Characterization of *Gadd45*-Deficient Embryonic Stem Cells and Development of Methods to Explore Repetitive DNA Elements

Dissertation

Zur Erlangung des Grades

Doktor der Naturwissenschaften

Am Fachbereich Biologie

der Johannes Gutenberg-Universität Mainz

Marcel Mišák

geboren am 28.10.1992 in Mannheim

Mainz, 2025

Dekan:

1. Berichterstatter:

2. Berichterstatter:

Tag der mündlichen Prüfung: 30.01.2025

1 Table of Contents

2 Summary

This thesis presents two investigations employing computational analysis, an approach that has become indispensable to study organisms' genomes and gene regulation. First, I characterized a mutant cell line using different types of sequencing data. Second, I developed a computational sequencing analysis tool for detection of enriched tandem repeats in DNA sequencing datasets. Moreover, I created and applied a workflow for the design of retrotransposon subfamily-discriminating quantitative PCR (qPCR) primers.

(I) GADD45 family proteins have previously been implicated in epigenetic gene regulation via a mechanism involving R-loops, three-stranded nucleic acid structures, composed of a DNA:RNA hybrid and a displaced single DNA strand. Moreover, these proteins have been shown to be involved in the demethylation of enhancers in embryonic stem cells (mESCs). To shed more light on the exact mechanism whereby GADD45 proteins regulate gene expression specifically in mESCs, I analyzed DNA sequencing datasets for triple-knockout (TKO) mESC mutant cell lines that lack the three GADD45 proteins. The resulting data suggests a possible role of GADD45 family proteins in modulating R-loops at enhancer regions, thereby regulating expression of genes involved in heart-development.

(II) I developed counTR, a computational tool designed to detect tandem repeat enrichment in short-read sequencing data. By applying counTR, I demonstrated its utility not only in uncovering protein binding sites on these repeats, but also in comparing length of telomeres, regions located at the ends of chromosomes composed of a special class of tandem repeats implicated in cellular aging. Moreover, I created and applied a workflow to design qPCR primers with the ability to discriminate between different subfamilies of Long Interspersed Nuclear Element-1 (LINE1) retrotransposons. Only one human LINE1 subfamily, L1HS, remains capable of *de novo* genome integration with potentially detrimental physiological consequences. These experimentally validated qPCR primers can aid in understanding how this and other LINE1 subfamilies are regulated and amplify.

3 Zusammenfassung

Diese Dissertation beschreibt zwei Untersuchungen, die computergestützte Analyse einsetzen – eine Methode, die mittlerweile unverzichtbar für das Studium von Genomen und Genregulation ist. Erstens charakterisiere ich eine Zelllinien-Mutante mithilfe von Sequenzierungsdaten. Zweitens habe ich ein Programm zur Erkennung angereicherter Tandem-Repeats in Sequenzierungsdaten und einen Arbeitsablauf für das Design von qPCR-Primern, die zwischen Subfamilien von Retrotransposons unterscheiden können, entwickelt und letzteren angewandt.

(I) Die Proteine der GADD45-Familie sind unter anderem für ihre Rolle in der epigenetischen Genregulation bekannt. In diesem Mechanismus spielen R-Loops, dreisträngige Nukleinsäurestrukturen bestehend aus einem DNA:RNA-Hybrid und einem DNA-Einzelstrang, eine wesentliche Rolle. Zudem wurden sie mit der Demethylierung von Enhancern in murinen embryonalen Stammzellen (mESCs) in Verbindung gebracht. Um die Funktion von GADD45 in der Regulierung von Genexpression speziell mESCs zu verstehen, habe ich Sequenzierungsdatensätze für eine mESC-Mutantenzelllinie, der die drei GADD45 Proteine fehlen, analysiert. Die Ergebnisse deuten darauf hin, dass die GADD45-Proteine die Formierung von R-Loops an Enhancer-Regionen modulieren und so die Expression herzentwicklungsspezifischer Gene regulieren könnten.

(II) Ich habe das Programm *counTR* für die Erkennung von angereicherten Tandem-Repeats in Short-Read-Sequenzierungsdaten entwickelt und seine Anwendung in der Identifikation von Proteinbindungen auf diesen Repeats sowie zum Vergleich der Längen von Telomeren demonstriert. Letztere sind die aus einer speziellen Klasse von Tandem-Repeats bestehenden Chromosom-Enden, die eine wichtige Rolle bei der Zellalterung spielen. Darüber hinaus habe ich einen Arbeitsablauf für das Design qPCR-Primern entwickelt, die in der Lage sind, zwischen verschiedenen Unterfamilien von Long Interspersed Nuclear Element-1 (LINE1)-Retrotransposons zu unterscheiden, und diesen angewant. Nur eine LINE1-Unterfamilie, L1HS, ist im Menschen zur Transposition in andere DNA-Loci fähig – mit potenziell schädlichen physiologischen Folgen. Diese speziell designten und experimentell validierten qPCR-Primer können dabei helfen zu verstehen, wie L1HS und andere LINE1-Unterfamilien reguliert werden und sich in Genomen amplifizieren.

4 Part I: Characterization of *Gadd45* TKO mESCs

4.1 Introduction

4.1.1 Eukaryotic regulation of transcription via regulatory DNA elements

The rate of gene transcription by RNA polymerases defines the amount of synthesized RNA (Nudler, 2009). This transcription process can, in eukaryotes, be regulated by promoters, enhancers, silencers and insulators (Smith et al., 2012) (Figure 4.1). Among these regulatory DNA elements, promoters are located at the 5′ termini of genes, near the transcription start site (TSS), and they contain specific binding sites for transcription factors (Haberle & Stark, 2018). These proteins bind directly to DNA, thereby facilitating (or sometimes hindering) the recruitment and activity of RNA polymerases (Latchman, 1996). Enhancers and silencers are regulatory DNA regions that, by direct interaction with promoters, can augment or restrict transcription, respectively (Kolovos et al., 2012). Finally, insulators prevent promiscuous gene regulation by enhancers or silencers (Raab & Kamakaka, 2010). These regulatory sequences on DNA are static in their location and can thus, by themselves, not regulate genes dynamically.

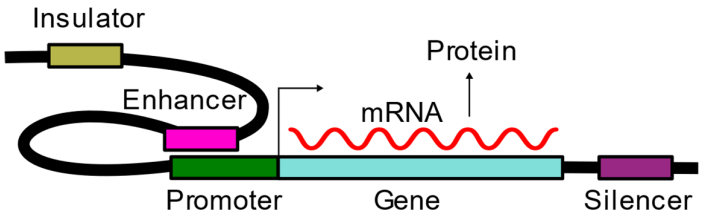


Figure 4.1: Schematic illustrating gene regulatory elements. Thick black line illustrates double-stranded DNA. Colored boxes represent gene regulatory elements. TSS represented by a right-pointing arrow. Figure inspired by a figure in Smith et al. (2012).

4.1.2 Epigenetic regulation of gene transcription in eukaryotes

The term *epigenetics* was coined in the 1940s by Conrad Waddington who defined it as “the branch of biology which studies the causal interactions between genes

and their products which bring the phenotype into being" (Waddington, 1968). The term has changed since then and is now generally accepted as "the study of changes in gene function that are mitotically and/or meiotically heritable and that do not entail a change in DNA sequence" (Wu & Morris, 2001). Such changes are called *epigenetic modifications* and encompass modified DNA bases and posttranslational modifications of histone tails, both of which can regulate or fine-tune gene expression (Kumar et al., 2018). Unlike regulatory sequences on DNA, epigenetic modifications can be dynamically set and removed. To regulate gene expression, epigenetic modifications require readers, writers and erasers, they are crucial in the shaping of functional differences between cell types and can change in response to environmental perturbations (Rahman & McGowan, 2022).

4.1.3 Chromatin and epigenetic regulation by histone modifications

In eukaryotic cells, chromosomal DNA is present in a highly organized structure made up of DNA, RNA, and proteins called *chromatin* (Gross et al., 2015). Here, ~147 base pairs of DNA are wrapped around histone protein octamers that are canonically formed by two copies each of histones H2A, H2B, H3, and H4, forming nucleosomes (Talbert & Henikoff, 2021) (Figure 4.2). The linker histone H1 binds to the nucleosomal DNA entry/exit site to stabilize the structure (Parseghian, 2015). Histone amino-terminal tails can be post-translationally modified by acetylation (ac), methylation (me), ubiquitination (ub) and phosphorylation, among others (Nathan et al., 2003; Shiio & Eisenman, 2003). Moreover, the canonical histone proteins can be replaced by histone variants, paralogous histone proteins which can also impact chromatin structure (Talbert & Henikoff, 2021). Chromatin can be divided into three different categories, based on its higher-order packaging, histone modifications and histone variants. The majority of chromatin is present as *heterochromatin* in a highly condensed, transcriptionally repressive state and characterized by histone modifications H3K9me3, H3K27me3 and H3K119ub (Morrison & Thakur, 2021). The second state, *euchromatin*, has a more open, transcriptionally permissive structure and is characterized by histone modifications H3K4me3, H3K27ac and H3K36me (Morrison & Thakur, 2021). The third state is called *centromeric chromatin* and mainly functions in cell division (Müller & Almouzni, 2017).

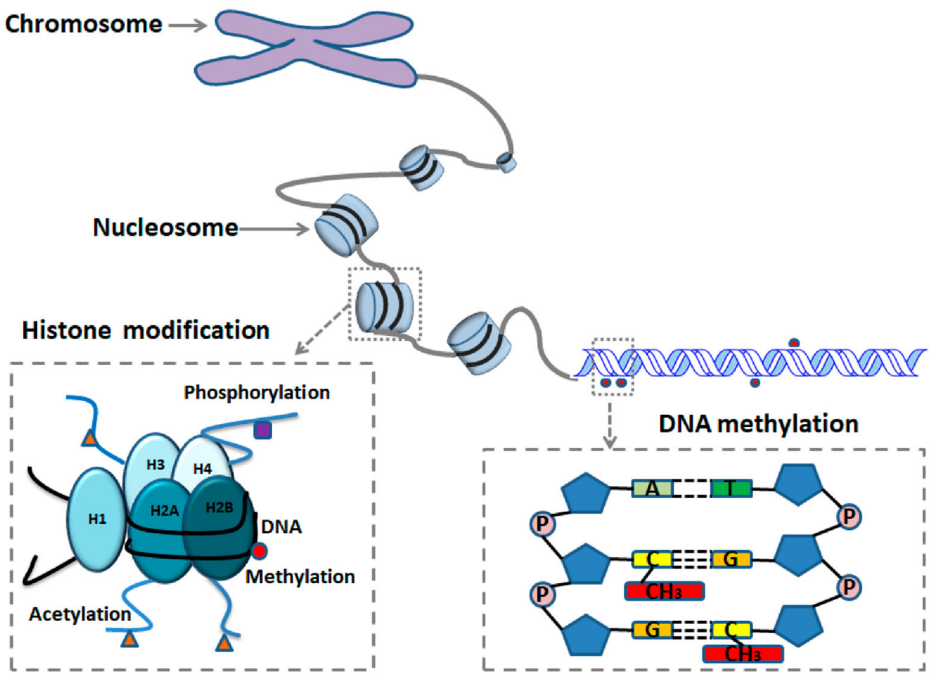


Figure 4.2: Major epigenetic modifications. Within cells, DNA is organized into chromosomes, composed of chromatin. Here, DNA is wrapped around histone proteins, forming nucleosomes. The histone proteins within those nucleosomes can be modified at their amino-terminal tails. Moreover, the cytosine DNA base can be modified by addition of a methyl group (CH_3), forming 5fC. Such DNA methylation commonly occurs symmetrically on opposing CpG dinucleotides (A: adenine, C: cytosine, G: guanine, T: thymine). Figure adapted from Ren et al. (2024) licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

While gene regulatory elements are static, they can be dynamically modified through histone modifications. For instance, H3K4me1 is generally present in enhancers (Kubo et al., 2024), whereas the presence of H3K27ac in enhancers distinguishes their active from their poised state by exclusive presence in the former (Creyghton et al., 2010). Meanwhile, H3K27me3 is considered to be a histone modification classically found in promoters to repress transcription (Jiang et al., 2024), while acetylation marks generally stimulate transcription (Vaid et al., 2020). Histone acetylation marks are written by histone acetyltransferases and erased by histone deacetylases, while histone methylation is catalyzed by histone methyltransferases and removed by histone demethylases (Yang et al., 2022). These histone modifications can then be recognized by reader proteins with downstream functions (Musselman et al., 2012) or directly influence the transcription rate of RNA polymerases (Tanny, 2014).

4.1.4 Epigenetic regulation by DNA methylation

DNA methylation refers to the addition of a methyl group to the DNA base cytosine, converting it to 5-methylcytosine (5mC) (Moore et al., 2013). Cytosine methylation is catalyzed by DNA methyltransferases (DNMTs) and occurs, with rare exceptions, within CpG dinucleotides (Davletgildeeva & Kuznetsov, 2024). CpG dinucleotides are palindromic, i.e. a 5'-CpG-3' on one strand base-pairs with a 5'-CpG-3' on the opposing strand. DNA methylation at CpG dinucleotides occurs typically on the cytosines on both DNA strands (Hua et al., 2024). After DNA replication and cell division, each daughter cell inherits one DNA strand already bearing methylated cytosines and a newly synthesized strand lacking the modification. The maintenance DNMT, DNMT1, can recognize such hemi-methylated sites and restore symmetrical methylation patterns by adding a methyl group to the cytosine on the unmethylated strand (Robert et al., 2003). Many eukaryotic gene promoters contain CpG islands (CGIs), short genomic regions with a high frequency of CpG dinucleotides (Moore et al., 2013). CGIs are commonly unmethylated, allowing for active gene transcription, but can become methylated, leading to gene silencing (Moore et al., 2013). Gene silencing by DNA methylation in promoters is achieved by inhibition of transcription factor binding or recruitment of repressor proteins (Jurkowska & Jeltsch, 2010). However, DNA methylation can also be found within the bodies of highly expressed genes, where it is thought to prevent spurious transcription initiation (Neri et al., 2017). DNA methylation also plays important roles in genomic imprinting, X-chromosome inactivation and the silencing of transposon expression (Sergeeva et al., 2023).

Cells can actively demethylate DNA by catalytically oxidizing 5mC to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC) via TET enzymes (Onodera et al., 2021) (Figure 4.3). The oxidized cytosine derivates 5fC and 5caC can then be replaced by unmodified cytosines by the base excision repair (BER) mechanism mediated by the protein TDG (Slyvka et al., 2017). While 5mC is commonly preserved after cell divisions through maintenance methylation by DNMT1, 5hmC, 5fC and 5caC are not (Slyvka et al., 2017) (Figure 4.3). Thus, these marks become gradually diluted out over successive cell cycles. Moreover, inhibition of DNMT1 can prevent maintenance methylation, leading to

passive DNA demethylation (He et al., 2017). Conversely, TET1 was also shown to mediate recruitment of Polycomb-related factors in order to mediate gene silencing (van der Veer et al., 2023). In a seemingly contradicting fashion, TET proteins can thus have a gene activating or gene repressive role.

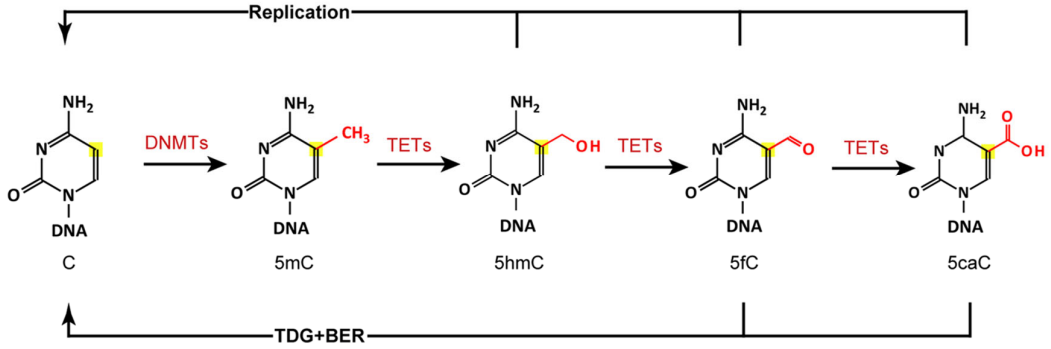


Figure 4.3: TET-mediated DNA demethylation. Cytosine is initially converted to 5mC by DNMTs. Through TET1-mediated oxidation, 5mC is sequentially transformed into 5hmC, 5fC, and 5caC. The modified bases 5hmC, 5fC, and 5caC can be gradually diminished through successive DNA replication rounds, while 5fC and 5caC can be actively restored to unmodified cytosine via TDG-mediated BER. The C5 position of cytosine is highlighted in yellow. Figure from Zhang et al. (2023) licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

4.1.5 R-Loops: Genome instability vs. epigenetic control

R-loops are three-stranded nucleic acid structures comprised of a DNA:RNA hybrid as well as a displaced single-stranded DNA (ssDNA) (Kim & Wang, 2021) (Figure 4.4). These structures can form cotranscriptionally, in *cis*, when a newly synthesized RNA re-anneals to its DNA template strand (Hegazy et al., 2020). *Cis* R-loops are frequently associated with RNA polymerase pausing and can also result from collisions between the transcription and replication machinery (Niehrs & Luke, 2020), particularly in large genes where transcription extends beyond a single cell cycle, making such collisions inevitable (Hegazy et al., 2020). As a result of hybridization between an RNA strand and a distant homologous DNA strand, R-loops can also form in *trans* (Niehrs & Luke, 2020). A prominent example of R-loop formation in *trans* occurs in the bacterial CRISPR mechanism, in which guide RNAs

are used to target specific proteins to homologous viral sequences, resulting in R-loops that facilitate cleavage and destruction of viral genomes (Pacesa et al., 2022).

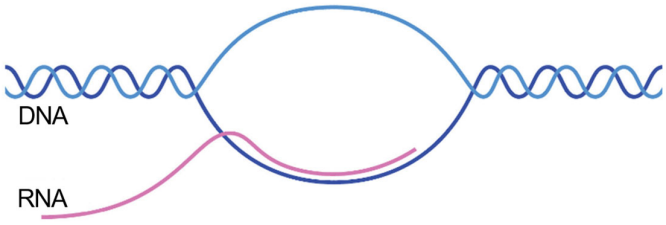


Figure 4.4: Schematic of an R-loop. R-loops are three-stranded nucleic acid structures composed of a DNA:RNA hybrid and a displaced ssDNA. Figure adapted from Hegazy et al. (2020).

Due to vulnerability of the exposed ssDNA and due to collisions between the DNA replication machinery and R-loops, R-loops are a source of genome instability (Stratigi et al., 2024). Hence, cells have developed strategies to promote R-loop resolution and regulate R-loop formation to ensure preservation of their genome integrity. R-loop removal is facilitated by helicases, topoisomerases, and ribonucleases (RNases) (Xu et al., 2024). Helicases like SETX and AQR unwind the DNA:RNA hybrids, thereby counteracting R-loop persistence (Yang et al., 2023). Topoisomerases TOP1, TOP2 and TOP3B alleviate torsional stress that can lead to R-loop formation, especially in highly transcribed regions, reducing the likelihood of accidental R-loop accumulation (Saha & Pommier, 2023). RNases H1 and H2 specifically degrade the RNA strand in DNA:RNA hybrids, thus resolving R-loops and restoring double-stranded DNA configuration (Zhao et al., 2018). These mechanisms minimize the potentially harmful impact of R-loops.

Originally exclusively considered to be harmful byproducts of transcription, R-loops have since been shown to function as epigenetic regulators of transcription. R-loops are capable of modulating gene expression through chromatin accessibility and recruitment of regulatory proteins (Fazzio, 2016). Via interaction with chromatin-modifiers, the presence of R-loops can result in histone modification or changes in DNA methylation. As a result, transcriptional activity may be induced or inhibited depending on chromatin context and genomic location. Additionally, R-loop presence can work as a barrier for RNA polymerase, inducing its pausing and thereby modulating gene expression (Zardoni et al., 2021). Thus, R-loops can be

termed *double-edged swords*. While posing risks to genome stability, they also play a role as epigenetic regulators that impact transcription dynamics.

4.1.6 The GADD45 protein family

The growth arrest and DNA damage-inducible 45 (GADD45) protein family is a group of small (17-18kDa), strongly acidic, evolutionarily conserved proteins. GADD45 family proteins are categorized into the L7Ae/L30e/S12e RNA-binding ribosomal protein superfamily (Huang et al., 2024; Sytnikova et al., 2011). In both humans and mice, this family consists of three paralogs: GADD45a, GADD45b, and GADD45g. By interacting with one another or by binding other cellular proteins, GADD45 proteins have the ability to form both homodimers and heterodimers (Kovalsky et al., 2001). Lacking intrinsic enzymatic activity, these proteins carry out their functions via recruitment of other cellular proteins and are involved in diverse processes, such as cell growth, DNA repair, apoptosis, and function as suppressors of tumorigenesis and autoimmune response (Sytnikova et al., 2011).

GADD45 family proteins play a key role in a number of cell signaling pathways, including pathways associated with stress response, cell cycle regulation through cyclin-dependent kinases (CDKs), and DNA repair. In response to environmental stresses, all three GADD45 family members are able to activate mitogen-activated protein kinase kinase kinase 4 (MEKK4) to induce apoptosis via the p38 and JNK pathways (Takekawa & Saito, 1998; Tamura et al., 2012). Moreover, GADD45a has been shown to directly interact with and activate p38 MAP kinase (Bulavin et al., 2003). Opposing the pro-apoptotic role of the GADD45 family proteins, GADD45b has also been shown to prevent JNK activation by inactivation of MAP kinase kinase 7 (MKK7) through direct binding and inhibition of its catalytic function, resulting in an antiapoptotic effect (Gupta et al., 2006).

4.1.7 GADD45a: An R-loop reader initiating TET1-mediated DNA demethylation

This lab previously described a mechanism in which GADD45a acts as an R-loop reader that binds directly to an R-loop formed in the promoter region of the *TCF21*

gene in the human HEK293T cell line (Arab et al., 2019) (Figure 4.5). This R-loop is formed by the long noncoding RNA (lncRNA) TARID transcribed in antisense direction to the *TCF21* gene. GADD45a recognition of this R-loop leads to the recruitment of TET1, triggering oxidative demethylation of the nearby DNA and, consequently, activation of TCF21 expression (Arab et al., 2019). This discovery highlights a new layer of gene regulation in which R-loops, long considered transcriptional byproducts, serve as regulatory elements capable of guiding epigenetic modifiers to specific genomic regions. Whether this mechanism operates more generally beyond a single cell type or promoter currently remains unknown.

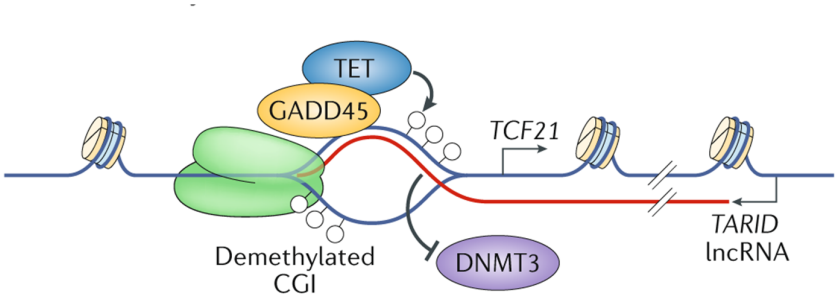


Figure 4.5: Illustration of GADD45's gene regulatory role in the *TCF21* locus. GADD45 binds to the R-loop formed by the antisense lncRNA TARID in the *TCF21* promoter and recruits TET, which induces oxidative demethylation. As an effect of R-loop formation in CGIs, de novo DNA methylation by DNMT3 may be inhibited. Figure adapted from Niehrs and Luke (2020).

4.1.8 GADD45 proteins are involved in the demethylation of enhancers in different embryonic mouse cell lines

In Schäfer et al. (2018), this lab showed that in mouse embryonic fibroblasts (MEFs), GADD45a binds to enhancers that are functionally dependent on the binding of the proteins C/EBPb and C/EBPd (Figure 4.6). Cooperatively with ING1, which binds to promoters occupied by the transcription factor E2F, GADD45a promotes the demethylation of such C/EBPb/d-binding enhancers via long-range chromatin loops. Mice deficient for both GADD45a and ING1 showed premature aging symptoms as a manifestation of an impaired GADD45a-ING1-C/EBP interaction.

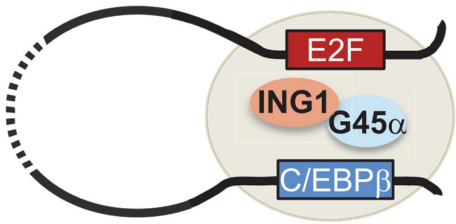


Figure 4.6: Model illustrating chromatin looping between distant GADD45a and ING1-bound regions in MEFs. GADD45a occupies C/EBP-dependent enhancers and interacts with ING1-bound promoters. Figure from Schäfer et al. (2018) licensed under CC BY-NC 4.0 (https://creativecommons.org/licenses/by-nc/4.0/).

In 2019, this lab explored the function of GADD45 proteins in mESCs – pluripotent cells that mirror the inner cell mass at implantation in early mouse development (Schüle et al., 2019). By generating GADD45 TKO mESC clones using CRISPR/Cas9 and comparing them to wild-type (WT) controls, the authors found that the absence of GADD45 genes led to aberrant hypermethylation at enhancers and other loci normally undergoing TET-mediated DNA demethylation (Schüle et al., 2019). Additionally, GADD45 TKO cells displayed a reduced population of "2C-like" cells, which resemble the two-cell stage embryo, a pivotal developmental phase marked by zygotic genome activation and widespread chromatin reorganization (Schüle et al., 2019). Mice deficient for *Gadd45a* and *Gadd45b* were shown to be sublethal, with surviving mice displaying phenotypic abnormalities characteristic of neural tube closure defects, reinforcing the critical role of GADD45 family members during embryogenesis (Schüle et al., 2019). Schüle et al. (2019) concluded that GADD45 proteins facilitate locus-specific DNA demethylation at TET-target enhancers and are required for normal cycling into the 2C-like state in mESCs, highlighting their importance in regulating early murine development. However, the mechanism remains unclear, and a direct link between the observed effects on enhancers and GADD45's R-loop binding activity (Arab et al., 2019) was not established.

4.1.9 Mapping chromatin-bound proteins genome-wide using sequencing methods

A comprehensive understanding of gene regulation by proteins or epigenetic marks requires genome-wide knowledge of their genomic target sites. Chromatin sites enriched for bound proteins or modified histones can be mapped on a genome-wide scale using methods like Chromatin Immunoprecipitation sequencing (ChIP-seq) or Cleavage Under Targets and Tagmentation sequencing. Such techniques are typically based on bulk short-read sequencing (Kaya-Okur et al., 2019; Nakato & Sakata, 2021), though single-cell adaptations are also available (Bartosovic et al., 2021; Grosselin et al., 2019). ChIP-seq involves first crosslinking the target of interest to chromatin, then fragmenting the chromatin, immunoprecipitating the target-bound fragments using an antibody, and finally purifying and sequencing the retrieved DNA fragments (Barski et al., 2007; Johnson et al., 2007). The more recent CUT&Tag method utilizes a primary antibody specific to the protein or histone modification of interest, followed by incubation with a secondary antibody to amplify the amount of antibodies binding at those sites (Kaya-Okur et al., 2019). A fusion construct of protein A and the Tn5 transposase then recognizes and binds the antibodies, positioning Tn5 to cut the DNA near the target and add sequencing adapters. After DNA purification, the resulting fragments are sequenced. (Kaya-Okur et al., 2019). Both methods result in enriched reads originating from binding-sites of the respective target and can be computationally analyzed by read mapping with subsequent peak calling of enriched regions (Cheng et al., 2024; Eder & Grebien, 2022). Moreover, comparing normalized read counts at identified peaks across different conditions allows for the detection of differential binding dynamics (Cheng et al., 2024; Eder & Grebien, 2022).

4.1.10 Genome-wide R-loop mapping by sequencing

Mapping genomic R-loop sites can be essential for understanding their role in gene regulation and genome stability. The laboratory of Frédéric Chédin developed the first R-loop mapping method called DNA-RNA immunoprecipitation sequencing (DRIP-seq) (Ginno et al., 2012). DRIP-seq uses the DNA:RNA hybrid-binding

antibody S9.6 to specifically enrich for chromatin regions containing these hybrid structures (Sanz & Chédin, 2019). Several additional R-loop mapping methods have been developed subsequently. R-loopBase (https://rloopbase.nju.edu.cn) catalogs eleven distinct methods, each relying on either the S9.6 antibody or mutated RNase H enzymes (or their hybrid-binding domains) that retain R-loop-binding capacity without degrading the RNA (Lin et al., 2022). These techniques differ in both methodology and the information they capture; some are limited to locating R-loops within the genome, while other (strand-specific) methods can additionally identify which DNA strand forms the DNA:RNA hybrid and which strand is left as displaced ssDNA (Chédin et al., 2021). Similar to ChIP-seq and CUT&Tag, R-loop mapping data is commonly analyzed using a workflow of read alignment, peak calling, and differential enrichment analysis. Unlike ChIP-seq and CUT&Tag protocols, which typically do not provide strand information, strand-specific R-loop techniques require reads to be separated by DNA strand before peak calling. This separation is essential to identify which DNA strand generates the enrichment. Using this information, it can be inferred which strand contains the DNA:RNA hybrid and which one is left as displaced ssDNA. While the approaches described in this section rely on bulk short-read sequencing, it should be noted that variants adapted for single-cell analysis and long-read sequencing have also emerged (Lu et al., 2023; Malig & Chédin, 2020).

4.1.11 Mapping of DNA methylation

Understanding which genomic loci are methylated and how their methylation status changes under different conditions is crucial for uncovering the role of DNA methylation in gene regulation. DNA methylation is commonly studied genome-wide with base-resolution by bisulfite sequencing techniques. In whole genome bisulfite sequencing (WGBS), bisulfite-treatment converts unmodified cytosines and also 5fC and 5caC to uracil while 5mC and 5hmC stays unaltered (Booth et al., 2013). When bisulfite-treated DNA is sequenced, unmodified cytosines, 5fC and 5caC are basecalled as thymines, while 5mC (but also 5hmC) is read as cytosine and can be detected as a methylated base by methylation caller programs after read mapping (Booth et al., 2013; Wreczycka et al., 2017). However, because WGBS typically

requires extensive sequencing coverage, its application can be relatively costly. Reduced representation bisulfite sequencing (RRBS) is another bisulfite sequencing-based approach that focuses on CpG-rich regions of the genome. In RRBS, genomic DNA is digested with the restriction enzyme MspI, which specifically recognizes and cuts at CCGG sites (Heyward & Sweatt, 2015). This selective digestion enriches regions dense in CpGs, including many CGIs. By limiting sequencing to these enriched fragments, RRBS protocols reduce the required sequencing coverage, making it a more cost-effective alternative to WGBS while still offering single-base resolution.

5mC DNA immunoprecipitation sequencing (DIP-seq), also called methylated DNA immunoprecipitation sequencing (MeDIP-seq), is an alternative approach for the genome-wide mapping of DNA methylation. 5mC DIP-seq enriches for methylated DNA fragments by using an antibody specific to 5mC, followed by high-throughput sequencing (Maamar et al., 2021). While 5mC DIP-seq does not provide single-base resolution, it is more cost-effective than WGBS, as it generally requires lower sequencing coverage. Moreover, unlike RRBS, 5mC DIP-seq is not confined to CpG-rich regions, enabling a more comprehensive assessment of the methylome. 5mC DIP-seq can be analyzed in a similar manner to ChIP-seq or CUT&Tag by read alignment, peak calling, and differential enrichment analysis (Li et al., 2010).

DNA methylation arrays offer another economical alternative to sequencing approaches by targeting only predefined CpGs. In these arrays, bisulfite-converted sample DNA fragments are hybridized to specific DNA probes for methylated and unmethylated fragments (Jiménez et al., 2021). Signal intensities from methylated and unmethylated probes can be compared to determine methylation levels at specific CpG sites (Jiménez et al., 2021). Specialized software packages are available for the analysis of such methylation arrays (Sahoo & Sundararajan, 2024).

4.1.12 Studying transcriptomes using sequencing methods

To understand how differential binding of chromatin- or DNA interactors and epigenetic marks affects the transcription of genes, transcriptomes of samples in different conditions, e.g. knockout and WT, can be compared. For this purpose, bulk

RNA sequencing (RNA-seq) is commonly employed. Typically, RNA-seq is performed by isolating RNA from a population of cells, reverse-transcription of the RNA into complementary DNA (cDNA), and sequencing (Hrdlickova et al., 2017). Although both short-read and long-read RNA-seq protocols are established, short-read RNA-seq remains widely used. After sequencing, reads are usually aligned to a reference genome, and normalized read counts are compared between conditions to identify differentially expressed genes (DEGs) (Deshpande et al., 2023).

In addition to transcriptome sequencing in bulk, RNA-seq can also be performed at the single cell level to profile gene expression within individual cells (Haque et al., 2017). Moreover, multiple nascent RNA sequencing methods are available, e.g. Global Run-On sequencing (GRO-seq) and Precision Run-On Sequencing (PRO-seq) both of which specifically measure transcripts that are actively being synthesized (Wang et al., 2018). When integrated with datasets profiling chromatin interactors or DNA modifications, transcriptome sequencing methods can reveal correlations with changes in gene expression and eventually aid in hypothesis generation for the underlying mechanisms driving those changes.

4.1.13 Thesis aims

GADD45 family proteins have been implicated in epigenetic gene regulation by (1) initiating TET1-mediated DNA demethylation after binding to R-loops (Arab et al., 2019) and (2) the DNA demethylation at enhancers in MEF cells and mESCs (Schäfer et al., 2018; Schüle et al., 2019). The aim of this first part of my thesis was to further investigate the mechanism by which GADD45 proteins regulate gene expression in mESCs. Specifically, I assessed whether the data would support a model in which GADD45 proteins bind to R-loops and facilitate TET1-mediated DNA demethylation and whether they influence epigenetic marks at enhancers. To achieve this, I characterized *Gadd45* TKO mESC lines by analyzing sequencing data generated by collaborators.

4.2 Results

4.2.1 Comparison of R-loop mapping methods

In this chapter, I analyze collaborator-generated R-loop mapping datasets to guide the selection of a suitable mapping method for comparing R-loop levels between WT and *Gadd45* TKO mESC lines in a future chapter. Numerous R-loop mapping sequencing strategies making use of a number of different molecular tools like the S9.6 antibody, catalytically dead RNase H1 (RNH) enzyme or RNH's DNA:RNA hybrid-binding protein domain (HBD) are available. While all R-loop mapping methods attempt to provide information on the genomic loci of R-loops, the methods can be separated into strand-specific and non-strand-specific R-loop mapping methods, depending on whether they provide information on the DNA strand containing the DNA:RNA hybrid as well as the looped-out ssDNA strand or not.

First, I compared non-strand-specific R-loop mapping methods. Khelifa Arab (this lab) established and applied the CUT&Tag method using the S9.6 antibody, HBD and enzymatically inactive E. coli RNH (ecRNH) in mESCs. I analyzed the newly generated CUT&Tag data and compared it to a publicly available DNA-RNA immunoprecipitation sequencing (DRIP-seq) dataset generated from mESCs (Sanz & Chédin, 2019) that I reanalyzed in a comparable manner. For each of the newly generated CUT&Tag datasets, peaks were called against an RNH-treated sample, whereas DRIP-seq peaks were called without control since no control sample was provided in the dataset. Visual inspection of two representative loci showed that the three CUT&Tag-based methods generate visually similar tracks, whereas the DRIP-seq signal differs considerably in shape (Figure 4.7). Peak intersection between the four methods indicated that ecRNH peaks are mostly a subset of S9.6 CUT&Tag or HBD CUT&Tag peaks, and DRIP-seq peaks have a rather small overlap with the CUT&Tag-based methods, whereas S9.6 CUT&Tag and HBD CUT&Tag substantially overlap (Figure 4.8a). All three CUT&Tag-based methods showed a substantial promoter bias (Figure 4.8b,c), most notably ecRNH, where more than fifty percent of called peaks localize to promoters (Figure 4.8c) and absolute quantification generally showed a low number of peaks outside of promoter regions (Figure 4.8d). Meanwhile, DRIP-seq also generated a relatively large fraction of

peaks within other gene regions (Figure 4.8b,c), most prominently towards the transcription termination site (Figure 4.8b). In absolute numbers, S9.6 CUT&Tag and HBD CUT&Tag had roughly twice as many promoter peaks as ecRNH or DRIP-seq (Figure 4.8d). The most striking difference in absolute peak numbers between all four methods was observed in distal intergenic regions, where S9.6 CUT&Tag generated approximately twice as many peaks as HBD CUT&Tag and six to seven times as many peaks as ecRNH or DRIP-seq. In terms of peak size, S9.6 CUT&Tag and HBD CUT&Tag generated the largest peaks, with median sizes around 450bp and means around 600bp.

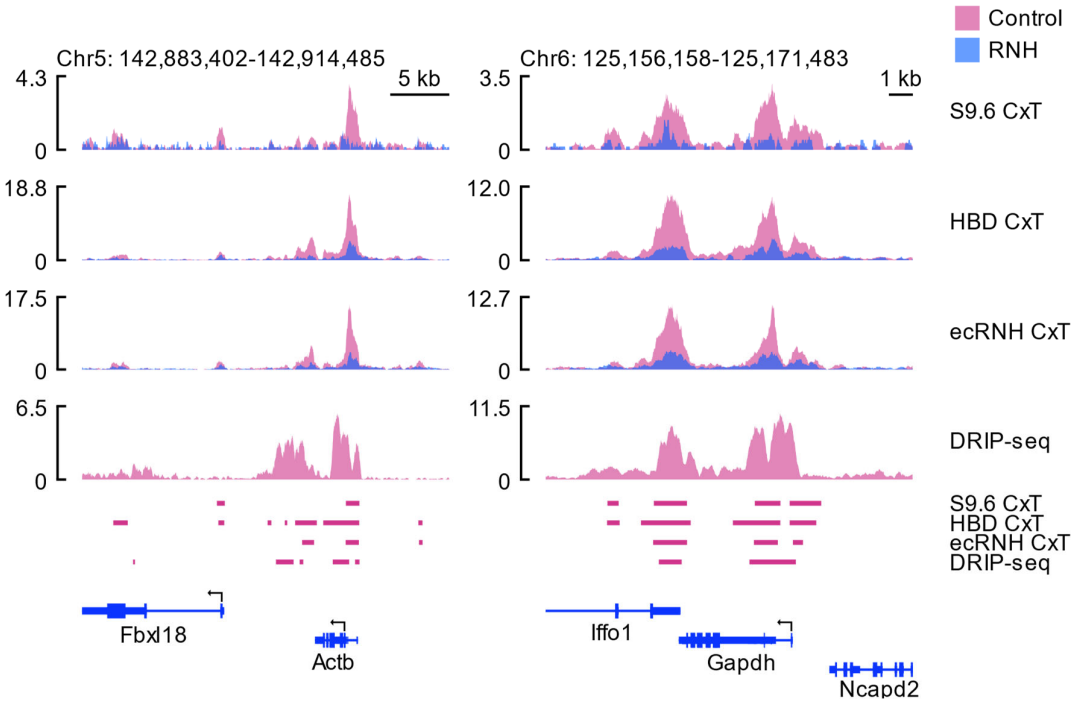


Figure 4.7: Visualization of selected loci for non-strand-specific R-loop mapping methods. R-loop mapping browser tracks for two representative R-loop containing loci around the *Actb* (left) and *Gapdh* gene (right). Upper part shows normalized read coverage for different untreated R-loop mapping methods and superimposed RNH-treated samples, if available. Bottom part shows R-loop mapping peaks along with annotated nearby genes. Peaks were called from the untreaded R-loop mapping sequencing data against the respective RNH sample, if available, otherwise against genomic background. (CxT: CUT&Tag).

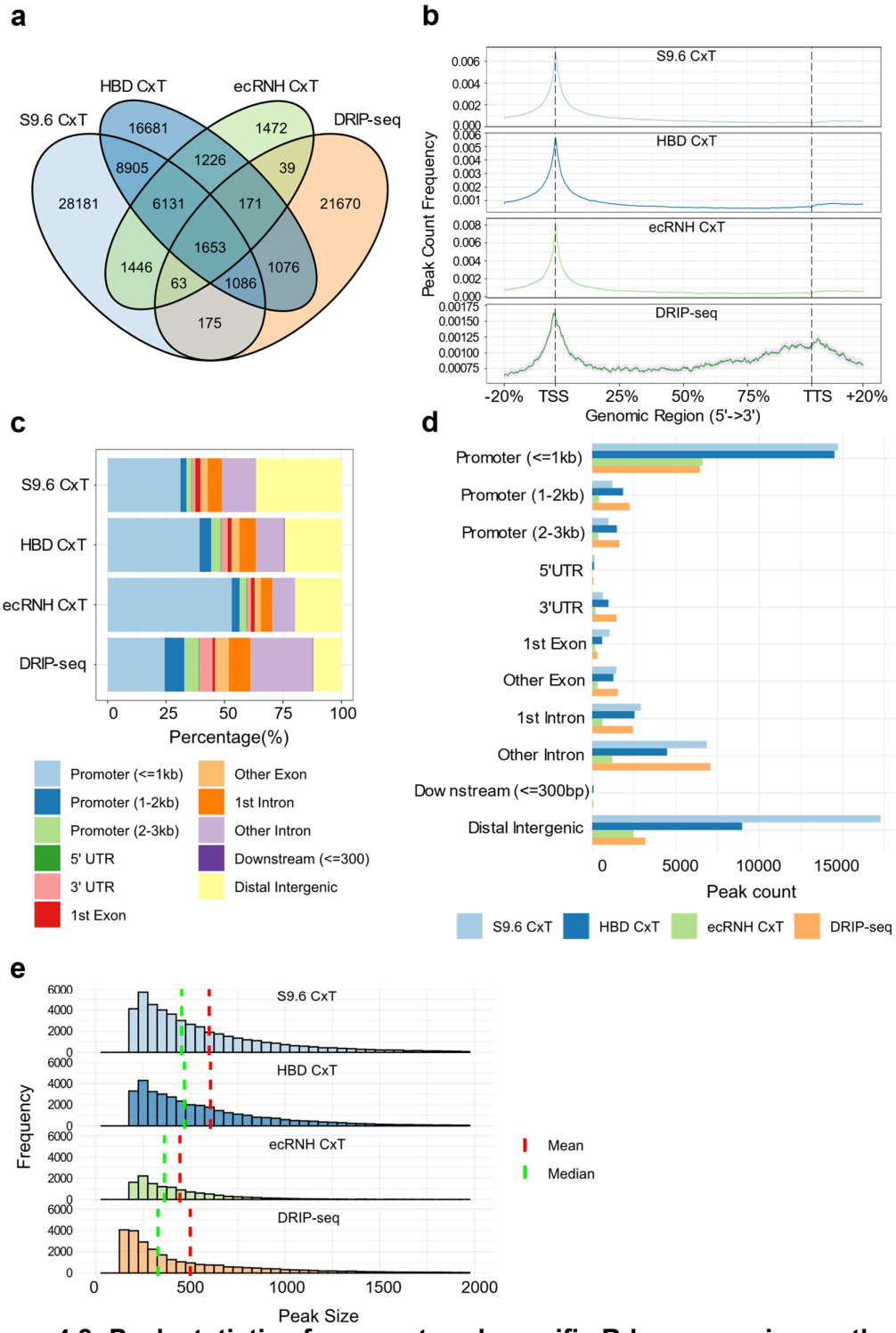


Figure 4.8: Peak statistics for non-strand-specific R-loop mapping methods. **(a)** Peak intersection, **(b)** peak distribution in annotated genes, **(c)** relative and **(d)** absolute genomic feature distribution and **(e)** peak size distribution for peaks up to 2000bp length for S9.6 CUT&Tag, HBD CUT&Tag, ecRNH CUT&Tag and DRIP-seq data. (CxT: CUT&Tag). Details on intersections of multiple genomic regions in Material and methods 6.1.9.

The previously described R-loop mapping methods do not provide any information in regard to the strand on which the DNA:RNA hybrid and the looped-out ssDNA strand reside. Therefore, Gaurav Joshi (this lab) established and applied a strand-specific DRIP-seq (strDRIP-seq) protocol with nuclease S1 treatment and chromatin fractionation via either restriction enzymes (REs) or sonication. Moreover, K. Arab performed the strand-specific Kethoxal-assisted single-stranded DNA sequencing (spKAS-seq) protocol (Wu et al., 2022) on mESCs. I developed customized pipelines to analyze these datasets. For both strDRIP-seq protocols, peaks were called against RNH-treated samples, whereas spKAS-seq peaks were called against an input sample due to poor RNH-sensitivity (not shown). All three methods generated visually distinct browser tracks, with the sonicated strDRIP-seq protocol having the clearest strand separation but often generating peaks in transcription direction that covered whole or large parts of genes (Figure 4.9). Intersection analysis showed a large agreement between the two strDRIP-seq protocols, but neither of them agreed well with the peaks generated using the spKAS-seq protocol (Figure 4.10a). Peaks called from the strDRIP-seq datasets had a transcription termination site (TTS) bias (Figure 4.10b), whereas spKAS-seq preferentially generated promoter peaks (Figure 4.10b,c,d) and sonicated strDRIP peaks showed an underrepresentation of promoter regions (Figure 4.10b). The most striking difference in absolute numbers was observed in gene body regions, where sonicated strDRIP-seq generated the most peaks (Figure 4.10d), which is consistent with the previously observed coverage of large parts of genes in the visual assessment of the tracks (Figure 4.9). spKAS-seq and RE-digested strDRIP-seq generated peaks with comparable sizes, with median peak sizes around 400bp and mean peak sizes of roughly 500bp. Meanwhile, sonicated strDRIP-seq generated much larger peaks, with a median around 750bp and a mean of more than 2000bp. In Table 4.1, I summarized the key findings derived from analyzing and comparing the tested R-loop mapping methods.

GADD45a has been shown to bind the R-loop-forming antisense lncRNA TARID in the human *TCF21* promoter and initiate TET1-mediated demethylation, thereby modulating *TCF21* transcription (Arab et al., 2019). Differential regulation of R-loop levels as a downstream effect of GADD45-binding has till date not been reported, but is plausible since R-loop formation is a transcription-dependent process. If

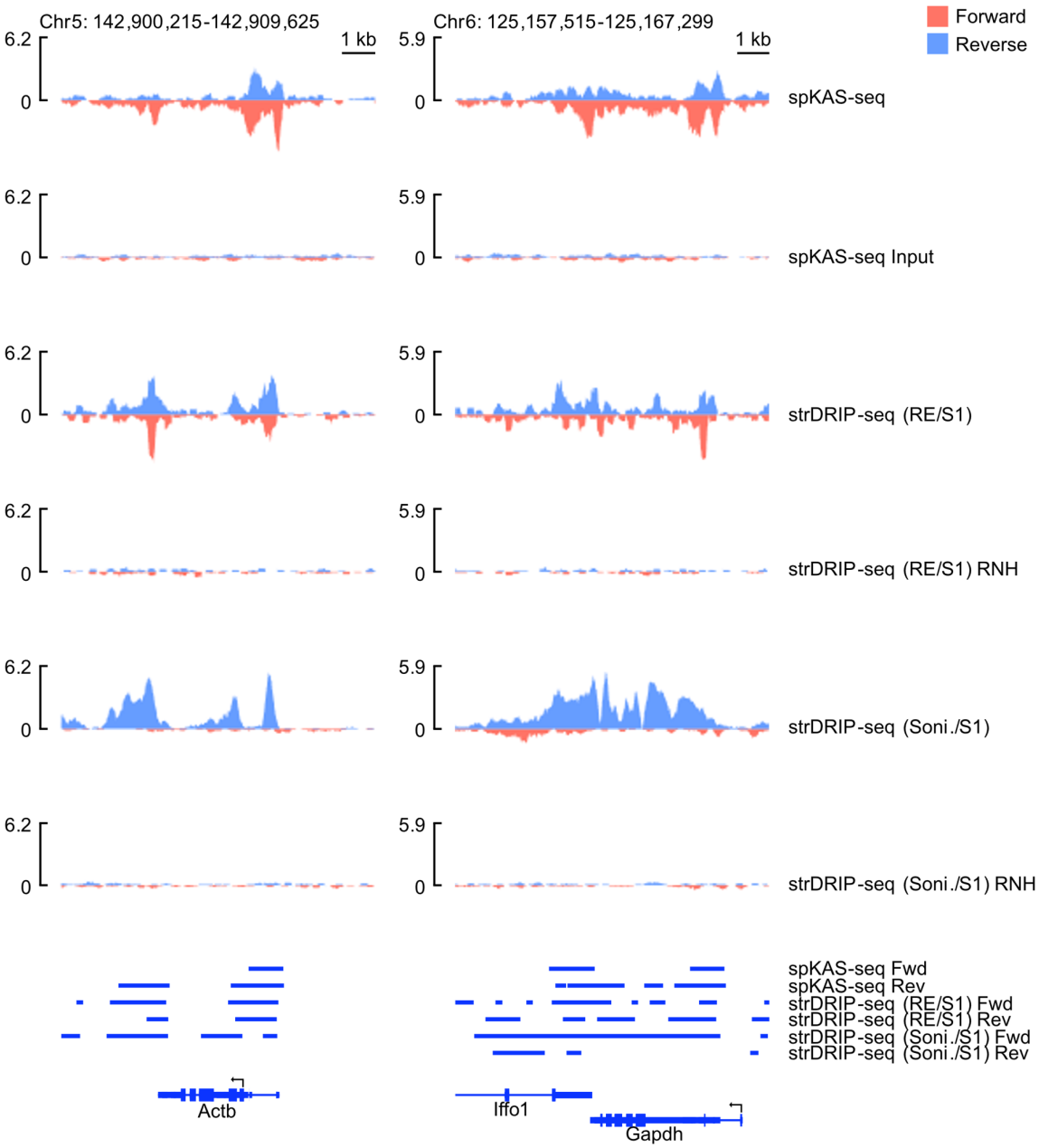


Figure 4.9: Visualization of selected loci for strand-specific R-loop mapping methods. R-loop mapping browser tracks for two representative R-loop containing loci around the *Actb* (left) and *Gapdh* gene (right). Upper part shows normalized read coverage for different untreated R-loop mapping methods and respective input or RNH-treated samples. Bottom part shows R-loop mapping peaks along with annotated nearby genes. Peaks were called from untreated R-loop mapping sequencing data against respective input or RNH-treated sample (Fwd: Forward strand, Rev: Reverse strand).

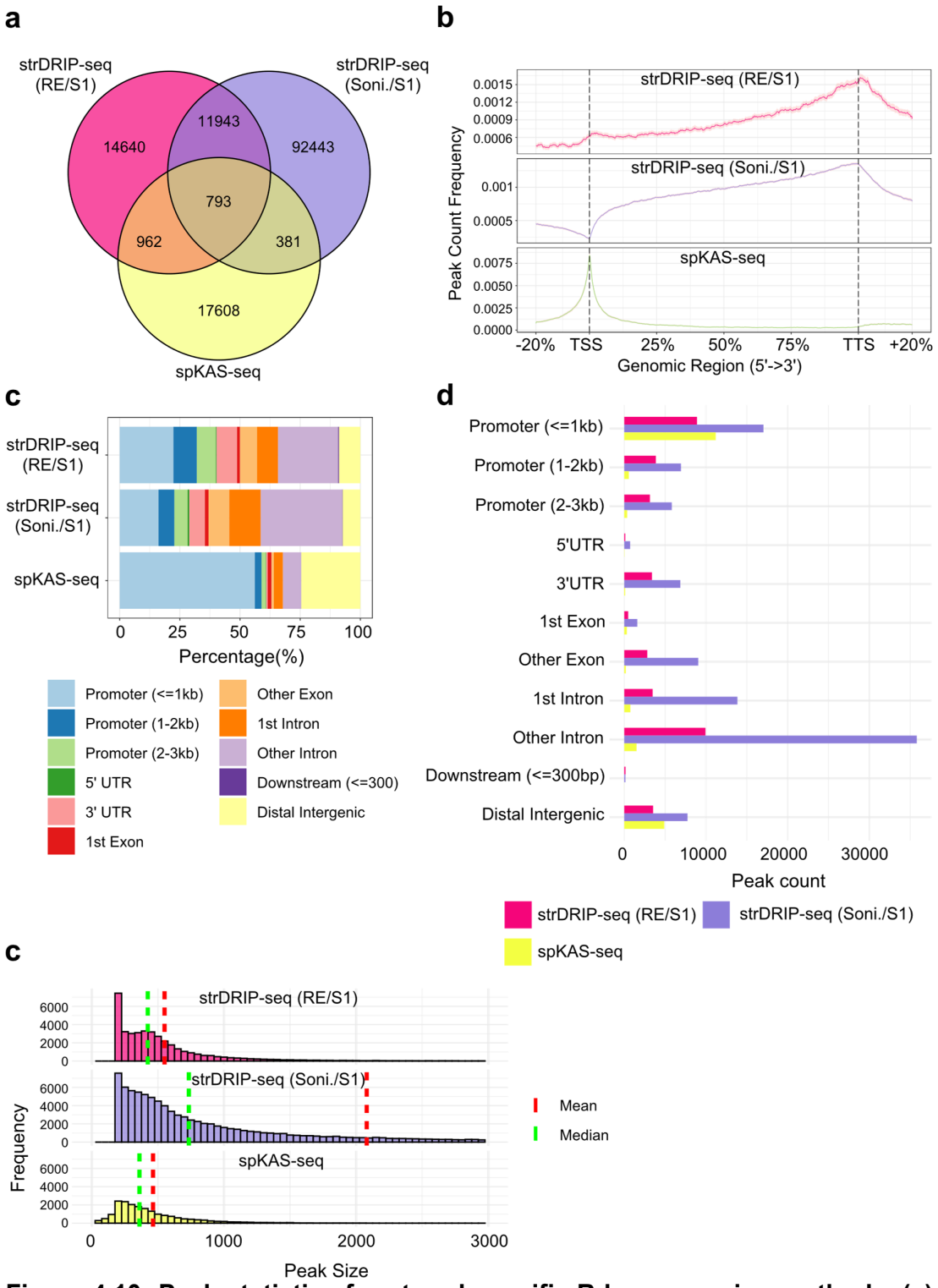


Figure 4.10: Peak statistics for strand-specific R-loop mapping methods. (a) Peak intersection, **(b)** peak distribution in annotated genes, **(c)** relative and **(d)** absolute genomic feature distribution and **(e)** peak size distribution for peaks up to 3000bp length for strDRIP-seq with RE digestion and nuclease S1 treatment (RE/S1) or sonication and nuclease S1 treatment (Soni./S1) and spKAS-seq data. Details on intersections of multiple genomic regions in Material and methods 6.1.9.

Gadd45 TKO mESCs are indeed affected by substantial differential promoter R-loop formation, it may be interesting to distinguish between R-loops formed by antisense lncRNAs and those formed in the same direction as gene transcription, thus a strand-specific protocol is required. Due to concerns over sonicated strDRIP-seq potentially capturing gene transcription signal and due to spKAS-seq's apparent superior ability to capture promoter R-loops, spKAS-seq was deemed to be the most suitable method to compare WT versus *Gadd45* TKO cell lines in a later chapter.

Table 4.1: Comparison of tested R-loop mapping methods. Comparison includes subjective rating of R-loop mapping capabilities of different genomic regions based on plots shown in Figure 4.8 and Figure 4.10 and visual assessment.

	S9.6 CxT	HBD CxT	ecRNH CxT	DRIP-seq	spkAS-seq	strDRIP-seq (RE/S1)	strDRIP-seq (Soni./S1)
Core reagent	S9.6 Ab	HBD	ecRNH	S9.6 Ab	N_3-kethoxal	S9.6 Ab	S9.6 Ab
Peaks	47,354	36,919	12,492	26,529	19,967	39,880	105,631
Mean peak size	599bp	445bp	606bp	499bp	463bp	550bp	2080bp
Median peak size	454bp	363bp	468bp	328bp	360bp	423bp	734bp
Strand-specific	no	no	no	no	yes	yes	yes
RNH-sensitivity	high	high	high	n.t.	poor	high	high
Promoter peaks*	++/++	++/++	+/++	+/++	+/++	+/+	++/-
Gene body peaks*	+/-	+/-	-/-	+/++	-/-	+/+	+/+
Gene 3' peaks*	-/-	-/-	-/-	+/+	-/-	+/++	+/++
Intergenic peaks*	++/++	+/+	-/+	-/+	+/+	+/+	++/+

CxT: CUT&Tag, Ab: Antibody, n.t.: not tested, ++: very high, +: high, -: low, *: first rating compares method to other tested methods in the same category (non-stranded/stranded) by absolute peak numbers, second rating compares peaks counts for the respective genomic region relative to other regions within that same method, vertical dashed line separates methods strand-specific and non-strand-specific methods.

4.2.2 Karyotyping and selection of *Gadd45* TKO mESC clones

To study the function of GADD45 proteins in mESCs, sequencing data from *Gadd45* TKO mutant mESC clones is compared with WT control clones in a later chapter. *Gadd45* TKO mutant cell lines are engineered to be unable to synthesize functional GADD45a/b/g proteins. Matched WT and *Gadd45* TKO mutant clone cell lines will hereafter be referred to as *Gadd45* TKO clone set. An mESC *Gadd45* TKO clone set that was previously generated in our lab (Schüle et al., 2019), hereafter *Gadd45* TKO clone set 1, turned out to be likely hypomorphic with incomplete deletion of the *Gadd45* coding sequence (unpubl. data). Therefore, Lars Schomacher (this lab) generated *Gadd45* TKO clone set 2 using CRISPR/Cas9 technology in two steps.

For *Gadd45* TKO clones, a *Gadd45g* single knockout (SKO) mutant was first created and then used as a parental clone for independent *Gadd45a* and *Gadd45b* knockouts in a second step. WT clones were analogously mock-treated and picked in two steps. Eventually, *Gadd45* TKO clone set 2 consisted of clones TKO86, TKO266, TKO273, TKO279, WT1, WT3, WT7 and WT12. The Gadd45 TKO clones were later confirmed to lack the complete coding sequence for both alleles of the three paralogous *Gadd45a*, *Gadd45b* and *Gadd45g* genes (not shown).

To detect potential undesired chromosomal copy number alterations in *Gadd45* TKO clone set 2, Khelifa Arab performed whole genome sequencing (WGS) with pooled WT and pooled *Gadd45* TKO samples. I estimated the chromosome-wise copy number differences between the two groups from normalized WGS read counts per chromosome (Figure 4.11a). The analysis suggested a 1:1.5 ratio in chromosome 8 copy numbers, hinting at a possible chromosome 8 trisomy in *Gadd45* TKO clones. Moreover, the mitochondrial genome showed a ~1:2.8 ratio, indicating mitochondrial gain in *Gadd45* TKOs or loss in WT. Since the analysis was performed on pooled samples for all *Gadd45* TKO and WT clones, and since it was not definitely clear whether the relative results indicated chromosome 8 gains in *Gadd45* TKO samples or chromosome 8 loss in WT, Gaurav Joshi validated the results by quantitative PCR (qPCR). In brief, he selected five loci on chromosome 8 and a control locus in an intergenic region on chromosome 10 and measured their amplification rate relative to the mean amplification rate of four independent loci on other chromosomes by qPCR for each clone in the *Gadd45* TKO clone set 2. The qPCR experiment indicated a 1.3 to 1.5-fold increased signal for the loci on chromosome 8 (Figure 4.11b), confirming that a substantial fraction of cells in all TKO clones from *Gadd45* TKO clone set 2 clones have chromosome 8 trisomy. Due to concerns over confounding effects resulting from the copy number differences influencing results in TKO versus WT comparisons and to test whether the copy number differences are a reproducible consequence of the lack of GADD45 proteins, Lars Schomacher generated *Gadd45* TKO clone set 3. In brief, he first generated and picked two independent *Gadd45g* SKO mESC mutants, SKO19 and SKO50 by using CRISPR/Cas9. These SKO mutants were then used as parental clones for independent *Gadd45a* and *Gadd45b* knockouts in a second step. SKO19 gave rise to TKO clones TKO259_ParSKO19, TKO271_ParSKO19,

TKO353_ParSKO19 whereas parental clone SKO50 was used to generate TKO81_ParSKO50, TKO83_ParSKO50, TKO330_ParSKO50 and TKO379_ParSKO50. Mock-treated WT control clones were generated and picked analogously in two steps. Parental WT clone Par2 was used to generate WT1_Par2, WT3_Par2, WT4_Par2 and WT5_Par2, whereas WT3_Par7, WT4_Par7, WT5_Par7 and WT6_Par7 were generated from parental WT clone Par7.

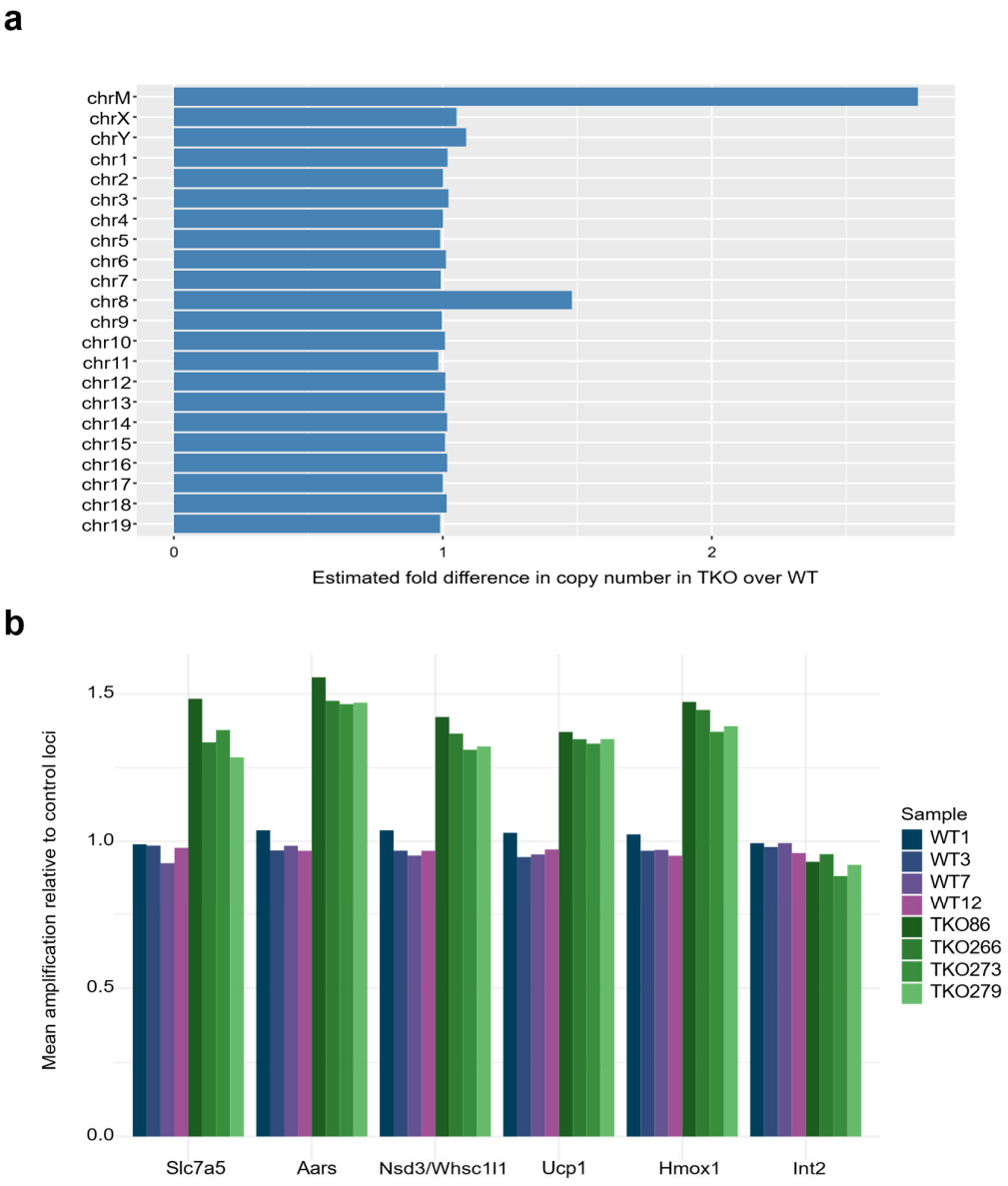


Figure 4.11: *Gadd45* TKO clones in *Gadd45* TKO clone set 2 have chromosome 8 trisomy. (a) Estimated fold differences in chromosome copy number from pooled WT and TKO WGS samples. **(b)** qPCR signal for five loci on chromosome 8 (*Slx7a5*, *Aars*, *Nsd3*/*Whsc1l1*, *Ucp1*, *Hmox1*).

To test for chromosomal copy-number alterations, Khelifa Arab performed a WGS for each clone in *Gadd45* TKO clone set 3. For each sample, I estimated the fold change in chromosome copy numbers in relation to the median chromosome copy number in all samples. The results suggested approximately a 20% reduction in chromosome 14 signal in WT clones with parental clone Par7, i.e. clones WT3_Par7, WT4_Par7, WT5_Par7, WT6_Par7 (Figure 4.12). Moreover, a roughly 30% reduction in chromosome Y signal was observed for all *Gadd45* TKO clones originating from SKO50, i.e. TKO81_ParSKO50, TKO83_ParSKO50, TKO330_ParSKO50 and TKO379_ParSKO50 (Figure 4.12). Additionally, WT clones originating from parental clone Par2, as well as *Gadd45* TKOs generated from parental clone SKO50 had a higher mitochondrial genome copy number than WT clones originating from Par7 and TKO clones generated from SKO19 (Figure 4.12). Gaurav Joshi performed RNA-seq using all clones from *Gadd45* TKO clone set 3 to confirm the successful knockout of *Gadd45* transcripts at the mRNA level. I analyzed the sequencing data and visualized the gene expression in the three *Gadd45* loci that were targeted by CRISPR/Cas9. The results indicated that for all but one TKO clones, the coding sequences of all three target genes *Gadd45a* (Supplementary Figure 4.1), *Gadd45b* (Supplementary Figure 4.2) and *Gadd45g* (Supplementary Figure 4.3) were successfully knocked out. TKO clone TKO259_ParSKO19 had minor but visible *Gadd45a* expression (Supplementary Figure 4.1) and while for TKO379_ParSKO50, the coding sequence for all three target loci was successfully knocked out, there was a visible up regulation of 3'UTR expression in *Gadd45b* (Supplementary Figure 4.2).

I thus decided to exclude clones WT3_Par7, WT4_Par7, WT5_Par7, WT6_Par7 due to increased loss of chromosome 14, TKO259_ParSKO19 due to visible Gadd45a expression and TKO379_ParSKO50 due to abnormal Gadd45b 3'UTR up regulation. Additionally, TKO353_ParSKO19 was excluded due to visual abnormalities of the cells' morphology (internal lab communication). This means in turn that clones WT1_Par2, WT3_Par2, WT4_Par2, WT5_Par2, and TKO271_ParSKO19, TKO81_ParSKO50, TKO83_ParSKO50, TKO330_ParSKO50 remained to be used in upcoming analyses. Since the latter three share an approximately 30% reduced chromosome Y signal, it was important to investigate which parts of chromosome Y were lost to assess the risks of potential

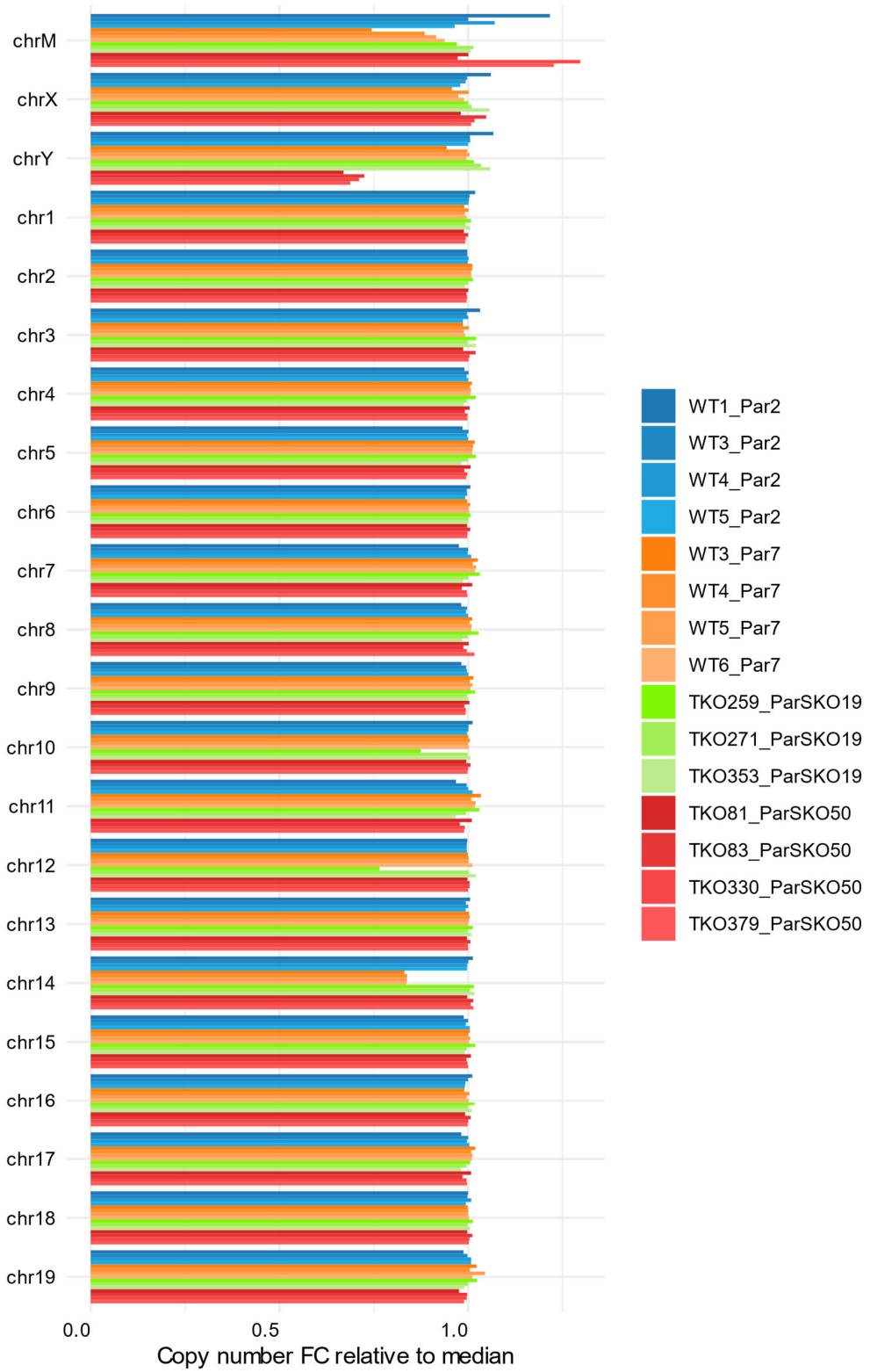


Figure 4.12: Clones in *Gadd45* TKO clone set 3 have chromosomal copy number alterations. Estimated copy number fold changes relative to per chromosome median signal of *Gadd45* TKO clone set 3 samples.

effects that might impact future WT versus TKO comparisons. Visual inspection of the mapped *Gadd45* TKO clone set 3 WGS data on chromosome Y indicated that large parts of chromosome Y (approximately chrY:8.4Mb-82.1Mb) have indeed been lost in TKO clones TKO81_ParSKO50, TKO83_ParSKO50, TKO330_ParSKO50 and TKO379_ParSKO50 (Figure 4.13). Importantly, regions towards both chromosome ends, (approximately chrY:0-8.4Mb and chrY: 82.1Mb-91.7Mb), including the region containing the important "sex-switch" gene *Sry* were confirmed to be present (Figure 4.13).

4.2.3 Differential gene expression in *Gadd45* TKO mESC clones

I aimed to understand the role of GADD45 family proteins in transcriptional regulation. Thus, to determine transcriptionally misregulated genes, as a first step, RNA-seq in WT and *Gadd45* TKO mESC was performed by Gaurav Joshi. I performed a differential expression analysis, which resulted in 1,056 up and 936 down regulated genes (FDR < 0.01). As expected, *Gadd45a* and *Gadd45b* were among the strongest down regulated genes by fold-change and false discovery rate (FDR) (Figure 4.14a). Principle component analysis of the samples resulted in a clustering of all samples by genotype (Figure 4.14b). It should be noted that TKO271_ParSKO19, the only TKO clone not affected by increased chromosome Y-loss, clustered separately from the other TKO clones in respect to principal component 2 (PC2). GO enrichment analysis and KEGG analysis of significantly up regulated genes resulted in an enrichment of synthesis of small molecules, mitochondrial energy generation and neuronal diseases (Figure 4.14c,e), whereas down regulated genes resulted in an enrichment of terms related to heart development and signaling pathways such as MAPK signaling (Figure 4.14d,f). Consistent down regulation of heart development genes across all *Gadd45* TKO clones was confirmed by visual assessment, with selected genes shown in Supplementary Figure 4.4. KEGG visualization of the diabetic cardiomyopathy pathway indicated an up regulation of all mitochondrial electron transport chain complexes (Supplementary Figure 4.5).

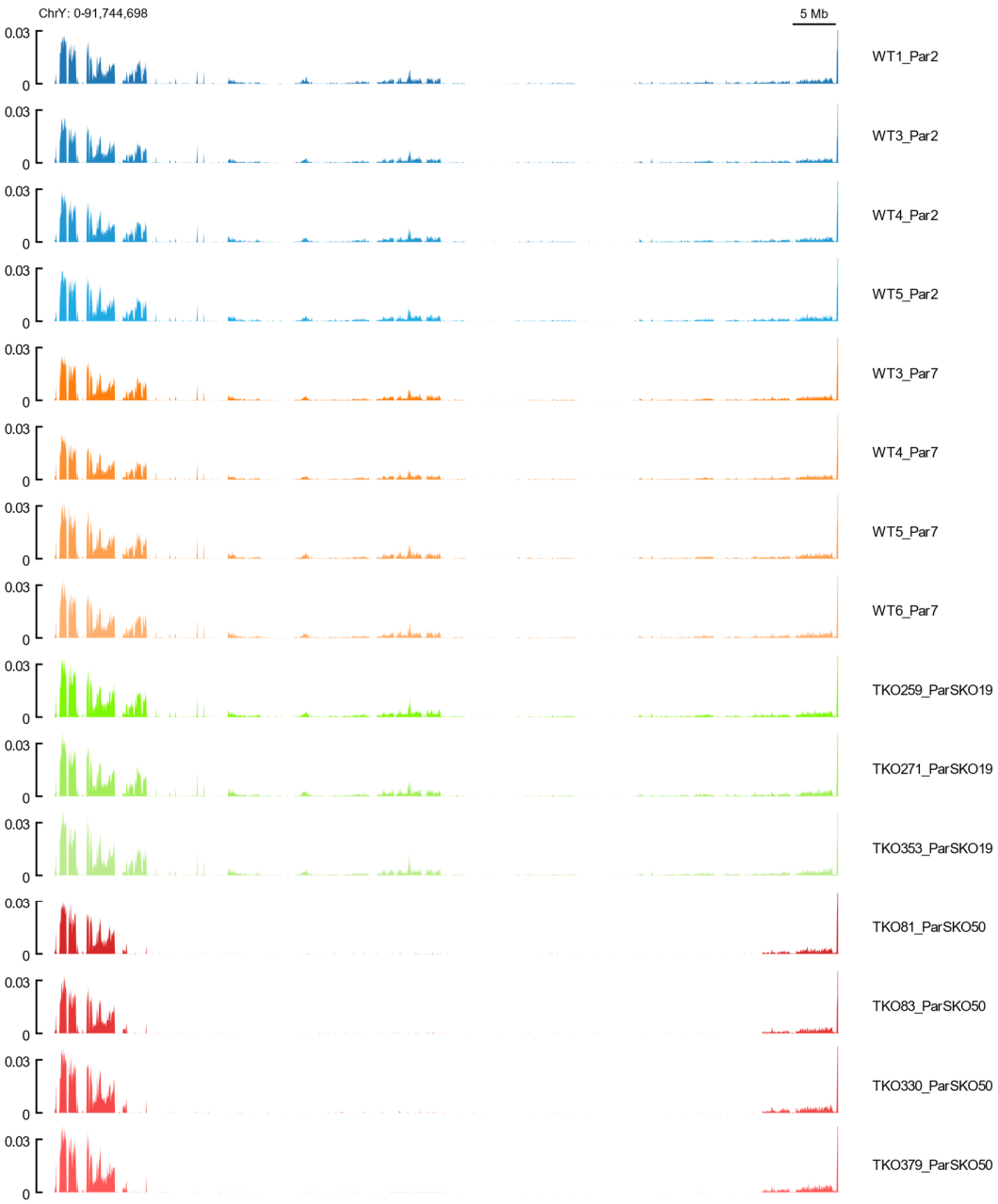


Figure 4.13: *Gadd45* TKO mESC clones originating from parental clone SKO50 lost large parts of chromosome Y. Visualization of PCR duplicate and multimapper-filtered WGS data for all *Gadd45* TKO mESCs from *Gadd45* TKO clone set 3. Tracks display the whole Y chromosome. Y axis shows CPM-normalized read counts. High-signal regions were clipped to avoid issues with the scaling of the tracks.

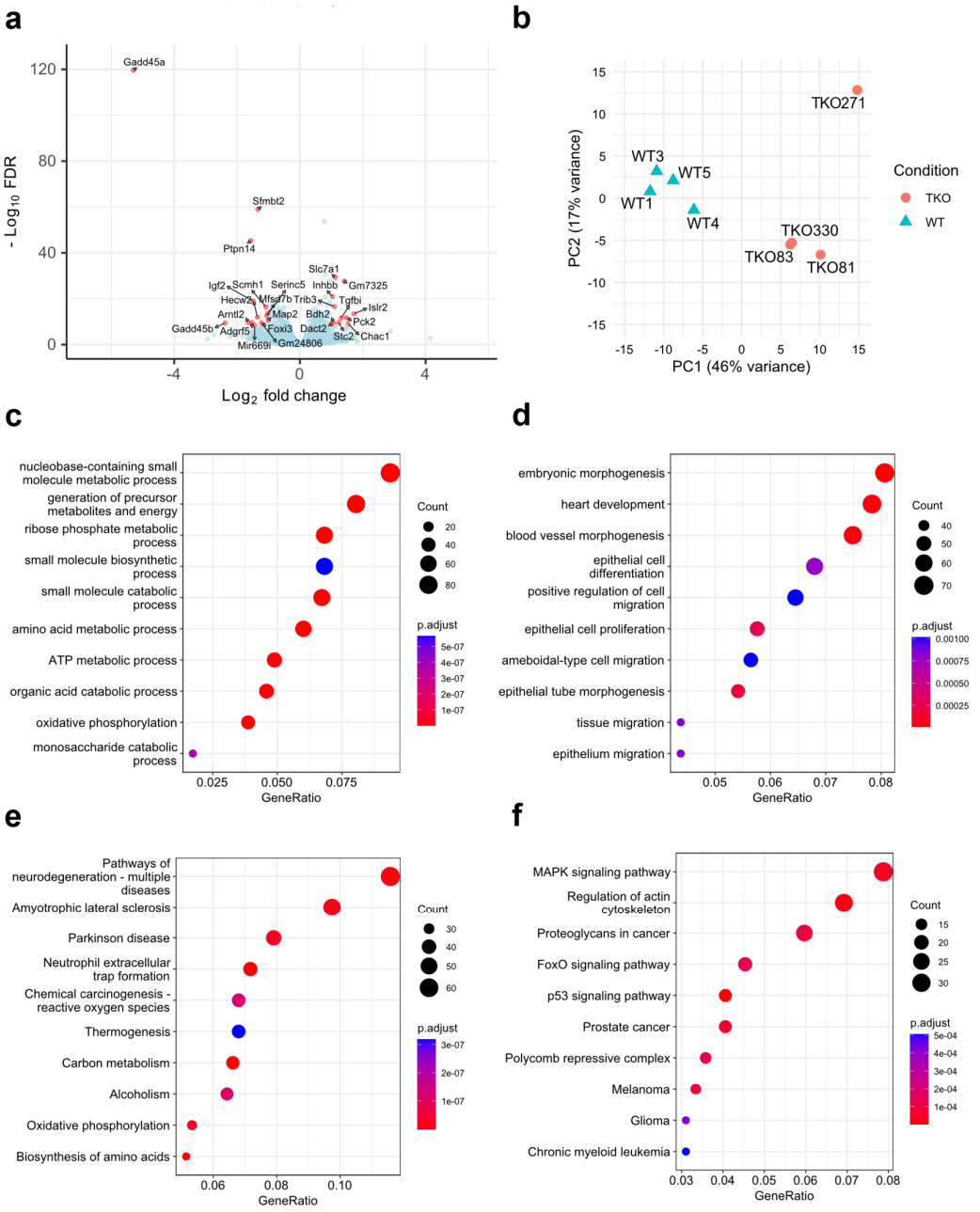


Figure 4.14: Differential expression analysis comparing *Gadd45* TKO mESC versus WT mESCs. (a) Volcano plot illustrating *Gadd45* TKO versus WT mESCs differential expression results with labeled top hits (q-value cutoff: 10^-8, FC cutoff: 1). **(b)** Principal component analysis of RNA-seq data (with abbreviated clone labels omitting parental origin) and GO enrichment analysis for **(c**) up and **(d)** down *Gadd45* TKO DEGs. KEGG pathway enrichment analysis for **(e)** up and **(f)** down *Gadd45* TKO DEGs. DEGs identified at an FDR < 0.01.

Previous unpublished results of this lab showed a reduced myocardium thickness in mice deficient for *Gadd45a* and *Tet1* (unpubl. results). Accordingly, I concentrated on the subset of down regulated genes corroborating the involvement of GADD45 proteins in heart development. To test whether the differentially down regulated heart development genes impact cardiac differentiation, Gaurav Joshi performed *in vitro* cardiac differentiation of WT and *Gadd45* TKO embryoid bodies (EBs). For this, he mimicked *in vivo* cardiac differentiation by inhibition of BMP4 and Wnt signaling at specific time points and performed RNA-seq at day 10 and 12 of differentiation. Ettore Zapparoli (this lab) mapped and counted resulting reads. I performed GO enrichment analysis of the differentially expressed genes in *Gadd45* TKO EBs (FDR < 0.01). Down regulated genes in *Gadd45* TKO EBs were enriched for GO terms associated with heart development (Figure 4.15a,b), suggesting a potential impairment in this process. WT EBs at day 10 of differentiation showed spontaneously beating regions, comparable to rhythmically contracting cardiomyocytes during *in vivo* heart development, while in *Gadd45* TKO EBs, the number of beating EBs was significantly reduced (Figure 4.15c). These results support the hypothesis that GADD45 family proteins are involved in and required for the cardiac differentiation process.

4.2.4 Mapping chromatin accessibility in *Gadd45* TKO mESCs

To understand whether the observed changes in gene expression in *Gadd45* TKO mESCs are linked to differential open chromatin, an Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq) was performed by Rintu Umesh (this lab). I called a total of 10,1615 consensus peaks for WT clones and performed a differential chromatin accessibility analysis, which resulted in 431 down and 582 up peaks (FDR < 0.05) with the “up” peaks displaying a preference to localize to promoter regions (Figure 4.16a,b). Motif search in the up ATAC-seq peaks resulted, among others, in an enrichment for motifs bound by the KLF family of transcription factors, including the pluripotency factor KLF4, as well as motifs bound by the nuclear receptors NR4A1, NR2F2 and ESRR (Figure 4.16c). “Down” ATAC-seq peaks showed an enrichment in motifs bound by AP-2, a transcription factor crucial for the development of neural crest cells, as well as motifs bound by

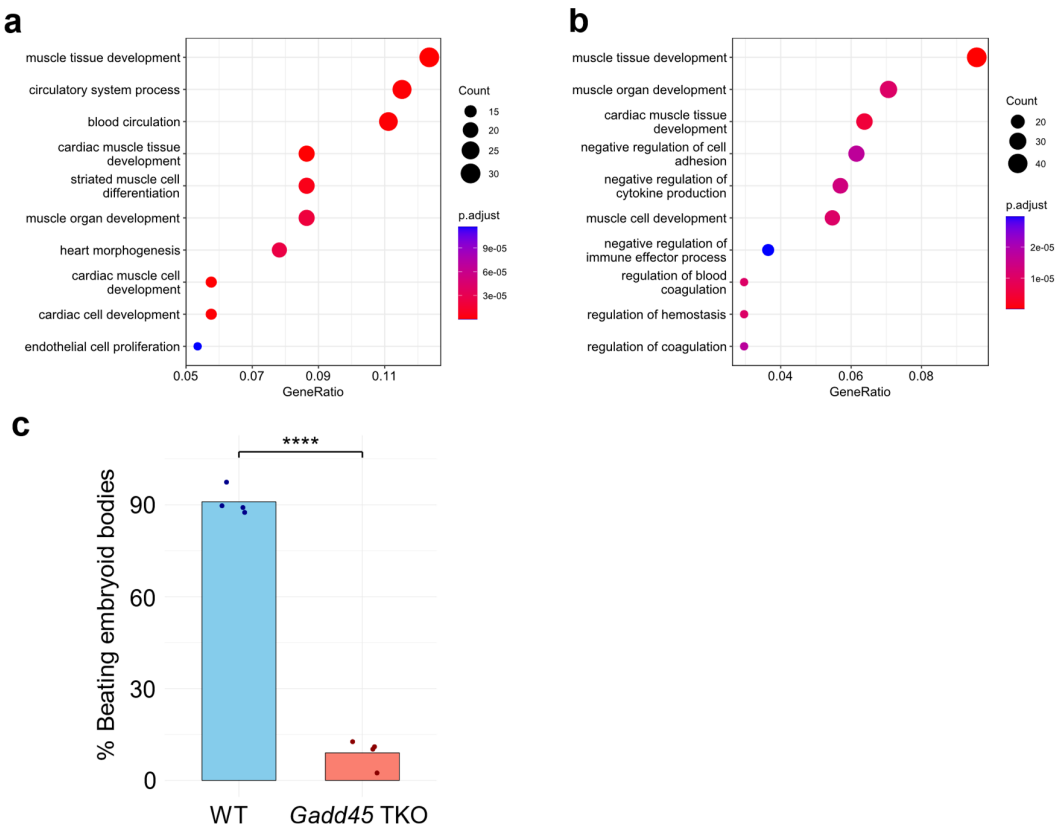


Figure 4.15: *Gadd45* TKOs have a cardiac differentiation defect. GO enrichment analysis for *Gadd45* TKO down regulated genes at **(a)** day 10 and **(b)** day 12 of cardiac differentiation. **(c)** Bar graph shows percentage of beating EBs at day 10 of *in vitro* cardiac differentiation of WT and *Gadd45* TKO mESCs. Experiment by Gaurav Joshi (~120 EBs per genotype were analyzed, statistical significance tested by two-tailed unpaired Student's t-test, ****: p-value <0.0001).

the pluripotency factor NANOG and the TEAD family of transcription factors involved in muscle and heart development (Figure 4.16d). GO enrichment analysis in up ATAC-seq peaks resulted in a functionally diverse set of terms (Figure 4.16e), while GO enrichment analysis in down ATAC-seq peaks resulted in terms related to development of different structures (Figure 4.16f). Intersection between differential ATAC-seq peaks and differentially expressed genes (DEGs) suggests a trend of up DEGs overlapping with up ATAC-seq peaks and down DEGs overlapping with down DEGs (Figure 4.17). This is expected since a state of increased open chromatin enables an increased binding of transcription factors, RNA polymerases and other factors, while decreased binding due to more condensed chromatin causes the opposite. However, the large majority of DEGs cannot be explained by direct intersection with differential open chromatin (Figure 4.17).

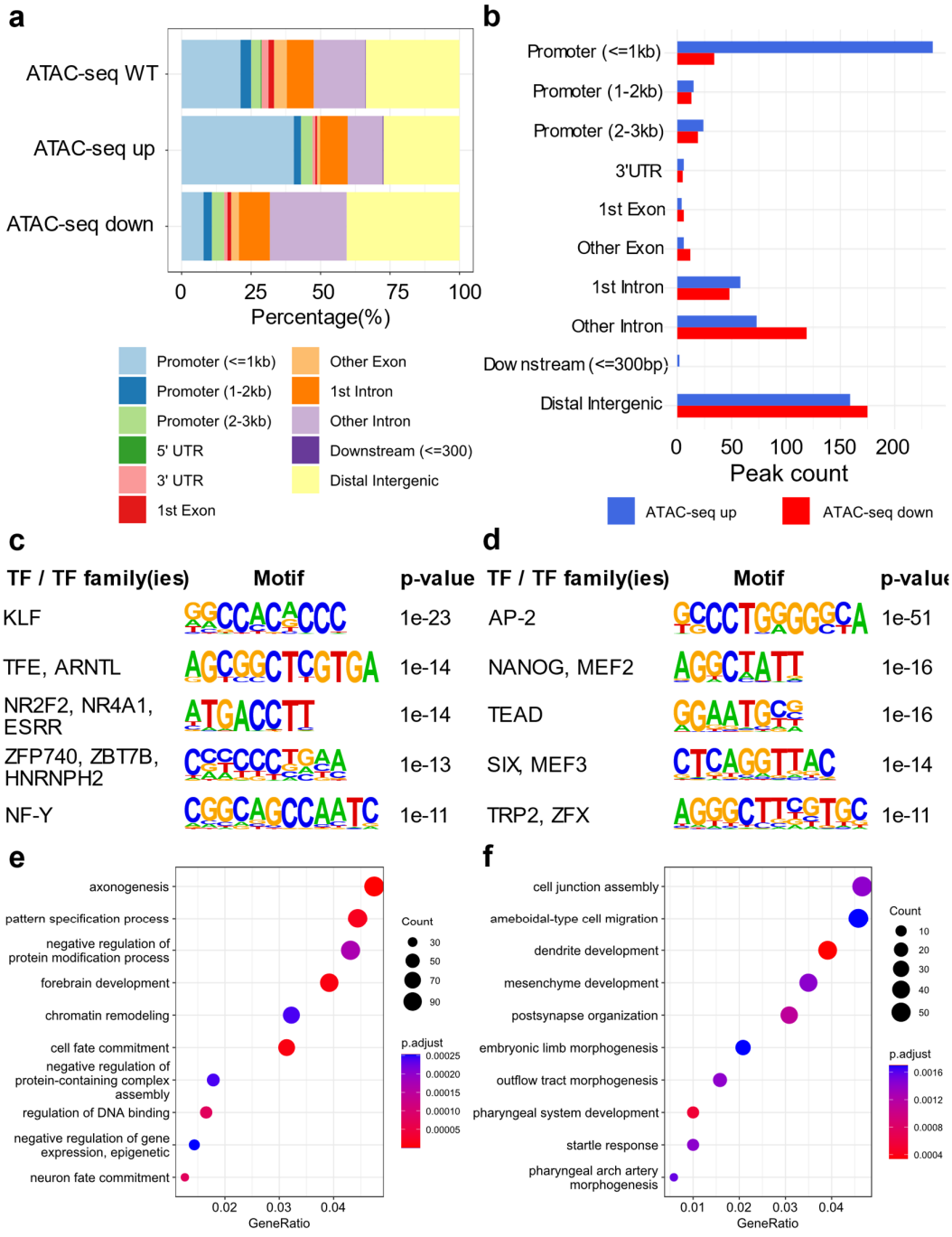


Figure 4.16: ATAC-seq analysis of *Gadd45* TKO mESCs. (a) Relative genomic feature distribution of WT consensus ATAC-seq peaks (ATAC-seq WT) for reference and *Gadd45* TKO up (ATAC-seq up) and down regulated (ATAC-seq down) peaks. **(b)** Absolute genomic feature distribution of ATAC-seq up and ATAC-seq down peaks. Motif enrichment analysis results in **(c)** up and **(d)** down regulated ATAC-seq peaks. GO enrichment analysis in **(e)** up and **(f)** down regulated ATAC-seq peaks.

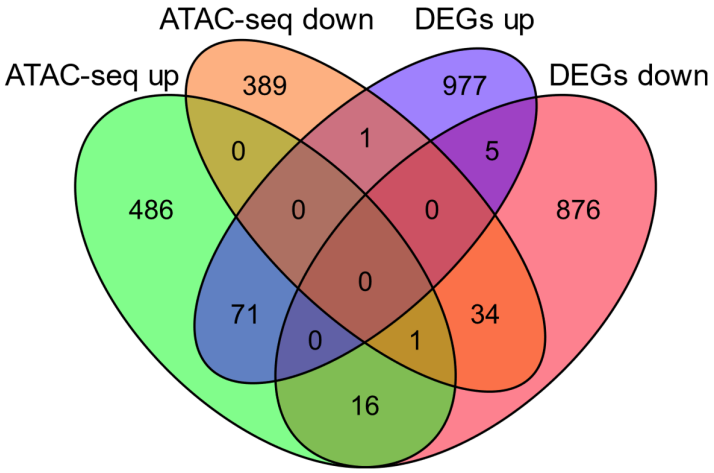


Figure 4.17: Majority of Gadd45 TKO mESC DEGs cannot be explained by differential open chromatin. Intersection of differential *Gadd45* TKO ATAC-seq peaks with *Gadd45* TKO DEGs. DEGs were extended 3kb upstream to include promoter regions in the overlap analysis. Details on intersections of multiple genomic regions in Material and methods 6.1.9.

4.2.5 R-loop mapping *Gadd45* TKO mESCs

GADD45a has been previously shown to directly bind to a promoter antisense R-loop, thereby inducing TET1-mediated DNA demethylation and inducing transcription (Arab et al., 2019). Whether this process can modulate the transcription of the DNA:RNA hybrid-forming RNA and alter cellular R-loop levels is unclear. To assess differential R-loop content in *Gadd45* TKO mESCs, Khelifa Arab performed spKAS-seq in WT and *Gadd45* TKO mESCs. I called a total of 19,967 consensus peaks in WT cells and differential comparison resulted in 109 spKAS-seq peaks with increased and 98 spKAS-seq peaks with decreased occupancy in *Gadd45* TKOs (FDR < 0.05). Differential peaks, similar to WT consensus peaks, mostly located to promoters or enhancers and R-loop orientation in respect to the closest gene did not appear to play a role (Figure 4.18a,b). GO enrichment analysis of up spKAS-seq peaks resulted in an enrichment of genes involved in genomic imprinting, whereas down peaks were enriched in terms related to rRNA processing (Figure 4.18c,d). Differential spKAS-seq peaks intersected with only few DEGs (Figure 4.18e). Given the limited overlap between altered R-loops and *Gadd45* TKO DEGs, it appears unlikely that aberrant R-loop levels within the loci of *Gadd45* TKO DEGs are responsible for their differential expression.

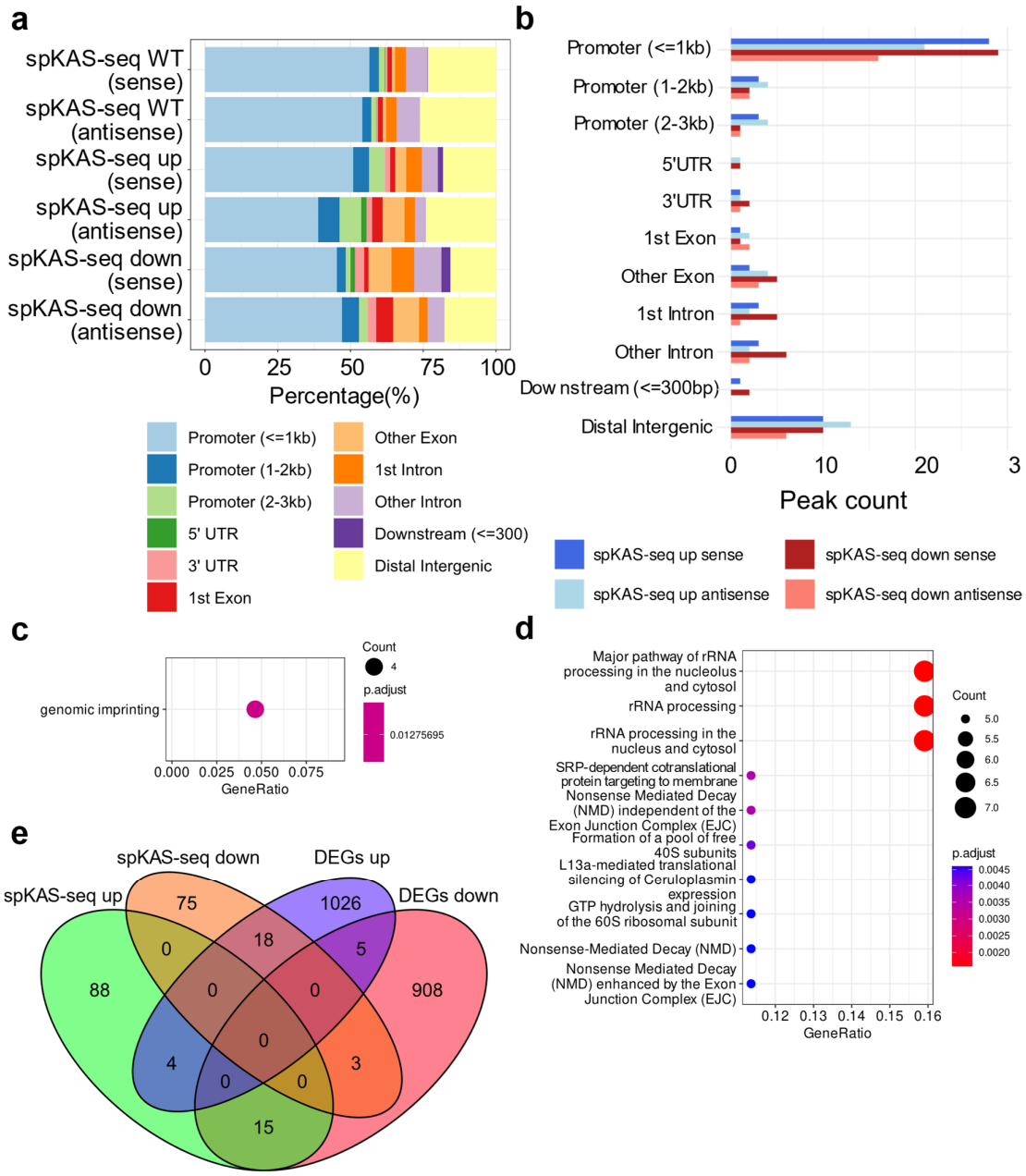


Figure 4.18: spKAS-seq analysis of *Gadd45* TKO mESCs. (a) Relative genomic feature distribution of WT consensus spKAS-seq peaks (spKAS-seq WT) for reference and *Gadd45* TKO up (spKAS-seq up) and down regulated (spKAS-seq down) peaks. **(b)** Absolute genomic feature distribution of *Gadd45* TKO spKAS-seq up and spKAS-seq down peaks. Sense and antisense indicate spKAS-seq peak orientation relative to closest annotated gene. GO enrichment analysis of **(c)** up and **(d)** down regulated spKAS-seq peaks. **(e)** Intersection of differential spKAS-seq peaks with DEGs in *Gadd45* TKO mESCs. DEGs were extended 3kb upstream to include promoter regions in the overlap analysis. Details on intersections of multiple genomic regions in Material and methods 6.1.9.

4.2.6 Impact of *Gadd45*-loss on TET1-binding and DNA methylation

GADD45a was previously shown to recruit TET1 in order to initiate TET1-mediated DNA demethylation at the human TCF21 promoter (Arab et al., 2019). To investigate whether TET1-binding is impaired in mESCs that lack GADD45 family proteins, Rintu Umesh performed TET1 CUT&Tag in WT and *Gadd45* TKO mESCs. My analysis resulted in 61,881 TET1 consensus peaks in WT mESCs. Interestingly, 6,403 peaks had significantly increased TET1 occupancy, while 1,513 peaks were significantly decreased in *Gadd45* TKO mESCs (FDR < 0.05). Increased TET1 peaks were predominantly enriched in promoters, whereas the feature distribution of depleted TET1 peaks sites was rather comparable to WT consensus TET1 peaks, albeit with a decreased promoter fraction (Figure 4.19a,b). GO enrichment analysis for TET1 up peaks yielded a wide variety of terms, whereas TET1 down peaks were associated with neuron-related processes (Figure 4.19c,d). TET1 up peaks intersected well with both up and down DEGs in Gadd45 TKOs (Figure 4.19e).

To test whether the increased TET1-binding in *Gadd45* TKOs may explain the DEGs in *Gadd45* TKO mESCs, I conducted a GO enrichment analysis for the DEGs intersecting with the *Gadd45* TKO TET1 up peaks. The resulting GO terms (Figure 4.20a,b) largely resembled the previously obtained GO terms for the DEGs without TET1 up peak intersection (Figure 4.14c,d). Increased TET1 binding could thus be responsible for the misregulation of these genes.

TET1 can function as a demethylase. The previously observed increase in TET1-binding at both *Gadd45* TKO up and down DEGs may thus result in depleted DNA methylation levels at these loci. Therefore, Yulia Kargapolova (this lab) compared DNA methylation levels in WT versus *Gadd45* TKO mESCs via Illumina Infinium Mouse Methylation BeadChip. Differential analysis was performed separately for (1) differentially methylated promoters, (2) differentially methylated gene bodies and (3) differentially methylated tiles. The analysis resulted in substantial differential DNA methylation in *Gadd45* TKOs: (1) 1,094 hyper and 3,251 hypomethylated promoters, (2) 782 hyper and 2,279 hypomethylated gene bodies and (3) 10,029 hyper and 13,305 hypomethylated tiles. Notably, hypomethylation was more prevalent than hypermethylation, particularly within gene bodies and promoters. I intersected promoters and gene bodies with differential DNA methylation with the

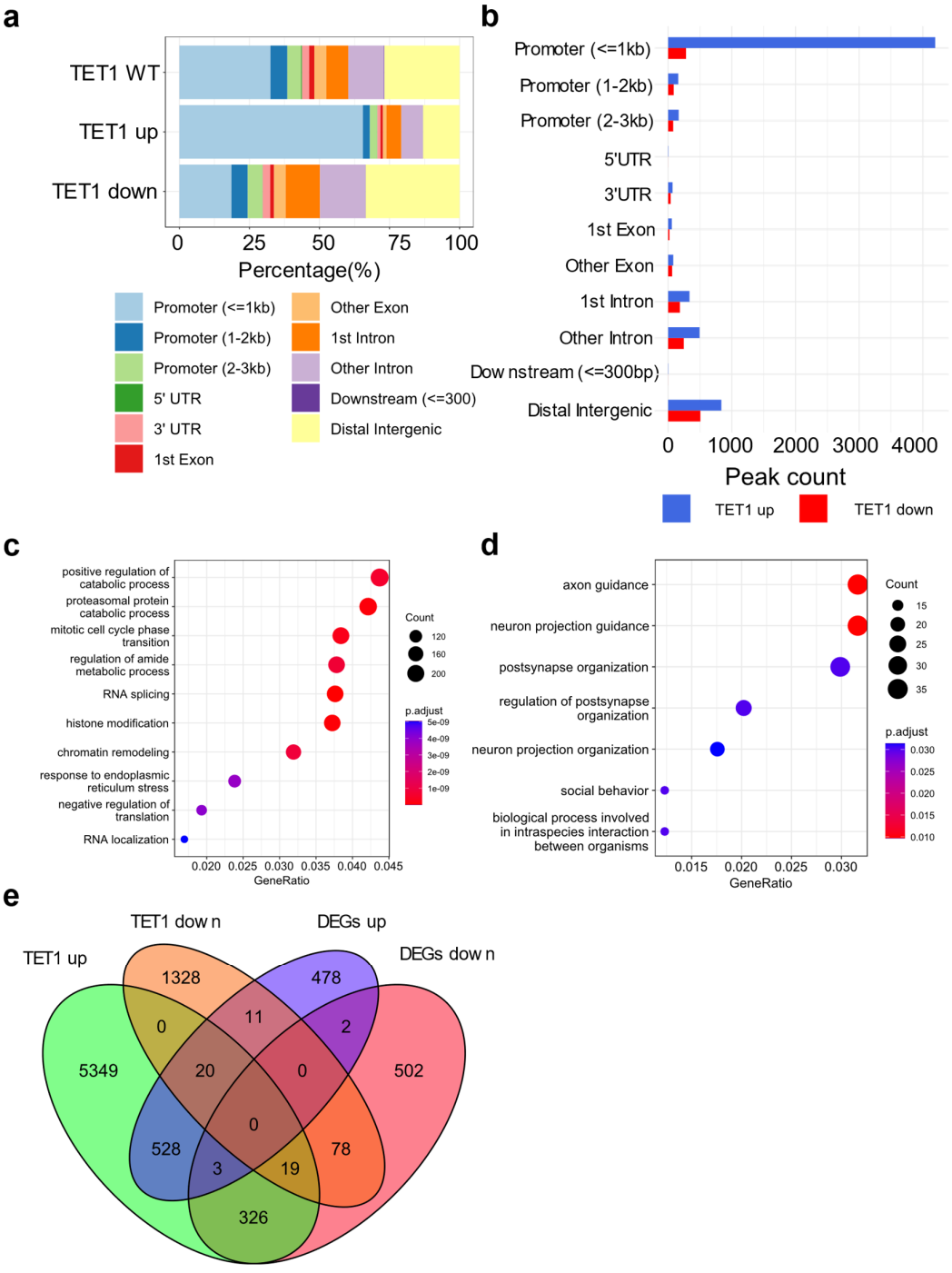


Figure 4.19: TET1 CUT&Tag analysis of *Gadd45* TKO mESCs. (a) Relative genomic feature distribution of WT consensus TET1 peaks (TET1 WT) for reference and *Gadd45* TKO up (TET1 up) and down regulated (TET1 down) peaks. **(b)** Absolute genomic feature distribution for up and down TET1 peaks. GO enrichment analysis of **(c)** up regulated and **(d)** down TET1 peaks. **(e)** Intersection of differential TET1 peaks with DEGs in *Gadd45* TKO mESCs. DEGs were extended 3kb upstream to include promoter regions in the overlap analysis. Details on intersections of multiple genomic regions in Material and methods 6.1.9.

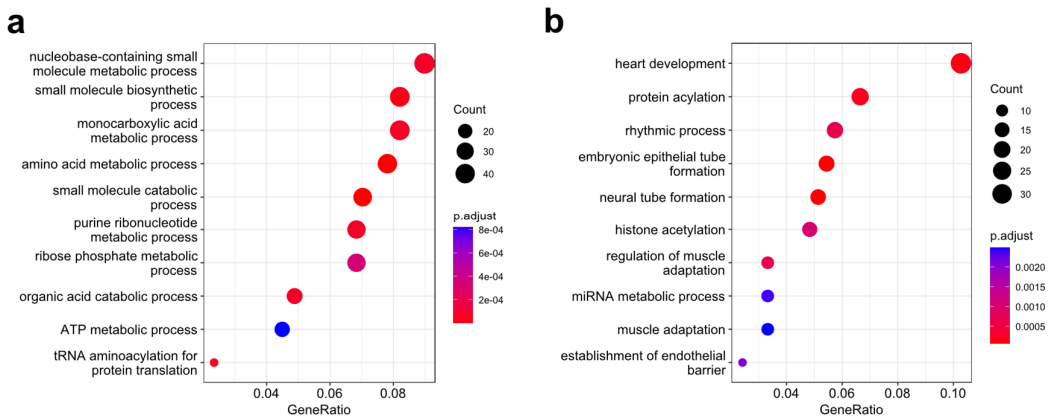


Figure 4.20: Association of increased TET1-binding with altered gene expression in *Gadd45* TKOs. GO enrichment analysis results for *Gadd45* TKO TET1 up peaks intersecting with *Gadd45* TKO (a) up and (b) down DEGs.

Gadd45 TKO up and down DEGs and TET1 up peaks. The large majority of DEGs did not directly intersect with hypo- or hypermethylated promoters (Figure 4.21a,b) or gene bodies (Figure 4.21c,d) in *Gadd45* TKOs. This suggests that shifts in DNA methylation on DEG loci likely have only a limited impact on their gene expression changes. Nevertheless, within the subset of genes that did intersect with altered methylation, up DEGs (Figure 4.21a,c) were more closely associated with increased TET1-binding and hypomethylation than down DEGs (Figure 4.21b,d). The observation of hypomethylation at genes with increased transcription and TET1 occupancy is consistent with TET1's role as a demethylase involved in gene activation. Conversely, genes with decreased transcription and increased TET1 binding showed a less pronounced intersection with hypomethylated regions. This is in line with TET1's gene-repressive role that functions independent of changes in DNA methylation. In summary, while differential TET1-binding may contribute to gene regulatory differences observed in *Gadd45* TKOs, its impact likely occurs largely independent of altered DNA methylation at these specific loci.

4.2.7 *Gadd45* TKO mESCs show dysregulated enhancer epigenomes

This laboratory has previously shown GADD45a to be involved in the demethylation of enhancers in mouse embryonic fibroblast (MEF) cells (Schäfer et al., 2018). H3K4me1 and H3K27ac are known as enhancer-defining histone marks (Cheng et al., 2014; S. Fu et al., 2018). H3K4me1 marks both active and poised enhancers

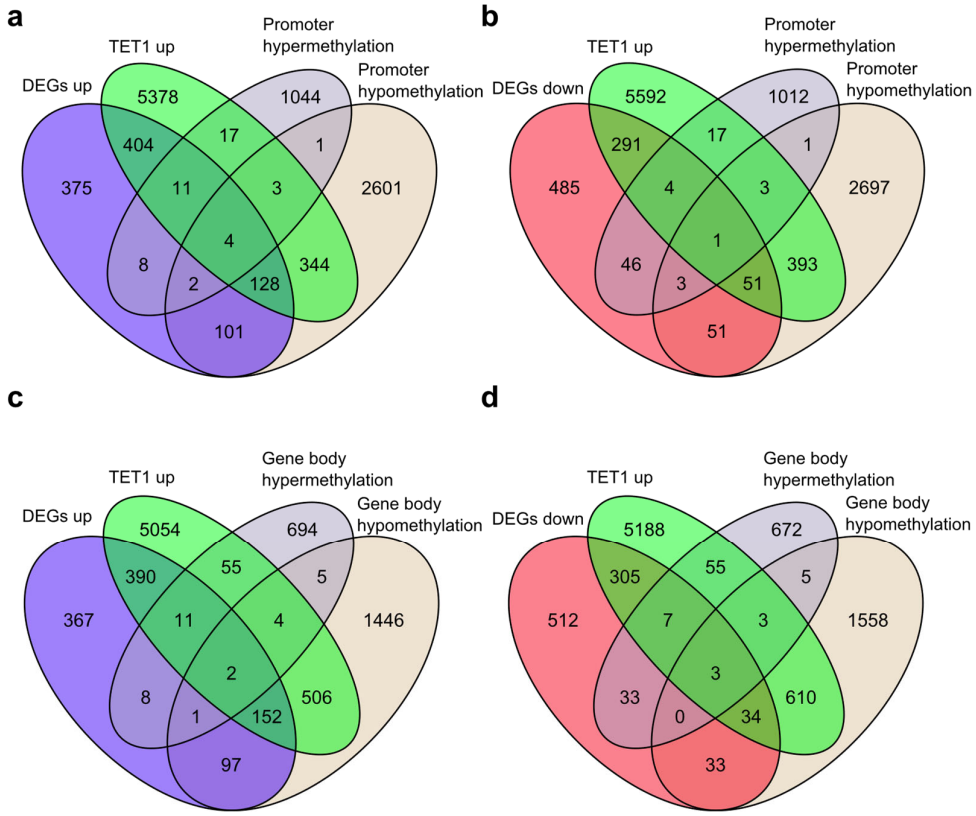


Figure 4.21: Intersection of *Gadd45* TKO DEGs with TET1 up peaks and regions of differential methylation. Intersection of Gadd45 TKO **(a)** up and **(b)** down DEGs with TET1 up peaks, hypermethylated and hypomethylated promoters. Intersection of *Gadd45* TKO **(c)** up and **(d)** down DEGs with TET1 up peaks, hypermethylated and hypomethylated gene bodies. Differentially methylated promoters and gene bodies were obtained by Yulia Kargapolova via Infinium Mouse Methylation BeadChip. Details on intersections of multiple genomic regions in Material and methods 6.1.9.

(Kubo et al., 2024), whereas H3K27ac distinguishes active from poised enhancers by its exclusive presence in the former (Creyghton et al., 2010). To investigate a possible role of GADD45 family proteins in writing or erasure of histone marks with functions in active enhancers, CUT&Tag experiments for H3K4me1 and H3K27ac were performed in WT and *Gadd45* TKO mESCs by Rintu Umesh.

I called a total of 20,655 H3K4me1 consensus peaks in WT mESCs. Despite being known as a classical enhancer mark, most H3K4me1 peaks in the WT consensus peaks localized to promoters, with distal intergenic regions being only the second most common genomic feature (Figure 4.22a,b). Enhancers are typically in intergenic regions or intronic (Panigrahi & O'Malley, 2021). Differential comparison revealed 2,341 peaks with significantly increased and 2,817 peaks with significantly decreased H3K4me1 occupancy in *Gadd45* TKO mESCs (FDR < 0.05). While the

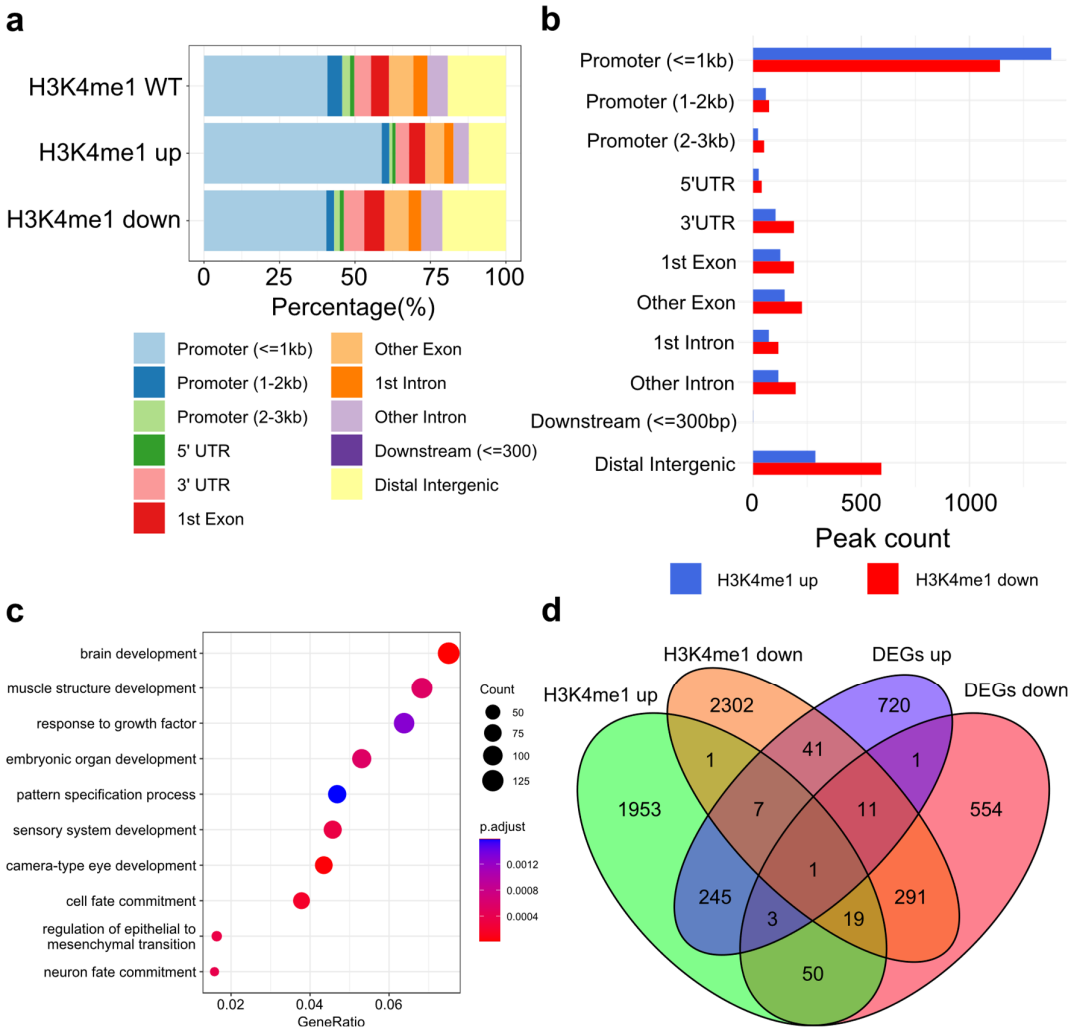


Figure 4.22: H3K4me1 CUT&Tag analysis of Gadd45 TKO mESCs. (a) Relative genomic feature distribution of WT consensus H3K4me1 peaks (H3K4me1 WT) for reference and *Gadd45* TKO up (H3K4me1 up) and down regulated (H3K4me1 down) peaks. **(b)** Absolute genomic feature distribution for up and down H3K4me1 peaks. **(c)** GO enrichment analysis of up regulated H3K4me1 peaks. **(d)** Intersection of differential H3K4me1 peaks with DEGs in Gadd45 TKO mESCs. DEGs were extended 3kb upstream to include promoter regions in the overlap analysis. Details on intersections of multiple genomic regions in Material and methods 6.1.9.

feature distribution of peaks with decreased H3K4me1 occupancy was comparable to that of WT consensus peaks, H3K4me1 up peaks tended to localize rather to promoter regions (Figure 4.22a). GO enrichment analysis of nearby genes did not reveal any enriched terms for decreased H3K4me1 peaks, but up H3K4me1 peaks showed an enrichment for development-related terms (Figure 4.22c). *Gadd45* TKO up DEGs intersected well with increased H3K4me1 occupancy, whereas down DEGs intersected with decreased H3K4me1 occupancy (Figure 4.22d).

For H3K27ac CUT&Tag, I called 12,423 consensus peaks in WT mESCs. 2,650 peaks had significantly increased and 1,221 peaks had significantly decreased H3K27ac occupancy in *Gadd45* TKO mESCs (FDR < 0.05). Similar to the H3K4me1 CUT&Tag experiment, most WT H3K27ac as well as differential H3K27ac peaks in Gadd45 TKO mESCs localized to promoters or distal intergenic regions (Figure 4.23a,b). Peaks with increased H3K27ac occupancy were enriched in blood vessel and prostate gland related terms (Figure 4.23c), while GO enrichment analysis in H3K27ac peaks with decreased occupancy did not result in enriched GO terms. Genes with increased expression in *Gadd45* TKO mESCs intersected well with H3K27ac up peaks, and genes with decreased expression had a high overlap with H3K27ac down peaks (Figure 4.23d).

To specifically test whether active enhancers are affected by differential DNA methylation in *Gadd45* TKO mESCs, I obtained a dataset in which active mESC enhancers were mapped genome-wide by formaldehyde-assisted isolation of regulatory elements coupled with self-transcribing active regulatory region sequencing (FAIRE-STARR-seq) (Glaser et al., 2021). Intersection of this active enhancer dataset with differentially methylated tiles from the Infinium Mouse Methylation BeadChip indicated that more than 10% of active enhancer regions in Gadd45 TKOs are hypomethylated (Figure 4.24b), whereas an intersection with hypermethylated tiles was observed to a lesser extent (Figure 4.24a). Heatmaps of the TET1, H3K4me1, H3K27ac CUT&Tag and spKAS-seq data in Gadd45 TKO mESC centered to the active enhancers showed an increased TET1, spKAS-seq and H3K4me1 and a decreased H3K27ac signal at these enhancers (Figure 4.25a). Occupancy signal profiles of the same datasets indicated weakly elevated TET1 and spKAS-seq signal at active enhancer regions with unchanged or decreased DNA methylation levels (Figure 4.25b). Moreover, H3K4me1 levels were globally increased at active enhancers, irrespective of DNA methylation changes (Figure 4.25b). H3K27ac levels were decreased at hypermethylated and unchanged active enhancers (Figure 4.25b). These data suggest that Gadd45 proteins may play a broader role in shaping the mESC enhancer chromatin landscape, extending beyond their previously described function in DNA demethylation of enhancers (Schüle et al., 2019).

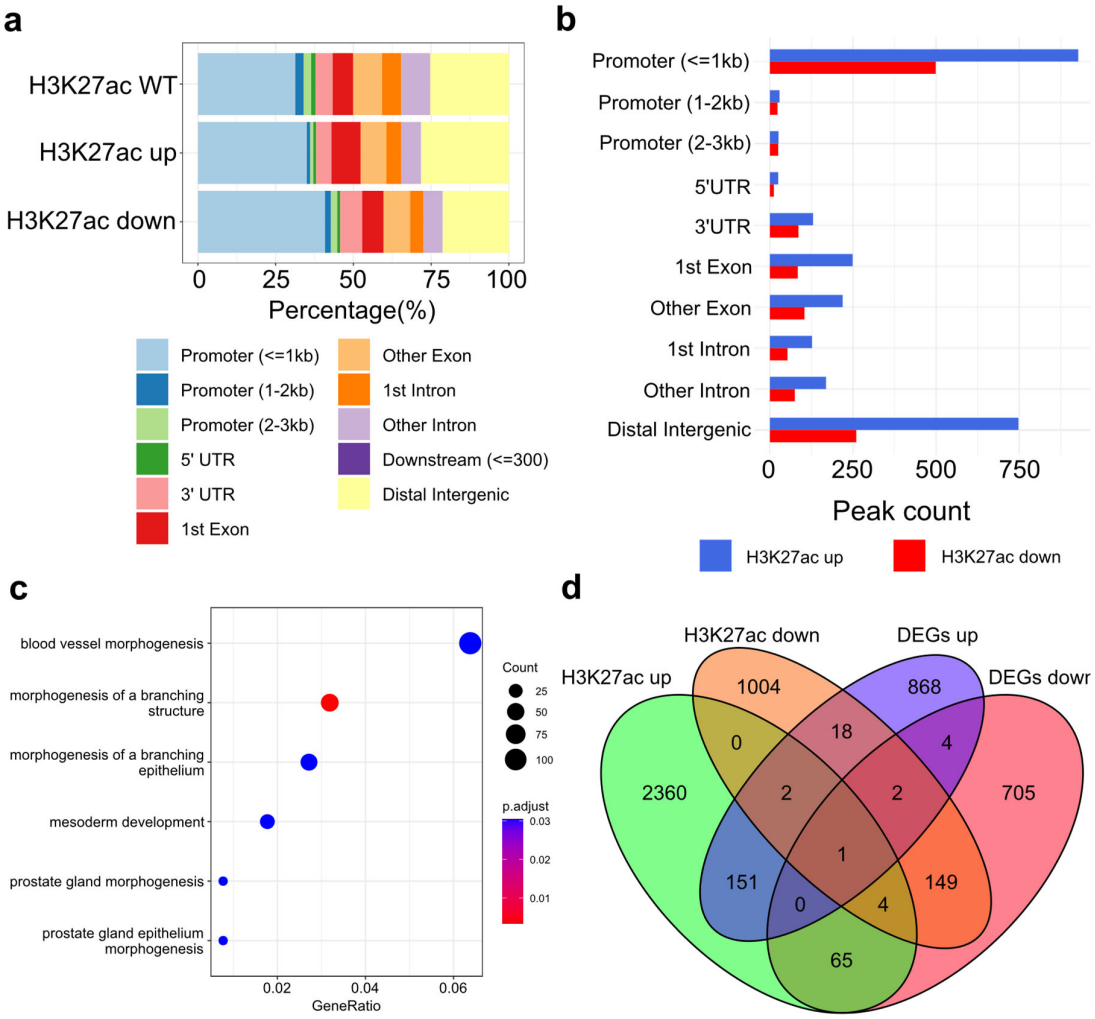


Figure 4.23: H3K27ac CUT&Tag analysis of *Gadd45* TKO mESCs. (a) Relative genomic feature distribution of WT consensus H3K27ac peaks (H3K27ac WT) for reference and *Gadd45* TKO up (H3K27ac up) and down regulated (H3K27ac down) peaks. **(b)** Absolute genomic feature distribution for up and down H3K27ac peaks. **(c)** GO enrichment analysis of up regulated H3K27ac peaks. **(d)** Intersection of differential H3K27ac peaks with DEGs in *Gadd45* TKO mESCs. DEGs were extended 3kb upstream to include promoter regions in the overlap analysis. Details on intersections of multiple genomic regions in Material and methods 6.1.9.

Taken together, these results indicate weakly elevated TET1 binding and hypomethylation at active enhancers in *Gadd45* TKOs. It remains unclear whether the observed mild increase in TET1-binding to demethylated active enhancers is sufficient to be the cause for their demethylation. Furthermore, *Gadd45* TKO mESCs exhibit globally increased H3K4me1 levels and decreased H3K27ac levels at active enhancers, with the latter predominantly occurring at hypermethylated active enhancers or active enhancers with unchanged DNA methylation. It remains

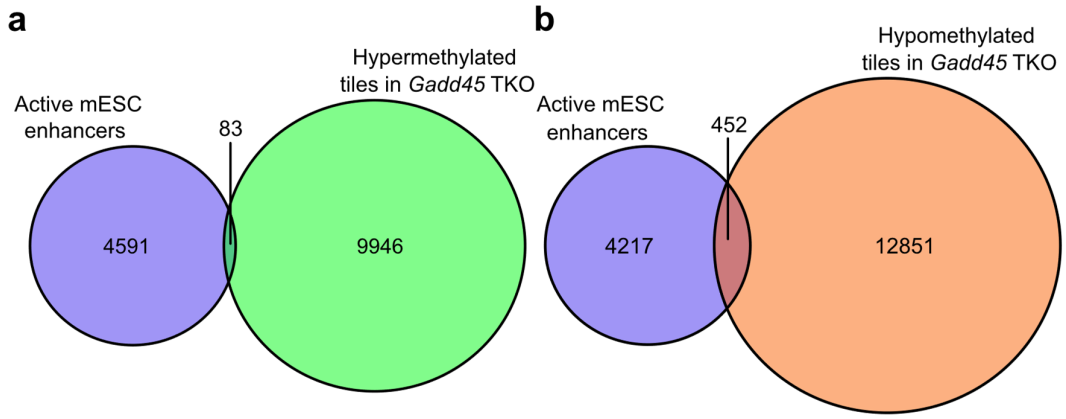


Figure 4.24: Active enhancers have increased hypomethylation in *Gadd45* TKO mESCs Active enhancer intersection with differential **(a)** hyper and **(b)** hypomethylated tiles in *Gadd45* TKO mESCs. Active enhancer regions were obtained from a published dataset, where Glaser et al. (2021) mapped active mESC enhancers by FAIRE-STARR-seq. Differentially methylated tiles were obtained by Yulia Kargapolova via Infinium Mouse Methylation BeadChip.

unclear whether the increased H3K4me1 levels, regardless of methylation status, and the decreased H3K27ac levels at hypermethylated or unchanged active enhancers are a direct result of a gene regulatory role of GADD45 proteins. However, the site-specific nature of the latter makes it more likely to be associated with such a role.

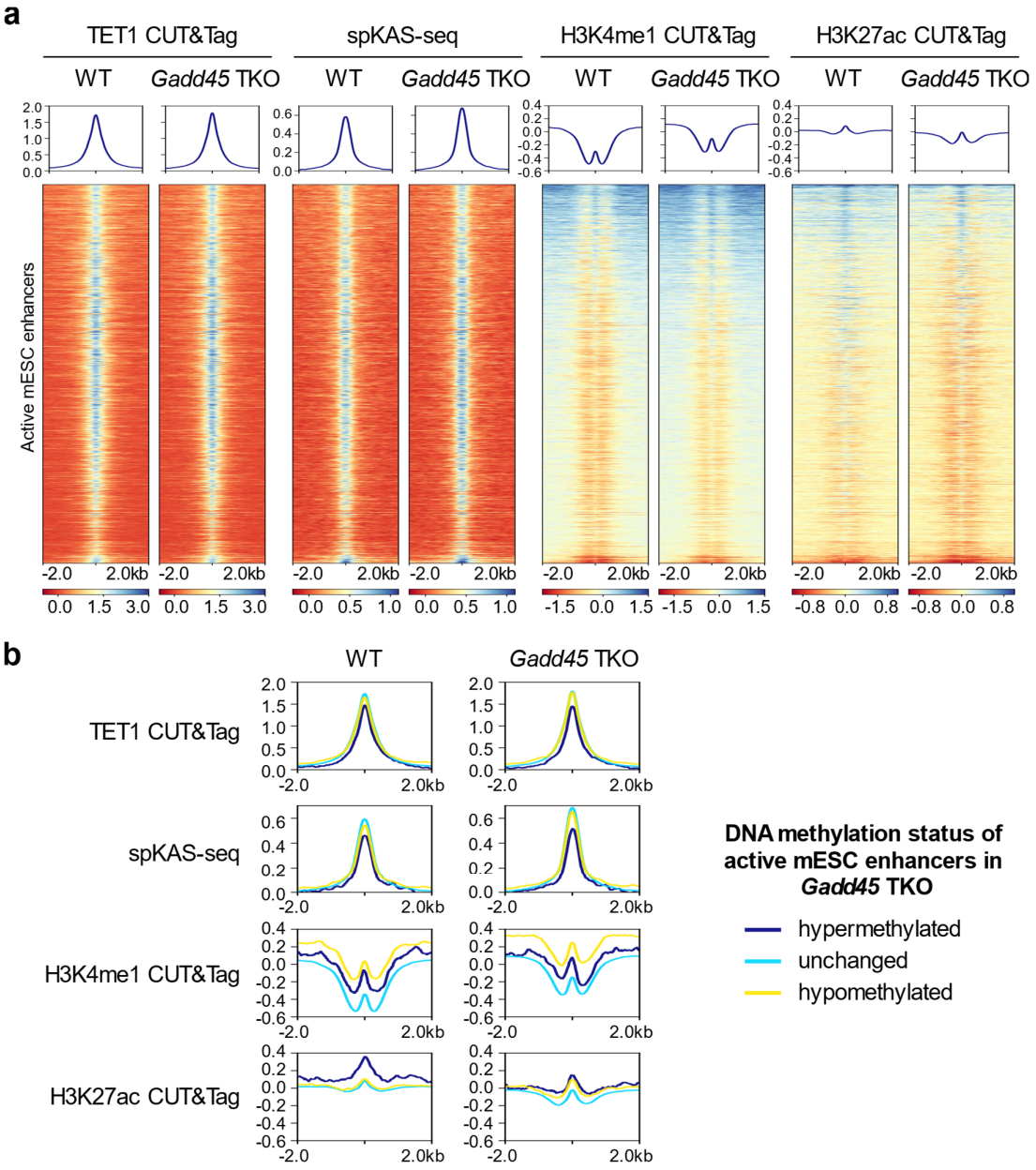


Figure 4.25: Enhancer chromatin is altered in *Gadd45* TKO mESCs. (a) Mean normalized signal with genomic heatmap for TET1, H3K4me1 and H3K27ac CUT&Tag signal for WT and *Gadd45* TKO mESCs at active enhancer loci. **(b)** Mean normalized signal for TET1, H3K4me1 and H3K27ac CUT&Tag signal for WT and *Gadd45* TKO at active enhancer loci that are hypermethylated, unchanged or hypomethylated in *Gadd45* TKOs. Active enhancers regions were obtained from a published dataset, where Glaser et al. (2021) mapped active mESC enhancers by FAIRE-STARR-seq. Enhancer methylation status in *Gadd45* TKOs was obtained by intersection with Infinium Mouse Methylation BeadChip data by Yulia Kargapolova.

4.3 Discussion

4.3.1 Comparison of R-loop mapping methods

To make an informed decision for an R-loop mapping method to be used to compare R-loop levels in *Gadd45* TKO versus WT mESCs, I first compared different stranded and non-stranded R-loop mapping methods with each other.

The 47,354 and 36,919 peaks I called for S9.6 CUT&Tag and HBD CUT&Tag, respectively are in the range of typically obtained peak numbers for R-loop mapping methods which lie roughly between 14,000 (Chen et al., 2019) to 70,000 (Sanz et al., 2016). Meanwhile, the 12,492 ecRNH CUT&Tag peaks I called are slightly below the lower end of that range. The large intersection between peaks called for the CUT&Tag-based methods and their lower intersection with the S9.6 antibody-based DRIP-seq peaks hinted at a strong bias introduced by usage of the CUT&Tag protocol rather than the antibody that determines its target. Moreover, the peak distribution plot showed a general promoter bias of all CUT&Tag based methods, whereas DRIP-seq also generates a substantial number of peaks that localize to genes' 3' regions. Since both, S9.6 CUT&Tag as well as DRIP-seq rely on usage of the S9.6 antibody, I expected that out of the three CUT&Tag based methods, results generated from S9.6 CUT&Tag would have the highest similarity to DRIP-seq. Interestingly, among all methods tested, the HBD CUT&Tag protocol demonstrated the largest overlap with DRIP-seq in terms of absolute peak numbers. The present experiment was performed without replicates and samples might be subject to fluctuations in quality of the results that cannot be judged with confidence without repeating the experiments. Nevertheless, a possible explanation is that out of S9.6 antibody, HBD and ecRNH, HBD best captures R-loops in genic 3' regions, indicated by a slight increase in peak numbers in genic 3' regions visible in Figure 4.8b These may intersect with genic 3' peaks captured by DRIP-seq. At the same time, this 3' bias is not reflected in the feature distribution analyses (Figure 4.8c,d), possibly since 3' R-loops of one gene can overlap with promoter or 5' R-loops of neighboring genes. As peaks are exclusively assigned to a single feature category and promoters or 5' regions are prioritized in such assignments, the contribution of 3' R-loops may be underrepresented in the distribution plots.

Next, I compared the strand-specific R-loop mapping methods spKAS-seq, stranded DRIP-seq (strDRIP-seq) with sonication and nuclease S1 treatment (Soni./S1) and strDRIP-seq with restriction enzyme digestion and nuclease S1 treatment (RE/S1). The 19,967 spKAS-seq peaks as well as the 39,880 strDRIP-seq (RE/S1) peaks I called are in the expected range of 14,000 (Chen et al., 2019) to 70,000 (Sanz et al., 2016) peaks, while the 100,703 peaks I obtained for strDRIP-seq (Soni./S1) are above that range, hinting at possible false positive peaks. strDRIP-seq (Soni./S1) had the clearest strand-separation of the three methods. However, due to the fact that called peaks often covered whole genes from start to end and mean peak sizes were roughly 4-fold higher than the mean peak sizes of the other methods, I reasoned that the method may be detecting short-lived R-loops that are a byproduct of the transcription process. While spKAS-seq and strDRIP-seq (RE/S1) mostly produced non-intersecting peaks, feature distribution indicated that spKAS-seq peaks mostly localized to promoters, while strDRIP-seq (RE/S1) largely localized to genic regions.

In summary, the choice of an R-loop mapping method considerably impacts the information provided when conducting such experiments. Differences between the methods particularly affect information on the strandedness of an R-loop as well as the ability of mapping R-loops in different genomic features. Such differences in R-loop mapping capabilities for R-loops residing in different genomic features have been observed in previous studies (Chédin et al., 2021). GADD45a was previously shown to bind an antisense promoter R-loop and positively regulate transcription of a nearby gene after inducing DNA demethylation (Arab et al., 2019). While thus far unreported in literature, such a process may also impact the transcription of the RNA forming the GADD45a-bound R-loop and thus impact R-loop levels. In this case, it is crucial to obtain information on R-loop strandedness, as GADD45a may specifically bind promoter antisense R-loops, which could be challenging to differentiate from nearby sense promoter R-loops in non-stranded R-loop mapping methods, where they may appear as a single peak. Given the advantage of obtaining strand-specific R-loop information, the observation that spKAS-seq effectively captures promoter-associated R-loops, it was decided to proceed with spKAS-seq for a WT versus *Gadd45* TKO comparison.

4.3.2 Selection of *Gadd45* TKO mutant mESC clones for experiments

Karyotyping analysis in *Gadd45* TKO clone set 2 showed that all *Gadd45* TKO clones were affected by chromosome 8 trisomy. All of these clones were generated from a single parental *Gadd45g* SKO clone that was previously picked and likely propagated the trisomy to all *Gadd45* TKO clones. *Gadd45* TKO clone set 3 WT and *Gadd45* TKO clones were generated from clones of two different parental origins, respectively. All WT clones generated from parental clone WT_Par7 were affected by chromosome 14-loss and thus removed, leaving four clones with parental WT_Par2 origin: WT1_Par2, WT3_Par2, WT4_Par2, WT5_Par2. For *Gadd45* TKO genotype, two out of three clones with ParSKO19 parental origin had to be removed due to visible *Gadd45a* expression and visual abnormalities. For *Gadd45* TKO clones of ParSKO50 origin, one clone was removed due to abnormal *Gadd45b* 3'UTR up regulation Moreover, all *Gadd45* TKO clones made from ParSKO50 turned out to suffer from chromosome Y-loss. Figure 4.26 illustrates the structure and location of protein coding genes on the mouse Y chromosome with indicates parts lost in *Gadd45* TKO ParSKO50 clones. Most protein coding genes, including the "sex switch" gene *Sry* are located on the short arm of chromosome Y. WGS data indicated that the short arm and the pseudoautosomal region (PAR) were not affected by the chromosome Y loss. Interestingly, both of these two regions are known to be potential crossover recombination regions i.e. capable of crossover with the X chromosome (Decarpentrie et al., 2016). Thus, an unsuccessful crossover event might have led to the loss of the remaining chromosome Y regions.

Sry holds particular significance as it is considered to be the defining gene of the Y chromosome (Bilen et al., 2013; Westemeier-Rice et al., 2024) and determinant of male somatic development (Koopman et al., 2016). *Sry* is the only Y chromosome gene necessary and on its own sufficient for healthy male somatic development (Koopman et al., 2016). Ensuring the presence of *Sry* is thus critical to avoid potential confounding effects during cardiac differentiation experiments. This is particularly important as the loss of chromosome Y has been implicated in cardiovascular diseases and heart failure (Sano et al., 2022). While Sano et al. (2022) have not linked these effects to particular genes on chromosome Y, these effects are unlikely to be caused by genes contained within the region lost in

ParSKO50 clones, as these genes are thus far only reported to function in spermatogenesis (Subrini & Turner, 2021). After confirmation of the presence of the important region containing the Sry gene in the three *Gadd45* TKO clones with ParSKO50 origin and due to the availability of one *Gadd45* TKO clone originating from ParSKO19 without the chromosomal abnormality, it was decided to proceed with experiments using *Gadd45* TKO clones TKO271_ParSKO19, TKO81_ParSKO50, TKO83_ParSKO50 and TKO330_ParSKO50.

A publication investigating publicly available mESC RNA-seq datasets lists for common chromosomal aberrations found four aberrations to be common in mESC cultures: chromosome 8 and chromosome 11 trisomy as well as loss of chromosome 10qB and chromosome 14qC-14qE (Ben-David & Benvenisty, 2012). Interestingly, I detected two out of these among the analyzed clones. The pervasiveness of such chromosomal abnormalities among the analyzed clones highlights the chromosomal fragility of mESC cell lines and the importance of karyotyping analyses in cell lines. Their oversight can influence comparisons, such as WT versus mutant analyses as a confounding effect and potentially lead to incorrect conclusions.

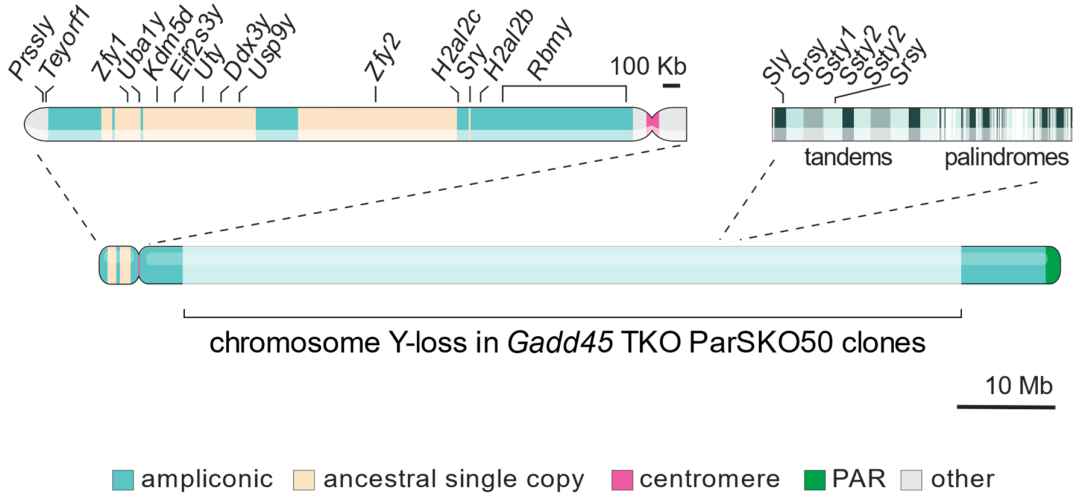


Figure 4.26: Structure and protein coding genes of the mouse Y chromosome. The specific chromosome Y region undergoing partial loss in ParSKO50 *Gadd45* TKO clones is indicated. Figure adapted from (Subrini & Turner, 2021) licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

4.3.3 Transcriptional dysregulation in *Gadd45* TKO mESCs is associated with cardiac differentiation defects

Previous *Gadd45* knockout clones established in our laboratory from *Gadd45* TKO clone set 1 were only found to contain 38 up and 97 down DEGs, respectively (Schüle et al., 2019). These mutants were later suspected to be hypomorphs (internal lab communication), which was likely the cause for the relatively low number of affected genes. In contrast, the newly generated *Gadd45* TKO mutants in Gadd45 TKO clone set 3 displayed substantial perturbations in gene expression, with 1,056 genes up regulated and 936 down regulated.

In the *Gadd45* TKO mutants from *Gadd45* TKO clone set 3, up regulated genes were notably enriched in pathways related to mitochondrial energy metabolism and neurodegenerative diseases, such as Parkinson's and Alzheimer's disease. GADD45 family proteins have previously been implicated in neurological disorders (Huang et al., 2024), including the neurodegenerative diseases Parkinson's (Yang et al., 2016) and Alzheimer's disease (Torp et al., 1998). Down regulated genes were associated with processes in heart development, epithelial cell differentiation, and the MAPK, FoxO, and p53 signaling pathways. These three signaling pathways are well-known for their interactions with GADD45 family proteins (Carter & Brunet, 2007; Jung et al., 2013; Yin et al., 2004). Moreover, GADD45b has been shown to negatively regulate JNK pathway-induced cardiac hypertrophy (Wang et al., 2008), while GADD45g is a suspected player in outflow tract development (Zhang et al., 2024) and known to induce cardiomyocyte death in the left ventricular reverse remodeling process (Iwahana et al., 2020; Lucas et al., 2015). Left ventricular reverse remodeling is a therapeutic goal for patients treated for cardiomyopathy (Verdonschot et al., 2018). The consistency between overrepresented functions of DEGs in Gadd45 TKO mutants and established literature strengthens the validity of these knockouts as a reliable model for studying the function of GADD45 family proteins. Interestingly, the volcano plot in Figure 4.14a showed the gene *Sfmbt2* to be the second strongest deregulated gene by FDR after Gadd45a itself. The *Sfmbt2* gene encodes a polycomb group protein and contains one of the largest microRNA (miRNA) clusters in the rodent genome (Inoue et al., 2017). miRNAs miR-669a and miR-669q were previously shown to prevent skeletal muscle differentiation in postnatal cardiac progenitor cells (Crippa et al., 2011). The observed enrichment

for heart development-related genes among the down DEGs as well as the cardiac differentiation defects observed *in vitro* may thus be a downstream effect of transcriptional down regulation of the miRNA cluster in *Sfmbt2*, resulting in aberrant skeletal muscle differentiation.

It should also be noted that three out of four *Gadd45* TKO mutants, all derived from the same parental clone, displayed a loss of chromosome Y, one of the four most common chromosomal aberrations in mESC cultures. Given that this loss occurred only in clones originating from the same parental source, it is likely unrelated to the absence of GADD45 family proteins. Despite confirming that the most relevant regions of chromosome Y are intact, it cannot be fully ruled out that this loss generally impacted the differential expression analysis between WT and Gadd45 TKO mESCs. Notably, a link has been established between hematopoietic Y chromosome-loss in men and an increased risk of cardiac fibrosis and heart failure (Sano et al., 2022). Nevertheless, visual confirmation of deregulated heart development genes along with consistent defects in cardiac differentiation in all *Gadd45* TKO clones, suggest that these phenotypes are not related to the observed chromosome Y-loss.

4.3.4 Epigenomic profiling of *Gadd45* TKO mESCs

In collaboration with Gaurav Joshi and Khelifa Arab, I attempted to map GADD45a's mESC chromatin binding sites through multiple ChIP-seq and CUT&Tag experiments, all of which were thus far unsuccessful (not shown). This can indicate a weak or transient binding of GADD45a, a localization to chromatin regions that are difficult to capture or no direct chromatin binding of GADD45a in mESCs. Epigenomic profiling sequencing experiments indicated that GADD45 TKO mESC chromatin has considerably differential open chromatin and R-loop levels if compared with WT clones. I could, however, not directly link a majority of these chromatin changes to differential gene expression by intersection analysis.

While this laboratory was thus far unable to show a direct interaction of GADD45 family proteins and TET1 in mESCs (not shown), sequencing analyses indicated a substantial overlap of sites with increased TET1-binding with both, up and down

DEGs in Gadd45 TKOs. I observed more than 6,000 sites with increased TET1-binding that largely occupy promoters. Intersection between these sites and *Gadd45* TKO DEGs was substantial and a GO enrichment analysis on intersecting genes resulted in similar GO terms as the GO enrichment analysis for the DEGs alone. The substantial overlap may hint at a regulatory role for TET1, though it might also be a consequence of the overall high prevalence of sites with increased TET1 occupancy. Nevertheless, the substantial increase in TET1 binding observed, coupled with the relatively few lost TET1 peaks in *Gadd45* TKO mESCs, argues against a model in which GADD45 primarily functions to recruit TET1 as proposed in Arab et al. (2019).

The chromatin modifications H3K4me1 and H3K27ac, which are known as classical enhancer modifications showed substantial perturbations in *Gadd45* TKO mESCs. In my analysis, most of these marks localized to promoters. Despite being known as active enhancer-specific or even active enhancer-defining, these marks have been reported to also occur in active promoters (Cheng et al., 2014; S. Fu et al., 2018). Moreover, it was previously reported that immunoprecipitation experiments co-immunoprecipitate interacting chromatin regions, such as enhancers and promoters, and can thus result in peaks associated with both, enhancers and their target promoter regions (Ibn-Salem & Andrade-Navarro, 2019). Likewise, the experiments conducted in this study may have been influenced by a similar phenomenon. For both of these activating histone modifications, their increased presence in *Gadd45* TKO mESCs substantially coincided with increased gene expression and their decreased presence coincided with down regulation of genes. Differential presence of these marks either in promoters, or incorrectly associated with promoters due to proximity with the enhancers regulating them, may thus partially be responsible for differential regulation of genes in *Gadd45* TKOs. An in depth-investigation of active mESC enhancers indicated increased TET1 and R-loop levels in *Gadd45* TKOs, specifically at enhancers with decreased methylation. This may suggest that, as a consequence of increased TET-binding, enhancers are hypomethylated, resulting in increased transcription of eRNAs at enhancer regions. Both, H3K4me1 and H3K27ac, had substantial differential active enhancer occupancy in *Gadd45* TKOs. While H3K4me1 occupancy was globally increased at active enhancers, a decreased H3K27ac occupancy was predominantly observed

at hypermethylated active enhancers or active enhancers with unchanged DNA methylation. The globally elevated H3K4me1 levels at active enhancer loci in Gadd45 TKO mESCs suggest that site-specific regulation of this histone mark through differential GADD45-binding is unlikely. The decreased H3K27ac levels observed specifically at hypermethylated or unchanged active enhancers, on the other hand, are due to their more site-specific nature, more likely to be associated with such a role of GADD45 family proteins. In summary, GADD45 proteins may regulate enhancer chromatin by different mechanisms: (1) by inhibition of TET1-binding and thus prevention of demethylation and R-loop formation and (2) by facilitating acetylation of H3K27 and negatively regulating DNA methylation.

4.3.5 Towards a model for GADD45-mediated gene regulation

The currently available data is insufficient to confidently support a distinct model of GADD45-mediated gene regulation. However, taken together, the sequencing experiment results suggest a function of one or more GADD45 family proteins in modulating enhancer chromatin. In Schäfer et al. (2018), this laboratory reported a role of GADD45a in the direct binding to and demethylation of C/EBP binding-dependent enhancers in mouse embryonic fibroblasts (MEFs). Similarly, the data analyzed in this thesis suggests a role for GADD45 in the modulation chromatin marks in mESC enhancers (Figure 4.27). Belonging to a ribosomal protein group, GADD45 proteins are known to be RNA-binding (Sytnikova et al., 2011) and may thus bind the eRNAs transcribed from enhancers. Moreover, a potential role of GADD45a involving its direct binding to eRNAs, where it is embedded within a dense assembly of proteins positioned between promoters and enhancers, may account for the difficulty in detecting GADD45a in protein-chromatin interaction-mapping methods. After eRNA-binding, GADD45 may inhibit R-loop formation of eRNAs and directly, or indirectly via cofactors, inhibit gene repressive TET1-binding and lead to activation of heart development genes. Meanwhile, in *Gadd45* TKO cells, due to the lack of GADD45's presence in enhancers, TET-binding is not inhibited, leading to excessive TET1-binding. TET1 may then demethylate enhancers, leading to increased enhancer R-loop-formation and repression of heart development genes.

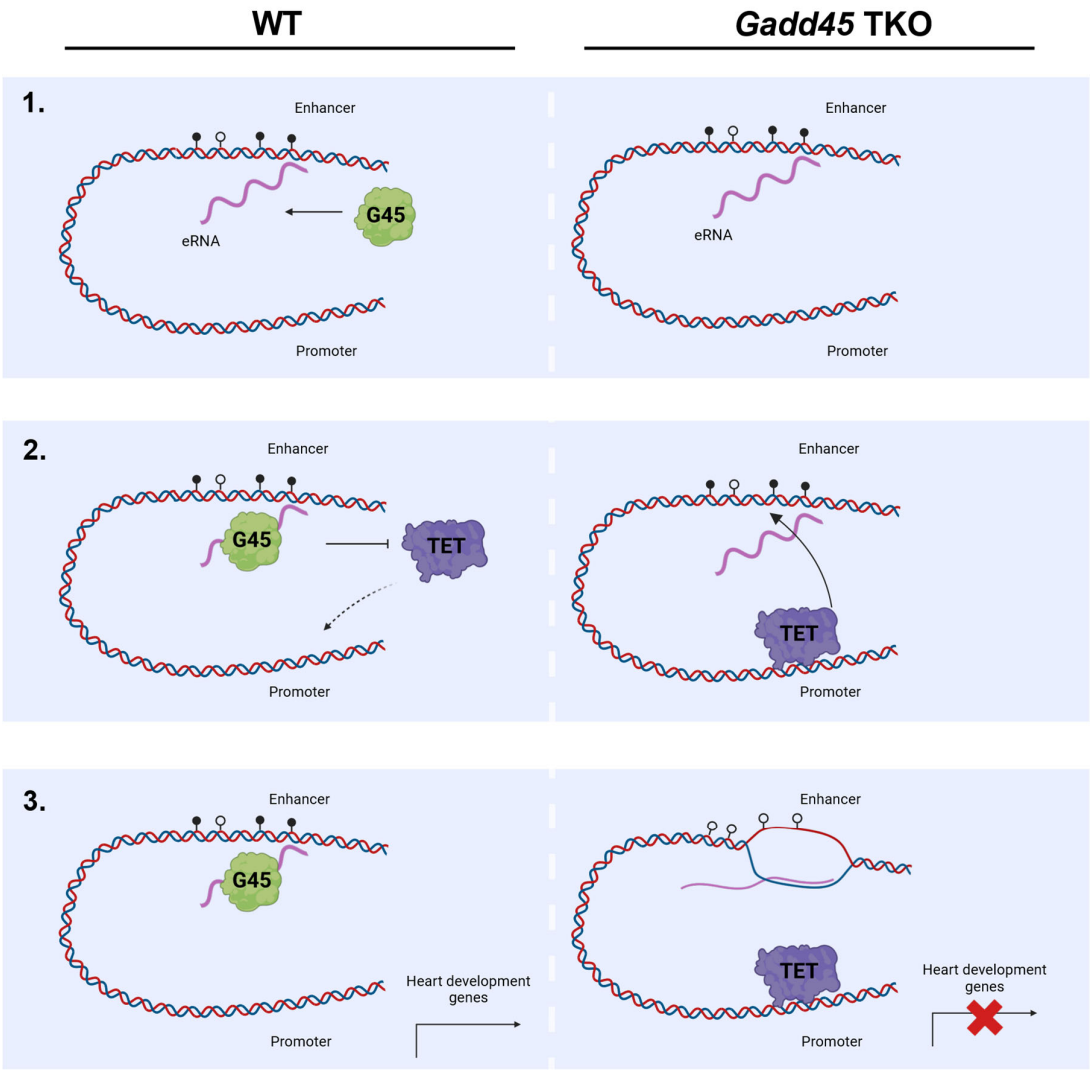


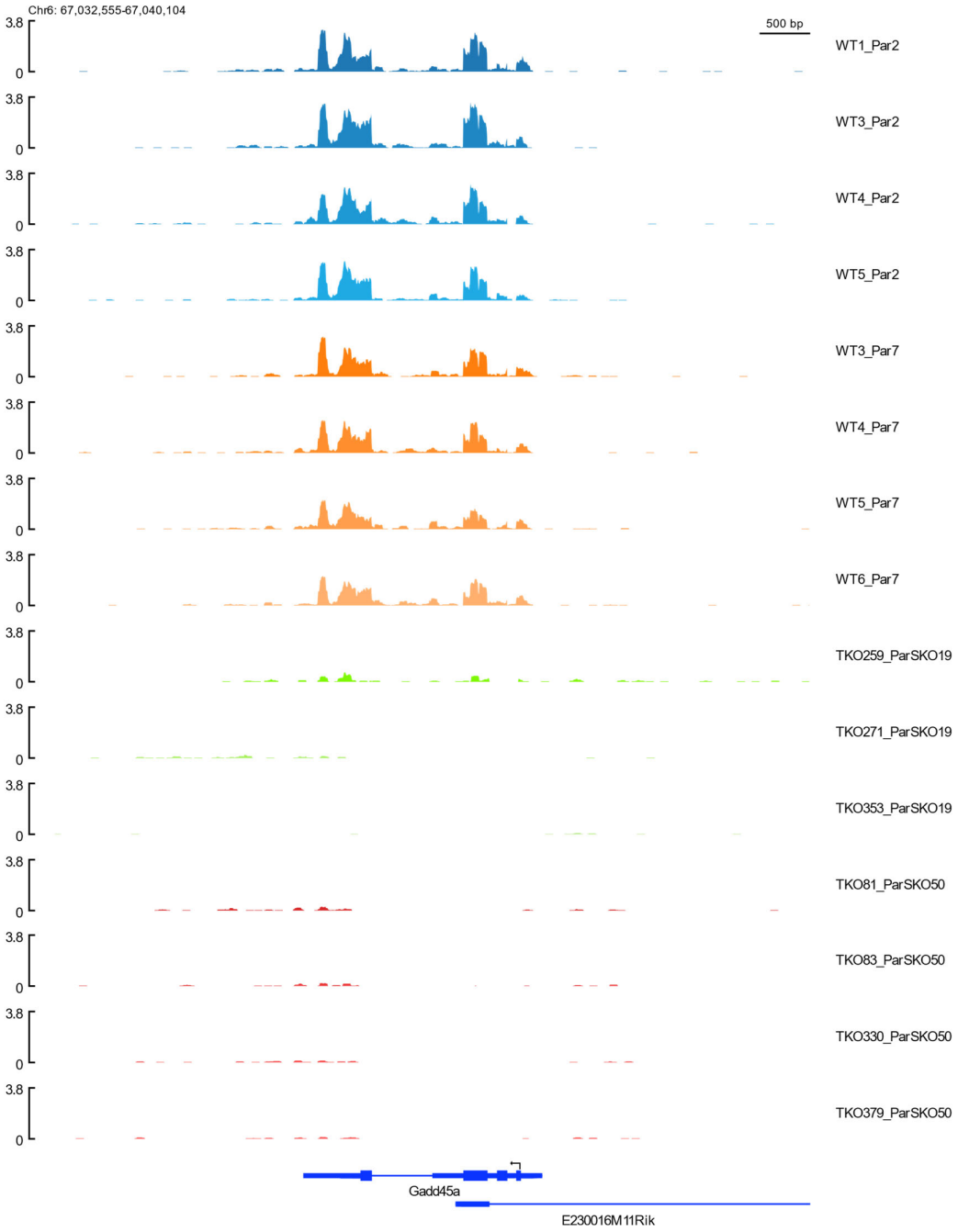
Figure 4.27: Model for the GADD45-mediated regulation of heart development gene expression. (Left) In WT cells, (1) GADD45 binds to enhancers in eRNAs, (2) inhibits promoter-binding of TET1, possibly via co-factors, and (3) keeps heart development genes active. (Right) In *Gadd45* TKO mESCs, (1) enhancers are not GADD45a-bound, thus (2) TET1 can bind to interacting promoters and (3) heart development genes are repressed due to excessive R-loop formation by eRNAs. Created with BioRender.com.

Of note, this model does not explain the observed decrease in H3K27ac at hypermethylated *Gadd45* TKO mESC enhancers. This observation may be explained by an alternative mechanism, in which GADD45 recruits C/EBP to a different subset of enhancers, thereby increasing H3K27ac levels and activating genes. In *Gadd45* TKOs, C/EBP may not be recruited to these enhancers,

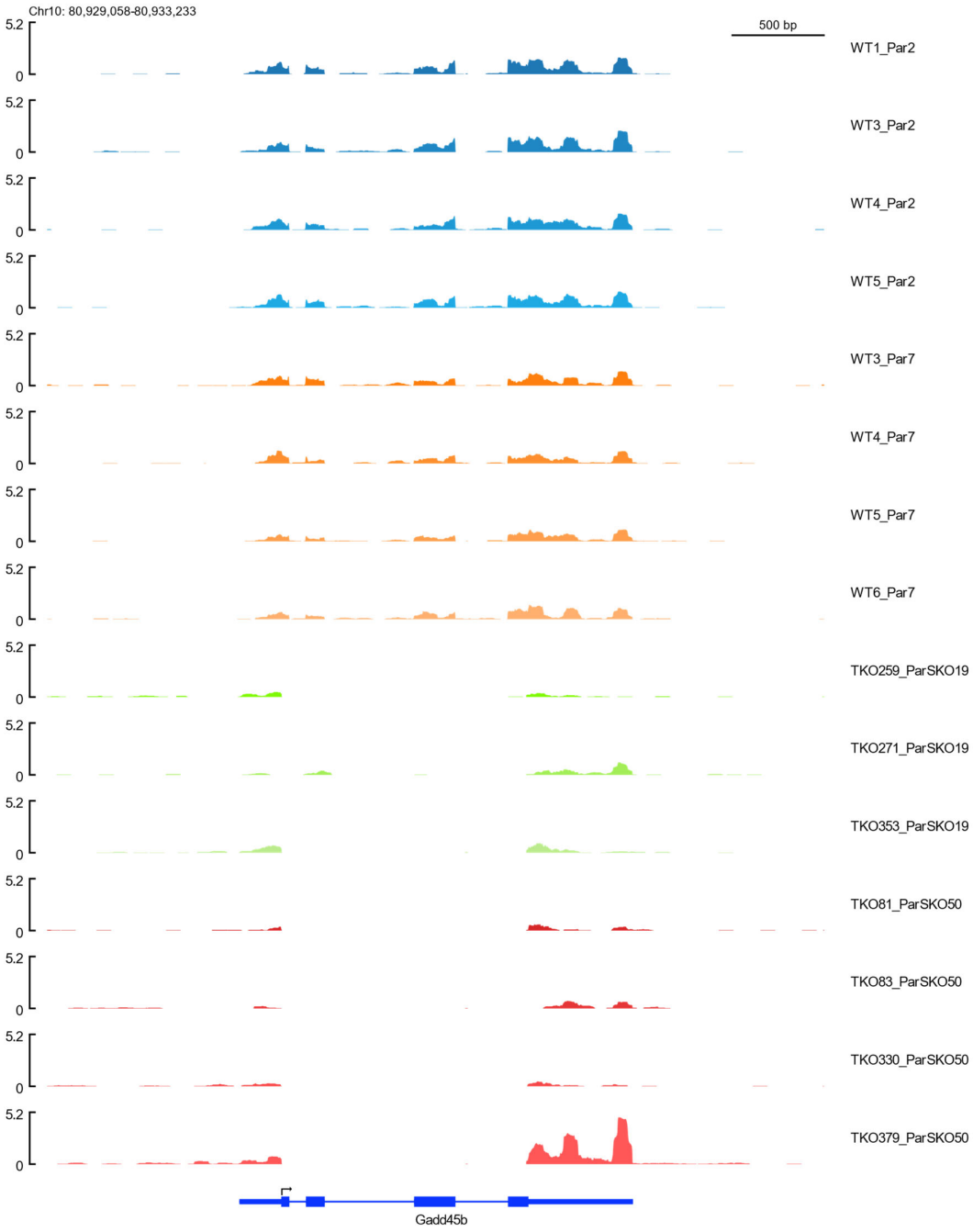
preventing their H3K27ac-modification. As a result, gene activation is impaired, and the associated genes fail to be properly expressed.

Several sequencing methods could be employed to test predictions of these models on a genome-wide scale. GADD45a-binding to (e)RNAs can be tested by RNA immunoprecipitation sequencing (RIP-seq) or cross-linking immunoprecipitation (CLIP)-based methods. Impaired or altered chromatin-chromatin interactions can be tested by high-throughput chromosome conformation capture (Hi-C) and comparable methods.

4.4 Supplementary material

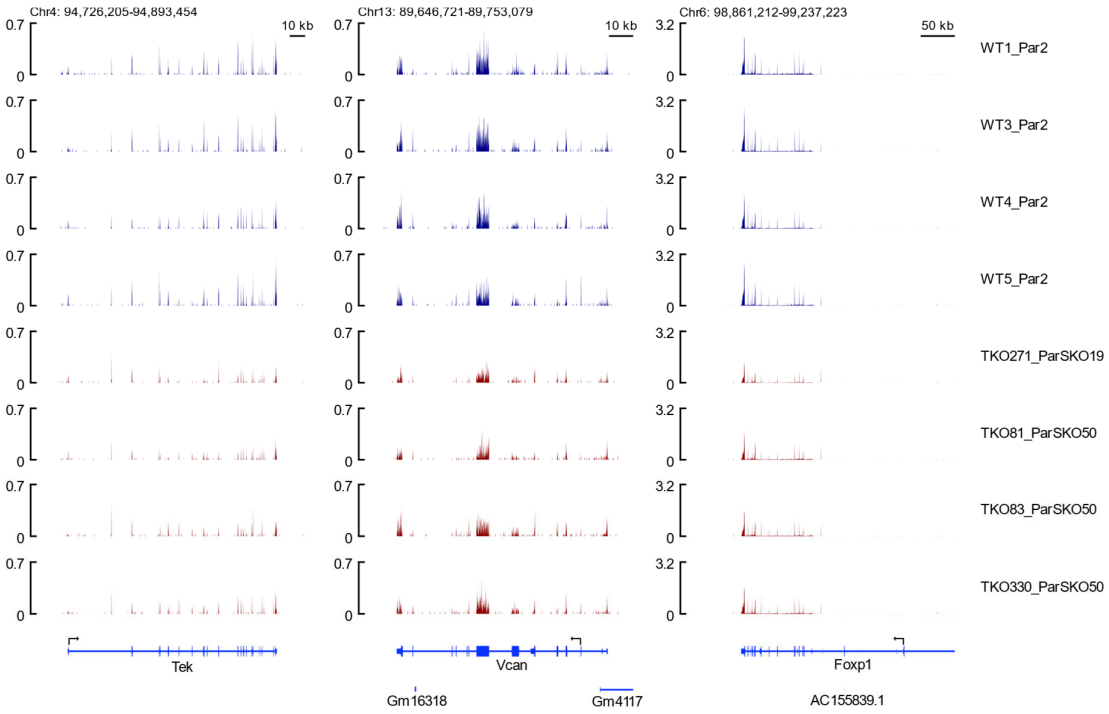


Supplementary Figure 4.1: All except for one *Gadd45* TKO clone set 3 TKO clones have a successfully knocked out *Gadd45a* coding sequence. Tracks show RNA-seq signal at the *Gadd45a* locus.

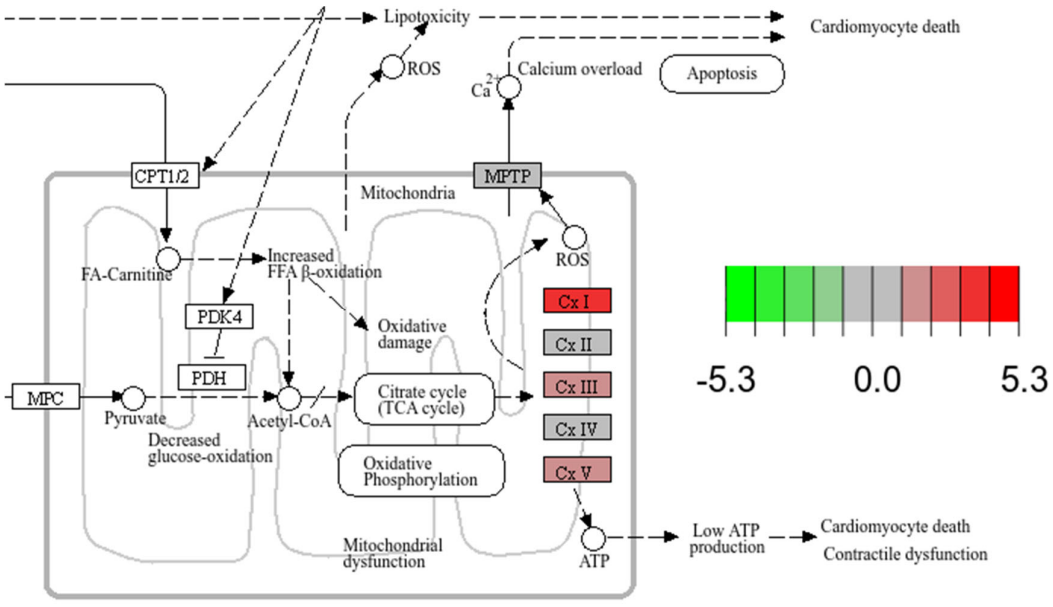


Supplementary Figure 4.2: *Gadd45* TKO clone set 3 TKO clones have a successfully knocked out *Gadd45b* coding sequence. Tracks show RNA-seq signal at the *Gadd45b* locus.

Supplementary Figure 4.3: *Gadd45* TKO clone set 3 TKO clones have a successfully knocked out *Gadd45g* coding sequence. Tracks show RNA-seq signal at the *Gadd45g* locus.



Supplementary Figure 4.4: Heart development genes are consistently down regulated in all *Gadd45* TKO clones. Tracks show RNA-seq signal in WT and *Gadd45* TKO clones for selected genes *Tek*, *Vcan* and *Foxp1*, all of which have known functions in heart development (Wilsbacher & McNally, 2016).



Supplementary Figure 4.5: Up regulation of electron transport chain complex genes in *Gadd45* TKOs may cause cardiomyocyte death. Cropped section of the diabetic cardiomyopathy KEGG pathway with indicated fold change in differential expression in *Gadd45* TKOs.

5 Part II: Development of methods to study repetitive DNA elements

5.1 Introduction

Repetitive DNA elements (or *repeats*) are patterns of nucleic acids occurring in multiple copies in the genome (Liao et al., 2023). In both, eukaryotes and prokaryotes, repeats cover a significant portion of genomes, and can be classified into two types based on the arrangement of the repeating units (Liao et al., 2023): Tandem repeats (or *satellite repeats*) and interspersed repeats (or *transposons*) (Figure 5.1). Tandem repeats are composed of highly similar DNA sequences repeated directly adjacent to each other (Trigiante et al., 2021). Interspersed repeats, on the other hand, are DNA sequences that can be found in multiple highly similar copies in distinct genomic loci as a result of their viral DNA-like ability to spread across the genome. (Genovese et al., 2018).

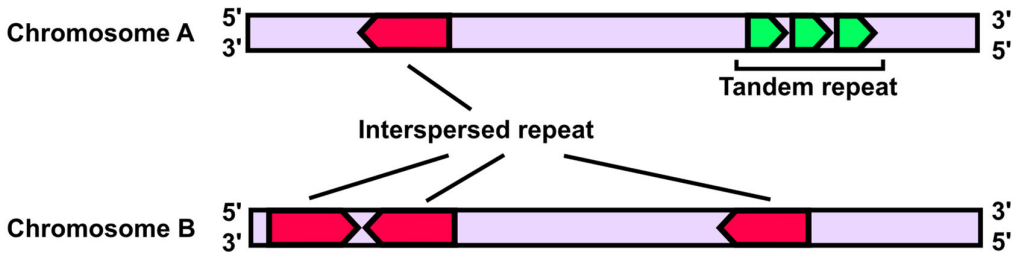


Figure 5.1: The two types of genomic repeats. Interspersed repeats consist of repeat units that are distributed across the genome, whereas tandem repeats are composed of repeat units directly adjacent to each other.

5.1.1 Tandem repeats

Tandem repeats (or *satellite repeats*) are patterns of nucleotides repeated numerous times in a head-to-tail manner (Eslami Rasekh et al., 2021). They are highly polymorphic and have been shown to occur in both prokaryotic and eukaryotic genomes (Liao et al., 2023). By the length of the repeat units forming the

repeat, tandem repeats can be further divided into different categories. While there is no scientific consensus on an exact categorization of tandem repeats, I will adhere to the definitions outlined in Eslami Rasekh et al. (2021): Short tandem repeats (STRs) (or microsatellites) are composed of repeat units shorter than seven bp, minisatellites consist of units of 7 to several hundred bp, and macrosatellites have hundreds to thousands of bp long repeat units (Figure 5.2).

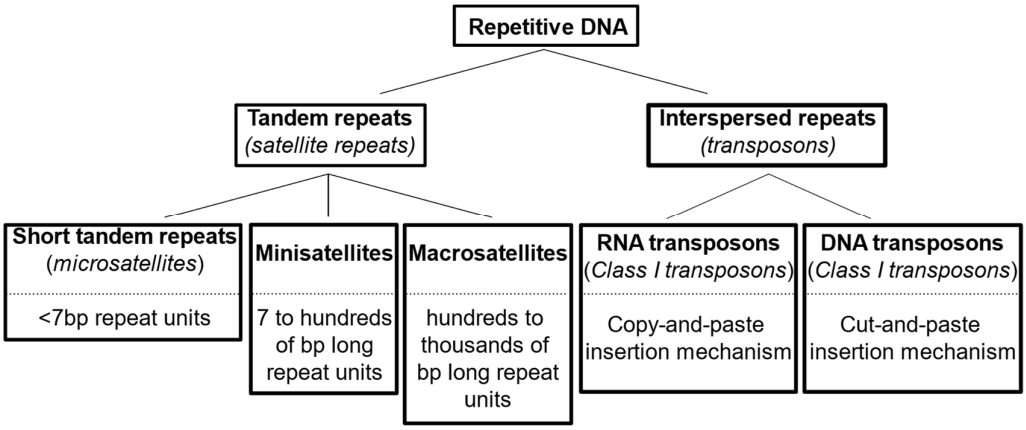


Figure 5.2: Repetitive DNA classification. Subclasses of DNA repeats with their distinguishing characteristics.

Due to their high copy number and exceptionally high mutation rate, STRs are of particular interest to biological and medical research (Liao et al., 2023). The majority of STR mutations result in a variation in the copy number of the repeat unit with multiple mechanisms contributing to STR length alterations (Verbiest et al., 2023). Out of these, the majority of alterations are thought to be a consequence of “strand slippage”, where misalignment of DNA strands during DNA replication can lead to stepwise repeat expansions (Verbiest et al., 2023). Such repeat expansions can affect both, somatic or germline cells, both of which can have deleterious consequences with the latter having the potential to affect offspring (Arning & Nguyen, 2021). STRs can be transcribed and produce a class of RNAs termed short tandem repeat-enriched RNA (strRNA) (Yap et al., 2018).

There is currently also no scientific consensus on the nomenclature of tandem repeats. Thus, for instance, the telomeric repeat (5’...TTAGGGTTAGGGTTA…3’) could be described to be comprised of 5’-TTAGGG-3’, 5’-TAGGGT-3’,

5’-AGGGTT-3’, 5’-GGGTTA-3’, 5’-GGTTAG-3’ or 5’-GGGTTA-3’ hexanucleotide repeat units. Indeed, different studies refer to the repeat unit of telomeric repeats as 5’-AGGGTT-3’ (Cysewski & Czeleń, 2009), 5’-TTAGGG-3’(Hinchie et al., 2024) or 5’-GGGTTA-3’ (Shiekh et al., 2022). In the scope of this work, I will be referring to tandem repeats by the lexicographically minimal string rotation of their repeat unit (i.e. the alphabetically first option). I would thus refer to the repeat unit of telomeric repeat as 5’-AGGGTT-3’.

5.1.2 Telomeric repeats and telomere maintenance

Telomeres are the genomic regions at the ends of linear chromosomes and, in vertebrates, consist of several kilobases of 5’-AGGGTT-3’ tandem repeats (Casari et al., 2022). In conjunction with the shelterin protein complex, the telomeric repeats form a protective cap that shields chromosome ends from DNA repair proteins, preventing telomere fusions and genome instability (Mir et al., 2020).

Due to the inability of DNA polymerases to replicate linear DNA in full-length, each cell division leads to the loss of a short terminal telomeric sequence (Rossiello et al., 2022). This “end replication problem” implies a shortening of the telomeres as a function of age, a process that in some cells, e.g. germ cells, is counteracted by a reverse transcriptase called telomerase (Mir et al., 2020). In most somatic cells or due to disease, telomerase activity can be silenced or be less efficient, and upon reaching a critical telomere length, affected cells undergo cell death or become senescent (Mir et al., 2020). Short telomeres have been linked to age-related degenerative diseases, such as Alzheimer’s disease or cardiovascular disorders and other diseases. In contrast, long telomeres are possibly associated with risks for several cancers (Mangaonkar & Patnaik, 2018; Vaiserman & Krasnienkov, 2020). While the links between shortened telomeres and diseases are often correlation-based and not necessarily causative (Rossiello et al., 2022), achieving telomere maintenance via telomerase expression or by recombination-based alternative lengthening of telomeres (ALT) has been shown to be a critical step in tumorigenesis (Lee et al., 2021). The subtelomeric regions upstream of the telomeres can initiate transcription via RNA Polymerase II, resulting in transcripts between 100nt and more than 9kb long transcripts termed Telomeric Repeat-

containing RNA (TERRA) that can thus contain large stretches of telomeric repeats (Chebly et al., 2022). Their proposed role as a negative regulator of telomerase remains so far unclear (Chebly et al., 2022). The mechanisms involved in telomere maintenance are subject to ongoing investigation and of particular interest to the fields of aging and cancer research.

5.1.3 Interspersed repeats

The discovery of interspersed repeats (or *transposons*) was first reported by Barbara McClintock in the early 1950s (McClintock, 1950). Interspersed repeats are selfish genetic elements that attempt to increase their copy numbers by novel integration into the host genome and can be divided into two groups: RNA transposons (or *Class I transposons*) and DNA transposons (or *Class II transposons*) (Muñoz-López & García-Pérez, 2010) (Figure 5.2). RNA transposons employ a so-called “copy-and-paste” insertion mechanism where RNA transposons are first transcribed from their host locus into RNA and translated into proteins, one of which is usually a reverse transcriptase (Bourque et al., 2018). The newly synthesized proteins then bind back to the transposon RNA, forming ribonucleoproteins (RNPs) that can reverse transcribe the RNA molecule into complementary DNA (cDNA) (Bourque et al., 2018). Subsequently, the cDNA can be inserted into a different locus within the host genome, completing the insertion cycle (Bourque et al., 2018). DNA transposons on the other hand work via a “cut-and-paste” insertion mechanism (Wells & Feschotte, 2020). Here, transposon genes are excised from their locus and inserted into a different genomic locus using a transposase enzyme that is encoded within some DNA transposons (Wells & Feschotte, 2020). Both human and mouse DNA transposons lack active transposases, hence DNA transposons are considered to be evolutionary remnants in these species (Liao et al., 2023).

5.1.4 Long-interspersed nuclear elements-1 (LINE1s)

Comprising ~17% of the human genome, LINE1 RNA transposons are considered to be the most representative class of all transposons (Bodak et al., 2014). In fact, all RNA transposons combined make up ~42% of the human genome, whereas all DNA transposons account for 3% (Bodak et al., 2014). Full-length (i.e. non-truncated) LINE1s are 6-7kb long elements with a 5'-untranslated region (5'UTR), an internal promoter, two open reading frames (ORFs) that code for the proteins ORF1p and ORF2p, as well a 3'-untranslated region (3'UTR) terminating in a poly(A) tail (Boissinot & Sookdeo, 2016). ORF1p is a 40 kDa protein thought to function as an RNA chaperone for the LINE1 parent RNA (Sil et al., 2023) whereas ORF2p is a 150 kDa protein responsible for the insertion of the LINE1 cDNA into the genome via its DNA endonuclease and reverse transcriptase functions (Goodier et al., 2007).

In principle, LINE1 cDNA can integrate into protein coding sequences of genes or in their regulatory elements, leading to mutated proteins, aberrant splicing events and gene misregulation (Zhang et al., 2020). Other than being implicated in cancer, LINE1s are known to play roles in metabolic, neuronal and autoimmune disorders (Zhang et al., 2020). Cells thus silence LINE1 expression by epigenetic mechanisms, most notably DNA methylation. (Lanciano et al., 2024). Abnormal epigenetic repression of LINE1s can lead to an increase in their transcription levels, a known hallmark of many cancer types (Perkiö et al., 2024). Sexton and Han (2019) list 26 currently known evolutionary LINE1 subfamilies in human. Out of these, only the most recent LINE1 subfamily L1HS remains fully retrotransposition-competent (Mir et al., 2015). The exact mechanism governing the LINE1 life cycle from transcription to integration is incompletely understood. (Baldwin et al., 2024; Mendez-Dorantes & Burns, 2023; Warkocki, 2023).

5.1.5 Biochemical assays to quantify repetitive sequences

Quantifying and comparing the expression level of repetitive elements, the extent to which a modification is present on repetitive DNA or RNA, or the copy number of a repeat in a genome can be a crucial method to understand a repeat's function or

the mechanism that regulates it. Biochemical assays for such analyses are direct, often highly specific and sensitive (Freen-van Heeren, 2021; Wang et al., 2013). Due to the absence of a sequencing step, they do not require expensive sequencing experiments with complex computational analyses.

Quantitative polymerase chain reaction (qPCR) is a cost effective, fast, sensitive and relatively high-throughput method that can be applied to compare levels of nucleic acid sequences in different samples (Petit-Marty et al., 2023). qPCR measures the amount of the PCR product present during a given PCR cycle using fluorescence and relies on target sequence-specific PCR primers (Petit-Marty et al., 2023). qPCR can also be applied as reverse transcription quantitative real-time PCR (RT-qPCR) to reverse-transcribed cellular RNA for comparison of RNA levels or after RNA enrichment, e.g. via immunoprecipitation, to study the presence of a target modification or protein, hereafter *targets*, on RNA (Martindale et al., 2020). Moreover, qPCR can be performed on DNA to compare levels of genomic DNA sequence content or after DNA or chromatin enrichment to study the presence of targets on DNA (Solomon et al., 2021). Quantitative fluorescent in situ hybridization (Q-FISH) is a fluorescent in situ hybridization (FISH)-based method to visualize and quantify the presence of a target nucleic acid sequence of interest in cells (Poon & Lansdorp, 2001). For this, fluorescently labeled peptide nucleic acid (PNA) probes complementary to a sequence of interest are used to bind to the target sequence and are subsequently visualized using a fluorescence microscope. This approach can be combined with flow cytometry (Flow-FISH) to measure the emitted signal in cell populations (Poon & Lansdorp, 2001). Both, Q-FISH and Flow-FISH are commonly used to study telomere lengths by probing for 5'-AGGGTT-3' repeats (Yu et al., 2024) but can be adapted to other repeats.

Multiple additional methods have been developed for the quantification and comparison of repetitive sequences (Yu et al., 2024). Similar to the previously mentioned approaches, they commonly rely on primers or probes specific to the repeats of interest. Since repetitive loci, by their very definition, are identical or highly similar to adjacent or distant loci, it can be difficult and often impossible to design PCR primers or oligonucleotide probes that are specific to a single repetitive region of interest. Nonetheless, it is in some cases possible to design primers or probes with a high specificity to target (sub)families of repeats of interest,

distinguishing them from other (sub)families. This is, however, a difficult process, as the amplification of unintended targets, that can share a high degree of homology, must be minimized. Moreover, such purely biochemical assays do not provide a genome-wide picture on a studied aspect, such as RNA expression, as they rely on specifically designed primers or oligonucleotides that only targets sequences of interest. Conversely, unbiased approaches that provide genome-wide pictures on repetitive and non-repetitive genomic elements generally require sequencing with advanced *in silico* analysis.

5.1.6 Studying repetitive sequences using sequencing approaches

Sequencing-based approaches, with the exception of targeted sequencing, commonly provide genome-wide pictures on gene expression, binding sites of a target of interest or the genome of an organism. Such approaches can also be used to specifically study repetitive elements.

Transcriptome sequencing technologies, such as RNA-seq or nascent RNA sequencing methods, can be used to identify and compare RNA expression (Gondane & Itkonen, 2023). These approaches produce quantitative amounts of sequencing reads for detected RNAs in cells' transcriptomes. While mostly used for the study of expression of annotated genes, such analyses can be adapted to study the expression of repetitive elements. For transposons, transcription into RNA is the first step of their life cycle, and therefore an essential prerequisite for RNP formation and retrotransposition. Hence, measuring transposon RNA levels is crucial to study their activity. Meanwhile, for tandem repeats, expression is rarely studied, with the exception of the short tandem repeat RNA (strRNA) TERRA. For other strRNAs, there is only rather recent evidence showing that these repeats can play functional roles (Ninomiya & Hirose, 2020; Yap et al., 2018).

Epigenetic marks as well as proteins attached to DNA or chromatin can have transcriptionally activating or repressing effects, whereas RNA modifications are known to modulate the stability of RNA molecules. Moreover, epigenetic marks on DNA or RNA can affect the affinity for nucleic acid binding proteins. Epigenetic profiling sequencing methods, such as ChIP-seq, DIP-seq, RNA

immunoprecipitation sequencing (RIP-seq) and others are commonly used to determine the localization of epigenetic marks, proteins or other targets on nucleic acid molecules (Chen, 2019). Such methods are commonly enrichment-based, i.e. loci of interaction produce an increased amount of sequencing reads that can be detected. Other than that, there are chemical conversion-based methods that, for instance, use bisulfite or pyridine borane treatment, which results in the specific conversion of either modified or unmodified bases that can be detected with base-resolution (Dai et al., 2024; Liu & Song, 2022). Likewise, such methods can be used to study targets of interest on RNAs transcribed from repetitive loci or on genomic loci containing repetitive elements to study their epigenetic regulation.

Whole-genome sequencing (WGS) is a DNA sequencing technique where the entire DNA in a sample of cells is read via sequencing and is commonly used for the detection of mutations, structural variations, karyotyping, or for genome assemblies (Ekblom & Wolf, 2014; Mareschal et al., 2021; Qin, 2019; Wheeler et al., 2022). In the study of repetitive sequences, such datasets can be used to detect differential telomere lengths, loci of tandem repeat expansion or novel integration sites of transposons (Bahlo et al., 2018; Lee et al., 2017; Savage et al., 2022).

Transcriptomics sequencing technologies, most notably RNA-seq, as well as WGS can be performed using either short-read or long-read, whereas read enrichment-based or chemical conversion-based epigenetic profiling techniques are currently only established using short-reads. Nevertheless, Oxford Nanopore-based PCR amplification-free RNA-seq or WGS datasets can be utilized for modification calling with base-resolution by applying accordingly pre-trained basecalling models (Wang et al., 2024; White & Hesselberth, 2022). Till date, however, few modifications can be reliably detected in practice using such models and enrichment-based and chemical conversion-based short-read sequencing techniques for epigenomic profiling remain the standard choice for such analyses. Likewise, short-read sequencing continues to be widely used for RNA-seq and WGS due to its superior basecalling accuracy. This advantage of short-reads does, however, come at the cost of challenges in the analysis of repetitive regions.

5.1.7 Challenges in the analysis of repetitive regions in enrichment-based short-read sequencing data

Standard analysis pipelines for the detection of differential expression of RNA or pipelines for the mapping or detection of differential presence of proteins and modifications on nucleic acid sequences by enrichment-based short-read experiments commonly rely on the accurate mapping of short-reads to reference genomes. After read mapping, normalized read numbers overlapping with annotated genes or *de novo* detected enrichment of reads called as peaks can be compared between different samples to detect differential expression or differential binding. While this approach mostly works well for non-repetitive regions, read mapping can be challenging for reads originating from repetitive loci (Figure 5.3).

Repeats are by their very definition composed of copies of sequences that are either directly adjacent to each other in the case of tandem repeats or in distant loci across the genome in the case of interspersed repeats. Moreover, different tandem repeats with a fully or partially identical sequence can often also be found at different genomic loci. Therefore, reads originating from repetitive loci can often not be uniquely assigned to a single genomic locus and are thus flagged as multimappers (or multimapping reads) by read mapping software. An additional source of error can stem from unmapped reads, i.e. reads that cannot be at all assigned to the reference genome. Different genomic regions are prone to produce reads that cannot be assigned: (1) Low-quality and variable genomic regions such as centromeres or telomeres are often hard-masked, i.e. replaced with stretches of Ns in reference genomes, (2) regions that are misrepresented due to genome assembly errors, and (3) regions subject to inter-individual variation, such as the HTT gene in the human genome that varies in the number of 5'-AGC-3' repeat unit copies in different individuals.

Due to the challenges associated with the analysis of repeats in enrichment-based short-read sequencing data, appropriate strategies are required.

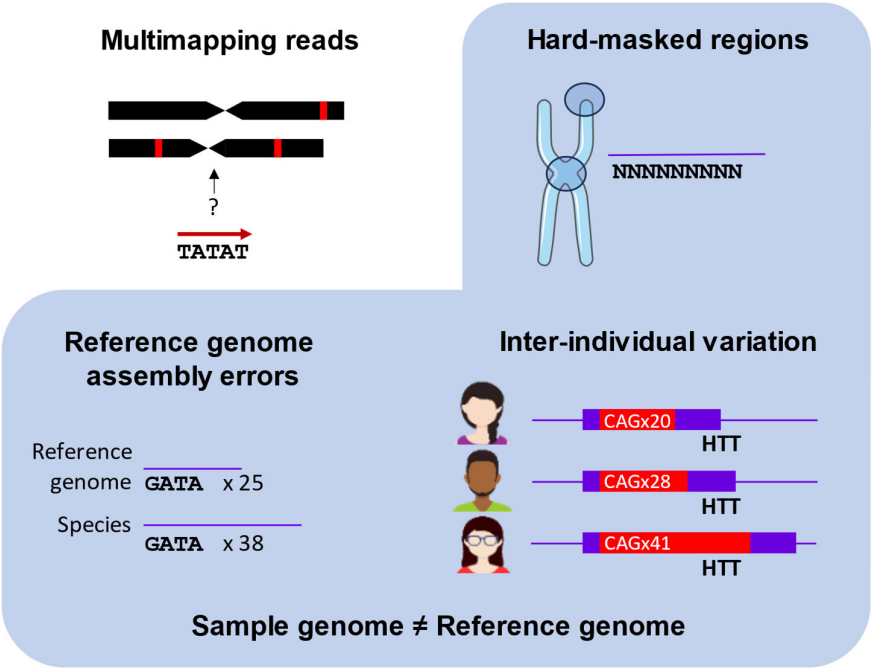


Figure 5.3: Illustration of common challenges in read mapping for the analysis of DNA repeats in short-read sequencing data. Hard-masked regions, reference genome assembly errors and inter-individual variation can be summarized as discrepancies between the genome of the sequenced sample and the reference genome.

5.1.8 Bioinformatics approaches for quantification-based tandem repeat analyses and their limitations

If a standard-type mapping-based analysis pipeline for the detection of enriched short-reads is employed, most of the previously discussed sources responsible for the generation of unmapped reads cannot be easily addressed. There are, however, methods to deal with multi-mapping reads (Deschamps-Francoeur et al., 2020): (1) Alignments of multimapping reads can be removed or ignored in the post mapping analysis. While this reduces the effect of reads of uncertain origin in downstream analyses, it leads to an underestimation of some repeats, potentially leading to false negatives in differential comparisons. (2) Multimapping read alignments can be retained in the sequencing data and thus, a read might be counted multiple times. This leads to a systematic overestimation of reads originating from repetitive elements and thus in downstream analyses to an overlooking of regions with differential coverage or in an overestimation of sites with differential coverage, if few sites of strong differential coverage sufficiently multimap to other (normally non-

differential) sites that can then pass thresholds to be called differential. (3) Counts of alignments of multimapping reads can be equally split across all mapping loci, ensuring that each read is counted only once. This, however, dilutes the effect of a differential locus between all valid alignments. (4) Other advanced methods attempt to estimate expected read coverage for multimapping loci from the coverage of surrounding regions or uniquely mapped reads and distribute multimappers accordingly. While the aforementioned methods offer a variety of ways to deal with multimapping reads and the choice of a suited method can substantially increase the accuracy of downstream analyses, none of them offers an ideal solution for the accurate comparison of multimapping reads. Researchers have thus come up with specialized solutions for the analysis of tandem repeat enrichment in sequencing data.

RepEnrich (Criscione et al., 2014) was developed for repeat enrichment analysis of sequencing data, such as RNA-seq or ChIP-seq, and makes use of reference genomes plus a corresponding user-supplied repeat annotation, such as RepeatMasker (http://www.repeatmasker.org). After read mapping, uniquely mapping reads are counted individually per repeat family whereas for multimappers, RepEnrich attempts a mapping to pseudogenomes that consist of a concatenation of all annotated repeats of each repeat subfamily annotation plus their flanking regions and spacer sequences. For reads that multimap between different repeat families, fractional counting is applied. Last, a counts table containing read counts for each repeat subfamily is produced and provided as an output. Repeat read counts can then be normalized using read count normalization strategies, such as DESeq2 (Love et al., 2014) or edgeR (Robinson et al., 2010) and repeat families with significantly differential read counts can be identified. Building upon RepEnrich and other bioinformatics innovations, RepEnTools (Choudalakis et al., 2024) was recently published. Similar to RepEnrich, RepEnTools aligns reads to a reference and counts repeats per annotated repeat family. For multimappers, one optimal alignment is randomly chosen and counted once. Finally, reads are normalized and enrichments can be detected. Both, RepEnrich and RepEnTools effectively address the problem with multimappers and are capable of detecting read enrichment for annotated repeats, including many tandem repeats and interspersed repeats. However, they do not address the previously discussed read mapping issues due

to discrepancies between the sequenced sample and the reference genomes, as both rely on read mapping to a linear reference genome. Moreover, RepEnrich development appears to be deprecated. (https://github.com/nerettilab/RepEnrich2). Both RepEnrich as well as RepEnTools should, in principle, be able to detect enriched reads at telomeres as long as an appropriate genome and repeat annotation that include non-masked telomeres are used. The human telomere-to-telomere (T2T) genome, which includes telomeric sequences, is a recent development (Nurk et al., 2022) and has not yet been widely adopted as a standard reference. A mouse T2T genome has thus far not been published.

Different tools have been developed for the comparison of total telomere lengths between different samples of WGS data. Such tools rely on the identification and quantification of reads of apparent telomeric origin. TelSeq estimates telomere length by counting reads containing at least *k* (default: 7) 5'-AGGGTT-3' repeat units and inserting the counts into an equation that estimates telomere length using these counts (Ding et al., 2014). Computel quantifies reads of apparent telomeric origin by first creating an index for and mapping reads to a telomeric pseudogenome (Nersisyan & Arakelyan, 2015). Read counts are subsequently inserted in an equation that estimates telomere length from these counts (Nersisyan & Arakelyan, 2015). Telomerecat (Farmery et al., 2018) and TelomereHunter (Feuerbach et al., 2019) combine information from mapped paired-end reads in subtelomeric regions, singleton reads, where only one of two mates maps, telomere-subtelomere junction-spanning reads and unmapped telomeric repeats identified via their 5'-AGGGTT-3' content to estimate telomere length. TelSeq, Computel, Telomerecat and TelomereHunter are all limited to the single use of comparing telomere content and do not support any other repeat quantifications.

5.1.9 Aim

My aim was to overcome some of the challenges in the study of tandem repeats by developing a Bioinformatics application that can faithfully detect and quantify tandem repeat enrichment in one group of short-read sequencing samples over another. In principle, such a program could be useful for e.g. relative comparison of estimated telomere length from WGS samples or the detection of tandem repeat-

binding targets of interest in ChIP-seq, CUT&Tag and other enrichment-based sequencing methods. In addition, I aimed to develop and apply a workflow to design a set of PCR primers that can discriminate between different subfamilies of LINE1 retrotransposons. Such primers could be useful to study LINE1 subfamily-specific expression or subfamily-specific targets on LINE1 sequences.

5.2 Results

5.2.1 Development of a bioinformatics application for *de novo* detection of enriched tandem repeat content

I developed counTR, a read mapping-free and highly parallelizable method to quantify tandem repeat enrichment without prior knowledge in quantitative sequencing data. counTR's basic rationale is outlined in Figure 5.4. In brief, counTR searches tandem repeats in raw reads using Phobos (Mayer et al., 2010). Detected tandem repeat-containing reads are grouped and quantified for each sample of interest and a read count matrix is produced. Downstream analysis of the read count matrix can be performed via standard read count normalization and differential analysis methods, such as DeSeq2 (Love et al., 2014), edgeR (Robinson et al., 2010), and limma-voom (Law et al., 2014). The analysis results in the reporting of statistically overrepresented or underrepresented tandem repeats in one sample group over another. Thus, by applying counTR, the user is able to detect differential tandem repeat content between groups of sequencing samples, completely independent of read mapping. By additionally grouping the repeats by length or perfection before counting, differences in quantities of imperfect or differently sized tandem repeats can be studied.

Figure 5.5 outlines the detailed program architecture and computational workflow of a counTR analysis. *processrepeats* is counTR's primary function and designed to run efficiently on high performance computing (HPC) clusters with many CPU cores, but also on workstation computers. After receiving a sequencing sample as input data, all processing is carried out in the RAM, thereby minimizing I/O operations to only input and output files and avoiding time-consuming writing to storage devices. The program reads the input file in chunks of *n* lines, and each chunk is submitted to a parallel process as soon as it is read, until all of *m* parallel processes are busy. This design should theoretically lead to an almost linear reduction of the *processrepeats'* runtime with increasing numbers of CPU cores, making it well-suited to run on HPC clusters with many CPU cores. The sequencing data chunks of *n* lines length are first piped into Phobos, which scans each read for tandem

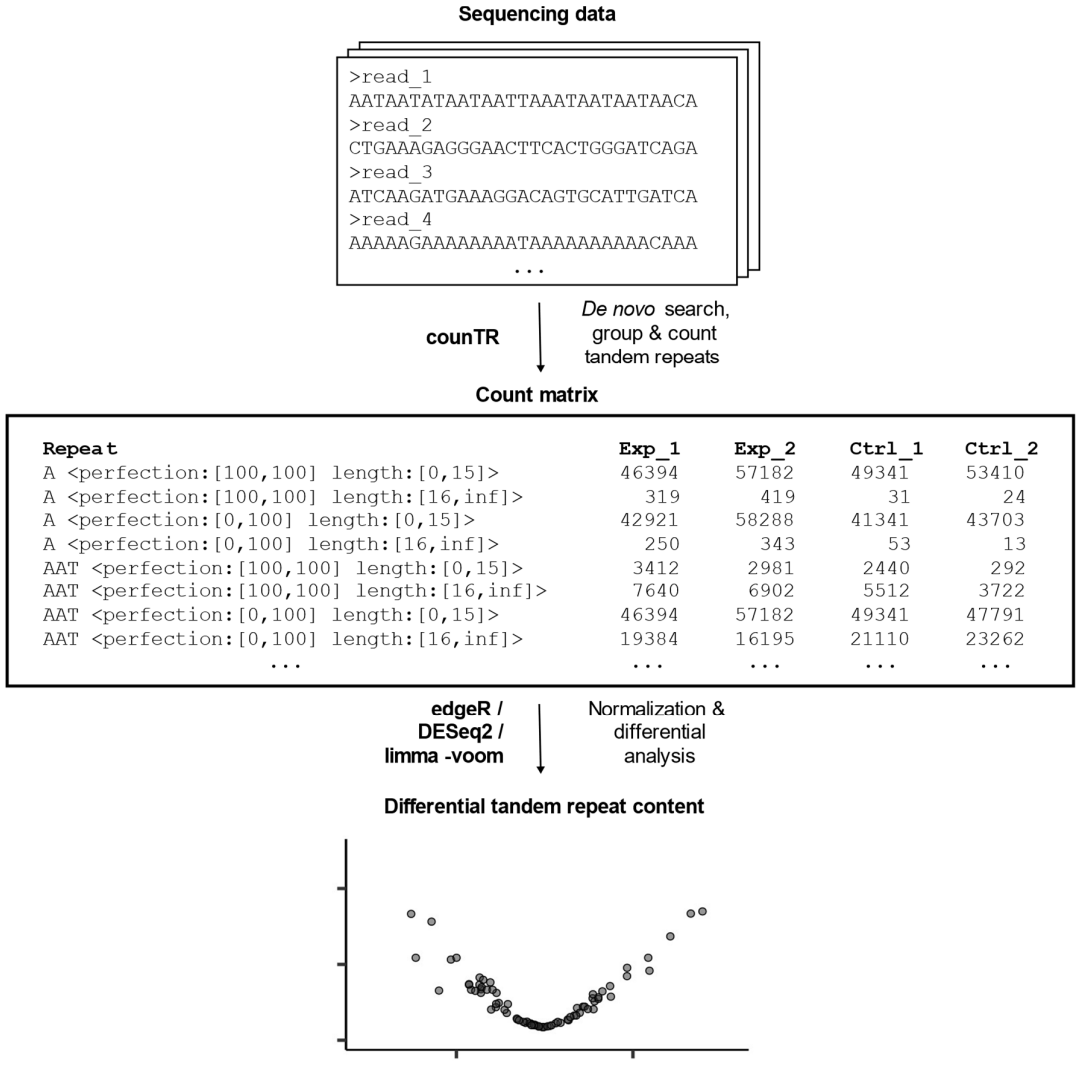


Figure 5.4: Basic rationale of a counTR analysis. Tandem repeats are de novo detected from raw reads, optionally grouped by length or repeat perfection, and counted. A count matrix containing repeat counts for each sample count is generated. The counts are then normalized by edgeR, DESeq2 or limma-voom and significantly differential tandem repeat content between different conditions is detected.

repeats and reports them along with various other statistics, such as repeat length and perfection. Next, repeats are filtered, grouped and counted group-wise depending on user-supplied parameters. By default, a detected repeat's group is identical to its detected repeat unit, e.g. all 5'-AGGGTT-3' repeat counts are reported in the "AGGGTT" group. Users can decide to additionally group repeats by length or perfection, e.g. "AGGGTT <perfection:[0,100) length:(80,100]>" (square or round brackets correspond to mathematical interval notation) or to group repeats with their reverse complement (Table 5.1). In the latter case, 5'-AGGGTT-3' repeats

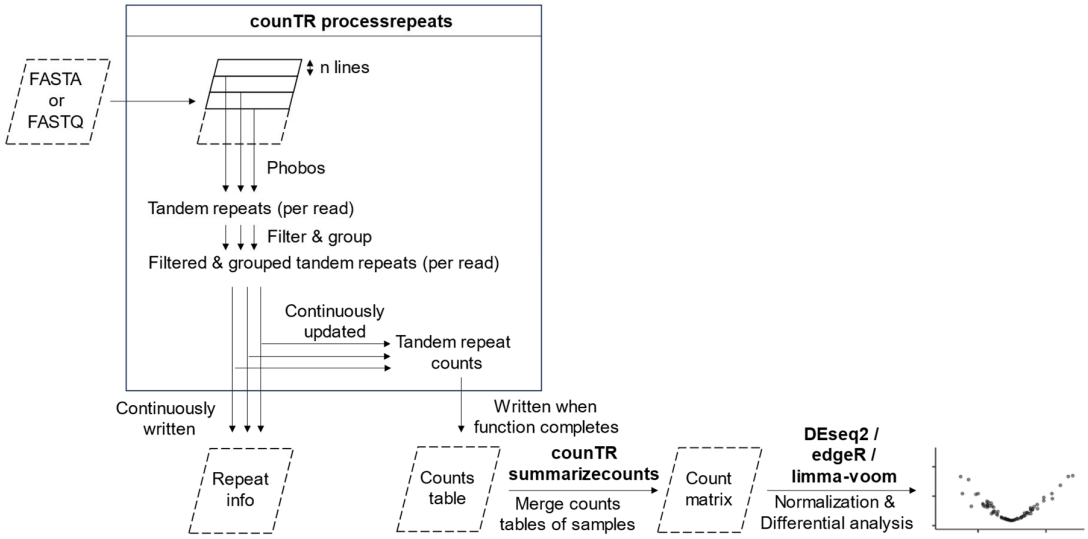


Figure 5.5: Detailed computational workflow of counTR. counTR's *processrepeats* function is run using FASTA or FASTQ files as input. The input file is read successively in chunks of *n* (user defined) lines, out of which *m* (user defined) are simultaneously processed in separate parallel processes. Initially, a chunk is internally converted into a modified FASTA file that is stored in the RAM and piped into Phobos, which detects tandem repeats contained within the reads. Resulting detected tandem repeats are grouped and optionally filtered. Detailed information on every detected repeat can be continuously written to a *repeat info* file. Simultaneously, an internal repeat counts list is continuously updated with the detected repeats of each chunk. Upon completion, the repeat counts list can be written to a *counts table* file. counTR *summarizecounts* can then be used to merge repeat counts tables of multiple samples into a count matrix, which can be used for read count normalization and differential analysis via DESeq2, edgeR or limma-voom.

that would normally be grouped into the "AGGGTT" group, would be grouped and counted together with 5'-AACCCT-3' repeats in the "AACCCT" group, since "AACCCT" comes alphabetically first. This function is useful in the analysis of non-stranded sequencing data, such as typical WGS or ChIP-seq experiments, where due to the lack of strand-specific enrichment, it is usually appropriate to summarize counts of repeats and their reverse complements. Detected repeats can also be written, along with detailed statistics on each repeat, to a continuously written *repeat info* file. Meanwhile, counts for each repeat group are continuously tracked in an internal repeat counts list that, upon completion of all chunks, is written to a *repeat counts* file. Repeat counts files of multiple samples are then merged into a count matrix via counTR *summarizecounts*. This step is necessary to add repeats that

Table 5.1: counTR *processrepeats* parameters

Parameter	Description
inputpath	Path to sequencing data file in fasta(.gz) or fastq(.gz) format.
outputdirectory	Directory where the output will be written to.
phobospath	Path to Phobos executable.
outputprefix	Prefix of output files, prefix will be taken from input file, if not set (default: None)
outputtype	Output to generate, countstable.txt (c), repeatinfo.txt (i), repeatinfo.txt.gz (g), concatenate the letters for multiple outputs, e.g. ci (countstable.txt and repeatinfo.txt) (default: c)
processes	Number of parallel processes to be used, to automatically set to maximum number of available logical cores, use 'auto' (default: auto)
grouping	Repeat grouping settings, example: 'perfection:[0,100)[100,100] length:[0,30)[30,inf]' (note the single quotation marks), if 'None', repeats will be only grouped by their motif (default: None)
groupingmotif	Motif to report for grouping, report the detected motif as is (detected), its reverse complement (rc), or combine forward and reverse complement (combine), all motifs are reported as their lexicographically minimal string rotation (default: detected)')
minperfection	Minimum perfection of a repeat to be considered (default: 0)
maxperfection	Maximum perfection of a repeat to be considered (default: 100)
minrepeatlength	Minimum repeat region length for a repeat to be considered (default: 0)
maxrepeatlength	Maximum repeat region length for a repeat to be considered (for infinite, set value to: inf) (default: inf)
minunitsize	Minimum repeat unit size for a repeat to be considered (default: 0)
maxunitsize	Maximum repeat unit size for a repeat to be considered (for infinite, set value to: inf) (default: inf)
mincopynumber	Minimum number of repeat unit copies in a repeat for a repeat to be considered (default: 0)
maxcopynumber	Maximum number of repeat unit copies in a repeat for a repeat to be considered (for infinite, set value to inf) (default: inf)
multirepreads	Which repeat to consider in case of reads with multiple repeats (after other filters have been applied), either all (consider all repeats for each read), none (ignore multi repeat reads), longest (only consider the longest repeat) or unique_longest (for each unique repeat unit, only consider the longest) (default: all)
readwhitelist	Path to list of read names that will not be filtered out, the rest is filtered (default: None)
readblacklist	Path to list of read names that will be filtered out, the rest is kept (default: None)
readchunksize	Approximate number of lines that are analyzed at once in a (parallel) process (default: 50000)
addphobosarguments	Add arguments to the default Phobos call (which is run with: --outputFormat 1 --reportUnit 1 --printRepeatSeqMode 2) Example: '--indelScore -4;--mismatchScore -5' (note the single quotation marks). This will run Phobos with: --outputFormat 1 --reportUnit 1 --printRepeatSeqMode 2 --indelScore -4 --mismatchScore -5 and thus change the parameters Phobos uses to align detected repeats to ideal repeats Warning: This command changes the way Phobos generates its output before it is passed to counTR and can result in unexpected behaviour, use with caution (default: None)

have not been detected in all samples as zero counts to the remaining samples. The count matrix can then be subjected to read count normalization and differential analysis. All available parameters for counTR's *processrepeats* and *summarizecounts* functions are listed in Table 5.1 and Table 5.2, respectively.

Table 5.2: counTR *summarizecounts* parameters

Parameter	Description
outputfile	path to output count matrix
inputpaths	countstable.txt files to be summarized into a count matrix
samplenames	list of sample names to be used in the resulting header in the same order as input files. If not set, input file names will be used (default: None)

While RepEnrich (Criscione et al., 2014) and RepEnTools (Choudalakis et al., 2024) detect both tandem and interspersed repeat enrichment through mapping-based approaches, counTR adopts a mapping-free design specialized for the detection of enriched tandem repeats. This approach makes counTR unsuitable for the analysis of interspersed repeats but enables it to detect and count tandem repeats divergent from reference genomes. Moreover, counTR can group tandem repeats by length or perfection, thereby revealing enrichment signatures within specific tandem repeat subpopulations. Key differences between counTR, RepEnrich and RepEnTools are summarized in Table 5.3, though the comparison does not include counTR's extended functionality including detailed per-repeat statistics, allowing for advanced analyses.

Table 5.3: Feature comparison between counTR, RepEnrich and RepEnTools.

	counTR (this work)	**RepEnrich** (Criscione et al., 2014)	**RepEnTools** (Choudalakis et al., 2024)
Detects tandem repeat enrichment	yes	yes	yes
Detects interspersed repeat enrichment	no	yes	yes
Detects telomeric repeat enrichment	yes	Reference genome-dependent	Reference genome-dependent
Reference genome- and repeat annotation-free	yes	no	no
Counts tandem repeat-containing reads divergent from reference genomes	yes	no	no
Subgroup tandem repeats (e.g. by perfection)	yes	no	no

5.2.2 counTR efficiently scales with increasing CPU core numbers

To benchmark counTR's multi-core scalability, I ran counTR on a representative ChIP-seq sample on an HPC cluster in separate runs with exponentially doubling CPU cores and measured the run times. As evident in Figure 5.6, a doubling of CPU cores corresponds to approximately a halving of program run time, indicating a linear scalability with respect to the number of CPU cores utilized. counTR's *summarizecounts* function was not benchmarked as it only combines the counts tables produced by *processrepeats* into a count matrix and thus usually completes within seconds.

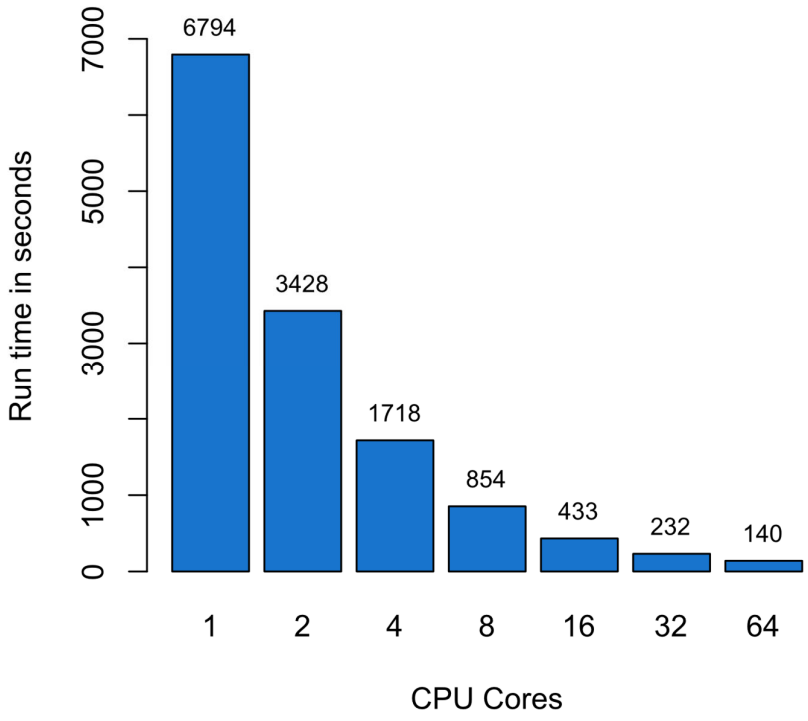


Figure 5.6: counTR run time scales efficiently with increasing number of CPU cores. counTR's *processrepeats* function's run time for a standard ChIP-seq sample of 28.6M 126bp reads using an HPC cluster. The function was run separately with 1, 2, 4, 8, 16, 32 and 64 CPU cores.

5.2.3 Detection of differential telomeric repeat content in simulated and experimental WGS data

To benchmark counTR's ability to detect differential tandem repeat content, I simulated triplicates of WGS short-reads for human chromosome 4, as well as scenarios where this chromosome is expanded by 10 kb of telomeric repeats or 10 kb of non-repetitive sequence. In brief, I first extracted chromosome 4 of the T2T-CHM13v2.0 assembly (T2T-chr4) and created modified versions of it where I expanded each chromosome end with either 5 kb of perfect telomeric repeats (T2T-chr4Δtelo) or genomic sequence extracted from the human DNAJA1 gene (T2T-chr4Δ*DNAJA1*) (Figure 5.7a). Using T2T-chr4, T2T-chr4Δtelo and T2T-chr4Δ*DNAJA1* as templates, I simulated 4M WGS reads per reference genome and performed a counTR analysis, where I compared T2T-chr4Δtelo and T2T-chr4Δ*DNAJA1* to T2T-chr4. Only the T2T-chr4Δtelo versus T2T-chr4, but not the T2T-chr4Δ*DNAJA1* versus T2T-chr4 comparison shows a statically significant differential repeat, namely the telomeric repeat (Figure 5.7b,c). Figure 5.7d shows the quantification of normalized telomeric repeat content for all three samples. There is no significant difference in telomeric repeat content between T2T-chr4 and T2T-chr4Δ*DNAJA1*, whereas T2T-chr4Δtelo shows the expected 3-fold increase in comparison to T2T-chr4 (Figure 5.7d).

Telomerecat and TelSeq are programs that estimate the combined length of all telomeres in a WGS sample. In Farmery et al. (2018), Telomerecat's authors compared the performance of their program to the performance of TelSeq. For this, they used a publicly available dataset published in Cai et al. (2014). Here, the authors performed WGS on *in vivo* mesenchymal stem cells (MSCs) extracted from a bone marrow donor, and on 1x, 8x, and 13x passaged MSCs as well as on induced pluripotent stem cells (iPSCs) generated from 1x passaged MSCs (Cai et al., 2014) (Figure 5.8a). MSCs contain no or only low levels of telomerase and therefore undergo telomere shortening with increasing passage numbers (Simonsen et al., 2002; Zimmermann et al., 2003), whereas iPSCs have restored telomerase activity and can thus be expected to have longer telomeres (Ma et al., 2023). Figure 5.8b and Figure 5.8c were taken from Farmery et al. (2018) and show Telomerecat's and TelSeq's telomere length estimations for aforementioned MSC samples. The authors demonstrated that their tool Telomerecat outperforms TelSeq

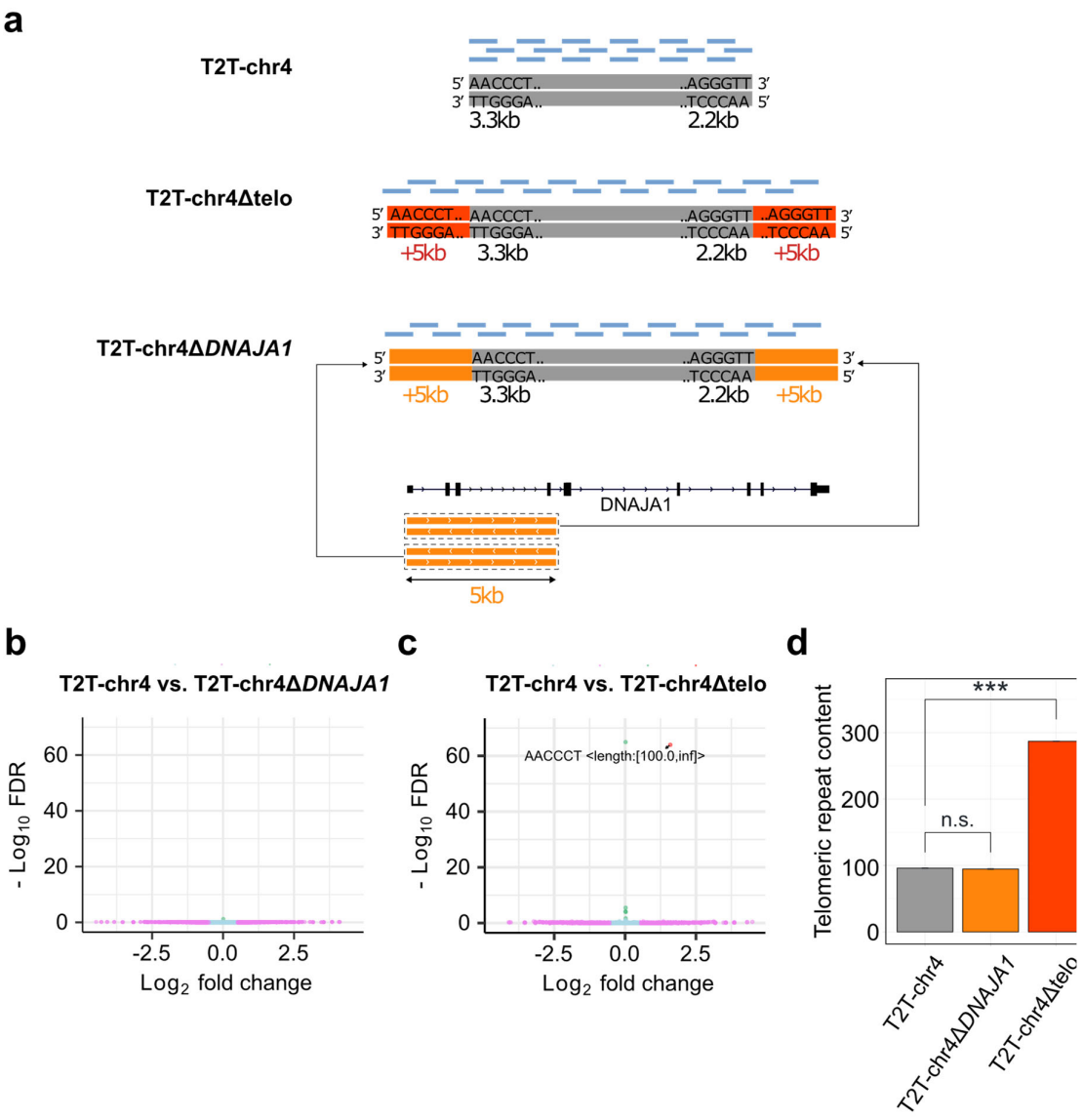


Figure 5.7: counTR detects differential telomeric repeat content in simulated WGS data. (a) Triplicates of WGS reads were simulated using chromosome 4 of the T2T-CHM13v2.0 assembly (T2T-chr4), a modified version where both telomeres were expanded by 5kb of perfect telomeric repeat sequence (T2T-chr4Δtelo), and a modified version where both telomeres were expanded by the first 5kb of the human *DNAJA1* gene or its reverse complement (T2T-chr4Δ*DNAJA1*) as illustrated. I ran counTR and plotted the differential tandem repeat content **(b)** in T2T-chr4Δtelo versus T2T-chr4 and **(c)** in T2T-chr4Δ*DNAJA1* versus T2T-chr4 (Light blue: n.s. and below FC cutoff, violet: n.s. and above FC cutoff, green: significant and below FC cutoff, red: significant and above FC cutoff, FDR cutoff: 0.05, FC cutoff: 0.5). **(d)** Quantification of normalized telomeric repeat content for T2T-chr4, T2T-chr4Δtelo and T2T-chr4Δ*DNAJA1*.

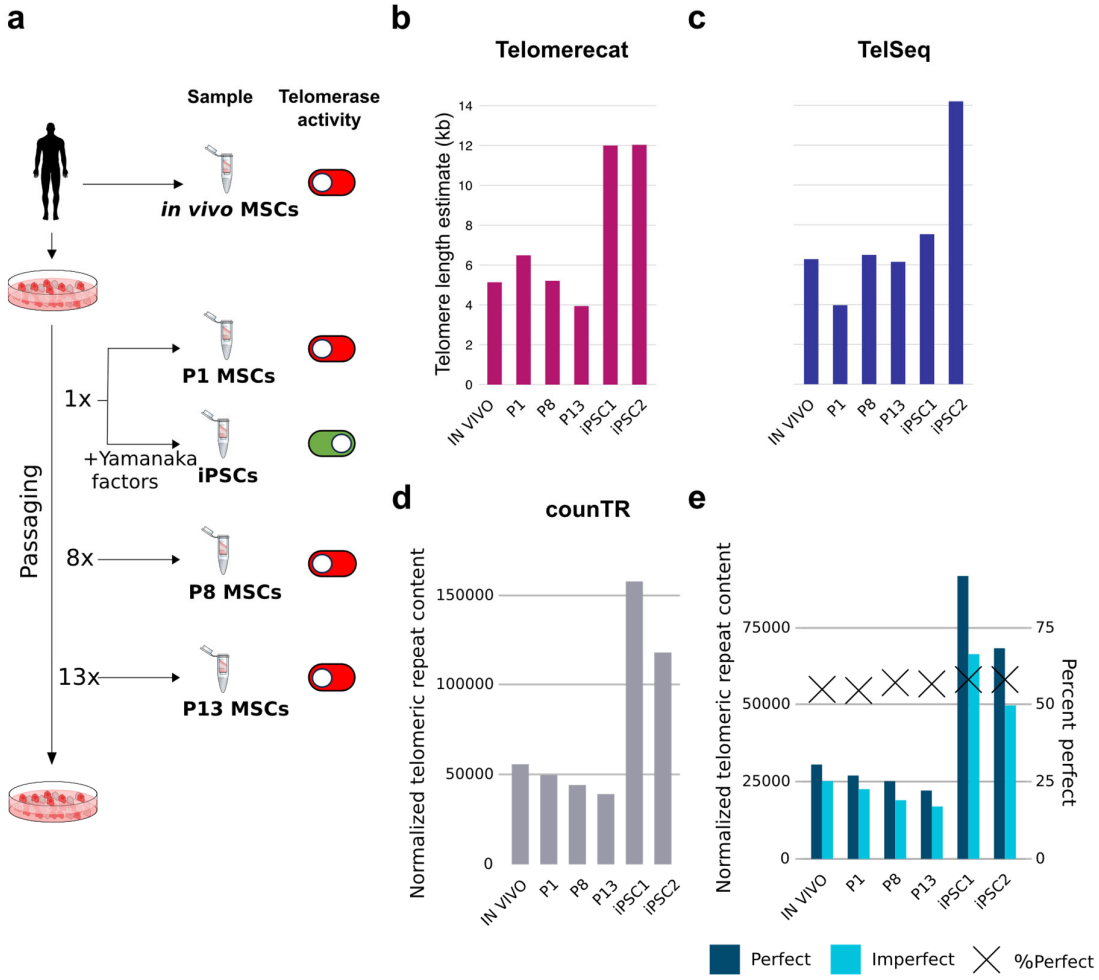


Figure 5.8: counTR accurately quantifies telomere length dynamics. (a) Experiment design for the WGS data generated by Cai et al. (2014), **(b)** Telomerecat and **(c)** TelSeq telomere length quantification for *in vivo*, once (P1), 8-times (P8), and 13-times (P13) passaged MSCs as and iPSCs (iPSC1 and iPSC2) for the WGS data generated by Cai et al. (2014). **(d)** Telomere content quantification for the same samples by counTR analysis. **(e)** counTR analysis with separation of reads with perfect or imperfect telomeric repeats. Panels b and c adapted from Farmery et al. (2018) licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

in the detection of the expected telomere dynamics of shortening telomeres with increasing passage numbers and increased telomere lengths in iPSCs (Figure 5.8b,c). I performed a counTR analysis using the same dataset and observed the expected telomere dynamics in all samples without exception (Figure 5.8d), while Telomerecat failed to detect the expectedly longer telomeres in the *in vivo* MSCs versus 1x passaged MSCs (Figure 5.8b). Moreover, counTR has the ability to separate repeats based on repeat perfection. Figure 5.8e shows the results for the

same dataset with separation of perfect and imperfect telomeric repeats in the 102bp reads. Both, perfect and imperfect telomeric repeat content drop with increasing passaging numbers and increase upon iPSC derivation (Figure 5.8e). Interestingly, the fraction of perfect telomeric repeat content increased both, with increasing passaging numbers, as well as after iPSC derivation (Figure 5.8e). This may suggest that (1) telomere portions that are the product of telomerase elongation have on average lower repeat complexity than native telomeric repeats formed during replication and (2) telomeric repeats towards the chromosome end have higher repeat complexity that is lost upon telomere shortening.

5.2.4 counTR detects telomere-binding of ADAR1 using cortical mouse neuron ChIP-seq data

To demonstrate that counTR can also be effectively applied to ChIP-seq and comparable data and, I investigated its ability to detect protein binding to tandem repeats. For this purpose, I focused on ADAR1, a protein with known roles in genome stability and regulation at telomeres, as a case study. The p110 isoform of ADAR1 has been shown to regulate R-loop formation at telomeres of cancer cell lines (Shiromoto et al., 2021). However, it has thus far not been demonstrated whether any of ADAR1's two isoforms, ADAR1 p110 or p150, possess the ability to directly bind to telomeres.

Marshall et al. (2020) generated an ADAR1 p150 ChIP-seq dataset from active and quiescent cortical mouse neurons derived from fear-conditioned mice that underwent either fear extinction training or were part of a control group. counTR analysis of quiescent neurons from the control group showed an enrichment of telomeric repeats over input control, for perfect and imperfect telomeric repeats of different lengths (Figure 5.9). Interestingly, among the analyzed tandem repeats, imperfect telomeric repeats spanning large portions of reads ("AACCCT <perfection:[0.0,100.0) length:(80.0,inf]>") were the only repeats to show statistically significant positive enrichment under an FDR cutoff of 0.01. This suggests that ADAR1 may specifically bind to telomeric repeats containing imperfections. Moreover, differential telomere-binding of ADAR1 between quiescent

or active neurons of fear extinction-trained or control group mice was also tested but did not yield statistically significant differences (not shown).

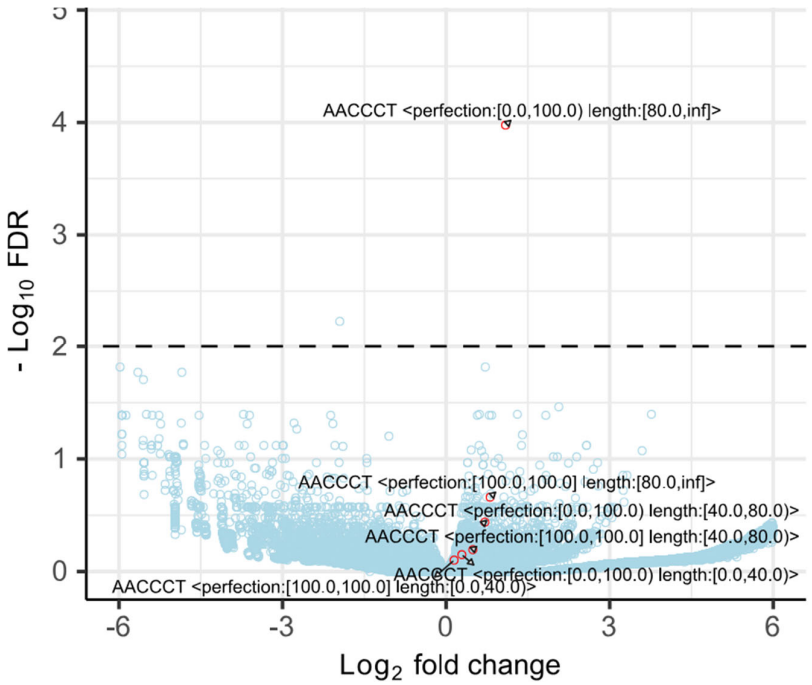


Figure 5.9: counTR analysis finds enriched telomeric repeats in ADAR1 ChIP-seq data from cortical mouse neurons. Volcano plot depicting ADAR1 ChIP-seq counTR analysis results of 126bp reads of mouse cortical neurons versus input control. Telomeric repeat counts are labeled with repeat perfection and length. Dashed line indicates FDR cutoff of 0.01.

5.2.5 Design and validation of human LINE1 subfamily-discriminating PCR primers

In order to study the expression of LINE1s or their enrichment in immunoprecipitation experiments with the help of qPCR experiments, primers targeting LINE1s are required. Since only the evolutionarily most recent LINE1 subfamily L1HS is retrotransposition-competent in humans (Beck et al., 2010) and evolutionary older LINE1 subfamilies are mostly considered to be DNA fossils (Wagstaff et al., 2018), it can be crucial to differentiate between families of different age. Thus, I developed a workflow to design LINE1 primers with a LINE1 subfamily-specific amplification preference (Figure 5.10, Material and methods 6.1.10) and

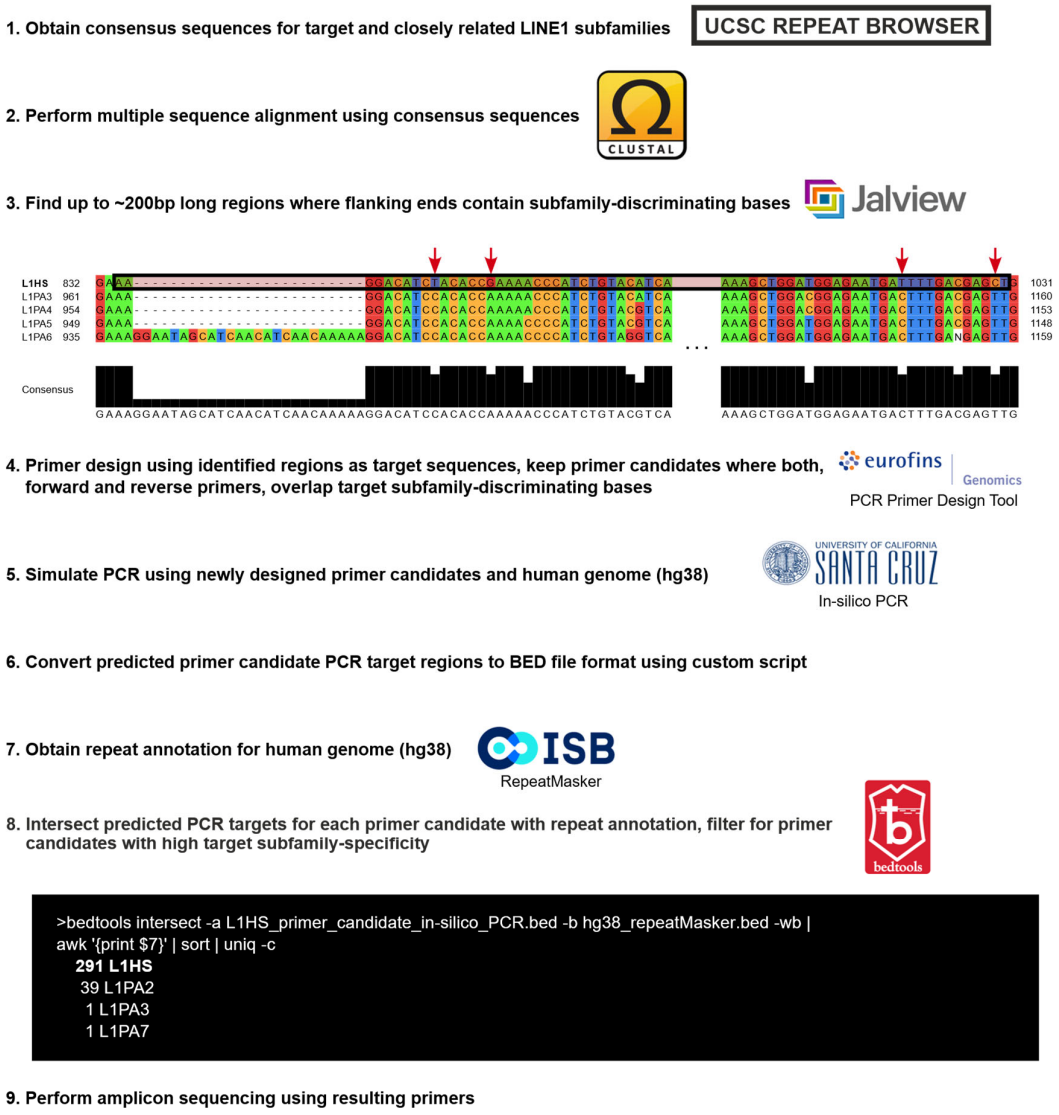


Figure 5.10: Steps taken to design LINE1 subfamily-discriminating PCR primers. Used programs and resources in each step are indicated with their respective logos.

attempted to design primers specific to LINE1 regions matching genomic DNA and mature mRNA regions for subfamilies L1HS, L1PA3, L1PA4, L1PA5, L1PA16 and L1PA17. In brief, I obtained consensus sequences for the LINE1 subfamilies between which I planned to distinguish, performed a multiple sequence alignment and manually inspected resulting alignments for regions up to ~200bp that, near both ends, contain subfamily-discriminating mutations and attempted to design primers that overlap both ends. When such primers could successfully be designed, I performed *in silico* PCR to predict loci of amplification and intersected them with

annotated repeats. If the loci largely overlapped with one LINE1 subfamily of interest, the primers were kept as candidate primers.

For each primer candidate, Amitava Basu (this lab) amplified human genomic DNA from human liver cells and ran the samples on an agarose gel, which produced specific products in the expected size range of roughly 200bp for each primer (Figure 5.11a), and subjected the samples to amplicon sequencing. All primer candidates except for Primer 5 turned out to primarily amplify LINE1 sequences (Figure 5.11b). While the results indicate that none of the primers is specific to a single LINE1 subfamily, primers 1, 2, 3, 6 and 7 predominantly amplified amplicons of either a single or two closely related families to more than ~65% and primer 4 amplifies subfamilies L1PA2, L1PA3, L1PA4, L1PA5 and L1PA6 to 10.1 – 22.7% each (Figure 5.11b). Amitava Basu performed RNA extraction on human liver samples and performed RT-qPCR to measure family-specific LINE1 expression using the primer candidates. The results showed a nearly absent signal for primer 7 and a relatively low signal for primers 1 and 6, whereas Primers 2, 3 and 4 showed comparably high signal (Figure 5.11c). Resulting LINE1 subfamily-specific primer sequences together with intended target regions experimentally determined target regions are listed in Table 5.4.

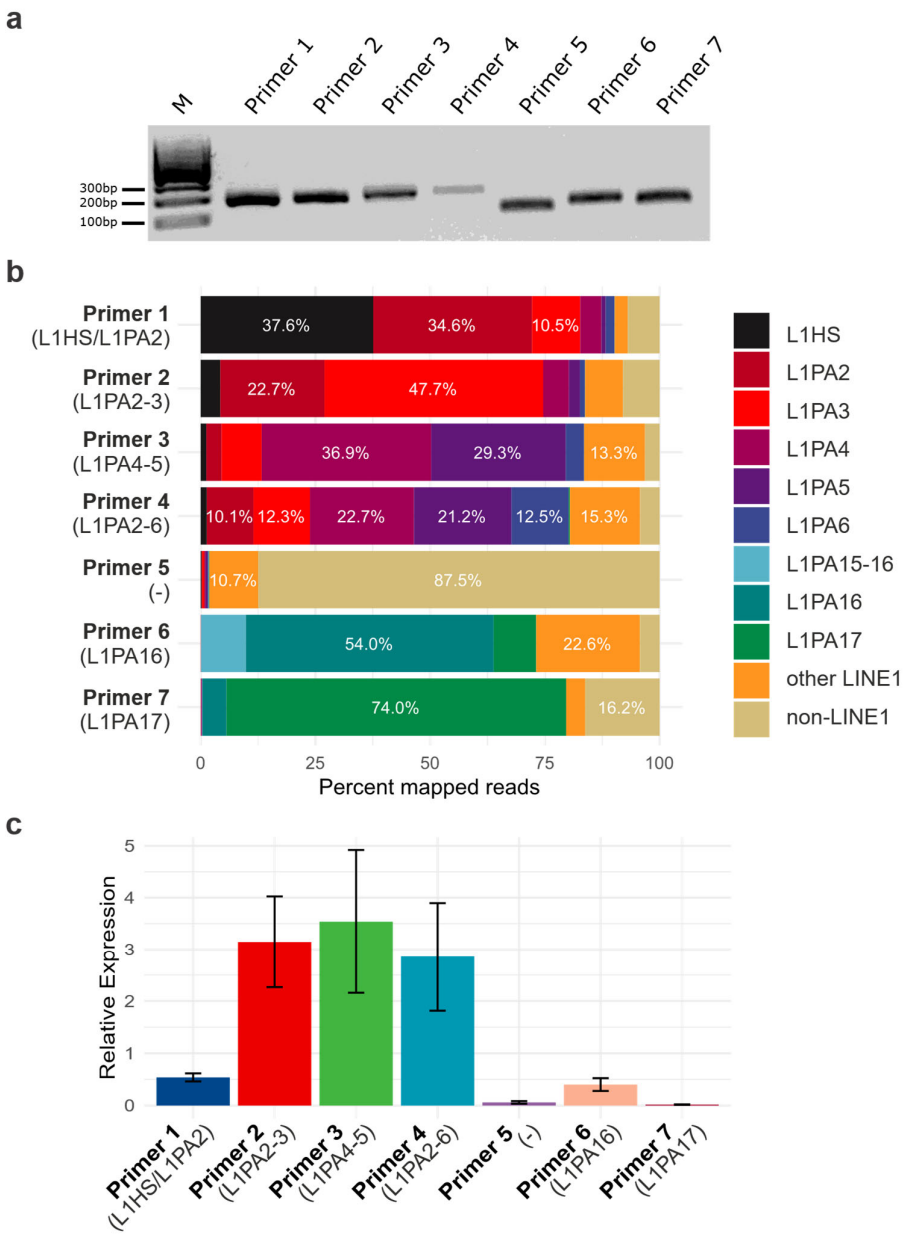


Figure 5.11: Newly designed primers discriminate between human LINE1 families. (a) PCR amplification was performed under standard conditions on human liver genomic DNA using the newly designed primers. Agarose gel showing amplification products with each individual primer. **(b)** Amplicon sequencing was performed on PCR products. Bar chart indicates percentage of all amplified fragments that map to LINE1 and non-LINE1 loci. **(c)** Primers were used to quantify LINE1 expression by RT-qPCR on human liver. Plot depicts measured relative expression normalized to GAPDH.

Table 5.4: LINE1 subfamily-discriminating primers.

Primer name	Intended target	Observed target subfamily(ies)	%GC	Tm	Primer sequence
Primer 1 (F)	L1HS (5'UTR)	L1HS, L1PA2	50.0	50.5	GACATCTACACCGAAAACCC
Primer 1 (R)	L1HS (ORF1)		40.9	50.7	TCGTCAAAATCATTCTCCATCC
Primer 2 (F)	L1PA3 (ORF1)	L1PA2, L1PA3	50.0	50.2	ACCAGCCACTGCAAAATC
Primer 2 (R)	L1PA3 (ORF2)		50.0	51.6	CCAATTTGCCAGTCTGTGTC
Primer 3 (F)	L1PA4 (ORF1)	L1PA4, L1PA5	55.0	54.6	ATGCACAAGCCTCAGTAGCC
Primer 3 (R)	L1PA4 (ORF1)		52.4	53.8	TCCATTCTCCCCGTCACTTTC
Primer 4 (F**)**	L1PA5 (5'UTR)	L1PA2, L1PA3, L1PA4,	52.6	51.7	TCCACACCAAAACCCCATC
Primer 4 (R**)**	L1PA5 (ORF1)	L1PA5, L1PA6	45.5	51.2	CTCGTCAAAGTCATTCTCCATC
Primer 5 (F**)**	L1PA16 (ORF2)	-	50.0	60.9	CCCTCAACAAACTAGGCATC
Primer 5 (R**)**	L1PA16 (ORF2)		50.0	62.8	CTTCCAGCTTTTGCCCATTC
Primer 6 (F**)**	L1PA16 (ORF2)	L1PA16	40.0	57.1	GACAAAGGTGACATTACAAC
Primer 6 (R**)**	L1PA16 (ORF2)		45.0	58.6	CTTGGGAGATTGTGTGTTTC
Primer 7 (F**)**	L1PA17 (ORF2)	L1PA17	45.0	60.0	AGAATGAAACTGGACCCCTA
Primer 7 (R**)**	L1PA17 (ORF2)		40.0	56.9	GTCCAGAAGAGTATTTCCTA

%GC: Percent GC content, Tm: Melting temperature**,** F: Forward primer, R: Reverse primer

5.3 Discussion

5.3.1 counTR is a novel approach to detect differential tandem repeat content in short-read sequencing data

I developed counTR, a novel alignment-free approach to detect enriched tandem repeat content in short-read sequencing data. In brief, counTR works alignment-free by *de novo* identifying tandem repeats inside reads produced via short-read sequencing experiments. Then, similar to RepEnrich (Criscione et al., 2014), detected tandem repeats are counted per repeat unit, which can optionally be grouped by repeat length and perfection. The resulting counts table can be normalized via established read count normalization methods and used for differential tandem repeat content analysis. This method can be suitable for different applications. Here, I explored its applicability to telomere length comparison and the detection of enriched tandem repeats in enrichment-based sequencing data.

Bioinformatics analyses are often carried out on HPC clusters containing hundreds or thousands of CPU cores and even contemporary standard workstation computers often contain between 4 and 8 CPU cores. Hence, it is crucial for well-designed programs that handle relatively large datasets, such as sequencing data, to make efficient use of these hardware resources and optimize run times. I implemented counTR's *processrepeats* function in a manner that instantly parallelizes downstream operations after a chunk of an input file has been read (Figure 5.5). Hence, a roughly linear scalability of the program's run time with increasing amounts of CPU cores can be expected. Indeed, separate counTR *processrepeats* runs with doubling CPU core numbers on the same representative ChIP-seq sample showed a near linear decrease in run time (Figure 5.6). It should, however, be noted that after a certain increase in CPU core numbers the run time performance can be expected to stagnate and potentially even worsen. This is expected to happen as soon as the parallelized functions processing the reads finish faster than a chunk of reads can be read from the input file, leading to idle CPU cores. This can, in turn, be counteracted by reducing the *readchunksize* parameter that controls the number of reads that are read at a time to enter parallel processing (Table 5.1).

5.3.2 Caveats in the usage of counTR and future directions

While counTR is a fast and effective approach for the unbiased detection of tandem repeat-enrichment in short-read sequencing data, users should still be aware of its limits.

Due to lack of read-mapping step filtering reads to the sequenced organism of interest, counTR is unable to discern the origin of reads. Thus, to prevent an overestimation of repeat counts, contamination of the sequenced samples should be ruled out prior analysis. This can, for instance, be accomplished in a complimentary mapping-based analysis with FASTQ Screen (Wingett & Andrews, 2018). Sequencing protocols often contain a PCR step to generate sufficient amounts of DNA for sequencing. This step may amplify certain fragments more than others and thus introduce a bias that cannot be addressed by counTR. Such biases can be prevented if sequencing is performed with either PCR-free protocols or reads are tagged with unique molecular identifiers (UMIs) before sequencing (Y. Fu et al., 2018; Smith et al., 2014). Such UMIs can be used to distinguish PCR duplicates from identical biologically meaningful reads and thus reduce PCR duplicate reads to a single read. However, the use of PCR-free protocols or UMIs is not always feasible and the decision to use such methods needs to be taken prior sequencing, thus many already publicly available sequencing datasets do not employ these strategies. The problem of a PCR duplicate bias in a counTR analysis is comparable to the bias introduced by PCR duplicates in conventional RNA-seq analyses, where their removal is generally considered to do more harm than good due to the removal of biologically meaningful reads (Parekh et al., 2016). Similarly, for counTR analyses, it is advisable to follow standard RNA-seq practices by ignoring PCR duplicate biases as long as PCR duplicate levels are not unusually high. Moreover, unlike RepEnrich or RepEnTools, which can perform differential analysis of all annotated repeats (Choudalakis et al., 2024; Criscione et al., 2014), counTR is limited to tandem repeats. Furthermore, counTR relies on Phobos, which is licensed under a non-open license and is only free of charge for academic use, while for commercial use a Phobos license is required. As for future directions, I have thus far not sufficiently explored counTR's applicability to single-cell or long-read sequencing data, to which it may be adapted.

5.3.3 counTR detects differential telomeric repeat content in WGS data analysis

Differences in telomere length between different WGS samples are reflected in differential telomeric repeat-containing read content. Thus, counTR's ability to compare tandem repeat content between different short-read sequencing data samples can be utilized to estimate a relative change in telomere length if applied specifically to the telomeric repeat in WGS data samples.

By benchmarking counTR's performance on simulated WGS samples, I found that it can reliably identify variations in telomere length. It should be noted, however, that the variability in repeat content among simulated replicates may be lower than in real data, potentially inflating counTR's apparent accuracy. A counTR analysis on telomeric repeats in WGS data of MSC samples with increasing passaging numbers and two iPSC-converted samples showed that counTR is able to reproduce the expected pattern of increasingly shortened telomeres as an effect of increasing passage numbers as well as extended telomeres in iPSC converted cells. Under the assumption that the telomeres of the sequenced samples indeed reflect the expected pattern, counTR outperforms TelSeq and Telomerecat, tools specifically designed to estimate telomere lengths.

5.3.4 Telomere shortening leads to an irreversible loss of mean telomeric repeat complexity

Unlike TelSeq and Telomerecat, to which I compared counTR in the previous chapter, counTR can incorporate repeat perfection into telomeric repeat content analysis. In my analysis, I split counting of telomeric repeat containing reads in a binary manner - in perfect and imperfect ones. It should be noted that for repeats that surpass the read length, such as the telomeric repeat, the relationship between detected imperfect and perfect repeats is impacted by the read length. The longer the read length, the higher the probability to detect an imperfection inside of a read, labeling the whole repeat imperfect. Despite this dependency on read length, the perfection metric can be used to make quantitative comparisons between samples of identical read length. The fraction of perfect telomeric repeats was generally

slightly higher than that of imperfect ones for the 102bp long reads. Interestingly, this fraction increased both, upon telomere shortening due to cell passaging as well as after telomere lengthening via telomerase in iPSCs. This observation can be explained by assuming (1) that the telomeres in *in vivo* MSCs are largely the result of DNA replication by DNA polymerases, rather than telomerase (Maestroni et al., 2017) and (2) the hypothesis that said telomeres are composed of telomeric repeats that become increasingly imperfect towards the direction of the chromosome ends (Figure 5.12). Upon telomere shortening, these less perfect repeat sections may get lost, thereby increasing the fraction of perfect telomeric repeats. Meanwhile, telomere lengthening via telomerase can restore telomere length, but may fail to restore the less perfection sections originally replicated by DNA polymerase-mediated DNA replication.

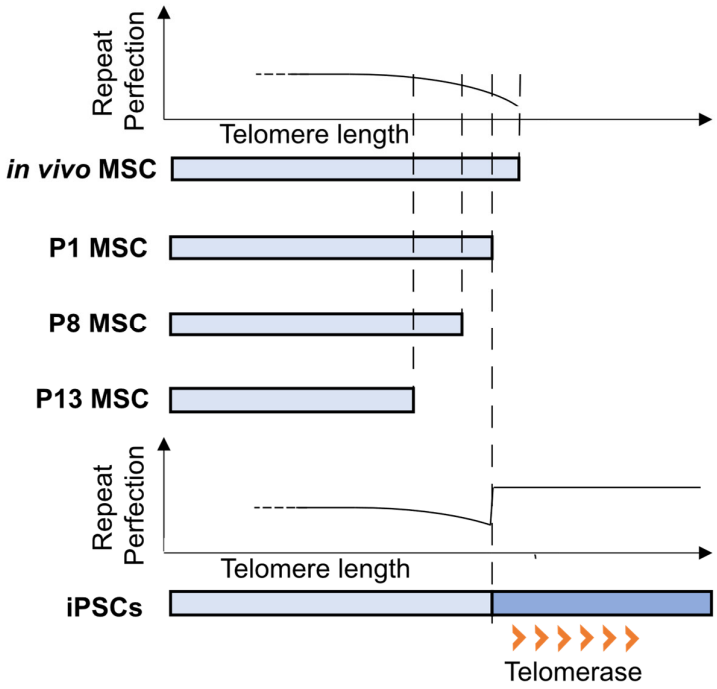


Figure 5.12: Hypothesis explaining the observed telomere dynamics in passaged and iPSC-converted MSCs. Telomeres are represented as bars with the right side of each bar representing the telomere end that at same time forms the end of the chromosome.

Telomere-shortening is an established hallmark of cellular aging and cultivation passaging can be regarded as a model system for the study of aging-related telomere attrition (Pilbauerova et al., 2021). The hypothesis can therefore be expanded to *in vivo* cellular aging which might lead to a loss of more complex telomeric repeat sections that cannot be restored via telomerase activity. Researchers are experimenting with telomerase overexpression as a means to combat aging and disease (Bernardes de Jesus & Blasco, 2013; El Maï et al., 2023). 2023). To my best knowledge, the metric of genomic repeat perfection has largely eluded the attention of researchers. Comparing attributes of rather perfect telomerase-lengthened telomeric repeats to more complex telomeric repeats that resemble DNA polymerase replicated telomeric repeats could thus be interesting for telomere-related studies in the field of aging research.

5.3.5 ADAR1 binds telomeres in cortical mouse neurons

In Shiromoto et al. (2021), the authors showed that ADAR1's p110 isoform regulates the formation of R-loops as well as the genome stability at telomeres in human cancer cell lines by editing A-C mismatches in DNA:RNA hybrids to I-C pairs. These R-loops are thought to be formed by the telomeric repeat-containing RNA (TERRA) that is transcribed from telomeric and subtelomeric regions and associates with the telomeres (Shiromoto et al., 2021). It was thus far, however, not shown whether any of ADAR1 isoforms can directly bind to telomeres. Marshall et al. (2020) generated an ADAR1 p150 ChIP-seq dataset in cortical mouse neurons. The authors had observed an increase of ADAR1 p150 binding to Z-DNA during fear extinction learning and knockdown of ADAR1 p150 led to an inhibition of fear extinction memory formation, whereas reintroduction of the protein rescued these effects (Marshall et al., 2020). I performed a counTR analysis on the same data to detect whether ADAR1 p150 can bind to telomeres in mouse neurons. Analysis of quiescent neurons from the control group resulted in the detection of significantly enriched imperfect telomeric repeats in the ADAR1 p150 ChIP-seq over input control, indicating direct binding of ADAR1 p150 to such imperfect telomeric repeats. Taken together with ADAR1 p110's reported activity on telomeres in human cancer cells (Shiromoto et al., 2021), it can be hypothesized that telomere-

binding of ADAR1 is conserved across species, tissues and ADAR1's p110 and p150 isoforms. At the same time, the results demonstrate that counTR can be successfully applied to detect tandem repeat occupancy of ChIP-seq targets.

5.3.6 Newly designed primers successfully discriminate between LINE1 subfamilies

I developed and applied a workflow to design subfamily-discriminating LINE1 PCR primers. By analysis of amplicon sequencing data generated using these primers, I showed that five out of seven primers (primers 1,2,3 and 6,7) amplified fragments of one or two LINE1 subfamilies to at least 65%. These primers can thus be used to specifically measure qPCR signals for these subfamilies. One primer (primer 5) turned out to amplify mostly non-LINE1 sequences and one primer (primer 4) turned out to have an amplification pattern that is rather balanced between LINE1 subfamilies L1PA2-L1PA6. The history of Old World primates is considered to have started about 21-25 million years ago and is associated with L1PA5-L1PA6 distribution (Khan et al., 2006; Protasova et al., 2021) whereas the earlier subfamilies L1PA2-L1PA5 and the recent L1HS subfamily amplified during the period of ape evolution (Mathews et al., 2003). Primer 4 could thus be used for a combined PCR amplification that is mostly exclusive to all LINE1 subfamilies that amplified exclusively after the evolutionary split into Old World primates, minus the most recent and only still retrotransposition-competent subfamily L1HS (Mathews et al., 2003).

The newly designed primers were then used to measure subfamily-specific LINE1 RNA expression. Subfamilies L1PA2-6 turned out to exhibit comparably high amounts of expression, whereas the combined expression for L1HS/L1PA2 and the expression of L1PA16 was roughly 6-fold lower. Meanwhile, L1PA17 expression was not measurable. The observation that the most recent LINE1 subfamily L1HS had lower expression levels than the older L1PA2-L1PA6 subfamilies is at first glance unexpected, but can be explained by differences in copy number. Each LINE1 subfamily within the L1PA2-L1PA6 group has approximately 4,000-12,000 copies, whereas L1HS comprises roughly 1,600 copies (Jiang et al., 2021; Khan et al., 2006). The mean expression levels of individual LINE1 loci are thus likely

comparable between L1HS and L1PA2-L1PA6 subfamilies and total RNA of the members of the latter group might only be higher due to increased copy numbers. Meanwhile, L1PA16 and L1PA17 are estimated to be present in around 9400 and 3300 copies, respectively (Khan et al., 2006). Thus, even if accounted for copy numbers, these much older subfamilies appear to be expressed to a lower degree, likely due to accumulation of mutations in regulatory elements. While elements belonging to the L1PA2-L1PA6 subfamilies can be expressed and are regulated by DNA-binding Krüppel-associated box domain-containing zinc finger proteins (KRAB-ZFPs) and KRAB-associated protein 1 (KAP1) (Castro-Diaz et al., 2014; Jacobs et al., 2014), their expression levels have to the best of my knowledge not been compared with each other, to L1HS or to other human LINE1 subfamilies.

6 Material and methods

6.1.1 Computational resources

All analyses were carried out either on an Ubuntu 18.04 LTS 64-bit workstation PC with an Intel Core i7-8650U CPU and 16GB of RAM or on an Ubuntu 22.04 LTS 64-bit HPC cluster running the SLURM workload manager consisting of 3 compute nodes with up to 128 Intel Xeon Gold 6252 cores and 2 TB of RAM.

6.1.2 Thesis writing

This thesis was written in Microsoft Word from Microsoft Office Professional Plus 2021. EndNote v.21 21.4.0.18113 was used as a citation manager. ChatGPT 4o and o1-preview (https://chatgpt.com/) were used to assist with the wording of individual sentences and smaller sections.

6.1.3 Chromosome copy number comparison using WGS data

Initial analysis was performed using NGSpipe2go's ChIP-seq pipeline (https://github.com/imbforge/NGSpipe2go) up until read mapping. Raw read quality was assessed with FASTQC v.0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and reads were subsequently mapped to the mm10 mouse reference genome using Bowtie v.2.4.5 (http://bowtie-bio.sourceforge.net/bowtie2) with options "--very-sensitive --end-to-end --fr --maxins 1000". Duplicate reads were removed using MarkDuplicates of Picard v.2.20 (https://broadinstitute.github.io/picard/). Per chromosome read numbers were obtained using idxstats from Samtools v.1.1.0 and used to calculate per chromosome RPKM values.

GADD45 clone set 2: For each of the two pooled WGS samples, each chromosome's RPKM value was normalized to the median RPKM value of all of the sample's chromosomes. Normalized RPKM values for the pooled *Gadd45* TKO clone sample were then divided by the normalized RPKM values for the WT sample.

GADD45 clone set 3: For each chromosome, the median RPKM value out of all samples was calculated. For each sample, each chromosome's RPKM was then divided by this sample-wide per chromosome median RPKM.

6.1.4 RNA-seq analysis

Analysis was performed using NGSpipe2go's RNA-seq pipeline (https://github.com/imbforge/NGSpipe2go). Raw read quality was checked using FASTQC v.0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and reads were aligned to the mm10 mouse reference genome via STAR v.2.7.10a (https://github.com/alexdobin/STAR), secondary read alignments were removed with Samtools v.1.1.0 (http://www.htslib.org). Reads were counted per gene using Subread featureCounts v.2.0.0 (https://subread.sourceforge.net) with stranded parameter "-s 2" and GENCODE GRCm38.p4 gene annotation. DEGs were identified in R v.3.1.4 using DESeq2 v.1.34.0 (https://bioconductor.org/packages/release/bioc/html/DESeq2.html) with independent gene filtering and without LFC shrinkage at a 0.01 FDR cutoff. GO term analysis of DEGs was conducted via ClusterProfiler v.4.8.2 (https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html) using only genes previously statistically tested for differential expression as universe.

6.1.5 ATAC-seq and CUT&Tag analyses

Analyses were performed using NGSpipe2go's ChIP-seq pipeline (https://github.com/imbforge/NGSpipe2go). Raw read quality was assessed with FASTQC v.0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and reads were subsequently mapped to the mm10 mouse reference genome using Bowtie v.2.4.5 (http://bowtie-bio.sourceforge.net/bowtie2) with options "--very-sensitive --end-to-end --fr --maxins 1000". Duplicate reads were removed using MarkDuplicates of Picard v.2.20 (https://broadinstitute.github.io/picard/). Peak calling was performed with MACS2 v.2.1.2 (https://github.com/macs3-project/MACS) and options "--bw 200 --keep-dup auto --broad --format BAMPE"

against IgG samples for CxT experiments and without control for ATAC-seq. Differential peaks were identified in R v.3.1.4 using the DiffBind package v.3.4.0 (https://bioconductor.org/packages/release/bioc/html/DiffBind.html) without control sample subtraction. For ATAC-seq, peaks were re-centered +-100bp around consensus summits, while for CxT experiments, re-centering was kept at the default +-200bp. Significantly differential peaks were identified at a 0.05 FDR cutoff. Motif analysis on differential ATAC-seq peaks was performed using HOMER v4.10 (http://homer.ucsd.edu/homer/motif/).

6.1.6 DRIP-seq analysis

Sequencing data was obtained from NCBI's SRA repository under accession number SRR2075686. Single-end reads were mapped to the mm10 mouse reference genome using Bowtie v.2.4.5 (http://bowtie-bio.sourceforge.net/bowtie2) with option "--very-sensitive". Peak calling was performed with MACS2 v.2.1.2 (https://github.com/macs3-project/MACS) and option "--format BAM" without control sample. Peaks with up to 300bp distance from each other were merged using bedtools merge (https://bedtools.readthedocs.io/).

6.1.7 strDRIP-seq analysis

Initial analysis was performed using NGSpipe2go's ChIP-seq pipeline (https://github.com/imbforge/NGSpipe2go) up until read mapping. Read quality of the raw single-end reads was assessed with FASTQC v.0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and reads were subsequently mapped to the mm10 mouse reference genome using Bowtie v.2.4.5 (http://bowtie-bio.sourceforge.net/bowtie2) with options "--very-sensitive --end-to-end --fr --maxins 1000". Duplicate reads were removed using MarkDuplicates of Picard v.2.20 (https://broadinstitute.github.io/picard/). Plus and minus strand reads were extracted using the "samtools view" command of Samtools v.1.10 with parameters "-F 16" and "-f 16", respectively. Peaks were called separately for the

plus and minus strand using MACS2 v.2.2.6 (https://github.com/macs3-project/MACS) and options “–format BAM --nomodel”.

6.1.8 spKAS-seq analysis

Paired-end reads were adapter trimmed using Trim Galore v. 0.6.10 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) and aligned to the mm10 genome using Bowtie v.2.5.1 (http://bowtie-bio.sourceforge.net/bowtie2). Plus strand read pairs were extracted using the “samtools view” command of Samtools v.1.16.1 with parameters “-b -f 146“ and “-b -f 98“ and merged with “samtools merge“-b -f 146“. Minus strand read pairs were processed analogously, but with “samtools view” parameters “-b -f 130“ and “-b -f 66“ before merging. Peaks were called separately for the plus and minus strand using MACS2 v.2.2.6 (https://github.com/macs3-project/MACS) and option ”–format BAMPE”. Differential peaks were called using a customized DiffBind script using DiffBind package v.3.4.0 (https://bioconductor.org/packages/release/bioc/html/DiffBind.html). The script runs the DiffBind “count” function separately for peaks on the plus and minus strand, then merges the data into a single object before performing differential analysis.

6.1.9 Intersection of genomic regions

Venn diagrams displaying intersection of genomic regions were created using the findOverlapsOfPeaks function of the ChIPpeakAnno v3.6.5 package with default settings. Note that such venn diagram-based illustrations of intersecting genomic loci in multiple datasets can lead to unexpected results, such as intersections between regions with increased and decreased binding. This can happen if, for example, up and down peaks both intersect with a single peak in a third dataset.

6.1.10 Design of human LINE1 subfamily-discriminating PCR primers

Consensus sequences for LINE1 subfamilies L1HS, L1PA3, L1PA4, L1PA5, L1PA6, L1PA15, L1PA16 and L1PA17 were obtained from the UCSC Repeat

Browser (Fernandes et al., 2020). Multiple sequence alignments (MSAs) were performed using Clustal Omega via the EMBL-EBI Job Dispatcher (Madeira et al., 2024) for L1HS, L1PA3, L1PA4, L1PA5, L1PA6 and L1PA14, L1PA15, L1PA16, L1PA17, respectively. Both MSAs were visually inspected for regions of up to ~200bp length that contained LINE1 subfamily-discriminating bases for families of interest at both ends. For such regions, primer design was attempted using Eurofins Genomics' PCR Primer Design Tool (https://eurofinsgenomics.eu/en/ecom/tools/pcr-primer-design) such that both primer pairs overlapped with the respective bases. For regions, for which such candidate primers could be designed successfully, *in silico* PCR (https://genome.ucsc.edu/cgi-bin/hgPcr) was performed using the GRCh38/hg38 genome. Output containing predicted *in silico* PCR amplified loci was converted to bed file format using a custom script and then intersected with annotated repeats in the hg38 4.0.5 RepeatMasker annotation (https://www.repeatmasker.org/faq.html). If the loci largely overlapped with one LINE1 subfamily of interest, primers with the according sequence were ordered for experimental validation.

6.1.11 Validation LINE1 subfamily-discriminating PCR primers

Amplicon sequencing was performed on PCR products generated using the primers. Raw read quality was assesed with FASTQC v.0.11.8 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and reads were subsequently mapped to the hg38 human reference genome using Bowtie v.2.3.4 (http://bowtie-bio.sourceforge.net/bowtie2) with options “--very-sensitive --end-to-end --fr --maxins 1000”. Read counting was performed using Subread featureCounts v.2.0.0 (https://subread.sourceforge.net/). All properly paired reads that overlapped with either a LINE1 family or non-LINE1 genomic loci were counted. Multi-mappers were counted once as long as they only map to a single family.

6.1.12 counTR implementation, usage and program availability

counTR is written in Python (tested on Python 3.9.7) using only standard Python libraries that come preinstalled with a default Python installation. Detection of tandem repeat containing reads relies on Phobos (https://www.ruhr-uni-bochum.de/spezzoo/cm/cm_phobos_download.htm) (Mayer, 2010) (tested using phobos_64_libstc++6 of the phobos-v3.3.12-linux package). At the time of writing, counTR only supports POSIX operating systems and was tested on Ubuntu 18.04 LTS 64-bit and Ubuntu 22.04 LTS 64-bit. counTR is available under https://github.com/mmisak/counTR and licensed under the GPL 3.0 license.

6.1.13 Benchmark of counTR's scalability in multicore CPU environments

counTR was run with parameters "--grouping 'perfection:[0,100)[100,100] length:[0,40),[40,80),[80,inf]' --outputtype c" on a HPC with 1, 2, 4, 8, 16, 32, 64 CPU cores and a constant 200GB of RAM on the FASTQ file available from NCBI's SRA repository under accession number SRR9140841.

6.1.14 Telomere expansion WGS simulation

The FASTA file for the human T2T-CHM13v2.0 genome was obtained from NCBI (https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_009914755.1). For T2T-chr4, chromosome 4 of the genome assembly was extracted. For T2T-chr4Δtelo and T2T-chr4ΔDNAJA1, T2T-chr4 was expanded at both ends using the cat command available in POSIX systems as illustrated in Figure 5.7a. In brief, for T2T-chr4Δtelo, 5kb of perfect AACCCT tandem repeat was added before the start of T2T-chr4 and 5kb of perfect AGGGTT to its end. For T2T-chr4ΔDNAJA1, the first 5kb of the 5' end of the human DNAJA1 gene was extracted and appended to the end of T2T-chr4. The same extracted sequence was reverse complemented and added before the start of the sequence. For short-read WGS simulation, wgsim (part of samtools 1.10) was run three times using parameters "-1 150 -2 150 -r 0 -R 0 -X 0 -e 0 -N 4000000" to generate triplicates of 4M 150bp paired-end short-read samples. counTR processrepeats was run for all R1 files of the simulated short-read

sequencing with parameters: “--grouping 'length:[0,40),[40,100),[100,inf]' --outputtype c --groupingmotif combine”. Downstream analysis was performed using edgeR as described below.

6.1.15 Telomere length dynamics analysis

Sequencing data for *in vivo* MSCs, passaged MSCs as well as MSC-derived iPSCs was obtained from NCBI’s SRA repository under accession number SRP032359. counTR processrepeats was run for each sample’s R1 file with parameters “--groupingmotif combine --outputtype c” and either “--grouping 'perfection:[0,100)[100,100] length:[0,40),[40,80),[80,inf]'” or “--grouping 'length:[0,40),[40,80),[80,inf]'” to perform runs with or without perfection grouping, respectively. Downstream analysis was performed using edgeR as described below.

6.1.16 ADAR1 ChIP-seq analysis in mouse cortical tissue

All ADAR1 ChIP-seq data as well as the corresponding input sample were obtained from the NCBI SRA repository under accession number SRP199704. counTR processrepeats was run for R1 files of each sample with parameters “--grouping 'perfection:[0,100)[100,100] length:[0,40),[40,80),[80,inf]' --outputtype c --groupingmotif combine”. Downstream analysis was performed using edgeR as described below.

6.1.17 edgeR downstream analysis of counTR output

Downstream analysis was performed using edgeR v.3.42.4 (https://bioconductor.org/packages/release/bioc/html/edgeR.html), repeats with fewer than 0.1 CPM were removed from the analysis, normalization factors were calculated using the TMM method and the robust dispersion estimation function estimateGLMRobustDisp was used, read counts were finally CPM normalized

before a general linear model was fit using the glmFit function and differential repeat counts were detected using the glmLRT function.

7 References

Arab, K., Karaulanov, E., Musheev, M., Trnka, P., Schäfer, A., Grummt, I., & Niehrs, C. (2019). GADD45A binds R-loops and recruits TET1 to CpG island promoters. *Nat Genet*, *51*(2), 217-223. https://doi.org/10.1038/s41588-018-0306-6

Arning, L., & Nguyen, H. P. (2021). Huntington disease update: new insights into the role of repeat instability in disease pathogenesis. *Med Genet*, *33*(4), 293-300. https://doi.org/10.1515/medgen-2021-2101

Bahlo, M., Bennett, M. F., Degorski, P., Tankard, R. M., Delatycki, M. B., & Lockhart, P. J. (2018). Recent advances in the detection of repeat expansions with short-read next-generation sequencing. *F1000Res*, *7*. https://doi.org/10.12688/f1000research.13980.1

Baldwin, E. T., van Eeuwen, T., Hoyos, D., Zalevsky, A., Tchesnokov, E. P., Sánchez, R., Miller, B. D., Di Stefano, L. H., Ruiz, F. X., Hancock, M., Işik, E., Mendez-Dorantes, C., Walpole, T., Nichols, C., Wan, P., Riento, K., Halls-Kass, R., Augustin, M., Lammens, A.,…Taylor, M. S. (2024). Structures, functions and adaptations of the human LINE-1 ORF2 protein. *Nature*, *626*(7997), 194-206. https://doi.org/10.1038/s41586-023-06947-z

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., & Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, *129*(4), 823-837. https://doi.org/10.1016/j.cell.2007.05.009

Bartosovic, M., Kabbe, M., & Castelo-Branco, G. (2021). Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. *Nat Biotechnol*, *39*(7), 825-835. https://doi.org/10.1038/s41587-021-00869-9

Beck, C. R., Collier, P., Macfarlane, C., Malig, M., Kidd, J. M., Eichler, E. E., Badge, R. M., & Moran, J. V. (2010). LINE-1 retrotransposition activity in human genomes. *Cell*, *141*(7), 1159-1170. https://doi.org/10.1016/j.cell.2010.05.021

Ben-David, U., & Benvenisty, N. (2012). High prevalence of evolutionarily conserved and species-specific genomic aberrations in mouse pluripotent stem cells. *Stem Cells*, *30*(4), 612-622. https://doi.org/10.1002/stem.1057

Bernardes de Jesus, B., & Blasco, M. A. (2013). Telomerase at the intersection of cancer and aging. *Trends Genet*, *29*(9), 513-520. https://doi.org/10.1016/j.tig.2013.06.007

Bilen, S., Okten, A., Karaguzel, G., Ikbal, M., & Aslan, Y. (2013). A 45 X male patient with 7q distal deletion and rearrangement with SRY gene translocation: a case report. *Genet Couns*, *24*(3), 299-305. https://www.ncbi.nlm.nih.gov/pubmed/24341145

Bodak, M., Yu, J., & Ciaudo, C. (2014). Regulation of LINE-1 in mammals. *Biomol Concepts*, *5*(5), 409-428. https://doi.org/10.1515/bmc-2014-0018

Boissinot, S., & Sookdeo, A. (2016). The Evolution of LINE-1 in Vertebrates. *Genome Biol Evol*, *8*(12), 3485-3507. https://doi.org/10.1093/gbe/evw247

Booth, M. J., Ost, T. W., Beraldi, D., Bell, N. M., Branco, M. R., Reik, W., & Balasubramanian, S. (2013). Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. *Nat Protoc*, *8*(10), 1841-1851. https://doi.org/10.1038/nprot.2013.115

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., Mager, D. L., & Feschotte, C. (2018). Ten things you should know about transposable elements. *Genome Biol*, *19*(1), 199. https://doi.org/10.1186/s13059-018-1577-z

Bulavin, D. V., Kovalsky, O., Hollander, M. C., & Fornace, A. J., Jr. (2003). Loss of oncogenic H-ras-induced cell cycle arrest and p38 mitogen-activated protein kinase activation by disruption of Gadd45a. *Mol Cell Biol*, *23*(11), 3859-3871. https://doi.org/10.1128/MCB.23.11.3859-3871.2003

Cai, J., Miao, X., Li, Y., Smith, C., Tsang, K., Cheng, L., & Wang, Q. F. (2014). Whole-genome sequencing identifies genetic variances in culture-expanded human mesenchymal stem cells. *Stem Cell Reports*, *3*(2), 227-233. https://doi.org/10.1016/j.stemcr.2014.05.019

Carter, M. E., & Brunet, A. (2007). FOXO transcription factors. *Curr Biol*, *17*(4), R113-114. https://doi.org/10.1016/j.cub.2007.01.008

Casari, E., Gnugnoli, M., Rinaldi, C., Pizzul, P., Colombo, C. V., Bonetti, D., & Longhese, M. P. (2022). To Fix or Not to Fix: Maintenance of Chromosome Ends Versus Repair of DNA Double-Strand Breaks. *Cells*, *11*(20). https://doi.org/10.3390/cells11203224

Castro-Diaz, N., Ecco, G., Coluccio, A., Kapopoulou, A., Yazdanpanah, B., Friedli, M., Duc, J., Jang, S. M., Turelli, P., & Trono, D. (2014). Evolutionally dynamic L1 regulation in embryonic stem cells. *Genes Dev*, *28*(13), 1397-1409. https://doi.org/10.1101/gad.241661.114

Chebly, A., Ropio, J., Baldasseroni, L., Prochazkova-Carlotti, M., Idrissi, Y., Ferrer, J., Farra, C., Beylot-Barry, M., Merlio, J. P., & Chevret, E. (2022). Telomeric Repeat-Containing RNA (TERRA): A Review of the Literature and First Assessment in Cutaneous T-Cell Lymphomas. *Genes (Basel)*, *13*(3). https://doi.org/10.3390/genes13030539

Chédin, F., Hartono, S. R., Sanz, L. A., & Vanoosthuyse, V. (2021). Best practices for the visualization, mapping, and manipulation of R-loops. *EMBO J*, *40*(4), e106394. https://doi.org/10.15252/embj.2020106394

Chen, J. Y., Zhang, X., Fu, X. D., & Chen, L. (2019). R-ChIP for genome-wide mapping of R-loops by using catalytically inactive RNASEH1. *Nat Protoc*, *14*(5), 1661-1685. https://doi.org/10.1038/s41596-019-0154-6

Chen, Y. (2019). The structural biology of the shelterin complex. *Biol Chem*, *400*(4), 457-466. https://doi.org/10.1515/hsz-2018-0368

Cheng, J., Blum, R., Bowman, C., Hu, D., Shilatifard, A., Shen, S., & Dynlacht, B. D. (2014). A role for H3K4 monomethylation in gene repression and partitioning of chromatin readers. *Mol Cell*, *53*(6), 979-992. https://doi.org/10.1016/j.molcel.2014.02.032

Cheng, S., Miao, B., Li, T., Zhao, G., & Zhang, B. (2024). Review and Evaluate the Bioinformatics Analysis Strategies of ATAC-seq and CUT&Tag Data. *Genomics Proteomics Bioinformatics*, *22*(3). https://doi.org/10.1093/gpbjnl/qzae054

Choudalakis, M., Bashtrykov, P., & Jeltsch, A. (2024). RepEnTools: an automated repeat enrichment analysis package for ChIP-seq data reveals hUHRF1 Tandem-Tudor domain enrichment in young repeats. *Mob DNA*, *15*(1), 6. https://doi.org/10.1186/s13100-024-00315-y

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., Boyer, L. A., Young, R. A., & Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A*, *107*(50), 21931-21936. https://doi.org/10.1073/pnas.1016071107

Crippa, S., Cassano, M., Messina, G., Galli, D., Galvez, B. G., Curk, T., Altomare, C., Ronzoni, F., Toelen, J., Gijsbers, R., Debyser, Z., Janssens, S., Zupan, B., Zaza, A., Cossu, G., & Sampaolesi, M. (2011). miR669a and miR669q prevent skeletal muscle differentiation in postnatal cardiac progenitors. *J Cell Biol*, *193*(7), 1197-1212. https://doi.org/10.1083/jcb.201011099

Criscione, S. W., Zhang, Y., Thompson, W., Sedivy, J. M., & Neretti, N. (2014). Transcriptional landscape of repetitive elements in normal and cancer human cells. *BMC Genomics*, *15*, 583. https://doi.org/10.1186/1471-2164-15-583

Cysewski, P., & Czeleń, P. (2009). Structural and energetic heterogeneities of canonical and oxidized central guanine triad of B-DNA telomeric fragments. *J Mol Model*, *15*(6), 607-613. https://doi.org/10.1007/s00894-008-0438-1

Dai, Q., Ye, C., Irkliyenko, I., Wang, Y., Sun, H. L., Gao, Y., Liu, Y., Beadell, A., Perea, J., Goel, A., & He, C. (2024). Ultrafast bisulfite sequencing detection of 5-methylcytosine in DNA and RNA. *Nat Biotechnol*. https://doi.org/10.1038/s41587-023-02034-w

Davletgildeeva, A. T., & Kuznetsov, N. A. (2024). The Role of DNMT Methyltransferases and TET Dioxygenases in the Maintenance of the DNA Methylation Level. *Biomolecules*, *14*(9). https://doi.org/10.3390/biom14091117

Decarpentrie, F., Ojarikre, O. A., Mitchell, M. J., & Burgoyne, P. S. (2016). Recombination between the mouse Y chromosome short arm and an additional Y short arm-derived chromosomal segment attached distal to the X chromosome PAR. *Chromosoma*, *125*(2), 177-188. https://doi.org/10.1007/s00412-015-0559-0

Deschamps-Francoeur, G., Simoneau, J., & Scott, M. S. (2020). Handling multi-mapped reads in RNA-seq. *Comput Struct Biotechnol J*, *18*, 1569-1576. https://doi.org/10.1016/j.csbj.2020.06.014

Deshpande, D., Chhugani, K., Chang, Y., Karlsberg, A., Loeffler, C., Zhang, J., Muszynska, A., Munteanu, V., Yang, H., Rotman, J., Tao, L., Balliu, B., Tseng, E., Eskin, E., Zhao, F., Mohammadi, P., P, P. Ł., & Mangul, S. (2023). RNA-seq data science: From raw data to effective interpretation. *Front Genet*, *14*, 997383. https://doi.org/10.3389/fgene.2023.997383

Ding, Z., Mangino, M., Aviv, A., Spector, T., Durbin, R., & Consortium, U. K. (2014). Estimating telomere length from whole genome sequence data. *Nucleic Acids Res*, *42*(9), e75. https://doi.org/10.1093/nar/gku181

Eder, T., & Grebien, F. (2022). Comprehensive assessment of differential ChIP-seq tools guides optimal algorithm selection. *Genome Biol*, *23*(1), 119. https://doi.org/10.1186/s13059-022-02686-y

Ekblom, R., & Wolf, J. B. (2014). A field guide to whole-genome sequencing, assembly and annotation. *Evol Appl*, *7*(9), 1026-1042. https://doi.org/10.1111/eva.12178

El Maï, M., Bird, M., Allouche, A., Targen, S., Şerifoğlu, N., Lopes-Bastos, B., Guigonis, J. M., Kang, D., Pourcher, T., Yue, J. X., & Ferreira, M. G. (2023). Gut-specific telomerase expression counteracts systemic aging in telomerase-deficient zebrafish. *Nat Aging*, *3*(5), 567-584. https://doi.org/10.1038/s43587-023-00401-5

Eslami Rasekh, M., Hernández, Y., Drinan, S. D., Fuxman Bass, J. I., & Benson, G. (2021). Genome-wide characterization of human minisatellite VNTRs: population-specific alleles and gene expression differences. *Nucleic Acids Res*, *49*(8), 4308-4324. https://doi.org/10.1093/nar/gkab224

Farmery, J. H. R., Smith, M. L., Diseases, N. B.-R., & Lynch, A. G. (2018). Telomerecat: A ploidy-agnostic method for estimating telomere length from whole genome sequencing data. *Sci Rep*, *8*(1), 1300. https://doi.org/10.1038/s41598-017-14403-y

Fazzio, T. G. (2016). Regulation of chromatin structure and cell fate by R-loops. *Transcription*, *7*(4), 121-126. https://doi.org/10.1080/21541264.2016.1198298

Fernandes, J. D., Zamudio-Hurtado, A., Clawson, H., Kent, W. J., Haussler, D., Salama, S. R., & Haeussler, M. (2020). The UCSC repeat browser allows discovery and visualization of evolutionary conflict across repeat families. *Mob DNA*, *11*, 13. https://doi.org/10.1186/s13100-020-00208-w

Feuerbach, L., Sieverling, L., Deeg, K. I., Ginsbach, P., Hutter, B., Buchhalter, I., Northcott, P. A., Mughal, S. S., Chudasama, P., Glimm, H., Scholl, C., Lichter, P., Frohling, S., Pfister, S. M., Jones, D. T. W., Rippe, K., & Brors, B. (2019). TelomereHunter - in silico estimation of telomere content and composition from cancer genomes. *BMC Bioinformatics*, *20*(1), 272. https://doi.org/10.1186/s12859-019-2851-0

Freen-van Heeren, J. J. (2021). Flow-FISH as a Tool for Studying Bacteria, Fungi and Viruses. *BioTech (Basel)*, *10*(4). https://doi.org/10.3390/biotech10040021

Fu, S., Wang, Q., Moore, J. E., Purcaro, M. J., Pratt, H. E., Fan, K., Gu, C., Jiang, C., Zhu, R., Kundaje, A., Lu, A., & Weng, Z. (2018). Differential analysis of chromatin accessibility and histone modifications for predicting mouse developmental enhancers. *Nucleic Acids Res*, *46*(21), 11184-11201. https://doi.org/10.1093/nar/gky753

Fu, Y., Wu, P. H., Beane, T., Zamore, P. D., & Weng, Z. (2018). Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. *BMC Genomics*, *19*(1), 531. https://doi.org/10.1186/s12864-018-4933-1

Genovese, L. M., Geraci, F., Corrado, L., Mangano, E., D'Aurizio, R., Bordoni, R., Severgnini, M., Manzini, G., De Bellis, G., D'Alfonso, S., & Pellegrini, M. (2018). A Census of Tandemly Repeated Polymorphic Loci in Genic Regions Through the Comparative Integration of Human Genome Assemblies. *Front Genet*, *9*, 155. https://doi.org/10.3389/fgene.2018.00155

Ginno, P. A., Lott, P. L., Christensen, H. C., Korf, I., & Chédin, F. (2012). R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. *Mol Cell*, *45*(6), 814-825. https://doi.org/10.1016/j.molcel.2012.01.017

Glaser, L. V., Steiger, M., Fuchs, A., van Bömmel, A., Einfeldt, E., Chung, H. R., Vingron, M., & Meijsing, S. H. (2021). Assessing genome-wide dynamic changes in enhancer activity during early mESC differentiation by FAIRE-STARR-seq. *Nucleic Acids Res*, *49*(21), 12178-12195. https://doi.org/10.1093/nar/gkab1100

Gondane, A., & Itkonen, H. M. (2023). Revealing the History and Mystery of RNA-Seq. *Curr Issues Mol Biol*, *45*(3), 1860-1874. https://doi.org/10.3390/cimb45030120

Goodier, J. L., Zhang, L., Vetter, M. R., & Kazazian, H. H., Jr. (2007). LINE-1 ORF1 protein localizes in stress granules with other RNA-binding proteins, including components of RNA interference RNA-induced silencing complex. *Mol Cell Biol*, *27*(18), 6469-6483. https://doi.org/10.1128/MCB.00332-07

Gross, D. S., Chowdhary, S., Anandhakumar, J., & Kainth, A. S. (2015). Chromatin. *Curr Biol*, *25*(24), R1158-1163. https://doi.org/10.1016/j.cub.2015.10.059

Grosselin, K., Durand, A., Marsolier, J., Poitou, A., Marangoni, E., Nemati, F., Dahmani, A., Lameiras, S., Reyal, F., Frenoy, O., Pousse, Y., Reichen, M., Woolfe, A., Brenan, C., Griffiths, A. D., Vallot, C., & Gérard, A. (2019). High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nat Genet*, *51*(6), 1060-1066. https://doi.org/10.1038/s41588-019-0424-9

Gupta, M., Gupta, S. K., Hoffman, B., & Liebermann, D. A. (2006). Gadd45a and Gadd45b protect hematopoietic cells from UV-induced apoptosis via distinct signaling pathways, including p38 activation and JNK inhibition. *J Biol Chem*, *281*(26), 17552-17558. https://doi.org/10.1074/jbc.M600950200

Haberle, V., & Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. *Nat Rev Mol Cell Biol*, *19*(10), 621-637. https://doi.org/10.1038/s41580-018-0028-8

Haque, A., Engel, J., Teichmann, S. A., & Lonnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med*, *9*(1), 75. https://doi.org/10.1186/s13073-017-0467-4

He, S., Sun, H., Lin, L., Zhang, Y., Chen, J., Liang, L., Li, Y., Zhang, M., Yang, X., Wang, X., Wang, F., Zhu, F., Chen, J., Pei, D., & Zheng, H. (2017). Passive DNA demethylation preferentially up-regulates pluripotency-related genes and facilitates the generation of induced pluripotent stem cells. *J Biol Chem*, *292*(45), 18542-18555. https://doi.org/10.1074/jbc.M117.810457

Hegazy, Y. A., Fernando, C. M., & Tran, E. J. (2020). The balancing act of R-loop biology: The good, the bad, and the ugly. *J Biol Chem*, *295*(4), 905-913. https://doi.org/10.1074/jbc.REV119.011353

Heyward, F. D., & Sweatt, J. D. (2015). DNA Methylation in Memory Formation: Emerging Insights. *Neuroscientist*, *21*(5), 475-489. https://doi.org/10.1177/1073858415579635

Hinchie, A. M., Sanford, S. L., Loughridge, K. E., Sutton, R. M., Parikh, A. H., Gil Silva, A. A., Sullivan, D. I., Chun-On, P., Morrell, M. R., McDyer, J. F., Opresko, P. L., & Alder, J. K. (2024). A persistent variant telomere sequence in a human pedigree. *Nat Commun*, *15*(1), 4681. https://doi.org/10.1038/s41467-024-49072-9

Hrdlickova, R., Toloue, M., & Tian, B. (2017). RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip Rev RNA*, *8*(1). https://doi.org/10.1002/wrna.1364

Hua, X., Zhou, H., Wu, H. C., Furnari, J., Kotidis, C. P., Rabadan, R., Genkinger, J. M., Bruce, J. N., Canoll, P., Santella, R. M., & Zhang, Z. (2024). Tumor detection by analysis of both symmetric- and hemi-methylation of plasma cell-free DNA. *Nat Commun*, *15*(1), 6113. https://doi.org/10.1038/s41467-024-50471-1

Huang, M., Wang, J., Liu, W., & Zhou, H. (2024). Advances in the role of the GADD45 family in neurodevelopmental, neurodegenerative, and neuropsychiatric disorders. *Front Neurosci*, *18*, 1349409. https://doi.org/10.3389/fnins.2024.1349409

Ibn-Salem, J., & Andrade-Navarro, M. A. (2019). 7C: Computational Chromosome Conformation Capture by Correlation of ChIP-seq at CTCF motifs. *BMC Genomics*, *20*(1), 777. https://doi.org/10.1186/s12864-019-6088-0

Inoue, K., Hirose, M., Inoue, H., Hatanaka, Y., Honda, A., Hasegawa, A., Mochida, K., & Ogura, A. (2017). The Rodent-Specific MicroRNA Cluster within the Sfmbt2 Gene Is Imprinted and Essential for Placental Development. *Cell Rep*, *19*(5), 949-956. https://doi.org/10.1016/j.celrep.2017.04.018

Iwahana, T., Okada, S., Kanda, M., Oshima, M., Iwama, A., Matsumiya, G., & Kobayashi, Y. (2020). Novel myocardial markers GADD45G and NDUFS5 identified by RNA-sequencing predicts left ventricular reverse remodeling in advanced non-ischemic heart failure: a retrospective cohort study. *BMC Cardiovasc Disord*, *20*(1), 116. https://doi.org/10.1186/s12872-020-01396-2

Jacobs, F. M., Greenberg, D., Nguyen, N., Haeussler, M., Ewing, A. D., Katzman, S., Paten, B., Salama, S. R., & Haussler, D. (2014). An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature*, *516*(7530), 242-245. https://doi.org/10.1038/nature13760

Jiang, J. C., Rothnagel, J. A., & Upton, K. R. (2021). Widespread Exaptation of L1 Transposons for Transcription Factor Binding in Breast Cancer. *Int J Mol Sci*, *22*(11). https://doi.org/10.3390/ijms22115625

Jiang, L., Huang, L., & Jiang, W. (2024). H3K27me3-mediated epigenetic regulation in pluripotency maintenance and lineage differentiation. *Cell Insight*, *3*(4), 100180. https://doi.org/10.1016/j.cellin.2024.100180

Jiménez, B. P., Kayser, M., & Vidaki, A. (2021). Revisiting genetic artifacts on DNA methylation microarrays exposes novel biological implications. *Genome Biol*, *22*(1), 274. https://doi.org/10.1186/s13059-021-02484-y

Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science*, *316*(5830), 1497-1502. https://doi.org/10.1126/science.1141319

Jung, H. J., Kim, H. L., Kim, Y. J., Weon, J. I., & Seo, Y. R. (2013). A novel chemopreventive mechanism of selenomethionine: enhancement of APE1 enzyme activity via a Gadd45a, PCNA and APE1 protein complex that regulates p53-mediated base excision repair. *Oncol Rep*, *30*(4), 1581-1586. https://doi.org/10.3892/or.2013.2613

Jurkowska, R. Z., & Jeltsch, A. (2010). Silencing of gene expression by targeted DNA methylation: concepts and approaches. *Methods Mol Biol*, *649*, 149-161. https://doi.org/10.1007/978-1-60761-753-2_9

Kaya-Okur, H. S., Wu, S. J., Codomo, C. A., Pledger, E. S., Bryson, T. D., Henikoff, J. G., Ahmad, K., & Henikoff, S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat Commun*, *10*(1), 1930. https://doi.org/10.1038/s41467-019-09982-5

Khan, H., Smit, A., & Boissinot, S. (2006). Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Res*, *16*(1), 78-87. https://doi.org/10.1101/gr.4001406

Kim, A., & Wang, G. G. (2021). R-loop and its functions at the regulatory interfaces between transcription and (epi)genome. *Biochim Biophys Acta Gene Regul Mech*, *1864*(11-12), 194750. https://doi.org/10.1016/j.bbagrm.2021.194750

Kolovos, P., Knoch, T. A., Grosveld, F. G., Cook, P. R., & Papantonis, A. (2012). Enhancers and silencers: an integrated and simple model for their function. *Epigenetics Chromatin*, *5*(1), 1. https://doi.org/10.1186/1756-8935-5-1

Koopman, P., Sinclair, A., & Lovell-Badge, R. (2016). Of sex and determination: marking 25 years of Randy, the sex-reversed mouse. *Development*, *143*(10), 1633-1637. https://doi.org/10.1242/dev.137372

Kovalsky, O., Lung, F. D., Roller, P. P., & Fornace, A. J., Jr. (2001). Oligomerization of human Gadd45a protein. *J Biol Chem*, *276*(42), 39330-39339. https://doi.org/10.1074/jbc.M105115200

Kubo, N., Chen, P. B., Hu, R., Ye, Z., Sasaki, H., & Ren, B. (2024). H3K4me1 facilitates promoter-enhancer interactions and gene activation during embryonic stem cell differentiation. *Mol Cell*, *84*(9), 1742-1752 e1745. https://doi.org/10.1016/j.molcel.2024.02.030

Kumar, S., Chinnusamy, V., & Mohapatra, T. (2018). Epigenetics of Modified DNA Bases: 5-Methylcytosine and Beyond. *Front Genet*, *9*, 640. https://doi.org/10.3389/fgene.2018.00640

Lanciano, S., Philippe, C., Sarkar, A., Pratella, D., Domrane, C., Doucet, A. J., van Essen, D., Saccani, S., Ferry, L., Defossez, P. A., & Cristofari, G. (2024). Locus-level L1 DNA methylation profiling reveals the epigenetic and transcriptional interplay between L1s and their integration sites. *Cell Genom*, *4*(2), 100498. https://doi.org/10.1016/j.xgen.2024.100498

Latchman, D. S. (1996). Inhibitory transcription factors. *Int J Biochem Cell Biol*, *28*(9), 965-974. https://doi.org/10.1016/1357-2725(96)00039-8

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol*, *15*(2), R29. https://doi.org/10.1186/gb-2014-15-2-r29

Lee, J. J., Lee, J., & Lee, H. (2021). Alternative paths to telomere elongation. *Semin Cell Dev Biol*, *113*, 88-96. https://doi.org/10.1016/j.semcdb.2020.11.003

Lee, M., Napier, C. E., Yang, S. F., Arthur, J. W., Reddel, R. R., & Pickett, H. A. (2017). Comparative analysis of whole genome sequencing-based telomere length measurement techniques. *Methods*, *114*, 4-15. https://doi.org/10.1016/j.ymeth.2016.08.008

Li, N., Ye, M., Li, Y., Yan, Z., Butcher, L. M., Sun, J., Han, X., Chen, Q., Zhang, X., & Wang, J. (2010). Whole genome DNA methylation analysis based on high throughput sequencing technology. *Methods*, *52*(3), 203-212. https://doi.org/10.1016/j.ymeth.2010.04.009

Liao, X., Zhu, W., Zhou, J., Li, H., Xu, X., Zhang, B., & Gao, X. (2023). Repetitive DNA sequence detection and its role in the human genome. *Commun Biol*, *6*(1), 954. https://doi.org/10.1038/s42003-023-05322-y

Lin, R., Zhong, X., Zhou, Y., Geng, H., Hu, Q., Huang, Z., Hu, J., Fu, X. D., Chen, L., & Chen, J. Y. (2022). R-loopBase: a knowledgebase for genome-wide R-loop formation and regulation. *Nucleic Acids Res*, *50*(D1), D303-D315. https://doi.org/10.1093/nar/gkab1103

Liu, Y., & Song, C. X. (2022). TAPS: The Development of a Direct and Base-Resolution Sequencing Method for DNA Methylation. *ACS Chem Biol*, *17*(10), 2683-2685. https://doi.org/10.1021/acschembio.2c00746

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Lu, B., Zhang, P., Chen, S., Benbarche, S., Castro, C., Fox, N., Abubakar, M., Biswas, J., Soldatov, R., Koche, R., & Abdel-Wahab, O. (2023). Genome-Wide Mapping of R-Loops in Single Cells Reveals Cell-Type Specific Roles for RLoops in Hematopoietic Cell Identity. *Blood*, *142*. https://doi.org/10.1182/blood-2023-180913

Lucas, A., Mialet-Perez, J., Daviaud, D., Parini, A., Marber, M. S., & Sicard, P. (2015). Gadd45gamma regulates cardiomyocyte death and post-myocardial infarction left ventricular remodelling. *Cardiovasc Res*, *108*(2), 254-267. https://doi.org/10.1093/cvr/cvv219

Ma, B., Martínez, P., Sánchez-Vázquez, R., & Blasco, M. A. (2023). Telomere dynamics in human pluripotent stem cells. *Cell Cycle*, *22*(23-24), 2505-2521. https://doi.org/10.1080/15384101.2023.2285551

Maamar, M. B., Sadler-Riggleman, I., Beck, D., & Skinner, M. K. (2021). Genome-Wide Mapping of DNA Methylation 5mC by Methylated DNA Immunoprecipitation (MeDIP)-Sequencing. *Methods Mol Biol*, *2198*, 301-310. https://doi.org/10.1007/978-1-0716-0876-0_23

Madeira, F., Madhusoodanan, N., Lee, J., Eusebi, A., Niewielska, A., Tivey, A. R. N., Lopez, R., & Butcher, S. (2024). The EMBL-EBI Job Dispatcher sequence analysis tools framework in 2024. *Nucleic Acids Res*, *52*(W1), W521-W525. https://doi.org/10.1093/nar/gkae241

Maestroni, L., Matmati, S., & Coulon, S. (2017). Solving the Telomere Replication Problem. *Genes (Basel)*, *8*(2). https://doi.org/10.3390/genes8020055

Malig, M., & Chédin, F. (2020). Characterization of R-Loop Structures Using Single-Molecule R-Loop Footprinting and Sequencing. *Methods Mol Biol*, *2161*, 209-228. https://doi.org/10.1007/978-1-0716-0680-3_15

Mangaonkar, A. A., & Patnaik, M. M. (2018). Short Telomere Syndromes in Clinical Practice: Bridging Bench and Bedside. *Mayo Clin Proc*, *93*(7), 904-916. https://doi.org/10.1016/j.mayocp.2018.03.020

Mareschal, S., Palau, A., Lindberg, J., Ruminy, P., Nilsson, C., Bengtzen, S., Engvall, M., Eriksson, A., Neddermeyer, A., Marchand, V., Jansson, M., Bjorklund, M., Jardin, F., Rantalainen, M., Lennartsson, A., Cavelier, L., Gronberg, H., & Lehmann, S. (2021). Challenging conventional karyotyping by next-generation karyotyping in 281 intensively treated patients with AML. *Blood Adv*, *5*(4), 1003-1016. https://doi.org/10.1182/bloodadvances.2020002517

Marshall, P. R., Zhao, Q., Li, X., Wei, W., Periyakaruppiah, A., Zajaczkowski, E. L., Leighton, L. J., Madugalle, S. U., Basic, D., Wang, Z., Yin, J., Liau, W. S., Gupte, A., Walkley, C. R., & Bredy, T. W. (2020). Dynamic regulation of Z-DNA in the mouse prefrontal cortex by the RNA-editing enzyme Adar1 is required for fear extinction. *Nat Neurosci*, *23*(6), 718-729. https://doi.org/10.1038/s41593-020-0627-5

Martindale, J. L., Gorospe, M., & Idda, M. L. (2020). Ribonucleoprotein Immunoprecipitation (RIP) Analysis. *Bio Protoc*, *10*(2), e3488. https://doi.org/10.21769/BioProtoc.3488

Mathews, L. M., Chi, S. Y., Greenberg, N., Ovchinnikov, I., & Swergold, G. D. (2003). Large differences between LINE-1 amplification rates in the human and chimpanzee lineages. *Am J Hum Genet*, *72*(3), 739-748. https://doi.org/10.1086/368275

Mayer, C., Leese, F., & Tollrian, R. (2010). Genome-wide analysis of tandem repeats in Daphnia pulex--a comparative approach. *BMC Genomics*, *11*, 277. https://doi.org/10.1186/1471-2164-11-277

McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A*, *36*(6), 344-355. https://doi.org/10.1073/pnas.36.6.344

Mendez-Dorantes, C., & Burns, K. H. (2023). LINE-1 retrotransposition and its deregulation in cancers: implications for therapeutic opportunities. *Genes Dev*, *37*(21-24), 948-967. https://doi.org/10.1101/gad.351051.123

Mir, A. A., Philippe, C., & Cristofari, G. (2015). euL1db: the European database of L1HS retrotransposon insertions in humans. *Nucleic Acids Res*, *43*(Database issue), D43-47. https://doi.org/10.1093/nar/gku1043

Mir, S. M., Samavarchi Tehrani, S., Goodarzi, G., Jamalpoor, Z., Asadi, J., Khelghati, N., Qujeq, D., & Maniati, M. (2020). Shelterin Complex at Telomeres: Implications in Ageing. *Clin Interv Aging*, *15*, 827-839. https://doi.org/10.2147/CIA.S256425

Moore, L. D., Le, T., & Fan, G. (2013). DNA methylation and its basic function. *Neuropsychopharmacology*, *38*(1), 23-38. https://doi.org/10.1038/npp.2012.112
Morrison, O., & Thakur, J. (2021). Molecular Complexes at Euchromatin, Heterochromatin and Centromeric Chromatin. *Int J Mol Sci*, *22*(13). https://doi.org/10.3390/ijms22136922
Müller, S., & Almouzni, G. (2017). Chromatin dynamics during the cell cycle at centromeres. *Nat Rev Genet*, *18*(3), 192-208. https://doi.org/10.1038/nrg.2016.157
Muñoz-López, M., & García-Pérez, J. L. (2010). DNA transposons: nature and applications in genomics. *Curr Genomics*, *11*(2), 115-128. https://doi.org/10.2174/138920210790886871
Musselman, C. A., Lalonde, M. E., Cote, J., & Kutateladze, T. G. (2012). Perceiving the epigenetic landscape through histone readers. *Nat Struct Mol Biol*, *19*(12), 1218-1227. https://doi.org/10.1038/nsmb.2436
Nakato, R., & Sakata, T. (2021). Methods for ChIP-seq analysis: A practical workflow and advanced applications. *Methods*, *187*, 44-53. https://doi.org/10.1016/j.ymeth.2020.03.005
Nathan, D., Sterner, D. E., & Berger, S. L. (2003). Histone modifications: Now summoning sumoylation. *Proc Natl Acad Sci U S A*, *100*(23), 13118-13120. https://doi.org/10.1073/pnas.2436173100
Neri, F., Rapelli, S., Krepelova, A., Incarnato, D., Parlato, C., Basile, G., Maldotti, M., Anselmi, F., & Oliviero, S. (2017). Intragenic DNA methylation prevents spurious transcription initiation. *Nature*, *543*(7643), 72-77. https://doi.org/10.1038/nature21373
Nersisyan, L., & Arakelyan, A. (2015). Computel: computation of mean telomere length from whole-genome next-generation sequencing data. *PLoS One*, *10*(4), e0125201. https://doi.org/10.1371/journal.pone.0125201
Niehrs, C., & Luke, B. (2020). Regulatory R-loops as facilitators of gene expression and genome stability. *Nat Rev Mol Cell Biol*, *21*(3), 167-178. https://doi.org/10.1038/s41580-019-0206-3
Ninomiya, K., & Hirose, T. (2020). Short Tandem Repeat-Enriched Architectural RNAs in Nuclear Bodies: Functions and Associated Diseases. *Non-Coding RNA*, *6*(1), 6. https://www.mdpi.com/2311-553X/6/1/6
Nudler, E. (2009). RNA polymerase active center: the molecular engine of transcription. *Annu Rev Biochem*, *78*, 335-361. https://doi.org/10.1146/annurev.biochem.76.052705.164655
Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y.,…Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, *376*(6588), 44-53. https://doi.org/10.1126/science.abj6987
Onodera, A., Gonzalez-Avalos, E., Lio, C. J., Georges, R. O., Bellacosa, A., Nakayama, T., & Rao, A. (2021). Roles of TET and TDG in DNA demethylation in proliferating and non-proliferating immune cells. *Genome Biol*, *22*(1), 186. https://doi.org/10.1186/s13059-021-02384-1
Pacesa, M., Loeff, L., Querques, I., Muckenfuss, L. M., Sawicka, M., & Jinek, M. (2022). R-loop formation and conformational activation mechanisms of Cas9. *Nature*, *609*(7925), 191-196. https://doi.org/10.1038/s41586-022-05114-0
Panigrahi, A., & O'Malley, B. W. (2021). Mechanisms of enhancer action: the known and the unknown. *Genome Biol*, *22*(1), 108. https://doi.org/10.1186/s13059-021-02322-1
Parekh, S., Ziegenhain, C., Vieth, B., Enard, W., & Hellmann, I. (2016). The impact of amplification on differential expression analyses by RNA-seq. *Sci Rep*, *6*, 25533. https://doi.org/10.1038/srep25533
Parseghian, M. H. (2015). What is the role of histone H1 heterogeneity? A functional model emerges from a 50 year mystery. *AIMS Biophys*, *2*(4), 724-772. https://doi.org/10.3934/biophy.2015.4.724
Perkiö, A., Pradhan, B., Genc, F., Pirttikoski, A., Pikkusaari, S., Erkan, E. P., Falco, M. M., Huhtinen, K., Narva, S., Hynninen, J., Kauppi, L., & Vähärautio, A. (2024). Locus-specific LINE-1 expression in clinical ovarian cancer specimens at the single-cell level. *Sci Rep*, *14*(1), 4322. https://doi.org/10.1038/s41598-024-54113-w
Petit-Marty, N., Casas, L., & Saborido-Rey, F. (2023). State-of-the-art of data analyses in environmental DNA approaches towards its applicability to sustainable fisheries management [Review]. *Frontiers in Marine Science*, *10*. https://doi.org/10.3389/fmars.2023.1061530
Pilbauerova, N., Soukup, T., Suchankova Kleplova, T., Schmidt, J., & Suchanek, J. (2021). The Effect of Cultivation Passaging on the Relative Telomere Length and Proliferation Capacity of Dental Pulp Stem Cells. *Biomolecules*, *11*(3). https://doi.org/10.3390/biom11030464

Poon, S. S. S., & Lansdorp, P. M. (2001). Quantitative Fluorescence In Situ Hybridization (Q-FISH). *Current Protocols in Cell Biology*, *12*(1), 18.14.11-18.14.21. https://doi.org/https://doi.org/10.1002/0471143030.cb1804s12
Protasova, M. S., Andreeva, T. V., & Rogaev, E. I. (2021). Factors Regulating the Activity of LINE1 Retrotransposons. *Genes (Basel)*, *12*(10). https://doi.org/10.3390/genes12101562
Qin, D. (2019). Next-generation sequencing and its clinical application. *Cancer Biol Med*, *16*(1), 4-10. https://doi.org/10.20892/j.issn.2095-3941.2018.0055
Raab, J. R., & Kamakaka, R. T. (2010). Insulators and promoters: closer than we think. *Nat Rev Genet*, *11*(6), 439-446. https://doi.org/10.1038/nrg2765
Rahman, M. F., & McGowan, P. O. (2022). Cell-type-specific epigenetic effects of early life stress on the brain. *Transl Psychiatry*, *12*(1), 326. https://doi.org/10.1038/s41398-022-02076-9
Ren, L., Pushpakumar, S., Almarshood, H., Das, S. K., & Sen, U. (2024). Epigenetic DNA Methylation and Protein Homocysteinylation: Key Players in Hypertensive Renovascular Damage. *Int J Mol Sci*, *25*(21). https://doi.org/10.3390/ijms252111599
Robert, M. F., Morin, S., Beaulieu, N., Gauthier, F., Chute, I. C., Barsalou, A., & MacLeod, A. R. (2003). DNMT1 is required to maintain CpG methylation and aberrant gene silencing in human cancer cells. *Nat Genet*, *33*(1), 61-65. https://doi.org/10.1038/ng1068
Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139-140. https://doi.org/10.1093/bioinformatics/btp616
Rossiello, F., Jurk, D., Passos, J. F., & d'Adda di Fagagna, F. (2022). Telomere dysfunction in ageing and age-related diseases. *Nat Cell Biol*, *24*(2), 135-147. https://doi.org/10.1038/s41556-022-00842-x
Saha, S., & Pommier, Y. (2023). R-loops, type I topoisomerases and cancer. *NAR Cancer*, *5*(1), zcad013. https://doi.org/10.1093/narcan/zcad013
Sahoo, K., & Sundararajan, V. (2024). Methods in DNA methylation array dataset analysis: A review. *Comput Struct Biotechnol J*, *23*, 2304-2325. https://doi.org/10.1016/j.csbj.2024.05.015
Sano, S., Horitani, K., Ogawa, H., Halvardson, J., Chavkin, N. W., Wang, Y., Sano, M., Mattisson, J., Hata, A., Danielsson, M., Miura-Yura, E., Zaghlool, A., Evans, M. A., Fall, T., De Hoyos, H. N., Sundstrom, J., Yura, Y., Kour, A., Arai, Y.,…Walsh, K. (2022). Hematopoietic loss of Y chromosome leads to cardiac fibrosis and heart failure mortality. *Science*, *377*(6603), 292-297. https://doi.org/10.1126/science.abn3100
Sanz, L. A., & Chédin, F. (2019). High-resolution, strand-specific R-loop mapping via S9.6-based DNA-RNA immunoprecipitation and high-throughput sequencing. *Nat Protoc*, *14*(6), 1734-1755. https://doi.org/10.1038/s41596-019-0159-1
Sanz, L. A., Hartono, S. R., Lim, Y. W., Steyaert, S., Rajpurkar, A., Ginno, P. A., Xu, X., & Chédin, F. (2016). Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals. *Mol Cell*, *63*(1), 167-178. https://doi.org/10.1016/j.molcel.2016.05.032
Savage, A. L., Iacoangeli, A., Schumann, G. G., Rubio-Roldan, A., Garcia-Perez, J. L., Al Khleifat, A., Koks, S., Bubb, V. J., Al-Chalabi, A., & Quinn, J. P. (2022). Characterisation of retrotransposon insertion polymorphisms in whole genome sequencing data from individuals with amyotrophic lateral sclerosis. *Gene*, *843*, 146799. https://doi.org/10.1016/j.gene.2022.146799
Schäfer, A., Mekker, B., Mallick, M., Vastolo, V., Karaulanov, E., Sebastian, D., von der Lippen, C., Epe, B., Downes, D. J., Scholz, C., & Niehrs, C. (2018). Impaired DNA demethylation of C/EBP sites causes premature aging. *Genes Dev*, *32*(11-12), 742-762. https://doi.org/10.1101/gad.311969.118
Schüle, K. M., Leichsenring, M., Andreani, T., Vastolo, V., Mallick, M., Musheev, M. U., Karaulanov, E., & Niehrs, C. (2019). GADD45 promotes locus-specific DNA demethylation and 2C cycling in embryonic stem cells. *Genes Dev*, *33*(13-14), 782-798. https://doi.org/10.1101/gad.325696.119
Sergeeva, A., Davydova, K., Perenkov, A., & Vedunova, M. (2023). Mechanisms of human DNA methylation, alteration of methylation patterns in physiological processes and oncology. *Gene*, *875*, 147487. https://doi.org/10.1016/j.gene.2023.147487
Sexton, C. E., & Han, M. V. (2019). Paired-end mappability of transposable elements in the human genome. *Mob DNA*, *10*, 29. https://doi.org/10.1186/s13100-019-0172-5
Shiekh, S., Mustafa, G., Kodikara, S. G., Hoque, M. E., Yokie, E., Portman, J. J., & Balci, H. (2022). Emerging accessibility patterns in long telomeric overhangs. *Proc Natl Acad Sci U S A*, *119*(30), e2202317119. https://doi.org/10.1073/pnas.2202317119

Shiio, Y., & Eisenman, R. N. (2003). Histone sumoylation is associated with transcriptional repression. *Proc Natl Acad Sci U S A*, *100*(23), 13225-13230. https://doi.org/10.1073/pnas.1735528100
Shiromoto, Y., Sakurai, M., Minakuchi, M., Ariyoshi, K., & Nishikura, K. (2021). ADAR1 RNA editing enzyme regulates R-loop formation and genome stability at telomeres in cancer cells. *Nat Commun*, *12*(1), 1654. https://doi.org/10.1038/s41467-021-21921-x
Sil, S., Keegan, S., Ettefa, F., Denes, L. T., Boeke, J. D., & Holt, L. J. (2023). Condensation of LINE-1 is critical for retrotransposition. *Elife*, *12*. https://doi.org/10.7554/eLife.82991
Simonsen, J. L., Rosada, C., Serakinci, N., Justesen, J., Stenderup, K., Rattan, S. I., Jensen, T. G., & Kassem, M. (2002). Telomerase expression extends the proliferative life-span and maintains the osteogenic potential of human bone marrow stromal cells. *Nat Biotechnol*, *20*(6), 592-596. https://doi.org/10.1038/nbt0602-592
Slyvka, A., Mierzejewska, K., & Bochtler, M. (2017). Nei-like 1 (NEIL1) excises 5-carboxylcytosine directly and stimulates TDG-mediated 5-formyl and 5-carboxylcytosine excision. *Sci Rep*, *7*(1), 9001. https://doi.org/10.1038/s41598-017-07458-4
Smith, E. N., Jepsen, K., Khosroheidari, M., Rassenti, L. Z., D'Antonio, M., Ghia, E. M., Carson, D. A., Jamieson, C. H., Kipps, T. J., & Frazer, K. A. (2014). Biased estimates of clonal evolution and subclonal heterogeneity can arise from PCR duplicates in deep sequencing experiments. *Genome Biol*, *15*(8), 420. https://doi.org/10.1186/s13059-014-0420-4
Smith, R. P., Lam, E. T., Markova, S., Yee, S. W., & Ahituv, N. (2012). Pharmacogene regulatory elements: from discovery to applications. *Genome Med*, *4*(5), 45. https://doi.org/10.1186/gm344
Solomon, E. R., Caldwell, K. K., & Allan, A. M. (2021). A novel method for the normalization of ChIP-qPCR data. *MethodsX*, *8*, 101504. https://doi.org/10.1016/j.mex.2021.101504
Stratigi, K., Siametis, A., & Garinis, G. A. (2024). Looping forward: exploring R-loop processing and therapeutic potential. *FEBS Lett*. https://doi.org/10.1002/1873-3468.14947
Subrini, J., & Turner, J. (2021). Y chromosome functions in mammalian spermatogenesis. *Elife*, *10*. https://doi.org/10.7554/eLife.67345
Sytnikova, Y. A., Kubarenko, A. V., Schäfer, A., Weber, A. N., & Niehrs, C. (2011). Gadd45a is an RNA binding protein and is localized in nuclear speckles. *PLoS One*, *6*(1), e14500. https://doi.org/10.1371/journal.pone.0014500
Takekawa, M., & Saito, H. (1998). A family of stress-inducible GADD45-like proteins mediate activation of the stress-responsive MTK1/MEKK4 MAPKKK. *Cell*, *95*(4), 521-530. https://doi.org/10.1016/s0092-8674(00)81619-0
Talbert, P. B., & Henikoff, S. (2021). Histone variants at a glance. *J Cell Sci*, *134*(6). https://doi.org/10.1242/jcs.244749
Tamura, R. E., de Vasconcellos, J. F., Sarkar, D., Libermann, T. A., Fisher, P. B., & Zerbini, L. F. (2012). GADD45 proteins: central players in tumorigenesis. *Curr Mol Med*, *12*(5), 634-651. https://doi.org/10.2174/156652412800619978
Tanny, J. C. (2014). Chromatin modification by the RNA Polymerase II elongation complex. *Transcription*, *5*(5), e988093. https://doi.org/10.4161/21541264.2014.988093
Torp, R., Su, J. H., Deng, G., & Cotman, C. W. (1998). GADD45 is induced in Alzheimer's disease, and protects against apoptosis in vitro. *Neurobiol Dis*, *5*(4), 245-252. https://doi.org/10.1006/nbdi.1998.0201
Trigiante, G., Blanes Ruiz, N., & Cerase, A. (2021). Emerging Roles of Repetitive and Repeat-Containing RNA in Nuclear and Chromatin Organization and Gene Expression. *Front Cell Dev Biol*, *9*, 735527. https://doi.org/10.3389/fcell.2021.735527
Vaid, R., Wen, J., & Mannervik, M. (2020). Release of promoter-proximal paused Pol II in response to histone deacetylase inhibition. *Nucleic Acids Res*, *48*(9), 4877-4890. https://doi.org/10.1093/nar/gkaa234
Vaiserman, A., & Krasnienkov, D. (2020). Telomere Length as a Marker of Biological Age: State-of-the-Art, Open Issues, and Future Perspectives. *Front Genet*, *11*, 630186. https://doi.org/10.3389/fgene.2020.630186
van der Veer, B. K., Chen, L., Custers, C., Athanasouli, P., Schroiff, M., Cornelis, R., Chui, J. S., Finnell, R. H., Lluis, F., & Koh, K. P. (2023). Dual functions of TET1 in germ layer lineage bifurcation distinguished by genomic context and dependence on 5-methylcytosine oxidation. *Nucleic Acids Res*, *51*(11), 5469-5498. https://doi.org/10.1093/nar/gkad231
Verbiest, M., Maksimov, M., Jin, Y., Anisimova, M., Gymrek, M., & Bilgin Sonay, T. (2023). Mutation and selection processes regulating short tandem repeats give rise to genetic and phenotypic diversity across species. *J Evol Biol*, *36*(2), 321-336. https://doi.org/10.1111/jeb.14106

Verdonschot, J. A. J., Hazebroek, M. R., Wang, P., Sanders-van Wijk, S., Merken, J. J., Adriaansen, Y. A., van den Wijngaard, A., Krapels, I. P. C., Brunner-La Rocca, H. P., Brunner, H. G., & Heymans, S. R. B. (2018). Clinical Phenotype and Genotype Associations With Improvement in Left Ventricular Function in Dilated Cardiomyopathy. *Circ Heart Fail*, *11*(11), e005220. https://doi.org/10.1161/CIRCHEARTFAILURE.118.005220

Waddington, C. H. (1968). Towards a theoretical biology. *Nature*, *218*(5141), 525-527. https://doi.org/10.1038/218525a0

Wagstaff, B. J., Wang, L., Lai, S., Derbes, R. S., & Roy-Engel, A. M. (2018). Reviving a 60 million year old LINE-1 element. *Gene Rep*, *11*, 74-78. https://doi.org/10.1016/j.genrep.2018.02.007

Wang, F., Pan, X., Kalmbach, K., Seth-Smith, M. L., Ye, X., Antumes, D. M., Yin, Y., Liu, L., Keefe, D. L., & Weissman, S. M. (2013). Robust measurement of telomere length in single cells. *Proc Natl Acad Sci U S A*, *110*(21), E1906-1912. https://doi.org/10.1073/pnas.1306639110

Wang, J., Wang, H., Chen, J., Wang, X., Sun, K., Wang, Y., Wang, J., Yang, X., Song, X., Xin, Y., Liu, Z., & Hui, R. (2008). GADD45B inhibits MKK7-induced cardiac hypertrophy and the polymorphisms of GADD45B is associated with inter-ventricular septum hypertrophy. *Biochem Biophys Res Commun*, *372*(4), 623-628. https://doi.org/10.1016/j.bbrc.2008.05.122

Wang, J., Zhao, Y., Zhou, X., Hiebert, S. W., Liu, Q., & Shyr, Y. (2018). Nascent RNA sequencing analysis provides insights into enhancer-mediated gene regulation. *BMC Genomics*, *19*(1), 633. https://doi.org/10.1186/s12864-018-5016-z

Wang, Z., Fang, Y., Liu, Z., Hao, N., Zhang, H. H., Sun, X., Que, J., & Ding, H. (2024). Adapting nanopore sequencing basecalling models for modification detection via incremental learning and anomaly detection. *Nat Commun*, *15*(1), 7148. https://doi.org/10.1038/s41467-024-51639-5

Warkocki, Z. (2023). An update on post-transcriptional regulation of retrotransposons. *FEBS Lett*, *597*(3), 380-406. https://doi.org/10.1002/1873-3468.14551

Wells, J. N., & Feschotte, C. (2020). A Field Guide to Eukaryotic Transposable Elements. *Annu Rev Genet*, *54*, 539-561. https://doi.org/10.1146/annurev-genet-040620-022145

Westemeier-Rice, E. S., Winters, M. T., Rawson, T. W., & Martinez, I. (2024). More than the SRY: The Non-Coding Landscape of the Y Chromosome and Its Importance in Human Disease. *Noncoding RNA*, *10*(2). https://doi.org/10.3390/ncrna10020021

Wheeler, M. M., Stilp, A. M., Rao, S., Halldorsson, B. V., Beyter, D., Wen, J., Mihkaylova, A. V., McHugh, C. P., Lane, J., Jiang, M. Z., Raffield, L. M., Jun, G., Sedlazeck, F. J., Metcalf, G., Yao, Y., Bis, J. B., Chami, N., de Vries, P. S., Desai, P.,…Reiner, A. P. (2022). Whole genome sequencing identifies structural variants contributing to hematologic traits in the NHLBI TOPMed program. *Nat Commun*, *13*(1), 7592. https://doi.org/10.1038/s41467-022-35354-7

White, L. K., & Hesselberth, J. R. (2022). Modification mapping by nanopore sequencing. *Front Genet*, *13*, 1037134. https://doi.org/10.3389/fgene.2022.1037134

Wilsbacher, L., & McNally, E. M. (2016). Genetics of Cardiac Developmental Disorders: Cardiomyocyte Proliferation and Growth and Relevance to Heart Failure. *Annu Rev Pathol*, *11*, 395-419. https://doi.org/10.1146/annurev-pathol-012615-044336

Wingett, S. W., & Andrews, S. (2018). FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Res*, *7*, 1338. https://doi.org/10.12688/f1000research.15931.2

Wreczycka, K., Gosdschan, A., Yusuf, D., Grüning, B., Assenov, Y., & Akalin, A. (2017). Strategies for analyzing bisulfite sequencing data. *J Biotechnol*, *261*, 105-115. https://doi.org/10.1016/j.jbiotec.2017.08.007

Wu, C., & Morris, J. R. (2001). Genes, genetics, and epigenetics: a correspondence. *Science*, *293*(5532), 1103-1105. https://doi.org/10.1126/science.293.5532.1103

Wu, T., Lyu, R., & He, C. (2022). spKAS-seq reveals R-loop dynamics using low-input materials by detecting single-stranded DNA with strand specificity. *Sci Adv*, *8*(48), eabq2166. https://doi.org/10.1126/sciadv.abq2166

Xu, Y., Jiao, Y., Liu, C., Miao, R., Liu, C., Wang, Y., Ma, C., & Liu, J. (2024). R-loop and diseases: the cell cycle matters. *Mol Cancer*, *23*(1), 84. https://doi.org/10.1186/s12943-024-02000-3

Yang, S., Winstone, L., Mondal, S., & Wu, Y. (2023). Helicases in R-loop Formation and Resolution. *J Biol Chem*, *299*(11), 105307. https://doi.org/10.1016/j.jbc.2023.105307

Yang, T. C., Wu, P. C., Chung, I. F., Jiang, J. H., Fann, M. J., & Kao, L. S. (2016). Cell death caused by the synergistic effects of zinc and dopamine is mediated by a stress sensor gene

Gadd45b - implication in the pathogenesis of Parkinson's disease. *J Neurochem*, *139*(1), 120-133. https://doi.org/10.1111/jnc.13728

Yang, Y., Zhang, M., & Wang, Y. (2022). The roles of histone modifications in tumorigenesis and associated inhibitors in cancer therapy. *J Natl Cancer Cent*, *2*(4), 277-290. https://doi.org/10.1016/j.jncc.2022.09.002

Yap, K., Mukhina, S., Zhang, G., Tan, J. S. C., Ong, H. S., & Makeyev, E. V. (2018). A Short Tandem Repeat-Enriched RNA Assembles a Nuclear Compartment to Control Alternative Splicing and Promote Cell Survival. *Mol Cell*, *72*(3), 525-540 e513. https://doi.org/10.1016/j.molcel.2018.08.041

Yin, F., Bruemmer, D., Blaschke, F., Hsueh, W. A., Law, R. E., & Herle, A. J. (2004). Signaling pathways involved in induction of GADD45 gene expression and apoptosis by troglitazone in human MCF-7 breast carcinoma cells. *Oncogene*, *23*(26), 4614-4623. https://doi.org/10.1038/sj.onc.1207598

Yu, H. J., Byun, Y. H., & Park, C. K. (2024). Techniques for assessing telomere length: A methodological review. *Comput Struct Biotechnol J*, *23*, 1489-1498. https://doi.org/10.1016/j.csbj.2024.04.011

Zardoni, L., Nardini, E., Brambati, A., Lucca, C., Choudhary, R., Loperfido, F., Sabbioneda, S., & Liberi, G. (2021). Elongating RNA polymerase II and RNA:DNA hybrids hinder fork progression and gene expression at sites of head-on replication-transcription collisions. *Nucleic Acids Res*, *49*(22), 12769-12784. https://doi.org/10.1093/nar/gkab1146

Zhang, B., Zhu, Y., Zhang, Z., Wu, F., Ma, X., Sheng, W., Dai, R., Guo, Z., Yan, W., Hao, L., Huang, G., Ma, D., Hao, B., & Ma, J. (2024). SMC3 contributes to heart development by regulating super-enhancer associated genes. *Exp Mol Med*, *56*(8), 1826-1842. https://doi.org/10.1038/s12276-024-01293-0

Zhang, X., Zhang, R., & Yu, J. (2020). New Understanding of the Relevant Role of LINE-1 Retrotransposition in Human Disease and Immune Modulation. *Front Cell Dev Biol*, *8*, 657. https://doi.org/10.3389/fcell.2020.00657

Zhang, X., Zhang, Y., Wang, C., & Wang, X. (2023). TET (Ten-eleven translocation) family proteins: structure, biological functions and applications. *Signal Transduct Target Ther*, *8*(1), 297. https://doi.org/10.1038/s41392-023-01537-x

Zhao, H., Zhu, M., Limbo, O., & Russell, P. (2018). RNase H eliminates R-loops that disrupt DNA replication but is nonessential for efficient DSB repair. *EMBO Rep*, *19*(5). https://doi.org/10.15252/embr.201745335

Zimmermann, S., Voss, M., Kaiser, S., Kapp, U., Waller, C. F., & Martens, U. M. (2003). Lack of telomerase activity in human mesenchymal stem cells. *Leukemia*, *17*(6), 1146-1149. https://doi.org/10.1038/sj.leu.2402962

8 List of Acronyms

5caC	5-carboxylcytosine
5fC	5-formylcytosine
5hmC	5-hydroxymethyllcytosine
5mC	5-methyllcytosine
A	adenine
ADAR1	adenosine deaminase acting on RNA 1
ALT	alternative lengthening of telomeres
AP-2	activating enhancer binding protein 2
AQR	aquarius intron-binding spliceosomal factor
ATAC-seq	assay for transposase-accessible chromatin with high-throughput sequencing
C	cytosine
Cas9	clustered regularly interspaced short palindromic repeats-associated protein 9
CDK	cyclin-dependent kinase
cDNA	complementary DNA
cGAS	cyclic guanosine monophosphate–adenosine monophosphate synthase
ChIP-qPCR	chromatin immunoprecipitation quantitative polymerase chain reaction
ChIP-seq	chromatin immunoprecipitation sequencing
CpG	cytosine(-phosphate-)guanine
CPM	counts per million
CPU	central processing unit
CRISPR	clustered regularly interspaced short palindromic repeats
CTCF	CCCTC binding factor
CUT&Tag	cleavage under targets and tagmentation sequencing
DEG	differentially expressed gene
DIP-seq	DNA immunoprecipitation sequencing
DNA	deoxyribonucleic acid
DNAJA1	DnaJ heat shock protein family (Hsp40) Member A1
DNMT1	deoxyribonucleic acid methyltransferase
DNMT	DNA methyltransferase
DRIP-seq	DNA-RNA immunoprecipitation sequencing
eRNA	enhancer ribonucleic acid
ESC	embryonic stem cell
ESRR	estrogen-related receptor
FACS	fluorescence-activated cell sorting
FC	fold change
FDR	false discovery rate
FISH	fluorescent in situ hybridization
Flow-FISH	fluorescent in situ hybridization with flow cytometry
FoxO	forkhead box protein O
G	guanine

GADD45	GADD45 protein family The growth arrest and DNA damage-inducible 45
GADD45a	growth arrest and DNA damage-inducible alpha
GADD45b	growth arrest and DNA damage-inducible beta
GADD45g	growth arrest and DNA damage-inducible gamma
GAPDH	glyceraldehyde-3-phosphate dehydrogenase
GB	gigabyte
GO	gene ontology
GPL	GNU General Public License
HBD	hybrid-binding protein domain
HEK293T	human embryonic kidney 293T cells
HPC	high performance computing
HTT	huntingtin
IgG	immunoglobulin G
IMB	Institute of Molecular Biology
IPP	International PhD Programme
I/O	input/output
iPSC	induced pluripotent stem cell
JNK	Jun N-terminal kinase
KAP1	Krüppel-associated box-associated protein 1
kb	kilobases
kDA	kilodalton
KEGG	Kyoto Encyclopedia of Genes and Genomes
KLF	Krüppel-like factor
KLF4	Krüppel-like factor 4
KRAB-ZFPs	Krüppel-associated box domain-containing zinc finger proteins
LFC	logarithmic fold change
LINE1	long interspersed nuclear element-1
LTS	long-term support
M	million
MAP	mitogen-activated protein
MAPK	mitogen-activated protein kinase
MAPKKK	mitogen-activated protein kinase kinase kinase
Mb	megabases
MEF	mouse embryonic fibroblast
MEKK4	mitogen-activated protein kinase kinase kinase 4
mESC	mouse embryonic stem cell
MKK7	mitogen-activated protein kinase kinase 7
mRNA	messenger ribonucleic acid
MSA	multiple sequence alignment
MSC	mesenchymal stem cell
NCBI	National Center for Biotechnology Information
NR2F2	nuclear receptor subfamily 2 group F member 2
NR4A1	nuclear receptor subfamily 4 group A member 1
ORF	open reading frame
PAR	pseudoautosomal region
PC	personal computer
PC1/2	principal component 1/2

PCR	polymerase chain reaction
PhD	doctor of philosophy
PNA	peptide nucleic acid
POSIX	portable operating system interface
PRC2	polycomb repressive complex 2
Q-FISH	quantitative fluorescent in situ hybridization
qPCR	quantitative PCR
R-ChIP	R-loop chromatin immunoprecipitation
RAM	random access memory
RE	restriction enzyme
RIP-seq	ribonucleic acid immunoprecipitation sequencing
RNA	ribonucleic acid
RNA-seq	ribonucleic acid sequencing
RNase	ribonuclease
RNH	ribonuclease H1
RNPs	ribonucleoproteins
RPKM	reads per kilobase per million mapped reads
rRNA	ribosomal ribonucleic acid
RT-qPCR	reverse transcription quantitative real-time polymerase chain reaction
S1	nuclease S1
SETX	probable helicase senataxin
SKO	single knockout
SLURM	simple Linux utility for resource management
Soni.	sonication
SRA	sequence read archive
ssDNA	single-stranded DNA
STR	short tandem repeat
strDRIP-seq	strand-specific DNA-RNA immunoprecipitation sequencing
T	thymine
T2T	telomere-to-telomere
TARID	transcription factor 21 antisense ribonucleic acid inducing promoter demethylation
TB	terabyte
TCF21	transcription factor 21
TEAD	TEA domain family transcription factor
TERRA	telomeric repeat-containing ribonucleic acid
TET	ten-eleven translocation methylcytosine dioxygenase
TET1	ten-eleven translocation methylcytosine dioxygenase 1
TKO	triple knockout
TMM	trimmed mean of M values
TOP1	topoisomerase I
TOP2	topoisomerase II
TOP3B	topoisomerase type II beta
TSS	transcription start site
TTS	transcription termination site
UCSC	University of California, Santa Cruz
UMI	unique molecular identifier

unpubl.	unpublished
UTR	untranslated region
VNTR	variable number tandem repeat
WGS	whole genome sequencing
WT	wild-type

9 Acknowledgements

10 Lebenslauf