

Determining gene specificity from multivariate single-cell RNA sequencing data

Nikhila P. Swarna¹, A. Sina Boeshaghi³, Elisabeth Rebboah^{4,5}, M. Grace Gordon^{6,7,8,9}, Pooja Kathail¹⁰, Taibo Li¹¹, Marcus Alvarez⁸, Chun Jimmie Ye^{7,8,9,12,13,14,15,16}, Barbara Wold¹, Ali Mortazavi^{4,5}, and Lior Pachter^{1,2,*}

¹Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA

²Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA, USA

³Department of Bioengineering, University of California at Berkeley, Berkeley, CA, USA

⁴Department of Developmental and Cell Biology, University of California at Irvine, Irvine, CA, USA

⁵Center for Complex Biological Systems, University of California Irvine, Irvine, CA, USA

⁶Biological and Medical Informatics Graduate Program, University of California, San Francisco, CA, USA

⁷Division of Rheumatology, Department of Medicine, University of California, San Francisco, CA, USA

⁸Institute for Human Genetics, University of California, San Francisco, CA, USA

⁹Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA, USA

¹⁰Center for Computational Biology, University of California, Berkeley, Berkeley, CA, USA

¹¹Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA

¹²Gladstone-UCSF Institute of Genomic Immunology, San Francisco, CA, USA

¹³Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, USA

¹⁴Baker Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA, USA

¹⁵Parker Institute for Cancer Immunotherapy, University of California, San Francisco, San Francisco, CA, USA

¹⁶Arc Institute, Palo Alto, CA, USA

*To whom correspondence should be addressed: lpachter@caltech.edu

Abstract

An important application of single-cell genomics experiments is to identify genes specific to biological categories or experimental conditions. Although numerous approaches have been proposed to identify such genes, we consider an axiomatic approach based on defining properties that a specificity measure should have. This leads us to develop ember (Entropy Metrics for Biological ExploRation), which we show is the only method satisfying four key desired properties for a specificity measure. Applying ember to eight tissues from eight founder mouse strains, we find that gene specificity is often unintuitive: canonical markers can be supplanted, housekeeping genes are context-dependent, and mouse strain can drive unexpected cell type switching. Unsupervised learning on entropy metrics uncovers shared genes specialized to male gonads and kidney, as well as genes specific to non-consecutive developmental stages in the kidney. To facilitate further exploration of gene specificity in mice, we have also developed a comprehensive specificity database, along with a web interface and API. Extending ember to a human PBMC dataset collected from 255 diverse individuals, we find that variation in PBMCs is largely localized to classical monocytes. We also find genes with unique specificity by sex, age and ancestral background. Together, these applications establish ember as a powerful tool and provide a roadmap for elucidating the impact of human genetic variation using the murine model.

Introduction

With the advent of combinatorial barcoding and split-pool technologies in single-cell RNA sequencing (scRNA-seq), experiments have become increasingly high-dimensional (1, 2). Modern experiments not only assay tens of thousands of genes, but can also accommodate replicates across multiple biological variables: sex, tissue, age, genotype, and cell states

(3–8). Although this expansion in experimental size paired with an increased depth of sequencing allows unprecedented resolution of cellular and molecular biology, there is a lack of tools to appropriately dissect datasets and extract meaningful insights (9–11).

One challenge in analyzing such multivariate data is the identification of genes that are highly specific to a given variable. Several measures of specificity have been proposed to address this challenge. For example, Tau, or the tissue specificity, index (TSI) summarizes gene specificity into a single number bounded between 0 and 1, with 0 indicating broad expression and 1 indicating perfect specificity (12, 13). The preferential expression measure (PEM) assigns positive or negative values to indicate over- or under-expression in a given tissue (14, 15), producing tissue-by-tissue scores that can then be aggregated. The Gini index, originally developed in economics to quantify inequality, has also been adapted to describe the distributional unevenness of gene expression across tissues (16, 17). Specificity indices have also been studied empirically (18). While such studies inform discrepancies between methods, they rely on the number of genes identified as specific to assess the quality of a method, ignoring its intrinsic properties (or lack of properties). In particular, the issue of specificity in the context of multivariate data seems to have been largely overlooked. Moreover, a major limitation of existing specificity measures in multivariate contexts is their lack of decomposability: the specificity does not decompose hierarchically in a coherent manner, and depends on the order in which variables are partitioned, preventing consistent application of indices such as TSI across multiple dimensions. Conversely, generalized linear model (GLM) frameworks such as DESeq2 (19) and EdgeR (20), offer decomposability but rely on parametric assumptions and can scale poorly.

Ideally, a specificity measure should satisfy several natural properties. It should be bounded between 0 and 1 in-

dicating least and most specificity respectively, be insensitive to the ordering of categories, vary smoothly with the data, and behave coherently under hierarchical refinements in multimodal data (*Supplementary Note*). We show that an entropy-based approach, grounded in an axiomatic information theory framework (21, 22), provides a natural foundation for such a measure. This approach has been considered before (23), however, its reliance on the Kullback–Leibler divergence breaks key desirable additivity properties. In addition to these natural properties, any programmatic implementation should leverage single-cell resolution, efficiently scale to high-dimensional data and be robust to biological replicates for ease of use in practice.

We propose a set of three biologically informative entropy metrics: Ψ (Psi): fraction of information explained by a partition, ψ_{block} (psi_block): specificity to a block, and ζ (Zeta): specificity to a partition. Detailed derivations and interpretations are provided in the *Methods* section.

Together with novel visualization strategies, these metrics provide a principled and interpretable toolkit for studying biological specificity in high-dimensional single-cell data. In fact, we show that our specificity measure is the only one that satisfies four basic requirements that a specificity measure should have (*Supplementary Note*). We implemented our approach in a Python package called ember: Entropy Metrics for Biological ExploRation. In the following sections, we apply ember to three multivariate single-cell datasets: snRNAseq of 8 tissues from 8 individuals for each of the 8 founder mouse strains (8cube founder data (6)), scRNAseq of murine kidney at 6 developmental time points (developmental kidney data (7)), and scRNAseq of human PBMCs from 255 diverse individuals (human PBMC data), demonstrating the framework’s ability to uncover novel condition-specific genes as well as larger-scale biological patterns.

Results

Gene specificity is unintuitive and hierarchical additivity allows for novel gene discovery.

Strain-specific genes. Applying ember to the 8cube founder data set (6), we identified genes that are strain-specific across all eight tissues (Fig. 1a). We identified *Cwc22*, which encodes spliceosome-associated protein 58, as a WSB/EiJ (WSBJ) specific gene. *Cwc22* has been documented to have heightened expression in WSBJ mice due to a copy number expansion at the *R2d2* locus, located approximately 6 Mb downstream from its paralog *R2d1* (26) (Supp. Fig 1c). We also identified the gap junction protein encoded by *Gjb4*, as specific to NZO/HILtJ (NZOJ) (Supp Fig 1d). It has been documented that *Gjb4* expression in diabetic NZOJ mice is enhanced by a reduced *miR-341-3p* expression (27).

We identified *Tas1r1* as a CAST/EiJ (CASTJ) specific gene, with a slight bias towards females (Supp Fig 1a). *Tas1r1* encodes a member of the taste receptor family of genes, and functions as a receptor of umami taste along with another member of the same family, *Tas1r3* (28). Interestingly, it has also been observed that this wild-derived

strain displays a unique oral preference to sugars, distinguishing CASTJ from other strains of mice and rats (29). When comparing transcriptomic differences between mouse strains mapped to the standard GRCh39 reference genome, we run the risk of falsely quantifying mapping artifacts as strain specificity because the GRCh39 assembly is derived from a C57BL/6J (B6J) background (6). To exclude the possibility that *Tas1r1*’s specificity to CASTJ mice could be the result of such a mapping artifact, we examined its coding sequence from the recent first release of the telomere-to-telomere (T2T) genomes for B6J and CASTJ (30, 31) (Supp. table 1). There is no evidence of frameshift or nonsense mutations and no alternative splicing, as the total CDS and protein length are conserved. This indicates that the *Tas1r1* gene in B6J is not a pseudogene in CASTJ. Between the strains, 19 SNPs occur in exonic regions, of which three cause amino acid substitutions that may alter protein structure. We observed a large 4396 base pair deletion in B6J compared to CASTJ in the fourth intron at the CASTJ T2T coordinates 4:153181901-153186296 (reverse strand). Large intronic deletions have been associated with over or under expression of genes (32–34), thus cluing why we may be observing this specificity of *Tas1r1* to CASTJ. The gene just upstream of *Tas1r1* on the reverse strand chromosome four, *Zbtb48*, also exhibits biased expression towards CASTJ, though not as striking as that of *Tas1r1* (Supp Fig 1b), further validating our finding.

In addition to strain-specific genes, we also observed genes that are specific to pairs of strains. *Or5m9* encodes an olfactory receptor specific to B6J and 129S1/SvImJ (129S1J). Another olfactory receptor *Or5p58* is specific to the wild-derived strains PWK/PhJ (PWKJ) and CASTJ, the most genetically distant from B6J of the eight founder strains. *Mir7001*, encodes a microRNA specific to the two diabetic mouse strains NZOJ and NOD/ShiLtJ (NODJ). Additionally, we found that gene specificity to mouse strain is variable across biological replicates. We applied a novel visualization strategy to depict ψ_{block} scores for each founder mouse strain and their mean and standard deviation values in compositional space as an octagon, where proximity to a vertex represents greater specificity (Fig. 1a, detailed explanation in *Methods*). These results show that ember can identify known and novel genes that are specific to one or more mouse strains.

Cell type markers and housekeeping genes. Our method ember can be used to find cell type markers and housekeeping genes, while taking into account all possible partitions in a multivariate dataset. We found markers for skeletal muscle satellite cells, a rare but essential stem cell population required for adult muscle repair, making up only 1% of all nuclei in the 8cube gastrocnemius dataset (6) (Fig. d). The top satellite stem cell marker to emerge from our analysis was *Calcr*, calcitonin receptor (Fig. 1b, 1d). *Calcr* maintains the muscle stem cell pool by keeping them in a dormant state and in their location via the *cAMP*-PKA pathway (35). We were surprised that *Pax7*, the canonical myogenic marker, did not appear in our list of top satellite cell markers. In our

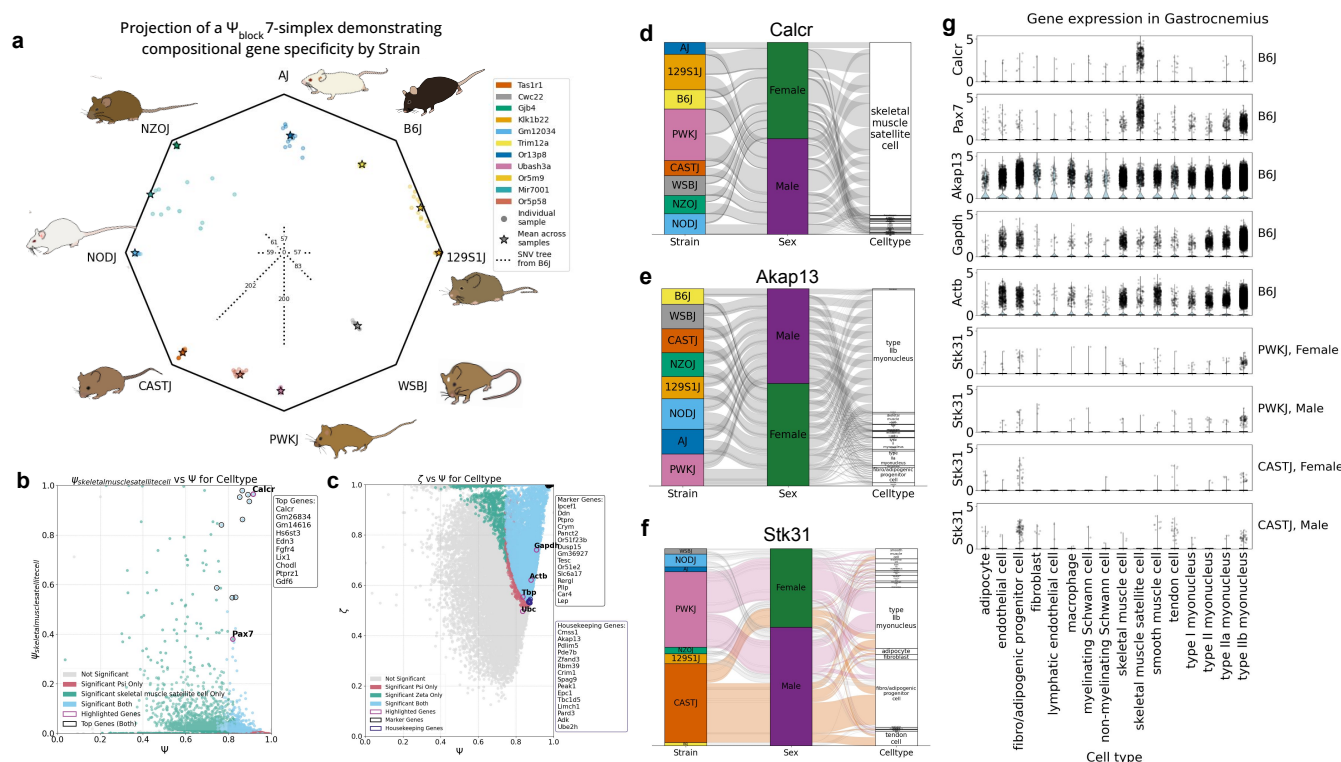


Fig. 1. Visualizing gene specificity and finding highly specific and non-specific genes **a.** Octagonal plot showing gene specificity in compositional space. The 8cube founder dataset (6) was partitioned by strain to generate ψ_{block} scores for each of the eight founder strains. Each gene is represented as a coordinate on a 7-simplex, where each vertex indicates perfect specificity to one strain. We project genes according to single nucleotide variant distance to B6J. Selected genes show specificity to one or two strains. Fifty balanced samples across biological replicates were used to generate ψ_{block} scores. Each sample is shown as a grey circle, and the Aitchison mean of these 50 samples is shown as a star. Mouse images from Saul *et al.* 2019 (24) **b.** Scatter plot showing gene specificity to skeletal muscle satellite cells in 8cube gastrocnemius (6), partitioned by cell type. Genes with high Ψ and high $\psi_{skeletal\ muscle\ satellite\ cells}$ are identified as marker genes (highlighted in black). Significance testing for both metrics is shown by color: genes significant for both are light blue and exhibit reliable specificity. Top marker genes are listed in the top-right text box. **c.** Scatter plot showing gene specificity in 8cube gastrocnemius (6), partitioned by cell type. Genes with high Ψ and high ζ are highly specific (highlighted in black), whereas those with high Ψ and low ζ are non-specific (highlighted in violet). Canonical housekeeping genes are highlighted in purple. Top cell type-specific genes are listed in the top-right text box, and top housekeeping genes are listed in the bottom-right text box. **d, e, f.** Alluvial plots depicting pseudo-bulked gene expression of *Calcr*, *Akap13* and *Stk31* across strains, sexes and cell types in 8cube gastrocnemius (6), generated using wompomp (25). **g.** Violin plots of gene expression in 8cube gastrocnemius (6), with gene names on the left and strain/sex annotations on the right.

analysis, *Pax7* was significant after permutation testing by $\psi_{satellite\ cells}$ but not by Ψ (Fig 1b), which means that *Pax7* has some biased expression in satellite cells, but partitioning by cell type does not explain its expression pattern. Examining total expression across all gastrocnemius cell types, we found that *Pax7* also has significant expression in type IIB myoneuceli, the most populous of the gastrocnemius cell types (Fig. 1g). We validated these findings by comparing to a gastrocnemius single-nucleus dataset generated by Alexander S. Ham *et al.* 2025 (36), where we observed the same expression of *Pax7* in both satellite cells and type IIB myoneuceli (Supp. Fig. 2).

Selecting for genes partitioned by cell type with high Ψ and low ζ , we found housekeeping genes for the 8cube gastrocnemius (6) (Fig. 1c). From our analysis, *Akap13*, which encodes an A kinase receptor protein, emerged as one of the top housekeeping genes by cell type (Fig. 1e). We found that *Akap13* outperforms several of the canonical housekeeping genes such as *Gapdh*, *Actb*, *Ubc* and *Tbp* (Fig 1c, 1g). Moreover, it has been documented that *Gapdh* mRNA is highly variable across tissues and cell types (37), making it an unreliable default housekeeping gene. This result illustrates the advantage of taking a context-aware approach to housekeep-

ing gene selection using ember.

Strain-driven cell type specificity. We also uniquely find genes that switch cell type specificity based on strain. *Stk31* encodes serine threonine kinase 31 which enables protein kinase activity. Bao *et al.* 2013 report that *Stk31* is exclusively expressed in spermatogenic cells of testes in mice, their study however is limited to mice of a C57BL/6J background (38). While in the 8cube dataset we observe that *Stk31* expression is generally highest in male gonads, specifically in CASTJ we observe some expression in multiple other tissues (Supp. Fig 1e). For instance in gastrocnemius, we observe that it is selectively expressed in fibro-adipogenic progenitor cells in CASTJ mice but in type IIB myonuclei in PWKJ mice (Fig. 1g). This strain-dependent expression reveals that genetic background can drive cell type specificity, challenging the notion of universal cell type markers. This especially motivates exercising caution when drawing cross-strain and cross-species conclusions from a mouse model. To fuel effective and informed use of mouse models in biological research, we have deposited specificity and gene expression information across the 8 diverse founder strains in 8 tissues into a publicly available and programmatically accessible API database and

website (*Data Availability*). Together, these results showcase the ability of ember as a tool to evaluate the efficacy of marker and housekeeping genes to a partition while also identifying novel strain-aware cell type-specific genes.

Unsupervised learning from entropy metrics can identify clusters of specific genes.

Clustering by tissue. We performed unsupervised machine learning using the leiden clustering algorithm on the entire 8cube data, clustering on ψ_{block} scores of each tissue (Fig. 2a) (6, 39). We found 4863 out of the 14181 genes with $\Psi > 0.45$, $\zeta > 0.45$ were specific to male gonads (34%). This is in agreement with previous work in humans by Djureinovic *et al.* 2014, which confirms that male gonads are the tissue with most tissue-specific genes (40). We also found clusters of genes that are specific to pairs of tissues involving male gonads (Table 1, Supp. tables 2-8). After performing a functional enrichment analysis (41) of the 226 genes with shared specificity to kidney and male gonads, we found 68 genes are regulated by a *Twist* transcription factor. *Twist1* and *Twist2* are paralogs in the *Twist* subfamily of basic loop-helix-loop proteins, which are essential for embryogenesis and are conserved across vertebrates and invertebrates (42, 43). The role of *Twist1* has been extensively studied in the kidney and is implicated in kidney disease (44–46). However, previous research has found that *Twist1* is not detected in the murine testes (47), so these genes with shared regulation by *Twist* family in both male gonads and kidney lead us to a novel line of biological questioning.

Serpina1f, which encodes serine peptidase inhibitor, clade A member 1f, is one of the 68 genes specific to male gonads and kidney and regulated by *Twist* (Fig 2b). *Serpina1f* has been previously identified as kidney-specific in a semi-quantitative PCR analysis study of adipose tissue, muscle, heart, lung, liver and kidney (48). We see from our data that *Serpina1f* is largely localized to proximal tubule epithelial cells in both males and females (Fig 2b, 2d). In male gonads, *Serpina1f* has been localized to mouse epididymis in an RNA in-situ hybridization study (49, 50) and to sperm cell membrane in a proteomics study (51). We also observed localization of *Serpina1f* to epididymis and spermatid in male gonads (Fig 2e). These results validate that *Serpina1f* is reproducibly and distinctly observed in both kidney and male gonads.

The only tissue with a cluster of highly specific genes ($\psi_{block} > 0.8$) other than male gonads was Liver (Fig 2a, Cluster 26, 230 genes). A functional enrichment analysis on this cluster of genes revealed regulation of large sets of these genes by six transcription factors: *Tcf-1*, *Cdx-2*, *Crx*, *Mef-2c*, *Sox-10* and *Cdp*. *Tcf-1* or *Hnf1a* encodes a liver-enriched transcription factor critical to several aspects of liver function (52–56). *Hpx* is one of the 26 genes that we identified as liver-specific and regulated by *Tcf-1* (Fig 2c, Supp. table 3, 8). *Hpx*, encodes hemopexin, a protein that binds heme and transports it to the liver for breakdown and iron recovery. While *Hpx* is visibly specific to liver across all 8 tissues (Fig

Table 1. Clusters of genes from Fig 2a. with specificity to male gonads and one other tissue.

Tissue	Cluster	Gene Count	Functional Enrichment Summary
Female Gonads	31	136	mitosis, chromosome segregation, reproductive processes
Kidney	24	226	Regulated by twist motif, peptidase activity, serine-type peptidase activity, serine hydrolase activity
Cortex, Hippocampus	27	208	Cell signaling, membrane receptor activity, GPCR activity, olfactory receptor activity
Diencephalon, Pituitary	18	363	Sensory perception, response to chemical stimulus, GPCR activity, olfactory receptor activity
Gastrocnemius	36	53	None

2c), at the strain level we noticed that *Hpx* shows a slight bias towards PWKJ mice, which are the most prone to develop hepatic inflammation of the founder strains (57). At the cell type level, we noticed localization of *Hpx* to hepatocytes across all strains (Fig 2f). These results show that we can find clusters of biologically meaningful genes with tissue specificity.

Clustering by age. We used ember to understand gene specificity in a developmental kidney dataset produced by Chen *et al.* 2025. In this data, the developing kidney is captured in six stages: E16.5, P0, week 3 (W3), week 12 (W12), week 52 (W52) and week 92 (W92), corresponding approximately to embryo, newborn, youth, adolescence, adult and old age in humans(7). Partitioning this data by age, we generated entropy metrics and then clustered genes by their ψ_{block} scores (Fig. 2g). This analysis revealed a cluster of genes specific to old age with high specificity to W92 and slightly lower specificity to W52 (Cluster 10, 486 genes, Supp. table 9). A gene set enrichment analysis using the molecular signatures database (MSigDB) on this cluster of genes revealed heightened immune activity (Supp. table 10) (58). This is in agreement with findings from the Chen *et al.* 2025 authors, who found steadily increasing immune cell proportions from W12 to W92, indicating an accumulation of immune cells in kidneys upon aging (7). One gene in this cluster of "aging-specific genes" is *Igkc*, which encodes immunoglobulin kappa constant. *Igkc* has previously been identified as an aging marker across several tissues, including kidney, in a proteomics study comparing 32 and 72 week old mice (59).

We found a cluster of genes specific to non-consecutive developmental stages. Cluster 22, a cluster of 222 genes, shows specificity to ages P0, W52 and W92. A gene set enrichment analysis of this cluster using MSigDB revealed heightened immune and inflammatory response (Supp. table 9, 11) (58). A study by Speer *et al.* 2025 found that neonatal sepsis lead to increased expression of renal tissue inflammation and injury biomarkers, consistent with acute kidney injury, and drawing some parallels between the molecular signatures of a new born mouse to that of an aged one (60), hinting at that biology that may be underlying this unique non-consecutive bimodal expression. Secretory leukocyte pro-

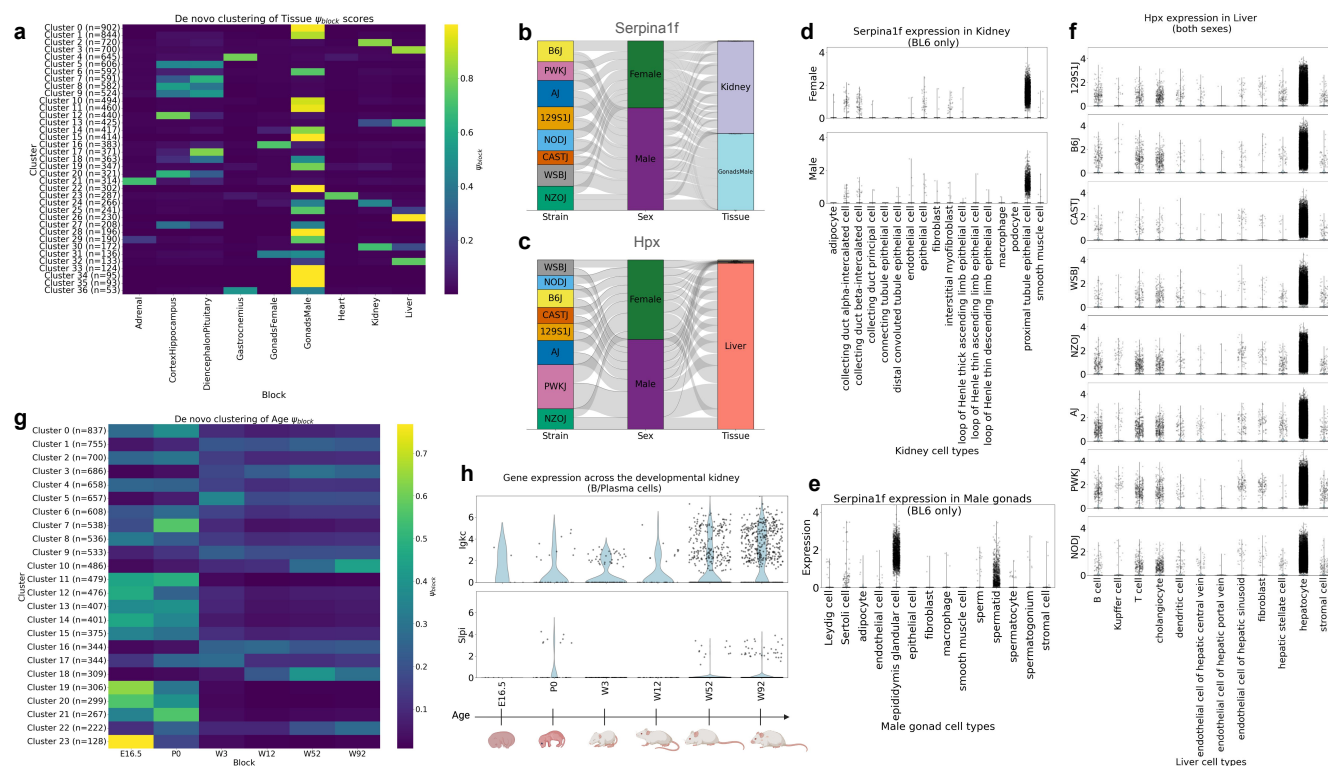


Fig. 2. Unsupervised clustering of ψ_{block} scores. **a.** Heatmap showing de novo unsupervised clustering of ψ_{block} scores generated from the 8cube dataset (6), partitioned by tissue. **b, c.** Alluvial plots depicting pseudo-bulked gene expression of *Serpina1f* and *Hpx* across strains, sexes and tissues in 8cube (6), generated using wompomp (25). **d.** Violin plots of *Serpina1f* gene expression in 8cube B6J kidney in each sex, across cell types (6). **e.** Violin plot of *Serpina1f* gene expression in 8cube B6J male gonads across cell types (6). **f.** Violin plots of *Hpx* gene expression in 8cube Liver in both sexes, across strains and cell types (6). **g.** Heatmap showing de novo unsupervised clustering of ψ_{block} scores generated from the developmental kidney dataset (7), partitioned by age. **h.** Violin plots of *Igkc* and *Slpi* gene expression in the B/Plasma cells of developmental kidney in both sexes, across ages (7), mouse developmental timeline image on x-axis created in <https://BioRender.com>.

tease inhibitor (*Slpi*), is a gene within this cluster of genes specific to new borns and old age, plays a regenerative role in the ischemic renal (I/R) injury model, through activation induced by *MIF-2/D-DT* (61). Together, these results show that applying unsupervised learning techniques to entropy metrics provides an unbiased methodology for identify clusters of genes with distinct and unexpected expression patterns.

Tissue and cell type are the dominant source of biological variation in a scRNAseq experiment.

Tissue-linked specificity. Generating entropy metrics for multiple partitions in a dataset allows unique insights into biological patterns and trends. We generated Ψ and ζ values partitioning the entire 8cube data by tissue, sex and mouse strain and their two and three way interaction terms (6). We extracted lists of genes most specific to these partitions ($\Psi > 0.75, \zeta > 0.75$) and from these seven sets of genes, we generated an upset plot to visualize gene membership (Fig. 3a) (62). We found genes to be most specific to tissue, along with a large set of genes with sex specificity linked to tissue. We also found that tissue specificity dominates strain specificity, with only 124 genes highly specific to strain and 1401 genes highly specific to tissue. To further validate this observed pattern, we performed a principal component analysis of gene expression data pseudo-bulked by mouse strain and tissue. PC1 and PC2 together account for more than half of the total variability (55.8% of total variance, 34.3%

and 21.5% for PC 1 and 2 respectively), indicating that the major biological factors driving variation are captured effectively within the first two principal components (Fig. 3b). The pseudo-bulked expression profiles show distinct clustering by tissue (shape), over mouse strain (color). In addition, a correlation cluster map of this pseudo-bulked data also revealed that tissues across strains had higher correlation to one another (Supp. Fig 3). These results show that tissue is the dominant source of biological variation, making genes with specificity to partitions other than tissue especially interesting.

The upset plot in Fig 3a isolates *Cyp4a14*, which encodes a cytochrome P450 monooxygenase, to be the only gene highly specific to sex:strain:tissue over other partitions (Fig. 3c). *Cyp4a14* shows biased expression towards female liver and kidneys in all strains except PWKJ, which as we have previously noted, are the most prone of the founder strains to develop hepatic inflammation and progress to fibrosis of the founder strains (57). Incidentally, it has been documented that *Cyp4a14* is implicated in the progression of hepatic fibrosis and that over-expression in a bile duct ligation murine model exacerbates fibrosis, while a knockout attenuates it (63). In this analysis we also found *Spink6* to be the only gene highly specific to tissue:strain over other partitions (Fig 3d). *Spink6* encodes serine peptidase inhibitor kazal type 6 and shows biased expression towards NODJ kidney.

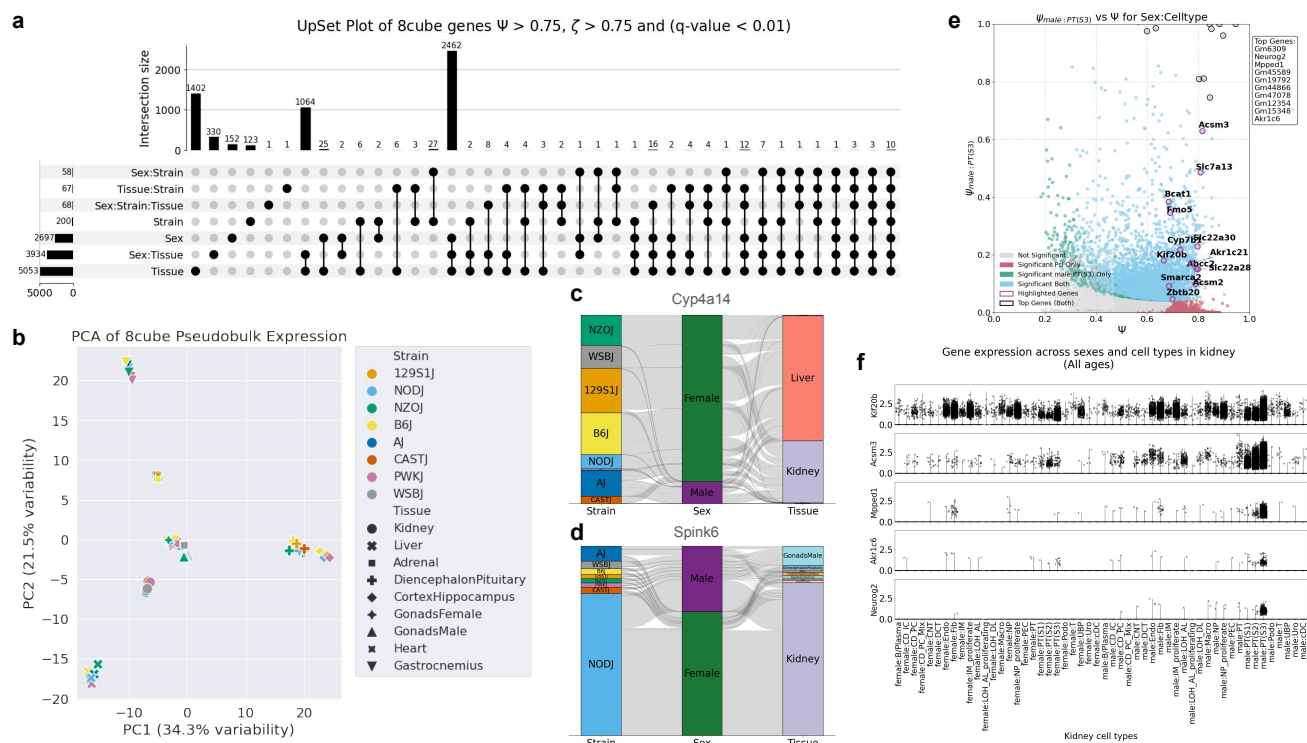


Fig. 3. Visualizing specificity patterns across partitions **a.** Upset plot generated from 8cube dataset (6) selecting highly specific genes partitioned by sex, strain, tissue and their 2-way and 3-way interaction terms. Thresholds used for highly specific gene selection are $\Psi > 0.75$ and $\zeta > 0.75$. Global testing correction was performed across all 7 partitioned and genes selected passed a significance threshold of 0.01. **b.** PCA plot of 8cube data pseudo-bulked by strain (color) and tissue (shape) (6). **c, d.** Alluvial plots depicting pseudo-bulked gene expression of *Cyp4a14* and *Spink6* across strains, sexes and tissues in 8cube (6), generated using wormwomp (25). **e.** Scatter plot showing gene specificity to male proximal tubule epithelial cells (Segment 3) in developmental kidney (7), partitioned by sex:celltype. Genes with high Ψ and high $\psi_{male:PTS3}$ are identified as marker genes (highlighted in black). Top differentially expressed genes in male PT(S3) cells identified by Chen *et al.* 2025 (7) are highlighted in purple. Significance testing via permutation testing for both metrics is shown by color: genes significant for both are light blue and exhibit reliable specificity. Top marker genes are listed in the top-right text box. **f.** Violin plots for *Kif20b* and *Acsn3*, identified as differentially expressed in male PT(S3) cells by (7) and *Mpped1*, *Akr1c6* and *Neurog2*, identified by ember as highly specific to male PT(S3) cells.

Cell-type-linked specificity. We observe a similar pattern at the tissue level, where a large proportion of biological variation within a tissue is explained by cell type rather than strain, 38.8% in 8cube kidney (6). In the developing kidney (7), we see that specificity by cell type dominates specificity by age and sex. (Supp. Fig 3a, b, c). In a tissue like the kidney, where it is interesting to study sexual dimorphism, being able to resolve specificity to sex from specificity to sex within a cell type is critical (64). Chen *et al.* 2025 identified that proximal tubule epithelial cells, specifically segment 3 (PTS3) present the largest number of sexually dimorphic genes (7). Chen *et al.* 2025 identified genes as sexually dimorphic in PTS3 cells with biased expression towards males using a non-parametric Wilcoxon rank sum through the Surate package.

We re-analyzed the developmental kidney dataset using ember to understand male-biased sexual dimorphism in PTS3 cells (7). We generated entropy metrics, partitioning the data by sex:celltype and overlaid the top genes identified by Chen *et al.* 2025 on a scatter plot showcasing specificity to male PTS3 cells (Ψ vs. $\psi_{male:PTS3}$). We noticed that these genes had $\Psi > 0.6$, and had a spread of $\psi_{male:PTS3}$ scores lower than 0.65 (Fig. 3e), indicating that while these genes may have high expression in male PTS3s, they are not truly specific to the block. Upon closer inspection we found

that one of the genes on this list, *Kif20b* along with having significant expression across various cell types, has biased expression towards males in PTS3 cells but towards females in loop of henle ascending limb proliferating cells, making it specific to neither sex nor cell type (Fig. 3f). We found that another gene on this list *Acsn3*, has high expression not only in male PTS3 cells but also in male PTS1 and male PTS2 (Fig 3f). Conversely, genes with high Ψ and $\psi_{male:PTS3}$ scores identified by ember show distinct and exclusive specificity to male PTS3 cells (Fig 3f). These results re-emphasize the advantage of leveraging decomposable gene-specificity when finding differentially expressed genes in a multivariate dataset.

Variation in humans. We generated entropy metrics for a human PBMC dataset of 255 individuals from four different ancestries: European, African, East Asian and Admixed American, partitioning by age, cell type, sex, ancestry and their two-way, three-way and four-way interaction terms (Fig 4a). We found, similar to mouse, that variation within PBMCs was dominated by cell type and that sex, age and ancestry related specificity are mostly linked to cell type. 105 genes were identified as specific to cell type or a cell-type-linked interaction term (Supp. table 12). We clustered these 105 genes on their ψ_{block} scores and found that not only is

cell type the dominant source of variation in PBMCs, but also this variation can be largely localized to classical monocytes (Fig. 4b). One of the 105 highly variable genes is *SLC11A1*, which encodes a member of the solute carrier family 11, a family of proton-coupled divalent metal ion transporters (Fig. 4c). Mutations in *SLC11A1* have been associated with susceptibility to diseases such as tuberculosis in Chinese population (65), kawasaki disease in Japanese population (66), cutaneous leishmaniasis in Pakistani population (67), type 2 diabetes mellitus in Iranian population (68) and many other ancestry specific variants that offer resistance or heightened susceptibility to disease (69).

Another gene in this list of 105 highly variable genes is *TREMI*, which encodes triggering receptor expressed on myeloid cells 1 and is an ancestral predictor for malaria susceptibility (Fig. 4d). A study of Columbian population showed lower malaria susceptibility in people of Native American ancestry, slightly higher susceptibility in European ancestry and oscillated between a protective effect and increased risk with no clear prediction in African ancestry (70). Another gene *ILB1* is on their list of ancestral malaria predictor genes as well as on our list of 105 highly variable genes linked to cell type.

Outside of cell-type-linked specificity, we found 20 genes uniquely specific to sex and 4 genes specific to sex:age, of which 18 were Y-chromosome-linked genes and the remaining 6 were X-chromosome-linked. We additionally found one gene, *RNF17*, which showed distinct specificity by sex:age:ancestry (Fig. 4e, f). *RNF17* encodes a protein containing a RING finger domain and has been previously considered testes specific. We found that *RNF17* shows biased expression in East Asian ancestry females in their twenties in human PBMCs.

Together, these results show that an entropy-guided approach is effective in understanding biological variance in human genetic data and can find known and novel partition-specific and highly variable genes. This analysis also draws to attention some limitations of the mouse model in understanding genetic variation in humans (24). We identified strain-specific genes across all tissues as well as within a tissue in the mouse 8cube data (6) (Fig. 1a), but such a parallel does not exist in humans. There are no genes that are exclusively ancestry-specific in human PBMCs. Instead, in humans we observe that ancestry specificity is linked to sex and cell type.

Discussion

Lists of marker genes or differentially expressed genes are often important takeaways from a scRNAseq experiment, but current methods lack a homogeneous definition of what constitutes a marker gene. An idealistic marker gene for a block in a partition should have distinct and exclusive expression in this block. Yet, even the long-noncoding RNA *Xist/XIST*, which is indispensable for X chromosome inactivation during early development in both humans and mice, has non-zero expression in normal male tissues (71). This brings to question whether such idealism, when trying to understand the

complexities of biology, may prevent us from truly learning what the transcriptomic orchestra of scRNAseq has to say.

A similar argument can be made for the definition of a housekeeping gene. An ideal housekeeping gene has been described as stably expressed irrespective of tissue type, developmental stage, cell cycle state, or external signal (72, 73). However, the inherently stochastic and bursty nature of transcription (74, 75) undercuts the existence of such a perfectly invariant housekeeping gene. Taking a more nuanced and context-aware approach to marker gene and housekeeping discovery, that accounts for the limitations of an experiment, such as tissues sampled, mouse strain used, sexes sampled, human populations sampled, and rarity and resolution of a cell type, may lead to more reproducible biological insights. Our tool ember's unbiased entropy-based approach, along with its attention to multivariate partitioning of a dataset allow for discovery of nuanced marker and housekeeping genes, quantifying specificity without assuming exclusivity.

A further challenge to interpreting nuanced gene expression lies in the availability of appropriate visualization tools. A violin plot allows for visualization at single-cell resolution, but is constrained to display expression within a partition. Alluvial plots allow for effective visualization of a gene's expression across multiple partitions, though at the cost of single-cell resolution (25, 76). To balance these trade-offs, we use both alluvial and violin plots in tandem. This complementary approach allowed us to highlight genes with intricate specificity uncovered by ember, providing interpretable and biologically meaningful visualizations. More broadly, our work aligns with critiques of visualization in single-cell genomics (77), emphasizing that the interpretability of biological results is inseparable from the limitations and choices of visualization strategy.

Our work provides a methodology to interpret gene specificity that is distribution agnostic, decomposable, robust to replicates and that fully leverages single-cell resolution, but it is not without its limitations. First, ember requires discrete partitioning of cells into blocks and is not equipped to deal with continuous variables. We recommend binning of continuous variables into biologically meaningful categories to get an approximate estimate of related gene specificity. Second, while ember is sensitive to lowly expressed genes, transcripts detected in very few cells yield unstable entropy estimates. As a result, filtering is required to remove genes expressed in fewer than 100 cells, which improves robustness but may inadvertently exclude biologically relevant or rare transcripts. Finally, computation of entropy metrics and p-values through sampling can be time consuming when dealing with dense data that have a large number of cells (>1000000 cells). We recommend using parallel computation and chunking strategies where possible to make analyses more manageable. Collectively, these constraints highlight areas for future methodological improvement.

When quantifying gene specificity in scRNA-seq experiments, a central challenge lies in placing results within a broader biological context. Frameworks that enable generalizable conclusions are essential for ensuring accessibility

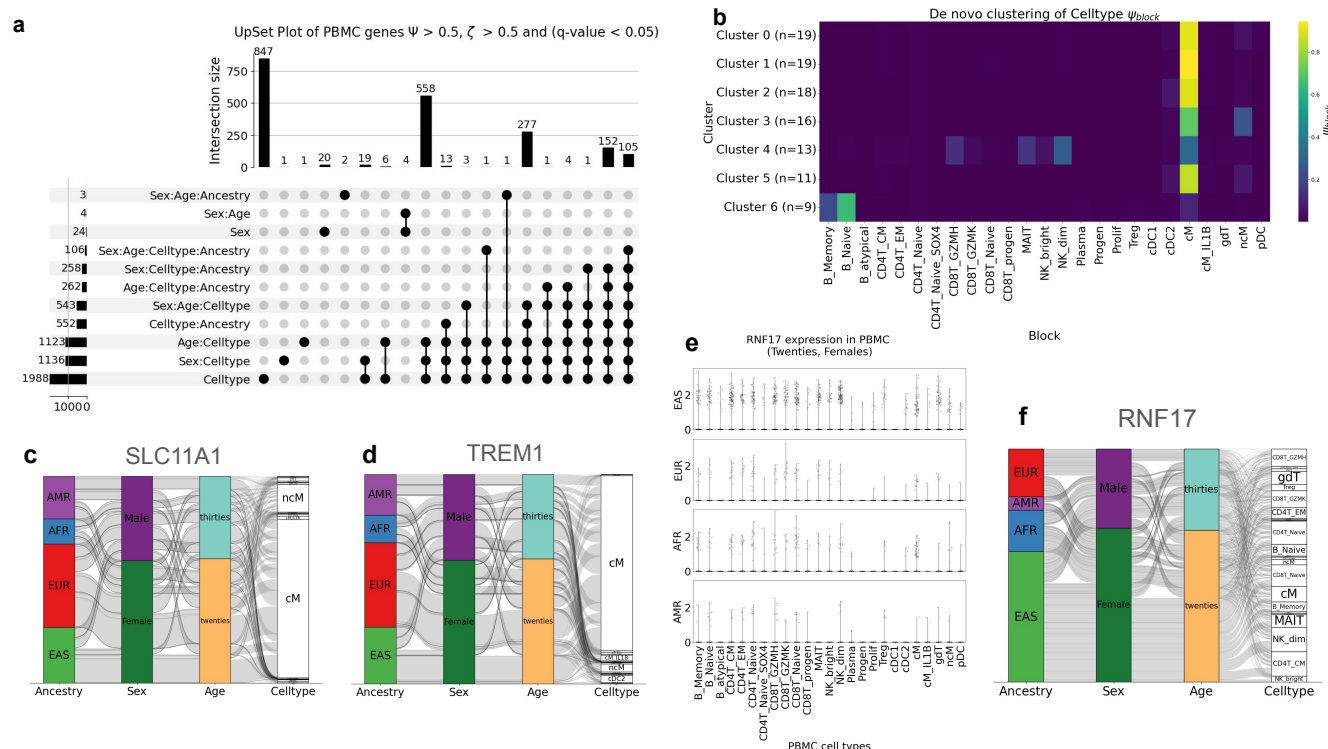


Fig. 4. Specificity in human PBMCs collected from 255 individuals from diverse ancestries. Upset plot generated from human PBMC dataset selecting highly specific genes partitioned by Sex, Age, Ancestry and Celltype and their 2-way, 3-way and 4-way interaction terms. Thresholds used for highly specific gene selection are $\Psi > 0.5$ and $\zeta > 0.5$. Global testing correction was performed across all 7 partitions and genes selected passed a significance threshold of 0.05. **c, d, f.** Alluvial plots depicting pseudo-bulked gene expression of *SLC11A1*, *TREM1* and *RNF17* respectively across ancestries, sexes, ages and cell types in human PBMCs. **e.** Violin plots of *RNF17* gene expression in 20-29 year-old females across ancestries and cell types in human PBMCs.

and reproducibility. Our entropy-based approach is particularly well-suited to revealing such organism-level patterns, distilling high-dimensional single-cell data into interpretable, large-scale conclusions. From our analyses, we found that the dominant source of biological variation within an organism is tissue or cell type. Specifically in the 8cube dataset, we conclude that tissue specificity outweighs genotype specificity. These findings align with conclusions made by Gilad and Mizrahi-Man in their reanalysis of the ENCODE data, that variability of genes by tissue is greater than variability by species (78). Taken together, these results reinforce the view that tissue is a primary driver of transcriptomic specificity.

Despite this broader pattern of specificity, it is wise to exercise caution when drawing cross-strain and cross-species conclusions. We observe in our analysis of the 8cube data (6) that there are many distinct differences in gene expression between strains. Strain-specific diversity limits the extent to which findings can be generalized. The heavy reliance on inbred strains exacerbates this issue, raising the possibility that observed specificity reflects artifacts of restricted genetic backgrounds rather than broadly conserved mechanisms. These concerns emphasize the danger of extrapolating from a single strain to complex, heterogeneous populations. We have deposited specificity and gene expression information across the 8 diverse founder strains into a publicly available and programmatically accessible API database and website (*Data Availability*), fueling the effective and informed

use of mouse models in biological research.

An important evolutionary question emerges from these observations: how do highly specific genes arise and change across lineages? Is there an adaptive advantage to maintaining a housekeeping role, or can specificity confer selective benefits? Moreover, how can a gene that is highly specialized in one mouse strain acquire a different function and specificity in another? The evolutionary trajectories of marker and housekeeping genes remain poorly understood (12, 79, 80). Addressing these questions may provide key insights into the origins of specificity and the selective pressures shaping gene expression programs across species. By providing a robust and quantitative framework for measuring and dissecting gene specificity, ember offers a way to begin tracing such evolutionary trajectories, enabling comparative analyses of how marker and housekeeping functions emerge, shift, and are maintained across contexts.

Methods

Ψ , ζ , and ψ_{block} scores. Let x_i be the count of a gene in cell i . Define the probability

$$p_i = \frac{x_i}{\sum_n x_j},$$

i.e., the proportion of the gene's total count found in cell i .

The **total entropy** E_T of the gene expression across all

cells is defined as:

$$E_T = - \sum_{i=1}^n p_i \log p_i,$$

where n is the number of cells.

Let n cells be partitioned into r blocks. We define Ψ based on the within-block and between-block decomposition of entropy. And let q_{jC} be the proportion of expression in cell j relative to the total in block C :

$$q_{jC} = \frac{x_j}{\sum_{k \in C} x_k} \quad \text{for } j \in C$$

Then, E_C is the **entropy within the block C** :

$$E_C = - \sum_{j \in C} q_{jC} \log q_{jC}.$$

The total **within entropy** E_W is the weighted sum:

$$E_W = \sum_C p_C E_C.$$

Since the total entropy E_T can be decomposed as:

$$E_T = E_B + E_W,$$

Then, we define **Ψ (Psi) as the fraction of information explained by a partition** of n cells:

$$\Psi = 1 - \frac{E_B}{E_T} = \frac{E_W}{E_T}.$$

We define the decomposable elements of E_W for each block as ψ_{block} , a measure of **specificity to a block**. Then ψ_{block} for block C is given by:

$$\psi_{block} = \frac{E_C}{E_W}.$$

Let (ψ_{blocks}) denote the set of ψ_{block} scores across all (r) blocks:

$$\psi_{blocks} = \psi_{block}^{(C)} : C = 1, \dots, r.$$

We define ζ as the normalized Kullback-Leibler divergence of the distribution of ψ_{blocks} to the uniform distribution. This metric quantifies **specificity to a partition**.

$$\zeta = 1 - \frac{H(\psi_{blocks})}{\log r}$$

Sampling over replicates. To ensure balanced representation of replicates across categories in an experimental design, we implemented a sampling strategy that generates subsets of samples with equal representation of experimental conditions within each category. For every unique category (e.g., genotype, strain, or age group), replicates were iteratively selected so that each condition (e.g., sex) was represented exactly once per sampling draw. The distinction between “category” and “condition” was made for conceptual clarity, although these variables can be defined interchangeably. Sampling was performed at the level of unique sample identifiers

recorded in the sample metadata (e.g., mouse_ids or sample_ids). Multiple balanced subsets (typically 100 draws) were generated while minimizing repeated selection of the same samples across draws by tracking sample usage frequency. Random seeds were used to ensure reproducibility, and the approach enforced complete balance by requiring at least one sample for every category-condition pair. (Supp. Fig 5 a,b,c)

Arithmetic mean and standard deviation values for Ψ and ζ were calculated to capture variation in specificity across biological replicates.

Since ψ_{block} values represent compositional fractions that sum to one within each partition, we aggregated replicate level ψ_{block} estimates across sampling draws using compositional data analysis principles. Specifically, the Aitchison mean was used to compute the central tendency of ψ_{block} values across draws, while the geometric standard deviation quantified variability in block-specific specificity estimates.

Octagon plot representation of block specificity. The octagon plot from Fig 1a was designed to visualize variation in specificity across biological replicates to the 8 different founder mouse strains. Because ψ_{block} values for a given partition represent compositional fractions that sum to one, gene-specificity profiles can be represented as points within a simplex. In this framework, each coordinate corresponds to the proportion of a gene’s specificity attributed to a particular block, such that all coordinates are non-negative and collectively sum to unity. In general, an n -simplex can be projected onto a regular $(n+1)$ -gon, yielding $n!/2$ unique 2D projections. For visualization of strain-level specificity across the eight founder mouse strains, the geometry of the 7-simplex was projected into two dimensions as one of 2,540 possible regular octagonal representations. We chose the projection ordered by distance in single nucleotide variants to B6J mice. Data used to generate this plot is available at the supplementary data link (*Supplementary Information*). Code to generate this compositional octagon plot is in the *Code Availability* section.

Generating p-values and q-values. Significance of Ψ , ψ_{block} , and ζ values were estimated by comparing observed mean values after sampling to null distributions generated from scrambling block labels. For each gene, empirical p -values were computed as the proportion of permuted values greater than or equal to the observed statistic.

All resulting p -values were combined across metrics and subjected to global false discovery rate (FDR) correction using the Benjamini–Hochberg procedure to obtain q -values.

Interpreting specificity from entropy metrics. To identify highly and lowly specific genes, we applied a standardized workflow consisting of data pre-processing, entropy metric calculation, and statistical filtering.

Pre-processing steps. For all three datasets analyzed, genes expressed in fewer than 100 cells were removed. Count data was normalized by the original data generators using the

default normalization workflows implemented in Scanpy and Seurat (81, 82).

Entropy metric workflow. Entropy-based specificity metrics were calculated using the ember Python package (see *Code Availability* for repository link). Gene-level entropy metrics were first computed for each dataset, after which genes with $q - \text{values} < 0.05$ were retained. Genes were then ranked according to their respective entropy metrics, and biologically meaningful thresholds (default 0.5 if no prior knowledge) were applied to identify the most specific genes within each partition or block.

(a) Specific to a block / “Marker” genes Metrics ψ_{block} and Ψ were used to identify genes with high specificity to individual blocks. Genes with the highest ψ_{block} and Ψ values were considered most specific to the selected block (e.g., skeletal muscle satellite cells in gastrocnemius, Fig. 1b).

(b) Specific to a partition Metrics ζ and Ψ were used to identify genes specific to a partition. Genes with the highest ζ and Ψ values were interpreted as most specific to the partition (e.g., gastrocnemius partitioned by cell type, Fig. 1c).

(c) Non-specific to a partition / “Housekeeping” genes Metrics ζ and Ψ were also used to identify non-specific or “housekeeping” genes. Genes with low ζ but high Ψ values were interpreted as least specific to any single partition (e.g., gastrocnemius partitioned by cell type, Fig. 1c).

Finding specificity trends using upset plots. To identify large-scale trends in gene specificity across multiple partitions, we computed Ψ and ζ scores for all relevant partitions, including interaction terms (e.g., two-way, three-way, or four-way combinations). Genes with $q - \text{value} < 0.05$ for both metrics were selected. Genes exhibiting strong specificity within each partition were selected using a biologically informed or default cutoff of 0.5 when no prior information was available. The sets of highly specific genes from each partition were then combined to define gene memberships, which were visualized using upset plots to reveal overlaps and sets of genes with shared specificity patterns across partitions (62). Data used to generate upset plots is available at the supplementary data link (*Supplementary Information*). Code to generate upset plots is in the *Code Availability* section.

Data availability

Mouse specificity explorer website for visualizing specificity and gene expression in the 8cube data (6): <https://mouseexplorer.onrender.com>. 8cubeDB API documentation: <https://eightcubedb.onrender.com/docs>. 8cubeDB API and mouse explorer website source code: <https://github.com/pachterlab/8cubeDB>. ember python package and documentation: <https://github.com/pachterlab/ember>.

Accessing raw data. The accessibility details of the 8cube dataset are described in the Data Availability section of (6)

and of the developmental kidney dataset in the Data Availability section of (7). The human PBMC data is available from the [IGVF portal](#) by searching for Jimmie Ye, UCSF. The Ham *et al.* 2025 data accessibility details are described in their Data Availability section (36).

Code availability. The code for reproducing the figures and analyses: https://github.com/pachterlab/SBRGKLAYWMP_2025. Data to reproduce the figures and analyses: <https://doi.org/10.5281/zenodo.17654220>.

Author contributions

N.P.S. contributed to conceptualization, performed analyses, developed the ember python package, mouse explorer website and 8cubeDB API, made figures, and drafted the manuscript. S.B. contributed to conceptualization. E.R. generated the 8-cube founder dataset and assisted with cell type annotations. M.G.G., P.K., T.L., M.A., and C.J.Y. generated the human PBMC dataset and performed cell type annotations. B.W. and A.M. contributed to supervision, data acquisition, and funding acquisition. L.P. contributed to conceptualization, supervision, and funding acquisition. All authors contributed to manuscript review.

Acknowledgments

We thank Conrad Oakes and Joseph Rich for help with optimizing alluvial plots. We thank Ruth Pachter and Juliet Lee for help with interpreting the coding and protein sequences for *Tas1r1*.

Funding

This work was funded by grant NIH UM1HG012077.

Competing Interests

No competing interests.

References

- Junyue Cao, Jonathan S. Packer, Vijay Ramani, Darren A. Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, Choli Lee, Scott N. Furlan, Frank J. Steemers, Andrew Adey, Robert H. Waterston, Cole Trapnell, and Jay Shendure. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, 357(6352):661–667, 2017. doi: 10.1126/science.aam8940.
- Alexander B. Rosenberg, Charles M. Roco, Richard A. Muscat, Anna Kuchina, Paul Sample, Zizhen Yao, Lucas T. Graybuck, David J. Peeler, Sumit Mukherjee, Wei Chen, Stephanie H. Pun, David L. Sellers, Bosiljka Tasic, and Georg Seelig. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, 360(6385):176–182, 2018. doi: 10.1126/science.aam8999.
- Shristi Pandey, Kimberle Shen, Seung-Hye Lee, Christopher J. Bohlen, Tracy J. Yuen, Brad A. Friedman, et al. Disease-associated oligodendrocyte responses across neurodegenerative diseases. *Cell Reports*, 40(8):111234, 2022. doi: 10.1016/j.celrep.2022.111234. Available at <http://research-pub.gene.com/OligoLandscape/>.
- Stefan Salcher, Gregor Sturm, Lena Horvath, Dominik Wolf, Andreas Pircher, and Zlatko Trajanoski. High-resolution single-cell atlas reveals diversity and plasticity of tissue-resident neutrophils in non-small cell lung cancer. *Cancer Cell*, 40(12):1503–1520.e8, 2022. doi: 10.1016/j.ccell.2022.10.010.
- Austin D. Reed, Sara Pensa, Adi Steif, Jack Stenning, Daniel J. Kunz, Linsey J. Porter, Kui Hua, Peng He, Alecia-Jane Twigger, Abigail J. Q. Siu, Katarzyna Kania, Rachel Barrow-McGee, Iain Goulding, Jennifer J. Gomm, Valerie Speirs, J. Louise Jones, John C. Marioni, and Walid T. Khaled. A single-cell atlas enables mapping of homeostatic cellular shifts in the adult human breast. *Nature Genetics*, 56:652–662, 2024. doi: 10.1038/s41588-024-01726-x. Open access.
- Elisabeth Rebboah, Ryan Weber, Elnaz Abdollahzadeh, Nikhila Swarna, Delaney K. Sullivan, Diane Trout, Fairlie Reese, Heidi Yahan Liang, Ghassan Filimban, Parvin Mahdipoor, Margaret Duffield, Romina Mojaverzargar, Erisa Taghizadeh, Negar Fattahi, Negar Mojtani, Haoran Zhang, Rebekah K. Loving, Maria Carrilli, A. Sina Boeshaghi, Shimako Kawauchi, Ingileif B. Hallgrímsdóttir, Brian A. Williams, Grant R. MacGregor, Lior Pachter, Barbara J. Wold, and Ali Mortazavi. Systematic cell-type resolved transcriptomes of 8 tissues in 8 lab and wild-derived mouse strains captures global and local expression variation. *bioRxiv*, 04 2025. doi: <https://doi.org/10.1101/2025.04.21.649844>. Preprint, not peer reviewed.
- Siqi Chen, Ruiyang Liu, Chia-Kuei Mo, Michael C. Wendt, Andrew Houston, Preet Lal, Yanyan Zhao, Wagma Caravan, Andrew T. Shinkle, Atieh Abedin-Do, Nataly Naser Al Deen, Kazuhito Sato, Xiang Li, André Luiz N. Targino da Costa, Yize Li, Alla Karpova, John M. Herndon, Maxim N. Artyomov, Joshua B. Rubin, Sanjay Jain, Xue Li, Sheila A. Stewart, Li Ding, and Feng Chen. Multi-omic and spatial analysis of mouse kidneys highlights sex-specific differences in gene regulation across the lifespan. *Nature Genetics*, 57:1213–1227, 04 2025. doi: <https://doi.org/10.1038/s41588-025-02161-x>.
- Ailsa M. Jeffries, Tianxiang Yu, Jennifer S. Ziegenfuss, Allie K. Tolles, Christina E. Baer, Cesar Bautista Sotelo, Yerin Kim, Zhiping Weng, and Michael A. Lodato. Single-cell transcriptomic and genomic changes in the ageing human brain. *Nature*, 2025. doi: 10.1038/s41586-025-09435-8.
- David Lahnemann, Johannes Köster, Ewa Szczurek, Davis J. McCarthy, Stephanie C. Hicks, Mark D. Robinson, Catalina A. Vallejos, Kieran R. Campbell, Niko Beerenwinkel, Ahmed Mahfouz, Luca Pinello, Pavel Skums, Alexandros Stamatakis, Camille Stephan-Otto Attolini, Samuel Aparicio, Jasmijn Baaijens, Marleen Balvert, Buys de Barbanson, Antonio Cappuccino, Giacomo Corleone, Bas E. Dutilh, Maria Florescu, Victor Guryev, Rens Holmer, and Alexander Schönhuth. Eleven grand challenges in single-cell data science. *Genome Biology*, 21:31, 2020. doi: 10.1186/s13059-020-1926-6.
- Malte D. Luecken and Fabian J. Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746, 2019. doi: 10.15252/msb.20188746.
- Tim Stuart and Rahul Satija. Integrative single-cell analysis. *Nature Reviews Genetics*, 20: 257–272, 2019. doi: 10.1038/s41576-019-0093-7.
- Itai Yanai, Hila Benjamin, Michael Shmoish, Vered Chalifa-Caspi, Maxim Shklar, Ron Ophir, Arren Bar-Even, Shirley Horn-Saban, Marilyn Safran, Eytan Domany, Doron Lancet, and Orit Shmueli. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, 21(5):650–659, 2005. doi: 10.1093/bioinformatics/bti042. All data available at <http://genecards.weizmann.ac.il/genenote/>; GEO accession GSE803.
- Anupama Roy, Himanshushekhkar Chaurasia, Baibhav Kumar, Naina Kumari, Sarika Jaiswal, Manish Srivastava, Mir Asif Iqbal, Ulavappa B. Angadi, and Dinesh Kumar. Feat: a comprehensive web-based expression atlas for functional genomics in tropical and sub-tropical fruit crops. *BMC Plant Biology*, 24:890, 2024. doi: 10.1186/s12870-024-05022-9.
- Lukasz Huminiński, Andrew T. Lloyd, and Kenneth H. Wolfe. Congruence of tissue expression profiles from gene expression atlas, sagemap and tissueinfo databases. *BMC Genomics*, 4:31, 2003. doi: 10.1186/1471-2164-4-31.
- Alec Barrett, Erdem Varol, Alexis Weinreb, Seth R. Taylor, Rebecca M. McWhirter, Cyril Cros, Berta Vidal, Manasa Basaravaju, Abigail Poff, Marc Hammarlund, et al. Integrating bulk and single cell rna-seq refines transcriptional profiles of individual C. elegans neurons. *eLife*, v1, 2025. doi: 10.7554/eLife.106183.1. Reviewed Preprint, published April 11, 2025.
- Lidia Ceriani and Paolo Verme. The origins of the gini index: extracts from variability and mutability (1912) by corrado gini. *The Journal of Economic Inequality*, 10:421–443, 2012. doi: 10.1007/s10888-011-9188-x.
- Steve O'Hagan, Marina Wright Muelas, Philip J. Day, Emma Lundberg, and Douglas B. Kell. Genegini: Assessment via the gini coefficient of reference “housekeeping” genes and diverse human transporter expression profiles. *Cell Systems*, 6(2):230–244.e1, 2018. doi: 10.1016/j.cels.2018.01.002.
- Nadezda Kryuchkova-Mostacci and Marc Robinson-Rechavi. A benchmark of gene expression tissue-specificity metrics. *Briefings in Bioinformatics*, 18(2):205–214, 2017. doi: 10.1093/bib/bbw008.
- Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15:550, 2014. doi: 10.1186/s13059-014-0550-8.
- Yunshun Chen, Lizhong Chen, Aaron T. L. Lun, Pedro L. Baldoni, and Gordon K. Smyth. edgeR v4: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets. *Nucleic Acids Research*, 53:gkaf018, 2025. doi: 10.1093/nar/gkaf018.
- Dmitrii Konstantinovich Faddeev. On the concept of entropy of a finite probabilistic scheme. *Uspekhi Matematicheskikh Nauk*, 11(1):227–231, 1956.
- A. Ya. Khinchin. *Mathematical Foundations of Information Theory*. Dover Publications, Mineola, New York, 2013. ISBN 9780486318448.
- Michael J. Casey, Jörg Fliege, Rubén J. Sánchez-García, and Ben D. MacArthur. An information-theoretic approach to single cell sequencing analysis. *BMC Bioinformatics*, 24: 311, 2023. doi: 10.1186/s12859-023-05331-0.
- Michael C. Saul, Vivek M. Philip, Laura G. Reinholdt, Elissa J. Chesler, and Center for Systems Neurogenetics of Addiction. High-diversity mouse populations for complex traits. *Trends in Genetics*, 35(7):501–514, July 2019. doi: 10.1016/j.tig.2019.04.003.
- Joseph Rich, Conrad Oakes, and Lior Pachter. Optimizing alluvial plots. *arXiv preprint arXiv:2509.03761*, September 2025. doi: 10.48550/arXiv.2509.03761.
- John P. Didion, Andrew P. Morgan, Amelia M.-F. Clayshulte, Rachel C. McMullan, Liran Yadgary, Petko M. Petkov, Timothy A. Bell, Daniel M. Gatti, James J. Crowley, Kunjie Hua, David L. Aylor, Ling Bai, Mark Calaway, Elissa J. Chesler, John E. French, Thomas R. Geiger, Terry J. Gooch, Theodore Jr. Garland, Alison H. Harrill, Kent Hunter, Leonard McMillan, Matt Holt, Darla R. Miller, Deborah A. O'Brien, Kenneth Paigen, Wenqi Pan, Lucy B. Rowe, Ginger D. Shaw, Petr Simecek, Patrick F. Sullivan, Karen L. Svenson, George M. Weinstock, David W. Threadgill, Daniel Pomp, Gary A. Churchill, and Fernando Pardo-Manuel de Villena. A multi-megabase copy number gain causes maternal transmission ratio distortion on mouse chromosome 2. *PLoS Genetics*, 11(2):e1004850, 02 2015. doi: <https://doi.org/10.1371/journal.pgen.1004850>.
- Anneke Gässler, Charline Quiclet, Oliver Kluth, Pascal Gottmann, Kristin Schwerbel, Anett Helms, Mandy Stadion, Ilka Wilhelmi, Wenke Jonas, Meriem Ouni, Frank Mayer, Joachim Spranger, Annette Schürmann, and Heike Vogel. Overexpression of gjb4 impairs cell proliferation and insulin secretion in primary islet cells. *Molecular Metabolism*, 41:101042, 06 2020. doi: <https://doi.org/10.1016/j.molmet.2020.101042>.
- Yoko Kusuhara, Ryusuke Yoshida, Tadahiro Ohkuri, Keiko Yasumatsu, Anja Voigt, Sandra Hübner, Katsumasa Maeda, Ulrich Boehm, Wolfgang Meyerhof, and Yuzo Ninomiya. Taste responses in mice lacking taste receptor subunit tr1r. *The Journal of Physiology*, 591(7): 1967–1985, 2013. doi: <https://doi.org/10.1113/jphysiol.2012.236604>.
- Anthony Scalfani, Austin S. Vural, and Karen Ackroff. Cast/eij and c57bl/6j mice differ in their oral and postoral attraction to glucose and fructose. *Chemical Senses*, 42(3):259–267, 02 2017. doi: <https://doi.org/10.1093/chemse/bjx003>.
- Junli Liu, Qilin Li, Yixuan Hu, Yi Yu, Kai Zheng, Dengfeng Li, Lexin Qin, and Xiaochun Yu. The complete telomere-to-telomere sequence of a mouse genome. *Science*, 386(6726): 1141–1146, 12 2024. doi: <https://doi.org/10.1126/science.adq8191>.
- Bailey Francis, Landen Gozashti, Kevin Costello, Takaaki Kasahara, Olivia S. Harrington, Jingtao Lilue, Mohab Helmy, Tadamu Kato, Anne Czechanski, Michael A. Quail, Iraad Bonner, Emma Dawson, Anne Ferguson-Smith, Laura Reinholdt, David J. Adams, and Thomas M. Keane. The structural diversity of telomeres and centromeres across mouse subspecies revealed by complete assemblies. *bioRxiv*, 10 2024. doi: <https://doi.org/10.1101/2024.10.24.619615>. Preprint, not peer reviewed.
- Maria Rigau, David Juan, Alfonso Valencia, and Daniel Rico. Intronic cnvs and gene expression variation in human populations. *PLoS Genetics*, 15(1):e1007902, 01 2019. doi: <https://doi.org/10.1371/journal.pgen.1007902>.
- Sheriar G. Hormuzdi, Risto Penttinen, Rudolf Jaenisch, and Paul Bornstein. A gene-targeting approach identifies a function for the first intron in expression of the $\alpha 1(i)$ collagen gene. *Molecular and Cellular Biology*, 18(6):3368–3375, 06 1998. doi: <https://doi.org/10.1128/mcb.18.6.3368>.
- Jin Wang, Xiaoyu Xi, Shifeng Zhao, Xiaolei Wang, Lixia Yao, Jinlin Feng, and Rong Han. Introns in the naa50 gene act as strong enhancers of tissue-specific expression in arabidopsis. *Plant Science*, 324:111422, 11 2022. doi: <https://doi.org/10.1016/j.plantsci.2022.111422>.
- Masahiko Yamaguchi, Yoko Watanabe, Takuji Ohtani, Shin'ichi Takeda, Hiroshi Yamamoto, and So-ichiro Fukada. Calcitonin receptor signaling inhibits muscle stem cells from escaping the quiescent state and the niche. *Cell Reports*, 13(2):302–314, 10 2015. doi: <https://doi.org/10.1016/j.celrep.2015.08.064>.
- Alexander S. Ham, Shuo Lin, Alice Tse, Marco Thürkauf, Timothy J. McGowan, Lena Jörin, Filippo Oliveri, and Markus A. Rüegg. Single-nuclei sequencing of skeletal muscle reveals subsynaptic-specific transcripts involved in neuromuscular junction maintenance. *Nature Communications*, 16:2220, 03 2025. doi: <https://doi.org/10.1038/s41467-025-12017-0>.
- Robert D. Barber, Dan W. Harmer, Robert A. Coleman, and Brian J. Clark. Gapdh as a housekeeping gene: analysis of gapdh mrna expression in a panel of 72 human tissues. *Physiological Genomics*, 21(3):389–395, 05 2005. doi: <https://doi.org/10.1152/physiolgenomics.00025.2005>.
- Jianqiang Bao, Shuiqiao Yuan, Ashley Maestas, Bhupal P. Bhetwal, Andrew Schuster, and Wei Yan. Stk31 is dispensable for embryonic development and spermatogenesis in mice. *Molecular Reproduction and Development*, 80(10):786, 2013. doi: 10.1002/mrd.22225.
- V. A. Traag, Ludo Waltman, and Nees Jan van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9:5233, 03 2019. doi: <https://doi.org/10.1038/s41598-019-41695-z>.
- D. Djureinovic, L. Fagerberg, B. Hallström, A. Danielsson, C. Lindskog, M. Uhlen, and F. Pontén. The human testis-specific proteome defined by transcriptomics and antibody-based profiling. *Molecular Human Reproduction*, 20(6):476–488, 06 2014. doi: <https://doi.org/10.1093/molehr/gau018>.
- Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. g:profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, 47(W1):W191–W198, 07 2019. doi: <https://doi.org/10.1093/nar/gkz369>.
- Inna Gitelman. Twist protein in mouse embryogenesis. *Developmental Biology*, 189(2): 205–214, 09 1997. doi: <https://doi.org/10.1006/dbio.1997.8614>.
- Hector L. Franco, José Casanovas, José R. Rodríguez-Medina, and Carmen L. Cadilla.

- Redundant or separate entities?—roles of twist1 and twist2 as molecular switches during gene transcription. *Nucleic Acids Research*, 39(4):1177–1186, 10 2010. doi: <https://doi.org/10.1093/nar/gkq890>.
44. Jiata Ren and Steven D. Crowley. Twist1: A double-edged sword in kidney diseases. *Kidney Diseases*, 6(4):247–257, 01 2020. doi: <https://doi.org/10.1159/000505188>.
 45. Qingfeng Wu, Shiren Sun, Lei Wei, Minna Liu, Hao Liu, Ting Liu, Ying Zhou, Qing Jia, Di Wang, Zhen Yang, Menglu Duan, Xiaoxia Yang, Peisong Gao, and Xiaoxuan Ning. Twist1 regulates macrophage plasticity to promote renal fibrosis through galectin-3. *Cellular and Molecular Life Sciences*, 79(3):137, 02 2022. doi: <https://doi.org/10.1007/s00182-022-04137-0>.
 46. Lianqin Sun, Lishan Liu, Juanjuan Jiang, Kang Liu, Jingfeng Zhu, Lin Wu, Xiaohan Lu, Zhimin Huang, Yanggang Yuan, Steven D. Crowley, Huijuan Mao, Changying Xing, and Jiata Ren. Transcription factor twist1 drives fibroblast activation to promote kidney fibrosis via signaling proteins prrx1/tnc. *Kidney International*, 106(5):840–855, 11 2024. doi: <https://doi.org/10.1016/j.kint.2024.07.028>.
 47. Yan Xu, Yixiang Xu, Lan Liao, Niya Zhou, Sarah M. Theissen, Xin-Hua Liao, Hoang Nguyen, Thomas Ludwig, Li Qin, Jarrod D. Martinez, Jun Jiang, and Jianming Xu. Inducible knock-out of twist1 in young and adult mice prolongs hair growth cycle and has mild effects on general health, supporting twist1 as a preferential cancer target. *The American Journal of Pathology*, 183(4):1281–1292, 10 2013. doi: <https://doi.org/10.1016/j.ajpath.2013.06.021>.
 48. Jibin Zhang, Jinsoo Ahn, Yeunsu Suh, Seongsoo Hwang, Michael E. Davis, and Kichoon Lee. Identification of CTLA2A, DEFB29, WFDC15B, SERPINA1F and MUP19 as novel tissue-specific secretory factors in mouse. *PLOS ONE*, 10(5):e0124962, 05 2015. doi: <https://doi.org/10.1371/journal.pone.0124962>.
 49. Xiao Wang, Fanyi Qiu, Junjie Yu, Meiyang Zhou, Anjian Zuo, Xiaojing Xu, Xiao-Yang Sun, and Zhengpin Wang. Transcriptome profiling of the initial segment and proximal caput of mouse epididymis. *Frontiers in Endocrinology*, 14:1190890, 05 2023. doi: <https://doi.org/10.3389/fendo.2023.1190890>.
 50. Shuanggang Hu, Guangxin Yao, Xiangqi Li, Zijiang Chen, and Yun Sun. Androgen receptor binding sites identified in mouse testis. *Acta Biochimica et Biophysica Sinica*, 45(8):795–797, 2013. doi: <https://doi.org/10.1093/abbs/gmt076>. Advance Access Publication 9 July 2013.
 51. Steve Dorus, Elizabeth R. Wasbrough, Jennifer Busby, Elaine C. Wilkin, and Timothy L. Karr. Sperm proteomics reveals intensified selection on mouse sperm membrane and acrosome genes. *Molecular Biology and Evolution*, 27(6):1235–1246, 06 2010. doi: <https://doi.org/10.1093/molbev/msq007>.
 52. Hui Qian, Xing Deng, Zhao-Wei Huang, Ji Wei, Chen-Hong Ding, Ren-Xin Feng, Xin Zeng, Yue-Xiang Chen, Jin Ding, Lei Qiu, Zhen-Lin Hu, Xin Zhang, Hong-Yang Wang, Jun-Ping Zhang, and Wei-Fen Xie. An hnf1 α -regulated feedback circuit modulates hepatic fibrogenesis via the crossstalk between hepatocytes and hepatic stellate cells. *Cell Research*, 25: 930–945, 07 2015. doi: <https://doi.org/10.1038/cr.2015.82>.
 53. Taro E. Akiyama, Jerrold M. Ward, and Frank J. Gonzalez. Regulation of the liver fatty acid-binding protein gene by hepatocyte nuclear factor 1 α (hnf1 α): Alterations in fatty acid homeostasis in hnf1 α -deficient mice. *Journal of Biological Chemistry*, 275(35):27117–27122, 09 2000. doi: [https://doi.org/10.1016/S0021-9258\(19\)61487-0](https://doi.org/10.1016/S0021-9258(19)61487-0).
 54. Qi Ni, Kai Ding, Ke-Qi Wang, Jin He, Chuan Yin, Jian Shi, Xin Zhang, Wei-Fen Xie, and Yong-Quan Shi. Deletion of hnf1 α in hepatocytes results in fatty liver-related hepatocellular carcinoma in mice. *FEBS Letters*, 591(11):1481–1491, 05 2017. doi: <https://doi.org/10.1002/1873-3468.12689>.
 55. David Q. Shih, Seamus Screenan, Karla N. Munoz, Lou Philipson, Marco Pontoglio, Moshe Yaniv, Kenneth S. Polonsky, and Markus Stoffel. Loss of hnf-1 α function in mice leads to abnormal expression of genes involved in pancreatic islet development and metabolism. *Diabetes*, 50(11):2472–2480, 11 2001. doi: <https://doi.org/10.2337/diabetes.50.11.2472>.
 56. Marco Pontoglio, Dominique Prié, Claire Cheret, Antonia Doyen, Christine Leroy, Philippe Froguel, Gilberto Velho, Moshe Yaniv, and Gérard Friedlander. Hnf1 α controls renal glucose reabsorption in mouse and man. *EMBO Reports*, 1(4):359–365, 2000. doi: <https://doi.org/10.1093/embo-reports/kv0071>.
 57. Giorgia Benegiamo, Giacomo V. G. von Alvensleben, Sandra Rodríguez-López, Ludger J. E. Goeminne, Alexis M. Bachmann, Jean-David Morel, Ellen Broeckx, Jing Ying Ma, Vinicius Carneira, Sameh A. Youssef, Nabil Azhar, Dermot F. Reilly, Katharine D’Aquino, Shannon Mullican, Maroun Bou-Sleiman, and Johan Auwerx. The genetic background shapes the susceptibility to mitochondrial dysfunction and nash progression. *Journal of Experimental Medicine*, 220(4):e20221738, 04 2023. doi: <https://doi.org/10.1084/jem.20221738>.
 58. Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell Systems*, 1(6):417–425, 12 2015. doi: <https://doi.org/10.1016/j.cels.2015.12.004>.
 59. Gregory R. Keele, Ji-Gang Zhang, John Szpyt, Ron Korstanje, Steven P. Gygi, Gary A. Churchill, and Devin K. Schweppe. Global and tissue-specific aging effects on murine proteomes. *Cell Reports*, 42(7):112715, 07 2023. doi: <https://doi.org/10.1016/j.celrep.2023.112715>.
 60. Esther M. Speer, Atilade A. Adediji, Joyce Lin, Alexandra Khorasanchi, Asma Rasheed, Maya Bhat, Kelly Mackenzie, Randolph Hennigar, Kimberly J. Reidy, and Robert P. Woronicki. Attenuation of acute kidney injury in a murine model of neonatal *Escherichia coli* sepsis. *Frontiers in Cellular and Infection Microbiology*, 14:1507914, 2025. doi: <https://doi.org/10.3389/fcimb.2024.1507914>.
 61. Akinobu Ochi, Dong Chen, Wibke Schulte, Lin Leng, Nickolas Moeckel, Marta Piecychna, Luisa Averdunk, Christian Stoppe, Richard Bucala, and Gilbert Moeckel. Mif-2/d-dt enhances proximal tubular cell regeneration through slpi- and atf4-dependent mechanisms. *American Journal of Physiology-Renal Physiology*, 313(3):F767–F780, 05 2017. doi: <https://doi.org/10.1152/ajprenal.00683.2016>.
 62. Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. Upset: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1983–1992, Dec 2014. ISSN 1941-0506. doi: <https://doi.org/10.1109/TVCG.2014.2346248>.
 63. Sha Li, Chenghai Wang, Xiaxia Zhang, and Wen Su. Cytochrome p450 omega-hydroxylase 4a14 attenuates cholestatic liver fibrosis. *Frontiers in Physiology*, 12:688259, 05 2021. doi: <https://doi.org/10.3389/fphys.2021.688259>.
 64. Lingyun Xiong, Jing Liu, Seung Yub Han, Junhyong Kim, Adam L. MacLean, and Andrew P. McMahon. Direct androgen receptor control of sexually dimorphic gene expression in the mammalian kidney. *Developmental Cell*, 58(21):2338–2358.e5, 11 2023. doi: <https://doi.org/10.1016/j.devcel.2023.09.019>.
 65. Baoping Hu, Yuhe Wang, Zhongtao Wang, Xue He, Li Wang, Dongya Yuan, Yongjun He, Tianbo Jin, and Shumei He. Association of slc11a1 polymorphisms with tuberculosis susceptibility in the chinese han population. *Frontiers in Genetics*, 13:899124, July 2022. doi: <https://doi.org/10.3389/fgene.2022.899124>.
 66. Kazunobu Ouchi, Yoichi Suzuki, Taro Shirakawa, and Fumio Kishi. Polymorphism of slc11a1 (formerly nramp1) gene confers susceptibility to kawasaki disease. *Journal of Infectious Diseases*, 187(2):326–329, January 2003. doi: <https://doi.org/10.1086/345878>.
 67. Mariam Sophie, Abdul Hameed, Akhtar Muneer, Azam J. Samdani, Saima Saleem, and Abid Azhar. Slc11a1 polymorphisms and host susceptibility to cutaneous leishmaniasis in pakistan. *Parasites & Vectors*, 10(12):12, January 2017. doi: <https://doi.org/10.1186/s13071-016-1954-y>.
 68. Zahra Kavian, Saman Sargazi, Mahdi Majidpour, Mohammad Sarhadi, Ramin Saravani, Mansour Shahraki, Shekoufeh Mirinejad, Milad Heidari Nia, and Maryam Piri. Association of slc11a1 polymorphisms with anthropometric and biochemical parameters describing type 2 diabetes mellitus. *Scientific Reports*, 13:6195, April 2023. doi: <https://doi.org/10.1038/s41598-023-33469-y>.
 69. Jenefer M. Blackwell, Tapasree Goswami, Carlton A. W. Evans, Dean Sibthorpe, Natalie Papo, Jacqueline K. White, Susan Searle, E. Nancy Miller, Christopher S. Peacock, Hiba Mohammed, and Muntaser Ibrahim. Slc11a1 (formerly nramp1) and disease resistance. *Cellular Microbiology*, 3(12):773–784, December 2001. doi: <https://doi.org/10.1046/j.1462-5822.2001.00150.x>.
 70. Jorge Eliécer Mario-Vásquez, Carlos Andrés Naranjo-González, Jehidy Montiel, Lina M. Zuluga, Ana M. Vásquez, Alberto Tobón-Castaño, Gabriel Bedoya, and Cesar Segura. Association of variants in il1b, tlr9, trem1, il10ra, and cd3g and native american ancestry on malaria susceptibility in colombian populations. *Infection, Genetics and Evolution*, 87: 104675, January 2021. doi: <https://doi.org/10.1016/j.meegid.2020.104675>.
 71. Dan N. Predescu, Babak Mokhlesi, and Sanda A. Predescu. X-inactive-specific transcript: a long noncoding rna with a complex role in sex differences in human disease. *Biology of Sex Differences*, 15:101, December 2024. doi: <https://doi.org/10.1186/s13293-024-00681-5>. Review article.
 72. Eli Eisenberg and Erez Y. Levanon. Human housekeeping genes, revisited. *Trends in Genetics*, 29(10):569–574, 2013. doi: <https://doi.org/10.1016/j.tig.2013.05.010>.
 73. Chintan J. Joshi, Wenfan Ke, Anna Drangowska-Way, Eyleen J. O’Rourke, and Nathan E. Lewis. What are housekeeping genes? *PLOS Computational Biology*, 18(7):e1010295, July 2022. doi: <https://doi.org/10.1371/journal.pcbi.1010295>. Version 2, Open Access, Peer-reviewed.
 74. Roy D. Dar, Brandon S. Razooyk, Abhyudai Singh, and Leor S. Weinberger. Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proceedings of the National Academy of Sciences*, 109(43):17454–17459, 10 2012. doi: <https://doi.org/10.1073/pnas.1213530109>.
 75. Ido Golding, Johan Paulsson, Scott M. Zawilski, and Edward C. Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–1036, 12 2005. doi: <https://doi.org/10.1016/j.cell.2005.09.031>.
 76. Martin Rosvall and Carl T. Bergstrom. Mapping change in large networks. *PLOS ONE*, 5(1):e8694, January 2010. doi: <https://doi.org/10.1371/journal.pone.0008694>.
 77. Tara Chari and Lior Pachter. The specious art of single-cell genomics. *PLOS Computational Biology*, 19(8):e1011288, August 2023. doi: <https://doi.org/10.1371/journal.pcbi.1011288>.
 78. Yoav Gilad and Orna Mizrahi-Man. A reanalysis of mouse encode comparative gene expression data. *F1000Research*, 4:121, 2015. doi: <https://doi.org/10.12688/f1000research.6536.1>.
 79. David Brawand, Magali Soumillon, Anamaria Necseulea, Philippe Julien, Gábor Csárdi, Patrick Harrigan, Manuela Weier, Angélica Liechti, Ayinuer Aximu-Petri, Martin Kircher, Frank W. Albert, Ulrich Zeller, Philipp Khaitovich, Frank Grützner, Sven Bergmann, Rasmus Nielsen, Svante Pääbo, and Henrik Kaessmann. The evolution of gene expression levels in mammalian organs. *Nature*, 478:343–348, 2011. doi: <https://doi.org/10.1038/nature10532>.
 80. Anamaria Necseulea and Henrik Kaessmann. Evolutionary dynamics of coding and non-coding transcriptomes. *Nature Reviews Genetics*, 15:734–748, 2014. doi: <https://doi.org/10.1038/nrg3802>.
 81. F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology*, 19:15, 2018. doi: <https://doi.org/10.1186/s13059-017-1382-0>.
 82. Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Eftymia Papalexi, William M. III Mauck, Yuhuan Hao, Marlon Stoekius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.e21, 2019. doi: <https://doi.org/10.1016/j.cell.2019.05.031>.