and thus to detect genes actively expressed above it. This allows us to define a sample-specific false positive threshold for calling active expression. We define a set of reference intergenic regions based on the compendium of curated RNA-Seq libraries in our database Bgee [3], allowing us to exclude actively expressed regions such as false negatives of gene annotation pipelines. Our method is used for calls in Bgee, and can be applied to any bulk or single-cell RNA-Seq library through the BgeeCall R package [23].

Results

A clean model organism: application to mouse data in Bgee

Bgee release 15.0 includes 566 RNA-Seq libraries for the house mouse *Mus musculus*, from 36 experiments, covering 164 healthy wild-type conditions (combinations of 71 anatomical structures and of 32 developmental stages – from zygote to >20 months old). The mouse genome is one of the best assembled and annotated among animals (all genomes information versions in Supplementary table 1), and one of the species with the highest amount of RNA-Seq data in Bgee (Supplementary table 2). Thus it presents a best case scenario. Mapping reads from all mouse libraries to intergenic regions, the distribution of log2(TPM) is deconvoluted into three Gaussians (Figure 1A): one very large but with low density, and two narrower which account for most of the TPM density. Considering the low overlap of each intergenic Gaussian with the log2(TPM) distribution from reads of all mouse libraries mapped to protein-coding genes (Figure 1B), we consider the right-most narrow Gaussian as representing the maximal transcription level of intergenic regions with no active expression in any condition sampled across mouse libraries. Thus we define the set of reference intergenic regions as all those with log2(TPM) lower than the $MITT_{mouse}$ for a region attributed to this Gaussian ($MITT_{mouse}$ log2(TPM) = -1.558037). Importantly, the selection of reference intergenic regions is quite consistent if only a subset of samples is used (Supplementary figure 1).

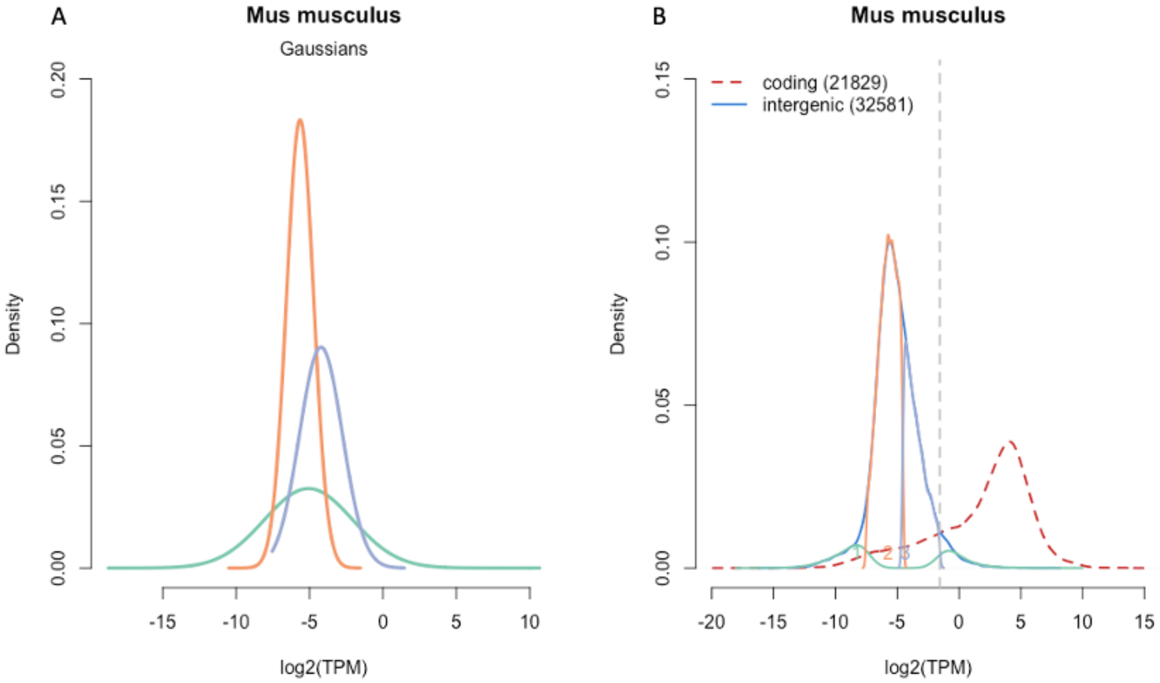

Figure 1: Deconvolution of log2(TPM) density distributions of intergenic regions, for mapped reads from all mouse libraries in Bgee release 15. **(A)** Gaussians from the deconvolution of log2(TPM). **(B)** Density profile of regions assigned to each Gaussian. In blue, all intergenic; in green, orange and light blue the density of regions assigned to each Gaussian. Dash red line all coding regions and dash line in gray $\mathrm{MITT}_{\mathrm{mouse}}$.

Using this set of reference intergenic regions to call actively expressed genes in each individual mouse library, we obtain a median of 68.11% of coding genes expressed, narrowly distributed across libraries, with an interquartile range of 62.97% to 70.51% (with $\alpha = 0.05$; Table 1). The minimum proportion is 37.96% of genes called expressed in a liver sample from Theiler stage 19 (library SRX1603153 [24]), while the maximum is 77.19% in an adult testis sample (library SRX1038931 [25]). The proportion of coding genes called expressed is rather stable using different p-value thresholds (Table 1). Without the deconvolution step used to define reference intergenic regions, slightly less coding genes are called actively expressed. Using a fixed threshold of 2 TPM, we obtain not only less calls of expression, but a much wider range of results depending on libraries. Unsurprisingly, the different intergenic-based methods correlate very well with each other (Pearson r of 0.90 to 0.99), while the fixed TPM method correlates more poorly (r = 0.70). This especially affects the set of libraries with the less expression calls, where there appears to be no correlation (Supplementary figure 2). Finally, expression calls with microarrays have larger variation between samples than any RNA-Seq based method.

Data & method	**Threshold**	**Sample number**	**Median**	**Minimum**	**Maximum**	**Interquartile range**
RNA-Seq deconvolution	$p \leq 0.001$	566	56.98 %	10.31 %	65.51 %	51.23% - 59.90 %
	$p \leq 0.01$	566	63.38%	26.03%	71.12%	57.70% - 65.81%
	$p \leq 0.05$	566	68.11%	37.96%	77.19%	62.97% - 70.51%
RNA-Seq no deconvolution	$p \leq 0.05$	566	66.31%	37.66%	75.48%	61.08% - 68.78%
RNA-Seq fixed threshold	TPM ≥ 2	566	52.95%	19.30%	61.95%	45.74% - 57.80%
Microarray gcRMA	Present or marginal calls	5358	56.59%	11.14%	85.30%	49.50 % - 63.38%
Microarray MAS5	Present or marginal calls	737	56.86%	23.24%	74.91%	45.23% - 62.63 %

Table 1: Distribution of calls of expressed (“present”) coding genes over mouse RNA-Seq libraries and Affymetrix microarrays in Bgee.

We evaluated the performance of these different methods using a mouse liver Ribo-Seq dataset. We use 89 libraries annotated in the Bgee database to the anatomical entity liver (Figure 2). The median true positive rate (TPR) differs only slightly with or without deconvolution; there is more difference when small p-values are used ($p \leq$ 1e-5), and then the deconvolution method performs slightly better. Using fixed TPM thresholds, we have less power (TPR) with the same level of specificity (FPR). The nominal p-values here could be affected by multiple testing, thus we also evaluated using a BH correction over all genes per library; this had very little impact (Supplementary figure 3). We also obtain consistent results with other benchmark data (Supplementary figure 4).

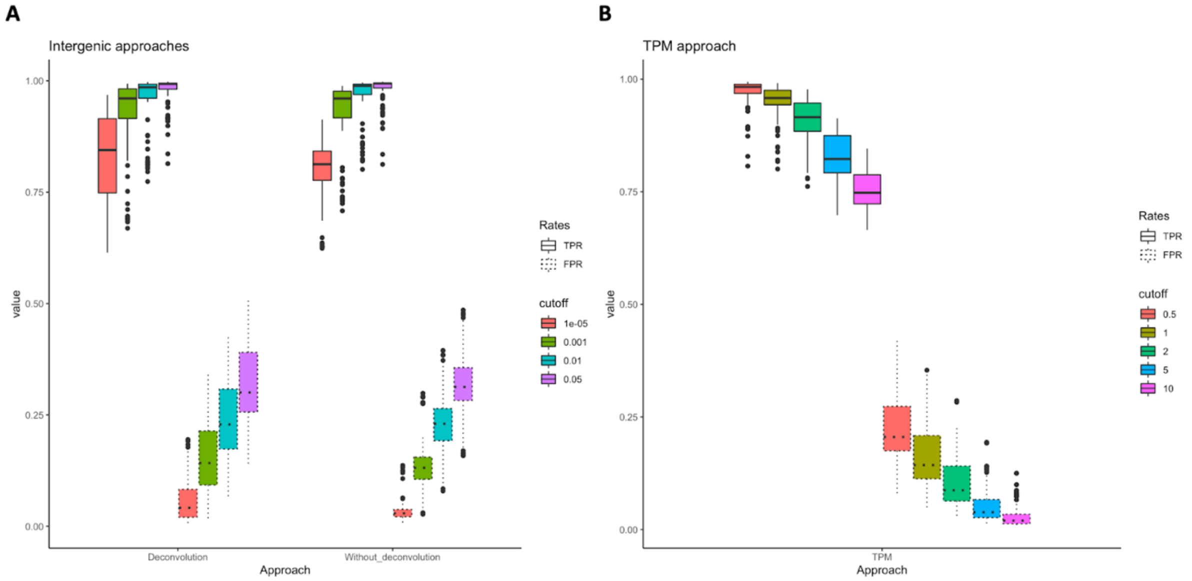

Figure 2: True discovery rate and false discovery rate for calls of gene expression, based on mouse Ribo-Seq. **(A)** Rates calculated per library using reference intergenic regions (deconvolution method) or without deconvolution; **(B)** Rate calculated based on TPM threshold. Colors correspond to different cutoffs (from $p \leq 0.05$ to $p \leq 0.00001$ and from TPM ≥ 0.5 to TPM ≥ 10) applied to each approach (deconvolution, without deconvolution or TPM threshold) and different line type (solid and dotted) correspond to the TPR or FPR.

Northern Pike: a non-model organism

The Northern pike *Esox lucius* is an Esociformes fish, studied as an outgroup to Salmoniformes [26] In contrast to *Mus musculus*, *Esox luciu* has few libraries in Bgee 15.0 (24 libraries from 24 conditions, from 16 anatomical structures and 2 developmental stages). We thus examined expression calls in such a case. Note that (Supplementary figure 5) we report results for all 52 species in Bgee 15.0.

One notable difference between non model and model species genomes is the annotation of non coding genes. If not annotated, these appear as actively expressed "intergenic" regions. In Ensembl release 102, the human genome GRCh38.p13 and mouse genome GRCm38.p6 include 23,982 and 16,060 annotated non coding genes, of which 16,896 and 9,972 are long non coding (lncRNA) genes. In the Northern pike genome Eluc_v4 there are 6,291 non coding genes, of which only 915 are lncRNA genes. It is thus probable that this annotation misses potentially expressed non coding genes.

Accordingly, we observe that the log2(TPM) distribution of intergenic regions is broader, and that it overlaps largely with the log2(TPM) distribution of protein-coding genes. This intergenic distribution is deconvoluted into four Gaussians (Figure 3A-B), and we define the $\text{MITT}_{\text{pike}}$ on the right of Gaussian 2 (log2(TPM) = -2.334883). As in mouse, our method produces consistent proportions of calls between libraries, with a median of 67.62% and an interquartile range of 65.40% - 72.16% of genes expressed at an α of 0.05. Using an α of 0.01 or 0.001 of intergenic among expressed genes reduces the median of expression as expected (Figure 3C). Here, using the whole set of intergenic regions without deconvolution has a huge effect, with the median number of protein coding genes called expressed at $\alpha = 0.05$ falling to 37.70%. The calls with a fixed threshold at 2 TPM are more variable with an interquartile range of 48.60% - 62.43%, and a median of 56.26%. In conclusion, the analysis on a species with less well annotated genome highlights the importance of the deconvolution of intergenic regions, allowing to define a cleaner set of reference intergenic regions to call genes expressed (also see Supplementary figure 6).

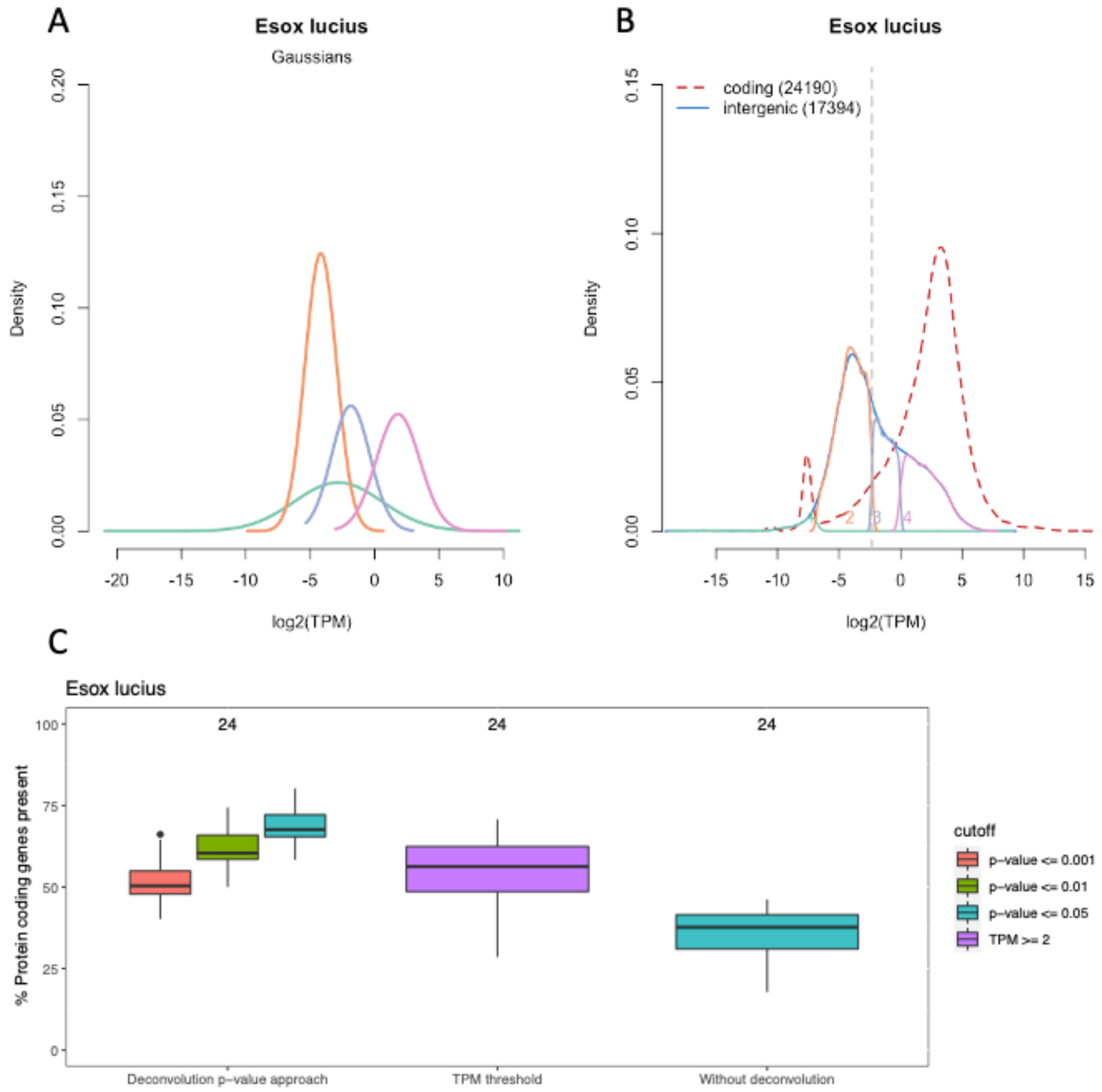

Figure 3: Expression of Northern pike genes. **(A)** Gaussians from the deconvolution of log2(TPM) density distributions for Northern pike libraries in Bgee 15.0. **(B)** density profile of regions assigned to each Gaussian. In blue, all intergenic; in green,orange, light blue and violet the density of regions assigned to each Gaussian. Dash red line all coding regions and dash line in gray $\text{MITT}_{\text{species.}}$ (log2(TPM) = -2.334883). **(C)** Distribution of calls of expressed ("present") coding genes over 24 Northern pike libraries. From left to right, results using thresholds

from reference (deconvoluted) intergenic with a α = 0.001, 0.01, 0.05, using a fixed 2 TPM threshold, and α = 0.05 without deconvolution.

Application to 52 species in Bgee: stable calls of expression

In the release 15.0 of the Bgee database, we used the pValue method with reference intergenic regions to classify genes as actively expressed or not across 15,516 curated RNA-Seq libraries from 52 animal species. As a comparison, we also applied both the method without deconvolution of intergenic regions and the fixed TPM cutoff method to call genes expressed or not in the same libraries (Figure 4). Of these, 4,847 are from human GTEx data, with a median of 57 libraries for the other species.

Consistent with the results on mouse and pike, a fixed threshold on TPM values led to large variation in the proportion of genes called expressed between samples of the same species (Figure 4A). Using all intergenic regions to estimate background expression led to very large differences between model and non model organisms, probably reflecting genome annotation quality (Figure 4B). For *H. sapiens*, *D. melanogaster*, *C. elegans*, and *M. musculus* the proportion of calls of expression was above 65%, with relatively little variation across samples for these species. For *G. aculeatus*, *A. calliptera*, *N. brichardi*, and *B. lanceolatum* a minority of genes are called present.

Our method provides calls of expression that were largely consistent across species, and among samples within a species (Figure 4C). A median proportion of 60% to 83% of genes are called expressed across species using α = 0.05. Only four non-model species stand out, *Astatotilapia calliptera*, *Neolamprologus brichardi*, *Branchiostoma lanceolatum* and *Latimeria chalumnae*. The low median proportion of protein coding genes could be explained by the low numbers of biological replicates over a few organs, where some anatomical entities are represented by just one sample.

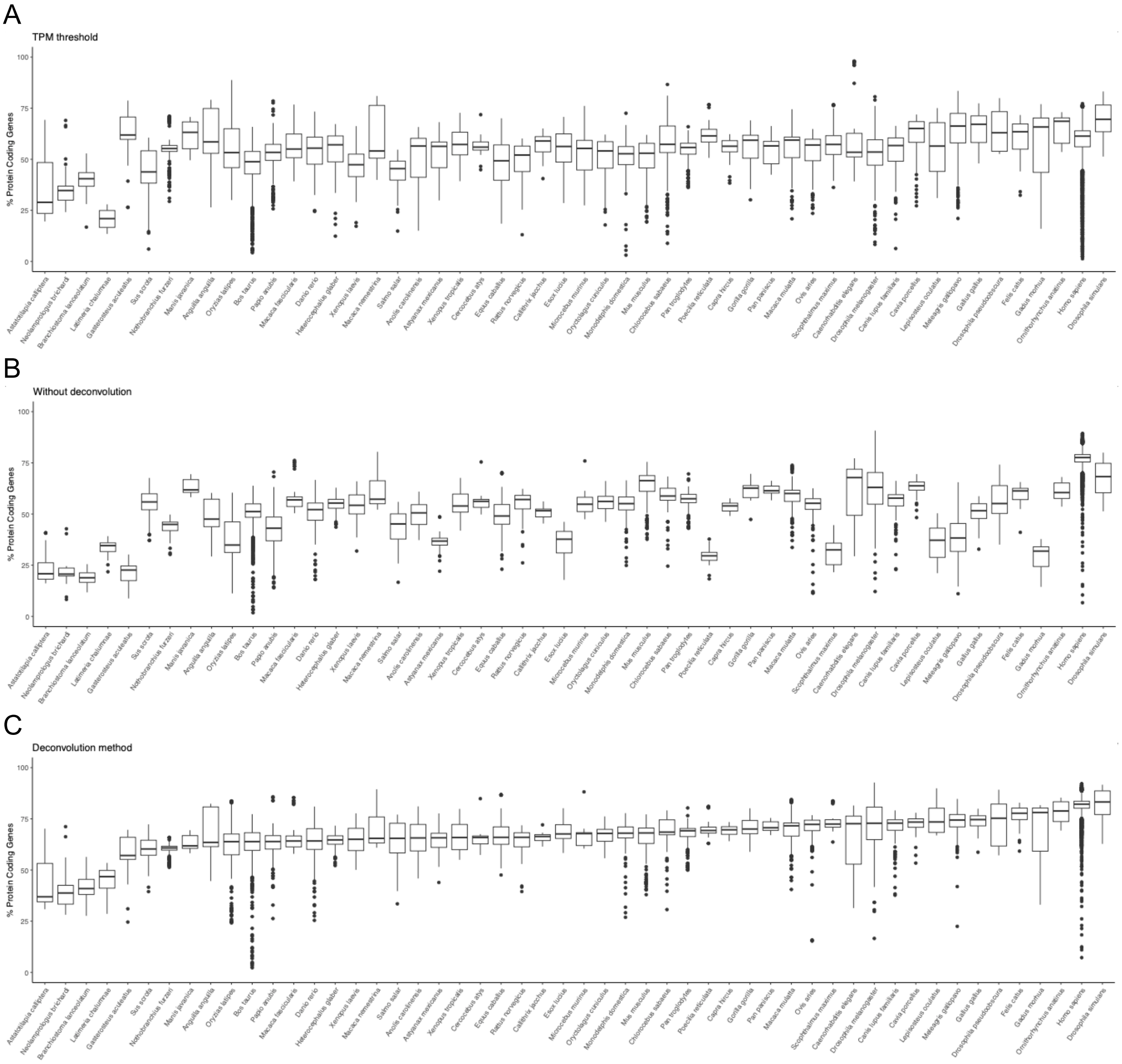

Figure 4: Proportion of genes called present per species for all RNA-Seq libraries in Bgee release 15. The boxplots are ordered by the median proportion of protein coding genes using our reference intergenic method. **(A)** Using a threshold of 2 TPM. **(B)** Using a background of all intergenic regions. **(C)** Using our reference intergenic method.

We used the proportion of non coding genes to check the role of annotation in detecting reference intergenic regions. Over all samples, the median proportion of genes called present using intergenic regions without deconvolution was strongly linked to the proportion of non-coding genes (Table 2). For other methods, the correlation was much weaker, confirming that our deconvolution method corrected the false negative calls issue (i.e., genes actively expressed but called absent because misannotated intergenic regions were detected at similar or higher expression levels).

Expression call method	Pearson r	Spearman ρ
Fixed TPM	0.198	0.111
Intergenic, no deconvolution	0.760	0.723
Intergenic, deconvolution	0.399	0.333

Table 2: Correlation between the median proportion of genes called present and the proportion of non-coding genes in the genome annotation, over all species and samples (15516 libraries). Thresholds used: ≥ 2 TPM or $\alpha = 0.05$.

A large experiment across anatomy: application to GTEx

GTEx [27] is the largest experiment in Bgee 15, with 4847 libraries in 75 human anatomical structures (curated subset of GTEx v6), which we will call "organs" here for simplicity. Bgee only includes libraries which were curated for healthy samples (e.g., no obesity, no disease, etc) [3], thus we expect a good consistency between samples. With a fixed threshold of 2 TPM there is large variation between organs, and a large variation between samples within organs (Figure 5A). On the other hand, our method provides very consistent call proportions between samples and also to some extent between organs (Figure 5B). Importantly, the differences between organs are consistent to expectations, e.g., testis has the highest proportion of genes actively expressed [28]–[30].

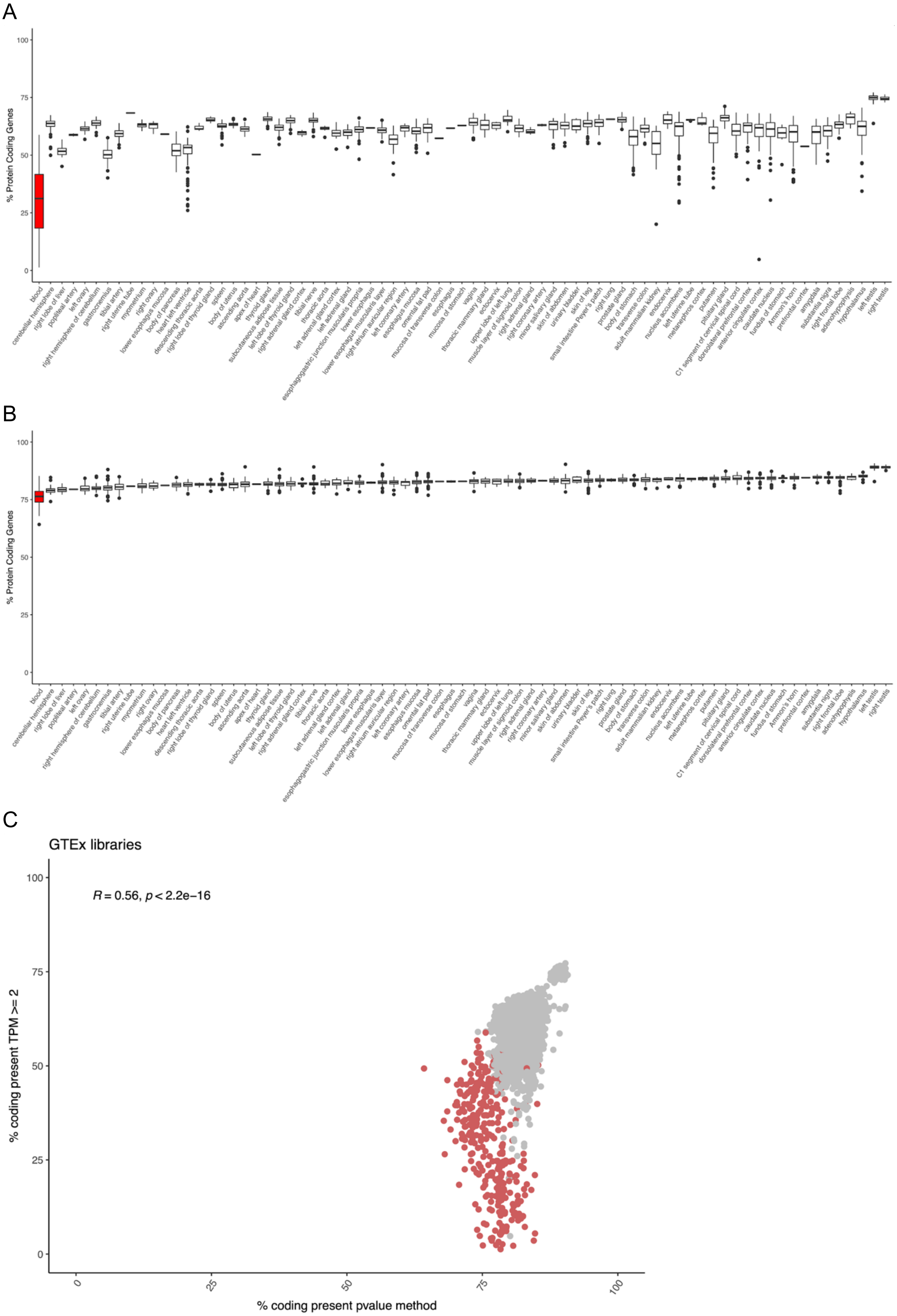

Figure 5: Distribution of calls of expressed ("present") coding genes over GTEx libraries in Bgee, per anatomical structure. In red, samples from blood. **(A)** Distribution on calls based on a fixed 2 TPM threshold. **(B)** Distribution

on calls based the p-value approach ($\alpha = 0.05$, reference intergenic). (**A-B**) The organs are ordered by the median proportion of calls using the p-value approach. **(C)** Relation between present calls by p-value, and by fixed TPM threshold; in the corner, Pearson correlation.

Finally, across all GTEx samples we observe a poor correlation between the proportion of genes called expressed with a fixed threshold compared to our method (Pearson r = 0.56) (Figure 5C). The correlation is even weaker after excluding blood samples (Pearson r = 0.43) (Supplementary figure 7). This shows that the method used to call expression can have a large impact on results.

Application to single cell RNA-Seq

The method that we present can also be applied to single-cell RNA-Seq with minor adaptations. For full length protocols we applied the method to two mouse experiments with a total of 1323 cells (from Smart-seq and Smart-seq2 protocols), and to two human experiments with a total of 158 cells. Compared to bulk data, the median proportion of coding genes called present was much lower, at 28% in human and 34% in mouse (Figure 6). This decrease in the proportion of genes called expressed can be explained by the high proportion of zero read count [32] per transcript in scRNA-Seq data, relative to bulk RNA-Seq (Table 3).

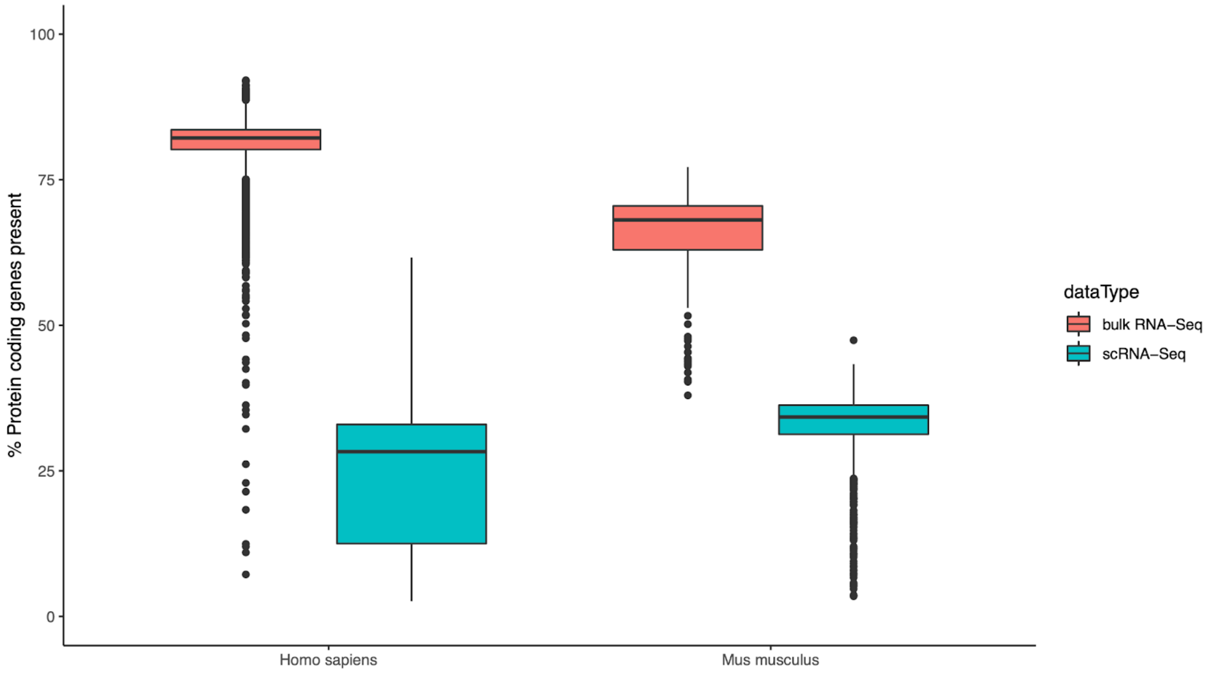

Figure 6: Proportion of protein coding genes expressed on full-length single cell vs. bulk RNA-Seq for human and mouse data. Here "scRNA-Seq" denotes full-length single cell RNA-Seq.

Species	**Data type**	**Library**	**Reference**	**Reads**	**Protein coding genes with 0 reads**
Mus musculus	bulk RNA-Seq	ERX1174321	[33]	232 M	11.1%
Mus musculus	scRNA-Seq	SRX259131	[34]	28 M	58.7%
Homo sapiens	bulk RNA-Seq	SRX618352	[27]	236 M	4.6%
Homo sapiens	scRNA-Seq	SRX2416177	[35]	14 M	68.3%

Table 3: Proportion of genes Ids with zero read counts for the libraries with more reads for RNA-Seq and scRNA-Seq integrated in Bgee 15 for mouse and human data. Here "scRNA-Seq" denotes full-length single cell RNA-Seq.

For droplet-based protocols, we focused on the 10X genomics platform and used 6 human testis libraries, 12 mouse spleen libraries, and 8 mouse blood libraries. These libraries provide expression information for 6,008 cells in human and 67,895 cells in mouse, after quality filtering [36]. Given the low average number of reads per cell relative to full length single-cell and bulk protocols [37], at the individual cell level we have no power, with extremely low CPMs and almost no distinction between genic and intergenic distributions (Figure 7A). We thus performed expression calls at the level of cell population rather than of individual cells. This is similar to bulk RNA-Seq samples resulting from an aggregation of a large population of cells. This allows us to recover satisfactory statistical power, with a clear difference between CPM levels of genic and intergenic regions (Figure 7B). In order to also provide information at the individual cell level, similar to the approach in [38], we propose that once a gene is called expressed in a cell population, it is called expressed in each cell of this population if the gene has at least one UMI count.

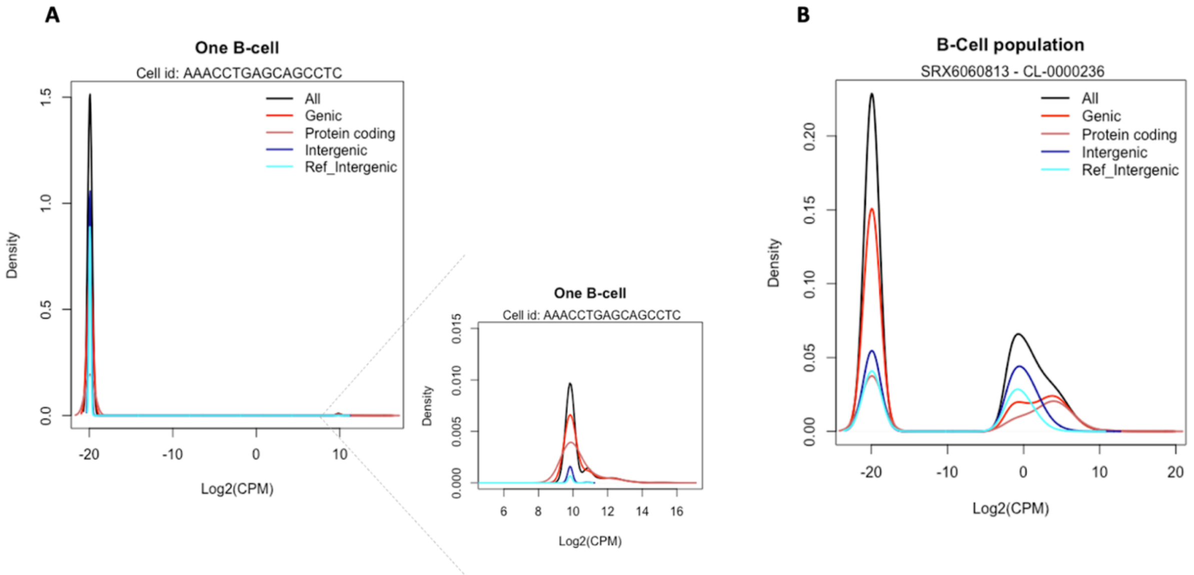

Figure 7: Density distribution of genic and intergenic regions from 10X single-cell RNA-Seq. **(A)** using the CPM of only one randomly picked B-cell; a zoom in panel A is performed in order to show the densities of the high expressed regions. **(B)** Density calculated using the counts per million (CPM) of a population of B-cells (cell ontology id: CL-0000236); Data from library SRX6060813 [39].

Overall, applying the reference intergenic method to these two different types of single-cell protocols yielded the same scale of expressed protein coding genes (Figure 8; Table 4), despite covering different cell types. The droplet-based protocols provided less dispersed results compared with full length protocols. In general, the use of our intergenic-based method provides stable results in calling expressed genes for different single-cell protocols.

Data & protocol	**Species**	**Median**	**Minimum**	**Maximum**	**Interquartile range**
scRNA-Seq full-length	Homo sapiens	28.293 %	2.59 %	61.633 %	12.50% - 32.962 %
scRNA-Seq full-length	Mus musculus	34.231 %	3.469 %	47.415 %	31.27% - 36.253 %
scRNA-Seq droplet-based	Homo sapiens	39.41 %	32.7 %	50.40%	35.79% - 41.33%
scRNA-Seq droplet-based	Mus musculus	30.22 %	23.43 %	34.39 %	27.80% - 32.02%

Table 4: Distribution of calls of expressed (“present”) coding genes over human and mouse single cell RNA-Seq libraries for different protocols.

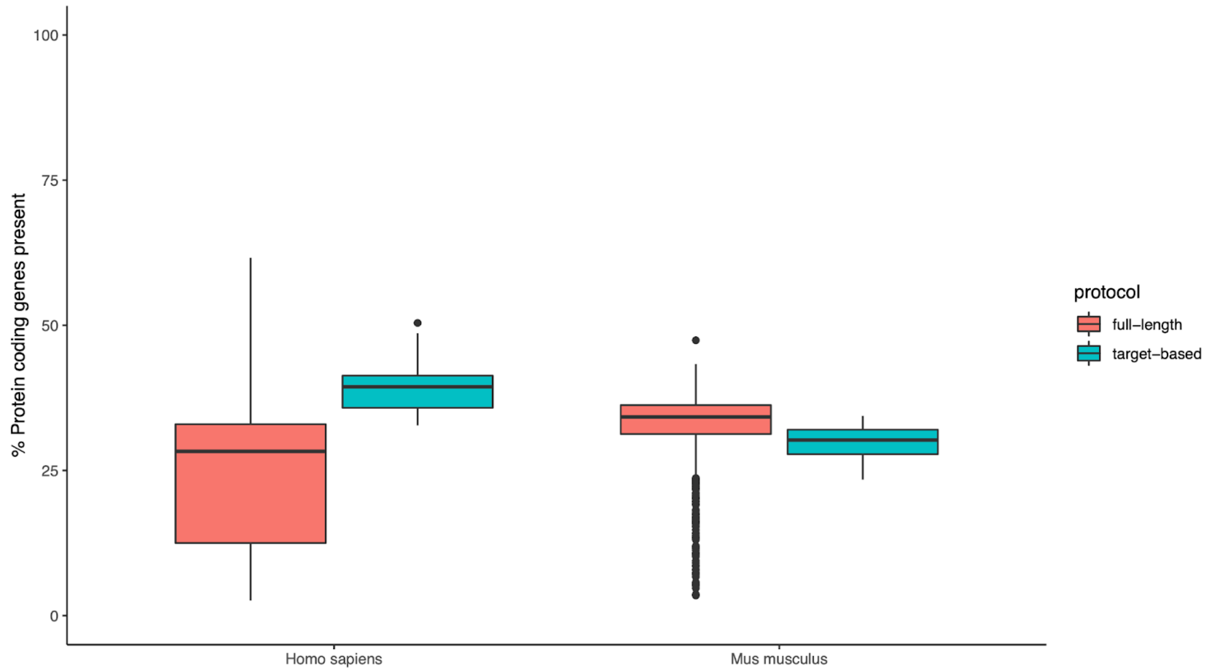

Figure 8: Comparison of proportion of expression calls in single cell RNA-Seq data using different protocols.

Using a simple CPM threshold (CPM ≥ 1) in droplet-based protocols (after summing the UMIs across cells that belongs to the same cell type and are from the same sample) to call expression genes we obtain an increase in the median expression (Supplementary figure 8), to 65.26% for human and 45.72% for mouse.

Performing functional enrichment on genes called expressed in the B-cell population identified (Figure 9) important and specific functional categories, such as: Hypoxia response via HIF activation – the transcription factor HIF has a fundamental role to help immune cells adapting to hypoxic environments [40]; the FAS signaling pathway, which plays an important role in the maintenance of immunological tolerance [41]; the oxidative stress response (ROS), which affects the maturation, activation and differentiation of B cells [42]; the toll receptor signaling pathway, which provides a mechanism for adaptive immune response in B cells, as well as activation and differentiation [43]; the transcription regulation by bZip transcription factors, which are regulators of B cell differentiation [44]. In summary, genes classified present by our method were clearly enriched for genes important for B cell biology, whereas a simple CPM threshold misses important pathways.

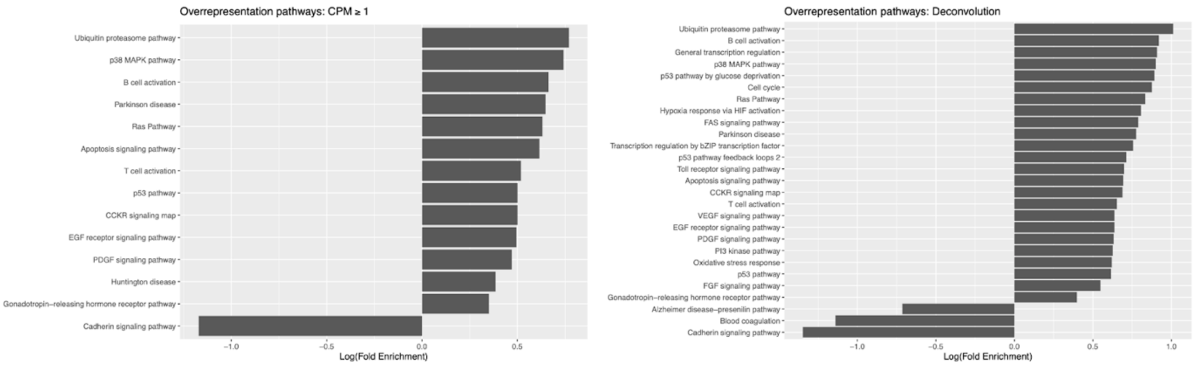

Figure 9: Enrichment test of functional pathways for genes called present in the B-cell population.The graphics represent the over and underrepresented pathways using **(A)** the genes called present using CPM ≥ 1 and **(B)** the genes called present by the reference intergenic method. All pathways shown have $p < 4.35e\text{-}3$.

Calls over multiple libraries

When multiple RNA-Seq libraries describe the same condition (e.g. the same organ), either from one experiment or several, it can be useful to make gene expression calls taking all of the available libraries into account (i.e., to obtain one call per gene and per condition). We applied three quantitative methods (see methods section), two derived from our reference intergenic approach, and one recently published Bayesian method [22], zigzag, which necessitates multiple libraries to make expression calls. We first evaluated method performance based on the reference datasets used in the zigzag publication [22], i.e. a set of genes known to be expressed in human lung based on epigenetic markers and in fly testis based on developmental genetic studies. The RNA-Seq data used are from Bgee samples matching the conditions of these benchmark datasets.

For human lung RNA-Seq, three samples were not included in the zigzag analysis because of their high variance (Supplementary figure 9). Indeed, when all 26 samples were used (Figure 10A) we detected a problem likely associated with how the variance of genes across samples is initialized. The high discrepancy in variance between samples impacted the convergence of the MCMC. Zigzag was designed for the initial values of variances for genes to be small, and in these libraries that is not the case. Given that, we could either remove from the zigzag analysis the libraries with the highest variance between genes, or the genes with the highest variance between libraries. We chose to remove libraries (Figure 10B). The high variance between samples is not a problem for the intergenic-based methods, because they are based on p-value or q-value computed in each sample independently. Hence, all samples can be included for the intergenic-based methods (Figure 10A).

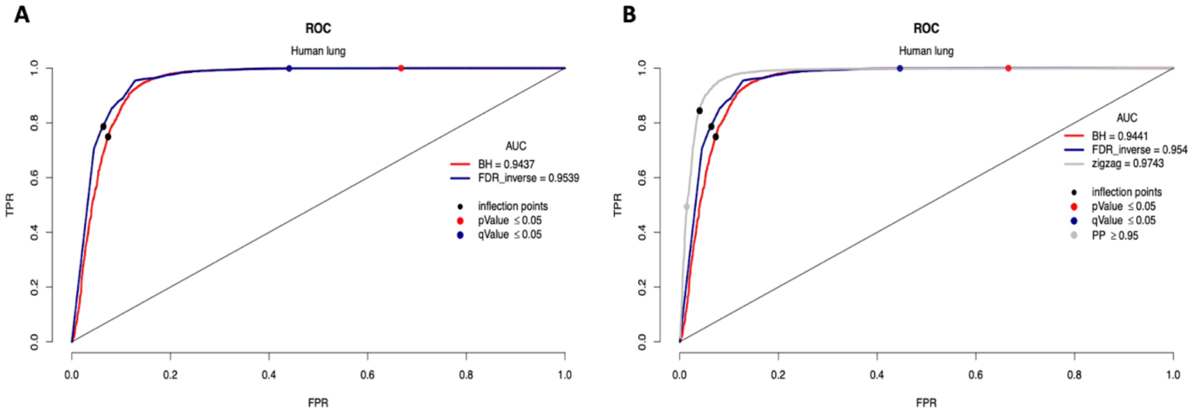

Figure 10: Comparison of the performance of three different methods to call active expression on a combination of libraries for human lung data. **(A)** Calls of expression using all 26 libraries. **(B)** Calls of expression restricted to 23 libraries for which zigzag could be run. True positive rates (TPR) as a function of False positive rates (FPR). The blue and red lines show the performance of the intergenic-based methods (FDR inverse and BH method) and the gray line shows the performance of zigzag. For each method we estimated the inflection point (black dots) and we show the standard thresholds for each method as colored dots.

All three methods perform very well based on AUC, with the top score for the Bayesian method, with an AUC of 0.97 (Figure 10) when just a subset of 23 samples is included in the analysis. On the other hand, the intergenic-based methods perform very well in both situations, with all 26 samples or just a subset of samples. For each method, we report two thresholds: the Extremum Distance Estimator (EDE) for inflection point calculated by using the coordinates of the TPR/FPR curve (black dots in Figure 10); and standard thresholds, i.e. for the Bayesian method a posterior probability PP ≥ 0.95 and for the intergenic based methods $q \leq 0.05$ or $p \leq 0.05$. With the inflection point thresholds, the false positive rates are low for all methods, with true positive rates of at least 75% (Table 5), but such thresholds will rarely be used in practice. On the other hand, with a standard threshold the true positive rate is much lower for the Bayesian method, while the false positive rates are much higher for the intergenic-based methods.

Method	Threshold	Number samples	TP	FP	TN	FN	TPR	FPR
BH	inflection point ($p \leq 3.28e\text{-}13$)	26	5408	323	4058	1812	0.749	0.074
FDR inverse	inflection point ($q \leq 3.54$ e-11)	26	5536	256	4179	1684	0.767	0.058
zigzag	inflection point (PP ≥ 0.707)	23	6097	179	4256	1123	0.844	0.040
BH	$p \leq 0.05$	26	7219	2925	1456	1	0.999	0.668
FDR inverse	$q \leq 0.05$	26	7210	1954	2481	10	0.998	0.440
zigzag	PP ≥ 0.95	23	3566	63	4372	3654	0.493	0.014

Table 5: Benchmark of combining multiple libraries for human lung data. Abbreviations: True positive (TP), False positive (FP), True negative (TN), False negative (FN), True positive rate (TPR), False positive rate (FPR).

We performed similar analyses with the fly testis benchmark. Surprisingly, with zigzag no true positive was found with PP ≥ 0.95 (Figure 11A). That can be explained by the difference in the log TPM distribution across the six samples, where 2 samples show high variance (Supplementary figure 10). In this method it is necessary to set thresholds in order to have identifiability constraints. For these samples we set the threshold α (not to be confused with the threshold of p-value testing) between 1 and 4 (Supplementary figure 10). The first threshold is important to catch the upper boundary for the inactive mean prior distribution that is also the lower boundary for the first active mean prior distribution, and the second threshold is important to define the lower boundary for a possible additional component of very highly expressed genes [22]. To set these thresholds using all the testis samples is difficult, given the dispersed density distribution of the data. Given that, and even if we were able to make an initial estimation using all samples, that estimation can be biologically misleading. To get a better estimation from zigzag we chose to remove these 2 samples with high variance for fly testis. After removing these two samples from the Bayesian analysis, all the approaches again perform quite similarly in terms of AUC, ranging from 0.98 for BH to 0.99 with the FDR inverse method (Figure 11B).

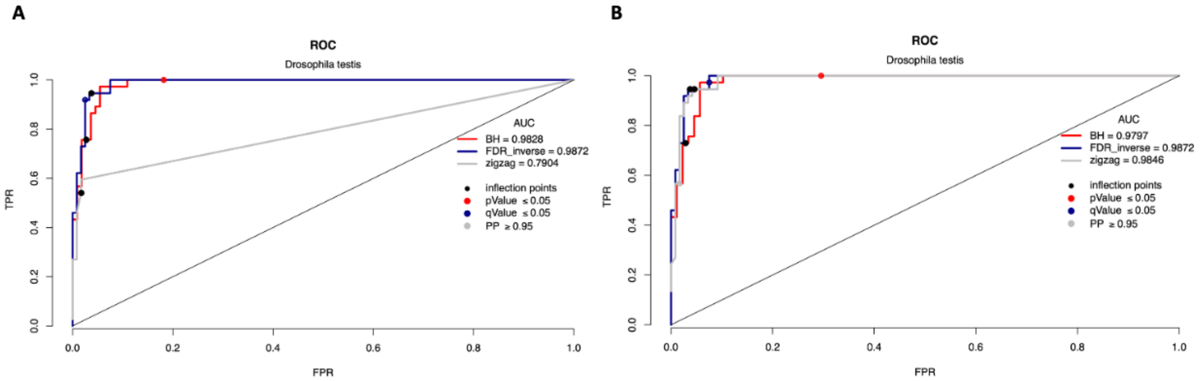

Figure 11: Comparison of the performance of three different methods to call active expression on a combination of libraries for fly testis data. **(A)** Calls of expression using all 6 libraries. **(B)** Calls of expression restricted to 4 libraries. True positive rates (TPR) as a function of False positive rates (FPR). The gray line shows the performance of zigzag, the blue and red lines show the performance of the intergenic-based methods (FDR inverse and BH method). For each method we estimated the inflection point (black dots) and we show the standard thresholds for each method as colored dots.

Using the inflection point thresholds again provides a very low false positive rate in all methods (Table 6), with a lower true positive rate for the BH method. Using the standard thresholds the Bayesian method again has a low true positive rate, while the BH method has a relatively high false positive rate compared with other methods. The FDR inverse method performs well on both criteria.

Method	Threshold	Number samples	TP	FP	TN	FN	TPR	FPR
BH	inflection point ($p \leq 0.0016$)	6	28	3	107	9	0.757	0.027
FDR inverse	inflection point ($q \leq 0.0086$)	6	35	4	116	2	0.946	0.033
zigzag	inflection point (PP $\geq$ 0.635)	4	35	6	114	2	0.946	0.05
BH	$p \leq 0.05$	6	37	20	90	0	1	0.182
FDR inverse	$q \leq 0.05$	6	34	3	117	3	0.919	0.025
zigzag	PP $\geq$ 0.95	4	21	2	118	16	0.568	0.017

Table 6: Benchmark of combining multiple libraries for fly testis data. Abbreviations: True positive (TP), False positive (FP), True negative (TN), False negative (FN), True positive rate (TPR), False positive rate (FPR).

Finally, we used mouse liver Ribo-Seq as an additional benchmark, since active translation by ribosomes provides a clear signal of activity of protein-coding genes that should only be seen in actively transcribed genes. Once again, all three methods display very good AUC values (Figure 12), from 0.94 for BH to 0.97 for zigzag. The Bayesian method performed very well using the inflection point threshold (Table 7). However with a standard threshold it had a very low true positive rate, while the BH method had a very high rate of false positives, and again the FDR inverse performs well.

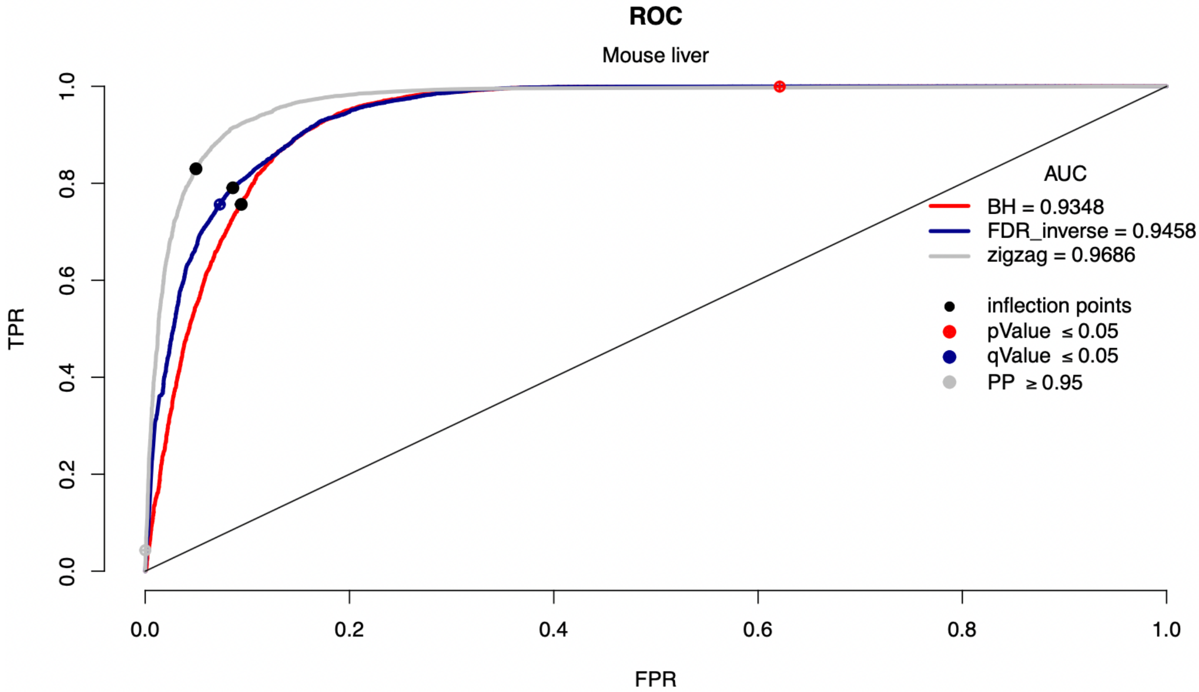

Figure 12: Comparison of the performance of three different methods to call active expression on a combination of libraries for mouse liver data. True positive rates (TPR) as a function of False positive rates (FPR). The gray line shows the performance of zigzag, the blue and red lines show the performance of the intergenic-based methods (FDR inverse and BH method). For each method we estimated the inflection point (black dots) and we show the standard thresholds for each method as colored dots.

Method	Threshold	Number samples	TP	FP	TN	FN	TPR	FPR
BH	inflection point ($p \leq$ 1.24e-7)	89	6476	1160	11166	2083	0.757	0.094
FDR inverse	inflection point ($q \leq 0.00066$)	89	6768	1147	12230	1791	0.791	0.086
zigzag	inflection point (PP $\geq$ 0.428)	89	7104	661	12716	1455	0.83	0.05
BH	$p \leq 0.05$	89	8555	7657	4669	4	0.999	0.62
FDR inverse	$q \leq 0.05$	89	6474	976	12401	2085	0.756	0.07
zigzag	PP $\geq$ 0.95	89	370	10	13367	8189	0.043	0.0007

Table 7: Benchmark of combining multiple libraries for mouse liver data. Abbreviations: True positive (TP), False positive (FP), True negative (TN), False negative (FN), True positive rate (TPR), False positive rate (FPR).

Over the three datasets, the FDR inverse approach provides a relatively high recovery of true positives while controlling the false discovery rate, especially relative to the other methods using a standard threshold, which is the expected usage of methods in practice. On the other hand, the BH approach can keep a high true positive rate even with much more stringent thresholds (Supplementary table 3). The FDR inverse approach is based on the q-value, which has the drawback of increased variation between individual samples (Figure 13). Zigzag performs well on the AUC criterion, but with a standard threshold has low sensitivity, and cannot be applied to all sets of libraries.

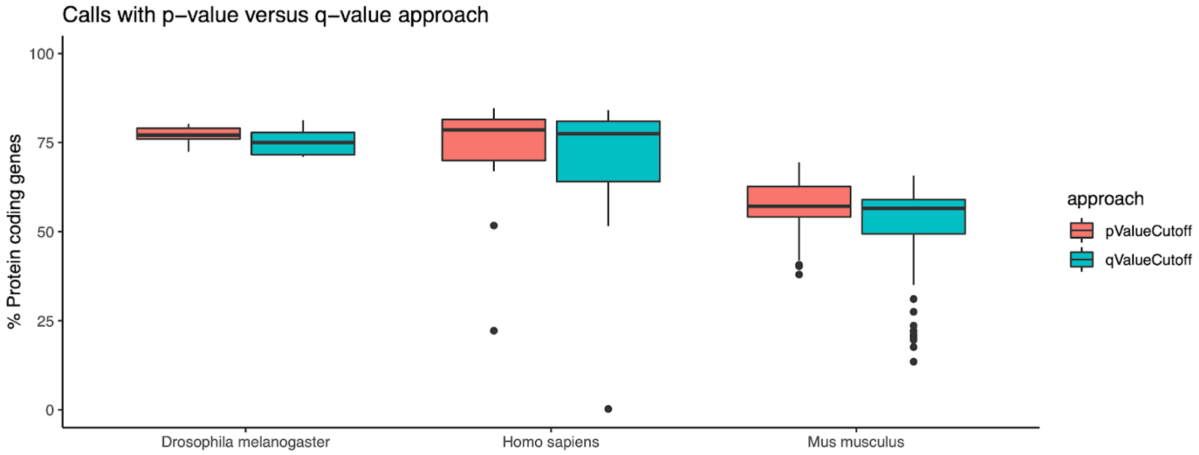

Figure 13: Proportion of protein coding genes called present using two different methods: $p \leq 0.05$ and $q \leq 0.05$. The data used for the analysis/plotting are the data used to benchmark combining multiple libraries: Drosophila testis samples, Human lung samples and Mouse liver samples.

Discussion

Determining when and where a gene is expressed is a fundamental level of information on gene function. To study this, we need to determine whether a gene is expressed or not in a given condition. Moreover, it is good practice to restrict quantitative analyses such as differential expression to analyzing the signal from expressed genes, not background transcriptional noise. Thus it is important to be able to infer reliably gene expression calls from quantitative RNA-Seq data. Ideally such a method should be robust to skewed expression among genes, poor genome annotation, and low numbers of replicates. As noted by Thompson et al. [22]: “Inferring the expression state (active or inactive) of a given gene from transcriptomic datasets is surprisingly difficult”.

Several methods have been proposed based on modeling the distribution of expression levels of genes [18] [21], [22]. In this work we propose an empirical approach to estimate background noise of RNA libraries, which can be applied even to a single library. We propose to use a set of reference intergenic regions to perform a call of expression per gene in a sample.

We are not the first to compare RNA-Seq counts mapping to exonic and other genome regions. Hebenstreit et al. [4] showed that intronic regions had intermediary expression levels, in-between those of intergenic regions and exonic regions. This is consistent with the expectation that extracted RNA will contain immature or partially mature mRNAs, as well as unannotated exons or exon boundaries within known genes. While this observation is in principle interesting [45], it is not helpful to estimate background transcriptional noise and call genes actively expressed. Hebenstreit et al. [4] reported the 90% quantile of the distribution of intergenic TPM values as a threshold above which genes could be called expressed with strong confidence. Of note, their intergenic regions had very low expression levels overall, spreading on the right-end of the TPM spectrum to overlap with exonic sequences. This is probably due to the inclusion of non-coding genes which were absent from the annotation version used: annotated non-coding RNA genes have increased from 9240 to 19291 [46], of which 13186 long non-coding RNA genes (from 3845). A later study explored the use of different thresholds on intergenic or intronic expression to call genes present [47], but it was based on only one RNA-Seq experiment, and did not provide any guidelines. The first study to our knowledge to use intergenic transcription to define gene expression established a 0.4 RPKM threshold balancing the numbers of false positives and false negatives, to define a set of human house-keeping genes [48]. Again, the distribution of RPKM values for intergenic regions spread was very large. These results and our own show that the use of intergenic regions to estimate false discovery rates in gene expression is promising, but that it is important to define reference intergenic sequences which are truly not actively expressed, thus excluding unannotated genes or other potentially transcribed regions.

To define these reference intergenic regions, we rely here on the curated healthy wild-type libraries in the Bgee database [3]. We propose that it is important to exclude tumor samples and immortalized cell lines from this step, since transcription can be deregulated in such cells, which could lead to wrong attributions of intergenic regions. It is also important in principle to include as diverse samples as possible, representing different cell states and types. Yet the method already works well in species with as few as two different anatomical structures, such as *Poecilia reticulata* with 45 libraries, as well as in some species with few libraries, such as *Drosophila pseudoobscura* with 10 libraries and just four anatomical structures (Supplementary figure 5). Our method is also robust to the use of different mapping or pseudo-mapping for quantification, as shown by comparing results between pseudo-aligners Kallisto [49] or Salmon [50] and the aligner HISAT2 [51] (see Methods section).

A median of 73% of all intergenic regions are selected as reference intergenic over all species (Supplementary figure 11). While there is a positive relation between this proportion and the proportion of non-coding genes in the genomes (Pearson r = 0.55), even for the less well annotated genomes we are able to recover >50% of intergenic regions in the reference. The intergenic regions that are not classified as reference present in general a higher GC content (Supplementary figure 12). On the other hand, it is consistent with known higher GC of genic regions [52], which might be missannotated as intergenic. There is no apparent chromosomal pattern of distribution of the reference intergenic regions.

For species present in Bgee we provide reference intergenic regions predefined, but it is straightforward to extend the approach to other species. In practice, we have only tested animal genomes. Other large eukaryotic genomes with large intergenic regions should behave similarly. We did not explore applicability to more compact genomes, such as yeasts or bacteria, but the lack of large intergenic regions might limit our power to define background expression.

The method which we propose appears robust to low coverage samples and to high variance between samples when combining libraries, and has much shorter computation times than the Bayesian method proposed by Thompson et al. [22]). Importantly, our method can be applied to single cell RNA-Seq data from different protocols with no change. Moreover, once the reference set of intergenic regions are defined for a species they can be applied to any existing or new libraries from this species; no posterior controls such as MCMC convergence or sampling are needed. Interestingly, the reference intergenic regions which we compute can be used as reference for other data, for example providing a baseline for promoter histone marks [53].

The two intergenic-based methods (pValue and qValue) presented in the library combination section can both be useful according to scientific questions, according to use cases. Given that the Bgee database annotates individual samples we use the p-value method, which has low variance between samples (Figure 13). Consequently, we use the BH method to combine libraries in Bgee. Relative to the Bayesian approach implemented in zigzag, this method can also be applied to samples with high variance, from different RNA-Seq experiments, and even from different anatomical entities (e.g. combine different brain parts to provide a “brain” call) or other sampling variables.

In conclusion, the approach that we present provides a robust, versatile, computationally efficient, and easy to scale method to distinguish genes which are actively expressed. It can be applied to single libraries as well as to sets, from one or several experiments, and to different bulk or single-cell protocols.

Methods

General approach

For a species under consideration we need an annotated reference genome sequence and a set of RNA-Seq libraries (Figure 14, in blue). In practice, we use curated healthy wild-type libraries from the Bgee database [3], but other sources can be used. From this genome we extract two sets of genomic regions for mapping of RNA-Seq reads: exons (coding and non coding), i.e., genic regions, and putative intergenic regions. The latter are defined as regions of DNA which are at least 0.5 kb from a gene annotation, and are at least 1 kb long. For regions which are longer than 20 kb, we keep only the 20 kb at the center of the region (i.e., 10 kb on each side of the center). These choices can be adapted to the structure and compactness of different genomes; here we focused on animal genomes. The definition of these putative intergenic regions is dependent on the quality of the genome annotation at this stage.

Reads from each library are mapped to the genic and putative intergenic regions, and the counts per region are recovered. In practice we use Kallisto (version 0.46.0) [54] to quantify the abundance of transcripts from RNA-Seq data using pseudo-mapping, but the principle would be similar for other read mapping tools (e.g., HISAT2 (version 2.2.0)) (Figure 15); we sum over transcripts per gene. We sum read counts over all libraries for a given species, followed by TPM normalization [15] to obtain the distributions of log2(TPM) for the genic and intergenic regions. Note that for log transformation we add a small pseudo-count to the read counts (1e-06). Within the genic regions we further distinguish the distribution of TPMs of protein-coding genes, which are usually reasonably well-annotated, even in non-model species.

Intergenic region selection

Publicly available genomes contain a variable proportion of ambiguous bases, coded as ‘N’. For the 52 species present in Bgee 15 this varies from 0% in *Caenorhabditis elegans* to 35.5% in *Gadus morhua* (Supplementary table 4). These Ns are often present in blocks of up to thousands of bases, and mostly affect intergenic regions. As absence of genome annotation is the only criterion to generate candidate intergenic regions, and reads will map poorly to regions with high proportion of N, such regions could bias downstream abundance quantifications. For instance, Kallisto pseudorandomly transforms all N bases to A, T, C, or G bases during the pseudo-mapping step. We thus generated a version of the intergenic regions from which we removed blocks of N longer than 31 bp (default k-mer size of Kallisto index) and all sequences containing more than 5% of N, as well as those smaller than 1 kb after N removal. This set is available at https://bgee.org/ftp/intergenic/1.0/ref_intergenic/; it contains a median of 0.000973% of Ns per region. We recommend performing a similar filtering for any new genome to analyze.

While some well-annotated species (e.g., human, mouse, fruit fly) have only a very small fraction of intergenic regions with high expression levels, other species, often non-model, present large fractions of "intergenic regions" with high expression levels (Supplementary figure 6). This is likely due to the poor annotation of various transcribed elements, such as non coding genes or very short coding genes. To define a strict set of intergenic regions, excluding these actively expressed regions, we deconvolute the summed intergenic log2(TPM) distribution into a set of Gaussian distributions, following the concept from Hebenstreit et al. [4]. We use Mclust [55] for deconvolution, fitting a Gaussian mixture with the number of components chosen by a Bayesian Information Criterion.

To define which of the Gaussians represent truly reference – i.e., non-expressed – intergenic regions, we use manual curation, although this could be automated in the future. For each species, the overlap of the intergenic log2(TPM) Gaussians to the summed log2(TPM) distribution of protein-coding regions is considered. The latter are used as a reference set of genes, which should be overall actively expressed in at least some conditions, and which should show little overlap of expression levels with true intergenic regions. From the deconvoluted Gaussian of these non expressed intergenic regions, we recover a threshold in TPM unit, namely the highest TPM value of any region classified as originating from this Gaussian, referred to in this paper as the *species-specific maximum inactively transcribed threshold*, or $\text{MITT}_{\text{species}}$. The final set of reference intergenic regions are defined as all intergenic regions with a lower TPM than $\text{MITT}_{\text{species}}$ (Figure 14A in blue and Figure 14B and C). Note that other intergenic regions are no longer used for the rest of the procedure.

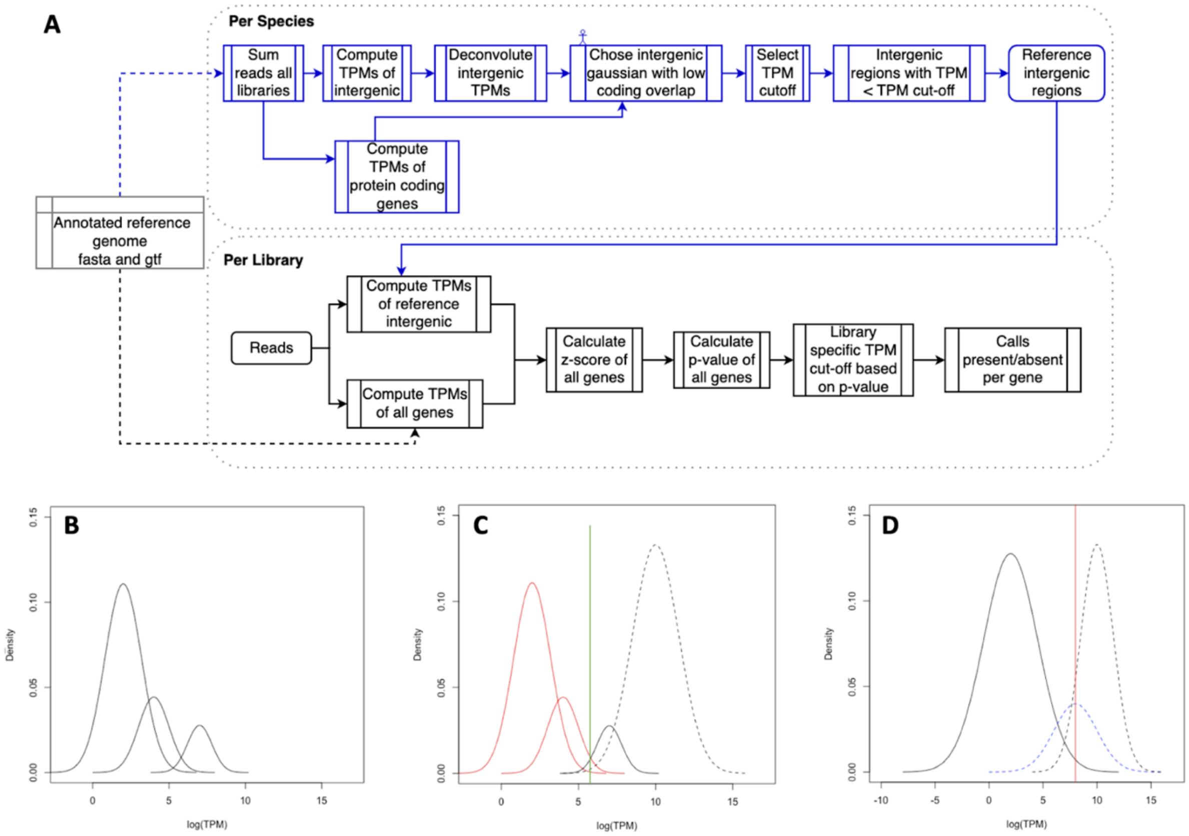
A
Per Species
Sum reads all libraries
Compute TPMs of intergenic
Deconvolute intergenic TPMs
Chose intergenic gaussian with low coding overlap
Select TPM cutoff
Intergenic regions with TPM < TPM cut-off
Reference intergenic regions
Compute TPMs of protein coding genes
Annotated reference genome fasta and gtf
Per Library
Reads
Compute TPMs of reference intergenic
Compute TPMs of all genes
Calculate z-score of all genes
Calculate p-value of all genes
Library specific TPM cut-off based on p-value
Calls present/absent per gene
B
C
D
Density
0.00
0.05
0.10
0.15
-10
-5
0
5
10
15
log(TPM)

Figure 14: Flow chart of information for calls of expression from RNA-Seq. **A)** In blue, steps done once per species; in black, steps done for each library separately. The figure uses "TPM" for simplicity, but any measure of normalized read counts can be used (e.g., FPKM, or CPM of UMI counts). **B-D**) illustrative examples of the distributions of TPMs: **B)** deconvolution of TPMs computed from all intergenic regions and from all libraries; **C)** using TPMs of coding genes from all libraries (dashed line), Gaussians of "non expressed intergenic" regions are defined (in red) and MITT$_{species}$ represented by a vertical green line; **D)** for one library, the distributions of TPMs for reference intergenic (full line), are used to define a minimum actively transcribed threshold (MATT$_{library}$), represented by a vertical red line, of expression for non coding genes (dashed line, blue), and coding genes (dashed line, black). In B and C we compute the density of log(TPM) for genes and intergenic, and normalize the two density curves for the number of genes and of intergenic regions, so that the two curves are directly comparable.

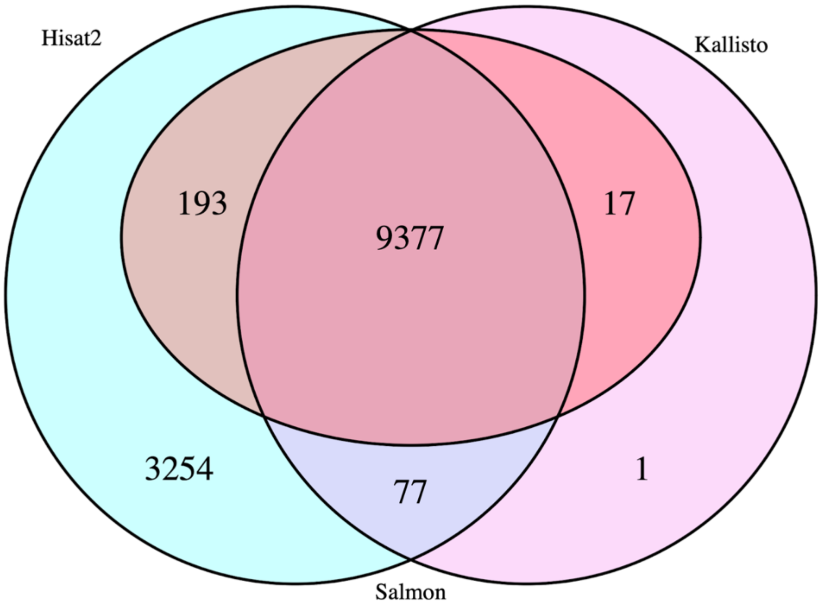

Figure 15: Classification of reference intergenic regions using alignment-based (HISAT2) or pseudo-mapping (Kallisto or Salmon) methods. Applied to 24 pike (*Esox lucius*) libraries integrated in Bgee 15.0. The match of reference intergenic regions selected from Salmon and Kallisto is 97%, between Salmon and HISAT2 the match is 74% and between Kallisto and HISAT2 the match is 73% of intergenic regions classified as reference intergenic.

Sample-specific expression threshold

In this work we introduce two methods, pValue and qValue (based on the corresponding concepts), to detect genes actively expressed in any standalone RNA-Seq library. The steps of either method are performed for each library separately in order to define a sample-specific threshold, or *minimum actively transcribed threshold* $\mathrm{MATT}_{\mathrm{library}}$, to classify active and inactive genes (Figure 14A, in black). We map reads both to transcripts and to the reference intergenic regions, and compute TPMs per gene (summing over transcripts) and per intergenic region – again using Kallisto in practice. The two reference intergenic methods for the single library are implemented in the BgeeCall R package [23].

pValue method

For each gene *i* in the library, we compute a Z-score in terms of standard deviations from the mean of reference intergenic regions.

$$ZScore_{gene_i} = \frac{(log2(TPM_{gene_i}) - mean(log2(TPM_{RefIntergenic})))}{sd(log2(TPM_{RefIntergenic}))}$$

Then for gene *i* in this library we calculate a p-value based on a null hypothesis of expression at a similar level to reference intergenic, estimated as a Normal distribution. In practice we use the pnorm() function in R:

$$pValue = pnorm(zScore, lower.tail{=}FALSE)$$

The library-specific TPM limit to call genes expressed is the minimum value of TPM where p-value $\leq \alpha$; in the Results we will use $\alpha = 0.05$ unless otherwise specified.

qValue method

We propose a q-value based approach as an alternative to the p-value presented above. The idea is that the distribution of TPM values from intergenic regions can be used to estimate the false positive rate at each TPM threshold. We calculate for each log2(TPM) value the area under the density distribution curve for the genic and intergenic regions. Based on that, the qValue parameter, for a corresponding log2(TPM) value, is the ratio between the numerical integration value to the background noise on non expressed regions and the sum of the numerical integration value to the background noise with the numerical integration value of the genic region.

In detail, for each individual sample, a numerical integration by linear interpolation is calculated from all density distribution of log2(TPM) values of genic and intergenic regions, as shown in the formula below:

$$integration_{region} = \int_{min_{dr}}^{max_{dr}} f(x).dx$$

where *f(x)* is a linear interpolated function of the density of the region, and the min_{dr} and max_{dr} are the limits across which the area is calculated taking into consideration the density of the region (*dr*); regions can be either genic or intergenic regions.

Then for each unique log2(TPM) value we calculate the numerical integration from the genic and intergenic density distribution curves. The integration is calculated from the log2(TPM) value until the maximum of the density distribution of each region.

$$unscaled_{region} = \int_{\log 2TPM_i}^{max_{dr}} f(x).dx$$

The integrated value is then scaled by the numerical integration of the all density regions.

$$scaled_{region} = \frac{unscaled_{region}}{integration_{region}}$$

Finally, the q-value$_i$ of each gene *i* is calculated following the formula:

$$qValue_i = \frac{scaled_{intergenic}}{scaled_{intergenic} + scaled_{genic}}$$

The library-specific TPM limit to call genes expressed is the minimum value of TPM where q-value $\leq \alpha$.

Combining multiple libraries

For some applications it can be of interest to determine genes which are actively expressed over a set of libraries. In order to combine information from multiple libraries, we tested three different methods (Table 8). The combination methods developed in this work (BH and FDR inverse) are implemented in the BgeeCall R package. We compared the performance of our two combination methods with the Bayesian method “zigzag” of Thompson et al. [22] (R Package version: zigzag_0.1.0 , and repository: https://github.com/ammonthompson/zigzag), across different conditions, such as developmental stages, sexes or strains. Indeed by design zigzag requires multiple libraries to infer gene expression state.

To validate these approaches and calculate true and false positive rates, we used the same datasets as in Thompson et al. [22], i.e. active/inactive genes defined from epigenomic data from *Homo sapiens* (human) lung [22] and genetic evidence from *Drosophila melanogaster* (fly) testis [22]. We also used a ribosome footprint from *Mus musculus* (mouse) liver [56] (see below).

Method	**Calls per library**	**Approach used to combine libraries**	**Reference**	**Availability**
pValue	Yes	BH	this work	BgeeCall package
qValue	Yes	FDR inverse	this work	BgeeCall package
zigzag	No	Bayesian inference	Thompson et al.[22]	zigzag package

Table 8: Two methods proposed in this work to call expressed genes at individual sample level then combine information from multiple libraries, and a third method from the literature to combine libraries. The two approaches proposed in this work use the reference set of intergenic regions to call expressed genes.

In the first approach, pValue, we simply combine information from multiple libraries by applying the Benjamini-Hochberg procedure (BH) [57] on the p-value computed per library as described above. In the second approach, to combine libraries based on the qValue computed per library (q_{library}) as described above, we propose an "FDR inverse" approach to control the false discovery rate. We define a q-value threshold taking into consideration the number N of libraries to combine and the $q_{library}$ threshold α used in single libraries:

$$FDR_{inverse} = 1 - ((1 - \alpha)^{1/N})$$

Mouse liver benchmark

We use a processed mouse liver Ribo-Seq dataset downloaded from GEO (GSE67305) to benchmark methods both at individual library level and in the process of combining multiple libraries to call genes present and absent. Active translation by ribosomes provides a clear signal of activity of protein-coding genes, so true positives were defined as the union of protein-coding genes detected in all samples with ≥ 1 RPKM and of protein-coding genes detected in at least one sample with ≥ 5 RPKM. True negatives were defined as protein-coding genes which matched neither condition.

Analysis of scRNA-Seq data (droplet-based protocols)

For droplet-based protocols dataset using the 10X Genomics technology, each sample was pseudo-mapped using Kallisto software and the output files were treated with the bustools [58] software (version 0.40.0). The bustools output was read into R with the BUSpaRse R package (version 1.3.1) and data analysis was performed in R.

The annotation of cells to cell types was based on the mapping provided by the original publication (using the barcodes information). To perform calls of expression per gene and cell type we used a "pseudo-bulk" approach [59] whereby the read counts of cells belonging to the same cell type and sample were summed before CPM normalization.

Functional enrichment analysis

We performed overrepresentation tests for PANTHER pathways [60] by using the annotation from the PANTHER database [61] http://pantherdb.org/ (PANTHER version 16.0 released on 2020-12-01). The Gene List Analysis tool was used with Fisher's exact test and FDR correction.

HISAT2

The mapping of RNA-Seq libraries from pike samples was performed using HISAT2 version 2.2.0. We used the default commands to build the index and to perform the mapping (scripts available on github repository). The sam output files generated by HISAT2 were converted into bam files and then we used salmon software to make the alignments quantification.

Salmon

The pseudo-mapping of RNA-Seq libraries from pike samples was performed using Salmon version 0.12.0. The index was built with k-mers of length 31, since all pike libraries have read lengths higher than 75bp. The commands to build and to make the pseudo-alignment were used by default (scripts available on github repository).

CG content

The CG content of the intergenic regions were quantified separately for regions classified as reference intergenic and other intergenic regions for all species using the SeqKit software [62] version 2.0.0 with following command:

```
seqkit fx2tab --name --gc species_X_Refintergenic.fa > species_X_Refintergenic_CG.tsv
seqkit fx2tab --name --gc species_X_other_intergenic.fa > species_X_other_intergenic_CG.tsv
```

Data access

All empirical data used in this work for RNA-Seq and single cell full-length protocols were obtained from Bgee 15.0 and are accessible through the BgeeDB R package [63]. The empirical data for droplet-based protocols (data not inserted in Bgee 15.0) are from the experiments SRP135999 [64] and SRP201320 [39].

All input files and specific code to reproduce the figures or statistical files of this paper are available at https://github.com/BgeeDB/Methods_RNASeq_expression_calls . The full Bgee pipeline for RNA-Seq is available at https://github.com/BgeeDB/bgee_pipeline/tree/master/pipeline/RNA_Seq, the pipeline for scRNA-Seq using full-length protocols is available at https://github.com/BgeeDB/bgee_pipeline/tree/develop/pipeline/scRNA_Seq/Full_Length_Protocols and the full droplet-based pipeline is available at (https://github.com/BgeeDB/bgee_pipeline/tree/develop/pipeline/scRNA_Seq/Droplet_based_Protocols). The BgeeCall R package used for methods in this work can be found linked to a specific tag (https://github.com/BgeeDB/BgeeCall/tree/calls_paper).

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

We thank all the Robinson-Rechavi group for their useful discussions on the methods implemented on the paper. We acknowledge support by the SIB Swiss Institute of Bioinformatics, the Canton de Vaud, and Swiss National Science Foundation grant 31003A_173048.

Authors' contributions

MR, JR and MRR designed the original approach. SSFC, JR, JW and MRR refined it. SSFC performed the graphic visualization as well as tables for the paper. JR, FBB and MRR wrote the first draft of the paper. SSFC, JR and MRR wrote the final version of the paper. All the authors contributed to result interpretation and discussion. All authors read and approved the final manuscript.

References

[1] F. Ji and R. I. Sadreyev, "RNA-seq: Basic Bioinformatics Analysis," *Curr. Protoc. Mol. Biol.*, vol. 124, no. 1, p. e68, 2018, doi: 10.1002/cpmb.68.

[2] "RNA-seq Data Analysis: A Practical Approach," *Routledge & CRC Press*. https://www.routledge.com/RNA-seq-Data-Analysis-A-Practical-Approach/Korpelainen-Tuimala-Somervuo-Huss-Wong/p/book/9781466595002 (accessed Jan. 27, 2022).

[3] F. B. Bastian *et al.*, “The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals,” *Nucleic Acids Res.*, vol. 49, no. D1, pp. D831–D847, Jan. 2021, doi: 10.1093/nar/gkaa793.

[4] D. Hebenstreit, M. Fang, M. Gu, V. Charoensawan, A. van Oudenaarden, and S. A. Teichmann, “RNA sequencing reveals two major classes of gene expression levels in metazoan cells,” *Mol. Syst. Biol.*, vol. 7, p. 497, Jun. 2011, doi: 10.1038/msb.2011.28.

[5] J. M. Raser and E. K. O’Shea, “Noise in Gene Expression: Origins, Consequences, and Control,” *Science*, Sep. 2005, doi: 10.1126/science.1105891.

[6] S. Pinheiro, S. Pandey, and S. Pelet, “Cellular heterogeneity: yeast-side story,” *Fungal Biol. Rev.*, vol. 39, pp. 34–45, Mar. 2022, doi: 10.1016/j.fbr.2021.11.005.

[7] A. Sanchez, S. Choubey, and J. Kondev, “Regulation of Noise in Gene Expression,” *Annu. Rev. Biophys.*, vol. 42, no. 1, pp. 469–491, May 2013, doi: 10.1146/annurev-biophys-083012-130401.

[8] W. J. Blake, M. KAErn, C. R. Cantor, and J. J. Collins, “Noise in eukaryotic gene expression,” *Nature*, vol. 422, no. 6932, pp. 633–637, Apr. 2003, doi: 10.1038/nature01546.

[9] W. J. Blake *et al.*, “Phenotypic consequences of promoter-mediated transcriptional noise,” *Mol. Cell*, vol. 24, no. 6, pp. 853–865, Dec. 2006, doi: 10.1016/j.molcel.2006.11.003.

[10] A. Raj and A. van Oudenaarden, “Stochastic gene expression and its consequences,” *Cell*, vol. 135, no. 2, pp. 216–226, Oct. 2008, doi: 10.1016/j.cell.2008.09.050.

[11] A. Becskei, B. B. Kaufmann, and A. van Oudenaarden, “Contributions of low molecule number and chromosomal positioning to stochastic gene expression,” *Nat. Genet.*, vol. 37, no. 9, pp. 937–944, Sep. 2005, doi: 10.1038/ng1616.

[12] A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi, “Stochastic mRNA synthesis in mammalian cells,” *PLoS Biol.*, vol. 4, no. 10, p. e309, Oct. 2006, doi: 10.1371/journal.pbio.0040309.

[13] J. M. Raser and E. K. O’Shea, “Control of stochasticity in eukaryotic gene expression,” *Science*, vol. 304, no. 5678, pp. 1811–1814, Jun. 2004, doi: 10.1126/science.1098641.

[14] L. Warren, D. Bryder, I. L. Weissman, and S. R. Quake, “Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 47, pp. 17807–17812, Nov. 2006, doi: 10.1073/pnas.0608512103.

[15] G. P. Wagner, K. Kin, and V. J. Lynch, “Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples,” *Theory Biosci. Theor. Den Biowissenschaften*, vol. 131, no. 4, Art. no. 4, Dec. 2012, doi: 10.1007/s12064-012-0162-3.

[16] M. Melé *et al.*, “Human genomics. The human transcriptome across tissues and individuals,” *Science*, vol. 348, no. 6235, pp. 660–665, May 2015, doi: 10.1126/science.aaa0355.

[17] K. Kin *et al.*, “The Transcriptomic Evolution of Mammalian Pregnancy: Gene Expression Innovations in Endometrial Stromal Fibroblasts,” *Genome Biol. Evol.*, vol. 8, no. 8, pp. 2459–2473, Aug. 2016, doi: 10.1093/gbe/evw168.

[18] G. P. Wagner, K. Kin, and V. J. Lynch, “A model based criterion for gene expression calls using RNA-seq data,” *Theory Biosci. Theor. Den Biowissenschaften*, vol. 132, no. 3, pp. 159–164, Sep. 2013, doi: 10.1007/s12064-013-0178-3.

[19] P. Moreno *et al.*, “Expression Atlas update: gene and protein expression in multiple species,” *Nucleic Acids Res.*, p. gkab1030, Nov. 2021, doi: 10.1093/nar/gkab1030.

[20] “An open RNA-Seq data analysis pipeline tutorial... | F1000Research.” https://f1000research.com/articles/5-1574/v1 (accessed Jan. 21, 2022).

[21] T. Hart, H. K. Komori, S. LaMere, K. Podshivalova, and D. R. Salomon, “Finding the active genes in deep RNA-seq gene expression studies,” *BMC Genomics*, vol. 14, p. 778, Nov. 2013, doi: 10.1186/1471-2164-14-778.

[22] A. Thompson, M. R. May, B. R. Moore, and A. Kopp, “A hierarchical Bayesian mixture model for inferring the expression state of genes in transcriptomes,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 117, no. 32, pp. 19339–19346, Aug. 2020, doi: 10.1073/pnas.1919748117.

[23] J. Wollbrett, S. F. Costa, J. Roux, M. R. Rechavi, and F. Bastian, *BgeeCall: Automatic RNA-Seq present/absent gene expression calls generation*. Bioconductor version: Release (3.14), 2022. doi: 10.18129/B9.bioc.BgeeCall.

[24] P. He *et al.*, “The changing mouse embryo transcriptome at whole tissue and single-cell resolution,” *Nature*, vol. 583, no. 7818, Art. no. 7818, 2020, doi: 10.1038/s41586-020-2536-x.

[25] J. Ruiz-Orera *et al.*, “Origins of De Novo Genes in Human and Chimpanzee,” *PLOS Genet.*, vol. 11, no. 12, Art. no. 12, Dec. 2015, doi: 10.1371/journal.pgen.1005721.

[26] E. B. Rondeau *et al.*, “The genome and linkage map of the northern pike (Esox lucius): conserved synteny revealed between the salmonid sister group and the Neoteleostei,” *PloS One*, vol. 9, no. 7, p. e102089, 2014, doi: 10.1371/journal.pone.0102089.

[27] GTEx Consortium, “Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans,” *Science*, vol. 348, no. 6235, Art. no. 6235, May 2015, doi: 10.1126/science.1262110.

[28] B. Xia *et al.*, “Widespread Transcriptional Scanning in the Testis Modulates Gene Evolution Rates,” *Cell*, vol. 180, no. 2, pp. 248-262.e21, Jan. 2020, doi: 10.1016/j.cell.2019.12.015.

[29] E. Witt, S. Benjamin, N. Svetec, and L. Zhao, “Testis single-cell RNA-seq reveals the dynamics of de novo gene transcription and germline mutational bias in Drosophila,” *eLife*, vol. 8, p. e47138, agosto 2019, doi: 10.7554/eLife.47138.

[30] “The molecular evolution of spermatogenesis across mammals | bioRxiv.” https://www.biorxiv.org/content/10.1101/2021.11.08.467712v1 (accessed Feb. 08, 2022).

[31] K. Krjutškov *et al.*, “Globin mRNA reduction for whole-blood transcriptome sequencing,” *Sci. Rep.*, vol. 6, no. 1, p. 31584, Aug. 2016, doi: 10.1038/srep31584.

[32] R. Jiang, T. Sun, D. Song, and J. J. Li, “Statistics or biology: the zero-inflation controversy about scRNA-seq data,” *Genome Biol.*, vol. 23, no. 1, p. 31, Jan. 2022, doi: 10.1186/s13059-022-02601-5.

[33] R. Neme and D. Tautz, “Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence,” *eLife*, vol. 5, p. e09977, doi: 10.7554/eLife.09977.

[34] Q. Deng, D. Ramsköld, B. Reinius, and R. Sandberg, “Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells,” *Science*, vol. 343, no. 6167, pp. 193–196, Jan. 2014, doi: 10.1126/science.1245316.

[35] J. Guo *et al.*, “Chromatin and Single-Cell RNA-Seq Profiling Reveal Dynamic Signaling and Metabolic Transitions during Human Spermatogonial Stem Cell Development,” *Cell Stem Cell*, vol. 21, no. 4, pp. 533-546.e6, Oct. 2017, doi: 10.1016/j.stem.2017.09.003.

[36] E. Z. Macosko *et al.*, “Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets,” *Cell*, vol. 161, no. 5, Art. no. 5, May 2015, doi: 10.1016/j.cell.2015.05.002.

[37] X. Wang, Y. He, Q. Zhang, X. Ren, and Z. Zhang, “Direct Comparative Analyses of 10X Genomics Chromium and Smart-seq2,” *Genomics Proteomics Bioinformatics*, vol. 19, no. 2, pp. 253–266, Apr. 2021, doi: 10.1016/j.gpb.2020.02.005.

[38] M. E. Ritchie *et al.*, “limma powers differential expression analyses for RNA-sequencing and microarray studies,” *Nucleic Acids Res.*, vol. 43, no. 7, p. e47, abril 2015, doi: 10.1093/nar/gkv007.

[39] H. G. Hilton *et al.*, “Single-cell transcriptomics of the naked mole-rat reveals unexpected features of mammalian immunity,” *PLoS Biol.*, vol. 17, no. 11, p. e3000528, Nov. 2019, doi: 10.1371/journal.pbio.3000528.

[40] A. F. McGettrick and L. A. J. O’Neill, “The Role of HIF in Immunity and Inflammation,” *Cell Metab.*, vol. 32, no. 4, pp. 524–536, outubro 2020, doi: 10.1016/j.cmet.2020.08.002.

[41] A. Yamada, R. Arakaki, M. Saito, Y. Kudo, and N. Ishimaru, “Dual Role of Fas/FasL-Mediated Signal in Peripheral Immune Tolerance,” *Front. Immunol.*, vol. 8, p. 403, Apr. 2017, doi: 10.3389/fimmu.2017.00403.

[42] Y. Tohyama, T. Takano, and H. Yamamura, “B cell responses to oxidative stress,” *Curr. Pharm. Des.*, vol. 10, no. 8, pp. 835–839, 2004, doi: 10.2174/1381612043452947.

[43] Z. Hua and B. Hou, “TLR signaling in B-cell development and activation,” *Cell. Mol. Immunol.*, vol. 10, no. 2, pp. 103–106, Mar. 2013, doi: 10.1038/cmi.2012.61.

[44] K. Igarashi, K. Ochiai, A. Itoh-Nakadai, and A. Muto, “Orchestration of plasma cell differentiation by Bach2 and its gene regulatory network,” *Immunol. Rev.*, vol. 261, no. 1, pp. 116–125, Sep. 2014, doi: 10.1111/imr.12201.

[45] S. Lee *et al.*, “Covering all your bases: incorporating intron signal from RNA-seq data,” *NAR Genomics Bioinforma.*, vol. 2, no. 3, Art. no. 3, Sep. 2020, doi: 10.1093/nargab/lqaa073.

[46] J. M. Mudge and J. Harrow, “Creating reference gene annotation for the mouse C57BL6/J genome assembly,” *Mamm. Genome*, vol. 26, no. 9, pp. 366–378, Oct. 2015, doi: 10.1007/s00335-015-9583-x.

[47] S. Harati, J. H. Phan, and M. D. Wang, “Investigation of factors affecting RNA-seq gene expression calls,” *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Int. Conf.*, vol. 2014, pp. 5232–5235, 2014, doi: 10.1109/EMBC.2014.6944805.

[48] D. Ramsköld, E. T. Wang, C. B. Burge, and R. Sandberg, “An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data,” *PLoS Comput. Biol.*, vol. 5, no. 12, Art. no. 12, Dec. 2009, doi: 10.1371/journal.pcbi.1000598.

[49] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, “Near-optimal probabilistic RNA-seq quantification,” *Nat. Biotechnol.*, vol. 34, no. 5, Art. no. 5, May 2016, doi: 10.1038/nbt.3519.

[50] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, “Salmon provides fast and bias-aware quantification of transcript expression,” *Nat. Methods*, vol. 14, no. 4, pp. 417–419, Apr. 2017, doi: 10.1038/nmeth.4197.

[51] D. Kim, J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg, “Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype,” *Nat. Biotechnol.*, vol. 37, no. 8, pp. 907–915, Aug. 2019, doi: 10.1038/s41587-019-0201-4.

[52] M. Sémon, D. Mouchiroud, and L. Duret, “Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance,” *Hum. Mol. Genet.*, vol. 14, no. 3, pp. 421–427, Feb. 2005, doi: 10.1093/hmg/ddi038.

[53] J. Liu, M. Frochaux, V. Gardeux, B. Deplancke, and M. Robinson-Rechavi, "Inter-embryo gene expression variability recapitulates the hourglass pattern of evo-devo," *BMC Biol.*, vol. 18, no. 1, p. 129, Sep. 2020, doi: 10.1186/s12915-020-00842-z.

[54] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter, "Near-optimal probabilistic RNA-seq quantification," *Nat. Biotechnol.*, vol. 34, no. 5, pp. 525–527, May 2016, doi: 10.1038/nbt.3519.

[55] L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery, "mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models," *R J.*, vol. 8, no. 1, pp. 289–317, Aug. 2016.

[56] P. Janich, A. B. Arpat, V. Castelo-Szekely, M. Lopes, and D. Gatfield, "Ribosome profiling reveals the rhythmic liver translatome and circadian clock regulation by upstream open reading frames," *Genome Res.*, vol. 25, no. 12, pp. 1848–1859, Dec. 2015, doi: 10.1101/gr.195404.115.

[57] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *J. R. Stat. Soc. Ser. B Methodol.*, vol. 57, no. 1, pp. 289–300, 1995.

[58] P. Melsted, V. Ntranos, and L. Pachter, "The barcode, UMI, set format and BUStools," *Bioinformatics*, vol. 35, no. 21, pp. 4472–4473, Nov. 2019, doi: 10.1093/bioinformatics/btz279.

[59] A. T. L. Lun and J. C. Marioni, "Overcoming confounding plate effects in differential expression analyses of single-cell RNA-seq data," *Biostat. Oxf. Engl.*, vol. 18, no. 3, pp. 451–464, Jul. 2017, doi: 10.1093/biostatistics/kxw055.

[60] H. Mi *et al.*, "PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D394–D403, Jan. 2021, doi: 10.1093/nar/gkaa1106.

[61] H. Mi and P. Thomas, "PANTHER Pathway: an ontology-based pathway database coupled with data analysis tools," *Methods Mol. Biol. Clifton NJ*, vol. 563, pp. 123–140, 2009, doi: 10.1007/978-1-60761-175-2_7.

[62] W. Shen, S. Le, Y. Li, and F. Hu, "SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation," *PLOS ONE*, vol. 11, no. 10, p. e0163962, May 2016, doi: 10.1371/journal.pone.0163962.

[63] A. Komljenovic, J. Roux, J. Wollbrett, M. Robinson-Rechavi, and F. B. Bastian, "BgeeDB, an R package for retrieval of curated expression datasets and for gene list expression localization enrichment tests." F1000Research, Aug. 07, 2018. doi: 10.12688/f1000research.9973.2.

[64] J. Guo *et al.*, “The adult human testis transcriptional cell atlas,” *Cell Res.*, vol. 28, no. 12, pp. 1141–1157, Dec. 2018, doi: 10.1038/s41422-018-0099-2.