An Atlas of Transposable Elements Associated with the Epigenome Across Human Cell Types

Jeffrey Hyacinthe

Quantitative Life Sciences

McGill University, Montreal, Canada

December 2024

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Abstract

Transposable elements (TEs) are DNA elements able to create copies of themselves across the genome and constitute about half of the human genome. TEs have long been overlooked partly due to the complexity of their analysis but also due to their overall low of activity. However, despite their silencing, there is a rising amount of evidence for their involvement in regulation such as their predominance in regulatory sequences and some cases of co-options like the *AIM2* gene coming from a MER41 TE. It was also observed that older TEs tend to have more regulatory potential and become more enhancer-like as they age in terms of epigenetic state, transcription factor binding potential or methylation levels. This highlights that a large part of TE's impact might come from their relationship with the epigenome rather than their expression. However, a comprehensive analysis of TEs and its relationship with the epigenome and especially the reference histone marks remain lacking.

In this work, I present an analysis of TEs and their association with the epigenome across human cell types through a large and varied dataset. First, we leveraged a new dataset from the international human epigenome consortium (IHEC) with over 4867 ChIP-seq samples across 6 histone marks and 175 cell annotations grouped into 47 cell categories to show that TEs have drastically different enrichment levels across histone marks. We observe that although TEs cover on average 55.8% of the histone peaks, TEs are generally depleted in repressive H3K9me3 histone mark, except for L1 where they are highly enriched. In contrast, MIRs were enriched in H3K4me1, H3K27ac and H3K27me3 and Alus were enriched in H3K36me3. We also find some significant differences in TE enrichment between cell types and that in 20% of the cases, these enrichments were cell-type specific. We report that at least 4% of health status comparison, within cell types

featuring both healthy and diseased samples, were significantly different. Finally, we used cell type specificity and extreme enrichments to identify 456 TE-cell type-histone association standout candidates.

Next, we used the results of our analysis to build an online portal with interactive access to the results and data. We realized that the true barrier to more widespread consideration of TEs may be accessibility and ease of analysis. Thus, we built an online portal that allows users to not only look for any TE's coverage and enrichment within any histone marks and cell types but also to rapidly analyze the TEs within their own data in a similar way and directly compare the results with the portal's IHEC dataset.

This work further supports the role of TE in genome regulation, highlights novel relationships between TEs and the epigenome and makes the analysis and interpretation of TEs easier and more accessible.

Résumé

Les transposons sont des séquences ADN ayant l'habilité de créé des copies d'elles même a travers le génome et compose environ la moitié du génome humain. Les transposons ont longtemps été ignorés en grande part du a la complexité de leur analyse, mais aussi a cause de leur bas niveau d'activité. Par contre, malgré leur répression, il y a de plus en plus de support pour leur implication dans la régulation tel leur prédominance dans les séquences régulatrices et dans certain cas, leur intégration au génome come le gène *AIM2* provenant d'un transposon MER41. Il fut aussi observé que les transposons plus anciens ont tendance à gagner du potentiel de régulation et devenir plus similaire a des amplificateurs avec l'âge. Cela amène l'idée que peut être que l'influence des transposons provienne de leur relation avec l'épigénome plutôt que leur expression. Par contre, une analyse compréhensive des transposons et leur relation avec l'épigénome, plus particulièrement avec les modifications histones demeure manquante.

Dans cette thèse, je présente une analyse des transposons et leur association avec l'épigénome à travers les différents types de cellules humaines grâce à une large banque de données. Premièrement, nous avons utilisé les nouvelles données du consortium international de l'épigénome humain (IHEC) contenant 4867 échantillons ChIP-seq à travers 6 modifications d'histones et 175 annotations de cellules regroupées dans 47 type de cellules. Nous démontrons que les transposons ont des niveaux d'enrichissement différents dépendamment de la modification d'histone avec lequel ils sont associés. Nous observons que les transposons couvrent 55.8% des pics d'histones, que transposons sont généralement appauvri en H3K9me3 sauf pour L1 qui est extrêmement enrichi. Par contraste, MIR est enrichi dans H3K4me1, H3K27ac et H3K27me3 et Alu enrichi pour H3K36me3. Nous trouvons des différences significatives en termes

d’enrichissement entres les types de cellules et dans 20% des cas, ces enrichissements sont spécifiques au type de cellule. Nous reportons qu’il y a des différences significatives entre l’état de santé dans un type de cellule pour aux moins 4% des comparaisons. Finalement, nous avons utilisé les associations extrême et spécifique entre cellule et transposons pour identifier 456 candidats pour association notable entre transposon, histone et type de cellule.

Ensuite, nous avons utilisé nos résultats pour développer un portail en ligne offrant un accès dynamique à nos résultats et donnés. Nous avons réalisé qu’une des barrières additionnelles à l’adoption des transposons dans les analyses génomiques pourrais aussi être l’accessibilité aux donnés. Donc, nous avons construits un portail qui permets aux chercheurs de non seulement, observer le croisement des histones par les transposons et types de cellules, mais aussi de lancer des analyses de leurs propres données et les comparer avec celles inclut dans le portail.

Le travail de cette thèse offre du support supplémentaire au rôle des transposons dans la régulation du génome, présente des relations entre les transposons et l’épigénome et facilite l’analyse et l’interprétation des transposons.

Table of Contents

List of Abbreviations

APC: Adenomatous polyposis coli

ATAC-seq: Assay for transposase-accessible chromatin with sequencing

BED: Browser Extensible Data

Bp: Base pair

ChIP-seq: Chromatin immunoprecipitation followed by sequencing

DHS: DNase I hypersensitive sites

DHS: DNase I hypersensitive sites

DMR: Differentially methylated region

DNA: DeoxyriboNucleic Acid.

DNMT: DNA methyltransferases

ENG: Endoglin gene

ERV: Endogenous retroviruses

GO: Gene ontology

GREAT: Genomic Regions Enrichment of Annotations Tool

GTEX: Genotype-Tissue Expression

GWAS: Genome Wide Association Study

HERV: Human endogenous retroviruses

HMM: Hidden markov model

IHEC: International human epigenome consortium

iPSC: Induced pluripotent stem cell

Kbp: Kilo basepair

KRAB-ZFP: Krüppel-associated box domain zinc finger proteins

KZFP: Krüppel-associated box domain zinc finger proteins

L1: LINE1

L2: LINE2

LINE: Long interspersed elements

LTR: Long terminal repeat

Mbp: Mega basepair

MIR: Mammalian wide interspersed repeat

Myr: Million Years

ORF: Open reading frames

PC: Principal component

PCA: Principal component analysis

RB1: Retinoblastoma 1

RNA: RiboNucleic Acid

Seq: Sequencing

SINE: Short interspersed elements

SVA: SINE/VNTR/Alu.

TE: Transposable element

TF: Transcription factor

TFBS: Transcription factor binding site

TSS: Transcription start site

UCSC: University of California, Santa Cruz

UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction

WGBS: Whole genome bisulfite sequencing

ZFP: Zinc finger protein

List of Figures

List of Tables

Acknowledgments

First, I would like to thank my supervisor Dr. Guillaume Bourque for his guidance and support. Guillaume was a pleasure to work with, always supportive, positive and insightful. Through him I have learned a lot and was blessed with a lot of opportunities, international consortium participation, from endless computational resources to the wonderful team at C3G and very pleasant conference travel opportunities. He granted me a lot of freedom and independence, which allowed me to try a many things, expand my skillset as a researcher and truly feel like my projects were my own. Thank you.

I would also like to thank my supervisory committee, Dr. Hamed Najafabadi and Dr. David Langlais, their insight really helped me during my PhD, sharing some of my schedule optimism but also sometimes bringing some much-needed reality checks. And honestly, one of their suggestions ended up being one of my favorite part of my project, so thank you.

I acknowledge C3G for their support and being an amazing group to work with. Special thanks to PO, David Lougheed and Hector for all the tech support. I also want to acknowledge Calcul Quebec and The Digital Research Alliance of Canada for their computing resources. I also thank the IHEC consortium for their data.

I must thank all my (current and former) lab members for their help and making all these years a joy! Pat really had the life and energy for the greatest ideas, I can't believe we had (live) Christmas trees! Audrey was also a joy to have around and gave some much-needed life to the QLS side of life. Cristian my desk neighbor! The pro linux guy that was always fun to talk with, on all subjects. Also was a great roommate to have during our conference trips, really enjoyed your company. Qinwei probably took way too much of my time in deep talks, always insightful, inspiring and I've learned a lot from him. Clément came like a storm, with an enthusiasm and energy I can only hope to match one day. As a senior, his insight, perspective and support have helped me tremendously. thank you all! Oh and the new students have also been amazing, good luck to you!

I wouldn't have made it where I am today if it wasn't for the wonderful Quantitative Life Sciences program. I am grateful for the rotation year opportunity that allowed me to find my great lab and acknowledge the funding. I must thank (former Co-director) Dr. Mathieu Blanchette for more or less having set me on this path through his undergrad bioinformatics courses. In fact, he might be where I even learned about the concept of research and bioinformatics as a field. Big thanks for that. Then I must thank (former director) Dr. Celia Greenwood. When I started in the program was brand new, not officially approved yet, we were the first cohort, the guinea pigs, and the ordeal seemed daunting. Celia was the one who brought all the positivity and reassurances that everything was going to work out. Throughout my PhD she also acted as a much-needed moral support. And as I write this thesis, clearly she was right. (coordinator) Alex was critical and just making sure everything was on track and making QLS a great program, thank you! Dr. Rob Sladek was critical toward the end of my PhD, in being both a support and guide in making sure that my thesis was a success and for that I am very grateful.

But QLS was not just the administrators, as a member of the first QLS cohort, the strong bonds that were built with the other students are what made me (and I am sure them) get through the first challenging years. The fact that we kept in touch even as we went our separate ways after joining our respective lab supports that. Even as the last of us write our thesis we still cheer and support each other and for that I am forever grateful. So Alex, Matt, Sara, YiXiao, Selin and Myriah a big thank you for everything. I have been fortunate to count you as friends and even, something like a family.

Extracurricular activities helped get my mind get out of the research but also make me a better-rounded individual. I'd like to thank the SAFE committee members for the good times and the great work! I would also like to thank Dr. Aimee Ryan who was involved in a lot of the groups that I joined. Thank you for all your involvement and support! I would also like to thank Dr. Loydie Majewska for her contribution to SAFE, her insight into the realities of being a black scientist, her support and getting me involved in the Canadian black scientist network. My time in the CBSN steering committee was very special. It was a unique environment that was particularly inspiring and the work we did was truly remarkable. Whenever I think about it, I cannot believe that I was involved in such an important initiative. For that I am forever grateful to Loydie, Maydianne and the steering community.

Big thanks to my university friends (Camberly, Chloe, Maryse, Gabriella, Laurie and Maky) who stuck with me throughout all those years and were always helpful, supportive and fun to hang out with!

Can't forget my high school gang (Kevin, Jonathan, Luca, Kristoff, Elsy) with whom I've now shared more than half of my life. You've helped shaped who I am today, you are all amazing, and very important to me, but I am sure that you already know that! Kevin, this is perhaps the time I finally become a new man.

Finally, I want to acknowledge my family that has just been a great environment. I thank my grandparents, that have always encouraged me and supported me in my long studies. I want to thank my brother Ray who also supported me, gave me a lot of life advice and with whom I had too many deep long discussions. And I must thank my parents Ogino and Katia who raised me, supported me, and always cheered for me even as my studies got rather long. I am truly grateful to you, for everything that you have and continue to do for me, this is all thanks to you!

Thesis Format

This manuscript-based thesis is comprised of 5 chapters. Chapter 1 is an introduction containing a literature review of topics relevant to my PhD work and thesis and presents its hypothesis and the objectives. Chapter 2 and 3 are original research chapters. Chapter 2 is a manuscript that was published in preprint on biorxiv. Chapter 3 is a manuscript to be submitted. Chapter 0 is a discussion of the results and their implications and contains preliminary work on expanding the work to DNA methylation. Chapter 0 describes new directions in which the work could be taken, potential expansions to the work and provides a conclusion.

Appendix A lists other manuscripts that the author contributed to. Appendix B andAppendix C contain the supplementary materials associated with chapter 2 and 0, respectively.

Contribution of Authors

Chapter 1 contains a literature review covering transposable elements, the epigenome and its main components and genomic tools. Chapter 0 and 0 contain a general discussion and future directions. These chapters were written by the author under the supervision of Guillaume Bourque.

Chapter 2 represents a manuscript authored by Jeffrey Hyacinthe and Guillaume Bourque. The thesis author developed the computational methods for the transposable element analysis, did the analysis, generated the figures. Jeffrey Hyacinthe wrote the manuscript with editing of Guillaume Bourque. Guillaume Bourque and the author conceived the study.

Chapter 3 represents a manuscript authored by Jeffrey Hyacinthe, David Lougheed and Guillaume Bourque. The thesis author contributed to the original tool concept, the generation of some of the data, developed the computational methods, did the analysis, generated the figures. Jeffrey Hyacinthe wrote the manuscript with editing of Guillaume Bourque. DL contributed to the online deployment of the tool.

Author contributions to other manuscripts are listed in Appendix A.

Chapter 1

Introduction

1.1 Transposable elements

1.1.1 Transposable Elements: The Long Elusive DNA Elements

Transposable elements (TEs) are DNA sequences with the ability to create copies of themselves and change position within the genome. From their discovery in the 1950s by Barbara McClintock[1], TEs, also known as mobile elements and repeats, have faced skepticism and pushback on their place and importance within the genome. Due to their large number of copies genomes yet their sequences not necessarily being translated into productive proteins to the host, TEs have long been dismissed as selfish elements or junk DNA[2,3]. However, a growing body of evidence is lending credence to TEs contribution to the regulatory network, evolution and disease[4–7]. In humans, TEs account for about 50% of the genome, but the vast majority of elements have lost their ability to transpose[8,9]. TEs are also generally repressed so that, those that still can, do not transpose and threaten genome integrity. Nonetheless, their remains within our genome can help trace back the evolution of our regulatory transcriptome and the rare novel transposition can lead to diseases, cancers and sometimes evolutionary innovation.

1.1.2 Family classification and characteristics

Transposable elements are classified in two classes depending on their transposition mechanism. DNA transposons which excise themselves before re-integrating the genome in another location. This is more akin to cut and paste as the original copy is removed. However, artefacts, errors,

repair and various mechanisms enable the number of copies to increase and for some partial remnants to remain at excisions sites[10,11]. The other main group are retrotransposons are elements that pass by an RNA intermediate which reverse transcribes back into DNA to reintegrates the genome in another location[12,13]. This mechanism is often referred to as copy and paste, the original copy remains and a new one is integrated in the genome.

DNA transposon cover about 3% of the genome and are no longer active but used to be 37 million years ago(Myr)[14]. Retrotransposons represents most of the transposable elements within the genome and are grouped into two large categories depending on the presence or absence of long terminal repeats (LTRs). LTR retrotransposons, commonly called human endogenous retroviruses (HERV or ERV), were inserted 25 Myr ago with limited activity and cover 8% of the human genome. Non retrotransposons make up the majority of all TEs. According to repeatmasker's classification[15], they include long interspersed elements (LINE) and short interspersed elements (SINE) classes. These classes each contain families such as LINE1 (L1) Alu and SVA (SINE/VNTR/Alu; a composite family). These 3 families constitute about 1/3 of the human genome and are the only ones with some reported current activity[13,16,17]. L1 and Alu, for instance, account for 60% of all interspersed repeats which seems to be specific to human as other organisms do not share that elevated concentration[8].

The families can be subdivided into subfamilies based on nucleotide insertion and deletions shared across the members of the subfamily. Since only a few elements successfully transfer, these elements own unique sequence is the one that propagates causing an expansion of that specific element, which eventually forms a subfamily. There is thus a form of hierarchical relationship between subfamilies as they will tend to originate from an older subfamily akin to a linear evolutionary sequence. For instance, all L1 subfamilies in human originate from a single lineage

over the past 40 Myr[18]. It is important to note that the subfamily names are attempts to group elements in a form of relatedness, however, the methods are not necessarily perfect and some annotations can sometimes feel inaccurate[19–21].

Thus TE families and subfamilies all have their own sequences, history and properties. L1 are LINEs of about 6 kilo base pairs (kbp) with two open reading frames (ORF1 and ORF2) which cover about 17% of the human genome through their over 500,000 copies[8]. However, the vast majority are no longer active due to incomplete reverse transcription, truncation and stop codons within ORFs[22]. The active L1 elements are mostly found within L1HS (or L1PA1) and L1PA2, the youngest L1PA subfamilies[8,23]. Alus are a primate specific family of 1 million 280 nucleotides long copies covering about 11% of the genome. The family is 65 Myr old and since it has no coding capacity, it depends on part of the L1 replication machinery to transpose. Active Alus are AluY, AluYa5 and AluYb8[24,25]. SVAs are even younger, particular to hominoid evolution (25MYR) with 3000 2 kbp long copies[26]. They are also non-autonomous and depend on L1 machinery. The more ancient families comprise only a small proportion of the human genome, nonetheless some families such as LINE2 (L2) and Mammalian wide interspersed repeat (MIR) have had major expansions[8,27].

1.1.3 General consequences and function

Expansions

The unique transposition ability of transposable elements cause consequences to the host genome. The most obvious one is genome expansions. As previously noted, many of the elements have thousands of instances, if not hundreds of thousands going up all the way to millions. Each instance brings with it additional base pairs to the genome that adds up over evolutionary timescales. L1

and Alu contributed about 750Mb or the human epigenome[8], over the past 6 Myr they have contributed to 8 Mb to the human genome[28]. But the TEs expansions are not a constant linear process, it often happens in spontaneous burst or expansions. For instance, AluYb is and ancient family (18-25 Myr) dating to early hominid which had for the longest time very low activity. Only recently, over the past few million years, did it have an expansion to ~2000 elements[29]. This is notable in that it suggests that the family remained able to transpose but did not explode until a copy became highly active. It is worth noting that a more active family would be deleterious and thus selected against and thus low transposition may be a beneficial features for TEs[30].

Germ line

While transposition may happen in any cell, it is important to remember that they may only be inherited and integrated when in the germ line. That is the only way for the transposition to pass on to the next progeny. Such events are fairly rare (estimated to be 1/8 individuals being born with a new transpositions), it is reported that there are about 1 SVA per 900 births, 1 L1 in 200 based on disease and 1 Alu per 20 births based on genome and disease[31].

Insertions

Another important thing to note is that TEs are not randomly distributed[32]. Since, as mentioned before, the different TEs have their own size, instance count, properties and expansions, it makes sense that TEs are not randomly distributed. In fact, ideal position are usually a balancing act between potential for future propagation and avoiding deleterious impact on host cell[33]. Some elements have even evolved their own mechanism to target those ideal sites[32]. Overall, TE insertions follow the rules of natural selection: deleterious insertions are discarded and only neutral

ones may remain, but they in turn may become unrecognizable though continuous mutations. For instance, L1 can be found in gene exons[34] but is rarely found within a gene[33,35].

There are a lot of ways by which TE insertions can impact genome structure, TEs can insert within a gene leading to a disrupted gene, ectopic recombination between non-allelic homologous retrotransposons may cause genomic rearrangements such as deletion or duplication of the intervening genomic sequence, and during duplication of a TE it is possible for its flanking regions on either side to also be copied and inserted leading to a 3' or 5' transduction[13,36].

But TEs' impact is not exclusive to insertion led disruptions, some TE proteins such as L1's ORF2p can induce DNA break and genome instability[37]. Accumulation of RNA transcript can also trigger innate immune response causing autoimmune diseases. The activation of interferon response is a supported property of ERV transcripts[38,39].

While, as previously mentioned, TEs can only be integrated and conserved if inserted in the germline, somatic transposition also happens. TE roles within soma includes early embryo and stem cells (pluripotency), expression within brain potentially for brain plasticity and TEs are often found in cancer and tumors since some insertions can disrupt tumor suppressors and oncogenes leading to cancer[40].

Repression

Since transposable elements have such deleterious potential, the genome has many mechanisms to silence them such as TRIM28-mediated transcription silencing, the repressive histone marks H3K9me3 and H3K27me3, DNA methylation and heterochromatin[41–44]. TEs are generally found within heterochromatin, inaccessible due to DNA methylation or H3K9me3[45] and are not expressed nor do they transpose. TEs are also targeted by Krüppel-associated box domain zinc

finger proteins (KZFP), a family of transcriptional regulators of higher vertebrates[45]. KRAB binding factors KAP1/TRIM28 as well as some KZFPs are able to repress TEs. The majority of KZFP associate with at least one TE while some TE subfamilies are recognized by multiple KZFP. With TEs being repressed by KZFP able to bind with them, a sort of arms race is established as TEs go through mutations that enables escape repression until KZFP genes evolve new zinc finger arrays that can recognize the escaped TE[45–47].

1.1.4 Currently known functional cases

TEs can impact the genome in many ways, from disrupting regulatory enhancers, promoters or regulatory elements all the way to the creation of new genes by insertion or the disruption of existing genes. While these can have positive effect on fitness, TE transposition and activity often causes diseases. Of the 124 disease causing insertions that have been reported, most are from insertional mutagenesis or aberrant splicing[34]. There has been reports of TEs being active in brain[48–50]. Through a retrotransposition method by Baillie et al. it was found that L1, Alu and SVA retotransposition occurred in human hippocampus and caudate nucleus[51]. While the extent of retrotransposition remains unclear, there is definitely TE activity within the brain. Some speculate that this retrotransposition and the resulting mosaicism may have to do with brain plasticity[48,52]. In fact, there is also reports of TEs being involved in brain diseases. Increased L1 transposition in neuron in schizophrenia[53]. Over expression of TE derived envelope (Env) proteins can be cytotoxic and has been linked to neurodegenerative diseases[33,54].

It is widely understood that genome instability is a hallmark of cancer. Thus, it is not surprising that TEs, with so many ways to disrupt the genome, are associated with cancer. One reported

disease causing insertion would be an L1 insertion into the adenomatous polyposis coli (APC) gene of a cancer patient[55]. Notably, that insertion was somatic as it was absent from the rest of the colon. This shows how an insertion in the wrong place can lead to issues even somatically. Another example is a de novo L1 insertion in intron 14 of the tumor suppressor retinoblastoma 1 (RB1) which caused aberrant RB1 splicing leading to retinoblastoma.

A standout context in which TEs appear to have evolutionarily contributed is human placenta and pregnancy. The placenta is a complex organ crucial to pregnancy that controls blood and nutrient exchange between fetus and mother. THE1B was identified as a ERV TE Using genetic editing, it was demonstrated that some TEs act as transcriptional enhancers to placental genes such as CSF1R and PSG5[56]. In addition, they identified LTR10A as an element that regulates endoglin gene (ENG) expression.

The immune system is a host defense system against retroviruses but also against retrotransposition. While viruses infect the cell from the outside, TEs generally remain within the cell. If TEs such as Alu are expressed in a cell, which can happen during viral infection[57], various nucleic acid forms of Alu may appear such as single stranded RNA, double stranded DNA, heteroduplex or double stranded DNA[58]. Since those a native to host cell, the innate immune system can tolerate a basal level of these free Alus. However, if their level grows to high or the immune system reacts too strongly, it may lead to autoimmune disorders[59]. In a study of TE's impact on flu infection, Chen et al. found that many TEs were upregulated post infection but that it did not correlate with viral load[60]. Furthermore, there was high inter-individual chromatin accessibility variability. They reported that there was some difference in the behavior and properties of sites with high or low variability and that KRAB-ZNF may have a role in immunity.

While TE activity is associated with a lot of diseases, there is also potential therapeutic or diagnostic usage. Since TE expression is often a sign that genome integrity has been compromised, TE measurements could be leveraged as a diagnostic as has been proposed for L1 in cancer[61].

Taken together, it is clear that TEs have had and continue to maintain a large impact on human evolution. And as the recent discovery of TEs potentially being the cause of the human tail loss[62] and SVA insertions affecting skin pigmentation[63] shows, more research into TEs is sure reveal new insight into human evolution, phenotypes and health.

1.2 Epigenome

1.2.1 The epigenome: Going Beyond the DNA Sequence

The epigenome is the set of marks, modification and chemical compounds that control and modulate the genome. Thus, epigenomics studies the phenotype changes and mechanisms that are not linked to the DNA sequence. While the DNA sequence contains the instructions for genes and their expression, there are multiple mechanisms that can further modulate gene expression. After all, if all cells in an organism possess the same DNA, how can they develop and behave differently? This is done through multiple mechanisms including modifying the chromatin, the physiological form of our genetic information[64]. The introduction of chromatin immunoprecipitation and sequencing (ChIP-seq) has enabled the profiling of the epigenome through measurement of histones marks and genomic sequences[64]. The epigenome is mostly investigated though histone marks, DNA methylation and chromatin state.

1.2.2 Histone marks

DNA is compacted through a complex coiling organization called chromatin. A major component of chromatin are the nucleosomes, protein complexes wrapped by DNA. A nucleosome is made of 2 subunits containing 4 histones (H2A, H2B, H3 and H4) and each of these histone can have several modifications at several positions. The main modifications (commonly referred as marks) are acetylation, methylation, phosphorylation and ubiquitylation and we mainly focused on acetylation and methylation, the most commonly studied modifications. Each of the modifications have different properties and lead to changes in the histone and thus DNA properties and interactions. For instance, acetylation usually occurs on lysine residues, neutralizing their positive charge and reducing the interaction potential[65].

Thus an example of a histone mark would be H3K27ac, which describes an acetylation (ac) of the 27th lysine (K27) on the histone H3. Histone marks are commonly the target of ChIP-Seq experiments to assess their binding sites and the regions in which they may have an influence. Thus there is a large amount of ChIP-seq data for histone marks. Specifically, H3K4me3, H3K4me1, H3K27ac, H3K36me3, H3K27me3, H3K9me3 are often assayed together for samples as they constitute a reference epigenome[66]. The reference epigenome defines that group of histones which can together describe the overall epigenetic state of the genome fairly well. The various combinations of histone and modifications form something akin to a "histone code"[64]. This histone code can be interpreted to determine the location of promoters, enhancers, gene activation and the status of other gene regulatory elements. Generally, H3K27ac is in enhancer and promoter regions, H3K4me1 is an enhancer mark, H3K4me3 is in active promoter regions, H3K36me3 is often found in gene bodies, and associated with gene splicing[67], H3K27me3 has a repressive role in promoter regions and H3K9me3 is mostly found in repressed regions[68].

Histone mark	Function and/or location
H3K27ac	Enhancer and promoter region
H3K4me1	Enhancer
H3K4me3	Promoter region; poised state
H3K36me3	Gene bodies
H3K27me3	Repression in promoter region; poised state
H3K9me3	Repression

Table 1. Summary of histone marks function and location within genome

Bernstein et al. added complexity to the histone interaction by showing regions with H3K4me3 and H3K27me3, an active and repressing histone, being found together[69]. These regions, termed bivalent, were particularly found in ES cells and would resolve through differentiation, providing a new "poised" state and a model of for an initial histone state and how pluripotency works at the histone level[69–71]. This highlight the overall importance of histones as they are dynamic, they can vary between cell type and condition, and can give an idea of how the genes from the fixed sequence will be expressed. Although histones have all this dynamic potential, they tend to be rather consistent within a cell type. Thanks to this, models were developed to impute histone mark tracks by using information from the other histones[72]. Imputed samples generally accurately predict the histone tracks but they display less variability.

One notable feature of histones is that they are inherited by progeny even if they although they are not part of the DNA sequence. It is commonly understood that histone positions are inherited by progeny even if the how remains unclear. It has been reported that histone are indeed conserved from parental inheritance[73], at least in yeast, but the mechanism remains unknown as it does not seem like the histone modifications alone are sufficient for inheritance[71,74].

1.2.3 DNA Methylation

DNA methylation is another major component that controls gene regulation, chromatin organization and thus cell identity and cell development. DNA methylation has a repressive role and is typically found on CpG dinucleotides. DNA methylation is established by DNA methyltransferases (DNMTs) which transfers a methyl group to the C'5 and prevents proper binding of transcriptional regulators[75]. DNA methylation is generally measured through whole genome bisulfite sequencing (WGBS) and often reported as differentially methylated regions (DMRs), regions with distinct methylation between samples.

Histones and methylation has some interplay, within normal cell, H3K4me3 blocks DNMTs while H3K9me3 and H3K27me3 recruit it. Thus the repressive methylated histone marks may also lead to DNA methylation. Cancer can impact things, it can disrupt or remove H3K4me3 enabling DNMTs, lead to the substitution of H3K4me3 by a repressive or other mark, or simply lead to the loss of all histones[75]. In the context of repressive H3K9 and H3K27, the relevant lysine methyltransferases, SUV39H1/2 and EZH2, respectively, interact with the DNA methylation DNMTs[75,76]. DNA methylation also contributes to chromatin formation through its interaction with the other epigenome elements such as histones and polycomb complexes.

In mammals, most CGs are methylated but those in promoters tend to be protected from methylation. Hypomethylation, which is losing methylation, is a hallmark of cancer as it disrupts the normal expression patterns[75]. DNA methylation is crucial to methylation maintenance as it is present on both strands and can thus be recovered when DNA copies are made (from the copied strand still bearing the methylation). DNA methylation is closely associated with cell type differentiation, methylation tends to be absent for specific cell types[77]. It was found that enhancers are generally demethylated and that each cell type has regions that are uniquely demethylated

compared to others[78]. Furthermore, the unmethylated site can have motifs relevant to cell type regulators[78] which makes everything come together. Unmethylated sites are accessible to binding factors and serve as enhancers to specific cell type genes and regulators since those demethylated regions are unique to select cell types.

1.2.4 Chromatin accessibility and states

One of the first component of the epigenome to be discovered would be chromatin and the distinction between euchromatin and heterochromatin, states that can actually be visually observed. As briefly alluded to, chromatin is the physiological form the genome and it was found that it can be found in two state: Euchromatin where the DNA is decondensed and is accessible, and heterochromatin, where the DNA is condensed, less accessible and genes are rarer[64].

Chromatin state can go beyond this binary state through profiling of the other epigenome elements (histones, DNA methylation, etc.) to detect regulatory elements. In particular, the combination of histone modifications can provide even more insight. Ernst et al. investigated the recurring patterns of chromatin mark combinations to define a model of 15 finer chromatin states called chromHMM[68,79]. These include: repressed, poised and active promoters, strong and weak enhancers, putative insulators, transcribed regions and large scale repressed and inactive domains[68]. These states sacrifice the direct measurement nature of histone ChIP-seq for a prediction of a more interpretable chromatin state. The chromHMM chromatin states are increasingly used in the field, but their interpretability and the ability to compare results across experiments will always depends on the model and underlying data used. Nonetheless, the predicted states and inferred cis regulatory make for testable predictions many of which were confirmed which supports the value of this expanded set of states.

1.3 The Interplay Between TEs and The Epigenome

Since TEs are genomic elements with dynamics controlled in part by the epigenome and in reverse, the epigenome can be affected by TE insertions and expression, there is potential for complex interplay between the two. And it seems that this relationship between TE and the epigenome has left quite an impact on human evolution. Even with assessment of chromatin accessibility (Open chromatin with DNAse 1 hypersensitivity), it was found that the majority of primate-specific regulatory sequences were derived by TEs[80]. Specifically, they found that 63% of DNase I hypersensitive sites (DHS) regions overlapped TEs in primate specific sequences, that 36% of the DHS associated TEs were statistically enriched for at least one TF and that TE sequences were active in cell type specific manner[80].

Generally, epigenetic repressors are used to repress TEs. In normal cells, TEs are methylated and thus silenced, but when DNA methylation is absent, TEs can transpose which can lead to genomic instability[81,82]. For instance, in brain cancer it was found that Alu levels of methylations decreased compared to normal cells[83]. It was also found that colon cancer cells were 10 times more unmethylated than normal cells[84]. Thus TEs seem to be less methylated in cancer cells. Another context in which there is loss of methylation is primordial germ cells (PGCs) and early embryo as it is essential for a return to pluripotent state[85].

In a more recent study of chromHMM epigenomic states relationship with TEs[86], it was found that TEs encompass a quarter of the regulatory epigenome. It was also found that 47% of elements could be found in active regulatory state, that SINEs were enriched for active marks and that TEs overlapped the heterochromatin epigenetic state the most out of all 15 states. The study also found

that SINE elements with CpG islands are prone to more DNA methylation but that the CpGs islands are lost from TEs as they age.

Zhuo et al. reports clear signal of TE-derived tissue specific putative enhancers and promoters in human and chimpanzee[87]. They identified LTR5 as putative promoters in IPSCs although that TE had also been reported as an enhancer in human embryonal carcinoma cells by Fuentes et al[88]. They also found NR2F1 binding to be correlated with enhancer signatures on the 3' end of SVA elements and report limited TE associated heterochromatin spread[87].

Thus it is clear that TEs, which are not expressed much but spread all across the genome, should be studied further. Especially their relationship with the epigenome which is likely much more complex than the repression it is known for. There seems to be a complimentary or synergetic potential between TEs and the epigenome. After all, if TE insertions can change the regulatory transcriptome, they can disrupt enhancers and promoters and thus the chromatin state of a region which could have its own cascade of effect.

1.4 Genomic Tools and datasets

1.4.1 TE Analysis Limitations

While extensive study of TEs may seem like something that should have been done already, TE's repetitive nature has unfortunately led them to a long neglect. Indeed, repetition of sequence in the genome leads to numerous challenges within most genomic analysis and methods. When it comes to genomic experiments we often target specific genes sites or targets but when studying a TE with extremely similar instances throughout the genome, targeting specific instances is not trivial[33]. These reads or sequences that cannot be placed in a unique location, often called *multi-mapped*

reads, are ambiguous and prone to biases. For these reasons, TE are often ignored or masked within genomic experiments[89,90]. As technologies improves, a lot of the current challenges can potentially be resolved. The increasing of sequencing length through long read sequencing is promising for the proper assessment and positioning of TEs. However, for some widely used experiments such as ChIP-seq, using smaller reads, resolving *multi-mapped* reads or clever TE measurements methods will remain a challenge worth tackling.

Usually, to be able to account for ambiguous reads, TE approaches will use methods prior to alignments where they can try to better place the reads or directly count the TE instances[89,90]. For analysis with processed genomic file formats, which public datasets usually make available, such as BEDs or Peak files, the approach is usually to intersect the samples peaks with RepeatMasker's[15] repeat positions.

1.4.2 TE Hub

To tackle the challenges of TE analysis, members of the TE community established TE Hub[91], an open and collaborative platform providing a reference point for all transposable elements methods and resources. TE Hub presents over 100 TE analysis methods, teaching resources and databases to help researchers get started on studying TEs. The methods are numerous because some focus on single TE families or specific solutions to specific problems. Creative heuristics are often needed to estimate TE measurements, which can lead to different coding languages and large dependencies. Thus, it is worth noting that it can often be complicated to install the tools, which are sometimes resource heavy and may need to be run on remote servers. While data sharing has made tremendous advances, methods and tools sharing can remain challenging due to external dependencies, different machines and the need for maintenance. Nonetheless, TE hub is a great resource for TE tools and a testament to the growing interest on TEs.

1.4.3 The UCSC Repeat (and Genome) Browser

The UCSC genome browser[92] is probably one of the most useful tool for the study and visualization of the genome. The UCSC browser makes available human (although other organisms are available) genome tracks with a large number of annotation and additional tracks. Among the tracks, the Repeat Masker[15] track, allows to visualize TE positions across the genome side by side with other annotations. Although the Repeat Masker track in the browser is useful, it is hard to glance at the overall general sequence and the environment of the elements as instances are spread across the genome, each in their own context and with their own sequence.

Fernandes et al. reversed the emphasis with the UCSC Repeat browser[93] which consists of a complete set of repeat reference sequence from RepeatMasker. The UCSC Repeat browser displays a consensus sequence of the TE of interest and alignments of the various instances across the genome to that consensus. It also provides processed tracks of some publicly available datasets of interest such as ChIP-seq data. Thus, the browser enables intuitive visualization of genomic data on TE consensus sequences. Since the repeat browser is online, there is no installation process and it works on any machine, furthermore it leverages the already familiar interface of the UCSC genome browser making it a straightforward and intuitive tool for TE investigation. Paradoxically, one of its limitation would be that although its interface and branding are closely associated to the UCSC genome browser, the repeat browser is actually a rather independent and standalone tool that does not have all the genome browser annotations and tracks available.

1.4.4 The International Human Epigenome Consortium and its Portal

The developments and improvements in sequencing technologies have allowed a large amount of genomic data to be generated. With the interest in the epigenome rising, many groups and

consortiums have generated large epigenome datasets. The international human epigenomce consortium (IHEC) was formed with the overall objective to understand the extent to which the epigenome has shaped human evolution. It is a consortium of international consortium with data contributed by 7 members: ENCODE, NIH Roadmap, CEEHRC, Blueprint, DEEP, AMED-CREST, and KNIH. Some of its goals are to coordinate the distribution of data across the research community with minimal restriction, coordinate development of bioinformatics standards and analytical tools to organize and display the epigenomic data generated by the consortium.

Perhaps as a realization of its goal, IHEC has an online portal[94], that provides access to over 7,000 reference epigenome datasets generated from over 600 tissues. The portal facilitates discovery through an intuitive grid interface, data visualization, straightforward sharing and connectivity to the UCSC browser for further analysis.

IHEC is currently preparing a new dataset, the EpiATLAS, which is a uniform reprocessing of samples from its many consortiums now including EpiHK and GIS in addition to the former 7 members. This dataset includes 5473 histone ChIP-seq samples, 645 WGBS DNA methylation samples and 1555 RNAseq samples which were grouped into 47 cell categories through cell labels harmonization. The data also has extensive metadata data such as health status, donor life stage, donor age, phenotypes and sex. This makes EpiATLAS one of the largest, most varied and promising uniformly processed dataset to date.

1.4.5 Other online datasets and tools

But making data accessible is not the only way to support reuse of scholarly data. Online visualization of results and online analysis are a great way to spread knowledge and encourage method adoption. For instance, the Genotype-Tissue Expression (GTEX) project makes available

genetic associations and gene expression and splicing in 838 individuals over 49 tissues [95–97]. The data is presented as dynamic plots which make it easy to interpret data and judge it before downloading or incorporating it in an experiment. The GTEX project has went on to increase its dataset and the experiments it covers. This is another advantage of browser tools; they can update while always being up to date for all users. However, this can lead to reproducibility issues so it is important to keep versioning and wise to push discrete updates.

Chip-Atlas is a data-mining suite for exploring epigenomic landscapes by fully integrating 419,000 ChIP-seq, ATAC-seq and Bisulfite-seq experiments[98]. Thus it has a large dataset (mining data from various sources, including non-human organism), but it also allows running experiments directly on the website. Although the data navigation and discovery is not as straightforward as the IHEC portal, the analysis and data download of their vast dataset is a great use case.

Another particularly relevant tool is the Genomic Regions Enrichment of Annotations Tool (GREAT)[99,100]. GREAT is a tool that predicts functions of cis-regulatory regions. While genes are usually well annotated with their biological function, non-coding regions are often devoid of such descriptions. GREAT infers the biological meaning of a set of genomic regions by analyzing the annotations of nearby genes. The tool is available online as a web page which takes BED file (or plain text) as input and outputs gene ontologies terms as well as closely associated genes relevant to the input regions. The GREAT result benefit from their webpage nature through connectivity with and a website like page linkage structure. GREAT served over 1 million jobs as of 2018 and has probably only gotten more popular since with its latest in 2022[100].

In brief, TEs are challenging to assess due to their repetitiveness and one of the ways to surmount that challenge are complex tools and methods, which can be hard to adopt. In parallel, there is a growing amount of epigenome data being made available and visual methods to select and observe

the data greatly contribute to accessibility and adoption. While there are some online TE browsing tools, it seems to be an area with a lot of untapped potential.

1.5 Hypothesis and Objectives

It is clear that TEs are an under-explored component of the human genome with high regulatory potential and impact on human health and disease. Due to their repression, TEs are not expressed much, however, due to their long history within the human genome, they also had the opportunity to integrate the genome and have a regulatory impact through the epigenome. The epigenome regulates gene expression across cells, independently from sequence, through changes in chromatin state, DNA methylation and histone marks. Together, TE and the epigenome appear promising for a larger regulatory and disease impact.

We hypothesize that TEs have unaccounted regulatory function, are critical to our proper understanding of regulation and disease and that their impact is associated with the epigenome. To support our hypothesis, we first made a comprehensive analysis of TEs overlap and enrichment across the cell types and histone marks within a new IHEC dataset of over 4000 uniformly reprocessed ChIP-seq samples. Next, we wanted to make sure that the resulting data from our analysis would be used beyond our observations and that our approach was in a privileged position to be widely used. We inspired ourselves by some of the various online genome tools to build a webtool that reports the results of our analysis in an intuitive and accessible manner and enables the analysis of TEs within user uploaded data. Finally, while we mostly investigated the relationship between TE and the epigenome through histones, we used a similar approach for some perspective on the relationship between TEs and DNA methylation.

Chapter 2

Large Scale Analysis of Transposable Elements Interaction with the Epigenome

Preface: bridging text between chapter 1 and chapter 2

Having introduced our main topics of interest, transposable elements and the epigenome, and established our hypothesis of there being some unaccounted for TE modulated genome regulatory function in chapter 1, we designed an experiment for a broad survey TEs and the epigenome.

In this chapter we leverage the EpiATLAS, a new large epigenome dataset from IHEC containing 4867 samples, including 47 cell types, 6 histone marks and healthy, disease and cancer cells to characterize the relationship between TE, cell types and the histone marks. We first broadly characterized the overlap between TEs and the histone marks and looked at the enrichments by comparing the overlaps to those from TSS distribution matching simulated controls. We then dived into TE subfamilies. Since the majority of subfamilies were depleted we focused on a subset of the most enriched subfamilies across all samples. For a full perspective on the TEs and their associations with the epigenome, we looked at association with TE age, cell types, cell type specificity and health status. Due to the large number of combinations cell types, TE families and histones, we found an overwhelming number of associations. We devised a candidate selection approach to highlight some of the most extreme and cell type specific cell types-histone-TE

associations and report them. Our hope is that some of these candidates will be TEs with regulatory function related to the cell type they are associated with.

This study was published as a preprint in biorxiv and is meant to be part of a collection of publications associated with the release of the EpiATLAS dataset.

Appendix B contains supplementary figures and tables descriptions.

Transposable elements impact the human regulatory landscape through cell type specific epigenomic associations

Jeffrey Hyacinthe[1], Guillaume Bourque[2,3,4,*]

[1] Quantitative Life Sciences, McGill University, Montréal, QC, Canada

[2] Department of Human Genetics, McGill University, Montréal, QC, Canada

[3] Canadian Center for Computational Genomics, McGill University, Montréal, QC, Canada

[4] Victor Phillip Dahdaleh Institute of Genomic Medicine at McGill University, Montréal, QC, Canada

[*] Correspondence: guil.bourque@mcgill.ca

2.1 Abstract

Transposable elements (TEs) are DNA sequences able to create copies of themselves within the genome. Despite their limited expression due to silencing, TEs still manage to impact the host genome. For instance, some TEs have been shown to act as cis-regulatory elements and be co-opted in the human genome. This highlights that the contributions of TEs to the host might come from their relationship with the epigenome rather than their expression. However, a systematic analysis that relates TEs in the human genome directly with chromatin histone marks across distinct cell types remains lacking. Here we leverage a new dataset from the International Human Epigenome Consortium with 4867 uniformly processed ChIP-seq experiments for 6 histone marks across 175 annotated cell labels and show that TEs have drastically different enrichments levels across marks. Overall, we find that TEs are generally depleted in H3K9me3 histone modification, except for L1s, while MIRs were highly enriched in H3K4me1, H3K27ac and H3K27me3 and Alus were enriched in H3K36me3. Furthermore, we present a generalised profile of the relationship between TEs enrichment and TE age which reveals a few TE families (Alu, MIR, L2) as diverging from expected dynamics. We also find significant differences in TE enrichment between cell types and that in 20% of the cases, these enrichments were cell-type specific. Moreover, we report that at least 4% of cell types-histone-TE combinations featured significant differences in enrichment between healthy and cancer samples. Notably, we identify 456 cell type-histone-TE triplets with strong cell-type specific enrichments. We show that many of these triplets are associated with relevant biological processes and genes expressed in the relevant cell type. These results further support a role for TE in genome regulation and highlight novel associations between TEs and histone marks across cell types.

2.2 Introduction

Transposable elements (TEs) are DNA sequences with the ability to duplicate themselves within the genome. This transposition is done through 2 main mechanisms which define TE classes. Retrotransposons are sequences that use an RNA intermediate which reverse transcribes back into DNA and integrates in another genomic location akin to a copy and paste approach; while DNA transposons excise themselves before inserting elsewhere using a cut and paste mechanism[101,102]. Within classes, TEs are grouped into families, which capture the elements origin and history, and also into subfamilies representing more closely related elements and finer divergence within those families[19,103,104]. Through their replication, TEs have proliferated through genomes and currently cover at least 50% of the human genome[15].

One reason why TEs have gathered increased attention is due to their involvement in host gene regulation. Indeed, while most TEs in the human genome have lost their transposition ability, some TEs have been found to be associated with enhancers[105]. For instance, they were shown to be associated with the core regulatory network of human embryonic stem cells[4,106] and some TEs have been co-opted into the human genome such as an ERV by the Aim2 gene[107]. Furthermore, about 750 SVAs were found to act as enhancers or promoters modulating gene expressions in pluripotent cells[6].

A recent investigation with 19 different cell types showed that TE encompassed one quarter of the regulatory epigenome and that 47% of TE instances could be found in an active state[86]. This study,

primarily centered on epigenetic states, highlighted that cell type specific TE associations could be detected. In other studies, the histone mark H3K9me3, a hallmark of heterochromatin, has been shown to be associated with repeat elements[108,109]. More targeted investigations of specific TEs such as Alus or L1 and histone marks have demonstrated that TEs can behave as enhancers[105,110] and in a cell-type specific manner[111]. These results demonstrate that cell type differentiation and function may be partly regulated by TEs. However, a comprehensive analysis that relates multiple TEs with reference histone marks across distinct cell types is lacking. Thus, we set out to investigate TE's associations with histone marks to better understand TEs and their relationship with the epigenome.

With the generation of a large epigenome dataset from the International Human Epigenome Consortium (IHEC)[94], it is now possible to investigate all TEs in the human genome, across reference histone marks in various cell types. The latest IHEC reprocessing EpiATLAS dataset makes available 4867 Chromatin Immunoprecipitation sequencing (ChIP-Seq) datasets from 6 histone marks (H3K4me1, H3K36me3, H3K9me3, H3K27me3, H3K4me3 and H3K27ac), from 175 annotated cell labels which were grouped into 47 cell types[112]. This dataset contains novel cells types for TE investigation, such as lymphocytes of B lineage and thyroid, has more samples of tissues previously reported to be associated with TEs such as twice the brain sample count from NIH Roadmap[66] , has a larger set of repressive mark data to contrast activating marks, contains healthy and disease samples from the same tissue and includes many replicates to characterise the variability of our observations.

Here, we present an overview of TE overlap found in the EpiATLAS dataset and a comprehensive map of TE subfamily enrichment across cell type and histone marks. We investigated the relationship between TE enrichment and TE age, we surveyed the TE enrichment across cell types and whether they changed depending on health status. We measured the cell type specificity of the TE enrichments and identified TE-cell type regulatory candidates in terms of specific and extreme associations. Some candidates were associated with relevant biological processes and reports of genes expression in the given cell type. Taken together, these results provide a consensus resource of the TE profile across cell types and histone marks.

2.3 Results

A large and varied comprehensive inter-consortium dataset

Our analysis leveraged the EpiATLAS dataset, a large uniformly processed dataset generated by a consortium of consortiums[94,112]. Specifically, we obtained 4867 ChIP-seq samples for 6 histone marks (H3K4me1, H3K36me3, H3K9me3, H3K27me3, H3K4me3 and H3K27ac), coming from 7 consortiums (Blueprint, CEEHRC, DEEP, AMED-CREST, NIH Roadmap Epigenomics, GIS, EpiHK) and prepared by the IHEC EpiATLAS integrative analysis group (Methods). Of the consortiums from which the data originated Blueprint & CEEHRC accounted for more than 2/3 of the samples (Fig 1A). Our dataset consisted of 175 different cell labels which were grouped into 47 broad cell types (Fig 1B). Compared to the NIH Roadmap reference epigenomes[66], this represents more than five times the number of samples (4867 vs 733), including 6 times the brain samples (476 vs 72) and introduced sample in novel cell types such as various lymphocytes (373)

(Fig 1B). It is also more than 3 times the number of samples from the more recent ENCODE epigenome dataset[113] (See Methods). In IHEC, each cell types contained on average 104 samples with a mean of 18 biological or technical replicates per assay. While most samples did not have all 6 histone marks, they were represented in comparable amounts (Fig 1C) with a larger number coming from H3K27ac (1481). Notably, the EpiATLAS dataset also included samples with different health conditions: 3557 healthy samples, 1007 cancer samples and 303 diseased samples (Fig 1D).

To make sure that the samples obtained from different consortiums were of good quality and comparable, we looked at their distribution of peaks relative to genes transcription start sites (TSS) and found that it remained consistent within assay across consortium (Supplemental Fig 1A). For instance, we find that 84.7% of H3K36me3 peaks were in intragenic regions across all consortiums. This was much lower for H3K9me3 samples at around 33.4%. The distinct distributions between assays combined with consistency across consortium indicated the data captured the similarly located regions across samples. Of the 4867 samples only 253 samples were deemed outliers based on peak distribution and were discarded (Supplemental Fig 1B, Methods). We visualised the 4614 retained samples through a UMAP dimension reduction on the peak counts within genome windows of 10 kilo base pairs (kbp) and found that, as expected, the assay type showed strongest clusters (Fig 1E, Methods). Within the assays, the consortium mostly mixed, but there were some consortium-specific clusters suggesting either cell-type or sample preparation effects (Supplemental Fig 2A). We also note that at that level, cell type associations were not clear, except within some H3K27ac and H3K4me1 datasets (Fig 1E-F, Supplemental Fig 2E). We used PCA as an alternative data visualization method and also found that assays form clusters based on

first 2 PCs, while cell types and consortium do not (Supplemental Fig 2B-C,F). The assay clusters are still visible between PC 2 and 3 (Supplemental Fig 2D).

Thus we make use of a more comprehensive dataset, which includes many replicates, novel healthy and disease tissues and is consistent even across consortiums, enabling our TE analysis to explore new associations and patterns.

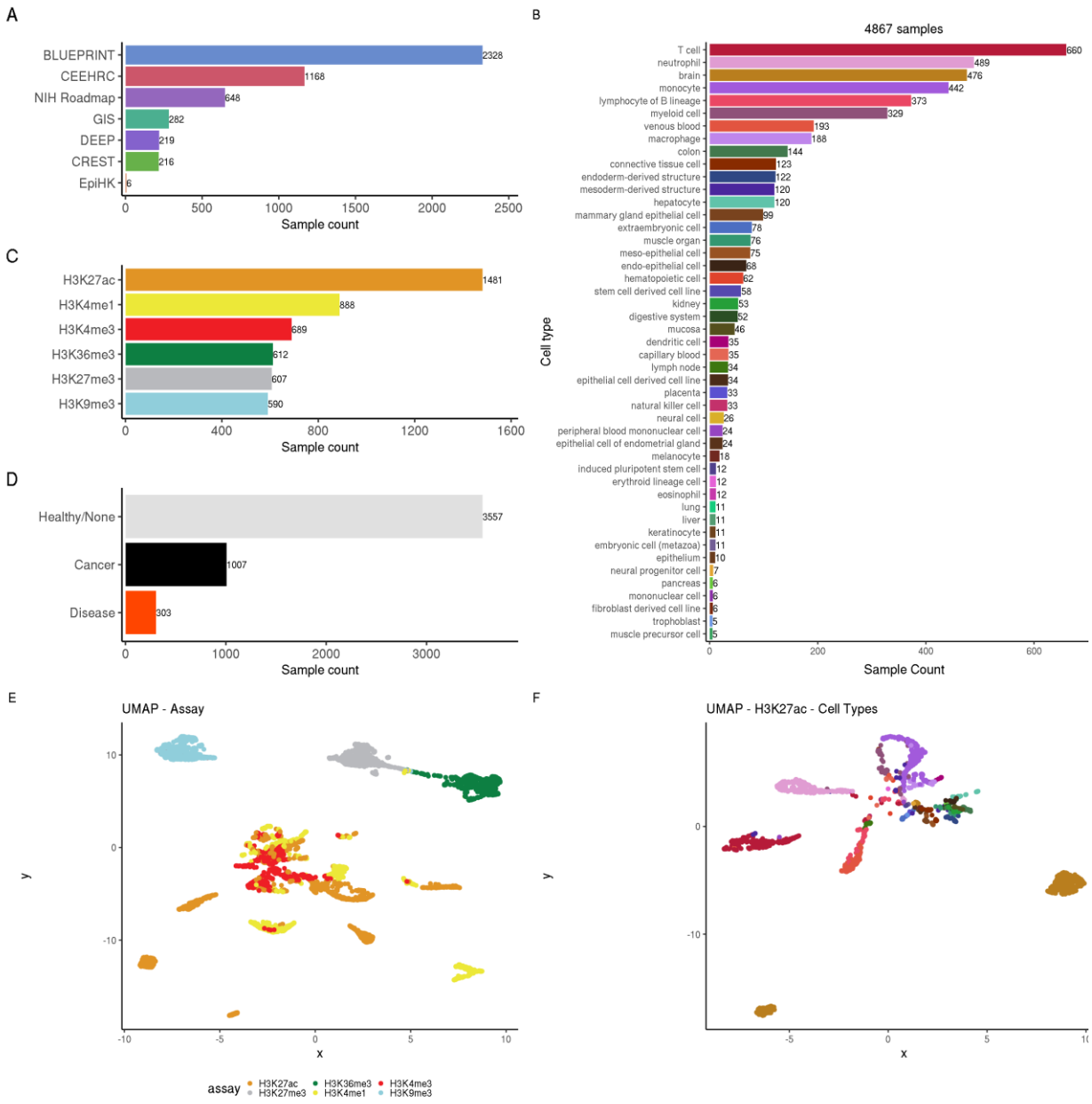

Figure 2.1 An expansive dataset obtained from IHEC.
A| Sample count by consortium **B**| Sample count for each cell type **C**| Proportion of samples in from the various assays. **D**|By health status **E**| UMAP on peak counts within 10kb regions across the genome for the 4614 samples. Only the 20,000 windows with most variance (excluding the first 1000) were used. Color represents the assay. **F**| Same as E but only for H3K27ac with cell type coloring from B.

TEs have distinct epigenetic marks association profiles

Next, to characterize the contribution of TEs to ChIP-seq peaks, we measured the percentage of peaks within each sample that overlapped such elements. We found that the overlap with TE families remained largely consistent within assays (Fig 2A). On average, 55.8% of peaks were found to overlap with TEs and we observed that the activating mark H3K4me1 (54.2%) and H3K27ac (50.7%) displayed more similar TE overlap profiles. In contrast, the repressing mark H3K9me3 had a distinctively higher TE overlap (78.3%), which is consistent with the reported role of H3K9me3 in TE repression[109,114,115]. We also found that most of the TE overlap came from the top 5 most common families (Alu, L1, MIR, L2, ERVL-MaLR, Supplemental Fig 3A). TEs can cause multi-mapped reads which can lead to analysis issues. To make sure that mappability was not impacting these observations, we used the mappability track from the UCSC Genome browser[116] (Methods). We found that mappability did not correlate with the overlap we observed and only Telo, SVA, Satellite, Centr and acro families featured a median mappability below 80% (Supplemental Fig 3A-B, 4). Next, we explored the cumulative and mean TE Length and the TE instance count and found that the TE overlap profile was most similar to the instance count (Supplemental Fig 3A, C-E).

To better capture TE enrichment, we identified TE families that featured significantly higher overlap (p-val < 0.005) than in distribution matched random simulations (Methods). Using the difference between the observed and the expected (obs-exp, which we called an enrichment), we noted distinct patterns such as an enrichment of L1 and ERV TEs in H3K9me3 and Alu in H3K36me3 (Fig 2B). The low enrichment also exposed that while there was high overlap with TE

families in the genome, most of it could be attributed to chance and that significant TE associations arose from a minority of regions.

Next, to detect more specific associations, we repeated this enrichment analysis using TE subfamilies instead of families. We found that the majority of TE subfamilies were not significantly enriched (Supplemental Fig 5A) and decided to focused on the most globally enriched subfamilies. This was done by adding up all significant subfamily enrichment across all samples and selecting the 164 subfamilies with the highest contribution (Supplemental Fig 5B, Methods). We then built a heatmap providing a comprehensive overview of TE enrichment for these 164 subfamilies across 4614 samples and six histone marks (Fig 2C). We observed that TE subfamilies within a family tend to feature similar enrichments, for instance, the enrichment in H3K9me3 appears to be generalised across many L1 subfamilies. The same is true for most family clusters. TE families also seem to have distinct histone mark preferences. For instance, within the top 20 most enriched subfamilies, H3K4me1, H3K27ac & H3K27me3 share a similar profile of MIR and L2 subfamilies while H3K36me3 features Alu and H3K9me3 L1 andd ERVL-MaLR (Supplemental Fig 5C).

This perspective further supports the antagonism between activating and repressing mark TE enrichments and highlighted several enrichment clusters (Fig 2C). The enrichment of L1 in H3K9me3 while it is depleted everywhere else may be linked to previously reported TE repressing mechanisms[49,117]. A similarly high association can be found with ERV elements, another family noted for being repressed[42,44,118]. The enrichment of Alu in H3K36me3 stands out and is further

investigated in the following sections. We also notice that MIR is highly enriched in marks that are depleted for the other TE families (H3K4me1, H3K27ac, H3K4me3), which may be due to their association with enhancers[110]. We found a widespread enrichment of ERVs across the different histones albeit at lower levels and less consistently. Since ERVs subfamilies are small and might not be the most enriched based on our obs–exp metric, we also explored enrichment based on fold change (observed / expected). We found that ERVs, were the most enriched according to fold change while Alu, L1 and MIR elements were most enriched according obs-exp (Supplemental Fig 6).

Taken together, these results suggest that TE families seem to have distinct preferences for histone mark enrichment and shows the complex, yet conserved relationship between TE and the epigenome.

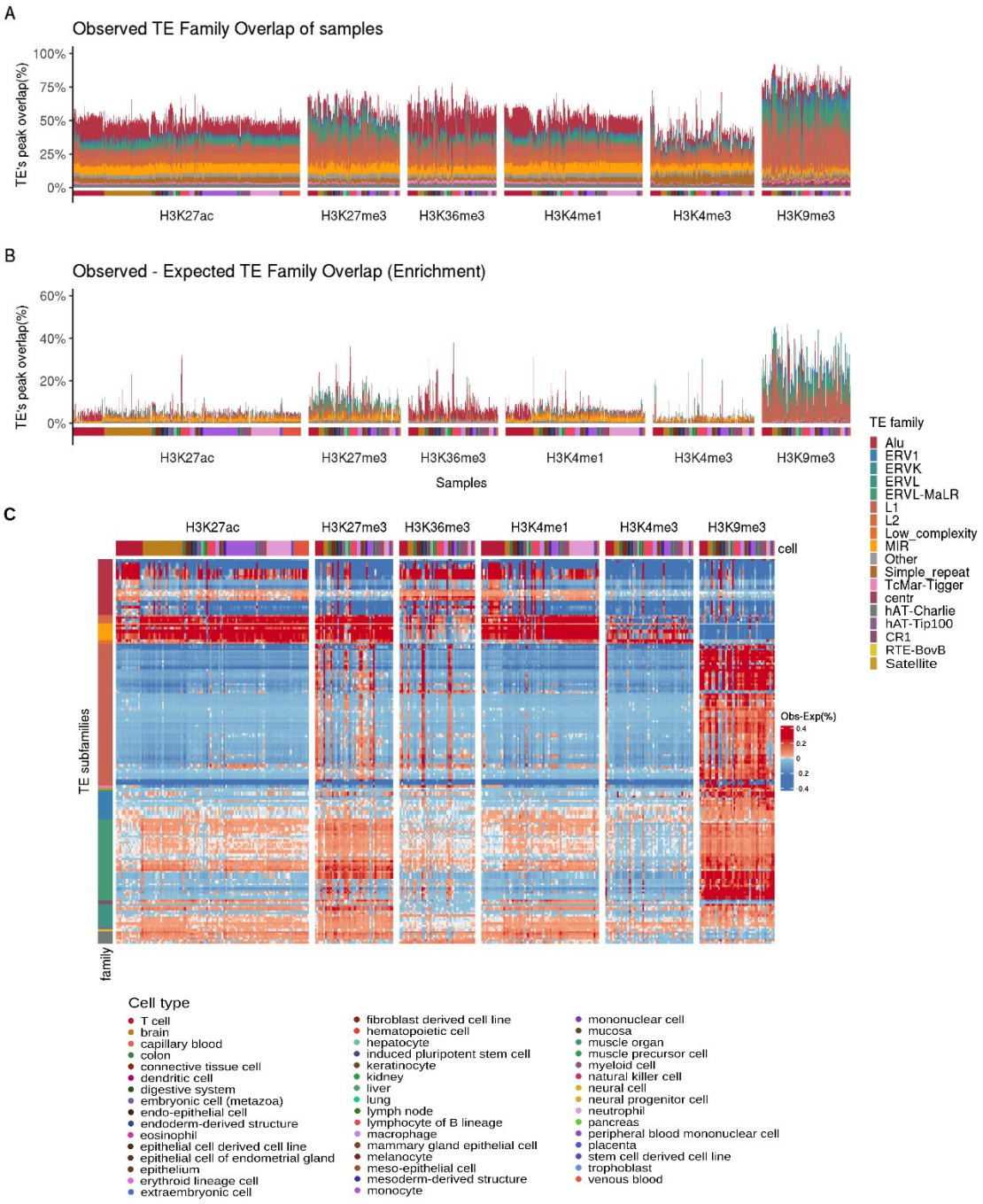

Figure 2.2 TE displays distinct association profiles with histone modifications

TE displays distinct association profiles with histone modifications. **A|** Percentage of peaks overlapping TEs colored by TE family for the 4614 samples. Samples are annotated by cell type. **B|** Difference between the Observed overlap (A) and the simulated background overlap for using only significantly enriched (p-val < 0.005) TEs. Same annotations as A. **C|** Heatmap of the TE subfamily enrichment of the selected 164 subfamilies for the 4614 samples. Enrichment (obs-exp in percent, including non-significantly enriched TEs). TE subfamilies are grouped by family (x axis), Samples are grouped by Assay (y axis) and annotated for cell type. The samples within groups are hierarchically clustered.

The different TE families display different enrichment dynamics as they age

Since older TE instances are more likely to have been either degraded by mutations or co-opted by the host genome[119], we expect TEs epigenetic profile to change based on their age. It was reported in Su et al.[105], that older Alus had a profile more like enhancers in terms of epigenetic state, conservation and TF binding potential. Pehrsson et al.[86] also showed that DNA methylation of Alu went down with age. To determine if the age relationship was something that could be generalised across all TE families, we performed linear regressions of TE enrichment (mean across cell types for each subfamily) as a function of TE age (Fig 3A-B, Supplemental Fig 7). We observed that within some families (L1, ERVL-MaLR), many of the younger subfamilies were depleted in some mark (H3K27me3, H3K36me3) while enriched in others (H3K9me3) (Fig 3A). At the same time, for L1 and ERVL-MaLR, we found a general pattern of TEs within a family approaching 0% enrichment as they age (0% enrichment being the equivalent of being near the random simulations). This is consistent with TEs being degraded into background as they age[120], which appears to be true whether they show an enrichment when they are young (H3K9me3) or a depletion (H3K27me3, H3K36me3).

While we found that some TE families tend towards no enrichment as they age, we also found a few families that diverged from the expected level. Alus show a distinct decline in enrichment in H3K4me3, H3K9me3 and H3K27me3 while for H3K36me3 it was an increase (Fig 3B, Supplemental Fig 7B). There are also many cases where there were little to no differences as they aged, such as for ERVK, ERVL and ERV1 (Supplemental Fig 7C). This exposes that the

relationship between TE age and enrichment in the epigenome is complex and varies depending on family. We also looked at subfamily average length and instance count as potential cofounders and found that there were some correlations with TE age (Supplemental Fig 8). This means that those properties could also be associated with the observed TE enrichment evolution over time. In addition, since younger TEs tend to be harder to characterize due to mappability issues (less time for mutations to generate unique reads), we investigated the mappability and TE age correlation and only found it to be noticeable within L1 and Alu (Supplemental Fig 9A). However, we note that the less mappable TEs were not necessarily less enriched (Supplemental Fig 9 B,C) and thus, unlikely to be the cause of the observed enrichments.

Overall, two contrasting age dynamics were found: some TE families become more enriched or depleted as they age and others tend towards the expected background. To better characterise these dynamics, we measured the absolute enrichment of old and young TEs. We split the top 50% oldest and youngest TE subfamilies per group (Methods). We observed once again that some families had diverged from expectation as they aged (Alu, L2, MIR) while others converged to expectation (L1, ERVL-MaLR) (Fig 3C). These results are also summarized with more TE families in Fig 3D.

Taken together, these results highlight the complex TE enrichment association with histone marks that evolved differently depending on the TE family.

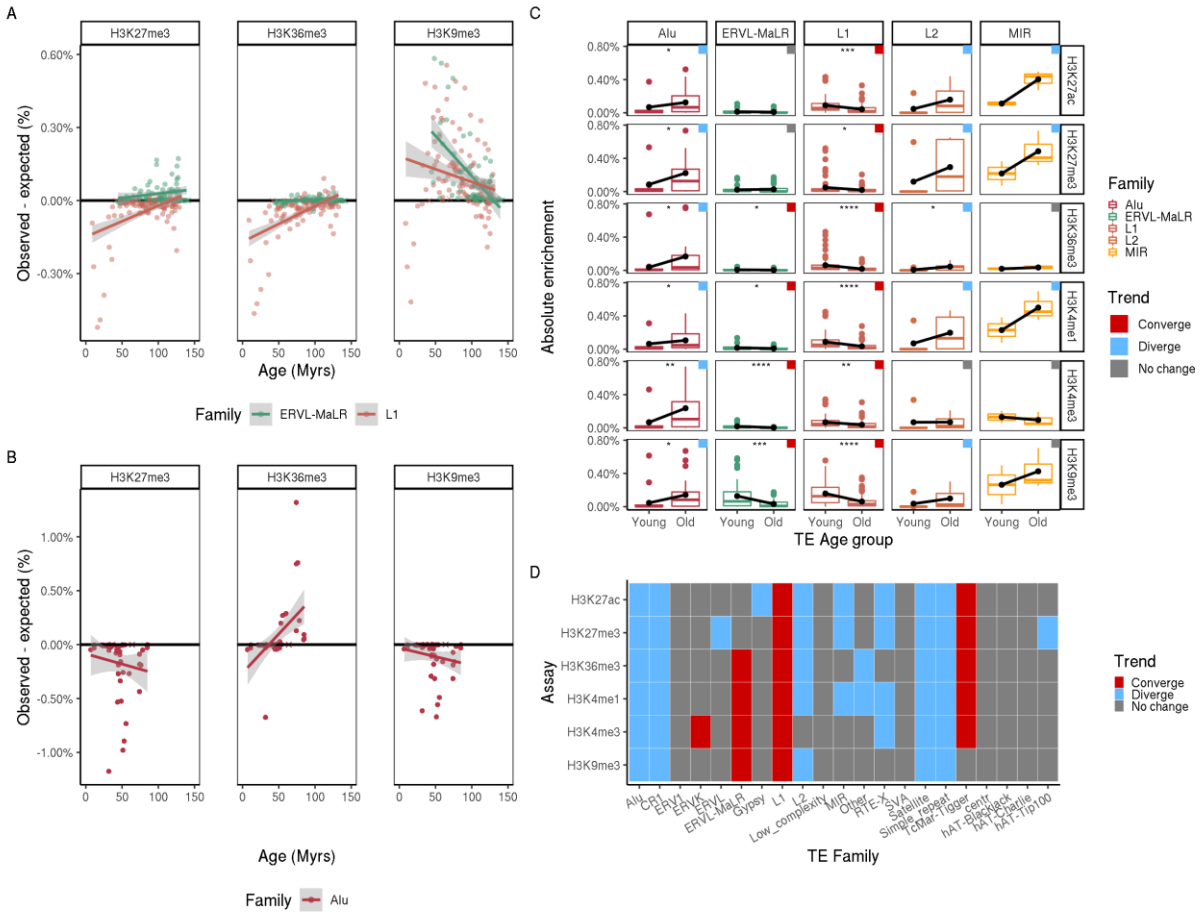

Figure 2.3 TE enrichment follows a family and context dependent age continuum.
A| TE Family enrichment (obs-exp%) in function of estimated age for L1 and ERVL-MaLR in H3K27me3, H3K9me3 and H3K36me3. Each point is the mean enrichment (obs-exp%) of a TE subfamily across all samples. Black line is at 0% enrichment. Line shows linear regression fit, crosses are small sized subfamilies excluded from regression., **B|** Same as A for Alu TE family. **C|** Absolute enrichment (distance from 0 obs-exp% enrichment) of Young and Old TE subfamilies. Black line connects the means of age groups. (*: p <= 0.05,**: p <= 0.01,***: p <= 0.001,****: p <= 0.0001) Color in corner represents the trend. **D|** Dynamic of TE enrichment between young and old. Categorises the observations of **C** by dynamics and includes additional families. No change: no significant change, Diverge: Older TE have higher Absolute enrichment, Converge: Older TE have lower Absolute enrichment thus converge to 0.

TE enrichment varies between cell types and histone mark contexts

Next, we were interested in whether TEs could be involved in distinct cellular profiles and we measured the association of enrichment with cell types. Since the TE enrichment was primarily associated with the assay, we first grouped all samples per assay and then performed t-tests of cell type's TE enrichment (significant only) against the mean for each histone mark and TE family pairs and sorted the cell types (Methods, Supplemental Table 1, 2).

For instance, looking at Alus within H3K36me3 due to their unexpected enrichment (Fig 2C), one notable cell type was the colon, which were significantly more enriched than the mean (10.9% average obs-exp versus 5.5% Fig 4A). We also observed high variability in some tissue's enrichment. Indeed, looking at brain, lymphocytes of B lineage or endoderm derived structure, we found some samples enriched from among the lowest across all cell types to the highest (Fig 4A). Since each of our cell types were in fact heterogeneous groupings, we then investigated the underlying original cell labels. We found that the variability was due, at least in part, to the underlying cell label within the assigned cell types (Fig 4B). This exposed a high variability of TE enrichment across cell labels, even within a given cell type. We found that the Alu enrichment within lymphocytes of B lineage was mainly driven by the *B cells* (mean 19.3%) and not observed in tonsil germinal center B cell (mean 1.1%). Conversely, the colon cell type had more consistent results and smaller variance in its enrichment.

We were also interested in the L1 enrichment observed in H3K9me3 (Fig 2C). There, brain had the highest enrichment (mean 24.1% versus 13.3% across cell types Fig 4C), which is line with

reports of high brain and L1 associations in different contexts. L1s are upregulated in neural cell differentiation, in stress (rat hippocampus), brain diseases and for mosaicism while also being repressed by H3K9me3[48,121–123]. In contrast to Alu in H3K36me3, keratinocyte, colon and lymphocytes of B lineage had lower enrichment relative to mean (Fig 4A-C). This shows that the TE associations with histone marks change between tissues. Notably, within the brain cell type we find that differently annotated samples could have drastically different levels of enrichment (37% to 3%) further supporting that better and finer cell type annotation may help better understand the contribution of TEs (Fig 4D).

Finally, we tested if the health status of the samples had an impact on the enrichment levels observed in the different cell types. Of the cell types-histone pairs with multiple health status available (2024, 33% of samples), we found that 218 cases (4% of the total) of TEs featuring significant differences between health statuses (Fig 4E, Supplemental Fig 10). We highlight the top 10 of the cell types-histone pairs that had the most significant health associated differences (Fig 4F). For instance, we find that in H3K4me1 brain samples TE enrichment was significantly distinct between healthy and cancer cells across 21 TE families. We also show the top 4 cases with the most significant differences between healthy and cancer samples (Fig 4G). In all those cases, we find that cancer samples had a significantly higher TE enrichment.

These results show that different TEs are associated with different cell types and that health status can also affect those associations.

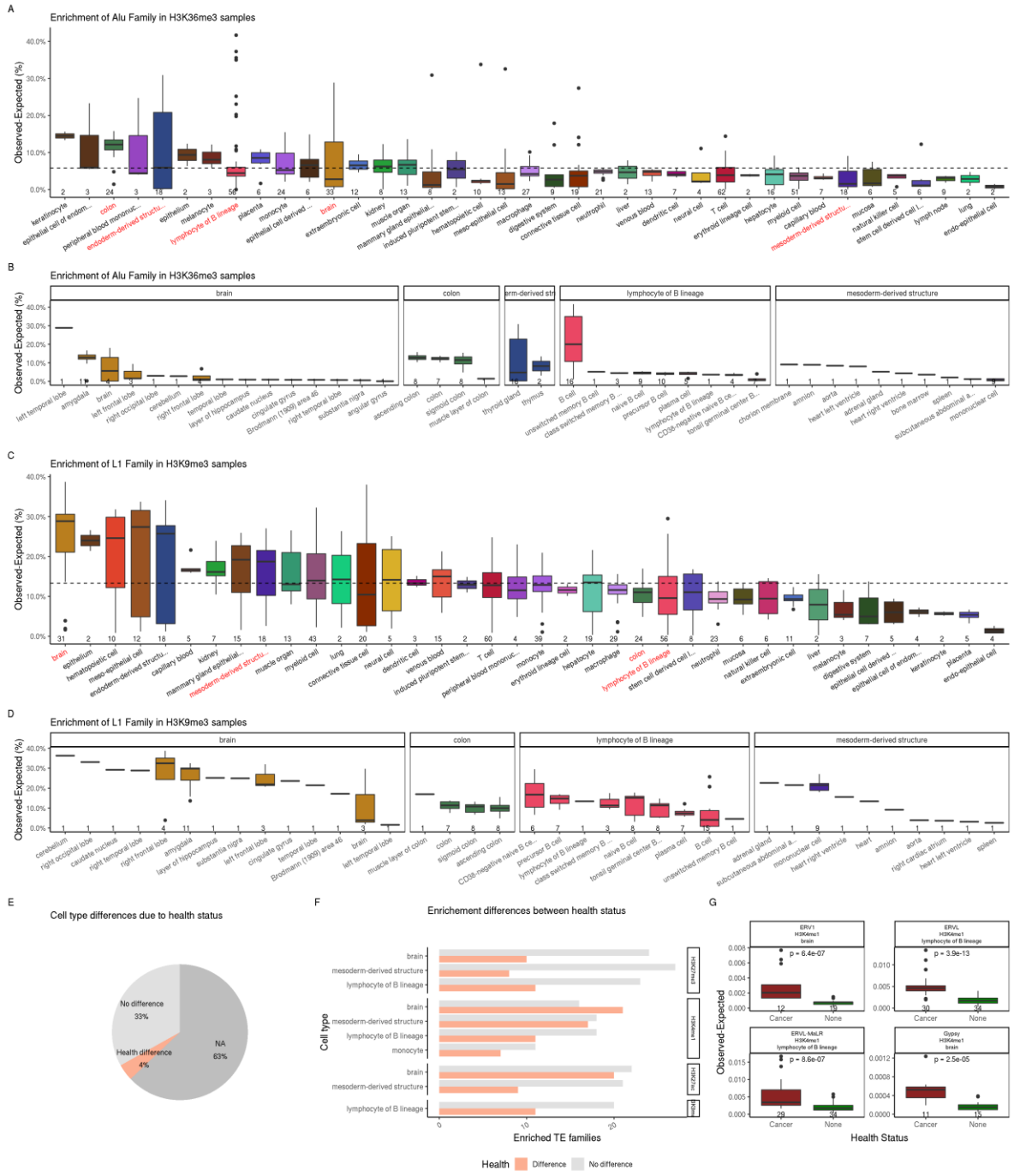

Figure 2.4 TE are enriched in histones in a cell type specific manner.
A| Mean Alu Enrichment in H3K36me3 samples grouped by cell type. Each point is the mean of all Alu enrichment (obs-exp%, significantly enriched only) for one sample. Boxplot shows distribution across samples, number of samples for each cell type listed below. Sorted by category Mean enrichment. Dashed line, overall mean. Only cell types including more than one enriched sample are shown. Highlighted in red are Cell types displayed in B. **B|** Underlying cell type label within select cell types displayed in **A. C-D|** Same as figure A and B for L1 in H3K9me3 **E|** Proportion of cases (cell type-histone pairs and TE combinations) where there were significant differences depending on sample's health status. **F|** Top 10

Cell type – Histone pairs with the most TE's featuring significant differences between health statuses. X axis is the number of TE families featuring the significant difference (orange) or having no difference (gray). **G**| Four examples of TE enrichment (obs-exp, significantly enriched only) significant differences within cell type – Histone pairs.

Identifying notable TE candidates from cell type specific enrichments

Having found clear associations between TEs and cell types, we wanted to know how specific these associations were and if we could leverage their specificity to identify notable associations. To get an idea of the cell type specificity of TE enrichments, we first grouped all samples by cell type and histone mark and measured the proportion of samples in which each of the select TEs were significantly enriched (Supplemental Fig 11A). We found that most TEs were enriched in only a few cell types, except for H3K9me3 samples enriched across most cell types. Next, we added up the number of cell types in which TE subfamilies were enriched, to measure the cell type specificity of TE enrichments and confirmed that it differed based on histone marks (Fig 5A). H3K9me3 featured multiple families that were enriched across many (>30) cell types while H3K4me3 and H3K27ac are highly specific with few (<30) cell types enriched. In particular, L1 and ERV TEs were enriched across cell types in H3K9me3, suggesting these TEs are repressed in a non-cell type specific way. We observed that the MIR family had the opposite trend. Although a family with few subfamilies, the number of cell types in which these subfamilies were enriched was high (~43) in H3K27ac and low (~10) in H3K9me3.

In summary, we found that across all marks, 1663 (21.8%) subfamilies were enriched in 1 to 5 cell types, 2827 (37.1%) in 6 to 30 cell types and 928 (12.2%) in 30 or more cell types (out of 47 ell

types, Fig 5B). We catalogued the TE subfamilies that were enriched in a cell type specific manner (enriched in 1-5 cell types) and found that ERV1 subfamilies had the most specific TE enrichments across all marks (Supplemental Fig 11B). Broken down by histone mark, we find that for H3K9me3 only 151 (11.9%) were enriched in 1 to 5 cell types compared to 361 (28.4%) for H3K27ac (Supplemental Fig 11C).

Finally to identify TE to cell type associations for further analyses, we selected TE candidates through two approaches: (i) cell type specific TE association with high enrichment (top specific obs-exp (TS.EN) or fold change (TS.FC)) and (ii) TEs that were much more enriched for a cell type than all the others (surplus obs-exp (S.EN) or fold change (S.FC)). For the top specific enrichment, we selected as candidates the TE-Cell type-Histone triplets that were more enriched than the 95th percentile of their histone group for either enrichment metric (Fig 5C and Supplemental Fig 12A). This led to the selection of 219 triplets based on obs-exp (TS.EN) and 219 triplets based on fold change (TS.FC, Supplemental Table 3). For instance, in H3K4me3, AluYb8 TE subfamily was selected since it had the highest fold change (14.19) in brain. Similarly, also in H3K4me4, L1PA7 was selected since it had the highest obs-exp enrichment (1.11%) in T-cell.

For the highest surplus, we identified the TEs that were much more enriched for a cell type than all the others by calculating the difference between the individual cell type TE enrichment and the mean across all cell types for each subfamily. We selected as candidate the top 20 TE-cell type pairs with the most difference in terms of obs-exp enrichment (S.EN, Supplemental Fig 11D) and Fold change (S.FC, Supplemental Fig 11E). The enrichment surplus was dominated by Alu

subfamilies associated with Brain, hematopoietic cells and mammary gland epithelial cells. For instance, AluY was 5.6% enriched for H3K9me3 in the brain, 4.5% more than the average across cell types (1.1%). Meanwhile, fold change surplus was mostly associated with GSAT centr repeats and ERV TEs. THE1C-int in H3K27ac had a 16.5 fold change enrichment while the mean fold change was only 2.8 fold. From all of this, we compiled a list of 456 cell type-histone-TE candidate triplets with their methods of identification (438 from cell type specific TE associations and 40 from the surplus criteria, with some being shared) (Fig 5D, Supplemental Table 3). Among the candidates, we note multiple previously observed associations such as MER11D in placenta[56,124], SVA_A for Stem cell being in line with pluripotent cell associations[6]and many Alus being enriched in brain[125,126].

These observations highlight how our candidate selection managed to recapture many previous associations and suggests that some of these new candidates could serve as cell type specific regulatory elements and are worth further investigation.

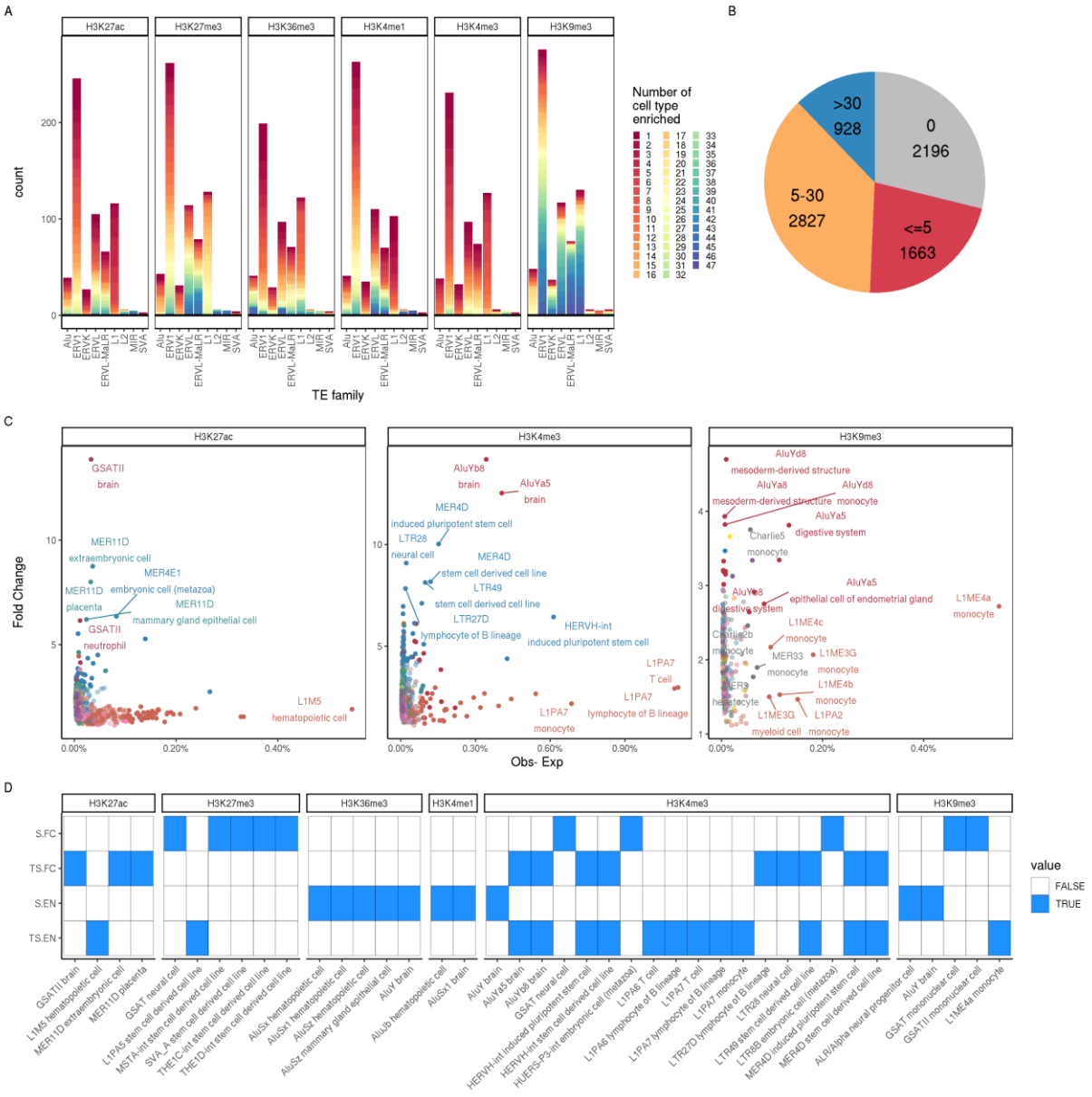

Figure 2.5 Identifying notable TE candidates from cell type specific enrichments.
A| Number of TE subfamilies within each family that were enriched for each number of cell types. X axis shows the different TE families, Y axis shows the count of enriched TE subfamilies and the color shows the number of cell types the enrichment was observed in (Red=Specific, Blue=Non-Specific).**B|** Pie chart of the number of subfamilies enriched in 3 bins of cell type numbers (or not enriched, gray) across all histones **C|** Fold change and obs–exp % enrichment of significantly enriched TE subfamilies per cell type of cell type specific subfamilies (enriched in 5 or less cell types, red segment in B). Labels are a random subset of the candidates: most enriched (95th percentile) points in terms of obs – exp or fold change. Only 3 histones are shown (H3K27ac, H3K4me3 and H3K9me3). **D|** Top 40 (10 selected per method) of the putative TE candidates annotated by the method of determination in y axis. S.FC is surplus fold change, TS.FC is top specific fold change, S.EN is surplus enrichment (obs-exp) and TS.EN is top specific enrichment (Obs-exp).

Identified TEs candidates are associated with relevant cell type biological processes and genes

Finally, we wanted to determine if we could link any of our 456 putative candidate TE subfamilies to potential activity through gene ontology associations. We merged the samples from a given candidate triplet to have one aggregate representative per context (merging of all peaks across samples from a cell type, assay and TE subfamily, Methods). For each merged sample we also generated a TE subfamily control keeping only the select TE's instances that did not have peaks in the aforementioned representative sample (Supplemental Fig 13). This was to check the importance of overlapping the histone within cell type for our triplets. Next, we looked at the biological processes terms that were significant compared to genomic background with GREAT[99,127] (Methods). For the 31 H3K27ac triplets, we found that there was some similarity in enriched terms between samples covering the same TEs, even across cell types (Fig 6A, Supplemental Table 4). Across marks, a similar pattern was observed for a subset of 209 triplets (Supplemental Fig 14, 15). It was clear that different processes were enriched within different triplets and thus, that TEs were, on some level, associated with different biological processes.

For example, we investigated MER11D in placenta and the associated biological processes and found enrichments for female pregnancy, ceramid and sphingolipids translocation (Supplemental Fig 14, 15A), all biological processes that are also involved in pregnancy and placenta[128,129]. We found a cluster of peaks within the PSG gene cluster (Fig 6B), near the *ABCB1* gene (and overlapping *RUNDC3B*) and near the *EPO* gene where a peak overlapped a candidate cis-regulatory elements (cCRE) (Supplemental Fig 16A, B). The *PSG* gene cluster stands for

pregnancy-specific glycoprotein is directly related to pregnancy, while *ABCB1* is associated with ceramid and transports proteins and *EPO* promotes blood cells and is secreted from fetal liver.

The enrichment of oxygen associated terms in MER51-int H3K27ac iPSCs also stood out as they were highly significant and high in fold change (cellular response to hypoxia, 4 hit, 1005 fold enrichment, pval= 9.37E-12; response to hyperoxia, 6 hits, 67 fold enrichment, pval=4.92E-10; regulation of release of cytochrome c from mitochondria, 8 hits, 42 fold enrichment, pval=1.95E-11; Supplemental 14, 15B). With the triplet (MER51-int H3K27ac iPSCs) containing 59 peaks, these terms accounted for between 6.78% to 13.56% of the peaks in this cell type. This is consistent with the fact that oxygen levels are known to be important in iPSCs and pluripotency[130]. When we looked at the genes that explained these biological processes enrichments, we found 3, *BNIP3*, *HDAC2* and *HGF*. Among those, we highlight *BNIP3* a gene linked to the 3 terms and near (within 50kb) 5 MER51-int associated peaks in H3K27ac. These peaks were missing from other cell types and were not found in H3K9me3 samples (Fig 6C). This shows an example of a TE, MER51-int, being associated with activating histone mark H3K27ac near a gene associated with biological processes relevant to a cell type of interest. We also note that these peaks did not overlap previously annotated cCREs, even if some were close (Supplemental Fig 16C). We found that AluY H3K4me3 peaks in brain were associated with autophagosome and many peaks around two brain related genes, *SYT11* and *RIT1* (Supplemental Fig 15C, 16D). Finally, we looked at genes near peaks and found multiple cases (20-31%; Fig 6D, star annotated genes) of the gene or protein product being expressed in the candidate's cell type. For instance, 32 peaks from L1PA7 H3K4me3 T-cell triplet were closest to *FAAH2* (Fig 6D) a gene highly expressed in T-cell according to The Human Protein Atlas (www.proteinatlas.org, Supplemental Fig 17A)[131]. We also

found *SLC25A18* (closest to 20 peaks) within H3K36me3 AluY Brain triplet being mainly expressed in brain and *ABR* (closest to 35 peaks) for H3K4me1 AluSx1 Brain triplet was also most expressed in brain (Supplemental 17B, C).

Taken together, these results show that through our candidate selection, we could find cases of TE instances in proximity to genes associated with biological processes relevant to their cell type.

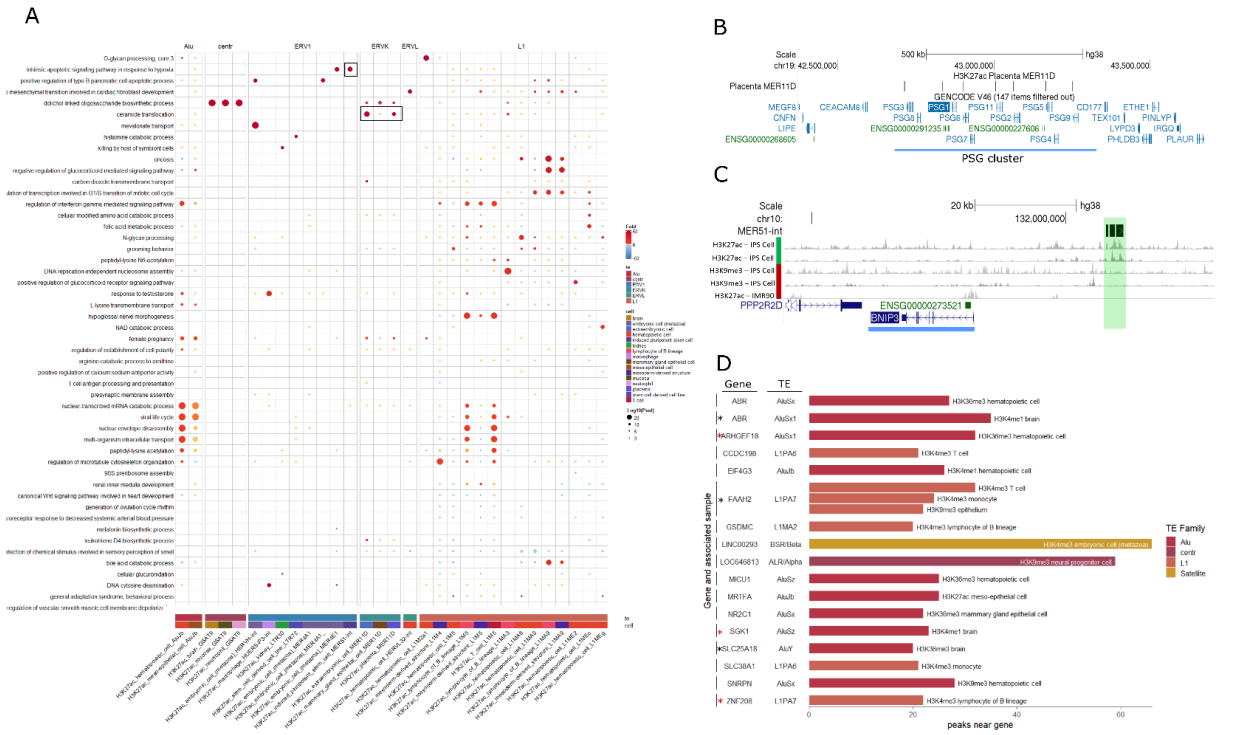

Figure 2.6 TEs associates with cell type relevant biological processes and genes.
A|Fold change enrichment difference from TE control of GO biological processes within H3K27ac candidate triplet (TE-Assay-Cell Type) samples. (fold change data – fold change of associated TE control) red enriched, yellow between -1 and 1, blue depleted. Circle sizes represents significance of the enrichment. Terms selected based on data, favored most enriched terms per candidates. In rectangle are some mentioned processes. **B|** Genome tracks around *PSG* gene cluster (blue highlight) showing a cluster of MER11D overlapping peaks from H3K27ac placenta sample **C|** Genome tracks around *BNIP3* gene (blue highlight) showing a few nearby MER51-int instances (green highlight) and IPSC H3K27ac and H3K9me3 tracks, as well as an H3K27ac IMR90 sample track, all from the IHEC portal (part of the underlying samples, but distinct from the actual data used due to the reprocessing). **D|** number of peaks near genes (within 50kb) found within candidate triplets samples. Colored by TE family, only top gene per sample shown. star: supported RNA/protein expression data, red star: weaker support.

2.4 Discussion

A number of recent studies have suggested a role for TEs in genome regulation [48,61,80,107,132–134]. In this context, a comprehensive study of TE and their association with the epigenome was needed. Here we investigated, 4614 samples spanning 175 cell annotations from 47 cell types and 6 histone marks. On average 55.8% of peaks were found to overlap with TE and that the overlap varied greatly depending on the histone mark investigated. This is in line with the observations by Pehrsson et al.[86] where they found varying degrees of TE overlap depending on epigenetic state. We also observed that H3K9me3, a less investigated repressive mark, had far more overlap with TEs (78.3%). We find that H3K9me3 is enriched especially in L1 and that this enrichment was not cell type specific. There was almost no enrichment in L1 for other histone marks. The role of H3K9me3 in silencing TE was established before[48,115,117] and was particularly highlighted in the context of brain tissue[49,123]. Our findings suggest that L1 is specifically being targeted by H3K9me3 histone modification across tissues to silence it.

While observing an association of TEs with a repressing mark was expected, we also found distinct enrichment of Alus with H3K36me3. Alus are one of the few TE families still able to transpose in the human genome[136]. In contrast to L1, they featured little association with H3K9me3. We also found that ERVs TEs were the most widespread TEs with some level of enrichment across most histone marks further supporting the ERV's contribution to the transcriptional landscape[80,137].

The TE-age association is a perspective that is worthy to explore as it relates to the time a TE had to be co-opted or decay within the genome. It was previously found that Alu and SINE become less methylated as they age[86], Alu become more preferred by H3K36me3 as they age[105], that older SINE are more in open chromatin and that generally TEs lose their motifs as they age[138]. Su et al. proposed Alus as proto-enhancers due to their general properties becoming more enhancer like as they aged and we were interested in seeing if that observation held true for other TEs. We found that while their general observation for Alu held true, the overall relationship between other TEs and TE age was complex. We find that in L1 and ERV TE families, whether they start enriched or depleted, TEs tend toward no enrichment as they get older. This can be interpreted as most TE degrading into background which is supported by previous results[138]. Notably, we found that some TE families (Alu, MIR, L2, Fig 3D), instead diverged from no enrichment as they got older, indicating they either became more enriched or depleted. This can be interpreted as most TE degrading into background which is supported by previous results[138]. Our results highlight varying TE family associations with histone marks and evolutionary trajectories over time.

The idea that TEs are involved in a cell type specific way has been supported by many studies [111,139] and here, with our large and varied dataset, we aimed to test and establish to what extent the association of TEs and the epigenome was cell type specific. We found that TEs can be enriched within histone marks in a cell type specific manner. While most cell types have a similar TE enrichment, select tissues displayed significantly distinct enrichments depending on TE and histone mark. Alus in H3K36me3 were significantly more enriched in colon than most other tissues. We also found a high enrichment within Lymphocytes. The importance of TEs in the immunity has been reported before [58,107,132,140] and here we also observe high variability which is

a factor of developing importance in the context of immunity[141]. L1 TEs were more enriched in brain than most other TEs within H3K9me3, which supports previous reports of L1 associations with brain[48,122,123] and H3K9me3[42,43,117] . We also found that in 4% of cell types-histone combinations cancer (or disease) samples could also feature significantly different TE enrichments than healthy ones. However, due to our datasets limited focus on health status contrast, we may be missing a lot of health associated TE enrichment differences.

We also measured cell type specificity of TE enrichments, through the number of cell types in which they were enriched. We found that TEs tended to be specific for most subfamilies in the activating marks while they were more non-specific subfamily in H3K9me3 (Fig 5A, supplemental Fig 13C). In contrast, we note that MIR elements were non-specific in H3K27ac while specific in H3K9me3 (Fig 5A). Given previous reports of MIR enhancer activity[110], we speculate that MIR is widely present in enhancers but selectively repressed for cell type specific purposes.

Finally, we used a set of criteria to identify 456 enriched cell type-histone-TE candidates (Fig 5D,G). We found that within our candidates, some biological processes were enriched in specific triplets. Among our candidates we identify MER11D in H3K27ac related to placenta and extraembryonic cells. The association between MER11 and placenta was observed before[56,124] and we found a cluster of MER11D associated peaks within the *PSG* gene cluster. We also highlight a MER51-int H3K27ac iPSCs enrichment in oxygen associated terms (Fig 6A). This enrichment came in part from MER51-int elements near *BNIP3* gene which were missing from repressive mark and some other cell types (Fig 6C). Looking directly at genes, we found that the genes closest

to our peaks, tended to have been reported as expressed within the triplet's cell type. However, the fact that the associations came from peak clusters near the same genes might be a confounding factor.

While our large and heterogeneous dataset enabled our comprehensive study, it also came with some limitations. First although a harmonized reprocessing was made through a singular pipeline for all samples, some batch effect from the different consortium could still be detected. However, since different consortium mostly investigated different cell types, the specific cell types within the categories would also influence such differences. Additionally, we observed many cell type specific enrichments but it would be interesting to perform a similar TE analysis on single cell data to better capture the differences between cell types which could be lost in our bulk and aggregated data. Finally, due to the ambiguity that comes from multi-mapped reads, only uniquely mapped reads were used for in this study. Since TEs can lead to multi-mapped reads[89], it is likely that some TE reads, and thus TE peaks, may have been lost leading to underestimating enrichments. It would be interesting to assess the exact TE contents lost from multi-map reads in future studies.

In summary, our data present an comprehensive overview of TE contents across histone marks and cell types. It shows the consistent yet complex relationship between TEs and the epigenome and further supports the implication of TEs in genome regulation.

2.5 Acknowledgements

This work was supported by a Canadian Institutes of Health Research (CIHR) program grant (CEE-151618) for the McGill Epigenomics Mapping Center, which is part of the Canadian Epigenetics, Environment and Health Research Consortium (CEEHRC) Network. G.B. is supported by a Canada Research Chair Tier 1 award and a FRQ-S, Distinguished Research Scholar award. We would like to acknowledge Calcul Quebec and the Digital Research Alliance of Canada for access to computing resources. We thank IHEC for making available reprocessed and harmonized epigenomic data from a large collection of human cell types[112].

2.6 Methods

Data Collection

The dataset was downloaded from the IHEC EpiATLAS integrative analysis sFTP server on January 23rd 2023 and also available on the IHEC data portal (https://epigenomesportal.ca/ihec/ , https://ihec-epigenomes.org/epiatlas/data/). The available ChIPSeq narrowPeak files for the 6 main histone marks (H3K4me1, H3K4me3, H3K27ac, H3K27me3, H3K9me3, H3K36me3) were downloaded. The narrowPeaks were obtained using GRCh38 reference build, with pval of 0.01. The data downloaded included samples from 8 consortiums (Blueprint, CEEHRC, ENCODE, DEEP, NIH Roadmap Epigenomics, AMED-CREST, GIS, EpiHK). The cell type and all metadata annotations was taken from the IHEC metadata harmonization v1.2. Only samples with an Epirr id within the metadata v1.2 were used (ENCODE samples were dismissed). Comparison were made with Roadmap complete Epigenomes while only considering the 6 main histone marks.

ENCODE dataset sample count for comparison was obtained by filtering within the Experiment matrix, epigenome dataset for the 6 main histones totalling 1426 available samples (https://www.encodeproject.org/, Jan 2024).

Quality control and sample selection

For each sample, we measured the distribution of peaks within regions relative to transcription start sites (TSS). Peaks were attributed to the first group they fell within among: TSS (within 1000 bp from TSS), Promoter (within 5000 bp upstream of TSS), Intragenic (Overlapping genes), Proximal (within 10 kbp from TSS), Distal (within 100 kbp from TSS) and desert (further than 100 kbp from TSS). The coverage of each region was compared between samples from the same histone mark and samples that were an outlier (according to boxplots, above (75% quantile) + 1.5 * interquartile range (IQR) or bellow (25% quantile) - 1.5 * IQR) in the covered percentage for 2 different regions were discarded. In addition, samples that were outliers in terms of peak count when grouped by histone marks were also discarded.

Dimension reduction on ChIP-seq samples

To have broad overview of our samples similarity and do dimension reduction to display the data, we counted the number of peaks within 10 kbp windows across the entire genome using bedtools[142] intersect for each sample. We sorted the windows by variance across the samples and kept the top 21,000 windows before discarding the top 1000 to protect against potential irregularities (e.g. regions with very high coverage). This resulted in 20,000 windows upon which UMAP was performed with r umap package using default arguments with 400 epochs.

Measuring TE enrichment

Histone mark peaks were annotated using the UCSC RepeatMasker track[15]. We resized the peaks to 200bp around their center and counted the number of times ChIP-seq peaks overlapped TE instances with bedtool intersect, when more than one TE overlapped the peak, the largest overlap was kept. As done before [80,143], to have a random baseline to compare against, we simulated for each sample a library of 200 bp random regions, with the same distribution of distances to nearest genes. This was done using the distribution of peaks described in Quality control and sample selection method. For each sample, the simulation was repeated 1000 times and we counted the incidence of observed count being higher than the random baseline for each repeat subfamily. A repeat subfamily was identified as significantly over-represented (enriched) when the over-represented incidence was greater than 995/1000 ($p < 0.005$). coverage percentages were measured as *peaks overlapping TE/sample (or simulated sample) peak count*. Observed – Expected metric was calculated by subtracting the expected coverage (resulting from mean across the 1000 simulations) from the observed coverage from the sample. Significant Observed – Expected was calculated using the only the repeat subfamilies identified as significantly overepresented (thus always positive, because observed in inherently higher than expected for each subfamily used).

Selection of the most enriched TE subfamilies

The most enriched TE were selected by doing a cumulative sum of the positive (>0) enrichment (observed-expected) TE subfamilies (excluding the simple_repeats) across all samples. Upon this

cumulative measure, the threshold was set as the value of the upper whisker of a boxplot, the value beyond which values are considered outliers. It was defined as *(75% quantile) + 1.5 * IQR.*

Correlation between samples and TEs

The correlation between samples was done with a matrix keeping the count of peaks found within the 20,000 10 kbp windows with highest variance (as described in the Dimension reduction on ChIP-seq samples section above) on which we grouped samples by assay and calculated the mean peak count per assay for each window. Then, the correlation between the assays was calculated with the Corr function in R.

The correlation between TEs was done with a matrix keeping the Observed – Expected (%) enrichment of the selected TE families for each sample on which we grouped the samples by assay and calculated the mean enrichment per assay for each TE family. Then the correlation between the assays was calculated with Corr function in R.

Estimation of TE Mappability, age and age categorization

TE mappability was calculated using the 50 bp Mappability track[116] from the UCSC Genome Browser, which is a conservative estimate of the true mappability since most of the reads in IHEC were targeting 75 bp and mappability increases with read length. The coverage of all TEs (also UCSC track) by the 50bp Unique Mappability (Umap 50) track gave us the proportion of each TEs that could be uniquely mapped which we used as mappability metric. The age estimates of each TE were based on the sequence divergence (milliDiv value from RepeatMasker) as described in

Bogdan et al.[132] We first divided the milliDiv value of each TE by 1000 and then by $2.2x10^{-9}$, the substitution rate of the human genome to calculate the age. The age of each family and subfamily was then obtained from the mean of all their instance's age. We categorized the subfamilies into Old or Young depending on the subfamily's age rank within the family. The youngest 50% were categorised young and oldest 50% old. For the dynamic groupings, if the difference between the old and young TE absolute enrichment (absolute value enrichment) was less than the young absolute enrichment (did not double or halve) and the p-value was larger than 0.05, the context (TE family for Histone mark) was deemed to have no change. If there was change, when the Old TEs absolute enrichment was larger than the Young TE's the context was categorized as diverging (moving away from 0 enrichment) and otherwise converging (approaching 0 enrichment).

Candidate selection

We selected TE candidates through two approaches: (i) cell type specific TE association with high enrichment (top specific obs-exp (TS.EN) or fold change (TS.FC)) and (ii) TEs that were much more enriched for a cell type than all the others (surplus obs-exp (S.EN) or fold change (S.FC)). For the top specific enrichment, we selected as candidates the TE-Cell type-Histone triplets that were more enriched than the 95^{th} percentile enrichment of their histone group for either enrichment metric (obs-expected or fold change). For the surplus method, we calculated the mean TE enrichment across all cell types and subtracted it from each cell type's enrichment. We selected the top 20 (for both obs-expected and fold change)

This resulted in a final set of 456 candidates. For a more restrained and manageable set, we selected a subset of the top 15 most enriched candidates per histone for top specific obs-exp and fold change using their respective metric (obs-exp and fold change, respectively). We thus had 90 top specific obs-exp candidates, 90 top specific fold change and the 40 surplus candidates. This resulted in a reduced set of 209 candidates (due to some candidates being identified by more than one method).

Identification of associated GO terms

To identify the GO terms associated with TEs depending on histone or cell type, we first merged all bed files by cell type and histone to have 1 representative per doublet (e.g. peaks from H3K9me3 in Brain cells). We then split the merged samples by TE to have the peaks associated with each TE for all doublets and resulting in triplets (e.g. peaks from H3K9me3 in brain cells overlapping L1PA4). As a control to observe the influence of TEs on their own, triplet controls were made by taking all instances of the given TE (straight from repeatmasker) and removing all those that overlapped peaks from the given triplet. These triplet controls represent the instances of the TE not within peaks of the cell type and histone (ex: L1PA4 peaks that are not in H3K9me3 brain peaks). We obtained the associated GO terms from these Triplets and associated controls using the R version of GREAT (rGREAT)[127]. We used the default configurations (5kb upstream, 1kb downstream, up to 1000kb) and background (genomic background) for the analysis. To account for the use of genomic background, the same analysis was made on the controls which were used to assess if the enrichment observed could be explained by our controls (TE itself or Histone-Cell type doublets). For visualisation purposes, GO term p-value was capped at 10E-200. The GO terms to show are selected by sorting the terms by p-value first and fold-change second. The genes supporting the GO term associations as well as the location of relevant peaks were

obtained by using the GREAT website (https://great.stanford.edu/great/public/html/index.php) version 4.04, hg38 and default configuration.

Visualisation

Figure generation was done using R[144], heatmaps were made using complexheatmap[145]

Chapter 3

Building a Web Tool for Transposable Element Enrichment Visualization and Analysis

Preface: bridging text between chapter 2 and chapter 3

In the experiment described in the previous chapter we did a comprehensive analysis of transposable elements in the epigenome across histone marks and cell types. We highlighted some of the most striking results and presented some putative regulatory TE candidates. However, with a set of more than 300 histones and TE family combinations, each with their measurements of 47 cell types, it was clear that we could not cover every noteworthy association within a publication. Furthermore, we only noted the standout association such as extreme TE enrichments or large TE enrichments differences between cell types, we could not even begin to cover all the subtler yet biologically striking results requiring specific cell type expertise.

In this chapter, we set out to make this data more accessible by building an online portal. Inspired by some of the tools we used (or failed to) during the project of Chapter 2, I determined that an online tool would be the best way to minimize friction and encourage adoption. We leveraged some of the results from Chapter 2 and built an interface to enable querying specific combinations of TE, cell type and histone marks. Our objective was for anyone who saw the results of Chapter 2 but was interested in a context that we did not cover in the main text to be able to query it

themselves. In this way, our results become more than our own conclusion but rather, serve as a resource. Once our tool could display our TE enrichment and overlap measurements we thought to expand the tool with an analysis functionality. We thought that an atlas was useful, but the ability to for other researchers to assess their own data and compare it to all the data available on the portal would be even better. The objective for the portal is thus two fold, first, to answer any TE curiosities of researchers and second, to ease in the adoption of measuring TE overlap and enrichments through a simple, no installation TE analysis functionality. Although Chapter 2 and 3 both use the EpiATLAS dataset, disparities between in results values may happen due to differences in the specific version used.

This manuscript is to be submitted.

TEExplorer: A Web Portal to Investigate TE-Epigenome Associations Across Human Cell Types

Jeffrey Hyacinthe[1], David R Lougheed[3,4], Guillaume Bourque[2,3,4,*]

[1] Quantitative Life Sciences, McGill University, Montréal, QC, Canada

[2] Department of Human Genetics, McGill University, Montréal, QC, Canada

[3] Canadian Center for Computational Genomics, McGill University, Montréal, QC, Canada

[4] Victor Phillip Dahdaleh Institute of Genomic Medicine at McGill University, Montréal, QC, Canada

[*] Correspondence: jeffrey.hyacinthe@mail.mcgill.ca

3.1 Abstract

Motivation

Transposable elements (TEs) are genomic sequences that can create copies of themselves and move within the genome. TEs cover about half of the human genome and there is a growing amount of studies that support their implication in genome regulation, their co-option and their association with cancer. However, their repetitive nature makes them challenging to analyze and thus often overlooked. The specialized tools needed to analyze TEs are often complicated to use or targeted to a specific TE family. In a previous study, we leveraged 4614 ChIP-seq samples from the EpiATLAS dataset and did a comprehensive analysis of the relationship of TEs with 6 histone marks and 47 human cell types. However, due to the number variable combinations, there were too many associations to highlight. We thought that those results could be useful to other researchers and that, in order to expand the consideration of TE involvement in genomic studies, they should be shared in an accessible and intuitive manner.

Results

Leveraging results from our previous study, we developed a web tool, TEExplorer, which makes available TE enrichment data for 57 cell types across 6 histone marks and works as an Atlas of TE overlap and TE enrichment across histone marks and cell types. The tool has 3 sections: the TE overview, TE subfamilies and Import sections. Within TE overview, the user can have a broad look at TE families and their overall overlap and enrichments across any of the 57 cell types and 6 histone marks. The TE subfamilies section allows the investigation of TE subfamilies within a Select TE family. The section reports the TE enrichments or overlap of the TE subfamilies within a select histone mark across cell types. Finally, the Import section allows users to upload their

own ChIP-seq BED file and obtain the TE overlap and enrichment of their own data and to compare their data with the EpiATLAS dataset. For the last 2 sections, the resulting overlap and enrichment can be downloaded as a table and 3 TE metrics can be visualized. With TEExplorer researchers with an interest in a particular histone mark, cell type or TE explore all the existing the associations found within the large EpiATLAS dataset with a dynamic interface.

Availability

Online portal: https://teexplorer.c3g.sd4h.ca

3.2 Introduction

Transposable elements (TEs) are DNA sequences with the ability to transpose and duplicate themselves, which cover about half of the human genome[102]. The ability to duplicate themselves has generated some interest in their potential to act as regulatory units that disseminated regulatory sequences throughout our genome[33]. However, due to our limited ability to characterize TEs because of their repetitive nature[89], much of their potential has only recently started to be explored. Moreover, uptake is still limited because their importance within regulation still isn't widely known and their analysis requires specialised tools which are often hard to use. Notably, one of the ways to spread scientific methods and approaches is to share data and results but it is just as important to make that the data is easily accessible and interpretable[146]. Some of the most widespread methods such as GREAT[99] and the UCSC genome browser[92] or datasets such as ENCODE[147], GTEX[96] leverage an online interface for ease of access, usage and sharing. These resources all highlight that when data and tools are made easy to use and interpret, it greatly expands their usage by researchers. The international human epigenome consortium (IHEC)[94] is a consortium of international consortiums with the overall objective to better understand the role of the epigenome in human evolution. The consortium makes available a large number of epigenome datasets including histone mark ChIP-seq samples. Recently the consortium has been working on a new uniformly reprocessed dataset, the EpiAtlas.

In our previous study[148] leveraging the IHEC EpiATLAS new dataset, we presented a comprehensive overview of the relationship between TEs, cell types and the epigenome. This included TE overlap with histone marks and TE enrichments measured by comparing observed overlaps relative to random control overlaps. However, we could only highlight some of the most

striking examples that we detected, and we believed that there are many more noteworthy associations. We wanted to make our results available in an easily navigable and interpretable manner and promote TE consideration by making similar TE analysis easy to initiate.

Here we present TEExplorer, a web portal that makes available TE enrichment data for 57 cell categories across 6 histone marks, which can be plotted and presented for 3 different metrics and investigated at the level of 60 TE family or 1,426 TE subfamilies. Notably, the portal also allows users to upload their own data for a rapid analysis of the TE overlap and enrichment of their own samples as well as to compare user results to those of the large IHEC consortium.

3.3 Dataset and Features

3.3.1 Data and structure

To enable our web tool, we used data from EpiATLAS dataset from the IHEC[94] and our previous TE study[148]. The data included 4614 Chromatin Immunoprecipitation sequencing (ChIP-Seq) samples from 57 cell types across 6 histone marks (H3K27ac, H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K9me3). We obtained the overall TE overlap from all the samples as well as their associated controls from the former TE assessment of the EpiATLAS dataset, a uniformly reprocessed dataset of samples from 8 consortiums within the IHEC. Through this we obtained 6,572,151 measurements of ChIP-Seq peak overlaps with TE and the associated expected overlap from controls. We also used RepeatMasker's[15] data for TE family information and the 50 bases pair UCSC mappability track[116] data for mappability estimations. The data was summarized into 2 databases that covered the two main TE data granularity, TE subfamily and TE family. The TE

subfamily database contained an entry for each TE subfamily per sample and its cell type, assay, observed overlap count, expected count from random simulations, total peak count in the sample and calculated metrics such as observed-expected count and fold change over expected. The TE family database was a pre-calculated aggregate of the former database where we grouped added up all overlap count within subfamilies to tally family level totals per samples. Following the tally, fold change is calculated through observed/expected. This allowed us to save computations on always adding subfamilies for all family level results.

The webtool is divided in 3 sections, the TE overview section for a broad look at TE families and their overall enrichments, the TE Subfamilies section to investigate a specific TE subfamily within a select histone mark across cell types and TE subfamilies and the Import section allows users to quickly obtain the TE measurements of their own data as well as to compare it with the EpiATLAS dataset.

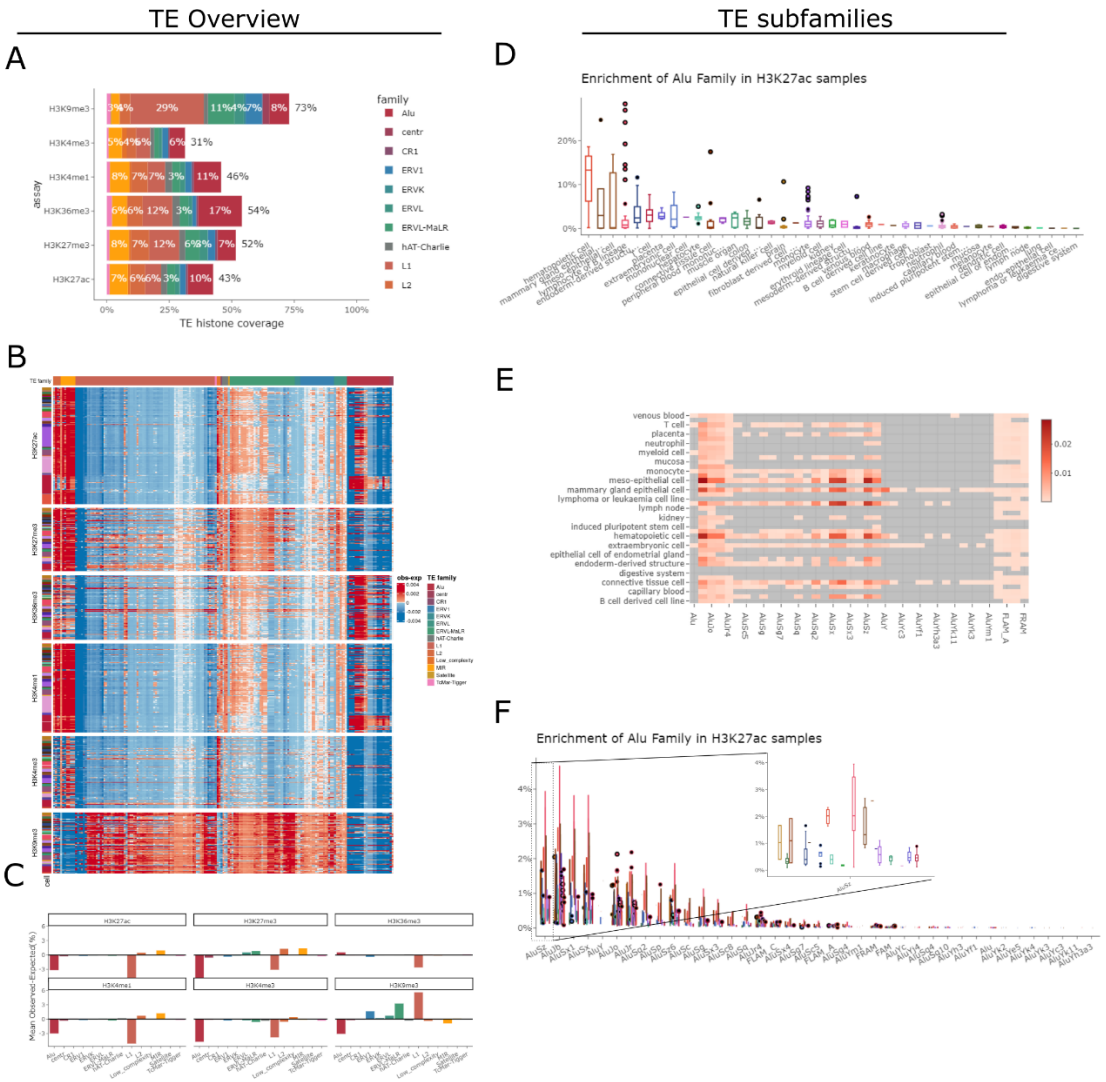

Figure 3.1 Overview of main plots from TE overview and TE subfamilies sections
A) Bar plot of mean TE overlap within the 6 histone marks broken down by TE family (fill color). The results shown depend on the user input query. **B)** Heatmap of TE subfamily enrichment of the samples matching the user input query, left bar annotations are cell type, top annotations are TE families. **C)** Mean enrichment of TE families for each selected assay **D)** Boxplot of TE Family enrichment of selected TE family (here Alu) across cell types sorted in descending mean order. **E)** TE subfamily enrichment heatmap of the TE subfamilies within a chosen family (here Alu) within cell types. **F)** TE subfamily enrichment boxplots within a chosen family (here Alu) across cell types (boxplot colors). Shown in zoom in on a specific subfamily (AluSz), the plots E, F, G are interactive and can zoom on select part for more clarity.

3.3.2 TE overview

The TE overview tab provides an overview of TE family enrichment across cell types and histone marks (Figure 3.1A,B,C). By default, this section displays data for all major TE families across all cell type and histone marks, but users can select specific TE families, histone marks and cell types of interest for more specific inquiry. After selecting the input and running the query, this section displays 3 plots. First a barplot the percentage of TE overlap for all selected histone and TE families (Figure 3.1A). Second, a heatmap of the TE enrichment relative to control (observed – expected) for all samples matching the query parameters across the TE subfamilies within the selected families (Figure 3.1B). By default, only the top TE subfamilies are shown but all subfamilies can also be toggled on. The third plot is a bar plot of the enrichment relative to control of TE families (Figure 3.1C).

With these plots users can get an idea of which TE families are most present in each cell type for the various histone marks. The heatmap enables the comparison of enrichment profiles between samples and cell types.

3.3.3 TE Subfamilies

The TE subfamily tab allows the investigation of specific TE families and their subfamilies coverage and enrichment across cell types. The user can select a TE family and a histone mark and obtain a breakdown of the TE enrichment or observed count across cell types and subfamilies. This section first reports basic information of the selected TE family (instance count, subfamily count and estimated mappability) and displays a boxplot of that family's enrichment or observed count for the selected TE family and histone across cell types (Figure 3.1D). It also reports the TE

enrichment or count for each subfamily within the select TE family across the cell types as a heatmap (Figure 3.1E) and boxplots (Figure 3.1F). Finally, the data from the query can be explored as a table and also downloaded.

Through this section, users can gain insight on the level of enrichment of a chosen TE across cell types and have a breakdown of the TE subfamilies contributing to this enrichment.

3.3.4 Import

The import tab allows users to import their own bed and get their TE overlap as well as TE enrichments and comparisons to the EpiATLAS data. The user can upload multiple bed files and select which cell type and assay (histone) to compare them to (this would usually from the same cell type as their samples, Figure 3.2A). When the analysis is launched, the uploaded files TE overlap is measured and compared to the EpiATLAS pre-generated random control TE overlaps (matching the cell type and assay query). The samples are also projected upon a UMAP plot to see if it aligns with the EpiATLAS data as a visual quality control(Figure 3.2B).

This section reports the TE overlap percentage broken down per TE family for each sample as well as those from the mean of all EpiATLAS samples matching the selected cell type and assay (Figure 3.2C). The user can then choose between 3 metrics for the other results plots: Observed – Expected, Observed count or Fold change. The web tool then displays a heatmap of the TE subfamilies for the uploaded samples in comparison to the EpiATLAS samples matching the query (Figure 3.2D), the mean TE family enrichment or mean observed count within each uploaded samples, and then two TE subfamilies breakdown as boxplots (Figure 3.2E) and a heatmap (Figure 3.2F), once again comparing the user data to the EpiATLAS samples. Finally, we also report a table of the data which includes the TE count for each subfamily per sample, the total peak count within the sample,

the expected count according to the chosen cell type and assay as well as two enrichment metrics, Observed - Expected and Fold change. This table can be downloaded to power further analysis or allow custom plotting.

These plots and data allow users to get an overview of the TE content of their own data and compare it to the existing EpiATLAs data. It can give an idea of which TE families and subfamilies are most present, if they are found more than expected. If all cell types are chosen, it can also be used to identify which cell type the uploaded cell type is most similar to.

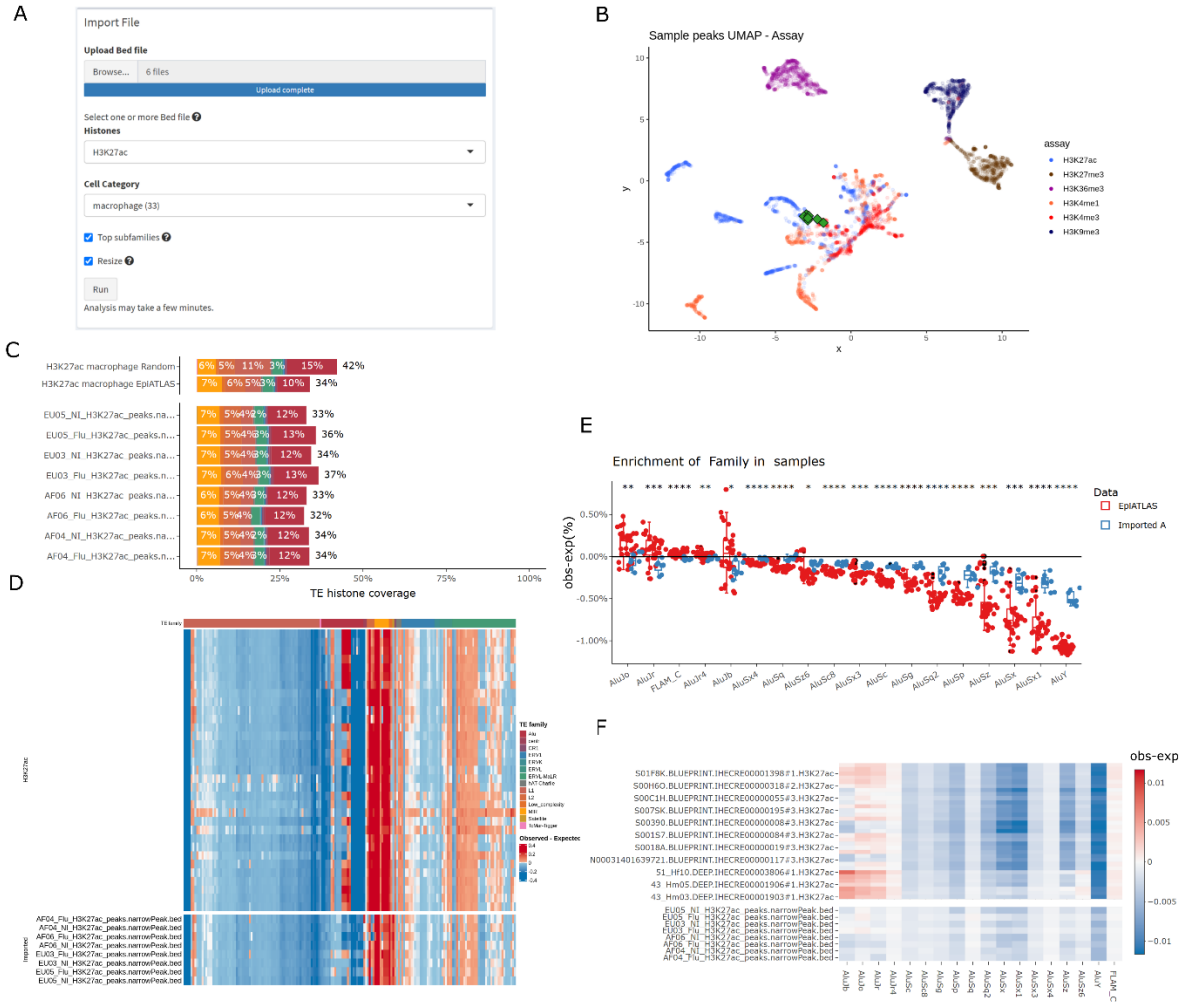

Figure 3.2 Main plots of the Import tab

A) Interface of data upload and input selection. **B)** UMAP on peak counts within 10kb regions across the genome for the EpiATLAS samples. Only the 20,000 windows with most variance (excluding the first 1000) were used. Color represents the assay. Overlaid green diamonds are the uploaded samples using the same method and regions. **C)** Bar plot of TE coverage the EpiATLAS data from selected histone and cell category and associated random simulation (top) and of the uploaded samples (bottom) **D)** TE enrichment relative to random controls of select subfamilies from EpiATLAS data (top) TE enrichment relative to the mean of random controls of the uploaded samples. **E)** Boxplot of the TE enrichment of TE subfamilies within a chosen family (here Alu). Compared between the EpiATLAS data and the uploaded samples (imported). **F)** heatmap of the TE enrichment of the TE subfamilies within a chosen family (here Alu) within samples. Shown in EpiATLAS samples (top) and the uploaded samples (bottom). The values shown in D, E and F can be Observed – Expected (shown), Observed count or Fold change. For all plots EpiATLAS samples used are a subset matching the chosen cell type and assay.

3.4 Implementation and methods

3.4.1 Implementation and data transformation

The webtool is powered by R[144] and R shiny and is available online at https://teexplorer.c3g.sd4h.ca. To improve performance, the data which were large tables were converted into sqlite databases[149].

For quality control purposes and to compare samples, we kept a sample peak approximation in the form of the total peak count within 10 kilobase pair (kbp) windows across the genome. These counts were tabulated into a table of 321186 rows per sample is reduced to the 20 000 most variable windows by variance and the index of these windows to reuse on user uploaded data. We ran a UMAP[150] (from umap R package) dimension reduction on those 20000 windows for all samples and show the first 2 dimensions in a plot that shows sample similarity and retain the model for user uploaded data projection.

3.4.2 Data import

The webtool can analyze bed and bed-like files such as narrow and broad peaks. The human reference used is hg38. Uploading up to 50mb of files is allowed. For each uploaded hg38 bed file, an overlap with the RepeatMasker TEs and with a template bed file of 10kbp windows across the genome are performed using genomicranges[151] and counts of peaks overlapping TEs and 10kbp windows respectively are obtained. By default, the uploaded peaks will be resized to 200bp in order to match the size of the peaks that were used for the EpiATLAS data. Of the 321186 10kbp windows, the 20 000 most variable windows from our data, as previously determined and saved, are kept. To compare the user uploaded samples to those from our data, we project the user samples

(with the R *predict* function applied to the retained umap model) onto the previously generated calculated UMAP plot on the EpiATLAS data. Since this is a computationally intensive step it is hidden by default and only calculated if the user expands the Quality control section.

The TE overlapping counts of uploaded samples are kept as a separate table with proportions being calculated through *peak overlapping TE count/peak count in sample* for each subfamily.

For the TE enrichments and comparisons between user and our data a new temporary table is created by combining a subset of our dataset that matches the chosen cell type and assay. The subset's expected counts are averaged across all samples and used as the user uploaded sample expected TE counts. This provides a random sample approximation without having to perform all simulation the computation.

3.4.3 Data visualization

The plots were generated with ggplot2, plotly for dynamic plots as well as complexheatmap[145] for the static heatmaps. The overall interface was powered by shinydashboard and shinyWidgets[152].

3.5 Conclusion

In summary, TEExplorer simplifies TE consideration by allowing researchers without TE expertise to explore which TEs are predominant within their experiment and how these TEs compare to expected levels. Thus, the portal can highlight whether TEs could have been overlooked within an experiment. From our previous study, we found more associations between TE, histone and cell type than we could ever highlight. This portal allows researchers with an

interest in a particular histone mark, cell type or TE interests to explore a vast array of associations directly. It can also quickly provide researchers with the TE overlap and enrichment of their own ChIP-seq BED files and show how it compares to the portal's data and the simulated controls used for the portal's analysis. While specialized tools will always be necessary for more in-depth TE analysis, our portal establishes a straightforward starting point towards greater TE consideration.

3.6 Acknowledgements

This work was supported by a Canadian Institutes of Health Research (CIHR) program grant (CEE-151618) for the McGill Epigenomics Mapping Center, which is part of the Canadian Epigenetics, Environment and Health Research Consortium (CEEHRC) Network. G.B. is supported by a Canada Research Chair Tier 1 award and a FRQ-S, Distinguished Research Scholar award. We would like to acknowledge Calcul Quebec, SecureData4Health and the Digital Research Alliance of Canada for access to computing resources. We thank the IHEC consortium for their EpiATLAS dataset.

Chapter 4

Discussion of Results

This thesis has shown the various ways in which TEs have and may continue to impact the human genome. In Chapters 2 and 3 we presented an overview of TE's association with the epigenome, in Chapter 3 we improved this overview by adding more metrics, performing analysis at the subfamily level and developing a dynamic interface to visualize the results for any subset of user interest. We have compiled and made these results accessible on a web portal and developed a process for other researcher to obtain comparable TE analysis. I believe that these results bring novel insight into the importance of TEs and their intricate relationship with the epigenome and that they will help expand their consideration in genomic research. In this chapter, I will discuss some more specific results related to my methods, the associations between TEs and histone, cell types, health and DNA methylation. I also discuss some use cases for my tool TEExplorer. Associated supplemental figures are in Appendix C

4.1 Distribution Matching Simulated Controls

The two data chapters leverage the implementation of a distance to TSS distribution matching set of simulated controls. This allowed us to have, by the virtue of matching the sample's peak distribution, more representative controls that could be generated for all samples. A common approach to controls in genomic experiments is to use randomized or scrambled simulated samples. These allow the comparison of observed measurements with what would be expected by chance. Since transposable elements are not randomly distributed, adjusting the distribution of our simulated controls such that it was not fully random, but rather representative of the sample was

an elegant solution. This ensured that peaks did not go in regions where they wouldn't be expected in the first place for TEs and were not biased by the nature of a general control. Although, the simulated samples were rather similar to each other, there were still some slight differences in terms of the TE constitution of these simulated samples, especially between assays, no doubt due to the distribution matching adjustment (Supplemental Figure S2. 1). In our implementation we used 6 region categories – TSS, promoter, intragenic, proximal, distal and desert – but these regions and the distances could be customized and fine-tuned to become more representative of samples. We used the commonly used ranges for these regions, but it might be interesting to study different values and their impact on results. Reusing the distribution measurements as a sample approximation proxy for quality control and comparison purposes was an innovative approach that may need to further explored and expanded.

4.2 The Complex Associations Between TE and the Histone Marks

We found that, as generally known[8], while TE covered a high proportion of the genome (mean of 55.8%), TE enrichment relative to expected levels was much lower (10.7%). In fact, TE subfamilies where found to be significantly enriched compared to simulated controls in less than 2% of cases. However, those enriched elements displayed clear patterns of association with histone marks. The repressive histone H3K9me3 had far more TE overlap (78.3%) than all other marks, and was also more enriched (27.1%). Different TE families were enriched for distinct histone marks; and we often observed enrichment contrasts between activating and deactivating marks within one mark and cell types had different levels of enrichments depending on which histone and TE family we looked at. For instance, we found that MIR an ancient TE family tended to be much more associated with activating histone marks, especially compared to L1 elements that were

highly repressed by H3K9me3, or Alus that were particularly enriched for H3K36me3. While the association between L1 and H3K9me3 was not entirely surprising, it is not entirely clear why Alus are enriched in H3K36me3. MIR was previously recognized as a family with positive correlation to tissue specific gene expression and exapted into enhancers[110]. Thus it is reassuring that we also found this distinct activating histone association. It was reported that Alus being in H3K36me3 regions may lead to Alu transcript expression[105], but there are very few reports on this pairing at this time. Interestingly, both Alus and H3K36me3 are heavily reported to be involved with splicing[153,154]. It might be worth investigating if there is a mechanism that connects Alus and H3K36me3 with mRNA splicing. It was also interesting to note that some of the TE subfamilies that we highlighted, such as AluYb8 and L1PA2(candidate), were some of the youngest TE subfamilies, which are reported to still be active[8,23–25].

The heatmaps of TE subfamily enrichment per sample featured in both Chapter 2 and 3, highlight how the different TE families had preference of associations for certain histone marks, but left me curious about the potential of combinations of histones. What about poised states of H3K27me3 with H3K4me3? Are the other histone marks also sometimes found together? These would be challenging experiments and would lead to an explosion of possibility (on an experiment that already featured too many variables). Through this line of thinking I have come to better understand the value of chromatin states and the chromHMM model that essentially takes the numerous combination possibilities and reduce them to a reasonable (15) set of states.

4.3 The Relationship Between TEs and Cell Types Depends on Histone Marks

In our results, it was striking to see that the TE enrichment varied so much across cell types. For example, for L1 in H3K9me3, brain was the most enriched at 24% (mean) meanwhile the lowest enrichment was in endo-epithelial cells at 1%. Furthermore, the results varied drastically depending on the specific TE-histone combination we studied. For instance, highlighting only the most enriched cell type (means enrichment, n>1): for Alu in H3K36me3, keratinocyte (14%); for Alu H3K9me3, hematopoietic cell (22%); for Alu H3K4me3, lymphocyte of B lineage (4%); for MIR in H3K27ac, digestive system (3%). This could suggest that the profile of histone mark-TE associations may contribute to cell type differentiation or proper gene regulation. It is worth mentioning that being able to explore all these combinations is a great use case for our tool TEExplorer. One of the more unexpected results that we found was that some TE enrichments within a cell type had very large variability. When we looked at the underlying cell labels within the cell types (the cell types were ontology supported groupings of original manual cell annotations), we found that distinct cell labels could explain some of the variability. While it is possible that there was some cofounder such as the source consortium or batch effect, this suggests that the cell type groupings might be too broad and cell type-TE enrichments are attached to finer cell type resolution.

It could also be that the variability has a functional purpose as it was shown for TEs involved in immunity[141]. To clarify the reality and significance of this variance, it may be interesting to use a dataset with many replicates per individual for a few cell types. With this, we could account for the inter-individual variability and the variability across replicates within a single individual. Furthermore, single cell sequencing technology could be leveraged to start get much finer cell type

resolution. With these we could avoid the aggregate measure from bulk samples and distinguish cell types within a tissue. For instance, to determine if there are differences between astrocytes, neurons and other brain cells.

4.4 TEs and their Association with Health and Diseases

Important applications of genomic studies include health and disease risk assessment and medical diagnostics. Since we had data from samples of healthy tissue, cancers and tissue with diseases, we leveraged that information to see if the health status was associated with any differences in TE enrichments. There have been many reports of TEs being associated with diseases[33,40,54,141], with some even being causal[55], but these are usually found in studies directly studying the disease. Here we reported many such associations across cell types and assays through an array like survey of all possibilities. While the set of samples featuring multiple health statuses (2024 samples) limited our power, we nonetheless found 4% that featured significant differences between health statuses. By case, we mean a comparison (Wilcoxon, pval <0.05) between the TE family enrichment of healthy samples and cancer or disease samples within a set cell type and assay. We found that the cancer samples generally had higher TE enrichment, which is in line with reports of TEs being more expressed in cancer[155]. Overall, we found the highest proportion of TEs featuring significant differences between health statuses within brain. This proportion was highest for the H3K4me1 (21/37, 57%, *differently enriched TE between health statuses / not differently enriched between health statuses*) mark and lowest for H3K9me3 (4/21, 19%). Other cell types with a notable proportion of TE enrichment differences between health statuses include monocytes, lymphocytes of B lineage and mesoderm-derived structure; which may support the involvement of TE in the

immune system[5,38,141,156]. Split by assay (each totaling 918 comparisons), we found 38:397 (cases of health difference: no health difference) in H3K27me3 samples, 67:358 for H3K4me1, 13:212 for H3K4me3, 45:308 for H3K27ac, 22:263 for H3K36me3 and 33:286 for H3K9me3. Given the limitations of the dataset (across all assays, more than 50% of cases were lacked the data to compare across health statuses (NA)), it is likely that our results underestimate the difference in association of TEs between healthy and disease or cancer data. Nonetheless, the results show there can be significant differences in TE enrichment depending on health status within a cell type. It would be interesting to see if the enrichments we've observed overlap disease associations found by GWAS. Some early work that was not taken further suggested that our TE candidates were associated relevant diseases according to GWAS, however we had not distinguished the health status of the samples. It could be worth revisiting our candidate selection and adding the layer of health status before assessing their trait association with GWAS and observing if health status leads to a difference in diseases and traits.

4.5 Identifying TEs with Potential Transcriptome Regulatory Function

Part of this work's objective was to identify TEs that may be co-opted or have genome regulatory functions. To do this we set up a system that highlighted TEs that were strongly associated with a cell type or were cell type specific and highly enriched. Through this we identified 456 potential regulatory TE candidates. A novel aspect of our selection strategy was to try and capture both observe-expected and fold change enrichment when selecting notable TEs. Usually projects select one of the two metrics, but we noticed that fold change and observed-expected highlighted different families. This is in part due to families having different instance counts and thus small

families reaching high fold changes more easily due to the low denominator. L1 and Alus generally did not reach high fold change, meanwhile it was clear that they systematically found in higher levels than expected (obs-exp).

Overall, some of the candidates that we highlighted recaptured existing research, such as MER11D in placenta[56,124] and Alus in brain[125,126], suggesting that some of the unsupported findings may also be worth investigating. According to Frost et al[56], the MER11D TEs possess GATA3 binding motifs and in the PSG region they are closely associated with enhancer H3K27ac, which we have also observed. One of the limitations of my experiments was that my research was fully computational, thus even through using other datasets or gene ontologies for function, the support I could provide was associations or correlations at best. Moving forward, it might be interesting to try some collaboration for wet lab confirmation of enhancer or promoter activity of some of our candidates.

4.6 Taking the TE Analysis to DNA Methylation

With Chapter 2 and 3 having found notable associations between TEs and histones, it would make sense to try a similar survey for DNA methylation, another important element of the epigenome. DNA Methylation is a repressive marker of the genome closely associated to gene expression. We performed such a survey on the epigenome data from the IHEC consortium as part of a figure and chapter contribution of the flagship publication of the EpiATLAS data reprocessing analysis (manuscript in preparation).

Our Study used 645 DNA methylation samples from the EpiATLAS dataset, which had already been converted into a matrix of percentage of coverage per CpG site for each sample. Since most samples had about 200 sites with no value (NA), we placed our selection threshold at 300 and

discard sites with more NA entries (Supplemental Figure S4.2 A). For the remaining sites, we selected a coverage of 80% for a CpG site to be considered methylated. Next, while considering only sites with >=80% coverage, we calculated the percentage of methylated samples with *methylated samples count (>=80%) / sample count (>0% <80%)* not including the NA values. We found a lot of sites with CpGs in 0 sample but also a lot of sites with CpG in a high (>95%) number of samples and we considered these constitutive (Supplemental Figure S4.2 B). We categorized sites based on how prevalent they were across samples: low for <25%, intermediate for >25% & <95% and high (or constitutive) for >95%. Thus, a CpG site categorized as high was methylated across nearly all samples while a site categorized as low was only methylated within less than 25% of samples. The majority of CpGs were of intermediate prevalence (but it had a much larger prevalence percent range), across the genome we found 4,083,000 constitutive, 17,093,368 intermediate and 3,701,616 low methylated CpG sites (Supplemental Figure S4.2 C).

We found that about 50% of all sites overlapped TEs, with a higher overlap for constitutive sites (~60%) (Supplemental Figure S4.3 A, Supplemental Figure S4.2 C, D). About 25% of sites that did not overlap TEs overlapped genes for constitutive and intermediate prevalence categories (Supplemental Figure S4.2 D). Low prevalence sites tended to be much less common in TEs and split between genes and other (non-characterized) regions. The TE families that overlapped CpG sites and their proportions were similar to those that overlapped the histones (for high and intermediate). However, some differences were Alu overlap being much higher in high prevalence CpG sites and MIR, L1 and ERVL-MaLR being lower in CpGs than histones (Supplemental Figure S4.3A,B). We next explored the position of sites relative to genes and found that the majority of sites were within 50kbp of gene TSS, and tended to have higher density close to the gene. Across the prevalence categories, the lowest prevalence sites tended to be concentrated near genes,

probably due to overlapping genes a lot. High and intermediate prevalence sites had much more similar profiles to each other. Interestingly, the only difference between CpGs overlapping TEs and those that did not was a dip near 0 for TE associated sites (Supplemental Figure S4.3C). This dip might be explained by TEs generally not overlapping genes

This preliminary work highlights a significant association between TEs and DNA methylation. Notably, while DNA methylation has a repressive role on TEs, its TE overlap profile is distinct from the repressive histone marks H3K27me3 and H3K9me3. The predominance of Alu in the constitutive CpG sites relative to the two histone mark and the absence of large L1 overlap found for H3K9me3 may suggest different repression mechanisms for these TE families. It would be interesting to take this work further and analyze these results at the subfamily level and to distinguish cell types as in the main chapters to see if cell type specificity is also observed here.

4.7 Applications of TE tool TEExplorer

One of the unexpected challenges that I faced within the experiments described in Chapter Large Scale Analysis of Transposable Elements Interaction with the Epigenome2 was that our analysis gave results for too many cases for us to cover. Our candidate selection process attempted to reduce all the relationships we observed to the most notable and potentially relevant ones. However, it was clear that not only was it possible for us to have missed important associations, but also that we were limited in the amount of relationships we could discuss within a publication. The TEExplorer tool allows users to query our results and investigate specific TE-hitstone-cell type relationships that they may be interested in. A common use case may be to confirm if an observation made has also been found in the EpiAtlas dataset. For instance, the tool was used to

identify if a specific L1 subfamily found to be enriched in lymphocytes could be supported by the epigenome. It was also used to look at if a certain ERV was enriched in any cell type or histone. Do note that the lack of specificity or result for the two examples is deliberate as they were not my own projects.

As exploratory investigation, TEExplorer was used to compare external macrophage data from patients with influenza to the EpiAtlas data. We found that the TE proportion was much higher in the influenza[157] dataset (subset of 8 H3K27ac samples AF04, AF06, EU03, EU05 infected and non-infected), however the broad profile of family enrichment or depletion was very similar between EpiAtlas and the external flue data. The specific TE subfamilies enrichment, we could determine that the two datasets had significant differences in enrichments (while the heatmap trends were similar, the boxplots highlighted significant differences). One major difference was the Alu subfamilies which were enriched in 4 subfamilies in the macrophages of the EpiAtlas data, but were not enriched for any of the shown families in the influenza data.

Chapter 5

Future Directions and conclusion

The work in this thesis has contributed to a better understanding of TEs and their relationship with the epigenome, however there were some limitations to the experiment and some results opened new questions. Some of these questions have been mentioned in the discussion: in this chapter I elaborate on some challenges that I think are worth exploring in the future.

5.1 Tackling the multi-mapped reads from a new angle

TEs are still held back by the problem of their repetitiveness and multi-mapped reads[89]. My experiments were no exception and did not consider multi-mapped reads. Long reads can help better place TEs and have a better idea of where the TEs are within a genome and assembly. However, a large amount of the publically available data are in post assembly and peak calling format BED, peaks etc. These provide genomic ranges of interest and may not have considered multi-mapped reads if the data generator did not go out of their way to. I believe that it would be interesting to be able to recapture the lost TEs from multi-mapped reads on all the currently available data which may have discarded them. While I can only speculate on the viability of the approach, I am interested in the idea of imputing BED files with multi-mapped reads re-inserted from BED files without them. With a large enough dataset it could be possible to compare the TEs detected with and without keeping multi-mapped reads and taking an imputation approach similar to ChromImpute[72] be able to predict the expected TEs with multi-mapped reads from a sample that discarded them. That is, if the multi-mapped reads exist in large enough quantity to support such an approach. My preliminary work has identified very low multi-mapped reads in human

epigenome data. Of course, approaches that can better handle multi-mapped reads would be instrumental to better TE data going forward, but I believe there might be overlooked value in finding a way to bring forward existing data.

5.2 Taking TEExplorer to new frontiers

TEExplorer is a useful tool and resource but there are a lot of features that I would like to add to a future version. The most important enhancement would be to expand the dataset. While the first version leverages data that I was already familiar with and had privileged access to, there is constantly more data being made available which could lead to better discoveries and insights. Even within the EpiATLAS dataset, I would be interested in adding the methylation dataset, thus TEExplorer would go beyond only characterizing histone marks. It may also be interesting to add the imputed histone marks dataset that were not initially available to me, but could provide much more additional data. Finally, I would be like to try to incorporate the ChromHMM[79] states in TEExplorer. While I was initially skeptical of their value relative to histone marks, I've come around to appreciate their interpretability and more nuanced state determination.

When it comes to user uploaded data, one of the limitation was that we do not create sample tailored controls for our enrichment measures. Instead we use an average of all the EpiATLAS samples matching the selected cell type and assay, which is much less computationally intensive. I believe that it works as an appropriate baseline, however I would like make it so user uploaded samples also get their own TSS distribution matching controls to make all the comparison fairer and avoid any dataset led biases. Another useful feature would be grouping of samples. While the TEExplorer allows comparisons between the user uploaded samples and the EpiATLAS dataset, it doesn't take into consideration that the user may have samples from multiple groups. It could be

useful to be able to assign groups to the user samples so that they can be compared to each other's (and the EpiAtlas data).

Moving forward I would also like to implement some form of result highlights to user uploaded analysis. This is to point out notable TE families or subfamilies similar to how I identified some TE candidates in Chapter 2. This change would improve accessibility because while the analysis does make available measurements for all subfamilies, if the user does not have a hypothesis, they might not know where to look or notice standout results.

Finally, while I set out to make this TEExplorer online for ease of access and use, I've also realized some of its downsides such as scalability, that is using TEExplorer on large amounts of user data, and the ability to integrate it within a larger pipeline. A standalone command line version of the method is something that I hope to work on.

5.3 Conclusion

In conclusion, we believed that that TEs had more epigenome associated regulatory functions than currently understood and that they were critical to our proper understanding of gene regulation and disease. We wanted to expand current knowledge of TEs and especially how they related to the epigenome. With our large scale analysis, we achieved our goal by presenting a comprehensive overview of the relationship between TEs and histone marks. We identified TE-cell type candidates some of which were already supported lending credence to viability of the unsupported ones. Finally, we developed TEExplorer to make this TE knowledge accessible and push forward TE analysis which should continue to evolve into the future. It is my hope that one day transposable elements will be characterized just as well as genes because I believe that even with everything that we've found, their known contribution will keep on expanding.

Bibliography

1. McClintock, B. The Origin and Behavior of Mutable Loci in Maize. *Proc Natl Acad Sci U S A* **36**, 344–355 (1950).
2. Orgel, L. E. & Crick, F. H. C. Selfish DNA: the ultimate parasite. *Nature* **284**, 604–607 (1980).
3. Rayan, N. A., del Rosario, R. C. H. & Prabhakar, S. Massive contribution of transposable elements to mammalian regulatory sequences. *Seminars in Cell & Developmental Biology* **57**, 51–56 (2016).
4. Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* **42**, 631–634 (2010).
5. Bogdan, L., Barreiro, L. & Bourque, G. Transposable elements have contributed human regulatory regions that are activated upon bacterial infection. *Philosophical Transactions of the Royal Society B: Biological Sciences* **375**, 20190332 (2020).
6. Barnada, S. M. *et al.* Genomic features underlie the co-option of SVA transposons as cis-regulatory elements in human pluripotent stem cells. *PLOS Genetics* **18**, e1010225 (2022).
7. Andrews, G. *et al.* Mammalian evolution of human cis-regulatory elements and transcription factor binding sites. *Science* **380**, eabn7930 (2023).
8. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
9. Senft, A. D. & Macfarlan, T. S. Transposable elements shape the evolution of mammalian development. *Nat Rev Genet* **22**, 691–711 (2021).

10. HICKMAN, A. B. & DYDA, F. Mechanisms of DNA Transposition. *Microbiol Spectr* **3**, MDNA3-0034–2014 (2015).
11. Hagemann, A. T. & Craig, N. L. Tn7 Transposition Creates a Hotspot for Homologous Recombination at the Transposon Donor Site. *Genetics* **133**, 9–16 (1993).
12. Luan, D. D., Korman, M. H., Jakubczak, J. L. & Eickbush, T. H. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* **72**, 595–605 (1993).
13. Cordaux, R. & Batzer, M. A. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10**, 691–703 (2009).
14. Pace, J. K. & Feschotte, C. The evolutionary history of human DNA transposons: Evidence for intense activity in the primate lineage. *Genome Res.* **17**, 422–432 (2007).
15. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0. (2013).
16. Deininger, P. L. & Batzer, M. A. Alu repeats and human disease. *Mol Genet Metab* **67**, 183–193 (1999).
17. Kazazian, H. H. *et al.* Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* **332**, 164–166 (1988).
18. Khan, H., Smit, A. & Boissinot, S. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Res* **16**, 78–87 (2006).
19. Carey, K. M., Patterson, G. & Wheeler, T. J. Transposable element subfamily annotation has a reproducibility problem. *Mobile DNA* **12**, 4 (2021).
20. Chen, X. *et al.* Cryptic endogenous retrovirus subfamilies in the primate lineage. 2023.12.07.570592 Preprint at https://doi.org/10.1101/2023.12.07.570592 (2023).

21. Chen, Y. *et al.* BERTE: High-precision hierarchical classification of transposable elements by a transfer learning method with BERT pre-trained model and convolutional neural network. 2024.01.28.577612 Preprint at https://doi.org/10.1101/2024.01.28.577612 (2024).
22. Szak, S. T. *et al.* Molecular archeology of L1 insertions in the human genome. *Genome Biology* **3**, research0052.1 (2002).
23. Myers, J. S. *et al.* A Comprehensive Analysis of Recently Integrated Human Ta L1 Elements. *The American Journal of Human Genetics* **71**, 312–326 (2002).
24. Bennett, E. A. *et al.* Active Alu retrotransposons in the human genome. *Genome Res.* **18**, 1875–1883 (2008).
25. Konkel, M. K. *et al.* Sequence Analysis and Characterization of Active Human Alu Subfamilies Based on the 1000 Genomes Pilot Project. *Genome Biology and Evolution* **7**, 2608–2622 (2015).
26. Wang, H. *et al.* SVA Elements: A Hominid-specific Retroposon Family. *Journal of Molecular Biology* **354**, 994–1007 (2005).
27. Krull, M., Petrusma, M., Makalowski, W., Brosius, J. & Schmitz, J. Functional persistence of exonized mammalian-wide interspersed repeat elements (MIRs). *Genome Res.* **17**, 1139–1145 (2007).
28. Waterson, R. H., Lander, E. S., Wilson, R. K., & The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69–87 (2005).
29. Han, K. *et al.* Under the genomic radar: The Stealth model of Alu amplification. *Genome Res* **15**, 655–664 (2005).

30. Li, T.-H. & Schmid, C. W. Alu's dimeric consensus sequence destabilizes its transcripts. *Gene* **324**, 191–200 (2004).

31. Xing, J. *et al.* Mobile elements create structural variation: Analysis of a complete human genome. *Genome Res* **19**, 1516–1526 (2009).

32. Sultana, T., Zamborlini, A., Cristofari, G. & Lesage, P. Integration site selection by retroviruses and transposable elements in eukaryotes. *Nat Rev Genet* **18**, 292–308 (2017).

33. Bourque, G. *et al.* Ten things you should know about transposable elements. *Genome Biology* **19**, (2018).

34. Hancks, D. C. & Kazazian, H. H. Roles for retrotransposon insertions in human disease. *Mobile DNA* **7**, 9 (2016).

35. Gotea, V. & Makałowski, W. Do transposable elements really contribute to proteomes? *Trends Genet* **22**, 260–267 (2006).

36. Moran, J. V., DeBerardinis, R. J. & Kazazian, H. H. Exon Shuffling by L1 Retrotransposition. *Science* **283**, 1530–1534 (1999).

37. Hedges, D. J. & Deininger, P. L. Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **616**, 46–59 (2007).

38. Kassiotis, G. & Stoye, J. P. Immune responses to endogenous retroelements: taking the bad with the good. *Nat Rev Immunol* **16**, 207–219 (2016).

39. Chiappinelli, K. B. *et al.* Inhibiting DNA Methylation Causes an Interferon Response in Cancer via dsRNA Including Endogenous Retroviruses. *Cell* **162**, 974–986 (2015).

40. Scott, E. C. *et al.* A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. *Genome Res.* **26**, 745–755 (2016).

41. Enriquez-Gasca, R. *et al.* Co-option of endogenous retroviruses through genetic escape from TRIM28 repression. *Cell Reports* **42**, (2023).

42. Maksakova, I. A. *et al.* H3K9me3-binding proteins are dispensable for SETDB1/H3K9me3-dependent retroviral silencing. *Epigenetics & Chromatin* **4**, 12 (2011).

43. Pezic, D., Manakov, S. A., Sachidanandam, R. & Aravin, A. A. piRNA pathway targets active LINE1 elements to establish the repressive H3K9me3 mark in germ cells. *Genes Dev.* **28**, 1410–1428 (2014).

44. Wang, Z. *et al.* Dominant role of DNA methylation over H3K9me3 for IAP silencing in endoderm. *Nat Commun* **13**, 5447 (2022).

45. Ecco, G., Imbeault, M. & Trono, D. A tale of domestication: the endovirome, its polydactyl controllers and the species-specificity of human biology. *Development* **144**, 2719–2729 (2017).

46. Imbeault, M., Helleboid, P.-Y. & Trono, D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**, 550–554 (2017).

47. Najafabadi, H. S. *et al.* C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nat Biotechnol* **33**, 555–562 (2015).

48. Ahmadi, A., De Toma, I., Vilor-Tejedor, N., Eftekhariyan Ghamsari, M. R. & Sadeghi, I. Transposable elements in brain health and disease. *Ageing Research Reviews* **64**, 101153 (2020).

49. Erwin, J. A., Marchetto, M. C. & Gage, F. H. Mobile DNA elements in the generation of diversity and complexity in the brain. *Nature Reviews Neuroscience* **15**, 497–506 (2014).

50. Evans, T. A. & Erwin, J. A. Retroelement-derived RNA and its role in the brain. *Seminars in Cell & Developmental Biology* **114**, 68–80 (2021).

51. Baillie, J. K. *et al.* Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* **479**, 534–537 (2011).

52. Bedrosian, T. A., Quayle, C., Novaresi, N. & Gage, Fred. H. Early life experience drives structural variation of neural genomes in mice. *Science* **359**, 1395–1399 (2018).

53. Bundo, M. *et al.* Increased L1 Retrotransposition in the Neuronal Genome in Schizophrenia. *Neuron* **81**, 306–313 (2014).

54. Li, W. *et al.* Human endogenous retrovirus-K contributes to motor neuron disease. *Sci Transl Med* **7**, 307ra153 (2015).

55. Miki, Y. *et al.* Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer Res* **52**, 643–645 (1992).

56. Frost, J. M. *et al.* Regulation of human trophoblast gene expression by endogenous retroviruses. 2022.04.26.489485 Preprint at https://doi.org/10.1101/2022.04.26.489485 (2022).

57. Russanova, V. R., Driscoll, C. T. & Howard, B. H. Adenovirus type 2 preferentially stimulates polymerase III transcription of Alu elements by relieving repression: a potential role for chromatin. *Mol Cell Biol* **15**, 4282–4290 (1995).

58. Stenz, L. The L1-dependant and Pol III transcribed Alu retrotransposon, from its discovery to innate immunity. *Mol Biol Rep* **48**, 2775–2789 (2021).

59. Ahmad, S. *et al.* Breaching Self-Tolerance to Alu Duplex RNA Underlies MDA5-Mediated Inflammation. *Cell* **172**, 797-810.e13 (2018).

60. Chen, X. *et al.* Transposable elements are associated with the variable response to influenza infection. *Cell Genomics* **3**, (2023).

61. Ardeljan, D., Taylor, M. S., Ting, D. T. & Burns, K. H. The Human Long Interspersed Element-1 Retrotransposon: An Emerging Biomarker of Neoplasia. *Clinical Chemistry* **63**, 816–822 (2017).

62. Xia, B. *et al.* On the genetic basis of tail-loss evolution in humans and apes. *Nature* **626**, 1042–1048 (2024).

63. Kamitaki, N. *et al.* A sequence of SVA retrotransposon insertions in ASIP shaped human pigmentation. *Nat Genet* 1–9 (2024) doi:10.1038/s41588-024-01841-4.

64. Allis, C. D. & Jenuwein, T. The molecular hallmarks of epigenetic control. *Nat Rev Genet* **17**, 487–500 (2016).

65. Lee, H.-T., Oh, S., Ro, D. H., Yoo, H. & Kwon, Y.-W. The Key Role of DNA Methylation and Histone Acetylation in Epigenetics of Atherosclerosis. *J Lipid Atheroscler* **9**, 419–434 (2020).

66. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

67. Bhattacharya, S. *et al.* The methyltransferase SETD2 couples transcription and splicing by engaging mRNA processing factors through its SHI domain. *Nat Commun* **12**, 1443 (2021).

68. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).

69. Bernstein, B. E. *et al.* A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell* **125**, 315–326 (2006).

70. Azuara, V. *et al.* Chromatin signatures of pluripotent cell lines. *Nat Cell Biol* **8**, 532–538 (2006).

71. Vastenhouw, N. L. & Schier, A. F. Bivalent histone modifications in early embryogenesis. *Curr. Opin. Cell Biol.* **24**, 374–386 (2012).

72. Ernst, J. & Kellis, M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat Biotechnol* **33**, 364–376 (2015).

73. Radman-Livaja, M. *et al.* Patterns and mechanisms of ancestral histone protein inheritance in budding yeast. *PLoS Biol* **9**, e1001075 (2011).

74. Moazed, D. Mechanisms for the Inheritance of Chromatin States. *Cell* **146**, 510–518 (2011).

75. Jin, B., Li, Y. & Robertson, K. D. DNA Methylation: Superior or Subordinate in the Epigenetic Hierarchy? *Genes & Cancer* **2**, 607–617 (2011).

76. Lehnertz, B. *et al.* Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. *Curr Biol* **13**, 1192–1200 (2003).

77. Loyfer, N. *et al.* A DNA methylation atlas of normal human cell types. *Nature* **613**, 355–364 (2023).

78. Ziller, M. J. *et al.* Charting a dynamic DNA methylation landscape of the human genome. *Nature* **500**, 477–481 (2013).

79. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**, 215–216 (2012).

80. Jacques, P.-É., Jeyakani, J. & Bourque, G. The Majority of Primate-Specific Regulatory Sequences Are Derived from Transposable Elements. *PLoS Genet* **9**, e1003504 (2013).

81. Ross, J. P., Rand, K. N. & Molloy, P. L. Hypomethylation of Repeated Dna Sequences in Cancer. *Epigenomics* **2**, 245–269 (2010).

82. Kulis, M. & Esteller, M. 2 - DNA Methylation and Cancer. in *Advances in Genetics* (eds. Herceg, Z. & Ushijima, T.) vol. 70 27–56 (Academic Press, 2010).

83. Xie, H. *et al.* Epigenomic analysis of Alu repeats in human ependymomas. *Proceedings of the National Academy of Sciences* **107**, 6952–6957 (2010).

84. Jordà, M. *et al.* The epigenetic landscape of Alu repeats delineates the structural and functional genomic architecture of colon cancer cells. *Genome Res.* **27**, 118–132 (2017).

85. Popp, C. *et al.* Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. *Nature* **463**, 1101–1105 (2010).

86. Pehrsson, E. C., Choudhary, M. N. K., Sundaram, V. & Wang, T. The epigenomic landscape of transposable elements across normal human development and anatomy. *Nat Commun* **10**, 1–16 (2019).

87. Zhuo, X., Du, A. Y., Pehrsson, E. C., Li, D. & Wang, T. Epigenomic differences in the human and chimpanzee genomes are associated with structural variation. *Genome Res* (2020) doi:10.1101/gr.263491.120.

88. Fuentes, D. R., Swigut, T. & Wysocka, J. Systematic perturbation of retroviral LTRs reveals widespread long-range effects on human gene regulation. *eLife* **7**, e35989 (2018).

89. Goerner-Potvin, P. & Bourque, G. Computational tools to unmask transposable elements. *Nature Reviews Genetics* **19**, 688–704 (2018).

90. Lerat, E., Casacuberta, J., Chaparro, C. & Vieira, C. On the Importance to Acknowledge Transposable Elements in Epigenomic Analyses. *Genes* **10**, 258 (2019).

91. Elliott, T. A. *et al.* TE Hub: A community-oriented space for sharing and connecting tools, data, resources, and methods for transposable element annotation. *Mobile DNA* **12**, 16 (2021).

92. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

93. Fernandes, J. D. *et al.* The UCSC repeat browser allows discovery and visualization of evolutionary conflict across repeat families. *Mobile DNA* **11**, 13 (2020).

94. Bujold, D. *et al.* The International Human Epigenome Consortium Data Portal. *cels* **3**, 496-499.e2 (2016).

95. The GTEx Consortium *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).

96. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580–585 (2013).

97. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).

98. Zou, Z., Ohta, T. & Oki, S. ChIP-Atlas 3.0: a data-mining suite to explore chromosome architecture together with large-scale regulome data. *Nucleic Acids Research* **52**, W45–W53 (2024).

99. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**, 495–501 (2010).

100. Tanigawa, Y., Dyer, E. S. & Bejerano, G. WhichTF is functionally important in your open chromatin data? *PLOS Computational Biology* **18**, e1010378 (2022).

101. Finnegan, D. J. Eukaryotic transposable elements and genome evolution. *Trends in Genetics* **5**, 103–107 (1989).

102. Wells, J. N. & Feschotte, C. A Field Guide to Eukaryotic Transposable Elements. *Annu Rev Genet* **54**, 539–561 (2020).

103. Ye, M. *et al.* Specific subfamilies of transposable elements contribute to different domains of T lymphocyte enhancers. *Proceedings of the National Academy of Sciences* **117**, 7905–7916 (2020).

104. Smit, A. F. A., Tóth, G., Riggs, A. D. & Jurka, J. Ancestral, Mammalian-wide Subfamilies of LINE-1 Repetitive Sequences. *Journal of Molecular Biology* **246**, 401–417 (1995).

105. Su, M., Han, D., Boyd-Kirkup, J., Yu, X. & Han, J.-D. J. Evolution of Alu Elements toward Enhancers. *Cell Reports* **7**, 376–385 (2014).

106. Ohnuki, M. *et al.* Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 12426–12431 (2014).

107. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science (New York, N.Y.)* **351**, 1083 (2016).

108. Feliciello, I., Sermek, A., Pezer, Ž., Matulić, M. & Ugarković, Đ. Heat Stress Affects H3K9me3 Level at Human Alpha Satellite DNA Repeats. *Genes* **11**, 663 (2020).

109. Mosch, K., Franz, H., Soeroes, S., Singh, P. B. & Fischle, W. HP1 Recruits Activity-Dependent Neuroprotective Protein to H3K9me3 Marked Pericentromeric Heterochromatin for Silencing of Major Satellite Repeats. *PLOS ONE* **6**, e15894 (2011).

110. Jjingo, D. *et al.* Mammalian-wide interspersed repeat (MIR)-derived enhancers and the regulation of human gene expression. *Mobile DNA* **5**, 14 (2014).

111. Cao, Y. *et al.* Widespread roles of enhancer-like transposable elements in cell identity and long-range genomic interactions. *Genome Res* **29**, 40–52 (2019).

112. International Human Epigenome Consortium, EpiATLAS - a reference for human epigenomic research. *In preperation*.

113. Moore, J. E. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).

114. Allshire, R. C. & Madhani, H. D. Ten principles of heterochromatin formation and function. *Nat Rev Mol Cell Biol* **19**, 229–244 (2018).

115. Kabi, M. & Filion, G. J. Heterochromatin: did H3K9 methylation evolve to tame transposons? *Genome Biology* **22**, 325 (2021).

116. Karimzadeh, M., Ernst, C., Kundaje, A. & Hoffman, M. M. Umap and Bismap: quantifying genome and methylome mappability. *Nucleic Acids Research* **46**, e120 (2018).

117. Seczynska, M., Bloor, S., Cuesta, S. M. & Lehner, P. J. Genome surveillance by HUSH-mediated silencing of intronless mobile elements. *Nature* 1–9 (2021) doi:10.1038/s41586-021-04228-1.

118. Collins, P. L., Kyle, K. E., Egawa, T., Shinkai, Y. & Oltz, E. M. The histone methyltransferase SETDB1 represses endogenous and exogenous retroviruses in B lymphocytes. *Proceedings of the National Academy of Sciences* **112**, 8367–8372 (2015).

119. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet* **18**, 71–86 (2017).

120. Blumenstiel, J. P. Birth, School, Work, Death, and Resurrection: The Life Stages and Dynamics of Transposable Element Proliferation. *Genes* **10**, 336 (2019).

121. Hunter, R. G., McEwen, B. S. & Pfaff, D. W. Environmental stress and transposon transcription in the mammalian brain. *Mobile Genetic Elements* **3**, e24555 (2013).

122. Erwin, J. A. *et al.* L1-associated genomic regions are deleted in somatic cells of the healthy human brain. *Nat Neurosci* **19**, 1583–1591 (2016).

123. Muotri, A. R. *et al.* Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* **435**, 903–910 (2005).

124. Bi, S., Gavrilova, O., Gong, D.-W., Mason, M. M. & Reitman, M. Identification of a Placental Enhancer for the Human Leptin Gene *. *Journal of Biological Chemistry* **272**, 30583–30588 (1997).

125. Jacob-Hirsch, J. *et al.* Whole-genome sequencing reveals principles of brain retrotransposition in neurodevelopmental disorders. *Cell Res* **28**, 187–203 (2018).

126. Mehler, M. F. & Mattick, J. S. Noncoding RNAs and RNA Editing in Brain Development, Functional Diversification, and Neurological Disease. *Physiological Reviews* **87**, 799–823 (2007).

127. Gu, Z. & Hübschmann, D. rGREAT: an R/bioconductor package for functional enrichment on genomic regions. *Bioinformatics* **39**, btac745 (2023).

128. Fakhr, Y., Brindley, D. N. & Hemmings, D. G. Physiological and pathological functions of sphingolipids in pregnancy. *Cellular Signalling* **85**, 110041 (2021).

129. Enthoven, L. F. *et al.* Effects of Pregnancy on Plasma Sphingolipids Using a Metabolomic and Quantitative Analysis Approach. *Metabolites* **13**, 1026 (2023).

130. Nit, K., Tyszka-Czochara, M. & Bobis-Wozowicz, S. Oxygen as a Master Regulator of Human Pluripotent Stem Cell Function and Metabolism. *Journal of Personalized Medicine* **11**, 905 (2021).

131. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

132. Bogdan, L., Barreiro, L. & Bourque, G. Transposable elements have contributed human regulatory regions that are activated upon bacterial infection. 9.

133. Kojima, S. *et al.* Mobile elements in human population-specific genome and phenotype divergence. 2022.03.25.485726 Preprint at https://doi.org/10.1101/2022.03.25.485726 (2022).

134. Barnada, S. M. *et al. Genomic Features Underlie the Co-Option of SVA Transposons as Cis-Regulatory Elements in Human Pluripotent Stem Cells*. 2022.01.10.475682 https://www.biorxiv.org/content/10.1101/2022.01.10.475682v1 (2022) doi:10.1101/2022.01.10.475682.

135. Saksouk, N., Simboeck, E. & Déjardin, J. Constitutive heterochromatin formation and transcription in mammals. *Epigenetics and Chromatin* **8**, (2015).

136. Mills, R. E., Bennett, E. A., Iskow, R. C. & Devine, S. E. Which transposable elements are active in the human genome? *Trends in Genetics* **23**, 183–191 (2007).

137. Cowley, M. & Oakey, R. J. Transposable Elements Re-Wire and Fine-Tune the Transcriptome. *PLOS Genetics* **9**, e1003234 (2013).

138. Du, A. Y., Chobirko, J. D., Zhuo, X., Feschotte, C. & Wang, T. Regulatory Transposable Elements in the Encyclopedia of DNA Elements. 2023.09.05.556380 Preprint at https://doi.org/10.1101/2023.09.05.556380 (2023).

139. Bogu, G. K., Reverter, F., Marti-Renom, M. A., Snyder, M. P. & Guigó, R. Atlas of transcriptionally active transposable elements in human adult tissues. *bioRxiv* 714212 (2019) doi:10.1101/714212.

140. Kong, Y. *et al.* Transposable element expression in tumors is associated with immune infiltration and increased antigenicity. *Nature Communications* **10**, 5228 (2019).

141. Chen, X. *et al.* Transposable elements are associated with the variable response to influenza infection. 2022.05.10.491101 Preprint at https://doi.org/10.1101/2022.05.10.491101 (2022).

142. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

143. Zhuang, Q. K.-W. *et al.* Sex Chromosomes and Sex Phenotype Contribute to Biased DNA Methylation in Mouse Liver. *Cells* **9**, 1436 (2020).

144. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing,.

145. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).

146. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).

147. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

148. Hyacinthe, J. & Bourque, G. Transposable elements impact the regulatory landscape through cell type specific epigenomic associations. 2024.08.07.606967 Preprint at https://doi.org/10.1101/2024.08.07.606967 (2024).

149. Müller, K., Wickham, H., James, D. A. & Falcon, S. *RSQLite: SQLite Interface for R*. (2024).

150. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at http://arxiv.org/abs/1802.03426 (2020).

151. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLOS Computational Biology* **9**, e1003118 (2013).

152. Perrier, V., Meyer, F. & Granjon, D. *ShinyWidgets: Custom Inputs Widgets for Shiny*. (2024).

153. Gal-Mark, N., Schwartz, S. & Ast, G. Alternative splicing of Alu exons—two arms are better than one. *Nucleic Acids Res* **36**, 2012–2023 (2008).

154. Payer, L. M. *et al.* Alu insertion variants alter mRNA splicing. *Nucleic Acids Res* **47**, 421–431 (2019).

155. Criscione, S. W., Zhang, Y., Thompson, W., Sedivy, J. M. & Neretti, N. Transcriptional landscape of repetitive elements in normal and cancer human cells. *BMC Genomics* **15**, 583 (2014).

156. Zhu, X., Fang, H., Gladysz, K., Barbour, J. A. & Wong, J. W. H. Overexpression of transposable elements is associated with immune evasion and poor outcome in colorectal cancer. *European Journal of Cancer* **157**, 94–107 (2021).

157. Aracena, K. A. *et al.* Epigenetic variation impacts individual differences in the transcriptional response to influenza infection. *Nat Genet* **56**, 408–419 (2024).

Appendices

Appendix A

The list of contribution of the thesis author to other works.

Zhuang, Q. K.-W.; Galvez, J. H.; Xiao, Q.; AlOgayil, N.; **Hyacinthe, J**.; Taketo, T.; Bourque, G.; Naumova, A. K. Sex Chromosomes and Sex Phenotype Contribute to Biased DNA Methylation in Mouse Liver. *Cells* **2020**, *9* (6), 1436. https://doi.org/10.3390/cells9061436.

Manuscript under preparation:

IHEC Consortium*, The EpiAtlas dataset release (working title)

Jeffrey Hyacinthe is an author member of the IHEC consortium

Appendix B

Supplementary material for Chapter 2 and its corresponding manuscript: Transposable elements impact the regulatory landscape through cell type specific epigenomic associations

Supplemental Figures

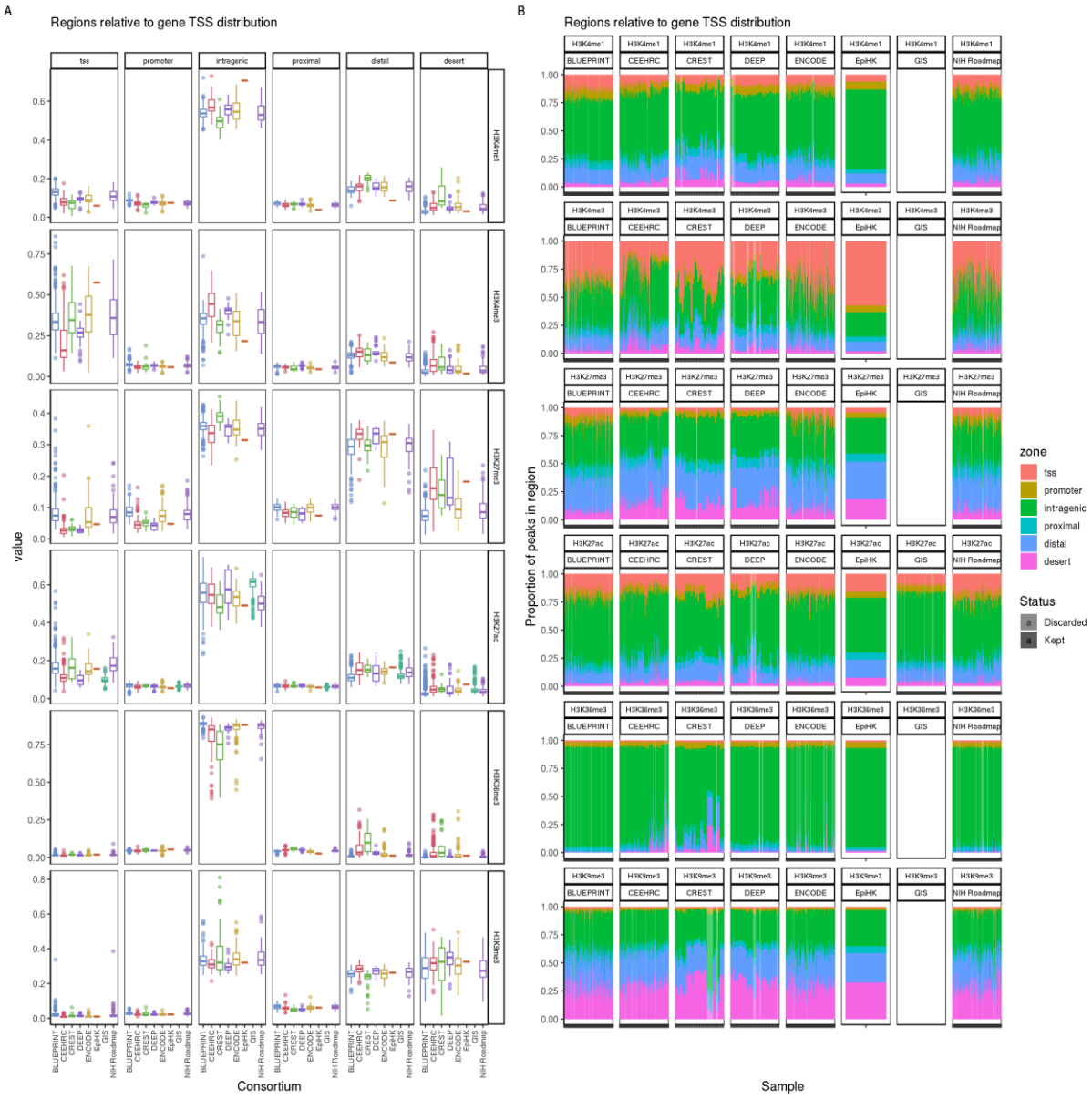

Figure S2. 1 Distribution of the peaks in regions relative to TSS

A) Distribution of peaks within annotations relative to gene TSS. Horizontal facets are the different regions, the Y axis is the proportion found within the region **B)** X axis are the samples faceted by their assay and consortium. Y axis is the proportion found within the colored regions. Color are the different regions, transparent samples are those discarded due to outlier samples.

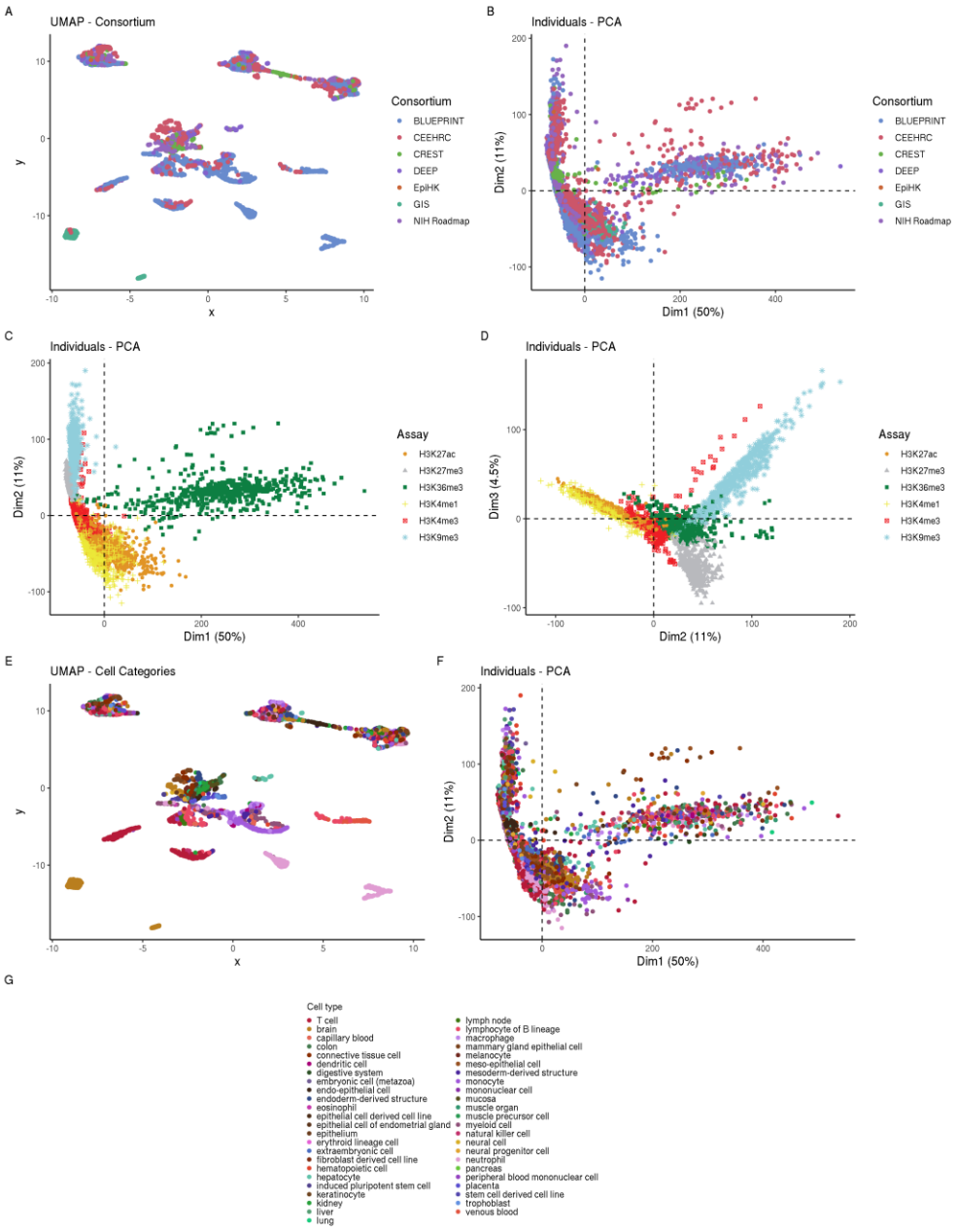

Figure S2. 2 ChIP-Seq samples data dimension reduction

A) UMAP on peak count within 10kb regions across full genome of the 4614 samples. Only the 20,000 windows with most variance (excluding the first 1000) were used colored by consortium **B)** PCA of first 2 PCs using same data as A colored by Sample's Consortium **C)** colored by Sample's assay **D)** PCA of 2nd and 3rd PC colored by sample's assay **E)** Same as A colored by cell type **F)**PCA of first 2 PCs as in B and C colored by Cell type **G)** Cell type Color legend of E and F

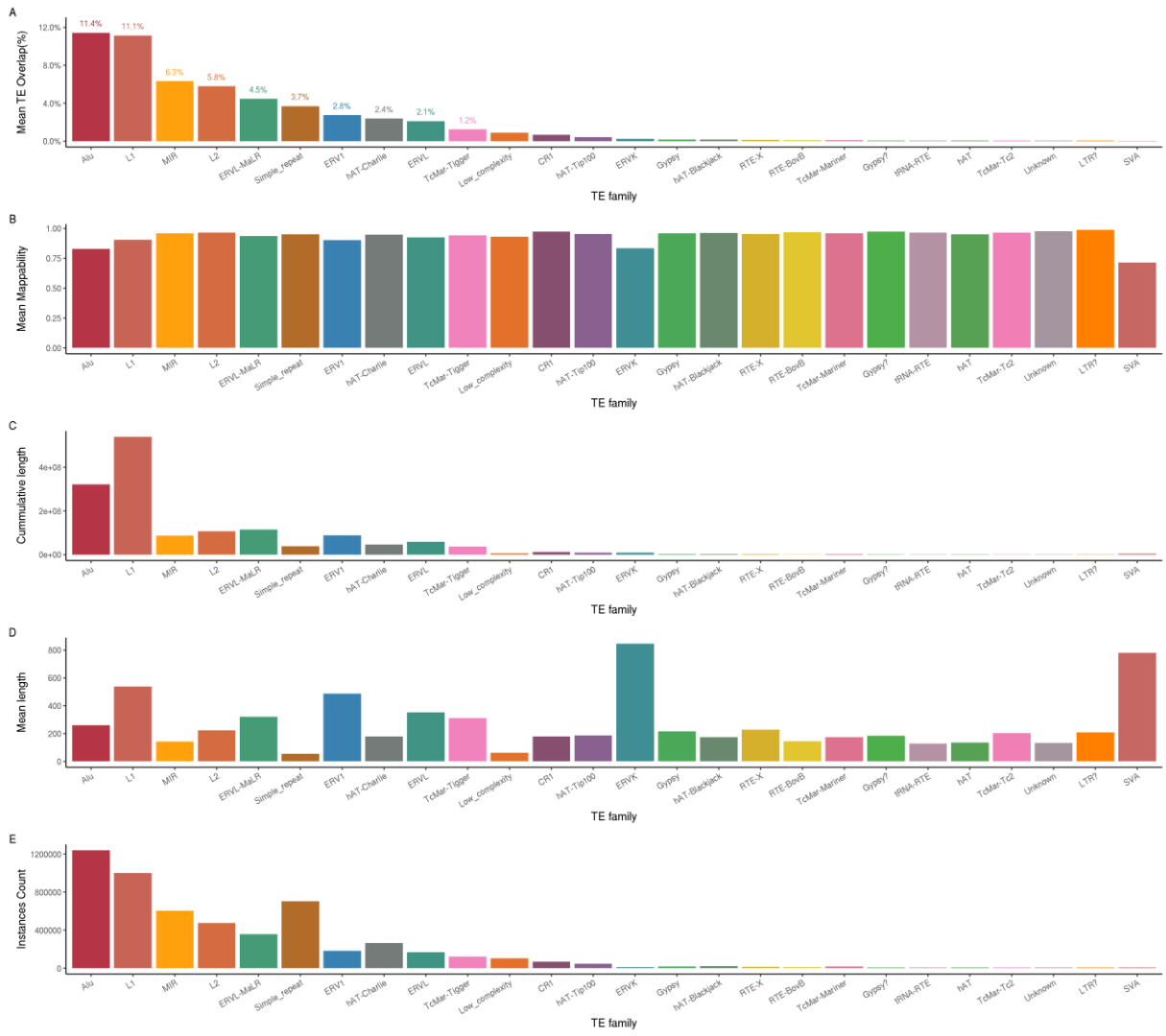

Figure S2. 3 TE Family Overlap and TE family properties.
A) Mean TE overlap across all samples **B)** TE Mappability estimates **C)** Cummulative bp length within the genome **D)** Mean TE instance length **E)** TE instance count. X axis sorted according to TE overlap(A). Only families containing at least 1000 instances shown. Centr family excluded due to extreme outlier mean length.

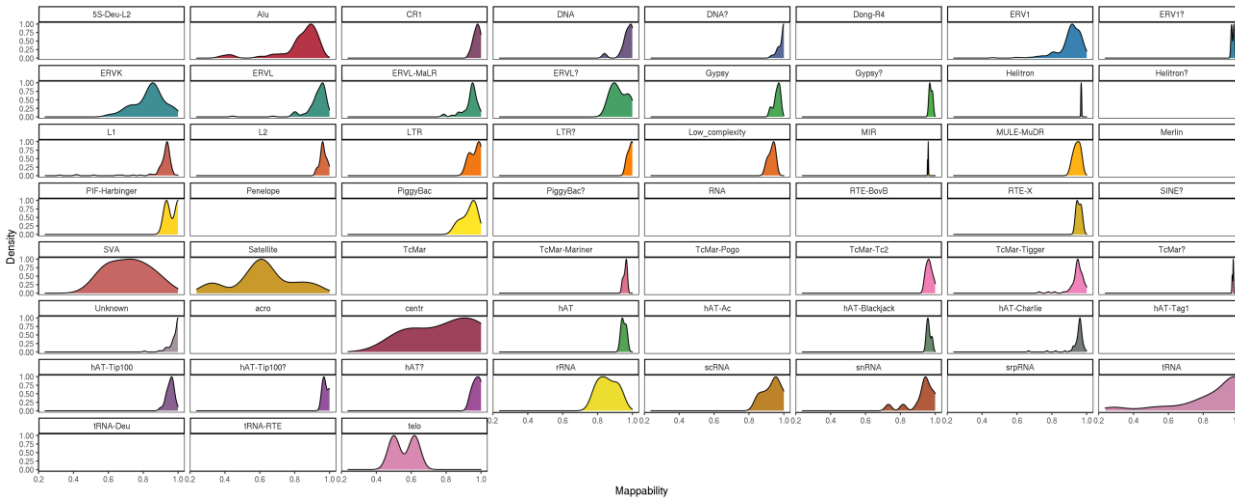

Figure S2. 4 TE Families Mappability distribution

Density distribution of TE instances Mappability (TE coverage by the 50bp Unique Mappability (Umap 50) track) within the TE families. (Empty if not enough instances)

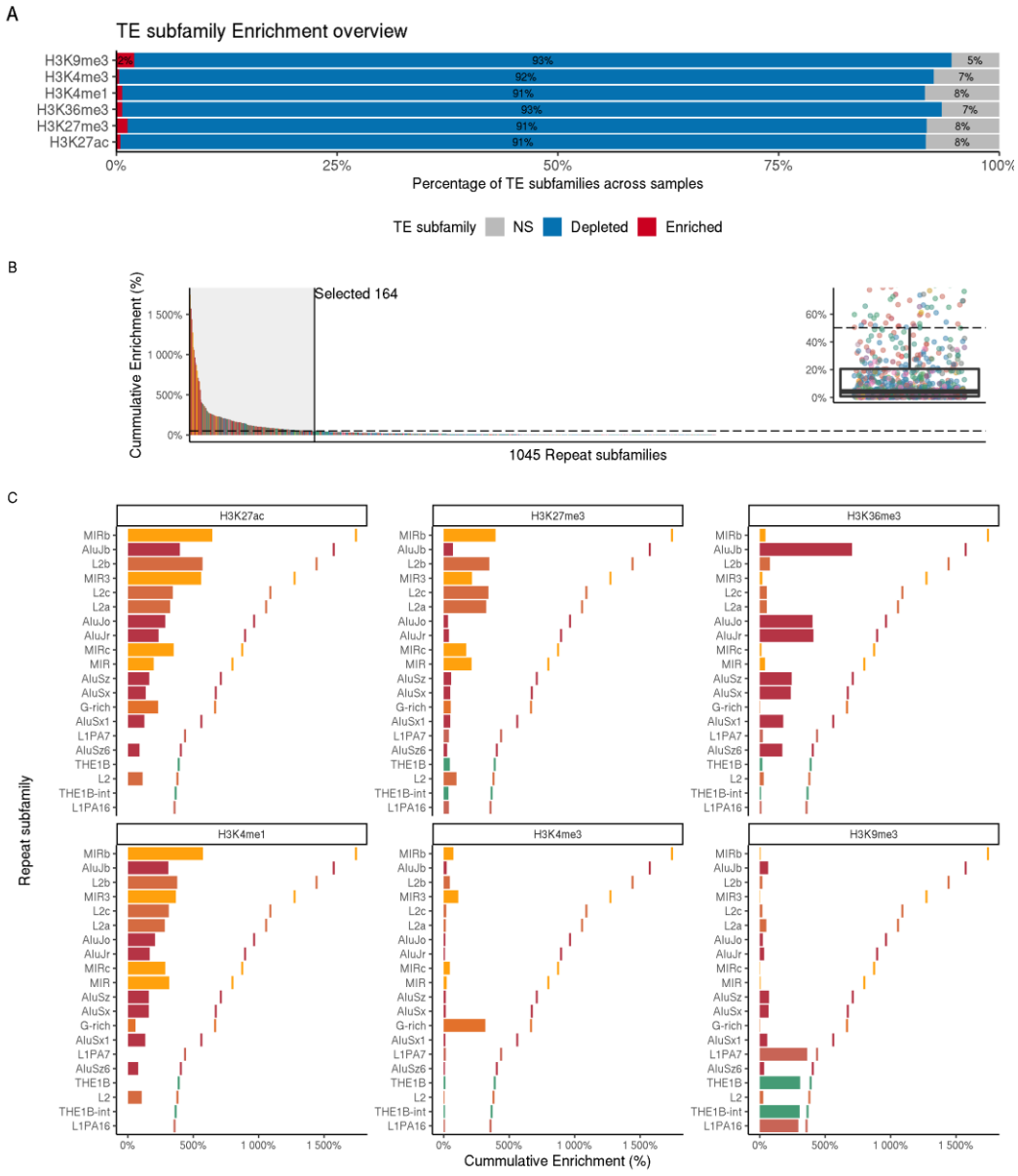

Figure S2. 5 Breakdown of TE Enrichments Within Subfamilies

A) Percentage of TE subfamilies enriched or depleted relative to random simulation across assays. Each TE subfamily is enrichment is recorded individually for each sample. NS are non significantly enriched TEs **B)**Sorted cumulative sum of the 1045 TE subfamilies enrichment (obs-exp, significantly enriched, and therefore >0, only) across all samples. The 164 TE subfamilies above the threshold line (dashed line) were selected for the downstream analysis. The threshold was selected as the upper whisker from the boxplot of the cumulative sum of TE enrichment **C)** Top 20 of Cummulative sum of TE subfamilies genome overlap across all samples grouped by Histone mark. Vertical line is the sum of the enrichment across all assays. Thus, the bar size relative to the vertical line represents the proportion of the enrichment coming from the given assay.

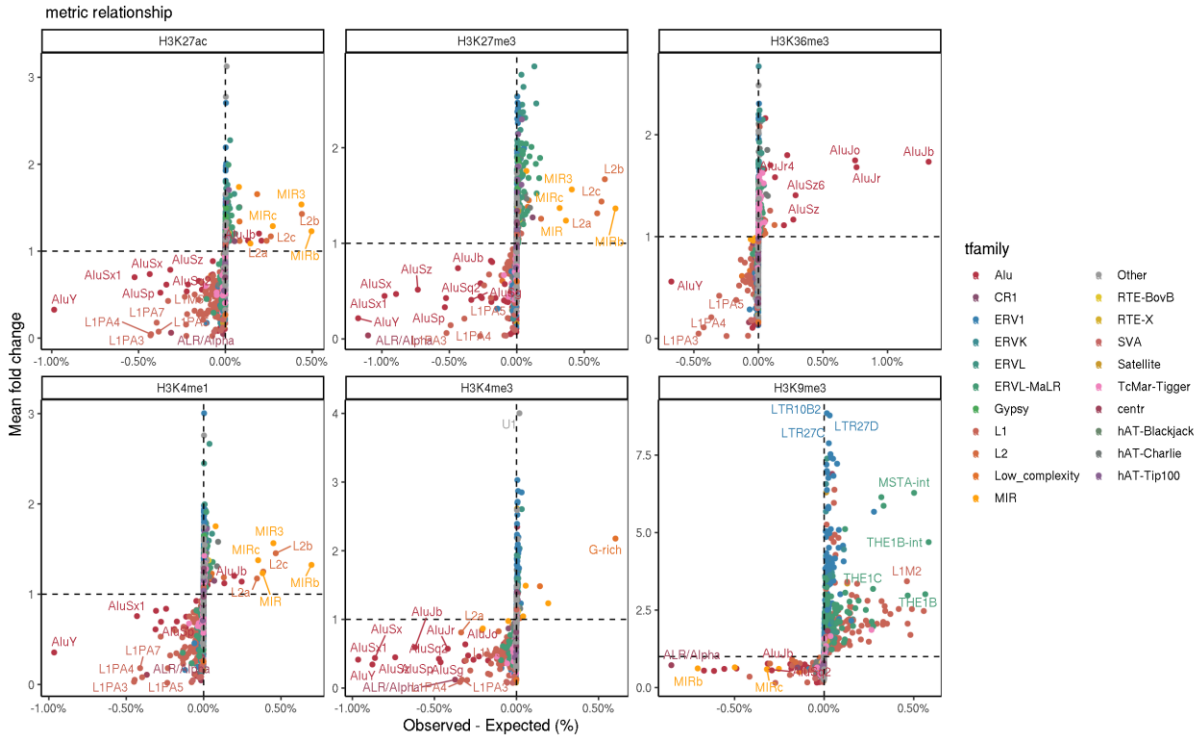

Figure S2. 6 Fold change Enrichment in function of Obs-Exp

Mean Fold change in function of observed – expected enrichment of TE subfamilies (mean across cell types). Colored by TE family.

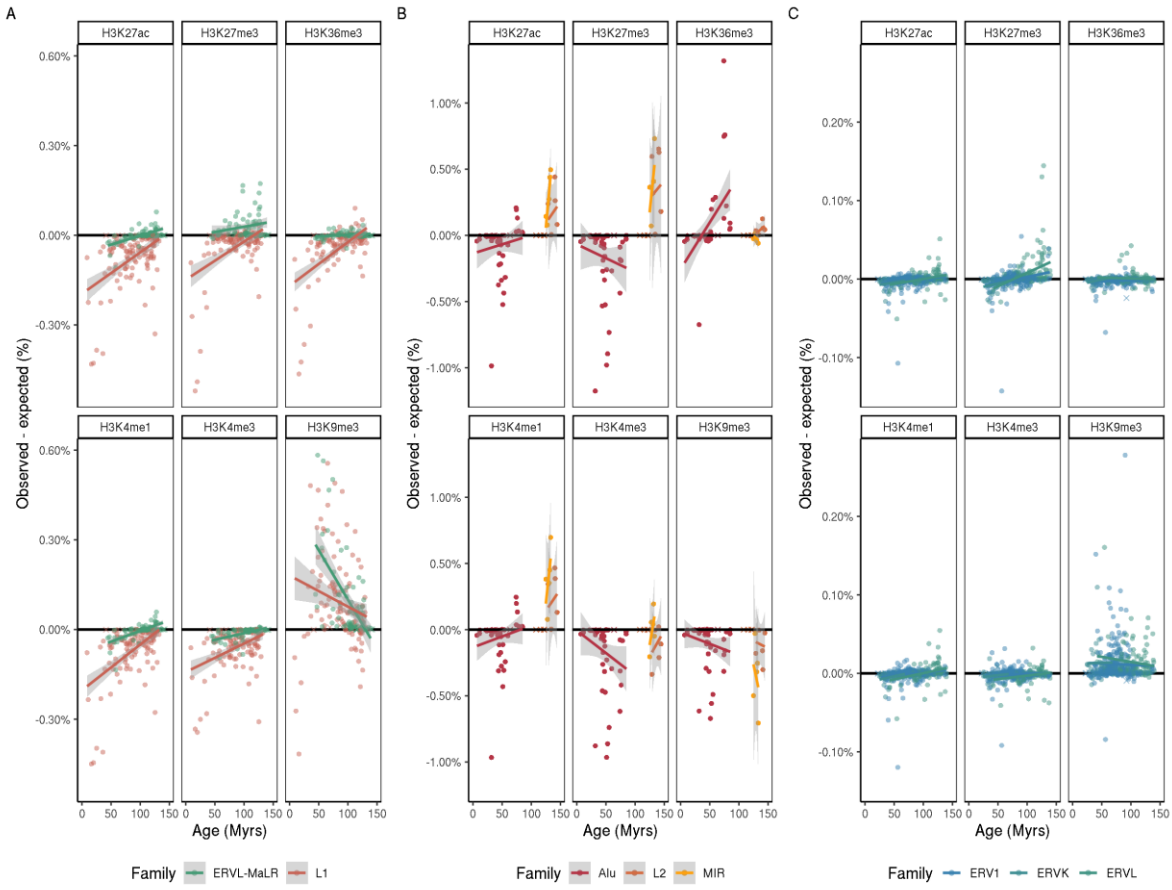

Figure S2. 7 Expanded regression set

A) TE Family enrichment (obs-expected%) in function of estimated age for L1, ERVL-MaLR across all 6 Histone marks. Black line is at 0 enrichment. Line shows linear regression fit, crosses are small sized subfamilies excluded from regression. **B)** For Alu,L2 and MIR **C)** For ERV1, ERVK and ERVL

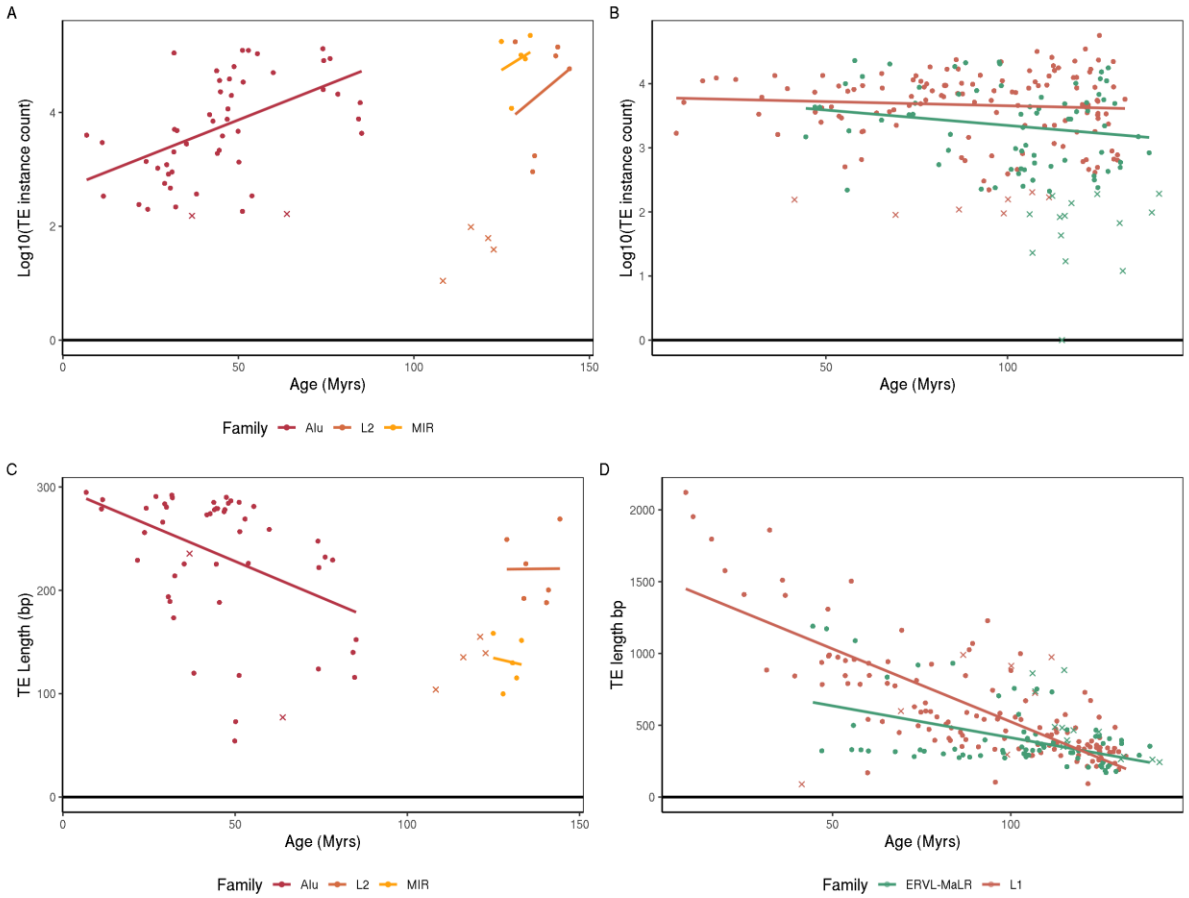

Figure S2. 8 Potential TE properties cofounding with age

A) Mean in instance count of TE subfamilies in function of their mean age estimate for Alu, MIR and L2. **B)** same as A for L1 and ERVL-MaLR. **C)** Mean TE Length of TE subfamilies in function of their mean age estimates for Alu, L2 and MIR. **D)** same as C for L1 and ERVL MaLR. X shapes are subfamilies too small (less than) in instance count and were not used for regression.

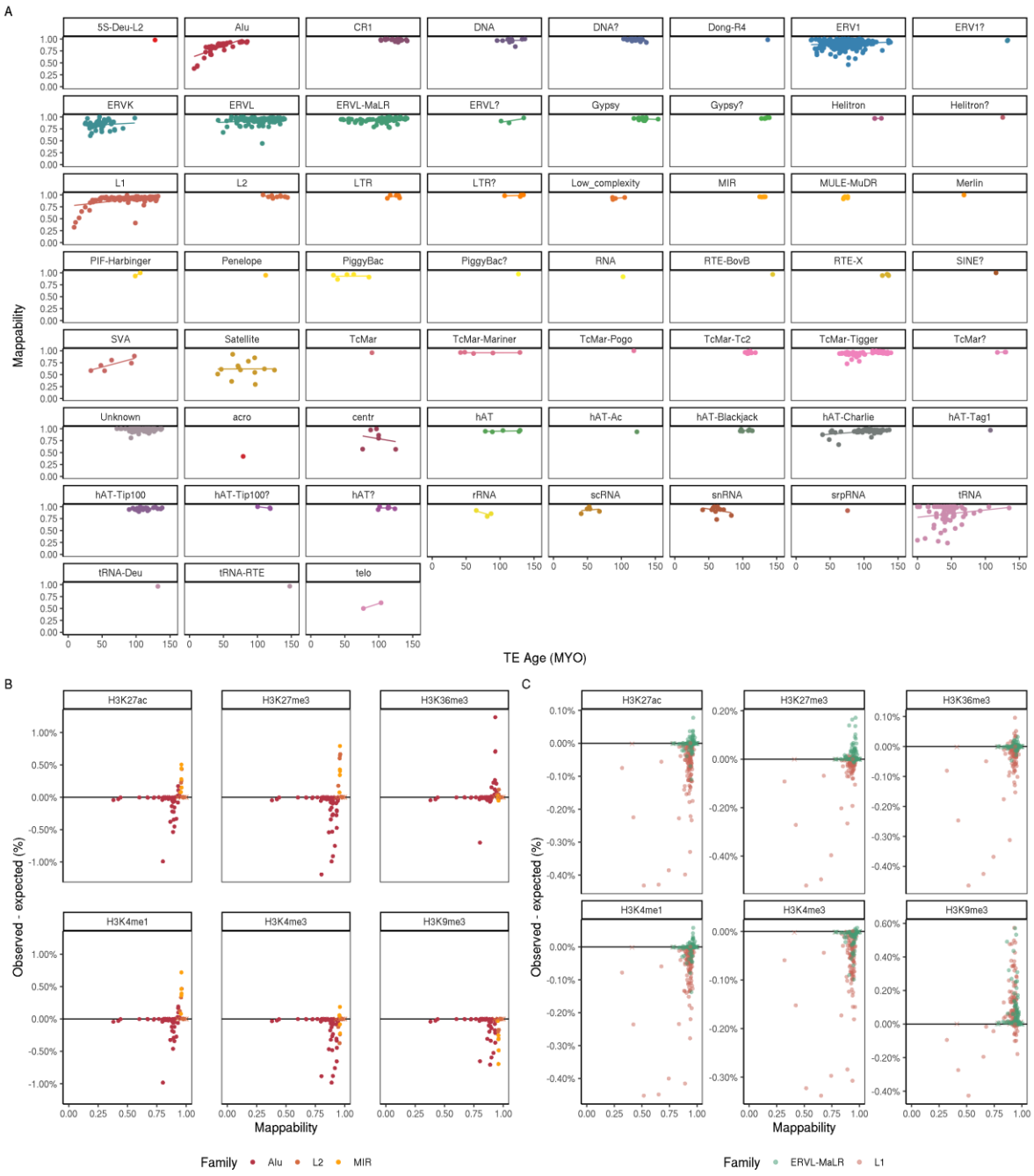

Figure S2. 9 TE Mappability and Age

A) TE Mappability in function of TE age. **B)** Enrichment in function of mappability for Alu, L2 and MIR **C)** Same as B for ERVL-MaLR and L1 TE families

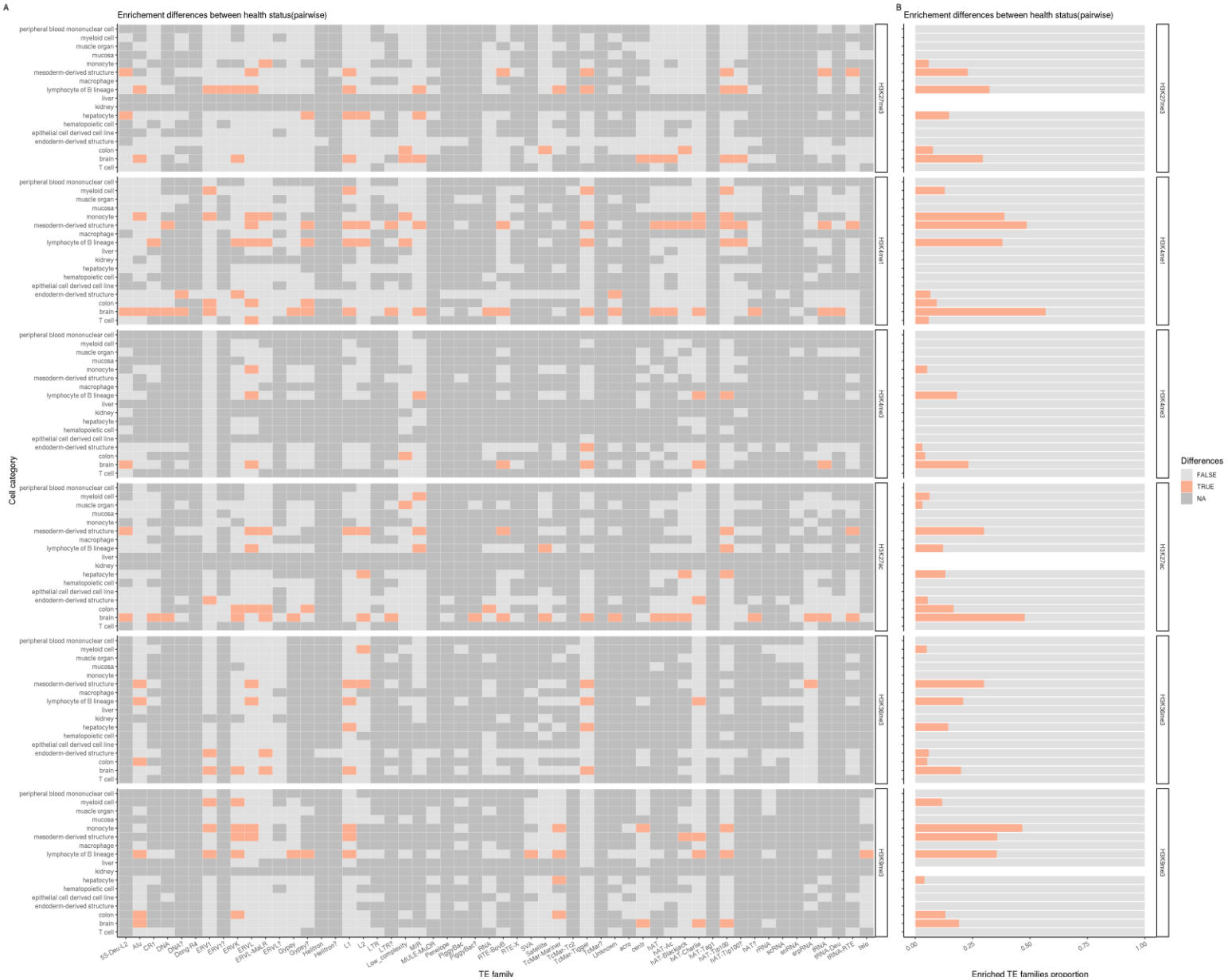
A
Enrichement differences between health status(pairwise)
B
Enrichement differences between health status(pairwise)
Cell category
peripheral blood mononuclear cell
myeloid cell
muscle organ
mucosa
monocyte
mesoderm-derived structure
macrophage
lymphocyte of B lineage
liver
kidney
hepatocyte
hematopoietic cell
epithelial cell derived cell line
endoderm-derived structure
colon
brain
T cell
H3K27me3
H3K4me1
H3K4me3
H3K27ac
H3K36me3
H3K9me3
SS-Dev-L2
Alu
CR1
DNA
DNA?
Dong-R4
ERV1
ERV1?
ERVK
ERVL
ERVL-MaLR
ERVL?
Gypsy
Gypsy?
Helitron
Helitron?
L1
L2
LTR
LTR?
Low_complexity
Mir
MULE-MuDR
Penelope
PiggyBac
PiggyBac?
RNA
RTE-BovB
RTE-X
SVA
Satellite
TcMar-Mariner
TcMar-Tc2
TcMar-Tigger
tRNA
Unknown
acro
cent
hAT
hAT-Ac
hAT-Blackjack
hAT-Charlie
hAT-Tag1
hAT-Tip100
hAT-Tip100?
hAT?
rRNA
scRNA
snRNA
srpRNA
tRNA
tRNA-Deu
tRNA-RTE
telo
TE family
0.00
0.25
0.50
0.75
1.00
Enriched TE families proportion
Differences
FALSE
TRUE
NA

Figure S2. 10 Cell type enrichments with significant differences between health statuses

A) Comparison between the cell type's health status for the listed TE families. Orange means there was at least 1 significant pairwise difference between the health status, light gray no significant difference, dark gray: Not enough data for the comparison (Only one health status available) **B)** Tally of the proportion TE families with health status differences (orange) or no difference (light gray) for each cell types (Aligned with A).

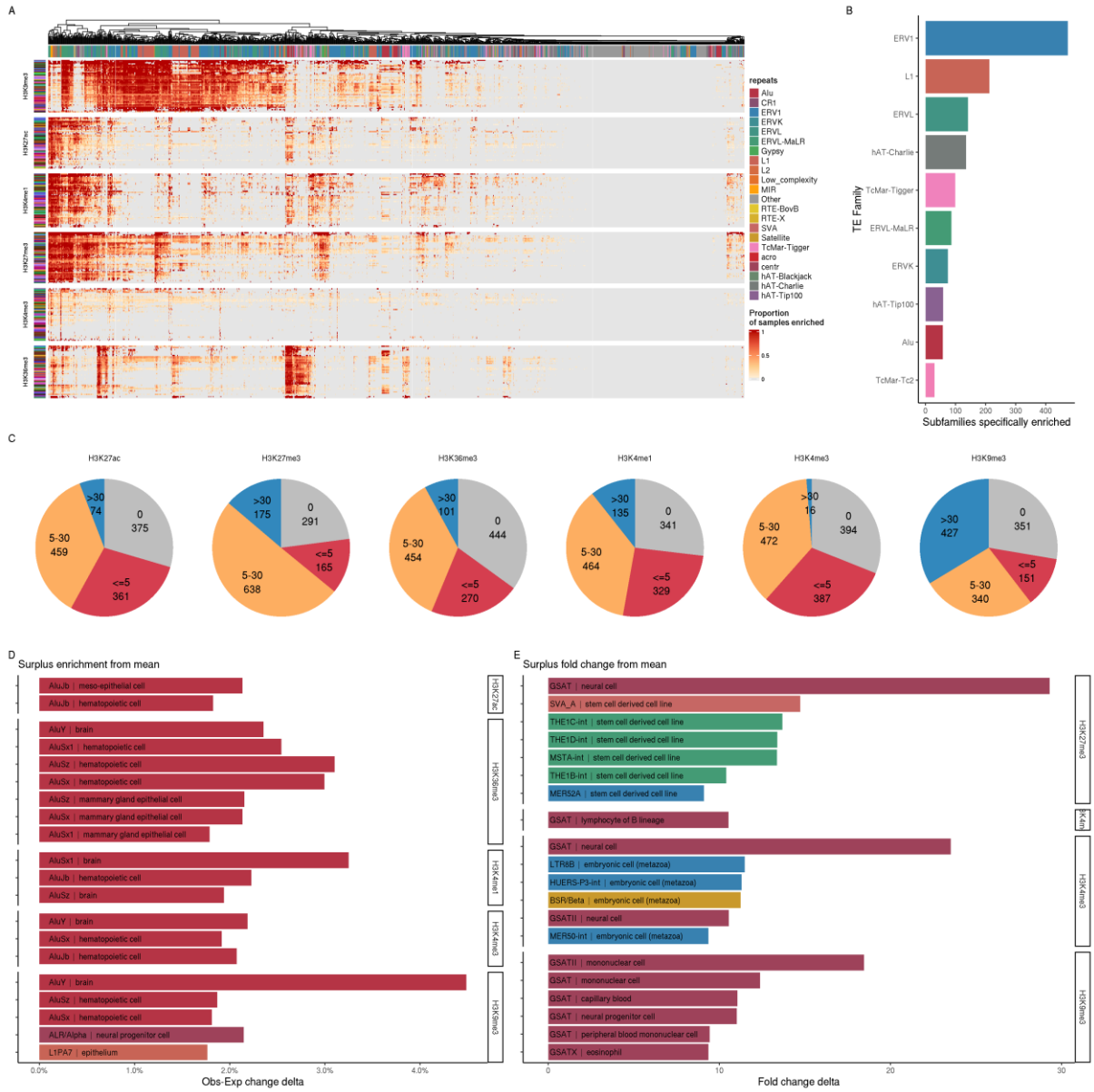

Figure S2. 11 Measurements of cell type specificity and surplus

A) Proportion of samples within each cell type with TE subfamily Enriched (pval<0.001). x axis shows the all TE subfamilies, y axis shows the 47 cell types grouped by histone mark. The TE subfamilies are hierarchically clustered.**B)** Top 10 families with most TE subfamilies (with at least 100 instances) enriched in less than 6 cell types (red part of fig 5B pie chart) **C)** Number of TE subfamilies that were enriched in bins of numbers of cell types. Separated by histone mark. **D)** Enrichment surplus of TE subfamily for given cell type. Surplus represent difference between the cell type enrichment and the mean across all cell types. **E)** Same as E for Fold change.

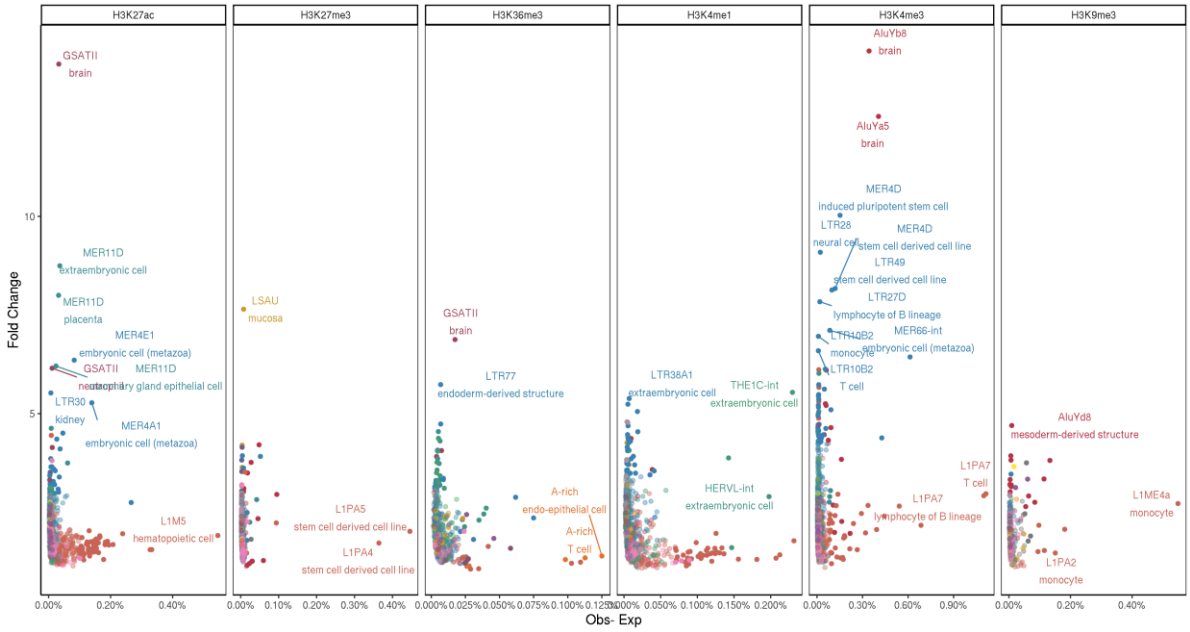

Figure S2. 12 Expanded set of Cell type specific TE-histone pairs
Fold change and Observed – Expected TE enrichment of TE subfamilies per cell type. Labeled are random subset of the candidates: most enriched (95th percentile) points in terms of Obs – expected or Fold change.

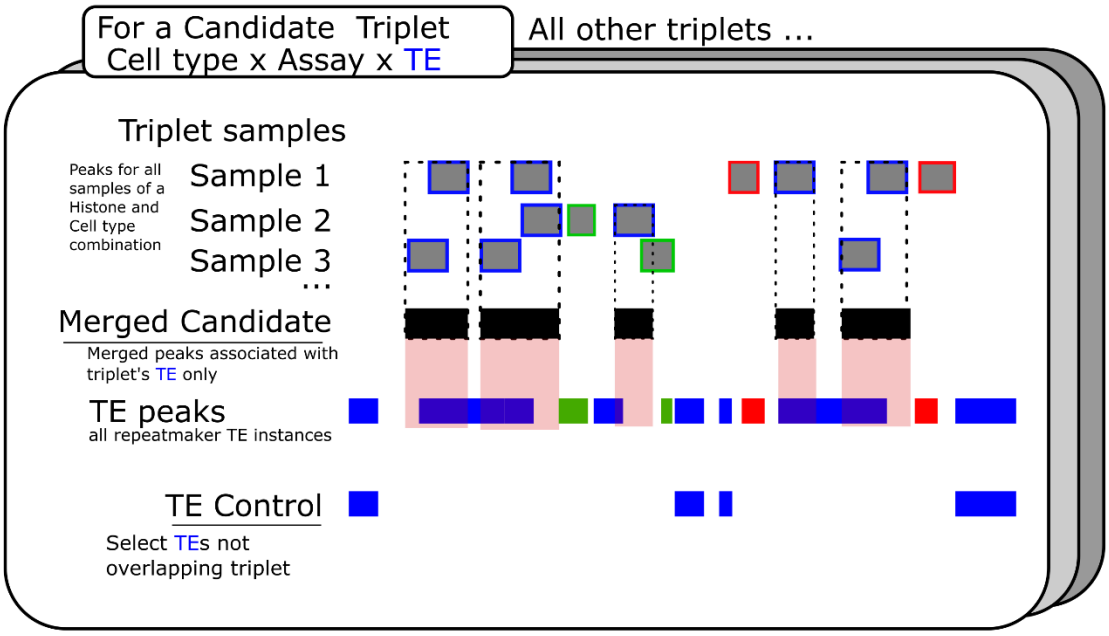

Figure S2. 13 Diagram of merged sample and control's generation
Diagram of merged sample and control's generation (Blue, green and red represent TE families, blue is the TE family of interest for this example) 1. Merge samples from merging all peaks for each Assay-cell type combinations and keeping only the peaks from select TE to compose merged candidates. 2. A TE control keeping only the select TE's instances that were not in the aforementioned triplet. A merged candidate file and associated TE control was generated for all candidates.

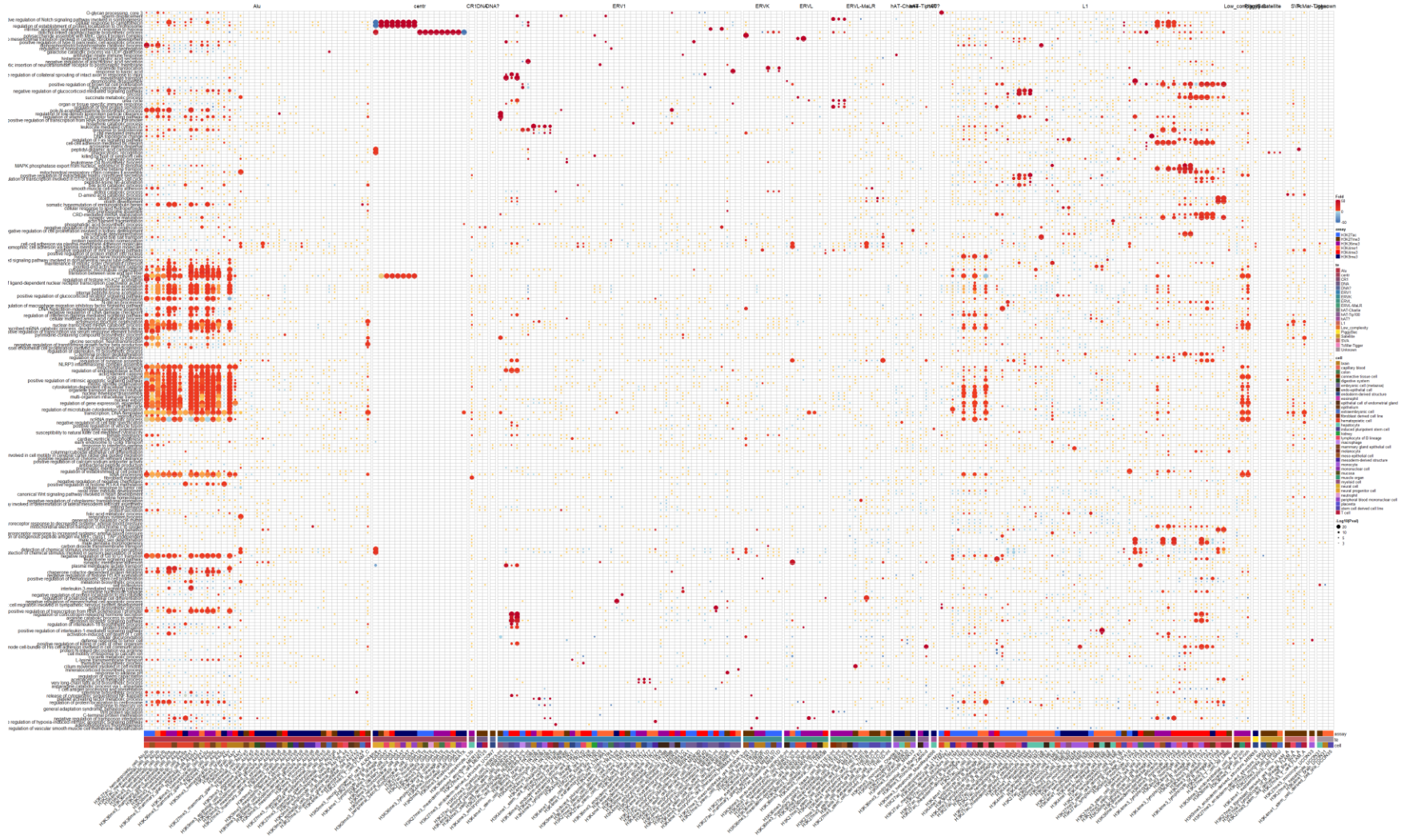
Alu
centr
CR1
DNA?
ERV1
ERVK
ERVL
ERVL-MaLR
hAT-Charlie
hAT-Tip100
L1
Low_complexity
Satellite
TcMar-Tigger
Field
assay
le
cell
Log10(Pval)

Figure S2. 14 GO biological process enrichments of TE candidate subset

Fold change enrichment difference from TE control of GO biological processes within for subset of 209 candidate triplet (TE-Assay-Cell Type) samples across histones. (fold change data – fold change of associated TE control) red enriched, yellow between -1 and 1, blue depleted. Circle sizes represents significance of the data. Terms selected based on data, favored most enriched terms per candidates. In rectangle are some mentioned processes.

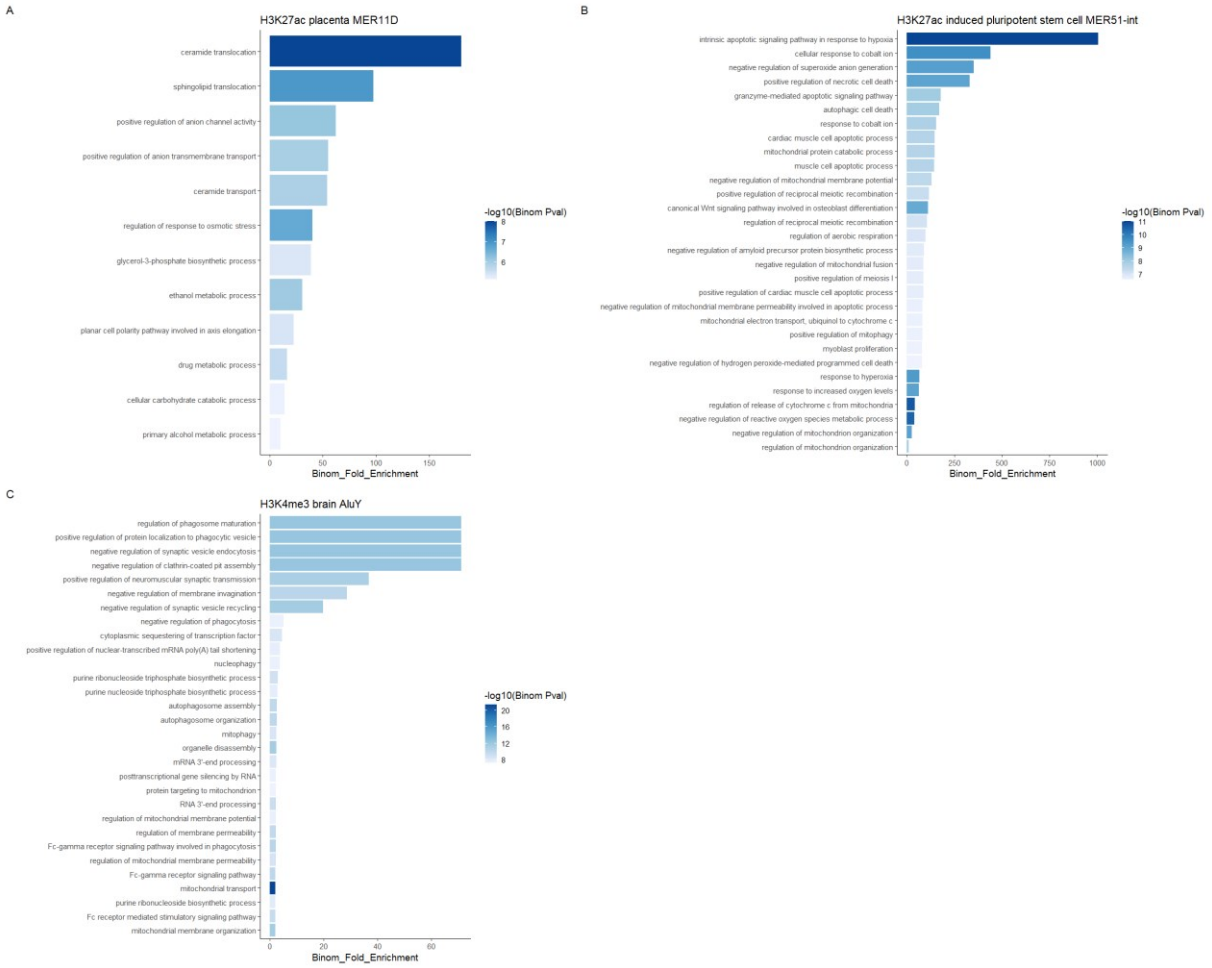

Figure S2. 15 GO biological processes enrichments within select candidate triplets
A) Fold change enrichment of GO biological processes within select candidate triplets within H3K27ac placenta MER11D triplet. **B)** Within H3K27ac IPS cell MER51-int triplet. **C)** Within H3K4me3 brain AluY. For all plots, showing up to the top 30 processes with fold enrichment >= 2 and –log10(pval) > 5.

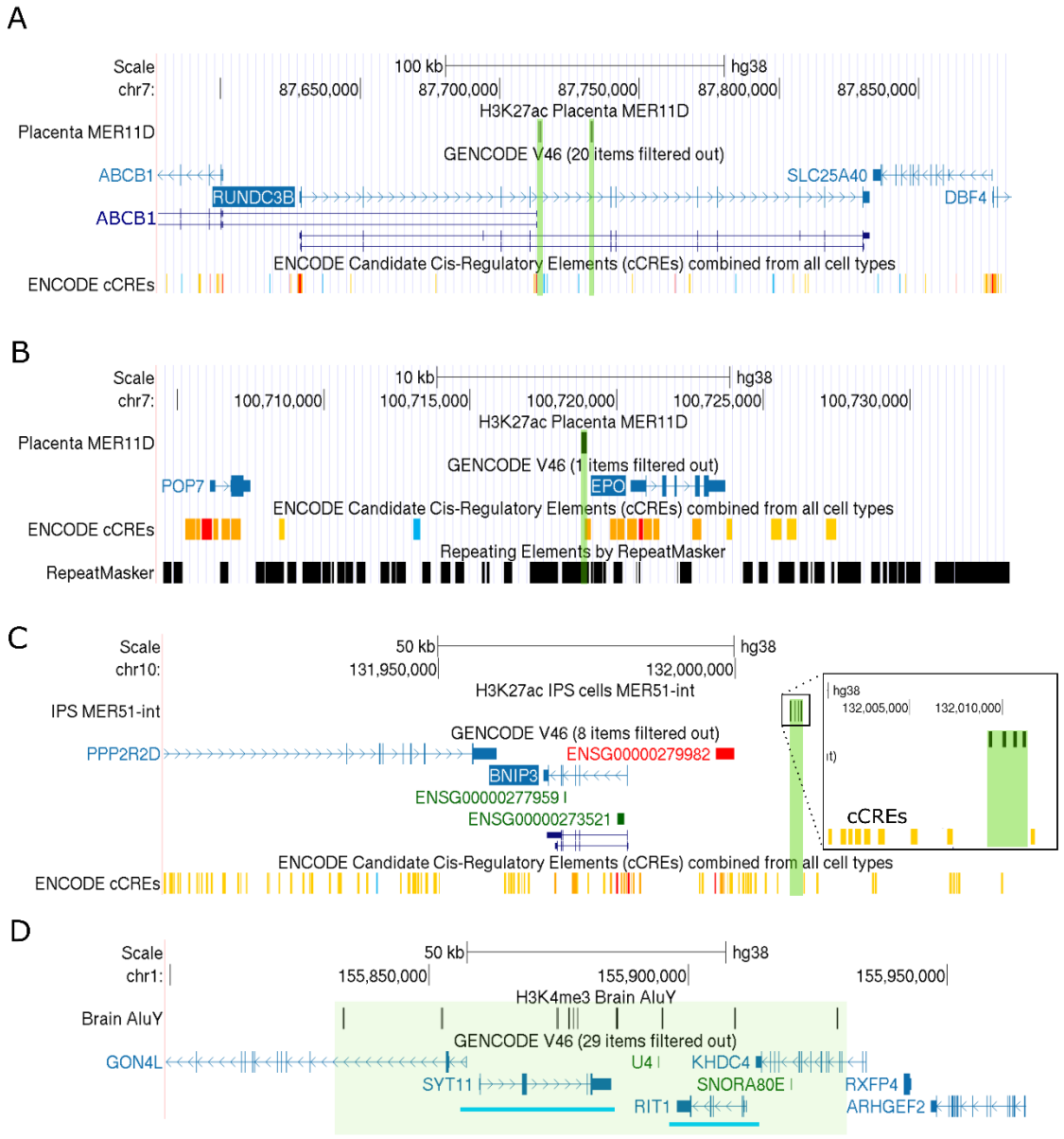

Figure S2. 16 Genome tracks of candidate associated peaks

A) H3K27ac Placenta peaks overlapping MER11D near ABCB1 gene. **B)** H3K27ac Placenta peak overlapping MER11D near *EPO* gene. **C)** H3K27ac IPS cells peaks overlapping MER51-int near *BNIP3* gene. Close up shows the peaks not overlapping cCREs. **D)** Cluster of H3K4me3 brain peaks overlapping AluY around *SYT11* and *RIT1* genes.

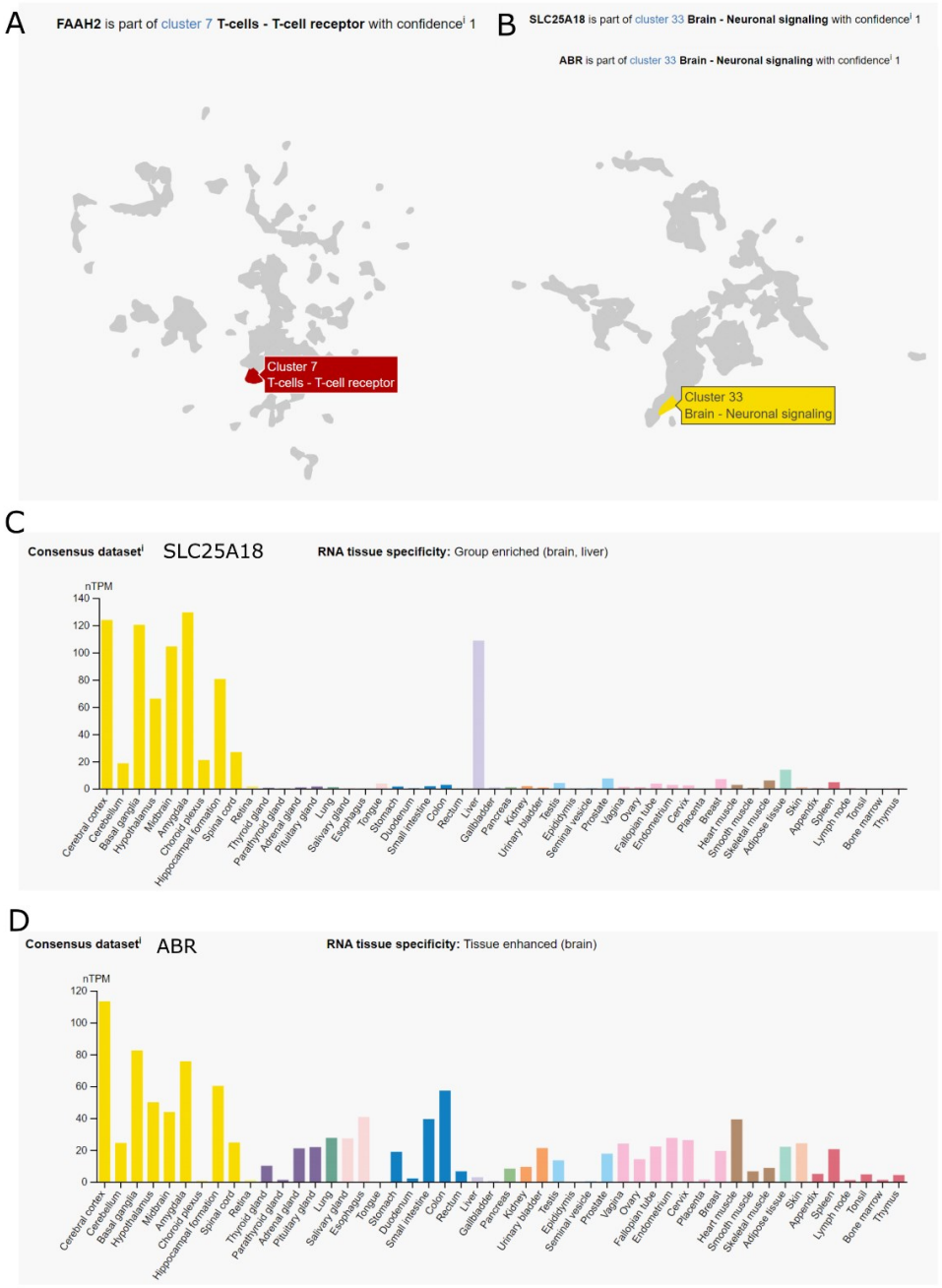

Figure S2. 17 RNA tissue expression of select genes
A) Single cell expression clustering of *FAAH2* gene from protein Atlas(https://www.proteinatlas.org/ENSG00000165591-FAAH2/single+cell+type) **B)** Expression clustering of *SCL25A18* and *ABR* genes, they both grouped in the same cluster. **C)** RNA normalized expression across tissues for *SLC15A18* gene(https://www.proteinatlas.org/ENSG00000182902-SLC25A18/tissue) **D)** Same as C for *ABR* gene(https://www.proteinatlas.org/ENSG00000159842-ABR/tissue). Images and data available from v23.proteinatlas.org

Supplemental Tables

Files available: https://bitbucket.org/hyacinthe_j/thesis-data/src/main/

Supplementary Table 1. Cell type TE enrichement. Summary statistics (Mean, Median, Max, Min, n) and rank (cell type per assay and TE family) of TE enrichments. Only TE subfamilies that were significantly enriched (Obs-Exp, thus positive) were used and added up for one TE family measurement per sample. The summary statistics are obtained from the summary of all samples grouped by assay and cell type.
STable1_summary_te_cell_data_enriched.csv

Supplementary Table 2
Cell type TE enrichment including non-significant (and depleted) elements Summary statistics (Mean, Median, Max, Min, n) and rank (cell type per assay and TE family) of TE enrichments. All TE subfamilies that were used were used and added up for one TE family measurement per sample. Negative values are possible for cases where Obs-Exp was depleted. The summary statistics are obtained from the summary of all samples grouped by assay and cell type. (same as table 1, but including non significantly enriched TEs)
STable2_summary_te_cell_data_all.csv

Supplementary Table 3 Select candidates table full set Candidate TE name, family and associated assay and cell type with the TE enrichment of that grouping as obs-exp (count_obs_exp_percent) and fold change (count_fold_change). For candidates identified by surplus, the difference from the mean (cell_mean_delta, cell_mean_fold_delta) are also listed. The sample count (n), number of times the observed count was higher than expected (time_over,1000 trial per sample) and resulting pvalue (pval) are listed. And the 4 candidate identifications are shown as true or false (top_obs, top_foldchange, obs_surplus, foldchange_surplus) with the valid method count (method_count) in the last column for each candidate.
STable3_TE_candidates_full_set.csv

Supplementary Table 4

Select candidates table 209 subset

Candidate TE name, family and associated assay and cell type with the TE enrichment of that grouping as obs-exp (count_obs_exp_percent) and fold change (count_fold_change). For candidates identified by surplus, the difference from the mean (cell_mean_delta, cell_mean_fold_delta) are also listed. The sample count (n), number of times the observed count was higher than expected (time_over,1000 trial per sample) and resulting pvalue (pval) are listed. And the 4 candidate identifications are shown as true or false (top_obs, top_foldchange, obs_surplus, foldchange_surplus) with the valid method count (method_count) in the last column for each candidate.

Subset visualised that was selected as a subset of the top 15 most enriched candidates per histone for top specific obs-exp and fold change using their respective metric (obs-exp and fold change, respectively). And keeping the 40 surplus candidates.

STable4_TE_candidates_subset.csv

Appendix C

Supplementary material for Chapter 0

Supplementary figures

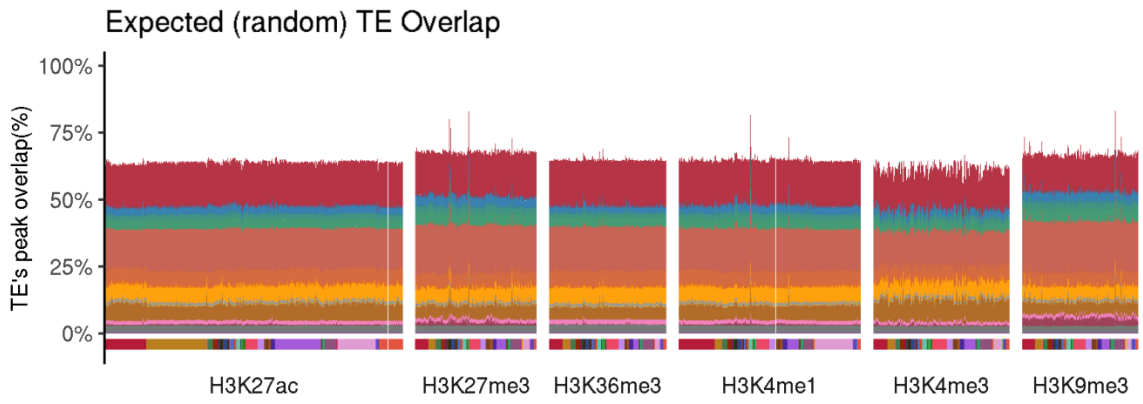

Figure S4.1 TE overlap from distribution matching controls

Overlap of peaks from simulated samples with TE family in 4614 samples. Samples are annotated by cell type

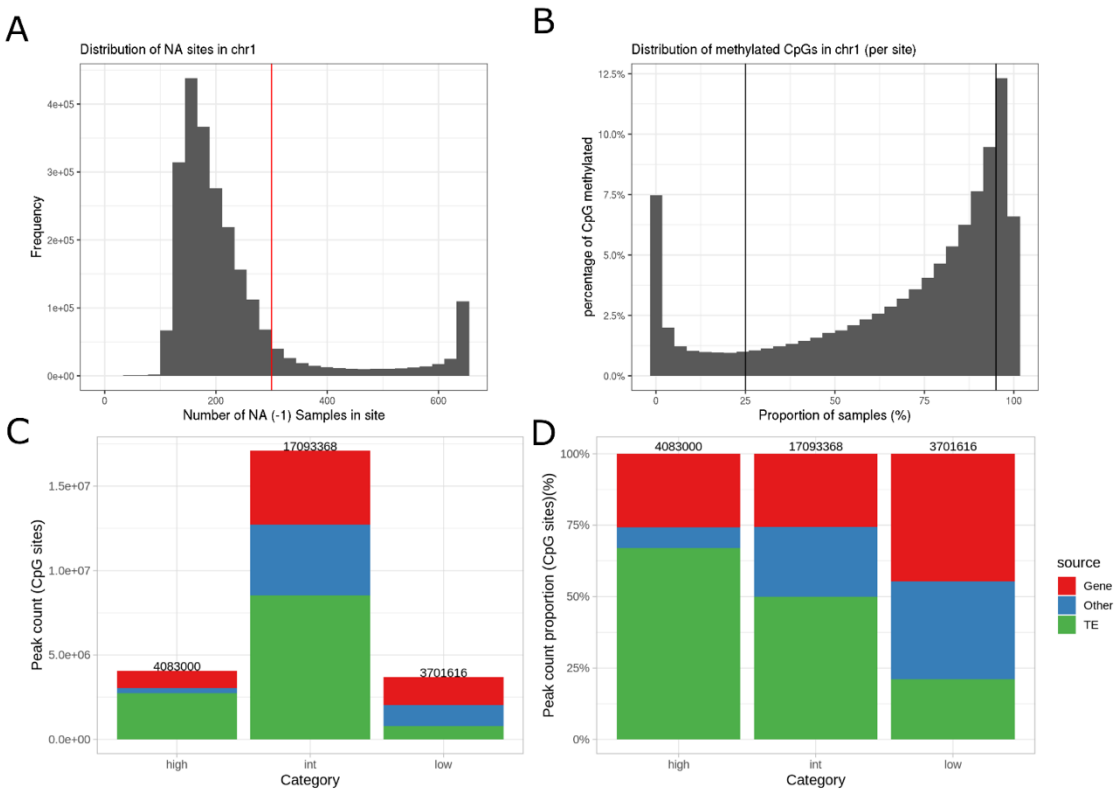

Figure S4.2 CpG sites determination thresholds and prevalence categories

A) distribution of the number of NA sites in chromosome 1, red line is the 300 selected threshold. Sites with less NA are kept. B) Distribution of the proportion of proportion of samples (prevalence) CpG sites were methylated in (>=80% coverage), category thresholds of under 25% for low and above 95% for high(vertical lines), intermediate in between. C) Number of CpG sites in the 3 prevalence categories, count listed on top of bars, colored by what the site overlapped between TE or gene, if either (source). D) same as D but displayed as percentages.

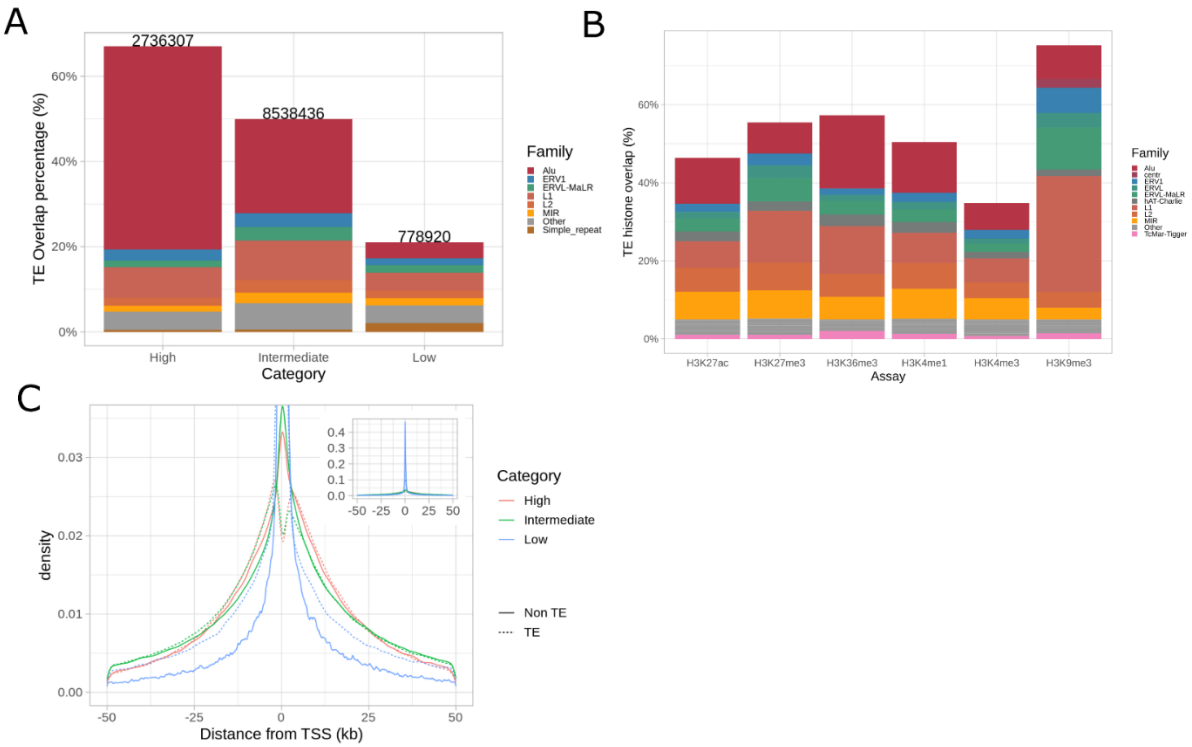

Figure S4.3 CpG site TE overlap and distance to TSS according to prevalence
A) CpG sites overlap with TEs depending on site's prevalence distribution Colors are TE families, Numbers at top are the numbers of sites. Note that High, the highest percentage, has a lower number of sites than intermediate. **B)** Mean TE overlap of histones (from chapter 2 and 3). TE family are the colors. **C)** Distance to nearest gene TSS for the CpG sites depending on prevalence category (color) and whether the site overlapped TEs (line type)