NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTAMENT OF LIFE SCIENCES

MARIANA DE ALMEIDA CAETANO PERDIGÃO LUZ

BSc in Cell and Molecular Biology

ASSESSING THE PREVALENCE OF TRANSCRIPTION READTHROUGH IN HEALTHY TISSUES THROUGH COMPUTATIONAL APPROACHES

MASTER IN MOLECULAR GENETICS AND BIOMEDICINE

NOVA University Lisbon

September, 2022

DEPARTAMENT OF
LIFE SCIENCES

ASSESSING THE PREVALENCE OF TRANSCRIPTION READTHROUGH IN HEALTHY TISSUES THROUGH COMPUTATIONAL APPROACHES

Mariana De Almeida Caetano Perdigão Luz

BSc in Cell and Molecular Biology

Adviser: Ana Rita Grosso
Assistant Professor, NOVA School of Science and Technology, NOVA University Lisbon

Co-advisers: Paulo Caldas
Junior Researcher, NOVA School of Science and Technology, NOVA University Lisbon

Examination Committee:

Chair: Maria Alexandra Núncio de Carvalho Ramos Fernandes
Assistant Professor with Aggregation, NOVA School of Science and Technology, NOVA University Lisbon

Rapporteurs: Miguel Casanova Vieira Parente
Senior Researcher, Institute of Molecular Medicine, University of Lisbon

Adviser: Ana Rita Grosso
Assistant Professor, NOVA School of Science and Technology, NOVA University Lisbon

MASTER IN MOLECULAR GENETICS AND BIOMEDICINE

NOVA University Lisbon
September, 2022

Assessing The Prevalence Of Transcription Readthrough In Healthy Tissues Through Computational Approaches

ACKNOWLEDGMENTS

I would like to thank Ana Rita Grosso for the unending support and immeasurable kindness. I'll always be grateful to you for accepting me into your lab, without you I wouldn't have been able to finish writing my thesis.

Thank you to Paulo Caldas, not only for your advice but also for the unending patience and good humour with which you helped me throughout this whole process.

Thank you also to the Computational Multi-Omics team for the warm welcome, especially Inês and Cátia Vanessa, without whom lunchtime wouldn't have been nearly as much fun.

Lastly, I would also like to thank my family and friends, especially my little chokos, for supporting me through the all-nighters and not resenting me for the cancelled plans.

ABSTRACT

Transcription termination is a crucial step in the creation of functional RNAs and proteins. In fact, inefficient termination has been shown to lead to the production of aberrantly long transcripts through a phenomenon called transcription readthrough (TRT), in renal cancer cells and cells undergoing several different types of stress (such as hyperosmotic, oxidative and heat shock). Despite the growing research interest in this topic, the prevalence of this phenomenon had yet to be studied in healthy tissues. Therefore, my goal was to assess the prevalence of TRT events in healthy human tissues and characterise them. Using computational approaches to analyse human transcriptome profiles revealed that approximately 19% of expressed genes in healthy human tissues produce readthrough (RT) transcripts. In roughly 2% of these TRT events, the RT transcript was found to not only include the intergenic region but to also overlap downstream genes. This analysis also revealed that, on average, genes that produced RT transcripts (RT genes) were more highly expressed than those that did not (NRT genes). Furthermore, RT genes were found to be significantly enriched for protein-coding genes and a significant overlap between RT genes across all tissues was observed, suggesting that TRT is not a stochastic phenomenon but instead determined by still unknown gene features. This work shows unequivocally that TRT is not only prevalent in cells undergoing stress or in disease but also in healthy human tissues.

Keywords: Readthrough transcription, transcription termination, non-coding RNAs, transcriptomics, bioinformatics

RESUMO

A terminação da transcrição é um passo crucial para a produção de RNA e proteínas funcionais e já foi demonstrado que a terminação ineficiente leva à produção de transcritos aberrantemente longos num fenómeno chamado transcrição aberrante ("readthrough transcription" ou TRT), tanto em células de cancro como em células sob condições diferentes de *stress* (tais como *stress* hiperosmótico, *stress* oxidativo e choque térmico). Apesar do crescente interesse neste tema, a prevalência deste fenómeno não foi ainda estudada em tecidos saudáveis, razão pela qual este trabalho se focou na deteção e caracterização de eventos de TRT. A utilização de abordagens computacionais para analisar perfis de transcritoma de tecidos humanos revelou que aproximadamente 19% dos genes expressos produzem transcritos aberrantes (RT, "readthrough transcripts"). Cerca de 2% destes transcritos incluíam não só a região intergénica, mas também genes a jusante. Em média, os genes que tinham originado transcritos RT eram mais expressos do que aqueles que não produziam transcritos RT, e estavam significativamente enriquecidos para genes codificadores de proteínas. Adicionalmente, os vários tecidos apresentavam uma grande quantidade de transcritos RT em comum, sugerindo que TRT não é um fenómeno aleatório, mas sim determinado por características ainda não conhecidas. Este trabalho mostra inequivocamente que TRT é predominante em tecidos humanos saudáveis.

Palavras-chave: Transcrição aberrante, terminação, RNAs não-codificantes, transcriptómica, bioinformática

CONTENTS

LIST OF FIGURES

ACRONYMS

CPA	Cleavage and polyadenylation complex.
CTD	Carboxyl-terminal domain.
DNA	Deoxyribonucleic acid.
FDR	False discovery rate.
FPKM	Kilobase per million mapped fragments.
GTEx	Genotype-Tissue Expression.
NRT gene	Gene that does not produce readthrough transcripts.
PAS	Polyadenylation site.
RNA	Ribonucleic acid.
RNAP	Ribonucleic acid polymerase.
RNA-seq	High-throughput ribonucleic acid sequencing.
RT gene	Gene produces readthrough transcripts.
RT transcript	Readthrough transcript.
TRT	Transcription readthrough.

1

INTRODUCTION

1.1 Transcription in eukaryotes

Transcription is a fundamental process to the cell that results in the creation of an RNA molecule complementary to the DNA molecule used as a template by the RNA polymerase (RNAP) and involves several regulatory mechanisms throughout its initiation, elongation, and termination steps (reviewed in Cramer, 2019).

In eukaryotic cells, the nuclear transcriptome is produced by three to five different types of RNAP, each responsible for the synthesis of a distinct subset of transcripts (reviewed in Cramer, 2019). RNAP I and III are mostly responsible for the synthesis of ribosomal and transfer RNAs. In particular, RNAP I synthesises preribosomal RNA 35S in *S. cerevisiae*, 45S in plants and 47S in mammals, which mature into the 25/28S, 18S and 5.8S ribosomal RNAs, that form the major components of the ribosome together with 5S ribosomal RNA (reviewed in Russell & Zomerdijk, 2005 and Sáez-Vásquez & Delseny, 2019). RNAP III synthesises 5S ribosomal RNA, transfer RNAs, and other small RNAs found in the nucleus and cytosol (reviewed in White, 2011). RNAP IV and V, both of which exist only in plants, are involved in the regulation of heterochromatin through the synthesis of small interfering RNAs in a process known as RNA-directed DNA methylation (Herr et al., 2005; Kanno et al., 2005; Onodera et al., 2005; Pontier et al., 2005; Wierzbicki et al., 2008).

RNAP II is estimated to be responsible for the transcription of around 85% of the human genome, synthesising precursor messenger RNAs (pre-mRNAs), along with most small nuclear RNAs (snRNAs) and microRNAs (reviewed in Sims et al., 2004). This multiprotein complex is characterised by the carboxyl-terminal domain (CTD) of RBP1, its largest subunit, which consists of tandem repeats of the consensus heptapeptide YSPTSPS (Tyr1-Ser2-Pro3-Thr4-Ser5-Pro6-Ser7), with the number of copies varying from 26 in yeast to 52 in mammals. The post-translational modification patterns of these amino acids have been shown to play an important role during the transcription process, such as phosphorylation of Ser2 and Ser5 coordinating the recruitment of factors during elongation and

phosphorylation of Tyr1 inhibiting termination by preventing termination factors from binding to the CTD (reviewed in Chapman et al., 2008; Eaton & West, 2020; Eick & Geyer, 2013; Lyons et al., 2020).

1.2 Termination of transcription by RNAP II

Although RNAP II is the most studied of the eukaryotic RNA polymerases, there is still much left to uncover, especially when it comes to the way it terminates transcription (reviewed in Eaton & West, 2020 and Proudfoot, 2016). Historically, most RNAP II research has focused on the initiation and elongation steps, with termination only becoming a bigger research interest in more recent years, rightfully so, as this stage is as important to the creation of functional RNAs and proteins as the initiation and elongation steps.

For most protein-coding genes, transcription termination is triggered when RNAP II encounters a polyadenylation site (PAS), comprised of the consensus sequence AAUAAA along with upstream U-rich and downstream U/GU-rich sequences, which induces the activation of the cleavage and polyadenylation complex (CPA) associated with its CTD domain, leading to pre-mRNA cleavage followed by polyadenylation. Two main theories explain this PAS-dependent termination, referred to as the allosteric/anti-terminator and torpedo models, but more recent experiments seem to point to the real termination mechanism unifying both (Eaton et al., 2020). According to this unified model, the slowing down of RNAP II after the PAS is caused by dephosphorylation of the elongation factor SPT5 by protein phosphatase 1 (PP1), and the phosphorylation of Thr4, which has been theorised to signify imminent termination of transcription. The stranded RNAP II is then easily terminated by the 5'-3' exoribonuclease XRN2, which induces dissociation from the DNA template through degradation of the unprotected 5' end of the RNA.

PAS-dependent termination is not, however, the only RNAP II termination process. Replication-dependent histone transcripts are not cleaved after a PAS is recognised, but instead when U7 snRNA, as part of a histone cleavage complex, recruits cleavage-polyadenylation specificity factor 73 (CPSF73), the same endonuclease that is responsible for cleavage in PAS-dependent termination. In the human U1 and U2 snRNA transcripts, termination occurs when RNAP II encounters a 3' box, which is a 13-16 nucleotides element located 9-19 nucleotides downstream of the 3' end, and relies on the Integrator complex, which is not conserved in lower eukaryotes and contains a paralog of CPSF73 called Integrator complex subunit 11 (INTS11) (Hernandez, 1985; O'Reilly et al., 2014).

Evidence also suggests that RNAP II is capable of terminating transcription at all points in the transcription cycle in a process deemed premature termination, including not only within the gene body (intragenic premature termination) but also in the vicinity of the transcription start site (TSS-linked premature termination) (reviewed in Kamieniarz-Gdula & Proudfoot, 2019). Premature termination is a

form of negative gene expression regulation, especially of transcriptional regulator genes such as *CSTF3*, which codes for a CPA subunit that stimulates 3' end cleavage and has been shown to undergo premature termination induced by high levels of the CSTF3 protein (Luo et al., 2013). While most events of premature termination lead to unstable transcripts that undergo rapid degradation, some generate stable transcripts that can have independent functions and are considered instances of alternative polyadenylation (Berkovits & Mayr, 2015).

The interaction of CPA components with RNAP II as it transcribes a PAS can also have a pausing effect on transcription, leading to a gradual release of RNAP II that is seemingly independent of PAS cleavage and only caused by RNAP II conformational changes (Zhang et al., 2015). Pausing of RNAP II can also be caused by chromatin structure, such as core nucleosomes or R-loops, which are structures caused by displacement of the sense DNA strand induced by the formation of RNA:DNA hybrids (reviewed in Skourti-Stathaki & Proudfoot, 2014). R-loops are favoured by defective splicing as well as the act of transcription due to the transient displacement of nucleosomes by RNAP II and are most often associated with G-rich regions (Reabansg et al., 1994). Besides leading to RNAP II pausing, R-loops have also been shown to cause low-level antisense transcription, which may result in the creation of localized patches of repressed chromatin, leading to further RNAP II pausing (Skourti-Stathaki et al., 2014).

However, it is not only the premature termination of transcription that is of significance, but also its inefficient termination, which leads to longer transcripts rather than shorter. Interestingly, this phenomenon might be the cause of up to 80% of cases of intergenic transcription (Agostini et al., 2021).

1.3 Readthrough transcription

Readthrough transcription (graphically depicted in Figure 1.1) is a form of abnormal transcription, in which inefficient 3' end cleavage of the nascent transcript leads to RNAP II continuing transcription uninterrupted past the 3' end of genes (recently reviewed in Morgan et al., 2022 and Rosa-Mercado & Steitz, 2022). This can lead to the production of novel transcripts, changes in epigenetic states, and altered 3D genome structure (Cardiello et al., 2018; Grosso et al., 2015; Heinz et al., 2018; Hennig et al., 2018; Rutkowski et al., 2015).

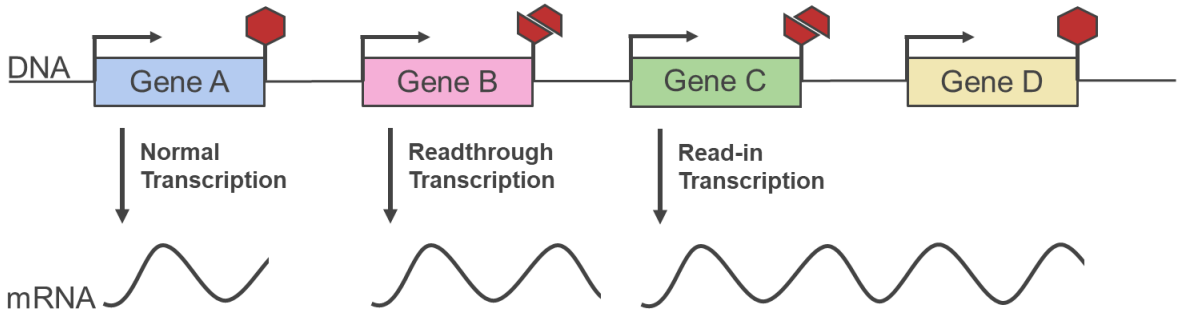

Figure 1.1. Graphical depiction of transcription readthrough (TRT). Readthrough transcription is a type of abnormal transcription, in which the inefficient 3' end cleavage of the nascent transcript leads to the creation of aberrant transcripts, referred to as readthrough transcripts. The downstream genes that are transcribed due to this inefficient transcription termination are referred to as read-in genes. In cases where this happens, readthrough transcription can also be called read-in transcription.

These novel transcripts, classified as long non-coding RNAs (lncRNAs), have been referred to as readthrough (RT) transcripts or as downstream-of-gene-containing transcripts in the literature. In this work, they are simply referred to as RT transcripts. Genes that produce RT transcripts are referred to as RT genes, while those that do not are referred to as NRT genes.

RT transcripts first came to notice when, in 2015, three different groups independently published their observations of widespread transcription readthrough (TRT) in cells under a variety of stress conditions (Grosso et al., 2015; Rutkowski et al., 2015; Vilborg et al., 2015). Since then, RT transcripts have become a hallmark of termination defects and the mammalian cellular stress response, having been detected in cells undergoing hyperosmotic stress (Vilborg et al., 2015), heat shock (Cardiello et al., 2018; Vilborg et al., 2017), oxidative stress (Vilborg et al., 2017) and hypoxia (Wiesel et al., 2018), as well as in clear cell renal cell carcinoma (ccRCC) cells (Grosso et al., 2015) and cells infected by the influenza and herpes simplex viruses (Bauer et al., 2018; Rutkowski et al., 2015).

1.4 Biogenesis of RT transcripts

RT transcripts remain chromatin-bound and have previously been found to reach lengths of up to 200 kb (Vilborg et al., 2015). In some cases, readthrough transcription continues beyond the intergenic region and overlaps downstream genes, which are referred to as read-in genes (Rutkowski et al., 2015).

It has been shown that RT transcripts are produced by approximately 10% of human protein-coding genes in a cell (Vilborg et al., 2015). When a given mammalian cell line is exposed to several different stress conditions, the RT transcripts seem to show significant overlap (Hennig et al., 2018;

Vilborg et al., 2017). It is unclear if the same is true across different cell types. The fact that these transcripts arise within minutes of exposure to the stress factor suggests that they are produced from genes already being actively transcribed. In line with this, whether a gene is activated or repressed once the stress response is triggered does not seem to affect RT transcript production, with both types of RT genes having been observed (Rosa-Mercado et al., 2021; Vilborg et al., 2015). However, active transcription or high gene expression levels alone do not seem to be enough to generate RT transcripts, as very highly expressed genes have been observed to not produce RT transcripts (Vilborg et al., 2015).

It has been found that, in the context of herpes simplex virus 1 infection, RT genes tend to have a weaker PAS, as well as a depletion of canonical PAS and GU-rich sequences in the transcript body, comparatively to NRT genes (Rutkowski et al., 2015). This evidence suggests that RT genes might be predisposed to produce RT transcripts.

A redistribution in RNAP II occupancy from gene bodies into intergenic regions was also noted in cells undergoing hyperosmotic stress, which is a possible explanation for the widespread transcriptional repression observed in cells afterwards (Amat et al., 2019; Rosa-Mercado et al., 2021). This redistribution was also noticed in the context of viral infection and heat shock (Bauer et al., 2018; Mahat et al., 2016). Taken together, these results suggest that RT transcript production might displace RNAP II to downstream regions.

The downstream regions of RT genes have also been shown to be slightly enriched, compared to that of NRT genes, for demethylated histone H3 lysine 79 and trimethylated histone H3 lysine 36, marks known to favour elongation (Godfrey et al., 2020; Tomson & Arndt, 2013), in both mouse myoblast cells and human fibroblasts (Hennig et al., 2018; Vilborg et al., 2017). However, in ccRCC, the inactivation of *SETD2*, which codes for a histone methyltransferase specific for lysine 36 on histone H3, has been connected to higher TRT levels and the expression of wild-type *SETD2* has been found to be sufficient to revert this effect. Further study is necessary to understand how TRT is impacted by these epigenetic modifications.

The splicing status of RT transcripts has not yet been clarified. However, studies have shown that unspliced pre-mRNAs tend to not undergo 3' end cleavage in both yeast and mammalian cells (Alpert et al., 2020; Herzel et al., 2018; Soucek et al., 2016), which has led to the theory that this might be the case for RT transcripts as well. Experiments have also shown, in differentiating mouse erythroblasts, that RT transcripts are produced by genes where nascent transcripts experience inefficient co-transcriptional splicing and that introducing a cryptic splice site increases 3' end cleavage efficiency (Alpert et al., 2020; Castillo-Guzman et al., 2020). Increased TRT levels were also observed when PladB, a splicing inhibitor, was present (Castillo-Guzman et al., 2020). Furthermore, splicing disruptions have been observed in several of the stress conditions in which TRT has been studied, including heat shock and infection by the herpes simplex virus 1 (Rutkowski et al., 2015; Shalgi et al.,

2014). Together, these findings suggest that disrupted splicing patterns can interfere with transcription termination by reducing the levels of 3' end cleavage and, therefore, might be one of the causes of TRT.

Other possible players in the production of RT transcripts are the CPA and the Integrator complex. The CPA machinery has been implicated as the depletion of CPSF73 has been associated with the production of RT transcripts (Cardiello et al., 2018). Knockdown of the Integrator complex subunit INTS11, a paralog of CPSF73, has been shown to have a similar effect (Rosa-Mercado et al., 2021).

1.5 RT transcript functions

Evidence suggests that some ncRNAs are responsible for supporting chromatin structure, among which were identified spliced transcripts with long 3' untranslated regions (UTRs) (Caudron-Herger et al., 2011; Hall et al., 2014). As RT transcripts are also ncRNAs with long 3' UTRs which are retained in the nucleus, it has been suggested that they might be a part of this group of ncRNAs that play a role in reinforcing the nuclear scaffold during the cellular stress response, though further proof is necessary (Vilborg & Steitz, 2016).

TRT events in ccRCC have also been found to lead to the production of chimeric transcripts, though it is not yet known how they are involved in tumorigenesis (Grosso et al., 2015).

It has also been speculated that RT transcripts might play a role in altering the expression of other genes through transcriptional interference, namely that of read-in genes (Mazo et al., 2007; Rosa-Mercado et al., 2021; Rutkowski et al., 2015). The production of RT transcripts as antisense transcripts might also affect the expression of genes of the opposite strand that suffer overlap (Muniz et al., 2017; Vilborg et al., 2015). In this case, RT transcripts might be a contributor to the state of transcriptional repression verified in cells under stress. This is further supported by gene ontology analyses of RT genes revealing enrichment for functions associated with transcriptional repression, which are not enriched in NRT genes (Rosa-Mercado et al., 2021).

1.6 Transcriptome profiling of healthy tissues

The study of aberrant transcription events such as TRT, and of transcription in general, has been made much easier in the last decades thanks to the rapid evolution of sequencing technologies. High-throughput RNA sequencing (RNA-seq) has become the dominant method in recent years, as it not only allows for the quantification of gene expression but also the identification of alternatively spliced genes, allele-specific expression and novel transcripts. RNA-seq consists of RNA isolation, reverse transcription to complementary DNA, preparation of a sequencing library and then sequencing on a

next-generation sequencing platform (Marioni et al., 2008; Wang et al., 2009). Reads are then either aligned to a reference genome or assembled *de novo*, which is especially useful for non-model organisms that have not yet been sequenced. Information on strand specificity is usually lost during the process unless specific measures are taken to define transcript direction.

Analysing the transcriptome has been made even more accessible by the fact that thousands of studies have chosen to make their RNA-seq data available for reuse by anyone in the scientific community by adding them to public repositories such as the Gene Expression Omnibus (Barrett et al., 2013), The Cancer Genome Atlas and the Genotype-Tissue Expression (GTEx) Project. It is this practice that has enabled this work, as well as many others, to be carried out, allowing for more analysis to be performed on the same datasets and more knowledge to be gained.

1.7 Main objective of the work

Despite the growing research interest in this topic, the prevalence of TRT in healthy mammalian tissues has yet to be explored. Hence, the main objective of this work is to uncover and assess the prevalence of transcriptional readthrough across human healthy tissues using publicly available transcriptome profiles and computational approaches.

To this aim, this work was divided into three main tasks:

- Asses the prevalence of RT transcripts in healthy human tissues using open-source tools and publicly available RNA-seq data from the GTEx project.
- Characterise features of RT and NRT genes and find associations with the likelihood of producing RT transcripts.
- Characterise flanking regions at the 3' end of RT and NRT genes and find associations with the likelihood of producing RT transcripts.

The output of this work will potentially provide the first steps to identify the mechanisms underlying the production of readthrough transcripts and ultimately shed light on their functional role and impact on human health.

2

METHODS

2.1 RNA-seq data from human samples

To explore the prevalence of TRT in healthy human individuals, a large collection of publicly available transcriptomes was analysed. These were taken from the GTEx Project, which is supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. Specifically, the RNA-seq data analysed in this work was obtained from the NCBI database of Genotypes and Phenotypes (dbGaP) study accession number phs000424.v8.p2 on 01/04/2021 by Ana Rita Grosso.

The GTEx Portal has collected data for 54 different tissues from 948 different donors with ages ranging between 20 and 70, totalling 17382 samples. It has also collected relevant clinical information for each of these individuals, which was anonymised in accordance with the NIH Genomic Data Sharing (GDS) Policies in order to protect their privacy. This information was used to filter out confounder effects.

As the goal of this work was to analyse healthy tissues, only samples from individuals with no terminal diseases and whose causes of death were classified as fast/violent according to the 4-point Hardy Scale were included. This guarantees that individuals with diagnosed illnesses that might affect TRT levels (e.g., cancer and dementia) are excluded. To avoid technical bias, only paired-end samples with similar sequencing protocols (using the Trizol method for RNA isolation) and a minimum threshold of sequencing coverage (60 million reads per sample) were considered. Moreover, to ensure the data was representative of the overall population cell culture samples and tissues enclosing less than 50 samples were excluded. After these filters were applied, the data left encompassed 23 different tissues from 236 different individuals with ages ranging from 20 to 70, totalling 2778 samples.

2.2 TRT detection

First, the BAM files obtained from the GTEx Portal were converted back to FASTQ using SAMtools v1.10 (Danecek et al., 2021). The transcriptomes were realigned to the reference genome (GRCh38, version 37, Ensembl 103), which was the most recent version available at the time (Cunningham et al., 2022), using the Spliced Transcripts Alignment to a Reference (STAR) software v2.7.8a (Dobin et al., 2013).

ARTDeco (Roth et al., 2020) was then used to identify and characterise TRT. This tool expands on previous methods by implementing three separate strategies: evaluating readthrough level, evaluating read-in level, and detecting novel transcripts. Only the latter was relevant to this work, as it allowed for downstream analysis of expression levels of genes and their readthrough regions. ARTDeco employs a rolling window approach to search for continuous coverage over a defined minimal length downstream of the 3' end of each gene. In this case, a rolling window of 500 bp, a minimum tail length of 2000 bp and a minimum coverage of 0.15 fragments per kilobase per million mapped fragments (FPKM) were used as thresholds. The realignment and ARTDeco pipeline were both run by Paulo Caldas.

Due to the lack of strand specificity, a significant amount of reads coming from genes being expressed in the opposite strand were mistakenly classified as RT transcripts. It was possible to identify the existence of these mislabelled transcripts through an initial visual inspection of transcriptomic profiles on GenomeBrowser (Kent et al., 2002) and then further analysis of transcriptomic profiles created using pyGenomeTracks v3.7 (Lopez-Delisle et al., 2021).

To discard these false positives, the list of RT transcripts identified by ARTDeco was filtered to discard those overlapping with genes in the opposite strand. This was done using the intersect function from BEDtools v2.30.0 (Quinlan & Hall, 2010), which screens for overlaps between two sets of genomic features. It is likely that this approach also removes a small number of true RT genes with close downstream neighbours in the opposite strand. However, it also ensures that the list of RT transcripts is robust, which is imperative when the aim is to characterise these transcripts and the genes that produce them. From the remaining list of RT transcripts, only those produced by expressed genes (defined as having a minimum FPKM of 1 in at least 25% of samples from a given tissue) were used for downstream analysis. Only the longest isoform detected for each gene was considered to ensure reads from the gene were not mistaken for reads from the downstream intergenic region.

2.3 Gene set enrichment analysis

Gene set enrichment analysis was performed on the lists of RT genes for each tissue and the list of RT genes in common in all tissues using the ToppFun tool from the ToppGene Suite (Chen et al., 2009), based on gene ontologies, phenotype and literature co-citation. For the lists of RT genes for each tissue, the list of expressed genes of the respective tissue was used as background. For the list of RT genes in common in all tissues, the list of all expressed genes in common in all tissues was used as background. Results were considered relevant when, after false discovery rate (FDR) adjustment (Benjamini & Hochberg, 1995) was applied, the FDR-adjusted p-value remained inferior to 0.05.

2.4 Data analysis and integration

Data analysis was carried out using Python v3.7.13 (Hunt, 2019) (refer to *Appendix* for the script). The data analysed was read using pandas v1.3.5 (McKinney, 2010; Reback et al., 2021) and kept as pandas DataFrame objects, which have integrated indexing and are easily manipulated. The module NumPy v1.21.5 (Harris et al., 2020) was used to convert data for easier analysis. The seaborn v0.11.2 (Waskom, 2021) and Matplotlib v3.5.1 (Caswell et al., 2021; Hunter, 2007) modules were used for plotting.

The significance of differences in expression levels between groups of genes was evaluated by the Mann-Whitney U Test (Mann & Whitney, 1947) using the SciPy v1.7.3 module (Virtanen et al., 2020). This test was also used to evaluate the difference between the distribution of intergenic lengths and gene classifications. All of the resulting p-values were corrected by the FDR method using the statsmodels v0.13.2 module (Seabold & Perktold, 2010), with differences being considered significant if the FDR-adjusted p-value was inferior to 0.05.

Correlations between the expression levels of RT genes and RT transcripts, between the length of the transcript tail and the expression levels of RT genes, and between the transcript tail length and distance to the closest downstream gene were all measured by the Pearson correlation coefficient using the SciPy module.

The SciPy module was also used to determine if there was an enrichment of protein-coding genes in the list of RT genes, employing Fisher's exact test (Fisher, 1922) for that purpose. Once again, p-values were corrected by the FDR method using the statsmodels module, being considered significant when inferior to 0.05.

Furthermore, the BEDtools intersect tool was also used to help determine read-in levels by searching for overlap between RT transcripts and genes, while the closest tool was used to evaluate the distance between a given RT transcript or genes and its respective closest downstream gene.

3

RESULTS AND DISCUSSION

3.1 TRT is pervasive in healthy tissues

To explore the occurrence of TRT in healthy human tissues, a large collection of transcriptomes made publicly available by the GTEx Project was analysed (see section *2.1. RNA-seq data obtained from human samples* for filtering criteria applied). Only paired-end samples with a minimum of 60 million reads from individuals with fast/violent causes of death and no terminal diseases were considered. To ensure representativeness, tissues were only studied if at least 50 samples met these requirements. In total, 2778 samples across 23 different tissues were studied (Figure 3.1): subcutaneous adipose tissue (ADPSUB), visceral adipose tissue (ADPVIS), aorta artery (ARTAOR), coronary arteries (ARTCOR), tibial arteries (ARTTIB), cerebellum (BRNCER), cerebral cortex (BRNCTX), breast tissue (BREAST), sigmoid colon (COLON), gastroesophageal junction (ESOGAST), oesophagus mucous membrane (ESOMUC), oesophagus muscle (ESOMUS), atriums (HEARTAP), left ventricle (HEARTLV), liver (LIVER), lung (LUNG), skeletal muscle (MUSSKL), tibial nerve (NERVE), pituitary gland (PITTY), skin exposed to the sun (SKINLL), skin not exposed to the sun (SKINSP), testicles (TESTIS) and thyroid (THYRD).

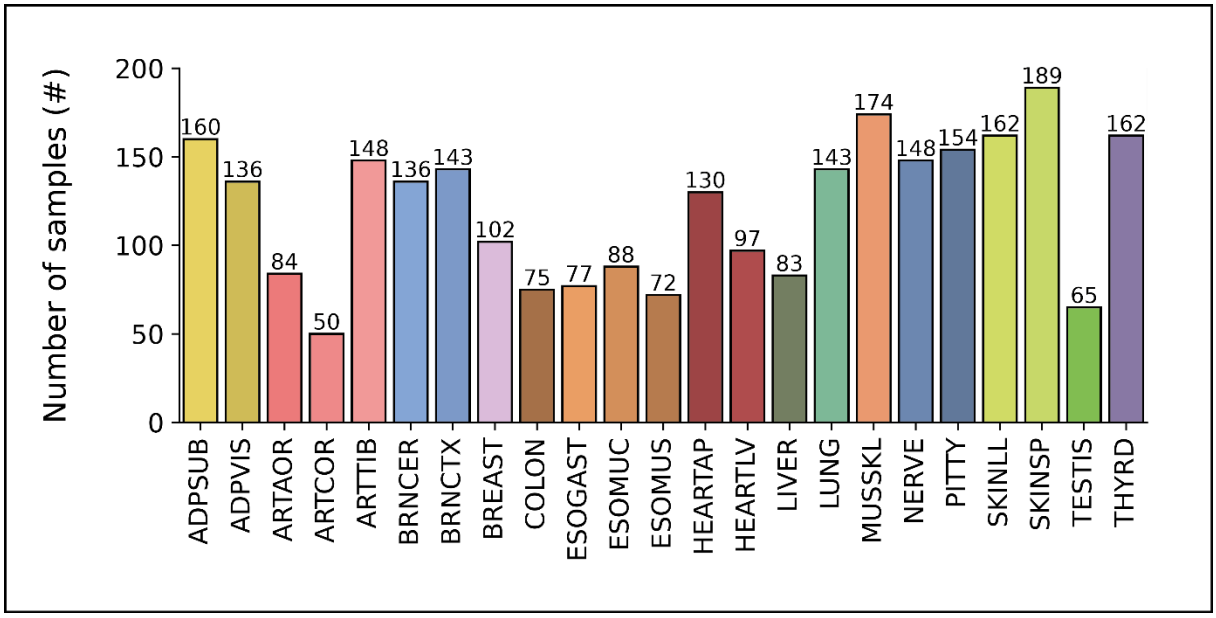

Figure 3.1. Number of samples analysed for each tissue. Barplot showing the number of samples assessed for the 23 tissues: ADPSUB – subcutaneous adipose tissue, ADPVIS – visceral adipose tissue, ARTAOR – aorta artery, ARTCOR – coronary arteries, ARTTIB – tibial arteries, BRNCER – cerebellum, BRNCTX – cerebral cortex, BREAST – breast tissue, COLON – sigmoid colon, ESOGAST – gastroesophageal junction, ESOMUC – oesophagus mucous membrane, ESOMUS – oesophagus muscle, HEARTAP – atriums, HEARTLV – left ventricle, LIVER – liver, LUNG – lung, MUSSKL – skeletal muscle, NERVE – tibial nerve, PITTY – pituitary gland, SKINLL – skin exposed to the sun (lower legs), SKINSP – skin not exposed to the sun (suprapubic), TESTIS – testicles, THYRD – thyroid.

ARTDeco (Roth et al., 2020) was used to search for RT transcripts in these samples. As the original data obtained from the GTEx Project lacks strand specificity, a significant number of genes being expressed in the opposite strand were mistakenly labelled by ARTDeco as RT transcripts produced by other genes (examples shown in Figure 3.2.A). To ensure no false positives were included in downstream analyses, putative RT transcripts which overlapped genes in the opposite strand were removed from consideration (section *2.2. Transcription readthrough detection* for further details). This filter did remove a significant number of expressed genes (approximately 28%) but it also ensured that the subsequent analyses were only performed on robust, unambiguous RT transcripts (examples in Figure 3.2.B).

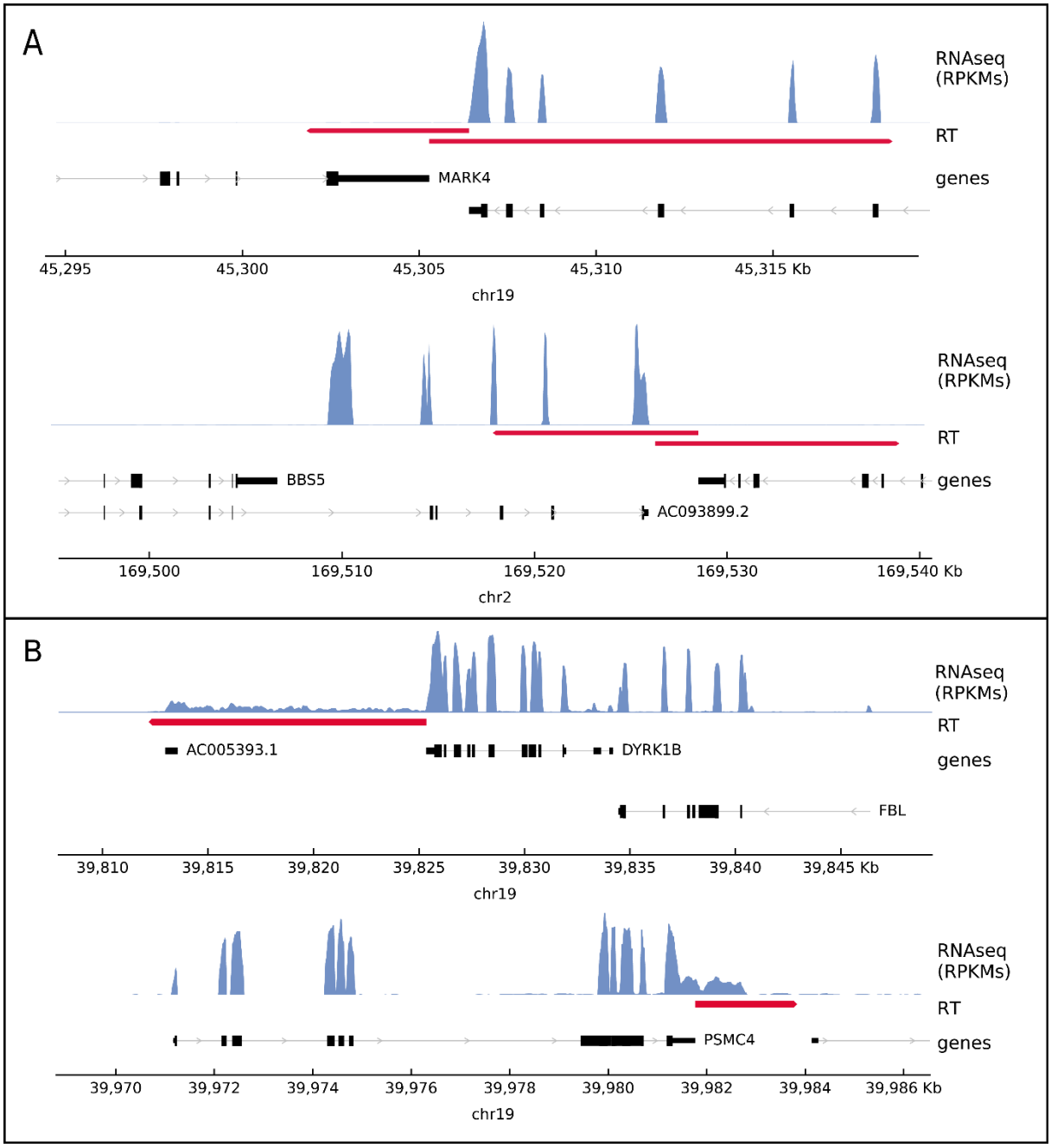

Figure 3.2. Genome browser tracks for example genes. A) Transcriptomic profiles from genes expressed in the skeletal muscle tissue, which show reads of a gene in the opposite strand being mislabelled as RT transcripts. **B)** Transcriptomic profiles from genes expressed in the skeletal muscle tissue, showing clear RT transcripts.

The resulting list of genes was further filtered to ensure that only relevant genes to the tissue being studied were analysed (section *2.4. Data analysis* for details). For this purpose, only genes which had a minimum FPKM of 1 in at least 25% of the samples in a given tissue were considered expressed. The total number of remaining expressed genes (Figure 3.3.A) varied from 6954 (skeletal muscle) up to 13084 (testicles).

A significant increase in read counts in the intergenic region immediately downstream of several expressed genes was observed, revealing the production of RT transcripts. The number of RT genes varied across tissues (Figure 3.3.A) and among samples (Figure 3.3.B), making the total number of RT genes detected higher than the average number per tissue. The number of genes that produced RT transcripts in each tissue varied from 932 (left ventricle) up to 3082 (testicles) (refer to *Appendix* for the full list of expressed genes and RT transcripts). This corresponded to approximately 19% of expressed genes in each tissue producing RT transcripts (Figure 3.3.A), with the percentage varying in specific tissues from roughly 13% (left ventricle) up to 27% (cerebellum).

The percentage of genes that were found to produce RT transcripts in this work is not completely out of line with the 10% of protein-coding genes that had been estimated to produce RT transcripts in neuroblastoma cells undergoing stress (Vilborg et al., 2015). The lower estimates obtained previously are easily explained by the fact that a higher minimum tail length was considered (5 kb instead of the 2 kb used in this study), the data was obtained from a different type of culture (neuroblastoma cell line instead of healthy tissue samples from various donors) and only protein-coding genes were considered for the analysis instead of the whole genome. Additionally, due to the high number of dubious cases that had to be filtered out, the absolute number of RT genes is probably underestimated in this work. In the future, it would be important to obtain strand-specific transcriptome profiles to confirm how extensive TRT is.

These results show that TRT is not a phenomenon exclusive to the mammalian cellular stress response and it is, in fact, pervasive in healthy human tissues.

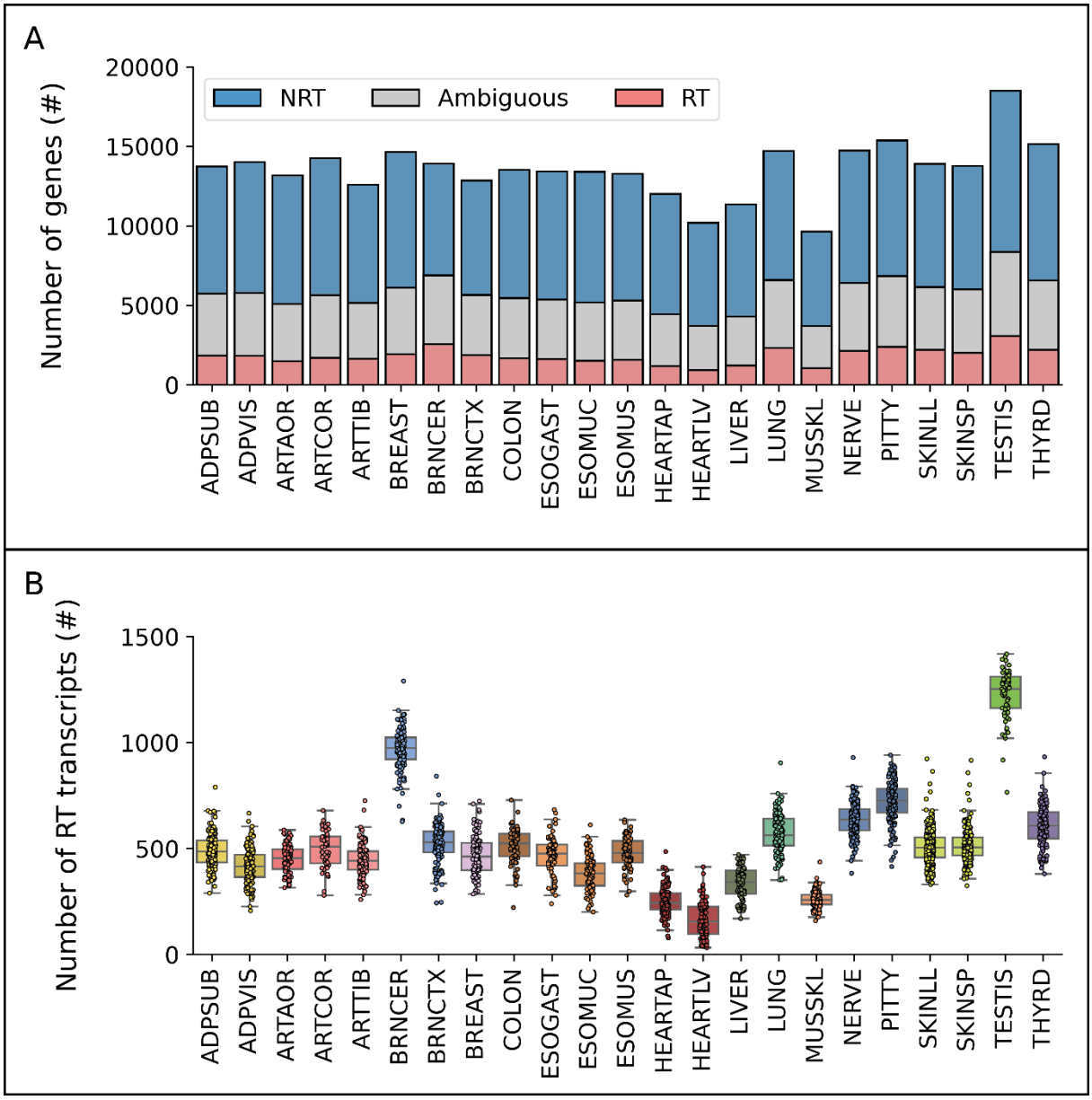

Figure 3.3. Transcription readthrough (TRT) is pervasive in healthy human tissues. A) Barplot showing all expressed genes classified in: RT genes, NRT genes and ambiguous genes (filtered out and not considered for further analysis). **B)** Boxplot showing the number of RT genes detected for each sample and tissue.

3.2 Characterisation of RT genes

Next, the differences in the average expression levels between genes that do and do not produce RT transcripts were analysed (Figure 3.4.A). The average expression of RT genes was found to be slightly higher than NRT genes (Mann-Whitney U Test FDR-adjusted p-value < 0.05 for all tissues except for the pituitary gland and oesophagus muscle). Moreover, the expression levels of RT transcripts

showed a moderate correlation with the expression of the respective gene (Pearson's r = 0.437, p-value < 0.001) (Figure 3.4.B).

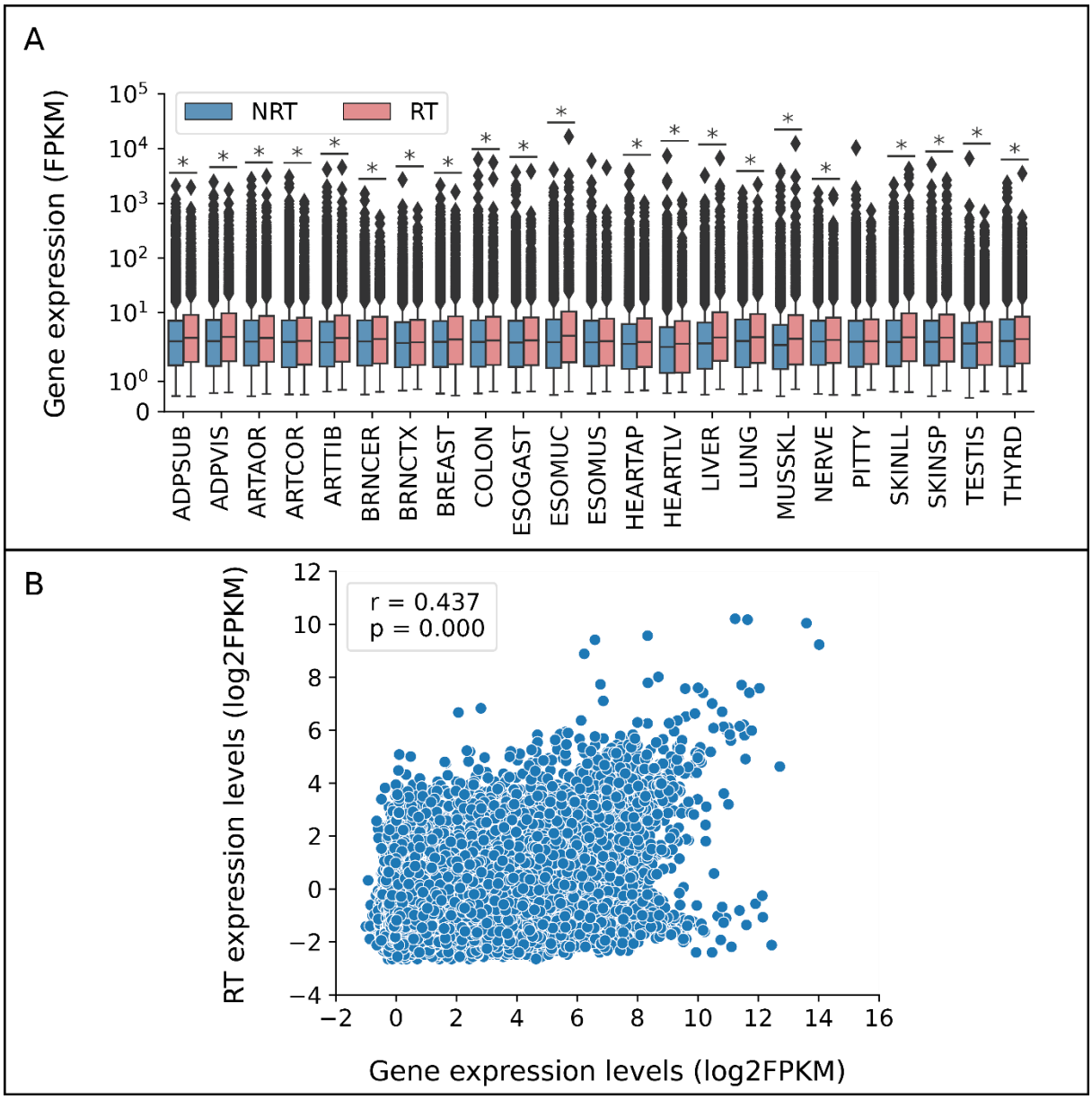

Figure 3.4. RT transcript production is associated with higher gene expression. A) Boxplot showing expression levels of RT and NRT genes for every tissue. **B)** Scatter plot comparing RT transcript expression levels (log2 FPKM) and gene expression levels (log2 FPKM). *****Mann-Whitney U Test FDR-adjusted p-value < 0.05.

These results suggest that RT genes are slightly more expressed than NRT genes, as previously suggested in the literature (Rosa-Mercado et al., 2021; Vilborg et al., 2015). It is possible that RT transcripts are more often produced by highly expressed genes simply because the likelihood of an error at the 3' end increases with the transcription frequency. It should also be taken into account that the

higher coverage of more highly expressed genes might allow the detection of RT transcripts on such genes.

Though it is clear from these results that the production of RT transcripts is associated with higher expression levels, it is very likely that other factors besides gene expression also play a role in determining RT transcript production. Otherwise, the difference between the expression levels of NRT and RT genes and the correlation between the expression levels of the gene and the respective transcript tail would be much more pronounced. This is in line with previous observations that, in cells undergoing stress, RT transcripts are produced by highly expressed genes, but high expression levels are not sufficient for RT transcript production (Vilborg et al., 2015).

To test this possibility, different features of genes that could explain the occurrence of TRT were explored. Analysis of the gene classifications of RT genes revealed that 85% of RT genes were protein-coding (Figure 3.5.A). Considering all expressed genes as background (Figure 3.5.B), this enrichment was deemed significant when Fisher's Exact Test was employed (FDR-adjusted p-value < 0.05).

As significant overlap of RT genes had previously been reported within the same mammalian cell line exposed to different stress conditions (Hennig et al., 2018; Vilborg et al., 2017), this possibility was also explored here. In accordance with the literature, it was observed that 229 genes produced RT transcripts in all 23 tissues, while only 2587 RT genes were exclusive to a single tissue (Figure 3.5.C).

Together, these results support the theory that TRT is not an entirely stochastic phenomenon and certain gene features may play a role in determining which genes produce RT transcripts.

Considering this, gene set enrichment analyses were performed to assess if RT genes were associated with specific biological pathways or functions. For this purpose, the ToppFun tool (Chen et al., 2009) was used to analyse the RT genes in each tissue and the RT genes in common across all tissues based on gene ontologies, phenotype and literature co-citation. A list of expressed genes for each tissue and a list of expressed genes in common across all tissues were used as the background, respectively. However, no enrichment was found for any molecular function, biological process, cellular component, pathway or human phenotype for any of the gene sets. This is contradictory to the enrichment for functions associated with transcriptional repression that had been previously found in cells undergoing hyperosmotic stress (Rosa-Mercado et al., 2021), though it is possible that this difference is simply caused by the fact that the data analysed here comes from healthy tissues.

These results suggest that gene features other than specific functions might play a role, such as a weaker PAS, inefficient co-transcriptional splicing, chromatin structure or epigenomic changes, all of which have been previously observed in cases of TRT (Alpert et al., 2020; Castillo-Guzman et al., 2020; Pan & Frank Huang, 2022; Rutkowski et al., 2015; Vilborg et al., 2015). It would be interesting to check

in the future if these and possible other features do indeed play a role in the production of RT transcripts in healthy tissues.

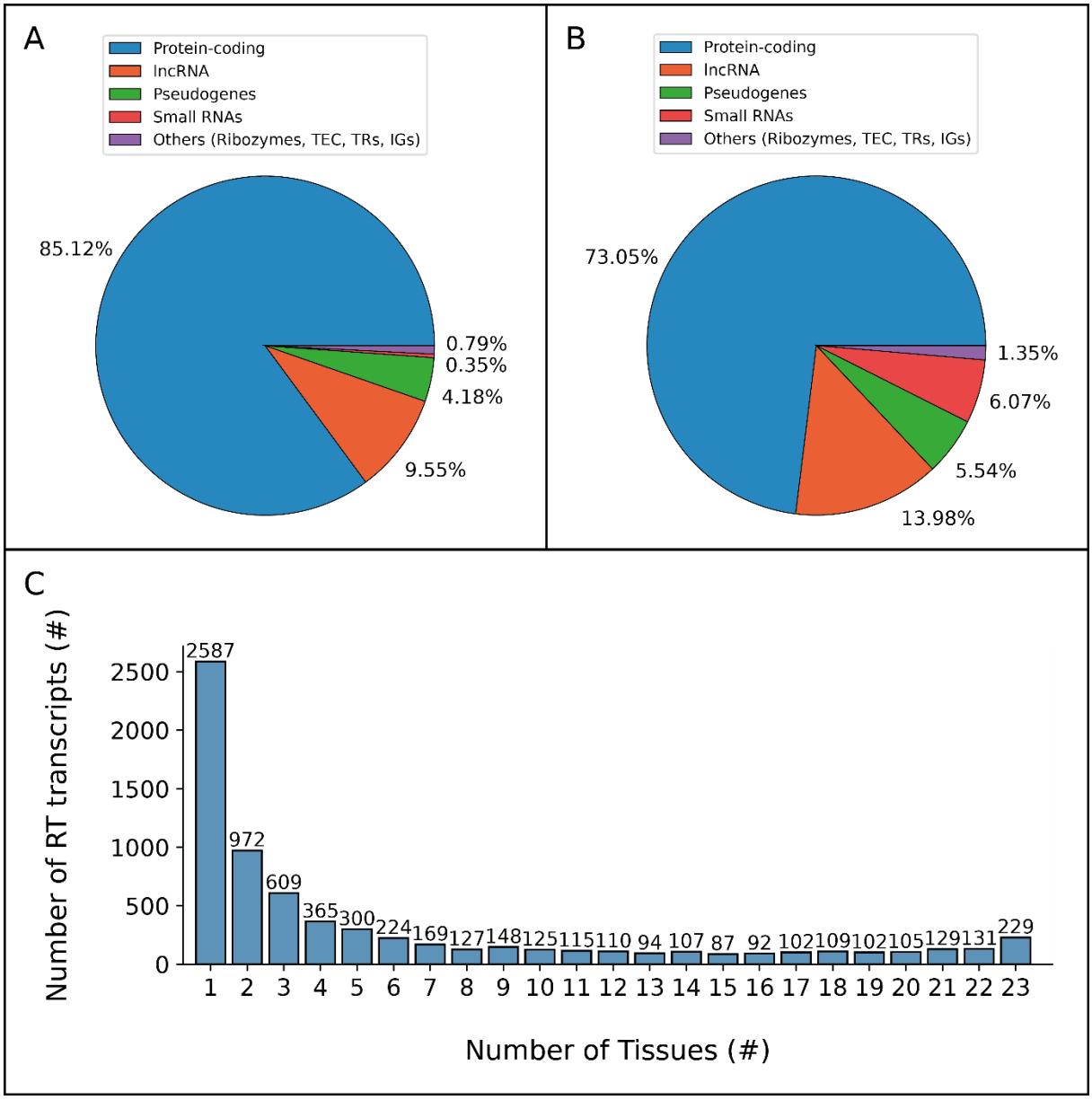

Figure 3.5. TRT is not a stochastic phenomenon. A) Pie plot showing the proportion of gene classifications for RT genes. **B)** Pie plot showing the proportion of gene classifications for all expressed genes. **C)** Barplot showing the number of tissues that each RT gene is expressed in.

3.3 Characterisation of transcript tails

As mentioned before, in stress conditions, RT transcripts have been found to overlap downstream genes (Grosso et al., 2015; Rutkowski et al., 2015). In this study, this phenomenon of read-in transcription was also observed in healthy human tissues in roughly 2% of TRT events. TRT events might simply be less common in healthy tissues compared to cells undergoing stress, but there is also a possibility that this low number is, in part, caused by the filtering of ambiguous genes, as all transcripts which overlapped genes in the opposite strand were removed from consideration. Still, even with such a low number of transcripts overlapping downstream genes, it was possible to observe cases of read-in where more than one gene was overlapped (Figure 3.6).

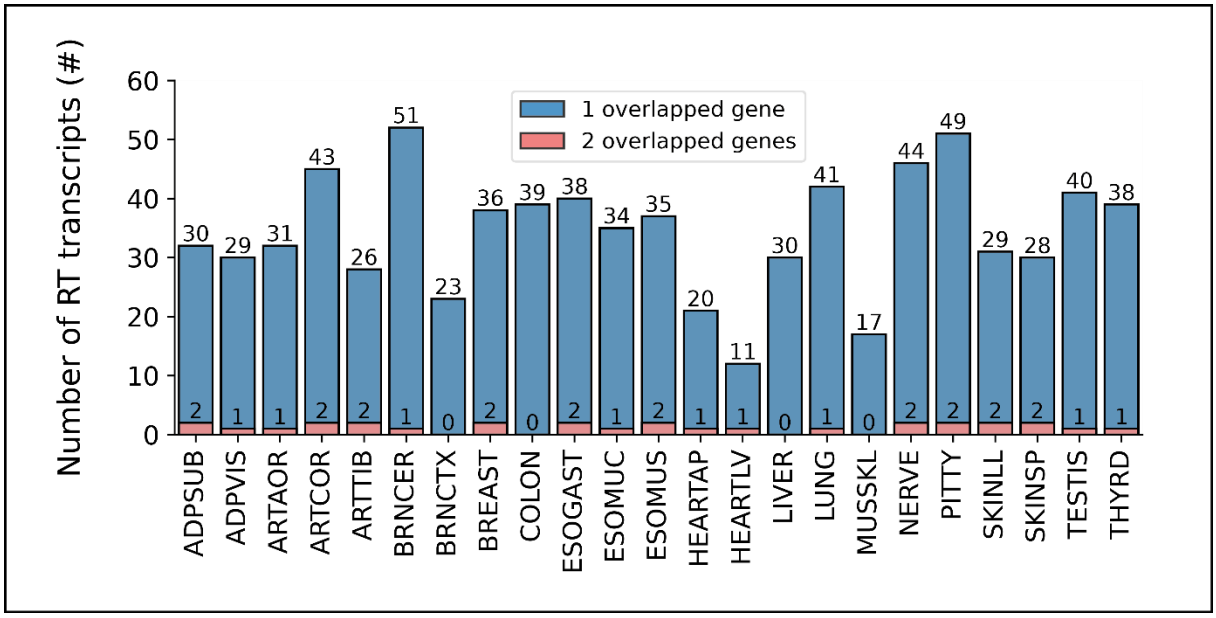

Figure 3.6. RT transcripts overlap downstream genes. Barplot showing the number of RT transcripts that overlapped downstream genes and how many genes were overlapped.

Read-in genes seem to mostly code for lncRNAs (Figure 3.7.A), which account for approximately 54% of these genes. This is in contrast with what was found for all expressed genes and for RT genes, where only approximately 14% and 10% of genes coded for lncRNAs, respectively (Figure 3.5.A and 3.5.B). When Mann-Whitney U Test was employed (FDR-adjusted p-value < 0.05), this difference in the proportion of lncRNAs was deemed significant. A possible explanation is the fact that these genes are significantly less expressed (Mann-Whitney U Test FDR-adjusted p-value < 0.05 for all tissues) than protein-coding genes (Figure 3.7.B). This difference in expression means that genes that code for lncRNAs probably are not as often blocked by the transcription machinery, which might make it slightly easier for RNAP II to continue propagating transcription over these genes.

Like with RT genes, gene set enrichment analyses were performed on read-in genes, but no enrichment for any particular function was found.

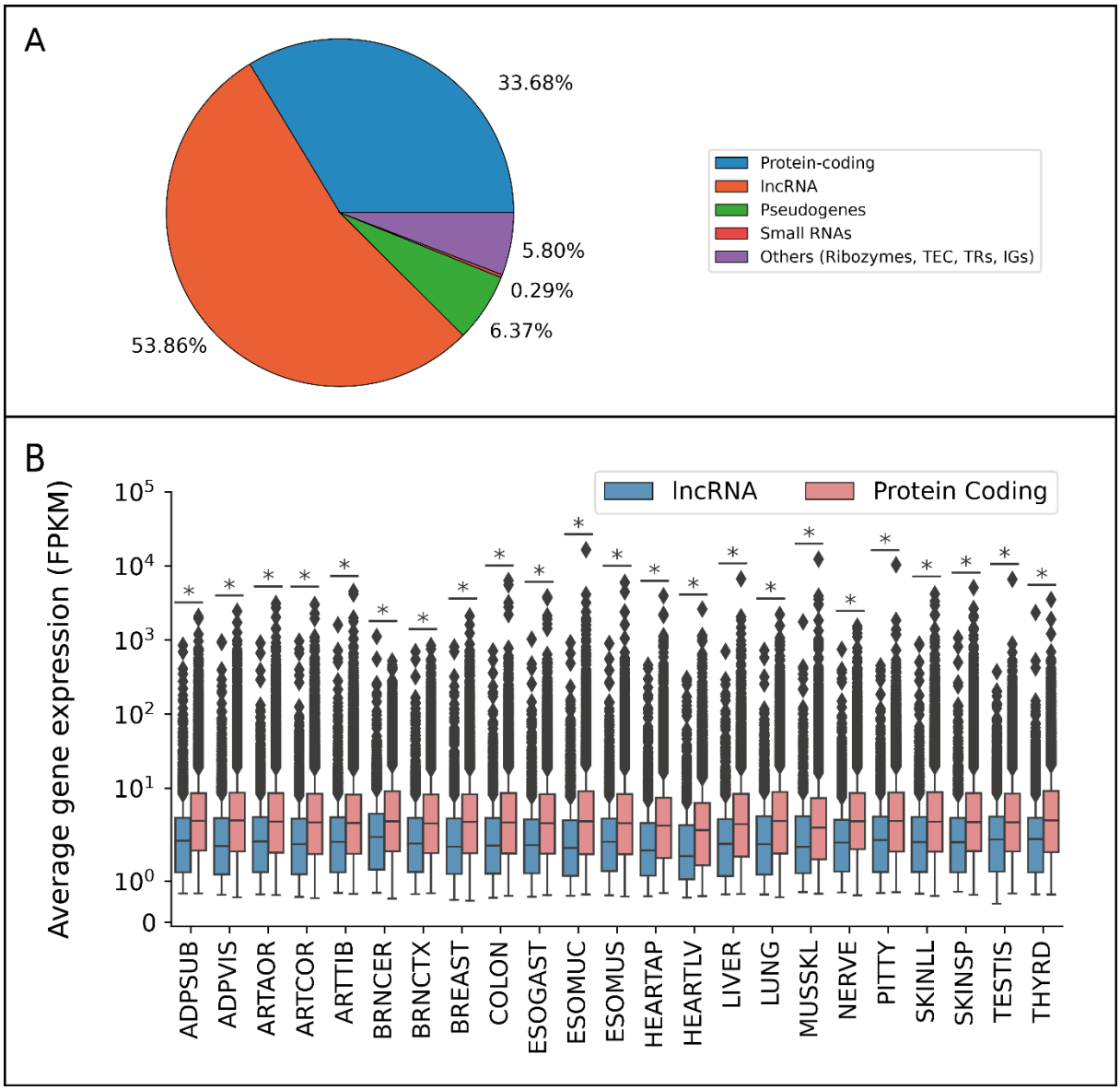

Figure 3.7. Read-in genes are enriched for lncRNAs. A) Pie plot showing the proportion of gene classifications for read-in genes. **B)** Boxplot showing expression levels of lncRNA and protein-coding genes. *****Mann-Whitney U Test FDR-adjusted p-value < 0.05.

To characterise the flanking regions at the 3' end of RT genes, the length of RT transcripts tails was analysed. This varies quite a lot among transcripts (Figure 3.8.A), with the minimum being 2 kb in all tissues (the set threshold for ARTDeco) and the reported maximum for each tissue ranging from 30 kb (skeletal muscle) up to 95 kb (breast tissue). The average length of the tail, however, is much more consistent, varying between 4491 bp (left ventricle) and 5978 bp (cerebellum). The length of the tail

seems to be neither correlated with the expression level of the transcript-producing gene (Pearson's r = 0.007, p-value = 0.154) nor with the expression level of the tail itself (Pearson's r = 0.105, p-value < 0.001). This suggests that expression level does not heavily influence TRT termination, which must instead be related to other features.

Next, the possibility that the transcription machinery of downstream genes could possibly act as a roadblock for TRT events, which was suggested by the enrichment of non-coding genes in read-in genes, was explored. For this, the distance between the 3' end of the RT transcript and the closest downstream expressed gene (that is not overlapped) was analysed. This distance is very heterogeneous, ranging between as little as 0 bp and more than 20000 kb (Figure 3.8.B). There is an observable difference in the number of closest genes belonging to the opposite or same strand in the 500 bp cohort, which is most probably caused by the filter applied to remove ambiguous genes. Still, the distribution of the distances between the RT transcript and the closest genes is not significantly different based on if the gene belongs to the same strand or not (Mann-Whitney U Test FDR-adjusted p-value > 0.91).

There was also no significant correlation between the tail length and the distance between the respective RT transcript and the closest downstream gene, whether that be in general (Pearson's r = 0.000, p-value = 0.938) or when taking into account which strand the closest gene belonged to (Pearson's r = 0.003, p-value = 0.763 for downstream genes in the same strand as the RT gene; Pearson's r = -0.003, p-value = 0.755 for those in the opposite strand).

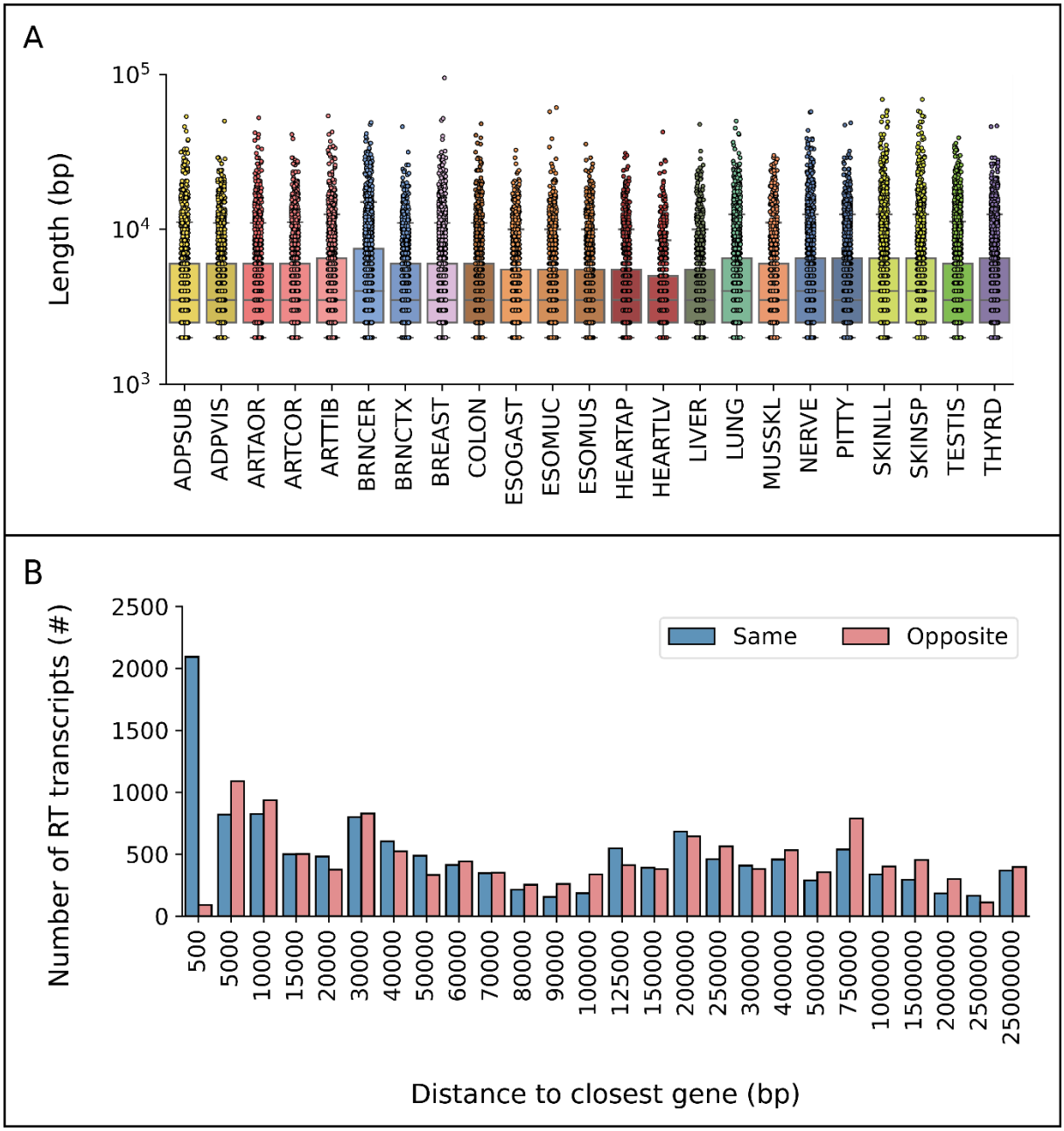

Figure 3.8. Distribution of RT transcript tail length and length of intergenic regions. A) Boxplot showcasing the length of RT transcript tails for every tissue. **B)** Barplot showing the distribution of the distance between the 3' end of RT transcripts and the closest downstream gene.

To understand if the distances between the RT gene and respective closest downstream genes were a feature that impacts TRT, the distance between the 3' end of all expressed genes and the closest expressed downstream gene was analysed (Figure 3.9.A). No significant differences were observed between the distance of NRT and RT genes to the respective closest downstream genes (Mann-Whitney U Test FDR-adjusted p-value > 0.05 for all tissues except testis). This goes against previous observations (Vilborg et al., 2017), where RT genes had been found to have closer downstream genes

than NRT genes. It is possible that this difference is due to the fact that the study focused on cells under pan-stress, as opposed to healthy tissues, but lab experiments comparing this metric in cells before and after undergoing stress would have to be performed to ascertain that.

A difference was observed between the distribution of these distances when downstream genes belonged to the same or opposite strand as the RT gene for the 500 bp cohort (Figure 3.9.B), most likely due to the filter applied to removed ambiguous genes, but this was not deemed significant (Mann-Whitney U Test FDR-adjusted p-value > 0.93).

Taken together with the detection of read-in transcription (Figure 3.6), these results seem to suggest that the existence of downstream genes in close proximity does not prevent the elongation of the RT transcript tail. This implies that other gene features must be involved in the termination of TRT. One possibility to study in the future would be an enrichment in the presence of PAS in the 3' end region of the transcript tails, which might lead to transcription termination. It would also be interesting to explore the chromatin structure near the 3' end of the transcript, as structures such as R-loops have already been shown to lead to RNAP II pausing (Skourti-Stathaki & Proudfoot, 2014), and so might be another player in the termination of TRT.

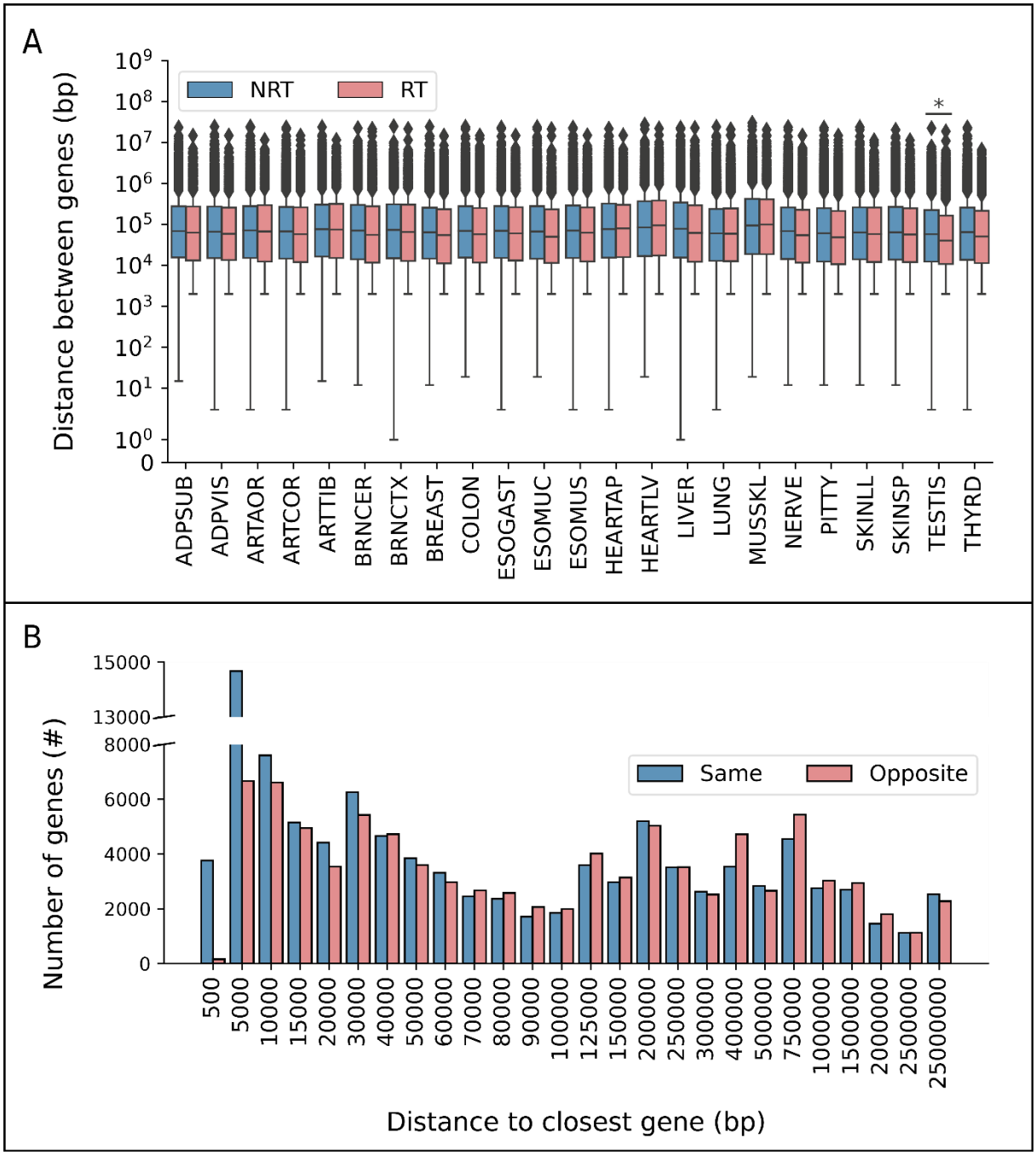

Figure 3.9. Distance between expressed genes. A) Boxplot showing the distance between NRT and RT genes and the respective closest downstream genes. **B)** Barplot showing the distribution of the distance between the 3' end of genes and the closest downstream gene. *****Mann-Whitney U Test FDR-adjusted p-value < 0.05.

4

CONCLUSION

In recent years, TRT has emerged as a hallmark of the mammalian cellular stress response, having been detected in cancer cells and under cellular stress conditions (e.g., viral infection, hypoxia, hyperosmotic stress, heat shock, and oxidative stress). This phenomenon is caused by inefficient 3' end cleavage of nascent RNAs, which leads to the synthesis of ncRNAs with long 3' UTR called RT transcripts. It has been speculated that these RT transcripts, which are retained in the nucleus, might serve the function of reinforcing the nuclear scaffold or altering the expression of other genes through transcriptional interference.

Until now, despite the growing research interest in this topic, TRT had not yet been studied in healthy tissues. Hence, this work focused on uncovering and assessing the prevalence of TRT across human healthy tissues, to better understand this phenomenon and its potential role in human health.

This work successfully showed that TRT is indeed pervasive in healthy human tissues, not only a characteristic of cells under stress conditions. Furthermore, genes that produced readthrough transcripts were found to be more highly expressed than those that did not and their expression was found to be slightly correlated with the coverage of the readthrough tail.

Also highlighted was the fact that most RT genes were present in more than one tissue and that RT genes were enriched for protein-coding genes comparatively to all expressed genes. However, gene set enrichment analyses revealed no enrichment for any pathways or functions, which suggests the existence of not yet known gene features that affect the production of RT transcripts.

It was also ascertained that TRT termination is neither influenced by expression levels nor the presence of expressed genes downstream, which would mean that termination must be determined by other gene features.

The next stage of research should be identifying the gene features that determine RT transcript production and TRT termination. It would also be important to determine if TRT levels are affected by demographics such as age, ethnic background and gender.

REFERENCES

Agostini, F., Zagalak, J., Attig, J., Ule, J., & Luscombe, N. M. (2021). Intergenic RNA mainly derives from nascent transcripts of known genes. *Genome Biology*, *22*(1), 1–19. https://doi.org/10.1186/S13059-021-02350-X/FIGURES/5

Alpert, T., Straube, K., Carrillo Oesterreich, F., & Neugebauer, K. M. (2020). Widespread Transcriptional Readthrough Caused by Nab2 Depletion Leads to Chimeric Transcripts with Retained Introns. *Cell Reports*, *33*(4). https://doi.org/10.1016/j.celrep.2020.108324

Amat, R., Böttcher, R., Le Dily, F., Vidal, E., Quilez, J., Cuartero, Y., Beato, M., De Nadal, E., & Posas, F. (2019). Rapid reversible changes in compartments and local chromatin organization revealed by hyperosmotic shock. *Genome Research*, *29*(1), 18–28. https://doi.org/10.1101/GR.238527.118

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., & Soboleva, A. (2013). NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, *41*(D1), D991–D995. https://doi.org/10.1093/NAR/GKS1193

Bauer, D. L. V., Tellier, M., Martínez-Alonso, M., Nojima, T., Proudfoot, N. J., Murphy, S., & Fodor, E. (2018). Influenza Virus Mounts a Two-Pronged Attack on Host RNA Polymerase II Transcription. *Cell Reports*, *23*(7), 2119-2129.e3. https://doi.org/10.1016/j.celrep.2018.04.047

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/J.2517-6161.1995.TB02031.X

Berkovits, B. D., & Mayr, C. (2015). Alternative 3′ UTRs act as scaffolds to regulate membrane protein localization. *Nature 2015 522:7556*, *522*(7556), 363–367. https://doi.org/10.1038/nature14321

Cardiello, J. F., Goodrich, J. A., & Kugel, J. F. (2018). Heat Shock Causes a Reversible Increase in RNA Polymerase II Occupancy Downstream of mRNA Genes, Consistent with a Global Loss in Transcriptional Termination. *Molecular and Cellular Biology*, *38*(18). https://doi.org/10.1128/MCB.00181-18

Castillo-Guzman, D., Hartono, S. R., Sanz, L. A., & Chédin, F. (2020). SF3B1-targeted Splicing Inhibition Triggers Global Alterations in Transcriptional Dynamics and R-Loop Metabolism. *BioRxiv*, 2020.06.08.130583. https://doi.org/10.1101/2020.06.08.130583

Caswell, T. A., Droettboom, M., Lee, A., Andrade, E. S. de, Hoffmann, T., Hunter, J., Klymak, J., Firing, E., Stansby, D., Varoquaux, N., Nielsen, J. H., Root, B., May, R., Elson, P., Seppänen, J. K., Dale, D., Lee, J.-J., McDougall, D., Straw, A., … Ivanov, P. (2021). matplotlib/matplotlib: REL: v3.5.1. *Zenodo*. https://doi.org/10.5281/ZENODO.5773480

Caudron-Herger, M., Müller-Ott, K., Mallm, J. P., Marth, C., Schmidt, U., Fejes-Tóth, K., & Rippe, K. (2011). Coding RNAs with a non-coding function: Maintenance of open chromatin structure. *Http://Dx.Doi.Org/10.4161/Nucl.2.5.17736*, *2*(5), 410–424. https://doi.org/10.4161/NUCL.2.5.17736

Chapman, R. D., Heidemann, M., Hintermair, C., & Eick, D. (2008). Molecular evolution of the RNA polymerase II CTD. *Trends in Genetics*, *24*(6), 289–296. https://doi.org/10.1016/J.TIG.2008.03.010

Chen, J., Bardes, E. E., Aronow, B. J., & Jegga, A. G. (2009). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Research*, *37*(Web Server issue). https://doi.org/10.1093/NAR/GKP427

Cramer, P. (2019). Organization and regulation of gene transcription. *Nature 2019 573:7772*, *573*(7772), 45–54. https://doi.org/10.1038/s41586-019-1517-4

Cunningham, F., Allen, J. E., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Austine-Orimoloye, O.,

Azov, A. G., Barnes, I., Bennett, R., Berry, A., Bhai, J., Bignell, A., Billis, K., Boddu, S., Brooks, L., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., … Flicek, P. (2022). Ensembl 2022. *Nucleic Acids Research*, *50*(D1), D988–D995. https://doi.org/10.1093/NAR/GKAB1049

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), 1–4. https://doi.org/10.1093/GIGASCIENCE/GIAB008

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/BIOINFORMATICS/BTS635

Eaton, J. D., Francis, L., Davidson, L., & West, S. (2020). A unified allosteric/torpedo mechanism for transcriptional termination on human protein-coding genes. *Genes & Development*, *34*(1–2), 132–145. https://doi.org/10.1101/GAD.332833.119/-/DC1

Eaton, J. D., & West, S. (2020). Termination of Transcription by RNA Polymerase II: BOOM! *Trends in Genetics*, *36*(9), 664–675. https://doi.org/10.1016/J.TIG.2020.05.008

Eick, D., & Geyer, M. (2013). The RNA polymerase II carboxy-terminal domain (CTD) code. *Chemical Reviews*, *113*(11), 8456–8490. https://doi.org/10.1021/CR400071F/ASSET/IMAGES/LARGE/CR-2013-00071F_0030.JPEG

Fisher, R. A. (1922). On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, *85*(1), 87. https://doi.org/10.2307/2340521

Godfrey, L., Crump, N. T., O'Byrne, S., Lau, I. J., Rice, S., Harman, J. R., Jackson, T., Elliott, N., Buck, G., Connor, C., Thorne, R., Knapp, D. J. H. F., Heidenreich, O., Vyas, P., Menendez, P., Inglott, S., Ancliff, P., Geng, H., Roberts, I., … Milne, T. A. (2020). H3K79me2/3 controls enhancer–promoter interactions and activation of the pan-cancer stem cell marker PROM1/CD133 in MLL-AF4 leukemia cells. *Leukemia 2020 35:1*, *35*(1), 90–106. https://doi.org/10.1038/s41375-020-0808-y

Grosso, A. R., Leite, A. P., Carvalho, S., Matos, M. R., Martins, F. B., Vítor, A. C., Desterro, J. M. P., Carmo-Fonseca, M., & de Almeida, S. F. (2015). Pervasive transcription read-through promotes aberrant expression of oncogenes and RNA chimeras in renal carcinoma. *ELife*, *4*(NOVEMBER2015). https://doi.org/10.7554/ELIFE.09214

Hall, L. L., Carone, D. M., Gomez, A. V., Kolpa, H. J., Byron, M., Mehta, N., Fackelmayer, F. O., & Lawrence, J. B. (2014). Stable C0T-1 repeat RNA is abundant and is associated with euchromatic interphase chromosomes. *Cell*, *156*(5), 907–919. https://doi.org/10.1016/j.cell.2014.01.042

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature 2020 585:7825*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Heinz, S., Texari, L., Hayes, M. G. B., Urbanowski, M., Chang, M. W., Givarkes, N., Rialdi, A., White, K. M., Albrecht, R. A., Pache, L., Marazzi, I., García-Sastre, A., Shaw, M. L., & Benner, C. (2018). Transcription Elongation Can Affect Genome 3D Structure. *Cell*, *174*(6), 1522-1536.e22. https://doi.org/10.1016/j.cell.2018.07.047

Hennig, T., Michalski, M., Rutkowski, A. J., Djakovic, L., Whisnant, A. W., Friedl, M. S., Jha, B. A., Baptista, M. A. P., L'Hernault, A., Erhard, F., Dölken, L., & Friedel, C. C. (2018). HSV-1-induced disruption of transcription termination resembles a cellular stress response but selectively increases chromatin accessibility downstream of genes. *PLOS Pathogens*, *14*(3), e1006954. https://doi.org/10.1371/JOURNAL.PPAT.1006954

Hernandez, N. (1985). Formation of the 3′ end of U1 snRNA is directed by a conserved sequence located downstream of the coding region. *The EMBO Journal*, *4*(7), 1827–1837. https://doi.org/10.1002/J.1460-2075.1985.TB03857.X

Herr, A. J., Jensen, M. B., Dalmay, T., & Baulcombe, D. C. (2005). RNA polymerase IV directs silencing of endogenous DNA. *Science*, *308*(5718), 118–120. https://doi.org/10.1126/SCIENCE.1106910/SUPPL_FILE/PAPV2.PDF

Herzel, L., Straube, K., & Neugebauer, K. M. (2018). Long-read sequencing of nascent RNA reveals coupling among RNA processing events. *Genome Research*, *28*(7), 1008–1019. https://doi.org/10.1101/GR.232025.117

Hunt, J. (2019). *A Beginners Guide to Python 3 Programming*. https://doi.org/10.1007/978-3-030-20290-3

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Kamieniarz-Gdula, K., & Proudfoot, N. J. (2019). Transcriptional Control by Premature Termination: A Forgotten Mechanism. *Trends in Genetics*, *35*(8), 553–564. https://doi.org/10.1016/J.TIG.2019.05.005

Kanno, T., Huettel, B., Mette, M. F., Aufsatz, W., Jaligot, E., Daxinger, L., Kreil, D. P., Matzke, M., & Matzke, A. J. M. (2005). Atypical RNA polymerase subunits required for RNA-directed DNA methylation. *Nature Genetics 2005 37:7*, *37*(7), 761–765. https://doi.org/10.1038/ng1580

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, and D. (2002). The Human Genome Browser at UCSC. *Genome Research*, *12*(6), 996–1006. https://doi.org/10.1101/GR.229102

Lopez-Delisle, L., Rabbani, L., Wolff, J., Bhardwaj, V., Backofen, R., Grüning, B., Ramírez, F., & Manke, T. (2021). pyGenomeTracks: reproducible plots for multivariate genomic datasets. *Bioinformatics*, *37*(3), 422–423. https://doi.org/10.1093/BIOINFORMATICS/BTAA692

Luo, W., Ji, Z., Pan, Z., You, B., Hoque, M., Li, W., Gunderson, S. I., & Tian, B. (2013). The Conserved Intronic Cleavage and Polyadenylation Site of CstF-77 Gene Imparts Control of 3′ End Processing Activity through Feedback Autoregulation and by U1 snRNP. *PLOS Genetics*, *9*(7), e1003613. https://doi.org/10.1371/JOURNAL.PGEN.1003613

Lyons, D. E., McMahon, S., & Ott, M. (2020). A combinatorial view of old and new RNA polymerase II modifications. *Https://Doi.Org/10.1080/21541264.2020.1762468*, *11*(2), 66–82. https://doi.org/10.1080/21541264.2020.1762468

Mahat, D. B., Salamanca, H. H., Duarte, F. M., Danko, C. G., & Lis, J. T. (2016). Mammalian Heat Shock Response and Mechanisms Underlying Its Genome-wide Transcriptional Regulation. *Molecular Cell*, *62*(1), 63–78. https://doi.org/10.1016/j.molcel.2016.02.025

Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Https://Doi.Org/10.1214/Aoms/1177730491*, *18*(1), 50–60. https://doi.org/10.1214/AOMS/1177730491

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., & Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, *18*(9), 1509–1517. https://doi.org/10.1101/GR.079558.108

Mazo, A., Hodgson, J. W., Petruk, S., Sedkov, Y., & Brock, H. W. (2007). Transcriptional interference: an unexpected layer of complexity in gene regulation. *Journal of Cell Science*, *120*(16), 2755–2761. https://doi.org/10.1242/JCS.007633

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, 56–61. https://doi.org/10.25080/MAJORA-92BF1922-00A

Morgan, M., Shiekhattar, R., Shilatifard, A., & Lauberth, S. M. (2022). It's a DoG-eat-DoG world—altered transcriptional mechanisms drive downstream-of-gene (DoG) transcript production. *Molecular Cell*, *82*(11), 1981–1991. https://doi.org/10.1016/J.MOLCEL.2022.04.008

Muniz, L., Deb, M. K., Aguirrebengoa, M., Lazorthes, S., Trouche, D., & Nicolas, E. (2017). Control of Gene Expression in Senescence through Transcriptional Read-Through of Convergent Protein-Coding Genes. *Cell Reports*, *21*(9), 2433–2446. https://doi.org/10.1016/j.celrep.2017.11.006

O'Reilly, D., Kuznetsova, O. V., Laitem, C., Zaborowska, J., Dienstbier, M., & Murphy, S. (2014). Human snRNA genes use polyadenylation factors to promote efficient transcription termination. *Nucleic Acids Research*, *42*(1), 264–275. https://doi.org/10.1093/NAR/GKT892

Onodera, Y., Haag, J. R., Ream, T., Nunes, P. C., Pontes, O., & Pikaard, C. S. (2005). Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation- dependent heterochromatin formation. *Cell*, *120*(5), 613–622. https://doi.org/10.1016/j.cell.2005.02.007

Pan, X., & Frank Huang, L. (2022). Multi-omics to characterize the functional relationships of R-loops with epigenetic modifications, RNAPII transcription and gene expression. *Briefings in Bioinformatics*, *23*(4). https://doi.org/10.1093/BIB/BBAC238

Pontier, D., Yahubyan, G., Vega, D., Bulski, A., Saez-Vasquez, J., Hakimi, M. A., Lerbs-Mache, S., Colot, V., & Lagrange, T. (2005). Reinforcement of silencing at transposons and highly repeated sequences requires the concerted action of two distinct RNA polymerases IV in Arabidopsis. *Genes & Development*, *19*(17), 2030–2040. https://doi.org/10.1101/GAD.348405

Proudfoot, N. J. (2016). Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science*, *352*(6291). https://doi.org/10.1126/SCIENCE.AAD9926/ASSET/4BAC16B8-0F78-40E3-AE87-7EA560D74D36/ASSETS/GRAPHIC/352_AAD9926_FA.JPEG

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/BIOINFORMATICS/BTQ033

Reabansg, M. E., Lebowitzs, J., Griffin111, J. A., & State, N. Y. (1994). THE JOURNAL OF BIOLOGICAL CHEMISTRY Transcription Induces the Formation of a Stable RNAoDNA Hybrid in the Immunoglobulin a Switch Region*. *Journal of Biological Chemistry*, *269*(34), 21850–21857. https://doi.org/10.1016/S0021-9258(17)31881-1

Reback, J., jbrockmendel, McKinney, W., Bossche, J. Van den, Augspurger, T., Cloud, P., Hawkins, S., Roeschke, M., gfyoung, Sinhrks, Klein, A., Petersen, T., Hoefler, P., Tratner, J., She, C., Ayd, W., Naveh, S., Garcia, M., Darbyshire, J., … Seabold, S. (2021). pandas-dev/pandas: Pandas 1.3.5. *Zenodo*. https://doi.org/10.5281/ZENODO.5774815

Rosa-Mercado, N. A., & Steitz, J. A. (2022). Who let the DoGs out? – biogenesis of stress-induced readthrough transcripts. *Trends in Biochemical Sciences*, *47*(3), 206–217. https://doi.org/10.1016/j.tibs.2021.08.003

Rosa-Mercado, N. A., Zimmer, J. T., Apostolidi, M., Rinehart, J., Simon, M. D., & Steitz, J. A. (2021). Hyperosmotic stress alters the RNA polymerase II interactome and induces readthrough transcription despite widespread transcriptional repression. *Molecular Cell*, *81*(3), 502-513.e4. https://doi.org/10.1016/j.molcel.2020.12.002

Roth, S. J., Heinz, S., & Benner, C. (2020). ARTDeco: Automatic readthrough transcription detection. *BMC Bioinformatics*, *21*(1), 1–22. https://doi.org/10.1186/S12859-020-03551-0/TABLES/1

Russell, J., & Zomerdijk, J. C. B. M. (2005). RNA-polymerase-I-directed rDNA transcription, life and works. *Trends in Biochemical Sciences*, *30*(2), 87–96. https://doi.org/10.1016/j.tibs.2004.12.008

Rutkowski, A. J., Erhard, F., L'Hernault, A., Bonfert, T., Schilhabel, M., Crump, C., Rosenstiel, P., Efstathiou, S., Zimmer, R., Friedel, C. C., & Dölken, L. (2015). Widespread disruption of host transcription termination in HSV-1 infection. *Nature Communications 2015 6:1*, *6*(1), 1–15. https://doi.org/10.1038/ncomms8126

Sáez-Vásquez, J., & Delseny, M. (2019). Ribosome Biogenesis in Plants: From Functional 45S Ribosomal DNA Organization to Ribosome Assembly Factors. *The Plant Cell*, *31*(9), 1945–1967. https://doi.org/10.1105/TPC.18.00874

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. *Proceedings of the 9th Python in Science Conference*, 92–96. https://doi.org/10.25080/MAJORA-92BF1922-011

Shalgi, R., Hurt, J. A., Lindquist, S., & Burge, C. B. (2014). Widespread inhibition of posttranscriptional splicing shapes the cellular transcriptome following heat shock. *Cell Reports*, *7*(5), 1362–1370. https://doi.org/10.1016/j.celrep.2014.04.044

Sims, R. J., Mandal, S. S., & Reinberg, D. (2004). Recent highlights of RNA-polymerase-II-mediated transcription. *Current Opinion in Cell Biology*, *16*(3), 263–271. https://doi.org/10.1016/J.CEB.2004.04.004

Skourti-Stathaki, K., Kamieniarz-Gdula, K., & Proudfoot, N. J. (2014). R-loops induce repressive chromatin marks over mammalian gene terminators. *Nature 2014 516:7531*, *516*(7531), 436–439. https://doi.org/10.1038/nature13787

Skourti-Stathaki, K., & Proudfoot, N. J. (2014). A double-edged sword: R loops as threats to genome integrity and powerful regulators of gene expression. *Genes & Development*, *28*(13), 1384–1396. https://doi.org/10.1101/GAD.242990.114

Soucek, S., Zeng, Y., Bellur, D. L., Bergkessel, M., Morris, K. J., Deng, Q., Duong, D., Seyfried, N. T., Guthrie, C., Staley, J. P., Fasken, M. B., & Corbett, A. H. (2016). Evolutionarily Conserved Polyadenosine RNA Binding Protein Nab2 Cooperates with Splicing Machinery To Regulate the Fate of Pre-mRNA. *Molecular and Cellular Biology*, *36*(21), 2697–2714. https://doi.org/10.1128/MCB.00402-16/SUPPL_FILE/ZMB999101333SO2.PDF

Tomson, B. N., & Arndt, K. M. (2013). The many roles of the conserved eukaryotic Paf1 complex in regulating transcription, histone modifications, and disease states. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, *1829*(1), 116–126. https://doi.org/10.1016/J.BBAGRM.2012.08.011

Vilborg, A., Passarelli, M. C., Yario, T. A., Tycowski, K. T., & Steitz, J. A. (2015). Widespread Inducible Transcription Downstream of Human Genes. *Molecular Cell*, *59*(3), 449–461. https://doi.org/10.1016/j.molcel.2015.06.016

Vilborg, A., Sabath, N., Wiesel, Y., Nathans, J., Levy-Adam, F., Yario, T. A., Steitz, J. A., & Shalgi, R. (2017). Comparative analysis reveals genomic features of stress-induced transcriptional readthrough. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(40), E8362–E8371. https://doi.org/10.1073/PNAS.1711120114

Vilborg, A., & Steitz, J. A. (2016). Readthrough transcription: How are DoGs made and what do they do? *Https://Doi.Org/10.1080/15476286.2016.1149680*, *14*(5), 632–636. https://doi.org/10.1080/15476286.2016.1149680

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … Vázquez-Baeza, Y. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods 2020 17:3*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics 2008 10:1*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/JOSS.03021

White, R. J. (2011). Transcription by RNA polymerase III: more complex than we thought. *Nature Reviews Genetics 2011 12:7*, *12*(7), 459–463. https://doi.org/10.1038/nrg3001

Wierzbicki, A. T., Haag, J. R., & Pikaard, C. S. (2008). Noncoding Transcription by RNA Polymerase Pol IVb/Pol V Mediates Transcriptional Silencing of Overlapping and Adjacent Genes. *Cell*, *135*(4), 635–648. https://doi.org/10.1016/j.cell.2008.09.035

Wiesel, Y., Sabath, N., & Shalgi, R. (2018). DoGFinder: A software for the discovery and quantification of readthrough transcripts from RNA-seq. *BMC Genomics*, *19*(1), 1–7. https://doi.org/10.1186/S12864-018-4983-4/FIGURES/3

Zhang, H., Rigo, F., & Martinson, H. G. (2015). Poly(A) Signal-Dependent Transcription Termination Occurs through a Conformational Change Mechanism that Does Not Require Cleavage at the Poly(A) Site. *Molecular Cell*, *59*(3), 437–448. https://doi.org/10.1016/j.molcel.2015.06.008

A

APPENDIX

The script used in this work, as well as the lists of genes and transcripts detected (both expressed in each tissue and in all tissues), can be found here: https://github.com/mpluz/mluzmgmb2022thesis.