Università degli Studi di Trento: Department of Cellular, Computational and Integrative Biology , 1–11
Date: 09/12/2024
Original Article

Human intronless genes: an updated analysis on evolution, peculiarities and associated diseases

Eylul Bulut,[1] Ilaria Massignani,[1] Sofia Pietrini[1] and Marte Toffoli[1]

[1]Department of Cellular, Computational and Integrative Biology, Università degli Studi di Trento, Trento, Italy

Abstract

The structure of eukaryotic genes typically consists of exons interrupted by intragenic non-coding DNA regions (introns) removed by RNA splicing to generate the mature mRNA. A fraction of genes, however, comprise a single coding exon with introns in their untranslated regions or are single-exon genes (SEGs), lacking introns entirely. Single-exon genes constitute approximately 12% of the human genome. Many studies, focused on SEGs functional analysis, have been conducted, often with a limited gene dataset taken into consideration. Here we present an updated functional analysis of single-exon genes using a more comprehensive and current gene dataset.

Our functional analysis of SEGs has displayed tissue-specific expression pattern, predominantly in testis or neuro-specific tissues, with a significant percentage of genes expressed exclusively in tissues classified as "high-turnover". The analysis also displayed an over-representation in genes associated with sensory perception, regulation activities, immune responses, protein homeostasis, transcription, and cellular communication. These characteristics translate into associated diseases, mainly neuropathies, developmental disorders, and cancer, where, for some of them, meaningful and not random gene-disease relationships have been found.

Moreover, SEGs display a higher susceptibility to m6A modifications compared to multi-exon genes (MEGs). However, this enrichment doesn't appear to translate into a stronger association between m6A modifications and disease. Despite the unique functional and molecular characteristics of SEGs, the strength of their correlation with diseases remains comparable to that of other genes.

Key words: Single-exon genes, SEGs, MEGs, Enrichment analysis, N6-methyladenosine

Introduction

Single-exon genes are characterized by the absence of introns in their nucleotide sequence. Introns are portions of non-coding DNA that can be removed through a splicing process.

In eukaryotic genomes, multi-exon genes are extremely common, they, indeed, constitute the majority of the human genome. The presence of introns, however, entails some disadvantages, including high energy consumption, a reduction in the rate of protein production and the risk of introducing errors during the splicing process.

Although single-exon genes are more efficient than intron containing ones, in organisms considered to be more evolved they are present within their genome with a percentage of about 3% (1). The evolutionary advantage of multi-exon genes (MEGs) is due to alternative splicing, which allows the generation of different proteins starting from a single gene, increasing variability. Furthermore, the splicing process allows us to distinguish between correct and defective coding messengers as they are able to identify premature stop codons. In fact, when a stop codon is positioned too close to the exonic junction complex, mRNA degradation is induced. This mechanism not only eliminates incorrect mRNAs, but also regulates some perfectly functional messengers (2).

The presence of introns is therefore essential in organisms with a complex embryogenesis process, in neuronal development and in immune responses. This raises questions about why such complex organisms retain a small percentage of single-exon genes.

A previous published study allowed us to characterize intron-deprived genes, providing information on their functional classification and on their associated diseases(1). Our study aims to reproduce the analyses carried out in the cited study, using a larger database of single-exon genes. The database used in our work includes over 2000 genes, of which 1962 can be used for functional analyses; this number corresponds to approximately three times the genes analyzed in the reference paper, allowing us to conduct more precise and in-depth analyses for a better characterization of this type of genes. This study mainly focuses on performing functional analyses on a larger database, with the aim of obtaining more accurate and complete results, and verify whether the results obtained in the reference paper remain confirmed.

In addition to the gene enrichment analysis also performed in the reference paper, further investigations have been conducted in order to characterize this group of genes in more depth. These analyses provided us with clues as to why these genes are still present within the genome. In particular, one of the analysis aims to verify whether the level of gene expression in the various tissues can represent a discriminating factor to determine the presence of introns.

We then searched for data regarding the diseases that are enriched in the paper (1) and conducted differential expression analysis on the datasets retrieved, to better understand the relevance of single-exon genes in disease phenotypes. The diseases investigated include: Colorectal Carcinoma, Familial Melanoma, Myelodysplastic syndrome and Acute Myelogenous Leukemia, Sertoli Cell-Only Syndrome, Williams Syndrome e Hermansky-Pudlak Syndrome type 6.

Another portion of the study focuses on the impact that RNA modifications can have on this type of genes, with a particular focus on the m6A modification (adenosine methylation at site 6), a modification that affects messenger RNA and involves the addition of a methyl group to the nitrogen atom in position 6 of the adenine ring.

This methylation is relevant in several stages of the RNA cycle, including transcription regulation, maturation, translation, degradation and mRNA stability. RNA m6A methylation may be involved in the regulation of physiological and pathological processes, including oncological mechanisms (3). This observation prompted us to investigate to what extent single-exon genes are related to this type of modification and whether there is a correlation between the amount of methylation and the diseases most associated with this set of genes.

The workflow we used for this purpose is described in the figure 1.

Materials and Methods

Our starting point for this project was the SinEx DataBase 2.0 (4), an updated version of the SinEx DB, which remains the most comprehensive database for single-exon genes built to date. SinEx provided an updated dataset of SEGs and a curated source of information, including GO functional categorizations for all our genes, Fig.2, which facilitated the preliminary analysis of the dataset. Data were accessed via the SinEx web interface, where we were able to download protein sequences of SEGs in FASTA format as well as SEGs functional assignation and SEI information.

After extracting the list of SEGs, we used Enrichr (5) for enrichment analysis, leading us to download and use data from external sources such as GTEx (6), for tissue expression, Orphanet (7), for correlated diseases analysis, and GO(8), for biological process annotation. Tissue expression data were further analyzed using data mining techniques implemented in Python.

Additional data was collected from the Gene Expression Omnibus (GEO)(9), enabling us to retrieve expression data for correlated diseases and perform differential expression analysis on this data using GEO2R. After identifying significantly differentiated single-exon genes for each disease, the OneGene-Web-Weaver (10) was used to expand each of these lists.

In the final phase of the project, we retrieved all human genes with m6A modifications via the M6A-Atlas v2.0 API (11). We converted all the data in .csv files through Python, we employed Ensembl BioMart (12) to retrieve some extra genes information for our dataset and we performed a first analysis on this data using R. Lastly, we used the results of the previous differential

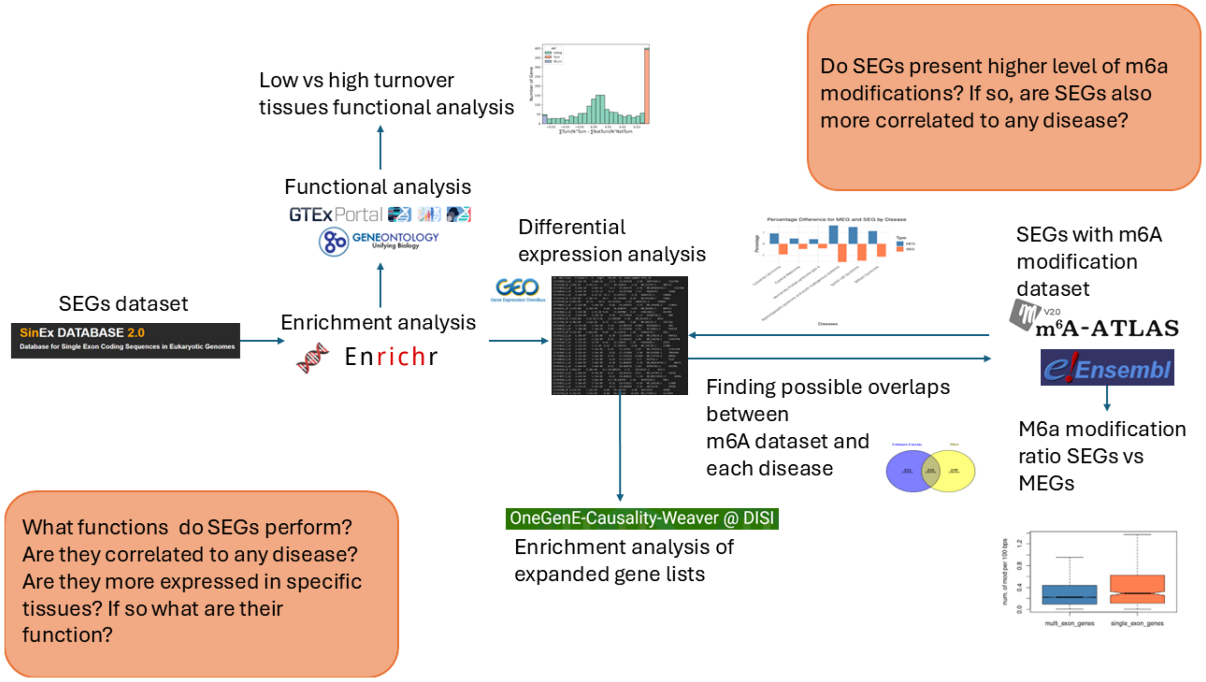
Low vs high turnover
tissues functional analysis
Do SEGs present higher level of m6a modifications? If so, are SEGs also more correlated to any disease?
Functional analysis
GTExPortal
GENEONTOLOGY
Unifying Biology
Differential
expression analysis
Percentage Difference for MEG and SEG by Disease
SEGs with m6A
modification
dataset
GEO
Gene Expression Omnibus
SEGs dataset
SinEx DATABASE 2.0
Database for Single Exon Coding Sequences in Eukaryotic Genomes
Enrichment analysis
Enrichr
V2.0
m6A-ATLAS
Ensembl
Finding possible overlaps
between
m6A dataset and
each disease
M6a modification
ratio SEGs vs
MEGs
What functions do SEGs perform? Are they correlated to any disease? Are they more expressed in specific tissues? If so what are their function?
OneGenE-Causality-Weaver @ DISI
Enrichment analysis of
expanded gene lists
multi_exon_genes
single_exon_genes

Fig. 1: Problem definition and solution pipeline

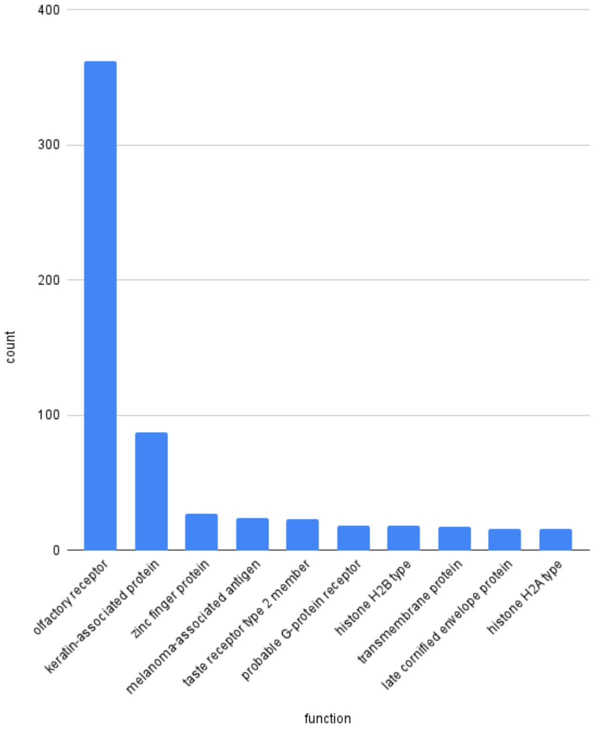

Fig. 2: GO functional categorization

expression analysis to perform a comparison between single-exon genes and multi-exon genes correlation to diseases. Our aim was to understand if sigle-exon genes affected by the m6A modification could be more positively correlated to diseases than other genes.

Identification and Preliminary Analysis of Single-Exon Genes

Using the Python programming language, the list of single-exon genes was extracted and used for subsequent analyses. In addition, a list of encoded proteins was generated, from which a histogram was produced to display the types of proteins most expressed.

Gene Enrichment Analysis

To verify the preferential expression of the genes present in the dataset, a gene enrichment analysis was performed using the EnrichR software (5). This software, using the GTEx Tissue database (6), identified the tissues in which the genes are most expressed, and also allowed us to perform Gene Onthology analyses (8) that provided data on the cellular components, biological processes and molecular functions in common in the gene set. The results include a table containing information such as p-values and rankings, which allowed the construction of a histogram representing the 10 more significant results.

Filtering and Analyzing Gene Expression Data

The database provided by GTEx was filtered to include:

- Gene expression level information for each tissue for all genes in the SEG dataset.
- A random set of multi-exon genes equal in number to the initial single-exon genes.

Each gene was labeled to indicate its origin (single-exon or multi-exon). Genes with zero expression levels in all tissues were eliminated. Another gene that we removed is CDC42EP3, this gene in fact has out of order coordinates that did not allow us to correctly visualize the PCA. This is due to the high level of expression of the gene in almost all tissues except the brain. The remaining data were standardized and a Principal Component Analysis (PCA) was performed using the Python Scikit-learn library (13). Each gene was represented with a different color based on its classification (single-exon or multi-exon).
A K-Nearest Neighbors clustering algorithm (14) was then applied to identify two distinct clusters separating single-exon genes from multi-exon genes.

Gene Expression Study in TurnOver and Non-TurnOver Tissues

To verify the distribution of gene expression in TurnOver and Non-TurnOver tissues, the GTEx database was used, which contains the gene expression levels for each gene in each tissue. The database was filtered to include only the genes of interest, excluding those with insufficient expression levels.
Three tissue lists were created: TurnOver, Non-TurnOver and Unknown, based on predefined categorizations:

- **Turnover tissue:** Adipose Subcutaneous, Adipose Visceral Omentum, Bladder, Colon Sigmoid, Colon Transverse, Colon Transverse Mixed Cell, Colon Transverse Mucosa, Esophagus Gastroesophageal Junction, Esophagus Mucosa, Liver, Liver Hepatocyte, Liver Mixed Cell, Liver Portal Tract, Lung, Skin Not Sun Exposed Suprapubic, Skin Sun Exposed Lower leg, Small Intestine Terminal Ileum, Small Intestine Terminal Ileum Lymphode Aggregate, Small Intestine Terminal Ileum Mixed Cell, Spleen, Stomach, Stomach Mixed Cell, Stomach Mucosa, Testis, Thyroid, Whole Blood.
- **Non-turnover tissue:** Brain Amygdala, Brain Anterior cingulate cortex BA24, Brain Caudate basal ganglia, Brain Cerebellar Hemisphere, Brain Cerebellum, Brain Cortex, Brain Frontal Cortex BA9, Brain Hippocampus, Brain Hypothalamus, Brain Nucleus accumbens basal ganglia, Brain Putamen basal ganglia, Brain Spinal cord cervical c-1, Brain Substantia nigra, Heart Atrial Appendage, Heart Left Ventricle, Kidney Cortex, Kidney Medulla, Muscle Skeletal, Nerve Tibial, Ovary, Pancreas, Pancreas Acini, Pancreas Islets, Pancreas Mixed Cell, Prostate, Vagina, Colon Transverse Muscularis, Stomach Muscularis.
- **Unknown:** Adrenal Gland, Artery Aorta, Artery Coronary, Artery Tibial, Breast Mammary Tissue, Cells Cultured fibroblasts, Cells EBV-transformed Lymphocytes, Cervix Ectocervix, Cervix Endocervix, Esophagus Muscularis, Fallopian Tube, Minor Salivary Gland, Pituitary, Uterus.

For each gene, the expression levels were transformed into percentages to determine the relative expression of the gene in each tissue.
Subsequently, the means of the percentages in TurnOver and Non-TurnOver tissues were calculated for each gene. We then established a threshold that defines within which margins a gene can be considered preferentially expressed in TurnOver or Non-TurnOver tissues. From the data thus obtained, it was constructed

a histogram that contained the difference between the means of the percentages in the TurnOver and Non-TurnOver tissues on the x-axis and the frequency of genes for each interval on the y-axis, finally the genes above the threshold were highlighted using different colors. Furthermore, a number of bins equal to 24 was used to construct the histogram (defined by Rice's rule). Finally, the percentages of genes expressed in the TurnOver, Non-TurnOver and Unknown tissues were calculated, using three different thresholds (95%, 90%, 85%).

Diseases Associated With Single-Exon Genes

Many studies have associated single-exon genes with various diseases, including neurological or developmental disorders, neuropathies, eye malformations, infertility-associated syndromes and several types of cancers (1). Further investigation using Orphanet confirmed this association, but failed to provide information about additional diseases that could update the table of correlated diseases previously estabilished in (1).
To extend the analysis, we then employed the Gene Expression Omnibus (GEO) (9), to search for data related to diseases already known to be correlated to SEGs. Our focus was on identifying suitable datasets for differential expression analysis. We were able to successfully retrieve and analyze data from six conditions out of approximately twenty correlated diseases:

- *Cancer:* Colorectal Carcinoma (15), Familial Melanoma (16), Myelodysplastic syndrome and Acute Myelogenous Leukemia (17);
- *Developmental Disoders:* Sertoli Cell-Only Syndrome (18), Williams Syndrome (19);
- *Other diseases:* Hermansky-Pudlak Syndrome type 6 (20).

Subsequently, the GEO2R web tool was used to compare the groups of samples within the GEO Series previously cited, in order to identify differentially expressed genes across experimental conditions. In each case, the comparison was conducted between control and disease groups.
The entire sets of results, including the corresponding volcano plots, were downloaded and filtered to isolate SEGs from the larger dataset, thereby obtaining a preliminary understanding of their relevance and significance in each specific disease.
From the six networks of single-exon genes associated with the previous conditions, five were later expanded using OneGenE-Causality-Weaver (10). Genes associated with Familial Melanoma were excluded from this analysis, due to insufficient significant information. Due to hardware limitations, lists of the top 25 to 30 differentially expressed single-exon genes for each disease were picked as the seeders of the network. The range of the OneGenE Relative Frequence to consider was always left at default settings, on the contrary the "Connected node search type" filter was set to "Only nodes shared between 2 or more seed nodes".
The resulting expanded gene lists were then subjected to enrichment analysis to identify potential pathways or functional associations.

An In-Depth Look At 6 Correlated Diseases

The top 25 most expressed single-exon genes in the gene list obtained by differential expression were presented separately for each disease using EnrichR(5). KEGG 2021 Human library (21) was used to understand the locations and roles of genes in biological processes with human pathway analysis. The bar graph and clustergram data resulting from the use of this library were analyzed.
Only processes with p-values of 0.05 and below were considered. A clustergram was used to analyze the expression profiles of genes and visualize these data. These results were used to study the Monarch Initiative resource's relationship with genes, pathways and diseases(22).

M6A Analysis

N6-methyladenosine (M6A) is the most abundant modification of mRNA, is essential for normal development and its dysregulation promotes cancer. The modification is installed at a consensus motif, typically represented as a DRACH motif (D = A/G/T, R = A/G, H = A/C/T), and is catalyzed by the methyl-transferase (MTase) complex, of which METTL3 and METTL14 are the core components (23). In literature, a peak of this DRACH motif is observed near the stop codon of mature mRNAs, while is quite evenly distributed elsewhere. Notably, hyper-methylated sites are strongly enriched within short internal exon of MEGs, but, intriguingly, the level of M6A is even higher in SEGs (24).
Despite the importance of this modification, no specific analysis has been conducted on single-exon genes.In fact, there isn't any information about the implications of m6A in diseases or in relation to single-exon genes in literature. To address this, we performed an analysis to explore whether m6A's association with diseases could provide new insights.
We started from M6A-Atlas v2.0 database (11), that contains all known m6A modifications. Using its API we were able to download a comprehensive file containing all human genes affected by m6A. This file was structured as a .csv file, where each gene, with its correlated information, was listed multiple times, corresponding to the number of modifications it has. Using Python we classified genes into SEGs and MEGs and extracted the number of m6A for each gene of both categories, resulting in 447 single-exon genes and 11754 multi-exon genes.
To gather some more needed information about the m6A genes, we used Ensembl BioMart (12), which grant us to obtain the length of each gene. This enable us to compute on R the ratio between the number of m6A modifications per 100 nucleotides. This process allowed us to visualize which of the two categories (SEG or MEG) exhibited a higher density of m6A modifications. (Figure 10)
After this preliminary analysis on the level of modification of single-exon genes, we still wanted to address whether m6A influences SEGs diseases correlation, since we lacked evidence on that. We considered the diseases for which we were able to obtain differential expression analysis data and we used R to compare:

- The percentage of all MEGs with m6A and all SEGs with m6A that appear to be present within the most significative genes correlated to diseases (p-value < 0.05);
- The percentage of significative MEGs and SEGs (p-value $<$ 0.05) that are present for each diseases.

This comparison aimed to determine if the two categories were more or less correlated to those six diseases and it was performed by calculating the difference between the percentage of disease-correlated genes subjected to m6A and the percentage of genes of the same category that were correlated to the same disease.
The same comparison was performed considering the most modified genes with at least one m6A per 100 bp.

Results

The enrichment analysis performed on the entirity of our SEGs dataset gave us a first overlook of what data we were working on and what data to consider in the first steps of our analysis. As shown in Fig. 3, we were able to obtain a list of the top 10, and later on also the top 25, tissues with the highest gene enrichment. An unusual high proportion of SEGs exhibit testis or neuro-specific expression patterns, according to the GTEx Tissue V8 2023 Table (6). The database shows, indeed, that testis account for 4 out of the top 10 most enriched tissues, with an age bracket spanning from 20 to 59 years. Neuro-specific tissues also are prominent, with the hypothalamus, nucleus accumbens (basal ganglia), cortex and nerve tibial displaying significant expression levels. Among these, the tibial nerve is the most enriched tissue in our dataset. Other noteworthy tissues include the bladder and skin – sun exposed (lower leg), which rank among the top 25.

Fig. 4, instead, illustrates the 10 main molecular functions associated with the analyzed gene set. A significant portion of our genes code for proteins responsible for sensory perception, such as the detection of chemical stimuli and the perception of smell. Another common functional group among SEGs includes genes associated in regulatory activities involved in the transcription process and immune response.

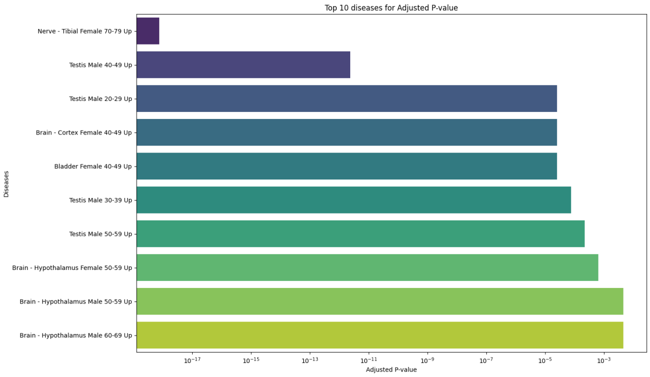

Fig. 3: Top 10 most enriched tissues, ordered by adjusted p-value

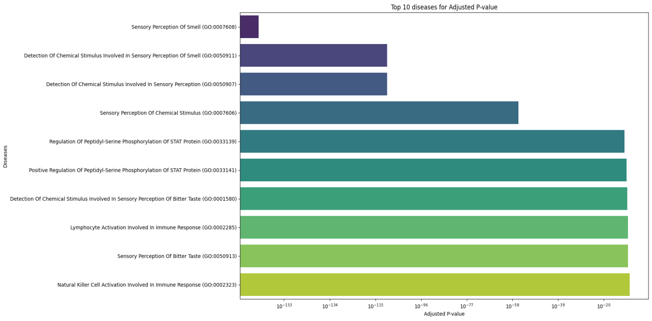

Fig. 4: Top 10 molecular functions associated with SEGs, ordered by adjusted p-value

Fig. 5 shows the Principal Component Analysis (PCA) performed on tissue expression data of the genes present in the SEG database and an equivalent set of multi-exon genes. As highlighted in the figure, the genes are not clearly separable into the two categories. This result is further confirmed by the application of the k-Nearest Neighbors (k-NN) algorithm, which produced an accuracy of 59.45%.

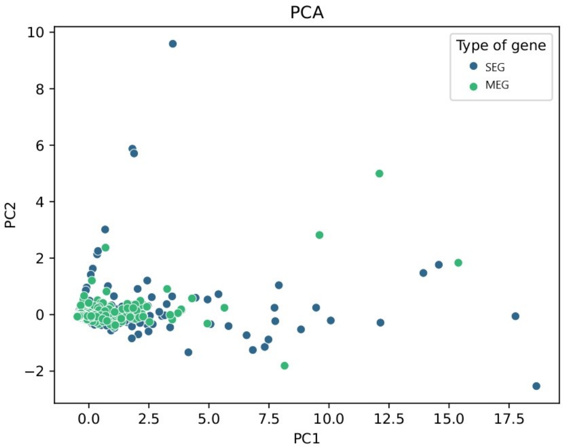

Fig. 5: PCA gene expression SEGs vs MEGs, each point in the figure corresponds to a gene, the colors identify the type of gene, blue if the gene is single exon, green if it is multi-exon.

Fig. 6 shows a histogram representing the distribution of gene expression between TurnOver and Non-TurnOver tissues. Genes preferentially expressed in TurnOver and Non-TurnOver tissues, with values above the pre-established threshold, are indicated in pink and blue respectively, while genes whose expression difference does not exceed the threshold are represented in green. Table 1 reports the quantities of genes classified as TurnOver and Non-TurnOver based on the different thresholds set, providing a detailed view of the distribution in the two categories.

Threshold	Turn	Non-Turn	% Turn	% Non-Turn
95%	397	43	23%	2,5%
90%	436	70	25%	4%
85%	482	93	27%	5,3%

Table 1. The table below shows the number of genes expressed exclusively in TurnOver or Non-TurnOver tissues, based on the selected threshold. Furthermore, their percentage within the analyzed set is indicated.

We subsequently started the functional analysis on genes related to the six disease previously listed. The differential expression analysis performed for each disease, showed significant results regarding our dataset. For Colorectal Carcinoma and Sertoli Cell-Only Syndrome the number of genes which exhibited an adjusted p-value of less than 0.05 was more than 450, with many of them having also a significant log-fold change value, as shown in Fig.7.

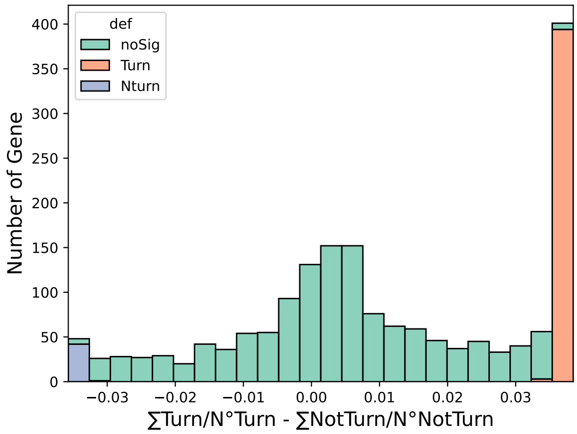

Fig. 6: The figure represents a histogram that illustrates the difference between the averages of the expression levels. In pink are highlighted the genes that are expressed exclusively in the Turnover tissues while in blue are the genes expressed exclusively in non-Turnover tissues. In green instead are represented all the genes that have not reached the 95% threshold necessary to be considered expressed in a specific category.

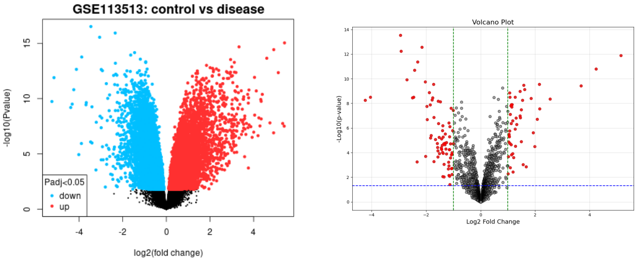

Fig. 7: Volcano plot resulting from differential expression analysis of Colorectal Carcinoma related genes. All correlated genes vs correlated SEGs only. The threshold for the adjusted p-value was set as 0.05. In the plot displaying only SEGs, only genes with $|log2fc| > 1$ were highlighted.

Fig. 8 shows the enrichment analysis performed for Colorectal Carcinoma(CC), the first 6 processes were found to have a relationship. Fig. 9 shows CLDN8 and CLDN23 SEGs were observed as the genes with the highest relationship with the evaluated processes. This result may indicate that CLDN8 and CLDN23 genes play an important role in the pathogenesis of CC disease. The pathways found to be significant according to the P-value and the genes highlighted in the clustergram analysis provide important clues about the tumor biology and microscopic environment of CC. The Tight Junction pathway we observed includes connections that regulate cellular barrier functions(25). CLDN8 and CLDN23 SEGs highlighted in the clustergram are also tight junction proteins and support our results. Loss of expression in these two genes is associated with tumor growth and metastasis(22).

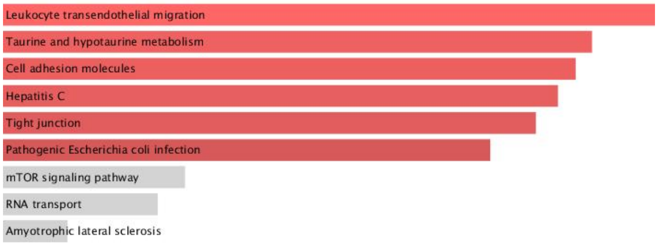

Fig. 8: Pathways obtained from the KEGG 2021 Human library analysis of Colorectal Carcinoma(CC)

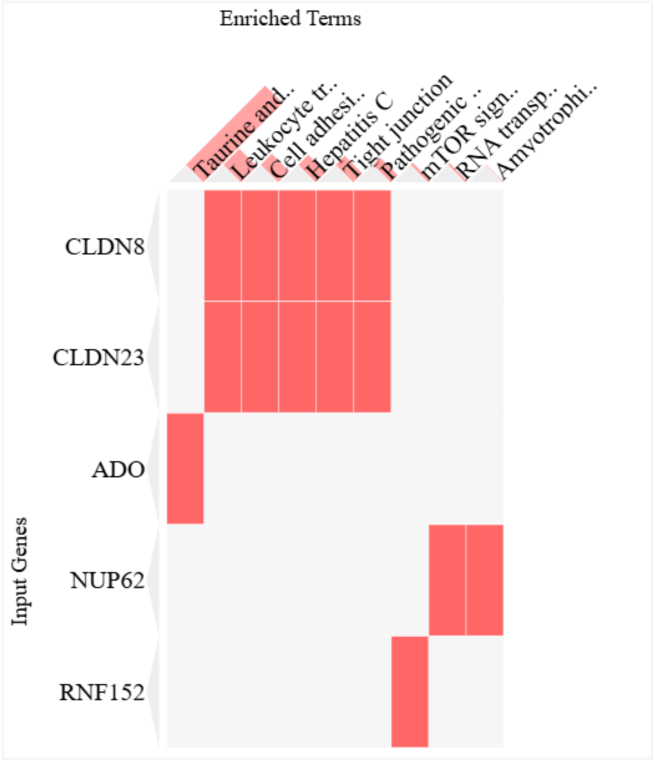

Fig. 9: Clustergram graph results in the KEGG 2021 human library. CLDN8 and CLDN23 genes were the most observed genes

In the analysis performed for Familial Melanoma(FM), only the Olfactory Transduction pathway was shown to be associated with the disease significantly. CLDN4 SEG was observed as the gene with the highest association with the evaluated processes. This result showed that CLDN4 gene plays a critical role in the development of melanoma, it helps to form tight junctions in epithelial cells and maintain cell integrity(22). The olfactory transduction pathway we observed is known to play a role in regulating cellular growth, differentiation and stress responses(26). The analysis performed for Hermansky-Pudlak Syndrome Type 6(HPS-6), it highlighted 9 different processes. In the clustergram, CREB5 AND ADRB2, SEGs were observed as the genes with the highest association with the evaluated processes. The cGMP-PKG Signaling and PI3K-Akt Signaling pathways we observed are known to be related to cellular growth, metabolism, survival and immune response(27). It is known that CREB5 and ADRB2 SEGs, which are densely found in the clustergram, play a role in regulating transcriptional activity and immune response(22).

In the analysis performed for Acute Myelogenous Leukemia(AML), it was observed that 3 different processes could be significant. In the clustergram, the PABPC3 SEG was observed as the gene with the highest association with the evaluated processes. This result showed that PABPC3 SEG could explain the molecular mechanism of the disease. The mRNA Surveillance and Splicesome pathways we observed play a role in correcting genetic errors and clearing unstable mRNAs(28). The PABPC3 gene, which is abundant in the clustergram, regulates mRNA stability, translation, and degradation(22).

In the analysis performed for Sertoli Cell-Only Syndrome(SCS), it was observed that 6 different processes could be associated with p-value ranks. In the clustergram , PRKACG and PDHA2, SEGs were observed as the genes with the highest connection with the evaluated processes. The Glycolysis/gluconeogenesis, Olfactory Transduction and Glucagon Signaling pathways, play a role in cell energy production and cellular signal transmission(29)(26). The PRKACG and PDHA2 genes, which are densely found in the clustergram, play a role in regulating intracellular biological processes(22).

In the analysis performed for Williams Syndrome(WS), it was observed that only 1 process could be associated with the p-value which is Complement and Coagulation Cascade. In the clustergram, the FZD2 SEG was observed as the gene with the highest connection with the evaluated processes. The Complement and Coagulation Cascade pathway we observed may increase vascular inflammation, endothelial damage, and clotting tendency(30). The FZD2 gene, which is abundant in the clustergram, may cause problems in cell proliferation and differentiation(22).

As a result of all these analyses, we can say that the effect of the genes we observed in the clustergram in these diseases is not random, but rather occurs due to a specific biological or pathological mechanism. The basis of these analyses is based on the statistical evaluation of gene expression data and the biological pathways in which these genes are enriched. In other words, the results are meaningful in a biological context and are not random. Once obtaining a basic understanding of what single-exon genes' role could be, especially in relation with possible correlated diseases, we moved on to the last portion of our analysis.

Fig. 10 shows that SEGs tend to be more affected by m6A modification than MEGs. On the x-axis we display the ratio's value, which represent the number of m6A modifications per 100 base pairs > 1. A Wilkson test was conducted to asses the statistical significance of the observable differences between the two boxplots. The test resulted in a rejection of the null hypothesis (the two means being equal), with a p-value of 2.844e-06, indicating that the difference between the two boxplots is statistically significant. After finding out a significant overlap between single-exon genes correlated with the diseases previously studied and single-exon genes presenting m6A modification, our last analysis aimed at studying a possible correlation between the two. Fig. 11 shows the result of the percentage difference between the m6A genes that are correlated with a specific disease and the percentage of genes of the same category (SEGs or MEGs) that are correlated to the same disease. We computed this difference for all our genes and for each disease. The resulted graph indicate that, generally, SEGs subjected to m6A tend to be less correlated to the six diseases than MEGs, though not at a significant level.

Given this outcome, we supposed that the amount of m6A in SEGs is negatively correlated to diseases. We then performed the same process on the most modified genes. We considered the ones with a number of m6A modifications per 100 bp larger than 1. The result is shown in Fig. 12 and tells us that some diseases presented a slightly increased percentage of correlation with SEGs with m6A (Colorectal Carcinoma, Hermansky Pudlak syndrome type 6, Myelodyspastic syndrome and AML, Sertoli Cell Syndrome) while the other diseases don't show any substantial difference in correlation between SEGs and MEGs.

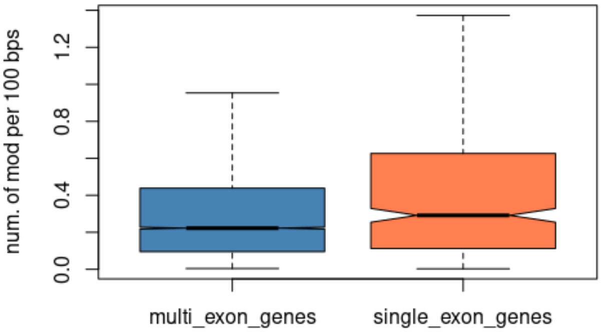

Fig. 10: Boxplots of SEGs and MEGs' ratio

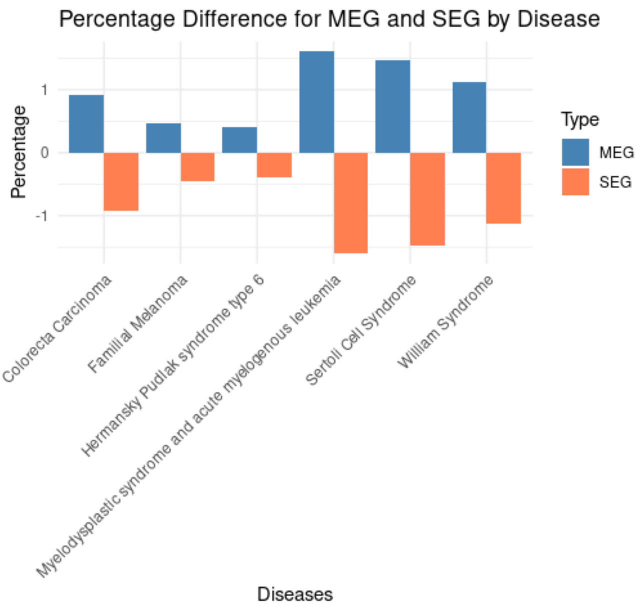

Fig. 11: Barplot of difference in correlation with diseases

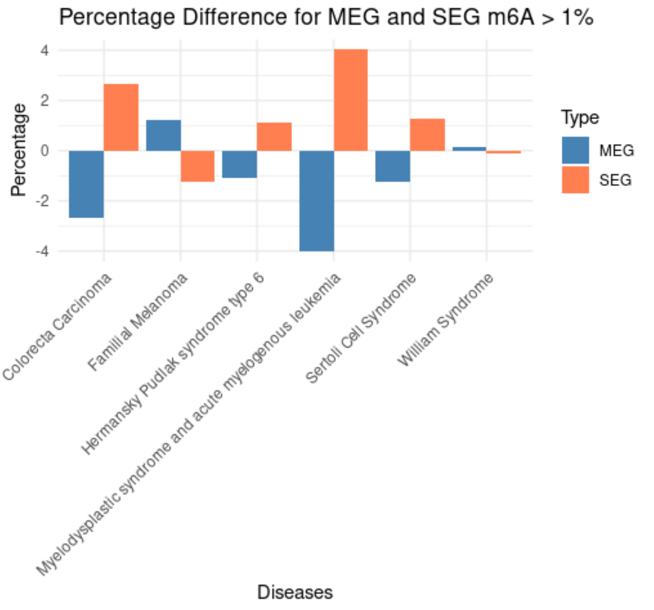

Fig. 12: Barplot of difference in correlation with diseases for SEGs and MEGs with at least a m6A modification per 100 bps

Discussion and Conclusion

The single-exon gene dataset shows significant enrichment in functions related to odor perception and translation regulation. This result is consistent with the prevalence of genes encoding for olfactory receptors (figure 2) in the gene set and with what is reported in the reference paper (1). The correlation suggests that these genes may have a specific role in sensory processing and fine-tuning of protein synthesis.

Gene enrichment analysis on tissues revealed a variety of involved tissues, including both tissues with high cell TurnOver and tissues with lower TurnOver. This indicates that single-exon genes can be expressed in very heterogeneous biological contexts. A more in-depth analysis (Fig. 6) highlighted that approximately 23% of the genes in the set are expressed exclusively in high-TurnOver tissues. These genes are mainly involved in odor perception, in the response processes to chemical stimuli, but also in gonadal development. Most of these genes, which encode olfactory receptors, are exclusively expressed in high TurnOver tissues, for biological and functional reasons related to the nature of olfactory perception and the need to maintain regenerative efficiency (31).

The data reported in the table 1 show an effective increase in the genes considered significant although not uniform. This is due to a high number of genes that present an expression level equal to 0 in the Non-TurnOver tissues.

The PCA analysis shows that gene expression profiles in different tissues do not allow to clearly distinguish single-exon genes from the others. This result is further confirmed by using the k-Nearest Neighbors algorithm, which presents an accuracy of 59.45%, a value close to what would be expected from a random classification. This suggests that gene expression levels, alone, do not represent an effective discriminating criterion for this categorization.

In the article we initially referenced (1), a single-exon gene was shown to be linked to each of the diseases mentioned. We, on the other hand, were able to link a list of genes related to 6 of these 20 diseases, that, in some cases reached the 400 units. We contributed to the gene and disease connection by providing in-depth information as a result of the Enrichr analysis we conducted. We were able to obtain a more complete set of information by doing pathway enrichment analysis on the expanded lists of genes generated by the OneGenE-web-weaver. One of the most significant findings involves processes related to cell division and genomic stability. Alterations in these pathways, such as chromosomal separation and mitotic regulation, frequently lead to cell cycle dysfunction. This is particularly evident in conditions like Colorectal Carcinoma, where such dysfunctions contribute to cancer development and progression. These processes are also implicated in autoimmune disorders and metabolic syndromes, underscoring their broad relevance across disease types. Protein regulation emerged as another recurring theme. Dysfunctions in mechanisms such as ubiquitination, deubiquitination, and ribosomal biogenesi are linked to neurodegenerative diseases and cancers, where errors in protein homeostasis can disrupt cellular integrity and metabolism. This is further supported by findings in diseases like Sertoli Cell Syndrome, where genes involved in protein synthesis and responses to metabolic stress appear to be critical. The analysis also revealed the central importance of DNA and RNA processes. In conditions like Myelodysplastic Syndrome, many genes were associated with transcription and DNA repair, processes essential for maintaining genomic integrity. These genes were predominantly expressed in the nucleus and intracellular organelles, emphasizing their role in orchestrating fundamental genetic activities. Another significant finding involves cellular communication and adhesion. In Hermansky-Pudlak Syndrome, genes regulating cell adhesion, intracellular trafficking, and intercellular communication were strongly associated with the disease. Developmental and metabolic disorders also revealed distinctive yet overlapping pathways. In developmental conditions like Williams Syndrome, genes involved in hormonal signaling, bone development, and sensory perception were highlighted. These pathways are critical for proper differentiation and developmental progression, and their dysregulation may lead to developmental disorders or hormonal imbalances.

Single-exon genes set appear to be more subject to m6A modification with statistical significance, but a further analysis has shown that the genes with m6A don't present an enhanced correlation with diseases with respect to the genes without m6A.

In conclusion, the study confirmed most of the results reported in the paper (1). Functional analyses also highlighted an enrichment in functions related to odor perception, attributable to the presence of a high amount of genes encoding olfactory receptors in the database used, which were not present in the database used in the cited study. Differential expression analyses identified several genes relevant to the pathological phenotype, compared to the single-exon gene highlighted in the first study. Furthermore, the study of m6A modifications allows to further characterize the set of genes, providing additional information on the correlation between pathological phenotype and quantity of modifications.

Limitations of the study

The availability and reachability of data regarding the various diseases we wanted to analyze, put a limit in the number of conditions we were able to study. We encountered numerous problems when searching for data in DISGENET, therefore we had to rely only on databases available on GEO.

For our enrichment analysis of single-exon genes and the diseases they are linked to, in the pathway graphs given according to the p-value order, only one pathway was observed in Familial Melanoma and Williams Syndrome. Since it is an important step in understanding whether these genes are strongly linked to diseases, a literature search was also conducted and the results were obtained.

Lastly, we wanted to conduct an additional analysis to validate whether the saturation of methylated motifs per gene was comparable to the previously computed ratio. This was made impossible by the temporary unavailability of Ensembl. We would have assessed whether the ratio between the total number of motifs and the total number of modification would give us comparable results to the m6A/length ratio for the two categories using BiocManager and Biostrings library. This new ratio would have been calculated by retrieving the coding sequence (cDNA) for each gene from Ensembl BioMart and counting the total number of DRACH motifs on each sequence. Using this result, we would have been able to compute the ratio

$$\frac{NumberOfMethylatedMotifs * 100}{TotalNumberOfMotifs}$$

for each gene and generate a boxplot to do a comparison with the one obtained from the first analysis.

Future directions

The findings from this study leave space for further exploration to investigate the biological role of single-exon genes. In particular, the tissue-specific expression of SEGs in high-TurnOver tissues and neuro-specific ones raises question about SEGs possible role in development, reproduction, or neurological processes not yet investigated.

A more detailed investigation on correlated diseases can be made with the availability of more complete data. Another possible route would be to investigate on whether or not SEGs could serve as potential biomarkers or therapeutic targets. In addition, re-analysis of differential expression analysis data using statistical tests can help to resolve whether these genes are randomly scattered or not.

Although SEGs exhibit an enrichment in m6A modifications, this study found no significant correlation between these modifications and stronger disease associations. The m6A analysis would benefit greatly from a more deep analysis taking into consideration more aspects of the genes, as considering different transcript and not just the Ensembl's canonical, or verify whether the position of m6A modifications on the sequence influences diseases correlation.

Authors Contributions

- **Eylul Bulut:** Performed literature's evaluation, Diseases associated with single-exon genes (enrichment analysis).
- **Ilaria Massignani:** Performed Literature's Evaluation, preliminary analysis of the SinEx dataset, Gene expression Study in TurnOver and Non-TurnOver Tissues, PCA Analysis based tissue expression analysis of SEGs vs MEGs.
- **Sofia Pietrini:** Performed Literature's Evaluation, preliminary analysis of the SinEx dataset, enrichment analysis of the SinEx dataset, preliminary tissue expression analysis of SEGs vs MEGs, searching of data related to diseases listed in (1), differential expression analysis of data retrieved, using of the OneGenE-web-weaver.
- **Marte Toffoli:** Performed Literature's Evaluation, Collected Data from SinEx dataset, Preliminary Analysis of the SinEx dataset, Comparison between SEGs and MEGs in m6A analysis, Correlation with Diseases in m6A analysis.

References

1. Ewa A. Grzybowska. Human intronless genes: Functional groups, associated diseases, evolution, and mrna processing in absence of splicing. *Biochemical and Biophysical Research Communications*, 424(1):1–6, 2012. URL: https://www.sciencedirect.com/science/article/pii/S0006291X12011874, doi:10.1016/j.bbrc.2012.06.092.
2. Carol L. Peebles and Steven Finkbeiner. RNA decay back in play. 10(9):1083–1084. doi:10.1038/nn0907-1083.
3. Yuanyuan An and Hua Duan. The role of m6a RNA methylation in cancer metabolism. 21(1):14. doi:10.1186/s12943-022-01500-4.
4. R Jorquera, C González, P T L C Clausen, B Petersen, and D S Holmes. Sinex db 2.0 update 2020: database for eukaryotic single-exon coding sequences. *Database*, 2021:baab002, 01 2021. arXiv:https://academic.oup.com/database/article-pdf/doi/10.1093/database/baab002/36136900/baab002.pdf, doi:10.1093/database/baab002.
5. Zhuorui Xie, Allison Bailey, Maxim V. Kuleshov, Daniel J. B. Clarke, John E. Evangelista, Sherry L. Jenkins, Alexander Lachmann, Megan L. Wojciechowicz, Eryk Kropiwnicki, Kathleen M. Jagodnik, Minji Jeon, and Avi Ma'ayan. Gene set knowledge discovery with enrichr. *Current Protocols*, 1(3):e90, 2021. URL: https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/cpz1.90, arXiv:https://currentprotocols.onlinelibrary.wiley.com/doi/pdf/10.1002/cpz1.90, doi:10.1002/cpz1.90.
6. John Lonsdale, Jeffrey Thomas, Mike Salvatore, and GTEx Consortium. The genotype-tissue expression (gtex) project. *Nature Genetics*, 45(6):580–585, Jun 2013. doi:10.1038/ng.2653.
7. SS Weinreich, R Mangon, JJ Sikkens, M E en Teeuw, and MC Cornel. [orphanet: a european database for rare diseases]. *Nederlands tijdschrift voor geneeskunde*, 152(9):518—519, March 2008. URL: http://europepmc.org/abstract/MED/18389888.
8. Seth Carbon and Chris Mungall. Gene ontology data archive, January 2024. doi:10.5281/zenodo.10536401.
9. Tanya Barrett, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L. Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva. Ncbi geo: archive for functional genomics data sets—update. *Nucleic Acids Research*, 41(D1):D991–D995, 11 2012. arXiv:https://academic.oup.com/nar/article-pdf/41/D1/D991/3678141/gks1193.pdf, doi:10.1093/nar/gks1193.
10. Francesco Asnicar, Luca Erculiani, Francesca Galante, Caterina Gallo, Luca Masera, Paolo Morettin, Nadir Sella, Stanislau Semeniuta, Thomas Tolio, Giulia Malacarne, Kristof

Engelen, Andrea Argentini, Valter Cavecchia, Claudio Moser, and Enrico Blanzieri. Discovering candidates for gene network expansion by distributed volunteer computing. In *2015 IEEE Trustcom/BigDataSE/ISPA*, volume 3, pages 248–253, 2015. doi:10.1109/Trustcom.2015.640.

11. Ma J Wei Z Wang Y Zhang Y Huang D Song B Meng J Rigden DJ Chen K. Liang Z, Ye H. m6a-atlas v2.0: updated resources for unraveling the n6-methyladenosine (m6a) epitranscriptome among multiple species. *Nucleic Acids Res.*, 2024 Jan 5. doi: doi:10.1093/nar/gkad691.
12. Peter W Harrison, M Ridwan Amode, Olanrewaju Austine-Orimoloye, Andrey G Azov, Matthieu Barba, If Barnes, Arne Becker, Ruth Bennett, Andrew Berry, Jyothish Bhai, Simarpreet Kaur Bhurji, Sanjay Boddu, Paulo R Branco Lins, Lucy Brooks, Shashank Budhanuru Ramaraju, Lahcen I Campbell, Manuel Carbajo Martinez, Mehrnaz Charkhchi, Kapeel Chougule, Alexander Cockburn, Claire Davidson, Nishadi H De Silva, Kamalkumar Dodiya, Sarah Donaldson, Bilal El Houdaigui, Tamara El Naboulsi, Reham Fatima, Carlos Garcia Giron, Thiago Genez, Dionysios Grigoriadis, Gurpreet S Ghattaoraya, Jose Gonzalez Martinez, Tatiana A Gurbich, Matthew Hardy, Zoe Hollis, Thibaut Hourlier, Toby Hunt, Mike Kay, Vinay Kaykala, Tuan Le, Diana Lemos, Disha Lodha, Diego Marques-Coelho, Gareth Maslen, Gabriela Alejandra Merino, Louisse Paola Mirabueno, Aleena Mushtaq, Syed Nakib Hossain, Denye N Ogeh, Manoj Pandian Sakthivel, Anne Parker, Malcolm Perry, Ivana Piližota, Daniel Poppleton, Irina Prosovetskaia, Shriya Raj, José G Pérez-Silva, Ahamed Imran Abdul Salam, Shradha Saraf, Nuno Saraiva-Agostinho, Dan Sheppard, Swati Sinha, Botond Sipos, Vasily Sitnik, William Stark, Emily Steed, Marie-Marthe Suner, Likhitha Surapaneni, Kyösti Sutinen, Francesca Floriana Tricomi, David Urbina-Gómez, Andres Veidenberg, Thomas A Walsh, Doreen Ware, Elizabeth Wass, Natalie L Willhoft, Jamie Allen, Jorge Alvarez-Jarreta, Marc Chakiachvili, Bethany Flint, Stefano Giorgetti, Leanne Haggerty, Garth R Ilsley, Jon Keatley, Jane E Loveland, Benjamin Moore, Jonathan M Mudge, Guy Naamati, John Tate, Stephen J Trevanion, Andrea Winterbottom, Adam Frankish, Sarah E Hunt, Fiona Cunningham, Sarah Dyer, Robert D Finn, Fergal J Martin, and Andrew D Yates. Ensembl 2024. *Nucleic Acids Research*, 52(D1):D891–D899, 11 2023. arXiv:https://academic.oup.com/nar/article-pdf/52/D1/D891/55040594/gkad1049.pdf, doi:10.1093/nar/gkad1049.
13. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
14. Antonio Mucherino, Petraq J. Papajorgji, and Panos M. Pardalos. *k-Nearest Neighbor Classification*, pages 83–106. Springer New York, New York, NY, 2009. doi:10.1007/978-0-387-88615-2_4.
15. Aling Shen, Liya Liu, Yue Huang, Zhiqing Shen, Meizhu Wu, Xiaoping Chen, Xiangyan Wu, Xiaoying Lin, Youqin Chen, Li Li, Ying Cheng, Jianfeng Chu, Thomas J Sferra, Lihui Wei, Qunchuan Zhuang, and Jun Peng. Down-regulating haus6 suppresses cell proliferation by activating the p53/p21 pathway in colorectal cancer. *Frontiers in cell and developmental biology*, 9:772077, 2021. URL: https://europepmc.org/articles/PMC8790508, doi:10.3389/fcell.2021.772077.
16. Miriam Potrony, Tariq Sami Haddad, Gemma Tell-Martí, Pol Gimenez-Xavier, Carlos Leon, Marta Pevida, Judit Mateu, Celia Badenas, Cristina Carrera, Josep Malvehy, Paula Aguilera, Sara Llames, Maria José Escámez, Joan A. Puig-Butillé, Marcela del Río, and Susana Puig. Dna repair and immune response pathways are deregulated in melanocyte-keratinocyte co-cultures derived from the healthy skin of familial melanoma patients. *Frontiers in Medicine*, 8, 2021. URL: https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2021.692341, doi:10.3389/fmed.2021.692341.
17. Peter Truong, Sylvie Shen, Swapna Joshi, Md Imtiazul Islam, Ling Zhong, Mark Raftery, Ali Afrasiabi, Hamid Alinejad-Rokny, Mary Nguyen, Xiaoheng Zou, Sarower Bhuyan, Chowdhury Sarowar, Elaheh Ghodousi, Olivia Stonehouse, Sara Mohamed, Cara Toscan, Patrick Connerty, Purvi Kakadia, Stefan Bohlander, and John Pimanda. Topors e3 ligase mediates resistance to hypomethylating agent cytotoxicity in acute myeloid leukemia cells. *Nature Communications*, 15, 08 2024. doi:10.1038/s41467-024-51646-6.
18. Agnieszka Malcher, Natalia Rozwadowska, Tomasz Stokowy, Tomasz Kolanowski, Piotr Jedrzejczak, Wojmir Zietkowiak, and Maciej Kurpisz. Potential biomarkers of nonobstructive azoospermia identified in microarray gene expression analysis. *Fertility and Sterility*, 100(6):1686–1694.e7, 2013. URL: https://www.sciencedirect.com/science/article/pii/S0015028213027854, doi:10.1016/j.fertnstert.2013.07.1999.
19. Ryo Kimura, Vivek Swarup, Kiyotaka Tomiwa, Michael Gandal, Neelroop Parikshak, Yasuko Funabiki, Masatoshi Nakata, Tomonari Awaya, Takeo Kato, Kei Iida, Shin Okazaki, Kanae Matsushima, Toshihiro Kato, Toshiya Murai, Toshio Heike, Daniel Geschwind, and Masatoshi Hagiwara. Integrative network analysis reveals biological pathways associated with williams syndrome. *Journal of Child Psychology and Psychiatry*, 60, 10 2018. doi:10.1111/jcpp.12999.
20. Takahiro Suezawa, Shuhei Kanagaki, Yohei Korogi, Kazuhisa Nakao, Toyohiro Hirai, Koji Murakami, Masatoshi Hagiwara, and Shimpei Gotoh. Modeling of lung phenotype of hermansky- pudlak syndrome type i using patient-specific ipscs. *Respiratory Research*, 22, 11 2021. doi:10.1186/s12931-021-01877-8.
21. Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Kegg as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 44(D1):D457–D462, 10 2015. arXiv:https://academic.oup.com/nar/article-pdf/44/D1/D457/9482226/gkv1070.pdf, doi:10.1093/nar/gkv1070.
22. Tim E Putman, Kevin Schaper, Nicolas Matentzoglu, Vincent P Rubinetti, Faisal S Alquaddoomi, Corey Cox, J Harry Caufield, Glass Elsarboukh, Sarah Gehrke, Harshad Hegde, Justin T Reese, Ian Braun, Richard M Bruskiewich, Luca Cappelletti, Seth Carbon, Anita R Caron, Lauren E Chan, Christopher G Chute, Katherina G Cortes, Vinícius De Souza, Tommaso Fontana, Nomi L Harris, Emily L Hartley, Eric Hurwitz, Julius O B Jacobsen, Madan Krishnamurthy, Bryan J Laraway, James A McLaughlin,

Julie A McMurry, Sierra A T Moxon, Kathleen R Mullen, Shawn T O'Neil, Kent A Shefchek, Ray Stefancsik, Sabrina Toro, Nicole A Vasilevsky, Ramona L Walls, Patricia L Whetzel, David Osumi-Sutherland, Damian Smedley, Peter N Robinson, Christopher J Mungall, Melissa A Haendel, and Monica C Munoz-Torres. The monarch initiative in 2024: an analytic platform integrating phenotypes, genes and diseases across species. *Nucleic Acids Research*, 52(D1):D938–D949, 11 2023. `arXiv:https://academic.oup.com/nar/article-pdf/52/D1/D938/55040743/gkad1082.pdf`, `doi:10.1093/nar/gkad1082`.

23. Nir Kwon Toth Barbosa Burel Brandis Rossmanith Le Hir Slobodin Schwartz Uzonyi, Dierks. Exclusion of m6a from splice-site proximal regions by the exon junction complex dictates m6a topologies and mrna stability. 2023.
24. Triboulet Robinson Liu Qi Sendinc Erde Gregory Richard I. Yang, Xin. Exon junction complex shapes the m6a epitranscriptome. *Nature Communications*, 2022.
25. Mills Clare Matter Karl Balda Maria S. Zihni, Ceniz. Tight junctions: from simple barriers to multifunctional molecular gates. *Nature Reviews Molecular Cell Biology*, 17(9):564–580, 2016. `doi:10.1038/nrm.2016.80`.
26. Sebastião Ana Maria Simoes de Souza Fabio Marques Antunes, Gabriela. Mechanisms of regulation of olfactory transduction and adaptation in the olfactory cilium. *PLoS One*, 9(8):e105531, 2014. URL: `https://pmc.ncbi.nlm.nih.gov/articles/PMC4140790/`, `doi:10.1371/journal.pone.0105531`.
27. Stepan Gambaryan. The role of no/sgc/cgmp/pkg signaling pathway in regulation of platelet function. *Cells*, 11(22), 2022. URL: `https://www.mdpi.com/2073-4409/11/22/3704`, `doi:10.3390/cells11223704`.
28. Wagner Eric J van Hoof, Ambro. A brief survey of mrna surveillance. *Journal Article*, 36(11):585–92, 2011. URL: `https://pmc.ncbi.nlm.nih.gov/articles/PMC3205232/`, `doi:10.1016/j.tibs.2011.07.005`.
29. Lei Q-Y Zhao S Guan K-L Xiong, Y. Regulation of glycolysis and gluconeogenesis by acetylation of pkm and pepck. *Cold Spring Harb Symp Quant Biol*, 76:285–9, 2011. URL: `https://pmc.ncbi.nlm.nih.gov/articles/PMC4876600/`, `doi:10.1101/sqb.2011.76.010942`.
30. Babiker HM. Chaudhry R, Usama SM. Physiology, coagulation pathways. *StatPearls*, 2013. URL: `https://www.ncbi.nlm.nih.gov/books/NBK482253/`.
31. Masami Watabe-Rudolph, Yvonne Begus-Nahrmann, André Lechel, Harshvardhan Rolyan, Marc-Oliver Scheithauer, Gerhard Rettinger, Dietmar Rudolf Thal, and Karl Lenhard Rudolph. Telomere shortening impairs regeneration of the olfactory epithelium in response to injury but not under homeostatic conditions. 6(11):e27801. `doi:10.1371/journal.pone.0027801`.
32. Francesco Asnicar, Luca Masera, Davide Pistore, Samuel Valentini, Valter Cavecchia, and Enrico Blanzieri. Onegene: Regulatory gene network expansion via distributed volunteer computing on boinc. In *2019 27th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*, pages 315–322, 2019. `doi:10.1109/EMPDP.2019.8671629`.