Characterizing the transposome and its activity

Dissertation

in Partial Fulfilment of the Requirements for the Degree of

“doctor rerum naturalium” (Dr. rer. nat.)

Submitted to the Council of the Faculty of Biological Sciences

of Friedrich Schiller University Jena

by M. Sc. Robert Schwarz

born on 28.03.1988 in Gera

Die vorliegende Arbeit wurde in der Zeit von Juli 2018 bis Januar 2023 am Leibniz-Institut für Alternsforschung – Fritz-Lipmann Institut e.V. (FLI) in Jena angefertigt und am 22.11.2023 verteidigt.

Gutachter:

1. Prof. Dr. Dr. Steve Hoffmann,
   Hoffmann Research Group,
   Leibniz Institute on Aging – Fritz Lipmann Institute, Jena, Germany
2. Prof. Dr. Christoph Englert,
   Englert Research Group,
   Leibniz Institute on Aging – Fritz Lipmann Institute, Jena, Germany
3. Prof. Dr. Ivo Große
   Hoffmann Research Group,
   Institute for Informatics – Martin Luther University Halle-Wittenberg, Halle, Germany

We absolutely must leave room for doubt

or there is no progress and there is no learning.

There is no learning without having to pose a question.

And a question requires doubt.

-

Richard Feynman

Abstract

Sequencing technologies enabled us to decode sequence of bases that shape the genome and to identify sequence stretches with biological functions, such as genes. Such technologies also enable us to study the expression and regulation base sequences. Intriguingly, transposable elements (TEs) can occupy a substantial proportion of a genome. TEs provide a comprehensive repertoire of (non)-coding sequences with biological functions that can potentially impact gene expression. However, quantification of TE expression remains challenging due to their high sequence similarity and the limited lengths of base sequences decoded by state-of-the-art sequencing technologies.

In my thesis, I evaluated five TE quantification software applications with respect to their performance in quantifying the expression of individual TEs. Originally, three of the five tools were unable to quantify the expression of individual TEs. I slightly modified these tools to enable the output of locus-specific quantification data. My tool evaluation was based on simulated datasets for model and non model organisms created using publicly available as well as self-implemented simulation software. Notably, we found that SalmonTE, a tool originally designed to asses TE expression at the family-level, could recover simulated TE expression fairly accurately upon modification. Thus, I showed that modified SalmonTE can be applied for reliable differential expression analyses in model and non-model organisms.

In the second part of my thesis, I applied modified SalmonTE to study (differentially) TE expression in blood, brain, and skin of mice of different ages (6 and 24 months). While previous family-level studies of TE expression identified up-regulation of TEs as a characteristic of aging, my results indicate that individual TEs are also commonly down-regulated during aging. Integration of transcription start site sequencing data identified TE regions, *i.e.*, stretches of expressed TEs that share common transcription start sites, to be nested in genes with highly tissue-specific functions. Co-regulation of TEs and host genes indicates potential biological functions of independently expressed TEs concerning the transcriptional regulation of genes involved in highly tissue-specific pathways. Analyses of the putative promoter regions of independently expressed TEs identified transcription factors of the Sox family as candidates controlling their regulation. Together, this study revealed the expression dynamics of individual TEs during aging and provides a comprehensive resource of independently expressed TEs. These data can be a promising starting point to intensify research into locus-specific TE expression to gain a better understanding of the biological functions, interactions, and regulation of TEs.

In the last part of my thesis, I developed an expression database on p53 and cell cycle-dependent gene regulation with an intuitive web interface. This database serves as a blueprint to make the expression data of TEs and their associated genes easily accessible to the scientific community.

Zusammenfassung

Sequenzierungstechnologien haben es uns ermöglicht, die Abfolge der Basen zu entschlüsseln, die das Genom formen, sowie Sequenzabschnitte mit biologischen Funktionen, wie z. B. Gene, zu identifizieren. Diese Technologien ermöglichen es uns auch, die Expression und Regulierung von Basensequenzen zu untersuchen. Interessanterweise können transponierbare Elemente (TEs) einen erheblichen Teil eines Genoms einnehmen. TEs bieten ein umfassendes Repertoire an (nicht)-kodierenden Sequenzen mit biologischen Funktionen, die möglicherweise die Genexpression beeinflussen können. Die Quantifizierung der TE-Expression bleibt jedoch aufgrund ihrer hohen Sequenzähnlichkeit und der begrenzten Länge der Basensequenzen, die mit modernen Sequenzierungstechnologien entschlüsselt werden können, eine Herausforderung.

In meiner Dissertation habe ich fünf TE-Quantifizierungssoftwareanwendungen im Hinblick auf ihre Leistung bei der Quantifizierung der Expression individueller TEs bewertet. Ursprünglich waren drei der fünf Tools nicht in der Lage, die Expression einzelner TEs zu quantifizieren. Ich habe diese Tools leicht modifiziert, um die Ausgabe von lokusspezifischen Quantifizierungsdaten zu ermöglichen. Meine Bewertung der Tools basierte auf simulierten Datensätzen für Modell- und Nicht-Modellorganismen, die mit öffentlich zugänglicher und selbst implementierter Simulationssoftware erstellt wurden. Insbesondere haben wir festgestellt, dass SalmonTE, ein Tool, das ursprünglich für die Bewertung der TE-Expression auf Familienebene entwickelt wurde, die simulierte TE-Expression nach einer Modifizierung ziemlich genau messen konnte. So konnte ich zeigen, dass das adaptierte SalmonTE für zuverlässige differentielle Expressionsanalysen in Modell- und Nicht-Modellorganismen eingesetzt werden kann.

Im zweiten Teil meiner Arbeit habe ich das adaptierte SalmonTE eingesetzt, um die TE-Expression in Blut, Gehirn und Haut von Mäusen unterschiedlichen Alters (6 und 24 Monate) zu untersuchen. Während frühere Studien zur TE-Expression auf Familienebene eine Hochregulierung von TEs als Merkmal des Alterns identifiziert haben, deuten meine Ergebnisse darauf hin, dass einzelne TEs während des Alterns auch häufig herunterreguliert werden. Durch die Integration von Daten zur Sequenzierung von Transkriptionsstartstellen wurden TE-Regionen identifiziert, d. h. Abschnitte exprimierter TEs, die gemeinsame Transkriptionsstartstellen aufweisen und in Genen mit sehr gewebespezifischen Funktionen eingebettet sind. Die Ko-Regulation von TEs und Wirtsgenen deutet auf mögliche biologische Funktionen unabhängig exprimierter TEs hin, die die Transkriptionsregulation von Genen betreffen, die an hochgradig gewebespezifischen Funktionen beteiligt sind. Analysen der potentiellen Promotorregionen unabhängig exprimierter TEs identifizierten Transkriptionsfaktoren der Sox-Familie als Kandidaten, die deren Regulierung

kontrollieren. Insgesamt hat diese Studie die Expressionsdynamik einzelner TEs während des Alterns aufgezeigt und bietet eine umfassende Ressource unabhängig exprimierter TEs. Diese Daten können ein vielversprechender Ausgangspunkt sein, um die Erforschung der lokusspezifischen TE-Expression zu intensivieren und ein besseres Verständnis der biologischen Funktionen, Interaktionen und der Regulierung von TEs zu erlangen.

Im letzten Teil meiner Dissertation habe ich eine Expressionsdatenbank für p53 und zellzyklusabhängige Genregulation mit einer intuitiven Webschnittstelle entwickelt. Diese Datenbank dient als Blaupause, um die Expressionsdaten von TEs und den mit ihnen verbundenen Genen der wissenschaftlichen Gemeinschaft leicht zugänglich zu machen.

Table of contents

Abstract ... iii

Zusammenfassung ... v

Table of contents ... vii

Introduction ... 1

1.1. Prolog ... 1

1.2. Transposable elements ... 2

1.2.1. Classification of transposable elements ... 2

1.2.2. TEs as integral part of the genome ... 4

1.2.3. TEs as architects of genome structure and instructors of gene regulation ... 4

1.2.4. TEs expression during aging ... 6

1.2.5. Quantification of Transposable element expression ... 7

1.3. Databases providing comprehensive expression profiles ... 9

1.4. Thesis focus and aim ... 11

Manuscripts ... 12

2.1. Overview of Manuscripts ... 12

2.2. Manuscript 1 (M1) – Locus-specific expression analysis of transposable elements ... 13

2.3. Manuscript 2 (M2) – Expression differences of transposable elements during aging affect major tissue-specific pathways ... 25

2.4. Manuscript 3 (M3) – TargetGeneReg 2.0: a comprehensive web-atlas for p53, p63, and cell cycle-dependent gene regulation ... 72

Discussion ... 85

3.1. Evaluation of tools with respect to locus-specific expression quantification of TEs ... 86

3.2. Locus-specific TE expression quantification during the process of aging ... 88

3.3. Blueprint of a data base for differential expression data ... 91

Conclusion & Outlook ... 93

Bibliography ... 95

Abbreviation 103
Appendix 104
  7.1. Manuscript 1 – Form 2 104
  7.2. Manuscript 2 – Form 2 104
  7.3. Manuscript 3 – Form 2 105
Danksagung 106
Statement of authorship 107

Introduction

1.1. Prolog

The blueprint of all living matter is encoded in the genome. The genome is full of components that orchestrate the process of live. Since the description of the deoxyribonucleic acid (DNA) structure by James Watson and Francis Crick [1], enormous efforts have been made to decipher the composition of the DNA code and its function. Thus, the development of a method for determining the base sequence of DNA by Frederick Sanger [2] and others was an important milestone in the history of genetics. Decades later, the development of next generation high-throughput sequencing methods advanced genome analysis to a new era by drastically increasing availability and reducing costs. Complementary sequencing protocols, *e.g.*, to study ribonucleic acid (RNA) expression (RNA-Seq) or DNA-associated factors via chromatin immunoprecipitation DNA sequencing, opened the door to get unprecedented insights into the dynamics of gene expression and their regulation.

The initial publication of the human genome reported that only 2% of the genome are protein coding genes [3], reinvigorating controversial debates about the functional role of the other 98%. Notably, the Human Genome Project estimated that ~45% of the human genome consists of transposable elements (TEs) [3]. Initially described by Barbara McClintock and termed “controlling elements” during the 1950s [4], TEs are genomic sequences that can change their location (DNA-transposons) or create copies of themselves (retrotransposons) within their host genomes. In fact, the accumulation of TEs can be observed across eukaryotes, *e.g.*, mouse (~39%, [5]), zebrafish (~50%, [6]), and maize (between ~64% and ~85%, [7, 8]). McClintock received the Nobel Prize for her groundbreaking discovery in 1983.

The activity of TEs is often associated with diseases like cancer [9-11], neurological impairments [12], or aging [13]. One of the mechanisms by which TEs can contribute to pathological or deteriorating processes is the recognition of TE products by cell defense mechanisms leading to the inflammatory processes [14, 15]. Furthermore, the *de novo* integration of TEs into the genome has the potential to negatively affect coding genes and impact the genome’s integrity [16, 17]. On the other hand, TEs may also have constructive roles in the genome’s architecture and provide regulatory components for genes in close proximity [18, 19]. For example, TEs provide landing platforms for proteins such as transcription factors (TF) [18, 20, 21] or provide molecules regulating the activity of closely located genes [22, 23].

Consequently, the biological role of TEs appears complex, and their repetitive occurrence in genomes complicates their analysis. Most of the current knowledge is based on analysis of either specific TE loci or entire groups of TE elements sharing sequence similarities. The latter approach, however, impedes the resolution of the precise genomic location of active TEs. In my thesis, I focus on the analysis of locus-specific TE activity on a genome-wide scale. In the following, I am providing a brief introduction concerning the categorization and regulation of TEs, before moving on to their biological implication and the quantification of TE expression.

1.2. Transposable elements

1.2.1. Classification of transposable elements

Briefly, transposable elements are components of the genome that have inserted themselves into new genomic loci. Barbara McClintock was the first scientist to discover such "jumping genes" in maize several decades ago [4]. Since then, with few exceptions, TEs have been discovered in almost all higher eukaryotes that have been sequenced [24, 25]. The whole set of TEs in a genome are referred to as the transposome. In general, TEs are separated into two classes based on their transposition mechanism (Figure 1). Elements of Class I (retrotransposons) use a copy-and-paste mechanism. These elements are transcribed into RNA, reverse-transcribed into DNA and ultimately integrated at a new locus of the host genome [26]. Therefore, each successful transposition event of a retrotransposon leads to a novel copy. On the other hand, Class II elements (DNA transposons) use a cut-and-paste mechanism, which means that the element is excised from the genome and integrated into a new genomic locus [27, 28].

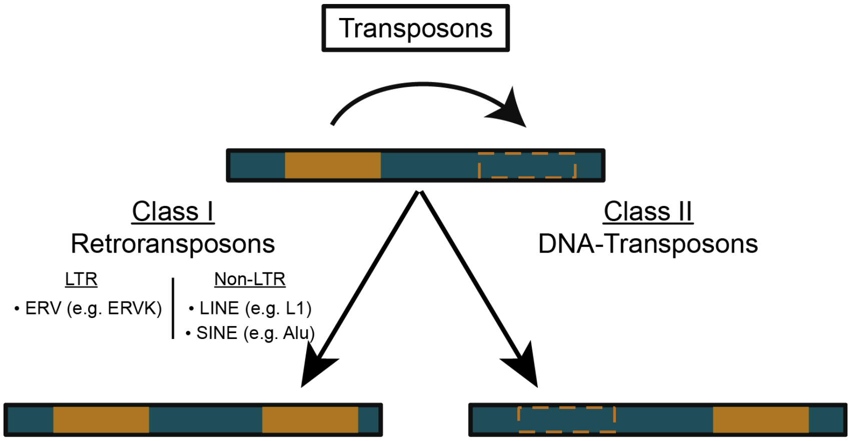


Figure 1 Classification of transposons. Transposons are categorized into two classes by their used transposition mechanism. Class I elements (retrotransposons) use a copy-paste mechanism to propagate within their host genome. Retrotransposons are further classified into LINE, SINE and LTR-retrotransposons. Class II elements (DNA-transposons) use a cut-paste mechanism to change the loci within the host genome. **Abbreviations:** ERV – endogenous retrovirus; ERVK –

endogenous retrovirus K; LINE – long interspersed nuclear element; LTR – long terminal repeats; SINE – short interspersed nuclear element

Retrotransposons are further subdivided into Long Terminal Repeat (LTR) retrotransposons and non-LTR-retrotransposons while the latter are further comprised Long Interspersed Nuclear Elements (LINE) and Short Interspersed Nuclear Elements (SINE). Copies with a common ancestor are further grouped into families, *e.g.*, LINE1 (L1) or primate-specific Alu elements, based on common structural features. Individual TE elements exhibiting high sequence similarity and sufficient differences to other instances within the same family are grouped into subfamilies.

Characteristically, LTR-retrotransposons have a body of protein-coding open reading frames (ORFs) enclosed by LTRs at their 5'- and 3'-ends [24, 29]. As this enclosure is reminiscent of the structure of endogenous retroviruses (ERV) [30, 31] it is assumed that LTRs are descendants of ERVs, which have lost the capability for extracellular replication [32, 33]. ORFs of LTR-retrotransposons encode the genes *gag*, *pol*, and *env,* while the coding region for the *env* gene is usually deleted [32] and is not essential for transposition [34]. The *gag* gene encodes a capsid and a nucleocapsid protein, while *pol* provides the blueprints for a protease, the reverse transcriptase (RT), and the integrase (IN) [24, 35]. Notably, LTR-regions contain promoter sequences facilitating binding of RNA polymerase (Pol) II [36, 37], and other transcription factor binding sites (TFBSs) [37]. Prototypically, the RNA of LTR elements is reverse transcribed into DNA in the cytosol of a cell, transferred to the nucleus, and integrated into the DNA by the IN [31]. Intact LTR-retrotransposons encode for all proteins necessary for their transposition, hence, they are classified as autonomous TEs.

In contrast, LINEs and SINEs lack LTRs at their 5' and 3'-ends (non-LTR-retrotransposons). Full length LINE elements do have a length of 6-7 kb [38] and contain a Pol II binding site within their 5'-end as well as two ORFs: ORF1 and ORF2. The ORF2 encodes for an endonuclease (EN) and the RT. Once translated, ORF1 and ORF2 build a ribonucleoprotein complex (RNP) that binds the LINE messenger RNA (mRNA) and transfers it to the nucleus. The EN nicks the DNA and the RT uses the free 3'OH as primer for the mRNA template in the RNP to reverse transcribe the RNA into the DNA beginning at the 3'-end [39-41]. Since intact LINEs encode for all proteins that are needed for the transposition they are also part of autonomous TEs. In comparison, SINEs do not have any ORF that code for proteins, which they can use for transposition (non-autonomous TEs) and rely on the transposition apparatus of LINEs [42-45]. While the origin of LINEs are uncertain [46], SINEs may derive from tRNA, 7SL RNA, and 5S RNA, contain a Pol III promoter [24], and have a length of 80-500 base pairs (bp) [38].

The majority of DNA transposons code for a transposase that is flanked by terminal inverted repeats (TIRs) of variable length. The TIR sequences are used to categorize these elements into nine super families [24]. The self-encoded transposase recognizes the TIRs, cuts out the DNA transposon sequence, and integrates it at another locus of the host genome. Copies of DNA transposons can be created during chromosome replication when they are reinserted in front of the replication fork [47].

1.2.2. TEs as integral part of the genome

The proliferation of TEs can have a fundamental impact on the size of host genomes. It has been reported that the genome size positively correlates with the amount of TEs contained in a genome [48, 49]. Retrotransposons usually dominate the TE content of the majority of eukaryotic genomes [3, 46, 50] due to their copy-and-paste mechanism. Notably, the integration process is not always perfect. For example, the integration of LINEs begins at the 3'end and an interruption during the reverse transcription process leads to 5'-end-truncated TE sequences [41, 51, 52]. Such 5'-end truncated elements may lack essential binding sites, so that further transpositions are impaired [53]. Interestingly, truncated ORF1 proteins of L1 elements provide for a suppression of full length L1 elements [54]. Indeed, the majority of TEs within human and mouse are immobilized [55] while only a small fraction of TEs are actually mobile in mammalian genomes [41, 47]. Besides, TEs have accumulated mutations over millions of years additionally impacting the transposition capability [56]. Mutations in TEs are useful to estimate the age of individual TEs. The Kimura distance [57] is a frequently used measure that serves as a proxy for the age of TEs. To calculate the Kimura distance of a TE, first a multiple sequence alignments of all members of the TE's family is carried out and the most abundant base at each position in the alignment is stored in a consensus sequence. Subsequently, the distance of each individual TE to its family consensus sequence can be calculated. The Distance serves to approximate the time that has passed since the individual transposition event. It is assumed that, the more recent a transposition took place, the fewer mutations are accumulated in an individual TE locus, *e.g.*, the smaller the Kimura distance. Ancient TEs spent a long time in the genome, so they accumulated more mutations, resulting in greater Kimura distances. According to this rationale, the distance can also be an indicator for the probability a TE is still mobile in the genome and was not inactivated by sequence alterations over time.

1.2.3. TEs as architects of genome structure and instructors of gene regulation

From an evolutionary perspective, TEs may be an important resource contributing to transcript diversity [58-60]. TEs can donate coding sequences they acquired and these can be used by their host (*i.e.*, domestication) [61, 62]. For example, the envelope gene of human endogenous retrovirus (HERV) subfamily W has been adopted and evolved into the gene *Syncytin* involved in the human placental morphogenesis [63]. Likewise, it has been demonstrated that the knockout of the TE-derived gene *Peg10* leads to early embryonic lethality in mice [64]. TEs can take parts of the genome along on their journey, resulting in exon shuffling that can allow genes to acquire new functions [65, 66].

Additionally, TEs can transduce host mRNA or facilitate chromosomal rearrangements leading to gene duplications [60, 67, 68]. Intriguingly, also somatic transposition events of TEs have been reported, which lead to the genomic mosaicism in neuronal cells [69-73]. It is proposed that between 0.6 [69] and 13.7 [72] new L1 insertions exist per neuron in human. Indeed, it has been reported that early-life experiences drive the expression of TEs in the brains of mice, for example, the lack of maternal care was shown to lead to an increased L1 activity in pubs [74]. Importantly, the TE activity experienced during development appears to be a determining factor for the TE activity in adult brains as well [75].

Moreover, TEs are loaded with TFBS [37] and thus potentially influence their genomic environment in *cis* by recruiting TFs or acting as alternative promoters for nearby genes [76, 77]. Domestication of TE-derived regulatory sequences at appropriate loci allows rewiring of gene regulatory pathways, which may be advantageous for species adaptation to environmental changes [78]. For example, CTCF is an important protein shaping the 3D genomic landscape, which is essential for gene regulation [79]. TEs are important resource that helped distributing CTCF binding sites throughout the mammalian genome [80].

A tight regulation of gene expression is essential for cell homeostasis and cell identity. In recent years, it became apparent that non-coding RNAs (ncRNAs) also contribute to this regulatory orchestra [81-84]. In this context, TEs contribute to the repertoire of ncRNAs [85, 86] and thus provide regulatory units that can impact gene expression in *trans* [87, 88]. The identification of species-specific TE-derived regulatory transcripts indicates a high regulatory innovation by TEs [18]. For example, transcripts from the B2 family in mice keep stress response genes (SRGs) in a poised state [22]. The induction of stress signals leads to a degradation of these transcripts that turns on the SRGs and enables a quick reaction of the cell. Enhancer RNAs represent another regulatory entity that acts on the 3D structure of the chromatin and intensifies the expression of the enhancer-associated genes. Indeed, multiple TE families show an enrichment of signals that are typical for enhancers [89]. For example, the HERV-H is a long noncoding RNA with an enhancer functionality that is essential for human embryonic cell identity [85, 90].

In summary, recent literature has demonstrated the capability of TEs to affect the genome and its regulation, *e.g.*, by rewiring gene regulation networks [91]. Theoretically, shuttling (copies of) coding or regulatory sequences thru the genome would have the advantage that functional entities do not need to evolve at multiple positions in the genome independently. Thus, it is supposed that TEs are an important resource to quickly adapt on environmental changes [47, 60], so that TEs could have been responsible for an accelerated genome evolution [21, 92, 93].

1.2.4. TEs expression during aging

While not entirely straightforward, aging may be defined as a progressive loss of molecular functions combined with decreased fertility and increased mortality [94]. The global improvement of life expectancy [95, 96] entails an uptick of age-associated diseases like cancer or neurological disorders [97]. Thus, our understanding of the molecular foundation of aging processes is critical for the prevention and treatment of age-related diseases. One of these mechanisms is the age-associated activation of TEs.

So far, multiple safeguards have been described that protect the cell against the activation of TEs. In addition to the epigenetic repression of TE regions via heterochromatinization or DNA methylation [98, 99], also post-transcriptional mechanisms involving short interfering RNAs and PIWI (P-element induced wimpy testis)-interacting RNA contribute to TE silencing by initiating their degradation [13]. Indeed, compromised small RNA and RNA interference pathways have been shown to substantially increase the TE content in the fruit fly *Drosophila melanogaster* [100].

It is believed that a gradual loss of repression during aging leads to the expression of TEs [101]. Once exported to the cytoplasm, cellular sensing mechanisms recognizing TE-RNA (*e.g.*, the retinoic acid-inducible gene I and the melanoma differentiation-associated gene 5) or the reverse-transcribed TE-DNA (*e.g.*, cyclic GMP-AMP synthase [cGAS] and absent in melanoma 2) trigger a sterile inflammation [13]. Thus, the expression itself, without reverse transcription, may already contribute to age-associated inflammation, also known as, “inflammaging” [102]. Indeed, the de-repression of LINEs in aging mice has been shown to result in an accumulation of TE cDNA copies in the cytosol, which triggers the cGAS DNA sensing pathway and leads to inflammation [103].

Although the overwhelming majority of TEs is affected by truncations [104, 105] and mutations [56, 106] leading to their inactivity, even the human genome still contains some fully functional mobile TEs [11, 106-108]. In principle, the activation of such elements could be detrimental for genome integrity [109, 110] and could afflict damages to the coding sequences of genes [11] or their regulation [10]. However, while L1 elements, for instance, are strongly expressed in many cancer types, there is little evidence that insertions of these elements actually contribute to this disease [40].

In summary, the transcription and genomic insertion of TEs has been described to be a characteristic hallmark of aging and age-related diseases. Given the large number of TEs and the wealth of proposed control mechanisms [111], however, it is unclear which TE is controlled by which molecular safeguards. Thus, to obtain a comprehensive picture on TE (dys-)regulation during aging as well as under healthy or diseased conditions it is essential to develop locus-specific analysis strategies that integrate many levels of genomic, transcriptomic, and epigenomic data.

1.2.5. Quantification of Transposable element expression

The first step towards a detailed understanding of TE regulation and expression is the development of suitable tools to assess the (differential) expression of TEs, *e.g.*, during aging or comparing healthy and diseased samples.

Transcription, *i.e.*, the production of RNA molecules based on a given DNA template, refers to a complex process involving many proteins generating both coding mRNA as well as ncRNA. LTR-retrotransposons, ERVs, and LINEs are typically transcribed by Pol II and frequently polyadenylated while SINEs are often transcribed by Pol III [87] and thus lack a 5' m7G-cap structure [112]. Thus, the TE-derived RNA molecules are detectable – in principle – by several transcriptome sequencing technologies. For this, the RNA is isolated, broken down into small pieces (*i.e.*, fragments), and reversely transcribed into cDNA. These fragments are amplified and read by the sequencing machine. These reads can then be mapped back to the reference genome for quantification. Nevertheless, the quantification of transcripts from TEs is still a challenging task for bioinformatics.

The repeated presence of TE copies in a genome can lead to reads that map to multiple loci in the reference genome with equal scores (multi-mapped reads) [50]. Inappropriate handling of multi-mapping reads can lead to wrong biological conclusions [113, 114]. Increasing the read length is one possibility to decrease the number of multi-mapping reads [115], as the likelihood of a read being unique grows with its length. To partially remedy this problem, current short-read sequencing technologies like Illumina enable the sequencing of paired-end reads. Here, a fragment of several hundred base pairs is sequenced from both ends in parallel. Aligning both paired-end reads under the constraint that the sequences are mapping in close vicinity to each other helps to reduce the multi-mapping problem [115]. It has been shown that longer reads as well as paired-end reads improve the assignability of reads that originate from TEs [116]. However, the analysis of evolutionarily young TEs, *i.e.*, elements exhibiting high sequence similarities (*i.e.*, small Kimura distance), may still be substantially obfuscated by multi-mapping reads. Analyses that simply discard multi-mapping reads bear the risk to underestimate TE expression and therefore miss the expression signal for entire families. Thus, TE quantification analyses based on unique reads are not recommended [117, 118]. In recent years, sophisticated algorithms to assign multi-mapping reads were developed and implemented in different computer programs, *e.g.*, TEtranscripts [118], TEtools [119], SalmonTE [120], SQuIRE [121], and Telescope [122] (Figure 2). In the following I will briefly explain the assignment concepts used by these tools.

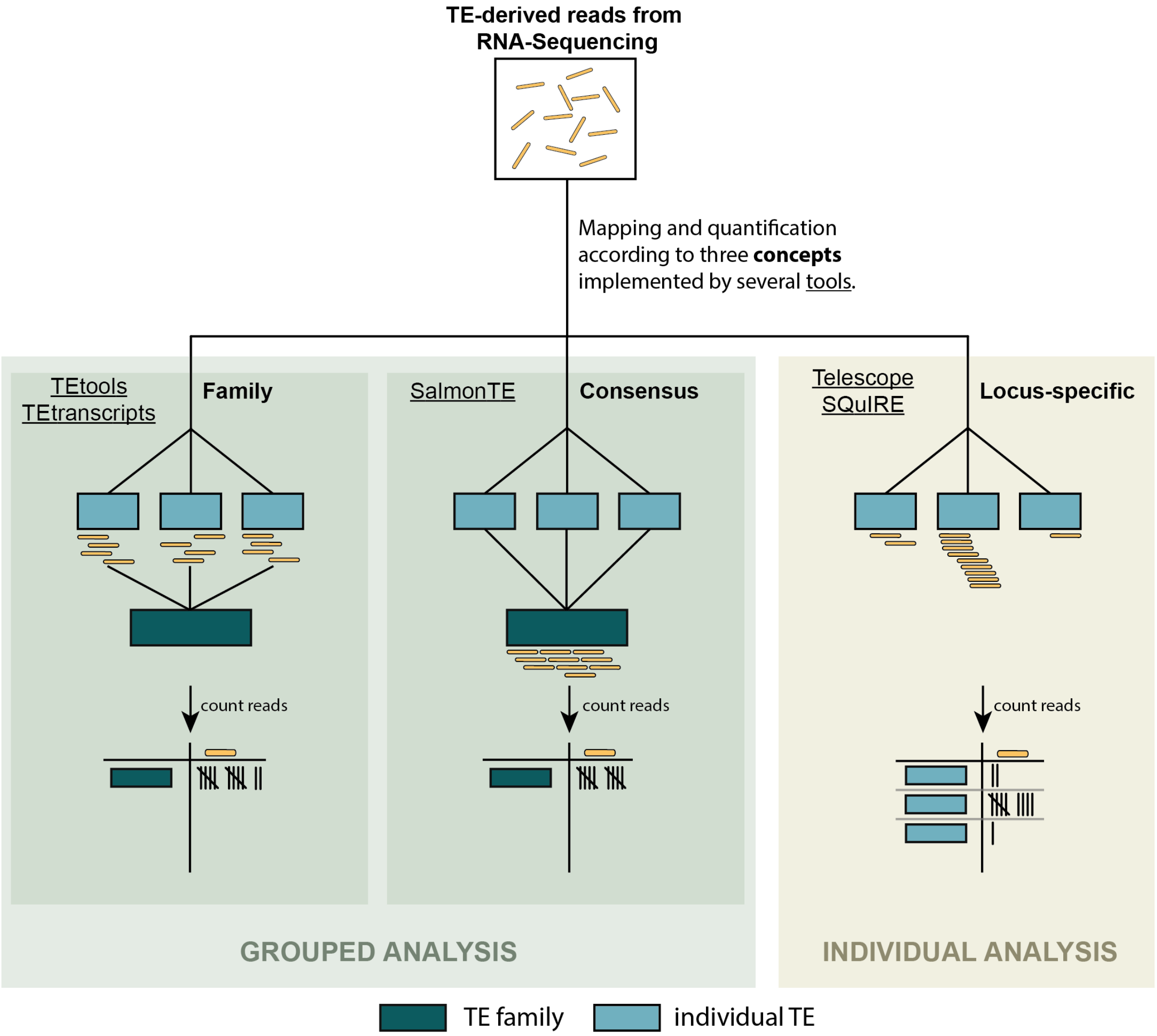


Figure 2 Assignment concepts of RNA-Seq derived reads. Reads derived from TEs can be analyzed to quantify TE expression in different ways. In grouped analysis, reads are either assigned to individual TEs (light blue) and subsequently all reads belonging to TEs of the same family (dark blue) are aggregated (TEtools, TEtranscripts), or a consensus sequence per TE family (dark blue) is first calculated and the reads are mapped to consensus sequence representing a TE family (SalmonTE). Individual analysis avoids aggregation steps and provides read counts for individual TE instance (light blue; Telescope, SQuIRE). The individual analysis has the advantage of obtaining coordinates of expressed individual TEs, which is lost in grouped analysis. **Abbreviations:** RNA – ribonucleic acid; TE – transposable element.

TEtranscripts has the advantage to quantify gene and TE expression in one run and is equipped with two modes. The first mode considers only unique reads while the second also handles multi-mapped reads. TEtranscripts quantifies the TE expression at the family level by estimating a combined abundance per TE family. In the “multi-mode”, all multi-mapped reads are weighted by the number of loci they were mapped to and subsequently assigned by an expectation-maximization (EM) algorithm. The EM-algorithm alternates between two steps, *e.g.*, the E- and M-step. The E-step calculates the fractional distribution of multi-mapped reads, which means the likelihood that the read comes from a certain TE instance. Different parameters can be considered for the likelihood calculation such as relative length of the TE, strand orientation, and the starting point of a read [120]. This relative abundance of reads is used in the M-step to update the relative abundance of each TE. These two steps are repeated for a certain number of iterations or until the program converges [118]. SalmonTE is based on salmon [123], an alignment-free mapper, which works with TE consensus sequences, hence, it produces family based counts. This tool does also use an EM algorithm for the assignment of multi mapped reads.

In contrast, TEtools considers the genomic sequences of individual TE loci and randomly assigns multi-mapping reads during the mapping procedure. Afterwards, read counts of all TE loci belonging to the same TE family are aggregated. However, TEtools exclusively uses TE sequences in their default reference, so that reads originating from genes for example can be miss assigned to TEs, which leads to an overestimation of the TE expression [117].

These three tools (TEtranscripts, TEtools, and SalmonTE) share a common drawback, namely the TE quantification at the family level. Consequently, the loci where transcripts originated from cannot be resolved and, thus, expression dynamics of individual TE loci remain in the dark. To bridge this gap, two additional tools, SQuIRE and Telescope, became available. Both tools aim to provide a TE locus-specific resolution and use an EM-algorithm for the assignment of multi-mapping reads. Other tools and strategies based on similar concepts are reviewed in [124].

1.3. Databases providing comprehensive expression profiles

In the past two decades, high-throughput sequencing technologies led to an accumulation of gene expression profiling data sets that cover thousands of genes simultaneously. Differential expression profiling data sets provide information on gene expression changes under certain conditions, *e.g.*, aging, disease, or treatment. Following increasingly improved guidelines of good scientific practice, the raw sequencing data of such studies is available through databases, such as Gene Expression Omnibus data base maintained at the National Center for Biotechnological Information (NCBI) [125], which enables researches to re-use the data.

Usually, sequencing data sets are created for specific scientific questions in a particular context where specific subsets of genes are studied in depth, which has the consequence that there is a large number of genes usually remains out of focus. However, such data sets provide an important resource to verify and compare expression profiles of genes of interest various conditions. This is a powerful method that offers the opportunity to derive new hypothesis or conclusions that are supported by multiple data sets, which is also known as meta-analysis [126-128]. However, leveraging large numbers of publicly available data sets can still be a huge effort for individual scientists to dig through tables from individual studies to obtain expression data on their genes of interest. In addition, data from different studies may have been analyzed differently, so that direct comparisons are limited. To accelerate discovery in the scientific communities, databases are generated through the collection and re-analysis of raw data. Such databases, for instance, can provide web interfaces for scientists to easily and quickly browse gene expression information across multiple data sets. For example, the expression atlas from the European Bioinformatics Institute provides comprehensive information about gene expression from thousands of studies [129]. Another example is the database from Mouse Genome Informatics [130], which provides mouse-specific expression data.

Given that TEs have been considered “junk” DNA for long, their study was often neglected [131]. In addition, the tardy development of appropriate methods to investigate TE expression likely contributed to the lack of standardized TE quantification pipelines and databases for TE expression. Since the burgeoning recognition that TEs have biological relevant functions, attempts have been made to combine the quantification of genes and TEs [118], which remains an ongoing development. Most of RNA-Seq data sets that are publicly available have not been investigated with respect to TE expression and provide a promising treasure for TE expression profiling on a large scale.

Given that databases on TE expression in different species, tissues, and specific conditions are still missing, it is difficult to obtain a comprehensive overview of TE expression. The emerging trend towards locus-specific expression analysis provides an opportunity to provide a powerful resource to the research community through the generation of a database that contains information about the expression of individual TE instances and their associated genes across species, tissues, and conditions.

1.4. Thesis focus and aim

TEs are ubiquitous in essentially all eukaryotic genomes, and their expression can be quantified with modern sequencing technologies. However, the technical limitation of read length and sequence similarities of TE copies poses a major challenge to assigning reads derived from TEs. Thus, most of our current knowledge about TE regulation comes from family-based analyses that lack information on the individual TE loci that express TE transcripts. Certainly, the biological functions of TEs in various contexts, *e.g.*, aging, cancer, or brain development, require a deeper understanding of their (dys)-regulation. Therefore, the localization of actively transcribed TEs is an important step to obtain insights into their regulation and biological roles. The aim of this thesis was to perform a genome-wide, locus-specific quantification of TE expression in an age-dependent setting. The specific points addressed in this thesis were on the following:

I. Evaluation of currently available TE quantification tools according to their performance with respect to the locus-specific quantification of TE expression.
II. Realization of a locus-specific TE quantification in different tissues (blood, brain, skin) of mice of different ages (6 and 24 months).
III. Development of a database structure with a web interface that can host expression data for an easy access by any scientists.

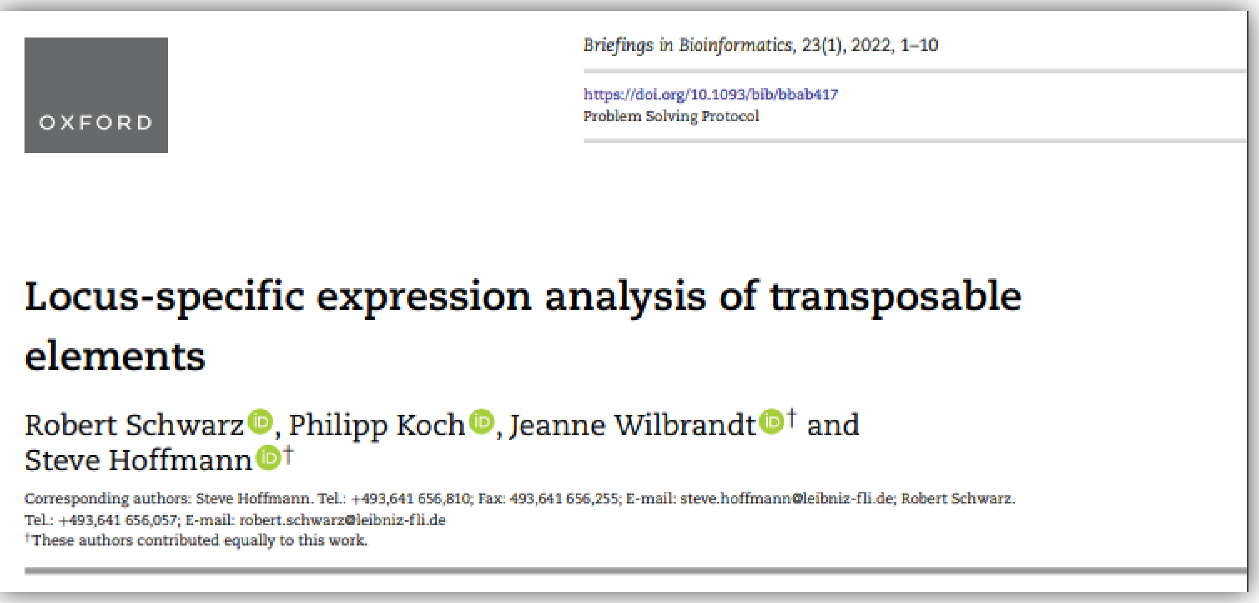
Manuscripts

2.1. Overview of Manuscripts

No.	Manuscript	Status; IF*
M1	**Locus-specific expression analysis of transposable elements** Schwarz R., Koch P., Wilbrandt J., Hoffmann S. Briefings in Bioinformatics, Volume 23, Issue 1, January 2022 This manuscript provides a comprehensive evaluation of TE quantification software in terms of their performance in quantifying locus-specific TE expression. For this, I manipulated TE family quantification tools to provide locus-specific expression information and evaluate their performances along with tools that already address the challenge of multi-mapping reads in a locus-specific manner. That manuscript concludes that the locus-specific expression analysis is sufficiently possible with currently available sequencing technologies and quantification tools.	published; 13.99
M2	**Expression differences of transposable elements during aging affect major tissue-specific pathways** Schwarz R., Koch P., Förste S., Groth M., Wilbrandt J., Fischer M., Hoffmann S. In this study I performed a comprehensive locus-specific TE quantification analysis in different tissues (brain, blood, and skin) of six and 24 months old mice. Beyond pervious studies, I indicate that TEs can be also down-regulated during aging and identify a set of TEs that are regulated by their own promoter. In addition, a co-regulation between TEs and their host genes was indicated that show highly tissue-specific expression patterns. Furthermore, a TFBS analysis suggests the involvement of Sox TFs in the regulation of independently expressed TEs. Overall, that study provides an interesting set of TEs that represents a striking starting point du investigate the relevance of TEs during aging.	submitted; -
M3	**TargetGeneReg 2.0: a comprehensive web-atlas for p53, p63, and cell cycle-dependent gene regulation** Fischer M., Schwarz R., Riege K., Decaprio J., Hoffmann S. NAR Cancer. 2022 Mar; 4(1): zcac009. This study provides a comprehensive web-atlas for p53, p63 and cell cycle dependent gene regulation created by analyzing datasets from multiple studies. In this project, I built a suitable data structure that allows both storage of the complex data and convenient accessibility. In addition, I designed and developed a website to make the data available to the public.	published; -

* IF - Impact factor

2.2. Manuscript 1 (M1) – Locus-specific expression analysis of transposable elements



OXFORD

Briefings in Bioinformatics, 23(1), 2022, 1–10

https://doi.org/10.1093/bib/bbab417
Problem Solving Protocol

Locus-specific expression analysis of transposable elements

Robert Schwarz, Philipp Koch, Jeanne Wilbrandt† and
Steve Hoffmann†

Corresponding authors: Steve Hoffmann. Tel.: +493,641 656,810; Fax: 493,641 656,255; E-mail: steve.hoffmann@leibniz-fli.de; Robert Schwarz.
Tel.: +493,641 656,057; E-mail: robert.schwarz@leibniz-fli.de
†These authors contributed equally to this work.

Summary:

The transcripts of TEs are part of the transcriptome, which can theoretically be measured with modern sequence technology, but is hampered by their repeated occurrence in the genome. This manuscript provides a comprehensive evaluation of TE quantification software in terms of their performance in quantifying locus-specific TE expression. For this, I manipulated TE family quantification tools to provide locus-specific expression information and evaluate their performances along with tools that already address the challenge of multi-mapping reads in a locus-specific manner. That manuscript concludes that the locus-specific expression analysis is sufficiently possible with currently available sequencing technologies and quantification tools.

Overview:

Manuscript No. 1

Manuscript title: Locus-specific expression analysis of transposable elements

Authors: Robert Schwarz, Philipp Koch, Jeanne Wilbrandt, Steve Hoffmann

Bibliographic information:

Robert Schwarz, Philipp Koch, Jeanne Wilbrandt, Steve Hoffmann, Locus-specific expression analysis of transposable elements, *Briefings in Bioinformatics*, Volume 23, Issue 1, January 2022, bbab417, https://doi.org/10.1093/bib/bbab417.

The candidate is:

☒ First author, ☐ Co-first author, ☐ Corresponding author, ☐ Co-author.

Status: published

Authors' contributions (in %) to the given categories of the publication:

Author	**Conceptual**	**Data analysis**	**Experimental**	**Writing the manuscript**
Schwarz R.	40%	100%	100%	50%
Koch P.	30%			20%
Wilbrandt J.				20%
Hoffmann S.	30%			10%
Total:	100%	100%	100%	100%

Briefings in Bioinformatics, 23(1), 2022, 1–10

https://doi.org/10.1093/bib/bbab417
Problem Solving Protocol

Locus-specific expression analysis of transposable elements

Robert Schwarz, Philipp Koch, Jeanne Wilbrandt[†] and Steve Hoffmann[†]

Corresponding authors: Steve Hoffmann. Tel.: +493,641 656,810; Fax: 493,641 656,255; E-mail: steve.hoffmann@leibniz-fli.de; Robert Schwarz. Tel.: +493,641 656,057; E-mail: robert.schwarz@leibniz-fli.de
[†]These authors contributed equally to this work.

Abstract

Transposable elements (TEs) have been associated with many, frequently detrimental, biological roles. Consequently, the regulations of TEs, e.g. via DNA-methylation and histone modifications, are considered critical for maintaining genomic integrity and other functions. Still, the high-throughput study of TEs is usually limited to the family or consensus-sequence level because of alignment problems prompted by high-sequence similarities and short read lengths. To entirely comprehend the effects and reasons of TE expression, however, it is necessary to assess the TE expression at the level of individual instances. Our simulation study demonstrates that sequence similarities and short read lengths do not rule out the accurate assessment of (differential) expression of TEs at the instance-level. With only slight modifications to existing methods, TE expression analysis works surprisingly well for conventional paired-end sequencing data. We find that SalmonTE and Telescope can accurately tally a considerable amount of TE instances, allowing for differential expression recovery in model and non-model organisms.

Key words: RNA sequencing; transposable elements; tool comparison; simulation; differential expression analysis

Introduction

The expression of transposable elements (TEs) has been repeatedly associated with various disorders including neurodegenerative [1, 2] and age-dependent diseases [3] or cancer [4, 5]. From an evolutionary perspective, however, expressed and reinserted TEs may play an advantageous role for the development of new genes by limiting gene conversion [6]. Likewise, it is suggested that TEs contribute to the heterogeneity and complexity of the brain [7]. While the activity of individual TEs is influenced by epigenomic factors such as DNA-methylation in vertebrates [8], a detailed understanding of the regulatory mechanisms is still missing. The major difference between TEs and other genomic features such as exons or lncRNAs is their high repetitiveness. Specifically, TE families contain long stretches of sequence that occur multiple times across the genome. Consequently, read aligners often face the challenge to correctly align TE reads to their locus of origin; i.e. the locus where the transcript read by the sequencer originated from. To deal with this multi-mapping read problem specialized tools have been developed in the past years.

Robert Schwarz is a PhD candidate in Bioinformatics at Friedrich-Schiller-University Jena, Germany, and is based at the Leibniz Institute on Aging—Fritz Lipmann Institute (FLI). His research focuses on the characterization of transposable elements and its activity.
Philipp Koch received his doctorate from the University of Jena in 2016. Since then, he has been working within the Core Facility Life Science Computing at the FLI on service-oriented bioinformatics analyses and on the development of software tools for molecular genetics.
Jeanne Wilbrandt received her doctorate in Biology from the University of Bonn in 2018. Her research focused on genomics and bioinformatics approaches. Currently, she works as Data Steward within the Core Facility Life Science Computing at the FLI.
Steve Hoffmann is group leader at the FLI and professor for Computational Biology at the Friedrich Schiller University Jena. His main research interest is the development of methods to investigate the regulation of genome activity.
Submitted: 2 July 2021; **Received (in revised form):** 24 August 2021

© The Author(s) 2021. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Table 1. Overview of compared TE expression methods adapted from Lanciano et al. [31]. EM- Expectation maximization; TE- transposable element

Tool	Level	Used alignment tool	Multi-mapper handling	Used references	Ref.
SalmonTE	Family	Salmon	EM-Algorithm	Consensus of families	[13]
Telescope	Instance	Free Choice	EM-Algorithm	Reference genome	[12]
TEtranscripts	Family	Free Choice	EM-Algorithm	Reference genome	[15]
SQuIRE	Instance	STAR	EM-Algorithm	Reference genome	[11]
TEtools	Family	Bowtie/Bowtie2	Random assignment	TE pseudogenome	[14]

The first important step to investigate this critically understudied part of genome regulation is the accurate and precise measurement of the expression of individual TE copies (TE instances). In this study, we systematically compare methods with regard to their ability to detect and quantify the expression of individual TE instances from simulated high-throughput sequencing data of three species (*Mus musculus*, house mouse; *Homo sapiens*, human; and *Nothobranchius furzeri*, turquoise killifish). Our analysis of the vertebrate model-organisms mice and human is complemented by the short-lived killifish *N. furzeri*, as it is quickly becoming an important model organism in aging research [9]. With an estimated TE content of 42.1%, its genome contains a considerable amount of TEs [10] and could be an interesting organism to study the regulation of TEs during aging. In contrast to the other two reference genomes used here, the assembly still is in a comparably early phase. Thus it provides TE expression benchmarks for genomes of lower quality.

Major obstacles for TE detection and quantification are the technical read length limitation of most RNA sequencing (RNA-Seq) experiments and the high sequence similarity of TEs. Since most TEs are too long for many sequencers to be read at once or already underwent RNA processing prior to library construction, many reads are expected to map to multiple instances of a TE, i.e. a TE family. In addition, low quality genomes render the analysis of repetitive elements particularly hard, as TEs may be misplaced or absent in the reference. Therefore, the analysis of TE expression has frequently been restricted to TE families, which often means that a consensus sequence per family is calculated and used as a reference. Consequently, the detection and analysis of individual TEs with pathological or physiological relevance remains a critical challenge for the investigation of sizable parts of genomes across all kingdoms of life. Notably, family-level investigations are also obfuscated when family members are not coordinately up- or down-regulated. Only recently, tools such as SQuIRE [11] and Telescope [12] became available to tackle TE expression analysis on instance-level.

Here, we investigate to which extent existing methods implemented in SalmonTE [13], TEtools [14], TEtranscripts [15], SQuIRE and Telescope (see Table 1) can be used to quantify locus-specific TE expression. We simulated RNA-Seq data for *M. musculus*, *H. sapiens*, and the non-model organism *N. furzeri*, because as it allows benchmarking of tool performances. In contrast to real data, exact expression values and expression differences are known and thus serve as a gold standard in all evaluations. To this end, we modified the three methods originally designed for family-level analyses to obtain expression estimates for individual TEs. Using DESeq2 [16], a tool to estimate differential expression from count data of high-throughput sequencing reads, we additionally investigate the ability to recover differentially expressed TEs (DETEs) based on the tools' alignments. Furthermore, our analysis provides insights into the relation of Kimura distances [17] and the ability to investigate expression levels of individual TE orders as defined by RepeatMasker [18]. In summary, our study provides a comprehensive assessment of the possibilities of DETE detection. This is an important step towards a better understanding of mechanisms underlying TE regulation in health, disease and aging.

Methods

The workflow of the tool evaluation is shown in Figure 1 and all in-house scripts, used in the following section, can be found at GitHub (simulation, evaluation and scripts: https://github.com/Hoffmann-Lab/TEdetectionEvaluation). Additionally, all command line calls are listed in Supplemental File 6.

Generation of repeat reference library

We used the repeat annotation of RepeatMasker of *M. musculus* (mm10, based on Repeat Library 20140131 downloaded in January 2020, https://www.repeatmasker.org/genomes/mm10/RepeatMasker-rm405-db20140131/mm10.fa.align.gz), *H. sapiens* (hg38, based on Repeat Library 20140131 downloaded in January 2020, https://repeatmasker.org/genomes/hg38/RepeatMasker-rm405-db20140131/hg38.fa.align.gz) and *N. furzeri* (Nfu_20150522 downloaded in January 2020, https://nfingb.leibniz-fli.de/data/raw/notho4/Nfu_20150522.dispersed_repeats.Nf-RepLib.20141117.align.gz), along with the reference genome of *M. musculus* mm10 (v102 downloaded in January 2021 from ftp://ftp.ensembl.org/pub/release-102/fasta/mus_musculus/dna/), *H. sapiens* hg38 (v102 downloaded in January 2021 from ftp://ftp.ensembl.org/pub/release-102/fasta/homo_sapiens/dna/) and *N. furzeri* Nfu_20150522 (downloaded from https://nfingb.leibniz-fli.de/data/raw/notho4/Nfu_20150522.softmasked_genome.fa.gz) [10] to generate a reference sequence library of TEs in FASTA format for each organism. Specifically, coordinates of TEs from the RepeatMasker annotation were converted into BED format and used to generate a reference library of nucleotide sequences for each annotated TE by using bedtools getfasta (v2.29.2–41-g4ebba703) [19]. Genomic position, Kimura distance, strand and TE categories are tracked for each instance throughout the evaluation pipeline via unique TE identifiers (TE ids, in the format chr|start|end|TE-repclass|TE-family|TE-subfamily|score|KimuraDistance). All following steps are based on these generated reference libraries.

Simulation of short read RNA-Seq data

In this study, we consider single-end (50 and 100 bp read length) as well as paired-end (100 bp read length) sequencing experiments. For either experimental setup, two distinct sets with five replicates each are generated. Throughout this study, the first set is considered a control (Set 1), while the second set contains 5% uniformly randomly drawn DETEs (Set 2; 2.5% up- and down-regulated, respectively). As a basis for our simulation,

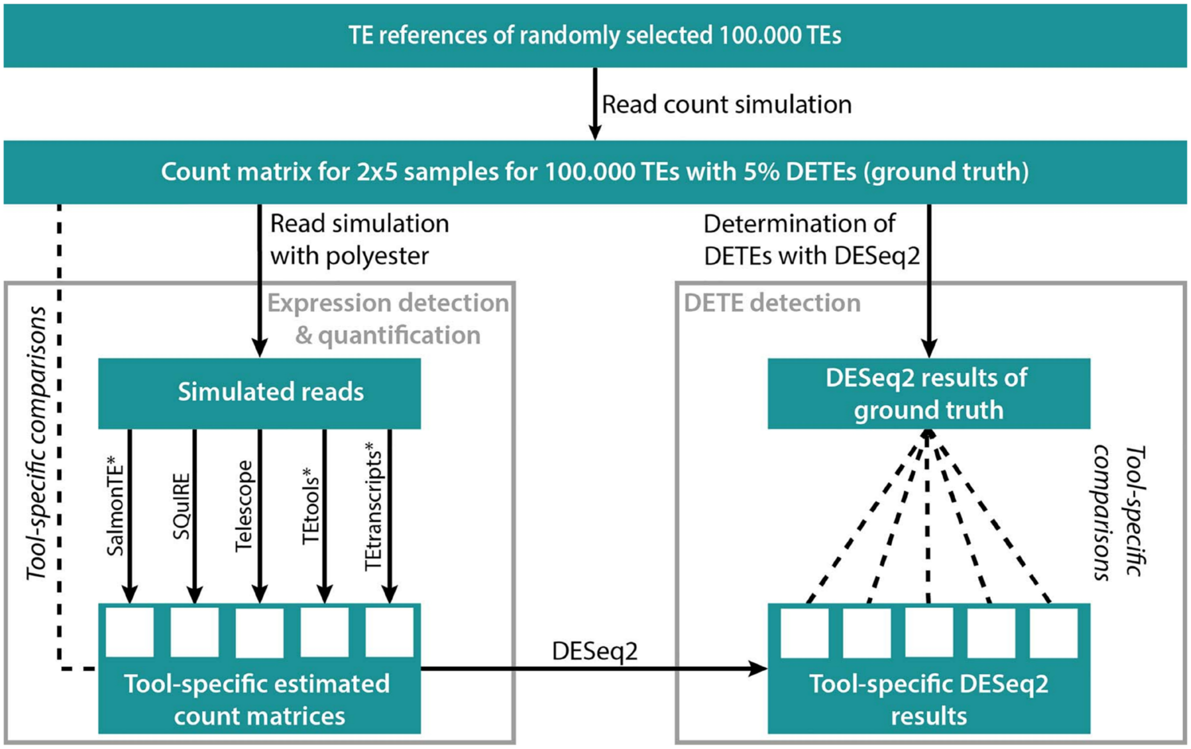

TE references of randomly selected 100.000 TEs
Read count simulation
Count matrix for 2x5 samples for 100.000 TEs with 5% DETEs (ground truth)
Read simulation with polyester
Determination of DETEs with DESeq2
Expression detection & quantification
DETE detection
Tool-specific comparisons
Simulated reads
DESeq2 results of ground truth
SalmonTE*
SQuIRE
Telescope
TEtools*
TEtranscripts*
Tool-specific comparisons
DESeq2
Tool-specific estimated count matrices
Tool-specific DESeq2 results

Figure 1. Workflow of tool evaluation. A count matrix for 100.000 randomly selected TEs was simulated, which was used to simulate reads with polyester. The tools SalmonTE*, SQuIRE, Telescope, TEtools* and TEtranscripts (* marks adapted tools) were applied to estimated counts per TE. The tool-specific estimated counts were compared with the ground truth (Expression detection & quantification). The ground truth of DETEs of the simulated TEs was determined with DESeq2 and compared to the tool-specific DESeq2 results (DETE detection). TE – Transposable element; DETEs – differentially expressed TEs.

we uniformly randomly drew 100,000 TEs, i.e. LINE, SINE, LTR or DNA elements, with at least 100 bp in length and a known Kimura distance from the reference library.

Polyester (v1.22.0) from the Bioconductor universe (v3.10) [20] was used to simulate RNA-Seq data in FASTQ format. It allows simulating GC-biases and sequencing errors based on Illumina sequencing error profiles that are provided with the polyester package. A mean read coverage of 20-fold per TE was simulated and the fragment length for the paired-end data was drawn from a Gaussian distribution with a mean of 250 bp (SD = 25 bp; default settings, see Supplemental methods). The number of simulated reads per TE and sample is stored in a count matrix, which serves as a reference in the evaluation process. This matrix was also used as input for DESeq2 (v1.26.0), to identify those TEs that can be detected as differentially expressed with a perfect read assignment.

An additional simulation was done for *M. musculus* using an in-house script implementing an alternative GC-bias unaware simulation strategy using quality values of real experiments to introduce sequencing errors (see Supplement methods).

Tool adaption, invocation and filtering of results

As described above, TEtools (v1.0.0), SalmonTE (v0.4) and TEtranscripts (v2.2.1) use different strategies to estimate TE expression at family-level (Table 1). We adapted the tools in order to evaluate their performance at the level of individual TEs instances and compare them with the dedicated instance-specific tools Telescope (v1.0.3) and SQuIRE (v0.9.9.92). As we did not change the algorithm of the tools, which are responsible for the assignment of multi-mapping reads, we do not expect an interfering of the outcomes.

By default, TEtools aligns reads to the instance-specific reference sequences and aggregates individual read counts afterwards to a family read count using a translation file. To suppress the aggregation step, we substituted the ids of TE families with ids of TE instances. Similarly, TEtranscripts uses an annotation file mapping TEs to their respective families. Again, we substituted the family names by TE ids to avoid the aggregation process. Since TEtranscripts and Telescope require precomputed alignments, simulated sequencing data was aligned with STAR (v2.7.6a) [21] according to the recommendation of TEtranscripts. Conversely, SalmonTE ships with an index for *M. musculus* and *H. sapiens* based on consensus sequences for each family. For our evaluation, we created an instance-specific index with Salmon (v0.9.1) [22] for each species instead, based on our repeat reference libraries. In the following, modified tools are referred to with an appended asterisk (*).

SQuIRE requires RepeatMasker's '.out' file format. To provide such a file, we translated the downloaded '.align' files into the '.out' format via an in-house script. This mapping is bijective, as the coordinates of each annotated TE are unique. From this, SQuIRE generates its own annotation file in BED format with SQuIRE-specific TE ids.

SQuIREs TE ids differ to ours, so that we cannot compare the results to the simulated counts by a simple merging process. However, both TE ids contain the genomic coordinates of the

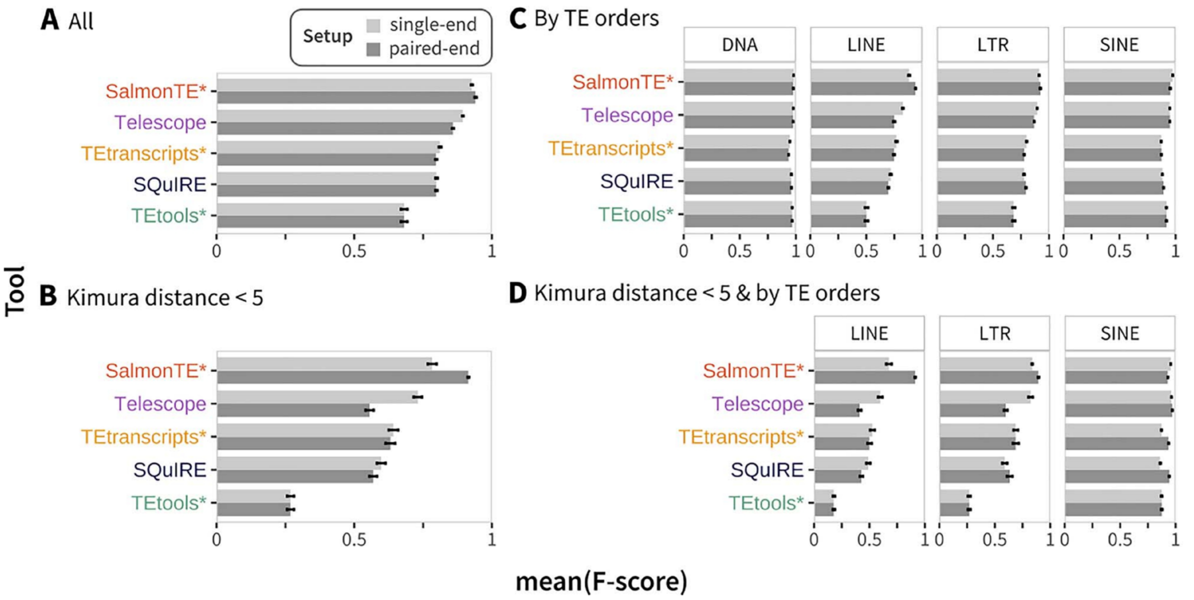


Figure 2. Comparison of TE expression detection in the *M. musculus* dataset. Mean F-scores were calculated across the ten replicates per tool and are given per setup (single-end 100 bp in light grey; paired-end 100 bp in dark grey) for (A) all TEs, (B) TEs with Kimura distance < 5, (C) orders of all TEs, and (D) orders of TEs with Kimura distance < 5. Note that in (D), DNA transposons are not shown, because no instance with a Kimura distance < 5 was present in the simulated data (see Supplemental File 1). TE — Transposable element.

respective TE. These coordinates are unique for each TE and allow to find the corresponding instances in both count tables; i.e. there is a one-to-one relationship of the entries in the count tables.

Except for the modifications described above, all tools were run with default settings. Subsequently, the outputs were parsed and aggregated across all samples with an in-house script to obtain instance-specific read count tables for each tool, which were used for all downstream comparisons. We removed all TEs with 10 or less reads summed up over all 10 samples. This cut-off was chosen as it translates to more than one read per TE and sample on average. Removing low count genes allows the mean–variance relationship in the data to be estimated with greater reliability and also reduces the number of statistical tests that need to be carried out in downstream analyses looking at differential expression [23].

Evaluation of the results

Expression detection and quantification

Throughout this study, a TE is considered to be detected by a given tool in a particular replicate if the reported read count is equal to or larger than five. This step is common praxis to eliminate noise produced by occasional misalignments of individual reads [24]. Using this binary measure we are able to categorize the results for each TE as true positive (TP), false positive (FP), and false negative (FN). Using these, the recall (sensitivity), precision, and F-score are calculated. Additionally, mean F-scores were separately calculated for TEs grouped by Kimura distances (binned with step sizes of 5) and by TE orders. Both distances and orders are given by the RepeatMasker annotation (Figure 2).

Furthermore, based on the count data generated by each tool we calculated the mean expression levels per TE i and Set

$$baseMean_i = \frac{\sum_1^j n_{ij}}{j}$$

,where n_{ij} (Read counts $\in K^{ixj}$) is the count of TE i in replicate j and compare them to the mean expression levels of the simulated TEs. Based on these base means, the coefficient of determination (r^2) was calculated from simulated and recovered read counts for true positives only (r^2(TP)). For visualization (Figure 3, Supplemental Figs 4–7), the logarithm of the mean expression levels was calculated and set to 0 if the original value was 0.

DETE detection

DETEs generated in our simulation may escape the detection by DESeq2 due to low expression, low expression fold changes and/or high dispersion. Additionally, DESeq2 might identify DETEs that were not simulated as such. To distill the set of DETEs that are detected by DESeq2 using the counts of an ideal aligner, we first ran DESeq2 on the simulated counts directly. Using the DESeq2 output, the subset of TEs detected as differentially expressed was used for further analysis as our ground truth. In this setup, a perfect aligner would have the power of 1. Subsequently, we ran DESeq2 with the count tables generated by all tools tested in this study. Afterwards, we evaluated the tools by comparing the output to the ground truth. The evaluation is based on true positive rates (TPRs) and the false discovery rates (FDRs) and is calculated as follows: (1) sort the DESeq2 result table in ascending order by the adjusted p-value; (2) count TPs and FPs in a cumulative manner; (3) use the cumulative values to calculate a TPR and FDR for each instance.

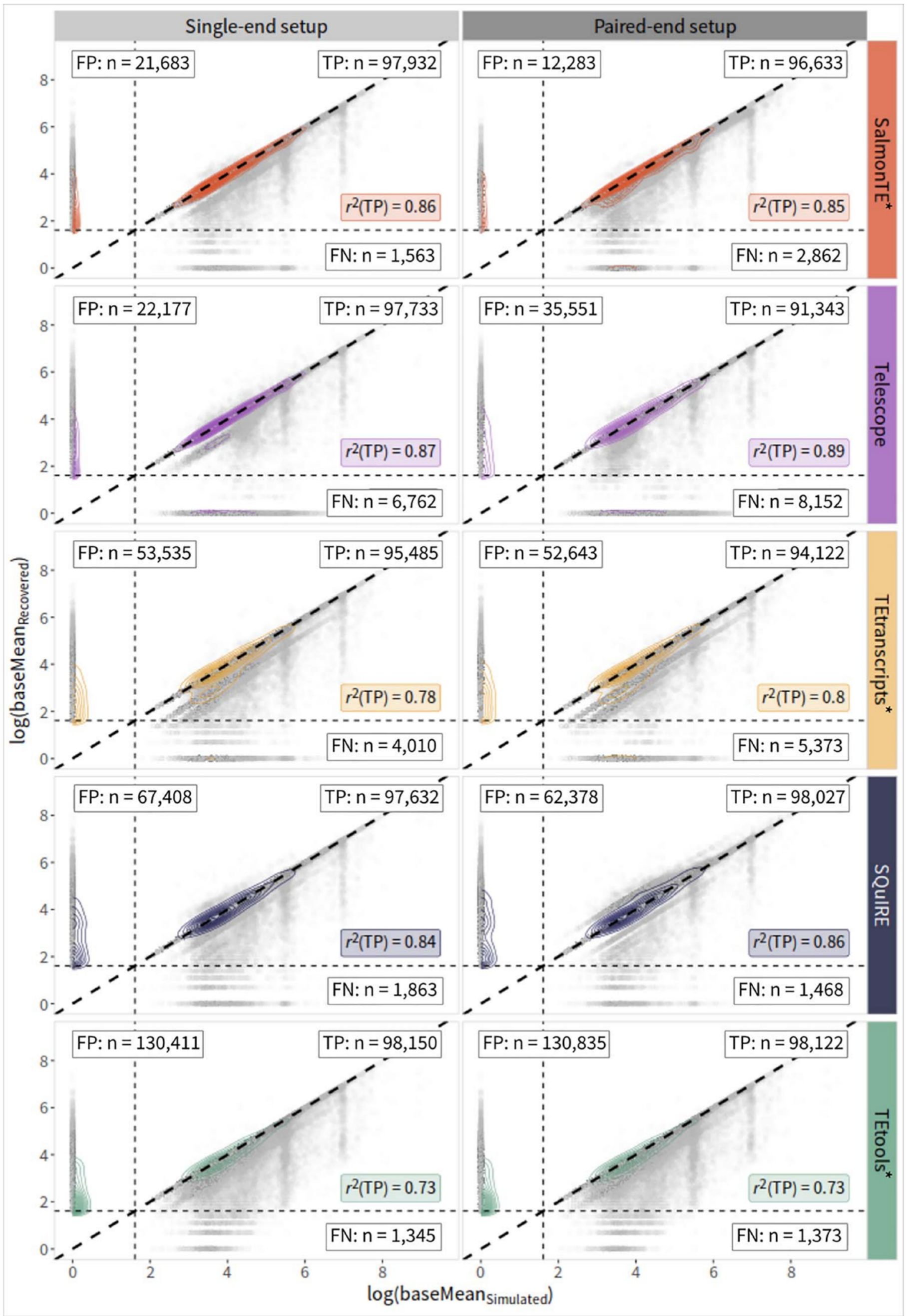


Figure 3. Comparison of recovered and simulated TE read counts of Set 2 of *M. musculus*. Scatter plots for simulated and recovered read counts for each tool (row) and sequencing setup (column; single-end: light-grey, paired-end: dark-grey). Dashed diagonal lines represent the perfect recovery (data points above: overestimation, points below: underestimation); dashed horizontal / vertical lines indicate the detection cut-offs to distinguish TP (upper right area), FP (upper left), and FN (lower right) at an expression value of 5. For each tool and setup, a coefficient of determination for TPs (r^2(TP)) is given (colored boxes) as well as counts of TEs considered as TP, FP, and FN (boxes in respective areas). TNs are here filtered out due to their high number. Note that data points lying on the horizontal dashed line are counted to the upper categories (TP or FP) and those on the vertical are counted either to FN or TP, due to usage of the R-package ggpmisc (v0.4.0) [25]. FN — false negative; FP — false positive; TE—transposable element; TP—true positive.

Ranking

The tools are ranked for each part of the evaluation (detection and quantification of TE expression, detection of differential TE expression), based on different categories within the evaluation parts (see Supplemental methods).

Results

The following results are based on the 100-bp polyester-based data sets, if not stated explicitly otherwise. The results of a complementary alternative simulation are shown in the Supplementary Material.

Simulation

After filtering for minimum read count (see Methods), 99 427 simulated expressed TEs were used for downstream analyses of *M. musculus*, 99 765 of *H. sapiens* and 99 235 of *N. furzeri* (Supplemental File 1). DESeq2 predicted 5 153 differentially expressed TEs in the *M. musculus* dataset (adjusted *p*-value threshold of 0.05), 5 174 in *H. sapiens* and 5 148 in *N. furzeri* when the simulated counts are used directly. These sets of DETEs are used as 'ground truth' of each species (see Methods).

Detection of TE expression

We first analyzed the tools' abilities to distinguish between truly expressed and silent TEs. Overall, similar observations can be made in *M. musculus* (Figure 2), *H. sapiens* and *N. furzeri* (Supplemental Figure 3). Across all species and sequencing setups, our results consistently indicate that tools using expectation maximization algorithms to assign multi-mapping reads perform better on average than TEtools*, which omits such a step. The overall improvement upon using paired-end data, as measured by the F-score (Supplemental File 2), appears to be surprisingly limited in all species when considering TEs across all Kimura distances (Figure 2A; Supplemental Figure 3A). With median F-scores from 0.93 (single-end) to 0.97 (paired-end) only SalmonTE* shows consistently improved F-scores across all species. In some cases, the single-end data delivers higher F-scores compared to paired-end data, e.g. Telescope for *M. musculus* (0.86 to 0.89, Figure 2A).

Consequently, the most substantial F-score increase comparing single-end (0.78) and paired-end (0.91) is observed for SalmonTE* for Kimura distances <5 in *M. musculus*. On the other hand, the F-score is significantly decreased for Telescope for the same set of TEs from 0.73 to 0.56 (Figure 2B). Tools using the STAR aligner (Telescope, TEtranscripts*, and SQuIRE) obtain higher F-scores for single-end than for paired-end data in *M. musculus*. However, in *H. sapiens* and *N. furzeri*, SQuIRE and TEtranscripts* show the expected improvement of F-scores using paired-ends for Kimura distances <5 (Supplemental Figure 3B).

Conversely, the length of single-end reads had a stronger impact. Compared to 50 bp single-end reads, the mean F-scores for the 100 bp single-ends improved from 0.8 to 0.82 across all tools in *M. musculus* (0.87 to 0.91 in *H. sapiens*, 0.76 to 0.82 in *N. furzeri*).

When considering F-scores for the four investigated TE classes (DNA, LINE, LTR and SINE) separately, best results are consistently obtained for DNA elements (Figure 2C, Supplemental Figure 3C). Despite its large number of DNA elements with a Kimura distance <5 ($n = 10\,916$), this is also true in *N. furzeri*. On the other hand, the lowest F-scores are observed for LINEs with a Kimura distance <5 in all species (Figure 2D; Supplemental Figure 3D, Supplemental File 2). Again, we also observe the strongest F-score increase for LINEs upon paired-end data usage for SalmonTE* in all species (from 0.67 to 0.91 in *M. musculus*, from 0.90 to 0.98 in *H. sapiens* and from 0.83 to 0.93 in *N. furzeri*).

The superior performance of SalmonTE* is also confirmed using the alternative simulation strategy. Importantly, the ranking of all tools is comparable using this alternative data, only SQuIRE and TEtranscripts* swap their ranks (see Supplemental File 5). Here, however, the tools appear to make slightly better use of paired-end information.

Quantification of TE expression

In terms of the tools' performances in quantifying TE expression, we evaluated the expression detection performance based on FP, TP, and FN counts, as well as r^2(TP), for single- and paired-end data. Results for *M. musculus* are shown in (Figure 3). Analogous data for the other species and simulations are shown in the Supplement (Supplementary Figs 4–7; Supplementary File 3). SalmonTE* and Telescope continuously show the lowest counts of FPs across all studied species and setups ranging from 3 028 in *H. sapiens* (SalmonTE) to 35 551 in *M. musculus* (Telescope). Surprisingly, in the case of Telescope, the numbers of FPs are consistently increase by using paired-end data. The differences between the tools are less pronounced regarding FNs. Here, TEtools* consistently yields the lowest count of FNs across all species and sequencing setups.

We observe a tendency of SalmonTE*, TEtranscripts* and TEtools* towards underestimating the TP counts. This is most pronounced in *N. furzeri* (Supplementary Figure 6) where almost half of simulated TEs (48%; median of all three tools) receive fewer reads than simulated while this is the case for only 24% of human TEs. Overestimation of TE expression appears to be most pronounced for TEs quantified with SQuIRE, which can be consistently observed in all species and sequencing setups.

Our analysis also indicates that the r^2(TP) values obtained with Telescope are the only ones consistently improving when paired-end data is used, while the other tools exhibit inconsistencies or, in the case of TEtools*, don't improve. The majority of the tools show slightly increased r^2(TP) for *M. musculus* and *N. furzeri*, and slightly decreased values in *H. sapiens*.

Differential TE expression

Subsequently, we evaluated the ability to detect expression changes with DESeq2 based on the tools' read count tables. For benchmarking, we used the FDRs and TPRs to analyze the DETE detection performances (Figure 4A and B; Supplemental Figure 8). Exact numbers for the recall are given in Supplemental File 4. In general, we observe that the ranking of the tools in this exercise is comparable for all genomes, sequencing- and simulation strategies.

At a fixed FDR of 0.1, SalmonTE* achieves the highest TPRs (0.81 to 0.99) across all data sets. With TPRs from 0.47 to 0.95, Telescope always takes the second rank. Both tools benefit from paired-end information. Conversely, TPRs across all data sets for TEtranscripts* (0.26 to 0.43) or TEtools* (0.24 to 0.61) are smaller and results do not substantially improve with paired-end reads. SQuIRE does not reach TPRs bigger than 0.05 for an FDR of 0.1 in all species and simulation strategies.

Overall performances are apparently impacted by the genome quality. Consequently, results in *H. sapiens* are generally better compared to *M. musculus* and *N. furzeri*. With the exception of individual performances for paired-end data, the results for *M. musculus* are generally better than for the non-model genome of *N. furzeri*. Especially Telescope shows a decline in performance when applied to the killifish transcriptome simulation. Again, for all species investigated here, SalmonTE* outperforms the other tools (cf. Supplemental Figure 8).

Calculating TPRs for TEs with a Kimura distance <5 (Figure 4B, Supplemental Figure 8), we observe that SalmonTE* and Telescope maintain their leading ranks. Again, paired-end data typically improves the results of both tools. The TPRs of SalmonTE* (0.73 to 0.97) and Telescope (0.70 to 0.98) indicate their overall suitability for the expression measurement of young elements in both, model and non-model organisms.

Given that SQuIRE ranks overall second in the quantification of TE expression, it is surprising that the tool shows a comparatively poor performance in the differential expression

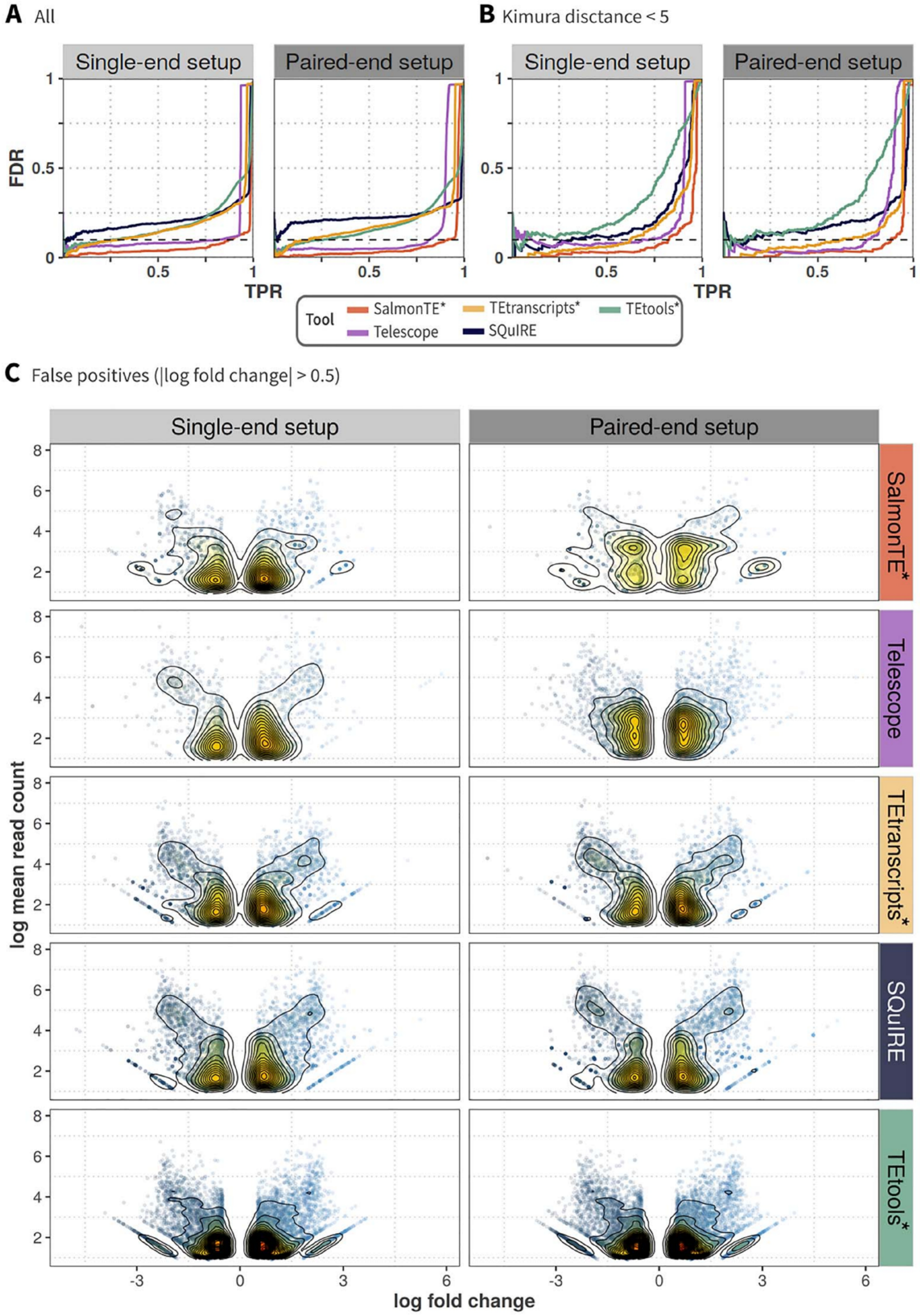


Figure 4. Comparisons of DETE detection performance and expression changes in all FPs of Set 2 versus Set 1 for single-end (left column in respective panel, light-grey) and paired-end setup (right column, dark-grey). (A, B) DETE detection performance (recovery of TEs simulated as differentially expressed in Set 2 compared to their expression in Set 1) is visualized as TPR in relation to FDR, shown per tool (lines) for (A) all detected TEs and (B) TEs with Kimura distance <5. The dashed horizontal lines represent a fixed FDR of 0.1. (C) Expression fold changes of FPs between Set 2 and 1 in contrast to mean read counts across all replicates for each tool (rows). Data points with a |log(fold change)| < 0.5 were removed for the sake of clarity. DETE – differentially expressed TE; FDR – false discovery rate; FP – false positive; TE transposable element; TPR – true positive rate.

exercise (TPR of 0.002 to 0.02). This result may be explained by the combination of a relatively high number of FPs and a stronger tendency for over-estimation of read counts (Figure 3; Supplemental File 3). To investigate the role of FPs in this phenomenon, we selected all TEs that were simultaneously wrongly detected in Set 1 and Set 2. This examination revealed populations of 1 571 and 1 518 TEs in the single- and paired-end setups, respectively, with comparably high read counts (mean count >20) and fold-changes |log(fold change)| >1, Figure 4C). Of these, 97% were in fact also wrongly detected as differentially expressed. Thus, we reason that the rather large number of FPs in combination with more pronounced mis-estimations of read counts could explain this result.

Discussion

While TEs have repeatedly been shown to play a role in pathological and physiological processes [3, 4, 26, 27], little is known about their expression patterns across different species, tissues and developmental stages. As the elevated expression of TEs has been observed during aging, a better understanding of molecular causes and consequences of TE dysregulation could, for instance, also yield new insights in age-related diseases and phenotypes [3, 28–30]. The lack of knowledge on TE regulation may be a consequence of a perceived lack of suitable methods to investigate the expression of repetitive regions of the genomes. Analyses on the level of TE families may only reveal transcriptional variation of single instances or sets of them if the changes are strong and consistent enough to compensate for contra-directional expression patterns of other family members. This may be exceptionally critical for families with multiple active instances. The most important shortcoming of family-level strategy, however, is the blindness regarding the precise genomic context in which the expression occurs. Since it is hard to imagine that all active instances of a TE family are governed by the same mechanisms or exert identical effects on cellular functions, it is critical to investigate TE expression at the level of single instances.

While achieving this goal is hampered by inherently high degrees of sequence similarity, technical, and, ultimately, financial limitations, our study explores to which extent the measurement of locus-specific TE expression is achievable with existing methods. Notably, three of the tools tested here are originally designed to work on the family-level only (SalmonTE, TEtranscripts and TEtools).

Detection of expression

The analysis of repetitive elements is critically obfuscated by multi-mapping reads and different strategies have been devised to assign these reads over the years [31]. Two of the methods tested here, implemented by Telescope and SalmonTE, involve read-generating models and maximum likelihood objectives for distributing multi-mapping reads to candidate loci. Of note, SalmonTE is based on Salmon and relies entirely on its quasi-read-mapping algorithm. Different solutions, also involving expectation maximization algorithms for the assignment of multi-mapping reads, are employed by SQuIRE and TEtranscripts. TEtools, also intended for use on the family-level only, omits such a step and assigns multi-mapping reads randomly to the TE pseudogenome (Table 1).

In general, we observe that the detection of expressed TEs works better with tools that employ expectation maximization steps, i.e. SalmonTE*, Telescope, TEtranscripts* and SQuIRE (Figure 2A). Telescope and TEtranscripts* work with pre-computed alignment files and recommended alignment parameters are the same for both tools. Even though Telescope and TEtranscripts* were thus called with the very same alignment files, their performances differed strongly. Thus, it is safe to assume that these differences are due to post-alignment calculations rather than the accurate assignment of reads to a genomic locus by the aligner. Apparently, SQuIRE's strategy to assign reads to multiple loci (Supplemental File 1) tends to increase the number of falsely detected expressed TEs. In turn, this has negative effects on the F-score statistics.

The analysis of repetitive genomic regions is substantially influenced by the amount of effective sequence information. Thus, paired-end setups should facilitate the detection of transcripts from many TEs [32]. In general, SalmonTE* is able to benefit the most from the additional sequence information in paired-end data. However, the degree to which individual tools take advantage of the additional sequence information varies strongly. Surprisingly, in the case of Telescope, paired-end data led to a drop of performance in detecting expressed TEs in all genomes and simulation strategies. This phenomenon might in part be explained by the tool's filtering strategies. By default, reads and read-pairs mapping to more than 100 possible loci are removed. In comparison with single-end, paired-end data typically reduces the number of multi-mappers such that fewer reads are removed by this filter [32]. Consequently, a higher number of read alignments are reported (shown by increased mapping rate, Supplemental File 1). On the flip side, the threshold might also substantially safeguard against misalignments and could explain the elevated number of FPs for paired-end data.

The Kimura distance [17] of a TE describes the sequence similarity to its family consensus sequence. Since sequence similarity plays a crucial role in tool performances, we evaluated the tools for distinct Kimura distances. As expected, we observe decreasing F-scores for elements with low Kimura distances (Figure 2B, Supplemental Figure 2), which can be mitigated by paired-end sequencing strategies. Naturally, this has consequences for exact measurement of elements from active families. Among young elements with a high sequence similarity (Kimura distance <5), LINE instances of *M. musculus* are especially difficult to track, as their similarity distribution is skewed to a Kimura distance of 0 (Supplemental Figure 1). In contrast to DNA transposons, families of LINE, SINE and LTR classes are still active in *M. musculus* [33]. The detection of young LINE instances appears to be more successful in *H. sapiens* and *N. furzeri*, since in these genomes the distribution of Kimura distances is not as strongly skewed to 0 indicating a reduced or less recent LINE activity (Supplemental Figure 1). On the other hand, all tools perform well for DNA transposons. In the case of *N. furzeri* this is a bit surprising, as this organism appears to have a very high number of young DNA transposons. Here, the cut-and-paste transposition mechanism of DNA transposons and rather small family sizes [34] appear to substantially ameliorate the multi-mapping read problem and its consequences.

While SalmonTE* came up as the top runner in most of our benchmarks, we noted some exceptions. Importantly, it did not recover the highest number of 'truly' expressed TEs (TP). This might be a drawback in all such cases where maximum sensitivity is of essence. Furthermore, SalmonTE* does not show the highest r^2 values for the count estimation of TP, as the underestimation of the counts is more pronounced compared to other tools (Supplementary File 3).

Quantification and detection of differential expression

In light of mounting evidence for the biological relevance of TEs in health and disease, we evaluated the applicability of the five methods for differential expression analysis. A critical factor for the reliable detection of differential expression is the accuracy of read count estimates. While the majority of the tools show a systematic bias, i.e. an underestimation, in single- and paired-end setups, paired data improves estimates on average (Figure 3, Supplemental Figures 4–7). This result can be expected as the

Table 2. Ranking of the tools concerning their performance of detection and quantification of TE expression and detection of differential expression. TE- transposable element

Tool	Detection of TE expression	Quantification of TE expression	Detection of differential TE expression
SalmonTE*	1	1	1
Telescope	2	4	2
SQuIRE	3	2	5
TEtools*	5	3	3
TEtranscripts*	4	5	4

number of unaligned or misaligned reads is reduced by additional paired-end information. Despite the fact that Telescope yields an increased number of FPs when paired-end data are used, it is able to substantially improve the read count estimates for truly expressed TEs, and shows the highest accuracy and precision (i.e. in *M. musculus*, Figure 3). The best performance in terms of detecting DETEs is observed for SalmonTE*.

On the flip side, SQuIRE's usability for the detection of DETEs appears to be limited by the assignment of reads to multiple loci (Supplemental File 1). Despite the second rank considering the quantification of TE expression (Table 2), a substantial number of FPs show such a high difference between Set 1 and Set 2 (Figure 4C) that they are called as DETEs. Consequently, the TPR for an FDR of 0.1 of SQuIRE lags behind the other evaluated approaches in this specific exercise.

Simulation

Simulations allow the systematic analysis of computational methods when the ground truth for actual data is unknown or difficult to obtain. On the flip side, simulated data cannot reflect reality in all its facets. For instance, unknown alternative transcription starts, termination sites, or post-transcriptional processes leading to RNA degradation lead to specific transcripts not covered by any annotation. Thus, simulations may not reach the level of complexity in real data. Also, it is essential to keep in mind that models and parameters accounting for phenomena such as GC-biases or sequencing errors are global approximations. However, for benchmarking alignment algorithms entirely relying on the sequence information of reference genomes and individual reads or read-pairs, such simulations provide indispensable insight into the tools' capabilities to deal with repetitive sequences.

Conclusion

Within the limits of our simulation study, a tool originally designed for family-level quantification, SalmonTE, emerges as the most convincing results. In addition to favorable results in detecting expressed TEs, SalmonTE* results enable a surprisingly high recall of differentially expressed TE transcripts. The general ranking of the tools regarding DETE detection (Table 2) based on TPRs for an FDR of 0.1 — SalmonTE* performing best, Telescope second, TEtools* third, followed by TEtranscripts* and SQuIRE — holds for all sequencing setups and studied species.

Arguably, the detection, quantification, and differential expression analysis of transcribed TEs remains one of the most challenging tasks in genome research. The misplacement or absence of instances from reference genomes, especially in the case of active TEs, insufficient read lengths, and high degrees of sequence similarity often restrain investigations of this biologically relevant class of RNA. Despite all technical difficulties, our analysis shows that an accurate and precise reference mapping of many individual TEs is already possible and encourages a more intensive research into this direction.

Key Points

- Accurate expression assessment of individual transposable elements is possible and can help to study their biological role more in detail, here demonstrated on simulated data.
- RNA-Seq protocols affect the detection of locus specific TE expression, however, even older protocols, e.g. single-end, are appropriate to get a comprehensive overview about individual TE expression.
- Detection of differentially expressed TE instances can be achieved with existing methods, partially with slight modifications.

Author contribution

All authors contributed to and approved the manuscript. S.H. and P.K. designed and supervised the study. S.H. and R.S. implemented the methods. R.S. carried out the analyses. R.S., P.K., J.W. and S.H. interpreted the data.

Data and materials availability

All in-house scripts, used in this study, can be found at GitHub (simulation, evaluation, and scripts: https://github.com/Hoffmann-Lab/TEdetectionEvaluation) and all command line calls are listed in Supplemental File 6. Data including references of transposable elements, simulated FASTQ files, and count tables of the tools are available at https://zenodo./record/4561751 [35]. The used FASTQ files for the alternative simulation are publicly available on GEO (single-end: GSM1944865, paired-end: GSM1716945) [36, 37].

Acknowledgements

We thank Martin Bens, Marco Groth, Konstantin Riege and Karol Szafranski for fruitful discussions and useful suggestions.

Funding

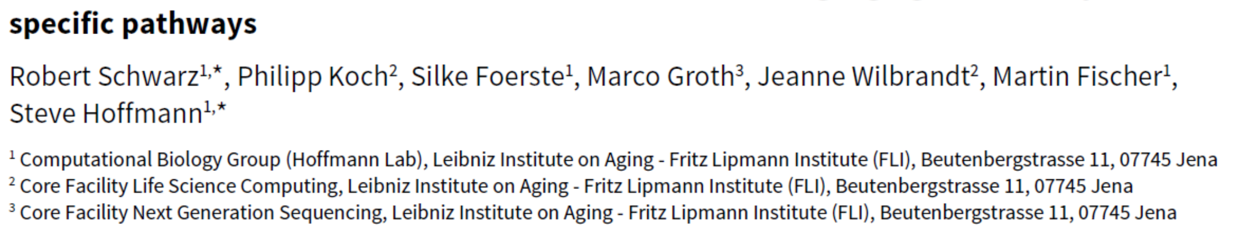
The work was supported by the Klaus Tschira Stiftung (project number: 03.136.2018) and the German Federal Ministry of Education and Research (de.NBI, 031L0106D). The

FLI is a member of the Leibniz Association and is financially supported by the Federal Government of Germany and the State of Thuringia.

References

1. Li W, Lee MH, Henderson L, *et al*. Human endogenous retrovirus-K contributes to motor neuron disease. *Sci Transl Med* 2015;**7**(307):307ra153.
2. Reilly MT, Faulkner GJ, Dubnau J, *et al*. The role of transposable elements in health and diseases of the central nervous system. *J Neurosci* 2013;**33**(45):17577–86.
3. De Cecco M, Ito T, Petrashen AP, *et al*. L1 drives IFN in senescent cells and promotes age-associated inflammation. *Nature* 2019;**566**(7742):73–8.
4. Scott EC, Gardner EJ, Masood A, *et al*. A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. *Genome Res* 2016;**26**(6):745–55.
5. Lock FE, Rebollo R, Miceli-Royer K, *et al*. Distinct isoform of FABP7 revealed by screening for retroelement-activated genes in diffuse large B-cell lymphoma. *Proc Natl Acad Sci U S A* 2014;**111**(34):E3534–43.
6. Kapusta A, Kronenberg Z, Lynch VJ, *et al*. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet* 2013;**9**(4):e1003470.
7. Erwin JA, Marchetto MC, Gage FH. Mobile DNA elements in the generation of diversity and complexity in the brain. *Nat Rev Neurosci* 2014;**15**(8):497–506.
8. Walsh CP, Chaillet JR, Bestor TH. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat Genet* 1998;**20**(2):116–7.
9. Kim Y, Nam HG, Valenzano DR. The short-lived African turquoise killifish: an emerging experimental model for ageing. *Dis Model Mech* 2016;**9**(2):115–29.
10. Reichwald K, Petzold A, Koch P, *et al*. Insights into sex chromosome evolution and aging from the genome of a short-lived fish. *Cell* 2015;**163**(6):1527–38.
11. Yang WR, Ardeljan D, Pacyna CN, *et al*. SQuIRE reveals locus-specific regulation of interspersed repeat expression. *Nucleic Acids Res* 2019;**47**(5):e27.
12. Bendall ML, de Mulder M, Iniguez LP, *et al*. Telescope: characterization of the retrotranscriptome by accurate estimation of transposable element expression. *PLoS Comput Biol* 2019;**15**(9):e1006453.
13. Jeong H-H, Yalamanchili HK, Guo C *et al*: An ultra-fast and scalable quantification pipeline for transposable elements from next generation sequencing data. 2018: 168–79.
14. Lerat E, Fablet M, Modolo L, *et al*. TEtools facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of piRNA genes. *Nucleic Acids Res* 2017;**45**(4):e17.
15. Jin Y, Tam OH, Paniagua E, *et al*. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* 2015;**31**(22):3593–9.
16. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**(12):550.
17. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 1980;**16**(2):111–20.
18. Smit AFA. *RHPG: RepeatMasker* http://repeatmasker.org. 1996.
19. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;**26**(6):841–2.
20. Huber W, Carey VJ, Gentleman R, *et al*. Orchestrating high-throughput genomic analysis with bioconductor. *Nat Methods* 2015;**12**(2):115–21.
21. Dobin A, Davis CA, Schlesinger F, *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**(1):15–21.
22. Patro R, Duggal G, Love MI, *et al*. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 2017;**14**(4):417–9.
23. Law CW, Alhamdoosh M, Su S, *et al*. RNA-seq analysis is easy as 1–2-3 with limma, Glimma and edgeR. *F1000Res* 2016;**5**.
24. Zhang C, Zhang B, Lin LL, *et al*. Evaluation and comparison of computational tools for RNA-seq isoform quantification. *BMC Genomics* 2017;**18**(1):583.
25. Aphalo PJ: *ggpmisc: Miscellaneous Extensions to 'ggplot2'* https://CRAN.R-project.org/package=ggpmisc. 2020.
26. Bourque G, Leong B, Vega VB, *et al*. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res* 2008;**18**(11):1752–62.
27. Bedrosian TA, Quayle C, Novaresi N, *et al*. Early life experience drives structural variation of neural genomes in mice. *Science* 2018;**359**(6382):1395–9.
28. Simon M, Van Meter M, Ablaeva J, *et al*. LINE1 Derepression in aged wild-type and SIRT6-deficient mice drives inflammation. *Cell Metab* 2019;**29**(4):871–85 e875.
29. Wood JG, Helfand SL. Chromatin structure and transposable elements in organismal aging. *Front Genet* 2013;**4**:274.
30. De Cecco M, Criscione SW, Peterson AL, *et al*. Transposable elements become active and mobile in the genomes of aging mammalian somatic tissues. *Aging (Albany NY)* 2013;**5**(12):867–83.
31. Lanciano S, Cristofari G. Measuring and interpreting transposable element expression. *Nat Rev Genet* 2020;**21**(12): 721–36.
32. Sexton CE, Han MV. Paired-end mappability of transposable elements in the human genome. *Mob DNA* 2019;**10**:29.
33. Huang CR, Burns KH, Boeke JD. Active transposition in genomes. *Annu Rev Genet* 2012;**46**:651–75.
34. Wicker T, Sabot F, Hua-Van A, *et al*. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* 2007;**8**(12):973–82.
35. Schwarz Robert. Individual TE expression detection [Data set]. Zenodo. 2021. https://doi.org/10.5281/zenodo.4561751.
36. Huhne R, Thalheim T, Suhnel J. AgeFactDB–the JenAge ageing factor database–towards data integration in ageing research. *Nucleic Acids Res* 2014;**42**(Database issue):D892–6.
37. Li X, Liu Y, Salz T, *et al*. Whole-genome analysis of the methylome and hydroxymethylome in normal and malignant lung and liver. *Genome Res* 2016;**26**(12):1730–41.

2.3. Manuscript 2 (M2) – Expression differences of transposable elements during aging affect major tissue-specific pathways



Expression differences of transposable elements during aging affect major tissue-specific pathways

Robert Schwarz[1,*], Philipp Koch[2], Silke Foerste[1], Marco Groth[3], Jeanne Wilbrandt[2], Martin Fischer[1], Steve Hoffmann[1,*]

[1] Computational Biology Group (Hoffmann Lab), Leibniz Institute on Aging - Fritz Lipmann Institute (FLI), Beutenbergstrasse 11, 07745 Jena
[2] Core Facility Life Science Computing, Leibniz Institute on Aging - Fritz Lipmann Institute (FLI), Beutenbergstrasse 11, 07745 Jena
[3] Core Facility Next Generation Sequencing, Leibniz Institute on Aging - Fritz Lipmann Institute (FLI), Beutenbergstrasse 11, 07745 Jena

Summary:

Transposable elements (TEs) are arguably the largest class of genetic material with an unclear biological role. In this study I performed a comprehensive locus-specific TE quantification analysis in different tissues (brain, blood, and skin) of six and 24 months old mice. Beyond pervious studies, I indicate that TEs can be also down-regulated during aging and identify a set of TEs that are regulated by their own promoter. In addition, a co-regulation between TEs and their host genes was indicated that show highly tissue-specific expression patterns. Furthermore, a TFBS analysis suggests the involvement of Sox TFs in the regulation of independently expressed TEs. Overall, that study provides an interesting set of TEs that represents a striking starting point du investigate the relevance of TEs during aging.

Overview:

Manuscript No. 2

Manuscript title: Expression differences of transposable elements during aging affect major tissue-specific pathways

Authors: Schwarz R., Koch P., Förste S., Groth M., Wilbrandt J., Fischer M., Hoffmann S.

Bibliographic information:

Robert Schwarz, Philipp Koch, Silke Förste , Marco Groth, Jeanne Wilbrandt , Martin Fischer, Steve Hoffmann, Expression differences of transposable elements during aging affect major tissue-specific pathways. (submitted)

The candidate is:

☒ First author, ☐ Co-first author, ☐ Corresponding author, ☐ Co-author.

Status: submitted

Authors' contributions (in %) to the given categories of the publication:

Author	**Conceptual**	**Data analysis**	**Experimental**	**Writing the manuscript**
Schwarz R.	50%	100%		45%
Förste S.			30%	
Groth M.			20%	
Fischer M.	10%			20%
Hoffmann S.	40%			30%
Other			50%	5 %
Total:	100%	100%	100%	100 %

Expression differences of transposable elements during aging affect major tissue-specific pathways

Robert Schwarz[1,]* (ORCID: 0000-0002-0654-0943), Philipp Koch[2] (ORCID: 0000-0003-2825-7943), Silke Förste[1], Marco Groth[3] (ORCID: 0000-0002-9199-8990), Jeanne Wilbrandt[2] (ORCID: 0000-0002-0363-3837), Martin Fischer[1] (ORCID: 0000-0002-3429-1876) and Steve Hoffmann[1,]* (ORCID: 0000-0002-5239-7201)

[1] Computational Biology Group (Hoffmann Lab), Leibniz Institute on Aging - Fritz Lipmann Institute (FLI), Beutenbergstrasse 11, 07745 Jena
[2] Core Facility Life Science Computing, Leibniz Institute on Aging - Fritz Lipmann Institute (FLI), Beutenbergstrasse 11, 07745 Jena
[3] Core Facility Next Generation Sequencing, Leibniz Institute on Aging - Fritz Lipmann Institute (FLI), Beutenbergstrasse 11, 07745 Jena

*To whom correspondence should be addressed; E-mail (Phone / Fax):

steve.hoffmann@leibniz-fli.de (+49 3641 656810 / +49 3641 656255),

robert.schwarz@leibniz-fli.de (+49 3641 656057 / -)

Abstract

Transposable elements (TEs) are arguably the largest class of genetic material with an unclear biological role. At the same time, it is increasingly appreciated that TEs play critical roles in various pathophysiological processes. Recent research suggests that the up-regulation of TEs is a characteristic of aging and could be a critical factor in the aging process. To investigate the aging dynamics of TE expression, we generated a transcription data set of mice (M. musculus) from three tissues (brain, blood, skin) using RNA-Seq and CAGE-Seq. This combination enabled the identification of independently expressed TEs with proper transcription start sites and putative TE promoters. Using a locus-specific analysis, we unexpectedly find that TEs are up- and down-regulated during aging to the same extent, challenging the narrative of an entirely detrimental role of TE expression. Strikingly, independently expressed TEs are substantially enriched in genes with highly tissue-specific functions such as synapse regulation in brain and cell-substrate junctions in skin. In the mouse brain, we identify highly tissue-specific genes such as the protocadherin-beta cluster to be affected by differential TE expression. Moreover, our data strongly suggest the involvement of Sox transcription factors in the regulation of TE expression. Our findings demonstrate the tissue-specific and age-dependent expression of individual TEs in mice that may be regulated by Sox transcription factors. These TEs are enriched in tissue-specific genes and show independent but strong co-regulation with their host genes. Thus, we provide a striking and consequential starting point to elucidate the full relevance of TEs during aging.

Introduction

Three-quarters of a century after Barbara McClintock's [1] groundbreaking discovery, our understanding of the biological roles of transposable elements (TEs) remains limited. TEs, colloquially called "jumping genes", either jump to a new position (DNA-transposons) or spread within their host genome via copy-paste mechanisms (retrotransposons) [2]. Such transposition events can have a critical impact on genome integrity and impair its functionality. Importantly, successful transposition events may also have substantial effects on genome regulation, as TEs harbor transcription factor binding sites potentially affecting gene expression in *cis* and *trans* [3, 4]. Given the high abundance of TE's in many genomes, it is quintessential to investigate the impact of TE accessibility and expression on cellular function [5, 6].

In the scientific literature, the expression of TEs is typically reflected in the context of deteriorating processes. For instance, the up-regulation of TEs has been associated with diseases like cancer [7–10], neurological disorders [11, 12], or aging [13, 14]. L1, a TE-superfamily within the TE class of long interspersed nuclear elements (LINEs) amounting to more than 20% of the human genome [2], have been of special interest because several of their members are still able to transpose in humans and mice. In this context, it has been shown that the escape of an L1 element from repression may result in a transposition event impairing the APC gene ultimately paving the way for colorectal cancer development [10].

Importantly, a successful transposition event is not required for having substantial effects on a cell [15]. For instance, TE-derived RNAs and DNAs alone can trigger the immune system via double-stranded

RNA and DNA detection mechanisms within the cytoplasm [16, 17]. Such immune responses to TE up-regulation have been reported for different malignomas [7] as well as in senescent cells [18]. In cancer, TE-triggered inflammation may even be a defense mechanism to suppress carcinogenesis and it has been suggested that down-regulation of TEs could protect some cancer tissues against the immune response [19].

Despite this disease-centric view of TEs, it is important to note that also healthy tissues show TE expression and transposition [7, 20]. For instance, the activity of TEs is regularly observed during brain development and is considered to be a major contributing factor to the mosaicism of the neuronal genome [12, 21–25]. Recently, a study on the effects of maternal care on the mouse brain established a link between the activity of TEs and psychosocial conditions [20]. Furthermore, it has been proposed that the expression of TEs during development also impacts TE expression in adult brains and may thus have long-term effects [26]. On the molecular level, gene regulatory functions of TEs have been suggested for B2 elements in brain, a TE superfamily that is part of the short interspersed nuclear element (SINE) TE class [3, 27]. Specifically, B2 elements might act in *trans* to keep the transcription machinery of stress response genes in a poised state [28]. Finally, TE activity triggered by environmental changes could enable somatic cells to overcome hurdles during lifetime [12]. The associations of TE activity in brain indicate tissue specificity and a tissue-specific accessibility to TEs has been demonstrated [29, 30]. Thus, it appears necessary to analyze data on multiple tissue types to obtain a comprehensive picture on the causes and effects of TE expression.

The repetitive nature of TEs renders systematic investigations of expression patterns, regulatory mechanisms, or potential functions challenging [31]. Much of our current understanding about TEs and their transcription is based on approaches that aggregate expression data on the level of TE superfamilies [7, 9, 13, 20] or are focused on specific TE elements [10, 32]. Aggregation approaches, however, easily miss the effects of individual elements (locus-specific) or subsets of elements. Here, we provide a locus-specific expression analysis to enable a more detailed characterization of TE expression and its biological consequences using a SalmonTE-based [33] analysis strategy that we described recently [34].

Results

Age- and tissue-specific TE expression based on RNA-Seq data

We performed 150 bp paired-end sequencing of rRNA-depleted RNA from blood, brain, and skin tissue samples of young (6 months) and old (24 months) male mice (*Mus musculus*; see Methods). Comparing the age-associated RNA-Seq expression data of TEs at the superfamily level, we observe comparably small mean $\log_2$ fold changes (L2FC) between the ages for individual tissues (Figure 1A; $\overline{L2FC_{brain}}$ = 0.001, $\overline{L2FC_{skin}}$ = -0.007, $\overline{L2FC_{blood}}$ = 0.014). Notably, at this resolution, we already see a tendency for the majority of TE superfamilies in skin to be down-regulated during aging. In contrast, the majority of TE superfamilies in blood show a tendency towards up-regulation. In brain, we observe a more balanced picture.

The superfamily-based analysis, however, largely prohibits the investigation of the expression dynamics within a single superfamily. Using our previously described and evaluated strategy [34], we identified differentially expressed TEs (DETEs) at the level of individual loci during aging (24 vs. 6 months) in three different tissues (blood, brain, skin; see Methods). In total, we detected between ~50,000 and 100,000 expressed TEs (brain = 46,834, skin = 96,457, blood = 97,960; Supplemental Table 1). Of this rather large number of TEs with expression signals, only a minority of elements (~50-1,000) were found to be significantly differentially expressed during aging (Figure 1B; Supplemental Table 1). Assignment of the detected TEs to their TE classes revealed a tissue-specific composition of expressed TEs and DETEs (Figure 1C). Interestingly, only 4 % (n=7,441) of the total 241,251 detected TEs (across all tissues) were detected in all three tissues (Supplemental Figure 1), indicating a pronounced tissue specificity of TE expression. This observation could be explained by the tissue-specific accessibility of DNA, *e.g.*, during development [30]. In brain, the majority of detected TEs belong to the LINE class (39%), almost twice as many compared with blood and skin (20.29% and 23.68%, respectively).

Beyond previous findings of TE superfamily-based reports [13, 18, 35–37], we discovered down-regulated TEs during aging (brain = 42, skin = 580, blood = 23; false discovery rate [FDR] $\leq$ 0.05; Benjamini and Hochberg) at the same order of magnitude as up-regulated ones (see Figure 1B; brain = 93, skin = 466, blood = 32; FDR $\leq$ 0.05; Benjamini and Hochberg). Additionally, the locus-specific expression analysis (see Methods) shows stronger regulatory TE dynamics during aging. In several cases, we observe L2FCs that are orders of magnitude larger than the mean L2FC of their respective superfamily, e.g., an L1Lx_II_orf2 element in brain (chr18:37378135-37382301; L2FC = 1.99 vs. $\overline{L2FC}_{L1}$ = 0.00036), an ERVK/IAPEz-int element in skin (chr3:51240387-51241612; L2FC = -7.13 vs. $\overline{L2FC}_{ERVK}$ = 0.012), and B4/RSINE1 element in blood (chr4:32516261-3251637; L2FC = -1.52 vs. $\overline{L2FC}_{B4}$ = 0.015). Again, the different counts of DETEs in the analyzed tissues indicate a more tissue-specific regulation of TEs during aging.

Standardized expression scores for the top 50 DETEs (sorted by FDR, see Methods) reveal distinct expression patterns between young and old mice within TE classes (Figure 1D). Importantly, individual TEs within the same superfamily were frequently regulated in opposite directions. Such patterns likely contribute to the comparably small L2FC at the superfamily level (Figure 1A) since the opposite effects can cancel out each other. In addition, the data indicate a tissue-specific regulation of individual TEs within superfamilies. For example, we observe that the majority of differentially expressed endogenous retrovirus-K (ERVK) elements in the top 50 DETEs were up-regulated in brain, whereas the majority was down-regulated in aged skin. Furthermore, the top 50 DETEs from skin, blood, and brain underscore a high tissue specificity of differential TE expression (Figure 1D). On the global level (see Figure 1C) as well as among the top 50 DETEs, L1 elements were more frequently differentially expressed in brain compared to blood and skin. In brain, we typically observe the up-regulation of L1Md elements, *i.e.*, members of a large and active L1 superfamily in mice [38, 39]. Of note, among differentially expressed L1Mds we predominantly observe ORF2 loci that contain sequence information for the multifunctional ORF2p protein, which carries the endonuclease and reverse transcriptase activities of L1.

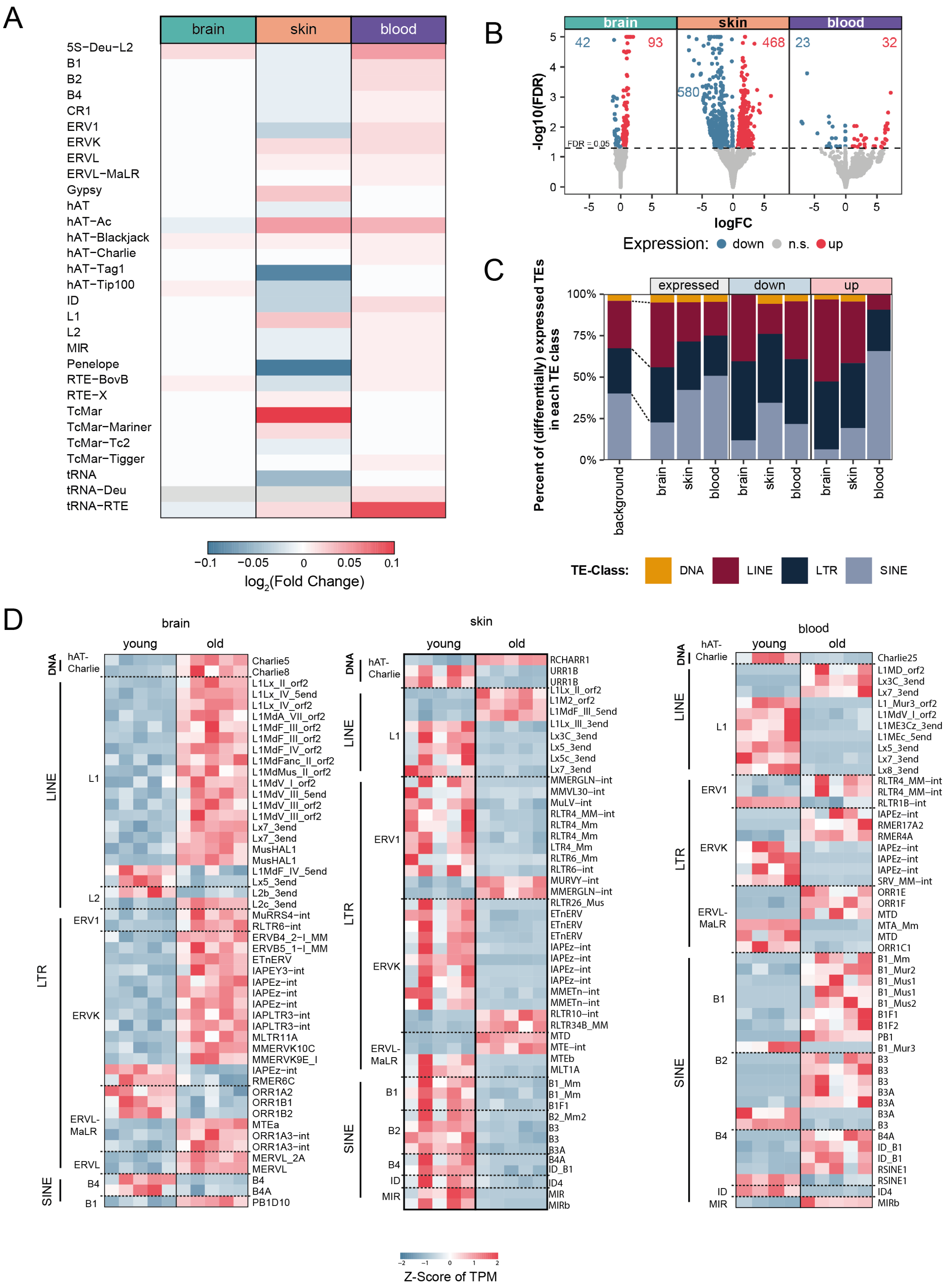

A
brain
skin
blood
5S−Deu−L2
B1
B2
B4
CR1
ERV1
ERVK
ERVL
ERVL−MaLR
Gypsy
hAT
hAT−Ac
hAT−Blackjack
hAT−Charlie
hAT−Tag1
hAT−Tip100
ID
L1
L2
MIR
Penelope
RTE−BovB
RTE−X
TcMar
TcMar−Mariner
TcMar−Tc2
TcMar−Tigger
tRNA
tRNA−Deu
tRNA−RTE
−0.1 −0.05 0 0.05 0.1
$\log_2$(Fold Change)
B
brain
skin
blood
42
93
468
580
23
32
FDR = 0.05
-log10(FDR)
0 1 2 3 4 5
-5 0 5
logFC
Expression: ● down ● n.s. ● up
C
expressed
down
up
Percent of (differentially) expressed TEs in each TE class
0% 25% 50% 75% 100%
background
brain
skin
blood
TE-Class: DNA LINE LTR SINE
D
brain
young
old
DNA
hAT-Charlie
LINE
L1
L2
LTR
ERV1
ERVK
ERVL-MaLR
ERVL
SINE
B4
B1
Charlie5
Charlie8
L1Lx_II_orf2
L1Lx_IV_5end
L1Lx_IV_orf2
L1MdA_VII_orf2
L1MdF_III_orf2
L1MdF_III_orf2
L1MdF_IV_orf2
L1MdFanc_II_orf2
L1MdMus_II_orf2
L1MdV_I_orf2
L1MdV_III_5end
L1MdV_III_orf2
L1MdV_III_orf2
Lx7_3end
Lx7_3end
MusHAL1
MusHAL1
L1MdF_IV_5end
Lx5_3end
L2b_3end
L2c_3end
MuRRS4−int
RLTR6−int
ERVB4_2−I_MM
ERVB5_1−I_MM
ETnERV
IAPEY3−int
IAPEz−int
IAPEz−int
IAPEz−int
IAPLTR3−int
IAPLTR3−int
MLTR11A
MMERVK10C
MMERVK9E_I
IAPEz−int
RMER6C
ORR1A2
ORR1B1
ORR1B2
MTEa
ORR1A3−int
ORR1A3−int
MERVL_2A
MERVL
B4
B4A
PB1D10
skin
young
old
DNA
hAT-Charlie
LINE
L1
LTR
ERV1
ERVK
ERVL-MaLR
SINE
B1
B2
B4
ID
MIR
RCHARR1
URR1B
URR1B
L1Lx_II_orf2
L1M2_orf2
L1MdF_III_5end
L1Lx_III_3end
Lx3C_3end
Lx5_3end
Lx5c_3end
Lx7_3end
MMERGLN−int
MMVL30−int
MuLV−int
RLTR4_MM−int
RLTR4_Mm
RLTR4_Mm
LTR4_Mm
RLTR6_Mm
RLTR6−int
MURVY−int
MMERGLN−int
RLTR26_Mus
ETnERV
ETnERV
ETnERV
IAPEz−int
IAPEz−int
IAPEz−int
IAPEz−int
MMETn−int
MMETn−int
RLTR10−int
RLTR34B_MM
MTD
MTE−int
MTEb
MLT1A
B1_Mm
B1_Mm
B1F1
B2_Mm2
B3
B3
B3A
B4A
ID_B1
ID4
MIR
MIRb
blood
young
old
DNA
hAT-Charlie
LINE
L1
LTR
ERV1
ERVK
ERVL-MaLR
SINE
B1
B2
B4
ID
MIR
Charlie25
L1MD_orf2
Lx3C_3end
Lx7_3end
L1_Mur3_orf2
L1MdV_I_orf2
L1ME3Cz_3end
L1MEc_5end
Lx5_3end
Lx7_3end
Lx8_3end
RLTR4_MM−int
RLTR4_MM−int
RLTR1B−int
IAPEz−int
RMER17A2
RMER4A
IAPEz−int
IAPEz−int
IAPEz−int
SRV_MM−int
ORR1E
ORR1F
MTD
MTA_Mm
MTD
ORR1C1
B1_Mm
B1_Mur2
B1_Mus1
B1_Mus1
B1_Mus2
B1F1
B1F2
PB1
B1_Mur3
B3
B3
B3
B3A
B3A
B3A
B3
B4A
ID_B1
ID_B1
RSINE1
RSINE1
ID4
MIRb
−2 −1 0 1 2
Z-Score of TPM

Figure 1 - Locus-specific quantification of TE expression. (A) Mean L2FC of TE superfamilies (rows) in brain (green), skin (orange), and blood (purple) compared between 24 months and 6 months old mice. **(B)** Volcano plots of individual TE expression in brain (green), skin (orange), and blood (purple) (L2FC of expressed TEs [x-axis] and their significance [y-axis]). Colored dots indicate differentially expressed TE loci (FDR ≤ 0.05; blue: down-regulated TE loci; red: up-regulated TE loci). **(C)** Proportion of TE classes among expressed or differentially expressed individual TE loci. \`background\` represents the proportion of TE classes among all TEs in the mouse genome. \`expressed\` (gray) denotes TEs with an expression signal, while \`up\` (red) and \`down\` (blue) denote differentially expressed TEs. **(D)** Heatmap of standardized expression scores (i.e., z-scores) derived from TPM of differentially expressed individual TEs (top 50, sorted by FDR; rows) in brain (green), skin (orange), and blood (purple) grouped by TE class and superfamily and clustered by up- and down-regulated individual TEs. The TE element annotation is based on the mm10 RepeatMasker annotation (version: open-4.0.5 - Repeat Library 20140131). **Abbreviations:** FDR – False discovery rate; L2FC – $\log_2$(fold change); TE – transposable element; TPM - transcripts per million.

In contrast, only few SINEs were (differentially) expressed in brain tissue. Instead, SINEs were strongly up-regulated in the blood of aging mice. This expression was predominantly sustained by B1, B2 and B4 elements. Concerning LTR elements, the picture is dominated by differentially expressed endogenous retroviruses (ERVs) typically up-regulated in aged brain tissue and down-regulated in aged skin. Among the down-regulated LTRs in brain are ERVL-MaLR elements of type ORR1A2, ORR1B1, and ORR1B2. Previous research has suggested that these elements harbor binding sites for the developmentally decisive transcription factor Tbx6 [40] and, in the case of ORR1A2, for the differentiation factor Klf4 [41]. The latter has been associated with aging and neurodegeneration [42, 43]. Moreover, the down-regulated ORR1A2 element itself is located at the opposite strand of an intron of *Pde10a*. *Pde10a* is a gene mainly expressed in brain and a target for psychiatric and neurodegenerative drug discovery [44]. These examples highlight that TEs with differential expression during aging are associated with tissue-specific hallmarks of aging, such as neurodegenerative processes.

In summary, the locus-specific analysis of TE expression reveals tissue-specific differences on the level of TE classes, superfamilies, subfamilies, and individual TEs. At the same time, the data indicate that the direction of regulation during aging is not the same for all members of a superfamily. In all three tissues, the count of up- and down-regulated TEs are of the same order of magnitude.

Independent expression of TEs

Many TEs are located in introns (n = 793,002 of ~4.2 million [18.75%]) or in close proximity of a gene (n = 2,627,938 [62%] within 100 kb up or downstream). Thus, we checked whether the observed TE expression is a mere consequence of the transcription of its host. Clearly, TEs located within an intron or downstream of a gene may be co-transcribed as a consequence of intron retention [45] or separate splicing processes [46]. Such host-initiated TE expression could entail a lack of transcription start sites (TSS) within or nearby the TE. On the other hand, host-independent TE expression would require a separate TSS at the TE. To distinguish between these cases, we applied Cap-analysis gene expression sequencing (CAGE-Seq) [47], an established method to identify TSS on a genome-wide scale [29], to the same samples used for the RNA-Seq. This enabled us to create a map of TSSs associated with TEs for all three tissues (see Methods; Supplemental Table 1). The enrichment patterns of TSSs in TE superfamilies (against the genomic TE background) underscore the observed highly tissue-specific TE expression (Figure 2A): the B1 superfamily is the only set of TEs that consistently accumulates TSSs in all three tissues.

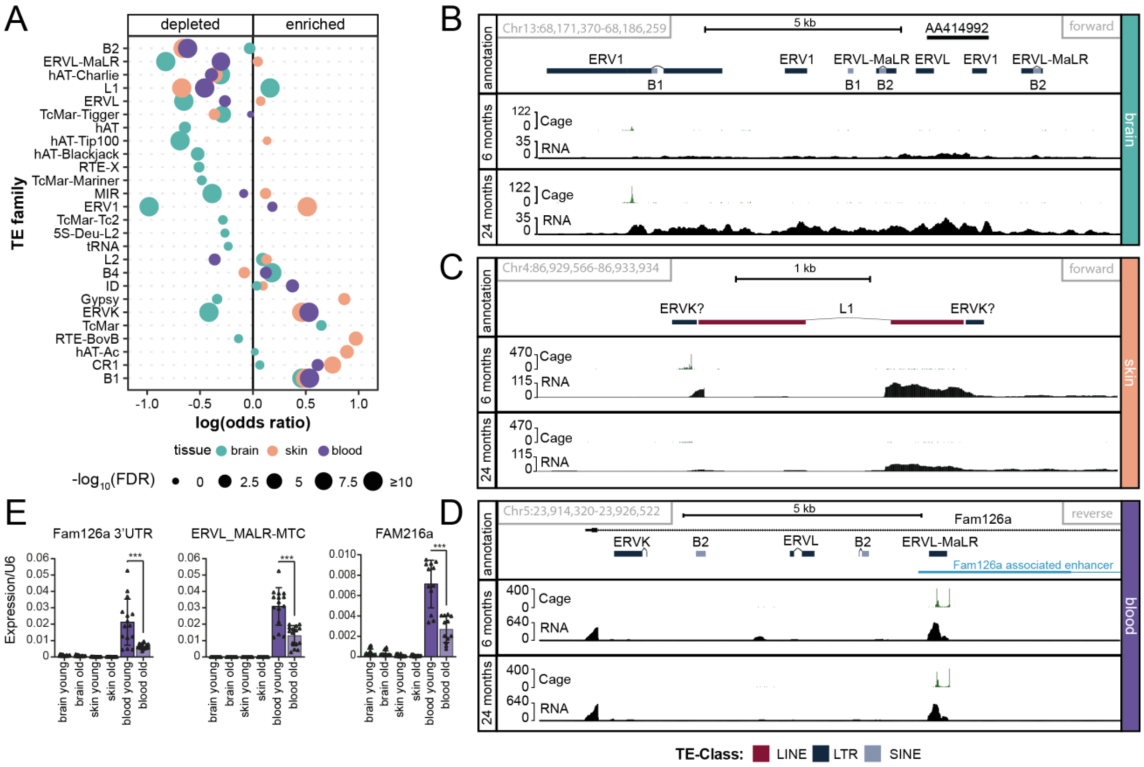


Figure 2 – Identification of independently expressed TEs. The intersection of CAGE-Seq peaks with TEs allows to predict TE-specific TSSs. **(A)** Significance (-$\log_{10}$(FDR), point size) of depleted (left) and enriched (right) TE superfamilies based on individual TEs with a TSS. Normalized by the number of TEs in the respective superfamily as given by the genomic annotation. The x-axis displays the log(odds ratio) in-set vs. in-genome (by count; see Methods). **(B-D)** CAGE- (green) and RNA-Seq (black) coverage tracks of genomic regions with putatively independently expressed TEs in brain (green), skin (orange), and blood (purple) for 6 (middle row of each panel) and 24 months old mice (last row). The first row shows the annotation of TEs, genes and enhancers in the

respective region. **(B)** Example of a region with multiple TEs that is jointly up-regulated during aging in brain. The TSSs indicates that the transcript starts in an ERV1 element (coverage track of CAGE). **(C)** Example of a region with individual TEs that is down-regulated during aging in skin. Transcription starts from an ERVK? element. **(D)** Example of an independently expressed TE (ERVL-MaLR) in blood that intersects with an enhancer (light blue) associated with the gene Fam126a. The independently expressed TE itself is located in the last intron of the gene Fam126a. **(E)** RT-qPCR analysis shows the tissue-specific expression and co-regulation of ERVL_MALR and Fam126a. Data are shown as mean ± s.e.m. and p-values are from a two-sided unpaired t-test (*** – p-value <0.001). **Abbreviations:** CAGE – Cap-analysis gene expression sequencing; FDR – False discovery rate; L2FC – $\log_2$(fold change); RT-qPCR – real-time quantitative polymerase chain reaction; TE – transposable element; TSS - transcription start site.

The overall strongest enrichment of TSSs is observed for LINE/RTE-BovB elements in skin (log(odds ratio) = 0.97, FDR = $2.8e^{-04}$; standard binomial test, corrected with Benjamini and Hochberg). In brain, we observe a reduction of putatively independently expressed TEs in the majority of TE superfamilies. In particular, we observe a depletion of TSS-carrying TEs in ERVL, ERVK and ERV1 superfamilies. In turn, independently expressed ERVs appear to be enriched in blood and skin. A particularly strong enrichment is observed for ERV1 superfamily members in skin. Just recently, it has been suggested that the expression of ERVs are critical means for controlling the inflammatory response to exogenous skin microbiota [48]. In brain, a notable exception is the TSS enrichment in the L1 superfamily (log(odds ratio) = 0.16, FDR = $1e^{-06}$; standard binomial test, corrected with Benjamini and Hochberg). Its elements are repeatedly associated with neuronal (dys-)functions in the literature [12, 21, 22, 24, 25].

To illustrate that the complexity of TE expression and associated TSSs is not reflected by the above summary statistics, we provide examples of age-dependently expressed TEs (Figure 2B-C). Here, neighboring TEs are either up- (Figure 2B) or down-regulated (Figure 2C) during aging according to RNA- and CAGE-Seq data. The example in brain shows that multiple TEs from the SINE and ERV classes appear to be expressed in a single transcript jointly up-regulated during aging (Figure 2B). A single down-regulated TSS indicates that an ERV1-associated promoter drives the expression of this TE structure.

Intriguingly, TEs have been found to be frequently associated with enhancers [30, 49–51]. Thus, we explored potential co-regulations of TEs and their coding host genes. One example for a potential co-regulation of independently expressed TEs and a host gene is found in the intron of Fam126a (chr5:23915277-24030312; Figure 2D). Here, an element of the ERVL-MaLR superfamily expressed in blood overlaps with an annotated enhancer. Our RNA-Seq data indicate that the host gene is down-regulated (L2FC = -0.80, FDR = $1.6e^{-04}$), while the TE shows a borderline down-regulation (L2FC = -0.01, p-value = 0.12). The CAGE-Seq signal intersecting with the ERVL-MaLR element also shows a tendency for down-regulation (L2FC = -0.093, p-value = 0.16). Indeed, we were able to confirm the differential expression of the TE, the 3'UTR of Fam126a as well as the entire host gene by RT-qPCR (see Methods, Figure 2E). These findings suggest that the ERVL-MaLR element may affect the overlapping Fam126a enhancer to elicit a co-regulation of the TSSs that give rise to Fam126a upstream and ERVL-MaLR itself.

Together, the CAGE-Seq data enable the distinction between TEs that possess their own TSS and TEs that most likely require co-transcription from a TSS belonging to a host gene or another TE. Our data corroborate tissue specificity also for independently expressed TEs, which include the specific enrichment of L1 expression in brain and the skin-specific expression of ERV1s.

Expression of independent TE regions

As illustrated in Figure 2B-D, the arrangement of TEs in the genome as well as their expression is complex. TEs of different subfamilies may occur in clusters or even overlap with each other in the genome. Therefore, we grouped closely spaced TEs (distance ≤ 500 bp) into TE regions. Analogously to our analysis for single TEs, TE regions overlapping with a CAGE-Seq-determined TSS were deemed to be independently expressed (Figure 3A). In total, we identified between 3,332 and 11,610 independently expressed TE regions (blood = 3,332, brain = 11,610, skin = 4,234) and categorized them into single-, double-, and multi-TE regions, i.e., regions that contain one, two, or more TEs (Supplemental Figure 2A). In agreement with our previous results, elements of the L1 superfamily made up the majority of all three TE region types in brain. In skin and blood, the strongest contribution came from B4 and B1 elements (Supplemental Figure 2B). Based on this categorization, we asked whether specific TE families are more frequently present in the body or at the 5'- or 3'-ends of multi-TE regions (Figure 3B). While we observed a clear overrepresentation of L1 elements (brain) and B1 elements (skin, blood), it was not restricted to or overrepresented at any location within the regions.

Next, we analyzed the distribution of TSSs within independent TE regions. In skin and blood, we observed a pronounced TSS frequency peak just at the beginning of TE regions steadily decreasing towards the end of the region (Figure 3C). Interestingly, these data provide evidence that the 5’-element of a chain of closely spaced TEs is more likely to carry a TSS than any other downstream element (Supplementary Figure 2E). Thus, downstream elements may frequently be co-regulated by regulatory regions located near the regions 5’-end and within the first element. In brain, the distribution was markedly different. In contrast to the two other tissues, TSSs were more frequently located at the 3'-end of the regions. Of note, previous work reported that L1 elements are frequently truncated at the 5’-end [52, 53] and transcription initiated at their 3’-end [29]. Thus, the overrepresentation of L1 elements in brain-expressed TEs offers an explanation for the unexpected enrichment of TSSs at the 3’-end of TE regions in brain.

We performed a DNA motif analysis upstream of TSSs of the TE regions to identify transcription factors that may be involved in the TE expression (see Methods). In all three tissues, a motif was enriched that was most similar to a binding site of Sox transcription factors, with brain showing the most significant enrichment (Figure 3D). Intriguingly, Sox motifs are most strongly enriched within the L1 superfamily (Supplemental Figure 3E), and, thus, we checked whether Sox motif-carrying L1 elements may explain the frequent occurrence of TSSs at the 3’-end of brain-expressed TE regions. To investigate the spatial relationship between TSS and the Sox-motif, we calculated a Sox/TSS ratio along the TE regions (Figure 3E; Supplemental Figure 3A-B). In brain, the Sox/TSS ratio strongly increases towards the 3’-end, indicating that initiation of TE transcription is linked to the Sox motif (Figure 3E). Moreover, we found that there is a rather large population of Sox-motif-carrying TEs with a length of 1,000 ± 50 bp (n=667, Figure 3F). Strikingly, all elements in this set are annotated by RepeatMasker [54] as 3'-ends of L1 subfamilies (Figure 3F, inset). In summary, our data indicate that a substantial amount of L1 expression in brain, but not in skin or blood (Supplemental Figure 3C-D), can be attributed to a specific type of L1 3’-ends which harbors binding sites for the Sox transcription factor family.

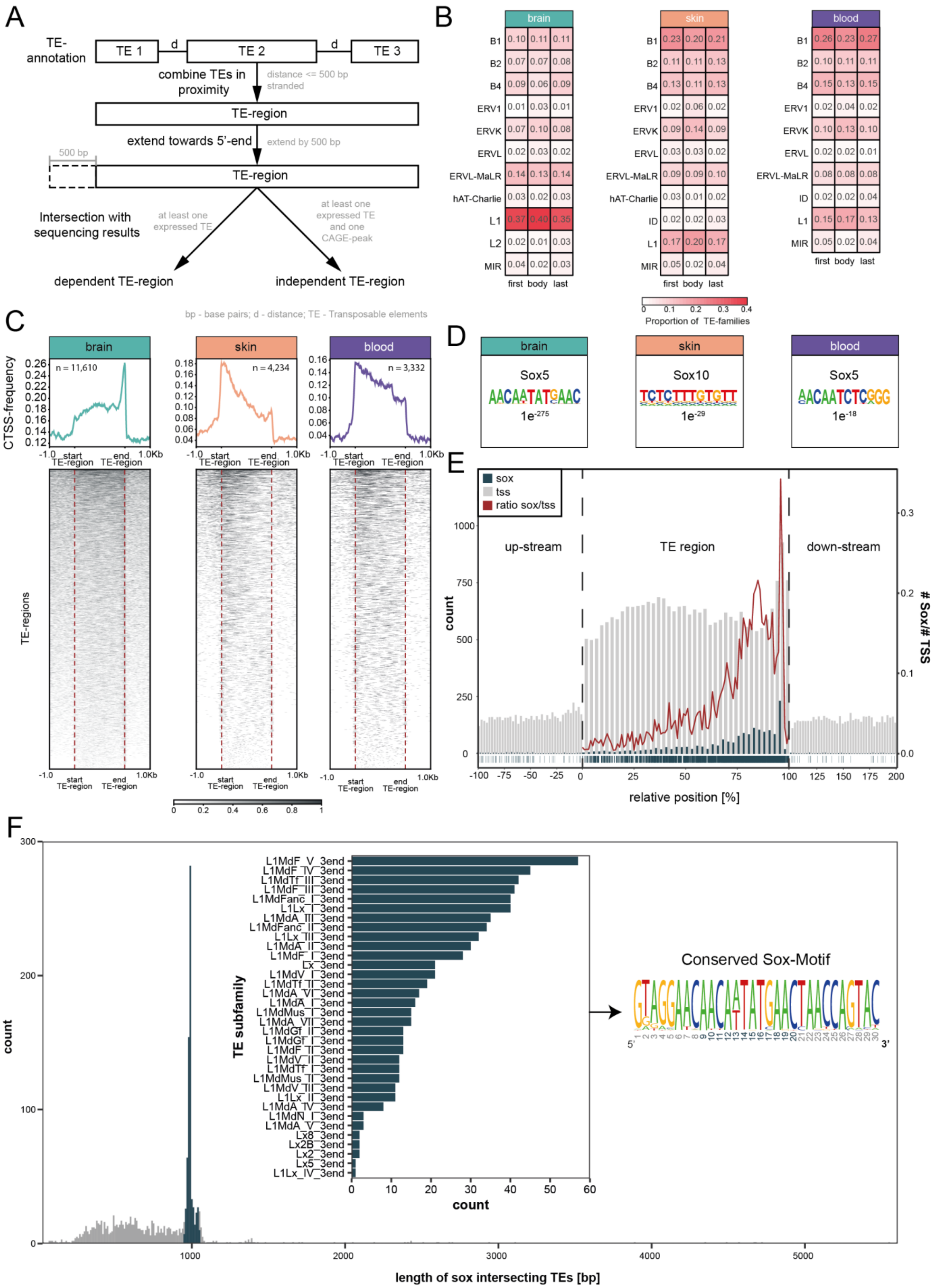

A
TE-annotation
TE 1
d
TE 2
d
TE 3
combine TEs in proximity
distance <= 500 bp stranded
TE-region
extend towards 5'-end
extend by 500 bp
500 bp
TE-region
Intersection with sequencing results
at least one expressed TE
at least one expressed TE and one CAGE-peak
dependent TE-region
independent TE-region
bp - base pairs; d - distance; TE - Transposable elements
B
brain
B1 0.10 0.11 0.11
B2 0.07 0.07 0.08
B4 0.09 0.06 0.09
ERV1 0.01 0.03 0.01
ERVK 0.07 0.10 0.08
ERVL 0.02 0.03 0.02
ERVL-MaLR 0.14 0.13 0.14
hAT-Charlie 0.03 0.02 0.03
L1 0.37 0.40 0.35
L2 0.02 0.01 0.03
MIR 0.04 0.02 0.03
first body last
skin
B1 0.23 0.20 0.21
B2 0.11 0.11 0.13
B4 0.13 0.11 0.13
ERV1 0.02 0.06 0.02
ERVK 0.09 0.14 0.09
ERVL 0.03 0.03 0.02
ERVL-MaLR 0.09 0.09 0.10
hAT-Charlie 0.03 0.01 0.02
ID 0.02 0.02 0.03
L1 0.17 0.20 0.17
MIR 0.05 0.02 0.04
first body last
blood
B1 0.26 0.23 0.27
B2 0.10 0.11 0.11
B4 0.15 0.13 0.15
ERV1 0.02 0.04 0.02
ERVK 0.10 0.13 0.10
ERVL 0.02 0.02 0.01
ERVL-MaLR 0.08 0.08 0.08
ID 0.02 0.02 0.04
L1 0.15 0.17 0.13
MIR 0.05 0.02 0.04
first body last
0 0.1 0.2 0.3 0.4
Proportion of TE-families
C
brain
skin
blood
CTSS-frequency
n = 11,610
n = 4,234
n = 3,332
-1.0 start TE-region end TE-region 1.0Kb
TE-regions
0 0.2 0.4 0.6 0.8 1
D
brain
Sox5
AACAATATGAAC
1e-275
skin
Sox10
TCTCTTTGTGTT
1e-29
blood
Sox5
AACAATCTCGGG
1e-18
E
sox
tss
ratio sox/tss
up-stream
TE region
down-stream
count
# Sox/# TSS
relative position [%]
-100 -75 -50 -25 0 25 50 75 100 125 150 175 200
0 250 500 750 1000
0.0 0.1 0.2 0.3
F
count
0 100 200 300
1000 2000 3000 4000 5000
length of sox intersecting TEs [bp]
TE subfamily
L1MdF_V_3end
L1MdF_IV_3end
L1MdTf_III_3end
L1MdF_III_3end
L1MdFanc_I_3end
L1Lx_I_3end
L1MdA_III_3end
L1MdFanc_II_3end
L1Lx_III_3end
L1MdA_II_3end
L1MdF_I_3end
Lx_3end
L1MdV_I_3end
L1MdTf_II_3end
L1MdA_VI_3end
L1MdA_I_3end
L1MdMus_I_3end
L1MdA_VII_3end
L1MdGf_II_3end
L1MdGf_I_3end
L1MdF_II_3end
L1MdV_II_3end
L1MdTf_I_3end
L1MdMus_II_3end
L1MdV_III_3end
L1Lx_II_3end
L1MdA_IV_3end
L1MdN_I_3end
L1MdA_V_3end
Lx8_3end
Lx2B_3end
Lx2_3end
Lx5_3end
L1Lx_IV_3end
0 10 20 30 40 50 60
count
Conserved Sox-Motif
GTAGGAACAACAATATGAACTAACCAGTAC
5' 3'

Figure 3 – Characterization of TE regions. Adjacent TEs can be co-expressed, therefore, individual TEs in close proximity were combined to TE regions and characterized. **(A)** Scheme of the definition of TE regions and its division into independently and dependently expressed TEs. **(B)** Proportion of TE superfamilies (rows) at respective positions (first, body, last; in columns) in TE regions with more than two members in brain (green), skin (orange), and blood (purple). **(C)** CAGE-Seq peak frequency across all TE regions in brain (green), skin (orange), and blood (purple) and their adjacent areas (≤ 1Kb). The frequencies result from the peak count (greyscale) across the TE regions as depicted below frequency plots. **(D)** Enriched Sox motifs in putative promoter regions (starting at TSS and ending 500 bp up-stream) of independently expressed TE regions in brain (green), skin (orange), and blood (purple) with the respective FDR. **(E)** Counts (left y-axis) of TSSs (gray) and Sox motifs (blue) and their ratio (red, right y-axis) across TE regions and their adjacent areas (up- and down-stream; ≤ 500 bp) in brain. **(F)** Length histogram of Sox-motif-carrying TEs in brain. Highlighted bars (blue) indicate individual TEs with a length of 950 to 1050 bp. The inset displays the member counts of subfamilies (rows) within the highlighted length interval. The motif logo shows the frequency of respective base-pair occurrence within the Sox-motif of those TEs. **Abbreviations:** bp – base-pair; CAGE – Cap-analysis gene expression sequencing; FDR – False discovery rate; Kb - kilo-base; L2FC – log2(fold change); TE – transposable element; TSS - transcription start site.

The expression of intronic TEs or TEs proximal to important regulatory elements may elicit effects on host genes or associated genes [26, 55–57]. To analyze this relation, we compared the RNA-Seq data between 24 and 6 months old mice for independently expressed TE regions and their hosts (Figure 4, see Methods). Overall, we detected between ~2,600 and 10,200 independent TE regions (brain = 10,195, skin = 3,244, blood = 2,604) that intersect with a gene. Accounting for multiple overlaps, between ~1,500 and 2,000 genes are potentially affected by independent TE region expression (brain = 1,788, skin = 2,047, blood=1,478) (Figure 4A).

In all three tissues, we observed a positive correlation between the L2FC of pairs of independent TE regions and their host genes (Figure 4B). In skin, a pronounced common up-regulation of multiple TSS-carrying TE regions was observed with the neighboring genes *Skint5* and *Skint11* (Figure 4B), located on opposite strands in an 800 kb region on chromosome 4. The genes are located at the 3’-end of an even longer cluster comprising all members of the paralogous *Skint* family (*Skint1-11*). Recent studies demonstrated that the Skint family regulates Vγ5Vδ1+ dendritic epidermal T-cells (DETC), the dominant T cell compartment in the epidermis [58, 59]. DETCs are of special relevance in keratinocyte proliferation, survival, and antimicrobial protection [60] and may thus play a critical role in the development of skin aging hallmarks and affect skin barrier function [61]. Interestingly, a second prominent cluster of differentially expressed TSS-carrying TE regions is located within the type I keratin family genes *Krt27*, *Krt28*, and *Krt35* (Figure 4B). Keratin genes are the largest subset of intermediate filament genes that arose from extensive evolutionary gene duplication events creating a diverse set of paralogs [62]. All three mentioned keratin genes appear to play a role in the hair follicle and its inner root sheath. Although highly variable, hair loss is commonly observed in aging mice [63]. In brain, one functionally interesting co-regulated pair is found at the locus of the Ras guanine nucleotide releasing factor 1 (*Rasgrf1*). The TE region as well as the host gene was up-regulated during aging (Figure 4B). Early studies have shown that RasGRF1's downstream signaling pathway is critical for the consolidation of long-term memory [64]. Additional evidence for a neuronal function of RasGRF1 has been provided through recent studies that found RasGRF1 to be important for axonal growth of cortical neurons from rats [65] and for regulating dendritic density in human stem cell-derived neurons [66]. However, *Rasgrf1*-deficient mice have been shown to age significantly slower than their wild-type counterparts and display strongly improved neuromuscular coordination [67]. Together, these data indicate that the co-regulation of TSS-carrying TE regions and proximal protein-coding genes may contribute to their tissue-specific and age-dependent expression dynamics.

Our data on blood was less conclusive as compared to the other tissues. However, we observed a strong down-regulation of one TE region located in the first intron of *Kcnq5*, a member of the KCNQ potassium channel family that did not coincide with a differential regulation of the host. Importantly, only one pair at the *Pla2g4e* locus displayed a clearly divergent regulation. While the *Pla2g4e* gene was up-regulated in brains of old mice (L2FC = 0.4, FDR = $6e^{-03}$), the TE region showed a trend towards down-regulation (L2FC=-0.4, p-value = $7e^{-04}$, n.s. after correction). Previously, the gene has been suggested to play a role in the development of Alzheimer's disease [68]. The overexpression of *Pla2g4e* in brain tissue of mice expressing amyloid precursor proteins (APP) led to the amelioration of disease associated impairments, *e.g.*, an improvement of memory [68]. Thus, there may be an antagonistic regulatory association between the TE_region_876806 (chr2:120217502-120225936) and *Pla2g4e* through which TE down-regulation could promote neuroprotective processes during aging.

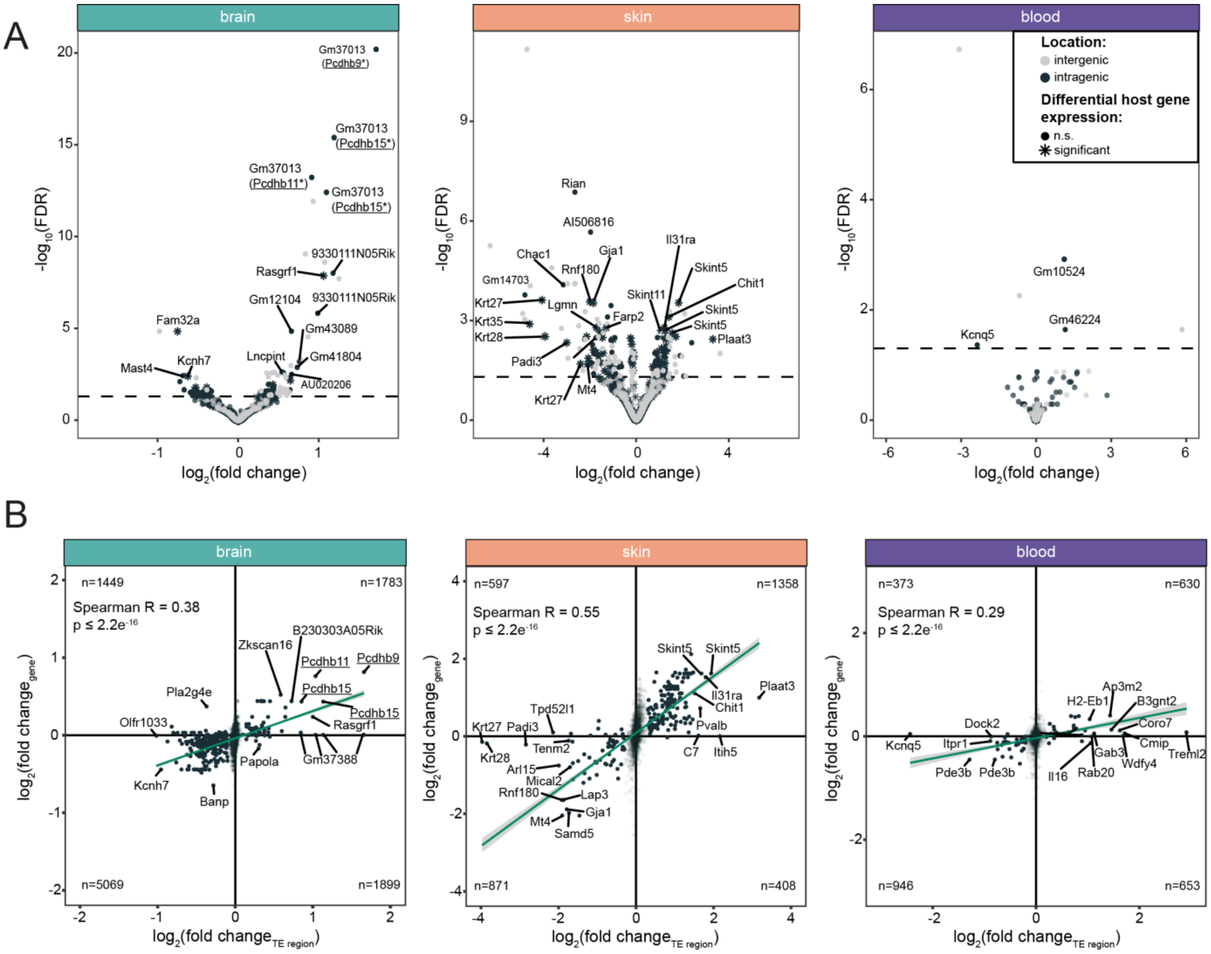


Figure 4 – Association of independent TE regions and host genes. (A) Volcano plots of expressed TE regions in brain (green), skin (orange), and blood (purple) showing the $\log_2$(fold change) (24- vs. 6-months old mice) and their significance (log10(FDR), y-axis). Each point indicates one independently expressed TE region (dark blue: TE region intersects with gene = intragenic; gray: TE region between genes = intergenic). Data points above the dashed line (FDR = 0.05) represent differentially expressed TE regions and the asterisk symbol indicates differential expression of their host gene (FDR ≤ 0.05). The most highly expressed TE regions in brain overlap with gene Gm37013, which spans a cluster of protocadherin genes. Protocadherin genes in close proximity to the respective TE regions are indicated in parentheses (* - differentially expressed). **(B)** Scatter plots showing the positive correlation between L2FCs of independently expressed TE regions (x-axis) and their host genes (y-axis) in brain (green), skin (orange), and blood (purple). Each data point indicates an independently expressed TE region that overlaps with a gene. The green line shows the best fit to the linear model. In the brain panel (green), the underlined protocadherin genes represent the expression of the respective protocadherin gene (y-axis) and the closest independently expressed TE region. **Abbreviations:** FDR – False discovery rate; L2FC – $\log_2$(fold change); TE – transposable element.

A marked up-regulation of multiple TE regions overlapping with Gm37013, the protocadherin alpha 4-gamma precursor gene, was observed in aged brains (Figure 4A and B). The gene spans an entire region of different protocadherins, subdivided in three separate gene clusters (α, β, and γ; chr18:36930184-37841870). Mechanistically, protocadherins are present in the synaptic membrane and thought to play a critical role in the neuronal signal transduction [69]. The neuron-specific combination of expressed protocadherins equips each neuron with a unique combination of cell-surface homophilic recognition molecules that result in self-avoidance [70]. This neuronal self-avoidance prevents dendrites and axons to connect to their own soma [71]. We find that the protocadherin cluster is loaded with independently expressed TE regions. The ones strongly up-regulated during aging are specifically found in the β-cluster (Figure 5A). The increased expression of TE regions is accompanied by the up-regulation of all β-cluster genes indicating potential co-regulation (Figure 5B). Two TE regions mainly composed of L1 elements (TE_region_728239, TE_region_728242) are located directly upstream of Pcdhb15 (chr18:37473540-37476340). Their up-regulation is supported by the RNA- and CAGE-Seq data. Additionally, TE_region_728239 shows a Sox-motif close to the TSS (Figure 5C). Intriguingly, the transcription start site of Pcdhb15 is down-regulated (Figure 5D blue arrow), while the transcript itself is up-regulated. Thus, our data suggest that the TE regions provide alternative transcription initiation sites for Pcdhb15 enabling its up-regulation during aging despite a down-regulation of its own TSS. The transcription of the protocadherin cluster is highly complex and the promoter usage of the α-cluster was recently found to be driven by stochastic processes guided by an antisense long non-coding RNA (lncRNA) [72]. Our data suggest the hypothesis that TSSs provided through TEs have a role in the stochastic promoter selection.

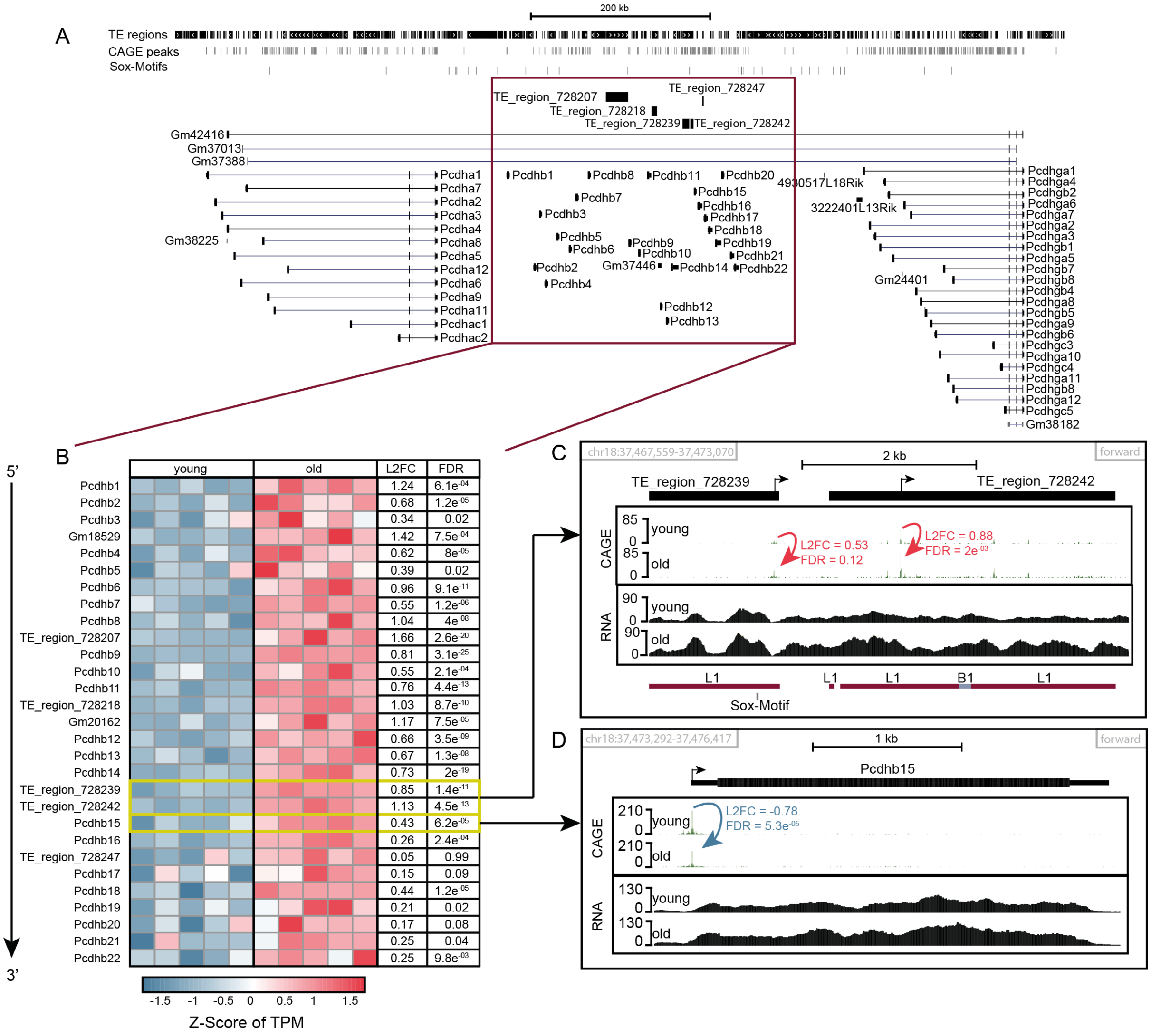


Figure 5 – Differentially expressed TE regions in protocadherin cluster in brain. (A) Genome-Browser-like overview of the protocadherin cluster with tracks for TE regions, CAGE-peaks, Sox-Motifs, expressed TE regions in brain, and gene annotations. **(B)** Heatmap of standardized expression scores derived from TPM of protocadherins and TE regions in the protocadherin beta cluster in young (6 months, left) and old (24 months, right) mice sorted by their genomic start position (5' → 3') in brain. L2FC and FDR values are displayed for each gene/TE region. **(C)** Genome-Browser-like view of the TE regions 728239 and 728242. The top track shows differentially expressed TE regions (black arrows indicate TSS and transcription direction) located up-stream of Pcdhb15 in brain. The CAGE and RNA coverage tracks for young and old mice are shown below. In the CAGE row, red arrows indicate the up-regulation in aged mice (positive L2FC) of TSSs that overlap with the TE regions. The last two tracks show the location of individual TEs (superfamily level) and a predicted Sox-motif in the first L1 element. **(D)** Genome-Browser-like view of the Pcdhb15 gene. At the top, the annotation of the gene Pcdhb15 is shown (black arrow indicates TSS and transcription direction). The CAGE and RNA coverage tracks for young and old mice are shown below. In the CAGE row, the blue arrow indicates the down-regulation in aged mice (negative L2FC) of the TSS that overlaps with Pcdhb15, while the RNA-Seq track below indicates an up-regulation of Pcdhb15 during aging. **Abbreviations:** CAGE – Cap-analysis gene expression sequencing; FDR – false discovery rate; L2FC – log2 fold change; TPM – transcripts per million; TE – Transposable element; TSS – transcription start site.

The co-regulation of independent TE regions and their host genes may point to a coupled functional role for common pathways. Such relations have already been shown for ncRNAs and their host genes [46]. Therefore, we investigated the biological role of genes that were affected by independently expressed TE regions. Against the background of all brain-expressed genes, a gene ontology (GO) analysis (see Methods) revealed a strong enrichment of genes with functions in neuronal synapses and signaling (Figure 6A). Surprisingly, the enrichment is substantially stronger for genes associated with independently expressed TE regions than for differentially expressed genes (DEGs) during aging (Supplemental Figure 4A). To analyze the influence of the genomic TE distribution on this result, we checked whether genes with neuronal functions harbor TEs as frequently as other brain-expressed genes. To this end, we counted the number of TEs in genes from the previously identified GO terms and compared them with the number of TEs in other randomly drawn expressed genes. Strikingly, our data show a strong and consistent accumulation of TEs in tissue-specific genes with neuronal functions (Figure 6B), while TE free genes are depleted in similar GO terms, *e.g.*, neuron to neuron synapse (Supplemental Figure 5). Analogous analyses in skin and blood corroborated that TEs appear to be enriched for localization in genes that belong to key tissue-specific pathways (Supplemental Figure 4B-C). In skin, for instance, the strongest enrichments are observed for regulation of Wnt signaling pathways and cell-substrate junctions. In contrast, we find the strongest enrichments for B cell activation and immune signaling pathways for blood.

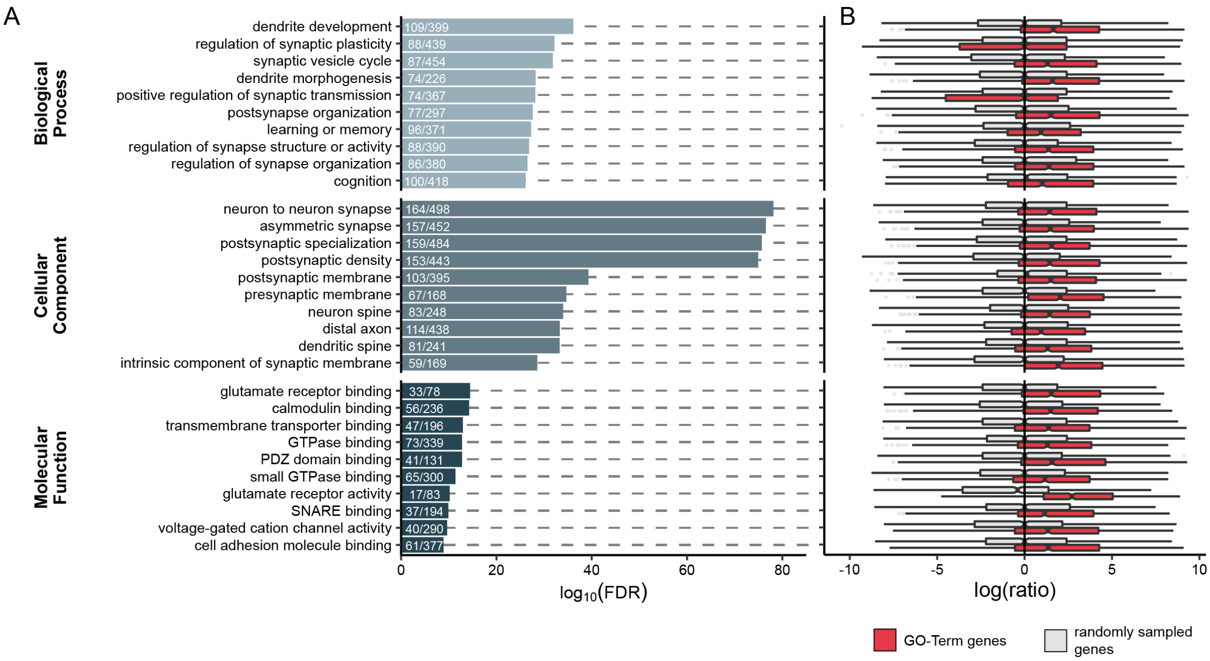


Figure 6 – GO term analyses of expressed genes intersecting independently expressed TE regions in brain. (A) Top 10 GO terms (sorted by FDR) of Biological Process, Cellular Component, and Molecular Function where genes overlapping with independently expressed TE regions in brain are enriched (background all detected genes in brain). The x-axis shows the significance (log10(FDR); Benjamini-Hochberg corrected) for each GO term (y-axis), while the numbers in each bar represent the count of genes overlapped by independent TE regions and the count of genes within each GO term. **(B)** Ratio of counts of intronic TEs in gene set of interest (red = GO term genes; gray = randomly sampled set with same size of GO term gene set) and a randomly selected gene for the gene set of expressed genes in brain. For each GO term, one gene was drawn from each set and the ratio was calculated, which was repeated 1,000 times (content of one box). The box plot center line represents the median, the upper and lower bounds correspond to the first and third quartiles, and the whiskers reach to 1.5 times the interquartile range. **Abbreviations:** FDR – false discovery rate; GO – Gene Ontology; TE – transposable elements.

In summary, we observe a strong enrichment of multiple tissue-specific functions in genes overlapping independent TE regions and the genes display a strong co-regulation with these TSS-carrying TE regions.

Discussion

Despite the potentially beneficial roles of TEs on an evolutionary scale, TE activity in somatic cells is mainly associated with the erosion of genome integrity and regulation promoting diseases [73–75]. In aging cells, it has been shown that the loss of (epigenomic) control over TEs leads to chronic sterile inflammation typically referred to as "inflammaging" [76]. To better assess the dynamics, causes, and potential effects of TE expression during aging on a genome-wide scale, we applied a locus-specific approach to characterize the expression of TEs [34] in three tissues of young and old mice. Thus, our study closes a gap between superfamily-level-based analyses and studies that focused on individual elements. Specifically, it demonstrates that the expression dynamics of multiple TE loci differ substantially from their superfamily-based averages (Figure 1). Hence, future research should intensify efforts to provide locus-specific data rather than aggregates at the TE class, superfamily, or subfamily level. Nevertheless, aggregation of transcribed TEs on the superfamily level clearly shows distinct expression patterns for the three analyzed tissues. Previous research reported that members of the L1 superfamily are active in the mouse brain and key drivers of genomic mosaicism in neurons [21, 77, 78]. Well in line, we observe a characteristic enrichment of expressed L1 superfamily members in that tissue.

In the context of aging, it was proposed that the relaxation of heterochromatin in gene-poor regions during aging makes TEs accessible and leads to increased TE activity [13, 35]. Our study clearly shows that TEs are about as often down-regulated as up-regulated (Figure 1B). While not too surprising at a first glance, this finding thwarts the notion of a categorically detrimental role of TE expression. In analogy to the reported TE down-regulation potentially helping cancer cells to hide from the immune system [19], one may surmise that TEs are sentinels for the (epi-)genomic integrity of a cell. The question arises to which extent age-related TE down-regulation could facilitate the emergence of diseases by diminishing the clearance of deregulated cells.

In addition to these potential global functions, our RNA-Seq analysis established that expressed TEs are frequently located intragenic of coding genes (Supplemental Table 1). To clearly distinguish TEs piggybacking on their host’s transcription, *e.g.*, through intron retention, from TEs with their own TSS, we performed a CAGE-Seq analysis. In addition to TSSs, CAGE-Seq enabled us to identify the putative promoters of expressed TEs. In all three tissues, the putative promoter regions significantly enriched DNA recognition motifs of the Sox transcription factor family (Figure 3D). The distribution of TSSs across regions with one or more closely spaced TEs in skin and blood indicated that it is the first element of a region that frequently serves as a starting point for transcription (Figure 3C). The marked difference of the TSS distribution in brain, *i.e.*, a more frequent occurrence of TSS at the regions’ 3’-ends, was associated with an increased presence of Sox motifs and L1 3’-end subfamilies of a characteristic length (Figure 3E). Thus, our data suggest that Sox transcription factors could be involved in the control of these regions. The strongest motif similarity was seen for Sox5, a transcription factor of the SoxD group. Sox5 was reported to be involved in controlling critical fate decisions for subtype-specific neuronal differentiation [79]. Further, it was shown that Sox5 (together with its sibling Sox6) is required for the activation of reversibly quiescent neural stem cells [80]. Moreover, the gene has been suggested to be involved in the development of autism spectrum disorders [81]. The strong and spatially correlated enrichment of these motifs near TSSs of TE’s begs the question whether the expression of TEs affects the function of this essential neuronal transcription factor.

We observed a strong positive correlation of age-related expression changes of TEs and their overlapping genes (Figure 4B). Our results indicate that TEs - despite having their own TSSs - are co-regulated with their host genes. It remains to be established whether TE and host gene expression might reinforce each other and which mechanisms are critical for this correlation, *e.g.*, by keeping the DNA in accessible configurations or by co-opting distal enhancers. Analyzing significantly differentially regulated pairs of TEs and host genes during aging, we identified tissue-specific loci with fundamental roles in synaptic signal transduction or critical immunological functions (Figure 6A). Notably, the recurrently affected Protocadherin, Keratin, and Skint genes are all organized in clusters. Moreover, Protocadherin and Keratin clusters exhibit remarkable evolutionary conservation [62, 82, 83]. The accumulation of co-regulated TEs in these regions poses the exciting question to which extent TEs facilitated their generation and still affect their regulation. One could speculate that a cluster of highly conserved genes is indeed an optimal pen to domesticate TEs and use their regulatory potential to orchestrate its expression. If so, the transposon would need to be reinstated as a controlling element.

Conclusion

In summary, our study demonstrates that the tissue-specific and independent expression of individual TEs in mice is strongly co-regulated with host genes. TEs with age-dependent expression dynamics are located in the neighborhood of genes with critical importance for the tissue function and marked relevance for aging phenotypes. We provide evidence that the Sox transcription factor family is a critical driver of TE expression - especially in brain tissue.

Methods

Mice

All mice were kept solely for aging until 24 months in a controlled environment and health status. Organs (brain, blood, skin) from 6- and 24-month-old C57BL/6JRj male mice were obtained from Janvier Labs.

Sequencing

Total RNA was extracted using the innuPREP RNA Mini Kit (Analytik Jena, Jena, Germany). Sequencing of RNA samples was performed using Illumina's next-generation sequencing methodology [84]. In detail, total RNA was quantified and quality checked using Tapestation 4200 Instrument in combination with RNA ScreenTape (both Agilent Technologies).

RNA-Seq libraries were prepared from 300 ng of input material (total RNA) using NEBNext Ultra II Directional RNA Library Preparation Kit in combination with NEBNext rRNA Depletion Kit (Human/Mouse/Rat) and NEBNext Multiplex Oligos for Illumina (Unique Dual Index UMI Adaptors RNA) following the manufacturer's instructions (New England Biolabs). Quantification and quality checked of libraries was done using an Agilent 4200 Tapestation Instrument and a DNA 1000 ScreenTape (Agilent Technologies). Libraries were pooled and sequenced on a NovaSeq 6000 using S1 300 cycle v1.5 reagents. System runs in 151 cycle/paired-end/standard loading workflow mode.

CAGE-Seq libraries were prepared from 1,700 - 5,000 ng of input material (total RNA) using CAGE Preparation Kit (Kabushiki Kaisha DNAFORM) following the manufacturer's instructions. For RNA derived from blood, pools of two or three samples were built up in order to achieve the quantity of 5,000 ng per library preparation reaction. Quantification and quality checked of libraries was done using an Agilent 2100 Bioanalyzer Instrument and a High Sensitivity DNA kit (Agilent Technologies). Libraries were pooled and sequenced on a NextSeq 500 using 75 cycle, high-output, v2.5 reagents. System runs in 81 cycle/single-end mode with spiking-in around 5 % of PhiX library (Illumina).

Sequence information was converted to FASTQ format using bcl2fastq (v2.20.0.422; default).

Quantification of gene and TE expression by RNA-Seq

Generation of SalmonTE reference index

TE sequences were extracted from the reference genome based on the RepeatMasker annotation of *Mus musculus* (mm10, based on Repeat Library 20140131, downloaded in January 2020, https://www.repeatmasker.org/genomes/mm10/RepeatMasker-rm405-db20140131/mm10.fa.align.gz) as described in [34] and combined with the gene annotation of *M. musculus* mm10 (v102 from http://ftp.ensembl.org/pub/release-102/fasta/mus_musculus/cdna/Mus_musculus.GRCm38.cdna.all.fa.gz). The Alu superfamily was relabeled to B1, as the Alu superfamily is the primate specific counterpart of the mouse specific B1. The generated sequence file served as input for the SalmonTE index generation with salmon (parameter: --type quasi -k 31) [85].

Alignment and expression quantification

Raw data was deduplicated for over-amplified PCR fragments based on uniqueness of read pair and UMI sequence. Reads were then mapped to the generated index using SalmonTE (v0.4) [33], with the expression measurement type was set to count (parameter: --exprtype=count). The expression matrix generated by SalmonTE was split-up; one for the genes and the other for TEs. The counts of the individual isoforms of a gene were summed-up to calculate the respective gene count. Features with less than or equal to ten reads in total across all samples were removed from the count matrices. DESeq2 (v1.34.0) [86] was separately applied to each count matrix to determine differentially expressed genes (DEGs) and differentially expressed TEs (DETEs). L2FC values were shrunken using the apeglm function [87] built into DESeq2. All TE instances which got an adjusted p-value assigned by DESeq2 were considered expressed in all downstream analyses. Aside from that, the raw counts were converted into transcripts per million (TPM; Equation 1-1) and subsequently scaled and centered for each gene to obtain z-scores (Equation 1-2).

$$TPM = \frac{Number\ of\ reads\ mapped\ to\ a\ gene * 10^3}{gene\ length\ in\ bp} * \frac{10^6}{\Sigma(\frac{Number\ of\ reads\ mapped\ to\ a\ gene * 10^3}{gene\ length\ in\ bp})}$$ **Equation 1-1**

$$z\ score = \frac{(TPM - \mu)}{\sigma}$$ **Equation 1-2**

Peak-calling and expression quantification by CAGE-Seq data

CAGE-seq captures transcripts with 5’-caps which are characterized by a methylated guanine. This guanine is appended to the mRNA right after its transcription and hence, it is not represented in the genomic sequence. Therefore, the raw reads were G-clipped with an in-house script. Then, we utilized Trimmomatic (v0.39) [88] (5nt sliding window approach, mean quality cutoff 20) for read quality trimming according to manual inspections of FastQC (v0.11.9) [89] reports. Cutadapt (v3.3) [90] was used to clip Illumina TruSeq adapter sequence from reads of young samples or Nextera adapter sequences from reads of old samples, respectively, as well as to clip mono- and di-nucleotide content. Subsequently, possible sequencing errors were detected and corrected using Rcorrector (v1.0.4) [91]. Further, ribosomal RNA (rRNA) transcripts were artificially depleted by read alignment against rRNA databases as performed by SortMeRNA (v2.1) [92]. The remaining high-quality reads were then aligned to the reference genome of *M. musculus* mm10 (v102 downloaded in January 2021 from ftp://ftp.ensembl.org/pub/release-102/fasta/mus_musculus/dna/). For this purpose, we used the splice-aware mapping software segemehl (v0.3.4) [93,94] with adjusted accuracy (95%). The resulting mappings of the young samples were filtered by samtools (v1.12) [95] for uniquely mapped reads. Brain samples were sequenced with a higher coverage than those of blood and skin, thus we performed downsampling of the brain samples to the level of skin using samtools (v.1.12). Finally, all sample-specific alignments were merged in a tissue-specific manner and then separated into forward and reverse aligned reads. PEAKachu (https://github.com/tbischler/PEAKachu, v0.2.0, default setting) was used to call strand-specific peaks in brain, blood, and skin. Called peaks with a distance ≤ 50 bp were merged with bedtools (v2.30.0-20-g484c0d4f-dirty) [96]. The genomic position at which most reads of a CAGE-peak start is defined as the TSS (transcription start site).

Start and end coordinates of peaks were used to extract read counts from the alignment files using featureCounts (v2.0.3) [97]. DESeq2 was applied to determine differential CAGE-peak expression. Intersection of peak coordinates with either gene or TE coordinates provides the gene or TE specific CAGE-peaks, respectively.

Enrichment analysis of TSSs in TEs

For each superfamily, the proportion of TSS-containing TEs belonging to this superfamily among all TSS-containing TEs (=target set) is calculated. The same was done for the genomic background of the TEs (background set). The ratio between the proportion of the superfamilies in the target and background set was calculated (odds ratio), and the standard binomial test was used to estimate the significance of the enrichment (corrected with Benjamini and Hochberg).

Analysis of TE regions

TE-regions and their characterization

TEs with a distance of 500 bp or less were merged to TE regions utilizing bedtools. The TE regions were 500 bp prolonged towards the 5′-end with bedtools. TE regions containing at least one expressed TE were defined as expressed TE regions. Such regions are further categorized into independently or dependently expressed TE region in case they either harbor a TSS or not, respectively. Independently expressed TE regions were categorized into single-, double-, and multi-TE regions according to the number of TEs that form the TE region. Only for multi-TE regions, the proportion of TE superfamilies at three positions of the regions was calculated, separately for each tissue. For the flanking positions (first/last), the single flanking TE was considered while the central position (i.e., body) was averaged over all remaining TEs of that TE region. The density of CAGE-peaks along the TE regions was calculated with deeptools (v.3.5.0) [98] (computeMatrix scale-regions; parameter: --missingDataAsZero, --afterRegionStartLength 1000, --regionBodyLength 2000, --beforeRegionStartLength 1000), whereas the score of each CAGE-peak was set to one.

Motif analysis of independently expressed TE regions

The sequence starting at the TSSs and extending to 500 bp up-stream of the TSSs is defined as the promoter region, thus TE regions with multiple TSSs contain multiple promoter sequences. HOMER [99] (findMotifsGenome.pl) was utilized to predict regulatory motifs within the promoter regions, using promoter coordinates and the reference genome (mm10 v102) as input. All genomic coordinates of Sox-motifs were extracted using scanMotifGenomeWide.pl from the HOMER suite. The intersection of Sox coordinates and TEs provided all individual TEs that contain at least one Sox motif. The relative positions of TSSs and Sox motifs within independent TE regions and their adjacent regions were determined using bedtools. In addition, the ratio of the amount of TSS and Sox-motifs at each relative position was calculated.

Overrepresentation analysis of GO terms

For expressed genes with at least one independently expressed TE region within their introns in sense direction and for all genes that do not have TEs in their introns (sense), separate GO term enrichment analyses were done with an in-house script (Fisher's exact test, corrected with Benjamini and Hochberg, significance cut-off at FDR ≤ 0.05). Only GO terms with 10 to 500 genes were considered in this analysis. Expressed genes were used as background in the first analysis, while all genes served as background for the second analysis.

TE enrichment in introns of genes

To test for enrichment of TEs within introns of genes, bootstrapping was performed as follows. The gene set of the GO term of interest represents the target set, while randomly selected expressed genes (of the same size as target set) represents the background. Then, one gene was drawn from each set, the numbers of TEs within introns were counted (restricted to the same strand) and the ratio was calculated (ratio = target_gene+0.1/background_gene+0.1). This procedure was repeated 1,000 times for each GO term.

RNA extraction and reverse transcription semi-quantitative real-time PCR (RT-qPCR)

Total cellular RNA was extracted using the innuPREP RNA Mini Kit (Analytik Jena, Jena, Germany) following the manufacturer's protocol. One-step reverse transcription and real-time PCR was performed with a

Quantstudio 5 using Power SYBR Green RNA-to-CT 1-Step Kit (Thermo Fisher Scientific, Waltham, USA) according to the manufacturer’s protocol. The following RT-qPCR primer sequences were used: Fam126a (forward: AGAGGTGTGAGCAGCAGGAT, reverse: TGCATTAGCAACCAGCAGAG), Fam126a-3’UTR (forward: GGGCTGCCTTCTGTACTTTG, reverse: ATGGCCAGTTCCAACAAGAC), MaLR_MTC (forward: CACCATGACCACAAGCTACG, reverse: GAACAAACCAGTGAGCAGCA).

Data Availability

Raw and processed data of RNA- and CAGE-Seq have been deposited in the Gene Expression Omnibus repository [100] and are accessible through GEO Series accession number GSE220773 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE220773). Full quantification results and positional data are stored at https://zenodo.org/ and are accessible via doi 10.5281/zenodo.7426786 (see Supplementary table 2).

Code Availability

All in-house scripts that were used in to analyze the data will be made available upon publication in a suited repository. All applied publicly available software is mentioned in the methods.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

The work was partially supported by the Klaus Tschira Stiftung (project number: 03.136.2018). The FLI is a member of the Leibniz Association and is financially supported by the Federal Government of Germany and the State of Thuringia.

Author Contributions

SH supervised the study. MG, MF and RS designed the Seq-experiments. SF performed the other experiments. RS and SH analyzed the data. SH, MF and RS interpret the data. All authors contributed to and approved the manuscript.

Acknowledgements

The Core Facility DNA sequencing (Ivonne Görlich, Cornelia Luge, and Martin Bens) of the FLI is gratefully acknowledged for their technical support in library preparation and sequencing. Furthermore, we thank Konstantin Riege and Arne Sahm for fruitful discussions and useful suggestions.

References

1. McClintock B. The origin and behavior of mutable loci in maize. Proc Natl Acad Sci U S A. 1950;36: 344–355.

2. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8: 973–982.

3. Fueyo R, Judd J, Feschotte C, Wysocka J. Roles of transposable elements in the regulation of mammalian transcription. Nat Rev Mol Cell Biol. 2022;23: 481–497.

4. Hermant C, Torres-Padilla M-E. TFs for TEs: the transcription factor repertoire of mammalian transposable elements. Genes Dev. 2021;35: 22–39.

5. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet. 2016;18: 71–86.

6. Almeida MV, Vernaz G, Putman ALK, Miska EA. Taming transposable elements in vertebrates: from epigenetic silencing to domestication. Trends Genet. 2022;38: 529–553.

7. Kong Y, Rose CM, Cass AA, Williams AG, Darwish M, Lianoglou S, et al. Transposable element expression in tumors is associated with immune infiltration and increased antigenicity. Nat Commun. 2019;10: 5228.

8. Criscione SW, Zhang Y, Thompson W, Sedivy JM, Neretti N. Transcriptional landscape of repetitive elements in normal and cancer human cells. BMC Genomics. 2014;15: 583.

9. Clayton EA, Wang L, Rishishwar L, Wang J, McDonald JF, Jordan IK. Patterns of Transposable Element Expression and Insertion in Cancer. Front Mol Biosci. 2016;3: 76.

10. Scott EC, Gardner EJ, Masood A, Chuang NT, Vertino PM, Devine SE. A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. Genome Res. 2016;26: 745–755.

11. Li W, Lee MH, Henderson L, Tyagi R, Bachani M, Steiner J, et al. Human endogenous retrovirus-K contributes to motor neuron disease. Sci Transl Med. 2015;7: 307ra153.

12. Erwin JA, Marchetto MC, Gage FH. Mobile DNA elements in the generation of diversity and complexity in the brain. Nat Rev Neurosci. 2014;15: 497–506.

13. De Cecco M, Criscione SW, Peterson AL, Neretti N, Sedivy JM, Kreiling JA. Transposable elements become active and mobile in the genomes of aging mammalian somatic tissues. Aging . 2013;5: 867–883.

14. Della Valle F, Reddy P, Yamamoto M, Liu P, Saera-Vila A, Bensaddek D, et al. RNA causes heterochromatin erosion and is a target for amelioration of senescent phenotypes in progeroid syndromes. Sci Transl Med. 2022;14: eabl6057.

15. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, et al. Ten things you should know about transposable elements. Genome Biol. 2018;19: 199.

16. Chen YG, Hur S. Cellular origins of dsRNA, their recognition and consequences. Nat Rev Mol Cell Biol. 2022;23: 286–301.

17. Simon M, Van Meter M, Ablaeva J, Ke Z, Gonzalez RS, Taguchi T, et al. LINE1 Derepression in Aged Wild-Type and SIRT6-Deficient Mice Drives Inflammation. Cell Metab. 2019;29: 871–885 e5.

18. De Cecco M, Ito T, Petrashen AP, Elias AE, Skvir NJ, Criscione SW, et al. L1 drives IFN in senescent cells and promotes age-associated inflammation. Nature. 2019;566: 73–78.

19. Zhao Y, Oreskovic E, Zhang Q, Lu Q, Gilman A, Lin YS, et al. Transposon-triggered innate immune response confers cancer resistance to the blind mole rat. Nat Immunol. 2021;22: 1219–1230.

20. Bedrosian TA, Quayle C, Novaresi N, Gage FH. Early life experience drives structural variation of neural genomes in mice. Science. 2018;359: 1395–1399.

21. Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, Lovci MT, et al. L1 retrotransposition in human neural progenitor cells. Nature. 2009;460: 1127–1131.

22. Muotri AR, Chu VT, Marchetto MCN, Deng W, Moran JV, Gage FH. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature. 2005;435: 903–910.

23. Perrat PN, DasGupta S, Wang J, Theurkauf W, Weng Z, Rosbash M, et al. Transposition-driven genomic heterogeneity in the Drosophila brain. Science. 2013;340: 91–95.

24. Upton KR, Gerhardt DJ, Jesuadian JS, Richardson SR, Sánchez-Luque FJ, Bodea GO, et al. Ubiquitous L1 Mosaicism in Hippocampal Neurons. Cell. 2015;161: 228–239.

25. Della Valle F, Thimma MP, Caiazzo M, Pulcrano S, Celii M, Adroub SA, et al. Transdifferentiation of Mouse Embryonic Fibroblasts into Dopaminergic Neurons Reactivates LINE-1 Repetitive Elements. Stem Cell Reports. 2020;14: 60–74.

26. Jönsson ME, Garza R, Sharma Y, Petri R, Sodersten E, Johansson JG, et al. Activation of endogenous retroviruses during brain development causes an inflammatory response. EMBO J. 2021;40: e106423.

27. Pehrsson EC, Choudhary MNK, Sundaram V, Wang T. The epigenomic landscape of transposable elements across normal human development and anatomy. Nat Commun. 2019;10: 1–16.

28. Cheng Y, Saville L, Gollen B, Isaac C, Belay A, Mehla J, et al. Increased processing of SINE B2 ncRNAs unveils a novel type of transcriptome deregulation in amyloid beta neuropathology. Elife. 2020;9. doi:10.7554/eLife.61265

29. Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, et al. The regulated retrotransposon transcriptome of mammalian cells. Nat Genet. 2009;41: 563–571.

30. Miao B, Fu S, Lyu C, Gontarz P, Wang T, Zhang B. Tissue-specific usage of transposable element-derived promoters in mouse development. Genome Biol. 2020;21: 255.

31. Lanciano S, Cristofari G. Measuring and interpreting transposable element expression. Nat Rev Genet. 2020;21: 721–736.

32. Philippe C, Vargas-Landin DB, Doucet AJ, van Essen D, Vera-Otarola J, Kuciak M, et al. Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. Elife. 2016;5. doi:10.7554/eLife.13926

33. Jeong H-H, Yalamanchili HK, Guo C, Shulman JM, Liu Z. An ultra-fast and scalable quantification pipeline for transposable elements from next generation sequencing data. 2018; 168–179.

34. Schwarz R, Koch P, Wilbrandt J, Hoffmann S. Locus-specific expression analysis of transposable elements. Brief Bioinform. 2022;23. doi:10.1093/bib/bbab417

35. De Cecco M, Criscione SW, Peckham EJ, Hillenmeyer S, Hamm EA, Manivannan J, et al. Genomes of replicatively senescent cells undergo global epigenetic changes leading to gene silencing and activation of transposable elements. Aging Cell. 2013;12: 247–256.

36. Van Meter M, Kashyap M, Rezazadeh S, Geneva AJ, Morello TD, Seluanov A, et al. SIRT6 represses LINE1 retrotransposons by ribosylating KAP1 but this repression fails with stress and age. Nat Commun. 2014;5: 1–10.

37. Green CD, Huang Y, Dou X, Yang L, Liu Y, Han J-DJ. Impact of Dietary Interventions on Noncoding RNA Networks and mRNAs Encoding Chromatin-Related Factors. Cell Rep. 2017;18: 2957–2968.

38. Voliva CF, Jahn CL, Comer MB, Hutchison CA 3rd, Edgell MH. The L1Md long interspersed repeat family in the mouse: almost all examples are truncated at one end. Nucleic Acids Res. 1983;11: 8847–8859.

39. Zhou M, Smith AD. Subtype classification and functional annotation of L1Md retrotransposon promoters. Mob DNA. 2019;10: 14.

40. Yasuhiko Y, Hirabayashi Y, Ono R. LTRs of Endogenous Retroviruses as a Source of Tbx6 Binding Sites. Front Chem. 2017;5: 34.

41. Bakoulis S, Krautz R, Alcaraz N, Salvatore M, Andersson R. Endogenous retroviruses co-opted as divergently transcribed regulatory elements shape the regulatory landscape of embryonic stem cells. Nucleic Acids Res. 2022;50: 2111–2127.

42. Stein D, Mizrahi A, Golova A, Saretzky A, Venzor AG, Slobodnik Z, et al. Aging and pathological aging signatures of the brain: through the focusing lens of SIRT6. Aging . 2021;13: 6420–6441.

43. Hsieh PN, Sweet DR, Fan L, Jain MK. Aging and the Krüppel-like factors. Trends Cell Mol Biol. 2017;12: 1–15.

44. Zagorska A, Partyka A, Bucki A, Gawalskax A, Czopek A, Pawlowski M. Phosphodiesterase 10 Inhibitors - Novel Perspectives for Psychiatric and Neurodegenerative Drug Discovery. Curr Med Chem. 2018;25: 3455–3481.

45. Gualandi N, Iperi C, Esposito M, Ansaloni F, Gustincich S, Sanges R. Meta-Analysis Suggests That Intron Retention Can Affect Quantification of Transposable Elements from RNA-Seq Data. Biology . 2022;11. doi:10.3390/biology11060826

46. Boivin V, Deschamps-Francoeur G, Scott MS. Protein coding genes as hosts for noncoding RNA expression. Semin Cell Dev Biol. 2018;75: 3–12.

47. Takahashi H, Lassmann T, Murata M, Carninci P. 5′ end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. Nat Protoc. 2012;7: 542–561.

48. Lima-Junior DS, Krishnamurthy SR, Bouladoux N, Collins N, Han S-J, Chen EY, et al. Endogenous retroviruses promote homeostatic and inflammatory responses to the microbiota. Cell. 2021;184: 3794-3811.e19.

49. Ye M, Goudot C, Hoyler T, Lemoine B, Amigorena S, Zueva E. Specific subfamilies of transposable elements contribute to different domains of T lymphocyte enhancers. Proc Natl Acad Sci U S A. 2020;117: 7905–7916.

50. Xie M, Hong C, Zhang B, Lowdon RF, Xing X, Li D, et al. DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. Nat Genet. 2013;45: 836–841.

51. Todd CD, Deniz Ö, Taylor D, Branco MR. Functional evaluation of transposable elements as enhancers in mouse embryonic and trophoblast stem cells. 2019 [cited 11 Oct 2022]. doi:10.7554/eLife.44344

52. Morrish TA, Gilbert N, Myers JS, Vincent BJ, Stamato TD, Taccioli GE, et al. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nat Genet. 2002;31: 159–165.

53. Suzuki J, Yamaguchi K, Kajikawa M, Ichiyanagi K, Adachi N, Koyama H, et al. Genetic Evidence That the Non-Homologous End-Joining Repair Pathway Is Involved in LINE Retrotransposition. PLoS Genet. 2009;5: e1000461.

54. A.F.A. Smit RH&. PG. RepeatMasker http://repeatmasker.org. 1996.

55. Enriquez-Gasca R, Gould PA, Rowe HM. Host Gene Regulation by Transposable Elements: The New, the Old and the Ugly. Viruses. 2020;12. doi:10.3390/v12101089

56. Fasching L, Kapopoulou A, Sachdeva R, Petri R, Jönsson ME, Männe C, et al. TRIM28 represses transcription of endogenous retroviruses in neural progenitor cells. Cell Rep. 2015;10: 20–28.

57. Brattås PL, Jönsson ME, Fasching L, Nelander Wahlestedt J, Shahsavani M, Falk R, et al. TRIM28 Controls a Gene Regulatory Network Based on Endogenous Retroviruses in Human Neural Progenitor Cells. Cell Rep. 2017;18: 1–11.

58. Barbee SD, Woodward MJ, Turchinovich G, Mention J-J, Lewis JM, Boyden LM, et al. Skint-1 is a highly specific, unique selecting component for epidermal T cells. Proc Natl Acad Sci U S A. 2011;108: 3330–3335.

59. Narita T, Nitta T, Nitta S, Okamura T, Takayanagi H. Mice lacking all of the Skint family genes. Int Immunol. 2018;30: 301–309.

60. Nielsen MM, Witherden DA, Havran WL. γδ T cells in homeostasis and host defence of epithelial barrier tissues. Nat Rev Immunol. 2017;17: 733–745.

61. Wang Z, Man M-Q, Li T, Elias PM, Mauro TM. Aging-associated alterations in epidermal function and their clinical significance. Aging . 2020;12: 5551–5565.

62. Ho M, Thompson B, Fisk JN, Nebert DW, Bruford EA, Vasiliou V, et al. Update of the keratin gene family: evolution, tissue-specific expression patterns, and relevance to clinical disorders. Hum Genomics. 2022;16: 1.

63. Ge Y, Miao Y, Gur-Cohen S, Gomez N, Yang H, Nikolova M, et al. The aging skin microenvironment dictates stem cell behavior. Proc Natl Acad Sci U S A. 2020;117: 5339–5350.

64. Brambilla R, Gnesutta N, Minichiello L, White G, Roylance AJ, Herron CE, et al. A role for the Ras signalling pathway in synaptic transmission and long-term memory. Nature. 1997;390: 281–286.

65. Umeda K, Negishi M, Katoh H. RasGRF1 mediates brain-derived neurotrophic factor-induced axonal growth in primary cultured cortical neurons. Biochem Biophys Rep. 2019;17: 56–64.

66. Cifelli JL, Berg KR, Yang J. Benzothiazole amphiphiles promote RasGRF1-associated dendritic spine formation in human stem cell-derived neurons. FEBS Open Bio. 2020;10: 386–395.

67. Borrás C, Monleón D, López-Grueso R, Gambini J, Orlando L, Pallardó FV, et al. RasGrf1 deficiency delays aging in mice. Aging . 2011;3: 262–276.

68. Pérez-González M, Mendioroz M, Badesso S, Sucunza D, Roldan M, Espelosín M, et al. PLA2G4E, a candidate gene for resilience in Alzheimer´s disease and a new target for dementia treatment. Prog Neurobiol. 2020;191: 101818.

69. Wu Q, Maniatis T. A striking organization of a large family of human neural cadherin-like cell adhesion genes. Cell. 1999;97: 779–790.

70. Lefebvre JL, Sanes JR, Kay JN. Development of dendritic form and function. Annu Rev Cell Dev Biol. 2015;31: 741–777.

71. Zipursky SL, Grueber WB. The molecular basis of self-avoidance. Annu Rev Neurosci. 2013;36: 547–568.

72. Canzio D, Nwakeze CL, Horta A, Rajkumar SM, Coffey EL, Duffy EE, et al. Antisense lncRNA Transcription Mediates DNA Demethylation to Drive Stochastic Protocadherin α Promoter Choice. Cell. 2019;177: 639–653.e15.

73. Casale AM, Liguori F, Ansaloni F, Cappucci U, Finaurini S, Spirito G, et al. Transposable element activation promotes neurodegeneration in a Drosophila model of Huntington’s disease. iScience. 2022;25: 103702.

74. Ravel-Godreuil C, Znaidi R, Bonnifet T, Joshi RL, Fuchs J. Transposable elements as new players in neurodegenerative diseases. FEBS Lett. 2021;595: 2733–2755.

75. Burns KH. Transposable elements in cancer. Nat Rev Cancer. 2017;17: 415–424.

76. Franceschi C, Capri M, Monti D, Giunta S, Olivieri F, Sevini F, et al. Inflammaging and anti-inflammaging: a systemic perspective on aging and longevity emerged from studies in humans. Mech Ageing Dev. 2007;128: 92–105.

77. Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, et al. Somatic retrotransposition alters the genetic landscape of the human brain. Nature. 2011;479: 534–537.

78. Singer T, McConnell MJ, Marchetto MC, Coufal NG, Gage FH. LINE-1 retrotransposons: mediators of somatic variation in neuronal genomes? Trends Neurosci. 2010;33: 345–354.

79. Lai T, Jabaudon D, Molyneaux BJ, Azim E, Arlotta P, Menezes JRL, et al. SOX5 controls the sequential generation of distinct corticofugal neuron subtypes. Neuron. 2008;57: 232–247.

80. Li L, Medina-Menéndez C, García-Corzo L, Córdoba-Beldad CM, Quiroga AC, Calleja Barca E, et al. SoxD genes are required for adult neural stem cell activation. Cell Rep. 2022;38: 110313.

81. Parikshak NN, Swarup V, Belgard TG, Irimia M, Ramaswami G, Gandal MJ, et al. Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. Nature. 2016;540: 423–427.

82. Zimek A, Weber K. The organization of the keratin I and II gene clusters in placental mammals and marsupials show a striking similarity. Eur J Cell Biol. 2006;85: 83–89.

83. Wu Q, Zhang T, Cheng JF, Kim Y, Grimwood J, Schmutz J, et al. Comparative DNA sequence analysis of mouse and human protocadherin gene clusters. Genome Res. 2001;11: 389–404.

84. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456: 53–59.

85. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14: 417–419.

86. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15: 550.

87. Zhu A, Ibrahim JG, Love MI. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. Bioinformatics. 2019;35: 2084–2092.

88. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30: 2114–2120.

89. Andrews S. FastQC A quality control tool for high throughput sequence data. 2010 [cited 7 Nov 2022]. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc

90. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17: 10–12.

91. Song L, Florea L. Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. Gigascience. 2015;4: 48.

92. Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics. 2012;28: 3211–3217.

93. Hoffmann S, Otto C, Doose G, Tanzer A, Langenberger D, Christ S, et al. A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. Genome Biol. 2014;15: R34.

94. Hoffmann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, Vogel J, et al. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comput Biol. 2009;5: e1000502.

95. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079.

96. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26: 841–842.

97. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30: 923–930.

98. Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 2016;44: W160–5.

99. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38: 576–589.

100. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30: 207–210.

Supplemental Tables

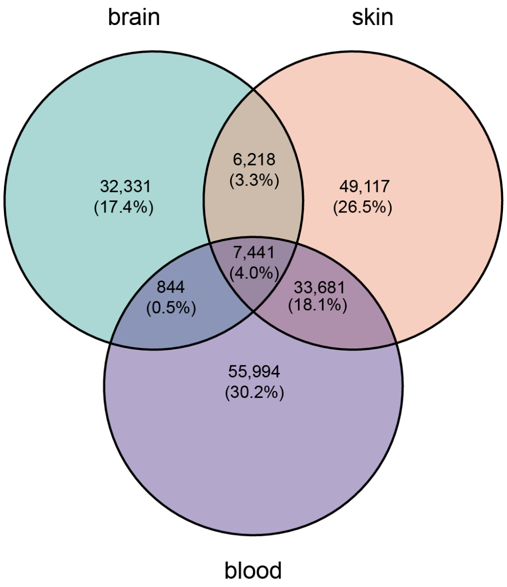
Supplemental Table 1 – Counts of detected (differentially) expressed TEs comparing 24 months and 6 months old male mice in different tissues (brain, skin, blood; rows) and counts of TEs that intersect with a CAGE transcription start site (CTSS). The total counts are further separated into TE counts for specific positions (columns). **Abbreviations:** TE – transposable element; TSS – transcription start site.

			Count at specific position (proportion of total count in %)				
	tissue	**total** (proportion of genomic TEs in %)	promoter (500 bp)	exon	intron	down-stream located (500 bp)	intergenic
expressed (RNA-Seq)	brain	**46,834** (1.11)	249 (0.53)	457 (0.98)	17,031 (36.36)	342 (0.73)	28,755 (61.40)
	skin	**96,457** (2.30)	563 (0.58)	1125 (1.17)	34,582 (35.85)	975 (1.01)	59,212 (61.30)
	blood	**97,960** (2.33)	504 (0.51)	799 (0.82)	37,122 (37.90)	852 (0.87)	58,683 (59.91)
differentially expressed (RNA-Seq)	brain	**135** (0.003)	1 (0.74)	4 (2.96)	43 (31.85)	2 (1.48)	85 (62.96)
	skin	**1,048** (0.02)	7 (0.67)	39 (3.72)	325 (31.01)	24 (2.20)	653 (62.31)
	blood	**55** (0.001)	1 (1.82)	1 (1.82)	15 (27.27)	1 (1.82)	37 (67.27)
CTSS intersected TEs (CAGE-Seq)	brain	**61,476** (1.46)	726 (1.18)	2,681 (4.36)	50,374 (81.94)	672 (1.09)	7,023 (11.42)
	skin	**8,572** (0.20)	455 (5.31)	1,281 (14.94)	3,766 (43.93)	166 (1.94)	2,904 (33.88)
	blood	**3,774** (0.09)	182 (4.82)	247 (6.54)	2,410 (63.86)	81 (2.15)	854 (22.63)

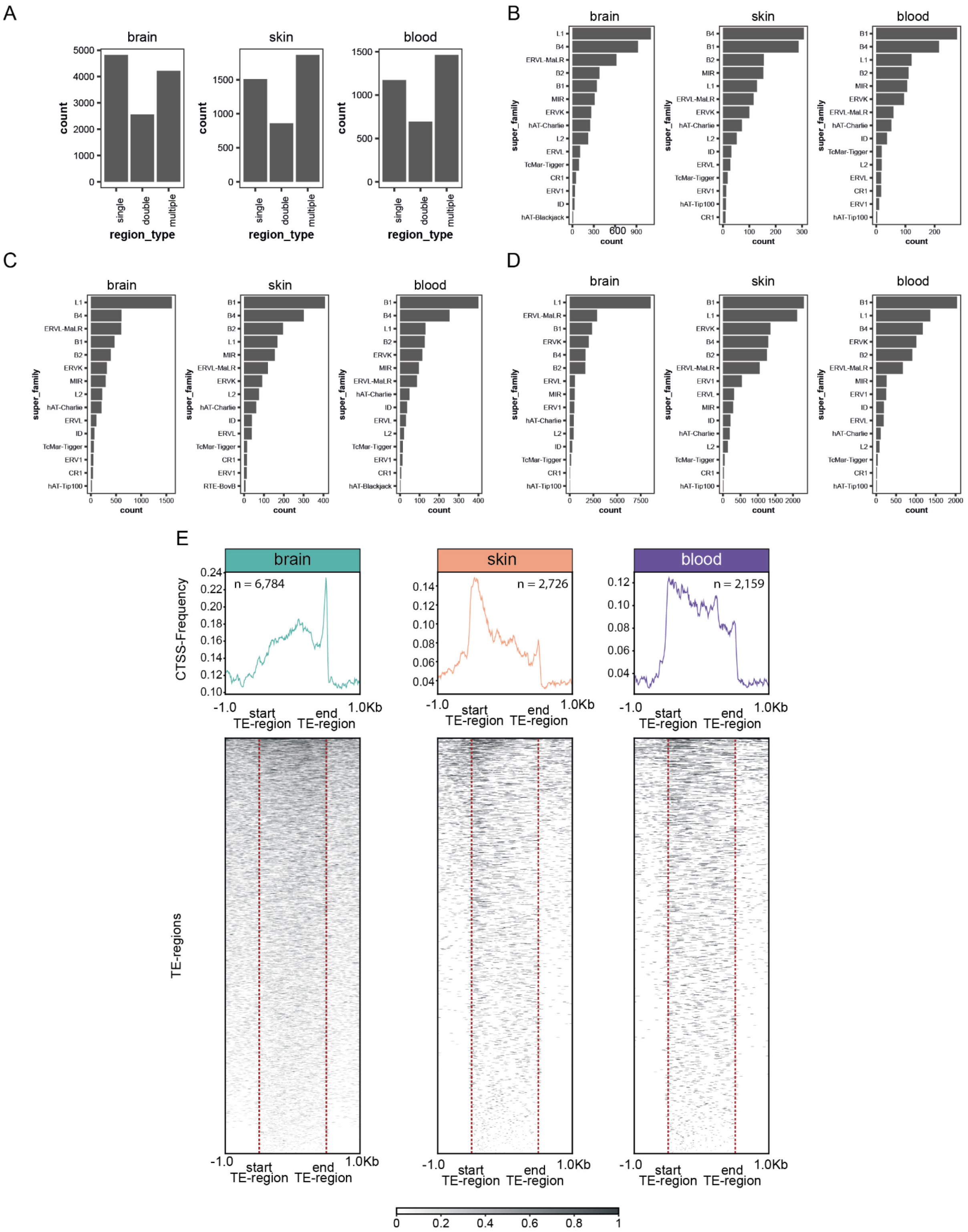
Supplemental Table 2 – Overview of additional files stored at https://zenodo.org/ (doi: 10.5281/zenodo.7426786)

File	**Description**
01_quantification_TEs.csv	DESeq2 results for individual TEs based on RNA-Seq
02_quantification_TE_region.csv	DESeq2 results for individual TE regions based on RNA-Seq
03_CAGE_quantification_TEs.csv	DESeq2 results for individual TEs based on CAGE-Seq
04_CAGE_quantification_TEs_region.csv	DESeq2 results for individual TE regions based on CAGE-Seq
TE_regions.bed	Annotation of TE regions in .bed format.

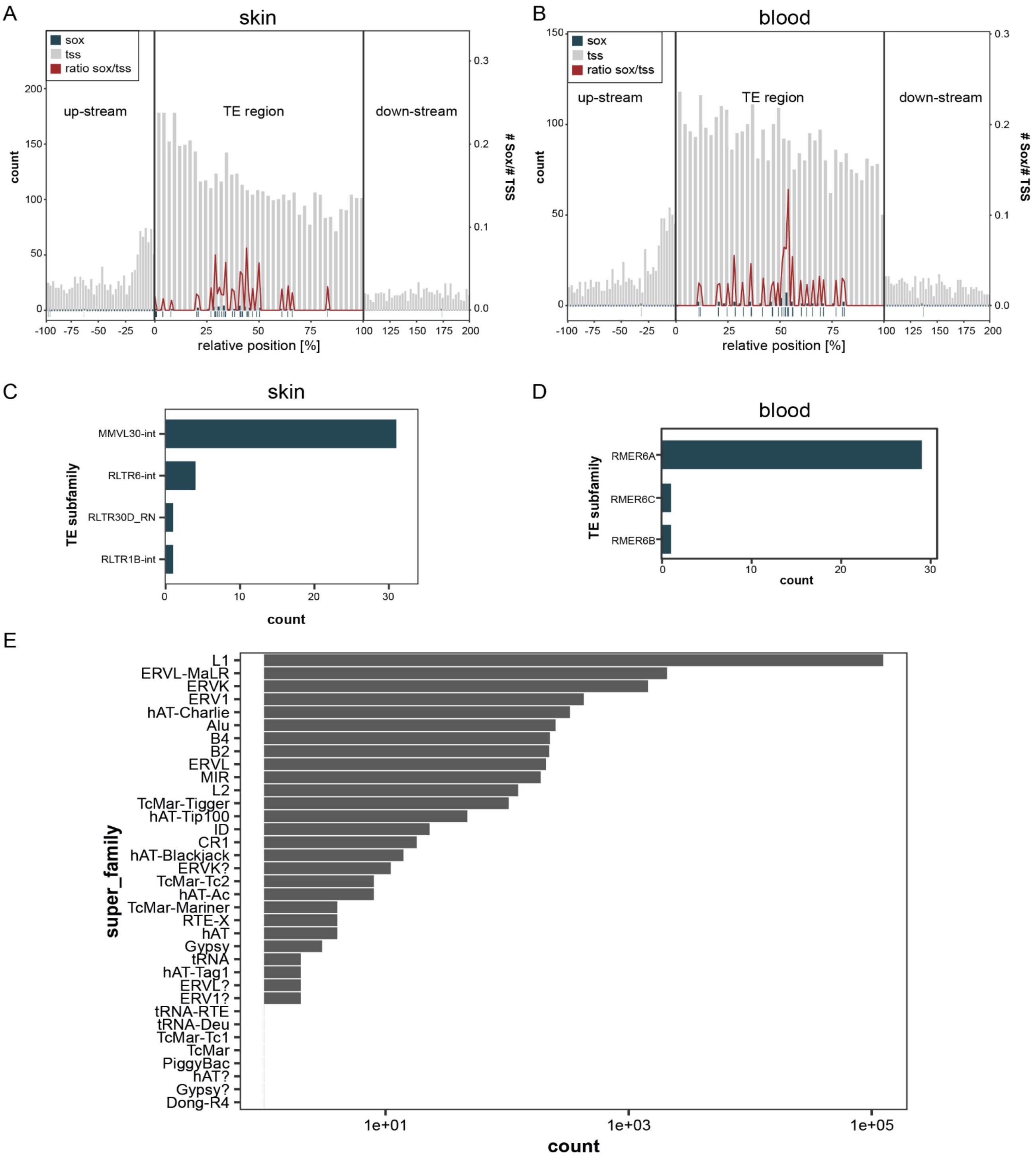
Supplemental Figures



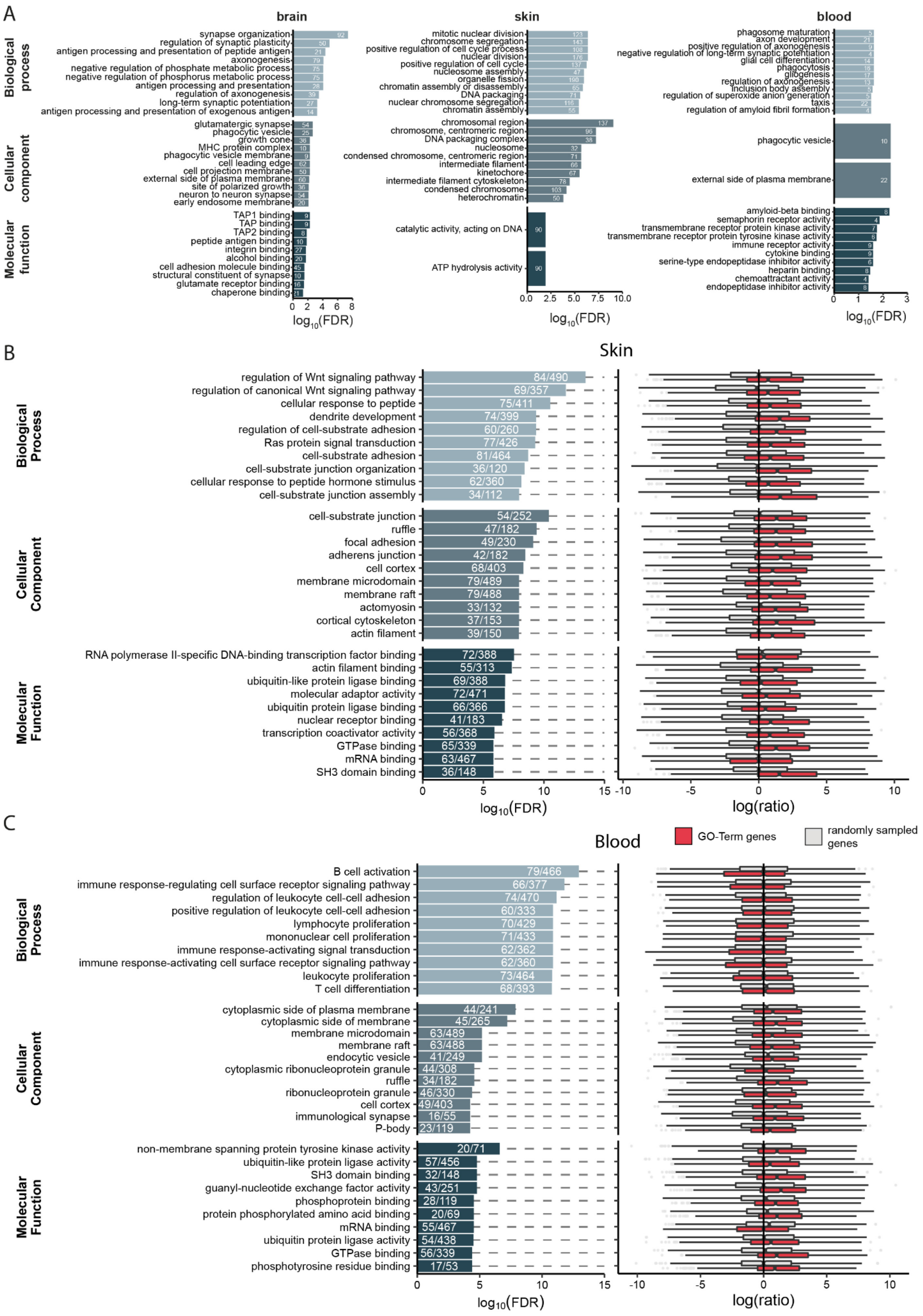
Supplemental Figure 1 – Tissue-specific expression of TEs. The Venn diagram shows the intersection of expressed TEs detected by RNA-Seq in brain (blue), skin (orange), and blood (purple) as counts and percentages. While the minority is expressed in multiple tissues, the absolute majority is expressed exclusively in one tissue. **Abbreviations:** TE – transposable element.



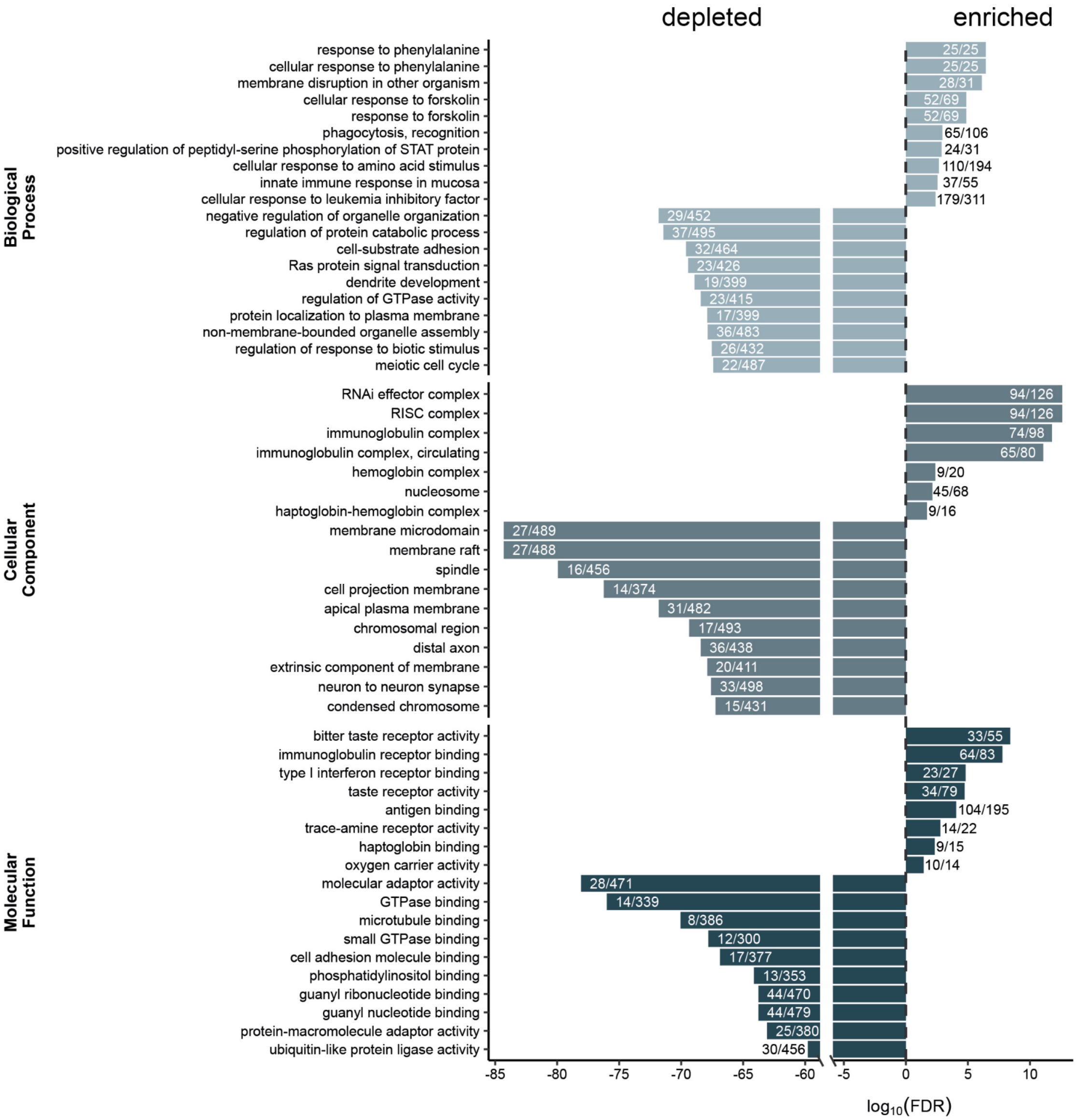
Supplemental Figure 2 - Characterizing the TE regions. TEs in close proximity (distance ≤ 500 bp) are merged to TE regions. TE regions are classified by the number of individual TEs that form the TE region (region type; single, double, and multi). **(A)** Region type (single, double, multi) count of independently expressed TE regions for blood, brain, and skin. **(B-D)** Counts of individual TEs (superfamily level) that are contained in single **(B)**, double **(C)** and multiple-TE regions **(D)**. **(E)** CAGE-Seq peak frequency across all multiple TE regions in brain (green), skin (orange), and blood (purple) and their adjacent areas (≤ 1Kb). The frequencies result from the peak count (greyscale) across the multiple TE regions as depicted below the frequency plots. **Abbreviations:** bp – base-pair; CAGE – Cap-analysis gene expression sequencing; FDR – False discovery rate; Kb - kilo-base; L2FC – log2(fold change); TE – transposable element; TSS - transcription start site.



Supplemental Figure 3 - Sox-Motif intersection with individual TE elements. (A-B) Counts of TSSs (gray) and Sox motifs (blue) and their ratio (red) across TE regions and their adjacent areas (up- and down-stream; ≤ 500 bp) in skin **(A)** and blood **(B).** The red line indicates the ratio between counts of Sox-motif and TSSs at each relative position in TE regions (right y-axis). The blue vertical lines under the y-axis indicate the relative position of the Sox motif within independent TE regions and their adjacent areas. **(C-D)** Counts of individual TE instances at the superfamily level within TE regions that intersect with a Sox motif in skin **(C)** and blood **(D). (E)** Counts of individual TE instances (superfamily level) that intersect with a predicted Sox motif in a genome wide view in brain. **Abbreviations:** bp – base-pair; TE – transposable element; TSS - transcription start site.

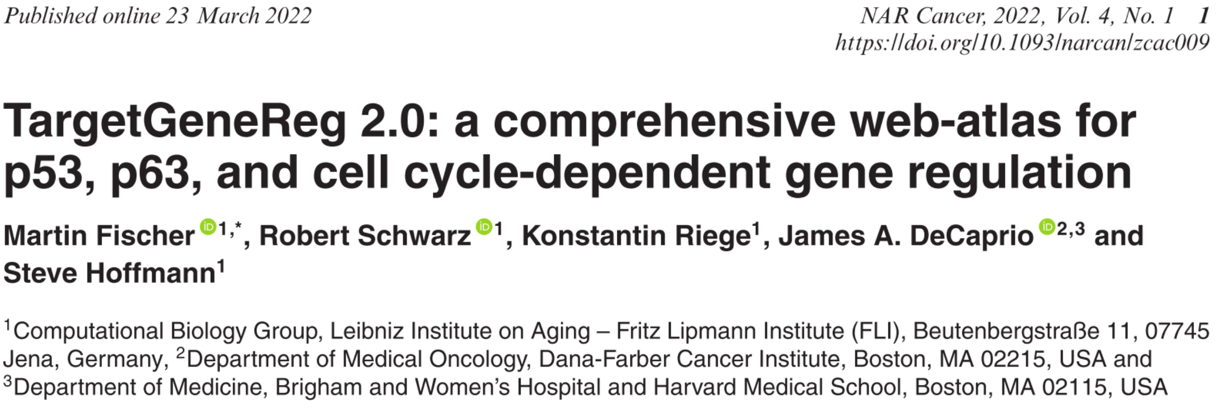

A
brain
skin
blood
Biological process
Cellular component
Molecular function
synapse organization 92
regulation of synaptic plasticity 50
antigen processing and presentation of peptide antigen 21
axonogenesis 79
negative regulation of phosphate metabolic process 75
negative regulation of phosphorus metabolic process 75
antigen processing and presentation 28
regulation of axonogenesis 39
long-term synaptic potentiation 27
antigen processing and presentation of exogenous antigen 14
glutamatergic synapse 54
phagocytic vesicle 25
growth cone 36
MHC protein complex 10
phagocytic vesicle membrane 9
cell leading edge 62
cell projection membrane 50
external side of plasma membrane 60
site of polarized growth 36
neuron to neuron synapse 54
early endosome membrane 20
TAP1 binding 9
TAP binding 9
TAP2 binding 8
peptide antigen binding 10
integrin binding 27
alcohol binding 20
cell adhesion molecule binding 45
structural constituent of synapse 10
glutamate receptor binding 16
chaperone binding 21
0 2 4 6 8
log10(FDR)
mitotic nuclear division 123
chromosome segregation 143
positive regulation of cell cycle process 108
nuclear division 176
positive regulation of cell cycle 137
nucleosome assembly 47
organelle fission 190
chromatin assembly or disassembly 65
DNA packaging 71
nuclear chromosome segregation 116
chromatin assembly 55
chromosomal region 137
chromosome, centromeric region 96
DNA packaging complex 38
nucleosome 32
condensed chromosome, centromeric region 71
intermediate filament 66
kinetochore 67
intermediate filament cytoskeleton 78
condensed chromosome 103
heterochromatin 50
catalytic activity, acting on DNA 90
ATP hydrolysis activity 90
0.0 2.5 5.0 7.5 10.0
log10(FDR)
phagosome maturation 5
axon development 21
positive regulation of axonogenesis 9
negative regulation of long-term synaptic potentiation 4
glial cell differentiation 14
phagocytosis 16
gliogenesis 17
regulation of axonogenesis 13
inclusion body assembly 5
regulation of superoxide anion generation 5
taxis 22
regulation of amyloid fibril formation 4
phagocytic vesicle 10
external side of plasma membrane 22
amyloid-beta binding 8
semaphorin receptor activity 4
transmembrane receptor protein kinase activity 7
transmembrane receptor protein tyrosine kinase activity 6
immune receptor activity 9
cytokine binding 9
serine-type endopeptidase inhibitor activity 6
heparin binding 8
chemoattractant activity 4
endopeptidase inhibitor activity 8
0 1 2 3
log10(FDR)
B
Skin
Biological Process
regulation of Wnt signaling pathway 84/490
regulation of canonical Wnt signaling pathway 69/357
cellular response to peptide 75/411
dendrite development 74/399
regulation of cell-substrate adhesion 60/260
Ras protein signal transduction 77/426
cell-substrate adhesion 81/464
cell-substrate junction organization 36/120
cellular response to peptide hormone stimulus 62/360
cell-substrate junction assembly 34/112
Cellular Component
cell-substrate junction 54/252
ruffle 47/182
focal adhesion 49/230
adherens junction 42/182
cell cortex 68/403
membrane microdomain 79/489
membrane raft 79/488
actomyosin 33/132
cortical cytoskeleton 37/153
actin filament 39/150
Molecular Function
RNA polymerase II-specific DNA-binding transcription factor binding 72/388
actin filament binding 55/313
ubiquitin-like protein ligase binding 69/388
molecular adaptor activity 72/471
ubiquitin protein ligase binding 66/366
nuclear receptor binding 41/183
transcription coactivator activity 56/368
GTPase binding 65/339
mRNA binding 63/467
SH3 domain binding 36/148
0 5 10 15
log10(FDR)
-10 -5 0 5 10
log(ratio)
C
Blood
GO-Term genes
randomly sampled genes
Biological Process
B cell activation 79/466
immune response-regulating cell surface receptor signaling pathway 66/377
regulation of leukocyte cell-cell adhesion 74/470
positive regulation of leukocyte cell-cell adhesion 60/333
lymphocyte proliferation 70/429
mononuclear cell proliferation 71/433
immune response-activating signal transduction 62/362
immune response-activating cell surface receptor signaling pathway 62/360
leukocyte proliferation 73/464
T cell differentiation 68/393
Cellular Component
cytoplasmic side of plasma membrane 44/241
cytoplasmic side of membrane 45/265
membrane microdomain 63/489
membrane raft 63/488
endocytic vesicle 41/249
cytoplasmic ribonucleoprotein granule 44/308
ruffle 34/182
ribonucleoprotein granule 46/330
cell cortex 49/403
immunological synapse 16/55
P-body 23/119
Molecular Function
non-membrane spanning protein tyrosine kinase activity 20/71
ubiquitin-like protein ligase activity 57/456
SH3 domain binding 32/148
guanyl-nucleotide exchange factor activity 43/251
phosphoprotein binding 28/119
protein phosphorylated amino acid binding 20/69
mRNA binding 55/467
ubiquitin protein ligase activity 54/438
GTPase binding 56/339
phosphotyrosine residue binding 17/53
0 5 10 15
log10(FDR)
-10 -5 0 5 10
log(ratio)

Supplemental Figure 4 - GO term analyses of differentially expressed genes and genes containing independently expressed TEs. (A) Top 10 GO terms (sorted by FDR) of Biological Process, Cellular Component, and Molecular Function of differentially expressed genes (background all expressed genes) in brain, skin and blood. The x-axis shows the significance ($\log_{10}$(FDR); Benjamini-Hochberg corrected) for each GO term (y-axis), while the numbers in each bar represent the count of DEGs. **(B)** On the left side, top 10 GO terms (sorted by FDR) of Biological Process, Cellular Component, and Molecular Function where genes overlapping with independently expressed TE regions in skin are enriched (background all detected genes in skin). The x-axis shows the significance ($\log_{10}$(FDR); Benjamini-Hochberg corrected) for each GO term (y-axis), while the numbers in each bar represent the count of genes overlapped by independent TE regions and the count of genes within each GO term. On the right side, ratio of counts of intronic TEs in gene set of interest (red = GO term genes; gray = randomly sampled set with same size of GO term gene set) and a randomly selected gene for the gene set of expressed gene in skin. For each GO term, one gene was drawn from each set and the ratio was calculated, which was repeated 1000 times (content of one box). The box plot center line represents the median, the upper and lower bounds correspond to the first and third quartiles, and the whiskers reach to 1.5 times the interquartile range. **(C)** on the left side, top 10 GO terms (sorted by FDR) of Biological Process, Cellular Component, and Molecular Function where genes overlapping with independently expressed TE regions in blood are enriched (background all detected genes in blood). The x-axis shows the significance ($\log_{10}$(FDR); Benjamini-Hochberg corrected) for each GO term (y-axis), while the numbers in each bar represent the count of genes overlapped by independent TE regions and the count of genes within each GO term. On the right side, ratio of counts of intronic TEs in gene set of interest (red = GO term genes; gray = randomly sampled set with same size of GO term gene set) and a randomly selected gene for the gene set of expressed gene in blood. For each GO term, one gene was drawn from each set and the ratio was calculated, which was repeated 1000 times (content of one box). The box plot center line represents the median, the upper and lower bounds correspond to the first and third quartiles, and the whiskers reach to 1.5 times the interquartile range. **Abbreviations:** FDR – false discovery rate; GO – Gene Ontology; TE – transposable elements.



Supplemental Figure 5 - GO term enrichment analysis of TE-free genes. Top 10 GO terms (sorted by FDR) of enriched and depleted TE-free genes in GO terms of the ontologies Biological Process, Cellular Component, and Molecular Function. The x-axis shows the significance ($\log_{10}$(FDR); Benjamini-Hochberg corrected) for each GO term (y-axis), while the numbers in each bar represent the count of TE-free genes and the count of genes within each GO term. **Abbreviations:** FDR – false discovery rate; GO – Gene Ontology; TE – transposable elements.

2.4. Manuscript 3 (M3) – TargetGeneReg 2.0: a comprehensive web-atlas for p53, p63, and cell cycle-dependent gene regulation



Published online 23 March 2022 *NAR Cancer, 2022, Vol. 4, No. 1* ***1***
https://doi.org/10.1093/narcan/zcac009

TargetGeneReg 2.0: a comprehensive web-atlas for p53, p63, and cell cycle-dependent gene regulation

Martin Fischer[1,*], Robert Schwarz[1], Konstantin Riege[1], James A. DeCaprio[2,3] and Steve Hoffmann[1]

[1]Computational Biology Group, Leibniz Institute on Aging – Fritz Lipmann Institute (FLI), Beutenbergstraße 11, 07745 Jena, Germany, [2]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA and [3]Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA

Summary:

The combined analysis of multiple datasets provides a valuable resource that fully realizes its power through public accessibility. This study provides a comprehensive web-atlas for p53, p63 and cell cycle dependent gene regulation created by analyzing datasets from multiple studies. In this project, I built a suitable data structure that allows both storage of the complex data and convenient accessibility. In addition, I designed and developed a website to make the data available to the public.

Overview:

Manuscript No. 3

Manuscript title: TargetGeneReg 2.0: a comprehensive web-atlas for p53, p63, and cell cycle-dependent gene regulation

Authors: Fischer M., Schwarz R., Riege K., Decaprio J., Hoffmann S.

Bibliographic information:

Martin Fischer, Robert Schwarz, Konstantin Riege, James A DeCaprio, Steve Hoffmann, TargetGeneReg 2.0: a comprehensive web-atlas for p53, p63, and cell cycle-dependent gene regulation, *NAR Cancer*, Volume 4, Issue 1, March 2022, zcac009, https://doi.org/10.1093/narcan/zcac009

The candidate is:

☐ First author, ☐ Co-first author, ☐ Corresponding author, ☒ Co-author.

Status: published

Authors' contribution (in %) to the given categories of the publication:

Author	**Conceptual**	**Data analysis**	**Writing the manuscript**
Fischer M.	100%		80%
Schwarz R.		50%	
Riege K.		50%	
Hoffmann S.			10%
DeCaprio J.			10%
Total:	100%	100%	100%

Published online 23 March 2022

NAR Cancer, 2022, Vol. 4, No. 1 **1**
https://doi.org/10.1093/narcan/zcac009

TargetGeneReg 2.0: a comprehensive web-atlas for p53, p63, and cell cycle-dependent gene regulation

Martin Fischer [1,*], Robert Schwarz [1], Konstantin Riege[1], James A. DeCaprio [2,3] and Steve Hoffmann[1]

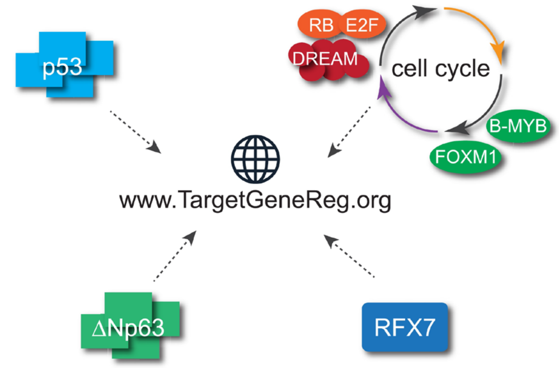
[1]Computational Biology Group, Leibniz Institute on Aging – Fritz Lipmann Institute (FLI), Beutenbergstraße 11, 07745 Jena, Germany, [2]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA and [3]Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA

Received December 09, 2021; Revised February 03, 2022; Editorial Decision March 04, 2022; Accepted March 07, 2022

ABSTRACT

In recent years, our web-atlas at www.TargetGeneReg.org has enabled many researchers to uncover new biological insights and to identify novel regulatory mechanisms that affect p53 and the cell cycle – signaling pathways that are frequently dysregulated in diseases like cancer. Here, we provide a substantial upgrade of the database that comprises an extension to include non-coding genes and the transcription factors ΔNp63 and RFX7. TargetGeneReg 2.0 combines gene expression profiling and transcription factor DNA binding data to determine, for each gene, the response to p53, ΔNp63, and cell cycle signaling. It can be used to dissect common, cell type and treatment-specific effects, identify the most promising candidates, and validate findings. We demonstrate the increased power and more intuitive layout of the resource using realistic examples.

GRAPHICAL ABSTRACT



INTRODUCTION

The cell proliferation cycle and the tumor suppressor p53 are closely linked and represent the most commonly dysregulated signaling pathways in cancer. Despite more than 40 years of research on p53 and many more on the cell cycle, we still lack a comprehensive understanding of the p53 and the cell cycle-dependent regulation of a surprisingly large number of genes. Several mechanisms have been proposed to explain the temporal regulation of hundreds of cell cycle genes (1,2) and the downstream targets of p53 (3–5), but the substantial overlap between p53 and the cell cycle render the analysis of individual genes difficult.

The expansion of publicly available high-throughput datasets has enabled a more detailed understanding of gene regulatory mechanisms and networks in recent years. We developed a meta-analysis approach to cross-validate results and to improve statistical power by integrating datasets derived from different experimental setups (6). The meta-analysis allows inferring p53 and cell cycle regulation of genes from multiple cell types and treatment conditions and derive common signature genes. It follows the intuitive idea that when multiple independent datasets agree on a finding it is more likely to be accurate and that the sum of available evidence provides the best guess for the truth. Previously, we employed this meta-analysis approach to chart the transcriptional programs of the cell cycle, human and mouse p53, the viral oncoprotein E7, and the transcription factor ΔNp63 (6–10). Key findings from these meta-analyses included that p53 serves as a transcriptional activator, while genes repressed by p53 were actually cell cycle-dependent genes (6). Specifically, while p53 up-regulates hundreds of genes directly through engaging with chromatin in physical contact with the gene locus, it also down-regulates the large group of cell cycle genes indirectly through its target gene *CDKN1A*. *CDKN1A* encodes for the cyclin-dependent kinase (CDK) inhibitor p21 that suppresses cyclin:CDK activity leading to the activation of the cell cycle *trans*-repressor complexes DREAM (DP, RB-like, E2F4 and MuvB) and RB:E2F (11–17). More-

---

*To whom correspondence should be addressed. Tel: +49 3641656876; Fax: +49 3641656255; Email: martin.fischer@leibniz-fli.de

© The Author(s) 2022. Published by Oxford University Press on behalf of NAR Cancer.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

over, our meta-analyses revealed that the transcription factor complexes RB:E2F, DREAM, and MMB:FOXM1 controlled essentially all of the cell cycle genes. The analysis revealed a small number of genes that were specifically activated by p53 and controlled within the cell cycle by RB:E2F and DREAM (6). However, transcriptome analyses suggest that larger subnetworks of the p53 and cell cycle-dependent gene regulation networks (GRN) are yet to be understood (5,6).

The target gene regulation (TargetGeneReg) database developed from the meta-analyses was made available through a web-based atlas at www.TargetGeneReg.org (6) to enable researchers to easily scrutinize the influence of the cell cycle and p53 on any gene of interest. Through www.TargetGeneReg.org, researchers can rapidly determine common as well as treatment, cell type, and species-specific regulations, identify promising targets, and validate findings. In numerous research projects, it is necessary to establish the degree to which a given gene is directly or indirectly affected by p53 and major cell cycle signaling pathways. Its easy-to-use interface enables researchers to quickly gather evidence about the extent and frequency their genes of interest are affected by these critical regulators. TargetGeneReg has been used for understanding cell cycle regulators, their signaling cues, and their disease relevance (18–26). Moreover, our database has helped to identify pathways that respond to drugs and stress conditions (27,28), among many other applications.

Alternative resources such as the p53 BAER hub and the Cyclebase v3.0 either focus on p53-dependent regulation or the influence of the cell cycle on gene expression, respectively (29,30). However, the integration of both layers of information necessary to understand p53's contributions to target gene regulation is not readily possible using these tools. Likewise, it is of interest for many researchers to study the target gene expression in other species such as *Mus musculus*. While Cyclebase v3.0 includes cell cycle-dependent gene regulation data from other species, the p53 BAER hub does not provide information beyond *Homo sapiens*.

The TargetGeneReg resource enabled us to compare the p53 GRN between mouse and human. Surprisingly, up-regulation by p53 displayed substantial evolutionary divergence, while down-regulation of cell cycle genes by p53–p21 is well conserved (9,31). Moreover, we employed the resource to compare the GRN of p53 to its sibling ΔNp63 and, in contrast to previous reports, we demonstrated that ΔNp63 minimally affects any direct p53 target. Instead, a large number of ΔNp63 targets were cell cycle genes, but the mechanistic link between ΔNp63 and the cell cycle remained unclear (10,32). Most recently, TargetGeneReg enabled the discovery of the transcription factor RFX7, an emerging tumor suppressor, as a novel node in the p53 GRN, proposing a mechanism for how p53 regulates several targets (33). RFX7 is linked to multiple lymphoid cancers (34), such as Burkitt lymphoma where we and others identified RFX7 as a potential cancer driver (35,36).

Here, we provide a major update for TargetGeneReg through a substantial expansion of the underlying data resources to include recent RNA-seq and ChIP-seq datasets on p53 and cell cycle regulation, inclusion of data resources on ΔNp63 and RFX7, an upgrade of the website, and visualizations of expanded ChIP-seq data through the UCSC Genome Browser.

MATERIALS AND METHODS

RNA-seq analysis pipeline

We used Trimmomatic (37) v0.39 (5nt sliding window approach, 5′ leading and mean quality cutoff 20) for read quality trimming according to inspections made from FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) v0.11.9 reports. Illumina adapters as well as mono- and di-nucleotide content were clipped using Cutadapt v2.10 (38). Potential sequencing errors were detected and corrected using Rcorrector v1.0.4 (39). Ribosomal RNA (rRNA) transcripts were artificially depleted by read alignment against rRNA databases through SortMeRNA v2.1 (40). The preprocessed data was aligned to the reference genome hg38, retrieved along with its gene annotation from Ensembl v102 (41), using the mapping software segemehl (42,43) v0.3.4 with adjusted accuracy (95%) and split-read option enabled. Mappings were filtered by Samtools v1.12 (44) for uniqueness and, in case of paired-end data, properly aligned mate pairs. Differential gene expression and its statistical significance was identified using DESeq2 v1.30.0 (45). Common thresholds of $|\log_2\text{fold-change}| \geq 0.25$ and FDR <0.05 were used to identify significantly differentially expressed genes.

Microarray analysis pipeline

All microarray datasets were available at a pre-processed stage at the Gene Expression Omnibus (GEO) and we re-analyzed the datasets with GEO2R to obtain fold expression changes and Benjamini Hochberg-corrected *P*-values (FDR) (46). Gene identifiers were mapped to Ensembl Gene IDs using the Ensembl annotation data v102 (41). Common thresholds of $|\text{log2fold-change}| \geq 0.25$ and FDR <0.05 were used to identify significantly differentially expressed genes.

Meta-analysis / generation of Expression Scores

Following our meta-analysis approach (6), *Expression Scores* for genes regulated by human and mouse p53 and ΔNp63 were calculated as the number of datasets that find the gene to be significantly up-regulated minus the number of datasets that find the gene to be significantly down-regulated by the respective transcription factor. Both, the *mouse p53* and the *ΔNp63 Expression Score* were published previously (9,10). The *p53 Expression Score 2.0* (human) contains two additional quality control measures. First, only datasets derived from at least two biological replicates have been considered for this updated score. Second, all datasets were removed that failed to identify at least 50 out of 116 direct p53 target genes that were most recurrently identified in a previous meta-analysis (3). The *Cell Cycle Expression Score* reflects the number of datasets that identified a gene as cell cycle-regulated gene. The *Cell Cycle Gene Category* is calculated by a majority vote of the nine datasets on cell cycle-dependent gene expression and is displayed for each gene that shows a *Cell Cycle Expression Score* $\geq$ 3. Precisely, each dataset that identified peak expression of

Table 1. Comparison of TargetGeneReg v2.0 features and resource properties to TargetGeneReg v1.0/1.1 (6,9), p53 BAER hub (26), and Cyclebase v3.0 (27). Numbers concern data from human except for when 'mouse' is indicated. While Cyclebase v3.0 contains multiple datasets on cell cycle-dependent gene regulation from various species, it contains only one dataset from human. ChIP-seq replicates were combined to single datasets for TargetGeneReg but kept separate for the p53 BAER hub. *NA* – not available, information was not provided in the respective publications

	TargetGeneReg v2.0	TargetGeneReg v1.0/1.1	p53 BAER hub	Cyclebase v3.0
# human genes	**37243**	18845	*NA*	*NA*
human genome version	**hg38**	hg19	hg19	*NA*
# p53 expression datasets	**57**	20	16	-
# p53 ChIP-seq tracks/datasets	32 / 28	15	**41**	-
p53RE prediction	**yes**	no	**yes**	-
# mouse p53 expression datasets	**15**	**15**	-	-
# mouse p53 ChIP-seq datasets	**9**	**9**	-	-
# cell cycle expression datasets	**9**	5	-	1
# DREAM ChIP-seq datasets	**17**	9	-	-
# RB ChIP-seq datasets	**6**	2	-	-
# MMB-FOXM1 ChIP-seq datasets	**22**	6	-	-
CHR and E2F motif predictions	**yes**	**yes**	-	-
# ΔNp63 expression datasets	**16**	-	-	-
# ΔNp63 ChIP-seq datasets	**20**	-	-	-
p63RE prediction	**yes**	-	-	-
RFX7 target gene prediction	**yes**	-	-	-
Genome browser visualizations	**yes**	no	**yes**	no

the gene in 'G1', 'G1/S', or 'S-phase' is grouped as 'G1/S', and peak expression in 'G2', 'G2/M', 'M', and 'M/G1' is grouped as 'G2/M'.

ChIP-seq data integration

Peak datasets and bigwigs from ChIP-seq experiments were retrieved from CistromeDB (47) ensuring a common data processing pipeline and thereby a direct comparability. Only RFX7 ChIP-seq data were taken from our recent study (33) as they were not yet available through CistromeDB. Bigwigs (ChIP-seq tracks) have been made available through track hubs for the UCSC Genome Browser (48). Notably, while ChIP-seq replicates are available as individual tracks in the track hubs, they have been jointly considered as one dataset for the generation of peak-of-peaks summaries. Precisely, when replicate experiments were available, all peaks were used that have been identified in at least two replicates. To identify overlapping and non-overlapping peaks, Bedtools 'intersect' was employed, and to generate the peak-of-peaks summaries, multiple peak files were combined using Bedtools 'multiinter' (49). The p53 and p63 ChIP-seq collections and summaries on human and mouse p53 and ΔNp63 have been published previously (9,10). Similarly, p53REs and p63REs were taken from our previous study (10).

RESULTS

Similar to TargetGeneReg v1.0 (6), TargetGeneReg v2.0 focuses on gene regulation by the tumor suppressor p53 in conjunction with the human cell cycle. An earlier upgrade (TargetGeneReg v1.1) introduced p53-dependent gene regulation and p53 binding data from mouse to TargetGeneReg (9), which, to our knowledge, is unique to the TargetGeneReg resources. Similarly, transcription factor binding data for central transcriptional cell cycle regulators, including the DREAM complex, RB, the MMB complex, and FOXM1, is unique to TargetGeneReg. The upgrade to version 2.0 not only expands the data, but also includes data on p53's oncogenic sibling ΔNp63 and the emerging tumor suppressor RFX7 (Table 1). In the following sections we provide more detailed information on the data and the applicability of the upgraded TargetGeneReg resource.

Gene regulation by p53

In the first version of TargetGeneReg, we integrated 20 datasets on p53-dependent gene regulation (6). Since the publication of TargetGeneReg v1.0, several additional high-throughput datasets with varying resolution and experimental strategies became available. To optimally use this additional data and strengthen the power of our resource, we have adjusted our quality control regiment. Specifically, we systematically searched the GEO database for RNA-seq and microarray datasets that employed experimental strategies known to affect p53 signaling. This search included experiments involving MDM2 inhibitors (Nutlin and RG7388), genotoxic and nucleolar stress inducers (Doxorubicin, 5-FU, Actinomycin D, Daunorubicin, Etoposide, Bleomycin, Camptothecin, and UV), viral oncoproteins (SV40 LT, HPV16 E6, and HPV18 E6), exogenous TP53 expression, TNFα, and senescence (oncogene and replication-induced). For inclusion in the updated resource, we required all datasets to comprise at least two biological replicates for both treatment and control conditions. Notably, all datasets we obtained were derived from cell line models. We integrated information from 64 RNA-seq and 35 microarray datasets derived to identify significantly differentially expressed genes. To verify the effects of the selected experiments on known p53-regulated genes, we used a benchmark dataset of 116 direct and highly responsive p53 target genes identified earlier based on 16 genome-wide analyses (3). All experiments that yielded <50 significantly differentially expressed benchmark targets were removed from further analysis (see Materials and Methods). This measure ensures focus on activation of the p53 pathway by the experimental setup and sufficient power to identify specific p53 activities (Figure 1A). A total of 44 RNA-

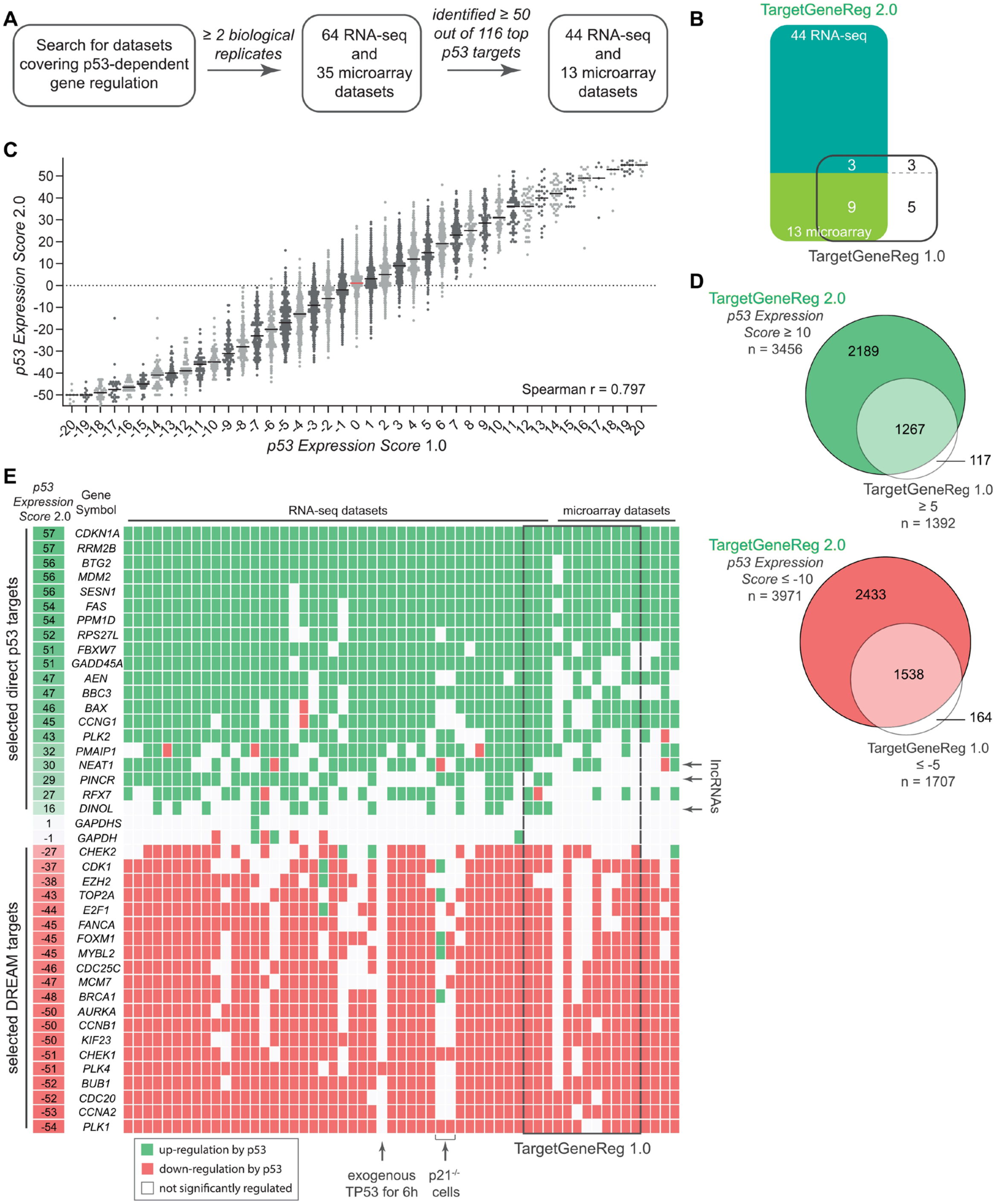


Figure 1. (**A**) Flow chart for the integration of datasets on p53-dependent gene regulation. (**B**) Datasets on p53-dependent gene regulation in TargetGeneReg v2.0 compared to TargetGeneReg v1.0. (**C**) The *p53 Expression Score* v2.0 from TargetGeneReg v2.0 compared to the *p53 Expression Score* v2.0 from TargetGeneReg v1.0 for 17 446 genes present in both databases. Genes are displayed by individual points. The median is indicated by a black line or a red line to highlight '0'. (**D**) Genes passing the recommended p53 Expression Score threshold to be considered high-recurrence genes that are up or down-regulated by p53 in TargetGeneReg v2.0 compared to TargetGeneReg v1.0. (**E**) The p53 Expression Score v2.0 and data from the underlying 57 individual datasets visualized for 20 selected direct p53 target genes, 20 selected targets of the DREAM complex, and the non-regulated *GAPDH* genes. It is indicated whether individual datasets were generated using RNA-seq or microarray. Individual datasets that were also present in TargetGeneReg v1.0 are indicated. The lncRNAs *NEAT1*, *PINCR*, and *DINOL* are highlighted as they were not available in TargetGeneReg v1.0. Details on three datasets in which most DREAM targets are not down-regulated by p53 are shown.

seq and 13 microarray datasets passed this control (Figure 1B).

The *p53 Expression Score*, based on these 57 datasets, was calculated for each gene by the number of datasets yielding significant up-regulation minus the number of datasets with significant down-regulation of the gene by p53. To calculate the score, we required a gene to be sufficiently expressed in at least three datasets. A gene was deemed to be expressed when DESeq2 was able to include it in the differential expression analysis, i.e. assign $\log_2$fold-change and FDR values. A direct comparison of the updated score (*p53 Expression Score* v2.0) with the initial one (*p53 Expression Score* v1.0) exhibits a strong correlation but also suggests an improved resolution (Figure 1C). Most importantly, while the previous version was limited to 18 845 protein-coding genes from hg19, the updated resource now provides a *p53 Expression Score* for 37 243 genes from hg38. The *p53 Expression Score* v1.0 had a minimum threshold of $\geq$5 and $\leq$–5 to consider genes with high confidence as being up and down-regulated by p53, respectively (6). In the case of the updated *p53 Expression Score* v2.0, respective thresholds of $\geq$10 and $\leq$–10 were passed by 3456 and 3971 genes (Figure 1D). To illustrate the advantage of the updated *p53 Expression Score*, we visualized it together with the underlying individual datasets for 20 selected direct p53 target genes and 20 selected targets of the DREAM complex (Figure 1E).

Of note, some datasets show few if any down-regulated DREAM targets. This is the case when p21 (CDKN1A) negative cells were used, since they were unable to reactivate the DREAM complex efficiently. Likewise, an experiment where exogenous TP53 was induced for only 6 h, an interval too brief to reactivate the cell cycle *trans*-repressor complexes, did not lead to a down-regulation of critical cell cycle genes. Moreover, our comparison identifies p53-dependent lncRNAs such as *DINOL* (50), *PINCR* (51) and *NEAT1* (52,53) excluded from the previous version because of insufficient data (Figure 1E). In addition to information on the p53-dependent regulation of thousands of non-coding RNAs, the updated *p53 Expression Score* v2.0 provides much more detailed information on p53-dependent regulation for hundreds of genes for which the *p53 Expression Score* v1.0 was inconclusive. Consequently, 1674 and 1917 genes that previously displayed a *p53 Expression Score* v1.0 between 5 and –5 now passed the threshold of a *p53 Expression Score* v2.0 of $\geq$10 and $\leq$–10, respectively, indicating a differential regulation by p53 with high recurrence.

The direct binding of p53 to the gene promoter is a crucial property of many genes up-regulated by this transcription factor (3–5). While TargetGeneReg v1.0 integrated 15 datasets on p53 genome binding (6), the updated version now integrates an expanded collection of 28 ChIP-seq datasets. While the previous version displayed only the number of datasets that identified p53 binding near a gene's TSS, the updated resource contains precise binding location information and visualizations thereof. To enable users to rapidly visualize the large number of 28 individual p53 ChIP-seq datasets, we provide a 'peak-of-peaks' data track representing a pile-up of p53 peak regions from individual datasets (Figure 2A). Therefore, the 'peak-of-peaks' track provides quick summary information on how many datasets identified p53 binding to any locus in the genome. In addition, the p53 response element (p53RE) most closely resembling a canonical p53RE is displayed for each peak-of-peaks, as described previously (10). In addition, the ChIP-seq tracks from all individual datasets can be displayed upon selection, providing a seamless visualization of cell type and treatment-specific information next to the summary data (Figure 2A and B). As established previously, any binding site with support from at least five datasets is considered to be of high recurrence (9,10). The website's 'Overview' section indicates for every gene whether it displays a high-recurrence binding site within 2.5 kb of a TSS, and whether a high-recurrence binding site is linked to the gene locus through a double-elite enhancer:gene association listed in the GeneHancer database (54).

Together, TargetGeneReg v2.0 provides information on p53-dependent gene regulation for twice as many genes from almost three times as many datasets in total and almost five times as many datasets that follow the tightened control measures. In addition, it provides almost twice as many p53 ChIP-seq datasets, predictions for the underlying p53RE, and precise location visualizations.

Cell cycle-dependent gene regulation

Cell cycle genes play essential roles in cell cycle progression and therefore are typical markers of proliferation that are dysregulated in many cancers (55). The tumor suppressor p53 down-regulates cell cycle genes to sustain cell cycle arrest. Based on TargetGeneReg v1.0, we consolidated the five cell cycle gene peak clusters defined by Whitfield *et al.* (56) to two major groups of cell cycle genes, namely G1/S and G2/M genes (6). Here, we expanded the previous resource's five datasets to include four additional datasets (Figure 3A). For all genes identified as cell cycle-dependently regulated in at least three of the nine datasets, we predicted whether the gene is a G1/S or a G2/M gene based on a majority vote by the nine datasets (see Materials and Methods). The website's 'Overview' section provides information on the number of datasets that suggest a gene to be driven by the cell cycle ('Cell Cycle Expression Score') and its classification prediction ('Cell Cycle Gene Category').

The two distinct groups of G1/S and G2/M genes are primarily characterized by E2F and CHR (cell cycle genes homology region) DNA recognition motifs in their promoters, respectively (2,6,57). The DREAM complex can bind to both E2F and CHR motifs through its respective subunits E2F4 and LIN54, while RB:E2F specifically binds G1/S cell cycle genes through E2F motifs. In contrast, the transcription factors B-MYB (also known as MYBL2) and FOXM1 associate with DREAM's LIN54-containing MuvB core complex later in the cell cycle to specifically activate the expression of G2/M genes through binding their CHR sequences (2). To allow a more comprehensive analysis of cell cycle-dependent regulation, we expanded the nine datasets on genome binding by DREAM complex components to 17. Similarly, we extended the two previous datasets on RB binding to six datasets, and the previous six datasets on MMB:FOXM1 (B-MYB:MuvB:FOXM1) binding to 22 datasets. Potential E2F and CHR motifs under respective RB and MMB:FOXM1 binding sites have been predicted using HOMER (58). The individual ChIP-

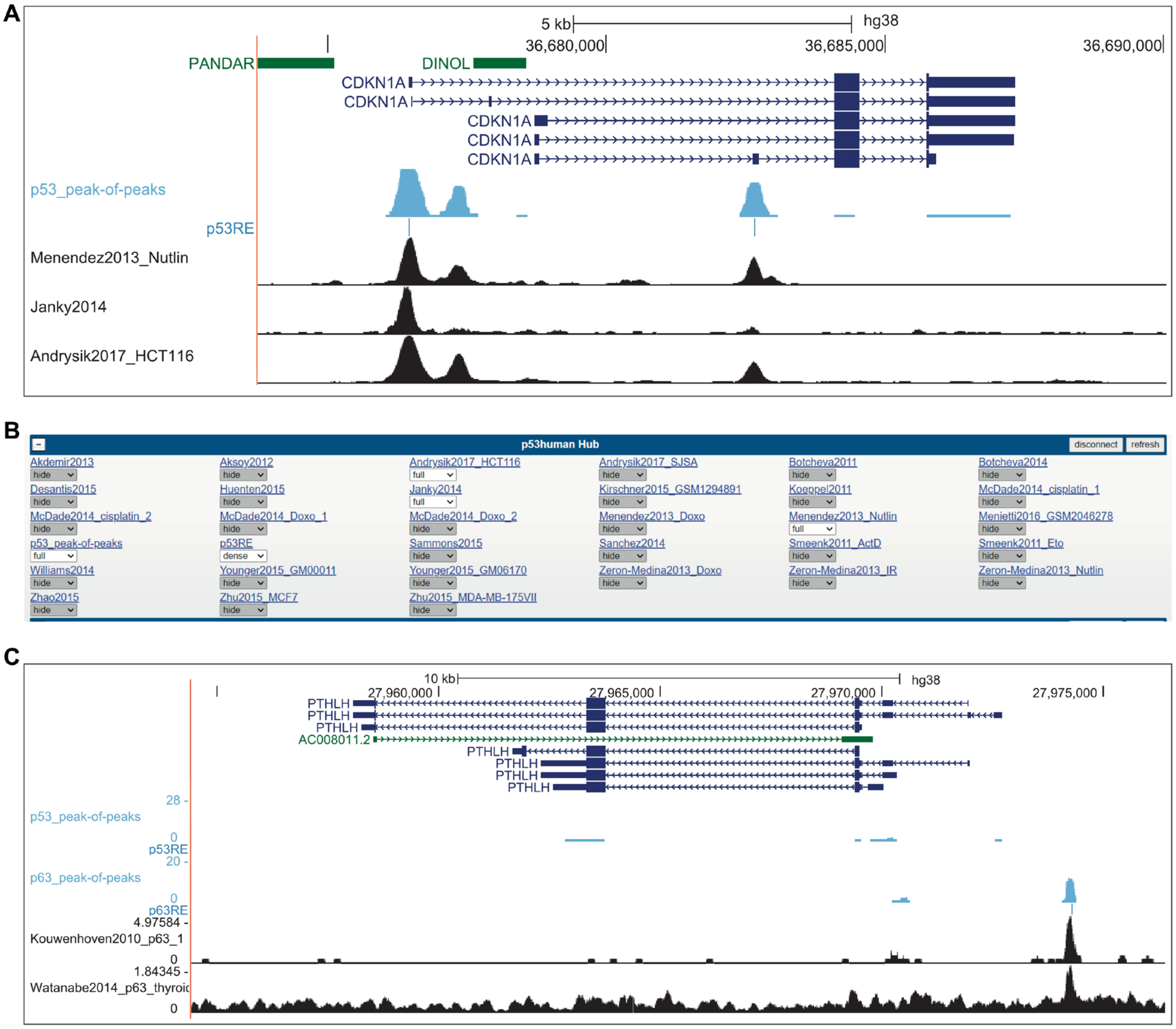


Figure 2. (**A**) Image of UCSC Genome Browser displaying the *CDKN1A* locus linked from TargetGeneReg v2.0. The blue tracks display the p53_peak-of-peaks summary and the most likely underlying p53RE that has been identified. Individual p53 ChIP-seq tracks can be selected from (**B**) the track hub that is loaded through the TargetGeneReg v2.0 linkage. Together, the visualization provides precise visualization of p53 binding sites locations and their underlying p53RE and enables seamless comparisons between summary data and individual datasets. (**C**) Image of UCSC Genome Browser displaying the *PTHLH* locus linked from TargetGeneReg v2.0. The blue tracks display the p53_peak-of-peaks and p63_peak-of-peaks summaries and the most likely underlying p53RE and p63RE. Individual p63 ChIP-seq tracks can be selected from the track hub hat is loaded through the TargetGeneReg v2.0 linkage, as shown for p53 above. *PTHLH* is a direct target of ΔNp63 but not p53, and the unique p63 binding site can be readily seen by comparing the peak-of-peaks summary binding data.

seq tracks, peak-of-peaks, and motif predictions are available through UCSC Genome Browser visualizations.

Gene regulation by mouse p53 and its difference to human p53

Previously, we employed our meta-analysis approach on mouse p53 synthesizing p53-dependent gene regulation data across 15 datasets, and we made the data available through TargetGeneReg v1.1 (9). Here, we integrated our database on p53-dependent gene regulation in mice (mm10) with the updated database on p53-dependent gene regulation in humans (hg38) described above. Therefore, TargetGeneReg v2.0 includes mouse p53-dependent gene regulation data for the one-to-one orthologs of 14 712 human genes. While there is a good correlation between the *mouse p53 Expression Score* and the *human p53 Expression Score* v2.0 for genes down-regulated by p53, the correlation for up-regulated genes is poor (Figure 3B), indicating a strong and a comparably low evolutionary conservation of p53 down and up-regulated genes, respectively. Similar results have already been reported for the first *p53 Expression Score* (9,31). Precise binding data (ChIP-seq) of mouse p53 is available through links to the UCSC Genome Browser

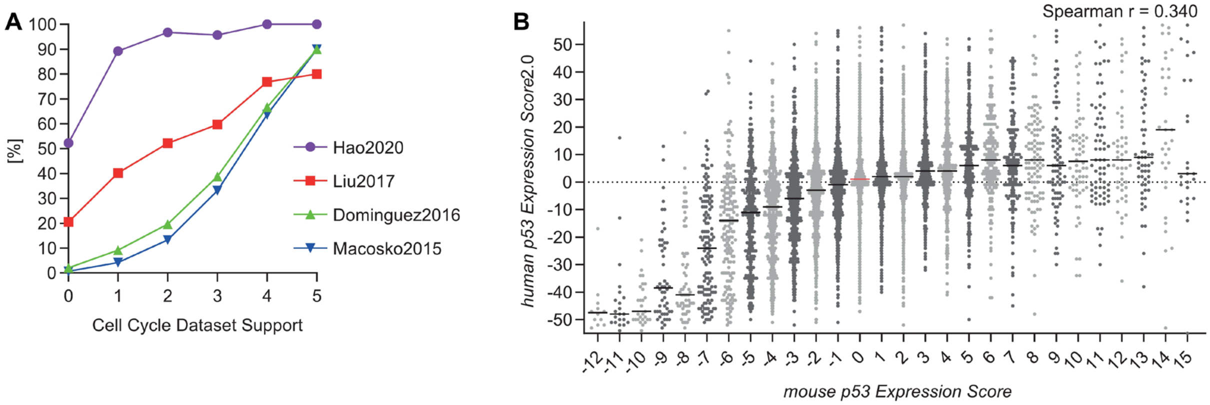


Figure 3. (**A**) Four new datasets on cell cycle-dependent gene expression have been added (63–66). Following our previous quality control (6), we tested whether the datasets were more likely to identify a gene as cell cycle gene when more datasets from TargetGeneReg v1.0 (X-axis) agreed on its cell cycle gene status. (**B**) The *mouse p53 Expression Score* (9) compared to the *human p53 Expression Score* v2.0 for 14 712 one-to-one orthologs with both scores. Genes are displayed by individual points. The median is indicated by a black line or a red line to highlight ‘0’.

with embedded track hubs similar to human protein binding data described above.

Gene regulation by ΔNp63

We previously employed our meta-analysis approach to provide a comprehensive resource for gene regulation by p53’s sibling ΔNp63, an essential oncoprotein in squamous cell carcinomas (10). Given its relevance to cancer and close connection to p53, we integrated our ΔNp63 database comprising 16 datasets on p63-dependent gene regulation and 20 ChIP-seq datasets on p63 DNA binding into TargetGeneReg v2.0. The *ΔNp63 Expression Score* and predictions of p63 targets (180 high-recurrence targets available in Table 1 from Riege *et al.* (10)) and potential p63 targets (comprising all genes bound and regulated by p63) are available through the website’s ‘Overview’ section. DNA binding data and identified p63 response elements (p63RE) are available through UCSC Genome Browser track hubs and enable a seamless comparison between individual p63 ChIP-seq tracks, summary data thereof (p63_peak-of-peaks), and respective data from p53. *PTHLH*, for instance, is a direct target gene of ΔNp63 but not of p53 (Figure 2C).

Expanding the p53 gene regulatory network through RFX7

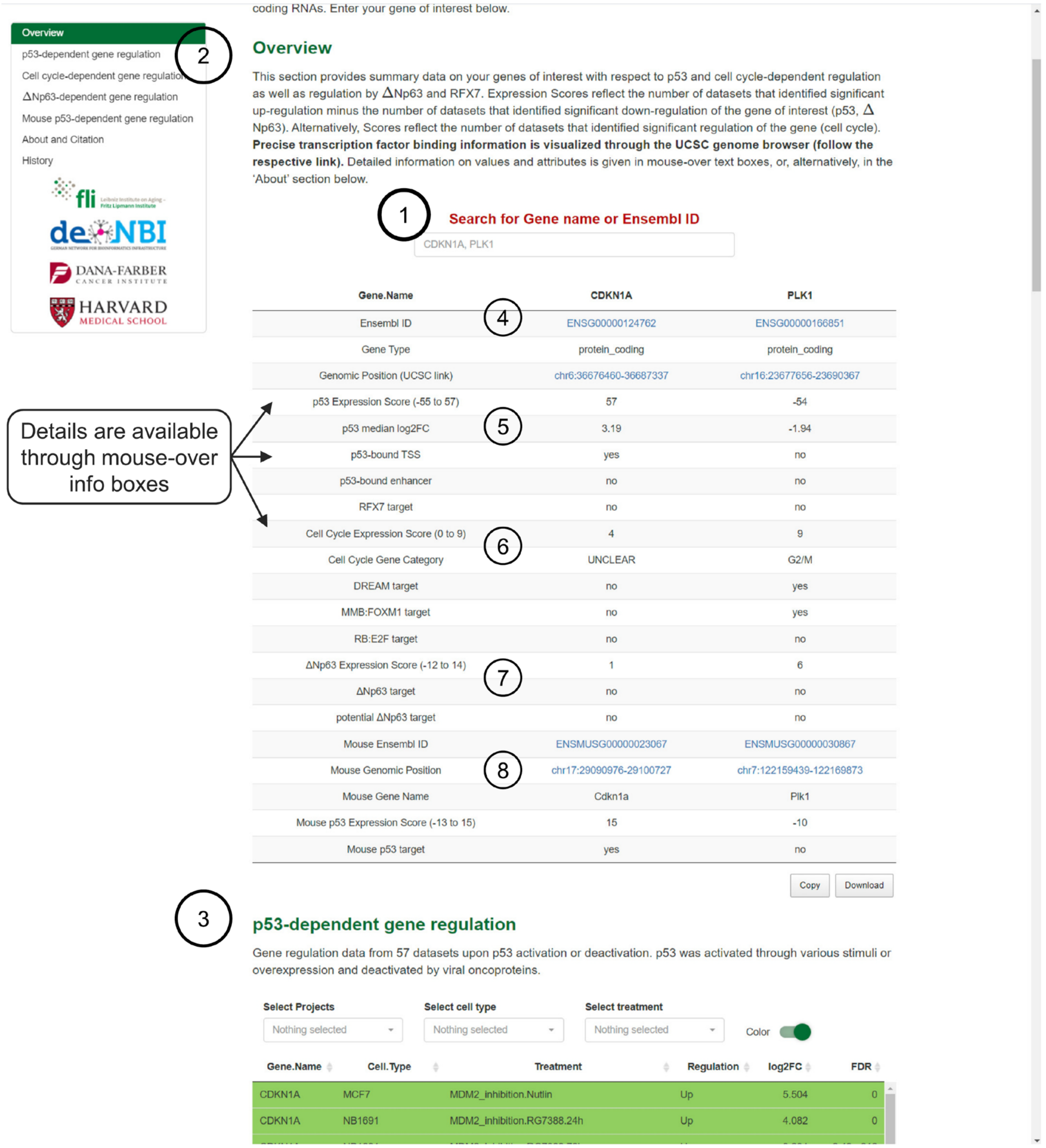
Complex cross-talk between signaling pathways impedes the identification of indirect gene regulatory mechanisms employed by p53. For example, following two decades of conflicting data on mechanisms of p53-dependent gene repression, p53 was found to serve as a transcriptional activator that represses genes indirectly, with its target p21 taking a predominant role through its profound influence on down-regulating the cell cycle genes (3,4,6,7,59). Recently, we identified the transcription factor and emerging tumor suppressor RFX7 as a vital node in the p53 transcriptional program. RFX7 orchestrates a subnetwork of tumor suppressor genes in response to cellular stress and p53 (33) and cooperates with p53 to inhibit the pro-survival kinases AKT and mTORC1 (60). Given the crucial role of the novel p53-RFX7 signaling axis in the p53 gene regulatory network and its potential importance to cancer biology, we included the data on RFX7 target genes in TargetGeneReg 2.0. The website’s ‘Overview’ section displays whether a gene has been predicted as an RFX7 target, offering a mechanistic explanation for its p53-dependent up-regulation. In addition, RFX7 ChIP-seq tracks are available through the UCSC Genome Browser visualizations.

Navigating the TargetGeneReg 2.0 website

Our updated resource is tailored to rapidly provide information on genes of interest entered in the main input field. The arguably strongest asset of the TargetGeneReg resource is summary data on gene regulation by various transcription factors generated through a synthesis of multiple individual datasets integrated by our meta-analysis approach. Therefore, the ‘Overview’ section situated at the top of the one-page website provides all summary information on the genes of interest that have been entered (Figure 4). Importantly, it enables a quick direct comparison of the summary data between multiple genes. More detailed information on gene regulation from the individual datasets, a pie chart illustration of the summary data, as well as volcano and box-plots of the dataset results are provided in the detailed sections below, which are equipped with sorting options to quickly identify the most relevant data. Precise transcription factor binding data visualized through the UCSC Genome Browser (48) are available through the genomic position links provided in the ‘Overview’ section for both human (hg38) and mouse (mm10).

DISCUSSION

The TargetGeneReg has gained a strong reputation in the p53 and cell cycle communities. It provides a deeper insight into p53 and cell cycle-dependent gene regulation mechanisms. The presented upgrade, TargetGeneReg v2.0, substantially improves this resource. Like its predecessor, the


Overview
p53-dependent gene regulation
Cell cycle-dependent gene regulatio
ΔNp63-dependent gene regulation
Mouse p53-dependent gene regulation
About and Citation
History
fli Leibniz Institute on Aging - Fritz Lipmann Institute
de.NBI
DANA-FARBER CANCER INSTITUTE
HARVARD MEDICAL SCHOOL
1
2
3
4
5
6
7
8
Details are available through mouse-over info boxes
Copy
Download
Color

coding RNAs. Enter your gene of interest below.

Overview

This section provides summary data on your genes of interest with respect to p53 and cell cycle-dependent regulation as well as regulation by ΔNp63 and RFX7. Expression Scores reflect the number of datasets that identified significant up-regulation minus the number of datasets that identified significant down-regulation of the gene of interest (p53, ΔNp63). Alternatively, Scores reflect the number of datasets that identified significant regulation of the gene (cell cycle). **Precise transcription factor binding information is visualized through the UCSC genome browser (follow the respective link).** Detailed information on values and attributes is given in mouse-over text boxes, or, alternatively, in the 'About' section below.

Search for Gene name or Ensembl ID

CDKN1A, PLK1

Gene.Name	CDKN1A	PLK1
Ensembl ID	ENSG00000124762	ENSG00000166851
Gene Type	protein_coding	protein_coding
Genomic Position (UCSC link)	chr6:36676460-36687337	chr16:23677656-23690367
p53 Expression Score (-55 to 57)	57	-54
p53 median log2FC	3.19	-1.94
p53-bound TSS	yes	no
p53-bound enhancer	no	no
RFX7 target	no	no
Cell Cycle Expression Score (0 to 9)	4	9
Cell Cycle Gene Category	UNCLEAR	G2/M
DREAM target	no	yes
MMB:FOXM1 target	no	yes
RB:E2F target	no	no
ΔNp63 Expression Score (-12 to 14)	1	6
ΔNp63 target	no	no
potential ΔNp63 target	no	no
Mouse Ensembl ID	ENSMUSG00000023067	ENSMUSG00000030867
Mouse Genomic Position	chr17:29090976-29100727	chr7:122159439-122169873
Mouse Gene Name	Cdkn1a	Plk1
Mouse p53 Expression Score (-13 to 15)	15	-10
Mouse p53 target	yes	no

p53-dependent gene regulation

Gene regulation data from 57 datasets upon p53 activation or deactivation. p53 was activated through various stimuli or overexpression and deactivated by viral oncoproteins.

Select Projects: Nothing selected
Select cell type: Nothing selected
Select treatment: Nothing selected

Gene.Name	Cell.Type	Treatment	Regulation	log2FC	FDR
CDKN1A	MCF7	MDM2_inhibition.Nutlin	Up	5.504	0
CDKN1A	NB1691	MDM2_inhibition.RG7388.24h	Up	4.082	0

① Start at this input field: Enter your genes of interest as gene symbols or ensembl IDs separated by commas.
② The one-page website contains several sections. Navigate through the navigation bar or by scrolling.
③ The additional sections provide more detailed information, such as data from individual datasets.
④ The overview section starts with general information on your genes of interest. Mind the genome position link!
⑤ Information on p53-dependent gene regulation are clustered together with the p53 network extender RFX7.
⑥ Information on cell cycle-dependent gene regulation are clustered together.
⑦ Information on ΔNp63-dependent gene regulation are clustered together.
⑧ Detailed information on mouse orthologs and gene regulation by mouse p53. Mind the genome position link!

Figure 4. The TargetGeneReg 2.0 website design. Start by entering your genes of interest as gene symbol or Ensembl gene ID. The overview section provides a helpful summary on the regulation of your genes of interest. Details on each point in the overview section is available through mouse-over boxes and through the 'About' section. The additional sections are available through the navigation bar in the upper left corner or by scrolling through this one-page website design. The additional sections contain more detailed information, such as data from the individual datasets and citation and history data for the resource.

resource is tailored to quickly retrieve information on the users' genes of interest and provides swift comparisons between genes and experimental conditions. TargetGeneReg v1.0 was a starting point to help the p53 and cell cycle communities to gain deep biological insights by providing reference points and a platform that enables users to quickly test whether their genes of interest are likely regulated by p53 or the cell cycle. The new version integrates more datasets, substantially improving its power. Specifically, TargetGeneReg now includes data on non-coding RNAs and provides information on the gene regulation by p53's oncogenic sibling ΔNp63 and the emerging tumor suppressor RFX7.

Visualization of the transcription factor binding data through the UCSC Genome Browser provides precise location information for the user to better interpret the potential consequences of the binding for their gene of interest. For example, p53 binding to intronic locations can induce alternative transcription start sites leading to transcript variants with shortened 5′ sequences, as reported for *MDM2* and *FBXW7* (61,62). Visualization of the strongest scoring underlying p53RE and p63RE provides an unprecedented depth of binding information.

While the summary data, such as the *Expression Scores* are particularly helpful to quickly assess the regulation of genes, it is critical to tally the characteristics of individual datasets and genes used for the generation of this summary. Importantly, a low *Expression Score* does not rule out p53 or cell cycle-dependent regulation. In addition to biological variability, such as cell line-specific differences, genes may evade the differential expression detection due to low transcript abundances, low but biologically relevant fold-changes, or methodological limitations (e.g. limited number of replicates and sequencing depth). For example, many non-coding RNAs have low expression levels and thus may have a low *p53 Expression Score* although they are actually strongly regulated by p53 (e.g. *DINOL* displayed in Figure 1E). From a statistical perspective, this situation would require the integration of more datasets to increase the statistical power. To address this limitation in spite of additional datasets, we display the 'p53 median log2FC'. The combination of a low *p53 Expression Score* and a high 'p53 median log2FC' might indicate that a gene evades differential expression detection due to a low overall expression level.

Together, with TargetGeneReg 2.0 we provide a comprehensive resource on p53-dependent regulation in humans and mice. Additional information on ΔNp63 and cell cycle-dependent gene regulation facilitates the discovery of further novel biological insights.

DATA AVAILABILITY

All data are available through the web-atlas at www.TargetGeneReg.org. Accession numbers of individual datasets are available through the website.

ACKNOWLEDGEMENTS

We gratefully acknowledge Bernd Senf from the FLI Core Facility Life Science Computing for help with the website's server backbone. We thank Arne Sahm for critical reading of the manuscript.

Author contributions: M.F. conceived the study. M.F. and S.H. supervised the work. M.F. and K.R. curated and analyzed the data. R.S., K.R., M.F. and S.H. designed the website. R.S. created the website. M.F., J.A.D. and S.H. interpreted the data. M.F., with the help of S.H. and J.A.D., wrote the manuscript. All authors read and approved the manuscript.

FUNDING

German Federal Ministry for Education and Research (BMBF) [031L016D to S.H.]; National Institutes of Health [R35CA232128 and P01CA203655 to J.A.D., in part]; the FLI is a member of the Leibniz Association and is financially supported by the Federal Government of Germany and the State of Thuringia.

Conflict of interests statement. J.A.D. received research funding from Rain Therapeutics. The remaining authors declare no competing interests.

REFERENCES

1. Sadasivam,S. and DeCaprio,J.A. (2013) The DREAM complex: master coordinator of cell cycle-dependent gene expression. *Nat. Rev. Cancer*, **13**, 585–595.
2. Fischer,M. and Müller,G.A. (2017) Cell cycle transcription control: DREAM/MuvB and RB-E2F complexes. *Crit. Rev. Biochem. Mol. Biol.*, **52**, 638–662.
3. Fischer,M. (2017) Census and evaluation of p53 target genes. *Oncogene*, **36**, 3943–3956.
4. Sullivan,K.D., Galbraith,M.D., Andrysik,Z. and Espinosa,J.M. (2018) Mechanisms of transcriptional regulation by p53. *Cell Death Differ.*, **25**, 133–143.
5. Sammons,M.A., Nguyen,T.-A.T., McDade,S.S. and Fischer,M. (2020) Tumor suppressor p53: from engaging DNA to target gene regulation. *Nucleic Acids Res.*, **48**, 8848–8869.
6. Fischer,M., Grossmann,P., Padi,M. and DeCaprio,J.A. (2016) Integration of TP53, DREAM, MMB-FOXM1 and RB-E2F target gene analyses identifies cell cycle gene regulatory networks. *Nucleic Acids Res.*, **44**, 6070–6086.
7. Fischer,M., Steiner,L. and Engeland,K. (2014) The transcription factor p53: not a repressor, solely an activator. *Cell Cycle*, **13**, 3037–3058.
8. Fischer,M., Uxa,S., Stanko,C., Magin,T.M. and Engeland,K. (2017) Human papilloma virus E7 oncoprotein abrogates the p53-p21-DREAM pathway. *Sci. Rep.*, **7**, 2603.
9. Fischer,M. (2019) Conservation and divergence of the p53 gene regulatory network between mice and humans. *Oncogene*, **38**, 4095–4109.
10. Riege,K., Kretzmer,H., Sahm,A., McDade,S.S., Hoffmann,S. and Fischer,M. (2020) Dissecting the DNA binding landscape and gene regulatory network of p63 and p53. *Elife*, **9**, e63266.
11. Quaas,M., Müller,G.A. and Engeland,K. (2012) p53 can repress transcription of cell cycle genes through a p21 WAF1/CIP1-dependent switch from MMB to DREAM protein complex binding at CHR promoter elements. *Cell Cycle*, **11**, 4661–4672.
12. Fischer,M., Grundke,I., Sohr,S., Quaas,M., Hoffmann,S., Knörck,A., Gumhold,C. and Rother,K. (2013) p53 and cell cycle dependent transcription of kinesin family member 23 (KIF23) is controlled Via a CHR promoter element bound by DREAM and MMB complexes. *PLoS One*, **8**, e63187.
13. Fischer,M., Quaas,M., Wintsche,A., Müller,G.A. and Engeland,K. (2014) Polo-like kinase 4 transcription is activated via CRE and NRF1 elements, repressed by DREAM through CDE/CHR sites and deregulated by HPV E7 protein. *Nucleic Acids Res.*, **42**, 163–180.
14. Fischer,M., Quaas,M., Nickel,A. and Engeland,K. (2015) Indirect p53-dependent transcriptional repression of Survivin, CDC25C, and PLK1 genes requires the cyclin-dependent kinase inhibitor

p21/CDKN1A and CDE/CHR promoter sites binding the DREAM complex. *Oncotarget*, **6**, 41402–41417.

15. Fischer,M., Quaas,M., Steiner,L. and Engeland,K. (2016) The p53-p21-DREAM-CDE/CHR pathway regulates G2/M cell cycle genes. *Nucleic Acids Res.*, **44**, 164–174.
16. Uxa,S., Bernhart,S.H., Mages,C.F.S., Fischer,M., Kohler,R., Hoffmann,S., Stadler,P.F., Engeland,K. and Müller,G.A. (2019) DREAM and RB cooperate to induce gene repression and cell-cycle arrest in response to p53 activation. *Nucleic Acids Res.*, **47**, 9087–9103.
17. Schade,A.E., Fischer,M. and DeCaprio,J.A. (2019) RB, p130 and p107 differentially repress G1/S and G2/M genes after p53 activation. *Nucleic Acids Res.*, **47**, 11197–11208.
18. Macedo,J.C., Vaz,S., Bakker,B., Ribeiro,R., Bakker,P.L., Escandell,J.M., Ferreira,M.G., Medema,R., Foijer,F. and Logarinho,E. (2018) FoxM1 repression during human aging leads to mitotic decline and aneuploidy-driven full senescence. *Nat. Commun.*, **9**, 2834.
19. Decaesteker,B., Denecker,G., Van Neste,C., Dolman,E.M., Van Loocke,W., Gartlgruber,M., Nunes,C., De Vloed,F., Depuydt,P., Verboom,K. *et al.* (2018) TBX2 is a neuroblastoma core regulatory circuitry component enhancing MYCN/FOXM1 reactivation of DREAM targets. *Nat. Commun.*, **9**, 4866.
20. Braun,L., Brenier-Pinchart,M.-P., Hammoudi,P.-M., Cannella,D., Kieffer-Jaquinod,S., Vollaire,J., Josserand,V., Touquet,B., Couté,Y., Tardieux,I. *et al.* (2019) The Toxoplasma effector TEEGR promotes parasite persistence by modulating NF-κB signalling via EZH2. *Nat. Microbiol.*, **4**, 1208–1220.
21. Werwein,E., Cibis,H., Hess,D. and Klempnauer,K.-H. (2019) Activation of the oncogenic transcription factor B-Myb via multisite phosphorylation and prolyl cis/trans isomerization. *Nucleic Acids Res.*, **47**, 103–121.
22. Diab,A., Gem,H., Swanger,J., Kim,H.Y., Smith,K., Zou,G., Raju,S., Kao,M., Fitzgibbon,M., Loeb,K.R. *et al.* (2020) FOXM1 drives HPV+ HNSCC sensitivity to WEE1 inhibition. *Proc. Natl. Acad. Sci. U.S.A.*, **117**, 28287–28296.
23. Ng,J.C.F., Quist,J., Grigoriadis,A., Malim,M.H. and Fraternali,F. (2019) Pan-cancer transcriptomic analysis dissects immune and proliferative functions of APOBEC3 cytidine deaminases. *Nucleic Acids Res.*, **47**, 1178–1194.
24. Bainor,A.J., Saini,S., Calderon,A., Casado-Polanco,R., Giner-Ramirez,B., Moncada,C., Cantor,D.J., Ernlund,A., Litovchick,L. and David,G. (2018) The HDAC-Associated Sin3B protein represses DREAM complex targets and cooperates with APC/C to promote quiescence. *Cell Rep.*, **25**, 2797–2807.
25. Pfister,K., Pipka,J.L., Chiang,C., Liu,Y., Clark,R.A., Keller,R., Skoglund,P., Guertin,M.J., Hall,I.M. and Stukenberg,P.T. (2018) Identification of drivers of aneuploidy in breast tumors. *Cell Rep.*, **23**, 2758–2769.
26. Gonyo,P., Bergom,C., Brandt,A.C., Tsaih,S.-W., Sun,Y., Bigley,T.M., Lorimer,E.L., Terhune,S.S., Rui,H., Flister,M.J. *et al.* (2017) SmgGDS is a transient nucleolar protein that protects cells from nucleolar stress and promotes the cell cycle by regulating DREAM complex gene expression. *Oncogene*, **36**, 6873–6883.
27. Hernandez Borrero,L., Dicker,D.T., Santiago,J., Sanders,J., Tian,X., Ahsan,N., Lev,A., Zhou,L. and El-Deiry,W.S. (2021) A subset of CB002 xanthine analogs bypass p53-signaling to restore a p53 transcriptome and target an S-phase cell cycle checkpoint in tumors with mutated-p53. *Elife*, **10**, e70429.
28. Min,M. and Spencer,S.L. (2019) Spontaneously slow-cycling subpopulations of human cells originate from activation of stress-response pathways. *PLOS Biol.*, **17**, e3000178.
29. Nguyen,T.-A.T., Grimm,S.A., Bushel,P.R., Li,J., Li,Y., Bennett,B.D., Lavender,C.A., Ward,J.M., Fargo,D.C., Anderson,C.W. *et al.* (2018) Revealing a human p53 universe. *Nucleic Acids Res.*, **46**, 8153–8167.
30. Santos,A., Wernersson,R. and Jensen,L.J. (2015) Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. *Nucleic Acids Res.*, **43**, D1140–D1144.
31. Fischer,M. (2021) Mice are not humans: the case of p53. *Trends Cancer*, **7**, 12–14.
32. Woodstock,D.L., Sammons,M.A. and Fischer,M. (2021) p63 and p53: collaborative partners or dueling rivals? *Front. Cell Dev. Biol.*, **9**, 701986.
33. Coronel,L., Riege,K., Schwab,K., Förste,S., Häckes,D., Semerau,L., Bernhart,S.H., Siebert,R., Hoffmann,S. and Fischer,M. (2021) Transcription factor RFX7 governs a tumor suppressor network in response to p53 and stress. *Nucleic Acids Res.*, **49**, 7437–7456.
34. Fischer,B.A., Chelbi,S.T. and Guarda,G. (2020) Regulatory factor X 7 and its potential link to Lymphoid cancers. *Trends Cancer*, **6**, 6–9.
35. López,C., Kleinheinz,K., Aukema,S.M., Rohde,M., Bernhart,S.H., Hübschmann,D., Wagener,R., Toprak,U.H., Raimondi,F., Kreuz,M. *et al.* (2019) Genomic and transcriptomic changes complement each other in the pathogenesis of sporadic Burkitt lymphoma. *Nat. Commun.*, **10**, 1459.
36. Grande,B.M., Gerhard,D.S., Jiang,A., Griner,N.B., Abramson,J.S., Alexander,T.B., Allen,H., Ayers,L.W., Bethony,J.M., Bhatia,K. *et al.* (2019) Genome-wide discovery of somatic coding and noncoding mutations in pediatric endemic and sporadic Burkitt lymphoma. *Blood*, **133**, 1313.
37. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
38. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10.
39. Song,L. and Florea,L. (2015) Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. *Gigascience*, **4**, 48.
40. Kopylova,E., Noé,L. and Touzet,H. (2012) SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, **28**, 3211–3217.
41. Howe,K.L., Achuthan,P., Allen,J., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Azov,A.G., Bennett,R., Bhai,J. *et al.* (2021) Ensembl 2021. *Nucleic Acids Res.*, **49**, D884–D891.
42. Hoffmann,S., Otto,C., Kurtz,S., Sharma,C.M., Khaitovich,P., Vogel,J., Stadler,P.F. and Hackermüller,J. (2009) Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput. Biol.*, **5**, e1000502.
43. Hoffmann,S., Otto,C., Doose,G., Tanzer,A., Langenberger,D., Christ,S., Kunz,M., Holdt,L.M., Teupser,D., Hackermüller,J. *et al.* (2014) A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. *Genome Biol.*, **15**, R34.
44. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
45. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
46. Clough,E. and Barrett,T. (2016) The Gene Expression Omnibus Database. In: *Methods in Molecular Biology*. NIH Public Access, Clifton, NJ, Vol. **1418**, pp. 93–110.
47. Zheng,R., Wan,C., Mei,S., Qin,Q., Wu,Q., Sun,H., Chen,C.-H., Brown,M., Zhang,X., Meyer,C.A. *et al.* (2019) Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Res.*, **47**, D729–D735.
48. Navarro Gonzalez,J., Zweig,A.S., Speir,M.L., Schmelter,D., Rosenbloom,K.R., Raney,B.J., Powell,C.C., Nassar,L.R., Maulding,N.D., Lee,C.M. *et al.* (2021) The UCSC Genome Browser database: 2021 update. *Nucleic Acids Res.*, **49**, D1046–D1057.
49. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
50. Schmitt,A.M., Garcia,J.T., Hung,T., Flynn,R.A., Shen,Y., Qu,K., Payumo,A.Y., Peres-da-Silva,A., Broz,D.K., Baum,R. *et al.* (2016) An inducible long noncoding RNA amplifies DNA damage signaling. *Nat. Genet.*, **48**, 1370–1376.
51. Chaudhary,R., Gryder,B., Woods,W.S., Subramanian,M., Jones,M.F., Li,X.L., Jenkins,L.M., Shabalina,S.A., Mo,M., Dasso,M. *et al.* (2017) Prosurvival long noncoding RNA PINCR regulates a subset of p53 targets in human colorectal cancer cells by binding to Matrin 3. *Elife*, **6**, e23244.
52. Adriaens,C., Standaert,L., Barra,J., Latil,M., Verfaillie,A., Kalev,P., Boeckx,B., Wijnhoven,P.W.G., Radaelli,E., Vermi,W. *et al.* (2016) p53 induces formation of NEAT1 lncRNA-containing paraspeckles that modulate replication stress response and chemosensitivity. *Nat. Med.*, **22**, 861–868.
53. Mello,S.S., Sinow,C., Raj,N., Mazur,P.K., Bieging-Rolett,K., Broz,D.K., Imam,J.F.C.C., Vogel,H., Wood,L.D., Sage,J. *et al.* (2017) Neat1 is a p53-inducible lincRNA essential for transformation suppression. *Genes Dev.*, **31**, 1095–1108.

54. Fishilevich,S., Nudel,R., Rappaport,N., Hadar,R., Plaschkes,I., Iny Stein,T., Rosen,N., Kohn,A., Twik,M., Safran,M. *et al.* (2017) GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database*, **2017**, bax028.
55. Whitfield,M.L., George,L.K., Grant,G.D. and Perou,C.M. (2006) Common markers of proliferation. *Nat. Rev. Cancer*, **6**, 99–106.
56. Whitfield,M.L., Sherlock,G., Saldanha,A.J., Murray,J.I., Ball,C.A., Alexander,K.E., Matese,J.C., Perou,C.M., Hurt,M.M., Brown,P.O. *et al.* (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, **13**, 1977–2000.
57. Müller,G.A., Stangner,K., Schmitt,T., Wintsche,A. and Engeland,K. (2017) Timing of transcription during the cell cycle: protein complexes binding to E2F, E2F/CLE, CDE/CHR, or CHR promoter elements define early and late cell cycle gene expression. *Oncotarget*, **8**, 97736–97748.
58. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
59. Fischer,M. (2016) p21 governs p53's repressive side. *Cell Cycle*, **15**, 2852–2853.
60. Coronel,L., Häckes,D., Schwab,K., Riege,K., Hoffmann,S. and Fischer,M. (2022) p53-mediated AKT and mTOR inhibition requires RFX7 and DDIT4 and depends on nutrient abundance. *Oncogene*, **41**, 1063–1069.
61. Barak,Y., Gottlieb,E., Juven-Gershon,T. and Oren,M. (1994) Regulation of mdm2 expression by p53: Alternative promoters produce transcripts with nonidentical translation potential. *Genes Dev.*, **8**, 1739–1749.
62. Sionov,R.V., Netzer,E. and Shaulian,E. (2013) Differential regulation of FBXW7 isoforms by various stress stimuli. *Cell Cycle*, **12**, 3547–3554.
63. Macosko,E.Z., Basu,A., Satija,R., Nemesh,J., Shekhar,K., Goldman,M., Tirosh,I., Bialas,A.R., Kamitaki,N., Martersteck,E.M. *et al.* (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, **161**, 1202–1214.
64. Dominguez,D., Tsai,Y.-H., Weatheritt,R., Wang,Y., Blencowe,B.J. and Wang,Z. (2016) An extensive program of periodic alternative splicing linked to cell cycle progression. *Elife*, **5**, e10288.
65. Liu,Y., Chen,S., Wang,S., Soares,F., Fischer,M., Meng,F., Du,Z., Lin,C., Meyer,C., DeCaprio,J.A. *et al.* (2017) Transcriptional landscape of the human cell cycle. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, 3473–3478.
66. Hao,Q., Zong,X., Sun,Q., Lin,Y.-C., Song,Y.J., Hashemikhabir,S., Hsu,R.Y.C., Kamran,M., Chaudhary,R., Tripathi,V. *et al.* (2020) The S-phase-induced lncRNA SUNO1 promotes cell proliferation by controlling YAP1/Hippo signaling pathway. *Elife*, **9**, e55102.

Discussion

TEs are present in virtually all eukaryotic genomes. Although the transposome varies in composition and magnitude along the tree of life, it has become evident that "jumping genes" are longtime companions of evolution that account for a substantial part of the known genomic code. Along with other non-coding but transcribed components of the genome, TEs were frequently considered as "junk" DNA. Recent research, however, gradually debunks this narrative. For instance, the Encyclopedia of DNA Elements project claimed that more than 80% of the human genome is functional, particularly outside of protein-coding genes [132]. While this assertion has sparked vicious criticism and raised questions about proper definitions of "function" [133], there is little doubt that non-coding transcripts have the potential to take over regulatory functions [82-84, 86]. Seven decades after the Nobel Prize-winning discovery of "controlling elements" by Barbara McClintock [4], TEs are still central to the genomes' "dark matter". Recent research suggests that McClintock's original characterization of the Ac-Ds family as a gene control system might well be true for other TEs as well. Given the vast repertoire of TFBSs nested within genomic TEs, they might well be essential drivers of evolution that help rewire gene regulatory networks. Furthermore, TEs may have been domesticated by their hosts in response to evolutionary pressures, *e.g.*, exposure to pathogens.

Since TEs compromise more than the half of the human genome [3], it is not surprising that at least a fraction of them is involved in regulatory processes of the host. In fact, a mounting number of studies reveal novel biological functions of TEs [22, 23, 37, 89, 134, 135]. In this light, the term "junk" DNA appears to be outdated and begs the question to which extent the line between the once invading elements and the host organism has already been crossed. Family level-based analyses of TE expression yielded important insights into the biological role of TEs [14, 74, 90, 136]. Naturally, the detection of differentially regulated TE expression in health, disease, or during aging is not sufficient to determine whether the TE transcription is cause or consequence. Additionally, it is necessary to investigate individual and hence locus-specific regulatory actions. By design, the analysis of (differential) TE expression using family-based approaches does not deliver the required level of resolution and renders functional follow-up experiments difficult to impossible. Moreover, expression data aggregated at the family-level may obfuscate important expression dynamics within TE families when family members are not coordinately up- or down-regulated.

In brief, accumulating reports of TE dysregulation in diseases such as cancer, neurodegenerative disorders, and aging in somatic cells reinforce the need of appropriate tools for locus-specific TE quantification. The aim of the first publication (M1) was a quantitative evaluation of different TE quantification strategies. Our benchmarks have indicated that a slight modification of an existing tool is sufficient to achieve surprisingly good benchmarks. Apparently, the diversity of related TE sequences is generally high enough that, in combination with EM methods, the number assigned reads allows an accurate estimation of the expression level. From this perspective, the study should encourage researchers to include TEs in their differential expression analyses.

3.1. Evaluation of tools with respect to locus-specific expression quantification of TEs

Ambiguously mapping reads are challenging for the quantification of expression. Hence, TE expression modules are generally not incorporated into standard transcript quantification pipelines. The tool evaluation in M1 was motivated by the lack of appropriate tools for locus-specific TE quantification at the beginning of the project. Therefore, we tried to answer the question whether software designed for the family based quantification is a sufficient basis for locus-specific TE expression analysis. While the project was ongoing, SQuIRE and Telescope became available claiming to have solved the locus-specific quantification problem. However, still no study was available evaluating the different tools with respect to locus specificity.

A high sequence similarity of particular TEs can be an indication for elements that are actively transposing within their host genome. Such TEs are especially prone to produce multi-mapping reads when short read sequencing technologies are used. Thus, active TEs are especially challenging for the alignment tools. Consistently, in contrast to the whole set of simulated TEs, the performances of the evaluated tools are decreased for elements with a small Kimura distance (≤ 5). However, the majority of TEs in the human genome are ancient [121] and accumulated a sufficient amount of sequence variations, *e.g.*, mutations, so that the majority of the tools accurately quantify their expression.

TEtools, originally developed for family-based analyses, is an example for a tool that addresses the multi-mapping read problem by randomly assigning multi-mapped reads to one of the identified loci. When used in locus-specific analyses, this strategy leads to a substantial overestimation of TE expression and thus a high number of false positive detected TEs. In comparison, tools that employ an EM strategy to solve the multi-mapping read problem perform much better. Also, EM-based tools consistently show an improved correlation between expected and detected read counts. In agreement with earlier reports, our simulation confirmed that the TE quantification performance is improved when paired-end data is used instead of single-end data [116]. Overall, SalmonTE outperforms all other tools in locus-specific expression analysis, which was surprising for us as it was also originally developed for family based quantification analysis.

Although we obtained similar results based on RNA-Seq data sets simulated with two distinct strategies, it must be noted that simulated data is a critical limitation of our study. While our simulations were based on specific models reflecting certain biases and errors of an RNA-Seq experiment, the simulated data must not be confused with real data. However, the similarity of our benchmark results across two distinct simulation strategies indicates that our results are sufficiently robust to make recommendations for the analysis of real RNA-Seq data.

The annotation of TEs is an ongoing endeavor. Thus, virtually all reference genomes have to be considered to be incomplete in terms of TE annotations [17, 107]. This fact leads to a further obstacle of our simulation, as we were able to simulate only annotated elements. Importantly, the human and mouse genomes also still contain actively transposing TEs [99, 106, 107, 137, 138], resulting in individual TE insertions that are unlikely to be reflected in the reference genomes and their annotations [107].Taking L1 elements into account, two individual human genomes differ on average at 285 sites [139]. The frequency of transpositions is assumed to be even higher in mice [140, 141]. In addition to the inter-individual variations, TEs are also active in somatic cells, *e.g.*, L1 elements are a driver of the genomic mosaicisms in brain-specific cells [69, 72, 142]. Obviously, these actively transposing elements are of special interest in the TE expression analysis. Unfortunately, they are also particularly hard to analyze and their exact measurements depends on multiple factors, *e.g.*, their length or the site of insertion, which cannot be simulated in a meaningful way.

As stated above, the quantification of TE expression works surprisingly well with minor adaptations of SalmonTE. However, our benchmarks also indicate that there is room for improvement, especially of the investigation of young elements. We assume that the observed shortcomings are best to be addressed by novel experimental strategies. Critically, the technical read length limitation of the Illumina sequencing platform is a main reason for the decreased performance quantifying young elements. For TEs with highly identical copies, reads are simply not long enough to span a sufficient number of polymorphisms. Third generation sequencing technologies, *e.g.*, PacBio SMRT seq [143, 144] and Oxford Nanopore [145], produce considerably longer reads that can help to overcome that problem. However, these technologies still suffer from higher error rates and lower throughput, compared to the short-read sequencing technologies, which is particularly challenging for the mapping process. Nevertheless, improvements of base calling algorithms and chemistries presumably provide a more accurate quantification in the future. Alternatively, a combination of short- and long-read sequencing technologies could obtain long reads that are subsequently corrected using the accurate short-read information [146].

In summary, M1 shows that locus-specific quantification of TEs is possible already with the sequencing technologies and bioinformatics tools that are currently available. Our proposed approach outlines a comparably convenient way to quantify TE expression under many different conditions. Notably, it encourages researchers to study TE expression by re-analyzing existing high-throughput sequencing data.

3.2. Locus-specific TE expression quantification during the process of aging

Age is a risk factor for many diseases [101, 147] and their effective prevention and treatment requires a detailed molecular understanding of the aging process at the molecular level. The investigation of the genome and its regulation is a major component of this complex endeavor. Recent studies have demonstrated that TEs play a more important role in aging than previously anticipated. The expression of certain TEs during aging can lead to sterile inflammation [14, 103] and the potential involvement of TEs in malignancies [9, 10] or neurological conditions [12, 40, 148] are evidence for a more fundamental role of the so-called "jumping genes". Despite such observations, it is not clear whether the dysregulation of TEs is a primary driver of the aging process [13]. In the genome, TEs are usually heavily methylated [149, 150] and silenced due to heterochromatin formation in somatic cells [151]. Since the chromatin architecture is dynamically changing during aging [152], silenced TEs can become accessible and reactivated. In line with this, the up-regulation of TE families has been reported during aging [14, 100, 137, 153] and in age-dependent diseases [9, 10, 148, 154]. However, family-based TE expression studies lack information on the expression dynamics of individual TEs and to not provide insight to their potential function as transcriptional regulators, *e.g.*, through the establishment of promoters or enhancers.

As stated in the introduction, the complexity of TEs and their potential biological functions is enormous. This underscores the need for locus-specific analyses to shed light on the expression profiles of individual TEs gain insights on the regulation of individual TEs. Our tool evaluation in M1 provided us the unique opportunity to apply the best performing tool that is currently available, *i.e.*, adjusted SalmonTE, to a data set of three different tissues (brain, skin, and blood) from aged mice to study TE expression at locus resolution. In particular, we compared six versus 24 month old male mice.

Intriguingly, our locus-specific expression analysis reveals complex TE expression patterns during aging. Beyond previous studies that detected an age dependent up-regulation of specific TE families in somatic cells [14, 137, 153], our data reveal a substantial number of TEs that are down-regulated in 24 compared to six month old mice. Importantly, individual TEs from the same family show divergent regulatory patterns during aging, which differ substantially from their family-based averages. The down-regulation of TEs evokes the question whether their suppression could have an aging-relevant role. Recently, it has been suggested that actively transcribed TEs, *i.e.*, members of LTR, LINEs, and SINEs, function as tumor suppressors in blind mole rats to compensate for a mutated p53 gene [136]. In brief, the authors provided evidence that TEs could act as an alarm system sensing cellular proliferation and triggering cell death via the cGAS-Sting pathway. The finding is reminiscent of other studies suggesting that cancer cells may down-regulate TEs to be invisible and protected against the host immune system [155, 156]. In blind mole rats, the leading cause for the activation of silenced TEs was attributed to a pervasive loss of DNA methylation in highly proliferating cells due to weakly functioning DNA methyltransferase (DNMT) 1 [136]. Interestingly, overexpression of DNMTs correlates with tumors aggressiveness [157, 158], which may lead to the silencing of TEs to avoid immune responses.

The epigenome is an important regulator of cell type-specific gene expression, and hence a quintessential maintainer of cell identity [159-161]. The cell type-specific configuration of chromatin necessarily also affects the accessibility of TEs. For example, TE-derived promoters have been shown to be used in a highly tissue-specific way in mouse development [134]. TE instances of specific TE families have been shown to overlap with enhancer-associated chromatin marks, *e.g.*, H3K4me1 in $CD8^+$ T-cells, and promote the expression of immune-related genes [89], also demonstrating the cell type specificity. Concordantly, we find that the majority of expressed TEs are exclusively expressed in one tissue. Just about 4% are expressed in all three tissues. Regarding the association of TEs and enhancers, we identified an expressed TE that overlaps with an enhancer which is associated with the *Fam126a* gene in blood. The expression of the TE and the gene was confirmed by reverse transcriptase semi-quantitative polymerase chain reaction and both, the TE and *Fam126a* displayed a trend of down-regulation during aging. The striking tissue specificity raised the question whether TEs merely piggyback on tissue-specific accessible regions or instead contribute to the regulation in the first place.

The genome-wide profiling of TSSs with CAGE-Seq allows identifying TEs that are independently expressed and regulated through their own promoter. Our enrichment analysis of independently expressed TEs corroborated the heterogeneous expression pattern across different tissues we observed in the RNA-Seq data. In skin, for example, we identified a tissue-specific enrichment of TSSs within several ERV families. Recently, a study suggested a “communication” between the skin and exogenous skin microbiota through ERV elements of host regulating inflammatory processes [162]. Moreover, our data indicates an exclusive enrichment of TSS within L1 (a LINE) elements in brain. LINEs are known to be mainly expressed in brain and have been implicated to have regulatory effects there [19, 163, 164]. Only the B1 family (a SINE) consistently enriched TSSs in all three tissues. This is an unexpected finding because TEs transcribed by RNA Pol III, like SINEs, do not contain the 5’ m7G-cap structure [112], and thus should be hidden from the CAGE-Seq approach. However, Alu elements, the primate-specific counterpart of B1, contain regulatory sequences which can be accessible for Pol II by mutations [165] and which offers an explanation for the TSSs we identified in the B1 family.

It is of note that PEAKachu [166], the tool we used for calling CAGE-Seq peaks, requires alignment files, which cannot be provided by SalmonTE, so that an allocation of multi-mapping reads did not take place. The exclusive use of uniquely aligned reads limits the identification of TSSs in TEs with high sequence similarity at their transcription initiation site. Consequently, the number of independently expressed TE regions is rather underestimated in our analysis.

Given that neighboring TEs in the genome often showed a continuous RNA-Seq expression signal, we grouped closely spaced TEs (distance ≤ 500bp) into TE regions. Analogously to individual TE loci, TE regions that intersected with a CAGE-Seq-derived TSS and showed sufficiently strong expression signals were considered to be independently expressed TE regions. Analyzing the distribution of the TSSs along the TE regions in skin and blood indicated that TE transcription is frequently initiated at the regions' 5'-end. This suggests, that the TE located 5' in a TE region frequently donates a TSS. In brain, however, this was not the case. Here, TSSs were most frequently located at the 3'-end of independent TE regions. The accurate localization of TSS within TEs facilitates the systematic inspection of putative TE promoter sequences.

Looking for common regulatory factors involved in the expression of TEs, we searched for TFBS motifs. TEs from all tissues showed a substantial enrichment of potential TFBS of the Sox family near their TSSs. The Sox family TFs share a high mobility group box domain that typically mediates DNA binding. Sox TFs are known to regulate neuronal differentiation and to be involved in adult neurogenesis [167]. Importantly, they may also have a role in the regulation of TEs. For instance, the expression and transposition of human L1 elements containing two Sox-binding sites within their 5' UTR were found to be negatively correlated with the expression of *Sox2* [71]. Another Sox family member contributing to neuronal development is *Sox5*, which is involved in controlling subtype-specific neuronal differentiation [168]. It has been reported that *Sox5* haploinsufficiency leads to the neurodevelopmental disorder Lamb-Shaffer syndrome [169] and *Sox5* may contribute to the development of autism spectrum disorders [170]. The enrichment of Sox TFBS in TE-derived brain-specific transcripts raised the question whether their expression mechanistically contribute to the regulatory roles of Sox TFs. Intriguingly, Sox motifs that are co-localized with TSSs in brain are likely caused by an L1 3'-end subfamily with a characteristic length between 950 and 1050 bp. In agreement with our finding, truncated L1 elements are known to enrich TSSs near their 3'-end [55]. In addition, L1 elements contain a weak polyadenylation site that leads to 3' read-through events [65, 171]. Thus, the independently expressed L1 3'-ends we identified may indicate a set of regulatory loci at which the L1 instance functions as a regulator for down-stream genes.

Our expression analysis indicates a surprising co-regulation of independently expressed TEs and their host genes, and, some of those pairs of TEs and host genes are located in gene clusters known to have fundamental roles in synaptic signal transduction or critical immunological functions. In brain, we identified the protocadherin cluster to be recurrently affected by differentially expressed TEs and genes. The expression of genes in the protocadherin cluster is highly randomized and depends on the expression of anti-sense RNAs [172]. The cluster equips each neuron with a unique set of cell surface proteins. This mechanism is critical for avoiding the connection of dendrites to their own soma [173]. Furthermore, TSS-carrying TEs are localized in the skint- and keratin clusters in skin. Genes of the skint family have an important role in the development of the dominant T cell compartment in the epidermis - Vγ5Vδ1dendritic epidermal T-cells, which are a subset of γδT cells that thwart against infections and tumor development [174-176]. The keratin cluster emerged from gene duplications and builds the largest subset of intermediate filament genes [177]. Keratin genes are responsible for keratin intermediate filaments that form important barriers, and mice lacking keratin genes exhibit severe epidermal barrier damage leading to death [178]. Our finding poses the exciting question about the role of TEs in stochastic expression of genes from these clusters. We hypothesize that TEs provide regulatory platforms that enable distinct expression patterns from these gene clusters in individual cells. In particular, the protocadherin and keratin clusters exhibit remarkable evolutionary conservation [177, 179, 180], with TEs providing on of the few possibilities to alter their regulation.

The co-regulation of TEs and their host genes suggests a common function in certain biological pathways. It is noticeable that genes associated with independent TEs are enriched in GO terms for highly tissue-specific function. Genes involved in neuronal synapse plasticity and connectivity were particularly enriched for independent TEs in brain. We found that those genes were particularly prone to the host TEs within their introns. Notably, neuronal activity was shown to trigger DNA double strand breaks (DSBs) that induce the expression of genes crucial for experience-driven changes to synapses, learning, and memory [181]. Since proteins encoded by TEs can induce DSBs, TEs may be involved in this process. In contrast to genes with tissue-specific functions, genes with general cell functions tasks, *e.g.*, genes encoding for RISC complex, immunoglobulin complex, and nucleosome, harbor less TEs within their introns. Thus we hypothesize that the accumulation of TEs may be evolutionarily beneficial in cell type-specific genes but less so in genes with general roles.

Taken together, the locus-specific characterization of TEs resolved the expression dynamics within TE families and revealed that TEs are as frequently down- as up-regulated during aging. The integration of CAGE-Seq and RNA-Seq data provides a catalogue of independently expressed TE regions and their associated genes, which are largely involved in tissue-specific processes. In addition, our analysis strongly suggests the involvement of Sox TFs in the regulation of TEs. Overall, our study challenges the narrative of an entirely detrimental role of TE expression during aging and suggests important roles for TEs in shaping the distinct transcriptional landscapes in tissues and individual cells.

3.3. Blueprint of a data base for differential expression data

Our locus-specific TE expression data sets from aged mice cover differential expression data on thousands of individual TEs. To make the data swiftly available and reusable by the research communities, I envision to create a web-based atlas. In M3, I implemented a data structure and web interface for a web atlas that allows easy access to the differential expression information for genes of interest. The web atlas created in M3, TargetGeneReg 2.0 (http://www.targetgenereg.org), provides information on p53 and cell cycle-dependent gene regulation and serves as a blueprint for web atlases that enable easy access to differential expression data. The web interface contains a search bar that allows the user to enter their gene of interest and get access to differential expression profiles and transcription factor binding data from multiple data sets.

The backbone of TargetGeneReg2.0 is based on shiny [182], an R package that allows to build interactive web applications within R. Through the R universe, packages, thousands of which have been made available by the R user community, can be integrated. One class of such packages integrates read-to-use data handling methods that simplify parts oft the website, *e.g.*, a highly-efficient search function. The ready-to-use data processing methods of TargetGeneReg2.0 are concerted to a specific data structure. Therefore, it is critical that each data set is structured identically. A ready-to-use supporter script structurers and merges data to ensure proper processing. The backend data structure allows for a seamless integration of new data without website shut-downs. In addition, the modular structure of the website and its backend provides a blueprint that can be adapted to provide any differential expression data of interest. Therefore, an extension of the data structure generated for M3 will enable the simultaneous accessibility of expression data on host genes and their associated TEs.

The modular design of the search engine in principle enables the integration of search requests considering individual TEs. However, fast changing identifiers of individual TEs constitute challenges to provide an intuitive search for specific TEs. The current data structure already deposits genomic coordinates of each gene, which would allow the integration of a search engine that is based on genomic coordinates rather than names or other identifiers. Thus, genomic coordinates may provide a convenient way to search for individual TEs of interest.

The web atlas could be extended to include results of publicly available RNA-Seq data sets, which would provide a solid basis for comprehensive analyses, *e.g.,* meta-analyses. The latter are powerful tools to gain deeper insights into molecular biological processes and mechanisms, with the advantage of increased statistical power compared to single-case studies [183]. Overall, the easy accessibility of TE expression data via a web-atlas would enable scientists to validate results and to develop new hypothesis regarding biological functions of TEs.

Conclusion & Outlook

As first part of this thesis, I evaluated TE quantification tools according to their performances with respect to locus-specific quantification of TEs based on comprehensive simulations. Within the limits of the simulation, a tool originally designed for family-level quantification of TEs, SalmonTE, outperformed all other tools following minor adaptations of the reference library. The results indicate that many individual TE instances can be quantified with sufficient confidence using currently available algorithms. In addition, the accurate quantification of individual TEs provides an opportunity for integration into standard expression quantification pipelines.

When I employed SalmonTE to assess the differential expression of individual TEs in young and old mice, I found that TEs are commonly down- and up-regulated during aging, challenging the narrative of TEs escaping repression during aging at large. The down-regulation of TEs in aged mice raises questions concerning their biological consequences. The question of the biological functions is reinforced by the integration of CAGE-Seq data. We uncovered stretches of expressed TEs (TE regions) sharing common TSSs, providing transcripts with unknown functions. Therefore, in the future, it would be of great interest to verify and extend the catalogue of independently expressed TE regions using long-read sequencing technologies, *e.g.*, PacBio SMRT seq or Oxford Nanopore. Additionally, a genome-wide assessment of TE-induced transcription termination sites (TTSs) could provide additional insights into the regulatory roles of TEs as they may provide alternative TTS for genes. Moreover, such analyses can help to annotate TE-induced transcripts and may contribute to a more comprehensive TE transcript catalogue.

In addition to down-regulated TEs, independently expressed TE regions are associated with highly tissue-specific genes, of which those associated with neuronal functions in brain are particularly interesting. The brain has an outstanding role during the evolution of humans. Its comparably fast evolution is difficult to explain with random base mutations model, especially in a species with such a small population size and long life span. TEs provide templates with potential functions that can be distributed throughout the genome and affect the expression of multiple genes. A quick distribution of TE-derived regulatory elements provides the evolutionary advantage that regulation mechanisms with similar functions do not need to evolve independently at multiple loci. TEs are already known as important contributors to genomic mosaicism in brain. Consequently, the TE composition can be highly variable in individual cells from the same brain. Therefore, it is likely that also the expression pattern is even more complex than shown in this study and requires locus-specific TE quantification in single-cell studies. Altogether, the independently expressed TE regions we identified will be a promising starting point to study their biological roles. Disruption of selected

TE regions, *e.g.*, using CRISPR-Cas technology, could help elucidate the effect on their host genes. Moreover, over-expression or silencing of Sox proteins could reveal their regulatory impact on individual TEs.

Notably, genome-wide analyses such as in the case of M2 combine data from thousands of loci and it can be difficult to assess the regulation of individual loci, such as a gene or TE of interest. Therefore, I envision to make the differential TE expression data easily accessible based on the blueprint web-atlas I generated in M3. In addition to data obtained from M2, I envision to feed that web-atlas with differential expression information on individual TEs across many more cell types and conditions through systematic re-analysis of publicly available RNA-Seq data sets using the modified SalmonTE I identified in M1. Such an extended database would provide a strong basis for meta-analyses of TE expression across multiple experimental setups. An easy availability through a web-atlas can enable scientists to validate results and develop new hypotheses on the regulation and function of TEs.

Overall, this thesis demonstrates the feasibility of locus-specific TE expression analyses and increases our understanding of the complexity of TE expression during aging. My locus-specific TE expression analysis challenges models that ascribe largely detrimental roles to TE expression during aging. Moreover, the tissue-specific co-regulation of TEs and their host genes highlights a potential influence of TE and host gene on each other. Therefore, the results encourage intensifying research into locus-specific TE expression analysis, to gain a better understanding of the biological functions, interactions, and regulation of TEs. This thesis shall motivate to reconsider the roles of TEs in development and disease.

Bibliography

1. Watson, J.D. and F.H. Crick, *Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid.* Nature, 1953. **171**(4356): p. 737-8.
2. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.
3. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
4. McClintock, B., *Controlling elements and the gene.* Cold Spring Harb Symp Quant Biol, 1956. **21**: p. 197-216.
5. Mouse Genome Sequencing, C., et al., *Initial sequencing and comparative analysis of the mouse genome.* Nature, 2002. **420**(6915): p. 520-62.
6. Howe, K., et al., *The zebrafish reference genome sequence and its relationship to the human genome.* Nature, 2013. **496**(7446): p. 498-503.
7. Schnable, P.S., et al., *The B73 maize genome: complexity, diversity, and dynamics.* Science, 2009. **326**(5956): p. 1112-5.
8. Jiao, Y., et al., *Improved maize reference genome with single-molecule technologies.* Nature, 2017. **546**(7659): p. 524-527.
9. Burns, K.H., *Transposable elements in cancer.* Nat Rev Cancer, 2017. **17**(7): p. 415-424.
10. Clayton, E.A., et al., *Patterns of Transposable Element Expression and Insertion in Cancer.* Front Mol Biosci, 2016. **3**: p. 76.
11. Scott, E.C., et al., *A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer.* Genome Res, 2016. **26**(6): p. 745-55.
12. Terry, D.M. and S.E. Devine, *Aberrantly High Levels of Somatic LINE-1 Expression and Retrotransposition in Human Neurological Disorders.* Front Genet, 2019. **10**: p. 1244.
13. Gorbunova, V., et al., *The role of retrotransposable elements in ageing and age-associated diseases.* Nature, 2021. **596**(7870): p. 43-53.
14. De Cecco, M., et al., *L1 drives IFN in senescent cells and promotes age-associated inflammation.* Nature, 2019. **566**(7742): p. 73-78.
15. Miller, K.N., et al., *Cytoplasmic DNA: sources, sensing, and role in aging and disease.* Cell, 2021. **184**(22): p. 5506-5526.
16. Romanish, M.T., C.J. Cohen, and D.L. Mager, *Potential mechanisms of endogenous retroviral-mediated genomic instability in human cancer.* Semin Cancer Biol, 2010. **20**(4): p. 246-53.
17. Bourque, G., et al., *Ten things you should know about transposable elements.* Genome Biol, 2018. **19**(1): p. 199.
18. Bakoulis, S., et al., *Endogenous retroviruses co-opted as divergently transcribed regulatory elements shape the regulatory landscape of embryonic stem cells.* Nucleic Acids Res, 2022. **50**(4): p. 2111-2127.
19. Wanichnopparat, W., et al., *Genes associated with the cis-regulatory functions of intragenic LINE-1 elements.* BMC Genomics, 2013. **14**: p. 205.
20. Bourque, G., et al., *Evolution of the mammalian transcription factor binding repertoire via transposable elements.* Genome Res, 2008. **18**(11): p. 1752-62.
21. Schmidt, D., et al., *Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages.* Cell, 2012. **148**(1-2): p. 335-48.

22. Cheng, Y., et al., *Increased processing of SINE B2 ncRNAs unveils a novel type of transcriptome deregulation in amyloid beta neuropathology.* Elife, 2020. **9**.
23. Zovoilis, A., et al., *Destabilization of B2 RNA by EZH2 Activates the Stress Response.* Cell, 2016. **167**(7): p. 1788-1802 e13.
24. Wicker, T., et al., *A unified classification system for eukaryotic transposable elements.* Nat Rev Genet, 2007. **8**(12): p. 973-82.
25. Wells, J.N. and C. Feschotte, *A Field Guide to Eukaryotic Transposable Elements.* Annu Rev Genet, 2020. **54**: p. 539-561.
26. Boeke, J.D., et al., *Ty elements transpose through an RNA intermediate.* Cell, 1985. **40**(3): p. 491-500.
27. Greenblatt, I.M. and R. Alexander Brink, *Transpositions of Modulator in Maize into Divided and Undivided Chromosome Segments.* Nature, 1963. **197**(4865): p. 412-413.
28. Rubin, G.M., M.G. Kidwell, and P.M. Bingham, *The molecular basis of P-M hybrid dysgenesis: The nature of induced mutations.* Cell, 1982. **29**(3): p. 987-994.
29. Dewannieux, M. and T. Heidmann, *Endogenous retroviruses: acquisition, amplification and taming of genome invaders.* Curr Opin Virol, 2013. **3**(6): p. 646-56.
30. McCarthy, E.M. and J.F. McDonald, *Long terminal repeat retrotransposons of Mus musculus.* Genome Biol, 2004. **5**(3): p. R14.
31. Nefedova, L. and A. Kim, *Mechanisms of LTR-Retroelement Transposition: Lessons from Drosophila melanogaster.* Viruses, 2017. **9**(4).
32. Mager, D.L. and J.P. Stoye, *Mammalian Endogenous Retroviruses.* Microbiol Spectr, 2015. **3**(1): p. MDNA3-0009-2014.
33. Magiorkinis, G., et al., *Env-less endogenous retroviruses are genomic superspreaders.* Proc Natl Acad Sci U S A, 2012. **109**(19): p. 7385-90.
34. Wilhelm, M. and F.X. Wilhelm, *Reverse transcription of retroviruses and LTR retrotransposons.* Cell Mol Life Sci, 2001. **58**(9): p. 1246-62.
35. Havecker, E.R., X. Gao, and D.F. Voytas, *The diversity of LTR retrotransposons.* Genome Biol, 2004. **5**(6): p. 225.
36. Criscione, S.W., et al., *Transcriptional landscape of repetitive elements in normal and cancer human cells.* BMC Genomics, 2014. **15**: p. 583.
37. Thompson, P.J., T.S. Macfarlan, and M.C. Lorincz, *Long Terminal Repeats: From Parasitic Elements to Building Blocks of the Transcriptional Regulatory Repertoire.* Mol Cell, 2016. **62**(5): p. 766-76.
38. Reilly, M.T., et al., *The role of transposable elements in health and diseases of the central nervous system.* J Neurosci, 2013. **33**(45): p. 17577-86.
39. Zhang, X., R. Zhang, and J. Yu, *New Understanding of the Relevant Role of LINE-1 Retrotransposition in Human Disease and Immune Modulation.* Front Cell Dev Biol, 2020. **8**: p. 657.
40. Burns, K.H., *Our Conflict with Transposable Elements and Its Implications for Human Disease.* Annu Rev Pathol, 2020. **15**: p. 51-70.
41. Ostertag, E.M. and H.H. Kazazian, Jr., *Biology of mammalian L1 retrotransposons.* Annu Rev Genet, 2001. **35**: p. 501-38.
42. Dewannieux, M., C. Esnault, and T. Heidmann, *LINE-mediated retrotransposition of marked Alu sequences.* Nat Genet, 2003. **35**(1): p. 41-8.
43. Dewannieux, M. and T. Heidmann, *L1-mediated retrotransposition of murine B1 and B2 SINEs recapitulated in cultured cells.* J Mol Biol, 2005. **349**(2): p. 241-7.
44. Raiz, J., et al., *The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery.* Nucleic Acids Res, 2012. **40**(4): p. 1666-83.
45. Jurka, J., *Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons.* Proc Natl Acad Sci U S A, 1997. **94**(5): p. 1872-7.

46. Senft, A.D. and T.S. Macfarlan, *Transposable elements shape the evolution of mammalian development.* Nat Rev Genet, 2021. **22**(11): p. 691-711.
47. Platt, R.N., 2nd, M.W. Vandewege, and D.A. Ray, *Mammalian transposable elements and their impacts on genome evolution.* Chromosome Res, 2018. **26**(1-2): p. 25-43.
48. Kidwell, M.G. and D.R. Lisch, *Transposable elements and host genome evolution.* Trends in Ecology & Evolution, 2000. **15**(3): p. 95-99.
49. Zhou, W., et al., *DNA methylation enables transposable element-driven genome expansion.* Proc Natl Acad Sci U S A, 2020. **117**(32): p. 19359-19366.
50. Treangen, T.J. and S.L. Salzberg, *Repetitive DNA and next-generation sequencing: computational challenges and solutions.* Nat Rev Genet, 2011. **13**(1): p. 36-46.
51. Scott, A.F., et al., *Origin of the human L1 elements: proposed progenitor genes deduced from a consensus DNA sequence.* Genomics, 1987. **1**(2): p. 113-25.
52. Malik, H.S., W.D. Burke, and T.H. Eickbush, *The age and evolution of non-LTR retrotransposable elements.* Mol Biol Evol, 1999. **16**(6): p. 793-805.
53. Hancks, D.C. and H.H. Kazazian, Jr., *Roles for retrotransposon insertions in human disease.* Mob DNA, 2016. **7**: p. 9.
54. Sokolowski, M., et al., *Truncated ORF1 proteins can suppress LINE-1 retrotransposition in trans.* Nucleic Acids Res, 2017. **45**(9): p. 5294-5308.
55. Faulkner, G.J., et al., *The regulated retrotransposon transcriptome of mammalian cells.* Nat Genet, 2009. **41**(5): p. 563-71.
56. Ohtani, H. and Y.W. Iwasaki, *Rewiring of chromatin state and gene expression by transposable elements.* Dev Growth Differ, 2021. **63**(4-5): p. 262-273.
57. Kimura, M., *A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences.* Journal of molecular evolution, 1980. **16**(2): p. 111-120.
58. Cowley, M. and R.J. Oakey, *Transposable elements re-wire and fine-tune the transcriptome.* PLoS Genet, 2013. **9**(1): p. e1003234.
59. Fablet, M. and C. Vieira, *Evolvability, epigenetics and transposable elements.* Biomol Concepts, 2011. **2**(5): p. 333-41.
60. Oliver, K.R. and W.K. Greene, *Transposable elements: powerful facilitators of evolution.* Bioessays, 2009. **31**(7): p. 703-14.
61. Marques, A.C., et al., *Emergence of young human genes after a burst of retroposition in primates.* PLoS Biol, 2005. **3**(11): p. e357.
62. Babarinde, I.A., et al., *Transposable element sequence fragments incorporated into coding and noncoding transcripts modulate the transcriptome of human pluripotent stem cells.* Nucleic Acids Res, 2021. **49**(16): p. 9132-9153.
63. Mi, S., et al., *Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis.* Nature, 2000. **403**(6771): p. 785-9.
64. Ono, R., et al., *Deletion of Peg10, an imprinted gene acquired from a retrotransposon, causes early embryonic lethality.* Nat Genet, 2006. **38**(1): p. 101-6.
65. Moran, J.V., R.J. DeBerardinis, and H.H. Kazazian, Jr., *Exon shuffling by L1 retrotransposition.* Science, 1999. **283**(5407): p. 1530-4.
66. Goodier, J.L., E.M. Ostertag, and H.H. Kazazian, Jr., *Transduction of 3'-flanking sequences is common in L1 retrotransposition.* Hum Mol Genet, 2000. **9**(4): p. 653-7.
67. Esnault, C., J. Maestre, and T. Heidmann, *Human LINE retrotransposons generate processed pseudogenes.* Nat Genet, 2000. **24**(4): p. 363-7.
68. Wei, W., et al., *Human L1 retrotransposition: cis preference versus trans complementation.* Mol Cell Biol, 2001. **21**(4): p. 1429-39.
69. Evrony, G.D., et al., *Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain.* Cell, 2012. **151**(3): p. 483-96.

70. Erwin, J.A., et al., *L1-associated genomic regions are deleted in somatic cells of the healthy human brain.* Nat Neurosci, 2016. **19**(12): p. 1583-1591.
71. Muotri, A.R., et al., *Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition.* Nature, 2005. **435**(7044): p. 903-10.
72. Upton, K.R., et al., *Ubiquitous L1 mosaicism in hippocampal neurons.* Cell, 2015. **161**(2): p. 228-39.
73. Bodea, G.O., E.G.Z. McKelvey, and G.J. Faulkner, *Retrotransposon-induced mosaicism in the neural genome.* Open Biol, 2018. **8**(7).
74. Bedrosian, T.A., et al., *Early life experience drives structural variation of neural genomes in mice.* Science, 2018. **359**(6382): p. 1395-1399.
75. Jönsson, M.E., et al., *Activation of endogenous retroviruses during brain development causes an inflammatory response.* EMBO J, 2021. **40**(9): p. e106423.
76. Ali, A., K. Han, and P. Liang, *Role of Transposable Elements in Gene Regulation in the Human Genome.* Life (Basel), 2021. **11**(2).
77. Kim, D.S. and Y. Hahn, *Identification of human-specific transcript variants induced by DNA insertions in the human genome.* Bioinformatics, 2011. **27**(1): p. 14-21.
78. Casacuberta, E. and J. Gonzalez, *The impact of transposable elements in environmental adaptation.* Mol Ecol, 2013. **22**(6): p. 1503-17.
79. Dixon, J.R., et al., *Chromatin architecture reorganization during stem cell differentiation.* Nature, 2015. **518**(7539): p. 331-6.
80. Diehl, A.G., N. Ouyang, and A.P. Boyle, *Transposable elements contribute to cell and species-specific chromatin looping and gene regulation in mammalian genomes.* Nat Commun, 2020. **11**(1): p. 1796.
81. Quinodoz, S. and M. Guttman, *Long noncoding RNAs: an emerging link between gene regulation and nuclear organization.* Trends Cell Biol, 2014. **24**(11): p. 651-63.
82. Pauli, A., J.L. Rinn, and A.F. Schier, *Non-coding RNAs as regulators of embryogenesis.* Nat Rev Genet, 2011. **12**(2): p. 136-49.
83. Chew, C.L., et al., *Noncoding RNAs: Master Regulators of Inflammatory Signaling.* Trends Mol Med, 2018. **24**(1): p. 66-84.
84. Kugel, J.F. and J.A. Goodrich, *Non-coding RNAs: key regulators of mammalian transcription.* Trends Biochem Sci, 2012. **37**(4): p. 144-51.
85. Lu, X., et al., *The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity.* Nat Struct Mol Biol, 2014. **21**(4): p. 423-5.
86. Fort, V., G. Khelifi, and S.M.I. Hussein, *Long non-coding RNAs and transposable elements: A functional relationship.* Biochim Biophys Acta Mol Cell Res, 2021. **1868**(1): p. 118837.
87. Chuong, E.B., N.C. Elde, and C. Feschotte, *Regulatory activities of transposable elements: from conflicts to benefits.* Nat Rev Genet, 2017. **18**(2): p. 71-86.
88. Karijolich, J., et al., *Genome-wide mapping of infection-induced SINE RNAs reveals a role in selective mRNA export.* Nucleic Acids Res, 2017. **45**(10): p. 6194-6208.
89. Ye, M., et al., *Specific subfamilies of transposable elements contribute to different domains of T lymphocyte enhancers.* Proc Natl Acad Sci U S A, 2020. **117**(14): p. 7905-7916.
90. Wang, J., et al., *Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells.* Nature, 2014. **516**(7531): p. 405-9.
91. Volff, J.N., *Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes.* Bioessays, 2006. **28**(9): p. 913-22.
92. Feschotte, C., *Transposable elements and the evolution of regulatory networks.* Nat Rev Genet, 2008. **9**(5): p. 397-405.

93. Ayarpadikannan, S. and H.S. Kim, *The impact of transposable elements in genome evolution and genetic instability and their implications in various diseases.* Genomics Inform, 2014. **12**(3): p. 98-104.
94. Kirkwood, T.B. and S.N. Austad, *Why do we age?* Nature, 2000. **408**(6809): p. 233-8.
95. Oeppen, J. and J.W. Vaupel, *Demography. Broken limits to life expectancy.* Science, 2002. **296**(5570): p. 1029-31.
96. Vaupel, J.W., *Biodemography of human ageing.* Nature, 2010. **464**(7288): p. 536-42.
97. de Almeida, A., T.P. Ribeiro, and I.A. de Medeiros, *Aging: Molecular Pathways and Implications on the Cardiovascular System.* Oxid Med Cell Longev, 2017. **2017**: p. 7941563.
98. Slotkin, R.K. and R. Martienssen, *Transposable elements and the epigenetic regulation of the genome.* Nat Rev Genet, 2007. **8**(4): p. 272-85.
99. Deniz, O., J.M. Frost, and M.R. Branco, *Regulation of transposable elements by DNA modifications.* Nat Rev Genet, 2019. **20**(7): p. 417-431.
100. Yang, N., et al., *Transposable element landscapes in aging Drosophila.* PLoS Genet, 2022. **18**(3): p. e1010024.
101. Lopez-Otin, C., et al., *The hallmarks of aging.* Cell, 2013. **153**(6): p. 1194-217.
102. Franceschi, C., et al., *Inflammaging and anti-inflammaging: a systemic perspective on aging and longevity emerged from studies in humans.* Mech Ageing Dev, 2007. **128**(1): p. 92-105.
103. Simon, M., et al., *LINE1 Derepression in Aged Wild-Type and SIRT6-Deficient Mice Drives Inflammation.* Cell Metab, 2019. **29**(4): p. 871-885 e5.
104. Levin, H.L. and J.V. Moran, *Dynamic interactions between transposable elements and their hosts.* Nat Rev Genet, 2011. **12**(9): p. 615-27.
105. Chu, C., et al., *Comprehensive identification of transposable element insertions using multiple sequencing technologies.* Nat Commun, 2021. **12**(1): p. 3836.
106. Brouha, B., et al., *Hot L1s account for the bulk of retrotransposition in the human population.* Proc Natl Acad Sci U S A, 2003. **100**(9): p. 5280-5.
107. Beck, C.R., et al., *LINE-1 retrotransposition activity in human genomes.* Cell, 2010. **141**(7): p. 1159-70.
108. Sassaman, D.M., et al., *Many human L1 elements are capable of retrotransposition.* Nat Genet, 1997. **16**(1): p. 37-43.
109. Gasior, S.L., et al., *The human LINE-1 retrotransposon creates DNA double-strand breaks.* J Mol Biol, 2006. **357**(5): p. 1383-93.
110. Hedges, D.J. and P.L. Deininger, *Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity.* Mutat Res, 2007. **616**(1-2): p. 46-59.
111. Jacobs, F.M., et al., *An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons.* Nature, 2014. **516**(7530): p. 242-5.
112. Deininger, P., *Alu elements: know the SINEs.* Genome Biol, 2011. **12**(12): p. 236.
113. Royo, H., M.B. Stadler, and A. Peters, *Alternative Computational Analysis Shows No Evidence for Nucleosome Enrichment at Repetitive Sequences in Mammalian Spermatozoa.* Dev Cell, 2016. **37**(1): p. 98-104.
114. Marinov, G.K., et al., *Pitfalls of mapping high-throughput sequencing data to repetitive sequences: Piwi's genomic targets still not identified.* Dev Cell, 2015. **32**(6): p. 765-71.
115. Chhangawala, S., et al., *The impact of read length on quantification of differentially expressed genes and splice junction detection.* Genome Biology, 2015. **16**(1).
116. Sexton, C.E. and M.V. Han, *Paired-end mappability of transposable elements in the human genome.* Mob DNA, 2019. **10**: p. 29.
117. Teissandier, A., et al., *Tools and best practices for retrotransposon analysis using high-throughput sequencing data.* Mob DNA, 2019. **10**: p. 52.

118. Jin, Y., et al., *TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets.* Bioinformatics, 2015. **31**(22): p. 3593-9.
119. Lerat, E., et al., *TEtools facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of piRNA genes.* Nucleic Acids Res, 2017. **45**(4): p. e17.
120. Jeong, H.-H., et al., *An ultra-fast and scalable quantification pipeline for transposable elements from next generation sequencing data.* 2018: p. 168-179.
121. Yang, W.R., et al., *SQuIRE reveals locus-specific regulation of interspersed repeat expression.* Nucleic Acids Res, 2019. **47**(5): p. e27.
122. Bendall, M.L., et al., *Telescope: Characterization of the retrotranscriptome by accurate estimation of transposable element expression.* PLoS Comput Biol, 2019. **15**(9): p. e1006453.
123. Patro, R., et al., *Salmon provides fast and bias-aware quantification of transcript expression.* Nat Methods, 2017. **14**(4): p. 417-419.
124. Lanciano, S. and G. Cristofari, *Measuring and interpreting transposable element expression.* Nat Rev Genet, 2020. **21**(12): p. 721-736.
125. Clough, E. and T. Barrett, *The Gene Expression Omnibus Database.* Methods Mol Biol, 2016. **1418**: p. 93-110.
126. Haidich, A.B., *Meta-analysis in medical research.* Hippokratia, 2010. **14**(Suppl 1): p. 29-37.
127. Mikolajewicz, N. and S.V. Komarova, *Meta-Analytic Methodology for Basic Research: A Practical Guide.* Front Physiol, 2019. **10**: p. 203.
128. Fischer, M. and S. Hoffmann, *Synthesizing genome regulation data with vote-counting.* Trends Genet, 2022. **38**(12): p. 1208-1216.
129. Papatheodorou, I., et al., *Expression Atlas: gene and protein expression across multiple studies and organisms.* Nucleic Acids Res, 2018. **46**(D1): p. D246-D251.
130. Ringwald, M., et al., *Mouse Genome Informatics (MGI): latest news from MGD and GXD.* Mamm Genome, 2022. **33**(1): p. 4-18.
131. Makalowski, W., *Genomics. Not junk after all.* Science, 2003. **300**(5623): p. 1246-7.
132. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
133. Graur, D., et al., *On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE.* Genome Biol Evol, 2013. **5**(3): p. 578-90.
134. Miao, B., et al., *Tissue-specific usage of transposable element-derived promoters in mouse development.* Genome Biol, 2020. **21**(1): p. 255.
135. Garcia-Perez, J.L., T.J. Widmann, and I.R. Adams, *The impact of transposable elements on mammalian development.* Development, 2016. **143**(22): p. 4101-4114.
136. Zhao, Y., et al., *Transposon-triggered innate immune response confers cancer resistance to the blind mole rat.* Nat Immunol, 2021. **22**(10): p. 1219-1230.
137. De Cecco, M., et al., *Transposable elements become active and mobile in the genomes of aging mammalian somatic tissues.* Aging (Albany NY), 2013. **5**(12): p. 867-83.
138. Mills, R.E., et al., *Which transposable elements are active in the human genome?* Trends Genet, 2007. **23**(4): p. 183-91.
139. Ewing, A.D. and H.H. Kazazian, Jr., *High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes.* Genome Res, 2010. **20**(9): p. 1262-70.
140. Maksakova, I.A., et al., *Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line.* PLoS Genet, 2006. **2**(1): p. e2.

141. Zhang, Y., et al., *Genome-wide assessments reveal extremely high levels of polymorphism of two active families of mouse endogenous retroviral elements.* PLoS Genet, 2008. **4**(2): p. e1000007.
142. Sanchez-Luque, F.J., et al., *LINE-1 Evasion of Epigenetic Repression in Humans.* Mol Cell, 2019. **75**(3): p. 590-604 e12.
143. Eid, J., et al., *Real-time DNA sequencing from single polymerase molecules.* Science, 2009. **323**(5910): p. 133-8.
144. Rhoads, A. and K.F. Au, *PacBio Sequencing and Its Applications.* Genomics Proteomics Bioinformatics, 2015. **13**(5): p. 278-89.
145. Quick, J., A.R. Quinlan, and N.J. Loman, *A reference bacterial genome dataset generated on the MinION portable single-molecule nanopore sequencer.* Gigascience, 2014. **3**: p. 22.
146. Au, K.F., et al., *Improving PacBio long read accuracy by short read alignment.* PLoS One, 2012. **7**(10): p. e46679.
147. Niccoli, T. and L. Partridge, *Ageing as a risk factor for disease.* Curr Biol, 2012. **22**(17): p. R741-52.
148. Guo, C., et al., *Tau Activates Transposable Elements in Alzheimer's Disease.* Cell Rep, 2018. **23**(10): p. 2874-2880.
149. Kochanek, S., D. Renz, and W. Doerfler, *DNA methylation in the Alu sequences of diploid and haploid primary human cells.* EMBO J, 1993. **12**(3): p. 1141-51.
150. Bestor, T.H., *The host defence function of genomic methylation patterns.* Novartis Found Symp, 1998. **214**: p. 187-95; discussion 195-9, 228-32.
151. Lippman, Z., et al., *Role of transposable elements in heterochromatin and epigenetic control.* Nature, 2004. **430**(6998): p. 471-6.
152. O'Sullivan, R.J. and J. Karlseder, *The great unravelling: chromatin as a modulator of the aging process.* Trends Biochem Sci, 2012. **37**(11): p. 466-76.
153. Li, W., et al., *Activation of transposable elements during aging and neuronal decline in Drosophila.* Nat Neurosci, 2013. **16**(5): p. 529-31.
154. Payer, L.M. and K.H. Burns, *Transposable elements in human genetic disease.* Nat Rev Genet, 2019. **20**(12): p. 760-772.
155. Cuellar, T.L., et al., *Silencing of retrotransposons by SETDB1 inhibits the interferon response in acute myeloid leukemia.* J Cell Biol, 2017. **216**(11): p. 3535-3549.
156. Sheng, W., et al., *LSD1 Ablation Stimulates Anti-tumor Immunity and Enables Checkpoint Blockade.* Cell, 2018. **174**(3): p. 549-563 e19.
157. Ma, H.S., et al., *Overexpression of DNA (Cytosine-5)-Methyltransferase 1 (DNMT1) And DNA (Cytosine-5)-Methyltransferase 3A (DNMT3A) Is Associated with Aggressive Behavior and Hypermethylation of Tumor Suppressor Genes in Human Pituitary Adenomas.* Med Sci Monit, 2018. **24**: p. 4841-4850.
158. Zhang, W. and J. Xu, *DNA methyltransferases and their roles in tumorigenesis.* Biomark Res, 2017. **5**: p. 1.
159. Ernst, J., et al., *Mapping and analysis of chromatin state dynamics in nine human cell types.* Nature, 2011. **473**(7345): p. 43-9.
160. Mikkelsen, T.S., et al., *Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.* Nature, 2007. **448**(7153): p. 553-60.
161. Hawkins, R.D., et al., *Distinct epigenomic landscapes of pluripotent and lineage-committed human cells.* Cell Stem Cell, 2010. **6**(5): p. 479-91.
162. Lima-Junior, D.S., et al., *Endogenous retroviruses promote homeostatic and inflammatory responses to the microbiota.* Cell, 2021. **184**(14): p. 3794-3811 e19.
163. Jönsson, M.E., et al., *Activation of neuronal genes via LINE-1 elements upon global DNA demethylation in human neural progenitors.* Nature Communications, 2019. **10**(1).

164. Petri, R., et al., *LINE-2 transposable elements are a source of functional human microRNAs and target sites.* PLoS Genet, 2019. **15**(3): p. e1008036.
165. Shankar, R., et al., *Evolution and distribution of RNA polymerase II regulatory sites from RNA polymerase III dependant mobile Alu elements.* BMC Evol Biol, 2004. **4**: p. 37.
166. Bischler, T. and K. Förstner, *PEAKachu.* https://github.com/tbischler/PEAKachu, 2021.
167. Stevanovic, M., et al., *SOX Transcription Factors as Important Regulators of Neuronal and Glial Differentiation During Nervous System Development and Adult Neurogenesis.* Front Mol Neurosci, 2021. **14**: p. 654031.
168. Lai, T., et al., *SOX5 controls the sequential generation of distinct corticofugal neuron subtypes.* Neuron, 2008. **57**(2): p. 232-47.
169. Zawerton, A., et al., *Widening of the genetic and clinical spectrum of Lamb-Shaffer syndrome, a neurodevelopmental disorder due to SOX5 haploinsufficiency.* Genet Med, 2020. **22**(3): p. 524-537.
170. Parikshak, N.N., et al., *Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism.* Nature, 2016. **540**(7633): p. 423-427.
171. Holmes, S.E., et al., *A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion.* Nat Genet, 1994. **7**(2): p. 143-8.
172. Canzio, D., et al., *Antisense lncRNA Transcription Mediates DNA Demethylation to Drive Stochastic Protocadherin alpha Promoter Choice.* Cell, 2019. **177**(3): p. 639-653 e15.
173. Lefebvre, J.L., J.R. Sanes, and J.N. Kay, *Development of dendritic form and function.* Annu Rev Cell Dev Biol, 2015. **31**: p. 741-77.
174. Barbee, S.D., et al., *Skint-1 is a highly specific, unique selecting component for epidermal T cells.* Proc Natl Acad Sci U S A, 2011. **108**(8): p. 3330-5.
175. Narita, T., et al., *Mice lacking all of the Skint family genes.* Int Immunol, 2018. **30**(7): p. 301-309.
176. Xiang, J., M. Qiu, and H. Zhang, *Role of Dendritic Epidermal T Cells in Cutaneous Carcinoma.* Front Immunol, 2020. **11**: p. 1266.
177. Ho, M., et al., *Update of the keratin gene family: evolution, tissue-specific expression patterns, and relevance to clinical disorders.* Hum Genomics, 2022. **16**(1): p. 1.
178. Kumar, V., et al., *A keratin scaffold regulates epidermal barrier formation, mitochondrial lipid composition, and activity.* J Cell Biol, 2015. **211**(5): p. 1057-75.
179. Zimek, A. and K. Weber, *The organization of the keratin I and II gene clusters in placental mammals and marsupials show a striking similarity.* Eur J Cell Biol, 2006. **85**(2): p. 83-9.
180. Wu, Q., et al., *Comparative DNA sequence analysis of mouse and human protocadherin gene clusters.* Genome Res, 2001. **11**(3): p. 389-404.
181. Madabhushi, R., et al., *Activity-Induced DNA Breaks Govern the Expression of Neuronal Early-Response Genes.* Cell, 2015. **161**(7): p. 1592-605.
182. Chang, W., et al., *shiny: Web Application Framework for R.* 2022.
183. Fischer, M. and S. Hoffmann, *Synthesizing genome regulation data with vote-counting.* Trends Genet, 2022.

Abbreviation

cGAS	cyclic GMP-AMP synthase
DNMT	deoxyribonucleic acid methyltransferase
DNA	deoxyribonucleic acid
DSB	double strand breaks
e.g.	exempli gratia (for example)
EM	expectation maximization
EN	endonuclease
ERV	endogenous retrovirus
HERV	human endogenous retrovirus
i.e.	id est (that is)
IN	integrase
LINE	long interspersed nuclear element
LTR	long terminal repeat
mRNA	messenger ribonucleic acid
ncRNA	non-coding ribonucleic acid
ORF	open reading frame
PIWI	P-element induced wimpy testis
Pol	polymerase
RNA	ribonucleic acid
RNA-Seq	ribonucleic acid sequencing
RNP	ribonucleoprotein complex
RT	reverse transcriptase
SINE	short interspersed nuclear element
SRG	stress response genes
TE	transposable element
TF	transcription factor
TFBS	transcription factor binding site
TIR	terminal inverted repeats
TSS	transcription start site
TTS	transcription termination site

Appendix

7.1. Manuscript 1 – Form 2

Manuskript Nr. 1

Kurzreferenz Schwarz *et al.* (2022), Briefings in Bioinformatics

Beitrag des Doktoranden / der Doktorandin

Beitrag des Doktoranden zu Abbildungen, die experimentelle Daten wiedergeben:

Abbildung(en) # Alle	☒	100 % (die in dieser Abbildung wiedergegebenen Daten entstammen vollständig experimentellen Arbeiten, die der Kandidat/die Kandidatin durchgeführt hat)
	☐	0 % (die in dieser Abbildung wiedergegebenen Daten basieren ausschließlich auf Arbeiten anderer Koautoren)
	☐	Etwaiger Beitrag des Doktoranden / der Doktorandin zur Abbildung: _____%

7.2. Manuscript 2 – Form 2

Manuskript Nr. 2

Kurzreferenz Schwarz *et al.*, eingereicht

Beitrag des Doktoranden / der Doktorandin

Beitrag des Doktoranden zu Abbildungen, die experimentelle Daten wiedergeben:

Abbildung(en) # Alle	☒	100 % (die in dieser Abbildung wiedergegebenen Daten entstammen vollständig experimentellen Arbeiten, die der Kandidat/die Kandidatin durchgeführt hat)
	☐	0 % (die in dieser Abbildung wiedergegebenen Daten basieren ausschließlich auf Arbeiten anderer Koautoren)
	☐	Etwaiger Beitrag des Doktoranden / der Doktorandin zur Abbildung: _____%

7.3. Manuscript 3 – Form 2

Manuskript Nr. 3

Kurzreferenz Fischer *et al.* (2022), NAR Cancer

Beitrag des Doktoranden / der Doktorandin

Beitrag des Doktoranden zu Abbildungen, die experimentelle Daten wiedergeben:

Abbildung(en) # Alle	☐	100 % (die in dieser Abbildung wiedergegebenen Daten entstammen vollständig experimentellen Arbeiten, die der Kandidat/die Kandidatin durchgeführt hat)
	☒	0 % (die in dieser Abbildung wiedergegebenen Daten basieren ausschließlich auf Arbeiten anderer Koautoren)
	☐	Etwaiger Beitrag des Doktoranden / der Doktorandin zur Abbildung: _____%

Danksagung

Zu Beginn möchte ich mich besonders bei meinem Betreuer Steve Hoffmann bedanken, der es mir ermöglichte, an diesem spannenden Thema zu arbeiten. Mit anhaltend konstruktivem Einfluss schaffte er, das Projekt in die Form zu bringen, wie sie hier niedergeschrieben steht. Vielen Dank für dein geduldiges und ruhiges Gemüt und deine Fähigkeit, mich immer wieder aus Tälern voller Zweifel herauszuziehen, sowie meine Motivation zu kitzeln.

Ein großer Dank geht auch an den besten Bürokollegen Philipp Koch, durch dessen Betreuung meiner Masterarbeit der Weg für diese Dissertation geebnet wurde. Vielen Dank für dein stets offenes Ohr und deine Hilfe bei allen Dingen, die es bedurfte, um diese Arbeit abzuschließen. Besonders hervorheben möchte ich seinen grünen Daumen, welcher es mir ermöglichte, täglich in einer grünen Oase zu sitzen. Weiterhin möchte ich mich bei Martin Fischer, Martin Bens, Konstantin Riege und Jeanne Wilbrandt, bedanken die mich stets mit produktiven Gesprächen auf meinem Weg begleitet haben. Darüber hinaus danke ich der gesamten Arbeitsgruppe, sowie den Serviceeinrichtungen Life Science Computing und Next Generation Sequencing für ihre stetige und bedingungslose Bereitschaft, mich auf meinem Weg zu unterstützen. Außerdem möchte ich mich bei meinem Promotionskomitee, bestehend aus Steve Hoffmann, Peter Stadler und Christoph Englert, herzlich für Ihre investierte Zeit und Mühe bedanken.

Ein unbeschreiblicher Dank geht an Jennifer Müller, die es mit ihrer besonnen Art immer wieder schaffte, mein Selbstbewusstsein zu stärken und aufkommende Zweifel im Keim zu ersticken. Ich danke dir unglaublich für dein geduldiges Zuhören und unermüdliches Unterstützen in allen Belangen meines Lebens.

Ein ganz besonderer Dank geht an Andreas Jörk, Rico Schubert, Robert Schenderlein, Eric Mittenzwei, Willi Meierhof, Kristin Gross, Carolin Voigt, Peter Voigt, Philipp Ziegenbein, Tobias Klement, Martin Grunert, Tina Serfling, Roman Lange – ein Freundeskreis, der mich seit nun mehr 20 Jahren begleitet und mich dazu inspirierte, den akademischen Weg zu gehen. Trotz vermehrter Absagen in letzter Zeit bin ich froh und besonders stolz, euch weiterhin als gute Freunde zu haben.

Zu guter Letzt möchte ich mich bei meinen Eltern, sowie meinem Bruder bedanken, die mich stets unterstützt und niemals meinen Weg angezweifelt haben.

Statement of authorship

I confirm that I am familiar with the relevant course of examination for doctoral candidates in the Faculty of Biological Sciences at Friedrich-Schiller-University in Jena. I have composed and written the dissertation by myself and I have acknowledged all additional assistance, personal communication, and sources within the work.

I have not enlisted the assistance of a doctoral consultant and no third parties have received either direct or indirect monetary benefits from me for work connected to the submitted dissertation. I have not submitted the dissertation in an exact or modified version for a state or other scientific examination.

Ich erkläre, dass mir die geltende Promotionsordnung der Fakultät für Biowissenschaften der Friedrich-Schiller-Universität in Jena bekannt ist. Ich versichere, dass ich die vorliegende Dissertation selbst angefertigt, keine Textabschnitte eines Dritten oder eigener Prüfungsarbeiten ohne Kennzeichnung übernommen und alle verwendeten Hilfsmittel, persönlichen Mitteilungen und Quellen in meiner Arbeit angegeben habe.

Ich bestätigte, dass ich die Hilfe eines Promotionsberaters nicht in Anspruch genommen habe und dass Dritte weder unmittelbar noch mittelbar geldwerte Leistungen von mir für Arbeiten erhalten haben, die in Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen. Ich versichere, dass ich die Dissertation weder in gleicher noch in ähnlicher Form zuvor als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht habe.

X

Robert Schwarz