Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Recounting the FANTOM CAGE–Associated Transcriptome

Eddie Luidy Imada, Diego F Sanchez, Leonardo Collado-Torres, et al.

Genome Res. published online February 20, 2020
Access the most recent version at doi:10.1101/gr.254656.119

P<P Published online February 20, 2020 in advance of the print journal.

Accepted Manuscript Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.

Open Access Freely available online through the *Genome Research* Open Access option.

Creative Commons License This manuscript is Open Access.This article, published in *Genome Research*, is available under a Creative Commons License (Attribution 4.0 International license), as described at http://creativecommons.org/licenses/by/4.0/.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**

To subscribe to *Genome Research* go to:
https://genome.cshlp.org/subscriptions

Published by Cold Spring Harbor Laboratory Press

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Recounting the FANTOM Cage Associated Transcriptome

Eddie-Luidy Imada[1,2,+], Diego Fernando Sanchez[1,+], Leonardo Collado-Torres[3], Christopher Wilks[4], Tejasvi Matam[1], Wikum Dinalankara[1], Aleksey Stupnikov[1], Francisco Lobo-Pereira[5], Chi-Wai Yip[6], Kayoko Yasuzawa[6], Naoto Kondo[6], Masayoshi Itoh[7], Harukazu Suzuki[6], Takeya Kasukawa[6], Chung-Chau Hon[6], Michiel JL de Hoon[6], Jay W Shin[6], Piero Carninci[6], FANTOM consortium , Andrew E Jaffe[3,8,9], Jeffrey T Leek[9], Alexander Favorov[1,10], Gloria R Franco[2], Ben Langmead[4,9,*], and Luigi Marchionni[1,*,+,a]

[1] Department of Oncology, Johns Hopkins University School of Medicine, Baltimore, MD, USA
[2] Departamento de Bioqu´ımica e Imunologia, ICB, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil
[3] Lieber Institute for Brain Development, Baltimore, MD, USA
[4] Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA
[5] Departamento de Biologia General, ICB, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil [6]RIKEN Center for Integrative Medical Sciences, Yokohama, Japan
[7] RIKEN, Preventive Medicine and Diagnostic Innovation Program, Yokohama, Japan
[8] Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[9] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[10] Laboratory of Systems Biology and Computational Genetics, VIGG RAS, Moscow, RF
[*,+] These authors contributed equally to this work
[a] Correspondence to Luigi Marchionni: marchion@jhu.edu

ABSTRACT

Long non-coding RNAs (lncRNAs) have emerged as key coordinators of biological and cellular processes. Characterizing lncRNA expression across cells and tissues is key to understanding their role in determining phenotypes including human diseases. We present here FC-R2, a comprehensive expression atlas across a broadly defined human transcriptome, inclusive of over 109,000 coding and non-coding genes, as described in the FANTOM CAGE-Associated Transcriptome (FANTOM-CAT) study. This atlas greatly extends the gene annotation used in the original *recount2* resource. We demonstrate the utility of the FC-R2 atlas by reproducing key findings from published large studies and by generating new results across normal and diseased human samples. In particular, we (a) identify tissue specific transcription profiles for distinct classes of coding and non-coding genes, (b) perform differential expression analysis across thirteen cancer types, identifying novel non-coding genes potentially involved in tumor pathogenesis and progression, and (c) confirm the prognostic value for several enhancers lncRNAs expression in cancer. Our resource is instrumental for the systematic molecular characterization of lncRNA by the FANTOM6 Consortium. In conclusion, comprised of over 70,000 samples, the FC-R2 atlas will empower other researchers to investigate functions and biological roles of both known coding genes and novel lncRNAs.

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Introduction

Long non-coding RNAs (lncRNAs) are commonly defined as transcripts longer than 200 nucleotides that are not translated into proteins. This definition is not based on their function, since lncRNAs are involved in distinct molecular processes and biological contexts not yet fully characterized (Batista and Chang 2013). Over the past few years, the importance of lncRNAs has clearly emerged, leading to an increasing focus on decoding the consequences of their modulation, studying their involvement in the regulation of key biological mechanisms during development, normal tissue and cellular homeostasis, and in disease (Batista and Chang 2013; Esteller 2011; Ling et al. 2015).

Given the emerging and previously underestimated importance of non-coding RNAs, the FANTOM consortium has initiated the systematic characterization of their biological function. Through the use of Cap Analysis of Gene Expression sequencing (CAGE-seq), combined with RNA-seq data from the public domain, the FANTOM consortium released a comprehensive atlas of the human transcriptome, encompassing more accurate transcriptional start sites (TSS) for coding and non-coding genes, including numerous novel long non-coding genes: the FANTOM CAGE-Associated Transcriptome (FANTOM-CAT) (Hon et al. 2017). We hypothesized that these lncRNAs can be measured in many RNA-seq datasets from the public domain and that they have been so far missed by the lack of a comprehensive gene annotation.

Although the systematic analysis of lncRNAs function is being addressed by the FANTOM consortium in loss of function studies, increasing the detection rate of these transcripts combining different studies is difficult because the heterogeneity of analytic methods employed. Current resources that apply uniform analytic methods to create expression summaries from public data do exist but can miss several lncRNAs because their dependency on a pre-existing gene annotation for creating the genes expression summaries (Lachmann et al. 2018; Tatlow and Piccolo 2016). We recently created *recount2* (Collado-Torres et al. 2017b), a collection of uniformly-processed human RNA-seq data, wherein we summarized 4.4 trillion reads from over 70,000 human samples from the Sequence Read Archive (SRA), The Cancer Genome Atlas (TCGA) (The Cancer Genome Atlas Research Network et al. 2013), and the Genotype-Tissue Expression (GTEx) (The GTEx Consortium 2013) projects (Collado-Torres et al. 2017b). Importantly, *recount2* provides annotation-agnostic coverage files that allow re-quantification using a new annotation without having to re-process the RNA-seq data.

Given the unique opportunity to access latest results to the most comprehensive human transcriptome (the *FANTOM-CAT* project) and the *recount2* gene agnostic summaries, we addressed the previous described challenges building a comprehensive atlas of coding and non-coding gene expression across the human genome: the *FANTOM-CAT/recount2* expression atlas (FC-R2 hereafter). Our resource contains expression profiles for 109,873 putative genes across over 70,000 samples, enabling an unparalleled resource for the analysis of the human coding and non-coding transcriptome.

Results

Building the *FANTOM-CAT/recount2* resource

The *recount2* resource includes a coverage track, in the form of a BigWig file, for each processed sample. We built the FC-R2 expression atlas by extracting expression levels from *recount2* coverage tracks in regions that overlapped unambiguous exon coordinates for the permissive set of *FANTOM-CAT* transcripts, according to the pipeline shown in Figure 1. Since *recount2*'s coverage tracks do not distinguish from between genomic strands, we removed ambiguous segments that presented overlapping exon annotations from both strands (see Methods section and Supp. Methods). After such disambiguation procedure, the remaining

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

1,066,515 exonic segments mapped back to 109,869 genes in *FANTOM-CAT* (out of the 124,047 starting ones included in the permissive set (Hon et al. 2017)). Overall, the FC-R2 expression atlas encompasses 2,041 studies with 71,045 RNA-seq samples, providing expression information for 22,116 coding genes and 87,763 non-coding genes, such as enhancers, promoters, and others lncRNAs.

Validating the *FANTOM-CAT/recount2* resource

We first assessed how gene expression estimates in FC-R2 compared to previous gene expression estimates from other projects. Specifically, we considered data from the GTEx consortium (v6), spanning 9,662 samples from 551 individuals and 54 tissues types (The GTEx Consortium 2013). First, we computed the correlation for the GTEx data between gene expression based on the FC-R2 atlas and on GENCODE (v25) gene model in *recount2*, which has been already shown to be consistent with gene expression estimates from the GTEx project (Collado-Torres et al. 2017b), observing a median correlation ≥ 0.986 for the 32,922 genes in common. This result supports the notion that our pre-processing steps to disambiguate overlapping exon regions between strands did not significantly alter gene expression quantification.

Next, we assessed whether gene expression specificity, as measured in FC-R2, was maintained across tissue types. To this end, we selected and compared gene expression for known tissue-specific expression patterns, such as keratin 1 (*KRT1*), estrogen receptor 1 (*ESR1*), and neuronal differentiation 1 (*NEUROD1*) (Figure 2). Overall, all analyzed tissue specific markers presented nearly identical expression profiles across GTEx tissue types between the alternative gene models considered (see Figure 2 and S1), confirming the consistency between gene expression quantification in FC-R2 and those based on GENCODE.

We also assessed whether there are genes that are not expressed in any of the normal tissues included in GTEx. Out of 109,869 genes, 681 (0.6%, see Figures S3 and S4), and they were over-represented in the FANTOM-CAT permissive set (chi-squared test, p-value $< 2.2\text{x}10^{-16}$).

Tissue-specific expression of lncRNAs

It has been shown that, although expressed at a lower level, enhancers and promoters are not ubiquitously expressed and are more specific for different cell types than coding genes (Hon et al. 2017). In order to verify this finding, we used GTEx data to assess expression levels and specificity profiles across samples from each of the 54 analyzed tissue types, stratified into four distinct gene categories: coding mRNA, intergenic promoter lncRNA (ip-lncRNA), divergent promoter lncRNA (dp-lncRNA), and enhancers lncRNA (e-lncRNA). Overall, we were able to confirm that these RNA classes are expressed at different levels, and that they display distinct specificity patterns across tissues, as shown for primary cell types by Hon et al. (Hon et al. 2017), albeit with more variability likely due to the increased cellular complexity present in tissues. Specifically, coding mRNAs were expressed at higher levels than lncRNAs (log_2 median expression of 6.6 for coding mRNAs, and of 4.1, 3.8, and 3.1 for ip-lncRNA, dp-lncRNA, and e-lncRNA, respectively). In contrast, the expression of enhancers and intergenic promoters was more tissue-specific (median = 0.41 and 0.30, respectively) than what observed for divergent promoters and coding mRNAs (median = 0.13 and 0.09, respectively) (Figure 3A). Finally, when analyzing the percentage of genes expressed across tissues by category, we observed that coding genes are, in general, more ubiquitous, while lncRNAs are more specific, with enhancers showing the lowest percentages of expressed genes (mean ranging from 88.42% to 41.98%, see Figure 3B), in agreement with the notion that enhancer transcription is tissue specific (Ong and Corces 2011).

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Differential expression analysis of coding and non-coding genes in cancer

We analyzed coding and non-coding genes expression in cancer using TCGA data. To this end, we compared cancer to normal samples separately for 13 tumor types, using FC-R2 re-quantified data. We further identified the differentially expressed genes (DEG) in common across the distinct cancer types (see Figure 4). Overall, the number of DEG varied across cancer types and by gene class, with a higher number of significant coding than non-coding genes (FDR $\leq$ 0.01, see Table 1). Importantly, a substantial fraction of these genes was exclusively annotated in the FANTOM-CAT meta-assembly, suggesting that relying on other gene models would result in missing many potential important genes (see Table 1). We then analyzed differential gene expression consensus across the considered cancer types. A total of 41 coding mRNAs were differentially expressed across all the 13 tumor types after global correction for multiple testing (FDR $\leq 10^{-6}$, see Supplemental table S1). For lncRNAs, a total of 28 divergent promoters, 4 intergenic promoters, and 3 enhancers were consistently up- or down-regulated across all the 13 tumor types after global correction for multiple testing (FDR $\leq$ 0.1, see Supplemental tables S2, S3, S4, respectively).

A usual task performed after differential gene expression analysis is to identify biological processes and pathways associated with the DEGs. To this end, gene set enrichment methods are usually employed, however, this requires detailed gene-to-function annotations, which are mostly lacking for lncRNAs. One possible way to assist prioritizing non-coding transcripts for follow up functional studies is to identify association with other features along the genome. As an example of this type of analysis, we have assessed the overlap between Single-Nucleotide Polymorphism (SNPs) associated with cancer in GWAS studies and the list of DEGs we identified. On average, the percentage of DEGs overlapping cancer SNPs ranged from 6.6% in dp-lncRNA to 10.21% in ip-lncRNA across the 13 cancer types (see Supplemental table S5).

Next, we reviewed the literature to identify functional correlates for these consensus genes. Most of the up-regulated coding genes (Supplemental Table S1) participate in cell cycle regulation, cell division, DNA replication and repair, chromosome segregation, and mitotic spindle checkpoints. Most of the consensus down-regulated mRNAs (Supplemental Table S1) are associated with metabolism and oxidative stress, transcriptional regulation, cell migration and adhesion, and with modulation of DNA damage repair and apoptosis.

Three down-regulated dp-lncRNA genes, *GAS1RR*, *RPL34-DT*, and *RAP2C-AS1*, were reported to be implicated in cancer (Supplemental Table S2). The first one controls epithelial-mesenchymal transition, the second is associated with tumor size increase, while the third is associated with urothelial cancer after kidney cancer transplantation (Zhou et al. 2016; Zhao et al. 2015; Shang et al. 2016). Among the up-regulated dp-lncRNAs (Supplemental Table S2), *SNHG1* has been implicated in cellular proliferation, migration and invasion of different cancer types, and to be strongly up-regulated in osteosarcoma, non-small lung cancer, and gastric cancer (Cao et al. 2013; Sun et al. 2017).

Among the ip-lncRNAs ubiquitously down-regulated (see Supplemental Table S3), *MIR99AHG* has been identified in many different tumor types including leukemia, breast, vulvar, prostate, and bladder cancer (Emmrich et al. 2014; Ni et al. 2016; Gökmen-Polar et al. 2016; Sun et al. 2014; Li et al. 2017). For instance, in vulvar squamous cell carcinoma, *MIR99AHG* and *MIR31HG* expressions are correlated and associated with tumor differentiation (Ni et al. 2016). Similarly, *MIR99AHG* down-regulation in ER positive breast cancer is associated with progression, recurrence, and metastasis (Gökmen-Polar et al. 2016). In contrast, increased expression of *SNHG17* (an ip-lncRNA, see Supplemental Table S3), was associated with short term survival in breast cancer, and with tumor size, stage, and lymph node metastasis in colorectal cancer (Zhao et al. 2014; Ma et al. 2017). In addition, *LINC01311*, another ip-lncRNA (Supplemental Table S3), was found to be up-regulated in liver cancer and metastatic prostate cancer (Zhu et al.

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

2016). Despite we did not identify any cancer association for common e-lncRNAs, one among those we identified, *LINC02884*, has been previously reported to be up-regulated in late-onset Alzheimer's disease (Humphries et al. 2015). Furthermore, the enhancers lncRNA class also yielded the lowest number of genes in common among all cancer types, reinforcing the concept that enhancers are expressed in a tissue specific manner (See Figure 3A and Supplemental Table S4).

Finally, we focused more in depth on prostate cancer (PCa) as a prototypical example, and we were able to confirm previous findings for both coding and non-coding genes (see Supplemental Figure S2). For coding genes, we confirmed differential expression for known markers of PCa progression and mortality, like *ERG*, *FOXA1*, *RNASEL*, *ARVCF*, and *SLC43A1* (Lin et al. 2011; Yu et al. 2010). Similarly, we also confirmed differential expression for non-coding genes, like *PCA3*, the first clinically approved lncRNA marker for PCa (Bussemakers et al. 1999; de Kok et al. 2002), *PCAT1*, a prostate-specific lncRNA involved in disease progression (Prensner et al. 2011), *MALAT1*, which is associated with PCa poor prognosis (Ren et al. 2013), *CDKN2B-AS1*, an anti-sense lncRNA up-regulated in PCa that inhibits tumor suppressor genes activity (Gutschner and Diederichs 2012; Kotake et al. 2011), and the *MIR135* host gene, which is associated with castration-resistant PCa (Huang et al. 2015).

Confirming prognostic enhancers

Chen and collaborators have recently surveyed enhancers expression in nearly 9,000 patients from TCGA (Chen et al. 2018), using genomic coordinates from the FANTOM5 project (Andersson et al. 2014), identifying 4,803 expressed genomic regions with prognostic potential in one or more TCGA tumor types. We therefore leveraged the FC-R2 atlas to identify prognostic coding and non-coding genes using both Univariate and Multivariate Cox proportional hazard models, comparing our results for e-lncRNAs with those reported by Chen and colleagues. To this end, we started by comparing gene annotations and genomic overlap between the studies. This was necessary because Chen and collaborators relied on the enhancer regions reported by Andersson et al. (Andersson et al.
2014), which is based on the observation of bi-directional transcription. Our resource, on the contrary, relies on the latest updated FANTOM-CAT annotation, which takes into account other features, such as the epigenetic context, when defining RNA categories. Out of the 4,803 genomic regions found prognostic by Chen and collaborators (Chen et al. 2018), we could unambiguously map 1218 regions to exons annotated in the FANTOM-CAT gene models for the four RNA categories we considered in our study (corresponding to a total of 1,046 unique genes). Overall, despite the mentioned differences in annotation and quantification (see Supplemental Table S6), we were still able to confirm prognostic value for 466 genes out of the 1,046 reported by Chen et al (Chen et al. 2018), including *"enhancer 22"* (*KLHDC7B-DT*), which was highlighted as a promising prognostic marker for kidney cancer (Supplemental Figure S5).

We then considered the FANTOM-CAT RNA classes across the different tumor types. We were able to identify a variable number of genes significantly associated with overall survival (FDR ≤ 0.05) in univariate Cox proportional hazards models (see Supplemental Tables S7-S10). Among the consensus DEGs identified across all tumor types, 40 out of 41 coding mRNAs, 25 out of 28 dp-lncRNAs, 4 out of 4 ip-lncRNAs, and 2 out of 3 e-lncRNAs were found to be associated with survival (see Supplemental Tables S11-S14). Kaplan-Meier curves for selected differentially expressed genes for each RNA category are shown in Supplemental Figure S6. Finally, we performed multivariable analysis controlling for relevant clinical and pathological

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

characteristics in each tumor type. Overall, despite a number of genes being associated with such variables, we obtained similar results (see Supplemental Tables S15-S22).

Discussion

The importance of lncRNAs in cell biology and disease has clearly emerged in the past few years and different classes of lncRNAs have been shown to play crucial roles in cell regulation and homeostasis (Quinn and Chang 2016). For instance, enhancers – a major category of gene regulatory elements, which has been shown to be expressed (Arner et al. 2015; Andersson et al. 2014) – play a prominent role in oncogenic processes (Sur and Taipale 2016; Herz et al. 2014) and other human diseases (Hnisz et al. 2013). Despite their importance, however, there is a scarcity of large-scale datasets investigating enhancers and other lncRNA categories, in part due to the technical difficulty in applying high-throughput techniques such as ChIP-seq and Hi-C over large cohorts, and to the use of gene models that do not account for them in transcriptomics analyses. Furthermore, the large majority of the lncRNAs that are already known – and that have been shown to be associated with some phenotype – are still lacking functional annotation.

To address these needs, the FANTOM consortium has first constructed the FANTOM-CAT meta-transcriptome, a comprehensive atlas of coding and non-coding genes with robust support from CAGE-seq data (Hon et al. 2017), then it has undertaken a large scale project to systematically target lncRNAs and characterize their function using a multi-pronged approach (Ramilowski et al. 2019). In a complementary effort, we have leveraged public domain gene expression data from *recount2* (Collado-Torres et al. 2017a,b) to create a comprehensive gene expression compendium across human cells and tissues based on the *FANTOM-CAT* gene model, with the ultimate goal of facilitating lncRNAs annotation through association studies. To this end, the FC-R2 atlas is already in use in the FANTOM6 project (http://fantom.gsc.riken.jp/6/) to successfully characterize lncRNA expression in human samples (Ramilowski et al. 2019).

In order to validate our resource, we have compared the gene expression summaries based on FANTOM-CAT gene models with previous, well-established gene expression quantifications, demonstrating virtually identical profiles across tissue types overall and for specific tissue markers. We have then confirmed that distinct classes of coding and non-coding genes differ in terms of overall expression level and specificity pattern across cell types and tissues. We also have observed a small subset of genes that were not expressed in the large majority of the samples analyzed in GTEx project. These genes were mostly classified as small RNAs and enhancers, which was expected, given that the RNA-seq libraries included in *recount2* did not target small RNAs, and enhancers are usually expressed at a lower level. We further reveal that this subset of genes not expressed in any normal tissue, is also associated with a lower level of support of the corresponding FANTOM-CAT gene models (Hon et al. 2017).

Furthermore, using the FC-R2 atlas, we were also able to identify mRNAs, promoters, enhancers, and other lncRNAs that are differentially expressed in cancer, both confirming previously reported findings, and identifying novel cancer genes exclusively annotated in the FANTOM-CAT gene models, which have been therefore missed in prior analyses with TCGA data. Finally, we confirmed the prognostic value for some of the enhancer regions recently reported by Chen and colleagues in the TCGA (Chen et al. 2018), by performing a systematic screening for survival association of both coding and non-coding genes that are quantifiable in the FC-R2 resource. Overall, we identified several genes with potential prognostic value across the analyzed cancer types in TCGA, however, further corroboratory studies in independent patient cohorts are necessary to validate these associations.

Collectively, by confirming findings reported in previous studies, our results demonstrate that the FC-R2 gene expression atlas is a reliable and powerful resource for exploring both the coding and non-coding transcriptome, providing compelling evidence and robust support to the notion that lncRNA gene classes, including enhancers and promoters, despite not being yet fully understood,

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

portend significant biological functions. Our resource, therefore, constitutes a suitable and promising platform for future large scales studies in cancer and other human diseases, which in turn hold the potential to reveal important cues to the understanding of their biological, physiological, and pathological roles, potentially leading to improved diagnostic and therapeutic interventions.

Finally, all results, data, and code from the FC-R2 atlas are available as a public tool. With uniformly processed expression data for over 70,000 samples and 109,873 genes ready to analyze, we want to encourage researchers to dive deeper into the study of ncRNAs, their interaction with coding and non-coding genes, and their influence on normal and disease tissues. We hope this new resource will help paving the way to develop new hypotheses that can be followed to unwind the biological role of the transcriptome as a whole.

Methods

Data and pre-processing.

The complete FANTOM-CAT gene catalog (inclusive of robust, intermediate, and permissive sets) was obtained from the FANTOM consortium within the frame of the FANTOM6 project (Ramilowski et al. 2019). The genes were annotated using official HGNC symbols (https://www.genenames.org) when available. For genes without HGNC symbols, we named them according to HGNC instructions (see Supplemental Table S23). The remaining genes were referred using the official ID from the consortium that annotated the gene (ENSEMBL/FANTOM). This catalog accounts for 124,245 genes supported by CAGE peaks and it includes those described by Hon et al. (Hon et al. 2017). In order to remove ambiguity due to overlapping among exons from distinct genes, the BED files containing the coordinates for all genes and exons were processed with the *GenomicRanges* R/Bioconductor package (Lawrence et al. 2013) to obtain exon coordinates disjoining. To avoid losing strand information from annotation we processed data using a two-step approach by first disjoining overlapping segments on the same strand and then across strands (Figure 5). The genomic ranges (disjoint exons segments) that mapped back to more than one gene were discarded. The expression values for these ranges were then quantified using *recount.bwtool* (Ellis et al. 2018) (code at https://github.com/LieberInstitute/marchionni_projects). The resulting expression quantification were processed to generate RangedSummarizedExperiment objects compatible with the *recount2* framework (Collado-Torres et al. 2017b,a) (code available from https://github.com/eddieimada/fcr2). Thus, the FC-R2 atlas provides expression information for coding and non-coding genes (including enhancers, divergent promoters, and intergenic lncRNAs) for 9,662 samples from the Genotype-Tissue Expression (GTEx) project, 11,350 samples from TCGA, and over 50,000 samples from the Sequence Read Archive (SRA).

Correlation with other studies.

To test if the pre-processing steps used for FC-R2 had a major impact on gene expression quantification, we compared our data to the published GTEx expression values obtained from *recount2* (version 2, https://jhubiostatistics.shinyapps.io/recount/). Specifically, we first compared the expression distribution of tissue specific genes across different tissue types and then computed the Pearson correlation for each gene in common across the original *recount2* gene expression estimates based on GENCODE and our version based on the FANTOM-CAT transcriptome.

Expression specificity of tissue facets.

We analyzed the expression level and specificity of each gene stratified by RNA category (i.e. mRNA, e-lncRNA, dp-lncRNA, ip-lncRNA) using the same approach described by Hon et al. (Hon et al. 2017) (see Supplemental Methods). Briefly, overall expression

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

levels for each gene were represented by the maximum transcript per million (TPM) values observed across all samples within each tissue type in GTEx. Gene specificity was based on the empirical entropy computed using the mean expression value across tissue types. The 99.99% confidence intervals for the expression of each category by tissue type were calculated based on TPM values. Genes with a TPM greater than 0.01 were considered to be expressed.

Identification of differentially expressed genes.

We analyzed differential gene expression in 13 cancer types, comparing primary tumor with normal samples using TCGA data from the FC-R2 atlas. Gene expression summaries for each cancer type were split by RNA category (coding mRNA, intergenic promoter lncRNA, divergent promoter lncRNA, and enhancer lncRNA) and then analyzed independently. A generalized linear model approach, coupled with empirical Bayes moderation of standard errors (Smyth 2004), was used to identify differentially expressed genes between groups. The model was adjusted for the three most relevant coefficients for data heterogeneity as estimated by surrogate variable analysis (SVA) (Leek and Storey 2007). Correction for multiple testing was performed across RNA category by merging the resulting p-values for each cancer type and applying the Benjamini-Hochberg method (Benjamini and Hochberg 1995). Overlapping between DEG and GWAS SNPs was performed using the FANTOM-CAT gene regions coordinates and the SNPs positions obtained from the GWAS Catalog (Buniello et al. 2018).

Prognostic analysis.

To evaluate the prognostic potential of the genes in FC-R2, we performed both multivariate and univariate Cox proportional hazards regression analysis separately for each RNA class (22106 mRNAs, 17,404 e-lncRNAs, 6,204 dp-lncRNAs, and 1,948 ip-lncRNAs) across each of the 13 TCGA cancer types with available survival follow-up information (see Supplemental Methods and Supplemental Table S24). Genes with FDR equal or less than 0.05, using Benjamini-Hochberg correction (Benjamini and Hochberg 1995) within each cancer type and RNA class, were deemed significant prognostic factors. We further analyzed the prognostic value of the consensus differentially expressed genes we identified comparing tumors to normal samples by intersecting the corresponding gene lists with those obtained by Cox proportional regression. Finally, in order to compare our results to previous prognostic analyses, we obtained data on enhancers position and prognostic potential from Chen et al. (Chen et al. 2018), performed a liftover to the hg38 genome assembly to match FC-R2 coordinates, and assessed the overlap between prognostic genes identified in the two studies.

Data and code availability.

All data is available from http://marchionnilab.org/fcr2.html. Expression data can be directly accessed through https://jhubiostatistics.shinyapps.io/recount/ and the *recount* Bioconductor package (v1.9.5 or newer) at https://bioconductor.org/packages/recount as *RangedSummarizedExperiment* objects organized by the Sequence Read Archive (SRA) study ID. The data can be loaded using R-programming language and is ready to be analyzed using Bioconductor packages or the data can be exported to other formats for use in another environment.

All code used in this manuscript is available for reproducibility and transparency at: https://github.com/eddieimada/fcr2 and https://github.com/LieberInstitute/marchionni_projects. A compressed archive with all scripts used is also distributed as Supplemental material (see Supplemental Codes).

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Acknowledgements

This publication was made possible through support from the NIH-NCI grants P30CA006973 (L.M. and A.F.), 1U01CA231776 (W.D. and L.M.), U01CA196390 (L.M. and A.S.), and R01CA200859 (W.D. and L.M.), and the NIH-NIGMS grant R01GM118568 (C.W. and B.L.), R21MH109956-01 (L.C.T. and A.E.J.), and the Department of Defense (DoD) office of the Congressionally Directed Medical Research Programs (CDMRP) award W81XWH-16-1-0739 (E.L.I. and L.M.), Russian Academic project 0112-2019-0001 and Russian Foundation for Basic Research project 17-00-00208 (A.F.), Fundação de Amparo a Pesquisa do Estado de Minas Gerais award BDS-00493-16 (E.L.I and G.R.F.). *recount2* and FC-R2 are hosted on SciServer, a collaborative research environment for large-scale data-driven science. It is being developed at, and administered by, the Institute for Data Intensive Engineering and Science (IDIES) at Johns Hopkins University. SciServer is funded by the National Science Foundation Award ACI-1261715. For more information about SciServer, visit http://www.sciserver.org/.

Author contributions statement

L.M. conceived the idea, L.M., E.L.I., A.F. and B.L. designed the study; E.L.I., D.F.S., T.M., W.D., A.S., L.C.T., and L.M. performed the analysis; E.L.I., D.F.S., F.P.L., G.R.F. and L.M. interpreted the results; L.C.T., C.W., C.Y., K.Y, N.K., M.I., H.S., T.K., C.C.H., M.H., J.W.S., P.C. A.E.J., J.T.L. and B.L. provided the data and tools; E.L.I., D.F.S., L.C.T., B.L. and L.M. wrote the manuscript; All authors reviewed and approved the manuscript.

Disclosure declaration

All authors declare no conflicts of interest.

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

References

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* **507**: 455–461.

Arner E, Daub CO, Vitting-Seerup K, Andersson R, Lilje B, Drabløs F, Lennartsson A, Ronnerblad M, Hrydziuszko O, Vitezic M, et al. 2015. Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science* **347**: 1010–1014.

Batista PJ, Chang HY. 2013. Long noncoding RNAs: cellular address codes in development and disease. *Cell* **152**: 1298–1307.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc* Series **B, 57**: 289–300.

Buja A, Eyuboglu N. 1992. Remarks on parallel analysis. *Multivar Behav Res* **27**: 509–540.

Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, et al. 2018. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**: D1005–D1012.

Bussemakers MJ, van Bokhoven A, Verhaegh GW, Smit FP, Karthaus HF, Schalken JA, Debruyne FM, Ru N, Isaacs WB. 1999. DD3: a new prostate-specific gene, highly overexpressed in prostate cancer. *Cancer Res* **59**: 5975–5979.

Cao WJ, Wu HL, He BS, Zhang YS, Zhang ZY. 2013. Analysis of long non-coding RNA expression profiles in gastric cancer. *World J Gastroentero* **19**: 3658–3664.

Chen H, Li C, Peng X, Zhou Z, Weinstein JN, Cancer Genome Atlas Research Network, Liang H. 2018. A Pan-Cancer Analysis of Enhancer Expression in Nearly 9000 Patient Samples. *Cell* **173**: 386–399.e12.

Collado-Torres L, Nellore A, Jaffe AE. 2017a. recount workflow: Accessing over 70,000 human RNA-seq samples with Bioconductor. *F1000Research* **6**: 1558.

Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, Jaffe AE, Langmead B, Leek JT. 2017b. Reproducible RNA-seq analysis using recount2. *Nat Biotechno* **35**: 319–321.

de Kok JB, Verhaegh GW, Roelofs RW, Hessels D, Kiemeney LA, Aalders TW, Swinkels DW, Schalken JA. 2002. DD3 (PCA3), a very sensitive and specific marker to detect prostate tumors. *Cancer Res* **62**: 2695–2698.

Ellis SE, Collado-Torres L, Jaffe A, Leek JT. 2018. Improving the value of public RNA-seq expression data by phenotype prediction. *Nucleic Acids Res* **46**: e54–e54.

Emmrich S, Streltsov A, Schmidt F, Thangapandi VR, Reinhardt D, Klusmann JH. 2014. LincRNAs MONC and MIR100HG act as oncogenes in acute megakaryoblastic leukemia. *Mol cancer* **13**: 171.

Esteller M. 2011. Non-coding RNAs in human disease. *Nat Rev Genet* **12**: 861–874.

Gokmen-Polar Y, Zavodszky M, Chen X, Gu X, Kodira C, Badve S. 2016. Abstract P2-06-05: LINC00478: A novel tumor suppressor in breast cancer. *Cancer Research* **76**: P2–06–05.

Gutschner T, Diederichs S. 2012. The hallmarks of cancer: a long non-coding RNA point of view. *RNA Biol* **9**: 703–719.

Herz HM, Hu D, Shilatifard A. 2014. Enhancer malfunction in cancer. *Mol cell* **53**: 859–866.

Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-Andre V, Sigova AA, Hoke HA, Young RA. 2013. Super-enhancers in the control of cell identity and disease. *Cell* **155**: 934–947.

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Hon CC, Ramilowski JA, Harshbarger J, Bertin N, Rackham OJL, Gough J, Denisenko E, Schmeier S, Poulsen TM, Severin J, et al. 2017. An atlas of human long non-coding RNAs with accurate 5' ends. *Nature* **543**: 199–204.

Huang X, Yuan T, Liang M, Du M, Xia S, Dittmar R, Wang D, See W, Costello BA, Quevedo F, et al. 2015. Exosomal miR-1290 and miR-375 as prognostic markers in castration-resistant prostate cancer. *Eur Urol* **67**: 33–41.

Humphries CE, Kohli MA, Nathanson L, Whitehead P, Beecham G, Martin E, Mash DC, Pericak-Vance MA, Gilbert J. 2015. Integrated whole transcriptome and DNA methylation analysis identifies gene networks specific to late-onset Alzheimer's disease. *J Alzheimers Dis* **44**: 977–987.

Kassambara A, Kosinski M. 2018. *survminer: Drawing Survival Curves using 'ggplot2'*. R package version 0.4.3.

Kotake Y, Nakagawa T, Kitagawa K, Suzuki S, Liu N, Kitagawa M, Xiong Y. 2011. Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15(INK4B) tumor suppressor gene. *Oncogene* **30**: 1956–1962.

Lachmann A, Torre D, Keenan AB, Jagodnik KM, Lee HJ, Wang L, Silverstein MC, Ma'ayan A. 2018. Massive mining of publicly available RNA-seq data from human and mouse. *Nat Commun* **9**: 1366.

Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9**: e1003118.

Leek JT, Storey JD. 2007. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics* **3**: 1724–1735.

Li S, Li B, Zheng Y, Li M, Shi L, Pu X. 2017. Exploring functions of long noncoding RNAs across multiple cancers through co-expression network. *Sci Rep* **7**: 754.

Lin DW, FitzGerald LM, Fu R, Kwon EM, Zheng SL, Kolb S, Wiklund F, Stattin P, Isaacs WB, Xu J, et al. 2011. Genetic variants in the LEPR, CRY1, RNASEL, IL4, and ARVCF genes are prognostic markers of prostate cancer-specific mortality. *Cancer Epidem Biomar* **20**: 1928–1936.

Ling H, Vincent K, Pichler M, Fodde R, Berindan-Neagoe I, Slack FJ, Calin GA. 2015. Junk DNA and the long non-coding RNA twist in cancer genetics. *Oncogene* **34**: 5003–5011.

Ma Z, Gu S, Song M, Yan C, Hui B, Ji H, Wang J, Zhang J, Wang K, Zhao Q. 2017. Long non-coding RNA SNHG17 is an unfavorable prognostic factor and promotes cell proliferation by epigenetically silencing P57 in colorectal cancer. *Mol BioSystems* **13**: 2350–2361.

Ni S, Zhao X, Ouyang L. 2016. Long non-coding RNA expression profile in vulvar squamous cell carcinoma and its clinical significance.

Oncol Rep **36**: 2571–2578.

Ong CT, Corces VG. 2011. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat Rev Genet* **12**: 283–293.

Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, Laxman B, Asangani IA, Grasso CS, Kominsky HD, et al. 2011. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat Biotechno* **29**: 742–749.

Quinn JJ, Chang HY. 2016. Unique features of long non-coding RNA biogenesis and function. *Nat Rev Genet* **17**: 47–62.

Ramilowski J, Yip CW, Agrawal S, Chang JC, Ciani Y, Kulakovskiy IV, Mendez M, Ooi JLC, Ouyang JF, Parkinson N, et al. 2019. Functional Annotation of Human Long Non-Coding RNAs via Molecular Phenotyping. *bioRxiv*: 700864.

Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Ren S, Liu Y, Xu W, Sun Y, Lu J, Wang F, Wei M, Shen J, Hou J, Gao X, et al. 2013. Long noncoding RNA MALAT-1 is a new potential therapeutic target for castration resistant prostate cancer. *J Urology* **190**: 2278–2287.

Shang D, Zheng T, Zhang J, Tian Y, Liu Y. 2016. Profiling of mRNA and long non-coding RNA of urothelial cancer in recipients after renal transplantation. *Tumor Biol* **37**: 12673–12684.

Smyth GK. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol* **3**: Article3.

Sun D, Layer R, Mueller AC, Cichewicz MA, Negishi M, Paschal BM, Dutta A. 2014. Regulation of several androgen-induced genes through the repression of the miR-99a/let-7c/miR-125b-2 miRNA cluster in prostate cancer cells. *Oncogene* **33**: 1448–1457.

Sun Y, Wei G, Luo H, Wu W, Skogerbø G, Luo J, Chen R. 2017. The long noncoding RNA SNHG1 promotes tumor growth through regulating transcription of both local and distal genes. *Oncogene* **36**: 6774–6783.

Sur I, Taipale J. 2016. The role of enhancers in cancer. *Nat Rev Cancer* **16**: 483–493.

Tatlow PJ, Piccolo SR. 2016. A cloud-based workflow to quantify transcript-expression levels in public cancer compendia. *Sci Rep* **6**: 39259.

The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**: 1113–1120.

The GTEx Consortium. 2013. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* **45**: 580–585.

Yu J, Yu J, Mani RS, Cao Q, Brenner CJ, Cao X, Wang X, Wu L, Li J, Hu M, et al. 2010. An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. *Cancer cell* **17**: 443–454.

Zhao J, Liu Y, Zhang W, Zhou Z, Wu J, Cui P, Zhang Y, Huang G. 2015. Long non-coding RNA Linc00152 is involved in cell cycle arrest, apoptosis, epithelial to mesenchymal transition, cell migration and invasion in gastric cancer. *Cell cycle* **14**: 3112-3123.

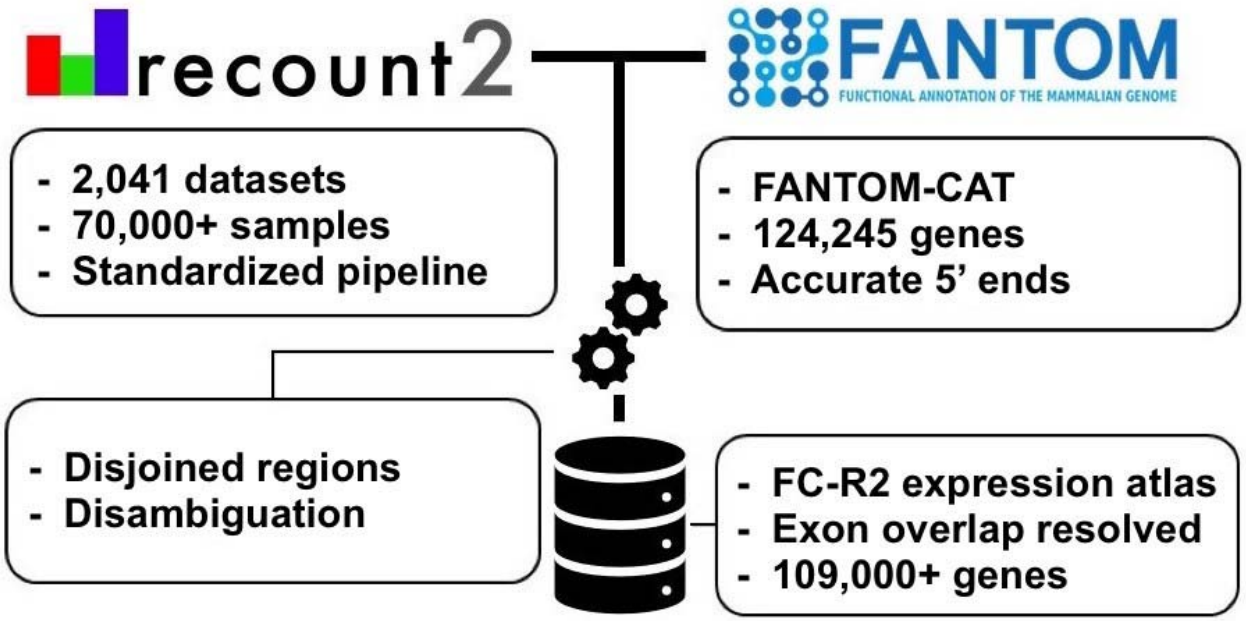
Zhao W, Luo J, Jiao S. 2014. Comprehensive characterization of cancer subtype associated long non-coding RNAs and their clinical implications. *Sci Rep* **4**: 6591.

Zhou M, Hou Y, Yang G, Zhang H, Tu G, Du Ye, Wen S, Xu L, Tang X, Tang S, et al. 2016. LncRNA-Hh strengthen cancer stem cells generation in twist-positive breast cancer via activation of hedgehog signaling pathway. *Stem cells* **34**: 55–66.

Zhu S, Li W, Liu J, Chen CH, Liao Q, Xu P, Xu H, Xiao T, Cao Z, Peng J, et al. 2016. Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR–Cas9 library. *Nat Biotechnol* **34**: 1279–1286.

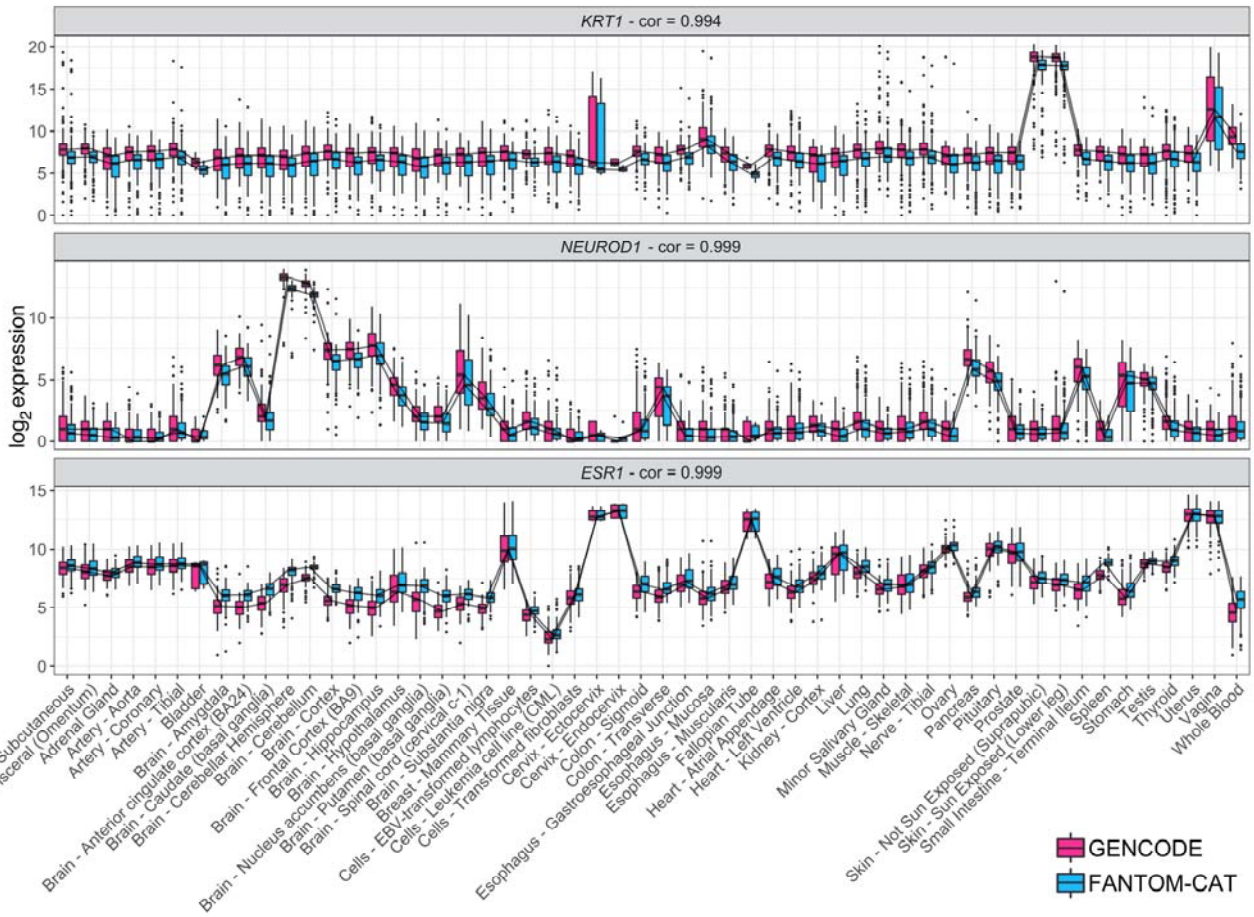
Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Figure 1



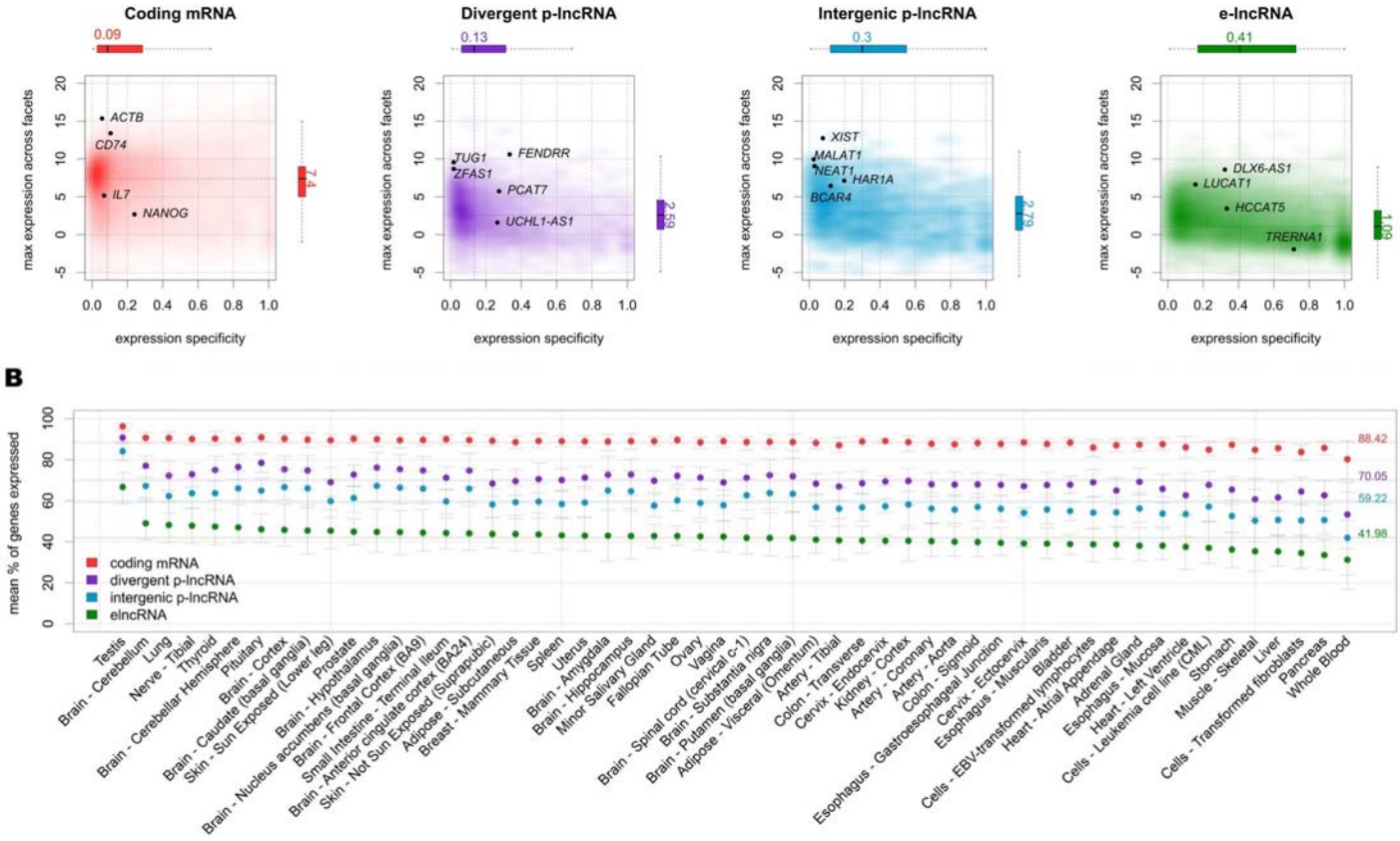
Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Figure 2



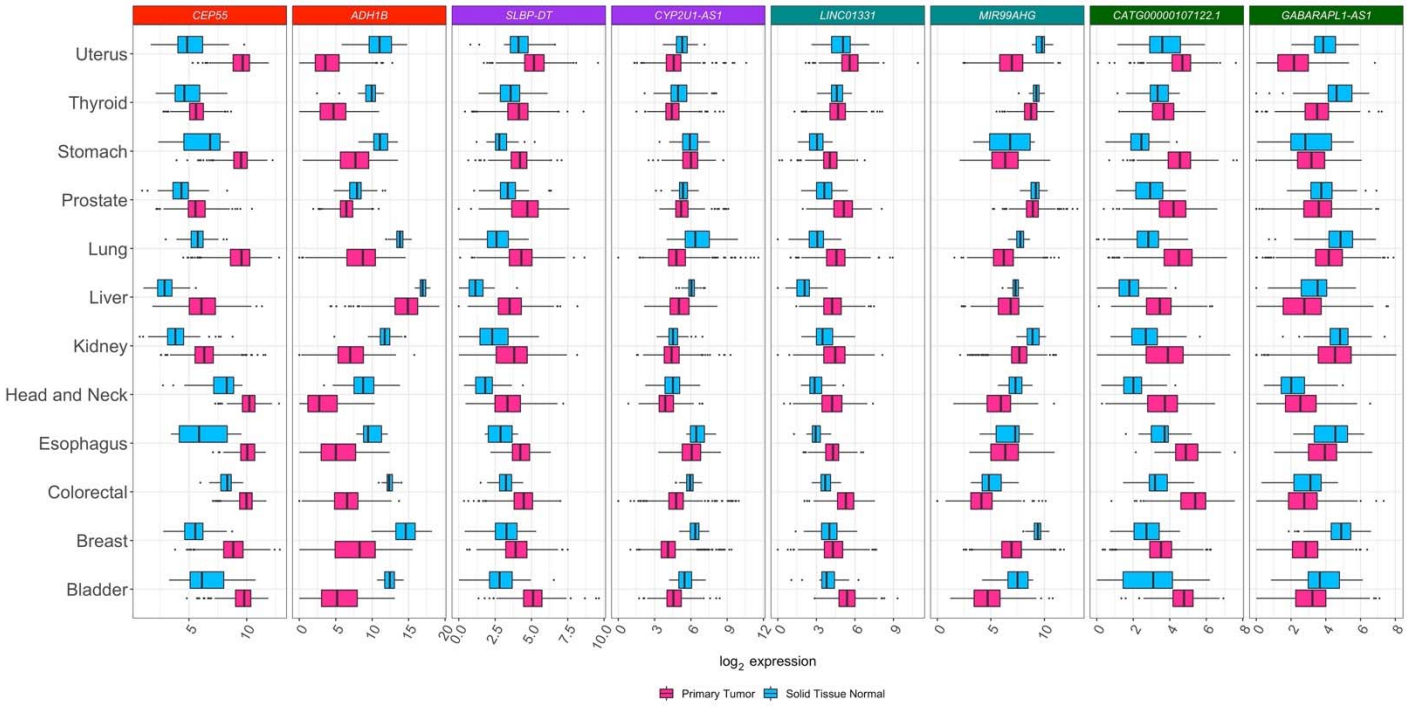
Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Figure 3



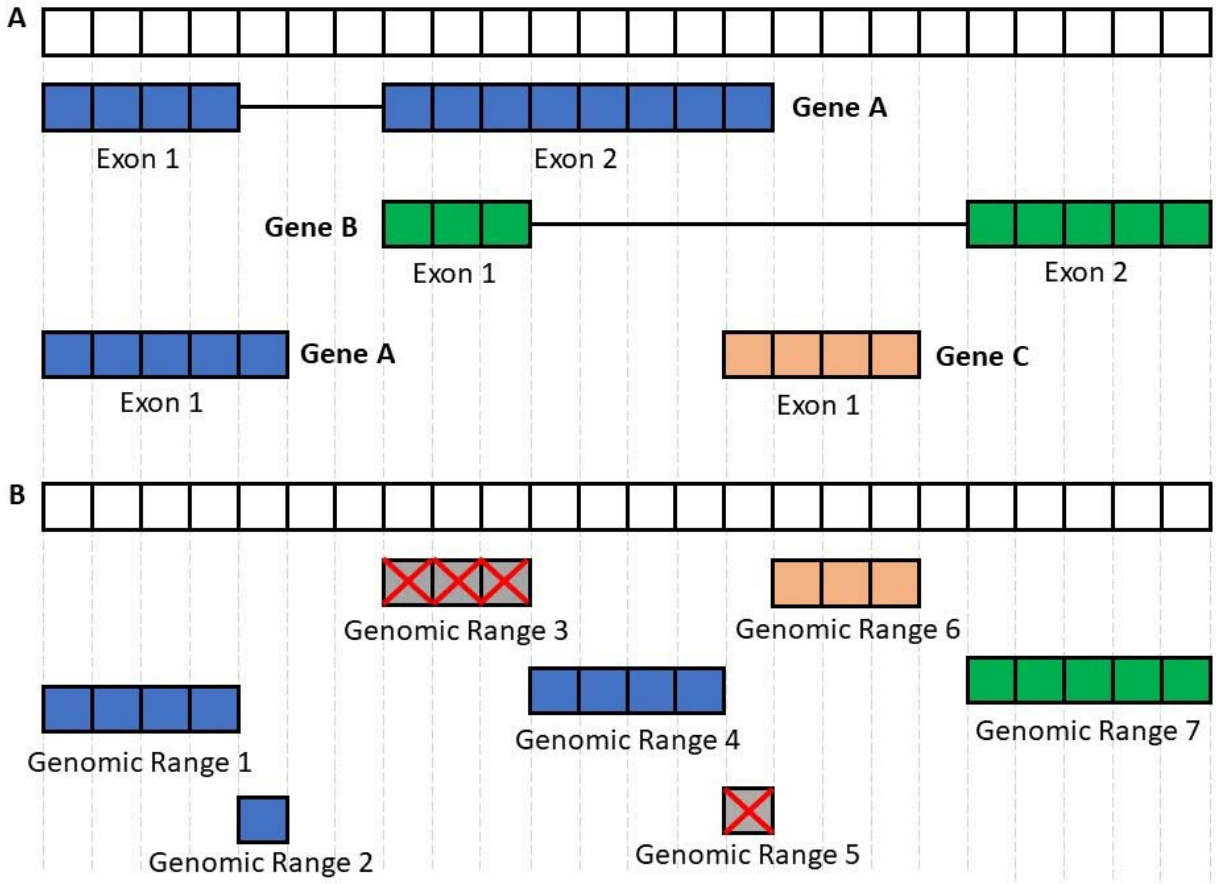
Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Figure 4



Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Figure 5



Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Figure Legends

Figure 1. Overview of the *FANTOM-CAT/recount2* resource development. FC-R2 leverages two public resources, the *FANTOM-CAT* gene models and *recount2*. FC-R2 provides expression information for 109,873 genes, both coding (22,110) and non-coding (87,693). This latter group encompasses enhancers, promoters, and others lncRNAs.

Figure 2. Tissue specific expression in GTEx. Log_2 expression for three tissue specific genes (*KRT1*, *NEUROD1*, and *ESR1*) in GTEx data stratify by tissue type using FC-R2 and GENCODE based quantification. Expression profiles are highly correlated and expressed consistently in the expected tissue types (*e.g.*, *KRT1* is most expressed in skin, *NEUROD1* in brain, and *ESR1* in estrogen sensitive tissue types like uterus, Fallopian tubes, and breast). Correlations are shown on top for each tissue marker. Center lines, upper/lower quartiles and whiskers represents the median, 25/75 percentiles and 1.5 interquartile range, respectively. Additional tissue specific markers are shown in Supplemental Figure S1.

Figure 3. Expression profiles across GTEx tissues. A) Expression level and tissue specificity across four distinct RNA categories. The Y-axis shows log_2 expression levels representing each gene using its maximum expression in GTEx tissues expressed as transcripts per million (TPM). The X-axis shows expression specificity based on entropy computed from median expression of each gene across the GTEx tissue types. Individual genes are highlighted in the figure panels. B) Percentage of genes expressed for each RNA category stratified by GTEx tissue facets. The dots represent the mean among samples within a facet and the error bars represent 99.99% confidence intervals. Dashed lines represent the means among all samples.

Figure 4. Differential expression for selected transcripts from distinct RNA classes across tumor types. Boxplots for selected differential expressed genes between tumor and normal samples across all 13 tumor types analyzed. For each tissue of origin, the most up-regulated (on the left) and down-regulated (on the right) gene for each RNA class is shown. Center lines, upper/lower hinges, and the whiskers respectively represent the median, the upper and lower quartiles, and 1.5 extensions of the interquartile range. Color coding on top of the figure indicates the RNA classes (red for mRNA, purple for dp-lncRNA, cyan ip-lncRNA, and green for e-lncRNA. These genes were select after global multiple testing correction across all 13 tumor types (see Supplemental Tables S1, S2, S3, and S4)

Figure 5. Processing the FANTOM-CAT genomic ranges. This figure summarizes the disjoining and exon disambiguation processes performed before extracting expression information from *recount2* using the FANTOM-CAT gene models. A) Representation of a genomic segment containing 3 distinct, hypothetical genes: gene A having two isoforms, and genes B and C with one isoform each. Each box can be interpreted as one nucleotide along the genome. Colors indicate the 3 different genes. B) Representation of disjoint exon ranges from example in Panel A. Each feature is reduced to a set of non-overlapping genomic ranges. The disjoint genomic ranges mapping back to two or more distinct genes are removed (crossed grey boxes). After removal of ambiguous ranges, the expression information for remaining ones is extracted from *recount2* and summarized at the gene level.

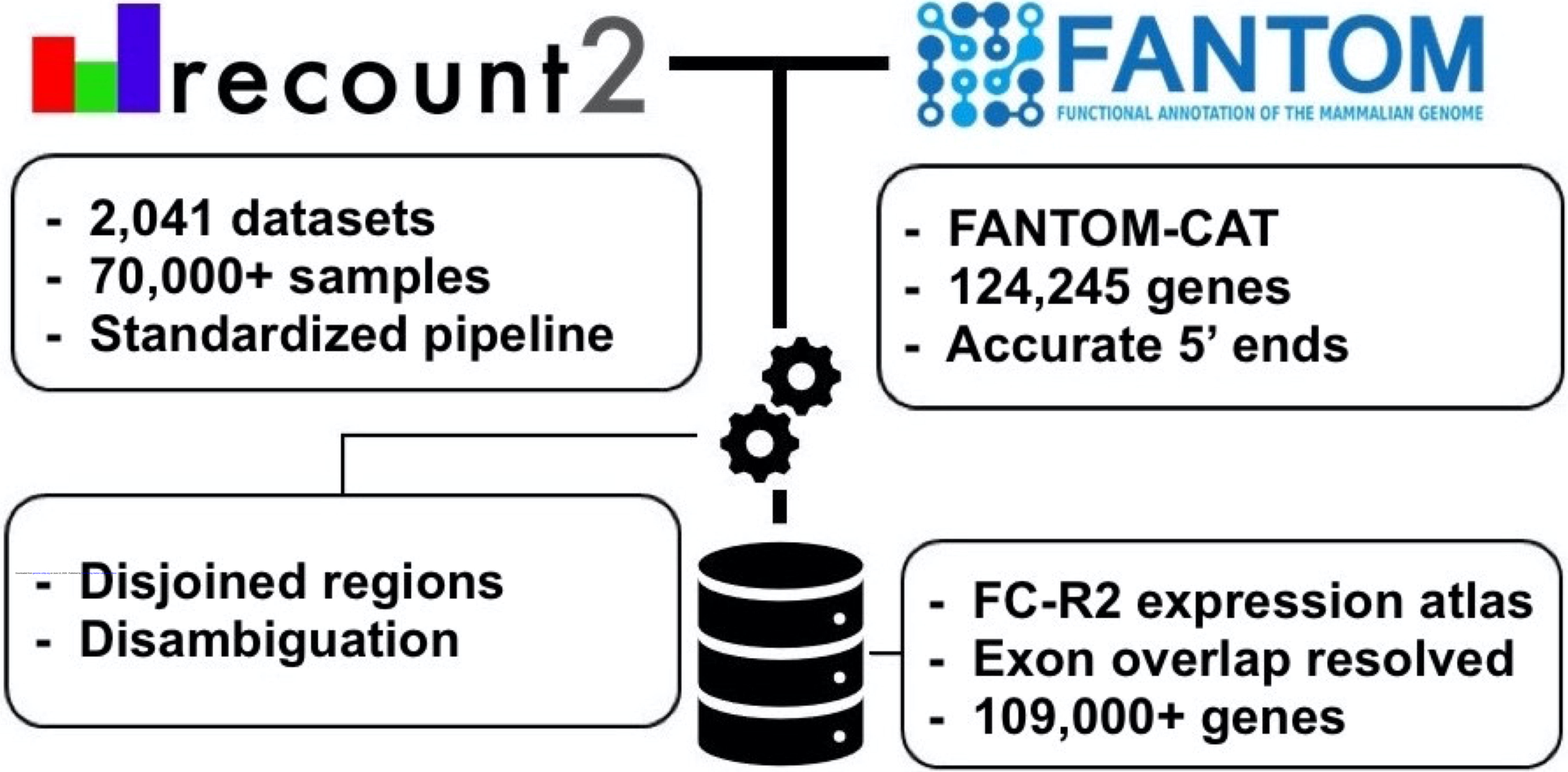
Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press

Table 1

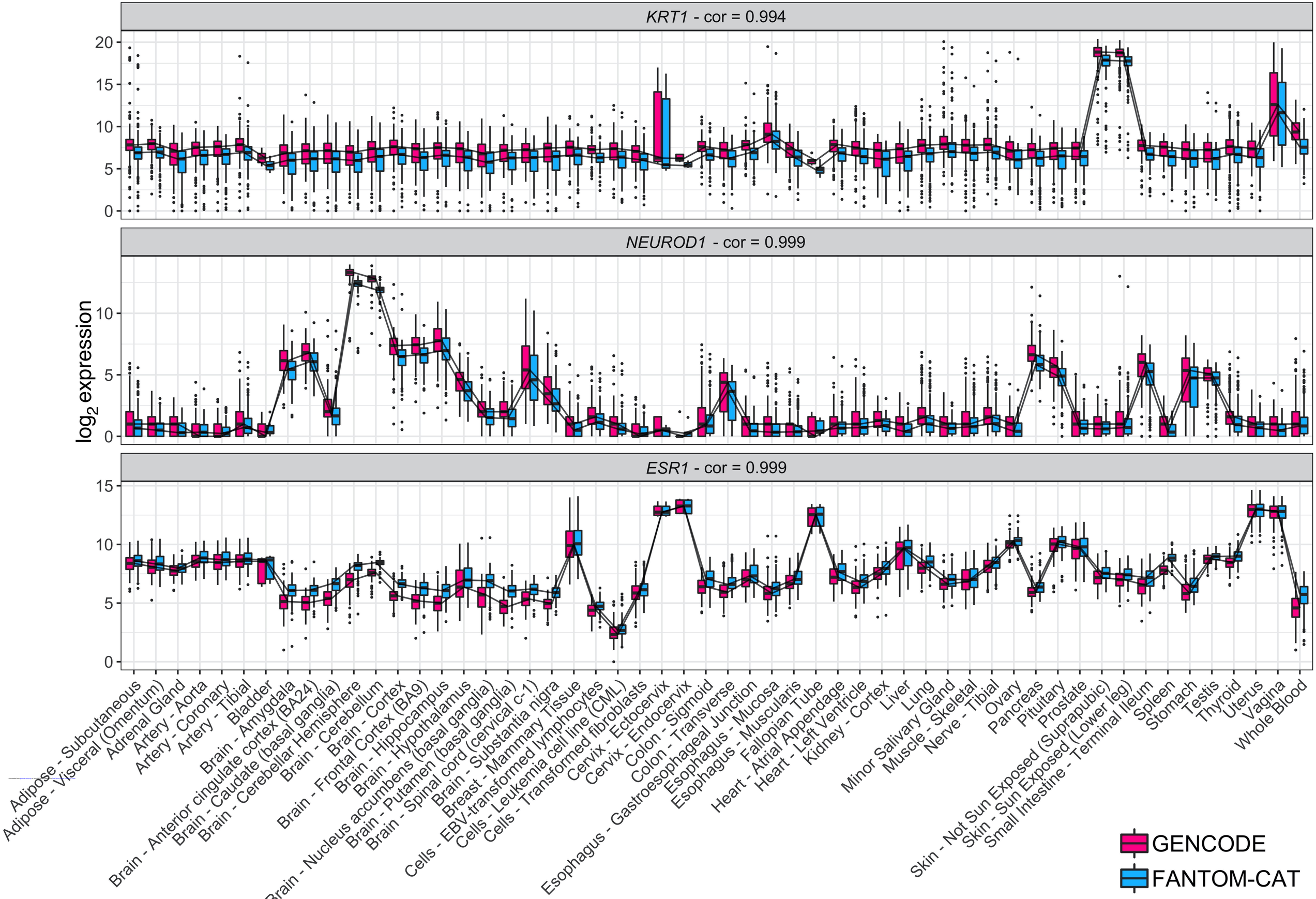
Cancer type	Total	dp-lncRNA		e-lncRNA		ip-lncRNA		mRNA	
		Up	Down	Up	Down	Up	Down	Up	Down
Bile	7010	200 (60)	313 (90)	186 (89)	203 (99)	47 (12)	84 (17)	2658 (106)	3319 (97)
Bladder	7680	344 (125)	319 (87)	140 (68)	149 (67)	65 (19)	82 (7)	3112 (201)	3469 (61)
Breast	15290	753 (291)	721 (202)	656 (377)	583 (305)	207 (50)	178 (32)	6109 (296)	6083 (244)
Colorectal	13685	490 (164)	592 (168)	381 (203)	400 (196)	130 (32)	160 (28)	5538 (371)	5994 (132)
Esophagus	4883	87 (21)	193 (50)	90 (38)	184 (103)	40 (11)	48 (2)	1921 (83)	2320 (77)
Head and Neck	10517	442 (138)	401 (96)	267 (139)	251 (112)	100 (23)	109 (18)	4329 (256)	4618 (53)
Kidney	15697	734 (238)	820 (281)	535 (299)	486 (209)	203 (45)	200 (48)	6349 (525)	6370 (114)
Liver	10554	346 (94)	395 (106)	230 (102)	248 (123)	90 (16)	112 (19)	4164 (174)	4969 (95)
Lung	17143	864 (338)	835 (304)	893 (512)	729 (396)	242 (76)	213 (39)	7523 (532)	5844 (212)
Prostate	13183	686 (287)	654 (218)	418 (254)	452 (214)	175 (55)	167 (30)	5153 (489)	5478 (128)
Stomach	11309	528 (213)	518 (164)	462 (291)	436 (240)	144 (51)	129 (22)	4509 (558)	4583 (89)
Thyroid	14264	752 (284)	804 (318)	527 (295)	594 (332)	161 (39)	174 (47)	5403 (189)	5849 (308)
Uterus	12906	641 (285)	713 (235)	454 (263)	612 (341)	210 (79)	225 (54)	5135 (335)	4916 (181)
Mean	11855	528 (195)	560 (178)	403 (225)	410 (211)	140 (39)	145 (28)	4762 (317)	4909 (138)
St. Dev	3650	237 (102)	218 (89)	225 (137)	189 (107)	67 (23)	55 (16)	1557 (167)	1234 (77)

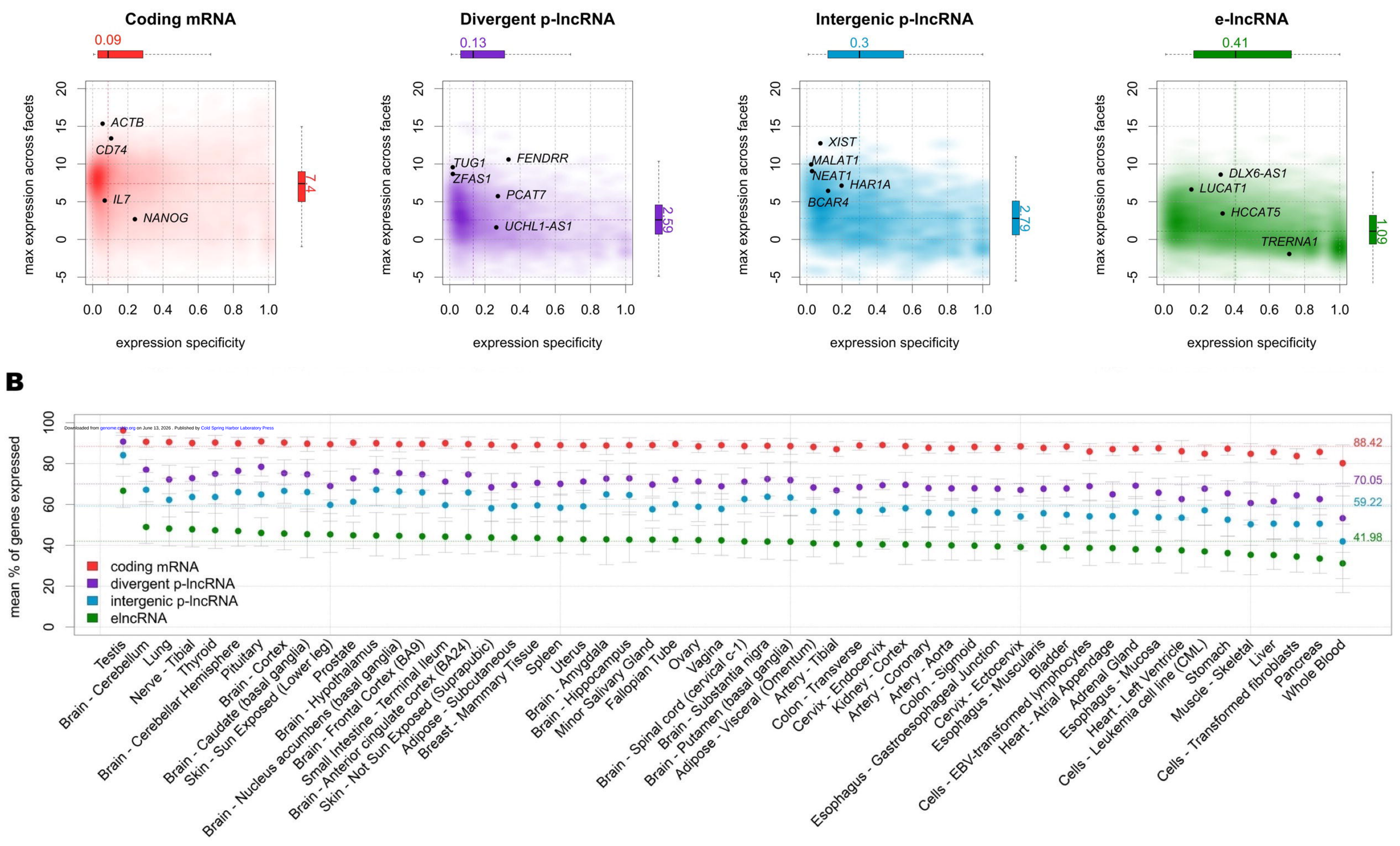
Tables Legends

Table 1. Differentially expressed genes in cancer. The table below summarizes the number of significant DEG (*FDR* < 0.01) between tumor and normal samples across the 13 cancer types analyzed for each gene class considered (coding mRNA, ip-lncRNA, dp-lncRNA, and e-lncRNA). Counts are reported separately for DEG up- and down-regulated in cancer, and values in parenthesis represents the number of genes exclusively annotated in the FANTOM-CAT gene model. Mean and standard deviation across cancer types is shown at the bottom.

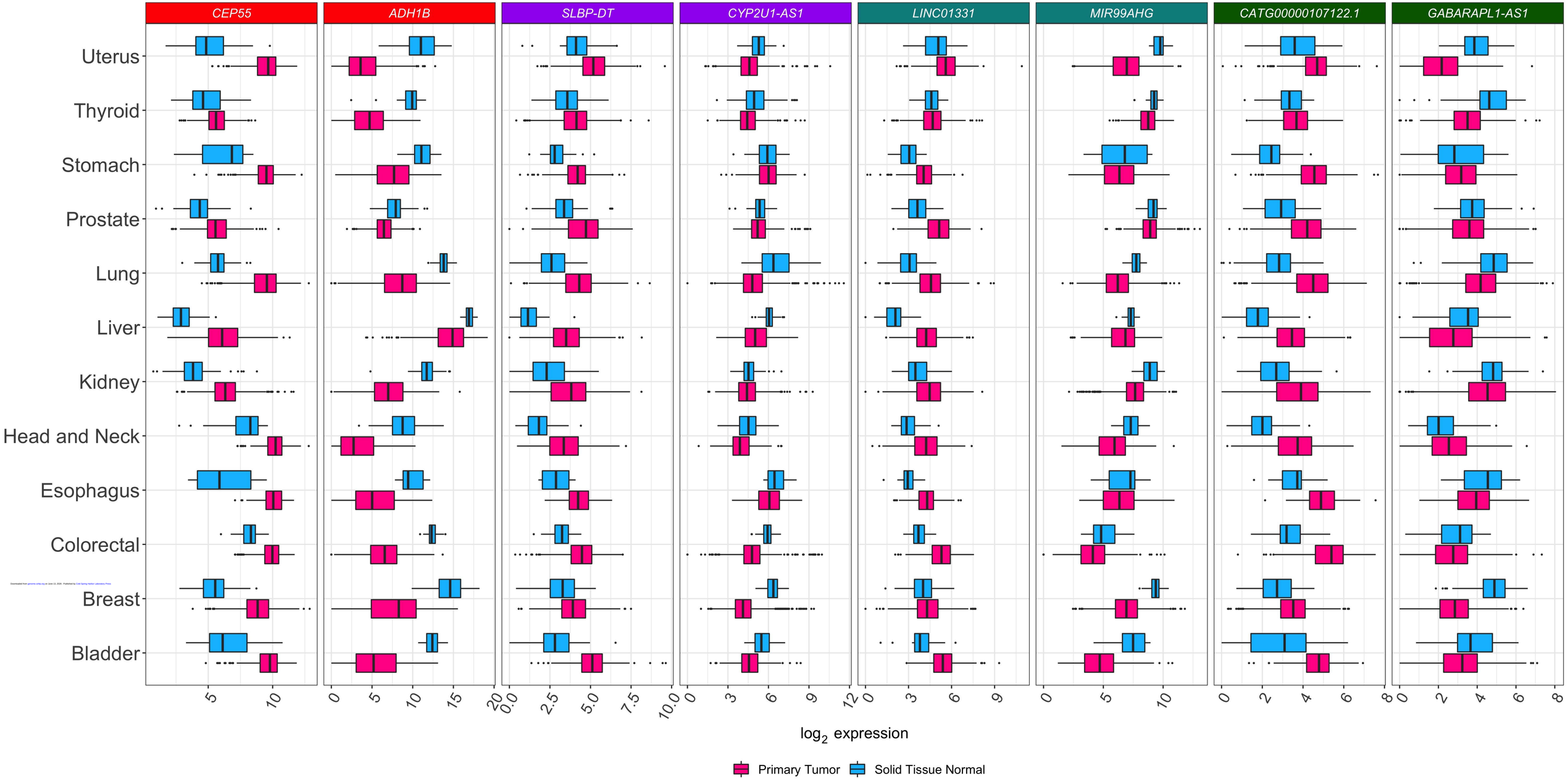


Downloaded from genome.cshlp.org on June 12, 2025 - Published by Cold Spring Harbor Laboratory Press

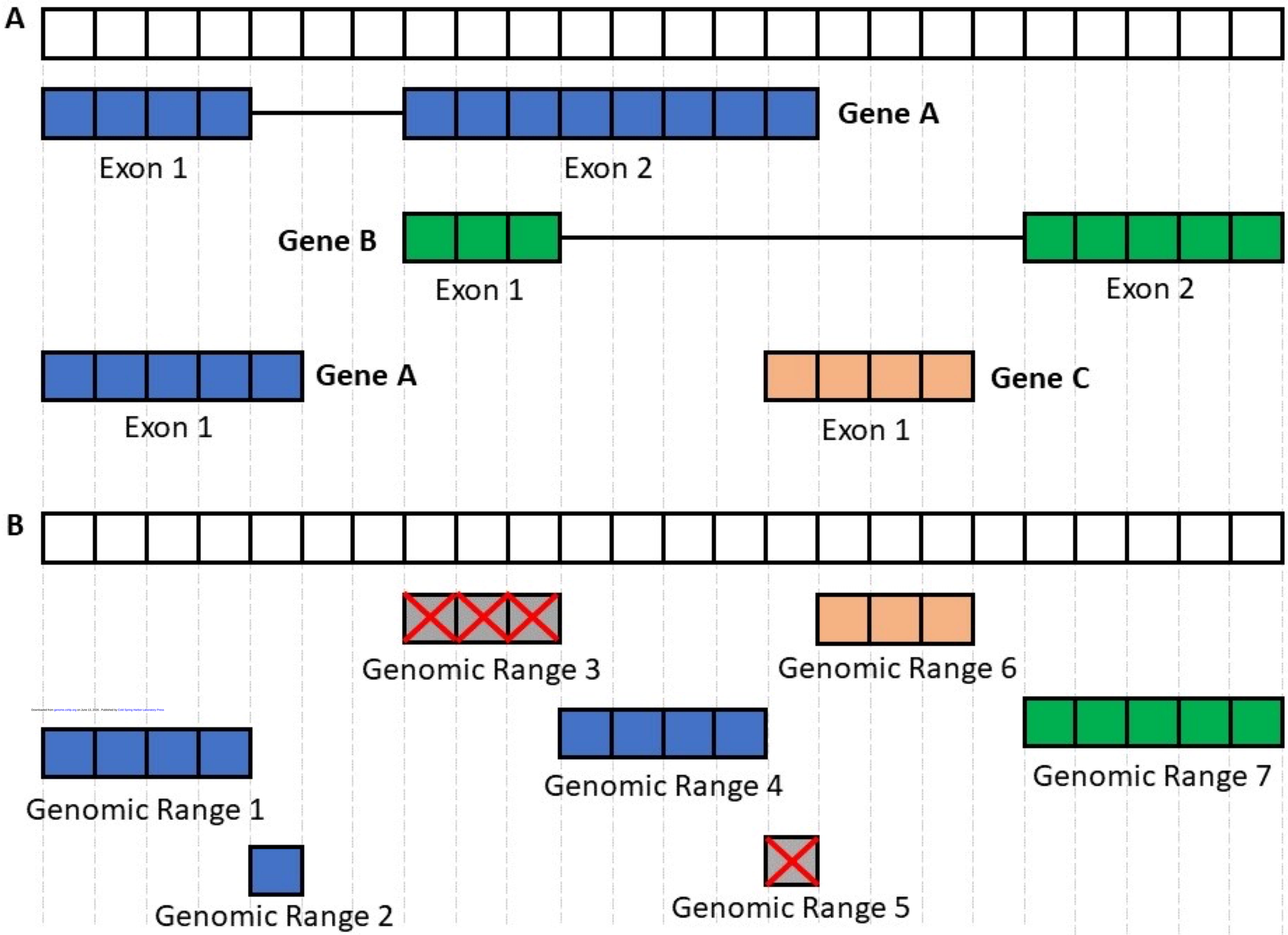

KRT1 - cor = 0.994
NEUROD1 - cor = 0.999
ESR1 - cor = 0.999
log2 expression
Adipose - Subcutaneous
Adipose - Visceral (Omentum)
Adrenal Gland
Artery - Aorta
Artery - Coronary
Artery - Tibial
Bladder
Brain - Amygdala
Brain - Anterior cingulate cortex (BA24)
Brain - Caudate (basal ganglia)
Brain - Cerebellar Hemisphere
Brain - Cerebellum
Brain - Cortex
Brain - Frontal Cortex (BA9)
Brain - Hippocampus
Brain - Hypothalamus
Brain - Nucleus accumbens (basal ganglia)
Brain - Putamen (basal ganglia)
Brain - Spinal cord (cervical c-1)
Brain - Substantia nigra
Breast - Mammary Tissue
Cells - EBV-transformed lymphocytes
Cells - Leukemia cell line (CML)
Cells - Transformed fibroblasts
Cervix - Ectocervix
Cervix - Endocervix
Colon - Sigmoid
Colon - Transverse
Esophagus - Gastroesophageal Junction
Esophagus - Mucosa
Esophagus - Muscularis
Heart - Atrial Appendage
Heart - Left Ventricle
Kidney - Cortex
Liver
Lung
Minor Salivary Gland
Muscle - Skeletal
Nerve - Tibial
Ovary
Pancreas
Pituitary
Prostate
Skin - Not Sun Exposed (Suprapubic)
Skin - Sun Exposed (Lower leg)
Small Intestine - Terminal Ileum
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
Whole Blood
GENCODE
FANTOM-CAT



Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press



Downloaded from genome.cshlp.org on June 13, 2026 - Published by Cold Spring Harbor Laboratory Press



Downloaded from genome.cshlp.org on June 13, 2026 . Published by Cold Spring Harbor Laboratory Press