EPFL

Thèse n° 10 774

# Meeting our Makers: Uncovering the *cis*-regulatory activity of transposable elements using statistical learning

Présentée le 19 janvier 2024

Faculté des sciences de la vie
Laboratoire de virologie et génétique
Programme doctoral en biologie computationnelle et quantitative

pour l'obtention du grade de Docteur ès Sciences

par

**Cyril David Son-Tuyên PULVER**

Acceptée sur proposition du jury

Prof. M. Brbic, présidente du jury
Prof. D. Trono, directeur de thèse
Dr Ö. Deniz, rapporteuse
Dr G. Cristofari, rapporteur
Prof. G. La Manno, rapporteur

École polytechnique fédérale de Lausanne

2024

L'animation du corps n'est pas l'assemblage l'une contre l'autre de ses parties - ni d'ailleurs la descente dans l'automate d'un esprit venu d'ailleurs, ce qui supposerait encore que le corps lui-même est sans dedans et sans « soi ».
Un corps humain est là quand, entre voyant et visible, entre touchant et touché, entre un oeil et l'autre, entre la main et la main se fait une sorte de recroisement, quand s'allume l'étincelle du sentant-sensible, quand prend ce feu qui ne cessera pas de brûler, jusqu'à ce que tel accident du corps défasse ce que nul accident n'aurait suffi à faire…

— Maurice Merleau-Ponty

A text or statement can thus be read as "containing" or "being about a [scientific] fact" when readers are sufficiently convinced that there is no debate about it and the processes of literary inscription are forgotten.
Conversely, one way of undercutting the "facticity" of a statement is by drawing attention to the (mere) processes of literary inscription which make the fact possible.

— Bruno Latour & Steve Woolgar

To Jonas, my ever-loving and supportive brother. Stay awesome.

# Acknowledgements

I would like to express my deep gratitude to my thesis supervisor Prof Didier Trono for encouraging me to work in this enthralling field at the crossroads of evolution and development as well as for welcoming genuinely interdisciplinary research in his lab. His infectious enthusiasm for scientific research and inexhaustible ability to generate ideas have time and again lifted me up in periods of doubt and creative drought. I am also appreciative of his unrelenting efforts in securing the funding that has supported this research.

I am equally indebted to Dr Julien Pontis, who patiently initiated me to the intricated regulatory models he and Didier had been developing over the preceding years, provided me with high-quality data and proved himself a precious colleague and friend.

The third person whithout whom none of this work would have seen the light of day is Dr Raphaël de Fondeville, whose steadfast ability to grasp concepts from fields he had not been in contact with since high-school never ceased to amaze me. Whether this or his empathetic support contributed the most, I do not know. Both were essential to see me through the finish line.

I am grateful to Dr Özgen Deniz, Dr Gael Cristofari, Prof Gioele La Manno and Prof Maria Brbic for accepting to serve on my committee. This entailed engaging with the present manuscript, providing constructive criticism, starting lively discussions and for some, traveling far away from their home.

I would like to thank Prof Felix Naef, Prof Guillaume Bourque and Prof David Suter for having served as experts on my first year candidacy exam.

I am indebted to Prof Michele De Palma - my doctoral program mentor - as well as Prof Patrick Barth and Prof Matteo Dal Peraro - both of whom served as presidents for my doctoral school - for their precious support.

My back-and-forth journey to Basel would not have unfolded that smoothly without the dedicated work of Séverine Reynard, who somehow always found ways to arrange all things administrative.

Completing this work would have proven impossibly tedious without the bioinformatics ground work heralded by Julien Duc (the `vim` master) and Evarist Planet at the Batcave. Together with Delphine Grun and Shaoline Sheppard, they expedited numerous thankless tasks with diligence and insight.

Clean data only comes to us through the expert hands of those conducting wet lab work, of which experiments are only the tip of the iceberg. I would like to thank Sandra Offner and Charlène Raclot for contributing to generating much of the experimental data analyzed in this thesis.

I extend my gratitude to past and present members of the Trono lab, particularly Dr Alexandre Coudray, my seemingly un-stessable office mate and climbing partner. We did send red-difficulty before graduation, and you even did it twice (or was it thrice?).

I will keep fond memories of the relief I found in discussing the more trying aspects of this academic journey with Danica Milovanovic, Dr Christina Ernst and Dr Romain Forey, all of whom - together with (no specific order) Dr Laia Simo Riudalbas, Dr Olga Rosspopoff, Dr Filipe Martins, Dr Christopher Playfoot, Dr Jonas de Tribolet-Hardy, Dr Pierre-Yves Helleboid, Dr Alexandra Iouranova, Dr Martina Begnis, Dr Priscilla Turelli, Dr Myriam Lamrayah and Dr Wayo Matsushima impressed me with their fearless conduct of (wet lab!) research and always had kind words as well as valuable advice to offer. The same applies to Dr Eunji Shin and Dr Joana Carlevaro, though their expertise was all things bioinformatics.

I would like to thank Dr Sina Nassiri for putting his trust in my scientific abilities at a time when I had little left, as he invited me to join him at Roche for an internship. This breath of fresh air was another instrumental turning point towards completing this thesis.

I would like to thank my medical team: PJ, AR, TS and DK, who I like to see as the mental

health equivalents of sports medicine physicians. Good and most importantly pleasurable intellectual work seldom emerges out of dark times.

Reaching the end, and more importantly doing so in a state of happiness, has only been possible thanks to my ever-resourceful support network. To my friends and family, I truly owe you for the completion of this work and so much more. Many of you would easily fit in several of the following groups, so please forgive me for satisfying my quasi-neurotic scientific classification impulses.

To Hannah, Ben and all my AAB friends, thank you for the discussions, breaks, lunches, aperos, dinners and nights out. Coming to the lab was all the more enjoyable knowing you would be around for the day.

Tom, Nina, Faustine and Luca; thank you for sharing your passion for climbing with me. There really is no such thing as real rock.

Koris and Basile; thank you for dedicating so much of your time to making music and playing records by my side. I did not expect my PhD to be the time I would perform at thermal baths.

Shout out to Pedro and Nelson from Plan B (ex-Pavillon), as well as TT, Fab, Ramon, Manolo, Nuno and Dora at Folklor for hosting me countless times in their DJ booths, and sponsoring me with top-class partying despite the PhD salary. Franco Scalisi: our last guitar lessons together were, plain and simple, therapy.

To the Tuesday is Danceday crew, and all of those I've connected with through dancing (a til-then surprisingly neglected aspect of my life): Jade, Mélina, Hervé, Léna, Julien, Alban, Guillaume, Nadine, Bhargav, Alessio, Caro, Clément, Lara; thank you.

Pietro, Nicola and Chiara; thank you for being family and never failing to invite me on January 1st.

To my high-school and EPFL mates Alain, Raphaël, Julien and Guillaume: I liked it so much I decided to leave last.

To Pauline and Margaux: thank you for these unforgettable Corsican Summers.

To Romaine, Icíar and Damien: thank you for being lifelong friends.

To all those who - someday, someplace - have had kind words for me: Félix G, Luca T, Estelle, Aureliano, Orestis, Marie, Jeremy, Félix D, Raphaël Sa, Clara, Minh Trực, Minh Đức, Jeannette

and those whom I have inevitably forgotten: thank you. I hope to return the favor.

Aude, thank you for having changed my life.

To my brother Jonas, my father Dominique and my mother Minh Huệ : I love you. Thank you for everything.

*Lausanne, 13 October 2023. And it was a Friday.* C. P.

# Abstract

The adaptation of organisms to their environment depends on the innovative potential inherent to genetic variation. In complex organisms such as mammals, processes like development and immunity require tight gene regulation. Complex forms emerge more often as a result of changes in gene regulation rather than gene products. Recent evidence accumulates to suggest that after spreading, transposable elements may become co-opted as *cis*-regulatory elements, thereby integrating neighboring genes into gene regulatory networks. Current methods for detecting *cis*-regulatory transposable elements rely on the joint exploration of multi-omics datasets, whose generation is both costly and time-consuming. We propose that modeling protein-coding gene expression (RNA-seq) as a function of the distribution of transposable elements into subfamilies as well as their respective genomic distances to promoters is sufficient to detect changes in transposable element-mediated *cis*-regulation. We leverage this model to show that far from solely affecting transcription during pre-implantation embryogenesis, evolutionarily recent transposable elements fine tune gene expression in *cis* throughout and beyond gastrulation while controlled by conserved master transcription factors. Moreover, we find that transposable elements optimally explain transcription at protein-coding gene promoters located within a 500kb range. Altogether, this work quantitatively shows that transposable elements disperse ready-for-use *cis*-acting platforms poised for integrating genes into regulatory networks controlled by conserved transcriptional and epigenetic regulators. Thus, metazoan adaptation appears to emerge from a rich genomic ecosystem whereby trans-

posable elements propagate in the gene pool in symbiosis with their cognate activators and controllers. Methods-wise, our work opens up new avenues for studying the regulatory role of transposable elements in the next-generation sequencing era.

Key words: transposable elements, transcription factors, gene regulation *cis*-regulatory elements, embryogenesis, gastrulation, endoderm, mesendoderm, germ layers, gene regulatory networks, epigenomics, transcriptomics, regulatory motif activity, RNA-seq, CRISPRi, CRISPRa, GATA6, EOMES, SOX15, LTR6, LTR5, SVA, PRIMA4-LTR, MER4A, MER4D

# Résumé

L'adaptation des organismes à leur environnement dépend du potentiel d'innovation permis par la variation génétique. Chez les organismes complexes, par exemple les mammifères, les processus vitaux comme le développement ou l'immunité requièrent une régulation précise de l'expression des gènes. Les formes complexes émergent principalement par des variations affectant la régulation des gènes plutôt que leurs produits. Des travaux récents suggèrent qu'après progagation, les éléments transposables sont parfois détournés pour servir d'éléments régulateurs en *cis*, résultant dans l'intégration de gènes alentours dans des réseaux de régulation transcriptionnelle. Les méthodes actuellement dédiées à la détection de l'activité *cis*-régulatrice des éléments transposables reposent sur l'exploration conjointe de jeux de données multi-omiques dont l'obtention peut s'avérer coûteuse. Nous proposons qu'une modélisation de l'expression des gènes codants - telle que produite par le séquençage à haut débit de l'ARN - basée d'une part sur la répartition des éléments transposables en sous-familles et d'autre part sur la distance les séparant des promoteurs permet de détecter des changements de régulation en *cis* imputables aux éléments transposables. Grâce à ce modèle, nous montrons que les éléments transposables récents sur le plan évolutif n'impactent pas seulement l'expression des gènes durant la phase pré-implantatoire du développement embryonnaire, mais aussi durant et après la gastrulation et ce sous le contrôle de facteurs de transcription conservés dits "maîtres". De plus, nous établissons que les éléments transposables expliquent des changements transcriptionnels émanant de promoteurs de gènes situés jusqu'à des distances de 500kb. Au travers d'une démarche quantitative, nous révélons

que les éléments transposables disséminent des plateformes *cis*-régulatrices prêtes à l'emploi susceptibles d'intégrer des gènes voisins dans des réseaux de régulation contrôlés par des facteurs transcriptionnels et épigénétiques conservés. Ces résultats suggèrent que l'adaptation métazoaire émerge d'un riche écosystème génomique constitué d'éléments transposables se propageant dans le pool génétique en symbiose avec leurs activateurs et répresseurs. Méthodologiquement, notre travail ouvre de nouveaux horizons pour l'étude de l'impact des éléments transposables sur la régulation à l'ère du séquençage à haut débit.

Mots clefs : éléments transposables, facteurs de transcription, régulation de l'expression des gènes, éléments *cis*-régulateurs, embryogenèse, gastrulation, endoderme, mésendoderme, feuillets embryonnaires, réseaux de régulation transcriptionnelle, épigénomique, transcriptomique, séquençage ARN à haut débit, inhibition via CRISPR, activation via CRISPR, GATA6, EOMES, SOX15, LTR6, LTR5, SVA, PRIMA4-LTR, MER4A, MER4D

# Contents

# 1 Introduction

## 1.1 The problem of adaptation

Organisms appear exquisitely designed for life in their environment (Paley, Eddy, & Knight, 1802/2006). This elementary yet remarkable observation becomes even more so upon noticing that this environment is in large part itself constituted of other organisms. In sum, organisms appear designed for life as collectives, as if ecosystems at large had been carefully planned by an omniscient and omnipotent Demiurge Engineer. Accounting for adaptation without appealing to the argument of design is one of the chief endeavours in Biology. A particularly fruitful mode of inquiry has been to identify processes unfolding at the timescale of lifespans that, when lined up end-to-end to timescales orders of magnitude greater, account as well or even better for adaptation than design (Lamarck, 1809/2011). That individual organisms spontaneously vary, that this variation is heritable, and that limited resources preclude geometric population growth mechanically lead to complex ecosystems constituted by remarkably varied and inter-dependent lifeforms (Darwin & Beer, 1859/2008; Darwin & Wallace, 1858).

Biologists, like most empirical scientists, are committed to identifying materialistic accounts of phenomena. Action at a distance, even when espousing mathematically formulated Laws of Nature with exquisite accuracy, leaves material causes to be desired. Hence, the Theory of Evolution immediately begged two questions upon formulation: that of the substrate of

heredity and that of the mechanistic principles causing it to vary.

Proteins, the molecular machines upon which Life revolves, were the first culprits in line, at a time when DNA was mostly considered a merely structural component of chromosomes (Griffith, 1928). It therefore came as a surprise that supplementing a non-virulent bacterial strain with DNA-rich and protein-depleted fractions purified from a heat-inactivated virulent strain resulted in virulence. Indeed, this strongly suggested that DNA, and not proteins, was the substrate of heredity (Avery, MacLeod, & McCarty, 1944). However, proteins kept the leading role in accounts of adaptation. The fruitfulness of the central dogma, stating that DNA directly codes for proteins thereby determining their structures and functions (Crick, 1958), cemented the assumption that variation at the level of protein-coding genes, most specifically the gradual accumulation and fixation of point mutations, underpinned adaptation (Britten & Davidson, 1971; King & Wilson, 1975).

But some recalcitrant observations resisted to that account of adaptation. The spontaneous cut-and-paste movement of genetic entities referred to as "jumping genes" was found to correlate with the color patterning of growing corn cobs (Feschotte, 2023; McClintock, 1950), suggesting far-reaching links between genome organization and development. That corn kernel coloring variegates as a function of the presence a transposon close to a particular gene hinted at two core tenets of developmental biology. First, that genetic variation needs not be gradual, but may instead arise in discrete steps whereby entire genetic loci are displaced at once. Second, that developmental patterning - and more generally ontogeny, a quasi-tautological requirement for the development of multi-cellular organisms from a single zygote - may be explained by changes in gene regulation. Lastly and perhaps most importantly, the sequences of homologous proteins between related yet phenotypically distinct species such as humans and chimps turned out to be nearly identical, casting doubt on whether variation at coding sequences of protein-coding genes could satisfactorily account for adaptation and the variety of life forms (Britten & Davidson, 1971; King & Wilson, 1975).

## 1.2 Repetitive DNA is granted a function

There are few things more attractive to biologists than explanations that dissolve multiple conundrums at once. We therefore hope that our reader will forgive us for momentarily adding to the ambient confusion by pointing to a seemingly unrelated yet highly puzzling feature of metazoan genomes: that they are littered with interspersed repetitive sequences (Britten & Davidson, 1969, 1971). Seminal DNA re-association experiments whereby DNA from two sources was radioactively labelled, melted to single strands and then re-annealed to double strands allowed for quantitatively probing into the repeat structure of metazoan genomes. Performing such experiments on DNA isolated from distantly versus intimately related species revealed that repetitive DNA was highly lineage-specific and evidently accumulated in bursts rather than gradually. Furthermore, the majority of repeats were found to be interspersed, i.e. to pepper metazoan genomes across vast distances as well as across chromosomes. This implies that most non-repeat genes are flanked by repeats, copies of which are located in the vicinity of other arbitrarily distant non-repeat genes. While one may be tempted to save the "evolution-by-gradual-mutations-at-protein-coding-genes" by dismissing the repeat content of metazoan genomes as a relatively unconsequential byproduct of the propagation of selfish germline parasites (Doolittle & Sapienza, 1980), one may alternatively attempt to find a common denominator between repetitive DNA, laws of variation, development, and speciation that mechanistically accounts for adaptation.

In two words, that common denominator is *gene regulation* (Britten & Davidson, 1969, 1971). Metazoans are - by definition - composed of differentiated cells, most of the time organized as tissues which in turn assemble into organs. Each of these cells carries the same genome, yet expresses a specialized set of proteins endowing it with a particular function. Since a cell's proteome depends on the set of protein-coding genes it expresses, it follows that some mechanism must exist to delineate that set. Moreover, as protein sets determining cell function virtually always overlap, the answer cannot be that cell type-defining sets of protein-coding genes simply co-locate proximally on the genome, i.e. in *cis*, thereby susceptible to coordinated

expression; for the sequential nature of the DNA molecule would then be at odds with the observed overlapping expression patterns. Thus, an intrinsic feature of genomes other than protein-coding gene adjacency is required to explain cell type-specific expression.

Repeats, owing to their genome-wide distribution and small sizes relative to protein-coding genes, provide just that kind of feature. Indeed, co-expression between any two protein-coding genes may in principle be mediated by the simultaneous recruitment of DNA-binding transcriptional regulators at highly similar repeats, each located proximally to one or several protein-coding genes. Generalizing that line of reasoning to any set of *n* protein-coding genes "wired" via a corresponding set of *m cis*-located repeats, one now possesses an elegant account of the coordinated expression of large sets of genes through a single to few *trans*-acting factors, i.e. factors encoded by genes arbitrarily distant from their genomic targets. Finally, since such factors must be encoded by genes, themselves susceptible to repeat-mediated *cis*-regulation, one may envision arbitrarily complex gene regulatory networks integrating cellular reactions to extrinsic cues as a coordinated transcriptional response. What role may such a system play in development and adaptation? Quoting the proponents of that remarkably insightful hypothesis:

> [...] metazoan organization arises through cellular ontogeny [i.e. cellular differentiation];
>
> ontogeny results from the operation of genetic regulatory programs;
>
> major events in metazoan evolution consist of changes in organization;
>
> thus the understanding of major events in evolution requires the examination of the origin of novel programs of gene regulation (Britten & Davidson, 1971).

A related argument can be made upon noticing that differentiated tissues differ more within organisms than homologous tissues across metazoan species do. If gene regulation can produce widely different forms from a single genome, then ascribing the subtler differences found at homologous organs across species to heritable differences in gene regulation becomes comparatively parsimonious (Carroll, 2008).

How then may gene regulation evolve within the context of *cis*-regulation? Which are the laws of variation that apply? Notwithstanding extreme structural variation events such as whole genome duplications, any combination of the following may in principle take place: (1) *de novo trans*-acting factor gene birth, e.g. through gene fusion; (2) *trans*-acting factor gene duplication; (3) *trans*-acting factor gene mutation; (4) *de novo cis*-regulatory element (CRE) birth, e.g. through maturation of a previously non-regulatory locus; (5) CRE duplication and (6) CRE mutation. Although notable exceptions exist (Imbeault, Helleboid, & Trono, 2017), comparative genomics studies indicate that (1), (2) and (3) are rare, probably owing to pleiotropic mosaïcism - i.e. that a single *trans*-acting factor often regulates many processes, each involving hundreds to thousands of *cis*-regulated target genes - such that the ensuing regulatory changes are unlikely to be tolerated (Carroll, 2008). In contrast, isolated events of type (4), (5) or (6) are much more likely to be tolerated as they are comparatively less prone to produce broad changes in gene expression and may instead generate subtler selectable phenotypic differences. The rapid rate at which repeats spread throughout genomes suggests that out of (4), (5) and (6), (5) probably contributes significantly. Indeed, cycles of rapid spread of CRE-containing repeats followed by periods of negative selection against detrimental insertions provide an attractive explanation for the fact that thousands of protein-coding genes are flanked by binding sites for only a handful of *trans*-acting factors, against the alternative view that each binding site emerged independently through gradual mutations.

## 1.3 Evidence of repeat-mediated *cis*-regulatory innovation

We now turn to empirical evidence supporting the notion that repeats fuel gene regulatory network innovation by providing raw *cis*-regulatory material upon spreading. We start by briefly reviewing how interspersed repeats spread, and then highlight the most relevant aspects for gene regulation.

Repeats cover more than half of the human genome (Hoyt et al., 2022; International Human Genome Sequencing Consortium, 2001; Nurk et al., 2022), a fraction comparable with that

of other metazoan genomes (Rosspopoff & Trono, 2023). The vast majority of interspersed repetitive sequences derive from a phylogenetically diverse class of genetic entities called transposable elements (TEs). TEs are uniquely characterized by an ability to mobilize within genomes, without leaving the host cell. This mobilization, called *transposition*, encompasses both cut-and-paste and copy-and-paste TE subfamily-specific mechanisms and is in essence parasitic, in that it requires components of the host cell machinery to complement the enzymatic activities of TE-encoded proteins. TEs are thus stretches of obligatory parasitic DNA encoding the information necessary - though not always sufficient - for their own mobility. A crucial step in transposition entails the transitory uncoupling of the integrant's genome from that of its host, either as an RNA (class I TEs, also called retrotransposons) or DNA (class II TEs, or DNA transposons) intermediate. Retrotransposons rely on reverse transcriptase to convert this RNA intermediate into integratable DNA. As a result, retrotransposons propagate in a copy-and-paste fashion, and have thereby littered metazoan genomes with vast numbers of copies in relatively short-timed retrotransposition bursts. Conversely, DNA transposons mobilize by excising from and subsequently reintegrating the genome of their host. As a result, DNA transposons "jump" in a cut-and-paste, non-replicative fashion that partly accounts for their paucity relative to retrotransposons in the human genome.

Typically, TE integrants contain DNA sequences necessary to attract the host transcriptional machinery, to code for transposition-specific proteins (class I and II transposons) and/or to generate copies of their genome as RNA intermediates (class I transposons). The vast majority of TE-derived sequences (thereafter simply referred to as "TEs") in the human genome bear mutations precluding the function and/or expression of these critical proteins. This, together with the fact that multiple layers of interconnected epigenetic and postranscriptional gatekeepers have evolved to control TE activity (J. De Tribolet-Hardy, Thorball, Forey, Planet, Duc, Coudray, et al., 2023; Deniz, Frost, & Branco, 2019; Imbeault et al., 2017), explains why human cells are not subjected to a constant transposition mayhem. Indeed, it will not have escaped our reader's attention that unchecked transposition might put metazoan genomes under considerable stress, not least because TE neo-insertions may disrupt coding sequences

and transcriptional regulation, e.g. through insertional mutagenesis (Solyom & Kazazian, 2012) or by serving as alternative promoters for oncogenes (Jang et al., 2019; Simó-Riudalbas et al., 2022), with potentially dire consequences for the host.

However, TE silencing is seldom absolute. A substantial fraction of TEs in the human genome retain the ability to recruit transcriptional regulators thereby influencing transcription locally. In such cases, TEs effectively function as *cis*-regulatory platforms for neighboring genes (Fig. 1.1) through TE-embedded reguatory sequences (TEeRS). Not all TE families are equivalent TEeRS suppliers, notably owing to differences in RNA polymerase usage. For instance, short interspersed nuclear elements (SINEs) are transcribed by RNA polymerase III - though intriguing exceptions have been reported (Horton, Kelly, Dziulko, Simpson, & Chuong, 2023) - while endogenous retroviruses (ERVs) and long interspersed nuclear elements (LINEs) are transcribed by RNA polymerase II, the enzymatic complex mediating eukaryotic protein-coding gene expression. One would thus expect long terminal repeats (LTRs) - the promoter sequences of ERVs - and the 5' untranslated regions (UTRs) of LINEs - the prototypical LINE promoters - to preferentially attract transcriptional activators. Indeed, evolutionarily young and thus functionally preserved LTRs and LINE 5' UTRs frequently form cell type-, tissue- and developmental stage-specific transcription factor (TF) binding hubs (Barakat et al., 2018; Bourque et al., 2008; Chuong, Elde, & Feschotte, 2016; Chuong, Rumi, Soares, & Baker, 2013; Ito et al., 2017; Kunarso et al., 2010; Lynch, Leclerc, May, & Wagner, 2011; Ohnuki et al., 2014; Pontis et al., 2019b; X. Sun et al., 2018; Sundaram et al., 2014; J. Wang et al., 2014) correlating with an enrichment for active chromatin features (Jacobs et al., 2014; Jacques, Jeyakani, & Bourque, 2013; Lynch et al., 2011; Pehrsson, Choudhary, Sundaram, & Wang, 2019; Pontis et al., 2019b; Sundaram et al., 2014). While the precise mechanism driving the expression of composite SINE-VNTR-Alu (SVA) TEs remains unindentified, they nonetheless recruit transcriptional activators at an LTR-derived sequence located at their 3' end (Pontis et al., 2019b).

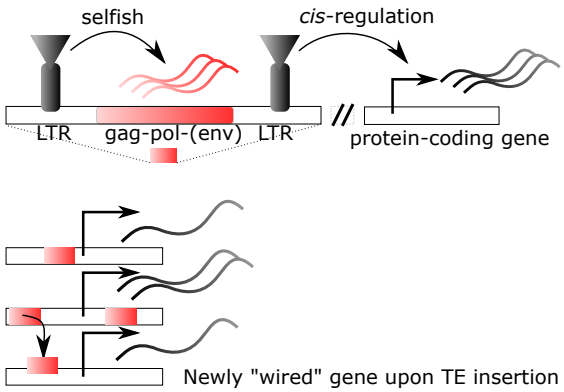



Figure 1.1 – The biochemical activity of TEs (here a class I LTR retrotransposon) first and foremost reflects their propensity for selfish propagation in the host genome. This initially selfish biochemical activity may become co-opted by the host to integrate or "rewire" a nearby gene into an existing gene regulatory networks.

## 1.4 Transposable elements affect gene regulation throughout early embryogenesis

Much of the most compelling evidence implicating TEs as CREs has been obtained in contexts related to early embryogenesis, such as embryonic stem cells (ESCs) isolated from developing embryos, embryonic carcinoma cell lines or induced pluripotent stem cells (iPSCs) reprogrammed from differentiated cells (Kunarso et al., 2010; Ohnuki et al., 2014; Pontis et al., 2019b; J. Wang et al., 2014). In fact, specific transcriptional and *cis*-regulatory TE activity patterns have been found to demarcate stages and models of early embryogenesis (Theunissen et al., 2016; J. Wang et al., 2014). Moreover, the advent of CRISPR-based epigenetic editing technologies (Gilbert et al., 2013) has allowed simultaneous perturbations affecting entire subfamilies and thousands of integrants to be tested (Fuentes, Swigut, & Wysocka, 2018b; Pontis et al., 2019b; Todd, Deniz, Taylor, & Branco, 2019). Reported impacts on gene regulation range from substantial (Fuentes et al., 2018b; Pontis et al., 2019b) to subtle (Todd et al., 2019) for TE subfamilies displaying otherwise canonical features of *cis*-regulatory activity, such as enrichment for enhancer histone marks, TF binding and frequent physical contacts with promoters. A recurring theme emerging from the study of TEs throughout early development is that the most active ones are also often strikingly recent (Chuong, Elde, & Feschotte, 2017;

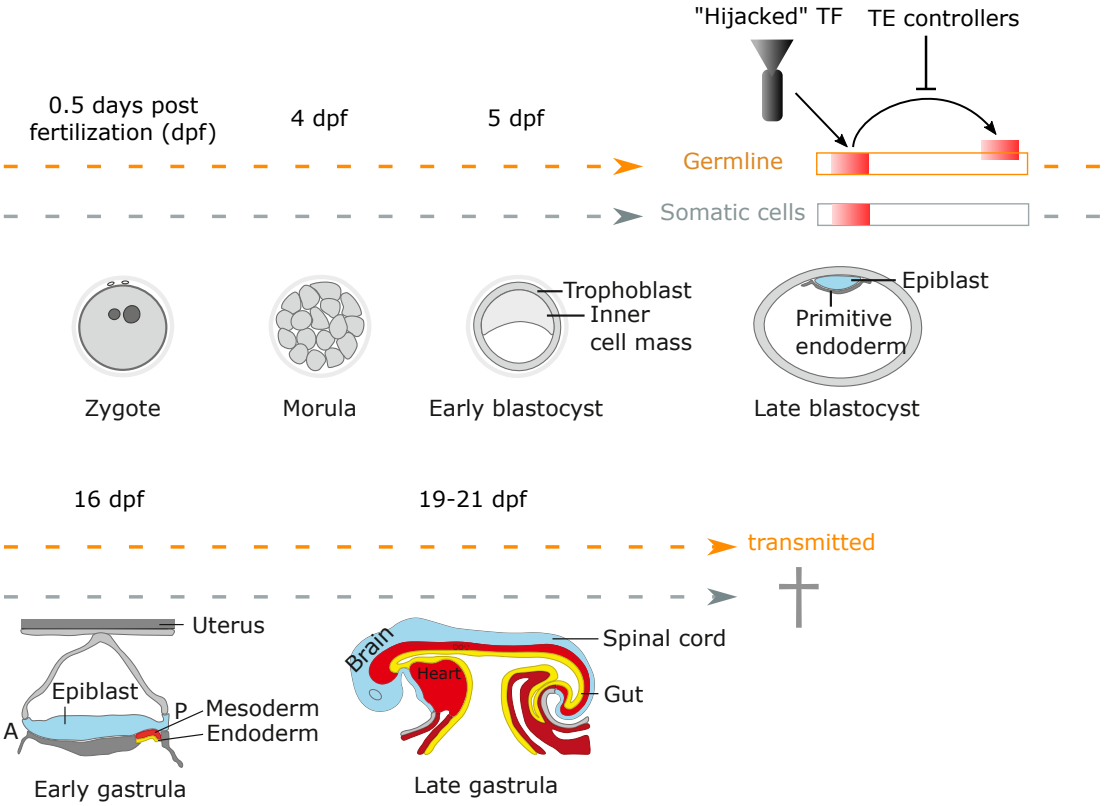



Figure 1.2 – TEs are selfish genetic parasites that "cheat" mammalian sexual reproduction to accumulate in the germline. They display particularly high activity levels in various models of early pluripotency, e.g. naive/primed hESCs, iPSCs, and are highly transcribed in cognate *bona fide* embryonic cells of pre-implantation (inner cell mass) and post-implantation (epiblast). This work further dissects the regulatoy role of TEs at these developmental stages and extends the inquiry to subsequent stages (gastrula) at which primordial germ cell committment may already have taken place. Figure adapted from (Van Den Brink & Van Oudenaarden, 2021)

Senft & Macfarlan, 2021) despite impacting expression at nearby genes through the action of conserved TFs. At face value, and assuming that the observed transcriptional and epigenetic patterns are of functional relevance for embryogenesis, these results can be interpreted as suggesting that upon spreading, TEs alter the gene regulatory networks governing development in a lineage-specific manner, thereby contributing to the evolution of forms, and thus to adaptation.

These results can however be interepreted in a completely different light. Indeed, taking a "selfish gene" view of the phenomena at hands (Chuong et al., 2017; Dawkins, 1976/2016; Doolittle & Sapienza, 1980; Williams & Dawkins, 1966/2019) - namely understanding TEs as

sequences whose sole purpose is to reach maximal frequency in the gene pool - draws one towards rather different conclusions. Under this view, and within the context of sexually-reproducing metazoans, TEs come forth as genetic parasites adapted for germline invasion, as somatic neo-insertions are lost to vertical transmission (Fig. 1.2). Under these considerations, the high activity of TEs during early embryogenesis can be explained by their selfish need for self-transcription at developmental windows permissive to germline neo-insertions, i.e. in lineages susceptible to give rise to germ cells. Accordingly, pluripotency TFs generally display elevated binding at TEs (Carter et al., 2022; Kunarso et al., 2010; Ohnuki et al., 2014; Pontis et al., 2019b; J. Wang et al., 2014). Moreover, ESCs are characterized by a paucity of repressive epigenetic marks relative to terminally differentiated cells. Thus, the pluripotency-specific erasure of epigenetic marks maintaining TEs silenced in differentiated tissues ought to increase TE activity in that context. Accordingly, metazoan genomes code for unusually rapidly evolving families of TFs whose main purpose appears to be TE control via direct binding in contexts where broad epigenetic repression is relaxed (Imbeault et al., 2017; Najafabadi et al., 2015; Rowe et al., 2010; Wells et al., 2023; Wolf et al., 2015). This "arms race" model (Jacobs et al., 2014) has in turn been reinterpreted as a "domestication model" to accomodate for the fact that many recent TE subfamilies are bound by sequence-specific transcriptional repressors predating them (J. De Tribolet-Hardy, Thorball, Forey, Planet, Duc, Coudray, et al., 2023; Imbeault et al., 2017), suggesting that TE controllers may act as tolerogenic agents for the spread of TEs, thereby cutting the fitness costs associated with TE-mediated gene regulatory network evolution (Britten & Davidson, 1969, 1971; Friedli & Trono, 2015; Imbeault et al., 2017; Pontis et al., 2019b) in a textbook case of multi-level selection (Doolittle, 2022).

## 1.5 Research questions

TEs are classified into approximately a thousand subfamilies according to mechanisms of transposition and phylogenetic considerations. Consequently, members of the same subfamily generally display high levels of sequence similarity and are thus likely to attract similar sets of transcriptional regulators in response to similar signaling cues. In simpler terms, a model

of TE-mediated GRN evolution predicts that protein-coding genes located in the vicinity of phylogenetically related TE integrants should exhibit detectable signs of co-regulation.

However, existing frameworks for studying TE-mediated *cis*-regulation usually consider protein-coding gene expression and TE-located epigenomic activity separately. Concordance is only tested in a secondary step, typically through proximity-informed enrichment tests (Lynch et al., 2011; Pontis et al., 2019b). Such approaches are hindered by several shortcomings. First, they require that multiple omics datasets (e.g. ChIP-seq, ATAC-seq and RNA-seq) be available for the context of interest, thus increasing experimental costs and complicating reanalyses of publically available data. Second, they are likely underpowered, as assessing differential gene expression and TE-mediated epigenetic activity independently unnecessarily increases the signal-to-noise ratio required for detecting TE-mediated *cis*-regulation by disregarding the genomic distance of TEs *vis-à-vis* protein-coding genes. Third, they do not provide quantitative estimates of the magnitude of TE-mediated *cis*-regulation, such as fold-changes per proximal integrants would. Fourth, TE transcription - which is routinely used in place or in addition to TE-located epigenomic activity - does not always correlate with TE-mediated *cis*-regulatory activity (Pontis et al., 2019b), and is hampered by the highly repetitive nature of TEs which biases read mapping against young subfamilies, in particular if short and single end reads are used (Sexton & Han, 2019).

We thus reasoned that if TEs truly act as collectives of CREs mediating coordinated gene expression, one should be able to mathematically model transcriptional changes at protein-coding genes as a function of their genomic distances to TEs, thereby quantifying the extant to which TEeRS contribute to variations in gene expression. The questions addressed in this work thus entail: (1) can one recover TE-mediated *cis*-regulation from transcriptional data - i.e. RNA-seq - alone via mathematical models of gene expression? (2) How do parameters such as TE evolutionary age, TE transcription, TE-protein-coding gene distance, TF binding at TEs and chromatin marks associate with TE-mediated *cis*-regulation? (3) May these parameters be of help to estimate TE-mediated *cis*-regulation at a higher resolution? (4) Under which previously unexplored contexts may TEs contribute to gene regulation? (5) Does the relevance of TE-

mediated *cis*-regulation decrease at developmental stages taking place after those modeled by ESCs, i.e. during and beyond gastrulation? (6) Which are the *trans*-acting TFs involved in TE-mediated *cis*-regulation? And finally, (7) what do the answers to these questions imply for our understanding of development, speciation and adaptation?

# 2 Primate-specific transposable elements shape transcriptional networks during human development

Julien Pontis[1*], Cyril Pulver[1], Christopher J. Playfoot[1], Evarist Planet[1], Delphine Grun[1], Sandra Offner[1], Julien Duc[1], Andrea Manfrin[2], Matthias P. Lutolf[2] and Didier Trono[1*]

Affiliations [1]Laboratory of Virology and Genetics, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland; and [2]Laboratory for Stem Cell Bioengineering, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland

*Corresponding authors: Didier Trono (didier.trono@epfl.ch) Julien Pontis (julien.pontis35@gmail.com)

Keywords: Primates, Transposable Elements, KRAB-Zinc Fingers, Gastrulation, *Cis*-regulatory elements, Genome Evolution, Single-cell Multi-Omic, Gastruloid, Embryonic Stem Cells.

Published as an article in Nature Communications (Pontis et al., 2022) https://doi.org/10.1038/s41467-022-34800-w. An early preprint version can be found at https://doi.org/10.1101/2021.08.18.456764

This work, mainly conducted by Julien Pontis, usefully lays the ground work for the next chapter of this thesis. My contribution amounts to most of the analyses underpinning sections 2.3.4, as well as some of the results presented in the following section, all of which leverage

early versions of *craTEs*, the algorithm presented in detail in the next chapter. I also provided substantial assistance for manuscript writing, including the for drafting the Introduction and Discussion sections.

## 2.1 Abstract

The human genome contains more than 4.5 million inserts derived from transposable elements (TEs), the result of recurrent waves of invasion and internal propagation throughout evolution. For new TE copies to be inherited, they must become integrated in the genome of the germline or preimplantation embryo, which requires that their source TE be expressed at these stages. Accordingly, many TEs harbor DNA binding sites for the pluripotency factors OCT4, NANOG, SOX2, and KLFs and are transiently expressed during embryonic genome activation. Here, we describe how many primate-restricted TEs have additional binding sites for lineage-specific transcription factors driving their expression during human gastrulation and later steps of fetal development. These TE integrants serve as lineage-specific enhancers fostering the transcription, amongst other targets, of KRAB-zinc finger proteins (KZFPs) of comparable evolutionary age, which in turn corral the activity of TE-embedded regulatory sequences in a similarly lineage-restricted fashion. Thus, TEs and their KZFP controllers play broad roles in shaping transcriptional networks during early human development.

## 2.2 Introduction

The human genome hosts some 4.5 million sequence inserts readily recognizable as derived from transposable elements (TEs). Most are retroelements, whether ERVs (endogenous retroviruses), LINEs (long interspersed nuclear elements), SINEs (short interspersed nuclear elements, which include primate-specific Alu repeats) or SVAs (SINE-VNTR-Alu, composites of an ERV and Alu, restricted to hominids), which replicate through a copy-and-paste mechanism with reverse transcription of an RNA intermediate followed by insertion of its DNA copy. TEs are increasingly recognized as major drivers of genome evolution owing to their

recombinogenic and regulatory potential, even though most are unable to spread further due to inactivating mutations (Chuong et al., 2017; Sundaram & Wysocka, 2020).

TEs are tightly controlled by epigenetic silencing mechanisms, yet many are expressed when these mechanisms are put on hold in the context of genome reprograming during gametogenesis or in preimplantation embryos, when transposition results in inheritable new integrants. Correspondingly, thousands of TEs, albeit mostly primate-restricted ERVs (HERV9, HERVK, HERVL, HERVH), SVAs, Alus, and young LINE-1s, display marks of open chromatin and are transcribed during embryonic genome activation (EGA) (Gao et al., 2018; Göke et al., 2015; Liu et al., 2019; Wu et al., 2018). The same subset of TEs is enriched in acetylated histone, a hallmark of enhancers, in human embryonic stem cells (hESCs) derived from the pre-implantation embryo (Kunarso et al., 2010; Pontis et al., 2019b; The FANTOM Consortium et al., 2014), where many are bound and controlled by pluripotency transcription factors (TFs) (Grow et al., 2015; Haring et al., 2021; Kunarso et al., 2010; Ohnuki et al., 2014; Pontis et al., 2019b; J. Wang et al., 2014). This broad induction of transcriptionally active TE loci likely contributes to the efficiency of EGA, and their *cis*-regulatory influences shape the gene regulatory landscape of the pre-implantation embryo. In a remarkable regulatory feedback loop, the pluripotency factor-mediated activation of primate-restricted TE-embedded enhancers leads to the expression of evolutionary contemporaneous Krüppel-associated box (KRAB)-containing zinc finger proteins (KZFPs), which act as sequence-specific repressors of these EGA-induced TEs (Greenberg & Bourc'his, 2019; Pontis et al., 2019b).

KZFP-induced heterochromatin formation and DNA methylation is often viewed as responsible for maintaining TE-embedded regulatory sequences in a repressed state at later stages of development and in adult tissues (Friedli & Trono, 2015; Greenberg & Bourc'his, 2019; H. Guo et al., 2014; Smith et al., 2014). However, it has become established that TE-embedded regulatory sequences (TEeRS) influence multiple aspects of human or mouse biology, and it is increasingly recognized that KZFPs exert profound influences on their actions(Chuong et al., 2017; Ecco et al., 2016; Friedli & Trono, 2015; Fueyo, Judd, Feschotte, & Wysocka, 2022; Playfoot et al., 2021; Turelli et al., 2020; Wolf et al., 2015; Yang, Wang, & Macfarlan, 2017). Here,

we report that many TEeRS induced during EGA are re-expressed during gastrulation and exhibit an open chromatin in fetal tissues, with a high degree of lineage specificity reflecting their activation no longer by pluripotency factors but by cell-type-restricted TFs. We also determine that many of these TE-derived enhancers are enriched near KZFP genes, the secondary stimulation of which is responsible for lineage-specific heterochromatin formation during germ layer formation. We thus conclude that evolutionary recent TEs and their KZFP controllers strongly influence and thereby confer a high degree of species specificity to the conduct of human gastrulation and fetal development.

## 2.3 Results

### 2.3.1 Cell-type-specific expression of primate-restricted TEs during human gastrulation

To examine the transcriptional state of TEs in the immediate post-implantation period, we analyzed single-cell transcriptome data from a gastrulating human embryo (Tyser et al., 2021). Gene expression patterns allowed the grouping of cells in clusters corresponding to epiblast, primitive streak, primordial germ cells (PGCs), ectoderm, to nascent, emergent, advanced, yolk sac and axial mesoderm, to endoderm and to early hematopoietic compartment (Fig. 2.2.1a, Fig. S1a). Overall, we could detect transcripts emanating from more than 100,000 TE loci, with a marked relative overrepresentation of primate-specific integrants (Fig. 2.2.1b, Fig. S1b). Many of these evolutionary recent TEs belong to the SVA, HERVK, HERVH, L1PA3, L1PA2 and L1Hs subgroups (Fig. 2.2.1c), previously found to be transcribed during EG(Göke et al., 2015; Pontis et al., 2019b). However, these EGA-induced subfamilies were more highly expressed in the primitive streak compared to the epiblast, consistent with de novo transcription, with some displaying further cell-specific patterns of expression (Fig. 2.2.1d, Fig. S1c-d). For instance, HERVH expression was broad but highest in PGC and axial mesoderm and low in hematogenic derivatives, whereas HERVK transcripts were most abundant in PGCs, endodermal cells, and blood progenitors (Fig. 2.2.1d, Fig. S1c-d). Interestingly, the evolutionary young LTR5Hs-

HERVK were expressed in both PGCs and endodermal cells, while the more ancient LTR5B-HERVK were detected only in the latter tissue (Fig. 2.2.1d, Fig. S1c-d). Moreover, HERVK11 and HERVS71/HERVK22 transcripts were abundant in the primitive streak and nascent mesoderm for the former and in definitive endoderm for the latter, but were not present in epiblast cells or previously detected during EGA (Fig. 2.2.1d, Fig. S1c-d). HERVIP10FH and HERV17 integrants were similarly de novo expressed in PGCs (Fig. 2.2.1d, Fig. S1c-d).

### 2.3.2 Evolutionarily recent TEs exert cell type-specific *cis*-regulatory influences during human development

Having documented the germ layer-specific expression of EGA-induced and several other TE subfamilies during human gastrulation, we next asked whether their stage- or lineage-restricted expression correlated with cell type-specific chromatin accessibility during human development. For this, we examined several in vivo and in vitro pre- and early post-implantation model systems. First, we re-analyzed chromatin accessibility dataset obtained from pre-implantation morula and blastocyst (Gao et al., 2018) and their respective in vitro derivatives naïve and primed human embryonic stem cells (hESC) (Pontis et al., 2019b), observing that many TEs expressed in specific cells of the gastrula exhibited some level of chromatin accessibility in the pre-implantation embryo and in hESC (Fig. 2.2a). Interestingly, most accessible subfamilies also have enhancer activity when tested in hESC with an episomal reporter assay13 (Supplementary Table 1). Next, we turned to in vitro differentiated hESC derivatives. We re-analyzed single-cell RNA-seq datasets generated from hESC-derived embryoid bodies (EBs) and chromatin accessibility studies performed in their purified primordial germ cells (PGCs) derivatives (D. Chen et al., 2019) (Fig. S2a), as well as transcriptome and chromatin data from hESC-derived endodermal cells (Lee et al., 2019). In addition, we experimentally profiled chromatin accessibility and RNA expression at the single nucleus level in gastruloid, a self-organizing elongated embryoid body recently proposed as an in vitro model to study some aspects of human gastrulation (Moris et al., 2020) (Fig. S2e-j). We found that in all these systems several TE subfamilies displayed cell-specific expression and

chromatin accessibility patterns that matched their expression in the gastrula. For instance, the chromatin of LTR5Hs and HERV17 integrants expressed in PGCs was opened in EB-derived PGCs (Fig. 2.2b, S2bd); as well, RNA levels and chromatin accessibility were matched for HERVH integrants in the axial mesoderm and in gastruloid-derived axial cells (Fig. 2.2b, S2kl); finally, the endoderm-expressed LTR5B/Hs-HERVK, LTR6B-HERVS71, and HERVH displayed an open chromatin in hESC-derived endoderm (Fig. 2.2b, S2c). Of note, in some of these cases chromatin accessibility was noted in hESCs and pre-implantation embryo, but it significantly increases in a TE-subfamily and lineage-specific fashion in the corresponding differentiated derivative.

We then turned to later developmental stages, by comparing single-cell ATAC-seq data generated from 15 organs containing 54 different cell types derived from 12- to 17-week-old human fetuses (Domcke et al., 2020) with those obtained in preimplantation embryo (Gao et al., 2018) and in vitro-differentiated hESCs (Pastor et al., 2018). This led us to three observations. First, while ERVs contribute only about 8% of the human genome through some 600,000 integrants, they accounted for a quarter of the 1 million chromatin-accessible loci detected in fetal tissues (Fig. 2.2c). Second, the most significantly accessible TE subfamilies were largely primate-restricted (Fig. 2.2d). Third, profiles derived from single-cell chromatin accessibility of endodermal fetal organs revealed differential accessibility for distinct TE subfamilies that corresponded to their germ layer-restricted expression (Fig. 2.2e).

Together, these results suggest a model whereby chromatin at evolutionarily recent TE loci is opened in the human embryo, with lineage-specific patterns of accessibility and expression in the gastrula influencing the chromatin and transcriptional landscape of later developmental stages.

### 2.3.3 Tissue-specific transcription factors control lineage-restricted TE expression during human gastrulation

While the combined expression and accessibility of many TEs in pre-implantation embryo and hESCs reflects their recognition by pluripotency factors (Fig. S3a), the patterns observed at later stages for EGA-induced TE expression suggested regulation by cell type-specific TFs. To probe this hypothesis, we examined the transcriptional changes induced at TEs by individual overexpression of 328 TFs in epiblast-derived human embryonic stem cells (hESC), available through a recent publication (Nakatake et al., 2020) (Fig. S3b,c). We found that, whereas more than 200,000 TE different loci were deregulated in the sum of all these experiments, each TF significantly induced only a restricted set of TE subfamilies (Fig. 2.3a, Fig. S3d, e). TEs previously noted to be transcribed during the minor and major waves of EGA, such as HERVL and HERVK respectively, were activated by their known cognate activators DUX4 and KLF4 (Fig. 2.3d, Fig. S3f), supporting the validity of our approach. By matching the binding profiles of 268 TFs as defined in various cellular contexts (Oki et al., 2018) with transcriptome studies performed in overexpressing hESCs (Nakatake et al., 2020), we identified 156 factors that could bind to and induce the expression of 667 TE subfamilies (Supplementary Data 1). We then determined that 92 of these TF-TE pairs were expressed in gastrulating human embryos (Fig. 2.3b). Among them, overexpression in hESC of TFs considered as markers of particular germ layers, such as Brachyury for early mesoderm, GATA6 for meso-endoderm and SOX17 for endoderm and PGCs, induced the TE subfamilies expressed in the corresponding cells of the gastrulating embryo (Fig. 2.3bcd). Furthermore, these patterns correlated with their binding specificity, with the meso-endoderm-specific GATA6 recognizing and inducing both LTR5Hs-HERVK and LTR5B-HERVK integrants, but the endoderm/PGC-specific SOX17 doing so only on LTR5Hs-HERVK, as observed in gastrulating embryos (Fig. 2.3cd, Fig. S3g).

### 2.3.4 Cell-type-specific TEeRS control gene expression during human gastrulation

Since chromatin accessibility is a known marker of *cis*-regulatory elements and reflects TF binding we asked whether the activation of TEeRS by cell-specific TFs controlled the expression of genes situated in their vicinity. To this end we developed an algorithm aimed at predicting *cis*-regulatory activity (Pulver et al., 2022). In brief, we used a linear regression model to seek a correlation between the presence of specific TE subfamily members in the proximity of deregulated genes, which we expressed as an enhancer activity prediction score (Fig. S4a). To validate this approach, we inhibited simultaneously SVA- and LTR5Hs-embedded transcriptional units by dCAS9-KRAB (CRISPRi)-mediated repression in naïve hESC. We then calculated the *cis*-regulatory activity prediction scores of TE subfamilies, and confirmed that those most significantly affected in this setting corresponded to SVAs and LTR5Hs HERVK subfamilies targeted by CRISPRi (Fig. S4b). Conversely, KLF4 overexpression in primed hESC resulted in an increased *cis*-regulatory activity prediction score for LTR5Hs-derived HERVK (Fig. S4c), supporting our previous observation that this TF binds to these units and induces their transcription and acquisition of the H3K27ac active chromatin mark (Nakatake et al., 2020; Pontis et al., 2019b).

We then applied our algorithm to the analysis of the 328 hESC-TF overexpression datasets. We proceeded to rank TE subfamilies according to their *cis*-regulatory activity prediction scores in each condition and found a strong correlation with both the percentage of up-regulated integrants and the overall level of induction of this TE subset for a given TF (Fig. 2.4a). Importantly, we also calculated the *cis*-regulatory activity prediction score of individual TE subfamilies in hESC-derived endodermal cells and found it to be significant for HERVK11, LTR5B- HERVK and HERVH (Fig. 2.4b), with our enhancer prediction scores correlating with increased H3K27ac loading, chromatin accessibility and expression of these TE integrants (Fig. 2.4c, Fig. S4d). To investigate experimentally the regulatory potential of a TEeRS located near a developmentally important gene, we targeted a putative enhancer harbored in an endodermal differentiation-specific TE (LTR6B) located upstream of the PRC2 subunit RBBP4

with CRISPRi in hESCs, and subjected these cells to an in vitro endodermal differentiation protocol (Fig. 2.4d, left panel). CRISPRi-induced repression of this LTR6B integrant resulted in RBBP4 downregulation (Fig. 2.4d, right panel). Of note, the RBBP4 gene is more highly expressed in human gastrula endoderm than in its murine counterpart, where this enhancer is absent because LTR6B is a primate-restricted ERV (Fig. S4e). To validate our observation functionally at the subfamily level, we also targeted the LTR5 consensus sequence with CRISPRi in hESC (Fig. S4f) and similarly proceeded to endodermal differentiation. This resulted in the downregulation of hundreds of genes located near the corresponding TE integrants (Fig. 2.4e). Furthermore, we applied our *cis*-regulatory prediction algorithm in this experimental setting and observed that the most significantly affected TE subfamily corresponded to CRISPRi-targeted LTR5B/Hs-derived HERVK (Fig. S4g). In addition, we were able to verify that the GATA6-binding sequence present in LTR5 conferred responsiveness to an enhancer-GFP reporter system in hESC-derived endoderm (Fig. 2.4f). Finally, by comparing gene expression in human and mouse endoderm we observed that, amongst TE subfamilies, LTR5Hs was one of the best predictor of human-specific *cis*-regulatory activity (Fig. S4h).

Together, these analyses and experimental data confirm that regulatory sequences hosted by TEs can act as species- and tissue-specific enhancers notably at play during early human development.

### 2.3.5 Primate-specific *cis*- and *trans*-regulators partner up to shape gene expression during human gastrulation

Of the hundred nearby genes downregulated upon LTR5 repression (Fig. S4i), 21 encode for KZFPs, a finding of interest since many members of this large family of DNA-binding proteins are responsible for silencing TEs through H3K9me3 deposition, histone deacetylation and DNA methylation (Ecco, Imbeault, & Trono, 2017). Because of their expansion by gene and segment duplications, many KZFP genes are grouped in clusters, notably on human chromosome 19. We observed that KZFP genes located in the same genomic cluster are often of similar evolutionary age and are generally surrounded by insertions of contemporaneous TE

subfamilies (Fig. 2.5a, Fig. S5ab). Moreover, we noted that while KZFP gene expression globally decreases during differentiation, primate-restricted clusters display coordinated upregulation in specific cells of human gastrula (Fig. S5cd). Most KZFP genes (19 genes, 17 of them primate-restricted) repressed upon LTR5 CRISPRi-mediated silencing during endodermal differentiation reside in one such cluster, which we found to be enriched in LTR5B inserts (Fig. 2.5a). Eleven of these KZFP genes, which include ZNF611, ZNF600, ZNF28, ZNF468 and ZNF320, were more expressed in human gastrula endodermal cells and during in vitro endodermal differentiation of hESCs (Fig. 2.5ab). Correspondingly, we determined that GATA6 bound directly to numerous LTR5B integrants during endodermal differentiation and induced expression of LTR5B-HERVK and nearby KZFP genes when overexpressed in hESC (Fig. 2.5b). Additionally, CRISPRa-mediated activation of LTR5 in NCCIT teratocarcinoma cells led to the induction of KZFP genes flanked by integrants belonging to this TE subset (Fig. S5e).

Interestingly, we observed strong changes in the H3K9me3 landscape of TE loci between hESC and hESC-derived endodermal cells and noted a correlation with overlapping targets of LTR5-controlled KZFP genes (Fig, 5c). We notably observed LTR5-dependent increases in H3K9me3 enrichment over the transiently expressed HERVK11 subfamily in primitive-streak/nascent mesoderm during in vitro endodermal differentiation (Fig. 2.5d). Furthermore, applying our *cis*-regulatory activity prediction algorithm in the setting of LTR5-repressed endoderm-foregut differentiation yielded the highest score for MER11-HERVK11 integrants. Accordingly, upon CRISPRi-mediated targeting of LTR5 in hESC-derived foregut we detected activation of several genes bearing MER11 inserts in their vicinity, including IQCG, DYSF and BAAT (Fig. S5f). These three genes were previously identified as co-expressed with MER11A in liver tissue (Pavlicev, Hiratsuka, Swaggart, Dunn, & Muglia, 2015), and the liver-specific BAAT, mutations of which are associated with familial hypercholanemia (Carlton et al., 2003), uses a MER11A LTR as its primary promoter (Cohen, Lock, & Mager, 2009). Most interestingly, we previously determined that the MER11A BAAT promoter is bound by the LTR5-stimulated endoderm-specific KZFPs ZNF468 and ZNF808 (Fig. 2.5f). This suggests a regulatory cascade where GATA6-mediated activation of LTR5 induces the expression of KZFPs repressing the

transcription of MER11A-controlled genes in differentiating endoderm.

More generally, our results suggest a model whereby master TFs trigger a chain reaction during gastrulation by activating primate-restricted TEs controlling KZFPs of similar evolutionary age, which in turn repress these and other TE-based *cis*-acting regulatory elements, contributing to shape the chromatin and transcriptional landscape of the different germ layers. Therefore, transcriptional networks at play [in] human development are controlled by a triangular relationship between canonical TFs, their primate-restricted TEeRS and evolutionarily related KZFPs countering their influences, the latter two endowing the regulome of human gastrulation and subsequent steps of fetal development with a high level of species-specificity.

## 2.4 Discussion

Together, these data demonstrate that transcriptional networks during human early development, while orchestrated by canonical TFs, are shaped by a partnership between primate-restricted TEeRS targeted by these activators and KZFP repressors countering their influences. As such, the regulome of human gastrulation displays a remarkable level of species-specificity reflecting the presence of both recently acquired TE integrants and KZFPs selected to tame their activity.

The broad expression of TEs such as HERVKs, HERVHs, SVAs and young LINE-1s during EGA and in the PGC lineage (Göke et al., 2015; Tang et al., 2015) is explained by their recognition by stem cell factors expressed during these periods, such as SOX2, OCT4, NANOG or KLF4/17 (Grow et al., 2015; Haring et al., 2021; Kunarso et al., 2010; Ohnuki et al., 2014; Pontis et al., 2019b; J. Wang et al., 2014). New TE integrants must seed the genome during the preimplantation period or germline formation to become inherited, hence their expression at these timepoints. Interestingly, even though HERVH has been suggested to represent a human embryonic stem cell marker (Santoni, Guerra, & Luban, 2012), we observed that HERVH integrants displayed strong accessibility and expression in other cell types during development including PGC, axial mesoderm and definitive endoderm.

Our data indicate that TEeRS regulate subsequent steps of embryonic development in part through the recruitment of TFs active in germ layer determination, such as GATA6 and SOX17. More generally, it is well established that TEs can harbor binding sites for a wide array of TFs active in differentiated tissues, and the expression of some somatic genes, for instance in the immune system, is driven by TEeRS (Bourque et al., 2008; Chuong et al., 2016; Ito et al., 2017; Sundaram et al., 2014). Additionally, primate-specific TEeRS are major contributors to *cis*-regulatory innovations in hESCs and adult liver (Jacques et al., 2013; Trizzino, Kapusta, & Brown, 2018). The *cis*-regulatory influences of these recently emerged TEeRS on human fetal development suggest that, in spite of the evolutionary constraints proposed by the hourglasses developmental model (Duboule, 1994), all stages of this process are subject to regulatory innovation, and it is likely that gastrulation is influenced by lineage-restricted TE-hosted enhancers in other species as well, as already noted for placentation (M.-a. Sun et al., 2021). Supporting this hypothesis, tissue-specific TE subfamily expression was recently observed during mouse gastrulation (He et al., 2021).

How the presence of binding sites for TFs typically active in differentiated tissues, which is predicted to promote neither the spread nor the inheritability of TEs, came to be selected through evolution is the object of much speculation (Britten & Davidson, 1971; Sundaram & Wang, 2018). However, we note here that primate-restricted ERVs are vastly overrepresented amongst TEs that harbor somatic TF binding sites, are expressed during gastrulation, and display an open chromatin state during fetal development. This is consistent with the observation that many act as enhancers or promoters in developing or differentiated organs (Pehrsson et al., 2019). As ERVs are derived from exogenous retroviruses that once replicated in somatic tissues and were largely endowed with oncogenic properties favoring their expansion and persistence, it maybe that the diversity of TF binding sites harbored by ERVs is just a consequence of their ancestry. But how was this feature maintained in evolution? Our finding that TEs activated during either EGA or gastrulation stimulate the transcription of KZFP genes, the products of which in turn repress these TEs and confer germ layer-specificity to their transcriptional influences, strongly suggests that KZFPs, rather than just the host side of an evolutionary arms

race (Jacobs et al., 2014), were instrumental in allowing for the preservation and exploitation of the broad regulatory potential of TEs in higher vertebrates.

## 2.5 Methods

### 2.5.1 hESC culture and differentiation

H9 and H1 human ESC line were maintained in mTSER plus on Matrigel and were passaged using TryplE in single cells. Endodermal differentiation was performed as in (Lee et al., 2019). Briefly, hESC were passaged at 80k cells/cm2 density in 12-well Matrigel-coated plate. When cells reached 80 – 90% confluence at 48-76h post-splitting, endodermal differentiation was initiated with 100 ng/ml of Activin A for 3 days, 5 µmol GSK-3 inhibitor (CHIR-99021) to activate the WNT pathway for the first day, and 0.5 µmol for the second day. Pancreatic differentiation was performed accordingly to a Stem Cell TechnologyTM protocol (Catalog #05120) with harvest after 3 days of endodermal differentiation followed by 3 days of foregut differentiation. Gastruloid differentiation was performed as in (Moris et al., 2020). Briefly, hESC were passaged at 40,000 cells/cm2 in mTSER replaced 48-76h post-splitting by Nutristem media with 3 µmol CHIR-99021. After 24h cells were split in single-cell passaging with TryplE, embryoid bodies were formed with 800 cells per well of low adherence 96-well plate Cell Star in E6 media supplemented with 3 µmol CHIR-99021 and 10 µmol Rock inhibitor for 18 hours, then replaced with E6 twice to be harvested at 72h.

### 2.5.2 CRISPRi experiments

sgRNA design was performed by taking the Dfam consensus of LTR5B common sequence (TTGCAGTTGAGATAAGAGGAAGG). Furthermore, sgRNA design for LTR6B (GGCTTTGGGCGTT-TATCAAT, TTGATAAACGCCCAAAGCCC, TATTACAAGGTGATAGATCC) perform to uniquely match chr1:33109510-33110065 (hg19). Specificity was predicted with the CRISPOR software v5.01 (Haeussler et al., 2016). hESCs in H9 media were transduced with dCAS9-KRAB lentiviral vector, selected, and maintained in puromycin (0.25 µg/mL) for 5 to 10 days before

differentiation experiments.

### 2.5.3 ChIP-qPCR and RT-qPCR

ChIP were performed as in (Pontis et al., 2019b) and primer use for HERVK11 were (Fw CCTTCCCATACTCGCAGTTC, Rv TGCATACAAGGACCAGCTCA) and for Neg (CCAATTTCGTGCCTCATTTT. TCAGCATGTCTCCTTTGCTG), RBBP4 (Fw ATGACCCATGCTCTGGAGTG, Rv GGACAAGTCGATGAATGCTGAAA) gene expression were normalized with ACTB (Fw CATGTACGTTGCTATCCAGGC, Rv CTCCTTAATGTCACGCACGAT). Reverse Transcriptions were performed using Thermo Scientific Maxima™ H Minus cDNA SynthesisMaster Mix ref #M1661.

### 2.5.4 ChIP-seq

Sequenced reads were aligned to the reference human genome hg19 with bowtie2 (Langmead & Salzberg, 2012). MACS 2.2.4 (Yong Zhang et al., 2008) was used for peak calling. Peaks were merged using bedtools v2.27.1 (Quinlan & Hall, 2010). FeatureCounts (Liao, Smyth, & Shi, 2014) was used to count uniquely mapped reads (MAPQ>10) on the peaks. Samtools tools v1.1 was used to convert in bam files. Library size correction was performed using the TMM method as implemented in the limma package of R, using the total number of aligned reads as size factor. All ChIP-seq binding locations from the literature were extracted from ChIP-Atlas database (Oki et al., 2018) containing datas for more than a thousand of chromatin associated proteins including more than 15'000 different datasets. Differential analysis on the uniquely mapped counts between conditions was performed with voom (Law, Chen, Shi, & Smyth, 2014). Heatmaps and profile averages were calculated using deeptools v3.5.1 (Ramírez, Dündar, Diehl, Grüning, & Manke, 2014) over 5kb windows around the peak/repeat center from bigwigs. Enrichment analysis over TE subfamilies was performed with HOMER software v4.10.4 (Heinz et al., 2010) and visualized in IGV v2.8.4 and ggplots v3.l.1 in R v4.l.2.

### 2.5.5 RNA-seq

Total RNA from cell lines was isolated with NucleoSpin™ RNA Plus kit (Machery-Nagel). cDNA was prepared with Maxima Reverse Transcriptase (Thermo Scientific). Sequencing libraries were performed with Illumina Truseq Stranded mRNA LT kit. Reads were mapped to the human (hg19) genome using Hisat2 v2.l.0 (Kim, Langmead, & Salzberg, 2015). Counts on genes and TEs were generated using featureCounts v1.6.2 (Liao et al., 2014) and only uniquely mapped reads with MAPQ >10 were kept. To avoid read assignation ambiguity between genes and TEs, a gtf file containing both, genes and TEs was provided to featureCounts. For repetitive sequences, an in-house curated version of the Repbase database was used (fragmented LTR and internal segments belonging to a single integrant were merged). Only uniquely mapped reads were used for counting on genes and repetitive sequences integrants. TEs overlapping exons or that did not have at least one sample with 3 reads were discarded from the analysis. Normalization for sequencing depth was done for both genes and TEs using the TMM method as implemented in the limma package of Bioconductor (Gentleman et al., 2004), with the counts on genes as library size. Finally, for each transgene, differential gene expression analysis was performed using Voom (Law et al., 2014) as it has been implemented in the limma package of Bioconductor v3.13 and assessing only genes (or TEs) that had 3 reads in at least one sample. A gene (or TE) was considered differentially expressed when the fold change between groups was bigger than 2 and the p-value was smaller than 0.05. A moderated t-test (as implemented in the limma package of R) was used to test significance. P-values were corrected for multiple testing using the Benjamini-Hochberg's method. For counting on TE subfamilies, we added up reads on repetitive sequences without filtering out for multi-mapped reads and added them up per subfamily.

### 2.5.6 *Cis*-regulatory activity estimation

To identify TE subfamilies exerting putative *cis*-regulatory activities directly from RNA-seq data, we modeled treatment vs control deviations in gene expression of protein coding genes as a linear combination of occurrences of nearby (within 50kb of the TSS of protein coding

genes) TE subfamily integrants (Pulver et al., 2022). We excluded TEs overlapping exons and TE subfamilies that colocalized less than 150 times near the promoters of protein coding genes. The coefficients of this linear regression problem can be interpreted as the deviations of logged gene expression values explained by the presence of one TE for each TE subfamily. We define these coefficients as the *cis*-regulator activity of TE subfamilies. Similar models were proposed to infer the activity of DNA motifs from gene expression data (Balwierz et al., 2014). To find TE subfamilies with significant *cis*-regulatory activities, we performed null significance hypothesis testing on the linear regression coefficients and accounted for multiple testing using the Benjamini Hochberg procedure.

### 2.5.7 Enhancer reporter system

We used a lentiviral vector containing a minimal promoter followed of GFP cDNA (FpG5, Addgene #69443) containing a full LTR from LTR5 subfamily amplified from a LTR5B (chr19:53226814-53227799, hg19). Then a single mutation was generated using Agilent Technologies QuikChange II XL (Cat#200522-5). H9 hESC were transduced by either of these enhancer-containing vectors followed by 3 days of endodermal differentiation and then analyze by FACS (FlowJo LCC v8.8.7).

### 2.5.8 ATAC-seq studies

Reads were aligned with bowtie2. Mitochondrial reads were removed before peak calling. Peak calling was done with MACS2 with q-value <10e-5, and using the –bampe option when PE reads.

### 2.5.9 Single-cell multi-omics

Cellranger-arc (Satpathy et al., 2019) was used to obtain counts on genes and peaks using default parameters. The hg38 reference genome provided by cellranger-arc was used. We identified five main clusters in this experimental system: TBXT-expressing axial mesoderm,

SOX2-expressing neuro-mesenchymal progenitors (NMP), and three different stages of paraxial mesoderm differentiation preferentially expressing TBX6 (somitic), or PAX3/TWIST1/SIX1 (early paraxial and pre-somitic) with a different degree of somitic HOX gene markers (Fig. 2.2c, Fig. S2b-e). Cellular clustering matched increased chromatin accessibility with the corresponding cell-type-specific TF binding sites, such as illustrated for TBXT (Fig. S2f).

### 2.5.10 Single-cell RNA-seq analyses

For the human embryo dataset, counts were obtained using cellranger (Molè et al., 2021) using a GTF of hg19 that contained both, genes and TEs. Only uniquely mapped reads on genes that were expressed in at least 1% of the samples and in a minimum 3 cells for TEs were kept. Then, for TEs not overlapping exons, counts were added up at subfamily level. Cells with less than 200 features and more than 25% of mitochondrial reads were removed. Seurat's SCTransform (Stuart et al., 2019) was used to normalize the data and correct for mitochondrial percentage and total number of reads biases. For embryoid-body time course differentiation, same method was apply except that only cells with more than 20% of mitochondrial reads were removed. UMAP were computed with Seurat's R package v4.1.0 with default parameters using the first 30 principal components as input. Of note, our and Tyler et al.'s UMAPs only present minimal differences, yet axial mesoderm and endoderm appear slightly closer when depicted with our analytical pipeline.

### 2.5.11 Single-cell ATAC-seq analyses

Single-cell ATAC-seq peaks data were downloaded from the Atlas Of Chromatin Accessibility During Development (https://atlas.brotmanbaty.org/bbi/human-chromatin-during-development/) (Domcke et al., 2020). Significance for TE family chromatin accessibility enrichment was assessed by random permutations: First, for each TE subfamily, the total number of detected peaks for a given cell type on the selected TE subfamily was computed using R GenomicRanges library v1.48.0. Then, TE subfamilies were randomly shuffled 10 times using bedtools with options –chrom and –noOverlapping and the total number of peaks for each permutation was

computed. The fold enrichment of the significant subfamilies was then plotted on a heatmap using R heatmap2 function for each cell type.

### 2.5.12 TE density profiling

TE bigwig densities were computed using python. First, TEs, were extracted from the TE database depending on their subfamily evolutionary ages in bed format and converted to bedgraph using the genomecov command of BedTools v2.27.1. Then, bedgraph signals were smoothed using a rectangular window of 10kb and written in bigwig format using the pyBigWig python library v0.3.18.

### 2.5.13 Ethics declaration

The commonly used H1 and H9 human embryonic stem cell lines were originally derived from embryos produced using in vitro fertilization for reproduction and no longer needed for this purpose, and donated through voluntary written consent for use in research. The import and use of these cells was authorized by Swiss Federal Office of Public Health after approval by the Canton of Vaud Ethics Committee (Authorization Number R-FP-S-2-0009-0000). All experiments reported in this manuscript were performed under Authorization Number R-FP-S-2-0014-0000.

### 2.5.14 Statistics & Reproducibility

No statistical method was used to predetermine sample size. No data were excluded from the analyses. The experiments were not randomized The Investigators were partly blinded to allocation during experiments and outcome assessment. All experiments contain at least 2 biologically independent replicates (for RNA-seq) and more than 3 for ChIP/RT-qPCR and FACS analysis. Fig 1d, S1d, 2a: p-value using non-parametric Wilcoxon rank sum test with Seurat algorithm. Fig 2a, c, d, S2c, S3a, S5b, Supp Data 1: p-value using Homer algorithm (annotatePeaks.pl). Fig S2c, 4a, 5a, Supp Data 1: p-value using a two-sided t.test. Fig 3ab,

S3def, 4a, S5f, Fig 4b, S4fg: p-value using two-sided t.test with correction for multiple testing using the Benjamini-Hochberg's method.

### 2.5.15 Data availability

The Single-cell multi-omics of gastruloid, RNA-seq of endodermal differentiated hESC with or without LTR5-targeting sgRNA generated have been deposited in the deposited in the Gene Expression Omnibus (GEO) database under accession number GSE181120. without restricted access.

All other genomic data of this study were extracted from the GEO database: GSE140021 for the 10X single-cell RNA-seq of hESC-derived embryoid body time course; GSE120648 for ATAC-seq of purified PGC during hESC-derived embryoid body differentiation; GSE117136 and GSE52657 for ChIP-seq and ATAC-seq from endodermal differentiation; GSE130418 for ATAC-seq of naïve and primed hESCs; Single-cell ATAC-seq from fetal organ: https://descartes.brotmanbaty.org/. E-MTAB-9388 for the SMART-seq single-cell RNA-seq of human gastrula from EBI database. DRA006296 for the overexpression dataset of TF in hESC from DDBJ database. hg19 reference genome wer used from UCSC. No restriction for dataset availability. Source data are provided as a Source Data file.

Code availability. All code used will be provided upon request and on Github.

### 2.5.16 Acknowledgments

We thank Shankar Srinivas, Antonio Scialdone and Elmir Mahammadov for providing single-cell expression raw data of the human gastrula and pseudotime data, the EPFL Genomics and the University of Lausanne Genomic Technologies facilities for help with sequencing, Alfonso Martinez Arias and Naomi Moris for discussing results obtained from their gastruloid protocol, and Alexandre Mayran for critical reading of the manuscript. This study was supported by grants from the Swiss National Science Foundation and the European Research Council (KRABnKAP, no. 268721; Transpos-X, no. 694658) to D.T.; by fellowships from the EPFL/Marie

Skłodowska-Curie Fund, the Association pour la Recherche sur le Cancer (ARC), and the Fondation Bettencourt to J.P.

### 2.5.17 Contributions

J.P. and D.T. conceived the study. J.P., D.T and C.Pu. wrote the manuscript; J.P. designed and performed all experiments with the technical help of S.O. and C.Pl; C.Pu. developed and applied the cis-regulatory activity algorithm; J.P., C.Pu, E.P., D.G and J.D. performed the bioinformatics analyses; A.M. and M.L. provided valuable advice for gastruloid experiments.

### 2.5.18 Corresponding authors

Correspondence should be addressed to Julien Pontis or Didier Trono

### 2.5.19 Competing interests

The authors declare no competing interests.

## 2.6 Figures

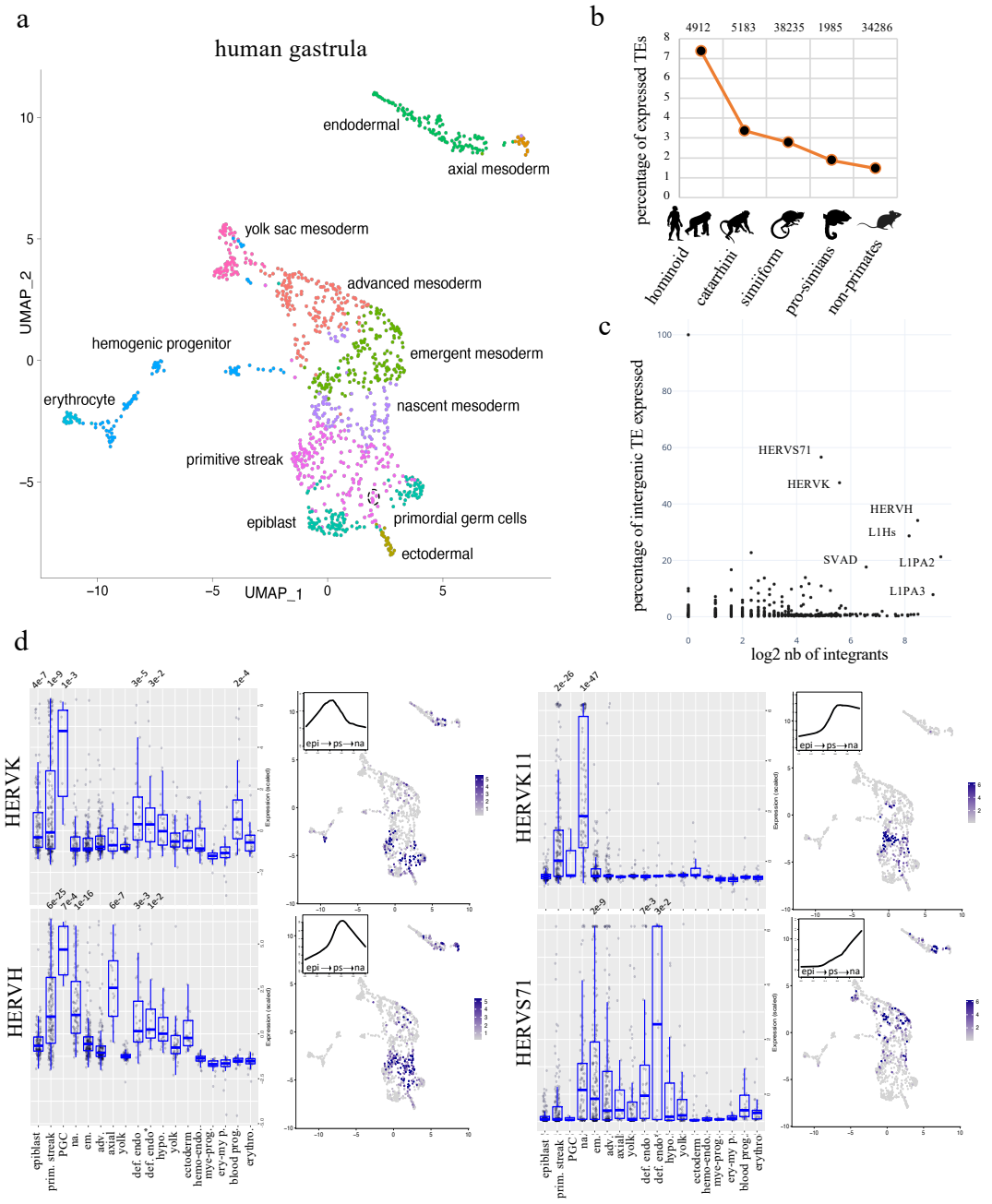



Figure 2.1 – **Cell-type specific expression of primate TEs during human gastrulation**

Figure 2.1 – **a**, Cellular composition of the human gastrula. UMAP (Uniform Manifold Approximation and Projection) based on gene expression in single cells from human gastrula; colors correspond to the different cell types identified in (Tyser et al., 2021). **b**, Age distribution of TEs expressed in human gastrula. All examined TE subfamilies (excluding DNA transposons) were assigned their evolutionary age category and percentage of expressed integrants from each age category was plotted. On top, the number of expressed integrants are indicated. **c**, Relative expression of TE subfamilies in human embryo, depicting number (x-axis) and percentage (y-axis) of integrants expressed from indicated subfamilies. **d**, Cell-type-specific expression of indicated TE subfamilies; each dot represents normalized expression of TE in one cell, grouped in boxplots corresponding to one specific cell type of human gastrula sub-clustering: epiblast (133 cells), primitive streak (prim. streak, 195 cells), primordial germ cells (PGC, 7 cells), nascent mesoderm (na. 98 cells), emergent mesoderm (em. 185 cells), advanced mesoderm (adv. 164 cells), axial mesoderm (axial, 23 cells), yolk mesoderm (yolk, 83 cells), definitive endoderm (def. endo. 35 cells), definitive endoderm non-proliferative (def. endo*, 18 cells), hypoblast (hypo, 29 cells), yolk endoderm (yolk, 53 cells), ectoderm (ectoderm 29 cells), hemogenic endothelium (hemo-endo. 37 cells), myeloid progenitor (mye-prog. 17 cells), erythro-myeloid progenitor (ery-my p. 28 cells), blood progenitor (blood prog. 29 cells), erythrocyte (erythro. 32 cells); pseudo-times are indicated in the upper left corner of the UMAPs; time points were extracted from (Tyser et al., 2021) including epibast (epi), primitive streak (ps) and nascent mesoderm (na) cells; their average expression values are indicated on the y-axis, and pseudo-times on the x-axis. Significant adjusted p-value of expressed TE subfamily in each cell type compared to all others are indicated on top of each boxplot (p-value are established using non-parametric Wilcoxon rank sum test).

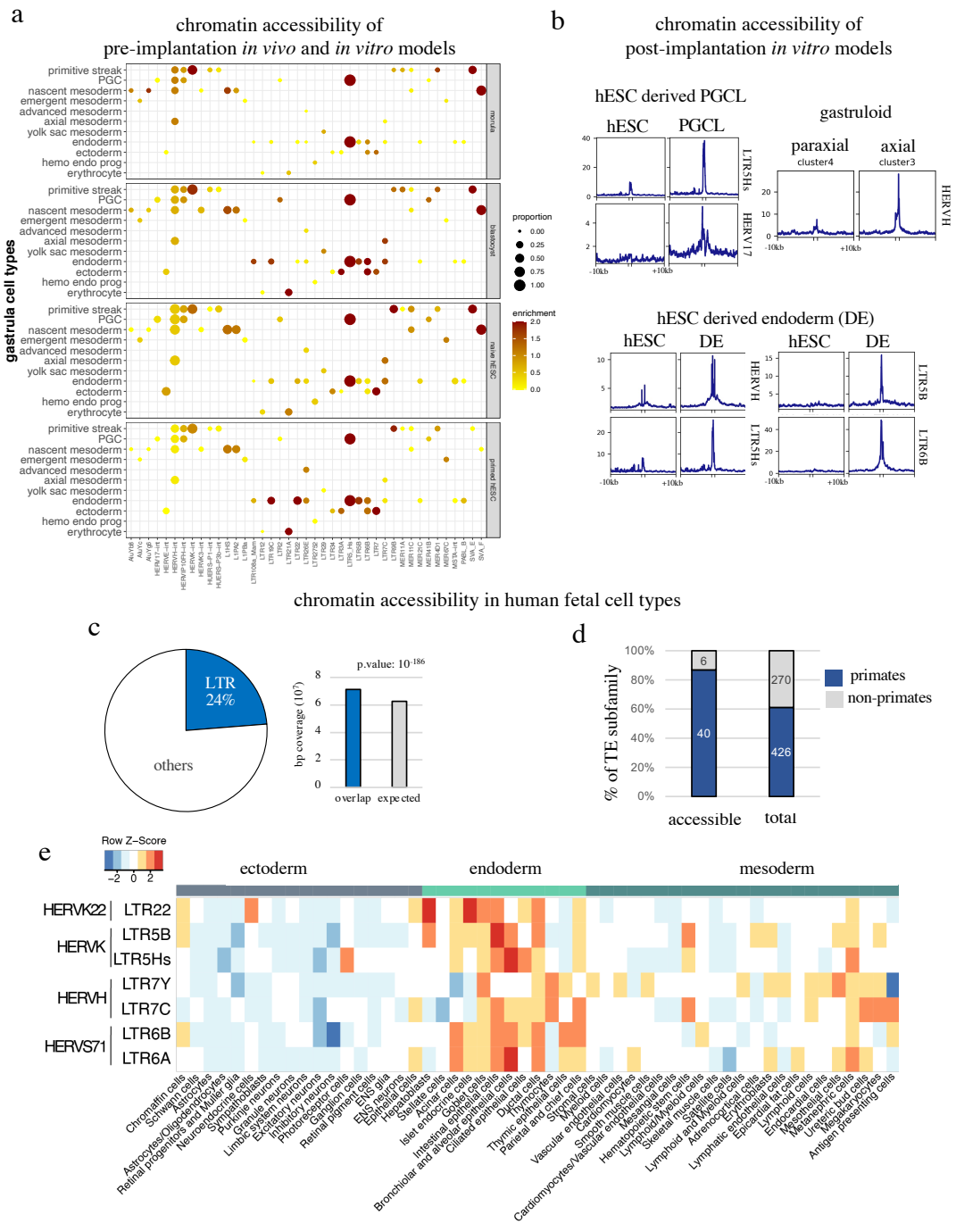



Figure 2.2 – **Evolutionarily recent TEs maintain their *cis*-regulatory potential during human gastrulation and fetal development**

Figure 2.2 – **a**, Chromatin accessibility of in vivo and in vitro preimplantation models. The x-axis represents the cell type-specific TE subfamilies in human gastrula (p.value <10e-5, p-value are established using non-parametric Wilcoxon rank sum test) that are also accessible in one of the in vivo or in vitro models (p.value <10e-3). The y-axis represents the different models, including chromatin accessibility data from morula/blastocysts (Gao et al., 2018) and primed/naïve hESCs; the size of the circles represents the number of accessibility sites overlapping with a specific TE subfamily, normalized by the number of elements in that subfamily; the color intensity represents the log enrichment relative to the random distribution of this overlap. **b**, Chromatin accessibility of post-implantation in vitro models. Line plot profiles of chromatin accessibility data at specific TE subfamily in several in vitro models; raw enrichment read average is display +/-10kb around the elements; axial cluster were analyzed from gastruloid perform in this study, EB-derived PGCL and hESC-derived endoderm (DE) and their corresponding hESC were re-analyzed from (D. Chen et al., 2019; Lee et al., 2019). **c**, Contribution of LTR TEs to chromatin accessibility at fetal stage. left panel, distribution of all chromatin accessible sites in human fetus (more than 1 million loci) and the proportion overlapping LTR; right panel, measured vs. expected contribution of LTR-derived TE (p-values are established using Homer algorithm). **d**, Relative accessibility of primate or non-primate TE integrants in human fetus re-analyzed from (Domcke et al., 2020; Gao et al., 2018; Pastor et al., 2018) indicating number of TE subfamilies with significant accessibility in at least one developmental context (>2 fold enrichment over random genome coverage, p-value <0.05 established using Homer algorithm). **e**, Enrichment over random distribution of selected TE subfamilies in indicated cell types during fetal development re-analyzed from (Domcke et al., 2020); unknown, maternal and placental cell types were removed.

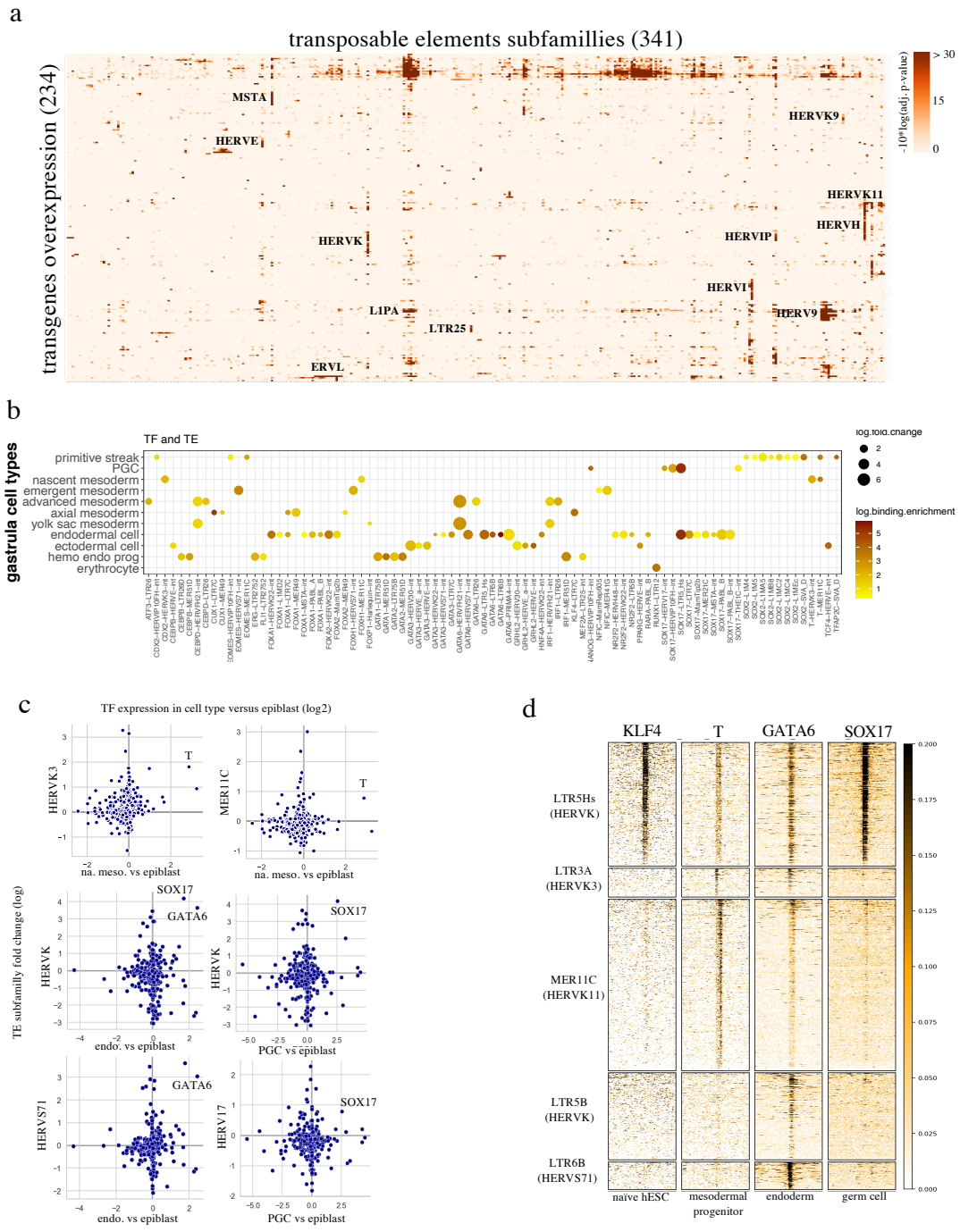



Figure 2.3 – **Evolutionary recent TEs act as cell-type-specific enhancers during human gastrulation and fetal development**

Figure 2.3 – **a**, Enrichment analysis of TE expression induced by transcription factor (TF) overexpression in hESC. Red color intensity corresponds to the enrichment (-10*log10[adjusted p-value]) of significantly up-regulated TEs over-representation for a given TE subfamily, reanalyzed from (Nakatake et al., 2020); only transgene overexpression conditions and TE subfamilies with an overrepresentation of increased expression (adjusted p-value <0.05, two-sided t.test with p.value correction for multiple testing using the Benjamini-Hochberg's method) among TEs expressed in each condition at least once are shown. *b*, Tissue-specific TEs are induced by tissue-specific TFs. Heatmap of paired tissue-specific TEs and TFs identified in the human gastrula (adjusted p-value <0.05, two-sided t.test with p.value correction for multiple testing using the Benjamini-Hochberg's method). This list was intersected with the 2,000 TFs and paired TEs identified to bind (adjusted p-value <0.05) and to induce expression of TE subfamilies (p-value <0.05); Dot size is proportional to the log of fold change of induction upon overexpression of TFs in hESCs; Color intensity corresponds to the log of binding enrichment. c, Scatter plot illustrating the coupling between germ layer-specific TFs and TE subfamilies. y-axis, TE subfamily log2 fold change expression of intergenic subfamily (excluding all reads overlapping a TE in an exon, an intron and /-10kb of protein coding gene bodies) induced by overexpressed TF in hESC; x-axis, log2 fold change expression of these TFs in human gastrula versus epiblast cells. *d*, Binding of transcription factors to their expression-sensitive TE subfamilies. Heat map of the binding profile of transcription factors to all TE elements of the indicated subfamilies. We performed ChIP-seq of KLF4 in naive hESCs; ChIP-seq of GATA6 and SOX17 were respectively re-analyzed from hESC-derived endodermal cells (Lee et al., 2019) and a germ cell line (Jostes et al., 2020). Black intensity reflects the binding strength of the transcription factors.

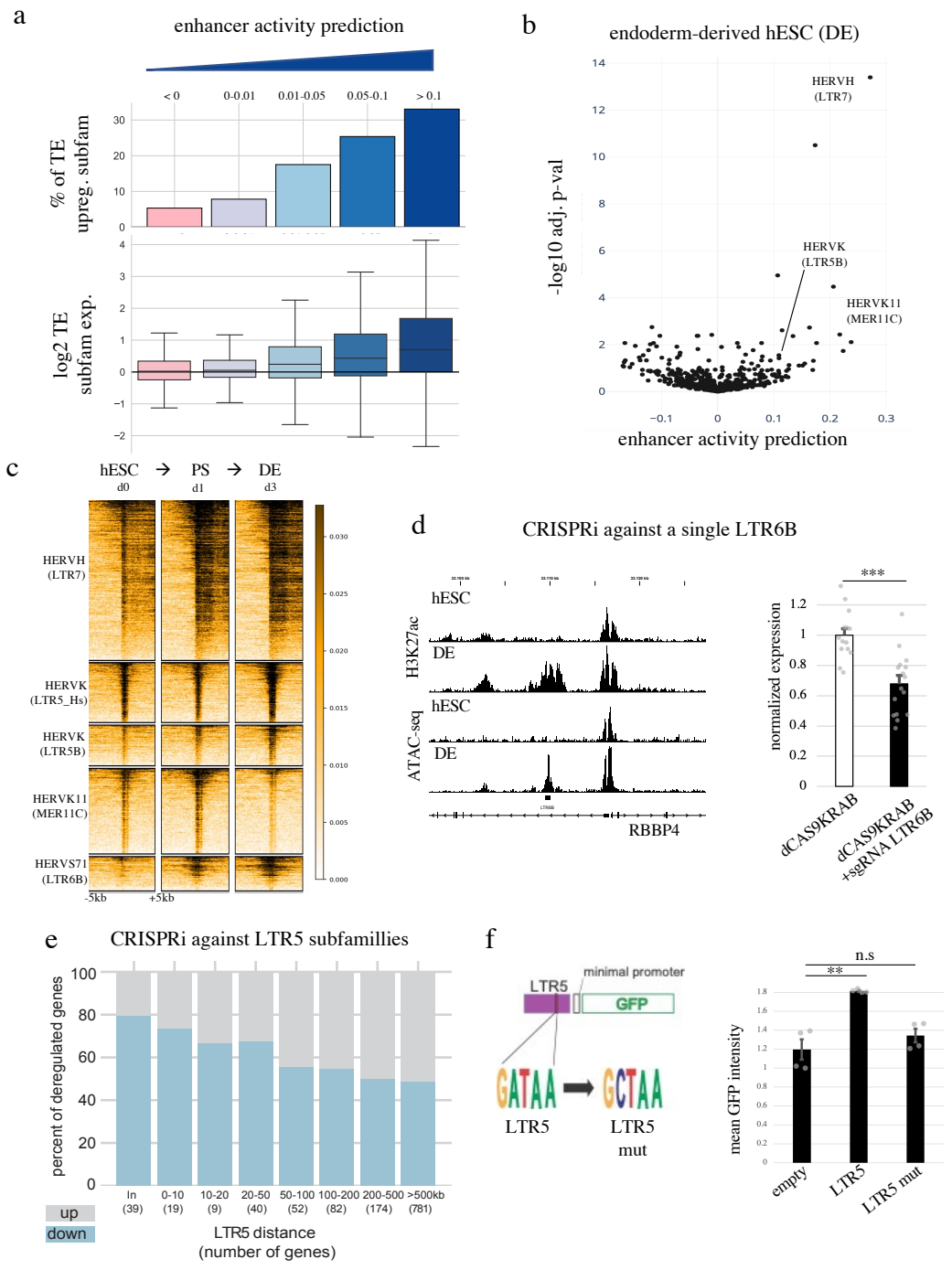



Figure 2.4 – **Cell-type-specific TEs control gene expression during gastrulation**

Figure 2.4 – **a**, Enhancer activity prediction of TE subfamilies correlates with their transcriptional activation. Enhancer activity prediction was established for all TE subfamilies in all overexpressed transgene conditions and grouped based on their activity values from the lowest to the highest (≤ 0 with 258071 values, 0-0.01 with 29810 values, 0.01-0.05 with 11411 values, 0.05-0.1 with 1687 value, >0.1 with 476 values); bar plots represent the percentage of transcriptionally induced TE subfamilies (adjusted p-value <0.05, at the subfamily level of intergenic TEs, two-sided t.test with p.value correction for multiple testing using the Benjamini-Hochberg's method) in each category; boxplots represent log2 fold TE subfamily add-up of normalized read count expression change. **b**, TE-derived enhancer activity prediction upon endoderm differentiation. Representation of enhancer activity prediction for all TE subfamilies after comparing the transcriptome of hESC and hESC-derived endodermal cells after 3 days of differentiation; x-axis represents the activity value and the y-axis, the -log10 adjusted p-value (establish by null significance hypothesis testing on the linear regression coefficients and accounted for multiple testing using the Benjamini Hochberg procedure). **c**, H3K27ac enrichment over TE subfamily during hESC-derived endodermal differentiation. Black intensity correlates to H3K27ac ChIP-seq signal +/-10kbp around all TEs from a named subfamily. d, RBBP4 is controlled by an LTR6B endoderm-specific enhancer. Left panel, genome browser of enhancer hallmark landscape (H3K27ac) (Loh et al., 2014) and chromatin accessibility profile (ATAC-seq) (Lee et al., 2019) of the promoter region of the RPPB4 gene in hESC and hESC-derived endoderm cells (DE); right panel, normalized RBBP4 RTqPCR result (over beta-actin and empty) of CRISPRi transduction with (+sgRNA) or without (open) sgRNAs targeting the LTR6B integrant upstream of RBBP4 followed by endodermal differentiation; error bars indicate SEM and p-value using a two-sided t-test (***: 4e-05) of 14 measurement generated by 4 biologically independent experiments. **e**, Impact of LTR5-targeting CRISPRi on gene expression during endodermal differentiation. Number of up- and downregulated genes (p-value <0.05. two-sided t.test) at an indicated distance from closest CRISPRi-targeted TE is shown (in: TE within a gene). **f**, LTR5 tissue-specific enhancer activity depends on GATA6. Left, schematic representation of the GFP-expressing vector harboring the LTR5B-derived enhancer fragment bound by GATA6 upstream of a minimal promoter; right, GFP activity illustrating the GATA6-dependent enhancer activity of LTR5B; error bars represent SEM and p-value using a two-sided t-test (**: 0.01, n.s : 0.3) of 4 measurements generated by 2 biologically independent experiments.

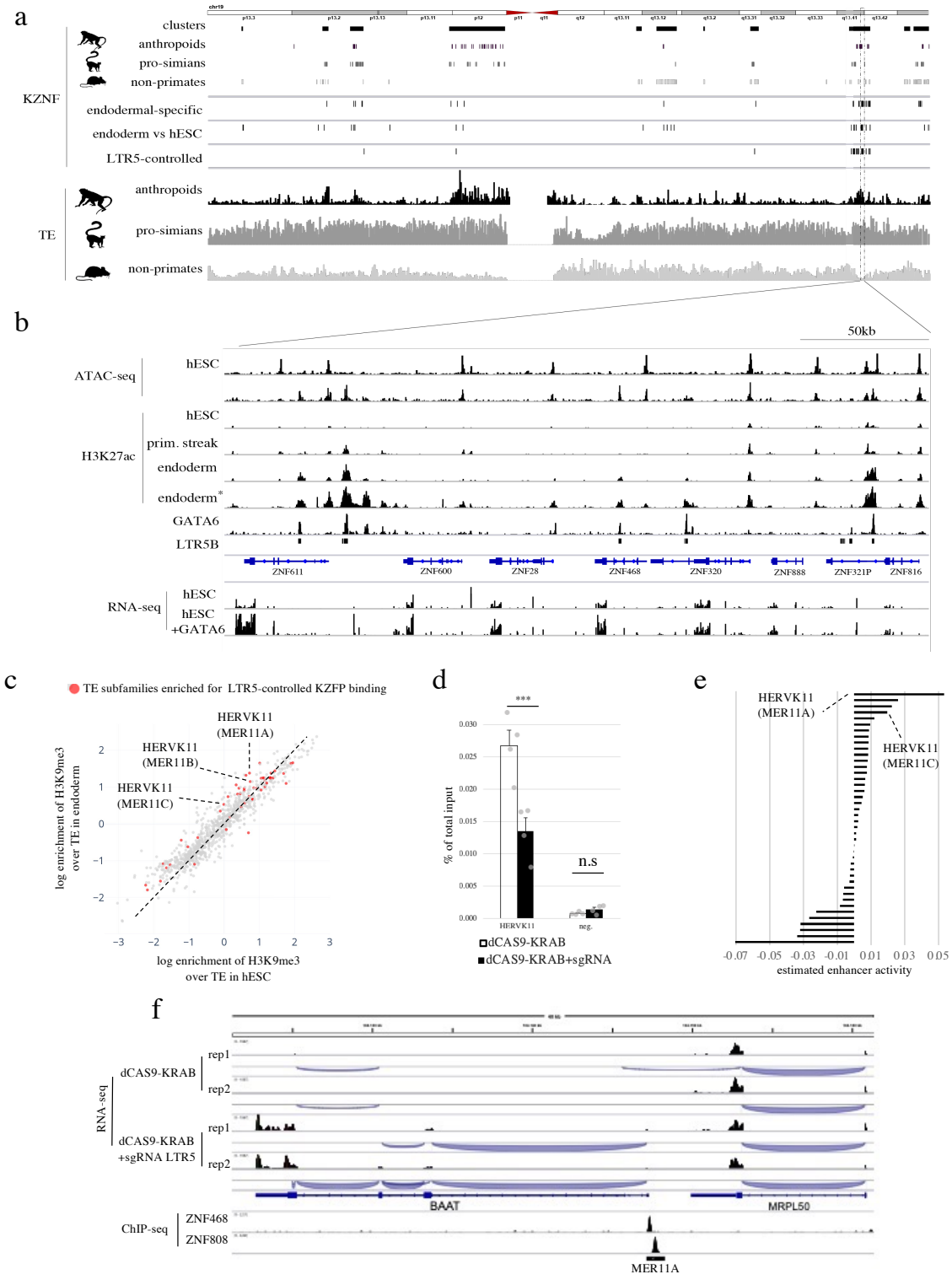



Figure 2.5 – **Primate specific *cis*- and *trans*-regulators partner up to control human gastrulation**

Figure 2.5 – **a**, Transcriptional KZFP regulation during endodermal differentiation, evolutionary age, and genomic distribution of TEs and KZFPs located on chromosome 19. Top, schematic representation of human chromosome 19 with KZFP gene clusters and individual KZFP. Middle, from top to bottom: KZFPs significantly upregulated in endodermal compared to other cells in human gastrula, in hESC-derived endoderm compared to hESC cells, and subjected to LTR5 control (as indicated by downregulation upon LTR5-targeting CRISPRi during endodermal differentiation with a p-value <0.05, used two-sided t.test). Bottom, TE density classified by evolutionary ages. b, Chromatin landscape of endodermal-specific KZFP cluster. Top, ATAC-seq and H3K27ac ChIP-seq profiles during endoderm differentiation at embryonic stem cell stage (hESC, day 0), primitive streak stage (prim. streak, day 1), and endodermal stage (endoderm, day 3) re-analyze from (Loh et al., 2014); H3K27ac (endoderm*) and GATA6 ChIP-seq correspond to the same kinetic dataset than the ATAC-seq re-analyzed from (Lee et al., 2019); bottom, RNA-seq performed in doxycycline-inducible GATA6 hESC line with (hESC+GATA6) or without (hESC) doxycycline during 48h, re-analyzed from (Nakatake et al., 2020). **c**, H3K9me3 differential enrichment in hESC vs hESC-derived endoderm. H3K9me3 TE subfamily enrichment was plotted for all subfamilies containing at least 10 integrants enriched in this mark in at least one condition; in red represent the TE subfamily significantly targeted by LTR5-controlled KZFPs. **d**, H3K9me3 loss at HERVK11 upon LTR5-mediated repression during endodermal differentiation. Panel represents H3K9me3 ChIP-qPCR upon LTR5 repression at this same HERVK11 elements; error bars indicate SEM and two-sided t.test were performed with 0.176 non-significant (n.s) or significant (***) 0.006 p-value results on biological quadruplicates. **e**, Cis-regulatory estimation of TE subfamily bound by LTR5-controlled KZFPs in hESC-derived foregut with or without LTR5 mediated repression. **f**, Transcriptional landscape of MER11A controlling BAAT expression in hESC-derived foregut with or without LTR5 mediated repression; bottom lanes are ChIP-seq of KZFP from (Imbeault, Helleboid, & Trono, 2017).

## 2.7 Supplementary information

# Fig. S1: Cell-type Specific Expression of Primates TEs during Human Gastrulation.

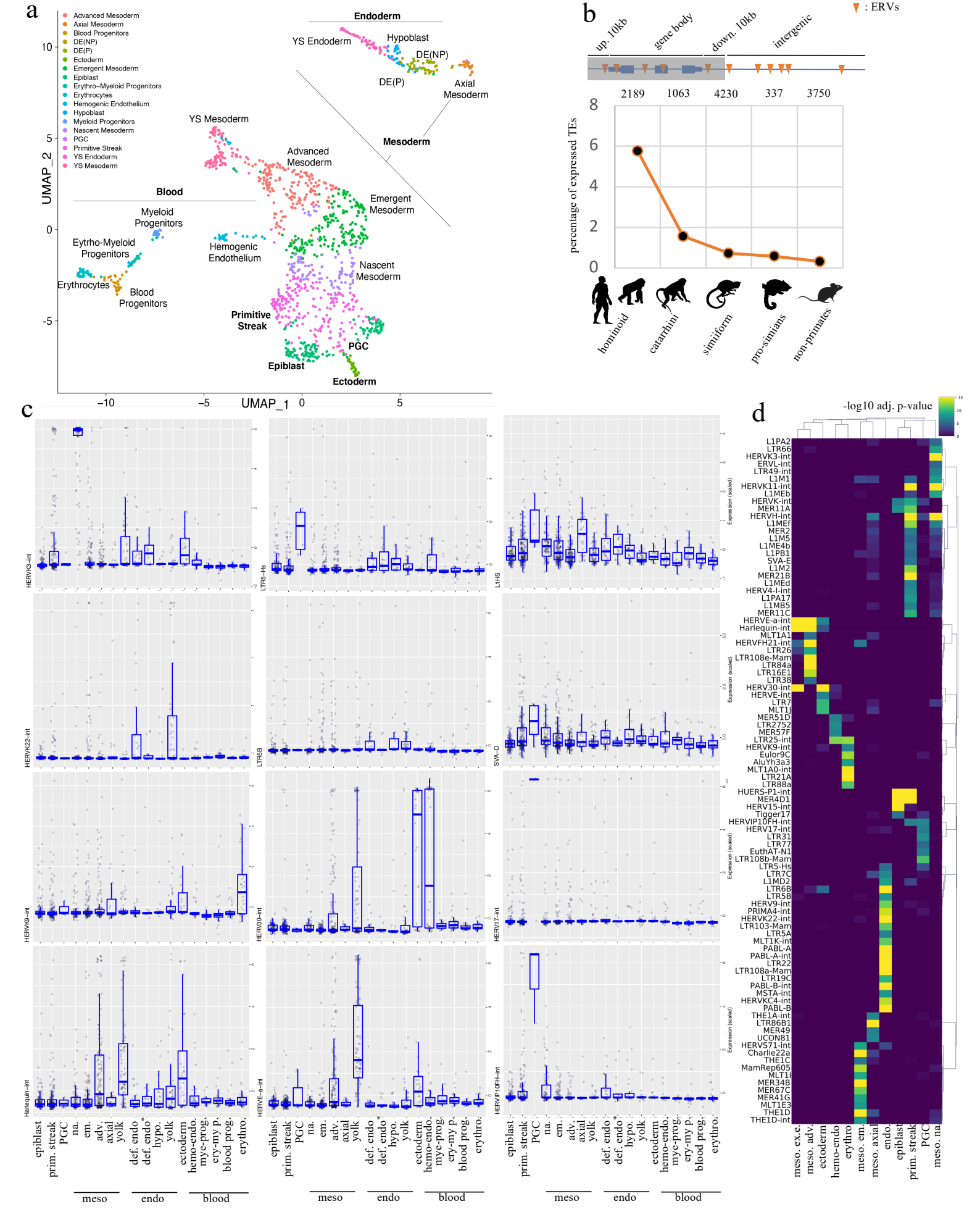
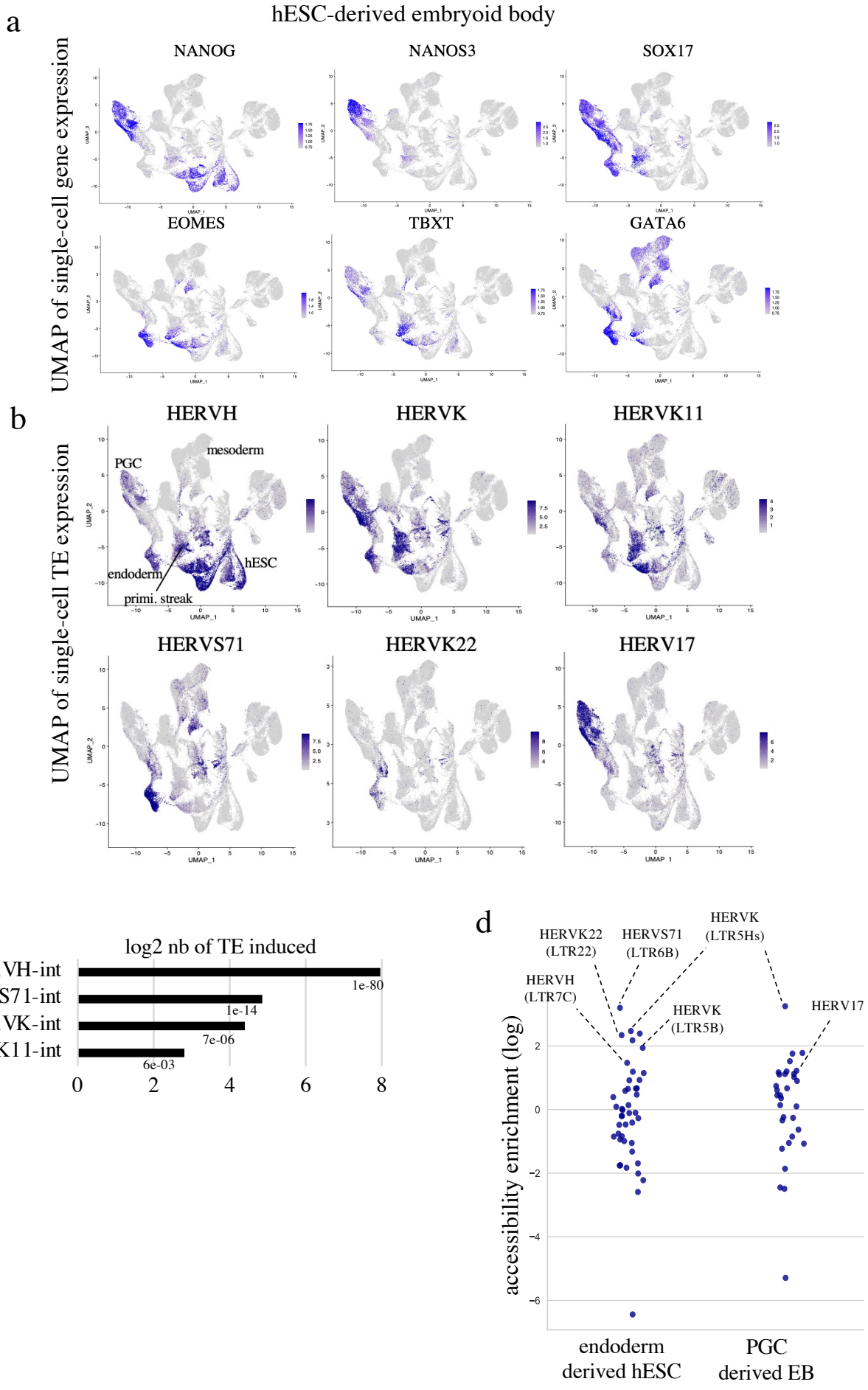


**Figure S1 for Fig 1.** ***Cell-type-specific expression of primate-restricted TEs during human gastrulation*****.**

**a**, Cellular composition of human gastrula. UMAP based on single-cell gene expression from human gastrula. Colors represent more detailed cell subtypes identified in[67]. **b**, Age distribution of expressed intergenic TEs in human gastrula. Each TE subfamily (excluding DNA transposons) was restricted to a specific evolutionary age category, and the percentage of expressed TE integrants in each was plotted. We excluded any TEs overlapping coding gene body and up/down-stream of a gene, thus removing TEs expressed due to readthrough or gene transcript inclusion. **c,** Cell-type-specific expression of TE subfamilies; boxplots with each dot representing the TE subfamily normalized expression in one cell. Cells were grouped in boxplots corresponding to one cell type of human gastrula sub-clustering: epiblast (133 cells), primitive streak (prim. streak, 195 cells), primordial germ cells (PGC, 7 cells), nascent mesoderm (na. 98 cells), emergent mesoderm (em. 185 cells), advance mesoderm (adv. 164 cells), axial mesoderm (axial, 23 cells), yolk mesoderm (yolk, 83 cells), definitive endoderm (def. endo. 35 cells), definitive endoderm non-proliferative (def. endo*, 18 cells), hypoblast (hypo, 29 cells), yolk endoderm (yolk, 53 cells), ectoderm (ectoderm 29 cells), hemogenic endothelium (hemo-endo. 37 cells), myeloid progenitor (mye-prog. 17 cells), erythro-myeloid progenitor (ery-my p. 28 cells), blood progenitor (blood prog. 29 cells), erythrocyte (erythro. 32 cells). **d,** Cell-type-specific expression of TE subfamilies; heatmap of the -log10 adjusted p-value of cell-type-specificity of TE subfamily expression; only TE subfamilies with adjusted p-value < $10^{e-5}$ were plotted (p-value are established using non-parametric Wilcoxon rank sum test).

**Fig. S2: Evolutionarily recent cell type-specific TEs have cis-regulatory potential during human development**



# Fig. S2: Evolutionarily recent cell type-specific TEs have cis-regulatory potential during human development

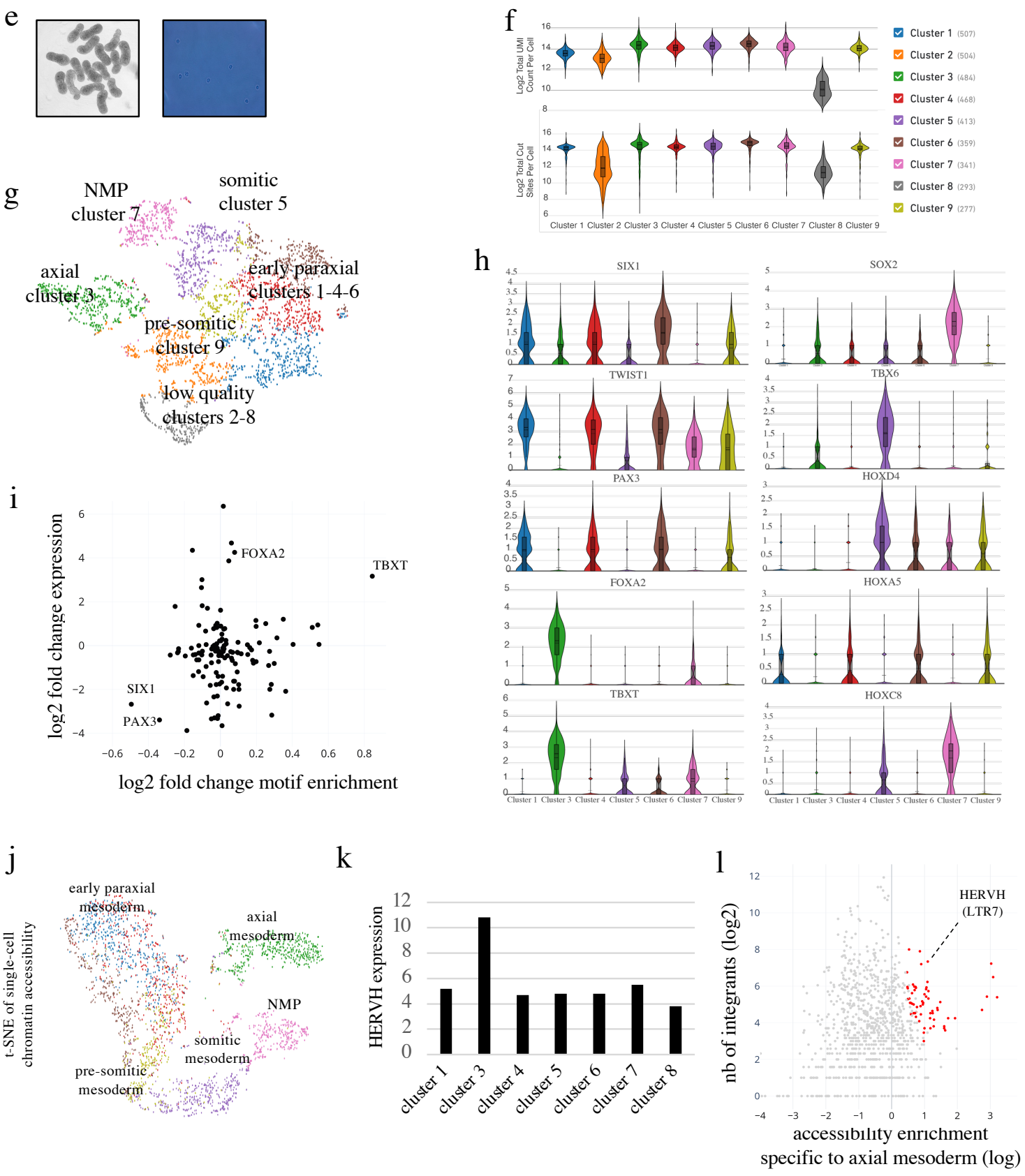
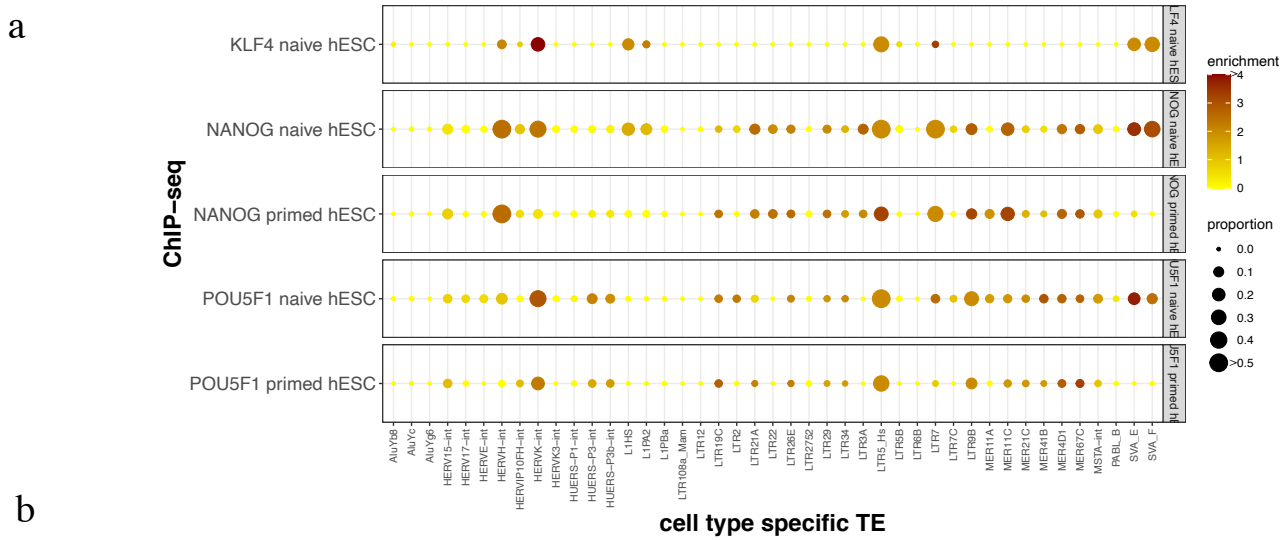
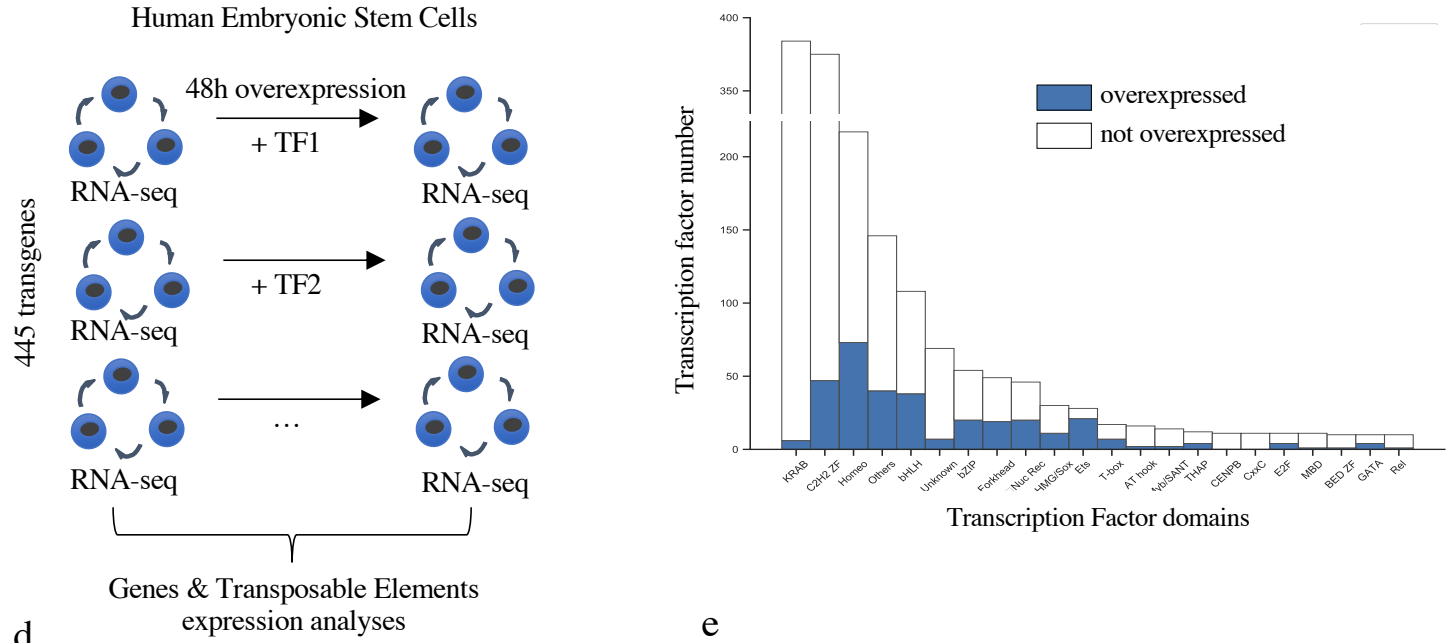


**Figure S2 for Fig 2. TEs are controlled by tissue-specific transcription factors**

**a,** Cell-type-specific expression of transcription factors during embryoid body differentiation. Each plot represents a UMAP defined on single-cell gene expression during differentiation of hESC into embryoid bodies over 5 days, re-analyzed from[26]; hESC corresponds to day 0; primitive streak (prim. streak) to days 1-2 with TBXT expression; cells at days 2-5 are stratified into PGC expressing NANOS3, NANOG, and SOX17, endoderm expressing SOX17 and GATA6, and mesoderm expressing GATA6 only. Color scale corresponds to level of relative TE subfamily expression based on normalized read counts. **b,** Cell-type-specific expression of TE subfamily during embryoid body differentiation. Each plot represents a UMAP of single-cell gene expression during *in vitro* differentiation of hESC into embryoid body over 5 days re-analyzed from[26]; hESC corresponds to day 0; primitive streak (prim. streak) to days 1-2 with TBXT expression; cells at days 2-5 are stratified in PGC expressing NANOS3, NANOG, and SOX17, endoderm expressing SOX17 and GATA6, and mesoderm expressing GATA6 only. Color scale corresponds to level of relative TE subfamily expression based on normalized read counts. **c,** Log2 number of expressed TE integrants induced upon hESC-derived endodermal differentiation; p-value enrichment is represented for each presented subfamily (two-sided t.test). **d,** Chromatin accessibility at endodermal and PGC-expressed TE subfamilies (adjusted p-value < 0.05) in hESC-derived endoderm and PGC, calculated over random genomic distribution and represented in natural log. **e,** Gastruloid formation and nuclei isolation. Left panel, picture of pooled elongated gastruloids used for the multi-omics experiment; right panel, trypan blue of purified nuclei used for the multi-omics experiment. **f,** UMI (unique molecular identifier) distribution of gene expression and transposase cut site counts of chromatin accessibility in each cluster defined by expression in single-nuclei RNA-seq and ATAC-seq. Clusters 2 and 8 contained lower amounts of RNA and/or ATAC-seq UMI/cut sites, hence were ignored for the rest of the analysis; clusters 1 and 4 look similar to cluster 6 but additionally express different cycling genes. Each cluster from 1-9 contains 507, 504, 484, 468, 413, 359, 341, 293, 277 cells respectively. **g,** UMAP of single-cell clustering of gastruloids based on gene expression; each color represents a different cluster illustrated in Fig. S2c. **h,** Violin plot of expression level of indicated cell-type-specific transcription factors for each cluster from Fig. S2c; cell-type-specific transcription factors were selected based on cluster-specificity of expression and DNA binding motif enrichment for their chromatin accessibility. Each cluster from 1-9 contains 507, 504, 484, 468, 413, 359, 341, 293, 277 cells respectively. **i,** Relative expression (*y-axis*) and motif enrichment at accessible chromatin (*x-axis*) of cell-type-specific transcription factors in axial mesodermal cells (cluster 3) comparted with other clusters.**j,** Chromatin accessibility in human gastruloids, analyzed at the single-cell level. t-SNE plot (t-distributed Stochastic Neighbor Embedding) representing chromatin accessibility clustering. Colors correspond to gene expression clustering from Fig. S2c-e. **k**, Expression profile of the HERVH subfamily in gastruloid-derived clusters. The *y-axis* corresponds to the normalized sum of accessible HERVH expression from axial mesoderm (315 loci) in each cluster. **l**, TE subfamily enrichment of chromatin accessibility (p-value < 0.05) in axial mesoderm. Each dot represents a TE subfamily*; x-axis*, natural log fold enrichment of TE loci compared to a random genomic distribution and *y-axis* represents the log2 number of accessible integrants. Red dots are TE subfamilies with a p-value of enrichment < 0.05.

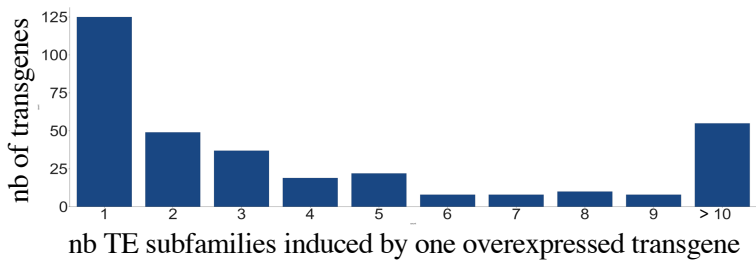
# Fig. S3: Tissue-specific transcription factors control cell-type-specificity of TE expression

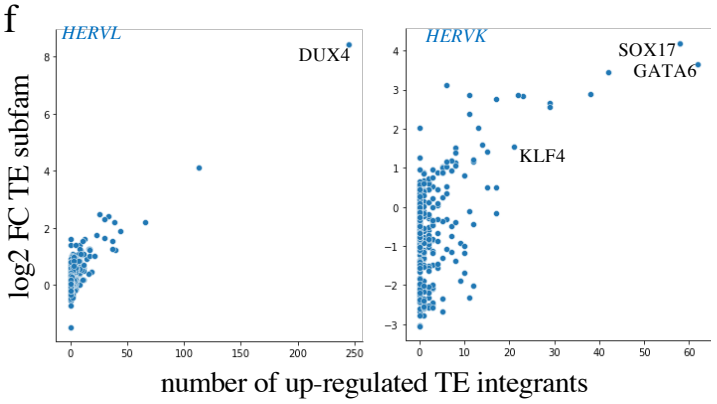



d

| | coding genes | transposable elements |
|---|---|---|
| total | 20'327 | 4'570'939 |
| expressed | 18'817 | 397'163 |
| deregulated | 18'428 | 216'160 |



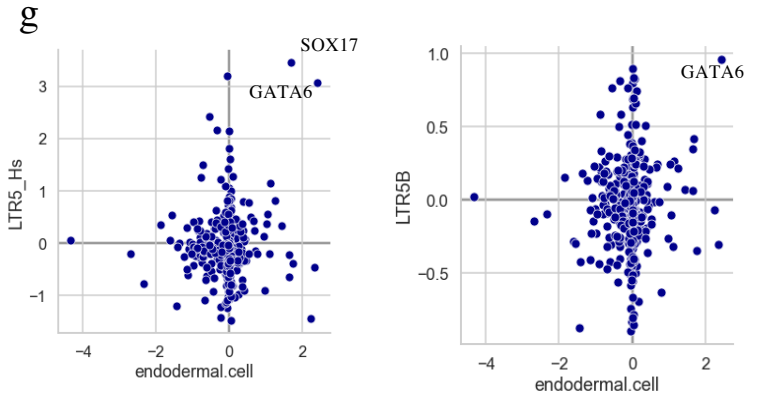
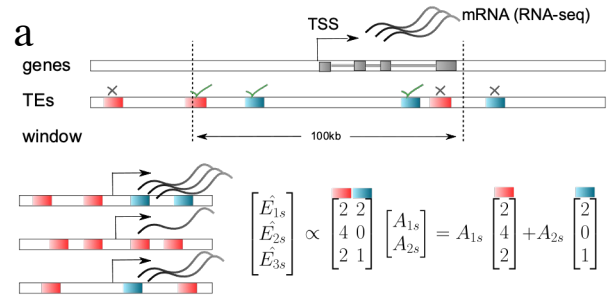
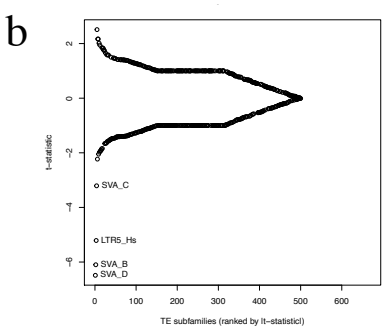



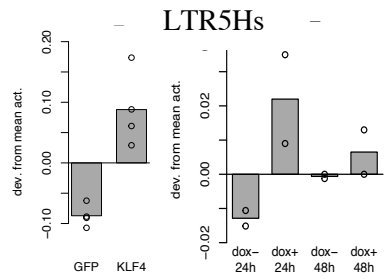
**Figure S3 for Fig 3. Evolutionary recent TEs act as cell-type-specific enhancers during human gastrulation and fetal development.**

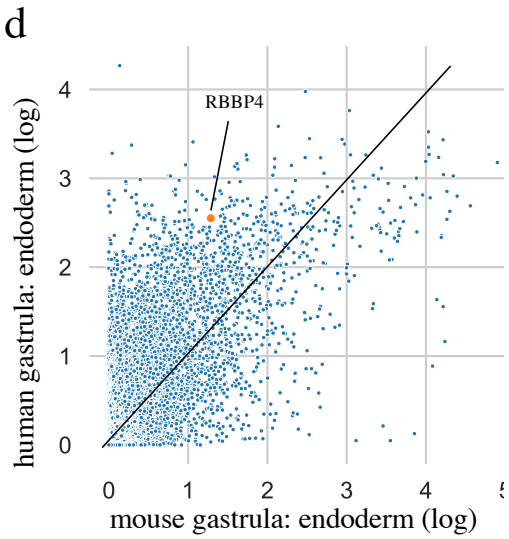
**a,** Binding of pluripotency factors at TE expressed in gastrula and accessible in pre-implantation *in vitro* or *in vitro* models. The *x-axes* correspond to the cell type-specific TE subfamilies in human gastrula (p.value < 10e-5) that are also accessible in one of the in vivo or in vitro models (p.value < 10e-3), the *y-axis* the ChIP-seq of pluripotency factors in primed/naïve hESCs; circle sizes represent the number of accessibility sites overlapping with a specific TE subfamily, normalized by the number of elements in that subfamily (p-value are established using Homer algorithm); color intensity represents the log enrichment relative to the random distribution of this overlap. **b,** Design of TF overexpression in hESC experiment performed in[31]. **c,** Overexpressed TF subtypes. Each bar represents the number of proteins harboring the indicated domain, with fraction in blue indicating those overexpressed in [31]. **d,** Sum of total, expressed and differentially expressed genes and TEs upon overexpression of 234 TFs in hESC (with adjusted p-value <0.05, two-sided t.test with p.value correction for multiple testing using the Benjamini-Hochberg's method). **e**, Number of TE subfamilies deregulated per overexpressed TF; each bar plot represents the number of TFs inducing the number of TE subfamilies as indicated on *x-axis* (adjusted p-value < 0.05 for significantly up-regulated TEs over-representation in a TE subfamily, two-sided t.test with p.value correction for multiple testing using the Benjamini-Hochberg's method). **f,** Scatter plot illustrating coupling between EGA-induced TFs and TEs. *y-axis*, log2 fold TE subfamily add-up of normalized read count expression change; *x-axis*, number of up-regulated TE integrants from that subfamily (2-fold with adjusted p-value <0.05, two-sided t.test with p.value correction for multiple testing using the Benjamini-Hochberg's method). **g,** Scatter plot illustrating correlation between germ layer-specific TEs and TFs. *y-axis*, log2 fold TE subfamily add-up of normalized read count expression induced by overexpressed transcription factors in hESC; *x-axis*, log2 fold change expression of these transcription factors in human gastrula versus epiblast cells.

# Fig. S4: Cell type-specific TE control gene expression during gastrulation



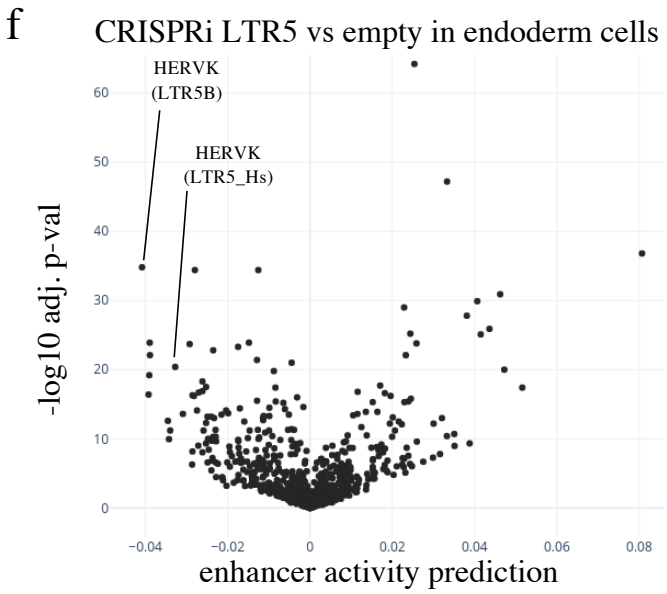



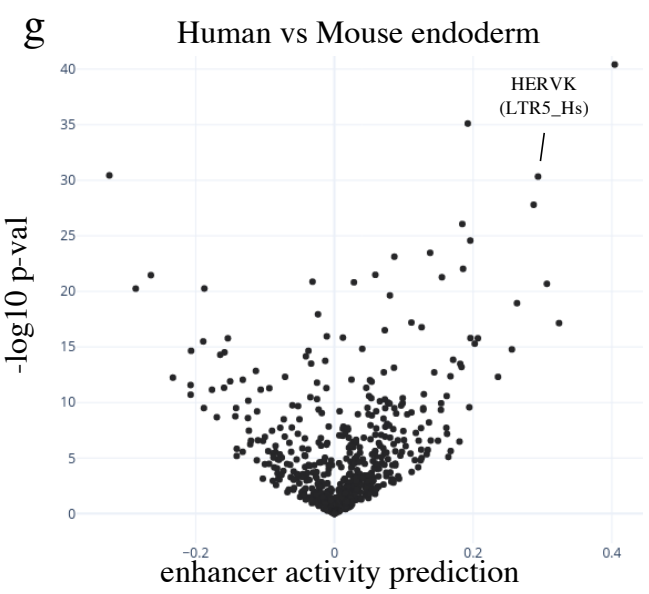




e

| TE subfamilly name | number of integrants | predicted targets |
|---|---|---|
| LTR5_Hs | 645 | 483 |
| LTR5B | 431 | 178 |
| LTR5A | 265 | 1 |
| LTR5 | 20 | 2 |



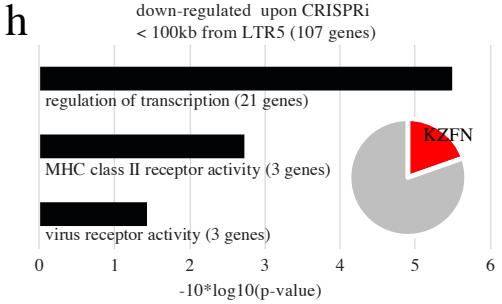



**Figure S4 for Fig 4. Cell-type-specific TE control gene expression during gastrulation**

**a,** TE subfamily-derived enhancer activity estimation schema. Regressive correlation is applied based on each coding gene expression level (Ex) and neighboring representation (100kb around TSS) of each TE subfamilies (T) in each TF-overexpressing hESC (E), allowing estimation of each TE subfamily relative enhancer activity (A). **b,** T-statistical tests of TE subfamily activity around deregulated genes upon CRISPRi-targeting SVA/LTR5Hs in naïve hESCs. **c,** Activity of LTR5Hs subfamily in the presence or absence of KLF4 overexpression in primed hESCs from[9], (left), and[31] at 24h or 48h overexpression (right). **d,** Scatterplot of gene expression comparison of human and mouse endoderm. Red dot highlights the *RBBP4* expression level; single-cell RNA-seq expression data of endodermal tissues of human gastrula[25] is compared to a single-cell RNA-seq of endodermal tissues of mouse gastrula (mixed stages) [70]. **e,** Predicted number of LTR5 loci targeted by CRISPRi (identified by CRISPOR[53]). **f,** TE-derived enhancer activity prediction upon CRISPRi-targeting LTR5 during endodermal differentiation. Representation of enhancer activity prediction for all TE subfamilies after comparing the transcriptome of hESC-derived endodermal cells with or without CRISPRi targeting LTR5 after 3 days of differentiation*; x-axis* represents the activity value, and *the y-axis* the -log10 adjusted p-value (establish by null significance hypothesis testing on the linear regression coefficients and accounted for multiple testing using the Benjamini Hochberg procedure). **g,** TE-derived enhancer activity prediction between human and mouse endodermal gastrulating cells. Representation of enhancer activity prediction for all TE subfamilies after comparing single-cell RNA-seq expression data of endodermal tissues of human gastrula[25] and the single-cell RNA-seq of endodermal tissues of mouse gastrula (mixed stages)[70]; *x-axis* represents the activity value and the *y-axis* the -log10 p-value (establish by null significance hypothesis testing on the linear regression coefficients). **h,** Gene Ontology of nearby LTR5-controlled genes. All down-regulated genes upon LTR5-targeting CRISPRi in a 100kbp window from LTR5 were selected; 2D-pie represents the proportion of 21 KZNF genes among the 107 down-regulated genes.

**Fig. S5: Primate specific cis- and trans-regulators partner up to control human gastrulation**

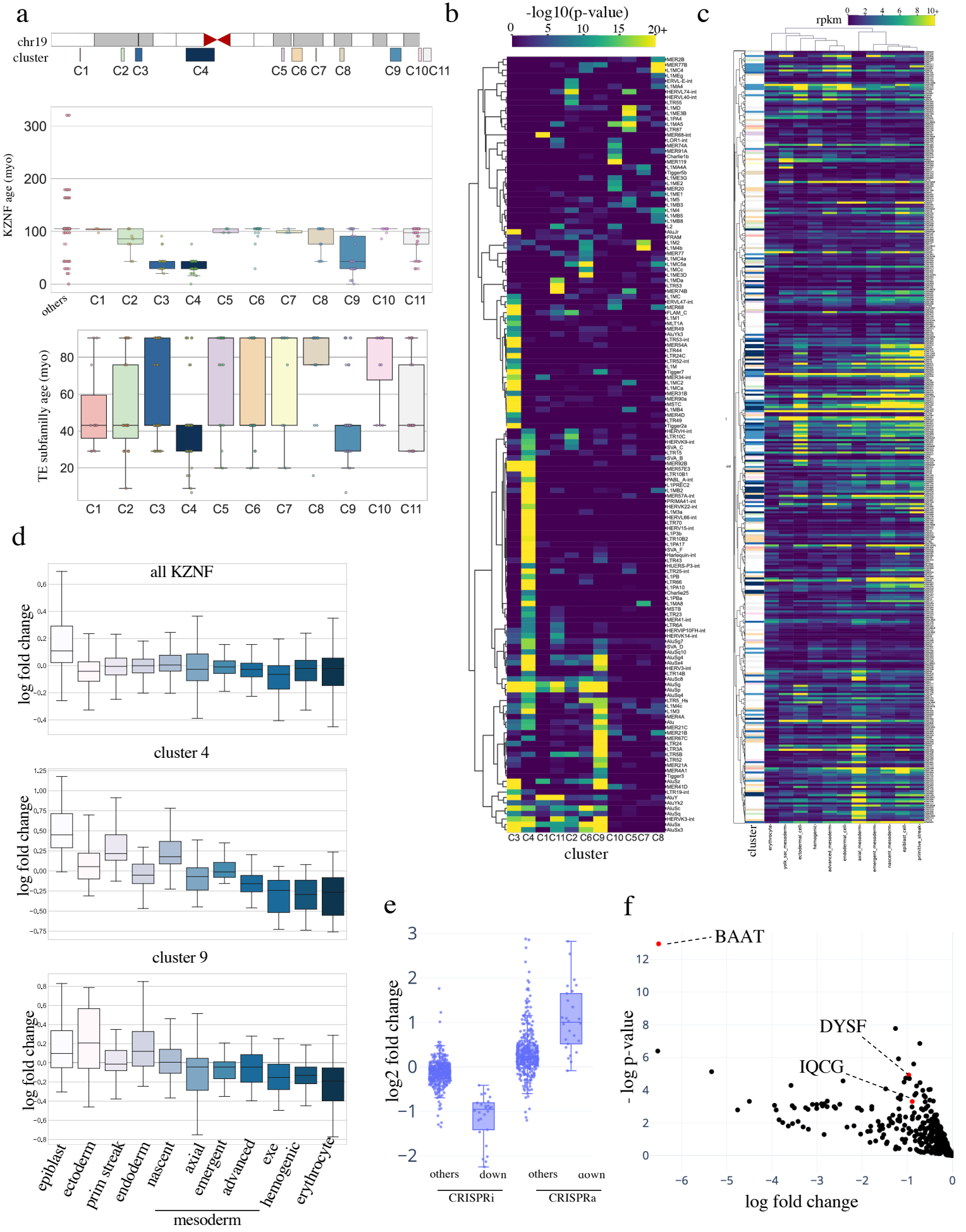


**Figure S5 for Fig 5. Primate specific cis- and trans-regulators partner up to control human gastrulation**

**a,** Evolutionary young *KZFP* gene clusters are enriched in contemporary TEs. Top, boxplot of *KZFP* genes evolutionary ages within 11 clusters (C1 to C11, with 4, 9, 20, 33, 5, 28, 3, 18, 37, 9, 39 genes respectively) on chromosome 19 or elsewhere in the genome (others with 152 genes); bottom, boxplot of evolutionary ages of TE subfamilies enriched in these *KZFP* gene clusters (p-value < $10^{e-4}$, with 8, 19, 39, 58, 10, 17, 4, 10, 28, 8, 17 TE subfamilies respectively from C1 to C11).

**b,** Heatmap of TE subfamily enrichment in *KZFP* gene clusters. Yellow intensity is proportional to significance (-10*log10(p-value established using Homer algorithm); only TE subfamily with a p-value enrichment of 10e-4, a fold change of 2 and at least 5 integrant inside one cluster were represented. **c**, Heatmap of *KZFP* genes expression in human gastrula re-analyzed from[67]. Yellow intensity is proportional to log normalized counts; only *KZFP* genes with at least 2 normalized counts were plotted. **d,** *KZFP* genes expression in human gastrula. Each boxplot represents log2 gene expression fold change of one cell type over the others (386 *KZFP* genes). Top panel shows the expression of all *KZFP* genes; middle and bottom panels depict *KZFP* genes in clusters 4 and 9, respectively. **e,** LTR5-controlled *KZFP* genes vs other genes during endodermal differentiation are activated upon CRISPRa in NCCIT cell line. Left, boxplots of *KZFP* genes significantly (down, 26 genes) or not (others, 339 genes) down-regulated upon LTR5-mediated repression during endodermal differentiation; on the right are presented the fold change upon LTR5-mediated activation in NCCIT cell line of the down (26 genes) and others (339 genes) KZFPs re-analyzed from[72]. **f,** Proximal MER11 genes downregulated upon CRISPRi targeting of LTR5 in hESC-derived foregut. The *x-axis* represents the logarithmic change; the *y-axis* represents the logarithmic p-value (two-sided t.test). Only genes located within 50kb of a MER11 and downregulated are shown. Red dots represent genes identified in in[35].

| Observed/Expected extracted from Barakat et al. 2018 | Primed hESC | | | | | Naive hESC | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| TE subfamillies | p_min-128 | p_128-256 | p_256-512 | p_512-1024 | p_1024-max | n_min-128 | n_128-256 | n_256-512 | n_512-1024 | n_1024-max |
| LTR7 | 0.79 | 1.46 | 2.76 | 4.99 | 8.90 | 0.95 | 0.80 | 0.00 | 13.66 | 0.00 |
| LTR21A | 0.99 | 1.25 | 1.58 | 0.00 | 0.00 | 0.80 | 2.27 | 2.10 | 8.57 | 0.00 |
| MER11C | 0.92 | 1.10 | 2.12 | 3.72 | 2.45 | 0.89 | 1.35 | 3.76 | 0.00 | 0.00 |
| MER4D1 | 0.95 | 1.29 | 1.16 | 2.20 | 2.11 | 0.86 | 1.92 | 2.50 | 2.56 | 0.00 |
| MER67C | 0.93 | 0.72 | 1.49 | 3.80 | 4.31 | 0.96 | 1.40 | 1.55 | 0.00 | 0.00 |
| L1PA2 | 0.92 | 1.52 | 2.33 | 1.47 | 1.91 | 0.96 | 1.31 | 1.31 | 1.55 | 0.69 |
| LTR34 | 0.91 | 2.47 | 0.70 | 0.00 | 2.78 | 0.96 | 0.00 | 5.41 | 0.00 | 0.00 |
| MER11A | 0.97 | 0.81 | 1.23 | 4.27 | 0.82 | 0.93 | 1.69 | 1.39 | 1.06 | 0.00 |
| MSTA-int | 0.91 | 1.06 | 2.11 | 3.00 | 3.84 | 1.04 | 1.03 | 0.00 | 0.00 | 0.00 |
| LTR29 | 0.96 | 0.91 | 1.53 | 0.80 | 3.83 | 1.00 | 0.79 | 1.47 | 1.50 | 0.00 |
| LTR5_Hs | 0.95 | 0.72 | 1.55 | 3.13 | 3.59 | 1.04 | 0.74 | 0.62 | 0.00 | 0.00 |
| LTR9B | 0.94 | 1.15 | 1.81 | 0.87 | 3.62 | 1.05 | 0.51 | 0.86 | 1.17 | 0.00 |
| LTR26E | 0.92 | 1.03 | 4.59 | 0.00 | 1.31 | 0.91 | 2.03 | 1.13 | 0.00 | 0.00 |
| LTR6B | 0.96 | 0.79 | 1.00 | 0.00 | 5.99 | 1.02 | 0.83 | 1.15 | 0.00 | 0.00 |
| LTR3A | 0.99 | 0.00 | 1.70 | 0.00 | 6.81 | 1.03 | 1.15 | 0.00 | 0.00 | 0.00 |
| HERVH-int | 0.95 | 1.01 | 1.27 | 2.17 | 3.34 | 1.04 | 0.48 | 1.32 | 0.00 | 0.00 |
| LTR5B | 1.01 | 0.81 | 1.03 | 1.07 | 1.03 | 0.98 | 1.12 | 0.78 | 3.17 | 0.00 |
| HUERS-P3b-int | 1.03 | 0.94 | 0.79 | 0.00 | 0.00 | 0.95 | 0.83 | 2.32 | 3.79 | 0.00 |
| AluYg6 | 1.04 | 0.70 | 0.89 | 0.00 | 0.00 | 0.87 | 2.25 | 0.89 | 3.64 | 0.00 |
| LTR12 | 1.00 | 0.58 | 1.48 | 3.09 | 0.00 | 1.01 | 0.44 | 2.46 | 0.00 | 0.00 |
| HERVIP10FH-int | 1.04 | 0.81 | 0.68 | 0.00 | 0.45 | 1.02 | 0.36 | 1.51 | 4.11 | 0.00 |
| LTR7C | 1.04 | 0.39 | 1.00 | 2.08 | 0.00 | 1.00 | 1.23 | 0.44 | 1.04 | 1.50 |
| LTR2 | 1.01 | 0.84 | 0.35 | 0.74 | 2.83 | 0.97 | 1.33 | 1.23 | 0.00 | 0.00 |
| MER41B | 0.98 | 1.10 | 0.85 | 2.11 | 1.69 | 1.07 | 0.68 | 0.00 | 0.00 | 0.00 |
| SVA_E | 1.05 | 0.43 | 0.79 | 0.61 | 0.20 | 0.93 | 1.66 | 1.15 | 1.57 | 0.00 |
| L1HS | 0.97 | 1.24 | 1.33 | 0.81 | 1.44 | 1.07 | 0.34 | 0.94 | 0.00 | 0.00 |
| HUERS-P1-int | 1.07 | 0.00 | 0.81 | 0.00 | 1.61 | 0.91 | 1.88 | 1.74 | 0.00 | 0.00 |
| HERV17-int | 1.02 | 1.05 | 0.45 | 0.93 | 0.00 | 0.86 | 2.79 | 0.78 | 0.00 | 0.00 |
| MER21C | 1.00 | 0.77 | 1.13 | 1.23 | 1.42 | 1.06 | 0.62 | 0.58 | 0.00 | 0.00 |
| LTR19C | 0.97 | 1.36 | 1.15 | 2.39 | 0.00 | 1.07 | 0.56 | 0.19 | 0.00 | 0.00 |
| L1PBa | 1.01 | 0.63 | 1.34 | 1.02 | 0.98 | 1.07 | 0.27 | 1.14 | 0.00 | 0.00 |
| LTR22 | 0.98 | 1.06 | 2.70 | 0.00 | 0.00 | 0.98 | 1.68 | 0.00 | 0.00 | 0.00 |
| PABL_B | 1.04 | 0.00 | 1.39 | 2.90 | 0.00 | 1.09 | 0.31 | 0.28 | 0.00 | 0.00 |
| HERVK-int | 1.01 | 0.65 | 0.99 | 2.75 | 0.00 | 1.11 | 0.13 | 0.36 | 0.00 | 0.00 |
| AluYc | 1.00 | 1.03 | 0.98 | 0.87 | 0.84 | 1.07 | 0.62 | 0.00 | 0.00 | 0.00 |
| HERVE-int | 1.02 | 0.72 | 1.23 | 0.00 | 1.23 | 1.06 | 0.80 | 0.00 | 0.00 | 0.00 |
| HERVK3-int | 1.07 | 0.67 | 0.00 | 0.00 | 0.00 | 0.95 | 1.43 | 1.70 | 0.00 | 0.00 |
| HUERS-P3-int | 1.00 | 1.47 | 0.62 | 0.00 | 0.00 | 1.09 | 0.49 | 0.00 | 0.00 | 0.00 |
| SVA_F | 1.08 | 0.25 | 0.36 | 0.56 | 0.24 | 1.06 | 0.61 | 0.31 | 0.00 | 0.00 |
| AluYb8 | 1.06 | 0.47 | 0.30 | 0.62 | 0.30 | 1.09 | 0.41 | 0.00 | 0.00 | 0.00 |
| HERV15-int | 1.11 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 1.03 | 1.07 | 0.00 | 0.00 |

**Supplementary Table1. Enhancer activity of TE subfamily in Human Embryonic Stem cells**

Data extracted from Barakat et al. 2018 supplemental data in primed and naive human embryonic stem cells (hESC). The most accessible subfamilies in Figure 2A were selected and their enhancer activity represented in this table when tested in hESCs with an episomal reporter assay.

# 3 Statistical learning quantifies transposable element-mediated *cis*-regulation

Cyril Pulver[1], Delphine Grun[1], Julien Duc[1], Shaoline Sheppard[1], Evarist Planet[1], Alexandre Coudray[1], Raphaël de Fondeville[2*], Julien Pontis[1,3*] and Didier Trono[1*]

* corresponding authors, equal contribution

Authors' email addresses, in order of appearance: cyril.pulver@epfl.ch, grundelphine@gmail.com, julien.duc@epfl.ch, shaoline.sheppard@gmail.com, evarist.planet@epfl.ch, alexandre.coudray@epfl.ch, raphael.defondeville@epfl.ch, julien.pontis25@gmail.com, didier.trono@epfl.ch

Author's affiliations: [1]School of Life Sciences, Swiss Federal Institute of Technology Lausanne (EPFL); [2]Swiss Data Science Center, Swiss Federal Institute of Technology Lausanne (EPFL); [3]Current address: SOPHiA GENETICS SA, La Pièce 12, CH-1180 Rolle, Switzerland, +41216941060

Published as a Research Article in Genome Biology (Pulver et al., 2023) https://doi.org/10.1186/s13059-023-03085-7. An early preprint version(Pulver et al., 2022) can be found at https://www.biorxiv.org/content/10.1101/2022.09.23.509180v1

I am the main contributor of this work, having devised and implemented the *craTEs* approach, generated all figures, drafted the original manuscript, replied to reviewers' comments and

revised the manuscript accordingly. My co-authors' contributions are listed in the Attributions section.

## 3.1 Abstract

Transposable elements (TEs) have colonized the genomes of most metazoans, and many TE-embedded sequences function as *cis*-regulatory elements (CREs) for genes involved in a wide range of biological processes from early embryogenesis to innate immune responses. Because of their repetitive nature, TEs have the potential to form CRE platforms enabling the coordinated and genome-wide regulation of protein-coding genes by only a handful of *trans*-acting transcription factors (TFs).

Here, we directly test this hypothesis through mathematical modeling and demonstrate that differences in expression at protein-coding genes alone are sufficient to estimate the magnitude and significance of TE-contributed *cis*-regulatory activities, even in contexts where TE-derived transcription fails to do so. We leverage hundreds of overexpression experiments and estimate that, overall, gene expression is influenced by TE-embedded CREs situated within approximately 500kb of promoters. Focusing on the *cis*-regulatory potential of TEs within the gene regulatory network of human embryonic stem cells, we find that pluripotency-specific and evolutionarily young TE subfamilies can be reactivated by TFs involved in post-implantation embryogenesis. Finally, we show that TE subfamilies can be split into truly regulatorily active versus inactive fractions based on additional information such as matched epigenomic data, observing that TF binding may better predict TE *cis*-regulatory activity than differences in histone marks.

Our results suggest that TE-embedded CREs contribute to gene regulation during and beyond gastrulation. On a methodological level, we provide a statistical tool that infers TE-dependent *cis*-regulation from RNA-seq data alone, thus facilitating the study of TEs in the next-generation sequencing era.

## 3.2 Introduction

The development and function of complex organisms relies on the tight regulation of gene expression at cellular and tissue levels. *Cis*-regulatory elements (CREs) are non-coding sequences that modulate the transcription of nearby genes in response to signaling cues, thereby contributing to the control of gene expression. Functionally, CREs operate through transcription factor (TF) recruitment and local chromatin remodeling (The ENCODE Project Consortium, 2012). Importantly, sequence specific TF-DNA binding allows for the simultaneous regulation of arbitrarily distant genes flanked by CREs carrying analogous TF binding sites (TFBS). Conceptually, the functional interactions implicating CREs, their target genes and their TF controllers form graph-like representations of the gene expression machinery known as gene regulatory networks (GRNs) (Britten & Davidson, 1971; Gerstein et al., 2012). Typically, one may represent CREs as edges connecting two types of nodes: TFs and the protein-coding genes they regulate. According to this view, cell-state and tissue-specific transcriptional programs - defined by specific sets of expressed TFs and accessible CREs - are thereby depicted by distinct GRN topologies. For example, the GRN of so-called "primed" human embryonic stem cells (hESCs), which resemble epiblast cells of the post-implantation embryo, is characterized by the expression and binding of OCT4, NANOG and SOX2 to pluripotency-specific CREs (Boyer et al., 2005). Changes in TF expression can alter GRN topology, thus polarizing cells towards a different state. For example, induced expression of Krüppel-like factor family (KLF) members in primed hESCs alters their GRN towards one resembling that of preimplantation-like "naïve" hESCs notably characterized by increased chromatin accessibility (Pontis et al., 2019b; Theunissen et al., 2016).

Whereas the repertoire of expressed TFs and accessible CREs varies across cell states within one organism, the genomic location of CREs with respect to their target genes varies across species. In fact, it has long been recognized that organisms evolve primarily through the emergence, spread and reorganization of CREs, i.e. modification of GRNs, (Britten & Davidson, 1971; King & Wilson, 1975; Wray, 2007) rather than through mutations affecting protein-coding

genes - including TFs - though exceptions to this tenet exist (Imbeault et al., 2017). GRNs may evolve through chromosomal or even genome-wide duplication events followed by divergence and specialization of the henceforth redundant regulatory sub-networks. However, large-scale duplications are too coarse to account for the fine-grained nuances in CRE compositions observed across the genomes of distinct species. Due to their important contribution to the size of most metazoan genomes, their intrinsic ability to recruit TFs and their potential for rapidly spreading ready-to-go regulatory modules throughout the genome of their host, transposable elements (TEs) have gained attention as a potential source of CREs (Britten & Davidson, 1971; Chuong et al., 2017; Feschotte, 2008).

TEs form a collection of genetic entities that autonomously or collectively code for the factors essential to their own mobility, a process known as transposition. Endogenous retroelements (EREs) propagate through retrotransposition, a copy-and-paste mechanism entailing the reverse transcription of an RNA intermediate encoded within the ERE sequence itself. In agreement with the replicative nature of retrotransposition, EREs constitute the vast majority of the approx. 4.5 million readily recognizable TE-derived sequences that contribute more than half of the human genome DNA content (Friedli & Trono, 2015; International Human Genome Sequencing Consortium, 2001). In contrast, DNA transposons propagate through a non-replicative cut-and-paste process and rely on genome replication to accumulate copies (Feschotte & Pritham, 2007). Both EREs and DNA transposons are further segregated into super/subfamilies (Bourque et al., 2018) forming sets of phylogenetically related integrants that use the same mechanism for transposition (Storer, Hubley, Rosen, Wheeler, & Smit, 2021). Seminal DNA reassociation studies demonstrated long before the Next Generation Sequencing era that most metazoan genomes were replete with repetitive sequences, some of which emerged in recent evolutionary times. Drawing from this line of work, Britten and Davidson famously reasoned that repetitive DNA may form a pool of potential CREs whose cycles of expansion followed by purifying selection fuels GRN evolution (Britten & Davidson, 1971). Consistent with this model, binding sites of conserved TFs and open chromatin regions enrich at evolutionarily young TE subfamilies, in particular in embryonic stem cells (ESCs), and more

occasionally in cancer cell lines and lymphoblastoid tissues (Bourque et al., 2008; Jacques et al., 2013; Pehrsson et al., 2019; Sundaram et al., 2014; Trizzino et al., 2018). Moreover, multiple functional studies support the regulatory potential of TEs, including evolutionarily recent integrants. For example, the majority of genes deregulated in human but not in mouse embryonic stem cells (mESCs) upon knockdown of the master regulator of pluripotency OCT4 are associated with EREs of the ERV1 family, for which an enhancer activity was confirmed by reporter assay (Kunarso et al., 2010). As well, the majority of species-specific enhancers in mouse and rat trophoblast stem cells overlap species-specific TE subfamilies, and a mouse specific subfamily (RLTR13D5) exhibits trophoblast stem cell-specific enhancer activity in a reporter assay (Chuong et al., 2013). Finally, the genetic excision of primate-specific MER41B integrants thwarts the functionality of a key innate immunity signaling cascade (Chuong et al., 2016) and hundreds of genes including stemness maintainers are downregulated upon epigenetic repression of the hominoid-specific SVA and LTR5-Hs subfamilies in hESCs (Pontis et al., 2019b). Together, these case studies suggest that evolutionarily recent EREs spread CREs upon which natural selection may act to fine-tune the GRNs of critical physiological processes such as embryogenesis and innate immunity (Chuong et al., 2017; Feschotte, 2008; Friedli & Trono, 2015). Despite accumulating evidence that some TE subfamilies form sets of functional CREs, no well-defined and genome-wide statistical framework has been proposed to estimate whether and how much TEs influence the expression of protein-coding genes. In addition, the identification of TE-embedded CREs currently relies on genome-wide epigenomic profiling, typically histone marks, TF binding, enhancer RNA (eRNA) production and chromatin accessibility (Bourque et al., 2008; Kunarso et al., 2010; Pehrsson et al., 2019; Sundaram et al., 2017; Trizzino et al., 2018). While these assays are instrumental to characterize exhaustively the involvement of TEs as CREs under specific biological contexts, performing them in pair with RNA-seq considerably increases experimental costs as well as the biological material required prior to sequencing. Thus, a statistical framework based on RNA-seq alone and capable of estimating which TE subfamilies serve as CREs would benefit the gene regulation research field for hypothesis generation and data interpretation at negligible additional costs.

The hypothesis that TEs influence the expression of protein-coding genes at the subfamily level has a corollary: one should be able to estimate the contribution of TEs to the expression of protein-coding genes by formulating a TE-centric mathematical model of gene regulation from basic principles of gene regulation. Analogous models have been developed to estimate the regulatory activity of TFBS motifs using transcriptomic data (Balwierz et al., 2014; Bussemaker, Foat, & Ward, 2007; Bussemaker, Li, & Siggia, 2001; FANTOM Consortium et al., 2009). These statistical approaches assume that DNA motifs or sequences - typically corresponding to TFBS - may regulate all promoters within which they are present with a quantitatively similar effect on gene expression. By analogy, as TEs evolved to attract the TFs necessary to trigger their own mobility, they can be conceptualized as larger regulatory sequences denoted as TE-embedded regulatory sequences (TEeRS). Thus, we took inspiration from the model of gene regulation championed by Britten and Davidson (Britten & Davidson, 1971) and hypothesized that phylogenetically related TE integrants may attract similar sets of transcriptional regulators and hence bear a similar regulatory influence on protein-coding genes located in their vicinity. Our system, coined ***craTEs*** (***cis***-**r**egulatory **a**ctivities of **T**ransposable **E**lement **s**ubfamilies), models variations in gene expression as a linear function of the susceptibility of protein-coding genes to the *cis*-regulatory activity of TE subfamilies. Here, we define activity as the variation in gene expression which can be attributed to the presence of integrants belonging to a set of phylogenetically related TEs within *cis*-regulatory distance of the gene promoter. In this study, we assume *a priori* that TE subfamilies form said sets. *craTEs* thereby enables the identification of *cis*-regulatory TE subfamilies from RNA-seq data alone, rooting it in the expression profile of protein-coding genes. Thus, *craTEs* adheres to a strict definition of *cis*-regulatory activity which requires an associated change in gene expression, in contrast with approaches restricted to profiling biochemical activity at TE loci (Jacques et al., 2013; Pehrsson et al., 2019; Sundaram et al., 2014; Sundaram et al., 2017; Trizzino et al., 2018).

In this study, we first show that *craTEs* accurately identifies *cis*-regulatory TE subfamilies from RNA-seq data alone. We demonstrate that it achieves this feat agnostically with respect to TE-derived transcription, with increased statistical power compared with standard

enrichment-based approaches, and in cases where changes in transcription at the corresponding subfamilies remain undetectable. We then leverage *craTEs* in conjunction with a large-scale TF perturbation RNA-seq dataset to estimate the maximal genomic distance up to which *cis*-regulatory TEs measurably contribute to the regulation of transcription genome-wide. Using the same dataset complemented with TF binding profiles and context-relevant TF knockout (KO) studies, we then identify novel regulatory links between TF expression and *cis*-regulatory TE activities throughout embryogenesis. Finally, we verify that *craTEs* detects biologically relevant regulatory phenomena by performing DNA binding, histone mark and chromatin accessibility profiling experiments. Overall, we present and validate *craTEs*, a simple mathematical model of TE-dependent gene regulation. *craTEs* recapitulates the findings of landmark case studies of TE-dependent *cis*-regulation and suggests previously unappreciated regulatory ties implicating TFs and TEs, particularly during and beyond gastrulation. These results support a model of GRN evolution whereby the spread of TEs provides an important supply of raw regulatory materials.

## 3.3 Results

### 3.3.1 *craTEs* models variations in gene expression as a linear combination of TE-encoded *cis*-regulatory elements

Using RNA-Seq data, we aimed to systematically uncover TE subfamilies that regulate the expression of protein-coding genes in *cis*. Integrants of the same TE subfamily share a high level of sequence similarity. Thus, they are predicted to exert a similar *cis*-regulatory influence on protein-coding genes located in their vicinity, for example through the simultaneous recruitment of a specific set of transcriptional regulators at multiple genomic loci. We thus assumed that the subfamily composition of TE integrants located within *cis*-regulatory distance of protein-coding genes contributes to a discernible fraction of the variation in gene expression (fig.3.1A) (Balwierz et al., 2014; Bussemaker et al., 2007). As a first approach, we set this distance to 50kb since differentially expressed (DE) genes were found to be enriched

within this range of epigenetically perturbed *cis*-regulatory LTR5-Hs and SVA TE subfamilies (Pontis et al., 2019b).

Considering two experimental conditions denoted as 1 and 2, for example "control cells" and "cells with transgene overexpression", we modeled the variation in gene expression $\Delta E$ of each of the $p$ protein-coding genes as a linear combination of the per-subfamily TE integrant counts $N_{pm}$ located within *cis*-regulatory distance of its promoters (see Methods). $N_{pm}$ represents the regulatory susceptibility (Bussemaker et al., 2007) of gene $p$ to TE subfamily $m$. We trade biological complexity for statistical simplicity by treating members of the same TE subfamily as "regulatory black boxes" of equal *cis*-regulatory potential. A well-known caveat of currently available ERE annotations is that integrants are often fragmented into multiple sequences (Friedli & Trono, 2015; Turelli et al., 2020), causing an artificial inflation of $N_{pm}$ and potentially deteriorating model performances. We therefore merged closely located (<100bp) ERE fragments of the same subfamily into single ERE integrants. LINEs, LTRs and SVAs were particularly prone to spurious fragmentation (fig. 3.1B), with numbers of integrants dropping by 8.3%, 7,9% and 15% respectively after merging. To define the regulatory susceptibility $N_{pm}$ of each gene $p$ to each TE subfamily $m$, we counted the number of integrants of subfamily $m$ falling within *cis*-regulatory distance of the promoters of $p$. We found that between 45.9% (LTRs) and 72.5% (SVAs) of all integrants were located within *cis*-regulatory distance, i.e. 50kb up/downstream, of at least one protein coding gene promoter (fig. 3.1B). In rare instances, TEs overlap gene exons. Since these are used to quantify RNA-seq reads, this may introduce a spurious association between the presence of an annotated TE integrant and gene expression. We addressed this by excluding TEs overlapping exons from the set of putatively *cis*-regulatory TEs susceptible to regulate the corresponding gene. Finally, we chose to emphasize TE-driven *cis*-regulation dependent on distal sequences, i.e. located more than 1.5kB up/downstream of a transcription start site, as the role of TEs as alternative promoters has been extensively studied elsewhere (Jang et al., 2019; Miao et al., 2020). Thus, we prevented TEs overlapping with promoters of gene $p$ from contributing to the set of regulatory susceptibilities $N_{pm}$ (fig. 3.1A-B). The combination of the last two filtering steps excluded 1.2% of TEs found within

*cis*-regulatory distance of protein-coding genes from $N$ (fig. 3.1B).

The main purpose of *craTEs* is the estimation of $\Delta A_{m\,2-1}$ which we define as the difference in *cis*-regulatory activity exerted by subfamily $m$ between conditions 1 and 2 (see equation 3.1). For the purpose of this study, we chose the convention that a positive *cis*-regulatory activity refers to an "enhancer-like" effect in condition 2 with respect to condition 1. Conversely, a negative *cis*-regulatory activity may reflect either the gain of a "silencer-like" effect or the loss of an "enhancer-like" effect in condition 2 versus condition 1. The *cis*-regulatory activity $\Delta A_{m\,2-1}$ has an intuitive interpretation: it is the quantity in expression that would be gained by any gene in condition 2 with respect to condition 1 upon insertion of an integrant of subfamily $m$ within *cis*-regulatory distance of one of its promoters. An independently and identically distributed Gaussian noise term centered around zero $\epsilon_{2+1}$ captures the variation in gene expression that is not accounted for by the linear model, and represents the sum of the noise terms corresponding to gene expression in each condition.

$$\Delta E_{p,2-1} = \Delta A_{0,2-1} + \sum_{m} N_{pm} \Delta A_{m,2-1} + \epsilon_{2+1} \tag{3.1}$$

*craTEs* estimates the vector of *cis*-regulatory TE subfamily activities $\Delta \hat{A}_{2-1}$ by minimizing the squared difference between the observed logged expression values $\Delta E_{2-1}$ and those modeled as linear combinations of the columns of the susceptibility matrix $N$, containing the regulatory susceptibilities $N_{pm}$. Further, *craTEs* assesses whether there is statistical evidence that $\Delta \hat{A_{m}}_{2-1}$ differs from zero: each component of $\Delta \hat{A}_{2-1}$ is tested against the null hypothesis $H_0 : \Delta A_{m\,2-1} = 0$, i.e. that there is no difference in activity between condition 1 and 2 for subfamily $m$, by means of a t-test (see Methods). This provides a measure of statistical significance for the estimated differences in TE-dependent *cis*-regulatory activities between conditions 1 and 2.

### 3.3.2 *craTEs* uncovers *cis*-regulatory TE subfamilies from RNA-seq data

We then assessed the ability of *craTEs* to detect *cis*-regulatory TE subfamilies under controlled experimental settings. In particular, we leveraged three RNA-seq datasets derived from experiments in which specific TE subfamilies were epigenetically silenced or activated, thus ablating their *cis*-regulatory effect on neighboring protein-coding genes (Deniz, Ahmed, Todd, Dawson, & Branco, 2019; Fuentes, Swigut, & Wysocka, 2018a; Pontis et al., 2019a). These datasets provide a biological "ground truth" against which we evaluated the output of *craTEs*. The targeted epigenetic modulation of specific genomic loci was achieved by means of the CRISPR interference or activation systems (Gilbert et al., 2013). CRISPRa/i relies upon a catalytically dead Cas9 domain (dCas9) that binds to DNA sequences complementary to user-defined guide RNAs (gRNAs). Once bound to the DNA, the dCas9-fused KRAB domain elicits the local deposition of repressive histone marks, thereby suppressing any enhancer activity exerted by the target site (CRISPRi). Conversely, the dCas9-fused VPR transactivator domain recruits a diverse set of transcriptional activators encompassing histone acetyltransferases upon DNA binding, thereby stimulating enhancer activity at the target site (CRISPRa) (Chavez et al., 2015; Memedula & Belmont, 2003). As TEs of the same subfamily exhibit high levels of sequence similarity, hundreds of related integrants can be targeted for activation/silencing by only a handful of carefully designed gRNAs (Fuentes et al., 2018b; Pontis et al., 2019b; Pontis et al., 2022). We have previously shown that the hominoid-specific LTR5-Hs and SVA TE subfamilies serve as enhancers in naïve hESCs, and that this *cis*-regulatory activity can be ablated by CRISPRi (Pontis et al., 2019b). We reanalyzed RNA-seq data from naïve hESCs where large fractions of the LTR5-Hs and SVA subfamilies were epigenetically silenced via CRISPRi across two independent experiments, each through a distinct guide RNA (g#1 and g#2). We applied *craTEs* to the vector $\Delta E_{CRISPRi-control}$ containing the differences in gene expression between each CRISPRi experiment and control naïve hESCs. The association between the differences in gene expression $\Delta E_{CRISPRi-control}$ and the *cis*-regulatory susceptibilities $N$ of promoters to TE subfamilies was statistically significant (g#1: p-val = $5.47 \cdot 10^{-146}$, F-test, g#2: p-val = 0, F-test), strongly suggesting an interrelation between changes in expression observed at protein-coding

genes and the genomic distribution of integrants for at least a subset of all TE subfamilies. After correcting for multiple testing using the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995), we uncovered 13 (g#1), resp. 39 (g#2) TE subfamilies with statistically significant differences in *cis*-regulatory activity (fig. 3.1C, Table S1), i.e. non-zero $\Delta A_{m,2-1}$ activity coefficients. Among these, LTR5-Hs, SVA B, C and D subfamilies displayed the largest and most statistically significant absolute estimated *cis*-regulatory activities. The negative activity values reflect the abrogation of the enhancer effect exerted by LTR5-Hs and SVAs in naïve hESCs by the CRISPRi system, and are best interpreted as the $log2$ fold-change in protein-coding gene expression attributable to the presence of a single integrant from the corresponding subfamily near the promoter of the given gene. Specifically, the expression of any given gene bearing an LTR5-Hs integrant in its *cis*-regulatory window decreases by an estimated fold-change contained within the 95% confidence interval $2^{-0.133\pm(1.96\cdot0.0095)} = [0.90; 0.92]$ i.e. by approximately 10% upon CRISPRi using g#1 (Table S1). We then applied *craTEs* to RNA-seq data generated from the CRISPRi-mediated repression of LTR5-Hs, LTR5A and LTR5B in the NCCIT human embryonal carcinoma cell line (Fuentes et al., 2018a) and found that LTR5-Hs and LTR5B showed the largest and most statistically significant absolute differences in *cis*-regulatory activity (fig. 3.1C), with the related LTR5A subfamily also displaying a weaker yet also statistically significant difference (Table S1). Conversely, *craTEs* uncovered LTR5-Hs, LTR5A and LTR5B as the TE subfamilies with the largest and most statistically significant absolute difference in *cis*-regulatory activity when applied to RNA-seq derived from NCCIT cells subjected to LTR5-Hs/LTR5A/LTR5B CRISPRa (fig. 3.1C), this time with positive activities mirroring the increased enhancer effect exerted by CRISPRa-targeted LTR5-Hs/LTR5A/LTR5B on neighboring genes. Thus, *craTEs* correctly inferred gains and losses of enhancer effect at the subfamilies targeted by CRISPRa/i and did so from the expression of protein-coding genes alone.

As it is well established that TEs are particularly active in hESCs (Macfarlan et al., 2012; Theunissen et al., 2016), we wondered whether *craTEs* would be able to recover TE-dependent *cis*-regulatory changes in other cellular contexts. A subset of LTR2B elements are marked

by the enhancer histone mark H2K27ac in various leukemia cell lines, including in chronic myelogenous leukemia-derived K562 cells (Deniz et al., 2020). We used *craTEs* to estimate the differences in TE-driven *cis*-regulatory activities between K562 cells where LTR2B were repressed via CRISPRi and their control counterparts. *craTEs* correctly identified LTR2B as significantly less active in LTR2B-CRISPRi K562 cells compared to control K562 cells (fig. 3.1C). Thus, *craTEs* recovers TE-dependent *cis*-regulatory mechanisms beyond the context of hESCs.

Next, we empirically verified whether the ability of *craTEs* to detect changes in *cis*-regulatory TE activity generalized beyond experiments of targeted TE repression via CRISPRi. TEs often exert *cis*-regulatory effects by serving as docking platforms for TFs. For example, the core pluripotency TF KLF4 is highly expressed in naïve hESCs, where it binds to LTR7, LTR5-Hs and SVA integrants (Pontis et al., 2019b). Interestingly, these subfamilies also display elevated levels of the enhancer histone mark H3K27ac in naïve hESCs. In contrast, primed hESCs generally express lower levels of KLF4 and TEs than their naïve counterparts (Pontis et al., 2019b; Theunissen et al., 2016). Using *craTEs*, we assessed the impact of KLF4 overexpression on TE-dependent *cis*-regulation in primed hESCs. *craTEs* identified LTR7, LTR5-Hs and SVA D as the most statistically significant and highly activated TE subfamilies upon KLF4 overexpression, thereby recapitulating our previous findings (Pontis et al., 2019b) agnostically with respect to epigenomics data and TE-derived transcripts (fig. 3.1C). Interestingly, we previously observed that the KLF4-dependent enhancer activity of SVAs in primed hESCs did not correlate with increased SVA transcription (fig. S3.1A) but instead with an accumulation of H3K27ac enhancer histone marks at SVA integrants (Pontis et al., 2019b). This suggests that *craTEs* detects TE-dependent *cis*-regulatory effects that would not be inferred from studying the variation in expression of TE integrants. Furthermore, overexpression of the repressive SVA-binder KRAB-zinc finger protein ZNF611 (Imbeault et al., 2017) in naïve hESCs abrogates the enhancer activity of SVAs (Pontis et al., 2019b). We used *craTEs* to estimate the differences in TE-dependent *cis*-regulation between ZNF611-overexpressing and control naïve hESCs. As expected, *craTEs* identified SVAs as the TE subfamilies with the most statistically significant and largest absolute differences in *cis*-regulatory activity between the two settings (fig.

3.1C), with negative activity values reflecting the loss of enhancer effect at SVAs upon ZNF611 overexpression. Of note, the proportion of the variance in gene expression explained by the distribution of TEs across *cis*-regulatory windows [2% - 12%, adj. $R^2$] (fig. 3.1D) overlapped with the typical fraction of explained variance reported upon modeling gene expression as a function of the distribution of TF binding motifs at gene promoters (Balwierz et al., 2014). Together, these results show that *craTEs* correctly identifies TE-dependent *cis*-regulatory activity changes beyond the context of targeted TE epigenetic perturbations and demonstrate its utility for identifying TE-dependent regulatory mechanisms under biological perturbations that affect TEs indirectly. In addition, *craTEs* identifies *cis*-regulatory TE subfamilies without resorting to mapping RNA-seq reads emanating from transcriptionally active TEs or performing epigenomics assays.

### 3.3.3 *craTEs* outperforms enrichment approaches based on differential expression analyses

The notion that differences in gene expression may reveal candidate *cis*-regulatory TEs has already been exploited in previous studies (Lynch et al., 2011; Pontis et al., 2019b) though the statistical methodologies differ from *craTEs* in key aspects. More specifically, these methods identify *cis*-regulatory TEs through a two-step process. First, differentially expressed (DE) genes are identified through ad-hoc statistical methods (Love, Huber, & Anders, 2014; M. D. Robinson, McCarthy, & Smyth, 2010). Then, per-subfamily scores for the enrichment of differentially expressed genes in the vicinity of TE integrants are computed. A high enrichment is reflected by a small probability (p-value) of finding more DE genes in the vicinity of a specific subfamily than the observed number of DE genes. We empirically compared the output of *craTEs* with that of the enrichment approach on the RNA-seq dataset whereby LTR5-Hs/SVA were silenced via CRISPRi (Pontis et al., 2019a). Using the enrichment approach, we found that DE genes whose expression fell under LTR5-Hs/SVA epigenetic repression (fig. S3.1B, p-val <0.05, Fisher's exact test, lenient DE calling) were statistically significantly enriched in the vicinity of LTR5-Hs, SVA B and SVA D integrants (Table S2, BH-adj. p-val <0.05, hypergeometric

test). Note that the DE enrichment approach failed to detect the regulatory link between gene downregulation and the TE subfamily SVA C (adj. p-val = 1, hypergeometric test), whereas these were identified by *craTEs* (fig. 3.1C, Table S1). Moreover, when correcting for multiple testing during differential expression analysis (fig. S3.1B, BH-adj. p-val <0.05, Fisher's exact test, stringent DE calling), DE genes were enriched near SVA D integrants, but not LTR5-Hs or other SVA subfamilies (Table S3). These results indicate that *craTEs* is more sensitive than DE enrichment approaches in the task of detecting *cis*-regulatory TE subfamilies from the expression of protein-coding genes. To assess whether this came at the cost of decreased specificity, we quantified the ability of *craTEs* - resp. the DE enrichment approach - to recover a ground truth set of *cis*-regulatory TE subfamilies upon CRISPRi-mediated repression of LTR5-Hs and SVAs either using g#1 or g#2 (fig. 3.1C, fig. S3.1D). Since no epigenomic data was available for g#1 (Pontis et al., 2019a), we leveraged ATAC-seq to generate chromatin accessibility profiles in naïve hESCs subjected to g#1-mediated CRISPRi against LTR5-Hs/SVAs. We then defined ground truth *cis*-regulatory TE subfamilies for each gRNA as those (1) with integrants directly targeted by the gRNA (LTR5-Hs, SVA A-F) (2) with enrichment for decreased ATAC-seq (g#1: {LTR5-Hs, SVA A-D}, Table S4; g#2: {LTR5-Hs, SVA A-F}, Table S5) and/or increased H3K9me3 upon CRISPRi (g#2: {LTR5-Hs, SVA A-F}, Table S6). We used the subfamily-specific BH-adjusted p-values computed according to *craTEs*, the lenient and the stringent DE enrichment approaches to classify subfamilies into two classes - *cis*-regulatory versus not *cis*-regulatory and subsequently computed the area under the receiver operating characteristic curves (AUCs) (fig. S3.1C). *craTEs* displayed higher AUCs than DE enrichment approaches for both gRNAs (g#1: 0.996 vs. {0.800, 0.600}, g#2: 1.0 vs. {0.85, 0.88}), noting that the low rates of true positives versus true negatives may partially explain the elevated AUCs. Overall, this suggests that by pooling information across all genes, and not just DE genes, *craTEs* offers increased statistical power over classical DE enrichment approaches in the task of identifying *cis*-regulatory TE subfamilies. Moreover, this emphasizes that *cis*-regulatory subfamilies as identified by *craTEs* agree with those displaying an enrichment based on differential context-matched epigenomic data.

*craTEs* estimates *cis*-regulatory TE activities by considering expression variations across hundreds of protein-coding genes. Consequently, *craTEs* does not require replicates to estimate TE subfamily *cis*-regulatory activities. To illustrate this, we reanalyzed the RNA-seq data derived from LTR5-Hs/SVA CRISPRi experiments (Pontis et al., 2019a). We treated each pair of LTR5-HS/SVA CRISPRi and control samples as a single experiment, in effect ignoring the information provided by the replicate structure. We applied *craTEs* to each of the four replicates for both gRNAs (fig. S3.1D). Consistent with our findings while accounting for replicates (fig. 3.1C), LTR5-Hs and SVA subfamilies collectively exhibited a statistically significant decrease in *cis*-regulatory activity upon CRISPRi across replicates, though not all of them passed the significance threshold. In addition, classifying subfamilies as *cis*-regulatory based on the measure of statistical significance reported by *craTEs* (BH-adj. p-val, t-test) yielded AUCs ranging from 0.87 to 0.99. Thus, whereas the discovery power of *craTEs* grows together with the number of replicates, the method can still uncover statistically significant changes in the *cis*-regulatory activities of TEs even in the absence of replicates. In contrast, any DE enrichment approach requires at least three samples due to the prerequisites of the DE analysis methods (Love et al., 2014; M. D. Robinson et al., 2010), and therefore cannot perform better than a random classifier in the absence of replicates (AUC = 0.5). In addition, *craTEs* not only quantifies the statistical significance of TE subfamily *cis*-regulatory activities but also provides a measure of the effect size through the estimated coefficient $\Delta \hat{A_{m,2-1}}$ which can be interpreted as the $log2$ fold-change in gene expression that would affect any gene upon insertion of an integrant from subfamily $m$ within its *cis*-regulatory window (fig. 3.1A). Overall, this case study suggests that *craTEs* is more powerful and more informative than DE-based enrichment approaches to discover *cis*-regulatory TE subfamilies from RNA-seq data, supporting the notion that TEs act as *cis*-regulatory fine-tuners, the dynamics of which may be overlooked when restricting the analysis to DE genes only.

### 3.3.4 Influential TE-embedded *cis*-regulatory information resides up to 500kb from gene promoters

In a first implementation of *craTEs*, we defined *cis*-regulatory regions as 50kb-long stretches of DNA directly adjacent to the 5' and 3' sides of protein-coding gene promoters. Though informed by previous work (Pontis et al., 2019b), this choice of genomic distance was based upon data corresponding to LTR5-Hs and SVA subfamilies in hESCs only and may not reflect the general range of action of *cis*-regulatory TEs across all subfamilies and cellular contexts. We therefore modified the *craTEs* model to weight the regulatory influence of each integrant $i$ on each gene $p$ as a continuous and decreasing function of its distance $d_{p,i}$ to the closest promoter of $p$ (fig. 3.2A). We defined the regulatory susceptibility $N_{pm}$ as:

$$N_{pm} = \sum_{i \in \{m,c\}} e^{-\frac{-d_{p,i}^2}{2L^2}} \tag{3.2}$$

where each weight was computed using a gaussian kernel applied to the integrant-promoter distances $d_{p,i}$. We considered all combinations of genes and integrants located on the same chromosome $c$. Note that integrants falling within exons or promoters of $p$ were excluded from $N_{pm}$. We computed 11 susceptibility matrices $N$ by varying the bandwidth of the gaussian kernel $L$ between 1 kilobase (kb) and 10 Gigabases (Gb) thus spanning the entire range of possible *cis*-regulatory distances. Setting $L$ to 1kb restricts *cis*-regulatory regions to the direct vicinity of gene promoters. In contrast, at 10 Gb, $L$ exceeds the length of human chromosomes by two orders of magnitude, thus yielding nearly equal regulatory susceptibility scores across genes located on the same chromosome (fig. S3.2A). We then tested which of these 11 matrices led to the smallest prediction error using 5-fold cross-validation. For the LTR5-Hs/SVA epigenetic repression, ZNF611 overexpression, KLF4 overexpression (Pontis et al., 2019a) and LTR2B epigenetic repression experiments (Deniz, Ahmed, Todd, Dawson, & Branco, 2019), the validation error was minimized for $L$ = 100kb or $L$ = 500kb (fig. 3.2B). As 95% of the area under a gaussian curve is contained within two standard deviations from its

mean (fig. S3.2B), this suggests that TEs encode discernible *cis*-regulatory information up to distances of approximately 200kb to 1 Million bases (Mb) from gene promoters. We note that errors estimated for small (1kb) and very large (>= 100Mb) values of $L$ were unstable due to the high degree of collinearity between predictors. Indeed, a small $L$ results in high numbers of zero-inflated columns in the $N$ matrix. Conversely, very large values of $L$ yield nearly equal weights for TE-gene pairs located on the same chromosome (fig. S3.2A). Both cases make the least squares problem ill-posed by making the matrix $N$ singular.

We wondered whether the optimal *cis*-regulatory bandwidths estimated from the four datasets treated thus far (fig. 3.2B) generalized to other TE subfamilies as well. We took advantage of a recently published RNA-seq dataset where hundreds of transgenes, mostly TFs, were overexpressed in primed hESCs through a dox-inducible system (Nakatake et al., 2020) (fig. 3.2C). We considered this dataset as a "perturbome" where each overexpressed transgene polarizes the primed hESC transcriptome towards a specific direction, e.g. towards the naïve hESC GRN or down a differentiation path. We used the same 5-fold cross-validation scheme to find the optimal value of $L$ for each transgene overexpression experiment (fig. 3.2D), this time comparing the prediction error for gene expression between *cis*-regulatory weight assignment informed by versus agnostic to hESC-specific topologically associating domains (TADs) (Dixon et al., 2015). In 436/441 transgene overexpression experiments, the optimal bandwidth $L$ took values between 50kb and 500kb. As most of the area below a gaussian curve is contained within two bandwidths from its mean (fig. S3.2B), TE subfamilies encode *cis*-regulatory information up to distances comprised between 100kb and 1Mb from the promoters of protein-coding genes in hESCs. In 187/441 transgene overexpression experiments, $L$ = 250kb led to the smallest cross-validation error, the majority of which (163/187) did not benefit from TAD-informed *cis*-regulatory weight assignment (fig. 3.2D), although the performance gap separating TAD-agnostic versus TAD-informed TE subfamily activity estimation was modest, as illustrated for GATA6, KLF4 and NEUROG1. Thus, TAD-agnostic *cis*-regulatory weight assignment according to a bandwidth of $L$ = 250kb can be chosen to weight *cis*-regulatory TE integrants such as to maximize predictability. As an illustration, with $L$ = 250kb, a TE located

250kb away from a gene promoter receives a weight of 0.61 (fig. S3.2B). The weight drops to 0.01 for a TE-promoter distance of 750kb resulting in a virtually negligible contribution to the *craTEs* model.

A predictor matrix $N$ based on TE contributions weighted by their distance to protein coding genes (fig. 3.2A) has two potential advantages over a predictor matrix $N$ computed from hard distance thresholds as we did when first validating the discovery power of *craTEs* (fig. 3.1A). First, the quality of the predictors is likely to improve, as the optimal distance until which *cis*-regulation affects gene expression is estimated directly from the expression data. In other words, a continuous and decreasing weighting function may better represent the regulatory potential of TEs on protein-coding genes than a hard threshold approach. Second, as we require that each TE subfamily included in $N$ sums up to a total regulatory potential greater than 150 (see the Methods section), the continuously decreasing weighting approach may allow for the inclusion of more TE subfamilies in the columns of $N$, leading to the discovery of previously overlooked statistically significant *cis*-regulatory TE subfamilies. We used the KLF4 overexpression RNA-seq dataset we previously generated (Pontis et al., 2019a) to illustrate these points. We replaced the regulatory susceptibilities $N_{pm}$ of matrix $N$ computed according to a hard distance threshold (fig. 3.1A) with those corresponding to the same subfamilies, this time computed either through TE-promoter distance weighting ($L$ = 250kB) or according to an approximately equivalent hard-thresholded *cis*-regulatory window width of 500kB (fig. 3.2A, eq. 3.2). All three models thus use the exact same number of predictors, i.e. cover the same TE subfamilies. Running *craTEs* with the weighted matrix $N$ computed with $L$ = 250kb increased the fraction of gene expression variation explained from 4.5% to 5.4% compared to using the matrix $N$ derived from 100kB-wide hard-thresholded *cis*-regulatory windows. As the number of predictors in $N$ remained unchanged, this suggests that the distance weighting approach better approximates the *cis*-regulatory potential of TE subfamilies than the hard distance threshold approach. Notably, as 500kB-wide hard-thresholded *cis*-regulatory windows explained 5.2% of the variance in gene expression, most of the increase in explained variance observed under the weighted ($L$ = 250kB) versus unweighted model (fig. 3.1) likely stems

from considering more distant TEs as putatively *cis*-regulatory. Next, we empirically evaluated whether allowing for the inclusion of TE subfamilies that passed the minimum per-subfamily regulatory potential with distance weighting (eq. 3.2) - but not with hard distance thresholding - would uncover additional biologically validated TE-dependent *cis*-regulatory changes. LTR7Y was identified as the most statistically significantly activated subfamily upon KLF4 overexpression in hESCs (fig. 3.2E), in agreement with previously published results (Pontis et al., 2019b) while it was absent from the model specified through hard distance thresholding (fig. 3.1C, Table S1). In addition, though also absent from the hard distance thresholding model, the primate-specific MER41G subfamily was found as statistically significantly and strongly activated in the distance-weighted model. Regarding the LTR5-Hs/SVA CRISPRi, ZNF611 overexpression and LTR2B CRISPRi experiments, using the distance-weighted matrix $N$ still uncovered LTR5-Hs/SVAs, LTR2B, resp. SVAs as differentially *cis*-regulatory (fig. S3.2E). To sum up, TE subfamilies typically encode *cis*-regulatory potential up to distances of approx. 500kb from the promoters of protein-coding genes, at least in the context of hESCs. This reinforces the notion that TEs form a layer of regulatory fine-tuners exerting a measurable impact on the expression of protein-coding genes.

### 3.3.5 TFs controlling gastrulation and organogenesis promote the *cis*-regulatory activity of evolutionarily young TE subfamilies activated during pluripotency

Having validated the ability of *craTEs* to agnostically recover well-established cases of TE-dependent *cis*-regulatory activities (Deniz et al., 2020; Fuentes et al., 2018b; Pontis et al., 2019b), we next aimed at characterizing the landscape of TF-induced TE-dependent *cis*-regulation in primed hESCs. As the epigenome of hESCs is markedly more open than that of differentiated cells (Hawkins et al., 2010), the number and strengths of the TF-TE regulatory interactions constituting the GRN of hESCs can be understood as upper bounds on those constituting the GRNs of differentiated tissues. We therefore applied *craTEs* to the "perturbome" dataset, where 441 transgenes, most of them TFs, were individually overexpressed in primed hESCs

for 48 hours through a dox-inducible system (fig. 3.2C) (Nakatake et al., 2020). Using the regulatory susceptibility matrix $N$ computed according to the best performing *cis*-regulatory bandwidth ($L$ = 250kb, fig. 3.2D), we estimated the changes in *cis*-regulatory TE activities associated with each dox-induced transgene overexpression experiment (additional file 1, Table S7). Dox-treatment alone and dox-induced GFP overexpression were not associated with any robust statistically significant change in *cis*-regulatory TE activity (fig. 3.3A, fig. S3.3A-B, Table S7) suggesting that neither the addition of doxycycline nor the metabolic cost entailed by strong transgene overexpression measurably altered the *cis*-regulatory activity of TE subfamilies. Interestingly, overexpression of the core pluripotency TF POU5F1 (also known as OCT4) was not associated with differential TE *cis*-regulatory activity (fig. S3.3A-B), suggesting that overexpressing an already highly expressed gene, namely POU5F1, in a cellular context that largely relies on it, i.e. primed hESCs, may not necessarily alter TE-dependent *cis*-regulation. Together, these results suggest that TE-dependent *cis*-regulatory activities inferred from the remaining transgene overexpression experiments are not driven by technical factors inherent to the system used but induced by the overexpressed transgene itself.

To reveal how TE-dependent *cis*-regulation relies on TF/transgene overexpression in hESCs, we performed hierarchical clustering on the matrix containing the statistical strengths of the estimated *cis*-regulatory TE activities (additional file 2, fig. S3.3C). Stratifying subfamilies according to size and evolutionary age (Pontis et al., 2019b) did not reveal any discernible bias regarding the distribution of statistically significant *cis*-regulatory activities (fig. S3.2C-D). Additionally, the directionality and statistical significance of TE-dependent *cis*-regulatory activities were robust to varying the *cis*-regulatory bandwidth $L$ (fig. S3.4, Table S7) and consistent across replicates when analysed invididually (fig. S3.5). Overexpressed TFs of the same family tended to cluster together - e.g NEUROD1, NEUROD2, NEUROG3, NEUROD4; PAX2, PAX5, PAX8; SNAI1, SNAI2, SNAI3; RUNX1, RUNX3; HES1, HEY1; LHX1, LHX5; GATA1, GATA2, GATA3 - whether considering effect size (fig. 3.3A, fig. S3.4) or statistical significance (fig. S3.3C) of the estimated differences in TE-dependent *cis*-regulatory activity. This suggests that commonalities in gene expression (Nakatake et al., 2020) likely driven by shared DNA

binding motifs (Ambrosini et al., 2020) were partially mirrored by similar *cis*-regulatory TE activity patterns. Experiments where the core trophoblast TF CDX2 (Bernardo et al., 2011; Strumpf et al., 2005) was overexpressed clustered away from all other experiments according to statistical significance (additional file 2, fig. S3.3C) but less so according to *cis*-regulatory activity estimates (fig. 3.3A). This may reflect both the widespread rewiring of TE-dependent *cis*-regulation as primed hESCs differentiate towards trophectodermal cells (Chuong et al., 2013) and a bias towards the detection of more differentially active *cis*-regulatory TE subfamilies in the CDX2 overexpression experiments due to a larger sample size compared to the other experiments (fig. S3.5) (Nakatake et al., 2020).

The LTR7 subfamily clustered away from all other TE subfamilies (additional file 2) and was statistically significantly less active in 186 out of 441 transgene overexpression experiments, making it the most frequently differentially active TE subfamily in this dataset (fig. S3.2D). Among overexpressed transgenes tied to a decrease of LTR7-dependent *cis*-regulatory activity, we found multiple TFs involved in post-implantation developmental stages (fig. 3.3A, Table S7), e.g. the meso/endodermal master TF GATA6 (Molkentin, 2000) and several homeobox-domain-containing TFs including PDX1 and RUNX1. LTR7 *cis*-regulatory activity also decreased upon overexpression of organ and tissue-specific TFs, e.g. NEUROD1, NEUROD2, NEUROD3, NEUROD4, MYT1, NR2E1, POU4F1, POU4F3, all involved in the formation of the nervous lineage (W. Li et al., 2008; Matsushita et al., 2017; Vahava et al., 1998; Vasconcelos et al., 2016; M. Zou, Li, Klein, & Xiang, 2012). Overexpression of TBX5, a key TF in the developing heart (Horb & Thomsen, 1999) also decreased the *cis*-regulatory activity of LTR7. Lastly, overexpression of TFs involved in the development and maintenance of the placenta e.g. CDX2, TEAD4 (Yagi et al., 2007) also led to a decrease of LTR7 *cis*-regulatory activity. Overall, inducing TFs tied to development and differentiation dampened the pluripotency-specific activity of LTR7 elements.

In contrast, we found rare transgenes (9/441) whose overexpression in primed hESCs led to an increase in LTR7 *cis*-regulatory activity (additional file 1, fig. 3.3A, Table S7). Overexpressing KLFs collectively increased the *cis*-regulatory activity of LTR7Y elements in agreement with

previous studies characterizing KLFs as inducers of LTR7Y enhancer activity in naïve hESCs (Pontis et al., 2019b). Interestingly, induction of KLF5 - but not of KLF1, KLF2 and KLF4 - increased the *cis*-regulatory activity of LTR7, matching previously reported visual inspections which revealed that among these, only KLF5-overexpressing cells retained an ESC-like phenotype 72 hours post-induction (Nakatake et al., 2020). By leveraging a large compendium of homogeneously reprocessed ChIP-seq data (Z. Zou, Ohta, Miura, & Oki, 2022), we confirmed that KLF4 binding was enriched at LTR7 and LTR7Y in various contexts related to primed hESCs (Lyu, Rowley, & Corces, 2018; Pontis et al., 2019b) (fig. 3.3B). MYB (also known as c-MYB), a TF involved in the maintenance of self-renewal in stem cells of the intestinal crypt, the bone marrow and the nervous system (Ramsay & Gonda, 2008) as well as the formation of stem-like memory CD8 T cells (Gautam et al., 2019), led to a marked increase in LTR7 *cis*-regulatory activity upon induction and displayed a modest enrichment in binding at related LTR7C integrants in monocytic-derived THP1 cells (Armenteros-Monterroso et al., 2019) (fig. 3.3B). Thus, MYB overexpression may reinforce self-renewal in the GRN of hESCs, a process tied to an increase in LTR7 *cis*-regulatory activity. More provocatively, this hints at the possible involvement of a MYB-LTR7 axis in the maintenance of self-renewal and stemness in the adult hematopoietic system. Our analysis thus suggests that a limited set of TFs linked to development and stemness may rely upon the enhancer potential of the LTR7 subfamily to establish, regulate and maintain these processes throughout development and adult life.

Other primate-specific TE subfamilies displayed partially overlapping patterns of *cis*-regulatory activity upon transgene overexpression. SVAs and LTR5-Hs were collectively activated by KLF4 and other KLFs (fig. 3.3A, Table S7) and enriched for KLF4 binding in hESCs (fig. 3.3B), consistent with previous work establishing the KLF4-dependent enhancer activity of these subfamilies in naïve hESCs (Pontis et al., 2019b). Interestingly, the *cis*-regulatory activity of LTR5-Hs and SVAs also increased upon overexpression of TFAP2C and NR5A1, both of which polarize hESCs towards the naïve state (Pastor et al., 2018; Yamauchi et al., 2020). Overall, these results suggest that recently emerged TE subfamilies form functional collections of enhancer-like CREs during pre-gastrulation embryogenesis.

We then wondered whether the overexpression of transgenes necessary for embryonic development during and after gastrulation was associated with an increase in *cis*-regulatory activity in recently emerged TE subfamilies. Overexpression of the core meso/endodermal TF GATA6 as well as other GATA family members increased the *cis*-regulatory activity of SVAs and LTR5-Hs (fig. 3.3A, Table S7), thereby resulting in the activation of a TE-dependent *cis*-regulatory network partially reminiscing that of naïve hESCs (Pontis et al., 2019b; Theunissen et al., 2016). This is surprising given that naïve hESCs resemble cells of the early blastocyst while GATA6 controls post-implantation developmental stages such as the formation of the mesoderm and the endoderm during gastrulation. Furthermore, overexpressing GATA family members increased the *cis*-regulatory activity of additional primate-specific TE subfamilies including the LTR5-Hs-related HERV-K subfamily LTR5B and the ERV1 subfamilies LTR6A, LTR6B, PRIMA4-LTR, MER4A1, MER4D and MER4D1. Importantly, LTR6B displayed the largest and most statistically significant increase in TE-dependent *cis*-regulatory activity along stem cell to endoderm differentiation across two independent datasets (fig. 3.3D, S3.3F) (Heslop, Pournasr, Liu, & Duncan, 2021a; Luo, Huangfu, & Beer, 2022; Luo et al., 2023). Moreover, GATA6 KO (Heslop et al., 2021a) reduced LTR6B *cis*-regulatory activity in differentiating endodermal cells, whereas GATA6 re-expression rescued it (fig. 3.3D). Along the same lines, GATA2 deletion in hematopoietic progenitor cells (Huang, Du, Shi, Chen, & Pan, 2017) decreased the *cis*-regulatory activity of LTR5-Hs and LTR5B (fig. S3.3D). Of note, GATA ChIP-seq peaks (Luo et al., 2022; Z. Zou et al., 2022) were strongly enriched at SVAs, LTR5-Hs, LTR5B, LTR6A, LTR6B, PRIMA4-LTR, MER4A1, MER4D and MER4D1 integrants across primitive streak-derived (Q. V. Li et al., 2019) as well as mesendodermal (Luo et al., 2022), mesodermal (Tsankov et al., 2015) - including blood-derived (Canver et al., 2017; Gertz et al., 2013; Mazumdar et al., 2015; Xu et al., 2012) - and placental lineage (Krendl et al., 2017) cells, with enriched binding at SVAs, LTR5-Hs, LTR5B, LTR6A and LTR6B extending to the endodermal lineage (Chia et al., 2015; Verzi et al., 2010) (fig. 3.3B). Together, these patterns of binding suggest that the changes in *cis*-regulatory activity observed upon GATA overexpression in hESCs and meso/endodermal differentiation result from the direct binding of GATA family members to primate-specific TE subfamilies. Interestingly, overexpressing EOMES, another regulator of germ layer formation

and mesoendodermal differentiation (Tosic et al., 2019), markedly increased the *cis*-regulatory activity of LTR6B elements (fig. 3.3A), at which it also displayed enriched binding in hESC-derived mesendodermal cells (Luo et al., 2022) (fig. 3.3B). Moreover, overexpression of SOX17, an additional regulator of endodermal differentiation (Séguin, Draper, Nagy, & Rossant, 2008), increased the *cis*-regulatory activity of LTR5-Hs, SVA-C and LTR6B (fig. 3.3A), while SOX17 binding was strongly enriched at SVAs, LTR7 and MER4A1 in germ cell-derived Tcam-2 cancer cells (Jostes et al., 2020) (fig. 3.3B), which share some phenotypic features with primordial germ cells (PGCs).

Evidence linking transcription during post-gastrulation embryogenesis with primate-specific TE-mediated *cis*-regulation extended beyond SOX17 to TFAP2C and SOX15, which both display elevated expression in the PGC lineage (F. Guo et al., 2015). Specifically, overexpression of SOX15 in hESCs markedly increased the *cis*-regulatory activity of LTR5-Hs (fig. 3.3A, Table S7), the latter exhibiting the largest statistically significant increase in *cis*-regulatory activity in human PGC-like cells (hPGCLCs) compared with cognate hESC-derived somatic cells (fig. 3.3C) (X. Wang et al., 2021a). Knocking out SOX15 in hPGCLCs led to a drop in LTR5-Hs *cis*-regulatory activity across two biological replicates (fig. 3.3C, fig. S3.3E), while SOX15 binding was strongly enriched at LTR5-Hs in hPGCLCs (fig. 3.3B). Additionally, inducing TFAP2C in hESCs increased the *cis*-regulatory activity of LTR5-Hs and SVAs (fig. 3.3A), both of which displayed a considerable enrichment in TFAP2C binding in hPGCLCs (D. Chen et al., 2019) and, interestingly, in cells of the ectoderm lineage (Lal et al., 2013; L. Li et al., 2019) (fig. 3.3B). In summary, evolutionarily recent and pre-implantation specific TE subfamilies form sets of CREs that regulate the expression of protein-coding genes in *cis* well past the epiblast stage, including during and after gastrulation, as evidenced firstly by increased *cis*-regulatory activity following germ layer-specific TF overexpression in hESCs, secondly by enriched TF binding in cells derived from the corresponding germ layers and thirdly by substantial stage-specific increases in *cis*-regulatory activity which were reverted upon germ layer-specific TF KO.

Older TE subfamilies that emerged prior to the speciation of primates also contribute to GRNs by donating CREs. Despite having spread before the speciation of amniotes hundreds of

millions of years ago, AmnSINE1 elements are retained in the genomes of extant amniotes including humans and mice (Hirakawa, Nishihara, Kanehisa, & Okada, 2009), and some AmnSINE1 elements were found to exert long-range enhancer effects on genes controlling brain development (Sasaki et al., 2008). We observed that in primed hESCs, overexpression of several homeobox domain-containing TFs, e.g. RUNX1, a regulator of hematopoietic ontogeny (M. J. Chen, Yokomizo, Zeigler, Dzierzak, & Speck, 2009), and PDX1, involved in pancreatic development (D'Amour et al., 2006), was associated with an increased AmnSINE1 *cis*-regulatory activity (fig. 3.3A, Table S7). Interestingly, AmnSINE1 elements are enriched within active enhancers in epigenomes derived from fetal human cell lines (Pehrsson et al., 2019). Lastly, MER135, an ancient subfamily of currently unidentified origin (Kojima, 2018) showed increased *cis*-regulatory activity upon overexpression of homeobox domain-containing TFs in primed hESCs (fig. 3.3A). More generally, these results hint that ancient TE subfamilies may retain their *cis*-regulatory potential at the subfamily level in extant species despite having colonized the genome of an evolutionarily distant common ancestor.

### 3.3.6 *Cis*-regulatory activities are more pronounced at epigenetically active TEs

We showed that *craTEs* agnostically uncovers SVAs and LTR5-Hs as the subfamilies with the most statistically significant and strongest loss of *cis*-regulatory activity upon CRISPRi-mediated epigenetic repression in naïve hESCs (fig. 3.1). However, it is highly likely that only a fraction of all integrants constituting a subfamily truly exert *cis*-regulatory effects. For example, integrants found within dynamic chromatin regions may be more differentially active than integrants located in stable chromatin regions. To empirically verify this hypothesis, we leveraged the epigenomic profiles matched to the LTR5-Hs/SVA CRISPRi RNA-seq dataset (Pontis et al., 2019a) and labeled the following integrants as "functional": those overlapping with genomic coordinates where loss of chromatin accessibility (ATAC-seq) or gain of the repressive histone mark H3K9me3 (ChIP-seq) were detected upon epigenetic repression of SVAs and LTR5-Hs. Conversely and by complementarity, we considered all other integrants from these subfamilies as "non-functional". We then expanded the weighted susceptibility matrix

$N$ ($L$ = 250kb) column-wise by splitting TE subfamilies into complementary functional and non-functional integrant subsets (fig. 3.4A). Finally, we used *craTEs* to jointly estimate the differences in *cis*-regulatory activity for functional and non-functional subsets of TE subfamilies upon epigenetic repression of LTR5-Hs and SVAs in naïve hESCs (fig. 3.4B). Functional subsets of SVAs and LTR5-Hs subfamilies displayed greater decreases in *cis*-regulatory activity upon epigenetic repression than complementary non-functional subsets. In addition, the estimated decrease in *cis*-regulatory activity was more pronounced for the subset of functional LTR5-Hs than that estimated for the corresponding unsplit subfamily. Of note, functional LTR5-Hs and SVA integrants tended to show slightly lower mappability scores (Sexton & Han, 2019) than non-functional integrants (fig. S3.7A), raising the concern that difficulties in ChIP-seq and/or ATAC-seq read assignment at repeats (Goerner-Potvin & Bourque, 2018) may drive functional versus non-functional integrant calling, thereby biasing TE subfamily *cis*-regulatory activity estimates. However, low mappability functional LTR5-Hs/SVA integrants still exhibited larger and more statistically significant decreases in estimated *cis*-regulatory activity upon CRISPRi than their low mappability non-functional counterparts (fig. S3.7A).

Next, we leveraged the epigenomics-informed adaptation of *craTEs* to test whether differences in TF binding or histone marks could single-out integrants with detectable changes in *cis*-regulatory activity in the context of TF overexpression experiments. To this end, we completed the matched transcriptomics and histone profiles available for KLF4 and ZNF611 overexpression in hESCs (Pontis et al., 2019a) by generating ChIP-seq profiles against KLF4 and ZNF611. We used *craTEs* to estimate differences in *cis*-regulatory activity for functional integrant subsets defined according to differences in histone marks or TF binding and focused on the main *cis*-regulatory TE subfamilies identified under KLF4 and ZN611 overexpression in hESCs, namely LTR7, LTR5-Hs and SVAs (fig. 3.4C). For both KLF4 and ZN611 overexpression experiments, histone mark-defined functional subsets had greater *cis*-regulatory activity than non-functional subsets, except for SVA-D under KLF4 overexpression, as well as SVA-A under ZNF611 overexpression. However, histone mark-defined functional subsets generally displayed only modest increases in *cis*-regulatory activity over unsplit subfamilies, at the cost

of a marked decrease in statistical significance. In contrast, TF-bound-defined functional integrants displayed increased *cis*-regulatory activity compared with the unsplit subfamily for all subfamilies except SVA-D under KLF4 overexpression, with increased significance in most cases, including when matching integrants for mappability prior to functional versus non-functional subfamily splitting (fig. S3.7B). Turning to cellular contexts featuring endogenous TF expression levels, we compared the usefulness of mesendodermal GATA6, EOMES and H3K27ac ChIP-seq peaks (Luo et al., 2022) for delineating the most active fraction of TE subfamilies in terms of *cis*-regulation. GATA6 and EOMES binding, whether considered individually or in combination, proved remarkably informative for discriminating *cis*-regulatory LTR6B integrants from their inactive counterparts during endoderm differentiation (fig. 3.4D, fig. S3.6A), and conversely aptly accounted for the decrease, resp. increase, in LTR6B, PRIMA4-LTR, MER4D, MER4D1 and LTR5-Hs observed upon GATA6 KO, resp. rescue, in iPSC-derived endodermal cells (Heslop et al., 2021a) (fig.3.4D). In contrast, H3K27ac peaks failed to single-out *cis*-regulatory LTR6B integrants in this context. Indeed LTR6B integrants devoid of the canonical enhancer histone mark exerted a more statistically significant *cis*-regulatory activity than LTR6B integrants overlapping H3K27ac peaks (fig. S3.6A). Similarly, SOX15 CUT&TAG peaks (X. Wang et al., 2021a) pinpointed LTR5-Hs integrants displaying increased *cis*-regulatory activity in hPGCLCs versus cognate somatic cells, a trend that was inverted upon upon SOX15 KO in hPGCLCs (fig. 3.4E). As implied by differing patterns in statistical significance at functional versus non-functional LTR5-Hs integrants, SOX15 binding better isolated *cis*-regulatory active LTR5-Hs than chromatin accessibility as assessed by ATAC-seq, although both epigenomic signals were in general agreement (fig. S3.6B). Importantly, we did not observe any noticeable difference in mappability between functional and non-functional integrants in ZNF611 overexpressing cell, endodermal cell (Luo et al., 2022) or hPGCLC-derived (X. Wang et al., 2021a) epigenomic data (fig. S3.7 C-E). Thus, TF binding appears to better single-out bona fide *cis*-regulatory integrants than changes in histone marks.

As both EOMES and GATA6 binding were associated with increased LTR6B, LTR6A, LTR5-Hs, MER4D, MER4D1 and PRIMA4-LTR *cis*-regulatory activity in the differentiating endoderm (fig.

3.3B), we performed multiple sequence alignment (MSA) of subfamily-restricted ChIP-seq peaks in search of EOMES and GATA6 TF binding motifs. MSA of GATA6 and EOMES peaks at LTR6A and LTR6B revealed consensus sequences containing several canonical GATA binding sites (fig. 3.4F, additional file 3). Fittingly, the DNA sequence of most LTR6B integrants adorned with GATA6/EOMES peaks contained recognizable GATA6 DNA binding motifs (fig. S3.6C), as assessed through motif search (Grant, Bailey, & Noble, 2011). Of note, MSA of EOMES peaks revealed additional GATA6 motifs in per-subfamily consensus sequences at LTR5-Hs, MER4D, MER4D1 and PRIMA4-LTR (additional file 3), while we failed to call consensus sequences from GATA6 ChIP-seq peaks at these subfamilies. This may stem from differences in ChIP-seq protocols and/or quality. Intriguingly, MSA of EOMES-bound TE regions did not reveal any canonical EOMES DNA binding site, a finding supported by the absence of EOMES motifs at LTR6B integrants subjected to motif search (fig. S3.6C). This suggests that GATA6 may first bind at primate-specific TE subfamilies in the differentiating endoderm, to then promote EOMES recruitment at these loci. Accordingly, LTR6B and MER4D1 integrants carrying GATA6 DNA-binding motifs were more *cis*-regulatory than those devoid of the motif in the differentiating endoderm, whether considering statistical significance or effect size (fig. S3.6D). The presence of a GATA6 motif also separated truly *cis*-regulatory from non *cis*-regulatory LTR6B integrants upon GATA6 KO/rescue in the differentiating endoderm but proved less discriminant than ChIP-seq-derived GATA6/EOMES binding at LTR6A, LTR5-Hs, MER4D and MER4D1 (fig. S3.6D). Overall, the agreement between GATA6/EOMES, resp. SOX15 binding, and primate-specific TE-mediated *cis*-regulation in endodermal fetal cells, resp. hPGCLCs, suggests that recently evolved TEs spread functional *cis*-regulatory platforms at which core TFs controlling post-gastrulation embryogenesis directly bind, in turn affecting protein-coding gene expression.

## 3.4 Discussion

The notion that some TE-derived sequences behave as bona fide CREs is supported by an ever-growing number of reports mostly relying on genome-wide profiles of promoter- or

enhancer-specific histone marks. However, whether that biochemical activity should be interpreted as evidence for an evolutionary process fostering the emergence of collections of CREs to the benefit of the host, or instead as a byproduct of the so-called "selfish" tendency of TEs for genome invasion is subject to debate (Bourque et al., 2018; Chuong et al., 2017; Feschotte, 2008). Still, if TEs truly spread functional CREs that become co-opted by the host through natural selection, one should at least be able to capture their effect on gene expression by modeling TE-dependent *cis*-regulation from basic gene regulation principles. Thus, we formulated *craTEs*, a system where differences in TE subfamily *cis*-regulatory activities are estimated in a single step from protein-coding gene expression. Using RNA-seq data derived from thoroughly characterized cases of TE-dependent *cis*-regulation, we showed that *craTEs* correctly identifies differentially active *cis*-regulatory TE subfamilies. Moreover, we could refine activity estimations by incorporating context-matched epigenomics data - e.g. TF binding or chromatin marks - into *craTEs*, highlighting that protein-coding gene expression and chromatin states at TEs fruitfully complement each other for uncovering TE-dependent *cis*-regulation. Crucially, *craTEs* does not rely on TE-derived reads and is thus well-suited for the post-hoc analysis of standard RNA-seq count tables that did not take TE transcripts into account during feature quantification. In addition, *craTEs* was able to identify *cis*-regulatory TE subfamilies (SVAs) in RNA-seq datasets where no difference in transcriptional activity for these TE subfamilies was previously detected (Pontis et al., 2019a). These results suggest that TE subfamilies form at least partially consistent sets of CREs modulating gene expression in a coordinated fashion genome-wide and more generally that TEs spread highly resembling and functional *cis*-regulatory sites thereby supplying the raw materials critical to the evolution of coordinated gene regulation. Note that *craTEs* cannot discriminate between waves of transposition whereby TE subfamilies spread sequences poised to acquire TFBS by gradual mutations from those whereby extant TFBS were intrinsic constituents of *de novo* transposed integrants. The former may be most relevant for older TE subfamilies e.g. AmnSINE1, MER121 and MER135 (Fueyo et al., 2022). Whereas *craTEs* relies on sequence similarity as encoded in the TE models used by Repeatmasker, relatedness across subfamilies is currently not modeled: each subfamily is considered as phylogenetically equidistant to all others which may hamper

sensitivity. One possible extension to *craTEs* entails encoding sequence similarity across subfamilies as a nearest-neighbour graph to constrain closely related TE subfamilies to receive similar *cis*-regulatory activities via penalized regression, e.g. the fused lasso (Tibshirani, Saunders, Rosset, Zhu, & Knight, 2005). We noticed that *craTEs* explains a fraction of the variation in gene expression ranging from approx. 2% to 12%, which is comparable to the proportion of variance in gene expression explained by previously published linear models of gene regulation based on putative TF binding sites at core promoters (Balwierz et al., 2014). In both cases, the low proportion of variance captured by the linear models is still sufficient for identifying statistically significant and relevant regulatory mechanisms from transcriptomic data alone. However, while the true mathematical function underlying the regulatory mechanism at play is most likely non-linear (Balwierz et al., 2014), the linear model proposed in this work is still useful. Inferred activity coefficients are interpretable and well-established statistical tests exist to determine whether differences in *cis*-regulatory activities statistically significantly deviate from zero. What is perhaps more impressive is that, in the context of this study at least, TE-dependent *cis*-regulation accounts for a fraction of the variation in gene expression comparable to that inferred from models of gene regulation based on the TFBS repertoire of core promoters. This observation underlines that TEs should not be ignored when attempting to delineate the regulatory programs orchestrating biological processes, in particular during embryogenesis.

We have also shown that *craTEs* identifies relevant *cis*-regulatory TE subfamilies with superior power compared to enrichment approaches based on differential expression analysis. In addition, *craTEs* readily identifies TE-dependent *cis*-regulatory changes in experiments limited to a single pair of samples, e.g. treatment versus control, whereas performing differential expression analysis requires at least a replicate in one of the conditions, i.e. three samples. This difference likely stems from how gene expression values are modeled in either method. DE methods model the distribution of gene expression values across conditions independently for each gene, although current methods now leverage information borrowing techniques to share information across genes within samples (Love et al., 2014). In effect, DE methods perform one

statistical test for each gene, resulting in tens of thousands of tests where the false discovery rate has to be controlled. Thus, any coordinated but mild difference in expression between co-regulated sets of genes is lost and cannot be used in the subsequent enrichment test. In contrast, *craTEs* leverages information across hundreds to thousands of genes to estimate the *cis*-regulatory activity of each TE subfamily in a single step. We leveraged epigenomic data - namely ATAC-seq and H3K9me3 ChIP-seq - to supplement gRNA complementarity in defining a ground truth set of differentially *cis*-regulatory TE subfamilies under LTR5-Hs/SVA CRISPRi in naïve hESCs. *craTEs* performed remarkably similarly to enrichment approaches based on epigenomic data, picking up subtle differences between CRISPRi gRNAs. Indeed, while LTR5-Hs and SVA A-D were enriched for epigenomic signal indicative of heterochromatin under both g#1 and g#2, SVA E/F only did so under g#2. One putative limitation of *craTEs* compared with enrichment approaches stems from the fact that it estimates exactly one activity coefficient per TE subfamily. Consequently, *craTEs* may fail to recover *cis*-regulatory TE subfamilies encompassing integrants exerting antagonistic *cis*-regulatory effects. In that case, enrichment approaches may benefit from treating increased and decreased expression and/or epigenomic signal separately, though this would likely require large sample sizes.

Next, we empirically determined that the typical range until which *cis*-regulatory TEs regulate their target promoters is 500kb. To this end, we applied a cross-validation procedure to a large scale primed hESCs perturbation dataset to select the *cis*-regulatory distance that minimized the error between true and predicted gene expression values as estimated by *craTEs*. To our knowledge, this is the first attempt aimed at quantitatively estimating such distance by aggregating transcriptomic data derived from hundreds of experimental perturbations. Thus, TE subfamilies exert *cis*-regulatory influences up to distances compatible with those typically separating enhancers from their target promoters, consistent with the notion that many TEeRS are, in fact, bona fide enhancers. Additionally, constraining TE-promoter *cis*-regulatory weights to TAD boundaries yielded similar predictions to TAD-agnostic distance weighting, though generally not ameliorating gene expression prediction.

We further characterized the landscape of TF overexpression-induced TE-dependent *cis*-

regulatory changes. TFs poised towards the GRN of the naïve hESC state, namely TFAP2C, KLFs and NR5A1, collectively bound to and increased the *cis*-regulatory activity of LTR5-Hs, SVA and LTR7Y subfamilies, which function as KLF4-responsive enhancers in naïve hESCs. Whether these newly-identified inducers of evolutionarily young and *cis*-regulatory TE subfamilies mediate their effect via direct binding or secondary transcriptional changes will require further work. Still, these results underline the importance of LTR5-Hs, SVA and LTR7Y subfamilies in the GRN of naïve pluripotency. Of note, whereas KLF5 overexpression was accompanied by increased LTR7 *cis*-regulatory activity, overexpression of KLF1, KLF2 and KLF4 was associated with a decrease in LTR7 *cis*-regulatory activity. This apparent discrepancy with the established role of LTR7 as KLF4-responsive enhancers (Carter et al., 2022; Ohnuki et al., 2014; Pontis et al., 2019b) may stem from differences in overexpression levels or hESC backgrounds, and may be explained by the dual involvment of KLF4 in both naïve pluripotency (Pontis et al., 2019b) and terminal differentiation in mesodermal (Feinberg et al., 2007) as well as endodermal (Katz et al., 2002) lineages. While the *cis*-regulatory activity of young TE subfamilies in pre-implantation embryogenesis is increasingly being recognized (Fueyo et al., 2022), the landscape of TE-dependent *cis*-regulation at later stages of human embryogenesis is still ill-defined. In the present study, we observed that inducing key regulators of gastrulation, germ layer/placental commitment and PGC differentiation – including GATA2, GATA6, EOMES and SOX15 - in hESCs increased the *cis*-regulatory activity of LTR5-Hs and SVA subfamilies together with other primate-specific TE subfamilies such as LTR5B, LTR6B, MER4A1, MER4D/D1 and PRIMA4-LTR. Importantly, binding of these TFs was enriched at the TE subfamilies they activated across various models of embryogenesis and differentiated tissues. TF binding as assessed by context-matched ChIP-seq/CUT&TAG profiles aptly discriminated truly *cis*-regulatory from inactive integrants during endoderm and hPGCLC differentiation, as well as upon KO/rescue of the corresponding TFs. Of note, we reported in a related manuscript (Pontis et al., 2022) that LTR5B, LTR5-Hs, LTR6A and LTR6B integrants are highly accessible in endodermal and mesodermal human fetal cells, though more rarely in ectodermal cells and that selected LTR6B integrants serve as enhancers for genes encoding key mesendodermal regulators. Finally, MSA of GATA6 and EOMES-bound LTR6B regions in the differentiating endoderm revealed a

GATA-rich consensus sequence, and GATA6 DNA-binding motifs uncovered through motif search recapitulated the functional versus non-functional dichotomy defined using ChIP-seq data for predicting gene expression. Thus, the *cis*-regulatory role played by primate-restricted TEs during pre-implantation embryogenesis appears maintained - if not reactivated - by developmental stage-specific TFs during subsequent steps of embryogenesis.

Lastly, we leveraged epigenomics data to test whether changes in chromatin states and evidence for direct TF binding could single-out *cis*-regulatory integrants from non-*cis*-regulatory integrants within TE subfamilies. Surprisingly, we found that across various experimental systems entailing TF overexpression, TF KO and endogenous TF expression, TF binding was better able to enrich for *cis*-regulatory integrants than changes in histone marks and/or chromatin accessibility, though how TF binding compares with context-matched chromatin states as defined using combinations of histone marks (Ernst & Kellis, 2010) remains to be seen. In the case of KLF4 overexpression, it is possible that the partial activation of compensatory TE-silencing mechanisms caused a divergence between chromatin and TF binding-derived *cis*-regulatory signals. Lastly, *craTEs* may benefit from the incorporation of STARR-ChIP-seq data, though whether fragment length and genome coverage shall prove appropriate for studying TE-mediated *cis*-regulation will have to be assessed.

## 3.5 Conclusion

That a simple mathematical model based on TE-promoter distances and the expression of protein-coding genes can infer TE-mediated *cis*-regulatory activities illustrates that as TEs spread, they rewire nearby protein coding genes into a web of regulatory dependencies which can be simultaneously fine-tuned by only a handful of transcriptional regulators. Furthermore, these recently emerged GRN components appear to regulate not only early embryogenesis, but also more advanced stages of development. For such vital and highly conserved events, the resulting speciation is only mechanistic owing to selective pressures. In those cases, the TE-dependent and species-specific CRE turnover is likely to result in equivalent phenotypic

adaptations across species, as reproductive/survival stakes leave little room for organismal novelty. However, in situations allowing for more phenotypic diversification, for instance in the brain, the rapidly evolving TE-based *cis*-acting regulome likely contributes to the emergence of new traits.

## 3.6 Methods

### 3.6.1 Cell culture

**Treatment protocol** Primed H1 were transduced with GFP or KLF4-containing lentiviral vectors and split after 48h then selected using blasticydin for the 3 following days. Naïve WIBR3dPE hESC cells in KN/2iL media were transduced with GFP or ZNF611-containing lentiviral vectors, split after 96h, then selected for a couple of passages with blasticydin on irradiated Mouse Embryonic Blasticidin-resistant (MMMbz).

**Growth protocol** Conventional (primed) human ESC lines were maintained in mTSER for H1 (Male) on Matrigel, for WIBR3 (Female) on irradiated inactivated mouse embryonic fibroblast (MEF) feeders in human ESC medium (hESM) and passaged with collagenase and dispase, followed by sequential sedimentation steps in hESM to remove single cells while naïve ES cells and primed H1 were passaged by Accutase in single cells. hES media composition: DMEM/F12 supplemented with 15% fetalbovine serum, 5% KnockOut Serum Replacement, 2 mM L-glutamine, 1% nonessential amino acids, 1% penicillin-streptomycin (Lonza), 0.1 mM $\beta$-mercaptoethanol and 4 ng/ml FGF2. Naïve media composition: 500 mL of medium was generated by including: 240 mL DMEM/F12, 240 mL Neurobasal, 5 mL N2 supplement, 10 mL B27 supplement, 2 mM L-glutamine, 1% nonessential amino acids, 0.1 mM $\beta$-mercaptoethanol, 1% penicillin-streptomycin, 50 $\mu$g/ml BSA. In addition for KN/2i media: PD0325901 (1 $\mu$M), CHIR99021 (1 $\mu$M), 20 ng/ml hLIF and Doxycycline (2 $\mu$g/ml).

### 3.6.2 ChIP-seq

Cells were cross-linked for 10 minutes at room temperature by the addition of one-tenth of the volume of 11% formaldehyde solution to the PBS followed by quenching with glycine. Cells were washed twice with PBS, then the supernatant was aspirated and the cell pellet was conserved in -80°C. Pellets were lysed, resuspended in 1mL of LB1 on ice for 10 min (50 mM HEPES-KOH pH 7.4, 140 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 10% Glycerol, 0.5% NP40, 0.25% Tx100, protease inhibitors), then after centrifugation resuspend in LB2 on ice for 10 min (10 mM Tris pH 8.0, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA and protease inhibitors). After centrifugation, resuspend in LB3 (10 mM Tris pH 8.0, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% NaDOC, 0.1% SDS and protease inhibitors) for histone marks and SDS shearing buffer (10 mM Tris pH8, EDTA 1mM, SDS 0.15% and protease inhibitors) for transcription factor and sonicated (Covaris settings: 5% duty, 200 cycle, 140 PIP, 20 min), yielding genomic DNA fragments with a bulk size of 100-300bp. Coating of the beads with the specific antibody and carried out during the day at 4°C, then chromatin was added overnight at 4°C for histone marks while antibody for transcription factor is incubated with chromatin first with 1% Triton and 150mM NaCl. Subsequently, washes were performed with 2x Low Salt Wash Buffer (10 mM Tris pH 8, 1 mM EDTA, 150mM NaCl, 0.15% SDS), 1x High Salt Wash Buffer (10 mM Tris pH 8, 1 mM EDTA, 500 mM NaCl, 0.15% SDS), 1x LiCl buffer (10 mM Tris pH 8, 1 mM EDTA, 0.5 mM EGTA, 250 mM LiCl, 1% NP40, 1% NaDOC) and 1 with TE buffer. Final DNA was purified with Qiagen Elute Column. Up to 10 nanograms of ChIPed DNA or input DNA (Input) were prepared for sequencing. Library was quality checked by DNA high sensitivity chip (Agilent). Quality controlled samples were then quantified by picogreen (Qubit 2.0 Fluorometer, Invitrogen). Cluster amplification and following sequencing steps strictly followed the Illumina standard protocol. Libraries were ligated with Illumina adaptors. Sequenced reads were demultiplexed to attribute each read to a DNA sample and then aligned to reference human genome hg19 with bowtie2 (Langmead & Salzberg, 2012). Peaks were called on mapped data using MACS2 (Yong Zhang et al., 2008). Differential analysis between conditions has been performed with VOOM (Law et al., 2014) using unique reads (filter for MAPQ <10), counted on the union of all

peaks of a same experiment. Samples were normalized for sequencing depth using the counts on the union peaks as library size and using the TMM method (M. D. Robinson et al., 2010) as it is implemented in the limma package of Bioconductor.

### 3.6.3 ATAC-seq

ATAC-seq was performed as previously described (Buenrostro, Giresi, Zaba, Chang, & Greenleaf, 2013) on primed WIRB3 and WIBR3dPE; naive WIBR3 and WIBR3dPE in 4iLA and KN/2iL media respectively; and in WIBR3dPE in KN/2iL media upon dCAS9-KRAB overexpression containing or not a guide RNA targeting SVA/LTR5Hs. Library were made using Nextera DNA Library Prep Kit (Illumina #FC-121-1030). ATAC-seq and DNase-seq reads were mapped to the human (hg19) genome using bowtie2 (Langmead & Salzberg, 2012). Mitochondrial reads were removed. Then accessible sites were called using MACS2 (Yong Zhang et al., 2008), only peaks with a score higher than 5 (–log10 p value) were kept. Then differential analysis between conditions was done using unique reads (filter for MAPQ <10), counted on the union of all peaks of a same experiment.

### 3.6.4 RNA-seq analysis

#### Mapping

Reads were mapped to the human (hg19) genome using hisat2 (Kim et al., 2015) with parameters `hisat2 -k 5 -seed 42 -p 7`.

#### Summarization

Counts on genes and TEs were generated using featureCounts (Liao et al., 2014). To avoid read assignation ambiguity between genes and TEs, a gtf file containing both was provided to featureCounts. For repetitive sequences, an in-house curated version of the Repbase database was used (fragmented EREs belonging to the same subfamily were merged). Only uniquely mapped reads were used for counting on genes and TEs. Finally, features that did not have at

least one sample with 20 reads were discarded from the analysis. Only features corresponding to protein-coding genes were kept, except when quantifying SVA-derived transcription for fig. S3.1A. Gene expression values pertaining to endoderm differentiation (48h vs 24h) (Luo et al., 2022) were obtained from GEO at accession number GSE213394. Gene expression values pertaining to hPGCLC differentiation with and without SOX15 KO (X. Wang et al., 2021a) were retrieved using recount3 (Wilks et al., 2021) and gene symbols were converted from genome assemblies hg38 to hg19 using ensembl Biomart (Cunningham et al., 2022).

**Normalization**

For input into *craTEs*, raw counts were transformed to transcripts per millions (TPM). A pseudocount equal to the fifth percentile of non-zero counts in the sample was added to each raw count before transformation to TPM and subsequent $\log_2$ transformation. For recount3-retrieved expression values, raw counts were used for filtering and the pre-computed TPM values were used.

### 3.6.5 ChIP-seq enrichment at TE integrants

ChIP-seq binding locations from published datasets were extracted from ChIP-Atlas (Z. Zou et al., 2022), except for the Wang et al. hPGCLCs datasets (X. Wang et al., 2021a) for which we downloaded `.narrowPeak` files directly from GEO at accession number GSE143345 and the Luo et al. endodermal differentiation datasets (Luo et al., 2022) which were processed from `fastq` files as described above. Enrichment analysis over TE subfamilies was performed with HOMER software v4.10.4 (Heinz et al., 2010), except for the Wang et al. (X. Wang et al., 2021a) and Luo et al. (Luo et al., 2022) datasets, for which we used `pyTEnrich` available at URL https://github.com/alexdray86/pyTEnrich as previously described (J. C. De Tribolet-Hardy, 2022; J. De Tribolet-Hardy, Thorball, Forey, Planet, Duc, Khubieh, et al., 2023). To build fig. 3.3B, we recovered the three top statistically significant enrichments for each selected pairs of TE-TF - excluding WNT3A - highlighted in fig. 3.3A. Enrichment values with p-val >1e-10 were filtered out. Cell type and germ layer assignments were hand curated by examining the

original publications, retrieved from the SRA run numbers. When applicable, we excluded enrichment values derived from perturbation experiments - e.g., knock-down of a particular gene - and kept control samples instead. We excluded an H3K4me1 ChIP-seq sample that was erroneously labeled as a GATA1 ChIP-seq sample in ChIP-Atlas.

### 3.6.6 Differential expression analysis-based *cis*-regulatory TE subfamily detection

DE analysis was performed using edgeR (M. D. Robinson et al., 2010). Starting from raw counts restricted to protein-coding genes, we performed library size normalization with the trimmed mean of M-values (TMM) normalization method (Mark D Robinson & Oshlack, 2010). We assumed that TMM-normalized counts follow a negative binomial distribution and estimated per-gene dispersions using the `estimateDisp` function from edgeR. We tested for differential expression using Fisher's exact test as implemented in the function `exactTest` from edgeR. We either considered DE genes as those with Benjamini-Hochberg adjusted p values <0.05 (stringent DE calling), or those with p values <0.05 (lenient DE calling). Next, using the hypergeometric distribution, we computed for each TE subfamily the probability of finding more DE genes within *cis*-regulatory distance of its integrants than what was observed (Lynch et al., 2011; Pontis et al., 2019b). We performed this last step separately for upregulated and downregulated genes. Finally, we gathered the results obtained for up/downregulated genes into a single table and accounted for multiple testing using the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995). We assessed how *craTEs* compared versus the DE enrichment approach by measuring their respective abilities to recover a ground truth set of *cis*-regulatory TE subfamilies in each of both LTR5-Hs/SVA CRISPRi experiments (Pontis et al., 2019a). For each biological replicate, we defined the ground truth set using two criteria: (1) complementarity with the gRNA used in the corresponding CRISPRi experiment and (2) increased heterochromatin marks and/or decreased chromatin accessibility upon treatment with CRISPRi. The resulting ground truth sets were: [g#1: {LTR5-Hs, SVA-A, SVA-B, SVA-C, SVA-D}]. [g#2: {LTR5-Hs, SVA-A, SVA-B, SVA-C, SVA-D, SVA-E, SVA-F}]. As considering TE

subfamilies as "sufficiently" *cis*-regulatory depends upon statistical significance and/or effect size thresholds, we used AUCs to systematically compare *craTEs* with competing approaches in the task of recovering truly *cis*-regulatory TE subfamilies. We used 1-(BH adj. p-values) as the probability of being classified as *cis*-regulatory. The AUC takes values between 0.5 and 1 and can be interpreted as the probability of having correctly ordered observations between classes such as to separate observations across both classes perfectly. An advantage of the AUC is that it allows for a detailed study of the relationship between sensitivity and specificity as the threshold for classification varies. Here, a perfect AUC = 1 would be reached in cases where ranking the adj. p-values yielded by *craTEs* ranks all TE subfamilies found in the ground truth as those with the most statistically significant changes in *cis*-regulatory activity.

### 3.6.7 *cis*-regulatory activity estimation for TE subfamilies (*craTEs*)

The *craTEs* model, available as an `R` package at URL https://github.com/pulvercyril/crates, was adapted from the motif activity response analysis (MARA) model of gene regulation (Balwierz et al., 2014). Let $E$ be the matrix of gene expression, with $P$ protein coding genes as rows, and $S$ samples as columns. $E_{ps}$ is the logged TPM expression value for gene $p$ in sample $s$. Let $N$ be the predictor/feature matrix with $P$ protein coding genes as rows and $M$ TE subfamilies as columns. $N_{pm}$ is regulatory susceptibility (Bussemaker et al., 2007) of protein-coding gene $p$ to TE subfamily $m$, and in the absence of weighting procedure is computed as the number of times an integrant belonging to TE subfamily $m$ is found in the vicinity of $p$. Let $A$ be the matrix of *cis*-regulatory TE subfamily activities, with $M$ TE subfamilies as rows and $S$ samples as columns. $A_{ms}$ is the *cis*-regulatory activity of TE subfamily $M$ in sample $S$. $A_{ms}$ can be seen as follows: if a TE integrant from subfamily $m$ is inserted in the vicinity of gene $p$, the expression of gene $p$ increases by the value $A_{ms}$. Then, the expression $E_{ps}$ of gene $p$ in sample $s$ is given by:

$$E_{ps} = c_p + d_s + \sum_m N_{pm} A_{ms} + \epsilon \tag{3.3}$$

where $c_p$ is a gene-specific constant representing basal transcription and $d_s$ is a sample-specific constant that models sample-specific batch effects such as PCR amplification biases. The model across samples and genes can be written as:

$$\begin{aligned}\begin{bmatrix} E_{1s} & \dots & E_{1S} \\ . & . & . \\ E_{Ps} & \dots & E_{PS} \end{bmatrix} &= \begin{bmatrix} c_1 & \dots & c_1 \\ . & . & . \\ c_P & . & c_P \end{bmatrix} + \begin{bmatrix} d_1 & \dots & d_S \\ . & . & . \\ d_1 & \dots & d_S \end{bmatrix} \\ &+ \begin{bmatrix} N_{11} & \dots & N_{1M} \\ . & . & . \\ N_{P1} & \dots & N_{PM} \end{bmatrix} \begin{bmatrix} A_{1s} & \dots & A_{1S} \\ . & . & . \\ A_{Ms} & \dots & A_{MS} \end{bmatrix} \\ &+ \begin{bmatrix} \epsilon & \dots & \epsilon \\ . & . & . \\ \epsilon & \dots & \epsilon \end{bmatrix}\end{aligned} \tag{3.4}$$

Column-centering $E$ sets $d_s$ to zero for each sample. Similarly, row-centering $E$ sets $c_p$ to zero for each gene. After row and column centering, the model becomes:

$$\begin{bmatrix} E'_{1s} & \dots & E'_{1S} \\ . & . & . \\ E'_{Ps} & \dots & E'_{PS} \end{bmatrix} = \begin{bmatrix} N_{11} & \dots & N_{1M} \\ . & . & . \\ N_{P1} & \dots & N_{PM} \end{bmatrix} \begin{bmatrix} A'_{1s} & \dots & A'_{1S} \\ . & . & . \\ A'_{Ms} & \dots & A'_{MS} \end{bmatrix} + \begin{bmatrix} \epsilon & \dots & \epsilon \\ . & . & . \\ \epsilon & \dots & \epsilon \end{bmatrix} \tag{3.5}$$

where $E'_{ps}$ represents the deviation in expression from the average expression for gene $p$ across all samples and $A'_{ms}$ the deviation in *cis*-regulatory activity from the average *cis*-regulatory activity for gene $p$ across all samples. The model is allowed to have a non-zero intercept,

therefore the model we fit is in effect:

$$\begin{bmatrix} E'_{1s} & ... & E'_{1S} \\ . & . & . \\ E'_{Ps} & ... & E'_{PS} \end{bmatrix} = \begin{bmatrix} 1 & N_{11} & ... & N_{1M} \\ . & . & . & . \\ 1 & N_{P1} & ... & N_{PM} \end{bmatrix} \begin{bmatrix} A_0 s' & A'_{1s} & ... & A'_{1S} \\ . & . & . & . \\ A_0 s' & A'_{Ms} & ... & A'_{MS} \end{bmatrix} + \begin{bmatrix} \epsilon & ... & \epsilon \\ . & . & . \\ \epsilon & ... & \epsilon \end{bmatrix} \tag{3.6}$$

MARA (Balwierz et al., 2014) uses using ridge regression and selects the regularization parameter $\lambda$ using 5-fold cross validation. $\lambda$ controls for overfitting by imposing a so-called "budget" on TE activities. This method addresses the curse of dimensionality (too many predictors with respect to the number of observations) and stability issues arising when there is a high collinearity in the space of predictors. However, the statistical significance of each predictor is more difficult to compute than in the standard linear regression setting. Additionally, in the MARA model, each activity deviates from a mean activity corresponding to a baseline regulatory state which can be hard to describe in biological terms. Instead, we chose to consider samples in pairs. We contrasted samples from condition 2 (e.g. treatment samples) with samples from condition 1 (e.g. control samples). Under the normalized MARA-like model:

$$E'_{ps} = A_{0s} + \sum_m N_{pm} A'_{ms} + \epsilon \tag{3.7}$$

We are interested in contrasting two samples labeled sample 1 and sample 2.

$$E'_{p2} - E'_{p1} = A_{02} + \sum_m N_{pm} A'_{m2} + \epsilon_2 - \left( A_{01} + \sum_m N_{pm} A'_{m1} + \epsilon_1 \right) \tag{3.8}$$

Therefore, we obtain:

$$\Delta E'_{p,2-1} = \Delta A_{0,2-1} + \sum_m N_{pm} \Delta A'_{m,2-1} + \epsilon_{2+1} \tag{3.9}$$

We used (eq. 3.9) as a model with identically and independently normal-distributed noise to estimate differences in activity between treatment and control samples. We then tested whether each estimated activity was t-distributed around 0. We controlled the false discovery rate using the Benjamini-Hochberg procedure. Paired replicates were treated by concatenating the vectors of differences in expression $\Delta E'_{p,2-1}$ for each pair. The susceptibility matrix $N$ was expanded row-wise accordingly.

### 3.6.8 Computing the regulatory susceptibilities of each gene to TE subfamilies

The genomic locations of TEs was derived from Repeatmasker RELEASE 20170127, based on the hg19/GRCh37 assembly of the human reference genome. RepeatMasker annotates TEs based on sequence similarity to a consensus sequence which tends to fragment partially degenerated integrants into multiple sequences. To avoid counting fragmented TEs several times, we merged TEs belonging to the same subfamily and the same strand separated by a genomic distance of less than 100 bp. The following steps were applied to each protein-coding gene (derived from ENSEMBL release 93 using Biomart) to designate the set of corresponding putatively *cis*-regulatory TEs. We defined gene promoter regions as clusters of transcription start sites (derived from ENSEMBL release 93 using Biomart) spaced by less than 1kb and extended by 500bp at their 5' and 3' ends. Next, we defined *cis*-regulatory windows as the union of promoter regions extended by 50kb at their 5' and 3' end. We identified all TEs present within *cis*-regulatory windows. We excluded TEs overlapping promoter regions as well as TEs overlapping exons. Finally, the remaining TEs were summed per subfamily to generate a vector representing the susceptibility of the gene to putatively *cis*-regulatory TEs.

### 3.6.9 Building the susceptibility matrix N

The TE susceptibility matrix summarizes the potential regulatory activity of TE subfamilies on protein-coding genes. $N$ was built by grouping integrants by subfamilies and summing them for each gene. Therefore, $N_{i,j}$ describes the number of integrants belonging to subfamily $j$ in the *cis*-regulatory window of gene $i$.

### 3.6.10 Weighting *cis*-regulatory TEs by their distance to gene promoters

To circumvent the need for a hard distance threshold, we weighted the regulatory potential of integrants by the distance separating them from gene promoters. Let $K$ be the number of integrants of TE subfamily $m$ present on the same chromosome as gene $p$. The regulatory potential of subfamily $m$ on gene $p$ is weighted by a gaussian kernel: $N_{pm} = \sum_K w_{pk}$, $w_{pk} = e^{-\frac{x_{pk}^2}{2L^2}}$ where:

1. $x_{pk}$ is the distance in base pairs between the center coordinate of TE integrant $k$ and the center coordinate of the closest promoter of gene $p$

2. $L$ is the standard deviation (i.e. bandwidth) of the gaussian kernel, in base pairs

### 3.6.11 Filtering $E$ and $N$

Each experiment, defined as the set of treatment versus control expression vectors that will eventually form matrix $E$, was subjected to a separate filtering procedure. Genes with raw count values of less than 10 in all samples were removed from $E$. A per-column pseudo-count computed as the fifth percentile of all non-zero values in the column was added to each entry in $E$. $E$ was transformed to transcript per millions (TPMs) and then log2-transformed. $E$ was column-centered, and then row-centered. TE subfamilies with a sum of susceptibility scores $\sum_p N_{pm}$ smaller than 150 were removed from $N$. To avoid confounding bona fide *cis*-regulatory changes with differences in expression directly attributable to experimental perturbations, e.g. KO or overexpression, we filtered out experimentally perturbed genes from

$E$ when applicable.

### 3.6.12 Estimating the optimal TE-promoter regulatory distance

To estimate the optimal distance until which TE subfamilies regulate gene expression in *cis*, we built several weighted susceptibility matrices $N$ by varying the values of $L$ between $10^3$ and $10^{10}$ base pairs and estimated the mean validation error using a 5-fold cross-validation on the gene space. The optimal value of $L$ was chosen as the one that minimized the mean validation error. To ensure that the validation errors were comparable, we kept the sets of TE subfamilies and protein-coding genes fixed across all weighted matrices $N$. To this end, we filtered $E$ and $N$ according to the unweighted matrix $N$ built with 100kB-wide *cis*-regulatory windows centered on gene promoters, as described above and in fig. 3.1. We then filtered each weighted susceptibility matrix $N$ according to the rows (protein coding-genes) and columns (TE subfamilies) contained in the unweighted susceptibility matrix $N$.

### 3.6.13 Splitting TE subfamilies between functional and non-functional fractions

Let $F$ bet the set of genomic ranges considered as functional. Each TE integrant from subfamily $m$ overlapping with at least one element in $F$ was assigned to the so-called "functional" fraction of subfamily $m$: $m_{functional}$. The matrix $N_{functional}$ was built as described above for $N$, considering $m_{functional}$ as a distinct subfamily. As splitting subfamilies into fractions may yield predictors, i.e. columns of $N_{functional}$, with too few putatively regulated genes to reliably estimate TE subfamily *cis*-regulatory activities, we applied the following procedure:

- TE subfamilies (including their functional fractions) that were excluded by the filtering procedure applied on $N$ described above were also excluded from $N_{functional}$.

- If either the functional or the non-functional fraction of a TE subfamily showed $\sum_p N_{pm} < 100$, both fractions were removed and replaced with the corresponding column in $N$, i.e. the vector of regulatory susceptibility scores $N_{pm}$ for the entire subfamily.

- We allowed some user-specified subfamilies to be "protected" from this filtering step. These subfamilies remained split between a functional and a non-functional fraction in $N_{functional}$ irrespective of the sum of their regulatory susceptibility scores.

### 3.6.14 Per integrant mappability scores

Coordinates of TE integrants from our curated hg19 TE database were converted to hg38 using the UCSC utilitary tool `liftOver` (Hinrichs, 2006) and thereafter shifted in the 5' direction by half of the genomic distance covered by a single read (single end mappability) or between the 5' end of the forward read and the 5' end of the reverse read (paired end). Average mappability scores over each integrant were computed using the UCSC utilitary tool `bigWigAverageOverBed` (Kent, Zweig, Barber, Hinrichs, & Karolchik, 2010). Mappability scores for hg38 (Sexton, Tillett, & Han, 2022) were queried as `.BigWig` files from the UCSC website at URL https://genome-euro.ucsc.edu, using the genome browser custom track (Raney et al., 2014) information at URL https://raw.githubusercontent.com/HanLabUNLV/TEmappability/master/hub.txt. We defined "low mappability", resp. "high mappability" integrants as those scoring below, resp. above the median mappability in their subfamily.

### 3.6.15 Multiple sequence alignment plots

Multiple sequence alignment (MSA) plots were made as previously described (Iouranova et al., 2022). In short: DNA sequences for integrants belonging to the indicated subfamilies were aligned using MAFFT (Katoh & Standley, 2013) with parameters `-reorder -auto`, and then merged using the `-merge` option. Positions in the alignment (columns) with more than 85% gaps were greyed out. ChIP-seq signals are scaled for each integrant (row) to the [0,1] interval before being superimposed on the alignments. Averaged (scaled) ChIP-seq signals across all integrants are plotted on top of the alignments.

### 3.6.16 Motif search

FASTA sequences for integrants belonging to the indicated subfamilies were scanned using FIMO (Grant et al., 2011) with default parameters. We used a zero-order background model computed over all TEs.

### 3.6.17 Statistical methods

The statistical significance of TE subfamily activities is evaluated through null hypothesis significance testing via a standard t-test, where the null hypothesis is $H0$ : the value of the associated linear regression coefficient (often referred to as $\beta$) is zero. All p-values reported in the manuscript are adjusted for multiple testing using the Benjamini Hochberg procedure, except when specified in the methods or main text. We reject the $H0$ when the adj. p-value $\leq 0.05$.

### 3.6.18 Declarations

#### Competing interests

Authors have no conflict of interest to declare.

#### Ethics approval

hESC usage has been approved by the Swiss Federal Office of Public Health, the Canton of Vaud Ethics Committee (Autorization Number R-FP-S-2-0009-0000) and registered in the European Human Pluripotent Stem Cell Registry (hPSCreg). Experimental methods comply with the Helsinki Declaration.

#### Consent for publication

Not applicable.

**Availability of data and materials**

*craTEs* (Pulver, 2023) is available as an open source R package at URL https://github.com/pulvercyril/crates and is distributed under the MIT License. The repository was archived on ZENODO upon submission at URL https://doi.org/10.5281/zenodo.8407480.

The TE annotation database RepeatMasker library RELEASE 20170127 can be found on the RepeatMasker website accessible at URL http://repeatmasker.org/libraries/RepeatMaskerMetaData-20170127.tar.gz.

The following RNA-seq datasets: naïve hESCs + CRISPRi against SVA/LTR5-Hs, primed hESCs + GFP or KLF4, naïve hESCs + GFP or ZNF611; ATAC-seq dataset: naïve hESCs + CRISPRi against SVA/LTR5-Hs g#2; ChIP-seq datasets: H3K9me3 in naïve hESCs + CRISPRi against SVA/LTR5-Hs g#2, H3K9me3/H3K27ac in primed hESCs + GFP or KLF4, H3K9me3/H3K27ac in naïve hESCs + GFP or ZNF611 can be found on the Gene Expression Omnibus (GEO) under accession number GSE117395 at URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE117395 (Pontis et al., 2019a, 2019b).

The RNA-seq datasets of NCCIT + CRISPRa/i against LTR5-Hs, LTR5A and LTR5B can be found on GEO under accession number GSE111337 at URL https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE111337 (Fuentes et al., 2018a, 2018b).

The RNA-seq dataset of K562 + CRISPRi against LTR2B can be found on GEO under accession number GSE136763 at URL https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE136763 (Deniz, Ahmed, Todd, Dawson, & Branco, 2019; Deniz et al., 2020).

The RNA-seq dataset of transgene overexpression in hESCs can be found on the DNA Data Bank of Japan (DDBJ) Sequence Read Archive (DRA) under SRA submission number DRA006296 at URL https://ddbj.nig.ac.jp/resource/sra-submission/DRA006296 (Nakatake et al., 2020).

The following RNA-seq and ChIP-seq datasets: RNA-seq during hESC-derived endoderm differentiation, ChIP-seq against EOMES, GATA6 and H3K27ac in hESC-derived mesendoderm

can be found on GEO under accession number GSE213394 at URL https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE213394 (Luo et al., 2022; Luo et al., 2023).

The following RNA-seq datasets: RNA-seq during iPSC-derived endoderm differentiation, with/without GATA6 KO or rescue can be found on GEO at accession number GSE156021 at URL https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE156021 (Heslop et al., 2021a, 2021b).

The following RNA-seq dataset: RNA-seq in GATA2 KO HPCs can be found on GEO at accession number GSE69797 at URL https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE69797 (Huang et al., 2015; Huang et al., 2017).

The following RNA-seq, CUT&TAG and ATAC-seq datasets: RNA-seq on hESC-derived hPGCLCs, hESC-derived somatic cells, SOX15 KO hPGCLCs, CUT&TAG against SOX15 in hPGCLCs, ATAC-seq in hPGCLCs can be found on GEO at accession number GSE143345 at URL https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE143345 (X. Wang et al., 2021a, 2021b).

The following ChIP-seq and ATAC-seq datasets: ChIP-seq against KLF4 in primed hESCs, ChIP-seq against ZNF611 in naïve hESCs, ATAC-seq in naïve hESCs + CRISPRi against SVA/LTR5-Hs g#1 can be found on the Gene Expression Omnibus (GEO) under accession number GSE208403 at URL https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?&acc=GSE208403.

The regulatory susceptibility matrix N, TEs vs. promoters, 100kB-wide windows can be found on ZENODO at URL https://doi.org/10.5281/zenodo.6707955

The regulatory susceptibility matrix N, TEs vs. promoters, weighted with $L = 2.5e5$kb can be found on ZENODO at URL https://doi.org/10.5281/zenodo.8117257

The regulatory susceptibility matrices N with functional fractions, TEs vs. promoters, weighted with $L = 2.5e5$kb can be found on ZENODO at URL https://doi.org/10.5281/zenodo.8117285

The regulatory susceptibility matrices N, TEs vs. promoters, weighted with $L$ in $[1e3kB, 1e10kB]$

can be found on ZENODO at URL https://doi.org/10.5281/zenodo.8117286

The code used to process the data and generate the figures can be found at URL https://renkulab.io/gitlab/crates and can be executed directly from the renkulab platform for reproducible data science, or alternatively locally after downloading `docker` images:

- https://renkulab.io/projects/crates/klf4-znf611-sva-crispri
- https://renkulab.io/projects/crates/promoter-te-subfamilies-matrix
- https://renkulab.io/projects/crates/hescs-activities

**Competing interests**

The authors declare that they have no competing interests.

**Funding**

This work was funded by the Swiss National Science Foundation (SNSF) (FNS 310030_108803, FNS 310030_192613), the European Research Council (ERC) (ERC 694658), the Swiss Data Science Center (SDSC) (SDSC C19-02) and the Ludwig Institute for Cancer Research.

**Author's contributions**

CP designed the research plan, analyzed the data and wrote the manuscript with biological supervision by DT, JP and statistical supervision by RF. DT, JP and RF all made substantial contributions to the manuscript. JP generated the ChIP-seq/ATAC-seq data, and processed the ChIP-seq data retrieved from ChIP-Atlas. CP, DG, JD, SS and EP transformed the raw RNA-seq data into count tables and processed the ChIP-seq/ATAC-seq data and provided the corresponding paragraphs in the Methods section. JD provided the code to merge fragmented EREs, to perform multiple sequence alignment and provided the corresponding paragraphs in the Methods section. AC provided early access to pyTEnrich.

**Acknowledgements**

We thank Charlène Raclot and Sandra Offner for technical support regarding wet lab experiments; Romain Forey, Eunji Shin, Paola Malsot, Felix Naef and members of Johan Jakobsson's research group at Lund University for scientific discussions; Nicolas Barrière, Cyril Matthey-Doret and the whole renku team for technical support regarding the renku platform and Séverine Reynard for administrative assistance.

## 3.7 Figures

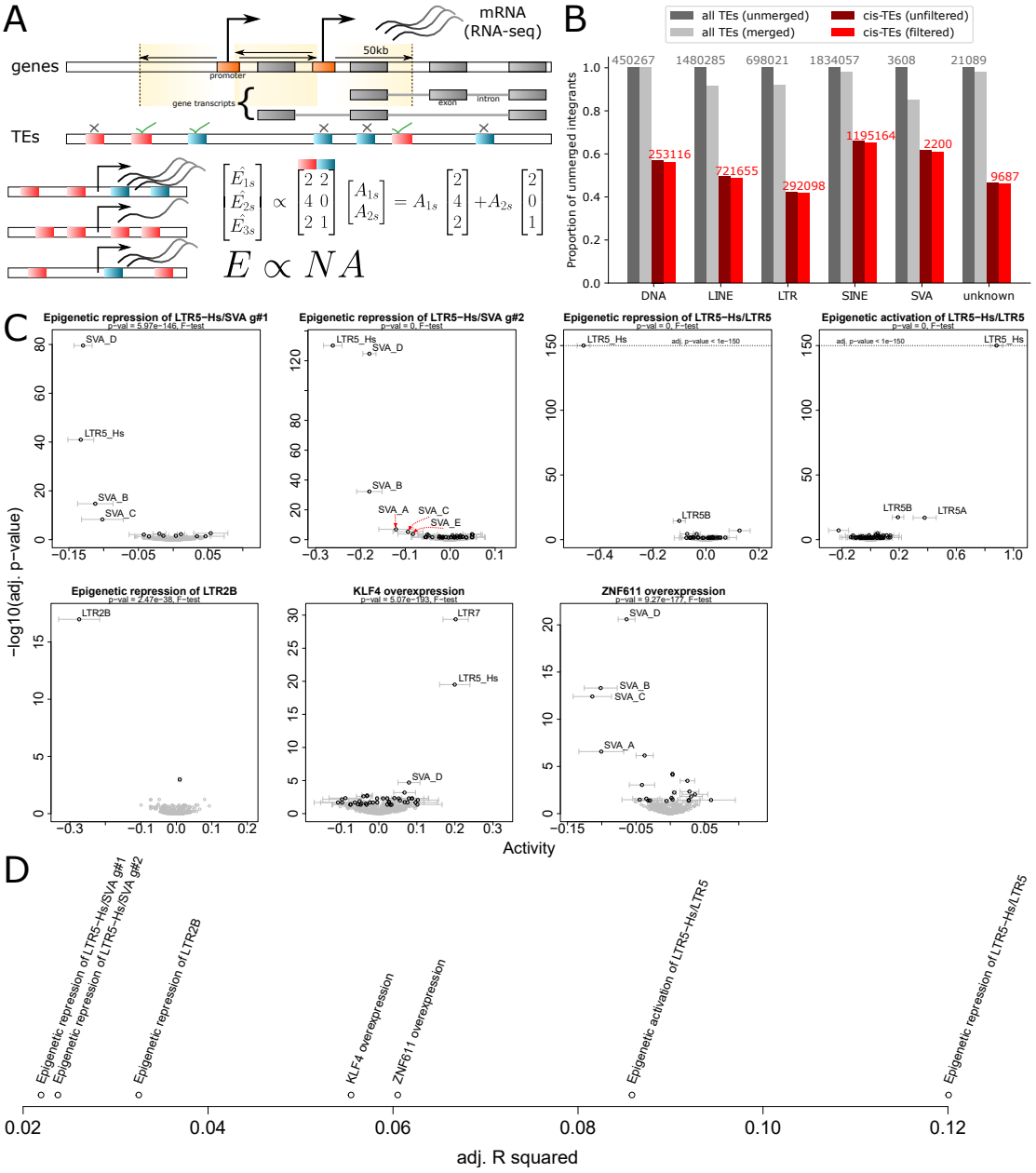



Figure 3.1 – ***craTEs* uncovers *cis*-regulatory TE subfamilies from RNA-seq.**

Figure 3.1 – **A** Overview of the *craTEs* model. Differences in expression [log(TPM)] for protein-coding genes between treatment and control samples (columns of matrix $E$) are modeled as a linear combination of the per-subfamily TE counts found in the *cis*-regulatory region (shaded beige) of each gene (columns of $N$). Differences in *cis*-regulatory activities for each treatment vs. control experiment (columns of $A$) are estimated by least squares. The *cis*-regulatory regions of each gene are defined as 50-kb long stretches of DNA 5' and 3' from promoter regions. *Cis*-regulatory regions exclude the exons (grey boxes) and promoters (orange boxes) of the genes they are assigned to. Grey bold lines: gene introns. Sequences of introns and exons: transcripts. **B** Proportion of integrants remaining at each step of the construction of N with respect to the original number of TEs present in the annotation (indicated in grey). "All TEs" refers to all integrants found in the the TE database "Repeatmasker RELEASE 20170127" (number of unmerged TEs are indicated in grey). "cis-TEs" refers to integrants found in *cis*-regulatory regions before ("unfiltered") and after ("filtered", numbers indicated in red) removing those overlapping exons and promoters of the corresponding gene. **C** Seven case studies exemplifying the estimation of the *cis*-regulatory activities of TE subfamilies from RNA-seq data. Black dots are TE subfamilies with statistically significant (BH-adj. p-value <0.05, t-test) differences in activities between the treatment and control groups. 95% confidence intervals for the estimated *cis*-regulatory activities are shown as grey bars. Grey dots are TE subfamilies with non-significant differences in activities. Subtitle: p-value from the F-test of overall significance in regression. From left to right: CRISPRi-mediated repression of LTR5-Hs and SVA integrants in naïve hESCs, gRNA #1 (g#1) $n = 3$ (3 treatment samples vs. 3 control samples) (Pontis et al., 2019a); CRISPRi-mediated repression of LTR5-Hs and SVA integrants in naïve hESCs, gRNA #1 (g#1) $n = 3$; CRISPRi-mediated repression of LTR5-Hs/A/B integrants in an embryonal carcinoma cell line (NCCIT), $n = 2$ (Fuentes, Swigut, & Wysocka, 2018a); CRISPRa-mediated activation of LTR5-Hs/A/B integrants in NCCIT, $n = 2$; CRISPRi-mediated repression of LTR2B integrants in K562, $n = 2$ (Deniz, Ahmed, Todd, Dawson, & Branco, 2019); overexpression of the pluripotency TF KLF4 in primed hESCs, $n = 4$ (Pontis et al., 2019a); overexpression of the SVA-targeting KZFP ZNF611 in naïve hESCs, $n = 2$ (Pontis et al., 2019a). **D** Proportion of variance of $E$ explained by *craTEs* for each experiment in **C**.

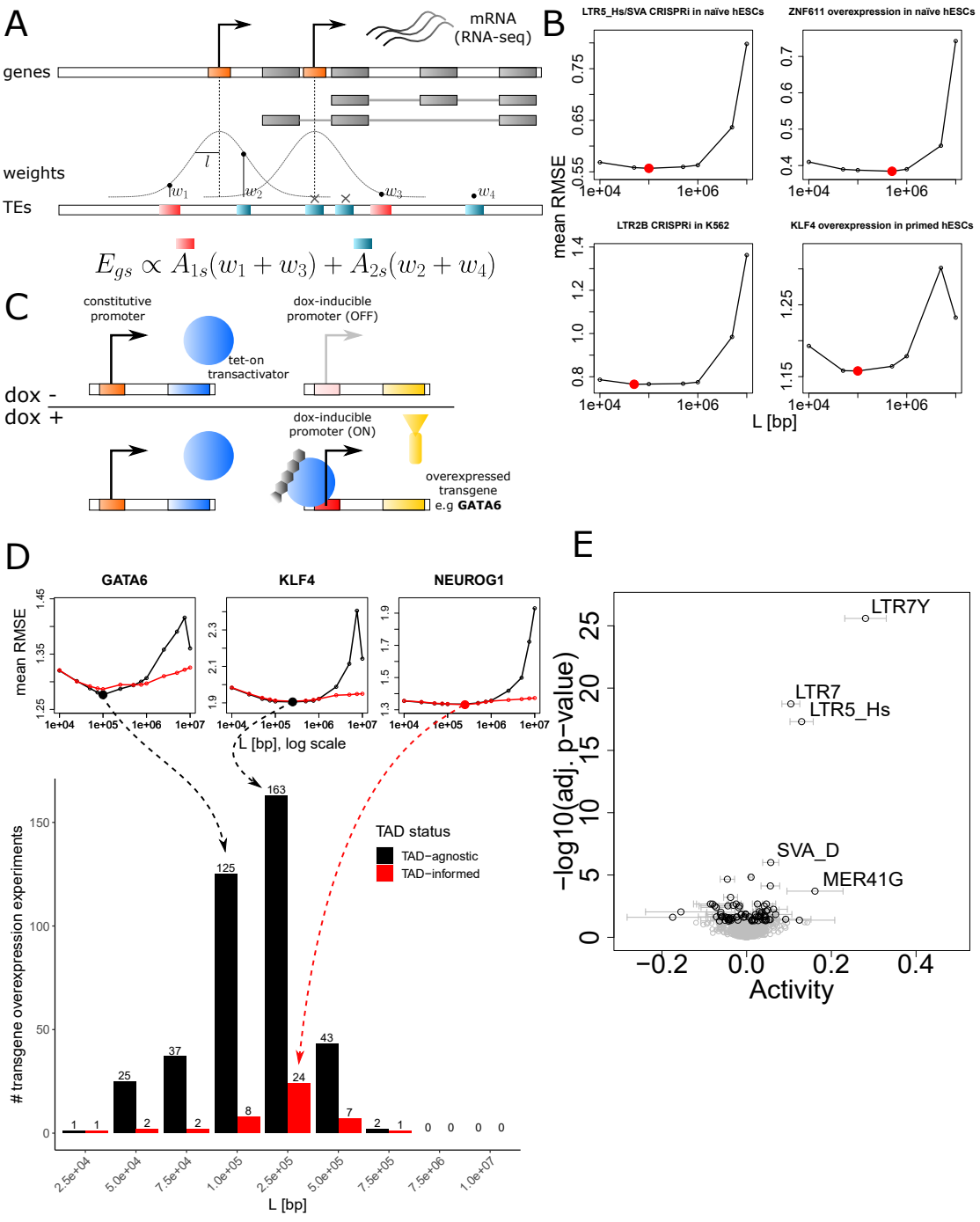



Figure 3.2 – **Influential TE-embedded *cis*-regulatory information resides up to 500kb from gene promoters**

Figure 3.2 – **A** Overview of the weighting process whereby the *cis*-regulatory influence of TEs decreases as a function of the distance to the closest promoter. The scheme depicts a protein-coding gene with two alternative promoters (in orange), coding for two alternative isoforms (in grey). Gaussian kernels with a maximum value of 1 and of varying bandwidth $L$ are centered on each promoter. Before being added to the corresponding element in the matrix $N$, each TE is weighted as a function of its distance to the closest gene promoter. TEs overlapping exons (grey boxes) and promoters (orange boxes) of the gene are excluded. **B** To find the bandwidth $L$ leading to smallest prediction error, the root mean-squared error (RMSE) was computed for each validation fold and averaged across the five folds over different values of $L$. **C** Overview of the experimental design of the hESC "perturbome" (Nakatake et al., 2020). hESC cell lines carrying a stably integrated dox-inducible transgene overexpression construct were established from individual cells. In each of the 441 transgene overexpression experiments, dox-treated samples (dox+) are compared to the same cell line in the absence of dox (dox-). Note that the number of replicates per experiment varies. **D** Histogram depicting the number of times each gaussian kernel bandwidth $L$ - either TAD-informed or agnostic - led to the smallest mean validation RMSE in a 5-fold cross-validation scheme for the 441 transgene overexpression experiments. TAD-informed (red): the *cis*-regulatory weights linking integrants to genes was restricted by topologically associating domain (TAD) boundaries. TAD-agnostic (black): TAD boundaries were not considered. Individual mean RMSE estimations for GATA6, KLF4 and NEUROG1 are showed as illustrative examples. **E** Estimation of the *cis*-regulatory activity of TE subfamilies upon KLF4 overexpression (Pontis et al., 2019a) using the matrix $N$ computed with $L$ = 250kb.

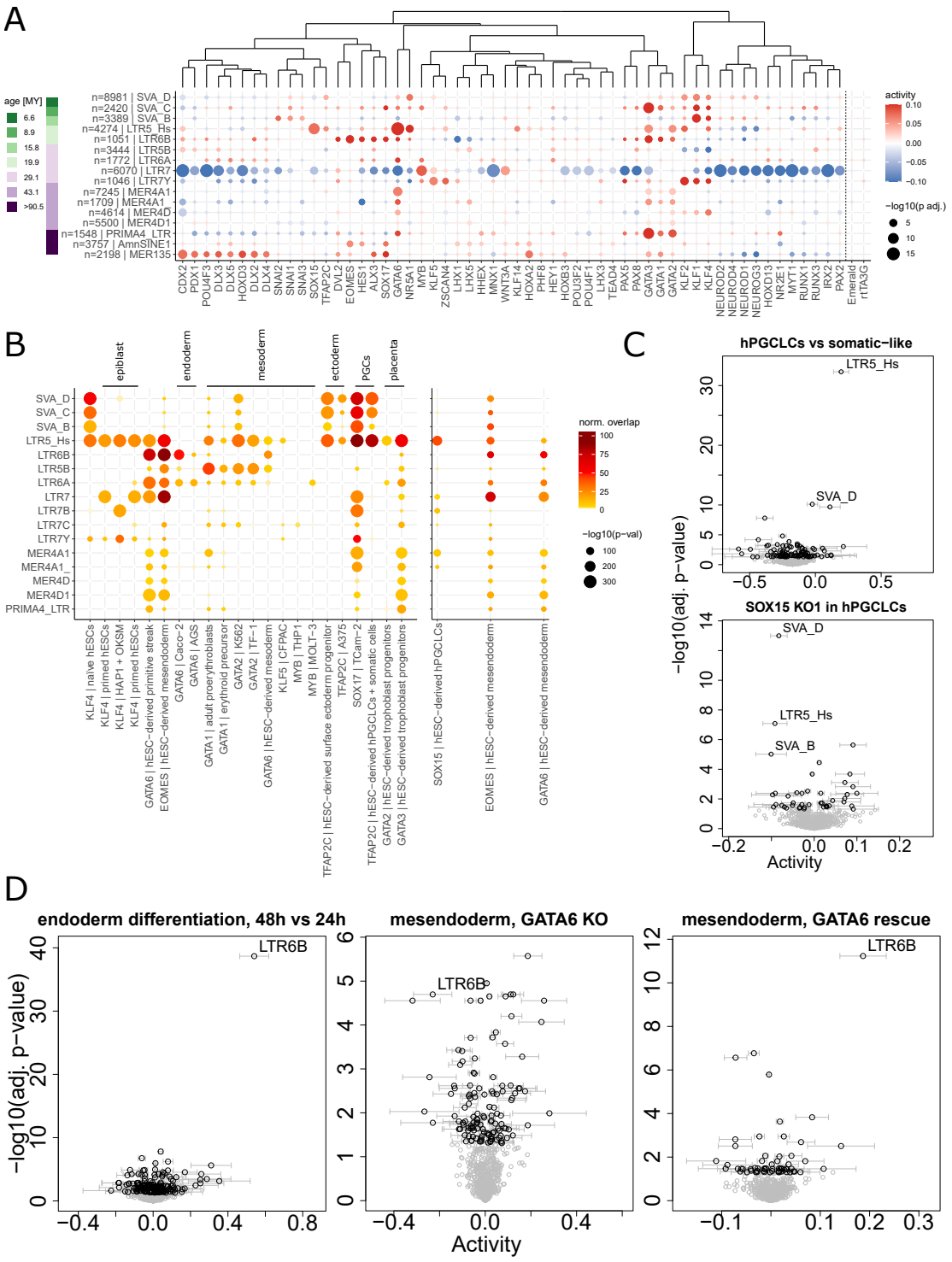



Figure 3.3 – **TFs controlling gastrulation and organogenesis promote the *cis*-regulatory activity of evolutionarily young TE subfamilies activated during pluripotency.**

Figure 3.3 – **A** TE subfamily *cis*-regulatory activities (color: activity coefficients; area: statistical significance) estimated from dox-induced transgene overexpression experiments at 48h in primed hESCs (Nakatake et al., 2020) using $N$ computed with L = 250kb. The number of replicates for each condition varies. Experiments were clustered using complete linkage hierarchical clustering on Euclidean distances computed from activity coefficients. Selected TE subfamilies were ordered by evolutionary age in millions of years, as previously estimated (Pontis et al., 2019b). The number of protein-coding genes with total *cis*-regulatory weights >0.13 (weight obtained at a distance of 2L, see fig. S3.2) is shown for each subfamily. The color labeling of the estimated activities was saturated at $|\Delta A| < 0.1$.**B** Top binding enrichment at selected evolutionarily young TEs by selected TFs controlling germ layer development. TEs (rows) were ordered as in **A**. ChIP-seq experiments (columns) were ordered by developmental stage or germ layer lineage. Color: number of peaks overlapping with subfamily-specific integrants, normalized for subfamily size. Area: statistical significance. Left: ChIP-seq peaks obtained from the ChIP-Atlas (Z. Zou, Ohta, Miura, & Oki, 2022). Right: ChIP-seq peaks obtained from (X. Wang et al., 2021a) and (Luo, Huangfu, & Beer, 2022). **C** Top: Estimated differences in TE-dependent *cis*-regulatory activities between hESC-derived EpCAM$^+$/INTEGRIN$\alpha$6$^+$ double positive (DP) hPGCLCs and double negative (DN) somatic cells at day 6 of differentiation, replicate #1 $n = 2$ (X. Wang et al., 2021a). Bottom: SOX15 KO DP hPGCLCs vs. DP hPGCLCs, day 6, $n = 2$. **D** Left: hESC-derived differentiating endoderm, 48h vs 24h, $n = 3$ (Luo, Huangfu, & Beer, 2022). Middle: iPSC-derived GATA6 KO mesendoderm vs. iPSC-derived mesendoderm, $n = 2$(Heslop, Pournasr, Liu, & Duncan, 2021a). Right: GATA6 rescue in iPSC-derived GATA6 KO mesendoderm vs. iPSC-derived GATA6 KO mesendoderm, $n = 2$.

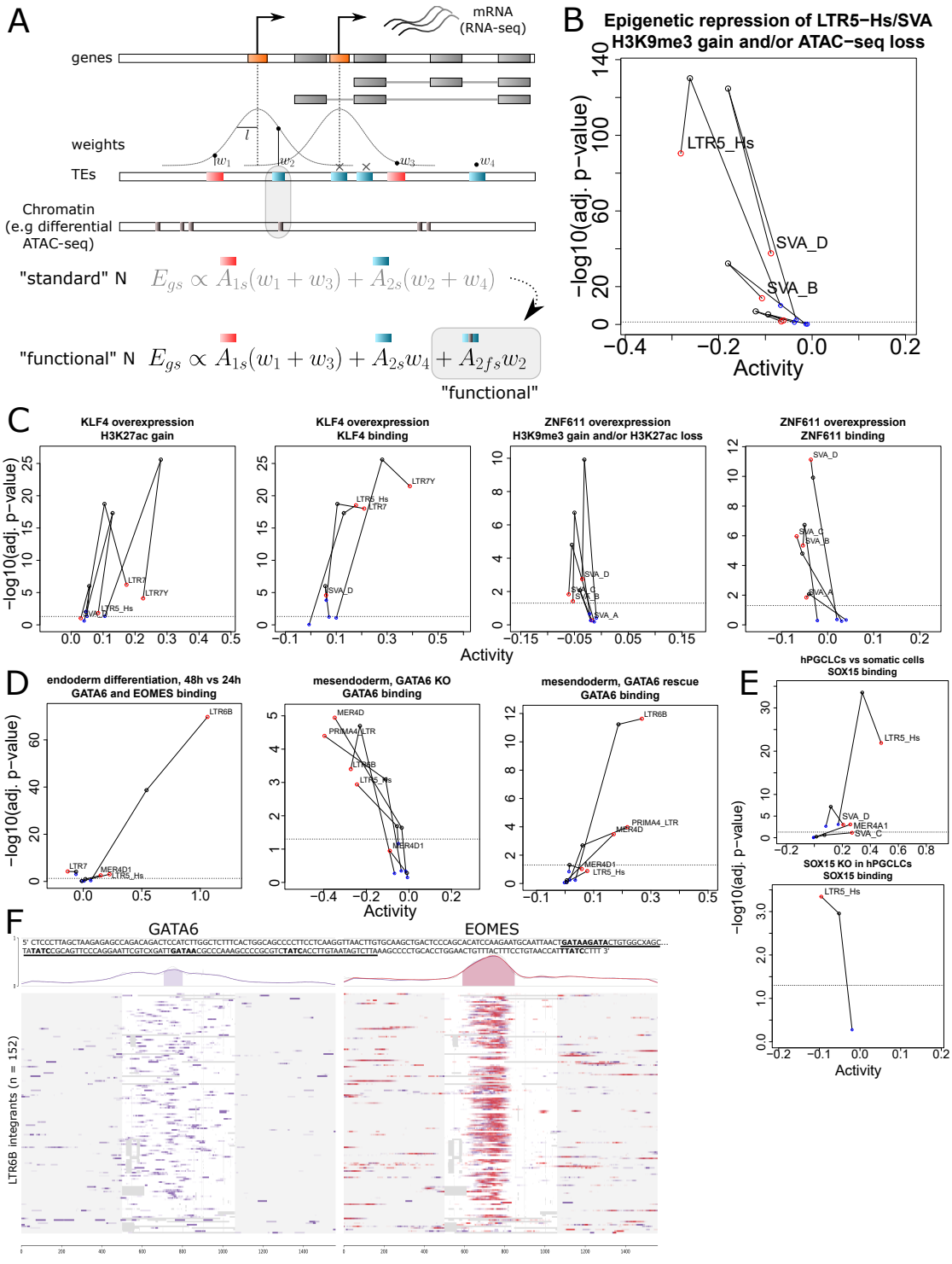



Figure 3.4 – ***Cis*-regulatory activities are more pronounced at epigenetically active TEs.**

Figure 3.4 – **A** Overview of the procedure whereby TE subfamilies are split between so-called "functional" and "non-functional" fractions based on additional evidence, e.g. differential chromatin accessibility. The regulatory susceptibility scores tying TE subfamilies to protein-coding genes are distributed between the functional and non-functional fractions of each TE subfamily, leading to an experiment-specific column-wise expansion of $N$. Concretely, functional and non-functional fractions of TE subfamilies are treated as independent TE subfamilies in the subsequent *cis*-regulatory activity estimation process. **B** Estimated differences in *cis*-regulatory activity for the functional (in red) and non-functional fractions (in blue) of LTR5-Hs and SVA subfamilies under CRISPRi-mediated epigenetic repression in naïve hESCs (Pontis et al., 2019a). The *cis*-regulatory activities for the unsplit subfamilies were estimated in a separated iteration of *craTEs*, using the standard distance-weighted $N$ matrix ($L = 250$kb), and are shown in black. The dotted line represents the significance threshold of BH-adjusted $p.val = 0.05$. Note that even though only selected subfamilies are plotted for clarity, all TE subfamilies were included in the fitting process. **C** Estimated differences *cis*-regulatory activity for the functional and non-functional fractions of selected TE subfamilies according to definitions of the functional state that are either based on differential chromatin states (1st and 3rd panels from the left) or differential TF binding (2nd and 4th panels from the left) at integrants (Pontis et al., 2019a). **D** Estimated differences in *cis*-regulatory activities for the functional (bound by both GATA6 and EOMES (Luo, Huangfu, & Beer, 2022)) vs. non-functional fractions of selected TE subfamilies during hESC-derived endoderm differentiation, 48h vs. 24h, $n = 3$ (Luo, Huangfu, & Beer, 2022) (left), functional (GATA6-bound (Luo, Huangfu, & Beer, 2022)) vs. non-functional fractions of selected TE subfamilies upon GATA6 KO in iPSC-derived mesendoderm, $n = 2$ (Heslop, Pournasr, Liu, & Duncan, 2021a) (center) and GATA6 rescue in GATA6 KO iPSC-derived mesendoderm, $n = 2$ (right). **E** Estimated differences in *cis*-regulatory activity for the functional (SOX15-bound) vs. non-functional fractions of selected TE subfamilies between DP hESC-derived hPGCLCs and DN somatic cells, day 6, $n = 2$ (X. Wang et al., 2021a) (top) and SOX15 KO in DP hESC-derived hPGCLCs (bottom). **F** Multiple sequence alignment (MSA) of all 152 LTR6B integrants considered by *craTEs* (central white rectangle). Grey patches within the central white rectangle indicate gaps. Sequences at loci found in grey rectangles flanking the MSA region are shown for convenience and were not aligned. The intensity of GATA6 (left, $n = 1$) and EOMES (right, $n = 2$) ChIP-seq signal is indicated at the corresponding genomic loci. The fraction of sequences adorned with ChIP-seq signal for each position is shown on top. Consensus sequences found underneath high-density ChIP-seq signal regions ($> \frac{1}{3}$ of sequences overlapping ChIP-seq reads) with the highest density of GATA6 (underlined), resp. EOMES (entire consensus) signal are reported, with GATA consensus DNA-binding sites in bold.

## 3.8 Supplementary information

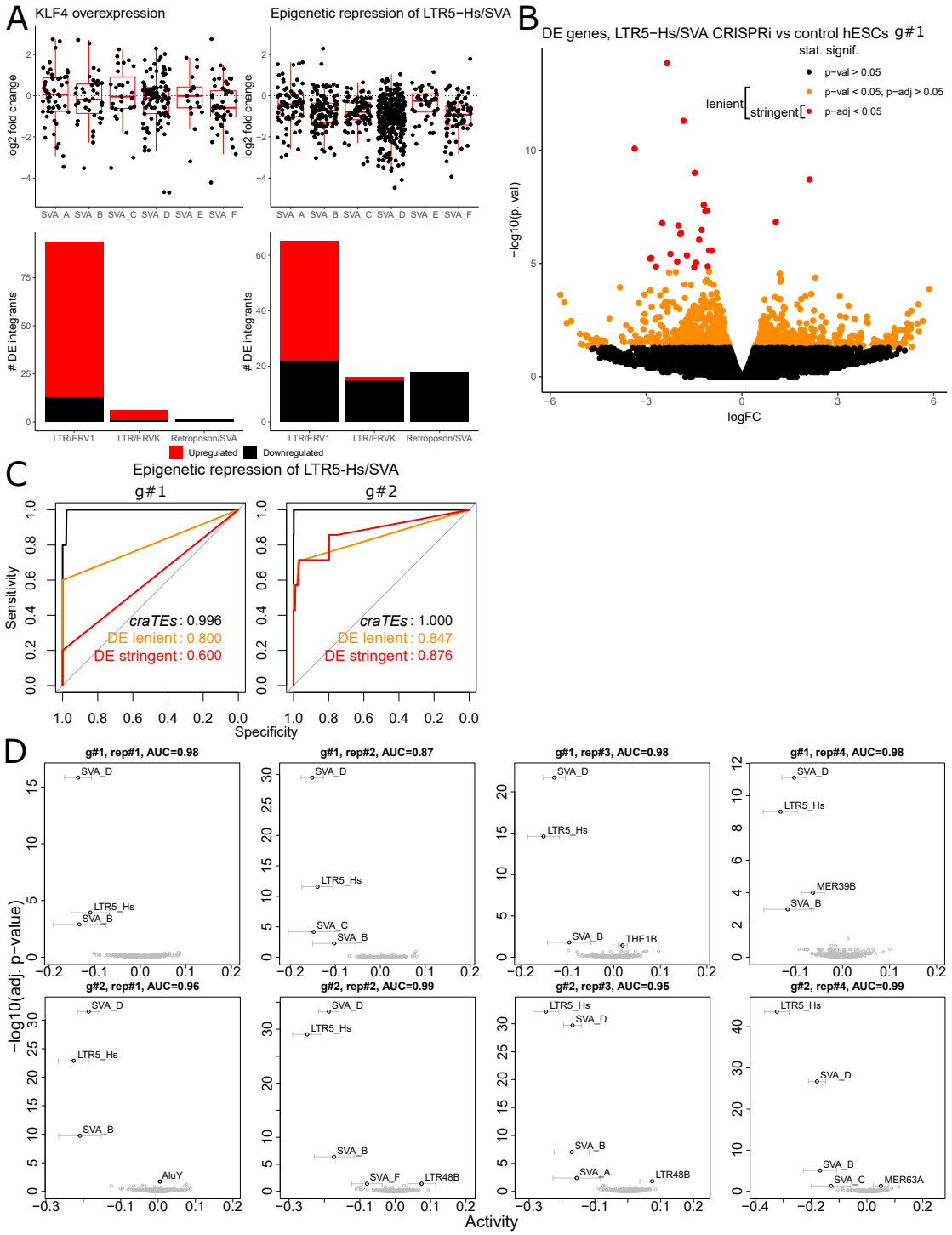



Figure S3.1 – ***craTEs* outperforms enrichment approaches based on differential expression analyses.**

Figure S3.1 – **A** Top: Log2 fold-change (RNA-seq) for SVA integrants (top) and number of differentially expressed (DE) integrants for ERV1, HERV-K and SVA TE classes (bottom) upon KLF4 overexpression and CRISPRi-mediated repression of LTR5-Hs/SVAs (Pontis et al., 2019a). **B** Volcano plot of DE analysis on protein-coding genes from the CRISPRi-mediated repression of LTR5-Hs/SVA using g#1. "Lenient" DE calling: all genes with unadjusted p-val <0.05, Fisher's exact test. "Stringent" DE calling: all genes with Benjamini-Hochberg adjusted (Benjamini & Hochberg, 1995) p-val <0.05. **C** ROC curves and AUCs for the classification of TE subfamilies as *cis*-regulatory vs. not *cis*-regulatory based on statistical significance using either *craTEs* or DE enrichment approaches. Ground truth: TE subfamilies that are (1) targeted by the gRNA and (2) display enrichment for differential ATAC-seq/ChIP-seq signal indicative of heterochromatin gain under CRISPRi-mediated repression of LTR5-Hs/SVAs (Pontis et al., 2019a). **D** Case study for the estimation of the TE subfamily *cis*-regulatory activities in a 1 vs. 1 sample setting ($n = 1$), using each of the paired replicates in the CRISPRi-mediated repression of LTR5-Hs and SVAs in naïve hESCs (Pontis et al., 2019a). AUCs as in **C**.

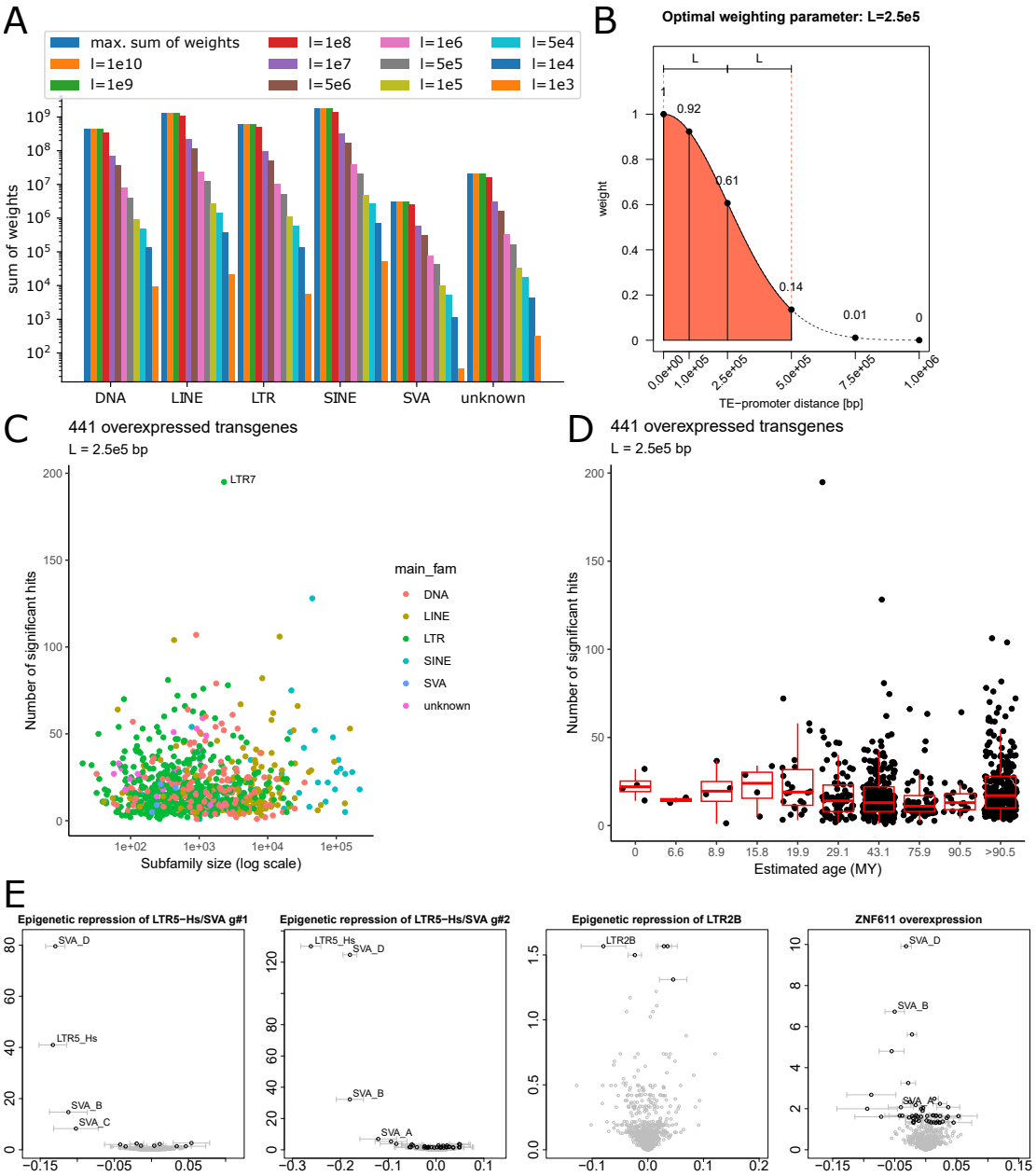



Figure S3.2 – **Related to figures 3.1 and 3.2.**

Figure S3.2 – **A** Sum of *cis*-regulatory weights as a function of the gaussian kernel width $L$ across each main TE families. The maximum *cis*-regulatory weight represents the limit case whereby each TE contributes to the regulation of each gene located on the same chromosome. **B** Illustration of the gaussian kernel corresponding to the optimal choice of $L = 250$kb. *cis*-regulatory weights for TE integrants located at selected distances from TSS are shown on the y-axis. The coloured area under the gaussian curve bounded by TE-promoter distances of 0 and $2L$ - i.e. 2 gaussian standard deviations - contains approx. 95% of the total area under the gaussian curve. **C** Number of times TE subfamilies reached statistical significance in the 441 transgene overexpression experiments (Nakatake et al., 2020) as a function of subfamily size or **D** evolutionary age. **E** Estimation of the *cis*-regulatory activity of TE subfamilies upon (left to right) CRISPRi-mediated epigenetic repression of LTR5-Hs/SVAs (Pontis et al., 2019a), CRISPRi-mediated epigenetic repression of LTR2B (Deniz, Ahmed, Todd, Dawson, & Branco, 2019), overexpression of ZNF611 (Pontis et al., 2019a) using the matrix $N$ computed with $L = 250$kb

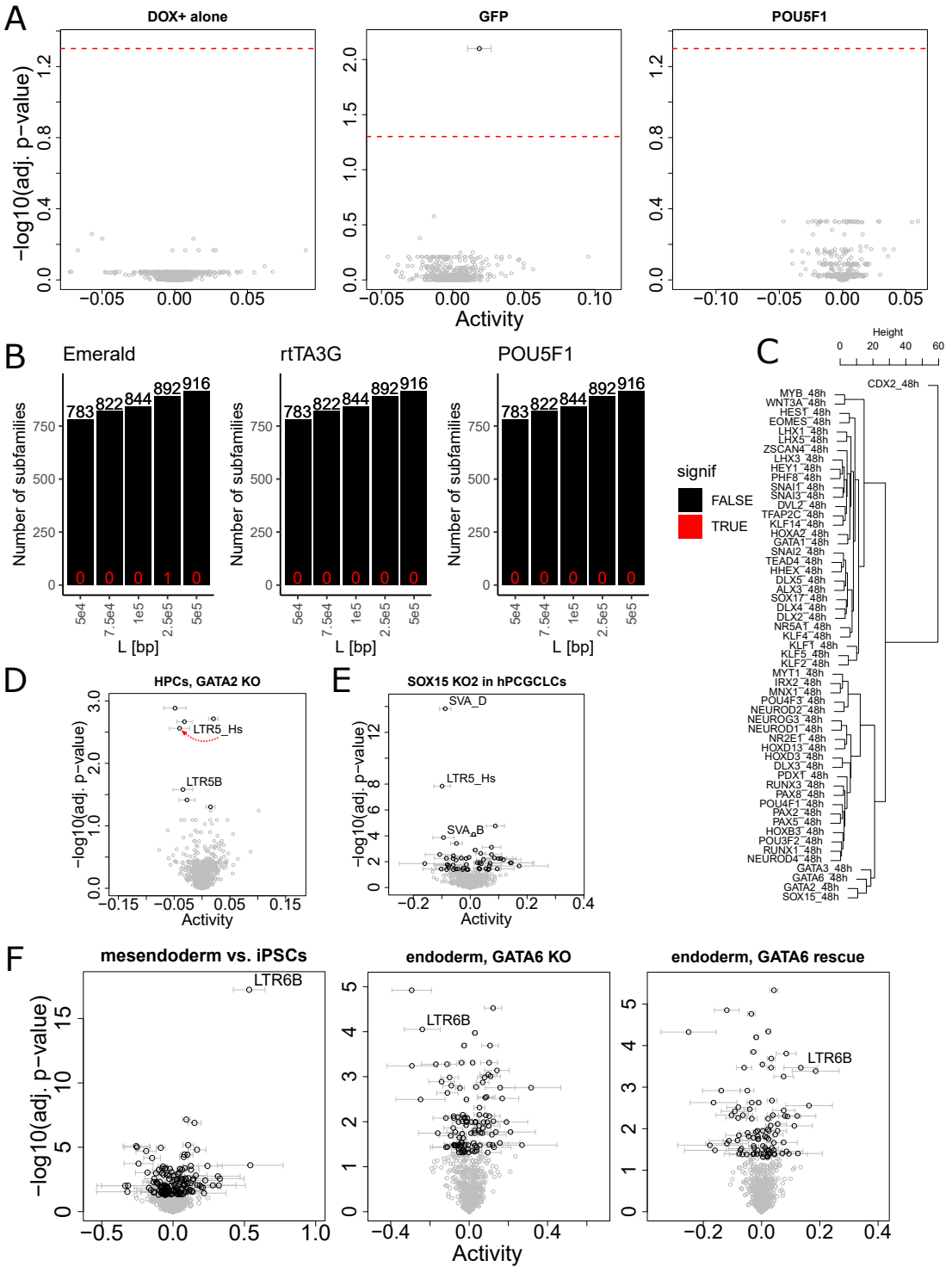



Figure S3.3 – **Related to figure 3.3.**

Figure S3.3 – **A** Estimation of the differences in TE subfamily *cis*-regulatory activity triggered by dox-treatment alone, $n = 1$(left), dox-induced GFP overexpression, $n = 2$ (middle) and dox-induced POU5F1 overexpression, $n = 3$ (right) compared to the corresponding untreated cell lines in the "perturbome" dataset (Nakatake et al., 2020). Dotted red line: statistical significance threshold. **B** Number of statistically significant differences (red), resp. non-significant differences (black) in TE-dependent *cis*-regulatory activities found for dox-induced GFP overexpression (left), dox-treatment alone (center) and dox-induced POU5F1 overexpression (right) found for distance-weighted susceptibility matrices $N$ derived using various bandwidths $L$. **C** Dendrogram obtained from performing complete linkage hierarchical clustering on Euclidean distances computed from statistical significance on the transgene overexpression experiments and TE subfamilies shown in fig. 3.3A. **D** Estimated differences in TE-subfamily *cis*-regulatory activities upon GATA2 KO in hematopoietic progenitor cells (HPCs), $n = 2$ (Huang, Du, Shi, Chen, & Pan, 2017), **E** DP hPGCLCs vs DN somatic cells at day 6 of differentiation, replicate 2 $n = 2$ (X. Wang et al., 2021a), **F** iPSC-derived mesendoderm vs. iPSCs, $n = 2$ (Heslop, Pournasr, Liu, & Duncan, 2021a) (left), GATA6 KO in iPSC-derived endoderm, $n = 2$ (center) and GATA6 rescue in GATA6 KO iPSC-derived endoderm, $n = 2$.

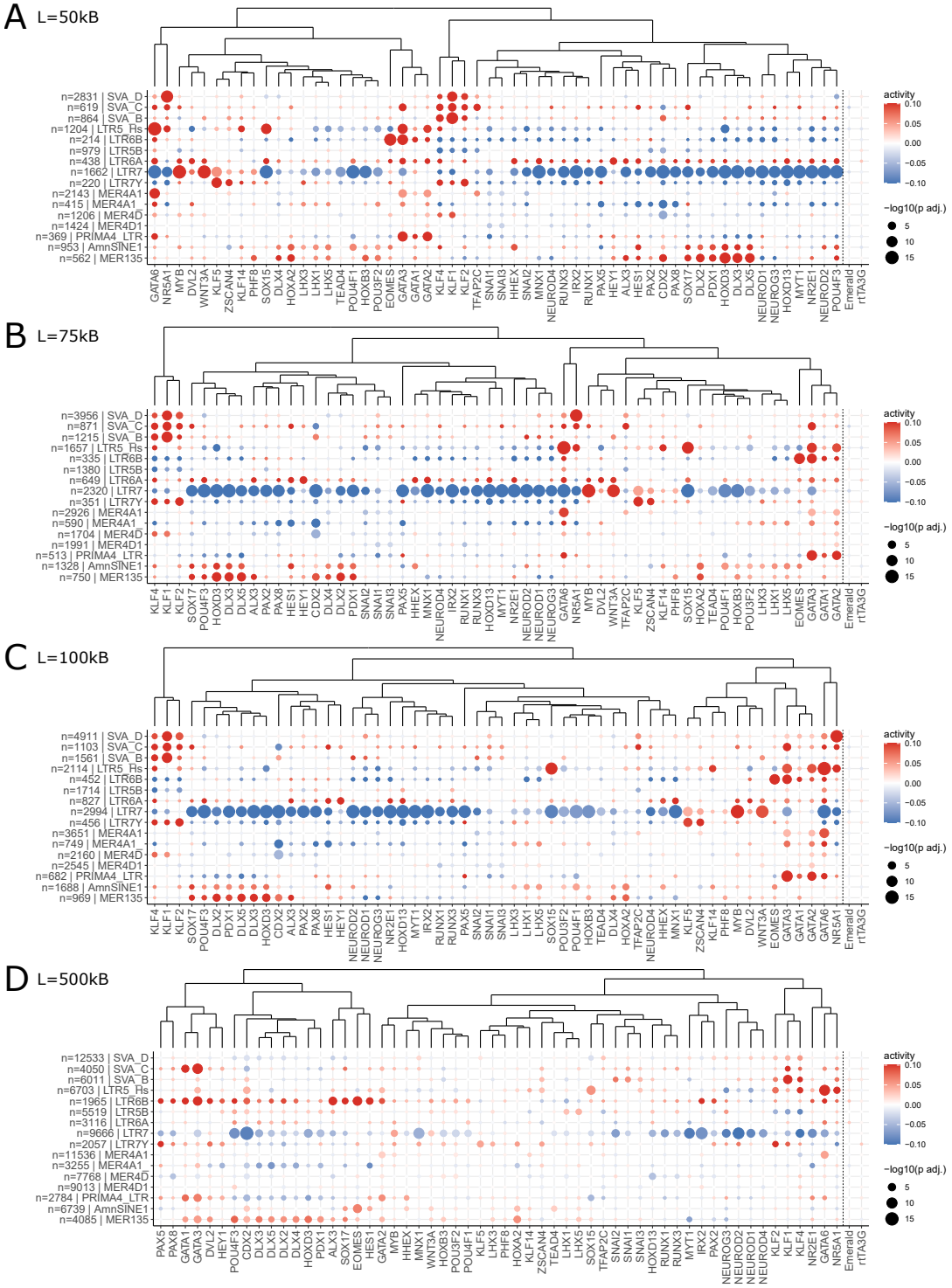



Figure S3.4 – **Estimated differences in TE subfamily *cis*-regulatory activities are robust to variations in $L$. A-D** TE-dependent *cis*-regulatory activities, as in fig. 3.3A, derived from $N$ matrices with different values of $L$.

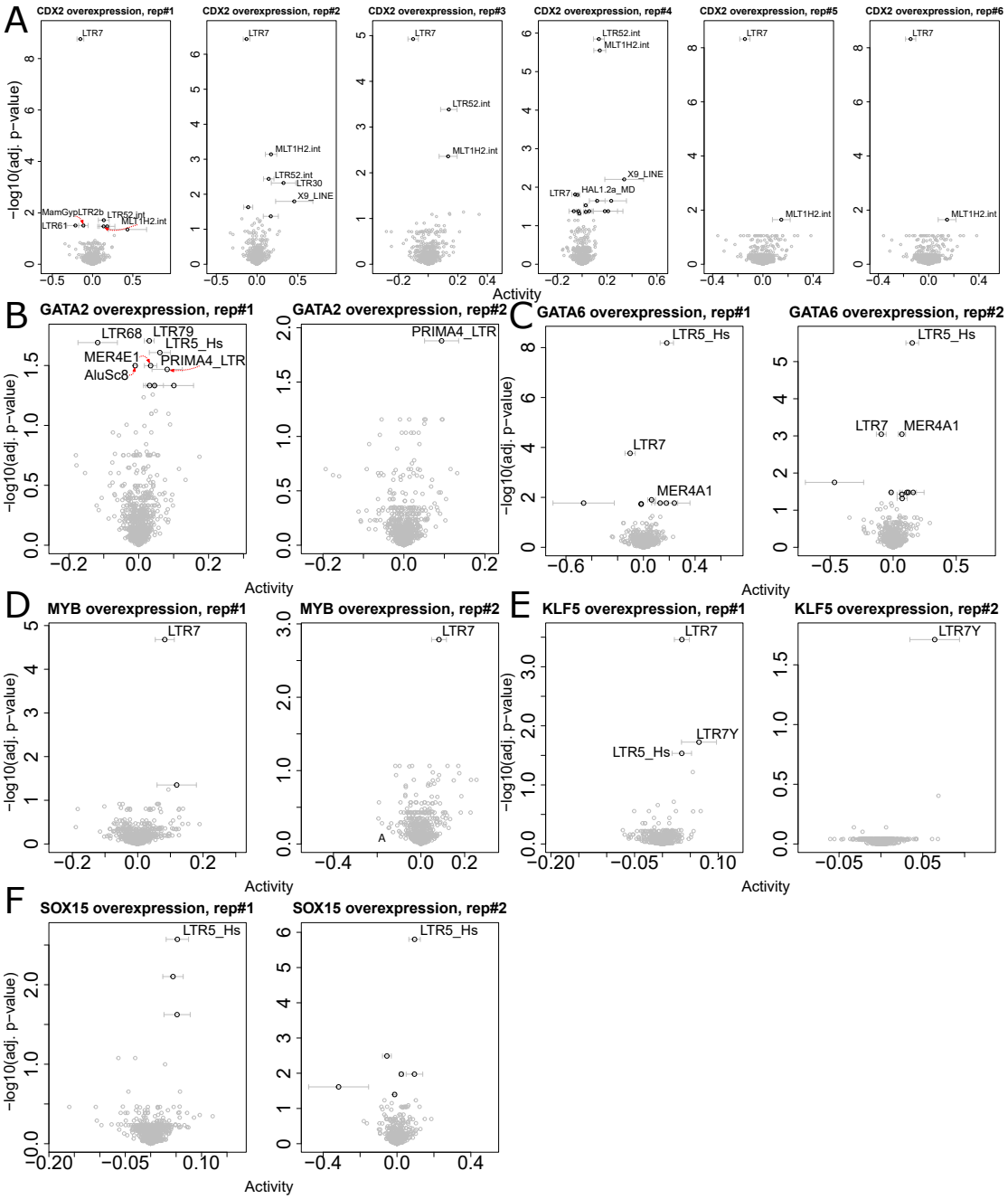



Figure S3.5 – **Individual replicates from the hESC "perturbome" dataset show consistent estimated differences in *cis*-regulatory activities. A** Estimated TE-dependent *cis*-regulatory activities across individual replicates ($n = 1$) of transgene overexpression experiments (Nakatake et al., 2020) for CDX2, **B** GATA2, **C** GATA6, **D** MYB, **E** KLF5 and **F** SOX15.

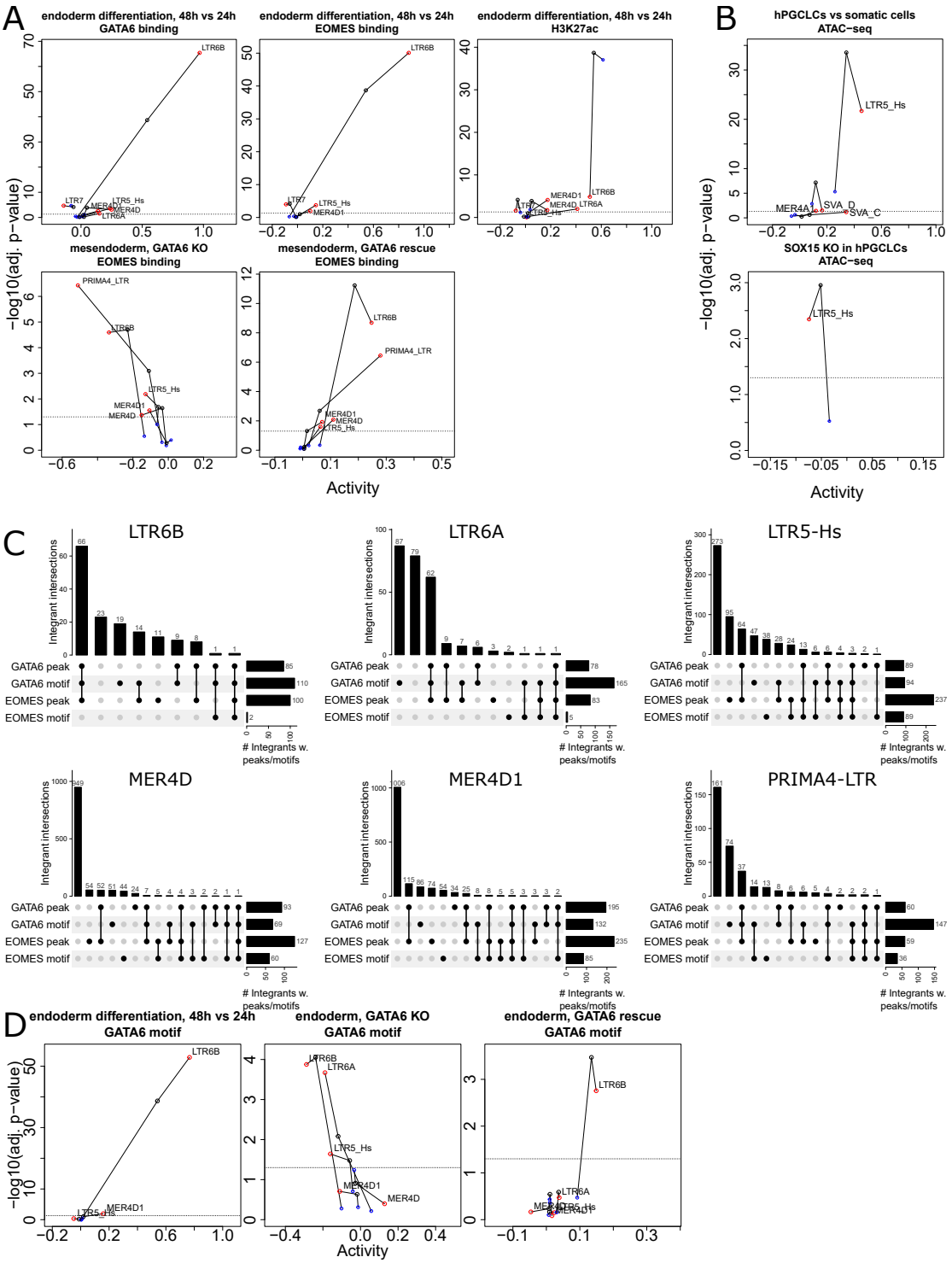



Figure S3.6 – **Related to figure 3.4.**

Figure S3.6 – **A** Estimated differences in *cis*-regulatory activities for the functional (top left: GATA6-bound, center: top center: EOMES-bound, top right: marked by H3K27ac, bottom: EOMES-bound (Luo, Huangfu, & Beer, 2022)) vs. non-functional fractions of selected TE subfamilies during hESC-derived endoderm differentiation, 48h vs. 24h, $n = 3$ (Luo, Huangfu, & Beer, 2022) (top) and GATA6 KO (bottom left) and rescue (bottom center) in iPSC-derived mesendoderm, $n = 2$ (Heslop, Pournasr, Liu, & Duncan, 2021a). **B** Functional (top: marked by ATAC-seq reads, bottom: SOX15-bound) vs. non-functional estimated differences in TE-dependent *cis*-reguatory activities in DP vs. DN hPGCLCs (top) and in SOX15KO vs. wild-type DP hPGCLCs (bottom) at day 6, $n = 2$ (X. Wang et al., 2021a). **C** Subfamily-restricted upset plots (Gu, Eils, & Schlesner, 2016) showing all intersections between the following sets of integrants: those overlapping GATA6 peaks (Luo, Huangfu, & Beer, 2022), GATA6 DNA-binding motifs as located by FIMO (Grant, Bailey, & Noble, 2011), EOMES peaks and EOMES-DNA binding motifs. **D** Estimated differences in *cis*-regulatory activities for functional (contains a GATA6 DNA-binding motif) vs. non-functional fractions of selected TE subfamilies during endoderm differentiation (Luo, Huangfu, & Beer, 2022) (left), upon GATA6 KO in iPSC-derived endoderm $n = 2$ (Heslop, Pournasr, Liu, & Duncan, 2021a) (center) and GATA6 rescue in iPSC-derived endoderm $n = 2$ (right).

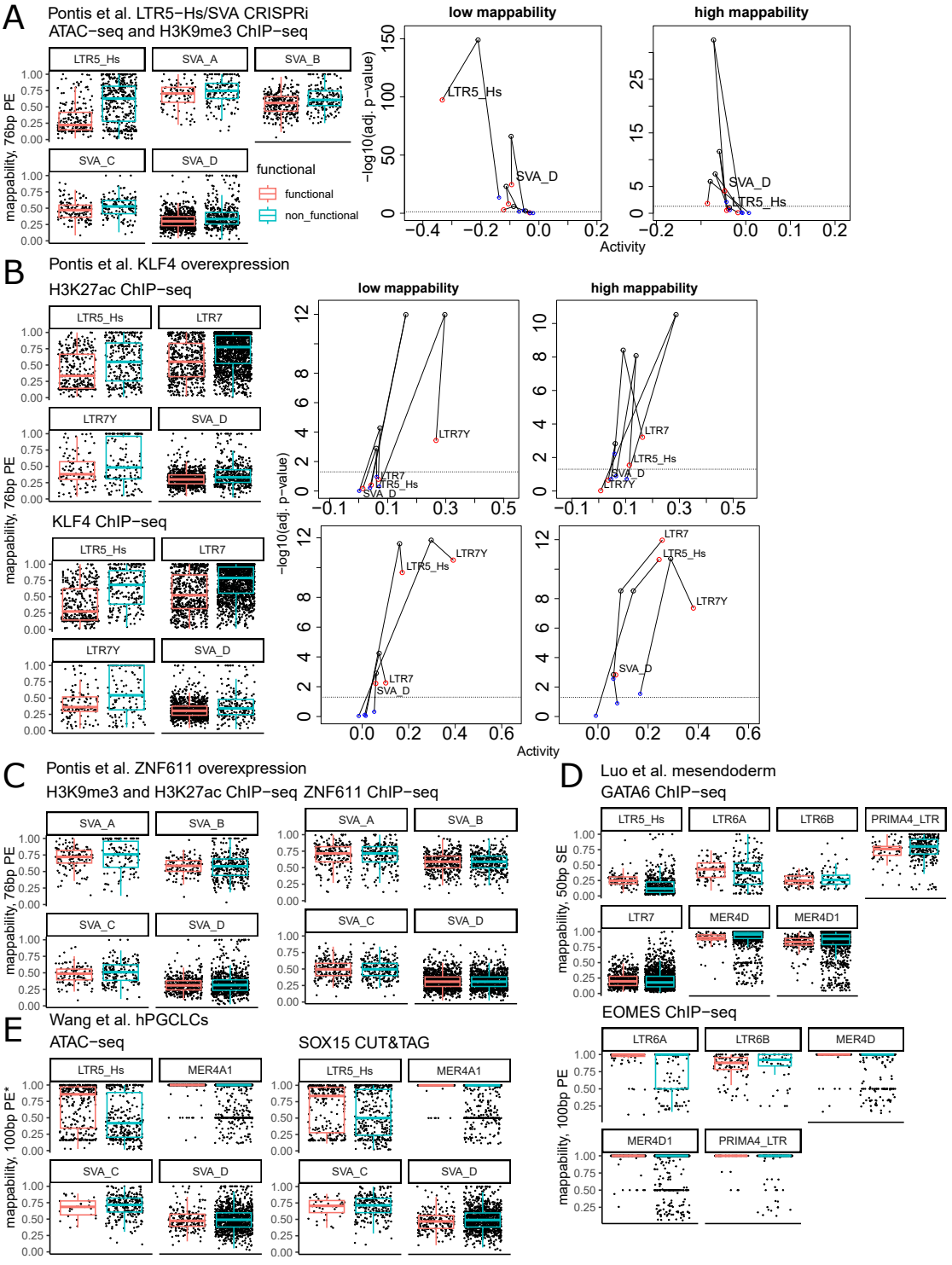



Figure S3.7 – **Epigenomics-informed estimates of TE subfamily *cis*-regulatory activities are robust across mappability levels**

Figure S3.7 – **A** Left: mappability scores (Sexton & Han, 2019) averaged over each functional (overlapping ATAC-seq loss and/or H3K9me3 gain loci, red) and non-functional (complement, blue) integrant from relevant *cis*-regulatory TE subfamilies (fig. 3.4A) upon CRISPRi-mediated epigenetic repression of LTR5-Hs/SVAs in naïve hESCs, g#2 (Pontis et al., 2019a). Boxplots: lower quartile, median and upper quartile of mappability scores averaged over each integrant. Right: estimated differences in *cis*-regulatory activities for mappability-matched functional (red) vs. non-functional (blue) fractions of relevant TE subfamilies upon CRISPRi-mediated repression of LTR5-Hs/SVAs, g#2, $n = 3$. Prior to the functional/non-functional split, each of the indicated subfamilies was split into a low, resp. high mappability subgroup using a median split over per-integrant mappability scores within the subfamily. **B** Mappability over functional (top left: H3K27ac gain, bottom left: KLF4 ChIP-seq peak) and non-functional integrants, and differences in *cis*-regulatory activities (right) for mappability-matched functional vs. non-functional fractions of relevant TE subfamilies upon KLF4 overexpression in primed hESCs, $n = 4$ (Pontis et al., 2019a). **C** Mappability over functional (left: H3k9me3 gain and/or H3K27ac loss, right: ZNF611 binding) vs. non-functional integrants upon ZNF611 overexpression in naïve hESCs (Pontis et al., 2019a). **D** Mappability over functional (top: GATA6 ChIP-seq peak, bottom: EOMES ChIP-seq peak) vs. non-functional integrants in hESC-derived mesendoderm (Luo, Huangfu, & Beer, 2022). **E** Mappability over functional (left: ATAC-seq peak, right: SOX15 CUT&TAG peak) vs. non-functional integrants in hPGCLCs (X. Wang et al., 2021a). The asterisk highlights that we used 100bp paired end mappability tracks, as 150bp paired end mappability tracks were not available.

# 4 Perspectives

## 4.1 Sharpening the tool: improving *craTEs*

Characterizing *craTEs* showed that a substantial fraction of the variation in protein-coding gene expression is accounted by the genomic distribution of TEs. In its most basic formulation, *craTEs* considers TEs of the same subfamily as regulatory black boxes endowed with equal regulatory potential. *craTEs* thus remains agnostic to the precise mechanisms underpinning the estimated *cis*-regulatory activities. However, supplementing the model with epigenomic data or regulatory motifs has at many occasions usefully refined those estimations. We effectively achieved this by splitting existing subfamilies into subgroups ("sub-subfamilies") according to relevant non-transcriptomic data. In effect, this amounted to redistributing columns of the regulatory susceptibility matrix $N$ thereby creating new predictors. May this approach be systematized? For instance, would segregating TE subfamilies according to TF binding enable the study of TE-mediated *cis*-regulation at higher resolutions, and perhaps more importantly, truer to one of the most likely mechanistic accounts of *cis*-regulation? A naive, brute-force approach would entail "exploding" TE subfamilies according to TF binding data, e.g. distributing each subfamily into subgroups according to the presence of TF binding motifs (LTR5-Hs becomes LTR5-Hs-KLF, LTR5-Hs-GATA, etc.). This would result in a number of predictors, or columns of $N$, equal to a multiplication of the number of subfamilies by

the number of TFs or regulatory motifs considered. While the ordinary least squares (OLS) resolution of the linear regression problem remained fruitful when considering a single TF or motif, i.e. when the number of columns in $N$ was doubled, this would assuredly fail for larger numbers of TFs, if only because the regression problem would quickly become high-dimensional (more predictors / *cis*-regulatory subgroups than observations / genes whose expression is modeled). This could be addressed through regularization (James, Witten, Hastie, & Tibshirani, 2013), but some difficulties would persist.

We preliminary explored such a modeling approach (data not shown) and found that many TF-TE combinations only take non-zero *cis*-regulatory susceptibility values for a handful of genes, sometimes less than ten, thus greatly compromising the prospects of obtaining stable activity estimates. Second, as distinct TFs often bind similar sequences due to structural commonalities or multimerization (Ambrosini et al., 2020), TF-TE predictors were hampered by high levels of collinearity. For instance, we attempted to model the transcriptional variation caused by KLF4 overexpression in hESCs as a function of TF binding motifs - amongst which KLFs - located within LTR7 elements and weighted by their distance to gene promoters. Despite restricting the analysis to a single TE subfamily, we were not able to recover LTR7-located KLF DNA binding motifs as those with the greatest *cis*-regulatory activity (data not shown) whether considering effect size or statistical significance.

However, an alternative approach may hold more promise for refining *craTEs*-estimated regulatory activities. In its current formulation, the model is blind to any sort of relationship *between* subfamilies, whether phylogenetic or functional. For example, integrants of the SVA subfamilies (SVA-A, SVA-B, ... SVA-F) highly resemble each other, at least much more than any SVA integrant resembles a MIR or Alu integrant, and all display KLF4 binding at their 3' end (Pontis et al., 2019b). Indeed, we observed that SVA *cis*-regulatory activities often correlated. We found similar patterns for LTR5 (LTR5-Hs, LTR5A, LTR5B) and MER4 (MER4A1, MER4D, MER4D1) families. Thus, encoding cross-subfamily information into *craTEs*, e.g. as a nearest-neighbor graph, may enable the smoothening of TE-mediated *cis*-regulatory activity estimates across the barriers imposed by the somewhat arbitrary subfamily classification

scheme (Tibshirani et al., 2005).

A much lower hanging fruit probably lies in constructing $N$ matrices for other species with sufficiently well-annotated genomes. Mouse comes as an obvious first candidate. This would somewhat alleviate the limits imposed by ethical considerations on the availability of human developmental data and allow for comparing TE-mediated *cis*-regulation across mammalian lineages. The parameters to consider before endeavoring to extend *craTEs* to a particular species are: genome assembly quality (are the genomes assembled into chromosomes, or do they remain scattered as scaffolds?); protein-coding gene annotation, including precise TSS localizations; TE annotations and the availability of pertinent RNA-seq data, ideally matched to epigenomic data. Other model organisms, such as drosophila or zebrafish meet these criteria and would thus be suitable to adapting *craTEs*.

As a last proposed improvement, the current TE subfamily annotation could benefit from further refinement before constructing the regulatory susceptibility matrix $N$, such as accounting for within subfamily distributions of sequences known to drive transcriptional and epigenomic activity. For instance, segregating L1 subfamilies between 5'UTR-containing vs. 5'-truncated fractions may reveal L1 5'UTR-mediated *cis*-regulatory activities being masked under the current L1 subfamily classification.

## 4.2 Unaddressed TE-associated functions

The measurable influence that TEs bear on genomes is not restricted to *cis*-regulation, however broad a definition of that term one chooses. TEs, including lineage-specific ones, are enriched at CTCF binding sites (Kunarso et al., 2010) and specific subfamilies show strong statistical associations with TAD boundaries (Diehl, Ouyang, & Boyle, 2020), suggesting that the spread of TEs may contribute to heritable variations in 3D-genome organization. More intriguingly, the LTR7-driven HERVH subfamily of primate-specific TEs appears endowed with *de novo* TAD boundary formation capabilities, with maintenance requiring active transcription (Yanxiao Zhang et al., 2019). This adds to the long list of pluripotency-specific HERVH-associated

transcriptional and epigenomic phenomena, such as the expression of long non-coding RNAs required for establishing and maintening pluripotency (Sexton et al., 2022; J. Wang et al., 2014). This also raises the possibility that the pluripotency-specific *cis*-regulatory activity of LTR7-HERVH is in part a byproduct of some orthogonal LTR7-HERVH process dependent on active transcription, thereby putting an evolutionary constraint on LTR7 *cis*-regulatory activity by virtue of sequence homology.

*craTEs* uses a human reference genome for relating TEs to protein-coding genes, and is therefore blind to polymorphic variation at TEs. Selected young HERVK, L1, SVA and SINE subfamilies retain the ability to retrotranspose (Mills, Bennett, Iskow, & Devine, 2007; Sultana et al., 2019) and are polymorphic in the human population owing to their recent appearance in the human lineage (Lanciano et al., 2023; Philippe et al., 2016). Our work has notably emphasized the potent *cis*-regulatory role of SVAs and LTR5-HERVK. We thus expect polymorphic variation at these elements to be associated with *cis*-regulatory consequences that are currently missed with *craTEs* (Van Bree et al., 2022).

## 4.3 Unexplained observations

We observed that old, conserved mammalian TE subfamilies such as MER135, MER121 and AmnSINE1 displayed recurrent increases in *cis*-regulatory activity upon overexpression of homeobox-domain containing TFs in hESCs. We were however not able to attribute this to direct TF binding, unlike *cis*-regulatory changes induced at young TE subfamilies by homeobox-devoid master regulators of development, although this may be due to a lack of TF binding data generated in appropriate biological contexts. The role played by these ancient subfamilies thus remains enigmatic, but may prove relevant for fetal development at timepoints taking place beyond gastrulation and concurrent with limb development and the activation of GRNs governed by Hox TFs.

## 4.4 Lessons learned

*craTEs*' ability to detect TE-mediated *cis*-regulation across a wide range of contexts cements the notion that as they spread, TEs disseminate ready-for-use TF binding platforms poised to fine tune the expression of nearby genes. Consistent with this view, we found that many of the subfamilies displaying large differences in *cis*-regulation were evolutionary young, which correlates with higher levels of intra-subfamily sequence homology as less time was available for genetic drift-induced divergence.

### 4.4.1 *craTEs*-estimated *cis*-regulatory activities, a new metric for studying TE-mediated gene regulation from transcriptomic data

*craTEs* was particularly apt at recovering TE-mediated *cis*-regulation from RNA-seq data alone, including in cases where TE transcription did not correlate with enhancer activity, as for SVAs under KLF4 overexpression in hESCs. This speaks for the usefulness of leveraging *craTEs* for exploring published RNA-seq data as a preliminary step in the characterization of TE-mediated *cis*-regulation. In addition, *craTEs* proved superior to approaches relying on differential expression of protein-coding genes followed by enrichment for TE proximity despite using the exact same input data, namely RNA-seq and the genomic location of genes and TEs.

### 4.4.2 Evolutionary conservation is not always an appropriate proxy for function

Modeling transcription is a complex task, not least due to the sheer number of molecular players involved: hundreds of TFs, thousands of genes, hundred of thousands of CREs engaging in highly-dynamic 3D interactions, not to mention that essential mechanisms such as those by which TFs locate their genomic targets are still poorly understood. As a consequence, feature selection often exploits conservation metrics as proxies for function (Balwierz et al., 2014). However, we found that evolutionarily young TEs detectably and robustly fine-tune protein coding gene expression under the action of deeply conserved TFs. This demonstrates the

limitations of using conservation as a proxy for functional relevance, for instance to palliate for the poor specificity of widely used TF DNA binding motifs.

### 4.4.3 TE-mediated *cis*-regulation surges during specific windows of transcriptional reorganization

Working for a long time in developmental genomics can induce a sort of cessity to insights imported from other domains. Most damagingly, one may build a personal picture of gene expression as the primary explanation to any sort of patterning observed during embryogenesis. This, however, is unlikely to be the case. Biophysical simulations suggest that complex cell migration patterns such as those taking place during gastrulation can emerge spontaneously out of collectives of individual cells behaving according to simple rules, e.g dividing at a specific rates, displaying a certain range of cell-to-cell interactive forces, etc. (Van Drongelen, Vazquez-Faci, Huijben, Van Der Zee, & Idema, 2018). This suggests that changes in gene expression in the developing embryo probably do not unfold linearly, but rather in bursts inducing sudden alterations to cellular parameters such as stickiness, proliferation rates or stiffness, before proliferation ensues under relatively stable gene expression programs. By the same token, TE-mediated *cis*-regulation is likely not linear throughout development, as evidenced by the short-lived and mesendoderm-specific surge in GATA6/EOMES-LTR6B activity we observed during endoderm differentiation.

### 4.4.4 Collectives of neo-insertions generate transcriptional variation

Our work underlines that TEs affect gene expression as collectives of homologous CRE platforms. Thus, while isolated TE neo-insertions may certainly become co-opted to the benefit of the host, our results support the view that gene regulation may evolve through waves of regulatory rewiring sustained by the genomic spread of TEs.

### 4.4.5 TE- and TE controller-mediated regulatory novelty persists during and beyond gastrulation

The magnitude of TE-mediated *cis*-regulation we and others observed in early embryogenesis almost certainly reflects the strong selective pressure exerted on TEs by sexual reproduction and its inevitable germline-somatic divide. Indeed, only germline neo-insertions may be transmitted vertically to the offspring, though rare and intriguing cases of emerging somatic-to-germline horizontal infectivity have been reported (Senti et al., 2023). This justifies interpreting TE-mediated *cis*-regulation in hESCs as a byproduct of the selfish transcriptional activity of TEs. However, we observed substantial TE-mediated *cis*-regulation during developmental stages at which germline committment may already have taken place, such as during formation of the definitive endoderm and during gastrulation (Saitou & Hayashi, 2021). While our results were largely obtained through the study of *in vitro* embryonic differentiation, which almost certainly differs from its *in vivo* counterpart, they nonetheless suggest that previously unappreciated levels of TE-mediated *cis*-regulation take place past germ cell committment, in developing somatic tissues. This suggests that lineage-specific waves of TE invasions contribute to CRE turnover, thereby shaping developmental GRNs controlled by otherwise conserved *trans*-acting factors. Metazoan adaptation thus appears to emerge out of a rich genomic ecosystem where mobile genetic elements, their cognate controllers and deeply conserved developmental TFs symbiotically propagate in the gene pool.

# 5 Reverse-engineering Science Studies for Life Sciences researchers

**Cyril Pulver**, Laboratory of Virology and Genetics, School of Life Sciences, EPFL

What is the shortest path to becoming a proficient researcher in Life Sciences? If socio-economical welfare depends even only marginally on technological breakthroughs buttressed by scientific discoveries (Economics, 2017), then one ought to pay attention. The question may be reformulated as follows: which skills should aspiring researchers develop during their training to become efficient discoverers of well-grounded scientific knowledge? Contemporary university curricula, in particular research-oriented Masters- and doctoral-level coursebooks, are attempts at reifying this training. There, one expectedly finds courses covering what the field considers as established theoretical frameworks and facts forming a bedrock of accepted knowledge from which new enquiries may begin (e.g. "Cancer biology"). These courses are sometimes supplemented with round-views of what the academic community considers "pressing questions" (e.g. "Hot topics in cancer"). Quite prominent as well are courses covering methodologies and techniques – most of the times imported from other fields – susceptible to yield solutions to unanswered questions (e.g. "Applied biomedical signal processing").

Is completing this disparate array of courses sufficient to shape zealous students into full-fledged researchers? Judging by the structure of extant curricula, one has to answer by a categorical "no". Becoming a proficient scientist requires extensive practical research expe-

rience, only attainable by engaging in hands-on lab work under the guidance of seasoned researchers. Indeed, postgraduate degrees are typically delivered only once an original research contribution attributed to the student receives approval by a committee of senior peers. Approximated as a summation over academic requirements, scientific training parallels Thomas Kuhn's *paradigm*: an interwoven meshwork of theoretical, technical and tacit know-how which simultaneously guides and corrals researchers as they work away at open problems in their fields (Kuhn, 1962; Ladyman, 2002). We may thus once again reformulate our opening question: what is the most efficient way of mastering these theoretical, technical and tacit skills?

Theory – and this includes background facts – ought to be accessible through textbooks and articles. Technical knowledge ought to be accessible through protocols, or when the latter fail to deliver on replicability, by direct communication with the research group who developed or applied the technique. In contrast, tacit know-how, precisely because little institutional effort is invested in making it explicit, may chiefly - if not only - be attained "on the job" by engaging in lab work by the side of proficient researchers[I]. What therefore distinguishes the means for attaining tacit know-how versus theory and technique is a relative lack of educational resources available independently from one's own practical experience. And since much of a beginner researcher's work is molded by local conditions such as research culture, levels of funding and mentoring styles, post-graduation proficiency in tacit know-how is prone to vary. By the same token, expounding tacit know-how may contribute to levelling out inequalities caused by the intrinsic variability of postgraduate education, ultimately enabling beginner and established researchers alike to see their work in a new light.

Which skills fit under the umbrella of tacit know-how? Our best guess is: anything that researchers classically hold as merely secondary to what they regard as the core essence of their work, namely their preferred flavor of the *scientific method*[II]. Indeed, acknowledging

[I] We define tacit know-how as unbeknownst to those skillful at it, yet explainable. This slightly differs from Polanyi's much discussed tacit knowledge which emphasizes subliminality, remains unspecifiable and may thus only be transmitted from teacher to student through side-to-side interactions in the workplace.

[II] We suspect that the extent to which the researchers are willing to disclose tacit know-how correlates with the degree to which they perceive it as orthogonal to the "actual study" of scientific objects themselves, namely facts

that academic success partially depends upon specialized knowledge and practices, some scientists have started to lift the veil on tacit know-how by disseminating educational resources on writing (Booth, Colomb, Williams, Bizup, & FitzGerald, 2016; Schimel, 2012; Turabian, 2018), communicating, management, graphic design, etc. (see for example the Chicago Guides to Writing, Editing, and Publishing). Rare and notable exceptions notwithstanding (Schimel, 2012) this body of work tends to frame the expounded skills as divorced from the scientific method. For instance, establishing matters of fact is often trivialized as mere data gathering, pictured as relatively straightforward contrasted with the highly intellectual and therefore valued process of devising testable conjectures (Booth et al., 2016; Laake, 2007; Laake, Benestad, & Olsen, 2015; Turabian, 2018) following normative algorithm-like procedures adapted from Popperian falsifiability (e.g. the hypothetico-deductive method) (Laake et al., 2015; Ladyman, 2002; Popper, 1935).

At this point, Life Sciences researchers prompted to compare their everyday lab schedule with those depictions of Science may be given pause. Indeed, they may fail to pinpoint a single paper explicitly endorsing Popperian falsification in the literature they cherish, thereby concluding that most of their peers are methodologically illiterate at best or hypocrites at worse. They may wonder why their senior colleagues remain unconcerned by the only sensible Popperian interpretation to be derived from the pile of failed experiments and inconclusive results clogging the notebooks and hard drives of the lab: that the working hypothesis simply should be abandoned. They may gradually show contempt for colleagues who tinker with vast amounts of loosely connected measurements in attempts to "build a story", or find "confirmatory" evidence supporting their favorite model. They may consequently turn to exemplars for guidance, in particular maxims attributed to illustrious scientists such as Claude Bernard (Curd, Cover, & Pincock, 2013) or Richard Feynman (Feynman, 1974), frequently presented as champions of a scientific method rooted in uncompromising philosophical doubt and razor-sharp critical thinking (Latour & Woolgar, 1988). Thus, our bewildered researcher, at first looking to bridge the gap between day-to-day lab work and polished – i.e.

and theories.

publishable - matters of facts is brought back full-circle to scientific textbook descriptions of the scientific method that belittle facts, aggrandize theories and often tunnel-vision on Physics as the preeminent science to be explained, at the detriment of other fields (Bonneuil & Pestre, 2015; Hacking, 2012).

What Science Studies have explored since their inception in the 1970s is the possibility that a myriad of factors - including but certainly not limited to the critical appraisal of experimental results - may better account for the robustness of established scientific facts and theories than hindsight and sometimes quite contrived appeals to rationality (Lakatos, 1978). Crucially, these factors are not circumvented to what researchers may explain away as biases to be controlled for, e.g. national interests, economic incentives, social norms, cognitive biases and field-specific hypes (Pestre, 2013). Rather, they include the development, operating and industrialization of measurement instruments and computing resources, the dissemination of mathematical and statistical frameworks, the preferential use of specific argumentative schemes and perhaps most importantly the multi-facetted social settings within which researchers painstakingly formulate, revise, refine, consolidate and circulate matters of fact. In a nutshell, Science Studies attempt to account for the emergence of scientific knowledge as the outcome of know-how-rich processes rather than as trickling down from metaphysical entities such as Reason and Logic.

Can Science Studies distill generalizable insights for today's Life Sciences researchers in training? We thereafter briefly illustrate how know-how laid bare by seminal Science Studies texts may be reverse-engineered for the benefit of practicing researchers. In *Laboratory Life: The Construction of Scientific Facts* (Latour & Woolgar, 1979/1986), Latour and Woolgar (LW) historicize what may at first seem like a prototypically ahistorical matter of fact: "[Thyrotropin Releasing Factor] TRF is Pyro-Glu-His-Pro-N2" (p. 280). LW stage a culture shock involving soon-to-be "Nobelized" researchers at the Salk Institute and a scientifically clueless observer versed in anthropology, sociology and philosophy, thereby capturing the tortuous negotiation process leading to the temporary and consensual settlement that characterizes successful research projects. By systematically mining text sources compiled from specialized scientific

literature, handwritten notes and transcripts of informal lab discussions, LW unveil that scientific matter-of-fact statements are positioned on a disputability scale ranging from highly speculative – i.e. controversial – to established – i.e. undisputed – assertions. The robustness of assertions is encoded in modalities, secondary statements that nuance the matter-of-fact statements. It thus follows that researchers can potentiate the robustness of statements by qualifying them with modalities. This rhetoric know-how becomes crucial for framing one's own experimental results in the highly controversial context of bleeding-edge research, an area of knowledge fraught with anomalies, inconsistencies and ensuing disputes (Collins, 1985; Shapin & Schaffer, 1985), such as when navigating the perilous peer-review process (as illustrated in Box 1). Rhetoric know-how aside, Science Studies have put a new and pragmatic spin on practices specific to contemporary lab research. Examples include justifying the opening of new research avenues, addressing replicability issues, appealing to and interpreting the output of measurement instruments, understanding the oft-forgotten contributions of technicians and efficiently disseminating one's own research results.

By endeavoring to explicate the solidity of scientific facts without resorting to cause-and-effect scientific reasoning, Science Studies have shed light on know-how until then confined to the labs and offices of the most prolific research groups. This now elucidated tacit know-how is ripe for reformulation and dissemination back into the scientific research community, where it may simultaneously accelerate learning, amend the unrealistic and often counter-productive view of "absolutely rational" research conveyed by some scientists, and nuance the prescriptive discourse emanating from classical Philosophy of Science. "Reverse-black-boxing" or - to use a term more familiar to MINT-minded readers - "reverse-engineering" the findings of Science Studies may prove particularly fruitful for Life Sciences research, which tends to capitalize on interdisciplinarity and technological innovation.

## 5.1 Box 1: a concrete reverse-engineering of Science Studies insights for Life Sciences researchers

We exemplify our project with a concrete case of peer-review drawn from our own work (Pulver et al., 2023). We detail how we drew on know-how presented in *Laboratory Life* to substantiate our rebuttal.

Briefly, the paper presents a mathematical procedure abbreviated as "*craTEs*" which estimates the activities – encoded as numbers on a continuous scale - of a first group of genetic entities called "transposable elements" on a second group of genetic entities called "protein-coding genes". After showing that the procedure recovers previously established cases of such activity, we propose to scale it to a large and recently published gene expression dataset (Nakatake et al., 2020) gathered from embryonic stem cells subjected to various experimental perturbations (Pulver et al., 2022 Figure 3A). We then discuss in broader terms what the activities we computed may imply for the field of gene regulation.

One reviewer took issue with some of the activities we computed, pointing out that they appeared at odds with an established matter of fact (at least in the small field studying transposable elements and their impact on gene regulation):

Reviewer: *There are certain instances from the analysis of the TF overexpression data in figure 3A and additional file 2 (heatmaps) which does [sic] not make sense.* ***For example, KLF4 seems not to enrich [sic] over LTR7***

*(not the same as what was shown in Ohnuki et al., (PNAS) and Carter et al., (eLife))*

This in turn threatened the validity of our procedure. We thus leveraged two rhetoric know-hows made explicit in *Laboratory Life*:

- We qualified the matter-of-fact statement "KLF4 activates LTR7 in embryonic stem cells" with **modalities that draw attention to the local circumstances in which the fact was**

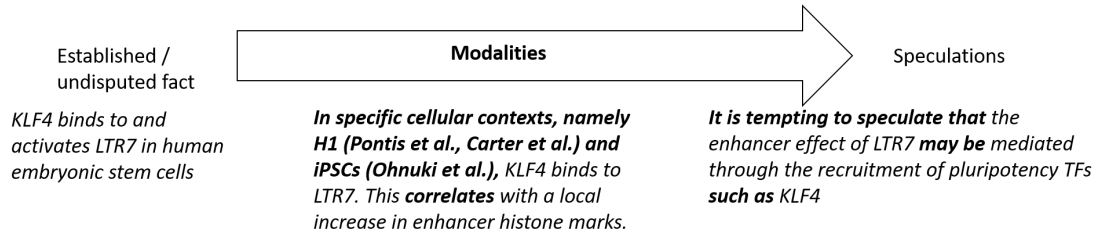
**reported** as well as **how these circumstances differ from those in which the dataset we analyzed was collected**. This fragilizes the generalizability of the statement, thereby managing an ontological space for our anomalous results (Latour and Woolgar, 1979/1986 p. 156).

- We superimposed "several statements or documents in such a way that all the statements were seen to relate to something outside of, or beyond, the reader's or author's subjectivity." (Latour and Woolgar, 1979/1986, p. 159). First, **we appealed to inscriptions reported in the recently published study (Nakatake et al., 2020) from which the dataset was drawn to safeguard the objectivity of our own mathematical approach**. Second, **we appealed to published literature ascribing an until then unmentioned function (controlling late gut development) to the scientific object at play (KLF4)**.

We color-coded the parts of our reply according to the know-how that applies:

Our rebuttal: "*This indeed comes as a peculiar observation, against the fact that we (Pontis et al., Cell Stem Cell 2019, this manuscript, fig. 1, fig. 3B) among others (Ohnuki et al., PNAS, Carter et al. Elife 2022)* ***have reported at several occasions*** *that KLF4 binds to and transcriptionally activates LTR7 in human embryonic and pluripotent stem cells. However,* ***KLF4 has also been characterized as a master regulator of endoderm differentiation (e.g terminal differentiation of goblet cells Katz et al., Development 2002)***. *It is thus possible that* ***Natakake et al's hESC cell line (SEES3, not the canonical H1 cell line analyzed in Carter et al., 2022)*** *reacts to* ***rapid and elevated overexpression*** *of KLF1, KLF2 and KLF4 by triggering some differentiation-specific mechanisms, thus polarizing them away from the ESC state.* ***Supporting that observation is the fact that unlike induction of KLF5, which leads to an "ESC" visual phenotype after 3 days of induction, induction of KLF1, KLF2 and KLF4 respectively lead to "spiky"/spiky/"floating, proliferation or died" phenotypes (Nakatake et al., Cell Reports 2020, Table S1). Accordingly, craTEs uncovers a KLF5-dependent increase in LTR7 activity, and KLF1, KLF2 and KLF4-dependent decreases in LTR7 activity.***"

More generally, one can schematize the use of modalities in the construction of scientific facts as reported by LW as follows:




### 5.1.1 Acknowledgements

We warmly thank Bhargav Srinivasa Desikan, Aude Fauvel, Damon Kutzin, Jonas Pulver and Didier Trono for engaging in fruitful discussions, making suggestions and welcoming spontaneous talks and presentations on the topic.

# Bibliography


Ambrosini, G., Vorontsov, I., Penzar, D., Groux, R., Fornes, O., Nikolaeva, D. D., … Bucher, P. (2020). Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study. *Genome Biology*, *21*(1), 114. doi:10.1186/s13059-020-01996-3

Armenteros-Monterroso, E., Zhao, L., Gasparoli, L., Brooks, T., Pearce, K., Mansour, M. R., … Williams, O. (2019). The AAA+ATPase RUVBL2 is essential for the oncogenic function of c-MYB in acute myeloid leukemia. *Leukemia*, *33*(12), 2817–2829. doi:10.1038/s41375-019-0495-8

Avery, O. T., MacLeod, C. M., & McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *Journal of Experimental Medicine*, *79*(2), 137–158. doi:10.1084/jem.79.2.137

Balwierz, P. J., Pachkov, M., Arnold, P., Gruber, A. J., Zavolan, M., & van Nimwegen, E. (2014). ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome Research*, *24*(5), 869–884. doi:10.1101/gr.169508.113

Barakat, T. S., Halbritter, F., Zhang, M., Rendeiro, A. F., Perenthaler, E., Bock, C., & Chambers, I. (2018). Functional Dissection of the Enhancer Repertoire in Human Embryonic Stem Cells. *Cell Stem Cell*, *23*(2), 276–288.e8. doi:10.1016/j.stem.2018.06.014

Benjamini, Y. & Hochberg, Y. (1995). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

Bernardo, A. S., Faial, T., Gardner, L., Niakan, K. K., Ortmann, D., Senner, C. E., … Pedersen, R. A. (2011). BRACHYURY and CDX2 Mediate BMP-Induced Differentiation of Human and Mouse Pluripotent Stem Cells into Embryonic and Extraembryonic Lineages. *Cell Stem Cell*, *9*(2), 144–155. doi:10.1016/j.stem.2011.06.015

Bonneuil, C. & Pestre, D. (2015). Le siècle des technosciences (depuis 1914). In D. Pestre & S. v. Damme (Eds.), *Histoire des sciences et des savoirs 3* (pp. 9–24). Paris: Éditions du Seuil.

Booth, W. C., Colomb, G. G., Williams, J. M., Bizup, J., & FitzGerald, W. T. (2016). *The craft of research* (Fourth edition). Chicago guides to writing, editing, and publishing. Chicago: The University of Chicago Press.

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., … Feschotte, C. (2018). Ten things you should know about transposable elements. *Genome Biology*, *19*(1), 199. doi:10.1186/s13059-018-1577-z

Bourque, G., Leong, B., Vega, V. B., Chen, X., Lee, Y. L., Srinivasan, K. G., … Liu, E. T. (2008). Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Research*, *18*(11), 1752–1762. doi:10.1101/gr.080663.108

Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., … Young, R. A. (2005). Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells. *Cell*, *122*(6), 947–956. doi:10.1016/j.cell.2005.08.020

Britten, R. J. & Davidson, E. H. (1969). Gene Regulation for Higher Cells: a Theory: New facts regarding the organization of the genome provide clues to the nature of gene regulation. *Science*, *165*(3891), 349–357. doi:10.1126/science.165.3891.349

Britten, R. J. & Davidson, E. H. (1971). Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. *The Quarterly Review of Biology*, *46*(2), 111–138. doi:10.1086/406830

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, *10*(12), 1213–1218. doi:10.1038/nmeth.2688

Bussemaker, H. J., Foat, B. C., & Ward, L. D. (2007). Predictive Modeling of Genome-Wide mRNA Expression: From Modules to Molecules. *Annual Review of Biophysics and Biomolecular Structure*, *36*(1), 329–347. doi:10.1146/annurev.biophys.36.040306.132725

Bussemaker, H. J., Li, H., & Siggia, E. D. (2001). Regulatory element detection using correlation with expression. *Nature Genetics*, *27*(2), 167–171. doi:10.1038/84792

Canver, M. C., Lessard, S., Pinello, L., Wu, Y., Ilboudo, Y., Stern, E. N., … Orkin, S. H. (2017). Variant-aware saturating mutagenesis using multiple Cas9 nucleases identifies regulatory elements at trait-associated loci. *Nature Genetics*, *49*(4), 625–634. doi:10.1038/ng.3793

Carlton, V. E. H., Harris, B. Z., Puffenberger, E. G., Batta, A. K., Knisely, A. S., Robinson, D. L., … Bull, L. N. (2003). Complex inheritance of familial hypercholanemia with associated mutations in TJP2 and BAAT. *Nature Genetics*, *34*(1), 91–96. doi:10.1038/ng1147

Carroll, S. B. (2008). Evo-Devo and an Expanding Evolutionary Synthesis: a Genetic Theory of Morphological Evolution. *Cell*, *134*(1), 25–36. doi:10.1016/j.cell.2008.06.030

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

Chavez, A., Scheiman, J., Vora, S., Pruitt, B. W., Tuttle, M., P R Iyer, E., … Church, G. M. (2015). Highly efficient Cas9-mediated transcriptional programming. *Nature Methods*, *12*(4), 326–328. doi:10.1038/nmeth.3312

Chen, D., Sun, N., Hou, L., Kim, R., Faith, J., Aslanyan, M., … Clark, A. (2019). Human Primordial Germ Cells Are Specified from Lineage-Primed Progenitors. *Cell Reports*, *29*(13), 4568–4582.e5. doi:10.1016/j.celrep.2019.11.083

Chen, M. J., Yokomizo, T., Zeigler, B. M., Dzierzak, E., & Speck, N. A. (2009). Runx1 is required for the endothelial to haematopoietic cell transition but not thereafter. *Nature*, *457*(7231), 887–891. doi:10.1038/nature07619

Chia, N.-Y., Deng, N., Das, K., Huang, D., Hu, L., Zhu, Y., … Tan, P. (2015). Regulatory crosstalk between lineage-survival oncogenes KLF5, GATA4 and GATA6 cooperatively promotes gastric cancer development. *Gut*, *64*(5), 707–719. doi:10.1136/gutjnl-2013-306596

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

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

Chuong, E. B., Rumi, M. A. K., Soares, M. J., & Baker, J. C. (2013). Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nature Genetics*, *45*(3), 325–329. doi:10.1038/ng.2553

Cohen, C. J., Lock, W. M., & Mager, D. L. (2009). Endogenous retroviral LTRs as promoters for human genes: a critical assessment. *Gene*, *448*(2), 105–114. doi:10.1016/j.gene.2009.06.020

Collins, H. (1985). *Changing Order: Replication and Induction in Scientific Practice*. Chicago: University of Chicago Press.

Crick, F. H. (1958). On protein synthesis. In *Symp Soc Exp Biol* (Vol. 12, p. 8). Issue: 138-63.

Cunningham, F., Allen, J. E., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., … Flicek, P. (2022). Ensembl 2022. *Nucleic Acids Research*, *50*(D1), D988–D995. doi:10.1093/nar/gkab1049

Pierre Duhem: Physical Theory and Experiment. (2013). In M. Curd, J. A. Cover, & C. Pincock (Eds.), *Philosophy of science: the central issues* (2nd ed, pp. 227–249). New York: W.W. Norton.

D'Amour, K. A., Bang, A. G., Eliazer, S., Kelly, O. G., Agulnick, A. D., Smart, N. G., … Baetge, E. E. (2006). Production of pancreatic hormone–expressing endocrine cells from human embryonic stem cells. *Nature Biotechnology*, *24*(11), 1392–1401. doi:10.1038/nbt1259

Darwin, C. & Beer, G. (2008). *On the origin of species* (Rev. ed). Oxford world's classics. New York: Oxford University Press. (Original work published 1859)

Darwin, C. & Wallace, A. (1858). On the Tendency of Species to form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection. *Journal of the*

*Proceedings of the Linnean Society of London. Zoology*, *3*(9), 45–62. doi:10.1111/j.1096-3642.1858.tb02500.x

Dawkins, R. (2016). *The selfish gene* (40th anniversary edition). Oxford landmark science. Oxford New York, NY: Oxford University Press. (Original work published 1976)

De Tribolet-Hardy, J. C. (2022). *KRAB zinc-finger proteins and their transposable element targets: between antagonism and cooperation* (Doctoral dissertation, EPFL). Publisher: Lausanne, EPFL.

De Tribolet-Hardy, J., Thorball, C. W., Forey, R., Planet, E., Duc, J., Coudray, A., … Trono, D. (2023). Genetic features and genomic targets of human KRAB-zinc finger proteins. *Genome Research*, *33*(8), 1409–1423. doi:10.1101/gr.277722.123

De Tribolet-Hardy, J., Thorball, C. W., Forey, R., Planet, E., Duc, J., Khubieh, B., … Trono, D. (2023). *Genetic features and genomic targets of human KRAB-Zinc Finger Proteins*. Genomics. doi:10.1101/2023.02.27.530095

Deniz, Ö., Ahmed, M., Todd, C. D. [Christopher D.], Dawson, M. A., & Branco, M. R. (2019). Endogenous retroviruses are a source of oncogenic enhancers in acute myeloid leukemia [RNA-Seq]. Accession number: GSE136763. Gene Expression Omnibus (GEO).

Deniz, Ö., Ahmed, M., Todd, C. D., Rio-Machin, A., Dawson, M. A., & Branco, M. R. (2020). Endogenous retroviruses are a source of enhancers with oncogenic potential in acute myeloid leukaemia. *Nature Communications*, *11*(1), 3506. doi:10.1038/s41467-020-17206-4

Deniz, Ö., Frost, J. M., & Branco, M. R. (2019). Regulation of transposable elements by DNA modifications. *Nature Reviews Genetics*, *20*(7), 417–431. doi:10.1038/s41576-019-0106-6

Diehl, A. G., Ouyang, N., & Boyle, A. P. (2020). Transposable elements contribute to cell and species-specific chromatin looping and gene regulation in mammalian genomes. *Nature Communications*, *11*(1), 1796. doi:10.1038/s41467-020-15520-5

Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., … Ren, B. (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature*, *518*(7539), 331–336. doi:10.1038/nature14222

Domcke, S., Hill, A. J., Daza, R. M., Cao, J., O'Day, D. R., Pliner, H. A., … Shendure, J. (2020). A human cell atlas of fetal chromatin accessibility. *Science*, *370*(6518), eaba7612. doi:10.1126/science.aba7612

Doolittle, W. F. (2022). All about levels: transposable elements as selfish DNAs and drivers of evolution. *Biology & Philosophy*, *37*(4), 24. doi:10.1007/s10539-022-09852-3

Doolittle, W. F. & Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature*, *284*(5757), 601–603. doi:10.1038/284601a0

Duboule, D. (1994). Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Development (Cambridge, England). Supplement*, 135–142.

Ecco, G., Cassano, M., Kauzlaric, A., Duc, J., Coluccio, A., Offner, S., … Trono, D. (2016). Transposable Elements and Their KRAB-ZFP Controllers Regulate Gene Expression in Adult Tissues. *Developmental Cell*, *36*(6), 611–623. doi:10.1016/j.devcel.2016.02.024

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

Economics, B. (2017). The Economic Contribution of the Institutions of the ETH Domain.

Ernst, J. & Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, *28*(8), 817–825. doi:10.1038/nbt.1662

FANTOM Consortium, Suzuki, H., Forrest, A. R. R., van Nimwegen, E., Daub, C. O., Balwierz, P. J., … Hayashizaki, Y. (2009). The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nature Genetics*, *41*(5), 553–562. doi:10.1038/ng.375

Feinberg, M. W., Wara, A. K., Cao, Z., Lebedeva, M. A., Rosenbauer, F., Iwasaki, H., … Jain, M. K. (2007). The Kruppel-like factor KLF4 is a critical regulator of monocyte differentiation. *The EMBO Journal*, *26*(18), 4138–4148. doi:10.1038/sj.emboj.7601824

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

Feschotte, C. (2023). Transposable elements: McClintock's legacy revisited. *Nature Reviews Genetics*. doi:10.1038/s41576-023-00652-3

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

Feynman, R. (1974). Cargo Cult Science. *Engineering and Science*, *37*(7), 10–13.

Friedli, M. & Trono, D. (2015). The developmental control of transposable elements and the evolution of higher species. *Annual Review of Cell and Developmental Biology*, *31*, 429–451. doi:10.1146/annurev-cellbio-100814-125514

Fuentes, D. R., Swigut, T., & Wysocka, J. (2018a). Systematic perturbation of retroviral LTRs reveals widespread long-range effects on human gene regulation. Accession number: GSE111337. Gene Expression Omnibus (GEO).

Fuentes, D. R., Swigut, T., & Wysocka, J. (2018b). Systematic perturbation of retroviral LTRs reveals widespread long-range effects on human gene regulation. *eLife*, *7*, e35989. doi:10.7554/eLife.35989

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

Gao, L., Wu, K., Liu, Z., Yao, X., Yuan, S., Tao, W., … Liu, J. (2018). Chromatin Accessibility Landscape in Human Early Embryos and Its Association with Evolution. *Cell*, *173*(1), 248–259.e15. doi:10.1016/j.cell.2018.02.028

Gautam, S., Fioravanti, J., Zhu, W., Le Gall, J. B., Brohawn, P., Lacey, N. E., … Gattinoni, L. (2019). The transcription factor c-Myb regulates CD8+ t cell stemness and antitumor immunity. *Nature Immunology*, *20*(3), 337–349. doi:10.1038/s41590-018-0311-z

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., … Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, *5*(10), R80. doi:10.1186/gb-2004-5-10-r80

Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K.-K., Cheng, C., … Snyder, M. (2012). Architecture of the human regulatory network derived from ENCODE data. *Nature*, *489*(7414), 91–100. doi:10.1038/nature11245

Gertz, J., Savic, D., Varley, K. E., Partridge, E. C., Safi, A., Jain, P., … Myers, R. M. (2013). Distinct properties of cell-type-specific and shared transcription factor binding sites. *Molecular Cell*, *52*(1), 25–36. doi:10.1016/j.molcel.2013.08.037

Gilbert, L. A., Larson, M. H., Morsut, L., Liu, Z., Brar, G. A., Torres, S. E., … Qi, L. S. (2013). CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. *Cell*, *154*(2), 442–451. doi:10.1016/j.cell.2013.06.044

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

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

Grant, C. E., Bailey, T. L., & Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics*, *27*(7), 1017–1018. doi:10.1093/bioinformatics/btr064

Greenberg, M. V. C. & Bourc'his, D. (2019). The diverse roles of DNA methylation in mammalian development and disease. *Nature Reviews Molecular Cell Biology*, *20*(10), 590–607. doi:10.1038/s41580-019-0159-6

Griffith, F. (1928). The Significance of Pneumococcal Types. *Journal of Hygiene*, *27*(2), 113–159. doi:10.1017/S0022172400031879

Grow, E. J., Flynn, R. A., Chavez, S. L., Bayless, N. L., Wossidlo, M., Wesche, D. J., … Wysocka, J. (2015). Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature*, *522*(7555), 221–225. doi:10.1038/nature14308

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, *32*(18), 2847–2849. doi:10.1093/bioinformatics/btw313

Guo, F., Yan, L., Guo, H., Li, L., Hu, B., Zhao, Y., … Qiao, J. (2015). The Transcriptome and DNA Methylome Landscapes of Human Primordial Germ Cells. *Cell*, *161*(6), 1437–1452. doi:10.1016/j.cell.2015.05.015

Guo, H., Zhu, P., Yan, L., Li, R., Hu, B., Lian, Y., … Qiao, J. (2014). The DNA methylation landscape of human early embryos. *Nature*, *511*(7511), 606–610. doi:10.1038/nature13544

Hacking, I. (2012). Introductory Essay. In T. S. Kuhn (Ed.), *The Structure of Scientific Revolutions*. University of Chicago Press.

Haeussler, M., Schönig, K., Eckert, H., Eschstruth, A., Mianné, J., Renaud, J.-B., … Concordet, J.-P. (2016). Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biology*, *17*(1), 148. doi:10.1186/s13059-016-1012-2

Haring, N. L., Van Bree, E. J., Jordaan, W. S., Roels, J. R., Sotomayor, G. C., Hey, T. M., … Jacobs, F. M. (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 Research*, *31*(4), 551–563. doi:10.1101/gr.265348.120

Hawkins, R. D., Hon, G. C., Lee, L. K., Ngo, Q., Lister, R., Pelizzola, M., … Ren, B. (2010). Distinct Epigenomic Landscapes of Pluripotent and Lineage-Committed Human Cells. *Cell Stem Cell*, *6*(5), 479–491. doi:10.1016/j.stem.2010.03.018

He, J., Babarinde, I. A., Sun, L., Xu, S., Chen, R., Shi, J., … Chen, J. (2021). Identifying transposable element expression dynamics and heterogeneity during development at the single-cell level with a processing pipeline scTE. *Nature Communications*, *12*(1), 1456. doi:10.1038/s41467-021-21808-x

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., … Glass, C. K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and b Cell Identities. *Molecular Cell*, *38*(4), 576–589. doi:10.1016/j.molcel.2010.05.004

Heslop, J. A., Pournasr, B., Liu, J.-T., & Duncan, S. A. (2021a). GATA6 defines endoderm fate by controlling chromatin accessibility during differentiation of human induced pluripotent stem cells. Accession number: GSE156021. Gene Expression Omnibus (GEO).

Heslop, J. A., Pournasr, B., Liu, J.-T., & Duncan, S. A. (2021b). GATA6 defines endoderm fate by controlling chromatin accessibility during differentiation of human-induced pluripotent stem cells. *Cell Reports*, *35*(7), 109145. doi:10.1016/j.celrep.2021.109145

Hinrichs, A. S. (2006). The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research*, *34*(90001), D590–D598. doi:10.1093/nar/gkj144

Hirakawa, M., Nishihara, H., Kanehisa, M., & Okada, N. (2009). Characterization and evolutionary landscape of AmnSINE1 in Amniota genomes. *Gene*, *441*(1-2), 100–110. doi:10.1016/j.gene.2008.12.009

Horb, M. & Thomsen, G. (1999). Tbx5 is essential for heart development. *Development*, *126*(8), 1739–1751. doi:10.1242/dev.126.8.1739

Horton, I., Kelly, C. J., Dziulko, A., Simpson, D. M., & Chuong, E. B. (2023). Mouse B2 SINE elements function as IFN-inducible enhancers. *eLife*, *12*, e82617. doi:10.7554/eLife.82617

Hoyt, S. J., Storer, J. M., Hartley, G. A., Grady, P. G. S., Gershman, A., De Lima, L. G., … O'Neill, R. J. (2022). From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science*, *376*(6588), eabk3112. doi:10.1126/science.abk3112

Huang, K., Du, J., Ma, N., Liu, J., Wu, P., Dong, X., … Pan, G. (2015). GATA2 –/– human ESCs undergo attenuated endothelial to hematopoietic transition and thereafter granulocyte commitment. *Cell Regeneration*, *4*(1), 4:4. doi:10.1186/s13619-015-0018-7

Huang, K., Du, J., Shi, X., Chen, Q., & Pan, G. (2017). GATA2 knockout study to investigated the role of GATA2 in human hematopoiesis. Accession number: GSE69797. Gene Expression Omnibus (GEO).

Imbeault, M., Helleboid, P.-Y., & Trono, D. (2017). KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature*, *543*(7646), 550–554. doi:10.1038/nature21683

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. doi:10.1038/35057062

Iouranova, A., Grun, D., Rossy, T., Duc, J., Coudray, A., Imbeault, M., … Trono, D. (2022). KRAB zinc finger protein ZNF676 controls the transcriptional influence of LTR12-related

endogenous retrovirus sequences. *Mobile DNA*, *13*(1), 4. doi:10.1186/s13100-021-00260-0

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

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

Jacques, P.-É., Jeyakani, J., & Bourque, G. (2013). The majority of primate-specific regulatory sequences are derived from transposable elements. *PLoS genetics*, *9*(5), e1003504. doi:10.1371/journal.pgen.1003504

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer Texts in Statistics. New York, NY: Springer New York. doi:10.1007/978-1-4614-7138-7

Jang, H. S., Shah, N. M., Du, A. Y., Dailey, Z. Z., Pehrsson, E. C., Godoy, P. M., … Wang, T. (2019). Transposable elements drive widespread expression of oncogenes in human cancers. *Nature Genetics*, *51*(4), 611–617. doi:10.1038/s41588-019-0373-3

Jostes, S. V., Fellermeyer, M., Arévalo, L., Merges, G. E., Kristiansen, G., Nettersheim, D., & Schorle, H. (2020). Unique and redundant roles of SOX2 and SOX17 in regulating the germ cell tumor fate. *International Journal of Cancer*, *146*(6), 1592–1605. doi:10.1002/ijc.32714

Katoh, K. & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780. doi:10.1093/molbev/mst010

Katz, J. P., Perreault, N., Goldstein, B. G., Lee, C. S., Labosky, P. A., Yang, V. W., & Kaestner, K. H. (2002). The zinc-finger transcription factor Klf4 is required for terminal differentiation of goblet cells in the colon. *Development*, *129*(11), 2619–2628. doi:10.1242/dev.129.11.2619

Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S., & Karolchik, D. (2010). BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, *26*(17), 2204–2207. doi:10.1093/bioinformatics/btq351

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357–360. doi:10.1038/nmeth.3317

King, M.-C. & Wilson, A. C. (1975). Evolution at Two Levels in Humans and Chimpanzees: Their macromolecules are so alike that regulatory mutations may account for their biological differences. *Science*, *188*(4184), 107–116. doi:10.1126/science.1090005

Kojima, K. K. (2018). Human transposable elements in Repbase: genomic footprints from fish to humans. *Mobile DNA*, *9*(1), 2. doi:10.1186/s13100-017-0107-y

Krendl, C., Shaposhnikov, D., Rishko, V., Ori, C., Ziegenhain, C., Sass, S., … Drukker, M. (2017). GATA2/3-TFAP2A/c transcription factor network couples human pluripotent stem cell differentiation to trophectoderm with repression of pluripotency. *Proceedings of the National Academy of Sciences*, *114*(45). doi:10.1073/pnas.1708341114

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Kunarso, G., Chia, N.-Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.-S., … Bourque, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature Genetics*, *42*(7), 631–634. doi:10.1038/ng.600

Laake, P. (Ed.). (2007). *Research methodology in the medical and biological sciences* (rev. and expanded version of the Norwegian ed.). Amsterdam Heidelberg: Elsevier, Academic Press.

Laake, P., Benestad, H. B., & Olsen, B. R. (Eds.). (2015). *Research in medical and biological sciences: from planning and preparation to grant application and publication*. Amsterdam ; Boston: Elsevier/Academic Press.

Ladyman, J. (2002). *Understanding philosophy of science*. London ; New York: Routledge.

Lakatos, I. (1978). *The methodology of scientific research programmes*. Cambridge: Cambridge University Press.

Lal, G., Contreras, P. G., Kulak, M., Woodfield, G., Bair, T., Domann, F. E., & Weigel, R. J. (2013). Human Melanoma Cells Over-Express Extracellular Matrix 1 (ECM1) Which Is Regulated by TFAP2C. *PLoS ONE*, *8*(9), e73953. doi:10.1371/journal.pone.0073953

Lamarck, J. B. P. A. d. M. d. (2011). *Philosophie zoologique: Ou exposition ; des considerations relative à l'histoire naturelle des animaux. Volume 1*. Cambridge: Cambridge University Press. (Original work published 1809)

Lanciano, S., Philippe, C., Sarkar, A., Pratella, D., Domrane, C., Doucet, A. J., … Cristofari, G. (2023). *Comprehensive locus-specific L1 DNA methylation profiling reveals the epigenetic and transcriptional interplay between L1s and their integration sites*. Genomics. doi:10.1101/2023.01.03.522582

Langmead, B. & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. doi:10.1038/nmeth.1923

Latour, B. & Woolgar, S. (1986). *Laboratory life: the construction of scientific facts*. Princeton, N.J: Princeton University Press. (Original work published 1979)

Latour, B. & Woolgar, S. (1988). *La Vie de laboratoire : la production des faits scientifiques*. Éditions La Découverte.

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

Lee, K., Cho, H., Rickert, R. W., Li, Q. V., Pulecio, J., Leslie, C. S., & Huangfu, D. (2019). FOXA2 Is Required for Enhancer Priming during Pancreatic Differentiation. *Cell Reports*, *28*(2), 382–393.e7. doi:10.1016/j.celrep.2019.06.034

Li, L., Wang, Y., Torkelson, J. L., Shankar, G., Pattison, J. M., Zhen, H. H., … Oro, A. E. (2019). TFAP2C- and p63-Dependent Networks Sequentially Rearrange Chromatin Landscapes to Drive Human Epidermal Lineage Commitment. *Cell Stem Cell*, *24*(2), 271–284.e8. doi:10.1016/j.stem.2018.12.012

Li, Q. V., Dixon, G., Verma, N., Rosen, B. P., Gordillo, M., Luo, R., … Huangfu, D. (2019). Genome-scale screens identify JNK-JUN signaling as a barrier for pluripotency exit and endoderm differentiation. *Nature Genetics*, *51*(6), 999–1010. doi:10.1038/s41588-019-0408-9

Li, W., Sun, G., Yang, S., Qu, Q., Nakashima, K., & Shi, Y. (2008). Nuclear Receptor TLX Regulates Cell Cycle Progression in Neural Stem Cells of the Developing Brain. *Molecular Endocrinology*, *22*(1), 56–64. doi:10.1210/me.2007-0290

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

Liu, L., Leng, L., Liu, C., Lu, C., Yuan, Y., Wu, L., … Lin, G. (2019). An integrated chromatin accessibility and transcriptome landscape of human pre-implantation embryos. *Nature Communications*, *10*(1), 364. doi:10.1038/s41467-018-08244-0

Loh, K. M., Ang, L. T., Zhang, J., Kumar, V., Ang, J., Auyeong, J. Q., … Lim, B. (2014). Efficient Endoderm Induction from Human Pluripotent Stem Cells by Logically Directing Signals Controlling Lineage Bifurcations. *Cell Stem Cell*, *14*(2), 237–252. doi:10.1016/j.stem.2013.12.007

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

Luo, R., Huangfu, D., & Beer, M. A. (2022). Dynamic network-guided CRISPRi screen reveals CTCF loop constrained enhancer function in cell state transitions. Accession number: GSE213394. Gene Expression Omnibus (GEO).

Luo, R., Yan, J., Oh, J. W., Xi, W., Shigaki, D., Wong, W., … Beer, M. A. (2023). Dynamic network-guided CRISPRi screen identifies CTCF-loop-constrained nonlinear enhancer gene regulatory activity during cell state transitions. *Nature Genetics*, *55*(8), 1336–1346. doi:10.1038/s41588-023-01450-7

Lynch, V. J., Leclerc, R. D., May, G., & Wagner, G. P. (2011). Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nature Genetics*, *43*(11), 1154–1159. doi:10.1038/ng.917

Lyu, X., Rowley, M. J., & Corces, V. G. (2018). Architectural Proteins and Pluripotency Factors Cooperate to Orchestrate the Transcriptional Response of hESCs to Temperature Stress. *Molecular Cell*, *71*(6), 940–955.e7. doi:10.1016/j.molcel.2018.07.012

Macfarlan, T. S., Gifford, W. D., Driscoll, S., Lettieri, K., Rowe, H. M., Bonanomi, D., … Pfaff, S. L. (2012). Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature*, *487*(7405), 57–63. doi:10.1038/nature11244

Matsushita, M., Nakatake, Y., Arai, I., Ibata, K., Kohda, K., Goparaju, S. K., … Ko, M. S. (2017). Neural differentiation of human embryonic stem cells induced by the transgene-mediated overexpression of single transcription factors. *Biochemical and Biophysical Research Communications*, *490*(2), 296–301. doi:10.1016/j.bbrc.2017.06.039

Mazumdar, C., Shen, Y., Xavy, S., Zhao, F., Reinisch, A., Li, R., … Majeti, R. (2015). Leukemia-Associated Cohesin Mutants Dominantly Enforce Stem Cell Programs and Impair Human Hematopoietic Progenitor Differentiation. *Cell Stem Cell*, *17*(6), 675–688. doi:10.1016/j.stem.2015.09.017

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

Memedula, S. & Belmont, A. S. (2003). Sequential Recruitment of HAT and SWI/SNF Components to Condensed Chromatin by VP16. *Current Biology*, *13*(3), 241–246. doi:10.1016/S0960-9822(03)00048-4

Miao, B., Fu, S., Lyu, C., Gontarz, P., Wang, T., & Zhang, B. (2020). Tissue-specific usage of transposable element-derived promoters in mouse development. *Genome Biology*, *21*(1), 255. doi:10.1186/s13059-020-02164-3

Mills, R. E., Bennett, E. A., Iskow, R. C., & Devine, S. E. (2007). Which transposable elements are active in the human genome? *Trends in Genetics*, *23*(4), 183–191. doi:10.1016/j.tig.2007.02.006

Molè, M. A., Coorens, T. H. H., Shahbazi, M. N., Weberling, A., Weatherbee, B. A. T., Gantner, C. W., … Zernicka-Goetz, M. (2021). A single cell characterisation of human embryogenesis identifies pluripotency transitions and putative anterior hypoblast centre. *Nature Communications*, *12*(1), 3679. doi:10.1038/s41467-021-23758-w

Molkentin, J. D. (2000). The Zinc Finger-containing Transcription Factors GATA-4, -5, and -6. *Journal of Biological Chemistry*, *275*(50), 38949–38952. doi:10.1074/jbc.R000029200

Moris, N., Anlas, K., Van Den Brink, S. C., Alemany, A., Schröder, J., Ghimire, S., … Martinez Arias, A. (2020). An in vitro model of early anteroposterior organization during human development. *Nature*, *582*(7812), 410–415. doi:10.1038/s41586-020-2383-9

Najafabadi, H. S., Mnaimneh, S., Schmitges, F. W., Garton, M., Lam, K. N., Yang, A., … Hughes, T. R. (2015). C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nature Biotechnology*, *33*(5), 555–562. doi:10.1038/nbt.3128

Nakatake, Y., Ko, S. B. H., Sharov, A. A., Wakabayashi, S., Murakami, M., Sakota, M., … Ko, M. S. H. (2020). Generation and Profiling of 2,135 Human ESC Lines for the Systematic Analyses of Cell States Perturbed by Inducing Single Transcription Factors. *Cell Reports*, *31*(7), 107655. doi:10.1016/j.celrep.2020.107655

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., … Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, *376*(6588), 44–53. doi:10.1126/science.abj6987

Ohnuki, M., Tanabe, K., Sutou, K., Teramoto, I., Sawamura, Y., Narita, M., … Takahashi, K. (2014). Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. *Proceedings of the National Academy of Sciences*, *111*(34), 12426–12431. doi:10.1073/pnas.1413299111

Oki, S., Ohta, T., Shioi, G., Hatanaka, H., Ogasawara, O., Okuda, Y., … Meno, C. (2018). Ch IP -Atlas: a data-mining suite powered by full integration of public Ch IP -seq data. *EMBO reports*, *19*(12), e46255. doi:10.15252/embr.201846255

Paley, W., Eddy, M., & Knight, D. M. (2006). *Natural theology: or, evidence of the existence and attributes of the deity, collected from the appearances of nature*. Oxford world's classics. Oxford ; New York: Oxford University Press. (Original work published 1802)

Pastor, W. A., Liu, W., Chen, D., Ho, J., Kim, R., Hunt, T. J., … Clark, A. T. (2018). TFAP2C regulates transcription in human naive pluripotency by opening enhancers. *Nature Cell Biology*, *20*(5), 553–564. doi:10.1038/s41556-018-0089-0

Pavlicev, M., Hiratsuka, K., Swaggart, K. A., Dunn, C., & Muglia, L. (2015). Detecting Endogenous Retrovirus-Driven Tissue-Specific Gene Transcription. *Genome Biology and Evolution*, *7*(4), 1082–1097. doi:10.1093/gbe/evv049

Pehrsson, E. C., Choudhary, M. N. K., Sundaram, V., & Wang, T. (2019). The epigenomic landscape of transposable elements across normal human development and anatomy. *Nature Communications*, *10*(1), 5640. doi:10.1038/s41467-019-13555-x

Pestre, D. (2013). *À Contre-Science: Politiques Et Savoirs des Sociétés Contemporaines*. Paris: Éditions du Seuil.

Philippe, C., Vargas-Landin, D. B., Doucet, A. J., Van Essen, D., Vera-Otarola, J., Kuciak, M., … Cristofari, G. (2016). Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. *eLife*, *5*, e13926. doi:10.7554/eLife.13926

Playfoot, C. J., Duc, J., Sheppard, S., Dind, S., Coudray, A., Planet, E., & Trono, D. (2021). Transposable elements and their KZFP controllers are drivers of transcriptional innovation in the developing human brain. *Genome Research*, *31*(9), 1531–1545. doi:10.1101/gr.275133.120

Pontis, J., Planet, E., Offner, S., Turelli, P., Duc, J., Coudray, A., … Trono, D. (2019a). Hominid-specific transposable elements and KRAB-ZFPs facilitate human embryonic genome activation and transcription in naïve hESCs. Accession number: GSE117395. Gene Expression Omnibus (GEO).

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

Pontis, J., Pulver, C., Playfoot, C. J., Planet, E., Grun, D., Offner, S., … Trono, D. (2022). Primate-specific transposable elements shape transcriptional networks during human development. *Nature Communications*, *13*(1), 7178. doi:10.1038/s41467-022-34800-w

Popper, K. R. (1935). *The Logic of Scientific Discovery*. Routledge.

Pulver, C. (2023). craTEs. doi:10.5281/zenodo.8407480

Pulver, C., Grun, D., Duc, J., Sheppard, S., Planet, E., Coudray, A., … Trono, D. (2023). Statistical learning quantifies transposable element-mediated cis-regulation. *Genome Biology*, *24*(1), 258. doi:10.1186/s13059-023-03085-7

Pulver, C., Grün, D., Duc, J., Sheppard, S., Planet, E., De Fondeville, R., … Trono, D. (2022). *Statistical learning quantifies transposable element-mediated cis-regulation*. Genomics. doi:10.1101/2022.09.23.509180

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

Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A., & Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Research*, *42*(W1), W187–W191. doi:10.1093/nar/gku365

Ramsay, R. G. & Gonda, T. J. (2008). MYB function in normal and cancer cells. *Nature Reviews Cancer*, *8*(7), 523–534. doi:10.1038/nrc2439

Raney, B. J., Dreszer, T. R., Barber, G. P., Clawson, H., Fujita, P. A., Wang, T., … Kent, W. J. (2014). Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*, *30*(7), 1003–1005. doi:10.1093/bioinformatics/btt637

Robinson, M. D. [M. D.], McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. doi:10.1093/bioinformatics/btp616

Robinson, M. D. [Mark D] & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), R25. doi:10.1186/gb-2010-11-3-r25

Rosspopoff, O. & Trono, D. (2023). Take a walk on the KRAB side. *Trends in Genetics*, *39*(11), 844–857. doi:10.1016/j.tig.2023.08.003

Rowe, H. M., Jakobsson, J., Mesnard, D., Rougemont, J., Reynard, S., Aktas, T., … Trono, D. (2010). KAP1 controls endogenous retroviruses in embryonic stem cells. *Nature*, *463*(7278), 237–240. doi:10.1038/nature08674

Saitou, M. & Hayashi, K. (2021). Mammalian in vitro gametogenesis. *Science*, *374*(6563), eaaz6830. doi:10.1126/science.aaz6830

Santoni, F. A., Guerra, J., & Luban, J. (2012). HERV-h RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. *Retrovirology*, *9*(1), 111. doi:10.1186/1742-4690-9-111

Sasaki, T., Nishihara, H., Hirakawa, M., Fujimura, K., Tanaka, M., Kokubo, N., … Okada, N. (2008). Possible involvement of SINEs in mammalian-specific brain formation. *Proceedings of the National Academy of Sciences*, *105*(11), 4220–4225. doi:10.1073/pnas.0709398105

Satpathy, A. T., Granja, J. M., Yost, K. E., Qi, Y., Meschi, F., McDermott, G. P., … Chang, H. Y. (2019). Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral t cell exhaustion. *Nature Biotechnology*, *37*(8), 925–936. doi:10.1038/s41587-019-0206-z

Schimel, J. (2012). *Writing science: how to write papers that get cited and proposals that get funded*. Oxford: Oxford University Press.

Séguin, C. A., Draper, J. S., Nagy, A., & Rossant, J. (2008). Establishment of Endoderm Progenitors by SOX Transcription Factor Expression in Human Embryonic Stem Cells. *Cell Stem Cell*, *3*(2), 182–195. doi:10.1016/j.stem.2008.06.018

Senft, A. D. & Macfarlan, T. S. (2021). Transposable elements shape the evolution of mammalian development. *Nature Reviews Genetics*, *22*(11), 691–711. doi:10.1038/s41576-021-00385-1

Senti, K.-A., Handler, D., Rafanel, B., Kosiol, C., Schlötterer, C., & Brennecke, J. (2023). *Functional Adaptations of Endogenous Retroviruses to the Drosophila Host Underlie their Evolutionary Diversification*. Evolutionary Biology. doi:10.1101/2023.08.03.551782

Sexton, C. E. & Han, M. V. (2019). Paired-end mappability of transposable elements in the human genome. *Mobile DNA*, *10*(1), 29. doi:10.1186/s13100-019-0172-5

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

Shapin, S. & Schaffer, S. (1985). *Leviathan and the Air-Pump: Hobbes, Boyle, and the Experimental Life*. Princeton University Press.

Simó-Riudalbas, L., Offner, S., Planet, E., Duc, J., Abrami, L., Dind, S., … Trono, D. (2022). Transposon-activated POU5F1B promotes colorectal cancer growth and metastasis. *Nature Communications*, *13*(1), 4913. doi:10.1038/s41467-022-32649-7

Smith, Z. D., Chan, M. M., Humm, K. C., Karnik, R., Mekhoubad, S., Regev, A., … Meissner, A. (2014). DNA methylation dynamics of the human preimplantation embryo. *Nature*, *511*(7511), 611–615. doi:10.1038/nature13581

Solyom, S. & Kazazian, H. H. (2012). Mobile elements in the human genome: implications for disease. *Genome Medicine*, *4*(2), 12. doi:10.1186/gm311

Storer, J., Hubley, R., Rosen, J., Wheeler, T. J., & Smit, A. F. (2021). The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA*, *12*(1), 2. doi:10.1186/s13100-020-00230-y

Strumpf, D., Mao, C.-A., Yamanaka, Y., Ralston, A., Chawengsaksophak, K., Beck, F., & Rossant, J. (2005). Cdx2 is required for correct cell fate specification and differentiation of trophectoderm in the mouse blastocyst. *Development*, *132*(9), 2093–2102. doi:10.1242/dev.01801

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., … Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell*, *177*(7), 1888–1902.e21. doi:10.1016/j.cell.2019.05.031

Sultana, T., Van Essen, D., Siol, O., Bailly-Bechet, M., Philippe, C., Zine El Aabidine, A., … Cristofari, G. (2019). The Landscape of L1 Retrotransposons in the Human Genome Is Shaped by Pre-insertion Sequence Biases and Post-insertion Selection. *Molecular Cell*, *74*(3), 555–570.e7. doi:10.1016/j.molcel.2019.02.036

Sun, M.-a., Wolf, G., Wang, Y., Senft, A. D., Ralls, S., Jin, J., … Macfarlan, T. S. (2021). Endogenous Retroviruses Drive Lineage-Specific Regulatory Evolution across Primate and Rodent Placentae. *Molecular Biology and Evolution*, *38*(11), 4992–5004. doi:10.1093/molbev/msab223

Sun, X., Wang, X., Tang, Z., Grivainis, M., Kahler, D., Yun, C., … Boeke, J. D. (2018). Transcription factor profiling reveals molecular choreography and key regulators of human

retrotransposon expression. *Proceedings of the National Academy of Sciences*, *115*(24). doi:10.1073/pnas.1722565115

Sundaram, V., Cheng, Y., Ma, Z., Li, D., Xing, X., Edge, P., … Wang, T. (2014). Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Research*, *24*(12), 1963–1976. doi:10.1101/gr.168872.113

Sundaram, V., Choudhary, M. N. K., Pehrsson, E., Xing, X., Fiore, C., Pandey, M., … Wang, T. (2017). Functional cis-regulatory modules encoded by mouse-specific endogenous retrovirus. *Nature Communications*, *8*(1), 14550. doi:10.1038/ncomms14550

Sundaram, V. & Wang, T. (2018). Transposable Element Mediated Innovation in Gene Regulatory Landscapes of Cells: Re-Visiting the "Gene-Battery" Model. *BioEssays*, *40*(1), 1700155. doi:10.1002/bies.201700155

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

Tang, W. W., Dietmann, S., Irie, N., Leitch, H. G., Floros, V. I., Bradshaw, C. R., … Surani, M. A. (2015). A Unique Gene Regulatory Network Resets the Human Germline Epigenome for Development. *Cell*, *161*(6), 1453–1467. doi:10.1016/j.cell.2015.04.053

The ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. doi:10.1038/nature11247

The FANTOM Consortium, Fort, A., Hashimoto, K., Yamada, D., Salimullah, M., Keya, C. A., … Carninci, P. (2014). Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nature Genetics*, *46*(6), 558–566. doi:10.1038/ng.2965

Theunissen, T. W., Friedli, M., He, Y., Planet, E., O'Neil, R. C., Markoulaki, S., … Jaenisch, R. (2016). Molecular Criteria for Defining the Naive Human Pluripotent State. *Cell Stem Cell*, *19*(4), 502–515. doi:10.1016/j.stem.2016.06.011

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(1), 91–108. doi:10.1111/j.1467-9868.2005.00490.x

Todd, C. D. [Christopher D], Deniz, Ö., Taylor, D., & Branco, M. R. (2019). Functional evaluation of transposable elements as enhancers in mouse embryonic and trophoblast stem cells. *eLife*, *8*, e44344. doi:10.7554/eLife.44344

Tosic, J., Kim, G.-J., Pavlovic, M., Schröder, C. M., Mersiowsky, S.-L., Barg, M., … Arnold, S. J. (2019). Eomes and Brachyury control pluripotency exit and germ-layer segregation by changing the chromatin state. *Nature Cell Biology*, *21*(12), 1518–1531. doi:10.1038/s41556-019-0423-1

Trizzino, M., Kapusta, A., & Brown, C. D. (2018). Transposable elements generate regulatory novelty in a tissue-specific fashion. *BMC Genomics*, *19*(1), 468. doi:10.1186/s12864-018-4850-3

Tsankov, A. M., Gu, H., Akopian, V., Ziller, M. J., Donaghey, J., Amit, I., … Meissner, A. (2015). Transcription factor binding dynamics during human ES cell differentiation. *Nature*, *518*(7539), 344–349. doi:10.1038/nature14233

Turabian, K. L. (2018). *A manual for writers of research papers, theses, and dissertations: Chicago Style for students and researchers* (9th edition) (W. C. Booth, G. G. Colomb, J. M. Williams, J. Bizup, & W. T. FitzGerald, Eds.). Chicago guides to writing, editing, and publishing. Chicago London: The University of Chicago Press.

Turelli, P., Playfoot, C., Grun, D., Raclot, C., Pontis, J., Coudray, A., … Trono, D. (2020). Primate-restricted KRAB zinc finger proteins and target retrotransposons control gene expression in human neurons. *Science Advances*, *6*(35), eaba3200. doi:10.1126/sciadv.aba3200

Tyser, R. C. V., Mahammadov, E., Nakanoh, S., Vallier, L., Scialdone, A., & Srinivas, S. (2021). Single-cell transcriptomic characterization of a gastrulating human embryo. *Nature*, *600*(7888), 285–289. doi:10.1038/s41586-021-04158-y

Vahava, O., Morell, R., Lynch, E. D., Weiss, S., Kagan, M. E., Ahituv, N., … Avraham, K. B. (1998). Mutation in Transcription Factor *POU4F3* Associated with Inherited Progressive Hearing Loss in Humans. *Science*, *279*(5358), 1950–1954. doi:10.1126/science.279.5358.1950

Van Bree, E. J., Guimarães, R. L., Lundberg, M., Blujdea, E. R., Rosenkrantz, J. L., White, F. T., … Jacobs, F. M. (2022). A hidden layer of structural variation in transposable elements

reveals potential genetic modifiers in human disease-risk loci. *Genome Research*, *32*(4), 656–670. doi:10.1101/gr.275515.121

Van Den Brink, S. C. & Van Oudenaarden, A. (2021). 3D gastruloids: a novel frontier in stem cell-based in vitro modeling of mammalian gastrulation. *Trends in Cell Biology*, *31*(9), 747–759. doi:10.1016/j.tcb.2021.06.007

Van Drongelen, R., Vazquez-Faci, T., Huijben, T. A., Van Der Zee, M., & Idema, T. (2018). Mechanics of epithelial tissue formation. *Journal of Theoretical Biology*, *454*, 182–189. doi:10.1016/j.jtbi.2018.06.002

Vasconcelos, F. F., Sessa, A., Laranjeira, C., Raposo, A. A., Teixeira, V., Hagey, D. W., … Castro, D. S. (2016). MyT1 Counteracts the Neural Progenitor Program to Promote Vertebrate Neurogenesis. *Cell Reports*, *17*(2), 469–483. doi:10.1016/j.celrep.2016.09.024

Verzi, M. P., Shin, H., He, H. H., Sulahian, R., Meyer, C. A., Montgomery, R. K., … Shivdasani, R. A. (2010). Differentiation-specific histone modifications reveal dynamic chromatin interactions and partners for the intestinal transcription factor CDX2. *Developmental Cell*, *19*(5), 713–726. doi:10.1016/j.devcel.2010.10.006

Wang, J., Xie, G., Singh, M., Ghanbarian, A. T., Raskó, T., Szvetnik, A., … Izsvák, Z. (2014). Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. *Nature*, *516*(7531), 405–409. doi:10.1038/nature13804

Wang, X., Veerapandian, V., Yang, X., Song, K., Xu, X., Cui, M., … Zhao, X.-Y. (2021a). The chromatin accessibility landscape reveals distinct transcriptional regulation in the induction of human primordial germ cell-like cells from pluripotent stem cells. Accession number: GSE143345. Gene Expression Omnibus (GEO).

Wang, X., Veerapandian, V., Yang, X., Song, K., Xu, X., Cui, M., … Zhao, X.-Y. (2021b). The chromatin accessibility landscape reveals distinct transcriptional regulation in the induction of human primordial germ cell-like cells from pluripotent stem cells. *Stem Cell Reports*, *16*(5), 1245–1261. doi:10.1016/j.stemcr.2021.03.032

Wells, J. N., Chang, N.-C., McCormick, J., Coleman, C., Ramos, N., Jin, B., & Feschotte, C. (2023). Transposable elements drive the evolution of metazoan zinc finger genes. *Genome Research*, *33*(8), 1325–1339. doi:10.1101/gr.277966.123

Wilks, C., Zheng, S. C., Chen, F. Y., Charles, R., Solomon, B., Ling, J. P., … Langmead, B. (2021). Recount3: summaries and queries for large-scale RNA-seq expression and splicing. *Genome Biology*, *22*(1), 323. doi:10.1186/s13059-021-02533-6

Williams, G. C. & Dawkins, R. (2019). *Adaptation and natural selection: a critique of some current evolutionary thought* (New Princeton Science Library edition). Princeton science library. Princeton: Princeton University Press. (Original work published 1966)

Wolf, G., Yang, P., Füchtbauer, A. C., Füchtbauer, E.-M., Silva, A. M., Park, C., … Macfarlan, T. S. (2015). The KRAB zinc finger protein ZFP809 is required to initiate epigenetic silencing of endogenous retroviruses. *Genes & Development*, *29*(5), 538–554. doi:10.1101/gad.252767.114

Wray, G. A. (2007). The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics*, *8*(3), 206–216. doi:10.1038/nrg2063

Wu, J., Xu, J., Liu, B., Yao, G., Wang, P., Lin, Z., … Sun, Y. (2018). Chromatin analysis in human early development reveals epigenetic transition during ZGA. *Nature*, *557*(7704), 256–260. doi:10.1038/s41586-018-0080-8

Xu, J., Shao, Z., Glass, K., Bauer, D. E., Pinello, L., Van Handel, B., … Orkin, S. H. (2012). Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis. *Developmental Cell*, *23*(4), 796–811. doi:10.1016/j.devcel.2012.09.003

Yagi, R., Kohn, M. J., Karavanova, I., Kaneko, K. J., Vullhorst, D., DePamphilis, M. L., & Buonanno, A. (2007). Transcription factor TEAD4 specifies the trophectoderm lineage at the beginning of mammalian development. *Development*, *134*(21), 3827–3836. doi:10.1242/dev.010223

Yamauchi, K., Ikeda, T., Hosokawa, M., Nakatsuji, N., Kawase, E., Chuma, S., … Suemori, H. (2020). Overexpression of Nuclear Receptor 5A1 Induces and Maintains an Intermediate State of Conversion between Primed and Naive Pluripotency. *Stem Cell Reports*, *14*(3), 506–519. doi:10.1016/j.stemcr.2020.01.012

Yang, P., Wang, Y., & Macfarlan, T. S. (2017). The Role of KRAB-ZFPs in Transposable Element Repression and Mammalian Evolution. *Trends in Genetics*, *33*(11), 871–881. doi:10.1016/j.tig.2017.08.006

Zhang, Y. [Yanxiao], Li, T., Preissl, S., Amaral, M. L., Grinstein, J. D., Farah, E. N., … Ren, B. (2019). Transcriptionally active HERV-h retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nature Genetics*, *51*(9), 1380–1388. doi:10.1038/s41588-019-0479-7

Zhang, Y. [Yong], Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., … Liu, X. S. (2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, *9*(9), R137. doi:10.1186/gb-2008-9-9-r137

Zou, M., Li, S., Klein, W. H., & Xiang, M. (2012). Brn3a/Pou4f1 regulates dorsal root ganglion sensory neuron specification and axonal projection into the spinal cord. *Developmental Biology*, *364*(2), 114–127. doi:10.1016/j.ydbio.2012.01.021

Zou, Z., Ohta, T., Miura, F., & Oki, S. (2022). ChIP-Atlas 2021 update: a data-mining suite for exploring epigenomic landscapes by fully integrating ChIP-seq, ATAC-seq and Bisulfite-seq data. *Nucleic Acids Research*, *50*(W1), W175–W182. doi:10.1093/nar/gkac199

# Cyril PULVER

Engineer in Life Sciences and Technology, EPFL
Adress: Place de la Gare 12, Renens, Switzerland
Phone: +41795811793
E-mail: cyril.pulver@epfl.ch
Linkedin: https://www.linkedin.com/in/cyril-pulver


## Education

- **École Polytechnique Fédérale de Lausanne** — Lausanne, VD
  *PhD student* — *January 2020 - present*
  - Doctoral School: Quantitative and Computational Biology (EDCB)
  - Laboratory of Virology and Genetics (head: Didier Trono)
- **École Polytechnique Fédérale de Lausanne** — Lausanne, VD
  *Master of Science in Life Sciences and Technology* — *September 2016 - July 2019*
  - Orientation: Molecular Medicine and Systems Biology (120 ECTS credits)
  - Graduated with an average of 5.73/6
- **École Polytechnique Fédérale de Lausanne** — Lausanne, VD
  *Bachelor of Science in Life Sciences and Technology* — *September 2013-June 2016*
  - Graduated with a bachelor cycle average of 5.07/6

## Publications

- Martins F, Rosspopoff O, Carlevaro-Fita J, Forey R, Offner S, Planet E, **Pulver C**, Pak H, Huber F, Michaux J, Bassani-Sternberg M, Turelli P, Trono D. (2023). KRAB zinc finger proteins ZNF587/ZNF417 protect lymphoma cells from replicative stress-induced inflammation. biorXiv
- **Pulver C**, Grun D, Duc J, Sheppard S, Planet E, de Fondeville R, Pontis J, Trono D. (2022). Statistical learning quantifies transposable element-mediated cis-regulation. biorXiv
- Pontis J, **Pulver C**, Playfoot CJ, Planet E, Grun D, Offner S, Duc J, Manfrin A, Lutolf MP, Trono D. (2022). Primate-specific transposable elements shape transcriptional networks during human development. *Nature Communications, Volume 13*, 7178

## Experience

- **Internship in Early Biomarker Development Oncology** — Hoffmann-La Roche, Basel
  *Data Scientist* — *November 2021 - October 2022*

  - Developed, implemented and evaluated statistical learning models for transcriptome-based biomarker discovery in the context of cancer immunotherapy

- **Master thesis in Michele De Palma's research unit** École Polytechnique Fédérale de Lausanne, VD
  *Bioinformatician* *September 2018 - July 2019*
  - Studied the role of macrophages in anti-EGFR therapy (follows-up on a lab immersion completed in the same group)

- **Industry internship at Geneva Biotech** Geneva Biotech, GE
  *Wet lab intern* *February 2018 - July 2018*
  - Devised a method for the delivery of large DNA cargo viral vectors for next generation T cell engineering

- **Lab immersion in Didier Trono's research unit** École Polytechnique Fédérale de Lausanne, VD
  *Bioinformatician* *July 2017 - December 2017*
  - Studied the evolution of KRAB zinc-finger genes in Neanderthals, Denisovans and Humans.

- **Lab immersion in Cathrin Brisken's research unit** École Polytechnique Fédérale de Lausanne, VD
  *Wet lab intern* *February 2017 - June 2017*
  - Assessed the impact of progesterone receptor overexpression on the proliferation of an estrogen positive breast cancer cell line.

- **International Genetically Engineered Machine (iGEM) competition 2015** Lausanne, VD, Boston, USA
  *EPFL team member* *February 2015 - October 2015*
  - EPFL's project: Logic Orthogonal gRNA Implemented Circuits (BioLOGIC) aimed at designing and characterizing a dCas9-mediated transistor which can be assembled into logic gates in live bacteria and yeasts.

- **Teaching** École Polytechnique Fédérale de Lausanne, VD, Müntschemier, BE, Hergiswil, LU
  *Teaching assistant* *September 2015 - present*
  - Assisting students in solving exercises weekly for physics and biology bachelor courses.
  - Gave introductory lectures in evolution, endocrinology and immunology to high school students preparing for the Biology Olympiads

## Skills

- Bioinformatics and computational approaches
  - NGS data processing and analysis (RNA-seq, ChIP-seq, CutTag, ATAC-seq, scRNA-seq, scATAC-seq)
  - Genome annotation
  - Data analysis
  - Statistical learning
- Programming Languages, softwares and platforms
  - Python, R, Unix Shell, High Performance Computing (SLURM), Apache Spark, Apache Hadoop, git, Docker, renku
- Languages
  - French (native speaker), English (excellent command), German (good command), Spanish (good command)

## Activities and Interests

- **Music project (Kūn)** Switzerland
  *Composer, DJ* *August 2014 - present*
  - Dancefloor oriented music project leading me to DJ at a professional level in various festivals and clubs in Switzerland and abroad such as Caprices (Crans-Montana), Electron (Geneva), Polaris (Verbier), D! Club (Lausanne, former residency), Halle W (Geneva), Shapes (Sandefjord), Jaeger (Oslo) , Audio (Geneva) and Folklor (Lausanne, current residency).