UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

The dark side of the genome:
Development of a computational pipeline to identify transposable elements with a functional role in genome regulation

Miguel Casanova Vieira Parente

Mestrado em Bioinformática e Biologia Computacional

Dissertação orientada por:
Doutor Nuno Luís Barbosa Morais
Professora Margarida Henriques da Gama Carvalho

2024

Acknowledgments

I would like to begin by expressing my heartfelt gratitude to my advisor and friend, Nuno Morais. Our scientific journey started together in 2001 in Prof. Maria Carmo Fonseca's group. Twenty years later, Nuno gave me the incredible opportunity to join the Disease Transcriptomics Lab at the IMM, and guided me as a supervisor for my master's degree in Bioinformatics and Computational Biology. Your enthusiasm, rigour, and meticulous attention to detail have been truly inspiring and have allowed me to grow in the fields of computational biology and statistics. Thank you for your invaluable guidance, for broadening my horizons, and most importantly, for keeping my passion for academia and scientific research alive.

A special thanks to Claire Rougeulle, head of the "Non-coding RNA, Differentiation and Development" group at the Epigenetics and Cell Fate Unit in Paris. With you, I've experienced some of the most fulfilling years of my scientific career. Your intelligence, kindness, and unwavering trust in me have motivated me to always give my best. I am also grateful to you for providing access to the unpublished datasets used in this project.

I would also like to extend my thanks to Simão Teixeira da Rocha, who served as an unofficial supervisor and, more importantly, a true friend. In your lab at SCERG, I had the opportunity to mentor younger students, continue working on human pluripotency projects, and hone my bioinformatics skills. Thank you for being such a close and dear friend over the past 15 years. Your creativity, genuine passion for science, kindness, and constant support for everyone around you are a source of inspiration.

In addition, I am deeply grateful to Margarida Gama Carvalho for taking on the role of internal supervisor at FCUL. Our paths first crossed in 2001 in the corridors of the "Instituto de Histologia e Embriologia," where you supervised my final-year biology project. It feels only fitting that both you and Nuno have now played key roles as supervisors for my master's thesis in bioinformatics. Thank you very much for your support.

Science, for me, is shaped by the people who surround you, challenge you, and inspire you to push boundaries every day. During my time at the Disease Transcriptomics Lab, I had the privilege of working alongside extraordinary colleagues, brilliant scientists, and wonderful friends. A huge thank you to Nuno Agostinho, Mariana Ferreira, Marta Bica, Rita Belo, José Ferrão, Rita Silva, Alexandre Afonso, and Jacek Marzec. You made each day at work exciting and enriching. Thank you for reigniting my sense of wonder, reminding me of the joy of discovery, and showing me that the future of science is in excellent hands.

Last but certainly not least, I want to express my deepest gratitude to my family, without whom I would never have reached this point. A big thank you to my parents and my brother for their unwavering support, for giving me the freedom to chase my dreams, and for being the best role models I could ask for. To my children, Rafael and Inês Casanova, thank you for inspiring me to be the best version of myself. I hope to always make you proud. Finally, to my partner, friend, and colleague, Anne-Valérie Gendrel, my sincerest and most heartfelt thanks. We have travelled far and wide, sharing much of this journey together. Thank you for your endless patience, support, and for our shared passion for transposable elements and epigenetics. Your help with this project and manuscript has been invaluable, and without your encouragement and support, this would not have been possible.

Resumo

A complexidade genómica em mamíferos está intimamente ligada à diversidade das espécies, refletindo-se na multiplicidade de tipos celulares especializados e na capacidade de adaptação a ambientes variados. Historicamente, acreditava-se que a complexidade genómica estava associada ao tamanho do genoma e ao número de genes codificadores de proteínas. Contudo, esta correlação foi refutada por exemplos como o axolote mexicano (*Ambystoma mexicanum*) e o peixe pulmonado australiano (*Protopterus aethiopicus*), que possuem genomas significativamente maiores do que o humano, sem apresentar uma maior complexidade genómica.

Este paradoxo levou à reavaliação dos fatores que contribuem para a complexidade genómica, destacando o papel da fração não codificante do genoma, que corresponde a cerca de 98% do genoma humano. Dentro desta fração, os elementos transponíveis (TEs) representam aproximadamente 50%. Os TEs são sequências genéticas capazes de se mover dentro do genoma, influenciando profundamente a sua estrutura e função. Classificam-se em retrotransposões (classe I), como as famílias LINE, SINE, SVA e ERV, e transposões de DNA (classe II), minoritários no genoma humano. Tradicionalmente considerados de "lixo genómico" ou elementos parasitas, os TEs são agora reconhecidos como agentes-chave na evolução dos genomas e na regulação da expressão génica. Fornecem sequências regulatórias, como promotores, enhancers e sítios de ligação para fatores de transcrição, podendo remodelar redes regulatórias e contribuir para a diversidade genética e fenotípica. Além disso, influenciam a organização tridimensional do genoma, afetando interações entre regiões distantes, que se pensa ter um papel crucial na regulação genómica.

Durante o desenvolvimento embrionário precoce e a neurogénese, os TEs exibem padrões de expressão dinâmicos, sugerindo funções regulatórias críticas nestes contextos. Por exemplo, subfamílias específicas de TEs estão associadas a estados de pluripotência “naive” e “primed” em células estaminais embrionárias humanas (hESCs), influenciando a identidade celular e o potencial de diferenciação deste modelo celular. Apesar da ideia que os TEs têm uma função importante nos seus hospedeiros, a sua atividade é uma espada de dois gumes: a sua desregulação pode resultar em mutações prejudiciais, instabilidade genómica e doenças, incluindo cancros e desordens neurológicas. Por isso, os organismos desenvolveram mecanismos para controlar a atividade dos TEs, como a metilação do DNA e modificações pós-traducionais das histonas.

O estudo e a compreensão dos TEs tem sido limitado pela reduzida disponibilidade e qualidade de ferramentas para os estudar. Nomeadamente, a análise computacional dos TEs apresenta desafios significativos devido à sua natureza repetitiva e alta similaridade de sequências entre cópias. São por isso necessárias abordagens bioinformáticas avançadas para mapear com precisão as sequências derivadas de TEs e avaliar a sua contribuição para a expressão génica e estrutura genómica. Ferramentas de sequenciação de nova geração, incluindo tecnologia de leituras longas, e algoritmos especializados estão a ser desenvolvidos para superar estas limitações.

O objetivo desta tese é o desenvolvimento de uma ferramenta computacional que identifique, de forma robusta e imparcial, subfamílias de elementos transponíveis com regulação específica em diferentes contextos biológicos ou patológicos. Implementámos um pipeline para a análise simultânea da expressão de genes e TEs, tanto ao nível de subfamílias como de loci individuais, com o intuito de explorar como os TEs contribuem para a complexidade genómica e para a regulação da expressão génica. Esta abordagem visa facilitar a caracterização sistemática do transcriptoma de TEs em diversos contextos,

fornecendo uma base sólida para futuros estudos funcionais sobre o seu impacto, especialmente no sistema nervoso.

Neste trabalho, analisámos dados de RNA-seq total de 12 amostras de hESCs da linha H9 feminina, cultivadas nos estados de pluripotência primed e naive. Realizámos o controlo de qualidade da sequenciação com o *FastQC* e remoção de adaptadores com o *Trim Galore*. As leituras foram alinhadas ao genoma humano de referência (hg38/GRCh38.p13) usando o alinhador *STAR*. Implementámos duas estratégias de mapeamento: retenção de leituras unicamente mapeadas para quantificar a expressão de TEs a nível de loci individuais; e atribuição aleatória de leituras multi-mapeadas para quantificar a expressão global de subfamílias de TEs. Criámos anotações personalizadas de TEs a partir do UCSC RepeatMasker, gerando ficheiros GTF com atributos específicos para cada TE, incluindo subfamília, família e classe. De seguida, as anotações de genes e TEs foram combinadas para permitir a quantificação simultânea da sua expressão, utilizando a ferramenta featureCounts. A análise de expressão diferencial foi realizada com o pacote *DESeq2*, após filtragem de genes e TEs com baixa expressão e normalização dos dados. Realizámos análises exploratórias, como clustering hierárquico e análise de componentes principais (PCA), para avaliar a separação entre as condições experimentais. Para identificar subfamílias de TEs relevantes, utilizámos várias abordagens: seleção baseada nos valores estatísticos da análise diferencial; cálculo da proporção de loci diferencialmente expressos em cada subfamília; e análise de enriquecimento funcional inspirada no GSEA, empregando o pacote fgsea. A associação genómica entre genes diferencialmente expressos e subfamílias de TEs foi avaliada através de testes de permutação com o pacote *regioneR*, e com o pacote *bedtoolsR*.

A nossa análise revelou perfis transcriptómicos distintamente separados entre hESCs primed e naive. Através de PCA e clustering hierárquico, observámos uma clara distinção das amostras com base no estado de pluripotência, indicando diferenças significativas na expressão génica global. Identificámos numerosos genes diferencialmente expressos entre os dois estados ($p<0,01$). Genes associados à pluripotência naive, como *DPPA3*, *KLF4*, *TBX3*, *DNMT3L* e *GATA6*, estavam significativamente sobre-expressos em hESCs naive. Em contraste, genes como *MYC*, *DUSP6*, *OTX2*, *ZIC2* e *DNMT3B* mostraram sobre-expressão nas hESCs primed.

Em relação aos TEs, identificámos subfamílias diferencialmente expressas específicas para cada estado de pluripotência. Nas hESCs naive, observámos sobre-expressão significativa de subfamílias como HERVK-int, LTR5_Hs/LTR5, já conhecidas por estarem associadas a este estado de pluripotência, várias subfamílias SVA (incluindo SVA_A, SVA_B, SVA_C) e *Alu*, bem como subfamílias jovens de LINE-1 (L1Hs e L1PA2). Estes TEs, sendo elementos transponíveis recentes e potencialmente ativos, podem desempenhar papéis regulatórios únicos no estado naive, influenciando genes chave na manutenção da pluripotência. Nas hESCs primed, identificámos sobre-expressão das subfamílias HERVH-int e LTR7, conhecidas por estarem associadas a este estado, além de subfamílias de LINE mais antigas e MIR. Estes elementos podem ter sido integrados nos mecanismos regulatórios das hESCs ao longo da evolução, influenciando a expressão génica de forma distinta. Estes resultados validam a eficácia do nosso pipeline em detectar diferenças coerentes com o conhecimento atual sobre marcadores específicos de cada estado.

Ao analisar a expressão de loci individuais de TEs, observámos que, embora algumas subfamílias apresentem regulação global, a expressão de elementos individuais pode variar consideravelmente. Isto indica uma regulação complexa que depende do contexto genómico local, possivelmente influenciada por fatores epigenéticos e interações com elementos regulatórios adjacentes.

A análise de enriquecimento funcional revelou que os TEs associados ao estado naive estão significativamente enriquecidos nas proximidades das regiões regulatórias dos genes sobre-expressos nesse estado ($p<0,01$). Especificamente, TEs como LTR5_Hs, HERVK-int e SVA mostraram acumulação próxima aos genes naive, sugerindo um papel como elementos *cis*-reguladores. Além disso, verificámos que a distância média entre estes TEs e os genes sobre-expressos em hESCs naive é significativamente menor em comparação com genes aleatórios, reforçando a hipótese de interação funcional direta. Para os TEs associados ao estado primed, o padrão foi menos pronunciado, embora alguns, como HERVH-int e LTR7, mostrassem proximidade aumentada a genes específicos deste estado. Estes resultados sugerem que os TEs podem contribuir diferencialmente para a regulação génica nos dois estados de pluripotência, possivelmente através de mecanismos dependentes do contexto celular e da disponibilidade de fatores de transcrição específicos. Estes resultados sugerem que os TEs desempenham um papel significativo na regulação da expressão génica em hESCs, contribuindo para a definição dos estados de pluripotência primed e naive. As subfamílias de TEs diferencialmente expressas parecem estar envolvidas na orquestração de redes regulatórias específicas de cada estado, atuando potencialmente como plataformas para a ligação de fatores de transcrição ou como elementos moduladores da cromatina.

O nosso estudo demonstra que os elementos transponíveis influenciam significativamente a modulação do transcriptoma das hESCs. A identificação de subfamílias de TEs diferencialmente expressas entre os estados primed e naive, e a sua associação com genes específicos de cada estado, indica que os TEs contribuem para a identidade celular e para a regulação dos programas de expressão génica. A metodologia desenvolvida oferece uma ferramenta poderosa para investigar o papel dos TEs em diversos contextos biológicos. A aplicação deste pipeline em doenças do neurodesenvolvimento e neurodegenerativas é promissora, dado que os TEs têm sido implicados na patogénese de condições como a síndrome de Rett, esclerose lateral amiotrófica e doença de Alzheimer. A análise da desregulação dos TEs nestes contextos poderá fornecer informação valiosa sobre a progressão das doenças e identificar novos biomarcadores e alvos terapêuticos.

Assim, o nosso trabalho aprofunda o conhecimento sobre a regulação dos TEs em células estaminais e abre novas perspetivas para a investigação em doenças neurológicas, onde a modulação dos TEs pode ter implicações clínicas significativas. A compreensão detalhada do papel dos TEs poderá, no futuro, contribuir para o desenvolvimento de estratégias inovadoras de diagnóstico e tratamento em patologias do sistema nervoso.

Palavras-chave: elementos de transposição, sequenciação de RNA, transcriptómica computacional, células estaminais embrionárias humanas.

Abstract

Transposable elements (TEs), which constitute nearly half of the human genome, have historically been viewed as parasitic sequences with little functional importance. However, growing evidence suggests that TEs play pivotal roles in gene expression and epigenetic modulation. This thesis aims to develop and implement a computational pipeline to explore the functional roles of TEs in gene regulation using human embryonic stem cells (hESCs), as a model system known for its relevance to early development and cellular therapies.

The methodology involves the analysis of total RNA sequencing (RNA-seq) datasets from both primed and naive hESCs, two distinct pluripotency states, which have been shown to have a remarkably distinct pattern of TE transcription. The pipeline includes preprocessing, alignment, and quantification of gene and TE expression using custom TE annotations. Differential expression analysis was performed at both the TE subfamily and individual loci levels to identify TEs that are differentially active in a given state and that could be important for its transcriptional regulation. Additionally, we applied statistical tests to assess the genomic association between differentially expressed TEs and differentially expressed genes.

Our analysis identified key TE subfamilies, including endogenous retroviruses (ERVs), *Alus*, SINE-VNTR-*Alus* (SVAs), and LINE elements, that are differentially expressed between naive and primed states of hESCs. Moreover, genomic proximity analysis revealed a potential regulatory relationship between these TEs and neighbouring genes, suggesting that TEs may serve as *cis*-regulatory elements contributing to cell state-specific gene expression profiles.

In conclusion, this work establishes a robust pipeline for the parallel analysis of gene and TE expression, providing new insights into the potential regulatory roles of TEs in hESCs. Future studies could apply this pipeline to other developmental systems or disease models, offering a broader understanding of TE-mediated regulation and its implications for genome function in health and disease.

Keywords: transposable elements, RNA sequencing, computational transcriptomics, human embryonic stem cells.

Table of contents

List of figures

List of tables

Abbreviation List

ALS - Amyotrophic lateral sclerosis
ASD - Autism spectrum disorder
bp - base pair
CAGE - Cap-analysis gene expression
ChIP - Chromatin immunoprecipitation
CRAN - The Comprehensive R Archive Network
DEA - Differential expression analysis
DEG - Differentially expressed gene
DETE - Differentially expressed transposable element
DNMT - DNA methyl-transferase
EDA - Exploratory data analysis
ERV - Endogenous retrovirus
ES - Enrichment score
GSEA - Gene set enrichment analysis
GTF - Gene Transfer Format
HERV - Human endogenous retrovirus
hESC - Human embryonic stem cell
HMT - Histone methyl-transferase
ICM - Inner cell mass
kb - Kilobase
KRAB-ZFPs - Krüppel-associated box domain zinc finger proteins
LINE - Long interspersed nuclear element
lncRNA - long non-coding RNA
log2FC - log2 fold-change
LTR - Long terminal repeat
MECP2 - methyl-CpG-binding protein 2
MERV - Mouse endogenous retrovirus
MLS - Multiple sclerosis
NES - Normalised enrichment score
NGS - next-generation sequencing
NPC - Neuronal precursor cell
ORF - Open reading frame
PC - Principal component
PCA - Principal component analysis
RIN - RNA integrity score
RNA pol - RNA polymerase
RNA-seq - RNA sequencing
RTT - Rett syndrome
SINE - Short interspersed nuclear element
STAR - Spliced Transcripts Alignment to a Reference
SVA - SINE variable-number tandem-repeat *Alu* element
TAD - Topological associated domain
TE - Transposable element
TF - Transcription factor
UTR - Untranslated region
VNTR - Variable number of tandem repeats

1. Introduction

1.1. How complex is defining genome complexity?

The rich diversity of mammalian species is closely related with an increase in the complexity of genomes. This translates into organisms displaying an increasing number of functionally specialised cell types, that confer an advantage in adapting to complex environments. For decades, the discussion around the parameters that define the complexity of genomes has been tightly coupled to its size and the number of protein-coding genes harboured within. Interestingly, our increased knowledge about genomes and gene composition quickly debunked the idea of a simple association between genome size and organism complexity. In fact, many animals, including the Mexican axolotl (*Ambystoma mexicanum*) and the Australian lungfish (*Protopterus aethiopicus*), have genomes that are, respectively, 10 to 14 times bigger than those of humans (Keinath et al. 2015; Meyer et al. 2021). But more than sheer genome size, scientists have seldom resorted to the argument of significant genomic information in the form of protein-coding genes. The completion of the draft sequence of the human genome project, in 2001 (International Human Genome Sequencing Consortium 2001), revealed that the number of protein-coding genes was as low as 20,000 genes, a figure that is unimpressive, compared to parasitic organisms like *Trichomonas vaginalis*, to *Mus musculus*, or even to plant species, like rice (*Oryza sativa*) (Straalen et al. 2011). Recently, state-of-the-art sequencing technologies producing a full, telomere-to-telomere sequence of the human genome, did not change the initial estimates: there are 63,494 predicted genes, 19,969 of which are predicted to be protein-coding (Nurk et al. 2021).

The paradox created by the lack of a linear association between genome size, number of protein-coding genes, and organism complexity is enticing and has garnered the attention of many generations of scientists. The last decades revealed that, more than size and number of genes, the architectural organisation of genomes and the orchestrated and timely regulation of gene expression seem to play a crucial role in the emergence of complex programs of cellular identity and function. The recent technological advancement on methodologies for understanding nuclear organisation, namely chromosome conformation capture technologies, revealed that the genome is not a one-dimensional sequence of nucleotides, but a complex and dynamic, three-dimensional structure, hierarchically organised into distinct functional domains (Jerkovic´ and Cavalli 2021). This organisation has a deep impact on how the activity of genes is regulated, strongly contributing to genome complexity.

In addition to genome architecture, alternative splicing, a process of selecting different combinations of splice sites to generate a set of distinct spliced mRNAs from the same gene, has been proposed to contribute to the evolution of phenotypic novelty, by promoting transcript diversification in the absence of an increased number of genes (Barbosa-Morais et al. 2012; Bush et al. 2017; Yang et al. 2021).

1.2. The non-coding fraction of the genome: the gold hidden between the rubble

The last decades have witnessed a notable shift in the discussion about genome complexity, pushing it far beyond coding sequences and the regulation of protein-coding genes. In fact, in mammalian species, only a very small fraction of the genome encodes for proteins. In humans, for example, less than 2% of the genome has coding potential (International Human Genome Sequencing Consortium 2001; Nurk et al. 2021). Interestingly, humans share the bulk of these coding regions with other mammals, which raises

the question of how a similar set of protein-coding genes can give rise to such an enormous biological variety and complexity. These observations have led researchers to question the role of the non-coding regions of the genome in the regulation of gene expression and the emergence of genetic and regulatory novelties. The bulk of the non-coding fraction of mammalian genomes consists of sequences of transposable elements (TEs) (International Human Genome Sequencing Consortium 2001; Nurk et al. 2021; Smit, AFA. et al. 2013), which represent approximately 50% of mammalian genomes (**Fig. 1.1A**). TEs are genetic elements that once had or still have the ability to transpose, that is, that are able to move from one location to another in the genome, independently of the host (Bourque et al. 2018; Senft and Macfarlan 2021; Wells and Feschotte 2020). For this reason, TEs are often considered parasitic genetic elements. Intriguingly, compared to protein-coding genes, TEs show a tremendous diversity across species (Senft and Macfarlan 2021; Wells and Feschotte 2020), which urges a better understanding of their function, in particular in species-specific regulatory innovations.

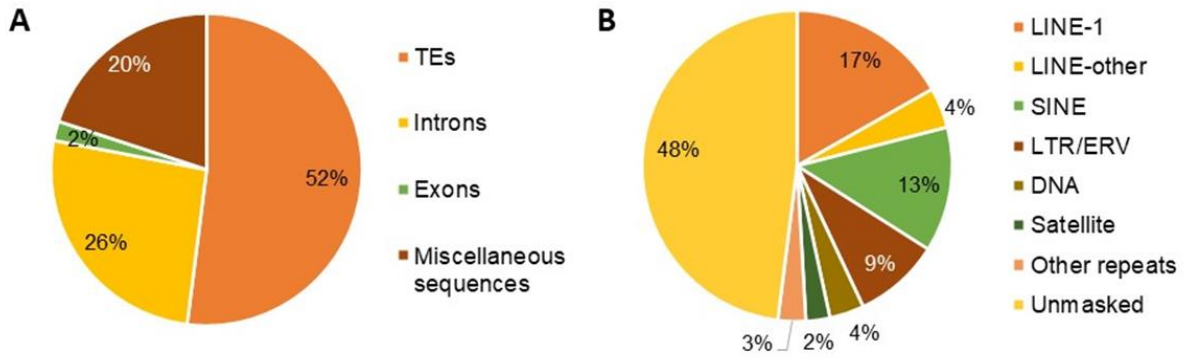


Fig. 1.1- Composition of the human genome. A) Pie chart indicating the percentages of repetitive and unique sequences. **B**) Pie chart indicating the proportions of different subclasses of TEs and other types of repeats. Unmasked represents unique sequences.

1.3. Classification of transposable elements

The origin of TEs is not completely known, but it is believed that they could be derived either from ancient viral infections, self-splicing group II introns from bacteria or cellular RNAs, such as tRNAs (Wells and Feschotte 2020). TEs that are established in the germline, are inherited vertically, across generations, leading to the colonisation of genomes with a myriad of different TE insertions. Their complex origin and propagation throughout evolutionary times lead to a staggering number of different TE groups populating most genomes (Bourque et al. 2018). The most basic division of eukaryotic TEs distinguishes between two major classes, based on their transposition intermediates: class I encompasses retrotransposons, elements that are able to replicate via an RNA intermediate, that is then reverse-transcribed into a cDNA intermediate before being integrated elsewhere in the genome; class II elements are DNA transposons, elements that get excised and move to a new genomic location. Whereas class I elements retain the original element and expand in the genome, by a "copy-and-paste" mechanism, class II TEs mobilise mainly through a "cut-and-paste" strategy. DNA transposons are inactive in most mammalian species and represent a minority within the mammalian repertoire of TEs, whereas retrotransposons are much more abundant (Garcia-Perez et al. 2016). For instance, in the human genome, class I TEs represent 45% of the genome, whereas class II elements constitute only 3.6% (**Fig. 1.1B**) (Smit, AFA. et al. 2013).

The phylogenetic tree of eukaryotic TEs is rich and complex: each of the two main classes can be further subdivided into subclasses, based on their mechanisms of replication and/or chromosomal integration,

and then into superfamilies, families and subfamilies (**Fig. 1.1B**; **Fig. 1.2**) (Wells and Feschotte 2020). The more detailed classification and grouping of TEs into subfamilies include elements that can be traced as descendants of a single ancestral unit (Britten and Kohne 1968; Bourque et al. 2018).

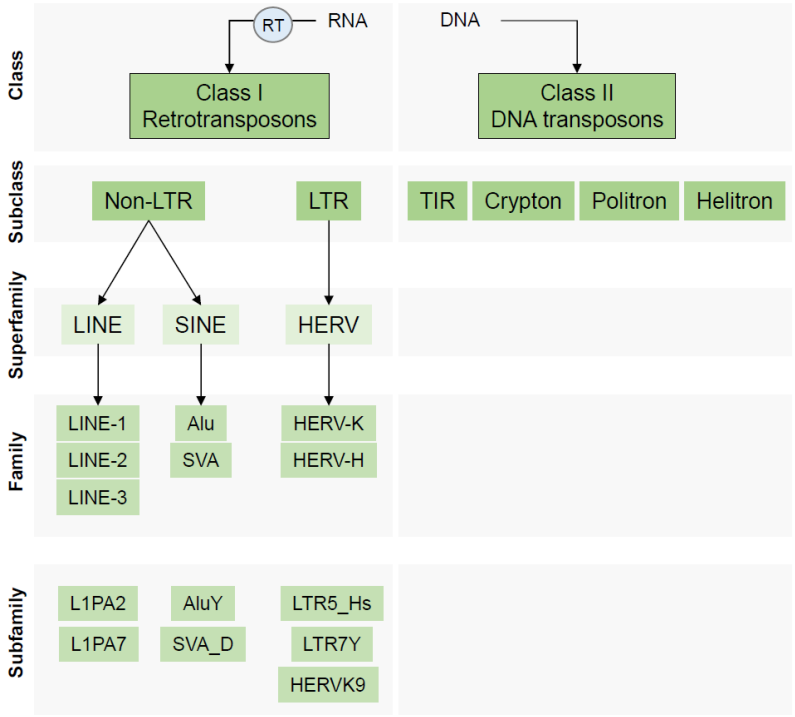


Fig. 1.2 - Classification of transposable elements (TEs) in the human genome by class, subclass, superfamily, family, and subfamily. Transposable elements are divided into Class I (RNA intermediates) and Class II (DNA intermediates) based on their transposition mechanisms. Class I elements transpose via reverse transcription (RT), while Class II elements use a DNA-based cut-and-paste mechanism. Selected examples of families and subfamilies from each superfamily, which will be described in subsequent sections, are indicated (adapted from Bourque et al. 2018).

1.3.1. Class I retrotransposons

Retrotransposons are broadly divided into long terminal repeat (LTR) elements and non-LTR elements.

1.3.1.1. Non-LTR retrotransposons

Non-LTR elements are structurally simpler and include autonomous long interspersed nuclear elements (LINEs) and non-autonomous short interspersed nuclear elements (SINEs) (comprising the primate-specific *Alu* repeats), which require LINE-derived proteins for retrotransposition. In addition, non-LTR elements include the non-autonomous Hominidae-specific composite SINE variable-number tandem-repeat *Alu* (SVA) elements (**Fig. 1.3**).

LINEs are typically 6-7 kb long and constitute the largest proportion of TE-derived sequences, representing 21.8% of the human genome (Smit, AFA. et al. 2013). These elements are composed of a 5’ untranslated region (UTR) containing a RNA pol II promoter (with sense and antisense activity), two coding open reading frames (ORFs), ORF1 and ORF2, and a 3’UTR containing the poly(A) signal (**Fig. 1.3A**) (Mita et al. 2018). ORF1 remains poorly characterised, but is thought to encode an RNA-binding protein with nucleic acid chaperone activity that forms an oligomeric product required for the recognition and transport of the template RNA to the nucleus (Richardson et al. 2015). ORF2 encodes a protein with endonuclease and reverse transcription activities, which are essential for the replication and

integration of these elements in the host genome. In the human genome, the LINE group is dominated by a single family, the LINE-1 family, which represents 80% of LINE elements and makes up 17% of the genome (Smit, AFA. et al. 2013).

SINEs are on average 200 bp long and constitute 13.4% of the human genome, making it the subclass with the largest number of elements. Their typical structure consists of a head, a body, and a tail (Kramerov and Vassetzky 2011). The head, located at the 5' end, is derived from 7SL RNAs or tRNAs and contains a RNA pol III promoter. The origin of the body is difficult to trace, but shares homology with certain LINE elements, varying significantly between SINE subfamilies. Lastly, the 3' tail is composed of short simple repeats of varying length and the poly(A) signal (Rodriguez-Terrones and Torres-Padilla 2018). A particular family of SINEs, the *Alu* elements, arose early in primate evolution by a process involving the fusion of two monomers derived from the 7SL RNA, flanking an A-rich sequence (**Fig. 1.3B**) (Kriegs et al. 2007). Their emergence coincided with the mammalian radiation and their expansion has been so successful that *Alu* repeats currently represent the family of TEs with the highest number of copies in mammalian genomes (> 1 million copies in the human genome) (Smit, AFA. et al. 2013). In a hominoid ancestor, a fusion between an *Alu* element, a variable number of tandem repeats (VNTR) and a SINE-R region derived from an LTR fragment of a human endogenous retrovirus K, gave rise to the SVA family (Wang et al. 2005; Wells and Feschotte 2020; Fueyo et al. 2022). Unlike other SINE families, SVA are ~2 kb long and rely on RNA pol II promoters for their transcription (**Fig. 1.3B**).

Genomic integration of non-LTR elements is achieved by target-primed reverse transcription, in which a LINE-encoded endonuclease generates a single-stranded nick that will allow the hybridization of the RNA template with the host DNA. This is followed by reverse transcription and the local integration of the cDNA strand (Wells and Feschotte 2020). Nonetheless, the vast majority of non-LTR elements have accumulated mutations, truncations or rearrangements, rendering them inactive for retrotransposition. It is however estimated that 80-100 copies from the LINE-1 family are fully intact and a handful, known as 'hot' LINE-1, are still highly active for retrotransposition. These elements account for most of the retrotransposition events in the human population, which can, in addition to their own RNA, retrotranspose RNAs from non-autonomous SINE and SVA elements (Brouha et al. 2003; Beck et al. 2010).

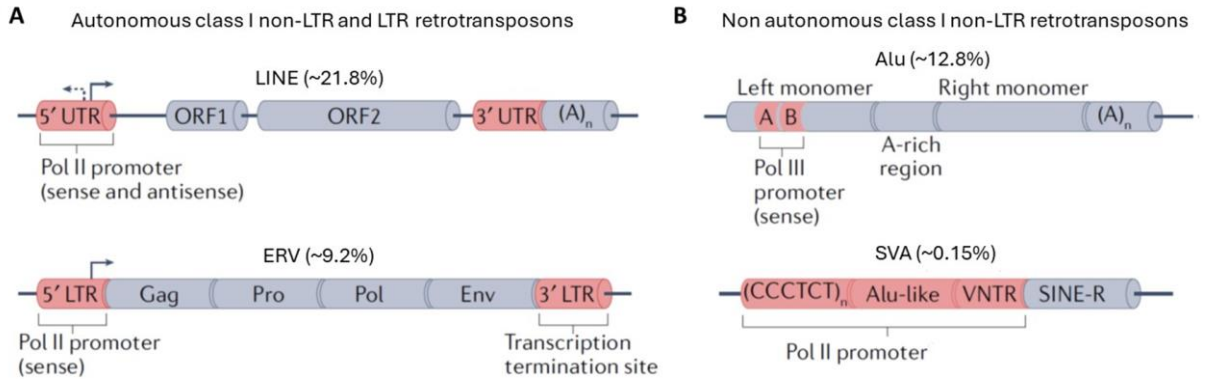


Fig. 1.3 - Structure of class I retrotransposons in the human genome. A) Typical structure of autonomous non-LTR (LINE) and LTR (ERV) retrotransposons. **B**) Typical structure of non autonomous non-LTR retrotransposons families, *Alu* and SVA. Pol II: RNA pol II promoter. Pol III: RNA pol III promoter (from Fueyo et al. 2022).

1.3.1.2. LTR retrotransposons

LTR retrotransposons, also known as endogenous retroviruses (ERVs), which constitute 9.2% of the human genome (Smit, AFA. et al. 2013), display more complex structures and mechanisms of replication, largely resembling retroviruses, from which they are believed to be derived (Eickbush and Malik 2007). This heterogeneous group is composed of various families, further divided into subfamilies, which have the ability to move around the genome by resorting to a cleavage and strand-transfer reaction catalysed by an integrase, much like what is observed for retroviruses (Wells and Feschotte 2020). A full length, autonomous, ERV has an average length of 7.5 kb and consists of two identical LTRs, which contain the *cis*-regulatory sequences including promoters, enhancers, termination and polyadenylation signals and flank the ORFs that encode the viral proteins (**Fig. 1.3A**). Autonomous LTR retrotransposons contain at least a set of two genes, *gag* and *pro-pol*, expressed as a single polycistronic RNA transcribed from a RNA Pol II promoter located within the 5' LTR sequence. These genes lead to the production of several proteins, including structural proteins, a protease for post-transcriptional processing of ERV derived proteins, a reverse transcriptase and an integrase (Wells and Feschotte 2020). In contrast to retroviruses, most ERVs have lost the fusogenic *env* gene, required to produce infectious virions. Unlike non-LTR retrotransposons, autonomous ERV elements are no longer competent for retrotransposition in the human genome; they can however still be transcriptionally active. Moreover, throughout evolution, ERVs are often reduced to a single LTR, or 'solo LTR', following ectopic recombination between the two LTRs. These solo LTRs are distributed abundantly across the genome, populating it with a vast arsenal of regulatory elements.

1.3.2. Class II DNA transposons

Class II elements, also known as DNA transposons, depend on a DNA intermediate for their mobilisation and transposition. They constitute one of the most diverse and widespread groups of TEs, and can be found in prokaryotes and eukaryotes alike. Their evolutionary success is such that some of the transposase genes they encode, are amongst the oldest and most abundant genes found in nature (Wells and Feschotte 2020). In humans, they represent ~3.6% of the genome (Smit, AFA. et al. 2013) and can be divided into two major groups, depending on the mechanism of transposition and the number of DNA strands cut during this process (Wicker et al. 2007; Rodriguez-Terrones and Torres-Padilla 2018).

A first group includes the *TIR* and *Crypton* subclasses and includes transposons that transpose via the classical "cut-and-paste" mechanism, requiring both DNA strands to be cleaved. These elements consist of a single ORF encoding a transposase flanked by two inverted terminal repeats (ITRs). The TE-encoded transposase binds the two ITRs, excising the TE element from the host genome and mediating its insertion at a different genomic site (**Fig. 1.4**).

The other group includes the *Helitrons* and *Polintons* subclasses. Unlike TEs from the previous group, these require only a single DNA strand to be cleaved for its transposition. This leads to the excising of a single-strand DNA circle from the lagging strand DNA template, which is then "pasted" into a target site. As such, TEs from this group transpose replicatively via a "peel-and-paste" mechanism (Rodriguez-Terrones and Torres-Padilla 2018).

Although DNA transposons are particularly active in bacteria and plants, there is no sign of recent activity from any DNA transposon subclasses in the human genome. Nevertheless, they had an important impact on the evolution of mammalian genomes (Feschotte and Pritham 2007).

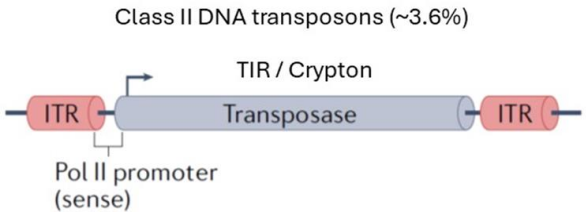


Fig. 1.4 - Structure of class II DNA transposons in the human genome. Structure of the TIR/Crypton subclass of DNA transposons in the human genome. ITR: inverted terminal repeats. Pol II: RNA pol II promoter (from Fueyo et al. 2022).

1.4. Sculpting genomes: how do TEs contribute to the genome of their hosts?

TEs were classically considered as junk DNA or as purely parasitic elements, self-propagating in the host genome. The pioneering and visionary work of Barbara McClintock on the role of "jumping genes" in the colour phenotype of maize kernels, more than 70 years ago, started exploring how important and ubiquitous is the role of TEs as major drivers of genetic and phenotypic diversity (McClintock 1950, 1951). Since this pioneering work, these elements have been shown to play major roles in genome evolution, structural variation, genome size expansion, spatial organisation, genetic diversity and gene regulation (Chalopin et al. 2015; Cordaux and Batzer 2009; Feschotte and Pritham 2007; Kruse et al. 2019).

The evolutionary success of TEs is partly justified by the delicate balance established between expression and repression by the host. Unchecked expression and transposition of TEs can have prejudicial effects, namely inducing mutations, disrupting genes, hindering the transcriptional coordination of gene networks and leading to the production of extranuclear retroviral DNA that can induce cellular toxicity (Bourque et al. 2018). For that reason, the host maintains a tight arms-race with TEs, keeping them in check to prevent deleterious changes (Molaro and Malik 2016). The maintenance of this delicate equilibrium has to safeguard that essential biological processes from the host are conserved while, at the same time, allowing the emergence of novel regulatory mechanisms that can present an evolutionary advantage.

Although the importance of TEs in the host genome might sound surprising, these elements possess a set of characteristics that justify their tremendous impact. TEs are ubiquitous and abundant genomic elements that carry a rich arsenal of molecular tools that render them autonomous for their propagation. As such, TEs are potent insertional mutagens, which can transpose and insert *de novo* into genes, disrupting their normal function (**Fig. 1.5**). When occurring in the germline, *de novo* transposition events can result in monogenic genetic diseases.

TEs also carry a number of regulatory sequences, such as promoters, enhancers and transcription termination signals. This is the case, for example, of the 5'UTR of LINEs or the LTRs flanking ERVs. Even though most of the TEs that populate the human genome have accumulated mutations and truncations that rendered them unable to transpose, they have been leaving behind these *cis*-regulatory sequences, which can recruit different transcription factors from the host, such as RNA polymerases, or act as binding sites for specific transcription factors (TFs). These sequences are widely distributed across the human genome and can have a strong impact on the regulation of host genomes. Indeed, some of these TE-derived sequences have been co-opted by their hosts and contribute to regulatory innovation in a species-specific manner (Chuong et al. 2017; Feschotte and Gilbert 2012; Ito et al. 2017; Sundaram

et al. 2014). Importantly, as TE families populate the genome of their host as a multitude of related copies, they can confer the same regulatory potential to a large group of genes, thereby contributing to the expansion of gene-regulatory networks (Sundaram and Wysocka 2020; Sundaram et al. 2014).

The mechanisms by which TE-derived elements can affect the host genome, either through *cis*- or *trans*-mechanisms, are diverse: creation of new or alternative promoters and enhancers; creation or disruption of TF binding sites; alteration of splicing patterns; production of regulatory non-coding RNAs; modification of the 3D organisation of the genome; generation of novel genes by fusion with TE-derived coding sequences; collateral effects of TE silencing (**Fig. 1.5**, reviewed in (Feschotte 2008; Chuong et al. 2017; Fueyo et al. 2022; Almeida et al. 2022)). A few examples will be further described in the next sections.

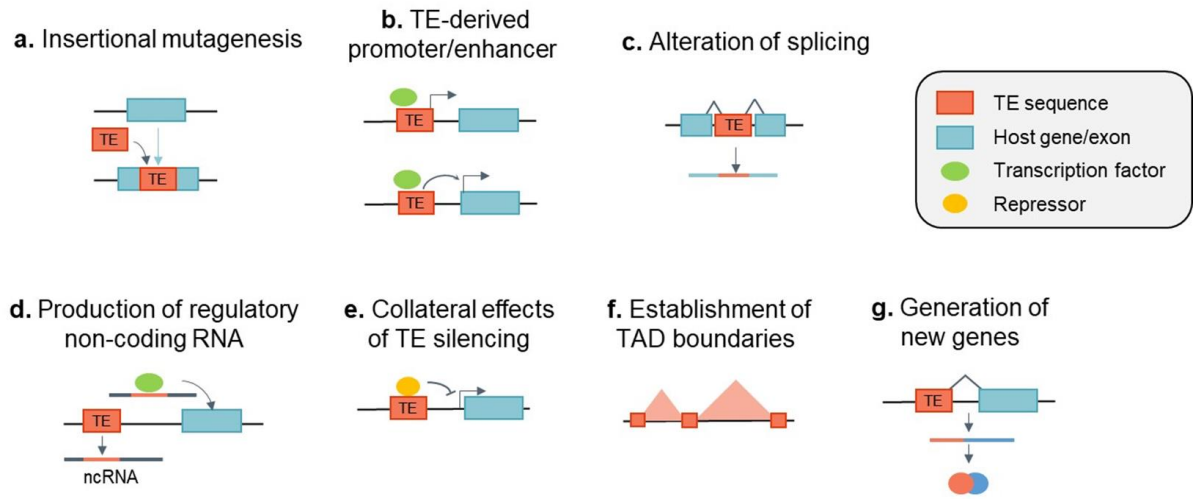


Fig. 1.5 - Examples of the influence of TEs on the genetic and transcriptional landscape of host genomes. Actively transposing TEs can be a source of mutations by inserting *de novo* into genes (a). Even without being active, TEs have the ability to alter the host genome through alteration of transcription (b-d), of gene expression regulation (b,e) and 3D chromatin architecture (f). Sequences derived from TEs can also be co-opted by the host for the generation of new genes (g) (adapted from Le Breton et al. 2024).

1.4.1. Transposition: jumping around the genome

The ability of TEs to change their position within a genome is remarkable. Transposition is an inherent ability of TEs that is kept in check by the host. For this reason, most of the TE elements on the human genome have lost the ability to transpose. As previously mentioned, LINE-1 elements are the only family of autonomous TEs still active in the human genome and capable of retrotransposition. Moreover, SINEs, which are non-autonomous elements as they do not encode any proteins, rely on the machinery produced by LINE-1 elements for their retrotransposition. Indeed, although they show a strong *cis*-preference, LINE-1-derived proteins, ORF1 and ORF2, are able to bind in trans SINE RNAs. When occurring in the germline, *de novo* transposition events can result in monogenic genetic diseases. To date, in humans, 124 LINE-1-mediated insertions of LINE-1, *Alu* or SVA elements, which result in genetic diseases, have been reported. Most of these reported disease-causing insertions inactivate gene function through integration into a coding exon or aberrant splicing (Hancks and Kazazian 2016).

Furthermore, misregulation of TEs is a hallmark of many cancers. Numerous cancers, including lung, colon, pancreatic and ovarian cancers, have been associated with misregulation of LINE-1 element, ORF1 protein expression and somatic LINE-1 retrotransposition events. While some *de novo* LINE-1 insertions have been shown to drive cancer (e.g. the *APC* tumor suppressor gene in colon cancer), the majority of somatic LINE-1 insertions in malignancies are spread across non-coding regions of the

genome. Further research is required to determine their potential functional impact (Payer and Burns 2019).

Interestingly, somatic retrotransposition events of LINE-1 have been shown to occur in the healthy human brain (Le Breton et al. 2024). While the exact rate of these events remains uncertain, this could have important implications for neuronal plasticity and diversity. However, the actual functional significance of these events in brain function remains an open question.

1.4.2. Mechanisms of *cis*-regulation of the host genome

Examples of how TEs are thought to have been co-opted into regulatory roles have built up in recent years, thanks to the development of massively parallel sequencing technologies. For instance, transcriptome analysis uncovered that a large proportion of mammalian tissue-specific or alternative promoters are derived from TE sequences. This was demonstrated using RNA sequencing (RNA-seq) analyses, which revealed that chimeric transcripts initiating within TE-derived promoters constitute a considerable fraction of mammalian transcriptomes, during early development and in many mammalian tissues (Macfarlan et al. 2012; Melé et al. 2015). Moreover, cap-analysis gene expression (CAGE)-seq analysis to map sites of transcriptional initiation uncovered surprisingly high amounts of RNA polymerase (RNA pol) II initiation within TEs in a range of human and mouse cell types and tissues (Faulkner and Carninci 2009).

Surveys of genome-wide binding patterns of TFs in humans and mice revealed that several families of TEs carry sequence motifs for specific TFs. This suggests that the amplification of a given TE family is correlated with the expansion of binding sites for certain TFs across the genome (Bourque et al. 2008). Related to this, an important study analysing the binding sites of 26 orthologous TFs in both human and mouse cells by chromatin immunoprecipitation (ChIP)-seq showed that between 2 and 40% of binding sites are found within TEs (Sundaram et al. 2014). Strikingly, this appears to occur in a species-specific manner as, for most TFs, the binding sites contributed by TEs are not conserved between humans and mice and have evolved independently within a given species. Moreover, several TEs had significantly expanded TF-binding sites only in one species, suggesting that TEs are an important driving force for species-specific regulatory innovation (Sundaram et al. 2014; Sundaram and Wysocka 2020).

Recently, a few studies have started to explore the contribution of TEs for the three-dimensional organisation of genome architecture. Mammalian genomes have been shown to be partitioned into topological associated domains (TADs), megabase scale chromatin domains of self-interaction between distant genomic regions (Nora et al. 2012; Dixon et al. 2012). Interestingly, several TE families have been shown to be enriched at TAD boundaries or harbour insulator activity. This is likely occurring through CTCF/cohesin binding to TEs (Bousios et al. 2020; Diehl et al. 2020). For example, it has been shown that at least 40% of the CTCF binding sites in the mouse genome (22.8% in humans) are derived from SINEs (Sundaram et al. 2014).

1.4.3. Beyond transposition and genomic regulation

Besides their role in the short- and long-range regulation of their surrounding genome environment, the influence of TEs on their hosts was demonstrated for a number of cellular and molecular mechanisms. Remarkably, the “domestication” of TE sequences led to the emergence of several key protein-coding genes, with both conserved and species-specific functions (Bourque et al. 2018). For example, the coding sequences from different TE families have been domesticated in multiple occasions to integrate

genes involved in V(D)J somatic recombination in the vertebrate immune system (*RAG1/2* genes), placental development in both humans and mice (*Syncytin* genes), host defence against exogenous retroviruses in mice (*Fv1* gene) and brain development (*Arc* gene) (Naville et al. 2016; Frank and Feschotte 2017; Almeida et al. 2022).

The exaptation of TEs for the repertoire of host genes has an even higher impact when considering the non-coding component of mammalian genomes. Indeed, genes producing long non-coding RNAs (lncRNAs) contain a much higher density of TE-derived sequences than protein-coding genes, with TEs comprising ~30% of the total lncRNA sequences in humans and mice, compared with ~0.3% for coding sequences (Kapusta et al. 2013). In fact, more than 70% of human lncRNA transcripts contain at least one exon of partial TE origin, which highlights the crucial contribution of TEs to the repertoire of lncRNA genes. Several of these TE-derived lncRNAs were shown to be implicated in the regulation of important epigenetic mechanisms operating during embryonic development, such as in the process of X-chromosome inactivation (*XIST* lncRNA) or in the maintenance of pluripotency during early embryonic development (described in the next section) (reviewed in Casanova et al. 2016).

TEs have also been suggested to act as effectors of the innate immune system, namely in the transcriptional response to interferon signalling. Interferons constitute the first line of defence against viral infections by creating a cellular response that hinders viral replication and signals for the presence of these pathogens (Platanias 2005; Barreiro et al. 2010). Interestingly, several studies have shown that in different mammalian lineages, the LTRs of distinct ERV families have been repurposed as interferon-inducible enhancers for the regulation of adjacent interferon-stimulated genes (Chuong et al. 2016; Rookhuizen et al. 2021). Thus, in a fascinating turn of evolutionary events, TEs have been repurposed to fight viral infections, from which a significant proportion of TEs have derived.

These striking examples highlight how the co-option of TE sequences can be occasionally repurposed for important host cellular function, contributing notably to immunity, antiviral defence and development.

1.5. Embryonic and neuronal development: a playground for TEs?

Owing to their sheer number, biology and genome-wide distribution, TEs have been repurposed during evolution, in a species-specific manner, to gain a notable influence over particular developmental and cellular contexts. Undoubtedly, early embryonic development, pluripotency and brain function are amongst the biological processes where the role of TEs seems to have a more striking impact.

In the early developing embryo, extensive genomic changes occur in order to correctly execute a new developmental program. This leads to an epigenetic reset that creates a window of opportunity for TEs to escape the control of the host, leading to waves of TE expression. Whereas this was initially seen as an uncoordinated derepression of repetitive sequences, we now know that the activity of TEs is highly dynamic during early development, showing both stage and TE type-specific regulation (**Fig. 1.6**, Göke et al. 2015). The functional impact of TEs during these stages is not completely understood, but it has been shown that transcriptional activation of LINE-1 elements is essential for regulating chromatin accessibility and developmental progression during mouse preimplantation development (Jachowicz et al. 2017), and for the transcriptional regulation of LINE-1 enriched genes (Lu et al. 2014, 2020)and self-renewal in mouse ESCs (Percharde et al. 2018). Similarly, transcriptional downregulation of an ERV subfamily, MERV-L, in the mouse zygote leads to a developmental arrest at the two-cell stage (Huang

et al. 2017). Nevertheless, it remains to be determined whether this phenotype is a consequence of the depletion of the MERV-L transcripts, their protein products, or of the chimeric transcripts that are initiated within MERV-L LTRs. These few examples barely scratch the surface on the profound roles the expression and epigenetic regulation of TEs play during the early stages of embryonic development. Future studies should be aimed at identifying their full involvement, either beneficial or deleterious, in this developmental stage.

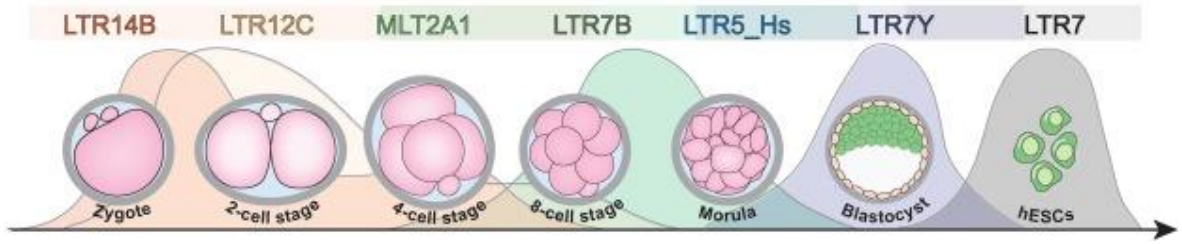


Fig. 1.6 - Example illustrating the transcriptional dynamics of different ERV subfamilies during human preimplantation development. ERVs from the HERVK14 (LTR14B) are expressed between the oocyte and the four-cell stage, LTR12C from the zygote to the eight-cell stage, HERVL (MLT2A1) around the 8-cell stage, LTR7B are found enriched in the eight-cell and morula, HERVK (LTR5_Hs) in morula and LTR7Y in the blastocyst. These ERV elements can regulate and initiate stage-specific transcription of the non-repetitive fraction of the transcriptome. Ultimately, the exaptation of distinct repeat elements in the molecular circuitry of specific lineages and/or developmental stages in different organisms created new mechanisms to regulate diverse cellular processes (from Casanova et al. 2016).

Whereas TEs are kept silenced in most somatic tissues, one organ escaping this rule is the brain, which shows an unusually high level of activity of TEs. Indeed, certain classes of TEs, in particular LINE-1 elements, are specifically active, at both the transcription and transposition levels in human neuronal precursor cells (NPCs), in neurons and glial cells from different regions of the human brain (Brattås et al. 2017; Coufal et al. 2009; Faulkner and Billon 2018; Upton et al. 2015). Higher TE activity in the brain has been associated with lower repression and the presence of brain-specific factors promoting TE expression (reviewed in Faulkner and Billon 2018). The amount of TE expression and levels of transposition in different brain regions and cell types remain unclear. Nevertheless, TE activity in this cellular context has been linked with the establishment of individual mosaicism in the brain and has been suggested to be important for brain development, neuronal diversity and functional connectivity. It is thus tempting to speculate that the transcriptional and transpositional dynamics of different classes of TEs during brain development and in the mature brain are important for the higher cognitive functions found in humans.

1.6. TEs as master regulators of naive and primed human pluripotency

Human embryonic stem cells (hESCs) are derived from early pre and post implantation embryos, capturing a particular stage of early embryonic development. These cells exhibit unlimited proliferation ability and have the potential to differentiate into various cell types of the embryonic and extraembryonic lineages. Because of these features, hESCs are considered as some of the most promising cellular candidates for the treatment of rare or incurable diseases, through the replacement of damaged cells in patients.

Interestingly, the expression of TEs was determined to be an important hallmark for defining the pluripotent state of hESCs (Theunissen et al. 2016). Pluripotency has been suggested to proceed through at least two distinct phases: naive and primed (Nichols and Smith 2009). Naive pluripotency represents a development stage corresponding to the inner cell mass (ICM) of the blastocyst, whereas primed

pluripotency corresponds to a more advanced stage found in the postimplantation epiblast. For this reason, naive and primed hESCs exhibit different developmental potentials, with naive cells being able to contribute to blastocyst chimaeras, whereas primed cells cannot. Moreover, these distinct states are characterised by marked differences at the morphological, molecular and metabolic levels. Remarkably, distinct subfamilies of primate-specific TEs have been found to exhibit dramatic state-specific expression profiles (Theunissen et al. 2016; Lu et al. 2014), and are now considered as markers for the naive and primed states. In the naive state, a mixed set of TEs is expressed, including different subfamilies of SVAs, as well as the LTR5/LTR5_Hs and HERVK-int subfamilies of ERVs (LTR5/LTR5_Hs correspond to the LTR regions and HERVK-int to the internal region of HERVK elements). Primed hESCs are characterised by a higher expression of ERVs from the LTR7 and HERVH-int subfamilies (LTR7 correspond to the LTR regions and HERVH-int to the internal region of HERVH elements). These elements have been suggested to play several *cis*-regulatory activities in hESCs, including acting as promoters for hESC-specific chimeric lncRNAs and protein-coding transcripts (Wang et al. 2014), or operating as state-specific enhancers, with a marked influence on the transcriptional state of neighbouring regions (Lu et al. 2014; Pontis et al. 2019). Thus, naive and primed hESCs are excellent models to test computational pipelines to explore the activity of distinct TE subfamilies.

1.7. Friends or foes: a permanent battle between TE expression and repression

Undoubtedly, TEs have been strongly influencing the way genomes are shaped. The evolutionary success and persistence of TEs in genomes reflects the delicate balance between their expression and repression in the genome of the hosts. Their expression has to warrant amplification of TEs, but not so extremely as to confer a fitness disadvantage to the host. Indeed, uncontrolled expression and transposition of TEs often have nefarious consequences on the genome of the host. As such, several host mechanisms exist to prevent TE expression and mobilisation. In somatic lineages, these include mechanisms of transcriptional silencing of chromatin, through the decoration of TEs with repressive chromatin modifications such as histone methylation of lysines 9 and 27 of histone H3 (H3K9me3 and H3K27me3, respectively), as well as DNA methylation (Walter et al. 2016; Mita and Boeke 2016; Almeida et al. 2022). The deposition of these marks depends on a series of enzymes, such as histone methyl-transferases (HMTs) and DNA methyl-transferases (DNMTs), which are crucial players for the long-term repression of TEs in host somatic tissues. These repressive mechanisms, in particular DNA methylation, are dynamically established during development and differentiation, which partly justifies the high transcriptional activity of different TE families observed in early preimplantation embryos and in ESCs (Seisenberger et al. 2013; Fort et al. 2014). The recruitment of HMTs and DNMTs to TEs is thought to occur in early embryonic cells, thanks to Krüppel-associated box domain zinc finger proteins (KRAB-ZFPs). Some KRAB-ZFPs were shown to bind different TE families (e.g. ERVs and LINEs) in a sequence-specific manner and to interact with the KAP1 protein, which acts as a scaffold for the recruitment of silencing complexes that include transcriptional repressors and epigenetic modifiers (Ecco et al. 2017; Imbeault et al. 2017). In germ cell lineages, where the above epigenetic mechanisms are less efficient, the host resorts to different mechanisms to control TEs, such as Piwi-interacting RNAs (piRNAs)-directed recruitment of epigenetic modifiers to TEs (Ernst et al. 2017).

The delicate equilibrium between TEs and their host is frequently broken in disease contexts. This often arises from compromised epigenetic mechanisms in the host leading to the aberrant expression and transposition of TEs. This has been shown to occur in various cancers, in particular of epithelial origin,

which results in cancer-specific transposition, but also in aberrant host gene transcription (reviewed in Payer and Burns 2019). Aberrant TE activity has also been documented in autoimmune diseases (Thomas et al. 2017) and in a myriad of neurological disorders, like amyotrophic lateral sclerosis (ALS), multiple sclerosis (MLS), schizophrenia, autism spectrum disorders (ASD) and Rett syndrome (RTT) (Muotri et al. 2010; Balestrieri et al. 2018; Gruchot et al. 2019; Küry et al. 2018; Saleh et al. 2019; Zhao et al. 2019). Hitherto, the levels of expression and transposition from different TE (sub)families in these different disease models remain poorly characterised. Moreover, whether enhanced TE activity is just a mere consequence of the disease or whether it could have an impact on pathogenicity remains to be formally tested.

1.8. Computational challenges for TE analysis

Despite the prevalence of TEs in genomes and their roles in physiological and pathological contexts, they have historically been neglected and often ignored from genomic studies. The reasons for this are deeply connected to their features: TE sequences are repetitive and interspersed; some TE insertions are polymorphic (some elements being present or absent from certain locus within a species, population or even individuals); TE transcripts are diverse and include full-length, short isoforms, chimeric RNA species (composed of TE and coding sequences) and even double-strand RNAs (**Fig. 1.7**), Lanciano and Cristofari 2020). Consequently, TE sequences are generally challenging to map in next-generation sequencing (NGS) datasets, resulting in their exclusion from further analysis. Moreover, genome assembly algorithms using short sequence reads struggle to correctly place TE sequences, resulting in imprecise annotations. To deal with these limitations, several bioinformatics tools have been developed to detect and analyse TEs. These tools tackle different aspects related to TEs, including classification and phylogeny, discovery and annotation of TEs in genomes, detection of polymorphic insertions and tools to characterise and predict the impact of TEs (Goerner-Potvin and Bourque 2018).

Unsurprisingly, the characterisation of the transcriptional dynamics of TEs in physiological and pathological contexts, also faces an inexorable problem. Mapping millions of reads derived from TE transcripts to a reference genome introduces a sizable amount of ambiguities in the mapping step. It is thus of paramount importance to carefully choose the parameters and algorithms to be used when the expression of TEs is investigated in sequencing datasets. Using simulated reads from high-throughput sequencing, a recent study benchmarked several alignment algorithms and mapping parameters (Teissandier et al. 2019). This study suggests a set of best practices to deal with repeat derived sequences and, importantly, how to deal with multi-mapped reads. These include the use of paired-end libraries to increase the uniqueness of sequenced fragments, the use of STAR (Dobin et al. 2013) for the alignment step, allowing the reporting of randomly one position for the multi-mapped reads and the use of featureCounts (Liao et al. 2014) to quantify the expression status of TE subfamilies. Whereas the above approach is robust when analysing the genome-wide dynamics of a given subfamily of TEs, it falters when inspecting the activity of individual TE copies. This is particularly obvious when mapping younger subfamilies of TEs. These groups had less time to diverge from the original TE copy from which they are derived and, as such, show a lower mappability. Strikingly, these younger TEs are also the ones usually showing a stronger activity and impact on genomic regulation. The low mappability of younger TEs results in a severe underestimation of their activity when using uniquely-mapped reads, and renders the study of these TEs at the individual level very challenging. Continuous development of computational tools to increase the efficiency of mapping strategies, together with the development of long-read sequencing technologies will surely stimulate, in an unprecedented manner, the mapping of TE elements, in particular in highly repetitive regions.

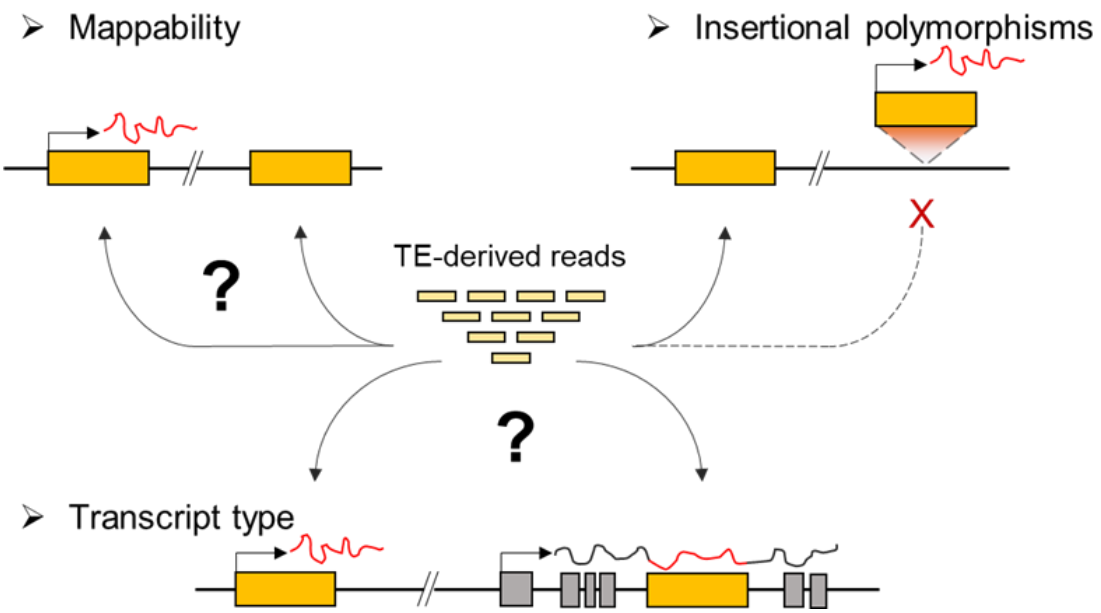


Fig. 1.7 - Major computational challenges when processing TE-derived reads in current generation sequencing platforms. Examples of the major difficulties faced when processing TE-derived reads: recently inserted TEs show low mappability due to their low sequence divergence; ongoing mobilisation of some TE families leads to high diversity of integrations sites and creates polymorphic TE insertions, not included in reference genomes; autonomous transcription of TEs can be easily confounded with chimeric transcripts or the expression of the gene into which a TE is inserted (adapted from Lanciano and Cristofari 2020).

2. Overall goal

Notwithstanding the increased interest in the biological and pathological functions of TEs, analysis of repetitive sequences still faces a significant amount of difficulties and limitations. The last decade has witnessed a steady development of computational pipelines and long read sequencing technologies, which have been pivotal in making the study of TEs more accessible. Nevertheless, a continuous effort on the development of "open-box" pipelines is crucial to endow researchers with decision-making support tools that help them make informed choices when exploring TEs. Importantly, existing tools offer limited strategies to shortlist TE subfamilies that are potentially associated with a given biological or pathological context.

With this in mind, the ultimate goal of this project was to create a tool to help identifying, in a robust and unbiased manner, TE subfamilies that show a context-specific regulation. For this, we started by developing and implementing a computational pipeline for the transcriptomic analysis of genes and TEs, at the subfamily and individual loci levels. Next, we focused on the implementation of distinct strategies to identify TE subfamilies, whose activity is potentially associated with particular biological or pathological contexts. Finally, we developed different approaches to test the association between differentially expressed genes and differentially expressed TEs. The computational strategy we developed could prove useful for the systematic characterization of the TEome in different physiological or pathological contexts. The functional impact of this upon normal or abnormal processes, in particular neurobiological, could then be tested in future experimental and functional studies.

3. Material and Methods

3.1. Experimental datasets

In this study, we used a total RNA-seq dataset that was previously generated in the group of Claire Rougeulle (Epigenetics and Cell Fate Center, Université de Paris, France) and has not yet been made publicly available. This dataset consists of 12 samples, corresponding to 6 different female H9 hESC lines[1] cultivated in two different conditions: under standard hESC culture conditions (primed pluripotency), using mTeSR1 medium (Stemcell Technologies) and Matrigel hESC-Qualified Matrix (Corning); reset to naive pluripotency using the NaïveCult Induction Kit (Stemcell Technologies) and Matrigel hESC-Qualified Matrix (Corning). As the original purpose of the dataset was to explore the role of the lncRNA *XACT* (Vallot et al. 2013) in primed to naive conversion of hESCs, 3 of the 6 H9 hESC lines carry an homozygous 90 kb deletion around the promoter region of *XACT*. As the effect of this mutation was found to have a very low impact on the transcriptomics profile of hESCs, we have decided to use this dataset to compare primed and naive pluripotency. Primed and naive hESCs have been shown to have a remarkably distinct and characteristic pattern of TE transcription (Theunissen et al. 2016). Thus, using a transcriptomics dataset of primed and naive hESCs constitutes an invaluable tool to test the proposed bioinformatics pipelines and assess how robust they are for the combined analysis of differentially expressed genes (DEGs) and differentially expressed transposable elements (DETEs).

For the preparation of this dataset, total RNA extraction was performed using Trizol (Life Technologies) followed by removal of genomic DNA contamination using TURBO DNA-free Kit (Life Technologies). Only samples showing a RNA integrity number (RIN) greater than 9 (TapeStation, Agilent Technologies) were used to prepare RNA-seq libraries. RNA concentrations were measured using a Qubit fluorometer (Invitrogen). For each sample, approximately 1000 ng of total RNA was used for library preparation using the Illumina TruSeq stranded total RNA Library Prep kit. Sequencing was performed in 100 bp paired-end reads using a Novaseq 6000 instrument (Illumina), with 90-100 millions paired reads per sample on average.

3.2. RNA-seq preprocessing and alignment to a reference genome

Quality control of the raw reads was performed with FASTQC v.0.11.9 (Andrews 2010) and sequencing adapters removed using Trim Galore v.0.6.6 (Krueger 2012). The genomic FASTA sequence for the human reference genome (hg38/GRCH38.p13) was downloaded from Gencode (Frankish et al. 2019). Adapter-free reads were then used to generate paired-end alignments using Spliced Transcripts Alignment to a Reference v.2.7.9a (STAR; Dobin et al. 2013). STAR is a fast RNA-seq read mapper that supports splice-junction and fusion read detection. This algorithm aligns reads by finding the Maximal Mappable Prefix (MMP) hits between reads (or read pairs) and the genome, using a Suffix Array Index. It has been developed to efficiently map different parts of a read to distinct genomic positions, thus allowing the detection of splicing or RNA-fusions. To maximise our ability to align reads coming from TEs, we have used two distinct strategies (**Table 3.1**), following the recommendations of a previous publication exploring tools and practices for TE analysis (Teissandier et al. 2019). A first strategy consists of retaining only uniquely-mapped reads, allowing the precise quantification of the activity of individual TE loci. This strategy suffers with the low mappability rate of evolutionarily young

[1] https://www.wicell.org/home/stem-cells/catalog-of-stem-cell-lines/wa09.cmsx

TE families, which are severely underestimated. A second strategy was used to assign randomly one position to multi-mapped reads. This strategy allows a robust quantification of the global expression levels of TE subfamilies, compromising the ability to identify which individual TEs are active.

Table 3.1 - Mapping strategies used for STAR aligner.

Mapping mode	Parameters	Goal
Uniquely-mapped reads	--outSAMtype BAM SortedByCoordinate --runThreadN 72 --alignIntronMax 500000 --alignMatesGapMax 500000 --alignEndsType EndToEnd --outFilterMultimapNmax 1 --outFilterMismatchNmax 999 --outFilterMismatchNoverLmax 0.03 --seedMultimapNmax 20000 --outSAMattributes All --outSAMmultNmax 1	Quantify expression of TEs at the individual loci level
Random assignment of multi-mapped reads	--outSAMtype BAM SortedByCoordinate --runThreadN 72 --outFilterMultimapNmax 5000 --outSAMmultNmax 1 --outFilterMismatchNmax 999 --outFilterMismatchNoverLmax 0.06 --outSAMattributes All --limitBAMsortRAM 40000000000	Quantification of global expression levels of TE subfamilies

3.3. Creation of custom TE annotations

A comprehensive annotation of gene and transcript models for a reference genome, is essential for transcriptomics studies, allowing to accurately count the number of reads mapping to genes. A common file format to store gene and transcript annotation information is the Gene Transfer Format (GTF). A GTF file contains the coordinates of genes and exons of a transcript, as well as information about the strand the transcript is generated from, gene name, coding portion of the transcript, alternative transcription start sites, amongst others. The GTF file for the comprehensive gene annotation on the reference chromosomes, scaffolds, assembly patches and alternate loci (haplotypes) was downloaded from Gencode (Gencode Release 39; Frankish et al. 2019). GTF files for TE annotations were customly generated from the UCSC RepeatMasker track using a custom-made function, *makeGTF*[2]. This function transforms the UCSC annotation file, creating GTF files that contain custom attributes for each annotated TE, including information about TE subfamily (e.g. L1Hs), family (e.g. LINE-1) and class (e.g. LINE). In addition, it creates an unique ID (e.g. L1Hs_dup1) that allows identifying each individual TE. Low complexity repeats and repetitive structural RNA (e.g. rRNAs, tRNAs, snRNAs, etc.) are removed from the GTF file. Two distinct GTF files are created to count reads per individual TEs or per TE subfamily. Gene and TE GTFs are then merged to allow simultaneous quantification of gene and TE expression.

3.4. Gene and TE expression quantification

To count the number of reads mapping to genes or TE, we have used featureCounts v.2.0.3 (Liao et al. 2014) and the custom-made GTF files described above. featureCounts is a program that counts how many reads map to genomic features, such as genes, exon, promoter and genomic bins. Uniquely-

[2] https://github.com/milcs40/TEScripts/blob/main/makeGTF.R

mapped reads or randomly multi-mapped reads covering gene exons, or TEs were counted using the following parameters: *featureCounts -g gene_name -M -p -C -s 2 -T 35*.

3.5. Differential expression analysis

3.5.1. R Statistical Software

All the analyses were performed in R Studio (v.1.4.1717; RStudio Team 2020), using R programming language (v.4.1.2; R Core Team 2021), which provides a wide range of statistical and visualisation solutions. R was used on most steps of the analysis, from filtering of count tables obtained after alignment and quantification of expression, until the production of the plots contained in this work. Several Bioconductor packages and packages from The Comprehensive R Archive Network (CRAN) were used to perform all analyses, including visualisation, normalisation, PCA, clustering, differential expression analysis, functional enrichment analysis and testing genomic association between genes and TEs. A summary of the main functions and packages used in this work is presented in **Table 3.2**.

3.5.2. Differential expression analysis using DESeq2

For differential expression analysis (DEA), we have used DESeq2 (Love et al. 2014). This widely used tool estimates the variance-mean dependence in count data from high-throughput sequencing datasets and tests for differential expression using generalised linear models based on a negative binomial distribution.

3.5.2.1. Filtering

Prior to DEA, the count tables obtained from *featureCounts* were filtered to discard lowly expressed genes or TEs, by keeping only those with more than 10 reads in at least 6 samples, and with more than 100 reads across all the 12 samples. For the random mapping strategy, the filtered count table contains 24,875 genes and 1,155 TE subfamilies. For the uniquely mapped strategy, the filtered count table contains 23,250 genes and 151,001 individual TE loci.

3.5.2.2. Count data normalisation

When preparing a library for RNA-seq, there are several biological and technical factors that can influence the library size and composition. In order to minimise the influence these have on measuring gene expression, it is essential to normalise samples to mitigate the influence of technical variability while maintaining biological variability. DESeq2 normalisation is achieved by calculating a median of ratios of gene counts relative to the geometric mean per gene. The counts for a gene in each sample is then divided by this mean. The median of these ratios in each sample is the estimated size factor for the respective sample. This will correct for sequencing depth and RNA composition bias, when only a small number of genes are very highly expressed in one condition but not another, for example. This normalisation is achieved using the *estimateSizeFactors()* function.

3.5.2.3. Exploratory data analysis

An important initial step, when dealing with high-dimensionality data like that generated by RNA-seq, is to perform exploratory data analysis (EDA). This consists of performing initial investigations of the

data in order to discover patterns, spot anomalies, test hypotheses, or check assumptions resorting to summary statistics and graphical representations. Several approaches and techniques of EDA exist, but in this work we have only used clustering and dimension reduction techniques, which help to create graphical representations of high-dimensional data containing many variables (genes/TEs, in this case).

We initially performed sample clustering, using either euclidean distances or pearson correlation. For this, normalised counts are first transformed using regularised logarithm (using the *rlogTransformation()* function from DESeq2), in order to remove the dependence of the variance on the mean, in particular the high variance on count data, when the mean for gene count is low. Clustering of samples, for both euclidean distances and pearson correlation, was performed using *pheatmap* and Ward's minimum variance method for clustering.

Next, we used Principal Component Analysis (PCA), a dimensionality reduction technique used to increase the interpretability of the data, while preserving the maximum amount of relevant information. PCA is based on the discovery of pairs of eigenvectors/eigenvalues, *i.e.* orthonormal vectors and their associated lengths, ordered by decreasing percentage of variance explained. By doing so, PCA allows projecting high dimensionality data into a new coordinate system, while preserving as much variance in the data as possible. This approach allows to visually identify clusters of samples with similar transcriptomic expression profiles. For performing PCA and visualisation, we have used the *PCAtools* package, using the regularised logarithm transformed count data and removing 10% of the genes showing the least variance across the samples.

3.5.2.4. Experimental design and identification of DEGs and DETEs

An essential step for DEA is to specify a design indicating how to model the samples. As such, a design formula has to be defined, that expresses the variables that will be used for modelling. For our analysis, we defined a design model that measures the effect of the pluripotent state of the samples (primed vs. naive), controlling for differences due to the genotype of the samples (*XACT* WT vs. *XACT* KO) and considering an interaction term to test whether the genotype could influence the results obtained by comparing pluripotent states. This was achieved using the following command: *DESeqDataSetFromMatrix(countData = countData, colData = sample_info, design = ~ 1 + genotype + state + genotype:state)*.

For DEA, DESeq2 assumes that gene (and TE) counts follow a negative binomial distribution, first calculating the dispersion estimates for gene (and TE) counts for the above model. Then using the "*state_naive_vs_primed*" coefficient determined in the design matrix, it will apply negative binomial generalised linear models to test the significance of the coefficient for each gene (and TE), using Wald test. This test will perform pairwise comparisons where the null hypothesis is that the log-fold change (the gene/TE model coefficient) is zero, indicating no significant differential expression between two groups. All of the above is accomplished with the *DESeq(DESeq2Object)* function. The results for the tests for each gene/TE are then obtained using: *results(DESeq2Object, independentFiltering = TRUE)*.

Table 3.2 - Summary of the main R packages used in this work.

Package	Version	Description
Data manipulation and visualisation tools		
tidyverse	v.1.3.2	Collection of open source packages for R that share an underlying design philosophy, grammar, and data structures of tidy data. It includes several data manipulation packages like *tidyr* and *dplyr*, and data analysis packages, such as *ggplot2* and *tibble* (Wickham et al. 2019).
data.table	v.1.14.4	Used for efficiently working with tabular data in R[3].
zeallot	v.0.1.0	Allows multiple, unpacking, or destructuring assignments in R[4].
ggplot2	v.3.3.6	R package dedicated to data visualisation. Used for declaratively creating graphs, based on The Grammar of Graphics[5].
ggrepel	v.0.9.1	Provides text and label geoms for *ggplot2* that help avoiding overlapping text labels[6].
Data exploration tools and differential expression analysis		
pheatmap	v.1.0.12	Used to draw clustered heatmaps in R[7].
PCATools	v.2.6.0	Bioconductor package providing functions for data exploration via PCA, including visualisation[8].
DESeq2	v.1.34.0	Bioconductor package that estimates variance-mean dependence in count data from high-throughput sequencing experiments and tests for differential expression by use of negative binomial generalised linear models (Love et al. 2014).
fgsea	v.1.20.0	Bioconductor package for fast preranked gene set enrichment analysis (GSEA). Enrichment analysis allows to determine whether a set of genes (or TEs) shows statistically significant and concordant differences between two given phenotypes (Korotkevich et al. 2021).
Genomic analysis Tools		
regioneR	v.1.26.1	Bioconductor package that provides a statistical framework based on customizable permutation tests to query the association between genomic region sets and other genomic features (e.g. DEGs and DETEs) (Gel et al. 2016).
refGenome	v.1.7.7	Allows importing and managing of genome annotation data from Ensembl genome browser and UCSC genome browser[9].
bedtoolsr	v.2.30.0-4	R package that provides a wrapper for bedtools functions, a widely used set of utilities for genomic analysis (Patwardhan et al. 2019).
GenomicRanges	v.1.46.1	Bioconductor package that allows to store and manipulate genomic intervals and variables defined along a genome (Lawrence et al. 2013).

[3] https://github.com/Rdatatable/data.table
[4] https://github.com/r-lib/zeallot
[5] https://ggplot2.tidyverse.org/
[6] https://github.com/slowkow/ggrepel
[7] https://github.com/raivokolde/pheatmap
[8] https://github.com/kevinblighe/PCAtools
[9] https://github.com/cran/refGenome

To define the lists of candidate DEGs and DETEs, we first need to determine statistical thresholds. In our analysis, we have used different strategies to select these parameters. For visualisation purposes, we have chosen to define a threshold for the Wald value. To decide on the threshold value, we have calculated the Wald statistic values in 100 permutations in which the labels of the 12 samples were randomly swapped (**Supp. Fig. 7.1**). The distribution of values of Wald statistics for these 100 permutations, was then compared to the observed values, and values above |3| were considered significant for differential expression. For experiments in which a list of up and downregulated DEGs or DETEs had to be defined, an adjusted p-value below 0.01 was used. The magnitude of differences in gene expression was measured in log2 fold-change. Values of 2 or 1 were considered to define differential expression of genes or TEs, respectively.

3.6. Selecting TE subfamilies

DEA generates a vast list of DEGs and DETEs (both at the subfamily and individual loci levels), between primed and naive H9 hESCs. In order to identify TE subfamilies that might have a more robust association with a particular pluripotent context, we have implemented a series of approaches based on the DEA for individual TEs or TE subfamilies.

For this, we started by developing a function (*createTETable*[10]) to create tables with information about TEs from UCSC or Repeatmasker annotation files. This function generates tables containing information about individual repeats, including TE subfamily, family and class, as well as genomic information, such as location, size and strandness. Three tables are generated: a table with basic individual repeat information; a table with information about individual repeats and statistics about family and classes; and a summarised table, with condensed information about repeat subfamilies.

3.6.1. Using randomly-mapped information and DEA at the TE subfamily level

We next employed different strategies to identify relevant TE subfamilies. First, using the results obtained from the DEA at the subfamily level, we selected the top up- and down-regulated subfamilies, selected based on the value for the Wald statistics.

3.6.2. Using uniquely-mapped information to measure proportion of DETE loci within each subfamily

Second, using information from the DEA of individual TEs (using uniquely-mapped reads), we measured the proportion of individual TE loci that are differentially expressed per subfamily in order to help discern between subfamilies that show a coordinate transcriptional regulation from those that show a loci specific regulation.

3.6.3. Using uniquely-mapped information and functional enrichment analysis

Third, we used an approach inspired in Gene Set Enrichment Analysis (GSEA) to identify subfamilies associated with a particular pluripotent context. GSEA is a computational method used to infer whether a defined list of genes shows statistically significant and concordant differences between two biological states (Subramanian et al. 2005). GSEA usually takes a list of genes ranked by a given statistic and tests

[10] https://github.com/milcs40/TEScripts/blob/main/createTETable.R

whether these genes belong to a gene set of interest (biological process, pathway, etc.), increasing or decreasing a running-sum statistic. After processing all the ranked list of genes, it determines an enrichment score (ES) and a statistical significance for each gene set tested, considering that if the gene set is enriched at the top or bottom of that list (that is, under or over-expressed), it is thought to be related to phenotypic differences.

Here, we have used a similar approach to test the functional enrichment of TE subfamilies with primed and naive pluripotent states. For this, we started by creating TEsets for all TE subfamilies. These subfamily specific TEsets include the unique name identifiers for all individual TEs belonging to the respective subfamily. Next, we created a list of individual TE loci ranked by the value of Wald statistic for DEA of naive vs. primed. This list was then used to perform functional enrichment analysis, using the *fgsea* Bioconductor package (Korotkevich et al. 2021) and the following command: *fgseaMultilevel(pathways = TEGenesets, stats = listOfRankedTEs, eps = 0.0, nPermSimple = 10000, minSize = 15, maxSize = Inf)*. This produced a list of TE subfamilies that are potentially associated with naive or primed hESCs.

3.7. Testing the genomic association between DEGs and selected TE subfamilies

Testing the genomic association between DEGs and DETEs is crucial to explore whether these groups are able to influence the expression of one another. Thus, we have employed different statistical and quantitative approaches to measure the spatial relations between these groups.

3.7.1. Measuring the association using regioneR

To identify and test meaningful associations between our set of DEGs and DETEs, we have used the Bioconductor package *regioneR* (Gel et al. 2016). *regioneR* provides a statistical framework using customizable permutation tests to assess the association between genomic regions. We started by using the package *refGenome* to extract the coordinates of the -10 kb to +2 kb regions around the start of all annotated genes. This gene universe was then subset in up and down DEGs (log2FoldChange > |1| and padj < 0.01). To implement *regioneR* in our analysis, we created the *permTestMod* function. This function starts by using the *overlapRegions* function of *regioneR* to count the number of TEs of any given subfamily, overlapping with the -10 kb to +2 kb regions of up or down DEGs. Next, the function implements the *resampleRegions* function from *regioneR*, which takes the list of genes of interest (either up or down DEGs) and the list of all annotated genes (gene universe), and provides a random sample from the gene universe, with size equal to the number of genes of interest. Then, the number of overlaps between the -10 kb to +2 kb regions of this randomised set of genes and the different TE subfamilies is calculated using the *overlapRegions* function. The permutation test is performed for 1000 randomisations, and the number of real observed overlaps is compared with the number of overlaps for the randomised permutations. A one-sided p value is calculated as the number of times that an overlap value for the randomised permutations is more extreme than the real observed overlap (+1), divided by the total number of permutations.

3.7.2. Measuring the median distance between DEGs and selected TE subfamilies

Alternatively, to test the association between DEGs and DETEs, we have measured the distances between the start of upregulated genes in naive hESCs (up DEGs) or upregulated genes in primed hESCs

(down DEGs) and the closest element of the set of top TE subfamilies, identified using our functional enrichment analysis strategy. Briefly, we used the package *refGenome* to extract the start positions (-1 bp to 0 bp) of all annotated genes. The subset of start positions for up and down DEGs (log2FoldChange > |1| and padj < 0.01) was subsequently extracted. To measure the distances between gene start and the closest TE element of the selected TE subfamilies, we have used the *closest* function of the R implementation of bedtools, *bedtoolsR*. The distribution of observed distances was compared to the distribution of distances to a random subset of genes, generated using the *resampleRegions* function from the *regioneR* package. To test whether the distribution of distances between our set of TE subfamilies and the start of up or down DEGs is statistically different from the distribution of distances between these TE subfamilies and the start of a random set of genes, we applied the Wilcoxon Rank-Sum test. This non-parametric test allowed us to assess whether the distributions were significantly different compared to a similar-sized set of random genes. Additionally, we used the Kolmogorov-Smirnov test to further evaluate differences in the overall distribution shapes. Both tests were adjusted for multiple comparisons using the FDR correction. Additionally, we calculated the proportion of distance difference, defined as the relative difference between the median distances of DEGs and the random gene set. This was calculated as:

$$Proportion\ of\ Difference\ = \frac{Median\ DEG\ distance\ -\ Median\ Random\ Distance}{Median\ Random\ Distance} \times 100$$

To ensure that both statistical and biological significance were considered, we used the proportion difference as a measure of effect size, complementing the p-value obtained from the Wilcoxon and Kolmogorov-Smirnov tests. The need to perform a combined battery of tests was crucial, as with samples of this size, solely relying on p-values can lead to claim support for results with no practical significance, as p-values quickly reach zero (Lin et al. 2013). By combining statistical tests with effect size measures, we ensured that only meaningful associations between TE subfamilies and DEGs in naive and primed hESCs were highlighted.

3.8. Code availability

The R Markdown notebooks and scripts used in this study are publicly available at https://github.com/milcs40/MsC_Transposable_Elements. This repository contains the full set of computational analyses performed in this work, including differential expression analysis, genomic association testing, statistical thresholding, and methods to identify potentially interesting TE subfamilies associated with naive and primed human embryonic stem cells. The code is provided to ensure transparency, reproducibility, and to serve as a resource for future studies on transposable elements and genome regulation.

4. Results and Discussion

4.1. Developing a pipeline for the parallel differential expression analysis of genes and TEs

The transcriptomic analysis of TEs is a challenging task and can benefit from the use of good quality data, namely, paired-end strand-specific total RNA-seq datasets, which increase the uniqueness of sequenced fragments and capture transcripts that are not necessarily polyadenylated (Teissandier et al. 2019). In addition, the use of well-characterised biological systems, in which expression of TEs has been previously explored, can help in critically assessing the results obtained using the computational pipeline developed within this project.

Therefore, in this study, we used a private RNA-seq dataset of primed and naive hESCs, where the expression of distinct subfamilies of TEs has been extensively documented (Göke et al. 2015; Theunissen et al. 2016). This dataset is composed of 12 samples, half corresponding to primed hESCs and the other half to reprogrammed naive hESCs. Within each of these groups, half of the samples are wild-type and the other half have a deletion of a 90 kb region of the *XACT* gene, a lncRNA gene located on the X chromosome (see Materials and Methods section).

For the analysis of TE expression, we implemented a pipeline that allows the simultaneous analysis of differential expression of both genes and TEs (**Fig. 4.1**). Moreover, we employed multiple strategies for mapping TEs: randomly assigning one position in the case of reads that have multiple mapping locations in the genome (random mapping); retaining only reads that map to a single location in the genome (unique mapping). The first strategy allows to quantify more accurately the global expression levels of TE subfamilies, as all multi-mapped reads are assigned to one random repetitive region of the genome. The second strategy allows to study individual TE insertions, in a locus-specific manner. This latter approach, while allowing to explore the transcriptional status of individual TEs, has the caveat of being inefficient at mapping reads derived from evolutionarily younger TEs subfamilies, as these are more identical in sequence.

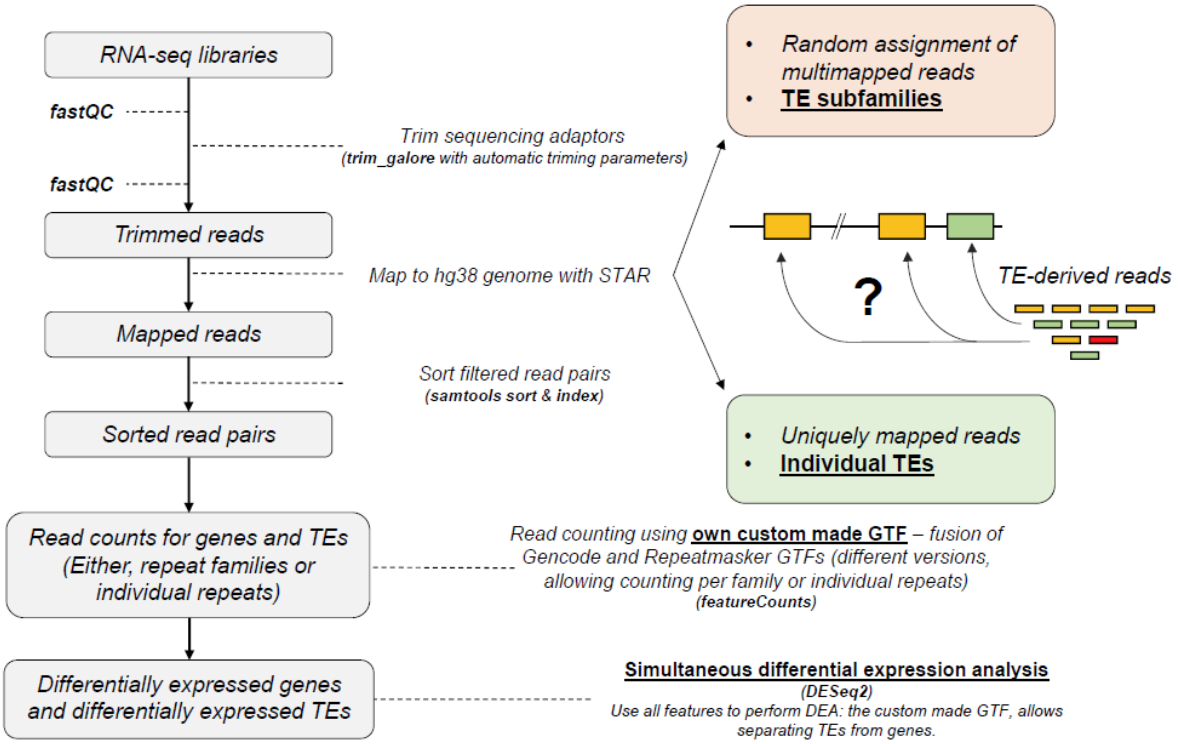


Fig. 4.1 - Pipeline used for analysing differential expression of genes and transposable elements. The diagram outlines the complete RNA-seq analysis pipeline, which includes quality control with fastQC, read trimming with Trim Galore, read mapping using STAR, read counting with custom made GTF files and differential expression analysis using DESeq2. The two mapping strategies for TE-derived reads are highlighted: random mapping to quantify expression at the TE family level, and unique mapping to assess expression of individual TE loci.

4.1.1. Creation of a custom made GTF file for human TEs

The efficient quantification of the number of reads assigned to TEs depends on the accurate annotation of these elements across the genome and the creation of GTF files that allow counting the number of reads mapped to each annotated TE. To this end, we have developed a R function, *makeGTF*, which takes TE annotations from either UCSC or Repeatmasker databases, removes low complexity repeats (e.g. rRNAs, tRNAs, snRNAs, etc.), labels each individual TE annotation with an unique identifier and creates an attribute column that includes, besides the unique identifier, information about the subfamily, family and class to which the TE belongs. In addition, it adds the prefix "TE_" to each TE, allowing to easily distinguish between genes and TEs in downstream analysis. By default, the function generates three different types of GTF files: an unmodified GTF file, a GTF file that allows counting TE subfamilies (**Table 4.1A**) and a GTF file that allows counting individual TE instances (**Table 4.1B**). The GTF files are then used by *featureCounts* to scan the BAM files containing the genomic coordinates of each read aligned to the reference genome, and counts the number of reads mapping to TEs. The GTF files allow using either the subfamily or the unique TE identifier attributes to count the number of reads mapping to each subfamily or to individual TEs, respectively.

Table 4.1 - Gene transfer format (GTF) tables with information of individual TE elements. GTF files contain multiple tab-separated fields, including name of the chromosome or scaffold (*seqname*), source of the annotation (*source*), name of the genomic feature type (*feature*; in this case, all TEs were labelled with '*exon*', as this is the feature that was used to quantify gene expression), start position of the feature (*start*), end position of the feature (*end*), Smith Waterman alignment score to a TE consensus sequence (*score*), relative orientation of the TE (*strand*), information about the coding frame (*frame*; as this does not apply to TEs, all TEs were labelled with '.') and a semicolon-separated list of tag-value pairs, providing additional information about each TE, including TE subfamily, unique instance, family and class (*attribute*). The attribute field is used to allow assigning the reads to TE subfamilies (**A**) or individual TE loci (**B**), by naming either the subfamily or the unique identifier with '*gene_name*', respectively.

A

chromosome	source	feature	start	end	score	strand	frame	attribute
chr1	hg38_rmsk	exon	10469	11447	3612	-	.	gene_name "TE_TAR1"; transcript_id "TAR1_dup1"; family_id "telo"; class_id "Satellite"
chr1	hg38_rmsk	exon	11505	11675	484	-	.	gene_name "TE_L1MC5a"; transcript_id "L1MC5a_dup4"; family_id "L1"; class_id "LINE"
chr1	hg38_rmsk	exon	11678	11780	239	-	.	gene_name "TE_MER5B"; transcript_id "MER5B_dup2"; family_id "hAT-Charlie"; class_id "DNA"
chr1	hg38_rmsk	exon	15265	15355	318	-	.	gene_name "TE_MIR3"; transcript_id "MIR3_dup12"; family_id "MIR"; class_id "SINE"
chr1	hg38_rmsk	exon	18907	19048	239	+	.	gene_name "TE_L2a"; transcript_id "L2a_dup30"; family_id "L2"; class_id "LINE"
chr1	hg38_rmsk	exon	19972	20405	994	+	.	gene_name "TE_L3"; transcript_id "L3_dup4"; family_id "CR1"; class_id "LINE"
chr1	hg38_rmsk	exon	20531	20679	270	+	.	gene_name "TE_Plat_L3"; transcript_id "Plat_L3_dup1"; family_id "CR1"; class_id "LINE"
chr1	hg38_rmsk	exon	21949	22075	254	+	.	gene_name "TE_MLT1K"; transcript_id "MLT1K_dup3"; family_id "ERVL-MaLR"; class_id "LTR"
chr1	hg38_rmsk	exon	23120	23371	787	-	.	gene_name "TE_MIR"; transcript_id "MIR_dup23"; family_id "MIR"; class_id "SINE"
chr1	hg38_rmsk	exon	23804	24038	312	+	.	gene_name "TE_L2b"; transcript_id "L2b_dup16"; family_id "L2"; class_id "LINE"

B

chromosome	source	feature	start	end	score	strand	frame	attribute
chr1	hg38_rmsk	exon	10469	11447	3612	-	.	gene_id "TAR1"; gene_name "TE_TAR1_dup1"; family_id "telo"; class_id "Satellite"
chr1	hg38_rmsk	exon	11505	11675	484	-	.	gene_id "L1MC5a"; gene_name "TE_L1MC5a_dup4"; family_id "L1"; class_id "LINE"
chr1	hg38_rmsk	exon	11678	11780	239	-	.	gene_id "MER5B"; gene_name "TE_MER5B_dup2"; family_id "hAT-Charlie"; class_id "DNA"
chr1	hg38_rmsk	exon	15265	15355	318	-	.	gene_id "MIR3"; gene_name "TE_MIR3_dup12"; family_id "MIR"; class_id "SINE"
chr1	hg38_rmsk	exon	18907	19048	239	+	.	gene_id "L2a"; gene_name "TE_L2a_dup30"; family_id "L2"; class_id "LINE"
chr1	hg38_rmsk	exon	19972	20405	994	+	.	gene_id "L3"; gene_name "TE_L3_dup4"; family_id "CR1"; class_id "LINE"
chr1	hg38_rmsk	exon	20531	20679	270	+	.	gene_id "Plat_L3"; gene_name "TE_Plat_L3_dup1"; family_id "CR1"; class_id "LINE"
chr1	hg38_rmsk	exon	21949	22075	254	+	.	gene_id "MLT1K"; gene_name "TE_MLT1K_dup3"; family_id "ERVL-MaLR"; class_id "LTR"
chr1	hg38_rmsk	exon	23120	23371	787	-	.	gene_id "MIR"; gene_name "TE_MIR_dup23"; family_id "MIR"; class_id "SINE"
chr1	hg38_rmsk	exon	23804	24038	312	+	.	gene_id "L2b"; gene_name "TE_L2b_dup16"; family_id

4.2. Exploratory analysis of hESC transcriptomic data

As explained above, our transcriptomics analysis followed two parallel and complementary approaches to quantify gene and TE expression: a random mapping approach, efficient at estimating the global transcriptional state of different TE subfamilies; and a unique mapping approach, allowing the measurement of the expression of individual TEs and a locus-specific perspective on the role and regulation of specific TEs insertions. The results of these two approaches will be presented and discussed in the next sections.

4.2.1. Primed and naive hESCs exhibit very distinct expression profiles

Naive and primed hESCs are very distinct cellular pluripotent states, corresponding to different developmental stages of the embryo (Nichols and Smith 2009). We have thus profited from this extensively studied cellular system to test our computational pipeline.

We started by exploring our data, in order to find patterns in our dataset. For this, we first measured the euclidean distances between samples to quantify the divergence between expression profiles, and observed a strong separation between naive and primed cells, forming two clear clusters (**Fig. 4.2A**). In addition, by performing a PCA, we also observed an obvious separation between naive and primed cells on the first principal component (PC) (**Fig. 4.2B**), indicating that the difference between states is the factor determining most of the variation in the dataset. Indeed, the variation explained in the PC1 ranged from 84.86% to 95.76%, in the unique mapping and random mapping approaches, respectively. Although the PC of both approaches explains slightly different percentages of the variance in the data,

as the underlying gene and TE expression is different, the difference between states is the factor that contributes the most for the variance in the data.

Intriguingly, the separation of the samples in PC2 changed, depending on the mapping strategy used. For the random mapping strategy, primed cells separate in primed WT and primed KO, even though the percentage of variation explained by PC2 is 0.81% (**Fig. 4.2B**). We further explored PC3, which explains 0.76% of the variation in the data, revealing that it separates naive WT and naive KO (**Supp. Fig. 7.2**). This suggests that the transcriptional differences imposed by the knock-out of the *XACT* gene are low compared to the pluripotent state of hESCs, which dominates the observed transcriptomic differences.

In contrast, in the PC2 for the unique mapping strategy, the primed samples all cluster together, whereas the naive hESCs separate in two clusters, depending on the genotype (**Fig. 4.2B**). In this case, the percentage of variation explained by PC2 is still rather low, at 2.08%. In addition, PC3, which explains 2% of the variance in the data, does not further separate samples into biologically relevant groups (**Supp. Fig. 7.2**). The differences observed might be explained by the high number of individual TEs that dominate PC2 in the unique mapping strategy. *XACT* is a very long lncRNA (~252 kb), with approximately 68% of its sequence derived from TEs like LINE and LTR elements, and showing a higher expression in naive hESCs (Vallot et al. 2013, 2017). It is possible that the clustering of naive samples in PC2 is influenced by the TEs contained within *XACT*, or under its regulation, although this hypothesis requires further investigation.

In summary, these observations show that for both mapping strategies, the bulk of the transcriptomic differences in our dataset is related to the cellular state of the cells, whereas the mutation of the *XACT* gene contributes little to the biological variation.

4.2.1.1. Differential expression analysis reveals genes and TE subfamilies differentially expressed in naive vs. primed hESCs

To infer global transcriptomic changes between primed and reprogrammed naive hESCs, we used generalised linear models (GLMs) implemented in DESeq2, using a design formula that defines the culture condition as the factor of interest to test for differential expression, but taking into account that some variation could be introduced from the different genotypes (see Materials and Methods). As we initially wanted to identify which subfamilies of TEs are globally deregulated between naive and primed hESCs, we decided to perform differential expression analysis (DEA) for both genes and TE subfamilies, using our random mapping approach. For determining differentially expressed genes (DEGs), a threshold of log2 fold-change (log2FC) of 2 and a minimum value for Wald statistics of 3 was considered. The statistical threshold was determined experimentally by calculating the distribution of Wald Statistics, in a series of 100 random permutations of sample labels (see Material and Methods, **Supp. Fig. 7.1**).

Using this approach, we identified a very large number of genes and TE subfamilies which are differentially expressed between the naive and primed states (**Fig. 4.3**). A positive value of log2FC represents genes/TEs more expressed in naive samples, whereas a negative value of log2FC represents genes/TEs more expressed in primed samples.

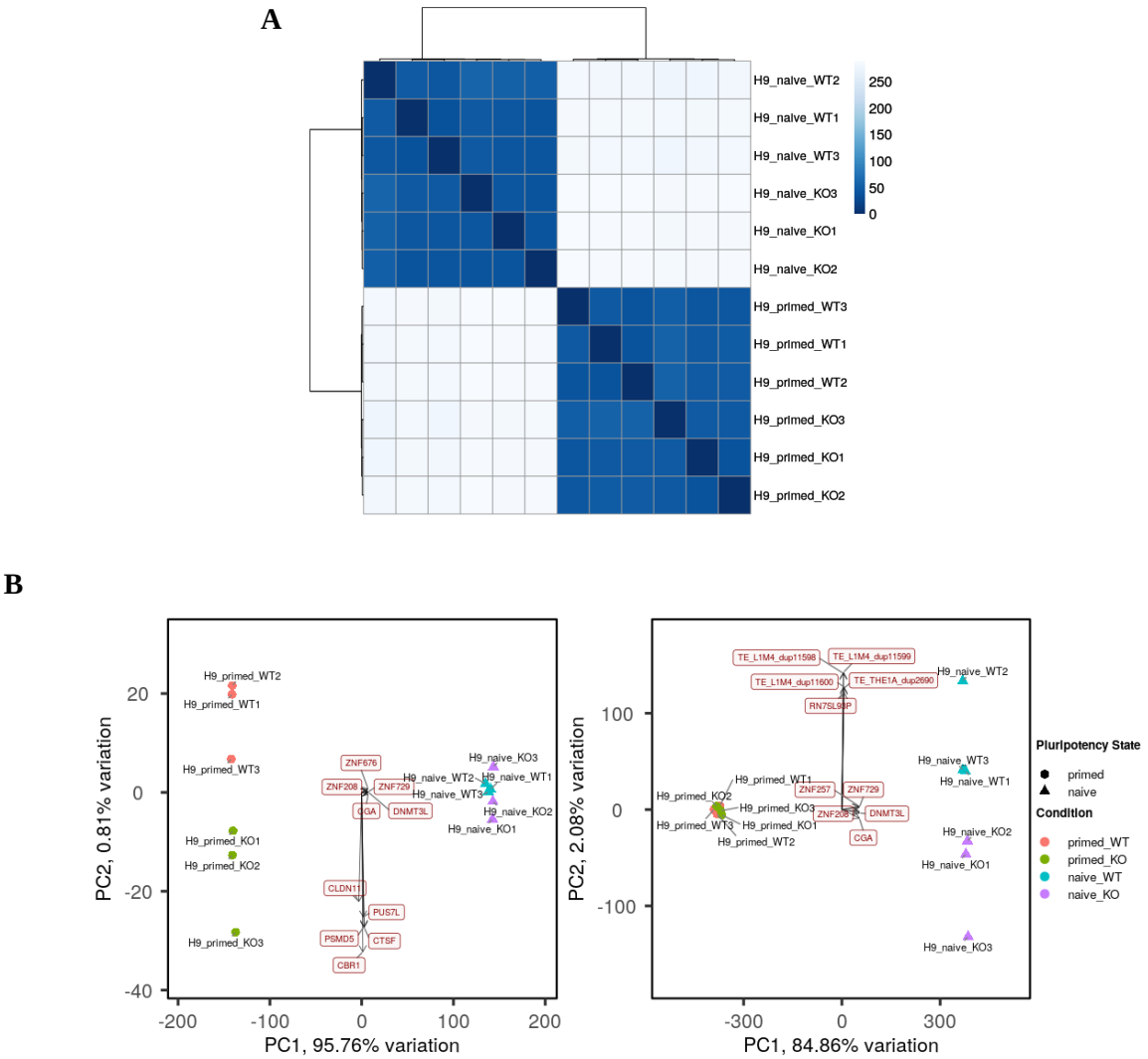


Fig. 4.2 - Distinct expression profiles of naive and primed hESCs datasets. A) Heatmap showing Euclidean distance measurements between naive and primed hESCs RNAseq datasets. The heatmap represents the Euclidean distance between samples based on normalised gene expression values, with clustering performed using the Ward.D2 method. Darker blue colours indicate closer proximity (more similar expression profiles), and lighter blue indicates greater distance between samples. The rows and columns represent individual samples, and hierarchical clustering was applied to group samples based on their expression similarities. **B**) PCA plots of the naive and primed hESCs datasets along the first (PC1) and the second principal components (PC2), for random mapping (left) and unique mapping (right) approaches. Samples are coloured by genotype and state (primed WT, primed KO, naive WT and naive KO) and shaped by pluripotency state (naive vs primed). Loadings for individual features (genes or TEs) are indicated by red labels and arrows, which show the contribution of each feature to the principal components. The percentage of explained variance by each PC is indicated on the respective axes.

Amongst the significant DEGs with the highest log2 fold-change, we found previously reported naive pluripotency and ground state marker genes, such as *DPPA3*, *DPPA5*, *KLF4*, *KLF17*, *DNMT3L*, *FGF4*, *GATA6*, *TBX3* and *IL6ST* (**Fig. 4.3; Fig. 4.4; Supp. Table 7.1**) (Guo et al. 2017; Messmer et al. 2019; Theunissen et al. 2016). In addition, we identified several genes encoding putative regulatory TFs, including *TRIM60*, and several KRAB-ZFP protein-coding genes, which have been shown to regulate the expression of different TE subfamilies in the genome, particularly in the early embryo (Imbeault et al. 2017; Ecco et al. 2017). In contrast, in primed hESCs, we observed upregulation of established marker genes of primed pluripotency, such as *MYC*, *DUSP6* and *PTPRZ1* (**Supp. Table 7.1**) (Messmer et al. 2019). Other known markers of pluripotency like *ZIC2*, *OTX2*, *SFRP2*, *DNMT3B* and *CD24*, were also strongly upregulated in primed hESCs (**Supp. Table 7.1**) (Messmer et al. 2019). Primed hESCs

also expressed a number of genes related to later developmental stages, including *SOX11*, *CYTL1*, *HMX2* or *THY1* (**Supp. Table 7.1**) (Messmer et al. 2019). Collectively, these observations demonstrate that the dataset used allows the unambiguous discrimination of primed and naive hESCs. Moreover, primed and naive hESCs reveal a remarkably distinct transcriptional program, characterised by a set of state-specific markers.

Allowing multi-mapped reads to be randomly assigned to a single target, might introduce bias in the analysis of the number of reads coming from a given gene. To confirm the robustness of using our random mapping approach for finding DEGs, we performed DEA using our unique mapping strategy. As expected, even if with this approach we include 151,001 individual TE loci in the DEA, the list of DEGs obtained is remarkably similar between the two strategies, validating our approach (**Supp. Fig. 7.3**). This result was expected, as most of the reads that map to multiple locations in the genome are generally associated with repeat-rich regions and not protein-coding genes.

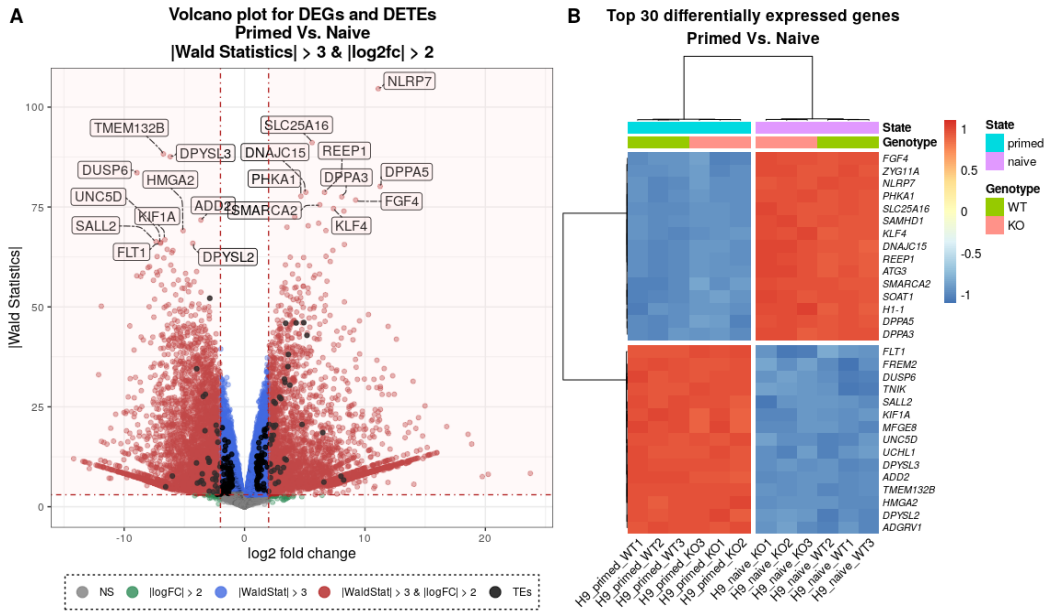


Fig. 4.3 - Primed and naive hESCs show very distinct transcriptomes. A) Volcano plot displaying expression fold changes (x-axis) and their significance (y-axis) for differentially expressed genes (DEGs) and transposable element subfamilies (DETEs) between naive and primed hESCs, using the random mapping approach. The horizontal red dashed line marks the log2 fold change threshold of 2, while vertical red dashed lines represent Wald statistics greater than 3. DEGs meeting the criteria of Wald > 3 and |logFC| > 2 are highlighted in red. DEGs with Wald > 3 and |logFC| < 2 are represented in blue. DEGs with Wald < 3 and |logFC| > 2 are represented in green. Non-significant (NS) DEGs are shown in grey. DETEs (subfamilies) are highlighted in black. Top DEGs are labelled. **B**) Heatmap displaying the top-30 differentially expressed genes between primed and naive hESCs. Rows represent genes, and columns represent samples, clustered using Ward's linkage method. Gene expression is scaled by row (z-score), with red representing high expression and blue representing low expression. Sample conditions are annotated by genotype (WT or KO) and pluripotency state (primed or naive).

The transcriptional dynamics of TEs in early embryonic development has been suggested to play a crucial role in the establishment of species- and stage-specific networks of transcriptional regulation (Göke et al. 2015; Yandım and Karakülah 2019). Importantly, naive and primed hESCs have been shown to express a unique TE signature, largely corresponding to early and late embryonic stages, respectively (see introduction) (Guo et al. 2017; Theunissen et al. 2016). With this in mind, we used our random mapping strategy to inspect the global expression of TE subfamilies in our dataset. DEA revealed a set of TE subfamilies preferentially upregulated in the naive state, notably HERVK-int, LTR5_Hs/LTR5, LTR7Y, LTR7B as well as several SVAs and *Alus* (**Fig. 4.5; Fig. 4.6; Supp. Table 7.2**). In contrast, in primed hESCs, we observed the upregulation of the LTR7 subfamily, which has been previously

identified as a marker of primed hESCs (Theunissen et al. 2016). In addition, we identified upregulation of other TE subfamilies, including HERV9NC and several LINE-1 subfamilies (**Fig. 4.6; Supp. Table 7.2**). These results are in line with previously reported observations and confirm that the random mapping strategy employed in this work is efficient for determining the global expression of TE subfamilies, concomitantly with gene expression.

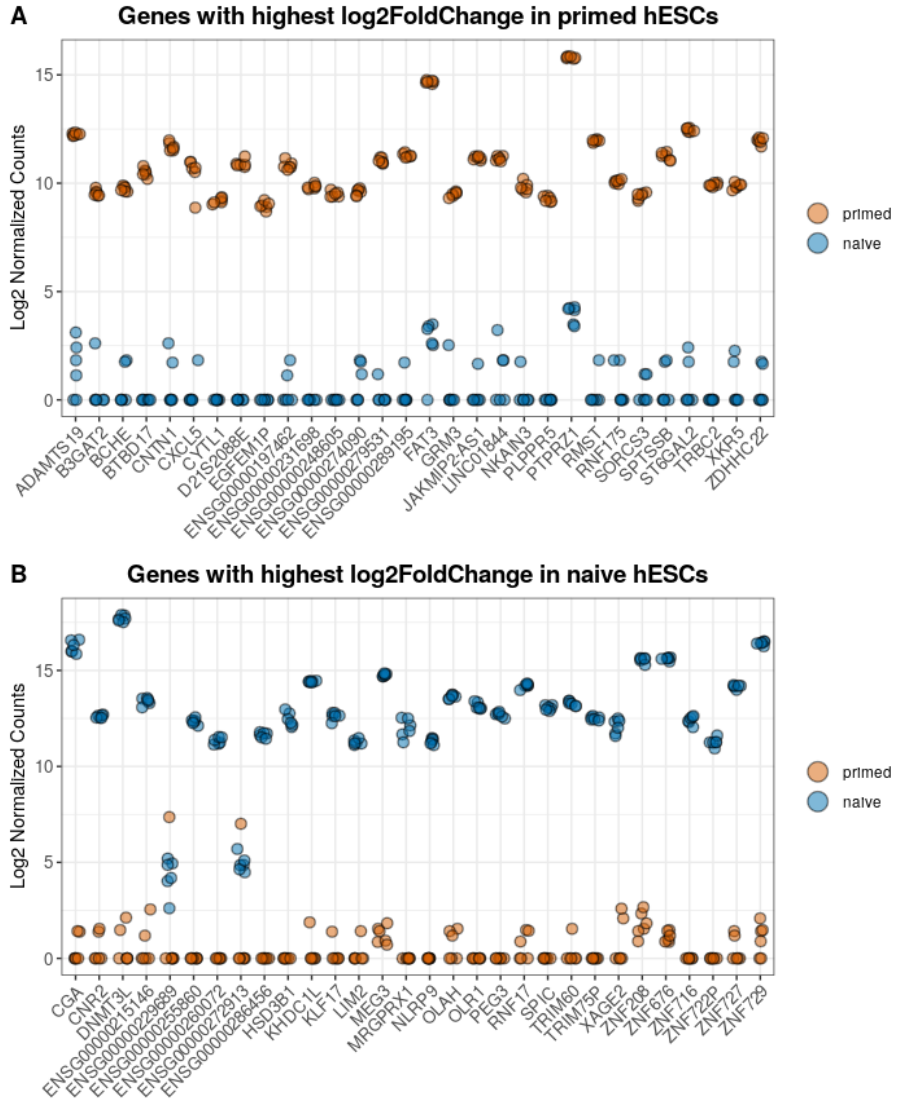


Fig. 4.4 - Top differentially expressed genes efficiently discriminate between pluripotent states. A) Dot plots showing normalised read counts (log2 scale) for the top-30 genes with the highest log2 fold change in primed hESCs. **B**) Dot plots showing normalised read counts (log2 scale) for the top-30 genes with the highest log2 fold change in naive hESCs. Each dot represents an individual sample, colour-coded based on pluripotency state (primed or naive). The position of the dots indicates the normalised read counts for each gene.

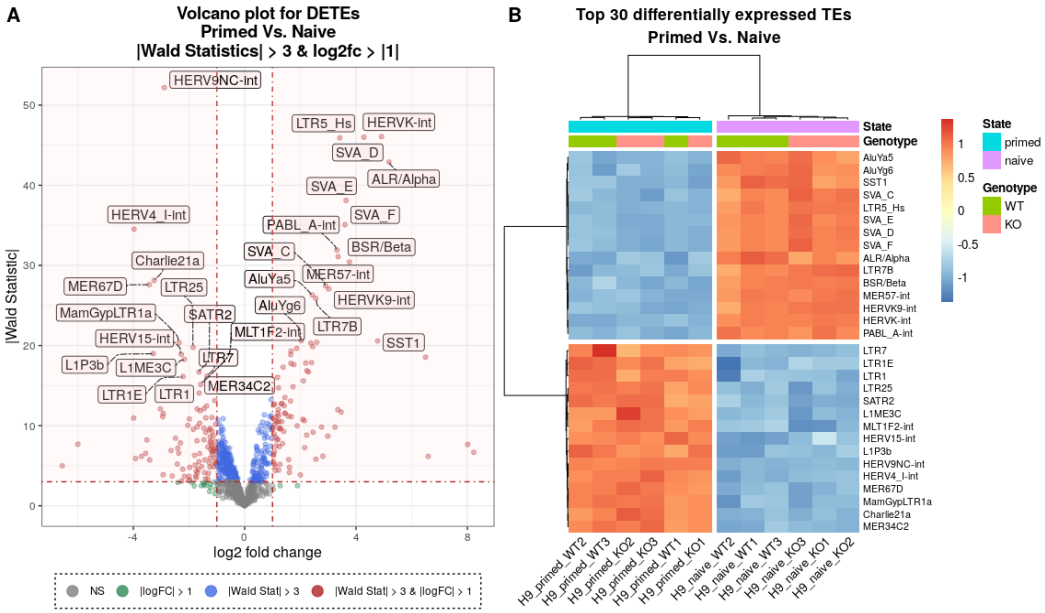


Fig. 4.5 - Primed and naive hESCs exhibit distinct profiles of TE expression at the subfamily level. A) Volcano plot displaying expression fold changes (x-axis) and their significance (y-axis) for transposable element subfamilies (DETEs) between naive and primed hESCs, using the random mapping approach. The horizontal red dashed line marks the log2 fold change threshold of 1, while vertical red dashed lines represent Wald statistics greater than 3. DETEs meeting the criteria of Wald > 3 and |logFC| > 1 are highlighted in red. DETEs with Wald > 3 and |logFC| < 1 are represented in blue. DETEs with Wald < 3 and |logFC| > 1 are represented in green. Non-significant (NS) DETEs are shown in grey. Top DETEs are labelled. **B**) Heatmap displaying the top-30 differentially expressed transposable element subfamilies between primed and naive hESCs. Rows represent TE subfamilies, and columns represent samples, clustered using Ward's linkage method. TE subfamily expression is scaled by row (z-score), with red representing high expression and blue representing low expression. Sample conditions are annotated by genotype (WT or KO) and pluripotency state (primed or naive).

4.2.1.2. The unique mapping strategy identifies individual TE loci that show differential transcriptional output in primed and naive hESCs

The transcriptional activity of individual TE loci might be distinct from the global activity of the subfamily to which they belong. In order to explore how individual TEs behave in primed and naive pluripotency, we performed DEA using our unique mapping approach. With this strategy, multi-mapped reads, which constitute a significant proportion of the total reads in short-read RNA-seq datasets, such as those from Illumina (Deschamps-Francoeur et al. 2020), are eliminated from our downstream analysis. As these reads are mostly derived from repetitive regions, TE subfamilies generally suffer from lower mapping percentages. This is particularly prevalent for younger subfamilies, in which the individual elements had less time to diverge from the original founder TE copy (Teissandier et al. 2019). As a consequence of this lower mappability and the fact that the bulk of the TE copies in the genome are transcriptionally inert, out of the 4,790,209 individual TEs annotated in the human genome, only 151,001 elements passed our threshold for a minimum of uniquely-mapped reads across all the samples in the dataset (see Materials and Methods).

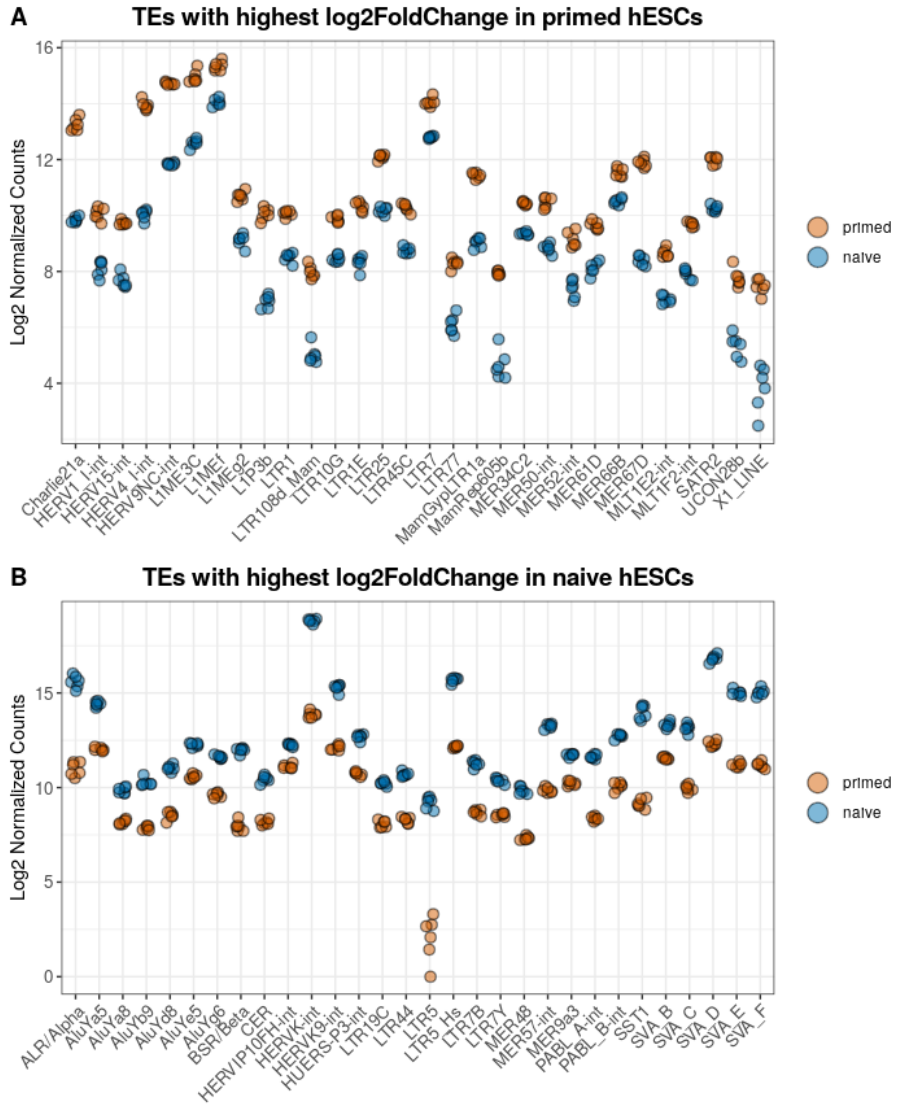


Fig. 4.6 - Top differentially expressed TE subfamilies efficiently discriminate between pluripotent states. A) Dot plots showing normalised read counts (log2 scale) for the top-30 TE subfamilies with the highest log2 fold change in primed hESCs. **B**) Dot plots showing normalised read counts (log2 scale) for the top-30 TE subfamilies with the highest log2 fold change in naive hESCs. Each dot represents an individual sample, colour-coded based on pluripotency state (primed or naive).

Similar to what we observed for the gene-centric analysis, a large number of individual TE elements were found to be differentially expressed between naive and primed hESCs (**Fig. 4.7**). In naive hESCs, the most differentially expressed TEs (DETEs) correspond to elements of the HERVK and L1PB subfamilies (**Fig. 4.7; Supp. Table 7.3**). Whereas the global upregulation of the HERVK subfamily is observed in naive contexts, the L1PB1 subfamily was not found to be differentially expressed in our random mapping analysis. This suggests that only a restricted subset of L1PB1 elements are strongly upregulated in naive hESCs, suggesting a local transcriptional regulation of specific insertions, rather than a coordinated upregulation of the whole subfamily. In primed hESCs, we observe overexpression of TEs belonging to primed-specific subfamilies, such as HERVH and HERV9NC. In addition, we observe overexpression of individual TEs belonging to multiple LINE-1 subfamilies, including L1M2, L1ME3C, L1MC5a and L1PA8A, which are globally upregulated in primed hESCs (**Fig. 4.7; Supp. Table 7.3**). Collectively, our data shows that primed and naive pluripotency are characterised by a distinct transcriptional profile of individual TEs. Whereas many identified DETEs belong to subfamilies that are associated with a specific pluripotent state, several other DETEs belong to subfamilies that are not globally associated with either primed or naive contexts. This suggests that the transcriptional

regulation of TEs is a complex and multifactorial process that might depend on the cellular context, but also on the availability of TFs to bind their regulatory sequences and the underlying chromatin status of individual TEs.

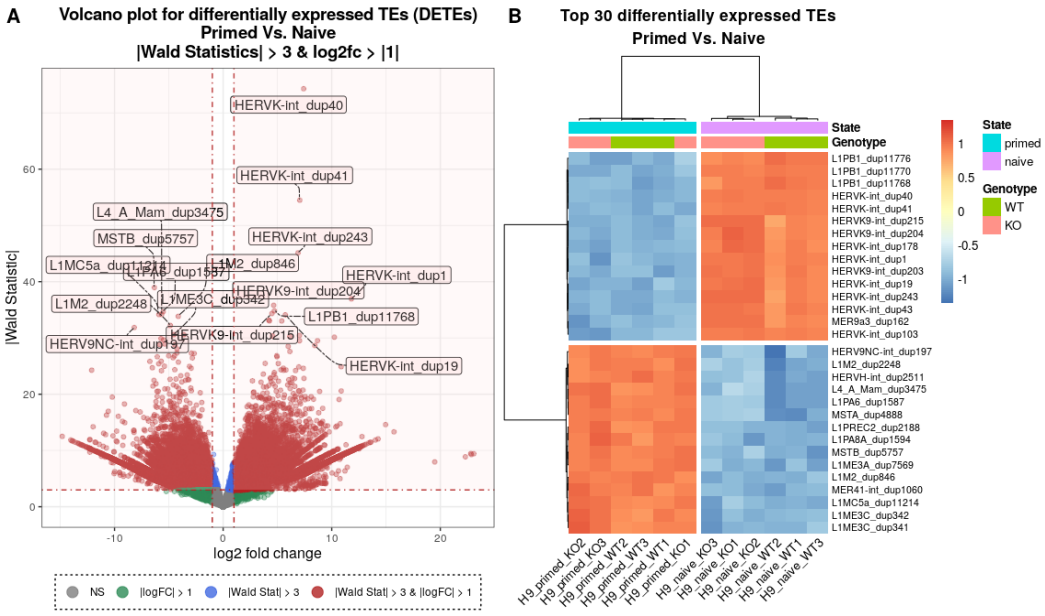


Fig. 4.7 - Primed and naive hESCs display distinct expression profiles of individual TE loci. A) Volcano plot displaying expression fold changes (x-axis) and their significance (y-axis) for differentially expressed transposable element loci (DETEs) between naive and primed hESCs, using the unique mapping approach. The horizontal red dashed line marks the log2 fold change threshold of 1, while vertical red dashed lines represent Wald statistics greater than 3. DETEs meeting the criteria of Wald > 3 and |logFC| > 1 are highlighted in red. DETEs with Wald > 3 and |logFC| < 1 are shown in blue. DETEs with Wald < 3 and |logFC| > 1 are represented in green. Non-significant (NS) DETEs are shown in grey. Top individual DETEs are labelled. **B**) Heatmap displaying the top-30 differentially expressed TE loci between primed and naive hESCs. Rows represent individual TE loci, and columns represent samples, clustered using Ward's linkage method. TE expression is scaled by row (z-score), with red representing high expression and blue representing low expression. Sample conditions are annotated by genotype (WT or KO) and pluripotency state (primed or naive).

4.3. Strategies to identify potentially interesting TE subfamilies

Transcriptomic studies, such as the one presented here, generally produce a wealth of data that is challenging to transform into biologically relevant information. Thus, different approaches combining a deep knowledge of the DEA methodology, statistics and biology, should be employed to try and refine the results and get an unbiased list of the most relevant differentially expressed targets.

With this in mind, we employed multiple strategies to define which TE subfamilies seem to be relevant in either primed or naive pluripotent contexts. A first strategy was focused on the top differentially expressed subfamilies identified with the random mapping approach. A second approach was based on using uniquely-mapped information, to identify the percentages of up and downregulated TE elements per subfamily. A third and final approach employed functional enrichment analysis and uniquely-mapped reads to identify subfamilies of TEs that are over-represented and might have an association with a particular pluripotent context.

4.3.1. Creation of TE information tables for downstream analysis

In order to test the importance of TEs and their subfamilies in distinct biological contexts, we needed a thorough and rich genomic information about individual TE elements and the subfamilies they belong to. For this, we developed a R function, *createTETable*, which takes TE annotations from either UCSC or Repeatmasker and creates three different TE tables: a table with information about individual TEs (**Table 4.2A**), a table with extended genomic information about individual TEs and their respective subfamilies, families and classes (not shown) and a summary table with information about annotated TE subfamilies (**Table 4.2B**). After eliminating low complexity repeats (e.g. rRNAs, tRNAs, snRNAs, etc.), our TE information tables comprise 1265 different TE subfamilies, organised in 47 families and 9 classes.

Table 4.2 - Annotation tables of individual TE loci and their respective subfamily, family, and class.

A) Example of 10 entries from the TE annotation table, detailing genomic location, size, strand information, the name of the TE instance and the corresponding subfamily, family, and class for each individual TE.

chromosome	start	end	size	strand	individualRepeat	subfamily	family	class
chr1	10469	11447	978	-	TAR1_dup1	TAR1	telo	Satellite
chr1	11505	11675	170	-	L1MC5a_dup4	L1MC5a	L1	LINE
chr1	11678	11780	102	-	MER5B_dup2	MER5B	hAT-Charlie	DNA
chr1	15265	15355	90	-	MIR3_dup12	MIR3	MIR	SINE
chr1	18907	19048	141	+	L2a_dup30	L2a	L2	LINE
chr1	19972	20405	433	+	L3_dup4	L3	CR1	LINE
chr1	20531	20679	148	+	Plat_L3_dup1	Plat_L3	CR1	LINE
chr1	21949	22075	126	+	MLT1K_dup3	MLT1K	ERVL-MaLR	LTR
chr1	23120	23371	251	-	MIR_dup23	MIR	MIR	SINE
chr1	23804	24038	234	+	L2b_dup16	L2b	L2	LINE

B) Example of 10 entries from the TE annotation table, providing genomic information for TE subfamilies and their corresponding families and classes, including the number of elements, total size (in kb), and percentage of genome occupancy (%).

Subfamily	number subfamily	size sub family (kb)	size sub family (%)	Family	number family	size family (kb)	size family (%)	Class	number class	size class (kb)	size class (%)
TAR1	165	102404	0	telo	441	308468	0.01	Satellite	9093	79808752	2.57
L1MC5a	18088	5191040	0.17	L1	1022089	549054102	17.71	LINE	1600696	674154623	21.75
MER5B	25817	2973688	0.1	hAT-Charlie	268836	47762034	1.54	DNA	511751	107069766	3.45
MIR3	88778	10135262	0.33	MIR	612281	87262743	2.82	SINE	1884272	416061066	13.42
L2a	176800	43859746	1.41	L2	482724	107628829	3.47	LINE	1600696	674154623	21.75
L3	46729	9065549	0.29	CR1	69112	12311859	0.4	LINE	1600696	674154623	21.75
Plat_L3	3225	512505	0.02	CR1	69112	12311859	0.4	LINE	1600696	674154623	21.75
MLT1K	17987	4013198	0.13	ERVL-MaLR	364558	116309238	3.75	LTR	769833	284211525	9.17
MIR	179429	28219968	0.91	MIR	612281	87262743	2.82	SINE	1884272	416061066	13.42
L2b	99844	18685072	0.6	L2	482724	107628829	3.47	LINE	1600696	674154623	21.75

4.3.2. Selecting top differentially expressed TE subfamilies based on a random mapping approach

Our first strategy consisted on selecting the top differentially expressed TE subfamilies, identified using a random mapping approach. For this, we sorted TE subfamilies using the Wald test value from our differential expression analysis, and selected a list of the top-15 TE subfamilies associated with either the naive or primed states (see section **4.2.1.1**, **Supp. Table 7.2**).

In order to explore how the individual TEs belonging to these subfamilies behave, we plotted their expression in primed vs. naive hESCs, using our unique mapping approach. The naive-associated TE subfamilies identified show a high number of upregulated individual elements in the naive context (**Fig. 4.8A**). These subfamilies include multiple SVAs subfamilies, HERVK-int and LTR5_Hs, which were also identified using the unique mapping strategy (**Fig. 4.7**). In contrast, the primed-associated TE subfamilies include the well characterised LTR7 (**Fig. 4.7**), as well as several LTR/ERV and LINE-1 subfamilies that were not identified using the unique mapping strategy and have a very low number of elements with mapped reads (**Fig. 4.8B**). Most of these LTR/ERV and LINE-1 subfamilies are low-copy and hence may not be very relevant. In this case, their global upregulation is likely driven by a small number of elements showing differential expression (e.g. LTR1, LTR25). As such, establishing a correlation between the global expression levels of subfamilies, obtained using a random mapping approach, with the transcriptional activity of individual TEs, obtained using a unique mapping approach, should be taken cautiously.

In summary, using this strategy, we are able to select TE subfamilies that are robustly differentially expressed between naive and primed hESCs. This provides a sound strategy to shortlist TE subfamilies based on their global transcriptional output, ignoring how individual TE elements behave. However, this approach has important limitations that have to be taken into consideration: first, it depends on defining empirical thresholds for determining what a differentially expressed TE subfamily is; second, it selects an arbitrary number of top subfamilies, disregarding other potentially relevant ones; third, TE subfamilies identified using a random mapping approach might not be representative, as they may be low-copy number subfamilies and/or include only a small number of misregulated elements.

4.3.3. Selecting TE subfamilies based on the percentage of up and downregulated TE instances

We next employed a simple approach of identifying the TE subfamilies with the highest percentage of individual TE elements overexpressed in either the naive or primed states (**Tables 4.3A-B**). For this, we used uniquely-mapped information to filter subfamilies with at least 30 elements with mapped reads and showing differential expression by considering an adjusted p-value lower than 0.05 and an absolute log2 fold-change of at least 1. We could immediately observe that the number of TE elements with mapped reads is lower than the number of total annotated TEs for any given subfamily. This can be justified by the low mappability of these repetitive elements, in particular those belonging to younger TE subfamilies, and with the fact that the majority of the TEs in the genome are transcriptionally silent.

Amongst the top subfamilies showing a higher percentage of upregulated elements, we found subfamilies typically associated with human naive pluripotency, including several SVAs, LTR5_Hs and HERVK-int (**Table 4.3A**). In contrast, the subfamilies showing a higher percentage of elements upregulated in primed hESCs (downregulated in naive hESCs) include LTR7 and HERVH-int, typically associated with primed pluripotency (**Table 4.3B**). Interestingly, the upregulation of TEs in naive hESCs seems to be more striking, with several subfamilies showing a global upregulation of most of its individual elements (from 60% to 90% of individual elements). In contrast, in primed hESCs, the upregulation of TE subfamilies is less pronounced (ranging from 50% to 68% of individual elements), with many subfamilies showing variable percentages of both up and downregulated elements. We can hypothesise that this is due to the lack of robust silencing mechanisms in naive hESCs, such as the DNA methylation machinery (Guo et al. 2017; Theunissen et al. 2016), which render the regulation of TE activity much more dependent on the presence of TFs in a given cellular state and less on epigenetic mechanisms to control their activity.

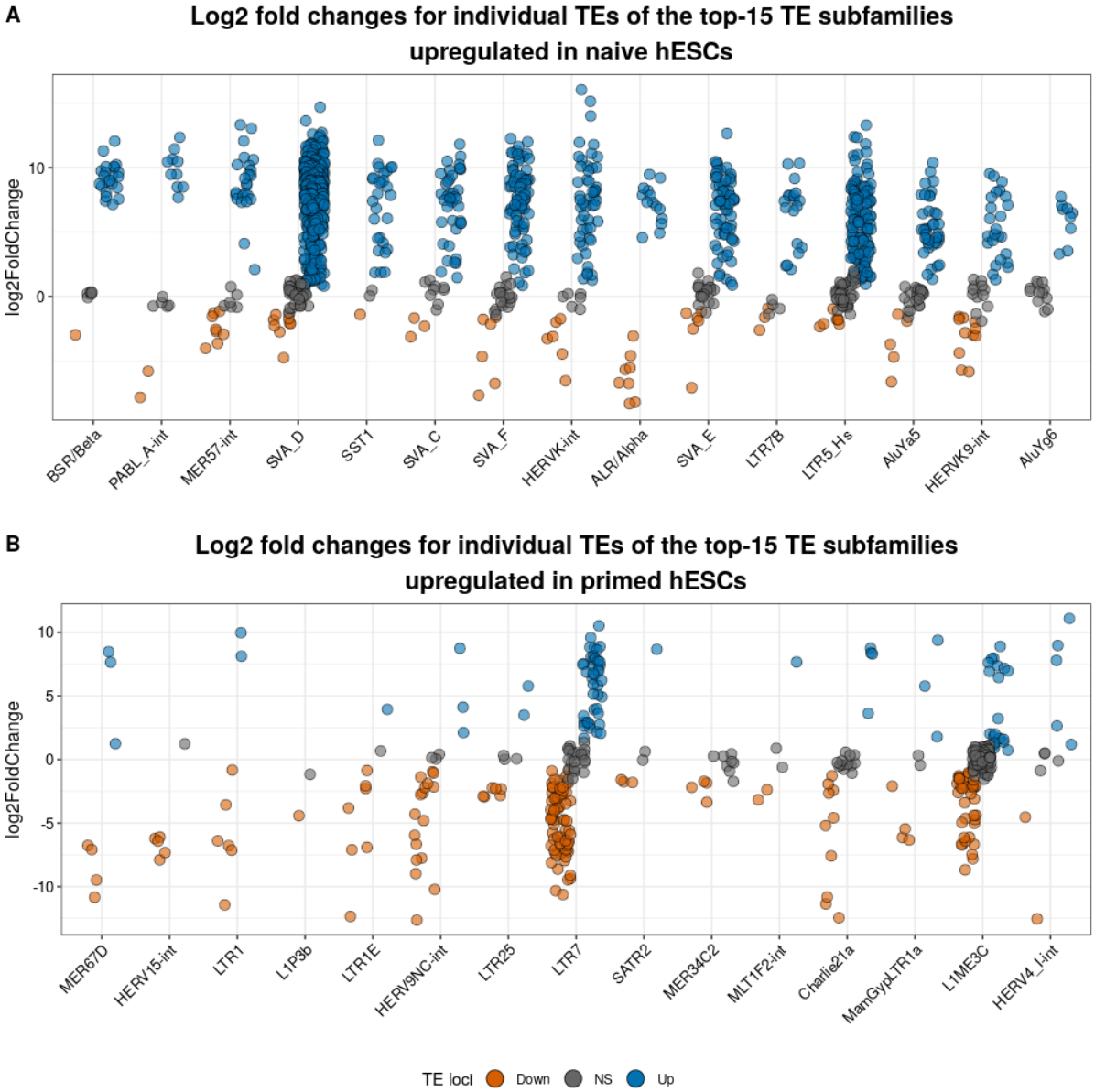


Fig. 4.8 - Expression of individual TE loci measured with the unique mapping approach follows the expression trend of their corresponding subfamilies measured by random mapping. A) Dot plot showing log2 fold changes for individual TE loci belonging to the top-15 TE subfamilies upregulated in naive hESCs, categorised as upregulated (blue), downregulated (orange), or not significantly changed (grey). **B**) Dot plot showing log2 fold changes for individual TE loci belonging to the top-15 TE subfamilies upregulated in primed hESCs (shown here as downregulated in naive), categorised as upregulated (blue), downregulated (orange), or not significantly changed (grey). Each point represents an individual TE locus.

In summary, this strategy provides a simple and straightforward approach to identify TE subfamilies whose elements are regulated in a coordinated manner, depending on the biological context. Whereas this approach is efficient when working with very distinct cellular contexts, like naive and primed hESCs, it might be ineffective when dealing with samples with more subtle differences. Moreover, it disregards the absolute values of the statistics for differential expression of individual TEs (like p-value and fold-changes), considering only if they are up or downregulated (i.e. log2FC > |1| and padj < 0.05). Finally, by discarding multi-mapped reads, this approach has a limited precision in estimating the number of elements truly affected per subfamily, which is particularly critical for young subfamilies with lower mappability.

Table 4.3 - Distinct TE subfamilies show different percentages of up and downregulated individual TEs in naive and primed contexts.

A) Table showing the top-15 subfamilies with highest percentage of upregulated TEs in naive hESCs. For each subfamily, the total number of elements in the genome, the number of expressed elements, the number and percentage of elements showing upregulation (Up), Downregulation (Down) or no significant fold change (NS), are indicated. Only subfamilies with more than 30 elements with detected reads are shown.

Subfamily	total number elements	number expressed elements	Up	Down	NS	% Up	% Down	% NS
SST1	665	31	28	1	2	90.32	3.23	6.45
SVA_D	1594	437	393	9	35	89.93	2.06	8.01
HERVK-int	296	64	52	6	6	81.25	9.38	9.38
SVA_C	529	53	41	3	9	77.36	5.66	16.98
SVA_F	1047	107	81	6	20	75.7	5.61	18.69
LTR5_Hs	717	186	139	7	40	74.73	3.76	21.51
SVA_E	715	90	66	6	18	73.33	6.67	20
SVA_B	874	70	51	5	14	72.86	7.14	20
HERVIP10FH-int	443	38	26	4	8	68.42	10.53	21.05
MER57-int	1348	38	24	8	6	63.16	21.05	15.79
MER61-int	1169	74	45	9	20	60.81	12.16	27.03
L1HS	1713	248	150	20	78	60.48	8.06	31.45
AluYa5	4058	66	39	5	22	59.09	7.58	33.33
LTR17	867	66	39	12	15	59.09	18.18	22.73
Harlequin-int	563	64	36	11	17	56.25	17.19	26.56

B) Table showing the top-15 subfamilies with highest percentage of upregulated TEs in primed hESCs. For each subfamily, the total number of elements in the genome, the number of expressed elements, the number and percentage of elements showing upregulation (Up), Downregulation (Down) or no significant fold change (NS), are indicated. Only subfamilies with more than 30 elements with detected reads are shown.

Subfamily	total number elements	number expressed elements	Up	Down	NS	% Up	% Down	% NS
L1M3e	854	32	22	3	7	68.75	9.38	21.88
LTR78B	3194	57	38	10	9	66.67	17.54	15.79
MamGypsy2-I	2333	52	32	5	15	61.54	9.62	28.85
HERVH-int	6137	346	204	117	25	58.96	33.82	7.23
L1M3f	711	41	23	5	13	56.1	12.2	31.71
LTR16A1	2751	33	18	6	9	54.55	18.18	27.27
LTR7	2485	137	71	42	24	51.82	30.66	17.52
Tigger14a	1255	35	18	4	13	51.43	11.43	37.14
MLT1H1	3862	55	28	13	14	50.91	23.64	25.45
MLT1G1	3721	57	29	9	19	50.88	15.79	33.33
THE1A-int	1508	105	53	22	30	50.48	20.95	28.57
LTR16	3112	32	16	7	9	50	21.88	28.12
THE1D-int	2763	102	51	21	30	50	20.59	29.41
MLT1K	17987	214	106	34	74	49.53	15.89	34.58
LTR78	4794	83	41	14	28	49.4	16.87	33.73

4.3.4. Selecting TE subfamilies based on functional enrichment analysis

Our previous approaches have some notable limitations: first, the arbitrary definition of threshold values (like those for fold-changes and adjusted p-values) can introduce a user bias in defining differential expression; second, looking at TE subfamilies as a whole does not allow distinguishing subfamilies that are globally regulated from those that have only a few elements that are highly differentially expressed; third, looking at individual TEs as isolated entities and their respective statistics for differential expression might hide small, but robust coordinated changes of TE subfamilies as a whole.

To address these caveats, we sought to employ an unbiased statistical method to identify subfamilies of TEs that may be relevant for a particular cellular state. For this, we performed functional enrichment analysis to discover subfamilies that are over- or under-represented amongst all individual TEs with detectable reads, using our unique mapping strategy. We started by creating "TE-sets" for each TE subfamily, comprising all the respective individual TEs with mapped reads. We next created a ranked list of individual TEs, sorted by the Wald test value for differential expression between naive and primed hESCs. This list was then used to check whether individual TEs from each subfamily concentrated in the extremes of the ranked list of differential expression. The accumulation of elements of a given subfamily at either the top (over-expressed) or bottom (under-expressed), suggests a possible association of the subfamily with the naive or primed states, respectively. Using this approach, we identified the top-20 subfamilies more significantly associated with either naive or primed hESCs (**Fig. 4.9**). Amongst the subfamilies over-expressed in naive cells (*i.e.* with high normalised enrichment scores (NES)), we observed SVAs, LTR5_Hs, HERVK-int, and several *Alu* subfamilies. This observation is in agreement with previous reports, suggesting these subfamilies are preferentially active in naive pluripotency. Moreover, we observed the enrichment of the human-specific L1Hs and primate-specific L1PA2, LINE-1 subfamilies, which have not been previously reported as being associated with the naive state and which were not identified with the previous approaches described above. In contrast, primed hESCs show upregulation for HERVH-int and LTR7, as previously documented. Moreover, several LINE subfamilies (belonging to the LINE-1, -2 and -3 families), as well as MIR, are upregulated in primed hESCs (**Fig. 4.9**). Altogether, this shows that this approach is robust at identifying subfamilies that are known to be associated with a given cellular context and, moreover, allows identifying additional TE subfamilies that might be missed using more conventional strategies.

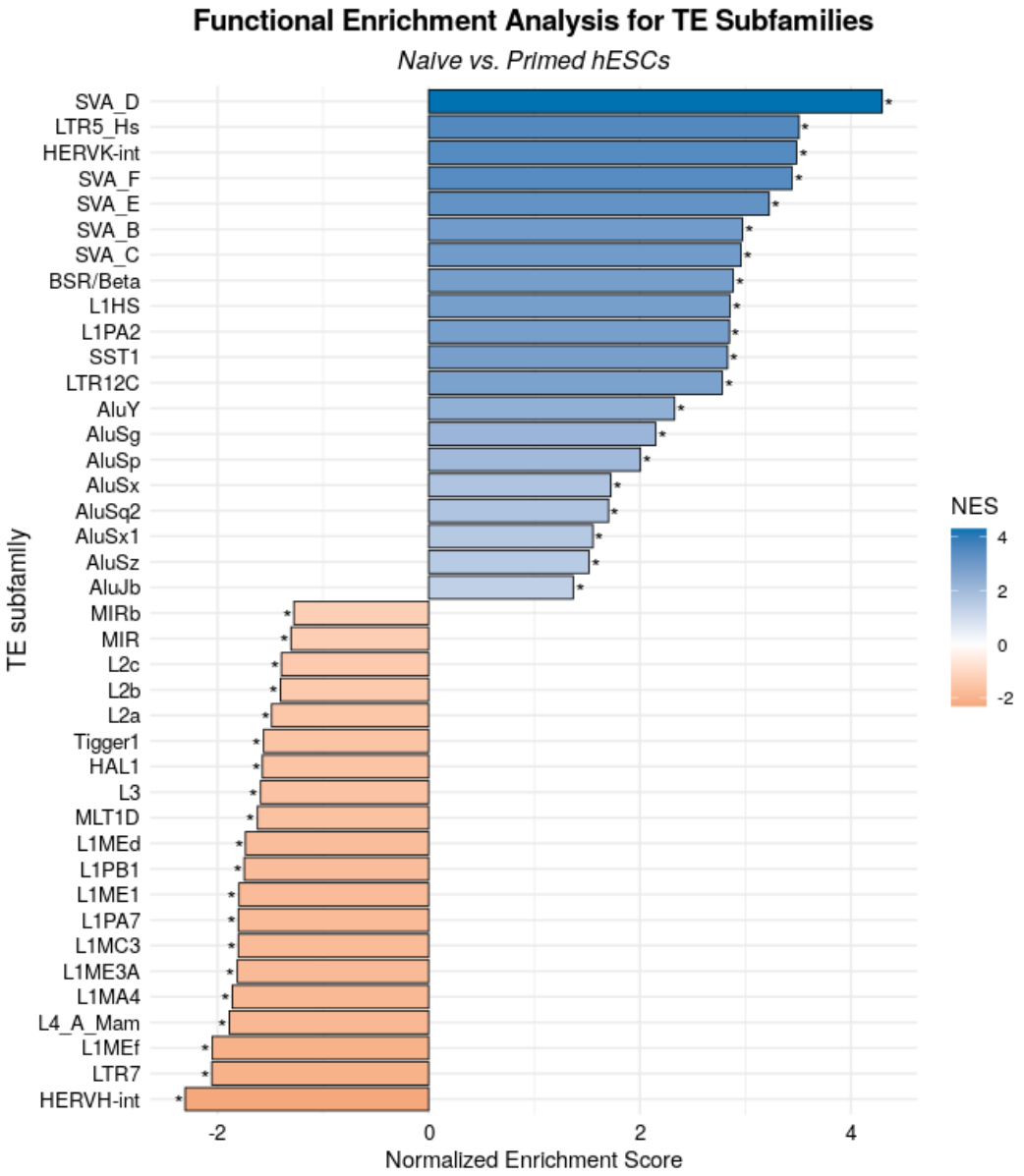


Fig. 4.9 - Functional enrichment analysis allows to identify, in an unbiased manner, TE subfamilies associated with naive and primed hESCs. The bar plot displays the normalised enrichment scores (NES) from functional enrichment analysis, identifying TE subfamilies over- or under-represented in naive and primed hESCs. Positive NES values (blue bars) indicate that TE subfamilies tend to be more highly expressed in naive hESCs, while negative NES values (orange bars) indicate that TE subfamilies tend to be more highly expressed in primed hESCs. The analysis was performed using gene set enrichment analysis (GSEA) with TE-sets, where each set contains all TEs with detected reads, from a particular subfamily. The asterisks (*) mark subfamilies with an adjusted p-value (padj) < 0.01, highlighting statistically significant enrichment with either the naive or primed state. Subfamilies are ordered by NES, with the top 20 enriched subfamilies for each condition shown.

In order to better understand how individual TEs of top differentially-enriched subfamilies behave, we next analysed the enrichment profiles for the TE subfamilies selected above (**Fig. 4.10**). For this we analysed enrichment plots, in which the individual members of each respective subfamily are represented along the x-axis as black lines, in the position they occupy in the ranked list of differential expression. The height of the lines corresponds to the value for the statistics considered (Wald statistics, in this case). As expected, SVA subfamilies, as well as LTR5_Hs and HERVK-int, show a clear accumulation of individual TEs at the top of the ranked list, supporting a major and coordinated activation of these subfamilies in naive hESCs (**Fig. 4.10A**). Interestingly, the same can be observed for the L1Hs and L1PA2 subfamilies, suggesting that these subfamilies might be interesting to explore

further in the context of naive pluripotency. The profiles of *Alu* subfamilies are more difficult to interpret, as their distribution along the ranked list of differential expression does not show a clear bias towards the bottom or the top. This can be explained by the fact that these subfamilies have a very high number of individual elements in the genome, making it harder to visually discriminate a trend. Nevertheless, the NES score, as well as the adjusted p-value suggest that globally they are associated with the naive state, but that a significant proportion of individual elements are upregulated in the primed state. These results were confirmed by plotting the log2-fold change of the expression levels of individual TEs for the top subfamilies enriched in naive hESCs (**Fig. 4.11A**).

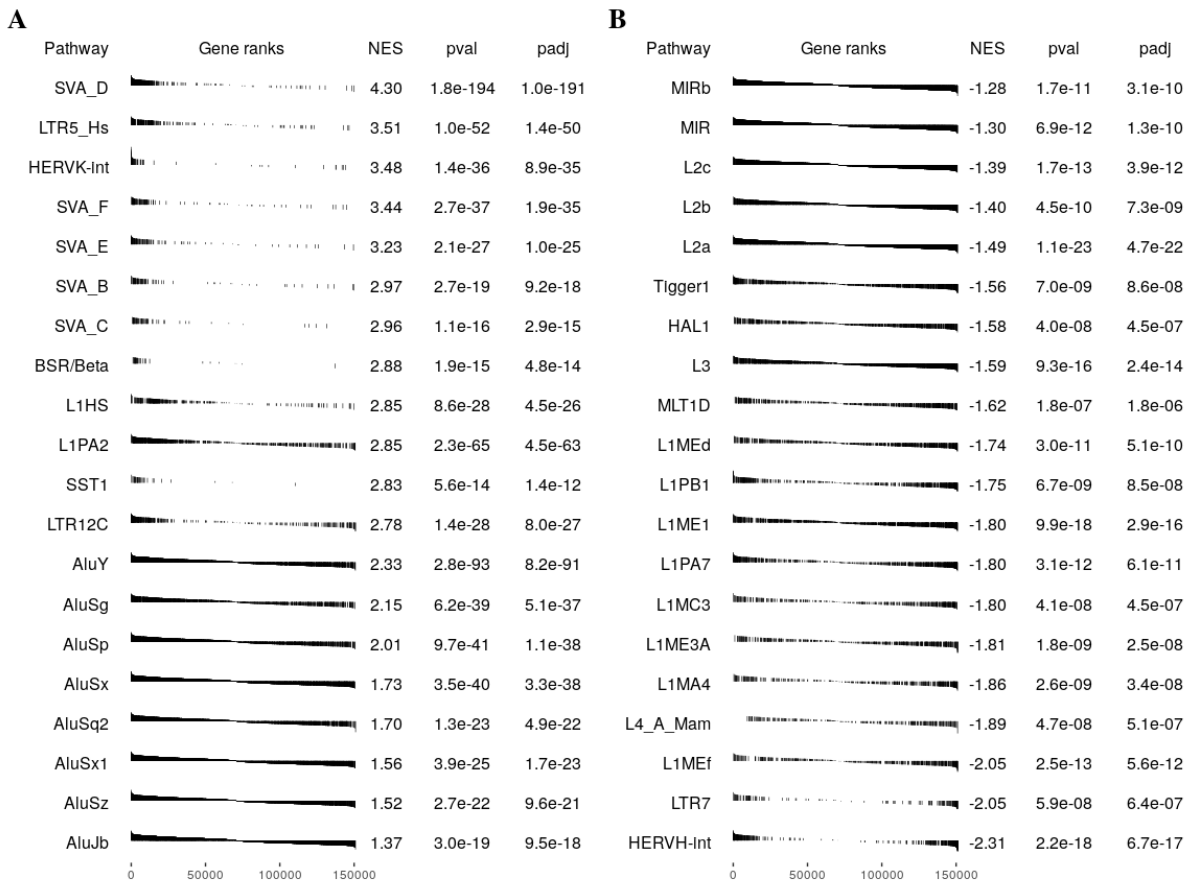


Fig. 4.10 - Enrichment plots for individual TEs from the top-40 differentially expressed subfamilies highlight the behaviour of individual elements in naive or primed contexts. Individual elements from the top-20 naive-associated subfamilies (**A**) or the top-20 primed-associated subfamilies (**B**) are represented along the x-axis as black vertical lines, ordered according to their normalised enrichment score (NES) and with the height of the line corresponding to their Wald statistical value. The p-value (pval) and adjusted p-value (padj) are indicated for each subfamily.

For the primed state, the scenario is not as clear-cut as for naive hESCs, and for many subfamilies the accumulation of TEs at the bottom part of the ranked list (indicating upregulation in primed state) is not clear and unambiguous. For example, for the primed-associated LTR7/HERVH-int subfamilies there is a clear accumulation of individual TEs at the bottom, but there is also a significant enrichment at the top of the ranked list of differential expression (**Fig. 4.10B; Supp. Fig. 7.4**). This indicates that the LTR7/HERVH-int subfamilies are preferentially activated in primed hESCs, and suggests that the regulation of individual TEs depends on the genomic context, with individual elements being upregulated in the primed, but also the naive state. Moreover, the lack of individual elements in the middle of the ranked list might indicate that most individual TEs are expressed and display a tight, context-specific regulation. These results were confirmed by plotting the log2-fold change of the expression levels of individual TEs for the top subfamilies enriched in primed hESCs (**Fig. 4.11B**).

The above observations have to be taken with a certain degree of caution, as functional enrichment analysis can be influenced by biases such as the size of the gene (or TE) sets and their expression levels. Larger sets or highly expressed subfamilies may disproportionately drive NES values, potentially introducing artefacts that confound the biological interpretation of the data (Wijesooriya et al. 2022). For example, the size of the "TE-sets" can influence the NES. This is particularly relevant with TEs, as different subfamilies are either low- or high-copy number and have different mappability, which in turn impact on the ability to measure expression changes. In order to test whether this could be influencing our results, we analysed how the number of elements with mapped reads for each TE subfamily influence the NES or their median expression (calculated using the mean expression of each individual TE of a given subfamily, across all samples) (**Supp. Fig. 7.5**). We could observe that the size of the "TE-set" does not have a major impact on either the NES or median expression of the corresponding subfamily (**Supp. Fig. 7.5A-B**). Additionally, to account for the bias that could be introduced by subfamilies with very high expression, we examined the relationship between NES and the median expression of each TE subfamily. While a few highly expressed subfamilies show elevated NES values, the majority do not exhibit a strong correlation, suggesting that highly expressed TEs do not systematically skew the enrichment analysis (**Supp. Fig. 7.5C**).

Overall, we show that employing a statistical approach of functional enrichment analysis is efficient to identify significantly enriched or depleted groups of TEs, associated with different cellular contexts. This strategy allows to validate subfamilies commonly associated with naive or primed pluripotency, and help identify other potential subfamilies that would have been disregarded with more conventional approaches. Whereas some bias can be introduced using this method, we believe it circumvents many of the limitations associated with differential expression analysis and constitutes a powerful tool to help researchers make informed decisions to test in a biological context.

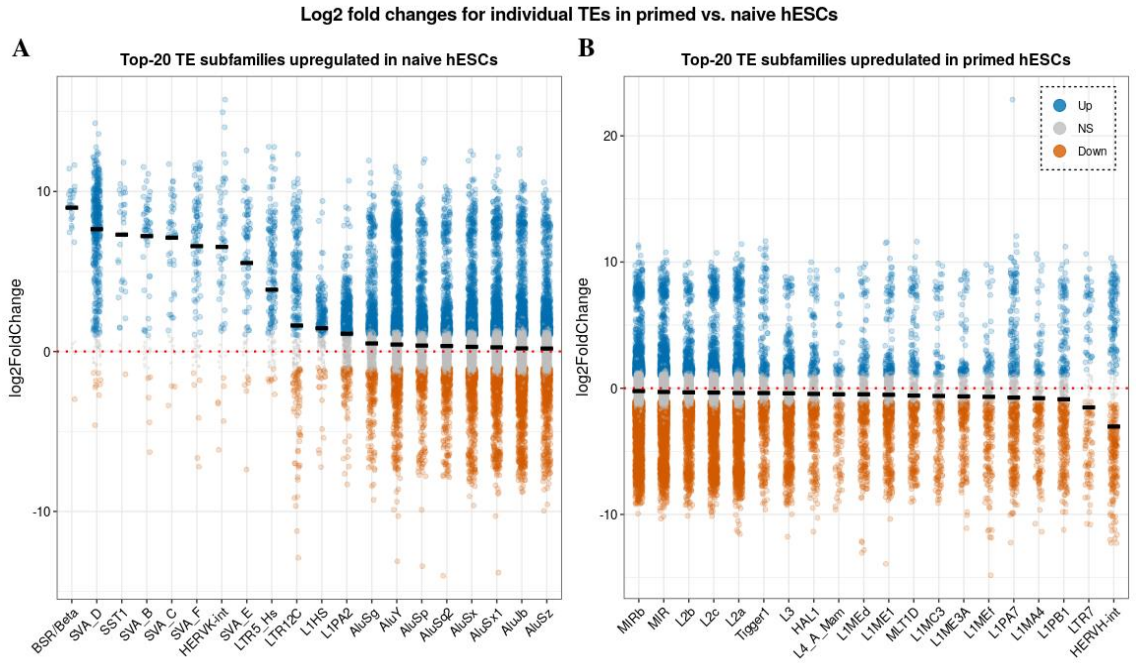


Fig. 4.11 - Plotting the log2 fold change of individual TEs from the top-20 upregulated subfamilies in naive and primed hESCs confirms the trend observed for each state. Dotplots showing the expression dynamics of individual TEs within the top-20 upregulated subfamilies in naive (**A**) and primed (**B**) hESCs. Blue dots indicate log2 fold change>2 (Up) in naive state, corresponding to TE loci upregulated in naive pluripotency. Red dots indicate log2 fold change<2 (Down) in naive state, corresponding to TE loci upregulated in primed pluripotency. Grey dots indicate non-significant (NS) fold change. Black horizontal lines indicate the median of the fold change within each subfamily.

4.4. Testing the association between DEGs and DETEs

TEs carry regulatory motifs that can be exapted by the host to function as regulatory sequences capable of controlling the transcription of nearby genes (Bourque et al. 2018). It has been suggested that the accumulation of TEs of particular subfamilies around developmentally-related genes, places them under an orchestrated network of transcriptional regulation (Sundaram and Wysocka 2020). In order to test this hypothesis, we decided to explore the genomic association between DEGs and DETEs in naive and primed pluripotency. Notably, we tested whether there is an increased density of TEs from selected subfamilies around the promoter regions of DEGs and whether the distance between DEGs and these TEs is lower than what is expected by chance. To narrow down the list of TE subfamilies to test, we decided to focus on the top-enriched subfamilies identified using our functional enrichment analysis strategy (see **Fig. 4.9; Fig. 4.11**).

4.4.1. Using permutations to test the genomic association between DEGs and DETEs

Identifying meaningful associations between different sets of genomic regions is of paramount importance for genomic studies. Due to the complexity of the genome, assigning a statistical significance to this association is not always a trivial task. To tackle this, we used regioneR (Gel et al. 2016), a bioconductor package that offers a fully customizable permutation test framework, and that has been previously used to test the association between DEGs and DETEs (Chelmicki et al. 2021).

To use regioneR in our study, we started by determining the number of up or down DEGs (4855 and 6702 genes, respectively, selected based on log2FC > |1| and padj < 0.01), showing an overlap with elements of any given TE subfamily. As the enhancer potential of individual TEs is thought to be correlated with their proximity to the regulatory regions of genes, we restricted our window size for defining an overlap to -10 kb to +2 kb of the gene start. Next, we created a random sample of genes from all annotated genes, with size equal to the number of considered DEGs. The number of overlaps between this random set of genes and all subfamilies was similarly assessed. The two previous steps were repeated 1000 times (1000 permutations). Finally, we calculated the permutation p-value for the association between up or down DEGs and each TE subfamily, by measuring the number of times that the random permutations had a more extreme value than the observed number of overlaps. Examples for 2 subfamilies are shown in **Supp. Fig. 7.6**.

Using this approach, we tested the association between genes upregulated in naive or primed contexts and the enrichment or depletion of selected TE subfamilies around their regulatory regions (**Fig. 4.12**; **Fig. 4.13**). For genes upregulated in naive pluripotency, we observe a significantly increased overlap with elements belonging to subfamilies active in naive pluripotency, such as LTR5_Hs/HERVK-int and several SVA and *Alu* subfamilies, amongst others (**Fig. 4.12A**). This suggests that the increased density of elements of these subfamilies around the regulatory regions of genes associated with naive pluripotency might be functionally linked to their context-specific transcriptional activation.

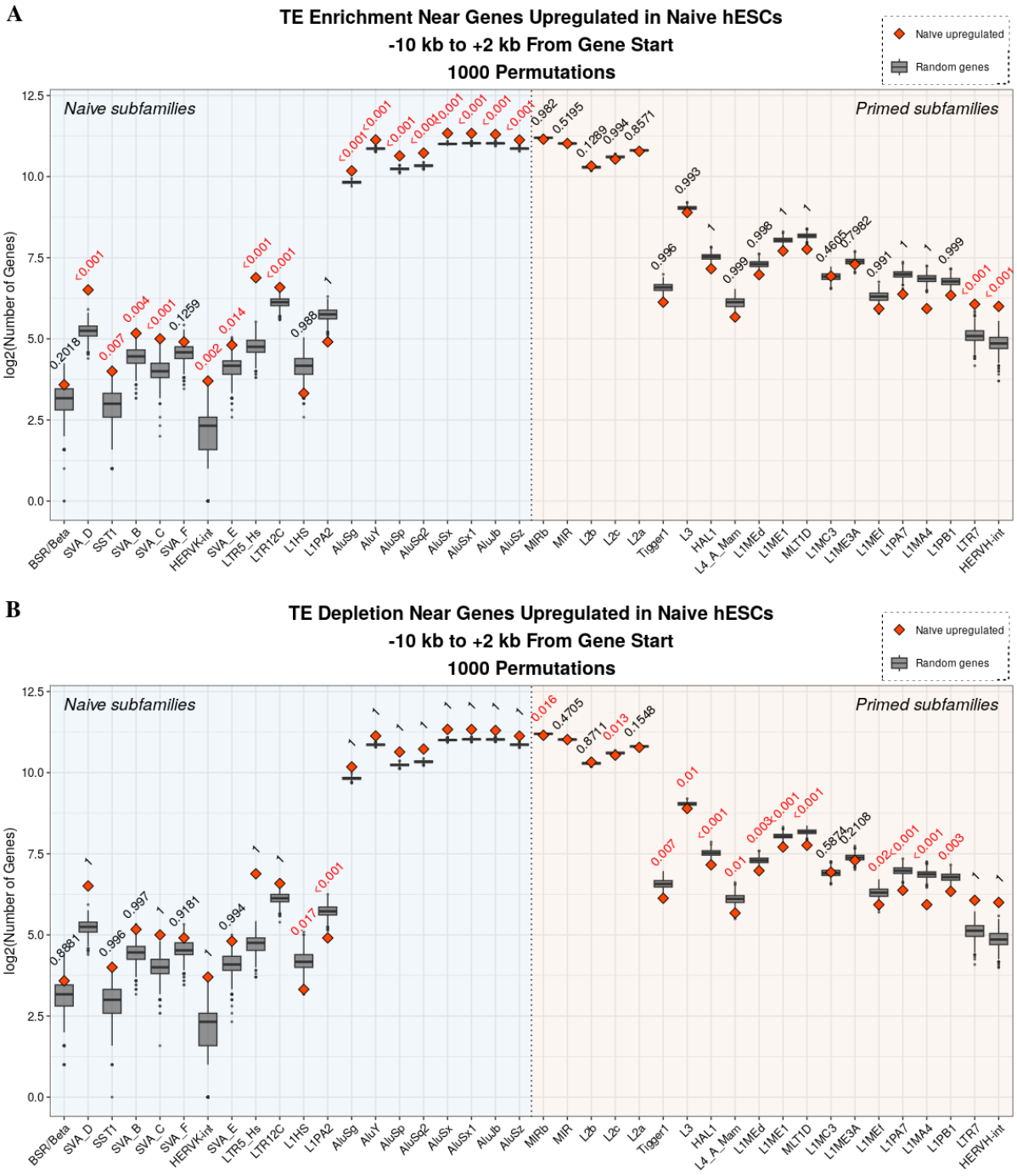


Fig. 4.12 - TE subfamilies associated with naive pluripotency show an increased accumulation around regulatory regions of genes upregulated in naive hESCs. A) Correlation between genes upregulated in naive hESCs and the local enrichment of selected naive-associated (blue half) or primed-associated (orange half) TE subfamilies near their regulatory regions (-10 kb to +2 kb from gene start). The analysis was performed using permutation-based tests, using the `*regioneR*` package. For each TE subfamily, the observed number of overlaps between TE instances and naive upregulated genes (orange diamonds) is compared against a distribution generated from 1000 random permutations of gene sets of equal size (boxplots). P-values represent the probability that the observed overlap is greater than or equal to the random overlap, and are highlighted in red when statistically significant ($p < 0.05$). **B)** Correlation between genes upregulated in naive hESCs and the local depletion of selected naive-associated (blue half) or primed-associated (orange half) TE subfamilies near their regulatory regions (-10 kb to +2 kb from gene start). For each TE subfamily, the observed number of overlaps between TE instances and naive upregulated genes (orange diamonds) is compared against a distribution generated from 1000 random permutations of gene sets of equal size (boxplots). P-values represent the probability that the observed overlap is lower than or equal to the random overlap and are highlighted in red when statistically significant ($p < 0.05$).

Intriguingly, we also observe an increased association between upregulated genes in naive cells and the LTR7/HERVH-int subfamilies, which we found to be globally upregulated in primed hESCs and have been described as primed-associated TE subfamilies. This might be explained by the fact that a significant proportion of individual elements are upregulated in naive hESCs (see **Table 4.3B; Fig. 4.11**), suggesting that a subset of TEs from these subfamilies might be functionally important for the transcriptional regulation of naive-associated genes. Indeed, a recent review on the role of the LTR7/HERVH-int subfamilies discusses how several subtypes seem to be involved with either primed or naive states, suggesting that this subfamily is important for all pluripotent contexts (Sexton et al. 2022).

In contrast, when we test the depletion of elements from our selected subfamilies from the regulatory regions of genes upregulated in naive hESCs, we observe a significant reduction of several LINE (including LINE-1s, LINE-2s and LINE-3s), as well as MIRb subfamilies, all of which show a preferential activation in primed pluripotency (**Fig. 4.12B**). The reduction of these TEs around the regulatory regions of naive-associated genes, might be interpreted as an evolutionary strategy to establish networks of transcriptional coordination that operate in a primed or naive context-specific manner. Interestingly, although being preferentially activated in naive hESCs, L1Hs and L1PA2 are also significantly reduced around the regulatory regions of genes over-expressed in this context. This can be related with the fact that these elements belong to evolutionarily young subfamilies, which comprise the most active elements in the human genome, including some that are competent for retrotransposition. As such, to limit the mutagenic potential of their uncontrolled retrotransposition, the host might employ several mechanisms to control these elements and reduce their accumulation around protein coding genes. Notwithstanding, this does not exclude the possibility that elements from this family play an important role in the genomic regulation of naive hESCs, for example through the 3D organisation of the genome. More studies will be needed to test this hypothesis.

We next investigated the enrichment or depletion of selected TE subfamilies around the regulatory regions of genes upregulated in primed hESCs (**Fig. 4.13**). Our analysis revealed an increased accumulation of the MIR and MIRb subfamilies, along with several LINE-2 and LINE-3 subfamilies, near the promoter regions of these genes (**Fig. 4.13A**). However, unlike the more pronounced enrichment observed for the naive-associated subfamilies, only a limited number of primed-associated subfamilies displayed a significant association with genes upregulated in primed hESCs. This suggests that the influence of TEs on the transcriptional program in primed hESCs may be less extensive than in naive hESCs, where we observed a greater proportion of individual elements upregulated within each naive-associated subfamily (see **Fig. 4.11**). Despite the lower density of primed-associated TEs near the regulatory regions of genes upregulated in primed hESCs, it remains possible that individual TE elements play key roles in regulating the molecular circuitry of primed-associated genes.

In addition, we saw a significantly reduced accumulation of several naive-associated subfamilies, including several SVAs, LTR5_Hs and *Alus* (**Fig. 4.13B**), around the promoter regions of primed-associated genes. Intriguingly, we also found a decreased accumulation of several subfamilies that were identified as preferentially active in primed contexts, including several LINE-1 subfamilies, such as L1PB1, L1MA4, L1PA7, L1MC3, L1ME1 and L1MEd. Similarly to what was discussed above, this might be due to the fact that these subfamilies identified as primed-associated show a high proportion of individual elements upregulated in either primed or naive contexts (see **Fig. 4.11**).

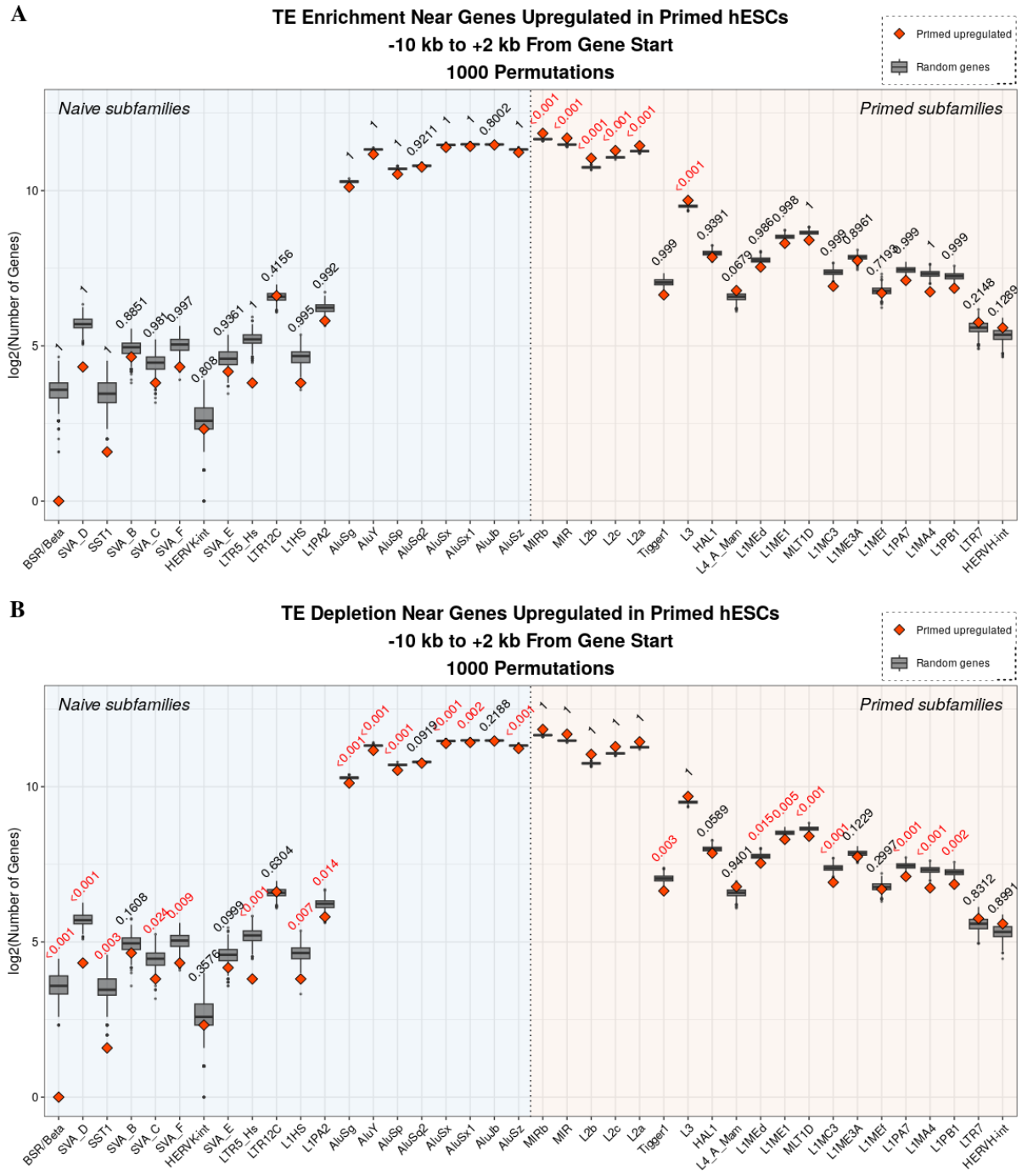


Fig. 4.13 - TE subfamilies associated with primed pluripotency show an increased accumulation around regulatory regions of genes upregulated in primed hESCs. A) Correlation between genes upregulated in primed hESCs and the local enrichment of selected naive-associated (blue half) or primed-associated (orange half) TE subfamilies near their regulatory regions (-10 kb to +2 kb from gene start). The analysis was performed using permutation-based tests with the `*regioneR*` package. For each TE subfamily, the observed number of overlaps between TE instances and primed upregulated genes (orange diamonds) is compared against a distribution generated from 1000 random permutations of gene sets of equal size (boxplots). P-values represent the probability that the observed overlap is greater than or equal to the random overlap and are highlighted in red when statistically significant ($p < 0.05$). **B**) Correlation between genes upregulated in primed hESCs and the local depletion of selected TE subfamilies near their regulatory regions (-10 kb to +2 kb from gene start). For each TE subfamily, the observed number of overlaps between TE instances and primed upregulated genes (orange diamonds) is compared against a distribution generated from 1000 random permutations of gene sets of equal size (boxplots). P-values represent the probability that the observed overlap is lower than or equal to the random overlap and are highlighted in red when statistically significant ($p < 0.05$).

Collectively, our data shows that TEs belonging to transcriptionally active subfamilies identified as being associated with primed or naive pluripotency distribute non-randomly around the regulatory regions of genes that are expressed in a context-specific manner. This supports the idea that certain TEs have been co-opted by the host to create networks of transcriptional regulation, which place context-specific genes under a common set of genetic and epigenetic regulators.

4.4.2. Measuring the distance between DEGs and DETEs

The previous approach focused on testing the potential role of TEs as enhancers located immediately upstream of gene promoters. In reality, the influence of regulatory regions and enhancers on the activity of genes can be felt across larger distances of the genome, either upstream or downstream, and even within the gene body. To have a complementary strategy of testing whether TEs can influence gene expression in primed or naive pluripotency, we decided to measure the distance between all TE elements from the top identified subfamilies to the start of genes upregulated in each of these pluripotent states.

For this, we started by using the *bedr* package (Haider et al. 2016), an R-based interface for the *BEDTools* package (Quinlan and Hall 2010), to estimate the distances between the gene start of naive or primed-upregulated genes and the closest element of the TE subfamilies selected using our functional enrichment analysis strategy (see **section 4.3.4**). These distances were then compared to the distance between an equal sized set of randomly selected genes and significant differences were evaluated using a combination of statistical tests and effect size measures (**Supp. Fig. 7.7**).

Unsurprisingly, and reflecting the data obtained in the previous section, genes overexpressed in naive pluripotency show a significantly reduced distance between their start and elements belonging to subfamilies active in naive pluripotency, such as LTR5_Hs/HERVK-int and several SVA and *Alu* subfamilies, amongst others (**Fig. 4.14; Supp. Table 7.4**). This supports the idea that TEs belonging to subfamilies that show a preferential global activation in a naive context might positively influence the expression of naive-associated genes, by accumulating closer and more frequently around these. Similarly to what was observed using our previous approach, young L1 subfamilies (L1HS and L1PA2) that are overexpressed in naive hESCs and contain elements still capable of transposition, are located further away from the gene start of naive-associated genes, which, might be due to host mechanisms to limit the mutagenic potential of these highly active TE subfamilies.

Furthermore, when we measured the distance between elements of subfamilies active in primed hESCs and genes upregulated in naive pluripotency, we could observe that these are located further away from the gene TSS, compared to what would be expected by chance (**Fig. 4.14**, right half). These results largely resemble those from our permutation tests (**Fig. 4.12**), with the notable difference of the LTR7 and HERVH-int subfamilies, which do not show significant differences in distance compared to a random set of genes.

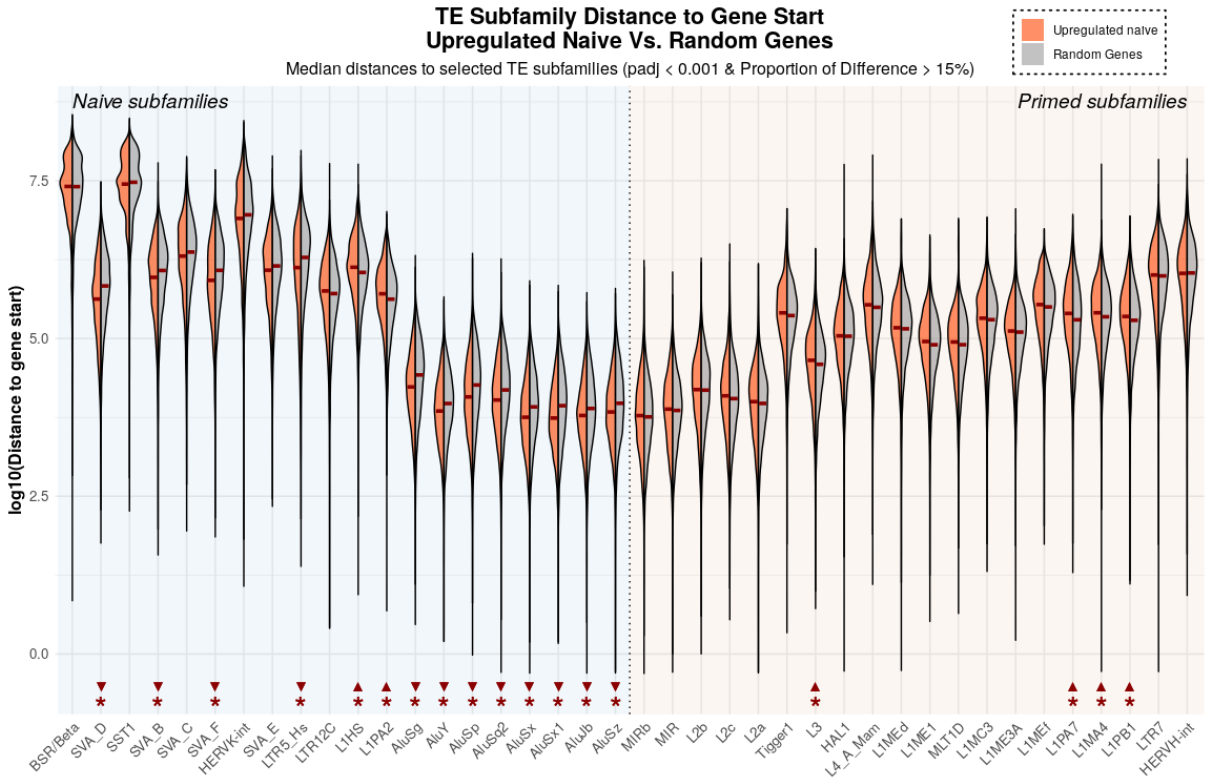


Fig. 4.14 - Naive-associated TE subfamilies show a decreased distance to the start of genes upregulated in naive hESCs. Split violin plots comparing the log10-transformed distance distributions (in bp) between the closest element from naive-associated (blue half) or primed-associated (orange half) TE subfamilies and the start of genes upregulated in naive hESCs (orange) compared to a random gene set of the same size (grey). The red lines represent the median distances, while the width of the violins indicates the density of data points. Red asterisks indicate subfamilies with significant differences between upregulated and random gene distances (FDR-adjusted Wilcoxon p-value < 0.001 and absolute distance proportion difference > 15%). Upward-pointing arrowheads indicate subfamilies where the distance is increased in genes upregulated in naive hESCs compared to random genes, while downward-pointing arrowheads indicate a decreased distance in genes upregulated in naive hESCs.

We next measured the median distances between the start of genes upregulated in primed hESCs and TE subfamilies associated with either naive or primed pluripotency (**Fig. 4.15; Supp. Table 7.5**). Similarly to the results obtained from our permutation tests, we observed a reduced distance between elements of the primed-associated MIR, MIRb and LINE-2 subfamilies, and the start of genes upregulated in a primed context.

In contrast, we observed an increased distance between elements of a few naive-associated subfamilies, including LTR5_Hs, HERVK-int, as well as a few SVA and *Alu* subfamilies, and the start of genes overexpressed in the primed state. Nevertheless, the increase in distances observed for these TE subfamilies is small compared to the distances to a random set of genes, and only the distance to the LTR5_Hs subfamily shows statistical significance. Future work should focus on discriminating the role of different subset of individual elements from a given subfamily, which are active in a context-specific manner, on the activity of primed- or naive-associated genes.

In summary, our analysis shows a trend for a reduced distance between TEs belonging to subfamilies identified as being associated with primed or naive pluripotency and genes expressed in the corresponding state. This strengthens the observations made in the previous section and paves the road for future analysis, focusing on exploring the role of TEs in genomic contexts that expand beyond the upstream regions of genes. This will include, besides different window sizes for upstream regions,

promoter, 5’ and 3’ UTRs, exons, introns and downstream regions and will allow to elucidate whether different subfamilies of TEs have employed different strategies to influence the expression of host genes.

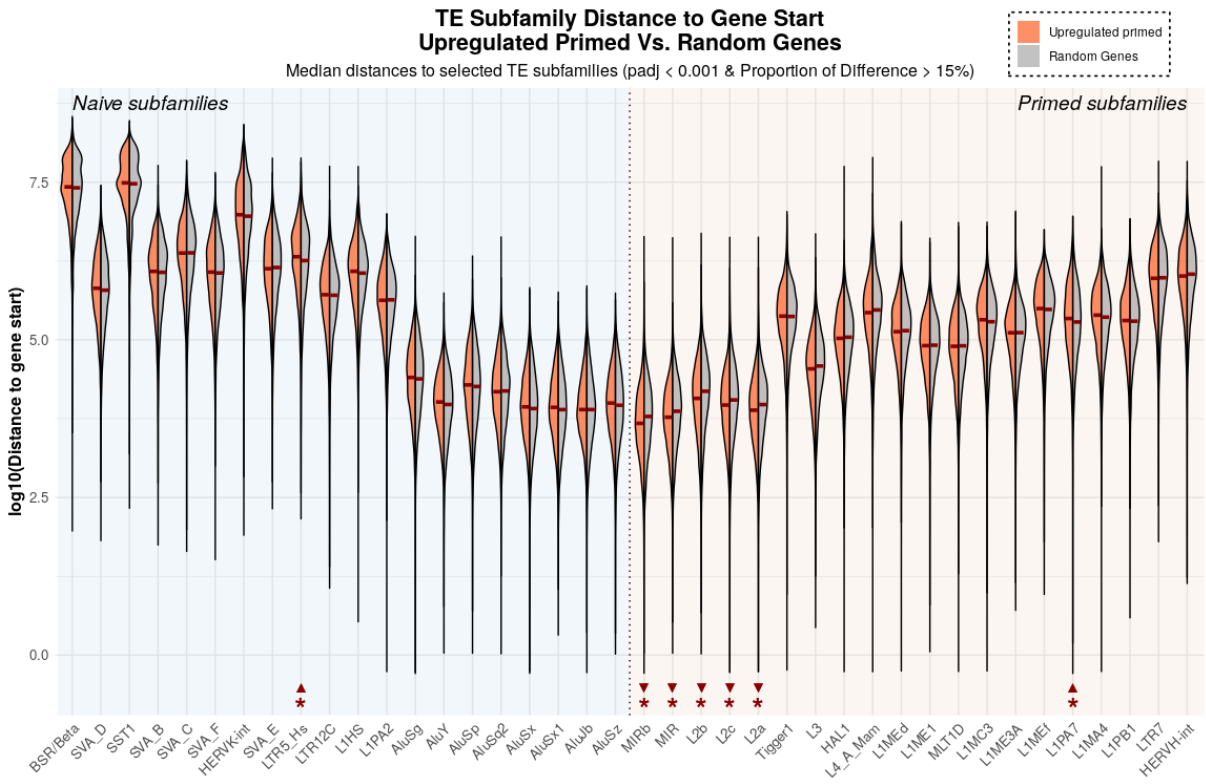


Fig. 4.15 - Primed-associated TE subfamilies show significant changes in distance to the start of genes upregulated in primed hESCs. Split violin plots comparing log10-transformed distance distributions (in bp) between the closest element from naive-associated (blue half) or primed-associated (orange half) TE subfamiliesand the start of genes upregulated in primed hESCs (orange) versus a random gene set of the same size (grey). The red lines represent the median distances, while the width of the violins indicates the density of data points. Red asterisks mark subfamilies with significant differences between primed upregulated genes and random gene distances (FDR-adjusted Wilcoxon p-value < 0.001 and absolute distance proportion difference > 15%). Upward-pointing arrowheads indicate subfamilies where the distance is increased in genes upregulated in primed hESCs compared to random genes, while downward-pointing arrowheads indicate a decreased distance in genes upregulated in primed hESCs.

5. Conclusion

5.1. Developing a pipeline for the parallel differential expression analysis of genes and TEs

In this work, we have developed a computational approach to measure the differential expression of genes and TEs from NGS data. This approach allows to quantify the global expression changes of TE subfamilies, but also allows diving into single TE loci in order to explore their individual expression. By combining these strategies, we further implemented a series of approaches that help in selecting interesting TE subfamilies for further functional analyses.

To test the robustness of our pipeline, we explored the gene and TE expression profiles of naive and primed hESCs. In doing so, we were able to confirm previous observations on the differential activity of TE subfamilies in naive and primed pluripotent contexts. We readily identified subfamilies that have been previously associated with each state: LTR5_Hs/HERVK, SVAs and *Alu* elements with the naive state; LTR7/HERVH with the primed state. In addition, our approach allowed us to identify other subfamilies that might have a biological role in determining the identity of either naive or primed hESCs, namely L1Hs, L1PA2 associated with the naive state, and MIR, LINE-2 and LINE-3 subfamilies, associated with the primed state. These identified subfamilies warrant further investigation into the role of TEs during human preimplantation development.

5.2. Limitations & future improvements of the pipeline

The pipeline presented here has some identified limitations and thus, shows a large margin for improvement and further development. Most of the shortcomings we identified are deeply related to the repetitive nature and genomic distribution of TEs.

First, the pipeline does not allow to distinguish TEs that are located within gene bodies, from those that are intergenic. This means that autonomous transcription of TEs can be easily confounded with TE-gene chimeric transcripts or the expression of the gene into which a TE is inserted. To tackle this, an extra step can be added when creating the custom-made TE GTF files, allowing the removal of all instances that overlap with annotated genes. This would allow focusing the analysis exclusively on intergenic TEs, minimising the number of potentially non-relevant TE instances and thus, false-positives that might obfuscate a biological role.

Second, the pipeline does not deal with scenarios in which multiple hits correspond to a single copy of a TE element. This can happen in distinct situations: the annotation for LTR-retrotransposons are split into the internal portion of the element and its LTR, in order to facilitate the identification of solo-LTRs (which is the case, for example, for the HERVH elements, split into LTR7 and HERVH-int); large deletions within a single element that causes it to be identified as multiple insertions; insertions of TEs within already existing TEs, leading to nested TEs. To deal with this limitation, the use of specialised tools to assemble annotations into complete TE copies (Bailly-Bechet et al. 2014) could be implemented.

Third, our strategy is currently developed to identify TE subfamilies that are globally associated with a particular biological or pathological context. As our study shows, several TE subfamilies show an heterogeneous response of their individual TE elements, with distinct subsets of elements being

activated in either the primed or naive states. Thus, future work should focus on identifying subsets of individual elements within TE subfamilies, and on exploring whether primed- or naive-associated subsets have distinct roles in the regulation of context-specific genes.

Related with the last point, when exploring the genomic association between DEGs and DETEs, besides discriminating the association of TE subsets that are context-specific with naive- or primed-associated genes, future iterations should expand the analysis of the potential regulatory role of individual TEs beyond the 10 kb upstream and 2 kb downstream window of the promoter, which would include enhancers as well as intronic and exonic sequences within the gene.

A large proportion of TE-derived transcripts are not polyadenylated. Because of this, our current pipeline depends on the availability of total RNA-seq datasets, which are not commonly produced in transcriptomics studies. Given that most publicly available datasets are derived from poly(A) RNAs, this limits our ability to apply this pipeline to large data collection projects, like GTEx[11] and TCGA[12], reducing our ability to explore the involvement of TEs in different developmental or pathological contexts. Thus, we envisage to develop a tool that does not require total RNA-seq and is able to test the genomic enrichment of certain TE subfamilies around regulatory regions of DEGs found in transcriptomics studies, without having to directly measure the transcriptional activity of TEs.

For this, profiting from the work developed on the annotation of TEs, we will create lists of genes that contain elements of a given TE subfamily, associated with a given gene feature (e.g. upstream region, promoter, UTRs, exons, introns, downstream region). We will then use these genesets, in a GSEA-like approach, to test whether DEGs show a genomic enrichment or depletion of particular TE subfamilies around gene features of interest. This tool will provide a useful resource that could be readily used by the scientific community to test the potential role of TEs on cell- and context-specific gene expression regulation.

5.3. Using the pipeline for tackling neurological diseases and future perspectives

The pipeline developed here can be readily used to explore the transcriptional dynamics of TEs in different physiological and pathological contexts, paving the way to test the impact of TEs in the regulation of gene expression programs and cellular identity. Indeed, we are currently using our computational approach to explore the TE expression landscape of different neurological disorders, where the misregulation of different groups of TEs has been documented (DeRosa et al. 2022). We are particularly interested in Rett Syndrome (RTT), a progressive neurodevelopmental disorder caused by mutations in the methyl-CpG-binding protein 2 (*MECP2*) gene, which predominantly affects girls and is one of the leading causes of intellectual disability in females. The *MECP2* gene encodes an epigenetic regulatory protein, MeCP2, which binds methylated cytosines in CG and CA contexts and interacts with transcriptional co-repressor complexes, to silence gene expression in a DNA methylation-dependent manner (Marano et al. 2021). Interestingly, MeCP2 has been suggested to have an important role in the regulation of TEs, namely of the L1Hs subfamily in the post-mortem brains of RTT patients (Muotri et al. 2010; Zhao et al. 2019). We will thus investigate, using publicly available total RNA-seq datasets and the computational approaches described in this thesis, the transcriptional landscape of TEs in RTT

[11] https://www.gtexportal.org/home/

[12] https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

neuronal cells, with the aim of elucidating the contribution of the misregulation of TEs for the pathogenesis of this disease. In particular, we will explore whether some of the common transcriptional changes observed in RTT could be driven by the epigenetic and transcriptional misregulation of TE families, subfamilies or individual elements.

6. References

Almeida, M. V., Vernaz, G., Putman, A. L. K., & Miska, E. A. (2022). Taming transposable elements in vertebrates: From epigenetic silencing to domestication. *Trends in Genetics*, *38*(6), 529–553. https://doi.org/10.1016/j.tig.2022.02.009

Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]*. *Available online at: Http://www.bioinformatics.babraham.ac.uk/projects/fastqc/*. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Bailly-Bechet, M., Haudry, A., & Lerat, E. (2014). "One code to find them all": A perl tool to conveniently parse RepeatMasker output files. *Mobile DNA*, *5*(1), 13. https://doi.org/10.1186/1759-8753-5-13

Balestrieri, E., Argaw-Denboba, A., Gambacurta, A., Cipriani, C., Bei, R., Serafino, A., Sinibaldi-Vallebona, P., & Matteucci, C. (2018). Human Endogenous Retrovirus K in the Crosstalk Between Cancer Cells Microenvironment and Plasticity: A New Perspective for Combination Therapy. *Frontiers in Microbiology*, *9*. https://doi.org/10.3389/fmicb.2018.01448

Barbosa-Morais, N. L., Irimia, M., Pan, Q., Xiong, H. Y., Gueroussov, S., Lee, L. J., Slobodeniuc, V., Kutter, C., Watt, S., Çolak, R., Kim, T., Misquitta-Ali, C. M., Wilson, M. D., Kim, P. M., Odom, D. T., Frey, B. J., & Blencowe, B. J. (2012). The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. *Science*, *338*(6114), 1587–1593. https://doi.org/10.1126/science.1230612

Barreiro, L. B., Marioni, J. C., Blekhman, R., Stephens, M., & Gilad, Y. (2010). Functional Comparison of Innate Immune Signaling Pathways in Primates. *PLOS Genetics*, *6*(12), e1001249. https://doi.org/10.1371/journal.pgen.1001249

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

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 Biology*, *19*(1), 199. https://doi.org/10.1186/s13059-018-1577-z

Bourque, G., Leong, B., Vega, V. B., Chen, X., Lee, Y. L., Srinivasan, K. G., Chew, J.-L., Ruan, Y., Wei, C.-L., Ng, H. H., & Liu, E. T. (2008). Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Research*, *18*(11), 1752–1762. https://doi.org/10.1101/gr.080663.108

Bousios, A., Nützmann, H.-W., Buck, D., & Michieletto, D. (2020). Integrating transposable elements in the 3D genome. *Mobile DNA*, *11*(1), 8. https://doi.org/10.1186/s13100-020-0202-3

Brattås, P. L., Jönsson, M. E., Fasching, L., Nelander Wahlestedt, J., Shahsavani, M., Falk, R., Falk, A., Jern, P., Parmar, M., & Jakobsson, J. (2017). TRIM28 Controls a Gene Regulatory Network Based on Endogenous Retroviruses in Human Neural Progenitor Cells. *Cell Reports*, *18*(1), 1–11. https://doi.org/10.1016/j.celrep.2016.12.010

Britten, R. J., & Kohne, D. E. (1968). Repeated Sequences in DNA. *Science*, *161*(3841), 529–540. https://doi.org/10.1126/science.161.3841.529

Brouha, B., Schustak, J., Badge, R. M., Lutz-Prigge, S., Farley, A. H., Moran, J. V., & Kazazian, H. H. (2003). Hot L1s account for the bulk of retrotransposition in the human population. *Proceedings of the National Academy of Sciences*, *100*(9), 5280–5285. https://doi.org/10.1073/pnas.0831042100

Bush, S. J., Chen, L., Tovar-Corona, J. M., & Urrutia, A. O. (2017). Alternative splicing and the

evolution of phenotypic novelty. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1713), 20150474. https://doi.org/10.1098/rstb.2015.0474

Casanova, M., Liyakat Ali, T. M., & Rougeulle, C. (2016). Enlightening the contribution of the dark matter to the X chromosome inactivation process in mammals. *Seminars in Cell & Developmental Biology*, *56*, 48–57. https://doi.org/10.1016/j.semcdb.2016.05.003

Chalopin, D., Naville, M., Plard, F., Galiana, D., & Volff, J.-N. (2015). Comparative Analysis of Transposable Elements Highlights Mobilome Diversity and Evolution in Vertebrates. *Genome Biology and Evolution*, *7*(2), 567–580. https://doi.org/10.1093/gbe/evv005

Chelmicki, T., Roger, E., Teissandier, A., Dura, M., Bonneville, L., Rucli, S., Dossin, F., Fouassier, C., Lameiras, S., & Bourc'his, D. (2021). m6A RNA methylation regulates the fate of endogenous retroviruses. *Nature*, *591*(7849), Article 7849. https://doi.org/10.1038/s41586-020-03135-1

Chuong, E. B., Elde, N. C., & Feschotte, C. (2016). Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science*, *351*(6277), 1083–1087. https://doi.org/10.1126/science.aad5497

Chuong, E. B., Elde, N. C., & Feschotte, C. (2017). Regulatory activities of transposable elements: From conflicts to benefits. *Nature Reviews Genetics*, *18*(2), 71–86. https://doi.org/10.1038/nrg.2016.139

Cordaux, R., & Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics*, *10*(10), 691–703. https://doi.org/10.1038/nrg2640

Coufal, N. G., Garcia-Perez, J. L., Peng, G. E., Yeo, G. W., Mu, Y., Lovci, M. T., Morell, M., O'Shea, K. S., Moran, J. V., & Gage, F. H. (2009). L1 retrotransposition in human neural progenitor cells. *Nature*, *460*(7259), 1127–1131. https://doi.org/10.1038/nature08248

DeRosa, H., Richter, T., Wilkinson, C., & Hunter, R. G. (2022). Bridging the Gap Between Environmental Adversity and Neuropsychiatric Disorders: The Role of Transposable Elements. *Frontiers in Genetics*, *13*. https://www.frontiersin.org/articles/10.3389/fgene.2022.813510

Deschamps-Francoeur, G., Simoneau, J., & Scott, M. S. (2020). Handling multi-mapped reads in RNA-seq. *Computational and Structural Biotechnology Journal*, *18*, 1569–1576. https://doi.org/10.1016/j.csbj.2020.06.014

Diehl, A. G., Ouyang, N., & Boyle, A. P. (2020). Transposable elements contribute to cell and species-specific chromatin looping and gene regulation in mammalian genomes. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-15520-5

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), Article 7398. https://doi.org/10.1038/nature11082

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Ecco, G., Imbeault, M., & Trono, D. (2017). KRAB zinc finger proteins. *Development*, *144*(15), 2719–2729. https://doi.org/10.1242/dev.132605

Eickbush, T. H., & Malik, H. S. (2007). Origins and Evolution of Retrotransposons. In *Mobile DNA II* (pp. 1111–1144). John Wiley & Sons, Ltd. https://doi.org/10.1128/9781555817954.ch49

Ernst, C., Odom, D. T., & Kutter, C. (2017). The emergence of piRNAs against transposon invasion to preserve mammalian genome integrity. *Nature Communications*, *8*(1), Article 1. https://doi.org/10.1038/s41467-017-01049-7

Faulkner, G. J., & Billon, V. (2018). L1 retrotransposition in the soma: A field jumping ahead. *Mobile DNA*, *9*(1), 1–18. https://doi.org/10.1186/s13100-018-0128-1

Faulkner, G. J., & Carninci, P. (2009). Altruistic functions for selfish DNA. *Cell Cycle*, *8*(18), 2895–2900. https://doi.org/10.4161/cc.8.18.9536

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics*, *9*(5), 397–405. https://doi.org/10.1038/nrg2337

Feschotte, C., & Gilbert, C. (2012). Endogenous viruses: Insights into viral evolution and impact on host biology. *Nature Reviews Genetics*, *13*(4), 283–296. https://doi.org/10.1038/nrg3199

Feschotte, C., & Pritham, E. J. (2007). DNA Transposons and the Evolution of Eukaryotic Genomes. *Annual Review of Genetics*, *41*(1), 331–368. https://doi.org/10.1146/annurev.genet.40.110405.090448

Fort, A., Hashimoto, K., Yamada, D., Salimullah, M., Keya, C. A., Saxena, A., Bonetti, A., Voineagu, I., Bertin, N., Kratz, A., Noro, Y., Wong, C.-H., de Hoon, M., Andersson, R., Sandelin, A., Suzuki, H., Wei, C.-L., Koseki, H., Hasegawa, Y., … Carninci, P. (2014). Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nature Genetics*, *46*(6), Article 6. https://doi.org/10.1038/ng.2965

Frank, J. A., & Feschotte, C. (2017). Co-option of endogenous viral sequences for host cell function. *Current Opinion in Virology*, *25*, 81–89. https://doi.org/10.1016/j.coviro.2017.07.021

Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonell Sala, S., Chrast, J., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., … Flicek, P. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, *47*(D1), D766–D773. https://doi.org/10.1093/nar/gky955

Fueyo, R., Judd, J., Feschotte, C., & Wysocka, J. (2022). Roles of transposable elements in the regulation of mammalian transcription. *Nature Reviews Molecular Cell Biology*, *23*(7), Article 7. https://doi.org/10.1038/s41580-022-00457-y

Garcia-Perez, J. L., Widmann, T. J., & Adams, I. R. (2016). The impact of transposable elements on mammalian development. *Development*, *143*(22), 4101–4114. https://doi.org/10.1242/dev.132639

Gel, B., Díez-Villanueva, A., Serra, E., Buschbeck, M., Peinado, M. A., & Malinverni, R. (2016). regioneR: An R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics*, *32*(2), 289–291. https://doi.org/10.1093/bioinformatics/btv562

Goerner-Potvin, P., & Bourque, G. (2018). Computational tools to unmask transposable elements. *Nature Reviews Genetics*, *19*(11), 688–704. https://doi.org/10.1038/s41576-018-0050-x

Göke, J., Lu, X., Chan, Y.-S., Ng, H.-H., Ly, L.-H., Sachs, F., & Szczerbinska, I. (2015). Dynamic Transcription of Distinct Classes of Endogenous Retroviral Elements Marks Specific Populations of Early Human Embryonic Cells. *Cell Stem Cell*, *16*(2), 135–141. https://doi.org/10.1016/j.stem.2015.01.005

Gruchot, J., Kremer, D., & Küry, P. (2019). Neural Cell Responses Upon Exposure to Human Endogenous Retroviruses. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.00655

Guo, G., von Meyenn, F., Rostovskaya, M., Clarke, J., Dietmann, S., Baker, D., Sahakyan, A., Myers, S., Bertone, P., Reik, W., Plath, K., & Smith, A. (2017). Epigenetic resetting of human pluripotency. *Development*, *144*(15), 2748–2763. https://doi.org/10.1242/dev.146811

Haider, S., Waggott, D., Lalonde, E., Fung, C., Liu, F.-F., & Boutros, P. C. (2016). A bedr way of genomic interval processing. *Source Code for Biology and Medicine*, *11*(1), 14. https://doi.org/10.1186/s13029-016-0059-5

Hancks, D. C., & Kazazian, H. H. (2016). Roles for retrotransposon insertions in human disease. *Mobile DNA*, *7*(1), 9. https://doi.org/10.1186/s13100-016-0065-9

Huang, Y., Kim, J. K., Do, D. V., Lee, C., Penfold, C. A., Zylicz, J. J., Marioni, J. C., Hackett, J. A., & Surani, M. A. (2017). Stella modulates transcriptional and endogenous retrovirus programs during maternal-to-zygotic transition. *eLife*, *6*. Scopus. https://doi.org/10.7554/eLife.22345

Imbeault, M., Helleboid, P.-Y., & Trono, D. (2017). KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature*, *543*(7646), Article 7646. https://doi.org/10.1038/nature21683

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860. https://doi.org/10.1038/35057062

Ito, J., Sugimoto, R., Nakaoka, H., Yamada, S., Kimura, T., Hayano, T., & Inoue, I. (2017). Systematic identification and characterization of regulatory elements derived from human endogenous retroviruses. *PLOS Genetics*, *13*(7), e1006883. https://doi.org/10.1371/journal.pgen.1006883

Jachowicz, J. W., Bing, X., Pontabry, J., Bošković, A., Rando, O. J., & Torres-Padilla, M.-E. (2017). LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. *Nature Genetics*, *49*(10), 1502–1510. https://doi.org/10.1038/ng.3945

Jerkovic´, I., & Cavalli, G. (2021). Understanding 3D genome organization by multidisciplinary methods. *Nature Reviews Molecular Cell Biology*, *22*(8), Article 8. https://doi.org/10.1038/s41580-021-00362-w

Kapusta, A., Kronenberg, Z., Lynch, V. J., Zhuo, X., Ramsay, L., Bourque, G., Yandell, M., & Feschotte, C. (2013). Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. *PLOS Genetics*, *9*(4), e1003470. https://doi.org/10.1371/journal.pgen.1003470

Keinath, M. C., Timoshevskiy, V. A., Timoshevskaya, N. Y., Tsonis, P. A., Voss, S. R., & Smith, J. J. (2015). Initial characterization of the large genome of the salamander Ambystoma mexicanum using shotgun and laser capture chromosome sequencing. *Scientific Reports*, *5*(1), Article 1. https://doi.org/10.1038/srep16413

Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M. N., & Sergushichev, A. (2021). *Fast gene set enrichment analysis* (p. 060012). bioRxiv. https://doi.org/10.1101/060012

Kramerov, D. A., & Vassetzky, N. S. (2011). Origin and evolution of SINEs in eukaryotic genomes. *Heredity*, *107*(6), Article 6. https://doi.org/10.1038/hdy.2011.43

Kriegs, J. O., Churakov, G., Jurka, J., Brosius, J., & Schmitz, J. (2007). Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. *Trends in Genetics*, *23*(4), 158–161. https://doi.org/10.1016/j.tig.2007.02.002

Krueger, F. (2012, March 14). *Babraham Bioinformatics—Trim Galore!* https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

Kruse, K., Díaz, N., Enriquez-Gasca, R., Gaume, X., Torres-Padilla, M.-E., & Vaquerizas, J. M. (2019). Transposable elements drive reorganisation of 3D chromatin during early embryogenesis. *bioRxiv*, 523712. https://doi.org/10.1101/523712

Küry, P., Nath, A., Créange, A., Dolei, A., Marche, P., Gold, J., Giovannoni, G., Hartung, H.-P., & Perron, H. (2018). Human Endogenous Retroviruses in Neurological Diseases. *Trends in Molecular Medicine*, *24*(4), 379–394. https://doi.org/10.1016/j.molmed.2018.02.007

Lanciano, S., & Cristofari, G. (2020). Measuring and interpreting transposable element expression. *Nature Reviews Genetics*. https://doi.org/10.1038/s41576-020-0251-y

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., & Carey, V. J. (2013). Software for Computing and Annotating Genomic Ranges. *PLOS Computational Biology*, *9*(8), e1003118. https://doi.org/10.1371/journal.pcbi.1003118

Le Breton, A., Bettencourt, M. P., & Gendrel, A.-V. (2024). Navigating the brain and aging: Exploring the impact of transposable elements from health to disease. *Frontiers in Cell and Developmental Biology*, *12*. https://doi.org/10.3389/fcell.2024.1357576

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

Lin, M., Lucas, H. C., & Shmueli, G. (2013). Research Commentary: Too Big to Fail: Large Samples and the p-Value Problem. *Information Systems Research*, *24*(4), 906–917.

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

Lu, J. Y., Shao, W., Chang, L., Yin, Y., Li, T., Zhang, H., Hong, Y., Percharde, M., Guo, L., Wu, Z., Liu, L., Liu, W., Yan, P., Ramalho-Santos, M., Sun, Y., & Shen, X. (2020). Genomic Repeats Categorize Genes with Distinct Functions for Orchestrated Regulation. *Cell Reports*, *30*(10), 3296-3311.e5. https://doi.org/10.1016/j.celrep.2020.02.048

Lu, X., Sachs, F., Ramsay, L., Jacques, P.-É., Göke, J., Bourque, G., & Ng, H.-H. (2014). The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nature Structural & Molecular Biology*, *21*(4), 423–425. https://doi.org/10.1038/nsmb.2799

Macfarlan, T. S., Gifford, W. D., Driscoll, S., Lettieri, K., Rowe, H. M., Bonanomi, D., Firth, A., Singer, O., Trono, D., & Pfaff, S. L. (2012). Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature*, *487*(7405), Article 7405. https://doi.org/10.1038/nature11244

Marano, D., Fioriniello, S., D'Esposito, M., & Della Ragione, F. (2021). Transcriptomic and Epigenomic Landscape in Rett Syndrome. *Biomolecules*, *11*(7), Article 7. https://doi.org/10.3390/biom11070967

McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences*, *36*(6), 344–355. https://doi.org/10.1073/pnas.36.6.344

McClintock, B. (1951). Chromosome Organization and Genic Expression. *Cold Spring Harbor Symposia on Quantitative Biology*, *16*, 13–47. https://doi.org/10.1101/SQB.1951.016.01.004

Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., Young, T. R., Goldmann, J. M., Pervouchine, D. D., Sullivan, T. J., Johnson, R., Segrè, A. V., Djebali, S., Niarchou, A., Consortium, T. Gte., Wright, F. A., Lappalainen, T., Calvo, M., Getz, G., … Guigó, R. (2015). The human transcriptome across tissues and individuals. *Science*, *348*(6235), 660–665. https://doi.org/10.1126/science.aaa0355

Messmer, T., von Meyenn, F., Savino, A., Santos, F., Mohammed, H., Lun, A. T. L., Marioni, J. C., & Reik, W. (2019). Transcriptional Heterogeneity in Naive and Primed Human Pluripotent Stem Cells at Single-Cell Resolution. *Cell Reports*, *26*(4), 815-824.e4. https://doi.org/10.1016/j.celrep.2018.12.099

Meyer, A., Schloissnig, S., Franchini, P., Du, K., Woltering, J. M., Irisarri, I., Wong, W. Y., Nowoshilow, S., Kneitz, S., Kawaguchi, A., Fabrizius, A., Xiong, P., Dechaud, C., Spaink, H. P., Volff, J.-N., Simakov, O., Burmester, T., Tanaka, E. M., & Schartl, M. (2021). Giant lungfish genome elucidates the conquest of land by vertebrates. *Nature*, *590*(7845), Article 7845. https://doi.org/10.1038/s41586-021-03198-8

Mita, P., & Boeke, J. D. (2016). How retrotransposons shape genome regulation. *Current Opinion in Genetics & Development*, *37*, 90–100. https://doi.org/10.1016/j.gde.2016.01.001

Mita, P., Wudzinska, A., Sun, X., Andrade, J., Nayak, S., Kahler, D. J., Badri, S., LaCava, J., Ueberheide, B., Yun, C. Y., Fenyö, D., & Boeke, J. D. (2018). LINE-1 protein localization and functional dynamics during the cell cycle. *eLife*, *7*, e30058. https://doi.org/10.7554/eLife.30058

Molaro, A., & Malik, H. S. (2016). Hide and seek: How chromatin-based pathways silence retroelements in the mammalian germline. *Current Opinion in Genetics & Development*, *37*, 51–58. https://doi.org/10.1016/j.gde.2015.12.001

Muotri, A. R., Marchetto, M. C. N., Coufal, N. G., Oefner, R., Yeo, G., Nakashima, K., & Gage, F. H. (2010). L1 retrotransposition in neurons is modulated by MeCP2. *Nature*, *468*(7322), 443–446. https://doi.org/10.1038/nature09544

Naville, M., Warren, I. A., Haftek-Terreau, Z., Chalopin, D., Brunet, F., Levin, P., Galiana, D., & Volff,

J.-N. (2016). Not so bad after all: Retroviruses and long terminal repeat retrotransposons as a source of new genes in vertebrates. *Clinical Microbiology and Infection*, *22*(4), 312–323. https://doi.org/10.1016/j.cmi.2016.02.001

Nichols, J., & Smith, A. (2009). Naive and Primed Pluripotent States. *Cell Stem Cell*, *4*(6), 487–492. https://doi.org/10.1016/j.stem.2009.05.015

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N. L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Blüthgen, N., Dekker, J., & Heard, E. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, *485*(7398), Article 7398. https://doi.org/10.1038/nature11049

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. (2021). *The complete sequence of a human genome* (p. 2021.05.26.445798). https://doi.org/10.1101/2021.05.26.445798

Patwardhan, M. N., Wenger, C. D., Davis, E. S., & Phanstiel, D. H. (2019). Bedtoolsr: An R package for genomic data analysis and manipulation. *Journal of Open Source Software*, *4*(44), 1742. https://doi.org/10.21105/joss.01742

Payer, L. M., & Burns, K. H. (2019). Transposable elements in human genetic disease. *Nature Reviews Genetics*, *20*(12), 760–772. https://doi.org/10.1038/s41576-019-0165-8

Percharde, M., Lin, C.-J., Yin, Y., Guan, J., Peixoto, G. A., Bulut-Karslioglu, A., Biechele, S., Huang, B., Shen, X., & Ramalho-Santos, M. (2018). A LINE1-Nucleolin Partnership Regulates Early Development and ESC Identity. *Cell*, *174*(2), 391-405.e19. https://doi.org/10.1016/j.cell.2018.05.043

Platanias, L. C. (2005). Mechanisms of type-I- and type-II-interferon-mediated signalling. *Nature Reviews Immunology*, *5*(5), Article 5. https://doi.org/10.1038/nri1604

Pontis, J., Planet, E., Offner, S., Turelli, P., Duc, J., Coudray, A., Theunissen, T. W., Jaenisch, R., & Trono, D. (2019). Hominoid-Specific Transposable Elements and KZFPs Facilitate Human Embryonic Genome Activation and Control Transcription in Naive Human ESCs. *Cell Stem Cell*, *24*(5), 724-735.e5. https://doi.org/10.1016/j.stem.2019.03.012

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. https://www.R-project.org/

Richardson, S. R., Doucet, A. J., Kopera, H. C., Moldovan, J. B., Garcia-Perez, J. L., & Moran, J. V. (2015). The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. In *Mobile DNA III* (pp. 1165–1208). John Wiley & Sons, Ltd. https://doi.org/10.1128/9781555819217.ch51

Rodriguez-Terrones, D., & Torres-Padilla, M.-E. (2018). Nimble and Ready to Mingle: Transposon Outbursts of Early Development. *Trends in Genetics*, *34*(10), 806–820. https://doi.org/10.1016/j.tig.2018.06.006

Rookhuizen, D. C., Bonte, P.-E., Ye, M., Hoyler, T., Gentili, M., Burgdorf, N., Durand, S., Aprahamian, F., Kroemer, G., Manel, N., Waterfall, J. J., Milne, R., Goudot, C., Towers, G. J., & Amigorena, S. (2021). *Induction of transposable element expression is central to innate sensing* (p. 2021.09.10.457789). bioRxiv. https://doi.org/10.1101/2021.09.10.457789

RStudio Team. (2020). *RStudio: Integrated Development for R*. Http://Www.Rstudio.Com/. http://www.rstudio.com/

Saleh, A., Macia, A., & Muotri, A. R. (2019). Transposable Elements, Inflammation, and Neurological Disease. *Frontiers in Neurology*, *10*. https://doi.org/10.3389/fneur.2019.00894

Seisenberger, S., Peat, J. R., Hore, T. A., Santos, F., Dean, W., & Reik, W. (2013). Reprogramming DNA methylation in the mammalian life cycle: Building and breaking epigenetic barriers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1609), 20110330. https://doi.org/10.1098/rstb.2011.0330

Senft, A. D., & Macfarlan, T. S. (2021). Transposable elements shape the evolution of mammalian development. *Nature Reviews Genetics*, *22*(11), Article 11. https://doi.org/10.1038/s41576-021-00385-1

Sexton, C. E., Tillett, R. L., & Han, M. V. (2022). The essential but enigmatic regulatory role of HERVH in pluripotency. *Trends in Genetics*, *38*(1), 12–21. https://doi.org/10.1016/j.tig.2021.07.007

Smit, AFA., Hubley, R, & Green, P. (2013). *RepeatMasker Open-4.0.* http://www.repeatmasker.org

Straalen, N. M. van, Roelofs, D., Straalen, N. M. van, & Roelofs, D. (2011). *An Introduction to Ecological Genomics* (Second Edition, Second Edition). Oxford University Press.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, *102*(43), 15545–15550. https://doi.org/10.1073/pnas.0506580102

Sundaram, V., Cheng, Y., Ma, Z., Li, D., Xing, X., Edge, P., Snyder, M. P., & Wang, T. (2014). Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Research*, gr.168872.113. https://doi.org/10.1101/gr.168872.113

Sundaram, V., & Wysocka, J. (2020). Transposable elements as a potent source of diverse *cis* -regulatory sequences in mammalian genomes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1795), 20190347. https://doi.org/10.1098/rstb.2019.0347

Teissandier, A., Servant, N., Barillot, E., & Bourc'his, D. (2019). Tools and best practices for retrotransposon analysis using high-throughput sequencing data. *Mobile DNA*, *10*(1), 52. https://doi.org/10.1186/s13100-019-0192-1

Theunissen, T. W., Friedli, M., He, Y., Planet, E., O'Neil, R. C., Markoulaki, S., Pontis, J., Wang, H., Iouranova, A., Imbeault, M., Duc, J., Cohen, M. A., Wert, K. J., Castanon, R., Zhang, Z., Huang, Y., Nery, J. R., Drotar, J., Lungjangwa, T., … Jaenisch, R. (2016). Molecular Criteria for Defining the Naive Human Pluripotent State. *Cell Stem Cell*, *19*(4), 502–515. https://doi.org/10.1016/j.stem.2016.06.011

Thomas, C. A., Tejwani, L., Trujillo, C. A., Negraes, P. D., Herai, R. H., Mesci, P., Macia, A., Crow, Y. J., & Muotri, A. R. (2017). Modeling of TREX1-Dependent Autoimmune Disease using Human Stem Cells Highlights L1 Accumulation as a Source of Neuroinflammation. *Cell Stem Cell*, *21*(3), 319-331.e8. https://doi.org/10.1016/j.stem.2017.07.009

Upton, K. R., Gerhardt, D. J., Jesuadian, J. S., Richardson, S. R., Sánchez-Luque, F. J., Bodea, G. O., Ewing, A. D., Salvador-Palomeque, C., van der Knaap, M. S., Brennan, P. M., Vanderver, A., & Faulkner, G. J. (2015). Ubiquitous L1 Mosaicism in Hippocampal Neurons. *Cell*, *161*(2), 228–239. https://doi.org/10.1016/j.cell.2015.03.026

Vallot, C., Huret, C., Lesecque, Y., Resch, A., Oudrhiri, N., Bennaceur-Griscelli, A., Duret, L., & Rougeulle, C. (2013). XACT, a long noncoding transcript coating the active X chromosome in human pluripotent cells. *Nature Genetics*, *45*(3), 239–241. https://doi.org/10.1038/ng.2530

Vallot, C., Patrat, C., Collier, A. J., Huret, C., Casanova, M., Liyakat Ali, T. M., Tosolini, M., Frydman, N., Heard, E., Rugg-Gunn, P. J., & Rougeulle, C. (2017). XACT Noncoding RNA Competes with XIST in the Control of X Chromosome Activity during Human Early Development. *Cell Stem Cell*, *20*(1), 102–111. https://doi.org/10.1016/j.stem.2016.10.014

Walter, M., Teissandier, A., Pérez-Palacios, R., & Bourc'his, D. (2016). An epigenetic switch ensures transposon repression upon dynamic loss of DNA methylation in embryonic stem cells. *eLife*,

5, e11418. https://doi.org/10.7554/eLife.11418

Wang, H., Xing, J., Grover, D., Hedges, D. J., Han, K., Walker, J. A., & Batzer, M. A. (2005). SVA Elements: A Hominid-specific Retroposon Family. *Journal of Molecular Biology*, *354*(4), 994–1007. https://doi.org/10.1016/j.jmb.2005.09.085

Wang, J., Xie, G., Singh, M., Ghanbarian, A. T., Raskó, T., Szvetnik, A., Cai, H., Besser, D., Prigione, A., Fuchs, N. V., Schumann, G. G., Chen, W., Lorincz, M. C., Ivics, Z., Hurst, L. D., & Izsvák, Z. (2014). Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature*, *516*(7531), 405–409. https://doi.org/10.1038/nature13804

Wells, J. N., & Feschotte, C. (2020). A Field Guide to Eukaryotic Transposable Elements. *Annual Review of Genetics*, *54*(1), 539–561. https://doi.org/10.1146/annurev-genet-040620-022145

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., & Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, *8*(12), Article 12. https://doi.org/10.1038/nrg2165

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

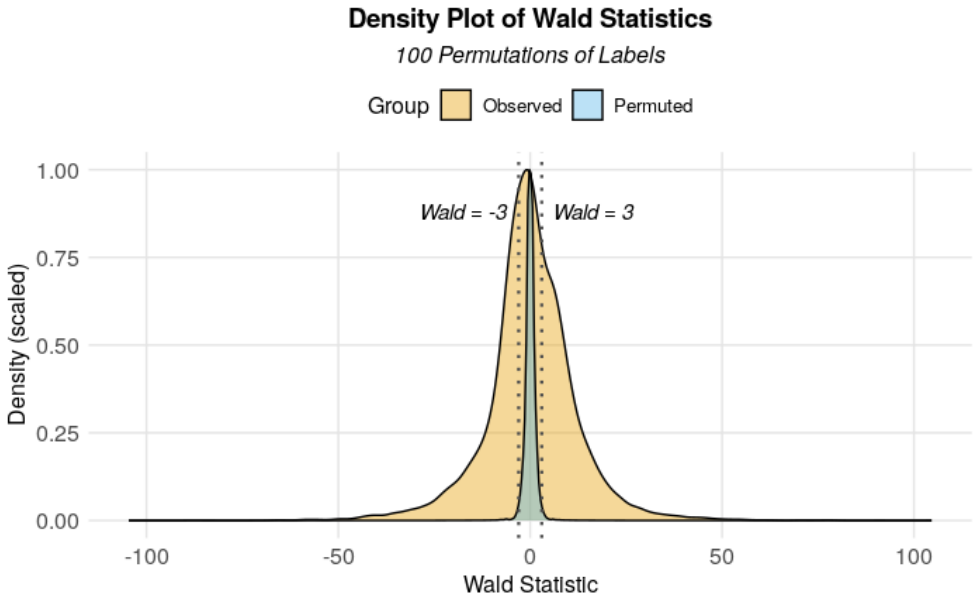
Wijesooriya, K., Jadaan, S. A., Perera, K. L., Kaur, T., & Ziemann, M. (2022). Urgent need for consistent standards in functional enrichment analysis. *PLOS Computational Biology*, *18*(3), e1009935. https://doi.org/10.1371/journal.pcbi.1009935

Yandım, C., & Karakülah, G. (2019). Expression dynamics of repetitive DNA in early human embryonic development. *BMC Genomics*, *20*(1), 439. https://doi.org/10.1186/s12864-019-5803-1

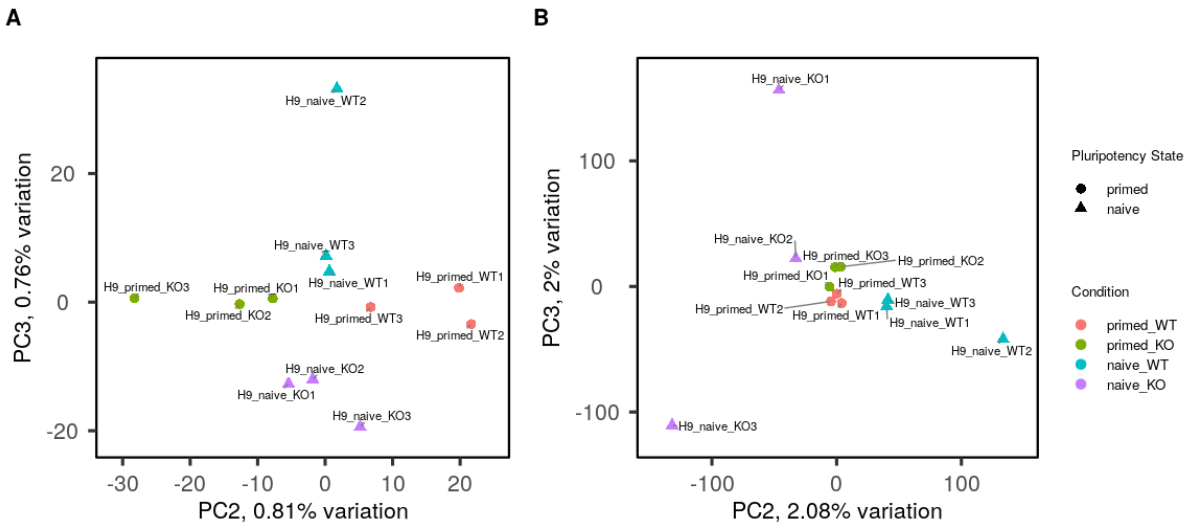
Yang, P., Wang, D., & Kang, L. (2021). Alternative splicing level related to intron size and organism complexity. *BMC Genomics*, *22*(1), 853. https://doi.org/10.1186/s12864-021-08172-2

Zhao, B., Wu, Q., Ye, A. Y., Guo, J., Zheng, X., Yang, X., Yan, L., Liu, Q.-R., Hyde, T. M., Wei, L., & Huang, A. Y. (2019). Somatic LINE-1 retrotransposition in cortical neurons and non-brain tissues of Rett patients and healthy individuals. *PLOS Genetics*, *15*(4), e1008043. https://doi.org/10.1371/journal.pgen.1008043

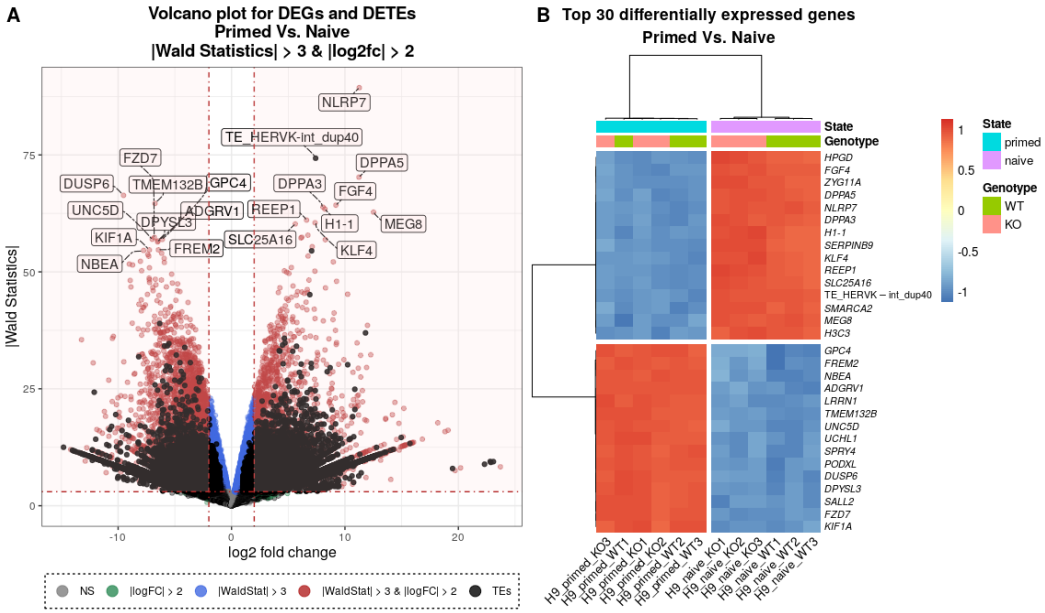
7. Annex



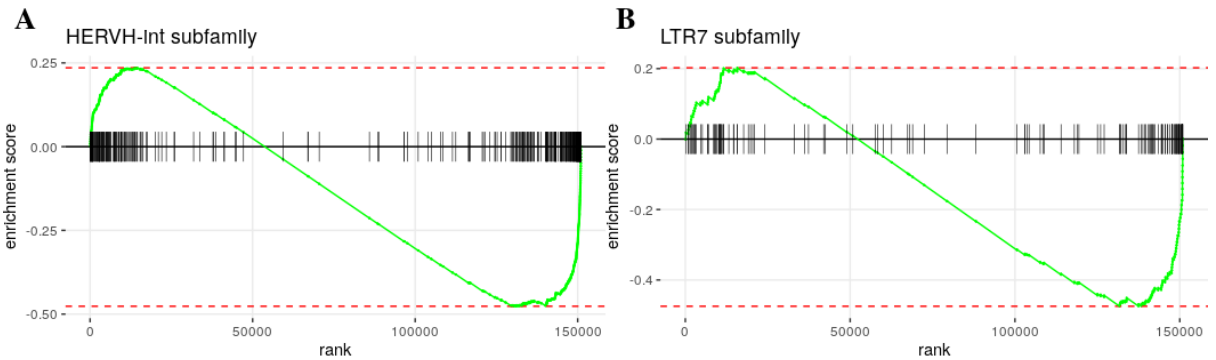
Supp. Fig. 7.1 - Determining Wald statistics for differential expression analysis. Density plot comparing the distribution of Wald statistics from observed data (orange) and permuted data (blue). The observed Wald statistics were calculated using the correct sample labels (primed vs. naive), while the permuted Wald statistics were calculated from 100 random permutations of the sample labels. Each distribution shows the scaled density of the Wald statistics. The observed distribution (orange) reflects the true state of the data, while the permuted distribution (blue) represents the null hypothesis, generated by randomly shuffling the sample labels, for comparison.The Wald statistics observed by chance allowed the selection of a statistical threshold for identifying differentially expressed genes (DEGs) and transposable elements (DETEs). The vertical dashed lines at Wald = -3 and Wald = 3 mark the threshold chosen to define significant differentially expressed genes (DEGs).



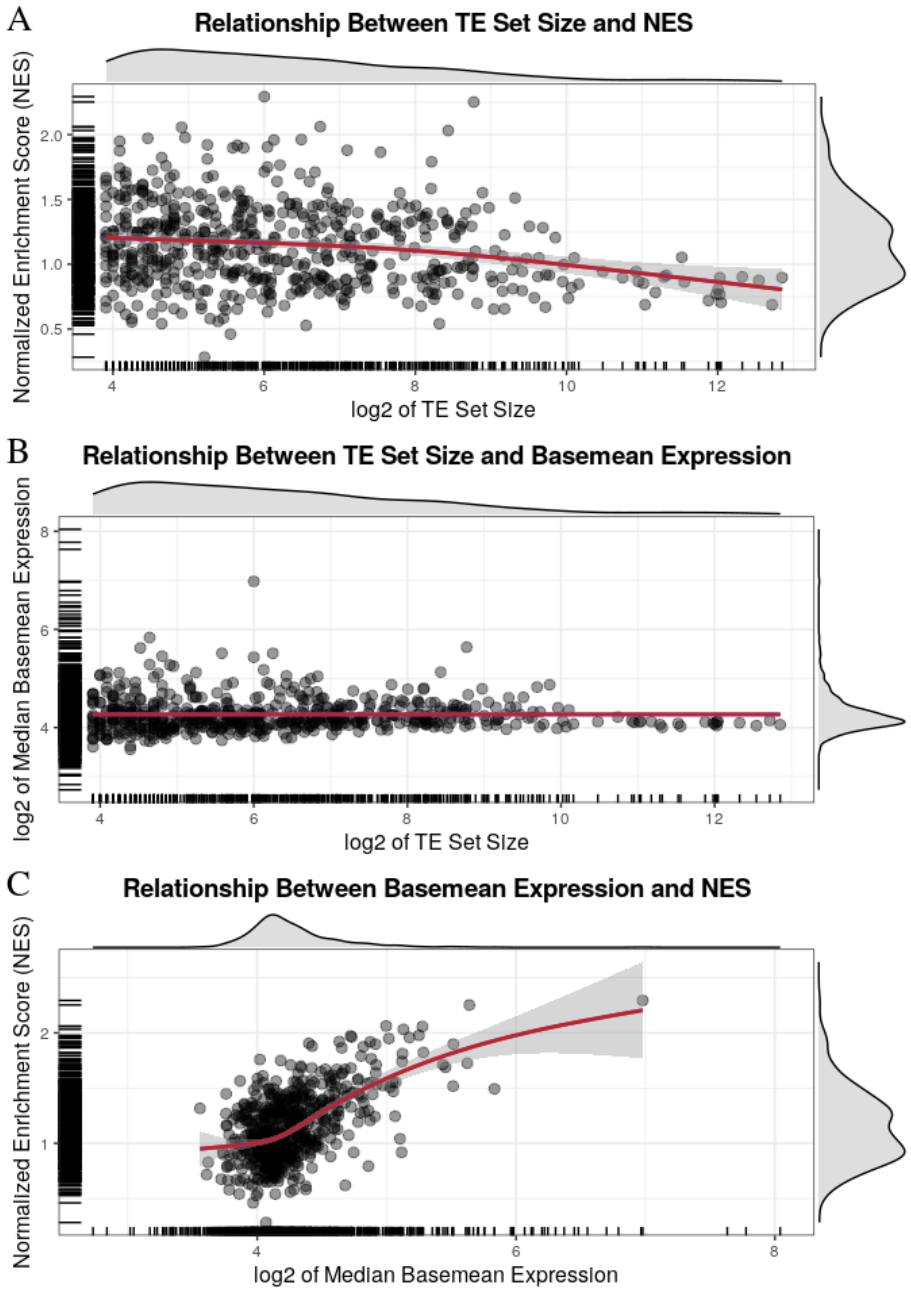
Supp. Fig. 7.2 - Distinct expression profiles of naive and primed hESCs. PCA plots of the naive and primed hESCs datasets along the second (PC2) and the third principal components (PC3), for random mapping (**A**) and unique mapping (**B**) approaches. Samples are colored by genotype and state (primed WT, primed KO, naive WT and naive KO) and shaped by pluripotency state (naive vs primed). The percentage of explained variance by each PC is indicated on the respective axes.



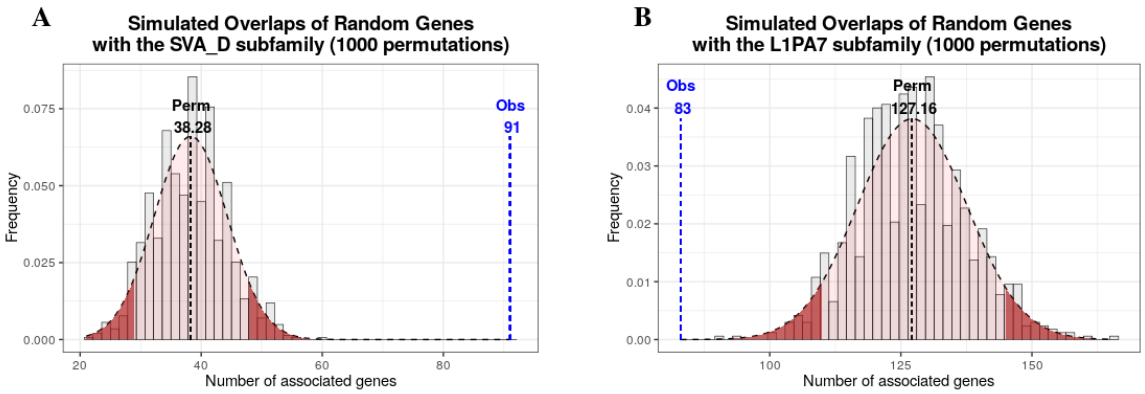
Supp. Fig. 7.3 - Differential expression analysis using the unique mapping approach gives comparable results to random mapping strategy. A) Volcano plot displaying expression fold changes (x-axis) and their significance (y-axis) for differentially expressed genes (DEGs) and individual TEs (DETEs) between naive and primed hESCs, using the unique mapping approach. The horizontal red dashed line marks the log2 fold change threshold of 2, while vertical red dashed lines represent Wald statistics greater than 3. DEGs meeting the criteria of Wald > 3 and |logFC| > 2 are highlighted in red. DEGs with Wald > 3 and |logFC| < 2 are represented in blue. DEGs with Wald < 3 and |logFC| > 2 are represented in green. Non-significant (NS) DEGs are shown in grey. All DETEs are highlighted in black. Top DEGs are labelled. B) Heatmap displaying the top-30 differentially expressed genes or individual TE between primed and naive hESCs. Rows represent genes or individual TEs, and columns represent samples, clustered using Ward's linkage method. Gene expression is scaled by row (z-score), with red representing high expression and blue representing low expression. Sample conditions are annotated by genotype (WT or KO) and pluripotency state (primed or naive).



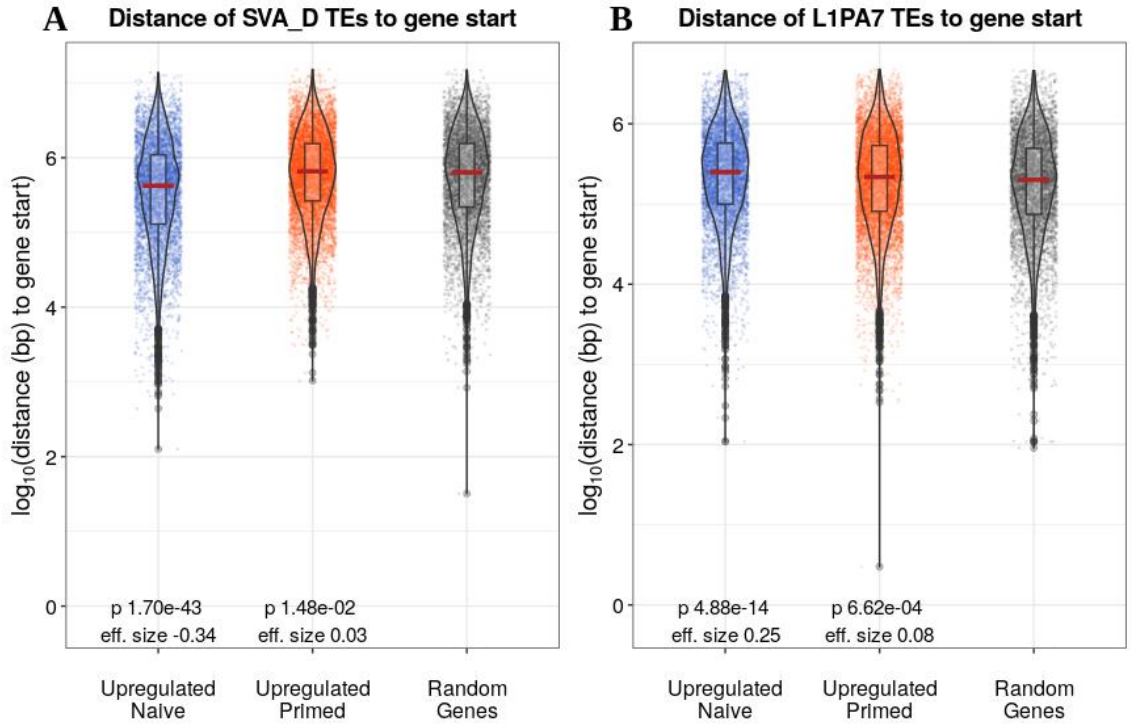
Supp. Fig. 7.4 - Individual elements from the LTR7/HERVH-int subfamilies are preferentially activated in primed hESCs. Gene set enrichment analysis for HERVH-int (**A**) and LTR7 TE subfamilies (**B**).



Supp. Fig. 7.5 - The size of the TE-sets does not have a major impact on the NES or expression levels of the corresponding subfamilies. Scatter plots comparing the size of the TE set with the NES (A) or with the median basemean expression across all samples (B), as well as the median basemean expression with the NES (C). Each point represents a TE subfamily. The red line indicates a fitted smoothing curve, with the shaded grey area representing the 95% confidence interval. Marginal density plots in each panel display the distribution of the variables along both axes (TE set size, NES, and median basemean expression). Rug plots along the axes mark individual data points.



Supp. Fig. 7.6 - Permutation tests for genomic association between SVA_D (naive-associated) or L1PA7 subfamilies (primed-associated) and genes upregulated in naive hESCs. The histograms show the distribution of simulated overlaps between random gene sets and the SVA_D (**A**) and L1PA7 (**B**) subfamilies, based on 1000 permutations. The observed number of overlaps between DEGs and the TE subfamilies (blue dashed line, labelled "Obs") is compared to the distribution of overlaps from the random gene sets (grey bars). The mean number of overlaps from the random permutations is indicated by the black dashed line (labelled "Perm"). The shaded red areas correspond to the tails of the distribution, showing more extreme overlap values beyond the 95th percentile.



Supp. Fig. 7.7 - Example of distance measurements between gene start of DEGs and elements of the SVA_D (naive-associated) and L1PA7 (primed-associated) subfamilies. The violin plots illustrate the log10-transformed distance distribution (in bp) from the gene start to the closest TE element for genes upregulated in naive (blue) or primed (orange) hESCs, compared to a random gene set of equal size (grey). Panel A represents the distances for the SVA_D subfamily, and Panel B represents distances for the L1PA7 subfamily. Red lines represent the median, and the width of the violin plot indicates the density of data points at different values. Statistical significance between upregulated gene groups and the random gene set was assessed using the Wilcoxon rank-sum test, with p-values and effect sizes annotated below each group. The effect size represents the proportionate difference in median distance between DEGs and the random gene set, providing a measure of practical significance in addition to the statistical test. These results provide insights into how TE proximity might influence gene regulation in distinct pluripotent states**.**

Supp. Table 7.1 - List of selected marker genes upregulated in naive hESCs and in primed hESCs. The table shows a selection of marker genes upregulated in naive (top table) or primed (middle and bottom tables) hESCs, using the random mapping strategy. For each gene, are reported: the name of the gene (gene_name); the mean base expression (baseMean), which represents the average normalised expression of the gene across all samples; the log2 fold change (log2FoldChange), which indicates the difference in expression levels between naïve and primed hESCs, where positive values indicate upregulation in naïve hESCs and negative values indicate upregulation in primed hESCs; the standard error of the log2 fold change (lfcSE); the Wald test statistic (stat) to assess significance; the unadjusted p-value (pvalue); and the adjusted p-value (padj) for multiple test corrections based on Benjamini-Hochberg (BH) procedure, which controls the false discovery rate (FDR).

Naïve hESCs marker genes						
gene_name	**baseMean**	**log2FoldChange**	**lfcSE**	**stat**	**pvalue**	**padj**
DPPA3	46340.26	8.077243	0.1041564	77.54917	0.00E+00	0.00E+00
DPPA5	57991.19	11.26276	0.1405004	80.16179	0.00E+00	0.00E+00
KLF4	10615.5	7.390404	0.09903644	74.62308	0.00E+00	0.00E+00
KLF17	3173.242	15.30998	1.183341	12.93793	2.75E-38	1.36E-37
DNMT3L	107564.2	18.56098	1.023562	18.13371	1.73E-73	1.57E-72
FGF4	16911.82	9.22395	0.1201835	76.74889	0.00E+00	0.00E+00
GATA6	1298.322	10.62586	0.5949236	17.86088	2.38E-71	2.10E-70
TBX3	2864.731	9.783727	1.058765	9.240693	2.45E-20	7.80E-20
IL6ST	20739.89	4.08806	0.08781371	46.55378	0.00E+00	0.00E+00
TRIM60	4984.352	15.97868	1.180354	13.53719	9.43E-42	4.99E-41
Primed hESCs marker genes						
gene_name	**baseMean**	**log2FoldChange**	**lfcSE**	**stat**	**pvalue**	**padj**
MYC	5546.194	-2.955311	0.1204393	-24.53777	5.84E-133	1.09E-131
DUSP6	44924	-8.93455	0.1069121	-83.56916	0.00E+00	0.00E+00
PTPRZ1	28596.29	-11.89426	0.2370423	-50.17781	0.00E+00	0.00E+00
ZIC2	5684.557	-6.675676	0.1339487	-49.83757	0.00E+00	0.00E+00
OTX2	3161.337	-1.337135	0.0631851	-21.16219	2.13E-99	2.72E-98
SFRP2	7369.454918	-5.790761207	0.128625258	-45.02040504	0.00E+00	0.00E+00
DNMT3B	73862.4	-1.090149	0.05488651	-19.86187	8.70E-88	9.58E-87
CD24	138857.7	-4.731889	0.1103416	-42.88401	0.00E+00	0.00E+00
Primed hESCs genes related to later developmental stages						
gene_name	**baseMean**	**log2FoldChange**	**lfcSE**	**stat**	**pvalue**	**padj**
SOX11	7726.353074	-7.023402602	0.119002093	-59.0191518	0.00E+00	0.00E+00
CYTL1	292.5497868	-11.40617295	1.185695443	-9.619816811	6.59E-22	2.19E-21
HMX2	644.4584	-9.889123	0.7440135	-13.29159	2.59E-40	1.33E-39
THY1	27220.11	-11.11297	0.5656974	-19.64472	6.41E-86	6.91E-85

Supp. Table 7.2 - List of the top-15 TE subfamilies upregulated in naive hESCs or in primed hESCs. The table shows the top-15 TE subfamilies upregulated in naïve (top table) or primed (bottom table) hESCs, obtained using the random mapping approach. For each subfamilies, are reported: the name of the subfamily (TE subfamily); the mean base expression (baseMean), which represents the average normalised expression of the subfamily across all samples; the log2 fold change (log2FoldChange), which indicates the difference in expression levels between naïve and primed hESCs, where positive values indicate upregulation in naïve hESCs and negative values indicate upregulation in primed hESCs; the standard error of the log2 fold change (lfcSE); the Wald test statistic (stat) to assess significance; the unadjusted p-value (pvalue); and the adjusted p-value (padj) for multiple test corrections based on Benjamini-Hochberg (BH) procedure, which controls the false discovery rate (FDR).

Top 15 TE subfamilies upregulated in naive hESCs						
TE subfamily	**baseMean**	**log2FoldChange**	**lfcSE**	**stat**	**pvalue**	**padj**
HERVK-int	238816.81729	4.91887493	0.10676095	46.073726	0.00E+00	0.00E+00
SVA_D	61308.99717	4.28366211	0.09313008	45.996548	0.00E+00	0.00E+00
LTR5_Hs	28750.46790	3.41377998	0.07434082	45.920666	0.00E+00	0.00E+00
ALR/Alpha	26381.02816	5.18540097	0.12088184	42.896443	0.00E+00	0.00E+00
SVA_E	17652.08989	3.63900388	0.09549093	38.108371	0.00E+00	0.00E+00
SVA_F	17976.77335	3.60291327	0.10273165	35.071112	1.86E-269	9.44E-268
PABL_A-int	1750.12707	3.31553013	0.10386773	31.920694	1.38E-223	5.23E-222
MER57-int	5359.86349	3.35974570	0.10804872	31.094729	2.84E-212	9.94E-211
BSR/Beta	2145.37562	3.76399034	0.12371749	30.424077	2.64E-203	8.61E-202
SVA_C	5085.02339	2.91017434	0.10603801	27.444634	8.05E-166	1.98E-164
HERVK9-int	22265.55068	3.02078203	0.11166526	27.052119	3.61E-161	8.49E-160
AluYa5	13107.10134	2.42366861	0.09210852	26.313186	1.35E-152	2.99E-151
LTR7B	1438.50368	2.54953433	0.09841416	25.906175	5.67E-148	1.20E-146
AluYg6	1961.15829	2.03368842	0.09862491	20.620434	1.80E-94	2.16E-93
SST1	8599.49134	4.77740242	0.23216068	20.578000	4.32E-94	5.16E-93
Top 15 TE subfamilies upregulated in primed hESCs						
TE subfamily	**baseMean**	**log2FoldChange**	**lfcSE**	**stat**	**pvalue**	**padj**
HERV9NC-int	15357.94994	-2.8856200	0.05531184	-52.170021	0.00E+00	0.00E+00
HERV4_I-int	8393.72202	-3.9735375	0.11513520	-34.511928	5.31E-261	2.51E-259
Charlie21a	5370.82756	-3.2507649	0.11556890	-28.128370	4.41E-174	1.16E-172
MER67D	2072.47711	-3.4290039	0.12423606	-27.600715	1.09E-167	2.72E-166
MamGypLTR1a	1648.16882	-2.3566969	0.11570416	-20.368299	3.20E-92	3.73E-91
LTR25	2770.86049	-1.8504017	0.09361805	-19.765436	5.91E-87	6.44E-86
L1P3b	595.72667	-3.2820374	0.17267023	-19.007546	1.48E-80	1.47E-79
HERV15-int	528.16244	-2.2800344	0.12055635	-18.912603	8.98E-80	8.85E-79
L1ME3C	19069.61710	-2.1629248	0.11828774	-18.285284	1.08E-74	9.99E-74
SATR2	2638.57540	-1.6436858	0.09846720	-16.692723	1.48E-62	1.14E-61
LTR7	12097.16071	-1.3512861	0.08347593	-16.187733	6.16E-59	4.44E-58
LTR1E	805.50961	-2.2142589	0.13721647	-16.136976	1.40E-58	1.01E-57
MLT1F2-int	534.60449	-1.5793575	0.10426824	-15.147062	7.92E-52	5.07E-51
LTR1	719.14052	-1.6374871	0.11641276	-14.066217	6.13E-45	3.43E-44
MER34C2	1019.34246	-1.0709540	0.08169174	-13.109697	2.90E-39	1.46E-38

Supp. Table 7.3 - List of the top-20 individual TE loci upregulated in naive hESCs or in primed hESCs. The table shows the top-20 individual TE loci upregulated in naïve (top table) or primed (bottom table) hESCs, using the unique mapping approach. For each loci, are reported: the name of the TE (TE loci); the mean base expression (baseMean), which represents the average normalised expression of the TE across all samples; the log2 fold change (log2FoldChange), which indicates the difference in expression levels between naïve and primed hESCs, where positive values indicate upregulation in naïve hESCs and negative values indicate upregulation in primed hESCs; the standard error of the log2 fold change (lfcSE); the Wald test statistic (stat) to assess significance; the unadjusted p-value (pvalue); and the adjusted p-value (padj) for multiple test corrections based on Benjamini-Hochberg (BH) procedure, which controls the false discovery rate (FDR).

Top 20 TE loci upregulated in naive hESCs						
TE loci	**baseMean**	**log2FoldChange**	**lfcSE**	**stat**	**pvalue**	**padj**
TE_HERVK-int_dup40	52078.2355	7.412943	0.09976848	74.30146	0.00E+00	0.00E+00
TE_HERVK-int_dup41	11606.4883	7.060836	0.12956959	54.49455	0.00E+00	0.00E+00
TE_HERVK-int_dup243	14679.5108	6.900207	0.15289303	45.13094	0.00E+00	0.00E+00
TE_HERVK-int_dup1	13149.1475	11.811891	0.31966597	36.95073	7.09E-299	6.47E-296
TE_HERVK9-int_dup204	2788.9808	4.655656	0.13002377	35.80619	8.85E-281	7.10E-278
TE_L1PB1_dup11768	1330.9627	4.712698	0.13538049	34.81076	1.67E-265	1.14E-262
TE_HERVK9-int_dup215	8678.9731	4.493187	0.13082686	34.34453	1.70E-258	1.08E-255
TE_HERVK-int_dup19	2252.4543	5.729057	0.16800586	34.10034	7.29E-255	4.49E-252
TE_HERVK9-int_dup203	2484.0094	4.205799	0.12679956	33.16887	3.03E-241	1.68E-238
TE_HERVK-int_dup178	2353.3142	4.264049	0.12908058	33.03401	2.64E-239	1.44E-236
TE_L1PB1_dup11776	845.6441	5.39536	0.16339461	33.02043	4.14E-239	2.24E-236
TE_HERVK-int_dup43	1130.4261	7.025472	0.23087078	30.43032	2.18E-203	8.10E-201
TE_MER9a3_dup162	851.7085	4.550588	0.14994322	30.34874	2.61E-202	9.59E-200
TE_L1PB1_dup11770	624.2397	6.141467	0.20283781	30.27772	2.25E-201	8.21E-199
TE_HERVK-int_dup103	3142.0576	10.24052	0.33969477	30.14624	1.20E-199	4.30E-197
TE_MER57-int_dup355	1766.4673	8.690326	0.29449122	29.50963	2.17E-191	7.22E-189
TE_HERVK-int_dup89	1206.4564	7.038918	0.23854899	29.50722	2.33E-191	7.72E-189
TE_HERVK-int_dup104	1332.8624	8.434475	0.29431416	28.65807	1.27E-180	3.84E-178
TE_LTR5B_dup439	1674.9596	5.986955	0.2100392	28.50399	1.05E-178	3.08E-176
TE_AluSz_dup96763	598.304	6.142673	0.22259398	27.59586	1.25E-167	3.37E-165
Top 20 TE loci upregulated in primed hESCs						
TE loci	**baseMean**	**log2FoldChange**	**lfcSE**	**stat**	**pvalue**	**padj**
TE_MSTB_dup5757	3142.0461	-6.33267	0.16255316	-38.95753	0.00E+00	0.00E+00
TE_L1PA6_dup1587	1291.1524	-5.522697	0.15931927	-34.66434	2.72E-263	1.82E-260
TE_L4_A_Mam_dup3475	1665.8136	-5.659464	0.16554045	-34.1878	3.67E-256	2.29E-253
TE_L1MC5a_dup11214	1385.5989	-5.91186	0.17306291	-34.16018	9.44E-256	5.83E-253
TE_L1M2_dup846	2471.6493	-4.1261	0.12193045	-33.83979	5.13E-251	3.06E-248
TE_L1M2_dup2248	1873.3993	-4.869188	0.15113167	-32.21819	9.82E-228	4.64E-225
TE_HERV9NC-int_dup197	2378.7869	-8.184548	0.25695638	-31.8519	1.24E-222	5.62E-220
TE_L1ME3C_dup342	6947.0044	-4.546596	0.14860951	-30.59425	1.46E-205	5.56E-203
TE_L1PREC2_dup2188	2103.159	-3.632435	0.11873734	-30.59219	1.55E-205	5.91E-203
TE_MSTA_dup4888	840.3756	-5.723731	0.19150105	-29.88877	2.75E-196	9.56E-194
TE_L1ME3A_dup7569	970.5264	-5.485789	0.18495636	-29.65991	2.53E-193	8.55E-191
TE_HERVH-int_dup2511	803.044	-5.359912	0.18537166	-28.91441	7.87E-184	2.46E-181
TE_L1PA8A_dup1594	1484.4264	-5.843994	0.20244871	-28.86654	3.14E-183	9.74E-181
TE_L1ME3C_dup341	5640.931	-4.439449	0.15440888	-28.75125	8.74E-182	2.67E-179
TE_MER41-int_dup1060	1233.203	-4.253501	0.15158979	-28.05928	3.08E-173	8.61E-171
TE_L1MEf_dup3311	2023.0843	-4.255071	0.15571597	-27.32585	2.09E-164	5.48E-162
TE_AluSq2_dup45891	2032.2446	-2.924998	0.10711293	-27.30761	3.45E-164	9.01E-162
TE_HERV9NC-int_dup205	9221.5458	-2.663978	0.09779418	-27.24066	2.14E-163	5.57E-161
TE_L1ME3A_dup4124	749.2363	-5.010081	0.18679149	-26.82178	1.80E-158	4.43E-156
TE_L1ME3Cz_dup3630	1443.4365	-4.483569	0.16725775	-26.80635	2.73E-158	6.67E-156

Supp. Table 7.4 - Statistical comparison of distances between naive upregulated genes and random genes for selected TE subfamilies. The table shows the results of statistical tests performed on the distances between TE subfamilies and both genes upregulated in naive hESCs and randomly selected genes (identical number of genes). For each TE subfamily, the Wilcoxon rank-sum test and the Kolmogorov-Smirnov (KS) test were applied to assess significant differences between the two groups. P-values were adjusted for multiple testing using the False Discovery Rate (FDR) method. Median distances between each TE subfamily and the upregulated genes and random genes are provided, along with the proportion of difference between the two groups, calculated as (median_observed - median_random) / median_random. Negative values in the proportion of difference indicate that genes upregulated in naive hESCs are located closer to the TE subfamily compared to random genes, while positive values indicate greater distances.

TE subfamilies	P-value Wilcoxon	P-value Wilcoxon FDR	P-value Kolmogorov Smirnov	P-value Kolmogorov Smirnov FDR	Observed median distances	Random median distances	Proportion of difference
BSR/Beta	2.89E-01	3.21E-01	6.16E-03	8.14E-03	25656231	25451240	0.008054264
SVA_D	2.41E-59	9.63E-58	0.00E+00	0.00E+00	421481	679888	-0.38007289
SST1	2.44E-03	3.37E-03	3.05E-04	5.21E-04	27933551	29869168	-0.06480318
SVA_B	5.20E-16	1.49E-15	6.89E-11	1.88E-10	930771	1195684	-0.2215577
SVA_C	2.29E-06	4.16E-06	7.52E-07	1.81E-06	2023581	2344796	-0.1369906
SVA_F	8.60E-24	3.82E-23	0.00E+00	0.00E+00	834061	1201688	-0.3059255
HERVK-int	5.14E-06	8.95E-06	3.06E-05	5.98E-05	7982692	9119846	-0.12469004
SVA_E	9.17E-09	2.04E-08	2.00E-06	4.09E-06	1207118	1411133	-0.14457532
LTR5_Hs	2.63E-34	1.75E-33	0.00E+00	0.00E+00	1327440	1924870	-0.31037421
LTR12C	1.34E-03	2.05E-03	2.35E-03	3.32E-03	563878	514706	0.09553415
L1HS	1.61E-15	4.29E-15	2.81E-12	8.87E-12	1345849	1113845	0.2082911
L1PA2	1.08E-16	3.59E-16	9.21E-12	2.70E-11	509954	419439	0.215800152
AluSg	1.65E-40	2.19E-39	0.00E+00	0.00E+00	17068.5	26440	-0.35444402
AluY	1.13E-21	4.10E-21	0.00E+00	0.00E+00	7054	9300	-0.24150538
AluSp	6.43E-37	6.43E-36	0.00E+00	0.00E+00	11869	18270	-0.35035577
AluSq2	5.91E-34	3.38E-33	0.00E+00	0.00E+00	10592	15295	-0.30748611
AluSx	2.86E-35	2.29E-34	0.00E+00	0.00E+00	5638	8217.5	-0.31390326
AluSx1	1.25E-40	2.19E-39	0.00E+00	0.00E+00	5462	8581	-0.36347745
AluJb	1.21E-22	4.82E-22	0.00E+00	0.00E+00	5993	7747	-0.22641022
AluSz	4.08E-26	2.04E-25	0.00E+00	0.00E+00	6843	9340	-0.26734475
MIRb	1.28E-02	1.65E-02	9.32E-02	1.06E-01	5986	5670	0.055731922
MIR	7.39E-02	8.96E-02	9.32E-02	1.06E-01	7592	7217	0.051960648
L2b	6.98E-01	6.98E-01	9.66E-01	9.90E-01	15383	15103	0.018539363
L2c	8.95E-04	1.43E-03	7.64E-03	9.79E-03	12326	11117	0.108752361
L2a	1.83E-01	2.09E-01	8.00E-02	9.65E-02	9899	9310	0.063265306
Tigger1	1.59E-07	3.03E-07	5.40E-05	9.62E-05	255026	230827	0.104836089
L3	4.52E-08	9.52E-08	8.39E-07	1.91E-06	45068	38865	0.159603757
HAL1	3.03E-01	3.28E-01	1.38E-01	1.48E-01	110233	108526	0.01572895
L4_A_Mam	2.68E-03	3.57E-03	7.32E-04	1.20E-03	341859	311671	0.096858546
L1MEd	2.88E-02	3.60E-02	9.80E-02	1.09E-01	147751	142404	0.037548103
L1ME1	1.59E-07	3.03E-07	5.40E-05	9.62E-05	89659	79584	0.126595798
MLT1D	3.39E-04	5.66E-04	8.60E-04	1.36E-03	88049	79763	0.103882753
L1MC3	1.70E-03	2.51E-03	1.61E-03	2.44E-03	209354	198803	0.05307264
L1ME3A	4.21E-01	4.32E-01	5.09E-01	5.35E-01	131135	126256	0.038643708
L1MEf	2.23E-03	3.19E-03	5.57E-03	7.62E-03	344435	315761	0.090809188
L1PA7	4.98E-16	1.49E-15	2.55E-15	8.72E-15	249790	197564	0.26434978
L1MA4	1.59E-10	3.97E-10	9.37E-07	2.02E-06	255746	220908	0.157703659
L1PB1	7.70E-10	1.81E-09	8.57E-08	2.20E-07	223602	194158	0.151649687
LTR7	4.11E-01	4.32E-01	1.87E-03	2.74E-03	1012778	981691	0.031666787
HERVH-int	1.77E-01	2.08E-01	5.53E-02	6.87E-02	1076492	1092508	-0.01465985

Supp. Table 7.5 - Statistical comparison of distances between primed upregulated genes and random genes for selected TE subfamilies. The table shows the results of statistical tests performed on the distances between TE subfamilies and both genes upregulated in primed hESCs and randomly selected genes (identical number of genes). For each TE subfamily, the Wilcoxon rank-sum test and the Kolmogorov-Smirnov (KS) test were applied to assess significant differences between the two groups. P-values were adjusted for multiple testing using the False Discovery Rate (FDR) method. Median distances between each TE subfamily and the upregulated genes and random genes are provided, along with the proportion of difference between the two groups, calculated as (median_observed - median_random) / median_random. Negative values in the proportion of difference indicate that genes upregulated in naive hESCs are located closer to the TE subfamily compared to random genes, while positive values indicate greater distances.

TE subfamilies	P-value Wilcoxon	P-value Wilcoxon FDR	P-value Kolmogorov Smirnov	P-value Kolmogorov Smirnov FDR	Observed median distances	Random median distances	Proportion of difference
BSR/Beta	1.38E-04	5.02E-04	3.59E-07	1.31E-06	26522016	25666058.5	0.033349784
SVA_D	1.52E-04	5.06E-04	1.08E-06	3.60E-06	656494	610903	0.074628869
SST1	6.22E-03	1.31E-02	5.87E-04	1.30E-03	30776402	29734638	0.035035369
SVA_B	2.06E-01	2.74E-01	1.94E-01	2.43E-01	1216235.5	1173983.5	0.035990284
SVA_C	6.55E-01	7.09E-01	3.80E-02	5.62E-02	2371161	2390632.5	-0.00814492
SVA_F	3.26E-02	5.67E-02	7.50E-02	1.07E-01	1183479.5	1145253	0.033378214
HERVK-int	6.92E-03	1.38E-02	1.21E-02	2.02E-02	9614517.5	9110347.5	0.05534037
SVA_E	4.46E-01	5.25E-01	2.97E-01	3.49E-01	1340970	1400096.5	-0.0422303
LTR5_Hs	2.52E-09	1.68E-08	1.54E-09	8.82E-09	2074722.5	1802128.5	0.151262244
LTR12C	2.60E-01	3.35E-01	4.07E-01	4.65E-01	516843.5	507601	0.018208199
L1HS	1.38E-03	3.45E-03	9.69E-03	1.69E-02	1217910	1140142.5	0.068208579
L1PA2	5.48E-01	6.09E-01	7.40E-01	7.59E-01	421199	429688	-0.0197562
AluSg	3.87E-02	6.19E-02	2.77E-04	6.53E-04	25247	23989	0.052440702
AluY	7.11E-07	3.16E-06	8.80E-11	5.86E-10	10323.5	9402	0.098011062
AluSp	1.07E-02	2.05E-02	2.30E-03	4.59E-03	19249.5	18172.5	0.059265374
AluSq2	8.88E-01	9.34E-01	6.83E-03	1.25E-02	15004	15417.5	-0.02682017
AluSx	1.15E-03	3.28E-03	2.35E-07	9.42E-07	8639	8107.5	0.065556583
AluSx1	1.77E-06	7.08E-06	2.84E-08	1.26E-07	8485	7798.5	0.088029749
AluJb	3.62E-02	6.03E-02	2.08E-08	1.04E-07	7822	7816.5	0.00070364
AluSz	1.54E-03	3.63E-03	2.73E-05	7.80E-05	9926.5	9125.5	0.087776012
MIRb	4.71E-20	9.42E-19	0.00E+00	0.00E+00	4712	6042	-0.22012579
MIR	1.61E-18	2.14E-17	2.22E-16	2.96E-15	5915	7306	-0.19039146
L2b	6.80E-26	2.72E-24	0.00E+00	0.00E+00	11740.5	15156.5	-0.22538185
L2c	1.51E-15	1.51E-14	1.08E-11	8.63E-11	9217	11045	-0.16550475
L2a	3.90E-13	3.12E-12	1.09E-12	1.09E-11	7634.5	9330	-0.18172562
Tigger1	9.57E-01	9.57E-01	9.44E-01	9.44E-01	237162.5	234164	0.012805128
L3	1.23E-07	6.16E-07	2.69E-06	8.28E-06	34632	38390.5	-0.09790182
HAL1	1.14E-01	1.69E-01	2.42E-02	3.87E-02	105537	109969	-0.04030227
L4_A_Mam	3.68E-04	1.13E-03	2.30E-03	4.59E-03	269822	297163.5	-0.09200827
L1MEd	2.05E-01	2.74E-01	1.03E-01	1.42E-01	134602	139590	-0.03573322
L1ME1	4.88E-01	5.58E-01	6.10E-01	6.60E-01	80836	82049	-0.01478385
MLT1D	9.53E-01	9.57E-01	6.53E-01	6.87E-01	78879	80242	-0.01698612
L1MC3	3.46E-03	7.68E-03	6.87E-03	1.25E-02	207788	192892.5	0.077221769
L1ME3A	3.52E-01	4.40E-01	5.25E-01	5.83E-01	129308	129805.5	-0.00383266
L1MEf	8.73E-02	1.34E-01	1.12E-01	1.49E-01	313103	303013.5	0.033297196
L1PA7	6.07E-08	3.47E-07	3.48E-05	9.29E-05	217212	191463	0.134485514
L1MA4	1.32E-03	3.45E-03	1.78E-04	4.46E-04	246003	228368.5	0.077219494
L1PB1	1.90E-01	2.71E-01	2.68E-01	3.25E-01	201857.5	197039.5	0.02445195
LTR7	4.23E-01	5.13E-01	1.37E-01	1.77E-01	942324	960507.5	-0.01893114
HERVH-int	1.82E-02	3.30E-02	2.82E-02	4.34E-02	1025656.5	1094604.5	-0.06298896