

1

2

3

4 **Gene size matters: What determines gene length in the human genome?**

5

6 Inês Lopes<sup>1</sup>, Gulam Altab<sup>1</sup>, Priyanka Raina<sup>1</sup>, João Pedro de Magalhães<sup>1\*</sup>

7

8 <sup>1</sup>Integrative Genomics of Ageing Group, Institute of Ageing and Chronic Disease, University of

9 Liverpool, Liverpool, L7 8TX, United Kingdom

10

11

12 **\* Corresponding Author:**

13 João Pedro de Magalhães; email for correspondence: [jp@senescence.info](mailto:jp@senescence.info)

14

15 **Abstract**

16

17 While it is expected for gene length to be influenced by factors such as intron number and  
18 evolutionary conservation, we have yet to fully understand the connection between gene length  
19 and function in the human genome.

20 In this study, we show that, as expected, there is a strong positive correlation between gene  
21 length and the number of SNPs, introns and protein size. Amongst tissue specific genes, we find  
22 that the longest genes are expressed in blood vessels, nerve, thyroid, cervix uteri and brain,  
23 while the smallest genes are expressed within the pancreas, skin, stomach, vagina and testis. We  
24 report, as shown previously, that natural selection suppresses changes for genes with longer  
25 lengths and promotes changes for smaller genes. We also observed that longer genes have a  
26 significantly higher number of co-expressed genes and protein-protein interactions. In the  
27 functional analysis, we show that bigger genes are often associated with neuronal development,  
28 while smaller genes tend to play roles in skin development and in the immune system.  
29 Furthermore, pathways related to cancer, neurons and heart diseases tend to have longer genes,  
30 with smaller genes being present in pathways related to immune response and  
31 neurodegenerative diseases.

32 We hypothesise that longer genes tend to be associated with functions that are important early  
33 in life, while smaller genes play a role in functions that are important throughout the organisms'  
34 whole life, like the immune system which require fast responses.

35

36

37

38

## 39 **Author Summary**

40 Even though the human genome has been fully sequenced, we still do not fully grasp all of its  
41 nuances. One such nuance is the length of the genes themselves. Why are certain genes longer  
42 than others? Is there a common function shared by longer/smaller genes? What exactly makes  
43 gene longer? We tried answering these questions using a variety of analysis. We found that,  
44 while there was not a particular strong factor in genes that influenced their size, there could be  
45 an influence of several gene characteristics in determining the length of a gene. We also found  
46 that longer genes are linked with the development of neurons, cancer, heart diseases and  
47 muscle cells, while smaller genes seem to be mostly related with the immune system and the  
48 development of the skin. This led us to believe that, whether the gene has an important function  
49 early in our life, or throughout our whole lives, or even if the function requires a rapid response,  
50 that its gene size will be influenced accordingly.

## 51 **Background**

52 With the sequencing of the human genome [1–3] there arose a great interest in understanding  
53 the relationship between genotype and phenotype, especially concerning human health [4,5].  
54 However, despite the recent advancements, we have yet to fully understand the human genome  
55 and its complexity [6].

56 Several studies have tried to decipher a connection between the length of a gene and its  
57 function. It is believed that genes that are more evolutionarily conserved are often associated  
58 with longer gene length and higher intronic burden [7–10]. In contrast, smaller gene length is  
59 often associated with high expression, smaller proteins and little intronic content [11]. This  
60 hypothesis is further supported by the house keeping genes, which are widely expressed and  
61 have characteristics similar to smaller gene length genes [12]. It was hypothesised that, due to  
62 this great levels of expression for smaller genes, there is selective pressure to maximize protein  
63 synthesis efficiency [11]. If that is the case, then the next question should be what functions  
64 serve longer genes to compensate for their expensive production of proteins. Gene length has  
65 been importantly associated with biological timing. The smaller genes produce smaller proteins  
66 faster, and these proteins often play a part in the regulation of longer proteins, which are  
67 expressed much later into the response. This allows for regulatory mechanisms to be set up in  
68 preparation for important protein expression [13]. On the other hand, longer genes have been  
69 associated with some important processes, including embryonic development [14] and  
70 neuronal processes [15]. Longer genes have also been previously shown to be related to  
71 diseases such as cancer, cardiomyopathies and diabetes [15].

72 In this present work, we used human genome data [16], to identify possible functions based on  
73 gene size. Correlation tests were used to search for relationships between gene length and other  
74 gene characteristics. In order to find the specific functions associated with gene size, the Gene  
75 Ontology (GO) and the KEGG Pathway were used. We observed that longer genes are expressed  
76 in the brain, heart diseases and cancer, while smaller genes mostly participate in the immune

77 system and in the development of the skin. Therefore, we hypothesize that genes with longer  
78 lengths are mostly associated with functions in the early development stages, while genes with  
79 smaller lengths have important roles in day-to-day functions.

## 80 Results

### 81 Longest and shortest genes

82 For all of the protein-coding transcripts in the human genome, a dataset was built selecting only  
83 the transcripts with the highest transcript length per gene (N=19,714 genes, S1 Table). Using  
84 mostly the transcript length for the rest of this analysis, stems from the fact that there is a very  
85 high correlation between the length of the longest transcript of a gene and its respective gene  
86 length (S1 Fig, Kendall test, tau = 0.72, p-value < 2.20E-16). The 5 biggest genes in terms of  
87 transcript length have all been studied previously, and we can see that they are associated with  
88 neuron functions [17–19], cardiac tissue [20] and cancer [21] (Table 1). However, the smallest  
89 genes might be annotation errors in the genome build.

90

91 **Table 1. List of the top 5 longest protein-coding transcripts in human.**

Transcript Stable ID	Gene	Gene name	Transcript Length	Exon Counts	Intron Counts	Number of SNPs	Protein size
<b>Longest Genes</b>							
ENST00000589042	ENSG00000155657	<i>TTN</i>	109224	363	362	74829	35991
ENST00000397910	ENSG00000181143	<i>MUC16</i>	43816	84	83	42852	14507
ENST00000262160	ENSG00000175387	<i>SMAD2</i>	34626	11	10	30781	467
ENST00000330753	ENSG00000185070	<i>FLRT2</i>	33681	2	1	28178	660
ENST00000609686	ENSG00000273079	<i>GRIN2B</i>	30355	13	12	98658	1484

92

93

94

95

96

97

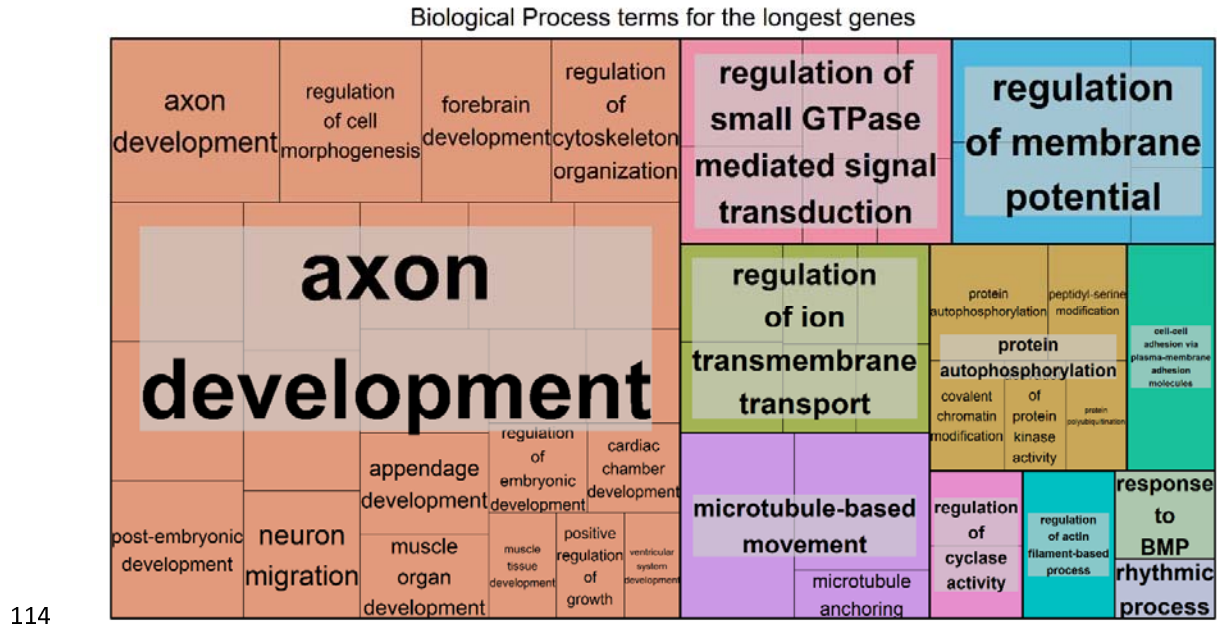
98 **Functional analysis**

99 One of the main objectives of the present study was to understand if gene function changed  
100 depending on the gene length. Keeping this in mind, and using a list of the top 5% protein  
101 coding genes with the longest and smallest transcript length, we performed an analysis, using  
102 tools like WebGestalt [22], DAVID [23,24], KEGG [25] and Molecular Signature Database [26,27].  
103 The results for KEGG Pathways, were colour coded for each boxplot based on their association  
104 with the terms we found most relevant (brain, cancer, heart, immune system, muscle,  
105 neurodegenerative disease, skin and other). For cases where there was no direct association, a  
106 literature search was done for relevant articles that might show that genes in those pathways  
107 were related to brain [28–47], cancer [48], immune system [49–53] and skin [54–58].

108 For genes with longer gene length (Fig 1), most of the biological functions found seem to be  
109 associated with the brain, specifically in regards to neurons. This can also be confirmed when  
110 looking at the Cellular Component (S2A Fig) and Molecular Function (S2B Figure), and at the  
111 similar results produced using DAVID (S2 Table).

112

113



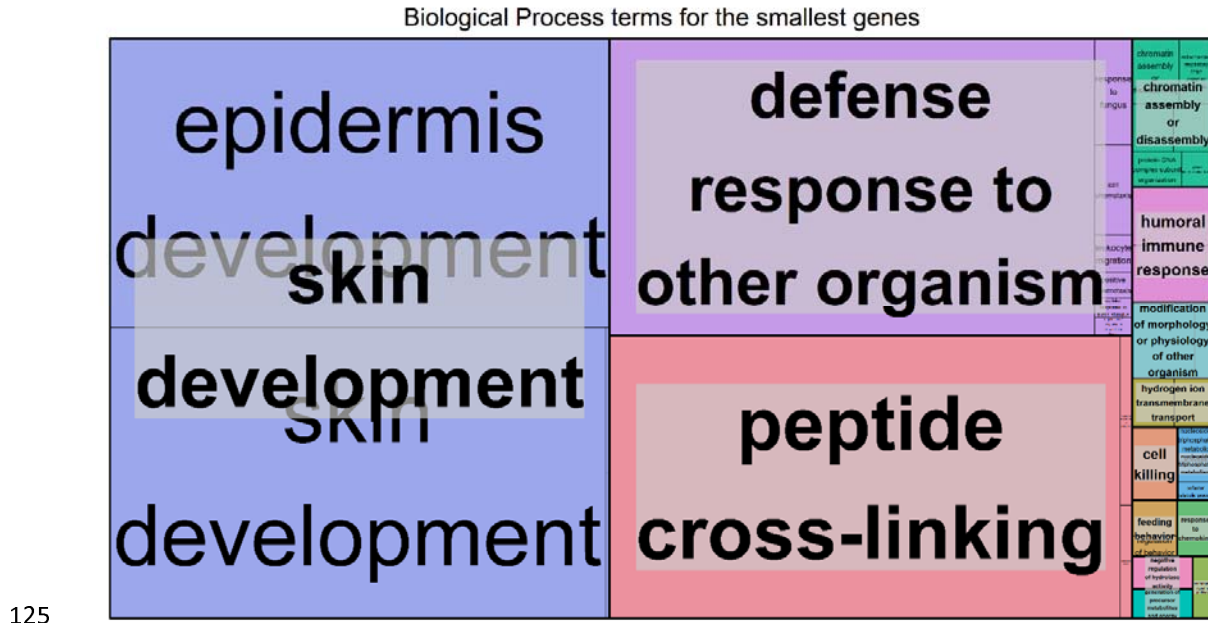
115 **Fig 1. Biological Process terms found associated to genes with the longest transcript**  
116 **length. Overrepresentation Enrichment Analysis was performed with WebGestalt [22]**  
117 **and the visualization tool REViGO [59] was used to produce this figure. The significance**  
118 **level was  $p < 0.05$  and the FDR was set at 0.05. FDR estimation was done using the**  
119 **Benjamini-Hochberg method.**

120

121 For the genes with smaller gene length (Fig 2), most of the biological functions found are related  
122 to skin and the immune system. Similarly to what we observed before, Cellular Component (S2C  
123 Fig), Molecular Function (S2D Fig) and DAVID (S2 Table) results supported this observation.

124





126 **Fig 2. Biological Process terms found associated to genes with the smallest transcript**  
127 **length. Overrepresentation Enrichment Analysis was performed with WebGestalt [22]**  
128 **and the visualization tool REViGO [59] was used to produce this figure. The significance**  
129 **level was  $p < 0.05$  and the FDR was set at 0.05. FDR estimation was done using the**  
130 **Benjamini-Hochberg method.**

131

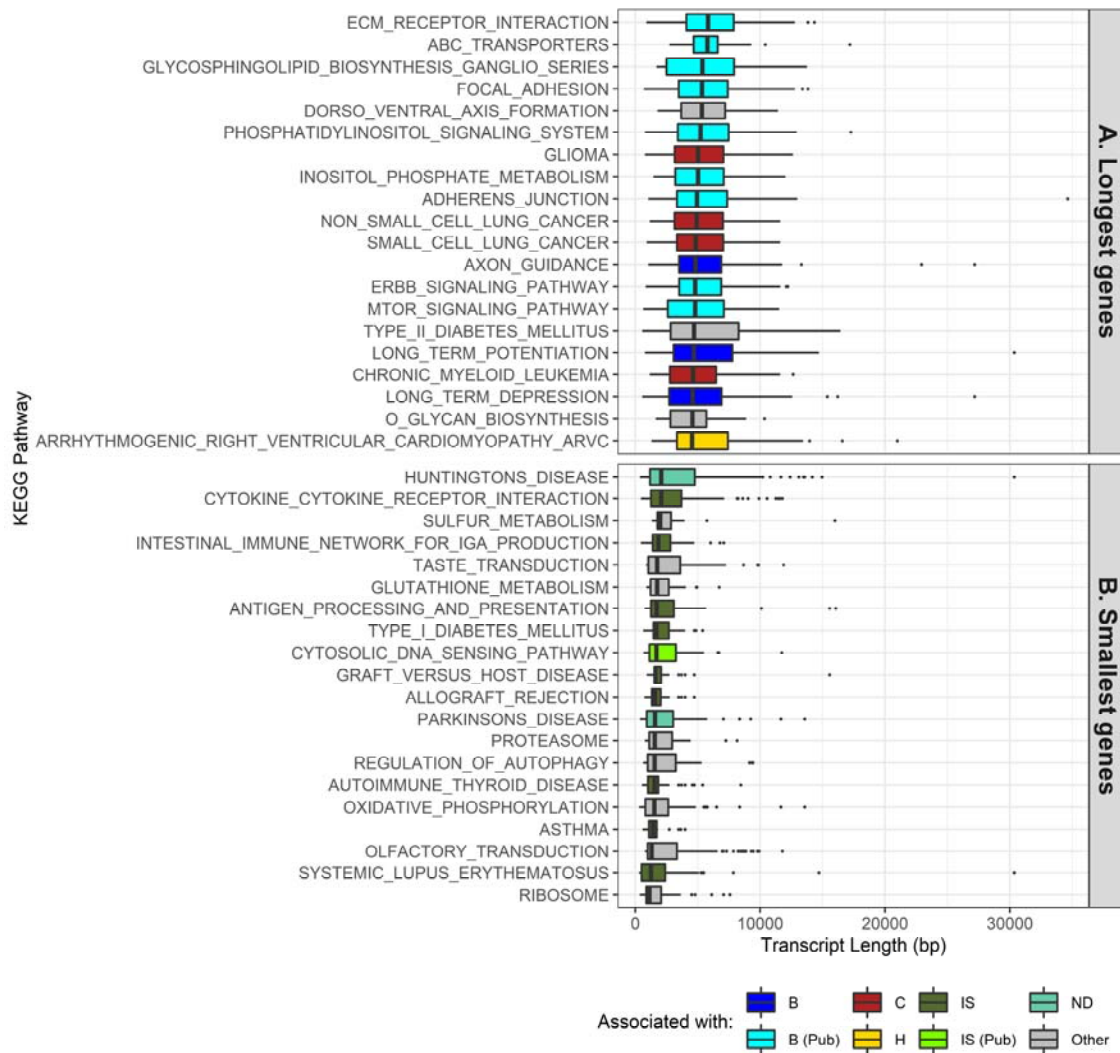
132 Additionally, while looking at the KEGG Pathways results for longest transcript length, we  
133 identified pathways associated with the brain, cancer, heart disease and muscle (Fig 3A, S3 Fig),  
134 while the pathways with the smallest transcript length are mostly associated with the immune  
135 system, a few of them were also associated with skin and neurodegenerative diseases (Fig 3B,  
136 S3 Fig).

137

138 The full KEGG Results (186 gene sets) can be found in the S3 Fig, and the KEGG Pathway IDs can  
139 be found in the S3 Table.

140

141



142

143

144 **Fig 3. Transcript length distribution per KEGG Pathway for the longest and smallest**  
 145 **genes. Colours illustrate what the KEGG pathway has been directly associated with (B for**  
 146 **Brain, C for Cancer, H for Heart, IS for Immune system and ND for Neurodegenerative**  
 147 **diseases), due to it being stated in the pathway itself, or indirectly associated with (Pub**  
 148 **tag), by means of literature references. KEGG Pathways and genes involved in said**  
 149 **pathways were obtained from the Molecular Signature Database [26,27]. A: Top 20**

150 **Pathways with the longest genes, ordered by median; B: Top 20 Pathways with the**  
151 **smallest genes, ordered by median.**

152

153

#### 154 **Gene properties correlate with transcript length**

155 In order to understand the relationship between transcript length and other gene  
156 characteristics, a correlation analysis was done. When looking at the number of SNPs for each  
157 transcript (Fig 4A), there was a significant positive correlation with transcript length (Kendall  
158 test,  $\tau = 0.45$ ,  $p\text{-value} < 2.20E-16$ ). Similar results were found, when comparing the number of  
159 SNPs per gene with gene length (S4A Fig, Kendall test,  $\tau = 0.49$ ,  $p\text{-value} < 2.20E-16$ ). After  
160 comparing the number of introns and the transcript length (Fig 4B), we found a weak significant  
161 positive correlation between these two variables (Kendall test,  $\tau = 0.35$ ,  $p\text{-value} < 2.20E-16$ ).  
162 The strongest positive correlation (Kendall test,  $\tau = 0.48$ ,  $p\text{-value} < 2.20E-16$ ) was associated  
163 with the protein size (Fig 4C), and the weakest correlation (Kendall test,  $\tau = 0.04$ ,  $p\text{-value} =$   
164  $3.06E-14$ ) was associated with the average gene expression (Fig 4D).

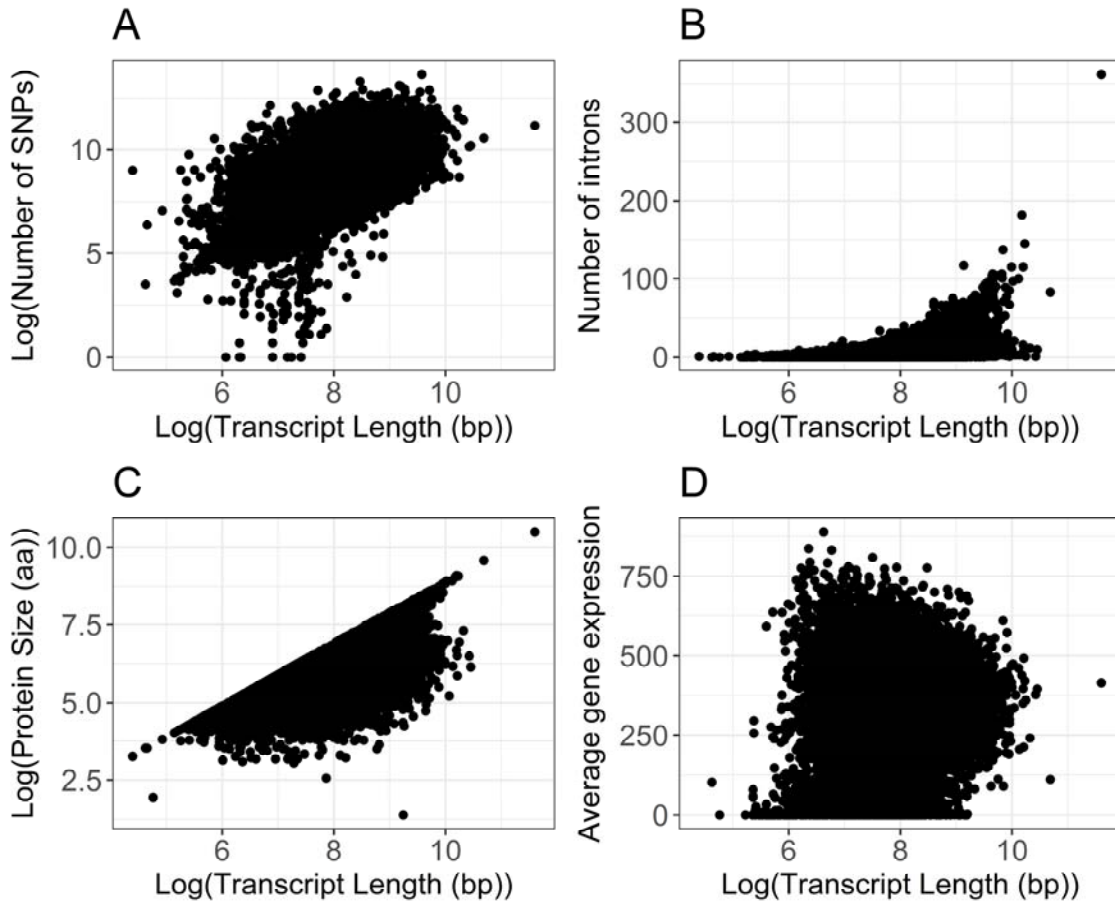
165

166

167

168

169



170

171

172 **Fig 4. Correlation analysis between Transcript Length (bp) and several other gene**  
173 **characteristics. All figures have been logarithmically transformed in order to help**  
174 **visualize their relationship and/or account for the skewing introduced by outliers. The**  
175 **original versions of the figures can be found in the S4B, S4C, S4D and S4E Fig. [A](#):**  
176 **Correlation between the log transformed number of SNPs and the log transformed**  
177 **Transcript Length (bp) (Kendall test, tau = 0.45, p-value < 2.20E-16). Number of SNPs and**  
178 **Transcript Length for each transcript were obtained using biomart; [B](#): Correlation**  
179 **between the number of introns and the log transformed Transcript Length (bp) (Kendall**  
180 **test, tau = 0.35, p-value < 2.20E-16). Number of introns and Transcript Length for each**  
181 **transcript were obtained using biomart; [C](#): Correlation between the log transformed**  
182 **Protein Size (aa) and the log transformed Transcript Length (bp) (Kendall test, tau = 0.48,**

183 **p-value < 2.20E-16). Protein Size and Transcript Length were obtained using biomart; D:**  
184 **Correlation between the Average Gene Expression and the log transformed Transcript**  
185 **Length (bp) (Kendall test, tau = 0.04, p-value = 3.06E-14). Average Gene Expression was**  
186 **obtained from the UCSC Genome browser, this value was derived from the total median**  
187 **expression level across all tissues and was based on the GTEx project. Transcript Length**  
188 **was obtained using biomart.**

189

190

191 Additionally, for the correlations with Transcript count (S4F Fig) and GC content (S4G Fig), we  
192 observed a weak significant positive correlation (Kendall test, tau = 0.22, p-value < 2.20E-16)  
193 and a weak significant negative correlation (Kendall test, tau = -0.19, p-value < 2.20E-16),  
194 respectively.

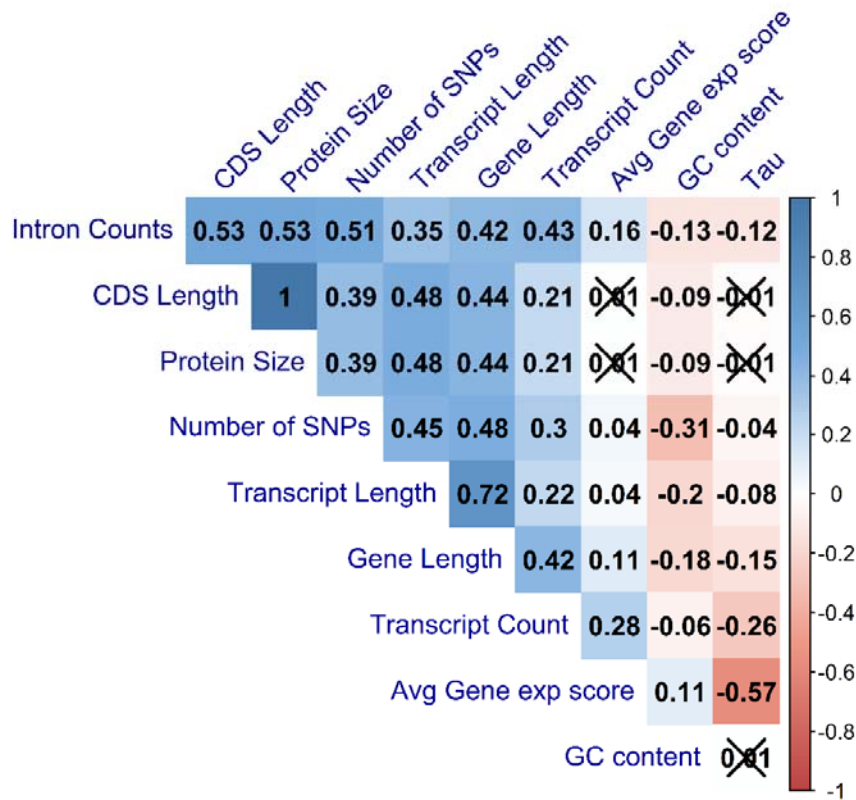
195

196 We were also interested in understanding the effect of transcript length in some particular  
197 mutations. We observed some strong statistically significant correlations between transcript  
198 length and synonymous (S4H Fig, Kendall test, tau = 0.44, p-value < 2.20E-16) and missense  
199 (S4I Fig, Kendall test, tau = 0.42, p-value < 2.20E-16) mutations. However, in case of nonsense  
200 mutations (S4J Fig, Kendall test, tau = 0.21, p-value < 2.20E-16) a weaker significant positive  
201 correlation with transcript length was observed. This was followed by the calculation of  
202 Missense/Synonymous (MIS/SYN) and Nonsense/Synonymous (NONS/SYN) rates in order to  
203 measure the functional importance of gene length. We observed that this ratios had similarly  
204 negative correlations with transcript length, with MIS/SYN having a weaker significant  
205 correlation (S4K Fig, Kendall test, tau = -0.07, p-value < 2.20E-16) than NONS/SYN (S4L Fig,  
206 Kendall test, tau = -0.19, p-value < 2.20E-16).

207

208 In order to better understand if the correlations found were solely due to the transcript length  
 209 or if other factors were influencing them, we built a correlation matrix with several gene  
 210 characteristics (Fig 5). We observed that properties like intron counts, CDS length, protein size,  
 211 number of SNPs and transcript count have some strong positive correlations amongst  
 212 themselves, some of which were stronger than any other correlation with transcript length. This  
 213 indicated that strong correlations with transcript length might not be due to the sole action of  
 214 transcript length itself, but rather due to a combined action between several gene  
 215 characteristics.

216



217

218

219 **Fig 5. Correlation matrix between gene properties. Kendall's test was used as a**  
 220 **measurement of correlation, with the numbers and the gradient of colours symbolizing**

221 **the Tau values for each comparison. Number of SNPs values is for each transcript. Values**  
222 **that are crossed out are not statistically significant. Values are clustered together based**  
223 **on their Tau values.**

224

225

## 226 **Distribution of transcript length and expression in human tissues**

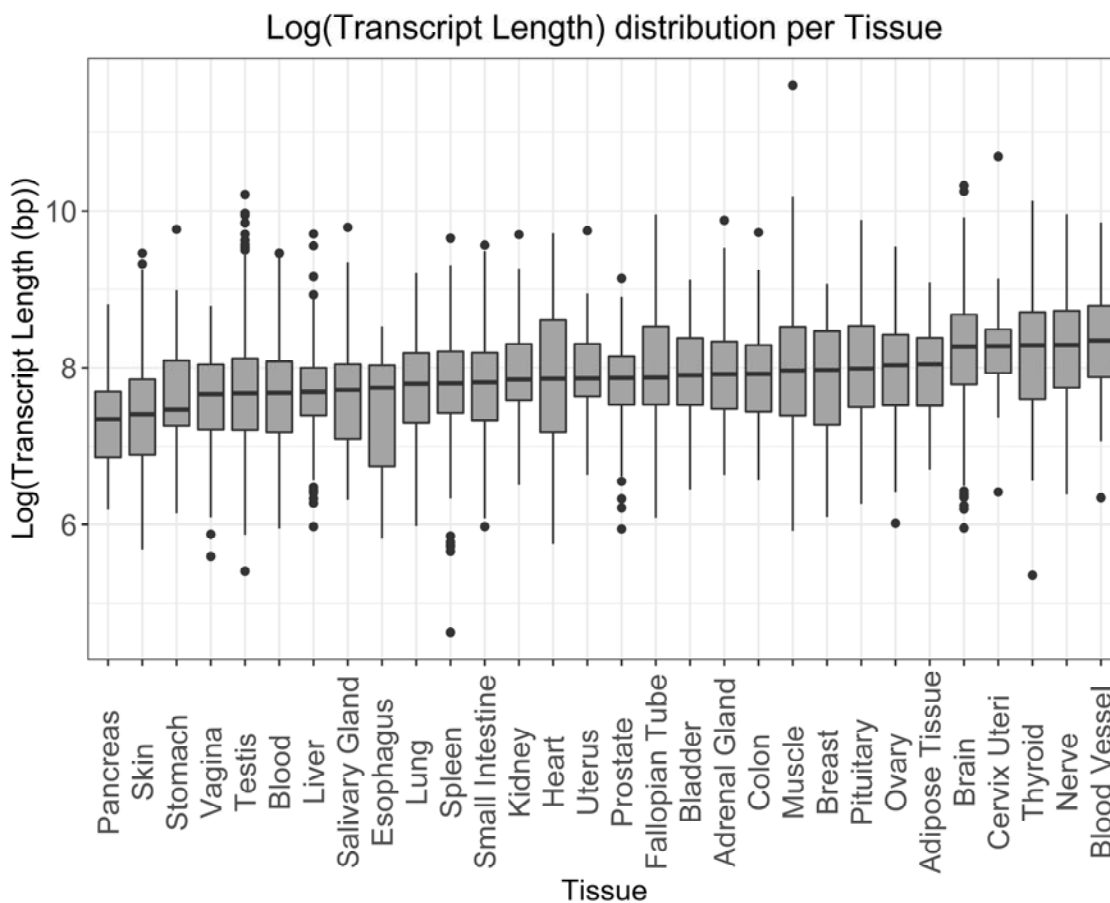
227 In this present work we have found that transcript length seems to peak at 2065 bp, with  
228 smaller transcripts being more common than longer ones (S5A Fig). As described previously [9],  
229 the distribution of the number of introns in the human genome (S5B Fig) has a mode of 3  
230 introns and there are very few genes with a large number of introns. The gene with the most  
231 introns is TTN, with 362 introns, which also leads the list of genes with the longest transcript  
232 length.

233 To better understand the distribution of transcript length in the human tissue specific genes, we  
234 used Tau values obtained from GTEx data [60]. Tau was used as a measure of tissue specificity,  
235 based on the expression profile in different tissues, with values ranging from 0, for broadly  
236 expressed genes, to 1, for tissue specific genes [61]. For genes with a Tau value above 0.8 (Fig 6,  
237 S6 Fig for the non-log transformed version), we observed that longer tissue specific genes are  
238 often associated with the blood vessel, nerve, thyroid, cervix uteri and brain, while smaller  
239 tissue specific genes are found in the pancreas, skin, stomach, vagina and testis.

240

241

242



243

244

245 **Fig 6. Log transformed Transcript length distribution for genes specifically expressed in**  
246 **the given Tissues. Tissue specificity was defined as a gene having a Tau specificity score**  
247 **greater than 0.8.**

248

249

## 250 **Ageing and transcript length**

251 Ageing is an important factor in our lives, and it affects most organisms. We were curious to see  
252 if, for genes related to ageing, the distribution of transcript length was significantly different  
253 than the rest of the protein-coding genes. We observed (S7A Fig and S7B Fig) that genes  
254 associated with ageing (N = 307) [62] have longer transcript lengths (median = 3517) when



255 compared with the rest of our dataset (median = 2956), and that this difference of medians was  
256 significant (Wilcoxon rank sum test, p-value = 0.00036).

257

258 To further understand if longer or smaller genes were more prominent with age, we used genes  
259 from ageing signatures obtained from a meta-analysis in human, mice and rat [60]. Genes from  
260 this signature were either overexpressed ( $N_{\text{Total}} = 449$ ,  $N_{\text{Brain}} = 147$ ,  $N_{\text{Heart}} = 35$ ,  $N_{\text{Muscle}} = 49$ ) or  
261 underexpressed ( $N_{\text{Total}} = 162$ ,  $N_{\text{Brain}} = 16$ ,  $N_{\text{Heart}} = 5$ ,  $N_{\text{Muscle}} = 73$ ) with age. Overall, the difference  
262 in medians for the distribution of transcript length in genes overexpressed (median = 3068) and  
263 underexpressed (median = 3026.5) with ageing was not observed to be significant (S7C Fig,  
264 Wilcoxon rank sum test, p-value = 0.81). However, tissue specific signatures showed that the  
265 brain favours smaller genes with age (S7D Fig, Wilcoxon rank sum test, p-value = 0.00086,  
266 median for overexpression in brain = 2651, median for underexpression in brain = 5824).

267

268

## 269 **Evolution and transcript length**

270 The relationship between intronic burden and evolution has been established before [9], but  
271 very few works approached this on a gene length front. Therefore we obtained the dN and dS  
272 values for three organisms paired with human, mouse (S8A Fig), gorilla (S8B Fig) and  
273 chimpanzee (S8C Fig), and we aimed to see how the distribution of transcript length happened  
274 in function of their dN/dS ratio. Overall, longer genes were associated with a dN/dS ratio lesser  
275 to 1 (median transcript length is 3294, 3377 and 3338 for mouse, chimpanzee and gorilla  
276 respectively), while smaller genes seem to be more associated with dN/dS ratios above or equal  
277 to 1 (median transcript length is 1171.5, 2229.5 and 2092 for mouse, chimpanzee and gorilla  
278 respectively) and the median of both groups was always significantly different (Wilcoxon rank  
279 sum test, p-value = 0.00073 for mouse and  $<2.2E-16$  for both gorilla and chimpanzee).

280

281

## 282 **Co-Expression Analysis and Protein-Protein Interactions**

283 Co-expression networks can help us to better understand the functions of genes that are often  
284 expressed together [63]. In order to see if the gene length influenced the amount of co-  
285 expressed partners, we used data from GeneFriends [64] (S4 Table). We observed a rather weak  
286 correlation between transcript length and the number of co-expression partners in our dataset  
287 (S9A Fig, Kendall Test,  $\tau = 0.10$ ,  $p\text{-value} < 2.2\text{E-}16$ ). However, despite this weak correlation,  
288 longer genes appeared to have more co-expressed gene partners than smaller genes (Fig 7A,  
289 Wilcoxon rank sum test,  $p\text{-value} < 2.2\text{E-}16$ , not-transformed figure in S9B Fig, median values of  
290 co-expression partners for longer genes = 2725, median values of co-expression partners for  
291 smaller genes = 32). We further analysed top and lowest hundred human co-expressed genes  
292 from the GeneFriends database (S4 Table) and observed that top highly co-expressed genes in  
293 the database have significantly higher transcript length (S9C Fig, Wilcoxon rank sum test,  $p\text{-}$   
294  $\text{value} = 0.00072$ , median = 3880) with respect to the bottom ones (median = 2587.5).

295

296 To determine if transcript length also influenced the number of protein-protein interactions, we  
297 used the protein-protein interaction data from BioGRID [65] (S5 Table). The results obtained  
298 were similar to the co-expression, where a weak correlation was observed between transcript  
299 length and the number of protein-protein interactions (S10A Fig, Kendall Test,  $\tau = 0.06$ ,  $p\text{-}$   
300  $\text{value} < 2.2\text{E-}16$ ).

301

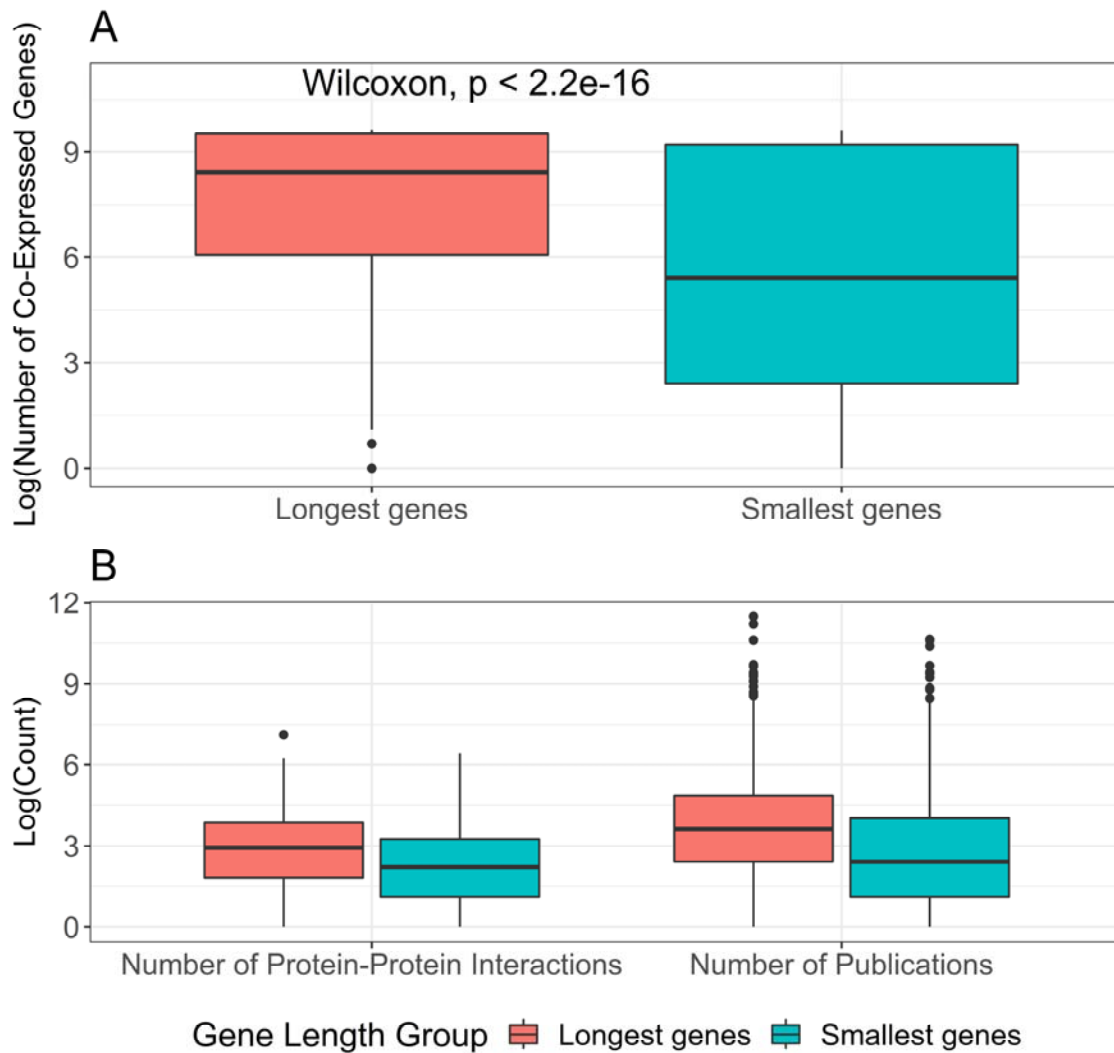
302 From such results, one would think that publication bias would have an effect on the number of  
303 interactions found. So, we obtained the number of publications for each gene studied here from  
304 PubMed and compared it to each gene length group and with the number of interactions (Fig

305 7B). We observed that the number of interactions and publications were significantly different  
306 between each gene length group (Wilcoxon rank sum test, p-value < 2.2E-16 for both  
307 comparisons), with both being higher for the group comprising of longer length genes. In order  
308 to assess the level of influence of publication bias in our protein-protein interaction dataset, we  
309 used correlations between the values of protein-protein interactions and the number of  
310 publications and we observed that, for both gene length groups, the correlations were not the  
311 strongest (Kendall test; Longest genes, tau = 0.26, p-value < 2.2E-16; Smallest genes, tau = 0.36,  
312 p-value < 2.2E-16), implying that while there might be some publication bias in effect, the  
313 strength of that effect is rather weak.

314

315 However, for the group of the longest genes, 208 (21%) entries were of zero value, while for the  
316 smallest group of genes, 544 (55%) entries were of zero value. This means that there were  
317 either no physical interactions for those genes, or that there were no entries in BioGRID for  
318 them. In order to account for this, and similarly to what we did for the co-expression analysis,  
319 we extracted the top 100 genes with the most and fewest protein-protein interactors (without  
320 null values) in our dataset and we observed the distribution of their transcript length. We  
321 observed that genes with the highest protein-protein interactions were longer (median  
322 transcript length = 3737), than genes with the lowest amount of protein-protein interactions  
323 (S10B Fig, Wilcoxon rank sum test, p-value = 0.039, median transcript length = 2764).

324



325

326

327 **Fig 7. Co-expression and protein-protein Interaction results pertaining to the longest and**  
328 **the smallest genes. The High group corresponds to the top 5% longest genes found in our**  
329 **original dataset ( $N_{\text{High}} = 986$ ), while the Low group corresponds to the top 5% smallest**  
330 **genes found in our original dataset ( $N_{\text{Low}} = 986$ ). **A:** Distribution of the Log transformed**  
331 **number of co-expressed genes for long genes and small genes. Number of co-expressed**  
332 **genes was obtained from data publicly available in GeneFriends [64]; **B:** Distribution of**  
333 **the number of protein-protein interactions and the number of publications for longer**  
334 **and smaller genes, all Log transformed. Number of protein-protein interactions was**  
335 **obtained from BioGRID [65] and the number of publications was obtained from PubMed.**

336

337

338

## 339 Discussion

340 With this work, we tried to elucidate what factors affected gene length and whether gene length  
341 had a role in determining the function of their proteins in the cell. Even looking at the 5 longest  
342 genes, we can get a small glimpse into one these objectives. *TTN* is the longest transcript in the  
343 human genome, and serves several important functions in the skeletal and cardiac muscles, and  
344 is often involved in structure, sensory and signalling responses [20,66,67]. The mucin *MUC16*  
345 (or CA125) is mostly known as a biomarker in ovarian cancer and is used to monitor patients as  
346 an indicator of cancer recurrence [21,68,69]. SMAD family member 2 (*SMAD2*) is thought to play  
347 a critical role in neuronal function [17] and to have a protective role in hepatic fibrosis [70]. The  
348 gene *FLRT2* is believed to have a role in tumour suppression in breast and prostate cancer  
349 [71,72] and, in mice models, *FLRT2* has been found as a guiding agent in neuronal and vascular  
350 cells [18,73]. For the *GRIN2B* gene, it has been shown to play an important role in the neuronal  
351 development and cell differentiation in the brain [19,74]. We cannot obtain any information at  
352 the moment pertaining to the function of the 5 smallest genes, since all of them are either novel  
353 and have yet to be properly studied, or could be annotation errors in the assembly.

354

355 In order to deeply understand the effects of gene length in protein function, we performed a  
356 functional analysis. For longer length genes, the GO terms obtained were mostly associated with  
357 neurons, for example terms like axon development, axon part, neuron to neuron synapse, actin  
358 and cell polarity [75] and GTPases [75]. For tissue specific genes, brain and nerve had the  
359 longest genes. Looking at the KEGG Pathways associated with the longest genes, the categories  
360 present are in the brain, cancer, heart diseases and muscle. Previous studies have associated  
361 longer length genes with neurons [76,77] and muscle [78]. Due to the very nature of longer  
362 genes, one expects high rates of mutation, not only due to their size, but also due to possible  
363 collisions between the RNA polymerase and the DNA polymerase, which causes instability and  
364 possible mutations [79]. It is not surprising to find associations between longer genes with

365 cancer [15] and hearth pathologies often caused by mutations in particularly long genes, like  
366 *DSC2* and *TTN* [80–82].

367 Looking at our smaller genes group, most of the GO terms provided were associated with the  
368 skin, for example skin development and cornified envelope, or with the immune system, for  
369 example, defence response to other organism and receptor agonist activity. Smaller tissue  
370 specific genes also have a major presence in the skin. With regards to the KEGG Pathways  
371 associated with the smaller genes, most pathways were involved in the immune system, with a  
372 few also being present in neurodegenerative diseases and in the skin. Previous studies have  
373 observed that most genes associated with immune functions are rather small in size [83].  
374 However, there are no studies to support the association of smaller genes with skin  
375 development. The categorization on the basis of published work has its advantages, but there is  
376 often overlapping of functions within these categories, for example, calcium signalling also  
377 happens in the muscle [84] and immune system [85], Wnt signalling pathway also has a role in  
378 cancer [86], TGF-beta signalling pathway can also be associated with the immune system [87],  
379 among others. In spite of this, our findings lead us to believe there is a disparity in gene sizes  
380 for genes that have a role or are present in tissues with very little to almost no development  
381 pos-natally (like neuron) and genes (not involved in housekeeping) that are quite frequently  
382 expressed during a human’s whole lifetime (like in skin development and immune response) or  
383 involved in providing functions with fast responses. Corroborating with our findings for the  
384 functional analysis, a recent preprint has showed that, with age, there is a downregulation of  
385 long transcripts and an upregulation of short transcripts, in a phenomena they named “length-  
386 driven transcriptome imbalance”, which in humans it affects the brain the most [88]. As we  
387 observed, smaller genes can be associated with the immune system and inflammation has a role  
388 in many ageing-related diseases [89], while longer genes are mostly associated with brain  
389 development, a function that happens early in life.

390

391 To understand whether there were factors that had an influence in gene length, we performed  
392 several correlation analysis. Overall there was no really strong correlation observed between  
393 the gene characteristics studied and transcript length. The biggest significant positive  
394 correlations were with protein size and number of SNPs, with transcript count, number of  
395 introns, GC content, and average gene expression having a weak significant positive correlation.  
396 Results of the correlation between average gene expression and transcript length were not in  
397 line with previous observations, which suggested that highly expressed genes are often smaller  
398 in length [11]. We also observed that among smaller genes, the average gene expression was, in  
399 fact, the highest (S4D Fig). However, genes with smaller lengths also had a great variability in  
400 the average gene expression values, and there was almost no correlation between transcript  
401 length and average gene expression. What has been stated in the previous studies is relevant,  
402 but the whole image is not captured properly. Rather than stating that the smaller genes are  
403 highly expressed, it is more accurate to say that smaller genes have a greater variability of levels  
404 of expression than longer genes. Similar to the correlation results for number of SNPs, both  
405 synonymous and missense mutations were also highly correlated with transcript length. It is  
406 particularly interesting that the correlation values were so high for missense mutations, since  
407 these may cause loss of function in the resulting protein. Likewise, it could be one of the reasons  
408 why the correlation between nonsense mutations and transcript length is weaker than the other  
409 two. Other works [9] have used the MIS/SYN and NONS/SYN ratios as a measure of functional  
410 importance, and we can, albeit faintly, observe here that longer genes appear to be more  
411 functionally important than smaller gene. The negative correlation between these ratios showed  
412 that longer genes may have more mechanisms in place to prevent loss of function mutations,  
413 when compared with synonymous mutations. Moreover, we also have to take account of  
414 “outliers” when looking into the correlation between transcript length and protein size (S4C  
415 Fig), specifically for longer genes. One would expect that for longer genes, the proteins produced  
416 would have a size comparable to their length and not be extremely small. However, after  
417 observing these outliers and we found that their protein size was rather small due to the



418 presence of very long 3'UTR regions. While these regions still account for the calculation of gene  
419 size, they are not translated into the protein, causing the presence of these “outliers”. Previous  
420 studies have shown that the brain has a preference for these long 3'UTR regions [90,91].

421

422 Interestingly, we also noticed that genes associated with ageing tend to be longer than the rest  
423 of the protein-coding genome. Moreover, we also showed that the overall (not tissue  
424 dependent) expression of genes with age appears to disregard transcript length, and that the  
425 brain seems to favour the expression of smaller genes with age. This last result, seems on par  
426 with the previously mentioned observations by Stoeger et al. [88], where they also witnessed  
427 the upregulation of smaller transcripts with age, especially in the brain. However, the results  
428 pertaining to the overall expression of genes with age seems to be different between what  
429 Stoeger et al. observed, with transcript length as an important source of ageing-dependent  
430 changes in values of expression, and what we observed based on Palmer et al. signatures of  
431 ageing [60], where transcript length does not influence the expression of genes with age. It is  
432 possible that these two works found two different sets of genes whose expression is affected in  
433 the ageing process. As such, further works should prove useful in dictating whether or not  
434 transcript length plays a major role in the expression of genes with age.

435

436 When comparing gene length with the dN/dS ratio for three organisms (Gorilla, Chimpanzee  
437 and Mouse), longer genes appeared to evolve under constraint, while for smaller genes there  
438 was a promotion for changes in the genes by natural selection. Previous studies have shown  
439 that, for genes classified as “old” (by virtue of having orthologues in older organisms), their  
440 length will be longer, they will have more introns and they evolve more slowly than smaller  
441 genes [7,8]. In terms of the co-expression analysis and protein-protein interactions, the longer  
442 genes, in general, had the most co-expression partners and protein-protein interactions. Further

443 validating our observations, we also saw that top hundred highest co-expression genes and PPI  
444 were longer in length as compared to lowest co-expression genes and PPI.

445

446 As a result of this work we have noticed that not all genes are studied with the same depth.  
447 Some genes have more information related to expression or function than others. We observed  
448 this especially within our 5% list of longest and smallest genes. Longer length genes had more  
449 functional information readily available than smaller ones. We can also observe that in the  
450 publication bias analysis for protein-protein interactions, where genes with longer lengths had  
451 more publications than smaller genes. Indeed, other groups have found that gene length can be  
452 an important predictor of the number of publications, and that novel genes are not often studied  
453 to their full capacity [92], while others have found that genetic associations tend to be more  
454 biased towards longer genes [93,94].

455

456 The present study has its own limitations. One of the limitations for this sort of study is that, the  
457 results might be “time-specific”. With new discoveries related to the human genome and its  
458 genes, the trends here observed might change, specifically when it concerns the currently  
459 extremely untapped field of smaller genes. Similarly as we previously noted, longer genes have a  
460 lot more information related to them, when compared with their smaller counterparts. While  
461 our findings with respect to the longer genes might be mostly reliable, we cannot show the same  
462 confidence in case of the smaller genes, considering that a lot of these genes were novel and  
463 have yet to be properly studied. However even after taking account of the above limitations, the  
464 present study still provides some very interesting insights pertaining to gene length and its  
465 possible role in early life development, diseases and response time in the human genome.

## 466 **Conclusion**

467 With this work we aimed to better understand the effects of gene length in gene function and  
468 factors that affected it. We observed that, for most of the factors studied, there was not a  
469 particularly strong correlation with transcript length. The strongest correlations here detected  
470 were associated with the number of SNPs and the protein size. We also showed that, for smaller  
471 genes, its association with high levels of expression is not entirely correct and that, instead,  
472 there is great variability of expression values among them. We also observed that longer genes  
473 appear to have the most co-expression partners and protein-protein interactions, in comparison  
474 to their smaller counterparts.

475 In case of the functional analysis, we observed that longer genes favoured functions in the brain,  
476 cancer, heart and muscle, while smaller genes are strongly associated with the immune system,  
477 skin and neurodegenerative diseases. This lead us to believe that gene length could be  
478 associated with the frequency of usage of the gene, with longer genes being less often used past  
479 the initial development and smaller genes playing a frequent role daily in the human body.

480

481

482

483

## 484 **Methods**

### 485 **Data retrieval and filtering**

486 All protein-coding human transcripts and genes ( $N_{\text{transcripts}} = 92696$ ), their length, transcript  
487 count and GC content were obtained using the biomart [16] website (GRCh38.p12, Ensembl 96,  
488 April 2019). Transcript length is defined by Ensembl as the total length of the exons in a gene  
489 plus its UTR regions lengths. Gene length was obtained using the R (version 3.5.2) package

490 EDASeq (version 2.14.1). Using R, the transcripts with the highest transcript length per gene  
491 were selected. In case of ties, due to multiple transcript having the same length per gene, we  
492 used some tags (APPRIS annotation was the principal one, if there was an entry in RefSeq or  
493 GENCODE) used by ensemble as a tie-breaker. Should that fail, the oldest transcript was chosen,  
494 by means of having a smaller numerical ID. Transcripts associated with PATCH locations or  
495 assemblies were removed from our dataset. For each transcript, we obtained data regarding  
496 their number of exons, CDS length, number of SNPs, synonymous (“synonymous\_variant”),  
497 missense (“missense\_variant”) and nonsense (“stop\_gained”) SNPs, protein length, dN and dS  
498 values, using the biomart (version 2.38.0) package in R. For the dN and dS values, only values  
499 associated with One to One orthologues were selected for the present analysis. Average  
500 expression was obtained from the UCSC Table browser tool [95], using expression as the group  
501 and the GTEx Gene track. Tissue specific Tau values of expression were obtained from a  
502 previous work [60]. The number of SNPs per gene was obtained using the Ensembl API, R and  
503 the httr (version 1.4.0) and jsonlite (version 1.6) packages.

504 The whole file produced and used in the analysis for this work can be found on the  
505 Supplementary Table 1 (N = 19714).

506 Gene names of genes related with ageing (N = 307) were obtained from GenAge (Build 19) [62].

507

## 508 **Statistical tests, graphs and other packages**

509 R and the function `corr.test` were used to perform the correlation tests. Due to the abundance of  
510 the data, there were a lot of ties in the ranks, which prevented the usage of Spearman’s  
511 correlation, so instead we chose to use the Kendall test for the correlations. The figures  
512 produced in this work were created using the `ggplot2` (version 3.2.0) package in R. Other  
513 packages used over the course of this work were: `corrplot` (version 0.84), `psych` (version  
514 1.8.12), `ggpubr` (version 0.2.1), `stringr` (version 1.4.0), `dplyr` (version 8.0.1), `plyr` (version 1.8.4)  
515 and `tidyr` (version 0.8.3).

516

## 517 **Functional Analysis**

518 WebGestalt (2019 release) [22] was used to do the Overrepresentation Enrichment Analysis for  
519 each of the gene ontology categories (Biological Process, Cellular Component and Molecular  
520 Function). The top 5% genes, with the highest and lowest gene length, were ran against the  
521 reference option of genome. The significance level was  $FDR < 0.05$  and the multiple test  
522 adjustment was done using the Benjamini–Hochberg method.

523 For confirmation of the results, the same two 5% lists were run on DAVID's [23,24] annotation  
524 clustering option, using the complete human genome as background. Only terms with p-value  
525 and FDR smaller or equal to 0.05 were considered. Default categories were used except for the  
526 category "UP\_SEQ\_FEATURE", since it was introducing a lot of redundant results.

527 To help better visualize the GO terms obtained from the analysis above described, the tool  
528 REViGO [59] was used. The p-values here considered were the FDR values obtained previously,  
529 with the human database option used for the GO terms.

530 In regards to the analysis done using the KEGG pathways, the grouping of genes and pathways  
531 was obtained from the Molecular Signature Database (version 6.2) [26,27,96–99], like it was  
532 done previously by another group [15]. Additionally, the colouring of the box plot was done  
533 based on the fact that the pathway in question is directly associated with the category (when  
534 the KEGG Pathway schematic shows cells from the category) or if they could be indirectly  
535 associated with the category (using available literature). For this last case, appropriate  
536 literature was selected if they mentioned elements of the KEGG Pathway being involved in said  
537 category.

538

### 539 **Co-Expression Analysis**

540 Co-expression correlation values were extracted from GeneFriends [64]. For each gene (N =  
541 19714), in the whole dataset and in the top 5% lists of genes with the longest and smallest  
542 transcript length (N = 986 for each list), the number of genes with correlation values superior or  
543 equal to 0.6 or smaller or equal to -0.6 were obtained using R. From our original dataset  
544 (N=19714 genes), 1046 genes were not present in GeneFriends (whole dataset), of which, 25  
545 missing genes were within the High group and 110 missing genes were within the Low group.

546 For obtaining the median values of genes present in the GeneFriends database, the co-  
547 expression values for each gene across the database were merged and this was followed by  
548 calculation of median values using R.

549

### 550 **Protein-Protein Interaction Analysis**

551 BioGRID (release 3.5.174) REST API [65] in conjugation with the R package httr was used to  
552 obtain all protein-protein interactions for the whole dataset and for the top 5% longest and  
553 smallest genes. All redundant and genetic interactions were removed from this analysis.

554 For the publication bias, the number of publications, in PubMed, per gene of each group was  
555 obtained using the Entrez Programming Utilities (E-utilities), and the R packages XML (version  
556 3.98-1.19), httr and biomaRt.

557

558

559

560 **Acknowledgements**

561 The authors wish to thank past and present members of the Integrative Genomics of Ageing

562 Group for useful suggestions and discussion, in particular Kasit Chatsirisupachai and Daniel

563 Palmer.

564

## 565 **References**

- 566 1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing  
567 and analysis of the human genome. *Nature*. 2001;409: 860–921. doi:10.1038/35057062
- 568 2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The Sequence of the  
569 Human Genome. *Science*. 2001;291: 1304–1351. doi:10.1126/science.1058040
- 570 3. International Human Genome Sequencing Consortium. Finishing the euchromatic  
571 sequence of the human genome. *Nature*. 2004;431: 931–45. doi:10.1038/nature03001
- 572 4. Gonzaga-Jauregui C, Lupski JR, Gibbs RA. Human genome sequencing in health and  
573 disease. *Annu Rev Med*. 2012;63: 35–61. doi:10.1146/annurev-med-051010-162644
- 574 5. Goldfeder RL, Wall DP, Khoury MJ, Ioannidis JPA, Ashley EA. Human Genome Sequencing  
575 at the Population Scale: A Primer on High-Throughput DNA Sequencing and Analysis. *Am*  
576 *J Epidemiol*. 2017;186: 1000–1009. doi:10.1093/aje/kww224
- 577 6. Simonti CN, Capra JA. The evolution of the human genome. *Curr Opin Genet Dev*.  
578 2015;35: 9–15. doi:10.1016/j.gde.2015.08.005
- 579 7. Vishnoi A, Kryazhimskiy S, Bazykin GA, Hannenhalli S, Plotkin JB. Young proteins  
580 experience more variable selection pressures than old proteins. *Genome Res*. 2010;20:  
581 1574–81. doi:10.1101/gr.109595.110
- 582 8. Wolf YI, Novichkov PS, Karev GP, Koonin E V., Lipman DJ. The universal distribution of  
583 evolutionary rates of genes and distinct characteristics of eukaryotic genes of different  
584 apparent ages. *Proc Natl Acad Sci*. 2009;106: 7273–7280. doi:10.1073/pnas.0901808106
- 585 9. Gorlova O, Fedorov A, Logothetis C, Amos C, Gorlov I. Genes with a large intronic burden  
586 show greater evolutionary conservation on the protein level. *BMC Evol Biol*. 2014;14: 50.  
587 doi:10.1186/1471-2148-14-50
- 588 10. Grishkevich V, Yanai I. Gene length and expression level shape genomic novelties.



- 589 Genome Res. 2014;24: 1497–503. doi:10.1101/gr.169722.113
- 590 11. Urrutia AO, Hurst LD. The signature of selection mediated by expression on human genes.  
591 Genome Res. 2003;13: 2260–4. doi:10.1101/gr.641103
- 592 12. Eisenberg E, Levanon EY. Human housekeeping genes are compact. Trends Genet.  
593 2003;19: 362–365. doi:10.1016/S0168-9525(03)00140-9
- 594 13. Kirkconnell KS, Magnuson B, Paulsen MT, Lu B, Bedi K, Ljungman M. Gene length as a  
595 biological timer to establish temporal transcriptional regulation. Cell Cycle. 2017;16:  
596 259–270. doi:10.1080/15384101.2016.1234550
- 597 14. Yang D, Xu A, Shen P, Gao C, Zang J, Qiu C, et al. A two-level model for the role of complex  
598 and young genes in the formation of organism complexity and new insights into the  
599 relationship between evolution and development. Evodevo. 2018;9: 22.  
600 doi:10.1186/s13227-018-0111-4
- 601 15. Sahakyan AB, Balasubramanian S. Long genes and genes with multiple splice variants are  
602 enriched in pathways linked to cancer and other multigenic diseases. BMC Genomics.  
603 2016;17: 225. doi:10.1186/s12864-016-2582-9
- 604 16. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018.  
605 Nucleic Acids Res. 2018;46: D754–D761. doi:10.1093/nar/gkx1098
- 606 17. Tao S, Sampath K. Alternative splicing of SMADs in differentiation and tissue  
607 homeostasis. Dev Growth Differ. 2010;52: 335–342. doi:10.1111/j.1440-  
608 169X.2009.01163.x
- 609 18. Yamagishi S, Hampel F, Hata K, del Toro D, Schwark M, Kvachnina E, et al. FLRT2 and  
610 FLRT3 act as repulsive guidance cues for Unc5-positive neurons. EMBO J. 2011;30: 2920–  
611 2933. doi:10.1038/emboj.2011.189
- 612 19. Hu C, Chen W, Myers SJ, Yuan H, Traynelis SF. Human GRIN2B variants in

- 613 neurodevelopmental disorders. *J Pharmacol Sci.* 2016;132: 115–121.  
614 doi:10.1016/j.jphs.2016.10.002
- 615 20. Ware JS, Cook SA. Role of titin in cardiomyopathy: from DNA variants to patient  
616 stratification. *Nat Rev Cardiol.* 2017;15: 241–252. doi:10.1038/nrcardio.2017.190
- 617 21. Felder M, Kapur A, Gonzalez-Bosquet J, Horibata S, Heintz J, Albrecht R, et al. MUC16  
618 (CA125): tumor biomarker to cancer therapy, a work in progress. *Mol Cancer.* 2014;13:  
619 129. doi:10.1186/1476-4598-13-129
- 620 22. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with  
621 revamped UIs and APIs. *Nucleic Acids Res.* 2019;47: W199–W205.  
622 doi:10.1093/nar/gkz401
- 623 23. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene  
624 lists using DAVID bioinformatics resources. *Nat Protoc.* 2009;4: 44–57.  
625 doi:10.1038/nprot.2008.211
- 626 24. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the  
627 comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 2009;37: 1–13.  
628 doi:10.1093/nar/gkn923
- 629 25. Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*  
630 2000;28: 27–30. doi:10.1093/nar/28.1.27
- 631 26. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set  
632 enrichment analysis: A knowledge-based approach for interpreting genome-wide  
633 expression profiles. *Proc Natl Acad Sci.* 2005;102: 15545–15550.  
634 doi:10.1073/pnas.0506580102
- 635 27. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular  
636 Signatures Database Hallmark Gene Set Collection. *Cell Syst.* 2015;1: 417–425.  
637 doi:10.1016/j.cels.2015.12.004

- 638 28. Kerrisk ME, Cingolani LA, Koleske AJ. ECM receptors in neuronal structure, synaptic  
639 plasticity, and behavior. *Prog Brain Res.* 2014;214: 101–31. doi:10.1016/B978-0-444-  
640 63486-3.00005-0
- 641 29. Lin T, Islam O, Heese K. ABC transporters, neural stem cells and neurogenesis – a  
642 different perspective. *Cell Res.* 2006;16: 857–871. doi:10.1038/sj.cr.7310107
- 643 30. Schnaar RL. Gangliosides of the Vertebrate Nervous System. *J Mol Biol.* 2016;428: 3325–  
644 3336. doi:10.1016/j.jmb.2016.05.020
- 645 31. Bauer H-C, Krizbai IA, Bauer H, Traweger A. “You Shall Not Pass”-tight junctions of the  
646 blood brain barrier. *Front Neurosci.* 2014;8: 392. doi:10.3389/fnins.2014.00392
- 647 32. Lasky JL, Wu H. Notch Signaling, Brain Development, and Human Disease. *Pediatr Res.*  
648 2005;57: 104R-109R. doi:10.1203/01.PDR.0000159632.70510.3D
- 649 33. Kwok JCF, Warren P, Fawcett JW. Chondroitin sulfate: A key molecule in the brain matrix.  
650 *Int J Biochem Cell Biol.* 2012;44: 582–586. doi:10.1016/j.biocel.2012.01.004
- 651 34. Russo D, Della Ragione F, Rizzo R, Sugiyama E, Scalabrì F, Hori K, et al. Glycosphingolipid  
652 metabolic reprogramming drives neural differentiation. *EMBO J.* 2018;37: e97674.  
653 doi:10.15252/embj.201797674
- 654 35. Massaly N, Francès B, Moulédous L. Roles of the ubiquitin proteasome system in the  
655 effects of drugs of abuse. *Front Mol Neurosci.* 2014;7: 99. doi:10.3389/fnmol.2014.00099
- 656 36. Zeng Y, Zhang L, Hu Z. Cerebral insulin, insulin signaling pathway, and brain  
657 angiogenesis. *Neurol Sci.* 2016;37: 9–16. doi:10.1007/s10072-015-2386-8
- 658 37. Funderburgh JL. Keratan Sulfate Biosynthesis. *IUBMB Life (International Union Biochem*  
659 *Mol Biol Life).* 2002;54: 187–194. doi:10.1080/15216540214932
- 660 38. Noelanders R, Vleminckx K. How Wnt Signaling Builds the Brain: Bridging Development  
661 and Disease. *Neurosci.* 2017;23: 314–329. doi:10.1177/1073858416667270

- 662 39. Dermietzel R, Spray DC. Gap junctions in the brain: where, what type, how many and  
663 why? *Trends Neurosci.* 1993;16: 186–192. doi:10.1016/0166-2236(93)90151-B
- 664 40. Grube M, Hagen P, Jedlitschky G. Neurosteroid Transport in the Brain: Role of ABC and  
665 SLC Transporters. *Front Pharmacol.* 2018;9. doi:10.3389/fphar.2018.00354
- 666 41. Monje FJ, Kim E-J, Pollak DD, Cabatic M, Li L, Baston A, et al. Focal Adhesion Kinase  
667 Regulates Neuronal Growth, Synaptic Plasticity and Hippocampus-Dependent Spatial  
668 Learning and Memory. *Neurosignals.* 2012;20: 1–14. doi:10.1159/000330193
- 669 42. Frere SG, Chang-Ileto B, Di Paolo G. Role of phosphoinositides at the neuronal synapse.  
670 *Subcell Biochem.* 2012;59: 131–75. doi:10.1007/978-94-007-3015-1\_5
- 671 43. Dickson EJ. Recent advances in understanding phosphoinositide signaling in the nervous  
672 system. *F1000Research.* 2019;8. doi:10.12688/f1000research.16679.1
- 673 44. Fisher SK, Novak JE, Agranoff BW. Inositol and higher inositol phosphates in neural  
674 tissues: homeostasis, metabolism and functional significance. *J Neurochem.* 2002;82:  
675 736–754. doi:10.1046/j.1471-4159.2002.01041.x
- 676 45. Stocker AM, Chenn A. The role of adherens junctions in the developing neocortex. *Cell*  
677 *Adh Migr.* 2015;9: 167–174. doi:10.1080/19336918.2015.1027478
- 678 46. Mei L, Nave K-A. Neuregulin-ERBB signaling in the nervous system and neuropsychiatric  
679 diseases. *Neuron.* 2014;83: 27–49. doi:10.1016/j.neuron.2014.06.007
- 680 47. Russo E, Citraro R, Constanti A, De Sarro G. The mTOR Signaling Pathway in the Brain:  
681 Focus on Epilepsy and Epileptogenesis. *Mol Neurobiol.* 2012;46: 662–681.  
682 doi:10.1007/s12035-012-8314-5
- 683 48. Ogretmen B. Sphingolipid metabolism in cancer signalling and therapy. *Nat Rev Cancer.*  
684 2018;18: 33–50. doi:10.1038/nrc.2017.96
- 685 49. Zhang T, de Waard AA, Wuhrer M, Spaapen RM. The Role of Glycosphingolipids in

- 686 Immune Cell Functions. *Front Immunol.* 2019;10. doi:10.3389/fimmu.2019.00090
- 687 50. Prentki M, Madiraju SRM. Glycerolipid Metabolism and Signaling in Health and Disease.  
688 *Endocr Rev.* 2008;29: 647–676. doi:10.1210/er.2008-0007
- 689 51. Seif F, Khoshmirsafa M, Aazami H, Mohsenzadegan M, Sedighi G, Bahar M. The role of  
690 JAK-STAT signaling pathway and its regulators in the fate of T helper cells. *Cell Commun*  
691 *Signal.* 2017;15: 23. doi:10.1186/s12964-017-0177-y
- 692 52. Le Floc'h N, Otten W, Merlot E. Tryptophan metabolism, from nutrition to potential  
693 therapeutic applications. *Amino Acids.* 2011;41: 1195–1205. doi:10.1007/s00726-010-  
694 0752-7
- 695 53. Barber GN. STING-dependent cytosolic DNA sensing pathways. *Trends Immunol.*  
696 2014;35: 88–93. doi:10.1016/j.it.2013.10.010
- 697 54. Taylor RG, Levy HL, McInnes RR. Histidase and histidinemia. Clinical and molecular  
698 considerations. *Mol Biol Med.* 1991;8: 101–16. Available:  
699 <http://www.ncbi.nlm.nih.gov/pubmed/1943682>
- 700 55. Ziboh VA, Miller CC, Cho Y. Metabolism of polyunsaturated fatty acids by skin epidermal  
701 enzymes: generation of antiinflammatory and antiproliferative metabolites. *Am J Clin*  
702 *Nutr.* 2000;71: 361s-366s. doi:10.1093/ajcn/71.1.361s
- 703 56. Fisher GJ, Voorhees JJ. Molecular mechanisms of retinoid actions in skin. *FASEB J.*  
704 1996;10: 1002–1013. doi:10.1096/fasebj.10.9.8801161
- 705 57. Iversen L, Kragballe K. Arachidonic acid metabolism in skin health and disease.  
706 *Prostaglandins Other Lipid Mediat.* 2000;63: 25–42. doi:10.1016/S0090-  
707 6980(00)00095-2
- 708 58. Slominski A, Zbytek B, Nikolakis G, Manna PR, Skobowiat C, Zmijewski M, et al.  
709 Steroidogenesis in the skin: Implications for local immune functions. *J Steroid Biochem*

- 710 Mol Biol. 2013;137: 107–123. doi:10.1016/j.jsbmb.2013.02.006
- 711 59. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO Summarizes and Visualizes Long Lists of  
712 Gene Ontology Terms. Gibas C, editor. PLoS One. 2011;6: e21800.  
713 doi:10.1371/journal.pone.0021800
- 714 60. Palmer D, Fabris F, Doherty A, Freitas AA, de Magalhães JP. Ageing Transcriptome Meta-  
715 Analysis Reveals Similarities Between Key Mammalian Tissues. bioRxiv [Preprint]. 2019;  
716 815381. doi:10.1101/815381
- 717 61. Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, et al. Genome-wide  
718 midrange transcription profiles reveal expression level relationships in human tissue  
719 specification. *Bioinformatics*. 2005;21: 650–659. doi:10.1093/bioinformatics/bti042
- 720 62. Tacutu R, Thornton D, Johnson E, Budovsky A, Barardo D, Craig T, et al. Human Ageing  
721 Genomic Resources: new and updated databases. *Nucleic Acids Res*. 2018;46: D1083–  
722 D1090. doi:10.1093/nar/gkx1042
- 723 63. van Dam S, Vösa U, van der Graaf A, Franke L, de Magalhães JP. Gene co-expression  
724 analysis for functional classification and gene-disease predictions. *Brief Bioinform*.  
725 2018;19: 575–592. doi:10.1093/bib/bbw139
- 726 64. van Dam S, Craig T, de Magalhães JP. GeneFriends: a human RNA-seq-based gene and  
727 transcript co-expression database. *Nucleic Acids Res*. 2015;43: D1124–D1132.  
728 doi:10.1093/nar/gku1042
- 729 65. Stark C. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*.  
730 2006;34: D535–D539. doi:10.1093/nar/gkj109
- 731 66. Chauveau C, Rowell J, Ferreiro A. A Rising Titan: TTN Review and Mutation Update. *Hum*  
732 *Mutat*. 2014;35: 1046–1059. doi:10.1002/humu.22611
- 733 67. Savarese M, Sarparanta J, Vihola A, Udd B, Hackman P. Increasing Role of Titin Mutations

- 734 in Neuromuscular Disorders. *J Neuromuscul Dis.* 2016;3: 293–308. doi:10.3233/JND-  
735 160158
- 736 68. Haridas D, Ponnusamy MP, Chugh S, Lakshmanan I, Seshacharyulu P, Batra SK. MUC16:  
737 molecular analysis and its functional implications in benign and malignant conditions.  
738 *FASEB J.* 2014;28: 4183–4199. doi:10.1096/fj.14-257352
- 739 69. Das S, Batra SK. Understanding the Unique Attributes of MUC16 (CA125): Potential  
740 Implications in Targeted Therapy. *Cancer Res.* 2015;75: 4669–4674. doi:10.1158/0008-  
741 5472.CAN-15-1050
- 742 70. Xu F, Liu C, Zhou D, Zhang L. TGF- $\beta$ /SMAD Pathway and Its Regulation in Hepatic  
743 Fibrosis. *J Histochem Cytochem.* 2016;64: 157–167. doi:10.1369/0022155415627681
- 744 71. Bae H, Kim B, Lee H, Lee S, Kang H-S, Kim SJ. Epigenetically regulated Fibronectin leucine  
745 rich transmembrane protein 2 (FLRT2) shows tumor suppressor activity in breast cancer  
746 cells. *Sci Rep.* 2017;7: 272. doi:10.1038/s41598-017-00424-0
- 747 72. Wu Y, Davison J, Qu X, Morrissey C, Storer B, Brown L, et al. Methylation profiling  
748 identified novel differentially methylated markers including OPCML and FLRT2 in  
749 prostate cancer. *Epigenetics.* 2016;11: 247–258. doi:10.1080/15592294.2016.1148867
- 750 73. Seiradake E, del Toro D, Nagel D, Cop F, Härtl R, Ruff T, et al. FLRT Structure: Balancing  
751 Repulsion and Cell Adhesion in Cortical and Vascular Development. *Neuron.* 2014;84:  
752 370–385. doi:10.1016/j.neuron.2014.10.008
- 753 74. Bell S, Maussion G, Jefri M, Peng H, Theroux J-F, Silveira H, et al. Disruption of GRIN2B  
754 Impairs Differentiation in Human Neurons. *Stem Cell Reports.* 2018;11: 183–196.  
755 doi:10.1016/j.stemcr.2018.05.018
- 756 75. Polleux F, Snider W. Initiating and Growing an Axon. *Cold Spring Harb Perspect Biol.*  
757 2010;2: a001925–a001925. doi:10.1101/cshperspect.a001925

- 758 76. Zylka MJ, Simon JM, Philpot BD. Gene Length Matters in Neurons. *Neuron*. 2015;86: 353–  
759 355. doi:10.1016/j.neuron.2015.03.059
- 760 77. Takeuchi A, Iida K, Tsubota T, Hosokawa M, Denawa M, Brown JB, et al. Loss of Sfpq  
761 Causes Long-Gene Transcriptopathy in the Brain. *Cell Rep*. 2018;23: 1326–1341.  
762 doi:10.1016/j.celrep.2018.03.141
- 763 78. Hosokawa M, Takeuchi A, Tanihata J, Iida K, Takeda S, Hagiwara M. Loss of RNA-Binding  
764 Protein Sfpq Causes Long-Gene Transcriptopathy in Skeletal Muscle and Severe Muscle  
765 Mass Reduction with Metabolic Myopathy. *iScience*. 2019;13: 229–242.  
766 doi:10.1016/j.isci.2019.02.023
- 767 79. Helmrich A, Ballarino M, Tora L. Collisions between Replication and Transcription  
768 Complexes Cause Common Fragile Site Instability at the Longest Human Genes. *Mol Cell*.  
769 2011;44: 966–977. doi:10.1016/j.molcel.2011.10.013
- 770 80. Corrado D, Link MS, Calkins H. Arrhythmogenic Right Ventricular Cardiomyopathy.  
771 Jarcho JA, editor. *N Engl J Med*. 2017;376: 61–72. doi:10.1056/NEJMra1509267
- 772 81. Maron BJ, Maron MS. Hypertrophic cardiomyopathy. *Lancet*. 2013;381: 242–255.  
773 doi:10.1016/S0140-6736(12)60397-3
- 774 82. Jefferies JL, Towbin JA. Dilated cardiomyopathy. *Lancet*. 2010;375: 752–762.  
775 doi:10.1016/S0140-6736(09)62023-7
- 776 83. Pipkin ME, Monticelli S. Genomics and the immune system. *Immunology*. 2008;124: 23–  
777 32. doi:10.1111/j.1365-2567.2008.02818.x
- 778 84. Kuo IY, Ehrlich BE. Signaling in Muscle Contraction. *Cold Spring Harb Perspect Biol*.  
779 2015;7: a006023. doi:10.1101/cshperspect.a006023
- 780 85. Vig M, Kinet J-P. Calcium signaling in immune cells. *Nat Immunol*. 2009;10: 21–27.  
781 doi:10.1038/ni.f.220



- 782 86. Zhan T, Rindtorff N, Boutros M. Wnt signaling in cancer. *Oncogene*. 2017;36: 1461–1473.  
783 doi:10.1038/onc.2016.304
- 784 87. Worthington JJ, Fenton TM, Czajkowska BI, Klementowicz JE, Travis MA. Regulation of  
785 TGF $\beta$  in the immune system: An emerging role for integrins and dendritic cells.  
786 *Immunobiology*. 2012;217: 1259–1265. doi:10.1016/j.imbio.2012.06.009
- 787 88. Stoeger T, Grant RA, McQuattie-Pimentel AC, Anekalla K, Liu SS, Tejedor-Navarro H, et al.  
788 Aging is associated with a systemic length-driven transcriptome imbalance. *bioRxiv*  
789 [Preprint]. 2019; 691154. doi:10.1101/691154
- 790 89. Goldberg EL, Dixit VD. Drivers of age-related inflammation and strategies for healthspan  
791 extension. *Immunol Rev*. 2015;265: 63–74. doi:10.1111/imr.12295
- 792 90. Wang L, Yi R. 3'UTRs take a long shot in the brain. *BioEssays*. 2014;36: 39–45.  
793 doi:10.1002/bies.201300100
- 794 91. Miura P, Shenker S, Andreu-Agullo C, Westholm JO, Lai EC. Widespread and extensive  
795 lengthening of 3' UTRs in the mammalian brain. *Genome Res*. 2013;23: 812–825.  
796 doi:10.1101/gr.146886.112
- 797 92. Stoeger T, Gerlach M, Morimoto RI, Nunes Amaral LA. Large-scale investigation of the  
798 reasons why potentially important genes are ignored. Freeman T, editor. *PLoS Biol*.  
799 2018;16: e2006643. doi:10.1371/journal.pbio.2006643
- 800 93. de Magalhães JP, Wang J. The fog of genetics: what is known, unknown and unknowable  
801 in the genetics of complex traits and diseases. *EMBO Rep*. 2019; e48054.  
802 doi:10.15252/embr.201948054
- 803 94. Mirina A, Atzmon G, Ye K, Bergman A. Gene Size Matters. *PLoS One*. 2012;7: e49093.  
804 doi:10.1371/journal.pone.0049093
- 805 95. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, et al. The UCSC

- 806           Table Browser data retrieval tool. *Nucleic Acids Res.* 2004;32: D493-6.  
807           doi:10.1093/nar/gkh103
- 808   96.   Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP.  
809           Molecular signatures database (MSigDB) 3.0. *Bioinformatics.* 2011;27: 1739–1740.  
810           doi:10.1093/bioinformatics/btr260
- 811   97.   Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*  
812           2000;28: 27–30. doi:10.1093/nar/28.1.27
- 813   98.   Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on  
814           genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 2017;45: D353–D361.  
815           doi:10.1093/nar/gkw1092
- 816   99.   Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for  
817           understanding genome variations in KEGG. *Nucleic Acids Res.* 2019;47: D590–D595.  
818           doi:10.1093/nar/gky962
- 819  
820  
821  
822  
823  
824  
825  
826  
827  
828

829

830

831

## 832 **Supporting information**

833 **S1 Table. Dataset with the highest protein-coding transcript length per Gene, in human.**

834 **S2 Table. Functional analysis results for WebGestalt and DAVID.**

835 **S3 Table. KEGG Pathway IDs used in Supplementary Figure 2.**

836 **S4 Table. Co-Expression results.**

837 **S5 Table. Number of Protein-Protein interactions and Publications in Pubmed for each**

838 **gene in the dataset.**

839 **S1 Fig. Functional analysis results for Cellular Component and Molecular Function.**

840 **S2 Fig. Transcript length distribution per KEGG Pathway.**

841 **S3 Fig. Correlation results for Number of SNPs, protein size, transcript count, GC content**

842 **and synonymous, missense and nonsense mutations against transcript length.**

843 **S4 Fig. Gene length and intron distribution in the human genome.**

844 **S5 Fig. Transcript length distribution for genes specifically expressed in the given tissues.**

845 **S6 Fig. Transcript length distribution for ageing related genes and for the rest of the**

846 **dataset.**

847 **S7 Fig. Evolution results for mouse, gorilla and chimpanzee.**

848 **S8 Fig. Co-expression results.**

849 **S9 Fig. Protein-protein interactions results.**

