Isoswitching drives the aging process in human brains

Beril Erdogdu[1,2], Hyun Joo Ji[1,3], Zoe C. Rudnick[1,2], Mihaela Pertea[1,2,3,5], Steven L. Salzberg[1,2,3,4,5]

[1]Center for Computational Biology, Johns Hopkins University, Baltimore, MD, USA
[2]Department of Biomedical Engineering, Johns Hopkins School of Medicine and Whiting School of Engineering, Baltimore, MD, USA
[3]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA
[4]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[5]Department of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD, USA

Abstract

Learning, reasoning, and working memory functions are attributed to the dorsolateral prefrontal cortex (DLPFC), a distinctive region of the human brain that is highly evolved in primates and exhibits notable variability among individuals. Environmental and genetic factors likely contribute to this variability, but little is known about how they influence changes within an individual brain across the lifespan as different cognitive tasks and challenges arise. Most genetic studies focus on DNA mutations or changes in overall gene expression levels. However, genes can also alter the form in which they are expressed through alternative splicing. Using RNA sequencing data from prenatal and postnatal human DLPFCs, we observed that many genes undergo dramatic shifts in their isoform preferences around the time of birth. We further found that thousands of genes continue to undergo gradual, temporally regulated changes in their preferred isoforms, a phenomenon we term 'isoswitching'. In this study, we present isoswitching as a major force in brain development, capable of accurately predicting human brain age from prenatal stages through late adulthood and beyond eighty years of age. This represents the first demonstration of brain age prediction based solely on RNA sequencing data. We also report isoswitching in the brain of a closely related primate, the rhesus macaque.

Introduction

From embryonic neurogenesis to senescence, the human brain evolves continuously through a precise orchestration of multiple cellular processes such as proliferation, axonal growth and pathfinding, selective cell death, and synaptic modulation[1,2]. Although these processes unfold in great complexity across both spatial and temporal dimensions, they are tightly regulated by genetic mechanisms that support the development and maintenance of a healthy brain[3,4]. These genetic mechanisms control processes as diverse as a child's initial acquisition of language and the gradual decline of memory with advancing age, despite all cells within an individual retaining essentially the same DNA sequence throughout life. The regional and temporal specificity of brain function is largely achieved through variations in gene expression, involving both transcriptional and post-transcriptional regulatory mechanisms of RNA[5].

The generation and utilization of alternative isoforms (transcripts) from the same gene is one mechanism known to drive functional diversity in the human brain across different life stages[6].

Differential transcript usage (DTU) refers to the change in the isoform preference for a given gene, without necessarily changing the total gene expression, and it plays a critical role in determining cell fate, function, and regulation across most eukaryotic tissues. The human brain exhibits a remarkably high rate of alternative splicing compared to other tissues, yet the corresponding DTU mechanisms and their functional consequences are still unknown[6-8]. Two major developmental stages in the human brain are marked by substantial shifts in the transcriptional landscape: first, during prenatal development, the brain transitions from rapid cell division and differentiation to specialized, organ-specific gene expression; and second, during early postnatal life, brain maturation occurs in coordination with environmental inputs[4]. Despite its clear relevance, the relationship between differential transcript usage and the processes underlying brain development and aging remains largely uncharacterized.

In our longitudinal study, we first examine the developmental implications of this phenomenon by comparing prenatal and postnatal human brain samples. By integrating RNA sequencing data from 341 human dorsolateral prefrontal cortex (DLPFC) samples sequenced at the Lieber Institute for Brain Development (LIBD), we present the first large-scale comparison between the prenatal and postnatal human brain transcriptomes, revealing thousands of genes where the dominant isoform shifts entirely between prenatal developmental stages and post-birth. Among these is *SNAP25*, a gene encoding a protein essential for synaptic plasticity and neurotransmitter release, known to produce two major variants, *SNAP25a* and *SNAP25b*[9]. Previous studies in murine organisms have demonstrated that the developmental transition from *SNAP25a* to *SNAP25b* is pivotal for cognitive development and overall fitness for survival[10-14]. Our study provides the first evidence of this isoform transition in the human brain, along with a detailed temporal characterization, between prenatal and early postnatal stages.

The DLPFC is one of the most advanced regions of the human brain, responsible for problem solving, attention, and working memory. It is also the most recently evolved brain region, with an extended maturation period that continues well into early adulthood[15,16]. This prolonged developmental timeline distinguishes humans from other species, highlighting both intriguing similarities and important differences in neurodevelopmental processes. Examining the most significant of the genes implicated in DLPFC development, we found that isoform shifts have a clear correlation with brain aging throughout all stages of human life, extending well beyond prenatal and early postnatal development. Using random forest regression models trained on the isoform usage matrix, we demonstrate for the first time that isoform usage patterns in the human brain consistently predict brain age ($R^2 > 0.90$) in multiple datasets, beginning from prenatal development and continuing through late adulthood, beyond 85 years of age.

The simple and interpretable structure of the random forest models allowed us to identify over 5,000 genes as important contributors to brain age prediction. Further, we augmented the random forest predictor with lasso regressors specific to different age groups in a stacked model approach, enabling us to evaluate which genes are particularly important at each life stage. While we explore the functional relevance of a selected subset of these important genes, including *PALM*, *ARL16*, *RRAGB*, and *AGAP3*, the remaining genes are presented as valuable candidates for future investigation. In support of our findings, we further validated the

predictive power of isoform shifts using an entirely independent dataset from Cardoso-Moreira *et al*. (2019) and extended this validation to a closely related primate, the rhesus macaque. Together, our results show that isoform shifts are a major instrument in guiding the brain's changes—both improvements and decline—throughout life and across the majority of genes. We refer to this dynamic reconfiguration of isoform dominance over time as 'isoswitching'.

Results

The Lieber Institute for Brain Development houses an extensive collection of postmortem brain samples, spanning various life stages from prenatal development through infancy, adolescence, adulthood, and up to 85 years of age. We used RNA-seq data from this collection that was analyzed in earlier studies[17,18] to explore the transcriptomic evolution of the human brain across different developmental stages and aging. Our study includes data from a total of 341 dorsolateral prefrontal cortex (DLPFC) samples, comprising 56 prenatal and 285 postnatal samples with no known history of psychiatric diseases. The dataset includes both male and female individuals, primarily of African American and Caucasian descent (Figure 1.a).

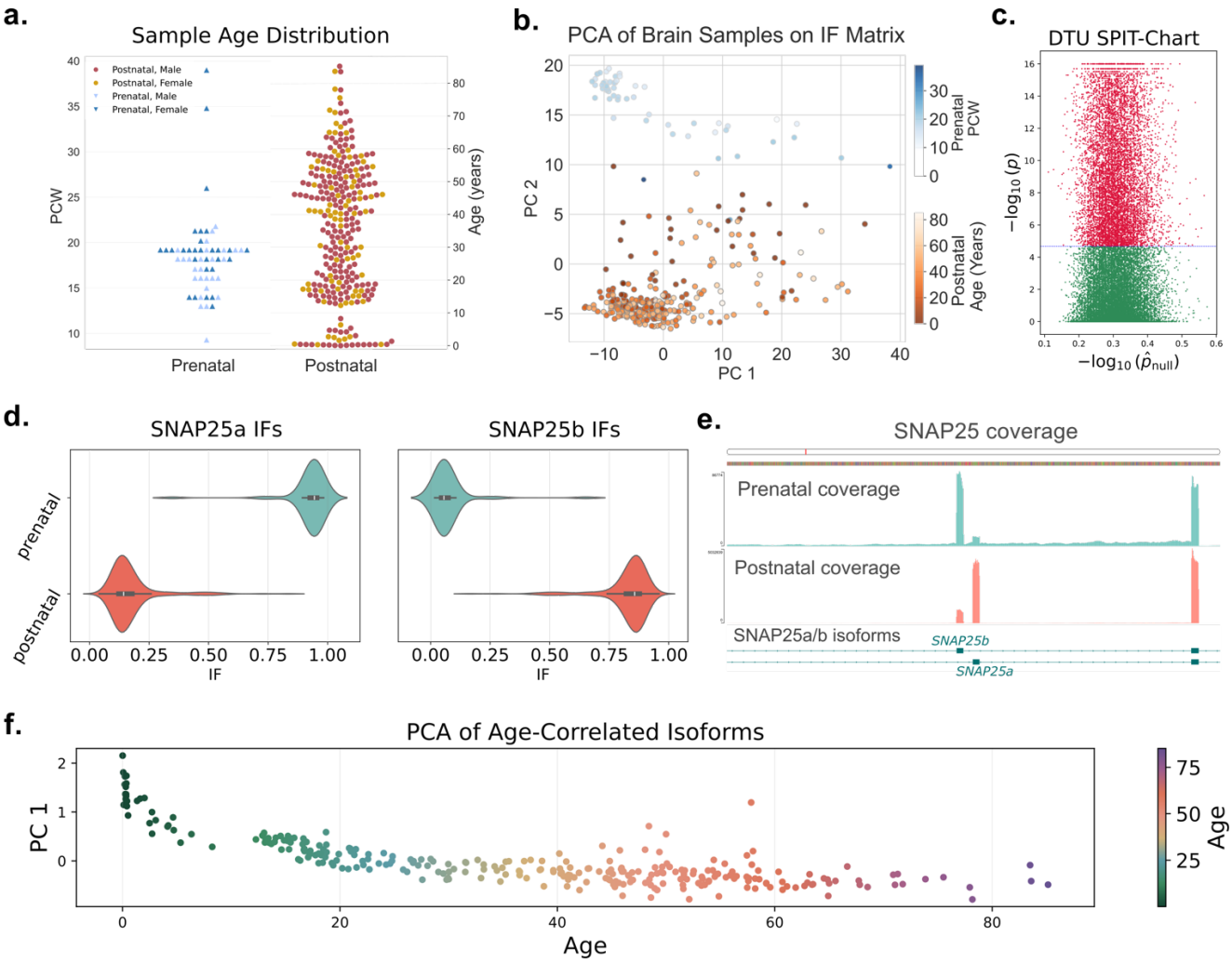

Figure 1.
a. Age distribution of all LIBD samples, separated by developmental group and colored by sex. Postnatal sample ages are shown in years (right y-axis), while prenatal sample ages are shown in post-conception weeks (PCWs, left y-axis). **b.** Principal Component Analysis (PCA) of the isoform fraction (IF) matrix for all LIBD brain samples. PC1 and PC2 are shown. Prenatal samples are colored in shades of blue, with darker shades indicating proximity to birth. Postnatal samples are colored in shades of orange, with darker tones near birth and lighter tones representing increasing age. **c.** SPIT-Chart from the DTU analysis comparing prenatal and postnatal groups. Each transcript is represented as a point. The y-axis shows the –log10(*p*-value) from the comparison, while the x-axis shows the null distribution from comparisons between random subsets of the control group. Significant transcripts are shown in red, and non-significant in green. **d.** Violin plots showing the distribution of isoform fraction (IF) values for the *SNAP25* variants across prenatal and postnatal samples. **e.** Coverage plot of SNAP25 switching dominant exons based on alignments of RNA-seq reads from prenatal (green, top) and postnatal (orange, bottom) samples. **f.** PCA of the IF matrix for DTU transcripts correlated with age. PC1 variation with age is displayed, with samples colored by age in years.

The role of alternative splicing in determining human cortical development has long been a subject of curiosity and investigation[19-22]. Differential transcript usage provides a simplified snapshot of complex alternative splicing mechanisms, capturing the most prominent of changes in isoform expression. As our initial step of analysis, we looked into obvious patterns of isoform dominance within different developmental groups of the LIBD brain samples. Isoform fractions (IF) for transcripts in the CHESS 3 gene catalog[23] were computed for all samples using the quantification of RNA-Seq reads. The IF value represents the ratio of total gene expression that any isoform accounts for in a sample; e.g., if an isoform accounts for 80% of the expression of a particular gene, its IF will be 0.80. To help minimize noise, we refined our analysis to focus on a reliable set of genes and transcripts with sufficient expression evidence (see Isoform Quantification and Filtering section in Methods). A total of 29,757 transcripts from 9,645 genes were compared across samples after filtering. Upon basic dimensionality reduction with Principal Component Analysis (PCA), we observed a striking separation between the prenatal and postnatal groups (Figure 1.b), suggesting systematic isoform shifts around the time of birth. Samples near the birth transition exhibited a weaker variance signal, positioning them closer to the opposite group; i.e., prenatal samples with higher post-conception weeks (PCW) tended to appear closer to early postnatal samples (infants near age 0).

To further investigate the differential transcript usage (DTU) events between the two groups, we conducted DTU analysis on the prenatal and postnatal samples using SPIT[24]. Developed for heterogeneous datasets and complex traits, SPIT is a conservative DTU detection tool that accounts for potential subgroups within cohorts. This ensures that changes affecting only a subset of a heterogeneous group, such as children or adolescents within the postnatal cohort, are not overlooked. To facilitate this comparison, we enabled subgroup analysis within the postnatal samples, given their broad age range and large sample size, which may reflect distinct isoform dominance patterns at different developmental stages.

We controlled for sex and ancestry as potential confounding factors. After excluding genes influenced by these variables, SPIT reported 12,374 transcripts from 5,443 genes involved in significant DTU events between prenatal and postnatal brain samples—an exceptionally large fraction, representing nearly half of all genes analyzed for DTU after filtering (Figure 1.c). Among these, 311 genes were flagged as having highly significant associations with developmental stage. *SNAP25*, a t-SNARE regulator of neurotransmitter release, displayed some of the most striking shifts between the two groups, showing nearly opposite IF distributions for its two isoforms, *SNAP25a* and *SNAP25b* (Figure 1.d).

The protein product of *SNAP25* is one of three proteins involved in the SNARE complex, which joins vesicles to the plasma membrane during exocytosis[25]. In higher eukaryotes, two isoforms of *SNAP25* exist, *SNAP25a* and *SNAP25b*. The associated protein products of the two isoforms are nearly identical, differing in only nine out of 206 amino acids[26]. The isoforms have similar functions; however, *SNAP25b* has been shown to release a larger pool of vesicles, thereby increasing the potential of exocytosis in cells and leading to a more efficient facilitation of synaptic transmission[27]. The transition from the predominant expression of *SNAP25a* to *SNAP25b* from early adulthood onwards in mouse brain development is well-established and

supports our findings[9-11,26,28,29]. In adult mouse brain, the developmentally correct expression of each isoform has been shown to be critical for synaptic maturation and neurotransmission plasticity. Bark *et al*.[10] demonstrated that when targeted mutations are induced to impair the transition between isoforms, affected mice largely die between 3 and 5 weeks of age, the period during which *SNAP25b* expression becomes dominant. While *SNAP25b* has been shown to be the primary isoform expressed in adult human brain[29], no studies have been done to indicate *SNAP25a* as the primarily expressed isoform in prenatal human brain or to investigate the temporal trajectory of the developmental isoform transition. In this work, we present the first comprehensive evidence of this transition.

Coverage patterns in the prenatal and postnatal samples, based on splice-aware RNA-Seq read alignment, highlight the precise exon switch that represents the transition from *SNAP25a* to *SNAP25b* (Figure 1.e).

Along with *SNAP25*, we examined the remaining 310 DTU genes flagged by SPIT to assess whether sample age accounted for some of the heterogeneity observed in the postnatal group. We first selected transcripts whose IF values correlated with aging (indicated by a Spearman correlation threshold of $|\geq 0.5|$) and used their IF values to perform a simple PCA. The clear separation of samples by age along PC1 is shown in Figure 1.f.

Isoswitching predicts brain age

In light of the sequential transition in isoform abundance patterns during development (Figure 1.b) and the time-dependent variation observed in the postnatal samples (Figure 1.f) based on the IF matrix of significant DTU genes, we asked whether these isoform shifts are temporally dynamic—exhibiting a continuous progression throughout life, including all prenatal and postnatal stages, rather than a discrete switch around the time of birth.

We began by analyzing the *SNAP25* transition in greater detail. The coverage plot in Figure 2.a includes prenatal samples (S0) and further subdivides the postnatal group into developmental stages: ages 0–13 (S1), 13–25 (S2), 25–60 (S3), and 60+ (S4). Following the birth transition, *SNAP25a* expression gradually declines, while *SNAP25b* continues to rise, showing a clear trend in S0, S1 and S2 and a more subtle but persistent movement in S3 and S4. Also considering the temporal IF trajectories of variants *SNAP25a* and *SNAP25b* (Figure 2.b), it is evident that their isoform dynamics are continuously shaped by time and show a clear association with sample age across the lifespan. While previous studies in model organisms have demonstrated the existence of this isoform shift, its extended and continuous nature over time has not been shown and may reflect a feature unique to humans. To assess whether this isoswitching pattern was specific to the brain rather than a broad range of human tissues, we examined *SNAP25* expression across tissues present in the Genotype-Tissue Expression (GTEx) collection[30]. The GTEx (V10) project offers RNA-Seq expression estimates from 19,788 samples across 54 tissues for most known human genes, and provides evidence that *SNAP25* is distinctly enriched in the brain (Figure 2.c).

These findings led us to explore whether brain age could be predicted solely from the isoform fractions of a selected set of genes. Remarkably, using the original set of 29,757 transcripts from 9,645 genes prior to any feature selection, we achieved a highly accurate prediction of brain age ($R^2 = 0.856$) in our initial investigation (Figure 2.d). Given the disproportionately large feature space relative to the number of samples, we opted for a simple learning framework and employed a random forest regressor[31-33] with 100 trees to avoid over-fitting. This exploratory model was trained on a randomly selected half of the LIBD brain samples, and tested on the other half. In this step, no more than half of the feature space was considered for each node split, and trees were grown to their maximum depth, producing homogeneous leaves to evaluate learning capacity. These parameters were later adjusted in the final model to facilitate interpolation and support feature selection.

Our primary objective in predicting brain age was to identify genes that undergo isoswitching with stable and well-defined trajectories across development and aging. These genes serve as valuable candidates for markers of brain aging, and the biological significance of their isoswitches provides insight into the mechanisms that regulate changes in the brain over time. We assessed feature significance within the decision forests using the Gini score, which quantifies the total reduction in impurity contributed by each feature across all tree splits, normalized over the forest and the feature space[32,33]. Using this metric, Paralemmin-1 (*PALM*) emerged as the top-ranked gene, contributing the most significant splits in the model. Coverage patterns of *PALM* (Figure 2.e) reveal a gradual emergence of an exon with minimal expression in the prenatal stage, becoming more pronounced in S1 and steadily increasing through later life stages (S2–S4). Examining the IF trajectories of the two prominent *PALM* variants—one containing the emerging exon and the other lacking it—reveals a nearly linear shift with age.

PALM encodes a phosphoprotein located on the cytoplasmic surface of plasma membranes that plays a role in the regulation of membrane dynamics. Previous studies have shown that alternative splicing of *PALM* regulates neurodevelopmental processes such as the maturation of dendritic filopodia into spines, and the trafficking and membrane localization of AMPA-type glutamate receptors (e.g., GluR1) and D3 dopamine receptors[34-36]. It has been demonstrated that alternative splicing involving the inclusion of exon 8 creates enhanced effects in these processes, with deletional mutational analysis indicating that amino acids 154–230 of paralemmin, corresponding to exon 8, show the strongest interactions with full-length D3[36].

Consistent with previous findings, we observed isoswitching patterns supporting the alternative splicing of exon 8, defined with respect to the MANE[37] isoform of *PALM*, where the predominantly expressed isoform transitions as a function of age from *PALM* transcript variant 2 (CHESS 3 ID: CHS.25148.2, RefSeq[38] ID: NM_001040134.2) to *PALM* transcript variant 1 (CHESS 3 ID: CHS.25148.1, RefSeq ID: NM_002579.3) (see Figure 2.f-h). We also examined the structural differences between the two isoforms at the protein level using ColabFold[39], and found that exclusion of exon 8 results in the loss of four adjacent beta sheets, as shown in the protein structure of the prenatal-dominant isoform CHS.25148.2 (Figure 2.g). Given that this region of the PALM protein exhibits the strongest inhibitory interaction with the D3 dopamine receptor, it is plausible that increased inclusion of this domain with age may contribute to a reduction in

dopamine signaling. We present evidence of this age-dependent isoform shift in humans, capturing the full trajectory of isoswitching across all stages of development in the brain.

Extending to a broad genetic landscape with predictive power

To extend our observations beyond individual examples like *PALM* and *SNAP25*, we next investigated how many genes with similarly strong and temporally consistent isoform shifts could be collected using our random forest regressor. Transcripts from these two genes were the top most significant features in the random forest model, respectively. After removing all isoforms of *PALM* and *SNAP25* and retraining the regressor on the same randomly selected half of the LIBD brain samples, the model maintained comparable predictive performance (R^2 = 0.80), with the top-ranking feature now corresponding to an isoform of *RECQL4*.

RECQL4 is a member of the RecQ helicase family and belongs to an enzyme group that is crucial for DNA stability and repair. Mutations in *RECQL4* have been directly associated with features of premature aging and an increased risk of cancer[40-42], and could also play a role in the alternative splicing of the its RNA molecule[43].

We then removed *RECQL4* and applied feature ablation by recursively eliminating all isoforms of the most important gene in each iteration, continuing this process until the model reached an unacceptably low level of performance[44]. Surprisingly, even after removing 5,000 genes, the model continued to perform well above random, retaining a meaningful level of predictive accuracy (Figure 3.a). These findings indicate that isoswitching relevant to brain aging is not confined to a few key genes, but is broadly distributed across the transcriptome, with many genes carrying predictive value.

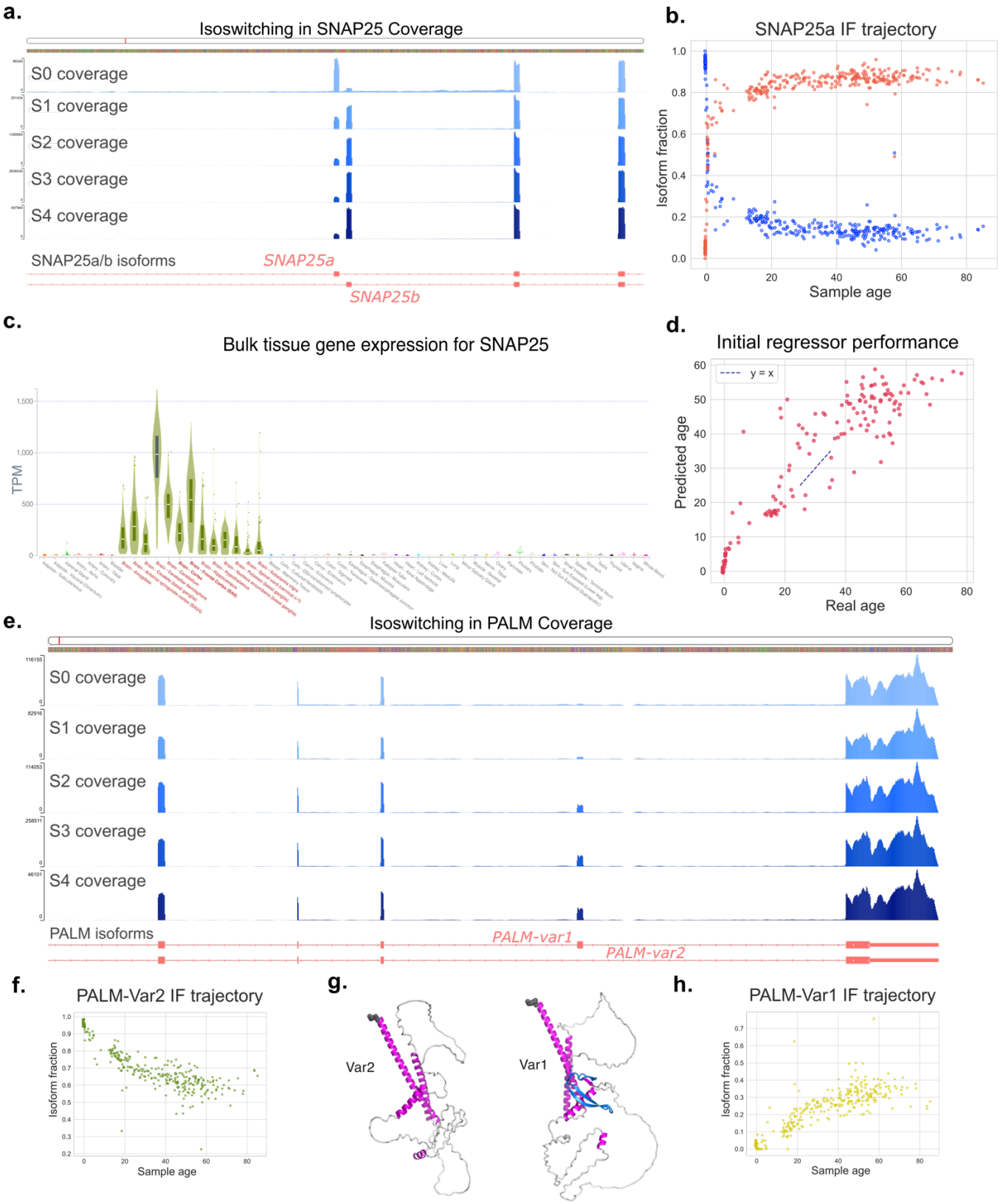

Figure 2.a. Coverage patterns of the shifting exons in *SNAP25*, based on alignments from all brain samples grouped as follows (top to bottom): ages 0–13 (S1), 13–25 (S2), 25–60 (S3), and 60+ (S4). Tracks

are colored by age group, with darker shades representing older ages. **b.** Isoform fraction (IF) trajectories of the *SNAP25a* (red) and *SNAP25b* (blue) variants. Each sample is represented by two points, one per variant. The x-axis shows sample age in years, and the y-axis shows the corresponding IF value. **c.** Expression profile of *SNAP25* across 54 human tissues from the GTEx project, illustrating enrichment in the brain. Transcript per million (TPM) values are shown as violin plots for each tissue. **d.** Actual vs. predicted ages from the initial exploratory random forest model, trained on the IF matrix from a random half of LIBD samples and tested on the remaining half. **e.** Coverage patterns for exon 8 of the *PALM* gene, based on alignments from all brain samples grouped into age groups as in (a). Coloring follows the same convention. **f.** Isoform fraction (IF) trajectory of *PALM* variant 2, which is dominant in prenatal samples. **g.** Protein structures of PALM variants. The additional beta sheets introduced by the merged exon are shown in blue. **h.** Isoform fraction (IF) trajectory of *PALM* variant 1, which is increasingly favored in older samples, shown using the same layout as in (f).

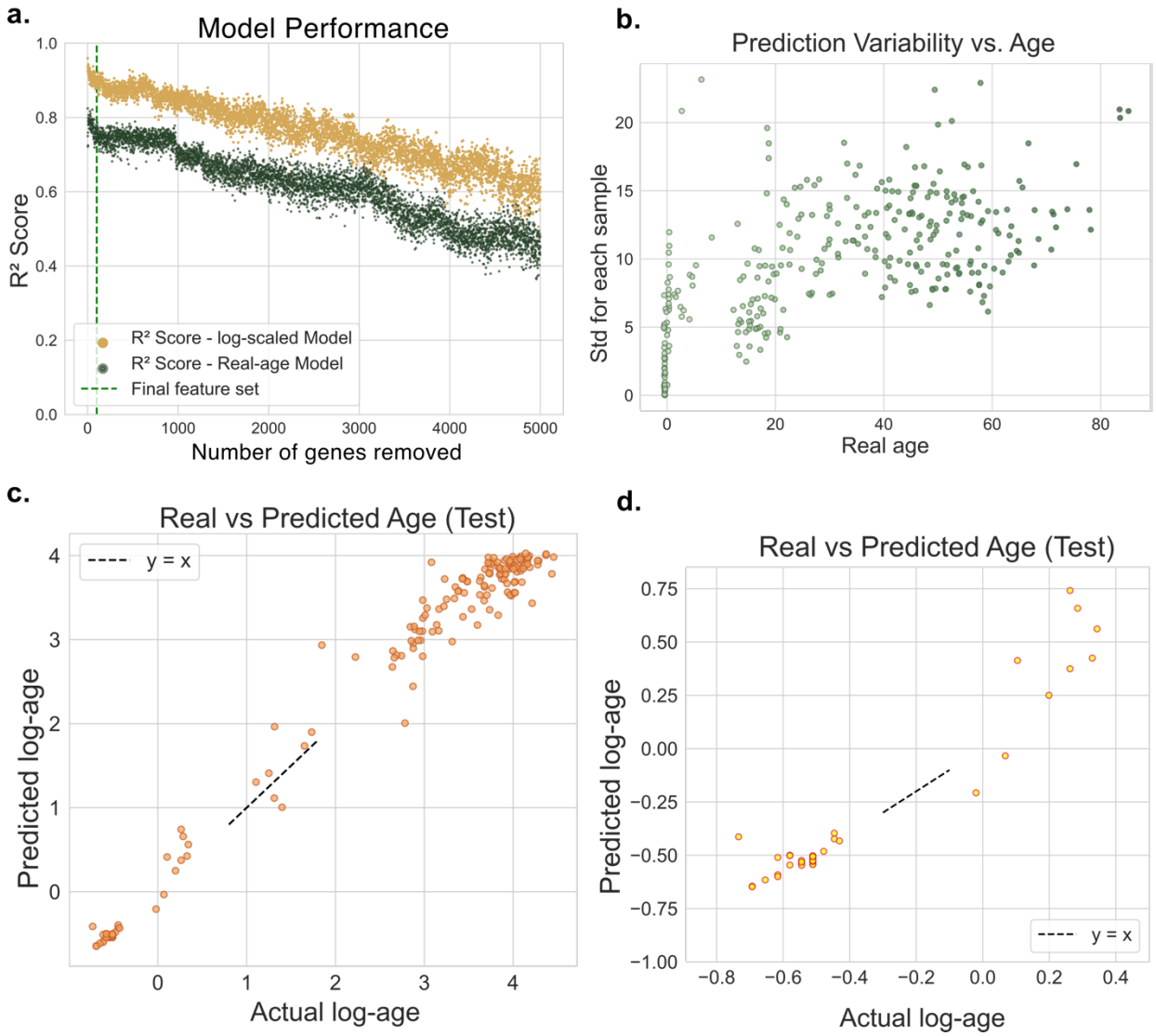

Figure 3.a. R^2 scores throughout the feature ablation process. At each step, the most important gene from the previous model was removed (without replacement), and the model was retrained. The process began with all transcripts remaining after initial filtering. The model trained on log-transformed ages is shown in yellow, and the model trained on raw age values is shown in green. **b.** Standard deviation of age predictions for each sample, based on outputs from the 100 decision trees in the random forest. Each point represents one sample, with sample age on the x-axis. **c.** Performance of the random forest model trained on log-scaled ages. The x-axis shows actual log-age, and the y-axis shows predicted log- age. **d.** Same performance plot as in (c), zoomed in to highlight predictions for prenatal and infant samples.

Scaling of brain age and time

A random forest regressor is an ensemble learning method composed of multiple decision trees, with 100 trees used in our models. Each tree independently generates a prediction for a given sample, and the final model output is calculated by averaging these individual predictions. To assess the model's confidence and stability, we examined the variance among the predictions made by different trees for each sample, providing a measure of uncertainty in the decision-making process. The standard deviation of tree predictions for the LIBD brain samples (Figure 3.b) shows a consistent increase in prediction uncertainty with age. This trend likely reflects two fundamental biological principles.

First, the biological significance of time varies substantially across different stages of life. A single year during prenatal development encompasses a dense sequence of cellular processes, including proliferation, migration, and differentiation. In contrast, a year in late adulthood is marked by slower biological change[45]. The difference between a newborn and a one-year-old is profound, while the difference between a seventy-year-old and a seventy-one-year-old is comparatively subtle and less predictable. Second, while early development tends to follow a structured and genetically guided trajectory, aging is a highly individual process shaped by the interaction of genetic, environmental, and lifestyle factors. Unlike development, where milestones occur in a relatively uniform pattern, aging progresses at different rates in different individuals[46]. Two people of the same chronological age may differ significantly in brain health, cognitive performance, and overall physical condition, reflecting the complexity and variability of aging[47,48].

While we cannot fully account for individual differences in the progression of aging due to limitations in available data and scientific knowledge, we can adjust for the biological significance of time across different age groups. Given our objective of identifying meaningful gene candidates that undergo significant and well-defined isoform shifts predictive of age, it is crucial to recognize that the biological impact of the same time unit is not uniform across the lifespan. An error of one year should not be weighted equally for a 60-year-old individual and an infant, as this would undervalue genes that are crucial for prenatal and early brain development.

An intuitive transformation to address this issue is to predict age on a logarithmic scale rather than using raw age values. This approach assigns greater importance to age differences in early life while gradually reducing their impact in later stages. To ensure all values remain positive, ages were incremented by 1 before applying the natural logarithmic function. The random forest model trained on log-transformed ages (referred to as log-ages hereafter) appropriately assigns greater weight to errors made during earlier stages of life and achieves an exceptional R^2 score of 0.96. After feature selection (see Methods), using only the isoform fraction matrix from the top 100 selected genes, the model achieves an even higher R^2 score of 0.977 (Figure 3.c). Focusing on the prenatal and infant sample group (ages -1 to 1), we observe that isoswitching not only predicts aging with high accuracy in postnatal stages, but also effectively tracks

developmental progression in prenatal and infant samples based on post-conception weeks (Figure 3.d).

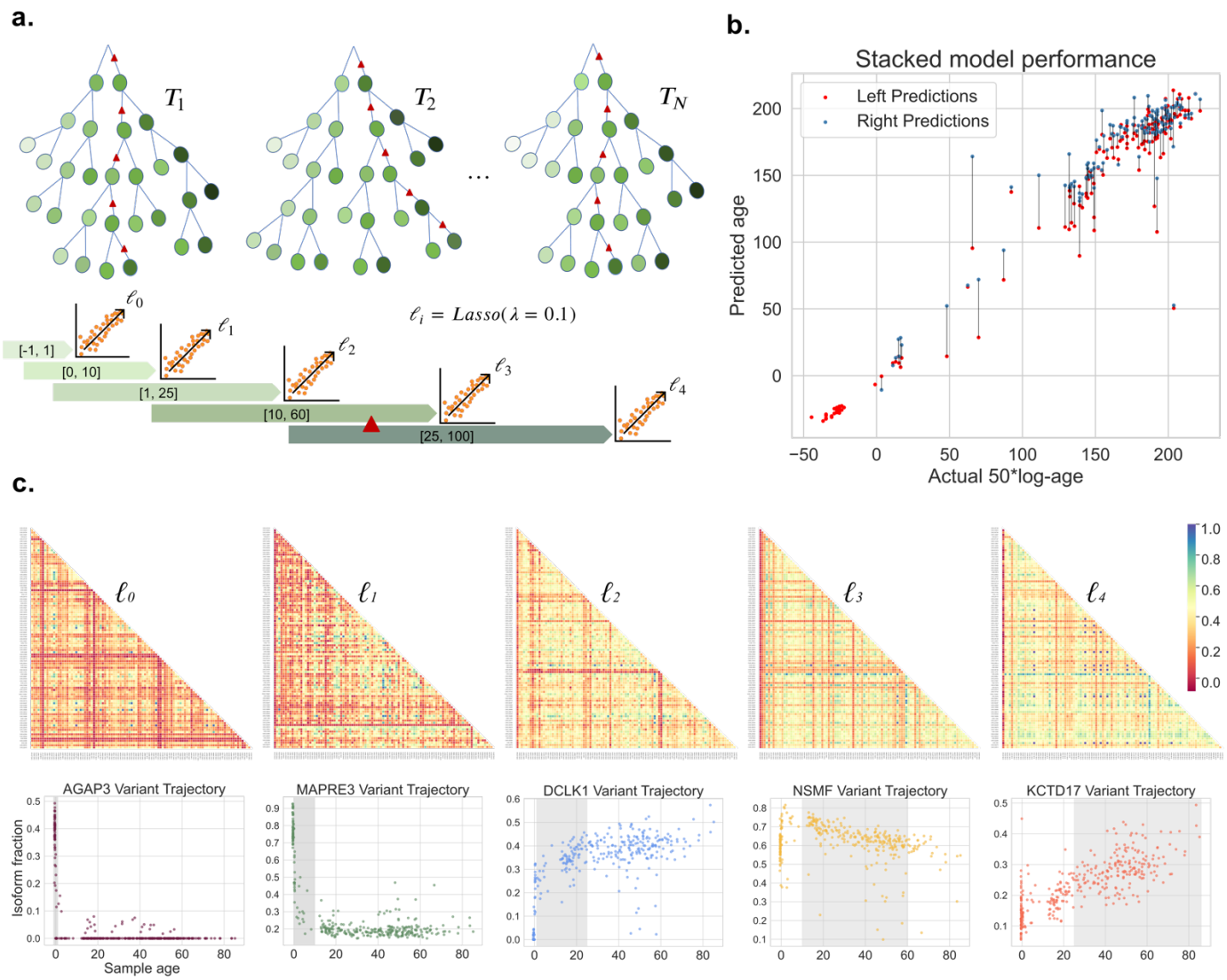

Figure 4.a. Schematic representation of the stacked model structure. An adult sample, represented by a red triangle, is traced through the model pipeline. First, each tree in the random forest predicts the sample's age, and these predictions are averaged to produce a final estimate. Based on this prediction, the sample is routed to one or more of the five lasso regressors, each corresponding to a specific age range: l_0: [–1, 1], l_1: [0, 10], l_2: [1, 25], l_3: [10, 60], and l_4: [25, 100]. In this example, the forest prediction falls within the range of both l_3 and l_4. **b.** Performance plot for the stacked model trained on log-ages. For samples that received predictions from two lasso regressors, the "younger" model's prediction is shown as the left point (red), and the "older" model's prediction is shown as the right point (blue). Points from the same sample are connected by lines. **c.** Heatmap of drops in accuracy (Spearman correlation) for lasso regressors when trained on every possible pair of genes from the top 100. Each cell represents the accuracy drop when only that gene pair is used. Lower drops (red) indicate stronger predictive power. For each lasso model, a gene was selected from the red tracks of the heatmap to display an example isoswitching trajectory beneath the corresponding panel. The age range associated with the lasso model is highlighted in gray in the trajectory plots.

Stacked model

Using our predictive model for estimating brain age across prenatal development, postnatal maturation, and aging, we identified a robust set of 100 genes that consistently contribute to its predictive accuracy. However, it is unlikely that all 100 genes contribute equally throughout the entire course of brain development and aging. Instead, distinct subsets are likely to be functionally relevant at different periods. For instance, a gene essential for neurogenesis and early cortical formation is unlikely to play a central role in the molecular processes associated with later-life decline, though some continuity between developmental and aging mechanisms may exist.

Thus, a crucial next step is to determine how and why each gene contributes to age prediction and to identify the specific life stages where its impact is most pronounced. By dissecting the temporal influence of these genes, we can gain deeper insights into their functional relevance across different phases of human brain maturation.

To better characterize the temporal dynamics of gene importance, we added an additional predictive layer on top of the random forest regressor. In this stacked modeling framework, the random forest first estimates the age of each sample. The predicted age is then used to assign the sample to one of five lasso regression models, each designed for a specific range of ages, designed to be overlapping: [-1, 1], [0, 10], [1, 25], [10, 60], and [25, 100]. It is important to note that all steps in both regressor layers use log-ages as input and output. However, corresponding real ages are used here for clearer presentation of the model structure.

Let $T = \{s_1, s_2, \dots s_{N/2}\}$ be the set of all training samples where N is the total number of brain samples. Each sample s_i has an isoform fraction (IF) vector of $X_i = [f_{t_1} \quad f_{t_2} \quad \cdots \quad f_{t_p}]$ with p elements where f_{t_j} represents the IF value from a selected transcript t_j. After feature selection, in each vector X_i there remained a total number of $p = 341$ transcripts from the 100 selected genes. For a training sample s_i, the random forest regressor predicts an age based on its X_i, which we will call $s_i{}'$, and this s_i' will fall within the boundaries of either 1 or 2 of the overlapping lasso regressor windows (Figure 4.a):

$$l_0 = [-1, 1], l_1 = [0, 10], l_2 = [1, 25], l_3 = [10, 60], l_4 = [25, 100]$$

Each of these lasso models is trained exclusively on the subset of samples corresponding to its respective age range, as determined by the random forest predictions. For example, l_0 will be trained on $\overline{X_0}$ where $\overline{X_0} \in \{X_i \ \forall \ i \ \text{ s.t. } \ -1 \leq s_i' \leq 1 \ \}$. The random forest regressor was constrained to have a minimum of 10 samples per leaf prior to the second regression step. This parameter was retained in all subsequent steps, as it proved effective in improving model generalization.

This framework serves two complementary purposes. First, it allows us to segment human brain development and aging into biologically meaningful time periods that may correspond to

distinct molecular transitions. Second, by using overlapping age ranges, the model provides a bridge between adjacent life stages, capturing smooth transitions rather than treating developmental phases as discrete, isolated units. Unlike prenatal samples—whose isoform profiles are strikingly distinct from postnatal ones—and the most elderly individuals, many samples fall within the overlapping range of two lasso regressors, allowing for interpolation or extrapolation from both models.

For instance, a sample with a predicted age of 20 years from the random forest is processed by both the [1, 25] and [10, 60] lasso models, receiving predictions from each. This overlap enables direct comparison between models trained on different temporal windows, allowing for a comparison of predictions depending on the chosen reference point in the temporal axis.

We selected lasso regression for this second layer due to its inherent ability to perform feature selection while strictly penalizing overfitting. Unlike standard regression models, lasso applies L1 regularization, which forces some regression coefficients to shrink to zero, effectively removing less relevant features from the model:

$$\min_{\beta} \left(\sum_{i=1}^{n} 50(y_i) - X_i\beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

where y_i is the log-age of the subject in sample s_i, X_i is the IF vector of sample s_i, p is the number of features, β is the vector of regression coefficients, and λ is the regularization parameter, which we set to 0.1. Log-ages were multiplied by an arbitrary factor of 50 to amplify age differences for clearer separation across regressors.

To evaluate the effectiveness of this framework, we applied it to the test set, where each sample was first processed by the random forest regressor and then routed to the appropriate lasso model for final age prediction. For samples that fell into overlapping age intervals and received predictions from two lasso models, the final estimate was calculated as the average of the two. As expected, for such samples the lasso model trained on the younger age range almost always produced lower age estimates than the model trained on the older range (Figure 4.b). Importantly, no unexpected discrepancies were observed between adjacent lasso models, indicating strong consistency in predictions across overlapping age windows. The overall coefficient of determination (R^2), based on the averaged predictions for all test samples, was 0.948 with log-ages and 0.872 when using raw age values.

The stacked modeling design not only provides separate regressors tailored to distinct age ranges, but also enables independent evaluation of each regressor's predictive accuracy. This structure allows us to move beyond a single global measure of age prediction performance and instead assess how individual genes contribute to prediction within specific developmental or aging intervals.

To identify the most influential genes, we conducted an exhaustive pairwise evaluation. From the set of 100 top-ranked genes, we considered all 4,950 possible gene pairs $\{g_u, g_v\}$ where $u \neq v$. For each gene pair, we trained and tested the lasso layer using only the isoform fraction vectors corresponding to transcripts from those two genes. Regularization parameter λ was set to 0.05 to account for the reduced feature space. The resulting Spearman correlation score was compared to that of the original model trained on all 100 genes. Gene pairs that produced minimal loss in correlation were interpreted as being highly informative on their own, suggesting they possess strong and independent predictive value for the corresponding range of brain age. Using this method, we identified genes that, when included in a randomly selected pair, preserved model accuracy nearly as well as the full 100-gene model. These genes represent the most robust contributors to age prediction across distinct developmental periods, appearing as prominent red tracks in the heatmaps shown in Figure 4.c. As the age categories widen and shift toward older groups across the lasso models from l_0 to l_4, prediction ambiguity increases, resulting in a gradual decline in the number of such stand-alone predictor genes.

From the red tracks in the heatmaps, we manually selected a subset of stand-alone predictor genes and inspected their IF trajectories across all LIBD samples. One representative example from each heatmap is shown below its corresponding panel in Figure 4.c. We observed a clear distinction in isoswitching behavior across developmental periods. In the lasso regressor trained on samples from the youngest subjects (prenatal to age 1), the most predictive genes exhibited a sharp, binary-like shift occurring precisely around the time of birth . In contrast, genes that contributed most to age prediction in later stages—such as those in the [10, 60] age range—showed gradual and continuous changes in isoform fractions over time. This difference suggests that some genes undergo abrupt regulatory transitions at birth, while others undergo gradual but sustained isoswitching that unfolds throughout life.

Testing with independent data

Having established a model with strong predictive performance across the human lifespan, we next examined whether the observed relationship between isoform shifts and brain aging is robust and reproducible. To answer this question, we selected an independently published dataset of postmortem brain samples spanning a broad age range. We aimed to evaluate whether the model could accurately predict ages in an external cohort, and to determine whether the 100 key genes identified in our original analysis exhibited similar isoform fraction trends in a different population.

For this external validation, we selected the dataset published by Cardoso-Moreira *et al*. (2019)[4], which provided an ideal test case for assessing the robustness of our findings in an independent cohort. Cardoso-Moreira *et al*. conducted a study on organ development across seven species using bulk RNA-Seq data collected at multiple time points, spanning both prenatal and postnatal life stages. Their study examined gene expression changes in seven organs (forebrain, hindbrain, heart, kidney, liver, ovary, and testis) and explored the relationship between evolution and development. They found that gene expression patterns in organs tend

to be more conserved across species during early development, with increasing divergence at later stages of life.

While the dataset includes only 53 human forebrain samples, with 32 being prenatal, it provides a valuable resource for studying developmental trajectories. Postmortem human brain data spanning a wide range of ages is exceptionally scarce, making this dataset particularly useful for validating our findings. We applied our random forest regressor—trained on the LIBD dataset and restricted to the top 100 predictive genes—to this external set of 53 samples, following the same quantification, normalization, and filtering procedures used for the LIBD RNA-Seq data. Our model predicted the log-ages of the samples with an R^2 score of 0.924, comparable to its performance on the LIBD samples ($R^2 = 0.948$). Similarly, applying the random forest regressor trained on the LIBD samples using real ages and the selected set of 100 genes resulted in strong performance, with an R^2 score of 0.840 (Figure 5.a), compared to 0.862 on the original LIBD test set.

We observed that the correlation between predicted and actual ages was noticeably lower for prenatal samples in this dataset. We attribute this to differences in the developmental timeframes represented in the two datasets. The LIBD dataset, on which our model was trained, captures a narrow window within prenatal development, whereas the Cardoso-Moreira *et al.* (2019) dataset spans both earlier and later stages of pregnancy. As a result, our model had less data on the range of isoform fraction patterns and their changes in the earliest stages of development.

Brain-Specificity of Isoform Shifts

After identifying a strong association between isoform shifts and brain aging, we next asked whether a comparable pattern could be detected in a different human organ. For this comparison, we selected the heart, which serves as a conceptual contrast to the brain. While the brain undergoes continuous structural and functional remodeling across the lifespan to meet evolving cognitive and physiological demands, the heart is required to maintain a fundamentally steady task and efficiency throughout life in order to support vital circulatory functions. Although both organs experience changes over time, a healthy brain retains lifelong neuroplasticity, whereas alterations in heart structure and gene expression are more often associated with aging-related cardiac pathologies[49-52].

For this analysis, we used RNA-seq data from 261 left-ventricle heart samples collected by the GTEx project[30], with ages ranging from 20 to 70. Although this dataset does not include prenatal or early postnatal samples, its large sample size provides a robust foundation for modeling the aging process and assessing whether the isoform dynamics observed in the brain are also present in the heart. Using a set of stably expressed genes across all samples (see Methods, Isoform Quantification and Filtering), we trained a separate model based on isoform fractions to evaluate age-related isoform shifts in the heart. The model performed poorly, showing little to no noticeable trend between predicted and actual ages, with an R^2 score of only 0.079 (Figure 5.b).

Figure 5.a. Performance of the random forest regressor trained on the LIBD training set and tested on an independent dataset from Cardoso-Moreira et al. (2019). Left: Model trained and applied on log-transformed ages. Right: Same model architecture trained and applied on real age values. **b.**

Performance plot of a random forest regressor trained on a randomly-chosen half of 261 GTEx heart left ventricle samples (ages 20–70) and tested on the remaining half. **c.** Coverage patterns for the isoswitching exon 8 in *PALM*, present in *PALM* var1 but not var2, based on RNA-seq alignments from 32 rhesus macaque brain samples (Cardoso-Moreira et al., 2019). Samples are grouped as follows (top to bottom): MS0 (prenatal), MS1 (ages 0–4), MS2 (ages 4–8), MS3 (ages 8–20), and MS4 (ages 20+). Tracks are colored by group, darkening with age. **e.** Same as (d), showing coverage of the isoform-switching region in *SNAP25*, which is conserved across human and macaque.

Validation in the Rhesus macaque brain

We next investigated whether similar patterns of isoswitching as a function of age could be observed in one of our closest relatives, the rhesus macaque. To explore this, we utilized the forebrain samples from the Cardoso-Moreira *et al*. (2019) study, which included 32 macaque brain samples, nine of which were prenatal. The ages of these samples ranged from 93 post-conception days to over 20 years.

Unlike the well-annotated human transcriptome, the macaque transcriptome is considerably less well-characterized, with fewer documented alternative isoforms per gene. As a result, direct quantification using Salmon failed to produce meaningful results due to the limitations of the existing annotation[53]. To address this, we constructed a more comprehensive transcript annotation by integrating gene models from GENCODE and NCBI's macaque annotations (see the Comprehensive Macaque Annotation section in Methods), allowing for a broader representation of isoform diversity. Using this improved annotation, we realigned the RNA-seq reads to the genome (see Alignment section in Methods) and focused our analysis on two of the strongest predictive genes in humans, *PALM* and *SNAP25*, to assess whether they exhibit similar isoswitching trajectories in macaques.

We partitioned the macaque samples into developmental stages following the same approach used for humans. Prenatal samples were labeled as MS0, while postnatal samples were categorized into MS1 (ages 0–4), MS2 (ages 4–8), MS3 (ages 8–20), and MS4 (ages 20+) based on currently proposed developmental milestones[54].

As shown in Figure 5.c, the same exon identified in *PALM* in humans—exon 8—exhibits minimal coverage during the prenatal stage in macaques, gradually emerging and increasing in expression over time. This exon is identical in length and results in the same protein-level amino acid changes in both species. A parallel pattern is observed for *SNAP25*, where the same exon shift leads to a nine–amino acid difference in the translated protein, consistent with our observations in humans (Figure 5.d). These findings align with previous work by Bark *et al*.[10,13], who documented the same *SNAP25* isoform variants in more distantly related species, including mice and chickens.

These coverage patterns provide strong evidence that the isoswitches identified in the human brain are also present in macaques. However, a larger dataset and a more comprehensive transcript annotation will be necessary for a genome-wide analysis of isoform dynamics during macaque forebrain development.

Functional implications of isoswitches with high model importance

To explore the functional consequences of transcript-level regulation, we analyzed protein products of age-associated isoform pairs to identify protein domains enriched at switch sites. We examined 1,000 isoform pairs whose IF trajectories were among the most predictive features for brain age. These pairs represent 1,000 high-importance genes that exhibit

dominant isoswitches across the age spectrum. This expanded gene set was obtained by increasing the threshold in our feature ablation procedure from 100 to 1,000 genes (see Feature Selection section in Methods). In each iteration, we first selected the most predictive isoform, as defined in Methods. We then identified a second isoform from the same gene whose isoform fraction trajectory had the strongest negative Spearman correlation with the first, reflecting the typical opposing patterns seen in isoswitching.

For 968 protein-coding genes in the list of 1,000 high-importance genes, we compared the protein sequences encoded by each isoform pair. We confirmed that 787 / 968 isoform pairs showed divergent protein sequences, ranging from single amino acid differences to large gaps spanning more than 100 amino acids (see Table 1). Among the various ways isoforms can diverge in their protein products, we focused on pairs exhibiting a single gap, which represents the inclusion or exclusion of a continuous sequence of coding exons, likely indicating a change in a specific protein domain.

We performed pairwise global (Needleman-Wunsch) alignment of the protein sequences encoded by the 787 isoform pairs identified earlier[55]. We identified 344 pairs exhibiting a single gap, with the gap length corresponding to the length difference between the two proteins. For a large majority (≈84.6%) of instances, we observed gaps of at least 10 amino acids, with some more than 50 amino acids (see Table 1).

Table 1. Difference in protein sequence lengths in proteins encoded by pairs of switching isoforms.

Length difference (aa)	Number of pairs	Proportion (%)
< 10	53	15.4
10–49	151	43.9
50–99	52	15.1
>= 100	88	25.6

To annotate the protein domains affected by these gaps–hereinafter called the isoform-specific sequences–we used Foldseek[56] for fast structural alignments against large protein databases. Structure-based alignments improve search sensitivity, as protein domains can be conserved structurally despite low sequence-level conservation. We searched ColabFold-predicted structures of the longer isoforms against the CATH50 database[39,57,58] (see Methods). The alignments were filtered to include only those with at least 90% coverage of the isoform-specific sequence and an E-value of ≤ 0.01.

We identified 134 out of 344 isoform-specific sequences that contained a CATH domain. Notably, 96 of these 134 domains shared protein structure superfamilies (see Methods). Seven superfamilies were repeated at least five times, with examples including the P-loop containing nucleotide triphosphate hydrolases superfamily and the leucine-rich repeat variant (LRV) superfamily, which appeared 9 and 10 times, respectively. Given that CATH (v4.3.0) contains over 500,000 domains and 6,000 superfamilies, the recurrence of the same domain or superfamily is unlikely to occur by chance.

Next, we highlight two genes with isoform changes that affect a P-loop domain, illustrating how transcript-level shifts can have biological consequences. While LRV domains appeared frequently in our search results, they remain less well studied in the context of brain function.

ADP ribosylation factor-like GTPase 16 (*ARL16*) is a protein-coding gene that encodes a GTPase responsible for hydrolyzing GTP into GDP. In ARL16, we observed an inverse relationship between isoforms CHS.23550.1 and CHS.23550.15 in our earlier predictive model, with CHS.23550.1 decreasing and CHS.23550.15 increasing with age. The first 80 amino acid residues of CHS.23550.1 are missing from CHS.23550.15, and this missing region contains two P-loop domains, commonly found in ATP- and GTP-binding proteins[59] (Figure 6.a-b). The P-loop is responsible for phosphate binding and is essential to ARL16's function as a GTPase. Structurally, P-loops appear as loops connecting a beta sheet to an alpha helix, and are often followed by another beta sheet (e.g., β – [P-loop] – α – β)[60,61]. Although the P-loop exhibits low sequence conservation, its structure is highly conserved[62], enabling reliable annotation of the structure using Foldseek.

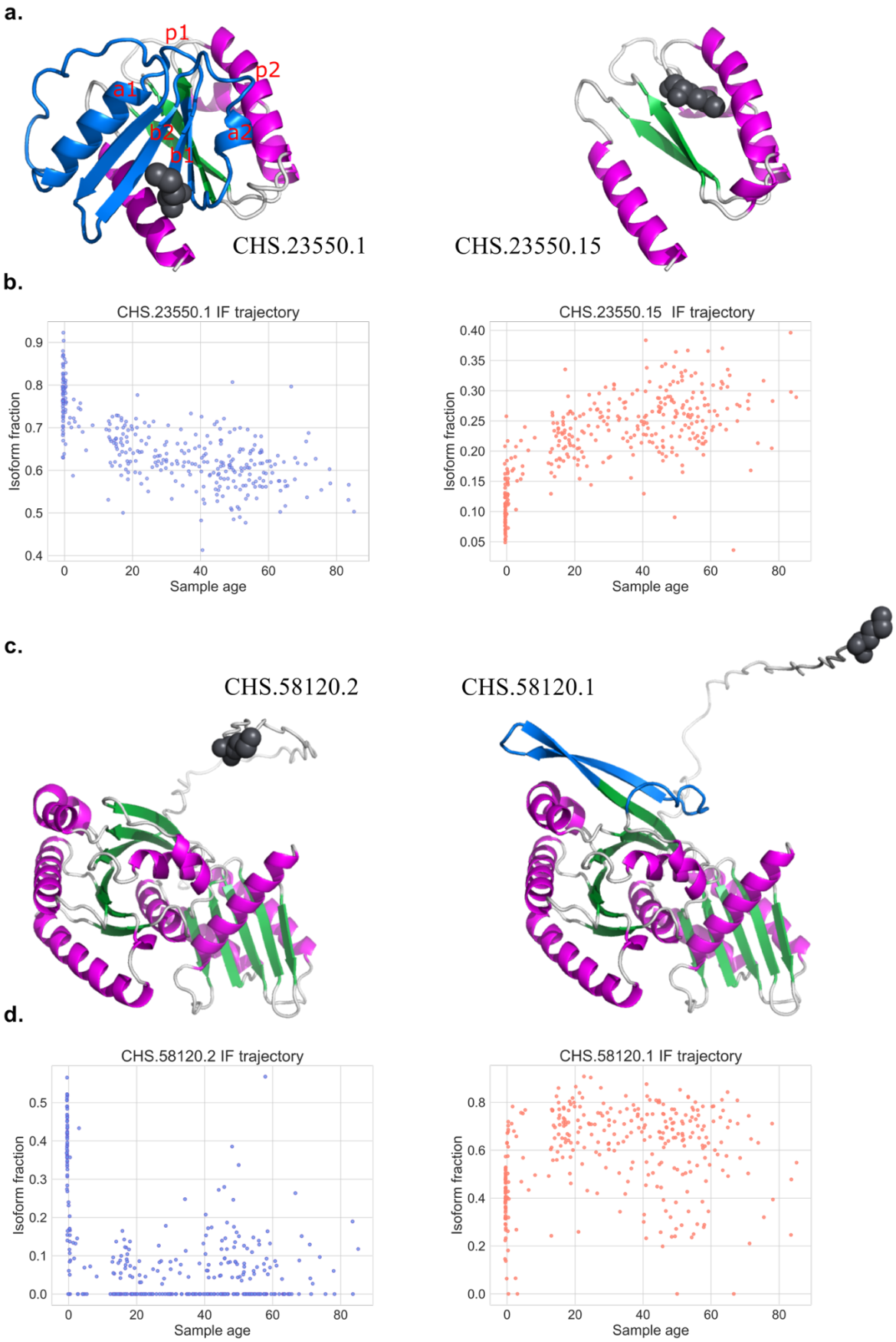
a.
p1
p2
a1
b2
a2
b1
CHS.23550.1
CHS.23550.15
b.
CHS.23550.1 IF trajectory
Isoform fraction
Sample age
CHS.23550.15 IF trajectory
Isoform fraction
Sample age
c.
CHS.58120.2
CHS.58120.1
d.
CHS.58120.2 IF trajectory
Isoform fraction
Sample age
CHS.58120.1 IF trajectory
Isoform fraction
Sample age

Figure 6. Aligned protein structures of isoform pairs for ARL16 and RRAGB. In each diagram, light grey, magenta, green, and blue structures are the loop, alpha helix, beta sheet, and isoform-specific regions respectively. Dark grey spheres indicate the N-terminus. **a.** Isoform CHS.23550.1 of ARL16 contains two additional P-loop domains (blue) compared to CHS.23550.15. The P-loops are manually labeled (p1, p2) using red text. **b.** Corresponding IF trajectories for CHS.23550.1 (blue) and CHS.23550.15 (orange). **c.** Isoform CHS.58120.1 of RRAGB, shown on the right, contains an elongation as well as an addition to a beta sheet (colored blue), compared to CHS.58120.2. **d.** Corresponding IF trajectories for CHS.58120.2 (blue) and CHS.58120.1 (orange).

The gene *RRAGB* (Ras-related GTP-binding B) is another example in which the dominant isoform undergoes changes in the P-loop domain. We confirmed an isoform shift from CHS.58120.2 to CHS.58120.1 for *RRAGB* and identified the 28aa residues missing from CHS.58120.2 as a P-loop domain (Figure 6.c,d). Like *ARL16*, *RRAGB* also encodes a GTPase, implying that changes in this domain may also affect GTP binding activity.

Previously, the shorter isoform of *RRAGB* was considered the dominant form expressed in most tissues, with the longer isoform expressed primarily in the brain[63]. While the expression of both short and long isoforms of *RRAGB* has been reported in the brain[64], the interaction between their IF levels has not been explored. Here we provide a novel finding that the longer isoform (CHS.58120.1) appears to be the dominant variant in the postnatal brain.

RRAGB is part of the Ragulator-Rag complex, which binds to and recruits mTORC1 to the lysosome, where cellular materials get broken down into nutrients[65]. mTORC1 promotes cellular anabolism, affecting processes like synaptic transmission and myelination[66,67]. *RRAGB* regulates mTORC1 in response to nutrient availability, ensuring mTORC1 persistence during nutrient deprivation to maintain vital organ function during starvation. Figlia et al. showed that each of *RRAGB*'s isoforms confers mTORC1 persistence through distinct mechanisms[64]. The functional difference between these two isoforms is attributed to the 28-amino acid addition in the longer isoform, which appears to impede GTP binding. Together with our DTU results, this suggests that the mechanism of action for RRAGB changes between fetal development and the postnatal brain.

Methods

Pairwise global alignment of protein sequences

We performed pairwise global alignment of two protein sequences using the pairwise-sequence-alignment Python module at https://github.com/aziele/pairwise-sequence-alignment. The module implements the Needleman-Wunsch algorithm for global sequence alignment and is part of the EMBOSS package[55].

For each alignment result, we checked whether the longest gap matched the length difference between the two sequences, excluding cases where the original sequences were of identical length. For all such cases, we saved the isoform-specific sequence and noted which of the two sequences was longer.

Structure-based domain searches

To search for protein domains intersecting our list of isoform-specific sequences, we used Foldseek[68] to align the predicted protein structures of the longer isoforms against the CATH50 database, which we obtained using the `foldseek databases` command. CATH50 comprises representative domain structure from CATH[58] (v4.3) and AlphaFold DB[69], clustered at 50% sequence identity. Most of the structures were downloaded from isoform.io, a public online database hosting ColabFold-predicted protein structures for 237,275 human isoforms[70]. We folded 36 isoforms that were omitted from the database due to their length (>2,000aa), using ColabFold on a Nvidia A100 GPU[39,57]. CHS.53345.5 was the only isoform that could not be folded due to its excessive length (>4,000aa).

We filtered the Foldseek results based on their E-values and alignment coordinates, requiring an E-value no larger than 0.01 and a minimum query coverage of 90%. Query coverage was calculated as the proportion of the isoform-specific sequence length that aligned. These filters ensured that only statistically significant domain hits with meaningful overlap in the gap region were retained.

We compiled two versions of the filtered Foldseek results: one including all domain hits and another containing only the top hit for each query. The latter was used to assess the enrichment of specific CATH superfamilies. The filtered output includes alignment coordinates, CATH domain ID, CATH superfamily ID, query coverage, E-value, and bit score for each hit. When multiple hits were identified for a single query, the domain with the lowest E-value was selected, with ties resolved arbitrarily. Domain and superfamily annotations were primarily obtained using the Foldseek-provided cath-v4_3_0.alphafold-v2.2022-11-22.tsv file. In cases where this information was unavailable, we referred to the cath-domain-list.csv file from the official CATH classification data available on the CATH website[58].

Feature selection

Feature selection presented several challenges because we wanted to avoid discarding features (genes) that might be important. We considered using backward elimination, where features are removed based on the decrease in model accuracy when a given feature is excluded from the training set[71]. If two features exhibit a highly similar association with the target variable, eliminating one has little impact on performance. Likewise, in forward selection, features are added based on their unique contribution to predictive power, meaning that strongly correlated features provide minimal additional information[71,72].

However, removing redundant isoform fractions solely based on redundancy is not ideal for our analysis. Consider a scenario where Gene A undergoes a linear isoform shift with aging, where isoform A_1 increases while A_2 decreases. If an entirely independent Gene B exhibits a similar pattern, with isoform B_1 increasing and B_2 decreasing, conventional feature elimination methods would likely discard one of these genes, as their predictive information is considered redundant.

Yet, if both genes undergo equally significant isoform shifts, capturing both is crucial for a comprehensive understanding of developmental and aging processes. Furthermore, whether these genes are functionally related within the same pathway or entirely independent, there is no way to determine a priori which isoform shift is more biologically relevant. Prematurely discarding one would risk losing valuable biological insight, emphasizing the need for a feature selection approach that preserves meaningful signals for downstream analysis.

Given these considerations, we chose to identify the strongest predictive genes one by one, as outlined in the Results section. The feature ablation approach[44], used to examine how predictive power is distributed across genes, also proved to be an effective method for feature selection. Starting with the most important features, we iteratively removed all isoforms associated with the top-ranked gene and retrained the model at each step. Although the procedure was intended to continue until model performance declined to a suboptimal level, this threshold was not breached even after the removal of a substantial number of genes (Figure 3.a). We ultimately selected the top 100 predictive genes for inclusion in our decision forest model—an arbitrary cutoff chosen to constrain the scope of downstream biological interpretation.

All training and testing splits were stratified to ensure a balanced distribution of samples across age ranges. To achieve this, we defined custom age bins based on the distribution and constraints of our dataset. These bins were more finely spaced in earlier developmental stages and were defined as follows: from −1 to −0.3 in intervals of 0.1, from −0.3 to 0 in a single 0.3 interval, from 0 to 1 year in intervals of 0.2, from 1 to 3 years in 1-year intervals, from 3 to 30 years in 3-year intervals, and from 30 to 100 years in 10-year intervals. This stratification step was kept in all subsequent models.

At each iteration, we trained a model on the training set using log-ages and identified the most important feature based on Gini scores. Each random forest regressor consisted of 10 estimator trees, with a minimum of 10 samples required per leaf. The top feature identified at each step corresponded to an isoform fraction vector associated with a parent gene. We then removed all

isoforms associated with that gene and repeated the process. Additionally, to assess the influence of log transformation, we trained separate models using both raw and log-age values at each step, comparing their R^2 scores to evaluate potential differences in predictive power. We observed minimal impact on the model's predictive accuracy, and no significant differences between the R^2 trends of the log-scaled and raw age models (Fig. 3.a).

Isoform Quantification and Filtering

RNA-Seq reads from all brain samples were mapped to transcripts in the CHESS 3[23] (v.3.1.1) gene catalog of the GRCh38 human genome (p.12) using Salmon[73] (version 1.10.0). For Salmon, the entire genome was used as a decoy sequence, and quantification was performed with the following options:
--validateMappings --mimicBT2 --recoverOrphans --gcBias.

Isoform quantifications were extracted as Transcripts Per Million (TPM) values and scaled using the "dtuScaledTPM" with the R package *tximport*[74]. These scaled read counts were used as the unit of expression in all downstream analyses. Quantified transcripts were filtered prior to DTU analysis with the SPIT pre-filtering module which applies the following criteria in order:

- An isoform must be expressed at a level of at least 1 Count Per Million (CPM) in a minimum of n_{small} samples, where n_{small} is set to the default 12.
- An isoform must have a positive read count in at least 20% of samples in both groups of analysis.
- A gene must have a minimum read count of 10 in at least 10 samples.
- An isoform must have an IF value greater than 0.1 in at least n_{small}samples.
- Following these filtering steps, a gene must be left with more than 1 isoforms.
- For each gene, the control group should exhibit a dominant isoform. This is measured by the same isoform having the highest IF in at least 50% of control samples.

In our analysis of transcriptomic changes across brain development and aging, we designated the prenatal group as the control and the postnatal group as the case. This setup was particularly suitable for SPIT, which is designed to identify potential substructures within the case group. Given the wide age range in the postnatal cohort—from infancy to late adulthood—isoform shifts may be more pronounced within specific age subsets than across the entire group. SPIT is capable of detecting such subgroup-specific patterns by accounting for bimodal or multimodal distributions within the case group, thereby preserving signals that might otherwise be overlooked in traditional DTU analyses. While changes in isoform usage are also expected during prenatal development, our dataset is largely composed of samples between 15 and 20 post-conception weeks, limiting the ability to detect meaningful clustering or temporal substructure within the prenatal group.

For the GTEx heart samples, RNA-Seq reads were again quantified using Salmon[73] with the RefSeq annotation (release 110). As this analysis did not involve a case–control design, we

focused on including only the most stably expressed genes to minimize noise in the predictive model. Genes were filtered based on their median CPM across all samples, retaining the top 1,000 genes with at least two annotated isoforms. Varying this threshold to 500 or 5,000 genes did not result in appreciable changes in model performance.

Alignment

To assess transcript coverage and examine stage-specific inclusion or exclusion of splice sites, we performed read alignments using HISAT2[75] in "splice-aware" mode for both human and macaque datasets. Human samples were aligned to the GRCh38.p12 reference genome, while macaque samples were aligned to the Mmul_10[76] genome assembly.

To generate a composite view of consistent alignment patterns across multiple samples, we used TieBrush[77] to aggregate alignments, followed by TieCov to compute coverage profiles. The resulting coverage data were then filtered, normalized, and visualized using the Integrative Genomics Viewer (IGV)[78] to provide a comprehensive picture of isoform usage across developmental stages.

RNA-Sequencing and Quality Control

We limited our analysis to control individuals, which were defined as those having no history of psychological issues, psychiatric care, or substance abuse. Tissue collection and sequencing protocols for the brain samples from LIBD have been previously described by Lipska *et al*[79]. and Jaffe *et al*[80]. The median RNA integrity number (RIN) of the control samples used in our analysis was 8.5, with over 80% of the samples having RIN values of 8 or higher. Samples with RIN values below 5 were excluded from the analysis. Additionally, we assessed the quality of the RNA-seq reads using FastQC[81] and MultiQC[82], which showed consistently high results, with 98% of our samples having the majority of their reads with a median per-base sequence quality score above 30.

Comprehensive Macaque Annotation

To process rhesus macaque samples, we used the Mmul_10 genome assembly[76,83-85], which was the most contiguous and accurate assembly available at the time of the study.

Although several alternative gene catalogues for Mmul_10 are available, substantial differences in covered loci and alternative splicing events made those inadequate for our analysis. Instead, we chose to adopt a more comprehensive genome annotation that will be released in a forthcoming study. This annotation represents a union of the Ensembl and NCBI annotations for the Mmul_10 assembly. This combined annotation comprises 158,009 transcripts and 50,083 genes, encompassing both protein-coding and non-coding biotypes.

Data availability

The source code supporting all modeling and statistical analyses, along with processed data, selected gene sets, and structural domain search results, has been deposited in Zenodo: https://doi.org/10.5281/zenodo.15335237

Acknowledgements
This work was supported in part by the U.S. National Institutes of Health under grants R01-MH123567, R01-HG006677, and R35-GM156470.

We would like to acknowledge Martin Steinegger and Milot Mirdita for their contributions to the functional analysis pipeline. We also thank Ales Varabyou for assembling the comprehensive Macaca mulatta transcriptome annotation.

ChatGPT (OpenAI) was used to assist with language refinement in the preparation of this manuscript.

References

1 Stiles, J. & Jernigan, T. L. The basics of brain development. *Neuropsychol Rev* **20**, 327-348 (2010). https://doi.org:10.1007/s11065-010-9148-4

2 Stiles, J. *The fundamentals of brain development: Integrating nature and nurture*. (Harvard University Press, 2008).

3 Douet, V., Chang, L., Cloak, C. & Ernst, T. Genetic influences on brain developmental trajectories on neuroimaging studies: from infancy to young adulthood. *Brain Imaging Behav* **8**, 234-250 (2014). https://doi.org:10.1007/s11682-013-9260-1

4 Cardoso-Moreira, M. *et al.* Gene expression across mammalian organ development. *Nature* **571**, 505-509 (2019). https://doi.org:10.1038/s41586-019-1338-5

5 Yamamoto, R. *et al.* Tissue-specific impacts of aging and genetics on gene expression patterns in humans. *Nature Communications* **13**, 5803 (2022). https://doi.org:10.1038/s41467-022-33509-0

6 Su, C. H., D, D. & Tarn, W. Y. Alternative Splicing in Neurogenesis and Brain Development. *Front Mol Biosci* **5**, 12 (2018). https://doi.org:10.3389/fmolb.2018.00012

7 Melé, M. *et al.* The human transcriptome across tissues and individuals. *Science* **348**, 660-665 (2015). https://doi.org:doi:10.1126/science.aaa0355

8 Raj, B. & Blencowe, B. J. Alternative splicing in the mammalian nervous system: recent insights into mechanisms and functional roles. *Neuron* **87**, 14-27 (2015).

9 Irfan, M. *et al.* SNAP-25 isoforms differentially regulate synaptic transmission and long-term synaptic plasticity at central synapses. *Scientific Reports* **9**, 6403 (2019). https://doi.org:10.1038/s41598-019-42833-3

10 Bark, C. *et al.* Developmentally regulated switch in alternatively spliced SNAP-25 isoforms alters facilitation of synaptic transmission. *J Neurosci* **24**, 8796-8805 (2004). https://doi.org:10.1523/jneurosci.1940-04.2004

11 Bark, I. C. & Wilson, M. C. Human cDNA clones encoding two different isoforms of the nerve terminal protein SNAP-25. *Gene* **139**, 291-292 (1994). https://doi.org:https://doi.org/10.1016/0378-1119(94)90773-0

12 Antonucci, F. *et al.* SNAP-25, a Known Presynaptic Protein with Emerging Postsynaptic Functions. *Frontiers in Synaptic Neuroscience* **8** (2016). https://doi.org:10.3389/fnsyn.2016.00007

13 Bark, I. C. Structure of the Chicken Gene for SNAP-25 Reveals Duplicated Exons Encoding Distinct Isoforms of the Protein. *Journal of Molecular Biology* **233**, 67-76 (1993). https://doi.org:https://doi.org/10.1006/jmbi.1993.1485

14 Gopaul, K. R. *et al.* Developmental Time Course of SNAP-25 Isoforms Regulate Hippocampal Long-Term Synaptic Plasticity and Hippocampus-Dependent Learning. *Int J Mol Sci* **21** (2020). https://doi.org:10.3390/ijms21041448

15 Gogtay, N. *et al.* Dynamic mapping of human cortical development during childhood through early adulthood. *Proceedings of the National Academy of Sciences* **101**, 8174-8179 (2004). https://doi.org:doi:10.1073/pnas.0402680101

16 Beveridge, N. J. *et al.* Maturation of the human dorsolateral prefrontal cortex coincides with a dynamic shift in microRNA expression. *Schizophr Bull* **40**, 399-409 (2014). https://doi.org:10.1093/schbul/sbs198

17 Collado-Torres, L. *et al.* Regional Heterogeneity in Gene Expression, Regulation, and Coherence in the Frontal Cortex and Hippocampus across Development and Schizophrenia. *Neuron* **103**, 203-216.e208 (2019). https://doi.org:https://doi.org/10.1016/j.neuron.2019.05.013
18 Jaffe, A. E. *et al.* Developmental regulation of human cortex transcription and its clinical relevance at single base resolution. *Nature Neuroscience* **18**, 154-161 (2015). https://doi.org:10.1038/nn.3898
19 Zhang, X. *et al.* Cell-Type-Specific Alternative Splicing Governs Cell Fate in the Developing Cerebral Cortex. *Cell* **166**, 1147-1162.e1115 (2016). https://doi.org:https://doi.org/10.1016/j.cell.2016.07.025
20 Weyn-Vanhentenryck, S. M. *et al.* Precise temporal regulation of alternative splicing during neural development. *Nature Communications* **9**, 2189 (2018). https://doi.org:10.1038/s41467-018-04559-0
21 Mazin, P. *et al.* Widespread splicing changes in human brain development and aging. *Mol Syst Biol* **9**, 633 (2013). https://doi.org:10.1038/msb.2012.67
22 Ham, S. & Lee, S.-J. V. Advances in transcriptome analysis of human brain aging. *Experimental & Molecular Medicine* **52**, 1787-1797 (2020). https://doi.org:10.1038/s12276-020-00522-6
23 Varabyou, A. *et al.* CHESS 3: an improved, comprehensive catalog of human genes and transcripts based on large-scale expression data, phylogenetic analysis, and protein structure. *Genome Biology* **24**, 249 (2023). https://doi.org:10.1186/s13059-023-03088-4
24 Erdogdu, B., Varabyou, A., Hicks, S. C., Salzberg, S. L. & Pertea, M. Detecting differential transcript usage in complex diseases with SPIT. *Cell Reports Methods* **4** (2024). https://doi.org:10.1016/j.crmeth.2024.100736
25 Sutton, R. B., Fasshauer, D., Jahn, R. & Brunger, A. T. Crystal structure of a SNARE complex involved in synaptic exocytosis at 2.4 Å resolution. *Nature* **395**, 347-353 (1998). https://doi.org:10.1038/26412
26 Johansson, J. U. *et al.* An ancient duplication of exon 5 in the Snap25 gene is required for complex neuronal development/function. *PLoS Genet* **4**, e1000278 (2008). https://doi.org:10.1371/journal.pgen.1000278
27 Sørensen, J. B. *et al.* Differential Control of the Releasable Vesicle Pools by SNAP-25 Splice Variants and SNAP-23. *Cell* **114**, 75-86 (2003). https://doi.org:https://doi.org/10.1016/S0092-8674(03)00477-X
28 Nagy, G. *et al.* Alternative splicing of SNAP-25 regulates secretion through nonconservative substitutions in the SNARE domain. *Mol Biol Cell* **16**, 5675-5685 (2005). https://doi.org:10.1091/mbc.e05-07-0595
29 Prescott, G. R. & Chamberlain, L. H. Regional and developmental brain expression patterns of SNAP25 splice variants. *BMC Neuroscience* **12**, 35 (2011). https://doi.org:10.1186/1471-2202-12-35
30 Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nature Genetics* **45**, 580-585 (2013). https://doi.org:10.1038/ng.2653
31 Heath, D., Kasif, S. & Salzberg, S. in *Advances in Psychology* Vol. 113 (eds Barbara Gorayska & Jacob L. Mey) 305-317 (North-Holland, 1996).

32 Breiman, L. Random Forests. *Machine Learning* **45**, 5-32 (2001). https://doi.org:10.1023/A:1010933404324
33 Breiman, L., Friedman, J., Stone, C. J. & Olshen, R. A. *Classification and Regression Trees*. (Taylor & Francis, 1984).
34 Arstikaitis, P. *et al.* Paralemmin-1, a modulator of filopodia induction is required for spine maturation. *Mol Biol Cell* **19**, 2026-2038 (2008). https://doi.org:10.1091/mbc.e07-08-0802
35 Kutzleb, C. *et al.* Paralemmin, a prenyl-palmitoyl-anchored phosphoprotein abundant in neurons and implicated in plasma membrane dynamics and cell process formation. *J Cell Biol* **143**, 795-813 (1998). https://doi.org:10.1083/jcb.143.3.795
36 Basile, M. *et al.* Paralemmin interacts with D3 dopamine receptors: implications for membrane localization and cAMP signaling. *Arch Biochem Biophys* **446**, 60-68 (2006). https://doi.org:10.1016/j.abb.2005.10.027
37 Morales, J. *et al.* A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* **604**, 310-315 (2022). https://doi.org:10.1038/s41586-022-04558-8
38 O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733-745 (2016). https://doi.org:10.1093/nar/gkv1189
39 Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nature Methods* **19**, 679-682 (2022). https://doi.org:10.1038/s41592-022-01488-1
40 Mo, D., Zhao, Y. & Balajee, A. S. Human RecQL4 helicase plays multifaceted roles in the genomic stability of normal and cancer cells. *Cancer Letters* **413**, 1-10 (2018). https://doi.org:https://doi.org/10.1016/j.canlet.2017.10.021
41 Singh, D. K. *et al.* The involvement of human RECQL4 in DNA double-strand break repair. *Aging cell* **9**, 358-371 (2010).
42 Lu, H. *et al.* Senescence induced by RECQL4 dysfunction contributes to Rothmund–Thomson syndrome features in mice. *Cell death & disease* **5**, e1226-e1226 (2014).
43 Colombo, E. A. *et al.* Novel physiological RECQL4 alternative transcript disclosed by molecular characterisation of Rothmund-Thomson Syndrome sibs with mild phenotype. *Eur J Hum Genet* **22**, 1298-1304 (2014). https://doi.org:10.1038/ejhg.2014.18
44 Wasserman, L. *All of Statistics: A Concise Course in Statistical Inference*. (Springer, 2005).
45 GIEDD, J. N. Structural Magnetic Resonance Imaging of the Adolescent Brain. *Annals of the New York Academy of Sciences* **1021**, 77-85 (2004). https://doi.org:https://doi.org/10.1196/annals.1308.009
46 Işıldak, U., Somel, M., Thornton, J. M. & Dönertaş, H. M. Temporal changes in the gene expression heterogeneity during brain development and aging. *Scientific Reports* **10**, 4080 (2020). https://doi.org:10.1038/s41598-020-60998-0
47 Libert, S., Chekholko, A. & Kenyon, C. (eLife Sciences Publications, Ltd, 2024).
48 Jylhävä, J., Pedersen, N. L. & Hägg, S. Biological Age Predictors. *EBioMedicine* **21**, 29-36 (2017). https://doi.org:10.1016/j.ebiom.2017.03.046
49 Marzola, P., Melzer, T., Pavesi, E., Gil-Mohapel, J. & Brocardo, P. S. Exploring the Role of Neuroplasticity in Development, Aging, and Neurodegeneration. *Brain Sci* **13** (2023). https://doi.org:10.3390/brainsci13121610

50 Power, J. D. & Schlaggar, B. L. Neural plasticity across the lifespan. *Wiley Interdiscip Rev Dev Biol* **6** (2017). https://doi.org:10.1002/wdev.216
51 Fleg, J. L. & Strait, J. Age-associated changes in cardiovascular structure and function: a fertile milieu for future disease. *Heart Fail Rev* **17**, 545-554 (2012). https://doi.org:10.1007/s10741-011-9270-2
52 Dai, D. F., Chen, T., Johnson, S. C., Szeto, H. & Rabinovitch, P. S. Cardiac aging: from molecular mechanisms to significance in human health and disease. *Antioxid Redox Signal* **16**, 1492-1526 (2012). https://doi.org:10.1089/ars.2011.4179
53 Zimin, A. V. *et al.* A new rhesus macaque assembly and annotation for next-generation sequencing analyses. *Biol Direct* **9**, 20 (2014). https://doi.org:10.1186/1745-6150-9-20
54 Alldritt, S. *et al.* Brain Charts for the Rhesus Macaque Lifespan. *bioRxiv* (2024). https://doi.org:10.1101/2024.08.28.610193
55 Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276-277 (2000). https://doi.org:10.1016/s0168-9525(00)02024-2
56 van Kempen, M. *et al.* Foldseek: fast and accurate protein structure search. *Biorxiv*, 2022.2002. 2007.479398 (2022).
57 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021). https://doi.org:10.1038/s41586-021-03819-2
58 Orengo, C. A. *et al.* CATH--a hierarchic classification of protein domain structures. *Structure* **5**, 1093-1108 (1997). https://doi.org:10.1016/s0969-2126(97)00260-8
59 Leipe, D. D., Wolf, Y. I., Koonin, E. V. & Aravind, L. Classification and evolution of P-loop GTPases and related ATPases11Edited by J. Thornton. *Journal of Molecular Biology* **317**, 41-72 (2002). https://doi.org:https://doi.org/10.1006/jmbi.2001.5378
60 Longo, L. M. *et al.* On the emergence of P-Loop NTPase and Rossmann enzymes from a Beta-Alpha-Beta ancestral fragment. *eLife* **9**, e64415 (2020). https://doi.org:10.7554/eLife.64415
61 Romero Romero, M. L. *et al.* Simple yet functional phosphate-loop proteins. *Proceedings of the National Academy of Sciences* **115**, E11943-E11950 (2018). https://doi.org:doi:10.1073/pnas.1812400115
62 Kinoshita, K., Sadanami, K., Kidera, A. & Go, N. Structural motif of phosphate-binding site common to various protein superfamilies: all-against-all structural comparison of protein–mononucleotide complexes. *Protein Engineering* **12**, 11-14 (1999). https://doi.org:10.1093/protein/12.1.11
63 Schürmann, A., Brauers, A., Massmann, S., Becker, W. & Joost, H. G. Cloning of a novel family of mammalian GTP-binding proteins (RagA, RagBs, RagB1) with remote similarity to the Ras-related GTPases. *J Biol Chem* **270**, 28982-28988 (1995). https://doi.org:10.1074/jbc.270.48.28982
64 Figlia, G. *et al.* Brain-enriched RagB isoforms regulate the dynamics of mTORC1 activity through GATOR1 inhibition. *Nature Cell Biology* **24**, 1407-1421 (2022). https://doi.org:10.1038/s41556-022-00977-x
65 Sancak, Y. *et al.* Ragulator-Rag complex targets mTORC1 to the lysosomal surface and is necessary for its activation by amino acids. *Cell* **141**, 290-303 (2010). https://doi.org:10.1016/j.cell.2010.02.024

66 Figlia, G., Gerber, D. & Suter, U. Myelination and mTOR. *Glia* **66**, 693-707 (2018). https://doi.org:10.1002/glia.23273
67 Lipton, J. O. & Sahin, M. The neurology of mTOR. *Neuron* **84**, 275-291 (2014). https://doi.org:10.1016/j.neuron.2014.09.034
68 van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* **42**, 243-246 (2024). https://doi.org:10.1038/s41587-023-01773-0
69 Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* **50**, D439-D444 (2021). https://doi.org:10.1093/nar/gkab1061
70 Sommer, M. J. *et al.* Structure-guided isoform identification for the human transcriptome. *Elife* **11**, e82556 (2022).
71 Draper, N. R. & Smith, H. *Applied Regression Analysis*. (Wiley, 1998).
72 Efroymson, M. A. Multiple regression analysis. *Mathematical methods for digital computers*, 191-203 (1960).
73 Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14**, 417-419 (2017). https://doi.org:10.1038/nmeth.4197
74 Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Res* **4**, 1521 (2015). https://doi.org:10.12688/f1000research.7563.2
75 Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* **37**, 907-915 (2019). https://doi.org:10.1038/s41587-019-0201-4
76 Warren, W. C. *et al.* Sequence diversity analyses of an improved rhesus macaque genome enhance its biomedical utility. *Science* **370** (2020). https://doi.org:10.1126/science.abc6617
77 Varabyou, A., Pertea, G., Pockrandt, C. & Pertea, M. TieBrush: an efficient method for aggregating and summarizing mapped reads across large datasets. *Bioinformatics* **37**, 3650-3651 (2021). https://doi.org:10.1093/bioinformatics/btab342
78 Robinson, J. T. *et al.* Integrative genomics viewer. *Nature Biotechnology* **29**, 24-26 (2011). https://doi.org:10.1038/nbt.1754
79 Lipska, B. K. *et al.* Critical Factors in Gene Expression in Postmortem Human Brain: Focus on Studies in Schizophrenia. *Biological Psychiatry* **60**, 650-658 (2006). https://doi.org:10.1016/j.biopsych.2006.06.019
80 Jaffe, A. E. *et al.* Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. *Nature Neuroscience* **21**, 1117-1125 (2018). https://doi.org:10.1038/s41593-018-0197-y
81 Andrews, S. (Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom, 2010).
82 Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047-3048 (2016).
83 Gokey, N. G. *et al.* Molecular analyses of mtDNA deletion mutations in microdissected skeletal muscle fibers from aged rhesus monkeys. *Aging Cell* **3**, 319-326 (2004). https://doi.org:10.1111/j.1474-9728.2004.00122.x

84 Hughes, J. F. *et al.* Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* **483**, 82-86 (2012). https://doi.org:10.1038/nature10843

85 Rogers, J. Genomic resources for rhesus macaques (Macaca mulatta). *Mamm Genome* **33**, 91-99 (2022). https://doi.org:10.1007/s00335-021-09922-z