EPFL

Thèse n° 9450

KRAB zinc-finger proteins and their transposable element targets: between antagonism and cooperation

Présentée le 13 juillet 2022

Faculté des sciences de la vie
Laboratoire de virologie et génétique
Programme doctoral en approches moléculaires du vivant

pour l'obtention du grade de Docteur ès Sciences

par

Jonas Caspar DE TRIBOLET-HARDY

Acceptée sur proposition du jury

Prof. D. M. Suter, président du jury
Prof. D. Trono, directeur de thèse
Prof. O. Barabas, rapporteuse
Prof. G. Andrey, rapporteur
Prof. J. Gräff, rapporteur

École polytechnique fédérale de Lausanne

2022

“It’s coexistence or no existence”

- Bertrand Russell

Acknowledgements

I would like to thank:

Prof. Didier Trono for giving me the opportunity to work on this project, for his advice and support.

The members of my jury Prof. Barabas, Prof. Andrey, and Prof. Gräff for their comments and the interesting discussion.

Prof. Sutter for presiding over my defense.

Pila for her support when things get difficult and for her scientific insights.

Sandra and Charlène for answering my innumerable questions with great patience and expertise.

Séverine for all the help and everything she does.

The Batcave for their help and for the IT infrastructure they provide.

My fellow PhD students and the Post-Docs for all the help and support, especially when writing this manuscript and preparing for my defense. It was a true privilege to work with such an exceptional group of people.

Bara for her work on the KRABopedia and her enthusiasm.

Past members of the Laboratory of Virology and Genetics for their work I was able to build on.

Anna Groner and Laura Occhipinti for their teaching and guidance, preparing me for this challenge.

My family and friends for their patience and support.

Mahsa for being Mahsa.

Table of contents

Table of figures

Abstract

There are 377 Krüppel-associated box (KRAB) domain-containing zinc finger proteins (KZFPs) in the human genome, making them the largest family of transcription factors. KZFPs are defined by a N-terminal KRAB domain and several zinc-finger domains arranged in an array at the C-terminus of the proteins. The zinc-finger domains each form sequence specific interactions with double stranded DNA, allowing the zinc-finger array to target specific genomic sequences. The KRAB domain, through its interaction with the protein TRIM28 (also known as KAP1) allows for the stable silencing of transcription in a genomic region. Together these two domains allow KZFPs to bind specific regions of the genome and generally lead to the formation of heterochromatin and silencing of transcription allthough other functions for some KZFPs have been recently reported. KZFPs usually target transposable elements (TEs), mobile genomic elements that can move through the genome either by cut-and-paste or copy-paste mechanisms and make up almost half of the human genome. Different KZFPs bind to specific regions of TEs, limiting their expression and thus mitigating some of the threat they pose to human development, allowing both the TE and its host to survive. In recent years it became apparent that both certain TEs and KZFPs have roles beyond the previously described threat and mitigation of that threat. TEs harbor gene regulatory regions allowing, through their spread, for a dissemination of those regions through the genome and for new gene regulatory networks to form. KZFPs can affect the transcription of genes located in proximity to their binding sites leading to changes in gene regulation as well. A multitude of regulatory roles involving KZFPs and TEs have been described in recent years making them an exciting field of study. A major hurdle in understanding both KZFPs and TEs is to identify the binding sites of KZFPs, as they reveal both the targets of the KZFP and how, if at all, a TE is regulated by KZFPs. To do so, we expanded on previous efforts and aimed to perform chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-seq) on every member of the human KZFP family. Here we present ChIP-seq experiments for 110 new KZFPs, which together with previously published data, allow us to study the binding of almost all human KZFPs (95%). The entirety of the data was analysed together to generate a coherent dataset and results for each KZFP are made available to the public on our web portal (https://tronoapps.epfl.ch/web/krabopedia/). The identified binding sites allowed us to witness the adaptation of the KZFP family to new TEs and showed how targeted sequences shift after segmental gene duplication events. Furthermore, we could corroborate that several KZFPs target the same TE subfamilies in a seemingly redundant fashion and show that unexpectedly these KZFPs arose independently in different genomic locations. These results represent a valuable tool for anyone studying KZFPs and should aid in the pursuit of both understanding KZFPs and TEs.

Keywords

KRAB zinc finger proteins, transposable elements, chromatin immunoprecipitation, evolution

Kurzfassung

Krüppel-assoziierte box (KRAB) Domäne-enthaltende Zink Finger Proteine (KZFPs) sind die grösste Familie von Transkriptionsfaktoren im menschlichen Genom, mit 377 Mitgliedern. Sie sind definiert durch eine KRAB Domäne an ihrem N-Terminus und mehreren Zink-Finger Domänen welche in einer Reihen and ihrem C-Terminus zu finden sind. Jede der Zink-Finger Domänen ermöglicht spezifische Interaktionen mit der DNA Doppelhelix und dadurch können die aneinandergereihten Zink-Finger Domänen spezifische Sequenzen des Genoms binden. Die KRAB Domäne interagiert mit dem TRIM28 Protein (auch bekannt als KAP1) und ermöglicht die stabile Unterbindung von Gentranskription. Diese zwei Domänentypen zusammen ermöglichen es KZFPs spezifische Regionen des Genoms zu binden und dort im allgemeine die Gentranskription durch die Etablierung von Heterochromatin einzuschränken, auch wenn kürzlich weitere Funktionen von einigen KZFPs entdeckt wurden. KZFPs binden meistens Transposons (TE), mobile Elemente welche sich entweder mit Ausschneiden-Einfügen oder Kopieren-Einfügen Mechanismen durch das Genom bewegen können und fast die Hälfte des menschlichen Genoms ausmachen. Unterschiedliche KZFPs binden spezifische Regionen von TEs und hemmen dadurch ihre Verbreitung, was die Gefahr von TEs für die menschliche Entwicklung mildert und das Überleben von Wirtorganisums sowie vom TE ermöglicht. Entdeckungen der vergangenen Jahre haben alternative Rollen von KZFPs und TEs zu Tage gebracht welche über diese Gefahreneindämmung hinausgehen. TEs enthalten Genregulationssequenzen welche sie durch ihre Ausbreitung im Genome verteilen, dies kann zur Entstehung neuer Regulationsnetzwerke führen. Ausserdem können KZFPs selbst die Regulation von Genen in der Nähe ihrer gebundenen Sequenzen beeinflussen und dadurch genregulatorische Rollen übernehmen. Mehrere solcher Rollen für KZFPs und TEs wurden in jüngere Vergangenheit beschrieben was KZFPs zu einem interessanten Forschungszweig macht. Eine der grossen Hürden für das Studium von KZFPs und auch TEs ist die Identifizierung der Sequenzen welche KZFPs binden, da sie zugleich die Zielsequenzen von KZFP enthüllen, sowie ob TEs von KZFPs reguliert werden. Um dies zu erreichen knüpften wir an vorhergehenden Anstrengungen an und haben uns vorgenommen Chromatin-Immunpräzipitationen gefolgt von massiver paralleler DNA-Sequenzierungen (ChIP-seq) für jedes Mitglied der menschlichen KZFP Familie zu machen. Wir präsentiere hier 110 neue ChIP-seq Experimente welche es uns erlauben, zusammen mit bereit publizierten Daten, beinahe jedes Mitglied (95%) der menschlichen KZFP Familie zu untersuchen. Die gesammelten Daten wurden zusammen mit den vorherigen analysiert um einen einheitlichen Datensatz zu generieren. Die Resultate dieser Analysen sind öffentlich zugänglich auf unserem Webportal (https://tronoapps.epfl.ch/web/krabopedia/). Die gebundenen Sequenzen welche wir identifizieren konnten ermöglichten uns zu zeigen wie sich die KZFP Familie an

neue TEs anpasst und wie die gebundenen Sequenzen sich nach KZFP-Genduplikationen verschieben. Wir können zudem bestätigen das viele TE-Unterfamilien mehrfach von verschiedenen KZFPs gebunden werden. Überraschenderweise können wir auch zeigen, dass diese scheinbar redundanten KZFPs unabhängig voneinander entstanden sind. Unsere Resultate sind ein wertvolles, frei zugängliches Werkzeug für das Studium von KZFPs und sollten eine grosse Hilfe für das bessere Verstehen von KZFPs und TEs darstellen.

Schlüsselwörter

KRAB Zink-Finger Proteine, Transposons, Chromatin-Immunpräzipitation, Evolution

Introduction

This thesis investigates the targets of Krüppel-associated box (KRAB) domain-containing zinc finger proteins (KZFPs) and the effects KZFPs have on these targets. Two, entangled, effects are focused on: transcriptional regulation and epigenetic modifications, both introduced here briefly before moving to KZFPs and their targets.

Transcriptional regulation

In 1958 Dr. Francis Crick first postulated that information in cells flows from DNA to RNA to Protein (Crick, 1958) and even though many exceptions to this so called “Central Dogma of Molecular Biology” have been found since (Crick, 1970; Morange, 2009; see retrotransposons later in this introduction), the transcription of genes into messenger RNA (mRNA) and subsequent translation into protein is still a major process in defining the state of a cell. In Eukaryotic cells, DNA is transcribed into RNA by three RNA polymerases (Pol) (Cramer et al., 2008). Pol I and Pol III synthesize the ribosomal RNAs and transfer RNAs, necessary for mRNA translation into proteins, and Pol II synthesizes mRNAs and a variety non-coding RNAs (Cramer, 2019). The process of transcription by different RNA polymerases and their regulation is reviewed in Cramer, 2019 and briefly summarize here. For transcription to take place Pol needs to be recruited to the promoter region of a gene by transcription initiation factors, where together with other proteins it will form a pre-initiation complex (PIC). The PIC will then open the DNA double strand at the promoter region and proceed to transcribe one strand of DNA, the template strand, into RNA. After the newly synthesize RNA reaches a certain length, the Pol will escape the promoter region and form an elongation complex together with another set of proteins called elongation factors. This complex will continue transcribing the template strand into RNA until Pol dissociates from DNA, terminating the transcription. In order for the polymerases to bind the promoter region its chromatin needs to be in a permissive state, namely depleted of nucleosomes and flanked by specialized +1 and -1 nucleosomes. Thus, chromatin state and modifications of it have a major impact on regulation of transcription.

Epigenetic chromatin modifications

Epigenetic modifications, from the Greek word epi for around or over, are inheritable modifications that can affect chromatin state and thus gene expression without altering the sequence of the DNA. The two most studied forms of epigenetic regulation are DNA methylation and histone modifications.

DNA methylation

DNA methylation involves the covalent addition of a methyl group to the fifth carbon of a cytosine (5mC) and occurs mostly on cytosine guanine pairs (CpGs). The p in CpGs stands for the phosphate linking bases together indicating that the C and G are on the same strand. The first quantification of 5mC in 1982 (Ehrlich et al., 1982) revealed that only 1% of the bases in the human genome are 5mC, even though this number can vary between tissues it remains a rare event. This is due to the fact that 5mC bases tend to be converted into thymidine through hydrolytic deamination (Singal and Ginder, 1999) thus making them mutagenic. DNA methylation is deposited by 3 proteins called DNA-methyltransferases DNMT1, DNMT3A and DNMT3B. Functionally DNA methylation negatively regulates gene expression by preventing the binding of transcription factors or by recruiting repressors (Moore et al., 2013). It was thought to be installed early in development and stable all along our lifespan (Hackett and Surani, 2013) however recent findings show that DNA methylation remains a dynamic process throughout our lives (Ciccarone et al., 2018; Farlik et al., 2016). DNMT3A and DNMT3B are de novo methyltransferases (Brunetti et al., 2017; Gagliardi et al., 2018) whereas DNMT1 serves as a maintenance enzyme targeting hetero-methylated DNA strands. Recently it has been shown in mouse that Dnmt1 can de novo methylate retrotransposons and that it colocalizes with Trim28 (Haggerty et al., 2021) indicating that KZFPs might recruit DNMT1 to methylate DNA. DNA methylation on transposable elements, which will be introduced later, is mainly deposited by PIWI interacting RNAs (piRNA) in early embryogenesis (reviewed in Lin et al., 2021; Ozata et al., 2019).

Histone modifications

In the nucleus DNA is organized in so called nucleosomes consisting of DNA wrapped around an octamer of histone proteins. These proteins can be modified in multiple locations and numerous ways such as methylation, acetylation, phosphorylation and ubiquitination, each with different influences on DNA structure and gene expression (Lennartsson and Ekwall, 2009). In this thesis we are mainly interested in modifications that govern the packaging of the nucleosomes, with the condensation of nucleosomes greatly affecting the accessibility of the DNA (Struhl, 1999). This accessibility has a large effect on gene transcription (Cramer, 2019) and is influenced by post-translational modifications of the histones. The two major modifications governing DNA condensation are acetylation and methylation of lysine on histone tails (Talbert and Henikoff, 2021). Acetylation is generally associated with open chromatin as it reduces the positive charge of the lysine in the histones and loosens the interactions between them and the negatively charged DNA (Talbert and Henikoff, 2021). Methylation on the other hand can be found on both open or closed chromatin and can have many effects not related directly related to DNA condensation as they serve more as scaffolds

for the binding of other effector proteins than directly affecting nucleosome-DNA interactions (Zhang et al., 2021). The most relevant modifications for this thesis are the trimethylation of lysine 9 and acetylation of lysine 27 on Histone 3 (H3K9me3 and H3K27ac), both of which lead to the formation of stably silenced regions that are maintained over cell divisions. (Talbert and Henikoff, 2021). There is also a synergistic effect between DNA methylation and repressive histone marks favouring the establishment of stably silenced region. The protein MeCP2 which recruits histone deacetylases and other complexes repressing transcription, specifically recognizes 5mC (Nan et al., 1998) leading to a reinforcement of the silencing of these regions.

Krüppel-associated box domain-containing zinc-finger proteins

Krüppel-associated box (KRAB) domain-containing zinc-finger proteins (KZFPs) are the largest family of transcription factors in humans (Lambert et al., 2019). They can be traced back to the early tetrapods (Imbeault et al., 2017) with most subsequent species hosting a large number of KZFPs (Figure 1). This paragraph represents an overview of the structure, function and evolutionary history of KZFPs.

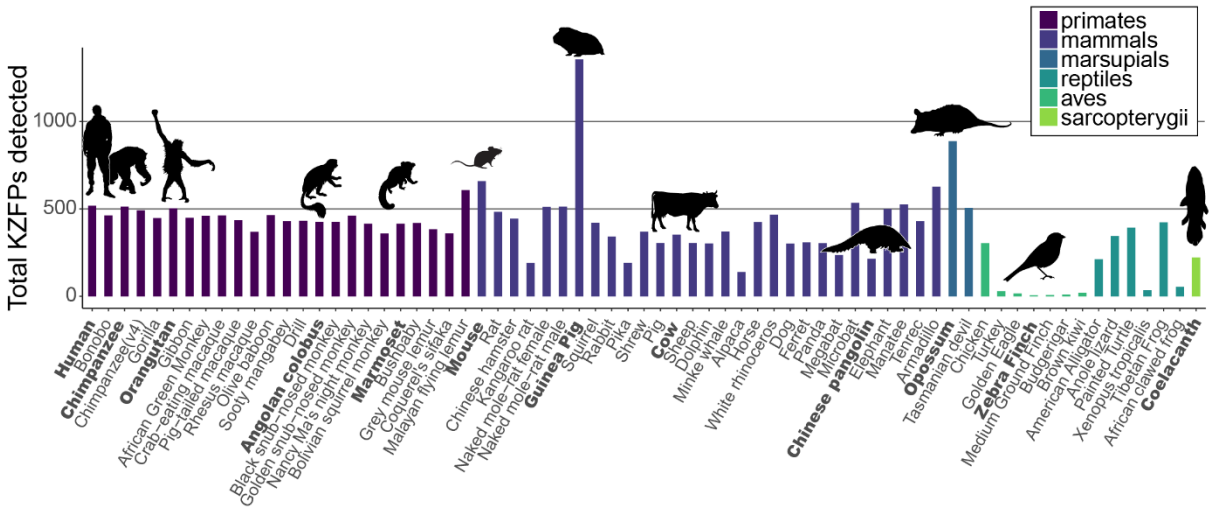


Figure 1: Number of detectable KRAB and zinc-finger domains across species.
Based unpublished work from Alexandre Coudray and Cyril Pulver using the methodology described in Imbeault et al., 2017. Silhouettes of random animals are shown (Keesey et al.) and labelled in bold.

Structure of KZFPs

KZFPs are defined by two domains: A N-terminal KRAB domain that spans approximately 75 amino acids and several C-terminal array of Cys_2-His_2 (C2H2) zinc-finger domains each 23 amino acids in length (Figure 2). KZFP genes are generally split into 3-4 exons, with the KRAB

domain split in 1 or 2 smaller exons and the C2H2 zinc-fingers located in a larger exon at the 3' end (Figure 2).

KRAB domain

KRAB domains are comprised of two modules, the KRAB-A module responsible for the repressive activity of a KZFP and the sometimes absent KRAB-B module which enhances KRAB-A activity (Ecco et al., 2017). The KRAB-A domain allows the recruitment of Tripartite Motif-Containing Protein 28 (TRIM28 also known as KAP1) which serves as a scaffold for the recruitment of several proteins involved in heterochromatin formation and DNA methylation such as the NuRD histone deacetylase complex, the histone methyltransferase SETDB1 or DNA methyltransferase DNMT1 (Ecco et al., 2017; Quenneville et al., 2011; Schultz et al., 2001, 2002).

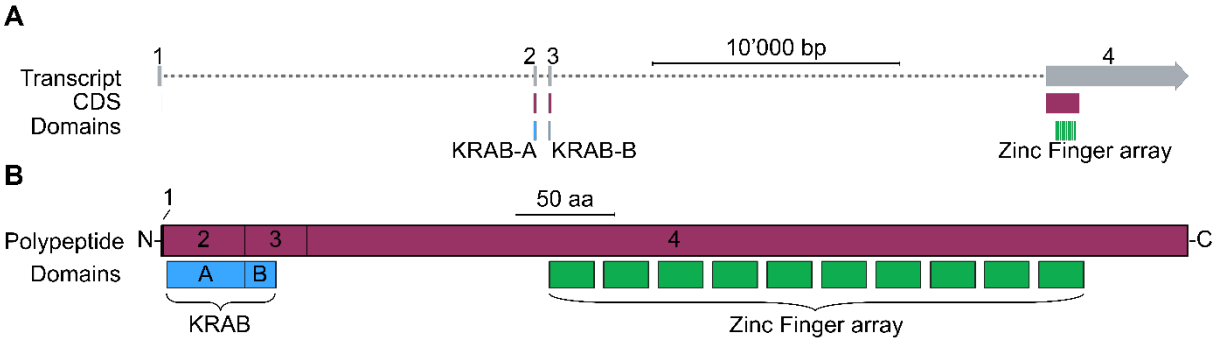


Figure 2: Structure of a human KZFP gene.
A) Transcript, Protein Coding Sequence (CDS) and domains for human ZNF626. Exons are shown in grey and numbered. Introns are depicted as dashed lines. The coding regions in each exon are shown in burgundy. The KRAB-A box and KRAB-B box are shown in blue, Zinc finger domains are shown in green. B) Amino acid sequence of ZNF626. The different segments and the exons they are encoded on are shown in burgundy and numbered. The domains are depicted below using the same colour code as in A). N and C indicate the N- and C-terminus of the polypeptide chain. Data from Ensembl and the HMMER web server (Cunningham et al., 2022; Potter et al., 2018).

Zinc-finger domain

Protein-coding KZFPs contain between 1 and more than 30 C2H2 zinc-finger domains with an average of 11.6 (data from ensembl: Cunningham et al., 2022). These domains interact with DNA in a sequence specific manner. They are of a simple modular nature, each comprised of a short two-stranded antiparallel β sheet and an alpha helix. The β sheet and α helix are coordinated by a Zinc ion interacting with two Cysteine (C2) and two Histidine (H2) amino acids, giving the domain its name (Figure 3A). The contacts with DNA are facilitated by 4 amino acids, 3 located in the α helix and one in the linker between the α helix and the β sheet (Figure 3B). The multiple zinc-fingers from a zinc-finger array wrap around the DNA, each contacting consecutive bases (Figure 3C). Every zinc-finger interacts with 3 bases and one additional base overlapping with the next zinc-finger (Elrod-Erickson et al., 1998). The collection of all

DNA-interacting amino acids from a zinc-finger array form the so-called zinc-finger print of a protein. The zinc-finger print can be used as a proxy for the binding behaviour of a zinc finger and comparisons of zinc-finger prints can be used to create evolutionary relationships (Imbeault et al., 2017). However all efforts to decipher a “recognition code” for zinc-fingers comparable with the genetic code have failed due to the interactions between different zinc-fingers int the same array (Wolfe et al., 2001). Having said that, two KZFPs with identical zinc-finger prints can nevertheless be expected to bind the same DNA sequences.

Additional domains

Other than the afore mentioned domains, KZFPs also contain a so-called linker region denoting a variable number of amino acids between the KRAB and zinc-finger domains. This region is highly unstructured, poorly conserved between different KZFPs and of variable length (Shen et al., 2021). However, a minimal length seems to be required for the correct functioning of the protein. Some KZFPs also carry additional domains, namely the Domain of Unknown Function 3669 (DUF3669) and the SCAN domain, the first only being present on KZFPs whereas latter also appears in zinc-finger containing proteins without a KRAB domain.

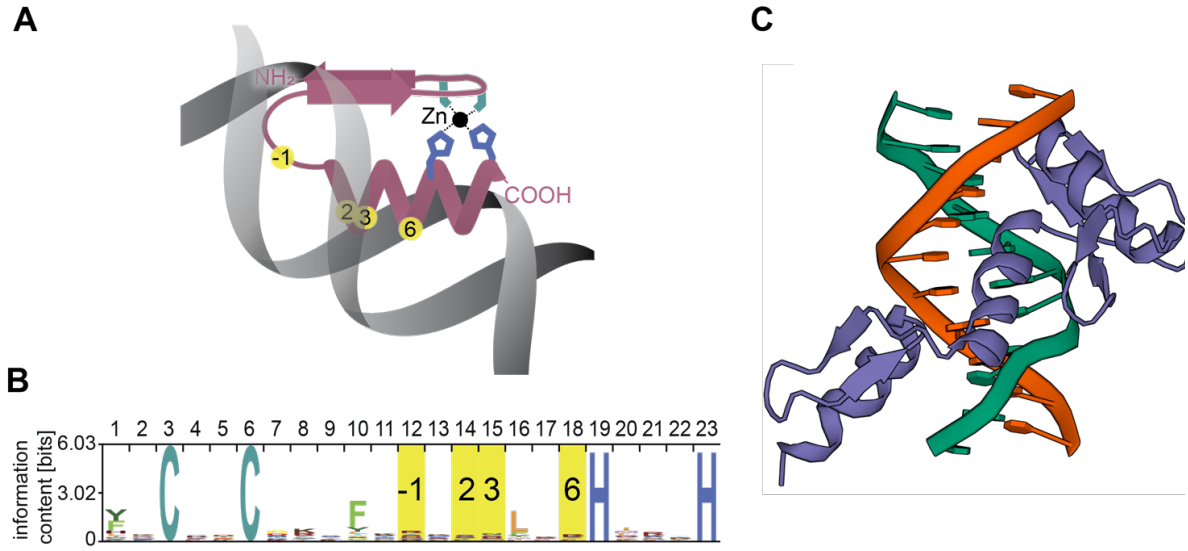


Figure 3: Zinc finger structure and Zinc-finger print.
A) Cartoon showing the structure of a single C2H2 zinc finger domain (burgundy) and its interactions with DNA (grey). The structurally crucial cysteines and histidines at positions 3,6,19 and 23 are highlighted in turquoise and blue respectively. The amino acids interacting with the DNA are labelled -1,2,3 and 6 according to their position relative to the α-helix and highlighted in yellow. They form the zinc-finger print of the protein. B) The consensus Hidden Markov Model (HMM) of a C2H2 zinc finger domain (PF00096 from El-Gebali et al., 2019). The height of the characters indicates the information content at a given position of the amino acid. The cysteines, histidines and DNA-interacting amino acids are coloured as in A. C) Interactions of 3 zinc-finger domains (purple) with DNA (orange and green) from Bank; Elrod-Erickson et al., 1998.

Functions

The commonly accepted function of KZFPs is the negative regulation of the expression of their target sequences, mostly transposable elements (TEs). The majority of KZFPs indeed bind TEs (Imbeault et al., 2017) and interact with TRIM28 (Helleboid et al., 2019) which is a key factor for the formation of heterochromatin and gene silencing. This repression of TEs has been shown for a number of mouse and humans KZFPs (Ecco et al., 2016; Fasching et al., 2015; Jacobs et al., 2014; Turelli et al., 2020; Wolf et al., 2015a). The function of this repression is not limited to solely prevent the spread of TEs but has also been reported for its roles in development and differentiation (Barde et al., 2013; Quenneville et al., 2011; Takahashi et al., 2019; Yang et al., 2017a; Zeng et al., 2012), metabolism (Lupo et al., 2013) and autophagy (Chauhan et al., 2013). Having said that, there is also a minority of KZFPs do not recruit TRIM28, some of which do not behave as repressors showing a heterogeneity of effects upon KZFP binding (Helleboid et al., 2019; Tycko et al., 2020). Additionally to their natural appearances, KRAB domains are widely used by researchers in a method called CRISPR inhibition (CRISPRi) (Gilbert et al., 2013) where a repressive KRAB domain is fused to a catalytically inactive Cas9 and targeted to a specific locus using guide RNAs. This allows the recruitment of TRIM28 silencing of those regions.

TRIM28 mediated functions

The main interactor of KZFPs, TRIM28 contains an N-terminal tripartite motif, giving it its name, and a C-terminal plant homeodomain (PHD) and bromodomain (Iyengar and Farnham, 2011). The tripartite motif is made up of three types of domains: a Really Interesting New Gene (RING), two B-box zinc fingers and coiled-coil domain (RBCC). The coiled-coil domain allows TRIM28 to form homodimers that can then interact with the KRAB domain of a KZFP or oligomerize with other TRIM28 dimers through the Bbox domain to form higher-order assemblies (Sun et al., 2019). Both the RING and PHD/Bromo domains have been shown to act as E3 small ubiquitin-like modifier (SUMO) ligases, which auto-SUMOylate several sites in the TRIM28 protein necessary for its repressive action (Sun et al., 2019). The main function of TRIM28 is to serve as a scaffold for Heterochromatin-Protein 1 (HP1), histone deacetylases (e.g. NuRD) and Histone 3 Lysine 9 (H3K9) methyltransferases (e.g., SETDB1). TRIM28 interacts with NuRD and SETDB1 through its PHD and bromodomain (Schultz et al., 2001, 2002) and with HP1 through a specific HP1 binding domain (HP1BD) located between the tripartite motif and the PHD and bromodomain (Lechner et al., 2000). These proteins in term facilitate the formation and spreading of heterochromatin and repression of transcription (Figure 4) (Nielsen et al., 1999; Ryan et al., 1999; Schultz et al., 2001, 2002). More recently additional functions of TRIM28 have been described (Randolph et al., 2022): There is the transcriptional activation for viral cell programs as a phosphorylation of the Serine 824 of

TRIM28 leads to the release of paused RNA-Polymerase II and viral transcription. Furthermore, a role in DNA damage response by being recruited to damaged sites and silencing transcription in affected regions has been described. Finally, in tumours TRIM28 has been reported to have an oncogenic capacity as it acts as E3 Ubiquitin and SUMO ligase on several tumour suppressor proteins such as p53 and AMPK (Randolph et al., 2022).

TRIM28 independent functions

A minority of KZFPs have been shown to not interact with TRIM28. First there is PRDM9 (Imai et al., 2017) which is designating hotspots of meiotic recombination by direct methylation of H3K4 and K36 through its PR/SET domain (Powers et al., 2016). It also contains an ancestral KRAB domain which is necessary for its function and potentially facilitates protein-protein interactions with components of the meiotic double-strand break machinery (Imai et al., 2017). The PRDM9 KRAB domain emerged 600 million years ago and is the likely ancestor of all human KRAB domains (Stubbs et al., 2011). Apart from PRDM9, several of the modern KZFPs have lost their ability to interact with TRIM28 (Helleboid et al., 2019), some of which have been shown to have activator instead of repressor functions (Tycko et al., 2020). Interestingly these activator KRAB domains resemble the original PRDM9 KRAB domain, suggesting that the ancestral KRAB domain might have been an activator domain (Tycko et al., 2020). Concerning the two other domains that can be observed on KZFPs: the SCAN and DUF3669 domains; Both the SCAN domain and the DUF3669 domains have been reported to allow for protein dimerization (Al Chiblak et al., 2019; Helleboid et al., 2019; Schumacher et al., 2000) and the DUF3669 domain has been shown to have repressive capabilities in early repression assays (Okumura et al., 1997; Williams et al., 1995). However, the exact functions of these domains is still under investigation.

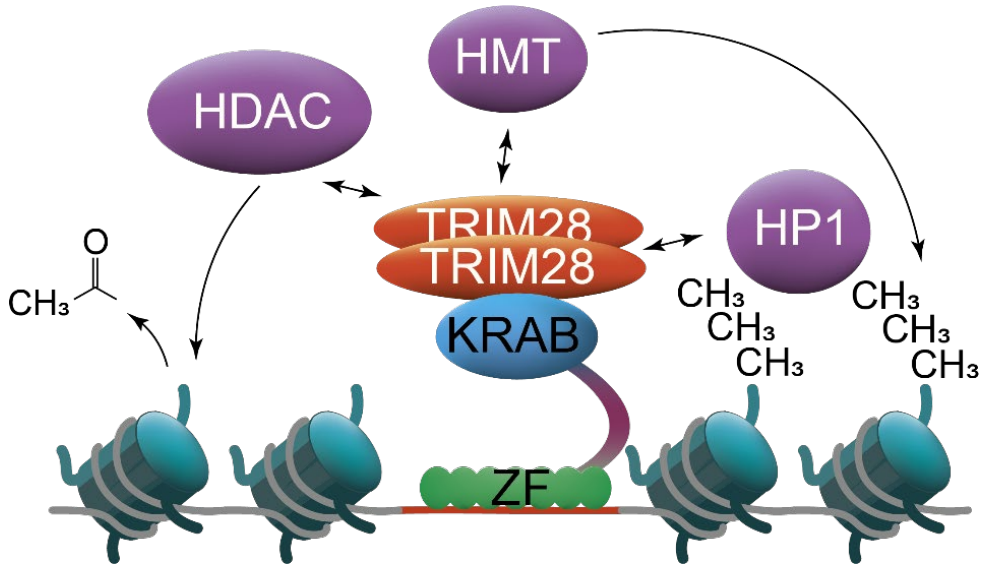


Figure 4: TRIM28 mediated effects of KZFPs on chromatin structure.
Upon recognition of its target sequence (red) by its zinc-finger domains (green, ZF) the KZFP will recruit a dimer of TRIM28 (orange) via its KRAB domain (blue). The TRIM28 dimer in term will serve as a scaffold for several proteins (purple). Histone deacetylases (HDAC, e.g. NuRD) and histone methyltransferases (HMTs, e.g. SETDB1) will deacetylate and methylated histone tails (e.g. Lysine 9 on Histone 3), whereas Heterochromatin protein 1 (HP1) will interact with both TRIM28 and trimethylated lysine 9 on Histone 3 (H3K9me3) to promote heterochromatin formation. Adapted from Randolph et al., 2022 and Wolf et al., 2015b.

Evolutionary history

The KZFP family is thought to have originated 600 million years ago with the appearance of PRDM9 (also known as Meisetz) (Birtle and Ponting, 2006). About 280 million years later the modern KZFPs emerged in the last common ancestor between the coelacanth (Latimeria chalumnae) and the tetrapods (Imbeault et al., 2017). Since their appearance a large number of potential KZFPs have been observed in subsequent species, with birds presenting a notable exception (Figure 1). Interestingly flying birds have simultaneously low numbers of KZFPs, high levels of TEs, and small genomes in general (Kapusta and Suh, 2017). The KZFP family has been continuously evolving since its emergence, resulting in a repertoire of KZFPs that is species specific (Figure 5). The rapid evolution of KZFPs primarily occurs through segmental duplications of either entire genes or parts of their zinc-finger arrays (Emerson and Thomas, 2009; Nowick et al., 2010; Stubbs et al., 2011). The occurrence of such duplications is most likely facilitated by the repetitive nature of KZFPs and their organization in clusters (see Figure 1 from the manuscript in Results I) allowing for the recombination between KZFPs and within their zinc finger arrays. This is reflected in the majority of young KZFPs being located on chromosome 19 (see Figure 1 from the manuscript in Results I) and in regions which are hotspots for copy number variations, making chromosome 19 a location for the rapid evolution and adaptation of the KZFP family (Lukic et al., 2014). This rapid evolution can occur because following a gene duplication one copy maintains the previous function of the protein, reducing the negative impact of modifications on the other copy, thus allowing new functions to arise more easily (see Figure 4 from the manuscript in Results I). Of note, this differentiation between two KZFP copies does not have to take the form of different zinc-finger prints, as the example of the two paralogs ZNF417 and ZNF587 shows (Turelli et al., 2020). In this case both ZNF417 and ZNF587 have maintained the same targets but have instead differentiated in the tissues they are expressed in, with ZNF417 and ZNF587 being expressed in different regions of the developing human brain.

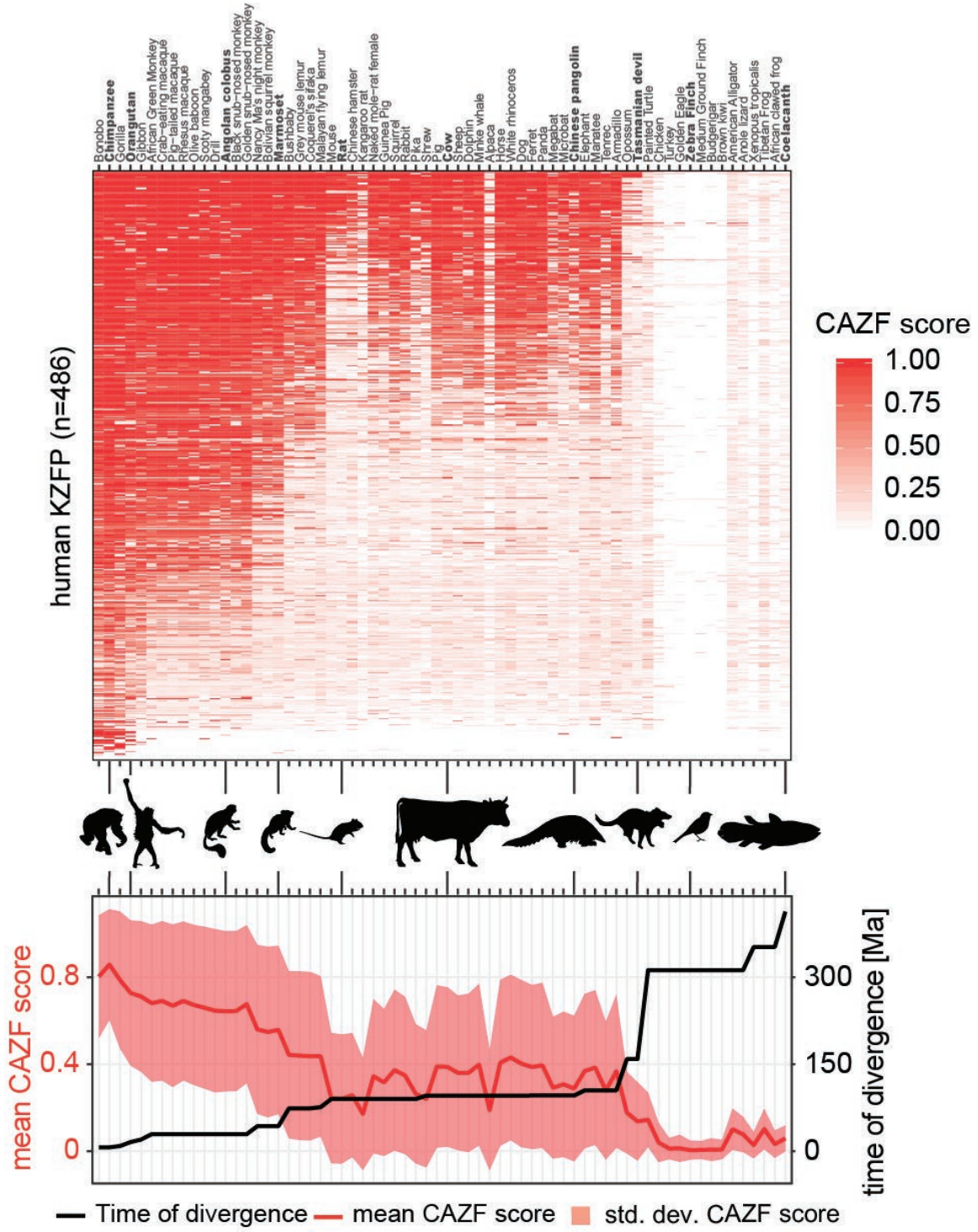


Figure 5: Conservation of human KZFPs across species.
The conservation of human KZFPs shows the continuous evolution of the KZFP family. To compare KZFPs a similarity score is calculated (CAZF score) which considers the number of matches across all residues in the zinc-finger print of KZFP (groups of four DNA-interacting amino acids from one zinc finger). A score of one necessitates an identical zinc-finger print but it does not translate to a fraction of identity thereafter. Top: Heatmap showing the highest CAZF score for the zinc-finger print of each human KZFP (rows) across different species (columns). Species are organized by their time of divergence from human (Kumar et al., 2017). Bottom: The left y-Axis shows the average highest CAZF score (red) across species with its standard deviation (pale red). The right y-axis shows the time of divergence between human and the species (black). Based on unpublished data from Alexandre Coudray and Cyril Pulver. Silhouettes are creative commons (Keesey et al.).

Transposable elements

Transposable elements (TEs) are mobile DNA elements that can insert themselves in new genomic locations. They were first discovered in 1950 by Barbara McClintock (McClintock, 1950) as she described so called mutable elements in maize kernels. What she observed were mobile, inheritable elements which led to the formation of heterochromatin and affected neighbouring genes responsible for kernel colour. Since their discovery, transposable elements have been found to be part of virtually any genome examined (Craig et al., 2015) and when the sequencing of the human genome was first published in 2001 (Venter et al., 2001) it became apparent that a large proportion of it is composed of TEs. Current estimates (Kojima, 2018) identify 48% of the human genome as TE-derived (Figure 6) a significantly larger proportion than the roughly 2% occupied by protein coding exons (Piovesan et al., 2019). Having said that, the majority of TEs in the human genome are inactive, with only a few copies of certain LINE1, SINE and SVA elements remaining active (Kazazian and Moran, 2017). Because of their large variety, TEs are classified into several groups based on their modes of transposition and sequence similarities.

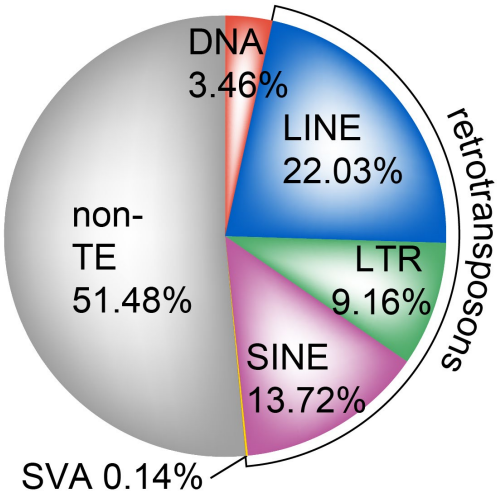


Figure 6: Transposable elements in the human genome.
Pie chart showing the occupation of the genome by different TE families. DNA transposons are shown in red. Retrotransposons include the Long Terminal Repeat (LTR) family in green and the three non-LTR families: Long Interspersed Elements (LINEs) in blue, Short Interspersed Elements (SINEs) in magenta and SINE-Variable Number of Tandem Repeats (VNTR)-Alus (SVAs) in yellow. RC/Helitrons were omitted due to them only occupying ~0.01% of the human genome. Based on data from Kojima, 2018.

Classification

The first major distinction to be made when classifying TEs is based on the template they use to transpose. DNA transposons move as pieces of DNA, generally leaving their initial location, whereas retrotransposons generate a copy of themselves via an RNA intermediate that is reverse-transcribed into DNA and inserted in a new genomic location. The second major distinction that can be made, is between elements which themselves encode the machinery for their transposition (autonomous) or elements that rely on high-jacking it from other elements (non-autonomous). Both DNA- and retro-transposons have autonomous and non-autonomous elements (Figure 7). Finally, other mechanistic and structural similarities are used to group TEs into families and sequence similarity is used to further subdivide those into subfamilies. Here an overview of the major classes of TEs is presented with an emphasis on important families.

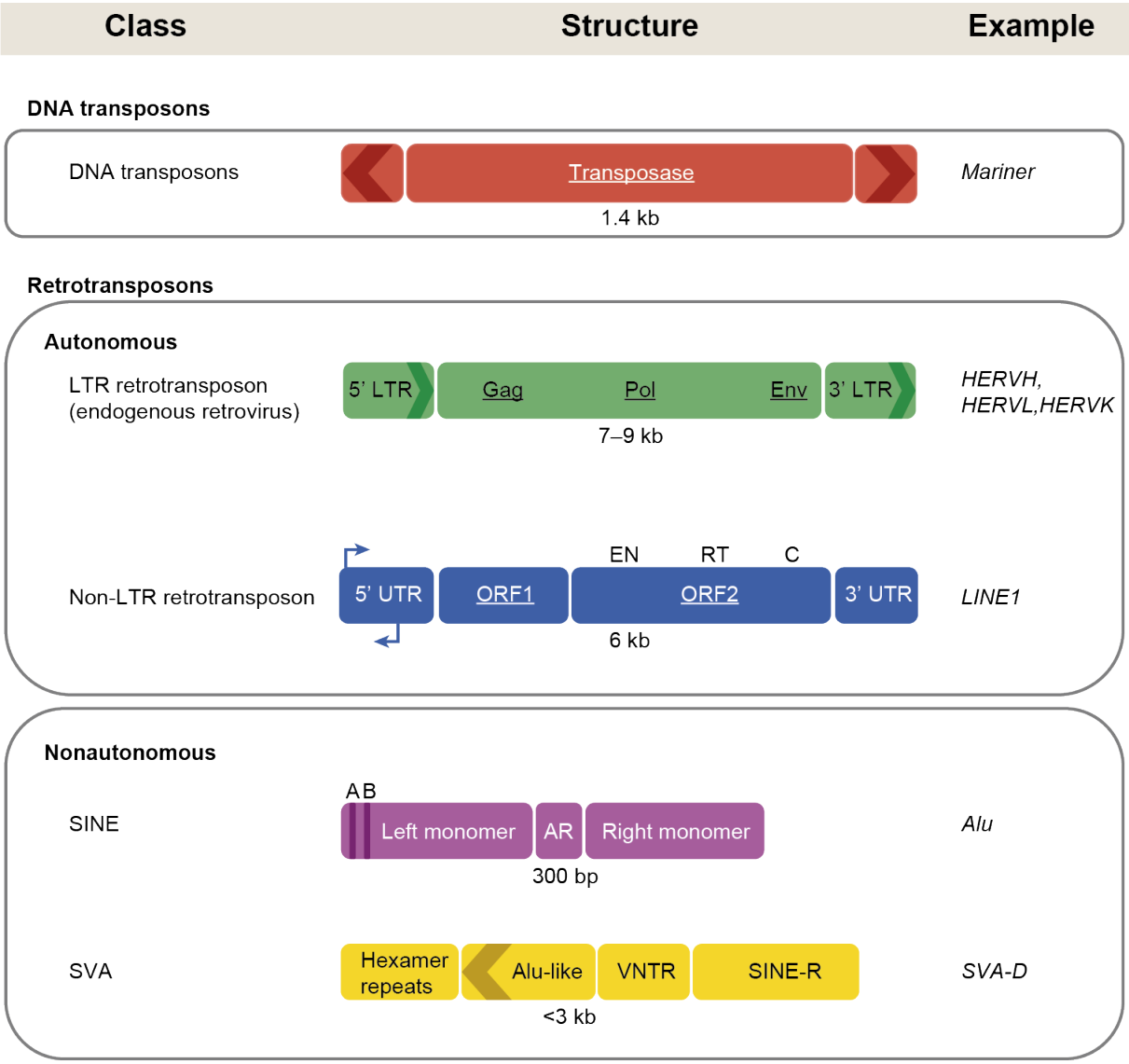


Figure 7: Structure of transposable elements.
An example of each main class of transposable elements is depicted with average genome length indicated underneath. DNA transposons (red), LTR retrotransposons (green), and LINEs (blue) are autonomous because they encode the proteins necessary for their transposition. SINE (purple) and SVA (yellow) nonautonomous retrotransposons spread with the help of LINE-encoded trans-acting functions. Protein-encoding segments are underlined. Abbreviations: LINE, long interspersed nuclear element; LTR, long terminal repeat; ORF, open reading frame; SINE, short interspersed nuclear element; SVA, SINE–VNTR–Alu; UTR, untranslated region. Adapted from Friedli and Trono, 2015.

DNA transposons

Even though DNA transposons occupy a much smaller proportion of the human genome than retrotransposons they are still immensely successful and can be found in all three branches of the tree of life: bacteria, archaea and eukaryotes (Craig et al., 2015). Structurally, most autonomous DNA transposons encode a transposase gene flanked by two recognition sites (Figure 7). The transposase protein will bind to those recognition sites, excise the DNA transposon and reinsert it at a new location (Craig et al., 2015). Non-autonomous DNA

transposons, so called Miniature Inverted-Repeat Transposable Elements (MITE), lack the transposase gene and only contain the recognition sites (Feschotte and Mouchès, 2000). This principle is also used in a lab setting to stably insert transgenes into a new genome, by flanking the transgene with recognition sites for a DNA transposase that is then ectopically expressed (Aronovich et al., 2011). DNA transposons are classified by the conserved catalytic sites of their transposase, which facilitate the breaking and joining of DNA strands. Consequently, there are four major groups of DNA transposons: DDE transposases, tyrosine-histidine-hydrophobic-histidine (HUH) transposases, tyrosine-transposases, and serine-transposases (Craig et al., 2015). There is no known DNA transposon currently active in the human genome however they are thriving in other mammals such as bats (Feschotte and Pritham, 2007).

LTR retrotransposons

Long Terminal Repeat (LTR) retrotransposons are the first major class of retrotransposons (Figure 7) and occupy around 9 percent of the human genome (Figure 6). They are related to retroviruses but lack the capacity to form viral particles and infect other cells, hence they are also called endogenous retroviruses (ERVs). LTR sequences, located at both ends of the element, are flanking the characteristic coding domains of exogenous retroviruses Gag, Pol and the often missing Env (Figure 7). These domains encode the necessary genes for viral particle formation but are often mutated and no longer functional in the vast majority of ERVs (Gifford et al., 2018). Transposition of LTR elements involves the formation of a double-stranded DNA intermediate by reverse transcription in the cytoplasm inside a virus-like particle, which is then re-shuttled into the nucleus and inserted into the genome (Figure 8). However, in the human genome the majority of LTRs (85%) are present in the form of so-called solo LTRs (Belshaw et al., 2007). These are the product of recombination between the repeated 5' and 3' LTR, leading to the excision of the internal coding region (Figure 9) and removing their ability to transpose. Despite a few LTR retrotransposons still having intact ORFs and transposition events have occurred in the common ancestor of human and chimpanzee, there is no evidence of active LTRs in the human genome (Cordaux and Batzer, 2009).

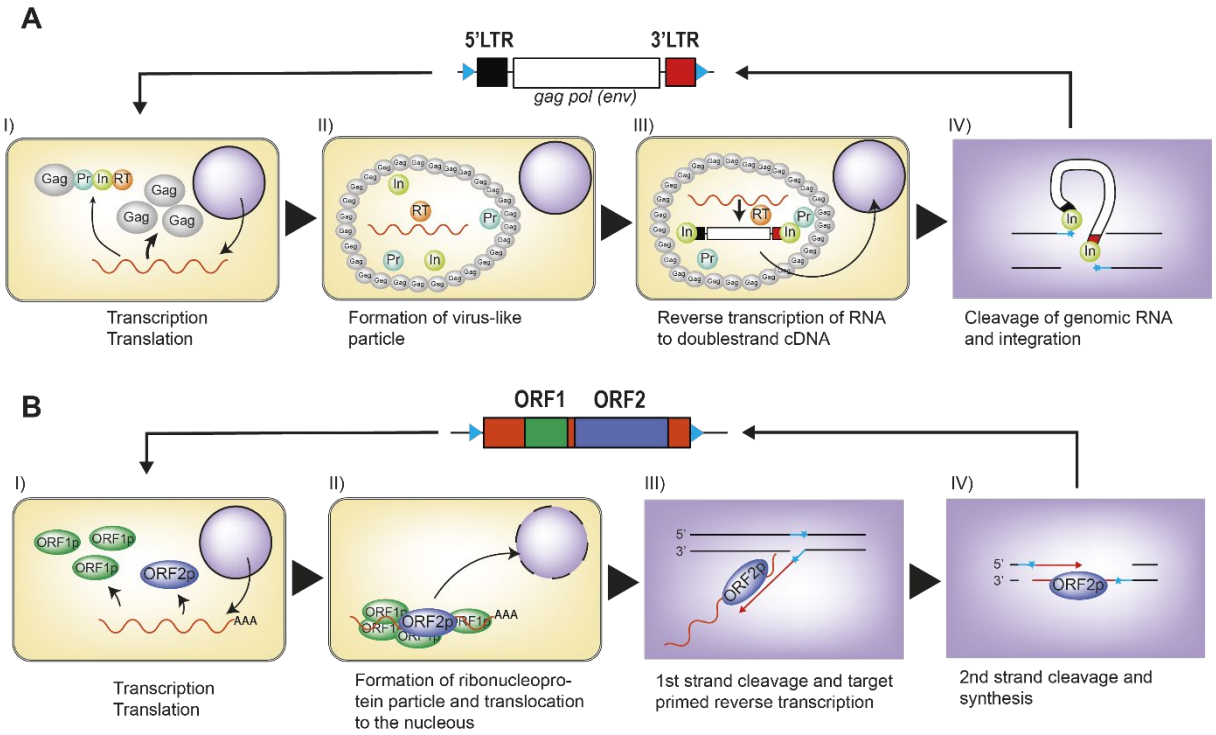


Figure 8: Long terminal repeat (LTR) and non-LTR retrotransposition mechanisms.
A) Top: Structure of a typical LTR element encoding for group specific antigen (gag), pol proteins and optionally env proteins flanked by a 5' and 3' LTR (black and red) and two target site duplications (TSDs) (light blue triangles). TDS represent the recognition site of the integrase being duplicated through DNA repair. Bottom: I) LTR mobilization starts with mRNA transcription and translation to yield mostly Gag and a few Gag–Pro–Pol fusion proteins. The fusion proteins consist of a Gag polyprotein (Gag), a protease (Pr), an integrase (In), and a reverse transcriptase (RT). II) Gag proteins build a virus-like particle in the cytoplasm (light brown) and encapsulate the fusion proteins, which are processed into separate mature proteins. III) The ERV mRNA is then reverse transcribed, generating a dsDNA. This dsDNA and the integrase build a preintegration complex which enters the nucleus (purple). IV) The integrase then creates a double-strand DNA break, followed by genomic integration of a new LTR copy. TSDs are indicated in light blue. B) Top: Structure of a full-length LINE1 element (red) encoding two proteins ORF1p (green) and ORF2p (blue) and flanked by TSDs (light blue triangles). Bottom: I) LINE1 mobilization begins with transcription of an LINE1 mRNA, which is translated to yield ORF1p and ORF2p. II) ORF1p, ORF2p, and the LINE1 mRNA form a ribonucleoprotein particle in the cytoplasm (light brown) that re-enters the nucleus (purple). III) The ORF2p endonuclease cleaves the first genomic DNA strand at its recognition site (light blue), while its reverse transcriptase uses a now free 3' OH group as a primer for reverse transcription of the LINE1 mRNA. IV) Following second-strand DNA cleavage, a new LINE1 copy is integrated into the genome and is typically flanked by TSDs (light blue). Adapted from Gerdes et al., 2016.

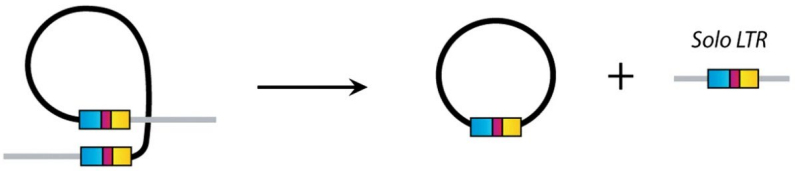


Figure 9: Solo LTRs.
Recombination between the two LTRs (blue-red-yellow box) of an LTR element resulting in the excision of the internal coding region (black) the formation of a solo LTR in the genome (grey). The different regions of an LTR are shown: The unique 3' region in blue, the repeat region in red, and the unique 5' region in yellow. From Gifford et al., 2018.

Non-LTR retrotransposons

The remaining retrotransposons, the so called non-LTR retrotransposons are the largest group of retrotransposons, occupying more than a third of the human genome (Figure 6). They encompass the Long and Short Interspersed Elements (LINEs and SINEs) and the SINE-Variable Number of Tandem Repeats (VNTR)-Alus (SVA) families. LINE elements are the autonomous elements of this group, encoding two proteins called ORF1p and ORF2p necessary for their transposition (Alisch et al., 2006). The much smaller and much more abundant ORF1p binds to the LINE RNA and acts as a chaperone, whereas ORF2p binds to the 3' end of the LINE RNA and leads to the simultaneous reverse transcription and integration of the LINE in the so-called target primed reverse transcription (Figure 8). In this process the ORF2p protein creates a nick in the target DNA and subsequently uses the freed-up hydroxyl group as a primer for reverse transcription of the RNA (Gerdes et al., 2016). This has the advantage of reverse transcribing and integrating the sequence in one step but also has the tendency to terminate prematurely, leading to many integrants being truncated in their 5' ends (Warren et al., 2008). For LINE elements these truncations notably lack the endogenous promoter of the elements which is located in the 5' end (Figure 7). Both SINE and SVA elements high-jack the ORF2p proteins from LINEs resulting in the reverse transcription and integration of their own RNAs using the same mechanism (Craig et al., 2015).

Effects of TE activity

It was the effects of TEs on neighbouring gene expression that led to their original discovery by Barbara McClintock in 1950 (McClintock, 1950). She observed the change in maize kernel colour caused by TE insertions leading to the silencing of neighbouring genes responsible for said colour. Since then a wide range of other effects have been described that can be broadly split into three categories: genomic re-arrangements, disruption and formation of genes, and formation of new gene regulatory networks (Figure 10). Most interesting for this project is if TEs affect neighbouring genes through embedded promoters or binding sites for enhancers or silencers as they can directly implicate KZFPs in gene regulation. KZFPs targeting those TEs

can either influence the activity of promoters and enhancers or be the trans molecule recruited to silencers in turn making KZFPs important players in these regulatory networks.

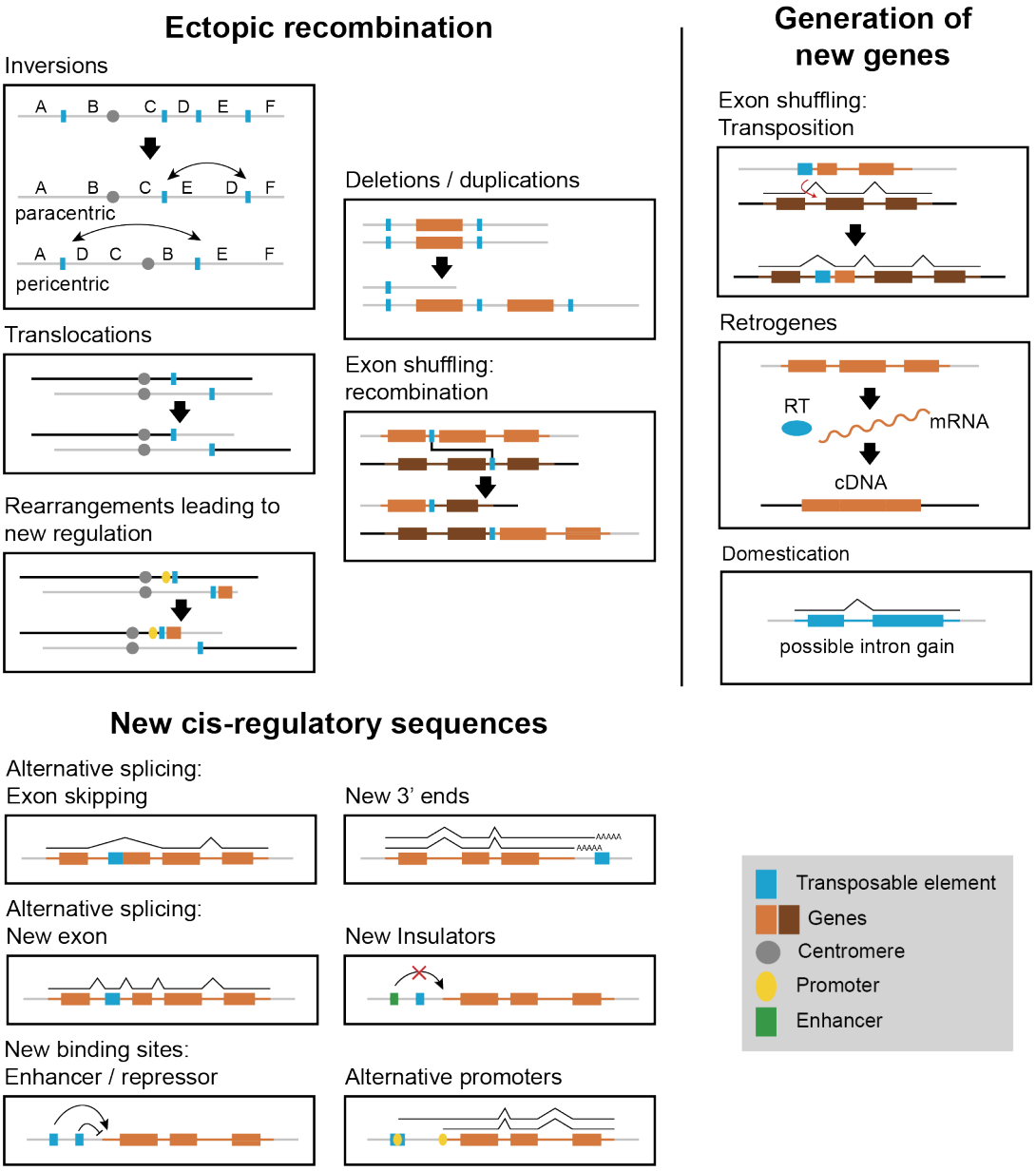


Figure 10: Genomic innovations and re-arrangements mediated by transposable elements. Summary on effects TEs can have on the structure of the genome and gene regulation. Adapted from Warren et al., 2015.

TEs as regulatory elements

Already when Britten and Davidson published “Gene Regulation for Higher Cells: A Theory” (Britten and Davidson, 1969) and introduced the concepts later to be named, protein-coding regions, promoters, enhancers and transcription factors, they suggested that the most efficient way for organisms to evolve is not to generate new gene functions but to make use of existing components in novel ways. Interestingly, they also immediately saw the potential of repetitive DNA sequences to serve as so-called sensor genes, now more commonly referred to as promoters and enhancers. The evolutionary benefit of TE embedded regulatory sequences (TEeRS) was thus identified early on, as they allow for the re-use and rearrangement of regulatory sequences. To date many TEeRS have been described, be they TE embedded promoters, enhancers, long non-coding RNAs, or insulators (reviewed in Fueyo et al., 2022). As a consequence, KZFPs with their ability to affect the chromatin state of TEs, are interesting candidates to act in trans on these TE embedded cis-regulatory elements.

TEs in disease

Given the potential of TEs and their TEeRS to promote evolutionary changes described in the previous paragraph, the risk involved with their activity should be apparent. Human diseases linked to TEs range from developmental issues to cancer to infertility (reviewed in Hancks and Kazazian, 2016; Ryan, 2004). All groups of TEs can be involved in disease, with Alus, LINE1, SVA and recently active LTR elements being the predominant culprits (Hancks and Kazazian, 2016). It needs to be stated at this point that despite these findings, there is no inherent advantage for TEs in causing disease. In contrast to viruses for example, whose dissemination can require cell lysis or is aided by disease symptoms such as sneezing and coughing, TEs are only transferred vertically to the next generation and need to be compatible with early life. Thus, disease-causing TE insertions are unlikely to propagate over long periods of time, which is in accordance with the observation that majority of disease-causing TE insertions are fairly recent.

Expression of TEs

In light of this interdependence between TEs and their host genomes, it is not surprising that TEs are generally silenced by repressive epigenetic marks such as DNA methylation and H3K9me3. Theoretically, any transposition of a TE that does not reach the germline does not result in its spread and not increase its evolutionary fitness, it only represents a risk for the host and thus the TE. However, there are two major epigenetic reprogramming events in mammalian genomes erasing most repressive marks and providing a permissive environment for TE expression. These occur in the early embryo after fusion of female and male gametes, and during development and migration of the precursor cells that will form future gametes (Cantone and Fisher, 2013). The reprogramming is necessary for the establishment of a

totipotent state of cells in these stages of development (Cantone and Fisher, 2013). As a consequence many KZFPs are highly active in embryonic stem cells and reproductive tissues (GTEx Consortium et al., 2017) along with other mechanism regulating TE expression such as the Piwi interacting RNA pathway (Lau et al., 2006). Other instances of increased TE expression are often associated with pathologies such as cancers and autoimmune disorders (Kazazian and Moran, 2017) or with cellular senescence (De Cecco et al., 2019). An exception to this is the brain which relative to other somatic tissues seems to be more permissive for TE expression and retrotransposition (Baillie et al., 2011). Despite the previously mentioned cases the main arena for the activity of TEs remains early development and germ cell precursors.

Quantification of TE expression

The quantification of RNA originating from repetitive elements is challenging for many of the reasons described for ChIP-seq data later. Sufficiently long and preferably paired-end reads are necessary to circumvent the problem of multi-mapping (Sexton and Han, 2019). This is quite relevant as many large publicly available datasets rely on short read sequencing such as single end 35bp, rendering the analysis difficult. In this case it becomes advisable to desist from quantifying individual TE loci but instead quantify the expression of a subfamily as whole. To this end the different reads on all the members of a subfamily can be aggregated to one value, similar to what is done for different transcripts of the same gene, alleviating the issue of multi-mapped reads. Alternatively, all reads can be mapped to consensus sequences of the TE subfamilies, treating each consensus like a mini genome and removing the potential for multi-mapping. Additionally, to quantifying the reads correctly, the location of the TE needs to be considered. Many TEs are located in introns or are at the 3' end of genes, this can allow for the transcription of the TE from the promoter of that surrounding gene, making the expression of TE completely coupled to the gene and not affected by TE specific regulation. In this case reads originating from such TEs should be removed from further analysis. If the library preparation allows to distinguish between the strands of RNA, intronic TEs that are on the opposite strand of their surrounding gene can still be analysed, as reads originating from either the gene or the TE transcription can be distinguished.

Chromatin Immunoprecipitation followed by massively parallel sequencing

A major part of this project is aimed at the identification of genomic target sites of transcription factors. In order to identify these DNA binding sites a commonly applied method is chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-seq) (Mardis, 2007). Here this method, as well as a few considerations which need to be taken when working with KZFPs, are briefly introduced.

Principle

In a ChIP-seq experiment, intact cells are treated with formaldehyde in order to covalently fix (crosslink) all DNA-interacting proteins to the DNA, thus conserving a snapshot of the state of the cell (Figure 11A). As the crosslinking is proximity based and proteins can randomly get crosslinked to DNA, a large number of cells (30 million in our protocol) are treated to average out these random events. After crosslinking, cells are lysed, the chromatin is extracted, fragmented into pieces of approximately 300bp, and the protein or histone mark of interest is immunoprecipitated (IP) using a specific antibody (Figure 11B). This IP allows for the isolation of the targeted protein and because of the crosslinking any DNA sequences attached to it. The quality of the antibody is a main determining factor of the quality of a ChIP, as DNA sequences originating from specific or non-specific antibody binding cannot be differentiated later. Furthermore, as there is always some unspecific DNA present in the IP, it is necessary to set aside an aliquot of the material before the IP to serve as an estimation of the background (Figure 11B, Input). After the IP the crosslinks for both the IP and the Input are reversed using heat, releasing the DNA fragments from the proteins. These fragments are then purified (Figure 11C), and ligated with adaptors for short read sequencing (Figure 11D). The reads obtained from the sequencing are subsequently aligned to a reference, usually the human genome, and the amount of reads for a given location are compared between the IP and the Input (Figure 11E). A statistical model (Zhang et al., 2008) is then used to determine regions with significantly higher reads in the IP than the Input (peaks), which are considered binding regions of the protein or regions carrying the probed histone marks.

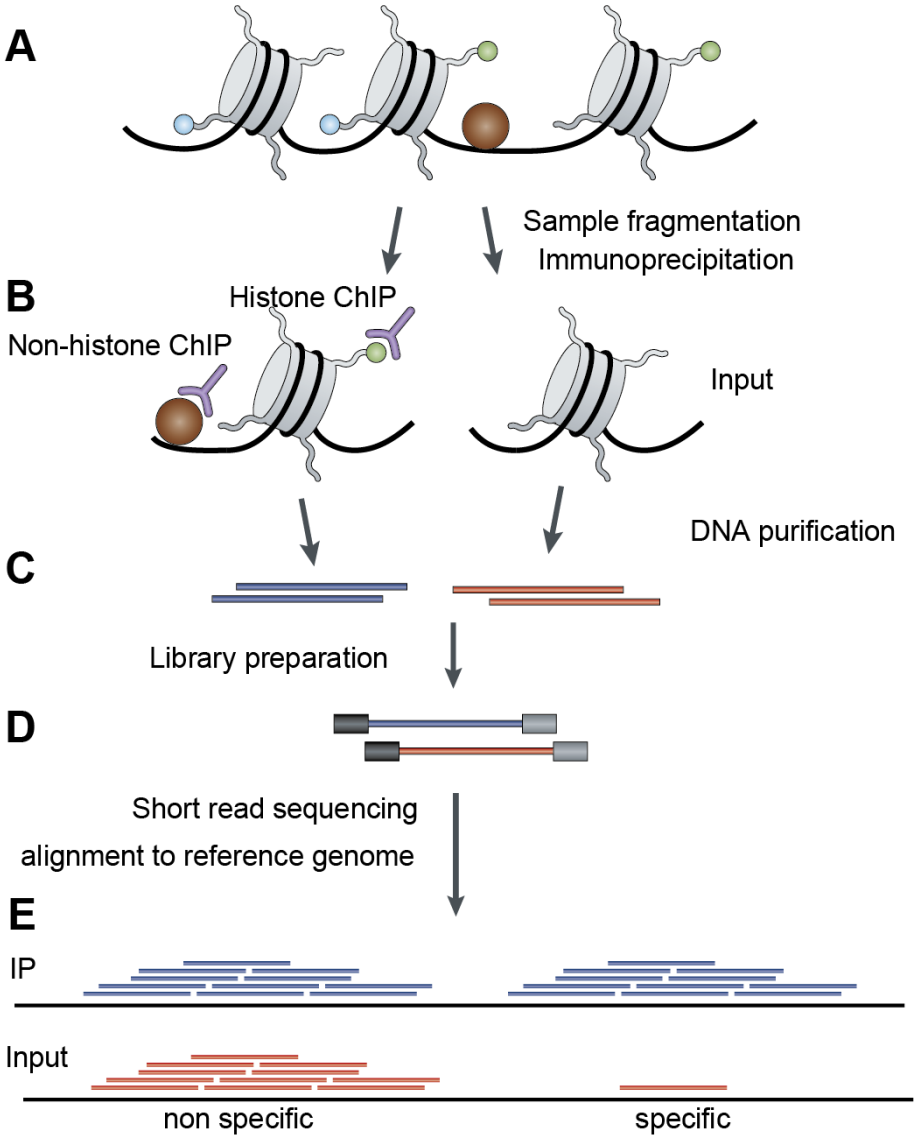


Figure 11: Overview of a ChIP-seq work flow.
A) Proteins are crosslinked to DNA using formaldehyde and the DNA is subsequently fragmented into pieces of approx. 300 bp. B) Samples are immunoprecipitated (IP) with specific antibodies (left) or kept as input samples (right). C) Crosslinking is reversed and DNA fragments are purified. D) Sequencing adaptors are ligated to the DNA fragments which are subsequently sequenced using massively parallel sequencing. E) Resulting reads are aligned against a reference sequence and IP and Input are compared to define specific and non-specific enrichments. DNA is shown in black, histones in grey, histone marks in blue or green, the protein of interest in brown, specific antibodies in purple, immunoprecipitated sequences in blue, input sequences in orange and sequencing adaptors as grey boxes. Adapted from Park, 2009.

ChIP-seq applied to KZFP studies

There are two major considerations when using ChIP-seq to study KZFPs. First it is very difficult to have specific antibodies against individual KZFPs. They are both in terms of their amino acids sequence and structurally extremely similar to each other, making it challenging to target only one at a time. To circumvent this issue, we employ ectopically expressed KZFPs that are tagged with a triple Human influenza hemagglutinin (HA) tag, a small epitope derived

from the human influenza virus against which excellent antibodies are available. This tag is added to the C-terminal end of the proteins and allows for highly specific IP while, likely not interfering with the protein fold based on X-ray crystallography of untagged zinc-fingers (Elrod-Erickson et al., 1998). Even though an influence of the HA tag on DNA binding cannot be excluded as a crystal structure is unavailable, similar results have been obtained using a different tag (GFP) (Schmitges et al., 2016). Furthermore superimposing ChIP-exo signal from endogenous TRIM28 with the ones from tagged KZFPs shows a close co-localization of the two signals (Imbeault et al., 2017), further diminishing the potential influence of the HA tag on DNA binding. Ectopic expression of KZFPs further enables the study of all KZFPs in the same cell type, even if many of them would not be expressed endogenously, a factor which is essential to ensure the comparability of different experiments. The second consideration is that KZFPs primarily bind to elements that are repeated in the genome. These regions present challenges for short read sequencing, as many reads cannot be assigned to a single locus (uniquely mapped) but instead could originate from multiple loci (multi-mapped). In traditional ChIP-seq experiments multi-mapped reads are discarded as they represent a minor proportion of the overall data and can lead to biased results, for KZFPs however this cannot be automatically assumed. The main solution for this challenge is to use longer reads when sequencing, thus increasing the amount of uniquely mapped reads. A study showed that using reads with at least 75bp in length from paired-end sequencing, where both ends of a fragment are sequenced, or 100bp reads from a single end, allow for the unique mapping of a vast majority of TEs (Sexton and Han, 2019). Paired-end reads are preferable over longer single end reads as the unique placement of one end allows the placement of the other, even if the latter is not in a uniquely mappable region, and having the information split in two paired reads ~150bp apart (when sequencing 300bp fragments) increases the chance of this occurring. In these cases where the reads are long enough it is still preferable to remove multi-mapped reads as they can interfere with the peak calling. Specifically, local enrichments can be artificially created due to most alignment software distributing multi-mapped reads randomly between possible locations for both IP and Input samples. If these distributions are uneven they can represent significant differences during peak calling and thus randomly create peaks. If shorter reads such as the popular 35bp single end reads have to be used it is advisable to group repeated sequences together and use subfamily levels for the analysis. For example, if binding to Alu elements is suspected, rather than relying on peak calling, it is preferable to align the Alu sequences as shown in the supplementary figures of the manuscript in the chapter Results I of this thesis and inspect the localization of both the input and the IP signal over the elements. This approach allows to identify enriched regions and is not influenced by multi-mapping reads as it eliminates the need for peak calling algorithms.

DNA binding motifs

DNA motifs represent the consensus binding sequence of transcription factor. They are calculated screening ChIP-seq peaks for common sequences (Grant et al., 2011) and have proven enormously helpful in the past. An identified motif allows to both screen for potential binding sites in silico and to verify experimentally identified binding sites. They are either represented as sequence logos with their information content (Schneider and Stephens, 1990) or as position weight matrices, which can be interpreted as showing the binding energy for each position and nucleotide (Stormo, 1990) (Figure 12). The information content (Bits) quantifies the information stored in a position regarding the question if the nucleotides there are conserved (always the same) or not. The higher the value for a nucleotide the more valuable it is in discerning conservation from a random distribution. The binding energy, simply put, is a representation of the likelihood of a specific nucleotide being bound relative to a background model, with positive values indicating higher likelihoods and negative value lower likelihoods. In other words, everything with a positive value is more likely to be found in a bound sequence while everything with a negative value is less likely. The advantage of position weight matrixes is that the background model can be varied for every single position whereas the information content assumes equal probabilities for each nucleotide at each position.

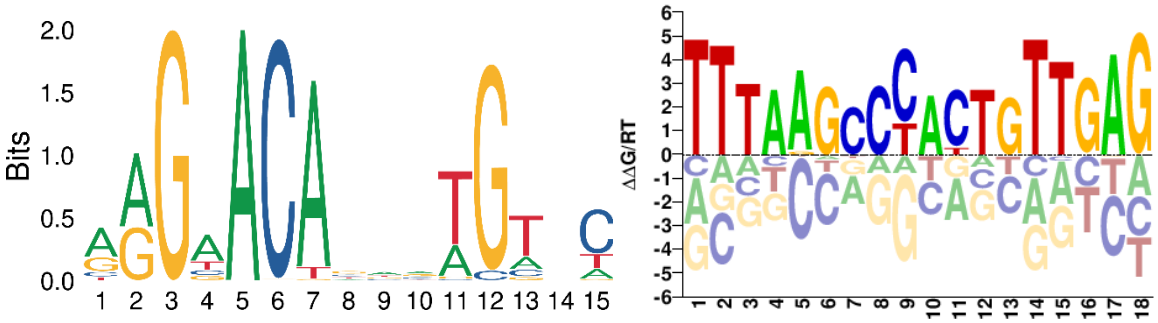


Figure 12: Examples of a sequence logo and a position weight matrix.
The x-axis shows the position in the DNA sequence in base pairs. Left: Sequence logo for AR (MA0007.2) (Castro-Mondragon et al., 2022). The y axis shows the information content for the displayed base at each position. Right Position weight matrix: The y-axis indicates the logarithm of the frequency of a given nucleotide divided by the background frequency. ZNF627 motif is taken as an example from the Cisbp database (Weirauch et al., 2014).

DNA binding motifs and TEs

The identification of DNA motifs located inside repetitive sequences poses both challenges and opportunities, which arise from the multiple, almost identical, copies of a TE throughout the genome. Generally, when identifying motifs in ChIP-seq peaks it can be assumed that only the binding site is conserved and everything else is distributed randomly. Consequently, with a large and diverse number of peaks anything that is conserved can be assumed to be important for binding. TEs break this assumption as very long sequences are conserved, thus motifs on TEs often do not reflect the nucleotides that are essential for binding but rather the consensus sequence of the TE in the region of binding. This however also harbors a tremendous opportunity with a slightly adjusted approach. Usually a random or weighted background distribution of nucleotides is assumed when calculating motifs (e.g. a 25% chance for either A, T, G or C). For TEs there is a unique opportunity as there are generally a large number of unbound sequences available. These sequences harbor point mutations, deletions or insertions that prevent binding and represent an ideal background for motif calculation. Thus, rather than looking for conservation in bound sites, looking for non-conserved nucleotides between bound and unbound sites reveals essential nucleotides for binding. Consequently this approach is very promising for TEs something which was also recognized by a recent review (Fueyo et al., 2022).

Aims of thesis

Both transposable elements (TEs) and KRAB zinc-finger proteins (KZFPs) are elephants hiding in plain sight, they represent immensely important parts of our genome yet they are routinely overlooked. TEs were considered junk-DNA for the longest time even though they make up almost half of the human genome. They were ignored in part, because their repeated and internally repetitive nature makes them difficult to study and even their detection with programs like RepeatMasker was initially aimed at removing (masking) them. Similarly, KZFPs are still routinely overlooked by the scientific community despite being the largest family of transcription factors in humans. They too are both repeated and repetitive, having many paralogs and being highly repetitive in their zinc-finger arrays. Making them difficult to work with as well, due to their tendency to recombine and the fact that obtaining specific antibodies for KZFPs, routinely done for other proteins, remains very challenging. In spite of these challenges more and more important roles of TEs were discovered, elevating them far beyond the junk status and putting a spotlight on KZFPs as their regulators. Key studies performed by our group and others started investigating the KZFP family systematically, revealing their roles as TE and gene regulators. The aim of this thesis is to extend this work, including KZFPs previously left behind and gain a complete overview of the KZFP family's targets and functions. In order to achieve this aim, I first set out to perform ChIP-seq on the KZFPs that had not been attempted yet or which had failed. This was to be followed by a re-analysis of all the data available with the aim to generate one unified data set and to make it available to the scientific community through a web portal (https://tronoapps.epfl.ch/web/krabopedia/). The complete dataset should allow the identification of new evolutionary and functional connections between KZFPs and TEs highlighting the different evolutionary pressures or alterative mechanisms of control at play. The overarching goal for all of these efforts being, to facilitate further study of both KZFPs TEs, and their interplay.

Results I

Contribution

The following manuscript, in preparation to be submitted for peer review, represents the main body of my thesis. I designed the study, designed and performed the experiments, analysed and interpreted the data and wrote the manuscript. Didier Trono designed the study and wrote the manuscript. Chrisitan W. Thorball provided processed data and wrote parts of the manuscript. Michael Imbeault and Sandra Offner performed experiments. Alexandre Coudray, Evarist Planet and Julien Duc provided processed data and developed methods used in the manuscript. Bara Khubieh aided with the web portal.

The genomic targets of the human KRAB-Zinc Finger Protein family

Jonas de Tribolet-Hardy[1], Alexandre Coudray[1], Christian W. Thorball[1], Michael Imbeault[1,2], Evarist Planet[1], Julien Duc[1], Bara Khubieh[1], Sandra Offner[1], Didier Trono[1]*

[1]*School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland*

[2]*Current address: Department of Genetics, University of Cambridge, United Kingdom*

**Corresponding author: didier.trono@epfl.ch*

Abstract

Krüppel-associated box (KRAB) domain-containing zinc finger proteins (KZFPs) are the largest family of transcription factors in humans (Lambert et al., 2018). They encompass 377 protein coding members, and are organized in several cluster throughout the genome. KZFPs represent an evolutionary response to transposable elements (TE), which are mobile DNA elements able to not only disrupt gene regulation and cause disease but also able to serve as regulatory elements. Here we present binding data from almost all human KZFPs (95%) allowing us to identify the targets of the majority of human KZFPs. Our work reveals that the KZFP family has expanded to adapt to the appearance of new TE subfamilies, targeting them with several KZFPs. We further show that the different KZFPs targeting the same TEs arose independently even though the expansion of KZFPs through local gene duplication can generate local groupings. Finally. All the data gathered on the human KZFPs is made available on our web portal the KRABopedia. Together these results represent a comprehensive overview of the binding behaviour of human KZFPs.

Introduction

Krüppel-associated box (KRAB) domain-containing zinc finger proteins consist of multiple C-terminal C2H2 Zinc Finger domains binding to DNA in a sequence specific manner and a N-terminal KRAB domain able to induce silencing of gene expression via interactions with TRIM28 (KAP1) (Urrutia, 2003). The KZFP protein family, consisting of 377 protein-coding members in humans (Table S1), has been continuously evolving ever since the emergence of KZFPs approximately 420 million years ago (Imbeault et al., 2017). The human KZFPs are not evenly distributed throughout the genome, but are predominantly organized in clusters and primarily located on chromosome 19. Even though the function of many KZFPs remains unknown, they are generally associated with repetitive elements (RE), most commonly with transposable elements (TEs); either to silence their transcription (Jacobs et al., 2014) or binding to cis-regulatory elements such as promoters or enhancers which are located within the TEs, so-called TE-embedded regulatory sequences (TEeRS) (Fueyo et al., 2022). TEs are REs that are able to translocate in the human genome and either create copies of themselves (retrotransposons) or leave their original position (DNA transposons). They constitute about 48% of the human genome and are organized in families and subfamilies (Kojima, 2018). The major retrotransposon families are Long Terminal Repeat (LTR), akin to retroviruses, Long and Short Interspersed Elements (LINE and SINE), and SINE-VNTR-Alus (SVA). Both LTR and LINE elements encode their own machinery for retrotransposition and are thus autonomous, whereas SINE and SVA elements are non-autonomous, diverting LINE proteins to retrotranspose their RNA. Uncontrolled TE activity is deleterious to an organism as new

insertions can disrupt the genome and cause disease (Durnaoglu et al., 2021; Hancks and Kazazian, 2016; Kim et al., 2016). As a consequence the activity of TEs needs to be controlled, be it by KZFPs or other mechanism such as piRNAs (Ozata et al., 2019) and the HUSH complex (Seczynska et al., 2021). However, the fact that TEs remain targeted by KZFPs millions of years after they are rendered inactive by mutations, indicates that they have acquired new roles beyond controlling the spread of TEs (Imbeault et al., 2017). Furthermore, co-opting TEs to disseminate regulatory sequences and influencing gene-regulatory networks can greatly accelerate evolutionary processes (Britten and Davidson, 1969), thus a certain level of TE activity increases the evolutionary fitness of an organism. As a consequence, any protein interacting with such sequences is primed to evolve regulatory functions itself. Several KZFPs have been found to be implicated in diverse cellular processes (Chen et al., 2019; Ecco et al., 2017; Hayashi and Matsui, 2006; Lupo et al., 2013; Quenneville et al., 2011; Takahashi et al., 2019; Wagner et al., 2000; Yang et al., 2017b; Zeng et al., 2012), nevertheless, a majority of them still have no reported function. A major hurdle in understanding KZFPs is the identification of their genomic binding sites. Previous efforts (Helleboid et al., 2019; Imbeault et al., 2017) have led to the characterization of more than half of human KZFPs. We present the results of the characterization of an additional 110 human KZFPs giving us data on 95 percent of human KZFPs. We analysed both newly generated and the previously published data and make all results openly available on our web portal.

Results

KZFP clusters show distinct times of expansion and levels of conservation

In order to have clear definitions we performed a census of the chromosomal distribution of human KZFPs. We report 467 juxtaposed KRAB and C2H2 poly-zinc finger domains, 377 of which are protein coding, organized in 31 clusters (Figure 1 and Table S1). Clusters were defined as a group of at least three genes separated by less than 250 kb, in accordance with previous methods (Huntley et al., 2006). With the aim of further characterizing those 31 clusters we evaluated the evolutionary age and degree of polymorphism in the human population for the KZFPs in each of them (Figure S1A and B). We observe that both the ages and degrees of polymorphism of KZFPs within the same cluster are significantly more similar than between clusters (ANOVA p.value < 2e-16 and 2.49e-6) and that the degrees of polymorphism are anti-correlated with the age (Spearman correlation = -0.46, p-value < 2.2e-16). This signifies that KZFP clusters have generated new KZFPs at distinct times and are conserved to different degrees in the human population, which is in accordance with the theory that chromosome 19 serves as the main birthplace of new KZFPs (Lukic et al., 2014).

The binding sites of 95 percent of human KZFPs have been profiled

In order to get a complete overview of the binding sites of human KZFPs we continued to perform chromatin immunoprecipitation (ChIP) as initiated by Imbeault et al., 2017. An additional 110 KZFPs were thus characterized (see Data Availability), resulting in a total of 351 out of the 377 protein coding human KZFPs characterized (Figure 2). We further incorporated experiments by other groups (Frietze et al., 2010; ENCODE Project Consortium, 2012; Yan et al., 2013; Schmitges et al., 2016; Venkataraman et al., 2018; Haring et al., 2021) performed using similar overexpression methods, leading to a total of 357 out 377 KZFPs, with replicates for 80 of them (Table S2). The number of identified peaks varied between experiments (Figure S2 and Table S2) and is not a sufficient indicator of the quality of a ChIP as we see experiments with a low number of peaks having very reproducible binding sites with consistent DNA binding motifs. Consequently, we determined enriched target sequences in the peaks to further analyse the data.

KZFPs are able to target the bulk of human TE subfamilies

To generalize and better understand the ChIP-seq data, enrichments over RE were calculated. The choice to include all REs and not restricting the analysis to TEs stems from the fact that a few KZFPs are enriched on non-TE REs. Enrichments of REs are particularly useful to evaluate ChIP-seq peaks as multiple bound elements of the same RE can fulfil a role similar to replicates and give high confidence in the binding specificity from a single experiment. The specificity of this binding can further be verified by aligning the REs and juxtaposing the original reads, allowing the identification of the targeted region of a RE. These analyses made available on our web portal, allow investigators to view, interpret ad verify individual KZFPs. Interestingly, we often see KZFPs targeting a few, mostly related, subfamilies more significantly than others. This presumably reflects the KZFPs binding to high affinity sites and then, potentially due to non-physiological expression levels, some lower affinity sites which would explain the high affinity sites being more conserved between replicates. In order to capture and generalize this phenomenon we define the targets within 10% of the log10 of the lowest False Detection Rate (FDR) as primary and the remaining as secondary targets, allowing us to compare results between experiments. When examining an overview of the results, we observe a vast majority of KZFPs are targeting TEs (Figure 3A and S3A). LTR and LINE families are the most frequently bound, while SINE and SVA families are targeted by a much lesser amount of KZFPs. Having said that, the amount of SVA binders is remarkable given their low frequency in the genome and will be investigated below. Finally, we have small number of KZFPs binding to DNA transposons as well as some mostly enriched on low complexity and simple repeat regions. If the binding to these non-TE regions is specific, it is hard to evaluate as alignments are not possible. However, KZFPs being involved in the formation of heterochromatin in

telomere and centromere regions is a plausible explanation for such a binding being specific and would represent another function of the KZFP family. When focusing on TEs, our results show that 97% of TEs in the human genome belong to subfamilies targeted by at least one KZFP, a number which is still 66% if only primary targets are considered (Figure 3B). The fact that KZFPs can bind to a vast majority of human TEs, reinforces the link between TEs and KZFPs, as both the majority of KZFPs bind TEs and the majority of TEs are bound by KZFPs.

Appearance of TEs and the KZFPs targeting them coincides for autonomous elements

When considering the time of the appearance of both KZFPs and their target TEs we see that a majority of them coincide (Figures 2D and S2C). For example, the TE families LINE1 and ERVL that emerged 105 million years ago are predominately bound by KZFPs that emerged around the same time, an observation that holds true for the later appearing ERV1 and ERVK families. Interestingly this coinciding appearance does not occur for SINE and SVA elements, both of which are targeted primarily by older KZFPs and both of which are non-autonomous. The lack of visible adaptation of the KZFPs family to SVAs could be explained by their recent appearance, however, the SINEs, more precisely the very successful Alu family are intriguing and merit further study. Specifically, as Alu elements are prominently featured in human diseases (Kim et al., 2016), the evolutionary cost exerted by those diseases must be balanced by some benefit, the nature of which remains unknown. Yet another interesting case are the LINE2 elements, half of which are targeted by KZFPs appearing at the same time as them, whereas the other half is targeted by younger KZFPs (Figure S2C). In contrast to LINE1s, LINE2s are completely inactive (Lovšin et al., 2001) and preventing their spreading has become unnecessary. This is reflected in many older KZFPs binding them, including some that have lost their ability to interact with TRIM28 (Figure S2). The conservation of older, and emergence of new KZFPs targeting LINE2 elements makes a regulatory function of both the KZFPs and the TE elements more likely, especially as several regulatory functions of LINE2 elements have been documented (Cao et al., 2019; Petri et al., 2019). Together these findings show how the emergence of new LINE and LTR elements lead to expansion of KZFPs, most likely to control the spread of those families and allow both the host and TEs to survive. The expansion is followed by a contraction as TEs are rendered inactive by mutations and only the elements conveying a fitness advantage are conserved. KZFPs retained during this contraction are potentially the ones that have taken on regulatory roles. Interestingly non-autonomous elements such as SINEs do not elicit such a KZFP response, thus indicating different evolutionary cost-benefit scenarios in these cases.

SVA are bound in the VNTR region

SVAs are overrepresented in our results when considering their low abundance in the genome, we find more KZFPs enriched on SVAs than the ten times more abundant SINE elements (Figure 2B) (Kojima, 2018). The question of the validity of those signals is important as SVA represent a young dynamic TE family that serves as enhancers in early embryogenesis (Haring et al., 2021; Pontis et al., 2019). When assessing the underlying signals of the SVA binding KZFPs, we see a number of them binding in the 5' repeats and the Alu-like portions of the SVAs (Figure 3A). Furthermore, the previously reported ZNF611 and ZNF91 (Jacobs et al., 2014) bind in the beginning of the Variable Number of Tandem Repeats (VNTR). However, the vast majority of signals are enriched in the later, more variable part of the VNTR. This region does not align between elements, thus a reference point at the end of the VNTR was chosen to assess the generalizability of peaks to the whole subfamily (Figure 3B). We can observe an increased signal relative to all input samples which together with the large number of other experiments we have at our disposal makes unspecific binding unlikely. To further asses these signals we investigated the strongest signal on the variable part of the VNTR stemming from ZNF141. We found the other highly enriched targets of ZNF141 L1PA3 and L1PA2 have well localized specific ZNF141 ChIP signals (Figure S3A). These signals also match a motif identified by (Weirauch et al., 2014) which can be found in a tandem organization in the at 5' end of these elements (Figure S3B), and is repeated in the SVA VNTR (Figure S3C) as well as another target of ZNF141, the SATR1 Satellite repeats which are 76bp imperfect tandem arrays (Hubley et al., 2016). Together these results show a strong rational for ZNF141 having a specificity for GC-rich tandem repeats making a non-biological origin of its VNTR signal unlikely and simultaneously highlight the use of the type of data available on our web portal. More generally, we show a situation where the very young SVA family is able to spread despite being targeted by multiple KZFPs.

Secondary targets reveal evolutionary history of KZFPs

Cases like the previously mentioned ZNF141 demonstrate the information contained in the less enriched secondary targets of a KZFP and due to the relations between the bound sequences, their value when evaluating the validity of primary targets. Interestingly, this can also be observed between different KZFPs when they are in the same genomic clusters. As all KZFPs genes are highly related and the family is constantly evolving, deciphering evolutionary relationships between individual members and finding conserved and novel targets can be challenging. Using the information of the secondary targets, we are able to infer evolutionary relationships of KZFPs that go beyond the resolution between age of a KZFP (Imbeault et al., 2017). To show this we can observe ZFP69 and ZFP69B located in cluster chr1.1 and both ~105 mio. years old (see KRABopedia). ZFP69 has a strong affinity for

mammalian specific LINE1 elements, traces of which can be seen in ZFP69B which is primarily enriched on the LTR HERVH-int (Figure 4A and S4A). When examining the binding locations on these elements (Figure S4B) specific bound locations for those elements can be found that carry their highly related motifs from Weirauch et al., 2014 (Figure S4C). In other instances, we see clusters that seem to evolve around different TE families. For example, clusters chr1.2 (Figure 4B) and chr5.2 (Figure 4C), where KZFPs bind different subfamilies of either LTRs or LINE1s. In order to give some directionality to these discovered relationships that go beyond the age of the KZFPs, we employ the number of miss-sense mutations the KZFPs carry in their C2H2 Zinc Finger domains. This reveals which clone after a duplication was more constrained and which was free to acquire new targets (Figure 4A, B and C). These examples show how the segmental gene duplication of KZFPs can lead to targeting of distinct elements and how we can follow that process by looking at the lowly enriched target subfamilies, allowing for a better understanding of the origin of individual KZFPs.

KZFPs targeting the same TEs arose independently

Given that we observe multiple KZFPs binding to the same TE subfamilies and we see proximal KZFPs sharing some of their targets, the question arises if KZFPs targeting the same TEs tend to colocalize. To evaluate this, we compare the number of KZFPs targeting the same TEs across all KZFPs or for the KZFPs in each cluster individually. We observe a median of 4 KZFPs targeting the same subfamily (only considering subfamilies targeted by at least one KZFP). This goes from 4 to 1.17 after normalizing the number of KZFPs by the number of clusters they are found in Figure 5A (left), which is significantly more than the expected 1.00 if KZFPs are distributed randomly across clusters (p-value < 0.0001, 10’000 iterations). Meaning that given the high number of clusters we would not expect KZFPs to colocalize if they are distributed randomly; however, given the retention of target affinity by neighbouring KZFPs discussed in the previous paragraph this is still surprising. To strengthen this observation, we repeat the same analysis only using the most redundantly targeted TE subfamilies with more than 15 KZFPs each (Figure 5A and S5A). The median of 1.79 is now significantly lower than random (p-value = 0.0274, 10’000 iterations) with an expected value of 1.86, meaning not only are the KZFPs not colocalizing, there seems to be a counter selection against it for the most frequently targeted TEs. To account for secondary targets being more evolutionary passengers than drivers and skewing these values towards a random distribution, we repeated the analysis only considering primary instead of all targets (Figure 5A right, and S5B). In this case the primary targets were still not colocalizing in clusters and instead followed a random model. This independence of KZFPs also manifests itself in the fact that KZFPs targeting the same subfamily generally do not do so in the same location on the TE. For example, the L1PA3 subfamily has binding sites all over (Figure 5B) and the median distance between two peaks

on the same TE integrant is 144 bp for all targeted TEs. These findings show that the different KZFPs targeting the same TE are not highly related duplications but arose independently and thus might have independent, not necessarily complementary, functions.

Repository for KZFP related information

In order to allow for the further investigation of individual KZFPs, we present all information on a central web portal, the KRABopedia (https://tronoapps.epfl.ch/web/krabopedia/). This will enable the analysis of KZFP targets as we as provide the means to assess the quality of the targets.

Discussion

The aim of this study was to complete the characterization of the binding sites of human KZFPs. We have generated data for an additional 110 KZFPs and re-analysed the previously published data to have a consistent dataset covering 95% of human KZFPs. Our findings confirm that KZFPs predominantly bind TEs, with autonomous TEs such as LTRs and LINE1 being more frequently targeted than non-autonomous elements such as SINEs. SVA are an exception to this as they are targeted by many KZFPs relative to their low abundance in the human genome. The difference between autonomous and non-autonomous elements is reinforced by comparing the time of appearance (age) of KZFP and their targets, showing that both coincide for autonomous elements, but not for non-autonomous elements. More precisely, SVAs and SINEs are being targeted by older KZFPs, some of which might not be repressors anymore (Helleboid et al., 2019; Tycko et al., 2020). This discrepancy implies that the spread of these elements was only possible if pre-existing KZFPs were present; to limit their toxicity or, actively aided by KZFPs with functions other than transcriptional repression. The fact that TEs with pre-existing KZFPs spread so efficiently through the genomes also implies the existence of either an additional fitness cost in targeting these elements, or more likely a fitness benefit in not doing so. Such a benefit of the presence and spread of TEs was observed for SVAs in early embryogenesis by Pontis et al., 2019. Furthermore, we were able to show that the TE subfamilies targeted by a KZFP reveal the evolutionary history of the KZFPs. We observe proximal KZFPs sharing the same targets but with different levels of affinities, one stronger and one weaker. This directly reflects the well-established model of KZFP evolution through segmental duplication followed by genetic drift (Lukic et al., 2014), where after a KZFP duplication one clone retains the original function allowing the other clone to acquire new targets. Interestingly this does not result in the clustering of KZFPs targeting the same TEs. Despite multiple KZFPs targeting the same subfamily, we do not see them colocalizing in the same cluster more often than a random distribution. This signifies that the multiple KZFPs targeting a TE subfamily are not merely a by-product of duplications but arose independently. Even though not surprising from an evolutionary perspective, as modification of the binding sites through genetic drift can produce any new binding site independent of the original target, the fact that we have several KZFPs targeting the same TEs arising independently is intriguing and leads to the question of the purpose of these multiple KZFPs. This purpose remains speculative but can range from protection against loss of KZFPs to individual KZFPs binding distinct TEeRS and fulfilling tissue or developmental stage specific roles. These findings highlight both the co-evolution of KZFPs and TEs, as very few KZFPs have TE independent binding sites, and also the fact that this co-evolution can take other forms than an arms-race between KZFP repressors and deleterious TEs. Our data supports a model where spread of

new TEs families such as LINEs or LTRs elicit an expansion and adaptation of the KZFP family, followed by slow inactivation of TEs and loss or adaptation of KZFPs targeting them. Only TEs and KZFPs with regulatory or other fitness-promoting functions should escape this process and be conserved. A circumstance which might be used by new TEs carrying bindings sites for those older regulatory KZFPs. Additionally, we can show KZFPs targeting the same elements arise independently and not locally through segmental gene duplication, the purpose of which remains mysterious.

Data Availability

All results regarding the TE targets for a KZFP as well as a verification through MSA plots, evolutionary ages and polymorphism are available on our web portal the KRABopedia (https://tronoapps.epfl.ch/web/krabopedia/). ChIP–seq and -exo data have been deposited in the Gene Expression Omnibus (GEO) database with the accession number GSE200964.

Acknowledgements

We thank Charlène Raclot, Mahsa Sanati and Kerim Benbouhafs for technical assistance, Bastien Mangeat and the Gene Expression Core Facility at EPFL for sequencing and SCITAS for computing infrastructure.

References

The references for this paper have been merged with the Bibliography of the thesis.

Figures

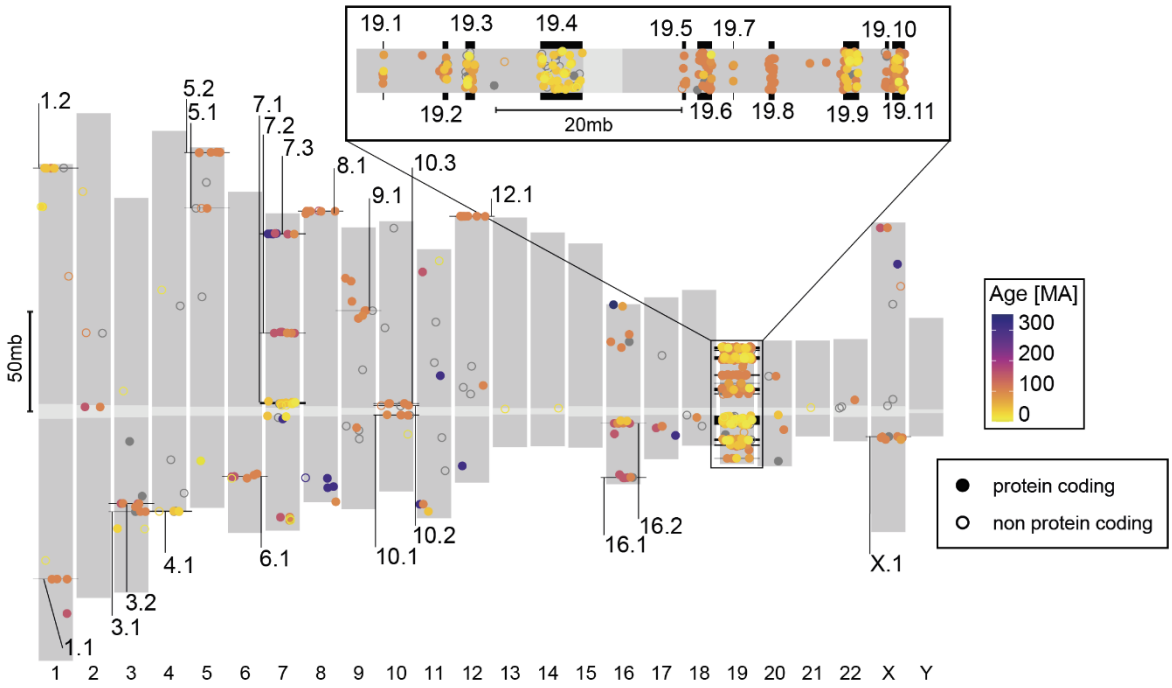
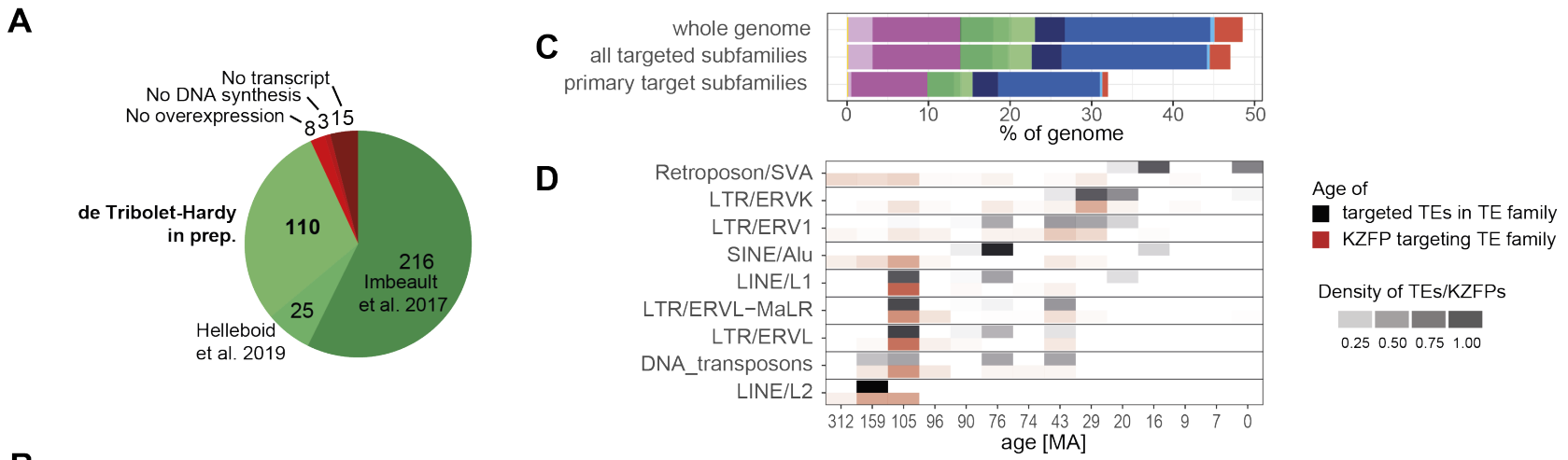


Figure 1: Map of human KRAB-Zinc Finger Proteins (KZFP)

Dots indicate relative chromosomal position of KZFP genes (defined by juxtaposed KRAB- and zinc finger-coding domains), with the colour code indicative of age (grey for unassigned) and numbered clusters pointed to in black. Non protein coding genes are indicated by a hollow circle. Higher magnification of chromosome 19 is presented on top.

Figure 2


A
No transcript
No DNA synthesis
No overexpression
8
3
15
de Tribolet-Hardy
in prep.
110
216
Imbeault
et al. 2017
25
Helleboid
et al. 2019
C
whole genome
all targeted subfamilies
primary target subfamilies
0
10
20
30
40
50
% of genome
D
Retroposon/SVA
LTR/ERVK
LTR/ERV1
SINE/Alu
LINE/L1
LTR/ERVL−MaLR
LTR/ERVL
DNA_transposons
LINE/L2
312 159 105 96 90 76 74 43 29 20 16 9 7 0
age [MA]
Age of
targeted TEs in TE family
KZFP targeting TE family
Density of TEs/KZFPs
0.25 0.50 0.75 1.00

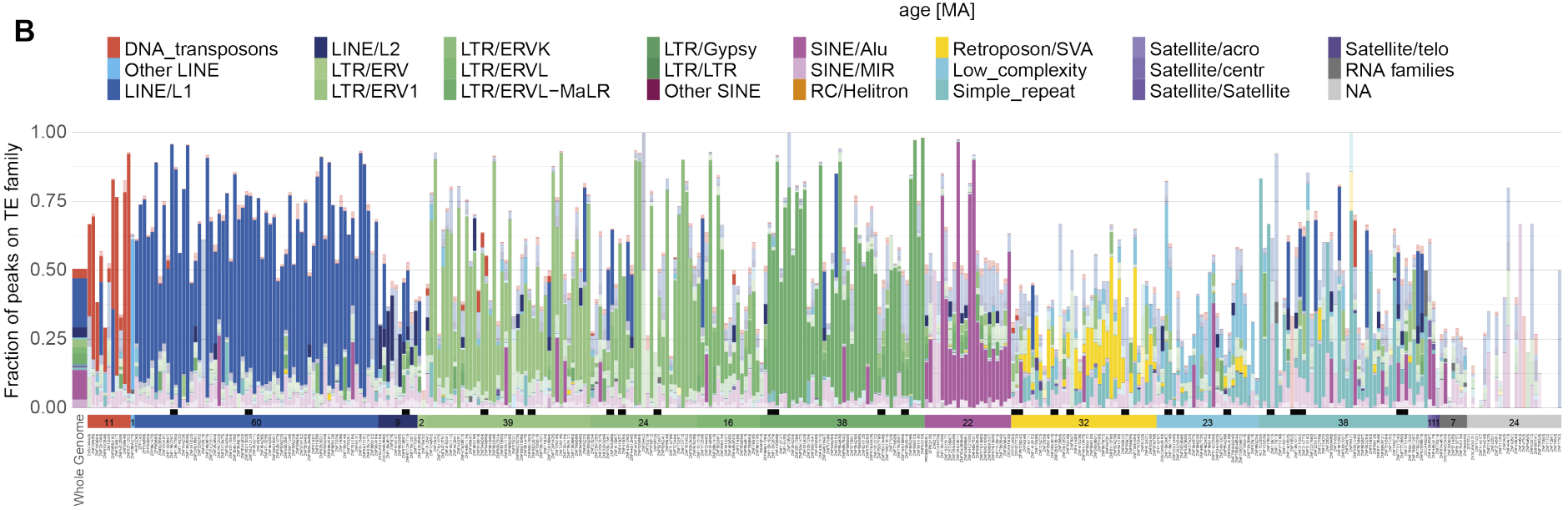

B
DNA_transposons
Other LINE
LINE/L1
LINE/L2
LTR/ERV
LTR/ERV1
LTR/ERVK
LTR/ERVL
LTR/ERVL−MaLR
LTR/Gypsy
LTR/LTR
Other SINE
SINE/Alu
SINE/MIR
RC/Helitron
Retroposon/SVA
Low_complexity
Simple_repeat
Satellite/acro
Satellite/centr
Satellite/Satellite
Satellite/telo
RNA families
NA
Fraction of peaks on TE family
1.00
0.75
0.50
0.25
0.00
11
1
60
9
2
39
24
16
38
22
32
23
38
111
7
24
Whole Genome

Figure 2: Genomic targets of human KRAB-Zinc Finger Proteins

A) Pie-chart of the data on all 377 protein-coding KZFPs. “No overexpression” indicates the number of KZFPs where the codon-optimized construct did not yield sufficient protein. “No transcripts” represents KZFPs with no annotated transcript containing both the KRAB and zinc finger domains simultaneously. “No DNA synthesis” indicates the number of KZFPs CDS that could not be synthesized, with a minimum of two tries. B) Bar graph showing the fraction of peaks over repetitive element (RE) families for all conducted experiments (x-axis), ordered by the most enriched family which are indicated by the horizontal bar below, along with the number of KZFPs for each category. Replicate experiments are indicated by black squares above the horizontal bar. Significant enrichments (FDR < 0.05) are shown in fully opaque colours whereas non-significant enrichments are transparent. The leftmost bar shows the genome percentage of the genome of each RE families, non-RE are not shown. The total number of peaks per experiment is indicated in brackets after the KZFP name below each bar. C) Bar graph showing the genome occupancy of targeted TE subfamilies. The upper bar shows the fractions of the genome covered by TEs, the central bar shows the coverage by all TE subfamilies which are targeted by a KZFPs (FDR < 0.05) and the lower bar shows the coverage of the primary subfamilies per KZFP (10% highest -log10[FDR]). Bars are coloured according to the TE families the subfamilies belong to, with the same colour code as in panel B. D) Age of KZFP and their target TEs. For each TE family (row) the ages of all KZFPs (red) that are highly enriched (as in C) on a subfamily belonging to that TE family are shown. The ages of those TE subfamilies are shown in black. If KZFPs are highly enriched on multiple subfamilies of the same family the most enriched is shown.

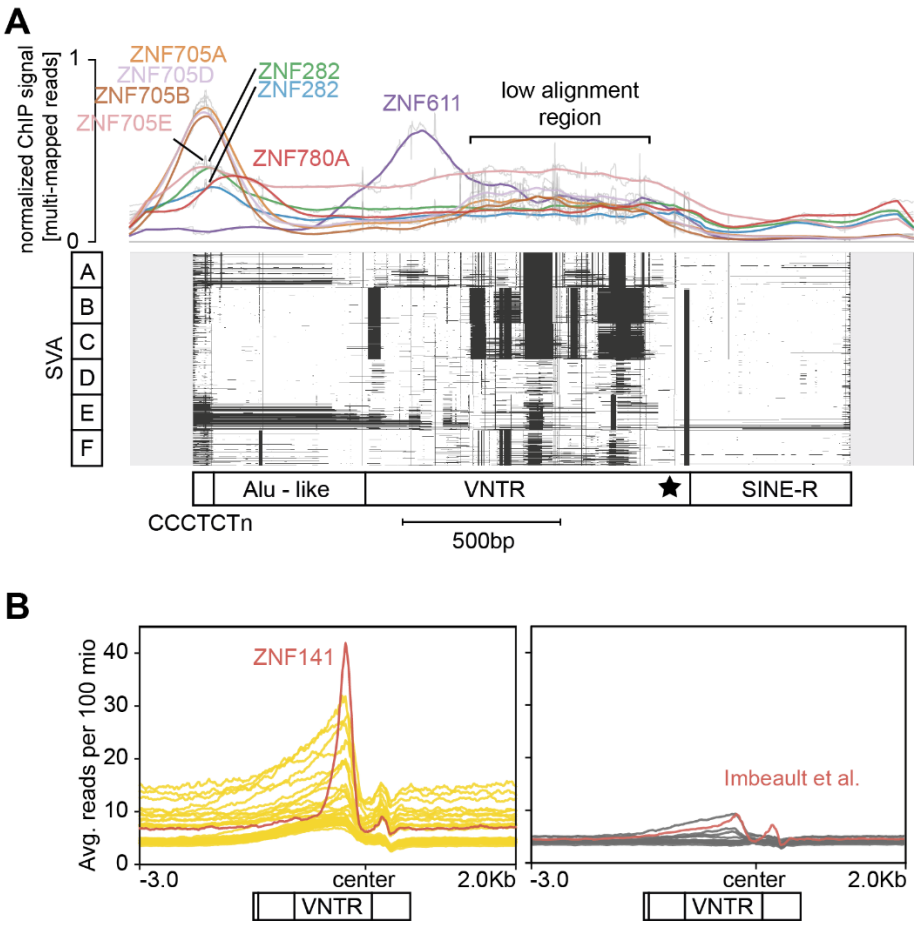


Figure 3: Binding of SVA elements

A) KZFP signal over the multiple sequence alignment (MSA) of SVA subfamilies A to F. Bottom: MSA plot of 100 of the longest SVA sequences for each subfamily indicated on the left, 200 bp of non-aligned extensions are added around elements shown in grey, white depicts aligned regions and black gaps in the alignment. For visibility, places in the alignment (columns) with more than 85% gaps where removed from the alignment. The approximate different domains of the SVAs are indicated below, adapted from (Hancks and Kazazian, 2010), the centre region for panel B is indicated by a star. Top: Line graph of the normalized cumulative reads for each position from indicated ChIP-seq and -exo experiments. B) Signal over the low alignment region of the remaining SVA binders centred on the 3' end of the VNTR without alignment of sequences. Left: ChIP signals for KZFPS enriched on SVAs (ZFP57, ZFP92, ZNF14, ZNF141, ZNF155, ZNF215, ZNF25, ZNF256, ZNF263, ZNF268, ZNF28, ZNF30, ZNF41, ZNF415, ZNF461, ZNF500, ZNF556, ZNF560, ZNF57, ZNF587B, ZNF597, ZNF624, ZNF641, ZNF689, ZNF699, ZNF747, ZNF812, ZNF813, ZNF852 and ZNF878) with the strongest signal for ZNF141 is shown in red. Right: Input signals for the presented ChIPs, the input from Imbeault et al., 2017 is shown in red. Multi-mapped reads where included for the signals in A and B.

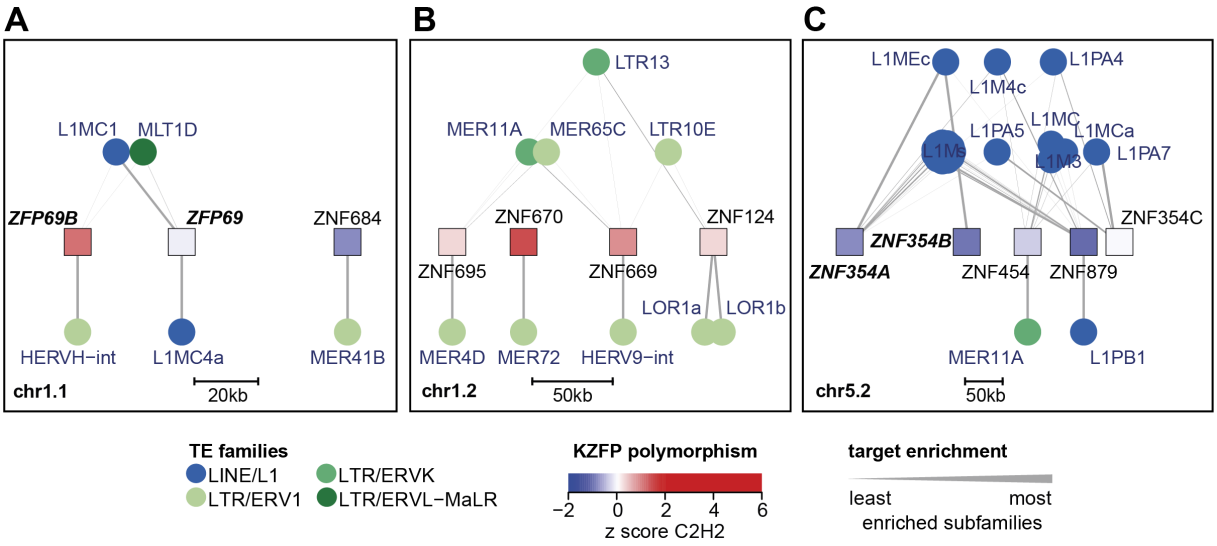


Figure 4: Secondary targets of KZFPs within the same cluster

Networks were targets (circles) of each KZFP (squares) are shown as connected edges and the amount of binding is represented by the line thickness. The thickest line for each KZFP represents the TE subfamily with the highest -log10(FDR) and then scales linearly to the lowest value. For visibility, only the best targets (below) and shared targets (above) are shown. The TE subfamilies are coloured according to their families, the KZFPs are coloured according to the number of C2H2 miss-sense mutation they carry in the human population with red being more polymorphic and blue more conserved KZFPs than average. A) shows cluster chr1.1. B) shows cluster chr1.2. C) shows cluster chr5.2.

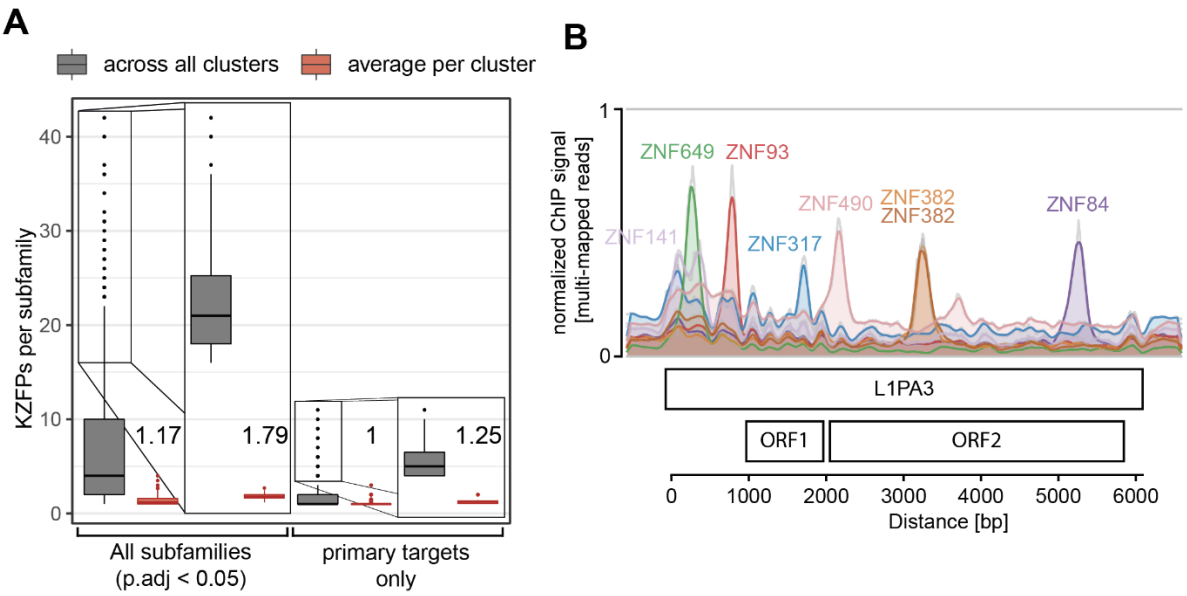


Figure 5: KZFPs targeting the same TE subfamilies do not cluster together

A) Boxplots showing the average number of KZFPs binding to a TE subfamily either for all KZFPs (grey) or in the same cluster (red). Left: All enriched subfamilies (FDR < 0.05), with a zoom on subfamilies with more than 15 KZFPs. Right: Only taking KZFPs which have the subfamily as a primary target (FDR within 10% of the log10[lowest FDR]). Clusters with no KZFPs binding the subfamilies are not considered for the averages. B) Binding sites of KZFPs on L1PA3 elements. 1000 L1PA3 elements were aligned (Figure S5D) and the normalized ChIP-seq and -exo signal is shown for each aligned position. The length of the L1PA3 as well as the location of its two Open Reading Frames (ORF1 and ORF2) are indicated below.

Supplementary Figures

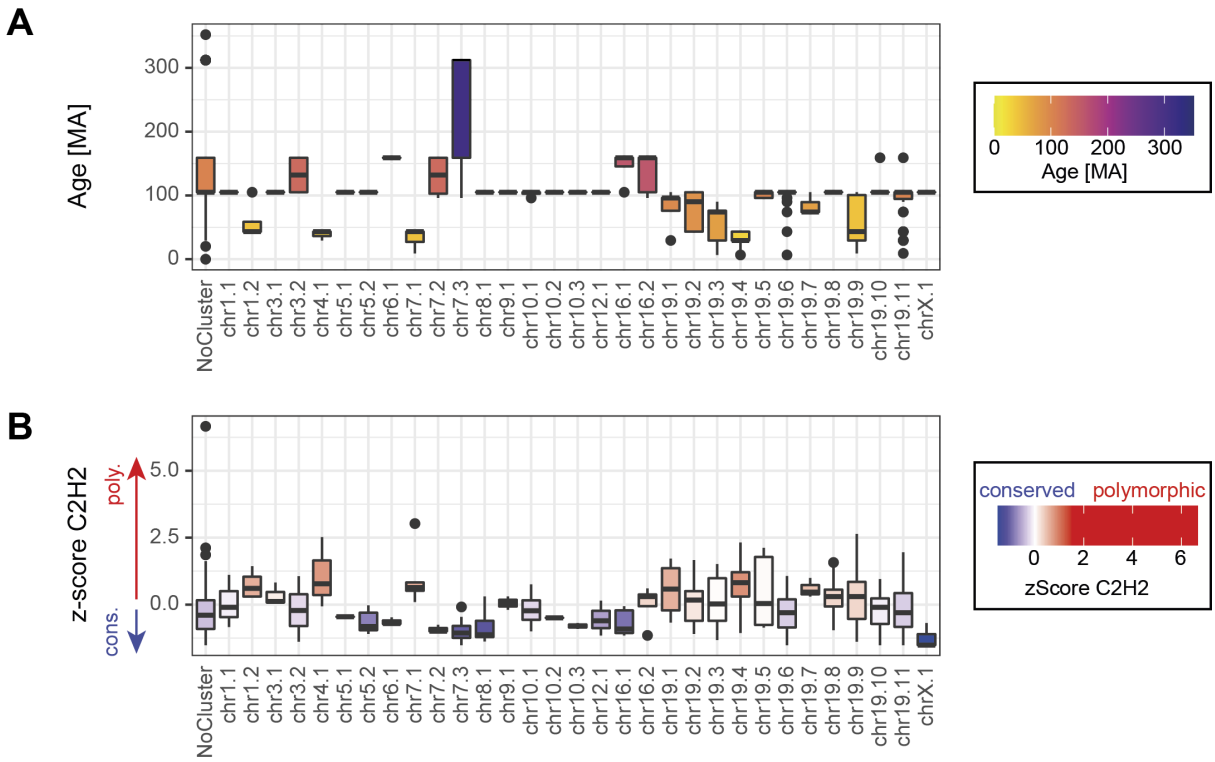


Figure S1: Age and polymorphism of KZFPs in clusters

A) Boxplots of the ages per cluster as defined in Figure 1, coloured by the median age of the KZFPs in the clusters. B) The z-score of the number of missense variants in the in the C2H2 residues of the KZFPs normalized to the number of ZF domains within their canonical transcript, with data obtained from The Genome Aggregation Database (gnomAD) (release-2.0.2). A positive score indicates a KZFP with above average polymorphisms (red) in the human population, a negative score indicates a below average, more conserved KZFPs (blue).

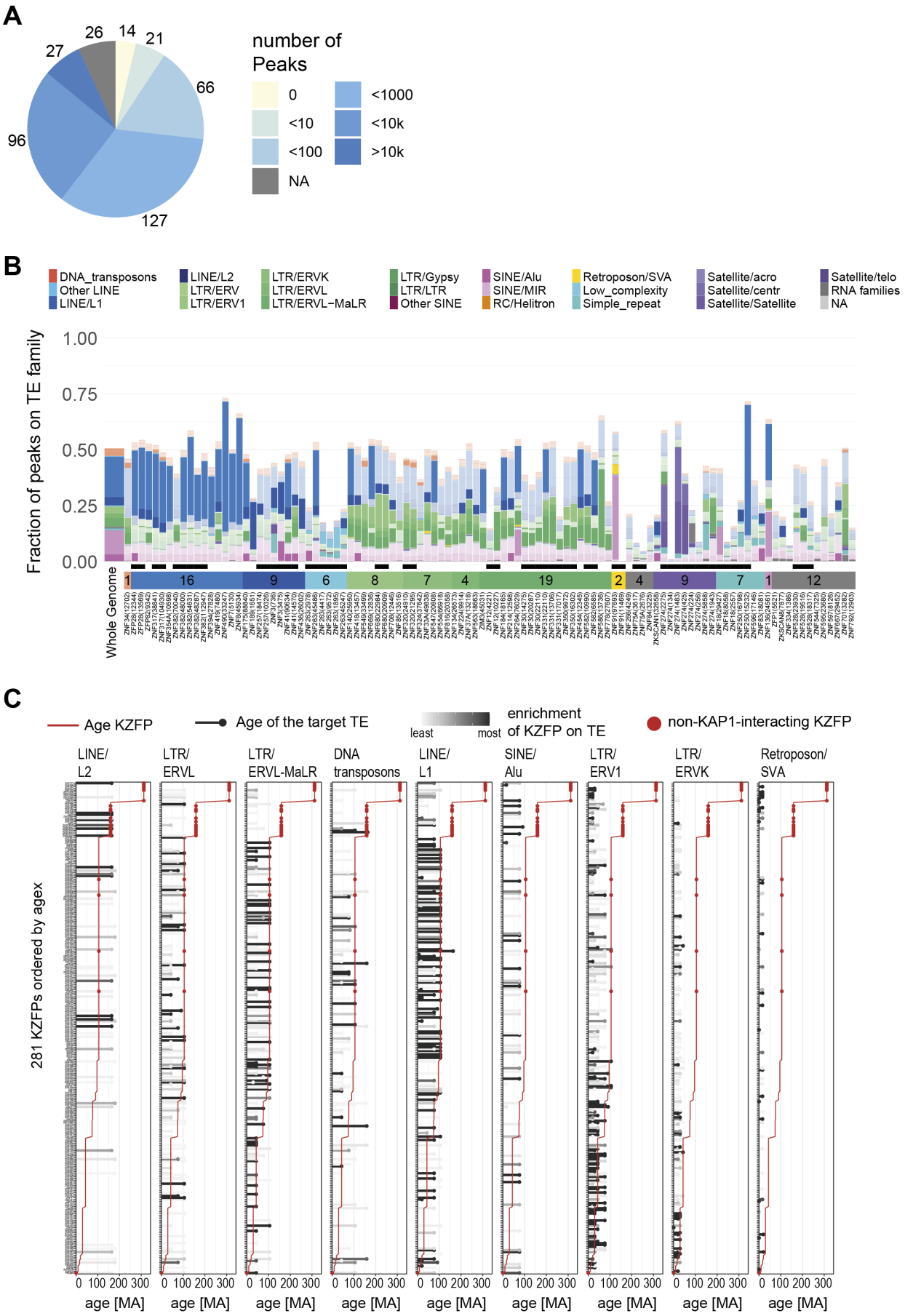

A
26
14
21
27
66
96
127
number of Peaks
0
<10
<100
<1000
<10k
>10k
NA
B
DNA_transposons
Other LINE
LINE/L1
LINE/L2
LTR/ERV
LTR/ERV1
LTR/ERVK
LTR/ERVL
LTR/ERVL−MaLR
LTR/Gypsy
LTR/LTR
Other SINE
SINE/Alu
SINE/MIR
RC/Helitron
Retroposon/SVA
Low_complexity
Simple_repeat
Satellite/acro
Satellite/centr
Satellite/Satellite
Satellite/telo
RNA families
NA
Fraction of peaks on TE family
1.00
0.75
0.50
0.25
0.00
Whole Genome
1
16
9
6
8
7
4
19
2
4
9
7
1
12
C
Age KZFP
Age of the target TE
least
most
enrichment of KZFP on TE
non-KAP1-interacting KZFP
LINE/ L2
LTR/ ERVL
LTR/ ERVL-MaLR
DNA transposons
LINE/ L1
SINE/ Alu
LTR/ ERV1
LTR/ ERVK
Retroposon/ SVA
281 KZFPs ordered by agex
0 100 200 300
age [MA]

Figure S2: Peaks, targets identified with external data and ages of KZFPs relative to the ages of their targets

A) Pie chart of the number of peaks per ChIP-seq or -exo for all 377 human KZFPs. In case of replicates the higher value is reported. B) Bar graph showing the fraction of peaks over repetitive element (RE) families for experiments conducted using different over-expression protocols (Table S2). Columns are ordered by the most enriched family which are indicated by the horizontal bar below, along with the number of KZFPs for each category. Replicate experiments are indicated by black squares above the horizontal bar. Significant enrichments (FDR < 0.05) are show in fully opaque colours where non-significant enrichments are transparent. The leftmost bar shows the genome occupancy of all RE families. The total number of peaks per experiment is indicated in brackets after the KZFP name below each bar. C) Detailed age comparison of KZFPs and their TE targets. KZFPs (rows) are ordered by age, shown as a red line. Their targets are split by family (excluding families targeted by less than 20 KZFPs) and shown as black or grey bars. The grey level of the TE targets shows the level of enrichment of the given KZFP for the subfamilies of that family with black showing the target with the highest -log10(FDR) linearly scaling to 0 (white). If the KZFP is enriched on several subfamilies of the same family the lowest FDR is shown. Red dots indicate KZFPs which are unlikely to interact with TRIM28 (Helleboid et al., 2019).

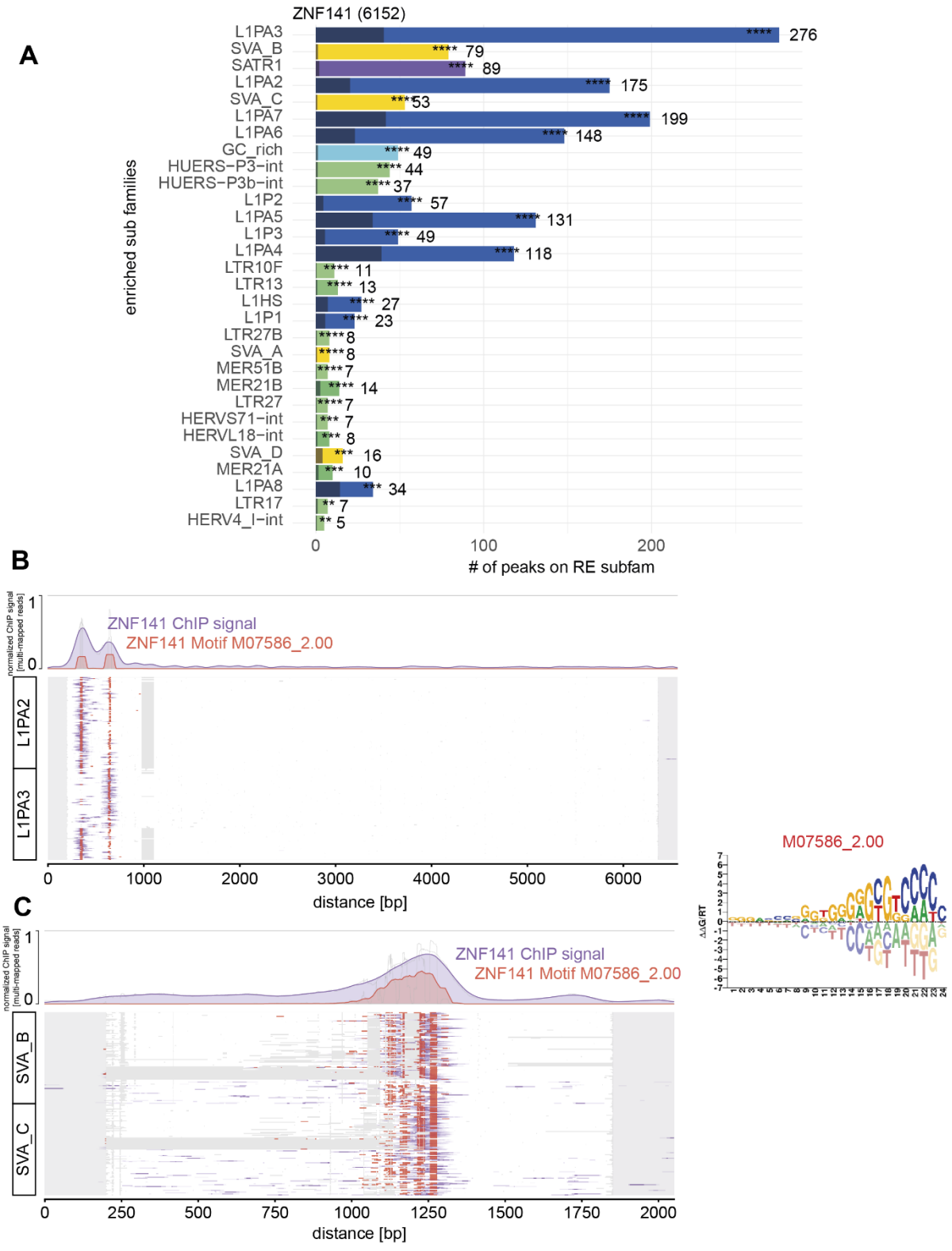

A
ZNF141 (6152)
enriched sub families
L1PA3 **** 276
SVA_B **** 79
SATR1 **** 89
L1PA2 **** 175
SVA_C *** 53
L1PA7 **** 199
L1PA6 **** 148
GC_rich **** 49
HUERS-P3-int **** 44
HUERS-P3b-int **** 37
L1P2 **** 57
L1PA5 **** 131
L1P3 **** 49
L1PA4 **** 118
LTR10F **** 11
LTR13 **** 13
L1HS **** 27
L1P1 **** 23
LTR27B **** 8
SVA_A **** 8
MER51B **** 7
MER21B **** 14
LTR27 **** 7
HERVS71-int *** 7
HERVL18-int *** 8
SVA_D *** 16
MER21A *** 10
L1PA8 *** 34
LTR17 ** 7
HERV4_I-int ** 5
0
100
200
# of peaks on RE subfam
B
normalized ChIP signal [multi-mapped reads]
ZNF141 ChIP signal
ZNF141 Motif M07586_2.00
L1PA2
L1PA3
0
1000
2000
3000
4000
5000
6000
distance [bp]
M07586_2.00
ΔΔG/RT
C
normalized ChIP signal [multi-mapped reads]
ZNF141 ChIP signal
ZNF141 Motif M07586_2.00
SVA_B
SVA_C
0
250
500
750
1000
1250
1500
1750
2000
distance [bp]

Figure S3: ZNF141 is binding to SVA VNTR

A) Enrichment of ZNF141 peaks over different repetitive element subfamilies (FDR < 0.01). The width of the coloured bars represents the number of peaks per subfamily also shown as a number on the right of the bar. The black transparent bars represent the expected number of peaks following a random distribution. The FDR of the enrichment is shown with stars (FDR < 0.0001 = ****, < 0.001=***, < 0.01=**, < 0.05=*, >= 0.05 = n.s). The y-axis is ordered by FDR. The number next to the title indicates the total number of peaks for the experiment. B) and C) Multiple sequence alignment (MSA) over the most enriched targets L1PA2 and 3 in B) and SVA_B and C in C). Up to 200 elements for the indicated targets where aligned, selecting first elements overlapping with a peak and then the longest elements. The signal of the ZNF141 ChIP was laid over the alignment in purple. The locations of the motif identified in (Weirauch et al., 2014) which is depicted on the right, is shown in red. The average signal normalized for each element (row-wise) can be seen as a line plot above the MSA plots.

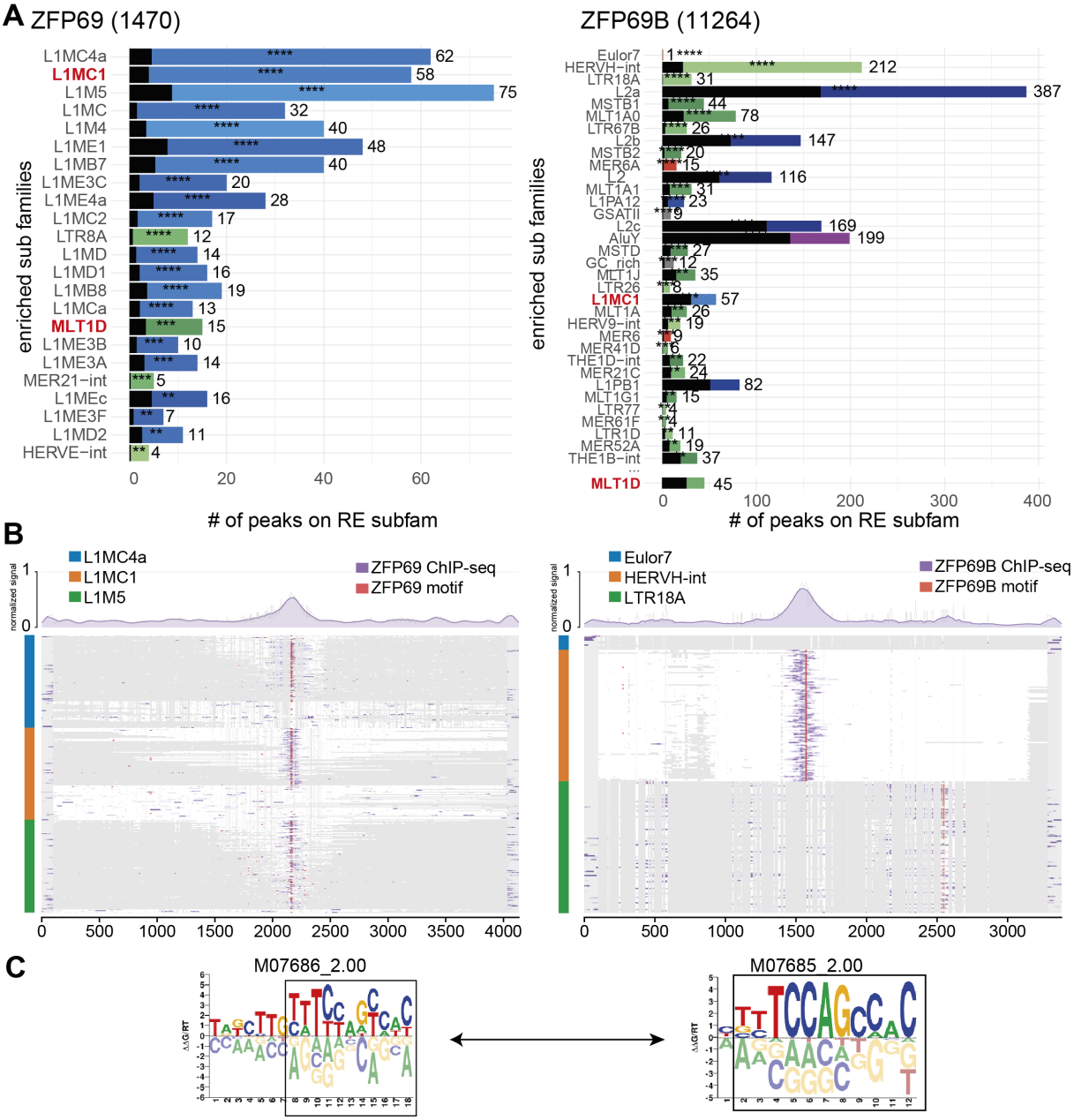

A
ZFP69 (1470)
ZFP69B (11264)
enriched sub families
# of peaks on RE subfam
B
ZFP69 ChIP-seq
ZFP69 motif
ZFP69B ChIP-seq
ZFP69B motif
normalized signal
C
M07686_2.00
M07685_2.00

Figure S4: ZFP69 and ZFP69B binding

A) Enrichment of peaks over different repetitive element subfamilies. Subfamilies with FDR < 0.01 are shown. The width of the coloured bars represents the number of peaks per subfamily also shown as a number on the right of the bar. The black transparent bars represent the expected number of peaks following a random distribution. The FDR of the enrichment is shown with stars (FDR < 0.0001 = ****, < 0.001=***, < 0.01=**, < 0.05=*, >= 0.05 = n.s). Rows are ordered by FDR. The number next to the title indicates the total number of peaks for the experiment. B) Multiple sequence alignment (MSA) over the 3 most enriched targets of ZFP69 (left) and ZFP95B (right). Up to 200 elements for the indicated targets (blue, orange and green) where aligned, selecting first elements overlapping with a peak and then the longest elements. The signal of ZFP69 and ZPF96B ChIPs was laid over their respective alignments in purple. The location of their motifs from panel C are shown in red. The normalized signal can be seen as a line plot above the MSA plot. C) Motifs from (Weirauch et al., 2014) for ZFP69 (left) and ZFP69B shown in panel B. Black rectangles indicate regions of high similarity.

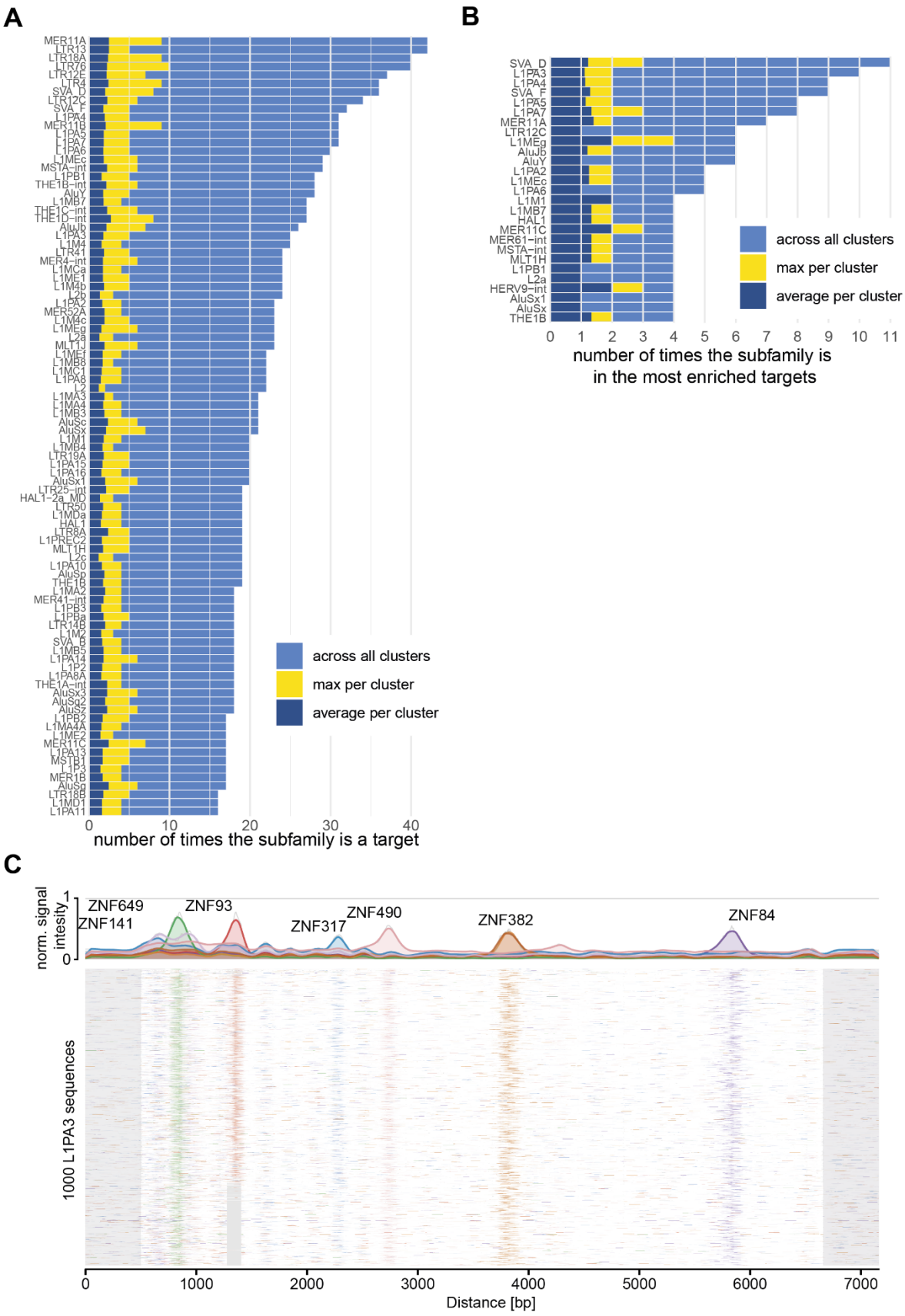

A
B
C
across all clusters
max per cluster
average per cluster
number of times the subfamily is a target
number of times the subfamily is in the most enriched targets
norm. signal intesity
ZNF649
ZNF141
ZNF93
ZNF317
ZNF490
ZNF382
ZNF84
1000 L1PA3 sequences
Distance [bp]

Figure S5: Localization of KZFPs targeting the same elements in clusters and on their target

A) and B) Bar graphs showing the number of KZFPs enriched on TE subfamilies (rows). The x-axis indicates the number of KZFPs enriched on the subfamily across all clusters (light blue), the highest number in any cluster (yellow) and the average number in the same cluster (dark blue). Clusters with no KZFPs binding the subfamily are not considered for the average. A) shows TE subfamilies with more than 15 enriched KZFPs with an FDR < 0.05. B) shows only the TE subfamilies per KZFPs within 10% of the highest -log10(FDR) for that KZFP. C) Multiple sequence alignment (MSA) for Figure 5B. 1000 of the longest L1PA3 elements were aligned. Alignments are shown in white, gaps in grey. Columns with more than 85% gaps were excluded to increase readability. The ChIP-seq and -exo signals for the indicated KZFPs are shown in colours and scaled for each row. The cumulative signal is shown as a line graph on top.

Tables

Table S1: Census of KZFPs

chr	chromosome
start	5' end of the identified KRAB domain
end	3' end of the last zinc finger domain in the zinc finger array
assigned gene	Gene name assigned to the KRAB-zinc-finger domain pair
strand	Defines the strand
classification	Classification of the assigned gene according to ensembl
age MA	Age of the assigned gene in million years
z_C2H2_miss	Normalized number of missense mutations in the C2H2 domains of the zinc fingers of the assigned gene. Values are standardized z-scores across all KZFPs with positive values indicating more polymorphic KZFPs and negative values more conserved KZFPs than the average
cluster	The KZFP cluster of the assigned gene

chr	start	end	assigned gene	strand	classification	age MA	z_C2H2_miss	cluster
chr1	23688470	23693655	ZNF436	-	protein coding	159	-0.910143028	noCluster
chr1	40922622	40929231	ZFP69B	+	protein coding	105	1.10855327	chr1.1
chr1	40954765	40961701	ZFP69	+	protein coding	105	-0.102664509	chr1.1
chr1	41006263	41013126	ZNF684	+	protein coding	105	-0.834441917	chr1.1
chr1	50307842	50312063	ZNF859P	+	unprocessed pseudogene	6.7	NA	noCluster
chr1	192962729	192963798	ZNF101P2	-	processed pseudogene	96	NA	noCluster
chr1	227834250	227843463	ZNF678	+	protein coding	29.4	1.858354753	noCluster
chr1	227884599	227894382	ZNF847P	-	unprocessed pseudogene	9.1	NA	noCluster
chr1	247150274	247163370	ZNF695	-	protein coding	43.2	0.301074751	chr1.2
chr1	247200762	247202833	ZNF670	-	protein coding	43.2	1.436591419	chr1.2
chr1	247263750	247265409	ZNF669	-	protein coding	105	0.90668364	chr1.2
chr1	247319903	247323109	ZNF124	-	protein coding	43.2	0.301074751	chr1.2
chr1	247353204	247363489	AL390728.4	-	transcribed unprocessed pseudogene	NA	NA	chr1.2
chr2	95814407	95818998	ZNF514	-	protein coding	105	1.209488085	noCluster
chr2	95843233	95847830	ZNF2	+	protein coding	159	-1.515751918	noCluster
chr2	132844867	132862838	AC098826.1	+	unprocessed pseudogene	NA	NA	noCluster
chr2	133070447	133076219	ZNF806	+	unprocessed pseudogene	105	NA	noCluster
chr2	203925656	203927117	AC023271.1	-	processed pseudogene	29.4	NA	noCluster
chr3	31836003	31837575	ZNF587P1	+	processed pseudogene	6.7	NA	noCluster
chr3	32030640	32032503	ZNF860	+	protein coding	29.4	-1.020253735	noCluster
chr3	40523386	40529435	ZNF619	+	protein coding	105	0.119392084	chr3.1
chr3	40552966	40558327	ZNF620	+	protein coding	NA	0.820168085	chr3.1
chr3	40570815	40574284	ZNF621	+	protein coding	105	0.041528084	chr3.1
chr3	42949519	42956795	ZNF662	+	protein coding	105	0.982384752	noCluster
chr3	44488078	44492059	ZNF445	-	protein coding	105	-1.385978584	chr3.2
chr3	44540715	44544154	ZNF852	-	protein coding	159	1.058085863	chr3.2
chr3	44609804	44612825	ZKSCAN7	+	protein coding	159	0.161318853	chr3.2
chr3	44673970	44685562	ZNF197	+	protein coding	105	-0.607338583	chr3.2
chr3	48302308	48311167	ZNF589	+	protein coding	NA	0.301074751	noCluster
chr3	75786097	75790881	ZNF717	-	protein coding	NA	NA	noCluster
chr3	101076124	101077332	ZNF90P1	+	processed pseudogene	9.1	NA	noCluster
chr4	59328	87317	ZNF595	+	protein coding	43.2	NA	chr4.1
chr4	59328	155306	ZNF718	+	protein coding	NA	NA	chr4.1
chr4	212446	248619	ZNF876P	+	transcribed unprocessed pseudogene	29.4	NA	chr4.1
chr4	262778	289938	ZNF732	-	protein coding	29.4	2.521640679	chr4.1
chr4	337576	369310	ZNF141	+	protein coding	43.2	-0.062290583	chr4.1
chr4	435425	466484	ZNF721	-	protein coding	NA	0.775029534	chr4.1
chr4	9385643	9388995	AC116655.1	-	unprocessed pseudogene	NA	NA	noCluster
chr4	26113507	26115431	AC097714.1	+	processed pseudogene	NA	NA	noCluster
chr4	103382447	103383409	AF213884.2	+	processed pseudogene	NA	NA	noCluster

chr	start	end	assigned gene	strand	classification	age MA	z_C2H2_miss	cluster
chr4	111337415	111339660	ZNF969P	+	processed pseudogene	6.7	NA	noCluster
chr5	23509587	23527843	PRDM9	+	protein coding	0	-0.397704737	noCluster
chr5	105878784	105880022	AC114940.1	-	processed pseudogene	NA	NA	noCluster
chr5	150275003	150278110	ZNF300	-	protein coding	105	-0.455936361	chr5.1
chr5	150310342	150322232	ZNF300P1	-	transcribed unprocessed pseudogene	105	NA	chr5.1
chr5	150329773	150332016	AC022106.1	-	unprocessed pseudogene	NA	NA	chr5.1
chr5	163282299	163283610	AC008432.1	-	processed pseudogene	NA	NA	noCluster
chr5	178139078	178154120	ZNF354A	-	protein coding	105	-0.81697243	chr5.2
chr5	178293250	178311253	ZNF354B	+	protein coding	105	-0.956728327	chr5.2
chr5	178339596	178359694	ZFP2	+	protein coding	105	NA	chr5.2
chr5	178373365	178392959	ZNF454	+	protein coding	105	-0.304534139	chr5.2
chr5	178454479	178460635	ZNF879	+	protein coding	105	-1.096484225	chr5.2
chr5	178503451	178506987	ZNF354C	+	protein coding	105	-0.029257371	chr5.2
chr6	27326096	27331075	ZNF204P	-	processed pseudogene	6.7	NA	noCluster
chr6	27419093	27425182	ZNF184	-	protein coding	105	-1.037639636	noCluster
chr6	28120044	28121762	ZKSCAN8	+	protein coding	159	-0.708273398	chr6.1
chr6	28212920	28214864	ZKSCAN4	-	protein coding	159	-0.737111917	chr6.1
chr6	28331474	28334035	ZKSCAN3	+	protein coding	159	-0.47756525	chr6.1
chr6	28962876	28967384	ZNF311	-	protein coding	105	-0.347791916	noCluster
chr6	29640504	29643830	ZFP57	-	protein coding	105	-1.212947473	noCluster
chr7	5096931	5105190	RBAK	+	protein coding	105	-0.47756525	noCluster
chr7	5160893	5167368	ZNF890P	-	transcribed unprocessed pseudogene	0	NA	noCluster
chr7	5879585	5887363	ZNF815P	+	transcribed unprocessed pseudogene	0	NA	noCluster
chr7	6683458	6694038	ZNF316	+	protein coding	159	NA	noCluster
chr7	6730523	6737436	ZNF12	-	protein coding	105	-1.39463014	noCluster
chr7	55990860	56007627	ZNF713	+	protein coding	312	-0.062290583	noCluster
chr7	56706637	56718384	AC095038.1	-	unprocessed pseudogene	NA	NA	noCluster
chr7	57187415	57194419	ZNF479	-	protein coding	20.2	-0.029257371	noCluster
chr7	57522171	57530230	ZNF716	+	protein coding	43.2	1.51229253	noCluster
chr7	62751759	62758764	ZNF733P	-	transcribed unprocessed pseudogene	20.2	NA	noCluster
chr7	62910202	62917202	ZNF734P	+	unprocessed pseudogene	20.2	NA	noCluster
chr7	63465995	63476268	AC115220.1	+	protein coding	NA	NA	chr7.1
chr7	63529274	63539681	ZNF727	+	protein coding	20.2	0.664440085	chr7.1
chr7	63673475	63680801	ZNF735	+	protein coding	NA	NA	chr7.1
chr7	63720604	63727826	ZNF679	+	protein coding	43.2	0.820168085	chr7.1
chr7	63796642	63809579	ZNF736	+	protein coding	43.2	0.099205121	chr7.1
chr7	63981204	64004804	ZNF680	-	protein coding	43.2	0.603879196	chr7.1
chr7	64151631	64169820	ZNF107	+	protein coding	43.2	0.528178084	chr7.1
chr7	64275301	64293852	ZNF138	+	protein coding	9.1	3.026314754	chr7.1
chr7	64377964	64389482	ZNF273	+	protein coding	43.2	0.482757418	chr7.1
chr7	64437209	64442488	ZNF117	-	protein coding	29.4	0.846122751	chr7.1
chr7	64852820	64865365	ZNF92	+	protein coding	43.2	-0.359589492	noCluster
chr7	99077289	99085090	ZNF789	+	protein coding	159	-0.996658584	chr7.2
chr7	99091196	99096459	ZNF394	-	protein coding	105	-0.996658584	chr7.2
chr7	99117815	99124019	ZKSCAN5	+	protein coding	159	-0.758740806	chr7.2
chr7	99159996	99171351	ZNF655	+	protein coding	96	NA	chr7.2
chr7	99627877	99631745	ZKSCAN1	+	protein coding	159	-1.515751918	noCluster
chr7	99668852	99672878	ZNF3	-	protein coding	159	-1.515751918	noCluster
chr7	148767601	148777803	ZNF786	-	protein coding	159	-0.088245249	chr7.3
chr7	148800730	148815434	ZNF425	-	protein coding	96	-0.463904899	chr7.3
chr7	148863255	148876734	ZNF398	+	protein coding	312.1	-1.061545251	chr7.3
chr7	148903793	148921679	ZNF282	+	protein coding	312.1	-1.152386584	chr7.3
chr7	148947777	148951449	ZNF212	+	protein coding	312.1	-0.910143028	chr7.3
chr7	148963915	148979341	ZNF783	+	protein coding	312.1	-1.061545251	chr7.3
chr7	149128878	149151322	ZNF777	-	protein coding	312.1	-1.515751918	chr7.3
chr7	149171645	149191200	ZNF746	-	protein coding	159	-1.515751918	chr7.3
chr8	192892	196338	ZNF596	+	protein coding	96	1.622403237	noCluster
chr8	7215398	7218752	ZNF705G	-	protein coding	312	6.659968091	noCluster
chr8	7806656	7810012	ZNF705B	+	protein coding	312	NA	noCluster
chr8	11967414	11970766	ZNF705D	+	protein coding	312	NA	noCluster
chr8	12213654	12217003	ZNF705C	+	NA	312	NA	noCluster
chr8	144773248	144776682	ZNF707	+	protein coding	105	-0.996658584	noCluster
chr8	145930955	145933674	AF186192.2	+	transcribed unprocessed pseudogene	NA	NA	chr8.1
chr8	145947196	145979697	ZNF251	-	protein coding	105	-0.607338583	chr8.1
chr8	145998674	146003543	ZNF34	-	protein coding	105	-1.212947473	chr8.1
chr8	146029031	146033723	ZNF517	+	protein coding	105	0.301074751	chr8.1
chr8	146054868	146068544	ZNF7	+	protein coding	105	-1.126431917	chr8.1
chr8	146106911	146115438	ZNF250	-	protein coding	105	-1.37599602	chr8.1
chr8	146201678	146225109	ZNF252P	-	transcribed unprocessed pseudogene	159	NA	chr8.1
chr9	35147592	35149172	AL353795.1	-	processed pseudogene	NA	NA	noCluster
chr9	39444541	39457000	ZNF658B-1	-	NA	NA	NA	noCluster
chr9	40772139	40784597	ZNF658	-	protein coding	105	NA	noCluster
chr9	41589560	41602018	ZNF658B-2	-	NA	NA	NA	noCluster
chr9	69830296	69848755	AL359955.1	+	unprocessed pseudogene	NA	NA	noCluster
chr9	95608674	95618594	ZNF484	-	protein coding	105	-0.218018583	noCluster
chr9	97054628	97063625	ZNF169	+	protein coding	96	-0.524755553	noCluster

chr	start	end	assigned gene	strand	classification	age MA	z_C2H2_miss	cluster
chr9	99491392	99500863	AL589843.1	+	lncRNA	NA	NA	chr9.1
chr9	99521161	99525908	ZNF510	-	protein coding	105	0.301074751	chr9.1
chr9	99580216	99607292	ZNF782	-	protein coding	105	-0.194423431	chr9.1
chr9	104162171	104171913	ZNF189	+	protein coding	105	NA	noCluster
chr9	114296024	114305438	ZNF483	+	protein coding	105	-1.350585857	noCluster
chr9	115805028	115812191	ZFP37	-	protein coding	105	0.301074751	noCluster
chr10	28627804	28629862	ZNF101P1	-	processed pseudogene	0	NA	noCluster
chr10	38120572	38127033	ZNF248	-	protein coding	105	-0.996658584	chr10.1
chr10	38172546	38185734	ZNF37CP-ZNF33CP	+	NA	NA	NA	chr10.1
chr10	38241081	38246468	ZNF25	-	protein coding	105	0.755281418	chr10.1
chr10	38305822	38345365	ZNF33A	+	protein coding	105	-0.039580249	chr10.1
chr10	38403688	38408217	ZNF37A	+	protein coding	96	-0.425655917	chr10.1
chr10	42830552	42850881	LOC441666	-	NA	NA	NA	chr10.2
chr10	43014691	43019180	ZNF37BP	-	transcribed_processed pseudogene	105	NA	chr10.2
chr10	43088084	43127863	ZNF33B	-	protein coding	105	-0.493786917	chr10.2
chr10	43971174	43978429	ZNF487	+	protein coding	105	NA	chr10.3
chr10	43991469	44020636	AL450326.2	+	unprocessed pseudogene	NA	NA	chr10.3
chr10	44052168	44063463	ZNF239	-	protein coding	105	-0.910143028	chr10.3
chr10	44104067	44112787	ZNF485	+	protein coding	105	-0.689921614	chr10.3
chr10	82057538	82059247	ZNF519P1	+	processed pseudogene	NA	NA	noCluster
chr10	132080719	132084868	AL157712.1	-	unprocessed pseudogene	NA	NA	noCluster
chr11	3380353	3392927	ZNF195	-	protein coding	43.2	-0.102664509	noCluster
chr11	6964319	6977744	ZNF215	+	protein coding	312	0.301074751	noCluster
chr11	7021185	7024060	ZNF214	-	protein coding	105	2.117901419	noCluster
chr11	23810843	23812195	AC100768.2	+	lncRNA	NA	NA	noCluster
chr11	40466820	40468691	AC090720.1	-	transcribed_processed pseudogene	NA	NA	noCluster
chr11	71527458	71530811	ZNF705E	-	protein coding	312	NA	noCluster
chr11	78095955	78096749	not_annotated_1	+	NA	NA	NA	noCluster
chr11	98435878	98437493	AP003038.1	-	processed pseudogene	NA	NA	noCluster
chr11	123596722	123598964	ZNF202	-	protein coding	159	-0.834441917	noCluster
chr11	129165290	129166338	ZNF123P	-	processed pseudogene	0	NA	noCluster
chr12	8326927	8330278	ZNF705A	+	protein coding	312	-0.910143028	noCluster
chr12	44402458	44403244	ZNF75BP	-	processed pseudogene	NA	NA	noCluster
chr12	47782932	47784031	LINC02156	+	lncRNA	NA	NA	noCluster
chr12	48736806	48739251	ZNF641	-	protein coding	105	-0.78902125	noCluster
chr12	58360849	58362098	AC084033.1	+	processed pseudogene	NA	NA	noCluster
chr12	133501967	133509741	ZNF605	-	protein coding	105	NA	chr12.1
chr12	133583633	133588061	ZNF26	+	protein coding	105	NA	chr12.1
chr12	133624546	133635500	ZNF84	+	protein coding	105	NA	chr12.1
chr12	133659694	133683168	ZNF140	+	protein coding	105	-0.607338583	chr12.1
chr12	133696890	133698378	ZNF891	-	protein coding	105	NA	chr12.1
chr12	133727619	133733533	ZNF10	+	protein coding	105	-1.152386584	chr12.1
chr12	133768080	133781098	ZNF268	+	protein coding	105	0.143089823	chr12.1
chr13	19041311	19059578	ZNF962P	-	unprocessed pseudogene	20.2	NA	noCluster
chr14	19170512	19188731	ZNF72P	-	unprocessed pseudogene	20.2	NA	noCluster
chr16	3166433	3170239	ZNF205	+	protein coding	159	-1.061545251	chr16.1
chr16	3188981	3191321	ZNF213	+	protein coding	159	-1.152386584	chr16.1
chr16	3336028	3340549	ZNF263	+	protein coding	159	-0.910143028	chr16.1
chr16	3363054	3367863	ZNF75A	+	protein coding	NA	-0.062290583	chr16.1
chr16	3486441	3490927	ZNF597	-	protein coding	105	-0.218018583	chr16.1
chr16	4802442	4810583	ZNF500	-	protein coding	159	-1.515751918	noCluster
chr16	25251229	25263347	ZKSCAN2	-	protein coding	159	1.209488085	noCluster
chr16	30543951	30545892	ZNF747	-	protein coding	159	NA	chr16.2
chr16	30566646	30569428	ZNF764	-	protein coding	159	0.603879196	chr16.2
chr16	30581368	30582875	ZNF688	-	protein coding	159	0.301074751	chr16.2
chr16	30594012	30596848	ZNF785	-	protein coding	96	0.301074751	chr16.2
chr16	30615653	30621278	ZNF689	-	protein coding	105	-1.152386584	chr16.2
chr16	31733952	31767101	ZNF720	+	protein coding	43.2	NA	noCluster
chr16	31895824	31928109	ZNF267	+	protein coding	43.2	-0.218018583	noCluster
chr16	68591906	68598583	ZFP90	+	protein coding	105	-1.37599602	noCluster
chr16	71482013	71488065	ZNF23	-	protein coding	105	0.073971417	noCluster
chr16	71509228	71512902	ZNF19	-	protein coding	NA	-0.304534139	noCluster
chr16	75200673	75204214	ZFP1	+	protein coding	105	-0.834441917	noCluster
chr16	89289570	89295030	ZNF778	+	protein coding	74	1.339261419	noCluster
chr16	90124281	90141873	PRDM7	-	protein coding	352	NA	noCluster
chr17	11881297	11887535	ZNF18	-	protein coding	312	-0.062290583	noCluster
chr17	15609706	15620586	ZNF286A	+	protein coding	159	NA	noCluster
chr17	16455193	16467133	ZNF287	-	protein coding	105	-1.256205251	noCluster
chr17	16525625	16538064	ZNF624	-	protein coding	105	-0.698179917	noCluster
chr17	52030317	52032682	AC023934.1	+	processed pseudogene	NA	NA	noCluster
chr18	9785120	9787745	ZNF415P1	+	processed pseudogene	NA	NA	noCluster
chr18	14104843	14124469	ZNF519	-	protein coding	96	0.704814011	noCluster
chr18	15254423	15273147	AP005901.4	+	unprocessed pseudogene	NA	NA	noCluster
chr19	2827620	2834775	ZNF554	+	protein coding	105	-0.218018583	chr19.1
chr19	2850590	2853821	ZNF555	+	protein coding	96	-0.667899472	chr19.1
chr19	2873499	2878297	ZNF556	+	protein coding	96	1.71416216	chr19.1

chr	start	end	assigned gene	strand	classification	age MA	z_C2H2_miss	cluster
chr19	2915525	2918197	ZNF57	+	protein coding	29.4	0.580586546	chr19.1
chr19	2933498	2939399	ZNF77	-	protein coding	76	1.360890308	chr19.1
chr19	7076397	7083740	ZNF557	+	protein coding	105	0.301074751	noCluster
chr19	8921971	8931976	ZNF558	-	protein coding	105	-0.910143028	noCluster
chr19	9267949	9272061	ZNF317	+	protein coding	105	-1.096484225	chr19.2
chr19	9406168	9413174	ZNF699	-	protein coding	105	0.171301417	chr19.2
chr19	9449874	9453971	ZNF559	+	protein coding	43.2	0.502944381	chr19.2
chr19	9489643	9492429	ZNF177	+	protein coding	105	-0.218018583	chr19.2
chr19	9522971	9529301	ZNF266	-	protein coding	90	1.663694752	chr19.2
chr19	9577170	9584995	ZNF560	-	protein coding	NA	-0.607338583	chr19.2
chr19	9638978	9644636	ZNF426	-	protein coding	90	0.466240812	chr19.2
chr19	9720923	9727841	ZNF561	-	protein coding	43.2	NA	chr19.2
chr19	9763485	9768805	ZNF562	-	protein coding	43.2	-0.607338583	chr19.2
chr19	9800861	9804490	ZNF812P	-	transcribed unprocessed pseudogene	NA	NA	chr19.2
chr19	9867398	9874078	ZNF846	-	protein coding	105	1.622403237	chr19.2
chr19	11725347	11728641	ZNF627	+	protein coding	74	-0.24397325	chr19.3
chr19	11759178	11763684	ZNF887P	+	transcribed unprocessed pseudogene	NA	NA	chr19.3
chr19	11793366	11797043	ZNF833P	+	transcribed unprocessed pseudogene	NA	NA	chr19.3
chr19	11832533	11836136	ZNF823	-	protein coding	76	-0.266683583	chr19.3
chr19	11867365	11870717	AC008543.2	-	unprocessed pseudogene	NA	NA	chr19.3
chr19	11888431	11893050	ZNF441	+	protein coding	43.2	-1.273508362	chr19.3
chr19	11915362	11918067	ZNF491	+	protein coding	15.8	NA	chr19.3
chr19	11941103	11944622	ZNF440	+	protein coding	29.4	1.209488085	chr19.3
chr19	11977025	11980114	ZNF439	+	protein coding	9.1	0.466240812	chr19.3
chr19	12014393	12017064	ZNF69	+	protein coding	29.4	NA	chr19.3
chr19	12058001	12061412	ZNF700	+	protein coding	NA	0.502944381	chr19.3
chr19	12087858	12090946	ZNF763	+	protein coding	6.7	1.391170752	chr19.3
chr19	12125677	12129101	ZNF433	-	protein coding	74	1.007618455	chr19.3
chr19	12154628	12157569	ZNF878	-	protein coding	76	-0.607338583	chr19.3
chr19	12184851	12188514	ZNF844	+	protein coding	29.4	1.51229253	chr19.3
chr19	12221125	12224519	ZNF788P	+	transcribed unprocessed pseudogene	NA	NA	noCluster
chr19	12243416	12246704	ZNF20	-	protein coding	NA	0.603879196	chr19.3
chr19	12256171	12258611	ZNF625	-	protein coding	74	-0.737111917	chr19.3
chr19	12296605	12299216	ZNF136	+	protein coding	74	0.021562956	chr19.3
chr19	12318152	12320224	AC012618.3	+	transcribed unprocessed pseudogene	NA	NA	chr19.3
chr19	12358100	12361177	ZNF44_alt	-	NA	NA	NA	chr19.3
chr19	12383239	12386891	ZNF44	-	protein coding	NA	0.179952973	chr19.3
chr19	12428721	12433519	ZNF563	-	protein coding	76	1.027805418	chr19.3
chr19	12460532	12463922	ZNF442	-	protein coding	76	-0.546777694	chr19.3
chr19	12490644	12494495	AC008758.3	-	transcribed unprocessed pseudogene	NA	NA	chr19.3
chr19	12501282	12504231	ZNF799	-	protein coding	76	NA	chr19.3
chr19	12540972	12544005	ZNF443	-	protein coding	76	0.982384752	chr19.3
chr19	12574122	12577658	ZNF709	-	protein coding	NA	-1.324507005	chr19.3
chr19	12637340	12639504	ZNF564	-	protein coding	76	-0.546777694	chr19.3
chr19	12691316	12694364	ZNF490	-	protein coding	43.2	-0.677216532	chr19.3
chr19	12734519	12740065	ZNF791	+	protein coding	90	-1.302007604	chr19.3
chr19	14805827	14830059	ZNF333	+	protein coding	NA	-0.062290583	noCluster
chr19	15932663	15934610	ZNF861P	-	unprocessed pseudogene	76	NA	noCluster
chr19	19788678	19791046	ZNF101	+	protein coding	43.2	-0.506403768	chr19.4
chr19	19822178	19825290	ZNF14	-	protein coding	43.2	0.301074751	chr19.4
chr19	19905274	19917871	ZNF506	-	protein coding	43.2	0.528178084	chr19.4
chr19	19944287	19946887	AC011477.1	+	transcribed unprocessed pseudogene	NA	NA	chr19.4
chr19	19989295	20003642	ZNF253	+	protein coding	29.4	0.301074751	chr19.4
chr19	20026094	20045681	ZNF93	+	protein coding	9.1	0.868833085	chr19.4
chr19	20116721	20135179	ZNF682	-	protein coding	43.2	0.846122751	chr19.4
chr19	20215053	20230956	ZNF90	+	protein coding	9.1	0.90668364	chr19.4
chr19	20295170	20308887	ZNF486	+	protein coding	29.4	2.319771049	chr19.4
chr19	20405903	20414488	AC078899.4	+	unprocessed pseudogene	NA	NA	chr19.4
chr19	20508157	20520759	ZNF826P_alt	-	NA	NA	NA	chr19.4
chr19	20577316	20592751	ZNF826P	-	transcribed unprocessed pseudogene	9.1	NA	chr19.4
chr19	20634067	20651342	AC008554.2	+	lncRNA	NA	NA	chr19.4
chr19	20726376	20736635	ZNF737	-	NA	29.4	1.139610136	chr19.4
chr19	20806838	20829205	ZNF626	-	protein coding	29.4	0.482757418	chr19.4
chr19	20975323	20990083	ZNF66	+	protein coding	29.4	1.209488085	chr19.4
chr19	21116835	21133084	ZNF85	+	protein coding	29.4	0.90668364	chr19.4
chr19	21216267	21241521	ZNF430	+	protein coding	43.2	0.13590869	chr19.4
chr19	21280996	21301448	ZNF714	+	protein coding	43.2	0.90668364	chr19.4
chr19	21349143	21366871	ZNF431	+	protein coding	43.2	-1.061545251	chr19.4
chr19	21404287	21428034	AC010620.2	-	unprocessed pseudogene	NA	NA	chr19.4
chr19	21475855	21493423	ZNF708	-	protein coding	43.2	1.598808086	chr19.4
chr19	21558045	21569928	ZNF738	+	protein coding	NA	NA	chr19.4
chr19	21587936	21609214	ZNF493	+	protein coding	29.4	1.51229253	chr19.4
chr19	21712465	21720784	ZNF429	+	protein coding	29.4	2.239023197	chr19.4
chr19	21820778	21839612	AC123912.3	+	unprocessed pseudogene	NA	NA	chr19.4
chr19	21908731	21927863	ZNF100	-	protein coding	43.2	0.796572933	chr19.4
chr19	21989721	22002017	ZNF43	-	protein coding	43.2	1.639789138	chr19.4

chr	start	end	assigned gene	strand	classification	age MA	z_C2H2_miss	cluster
chr19	22153002	22171705	ZNF208	-	protein coding	6.7	1.663694752	chr19.4
chr19	22255616	22272651	ZNF257	+	protein coding	29.4	0.796572933	chr19.4
chr19	22362097	22379522	ZNF676	-	protein coding	15.8	0.041528084	chr19.4
chr19	22486565	22501475	ZNF729	+	protein coding	29.4	0.942307693	chr19.4
chr19	22574416	22586308	ZNF98	-	protein coding	43.2	0.161318853	chr19.4
chr19	22646730	22655464	ZNF209P	+	unprocessed pseudogene	6.7	NA	chr19.4
chr19	22836114	22847968	ZNF492	+	protein coding	43.2	0.452476973	chr19.4
chr19	22937227	22952120	ZNF99	-	protein coding	29.4	0.820168085	chr19.4
chr19	23031068	23041202	ZNF723	+	protein coding	15.8	NA	chr19.4
chr19	23157793	23171247	ZNF728	-	protein coding	29.4	0.041528084	chr19.4
chr19	23316887	23329672	ZNF730	+	protein coding	29.4	0.301074751	chr19.4
chr19	23404795	23415089	ZNF724	-	protein coding	NA	NA	chr19.4
chr19	23494122	23525689	ZNF91_alt	-	NA	NA	NA	chr19.4
chr19	23541517	23557560	ZNF91	-	protein coding	NA	-0.20990775	chr19.4
chr19	23674387	23688777	ZNF725P	-	unprocessed pseudogene	6.7	NA	chr19.4
chr19	23836060	23845954	ZNF675	-	protein coding	29.4	0.796572933	chr19.4
chr19	23925883	23938347	ZNF681	-	protein coding	29.4	1.126905055	chr19.4
chr19	23990633	24016337	RPSAP58	+	processed pseudogene	NA	NA	chr19.4
chr19	24102180	24118949	ZNF726	+	protein coding	29.4	2.247674753	chr19.4
chr19	24288747	24311453	ZNF254	+	protein coding	43.2	1.838389624	chr19.4
chr19	35033456	35036291	ZNF807	-	unprocessed pseudogene	105	NA	chr19.5
chr19	35173681	35176451	ZNF302	+	protein coding	105	2.117901419	chr19.5
chr19	35230052	35232903	ZNF181	+	protein coding	105	-0.855087675	chr19.5
chr19	35249950	35260454	ZNF599	-	protein coding	105	0.041528084	chr19.5
chr19	35422776	35435724	ZNF30	+	protein coding	96	1.777246419	chr19.5
chr19	35449006	35451892	ZNF792	-	protein coding	96	-0.758740806	chr19.5
chr19	36673376	36699536	ZNF565	-	protein coding	NA	-1.061545251	chr19.6
chr19	36831137	36853134	ZFP14	-	protein coding	105	-1.236240122	chr19.6
chr19	36883657	36898909	ZFP82	-	protein coding	105	-0.153131916	chr19.6
chr19	36939968	36964354	ZNF566	-	protein coding	105	-0.737111917	chr19.6
chr19	37037875	37045692	ZNF529	-	protein coding	105	0.301074751	chr19.6
chr19	37100834	37118425	ZNF382	+	protein coding	105	-1.515751918	chr19.6
chr19	37129614	37149321	ZNF461	-	protein coding	105	-0.607338583	chr19.6
chr19	37203293	37211543	ZNF567	+	protein coding	105	-0.996658584	chr19.6
chr19	37238680	37253327	ZNF850	-	protein coding	105	-1.044722781	chr19.6
chr19	37309577	37314689	ZNF790	-	protein coding	105	0.755281418	chr19.6
chr19	37339364	37369160	ZNF345	+	protein coding	43.2	NA	chr19.6
chr19	37382402	37399355	ZNF829	-	protein coding	105	0.099205121	chr19.6
chr19	37427653	37441963	ZNF568	+	protein coding	NA	0.179952973	chr19.6
chr19	37482080	37488600	ZNF568_alt	+	NA	NA	NA	chr19.6
chr19	37581902	37619888	ZNF420	+	protein coding	105	-0.559527355	chr19.6
chr19	37642496	37647251	ZNF585A	-	protein coding	105	-0.879862584	chr19.6
chr19	37676134	37681046	ZNF585B	-	protein coding	105	-0.24397325	chr19.6
chr19	37726456	37734554	ZNF383	+	protein coding	105	-0.855087675	chr19.6
chr19	37838097	37854752	ZNF875	+	protein coding	NA	NA	noCluster
chr19	37870027	37880763	ZNF527	+	protein coding	105	-0.855087675	chr19.6
chr19	37903507	37917274	ZNF569	-	protein coding	105	-0.506403768	chr19.6
chr19	37966788	37976084	ZNF570	+	protein coding	105	-0.194423431	chr19.6
chr19	38023263	38028727	ZNF793	+	protein coding	105	0.603879196	chr19.6
chr19	38055505	38074992	ZNF571	-	protein coding	105	0.528178084	chr19.6
chr19	38090532	38104152	ZNF540	+	protein coding	90	0.194202594	chr19.6
chr19	38125893	38135631	ZFP30	-	protein coding	105	-0.304534139	chr19.6
chr19	38159296	38167290	ZNF781	-	protein coding	6.7	NA	chr19.6
chr19	38188946	38200717	ZNF607	-	protein coding	74	0.301074751	chr19.6
chr19	38229410	38262324	ZNF573	-	protein coding	96	1.066054401	chr19.6
chr19	40513186	40521658	ZNF546	+	protein coding	105	0.466240812	chr19.7
chr19	40539520	40554697	ZNF780B	-	protein coding	74	0.993199196	chr19.7
chr19	40579688	40589138	ZNF780A	-	protein coding	74	0.301074751	chr19.7
chr19	44341210	44353275	ZNF283	+	protein coding	105	-0.910143028	chr19.8
chr19	44376727	44384282	ZNF404	-	protein coding	105	0.560621418	chr19.8
chr19	44417601	44423845	ZNF45	-	protein coding	105	-0.218018583	chr19.8
chr19	44469107	44471569	ZNF221	+	protein coding	105	0.949941418	chr19.8
chr19	44495705	44501443	ZNF155	+	protein coding	105	1.457237176	chr19.8
chr19	44512947	44515517	ZNF230	+	protein coding	105	0.560621418	chr19.8
chr19	44531180	44537000	ZNF222	+	protein coding	105	0.301074751	chr19.8
chr19	44564613	44571331	ZNF223	+	protein coding	105	0.073971417	chr19.8
chr19	44585173	44591314	ZNF284	+	protein coding	105	0.301074751	chr19.8
chr19	44604959	44612502	ZNF224	+	protein coding	105	0.301074751	chr19.8
chr19	44622346	44636705	ZNF225	+	protein coding	105	0.514819065	chr19.8
chr19	44652929	44662191	ZNF234	+	protein coding	105	0.502944381	chr19.8
chr19	44676246	44681746	ZNF226	+	protein coding	105	0.942307693	chr19.8
chr19	44732604	44740965	ZNF227	+	protein coding	105	-0.767646819	chr19.8
chr19	44770356	44778739	ZNF233	+	protein coding	105	0.560621418	chr19.8
chr19	44791388	44803879	ZNF235	-	protein coding	105	-0.347791916	chr19.8
chr19	44831678	44844688	ZNF112	-	protein coding	105	-0.956728327	chr19.8
chr19	44890720	44896624	ZNF285	-	protein coding	105	1.572853419	chr19.8

chr	start	end	assigned gene	strand	classification	age MA	z_C2H2_miss	cluster
chr19	44932501	44936535	ZNF229	-	protein coding	105	0.087330437	chr19.8
chr19	44971878	44977358	ZNF285B	+	unprocessed pseudogene	105	NA	chr19.8
chr19	44980648	44983620	ZNF180	-	protein coding	105	NA	chr19.8
chr19	48785633	48790111	ZNF114	+	protein coding	96	-0.607338583	noCluster
chr19	50542423	50550292	ZNF473	+	protein coding	105	-0.506403768	noCluster
chr19	52084649	52091663	ZNF175	+	protein coding	105	0.161318853	noCluster
chr19	52376114	52381762	ZNF577	-	protein coding	105	1.079714752	chr19.9
chr19	52394032	52400225	ZNF649	-	protein coding	105	-0.607338583	chr19.9
chr19	52417931	52420950	AC011460.1	-	unprocessed pseudogene	NA	NA	chr19.9
chr19	52443467	52448741	ZNF613	+	protein coding	105	-0.689921614	chr19.9
chr19	52468433	52472378	ZNF350	-	protein coding	105	0.528178084	chr19.9
chr19	52496138	52505528	ZNF615	-	protein coding	105	0.301074751	chr19.9
chr19	52519167	52521741	ZNF614	-	protein coding	105	-0.78902125	chr19.9
chr19	52536990	52544853	ZNF432	-	protein coding	105	0.301074751	chr19.9
chr19	52568437	52580331	ZNF841	-	protein coding	43.2	-0.463904899	chr19.9
chr19	52616502	52627296	ZNF616	-	protein coding	43.2	0.301074751	chr19.9
chr19	52658199	52663838	ZNF836	-	protein coding	43.2	-0.077430805	chr19.9
chr19	52785369	52794588	ZNF766	+	protein coding	43.2	-0.607338583	chr19.9
chr19	52817411	52826184	ZNF480	+	protein coding	43.2	0.301074751	chr19.9
chr19	52856940	52870155	ZNF610	+	protein coding	43.2	1.890798086	chr19.9
chr19	52876369	52888627	ZNF880	+	protein coding	43.2	0.580586546	chr19.9
chr19	52909165	52919986	ZNF528	+	protein coding	43.2	1.996779642	chr19.9
chr19	52937213	52942681	ZNF534	+	protein coding	96	NA	chr19.9
chr19	53007913	53015998	ZNF578	+	protein coding	9.1	1.209488085	chr19.9
chr19	53050770	53059278	ZNF808	+	protein coding	29.4	0.383657781	chr19.9
chr19	53079153	53087273	ZNF701	+	protein coding	29.4	2.636994753	chr19.9
chr19	53095360	53100660	ZNF137P	+	transcribed unprocessed pseudogene	9.1	NA	chr19.9
chr19	53116215	53122309	ZNF83_alt	-	NA	NA	NA	chr19.9
chr19	53155743	53183669	ZNF83	-	protein coding	NA	-1.385978584	chr19.9
chr19	53207538	53217388	ZNF611	-	protein coding	29.4	0.161318853	chr19.9
chr19	53268440	53277948	ZNF600	-	protein coding	29.4	-1.288648584	chr19.9
chr19	53301707	53311380	ZNF28	-	protein coding	76	0.543318307	chr19.9
chr19	53342807	53352460	ZNF468	-	protein coding	76	0.502944381	chr19.9
chr19	53365113	53367406	AC010487.1	-	unprocessed pseudogene	NA	NA	chr19.9
chr19	53383908	53391500	ZNF320	-	protein coding	76	-0.855087675	chr19.9
chr19	53409022	53418569	ZNF888	-	protein coding	NA	NA	chr19.9
chr19	53452684	53456124	ZNF816	-	protein coding	NA	1.139610136	chr19.9
chr19	53571353	53578430	ZNF160	-	protein coding	90	-1.243227917	chr19.9
chr19	53611659	53619686	ZNF415	-	protein coding	90	0.961738994	chr19.9
chr19	53643012	53652614	ZNF347	-	protein coding	43.2	0.407946908	chr19.9
chr19	53667167	53678818	ZNF665	-	protein coding	90	1.10855327	chr19.9
chr19	53707274	53716979	ZNF818P	+	NA	NA	NA	chr19.9
chr19	53740368	53747144	ZNF677	-	protein coding	105	-0.062290583	chr19.9
chr19	53848764	53857489	ZNF845	+	protein coding	29.4	0.737113151	chr19.9
chr19	53879028	53887302	ZNF525	+	protein coding	20.2	1.339261419	chr19.9
chr19	53905323	53914243	ZNF765	+	protein coding	15.8	0.301074751	chr19.9
chr19	53952770	53960808	ZNF761	+	protein coding	29.4	NA	chr19.9
chr19	53989891	53996278	ZNF813	+	protein coding	29.4	-0.153131916	chr19.9
chr19	54074863	54081197	ZNF331	+	protein coding	105	-0.910143028	chr19.9
chr19	56884531	56889317	ZNF542P	+	transcribed unprocessed pseudogene	NA	NA	chr19.10
chr19	56895286	56901864	ZNF582	-	protein coding	105	-0.102664509	chr19.10
chr19	56925333	56935653	ZNF583	+	protein coding	105	-1.515751918	chr19.10
chr19	56952548	56972178	ZNF667	-	protein coding	105	-0.347791916	chr19.10
chr19	57027649	57037299	ZNF471	+	protein coding	105	0.949941418	chr19.10
chr19	57058882	57066575	ZFP28	+	protein coding	NA	-1.096484225	chr19.10
chr19	57085768	57089894	ZNF470	+	protein coding	105	0.528178084	chr19.10
chr19	57125185	57134119	ZNF71	+	protein coding	105	NA	chr19.10
chr19	57285940	57293441	ZIM2	-	protein coding	159	-0.062290583	chr19.10
chr19	57646294	57649960	ZIM3	-	protein coding	96	1.292071116	chr19.11
chr19	57705248	57724228	ZNF264	+	protein coding	105	-0.257948839	chr19.11
chr19	57755296	57765950	ZNF805	+	protein coding	105	-0.956728327	chr19.11
chr19	57795924	57803403	ZNF460	+	protein coding	105	-1.152386584	chr19.11
chr19	57835055	57840581	ZNF543	+	protein coding	105	0.999854239	chr19.11
chr19	57865098	57869148	ZNF304	+	protein coding	105	-0.834441917	chr19.11
chr19	57883155	57889544	ZNF547	+	protein coding	105	0.301074751	chr19.11
chr19	57908421	57911235	ZNF548	+	protein coding	96	-0.524755553	chr19.11
chr19	57929285	57932840	ZNF17	+	protein coding	43.2	-0.447030348	chr19.11
chr19	57953258	57956835	ZNF749	+	protein coding	43.2	0.580586546	chr19.11
chr19	57984659	57987148	ZNF772	-	protein coding	105	1.027805418	chr19.11
chr19	58002844	58005513	ZNF419	+	protein coding	105	1.952735359	chr19.11
chr19	58016033	58018847	ZNF773	+	protein coding	105	0.502944381	chr19.11
chr19	58046517	58050277	ZNF549	+	protein coding	96	0.161318853	chr19.11
chr19	58058176	58067719	ZNF550	-	protein coding	NA	0.041528084	chr19.11
chr19	58083555	58087292	ZNF416	-	protein coding	90	-1.020253735	chr19.11
chr19	58099912	58102718	ZIK1	+	protein coding	96	-1.515751918	chr19.11
chr19	58115653	58118684	ZNF530	+	protein coding	29.4	0.301074751	chr19.11

chr	start	end	assigned gene	strand	classification	age MA	z_C2H2_miss	cluster
chr19	58130123	58132753	ZNF134	+	protein coding	105	-0.425655917	chr19.11
chr19	58146002	58153531	ZNF211	+	protein coding	96	-0.689921614	chr19.11
chr19	58196632	58199638	ZNF551	+	protein coding	96	0.430848084	chr19.11
chr19	58213011	58216341	ZNF154	-	protein coding	105	0.846122751	chr19.11
chr19	58231704	58234702	ZNF671	-	protein coding	105	-0.304534139	chr19.11
chr19	58262158	58266700	ZNF776	+	protein coding	96	-1.112012658	chr19.11
chr19	58287916	58291140	ZNF586	+	protein coding	96	-0.607338583	chr19.11
chr19	58319425	58324782	ZNF552	-	protein coding	29.4	-0.062290583	chr19.11
chr19	58350396	58353926	ZNF587B	+	protein coding	9.1	0.301074751	chr19.11
chr19	58367480	58371572	ZNF587	+	protein coding	74	0.580586546	chr19.11
chr19	58383872	58388404	ZNF814	-	protein coding	43.2	0.631406872	chr19.11
chr19	58419853	58423548	ZNF417	-	protein coding	74	-0.001729694	chr19.11
chr19	58437541	58441916	ZNF418	-	protein coding	105	NA	chr19.11
chr19	58452300	58455422	ZNF256	-	protein coding	105	0.179952973	chr19.11
chr19	58489689	58500083	ZNF606	-	protein coding	105	-1.096484225	chr19.11
chr19	58572953	58579820	ZNF135	+	protein coding	105	-0.607338583	chr19.11
chr19	58697084	58724470	ZNF274	+	protein coding	NA	-1.515751918	chr19.11
chr19	58757672	58774428	ZNF544	+	protein coding	29.4	0.755281418	chr19.11
chr19	58797088	58806641	ZNF8	+	protein coding	105	-1.256205251	chr19.11
chr19	58921337	58929097	ZNF584	+	protein coding	105	-0.607338583	chr19.11
chr19	58944695	58948534	ZNF132	-	protein coding	105	-0.340158191	chr19.11
chr19	58965068	58967820	ZNF324B	+	protein coding	105	-0.910143028	chr19.11
chr19	58980552	58983368	ZNF324	+	protein coding	105	-1.515751918	chr19.11
chr19	58991006	58992075	ZNF446	+	protein coding	159	-0.304534139	chr19.11
chr20	2463809	2473465	ZNF343	-	protein coding	NA	-0.359589492	noCluster
chr20	18286330	18297388	ZNF133	+	protein coding	105	-0.86688525	noCluster
chr20	25655635	25666752	ZNF337	-	protein coding	43.2	0.301074751	noCluster
chr20	45113105	45121261	ZNF840P	+	unprocessed pseudogene	NA	NA	noCluster
chr20	45129943	45133373	ZNF334	-	protein coding	105	-0.218018583	noCluster
chr21	14467679	14486105	ZNF355P	-	unprocessed pseudogene	20.2	NA	noCluster
chr22	16634902	16653111	not_annotated_2	+	NA	NA	NA	noCluster
chr22	17336693	17354865	ZNF402P	+	processed pseudogene	NA	NA	noCluster
chr22	20754927	20761057	ZNF74	+	protein coding	105	-0.758740806	noCluster
chrX	46322188	46333009	KRBOX4	+	protein coding	76	NA	noCluster
chrX	46359298	46388338	ZNF674	-	protein coding	105	-0.524755553	noCluster
chrX	47269680	47272948	ZNF157	+	protein coding	105	-0.607338583	noCluster
chrX	47306861	47315791	ZNF41	-	protein coding	105	-1.195135447	noCluster
chrX	47747403	47776025	ZNF81	+	protein coding	105	-1.515751918	chrX.1
chrX	47835625	47842805	ZNF182	-	protein coding	NA	-1.515751918	chrX.1
chrX	47917801	47920318	ZNF630	-	protein coding	105	-0.689921614	chrX.1
chrX	63364943	63366923	AL355852.1	-	processed pseudogene	NA	NA	noCluster
chrX	66593423	66595251	AL049641.1	+	processed pseudogene	NA	NA	noCluster
chrX	114298540	114299145	AL121878.1	-	processed pseudogene	NA	NA	noCluster
chrX	123306937	123308270	ZIK1P1	-	processed pseudogene	105	NA	noCluster
chrX	134421104	134425055	ZNF75D	-	protein coding	312	-0.062290583	noCluster
chrX	152610134	152613427	ZNF275	+	protein coding	159	-1.515751918	noCluster
chrX	152684136	152686945	ZFP92	+	protein coding	105	-1.515751918	noCluster

Table S2: Summary of experiments

Name	KZFP gene name
Status	Experimental status of a gene: PubM= Data published, No transcript= No suitable transcript containing both KRAB and ZF domains, No overexpression= Low protein yield when overexpressing the protein in HEK293T cells, No DNA synthesis=Synthesis of cDNA failed
Peaks	Average number of peaks called for all replicates
Ensembl gene id	Ensemble gene ID of the KZFP
External synonym	Synonyms of the KZFP name
GSE	GEO accession number

Name	Status	Peaks	Ensembl gene id	External synonym	GSE
AC115220.1	PubM	6	NA	NA	GSE200964
KRBOX4	No transcript	NA	ENSG00000147121	FLJ20344; ZNF673	
PRDM7	No transcript	NA	ENSG00000126856	ZNF910	
PRDM9	PubM	588261	ENSG00000164256	KMT8B; MSBP3; PFM6; ZNF899	GSE78099
RBAK	PubM	610	ENSG00000146587	ZNF769	GSE78099
ZFP1	PubM	42	ENSG00000184517	FLJ34243; ZNF475	GSE104247; GSE78099
ZFP14	PubM	797	ENSG00000142065	KIAA1559; ZNF531	GSE78099
ZFP2	No transcript	NA	ENSG00000198939	FLJ21628; ZNF751	
ZFP28	PubM	858	ENSG00000196867	KIAA1431; mkr5	GSE76494; GSE120539
ZFP30	PubM	950	ENSG00000120784	KIAA0961; ZNF745	GSE200964
ZFP37	PubM	1	ENSG00000136866	ZNF906	GSE200964
ZFP57	PubM	16815	ENSG00000204644	bA145L22; bA145L22.2; C6orf40; ZNF698	GSE78099
ZFP69	PubM	1470	ENSG00000187815	FLJ16030; ZFP69A; ZKSCAN23A; ZNF642; ZSCAN54A	GSE78099
ZFP69B	PubM	11264	ENSG00000187801	FLJ34293; ZKSCAN23B; ZNF643; ZSCAN54B	GSE78099
ZFP82	PubM	105	ENSG00000181007	KIAA1948; MGC45380; ZNF545	GSE76494; GSE200964
ZFP90	PubM	680	ENSG00000184939	KIAA1954; NK10; ZNF756	GSE78099
ZFP92	PubM	46	ENSG00000189420	ZNF897	GSE200964
ZIK1	PubM	320	ENSG00000171649	ZNF762	GSE78099
ZIM2	PubM	10	ENSG00000269699	ZNF656	GSE78099
ZIM3	PubM	28384	ENSG00000141946	ZNF657	GSE76494; GSE78099
ZKSCAN1	PubM	184	ENSG00000106261	KOX18; PHZ-37; ZNF139; ZNF36; ZSCAN33	GSE31477; GSE200964
ZKSCAN2	PubM	2434	ENSG00000155592	FLJ23199; ZNF694; ZSCAN34	GSE78099
ZKSCAN3	PubM	131	ENSG00000189298	ZF47; Zfp47; ZNF306; ZNF309; ZSCAN35	GSE78099
ZKSCAN4	PubM	9	ENSG00000187626	FLJ32136; P1P373C6; p373c6.1; ZNF307; ZNF427; ZSCAN36	GSE120539
ZKSCAN5	PubM	6183	ENSG00000196652	ZFP95; ZNF914; ZSCAN37	GSE78099
ZKSCAN7	PubM	59	ENSG00000196345	FLJ12738; ZNF167; ZNF448; ZNF64; ZSCAN39	GSE120539
ZKSCAN8	PubM	4690	ENSG00000198315	LD5-1; ZNF192; ZSCAN40	GSE104247; GSE120539
ZNF10	PubM	3167	ENSG00000256223	KOX1	GSE78099
ZNF100	PubM	1797	ENSG00000197020		GSE78099
ZNF101	PubM	2041	ENSG00000181896	DKFZp570I0164; HZF12	GSE78099
ZNF107	No over-expression	NA	ENSG00000196247	smap-7; ZFD25; ZNF588	
ZNF112	PubM	1	ENSG00000062370	ZFP112; ZNF228	GSE200964
ZNF114	PubM	107	ENSG00000178150	MGC17986	GSE78099
ZNF117	No transcript	NA	ENSG00000152926	H-plk; HPF9	
ZNF12	PubM	727	ENSG00000164631	GIOT-3; KOX3; ZNF325	GSE51142; GSE104247; GSE78099
ZNF124	PubM	852	ENSG00000196418	HZF-16; HZF16	GSE78099
ZNF132	PubM	329	ENSG00000131849	pHZ-12	GSE78099
ZNF133	PubM	8607	ENSG00000125846	pHZ-13; pHZ-66; ZNF150	GSE78099
ZNF134	No transcript	NA	ENSG00000213762	pHZ-15	GSE76494
ZNF135	PubM	1508	ENSG00000176293	pHZ-17; ZNF61; ZNF78L1	GSE78099
ZNF136	PubM	4824	ENSG00000196646	pHZ-20	GSE76494; GSE200964

Name	Status	Peaks	Ensembl gene id	External synonym	GSE
ZNF138	PubM	8	ENSG00000197008	pHZ-32	GSE120539
ZNF14	PubM	111	ENSG00000105708	GIOT-4; KOX6	GSE200964
ZNF140	PubM	32	ENSG00000196387	pHZ-39	GSE76494; GSE200964
ZNF141	PubM	6152	ENSG00000131127	D4S90; pHZ-44	GSE78099
ZNF154	PubM	228	ENSG00000179909	pHZ-92	GSE78099
ZNF155	PubM	66	ENSG00000204920	pHZ-96	GSE200964
ZNF157	PubM	1318	ENSG00000147117	HZF22	GSE78099
ZNF160	PubM	6036	ENSG00000170949	F11; FLJ00032; HKr18; HZF5; KIAA1611; KR18	GSE120539
ZNF169	PubM	2106	ENSG00000175787	MGC51961	GSE78099
ZNF17	PubM	2336	ENSG00000186272	FLJ40864; FLJ46058; FLJ46615; HPF3; KIAA1947; KOX10	GSE78099
ZNF175	PubM	13	ENSG00000105497	OTK18	GSE76494; GSE200964
ZNF177	PubM	199	ENSG00000188629		GSE200964
ZNF18	PubM	317	ENSG00000154957	HDSG1; KOX11; Zfp535; ZKSCAN6; ZNF535; ZSCAN38	GSE76494; GSE97661; GSE78099
ZNF180	PubM	864	ENSG00000167384	HHZ168	GSE78099
ZNF181	PubM	55	ENSG00000197841	HHZ181; MGC44316	GSE78099
ZNF182	PubM	15177	ENSG00000147118	HHZ150; KOX14; Zfp182; ZNF21	GSE78099
ZNF184	PubM	249	ENSG00000096654		GSE76494; GSE78099
ZNF189	PubM	4615	ENSG00000136870		GSE104247; GSE78099
ZNF19	PubM	18325	ENSG00000157429	KOX12; MGC51021	GSE78099
ZNF195	PubM	510	ENSG00000005801		GSE120539
ZNF197	PubM	520	ENSG00000186448	D3S1363E; P18; ZKSCAN9; ZNF166; ZSCAN41	GSE78099
ZNF2	PubM	3234	ENSG00000275111	A1-5; Zfp661; ZNF661	GSE78099
ZNF20	PubM	41	ENSG00000132010	KOX13	GSE120539
ZNF202	PubM	12405	ENSG00000166261	ZKSCAN10; ZSCAN42	GSE78099
ZNF205	PubM	16801	ENSG00000122386	Zfp13; ZNF210	GSE78099
ZNF208	No over-expression	NA	ENSG00000160321	PMIDP; ZNF95	
ZNF211	PubM	562	ENSG00000121417	CH2H2-25; ZNF-25	GSE78099
ZNF212	PubM	192	ENSG00000170260	C2H2-150	GSE78099
ZNF213	PubM	NA	ENSG00000085644	CR53; ZKSCAN21; ZSCAN53	GSE120539
ZNF214	PubM	20	ENSG00000149050		GSE76494; GSE78099
ZNF215	PubM	270	ENSG00000149054	ZKSCAN11; ZSCAN43	GSE120539
ZNF221	PubM	NA	ENSG00000159905		GSE200964
ZNF222	PubM	1660	ENSG00000159885		GSE78099
ZNF223	PubM	157	ENSG00000178386		GSE78099
ZNF224	PubM	482	ENSG00000267680	BMZF-2; KOX22; ZNF255; ZNF27	GSE76494; GSE78099
ZNF225	PubM	202	ENSG00000256294		GSE78099
ZNF226	PubM	5	ENSG00000167380		GSE200964
ZNF227	PubM	69	ENSG00000131115		GSE200964
ZNF229	PubM	NA	ENSG00000278318		GSE200964
ZNF23	PubM	1093	ENSG00000167377	KOX16; Zfp612; ZNF359; ZNF612	GSE200964
ZNF230	PubM	1285	ENSG00000159882	FDZF2	GSE200964
ZNF233	PubM	1	ENSG00000159915	FLJ38032	GSE200964
ZNF234	PubM	25	ENSG00000263002	HZF4; ZNF269	GSE200964
ZNF235	PubM	18	ENSG00000159917	ANF270; HZF6; ZFP93; ZNF270	GSE78099
ZNF239	No transcript	NA	ENSG00000196793	HOK-2; MOK2	
ZNF248	PubM	9676	ENSG00000198105	bA162G10.3	GSE78099
ZNF25	PubM	59	ENSG00000175395	FLJ31890; KOX19; Zfp9	GSE78099
ZNF250	PubM	5	ENSG00000196150	MGC9718; ZFP647; ZNF647	GSE49402; GSE76494; GSE200964
ZNF251	PubM	21611	ENSG00000198169		GSE120539
ZNF253	PubM	1675	ENSG00000256771	BMZF-1; FLJ90391; ZNF411	GSE200964
ZNF254	PubM	696	ENSG00000213096	BMZF-5; HD-ZNF1; ZNF539; ZNF91L	GSE78099
ZNF256	PubM	148	ENSG00000152454	BMZF-3	GSE120539
ZNF257	PubM	38907	ENSG00000197134	BMZF-4	GSE76494; GSE78099
ZNF26	PubM	4804	ENSG00000198393	FLJ20755; KOX20	GSE78099
ZNF263	PubM	8816	ENSG00000006194	FPM315; ZKSCAN12; ZSCAN44	GSE51142; GSE76494; GSE19235; GSE31477; GSE78099
ZNF264	PubM	204	ENSG00000083844	KIAA0412	GSE76494; GSE78099
ZNF266	PubM	1416	ENSG00000174652	HZF1	GSE51142; GSE78099
ZNF267	PubM	1398	ENSG00000185947	HZF2	GSE78099
ZNF268	PubM	6	ENSG00000090612	HZF3	GSE200964
ZNF273	PubM	11075	ENSG00000198039	HZF9	GSE78099
ZNF274	PubM	581	ENSG00000171606	ZKSCAN19; ZSCAN51	GSE31477; GSE104247; GSE78099
ZNF275	No transcript	NA	ENSG00000063587		
ZNF28	PubM	12623	ENSG00000198538	DKFZp781D0275; KOX24	GSE78099
ZNF282	PubM	19316	ENSG00000170265	HUB1	GSE78099
ZNF283	PubM	45456	ENSG00000167637		GSE78099
ZNF284	PubM	19511	ENSG00000186026	DKFZp781F1775	GSE78099
ZNF285	PubM	85	ENSG00000267508	ZNF285A	GSE78099
ZNF286A	PubM	273	ENSG00000187607	KIAA1874; ZNF286	GSE200964
ZNF287	PubM	2323	ENSG00000141040	ZKSCAN13; ZSCAN45	GSE78099
ZNF3	PubM	76	ENSG00000166526	A8-51; FLJ20216; HF.12; KOX25; PP838; Zfp113	GSE51142; GSE104247; GSE78099
ZNF30	PubM	88	ENSG00000168661	DKFZp686N19164; FLJ20562; KOX28	GSE76494; GSE78099
ZNF300	PubM	8466	ENSG00000145908		GSE78099

Name	Status	Peaks	Ensembl gene id	External synonym	GSE
ZNF302	PubM	391	ENSG00000089335	ZNF135L; ZNF140L; ZNF327	GSE78099
ZNF304	PubM	1472	ENSG00000131845		GSE78099
ZNF311	PubM	64	ENSG00000197935		GSE78099
ZNF316	PubM	8	ENSG00000205903	ENST00000305834; MZF-3	GSE200964
ZNF317	PubM	6475	ENSG00000130803		GSE76494; GSE78099
ZNF320	PubM	3269	ENSG00000182986	DKFZp686G16228; ZFPL	GSE76494; GSE78099
ZNF324	PubM	304	ENSG00000083812	ZF5128; ZNF324A	GSE76494; GSE78099
ZNF324B	PubM	17	ENSG00000249471	FLJ45850	GSE78099
ZNF331	PubM	829	ENSG00000130844	RITA; ZNF361; ZNF463	GSE76494; GSE104247; GSE78099
ZNF333	PubM	1204	ENSG00000160961	KIAA1806	GSE78099
ZNF334	PubM	363645	ENSG00000198185	bA179N14.1	GSE104247; GSE78099
ZNF337	PubM	1604	ENSG00000130684	dJ694B14.1	GSE78099
ZNF33A	PubM	1165	ENSG00000189180	FLJ23404; KIAA0065; KOX31; KOX5; ZNF11A; ZNF33; ZZAPK	GSE76494; GSE78099
ZNF33B	PubM	134	ENSG00000196693	KOX2; KOX31; ZNF11B	GSE78099
ZNF34	PubM	5	ENSG00000196378	KOX32	GSE76494; GSE120539
ZNF343	PubM	3805	ENSG00000088876	MGC10715	GSE78099
ZNF345	No transcript	NA	ENSG00000251247	HZF10	
ZNF347	PubM	3751	ENSG00000197937	ZNF1111	GSE200964
ZNF350	No over-expression	NA	ENSG00000256683	ZBRK1; ZFQR	GSE76494
ZNF354A	PubM	1058	ENSG00000169131	EZNF; HKL1; KID-1; KID1; TCF17	GSE76494; GSE78099
ZNF354B	PubM	19	ENSG00000178338	FLJ25008; KID2	GSE78099
ZNF354C	PubM	820	ENSG00000177932	KID3	GSE200964
ZNF37A	PubM	53	ENSG00000075407	KOX21; ZNF37	GSE76494; GSE200964
ZNF382	PubM	12574	ENSG00000161298	FLJ14686; KS1	GSE76494; GSE78099
ZNF383	PubM	1641	ENSG00000188283	FLJ35863; Zfp383	GSE78099
ZNF394	PubM	575	ENSG00000160908	FLJ12298; ZKSCAN14; ZSCAN46	GSE76494; GSE200964
ZNF398	PubM	230	ENSG00000197024	KIAA1339; P51; P71; ZER6	GSE78099
ZNF404	PubM	10	ENSG00000176222		GSE200964
ZNF41	PubM	490	ENSG00000147124	MGC8941; MRX89	GSE76494; GSE200964
ZNF415	PubM	236	ENSG00000170954		GSE200964
ZNF416	PubM	NA	ENSG00000083817	FLJ20557	GSE120539
ZNF417	PubM	967	ENSG00000173480	MGC34079	GSE78099
ZNF418	PubM	341	ENSG00000196724	FLJ31551; KIAA1956	GSE76494; GSE78099
ZNF419	PubM	70	ENSG00000105136	ZAPHIR; ZNF419A	GSE76494; GSE78099
ZNF420	PubM	586	ENSG00000197050	FLJ32191	GSE200964
ZNF425	PubM	6565	ENSG00000204947		GSE78099
ZNF426	PubM	1	ENSG00000130818	MGC2663	GSE120539
ZNF429	PubM	6397	ENSG00000197013		GSE78099
ZNF43	PubM	1112	ENSG00000198521	HTF6; KOX27; ZNF39L1	GSE200964
ZNF430	PubM	1315	ENSG00000118620	FLJ13659	GSE78099
ZNF431	PubM	227	ENSG00000196705	KIAA1969	GSE78099
ZNF432	PubM	211	ENSG00000256087	KIAA0798	GSE78099
ZNF433	PubM	1065	ENSG00000197647	FLJ40981	GSE78099
ZNF436	PubM	88981	ENSG00000125945	KIAA1710; Zfp46	GSE76494; GSE200964
ZNF439	PubM	40	ENSG00000171291	DKFZp571K0837	GSE78099
ZNF44	PubM	2941	ENSG00000197857	KOX7; ZNF504; ZNF55; ZNF58	GSE78099
ZNF440	PubM	12845	ENSG00000171295	FLJ37933	GSE78099
ZNF441	PubM	21035	ENSG00000197044	FLJ38637	GSE78099
ZNF442	PubM	51	ENSG00000198342	FLJ14356	GSE78099
ZNF443	PubM	108	ENSG00000180855	ZK1	GSE78099
ZNF445	PubM	4178	ENSG00000185219	ZKSCAN15; ZNF168; ZSCAN47	GSE78099
ZNF446	PubM	3	ENSG00000083838	FLJ20626; ZKSCAN20; ZSCAN52	GSE120539
ZNF45	PubM	63	ENSG00000124459	ZNF13	GSE78099
ZNF454	PubM	268	ENSG00000178187	FLJ37444	GSE76494; GSE78099
ZNF460	PubM	7521	ENSG00000197714	HZF8; ZNF272	GSE78099
ZNF461	PubM	16478	ENSG00000197808	GIOT-1; MGC33911	GSE200964
ZNF468	PubM	7725	ENSG00000204604		GSE78099
ZNF470	PubM	147	ENSG00000197016	CZF-1; FLJ26175	GSE200964
ZNF471	PubM	3238	ENSG00000196263	KIAA1396; Z1971; Zfp78	GSE200964
ZNF473	PubM	3	ENSG00000142528	DKFZP434N043; HZFP100; KIAA1141	GSE200964
ZNF479	PubM	38813	ENSG00000185177	KR19	GSE78099
ZNF480	PubM	1134	ENSG00000198464	MGC32104	GSE78099
ZNF483	PubM	585	ENSG00000173258	KIAA1962; ZKSCAN16; ZSCAN48	GSE78099
ZNF484	PubM	392	ENSG00000127081	BA526D8.4; FLJ33884	GSE78099
ZNF485	PubM	8303	ENSG00000198298		GSE78099
ZNF486	PubM	47	ENSG00000256229	KRBO2; MGC2396	GSE78099
ZNF487	PubM	392	ENSG00000243660	KRBO1; ZNF487P	GSE78099
ZNF490	PubM	10001	ENSG00000188033	KIAA1198	GSE76494; GSE200964
ZNF491	No transcript	NA	ENSG00000177599	FLJ34791	
ZNF492	PubM	452	ENSG00000229676	KIAA1473; ZNF115	GSE78099
ZNF493	PubM	186	ENSG00000196268	FLJ36504	GSE200964
ZNF500	PubM	137	ENSG00000103199	KIAA0557; ZKSCAN18; ZSCAN50	GSE200964
ZNF506	PubM	9545	ENSG00000081665	DKFZp761G1812	GSE78099
ZNF510	PubM	819	ENSG00000081386	KIAA0972	GSE200964
ZNF514	PubM	68	ENSG00000144026	FLJ14457	GSE200964
ZNF517	PubM	NA	ENSG00000197363		GSE200964

Name	Status	Peaks	Ensembl gene id	External synonym	GSE
ZNF519	PubM	22300	ENSG00000175322	FLJ36809; HsT2362	GSE200964
ZNF525	PubM	684	ENSG00000203326	KIAA1979	GSE78099
ZNF527	PubM	98	ENSG00000189164	KIAA1829	GSE78099
ZNF528	PubM	1073	ENSG00000167555	KIAA1827	GSE76494; GSE78099
ZNF529	PubM	220	ENSG00000186020	KIAA1615	GSE78099
ZNF530	PubM	1037	ENSG00000183647	KIAA1508	GSE78099
ZNF534	PubM	42925	ENSG00000198633	FLJ25344; KRBO3	GSE78099
ZNF540	PubM	76	ENSG00000171817	DKFZp547B0714	GSE78099
ZNF543	PubM	1007	ENSG00000178229	DKFZp434H055	GSE78099
ZNF544	PubM	2	ENSG00000198131	AF020591	GSE104247; GSE200964
ZNF546	PubM	5	ENSG00000187187	MGC43537; ZNF49	GSE120539
ZNF547	PubM	2508	ENSG00000152433	FLJ31100	GSE76494; GSE78099
ZNF548	PubM	263	ENSG00000188785	FLJ32932	GSE78099
ZNF549	PubM	662	ENSG00000121406	FLJ34917	GSE76494; GSE78099
ZNF550	PubM	154	ENSG00000251369	MGC41917	GSE78099
ZNF551	PubM	7	ENSG00000204519	DKFZp686H1038	GSE200964
ZNF552	PubM	19	ENSG00000178935	FLJ21603	GSE78099
ZNF554	PubM	2234	ENSG00000172006	FLJ34817	GSE76494; GSE200964
ZNF555	PubM	1246	ENSG00000186300	MGC26707	GSE78099
ZNF556	PubM	100	ENSG00000172000	FLJ11637	GSE200964
ZNF557	PubM	882	ENSG00000130544	MGC4054	GSE78099
ZNF558	PubM	2581	ENSG00000167785	FLJ30932	GSE78099
ZNF559	PubM	589	ENSG00000188321	MGC13105	GSE200964
ZNF560	PubM	9601	ENSG00000198028	FLJ31986	GSE200964
ZNF561	PubM	598	ENSG00000171469	MGC45408	GSE78099
ZNF562	PubM	346	ENSG00000171466	FLJ20079	GSE78099
ZNF563	PubM	507	ENSG00000188868	FLJ34797	GSE76494; GSE200964
ZNF564	PubM	630	ENSG00000249709	MGC26914	GSE78099
ZNF565	PubM	7642	ENSG00000196357	FLJ36991	GSE78099
ZNF566	PubM	505	ENSG00000186017	FLJ14779; MGC12515	GSE78099
ZNF567	PubM	1248	ENSG00000189042	MGC45586	GSE78099
ZNF568	PubM	229	ENSG00000198453	DKFZp686B0797	GSE200964
ZNF569	PubM	54	ENSG00000196437	FLJ32053; ZAP1; Zfp74	GSE200964
ZNF57	PubM	93	ENSG00000171970	ZNF424	GSE200964
ZNF570	PubM	182	ENSG00000171827	FLJ30791	GSE78099
ZNF571	PubM	455	ENSG00000180479	HSPC059	GSE78099
ZNF573	PubM	14960	ENSG00000189144	FLJ30921	GSE78099
ZNF577	PubM	801	ENSG00000161551	MGC4400	GSE120539
ZNF578	PubM	512	ENSG00000258405	FLJ31384	GSE120539
ZNF582	PubM	264	ENSG00000018869	FLJ30927	GSE76494; GSE78099
ZNF583	PubM	493	ENSG00000198440	FLJ31030	GSE200964
ZNF584	PubM	290	ENSG00000171574	FLJ39899	GSE78099
ZNF585A	PubM	6468	ENSG00000196967	FLJ23765; Zfp27	GSE78099
ZNF585B	PubM	192	ENSG00000245680	FLJ14928; SZFP41; Zfp27	GSE200964
ZNF586	PubM	365	ENSG00000083828	FLJ20070	GSE76494; GSE200964
ZNF587	PubM	206	ENSG00000198466	FLJ14710; FLJ20813; UBF-fl; ZF6	GSE78099
ZNF587B	PubM	72	ENSG00000269343		GSE200964
ZNF589	PubM	NA	ENSG00000164048	SZF1	GSE200964
ZNF595	No DNA synthesis	NA	ENSG00000272602	FLJ31740	GSE76494
ZNF596	No over-expression	NA	ENSG00000172748		GSE76494
ZNF597	PubM	61	ENSG00000167981	FLJ33071; HIT-4	GSE97661; GSE120539
ZNF599	PubM	46	ENSG00000153896	FLJ30663	GSE200964
ZNF600	PubM	2324	ENSG00000189190	DKFZp686F06123; KR-ZNF1	GSE200964
ZNF605	PubM	10058	ENSG00000196458		GSE78099
ZNF606	PubM	91537	ENSG00000166704	FLJ14260; KIAA1852; ZNF328	GSE200964
ZNF607	PubM	600	ENSG00000198182	FLJ14802; MGC13071	GSE200964
ZNF610	PubM	2848	ENSG00000167554	FLJ36040	GSE78099
ZNF611	PubM	3790	ENSG00000213020	MGC5384	GSE78099
ZNF613	PubM	707	ENSG00000176024	FLJ13590	GSE78099
ZNF614	PubM	97	ENSG00000142556	FLJ21941	GSE78099
ZNF615	PubM	56	ENSG00000197619	FLJ33710	GSE78099
ZNF616	PubM	1173	ENSG00000204611	MGC45556	GSE78099
ZNF619	PubM	63	ENSG00000177873	FLJ90764	GSE78099
ZNF620	PubM	12	ENSG00000177842	MGC50836	GSE78099
ZNF621	PubM	125	ENSG00000172888	FLJ45246	GSE78099
ZNF624	PubM	239	ENSG00000197566	KIAA1349	GSE200964
ZNF625	PubM	3021	ENSG00000257591		GSE200964
ZNF626	PubM	241	ENSG00000188171		GSE78099
ZNF627	PubM	1202	ENSG00000198551	FLJ90365	GSE78099
ZNF630	PubM	209	ENSG00000221994	BC037316; dJ54B20.2; FLJ20573; MGC138344	GSE200964
ZNF641	PubM	180	ENSG00000167528	FLJ31295	GSE78099
ZNF649	PubM	4191	ENSG00000198093	FLJ12644	GSE78099
ZNF655	No transcript	NA	ENSG00000197343	VIK; VIK-1	
ZNF658	No over-expression	NA	ENSG00000274349	DKFZp572C163; FLJ32813; MGC35232	
ZNF66	PubM	126	ENSG00000160229	FLJ16537; ZNF66P	GSE200964
ZNF662	PubM	355	ENSG00000182983	FLJ45880	GSE78099

Name	Status	Peaks	Ensembl gene id	External synonym	GSE
ZNF665	PubM	129	ENSG00000197497	FLJ14345; ZFP160L	GSE200964
ZNF667	PubM	41	ENSG00000198046	FLJ14011	GSE76494; GSE78099
ZNF669	PubM	265	ENSG00000188295	FLJ12606	GSE76494; GSE78099
ZNF670	PubM	21	ENSG00000277462	MGC12466	GSE200964
ZNF671	PubM	3849	ENSG00000083814	FLJ23506	GSE78099
ZNF674	PubM	4618	ENSG00000251192	MRX92; ZNF673B	GSE78099
ZNF675	PubM	9280	ENSG00000197372	TBZF; TIZ	GSE78099
ZNF676	PubM	13935	ENSG00000196109		GSE200964
ZNF677	PubM	175684	ENSG00000197928	MGC48625	GSE200964
ZNF678	No transcript	NA	ENSG00000181450	MGC42493	
ZNF679	PubM	43428	ENSG00000197123	MGC42415	GSE200964
ZNF680	PubM	20156	ENSG00000173041	FLJ90430	GSE76494; GSE78099
ZNF681	PubM	476	ENSG00000196172	FLJ31526	GSE78099
ZNF682	PubM	2336	ENSG00000197124	BC39498_3	GSE78099
ZNF684	PubM	613	ENSG00000117010	MGC27466	GSE78099
ZNF688	PubM	NA	ENSG00000229809		GSE200964
ZNF689	PubM	16	ENSG00000156853	FLJ90415; TIPUH1	GSE200964
ZNF69	PubM	48	ENSG00000198429	Cos5	GSE78099
ZNF695	PubM	8721	ENSG00000197472	SBZF3	GSE78099
ZNF699	PubM	55	ENSG00000196110	FLJ38144; hang	GSE200964
ZNF7	PubM	334	ENSG00000147789	HF.16; KOX4	GSE104247; GSE78099
ZNF700	PubM	516	ENSG00000196757	DKFZp434I1610	GSE200964
ZNF701	PubM	584	ENSG00000167562	FLJ10891	GSE97661; GSE78099
ZNF705A	PubM	92245	ENSG00000196946	FLJ16353	GSE200964
ZNF705B	PubM	46544	ENSG00000215356		GSE200964
ZNF705D	PubM	38672	ENSG00000215343		GSE200964
ZNF705E	PubM	484	ENSG00000214534		GSE200964
ZNF705G	PubM	8	ENSG00000215372		GSE78099
ZNF707	PubM	2722	ENSG00000181135		GSE78099
ZNF708	PubM	1021	ENSG00000182141	KOX8; ZNF15; ZNF15L1	GSE76494; GSE78099
ZNF709	PubM	3	ENSG00000242852	FLJ38281	GSE200964
ZNF71	PubM	52	ENSG00000197951	Cos26; EZFIT	GSE200964
ZNF713	PubM	NA	ENSG00000178665	FLJ39963	GSE200964
ZNF714	PubM	786	ENSG00000160352		GSE78099
ZNF716	PubM	1351	ENSG00000182111	FLJ46189	GSE78099
ZNF717	PubM	2	ENSG00000227124	X17; ZNF838	GSE200964
ZNF718	PubM	44	ENSG00000250312	FLJ90036	GSE200964
ZNF720	No transcript	NA	ENSG00000197302		
ZNF721	PubM	5914	ENSG00000182903	KIAA1982	GSE200964
ZNF723	PubM	16994	ENSG00000268696	ZNF723P	GSE200964
ZNF724	PubM	426	ENSG00000196081	ZNF724P	GSE200964
ZNF726	PubM	145	ENSG00000213967	ZNF92P3	GSE200964
ZNF727	PubM	494	ENSG00000214652	ZNF727P	GSE200964
ZNF728	PubM	146	ENSG00000269067		GSE200964
ZNF729	PubM	1	ENSG00000196350		GSE120539
ZNF730	PubM	1164	ENSG00000183850		GSE78099
ZNF732	PubM	67	ENSG00000186777	FLJ59067	GSE200964
ZNF735	PubM	5127	ENSG00000223614	ZNF735P	GSE200964
ZNF736	PubM	9963	ENSG00000234444		GSE78099
ZNF738	No transcript	NA	ENSG00000172687		
ZNF74	PubM	96	ENSG00000185252	Cos52; Zfp520; ZNF520	GSE78099
ZNF746	PubM	316	ENSG00000181220	FLJ31413; PARIS	GSE120539
ZNF747	PubM	340	ENSG00000169955	MGC2474	GSE200964
ZNF749	PubM	210	ENSG00000186230	FLJ16360	GSE78099
ZNF75A	PubM	130	ENSG00000162086	FLJ31529	GSE97661; GSE200964
ZNF75D	PubM	10639	ENSG00000186376	D8C6; ZKSCAN24; ZNF75; ZNF82; ZSCAN28	GSE78099
ZNF761	PubM	402	ENSG00000160336	FLJ16231; FLJ35333; KIAA2033	GSE200964
ZNF763	PubM	4	ENSG00000197054	ZNF440L	GSE120539
ZNF764	PubM	32	ENSG00000169951	MGC13138	GSE78099
ZNF765	PubM	4148	ENSG00000196417		GSE78099
ZNF766	PubM	17616	ENSG00000196214		GSE78099
ZNF77	PubM	301	ENSG00000175691	pT1	GSE78099
ZNF772	PubM	8	ENSG00000197128	DKFZp686I1569	GSE120539
ZNF773	PubM	2	ENSG00000152439	MGC4728; ZNF419B	GSE200964
ZNF776	PubM	509	ENSG00000152443	FLJ38288	GSE78099
ZNF777	PubM	7286	ENSG00000196453	KIAA1285	GSE78099
ZNF778	PubM	4357	ENSG00000170100	FLJ31875	GSE76494; GSE78099
ZNF780A	PubM	1522	ENSG00000197782	ZNF780	GSE78099
ZNF780B	PubM	1287	ENSG00000128000	ZNF779	GSE200964
ZNF781	No transcript	NA	ENSG00000196381	FLJ37549	
ZNF782	PubM	81	ENSG00000196597	FLJ16636	GSE78099
ZNF783	PubM	3096	ENSG00000204946	DKFZp667J212	GSE78099
ZNF785	PubM	14	ENSG00000197162	FLJ32130	GSE78099
ZNF786	PubM	5643	ENSG00000197362	DKFZp762I137	GSE78099
ZNF789	PubM	664	ENSG00000198556		GSE78099

Name	Status	Peaks	Ensembl gene id	External synonym	GSE
ZNF790	PubM	45	ENSG00000197863	FLJ20350; MGC62100	GSE78099
ZNF791	PubM	164	ENSG00000173875	FLJ90396	GSE78099
ZNF792	PubM	2256	ENSG00000180884	FLJ38451	GSE104247; GSE78099
ZNF793	PubM	11636	ENSG00000188227		GSE78099
ZNF799	PubM	566	ENSG00000196466	HIT-40; MGC71805; ZNF842	GSE78099
ZNF8	PubM	131	ENSG00000278129	HF.18; Zfp128	GSE76494; GSE78099
ZNF805	PubM	266	ENSG00000204524		GSE78099
ZNF808	PubM	3426	ENSG00000198482		GSE78099
ZNF81	PubM	1549	ENSG00000197779	HFZ20; MRX45	GSE78099
ZNF813	PubM	421	ENSG00000198346	FLJ16542	GSE200964
ZNF814	PubM	NA	ENSG00000204514		GSE200964
ZNF816	PubM	288	ENSG00000180257	ZNF816A	GSE76494; GSE78099
ZNF823	PubM	56673	ENSG00000197933	HSZFP36	GSE78099
ZNF829	PubM	16	ENSG00000185869	DKFZp779O175	GSE200964
ZNF83	No transcript	NA	ENSG00000167766	FLJ11015; HPF1; ZNF816B	GSE51142
ZNF836	PubM	103	ENSG00000196267	FLJ16287	GSE200964
ZNF84	PubM	12151	ENSG00000198040	HPF2	GSE51142; GSE78099
ZNF841	PubM	145	ENSG00000197608	LOC284371	GSE200964
ZNF844	No over-expression	NA	ENSG00000223547	FLJ14959	
ZNF845	PubM	680	ENSG00000213799		GSE200964
ZNF846	PubM	405	ENSG00000196605		GSE78099
ZNF85	PubM	225	ENSG00000105750	HPF4; HTF1	GSE76494; GSE78099
ZNF850	PubM	286	ENSG00000267041	ZNF850P	GSE120539
ZNF852	PubM	45	ENSG00000178917		GSE200964
ZNF860	PubM	204	ENSG00000197385		GSE78099
ZNF875	PubM	290	ENSG00000181666	HKR1	GSE78099
ZNF878	PubM	789	ENSG00000257446		GSE200964
ZNF879	PubM	1475	ENSG00000234284	DKFZp686E2433	GSE78099
ZNF880	PubM	102	ENSG00000221923		GSE78099
ZNF888	No over-expression	NA	ENSG00000213793		
ZNF891	PubM	105	ENSG00000214029		GSE78099
ZNF90	PubM	9847	ENSG00000213988	HTF9	GSE78099
ZNF91	No DNA synthesis	NA	ENSG00000167232	HPF7; HTF10	GSE162571
ZNF92	PubM	153	ENSG00000146757	HPF12; TF12	GSE200964
ZNF93	PubM	4184	ENSG00000184635	FLJ12488; HPF34; TF34; ZNF505	GSE78099
ZNF98	No over-expression	NA	ENSG00000197360	F7175; ZNF739	
ZNF99	No DNA synthesis	NA	ENSG00000213973	C19orf9; MGC24986	

Methods

Census of the human KRAB Zinc Finger protein clusters

KZFP pairs were detected and their age defined as described in (Imbeault et al., 2017). In short, the human genome (hg19) was translated in 6 reading frames and scanned for zinc finger and KRAB domains using Hidden-Markov-Models (Pfam (El-Gebali et al., 2019): KRAB (PF01352) and zf-C2H2 (PF00096)). Hits for KRAB and zinc finger domains were combined based on proximity and strandness and then manually curated and integrated with existing gene or pseudogene annotations. Their age is based on sequencing similarity with orthologues in other species. The KZFP clusters were defined as having at least 3 KZFPs that are no more than 250 kb apart from the centre of another member, consistent with (Huntley et al., 2006). The clusters are named after their chromosome and then numbered starting from the short arm of the chromosome. The size of chromosomes and positions of centromeres were taken from UCSC genome browser annotation data for hg19 (Haeussler et al., 2019).

Human genetic variation data

Human genetic exome and whole genome sequencing data were obtained from The Genome Aggregation Database (gnomAD) (Karczewski et al., 2019; Lek et al., 2016) (release-2.0.2) for 123,136 and 15,496 individuals, respectively. The released genetic data was processed and filtered through several steps to guarantee that only high-quality variants were included. First, all variants +/- 1kb around the KZFP canonical transcripts as defined by Ensembl (v75, hg19) were extracted and filtered for variant quality, thus only retaining variants annotated as “PASS”. Second, all indels were normalized and multiallelic variants split using BCFTOOLS (v1.8) and reannotated with the Variant Effect Predictor (McLaren et al., 2016) and LOFTEE (v0.3beta). Third, all missense and loss of function (LoF) variants, defined as either frameshift, stop-gain or splice variants, were extracted from both the exome and whole genome datasets and either low confidence or flagged LoF variants were removed. The latter was primarily due to LoF variants found in the last 5% of the canonical transcript. Since genomic sequencing methods can yield variable coverage of genetic regions, especially when it comes to exome sequencing that is dependent on the capture of previously annotated protein-coding genes, we excluded all canonical transcripts having an average per-base coverage < 20x. Thus, bringing the total number of included KZFPs to 361. Furthermore, exons with an average per-base coverage < 20x were also removed, and the lengths of the coding sequences used later for normalizations were adjusted accordingly. Finally, the filtered exome and genome datasets were combined, and the allele counts and frequencies for all variants were recalculated, prior to the removal of all singletons (allele count = 1) to hinder inflation of observed mutational events due to potential technical artefacts.

Domain and site specifications

The genomic positions of the C2H2 zinc finger domains were obtained from the Ensembl database (v75, hg19). For each KZFP, only the ones from the canonical transcripts (as defined by Ensembl) were considered. The positions of the specific amino acids within these domains were computationally annotated. Z scores for the cysteine and histidine (C2H2) residues were calculated with the number of missense variants normalized to the number of zinc finger domains within the canonical transcript of each KZFP. For missense and LoF variants spanning either the whole CDS or a full protein domain, the number of variants per gene, x, was normalized by the length of the canonical coding sequence prior to Z score transformations.

$$Zscore = \frac{\left(x - mean(variantcount)\right)}{sd(variantcount)}$$

Cell Lines

HEK293T cells overexpressing HA-tagged KZFPs were generated as described in Imbeault et al., 2017. In short, cDNAs from the human KZFPs were codon-optimized and synthesized using the GeneArt service from ThermoFisher (former Life Technologies). Sequences were cloned into the doxycycline inducible expression vector pTRE-3HA which yields a C-terminally tagged proteins. Stable cell lines were generated using Lentivector transduction of mycoplasma free HEK293T cells as described on http://tronolab.epfl.ch. Presence and integrity of the integrated plasmids were verified using Sanger sequencing (primers: CMV1f: GGAGGCCTATATAAGCAGAGCTCGT, PGK4b: CGAACGGACGTGAAGAATGTGCGAGA) and KZFP expression was verified via western blot with an anti-HA antibody (ref. 12013819001, Roche) after >48h induction with 1ng/ml doxycycline. HEK293T cells were chosen in order to have a consistent cell line and genomic background for all conducted experiments.

ChIP-seq

Chromatin was prepared as described in Imbeault et al., 2017 and ChIP-seq was performed as described in Iouranova et al., 2022. In short: 30 mio KZFP expressing HEK293T cells were used after induction for more than 48h with 1ng/ml doxycycline. Cells were crosslinked with 1% methanol free formaldehyde for 10min before nuclear extraction followed by sonication in a Covaris E220 sonicator resulting in DNA fragments between 200-500bp. IP was performed overnight using 15µg anti-HA.11 antibody (BioLegend ref: 901503) coupled to 75ul Dynabeads Protein G (Invitrogen ref: 10009D). 10 ng of material for both total inputs and chromatin immunoprecipitated samples were used for library preparation. After end-repair and A-tailing,

Illumina IDT indexes were ligated to the samples. Aliquots were tested in qPCR to determine the optimal number of PCR cycles needed to amplify each library without reaching saturation. Libraries were size-selected using Ampure XP beads (Beckman Coulter), quality-checked on a Bioanalyzer DNA high sensitivity chip (Agilent) and quantified with a Qubit dsDNA HS assay kit (Qubit 2.0 Fluorometer, Invitrogen) using Illumina adapters. Libraries were sequenced as indicated on GEO: GSE200964.

Processing of ChIP-seq and ChIP-exo data

Both previously published (Helleboid et al., 2019; Imbeault et al., 2017) and new data were processed together. Reads were mapped to the human genome assembly hg19 using Bowtie2 short read aligner v2.3.5.1 (Langmead and Salzberg, 2012), using the --sensitive-local parameter. Prior to peak calling multi-mapped reads (MAPQ < 10), blacklisted regions and regions with high levels in input samples (greylist) defined by the R package GreyListChIP (Brown, 2022) were removed .Peaks were called using MACS2 v2.2.4 (Zhang et al., 2008) with defaults parameters except for *-q 0.01* and *--keep-dup all*.

External ChIP-seq data

To find KZFP ChIP-seqs done by others the programmatic access to GEO (Barrett et al., 2013) eSearch and eFetch functions were used to search and retrieve submissions containing any KZFP name but not the keywords “RNA” or “H3K”. The resulting hits were then manually curated using the GEOquery R package (Davis, 2022) in order to get bed files from ChIP-seq experiments. Peaks not called on hg19 were lifted over to hg19 using liftover from rtracklayer (Lawrence et al., 2022) and chain files from UCSC (Haeussler et al., 2019).

Enrichment on repeats

Repeat enrichment analyses from ChIP data were performed using pyTEnrich (https://alexdray86.github.io/pyTEnrich). Repeat annotations were obtained from UCSC (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/rmsk.txt.gz). The genome subset for the enrichment was generated by identifying regions with 0 coverage for each ChIP-seq and ChIP-exo using bedtools genomcov, merging regions with less than 100 bp distance and overlapping the resulting files with bedtools intersect. The resulting intervals were then filtered to be bigger than 40kb and removed from the analysis along with the Y-chromosome which is absent in HEK293T cells. Enrichments with FDR < 0.05 are considered significant. In order to normalize FDR between experiments -log10(FDR) were divided by their maximum yielding a scale from 0 to 1 or least to most enriched, values above 0.9 on this scale are considered most enriched.

Multiple sequence alignment plot and line plots

Multiple sequence alignment (MSA) plots where made as described in Iouranova et al., 2022. In short: FASTA sequences for the indicated subfamilies were extracted from the hg19 genome assembly, aligned using individually using MAFFT (Katoh and Standley, 2013) with parameters --reorder –auto, and then merged together using MAFFT's -merge option. To increase readability, positions in the alignment (columns) with more than 85% gaps were removed. To capture signal at the border the alignments are extended by 200-500bp of unaligned sequences. ChIP-seq and -exo signals are scaled for each line (row) to the [0,1] interval before being superimposed on the alignments. Average ChIP-seq signals across all rows are plotted on top of the alignments or without alignment for Figures 2 and 5. Motives were taken from Cisbp (Weirauch et al., 2014) converted to position weight matrixes and scanned for in the human genome (hg19) using PWMscan (Ambrosini et al., 2018) with default settings. Line plots in Figure3B were generated using deeptools plotProfile (Ramírez et al., 2016). SVA for all subfamilies (A-F) were centred on a well conserved region on the edge of the VNTR with the consensus sequence ACTAAGAAAAATTCTTCTGCCTTGGG.

Results II

This chapter represents additional results which were not included in the manuscript from the previous chapter in order to preserve a clear narrative, but are relevant for the discussion and perspectives of this thesis. Two major analyses associated with ChIP-seq data from transcription factors were not fully addressed in the manuscript: the identification of DNA binding motifs for the characterized KZFPs and the correlation between the expression of KZFPs and their targets. The following results represent efforts to answer these questions.

Key nucleotides for KZFP binding can be identified by comparing bound and unbound TE sequences

Considerable efforts were previously put into identifying the DNA binding motifs of KZFPs, culminating in the Cisbp database (Weirauch et al., 2014). These efforts were focused on a KZFP centric approach using the zinc-finger print of each KZFP to generate predictions which were then optimized with the experimental data from ChIP-seq. Here the results of an alternative, TE-centric approach are presented as a case study. We focus on ZNF627, a KZFP binding Tigger1 DNA transposons with high specificity and fidelity with 80% of its identified peaks covering the elements (Imbeault et al., 2017). First, we confirmed these findings by repeating the ChIP and checking for H3K9me3 deposition at target sites in cells overexpressing ZNF627 (Figure 13A). We focus on H3K9me3 as deposition of H3K9me3 together with the loss of H3K27ac it is the main downstream effect of KZFP binding and the deposition of new H3K9me3 is easier to observe than the loss of H3K27ac which depends on H3K27ac being present in a location to start with. Subsequently all Tigger1 elements (excluding fragments that do not contain the ZNF627 binding region) were clustered according to the ChIP-seq signal at the ZNF627 binding site, resulting in three clusters with either high (cluster1), medium (cluster2) or low (cluster3) ZNF627 binding (Figure 13A). The sequencing logo for the approximate ZNF627 binding site for cluster1 (Figure 13B) and cluster 3 (Figure 13C) revealed that consensus sequence for both regions are very similar with no apparent nucleotides differentiating binding of ZNF627 in cluster1 compared to non-binding in cluster3. However, if the position weight matrix for this region is calculated using the nucleotide frequencies of the low bound sequences (cluster3) as background for the bound regions (cluster1) we can identify several nucleotides (positions 18, 31, 35, 38 and 41 Figure 13D) necessary for efficient binding to cluster1. These nucleotides thus cannot be absent for binding to occur but their presence alone does not guarantee binding. For example, if a sequence has a G at position 35 it cannot be said if it will be bound by ZNF627 or not, as many unbound sequences also have Gs, but if it does have anything else than a G at that position binding is

highly unlikely. This allows for the identification of the key nucleotides necessary for ZNF627 binding and represents a promising approach to be further explored and used on more KZFPs.

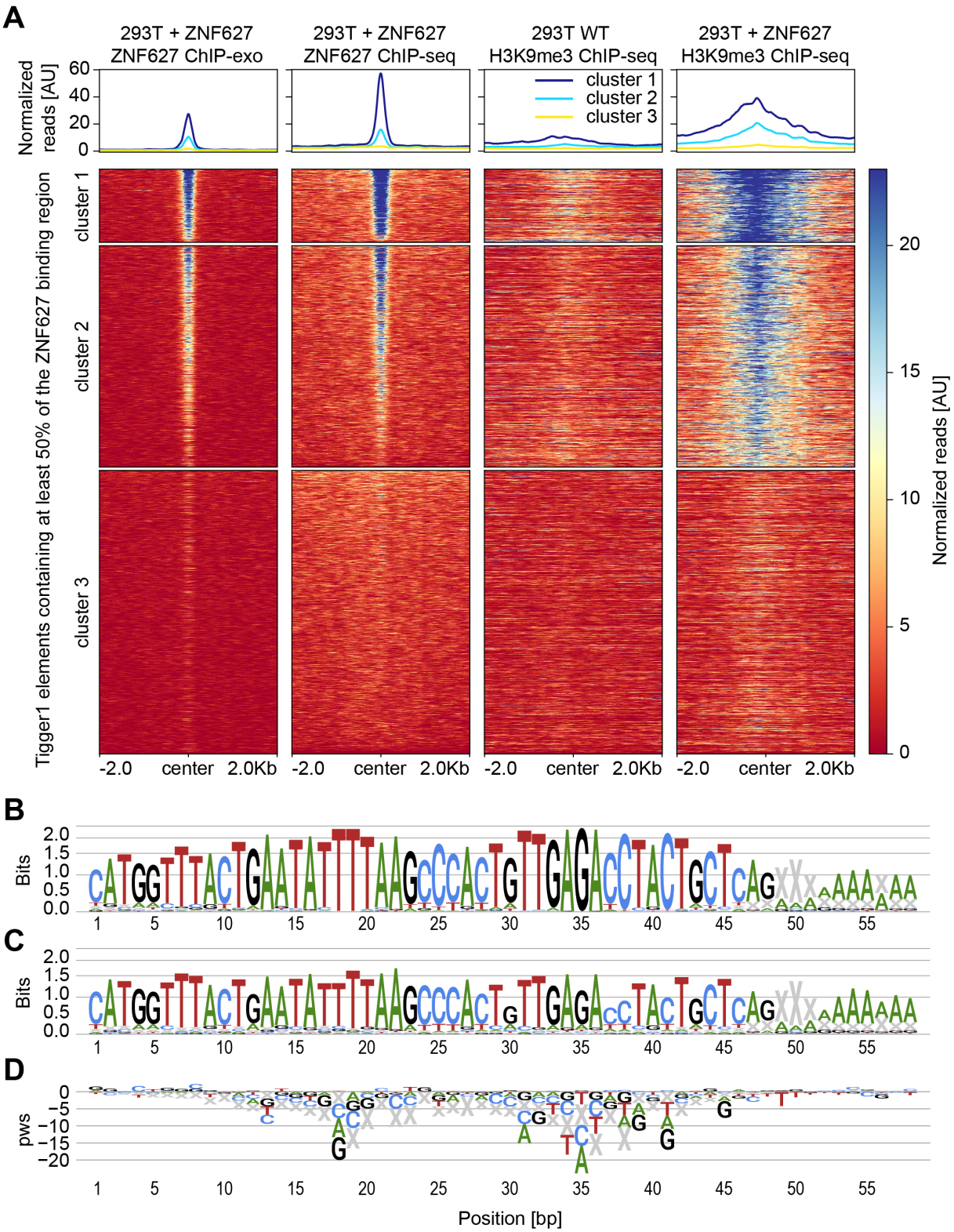


Figure 13: Identification of key nucleotides in ZNF627 binding.
A) Clustering of bound and unbound Tigger1 elements. Tigger1 elements carrying at least 50% of the approximate ZNF627 binding region (60bp) were aligned on this binding region and are the same for all four experiments (columns) shown. The signal from two independent experiments (ZNF627 ChIP-exo and ZNF627 ChIP-seq) in HEK293T cells overexpressing

ZNF627 (293T + ZNF627) were used to differentiate 3 clusters with either high (cluster1), medium (cluster2) or low (cluster3) signal. ZNF627 binding was confirmed by comparing H3K9me3 ChIP-seq in ZNF627 expressing (293T + ZNF627) to wildtype (293T WT) cells showing a marked increase at ZNF627 binding sites. The plots were generated using deeptools2 (Ramírez et al., 2016), ChIP-exo was previously performed by Imbeault et al., 2017. B-C) Sequence logo of the approximate binding site of ZNF627 in cluster1 (B) or cluster3 (C). The sequences from cluster1 depicted in A) were aligned and the frequency of each nucleotide as well as gaps in the alignment (X) were used to calculate the information content of each position of the sequence. D) Position weight matrix showing the position weight score (pws) for the nucleotide frequencies shown in B) and C). The height of each character represents the logarithm of the frequency of that character in cluster1 divided by its frequency in cluster3, for each position. Positive values indicate nucleotides that only occur in bound sequences (cluster1) while negative values indicate nucleotides that only occur in unbound sequences (cluster3).

KZFP expression does not correlate with the expression of their TE targets in differentiated or tumour tissues

After having identified the TE targets of a majority of KZFPs in the previous chapter we asked whether there was a correlation between KZFP expression and their identified TE targets both in physiological and pathological conditions. To answer this, correlations between KZFP and TE expression were calculated for healthy and tumour tissues using the Genotype-Tissue Expression project (GTEx) and The Cancer Genome Atlas (TCGA) datasets respectively (Figure 14). Both datasets show an overall lack of strong correlations between KZFPs and their TE targets, as medians are well contained in the -0.5 to 0.5 interval. If we compare the values for targets (red) with those for random non-targets (blue) we can observe that a few KZFPs do show differences but only for the tumour tissues from the TCGA dataset (Figure 14A). For the vast majority of tumour tissues and for all of the healthy tissues from the GTEx dataset (Figure 14B) however, the correlations were very similar between targets and non-targets. In general, it can thus be stated that the variations of KZFPs expression levels are not reflected by TE expression levels. The fact that certain tumour samples show a correlation might indicate a breakdown of normal gene regulation like local DNA demethylation or a more stem-like character of these tumours.

Figure 14: Correlation of KZFP expression with their target TEs.
Boxplots showing the spearman correlations of human KZFP expression values (log2 counts per million) with aggregated expression of their most enriched target TE subfamily (red) or random non-target TE subfamilies (blue). Each boxplot represents correlations from A) 33 types of tumour sample from the The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013) or B) 30 tissue types from healthy human donors from the Genotype-Tissue Expression project (GTEx Consortium et al., 2017). Outliers were omitted to increase readability.

KZFPs expression levels are fairly stable across tissues and generally not correlated based on KZFP cluster or target

The previous lack of correlation between KZFPs and their targets might be to a certain degree due to the similar expression level of KZFPs across tissues (Figure 15). Despite some tissue specificity of KZFP expression, such as higher levels in testis or lower levels in Liver hepatocellular carcinoma (LIHC), these differences are often still smaller than the differences in expression level between KZFPs. These differences in expression levels between KZFPs, which can be seen as 3 distinct clusters both for pathological (Figure 15A) and physiological (Figure 15B) suggests a certain level of co-regulation between KZFPs. Given that we had mostly observed KZFP targeting the same subfamily to be localized in different genomic clusters (see Figure 5 of the manuscript in the Results I chapter) we hypothesized that there is some co-regulation of KZFPs in the same genomic clusters. As a consequence, localizing KZFPs in different clusters would allow for distinct expression patterns of KZFPs targeting the same elements. To answer this question, we used the previously mentioned datasets from TCGA and GTEx and clustered KZFPs according to their expression (Figure 15). Our results show that expression levels are neither cluster nor age specific, thus KZFPs are regulated individually rather than on the cluster level. Given this result the question arose if on the other hand the expression of KZFPs targeting the same TE subfamilies are correlated. Upon investigation we can generally see no correlation between KZFPs targeting the same TE subfamily (Figure 16). However, for the few cases where we see consistent strong correlations such as MLT1I and Tigger2. The KZFPs targeting these elements RBAK and ZNF12 for MLT1I, and ZNF324 and ZNF324B for Tigger2 are located in close proximity in cluster 19.10 or on the short arm of chromosome 7. Thus, even though we generally do not have a correlation within clusters or with KZFPs targeting the same elements, if we see such correlations they are likely due to KZFPs being located in close proximity to each other.

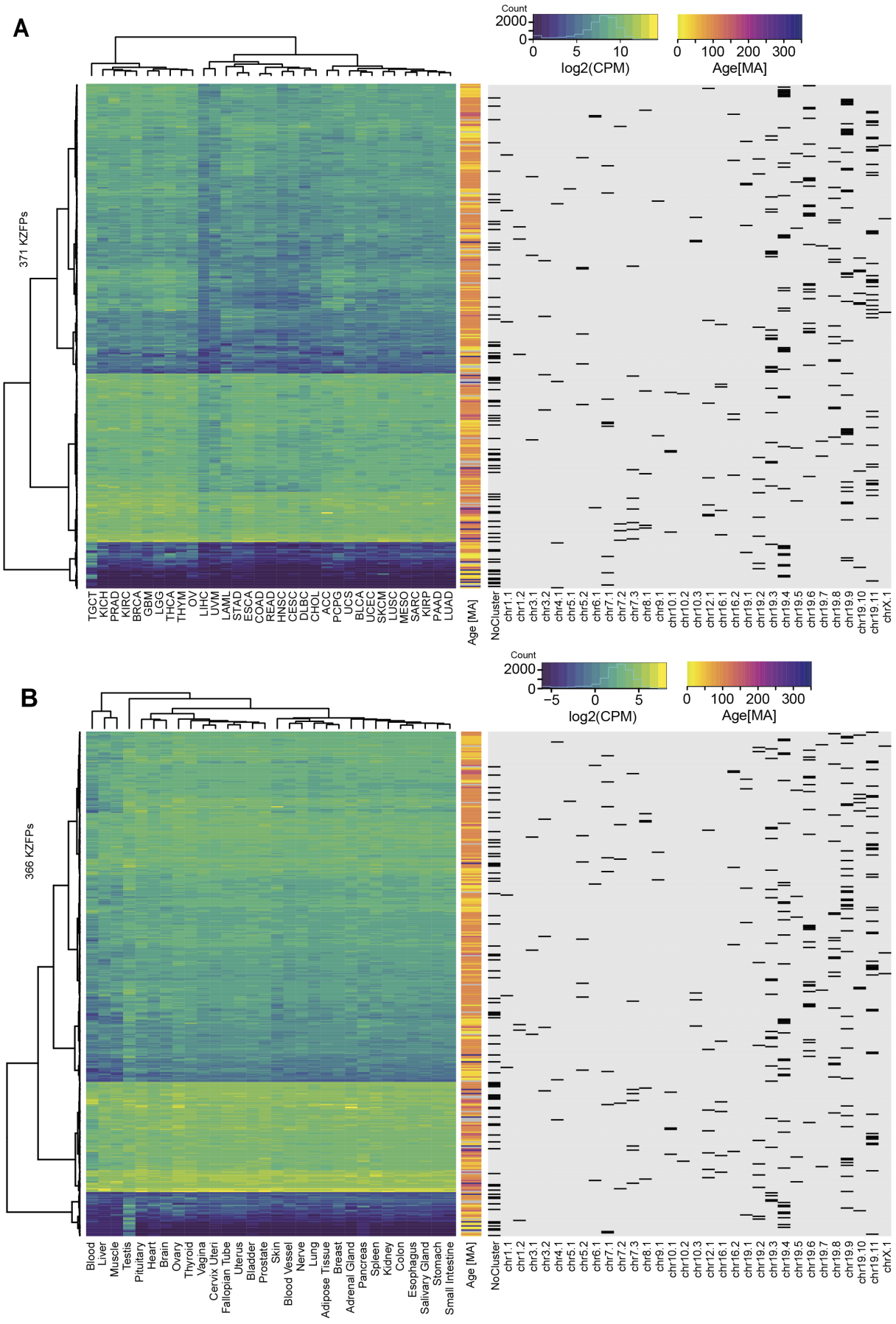

A
Count
2000
0
0 5 10
log2(CPM)
0 100 200 300
Age[MA]
371 KZFPs
TGCT KICH PRAD KIRC BRCA GBM LGG THCA THYM OV LIHC UVM LAML STAD ESCA COAD READ HNSC CESC DLBC CHOL ACC PCPG UCS BLCA UCEC SKCM LUSC MESO SARC KIRP PAAD LUAD
Age [MA]
NoCluster chr1.1 chr1.2 chr3.1 chr3.2 chr4.1 chr5.1 chr5.2 chr6.1 chr7.1 chr7.2 chr7.3 chr8.1 chr9.1 chr10.1 chr10.2 chr10.3 chr12.1 chr16.1 chr16.2 chr19.1 chr19.2 chr19.3 chr19.4 chr19.5 chr19.6 chr19.7 chr19.8 chr19.9 chr19.10 chr19.11 chrX.1
B
Count
2000
0
−5 0 5
log2(CPM)
0 100 200 300
Age[MA]
366 KZFPs
Blood Liver Muscle Testis Pituitary Heart Brain Ovary Thyroid Vagina Cervix Uteri Fallopian Tube Uterus Bladder Prostate Skin Blood Vessel Nerve Lung Adipose Tissue Breast Adrenal Gland Pancreas Spleen Kidney Colon Esophagus Salivary Gland Stomach Small Intestine
Age [MA]
NoCluster chr1.1 chr1.2 chr3.1 chr3.2 chr4.1 chr5.1 chr5.2 chr6.1 chr7.1 chr7.2 chr7.3 chr8.1 chr9.1 chr10.1 chr10.2 chr10.3 chr12.1 chr16.1 chr16.2 chr19.1 chr19.2 chr19.3 chr19.4 chr19.5 chr19.6 chr19.7 chr19.8 chr19.9 chr19.10 chr19.11 chrX.1

Figure 15: Expression of KZFPs in healthy and tumour tissues.
Left: Heatmap of expression values in counts per million [log2(CPM)] for KZFPs (rows) in different tissues: A) Tumour tissues from the The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013) or B) tissues from the Genotype-Tissue Expression project (GTEx Consortium et al., 2017). Both rows and columns were hierarchically clustered using a Ward function (Ward, 1963) with Eucledian distance. Center: Bar depicting the age of each KZFPs. Right: Heatmap showing the membership of each KZFP to a cluster (black). ACC= Adrenocortical carcinoma, LAML= Acute Myeloid Leukemia, BLCA= Bladder Urothelial Carcinoma, BRCA= Breast invasive carcinoma, CESC= Cervical squamous cell carcinoma and endocervical adenocarcinoma, CHOL= Cholangiocarcinoma, COAD= Colon adenocarcinoma, DLBC= Lymphoid Neoplasm Diffuse Large B-cell Lymphoma, ESCA= Esophageal carcinoma, GBM= Glioblastoma multiforme, HNSC= Head and Neck squamous cell carcinoma, KICH= Kidney Chromophobe, KIRC= Kidney renal clear cell carcinoma, KIRP= Kidney renal papillary cell carcinoma, LGG= Brain Lower Grade Glioma, LIHC= Liver hepatocellular carcinoma, LUAD= Lung adenocarcinoma, LUSC= Lung squamous cell carcinoma, MESO= Mesothelioma, OV= Ovarian serous cystadenocarcinoma, PAAD= Pancreatic adenocarcinoma, PCPG= Pheochromocytoma and Paraganglioma, PRAD= Prostate adenocarcinoma, READ= Rectum adenocarcinoma, SARC= Sarcoma, SKCM= Skin Cutaneous Melanoma, STAD= Stomach adenocarcinoma, TGCT= Testicular Germ Cell Tumours, THCA= Thyroid carcinoma, THYM= Thymoma, UCEC= Uterine Corpus Endometrial Carcinoma, UCS= Uterine Carcinosarcoma, UVM= Uveal Melanoma.

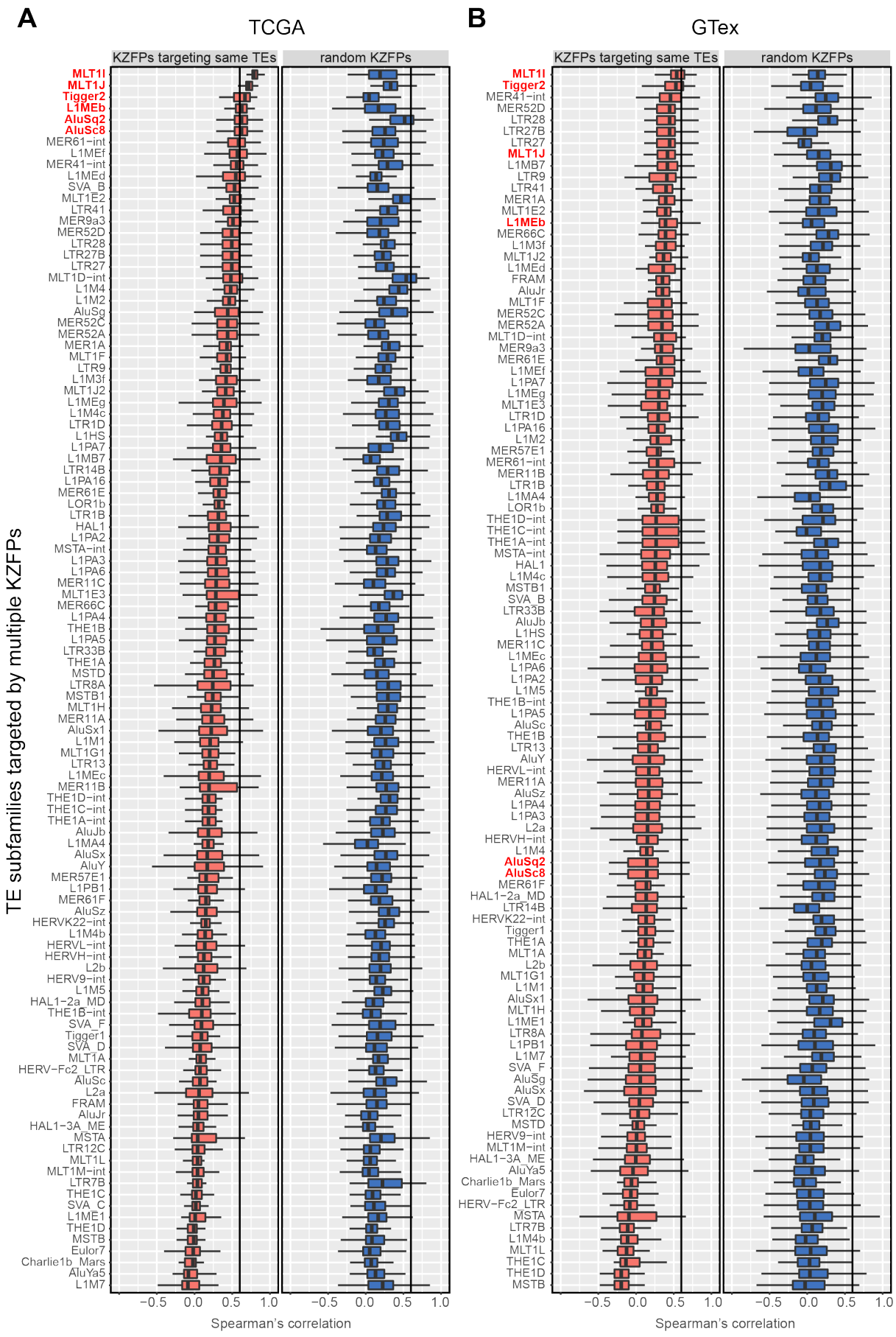

A
TCGA
KZFPs targeting same TEs
random KZFPs
TE subfamilies targeted by multiple KZFPs
MLT1I
MLT1J
Tigger2
L1MEb
AluSq2
AluSc8
MER61−int
L1MEf
MER41−int
L1MEd
SVA_B
MLT1E2
LTR41
MER9a3
MER52D
LTR28
LTR27B
LTR27
MLT1D−int
L1M4
L1M2
AluSg
MER52C
MER52A
MER1A
MLT1F
LTR9
L1M3f
MLT1J2
L1MEg
L1M4c
LTR1D
L1HS
L1PA7
L1MB7
LTR14B
L1PA16
MER61E
LOR1b
LTR1B
HAL1
L1PA2
MSTA−int
L1PA3
L1PA6
MER11C
MLT1E3
MER66C
L1PA4
THE1B
L1PA5
LTR33B
THE1A
MSTD
LTR8A
MSTB1
MLT1H
MER11A
AluSx1
L1M1
MLT1G1
LTR13
L1MEc
MER11B
THE1D−int
THE1C−int
THE1A−int
AluJb
L1MA4
AluSx
AluY
MER57E1
L1PB1
MER61F
AluSz
HERVK22−int
L1M4b
HERVL−int
HERVH−int
L2b
HERV9−int
L1M5
HAL1−2a_MD
THE1B−int
SVA_F
Tigger1
SVA_D
MLT1A
HERV−Fc2_LTR
AluSc
L2a
FRAM
AluJr
HAL1−3A_ME
MSTA
LTR12C
MLT1L
MLT1M−int
LTR7B
THE1C
SVA_C
L1ME1
THE1D
MSTB
Eulor7
Charlie1b_Mars
AluYa5
L1M7
−0.5
0.0
0.5
1.0
Spearman's correlation
B
GTex
KZFPs targeting same TEs
random KZFPs
MLT1I
Tigger2
MER41−int
MER52D
LTR28
LTR27B
LTR27
MLT1J
L1MB7
LTR9
LTR41
MER1A
MLT1E2
L1MEb
MER66C
L1M3f
MLT1J2
L1MEd
FRAM
AluJr
MLT1F
MER52C
MER52A
MLT1D−int
MER9a3
MER61E
L1MEf
L1PA7
L1MEg
MLT1E3
LTR1D
L1PA16
L1M2
MER57E1
MER61−int
MER11B
LTR1B
L1MA4
LOR1b
THE1D−int
THE1C−int
THE1A−int
MSTA−int
HAL1
L1M4c
MSTB1
SVA_B
LTR33B
AluJb
L1HS
MER11C
L1MEc
L1PA6
L1PA2
L1M5
THE1B−int
L1PA5
AluSc
THE1B
LTR13
AluY
HERVL−int
MER11A
AluSz
L1PA4
L1PA3
L2a
HERVH−int
L1M4
AluSq2
AluSc8
MER61F
HAL1−2a_MD
LTR14B
HERVK22−int
Tigger1
THE1A
MLT1A
L2b
MLT1G1
L1M1
AluSx1
MLT1H
L1ME1
LTR8A
L1PB1
L1M7
SVA_F
AluSg
AluSx
SVA_D
LTR12C
MSTD
HERV9−int
MLT1M−int
HAL1−3A_ME
AluYa5
Charlie1b_Mars
Eulor7
HERV−Fc2_LTR
MSTA
LTR7B
L1M4b
MLT1L
THE1C
THE1D
MSTB
−0.5
0.0
0.5
1.0
Spearman's correlation

Figure 16: Correlation of expression of KZFPs targeting the same TE subfamilies.
For each TE subfamily targeted by more than one KZFP (rows) the KZFPs which are most enriched (see "Enrichment on repeats" in the Methods section) were extracted and pairwise Spearman correlations were calculated for each tissue or tumour type (red). The correlations between different KZFPs and across different tissues or tumour types are shown as boxplots. The same procedure was repeated using the same number of randomly selected KZFPs not targeting that subfamily (blue). Subfamilies with strong median correlations (>0.6) in one of the two datasets are shown in red and the 0.6 cut-off is shown by a black vertical line. A) Correlations across the TCGA dataset (Weinstein et al., 2013). B) Correlations across the GTex dataset (GTEx Consortium et al., 2017).

Discussion

With the characterization of 95% of human KZFPs by CHIP-seq, a comprehensive overview of the family was achieved. We identified targets for the majority of KZFPs and could use the repetitive nature of TEs to verify those targets, with highly similar elements repeatedly being bound in the same location serving as replicates for the affinity and specificity of a KZFP. Findings, that were further confirmed with the available ChIP replicates and H3K9me3 data. Intriguingly, we could not observe a correlation between KZFP expression and their target TEs, despite various reports on the impact on expression of KZFP binding. A potential reason for this is that differentiated tissues were examined instead of focusing on early development. New TE insertions are only able to propagate if they occur in the germline, consequently there is not no fitness advantage in them being transcribed in somatic tissues even if a lower presence of KZFPs would permit it. Thus, changes in KZFPs mediated repression might not affect TE expression levels as the necessary activator signals are absent. This possibility however begs the question why KZFP are expressed in differentiated tissues. Putting aside the very plausible possibility that KZFP fulfill other roles than the immediate control of TE expression, it is also possible that the variation of KZFP expression present within the tissues is not big enough to observe effects on TE expression. However, experiments with homozygous knock-outs of individual KZFPs did not show any change in TE expression either (data not shown). An explanation for this lies in the high levels of redundancy with which TEs are targeted by KZFPs, resulting in any change in expression of one KZFP being potentially compensated by several other KZFPs. Interestingly we observed that the KZFPs targeting the same TEs do not co-localize in the same genomic cluster. Even though there is nothing forcing clustered KZFPs to target similar elements, as even highly related binding motifs can be found on very distinct TEs (Manuscript Figure S4), it is still remarkable as it shows that the multiple KZFPs targeting the same TEs are not merely a by-product of gene duplications. This observation also sheds a new light on the results reported by Wolf et al., 2020 where homozygous knock-outs in mice for two major KZFP clusters did not lead to any significant increase in TE transcription, implying that KZFPs in other clusters were able to compensate the loss. Even though the purpose of multiple KZFP targeting the same TEs requires further investigation, it can already be stated that there is no strong link between the clusters where a KZFP resides and its expression level, at least in differentiated tissues. The reverse is not true however, in the rare cases where expression of KZFPs is strongly correlated, we generally see their genes being located in close proximity to each other. Having said that, if disseminating KZFPs in different cluster serves a gene-regulatory role we do not see it reflected in a correlated expression of genes in the same cluster. Interestingly and in contrast to what was just discussed, a clear evolutionary history can be observed when looking at the targets of KZFPs that are located in the same cluster.

KZFPs very often share targets with their neighbour but at lower affinity, likely due to gene duplication events followed by genetic drift, leading to a new target sequence of one copy that is still related to the old one. Thus, even though this process does not lead to the clustering of KZFPs targeting the same TEs, it still occurs and can be readily observed. This discrepancy of on one hand KZFP targets not being related to genomic location and on the other hand neighbouring KZFPs showing traces of shared targets shows the independent rise of KZFPs against the same target. This independence in term implies distinct, non-redundant, roles for these different KZFPs that merit further investigation.

Concerning the difficulty to see the effect of KZFPs on their targets by using RNA-seq data, it might be more fitting to rely on H3K9me3 ChIP-seq instead. Even though less available and technically more difficult to generate than RNA-seq, H3K9me3 signals have thus far been more reliably connected to KZFP expression. The results for ZNF627 serve as an example for this, as binding correlates nicely with H3K9me3 deposition; however, expression of targets stays constant (data not shown). Relying on H3K9me3 circumvents the need for the TE to be expressed to see a down regulation upon over expression of a KZFP. Additionally, if the resolution of the H3K9me3 ChIP allows for it, signals from several KZFPs binding the same target at different locations could be distinguished. Studying H3K9me3 might also reveal KZFP activities not aimed at controlling expression but to influence chromatin state for alternative reasons. An example for this can be found in a recent study which shows how H3K9me3 deposition on TEs in mice influences CTCF binding and consequently chromatin folding (Gualdrini et al., 2022).

Identifying the genomic targets of a KZFPs is a key step in understanding its function and here we have generated a unique dataset to study it. We can show the interconnected evolution of KZFPs and TEs; how the KZFP family adapts to interact with new, active TEs, how older, inactive TEs, are bound by KZFPs with non-repressive functions and how certain new TEs might be using older pre-existing KZFPs to facilitate their spread. This lets us envision a KZFP-TE evolutionary cycle (Figure 17) where a new TE (Figure 17A) leads to the adaptation of several new KZFPs (Figure 17B). The KZFPs enabling the TE to persist and even spread in the organism without causing catastrophic damage (Figure 17C). This can result in an escape of the TE from KZFP control (Figure 17Di) resulting in an arms-race scenario starting the cycle anew. Alternatively, if the controlled situation is maintained, TEs will over time be inactivated by random mutations leading to a disappearance of KZFPs targeting them with the exception of TEs and or KZFPs that have taken on regulatory roles (Figure 17Dii). These KZFPs with alternative roles can then potentially be used by new TEs, allowing them to spread and leading again to an adaptation of the KZFP (Figure 17E).

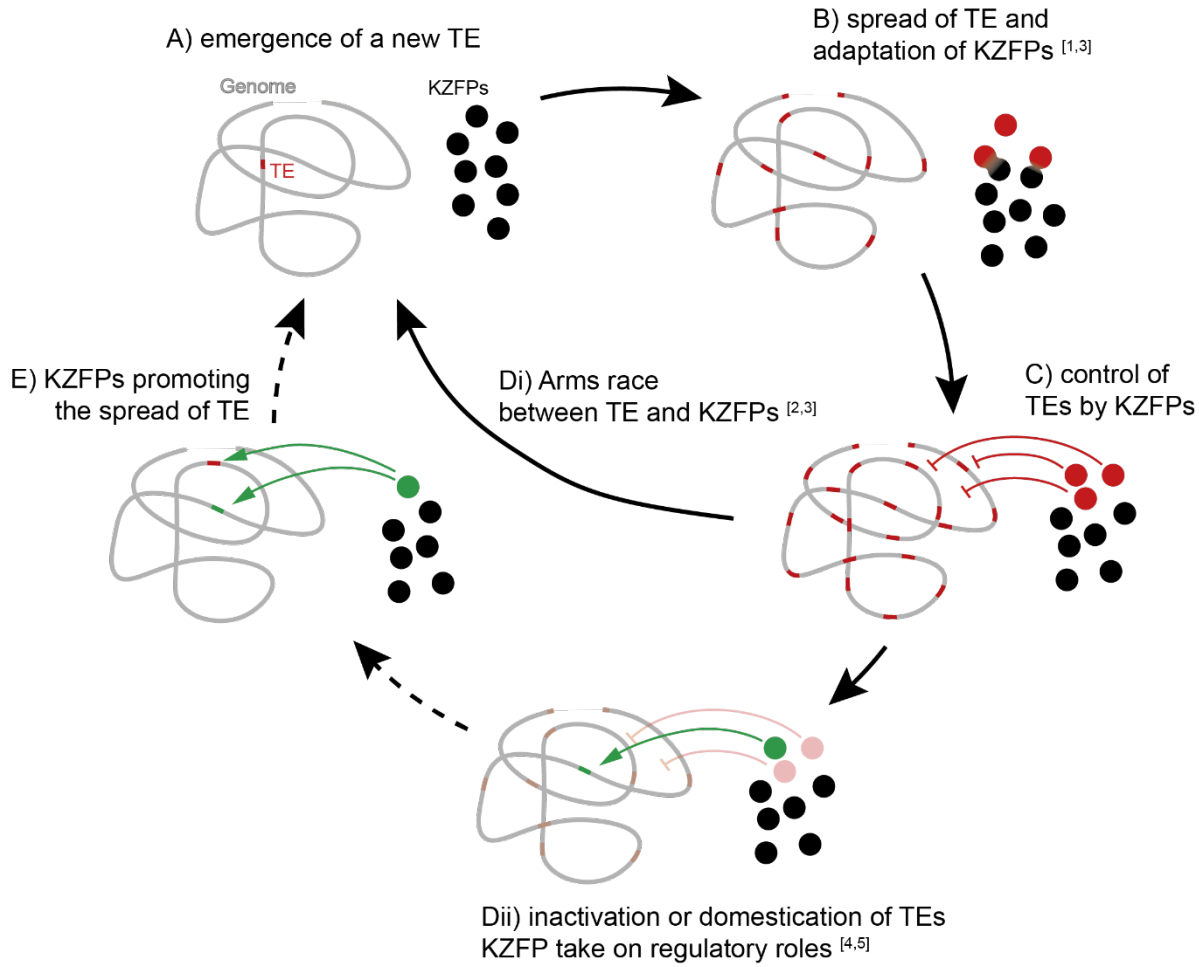


Figure 17: Model of KZFP-TE co-evolution.
A) A new TE (red) emerges in the genome (grey), not recognized by the existing pool of KZFPs (black dots). B) As the TE spreads the pool of KZFPs adapts through segmental gene duplication and genetic drift generating new genes that recognize the TEs (red dots). This process occurs independently at several locations C) Several KZFPs bind to and control the new TEs leading to a stable situation where the host can survive while often the TE is still able to spread with low frequency. Di) If the TE adapts to evade KZFP control it can spread again and start a new cycle leading to an arms race between KZFPs and TE. Dii) Alternatively with time, the vast majority of TE insertions become inactive through mutation and recombination, leading to a loss of repressive KZFPs (pale red). TEs however that have evolved regulatory or other beneficial functions will be conserved through this process (green) which in term can lead to a new regulatory role of the KZFPs binding them which is not necessarily repressive (green dot). E) New TEs (red) might use these regulatory KZFPs (green) facilitating their spread. [1] Thomas and Schneider, 2011 [2] Jacobs et al., 2014 [3] Imbeault et al., 2017 [4] Helleboid et al., 2019 [5] Tycko et al., 2020.

We can see several stages of this model in the human genome: The recently active or still active LTR and LINE1 elements are being targeted by several KZFPs each, corresponding to a situation where KZFPs are newly controlling the TEs or being evaded in an arms race scenario (Imbeault et al., 2017; Jacobs et al., 2014). The fact that these TEs are still expressed despite being targeted by multiple KZFPs showcases that TE expression cannot be solely deleterious for the host, else they would be completely silenced as soon as the tools to do so

are available. Much more likely we are observing a situation where a certain, controlled amount of TE expression is advantageous or at least well tolerable for the host. This advantage had first been hypothesized by Britten and Davidson in 1969 where they see TEs as an ideal tool for the dissemination of regulatory sequences, but can also be independent of transposition and simply due to TE embedded cis-regulatory elements requiring TE expression (Iouranova et al., 2022; Pontis et al., 2019). If a controlled situation between KZFPs and TEs persists a gradual loss of non-beneficial TEs and TE-KZFP interactions will occur. The fact that older LINE2 elements which are highly mutated and fragmented are bound by much less KZFPs, many of which have lost their repressive capabilities, suggests gene regulatory roles of the remaining TE-KZFP interactions. Finally, we also see that the very young and currently spreading SVAs are bound by several older KZFPs suggesting that SVAs might spread with the help of those KZFPs and not despite them.

In summary, through identification of the targets of most of human KZFPs we gained new insights in the evolution and function of the KZFP family. We see adaptations of the KZFPs to the spread of new TE families and we can observe the evolution through local gene duplications. Surprisingly these local duplications lead to neither an aggregation of KZFPs sharing targets or expression patterns. Together our results, made available on our web portal (https://tronoapps.epfl.ch/web/krabopedia/), provide an open resource for the future exploration of KZFPs and the TEs they control.

Perspectives

Moving forward with the investigation of human KZFPs, there are several avenues that merit exploration.

Fostering interactions with other researchers through our web portal

One of the aims of this thesis was to make the KZFP gene family more accessible to the scientific community. Our web portal the KRABopedia will enable anyone to quickly gain access to all our data concerning KZFPs, allowing them to view, interpret and verify our findings. This represents an opportunity for us to directly interact with people viewing and using our results, allowing us to adapt and expand in response to new ideas and findings. The aim is to create a central repository for anyone interested in KZFP functions or KZFPs mediated TE regulation. Currently, KZFPs detected in a genetic screen are often not pursued, as no specific information is available, by changing this we hope to facilitate impactful new findings. A better understanding of KZFPs should also assist research on TEs, with its frequently reported findings regarding their regulatory potential (Barnada et al., 2022; Bodea et al., 2022; Bonaventura et al.; Carter et al., 2022; Du et al., 2022; Iouranova et al., 2022; Kaemena et al., 2022).

Identification of KZFP targets - future steps

DNA binding motifs

As described in the Results II chapter, the data we generated allows to revisit the DNA binding motifs of KZFPs from a TE centric standpoint. The identification of key nucleotides for binding, through comparison of bound and unbound elements, could represent a next step in understanding KZFP function, and potentially shed light on how the sequence specificity of zinc-finger arrays arises. The key challenges for this undertaking are to identify both bound and unbound regions with high fidelity. In the example shown in the chapter Results II many manual steps where involved that would have to be generalized and probably require the development of new bioinformatic tools. However, the prospect of having highly accurate, generalizable DNA binding motifs is quite exciting and worth some investment.

Effects of KZFPs on targets

The fact that we do not see correlations between the expression of KZFPs and their targets merits further study. It is a common phenomenon and has been observed by several people, however rarely been published with the report by Wolf et al., 2020 being an exception. We ourselves have refrained from adding the results from our transcriptional analysis to our manuscript as it would divert attention from our main message. However, we believe that a thorough investigation of the subject would be of great benefit for the whole field and merit a

separate publication. A frequent explanation for discrepancy is the redundancy of KZFPs targeting the same TEs, our dataset should allow to identify redundant KZFP and target them simultaneously. However, another approach which has been effective is to rely on H3K9me3 as a readout of KZFP activity instead of transcription. When investigating the targets of a KZFP after modifying KZFP levels, changes in the expression of targets are often absent, but changes in H3K9me3 levels can readily be observed. If neither changing the model system to one with more active TEs, such as human embryonic stem cells, nor targeting multiple KZFPs simultaneously lead to satisfying results, it might be advisable to systematically acquire H3K9me3 ChIP-seq for each KZFP in the same manner the HA ChIP-seqs were conducted. In hindsight matching H3K9me3 ChIP-seq for each overexpression experiment would have been the most useful additional data to have when determining the targets of our HA ChIP-seq experiments. Acquiring this data, using the cell lines that are already on our possession, seems to be a key step in understanding this lack of correlation.

KZFPs with interesting targets

Our work revealed several intriguing KZFPs targeting specific TEs which merit further investigation. Here I would like to elaborate on these, while also highlighting how the data available on our web portal can be used to do so.

DNA transposons

As previously stated no DNA transposons are currently active in the human genome, however we find several KZFPs to be highly enriched specifically on DNA transposons (Manuscript Figure 2B). A look at the enrichments and multiple sequence alignments (MSA) available on the KRABopedia shows several interesting patterns. We have three KZFPs ZNF285, ZNF324B, and ZNF599, all of which have a very low number of peaks in total but a large fraction of those fall on DNA transposons. If we consult the MSA plots we see that the binding occurs on a few small fragments of the subfamily, whereas the majority of the TE elements are not targeted anymore. This makes both the KZFPs and those TE fragments prime suspects for having evolved to fulfil regulatory roles. Similarly, we see ZNF180 and ZNF627 binding only DNA transposons, but compared to the previous mentioned KZFPs they target many more integrands. It is unclear if there is a purpose in each of those targets or if the irrelevant ones did not have sufficient time to be removed by genetic drift and in time a situation as for ZNF285, ZNF324B and ZNF599 will be established. Finally, ZNF23 and ZNF585B show interesting evolutionary trajectories as both of them target LINE elements as well as DNA transposons, revealing a drift away from targeting LINE elements towards DNA transposons. This is nicely exemplified with ZNF585B which has a paralog ZNF585A binding LINE elements. The MSA plot for ZNF585B shows a clear signal over HSMAR2, while the LINE1 elements are bound very sporadically. Thus, it is reasonable to assume that after a duplication event of the

ancestral ZNF585 gene the ZNF585A copy continued binding LINE1 elements whereas ZNF585B evolved to target HSMAR2 elements. A follow-up study on both the targeted DNA transposons and their KZFPs could reveal these unreported regulatory functions.

LINE2 elements

Even more so than DNA transposons, LINE2 elements are highly fragmented and have been exposed to random mutations rendering them completely inactive. KZFPs targeting them in order to limit their spread have been superfluous for millions of years, thus any remaining specific targeting of LINE2 elements should have alternative reasons. More precisely the conservation of both the KZFP and its LINE2 embedded binding site, indicates that mutations in either of those are deleterious for the host and have been eliminated by natural selection. This moves the KZFPs targeting LINE2 elements far away from seeing them as repressors of TE expression and more into the field of transcription factors. In this context it is not surprising that we see two KZFPs also carrying a SCAN domain targeting LINE2 elements: ZKSCAN1 and ZKSCAN8. This additional domain can mediate other functions than recruitment of TRIM28, explaining the conservation of these KZFPs. Similar to ZNF285, ZNF324B, and ZNF599 for DNA transposons, ZNF3 has only a few peaks on very degenerated L2 elements, again hinting at a regulatory function of those TEs, as only very specific elements are targeted. On the other side of the spectrum we have ZNF189 which still binds many LINE2 elements across several subfamilies, looking much more like a classical repressor of LINE2 expression and making a regulatory function of a single targeted sequence less likely. This could on one hand indicate that, even though most certainly unable to retrotranspose, LINE2 expression still requires some broad control to limit deleterious effects. On the other hand, the conservation of the ZNF189 binding site across so many elements could also be due to a structural or more general regulatory role of ZNF189. In between ZNF3 and ZNF189 there is ZNF662 which binds elements of several LINE2 subfamilies but only a few at a time, further investigation of the prevalence of the ZNF662 binding sites and the location of the targeted integrands is necessary to better understand this case. Finally, for LINE2 elements we can again see the remnants of KZFP evolution in their targets with ZNF677, ZNF716 and ZNF77 all showing low affinity for LINE2 elements and much higher affinity for younger LINE1 or LTR elements. This highlights how KZFPs that were initially targeting LINE2 elements have adapted to target new TEs as LINE2 elements were continuously mutating, rendering their control through these KZFPs superfluous. Altogether, as LINE2 elements have already been found to have regulatory functions in the brain (Petri et al., 2019), our findings should allow for a better understanding of such established results and enable further LINE2 mediated regulatory functions to be revealed.

Alu elements

The peculiar situation of SINEs, specifically Alu elements is that they are primarily targeted by KZFPs older than themselves. A closer investigation of a few Alu binders such as ZNF135, ZNF284, ZNF441 and ZNF460 reveals that they bind almost exclusively Alus across several subfamilies and many elements, which makes them appear to be classical repressors that arose to control the spread of an element. It seems thus that the tools to prevent the spread of these subfamilies of Alus were present before they appeared. This poses two questions for further investigation: First, are the ages of the Alus wrong and secondly are the KZFPs targeting those elements hindering or aiding in their dissemination through the genome? If these Alus are actually older than we date them the KZFP data would perfectly align with the one for other TEs such as LINE1s and LTRs. The ages we use are determined by dfam (Hubley et al., 2016) where they look across species for the first appearance of an element. This might be problematic for short Alu elements as alignments might fail. However, if we assume the age is correct this would mean that the spread of Alus was not in spite of TEs but more likely thanks to KZFPs. Only Alus with a KZFP binding them were able to spread, which would be an interesting new evolutionary trajectory to investigate.

SVA elements

Similar to Alu elements, SVA elements are targeted by KZFPs that existed before their appearance. However contrary the Alu elements there is no concern with the ageing of either the TE nor the KZFPs as both are drastically different. Thus, Investigating the effect of these KZFPs on their bound elements to see if instead of preventing they are facilitating their spread represents an exciting prospect to unravel the evolution these TEs.

List of abbreviations

Abbreviation	Explanation
293T	Human embryonic kidney cells 293 T
5mC	5-methyl Cysteine
A	Adenine
AA	Amino Acid
ACC	Adrenocortical carcinoma
AU	Arbitrary Units
BLCA	Bladder Urothelial Carcinoma
bp	base pairs
BRCA	Breast invasive carcinoma
C	cytosine
C2H2	2 Cysteine – 2 Histidine
Cas9	CRISPR-associated 9
CAZF score	Coudray Alexandre Zinc Finger score
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
ChIP	Chromatin Immunoprecipitation
ChIP-exo	ChIP with exonuclease treatment
ChIP-seq	ChIP followed by massively parallel sequencing
CHOL	Cholangiocarcinoma
COAD	Colon adenocarcinoma
CpG	Cystein-phosphate-Guanine
CPM	Counts per million
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CRISPRi	CRISPR inhibition
dCas9	dead Cas9
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
DNA	Deoxyribonucleic acid
DNMT1	DNA methyltransferase 1
DUF	Domain of Unknown Function
ERV	Endogenous retrovirus
ESCA	Esophageal carcinoma
G	guanine
GBM	Glioblastoma multiforme
GTEx	Genotype-Tissue Expression
H3K27ac	Histone 3 Lysine 27 acetyl
H3K36me	Histone 3 Lysine 36 methyl
H3K4me	Histone 3 Lysine 4 methyl
H3K9me3	Histone 3 Lysine 9 methyl 3
HA	Human influenza hemagglutinin
HDAC	Histone deacetlyases
HEK293T	Human embryonic kidney cells 293 T
HMM	Hidden Markov Model
HMT	Histone Methyltransferases
HNSC	Head and Neck squamous cell carcinoma
HP1	Heterochromatin-Protein 1
HUH	histidine-hydrophobic-histidine
IP	Immunoprecipitation
KAP1	KRAB-associated protein 1 (a.k.a. TRIM28)
KICH	Kidney Chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
KRAB	Krüppel-associated box

KZFP	Krüppel-associated box domain-containing zinc-finger proteins
LAML	Acute Myeloid Leukemia
LGG	Brain Lower Grade Glioma
LIHC	Liver hepatocellular carcinoma
LINE	Long Interspersed Elements
log2	logarithm at base 2
LTR	Long Terminal Repeats
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MESO	Mesothelioma
NuRD	Nucleosome Remodeling Deacetylase
ORF	Open reading frame
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PCPG	Pheochromocytoma and Paraganglioma
PHD	plant homeodomain
PRAD	Prostate adenocarcinoma
pws	position weight score
RBCC	RING Bbox Coiled Coil
READ	Rectum adenocarcinoma
RING	Really Interesting New Gene
RNA	Ribonucleic acid
RT	reverse transcription
SARC	Sarcoma
SINE	Short Interspersed Elements
SKCM	Skin Cutaneous Melanoma
STAD	Stomach adenocarcinoma
SUMO	Small Ubiquitin-like Modifier
SVA	SINE-VNTR-Alu
T	thymine
TCGA	The Cancer Genome Atlas
TE	Transposable element
TEeRS	TE embededed regulatory sequences
TGCT	Testicular Germ Cell Tumors
THCA	Thyroid carcinoma
THYM	Thymoma
TRIM28	Tripartite motif-containing 28 (a.k.a. KAP1)
TSD	Target site duplications
UCEC	Uterine Corpus Endometrial Carcinoma
UCS	Uterine Carcinosarcoma
UTR	Untranslated region
UVM	Uveal Melanoma
VNTR	Variable Number of Tandem Repeats
MITE	Miniature Inverted-Repeat Transposable Elements

Biblography

Al Chiblak, M., Steinbeck, F., Thiesen, H.-J., and Lorenz, P. (2019). DUF3669, a “domain of unknown function” within ZNF746 and ZNF777, oligomerizes and contributes to transcriptional repression. BMC Mol. Cell Biol. *20*, 60. https://doi.org/10.1186/s12860-019-0243-y.

Alisch, R.S., Garcia-Perez, J.L., Muotri, A.R., Gage, F.H., and Moran, J.V. (2006). Unconventional translation of mammalian LINE-1 retrotransposons. Genes Dev. *20*, 210–224. https://doi.org/10.1101/gad.1380406.

Ambrosini, G., Groux, R., and Bucher, P. (2018). PWMScan: a fast tool for scanning entire genomes with a position-specific weight matrix. Bioinformatics *34*, 2483–2484. https://doi.org/10.1093/bioinformatics/bty127.

Aronovich, E.L., McIvor, R.S., and Hackett, P.B. (2011). The Sleeping Beauty transposon system: a non-viral vector for gene therapy. Hum. Mol. Genet. *20*, R14–R20. https://doi.org/10.1093/hmg/ddr140.

Baillie, J.K., Barnett, M.W., Upton, K.R., Gerhardt, D.J., Richmond, T.A., De Sapio, F., Brennan, P.M., Rizzu, P., Smith, S., Fell, M., et al. (2011). Somatic retrotransposition alters the genetic landscape of the human brain. Nature *479*, 534–537. https://doi.org/10.1038/nature10531.

Bank, R.P.D. RCSB PDB - 1A1L: ZIF268 ZINC FINGER-DNA COMPLEX (GCAC SITE). ZIF268 ZINC FINGER-DNA COMPLEX GCAC SITE.

Barde, I., Rauwel, B., Marin-Florez, R.M., Corsinotti, A., Laurenti, E., Verp, S., Offner, S., Marquis, J., Kapopoulou, A., Vanicek, J., et al. (2013). A KRAB/KAP1-miRNA cascade regulates erythropoiesis through stage-specific control of mitophagy. Science *340*, 350–353. https://doi.org/10.1126/science.1232398.

Barnada, S.M., Isopi, A., Tejada-Martinez, D., Goubert, C., Patoori, S., Pagliaroli, L., Tracewell, M., and Trizzino, M. (2022). Genomic features underlie the co-option of SVA transposons as cis-regulatory elements in human pluripotent stem cells. 2022.01.10.475682. https://doi.org/10.1101/2022.01.10.475682.

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., et al. (2013). NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. *41*, D991–D995. https://doi.org/10.1093/nar/gks1193.

Belshaw, R., Watson, J., Katzourakis, A., Howe, A., Woolven-Allen, J., Burt, A., and Tristem, M. (2007). Rate of Recombinational Deletion among Human Endogenous Retroviruses. J. Virol. *81*, 9437–9442. https://doi.org/10.1128/JVI.02216-06.

Birtle, Z., and Ponting, C.P. (2006). Meisetz and the birth of the KRAB motif. Bioinformatics *22*, 2841–2845. https://doi.org/10.1093/bioinformatics/btl498.

Bodea, G.O., Ferreiro, M.E., Sanchez-Luque, F.J., Botto, J.M., Rasmussen, J., Rahman, M.A., Fenlon, L.R., Gubert, C., Gerdes, P., Bodea, L.-G., et al. (2022). LINE-1 retrotransposon activation intrinsic to interneuron development. 2022.03.20.485017. https://doi.org/10.1101/2022.03.20.485017.

Bonaventura, P., Alcazer, V., Mutez, V., Tonon, L., Martin, J., Chuvin, N., Michel, E., Boulos, R.E., Estornes, Y., Valladeau-Guilemond, J., et al. Identification of shared tumor epitopes from endogenous retroviruses inducing high-avidity cytotoxic T cells for cancer immunotherapy. Sci. Adv. *8*, eabj3671. https://doi.org/10.1126/sciadv.abj3671.

Britten, R.J., and Davidson, E.H. (1969). Gene Regulation for Higher Cells: A Theory. Science *165*, 349–357. https://doi.org/10.1126/science.165.3891.349.

Brown, G. (2022). GreyListChIP: Grey Lists -- Mask Artefact Regions Based on ChIP Inputs (Bioconductor version: Release (3.14)).

Brunetti, L., Gundry, M.C., and Goodell, M.A. (2017). DNMT3A in Leukemia. Cold Spring Harb. Perspect. Med. *7*, a030320. https://doi.org/10.1101/cshperspect.a030320.

Cantone, I., and Fisher, A.G. (2013). Epigenetic programming and reprogramming during development. Nat. Struct. Mol. Biol. *20*, 282–289. https://doi.org/10.1038/nsmb.2489.

Cao, Y., Chen, G., Wu, G., Zhang, X., McDermott, J., Chen, X., Xu, C., Jiang, Q., Chen, Z., Zeng, Y., et al. (2019). Widespread roles of enhancer-like transposable elements in cell identity and long-range genomic interactions. Genome Res. *29*, 40–52. https://doi.org/10.1101/gr.235747.118.

Carter, T.A., Singh, M., Dumbović, G., Chobirko, J.D., Rinn, J.L., and Feschotte, C. (2022). Mosaic cis-regulatory evolution drives transcriptional partitioning of HERVH endogenous retrovirus in the human embryo. ELife *11*, e76257. https://doi.org/10.7554/eLife.76257.

Castro-Mondragon, J.A., Riudavets-Puig, R., Rauluseviciute, I., Berhanu Lemma, R., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Manosalva Pérez, N., et al. (2022). JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Res. *50*, D165–D173. https://doi.org/10.1093/nar/gkab1113.

Chauhan, S., Goodwin, J.G., Chauhan, S., Manyam, G., Wang, J., Kamat, A.M., and Boyd, D.D. (2013). ZKSCAN3 Is a Master Transcriptional Repressor of Autophagy. Mol. Cell *50*, 16–28. https://doi.org/10.1016/j.molcel.2013.01.024.

Chen, W., Schwalie, P.C., Pankevich, E.V., Gubelmann, C., Raghav, S.K., Dainese, R., Cassano, M., Imbeault, M., Jang, S.M., Russeil, J., et al. (2019). ZFP30 promotes adipogenesis through the KAP1-mediated activation of a retrotransposon-derived Pparg2 enhancer. Nat. Commun. *10*, 1809. https://doi.org/10.1038/s41467-019-09803-9.

Ciccarone, F., Tagliatesta, S., Caiafa, P., and Zampieri, M. (2018). DNA methylation dynamics in aging: how far are we from understanding the mechanisms? Mech. Ageing Dev. *174*, 3–17. https://doi.org/10.1016/j.mad.2017.12.002.

Cordaux, R., and Batzer, M.A. (2009). The impact of retrotransposons on human genome evolution. Nat. Rev. Genet. *10*, 691–703. https://doi.org/10.1038/nrg2640.

Craig, N.L., Chandler, M., Gellert, M., Lambowitz, A., Rice, P.A., and Sandmeyer, S. (2015). Mobile DNA III.

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

Cramer, P., Armache, K.-J., Baumli, S., Benkert, S., Brückner, F., Buchen, C., Damsma, G., Dengl, S., Geiger, S.R., Jasiak, A.J., et al. (2008). Structure of Eukaryotic RNA Polymerases.

Annu. Rev. Biophys. *37*, 337–352. https://doi.org/10.1146/annurev.biophys.37.032807.130008.

Crick, F. (1970). Central Dogma of Molecular Biology. Nature *227*, 561–563. https://doi.org/10.1038/227561a0.

Crick, F.H. (1958). On protein synthesis. Symp. Soc. Exp. Biol. *12*, 138–163. .

Cunningham, F., Allen, J.E., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Austine-Orimoloye, O., Azov, A.G., Barnes, I., Bennett, R., et al. (2022). Ensembl 2022. Nucleic Acids Res. *50*, D988–D995. https://doi.org/10.1093/nar/gkab1049.

Davis, S. (2022). GEOquery: Get data from NCBI Gene Expression Omnibus (GEO) (Bioconductor version: Release (3.14)).

De Cecco, M., Ito, T., Petrashen, A.P., Elias, A.E., Skvir, N.J., Criscione, S.W., Caligiana, A., Brocculi, G., Adney, E.M., Boeke, J.D., et al. (2019). L1 drives IFN in senescent cells and promotes age-associated inflammation. Nature *566*, 73–78. https://doi.org/10.1038/s41586-018-0784-9.

Du, A.Y., Zhuo, X., Sundaram, V., Jensen, N.O., Chaudhari, H.G., Saccone, N.L., Cohen, B.A., and Wang, T. (2022). Evolution of transposable element-derived enhancer activity. 2022.03.16.483999. https://doi.org/10.1101/2022.03.16.483999.

Durnaoglu, S., Lee, S.-K., and Ahnn, J. (2021). Human Endogenous Retroviruses as Gene Expression Regulators: Insights from Animal Models into Human Diseases. Mol. Cells *44*, 861–878. https://doi.org/10.14348/molcells.2021.5016.

Ecco, G., Cassano, M., Kauzlaric, A., Duc, J., Coluccio, A., Offner, S., Imbeault, M., Rowe, H.M., Turelli, P., and Trono, D. (2016). Transposable Elements and Their KRAB-ZFP Controllers Regulate Gene Expression in Adult Tissues. Dev. Cell *36*, 611–623. https://doi.org/10.1016/j.devcel.2016.02.024.

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

Ehrlich, M., Gama-Sosa, M.A., Huang, L.-H., Midgett, R.M., Kuo, K.C., McCune, R.A., and Gehrke, C. (1982). Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. Nucleic Acids Res. *10*, 2709–2721. https://doi.org/10.1093/nar/10.8.2709.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., et al. (2019). The Pfam protein families database in 2019. Nucleic Acids Res. *47*, D427–D432. https://doi.org/10.1093/nar/gky995.

Elrod-Erickson, M., Benson, T.E., and Pabo, C.O. (1998). High-resolution structures of variant Zif268-DNA complexes: implications for understanding zinc finger-DNA recognition. Struct. Lond. Engl. 1993 *6*, 451–464. .

Emerson, R.O., and Thomas, J.H. (2009). Adaptive Evolution in Zinc Finger Transcription Factors. PLOS Genet. *5*, e1000325. https://doi.org/10.1371/journal.pgen.1000325.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74. https://doi.org/10.1038/nature11247.

Farlik, M., Halbritter, F., Müller, F., Choudry, F.A., Ebert, P., Klughammer, J., Farrow, S., Santoro, A., Ciaurro, V., Mathur, A., et al. (2016). DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation. Cell Stem Cell *19*, 808–822. https://doi.org/10.1016/j.stem.2016.10.019.

Fasching, L., Kapopoulou, A., Sachdeva, R., Petri, R., Jönsson, M.E., Männe, C., Turelli, P., Jern, P., Cammas, F., Trono, D., et al. (2015). TRIM28 Represses Transcription of Endogenous Retroviruses in Neural Progenitor Cells. Cell Rep. *10*, 20–28. https://doi.org/10.1016/j.celrep.2014.12.004.

Feschotte, C., and Mouchès, C. (2000). Evidence that a Family of Miniature Inverted-Repeat Transposable Elements (MITEs) from the Arabidopsis thaliana Genome Has Arisen from a pogo-like DNA Transposon. Mol. Biol. Evol. *17*, 730–737. https://doi.org/10.1093/oxfordjournals.molbev.a026351.

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

Friedli, M., and Trono, D. (2015). The Developmental Control of Transposable Elements and the Evolution of Higher Species. Annu. Rev. Cell Dev. Biol. *31*, 429–451. https://doi.org/10.1146/annurev-cellbio-100814-125514.

Frietze, S., Lan, X., Jin, V.X., and Farnham, P.J. (2010). Genomic targets of the KRAB and SCAN domain-containing zinc finger protein 263. J. Biol. Chem. *285*, 1393–1403. https://doi.org/10.1074/jbc.M109.063032.

Fueyo, R., Judd, J., Feschotte, C., and Wysocka, J. (2022). Roles of transposable elements in the regulation of mammalian transcription. Nat. Rev. Mol. Cell Biol. 1–17. https://doi.org/10.1038/s41580-022-00457-y.

Gagliardi, M., Strazzullo, M., and Matarazzo, M.R. (2018). DNMT3B Functions: Novel Insights From Human Disease. Front. Cell Dev. Biol. *6*. .

Gerdes, P., Richardson, S.R., Mager, D.L., and Faulkner, G.J. (2016). Transposable elements in the mammalian embryo: pioneers surviving through stealth and service. Genome Biol. *17*, 100. https://doi.org/10.1186/s13059-016-0965-5.

Gifford, R.J., Blomberg, J., Coffin, J.M., Fan, H., Heidmann, T., Mayer, J., Stoye, J., Tristem, M., and Johnson, W.E. (2018). Nomenclature for endogenous retrovirus (ERV) loci. Retrovirology *15*, 59. https://doi.org/10.1186/s12977-018-0442-1.

Gilbert, L.A., Larson, M.H., Morsut, L., Liu, Z., Brar, G.A., Torres, S.E., Stern-Ginossar, N., Brandman, O., Whitehead, E.H., Doudna, J.A., et al. (2013). CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. Cell *154*, 442–451. https://doi.org/10.1016/j.cell.2013.06.044.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017–1018. https://doi.org/10.1093/bioinformatics/btr064.

GTEx Consortium, Aguet, F., Brown, A.A., Castel, S.E., Davis, J.R., He, Y., Jo, B., Mohammadi, P., Park, Y., Parsana, P., et al. (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204. .

Gualdrini, F., Polletti, S., Simonatto, M., Prosperini, E., Pileri, F., and Natoli, G. (2022). H3K9 trimethylation in active chromatin restricts the usage of functional CTCF sites in SINE B2 repeats. Genes Dev. https://doi.org/10.1101/gad.349282.121.

Hackett, J.A., and Surani, M.A. (2013). DNA methylation dynamics during the mammalian life cycle. Philos. Trans. R. Soc. B Biol. Sci. *368*, 20110328. https://doi.org/10.1098/rstb.2011.0328.

Haeussler, M., Zweig, A.S., Tyner, C., Speir, M.L., Rosenbloom, K.R., Raney, B.J., Lee, C.M., Lee, B.T., Hinrichs, A.S., Gonzalez, J.N., et al. (2019). The UCSC Genome Browser database: 2019 update. Nucleic Acids Res. *47*, D853–D858. https://doi.org/10.1093/nar/gky1095.

Haggerty, C., Kretzmer, H., Riemenschneider, C., Kumar, A.S., Mattei, A.L., Bailly, N., Gottfreund, J., Giesselmann, P., Weigert, R., Brändl, B., et al. (2021). Dnmt1 has de novo activity targeted to transposable elements. Nat. Struct. Mol. Biol. *28*, 594–603. https://doi.org/10.1038/s41594-021-00603-8.

Hancks, D.C., and Kazazian, H. (2010). SVA retrotransposons: Evolution and genetic instability. Semin. Cancer Biol. *20*, 234–245. https://doi.org/10.1016/j.semcancer.2010.04.001.

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

Haring, N.L., van Bree, E.J., Jordaan, W.S., Roels, J.R.E., Sotomayor, G.C., Hey, T.M., White, F.T.G., Galland, M.D., Smidt, M.P., and Jacobs, F.M.J. (2021). ZNF91 deletion in human embryonic stem cells leads to ectopic activation of SVA retrotransposons and up-regulation of KRAB zinc finger gene clusters. Genome Res. *31*, 551–563. https://doi.org/10.1101/gr.265348.120.

Hayashi, K., and Matsui, Y. (2006). Meisetz, a novel histone tri-methyltransferase, regulates meiosis-specific epigenesis. Cell Cycle Georget. Tex *5*, 615–620. https://doi.org/10.4161/cc.5.6.2572.

Helleboid, P.-Y., Heusel, M., Duc, J., Piot, C., Thorball, C.W., Coluccio, A., Pontis, J., Imbeault, M., Turelli, P., Aebersold, R., et al. (2019). The interactome of KRAB zinc finger proteins reveals the evolutionary history of their functional diversification. EMBO J. *38*, e101220. https://doi.org/10.15252/embj.2018101220.

Hubley, R., Finn, R.D., Clements, J., Eddy, S.R., Jones, T.A., Bao, W., Smit, A.F.A., and Wheeler, T.J. (2016). The Dfam database of repetitive DNA families. Nucleic Acids Res. *44*, D81–D89. https://doi.org/10.1093/nar/gkv1272.

Huntley, S., Baggott, D.M., Hamilton, A.T., Tran-Gyamfi, M., Yang, S., Kim, J., Gordon, L., Branscomb, E., and Stubbs, L. (2006). A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors. Genome Res. *16*, 669–677. https://doi.org/10.1101/gr.4842106.

Imai, Y., Baudat, F., Taillepierre, M., Stanzione, M., Toth, A., and de Massy, B. (2017). The PRDM9 KRAB domain is required for meiosis and involved in protein interactions. Chromosoma *126*, 681–695. https://doi.org/10.1007/s00412-017-0631-z.

Imbeault, M., Helleboid, P.-Y., and Trono, D. (2017). KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. Nature *543*, 550–554. https://doi.org/10.1038/nature21683.

Iouranova, A., Grun, D., Rossy, T., Duc, J., Coudray, A., Imbeault, M., de Tribolet-Hardy, J., Turelli, P., Persat, A., and Trono, D. (2022). KRAB zinc finger protein ZNF676 controls the transcriptional influence of LTR12-related endogenous retrovirus sequences. Mob. DNA *13*, 4. https://doi.org/10.1186/s13100-021-00260-0.

Iyengar, S., and Farnham, P.J. (2011). KAP1 Protein: An Enigmatic Master Regulator of the Genome *. J. Biol. Chem. *286*, 26267–26276. https://doi.org/10.1074/jbc.R111.252569.

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

Kaemena, D.F., Yoshihara, M., Ashmore, J., Beniazza, M., Zhao, S., Bertenstam, M., Olariu, V., Katayama, S., Okita, K., Tomlinson, S.R., et al. (2022). B1 SINE-binding ZFP266 impedes reprogramming through suppression of chromatin opening mediated by pioneering factors. 2022.01.04.474927. https://doi.org/10.1101/2022.01.04.474927.

Kapusta, A., and Suh, A. (2017). Evolution of bird genomes—a transposon's-eye view. Ann. N. Y. Acad. Sci. *1389*, 164–185. https://doi.org/10.1111/nyas.13295.

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. BioRxiv 531210. https://doi.org/10.1101/531210.

Katoh, K., and Standley, D.M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. *30*, 772–780. https://doi.org/10.1093/molbev/mst010.

Kazazian, H.H., and Moran, J.V. (2017). Mobile DNA in Health and Disease. N. Engl. J. Med. *377*, 361–370. https://doi.org/10.1056/NEJMra1510092.

Keesey, M., Monger, G., Wong, Y., Groom, R., Araujo, R., Taver, S., Werning, S., and Shyamal, L. http://www.phylopic.org.

Kim, S., Cho, C.-S., Han, K., and Lee, J. (2016). Structural Variation of Alu Element and Human Disease. Genomics Inform. *14*, 70–77. https://doi.org/10.5808/GI.2016.14.3.70.

Kojima, K.K. (2018). Human transposable elements in Repbase: genomic footprints from fish to humans. Mob. DNA *9*, 2. https://doi.org/10.1186/s13100-017-0107-y.

Kumar, S., Stecher, G., Suleski, M., and Hedges, S.B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol. Biol. Evol. *34*, 1812–1819. https://doi.org/10.1093/molbev/msx116.

Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., and Weirauch, M.T. (2018). The Human Transcription Factors. Cell *172*, 650–665. https://doi.org/10.1016/j.cell.2018.01.029.

Lambert, S.A., Yang, A.W.H., Sasse, A., Cowley, G., Albu, M., Caddick, M.X., Morris, Q.D., Weirauch, M.T., and Hughes, T.R. (2019). Similarity regression predicts evolution of transcription factor sequence specificity. Nat. Genet. 1. https://doi.org/10.1038/s41588-019-0411-1.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.

Lau, N.C., Seto, A.G., Kim, J., Kuramochi-Miyagawa, S., Nakano, T., Bartel, D.P., and Kingston, R.E. (2006). Characterization of the piRNA Complex from Rat Testes. Science *313*, 363–367. https://doi.org/10.1126/science.1130164.

Lawrence, M., Carey, V., and Gentleman, R. (2022). rtracklayer: R interface to genome annotation files and the UCSC genome browser (Bioconductor version: Release (3.14)).

Lechner, M.S., Begg, G.E., Speicher, D.W., and Rauscher, F.J. (2000). Molecular Determinants for Targeting Heterochromatin Protein 1-Mediated Gene Silencing: Direct Chromoshadow Domain–KAP-1 Corepressor Interaction Is Essential. Mol. Cell. Biol. *20*, 6449–6465. https://doi.org/10.1128/MCB.20.17.6449-6465.2000.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291. https://doi.org/10.1038/nature19057.

Lennartsson, A., and Ekwall, K. (2009). Histone modification patterns and epigenetic codes. Biochim. Biophys. Acta BBA - Gen. Subj. *1790*, 863–868. https://doi.org/10.1016/j.bbagen.2008.12.006.

Lin, Y., Zheng, J., and Lin, D. (2021). PIWI-interacting RNAs in human cancer. Semin. Cancer Biol. *75*, 15–28. https://doi.org/10.1016/j.semcancer.2020.08.012.

Lovšin, N., Gubenšek, F., and Kordi, D. (2001). Evolutionary Dynamics in a Novel L2 Clade of Non-LTR Retrotransposons in Deuterostomia. Mol. Biol. Evol. *18*, 2213–2224. https://doi.org/10.1093/oxfordjournals.molbev.a003768.

Lukic, S., Nicolas, J.-C., and Levine, A.J. (2014). The diversity of zinc-finger genes on human chromosome 19 provides an evolutionary mechanism for defense against inherited endogenous retroviruses. Cell Death Differ. *21*, 381–387. https://doi.org/10.1038/cdd.2013.150.

Lupo, A., Cesaro, E., Montano, G., Zurlo, D., Izzo, P., and Costanzo, P. (2013). KRAB-Zinc Finger Proteins: A Repressor Family Displaying Multiple Biological Functions. Curr. Genomics *14*, 268–278. https://doi.org/10.2174/13892029113149990002.

Mardis, E.R. (2007). ChIP-seq: welcome to the new frontier. Nat. Methods *4*, 613–614. https://doi.org/10.1038/nmeth0807-613.

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

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122. https://doi.org/10.1186/s13059-016-0974-4.

Moore, L.D., Le, T., and Fan, G. (2013). DNA Methylation and Its Basic Function. Neuropsychopharmacology *38*, 23–38. https://doi.org/10.1038/npp.2012.112.

Morange, M. (2009). The Central Dogma of molecular biology. Resonance *14*, 236–247. https://doi.org/10.1007/s12045-009-0024-6.

Nan, X., Ng, H.-H., Johnson, C.A., Laherty, C.D., Turner, B.M., Eisenman, R.N., and Bird, A. (1998). Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature *393*, 386–389. https://doi.org/10.1038/30764.

Nielsen, A.L., Ortiz, J.A., You, J., Oulad-Abdelghani, M., Khechumian, R., Gansmuller, A., Chambon, P., and Losson, R. (1999). Interaction with members of the heterochromatin protein 1 (HP1) family and histone deacetylation are differentially involved in transcriptional silencing by members of the TIF1 family. EMBO J. *18*, 6385–6395. https://doi.org/10.1093/emboj/18.22.6385.

Nowick, K., Hamilton, A.T., Zhang, H., and Stubbs, L. (2010). Rapid sequence and expression divergence suggest selection for novel function in primate-specific KRAB-ZNF genes. Mol. Biol. Evol. *27*, 2606–2617. https://doi.org/10.1093/molbev/msq157.

Okumura, K., Sakaguchi, G., Naito, K., Tamura, T., and Igarashi, H. (1997). HUB1, a novel Krüppel type zinc finger protein, represses the human T cell leukemia virus type I long terminal repeat-mediated expression. Nucleic Acids Res. *25*, 5025–5032. https://doi.org/10.1093/nar/25.24.5025.

Ozata, D.M., Gainetdinov, I., Zoch, A., O'Carroll, D., and Zamore, P.D. (2019). PIWI-interacting RNAs: small RNAs with big functions. Nat. Rev. Genet. *20*, 89–108. https://doi.org/10.1038/s41576-018-0073-3.

Park, P.J. (2009). ChIP–seq: advantages and challenges of a maturing technology. Nat. Rev. Genet. *10*, 669–680. https://doi.org/10.1038/nrg2641.

Petri, R., Brattås, P.L., Sharma, Y., Jönsson, M.E., Pircs, K., Bengzon, J., and Jakobsson, J. (2019). LINE-2 transposable elements are a source of functional human microRNAs and target sites. PLOS Genet. *15*, e1008036. https://doi.org/10.1371/journal.pgen.1008036.

Piovesan, A., Antonaros, F., Vitale, L., Strippoli, P., Pelleri, M.C., and Caracausi, M. (2019). Human protein-coding genes and gene feature statistics in 2019. BMC Res. Notes *12*, 315. https://doi.org/10.1186/s13104-019-4343-8.

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

Potter, S.C., Luciani, A., Eddy, S.R., Park, Y., Lopez, R., and Finn, R.D. (2018). HMMER web server: 2018 update. Nucleic Acids Res. *46*, W200–W204. https://doi.org/10.1093/nar/gky448.

Powers, N.R., Parvanov, E.D., Baker, C.L., Walker, M., Petkov, P.M., and Paigen, K. (2016). The Meiotic Recombination Activator PRDM9 Trimethylates Both H3K36 and H3K4 at Recombination Hotspots In Vivo. PLoS Genet. *12*, e1006146. https://doi.org/10.1371/journal.pgen.1006146.

Quenneville, S., Verde, G., Corsinotti, A., Kapopoulou, A., Jakobsson, J., Offner, S., Baglivo, I., Pedone, P.V., Grimaldi, G., Riccio, A., et al. (2011). In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions. Mol. Cell *44*, 361–372. https://doi.org/10.1016/j.molcel.2011.08.032.

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-

sequencing data analysis. Nucleic Acids Res. *44*, W160–W165. https://doi.org/10.1093/nar/gkw257.

Randolph, K., Hyder, U., and D'Orso, I. (2022). KAP1/TRIM28: Transcriptional Activator and/or Repressor of Viral and Cellular Programs? Front. Cell. Infect. Microbiol. *12*, 834636. https://doi.org/10.3389/fcimb.2022.834636.

Ryan, F.P. (2004). Human Endogenous Retroviruses in Health and Disease: A Symbiotic Perspective. J. R. Soc. Med. *97*, 560–565. https://doi.org/10.1177/014107680409701202.

Ryan, R.F., Schultz, D.C., Ayyanathan, K., Singh, P.B., Friedman, J.R., Fredericks, W.J., and Rauscher, F.J. (1999). KAP-1 corepressor protein interacts and colocalizes with heterochromatic and euchromatic HP1 proteins: a potential role for Krüppel-associated box-zinc finger proteins in heterochromatin-mediated gene silencing. Mol. Cell. Biol. *19*, 4366–4378. https://doi.org/10.1128/MCB.19.6.4366.

Schmitges, F.W., Radovani, E., Najafabadi, H.S., Barazandeh, M., Campitelli, L.F., Yin, Y., Jolma, A., Zhong, G., Guo, H., Kanagalingam, T., et al. (2016). Multiparameter functional diversity of human C2H2 zinc finger proteins. Genome Res. *26*, 1742–1752. https://doi.org/10.1101/gr.209643.116.

Schneider, T.D., and Stephens, R.M. (1990). Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. *18*, 6097–6100. https://doi.org/10.1093/nar/18.20.6097.

Schultz, D.C., Friedman, J.R., and Rauscher, F.J. (2001). Targeting histone deacetylase complexes via KRAB-zinc finger proteins: the PHD and bromodomains of KAP-1 form a cooperative unit that recruits a novel isoform of the Mi-2α subunit of NuRD. Genes Dev. *15*, 428–443. https://doi.org/10.1101/gad.869501.

Schultz, D.C., Ayyanathan, K., Negorev, D., Maul, G.G., and Rauscher, F.J. (2002). SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. Genes Dev. *16*, 919–932. https://doi.org/10.1101/gad.973302.

Schumacher, C., Wang, H., Honer, C., Ding, W., Koehn, J., Lawrence, Q., Coulis, C.M., Wang, L.L., Ballinger, D., Bowen, B.R., et al. (2000). The SCAN Domain Mediates Selective Oligomerization*. J. Biol. Chem. *275*, 17173–17179. https://doi.org/10.1074/jbc.M000119200.

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

Sexton, C.E., and Han, M.V. (2019). Paired-end mappability of transposable elements in the human genome. Mob. DNA *10*, 29. https://doi.org/10.1186/s13100-019-0172-5.

Shen, P., Xu, A., Hou, Y., Wang, H., Gao, C., He, F., and Yang, D. (2021). Conserved paradoxical relationships among the evolutionary, structural and expressional features of KRAB zinc-finger proteins reveal their special functional characteristics. BMC Mol. Cell Biol. *22*, 7. https://doi.org/10.1186/s12860-021-00346-w.

Singal, R., and Ginder, G.D. (1999). DNA Methylation. Blood *93*, 4059–4070. https://doi.org/10.1182/blood.V93.12.4059.

Stormo, G.D. (1990). [13] Consensus patterns in DNA. In Methods in Enzymology, (Academic Press), pp. 211–221.

Struhl, K. (1999). Fundamentally Different Logic of Gene Regulation in Eukaryotes and Prokaryotes. Cell *98*, 1–4. https://doi.org/10.1016/S0092-8674(00)80599-1.

Stubbs, L., Sun, Y., and Caetano-Anolles, D. (2011). Function and Evolution of C2H2 Zinc Finger Arrays. Subcell. Biochem. *52*, 75–94. https://doi.org/10.1007/978-90-481-9069-0_4.

Sun, Y., Keown, J.R., Black, M.M., Raclot, C., Demarais, N., Trono, D., Turelli, P., and Goldstone, D.C. (2019). A Dissection of Oligomerization by the TRIM28 Tripartite Motif and the Interaction with Members of the Krab-ZFP Family. J. Mol. Biol. *431*, 2511–2527. https://doi.org/10.1016/j.jmb.2019.05.002.

Takahashi, N., Coluccio, A., Thorball, C.W., Planet, E., Shi, H., Offner, S., Turelli, P., Imbeault, M., Ferguson-Smith, A.C., and Trono, D. (2019). ZNF445 is a primary regulator of genomic imprinting. Genes Dev. *33*, 49–54. https://doi.org/10.1101/gad.320069.118.

Talbert, P.B., and Henikoff, S. (2021). The Yin and Yang of Histone Marks in Transcription. Annu. Rev. Genomics Hum. Genet. *22*, 147–170. https://doi.org/10.1146/annurev-genom-120220-085159.

Thomas, J.H., and Schneider, S. (2011). Coevolution of retroelements and tandem zinc finger genes. Genome Res. *21*, 1800–1812. https://doi.org/10.1101/gr.121749.111.

Turelli, P., Playfoot, C., Grun, D., Raclot, C., Pontis, J., Coudray, A., Thorball, C., Duc, J., Pankevich, E.V., Deplancke, B., et al. (2020). Primate-restricted KRAB zinc finger proteins and target retrotransposons control gene expression in human neurons. Sci. Adv. *6*, eaba3200. https://doi.org/10.1126/sciadv.aba3200.

Tycko, J., DelRosso, N., Hess, G.T., Aradhana, Banerjee, A., Mukund, A., Van, M.V., Ego, B.K., Yao, D., Spees, K., et al. (2020). High-Throughput Discovery and Characterization of Human Transcriptional Effectors. Cell *183*, 2020-2035.e16. https://doi.org/10.1016/j.cell.2020.11.024.

Urrutia, R. (2003). KRAB-containing zinc-finger repressor proteins. Genome Biol. *4*, 231. https://doi.org/10.1186/gb-2003-4-10-231.

Venkataraman, A., Yang, K., Irizarry, J., Mackiewicz, M., Mita, P., Kuang, Z., Xue, L., Ghosh, D., Liu, S., Ramos, P., et al. (2018). A toolbox of immunoprecipitation-grade monoclonal antibodies to human transcription factors. Nat. Methods *15*, 330–338. https://doi.org/10.1038/nmeth.4632.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science *291*, 1304–1351. https://doi.org/10.1126/science.1058040.

Wagner, S., Hess, M.A., Ormonde-Hanson, P., Malandro, J., Hu, H., Chen, M., Kehrer, R., Frodsham, M., Schumacher, C., Beluch, M., et al. (2000). A broad role for the zinc finger protein ZNF202 in human lipid metabolism. J. Biol. Chem. *275*, 15685–15690. https://doi.org/10.1074/jbc.M910152199.

Ward, J.H. (1963). Hierarchical Grouping to Optimize an Objective Function. J. Am. Stat. Assoc. *58*, 236–244. https://doi.org/10.1080/01621459.1963.10500845.

Warren, I.A., Naville, M., Chalopin, D., Levin, P., Berger, C.S., Galiana, D., and Volff, J.-N. (2015). Evolutionary impact of transposable elements on genomic diversity and lineage-specific innovation in vertebrates. Chromosome Res. *23*, 505–531. https://doi.org/10.1007/s10577-015-9493-5.

Warren, W.C., Hillier, L.W., Marshall Graves, J.A., Birney, E., Ponting, C.P., Grützner, F., Belov, K., Miller, W., Clarke, L., Chinwalla, A.T., et al. (2008). Genome analysis of the platypus reveals unique signatures of evolution. Nature *453*, 175–183. https://doi.org/10.1038/nature06936.

Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. *45*, 1113–1120. https://doi.org/10.1038/ng.2764.

Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell *158*, 1431–1443. https://doi.org/10.1016/j.cell.2014.08.009.

Williams, A.J., Khachigian, L.M., Shows, T., and Collins, T. (1995). Isolation and characterization of a novel zinc-finger protein with transcription repressor activity. J. Biol. Chem. *270*, 22143–22152. https://doi.org/10.1074/jbc.270.38.22143.

Witherspoon, D.J., Watkins, W.S., Zhang, Y., Xing, J., Tolpinrud, W.L., Hedges, D.J., Batzer, M.A., and Jorde, L.B. (2009). Alu repeats increase local recombination rates. BMC Genomics *10*, 530. https://doi.org/10.1186/1471-2164-10-530.

Wolf, G., Yang, P., Füchtbauer, A.C., Füchtbauer, E.-M., Silva, A.M., Park, C., Wu, W., Nielsen, A.L., Pedersen, F.S., and Macfarlan, T.S. (2015a). The KRAB zinc finger protein ZFP809 is required to initiate epigenetic silencing of endogenous retroviruses. Genes Dev. *29*, 538–554. https://doi.org/10.1101/gad.252767.114.

Wolf, G., Greenberg, D., and Macfarlan, T.S. (2015b). Spotting the enemy within: Targeted silencing of foreign DNA in mammalian genomes by the Krüppel-associated box zinc finger protein family. Mob. DNA *6*, 17. https://doi.org/10.1186/s13100-015-0050-8.

Wolf, G., Iaco, A. de, Sun, M.-A., Bruno, M., Tinkham, M., Hoang, D., Mitra, A., Ralls, S., Trono, D., and Macfarlan, T.S. (2020). Non-essential function of KRAB zinc finger gene clusters in retrotransposon suppression. 2020.01.17.910679. https://doi.org/10.1101/2020.01.17.910679.

Wolfe, S.A., Grant, R.A., Elrod-Erickson, M., and Pabo, C.O. (2001). Beyond the “Recognition Code”: Structures of Two Cys2His2 Zinc Finger/TATA Box Complexes. Structure *9*, 717–723. https://doi.org/10.1016/S0969-2126(01)00632-3.

Yan, J., Enge, M., Whitington, T., Dave, K., Liu, J., Sur, I., Schmierer, B., Jolma, A., Kivioja, T., Taipale, M., et al. (2013). Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. Cell *154*, 801–813. https://doi.org/10.1016/j.cell.2013.07.034.

Yang, P., Wang, Y., Hoang, D., Tinkham, M., Patel, A., Sun, M.-A., Wolf, G., Baker, M., Chien, H.-C., Lai, K.-Y.N., et al. (2017a). A placental growth factor is silenced in mouse embryos by the zinc finger protein ZFP568. Science *356*, 757–759. https://doi.org/10.1126/science.aah6895.

Yang, P., Wang, Y., Hoang, D., Tinkham, M., Patel, A., Sun, M.-A., Wolf, G., Baker, M., Chien, H.-C., Lai, K.-Y.N., et al. (2017b). A placental growth factor is silenced in mouse embryos by the zinc finger protein ZFP568. Science *356*, 757–759. https://doi.org/10.1126/science.aah6895.

Zeng, Y., Wang, W., Ma, J., Wang, X., Guo, M., and Li, W. (2012). Knockdown of ZNF268, which Is Transcriptionally Downregulated by GATA-1, Promotes Proliferation of K562 Cells. PLOS ONE *7*, e29518. https://doi.org/10.1371/journal.pone.0029518.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137. https://doi.org/10.1186/gb-2008-9-9-r137.

Zhang, Y., Sun, Z., Jia, J., Du, T., Zhang, N., Tang, Y., Fang, Y., and Fang, D. (2021). Overview of Histone Modification. In Histone Mutations and Cancer, D. Fang, and J. Han, eds. (Singapore: Springer), pp. 1–16.

Curriculum Vitae

Jonas de Tribolet-Hardy

jonasdetribolet@gmail.com
Rue du Maupas 42b
CH-1004 Lausanne
+41 79 399 37 97
Date of Birth: 19.09.1988

Citizenship: Swiss and Swedish
Languages: German(native)
English(fluent)
French(fluent)
Swedish(bilingual)

Experience

PhD in genomics 03.2018 – 07.2022
Laboratory of Virology and Genetics, EPFL (Lausanne, CH)
Thesis title: KRAB zinc-finger proteins and their transposable element targets: between antagonism and cooperation

Research Assistant in oncology 11.2016 – 12.2017
Department of Biosystems Science and Engineering, ETHZ (Basel, CH)

EPFL Master Project in oncology 05.2014 – 02.2015
Lab of Prof. Myles Brown, Dana-Farber Cancer Institute (Boston MA, USA)
Project title: Mechanisms leading to castration resistance in prostate cancer

Intern in research and development 07.2013 – 01.2014
Evolva SA (Reinach BL, CH)

Summary of technical expertise

Programming/ Data analysis:	R, Bash, Matlab, ImageJ Macro language, Python, C++.
Genomics:	Generation and analysis of ChIP- and RNA-seq data.
Cell biology:	Mammalian and bacterial cell culture, including hESCs.
Molecular biology:	Lentivector production and transduction, techniques for manipulating and measuring gene expression.
Biochemistry:	Chromatin preparation, protein expression and purification.
Biosafety:	4 years of experience working in Biosafety level 2 conditions.

Education

2009-15 B.Sc & M.Sc. in Life Science and Technologies at EPFL (Lausanne, CH)
Orientation: Molecular Medicine.
2007 Pre-university education in Basel-Stadt(CH)

Scientific Publications

ARv7 Represses Tumor-Suppressor Genes in Castration-Resistant Prostate Cancer
Cato L*, **de Tribolet-Hardy J***, *et al.* **Cancer Cell**. 2019 Feb. PMID: 30773341. * These authors contributed equally

Additional Information

Scientific Publications (cont.)

KRAB zinc finger protein ZNF676 controls the transcriptional influence of LTR12-related endogenous retrovirus sequences. Alexandra Iouranova, Delphine Grun, Tamara Rossy, Julien Duc, Alexandre Coudray, Michael Imbeault, **Jonas de Tribolet-Hardy**, Priscilla Turelli, Alexandre Persat, Didier Trono. **Mobile DNA**. 2022 Jan 18;13 (1): 4

TransCONFIRM: Identification of a Genetic Signature of Response to Fulvestrant in Advanced Hormone Receptor-Positive Breast Cancer.
Jeselsohn R, Barry WT, Migliaccio I, Biagioni C, Zhao J, **De Tribolet-Hardy J**, *et al.*
Clin Cancer Res. 2016 Dec 1;22(23):5755-5764

TRIM24 Is an Oncogenic Transcriptional Activator in Prostate Cancer.
Groner AC, Cato L, **de Tribolet-Hardy J**, Bernasocchi T, Janouskova H, Melchers D, *et al.* **Cancer Cell**. 2016 Jun 13;29(6):846-58

Awards

2020 Best poster: Joint EPFL-UNIL PhD Retreat
2015 Best master project: EPFL department of molecular medicine

Other interests

philosophy, painting, digital art, weightlifting, cycling