מכון ויצמן למדע

WEIZMANN INSTITUTE OF SCIENCE

Thesis for the degree

Master of Science

Submitted to the Scientific Council of the
Weizmann Institute of Science
Rehovot, Israel

עבודת גמר (תזה) לתואר

מוסמך למדעים

מוגשת למועצה המדעית של
מכון ויצמן למדע
רחובות, ישראל

By
Maayan Klein Yoles

מאת
מעין קליין יולס

הספציפיות הרקמתית של גנים אנושיים בהיבטים של תכולת מידע ואבולוציה

Tissue Specificity of Human Genes Through the Lens of Information Content and Evolution

Advisor:
Prof. Yitzhak Pilpel

מנחה:
פרופ' יצחק פלפל

January 2026

שבט תשפ"ו

Table of Content

Abstract

Gene expression spans from broadly expressed housekeeping genes to highly tissue-specific ones, yet a large fraction of genes exhibit intermediate expression patterns, with elevated expression in some tissues and repression in others. How regulatory features scale across this spectrum, and whether intermediate expression patterns reflect distinct regulatory demands, remains unclear.

In this thesis, we investigate tissue specificity of human genes through the lens of expression complexity and information content. Using genome-wide expression data, we quantified tissue specificity with the tau tissue specificity score and systematically examined how diverse regulatory features scale across the full specificity range. Rather than focusing on individual genes, we compared regulatory architectures relative to one another, aiming to identify global trends governing gene regulation.

We find a consistent, non-monotonic relationship between tissue specificity and regulatory complexity. Genes with intermediate tissue specificity exhibit the highest regulatory information content, including longer untranslated regions and increased enhancer associations, whereas both broadly expressed and highly tissue-specific genes display simpler regulatory architectures. These patterns are robust to alternative specificity metrics.

Incorporating an evolutionary perspective reveals that tissue specificity and regulatory complexity are also shaped by gene age, with older genes tending toward broader expression across body tissues and intermediate-aged genes exhibiting the strongest signatures of regulatory complexity. Extending the analysis to mouse further shows that the relationship between expression complexity and regulatory information is conserved across species, despite divergence in individual regulatory features. Together, these findings suggest that tissue specificity reflects the informational demands of gene expression programs, and that regulatory architectures scale accordingly.

Introduction

Tissue specificity as a fundamental property of genes

Gene expression in each tissue and cell type in our body is highly intricate and regulated, enabling genetically identical cells that all emerged from the same gamete to acquire distinct identities and functions. The human genome contains approximately 20,000 protein-coding genes, whose expression patterns range from ubiquitous across nearly all tissues to highly restricted and tissue-specific. Genes are often classified as either housekeeping, meaning ubiquitously expressed in all cells and performing essential and core cellular functions, or tissue-specific, associated with specialized roles and functions.[1–3] Between these two extremes lies a large and diverse class of genes with intermediate expression breadth: genes that are highly expressed in some tissues but weakly expressed or silent in others. Despite constituting a substantial fraction of the genome, these intermediately expressed genes have received less systematic attention, and their biological roles and regulatory logic remain less well characterized.

Quantifying expression breadth and tissues specificity

To move beyond the simple classification of genes as either housekeeping or tissue-specific, Yanai et al. introduced the tau score, a quantitative measure of tissue specificity ranging from 0 to 1, using the formula:

$$\tau = \frac{\sum_{i=1}^{N}(1 - x_i)}{N - 1}$$

In the formula, N is the number of tissues, x_i is the normalized *log₁₀*-transformed expression data in each tissue, divided by the maximum expression value if the gene.

Tau score values close to 0 indicate ubiquitous expression and values close to 1 indicate strong tissue specificity. Importantly, many genes exhibit intermediate tau values, reflecting more complex expression patterns. Yanai et al. further demonstrated that this mid-range class comprises a substantial portion of the genes in the genome and suggested that it may encode critical functional information.[4]

Regulatory architecture as an information-encoding system

The expression pattern of each gene is encoded by a combination of regulatory elements that act at multiple levels. These include cis-regulatory elements, such as promoters, enhancers, untranslated regions, and introns, as well as trans-acting factors, including transcription factors and microRNAs. Together, these components determine where, when, and to what extent a gene is expressed. While many studies have examined individual regulatory features in isolation, a comprehensive view of how regulatory architecture scales with expression complexity is still lacking.

Minimum description length and information content in gene regulation

The principle of minimum description length (MDL) was introduced under information theory. MDL states that the best representation of a system is the one that minimizes the amount of information required to describe it[5,6]. More generally, MDL formalizes the idea that simpler, more regular patterns are more compressible, whereas complex patterns require longer descriptions. Applied to gene expression, we can examine the expression patterns of genes and ask how complex or compressible they are. This framework allows expression patterns to be viewed as informational objects with varying complexity. Genes with simple expression patterns, such as uniform expression across tissues or expression restricted to a single tissue, can be described concisely by simple rules – for example, "express in all tissues" or "express only in tissue X". And in contrast, genes exhibiting heterogeneous expression across multiple tissues require more detailed descriptions – for instance, “express in tissues A, D, and F but not or lowly in B, C, or E”.

From an MDL perspective, such complex expression patterns are expected to require richer regulatory architectures and elaborate control mechanisms to encode and maintain them. This conceptual framework leads to the central hypothesis of this thesis: the complexity of a gene’s expression pattern is reflected in the amount and structure of regulatory information required to describe it.

Viewing gene regulation through the lens of MDL provides a principled way to connect biological complexity with informational and evolutionary cost. This framing allows gene regulation to be viewed as an information-encoding system, and tissue specificity as a proxy for expression complexity.

Aims of the thesis

The aim of this thesis is to investigate gene expression tissue specificity through the lens of information theory, and specifically the principle of minimum description length. We examine whether the complexity of gene expression patterns, as quantified by the tau score, is reflected in the amount and organization of regulatory information required to encode them. To this end, we systematically analyze a wide range of cis- and trans-regulatory features across the full spectrum of tissue specificity in the human genome. We further assess the robustness and biological relevance of these patterns using paralogous genes, predictive modeling, and evolutionary analyses. Finally, we extend our framework to the mouse (Mus musculus) to evaluate the generality and evolutionary conservation of the observed relationships.

Results

Chapter 1: Gene expression breadth is reflected by tau score

Tau score distribution across the human genome is bimodal

To explore how gene expression breadth is organized at the genome-wide level, we began by calculating and analyzing tau scores for all human protein-coding genes, a well-established and robust summary measure of expression breadth and variation.

To calculate the tau score for protein-coding human genes, we used RNA-seq expression data from the Human Protein Atlas program. The program provides two RNA-seq datasets covering all human genes: one based on bulk tissue expression across 40 tissues, and another based on single-cell expression across 81 cell types.[7,8] Accordingly, each gene has two possible tau scores: a single-cell-based score and a bulk-tissue-based score.

The tau score for each gene is calculated using the formula:

$$\tau = \frac{\sum_{i=1}^{N}(1 - x_i)}{N - 1}$$

The normalized expression values in the data were first log_{10}-transformed after adding 1 to all values, to avoid log of zero.

In this formula, x_i represents the normalized expression value in tissue/cell type i after the log transformation and dividing by the maximum expression value of that gene, and N is the total number of tissues/cell types. We divide by $N - 1$ because x_i is the result of dividing all expression values of a gene by its maximum, resulting in 1 for that value and therefore an addition of 0 to the summed total.

Thus, the tau score reflects both the number of tissues in which a gene is expressed and the variability in its expression levels. As noted above, the score ranges from just above 0 for genes expressed uniformly across all tissues to 1 for genes expressed in a single tissue only. A true zero is only possible for genes absent from the analysis, since a perfectly unified expression does not exist in biological data. For downstream analyses, we retained only genes that had both a single-cell-based tau score and a tissue-based tau score, resulting in 18,235 genes with both specificity measurements.

After calculating both tau scores, we found their distributions to be bimodal and highly correlated, with a Spearman correlation of 0.89 (fig 1.1.C). In a complementary analysis, we also included genes that had a tau score in only one of the two datasets, in order to assess the impact of incomplete sampling across tissues and cell types (fig 1.1.D) and have found it weakens the correlation and that genes missing from one dataset tend to have high tau score in the other. Genes with high tissue-tau but single cell tau of 0 are most likely expressed in cell types that were not captured in the single-cell data but are present in the corresponding bulk tissue. Conversely, genes that receive only a single-cell tau score likely belong to cell populations that were missed or under-represented in the bulk tissue sampling. As shown in the plot, the first scenario is more prevalent, indicating that limited single-cell coverage is the dominant source of missing specificity measurements.

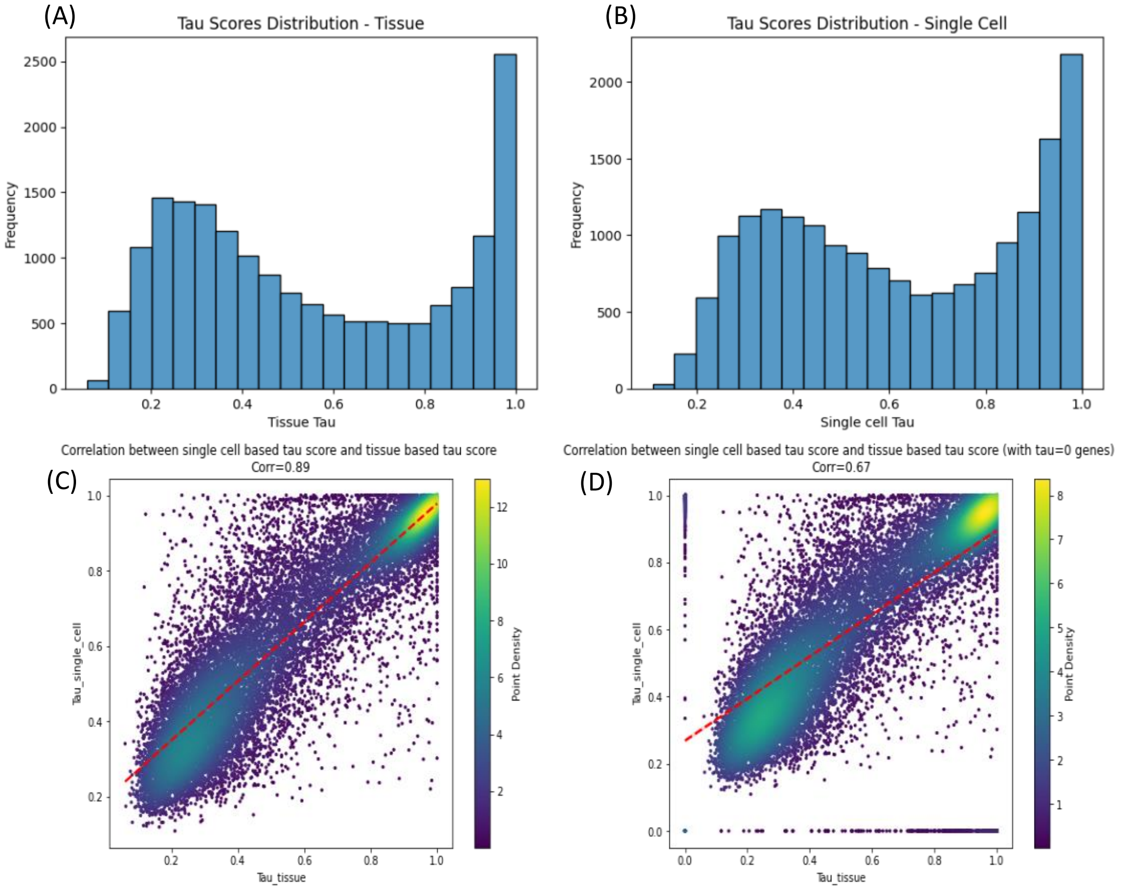


Figure 1.1. **The distribution of tau scores across the human genome.** (A) The tau score distribution based on bulk tissue RNA seq. (B) The tau scores distribution based on RNA seq single-cell data. (C) A scatterplot of the gene's tau score based on the tissue and the single cell data, with Pearson's correlation of 0.89 between the 2. In the plot, each dot is a gene. (D) The same scatterplot, including genes of tau=0 that we later neglect from the analysis.

Although the two tau scores for each gene are generally similar, the datasets used to calculate them differ in their coverage of the human body. At first glance, expression data across 81 cell types might seem more comprehensive than expression across 40 tissues. However, since each tissue is composed of multiple cell types, usually more than two, this simple sanity check suggests that the single-cell dataset represents a more limited view of the full cellular diversity of the human body.

Therefore, from this point onward we present analyses based on the bulk tissue tau scores, unless noted otherwise. All analyses were performed independently using both tissue-based and single-cell-based tau scores, and in most cases, they yielded highly consistent results. Cases in which the two measures yielded different results are explicitly addressed.

The tau score reflects the number of tissues expressing the gene and the expression levels variability

To demonstrate that the tau score agrees nicely with the breadth of gene expression, for both the number of expressing tissues and the variability in the gene levels found in these tissues, we plotted again the distribution of the tau scores of the human coding genes, and colored each gene by the number of tissues it is found to be expressed in, at least to a minimum level of 1 nTPM.

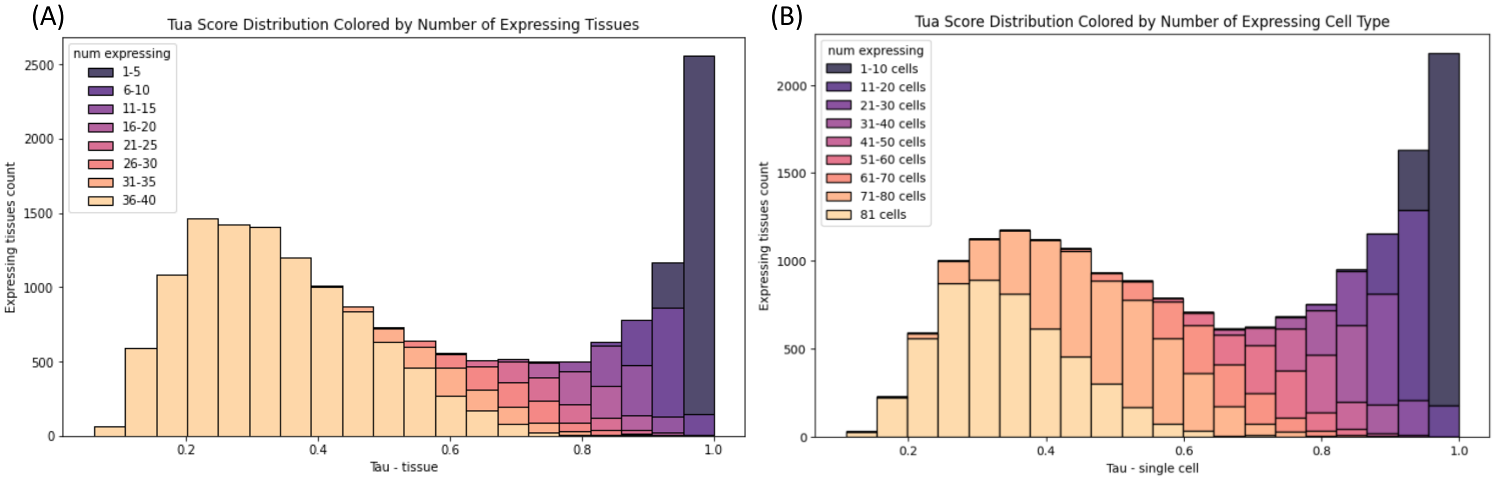


Figure 1.2. **The distribution of tau scores across the human genome stacked by number of expressing tissues and cell types.** Bars are colored by the number of tissues (A) or cell types (B) expressing the gene.

The plots show that for genes with tau values below 0.5, most are expressed in all or most sampled tissues, though their expression levels differ substantially across them. This leads to almost a 0.5 spread in tau specificity scores among genes that are expressed in the same number of tissues. Within this broadly expressed group, genes with more uniform expression have lower tau values, while increasing variability in expression drives tau upward. For genes with tau > 0.5, the number of expressing tissues becomes more variable and generally decreases as tau increases, reflecting increasing tissue specificity.

Large gene families do not disproportionately contribute to specific tau score ranges

Before examining regulatory features across tau score groups, we asked whether gene family size might bias the tau score distribution. Specifically, we considered the possibility that a large gene family could dominate a particular tau range, thereby introducing signals driven by shared family characteristics rather than general principles of gene regulation.

We used the HGNC database [9] to assign each protein-coding gene to its corresponding family or families. Most genes belong to at least one gene family, and some belong to more than one, since such families can be defined by either structural or functional features. Among protein-coding genes, the largest families are structure-based, as demonstrated by the two biggest ones: the “Zinc finger C2H2-type” family with 530 members and the “Ankyrin repeat domain-containing” family with 115 members.

The aim of our analysis was to determine whether large gene families disproportionately influence specific bins in the tau score distribution, potentially introducing bias. To address this, we assigned each gene to its corresponding family or families and categorized them by family size. We visualized the tau score distribution as a histogram, with bins colored by gene family size. When a gene belonged to multiple families, it was assigned the size of its largest family, since our concern was the influence of large families on the distribution. This was important because uneven representation of genes from large family for a specific tau range could influence the analysis in case of shared features among its members. We learned that genes from families of all sizes were evenly distributed across the tau score histogram, indicating that no significant bias was introduced by them.

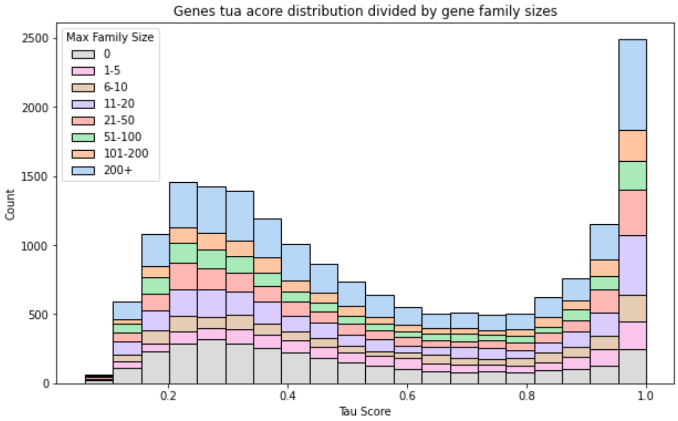


Figure 1.3. **The distribution of tau scores across the human genome stacked by gene family size.** For genes assigned to more than a single family, we considered the biggest one.

While no general trends were observed across large gene families as a group, it remained possible that an individual expanded family could make a disproportionately large contribution to a specific tau score bin, thereby distorting the interpretation. We therefore asked whether any single gene family was overrepresented in a particular region of the tau distribution.

To address this, we focused on gene families with 70 or more members, over 10% of the smallest tau bin in the histogram, and visualized the tau score distribution for each family as a heatmap. We also included all proteins with no family affiliation, grouped under family "nan" in the plot, as they might tilt the analysis as well. This analysis revealed no significant enrichment of any family within a specific tau bin, and the overall spread of genes across the tau range recapitulates the global bimodal distribution of tau scores observed across the human genome. Together, these results indicate that neither gene family size nor individual large families introduce a substantial bias into the observed tissue specificity patterns.

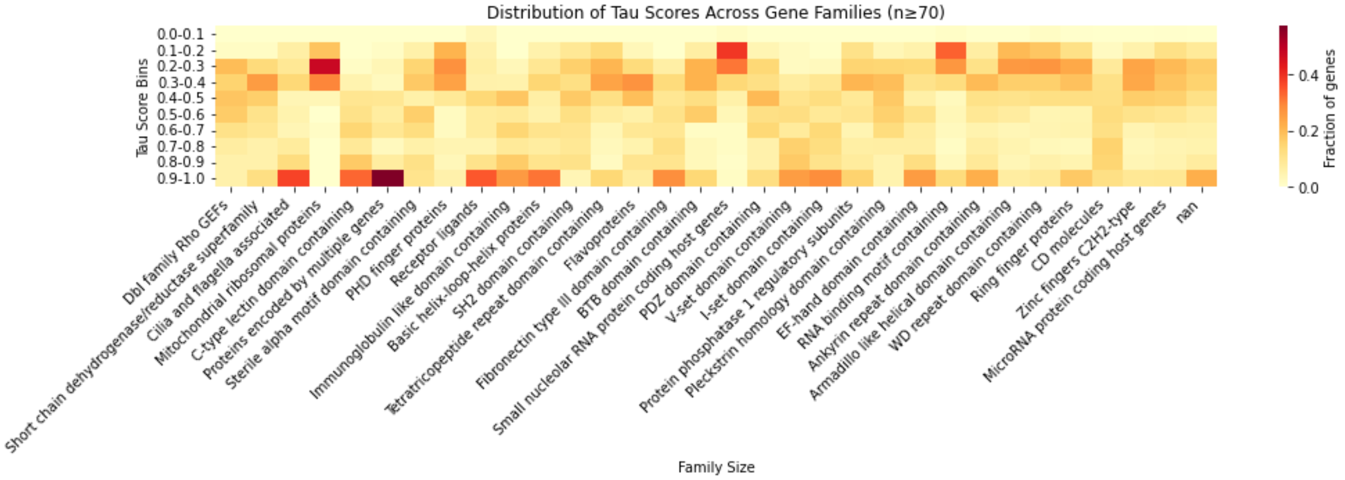
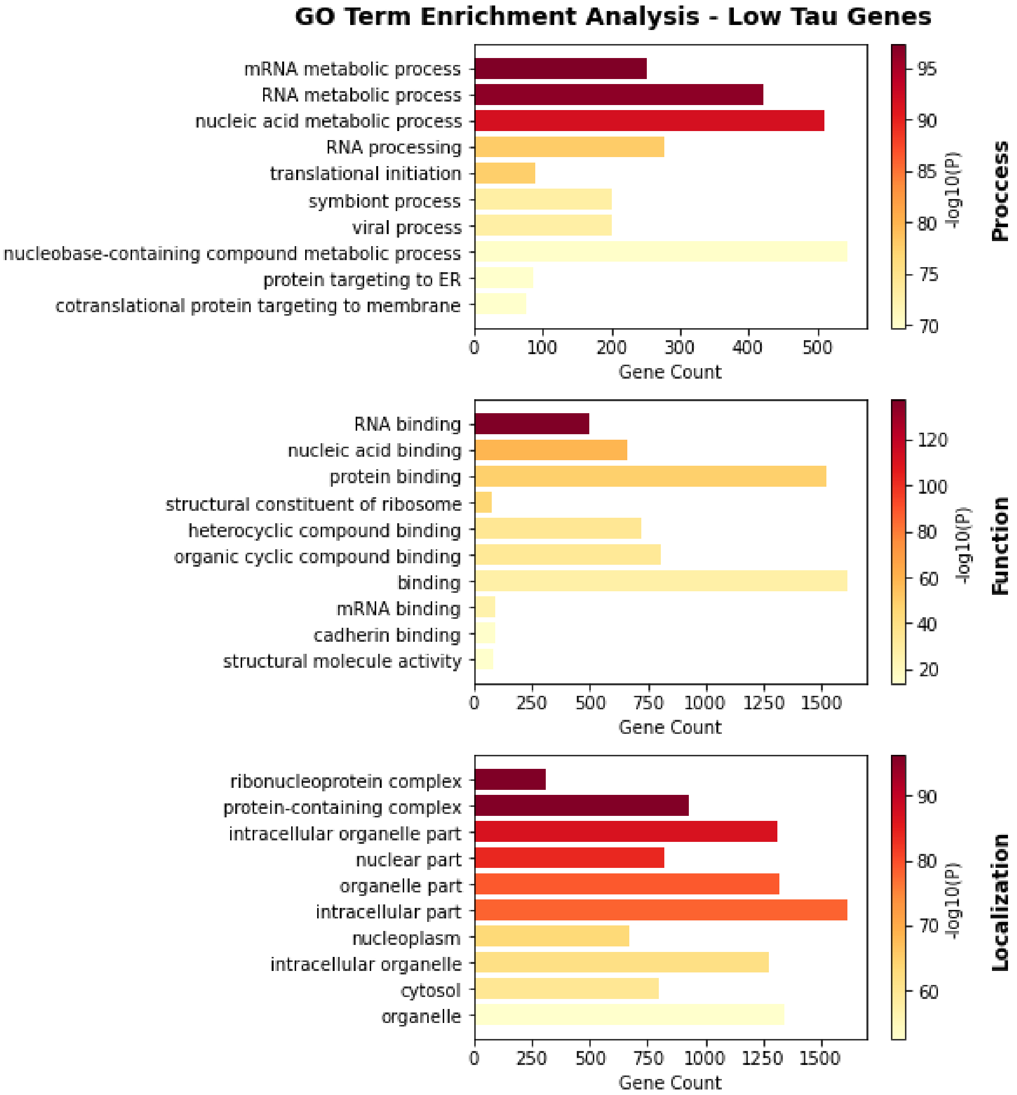


Figure 1.4. **The distribution of tau scores across the biggest gene families.** Each column represents the distribution of tau scores for all genes assigned to a gene family of 70 genes or more. The families are ordered in ascending size from left to right. The biggest "family" is marked "nan" as it contains all genes not assigned to a family, as we wanted to make sure there is no trend among these genes either.

Low tau genes are enriched for fundamental processes whereas high tau genes are enriched for sperm and gamete related processes

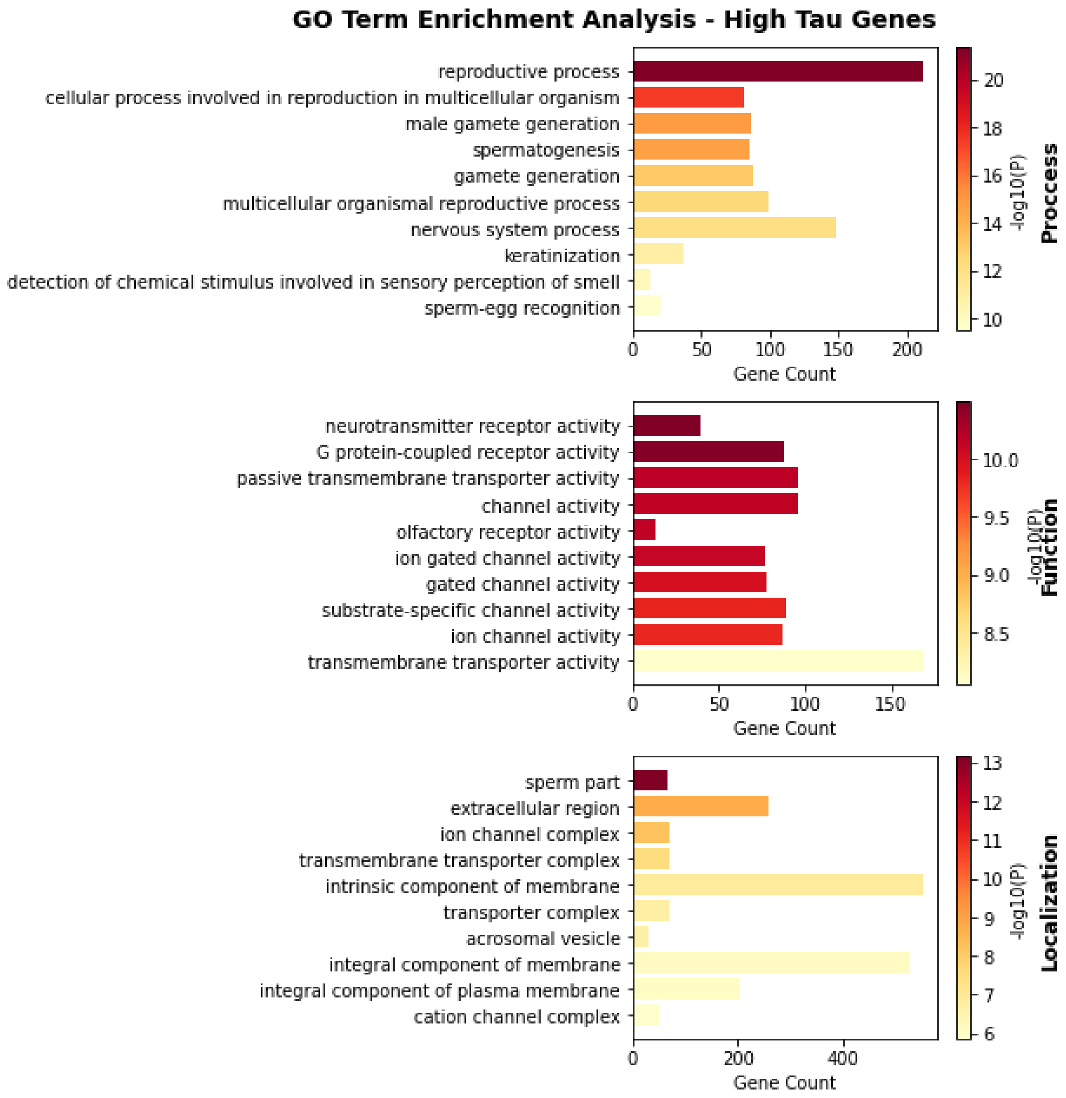
We next wanted to explore the meaning of the extreme tau scores at the peaks of the bimodal distribution. We performed ranked enrichment analysis, using Gorilla (Gene ontology enrichment analysis and visualization tool) with the genes ranked by their tau score[10,11]. We found the genes with low tau scores are enriched for fundamental cell processes and functions, like RNA and mRNA metabolic processes and binding across many cell departments, that are necessary for the function and survival of every living cell. They are also found in many different cell components.

Figure 1.5. **Go term enrichment analysis for low tau genes.** The enrichment analysis was performed using the Gorilla tool. The X axis represents the number of genes belonging to the GO category and the color represents the -log10(P) of the P values post FDR correction.



Genes with high tau values were significantly enriched for specialized biological processes, including reproduction (spermatogenesis) and keratinization, consistent with their tissue-restricted expression patterns. In contrast to low tau genes, the enriched categories for high tau genes contained fewer genes and showed lower overall counts per category, reflecting the fact that highly specialized functions are confined to a limited number of tissues and cell types. This is biologically expected, as tissue-specific pathways are inherently rarer and involve smaller gene sets compared to fundamental cellular processes that are shared across many tissues.

Figure 1.6. **Go term enrichment analysis for high tau genes.** The enrichment analysis was performed using the Gorilla tool**.** The X axis represents the number of genes belonging to the GO category and the color represents the -log10(P) of the P values post FDR correction.



Genes with midrange tau values are enriched for cell migration and immune related processes but with lower significance

We next examined genes with mid-range tau scores, corresponding to the region between the two extremes of the bimodal distribution. Unlike the low- and high-tau analyses, which are naturally defined by the gene ranks in an ascending or descending manner, the definition of “mid-range tau” is less straightforward and can be approached in several ways: (i) as a tau value of approximately 0.5, representing the mathematical midpoint between 0 and 1; (ii) as the central portion of the distribution around the median or mean tau value; or (iii) as the transition zone in which the number of expressing tissues begins to decrease while variability in expression still contributes substantially to the score as visually seen in figure 1.2

We applied all three definitions and observed similar, though not identical, enrichment patterns across them. Eventually, we defined mid-range tau relative to a value of 0.5, the midpoint of the possible tau scale. Rather than applying a strict cutoff, we used a ranked enrichment approach in which genes were ordered by their distance from tau = 0.5. This avoids reliance on arbitrary thresholds and ensures that the analysis is not dependent on the exact shape of the tau distribution in a given dataset. This also provides a stable reference point for future analyses.

Enrichment analysis of these genes showed that, although some functional categories, cellular components, and biological processes are enriched, the statistical significance is much lower than in the extreme tau groups. The main themes that do appear are related to immune functions and cell migration. The weaker enrichment signal suggests that mid-tau genes are not strongly tied to a specific subset of biological processes. Instead, they represent a broader expression pattern that is shared across multiple tissues and cell types, rather than defining a single cellular role.

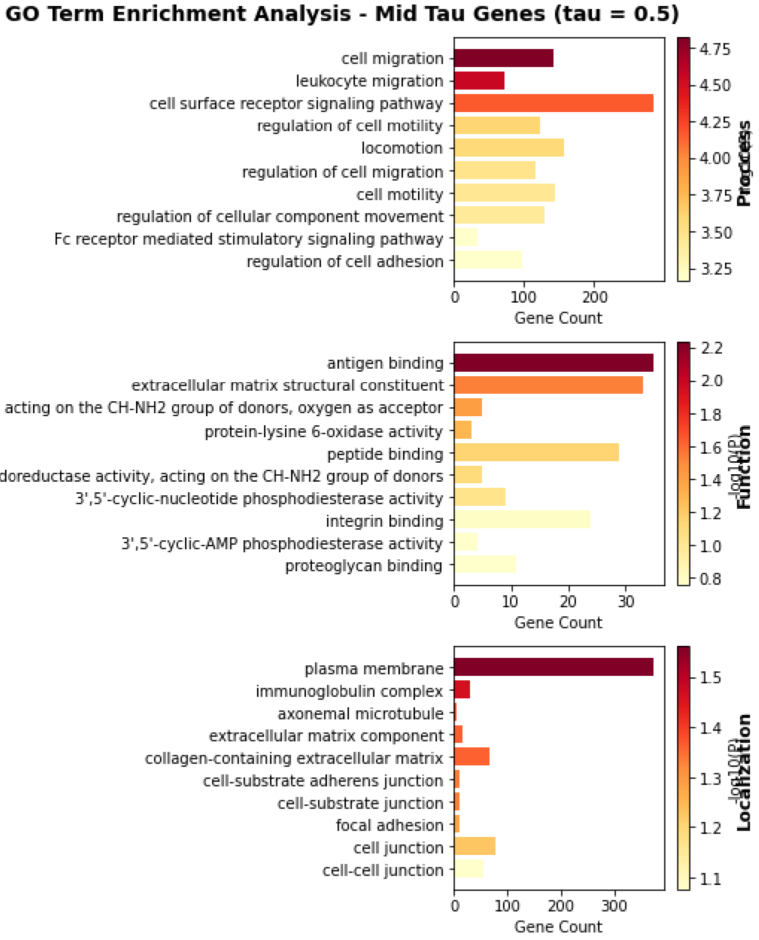


Figure 1.7. **Go term enrichment analysis for mid-range tau genes.** The enrichment analysis was performed using the Gorilla tool**.** The X axis represents the number of genes belonging to the GO category and the color represents the -log10(P) of the P values post FDR correction.

Genes expressed in underrepresented tissues in the HPA datasets lack tau scores

We only included in our analysis genes for which tau scores could be calculated using both tissue bulk expression and single-cell data. This approach yielded 18,235 genes, with about 1,900 human protein coding genes missing. We investigated whether these excluded genes share any common characteristics. These missing genes could be explained by the method used to build the HPA dataset. The Human Protein Atlas data utilized in our study comprises 40 tissues and 81 cell types. While comprehensive, these numbers do not capture the full diversity of human tissue and cell types, inevitably leading to some gaps in gene coverage. For example, about 600 of the excluded genes are associated with the olfactory signaling super-pathway, as we discovered using GeneCards' Gene Analytics tool[12]. This finding is consistent with our expectations, given that nasal tissue was absent from both the bulk tissue samples and single-cell datasets. The olfactory receptor gene family is extensive, yet our dataset includes only 21 of its members.

In addition, many of the missing genes are affiliated to different brain parts, but the data only captures 2 brain tissues (cerebral cortex and choroid plexus) and the single cell data includes about 8 different brain cell types.

Besides the absence of some tissues from the data, certain genes are primarily expressed under specific conditions, such as different stress conditions or diseases, or particular developmental stages, including embryonic or early postnatal development. As the Human Protein Atlas data are derived from healthy adult donors, genes with condition-specific or developmental expression patterns are likely to be underrepresented.

Down-sampling of tissue expression data reveals stabilization of tau scores

The final technical question we addressed in this preliminary stage of the work was regarding the robustness of the tau scores to omissions of tissues from the calculation. Single-cell–based specificity scores were highly correlated with tissue-based scores, demonstrating strong consistency between scores derived from 40 tissues and 81 cell types. However, we asked whether the inclusion of additional tissues or cell types in the future could substantially alter these results and potentially shift the analysis.

This question can be framed more broadly: if the number of human tissues and cell types is finite, then in principle there exists a “true tau” score that fully captures a gene’s expression specificity. We sought to estimate how close our current measurements estimate this value and how sensitive the tau score is to the addition of new expression information.

To address this, we applied a down-sampling approach. For each dataset, we randomly sampled subsets of tissues or cell types, ranging from 1 to the full set, repeating this process 500 times for each subset size.

For each sample, we recalculated tau scores and computed their Pearson correlation with the tau scores derived from the full dataset. We then calculated the mean correlation and standard deviation across the 500 replicates for each subset size, and then plotted the number of sampled tissues against the mean correlation.

Plotting the number of sampled tissues against the mean correlation revealed a saturation curve: correlation increased rapidly with the first few tissues and reached a plateau around ~45 tissues, indicating stabilization of the tau scores at that number of sampled tissues. This trend was more pronounced for single-cell–based tau scores. These results suggest that sampling approximately ~45 tissues is sufficient to obtain stable tau estimates, and that additional tissues have limited impact. Overall, this supports the robustness and reliability of our tau score calculations as it suggests that even if more tissues are sampled in future, the two scores estimation based on current tissues will likely be minorly affected.

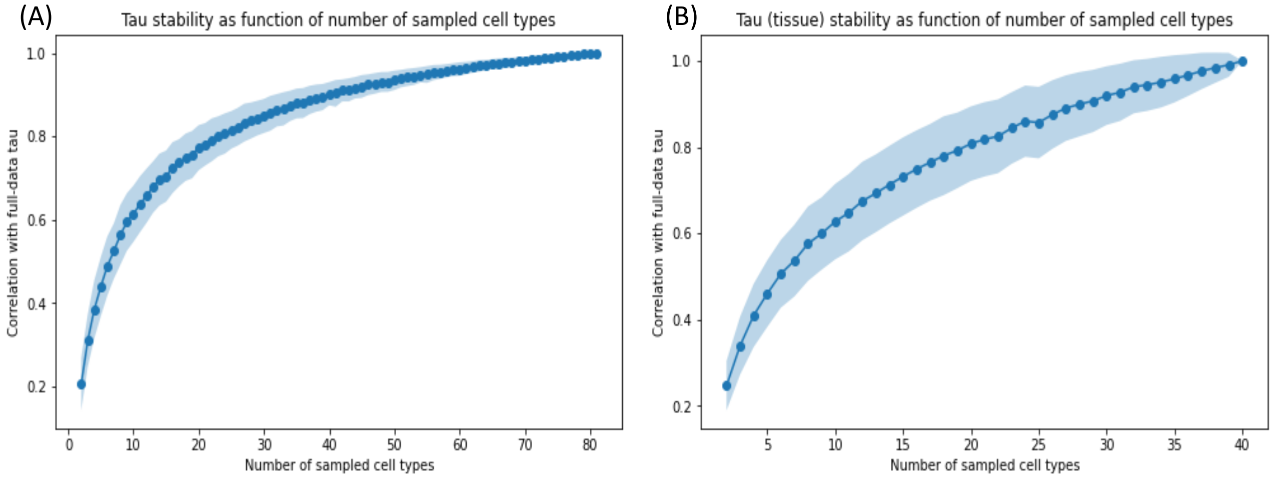


Figure 1.8. **Correlation between sample-based tau score to the final calculated one for number of cell types (A) and tissues (B)**. The x-axis shows the number of sampled tissues/cell types, and the y-axis shows the mean correlation (across 500 repetitions) between tau scores calculated from the sampled subset and the final tau scores based on the full dataset. The shaded area represents the standard deviation across repetitions.

Chapter 2: Greater expression complexity requires more regulatory information

Mid-range tau genes are expected to have the most regulatory content, under the MDL principal

As we have shown so far, between the extreme expression patterns of house-keeping genes and tissue specific one lies a large class of genes with intermediate tissue specificity, expressed highly in some tissues, yet minimally or not at all in others. Despite their abundance, it is unclear whether the characteristics of intermediately expressed genes simply represent a midpoint between broadly and narrowly expressed genes, or whether they exhibit distinct qualities and features. We sought to explore these features, focusing on the cis- and trans-regulatory elements and interactions with other players in the cell.

Patterns of gene expression are governed by multiple interconnected layers of regulation. This regulatory architecture includes many elements – the gene's CDS, introns, 3'UTR and 5'UTR, enhancers, miRNAs and more, each of these plays a role in shaping the correct gene expression pattern.

Under the MDL framework, house-keeping and tissue-specific genes are expected to require less regulatory content to achieve their expression patterns. As we mentioned before, patterns such as "express in all cells" or "express only in cell type X" is simpler than "express highly in cell types A and B, less in cell type C, minimally in cell D, and not at all in cell types E-G".

For instance, since regulatory information may be embedded within the gene's structure, it would be expected that house keeping and tissue specific genes be relatively short in comparison to mid-range tau ones. And indeed, prior studies have showed broadly expressed house-keeping genes are typically shorter; an observation thought to reflect selection against the energetic cost of long transcripts[13]. In contrast, genes with more complex expression patterns may encode additional regulatory instructions within their structure, resulting in increased length.

We hypothesized that the regulatory information content behaves similarly and would be higher for genes of mid-range tau score and complex expression patterns. To test this hypothesis, we divided the tau score distribution into 20 equally spaced bins and calculated the mean values of many regulatory features for those bins. The next plots will showcase the trends in which these features behave relatively to their tissue specificity score.

Genes with a mid-range tau score have longer structural features and more associated enhancers

To explore the connection between structural features and the tau score, we used the BioMart data mining tool[14,15] to find the transcript length, intron length and UTRs length for all human coding genes. For each gene we calculated these values, and after dividing all genes into 20 bins based on their tissue specificity scores, we calculated the mean value per bin.

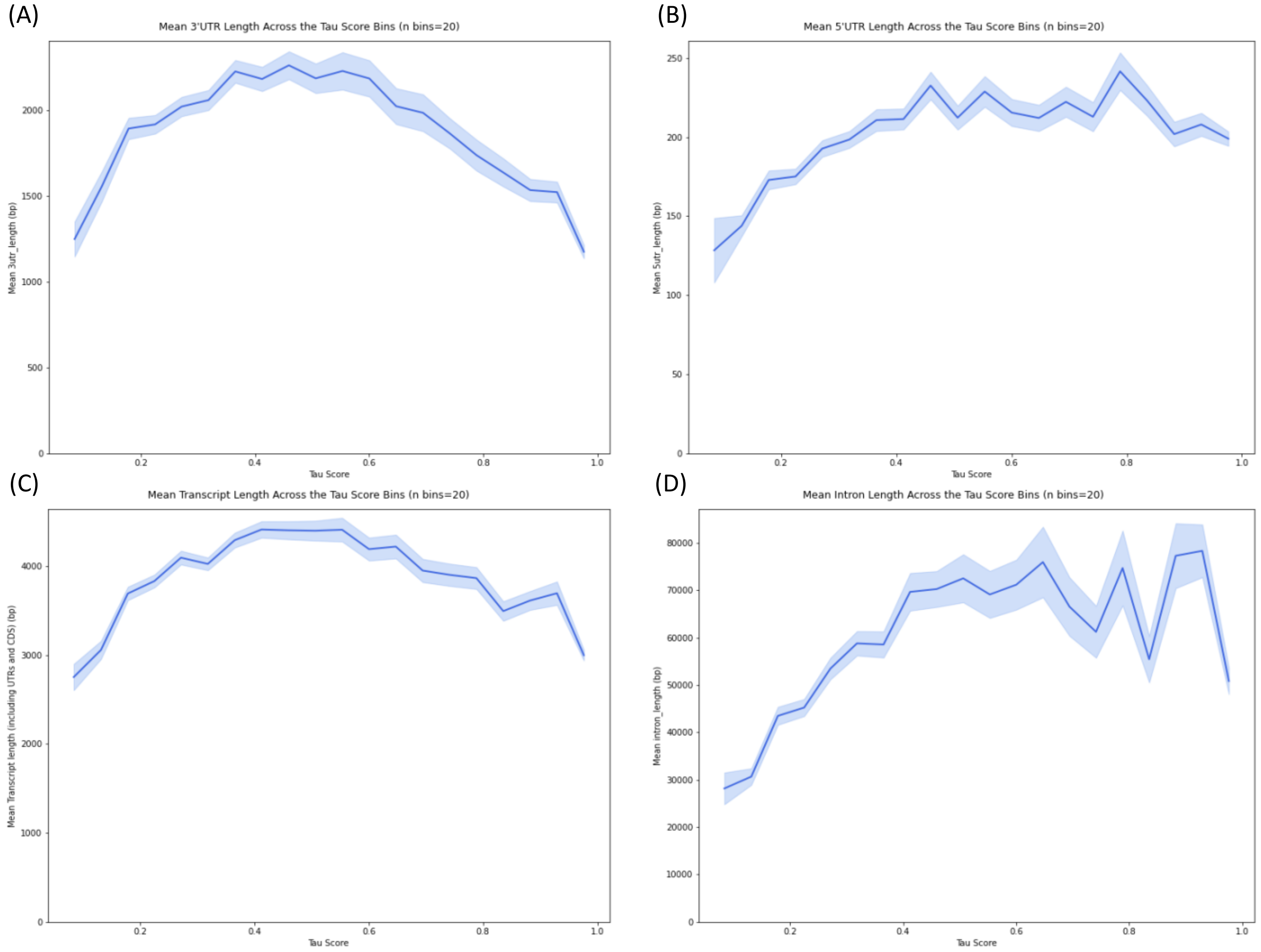


Figure 2.1. **Mean values of 3' UTR** (A) **and 5' UTR** (B) **lengths, transcript length** (C) **and intron length** (D) **across 20 tau score bins.** Shaded areas represent the Standard Error of the Mean (SEM).

Consistent with the MDL principal, we have found these features show an inverted U-shape peeking for the genes with intermediate tissue specificity scores indicating a more complex expression pattern. The U-shape is symmetrical for the 3'UTR length, and for the transcript length, intron length and 5'UTR the broadly expressed genes are the shortest. The 3'UTR length showed the most pronounced trend of peaking among genes with intermediate specificity. This aligns with its known regulatory roles, as 3' UTRs often serves as hubs for post-transcriptional control, including microRNA and other regulatory elements, while introns can harbor miRNA binding sites and additional regulatory elements. Among all features, 3' UTR length showed the strongest relationship with tau, highlighting its regulatory importance.

These findings highlight gene structure, particularly 3'UTRs, as a proxy for embedded regulatory information content. Longer genes appear to accommodate more intricate regulatory mechanisms, consistent with MDL predictions and what is shown in the literature.

We next asked whether the similar behavior of structural features reflects a direct coupling within gene architecture. Specifically, we tested whether genes with longer 3'UTRs also tend to have longer 5'UTRs. If so, correlations between features could trivially explain the coordinated trends observed across tau scores.

To address this, we examined the relationship between 3'UTR and 5'UTR length across all genes and have observed no association between them (Pearson's r = 0.15). This indicates that variation in UTR length is largely independent across transcript ends and cannot account for the parallel inverted U-shaped patterns observed for multiple structural features.

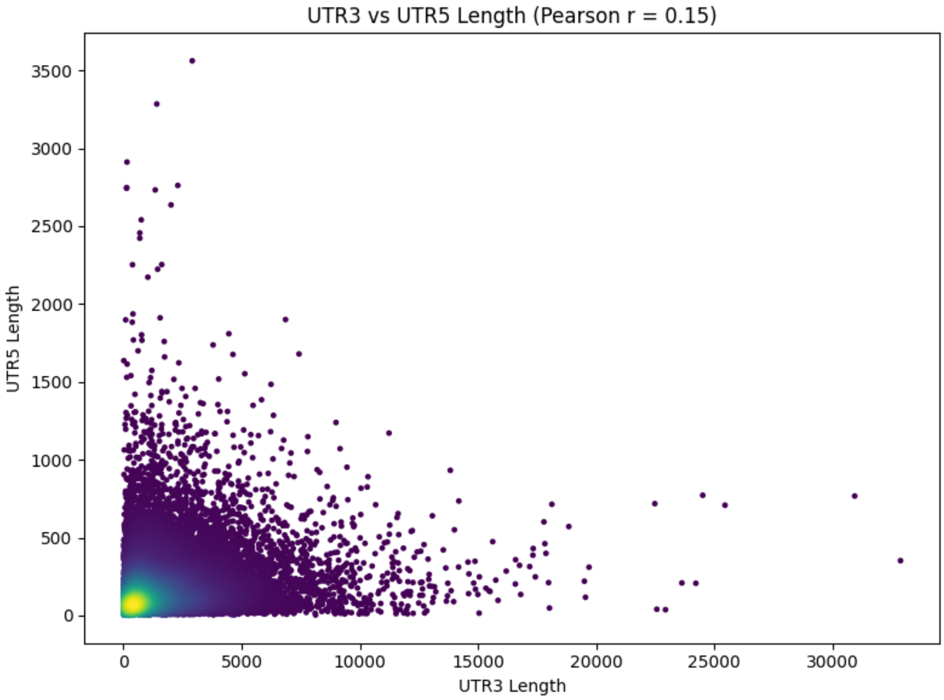


Figure 2.2**. Relationship between 3'UTR and 5'UTR length.** Each point represents a gene, plotted by its 3'UTR length (x-axis) and 5'UTR length (y-axis). Points are colored by local point density. Although both features span several orders of magnitude, no strong linear relationship is observed (Pearson's r = 0.15), indicating that 3' and 5' UTR lengths vary largely independently across genes.

To further explore this trend, we also checked the enhancers count per gene. Candidate cis-regulatory elements annotations and gene-enhancer links were obtained from GeneHancer[16], which integrates data from multiple sources and assigns confidence scores to predicted gene targets. We performed all analyses using both the full dataset (419,020 cis-regulatory elements) and the high-confidence "Double Elite" subset (122,815 elements), in which both the enhancer itself and its link to the target gene are supported by at least two independent sources, providing stronger evidence for the regulatory association. In the full dataset, each gene was linked to an average of ~55 enhancers, and each enhancer to ~7 genes; in the Double Elite subset, each gene was linked to ~7 enhancers, and each enhancer to ~6 genes. Results were consistent across both sets; primary analyses used the elite set. The number of enhancers required for the genes compared to their breadth of expression once again peaks among genes with intermediate tissue specificity, consistent with the MDL principle.

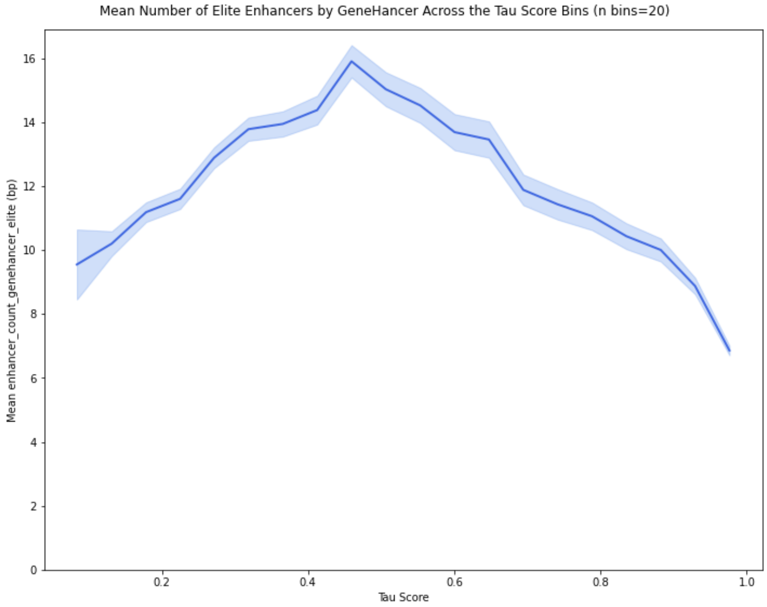


Fig 2.3. **Relationship between tau score and enhancers count per gene.** The mean enhancer counts per tau score bin. The values shown are for the bins assigned to the 20 bins based on their tau score, and the shaded areas around the lines are the Standard Error of the Mean (SME) values.

House-keeping genes are prone to having CPG islands

We next examined the number of associated CpG island for each gene[17] to evaluate its relationship with the tau score. CpG islands are genomic regions enriched in cytosine–phosphate–guanine (CpG) dinucleotides, typically located near gene promoters. They are generally associated with unmethylated DNA and contribute to maintaining an open chromatin state that supports gene transcription. This property explains why housekeeping genes, which are broadly and consistently expressed across tissues, are known to contain a higher CpG islands content[18,19].

To systematically explore the relationship between a gene's expression breadth and it's CpG island content, we assigned CpG islands to their corresponding genes using the accepted definition of what constitutes a CpG island as was proposed in 1987 by Gardiner-Garden and Frommer[20], including the island being at least 200 bp long with a CG content of at least 50% and an observed CpG/expected CpG in excess of 0.6 , where the Obs/Exp ratio is defined as $\frac{\text{number of CpG dinucleotides} \times \text{sequence length (in bp)}}{\text{number of C nucleotides} \times \text{number of G nucleotides}}$.

Our analysis shows an almost linear relationship between the mean CpG Obs/Exp ratio and the tissue specificity score across the genome. This high CpG content in promoters of house keeping genes is very well known[21]. Such house keeping genes promoters are typically unmethylated, a state that allows their broad expression. Using a scatterplot to better see the distribution of the data points, two distinct and differently looking gene groups emerge: genes that contain CpG islands and genes that do not. The latter group—genes without CpG islands—was strongly enriched for tissue specific genes, and therefore substantially influenced the overall mean trend. When excluding genes without CpG islands, we observed that broadly expressed genes tend to have CpG islands, but their CpG content hardly increases further with expression breadth. This pattern supports the notion that the presence of CpG islands, rather than their Obs/Exp ratio value, is the key feature distinguishing ubiquitously expressed genes from tissue-specific ones. It seems to be a binary feature that decreases with the rising specificity of the gene. The leaner trend we sought to have found was an artifact of combining these two separated gene groups into a single analysis.

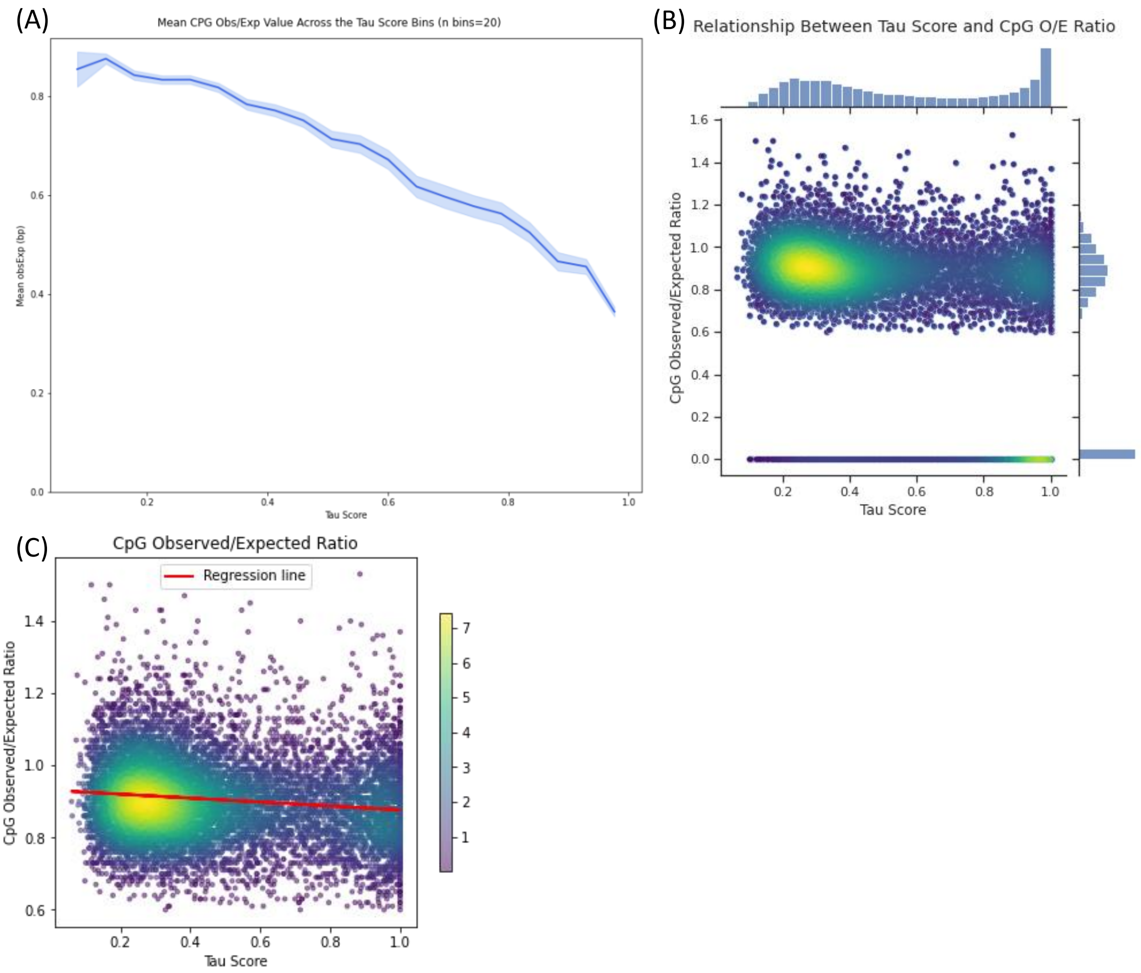


Fig 2.4. **Relationship between tau score and CpG content in genes.** (A) The mean CpG observed/expected value per tau score bin. The values shown are for the bins assigned to the 20 bins based on their tau score, and the shaded areas around the lines are the Standard Error of the Mean (SME) values. (B) A scatterplot of the tau score and observed/expected ratio for all genes, so that each gene is a dot. Genes with no CpG islands are represented as 0 on the y axis. (C) The same scatterplot, including only values for genes annotated to have CpG islands.

Broadly expressed genes interact more with miRNA and proteins

Beyond cis-regulatory control, gene expression is also shaped by post-transcriptional and post-translational mechanisms that influence RNA stability, protein interactions, and protein degradation. We therefore asked whether post-transcriptional regulation through miRNA targeting and protein–protein interaction complexity follow similar MDL-related trends and scale with tissue specificity and expression complexity.

To quantify protein interaction complexity, we used interaction counts from the STRING database[22], which integrates physical interactions and functional associations from experiments, curated knowledge, and computational predictions. For miRNA binding cites count we used TarBase database[23], an experimentally supported database of protein targets for miRNA molecules.

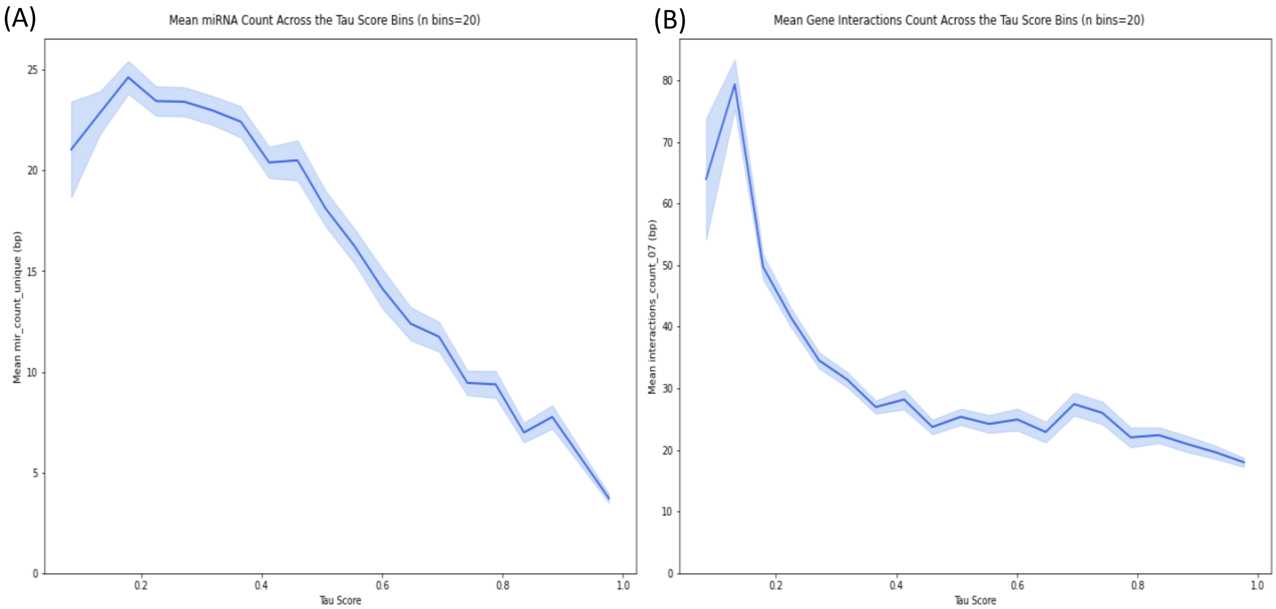


Fig 2.5. **Relationship between tau score and protein interactions count per gene.** (A) The mean protein-protein interactions count per tau score bin. (B) The mean number of miRNAs targeting a gene per tau score bin. In both plots, the values shown are for the bins assigned to the 20 bins based on their tau score, and the shaded areas around the lines are the Standard Error of the Mean (SME) values.

The miRNA count shows a slight rise at the low tau values, reaching a small peak, after which it decreases progressively toward the tissue-specific end of the distribution. The protein-protein interactions count shows similar trends but in a sharper manner. Both plots show that interaction counts are highest for broadly expressed genes, with a slightly different shape of the decline – for protein–protein interactions, the decrease more gradual and roughly linear, whereas for miRNA interactions the drop is steeper at the start and then flattens. Despite these differences in shape, the shared pattern is clear: genes with broad expression tend to accumulate more interaction partners than tissue-specific genes.

Overall, interaction complexity is highest among broadly expressed genes, despite the non-monotonic behavior at the extreme low-tau range. Unlike previous results, where complex expression patterns increase the need for additional regulatory elements, other factors could explain this trend:

First, protein–protein interactions and miRNA-mediated regulation function as fine-tuning mechanisms—a knob-like adjustment compared to the switch-like effects of enhancers or 3'UTRs. They do not determine whether a gene is expressed, but rather modulate activity, stability, and degradation at a post-transcriptional level. Such precision control is mostly relevant for genes that produce large amounts of protein, where a precise expression regulation could be very energetically expensive. For such genes, the mechanism could be keeping the gene at an "on" default and adjusting the outcome if necessary.

Second, proteins expressed in many cell types must be able to interact with a broad range of partners to maintain their general functionality and efficiency.

Third, as shown in the next chapter, housekeeping genes tend to be evolutionarily older, giving them more time to accumulate interaction partners. In contrast, younger proteins have had less evolutionary time to develop extensive interaction networks.

Correlations between the genes' features are weak despite shared trends

Some of the signals we identified show similar directionality and overall shape. Both the number of protein–protein interactions and miRNA targeting display a modest increase followed by a pronounced decrease across the tau spectrum. Likewise, UTR lengths and enhancer count exhibit similar inverted U-shaped patterns, peaking in genes with intermediate tissue specificity. We therefore asked whether these non-monotonic trends reflect a shared underlying signal or redundancy between features. To address this and assess the degree of redundancy between features, we examined pairwise Spearman correlations among all regulatory and structural measures. For visualization, features were hierarchically clustered based on their correlation profiles.

Overall, correlations between features were modest, with most pairs showing weak or near-zero association. Even features that display similar inverted U-shaped relationships with tissue specificity do not strongly correlate with one another. This was somewhat unexpected given the similarity in their global trends, and highlights that these features capture largely independent aspects of regulatory complexity rather than redundant signals.

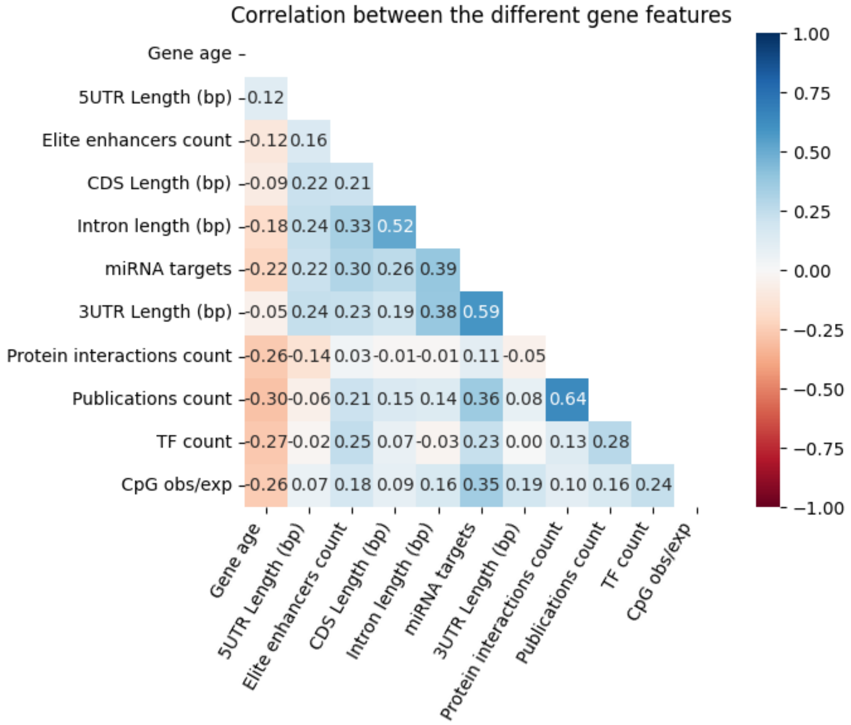


Figure 2.6. **Correlation matrix between the different gene features**. Pairwise Pearson correlations between all analyzed features are shown, with correlation coefficients indicated by both color and value.

Paralogue genes tend to have a similar tau score and expression patterns

Paralogue genes are genes that originated from duplication events[24] and therefore often share detectable sequence similarity and sometimes related functions and features. Since duplicated genes can retain overlapping regulatory programs, we hypothesized that some paralogues may also share similar expression patterns and therefore tissue specificity scores.

Paralogous gene annotations were obtained from Ensembl via BioMart. In total, the dataset contained 102,217 human paralogue pairs, involving ~14,000 genes, with many genes participating in multiple paralogous relationships. Paralogues span a wide range of sequence similarity and to focus on confident paralogous relationships, we also applied a commonly used sequence similarity threshold of 35% protein identity[25], yielding a filtered set of 20,448 paralogue pairs comprising ~8,400 genes. Importantly, all paralogue-based analyses produced highly similar results regardless of whether the full or thresholded dataset was used, indicating that the observed trends are robust to paralogue definition. In the analyses presented here, we therefore focus on the thresholded paralogue set, as it provides a conservative and high-confidence view of paralogue relationships.

To test this, we examined all paralogue gene pairs in the human genome and quantified how often both genes in a pair fall into similar tau score ranges. We divided the tau score into seven bins and placed each paralogue pair into a two-dimensional matrix according to the tau bin of gene A and gene B of the pair. After counting the number of observed paralogue pairs in each bin combination, we calculated the maximum possible number of pairs for each cell by multiplying the number of genes in the corresponding x-bin with the number in the y-bin. The resulting heatmap displays the percentage of observed paralogue pairs relative to this possible maximum.

In the heatmap, the diagonal is clearly darker than the rest of the plot, indicating that paralogues are more likely to share similar tau scores. Although the numbers are small, 0.055% on the diagonal peak vs. 0.005% elsewhere, the enrichment is consistent and supports the conclusion that paralogue genes tend to exhibit similar levels of tissue specificity, rather than different ones. The diagonal of the heatmap is consistently enriched relative to off-diagonal entries, demonstrating that paralogous genes are more likely to occupy similar tau score ranges than different ones. This enrichment is strongest for high tau bins and peaks for the intermediate bin of 0.6-0.73, suggesting that similarity in tissue specificity is particularly well preserved among paralogues with more complex or restricted expression patterns.

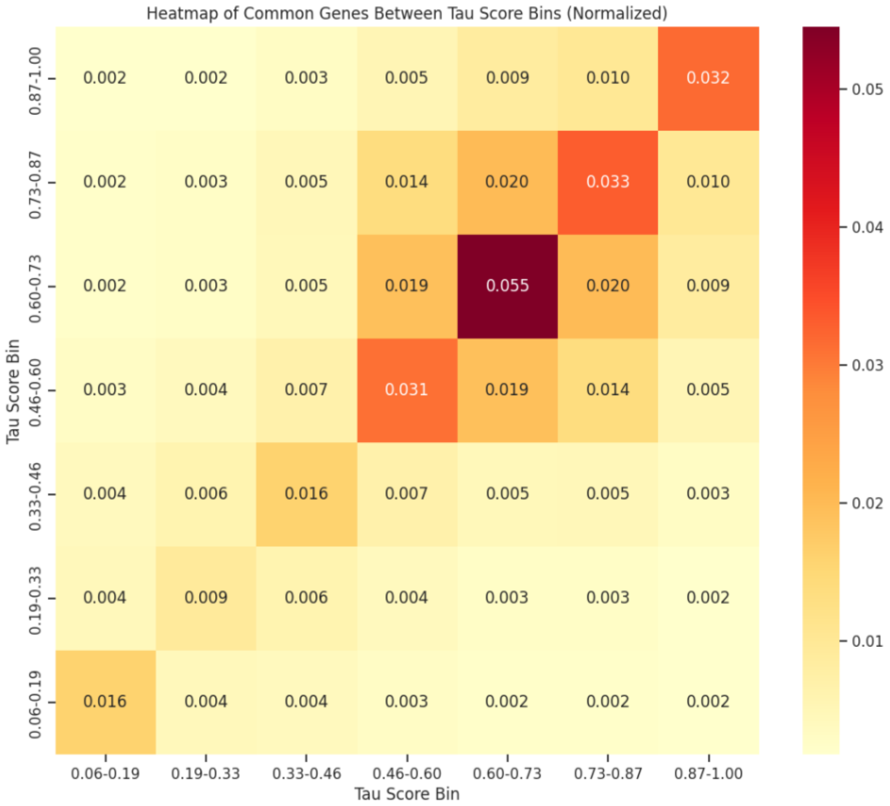


Figure 2.7. **Heatmap of the percentage of paralogous genes pair between tau score bins, out of all possible pairs**. The heatmap shows, for each pair of tau score bins, the percentage of observed paralogue gene pairs out of the maximum possible number of pairs for that bin combination. For each cell, the maximum possible value was calculated as the number of genes in the x-axis bin multiplied by the number of genes in the y-axis bin (adjusted appropriately for the diagonal).

While paralogous genes generally tend to have similar tissue specificity, it is also interesting to examine cases that deviate from this trend. We therefore examined the top pairs with the largest differences in tau scores. Two illustrative examples are SEPTIN12–SEPTIN2 ($\Delta\tau \approx 0.94$) and PAPOLB–PAPOLA ($\Delta\tau \approx 0.90$). In both cases, the broadly expressed paralogue retains low tissue specificity, whereas the high-tau paralogue is predominantly expressed in the testis. In both cases, the broadly expressed paralogue exhibits a substantially higher regulatory burden than its tissue-specific counterpart. For example, PAPOLA has nearly three times more elite enhancers than PAPOLB (28 versus 10), is regulated by far more transcription factors (36 versus 2) and is longer. In agreement with our prior analysis, the broadly expressed gene participates in a larger number of protein–protein interactions (56 versus 36) and shows extensive miRNA targeting (75 interactions versus none). A similar pattern is observed for the SEPTIN12–SEPTIN2 pair. These observations suggest that paralogue specialization toward a single tissue can be accompanied by a reduction in regulatory complexity. Although, from an MDL perspective, these paralogues might be expected to retain similar regulatory burden since they share a relatively simple expression patterns – a very broad and uniform one and a very tissue specific one. This divergence could reflect the lower regulatory information required to support highly tissue-specific expression, particularly in the testis. This possibility is further discussed in Chapter 6.

We further examined paralogous gene pairs with intermediate differences in tissue specificity and identified several cases in which the tau score difference was close to 0.5, with one gene exhibiting higher expression complexity across tissues. One such example is the MYH11–MYH1 pair. MYH11 has a tissue tau score of ~0.48, whereas MYH1 is expressed almost exclusively in skeletal muscle with a tau score of ~0.98. Consistent with this difference, MYH11 exhibits substantially greater regulatory complexity, including more elite enhancers (41 versus 2), much longer introns (147014 versus 20213), and a markedly longer 3'UTR (856 versus 109). As expected, MYH1 shows fewer miRNA interactions, while protein–protein interaction counts are similar between the two genes. Unlike earlier examples in which paralogues share the same evolutionary age, MYH1 and MYH11 differ in age assignment (phylostrata 17 and 12, respectively), illustrating how paralogues can diverge at different evolutionary times and acquire distinct expression and regulatory profiles.

These outlier cases prompted us to ask whether highly tissue-specific genes are biased toward tissues more generally. To address this, we examined the distribution of main expressing tissues for the genes in our data. Because the main expressing tissue is meaningful primarily for genes with high tau scores, we restricted the analysis to paralogue pairs in which both genes had tau ≥ 0.8. The results sow the testis are the most frequent dominant tissue, closely followed by the bone marrow and skeletal muscle. This observation suggests that the testis is a major contributor to high tissue specificity.

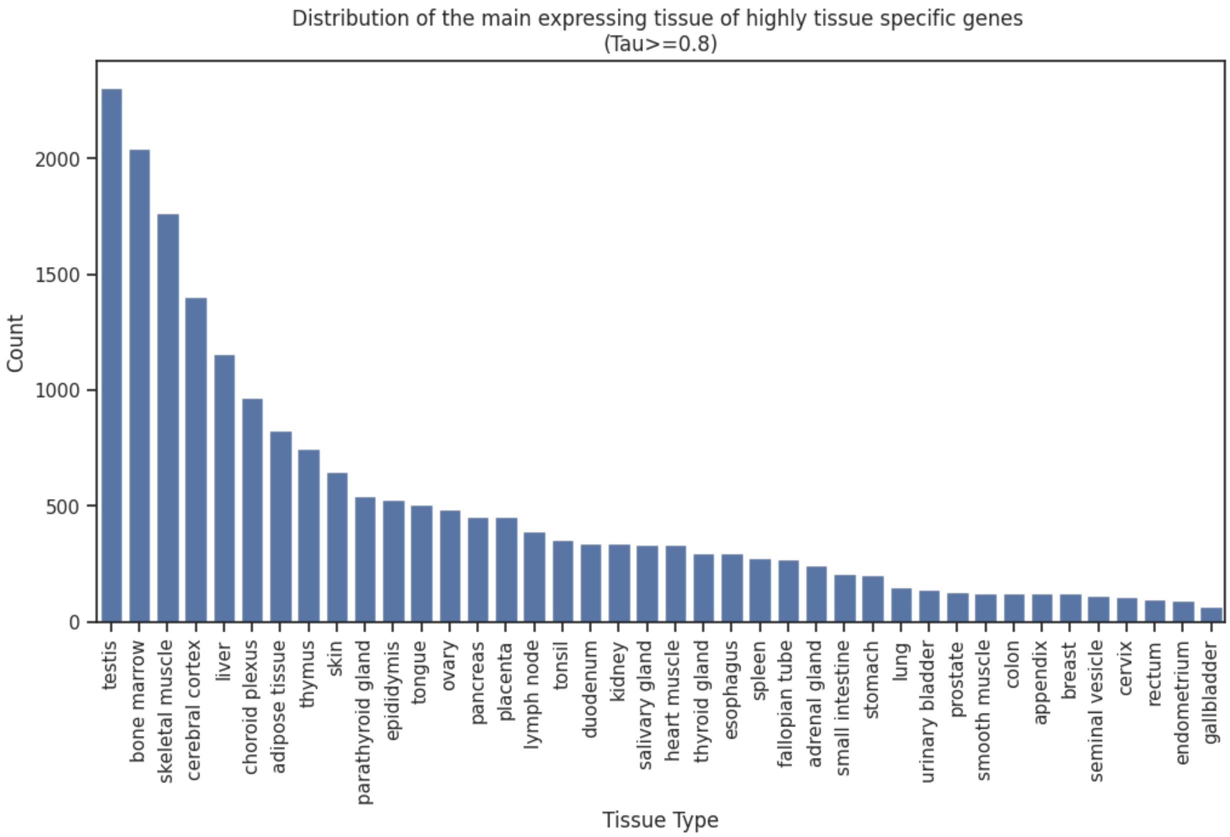


Figure 2.8. **the distribution of the main expressing tissue of genes with specificity score of 0.8 and higher**.

We next asked whether paralogue genes tend to share not only similar levels of tissue specificity but also the main expressing tissue. To assess whether this overlap exceeds what would be expected by chance, we used a permutation-based randomization test: For each paralogue pair, we kept the main expressing tissue of Gene A fixed and randomly reassigned the main expressing tissue of Gene B by shuffling the original tissue labels across all genes. Shuffling was used rather than assigning a tissue at random from a list, in order to preserve the overall tissue distribution and maintain the same probabilities as in the real dataset. We repeated this 10,000 times and, for each iteration, counted how many pairs had both paralogues expressed mainly in the same tissue.

The real number of pairs in which both paralogues shared the same main expressing tissue was far higher than in any of the randomized iterations, indicating that this pattern is extremely unlikely to arise by chance. We therefore conclude that paralogue genes with high tau scores tend to be expressed predominantly in the same tissue. We also applied McNemar’s test to evaluate whether the number of paralogous gene pairs sharing the same main expressing tissue was higher than expected by chance. The test confirmed a highly significant deviation from randomness, with a p-value approaching zero.

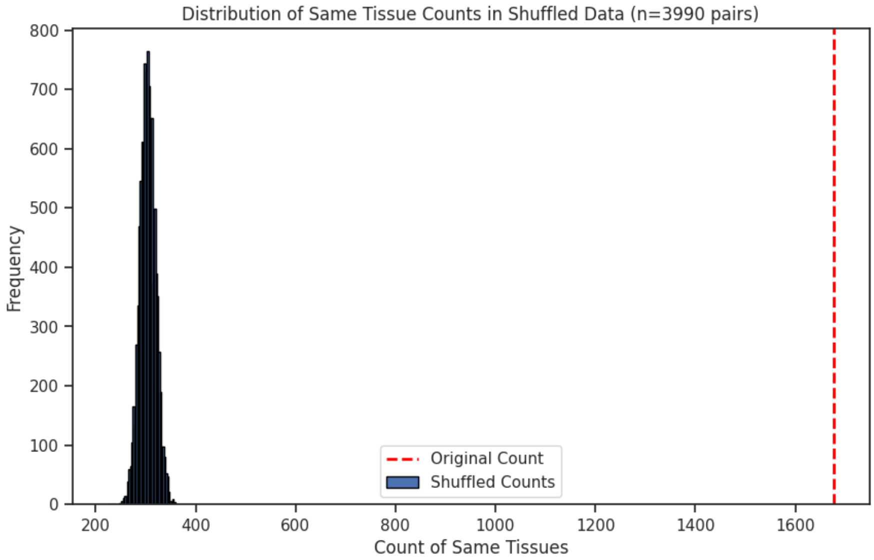


Fig 2.9. **Distribution of paralogues genes expressed mainly in the same tissue across 10,000 shuffles**. The plot shown the distribution of the same-tissue pairs after 10,000 randomizations in black, and in dotted red the actual count in the data.

Predicting Tau Scores from Gene Features Shows Limited Accuracy

After characterizing many gene-level features associated with tissue specificity, we asked whether these features were sufficient to predict the tau score. The hypothesis was that, if the features truly reflected mechanisms shaping expression patterns, a model trained on them should be able to reconstruct tau with reasonable accuracy. We therefore trained machine-learning models using all available features and evaluated their performance.

One of the central difficulties arose from the nature of the feature–tau relationships. Some features showed linear or binary behavior with respect to tau, but others followed non-linear patterns and especially inverted U-shaped trends. In such cases, genes with very low and very high tau values exhibit similar feature values, while genes in the middle of the tau range have distinct ones. This creates ambiguity for the model: two genes that are biologically opposite in tissue specificity can appear very similar in feature space. Although the inverted U-shaped relationships were not perfectly symmetrical, meaning in theory the extremes could be somewhat distinguished, in practice the signal was not strong enough to guide accurate prediction.

This challenge interacted with the distribution of the tau values themselves. Tau is strongly bimodal, with most genes either globally expressed or highly tissue-specific and relatively few in between. Faced with both ambiguous feature patterns and a distribution dominated by extremes, the model favored conservative predictions and avoided placing genes at the tails of the spectrum. As a result, the predicted tau values' distribution became more centered and substantially less bimodal than the true distribution, reflecting the model's tendency to minimize high-risk errors rather than commit to extreme predictions. Overall, the model achieved only modest performance ($R^2 = 0.47$), suggesting that the curated features are not sufficient on their own to fully reconstruct tissue-specific expression patterns. Additional dimensions of biological information — beyond what was used here — are likely required to accurately predict tau.

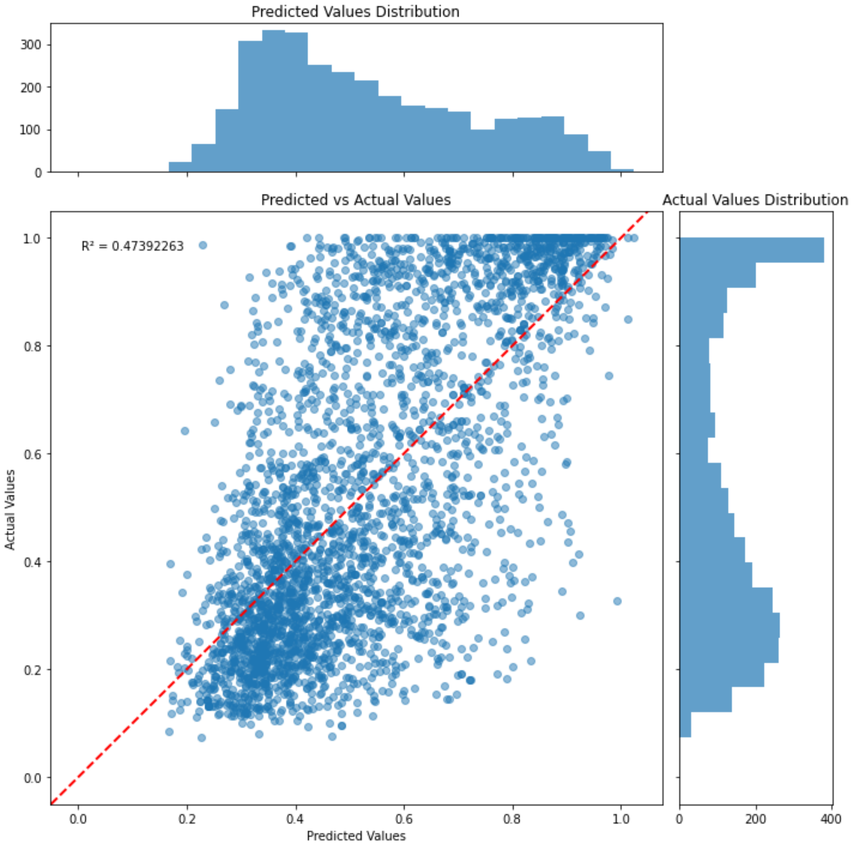


Fig 2.10 **The model's predicted tau scores compared to the actual tau scores**. The x-axis represents the predicted values and the y-axis the true tau values, and each dot is a gene. The diagonal line is the x=y line. One the sides of the plot are the distributions of the values for each axis.

Chapter 3: Gene expression and regulation reflect its evolutionary age

Evolutionary age shapes expression specificity: older genes show broad expression and younger genes are tissue-specific

The next aspect of genes we sought to explore was their evolutionary age and whether it is linked to expression patterns and regulatory elements patterns. To address this, we relied on the work of Thomas et al., 2018, which aimed to determine the age of all human coding genes. Gene age was estimated by searching multiple ortholog databases to identify the earliest ortholog. Each gene was then assigned to one of 19 phylostrata (from phylo, related to evolution, and strata, layers) based on the most distantly related species in which an ortholog is detected, ranging from genes shared by all living organisms (phylostratum 1) to primate-specific genes (phylostratum 19) ,with the final classification determined by majority vote across all database results.[26]

We first plotted the tau score density across the different ages. To complement this, we also created a quantitative plot showing the actual number of genes.

The first plot represents the density for each phylostrata individually, whereas for the quantitative one we divided tau scores into seven equally spaced bins, and the age phylostrata were grouped into seven biologically relevant categories.

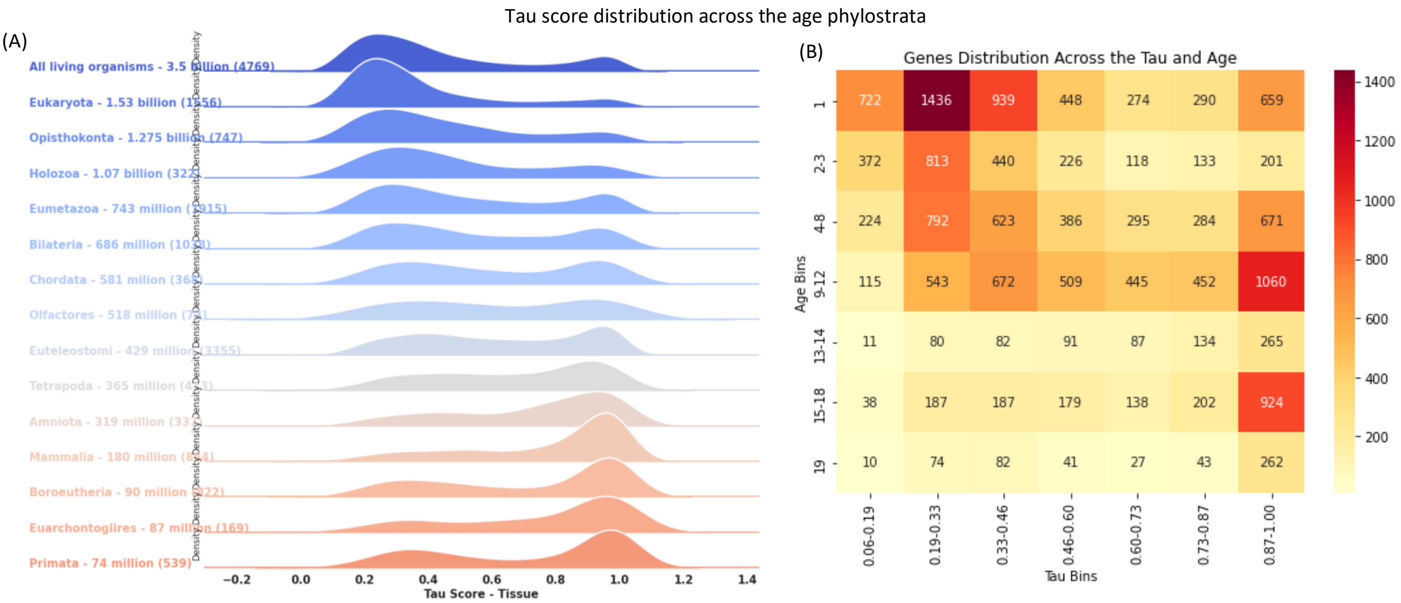


Figure 3.1. **The tau score distribution across the age phylostrata.** (A) The plot depicts the distribution of tau scores for genes across the 19 phylostrata, ordered from oldest (top) to youngest (bottom). Only 15 phylostrata are shown, as four had no assigned genes. Each phylostratum is labeled with its name, estimated age, and the number of assigned genes in parentheses. The x-axis represents the tau score, while the y-axis denotes density. (B) Quantitative plot of the number of genes for each tau score bin (bins=7) and age group (n=7). Each cell is labeled and colored by number of genes in it.

The plots show a fascinating trend – tau scores gradually shift over time from low scores indicating house-keeping genes, to high tau scores indicating tissue specificity. Older and more reserved genes tend to have a broad and unified expression patterns whereas younger and newer genes tend to be expressed in a smaller variety of tissues.

When examining the ridgeline plot, another detail emerges: while the overall trend is highly consistent, two exceptional groups stand out. One consists of old genes that are surprisingly tissue-specific, while the other includes relatively young genes with unexpectedly low tau scores.

To further characterize these two gene groups, we looked for GO tern enrichment for function, process and cell components localization. We once again used Gorilla and the list of all genes with a tau score as background[10].

The "specializing" genes – old genes with high tau scores

We refer to these genes as specializing genes because, although older genes generally tend to be expressed across many tissues, these genes instead became restricted to specific ones. It seems that, over time, they “used” their long presence in the genome to refine their expression toward particular tissues rather than expanding it as most of their counterparts did. A possible explanation could suggest that these genes are results of duplication events followed by neu- or sub-functionalization of one of the copies, creating a gene with a specific specialty from a more general one.

When trying to find a common GO categories for these genes, we have found then to be highly related mostly to iron channels and transport processes. For this analysis we used as background both the full list of human genes included in our study and the genes of high tau scores. These enrichment analysis results do not correspond to the enrichment pattern observed for all high-tau genes, suggesting that this functional bias is unique to genes that are both old and tissue-specific.

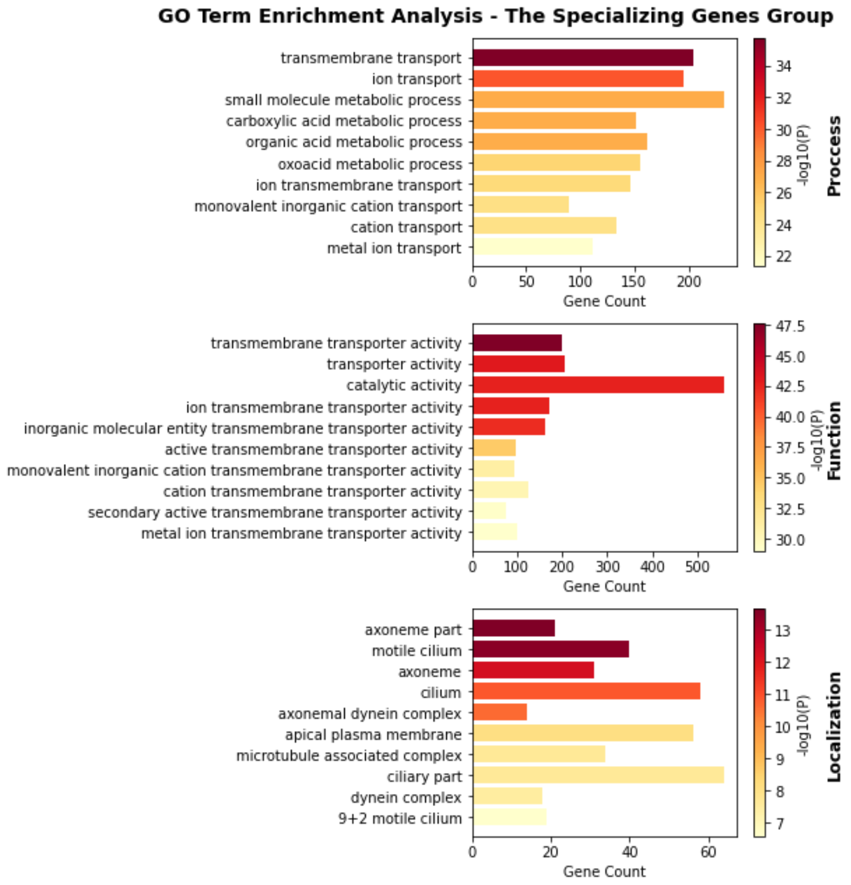


Figure 3.2. **Gorilla enrichment analysis for the specializing genes group.** The enrichment analysis was performed using the Gorilla tool. The X axis represents the number of genes belonging to the GO category and the colour represents the -log10(P) of the P values post FDR correction.

The "rapidly integrated genes" – young genes with low tau scores

The second outstanding gene group the emerges from the ridgeline plot are genes relatively young that are broadly expressed. We named them "successful gene" since their expression is embedded in most of the human cells despite not having much time in the genome. We once again ran the enrichment analysis twice – using the entire coding genes of the genome as background, and a second one with house-keeping genes as background. Using the second analysis, we wanted to check whether the features we have found are unique to that specific gene group or to genes presenting a low tau score in general.

We found that young genes with low tau scores form a distinct group of rapidly integrated genes, broadly expressed across most tissues despite their relatively recent evolutionary origin. To explore their characteristics, we used the same methodology as we did for the specializing gens and performed two enrichment analyses: one using the entire genome as a background and another comparing only genes with low tau scores against the full gene set. The second analysis aimed to determine whether the observed functional enrichments are specific to this group of young, broadly expressed genes, or represent a general property of genes with low tau values. We have found enrichment for regulation of transcription, RNA biosynthesis and DNA binding. It was also greatly enriched to the nucleus, leading us to think these gene are responsible for expression regulation.

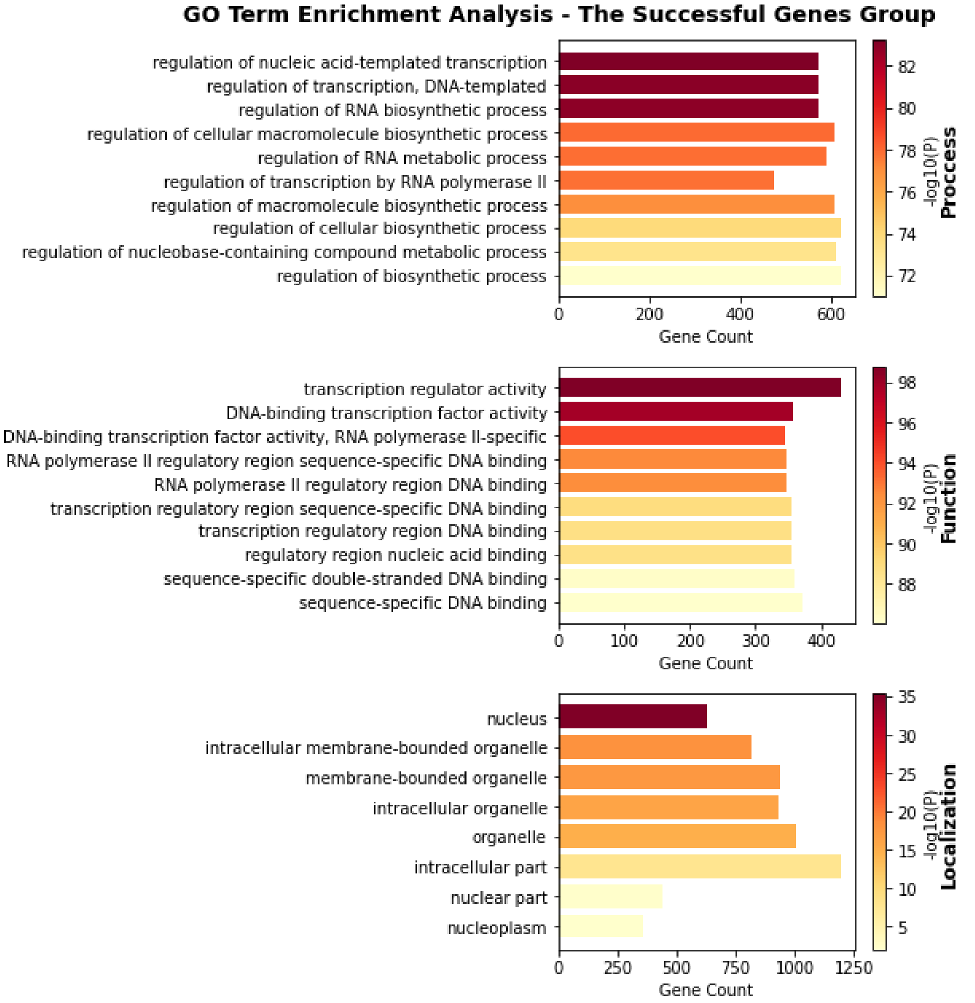


Figure 3.3. **Gorilla enrichment analysis for the rapidly integrated genes group.** The enrichment analysis was performed using the Gorilla tool. The X axis represents the number of genes belonging to the GO category and the color represents the -log10(P) of the P values post FDR correction.

Tissue specificity and regulatory information content vary across evolutionary age

Next, we asked whether regulatory architecture also changes systematically with gene age across evolutionary groups. We wanted to repeat the analysis previously done, of tau score against many regulatory features, and add time as a third dimension. Our aim was to explore whether the evolutionary age of a gene will reveal another layer of organization and meaning.

Overall, we have found no clear trend among cis- regulatory elements over time, as could be demonstrated by 3'UTR and 5'UTR. Having said that, among trans- regulatory elements we have found much stronger and clearer trends. Enhancers show a weak trend, the number of gene interactions and the number of TFs.

Examining the 3'UTR and intron lengths with respect to age and tau show no trend

We first thought to check one of our strongest signals of information content that rises with the rising complexity of gene expression, which is the mean 3'UTR length.

To examine evolutionary age effects at a biologically meaningful resolution while maintaining a sufficient group size for robust analysis, we grouped the 19 phylostrata into broader evolutionary bins. This grouping reflects major evolutionary transitions rather than absolute divergence times:

- Age 1 (Phylostratum 1; shared in all living organisms) include genes that originated at the base of cellular life, representing universal molecular machinery shared across all organisms.
- Age 2 (Phylostrata 2–3; shared only in Eukaryota and Opisthokonta) encompasses genes that arose with early eukaryotes and opisthokonts, coinciding with the emergence of compartmentalized cells and basic eukaryotic regulatory frameworks.
- Age 3 (Phylostrata 4–8; shared only in Holozoa, Metazoa, Eumetazoa, Bilateria, and Deuterostomia) span the transition to multicellularity and early animal lineages.
- Age 4 (Phylostrata 9–12; shared only in Chordata, Olfactores, Craniata and Euteleostomi) captures chordate and early vertebrate innovations.
- Age 5 (Phylostrata 13–14; shared only in Tetrapoda and Amniota) – includes tetrapod and amniote-specific genes, reflecting adaptations to terrestrial life and the consolidation of vertebrate organ systems.
- Age 6 (Phylostrata 15–18; shared only in Mammalia, Eutheria, Boroeutheria and Euarchontoglires) group mammalian and closely related lineages up to, but excluding, primates, capturing genes that arose during the mammalian radiation and its immediate evolutionary context.
- Age 7 (Phylostratum 19; shared only in Primata) is kept as a separate bin to isolate primate-specific genes, which represent the most recent evolutionary innovations in this dataset.

When grouping the genes of these 7 evolutionary phases, we have found the trend is shown mostly at the first four groups, and it gets less coherent as gene ages progress. At the oldest Phylostratum, no signal is seen. Besides the trend getting weaker and harder to detect as evolution progresses, the tau distribution of the genes shifts as expected – from mostly broadly expressed genes to genes with a more complex pattern and finally genes with minimal expressing tissues.

Taken together, these results indicate that the inverted U-shape relationship between 3'UTR length and tissue specificity is most apparent in older and mid-aged genes and becomes progressively weaker in younger evolutionary strata. This suggests that the regulatory complexity reflected by the 3'UTR length emerges and stabilizes over long evolutionary timescales, while recently evolved genes have not yet accumulated such post-transcriptional regulatory architecture.

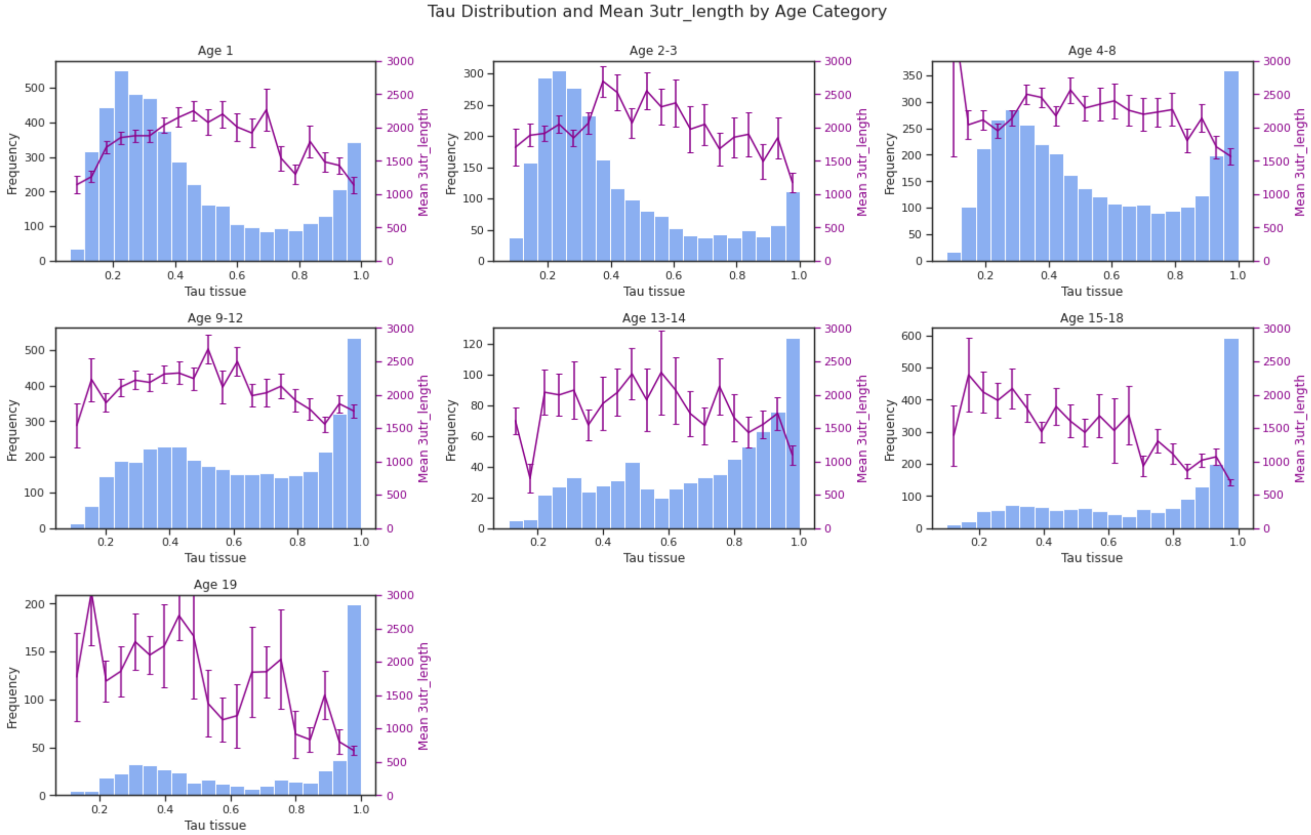


Figure 3.4. **The mean 3'UTR length per tau score bin (bin=20) across 7 age layers**. The plots are ordered from oldest genes at the upper left to oldest at the bottom. Each plot shows the tau distribution of the genes in that age bin, and in purple and the right Y axis the mean 3'UTR length.

We repeated this analysis using all other cis-element features presented so far, including intron length, CDS length and 5'UTR length, and have found very similar results. The inverted U-shape trend is found in the older genes and fades as gene age progresses up to the final phylostrata where we see a very fluctuating noise. These results led us to conclude age is not adding additional layer of information that determines the physical features of a gene.

Enhancer and protein–protein interaction complexity depend jointly on gene age and tissue specificity

The previous findings led us to ask if the evolutionary age of genes correlates better with the more complex and indirect regulatory features. So, we repeated the analysis using the number of enhancers. Once again we have used the 7 binned age categories and plotted the mean tested feature for each binned group. When looking at the mean number of enhancers per tau bin and evolutionary age bin combined {right?}, we have found that the highest count of annotated enhancers tends to display not only for the gene of mid-range tissue specificity genes, but also to genes of mid age, not the very old or very young ones. Once again, the trends are harder to detect among the very young genes, due to a very low amount of non tissue-specific genes in that group. But even so, at the 6th plot (ages 15-18) the trend of peak mean enhancer count for mid-tau genes is visible. Besides the inverted U shape, we see another interesting trend – the high of the peak is higher for the genes of the middle ages, compared to the oldest and youngest. The right Y axis on the plot has the same range in all plots, and the 3rd and 4th bins reach the most enhancers at the peak. We were amazed to find that information content is the highest in the mid-point of both axis – tau and age.

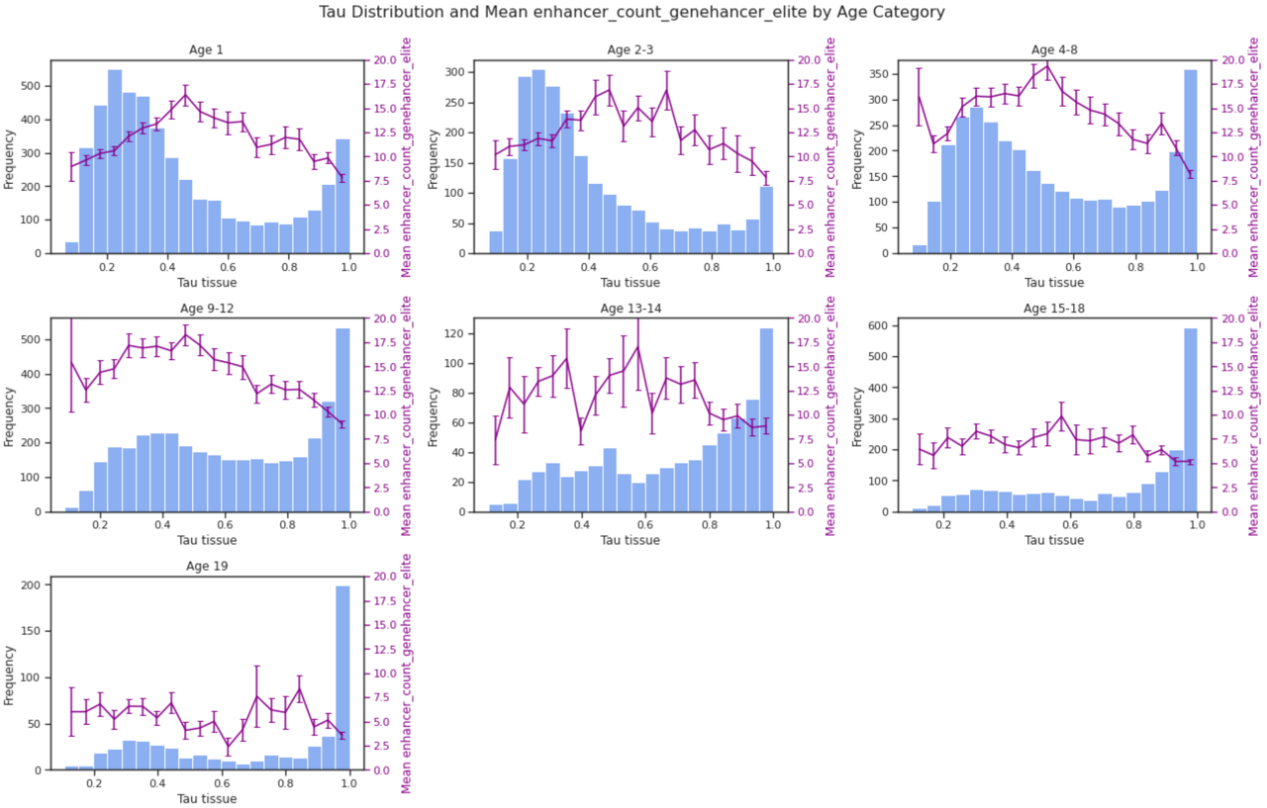


Figure 3.5. **The mean enhancers count per tau score bin (bin=20) across 7 age layers**. The plots are ordered from oldest genes at the upper left to oldest at the bottom. Each plot shows the tau distribution of the genes in that age bin, and in purple and the right Y axis the mean enhancers count.

We then checked the number of protein-protein interactions using the same method. Here we have found a new trend – the number of protein-protein interactions is higher for broadly expressed genes that are old and conserved. The steep drop, followed by a plateau, of the number if interactions, is clearly seen at the oldest gene group. At the next two plots the starting point is lower, and the platoon is the same. The 4th plot showcases a drop right at the beginning followed by a straight line, and for the rest of them the line is noisy but stable. So, all plot stable around ~20 protein interactors, but for the older genes the house keeping group is higher. It appears the effect of the breadth of expression corelating with the number of protein-protein interactions is true for conserved genes but not for newer ones. To get to the maximal number of interactions, a protein must be both broadly expressed and old.

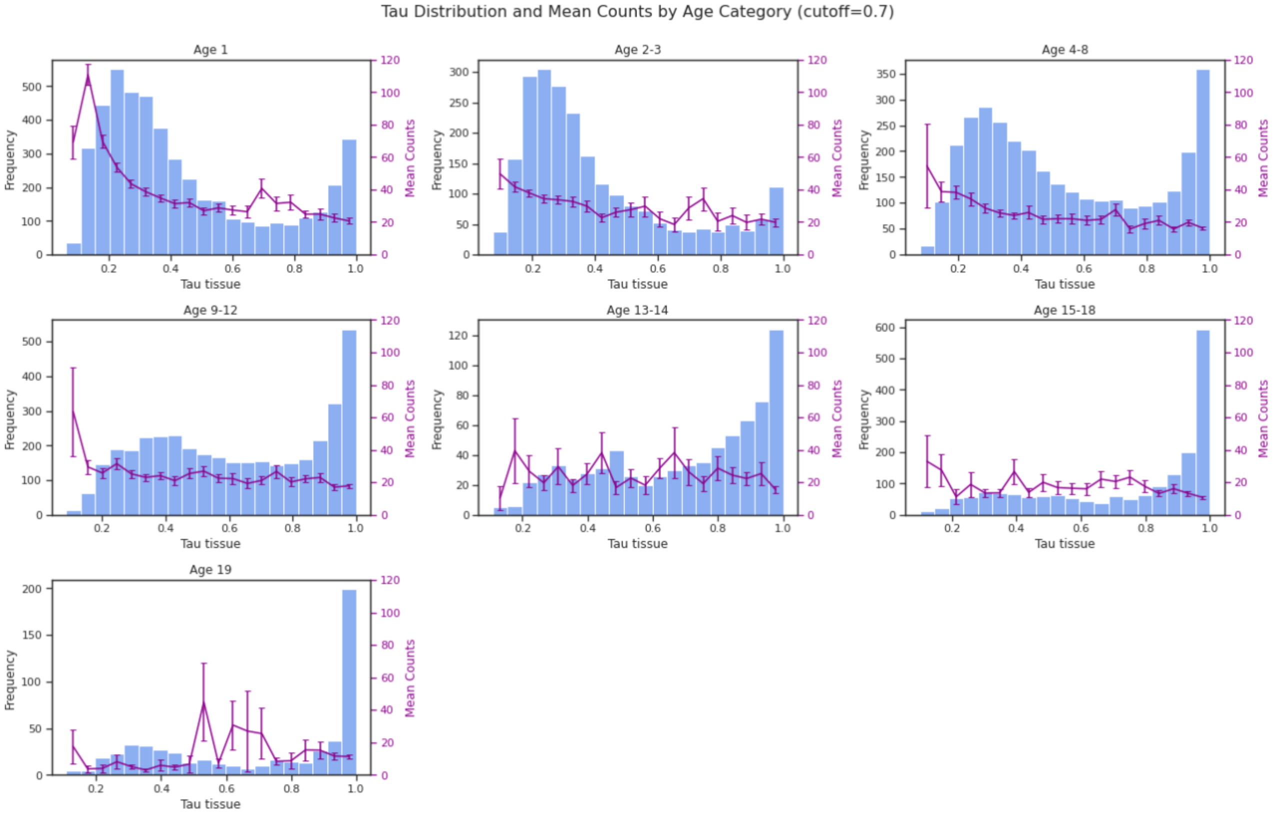


Figure 3.6. **The mean number of protein-protein interactions count per tau score bin (bin=20) across 7 age layers**. The plots are ordered from oldest genes at the upper left to oldest at the bottom. Each plot shows the tau distribution of the genes in that age bin, and in purple and the right Y axis the mean protein-protein interactions count.

Chapter 4: The MDL principle is conserved in mouse

To further evaluate the generality of our MDL principle, we asked whether the same patterns observed in humans could also be detected in another species. We chose the mouse (Mus musculus) because it is one of the most extensively studied model organisms and offers high-quality expression resources that allow meaningful comparison and replication of our analyses. In this chapter, our goal is to reproduce our findings in mouse and test whether the relationship between expression-pattern complexity and the amount of regulatory information required to encode it holds across species.

Tau scores in mouse mirror the bimodal patterns observed in humans

Our first goal was to compute tau scores for mouse genes. For this, we used the Tabula Muris Senis dataset, a public single-cell transcriptomic atlas containing over 500,000 cells from 23 tissues and organs. We downloaded the raw data and processed it using Seurat[27]. Since each annotated cell type was represented by many individual cells, we calculated the mean expression per cell type for each gene. As in the human dataset, we focused our analysis on genes with non-zero tau values.

The resulting distribution of mouse tau scores revealed a strongly bimodal pattern, highly similar to the distribution we observed in humans. We then checked whether the breadth of expression as demonstrated by the number of expressing the cell types is also still reflected by the mouse's tau scores and found another agreement.

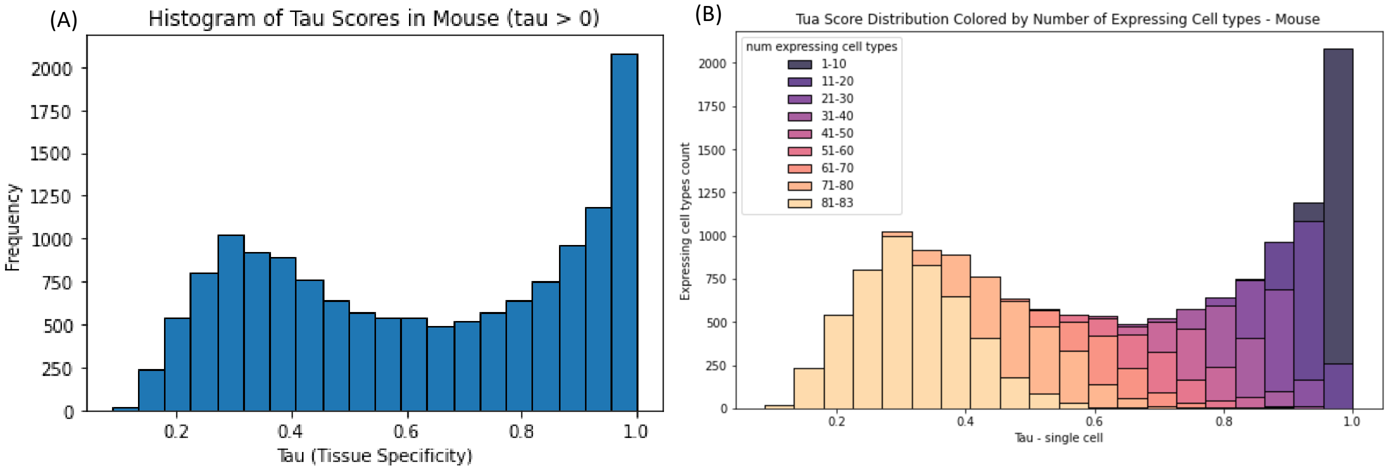


Figure 4.1. **The distribution of tau scores across the mouse genome**. (A) The tau scores are based on RNA seq single-cell data. (B) The distribution of tau scores across the human genome stacked by number of expressing tissues. Bars are colored by the number of tissues expressing the gene.

We next examined whether tau scores are conserved between humans and mice. Using Ensembl homology annotations, we mapped one-to-one orthologous gene pairs and compared their tau values. The correlation between human and mouse tau scores was high (r = 0.87), forming a clear diagonal with a limited number of outliers. This indicates that genes that are broadly expressed or tissue-specific in humans tend to exhibit similar expression patterns in mouse, supporting evolutionary conservation of tissue specificity.

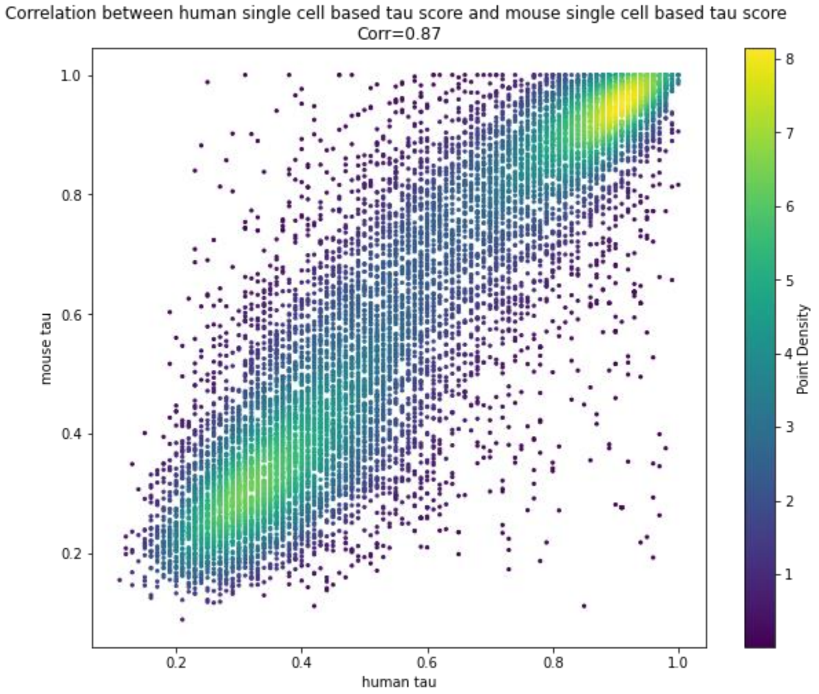


Figure 4.2. **Conservation of tissue specificity between human and mouse orthologs**. Comparison of tau scores for one-to-one human–mouse orthologous gene pairs (n = 11,235). Each point represents a gene pair, with human tau scores on the x-axis and mouse tau scores on the y-axis.

3'UTR length in mouse follows MDL-associated trends – maximized at mid-ranged tau scores

We wanted then to examines is the MDL principle discovered in human – the mid-ranged tau score genes have highest information content. Unfortunately, we lack in mouse detailed information such as enhancer lists and their association to genes. Yet several parameters of information content per genes can be more easily deduced. We started from 3' UTR length which can be computed for each gene, and that showed a clear relationship with tau in humans.

To test whether the mouse transcriptome exhibits the same MDL-associated trends observed in humans, we examined the length of the 3'UTR of mouse protein-coding genes. In humans, 3'UTR length showed one of the strongest signals of regulatory complexity, being shorter in housekeeping and highly tissue-specific genes and longest among genes with intermediate expression breadth. The 3’UTR was chosen because it represents a major regulatory hotspot, containing binding sites for RNA-binding proteins and microRNAs and influencing mRNA stability and localization. As part of the gene sequence itself, it provides a direct proxy for cis-regulatory information encoded within the transcript.

Using BioMart annotations, we calculated 3’UTR lengths for mouse genes, binned genes by tau score, and computed the mean 3’UTR length per bin. The resulting profile closely recapitulated the human pattern, exhibiting a clear inverted U-shape with a peak at intermediate tau values. This indicates that the relationship between expression-pattern complexity and regulatory information content observed in humans is also present in mouse.

Because tau scores are highly conserved between the two species, we next asked whether this similarity could simply reflect broad conservation of gene architecture across orthologs. If this were the case, genes with long or short 3’UTRs in humans would be expected to show similar lengths in mouse. To test this, we compared 3’UTR lengths of one-to-one human–mouse orthologs. While a weak positive correlation was observed, many orthologous gene pairs showed substantial divergence in 3’UTR length, including cases where one species harbors a long 3’UTR and the other a very short one.

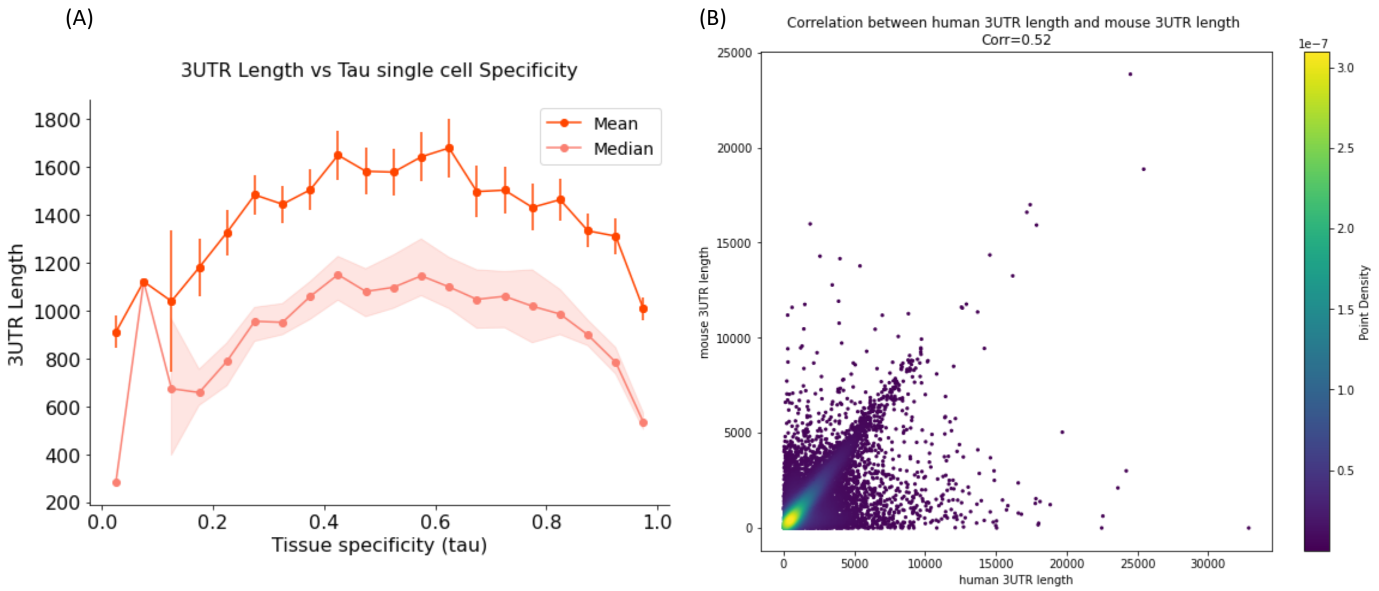


Figure 4.3. **Relationship between tissue specificity and 3’UTR length in mouse**. (A) Mean and median 3’UTR length per tau score bin. Shaded areas indicate variability across bins. (B) Scatterplot showing gene-wise mouse tau scores versus 3’UTR length, with points colored by local density.

Despite this divergence at the level of individual orthologs, the inverted U-shaped relationship between tau and 3’UTR length is preserved at the genome-wide level in both species. This indicates that the MDL-associated pattern does not arise simply from conserved gene architecture, but instead reflects an independently maintained relationship between expression-pattern complexity and regulatory information content in each species.

Comparison of regulatory features and tissue specificity for human and mouse orthologs

To further investigate the relationship between tissue specificity and regulatory features, we attempted to leverage human–mouse divergence. Because humans and mice share a recent common ancestor, differences in regulatory features between orthologous genes may reflect changes that accumulated after their evolutionary split. We focused on 3’UTR length and asked whether shifts in tissue specificity between orthologs are accompanied by corresponding changes in this feature.

Directly correlating differences in tau with differences in 3’UTR length is not straightforward, as tau does not scale monotonically with expression complexity. Transitions from low to intermediate tau represent a very different change than transitions from intermediate to high tau, and simple delta-based comparisons would therefore mix distinct signals. To address this, we divided tau scores into three categories (low, intermediate, and high) and grouped orthologous gene pairs into nine possible transition classes, representing all combinations of category changes (or lack thereof) between human and mouse. We then examined whether specific transitions in tau category were associated with the expected changes in 3’UTR length. For example, we could have expected that if a gene features a high tau in mouse and a mid-range tau in human, it may show a longer 3’UTR in human. However, across all transition classes, we did not observe a consistent or directional trend in 3’UTR length differences. Human genes have longer 3'UTR length on average and the is the trend seen in all categories, where the human-mouse delta is mostly positive and scattered in a cloud around the x axis.

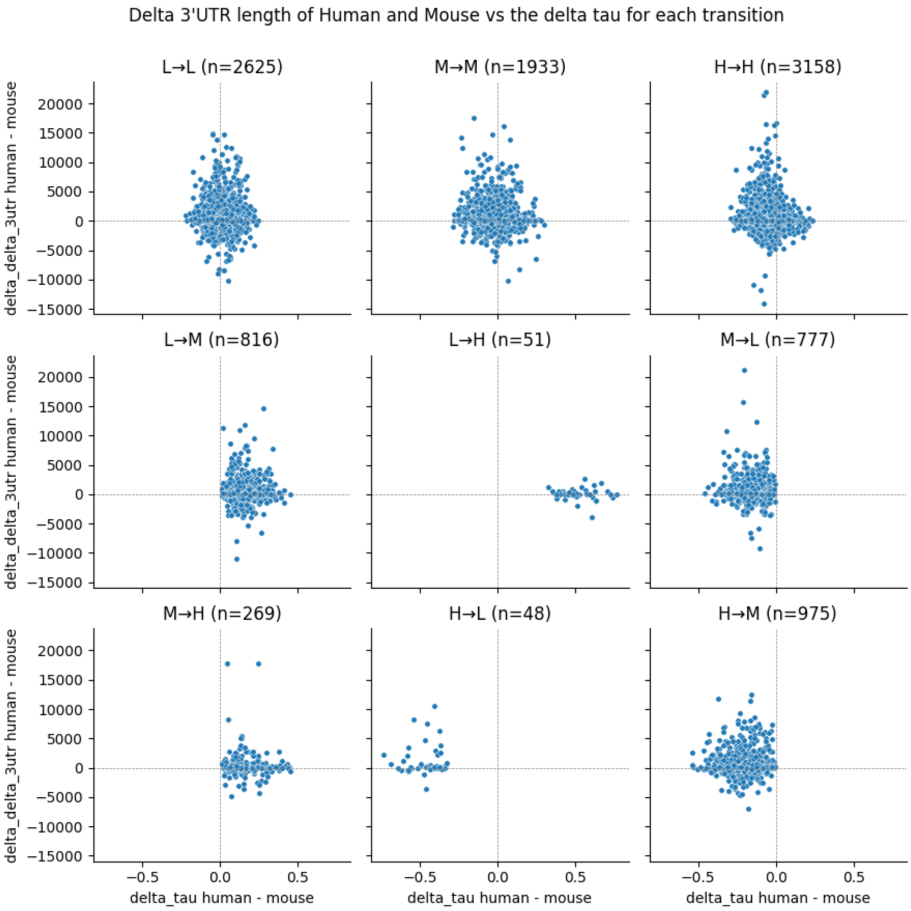


Figure 4.4. **Changes in 3’UTR length across tau category transitions between human and mouse orthologs**. Orthologous gene pairs were grouped according to transitions between low, intermediate, and high tau categories. Scatterplots show differences in 3’UTR length between species for each transition class.

Chapter 5: Genes that are “disallowed” in a single tissue in human and in mouse

Selective tissue-specific repression defines a distinct class of broadly expressed genes

As part of our effort to uncover rules governing gene expression breadth and regulation, we next focused on genes that are selectively repressed in specific tissues, called disallowed genes. Unlike housekeeping genes, which are broadly expressed, and tissue-specific genes, which are active in a restricted set of contexts, these genes are characterized by a pattern of broad expression combined with a pronounced exception in a single tissue. Such disallowed genes are of particular interest in the context of information content, as encoding an expression program of the form “expressed everywhere except in tissue X” might require both activation rules and an explicit repressive instruction, implying increased regulatory complexity compared to housekeeping genes. We therefore asked whether genes exhibiting this tissue-specific repression pattern differ in their regulatory architecture.

To identify disallowed genes, we applied the approach described by Pullen et al., who characterized genes selectively repressed in pancreatic islets [28] – for each gene, we searched for a tissue in which its expression showed a log fold change of at least 2 relative to all other tissues, and then applied a more stringent filter, retaining only genes with a median log fold change of at least 5. We applied this method using both the bulk tissue expression data and the single cell expression data, composed of 41 tissues and 81 cell types respectively, and the corresponding tau scores. When applying this method, this initial screen identified over 100 candidate genes for the 2 data combined, and the secondary filter yielded a high-confidence set of 39 disallowed genes for the tissue data and 45 for the single cell data. Of the two lists of disallowed genes, only a single gene overlapped: PKM. PKM encodes pyruvate kinase, a key glycolytic enzyme that catalyzes the transfer of a phosphoryl group from phosphoenolpyruvate to ADP, generating ATP and pyruvate. At the tissue level, PKM is disallowed in the liver, and at the single-cell level it is specifically disallowed in hepatocytes, the primary functional cell type of the liver.

The distributions pf tau scores of these disallowed genes looks closer to a normal distribution than to the bimodal genome-wide distribution.

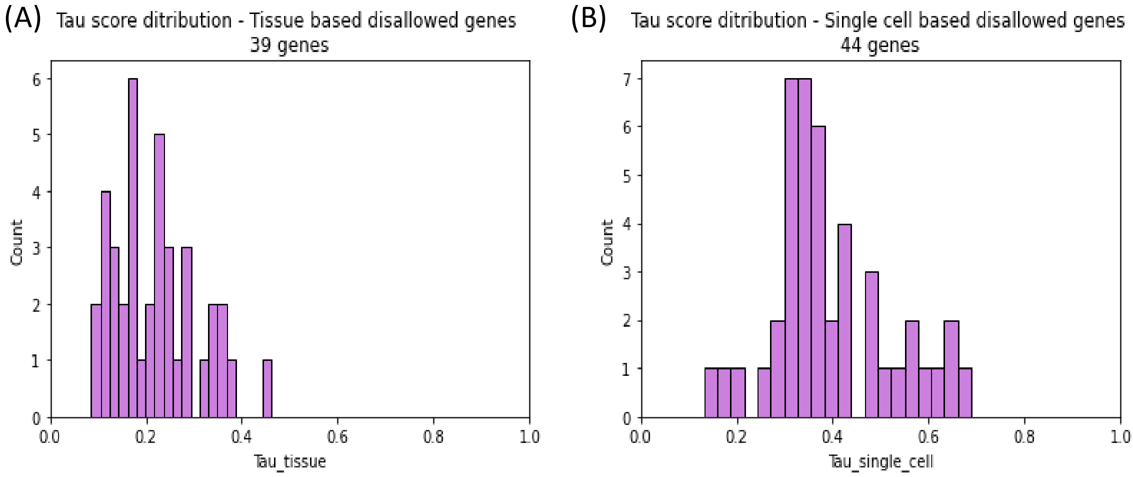


Figure 5.1. **The distribution of tau scores of the disallowed genes.** (A) The tau score distribution of disallowed genes based on bulk tissue RNA seq. (B) The tau score distribution of disallowed genes based on RNA seq single-cell data.

Disallowed genes form a statistically distinct subset of genes

To determine whether disallowed genes represent a genuine biological phenomenon rather than an expected consequence of stochastic expression variability, we performed a series of statistical tests using the gene-wise normalized expression values. Each test was performed twice – using the single cell data and the bulk tissue one.

First, for each gene, expression across tissues was converted to z-scores relative to that gene's own distribution, allowing direct comparison of expression extremeness independent of absolute expression level. We first compared the z-scores of disallowed gene–tissue pairs to the global distribution of all gene–tissue z-scores. Disallowed pairs were found to occupy the extreme left tail of the distribution, with significantly lower values than expected by chance (Mann–Whitney U test, one-sided, p= 5.88×10^{-25} for tissue and 3.51×10^{-23} for single cell), demonstrating that these genes are expressed at unusually low levels in their disallowed tissues relative to typical gene–tissue behavior.

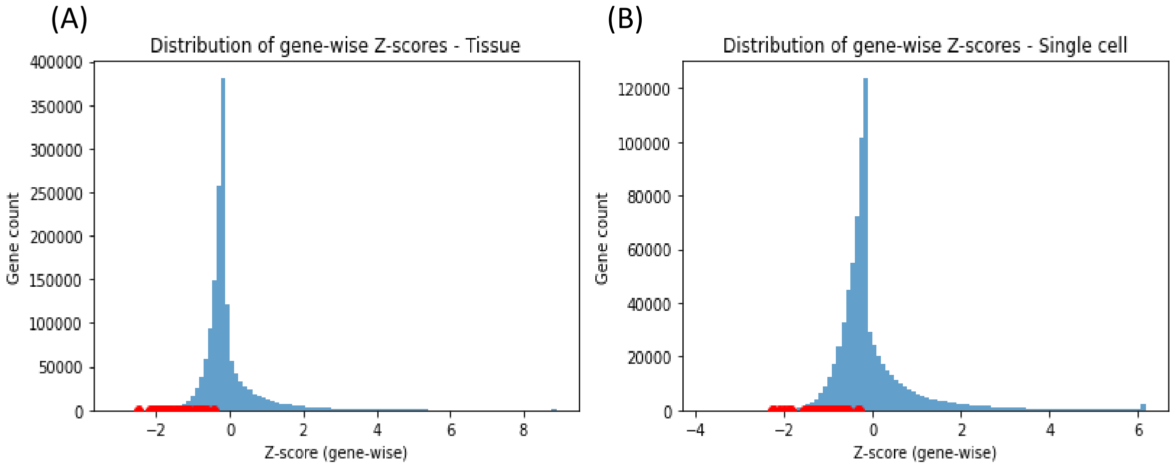


Figure 5.2. **The distribution of Z scores of all expression levels and of the disallowed genes.** (A) The per gene Z scores distribution of the tissue based data, and the disallowed genes marked in red. (B) The per gene Z scores distribution of the single cell based data, and the disallowed genes marked in red.

To exclude the possibility that this result simply reflects the fact that disallowed genes were selected based on low expression and therefore this is not a unique phenomenon, we performed a more stringent control by comparing disallowed gene–tissue z-scores to the distribution of per-gene minimum z-scores across all genes. This analysis tests whether disallowed genes are more strongly repressed than what is typical even for the lowest-expressing tissue of a gene. Disallowed genes remained significantly more extreme under this conservative null (Mann–Whitney U test, one-sided, p=1.07×10^{-13} for tissue and p=4.14×10^{-4} for single cell), indicating that their repression exceeds the level expected from normal gene-wise variability.

Together, these analyses demonstrate that disallowed genes are not statistical outliers or artifacts of thresholding but constitute a distinct subset of genes exhibiting a specific repression in a single tissue. This supports the existence of active, tissue-specific repression programs and establishes disallowed genes as a biologically meaningful class suitable for further investigation of their regulatory architecture and information content.

Distribution of disallowed tissues and cell types shows no clear mechanism

We next examined the tissues and cell types in which disallowed genes are selectively repressed. The identified disallowed contexts span a broad range of tissues and cell types, with distal tubular cells (specialized cells in the kidney) and late spermatids showing highest count in single cell data and the bone marrow showing the highest count in the bulk tissues data.

Although certain tissues appear more frequently than others as being the absence tissue of the disallowed genes, these contexts do not share an obvious developmental origin or physiological function. Instead, disallowed genes are observed across diverse biological settings, indicating that selective repression of broadly expressed genes is not restricted to a specific tissue type. This heterogeneity suggests that disallowed genes reflect a general regulatory phenomenon rather than a tissue-specific mechanism. Another explanation can be the sampling process in these particular cells and tissues or their processing. Notably, although testis and early spermatids are prominent in both analyses, they correspond to different sets of disallowed genes, indicating that dominant tissues and dominant cell types do not overlap at the gene level.

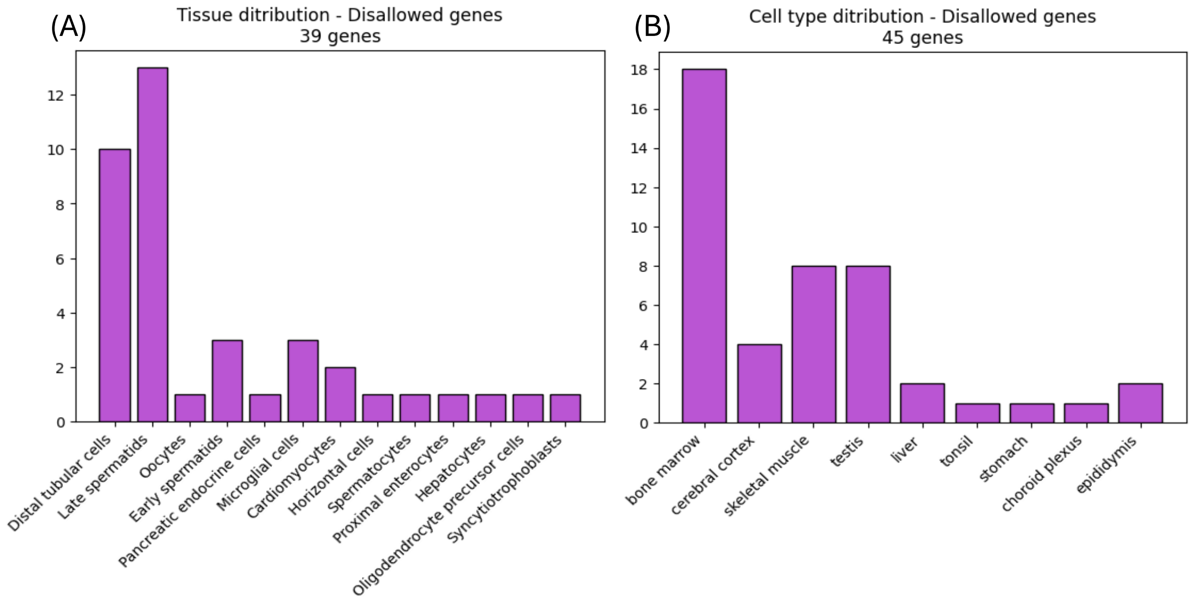


Figure 5.3. **Distribution of disallowed tissues and cell types**. (A) Number of disallowed genes per cell type identified from single-cell RNA-seq data (45 genes). (B) Number of disallowed genes per tissue identified from bulk tissue RNA-seq data (39 genes).

Selective tissue-specific repression is found in mouse

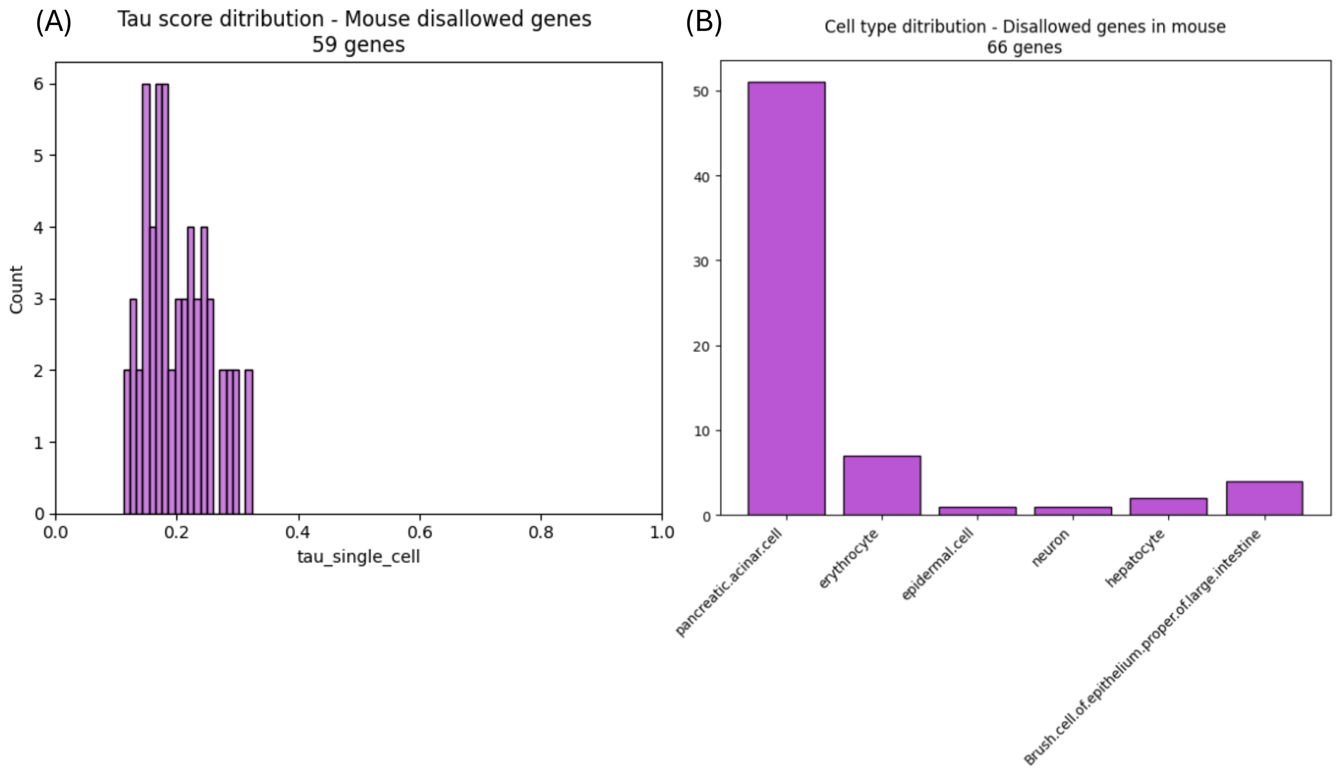
To test whether the properties of disallowed genes observed in human are conserved across species, we applied the same identification procedure to mouse single-cell expression data. Our aim was to assess whether the defining characteristics of disallowed genes—broad expression combined with selective repression in a single cell type—are also observed in mouse.

After repeating the analysis pipeline we have identified 66 disallowed gene in mouse, 59 had a tau score. As in the human data, mouse disallowed genes occupy a restricted range of low tau values, consistent with overall broad expression across cell types. Their tau distribution does not span the full range of tissue specificity, reflecting the constraint that disallowed genes must be expressed in most contexts while being selectively repressed in one.

We next examined the identity of the disallowed cell types. The vast majority of disallowed genes were associated with pancreatic acinar cells, the primary functional units of the pancreas's exocrine system, responsible for synthesizing, storing, and secreting essential digestive enzymes as inactive precursors that activate in the duodenum to break down food.

As in human, these cell types do not share an obvious common lineage or function that could explain this unique expression pattern, suggesting that selective repression of broadly expressed genes night not be restricted to a specific biological system.

Figure 5.4. **Disallowed genes in mouse single-cell data**. (A) Distribution of tau scores for disallowed genes identified in mouse single-cell expression data (59 genes). (B) Distribution of cell types in which disallowed genes are selectively repressed (66 gene–cell type pairs).



We identified a single human–mouse ortholog pair that is disallowed in both species, ITM2B, and only in the human single-cell disallowed set. The produces protein is a transmembrane protein involved in regulating amyloid precursor protein (APP) processing. The gene shows comparable tissue specificity in human and mouse (τ = 0.19 in human and τ = 0.17 in mouse), indicating conservation at the level of expression breadth. However, the context of disallowance differs between species: ITM2B is disallowed in early spermatids in human, whereas in mouse it is disallowed in pancreatic acinar cells.

The near absence of overlap between disallowed genes across species, together with the divergence in the specific disallowed cell types, suggests that disallowance is not a conserved gene-level property. Instead, it could reflect a context-dependent regulatory outcome that emerges from species-specific cellular programs and regulatory architectures.

Disallowed genes show distinct regulatory properties relative to tau-matched genes

Having established that disallowed genes form a distinct class of broadly expressed genes with selective tissue-specific repression, we next asked whether their regulatory architecture differs from that of other genes, and whether the regulatory burden required to achieve disallowed expression patterns is unique. This question must be addressed cautiously, as disallowed genes are inherently constrained to low tau values and therefore share many properties with other broadly expressed genes. A direct comparison to the full gene population would confound effects related to selective repression with effects driven simply by low tissue specificity.

To overcome this limitation, we designed a matched randomization test in which disallowed genes were compared only to genes with similar tau values that are otherwise not “disallowed”. For each disallowed gene, we defined a local pool of genes whose tau scores fell within a narrow window (±0.05) around that gene’s tau value. These pools were large, with a mean size of approximately 2,200 genes per disallowed gene, ensuring robust sampling. In each permutation, one gene was randomly sampled from the tau-matched pool for each disallowed gene, and the mean value of a given regulatory feature was computed across the sampled set. Repeating this procedure 10,000 times generated a null distribution of expected mean feature values under tau-matched sampling.

We applied this framework to a range of regulatory and structural features, including enhancer counts, gene length, untranslated region lengths, intron length, microRNA targeting, and evolutionary age. For each feature, the true mean value observed among disallowed genes was compared to the corresponding tau-matched null distribution. This approach explicitly tests whether disallowed genes differ from other genes beyond what is expected based on tissue specificity alone.

In the bulk tissue analysis, several features showed pronounced increases relative to the tau-matched expectation. Notably, disallowed genes exhibited substantially higher microRNA targeting, longer coding sequences, and longer 3’UTRs than expected for genes with similar tau scores. Other features showed more modest increases or fell within the expected range, while none showed a systematic decrease.

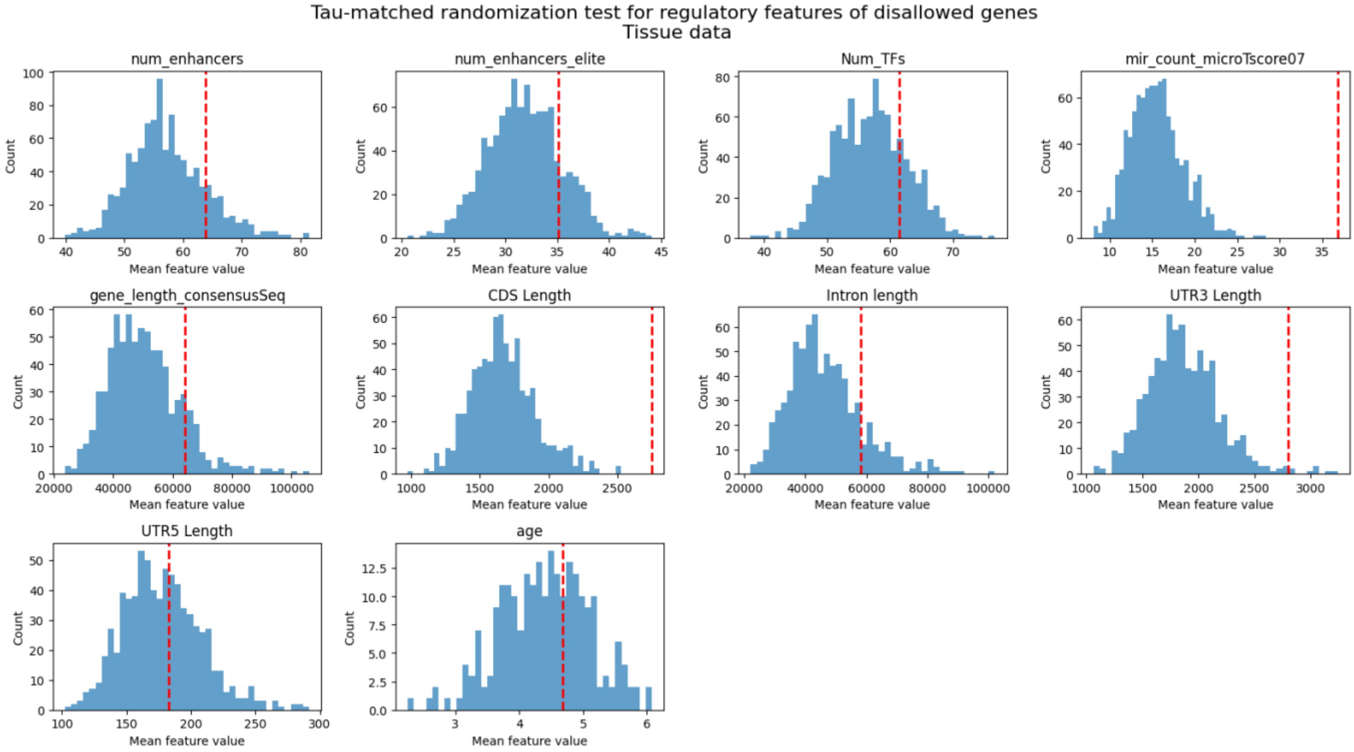


Figure 5.5. **Tau-matched randomization test for regulatory features of disallowed genes in tissue data**. For each regulatory feature, the histogram shows the null distribution of mean feature values of sampling tau-matched (±0.05) genes across 10,000 permutations. The red line indicates the true mean value observed for disallowed genes.

Together, these results suggest that selective repression at the tissue level is associated with an increased regulatory burden compared to other genes with similar expression breadth.

In contrast, the single-cell analysis revealed an almost opposite pattern. Most examined features showed slight decreases relative to the tau-matched null distribution, and several exhibited marked reductions, including 3'UTR length, coding sequence length, and intron length. These findings suggest that, at the single-cell level, disallowed expression patterns may require less regulatory information than expected for genes with comparable tissue specificity.

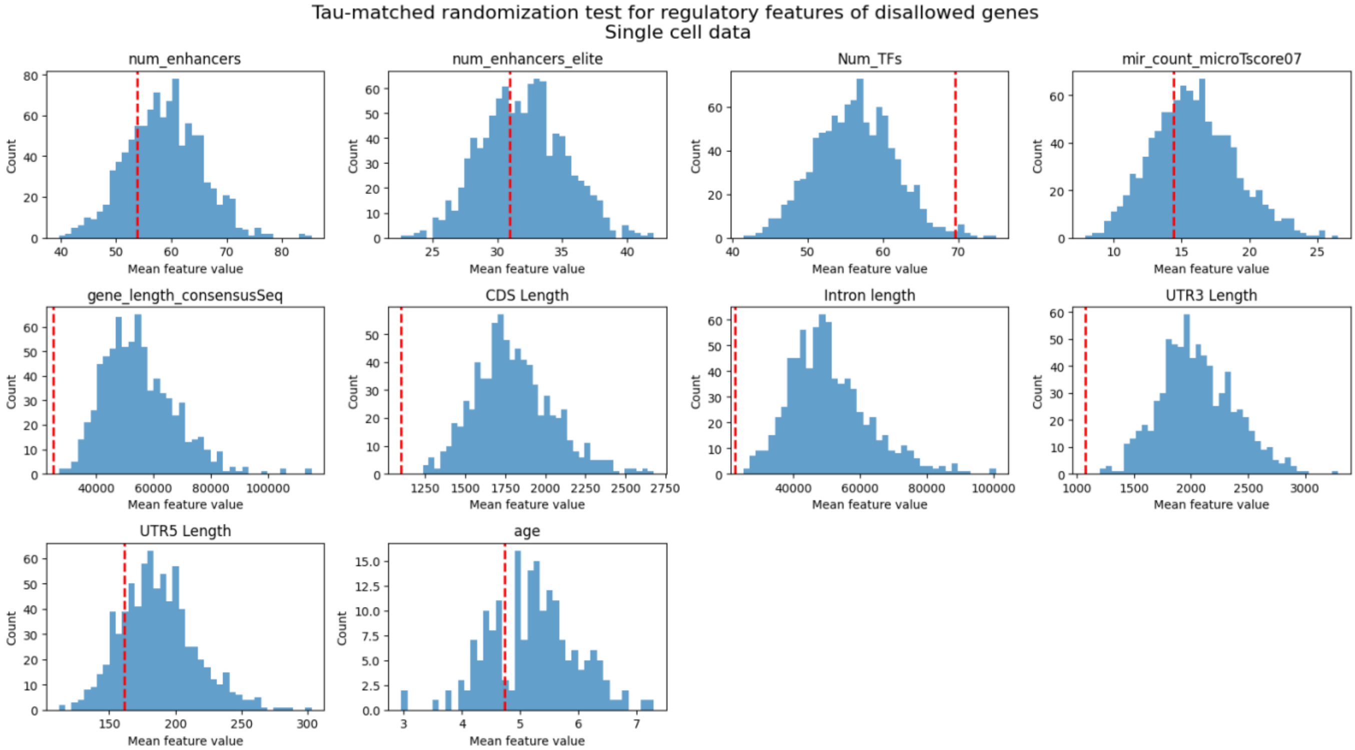


Figure 5.6. **Tau-matched randomization test for regulatory features of disallowed genes in single cell data**. For each regulatory feature, the histogram shows the null distribution of mean feature values of sampling tau-matched (±0.05) genes across 10,000 permutations. The red line indicates the true mean value observed for disallowed genes.

Taken together, the tissue-level and single-cell results likely reflect different regulatory contexts rather than a direct contradiction. Although disallowed genes were defined using a similar framework in both datasets, the biological meaning of disallowance at the tissue and single-cell levels may not be equivalent. At the tissue level, a gene classified as disallowed must be repressed across all cells contributing to the bulk tissue signal. Given the inherent cellular heterogeneity of most tissues, including multiple differentiated cell types and, this may impose a particularly stringent regulatory requirement. The enrichment for extensive miRNA targeting and longer 3’UTRs among tissue-disallowed genes is therefore consistent with a possible role for post-transcriptional mechanisms in enforcing widespread repression. In this context, miRNA-mediated regulation could provide a means to suppress low-level or residual expression across diverse cellular environments.

In contrast, disallowance observed at the single-cell level reflects gene exclusion within a specific cell identity. Such repression may occur later during differentiation and within a more narrowly defined regulatory program, potentially requiring fewer regulatory elements as the disallowed gene has a very unified expression pattern across all cells but one. The reduced regulatory burden observed for single-cell–disallowed genes may therefore indicate that cell-type–specific silencing can, in some cases, be achieved through a very simple regulatory architecture.

Importantly, these interpretations remain speculative and should be viewed as working hypotheses rather than definitive conclusions. The extent to which tissue-level and single-cell–level disallowance rely on distinct regulatory strategies, and the degree of overlap between the corresponding gene sets, will require further investigation. However, such analyses are beyond the scope of this thesis.

Chapter 6: The X chromosome abnormality presents a possible higher-level genomic organization

Chromosome X abnormality

So far, our analyses have examined the regulatory demands of individual genes across the genome. However, we observed that these principles may also extend to higher-order genomic organization. In agreement with previous studies, we found that genes located on the X chromosome (chrX) exhibit a significant bias toward tissue-specific expression[29,30]. Specifically, 25.3% of chrX genes exhibited high tissue specificity (tau ≥ 0.95), compared to 14.5% genome-wide. Focusing on these tissue-specific chrX genes, we found that they are associated with a significantly lower number of enhancers compared to all tissue-specific genes across the genome (6.1.B). This indicates that the regulatory demand for tissue-specific genes on chrX is lower than expected, even given their restricted expression patterns.

According to our hypothesis, genes of extreme tau scores, both low and high, require less regulatory elements to achieve their desired expression pattern. Since genes on chrX tend to be very tissue specific, we next examined the fraction of enhancers sequences in each chromosome. In the figure (6.1.B), the chromosomes are ordered by size and colored by enhancer sequence content, and chrX stands out as having much less enhancers in it.

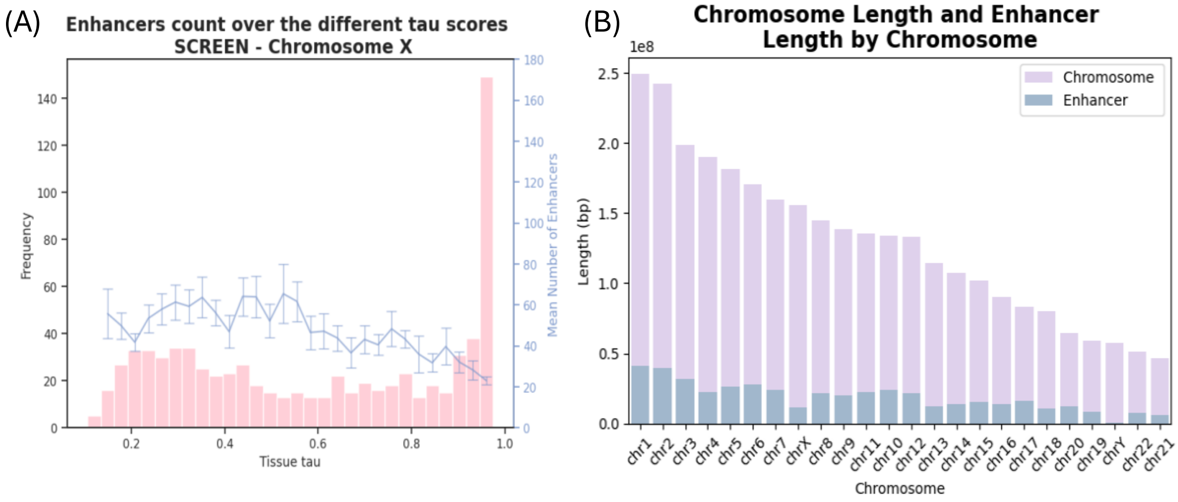


Figure 6.1. **ChrX tissue specificity and enhancers content**. (A) Histogram of the distribution of the tau score of chrX genes (in pink) and overlaid in blue is the mean number of enhancers per tau bin. (B) Histogram of the human chromosomes length, with the fraction of enhancers sequences colored in blue. The chromosomes are ordered by length (bp).

To understand this further, we asked whether these genes are enriched for specific tissues. As previously observed[31], we found a strong enrichment for testis-specific expression: approximately 65% of the tissue-specific genes on chrX are testis-specific.

These findings reveal that chrX harbors a disproportionately high number of testis-specific genes with lower-than-expected enhancers count. This pattern supports that MDL principles may apply not only to individual gene expression patterns but also to higher-order genomic organization. From an MDL perspective, when many nearby genes share a simple expression rule, such as “express in testis”, the collective regulatory burden can be reduced, minimizing the description length required. Such compression at the chromosome level could reflect a broader organizational principle, potentially shaping how regulatory programs are spatially arranged across the genome. While this observation aligns with the MDL framework, alternative explanations for the low enhancers count on chrX are possible and warrant further investigation.

Chapter 7: Exploring the validity of the tau score

Evaluating the tau score as a measure of tissue specificity

Tissue and cell-type specificity are inherently complex properties of gene expression, and no single scalar metric can fully capture all their aspects. In this work, the tau score was not used as a complete description of expression behavior, but rather as a robust summary measure of expression breadth and variability that allows systematic, genome-wide comparisons. Importantly, most analyses presented here focus on relative trends across the tau spectrum rather than on precise tau values for individual genes.

Concerns regarding the robustness of tau have been addressed extensively in previous benchmarking studies. In particular, Kryuchkova-Mostacci and Robinson-Rechavi[32] showed that tau performs consistently across tissue subsampling, normalization schemes, and datasets with varying numbers of tissues and expression distributions, and ranks among the most stable metrics for identifying tissue-specific genes.

During our work, the Gini index[33] came up as one of the closest alternative measures. Therefore, we directly compared tau and the Gini index in our datasets and observed a very strong agreement between the two measures. Both metrics show similar distributions across genes and are highly correlated. This concordance suggests that the main conclusions of this study are not driven the tau score properties, but rather reflect the properties of gene expression.

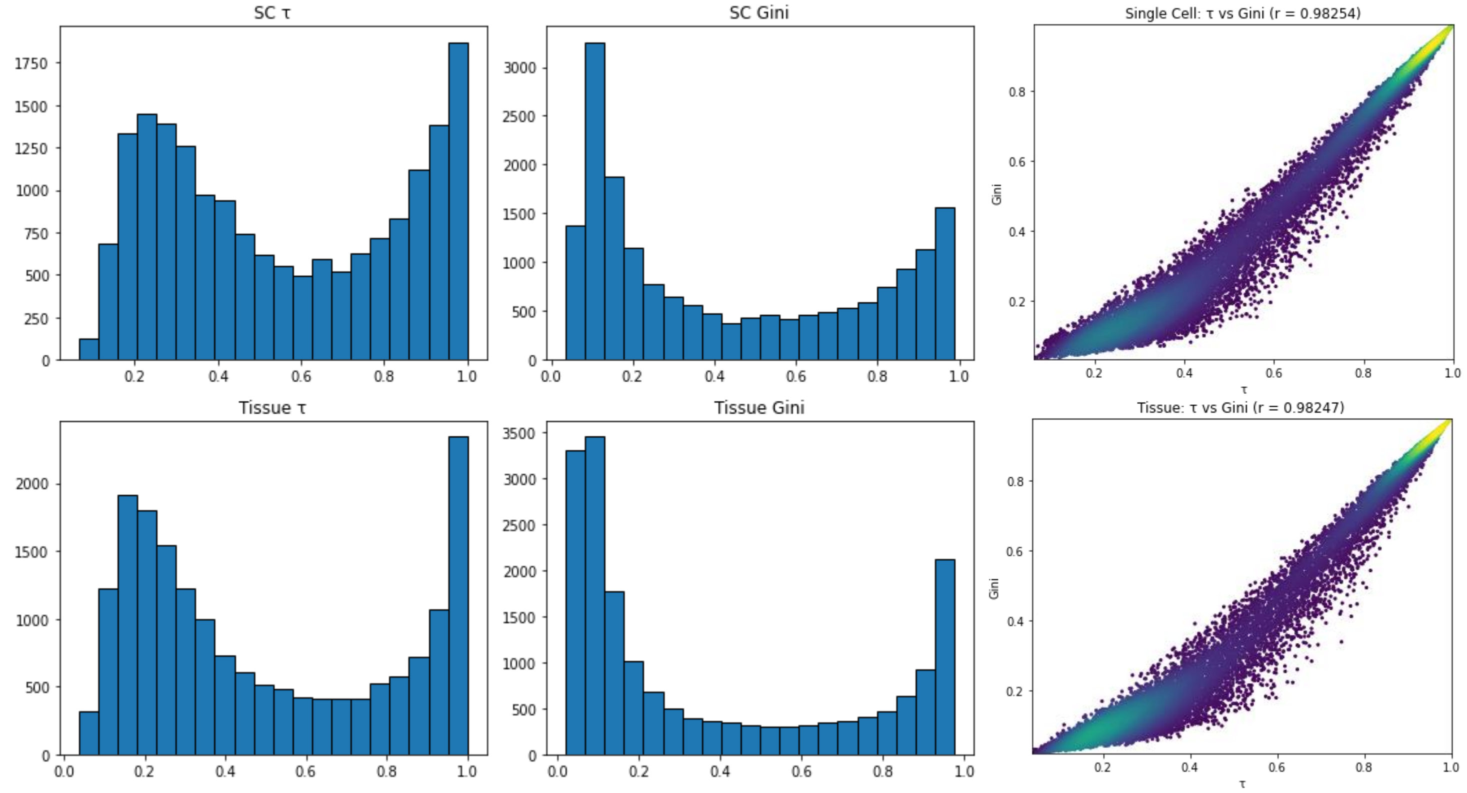


Figure 7.1. **Comparison between tau score and the Gini index as measures of tissue specificity.** Top panels show the genome-wide distributions of tau scores (A) and Gini index values (B), and a gene-wise comparison between the two measures, with each point representing a gene and colored by local point density (C). Bottom panels show the corresponding distributions and gene-wise comparison for tissue-based tau and Gini scores (D–F).

To further examine what the tau score captures, we tested its relationship with the variability of a gene's expression across tissues or cell types. For each gene, we calculated the standard deviation of its expression profile across the sampled tissues and cell types, and correlated these values with the corresponding tau scores. In both datasets, the biggest variation and the highest STD values belonged to genes with intermediate tau scores, indicating that these scores indeed represent the most heterogeneous expression profiles. This supports the view that tau reflects the spread and unevenness of its expression across tissues and cells.

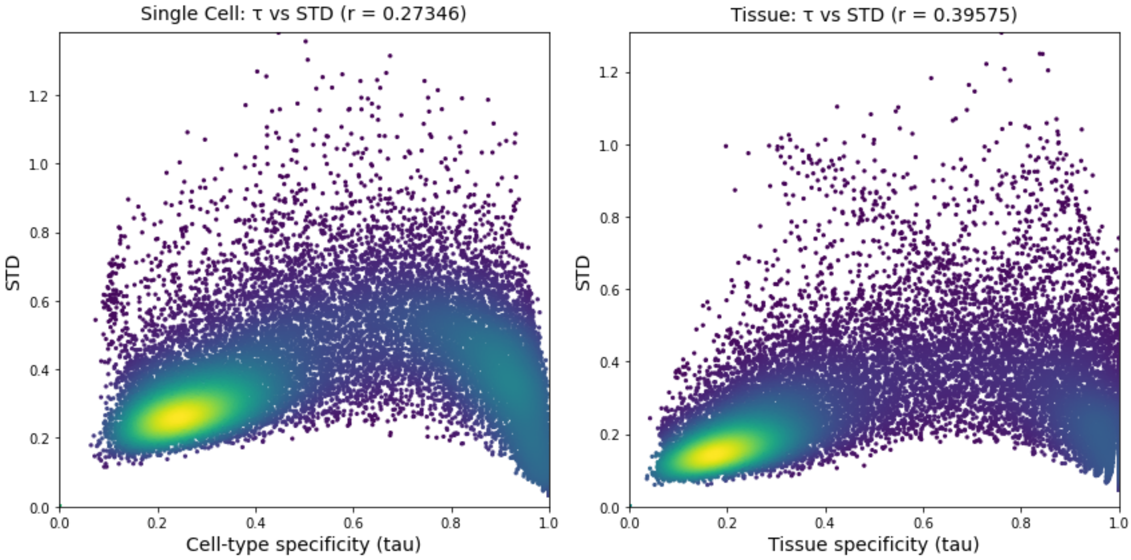


Figure 7.2. **Relationship between tau score and expression STD across tissues and cell types**. Scatterplots show gene-wise tau scores versus the standard deviation of expression across single-cell types (A) and bulk tissues (B). Each point represents a gene and is colored by local point density. Pearson correlations were computed after excluding missing values.

Highly studied genes do not drive tau-associated trends

Because many regulatory features analyzed in this work rely on existing annotations, we asked whether the observed trends could be driven by biases in scientific attention. Genes that are studied more extensively may have more annotated interactions or regulatory elements, potentially inflating apparent regulatory complexity independently of biology.

To assess this, we first checked the number of published papers that contain a gene name, using gene symbol, in the PubMed website[34] but have learnt that this methos is inaccurate; some genes have multiple names, some have names that comprise other commonly used words (like a gene named IMPACT) and overall this analysis could not count unindexed textual content. We then moved to the Gene2Pubmed dataset by NCBI[35], to quantify number of publications per gene, excluding large-scale screens studies, defined as those studying more than 100,000 genes at once. As expected, broadly expressed genes are associated with substantially more publications, reflecting their central role in basic cellular processes and frequent use as experimental controls. Importantly, despite this strong publication bias towards broadly expressed genes, these genes do not exhibit elevated values for most regulatory features examined in this study. When computing correlations between publication count and the other features examined in this work, we found that these associations are generally weak (ranging from −0.07 to 0.13), with the exception of protein–protein interaction count, which shows a moderate correlation (r = 0.47).

Crucially, we did not observe increased publication counts for genes with intermediate tissue specificity, even though these genes show the highest regulatory complexity. If annotation bias were driving the observed non-monotonic patterns, intermediate-tau genes would be expected to be over-represented in the literature. Instead, publication counts decline monotonically with increasing tau.

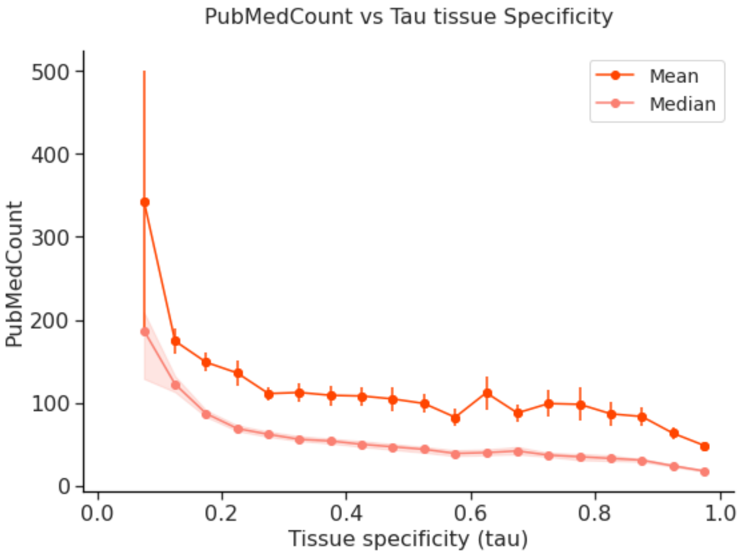


Figure 7.3. **Scientific attention as a function of tissue specificity**. Mean and median number of publications per gene as a function of tau score, based on Gene2PubMed associations after excluding large-scale screens.

This relationship is stable across decades throughput the 20th and 21st Centuries, indicating that the lack of enrichment for intermediate-tau genes is not specific to a particular research era (Figure 7.4). Together, these results argue that the central findings of this study are unlikely to be artifacts of differential scientific attention and further support the validity of tau as a meaningful organizing axis for gene expression complexity.

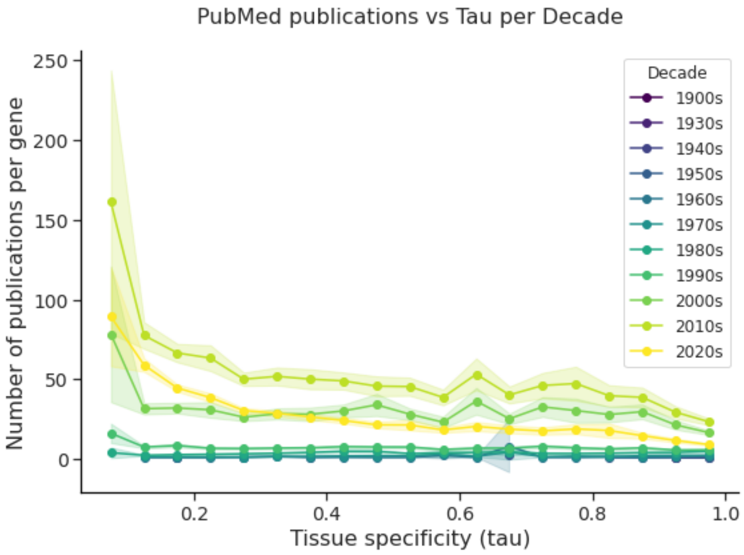


Figure 7.4. **Scientific attention across decades as a function of tissue specificity.** Mean and median number of publications per gene per tau score bin, stratified by decade of publication.

Finally, we asked whether a small number of extremely highly-studied genes could disproportionately influence the mean-based bibliometric analyses. Such genes might reflect community-wide experimental practices, for example commonly used control genes, and could obscure broader trends. Indeed, we identified a single major outlier—TP53, a central cancer- and cell-cycle–related gene—with a substantially higher publication count than all others. Aside from this exception, the most highly studied genes are distributed broadly across the tau spectrum, rather than clustering at low or high tissue specificity. The overall distribution of publication counts per gene is relatively uniform across tau values, with a slight bimodal pattern and modest enrichment at low tau values that mirrors the global tau distribution (Figure 7.5). This indicates that neither extreme outliers nor concentrated clusters of highly studied genes are responsible for the observed relationships between tissue specificity and regulatory features.

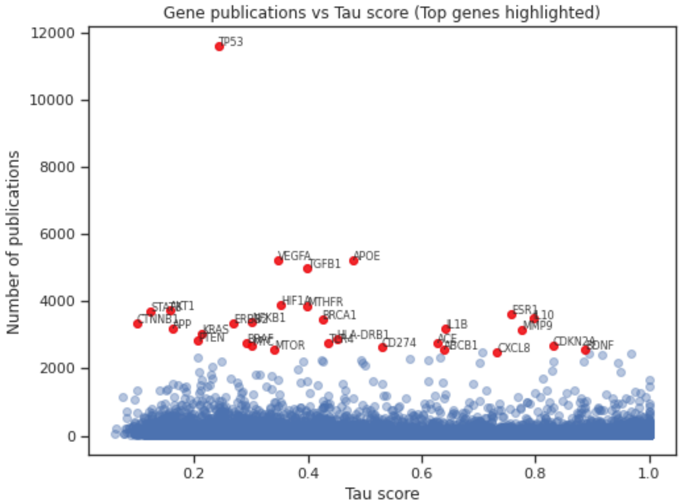


Figure 7.5. **Distribution of highly studied genes across the tau spectrum**. Scatterplot showing the number of publications per gene as a function of tau score. Each point represents a gene; the most highly studied genes are highlighted and labeled.

Discussion

In this work, we examined tissue specificity of human genes through the lens of information content, using the tau score as a continuous measure for expression pattern complexity. Rather than classifying genes as either housekeeping or tissue-specific, we analyzed the full spectrum of expression patterns across tissues and cell types. Across multiple regulatory and structural features, we observed a consistent non-monotonic relationship: genes with intermediate tissue specificity exhibit the highest regulatory complexity, while both broadly expressed and highly tissue-specific genes show simpler regulatory architectures.

This non-monotonic pattern was consistently observed across diverse regulatory and structural gene features, including gene architecture, enhancer associations, and evolutionary context, and was independently reproduced in mouse. Notably, although these features exhibit similar trends with tissue specificity, they show only weak pairwise correlations with one another, indicating that the observed signal does not arise from redundancy among measurements. Instead, the repeated emergence of the same pattern across largely independent features supports the existence of a robust and general relationship between expression-pattern complexity and regulatory information content.

Although our analyses do not establish a mechanistic role for the minimum description length (MDL) principle, they are consistent with its predictions: expression programs that cannot be described by simple rules appear to require greater regulatory information.

Incorporating evolutionary age adds a temporal dimension to this framework. The results also show that broadly expressed genes tend to be evolutionarily older, whereas younger genes are more often highly tissue-specific. However, regulatory complexity does not increase monotonically with age. Instead, genes of intermediate evolutionary age exhibit the highest overall regulatory burden, mirroring the pattern observed for intermediate tissue specificity. In contrast, interactions with fine-tuning regulatory layers such as microRNAs and protein–protein interactions appear to depend on a combination of both gene age and expression breadth, suggesting that these features accumulate gradually over evolutionary time rather than directly encoding expression-pattern complexity.

Analysis of disallowed genes highlights selective repression as a distinct and complex phenomenon. These genes are broadly expressed yet selectively repressed in a single tissue, implying expression programs that encode explicit negative regulation. Although disallowance is clearly non-random and statistically significant, we observed very limited overlap between disallowed genes identified at the bulk tissue level and those identified in single-cell data, as well as minimal conservation between human and mouse.

Moreover, neither the identity of disallowed genes nor the tissues or cell types in which repression occurs showed strong consistent agreement across datasets or species.

We initially hypothesized that disallowed expression patterns would require increased regulatory burden, reflecting the need to encode explicit repression. However, regulatory features associated with disallowed genes showed mixed and sometimes opposing trends depending on data resolution. At the tissue level, disallowed genes were enriched for post-transcriptional regulatory features, particularly microRNA targeting and longer 3'UTRs, whereas in single-cell analyses several structural features appeared reduced relative to tau-matched expectations. One possible explanation is that disallowance reflects different regulatory challenges at different biological scales: repression across all cells within a tissue may require more robust regulatory mechanisms, while repression confined to a single cell type may be achieved later in differentiation and with fewer regulatory components. Consistent with this interpretation, paralogous gene pairs that diverged toward strong tissue specificity often exhibited reduced regulatory complexity. These observations remain speculative and suggest that gene disallowance represents a flexible, context-dependent regulatory outcome rather than a single conserved mechanism.

Chromosome X and the testis emerged as a notable exception in this study. We found that the testis is often the main expressing tissue for highly tissue-specific human genes, and that testis-specific genes are strongly enriched on chromosome X. X-linked genes also tend to exhibit high tau scores, reflecting narrow and specialized expression patterns, and are associated with fewer enhancers than expected based on chromosome size. This reduced enhancer load suggests that alternative regulatory strategies may operate on chromosome X. Together, these observations point to a higher-order organizational principle in which chromosome X clusters genes with similar expression and regulatory requirements, particularly in relation to testis-specific programs.

Several limitations regarding our work should be noted. Tau compresses diverse expression patterns, and regulatory complexity was inferred from annotated features rather than measured directly. In our work, we don't directly quantify the MDL of the genes, and using tau lacks complexity as it does not capture the complexity of expression between tissue that are more related than others. In another work from the lab, regulatory information was quantified using a measure termed tMDL (tree-aware MDL), which incorporates tissue and cell lineage structure and estimates the number of expression changes using a parsimonious framework. The analyses presented here demonstrate correlation rather than causation and establishing causality would likely require large-scale perturbation experiments. Accordingly, the MDL framework presented here should be viewed as a descriptive and unifying perspective rather than a mechanistic explanation.

In addition, confidence based on the replication of the results in mice should be limited as the evolutionary distance between humans and mice is small. Do strongly support our finding, they should be examined using a model animal further down the evolutionary tree.

Future work could test whether tissue specificity has a spatial genomic component shared by closely located genes, extend the analysis to more distantly related species, and aim to quantify regulatory information more directly. In addition, further investigation will be required to clarify the regulatory mechanisms underlying gene disallowance and explore how they differ across biological scales.

Declaration of specific contributions

Figures	Data collected by	Data analyzed by	Notes
1.1-1.2, 1.5-1.7	Me	Me	Data collected and analyzed with guidance from Ruthie Golomb (PhD student)
2.1-2.3, 2.6	Me	Me	Data collected and analyzed with guidance from Ruthie Golomb (PhD student)
2.7-2.8	Simon Fishilevich	Me	
3.2-3.6	Me	Me	Data collected and analyzed with guidance from Ruthie Golomb (PhD student)
4.1	Tsviya Olender	Tsvia Olender and Me	Data collected and analyzed by Tsvia, further analysis was done by me
5.1, 5.5-5.6	Me	Me	Data collected and analyzed with guidance from Ruthie Golomb (PhD student)
6.1	Ruthie Golomb, Bar Cohen	Bar Cohen, Sapir Savariego and Me	Data collected and analyzed with guidance from Ruthie Golomb (PhD student), analysis by Bar Cohen (MSc student), Sapir Savariego (MSc student) and me
7.1	Me	Me	Data collected and analyzed with guidance from Ruthie Golomb (PhD student)
7.2	Me	Me	Data analysis in collaboration with Yotam Zigler (MSc student)
7.3-7.5	Me	Me	Data collected by me, with guidance from Dr. Noam Hadar and Dvir Dahari

All other figures contain data collected and analyzed by me.

Methods

Tau score

Tau score calculation

The tau scores were calculated based on the single-cell and tissue level RNA seq datasets from the Human Protein Atlas program.

On the raw expression values, we used a threshold of nTPM=1, and any value smaller was considered 1. We then added 1 to the gene levels and log transformed them and normalized each gene's expression values to the highest value for that gene. Using that expression values, we calculated the, based on the formula presented by Yanai et al., Bioinformatics, 2005:

$$\tau = \frac{\sum_{i=1}^{N}(1 - x_i)}{N - 1}$$

In the formula, x_i is the normalized expression value per tissue/cell and N is the number of tissues/cell types in the data.

We included in further analysis only gene with a tau score for both a single-cell tau score and tissue tau score, resulting in 18,235 genes.

Gene families

Gene family annotation

Gene family membership was obtained from HGNC and mapped to the gene symbols, gene-to-family assignments, and gene family metadata.

Gene length feature

We used the downloadable data from the Ensembl project, available through the BioMart data mining tool, to find the genes' transcript length, intron length and UTRs length. We used the "Ensembl Genes 113" database for human genes (GRCh38.p14). For the intron length we used the following formula:

Intron length = abs(transcript start (bp) – transcript end (bp) + 1) – transcript length including UTRs and CDS

Enhancers count per gene

The find the number of enhancers associated to a gene, we used 2 different databases:

GeneHancer (a database by GeneCards, integrating 9 databases of both experimentally validated and predicted interactions) and SCREEN (a database by ENCODE).

miRNA targeting annotation and counting

miRNA–gene interactions were obtained from the TarBase v9 database. Interactions were filtered using a microT score threshold of 0.7 to retain higher-confidence regulatory associations. For each gene, we counted the number of unique miRNAs targeting it and genes without annotated interactions were assigned a value of zero.

Protein–protein interaction analysis

To examine the relationship between gene interaction hubness and tissue specificity, we used protein–protein interaction data from the STRING database (v12.0). Human interaction data were downloaded from STRING (9606.protein.links.v12.0.txt), which provides pairwise protein interactions along with a combined confidence score. STRING protein identifiers were mapped to human gene symbols using the STRING aliases file (9606.protein.aliases.v12.0.txt), restricting mappings to Ensembl HGNC symbols. In cases where multiple protein identifiers mapped to the same gene symbol, a single representative protein was selected based on UniProt annotations to avoid duplication.

Interaction networks were filtered using three confidence thresholds: ≥0.4 (medium confidence), ≥0.7 (high confidence), and ≥0.9 (very high confidence). In this work we used the high confidence threshold of 0.7 as recommended by STRING.

CpG island annotation and count

We used the UCSC dataset regarding CpG islands' genetic location. CpG islands are determined by GC content of 50% or more, sequence length greater than 200bp and Obs/Exp ratio of 0.6 or higher. To annotate then to the corresponding coding genes, we used the UCSC coding genes dataset. We used the gene's TSS (transcription start site) and a window of 3000bp in total, meaning 1500bp upstream and downstream to it.

Predicting tissue specificity from gene features using machine learning

To test whether tissue specificity can be predicted from the regulatory and structural features used throughout this thesis, we trained a supervised regression model to predict gene-level tau scores. We used a combined gene feature table as input. The target variable was the tissue-based tau score, and the features included the remaining gene-level annotations after excluding identifier columns, other tau-derived variables, and gene age, which is highly correlated with tau and could therefore act as a shortcut feature. The resulting feature matrix was randomly split into training (80%) and test (20%) sets.

We trained a gradient-boosted decision tree regressor (LightGBM, LGBMRegressor) using default parameters and a fixed random seed. Model performance was evaluated on the held-out test set using mean squared error (MSE), root mean squared error (RMSE), and the correlation between predicted and observed tau values (computed as $\surd R^2$). Predicted versus observed values were visualized in a scatter plot with a y = x reference line.

Gene evolutionary age

Gene evolutionary age was obtained from the phylostratigraphic annotation of Thomas et al. (2018), which assigns each human protein-coding gene to one of 19 phylostrata based on the most distantly related species in which an ortholog is detected. Orthology was assessed across multiple databases, and final age assignment was determined by majority vote. Phylostrata range from universally conserved genes (phylostratum 1) to primate-specific genes (phylostratum 19).

Mouse expression and genomic data

Human–mouse orthology mapping

Human–mouse homolog pairs were obtained from Ensembl Compara (release 115) using the "protein_default" homologies table for Mus musculus.

Mouse tau score calculation (single-cell atlas)

Mouse single-cell expression data (cell-type–averaged, pre-normalized) was downloaded from Tabule Muris and used to compute a mouse single-cell tau score. Expression values were thresholded at 1 and log-transformed as consistent with the preprocessing applied to the human HPA data.

Protein-coding gene filtering

To restrict analysis to canonical protein-coding genes, we used BioMart mouse gene annotations and filtered for Ensembl canonical transcripts, GENCODE primary annotation, and protein_coding.

Mouse 3’UTR length

Mouse 3’UTR sequences were downloaded from BioMart and parsed from the FASTA-like export format. 3’UTR length was computed as sequence length in base pairs.

Gini coefficient

The Gini coefficient was used as an alternative measure of expression inequality across tissues. For each gene, it was computed from the normalized expression vector as

$$G = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} |x_i - x_j|}{2N \sum_{i=1}^{N} x_i}$$

where $x_i$is the expression level in tissue or cell type i, and $N$is the total number of tissues or cell types. Higher Gini values indicate more uneven, tissue-restricted expression, whereas lower values reflect broader expression.

Publications count per gene – Gene2Pubmed

To estimate how extensively each gene has been studied, we used NCBI's gene2pubmed annotation table, which links Entrez GeneIDs to PubMed IDs. We restricted the table to human entries (tax_id = 9606) and, for each gene, counted the number of unique associated PubMed IDs. GeneIDs were then mapped to our gene symbols (HGNC).

Acknowledgements

I would like to thank Prof. Tzachi Pilpel for welcoming me into the lab, for giving me the opportunity to take part in great and exciting science, and for his guidance along the way.

I would like to thank Ruthie Golomb for being a mentor, a colleague, and most importantly, a friend.

I am grateful to the wonderful scientists of the Pilpel lab, especially Meir Sylman, Noam Hadar, Dvir Dahari, and Orna Dahan.

Finally, I would like to thank my family — my husband, my parents, and my in-laws — for their constant support. And last but not least, my children, who teach me humility and remind me every day what is the meaning of life.

References

1. Dezsö, Z. *et al.* A comprehensive functional analysis of tissue specificity of human gene expression. *BMC Biol* 6, 1–15 (2008).
2. Joshi, C. J., Ke, W., Drangowska-Way, A., O'Rourke, E. J. & Lewis, N. E. What are housekeeping genes? *PLoS Comput Biol* 18, e1010295 (2022).
3. Zhu, J. *et al.* Identification of Tissue-Specific Protein-Coding and Noncoding Transcripts across 14 Human Tissues Using RNA-seq. *Sci Rep* 6, 1–11 (2016).
4. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* 21, 650–659 (2005).
5. Rissanen, J. Modeling by shortest data description. *Automatica* 14, (1978).
6. Roos, T. Minimum Description Length Principle. in *Encyclopedia of Machine Learning and Data Mining* (2016). doi:10.1007/978-1-4899-7502-7_894-1.
7. Downloadable data - The Human Protein Atlas. https://www.proteinatlas.org/about/download.
8. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* 347, (2015).
9. HGNC Database, H. G. N. C. (HGNC), D. of H. L. R. C. C. 0PT, U. K. www. genenames. org. HGNC Database. (2024).
10. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10, 1–7 (2009).
11. Eden, E., Lipson, D., Yogev, S. & Yakhini, Z. Discovering motifs in ranked lists of DNA sequences. *PLoS Comput Biol* 3, 0508–0522 (2007).
12. Fuchs, S. B. A. *et al.* GeneAnalytics: An Integrative Gene Set Analysis Tool for Next Generation Sequencing, RNAseq and Microarray Data. *OMICS* 20, 139–151 (2016).
13. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes are compact. *Trends in Genetics* 19, 362–365 (2003).
14. Smedley, D. *et al.* The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res* 43, W589–W598 (2015).
15. BioMart. https://www.ensembl.org/info/data/biomart/index.html.
16. Fishilevich, S. *et al.* GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)* 2017, (2017).
17. Perez, G. *et al.* The UCSC Genome Browser database: 2025 update. *Nucleic Acids Res* 53, D1243–D1249 (2025).
18. James Kent, W. *et al.* The Human Genome Browser at UCSC. *Genome Res* 12, 996 (2002).
19. UCSC Genome Browser: Annotation Database. https://genome.ucsc.edu/goldenpath/gbdDescriptionsOld.html.
20. Gardiner-Garden, M. & Frommer, M. CpG Islands in vertebrate genomes. *J Mol Biol* 196, 261–282 (1987).
21. Saxonov, S., Berg, P. & Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A* 103, 1412–1417 (2006).
22. Szklarczyk, D. *et al.* The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res* 51, D638–D646 (2023).
23. Skoufos, G. *et al.* TarBase-v9.0 extends experimentally supported miRNA–gene interactions to cell-types and virally encoded miRNAs. *Nucleic Acids Res* 52, D304–D310 (2024).

24. Homology: Orthologs and Paralogs. https://www.nlm.nih.gov/ncbi/workshops/2023-08_BLAST_evol/ortho_para.html.
25. Rost, B. Twilight zone of protein sequence alignments. *Protein Eng* 12, 85–94 (1999).
26. Litman, T. & Stein, W. D. Obtaining estimates for the ages of all the protein-coding genes and most of the ontology-identified noncoding genes of the human genome, assigned to 19 phylostrata. *Semin Oncol* 46, 3–9 (2019).
27. Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* 42, 293–304 (2024).
28. Pullen, T. J. *et al.* Identification of genes selectively disallowed in the pancreatic islet. *Islets* 2, 89–95 (2010).
29. Deng, X. *et al.* Evidence for compensatory upregulation of expressed X-linked genes in mammals, Caenorhabditis elegans and Drosophila melanogaster. *Nature Genetics 2011 43:12* 43, 1179–1185 (2011).
30. Hurst, L. D. *et al.* The Constrained Maximal Expression Level Owing to Haploidy Shapes Gene Content on the Mammalian X Chromosome. *PLoS Biol* 13, e1002315 (2015).
31. Lercher, M. J., Urrutia, A. O. & Hurst, L. D. Evidence That the Human X Chromosome Is Enriched for Male-Specific but not Female-Specific Genes. *Mol Biol Evol* 20, 1113–1116 (2003).
32. Kryuchkova-Mostacci, N. & Robinson-Rechavi, M. A benchmark of gene expression tissue-specificity metrics. *Brief Bioinform* 18, 205–214 (2017).
33. Gini coefficient - Wikipedia. https://en.wikipedia.org/wiki/Gini_coefficient.
34. PubMed. https://pubmed.ncbi.nlm.nih.gov/.
35. Index of /gene/DATA. https://ftp.ncbi.nlm.nih.gov/gene/DATA/.