frontiers in Genetics

ORIGINAL RESEARCH
published: 11 February 2021
doi: 10.3389/fgene.2021.559998

Gene Size Matters: An Analysis of Gene Length in the Human Genome

*Inês Lopes, Gulam Altab, Priyanka Raina and João Pedro de Magalhães**

Integrative Genomics of Ageing Group, Institute of Ageing and Chronic Disease, University of Liverpool, Liverpool, United Kingdom

While it is expected for gene length to be associated with factors such as intron number and evolutionary conservation, we are yet to understand the connections between gene length and function in the human genome. In this study, we show that, as expected, there is a strong positive correlation between gene length, transcript length, and protein size as well as a correlation with the number of genetic variants and introns. Among tissue-specific genes, we find that the longest transcripts tend to be expressed in the blood vessels, nerves, thyroid, cervix uteri, and the brain, while the smallest transcripts tend to be expressed in the pancreas, skin, stomach, vagina, and testis. We report, as shown previously, that natural selection suppresses changes for genes with longer transcripts and promotes changes for genes with smaller transcripts. We also observe that genes with longer transcripts tend to have a higher number of co-expressed genes and protein-protein interactions, as well as more associated publications. In the functional analysis, we show that bigger transcripts are often associated with neuronal development, while smaller transcripts tend to play roles in skin development and in the immune system. Furthermore, pathways related to cancer, neurons, and heart diseases tend to have genes with longer transcripts, with smaller transcripts being present in pathways related to immune responses and neurodegenerative diseases. Based on our results, we hypothesize that longer genes tend to be associated with functions that are important in the early development stages, while smaller genes tend to play a role in functions that are important throughout the whole life, like the immune system, which requires fast responses.

Keywords: genomics, transcripts, gene expression, immune system, mRNA, SNPs

OPEN ACCESS

Edited by:
Kimberly Glass,
Harvard Medical School, United States

Reviewed by:
Jay Brown,
University of Virginia, United States
Daniel Rico,
Newcastle University, United Kingdom

****Correspondence:***
João Pedro de Magalhães
jp@senescence.info;
aging@liverpool.ac.uk

Specialty section:
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

Received: *07 May 2020*
Accepted: *06 January 2021*
Published: *11 February 2021*

Citation:
Lopes I, Altab G, Raina P and de Magalhães JP (2021) Gene Size Matters: An Analysis of Gene Length in the Human Genome. Front. Genet. 12:559998. doi: 10.3389/fgene.2021.559998

INTRODUCTION

With the sequencing of the human genome (Lander et al., 2001; Venter et al., 2001; International Human Genome Sequencing Consortium, 2004) arose a great interest in understanding the relationship between genotype and phenotype, especially concerning human health (Gonzaga-Jauregui et al., 2012; Goldfeder et al., 2017). However, despite the recent advancements, we have yet to fully understand the human genome and its complexity (Simonti and Capra, 2015).

Several studies have tried to decipher the connections between the length of a gene and its functions. It is believed that genes that are more evolutionarily conserved are often associated with longer gene length and higher intronic burden (Wolf et al., 2009; Vishnoi et al., 2010; Gorlova et al., 2014; Grishkevich and Yanai, 2014). In contrast, a shorter gene length is associated with a high expression, smaller proteins, and little intronic content (Urrutia and

Hurst, 2003). This hypothesis is further supported by housekeeping genes, which are widely expressed and have characteristics similar to genes with shorter length (Eisenberg and Levanon, 2003). It was hypothesized that, due to the great levels of expression in smaller genes, there is selective pressure to maximize protein synthesis efficiency (Urrutia and Hurst, 2003). If that is the case, then the question remains regarding which functions benefit longer genes to compensate for their more expensive production of proteins.

Gene length has been associated with biological timing. In response to stimuli, smaller genes produce proteins faster, and these smaller proteins often play a part in the regulation of longer proteins, which, in turn, are expressed later in the response. This allows for regulatory mechanisms to be set up in preparation for the expression of important proteins (Kirkconnell et al., 2017). Indeed, longer genes have been associated with important biological processes, including embryonic development (Yang et al., 2018) and neuronal processes (Sahakyan and Balasubramanian, 2016). Longer genes have also been previously shown to be related to diseases such as cancer, cardiomyopathies, and diabetes (Sahakyan and Balasubramanian, 2016).

In this work, we used human genome data to identify possible functions associated with gene size, with a focus on protein-coding regions and genes. Correlation tests were used to identify relationships between gene length and other gene and protein characteristics. We observed that longer genes are expressed in the brain, heart diseases, and cancer, while smaller genes mostly participate in the immune system and in the development of the skin. Therefore, we hypothesize, based on our results, that genes with longer transcripts are mostly associated with functions in the early development stages, while genes with smaller transcripts have important roles in day-to-day functions.

MATERIALS AND METHODS

Data Retrieval and Filtering

All protein-coding human transcripts and genes ($N_{\text{transcripts}}$ = 92,696), their length, transcript count, and GC content were obtained using the BioMart (Zerbino et al., 2018) website (GRCh38.p12, Ensembl 96, April 2019). Transcript length is defined by Ensembl as the total length of the exons in a gene plus the lengths of its untranslated (UTR) regions. Gene length was obtained using the R (version 3.5.2) packages biomart (version 2.38.0) and GenomicRanges (version 1.34.0) and based on the code for the getGeneLengthAndGCContent function of the EDASeq (version 2.14.1) package. To avoid any discrepancies between how Ensembl defines transcript length and our calculated gene length, we extracted the start and end positions for all exons+UTR regions of all transcripts and calculated gene length based on the combined length between those regions. Using R, the transcripts with the highest transcript length per gene were selected. In the case of ties, due to multiple transcripts having the same length per gene, we used some tags (APPRIS annotation was the principal one, if there was an entry in RefSeq or GENCODE) used by Ensembl as a tiebreaker. Should that fail, the oldest transcript was chosen, by means of having a smaller numerical ID. Transcripts associated with PATCH locations or assemblies were removed from our dataset. For each transcript, we obtained data regarding their number of exons, coding sequence (CDS) length, the number of single nucleotide polymorphisms (SNPs), synonymous ("synonymous_variant"), missense ("missense_variant"), and nonsense ("stop_gained") SNPs, protein length, and the dN and dS values using the biomart package. For the dN and dS values, only those associated with one-to-one orthologs were selected for the present analysis. Average expression was obtained from the USCS Table browser tool (Karolchik et al., 2004) using expression as the group and the GTEx Gene track. Tissue-specific tau values of expression were obtained from another work (Palmer et al., 2019). The number of SNPs per gene was obtained using the biomart package and website.

The whole file produced and used in the analysis for this work can be found in **Supplementary Table S1** (N = 19,714).

A second dataset was built using the APPRIS annotation (Rodriguez et al., 2013) provided by Ensembl. Transcripts were first selected for each gene, based on whether they were identified as the principal isoform. Transcripts with entries in RefSeq were used as a tiebreaker. Should there still be duplicates, the oldest transcript was used. This dataset can be found in **Supplementary Table S2A** (N = 19,702).

Genes related with aging (N = 307) were obtained from GenAge (Build 19; Tacutu et al., 2018).

Statistical Tests, Graphs, and Other Packages

R and the function corr.test were used to perform the correlation tests. Due to the abundance of data, there were a lot of ties in the ranks, which prevented the usage of Spearman's correlation, so instead we chose to use the Kendall test for the correlations. Partial correlations were done using the ppcor (version 1.1) package. For multivariable regression, the lm function in R was used based on the following formula: Number of PPI ~ Transcript Length + Number of publications. The figures produced in this work were created using the ggplot2 (version 3.2.0) package in R. The other packages used over the course of this work were: corrplot (version 0.84), psych (version 1.9.12.31), ggpubr (version 0.2.1), cowplot (version 1.0.0), stringr (version 1.4.0), dplyr (version 0.8.5), plyr (version 1.8.4), and tidyr (version 0.8.3).

Functional Analysis

WebGestalt (2019 release; Liao et al., 2019) was used to do the overrepresentation enrichment analysis for each of the Gene Ontology (GO) categories (biological process, cellular component, and molecular function). The top 5% genes, with the highest and the lowest gene lengths, were ran against the reference option of genome. The significance level was FDR < 0.05 and the multiple test adjustment was done using the Benjamini-Hochberg method.

For confirmation of the results, the same two 5% lists were run on DAVID's (Huang et al., 2009a,b) annotation clustering

option using the complete human genome as background. Only those terms with a value of p and false discovery rate (FDR) smaller or equal to 0.05 were considered. Default categories were used, except for the category "UP_SEQ_FEATURE" since it was introducing a lot of redundant results.

To help better visualize the GO terms obtained from the analysis described above, the tool REViGO (Supek et al., 2011) was used. The values of p here considered were the FDR values obtained previously, with the human database option used for the GO terms.

With regard to the analysis done using the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, the grouping of genes with pathways was obtained from the Molecular Signature database (version 6.2; Kanehisa and Goto, 2000; Subramanian et al., 2005; Liberzon et al., 2011, 2015; Kanehisa et al., 2017, 2019), as was done previously by another group (Sahakyan and Balasubramanian, 2016). For each gene, the length used was that of the corresponding transcript from our dataset. Additionally, the coloring of the box plot was done based on the fact that the pathway in question is directly associated with the category (when the KEGG pathway schematic shows cells from the category) or if they could be indirectly associated with the category (using available literature). For this latter case, the appropriate literature was selected if the said literature mentioned elements of the KEGG pathway being involved in the said category. The categorization on the basis of published work has its advantages, but there is often overlapping of functions within these categories; for example, calcium signaling also happens in the muscle (Kuo and Ehrlich, 2015) and immune system (Vig and Kinet, 2009), the Wnt signaling pathway also has a role in cancer (Zhan et al., 2017), and the TGF-beta signaling pathway can also be associated with the immune system (Worthington et al., 2012), among others.

While the KEGG pathways used in this work did not incorporate all of the genes in our dataset, these KEGG pathways were considered canonical by the MSigDB (Subramanian et al., 2005; Liberzon et al., 2015), which would provide more certainty that our results are genuine by removing a lot of the ambiguity around the existence of several of our genes.

To further our understanding of the influence of gene length in the immune system, we analyzed data from Reactome (version 75; Jassal et al., 2019). Data were extracted based on the ontology levels. If we assume that the immune system is the first ontology level, then the second level would include its child terms (innate immune system, adaptive immune system, and cytokine signaling), with the third level including the child terms of the second level, and so on.

Co-Expression Analysis

Co-expression correlation values were extracted from GeneFriends (van Dam et al., 2015). For each gene (N = 19,714), in the whole dataset and in the top 5% lists of genes with the longest (High group) and the smallest (Low group) transcript lengths (N = 986 for each list), the number of genes with correlation values superior or equal to 0.6 or smaller or equal to −0.6 were obtained using R. From our original dataset (N = 19,714 genes), 1,046 genes were not present in GeneFriends (whole dataset), of which 25 missing genes were within the High group and 110 missing genes were within the Low group.

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

Protein-Protein Interaction Analysis

BioGRID (release 3.5.174) REST API (Stark, 2006), in combination with the R package httr (version 1.4.0), was used to obtain all protein-protein interactions for the whole dataset and for the top 5% longest and smallest genes. All redundant and genetic interactions were removed from this analysis.

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

RESULTS

Longest and Shortest Genes

From all of the protein-coding genes in the human genome, a dataset was built selecting only the longest transcripts for each gene (N = 19,714 genes; **Supplementary Table S1**). As our focus is on protein-coding regions, we used the transcript length in our analysis, owing to the fact that there is a very strong correlation between the length of the longest transcript of a gene and its respective gene length (Kendall test: $\tau = 0.72$, $p < 2.20\text{E}{-16}$; **Supplementary Figure S1**). The five biggest genes in terms of transcript length have been studied previously, and they are associated with neuronal functions (Tao and Sampath, 2010; Yamagishi et al., 2011; Hu et al., 2016), cardiac tissue (Ware and Cook, 2017), and cancer (Felder et al., 2014; **Table 1**). However, the smallest genes might be annotation errors in the genome build.

We also built an additional dataset using the principal isoform based on APPRIS annotation (N = 19,702; **Supplementary Table S2A**). We found an overlap of 14,955 transcripts between both datasets, resulting in a 75% overlap with the original transcript length-based dataset. Furthermore, the results presented in the following section were vastly similar to those obtained using the APPRIS dataset (**Supplementary Table S2** and **Supplementary Figure S2**), and therefore we focused our analysis and discussion on the transcript length-based dataset.

Functional Analysis

One of the main objectives of the present study was to understand whether gene function was associated with gene length. Keeping this in mind, and using a list of the top 5% protein-coding genes with the longest and smallest transcript lengths, we performed a functional analysis using tools like WebGestalt (Liao et al., 2019), DAVID (Huang et al., 2009a,b), KEGG (Kanehisa and Goto, 2000), and Molecular Signature Database (Subramanian et al., 2005; Liberzon et al., 2015).

TABLE 1 | List of the top 5 longest protein-coding transcripts in the human genome.

Transcript stable ID	Gene ID	Gene name	Transcript length	Gene length	Exon count	Intron count	Number of SNPs	Protein size
Longest genes								
ENST00000589042	ENSG00000155657	*TTN*	109,224	118,976	363	362	69,258	35,991
ENST00000397910	ENSG00000181143	*MUC16*	43,816	43,830	84	83	38,498	14,507
ENST00000262160	ENSG00000175387	*SMAD2*	34,626	36,426	11	10	26,668	467
ENST00000330753	ENSG00000185070	*FLRT2*	33,681	34,901	2	1	25,451	660
ENST00000609686	ENSG00000273079	*GRIN2B*	30,355	30,941	13	12	90,195	1,484

For the genes with longer transcript lengths (**Figure 1**), most of the biological functions found seem to be associated with the brain, specifically with regard to neurons. This can also be confirmed when looking at the cellular component (**Supplementary Figure S3A**) and molecular function (**Supplementary Figure S3B**) and at the similar results produced using DAVID (**Supplementary Table S3**). The top 10% longest protein-coding genes produced similar results (**Supplementary Table S3** and **Supplementary Figures S3E–G**).

For the genes with shorter transcript lengths (**Figure 2**), most of the biological functions found are related to the skin and the immune system. Cellular component (**Supplementary Figure S3C**), molecular function (**Supplementary Figure S3D**), and the DAVID (**Supplementary Table S3**) results support this observation. Again, the top 10% smallest protein-coding genes produced similar results (**Supplementary Table S3** and **Supplementary Figures S3H–J**).

The results for the KEGG pathways (**Figure 3A** and **Supplementary Figure S4**) were color coded for each box plot based on their association with the terms we found most relevant (brain, cancer, heart, immune system, muscle, neurodegenerative disease, skin, and others). For cases where there was no direct association, a literature search was done for relevant articles that showed that those pathways were related to the brain (Dermietzel and Spray, 1993; Fisher et al., 2002; Funderburgh, 2002; Lasky and Wu, 2005; Lin et al., 2006; Frere et al., 2012; Kwok et al., 2012; Monje et al., 2012; Russo et al., 2012; Bauer et al., 2014; Kerrisk et al., 2014; Massaly et al., 2014; Mei and Nave, 2014; Stocker and Chenn, 2015; Schnaar, 2016; Zeng et al., 2016; Noelanders and Vleminckx, 2017; Grube et al., 2018; Russo et al., 2018; Dickson, 2019), cancer (Ogretmen, 2018), immune system (Prentki and Madiraju, 2008; Le Floc'h et al., 2011; Barber, 2014; Seif et al., 2017; Zhang et al., 2019), and the skin (Taylor et al., 1991; Fisher and Voorhees, 1996; Iversen and Kragballe, 2000; Ziboh et al., 2000; Slominski et al., 2013). From the total 19,714 genes in our dataset, 5,203 (26%) were annotated with associations to KEGG pathways. The number of genes per pathway can be found in **Supplementary Table S4**.

Looking at the KEGG pathway results for the longest transcript lengths, we identified pathways associated with the brain, cancer, heart disease, and muscle (**Figure 3A** and **Supplementary Figure S4**), while the pathways with the shortest transcript lengths are mostly associated with the immune system; a few of them were also associated with the skin and neurodegenerative diseases (**Figure 3B** and **Supplementary Figure S4**).

The full KEGG results (186 gene sets) can be found in **Supplementary Figure S4**, and the KEGG pathway IDs can be found in **Supplementary Table S4**.

Finally, we wished to further understand the role of transcript length in the several functions of the immune system. Using the pathway database Reactome (Jassal et al., 2019), we investigated the distribution of transcript length at several ontology levels of the immune system. For the second ontology level, genes of the innate immune system (median = 2,830) and of the cytokine signaling pathways (median = 2,890) are significantly smaller than the genes from the adaptive immune system (median = 3,112), although the difference between them was not substantial (**Supplementary Figure S5A**). Regarding the third (**Supplementary Figure S5B**) and fourth (**Supplementary Figure S5C**) ontology levels, shorter genes appear to participate in the antimicrobial peptide pathway (defensins). Interestingly, part of the complement pathway (activation of C3 and C5; **Supplementary Figure S5C**) includes longer genes, but overall, the complement pathway (**Supplementary Figure S5B**) appears to be on the shorter-transcript-length side.

Gene Properties Correlated With Transcript Length

In order to understand the relationship between transcript length and other gene characteristics, a correlation analysis was performed. When looking at the number of SNPs for each transcript (**Figure 4A**), there was a significant positive correlation with transcript length (Kendall test: $\tau = 0.45$, $p < 2.20\text{E}{-16}$). Similar results were found when comparing the number of SNPs per gene with gene length (Kendall test: $\tau = 0.49$, $p < 2.20\text{E}{-16}$; **Supplementary Figure S6A**). After comparing the number of introns and the transcript length (**Figure 4B**), we found a weak but significant positive correlation between these two variables (Kendall test: $\tau = 0.35$, $p < 2.20\text{E}{-16}$). The strongest positive correlation (Kendall test: $\tau = 0.48$, $p < 2.20\text{E}{-16}$) was the association with protein size (**Figure 4C**); the weakest correlation (Kendall test: $\tau = 0.04$, $p = 3.06\text{E}{-14}$) was the association with average gene expression (**Figure 4D**).

Additionally, for the correlations with the transcript count (**Supplementary Figure S6G**) and GC content (**Supplementary Figure S6H**), we observed a weak but significant positive correlation (Kendall test: $\tau = 0.22$, $p < 2.20\text{E}{-16}$) and a weak significant negative correlation (Kendall test: $\tau = -0.19$, $p < 2.20\text{E}{-16}$), respectively.

We were also interested in understanding the effect of transcript length in specific mutations. We observed some strong and

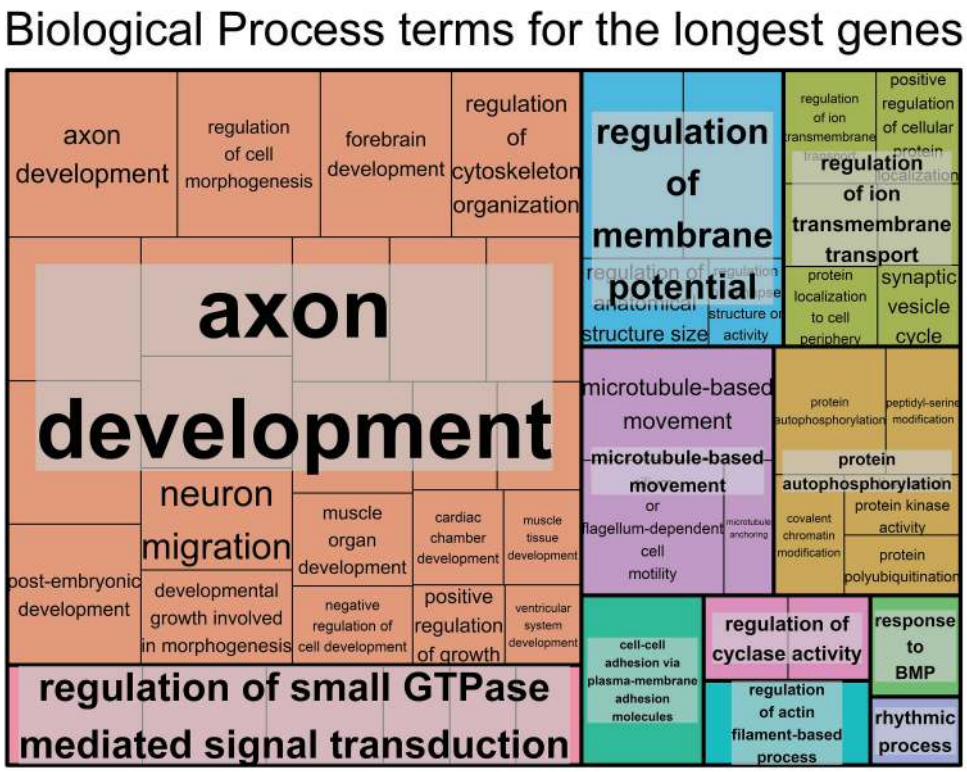

FIGURE 1 | Biological process Gene Ontology (GO) terms found associated with genes with the longest transcript length. Individual compartments represent each biological process term, and compartments were grouped based on semantic similarity. Compartment proportion is based on the false discovery rate (FDR) value. Overrepresentation enrichment analysis was performed with WebGestalt (Liao et al., 2019) and the visualization tool REViGO (Supek et al., 2011) was used to produce this figure. The significance level was $p < 0.05$ and the FDR was set at 0.05. FDR estimation was done using the Benjamini-Hochberg method. Full data available in **Supplementary Table S2A**.

statistically significant correlations between transcript length and synonymous (Kendall test: $\tau = 0.44$, $p < 2.20E{-}16$; **Supplementary Figure S6I**) and missense (Kendall test: $\tau = 0.42$, $p < 2.20E{-}16$; **Supplementary Figure S6J**) mutations. However, in the case of nonsense mutations (Kendall test: $\tau = 0.21$, $p < 2.20E{-}16$; **Supplementary Figure S6K**), a weaker but significant positive correlation with transcript length was observed. This was followed by the calculation of missense/synonymous (MIS/SYN) and nonsense/synonymous (NONS/SYN) rates in order to measure the functional importance of gene length. We observed that these ratios had similarly negative correlations with transcript length, with MIS/SYN having a weaker significant correlation (Kendall test: $\tau = -0.07$, $p < 2.20E{-}16$; **Supplementary Figure S6L**) than NONS/SYN (Kendall test: $\tau = -0.19$, $p < 2.20E{-}16$; **Supplementary Figure S6M**).

In order to better understand whether the correlations found were solely due to the transcript length or whether other factors were influencing them, we built a correlation matrix with several gene characteristics (**Figure 4E**). We observed that properties like intron counts, CDS length, protein size, number of SNPs, and transcript counts have strong positive correlations among themselves, some of which were stronger than any other correlations with transcript length. This indicates that the strong correlations with transcript length might not be due to the sole action of the transcript length itself, but rather due to the combined effects between several gene characteristics that also correlate with each other. Furthermore, partial correlations were performed for the number of SNPs (Kendall test: $\tau = 0.27$, $p < 2.20E{-}16$), intron count (Kendall test: $\tau = -0.02$, $p = 2.33E{-}05$), protein size (Kendall test: $\tau = 0.34$, $p < 2.20E{-}16$), average gene expression (Kendall test: $\tau = 0.02$, $p = 5.68E{-}07$), transcript count (Kendall test: $\tau = 0.05$, $p < 2.20E{-}16$), and GC content (Kendall test: $\tau = -0.09$, $p < 2.20E{-}16$) against transcript length while accounting for these other variables. The differences in the tau values between the correlations and partial correlations further illustrate that these variables are not independent of each other.

Distribution of Transcript Length and Expression in Human Tissues

In this present work, we have found that the transcript length seems to peak at 2,065 bp, with smaller transcripts being more common than longer ones (**Supplementary Figure S7A**). As described previously (Gorlova et al., 2014), the distribution of the number of introns in the human genome (**Supplementary Figure S7B**) has a mode of three introns,

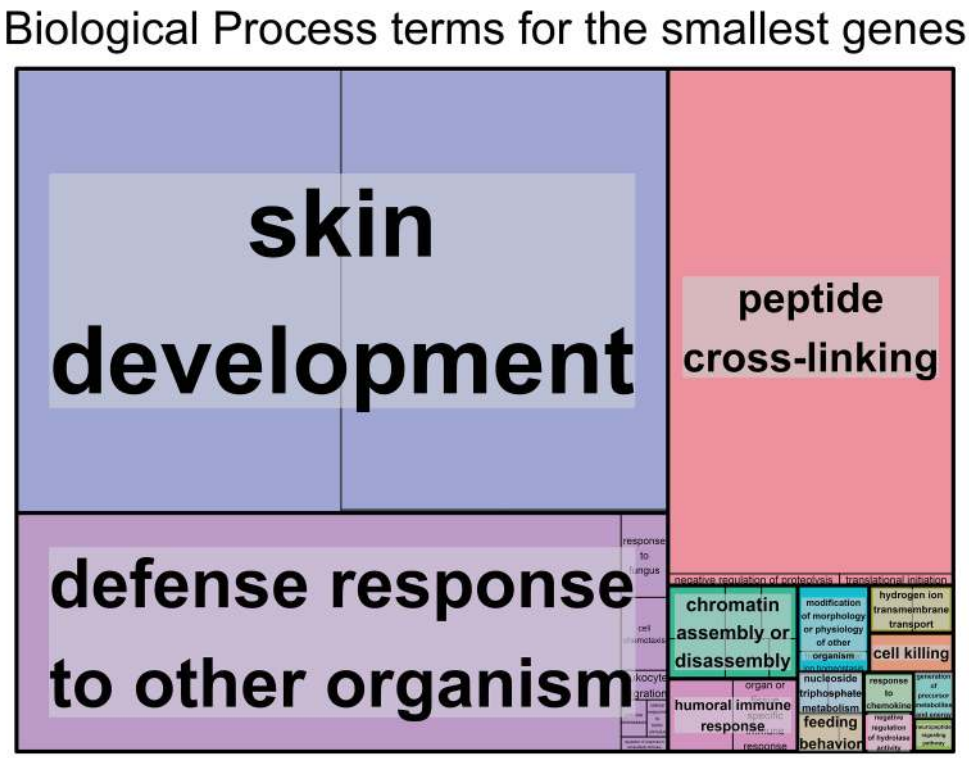

FIGURE 2 | Biological process Gene Ontology (GO) terms found associated with genes with the smallest transcript length. Individual compartments represent each biological process term, and compartments were grouped based on semantic similarity. Compartment proportion is based on the false discovery rate (FDR) value. Overrepresentation enrichment analysis was performed with WebGestalt (Liao et al., 2019) and the visualization tool REViGO (Supek et al., 2011) was used to produce this figure. The significance level was $p < 0.05$ and the FDR was set at 0.05. FDR estimation was done using the Benjamini-Hochberg method. Full data available in **Supplementary Table S2B**.

and there are very few genes with a large number of introns. The gene with the most introns is TTN, with 362 introns, which is also the gene with the longest transcript length.

To better understand the distribution of transcript length in human tissue expression, we used tau values obtained from GTEx data (Palmer et al., 2019). Tau was used as a measure of tissue specificity, based on the expression profile in different tissues, with values ranging from 0, for broadly expressed genes, to 1, for tissue-specific genes (Yanai et al., 2005). For genes with a tau value above 0.8 (**Figure 5** and **Supplementary Figure S8** for the non-log-transformed version), we observed that longer tissue-specific genes are often associated with the blood vessel, nerve, thyroid, cervix uteri, and brain, while smaller tissue-specific genes are found in the pancreas, skin, stomach, vagina, and testis.

Aging and Transcript Length

Aging is an important factor in our lives, and it affects most organisms. We were curious to see whether, for genes related to aging, the distribution of transcript length was significantly different from the rest of the protein-coding genes. We observed (**Supplementary Figures S9A,B**) that the genes associated with aging (N = 307; Tacutu et al., 2018) have longer transcript lengths (median = 3,517) when compared with the rest of our dataset (median = 2,956) and that this difference of medians was significant (Wilcoxon rank-sum test: p = 0.00036).

To further understand whether longer or smaller genes were more prominent with age, we used genes from aging gene expression signatures obtained from a meta-analysis in human, mouse, and rat tissues (Palmer et al., 2019). Genes from this signature were either overexpressed (N_{Total} = 449, N_{Brain} = 147, N_{Heart} = 35, N_{Muscle} = 49) or underexpressed (N_{Total} = 162, N_{Brain} = 16, N_{Heart} = 5, N_{Muscle} = 73) with age. Overall, the difference in the medians for the distribution of transcript length in the genes overexpressed (median = 3,068) and underexpressed (median = 3,026.5) with age was not observed to be significant (Wilcoxon rank-sum test: p = 0.81; **Supplementary Figure S9C**). However, the tissue-specific signatures of aging show that the brain favors smaller genes with age (Wilcoxon rank-sum test: p = 0.00086, median for overexpression in the brain = 2,651, median for underexpression in the brain = 5,824; **Supplementary Figure S9D**).

Evolution and Transcript Length

The relationship between intronic burden and evolution has been established before (Gorlova et al., 2014), but very few studies approached this on a gene length front. Therefore,

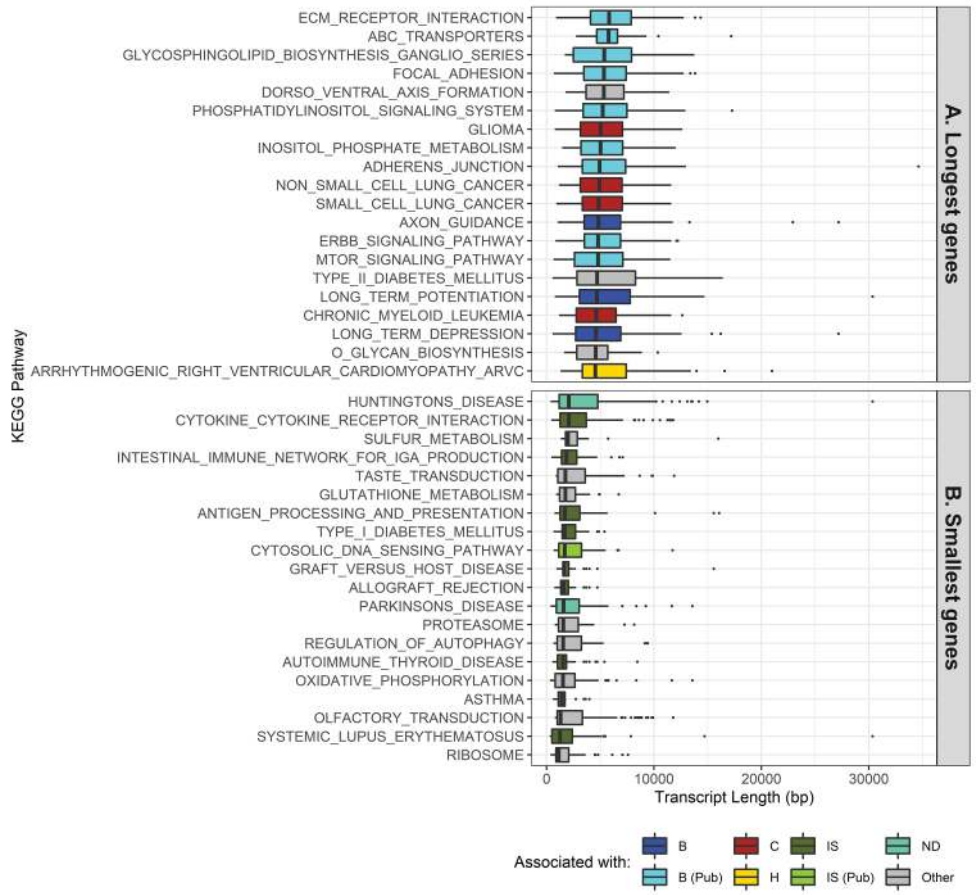

FIGURE 3 | Transcript length distribution per Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway for the longest and smallest genes. *Colors* illustrate what the KEGG pathway has been directly associated with (*B* for brain, *C* for cancer, *H* for heart, *IS* for immune system, and *ND* for neurodegenerative diseases), due to it being stated in the pathway itself or indirectly associated with (Pub tag), by means of literature references. KEGG pathways and genes involved in said pathways were obtained from the Molecular Signature Database (Subramanian et al., 2005; Liberzon et al., 2015). **(A)** Top 20 pathways with the longest genes, ordered by median. **(B)** Top 20 pathways with the smallest genes, ordered by median.

we obtained the dN and dS values for three organisms paired with human – mouse (**Supplementary Figure S10A**), gorilla (**Supplementary Figure S10B**), and chimpanzee (**Supplementary Figure S10C**) – and we aimed to see how the distribution of transcript length happened in function of their dN/dS ratios. Overall, longer genes were associated with a dN/dS ratio less than 1 (the median transcript lengths are 3,294, 3,377, and 3,338 for mouse, chimpanzee, and gorilla, respectively), while smaller genes seem to be more associated with dN/dS ratios above or equal to 1 (the median transcript lengths are 1,171.5, 2,229.5, and 2,092 for mouse, chimpanzee, and gorilla, respectively), and the medians for both transcript length groups were always significantly different (Wilcoxon rank-sum test: $p = 0.00073$ for mouse and $p < 2.2E{-}16$ for both gorilla and chimpanzee).

Co-Expression Analysis and Protein-Protein Interactions

Co-expression networks can help us better understand the functions of genes that are often expressed together and thus tend to be functionally related (van Dam et al., 2018). In order to determine whether gene length influenced the amount of co-expressed partners, we used data from GeneFriends (van Dam et al., 2015; **Supplementary Table S5**). We observed a weak correlation between transcript length and the number of co-expression partners in our dataset (Kendall test: $\tau = 0.10$, $p < 2.2E{-}16$; **Supplementary Figure S11A**). However, despite this weak correlation, longer genes appear to have more co-expressed gene partners than do smaller genes (Wilcoxon rank-sum test: $p < 2.2E{-}16$; **Figure 6A**; not-transformed figure in **Supplementary Figure S11B**, median values of the co-expression partners for longer genes = 2,725, median values of the co-expression partners for smaller genes = 32). We further analyzed the top and lower hundred human co-expressed genes from the GeneFriends database (**Supplementary Table S5**) and observed that the top highly co-expressed genes in the database have significantly longer transcript lengths (Wilcoxon rank-sum test: $p = 0.00072$, median = 3,880; **Supplementary Figure S11C**) with respect to the bottom ones (median = 2,587.5).

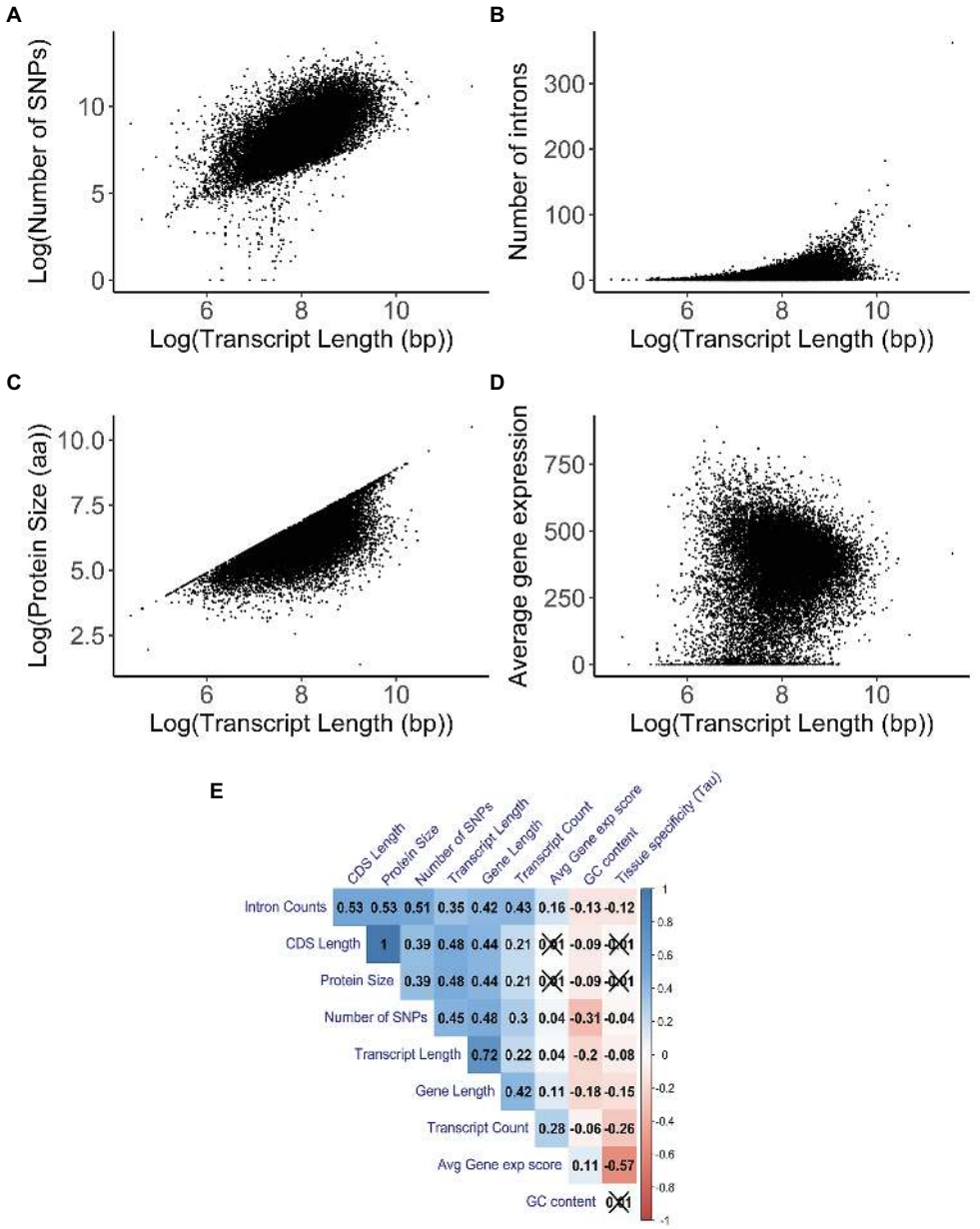

FIGURE 4 | Correlation analysis between transcript length (in base pairs) and other gene characteristics. A 2D density plot can be found in **Supplementary Figure S4B**. Parts **(A–D)** have been logarithmically transformed in order to help visualize their relationship and/or account for the skewing introduced by outliers. The original versions of the figures can be found in **Supplementary Figures S4C–F**. **(A)** Correlation between the log-transformed number of single nucleotide polymorphisms (SNPs) and the log-transformed transcript length (in base pairs; Kendall test: $\tau = 0.45$, $p < 2.20E{-}16$). The number of SNPs and the transcript length for each transcript were obtained using biomart. **(B)** Correlation between the number of introns and the log-transformed transcript length (in base pairs; Kendall test: $\tau = 0.35$, $p < 2.20E{-}16$). The number of introns and the transcript length for each transcript were obtained using biomart. **(C)** Correlation between the log-transformed protein size (in amino acids) and the log-transformed transcript length (in base pairs; Kendall test: $\tau = 0.48$, $p < 2.20E{-}16$). Protein size and transcript length were obtained using biomart. **(D)** Correlation between the average gene expression and the log-transformed transcript length (in base pairs; Kendall test: $\tau = 0.04$, $p = 3.06E{-}14$). Average gene expression was obtained from the UCSC Genome browser; this value was derived from the total median expression level across all tissues and was based on the GTEx project. Transcript length was obtained using biomart. **(E)** Correlation matrix between gene properties. Kendall's test was used as a measurement of correlation, with the *numbers* and the *gradient of colors* symbolizing the tau values for each comparison. The values for the number of SNPs are for each transcript. Values that are *crossed out* are not statistically significant. Values are clustered together based on their tau values.

To determine whether transcript length also influenced the number of protein-protein interactions, we used the protein-protein interaction data from BioGRID (Stark, 2006; **Supplementary Table S6**). The results obtained were similar to the co-expression, where a weak correlation was observed between transcript length and the number of protein-protein

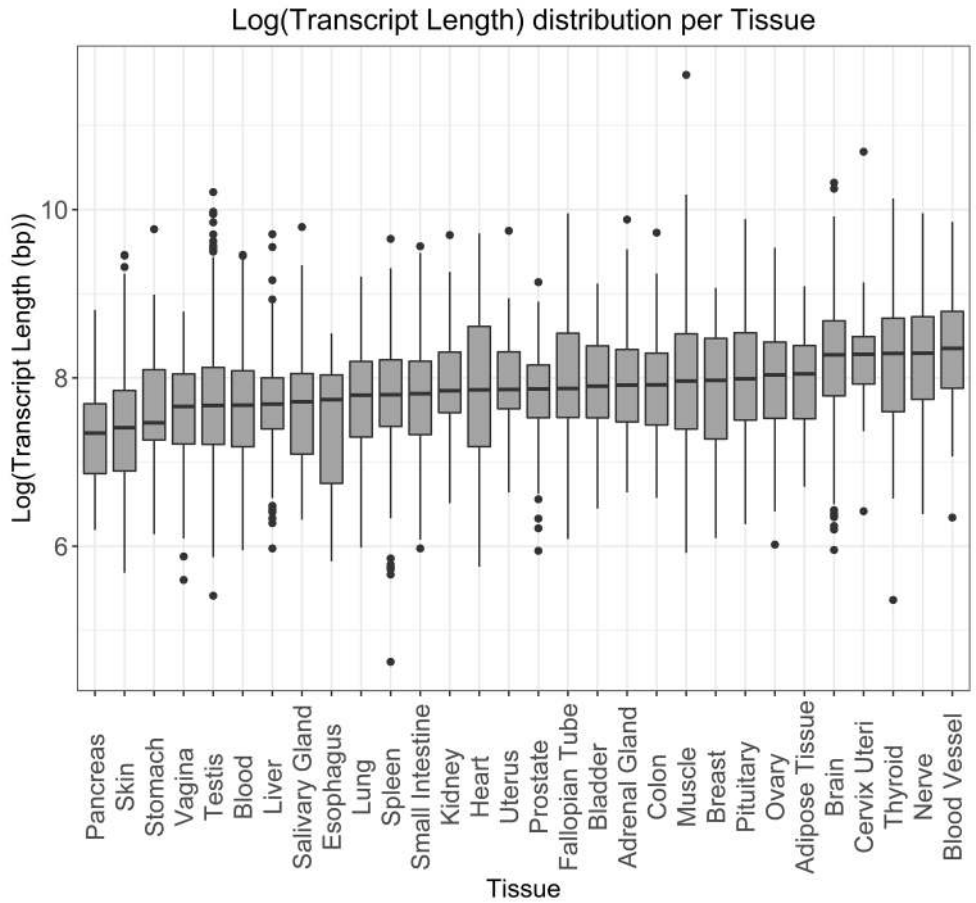

FIGURE 5 | Log-transformed transcript length distribution for genes specifically expressed in the given tissues. Tissue specificity was defined as a gene having a tau specificity score greater than 0.8.

interactions (Kendall test: τ = 0.06, p < 2.2E−16; **Supplementary Figure S12A**).

To ascertain that the interactions found were not due to publication bias, we obtained the number of publications for each gene from PubMed and compared these to gene length and to the number of interactions (**Figure 6B**). We observed that the number of interactions and publications were significantly different between each gene length group (Wilcoxon rank-sum test: p < 2.2E−16 for both comparisons), with both being higher for the group comprising longer length genes. In order to assess the level of influence of publication bias in our protein-protein interaction dataset, we used correlations between the values of protein-protein interactions and the number of publications, and we observed that, for both gene length groups, the correlations were not the strongest (Kendall test: longest genes, τ = 0.26, p < 2.2E−16; smallest genes, τ = 0.36, p < 2.2E−16), implying that while there might be some publication bias in effect, the strength of that effect is rather weak. To further support this, we carried out a multivariable regression to determine the influence of the number of publications and transcript length in the number of protein-protein interactions (PPI) found (p of F statistic = 1.936E−10). We again observed that the number of publications did not have a significant association with the number of PPI (t = 1.41, p = 1.59E−1), unlike transcript length (t = 6.55, p = 6.06E−11). This resulted in the model being extremely inaccurate (R^2 = 0.0022, residual standard error = 78.05).

In the group of the longest genes, 208 (21%) entries had zero protein-protein interactions, while for the smallest group of genes, 544 (55%) entries had zero protein-protein interactions. This means that there were either no physical interactions for those genes or that there were no entries in BioGRID for them. So as to account for this, and similarly to what we did for the co-expression analysis, we extracted the top 100 genes with the most and fewest protein-protein interactors (without null values) in our dataset and observed the distribution of their transcript length. We observed that the genes with the most protein-protein interactions were longer (median transcript length = 3,737) than the genes with the smallest amount of protein-protein interactions (Wilcoxon rank-sum test: p = 0.039, median transcript length = 2,764; **Supplementary Figure S12B**).

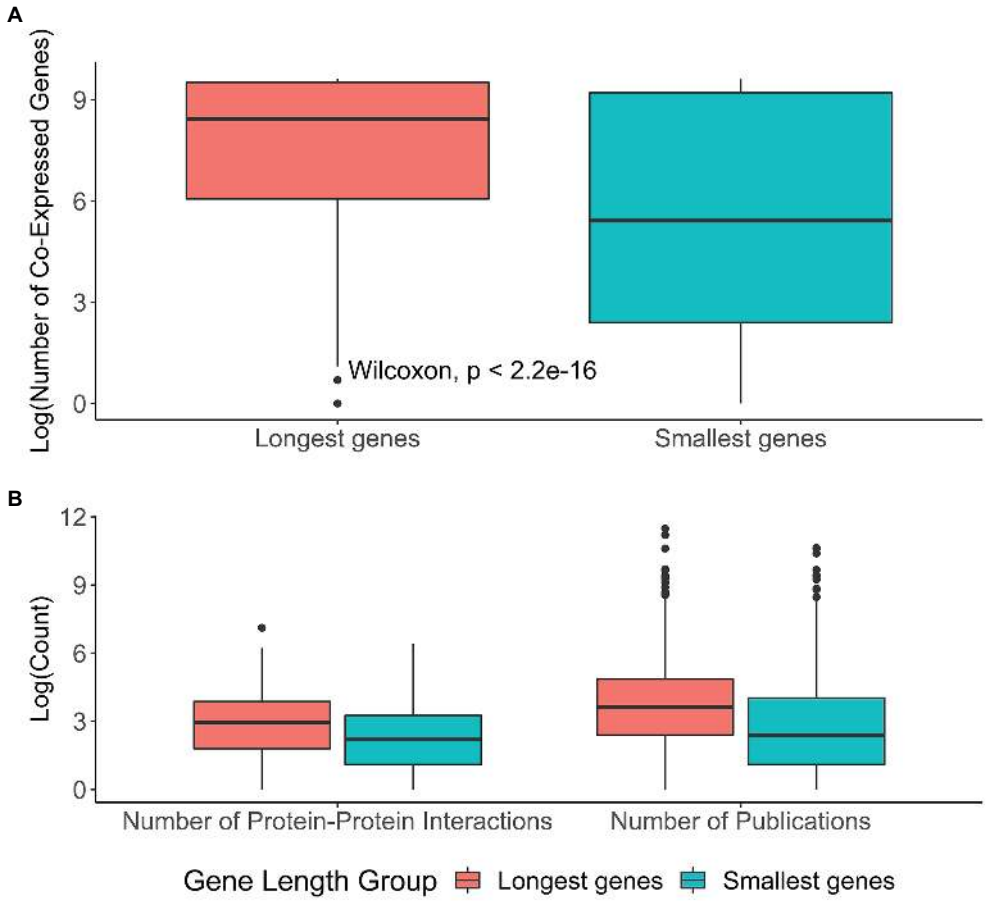

FIGURE 6 | Co-expression and protein-protein interaction results pertaining to the longest and the smallest genes. The High group corresponds to the top 5% longest genes found in our original dataset (N_{High} = 986), while the Low group corresponds to the top 5% smallest genes found in our original dataset (N_{Low} = 986). **(A)** Distribution of the log-ransformed number of co-expressed genes for the long genes and the small genes. The number of co-expressed genes were obtained from data publicly available in GeneFriends (van Dam et al., 2015). **(B)** Distribution of the number of protein-protein interactions and the number of publications for longer and smaller genes, all log-transformed. The number of protein-protein interactions were obtained from BioGRID (Stark, 2006) and the number of publications obtained from PubMed.

DISCUSSION

With this work, we tried to elucidate the factors associated with gene length and, in particular, whether gene length is associated with the function of the resulting proteins in the cell. Even looking at the five longest genes, we can get a small glimpse into the possible functions associated with gene length. *TTN* is the longest transcript in the human genome and serves several important functions in the skeletal and cardiac muscles and is often involved in structure, sensory, and signaling responses (Chauveau et al., 2014; Savarese et al., 2016; Ware and Cook, 2017). The mucin *MUC16* (or CA125) is mostly known as a biomarker in ovarian cancer and is used to monitor patients as an indicator of cancer recurrence (Felder et al., 2014; Haridas et al., 2014; Das and Batra, 2015). Furthermore, MUC16 normally functions as a protector of epithelial cells (Haridas et al., 2014). SMAD family member 2 (*SMAD2*) is thought to play a critical role in neuronal function (Tao and Sampath, 2010) and to have a protective role in hepatic fibrosis (Xu et al., 2016). The gene *FLRT2* is believed to have a role in tumor suppression in breast and prostate cancers (Wu et al., 2016; Bae et al., 2017), and in mouse models, *FLRT2* has been found as a guiding agent in neuronal and vascular cells (Yamagishi et al., 2011; Seiradake et al., 2014). For the *GRIN2B* gene, it has been shown to play an important role in neuronal development and in cell differentiation in the brain (Hu et al., 2016; Bell et al., 2018). We cannot obtain any information at the moment pertaining to the function of the five smallest genes since all of them are either novel and have yet to be properly studied and, indeed, could be annotation errors in the human genome assembly.

In order to understand the effects of gene length on protein function, we performed a functional analysis. For longer genes, the GO terms obtained were mostly associated with neurons; for example, terms like axon development, axon part, neuron-to-neuron synapse, actin and cell polarity (Polleux and Snider, 2010), and GTPases (Polleux and Snider, 2010). For tissue-specific genes, the brain and nerves also had the longest genes. Looking at the KEGG pathways associated with the longest genes, the categories present are in the brain, cancer, heart diseases, and muscle. Previous studies have associated longer genes with neurons

(Zylka et al., 2015; Takeuchi et al., 2018; McCoy and Fire, 2020) and muscle (Hosokawa et al., 2019). Due to the very nature of longer genes, one expects high rates of mutation, not only due to their size but also due to possible collisions between the RNA polymerase and the DNA polymerase, which cause instability and possible mutations (Helmrich et al., 2011). It is not surprising to find associations between longer genes and cancer (Sahakyan and Balasubramanian, 2016) and heart pathologies often caused by mutations in particularly long genes, like *DSC2* and *TTN* (Jefferies and Towbin, 2010; Maron and Maron, 2013; Corrado et al., 2017).

Looking at smaller genes, most of the GO terms were associated with the skin, for example skin development and cornified envelope, or with the immune system, for example, defense response to other organisms and receptor agonist activity. Smaller tissue-specific genes also have a major presence in the skin. With regard to the KEGG pathways associated with the smaller genes, most pathways were involved in the immune system, with a few also being present in neurodegenerative diseases and in the skin. In addition, the Reactome immune system pathway with the smallest genes was found to be associated with antimicrobial peptides (defensins). Defensins are small peptides that play a role in innate immunity and have been found to be expressed in several mammal species (Holly et al., 2017). Moreover, the complement cascade, another important function in the innate immune system, was also associated with smaller genes for the most part, with a specific pathway (activation of C3 and C5) being the exception. Previous studies have observed that most genes associated with immune functions are rather small in size (Pipkin and Monticelli, 2008). However, to our knowledge, there are no previous studies to support the association of smaller genes with skin development.

In spite of this, our findings led us to believe that there is a disparity in gene sizes for genes that have a role or are present in tissues with very little to almost no development postnatally (like neurons) and for genes (not involved in housekeeping) that are quite frequently expressed during a human's whole lifetime (like in skin development and immune response) or are involved in functions that require fast responses. Corroborating our findings for the functional analysis, a recent preprint has showed that, with age, there is a downregulation of long transcripts and an upregulation of short transcripts, in a phenomena they named "length-driven transcriptome imbalance," which, in humans, affects the brain the most (Stoeger et al., 2019). As we observed, smaller genes can be associated with the immune system – and inflammation has a role in many aging-related diseases (Goldberg and Dixit, 2015) – while longer genes are mostly associated with brain development, a function that happens early in life.

In terms of gene characteristics, there was no strong correlation with transcript length. The strongest positive correlations were with protein size and number of SNPs, with transcript count, number of introns, GC content, and average gene expression having a weak significant positive correlation.

The correlation we observed between average gene expression and transcript length was not in line with previous observations, which suggested that highly expressed genes are often shorter in length (Urrutia and Hurst, 2003). We observed that, among smaller genes, the average gene expression was, in fact, the highest (**Supplementary Figure S6F**). However, shorter genes also had a great variability in the average gene expression values, and there was almost no correlation between transcript length and average gene expression. What has been stated in the previous studies is relevant, but the whole image is not captured properly. Rather than stating that the smaller genes are highly expressed, it is more accurate to say that smaller genes have a greater variability of levels of expression than longer genes. Previous studies have shown that the length of messenger RNA (mRNA) will affect the translation dynamics, with smaller mRNAs producing more proteins (in both normal and cancer cells) than their longer counterparts, a phenomenon that may be due to energy conservation in the case of translation errors or misfolding (Valleriani et al., 2011; Wang et al., 2013; Guo et al., 2015). This matches with what we observed as well: smaller genes will have a more important role in day-to-day functions due to their ability to be rapidly translated and in more numbers; on the other hand, longer genes are important for early life development, especially in the brain and heart, where it would be worth spending more energy in more long-lasting, robust functions.

While the observation that the number of SNPs is correlated with transcript length is not surprising since, logically, longer genes will have a higher probability of accumulating more mutations than smaller genes, it is unexpected that the correlation was not stronger. To further explore this, we used different mutation ratios and observed their relationship with transcript length. Similar to the correlation results for the number of SNPs, both synonymous and missense mutations were also highly correlated with transcript length. It is particularly interesting that the correlation values were so high for missense mutations since these may cause loss of function in the resulting protein. Likewise, it could be one of the reasons why the correlation between nonsense mutations and transcript length is weaker than that for synonymous and missense mutations. Other works (Gorlova et al., 2014) have used the MIS/SYN and NONS/SYN ratios as a measure of functional importance, where, if a gene is important in terms of its function, it will be less tolerant toward the accumulation of missense and nonsense mutations. We observed that there was a, albeit faint, negative correlation between the ratios MIS/SYN and NONS/SYN and transcript length, which, based on the notions in the work of Gorlova et al., would imply that longer genes appear to be more functionally important than smaller genes. The negative correlation between the ratios MIS/SYN and NONS/SYN showed that longer genes may have more mechanisms in place to prevent loss-of-function mutations when compared with synonymous mutations. Moreover, we have to take account of "outliers" when looking into the correlation between transcript length and protein size (**Supplementary Figure S6E**), specifically for longer genes. One would expect that, for longer genes, the proteins produced would have a size comparable to their length and not be extremely small. However, some long transcripts result in small proteins due to the presence of very long 3'

UTR regions. While these regions still account for the calculation of transcript and gene size, they are not translated into the protein, causing the presence of these "outliers." Previous studies have shown that the brain has a preference for these long 3' UTR regions (Miura et al., 2013; Wang and Yi, 2014).

Interestingly, we noticed that genes associated with aging tend to be longer than the rest of the protein-coding genome. Moreover, we also showed that the overall (not tissue-dependent) expression of genes with age appears to be unrelated to transcript length and that the brain seems to favor the expression of smaller genes with age. The latter result is in line with the observations by Stoeger et al. (2019) who also witnessed an upregulation of smaller transcripts with age, especially in the brain. These results also make sense because small genes are often associated with immune function, which is often upregulated with age (de Magalhães et al., 2009). Furthermore, Palmer et al. (2019) observed that the genes overexpressed with age in the brain were mostly associated with immune functions. However, our results pertaining to the overall expression of genes with age are different from what Stoeger et al. (2019) observed that transcript length is an important source of aging-dependent changes in expression. It is possible that these differences between our results and those of Stoeger et al. (2019) are due to differences in the underlying datasets.

When comparing gene length with the dN/dS ratio for three organisms (gorilla, chimpanzee, and mouse), longer genes appear to evolve under stronger evolutionary constraints. Previous studies have shown that, for genes classified as "old" (by virtue of having orthologs in older organisms), their length will be longer, they will have more introns, and they evolve more slowly than smaller genes (Wolf et al., 2009; Vishnoi et al., 2010).

In terms of the co-expression analysis and protein-protein interactions, the longer genes, in general, had the most co-expression partners and protein-protein interactions. Further validating our observations, we also saw that the top hundred highest co-expression genes and PPI were longer in length as compared to the lowest co-expression genes and PPI. In light of these results, it is important to point out that gene length is a potential bias in large-scale genomic and systems biology studies that scientists should be aware of.

Not all genes are studied at the same depth. Some genes have more information related to expression or function than others. We observed this especially within our list of the 5% longest and smallest genes. Longer length genes had more functional information than shorter ones. We also observed that longer genes have more associated publications than smaller genes. Indeed, other groups have found that gene length can be an important predictor of the number of publications and that novel genes are not often studied to their full capacity (Stoeger et al., 2018), while others have found that genetic associations tend to be more biased toward longer genes (Mirina et al., 2012; de Magalhães and Wang, 2019).

The present study has its limitations. One of the limitations for this sort of study is that the results might be "time-specific." With new discoveries related to the human genome and its genes, the trends here observed might change, specifically when it concerns the currently untapped field of smaller genes. Similarly, as we previously noted, longer genes have a lot more information related to them when compared with their smaller counterparts. While our findings with respect to the longer genes might be more reliable, we cannot have the same confidence in the case of the smaller genes, considering that a lot of these genes have yet to be properly studied. Moreover, while smaller genes might be annotation errors, they could also be pseudo-genes or even non-coding genes that possess an open reading frame (ORF) that were missed when Ensembl scanned for pseudo-genes. Even after taking into account the above limitations, however, the present study still provides novel insights pertaining to gene length and its possible role in early life development, diseases, and response time in the human genome.

CONCLUSION

In this work, we aimed to further understand the relationships between gene length (mostly using the length of a gene's longest transcript as a proxy for gene length) and gene function as well as factors associated with gene length. We observed that, for most of the factors studied, there was not a particularly strong correlation with transcript length. The strongest correlations were observed with the number of SNPs and protein size. We also showed that, for smaller genes, their association with high levels of expression is not entirely correct and that, instead, there is great variability of expression values among them. We also observed that longer genes appear to have more co-expression partners and protein-protein interactions in comparison to their smaller counterparts.

At the functional level, we observed that longer genes tend to be associated with functions in the brain, cancer, heart, and muscle, while smaller genes are associated with the immune system, skin, and neurodegenerative diseases. This led us to believe that gene length could be associated with the frequency of usage of the gene, with longer genes being less often used past development and smaller genes playing a frequent role daily in the human body, like the immune system.

In conclusion, gene size does matter: longer genes tend to have more SNPs, are more likely to be important in development, have more interactions, and, ultimately, are more studied.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material** and in a GitHub repository (https://github.com/maglab/GeneLength_supplementary), further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

JM, GA, and IL conceived the study. IL and PR performed the analysis. IL prepared the figures. JM, PR, GA, and IL

drafted and finalized the paper. All authors contributed to the article and approved the submitted version.

FUNDING

IL is supported by a BBSRC grant (BB/R014949/1) to JM. GA is supported by the MRC-Arthritis Research UK Centre for Integrated Research into Musculoskeletal Ageing (CIMA), funded by the Medical Research Council and Versus Arthritis (grant number: MR/R502182/1). PR is supported by a Welcome Trust grant (208375/Z/17/Z) to JM. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

ACKNOWLEDGMENTS

We wish to thank past and present members of the Integrative Genomics of Ageing Group for useful suggestions and discussion, in particular Thomas Duffield, Kasit Chatsirisupachai, and Daniel Palmer.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.559998/full#supplementary-material

REFERENCES

Bae, H., Kim, B., Lee, H., Lee, S., Kang, H. -S., and Kim, S. J. (2017). Epigenetically regulated Fibronectin leucine rich transmembrane protein 2 (FLRT2) shows tumor suppressor activity in breast cancer cells. *Sci. Rep.* 7:272. doi: 10.1038/s41598-017-00424-0

Barber, G. N. (2014). STING-dependent cytosolic DNA sensing pathways. *Trends Immunol.* 35, 88–93. doi: 10.1016/j.it.2013.10.010

Bauer, H. -C., Krizbai, I. A., Bauer, H., and Traweger, A. (2014). "You Shall Not Pass"-tight junctions of the blood brain barrier. *Front. Neurosci.* 8:392. doi: 10.3389/fnins.2014.00392

Bell, S., Maussion, G., Jefri, M., Peng, H., Theroux, J. -F., Silveira, H., et al. (2018). Disruption of GRIN2B impairs differentiation in human neurons. *Stem Cell Rep.* 11, 183–196. doi: 10.1016/j.stemcr.2018.05.018

Chauveau, C., Rowell, J., and Ferreiro, A. (2014). A rising titan: TTN review and mutation update. *Hum. Mutat.* 35, 1046–1059. doi: 10.1002/humu.22611

Corrado, D., Link, M. S., and Calkins, H. (2017). Arrhythmogenic right ventricular cardiomyopathy. *N. Engl. J. Med.* 376, 61–72. doi: 10.1056/NEJMra1509267

Das, S., and Batra, S. K. (2015). Understanding the unique attributes of MUC16 (CA125): potential implications in targeted therapy. *Cancer Res.* 75, 4669–4674. doi: 10.1158/0008-5472.CAN-15-1050

de Magalhães, J. P., Curado, J., and Church, G. M. (2009). Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics* 25, 875–881. doi: 10.1093/bioinformatics/btp073

de Magalhães, J. P., and Wang, J. (2019). The fog of genetics: what is known, unknown and unknowable in the genetics of complex traits and diseases. *EMBO Rep.* 20:e48054. doi: 10.15252/embr.201948054

Dermietzel, R., and Spray, D. C. (1993). Gap junctions in the brain: where, what type, how many and why? *Trends Neurosci.* 16, 186–192. doi: 10.1016/0166-2236(93)90151-B

Dickson, E. J. (2019). Recent advances in understanding phosphoinositide signaling in the nervous system. *F1000Research* 8:278. doi: 10.12688/f1000research.16679.1

Eisenberg, E., and Levanon, E. Y. (2003). Human housekeeping genes are compact. *Trends Genet.* 19, 362–365. doi: 10.1016/S0168-9525(03)00140-9

Felder, M., Kapur, A., Gonzalez-Bosquet, J., Horibata, S., Heintz, J., Albrecht, R., et al. (2014). MUC16 (CA125): tumor biomarker to cancer therapy, a work in progress. *Mol. Cancer* 13:129. doi: 10.1186/1476-4598-13-129

Fisher, G. J., and Voorhees, J. J. (1996). Molecular mechanisms of retinoid actions in skin. *FASEB J.* 10, 1002–1013. doi: 10.1096/fasebj.10.9.8801161

Fisher, S. K., Novak, J. E., and Agranoff, B. W. (2002). Inositol and higher inositol phosphates in neural tissues: homeostasis, metabolism and functional significance. *J. Neurochem.* 82, 736–754. doi: 10.1046/j.1471-4159.2002.01041.x

Frere, S. G., Chang-Ileto, B., and Di Paolo, G. (2012). Role of phosphoinositides at the neuronal synapse. *Subcell. Biochem.* 59, 131–175. doi: 10.1007/978-94-007-3015-1_5

Funderburgh, J. L. (2002). Keratan Sulfate biosynthesis. *IUBMB Life* 54, 187–194. doi: 10.1080/15216540214932

Goldberg, E. L., and Dixit, V. D. (2015). Drivers of age-related inflammation and strategies for healthspan extension. *Immunol. Rev.* 265, 63–74. doi: 10.1111/imr.12295

Goldfeder, R. L., Wall, D. P., Khoury, M. J., Ioannidis, J. P. A., and Ashley, E. A. (2017). Human genome sequencing at the population scale: a primer on high-throughput DNA sequencing and analysis. *Am. J. Epidemiol.* 186, 1000–1009. doi: 10.1093/aje/kww224

Gonzaga-Jauregui, C., Lupski, J. R., and Gibbs, R. A. (2012). Human genome sequencing in health and disease. *Annu. Rev. Med.* 63, 35–61. doi: 10.1146/annurev-med-051010-162644

Gorlova, O., Fedorov, A., Logothetis, C., Amos, C., and Gorlov, I. (2014). Genes with a large intronic burden show greater evolutionary conservation on the protein level. *BMC Evol. Biol.* 14:50. doi: 10.1186/1471-2148-14-50

Grishkevich, V., and Yanai, I. (2014). Gene length and expression level shape genomic novelties. *Genome Res.* 24, 1497–1503. doi: 10.1101/gr.169722.113

Grube, M., Hagen, P., and Jedlitschky, G. (2018). Neurosteroid transport in the brain: role of ABC and SLC transporters. *Front. Pharmacol.* 9:354. doi: 10.3389/fphar.2018.00354

Guo, J., Lian, X., Zhong, J., Wang, T., and Zhang, G. (2015). Length-dependent translation initiation benefits the functional proteome of human cells. *Mol. BioSyst.* 11, 370–378. doi: 10.1039/C4MB00462K

Haridas, D., Ponnusamy, M. P., Chugh, S., Lakshmanan, I., Seshacharyulu, P., and Batra, S. K. (2014). MUC16: molecular analysis and its functional implications in benign and malignant conditions. *FASEB J.* 28, 4183–4199. doi: 10.1096/fj.14-257352

Helmrich, A., Ballarino, M., and Tora, L. (2011). Collisions between replication and transcription complexes cause common fragile site instability at the longest human genes. *Mol. Cell* 44, 966–977. doi: 10.1016/j.molcel.2011.10.013

Holly, M. K., Diaz, K., and Smith, J. G. (2017). Defensins in viral infection and pathogenesis. *Annu. Rev. Virol.* 4, 369–391. doi: 10.1146/annurev-virology-101416-041734

Hosokawa, M., Takeuchi, A., Tanihata, J., Iida, K., Takeda, S., and Hagiwara, M. (2019). Loss of RNA-binding protein Sfpq causes long-gene transcriptopathy in skeletal muscle and severe muscle mass reduction with metabolic myopathy. *iScience* 13, 229–242. doi: 10.1016/j.isci.2019.02.023

Hu, C., Chen, W., Myers, S. J., Yuan, H., and Traynelis, S. F. (2016). Human GRIN2B variants in neurodevelopmental disorders. *J. Pharmacol. Sci.* 132, 115–121. doi: 10.1016/j.jphs.2016.10.002

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37, 1–13. doi: 10.1093/nar/gkn923

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature* 431, 931–945. doi: 10.1038/nature03001

Iversen, L., and Kragballe, K. (2000). Arachidonic acid metabolism in skin health and disease. *Prostaglandins Other Lipid Mediat.* 63, 25–42. doi: 10.1016/S0090-6980(00)00095-2

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., et al. (2019). The reactome pathway knowledgebase. *Nucleic Acids Res.* 48, D498–D503. doi: 10.1093/nar/gkz1031

Jefferies, J. L., and Towbin, J. A. (2010). Dilated cardiomyopathy. *Lancet* 375, 752–762. doi: 10.1016/S0140-6736(09)62023-7

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361. doi: 10.1093/nar/gkw1092

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., and Tanabe, M. (2019). New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* 47, D590–D595. doi: 10.1093/nar/gky962

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., et al. (2004). The UCSC table browser data retrieval tool. *Nucleic Acids Res.* 32, D493–D496. doi: 10.1093/nar/gkh103

Kerrisk, M. E., Cingolani, L. A., and Koleske, A. J. (2014). ECM receptors in neuronal structure, synaptic plasticity, and behavior. *Prog. Brain Res.* 214, 101–131. doi: 10.1016/B978-0-444-63486-3.00005-0

Kirkconnell, K. S., Magnuson, B., Paulsen, M. T., Lu, B., Bedi, K., and Ljungman, M. (2017). Gene length as a biological timer to establish temporal transcriptional regulation. *Cell Cycle* 16, 259–270. doi: 10.1080/15384101.2016.1234550

Kuo, I. Y., and Ehrlich, B. E. (2015). Signaling in muscle contraction. *Cold Spring Harb. Perspect. Biol.* 7:a006023. doi: 10.1101/cshperspect.a006023

Kwok, J. C. F., Warren, P., and Fawcett, J. W. (2012). Chondroitin sulfate: a key molecule in the brain matrix. *Int. J. Biochem. Cell Biol.* 44, 582–586. doi: 10.1016/j.biocel.2012.01.004

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi: 10.1038/35057062

Lasky, J. L., and Wu, H. (2005). Notch signaling, brain development, and human disease. *Pediatr. Res.* 57, 104R–109R. doi: 10.1203/01.PDR.0000159632.70510.3D

Le Floc'h, N., Otten, W., and Merlot, E. (2011). Tryptophan metabolism, from nutrition to potential therapeutic applications. *Amino Acids* 41, 1195–1205. doi: 10.1007/s00726-010-0752-7

Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z., and Zhang, B. (2019). WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* 47, W199–W205. doi: 10.1093/nar/gkz401

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. *Cell Syst.* 1, 417–425. doi: 10.1016/j.cels.2015.12.004

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27, 1739–1740. doi: 10.1093/bioinformatics/btr260

Lin, T., Islam, O., and Heese, K. (2006). ABC transporters, neural stem cells and neurogenesis – a different perspective. *Cell Res.* 16, 857–871. doi: 10.1038/sj.cr.7310107

Maron, B. J., and Maron, M. S. (2013). Hypertrophic cardiomyopathy. *Lancet* 381, 242–255. doi: 10.1016/S0140-6736(12)60397-3

Massaly, N., Francès, B., and Moulédous, L. (2014). Roles of the ubiquitin proteasome system in the effects of drugs of abuse. *Front. Mol. Neurosci.* 7:99. doi: 10.3389/fnmol.2014.00099

McCoy, M. J., and Fire, A. Z. (2020). Intron and gene size expansion during nervous system evolution. *BMC Genomics* 21:360. doi: 10.1186/s12864-020-6760-4

Mei, L., and Nave, K. -A. (2014). Neuregulin-ERBB signaling in the nervous system and neuropsychiatric diseases. *Neuron* 83, 27–49. doi: 10.1016/j.neuron.2014.06.007

Mirina, A., Atzmon, G., Ye, K., and Bergman, A. (2012). Gene size matters. *PLoS One* 7:e49093. doi: 10.1371/journal.pone.0049093

Miura, P., Shenker, S., Andreu-Agullo, C., Westholm, J. O., and Lai, E. C. (2013). Widespread and extensive lengthening of 3' UTRs in the mammalian brain. *Genome Res.* 23, 812–825. doi: 10.1101/gr.146886.112

Monje, F. J., Kim, E. -J., Pollak, D. D., Cabatic, M., Li, L., Baston, A., et al. (2012). Focal adhesion kinase regulates neuronal growth, synaptic plasticity and hippocampus-dependent spatial learning and memory. *Neurosignals* 20, 1–14. doi: 10.1159/000330193

Noelanders, R., and Vleminckx, K. (2017). How Wnt signaling builds the brain: bridging development and disease. *Neuroscience* 23, 314–329. doi: 10.1177/1073858416667270

Ogretmen, B. (2018). Sphingolipid metabolism in cancer signalling and therapy. *Nat. Rev. Cancer* 18, 33–50. doi: 10.1038/nrc.2017.96

Palmer, D., Fabris, F., Doherty, A., and Freitas, A. A., and de Magalhães, J. P. (2019). Ageing transcriptome meta-analysis reveals similarities between key mammalian tissues. bioRxiv[Preprint]. 815381. doi: 10.1101/815381

Pipkin, M. E., and Monticelli, S. (2008). Genomics and the immune system. *Immunology* 124, 23–32. doi: 10.1111/j.1365-2567.2008.02818.x

Polleux, F., and Snider, W. (2010). Initiating and growing an axon. *Cold Spring Harb. Perspect. Biol.* 2:a001925. doi: 10.1101/cshperspect.a001925

Prentki, M., and Madiraju, S. R. M. (2008). Glycerolipid metabolism and signaling in health and disease. *Endocr. Rev.* 29, 647–676. doi: 10.1210/er.2008-0007

Rodriguez, J. M., Maietta, P., Ezkurdia, I., Pietrelli, A., Wesselink, J. J., Lopez, G., et al. (2013). APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Res.* 41, D110–D117. doi: 10.1093/nar/gks1058

Russo, D., Della Ragione, F., Rizzo, R., Sugiyama, E., Scalabrì, F., Hori, K., et al. (2018). Glycosphingolipid metabolic reprogramming drives neural differentiation. *EMBO J.* 37:e97674. doi: 10.15252/embj.201797674

Russo, E., Citraro, R., Constanti, A., and De Sarro, G. (2012). The mTOR signaling pathway in the brain: focus on epilepsy and epileptogenesis. *Mol. Neurobiol.* 46, 662–681. doi: 10.1007/s12035-012-8314-5

Sahakyan, A. B., and Balasubramanian, S. (2016). Long genes and genes with multiple splice variants are enriched in pathways linked to cancer and other multigenic diseases. *BMC Genomics* 17:225. doi: 10.1186/s12864-016-2582-9

Savarese, M., Sarparanta, J., Vihola, A., Udd, B., and Hackman, P. (2016). Increasing role of titin mutations in neuromuscular disorders. *J. Neuromuscul. Dis.* 3, 293–308. doi: 10.3233/JND-160158

Schnaar, R. L. (2016). Gangliosides of the vertebrate nervous system. *J. Mol. Biol.* 428, 3325–3336. doi: 10.1016/j.jmb.2016.05.020

Seif, F., Khoshmirsafa, M., Aazami, H., Mohsenzadegan, M., Sedighi, G., and Bahar, M. (2017). The role of JAK-STAT signaling pathway and its regulators in the fate of T helper cells. *Cell Commun. Signal.* 15:23. doi: 10.1186/s12964-017-0177-y

Seiradake, E., del Toro, D., Nagel, D., Cop, F., Härtl, R., Ruff, T., et al. (2014). FLRT structure: balancing repulsion and cell adhesion in cortical and vascular development. *Neuron* 84, 370–385. doi: 10.1016/j.neuron.2014.10.008

Simonti, C. N., and Capra, J. A. (2015). The evolution of the human genome. *Curr. Opin. Genet. Dev.* 35, 9–15. doi: 10.1016/j.gde.2015.08.005

Slominski, A., Zbytek, B., Nikolakis, G., Manna, P. R., Skobowiat, C., Zmijewski, M., et al. (2013). Steroidogenesis in the skin: implications for local immune functions. *J. Steroid Biochem. Mol. Biol.* 137, 107–123. doi: 10.1016/j.jsbmb.2013.02.006

Stark, C. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* 34, D535–D539. doi: 10.1093/nar/gkj109

Stocker, A. M., and Chenn, A. (2015). The role of adherens junctions in the developing neocortex. *Cell Adhes. Migr.* 9, 167–174. doi: 10.1080/19336918.2015.1027478

Stoeger, T., Gerlach, M., Morimoto, R. I., and Nunes Amaral, L. A. (2018). Large-scale investigation of the reasons why potentially important genes are ignored. *PLoS Biol.* 16:e2006643. doi: 10.1371/journal.pbio.2006643

Stoeger, T., Grant, R. A., McQuattie-Pimentel, A. C., Anekalla, K., Liu, S. S., Tejedor-Navarro, H., et al. (2019). Aging is associated with a systemic length-driven transcriptome imbalance. bioRxiv[Preprint]. 691154. doi: 10.1101/691154

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* 6:e21800. doi: 10.1371/journal.pone.0021800

Tacutu, R., Thornton, D., Johnson, E., Budovsky, A., Barardo, D., Craig, T., et al. (2018). Human ageing genomic resources: new and updated databases. *Nucleic Acids Res.* 46, D1083–D1090. doi: 10.1093/nar/gkx1042

Takeuchi, A., Iida, K., Tsubota, T., Hosokawa, M., Denawa, M., Brown, J. B., et al. (2018). Loss of Sfpq causes long-gene transcriptopathy in the brain. *Cell Rep.* 23, 1326–1341. doi: 10.1016/j.celrep.2018.03.141

Tao, S., and Sampath, K. (2010). Alternative splicing of SMADs in differentiation and tissue homeostasis. *Develop. Growth Differ.* 52, 335–342. doi: 10.1111/j.1440-169X.2009.01163.x

Taylor, R. G., Levy, H. L., and McInnes, R. R. (1991). Histidase and histidinemia. Clinical and molecular considerations. *Mol. Biol. Med.* 8, 101–116.

Urrutia, A. O., and Hurst, L. D. (2003). The signature of selection mediated by expression on human genes. *Genome Res.* 13, 2260–2264. doi: 10.1101/gr.641103

Valleriani, A., Zhang, G., Nagar, A., Ignatova, Z., and Lipowsky, R. (2011). Length-dependent translation of messenger RNA by ribosomes. *Phys. Rev. E* 83:042903. doi: 10.1103/PhysRevE.83.042903

van Dam, S., Craig, T., and de Magalhães, J. P. (2015). GeneFriends: a human RNA-seq-based gene and transcript co-expression database. *Nucleic Acids Res.* 43, D1124–D1132. doi: 10.1093/nar/gku1042

van Dam, S., Võsa, U., van der Graaf, A., Franke, L., and de Magalhães, J. P. (2018). Gene co-expression analysis for functional classification and gene-disease predictions. *Brief. Bioinform.* 19, 575–592. doi: 10.1093/bib/bbw139

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., et al. (2001). The sequence of the human genome. *Science* 291, 1304–1351. doi: 10.1126/science.1058040

Vig, M., and Kinet, J.-P. (2009). Calcium signaling in immune cells. *Nat. Immunol.* 10, 21–27. doi: 10.1038/ni.f.220

Vishnoi, A., Kryazhimskiy, S., Bazykin, G. A., Hannenhalli, S., and Plotkin, J. B. (2010). Young proteins experience more variable selection pressures than old proteins. *Genome Res.* 20, 1574–1581. doi: 10.1101/gr.109595.110

Wang, L., and Yi, R. (2014). 3'UTRs take a long shot in the brain. *BioEssays* 36, 39–45. doi: 10.1002/bies.201300100

Wang, T., Cui, Y., Jin, J., Guo, J., Wang, G., Yin, X., et al. (2013). Translating mRNAs strongly correlate to proteins in a multivariate manner and their translation ratios are phenotype specific. *Nucleic Acids Res.* 41, 4743–4754. doi: 10.1093/nar/gkt178

Ware, J. S., and Cook, S. A. (2017). Role of titin in cardiomyopathy: from DNA variants to patient stratification. *Nat. Rev. Cardiol.* 15, 241–252. doi: 10.1038/nrcardio.2017.190

Wolf, Y. I., Novichkov, P. S., Karev, G. P., Koonin, E. V., and Lipman, D. J. (2009). The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc. Natl. Acad. Sci.* 106, 7273–7280. doi: 10.1073/pnas.0901808106

Worthington, J. J., Fenton, T. M., Czajkowska, B. I., Klementowicz, J. E., and Travis, M. A. (2012). Regulation of TGFβ in the immune system: an emerging role for integrins and dendritic cells. *Immunobiol.* 217, 1259–1265. doi: 10.1016/j.imbio.2012.06.009

Wu, Y., Davison, J., Qu, X., Morrissey, C., Storer, B., Brown, L., et al. (2016). Methylation profiling identified novel differentially methylated markers including OPCML and FLRT2 in prostate cancer. *Epigenetics* 11, 247–258. doi: 10.1080/15592294.2016.1148867

Xu, F., Liu, C., Zhou, D., and Zhang, L. (2016). TGF-β/SMAD pathway and its regulation in hepatic fibrosis. *J. Histochem. Cytochem.* 64, 157–167. doi: 10.1369/0022155415627681

Yamagishi, S., Hampel, F., Hata, K., del Toro, D., Schwark, M., Kvachnina, E., et al. (2011). FLRT2 and FLRT3 act as repulsive guidance cues for Unc5-positive neurons. *EMBO J.* 30, 2920–2933. doi: 10.1038/emboj.2011.189

Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., et al. (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* 21, 650–659. doi: 10.1093/bioinformatics/bti042

Yang, D., Xu, A., Shen, P., Gao, C., Zang, J., Qiu, C., et al. (2018). A two-level model for the role of complex and young genes in the formation of organism complexity and new insights into the relationship between evolution and development. *EvoDevo* 9:22. doi: 10.1186/s13227-018-0111-4

Zeng, Y., Zhang, L., and Hu, Z. (2016). Cerebral insulin, insulin signaling pathway, and brain angiogenesis. *Neurol. Sci.* 37, 9–16. doi: 10.1007/s10072-015-2386-8

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., et al. (2018). Ensembl 2018. *Nucleic Acids Res.* 46, D754–D761. doi: 10.1093/nar/gkx1098

Zhan, T., Rindtorff, N., and Boutros, M. (2017). Wnt signaling in cancer. *Oncogene* 36, 1461–1473. doi: 10.1038/onc.2016.304

Zhang, T., de Waard, A. A., Wuhrer, M., and Spaapen, R. M. (2019). The role of Glycosphingolipids in immune cell functions. *Front. Immunol.* 10:90. doi: 10.3389/fimmu.2019.00090

Ziboh, V. A., Miller, C. C., and Cho, Y. (2000). Metabolism of polyunsaturated fatty acids by skin epidermal enzymes: generation of antiinflammatory and antiproliferative metabolites. *Am. J. Clin. Nutr.* 71, 361s–366s. doi: 10.1093/ajcn/71.1.361s

Zylka, M. J., Simon, J. M., and Philpot, B. D. (2015). Gene length matters in neurons. *Neuron* 86, 353–355. doi: 10.1016/j.neuron.2015.03.059

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.