TOOLS AND RESOURCES

voyAGEr: free web interface for the analysis of age-related gene expression alterations in human tissues

Short running title:

Age-related gene changes profiling app

Arthur L. Schneider[1,2,†]
arthur.schneider@centraliens-nantes.org

Rita Martins-Silva[1,†]
rita.silva@medicina.ulisboa.pt

Alexandre Kaizeler[1,†]
alexandre.afonso@medicina.ulisboa.pt

Nuno Saraiva-Agostinho[1,3]
nunodanielagostinho@gmail.com

Nuno L. Barbosa-Morais[1,*]
nmorais@medicina.ulisboa.pt

[1] Instituto de Medicina Molecular João Lobo Antunes, Faculdade de Medicina, Universidade de Lisboa, Av. Prof. Egas Moniz, 1649-028 Lisbon, Portugal

[2] *Present address:* Sia Partners, 4 Rue Voltaire, 44000 Nantes, France

[3] *Present address:* European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, CB10 1SD, Cambridge, UK

[†] Equal contributions

* Corresponding author

Abstract

We herein introduce voyAGEr, an online graphical interface to explore age-related gene expression alterations in 49 human tissues. voyAGEr offers a visualisation and statistical toolkit for the finding and functional exploration of sex- and tissue-specific transcriptomic changes with age. In its conception, we developed a novel bioinformatics pipeline leveraging RNA sequencing data, from the GTEx project, encompassing more than 900 individuals. voyAGEr reveals transcriptomic signatures of the known asynchronous ageing between tissues, allowing the observation of tissue-specific age-periods of major transcriptional changes, associated with alterations in different biological pathways, cellular composition, and disease conditions.

Notably, voyAGEr was created to assist researchers with no expertise in bioinformatics, providing a supportive framework for elaborating, testing and refining their hypotheses on the molecular nature of human ageing and its association with pathologies, thereby also aiding in the discovery of novel therapeutic targets. voyAGEr is freely available at https://compbio.imm.medicina.ulisboa.pt/app/voyAGEr.

Keywords

Humans; Ageing; Computational Biology; Transcriptome; Gene Expression Profiling

Main text

Introduction

The ageing-associated progressive loss of proper tissue homeostasis maintenance makes age a prevalent risk factor for many human pathologies, including cancer, neurodegenerative and cardiovascular diseases [1–3]. A better comprehension of the molecular mechanisms of human ageing is thus required for the development and effective application of therapies targeting its associated morbidities.

Dynamic transcriptional alterations accompany most physiological processes occurring in human tissues [4]. Transcriptomic analyses of tissue samples can thus provide snapshots of cellular states therein and insights into how their modifications over time impact tissue physiology. A small proportion of transcripts has indeed been shown to vary with age in tissue- [5] and sex-specific [6–11] manners. Such variations reflect dysregulations of gene expression that underlie cellular disfunctions [5].

Many studies analysed the age-related changes in gene expression in rodent tissues [12–18], emphasising the role in ageing of genes related to inflammatory responses, cell cycle or the electron transport chain. However, while it is possible to monitor the modifications in gene expression over time in those species by sequencing transcriptomes of organs of littermates at different ages, as a close surrogate of longitudinality, such studies cannot be conducted in humans for ethical reasons. Indeed, most studies aimed at profiling ageing-related gene expression changes in human tissues focused on a single tissue (e.g. muscle [19–21], kidney [22], brain [23–27], skin [28,29], blood [30,31], liver [32], retina [33]) and/or are limited to a comparison between young and old individuals [20,21,28,32,33], failing to fully capture the changes of the tissue-specific gene expression landscape throughout ageing [5]. A few studies were nonetheless led on more than one tissue in humans, from post-mortem samples [34,35] and biopsies [36,37], and in mice [18,36] and rats [38]. The age-related transcriptional profiles derived therein, either from regression [18,34,35,37] or comparison between age groups [36,38], highlight an asynchronous ageing of tissues (discussed in [39]), with some of them more affected by age-related gene expression changes associated with biological mechanisms known to be

78 impacted by ageing such as mitochondrial activity or metabolic homeostasis. In particular,
79 tissue-specific periods of major transcriptional changes in the fifth and eighth decades of
80 human lifespan have been revealed [34], reflecting the so-called digital ageing [39], consistent
81 with what is observed in mice [17,18]. Furthermore, despite outlining the tissue specificity of
82 the transcriptomic signatures of human ageing, some synchronisation was found between
83 tissues like lung, heart and whole blood, which exhibit a co-ageing pattern [35]. Nevertheless,
84 as each study followed its own specific procedures, from sample collection to data
85 processing, results from these analyses are hard to compare with one another.
86 Processed data from those studies have not been made easily accessible and
87 interpretable to researchers lacking computational proficiency but aiming to use them to
88 test their novel hypotheses. To fill this void, we have developed voyAGEr, a web application
89 providing flexible visualisation of comprehensive functional analyses of gene expression
90 alterations occurring in 49 human tissues with age in each biological sex. We leverage the
91 large RNA-seq dataset from the Genotype-Tissue Expression (GTEx) project [40], encompassing
92 tissue samples from hundreds of donors aged from 20 to 70 years, with a pipeline for gene
93 expression profiling with an optimised temporal resolution. voyAGEr allows to investigate
94 ageing from two perspectives: (i) gene-centric – how each gene's tissue-specific expression
95 progresses with age; and (ii) tissue-centric – how tissue-specific transcriptomes change with
96 age. Additionally, voyAGEr enables the examination of modules of co-expressed genes
97 altered with age in 4 tissues (brain cortex, skeletal muscle, left ventricle of the heart, whole
98 blood), namely their enrichment in specific cell types, biological pathways, and association
99 with diseases. We therefore expect voyAGEr to become a valuable support tool for
100 researchers aiming to uncover the molecular mechanisms underlying human ageing.
101 Moreover, being open-source, voyAGEr can be adapted by fellow developers to be used with
102 alternative datasets (e.g. from other species) or to incorporate other specific functionalities.
103 voyAGEr is freely available at
104 https://compbio.imm.medicina.ulisboa.pt/app/voyAGEr.

105 # Results

106

107 voyAGEr's interactive exploration of tissue-specific gene expression landscapes in ageing is
108 based on sequential fitting of linear models (v. Materials and Methods) to estimate, for each
109 gene in each tissue:

110 (i) the *Age* effect, i.e., how the age-associated changes in gene expression evolve
111 with age itself;
112 (ii) the *Sex* effect, i.e., how the differences in gene expression between sexes evolve
113 with age;
114 (iii) the *Age&Sex* interaction effect, i.e., how the differences between sexes of age-
115 associated changes in gene expression evolve with age.

116 We named our approach Shifting Age Range Pipeline for Linear Modelling (ShARP-LM).
117 Briefly, this method consists in performing differential gene expression (with gene
118 expression as a function of the donors' *Age*, *Sex* and *Age&Sex* interaction) in moving age
119 windows spanning 16 years. By considering the percentage of genes altered in each age
120 range, we can highlight age periods of major tissue-specific transcriptomic alterations
121 (**Figure 1**).

122

123 ## *Gene*-centric analyses of human tissue-specific expression changes across age

124 The progression of tissue-specific expression of a particular gene across age can be
125 examined in voyAGEr's *Gene* tab. By entering its HGNC symbol in the Gene selector, the user
126 has access to graphical summaries of the gene's tissue-specific expression (sub-tab *Profile*)
127 (**Figure 2A**) and significance of age-related changes in its expression due to the *Age*, *Sex*, and
128 *Age&Sex* effects (sub-tab *Alteration*) (**Figure 2B**) across age. Results can be displayed in a
129 heatmap for all tissues or in a scatter plot for a chosen individual tissue (**Figure 2C**). When
130 the gene is studied in a single tissue, the user can graphically and statistically profile the
131 association of the donors' sex and reported conditions (e.g., history of heart attack or
132 pneumonia) with the gene's expression profile. A table summarizing the donors' metadata is
133 also shown (**Figure 2C**). The user can interactively select donors of interest on the scatter
134 plot and further examine their information in the automatically subsetted table.

135 An example of a process whose molecular mechanisms are of particular interest to
136 researchers in the ageing field is cellular senescence. Senescence is a stress-induced cell
137 cycle arrest limiting proliferation of potentially oncogenic cells but progressively creating an
138 inflammatory environment in tissues as they age [41,42]. *CDKN2A,* that encodes cell cycle
139 regulatory protein p16^{INK4A} known to accumulate in senescent cells [43,44], has its expression
140 increased with age in the vast majority of tissues profiled (**Supplementary Figure 1A**).
141 Similarly, reduced levels of proliferation markers, such as *PCNA* [45] and *MKI67* [46], can be
142 studied as putative markers of ageing of certain tissues. These genes have their expression
143 altered with age in the lung and display a similar expression profile (decreasing from 25 to
144 30 years old (y.o.), constant between 35 and 50 y.o. and decreasing in older ages) (**Figure**
145 **2C**). However, these trends appear to vary according to the donor's history of non-
146 metastatic cancer (**Supplementary Figure 1B**), illustrating voyAGEr's use in helping to find
147 associations between gene expression and age-related diseases.
148 On a different note, sex biases have been reported in the expression of *SALL1* and *DDX43* in
149 adipose tissue and lung, respectively [10]. voyAGEr allows to not only recapitulate those
150 observations but also assess the temporal window where these changes occur
151 (**Supplementary Figure 2**).
152

153 ## *Tissue*-specific assessment of gene expression changes across age

154 ### *Peaks of gene expression alterations*

155 The landscape of global tissue-specific gene expression alteration across age can be
156 examined in voyAGEr's *Tissue* main tab. A heatmap displaying, for all tissues, the statistical
157 significance over age (v. Materials and Methods) of the proportion of genes altered with
158 *Age*, *Sex* or *Age&Sex* (depending on the user's interest) is initially shown (**Figure 3A**),
159 illustrating the aforementioned asynchronous ageing of tissues observed for humans and
160 rodents [18,32–36,39].
161 The user can then spot the age periods with the most significant gene expression alterations
162 in a selected tissue (**Figure 3B a**), and identify the associated altered genes (**Figure 3B b**).
163 The user can also plot the expression of a given gene of interest across age together with the
164 significance of its expression modification with *Age*, *Sex* or *Age&Sex* (**Figure 3B c**).

165 An example of voyAGEr's capabilities is illustrated in **Supplementary Figure 3**, showing
166 substantial transcriptomic alterations in the uterus from the late forties to the early fifties,
167 overlapping with the age distribution of menopausal onset[47], which could explain the
168 observed molecular modifications.
169 It is also possible to visualise tissues with more than one period of transcriptomic changes,
170 and to individually inspect these periods. As an example, the subcutaneous adipose tissue
171 appears to go through two main periods of transcriptional changes with age: a major one at
172 the late 20's (~13 % of altered genes), and a minor one after 45 years (~5% of altered genes)
173 (**Figure 3B a**). The most altered genes in this first peak appear to have their expression
174 modified only at this precise age period (e.g. *PRELID1*, *RUNX1T1*, *TUBB4B*, *FGFRL1* and
175 *MALSU1*). Similarly, mitochondrial genes (e.g. *MT-CYB*, *MT-ND4*, *MT-ATP6*, *MT-ND2*) (**Figure**
176 **3B b**) appear to be the most altered genes in the second peak (**Figure 3B c**). This particularity
177 suggests that different sets of genes drive the periods of major transcriptional changes,
178 which begs assessing if they reflect the activation of distinct biological processes.
179

180 ## Gene Set Enrichment

181 The user can explore the biological functions of the set of genes underlying each peak of
182 transcriptomic changes by assessing their enrichment in curated pathways from the
183 Reactome database [48]. voyAGEr performs Gene Set Enrichment Analysis (GSEA) [49] and the
184 user can visualise heatmaps displaying the evolution over age of the resulting normalised
185 enrichment score (NES[49], reflecting the degree to which a pathway is over- or under-
186 represented in a subset of genes) for a given tissue, effect (*Age*, *Sex* or *Age&Sex*) and
187 Reactome pathway (all, or user-selected) (**Figure 4A a**). To reduce redundancy and facilitate
188 the understanding of their biological relevance, we clustered those pathways into families
189 that also include Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [50] and Gene
190 Ontology (GO) Biological Processes of level 3 [51]. Briefly, we clustered gene sets from the 3
191 sources based on the overlap of their genes (v. Materials and Methods), thereby creating
192 families of highly functionally related pathways. Taking advantage of the complementary
193 and distinct terminology in Reactome, KEGG and GO, the user can interpret each family's
194 broad biological function by looking at the word cloud of its most prevalent terms, and
195 browsing the list of its associated pathways (**Figure 4A b**). For example, although most
196 pathways enriched in the two aforementioned peaks of altered genes in subcutaneous

197 adipose tissue were different, there is an overlap of pathways related to metabolism,
198 including various mitochondrial processes (**Figure 4A)**. This highlights the importance of
199 integrating individual gene data with pathway enrichment analysis to garner more
200 comprehensive insights on the biological relevance of those changes.
201 The peaks of transcriptomic changes can also be examined for enrichment in a user-provided
202 gene set (**Figure 4B**). As expression of senescence-related *CDKN2A* is increased in the
203 subcutaneous adipose tissue with age (**Supplementary Figure 1A),** we hypothesised that
204 other senescence-associated genes may be augmented too. Thus we used that voyAGEr
205 functionality, using the Senequest [42] geneset (supported by at least 4 sources) to test it,
206 observing no significant alterations (**Figure 4B**).
207

208 Modules of co-expressed genes

209 voyAGEr also allows functional analyses of modules of co-expressed genes (i.e. genes with
210 highly correlated expression across samples, defined by weighted correlation network
211 analysis [52]). Genes in the same module are likely to be co-regulated and share biological
212 functions or associations with phenotypical or pathological traits [53]. Those modules may also
213 act as markers of core transcriptional features of cellular activity and identity [54].
214 Concretely, voyAGEr enables the user to visualise how the expression of modules of genes
215 that are associated with a specific cell type, biological pathway or disease progresses over
216 age in a specific tissue. After selecting one tissue of interest, the user has, for each module,
217 access to four levels of information:
218 (i) *Expression*: its eigengene expression progression over age (**Figure 5C**);
219 (ii) *Cell types*: its enrichment in specific cell types, based on cell type signatures
220 found in the literature (**Figure 5C**);
221 (iii) *Pathways*: its enrichment in Reactome pathways;
222 (iv) *Diseases*: its enrichment in disease markers, based on gene-disease associations
223 from DisGeNET [55,56], calculated with both the `disgenet2r` package [55] and with
224 Fisher's tests (**Figure 5B**).
225 By default, for each tissue, results are displayed in the form of heatmaps of expression
226 (centred and scaled) illustrating how all modules evolve with age (**Figure 5A**). The user also
227 has the possibility to select a module of interest and see its eigengene progression over age

228 in a scatter plot (**Figure 5C**), lists of its association with diseases and pathways ordered by
229 significance, and a TreeMap for its cell type enrichments (**Figure 5C**). In the example of
230 **Figure 5C**, the "MEblue" module, comprising 241 genes co-expressed in the brain cortex,
231 shows significant enrichment in astrocyte markers. The apparent increase of this module's
232 expression with age may reflect the known age-related changes in astrocyte activation [57] and
233 the relative weakening of neuronal activity ("MEyellow" and "MEpink" modules).
234 As in the *Gene* tab, the user can separate donors based on their sex and associated medical
235 conditions in the scatter plot of the eigengene expression progression. On the Pathways and
236 Diseases-Manual tabs below the main plot, the user can also visualise the contingency table
237 from that specific disease/pathway, on the corresponding column.

238 Discussion

239

240 voyAGEr provides a framework to examine the progression of gene expression over age in
241 several human tissues, serving as a valuable resource for the ageing research community. In
242 particular, it helps to identify tissue-specific age periods of major transcriptomic alterations.
243 The results of our analyses show the complexity of human biological ageing by stressing its
244 tissue specificity [34] and non-linear transcriptional progression throughout lifetime, consistent
245 with previous results from both proteomic [58] and transcriptomic [28,34] analyses. By revealing
246 and annotating the age-specific transcriptional trends in each tissue, voyAGEr aims at
247 assisting researchers in deciphering the cellular and molecular mechanisms associated with
248 the age-related physiological decline across the human body.
249 Due to the tissue-specific nature of the pre-processing steps (v. Read count data pre-
250 processing in the Methods section), and given that most of the plotted gene expression
251 distributions are centred and scaled by tissue, it is important to note that voyAGEr may not
252 be always suited for direct comparisons between different tissues. For instance, it does not
253 allow to directly ascertain if a gene exhibits different expression levels in different tissues or
254 if the expression of a particular gene in one tissue changes more drastically with age than in
255 another tissue. Furthermore, we must emphasise that the majority of GTEx donors
256 contributed samples to multiple tissues (**Supplementary Figure 4A**), potentially introducing
257 biases and confounders when comparing gene expression patterns between tissues. Our
258 analyses of variance (**Supplementary Figure 4B**) and downsampling to control for common

259 donors (**Supplementary Figures 4C-E**) suggest very limited global confounding between the
260 impacts of donor and age on gene expression and that any potential cross-tissue bias not to
261 depend much on the proportion of common donors (**Supplementary Figure 4E**). However,
262 this effect must be taken into account when comparing specific pairs of tissues (e.g., Colon –
263 Transverse and Whole Blood, **Supplementary Figure 4D**).
264 Additionally, voyAGEr allows to scrutinise and visually display the tissue-specific differences
265 in gene expression between biological sexes across age. Biological sex is an important factor
266 in the prevalence of ageing-associated diseases, as well as their age of onset, progression
267 [6,7,59] and sex-related differences in gene expression[8–11]. By profiling the age distribution of
268 such differences, voyAGEr can lead to a better understanding of their influence in the
269 aetiology of the sex specificities of human ageing. For instance, we were able to corroborate
270 findings on the sex-differential transcriptome of adult humans by Gershoni et al. [9], with
271 voyAGEr emphasising its tissue-specificity and allowing to discriminate the ages at which
272 sex-related biases appear to be more prevalent (**Supplementary Figure 5**).
273 Nonetheless, it is essential to interpret sex chromosome-specific gene results in voyAGEr
274 with caution. For instance, we observe elevated expression of Y-chromosome-specific *DDX3Y*
275 in males, whilst its female expression (expected to be zero) is very low, in the range of what
276 can be considered background noise (**Supplementary Figure 6**). Its age-related alterations
277 exhibit a distinctive peak around the age of 40 apparently driven by subtle changes of gene
278 expression in female samples, illustrating the need for the abovementioned caution.
279 One of the limitations of voyAGEr is that most GTEx tissue donors had health conditions and
280 their frequency increased with age, preventing us from defining a class of healthy individuals
281 and identifying age-associated transcriptomic changes that could be more confidently
282 proposed to happen independently of any disease progression. Whilst the large sample sizes
283 and inherent biological variability among individuals, reflected in the diversity of condition
284 combinations, are expected to mitigate significant confounding effects, voyAGEr also allows
285 users to evaluate how tissue-specific gene expression trends vary according to the donors'
286 diverse conditions (see **Supplementary Figure 1B**).
287 The development of voyAGEr was accompanied by that of a pipeline, ShARP-LM, that
288 facilitates the holistic depiction of the transcriptional landscapes of adult human tissues
289 throughout ageing with a yearly age resolution. We take advantage of the
290 comprehensiveness of the transcriptome collection from human tissues from the GTEx

291 project to make our analyses a valid surrogate of a currently undoable longitudinal study. It
292 confers our method enough statistical robustness to mitigate the inter-individual differences
293 and deal with the non-uniform distribution of the donors' ages. Nevertheless, it is worth
294 highlighting that the age distribution of donors does impact the statistical power for
295 detecting transcriptional changes. Consequently, we are more likely to identify significant
296 alterations (with p-value < 0.05 in our gene-centric analyses) within age ranges that are
297 more prevalent in our sample population, often characterised by older individuals
298 (**Supplementary Figure 7)**. When downsampling to ensure a balanced age distribution, a loss
299 of statistical power is apparent but a considerable positive correlation with the original
300 results is maintained and a substantial number of significant alterations remain so
301 (**Supplementary Figure 8**). This limitation is likely to be overcome by the accumulation of
302 transcriptomes of human tissues in public databases, promising a gradual increase in
303 accuracy and age resolution with which human transcriptomic ageing can be profiled.
304 Similarly, the expanding collection of single-cell transcriptomes in public databases is
305 yielding improved gene markers for an increasing diversity of human cell types, enhancing
306 the usefulness of leveraging bulk transcriptomes to study the impact of ageing on the
307 cellular composition of human tissues, for which the co-expression module approach in
308 voyAGEr provides a proof-of-concept.
309 Nonetheless, it is imperative to approach the module-based analysis with caution, as direct
310 and literal interpretations may be misleading. For instance, it is not uncommon to observe
311 an enrichment of "Rheumatoid Arthritis" in modules associated with various immune cell
312 types in anatomical locations, such as the brain cortex, where the disease does not directly
313 manifest. If a specific module associated with a condition like "Liver Cirrhosis" exhibits an
314 age-related increase in the brain cortex, of course this does not mean such disease ever
315 occurs within older brains. Nevertheless, we consider that the module-based approach can
316 serve as a valuable resource for generating hypotheses.
317 Given its open-source nature, voyAGEr is envisaged to be a continually evolving resource,
318 able to accommodate new data and expand its functionalities, namely by incorporating
319 additional tissues into the modules section and integrating perturbagen data for inference of
320 molecular causes underlying observed gene expression alterations and small molecules to
321 target them for therapeutic purposes [60,61].

322 As an *in silico* approach with no experimental validation for its results, voyAGEr is meant to
323 be a discovery tool, supporting biologists in the exploration of a large transcriptomic dataset,
324 thereby generating, refining or preliminary testing hypotheses, laying the groundwork for
325 subsequent experimental research. It can be an entry point for projects aiming at better
326 understanding the tissue- and sex-specific transcriptional alterations underlying human
327 ageing, to be followed by targeted studies focusing on the functional roles of the most
328 promising markers identified therein in the physiology of ageing. Those marker genes can
329 contribute to the development of more robust and cross-tissue gene signatures of ageing [62]
330 and the expansion of age-related gene databases [63,64].
331 Moreover, the observed diverse and asynchronous changes in gene expression between
332 tissues over the human adult life provides potentially relevant information for the design of
333 accurate diagnostic tools and personalised therapies. On one hand, identifying those
334 changes' association with specific disorders could have a prognostic value by enabling the
335 identification of their onset before clinical symptoms manifest [36]. On the other hand,
336 computational screening of databases of genetic and pharmacologically induced human
337 transcriptomic changes could help to infer their molecular causes and uncover candidate
338 drugs to delay these effects [60,61,65,66].

Methods

Development platforms

Data analysis was performed in R (version 4.1.2) and the application developed with R package `Shiny` [67]. voyAGEr's outputs are plots and tables, generated with R packages `highcharter` [68] and `DT` [69], respectively, that can easily be downloaded in standard formats (png, jpeg, and pdf for the plots; xls and csv for the tables).

voyAGEr was deployed using Docker Compose and ShinyProxy 2.6.1 in a Linux virtual machine (64GB RAM, 16 CPU threads and 200GB SSD) running in our institutional computing cluster.

Read count data pre-processing

The matrix with the RNA-seq read counts for each gene in each GTEx v8 sample was downloaded from the project's data portal (https://www.gtexportal.org/) [40].

From the 54 tissues available from GTEx v8, five were discarded (kidney medulla, fallopian tube, bladder, ectocervix, endocervix) due to low (<50) numbers of samples.

Read count data for each tissue were then pre-processed separately. We started by filtering out genes deemed uninformative due to their very low expression across samples: only genes with at least 1 CPM in at least 40% of the samples were kept for analysis (the number of genes analysed for each tissue can be found in **Supplementary Table 1**)[70]. Read counts for those kept genes were used to calculate normalisation factors to scale the raw library sizes, using function `calcNormFactors` from `edgeR` [70] that implements the trimmed mean of M-values [71]. Read counts were subsequently normalised and log-transformed with the `voom` function [72] from package `limma` [73].

However, it is well-established that batch effects, which may stem from variations in sample treatment prior to RNA-seq library preparation, can introduce spurious gene expression differences between samples and result in confounding factors[74]. We therefore conducted an impartial and systematic search for potential batch effects. Firstly, we performed a principal component analysis of gene expression for each tissue, using the `prcomp` package.

367 We quantified the relation between each condition associated with every sample[1] and the
368 first two principal components, by computing Spearman correlations (for numerical
369 conditions), t-tests (for binary categorical conditions), or analysis of variance (ANOVA) tests
370 (for variables with more than two possible values, and, in the case of numerical variables,
371 fewer than 15 unique values). Conditions that surpassed defined empirical thresholds (p-
372 value < 0.05, Spearman correlation > 0.3, t-test > 10, and ANOVA > 20) were flagged as
373 potential batch effects. Except in brain tissues, the COHORT variable (i.e., whether the
374 samples were collected from organ donors or *post mortem*) appeared to be the main batch
375 effect, with ripple effects on numerous other related conditions (**Supplementary Figure 9**).
376 Moreover, SMRIN (sample's RNA integrity number), DTHHRDY (death classification based on
377 the 4-point Hardy Scale), and MHSMKYRS (smoke years) consistently emerged as conditions
378 associated with the primary axes of variance. The number of genes detected in each sample,
379 determined by the filtration step described above, was also identified as a significant
380 contributor to the primary data variance. We therefore corrected for these five conditions,
381 on a tissue-by-tissue basis, by adapting the `removeBatchEffect` function from the
382 `limma` package [73]. Specifically, we employed linear models to estimate the contributions to
383 gene expression of each of those factors and subtracted such contributions from the original
384 logCPM matrix. To ensure biological interpretability of the results, we offset the resulting
385 values to the minimum value in the non-corrected matrix. To prevent sample loss due to
386 missing values for the aforementioned five conditions and since the number of missing
387 values was relatively low compared to the total number of samples, imputation was carried
388 out using the `mice` package[75].
389 The resulting matrix of logCPM-corrected values was used for all downstream analyses. As an
390 illustrative example of the importance of batch removal, the expression of surfactant factor
391 *SFTPA2* was found to be associated with donors on a ventilator[76]. Without batch correction,
392 *SFTPA2* expression would have been associated with age due to higher prevalence of such
393 cases among older individuals (**Supplementary Figure 10**).
394

[1] According to the annotated variables for the dbGaP Study Accession phs000424.v8.p2, listed in https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/GetListOfAllObjects.cgi?study_id=phs000424.v8.p2&object_type=variable

395 ## ShARP-LM

396 To model the changes in gene expression with age, we developed the Shifting Age Range
397 Pipeline for Linear Modelling (ShARP-LM). For each tissue, we fitted linear models to the
398 gene expression of samples from donors with ages within windows with a range of 16 years
399 shifted through consecutive years of age (i.e. in a sliding window with window size = 16 and
400 step size = 1 years of age). This was the minimum age span needed to guarantee the
401 presence of more than one sample per window, across all considered tissues. As samples at
402 the ends of the dataset's age range would be thereby involved in fewer linear models, we
403 made the window size gradually increase from 11 to 16 years when starting from the
404 "youngest" samples and decrease from 16 to 11 years when reaching the "oldest"
405 (**Supplementary Figure 11)**.
406 Function `lm` from `limma` was used to fit the following linear model for gene expression
407 (GE):

408 $$GE \sim E_0 + \alpha.Age + \beta.Sex + \chi.Age\&Sex + \varepsilon$$

409 For each gene, α, β and χ are the coefficients to be estimated for their respective
410 hypothesised effects. For each sample, *Age* in years and *Sex* in binary were centred and
411 *Age&Sex* interaction given by their product. The coefficients E_0 and ε are thus the expression
412 of the average sample (i.e., with average sex and average age) and the error term.
413 For each gene in each model (i.e., each age window in each tissue), we retrieved the t-
414 statistics of differential expression associated with the three relevant variables and their
415 respective p-values. We considered the average age of the samples' donors within the age
416 window as the representative age of the observed expression changes.
417 In summary, for a given tissue and variable (*Age*, *Sex* and *Age&Sex*), ShARP-LM yields t-
418 statistics and p-values over age for all genes, reflecting the magnitude and significance of the
419 changes in their expression throughout adult life.
420

421 ## *Gene*-centric visualisation of tissue-specific expression changes across age

422 For visualisation purposes, the trend of each gene's expression progression over age in each
423 tissue was derived through Local Polynomial Regression Fitting, using R function `loess` on
424 logCPM values (**Figures 1, 2C, 3B c**). For summarizing in a heatmap a given gene's expression
425 across age in multiple tissues (**Figure 2A**) or the expression of several genes across age in a

426 specific tissue, each gene's regression values are centred and scaled, using R function
427 `scale`.
428 For summarizing in a heatmap the significance of a given gene's expression changes over age
429 in multiple tissues (**Figure 2B**), cubic smoothing splines are fitted to $-\log_{10}(p)$, with p being
430 the t-statistic's p-value, with R function `smooth.spline`.
431

432 Tissue-specific quantification of global transcriptomic alterations across age

433 To assess the global transcriptomic impact of each of the three modelled effects in each
434 tissue across age, we analysed the progression over age (i.e. over consecutive linear models)
435 of the percentage of genes whose expression is significantly altered (t-statistic's p-value ≤
436 0.01) by each effect (**Figure 3B**). To evaluate the significance of each percentage and assess
437 if high percentages can be confidently associated with major transcriptomic alterations, we
438 controlled for their false discovery rate (FDR) by randomly permuting the samples' ages and
439 sexes within each age window fifty thousand times and performing ShARP-LM on each
440 randomised dataset. We were then able to associate an FDR to each percentage of altered
441 genes by comparing it with the distribution of those randomly generated (**Figure 3B a**).
442 We also applied a linear model across the entire age range, thereby providing users with
443 more insight and supporting evidence into how a specific gene changes with age. For
444 visualisation purposes, we incorporated a dashed orange line, with the logFC per year for the
445 Age effect as slope, in the respective scatter plots (**Figure 3B c**). We depict the Sex effect
446 therein by prominent dots on the average samples, with pink and blue denoting females and
447 males, respectively.
448

449 Gene Set Enrichment Analysis (GSEA)

450 For each *Peak* of significant gene expression modifications, we performed GSEA [49] on the
451 ordered (from the most positive to the most negative) t-statistics of differential expression
452 for the respective tissue and age, using R package `fgsea` [77] and the `Reactome` database [48].
453 We extracted the GSEA normalised enrichment score (NES), which represents the degree to
454 which a certain gene set is overrepresented at the extreme ends of the ranked list of genes.
455 A positive NES corresponds to the gene set's overrepresentation amongst up-regulated
456 genes within the age window, whereas a negative NES signifies its overrepresentation

457 amongst down-regulated genes. The NES for each pathway was used in subsequent analyses
458 as a metric of its up- or down-regulation in the Peak. The resulting normalised enrichment
459 score (NES) for each pathway was used in subsequent analyses as a metric of its over- or
460 down-representation in the *Peak*.
461 To optimise computational efficiency and minimise redundancy in the analysed pathways,
462 we only considered pathways containing a minimum of 15 genes and up to 500 genes, as
463 suggested in the GSEA User Guide[78]. For the sake of clarity in voyAGEr's visual
464 representations, we only included pathways with a p-adjusted value less than or equal to
465 0.05, further narrowing it down to pathways within the top 1% of p-adjusted values.
466 Additionally, we exclusively featured pathways with at least one significant age Peak (FDR <=
467 0.05), as illustrated in **Figure 4A**.
468

469 ## Families of pathways

470 To reduce pathway redundancy and facilitate the assessment of their biological relevance in
471 results' interpretation, we created an unifying representation of pathways from Reactome
472 [48] ,KEGG [50], and level 3 Biological Processes from GO [51], by adapting a published pathway
473 clustering approach [79] to integrate them into families.
474 The approach relies on the definition of a hierarchy of pathways based on the numbers of
475 genes they have in common. For each two pathways P_i and P_j, respectively containing sets of
476 genes G_i and G_j, we computed their overlap index (OI) [79], defined as follows:
477 $$OI_{ij} = |G_i \cap G_j| / \min(|G_i|, |G_j|)$$
478 Where $|G_i|$ is the number of genes in set G_i and $|G_i \cap G_j|$ is the number of genes in common
479 between G_i and G_j. $OI_{ij} = 1$ would therefore indicate that P_i and P_j are identical in gene
480 composition or that one is a subset of the other. On the contrary, $OI_{ij} = 0$ would mean that P_i
481 and P_j have no genes in common. To ease the computational work, we removed from the
482 analysis pathways that are subsets of larger pathways (i.e. each pathway whose genes are all
483 present in another pathway).
484 From the OI_{ij} matrix, from which each row is a vector with the gene overlaps of pathway *i*
485 with each of the pathways, we computed the matrix of Pearson's correlation between all
486 pathways' overlap indexes with R function `cor`. That matrix was finally transformed into
487 Euclidean distances with R function `dist`, allowing for pathways to be subsequently

488 clustered with the complete linkage method with R function `hclust`. The final dendrogram
489 was empirically cut in 10 clusters (**Supplementary Figure 12**). Pathways that were initially
490 excluded from the computation for being subsets of others were added to the clusters of
491 their respective *parent* pathways. Each *daughter* pathway with more than one parent was
492 assigned to the cluster of the parent with the smallest number of genes, thereby maximizing
493 the daughter-parent similarity. The `data.table` R package, for fast handling of large
494 matrices, was used in this analysis [80].
495

496 ## Gene co-expression modules

497 Gene co-expression modules were defined with R package `WGCNA`[52]. For each considered
498 tissue, the process began with the identification of a set of informative genes that exhibit
499 high variability across samples (referred to as variable A), thus improving module definition.
500 Next, we calculated the bicorrelation matrix for the expression of all selected genes with the
501 `bicor` function. We then applied soft thresholding by raising all correlation values to the
502 power of β, accentuating stronger correlations. The value of β = 12 was chosen in
503 accordance to the WGCNA FAQs[81], and after confirmation of a free-scale topology using the
504 `pickSoftThreshold` function. We generated the dissimilarity matrix by subtracting the
505 output of the `TOMsimilarity` function from 1. Gene co-expression modules were then
506 defined using a static tree-cutting algorithm, implemented via the `cutreeStaticColor`
507 function, requiring as parameters a minimum number of genes per module (referred to as
508 variable B) and the tree-cutting height (variable C).
509 The empirical determination of parameters A to C was guided by the following principles: (i)
510 maximizing cell type signature enrichment, (ii) minimizing the number of genes per module,
511 and (iii) ensuring that the modules' eigengenes exhibited age-related variability. Different
512 combinations of these variables were exhaustively tested. To maintain biological relevance,
513 modules consisting of non-assigned genes or genes lacking substantial supporting evidence,
514 such as pseudogenes, were excluded.
515 Maximizing cell type enrichment in the modules, with a focus on known markers for specific
516 cell types, has previously been proven successful in unveiling core transcriptional features of
517 cell types in the human central nervous system [53]. Cell type enrichment analysis relied on
518 Fisher tests, providing odds enrichment scores and significance values (p-values). This
519 involved comparing module genes with the signature, considering as background of genes

520 those included in the module definition for each tissue. We prioritised modules of genes for
521 which we obtained at least one significant result for each cell type (odds ratio (OR) > 1 and
522 p-value < 0.05). The cell type signatures employed in this analysis were sourced from
523 MSigDB's C8 collection[49], and in the case of Whole Blood, additionally from LM22[82].
524 Specific variance thresholds (variable A) were employed: 0.5, 0.4, 0.35, and 0.9 for Brain –
525 Cortex, Muscle – Skeletal, Heart – Left Ventricle, and Whole Blood, respectively. The
526 minimum number of genes per module (variable B) was set at 15, 20, 20, and 15,
527 respectively. Tree-cutting heights (variable C) of 0.95, 0.98, 0.99, and 0.97 were respectively
528 used.
529 Each module is characterised by a set of genes and an eigengene, represented by the first
530 principal component obtained through singular-value decomposition of the module's gene
531 expressions. Subsequently, voyAGEr facilitates an evaluation of cell type enrichment, as
532 described earlier, and enrichment in biological pathways and diseases. The enrichment of
533 modules in cell types, Reactome pathways, and diseases (extracted from DisGeNET
534 [55,56]database version 7.0) was quantified using Fisher's tests. For disease enrichment, the
535 function `disease_enrichment` from the `disgenet2r` package[55] was employed,
536 utilising their curated set of diseases. The significance of these enrichments was determined
537 through p-value/FDR adjustment using Benjamini-Hochberg correction. For visual clarity,
538 only pathways and diseases displaying significant enrichment (p <= 0.05) in at least one
539 module were considered.
540 The four tissues (Brain - Cortex, Muscle - Skeletal, Heart - Left Ventricle, and Whole Blood)
541 covered by the Module section of voyAGEr were selected due to their relatively high sample
542 sizes and availability of comprehensive cell type signatures. The increasing availability of
543 human tissue scRNA-seq datasets (e.g., through the Human Cell Atlas[83]) will allow future
544 updates of voyAGEr to encompass a wider range of tissues.
545

546 Data and Code availability

547 Processed GTEx v8 RNA-seq data (read count tables) were downloaded from the project's
548 data portal (https://www.gtexportal.org/). Donor metadata were obtained from
549 dbGaP - database of Genotypes and Phenotypes (Accession phs000424.v9.p2 project ID
550 13661). voyAGEr's output tables can be directly downloaded in standard xls and csv formats.

The complete source code for voyAGEr, including pre-processing and Shiny app, can be accessed on GitHub at the following repository:
https://github.com/DiseaseTranscriptomicsLab/voyAGEr.

Acknowledgements

We thank iMM colleagues Joana Neves and Luísa Lopes, as well as all members from the Disease Transcriptomics Lab, for providing valuable feedback on the manuscript. We also thank the very knowledgeable anonymous reviewers designated by *eLife* for their constructive and insightful suggestions and criticisms to the first version of this manuscript; their feedback greatly contributed to much improved versions of both the article and voyAGEr itself.

Funding: This work was supported by European Molecular Biology Organization (EMBO Installation Grant 3057), FCT - Fundação para a Ciência e a Tecnologia I.P. (FCT Investigator Starting Grant IF/00595/2014, CEEC Individual Assistant Researcher contract CEECIND/00436/2018, PhD Studentships SFRH/BD/131312/2017, UI/BD/153368/2022 and UI/BD/153363/2022, projects LA/P/0082/2020 and UIDP/50005/2020), European Union (EU) Horizon 2020 Research and Innovation Programme (RiboMed 857119), and GenomePT project (POCI-01-0145-FEDER-022184), supported by COMPETE 2020 – Operational Programme for Competitiveness and Internationalization (POCI), Lisboa Portugal Regional Operational Programme (Lisboa2020), Algarve Portugal Regional Operational Programme (CRESC Algarve2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF), and by FCT - Fundação para a Ciência e a Tecnologia.

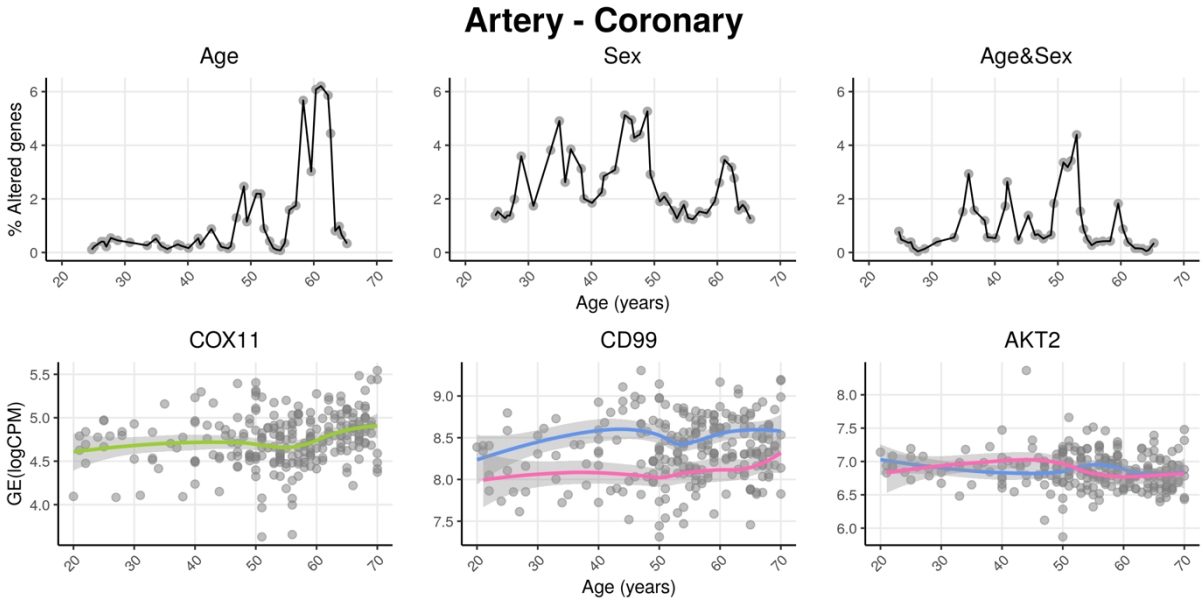
Authors' contributions

A.S. and N.L.B.-M. designed the study, with contributions from R.M.-S. and A.K.; A.S., R.M.-S, and A.K. performed all the development and bioinformatics work, under the supervision of N.L.B.-M.; N.S.-A. set up the web and database implementation; A.S., R.M.-S., A.K., and N.L.B.-M. wrote the manuscript, with contributions from N.S.-A..

Conflict of Interest statement

The authors declare no conflict of interest.

583 ## Figures and Tables



584 **Figure 1**: *voyAGEr profiles tissue-specific age-associated changes in gene expression and*
585 *their differences between sexes.*
586 For each of the 49 human tissues in GTEx, gene expression was linearly modelled in windows
587 spanning 16 years centred in consecutive years of age, to estimate the effects thereon of
588 *Age*, *Sex* and the interaction between them, i.e., how the *Sex* effect changes with age,
589 equivalent to how the *Age* effect differs between sexes (v. Materials and Methods). In each
590 age window, the percentage of genes with expression significantly altered by of each of
591 those effects gives their respective transcriptomic impact (upper panels). voyAGEr thereby
592 identifies the age periods at which major gene expression changes occur in each tissue. For
593 example, in coronary artery: major age-related transcriptional alterations are found at
594 around 60 years of age (upper left panel), illustrated by the behaviour of *COX11* (bottom left
595 panel); major gene expression differences between males and females happen across the
596 considered age range (upper centre panel), as illustrated by *CD99* (bottom centre panel);
597 major differences between sexes in age-related gene expression alterations happen across
598 the considered age range (upper right panel), as illustrated by *AKT2* (bottom right panel).
599 Solid loess lines in the bottom panels (green for all donors, pink for females, blue for males).
600 Gene expression (GE) in $\log_2$ of counts per million (logCPM).

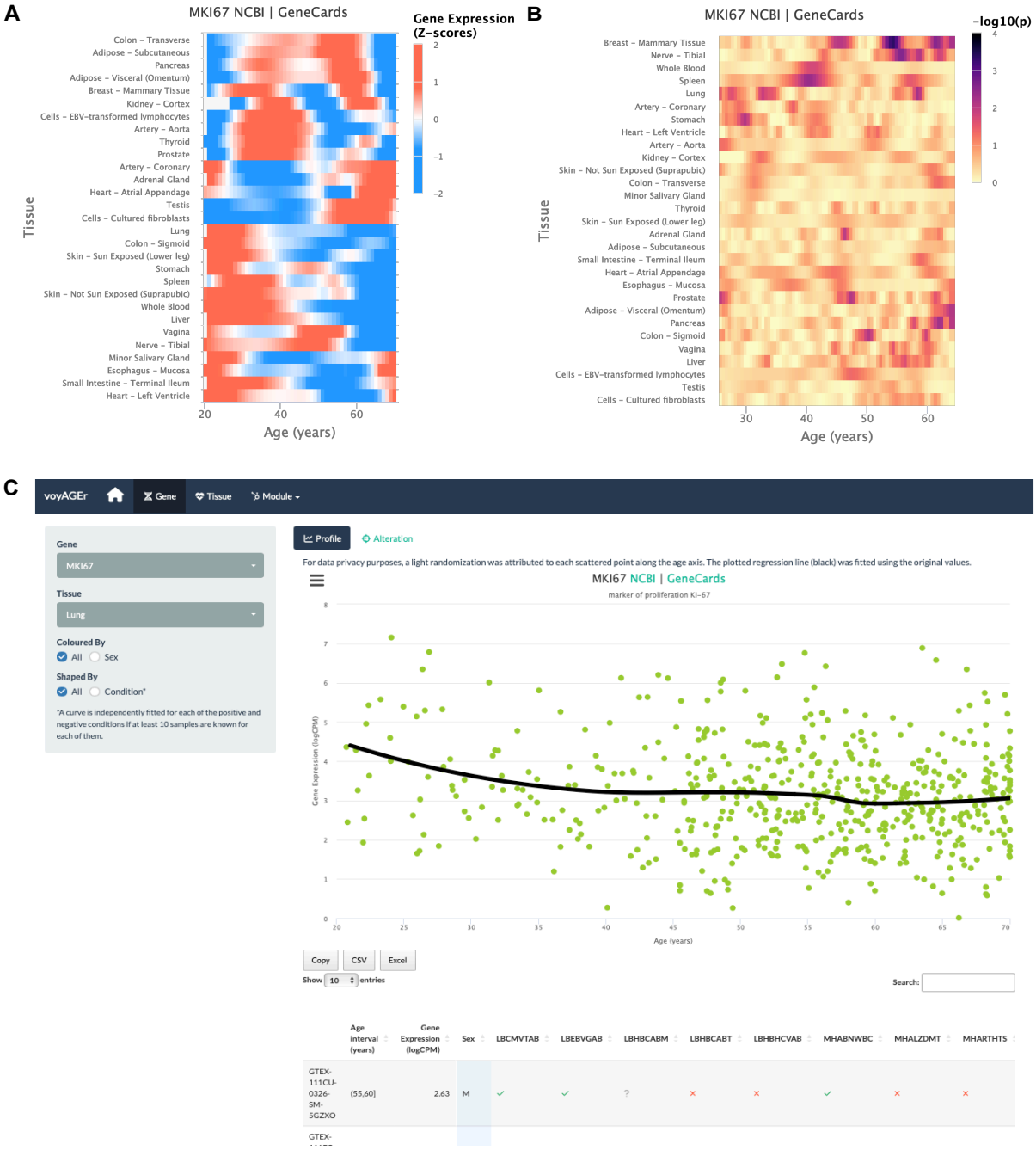


Figure 2: *Gene-centric analyses of expression alterations across age*

A - Heatmap of *MKI67* expression across tissues over age. **B** - Heatmap of significance of *Age*-associated *MKI67* expression alterations over age. p-values are for the moderated t-statistics of differential gene expression associated with the Age effect (v. the ShARP-LM approach in Materials and Methods). Notably, transcriptional changes are observed in lung (mid 20's, early 30's and after 55). **C** - voyAGEr's *Gene* tab interface. *MKI67* expression in lung is inspected. Donors' information is shown in a table and the scatter plot can be interactively adjusted according to the donors' condition of interest **(Supplementary Figure 1B)**.

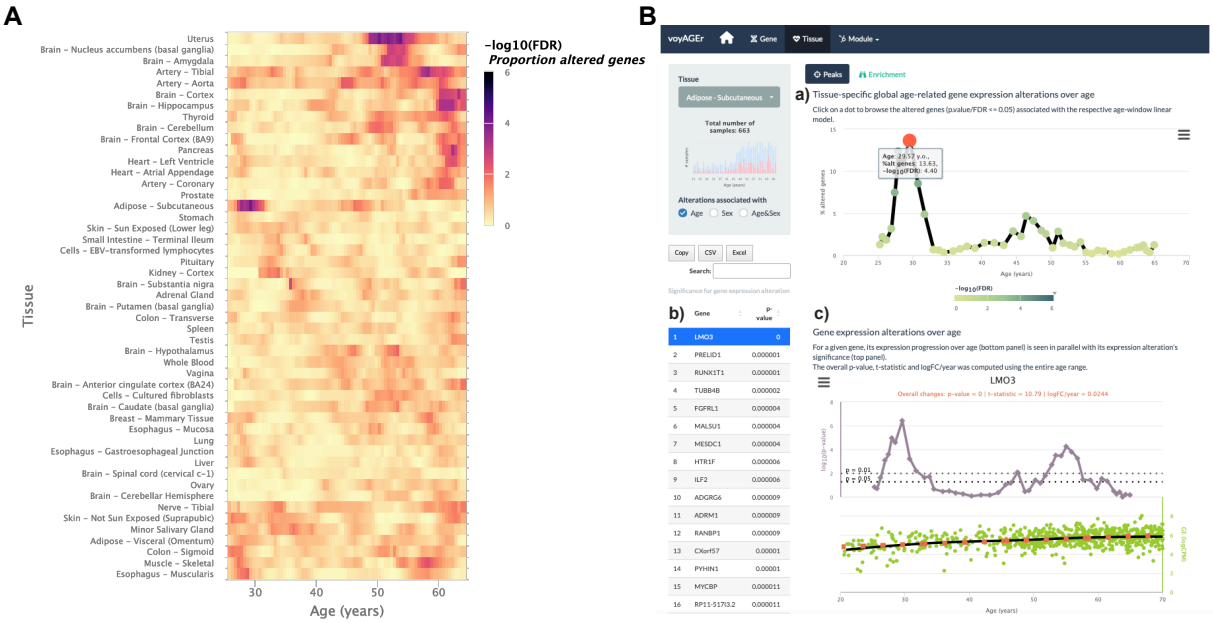


Figure 3: *Tissue-specific assessment of gene expression changes across age*

A - Heatmap of significance (FDR based on random permutations of age, v. Materials and Methods) over age of the proportion of genes with expression significantly altered with *Age* in the 49 analysed tissues. **B** - Exploration of gene expression changes across age in Subcutaneous Adipose tissue: (a) Percentage of genes with significantly altered expression with *Age* over age. Two main peaks of transcriptional changes are noteworthy, a major one at the late 20's and a minor one after 45; (b) Clicking on the dot of a specific age (29.57 y.o. in plot *a*) gives access to the list of the most altered genes at that age, ordered by statistical significance of expression changes (p-value of moderated t-statistic). (c) Plot of expression of the chosen top gene in the table in *b* across age (bottom) in parallel with the significance of its expression alterations with *Age*. The expression of *LMO3* significantly increases at around 30 years old, concomitantly with the first peak of transcriptomic changes with *Age*.

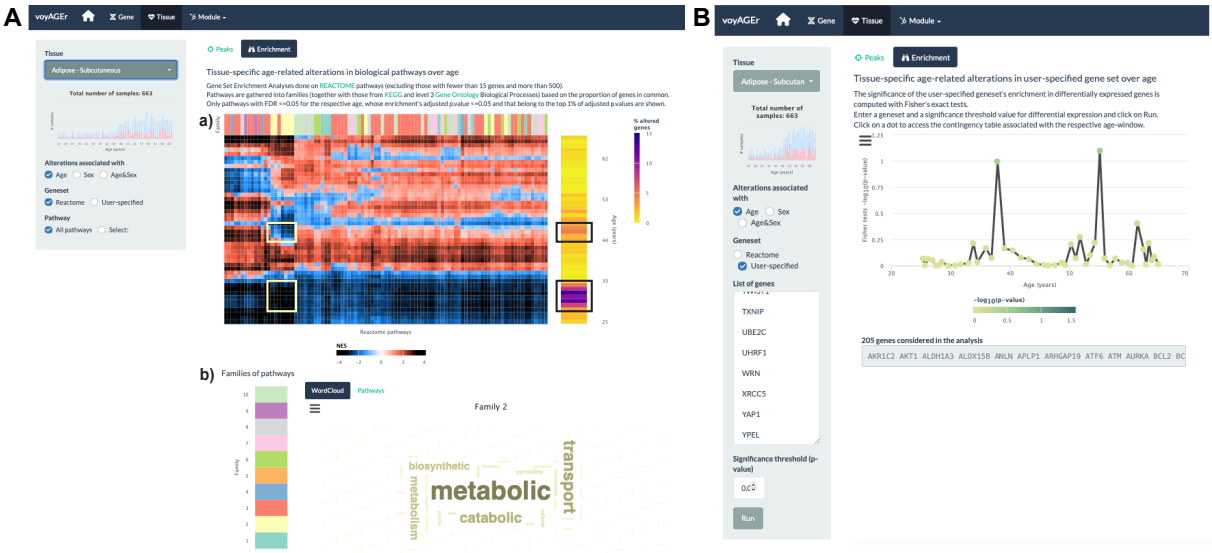


Figure 4: *Tissue-specific assessment of pathway expression changes across age in the human Subcutaneous Adipose tissue*

A - Heatmap depicting the normalised enrichment scores (NES) of Reactome pathways associated with specific tissues and effects. Pathways are classified into 10 families (a), which can be characterised by their frequently occurring terms (b), providing insights into their biological functions. Only pathways significantly linked to gene expression changes in at least one age window (FDR ≤ 0.01) are displayed. Black squares indicate the two age periods with prominent transcriptional changes, while yellow squares denote pathways common to both peaks, primarily belonging to family 2. Word cloud analysis (b) reveals that family 2 pathways are mainly related to metabolism. **B** - Enrichment of a user-provided gene set, given by the significance of Fisher's tests, in genes altered with *Age* throughout ageing (based on a user-defined p-value threshold). Here, the given gene set is composed of genes from Senequest [42] whose link with senescence is supported by at least 4 sources. In this case, there are no significant peaks.

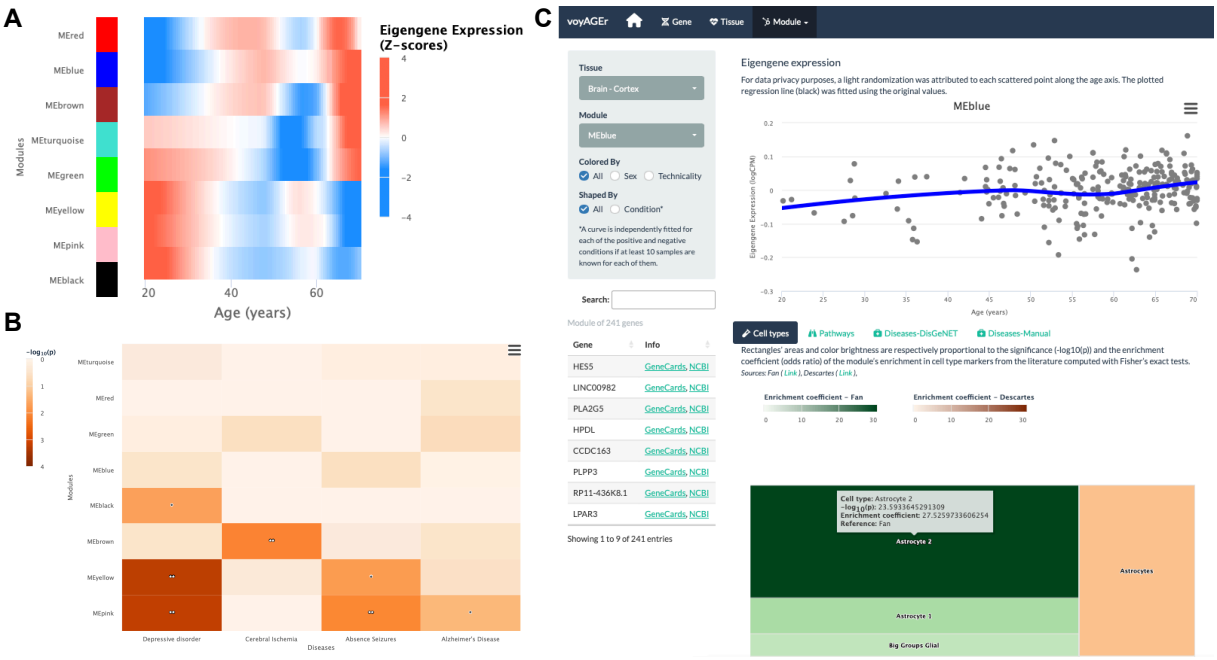
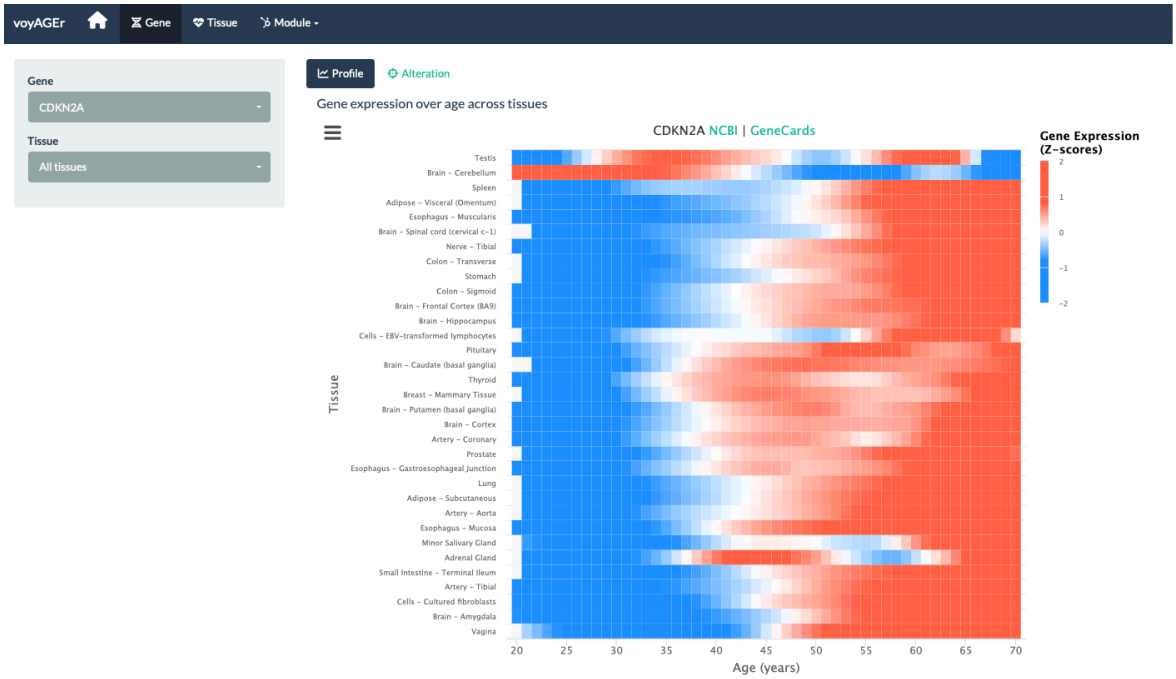


Figure 5: *Tissue-specific assessment of age-associated progression of modules of co-expressed genes*

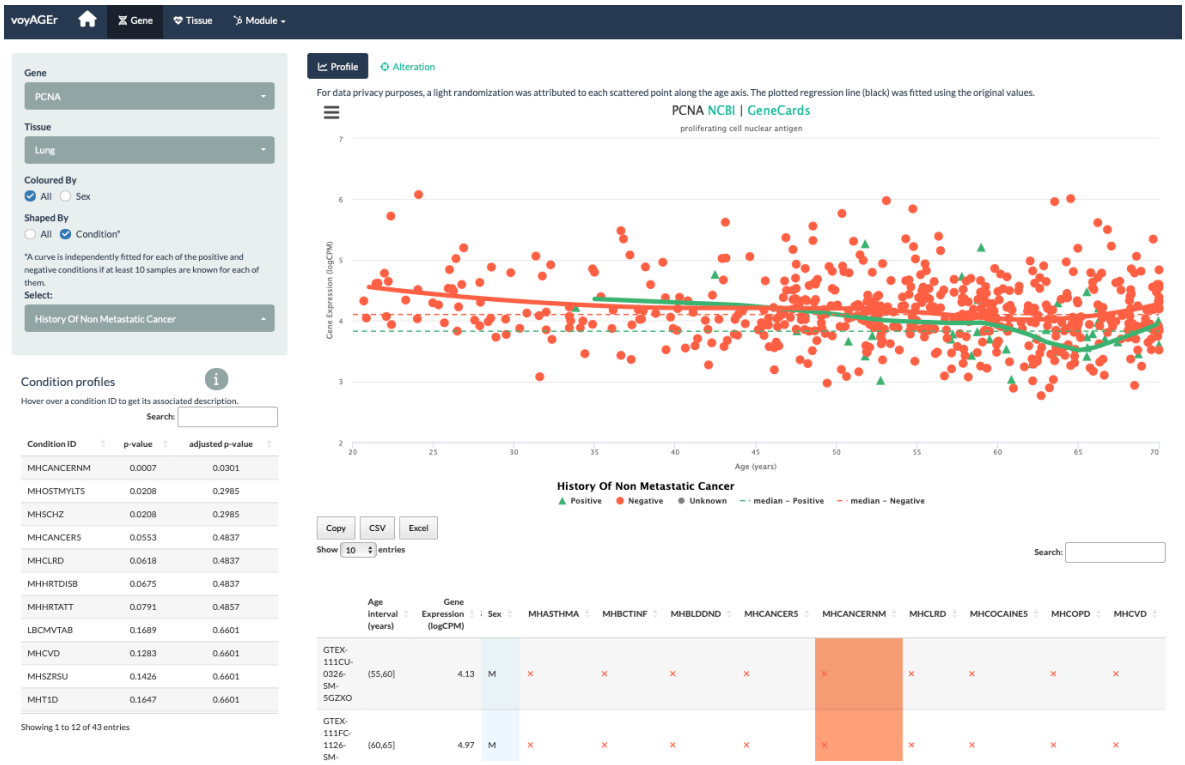
A - Heatmap of eigengene expression for all the modules of co-expressed genes in the brain cortex over age. **B** - Heatmap of association of the modules with four selected diseases, computed with the disgenet2r package[55]. **C** - Scatter plot (above) of eigengene expression over age, in all brain cortex samples, for a selected module of 241 genes (MEblue). The eigengene expression is derived from the 1st component of the single value decomposition of its genes' expression. This module was analysed, with Fisher's tests, for enrichment in cell types, based on markers from the literature (Fan[84] and Descartes[85]) and found to be associated with astrocytes, as can be observed by the TreeMap below (where each rectangle's area and darkness are proportional to the significance of its association with a cell type and its colour linked to the markers' source study).

Supplementary Figures and Tables

A

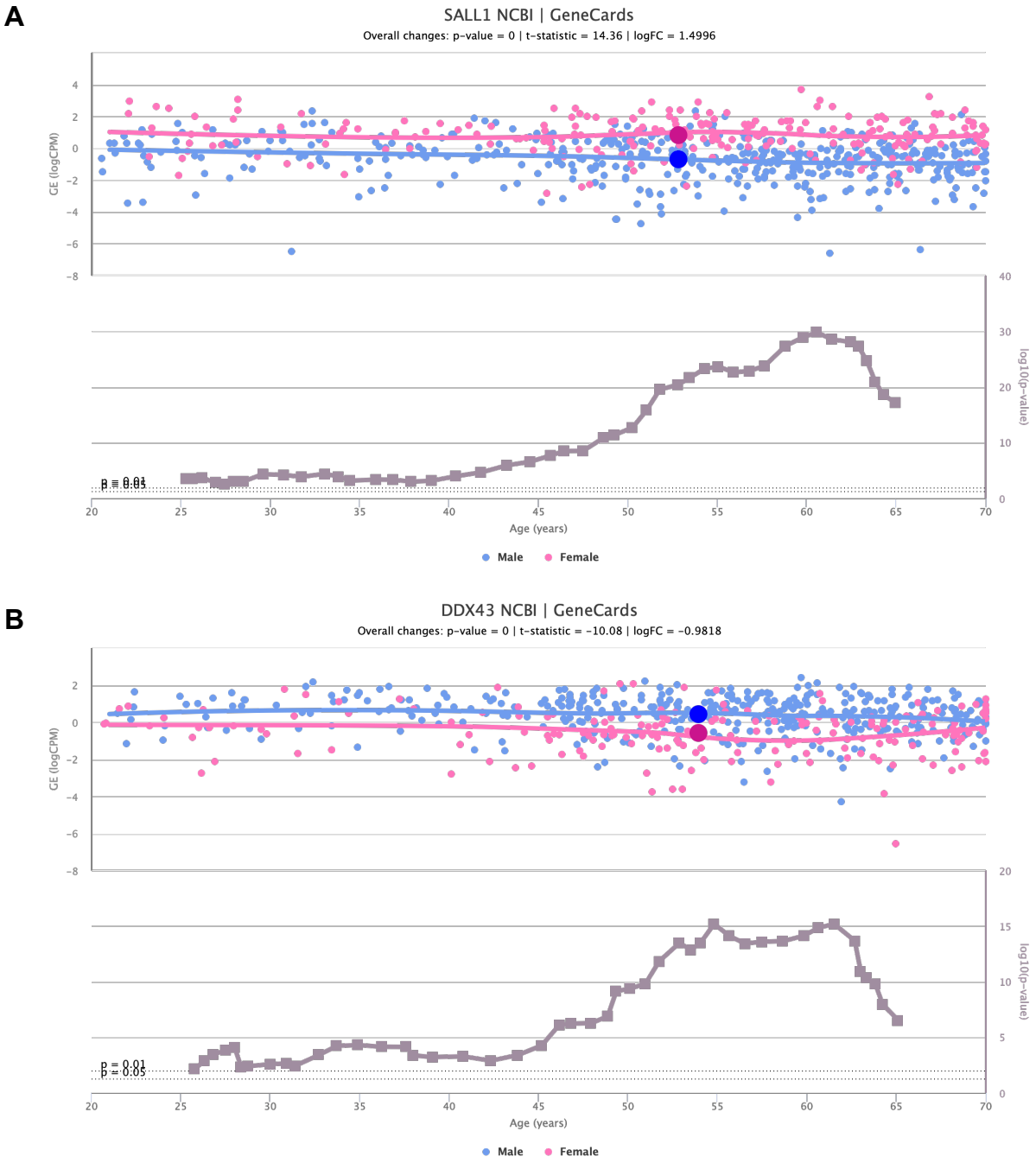


B

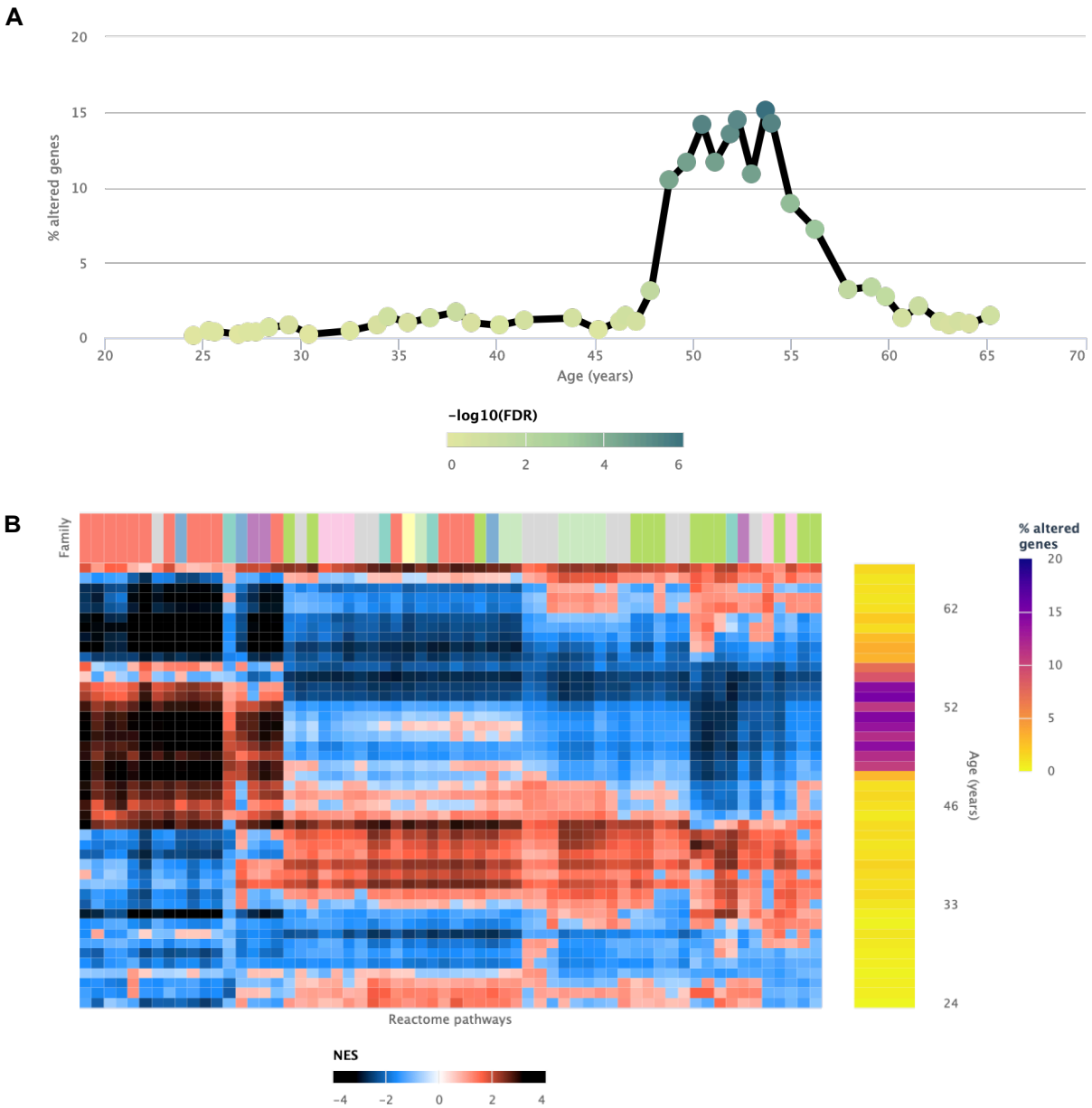


Supplementary Figure 1: *voyAGEr's Gene tab interface*.

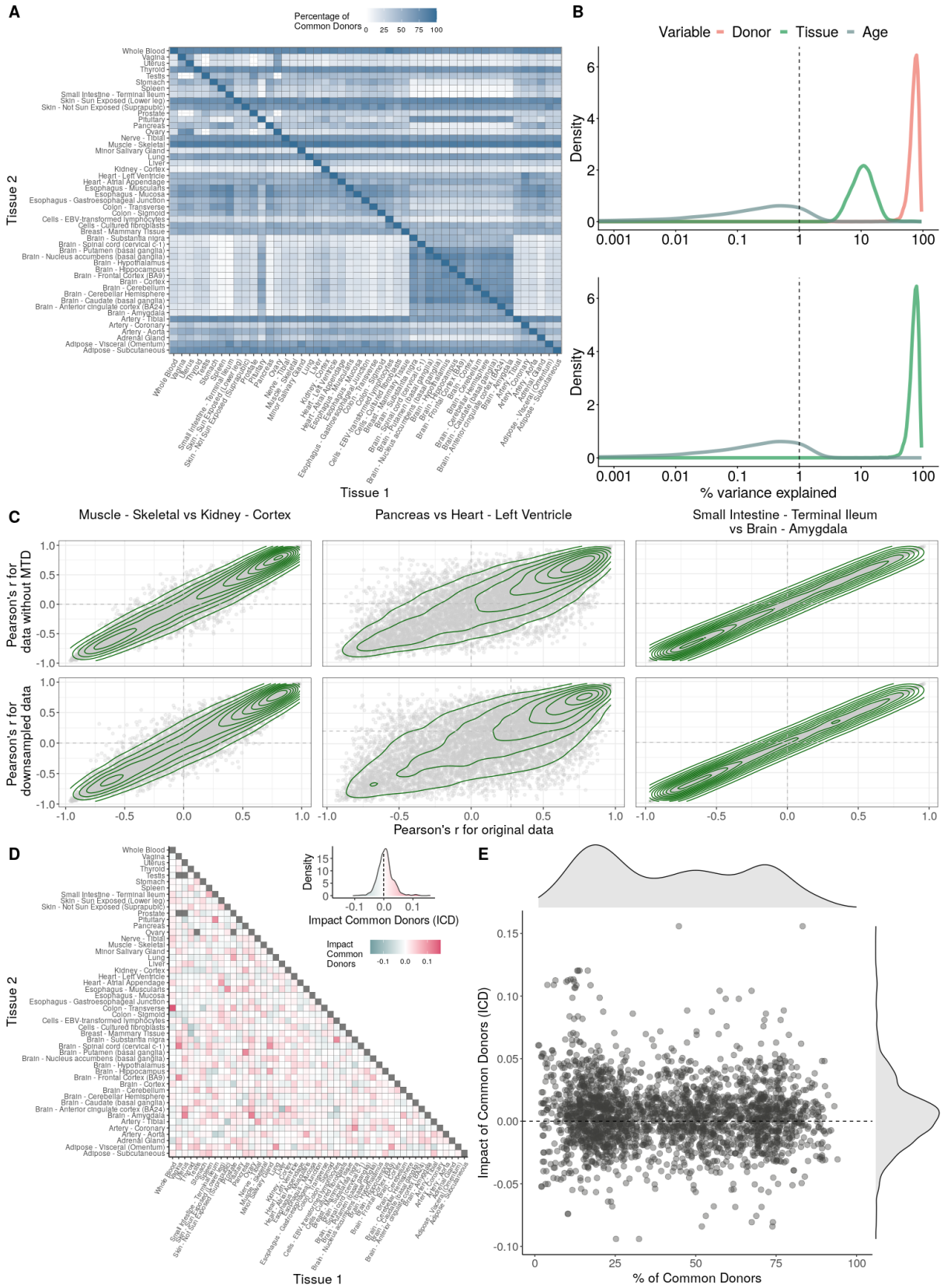
A - Heatmap of *CDKN2A* expression across tissues over age. **B** - *PCNA* expression in lung over age and its association with the donor's history of non-metastatic cancer.



Supplementary Figure 2: *Sex-specific SALL1 and DDX43 expression alterations over age.* Scatter plots of *SALL1* (**A**) and *DDX43* (**B**) expression over age (above), in adipose tissue and lung, respectively, coloured based on the donor's biological sex (male in blue and female in pink). The scatter plots of gene expression are shown together with line plots of the significance of their difference between sexes (below).



658 **Supplementary Figure 3**: *Transcriptomic alterations in the uterus coincident with the onset of*
659 *menopause.*
660 voyAGEr enables the identification of a prominent age period of transcriptional alterations in
661 the uterus, which coincides with the onset of menopause[47] The genes exhibiting significant
662 changes (**A**) reflect the transformations observed in biological processes (**B**), marked by a
663 distinct switch in enrichment scores. Positively enriched pathways become negative, and
664 vice-versa during this age frame.



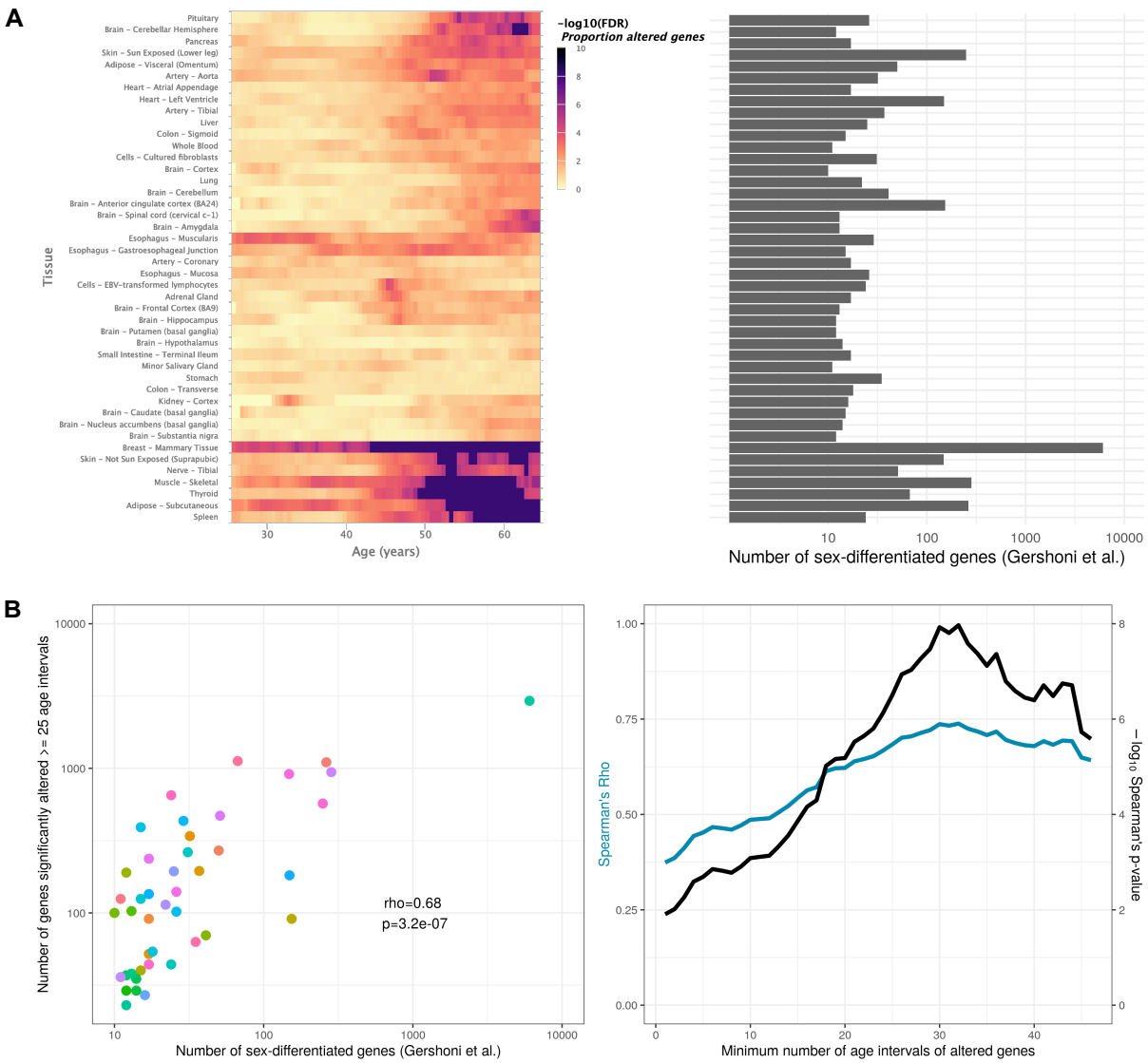
665 **Supplementary Figure 4**: *Impact of donor overlap between tissues in age-associated trends*

666 **A** - Heatmap depicting the percentage of common donors between pairs of tissues. A given

667 square illustrates the percentage of all samples of tissue in the x axis (Tissue *1*) that is in

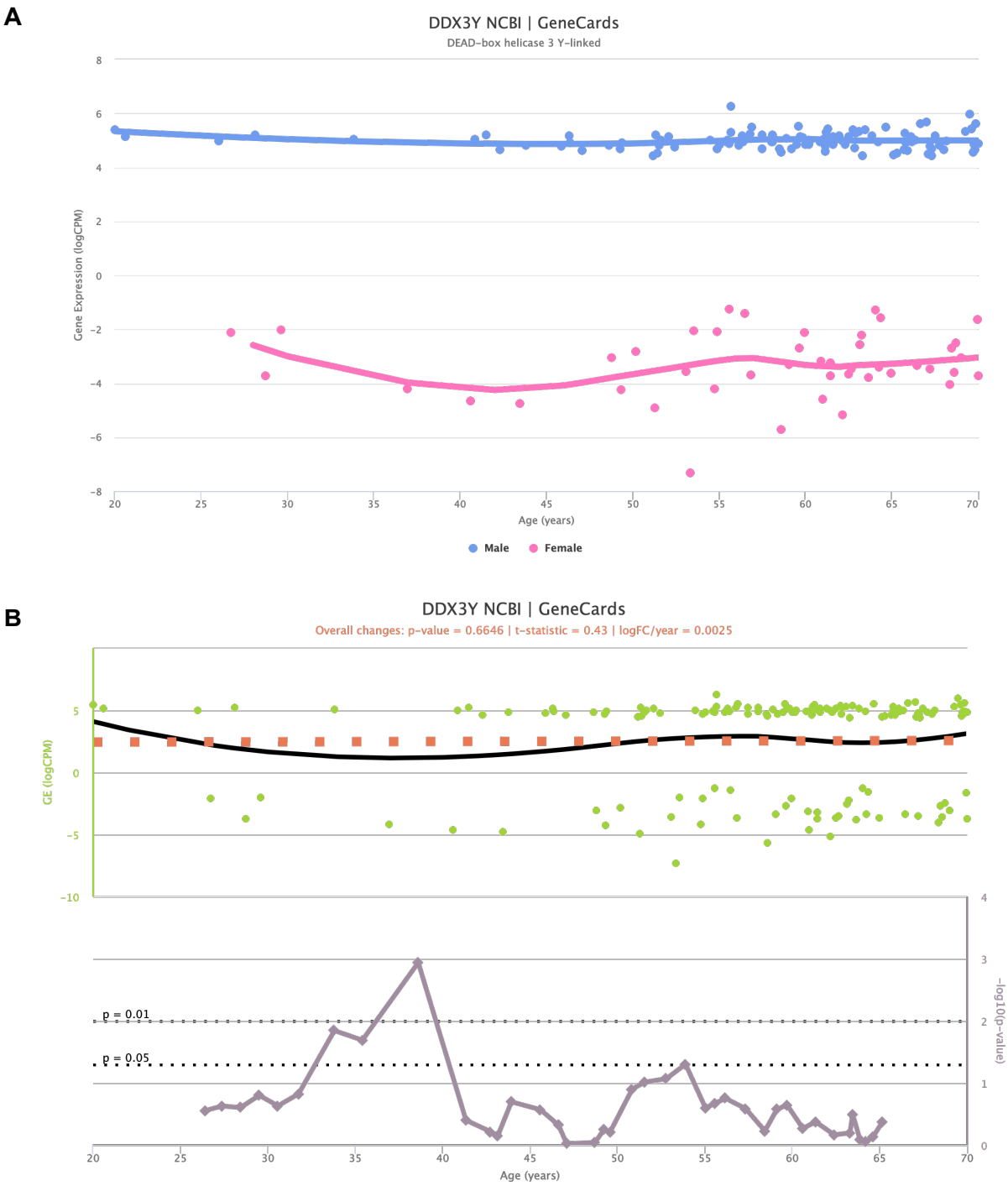
common with the tissue in the y axis (Tissue *2*). B - Assessment of the relative contributions of different sources to the dataset's variance. (Left panel) Tissue accounts for approximately 90% of the total variance, while donor contributes around 10%; age has a minimal impact (1%), likely due to the relative subtlety of its effects on gene expression and to the tissue specificity of ageing dynamics. (Right panel) Removal of the donor variable does not transfer variance to age, suggesting limited confounding between the two variables. **C** - Impact of the relative proportion of common donors on gene expression correlation between tissue pairs. Panels A, B, and C showcase the tissue pairs with the highest (Muscle Skeletal / Kidney Cortex), median (Pancreas / Heart Left Ventricle), and lowest (Small Intestine / Brain Amygdala) percentages of common donors, respectively. The left panels illustrate gene-by-gene Pearson's correlations of gene expression between the two tissues, comparing the scenarios with (x-axis) and without (y-axis) the removal of common donors. The right panels depict the same comparisons, but with random downsampling (y-axis) in both tissues based on the proportions defined for common donor removal. The depicted examples show that the outcomes are comparable when removing common donors or employing random downsampling. **D** - Comparison of the impacts of removing common donor samples and random downsampling across tissue pairs. The process of removing common donors involved the identification and removal of samples from shared donors while maintaining the original sample size imbalance between tissues. As this process inherently involves downsampling, which may impact results, we performed additional downsampling by randomly removing samples from both tissues according to the proportions defined for the removal of common donors. The heatmap is coloured based on whether the removal of common donors has a greater (red) or lesser impact (blue) than random downsampling. The values depicted in the heatmap, denoted as the Impact of Common Donors (ICD), are computed for each tissue pair. This calculation involves several steps: first, by determining the absolute difference in Pearson's correlation for each gene's mean expression within each age window from the ShARP-LM pipeline, between the original data and the subset of data without common donors (DiffWoCD) or with random downsampling (DiffRD). Subsequently, the medians of DiffWoCD and DiffRD are computed, and the difference between these median values provides the ICD for each tissue pair. Due to the unidirectional nature of correlation (i.e., the results for tissue 1 vs tissue 2 mirror those for tissue 2 vs tissue 1), the resulting matrix is triangular in form. Grey tiles denote NA values, i.e., where

700 the tissue-tissue comparison does not have a meaning, namely self-self and between sex-
701 specific tissues. Top right insert: density line ("smoothed" histogram) of all ICD values. **E** -
702 Scatter plot relating the Impact of Common Donors (ICD, values in heatmap D) with the
703 respective percentage of common donors (values in heatmap A).



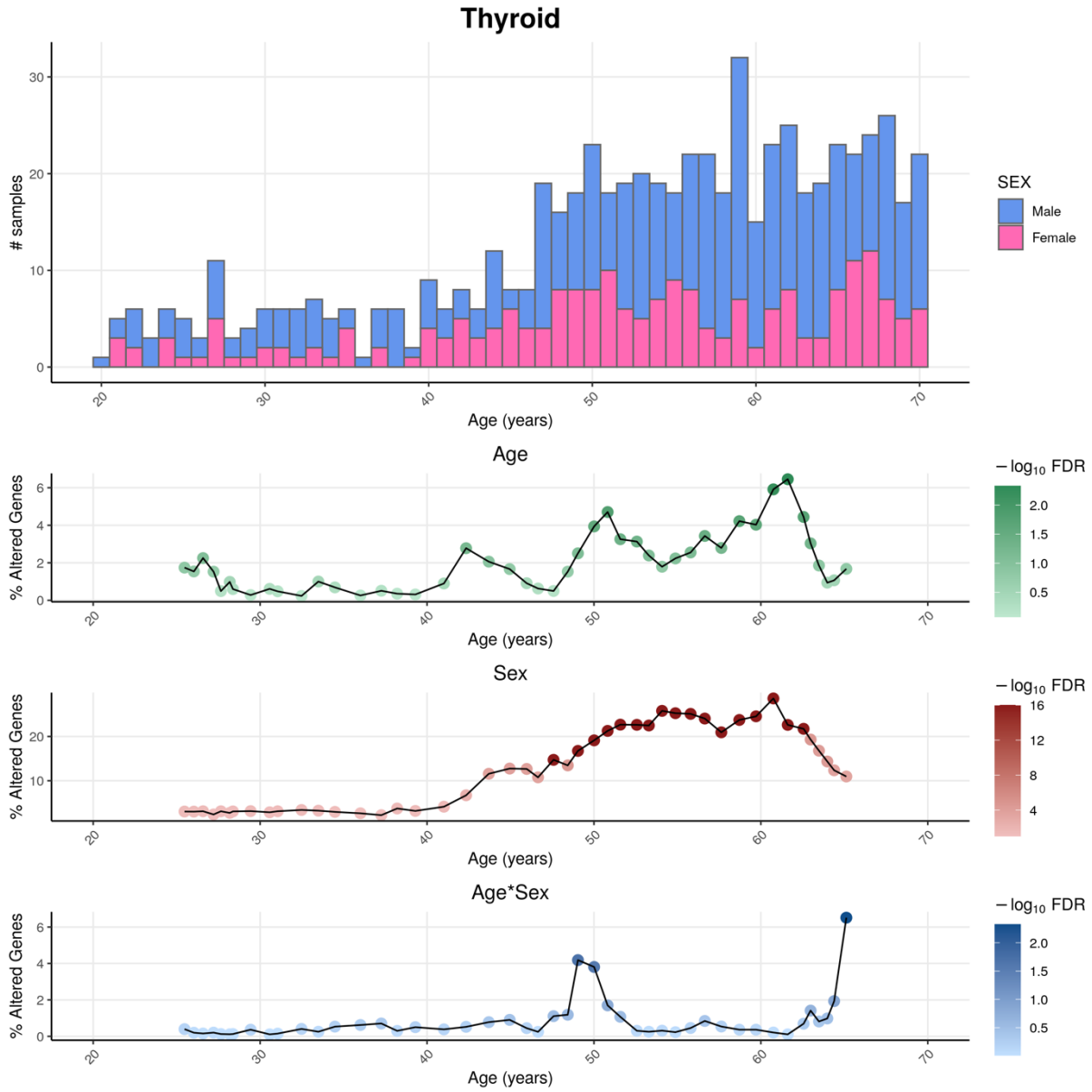
Supplementary Figure 5: *Tissue-specific sex-differentiated expression.*

A - Heatmap of significance over age, for each tissue, of the proportion of altered genes between sexes (left), in parallel with the number of sex-differentiated genes highlighted by Gershoni et al. (v. Figure 1 in [9]) (right). Mammary is corroborated as the tissue with the most sex-differentiated gene expression. For tissues like Adipose - Subcutaneous, Skin - Not Sun Exposed, Thyroid, Nerve - Tibial, Muscle - Skeletal, and Spleen, sex-related biases in expression mainly arise in late life. **B** - Scatter plot relating, for each tissue (each dot), the number of sex-differentiated genes found by Gershoni et al. with the number of genes found by voyAGEr to be altered (p-value <=0.05) between sexes in at least 25 age intervals (left). This correlation is significant (p-value <=0.05) regardless of the considered minimum number of age intervals a gene must be altered in (right).

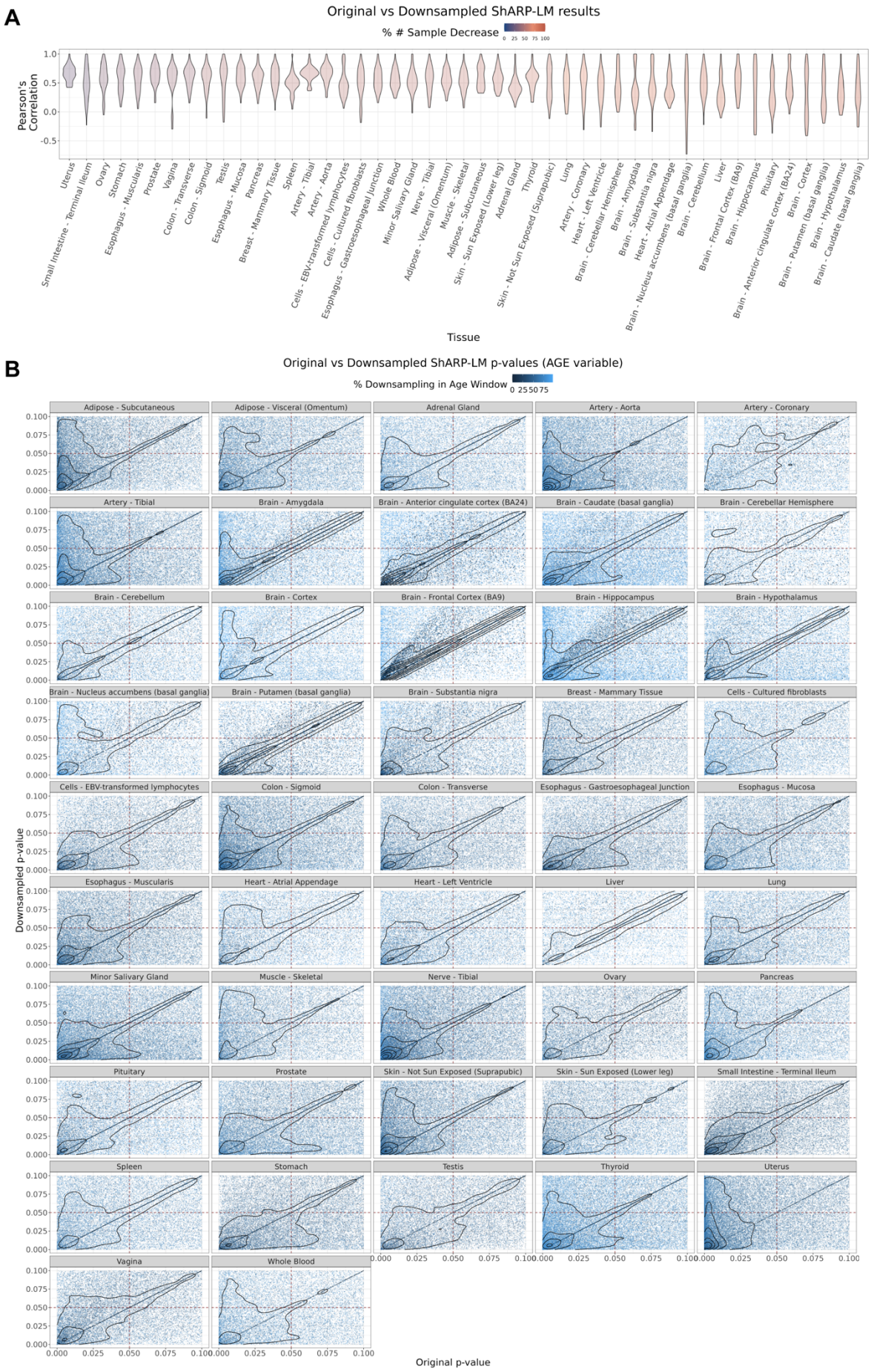


Supplementary Figure 6: *Impact of sex-specific genes in the interpretation of voyAGEr results.*

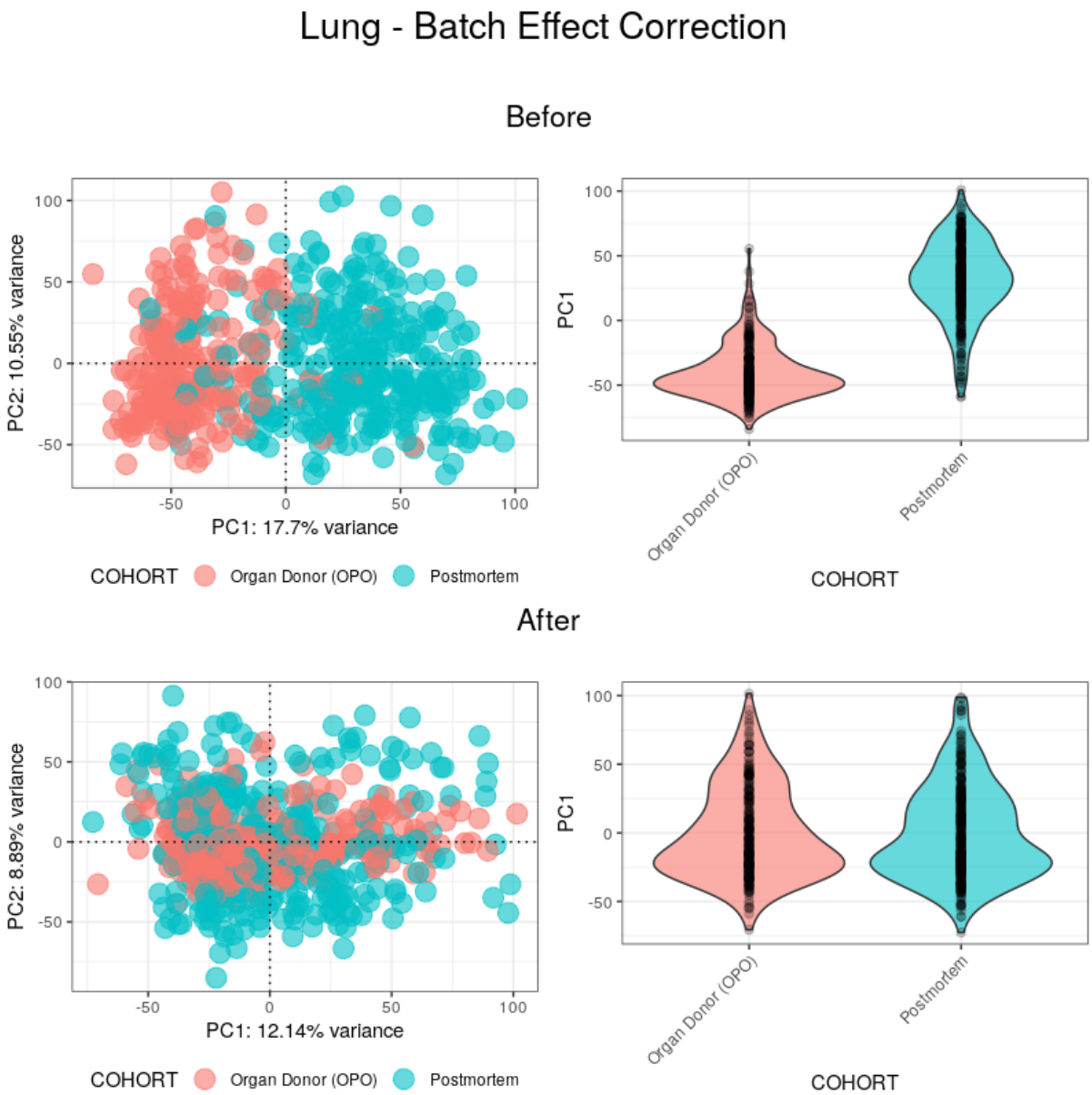
Representation of the expression of Y-chromosome-specific gene *DDX3Y*, stratified by sex (**A**), along with the corresponding significance indicating age-related alterations (**B**). Despite the non-zero expression values for female samples, they fall within a range deemed of background noise (-2 to -8 logCPM). Notably, when assessing the significance of age-related changes in the gene, a distinct peak emerges around the age of 40. This peak is spurious, as it primarily mirrors changes of gene expression in females.



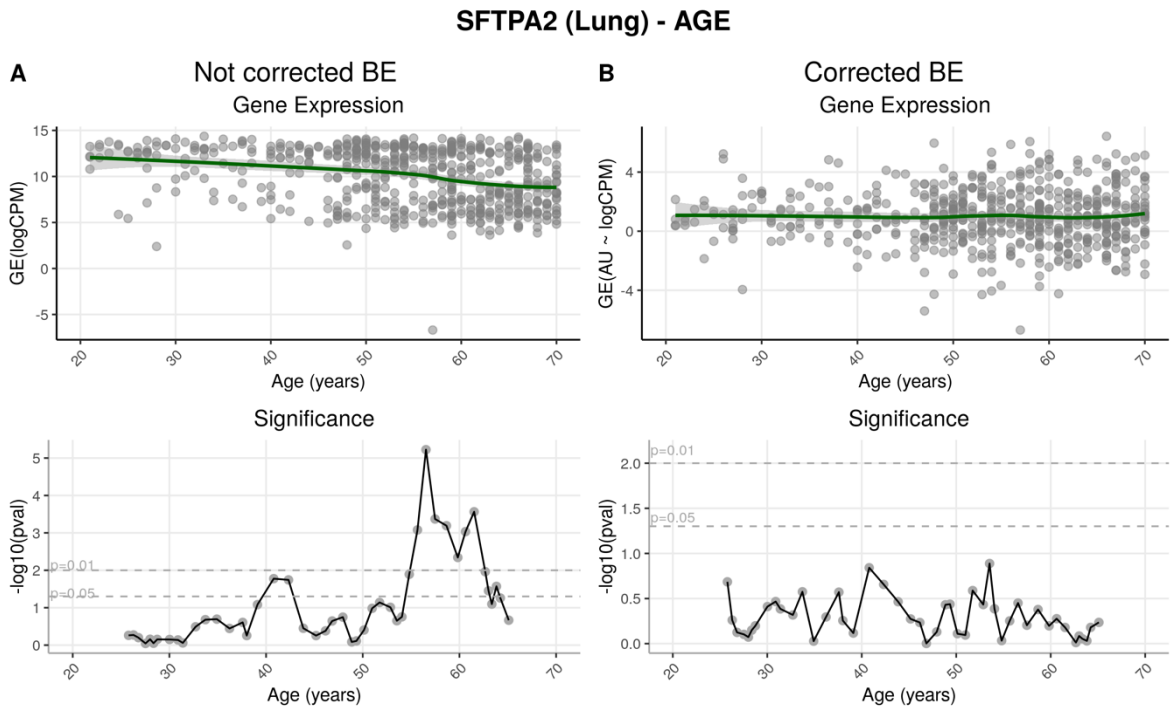
723 **Supplementary Figure 7**: *Distributions of GTEx donors' ages and its impact on the statistical*
724 *power of detection of differential gene expression.*
725 For thyroid as an example tissue, the sex-differentiated distribution of GTEx samples' ages
726 (v8) (above) are shown together with the percentage of altered genes over age (below) for
727 the three variables (*Age*, *Sex* and *Age&Sex*), as well as their statistical significance, given by
728 the ShARP-LM approach. The statistical power for detecting transcriptional changes is
729 influenced by the age distribution of donors, leading to a higher likelihood of identifying
730 significant alterations (FDR <= 0.05) in age ranges that are more prevalent in the sample
731 population, typically associated with older individuals.


A
Original vs Downsampled ShARP-LM results
% # Sample Decrease
0 25 50 75 100
Pearson's Correlation
1.0
0.5
0.0
-0.5
Uterus
Small Intestine - Terminal Ileum
Ovary
Stomach
Esophagus - Muscularis
Prostate
Vagina
Colon - Transverse
Colon - Sigmoid
Testis
Esophagus - Mucosa
Pancreas
Breast - Mammary Tissue
Spleen
Artery - Tibial
Artery - Aorta
Cells - EBV-transformed lymphocytes
Cells - Cultured fibroblasts
Esophagus - Gastroesophageal Junction
Whole Blood
Minor Salivary Gland
Nerve - Tibial
Adipose - Visceral (Omentum)
Muscle - Skeletal
Adipose - Subcutaneous
Skin - Sun Exposed (Lower leg)
Adrenal Gland
Thyroid
Skin - Not Sun Exposed (Suprapubic)
Lung
Artery - Coronary
Heart - Left Ventricle
Brain - Cerebellar Hemisphere
Brain - Amygdala
Brain - Substantia nigra
Heart - Atrial Appendage
Brain - Nucleus accumbens (basal ganglia)
Brain - Cerebellum
Liver
Brain - Frontal Cortex (BA9)
Brain - Hippocampus
Pituitary
Brain - Anterior cingulate cortex (BA24)
Brain - Cortex
Brain - Putamen (basal ganglia)
Brain - Hypothalamus
Brain - Caudate (basal ganglia)
Tissue
B
Original vs Downsampled ShARP-LM p-values (AGE variable)
% Downsampling in Age Window
0 25 50 75
Adipose - Subcutaneous
Adipose - Visceral (Omentum)
Adrenal Gland
Artery - Aorta
Artery - Coronary
Artery - Tibial
Brain - Amygdala
Brain - Anterior cingulate cortex (BA24)
Brain - Caudate (basal ganglia)
Brain - Cerebellar Hemisphere
Brain - Cerebellum
Brain - Cortex
Brain - Frontal Cortex (BA9)
Brain - Hippocampus
Brain - Hypothalamus
Brain - Nucleus accumbens (basal ganglia)
Brain - Putamen (basal ganglia)
Brain - Substantia nigra
Breast - Mammary Tissue
Cells - Cultured fibroblasts
Cells - EBV-transformed lymphocytes
Colon - Sigmoid
Colon - Transverse
Esophagus - Gastroesophageal Junction
Esophagus - Mucosa
Esophagus - Muscularis
Heart - Atrial Appendage
Heart - Left Ventricle
Liver
Lung
Minor Salivary Gland
Muscle - Skeletal
Nerve - Tibial
Ovary
Pancreas
Pituitary
Prostate
Skin - Not Sun Exposed (Suprapubic)
Skin - Sun Exposed (Lower leg)
Small Intestine - Terminal Ileum
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
Whole Blood
Downsampled p-value
0.000
0.025
0.050
0.075
0.100
Original p-value

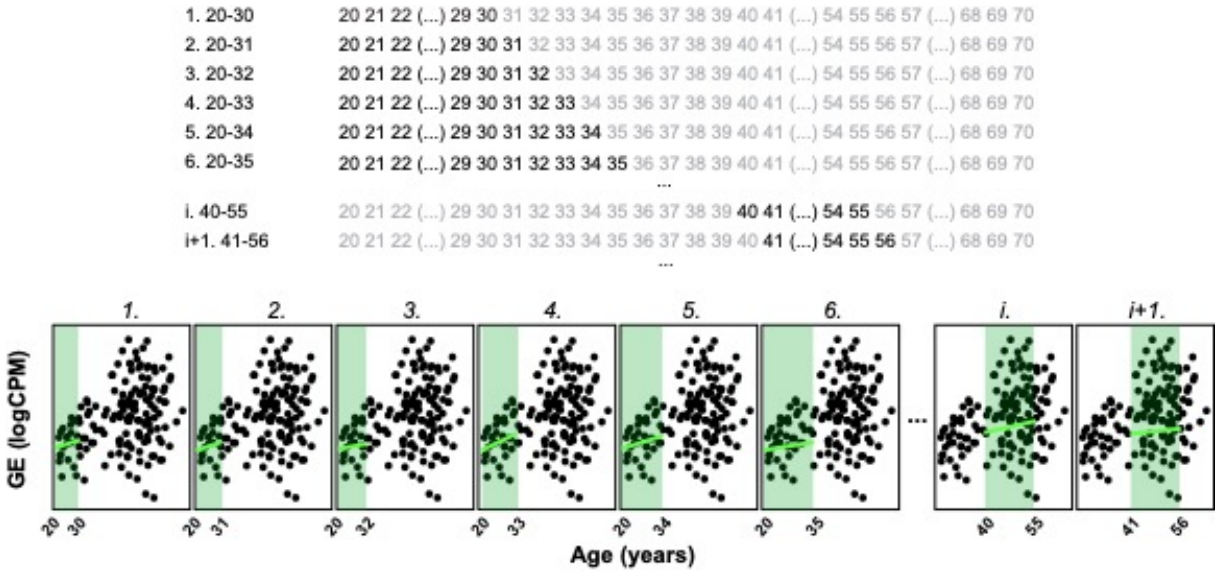
733 **Supplementary Figure 8**: *Effect of downsampling in ShARP-LM results*.
734 Downsampling was performed by reducing the number of samples in a given window to the
735 minimum available sample size within all age windows for a given tissue. **A** – Per tissue violin
736 plots of gene-wide distributions of Pearson's correlation coefficients between original and
737 downsampled logFC values for the Age variable across age windows, with tissues coloured by
738 and ordered by increasing percentage of downsampling-associated reduction in the number
739 of samples. **B** – Density scatter plots of comparison of associated original and downsampled
740 p-values for each tissue, coloured by the downsampling percentage in each age window,
741 highlighting the low range of p-values (from 0 to 0.1). Despite changes in logFC with
742 downsampling, a considerable correlation in significance is maintained, although
743 downsampling naturally results in a loss of statistical power, evident by the shift of points
744 towards the first quadrant (dashed lines: p-value = 0.05).



Supplementary Figure 9: *Batch effect correction applied to lung samples.* (**Top**) Before batch effect correction, 17.7% of the variance between samples appears to be mainly explained by the COHORT variable. (**Bottom**) After batch effect correction, PC1 no longer appears to be explained by the COHORT variable, with its weights haveing a similar distribution between batches.



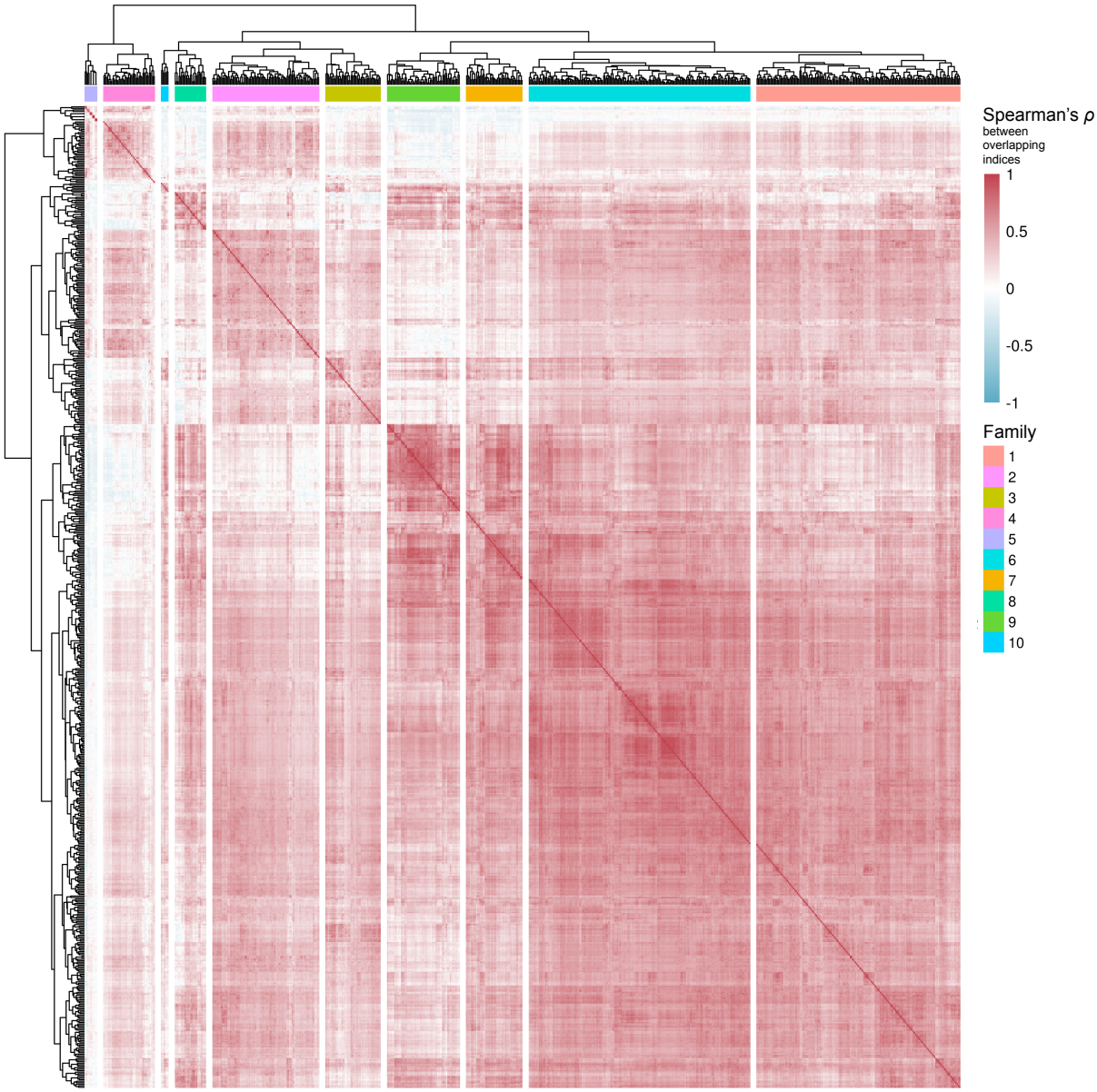
Supplementary Figure 10: *Effects of batch effect correction on gene-centric (SFTPA2) analysis*. **A** - Before batch effect correction, *SFTPA2* expression diminishes throughout ageing. This change is particularly significant after around 55 years of age. **B** - After correction, expression of this gene shows no significant alterations across age ranges.



754 **Supplementary Figure 11**: *Principle of the ShARP-LM method.*

755 ShARP-LM involves fitting linear models to gene expression data for each tissue, with
756 samples from donors grouped into age windows of 16 years each, shifted consecutively
757 through individual years of age using a sliding window approach. As samples at the ends of
758 the dataset's age range would thereby be involved in fewer linear models, we made the
759 window size gradually increase from 11 to 16 years when starting from the "youngest"
760 samples and decrease from 16 to 11 years when reaching the "oldest". The example shows
761 the expression of an arbitrary gene over age, with 8 age windows shown. Coloured
762 rectangles represent the age windows and coloured lines the linear modelling of the gene
763 expression over age in each of them.

764



765 **Supplementary Figure 12:** *Clustering of pathways from Reactome and KEGG, and level 3*
766 *Gene Ontology Biological Processes.*
767 Based on their common genes (v. Materials and Methods), pathways from curated
768 databases were clustered in 10 families.

Tissue	#Genes
Adipose - Subcutaneous	15519
Adipose - Visceral (Omentum)	15623
Adrenal Gland	15332
Artery - Aorta	15172
Artery - Coronary	15517
Artery - Tibial	14714
Brain - Amygdala	15816
Brain - Anterior cingulate cortex (BA24)	15867
Brain - Caudate (basal ganglia)	16136
Brain - Cerebellar Hemisphere	16207
Brain - Cerebellum	16473
Brain - Cortex	16178
Brain - Frontal Cortex (BA9)	16020
Brain - Hippocampus	15910
Brain - Hypothalamus	16412
Brain - Nucleus accumbens (basal ganglia)	16153
Brain - Putamen (basal ganglia)	15768
Brain - Spinal cord (cervical c-1)	15838
Brain - Substantia nigra	15796
Breast - Mammary Tissue	16311
Cells - Cultured fibroblasts	13737
Cells - EBV-transformed lymphocytes	13743
Colon - Sigmoid	15635
Colon - Transverse	16491
Esophagus - Gastroesophageal Junction	15459
Esophagus - Mucosa	15509
Esophagus - Muscularis	15355
Heart - Atrial Appendage	15003
Heart - Left Ventricle	14036
Kidney - Cortex	16054
Liver	14376
Lung	16616
Minor Salivary Gland	16412
Muscle - Skeletal	13331
Nerve - Tibial	16468
Ovary	15987
Pancreas	14857
Pituitary	17445
Prostate	17065
Skin - Not Sun Exposed (Suprapubic)	16079
Skin - Sun Exposed (Lower leg)	16097
Small Intestine - Terminal Ileum	16901
Spleen	16597
Stomach	15990
Testis	21913
Thyroid	16704
Uterus	15946
Vagina	16538
Whole Blood	13066

769 **Supplementary Table 1**: *Number of studied genes per tissue.*

770 For each tissue, we kept the genes with at least 1 CPM of expression in at least 40% of the

771 samples.

772 ## References

773 1. Campisi, J. Aging, Cellular Senescence, and Cancer. *Annu. Rev. Physiol.* **75**, 685–705
774 (2013).
775 2. Niccoli, T. & Partridge, L. Ageing as a Risk Factor for Disease. *Curr. Biol.* **22**, R741–R752
776 (2012).
777 3. Wyss-Coray, T. Ageing, neurodegeneration and brain rejuvenation. *Nature* **539**, 180–186
778 (2016).
779 4. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The Hallmarks of
780 Aging. *Cell* **153**, 1194–1217 (2013).
781 5. Stegeman, R. & Weake, V. M. Transcriptional Signatures of Aging. *J. Mol. Biol.* **429**, 2427–
782 2437 (2017).
783 6. Tower, J. Sex-Specific Gene Expression and Life Span Regulation. *Trends Endocrinol.*
784 *Metab.* **28**, 735–747 (2017).
785 7. Austad, S. N. & Fischer, K. E. Sex Differences in Lifespan. *Cell Metab.* **23**, 1022–1033
786 (2016).
787 8. Mele, M. *et al.* The human transcriptome across tissues and individuals. *Science* **348**,
788 660–665 (2015).
789 9. Gershoni, M. & Pietrokovski, S. The landscape of sex-differential transcriptome and its
790 consequent selection in human adults. *BMC Biol.* **15**, 7 (2017).
791 10. Kassam, I., Wu, Y., Yang, J., Visscher, P. M. & McRae, A. F. Tissue-specific sex differences
792 in human gene expression. *Hum. Mol. Genet.* **28**, 2976–2986 (2019).
793 11. Mayne, B. T. *et al.* Large Scale Gene Expression Meta-Analysis Reveals Tissue-Specific,
794 Sex-Biased Gene Expression in Humans. *Front. Genet.* **7**, (2016).
795 12. Zahn, J. M. *et al.* AGEMAP: A Gene Expression Database for Aging in Mice. *PLoS Genet.* **3**,
796 e201 (2007).
797 13. Kimmel, J. C. *et al.* Murine single-cell RNA-seq reveals cell-identity- and tissue-specific
798 trajectories of aging. *Genome Res.* **29**, 2088–2103 (2019).
799 14. Benayoun, B. A. *et al.* Remodeling of epigenome and transcriptome landscapes with
800 aging in mice reveals widespread induction of inflammatory responses. *Genome Res.* **29**,
801 697–709 (2019).

802 15. Lui, J. C., Chen, W., Barnes, K. M. & Baron, J. Changes in gene expression associated with
803 aging commonly originate during juvenile growth. *Mech. Ageing Dev.* **131**, 641–649
804 (2010).
805 16. Shavlakadze, T. *et al.* Age-Related Gene Expression Signature in Rats Demonstrate Early,
806 Late, and Linear Transcriptional Changes from Multiple Tissues. *Cell Rep.* **28**, 3263-
807 3273.e3 (2019).
808 17. The Tabula Muris Consortium. A single-cell transcriptomic atlas characterizes ageing
809 tissues in the mouse. *Nature* **583**, 590–595 (2020).
810 18. Schaum, N. *et al.* Ageing hallmarks exhibit organ-specific temporal signatures. *Nature*
811 **583**, 596–602 (2020).
812 19. Zahn, J. M. *et al.* Transcriptional profiling of aging in human muscle reveals a common
813 aging signature. *PLoS Genet.* **preprint**, e115 (2005).
814 20. Welle, S., Brooks, A. I., Delehanty, J. M., Needler, N. & Thornton, C. A. Gene expression
815 profile of aging in human muscle. *Physiol. Genomics* **14**, 149–159 (2003).
816 21. Jozsi, A. C. *et al.* Aged human muscle demonstrates an altered gene expression profile
817 consistent with an impaired response to exercise. *Mech. Ageing Dev.* **120**, 45–56 (2000).
818 22. Rodwell, G. E. J. *et al.* A Transcriptional Profile of Aging in the Human Kidney. *PLoS Biol.*
819 **2**, e427 (2004).
820 23. Galatro, T. F. *et al.* Transcriptomic analysis of purified human cortical microglia reveals
821 age-associated changes. *Nat. Neurosci.* **20**, 1162–1171 (2017).
822 24. Olah, M. *et al.* A transcriptomic atlas of aged human microglia. *Nat. Commun.* **9**, 539
823 (2018).
824 25. Berchtold, N. C. *et al.* Gene expression changes in the course of normal brain aging are
825 sexually dimorphic. *Proc. Natl. Acad. Sci.* **105**, 15605–15610 (2008).
826 26. Lu, T. *et al.* Gene regulation and DNA damage in the ageing human brain. *Nature* **429**,
827 883–891 (2004).
828 27. Işıldak, U., Somel, M., Thornton, J. M. & Dönertaş, H. M. Temporal changes in the gene
829 expression heterogeneity during brain development and aging. *Sci. Rep.* **10**, 4080 (2020).
830 28. Haustead, D. J. *et al.* Transcriptome analysis of human ageing in male skin shows mid-life
831 period of variability and central role of NF-κB. *Sci. Rep.* **6**, 26846 (2016).
832 29. Holzscheck, N. *et al.* Multi-omics network analysis reveals distinct stages in the human
833 aging progression in epidermal tissue. *Aging* **12**, 12393–12409 (2020).

834 30. Nakamura, S. *et al.* Identification of blood biomarkers of aging by transcript profiling of
835 whole blood. *Biochem. Biophys. Res. Commun.* **418**, 313–318 (2012).
836 31. Harries, L. W. *et al.* Human aging is characterized by focused changes in gene expression
837 and deregulation of alternative splicing. *Aging Cell* **10**, 868–878 (2011).
838 32. Thomas, R. Age-Associated Changes in Gene Expression Patterns in the Liver,. *J.*
839 *Gastrointest. Surg.* **6**, 445–454 (2002).
840 33. Yoshida, S., Yashar, B. M., Hiriyanna, S. & Swaroop, A. Microarray analysis of gene
841 expression in the aging human retina. *Invest. Ophthalmol. Vis. Sci.* **43**, 2554–60 (2002).
842 34. Gheorghe, M. *et al.* Major aging-associated RNA expressions change at two distinct age-
843 positions. *BMC Genomics* **15**, 132 (2014).
844 35. Yang, J. *et al.* Synchronized age-related gene expression changes across multiple tissues
845 in human and the link to complex diseases. *Sci. Rep.* **5**, 15145 (2015).
846 36. Aramillo Irizar, P. *et al.* Transcriptomic alterations during ageing reflect the shift from
847 cancer to degenerative diseases in the elderly. *Nat. Commun.* **9**, 327 (2018).
848 37. Glass, D. *et al.* Gene expression changes with age in skin, adipose tissue, blood and brain.
849 *Genome Biol.* **14**, R75 (2013).
850 38. Yu, Y. *et al.* A rat RNA-Seq transcriptomic BodyMap across 11 organs and 4
851 developmental stages. *Nat. Commun.* **5**, 3230 (2014).
852 39. Rando, T. A. & Wyss-Coray, T. Asynchronous, contagious and digital aging. *Nat. Aging* **1**,
853 29–35 (2021).
854 40. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–
855 585 (2013).
856 41. Van Deursen, J. M. The role of senescent cells in ageing. *Nature* **509**, 439–446 (2014).
857 42. Gorgoulis, V. *et al.* Cellular Senescence: Defining a Path Forward. *Cell* **179**, 813–827
858 (2019).
859 43. Gil, J. & Peters, G. Regulation of the INK4b–ARF–INK4a tumour suppressor locus: all for
860 one or one for all. *Nat. Rev. Mol. Cell Biol.* **7**, 667–677 (2006).
861 44. Erickson, S. *et al.* Involvement of the Ink4 proteins p16 and p15 in T-lymphocyte
862 senescence. *Oncogene* **17**, 595–602 (1998).
863 45. Narita, M. *et al.* Rb-Mediated Heterochromatin Formation and Silencing of E2F Target
864 Genes during Cellular Senescence. *Cell* **113**, 703–716 (2003).

865 46. Sun, X. & Kaufman, P. D. Ki-67: more than a proliferation marker. *Chromosoma* **127**,
866 175–186 (2018).
867 47. Kaczmarek, M. Variation in Age at Natural Menopause among Polish Women in Relation
868 to Biological and Social Factors. (2014) doi:10.13140/2.1.2610.1449.
869 48. Croft, D. *et al.* The Reactome pathway knowledgebase. *Nucleic Acids Res.* **42**, D472–D477
870 (2014).
871 49. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for
872 interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–
873 50 (2005).
874 50. Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**,
875 27–30 (2000).
876 51. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource.
877 *Nucleic Acids Res.* **32**, 258D – 261 (2004).
878 52. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network
879 analysis. *BMC Bioinformatics* **9**, 559 (2008).
880 53. van Dam, S., Võsa, U., van der Graaf, A., Franke, L. & de Magalhães, J. P. Gene co-
881 expression analysis for functional classification and gene–disease predictions. *Brief.*
882 *Bioinform.* bbw139 (2017) doi:10.1093/bib/bbw139.
883 54. Kelley KW, Nakao-Inoue H, Molofsky AV, O. M. Variation among intact tissue samples
884 reveals the core transcriptional features of human CNS cell classes. *Nat. Neurosci.* **21**,
885 1171–1184 (2018).
886 55. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update.
887 *Nucleic Acids Res.* (2019) doi:10.1093/nar/gkz1021.
888 56. Piñero, J. *et al.* DisGeNET: a comprehensive platform integrating information on human
889 disease-associated genes and variants. *Nucleic Acids Res.* **45**, D833–D839 (2017).
890 57. Palmer, A. L. & Ousman, S. S. Astrocytes and Aging. *Front. Aging Neurosci.* **10**, 337
891 (2018).
892 58. Lehallier, B. *et al.* Undulating changes in human plasma proteome profiles across the
893 lifespan. *Nat. Med.* **25**, 1843–1850 (2019).
894 59. Khramtsova, E. A., Davis, L. K. & Stranger, B. E. The role of sex in the genomics of human
895 complex traits. *Nat. Rev. Genet.* **20**, 173–190 (2019).

896 60. Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First
897 1,000,000 Profiles. *Cell* **171**, 1437-1452.e17 (2017).
898 61. de Almeida BP, Saraiva-Agostinho N, B.-M. N. cTRAP: Identification of candidate causal
899 perturbations from differential gene expression data. Preprint at (2020).
900 62. de Magalhães, J. P., Curado, J. & Church, G. M. Meta-analysis of age-related gene
901 expression profiles identifies common signatures of aging. *Bioinformatics* **25**, 875–881
902 (2009).
903 63. Tacutu, R. *et al.* Human Ageing Genomic Resources: new and updated databases. *Nucleic*
904 *Acids Res.* **46**, D1083–D1090 (2018).
905 64. Craig, T. *et al.* The Digital Ageing Atlas: integrating the diversity of age-related changes
906 into a unified resource. *Nucleic Acids Res.* **43**, D873–D878 (2015).
907 65. Dönertaş, H. M., Fuentealba Valenzuela, M., Partridge, L. & Thornton, J. M. Gene
908 expression-based drug repurposing to target aging. *Aging Cell* **17**, e12819 (2018).
909 66. Janssens, G. E. *et al.* Transcriptomics-Based Screening Identifies Pharmacological
910 Inhibition of Hsp90 as a Means to Defer Aging. *Cell Rep.* **27**, 467-480.e6 (2019).
911 67. https://shiny.rstudio.com/gallery/.
912 68. Kunst, J. highcharter: A Wrapper for the 'Highcharts' Library.
913 69. Yihui, X., Joe, C. & Xianying, T. DT: A Wrapper of the JavaScript Library 'DataTables'.
914 70. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for
915 differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–
916 140 (2010).
917 71. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression
918 analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
919 72. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model
920 analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
921 73. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing
922 and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
923 74. García-Pérez, R. *et al.* The landscape of expression and alternative splicing variation
924 across human traits. *Cell Genomics* **3**, 100244 (2023).
925 75. Buuren, S. V. & Groothuis-Oudshoorn, K. **mice** : Multivariate Imputation by Chained
926 Equations in *R. J. Stat. Softw.* **45**, (2011).

927 76. McCall, M. N., Illei, P. B. & Halushka, M. K. Complex Sources of Variation in Tissue
928 Expression Data: Analysis of the GTEx Lung Transcriptome. *Am. J. Hum. Genet.* **99**, 624–
929 635 (2016).
930 77. Korotkevich, G. *et al.* Fast gene set enrichment analysis. *bioRxiv* 060012 (2021)
931 doi:10.1101/060012.
932 78. GSEA-MSIGDB. Gene Set Enrichment Analysis (GSEA) User Guide.
933 79. Chen, Y.-A. *et al.* Integrated Pathway Clusters with Coherent Biological Themes for
934 Target Prioritisation. *PLoS ONE* **9**, e99030 (2014).
935 80. Dowle, M. & Srinivasan, A. data.table: Extension of 'data.frame'.
936 81. Langfelder, P. & Horvath, S. WGCNA Frequently Asked Questions.
937 82. Chen, B., Khodadoust, M. S., Liu, C. L., Newman, A. M. & Alizadeh, A. A. Profiling Tumor
938 Infiltrating Immune Cells with CIBERSORT. in *Cancer Systems Biology* (ed. Von Stechow,
939 L.) vol. 1711 243–259 (Springer New York, 2018).
940 83. Regev, A. *et al.* The Human Cell Atlas. *eLife* **6**, e27041 (2017).
941 84. Fan, X. *et al.* Spatial transcriptomic survey of human embryonic cerebral cortex by single-
942 cell RNA-seq analysis. *Cell Res.* **28**, 730–745 (2018).
943 85. Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science* **370**, eaba7721 (2020).