THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

L1 is dynamic during neurogenesis

Carmen Salvador-Palomeque

M.S. Biochemistry, Molecular Biology and Biomedicine

B.S. Biology. Speciality: Genetics

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2018

Faculty of Medicine

ABSTRACT

L1 retrotransposition is a significant source of endogenous mutagenesis in the human genome. This process is driven by a handful of highly active source L1s in each individual. L1 mobilization can occur in natural pluripotent cells, such as the stem cells present in the early embryo, and during the artificial generation of human induced pluripotent stem cells (hiPSCs) via cellular reprogramming. hiPSCs have been proposed for use in regenerative medicine, rendering an understanding of L1 retrotransposition and mutagenesis during hiPSC induction and cultivation essential.

Although endogenous L1 insertions are known to arise in hiPSC lines [1], more information is required regarding L1 mobilisation and insertional patterns in order to elucidate the regulation and impact of L1 activity in this context. For example, transductions can be useful to find active, retrotransposition-competent L1s (RC-L1s) responsible for *de novo* retrotransposition events, and infer relationships between progenitor and offspring L1 copies. It has also been reported that endogenous retrotransposition in hiPSCs may generate an elevated fraction of full-length insertions [1, 2]. Crucially, L1 insertions bearing transductions have as yet not been identified in hiPSCs or embryonic stem cells (ESCs). Additionally, L1 promoters are methylated by the host genome to reduce their potential to initiate L1 mRNA transcription [3]. The genome-wide methylation state of L1 families, and specific L1 loci, has been analysed in ESCs and hiPSCs, and in neural stem cells (NSCs) [4-10]. However, the methylation state of individual L1 loci, including *de novo* L1 insertions and their progenitor elements, has to date not been reported.

In the research described in this thesis, I identified a full-length *de novo* L1 insertion in a cultured hiPSC line. This L1 insertion was not present in the matched parental human dermal fibroblast (HDF) line and, as a result, was annotated as a reprogramming-associated event. Notably, the L1 carried transduced genomic sequences at its 5' and 3' termini. Using these unique transduced sequences, we identified the associated source (donor) L1 element, which had previously been reported to retrotranspose *in vitro* [11]. This donor L1 was part of an extended group of closely related L1s identified via their shared 3' transductions – an L1 transduction family [12]. Interestingly, the ancestral L1 copy, or lineage progenitor L1, of this family was previously reported as being inactive *in vitro* [13]. However, we found that the *de novo* L1 insertion, its immediate donor L1, and some other members of the transduction family, retrotransposed efficiently *in vitro*, indicating that these L1s comprise a lineage still capable of mobilisation, including during reprogramming.

I next analyzed DNA methylation amongst the L1 transduction family prior to fibroblast reprogramming, in hiPSCs, and during neuronal differentiation. Notably, members of the family were demethylated during reprogramming and were then progressively methylated during neuronal differentiation. The *de novo* L1 insertion was hypomethylated compared to its donor L1, and the donor L1 was in turn less methylated than the older family members.

This thesis therefore identifies an extended L1 transduction family that is mobile during reprogramming, likely as a result of specific relaxation of DNA methylation via an unknown mechanism. Using an *in vitro* system, my experimental data allows us to predict dynamic L1 methylation patterns during neurodifferentiation *in vivo*. Finally, my work further highlights L1 mutagenesis as a potential obstacle to the use of hiPSCs in biomedical applications.

DECLARATION BY AUTHOR

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, financial support and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my higher degree by research candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis and have sought permission from co-authors for any jointly authored works included in the thesis.

PUBLICATIONS DURING CANDIDATURE

Conference Abstracts:

-2018 Translational Research Institute (TRI) poster symposium. Brisbane, Australia. “Dynamic L1 methylation during reprogramming and neurodifferentiation”. Salvador-Palomeque C, Richardson SR, Faulkner GJ.

-2018 Keystone Symposia: Mobile Elements and Genome Plasticity, New Mexico, United States of America. “An L1 family lineage active during human hiPSC reprogramming”. Salvador-Palomeque C, Richardson SR, Sanchez-Luque FJ, Faulkner GJ.

-2017 Cold Spring Harbour Laboratory (CSHL) Meeting: The Mobile Genome, Cold Spring Harbour, New York, United States of America. “The role of L1 activity in Rett Syndrome”. Salvador-Palomeque C, Morell S, Richardson SR, Faulkner GJ.

-2016 International Congress of Transposable Elements (ICTE), Saint Malo, France. “L1 activity in Rett Syndrome”. Salvador-Palomeque C, Morell S, Richardson SR, Faulkner GJ.

-2015. EMBO Symposium. The Mobile Genome. Heidelberg. Germany. “Tex19.1 regulates LINE-1 ubiquitylation, protein stability and retrotransposition in the mammalian germline”. Garcia-Canadas M, Maclennan M, Salvador-Palomeque C, Adams I, Garcia-Perez JL.

-2015 6th International Postgraduate Symposium in Biomedical Sciences at The University of Queensland, Brisbane, Australia. “LINE-1 is dynamic during human neurogenesis”. Salvador-Palomeque C, Richardson SR, Faulkner GJ.

-2015 Translational Research Institute (TRI) poster symposium. Brisbane, Australia. “LINE-1 is dynamic during neurogenesis”. Salvador-Palomeque C, Richardson SR, Faulkner GJ.

Publications:

Upton, K. R., D. J. Gerhardt, J. S. Jesuadian, S. R. Richardson, F. J. Sanchez-Luque, G. O. Bodea, A. D. Ewing, **C. Salvador-Palomeque**, M. S. van der Knaap, P. M. Brennan, A. Vanderver and G. J. Faulkner (2015). "Ubiquitous L1 mosaicism in hippocampal neurons." Cell.

MacLennan, M., M. Garcia-Canadas, J. Reichmann, E. Khazina, G. Wagner, C. J. Playfoot, **C. Salvador-Palomeque**, A. R. Mann, P. Peressini, L. Sanchez, K. Dobie, D. Read, C. C. Hung, R. Eskeland, R. R. Meehan, O. Weichenrieder, J. L. Garcia-Perez and I. R. Adams

(2017). "Mobilization of LINE-1 retrotransposons is restricted by Tex19.1 in mouse embryonic stem cells." Elife.

Sanchez-Luque, F.J., Kempen, M-J.H.C., Gerdes, P., Vargas-Landin, D.B., Richardson, S.R., Troskie, R-L., Jesuadian, J.S., Cheetham, S.W., Carreira, P.E., **Salvador-Palomeque, C.**, García-Cañadas, M., Muñoz-Lopez, M., Sanchez, L., Lundberg, M., Macia, A., Heras, S.R., Brennan, P.M., Lister, R., Garcia-Perez, J.L., Ewing, A.D., Faulkner, G.J. (2019). "LINE-1 evasion of epigenetic repression in humans." Molecular Cell.

Cano D, Macía A, Sanchez-Carnerero V, Garcia-Cañadas M, Heras SR, **Salvador-Palomeque C**, Moran JV, García-Perez JL. "The role of p53 on LINE-1 expression and engineered retrotransposition". (Manuscript in preparation).

Publications included in this thesis

Salvador-Palomeque, C., Sanchez-Luque, F.J., Fortuna, P.R.J., Ewing, A.D., Wolvetang, E.J., Richardson, S.R., Faulkner, G.J. (2019) "Dynamic methylation of an L1 transduction family during reprogramming and neurodifferentiation." Molecular and Cellular Biology.

Contributor	Statement of contribution
Carmen Salvador-Palomeque	Conception and design (60%) Analysis and interpretation (55%) Drafting and production (65%)
Geoffrey J. Faulkner	Conception and design (20%) Analysis and interpretation (20%) Drafting and production (20%)
Sandra Richardson	Conception and design (10%) Analysis and interpretation (15%) Drafting and production (10%)
Francisco J. Sanchez-Luque	Conception and Design (10%) Analysis and interpretation (10%) Drafting and production (5%)

Contributions by others to the thesis

Significant contributions: Carmen Salvador-Palomeque, Geoffrey J. Faulkner and Sandra Richardson contributed to the conception, design of the project, analysis and interpretation of research data and drafting and presentation of results.

Francisco J. Sanchez-Luque contributed to the design of one relevant part of this PhD which will be published. Patrick R.J. Fortuna performed the tissue culture experiments of neurodifferentiation and associated immunocytochemistry-based cellular characterisation.

Statement of parts of the thesis submitted to qualify for the award of another degree

“None”.

Research Involving Human or Animal Subjects

Induced pluripotent stem cell lines were derived under UQ ethics approval #201200557. Copies of the ethics approval letters are included in the thesis as an appendix.

ACKNOWLEDGEMENTS

I thank members of the Faulkner laboratory for helpful advice and discussion, especially Prof. Geoffrey Faulkner, Dr. Sandra Richardson, Dr. Francisco Jose Sanchez-Luque, Dr. Adam Ewing, Dr. Patricia Carreira, Dr. Seth Cheetham and Dr. Santiago Morell-Hita. I thank Dr. Jose Luis Garcia-Perez and his laboratory for helpful advice and discussion. I thank Prof. Ernst Wolvetang and Mr. Patrick Fortuna for their collaboration with us. I thank the TRI flow cytometry core facility.

FINANCIAL SUPPORT

My research was supported by a University of Queensland International (UQI) postgraduate research scholarship.

KEYWORDS

L1or LINE-1: long interspersed element 1

hiPSCs: human induced pluripotent stem cells

ESCs: embryonic stem cells

AUSTRALIAN AND NEW ZEALAND STANDARD RESEARCH CLASSIFICATIONS (ANZSRC)

ANZSRC code: 060408 Genomics 60%

ANZSRC code: 060410 Neurogenetics 20%

ANZSRC code: 060404 Epigenetics (incl. Genome Methylation and Epigenomics) 20%

FIELDS OF RESEARCH (FOR) CLASSIFICATION

FoR code: 0604 Genetics, 60%

FoR code: 0601 Biochemistry and Cell Biology, 40%

DEDICATIONS

One day I made a promise…To my father, Antonio R. Salvador-Bedmar

Besides being thousands of kilometres away from me during the entire realisation of this PhD, I would like to dedicate this thesis to my family, especially to my parents, Mari Carmen and Antonio. To their omnipresence. To their human values, to fulfil me with curiosity and teaching me how to look at the word with a different perspective, because, with all of their examples, I decided to undertake this endeavour without surrendering, and I could taste and enjoy every minute of this dreamed journey. To Manolo and his unforgettable greatness, to the freedom he always gave me without clipping my wings, always thanks. To my siblings, Antonio and Julia and to my two little nieces Martina and Lia, it is when I look at them when I realise all the time I have been investing on this PhD. To my cousins and friends from Spain. Especially to Yoyo, Angela, Rocio and Eli. To all my "Aussie" friends, my family here, especially to: Robin-Lee and Adriana the nice breeze I needed inside and outside of the lab. To Gabi, Santi, Caroline, Rodrigo, Carol, Tom and Pat. My fireflies giving me energy and light to keep on sparking. Special thanks to Mick, for all his magnificent support in all of the senses.

To my mentor, Dr Garcia-Perez who introduced me to the amazing world of retrotransposons. Thanks for helping me to jump to the other side of the world and introducing me to Prof Faulkner, without the two of you, this PhD could have not been performed. To Prof Faulkner for giving me the opportunity of experiencing this life-changing adventure, surely, one of the most remarkable ones of my life. To Dr Richardson, for your patience and way of thinking. To all the LINE-1 lab members from Faulkner lab and Garcia-Perez lab, especially to Dr Carreira and Dr Garcia-Canadas. To Prof Wolvetang and his lab for their help. To Dr Rodriguez Niedenfuhr, my first supervisor. To all the difficulties I experienced along the way because they only showed me how to become stronger.

To all those curious minds that one day wondered and contributed to increase knowledge creating science and helped directly or indirectly to the realisation of this thesis. Special thanks to Barbara McClintock, for her persistency that I took as an example to follow. To all of those ones who are still passionate about doing so, fighting against the odds. To the necessity of knowing the unknown because, as the most special woman I have ever met says: '*Knowledge is the only thing no one can take away from you'* -M.C. Palomeque-Lizana.

TABLE OF CONTENTS

ABBREVIATIONS

ASP: Antisense Promoter

AT: Ataxia Telangiectasia

BAC: Bacterial Artificial Chromosome

Bp: Base pair

ChIP: Chromatin Immunoprecipitation

CNS: Central Nervous System

CNV: Copy Number Variation

CRC: Colorectal Cancer

CypA: Cyclophilin A

DG: Dentate Gyrus

DNA: Deoxyribonucleic Acid

DNMTs: DNA Methyltransferases

DNMT3L: DNA methyltransferase 3-like protein

DSB: Double-Strand DNA Breaks

EGFP: Enhanced Green Fluorescent Protein

EN: Endonuclease

ESC: Embryonic Stem Cell

FACS: Fluorescence Activated Cell Sorting

Fwd: Forward

GO: Gene Ontology

GPC: Glial Precursor Cell

HDAC: Histone Deacetylase

HDFs: Human Dermal Fibroblasts

hESCs: human Embryonic Stem Cells

hiPSC: human induced Pluripotent Stem Cell

HRG: Human Reference Genome

HIV: Immunodeficiency Virus Type I

Kb: Kilobase

Kbp: Kilo basepair

KDa: KiloDalton

KO: Knock Out

L1 or LINE-1: Long interspersed element 1

LNA: Locked Nucleic Acid

LTR: Long Terminal Repeat

MALBAC: Multiple Annealing and Looping Based Amplification Cycles

MDA: Multiple Display Amplification

ME: Mobile Elements

MELT: Mobile Element Locator Tool

Non-Ref: Non-Reference

NPC: Neural Precursor Cell

NSC: Neural Stem Cell

Nt: nucleotide

ORF: Open Reading Frame

ORF1p: ORF1 protein

ORF2p: ORF2 protein

PCR: Polymerase Chain Reaction

PGCs: Primordial Germ Cells

RC-L1: Retrotransposition-competent L1

RC-seq: Retrotransposon Capture sequencing

RNA: Ribonucleic Acid

RNP: Ribonucleoprotein

RPE: Retinal Pigment Epithelium

RT: Reverse transcriptase

RTT: Rett syndrome

SGZ: Subgranular Zone

SINE: Short interspersed element

VNTR: Variable Number of Tandem Repeats

SVA: SINE-VNTR-*Alu*

SVZ: Subventricular Zone

Ta: Transcribed, subset a

TC: Time Course

TEs: Transposable Elements

TF: Transcription Factors

TPRT: Target-Primed Reverse Transcription

Trunc-Non-Ref: Truncated Non-Reference

TS-ATLAS: Transduction-Specific Amplification Typing of L1 Active Subfamilies

USC: Unique Sequence Component

UTR: Untranslated Region

VNTR: Variable Tandem Number Repeats

WGA: Whole Genome Amplification

WGS: Whole Genome Sequencing

LIST OF FIGURES

.

LIST OF TABLES

INTRODUCTION

1 Types of mammalian mobile DNA.

Mobile DNA has played a major evolutionary role in shaping eukaryotic genomes [15]. These sequences, also known as “jumping genes”, are classified by the intermediate they use to mobilise: either DNA (transposons) or RNA (retrotransposons) (**Figure 1**) [16]. DNA transposons use a cut-and-paste mechanism involving a transposase, entirely comprise fossils of ancient elements in the human genome, and are inactive in humans [15]. Retrotransposons use a copy-and-paste mechanism and are found in the vast majority of mammalian genomes. Retrotransposons are also classified by whether they are autonomous or non-autonomous, i.e. if they encode their own proteins to mobilise independently, or if they rely on proteins from other elements to jump [17]. Autonomous retrotransposons are further divided into long terminal repeat (LTR) retrotransposons, which resemble retroviruses in both their mobility mechanism and structure [18], and non-LTR retrotransposons, which includes the prime protagonist of this thesis: Long interspersed element (LINE-1 or L1) (**Figure 1**).

L1 is the only autonomous, active human retrotransposon [15, 16]. Approximately ~45% of the human genome is composed of retrotransposons; ~17% is occupied by L1 copies [15]. Non-autonomous human retrotransposons include the short-interspersed elements (SINEs) *Alu* and SVA, which rely on the machinery of L1 to mobilise [19, 20]. One million *Alu* copies account for ~10% of the human genome [15, 16, 19, 21, 22]. By contrast, genomes of other organisms such as fruit fly, worm, and Arabidopsis may comprise less than 10% transposed elements, while some plant genomes contain >80% [15, 16, 23-25]. Thus, mobile DNA elements are dynamic molecular entities that change eukaryotic genome landscapes by their amplification and contribution to structural variation, and are therefore key components of genome evolution [16].

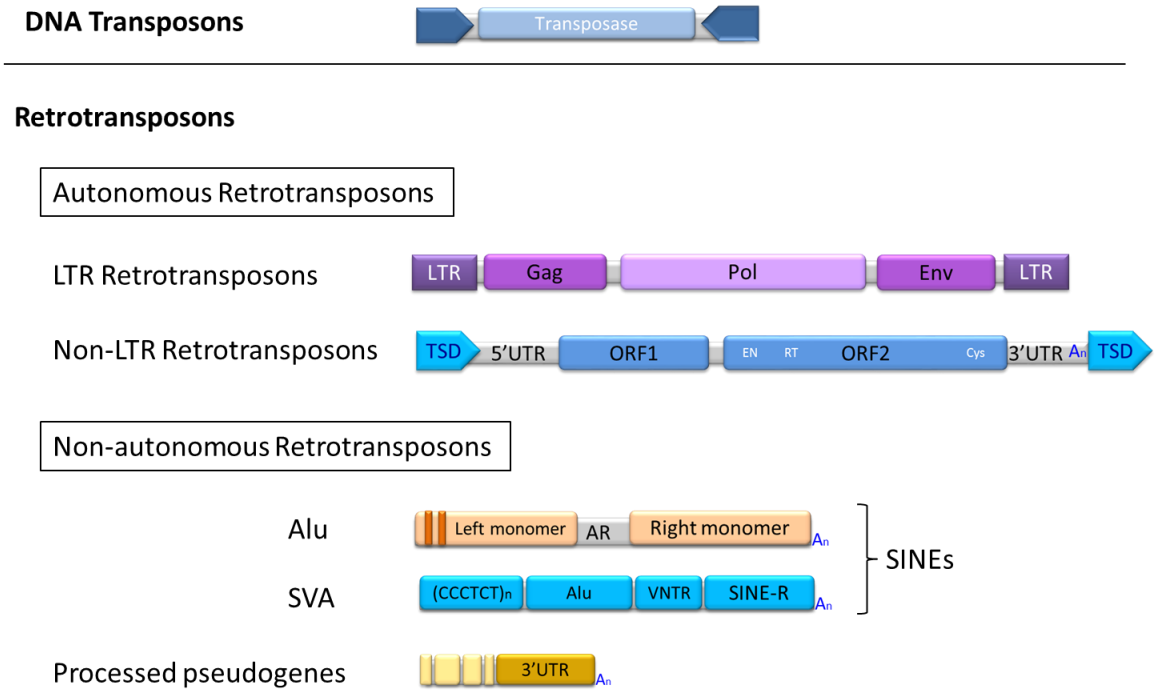


Figure 1: Types of human mobile elements: Displayed are the different types of mobile elements found in humans. This figure was adapted from [17].

2 L1 retrotransposons.

Approximately 500,000 L1 copies are present in the human reference genome (HRG) [15]. To estimate how many retrotransposition-competent L1s (RC-L1s) were present in the human population, Brouha *et al.* [13] i) interrogated the HRG to identify L1s with intact ORFs and ii) estimated the allele frequency of these L1s in the human population. Using genomic DNA and bacterial artificial chromosomes (BAC) DNA, the authors PCR amplified 82 out of 90 of the L1s with intact ORFs identified in the HRG, 67 clones from human gDNA and 15 from BAC DNA from 2 independent PCRs. As nearly 50% of the cloned L1s were retrotransposition competent in an *in vitro* retrotransposition assay [22], and considering that the HRG version used only represented 95% of a complete haploid genome, the authors estimated around 93 RC-L1s were found per diploid genome. In a second estimation, taking into account reports of polymorphic L1s in the human population not present in the HRG [26-29] and the estimated allele frequencies of 40 active reference L1s, the authors arrived at an average of 66 RC-L1s per average diploid genome [13]. Based on their estimates using these 2 different approaches, and the likely existence of undiscovered L1s, Brouha *et al.* reported that each individual human carries between 80 and 100 RC-L1s.

Additionally, Brouha *et al.* identified that 84% of assayed retrotransposition capability was the result of 6 highly active, or 'hot', L1s. The definition used to define "hot" L1s in this study was based on observing one third of the retrotransposition activity of the disease causing L1 insertion $L1_{RP}$ [30, 31]. Hence, the authors concluded that most of L1 retrotransposition in the human population stems from these "hot" L1s [13].

Subsequently, Beck *et al.* expanded upon the work of Brouha *et al.* by sequencing fosmid libraries generated using genomic DNA from six individuals from diverse geographic populations, identifying polymorphic L1 copies absent from the HRG. Differing from Brouha *et al*., Beck et al. used a PCR-free approach to clone L1 sequences. In this study, L1s were cloned directly from the produced libraries after enzymatic digestion. This approach eliminated the risk of PCR-induced substitutions being introduced into the L1 sequence, and therefore improved the accuracy of *in vitro* L1 activity estimates. A new set of 68 full-length polymorphic L1s was identified, of which half were highly active or "hot" in the retrotransposition assay. In this case, the term "hot" L1 was defined based on those elements that exhibited more than 10% of the retrotransposition activity of L1.3 [27]. It was demonstrated that the number of "hot" L1s can vary from person to person, and some were found only in specific populations [11]. Moreover, specific L1 copies can present heterogeneity in their activity due to allelic variants [32]. Therefore, it still is not fully known how many "hot" L1s are found in the human population. Most active L1s belong to the hominid specific L1-Ta family [30, 33] which was first identified in human teratocarcinoma cells and was termed "Ta" (for transcribed, subset a [34]). The remaining L1 copies (>99.9%) are defective due to 5' truncations, inversions, deletions, and other mutations [13, 15, 16].

A retrotransposition-competent L1-Ta is 6 kilo basepair (kbp) in length and contains a 5' untranslated region (UTR), two non-overlapping Open Reading Frames (ORF1 and ORF2) that encode proteins required for L1 mobility, and a 3'UTR that is punctuated by a poly(A) tail [26, 35]. The L1 5'UTR has an internal RNA polymerase II promoter that directs transcription from the 5' end of the element [36] and presents *cis*-acting binding sites for multiple transcription factors (TF) [37-42]. The 5'UTR also contains a potent antisense promoter (L1 ASP) [43, 44]. Another promoter was recently discovered in the 3'UTR region. Transcripts initiated at this 3' promoter are found in a wide variety of somatic tissues, including brain [45]. A relatively recent study demonstrated the presence of a primate specific ORF, called ORF0. This ORF0 is in the 5'UTR, anti-sense orientated with respect to the L1 copy [46]. ORF1 and ORF2 are separated by a ~60bp spacer [22]. ORF1 encodes a 40kDa protein, called ORF1p with RNA binding domain and chaperone activity [47, 48].

ORF2 encodes a 150kDa protein called ORF2p, with endonuclease (EN) [49] and reverse transcriptase (RT) activities [50]. These activities are required for canonical L1 retrotransposition [22].

3 L1 retrotransposition mechanism.

L1 retrotransposition begins when RNA polymerase II initiates transcription of a full-length L1 from its internal 5' sense promoter [36] (**Figure 2**). The L1 mRNA is transported to the cytoplasm where the translation of ORF1p and ORF2p [16] occurs via an unconventional termination-reinitiation mechanism [51, 52]. ORF1p and ORF2p exhibit *cis* preference, meaning that in the cytoplasm the two proteins bind to the L1 mRNA that encoded them [53] to form a ribonucleoprotein particle (RNP), the hypothesised retrotransposon intermediate [54-58]. The L1 RNP then migrates from the cytoplasm to the nucleus using a mechanism that is still unknown [16] but can occur independently of cell division [10, 59]. Once the RNP enters the nucleus, integration of the L1 sequence into the host genome is thought to proceed in most cases via Target Primed Reverse Transcription (TPRT). The TPRT mechanism was discovered by Luan *et al.*, studying the retrotransposon R2 in *Bombyx mori* [60]. In this process, the L1 EN cleaves the first strand of the new L1 genomic locus with a preference for 5'-TTTT/A-3' motifs [61]. The EN-induced DNA break creates a free 3'-OH group, which is then used as a primer for reverse transcription of the L1 mRNA by the L1 RT [62]. The second DNA strand is usually cleaved some distance downstream of the initial cut [63] leaving single stranded regions that are eventually filled, forming direct repeats named target site duplications (TSDs) [64]. The end result of L1 integration is a new insertion flanked by these TSDs, which are characteristic of TPRT. Second strand cleavage can also occur directly opposite the first strand cleavage site, giving rise to blunt insertions without TSDs [64].

In addition to TPRT, there are other less common, and less understood, mechanisms of L1 mobilisation that deviate from the canonical mechanism of retrotransposition [62]. The EN-independent pathway is one of these examples [65]. In that case, L1 takes advantage of existing double-strand breaks (DSBs) in genomic DNA (gDNA) to initiate reverse transcription from the broken DNA fragment [66]. Products of this EN independent process typically lack hallmarks (i.e., TSDs, canonical L1 cleavage motif) of TPRT mediated insertions [65, 66].

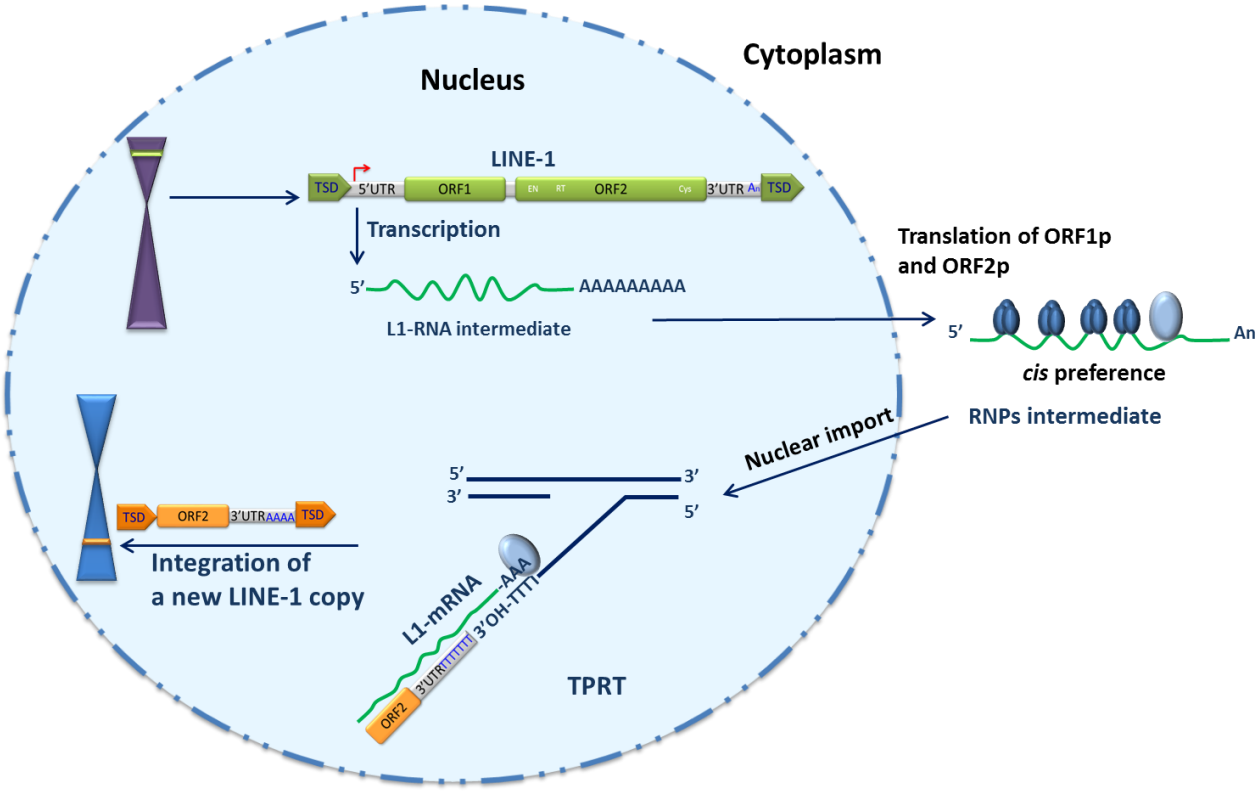


Figure 2: The L1 retrotransposition cycle. A full-length L1 is transcribed from its original location in the genome, represented in green situated in the purple chromosome; the transcript is exported and translated to produce ORF1p (represented in small blue circles) and ORF2p (represented in big blue circle). Both proteins bind to the L1 mRNA to form an RNP. The L1 RNP enters the nucleus and initiates TPRT in another region of the genome (represented as a turquoise chromosome). The L1 EN nicks the DNA, generating a free 3'-OH' residue that is used as a primer by the L1 RT to reverse transcribe the L1 mRNA template (green). The mechanism of second strand cleavage, second strand cDNA synthesis, and the completion of L1 integration are currently unclear. In this representation, TPRT results in a new L1 5'-truncated copy (orange bar in the turquoise chromosome). New L1 insertions are usually flanked by TSDs (dark orange arrows). This figure was adapted from [17, 67, 68].

4 Impact of L1 retrotransposition on the mammalian genome.

Retrotransposition is a major cause of genome structural variation, and is disproportionately likely to bring about phenotypic change. Unusually, for example, L1 retrotransposition is the source of resistance to human immunodeficiency virus type 1 (HIV) in owl monkeys. Owl monkey cells express a TRIM5–Cyclophilin A (CypA) fusion protein that blocks HIV-1 infection. This was the first example found in vertebrates of a chimeric gene generated by exon shuffling, where L1 *trans* mobilised CypA cDNA into the TRIM5

locus [69]. However, most new L1 insertions are neutral or deleterious. Since the original discovery by Haig Kazazian and colleagues in 1988 that L1 retrotransposition events could cause disease in humans, when they discovered *de novo* L1 insertions in the clotting Factor VIII gene of 2 unrelated patients with Haemophilia A [70], ~100 disease-causing mutations have been attributed to L1-mediated retrotransposition [3, 71, 72]. By disrupting exons [70] L1 insertions can disrupt gene coding sequences, induce mis-splicing, exon skipping and frameshift mutations [67, 73]. Moran *et al.* for example demonstrated, performing *in vitro* experiments, that engineered L1s (see below section 9.1) modified with artificial splice acceptor sites placed in the indicator cassette, can land in coding genes, oriented in sense or antisense [74]. That new L1 insertions can have such profound effects on gene function and expression means they have a disproportionate impact per event when compared to the 500,000 fixed L1 copies already found in the human genome, as most of the latter elements have been subject to aeons of evolutionary selection.

In addition, recent studies discovered that the L1 EN may produce double strand breaks that can contribute to genomic instability [75, 76]. Intronic L1 insertions can also impair genes via several routes (**Figure 3**). For example, due to the adenosine rich-content of L1 sequence, target genes can be inhibited by premature mRNA polyadenylation, transcriptional pausing or inhibition of transcription elongation that naturally affects L1s as well [77, 78]. Less commonly, antisense intronic L1 insertions can interrupt genes through a phenomenon known as gene breaking [79]. Due to these effects, L1 is considered a major factor in mammalian gene structural and functional evolution [17, 80].

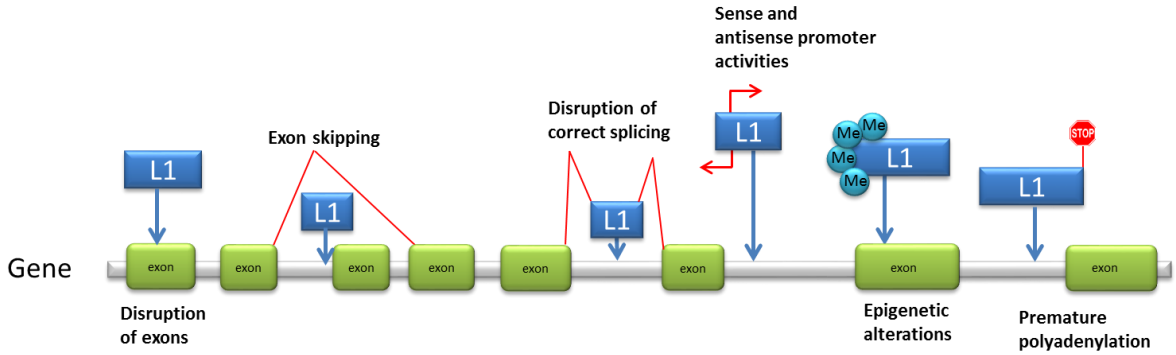


Figure 3: Schematic representation of the different ways by which L1 insertion can alter gene composition. A hypothetical gene locus is used to represent six different scenarios for L1 insertions: disruption of exons, exon skipping, incorrect splicing, provision of new sense and antisense promoter activities, epigenetic alterations and premature polyadenylation.

5 L1 retrotransposition can involve transductions.

L1 insertions carrying transductions (**Figure 4**) are of particular importance to this thesis. Transductions can flank L1 sequences that are mobilised along with the L1 into a new genomic location. These often unique "tag" sequences carry the necessary information, if long enough, to help identify the source element of a *de novo* retrotransposition event. Both 5' and 3' transductions can occur and these are generated in different ways; 3' transductions are produced when the canonical L1 Poly-A signal is not used, generating an L1 mRNA where polyadenylation is directed by an alternative downstream Poly-A signal [74, 81, 82]. By contrast, 5' transductions are far less common because they only appear if the L1 sense promoter, or another adjacent promoter, initiates transcription upstream of the L1 position +1, and the consequent L1 mRNA undergoes retrotransposition without being 5' truncated [17]. L1 transductions can include exons and promoters and could therefore generate new genes [83].

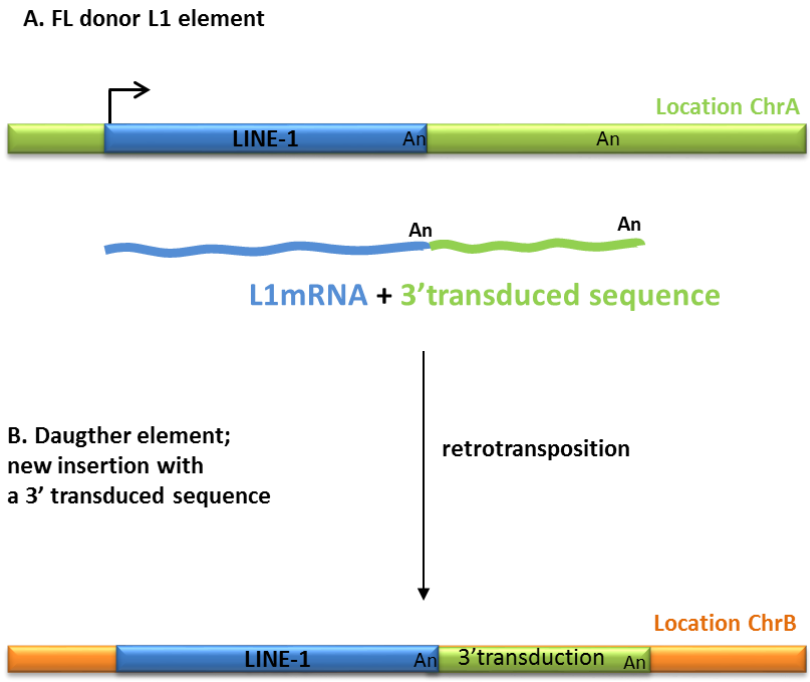


Figure 4: A L1 3' transduction. A. A full length donor L1 is represented as a blue rectangle at chromosomal location A, which is represented as a green rectangle. An L1 mRNA containing RNA transcribed from the 3' flank is reverse transcribed and integrated in a new location (ChrB, represented in orange) during retrotransposition. **B**. Daughter element; new insertion with a 3' transduced sequence. The new L1 insertion is represented as a blue rectangle in chromosomal location B (ChrB) as an orange rectangle. The 3' transduction is represented as a green rectangle in the new chromosomal location ChrB. Poly-A tracts are represented as An.

3’ transductions are highly variable in length but have an average size of 207 nucleotides [8, 81]. These sequences provide L1 insertions with a molecular link to their parental L1 and therefore allow us to track L1 elements that are still active [12]. Notably, the average length of human L1 3’ transductions is significantly less than would be predicted by random incidence of suitable Poly-A signals on the genome, either due to less efficient retrotransposition of these chimeric L1 mRNAs, or because L1 integrates preferentially in AT-rich regions [74, 81], increasing the likelihood of a strong Poly-A signal being located downstream of mobile L1s. The first example of an L1 3’ transduction was found to accompany a pathogenic, exonic insertion in the dystrophin gene of a muscular dystrophy patient [84]. The authors of this study referred to the transduction as a Unique Sequence Component (USC). The transduced sequence indicated a donor L1, known as LRE2, which was found to mobilise *in vitro*. Additional evidence for 3’ transductions was then obtained from an L1 retrotransposition assay conducted in cultured cells [74]. In this study, Moran *et al.* demonstrated that 2 more donor L1s responsible for pathogenic insertions could mobilize *in vitro*, and further identified the major functional domains necessary for efficient retrotransposition [22]. High frequency L1-mediated transduction was shown using a 3’ genetrap cassette containing an L1 Poly-A signal placed upstream of the reporter cassette gene (please refer to section 9.1 to see retrotransposition assay rationale). This vector system also contained a SV40 Poly-A signal which was moved downstream of the L1 copy and reporter cassette [74]. To select cells which carried retrotransposition events, G418 was added to the cell culture media. The elegant design of this system allowed antibiotic expression, and hence G418 resistant ($G418^R$) colonies, only if cells contained engineered L1 retrotransposition events generated from mRNAs employing the SV40 Poly-A signal. Cells containing retrotransposition events involving mRNAs polyadenylated at the direction of the L1 Poly-A signal were, by contrast, not G418 resistant. This engineered system allowed Moran *et al.* to specifically assess the characteristics of L1-mediated 3’ transductions *in vitro*, including their capacity to include exons and gene regulatory elements [74]. A speculative inference from these observations is that the weak native human L1 Poly-A signal supports genetic diversification during evolution, assisting retention of mobile L1 sequences in our ancestral genomes, at the cost of reduced L1 retrotransposition efficiency.

Analysing the HRG, two subsequent studies estimated the frequency of L1 mediated-3’ transductions *in vivo*, calculating the phenomenon was responsible for 0.6% [81] and 1% [82] of the human genome, respectively. The discrepancy in these percentages likely arose from the first study, Goodier *et al.*, searching for L1 sequences bearing 100 nucleotides on

their 3' end before the downstream TSD [81], whereas the other study only considered full-length L1 copies with 3' transduced sequences [82]. Goodier *et al.* also discovered the first evidence for 3' transduction families in human and mouse genomes, based on transductions with the same nucleotide sequence being presented by distinct L1 insertions and indicating a common progenitor element [81]. *In vivo* evidence for human L1-mediated transductions has subsequently been obtained from various normal and abnormal cellular contexts, including cancer (e.g. colon, hepatocellular, prostate, ovarian, lung) and in mice [6, 7, 9, 11, 12, 81, 82, 84-92].

5.1 Tracing internal mutations and 3' transductions as a means to identify active progenitor L1s.

Mutations can be encountered anywhere along a given L1 sequence, due to errors in L1 reverse transcription [93], as well as mutations arising in subsequent generations. These mutations can be used to group L1 sequences into families. However, human-specific subfamilies are evolutionary young, and their sequences are highly similar to each other. For example, the sequences of members of the L1 pre-Ta and –Ta family are, on average, >99% similar to other members of their family [94, 95]. Therefore, it is hard but not impossible to distinguish progenitor/offspring relationships between L1s without additional information, such as a transduction. Single-nucleotide resolution analyses of L1 sequences has been performed with a variety of approaches. However, it remains challenging to accurately sequence the entirety of a full-length L1 copy, or even one that is not heavily 5' truncated, requiring PCR and capillary (or PacBio) sequencing to validate insertions. For example, Scott *et al.* identified a tumour-specific L1 insertion in the tumour suppressor gene APC, which can result in Colorectal Cancer (CRC) if both copies are mutated [9, 96, 97]. In order to identify the source element which produced the insertion, Scott *et al.* analysed the internal mutational profile of the insertion, and compared it to a total of 295 full-length donor L1s that were identified in the genome of the CRC patient carrying the APC L1 mutation. These donor L1s carried between 7 and 147 unique mutations each. By evaluating the mutation profile of the tumour-specific L1 insertion in the APC gene, additional related L1 copies were identified apart from the source or progenitor element [9].

A more recent study by Gardner *et al.* analysed additional human genomes, sequenced by prior studies focused on population and clinical genetics, to identify mobile element insertions. The bioinformatic tool Mobile Element Locator Tool (MELT) used and developed by Gardner *et al.* was created to analyse high coverage Illumina whole genome

sequencing (WGS) data from 2534 human genomes [87, 98]. The interior L1 mutations identified, and subfamily analysis performed, by MELT confirmed the distribution of active L1s in the 1000 Genomes Project data sets [98] were similar to those observed in previous studies, providing additional validation for this method. Gardner *et al.* also identified 38 full-length donor L1s responsible for 121 L1 insertions bearing 3' transductions. Overall, 2.9% of L1 insertions carried 3' transductions and, notably, 56.2% of these were traced to 3 highly active progenitor L1s.

Gardner *et al.* then combined their 3' transduced-based analysis with data from previous studies of human cancer genomes, identifying a total of 113 full-length donor L1s. Of these, 40.7% were active exclusively in the germ line, 42.5% were only active in cancer and 16.8% were active in both the germline and cancer [87]. Moreover, using prior retrotransposition assay data for a subset of the donor L1s [11, 13, 22, 29], they studied if there was any correlation between the endogenous activity of a given L1 in the germ line or cancer with its activity in engineered cultured cell assays. In addition to donor L1s that were highly active in all three of the studied scenarios (germ line, cancer and *in vitro*), e.g. LRE3, other elements were only active in one or two contexts, such as a donor L1 on chromosome 22, which was highly active in cancer genomes [8]. Other studies, such as Nguyen *et al.* have also tested the retrotransposition competency, in a cultured cell assay, of donor L1s that mobilise in human cancer genomes [6], as did an analysis of murine hepatocarcinogenesis by retrotransposon capture sequencing (RC-seq), which found another donor L1 mobile in cancer due to an insertion carrying a 3' transduction and tested its retrotransposition capability *in vitro* [7].

3' transduced sequences have also been used to extensively identify lineage progenitor L1s found amongst the human population. Of the 68 full-length polymorphic L1s identified by Beck *et al.* [11], 25 carried a 3' transduction and, in turn, 17 of these could be grouped into L1 transduction families. A likely source element was found for one of these subfamilies and for two other cases [29, 99] 3' transductions were shared with pathogenic L1-Ta insertions [11]. Another system developed to detect active L1 transduction families is Transduction-Specific Amplification Typing of L1 active subfamilies (TS-ATLAS), which built upon the transposon display system ATLAS (Amplification Typing of L1 active subfamilies) designed for selective and specific amplification of L1 junctions at both termini of elements that are not present in the HRG [100]. In this study, Macfarlane *et al.* amplified L1s containing shared 3' transduced sequences and discovered new members of 3 L1 transduction families: $L1_{RP}$, AC002980 and LRE3 [29, 99, 101, 102]. Furthermore, a likely lineage

progenitor element for a disease-causing insertion ($L1_{RP}$) was identified and found to have high activity in an *in vitro* retrotransposition assay [12]. Finally, the authors showed evidence for variable Poly-A tract length in the mentioned L1 transduction lineages, and demonstrated highly variable allele frequencies amongst individuals, showing that retrotransposition continues to play a role in human genetic evolution [12].

Retrotransposition events, including those carrying transductions, can occur during embryogenesis and in somatic cells. For example, van den Hurk *et al.* reported a pathogenic L1 insertion into the CHM gene, responsible for the progressive eye disease choroideremia [103]. The 3' transductions carried by this insertion elucidated a donor L1 elsewhere on the genome. Through insertion-side specific PCR assays, the authors then determined that the CHM L1 insertion arose during the embryonic development of the affected patient's mother, hence demonstrating endogenous L1 mobilisation during human embryogenesis [103]. Single-cell genomic analyses of human neurons have also revealed somatic L1 mutations carrying transductions [89, 104], enabling the tracing of cell lineage clones in the healthy human brain [89] and the identification of donor L1s potentially active during neurodevelopment due to cell-type and locus-specific euchromatinisation [105, 106]. One L1 insertion carried a 3' transduced sequence of 614 bp, which allowed the identification of the source element [89], whilst another insertion carried a 5' transduced sequence of 101 bp that similarly indicated a distal donor L1 [104]. In mice, among insertions that occurred during early embryonic development, Richardson *et al.* identified three L1 insertions carrying 3' transduced sequences out of eleven *de novo* L1 insertions discovered by RC-seq [88, 107], and demonstrated the activity *in vitro* of two of these *de novo* L1 insertions. These analyses collectively demonstrate endogneous donor L1 activity in the mammalian germline, early embryo and tumour cells.

6 L1 mobilisation in pluripotent human cells.

To be inherited by progeny, new retrotransposon insertions must occur in cells contributing to the parental germline, either prior to germ cell specification in the early embryo, or later in development, during oogenesis or spermatogenesis [29, 103, 108-111]. New heritable L1 insertions are still occurring in the human population [70, 112, 113]. Despite recent work in mouse suggesting most of these events arise during embryogenesis or primordial germ cells [88], the developmental timing of most new heritable L1 insertions remains unclear and, aside from case studies involving pathogenic L1 mutations [70, 88,

103], our knowledge in this area is mostly derived from studies of engineered and endogenous L1 activity in cultured pluripotent cells.

6.1 L1 activity in hESCs.

Human embryonic stem cells (hESCs) are pluripotent cells derived from the blastocyst inner cell mass [114]. They can differentiate to the 3 embryonic germ layers, endoderm, mesoderm and ectoderm, and have been used extensively to model likely L1 activity during very early human development. Several studies have reported elevated L1 expression in hESCs [1, 2, 45, 115, 116]. For instance, Garcia-Perez *et al.* reported engineered L1 mobilisation in 3 different hESC lines and, by recovering the genomic coordinates of 7 insertions, demonstrated that L1 can insert into genes expressed in hESCs, and that some of those insertions can contain deletions in their nucleotide sequence. Remarkably, hESC lines containing an engineered L1 retrotransposition event could be differentiated into embryoid bodies and express mRNAs characteristic of the 3 germ layers. Hence, these data suggested that L1 retrotransposition could happen at early stages in human embryogenesis [115]. Wissing *et al.* elucidated full-length L1 mRNAs in two hESC lines, H9 and H13. Moreover, they observed decreased DNA methylation in 20 CpG dinucleotides located within the L1 5'UTR promoter region [2]. Using RC-seq, Klawitter *et al.* investigated endogenous L1, *Alu* and SVA mobilisation during the cultivation of 3 hESC lines. One *Alu* retrotransposition event was found in a cultured H9 cell population that was absent from a prior passage, supporting the view that L1, and L1-mediated, mobilisation was possible in the blastocyst [1, 103, 115]. This study did not identify any endogenous L1 insertions in the 3 hESC lines analysed, an outcome the authors ascribed to a lack of clonal expansion causing extensive genomic heterogeneity in the cell populations [1]. In sum, the consistent detection of L1 activity in hESCs suggests but does not prove that the early embryo is a major niche for heritable L1 insertions to arise in humans.

6.2 L1 activity in hiPSCs.

Human induced pluripotent stem cells (hiPSCs) are produced via directed ectopic expression of transcription factors, e.g. Sox2, Oct4, Nanog, Klf4 and c-Myc, in fibroblasts or other cell types [117]. Like hESCs, hiPSCs are capable of unlimited proliferation and can also generate the three primary germ layers. For biomedical applications, hiPSCs have the advantage of being compatible with the immune system of transplant recipients, thereby circumventing the problem of host immune rejection. Another advantage of hiPSCs is that they are free of the ethical barriers related to the use of materials derived from human

embryos [118]. Nonetheless, during reprogramming or expansion of hiPSCs *in vitro*, genetic and epigenetic aberrations can occur [119-123], including L1 mobilisation [1]. One explanation for L1 activity here is that during hiPSC reprogramming chromatin relaxation and epigenome-wide remodelling occur [124], providing a window of opportunity for L1 to retrotranspose [1, 5, 115] when L1 promoters are hypomethylated and undergo transcriptional activation.

The first reported study of L1 mobilisation in hiPSCs was performed by Wissing *et al.*, who demonstrated that reprogramming somatic cells into hiPSCs allowed mobilisation of engineered L1 reporter constructs [2]. Apart from showing that full-length L1 mRNA transcripts were upregulated and that L1 ORF1p was present in hESCs, they also studied L1 regulation in hiPSCs, finding hypomethylation of the L1-Ta family 5'UTR in 3 hiPSC lines, when compared to their parental human dermal fibroblast (HDF) cell lines. Wissing *et al.* also observed the same hypomethylation amongst 3 previously identified hot RC-L1s. To confirm endogenous retrotransposition could also occur in hiPSCs, Klawitter *et al.* applied RC-seq to a panel of 8 hiPSC lines and their parental fibroblast or endothelial cell lines [1], detecting and PCR validating 10 *de novo* retrotransposition events (7 L1, 2 *Alu*, 1 SVA). Four of these *de novo* retrotransposition events landed in protein-coding genes that are expressed in pluripotent stem cells [1] and, notably, 4/7 L1 insertions were full-length. Although none of these insertions carried transductions, Klawitter *et al.* tested two of them in the cultured cell retrotransposition assay [22] and found both retained mobility *in vitro*, thus showing that *de novo* L1 insertions could further propagate during hiPSC cultivation. Interestingly, in a previous study of hiPSCs, no *de novo* retrotransposition events were identified via Whole Genome Sequencing (WGS) applied to 9 hiPSC lines [125]. Another study reported de novo L1 insertions in hiPSCs via targeted L1 sequencing but could not PCR validate these [126]. Sequencing depth, bioinformatic parameters and PCR validation methods, potentially along with stem cell culture conditions and population bottlenecks in cell culture may explain these discrepancies [127].

Importantly, Klawitter *et al.* followed a *de novo* L1 insertion found in an intron of the CADPS2 gene of one hiPSC line, calculating the prevalence of the mutation alongside CADPS2 mRNA abundance via quantitative PCR after extended time in cell culture. CADPS2 expression and prevalence of the L1 insertion were anticorrelated, suggesting the mutation may have interfered with CADPS2 expression, although the mechanism of this interference was unresolved [1]. Given this observation was correlative, and the hiPSC line was not differentiated into a cell type where CADPS2, a key factor in neuronal biology [128],

was critical to cellular function, additional work is required to assess whether endogenous retrotransposition arising during hiPSC derivation has a functional impact.

6.3 The potential impact of L1 mobilisation on hiPSC biomedical applications.

The ability to generate hiPSCs has greatly impacted the stem cell field, opening multiple avenues in biological research and biomedical applications, including regenerative medicine. For example, hiPSCs have been used to create sheets of retinal pigment epithelium (RPE) to successfully treat patients with age-related macular degeneration [129]. hiPSC genome stability during culture, including perturbations caused by L1 mobilisation, nonetheless needs to be evaluated further before they are widely used as a source of cells in medical therapies [130]. Moreover, although the use of hiPSCs is becoming more and more widespread in research, consistency in quality control is often highly variable [131]. The hiPSC field has taken positive steps to introduce these potential issues. For instance, c-Myc is less commonly used now as a reprogramming factor due to its potential to act as a proto-oncogene [132]. Reprogramming technology has also been optimised to use non-integrating delivery vectors to circumvent the possibility of insertional mutagenesis [133]. In my view, it is remarkable that most of the endogenous L1 insertions found by Klawitter *et al.* were full-length and retained retrotransposition capability [1]. This observation suggests L1 may be an important and underrated source of mutagenesis during hiPSC production and cultivation, where the insertional pattern of *de novo* L1 insertions is unknown and largely random, which should be considered when using hiPSCs in biomedical therapies.

7 L1 mobilisation in the neuronal lineage.

Barbara McClintock first described somatic genome mosaicism due to mobile DNA activity in maize more than 70 years ago [134]. Nonetheless, retrotransposition was for many years thought to primarily occur in the germline, following the ‘selfish DNA’ model where retrotransposons would seek to propagate in situations where their transmission to subsequent generations was feasible [108]. Recently, however, it was discovered that L1 retrotransposition can also produce somatic mosaicism at discrete times during development, perhaps at a frequency actually higher than that of germline events [29, 88, 103, 110, 111, 115, 135].

Of particular relevance to this thesis is the surprising discovery by Muotri *et al.* of L1 retrotransposition events in the mammalian brain, and particularly in the neuronal lineage [136]. To place this observation in context, in the central nervous system (CNS), the lineage precursors of all neuronal and glial cells are neural stem cells (NSCs) [137]. These cells have the capacity to self-renew and differentiate into neurons, astrocytes and oligodendrocytes [114]. In the adult brain, NSCs are limited to neurogenic niches, such as the subventricular zone (SVZ), and the subgranular zone (SGZ) of the hippocampal dentate gyrus (DG) [138]. These resident NSC populations are of importance here, as the neuroanatomical regions they belong to were observed by Muotri *et al.*, and later works, to support L1 mobilisation. Muotri *et al.* found that rat neuronal precursor cells (NPCs) derived from hippocampal NSCs supported L1 mobility, as indicated by an engineered L1-EGFP (enhanced green fluorescent protein) reporter [31]. Furthermore, Muotri *et al.* saw L1 retrotranspose at a low frequency in NSCs, but observed an increase of L1 retrotransposition events only 48 hours after differentiation [136]. Coufal *et al.* subsequently demonstrated that human NPCs can also support retrotransposition [4]. In this case, cells were derived from human foetal brain and hESCs, and retrotransposition was again measured using an engineered human L1-EGFP reporter. In addition, Coufal *et al.* developed a quantitative multiplex polymerase chain reaction (TaqMan qPCR) assay to check the copy number variation (CNV) of endogenous L1s. They estimated 80 additional L1 ORF2 copies per cell in the hippocampus as a result of somatic retrotransposition, normalised to a plasmid control [4]. Additionally, Macia *et al.* demonstrated exogenous mobilisation can occur in post-mitotic human neurons using the L1-EGFP reporter [10]. These works, predominantly employing engineered L1 retrotransposition assays, broadly established the neuronal specificity of L1 mobilisation.

Somatic L1 retrotransposition may be enabled in the brain by two main regulatory pathways: Sox2/HDAC1 and Wnt-mediated TCF/LEF transcriptional activation [40]. The transcription factor Sox2, which is essential to maintain self-renewal of undifferentiated hESCs, binds to the L1 5'UTR and acts as a repressor, inhibiting L1 expression [4]. If the Wnt signalling cascade is activated, Sox2, a negative regulator of neuronal differentiation, is removed from the L1 promoter, activating L1 expression. Sox2 down regulation also promotes the expression of NeuroD1, a neurogenic transcription factor which promotes neuronal differentiation in hippocampus [40]. DNA methylation also regulates L1 promoter activity [36] by altering the capacity of transcription factors to bind the L1 5'UTR. MeCP2, for example, is a transcriptional repressor and DNA-methyl-binding protein that is particularly

important for the normal function of nerve cells [139] . MeCP2 mutations are responsible for the X-linked severe neurodevelopmental disorder Rett syndrome [140]. Interestingly, MeCP2 and L1 expression are inversely correlated throughout NSC differentiation and neuronal maturation, suggesting that MeCP2 modulates L1 activity during neural development [141, 142]. MeCP2 mutations in L1-EGFP transgenic mice also appeared to elevate neuronal L1 retrotransposition rates *in vivo* [136]. Thus, epigenetic regulation and transcription factors together modulate dynamic L1 activity in the brain.

Despite advances showing potential for L1 retrotransposition in neurons, and insights into how that process is regulated, the seminal studies by Muotri *et al.* and Coufal *et al.* were subject to significant caveats. Firstly, the L1-EGFP reporter system, introduced as a transgene and relying heavily on heterologous promoter sequences, may not recapitulate the regulatory landscape encountered by endogenous L1 retrotransposons *in vivo*. Secondly, the L1 CNV assay developed by Coufal *et al.* and based on qPCR does not characterise the sequences or genomic location of L1 insertions. To address these shortcomings, and directly detect endogenous L1 insertions in neurons, Baillie *et al.* developed a new technique called retrotransposon capture sequencing (RC-seq) [107]. RC-seq employs hybridisation to nucleic acid probes to enrich Illumina libraries for L1-genome 5’ and 3’ junction sequences, and has now been employed in mouse and human samples, in normal tissues, cancers and cultured stem cells [1, 6, 7, 88, 91, 107, 143]. By applying RC-seq to “bulk” DNA extracted from human hippocampus samples, Baillie *et al.* mapped the genomic integration site of thousands of putative somatic L1, showing L1 mobilisation into active protein-coding genes expressed in the brain, and therefore demonstrating endogenous L1 retrotransposition in this context.

Single-cell genomic analyses using RC-seq, other target approaches, and WGS, subsequently identified somatic L1 insertions in cortical and hippocampal neurons [89, 104, 143, 144]. Whole genome amplification (WGA) also generated sufficient material for PCR validation of L1 insertions in individual neurons, and quantification of how often somatic retrotransposition occurs in the brain. Notably, although these single-cell genomic analyses all reported neuronal L1 insertions, the estimated rate of L1 mobilisation remains debated, ranging from <0.1 to ~10 insertions per neuron, depending on the method, anatomical region and assumptions made during bioinformatic analysis [127]. In one study, Upton *et al.* reported a large number of somatic L1 insertions in hippocampal neurons and glia, and highlighted the potential for these events to generate a functional impact as they landed mainly in euchromatic regions of the genome [143]. Insertions were particularly enriched at

genes transcribed in the hippocampus and also in neuronal stem cell enhancer elements [143]. Given L1 integrates nearly randomly genome-wide [145, 146], these patterns may have been a result of post-integration selection. In another study, Evrony *et al.* identified two somatic L1 insertions in the cortex of an individual and, through lineage tracing, determined that both events occurred during brain development [89]. Erwin *et al.* found that rearrangements involving germline L1 copies in neurons may produce deletions in neuronally expressed genes [144]. Finally, Hazen *et al.* used somatic cell nuclear transfer to clonally amplify mouse olfactory neuron genomes, avoiding WGA, and finding 0-2 L1 insertions in 6 neuronal clones [147]. These studies collectively established the occurrence of endogenous L1 retrotransposition in the mammalian brain.

7.1 Potential impacts of L1 retrotransposition in neurogenesis.

Genomic stability is, for the most part in somatic tissues, clearly essential to protect cellular integrity and prevent neoplastic transformation [16]. Nonetheless, mitotic errors during DNA replication and recombination can cause intercellular genomic differences, i.e. somatic genome mosaicism [80, 108]. Notably, most of the early reported examples of somatic genome mosaicism in humans were associated with pathogenesis, including cancer and developmental syndromes [148]. There are some exceptions to this rule among healthy cells, such as V(D)J recombination in the adaptive immune system [149], where the domesticated transposase genes RAG1 and RAG2 produce somatic rearrangements, creating different receptors to elicit an immune response to foreign antigens [150]. As for lymphocytes, somatic mosaicism driven by retrotransposition and other genomic abnormalities [151-153] might be a source for functional diversification among neurons [143] and, to speculate, a route to neuronal plasticity underpinning cognition [108]. This process may be influenced by environmental factors, giving a possible mechanism for generating neuronal diversity in response to changes in the environment [108].. Moreover, in terms of hominid evolution, different L1 families are active in our nearest ancestors [33, 154] and, potentially, more or less mobile than L1-Ta in humans. Though it is interesting to question the role of L1 in many of the grander aspects of brain function, such as the evolution of cognition among hominids [80], we still lack essential information regarding L1 activity in the human brain, such as whether this phenomenon varies significantly amongst individuals, whether there is any correlation with ageing and, indeed, if certain neurodevelopmental stages accommodate somatic retrotransposition more than others. It is entirely possible that the main contribution of L1 to neuronal diversity involves cell-specific regulation of neuronal

genes by nearby L1 copies, which may be polymorphic [106, 155, 156]. Finally, despite reports of elevated L1 activity in neurological conditions, such as schizophrenia [157], it is still very unclear whether L1-driven mosaicism in the brain plays any role in psychiatric illnesses or neurodegeneration [158].

8 Obtaining a view of L1 in neurodifferentiation and neurological disease via hiPSCs.

Accessing endogenous neuronal cell populations from humans during neurodevelopment can be extremely challenging. As an alternative, hiPSCs can be derived from healthy and diseased individuals and then differentiated *in vitro* to yield cell types of interest, including neurons [117]. hiPSCs are also themselves roughly akin to hESCs in their genome-wide transcriptional profile and functional properties [159], enabling studies of early human embryogenesis with hiPSCs. In addition, hiPSCs provide an excellent model to study various stages of neurogenesis and neuronal maturation in healthy individuals and in the context of neurological disease because, complemented with post-mortem brain tissues, they allow us to obtain data by different approaches, enhancing the strength of conclusions. Therefore, hiPSCs are useful for studies of human neurobiology, and also can be used to test compounds that may modulate molecular processes, such as L1 retrotransposition, *in vitro*.

9 How to detect L1 retrotransposition:

9.1 Engineered L1 reporter assays in cultured cells and transgenic animals.

In 1996, Moran *et al.* reported a seminal method to create and detect engineered L1 retrotransposition events in cultured mammalian cells [22]. In this work, the authors transfected cells with vectors based on the L1 elements L1.2 and LRE2 to determine whether these could jump *in vitro*, and then examined their integration sites. Detection of new insertions was enabled by constructs where L1s were tagged with an indicator gene. Here, the indicator was an antibiotic resistance gene placed at the 3' end of the L1, antisense to the L1, disrupted by an intron in the same orientation as the L1 and carrying its own Poly-A signal. In sense to the L1, and placed 3' of the antisense oriented reporter gene, was an SV40 Poly-A signal. In this elegant system, a new retrotransposition event will be observed

only if the element was transcribed, processed by splicing, and then reverse transcribed and integrated into the host genome (**Figure 5**). Using this assay, Moran *et al.* demonstrated L1 mobilisation *in vitro* and, moreover, performed various mutational analyses showing that changes to key residues in ORF2 abrogated ORF2p reverse transcriptase activity. ORF1 mutations also reduced the retrotransposition rate [22]. Finally, Moran *et al.* found TSDs and other hallmarks of TPRT at the L1 integration sites. This study greatly illuminated the mechanistic basis for L1 mobilisation in mammalian cells, and provided a key method now used widely in the field to assess the retrotransposition competence of a given L1.

Subsequently, Ostertag *et al.* [31] used enhanced green fluorescent protein (EGFP) as an indicator gene, instead of an antibiotic resistance cassette, to assay L1 retrotransposition in cultured HeLa cells. They detected retrotransposition (i.e. EGFP expression) 48hr post-transfection. Using this L1-EGFP system, Ostertag *et al.* could track retrotransposition in individual cells in real time, greatly expediting calculations of retrotransposition rate in comparison to the antibiotic-resistance based L1 reporter, which requires at least 14 days in culture before cells are fixed and stained [22]. The assay is also sufficiently sensitive to detect changes in the retrotransposition rates among similar L1 elements [31].

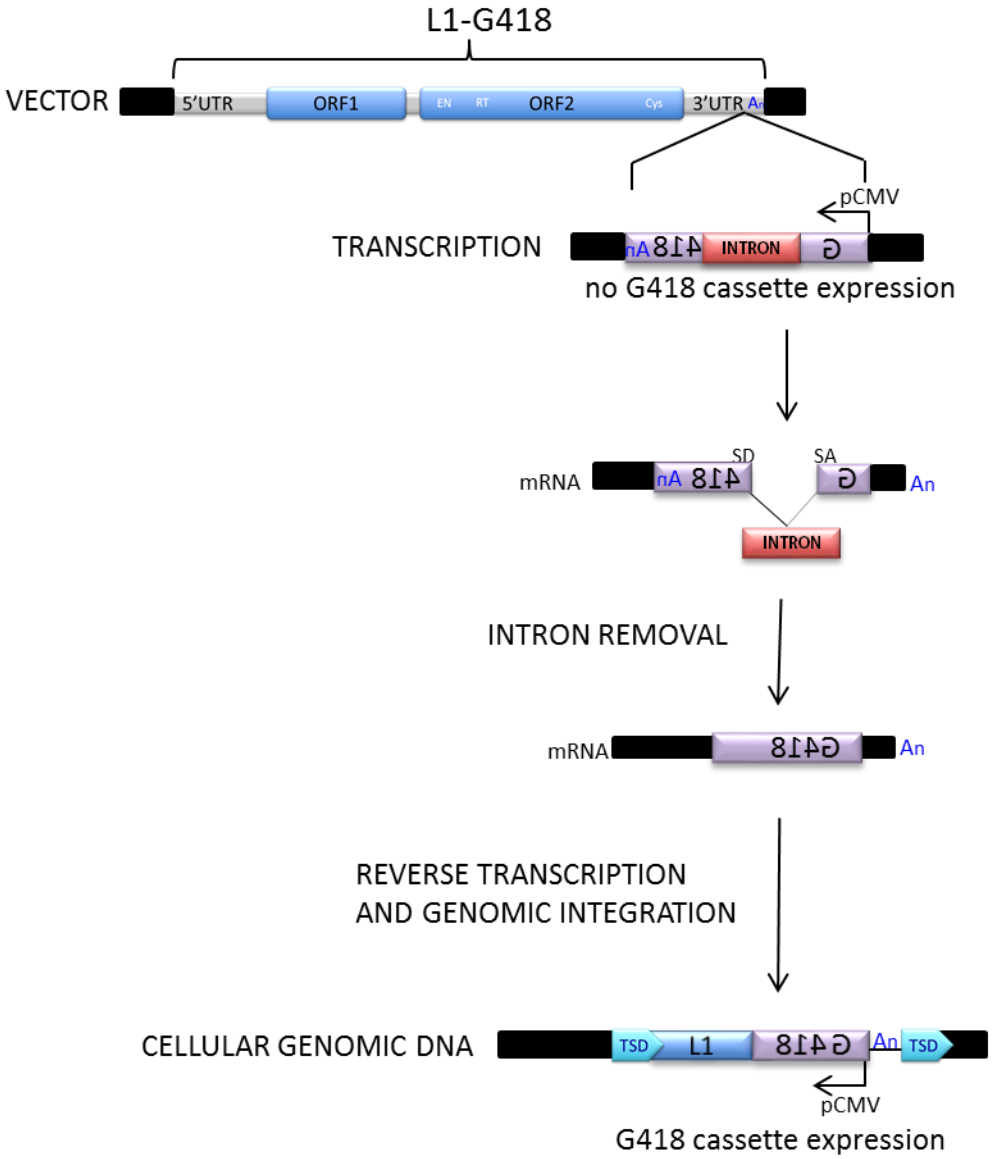


Figure 5: Retrotransposition assay schematic. The vector L1-G418 is represented at top, where L1 ORFs, L1 UTRs and the backbone are coloured blue, grey and black, respectively. Below the L1 is displayed the G418 expression cassette (light purple) tagged with an intron (red) and with a Poly-A signal (An). Also shown are the steps that the system needs to undergo to observe a new retrotransposition event, resulting in G418 gene expression: Transcription, splicing of the intron, reverse transcription and genomic integration and expression from the integrated genomic G418 (purple rectangle). This figure was adapted from [136]**.**

As an additional control, Ostertag *et al.* built upon an intron removal PCR assay [22] from Moran *et al.* where DNA extracted from cells transfected with the reporter plasmid was used as the input template for a PCR reaction, where primers were placed at the beginning and end of the reporter gene flanking the intron. The resultant DNA fragments were diagnostic for L1 retrotransposition, as an amplicon lacking the EGFP intron indicated retrotransposition had occurred.

Transgenic L1-EGFP reporter assays have been used in rodents to study L1 mobilisation *in vivo*, which has mainly been found in early embryonic states and in the male germline [110, 111, 135, 136, 160-162]. In these experiments, the spatiotemporal extent and frequency of retrotransposition can be determined by counting the number of EGFP positive cells in sections of animal tissue by microscopy, or by sorting EGFP positive cells by flow cytometry. These rodent models have been used to test the activity of numerous human L1s, including $L1_{RP}$ [110, 111, 160] and LRE3 [111, 161], as well as mouse L1s including TG_F21 [111] and the synthetic codon-optimised ORFeus_Mm [163], which supports efficient transcription elongation of ORF1 and ORF2 sequences [78]. Transgenic L1s have been tested with their expression driven by their endogenous promoter, or by a heterologous promoter [110, 111, 135, 136, 160-162], often showing mobilisation in either design. Conversely, there are two important caveats that must be considered when evaluating the retrotransposition rate of transgenic L1s. Firstly, the location of the transgene, which has only been established by some studies [135, 162, 163], may impact its expression. Secondly, epigenetic silencing of the reporter cassette upon retrotransposition can also occur [164]. For these reasons, engineered L1 reporter systems provide proof-of-principle evidence for L1 mobilisation *in vitro* but require corroboration by other approaches to assess endogenous retrotransposition patterns.

9.2 Whole-genome and targeted approaches to map L1 insertions.

Two general strategies are commonly used to identify endogenous L1 retrotransposition events: WGS and targeted sequencing of L1-genome junctions [113]. Several approaches, including RC-seq, are encompassed by the latter category [90, 100, 107, 112, 127, 165]. All involve bioinformatic analysis followed by PCR validation and Sanger sequencing of putative insertions to confirm they are truly *de novo* [166-168]. It is particularly important to recall that L1s are not fixed in the human population and, as a result, polymorphic L1 insertions are sometimes called as *de novo* by the initial sequencing analysis and this annotation needs to be corrected by PCR experiments [169-171]. PCR validation is further aided by characterisation of TPRT hallmarks to confirm retrotransposition, as opposed to another molecular process, has occurred to generate an L1 copy at a particular genomic locus. The length (6kb) and genomic copy number of L1 presents major challenges in trying to identify new L1 insertions, given that Illumina reads and insert sizes are typically ~150bp and <700bp, respectively. These features mean that Illumina sequencing cannot resolve the whole length of most new L1 insertions. As a result,

WGS and targeted sequencing approaches both rely on the detection of L1-genome junctions for the identification of L1 copies.

The main drawback, which can be prohibitive depending on the application, of WGS versus targeted approaches, is its cost. Otherwise, WGS presents a major advantage as it has the potential to identify both genomic junctions of a given L1 insertion, irrespective of whether the L1 is 5' truncated or inverted, or carries a transduction [70, 74, 81, 82, 172]. The detection of both ends of an L1 insertion allows characterisation of TPRT hallmarks *a priori*, allowing bioinformatics filtering of chimeric molecules generated during Illumina sequencing. Another, lesser but still important advantage of WGS, is its ubiquity; many labs use this technique, simplifying interpretation and dissemination of results. Conversely, the lower cost of targeted L1 sequencing approaches means that they can be applied to much larger cohorts, or applied at higher depth to bulk tissue samples to identify somatic L1 insertions or subclonal events in tumours [91, 107] and cultured cell lines [1, 6], as has been performed extensively with RC-seq in the work reported in this thesis.

9.3 The hunt for endogenous retrotransposition: RC-seq.

Retrotransposon capture sequencing (RC-seq) was first developed in the Faulkner laboratory to map somatic L1 insertions in human brain samples at single-nucleotide resolution [107]. RC-seq employs DNA capture probes targeting the 5' and 3' extremes of the human L1-Ta consensus sequence to enrich Illumina libraries for L1-genome junctions, including those of new L1 insertions (**Figure 6**). The method has evolved over several iterations to achieve higher capture efficiencies with different RC-seq probe designs [91, 107, 143], has been applied to cultured human stem cells and tumours [1, 6, 7, 91, 173, 174], mouse samples [7, 88], and to individual human hippocampal neuron genomes after WGA [143]. RC-seq can identify full-length L1s, which can be detected at both their L1-genome junctions, as well as 5' truncated L1s, which are detected by RC-seq probes only at their 3' junction [107, 143].

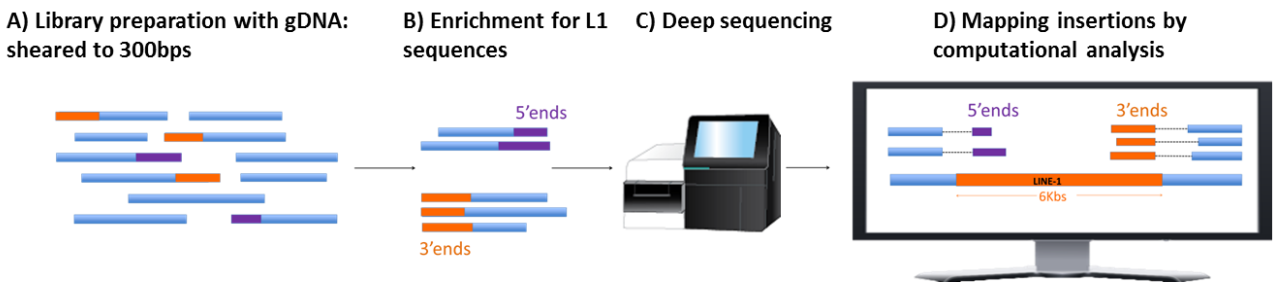


Figure 6: **RC-seq workflow**. A) Library preparation with Illumina adapters. B) Enrichment for L1-genome junction sequences using biotinylated capture probes. C) Enriched libraries

are processed and sequenced in paired-end 2×150mer mode on an Illumina platform. D) Computational analysis.

Empty/filled, 5' junction and 3' junction PCR assays are commonly used in the L1 field to confirm the presence of a *de novo* L1 insertion [166]. In the empty/filled assay, the filled site represents the L1 insertion. If the insertion is somatic or heterozygous, an empty site representing the allele lacking an L1 insertion will also be amplified. As their names suggest, 5' and 3' junction PCR assays specifically target L1-genome junctions. By Sanger sequencing the amplicons generated by these assays, one can assess whether a given L1 insertion carries TPRT hallmarks, including TSDs, a 3' Poly-A tract and integration at degenerate L1 EN motif. A common scenario for PCR validation of an L1 insertion is displayed in **Figure 7**.

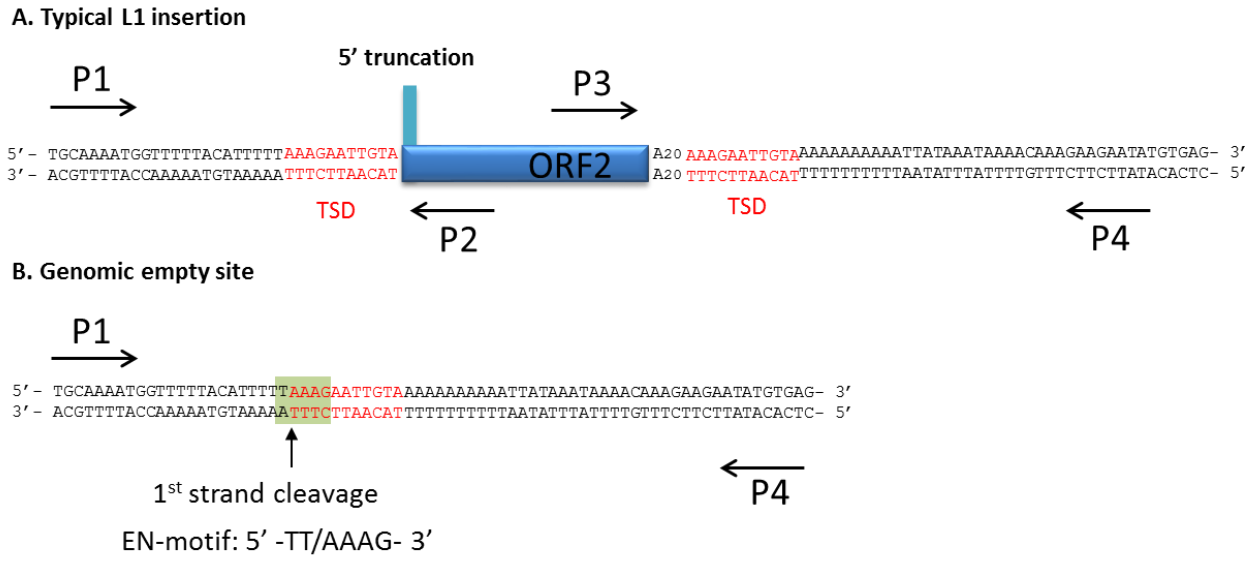


Figure 7: Structure and PCR validation of an L1 insertion. A) A typical L1 insertion with a 5´ truncation, TSDs (red nucleotides) and a Poly-A tract. P1, P2, P3 and P4 indicate primers used PCR amplify and validate the insertion. P1 and P4 are used in an Empty/filled PCR assay, P1 and P2 are used in a 5' junction PCR assay and P3 and P4 are used in a 3' junction PCR assay. B) The empty site prior to L1 integration, with the EN cleavage site highlighted in green. This figure was adapted from [166].

10 Disease-causing L1 insertions.

L1 insertions can disrupt gene function and expression, and cause pathogenesis [17, 71, 72]. To date, more than 100 cases of disease mediated by L1 have been described in the literature [17, 71]. The first of these were reported in ground breaking work by Kazazian

et al. in 1988. In this study, two independent L1 insertions were found in exon 14 of the blood clotting Factor VIII gene, causing haemophilia A [70]. Another remarkable early example of L1 as a mutagen was found by Miki *et al.*, who observed an exonic L1 insertion into the tumour suppressor gene APC of a colon cancer patient [9, 85]. The impact of L1 insertions in tumours has since been considered in numerous cancer types by genomic analyses, including in lung cancer [90], ovarian carcinoma [6], hepatocellular carcinoma (HCC) [7, 91], colorectal cancer [9, 174], oesophageal adenocarcinoma [175] and glioblastoma [173] and others [17, 86]. However, very few exonic insertions of the kind reported by Miki *et al.* have been found by these genome-wide studies [8, 35, 85, 86, 176], although it has been proposed that intronic L1 insertions, and those occurring after tumorigenesis, may also be important to cancer patient clinical outcomes [6, 91].

Pathogenic L1 insertions can be helpful to identify RC-L1s [26], as in the case of a rearranged L1 insertion found in the dystrophin gene of a muscular dystrophy patient, where a 3' transduced sequence allowed the authors, for the first time, to trace its source RC-L1 [84]. L1 insertions causing disease can also shed light on the mechanisms by which mobile elements can disrupt coding gene sequences [17, 71] and generate new transcriptional units through a phenomena called "gene breaking" [79]. Additionally, the adenosine-rich composition of L1 sequences can promote premature protein-coding gene mRNA polyadenylation, and attenuate gene expression by causing RNA polymerase II pausing [77, 78]. Ongoing L1 activity can impact host genome function in a variety of different cell types and developmental stages, as has been shown by a range of studies [70, 85, 103, 174], and new scenarios for L1 causing pathogenesis are likely to continue to be elucidated. In this way, we obtain a better understanding of L1 biology and, perhaps in the future, will be able to develop treatments or prognostic markers based on L1 activity, for example in cancer [177].

11 Mechanisms of L1 repression: the host genome defends itself.

L1 is repressed by the host genome at multiple steps of the retrotransposition cycle. To be able to mobilise, L1s need to initiate and transcribe a full-length mRNA from the internal promoter located in their 5'UTR. The majority of this promoter is located in the first 100bp of the L1-Ta 5'UTR [36]. Although the defences against L1 retrotransposition are multilayered, epigenetic repression, as a first-line obstacle to transcription, is arguably the

most significant [167]. Central to L1 epigenetic regulation is a CpG island present in the L1 5'UTR [178] that, via DNA methylation, can mediate L1 repression. L1 methylation has been investigated in various contexts, either looking genome-wide [4, 5] or at specific L1 loci [6-9]. One key study investigated the genome-wide methylation state of the L1-Ta family in 3 hiPSC lines and 2 hESC lines, finding lower methylation in hiPSCs and in hESCs when compared to parental fibroblasts. The same work also studied the methylation state of 3 "hot" RC-L1 loci, also observing hypomethylation in pluripotent cells [2]. Significant genome-wide L1-Ta hypomethylation has also been demonstrated in hiPSCs generated from fibroblasts and human cord blood endothelial cells (hCBECs) [1]. L1-Ta promoter hypomethylation in hiPSCs [164], hESCs [179] and in hESC-derived neural progenitor cells (NPCs) [4] coincides with elevated L1 mRNA levels, establishing a relationship between L1 promoter hypomethylation and L1 transcription during development. In cancer, DNA methylation analyses targeting specific RC-L1s in the same tumours where they have generated somatic L1 insertions have been performed by several studies [6-9]. For example, an RC-L1 on chromosome 17 was found by one study to be responsible for an exonic insertion in the APC gene of a colorectal cancer patient and, in the normal and tumour samples of this individual, the RC-L1 5'UTR was hypomethylated [9]. In another study, L1 locus-specific bisulfite sequencing indicated hypomethylation of RC-L1s mobile in HCC patient samples, and a mouse model of HCC [7]. These works collectively show how the absence or depletion of DNA methylation from RC-L1 promoter sequences is a precursor to L1 mutagenesis.

DNA methyltransferases (DNMTs), including DNMT1, DNMT3a and DNMT3b, establish and maintain DNA methylation [180]. As mentioned in previous sections, MeCP2 inhibits L1 retrotransposition by binding to methylated CpG dinucleotides in the L1 promoter [141]. Additionally, the nucleosome and deacetylase multiprotein complex (NuRD) is enriched at the L1 promoter, along with the histone deacetylases (HDACs) HDAC1 and HDAC2 [181]. E2F-Rb family complexes influence the epigenetic regulation of human and mouse L1s via nucleosome modifications and recruit HDAC1 and HDAC2 during early embryonic development. Hence, L1 sequences are centres for heterochromatin formation in a Rb family-dependent manner [182, 183]. Otherwise, in primordial germ cells (PGCs), deletion of the *de novo* methyltransferase 3-like protein (DNMT3L) results in overexpression of L1 mRNA in the mouse male germline, indicating it may be required to methylate L1s *de novo* [184].

To control L1 transcription and further attenuate the mobilisation of RC-L1s, numerous transcription factors regulate L1 expression, such is the SRY proteins SOX2 and SOX11 [38, 136], YY1 [41], RUNX3 [37], p53 [185] and NeuroD1 [40]. In addition, L1 can be transcriptionally silenced by Kruppel-associated protein, zinc-finger protein (KAP1, also known as TRIM28). Another protein described to repress L1 mobility by ribosylating KAP1, a cofactor which serves as a scaffold for heterochromatin complex comprising the SETDB1 [186], is the protein deacylase and mono-ADP ribosyltransferase Sirtuin (SIRT6). SIRT6 helps KAP1 to interact with the heterochromatin factor HP1α. Therefore, SIRT6 contributes to the packaging of L1 into transcriptionally repressive heterochromatin [187]. Another transcription factor, also a zinc finger protein, is PLZ, which maintains and mediates epigenetic silencing of L1 in germ cells, progenitor and haematopoietic stem cells [188]. Hence, transcriptional repression is a second major layer of L1 control.

If, however, L1 achieves transcription of a full-length L1 mRNA, its mobility can still be limited at the post-transcriptional level. MicroRNAs, a class of small RNA, can for example suppress endogenous retrotransposition, such as in the case of miR-128, which inhibits L1 mobility by binding to L1 RNA [189]. The biogenesis of microRNAs mediated by the Microprocessor (Drosha-DGCR8) complex also leads to L1 mRNA degradation [190]. L1 mRNA and L1-encoded proteins accumulate in cells lacking a functional Microprocessor [190]. The piwi-interacting RNA (piRNA) pathway also seems to be involved in adaptive methylation of retrotransposons [191, 192]. A recent study showed the possible involvement of piwi proteins in controlling L1 retrotransposition in primate iPSCs [159]. Interestingly, these proteins act upstream of DNMT3L upon mouse retrotransposons [191, 192]. Piwi proteins also interact with the helicase MOV10, playing a role in silencing retrotransposons in the mouse germline [193]. At the post-translational level, ORF1p phosphorylation is necessary for L1 retrotransposition [194]. Novel anti-L1 retrotransposition mechanisms have also been shown to involve several proteins of the cytidine deaminase family, APOBEC3A, APOBEC3B and APOBEC3F [195, 196]. For example, APOBEC3A deaminases transiently exposed single-strand DNA during the process of L1 integration [197]. Additionally, RNase H2 has been reported to promote L1 mobilisation, probably acting on L1 mRNA:cDNA hybrid molecules, facilitating second strand production during TPRT [198]. Inhibition of retrotransposition has also been demonstrated in stress granules when promoting the sequestration of L1 RNPs [199]. Altogether, these studies demonstrate the complexity and importance of host genome defences against retrotransposition.

12 To recapitulate.

Early embryogenesis provides a major developmental niche for heritable L1 retrotransposition events to arise in mammals [88, 103, 200]. Cultivated human embryonic stem cells (hESCs) and induced pluripotent stem cells (hiPSCs) resembling the cells of the embryonic inner cell mass also express L1 mRNAs and support engineered and endogenous L1 retrotransposition [1, 10, 115, 116, 179]. *De novo* L1 insertions arising during embryogenesis or later development can cause somatic mosaicism [88, 111, 127, 135]. In particular, somatic L1 insertions have been reported in brain tissue [4, 89, 104, 107, 143, 144,Hazen, 2016 #302, 147]. Engineered L1 reporter genes also mobilise *in vitro* during neurogenesis and in post-mitotic neurons [4, 10, 136]. Importantly, the L1-Ta subfamily is hypomethylated in hESCs and hiPSCs, when compared to neurons and other differentiated cells, suggesting genome-wide developmental enforcement of L1-Ta promoter methylation [1, 4, 10, 116]. However, the related temporal profiles of DNA methylation and somatic retrotransposition for individual RC-L1s during neurogenesis are unclear.

Hence, in this thesis, I identified a reprogramming-associated *de novo* L1 insertion in a cultivated hiPSC line. The L1 insertion was traced to a “hot” donor RC-L1 that was part of an extended and recently active transduction family. I then measured locus-specific DNA methylation for the individual *de novo*, donor and transduction lineage family L1 promoters, as well as the L1-Ta subfamily overall, at multiple points of neurodifferentiation. These experiments significantly elucidate the dynamic temporal profile of epigenetic L1 repression applied to new and extant L1 insertions during neurogenesis *in vivo*.

CHAPTER 1: DETECTION AND CHARACTERISATION OF ENDOGENOUS L1 MOBILISATION IN hiPSCs.

hiPSCs serve as an *in vitro* model of the inner cell mass of human blastocysts [114, 117, 118]. In principle, heritable L1-mediated retrotransposition events should occur during gametogenesis or during the early stages of embryogenesis, where new L1 copies can be transmitted to progeny [201]. In recent work, Richardson et al., demonstrated that endogenous L1 retrotransposition occurs in mouse primordial germ cells and pluripotent embryonic cells [88]. L1 is also expressed in human male and female germ cells and hESCs [1, 2, 103, 202]. Consistently, studies of endogenous and engineered L1 mobilisation suggest that retrotransposition can occur in the germ line and during early embryonic development, as well as in certain somatic tissues [4, 29, 70, 136, 160]. hESCs and hiPSCs most likely have permissive epigenomes where L1 elements can potentially find a window of opportunity to mobilise [2]. Although reprogramming triggers retrotransposition [1], further investigation of L1 mobilisation and insertional pattern in hiPSCs is needed to fully understand the regulation and impact of L1 activity on the genomes of reprogrammed cell lines. Notably, *de novo L1* insertions carrying 3' transductions have not been described in hESCs or in hiPSCs, which has to date hindered efforts to identify specific RC-L1s that are active in these biological contexts, and infer relationships between parental and offspring L1 copies.

In the work reported in this chapter, I investigated endogenous L1 activity in hiPSCs. I performed bulk RC-seq using gDNA from two hiPSCs lines, and analysed these data in conjunction with previously published RC-seq data generated from the corresponding fibroblast populations [1]. The two hiPSC lines employed, hiPSC-CRL1502 (female) and hiPSC-CRL2429 (male), were derived from fibroblasts from healthy patients with oriP/EBNA1-based pCEP4 episomal vectors: pEP4EO2S, CK2MEN2L and pEP4EO2-SET2K [203]. Using RC-seq, I compared each hiPSC line (labelled as time point 1: T_1) with the corresponding parental fibroblasts (T_0) to identify insertions present in the hiPSCs but not the fibroblasts, representing presumed *de novo* insertions arising during or after reprogramming. RC-seq was also applied to each hiPSC line after prolonged neurodifferentiation *in vitro* (T_2-T_6). Germline insertions present in the fibroblast genomes prior to reprogramming were identified and filtered. I sought to characterise reprogramming-associated retrotransposition events carrying transduced sequences that could allow me to identify the corresponding donor RC-L1s. I identified and characterised a single

reprogramming-associated L1 retrotransposition event carrying both 5' and 3' transductions, allowing identification of its immediate donor L1. No *de novo* L1 insertions were found by RC-seq in the latter stages of neurodifferentiation at the detection thresholds used here.

1.1: Identification of putative retrotransposition events arising during hiPSC reprogramming or during neurodifferentiation.

To identify endogenous *de novo* L1 mobilisation events in hiPSCs and during neurodifferentiation, seven time points (T_0-T_6) from two cell lines were assessed in parallel. The time points consisted of fibroblasts (T_0), hiPSCs (T_1), neural epithelium (T_2), neural rosettes, 29 days after neural induction, denoting immature neurons (T_3) and three stages of prolonged neuronal maturation (T_{4-6}) at differentiation days 72, 112 and 156, respectively. Immunocytochemistry was performed on the various time points to verify cells expressed markers consistent with their annotation (**Figure 8**).

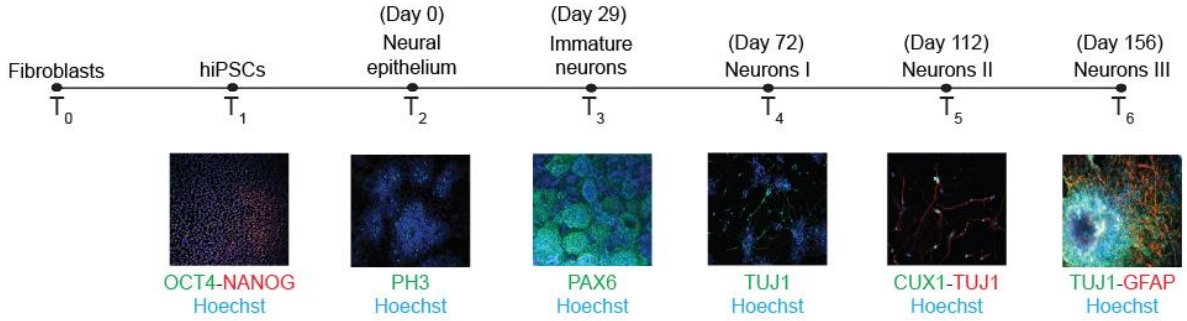


Figure 8. Above: schematic of the experimental approach of neurodifferentiation of hiPSCs reprogrammed from fibroblasts. TP, time point; T_0, fibroblasts; T_1, hiPSCs, 18 days post differentiation; T_2, neural epithelium, 29 days post differentiation; T_3, neural rosettes and immature neurons, 84 days post differentiation; T_4, neural rosettes and neurons I, 112 days post differentiation; T_5, neural rosettes and neurons II, 156 days post differentiation; T_6, neural rosettes and neurons III. **Below:** Characterisation of each time point: T_1, OCT4, NANOG and Hoechst; T_2, PH3 and Hoechst; T_3, PAX6 and Hoechst; T_4, TUJ1 and Hoechst; T_5, CUX1, TUJ1 and Hoechst; T_6, TUJ1, GFAP and Hoechst by immunocytochemistry. **Note:** The experiments shown were performed by Prof. Wolvetang's laboratory.

Next, RC-seq was performed on bulk genomic DNA (gDNA) extracted from the hiPSC-CRL1502 and hiPSC-CRL2429 cell lines at each time point. A cohort of 22 putative novel L1 insertions, denoted as "TC" for time course, and one SVA insertion, were identified by a stringent TEBreak analysis (https://github.com/adamewing/tebreak) and annotated as either polymorphic (previously published or present at T_0) or *de novo* (only present at T_1 or

later in one time course and not previously published in Klawitter et al., [1]) (**Table 1A, B, C)**.

According to the RC-seq output data table, from these 22 candidates, 17 were from the L1 family Ta [34], and 5 of them were from the family pre-Ta [11, 13]. Four putative *de novo* insertions landed in four different gene regions. Seven were detected at both L1-genome junctions, 5' and 3', and insertion TC_22 was only detected at its 3' end.

Table 1: List of putative *de novo* insertions. The following table shows the list of putative *de novo* insertions found in each cell line. Refer to **Appendix 4** for TSD_5prime, TSD_3prime, Genomic_Consensus_5p and Genomic_Consensus_3p sequence data.

Putative *de novo* L1 insertions	Chr	Superfamily	Subfamily	5_Prime_End	3_Prime_End
Timecourse_1	10	L1	L1Ta	130448104	130448098
Timecourse_2	12	L1	L1preTa	60413508	60413466
Timecourse_3	12	L1	L1Ta	89804489	89804494
Timecourse_4	12	L1	L1Ta	107960664	107960593
Timecourse_5	14	L1	L1Ta	55416805	55416749
Timecourse_6	18	L1	L1preTa	67763593	67763585
Timecourse_7	21	SVA	SVA_D	37160955	37160975
Timecourse_8	22	L1	L1Ta	32311246	32311234
Timecourse_9	4	L1	L1Ta	182778720	182778704
Timecourse_10	6	L1	L1preTa	134707566	134707605
Timecourse_11	6	L1	L1Ta	140975548	140975543
Timecourse_12	7	L1	L1Ta	69828257	69828287
Timecourse_13	9	L1	L1preTa	68202772	68202756
Timecourse_14	9	L1	L1Ta	83150175	83150175
Timecourse_15	X	L1	L1Ta	138146435	138146428
Timecourse_16	1	L1	L1Ta	59676272	59676261
Timecourse_17	1	L1	L1Ta	147333046	147333035
Timecourse_18	1	L1	L1Ta	231719300	231719315
Timecourse_19	12	L1	L1Ta	22116883	22116981
Timecourse_20	12	L1	L1Ta	75150880	75150913
Timecourse_21	3	L1	L1preTa	180758683	180758695
Timecourse_22	5	L1	L1Ta	37709844	37709844
Timecourse_23	7	L1	L1Ta	64554115	64554153

Table 1. A. Putative *de novo* L1 and SVA insertions: List of putative *de novo* candidates and their nomenclature. **Chr:** Chromosome where the putative *de novo* insertions mapped. **Superfamily:** The broadest retroelement classification. **Subfamily:** The L1-Ta subfamily is the most active in humans [13]. **5_Prime_End:** L1 5' end genomic coordinate (hg19 reference genome). **3_Prime_End:** L1 3' end genomic coordinate.

Putative *de novo* L1 insertions	Orient_5p	Orient_3p	TE_Align_Start	TE_Align_End	Inversion	Sample_count	Split_reads_5prime	Split_reads_3prime
Timecourse_1	-	+	4866	6054	Y	1	3	1
Timecourse_2	-	-	3592	6126	N	5	8	1
Timecourse_3	-	-	3947	6093	N	2	5	1
Timecourse_4	+	+	497	6059	N	2	4	1
Timecourse_5	+	+	1790	6057	N	5	1	7
Timecourse_6	-	-	606	6102	N	2	3	1
Timecourse_7	-	-	339	1434	N	5	1	9
Timecourse_8	+	+	244	6121	N	1	1	4
Timecourse_9	+	+	5556	6121	N	5	4	9
Timecourse_10	+	+	3	5994	N	2	4	1
Timecourse_11	+	+	5099	6121	N	8	22	14
Timecourse_12	-	-	5769	6121	N	2	5	1
Timecourse_13	+	+	5073	6126	N	5	13	7
Timecourse_14	None	-	3964	6065	NA	1	0	4
Timecourse_15	+	+	1569	6121	N	4	8	1
Timecourse_16	+	+	5332	6117	N	2	1	1
Timecourse_17	+	+	5538	6048	N	2	1	1
Timecourse_18	-	-	1	6118	N	4	10	4
Timecourse_19	+	+	1481	6048	N	2	1	1
Timecourse_20	+	-	5009	6121	Y	2	1	1
Timecourse_21	-	-	5550	6045	N	2	1	1
Timecourse_22	None	-	6002	6053	NA	2	0	3
Timecourse_23	-	-	3939	6121	N	2	2	1

Table 1. B. Putative *de novo* L1 and SVA insertions: List of the putative *de novo* candidates and their nomenclature. **Orient_5p:** Orientation of the 5' end of the insertion relative to the reference genome. **Orient 3p:** Orientation of the 3' end of the insertion. **TE_Align_Start:** Start position of the new insertion relative to L1 or SVA consensus. **TE_Align_End:** End position of the new insertion. **Inversion:** Presence or absence of an inversion in the new L1 insertion. **Sample_count:** Number of samples in which an insertion was detected. **Split_reads_5prime:** Number of unique RC-seq reads supporting the 5' end of the insertion. **Split_reads_3prime:** Number of unique RC-seq reads supporting 3' end of the insertion.

Putative *de novo* L1 insertions	GeneRegion	Sample_support
Timecourse_1	NA	hiPS_CRL2429: TP1.p40
Timecourse_2	NA	hiPS_CRL1502: TP1.p40, TP1.p76, TP2, TP4, TP6
Timecourse_3	NA	hiPS_CRL2429: TP1.p11, TP4
Timecourse_4	BTBD11,+,gene,transcript	hiPS_CRL2429: TP1.p40, TP6
Timecourse_5	WDHD1,-,gene,transcript	hiPS_CRL2429: TP1.p11, iPSC.C11.P1.p70, TP3, TP4, TP6
Timecourse_6	RTTN,-,gene,transcript	hiPS_CRL2429: TP1.p40, TP4
Timecourse_7	RUNX1,-,gene,transcript	hiPS_CRL2429: TP1.p11, TP1.p40, TP1.p70, TP3, TP6
Timecourse_8	NA	hiPS_CRL2429: TP6
Timecourse_9	NA	hiPS_CRL1502: TP1.p15, TP1.p40, TP3. TP4, TP6
Timecourse_10	NA	hiPS_CRL2429: TP1.p40, TP6
Timecourse_11	NA	hiPS_CRL1502: TP1.p15, TP1.p40, TP1.p76, TP2, TP3, TP4, TP5, TP6
Timecourse_12	AUTS2,+,gene,transcript	hiPS_CRL1502: TP1.p40, TP1.p76
Timecourse_13	NA	hiPS_CRL1502: TP1.p40, TP1.p76, TP3, TP5, TP6
Timecourse_14	NA	hiPS_CRL1502: TP1.p76
Timecourse_15	FGF13,-,gene,transcript	hiPS_CRL1502: TP1.p15, iPSC.C32.TP1.p76, TP3, TP4
Timecourse_16	NA	hiPS_CRL2429: TP1.p40, T.P 6
Timecourse_17	NA	hiPS_CRL2429: P1.p40, TP6
Timecourse_18	TSNAX-DISC1,+,gene,transcript	hiPS_CRL2429: TP1.p40, T.P2, TP3, TP6
Timecourse_19	NA	hiPS_CRL2429: TP1.p40, T.P3
Timecourse_20	NA	hiPS_CRL1502: TP1.p40, TP1.p76,
Timecourse_21	SOX2-OT,+,gene,transcript	hiPS_CRL1502: TP1.p40, TP6
Timecourse_22	WDR70,+,gene,transcript	hiPS_CRL2429: TP1.p70, TP2
Timecourse_23	RP11-460N20.5,+,gene,transcri	hiPS_CRL2429: TP3, TP6

Table 1.C. Putative *de novo* L1 and SVA insertions: List of the putative *de novo* candidates and their nomenclature. **Gene_Region:** In case of finding an insertion within a gene, this column indicates the gene name. **Sample_Support:** Samples in which the TEBreak pipeline found each insertion.

1.2: Manual analysis of RC-seq reads indicating putative *de novo* L1 insertions.

Each insertion was analysed *in silico* (**Appendix 1**). These analyses consisted of looking for the molecular features of TPRT [60]: Target Site Duplications (TSDs), integration at a degenerate endonuclease motif [49, 61] and a Poly-A tract [26, 35] upstream of the 3' TSD [64]. This part of the analysis was conducted to discard possible chimeras or PCR artefacts that passed the filters of the TEBreak pipeline [166]. Such artefacts can result from the random joining of two or more pieces of DNA that can be generated during multiple different steps in Illumina library preparation [166]. In many cases, the formation of chimeras is facilitated by the presence of microhomologies amongst two DNA fragments. For example, the putative de *novo* L1 insertion candidate TC_20 was called by TEBreak (https://github.com/adamewing/tebreak) as an inverted insertion, but the analysis of its sequence showed it was a recombination event or a PCR artefact. In addition, if a putative *de novo* insertion was found in one of the two hiPSC lines or its derivatives, it should not be present in the other cell line. It would be very unlikely that two different independent events

would have happened at the same genomic location and carry identical molecular features. Hence, insertions found in both cell lines would be considered as polymorphic germ line insertions, as there are many L1s in the human population that are not present in the reference genome sequence [8, 11, 86, 90, 91, 112, 113, 170, 171, 176, 204, 205].

RC-seq produces ~150bp paired-end reads that for a given insertion span the 3' L1-genome junction and, for full-length L1 insertions, the 5' L1-genome junction. Each read therefore contains a part of the L1 sequence and a part of its flanking genomic region. I used the sequencing output data from the RC-seq **Appendix 4**, Insert Consensus_5p, for the study of the 5' end and Insert_Consensus_3p, for the study of the 3' end. In both cases, the analysis consists of checking for an L1 sequence (either 5' end or 3' end) and its flanking regions. Below it is explained how the sequence of a new insertion (TC_18, shown as an example) should appear (**Figure 9**) in our study. The necessary steps to perform this *in silico* analysis are detailed in material and methods.

Insert consensus_5p:

CCATATTTTACATAATTTATAATTTATGAAATAGAATATTTGAATAATAAACCAATTCCATTTTGAAGGCTTCTGATGTAGTACTCTGT
TAAAAAAAAAAAAAAAAACTACTTGAGAAAAGTATGGATTGACTATATTGGAAGTTGCAAGGCCTGAGGAATGTTTTCCCGTGATTTTA
GTCCCTCTCATCAGTGTTCTATGCCTCAGTTCTGGTAACCCCAGAATGACGCTGTTACCTGACAGTATTCTAATGAAGATTAAAGAAAT
GACATCTGAAATAATGGAGGGGAGGAGCCAAGATGGCCGAATAGGAACAGCTCCGGTCTACAGCTCCCAGCGTGAGCGACGCAGAAGAC
GGTGATTTCTGCATTTCCATCTGAGGTACCGGGTTCATCTCACTAGGGAGTGCCAGACAGTGGGCGCAGGCCAGTGTGTGTGCGCACCG
TGCGCGAGCCGAAGCAGGGCGAGGCATTGCCTCACCTGGGAAGCGCAAGGGGTCAGGGAGTTCCCTTTCCGAGTCAAAGAAAGGGGTGA
CGGACGCACCTGGAAAATCGGGTCACTCCCAC

Insert consensus_3p:

AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAGAAATGACATCTGTTAATCTTATTAAGTTC
GCTGAGGCAGGATGAGGCTGGGTGCTGGTAAATGGAGCAGATAAATGATCTGAGACATCTACACTTTATTTTTCATGTTGAAAAGTCAT
CATAGAGAAACTTTCTCCTTTGGT

Figure 9: Example of a good candidate for PCR validation (sequences from TC_18 insertion). Insert consensus 5p: The genomic consensus sequence of the 5' end of a possible new L1 insertion should have (from 5' to 3') a gDNA sequence (represented in black letters with 259 nucleotides in this case) from the same genomic location as referred in the table, a Target Site Duplication (represented in blue letters and underlined (in this case, 16 nucleotides)) and a part of the L1 sequence (represented in red letters with 281 nucleotides in this case). We also observe a 5' transduced sequence (represented in green letters having 10 nucleotides). Depending if the L1 element was truncated or not we will find the 5' end of the L1 sequence at different positions with respect to the L1 reference sequence. TSDs are annotated by TEBreak (https://github.com/adamewing/tebreak) in **Appendix 4** but then they need to be analysed by manual inspection of the different genomic consensus sequences. **Insert consensus 3p:** The genomic consensus sequence of the 3' end of a possible new L1 insertion should have (from 5' to 3') L1 sequence which may begin at different positions with respect to the L1 reference sequence, but which should always end with a Poly-A tract (represented in red letters with 56 nucleotides). The Poly-A tract should be followed by a TSD (identical to the one we observed on the 5' end insert consensus sequence data; it is represented in blue letters and underlined with 16 nucleotides) and the gDNA sequence from the same genomic location as referred in the

table and present on the 5'end consensus sequence (represented in black letters with 130 nucleotides in this case).

1.3: PCR validation of endogenous L1 insertions.

After the analysis of the sequencing reads of the 23 putative *de novo* insertions, 15 were discarded due to being located in repetitive elements or showing obvious signs of being artefacts (see Materials and Methods). and the 8 remaining candidates were selected for PCR validation (**Table 2**): TC_2, TC_11, TC_13 TC_17, TC_18, TC_21, TC_22 and TC_23. For these 8 candidates, PCR primers were designed (**Figure 10, Appendix 2**) and then Empty-Filled and 3'junction PCR assays were performed (**Figures 11-14**).

Nomenclature	Expected Filled size (bps)	Expected Empty size (bps)	Expected 3'junction size (bps)
TC_2	2803	269	165
TC_11	1245	291	210
TC_13	1235	249	235
TC_17	895	385	181
TC_18	6213	96	294
TC_21	860	365	204
TC_22	425	374	286
TC_23	2815	633	194

Table 2. Expected sizes of the amplification products from the time course L1 candidate insertion PCRs.

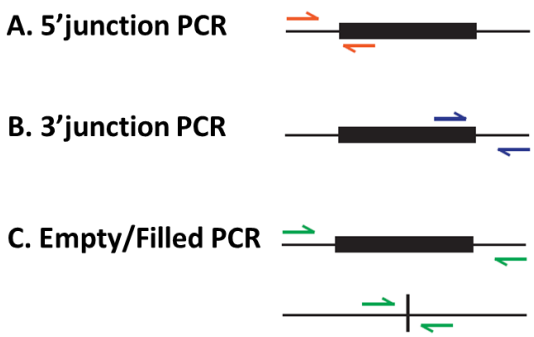


Figure 10: Schematic representation of PCR validations commonly attempted in the L1 field. Black boxes represent L1 elements. **A. 5' junction PCR.** Orange arrows represent primer position with respect the L1 sequence at the 5' end. **B. 3' junction PCR.** Blue arrows represent primer position with respect the L1 sequence at the 3'end. **C. Empty-Filled PCR.** Green arrows represent primer position with flanking the L1 copy. Above: Filled site. Below: Empty site.

The empty/filled PCR validation assay gel panel results (**Figures 11, 12 and 13**) confirmed validated for 3 of the 8 candidate L1 insertions. One insertion, TC_18, was a *de novo* retrotransposition event, whilst the other two were present at T_0 and were therefore polymorphic. Five L1 insertion candidates (TC_17, TC_22, TC_2, TC_21, TC_23) did not produce any validation products (**Figure 11**).

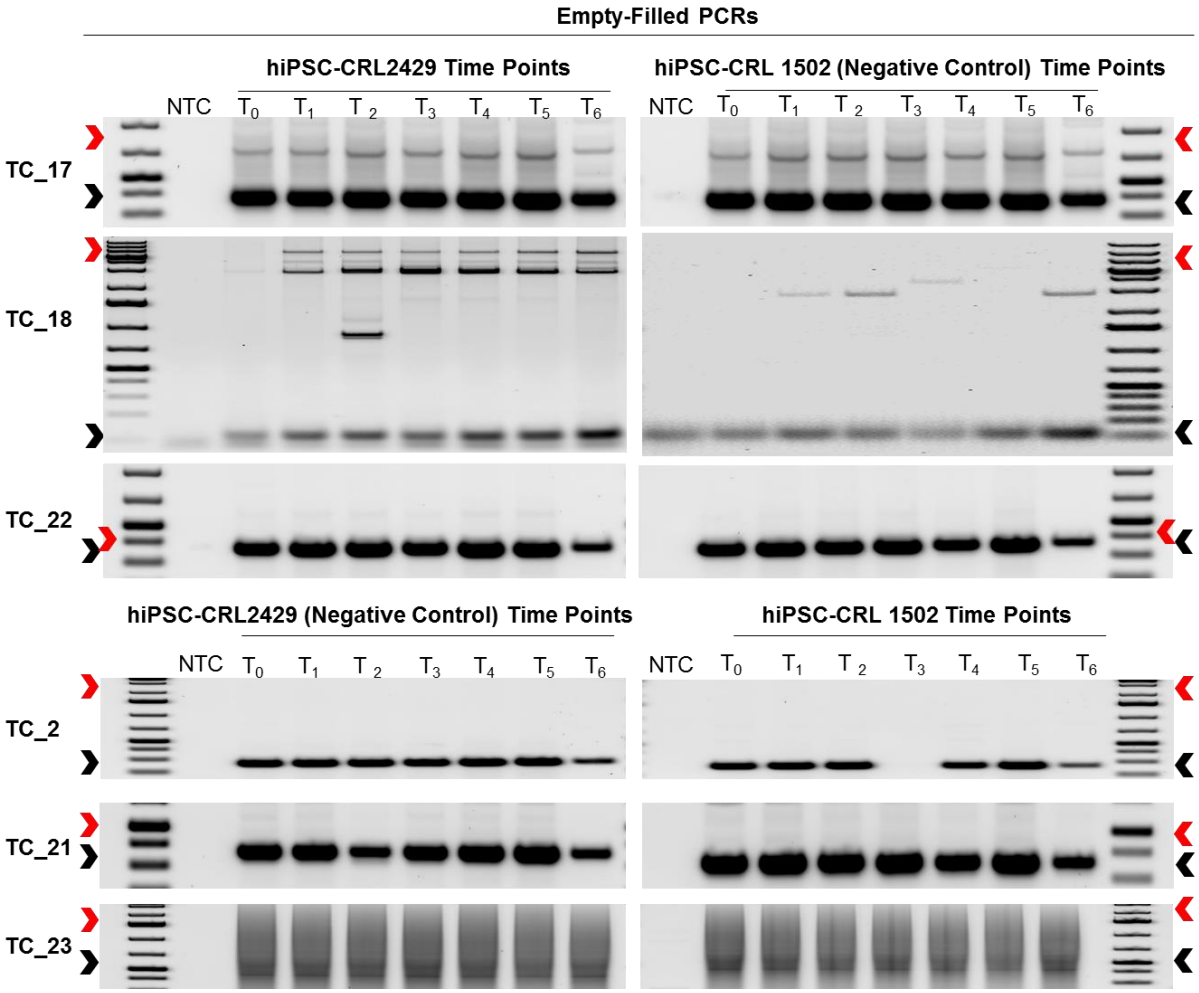


Figure 11: Empty-filled PCRs. On the left hand of the agarose gel panels are written the names of each putative *de novo* L1 insertion candidate. Insertions TC_17, TC_18 and TC_22, should be present in the cell line hiPSC-CRL2429 but absent in the cell line hiPSC-CRL1502. Insertions TC_2, TC_21 and TC_23, should be present in the cell line hiPSC-CRL1502 but absent in the cell line hiPSC-CRL2429. **Colour code:** Red arrows, expected filled size of the amplicons. Black arrows, expected empty size of the amplicons.

As expected from the results of the RC-seq output data table, the amplification product for TC_18 was present only in the cell line hiPSC-CRL2429, was around 6 Kb in length, and there was no amplification product at T_0. TC_11 and TC_13 were present at T_0 as well as the hiPSCs (T_1), making them polymorphic insertions. **Figure 12** and **Figure 13**

show the empty-filled PCR validation panels of the insertion TC_11 and TC_13, respectively. The cell line hiPSC-CRL2429 was used in both cases as a negative control. The amplification products in both cases at T_0 in the cell line hiPSC-CRL1502 confirmed that the candidates were not *de novo* retrotransposition events. The polymorphic insertions TC_11 and TC_13 were detected by RC-seq in hiPSCs or T_1 but not in HDFs or T_0. RC-seq failed to detect this insertion at T_0. Hence, this insertion was a false-negative result from the RC-seq approach. This fact emphasizes the necessity for PCR validation of L1 insertions.

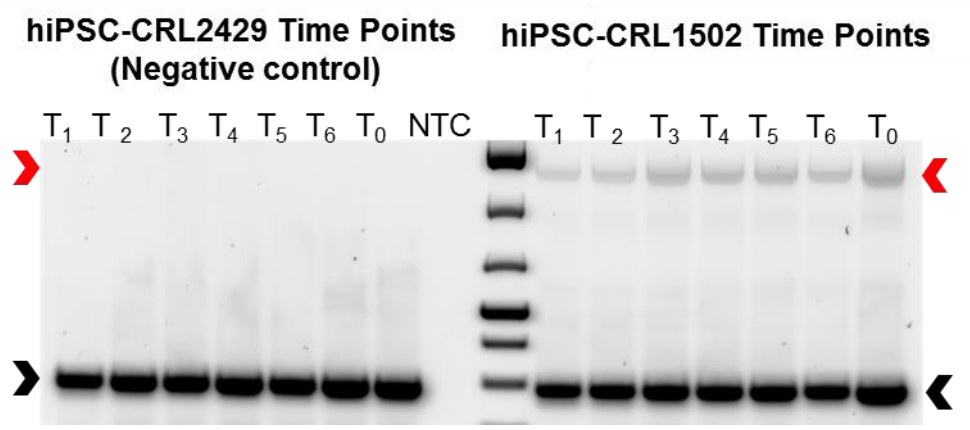


Figure 12: **Empty-filled PCR validation of the polymorphic insertion TC_11.** The insertion TC_11 found in T_0-T_6 in the cell line hiPSC-CRL1502. **Colour code:** Red arrow, size of the expected filled site band (1245 base pairs (bp)); Black arrow, size of the expected empty site band (291 bp).

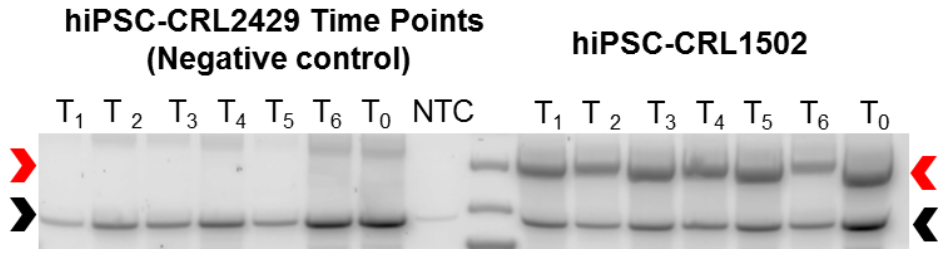


Figure 13: Empty-filled PCR validation of the polymorphic insertion TC_13. The insertion TC_13 found in the T_1, T_3, T_5 and T_6 derived from the cell line hiPSC-CRL1502. **Colour code:** Red arrow, size of the expected filled site band (1235 base pairs (bp)); Black arrow, size of the expected empty site band (249 bp).

The results obtained from the 3' junction PCR assay complemented results from the Empty-filled PCR assay (**Figure 14**). Amplification of the 3' end can be seen in the case of TC_18 in T_1 but not in T_0 in the hiPSC-CRL2429 cell line, and is entirely absent from the hiPSC-CRL1502 cell line and its derivatives. A 3' junction PCR was not performed in the case of TC_11 and TC_13 as these were clearly polymorphic from the Empty-filled assay.

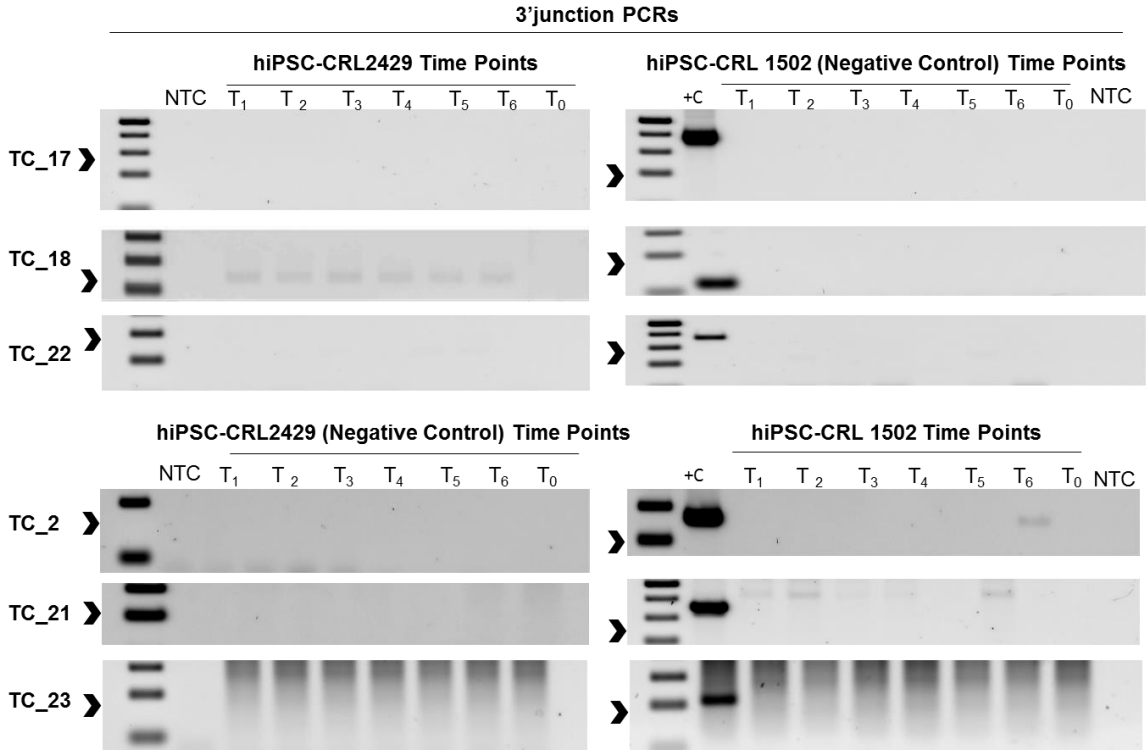


Figure 14: 3'Junction PCRs. On the left hand of the agarose gel panels are written the names of each putative *de novo* L1 insertion candidate. Insertions TC_17, TC_18 and TC_22, should be present in the cell line hiPSC-CRL2429 but absent in the cell line hiPSC-CRL1502. Insertions TC_2, TC_21 and TC_23, should be present in the cell line hiPSC-CRL1502 but absent in the cell line hiPSC-CRL2429. **Colour code:** Black arrows, expected size of the 3'junction amplicons. +C; positive PCR control using Fwd and Rv L1 primers with gDNA from T_0 to prove that the PCR was successful.

1.4: Sequence characterisation of PCR validated L1 insertions.

Once TC_11, TC_13, and TC_18 were confirmed as genuine L1 insertions, their filled site PCR amplification products were cloned in Topo-XL vector and capillary/Sanger sequenced (**Appendix 2**) to fully characterise their structural features (**Figure 15-17**).

Sequence analysis of the insertion TC_18 (also named as *de novo* L1 in this thesis) demonstrated a *de novo* full-length retrotransposition event of 6118 nucleotides which carried 2 transduced sequences flanking the L1 copy (**Figure 15**). The 5' transduced sequence observed was 10 nucleotides in length from the gDNA of its immediate donor element (called here the "donor L1") a previously-described "hot" polymorphic L1 [11], and the 3' transduced sequence presented 44 nucleotides from the gDNA of a reference L1 element [13], which we called the "lineage progenitor L1", being the source or parental element of both the *de novo* and donor L1s. Consistently, the lineage progenitor L1 presented only one Poly-A tract. The *de novo* L1 TC_18 presented two Poly-A tracts: the

first one was 17 nucleotides and the second one was 33 nucleotides. The TSDs of the TC_18 insertion were 16 nucleotides and the EN-motif was 5'-TT/AAAG-3'. The insertion was antisense to chromosome 1 (chromosomal location: Chr1:231,719,316). The family of the insertion was L1-Ta [30, 33]. This insertion landed in the TSNAX-DISC1 gene which is an RNA Gene and is classified as a RNA molecule non-translated into protein, a non-coding RNA (ncRNA). Diseases associated with TSNAX-DISC1 include Schizophrenia (http://www.genecards.org/cgi-bin/carddisp.pl?gene=TSNAX-DISC1&keywords=TSNAX).

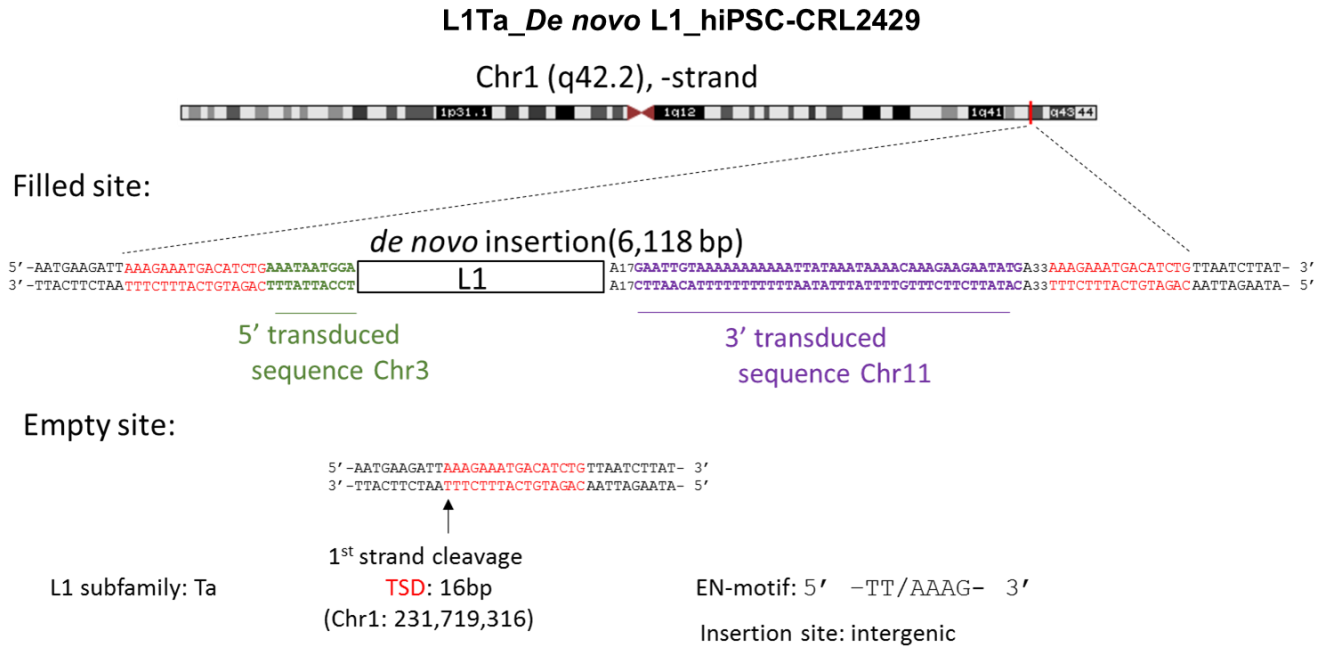


Figure 15: **Genomic location and structural features of the insertion TC_18 (*de novo* L1) detected by RC-seq in the cell line hiPSC-CRL2429. Filled site.** This figure shows a full-length L1-Ta element [34]. The insertion is 6118 nucleotides in length and contains 2 transduced sequences, one in the 5' end of 10 nucleotides (green letters) and another one in the 3'end of 44 nucleotides (purple letters). Black letters indicate genomic DNA (gDNA) where the insertion landed; Coordinates (Chr1:231,719,316), indicate the genomic position of insertion site. Red letters show the sequence of the TSDs flanking the L1 upon retrotransposition. Following the L1 sequence is a first Poly-A tract of 17 nucleotides in length and a second Poly-A tract of 33 nucleotides. **Empty site.** Red letters indicate TSDs; Black letters indicate gDNA; Black arrow, site of first strand cleavage by the L1 endonuclease.

Sequence analysis of the insertion TC_11 showed a 5' truncated polymorphic L1 element of 954 nucleotides (**Figure 16**). The L1 element was accompanied by a Poly-A tract of 44 nucleotides. The TSDs of the insertion were 6 nucleotides in length and the EN-motif was 5'-TT/GAAA-3'. The insertion was oriented in sense to chromosome 6, chromosomal

location: Chr6:140,975,331. The family of the insertion was L1-Ta [34]. This insertion landed in an intergenic region.

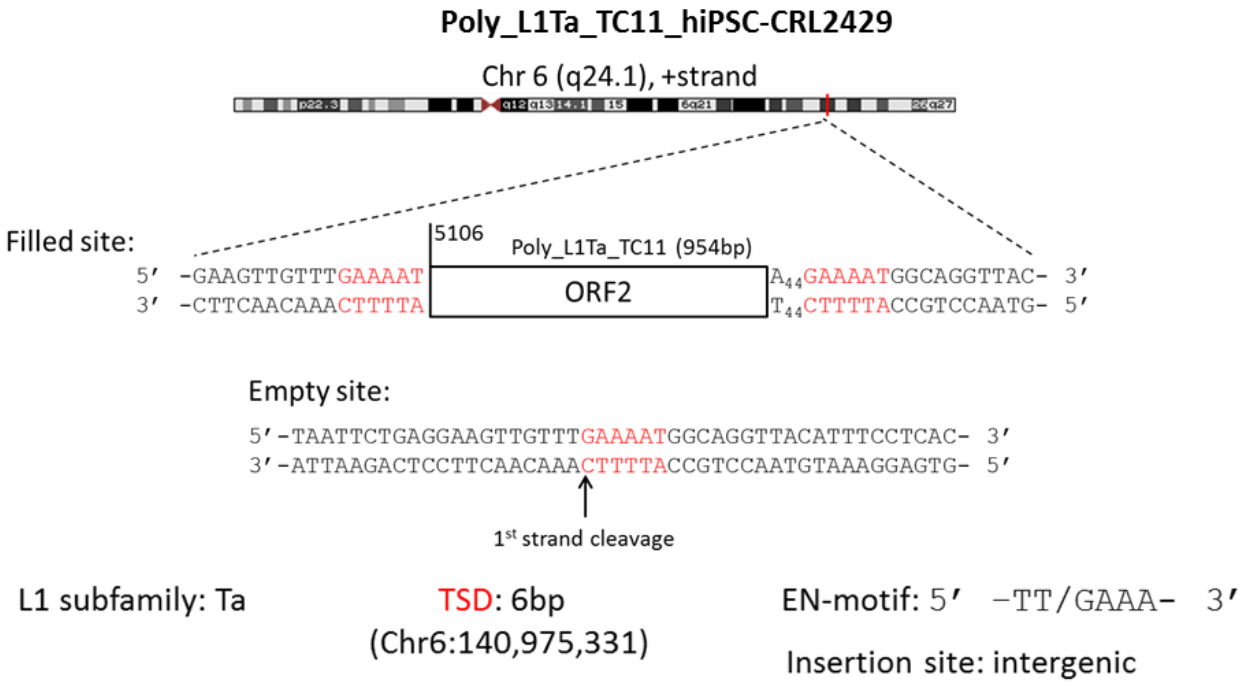


Figure 16: Genomic location and structural features of the polymorphic insertion TC_11 detected by RC-seq in the cell line hiPSC-CRL1502. Filled site. This figure shows the 3' end the polymorphic L1 element (ORF2) from the family Ta [34]. The insertion is 954 nucleotides in length and is truncated at the position 5106 with respect to a full-length L1 reference sequence. Black letters indicate genomic DNA (gDNA) where the insertion landed; Coordinates (Chr6:140,975,331), indicate the location of the insertion site. Red letters indicate TSDs flanking the L1 sequence upon retrotransposition. A Poly-A tract of 28 nucleotides follows the L1 sequence. **Empty site.** Red letters indicate the TSD; Black letters indicate gDNA sequence. Black arrow indicates the site of first strand cleavage by the L1 endonuclease.

Sequence analysis of the insertion TC_13 revealed a 5' truncated polymorphic L1 element of 986 nucleotides with an inversion of 171 nucleotides and a deletion of 177 nucleotides (**Figure 17**). The L1 element presented a Poly-A tract of 28 nucleotides. The TSDs of the insertion were 17 nucleotides in length and the EN-motif was 5'-TT/AAAA-3'. The insertion landed in the positive strand of chromosome 9, chromosomal location: Chr9:68,202,547. The family of the insertion was L1 pre-Ta [34]. This insertion landed in an intergenic region.

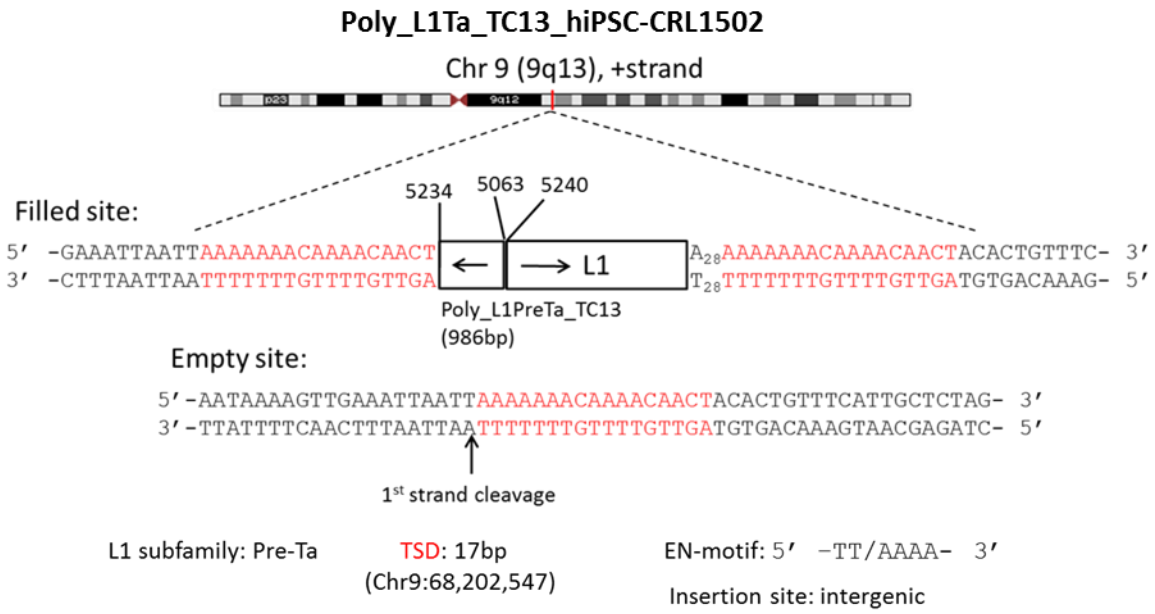


Figure 17: Genomic location and structural features of the polymorphic insertion TC_13 detected by RC-seq in the cell line hiPSC-CRL1502. Filled site: This insertion consists of a polymorphic L1 element (ORF2) from the L1 pre-Ta family. The insertion TC_13 is 986 nucleotides in length. This insertion has an inversion (nucleotides 5234 to 5063 with respect to a full-length L1 reference sequence) deletion (the sequence that is deleted is nucleotides 5063 to 5240). Black letters indicate gDNA where the insertion landed; Coordinates (Chr9:68,202,547), indicate the insertion site. Red letters indicate TSDs flanking the L1 sequence. A Poly-A tract of 44 nucleotides follows the L1 sequence. **Empty site:** Red letters indicate TSDs; Black letters indicate gDNA; Black arrow indicates the site of first strand cleavage by the L1 endonuclease.

CONCLUSIONS OF CHAPTER 1:

Of the 23 candidate *de novo* insertions detected by RC-seq [107], 8 were assessed to be strong candidates for PCR validation. One insertion, TC_18 (*de novo* L1) was successfully validated, characterised, and sequenced. The *de novo* L1 insertion was present in hiPSCs, throughout the neurodifferentiation time course, and was absent from the corresponding parental fibroblasts. As predicted by the RC-seq output data table, the insertion was full-length and was absent from the control cell line hiPSC-CRL1502. The sequencing results for the *de novo* L1 insertion demonstrated the presence of 2 transduced sequences flanking the L1 copy. This *de novo* L1 insertion was the first retrotransposition event reported in hiPSCs with both 5' and 3' transduced sequences. The nucleotide sequence of the 3' transduction allowed identification of the donor and lineage progenitor L1s that gave rise to the *de novo* L1 insertion. The identification of an extended L1 transduction family generated by the lineage progenitor L1 is described in Chapter 2.

CHAPTER 2: IDENTIFICATION AND CHARACTERISATION OF AN EXTENDED L1 TRANSDUCTION FAMILY.

In this chapter, I discovered that the *de novo* L1 insertion (TC_18) was part of an extended L1 transduction family comprising at least 14 elements. Various members of the transduction family, named here as L1_$_{1\text{-}14}$, have been previously described [8, 11, 86, 90, 91, 112, 113, 170, 171, 176, 204, 205] (**Table 3**) but, importantly, the relationships amongst them have not previously been elucidated. Several members of the family are still potentially active in the human population and, as demonstrated by TC_18, can generate reprogramming-associated L1 insertions in hiPSCs.

2.1: Identification of L1_$_{1\text{-}14}$ transduction family members.

Below is shown a summary of the L1_$_{1\text{-}14}$ family members described in this thesis, including the nomenclature I am using, their chromosomal location and whether they are reference L1s or non-reference elements.

Nomenclature	Coordinates (hg19)	Chr. location	Ref. element	Full-length
***De novo* L1 (TC_18)**	Chr1:231,719,316	q42.2	No	Yes
Lineage progenitor	Chr11:95,169,381	q21	Yes	Yes
Donor L1	Chr3:38,626,082	q22.2	No	Yes
Ref_Chr7_q21.3	Chr7:96,475,963	q21.3	Yes	Yes
Ref_Chr1_p31.1_a	Chr1:84,518,060	p31.1	Yes	Yes
Non-ref_Chr3_p24.3	Chr3:20,748,904	p24.3	No	Yes
Non-ref_Chr3_p12.2_a	Chr3:80,590,176	p12.2	No	Yes
Non-ref_Chr3_p12.2_b	Chr3:82,144,869	p12.2	No	Yes
Non-ref_ChrX_p11.4	ChrX:38,097,551	p11.4	No	Yes
Ref_Chr1_p31.1_b	Chr1:83,125,969	p31.1	Yes	No
Ref_Chr9_p23	Chr9:12,556,931	p23	Yes	No
Non-ref_Chr17_q12	Chr17:32,813,609	q12	No	No
Non-ref_Chr1_p22.2	Chr1:90,914,512	p22.2	No	No
Non-ref_Chr4_q12	Chr4:53,628,490	q12	No	No

Nomenclature: Ref, reference L1 element; Non-Ref, Non-Reference L1 element; 3' transd, 3' transduction

Table 3: **List of L1_$_{1\text{-}14}$ transduction family members**.

2.1.1: Identification of transduction family members present in the reference genome.

The 3' transduction carried by TC_18 was crucial to finding the other members of L1_$_{1\text{-}14}$: 5'-GAATTGTAAAAAAAAAAATTATAAATAAAACAAAGAAGAATATG-3'. When this 44nt sequence was mapped to the hg19 reference genome using BLAT in the UCSC genome browser, 5 different genomic loci presented full-length alignments with ≤2 mismatches [11]: Chr1:84,518,060, Chr11:95,169,381, Chr1:83,125,969, Chr9:12,556,931, and Chr7:96,475,963. *In silico* analysis in each case revealed an L1 immediately upstream of the transduced sequence, a first Poly-A tract and a second Poly-A tract, with the exception of the example of the alignment to Chr11:95,169,381, which had no second Poly-A tract. The lack of a second Poly-A tract and the precedence of the nucleotide sequence of the 3' transduced sequence indicated that this element, the lineage progenitor L1, is the likely original source element which gave rise to this L1 family. The 5 family members of L1_$_{1\text{-}14}$ identified in the reference genome comprised 3 full-length elements, named as Ref_Chr1_p31.1_a, lineage progenitor L1, and Ref_Chr7_q21.3, and two 5' truncated L1s, denoted as Ref_Chr1_p31.1_b and Ref_Chr9_p23.

Primers were designed to PCR amplify the lineage progenitor L1 and to subsequently clone it and then test its retrotransposition activity *in vitro*. The element was successfully amplified (**Figure 18**) and Sanger sequenced. Interestingly, this element was previously reported to be inactive by Brouha *et al.* [13], probably due to allelic variants disabling the version of the lineage progenitor L1 tested by that study. The lack of amplification product in the empty site demonstrated the homozygosity of the lineage progenitor in the cell line hiPSC_CRL2429. After Sanger sequencing, two different alleles were identified (**Results section 2.4**).

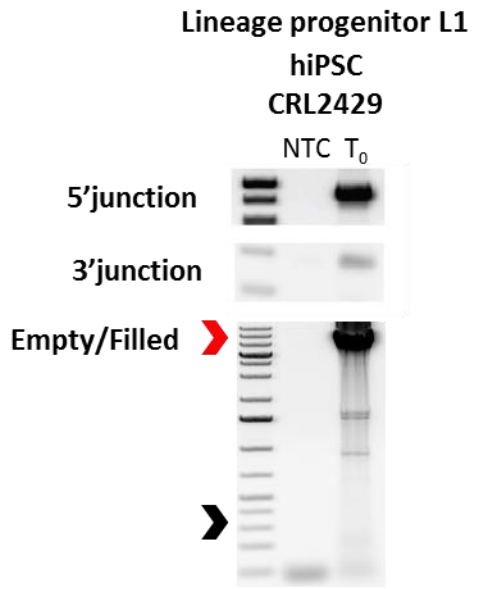


Figure 18: PCR validation panel of the lineage progenitor L1. 5' junction PCR, 3' junction PCR and empty/filled PCR gels. Amplification was observed in T_0 in the cell line hiPSC-CRL2429 in 5' junction PCR, 3' junction PCR and in the filled site (red arrow) of the empty/filled PCR assay. The absence of amplification product in the empty site confirmed the homozygosity of this element. NTC, non-template control.

The other full-length elements, Ref_Chr7_q21.3 and Ref_Chr1_p31.1_a, were flanked by repetitive sequences. Nevertheless, primers were designed, and amplification was attempted. The primers for Ref_Chr7_q21.3 produced an amplicon of the expected size, but subsequent cloning and Sanger sequencing of this amplicon failed (**Figure 19**). The PCR for Ref_Chr1_p31.1_a produced a high number of non-specific amplification products (data not shown).

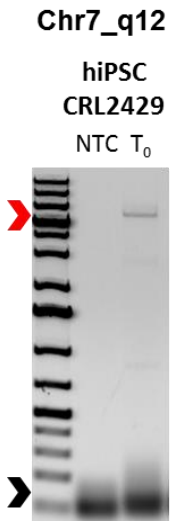


Figure 19: **Empty/filled PCR validation of Ref_Chr7_q21.3 element.** Amplification was observed in T_0 in the cell line hiPSC-CRL2429 in the filled site (red arrow) of the empty/filled PCR assay and in the empty site (black arrow). NTC, non-template control.

2.1.2: Identification of transduction family members absent from the reference genome.

Once reference genome members of L1_$_{1\text{-}14}$ were identified, I sought to find additional polymorphic family members that were absent from the reference genome. Analysis of the RC-seq data produced by Klawitter *et al.* [1] for cell lines hiPSC_CRL2429, hiPSC_CRL1502, and others, showed 5 candidate L1 insertions that carried the 3' transduced sequence shared by the transduction family. Three of these elements were amplified by PCR and Sanger sequenced from fibroblasts corresponding to cell line hiPSC-CRL2429. The elements were: the donor L1 (Non-ref), Non-ref_Chr3_p24.3 and Non-ref_Chr3_p12.2_b and their genomic locations were: Chr3:38,626,082, Chr3:20,748,904, Chr3:82,144,869, respectively. As mentioned in Chapter 1, the donor L1 was previously described in Beck *et al*., [11] and was identified to be the immediate progenitor of the *de novo* L1 insertion. The activity of the donor L1 and Non-ref_Chr3_p24.3 was further assessed in a cultured cell retrotransposition assay (see **Results section 2.5**). PCR amplification was also attempted for the other two potential full-length L1s, Non-ref_Chr3_p12.2_a and Non-ref_Chr17_q12, with genomic locations Chr3:80,590,176 and Chr17:32,813,609, respectively, using as a template genomic DNA from additional cell lines in which they were detected by Klawitter *et al.* [1] by at least one sequencing read. PCR validation was not immediately successful for these elements. As they carried the hallmarks of retrotransposition, as well as the 3' transduction shared amongst the L1_$_{1\text{-}14}$ family, I did not pursue further optimisation of the PCR conditions.

Identification of the donor L1 as the immediate progenitor of the *de novo* L1 insertion was possible due to the 10 nucleotides of 5' transduced sequence present adjacent to the *de novo* L1. This sequence fully matched the flanking genomic sequence of the polymorphic element previously identified by Beck *et al.* on Chromosome 3 [11]. Hence, it was possible to determine that the remaining family members lacked this unique upstream region. PCR validation against the donor L1 demonstrated that this element was present in the parental fibroblasts of the cell line hiPSC-CRL2429, where the *de novo* insertion was discovered. The empty/filled PCR validation assay in hiPSC_CRL2429 revealed a filled site of around ~6 Kb in length and amplification of the expected empty site band of 130 nt, confirming the heterozygosity of the element (**Figure 20**). This polymorphic element was previously characterised by Beck *et al.* [11] with nucleotide reference number AC211854. It was described to be an active element showing 101% of the *in vitro* activity of L1.3 [11].

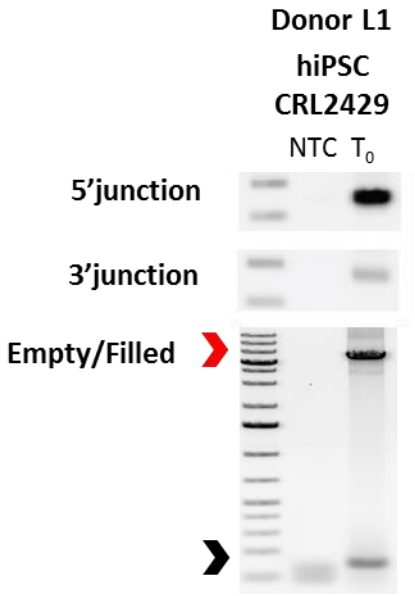


Figure 20: PCR validation panel of the donor L1. 5' junction PCR, 3' junction PCR and Empty/Filled PCR gels. Amplification was observed in T_0 in the cell line hiPSC-CRL2429 in the 5' junction PCR, the 3' junction PCR, in the filled site (red arrow) and in the empty site (black arrow) of the empty/filled PCR assay. The presence of amplification product in the empty site confirmed the heterozygosity of this element. NTC, non-template control.

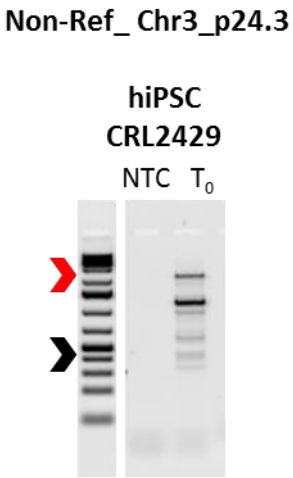


Figure 21: **Empty/filled PCR validation of Non-Ref_ Chr3_p24.3 element.** Amplification is observed in T_0 in the cell line hiPSC-CRL2429 in the filled site (red arrow) of the empty/filled PCR assay and in the empty site (black arrow). NTC, non-template control.

The polymorphic full-length L1 Non-ref_Chr3_p24.3 was also successfully PCR validated via an empty/filled PCR assay (**Figure 21**), presenting the expected amplification product in the filled site. The amplification product observed in the empty site, with 428 nucleotides, confirmed the heterozygosity of the element. This element was described in Beck *et al.* as being inactive, with nucleotide reference number: AC203662 [11]. The full-length L1 Non-ref_Chr3_p12.2_b [8, 86, 91, 113, 170, 176, 205] was not present in the cell

lines hiPSC-CRL2429 or hiPSC-CRL1502 and therefore it was PCR amplified using gDNA from Hela cells and an empty/filled PCR assay (**Figure 22**). The amplification product observed in the empty site with 386 nucleotides of fragment size confirmed the presence of an empty-site allele in HeLa cells and the L1 insertion was full-length.

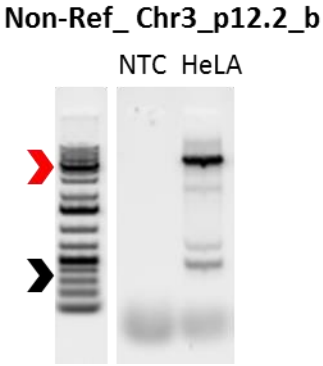


Figure 22: Empty/filled PCR validation of Non-Ref_Chr3_p12.2_b. Amplification is observed in Hela cells of the filled site (red arrow) from the empty/filled PCR assay and of the empty site (black arrow). NTC, non-template control.

According to the RC-seq data from Klawitter *et al.* paper [1] the other full-length element, named as Non-ref_Chr3_p12.2_a, was located at Chr3:80,590,176 and has also been described in other studies [8, 11, 87, 90, 91, 113, 176]. The final family member identified from the data of Klawitter *et al.* was named as Non-ref_Chr17_q12 [1, 86, 90, 91, 112, 171, 176, 204], was located at chromosomal location Chr17:32,813,609 and, according to sequencing reads obtain from 2 different studies [1, 176], this element was 5' truncated and inverted/deleted (**Figure 26**).

2.2: Identification of transduction family members absent from the reference genome and detected by other studies.

The L1_$_{1\text{-}14}$ transduction family was further elucidated by directly aligning the 3' transduced sequence carried by the *de novo* L1 to previous WGS and RC-seq reads generated by the Faulkner laboratory [1, 6, 7, 91, 143, 173]. Three additional polymorphic family members were found. According to the sequencing data, one was a full-length L1 and two were 5' truncated. The full-length element was previously described [8, 170] at the chromosomal location ChrX:38,097,551 and was named here as Non-ref_ChrX_p11.4. One non-reference 5' truncated element was named Non-ref_Chr1_p22.2 [8, 86, 91, 113, 170, 205] located in Chr1:90,914,512. The other non-reference truncated element was named

Non-ref_Chr4_q12 [8, 86, 90, 91, 113, 170, 176] and was located at Chr4:53,628,490.

Interestingly, two other L1 insertions were previously described by Tubio *et al.* [8] in an analysis of tumour genomes as “solo” L1 insertions, which could not be partnered with a matching source L1. These two L1 insertions each carried a relatively short, 18 nucleotide (GAATTGTAAAAAAAAAAA) 3’ transduced sequence that matches the lineage progenitor L1. The insertions have the identifiers PD7355f and PD7355a in Tubio *et al.* and they were in the chromosomal locations: Chr10:110,250,765 and Chr3:70,096,455, respectively. According to the Tubio *et al.* analysis, these elements were 5’ truncated. By characterising the L1_$_{1\text{-}14}$ transduction family, the origins of these two annotated solo L1 insertions are now clear, showing that the L1_$_{1\text{-}14}$ lineage is active in cancer.

2.3: Characterisation of an extended transduction family.

Altogether, the transduction family L1_$_{1\text{-}14}$ accounted for 14 elements, as well as two related instances carrying 18 nucleotides of the common 3’ transduced sequence, considered to be separate as they were identified only in tumour samples [8]. All of these corresponded to the L1-Ta subfamily [34]. Nine of the 14 were full-length L1s and 5 were 5’ truncated. Five were present in the reference genome and 9 were non-reference L1s (**Table 4A**). The length of the 5’ transduced sequences varied between 10 and 539 nucleotides. Ref_Chr7_q21.3 and Non-ref_Chr3_12.2_a presented the smallest first Poly-A tract length, with 16 adenines, whilst Non-ref_Chr17_q12 had the longest, with 46 adenines. The observed variation in L1 Poly-A tract length is thought to be caused by non-random degradation during DNA replication [206]. The length of the 3’ transduced sequence varied between 19 and 75 nucleotides amongst all the family members, and the length of the second Poly-A tract sequences ranged from 15 adenines in the Non-ref_Chr1_p22.2 to 80 adenines in the Non-ref_Chr3_12.2_a. Overall, the second Poly-A tract was longer than the first Poly-A tract (**Table 4A**), in line with expectations [127].

Nomenclature	Family	Length (nt)	1st Poly-A	3' transd length (nt)	2nd Poly-A
Lineage progenitor L1	L1-Ta	6043	29		Absent
Donor L1	L1-Ta	6041	17	44	21
***De novo* L1 (TC_18)**	L1-Ta	6118	17	44	33
Ref_ Chr1_p31.1_a	L1-Ta	6046	29	75	29
Ref_ Chr7_q21.3	L1-Ta	6016	16	42	17
Non-ref_ Chr3_p24.3	L1-Ta	6042	22	27	22
Non-ref_ Chr3_p12.2_a	L1-Ta	6041	16	41	80
Non-ref_ Chr3_p12.2_b	L1-Ta	6041	43	35	25
Non-Ref_ChrX_p11.4	L1-Ta	6121	18	43	18
Ref_Chr1_p31.1_b	L1-Ta	1485	42	51	23
Ref_Chr9_p23	L1-Ta	2703	29	41	29
Non-ref_Chr17_q12	L1-Ta	1011	46	19	21
Non-ref_Chr1_p22.2	L1-Ta	144	28	40	15
Non-ref_Chr4_q12	L1-Ta	173	18	44	26

Nomenclature: Ref, reference L1 element; Non-Ref, Non-Reference L1 element; 3' transd, 3' transduction

Table 4: A. Features of transduction family L1_1-14. In the table it is shown: Location; family; nucleotide length; first Poly-A tract length, 1st PolyA; 3' transduction length, 3' transduction length (nt); second Poly-A tail length, 2nd Poly-A.

Surveying the literature for *in vitro* estimates of L1 retrotransposition efficiency for full-length members of the transduction family revealed measurements for the lineage progenitor, donor L1, *de novo* L1 (same nucleotide sequence as the donor L1) and Non-ref_Chr3_p24.3. Non-ref_Chr3_p24.3 was reported to be inactive by Beck *et al* [11]. In the same study, the donor L1 was found to present 101% the activity of L1.3 in the retrotransposition assay. The lineage progenitor L1 was described to be virtually inactive, with 1.2% of the activity of the hot element $L1_{RP}$ [13] (**Table 4B**). Most of the elements were inserted in intergenic regions (**Table 4C**), except the donor L1, which was located in an intron of the gene SCN5A, and the *de novo* L1 insertion, which was positioned in an intron of the transcript TSNASX-DISC1.

Nomenclature	Previously reported	ID previously reported	Activity previously described
Lineage progenitor L1	Reference genome and Brouha *et al.[13]*	AP000652	1.2% of L1RP
Donor L1 (Non-ref)	Beck et al. [11], Ewing et al 2010.[112], Ewing et al.2011.[113], Helman et al.[176], Kuhn et al.[204], Lee et al.[86], Shulka et al.[91], Stewart et al.[205], Sudmant et al.[170], Tubio et al.[8], Wang et al.[171] and Iskow et al.[90]	AC211854	101% of L1.3
***De novo* L1 (TC_18)**	This study		
Ref_ Chr1_p31.1_a	Reference genome		
Ref_ Chr7_q21.3	Reference genome		
Non-ref_ Chr3_p24.3	Beck et al.[11], Ewing et al.2011. [113], Helman et al. [176], Lee et al. [86], Shukla et al. [91], Tubio et al. [8], Wang et al. [171] and Iskow et al. [90]	AC203662	Inactive
Non-ref_ Chr3_p12.2_a	Gardner et al.[87], Tubio et al. [8], Beck et al. [11], Ewing et al.2011[113], Helman et al. [176], Shukla et al. [91] and Iskow et al. [90]		
Non-ref_ Chr3_p12.2_b	Ewing et al. 2011[113], Tubio et al. [8], Helman et al. [176], Shukla et al. [91], Stewart et al. [205], Sudmant et al. [170] and Lee et al. [86]		
Non-ref_ChrX_p11.4	Tubio et al. [8] and Sudmant et al. [170]		
Ref_Chr1_p31.1_b	Reference genome		
Ref_Chr9_p23	Reference genome		
Non-ref_Chr17_q12	Iskow et al. [90], Ewing et al 2010. [112], Helman et al. [176], Kuhn et al.[204], Lee et al. [86], Shulka et al. [91] and Wang et al. [171]		
Trunc-Non-ref_Chr1_p22.2	Tubio et al. [8], Ewing et al.2011. [113], Lee et al. [86], Shulka et al. [91], Stewart et al. [205] and Sudmant et al. [170]		
Non-ref_Chr4_q12	Tubio et al. [8], Ewing et al.2011. [113], Helman et al. [176], Lee et al. [86], Shulka et al. [91], Sudmant et al. [170] and Iskow et al. [90]		

Nomenclature: Ref, reference L1 element; Non-Ref, Non-Reference L1 element.

Table 4: B. Features of transduction family L1$_{1\text{-}14}$..

Nomenclature	TSD	Endo-Motif (5'-3')	Gene Region
Lineage progenitor L1	AAAGAATTGTA	TT/AAAG	Intergenic
Donor L1 (Non-ref)	AGAATGAGTAAATAATG	AC/AGAA	SCN5A (intron)
***De novo* L1 (TC_18)**	AAAGAAATGACATCTG	TT/AAAG	TSNAX-DISC1 (intron;transcript variant)
Ref_ Chr1_p31.1_a	AGAAAAACAAATCA	AT/AGAA	Intergenic
Ref_ Chr7_q21.3	GAAAGTTCCAGTTGC	AT/GAAA	intergenic
Non-ref_ Chr3_p24.3	TAAAGACAC	GT/TAAA	Intergenic
Non-ref_ Chr3_p12.2_a	GAAAATGGAATGGG	AT/GAAA	Intergenic
Non-ref_ Chr3_p12.2_b	AGAAATAATAATTTCC	TT/AGAAA	Intergenic
Non-ref_ChrX_p11.4	AAAAGCGATATG	AT/AAAA	Intergenic
Ref_Chr1_p31.1_b	AAAAAAAATGGTTCATGC	TT/AAAA	Intergenic
Ref_Chr9_p23	GAAAAGTATTGTATTG	AA/GAAA	Intergenic
Non-ref_Chr17_q12	AAGAAGGTAAGATGG	TT/AAGA	Intergenic
Non-ref_Chr1_p22.2	AAAAAGCTCTTTCAG	TC/AAAA	Intergenic
Non-ref_Chr4_q12	TAAATTACAGGTTA	TT/TAAA	Intergenic

Nomenclature: Ref, reference L1 element; Non-Ref, Non-Reference L1 element; Trunc, Truncated L1 element

Table 4. C. Sequence features of L1_1-14. transduction family members. Target site duplications, TSDs; Endonuclease motif, Endo-Motif (5'-3').

2.3.1: Sequence analysis of reference elements.

The internal sequences of the reference element members of L1_1-14 (lineage progenitor, Ref_Chr7_q21.3, Ref_Chr1_p31.1_a, Ref_Chr1_p31.1_b and Ref_Chr9_p23) were obtained from the UCSC genome browser after BLAT searching [207] the 3' transduced sequence. Then, L1 sequences upstream of the transduced sequence were identified. The element Ref_Chr7_q21.3 contained a Poly-A tract of 16 nucleotides and the second Poly-A was 17 nucleotides in length, bracketing a 3' transduced sequence of 42 nucleotides. The element Ref_Chr1_p31.1_a presented a first and a second Poly-A tract of 29 nucleotides in each case. This element had also a 5' transduced sequence 539 nucleotides in length (**Figure 26**). This 5' transduced sequence belonged to the genomic flanking DNA upstream of the lineage progenitor L1. A single untemplated guanine preceded this 5' transduction upstream of the Ref_Chr1_p31.1_a element, indicating capping of the

mRNA template, and used a transcription start site in the 5' long terminal repeat (LTR7Y) sequence corresponding to a HERV-H provirus integrated ~126kb upstream of the lineage progenitor L1. Interestingly, this 5' transduction contained 3 exons (176, 356 and 7 nucleotides) resulting from 2 RNA splicing events. The element Ref_Chr1_p31.1_b was 5' truncated, and was 1485 nucleotides in length. The element Ref_Chr1_p31.1_b presented a first Poly-A tract of 42 nucleotides and a second Poly-A tract of 23 nucleotides, and carried a 3' transduced sequence of 51 nucleotides. The element Ref_Chr9_p23 was truncated at L1 position 4073, with an inversion of 754 nucleotides. This 5' truncated element was therefore 2703 nucleotides in length, with first and second Poly-A tracts of 29 nucleotides, and a 41 nucleotide 3' transduction.

2.3.2: Sequence characterisation of the lineage progenitor and donor L1.

After PCR amplification, the filled site products of the lineage progenitor and donor L1s were cloned and Sanger sequenced, as previously described for the de *novo* L1 and the other polymorphic insertions such as Non-ref_Chr3_p24_24.3 and Non-ref_Chr3_p12.2_b. This allowed verification of their genomic location and their structural features (**Figure 23 and 24**).

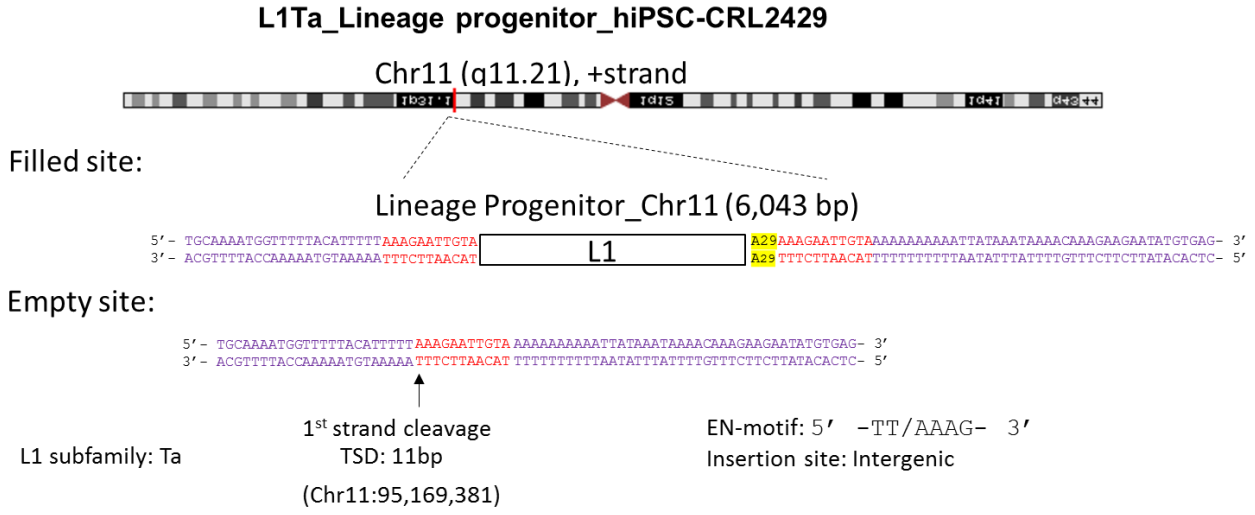


Figure 23: **Genomic location and structural features of the lineage progenitor L1. Filled site.** This figure shows a full-length L1 from the family L1-Ta [34, 207]. The insertion is 6043bp in length and is upstream of the characteristic genomic sequence (purple letters) which is found in the 3' transductions of the rest of the family members. Insertion site coordinates Chr11:95,169,381. The TSDs, red letters, flank the L1 sequence and after the L1 sequence is a Poly-A tract of 29 nucleotides in length (bright yellow). **Empty site.** Red letters, TSD; Purple letters, gDNA from the Chromosome; Black arrow, site of first strand cleavage by the L1 endonuclease.

Comprehensive sequence analysis of the lineage progenitor L1 indicated a full-length L1 of 6043 nucleotides (**Figure 23**) that did not carry any transduced sequence and presented a single Poly-A tract of 29 nucleotides. The TSDs of the insertion were 11nt and the EN-motif was 5' TT/AAAG 3'. The insertion was situated in the positive strand of chromosome 11, chromosomal location: Chr11:95,169,381. This insertion was in an intergenic region. Sequence analysis of the donor L1, which bore the *de novo* L1 insertion, indicated a full-length element of 6041 nucleotides (**Figure 24**) with a transduced sequence of 44 nucleotides and two Poly-A tracts of 17 and 21 nucleotides, respectively. The TSDs of the insertion were 17 nucleotides long and the EN-motif (AC/AGAA) only loosely matched the canonical endonuclease motif: The insertion was not located on the reference genome [11] and was situated on the sense strand of chromosome 3, chromosomal location: Chr3:38,626,082, within an intron of the gene SCN5A. The protein encoded by this gene is an integral membrane protein and tetrodotoxin-resistant voltage-gated sodium channel subunit. This protein is found primarily in cardiac muscle and is responsible for the initial upstroke of the action potential in an electrocardiogram.

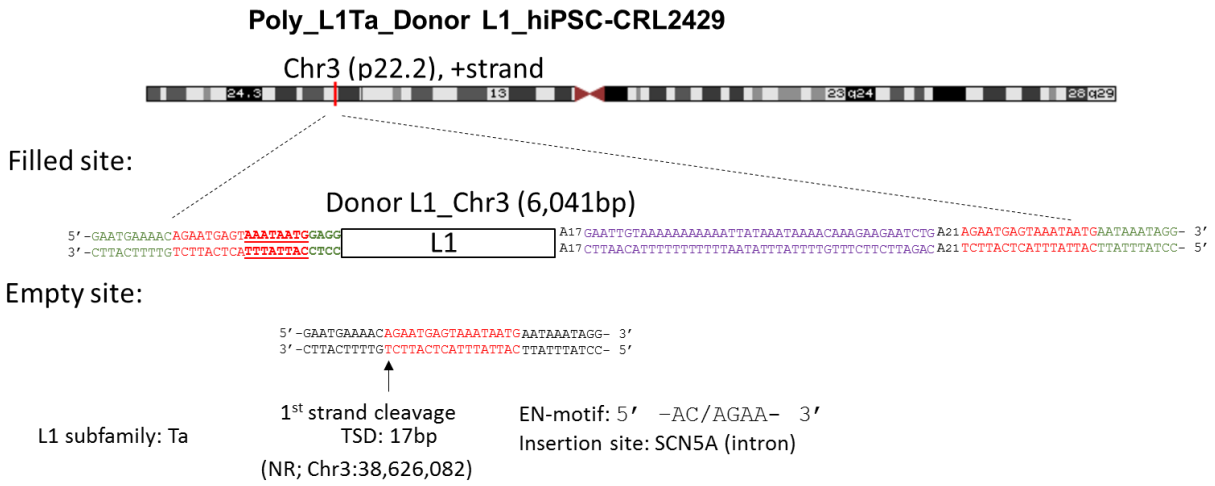


Figure 24: Genomic location and structural features of the donor L1. Filled site. This figure shows a full-length L1 (from the family L1-Ta [34]). The insertion is 6041 nucleotides in length and contains a 3' transduced sequence (purple letters) of 44 nucleotides and genome flanking the L1 copy (green letters). Highlighted in green, immediately upstream of the element, is 10 nucleotides of genomic sequence that was incorporated into a 5' transduction found on the *de novo* L1 insertion, from these 10 nucleotides; the 4 nucleotides flanking the L1 copy at the 5' end were untemplated nucleotides and the rest were part of the TSD situated upstream of the L1 element (bold letters). Red letters indicate TSDs flanking the L1 sequence. Following the L1 sequence is the first Poly-A tract of 17 nucleotides in length and a second Poly-A tract of 21 nucleotides. **Empty site.** Red letters indicate TSDs; Green letters indicate gDNA from the Chromosome; Black arrow indicates the site of first strand cleavage by the L1 endonuclease.

2.3.3: Sequence analysis of full-length transduction family members.

To fully characterise the sequences of the transduction family members, I performed independent PCR reactions using gDNA extracted from the cell line hiPSC-CRL2429 where the *de novo* L1 insertion was discovered, followed by Sanger sequencing, to discard PCR-induced mutations and distinguish possible allelic variants from elements that were determined to be homozygous due to lack of an empty-site amplicon. I obtained a consensus sequence for each L1. To reconstruct the consensus sequence from each element, avoiding PCR induced mutations from individual clones, non-mutated nucleotide sequence fragments were chosen to reconstruct the original consensus for each element. The resulting clones were verified by Sanger sequencing using 12 different primers covering the entire L1 sequence (**Figure 25**).

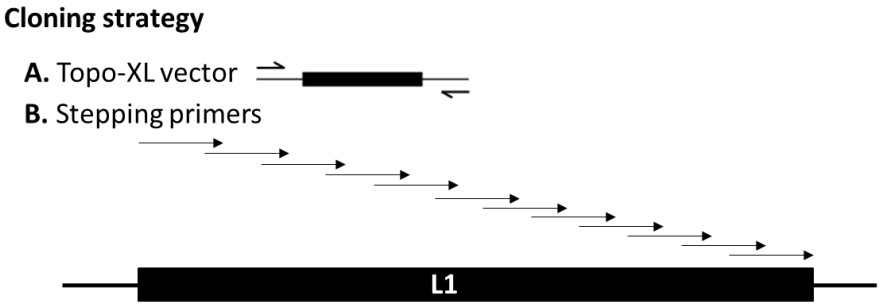


Figure 25: Cloning strategy diagram. A. Representation of the Topo-XL vector with an inserted L1 element. **B.** Representation of the overlapping sequence reads generated from the 12 stepping primers designed across the L1 sequence.

Once I had obtained the entire sequence of the desired L1, I manually assembled each of the 12 independent Sanger sequencing reactions of ~500 nucleotides in length, which overlapped their neighbouring sequencing reads, excluding the first and the last sequencing reactions that only overlapped with the next and with the previous sequencing reaction, respectively, to create a consensus ~6 Kb of each L1. Then I aligned the consensus L1 sequences to one another doing a multiple sequence alignment using Clustal omega software (https://www.ebi.ac.uk/Tools/msa/clustalo/). Additionally, to unambiguously determine their retrotransposition activity, the exact sequence of each element was rebuilt by strategic restriction digest and reassembly of the PCR amplification products. In this way, a clone of each element that was identical to the consensus for that element was generated. Each element was reconstructed in the retrotransposition indicator backbone prior to *in vitro* retrotransposition activity assessment.

2.4: Comparative nucleotide and amino acid sequence analysis of the transduction family L1 [1-14].

I next compared the transduction lengths and Poly-A tract lengths, as well as internal L1 sequences where available, of the 14 transduction family members. For the *de novo*, lineage progenitor, and donor L1s, as well as Non-ref_Chr3_p24.3, I obtained accurate sequences for the alleles present in hiPSC-CRL2429 as described above (https://www.ebi.ac.uk/Tools/msa/clustalo/) (please refer to material and methods for more details). I also analysed the reference nucleotide sequence of 2 full-length members of the family (Ref_Chr1_p31.1_a and Ref_Chr7_p21.3) found in the reference genome that, because of their genomic location in repetitive sequences, could not be PCR amplified and cloned, as described above. Two other truncated reference L1s carrying the distinguishing 3' transduced sequence were identified: Ref_Chr1_p31.1_b and Ref_Chr9_p23. In addition, three full-length (Non-ref_Chr3_p12.2a, Non-ref_Chr3_p12.2b and Non-ref_ChrX_p11.4) and three truncated non-reference elements (Non-ref_Chr17_q12, Non-ref_Chr1_p22.2 and Non-ref_Chr4_q12), were included, although the internal sequences of these elements were not available for comparison. Where available, the nucleotide and amino acid changes found among the family members were annotated relative to the previously-described "hot" L1.3 element, relationships among family members inferred from 3' transduction length, and the presence of 5' transductions, as indicated in **Figure 26**. The features of these insertions are also summarised in **Table 4**.

Notably, two allelic variants of the lineage progenitor L1 carried by the homozygous individual CRL2429 were identified by the presence of 4 distinct nucleotides along the L1 sequence. Both allelic variants shared, with the rest of the family members, several mutations along the L1 sequence. There were 6 other mutations present in both alleles, in the 3'UTR and in the ORFs as indicated in **Figure 26.** Additionally, there were unique mutations carried only by one of the alleles. Allele 1 (Lineage progenitor Allele 1) contained 1 mutation in the 5' UTR and 2 mutations in ORF2 (Q159H and D523H), one of which (D523H) was located in the ORF2p RT domain (**Figure 26**). Allele 2 (Lineage progenitor Allele 2) contained a unique silent mutation not shared by the rest of the family members, located in the ORF2 sequence (**Figure 26**). I also observed that the donor L1 and the *de novo* L1 had identical nucleotide sequences. The element Non-ref_Chr3_p24.3 presented a stop codon at the amino acid position 559 and was hence predicted to lack retrotransposition capability. The *de novo* L1 did not present deleterious mutations in

conserved functional residues, indicating that this insertion (and its donor L1) could potentially retain retrotransposition competence.

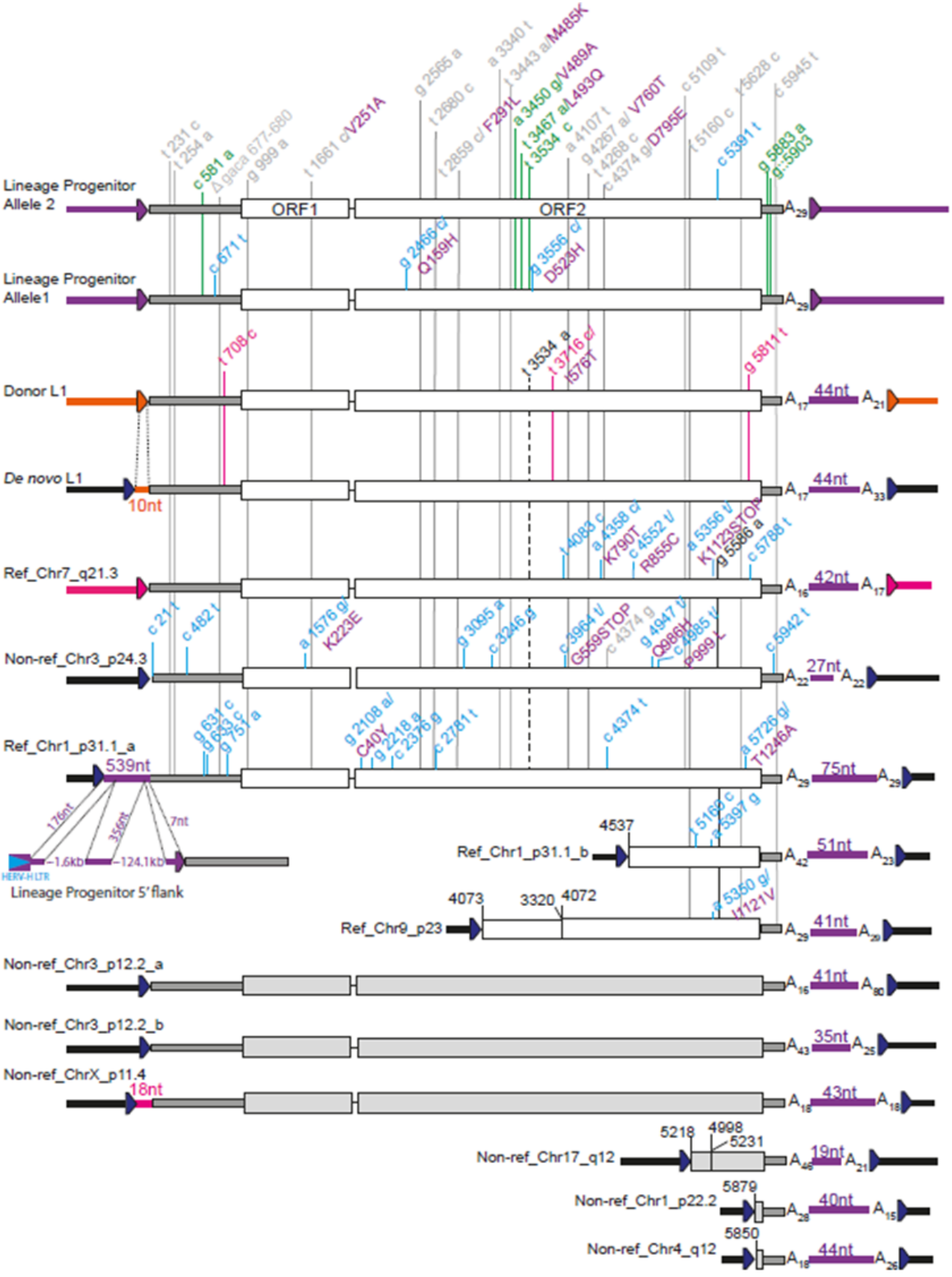


Figure 26: Schematic representation of transduction family L1_1-14 members. The nucleotide position relative to the sequence of L1.3 is indicated with numbers on each L1 element; changes present in all the different family members (grey), changes unique to one element (blue), changes present in donor L1 and *de novo* L1 (pink), changes present in some elements but not in all of the family members (black dashed lines). Specific amino acid changes relative to the L1.3 sequence (purple letters, next to each nucleotide change). Stop codon mutations (STOP). TSDs (dark blue triangles flanking the L1 copies); 5'UTR and 3'UTR (left and right grey boxes respectively); ORF1 and ORF2 (white boxes for sequences which are known and grey boxed for polymorphic L1s with unknown sequences); First and second PolyA (An); 3' transduced sequences with gDNA from Lineage progenitor element (purple line) with different nucleotide lengths depending on its case (purple line); 5' transduced sequences in the *de novo* L1 element (orange line, 10 nucleotides), Ref_Chr1_p31.1_a (purple, 539 nucleotides) and Non-ref_ChrX_p11.4 (pink line; 18 nt) with gDNA from each immediate donor element. The donor elements that originated the elements presenting the mentioned 5' transduced sequences were Donor L1 (orange line), Ref_Chr7_q21.3 (pink line, represents the gDNA that was transduced at the 5'genomic flanking DNA) and lineage progenitor (purple line, represents the gDNA that was transduced at the 5'genomic flanking DNA), the elements that were originated from these parental elements were: *De novo* L1, Non-ref_ChrX_p11.4 and Ref_Chr1_p31.1_a respectively.

2.5: Analysis of retrotransposition activity in cultured cells.

To evaluate the mobility and ORF1p and ORF2p biochemical activity of several members of the L1_1-14 transduction family, I used an established cell culture-based reporter assay [22]. Briefly, in this retrotransposition assay an L1 of interest is cloned into a suitable vectors for retrotransposition that contains an L1 driven by a promoter, in this case, its native internal promoter (**Figure 27**), and a reporter gene encoding antibiotic resistance that is in antisense orientation with respect to the L1 copy and is interrupted by an intron, in this case, neomycin phosphotransferase (NEO). In this reporter cassette, the intron is sense orientated with respect to the L1 copy and the reporter cassette contains its own polyadenylation signal. Hence, the reporter cassette is active and expresses antibiotic resistance only after splicing and retrotransposition, meaning that each new colony under antibiotic selection represents the product of a novel retrotransposition event.

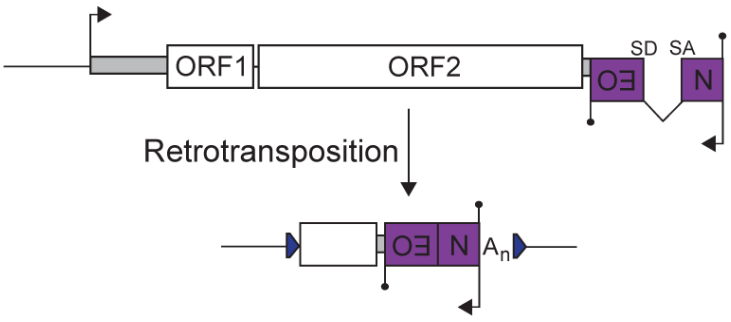


Figure 27: **Retrotransposition reporter cassette schematic.** Representation of the unspliced retrotransposition reporter cassette, neomycin phosphotransferase gene or G418 (purple box, named NEO), after the 3'UTR (right grey box) and spliced cassette activated upon retrotransposition, plus a polyadenylation signal [22]. The cassette is in antisense orientation relative to the L1 and carries its own polyadenylation signal, and is interrupted by an intron with splice donor and acceptor sites in sense orientation relative to the L1.

Hela cells are well known in the L1 field to support endogenous and engineered L1 retrotransposition [7, 22], hence I used Hela cells here to test the activity of the L1s. The RTSN assay rationale is represented below (**Figure 28**).

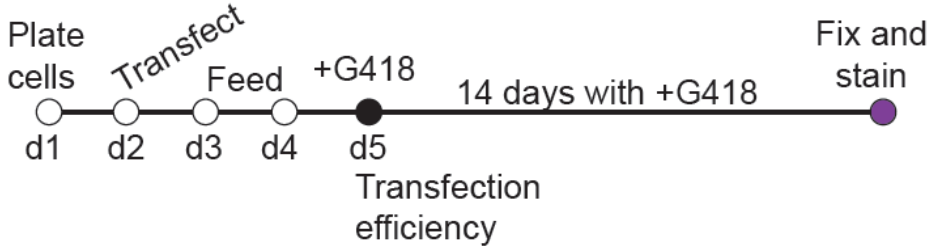


Figure 28: **RTSN assay timeline.** Circles represent relevant days during the performance of the assay. Day 1 (d1), plate cells; day 2 (d2), transfection; day 3 (d3) and day 4 (d4) feed with media, day 5 (d5), start antibiotic selection (neomycin, +G418); days 6 to day 17, antibiotic added in media every other day until day 18 when the cells were fixed and stained for colony counting.

The elements tested in this retrotransposition assay included: L1.3 as a positive control [28], L1.3 RT- as a negative control (reverse transcriptase mutant) [53], the donor L1, which was the parental element of the *de novo* L1 (bearing the same nucleotide sequence as the donor L1, and therefore not cloned and tested independently), the two different alleles of the lineage progenitor L1, and Non-Ref_Chr3, which was discovered in the RC-seq output data table of Klawitter *et al.* [1] and carried a predicted disabling ORF2 mutation. The retrotransposition activity of the elements is relative to L1.3. We observed that the lineage progenitor Allele 2 construct exhibited the highest frequency of retrotransposition activity. Lineage progenitor Allele 2 exhibited higher retrotransposition activity than Lineage progenitor Allele 1 from the same locus but less activity than the donor L1, and hence the *de novo* L1. The donor L1 was also more active than L1.3. Compared to the positive control L1.3, the lineage progenitor Allele 1 presented less activity. Non-Ref_Chr3 did not show any L1 activity, as expected, given its mutated amino acid sequence (**Figure 29**). Hence, these results prove the retrotransposition capability of *de novo* L1, donor L1, and 2 different alleles of the lineage progenitor, which gave rise to the entire family.

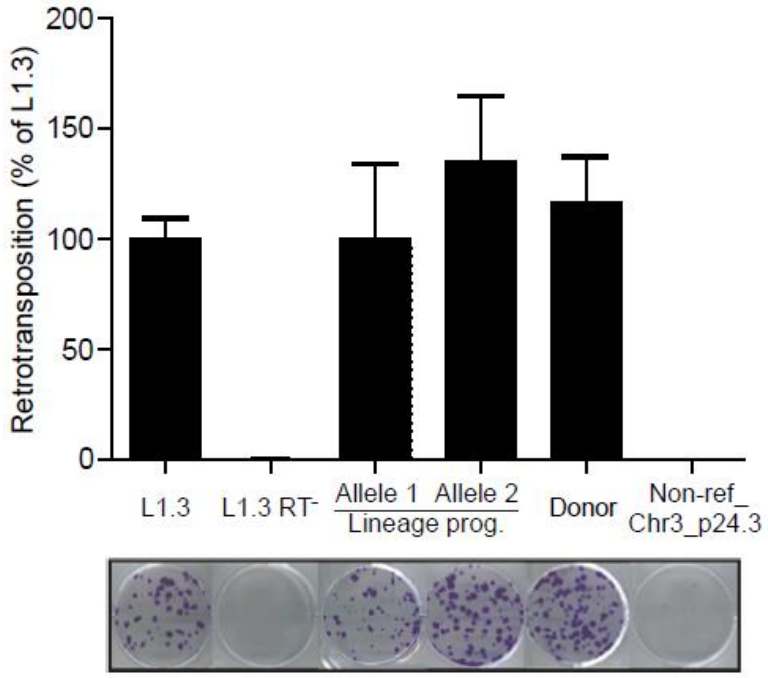


Figure 29: RTSN cultured-cell assay shows frequent mobilisation of L1s driven by their native promoter. Below the histograms: RTSN assay 6 well plate pictures for each element. Positive control: L1.3 wild-type human L1 reference element; Negative control: RT- reference human L1 (L1.3 mutated) element deficient at the reverse transcriptase domain (RT-). This assay was repeated 3 times (biological replicates) with similar results. Colony number values in each construct were normalized to wild-type L1, L1.3 reference element. Values represent the mean ± SD of colonies counted in 3 biological replicates with three technical replicates, n=9.

CONCLUSIONS OF CHAPTER 2:

In this chapter, I described an extended L1 transduction family. Although most of the members of this family were found by previous publications, this is the first time that the extent of the family has been established, with the lineage progenitor identified as its source element. Here, 14 different members of the transduction family have been identified, and the internal sequences of 8 members were thoroughly characterised. Retrotransposition competence in a cultured cell assay was observed for both alleles of the lineage progenitor L1, in the donor L1 and hence in the identical *de novo* L1 element. Therefore, this family contains members of which are still active in the human population, and can generate reprogramming-associated insertions in hiPSCs, such as the case of the *de novo* L1 insertion, which is full-length and still capable of mobilisation.

CHAPTER 3: DYNAMIC METHYLATION OF THE TRANSDUCTION FAMILY AND *DE NOVO* L1 INSERTION DURING NEURODIFFERENTIATION.

To achieve mobilisation, a given L1 first needs to be transcribed from its 5'UTR sense promoter. In turn, the host genome silences L1 by methylation of the 5'UTR, reducing or eliminating L1 mRNA production and therefore lessening the probability of retrotransposition [3]. Previous studies have investigated the overall methylation status of the L1-Ta family in various cellular contexts, including NSCs, hESCs and hiPSCs [4, 5]. Other more recent studies [6-9] have investigated the methylation status of particular L1 copies via locus-specific PCR and bisulfite sequencing. L1 has been reported as active in NSCs and mature neurons [4, 10, 89, 104, 136, 143]. However, the methylation status of a *de novo* L1 insertion and its donor element during neurodifferentiation has not been examined to date.

In this chapter, I investigate the methylation status of several members of the L1_$_{1\text{-}14}$ transduction family, including the *de novo* L1, its donor and the lineage progenitor L1, in fibroblasts, hiPSCs, and during neuronal differentiation. Additionally, I sought to determine whether the *de novo* L1 insertion was hypomethylated upon retrotransposition and if its donor was hypomethylated at the time point where the insertion occurred.

3.1: L1 promoter methylation changes during neurodifferentiation.

I employed targeted bisulfite sequencing (**Figure 30**) to evaluate promoter methylation of the *de novo*, donor and lineage progenitor L1s in the hiPSC-CRL2429 and hiPSC-CRL1502 cell lines. Primers internal to the L1-Ta family 5'UTR were used to study the L1-Ta subfamily as a comparison. A schematic representation of the CpG dinucleotides within the L1 CpG island, as well as a representation of the bisulfite sequencing by locus specific PCR assay, are provided in **Figures 30 and 31.**

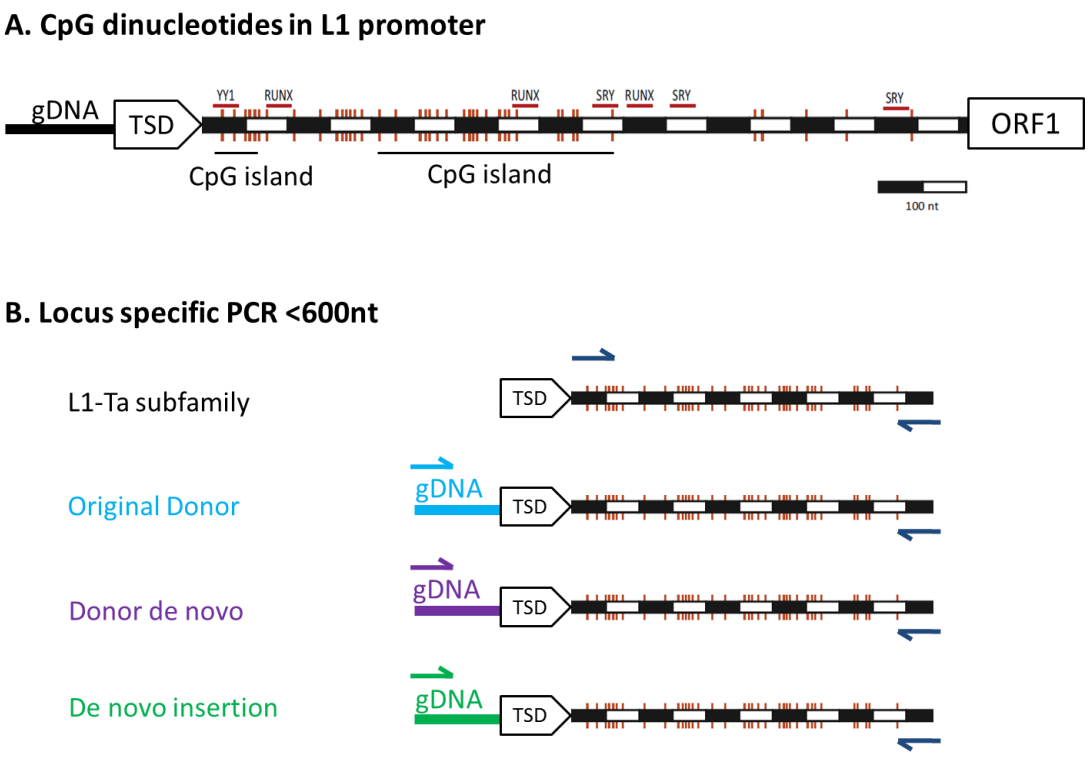


Figure 30: Schematic representation of CpG dinucleotides in the 5'UTR of an L1 element and bisulfite sequencing work flow followed in this chapter. A. CpG dinucleotide positions in the L1 promoter representation: gDNA, black line; TSD, white triangle on the left side; 2 CpG islands; different transcription factors that bind to the L1 promoter, YY1, RUNX and SRY; CpG dinucleotides, red vertical lines; ORF1, white box on the right side. **B**. Locus specific PCR. Specific primers and gDNA are represented in matching colours: overall L1 population, in this case the L1-Ta subfamily, dark blue primers aligning to the 5'UTR of the L1 copy; lineage progenitor, light blue; donor L1, purple; *de novo* L1, green. Elements of this figure were adapted from Faulkner & Garcia-Perez [127].

For technical reasons, the Illumina MiSeq platform cannot fully span fragments bigger than 600 nucleotides, as paired-end sequencing is limited to 2×300bp reads. To include genomic DNA upstream of each specific L1 copy, and therefore be able to unequivocally assign reads to that L1, meant a maximum of 35 CpG dinucleotides in the L1 5'UTR could be studied for each elements (**Figure 31**).

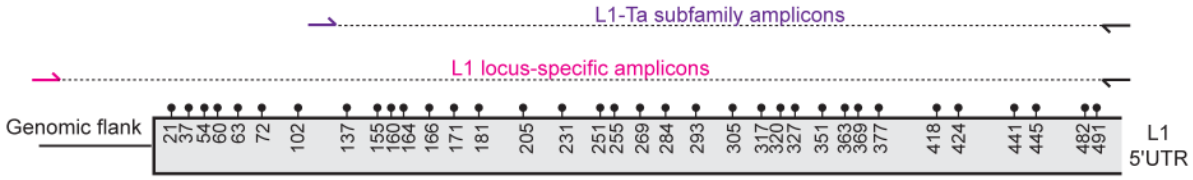


Figure 31: Schematic representation of the L1-specific bisulfite sequencing analysis. Black circles represent CpG nucleotide positions. Black arrows represent the reverse primer that was used in all the PCR amplification reactions. Pink arrow represents the location were the primers were designed to perform locus specific PCR with a maximum of 600 nucleotides

fragment size. Purple arrow represents the location were the primers were designed to perform the study of the overall L1-Ta population. Note: the L1-Ta subfamily amplicon covers 28 CpGs, whilst the L1-specific reactions cover 35 CpGs.

3.1.1: L1 promoter methylation assessment in the cell line hiPSC-CRL2429.

All the studied elements, including the L1-Ta subfamily, were less methylated in hiPSCs as compared to fibroblasts in hiPSC-CRL2429, as expected (**Table 5**, **Figure 32**). They then gradually became methylated during neurodifferentiation. The donor L1 exhibited lower levels of CpG methylation (31.1%) than the L1-Ta subfamily and lineage progenitor L1 (67.3% and 67.0%, respectively). This methylation profile was consistent with results obtained from hiPSC-CRL1502 (see below). Below is summarised the CpG methylation percentages obtained in each cell lines (**Table 5**).

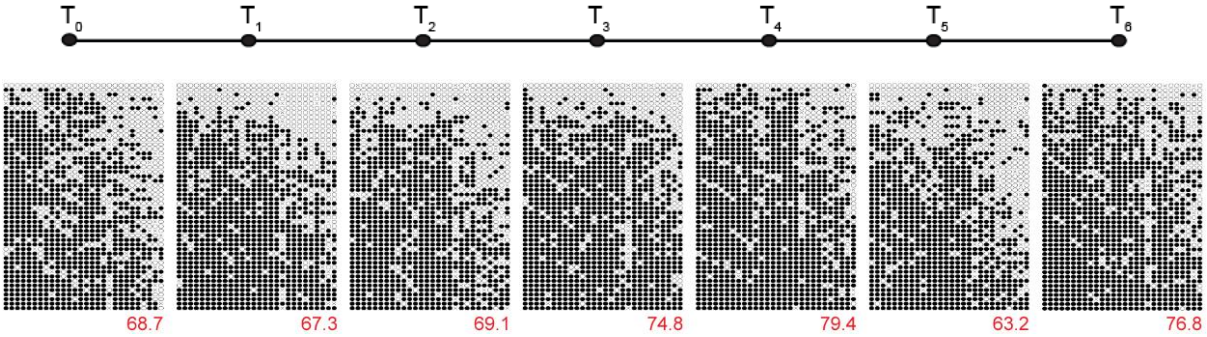
Cell line					
hiPSC-CRL2429	**Total CpG:**	**L1-Ta subfamily**	**Lineage progenitor**	**Donor L1**	***De novo* L1**
T_0	Fibroblasts (HDFs)	68.7	81.2	66.6	Not present
T_1	hiPSCs	67.3	67.0	31.1	*
T_2	Neural epithelial	69.1	82.2	45.3	40.2
T_3	Immature neurons and neural rosettes	74.8	84.4	53.8	34.4
T_4	Neurons I and neural rosettes	79.4	88.6	69.3	53.0
T_5	Neurons II and neural rosettes	63.2	65.8	39.6	29.3
T_6	Neurons III and neural rosettes	76.8	89.5	61.0	51.1
hiPSC-CRL1502	**Total CpG:**	**L1-Ta subfamily**	**Lineage progenitor**	**Donor L1**	***De novo* L1 Not present**
T_0	Fibroblasts (HDFs)	69.8	83.2	59.5	
T_1	hiPSCs	65.4	72.4	45.1	
T_2	Neural epithelial	69.7	75.9	36.9	
T_3	Immature neurons and neural rosettes	73.3	79.9	41.0	
T_4	Neurons I and neural rosettes	75.8	84.4	55.0	
T_5	Neurons II and neural rosettes	67.7	74.0	59.1	
T_6	Neurons III and neural rosettes	70.1	81.7	70.1	

Table 5: Summary of the total CpG methytlation percentages in hiPSC-CRL2429 and in hiPSC-CRL1502. T, time point. *; I did not recover enough distinct amplicons to reliably characterise this promoter.

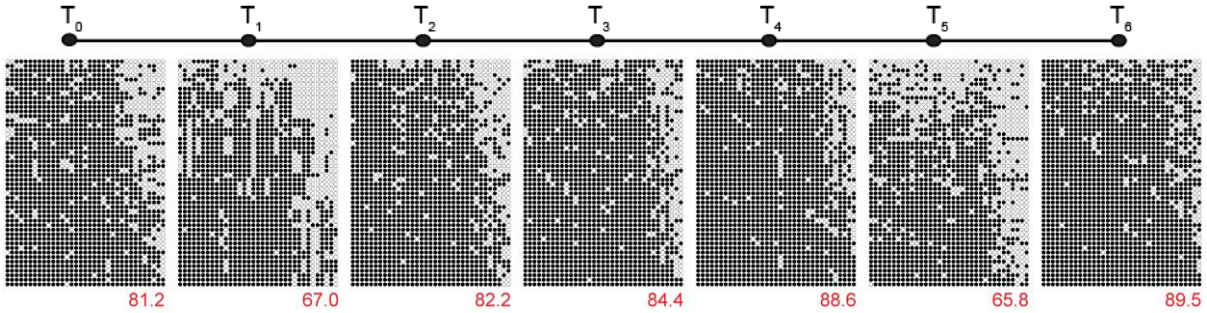
In hiPSC-CRL2429, and its derivatives, the *de novo* L1 exhibited the lowest methylation percentage when compared to the donor L1, lineage progenitor and to the L1-Ta subfamily across all time points where it could be detected (**Figure 32**). However, a sufficient cohort of distinct amplicons of the *de novo* L1 could not be generated in the hiPSCs, where it originated, to reliably characterise this promoter. The *de novo* was hypomethylated during early stages of differentiation, showing a minimum of 34.4% CpG dinucleotides methylated within the immature neurons and neural rosettes (T_3) (**Figure 32**). From T_2 to T_6, the *de novo* L1 was dynamically methylated, with values of 40.2% (T_2), 34.4% (T_3), 53.0% (T_4), 29.3% (T_5) and 51.1% (T_6). The drop at T_5 from T_4 was a general trend among all of the different promoters analysed and was statistically significant for each (paired t-test; $p<0.001$). I also observed significant re-methylation in T_6 (paired t-test; $p<0.01$) among all elements (**Figure 32 and 33**).

The donor L1 exhibited the lowest levels of CpG dinucleotide methylation in fibroblasts, 66.6%, compared to the overall L1-Ta subfamily, 68.7%, and to the lineage progenitor L1, 81.2% (**Table 5**). Indeed, the donor L1 promoter was fully unmethylated (CpGs in a row all represented as white circles) in numerous hiPSCs (T_1), the context where the *de novo* L1 insertion was most likely to have occurred **(Figures 32 and 34).** By contrast, the lineage progenitor L1 was more methylated than the donor and *de novo* L1 across all the different time points. Additionally, L1 promoter methylation observed amongst the overall L1-Ta subfamily were higher in all the different time points than the values obtained for the donor L1 or the *de novo* L1 (**Table 5**). One potential explanation for this trend is the presence of a FOX (forkhead box) protein, a class of pioneer transcription factor, binding site in the 10bp 5' transduction carried by the *de novo* L1 [208]. This binding site may have facilitated hypomethylation of the donor L1 and *de novo* L1, in comparison to the lineage progenitor L1, which did not carry the upstream FOX binding motif.

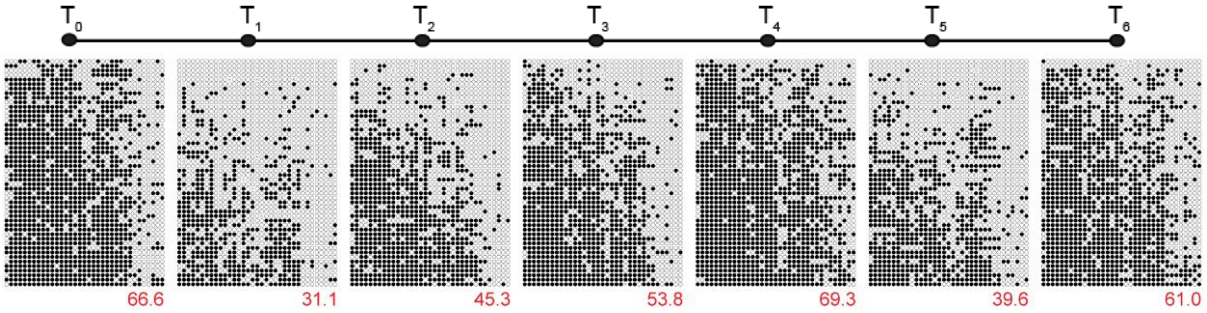
A. L1-Ta subfamily, hiPSC-CRL2429 cell line



B. Lineage progenitor, hiPSC-CRL2429 cell line



C. Donor L1, hiPSC-CRL2429 cell line



D. *De novo* L1, hiPSC-CRL2429 cell line

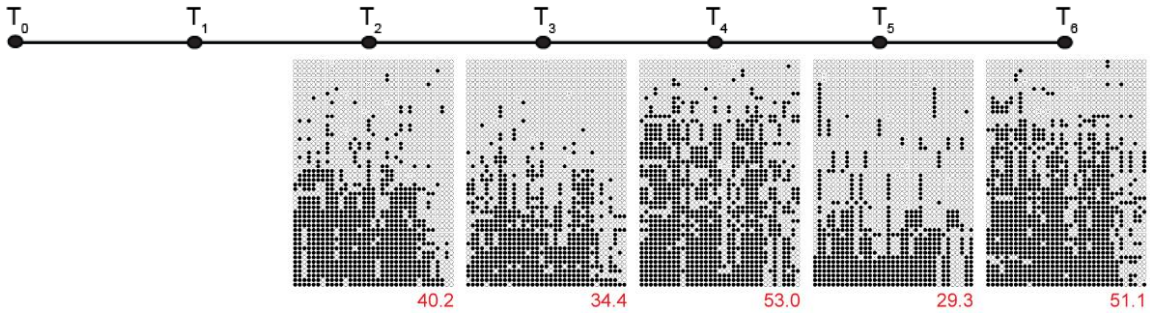


Figure 32: Methylation of L1_1-14 family member 5'UTRs in the hiPSC-CRL2429 cell line. Cartoons represent the methylation status of 50 randomly selected bisulfite sequencing reads comprising locus-specific PCR products at the 5' junctions of L1 sequences.

Transduction family members were analysed during hiPSC generation and neurodifferentiation in the hiPSC-CRL2429 cell line. Each column corresponds to a different time point (studied cell type, or time point, is indicated in the top of each column) (T_0= fibroblasts; T_1= hiPSCs; T_2= Neural epithelial; T_3= Immature neurons and neural rosettes; T_4= Neurons I and neural rosettes; T_5= Neurons II and neural rosettes; T_6= Neurons III and neural rosettes); each line contains the same set of L1 5'UTR CpG dinucleotides: **A.** L1-Ta subfamily population; **B.** Lineage progenitor; **C.** Donor L1; **D.** *de novo* L1 insertion. Red numbers below each panel indicates the CpG percentage in each time point for each studied 5'UTR.

The following graphs (**Figure 33**) summarise the methylation dynamics of each CpG dinucleotide in each targeted element from T_0 to T_6. Note the low levels of methylation of the *de novo* L1 during the early stages of differentiation (**Figure 33, C-D**) and its progressive methylation from neural epithelial (T_2) to mature neurons I, neural rosettes T_4 and the drop at T_5 to get methylated again in T_6 mature neurons III, neural rosettes. Also note that differences in methylation amongst the various elements are due to broad changes, where the vast majority of CpGs are differentially methylated, rather than just one or two (e.g. compare the donor and lineage progenitor L1s in **Figure 33B**).

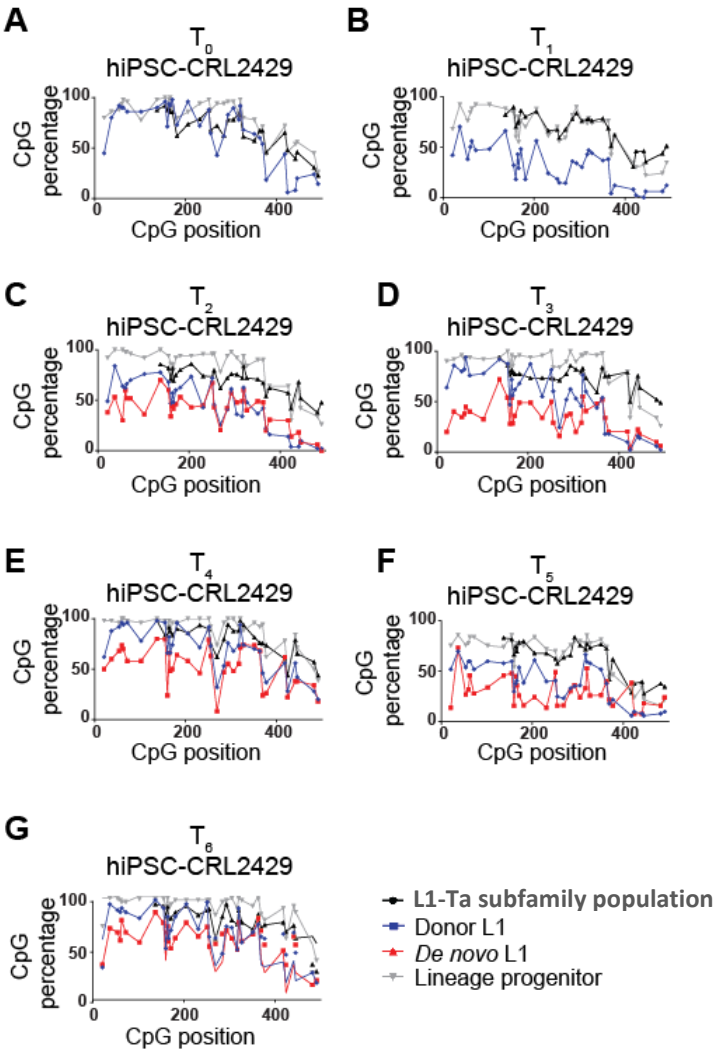
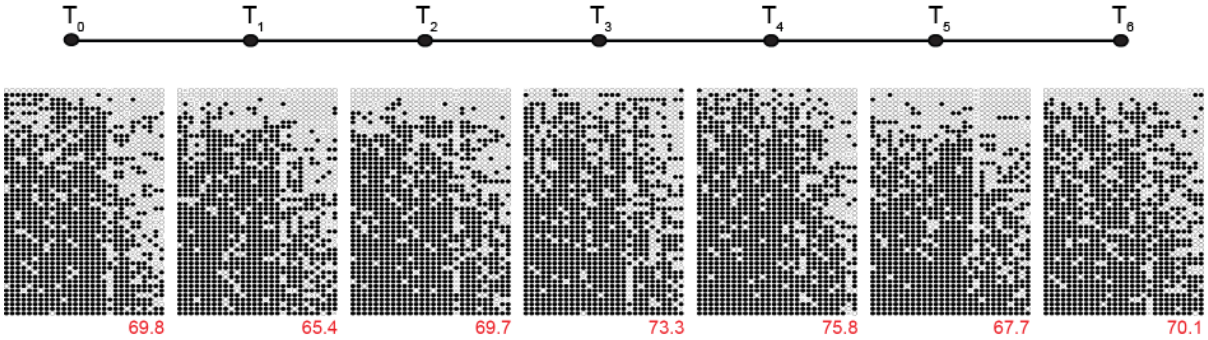


Figure 33: L1 promoter CpG methylation levels during neurodifferentiation time courses graphs I. A to **G** figures contain the methylation level (y-axis) in each time point the cell line hiPSC-CRL2429 along the L1 sequence (x-axis) of the different CpG dinucleotides. Graph legends: L1-Ta subfamily population (black lines); Donor L1 (blue lines); *de novo* L1 (red lines); Lineage progenitor (grey lines).

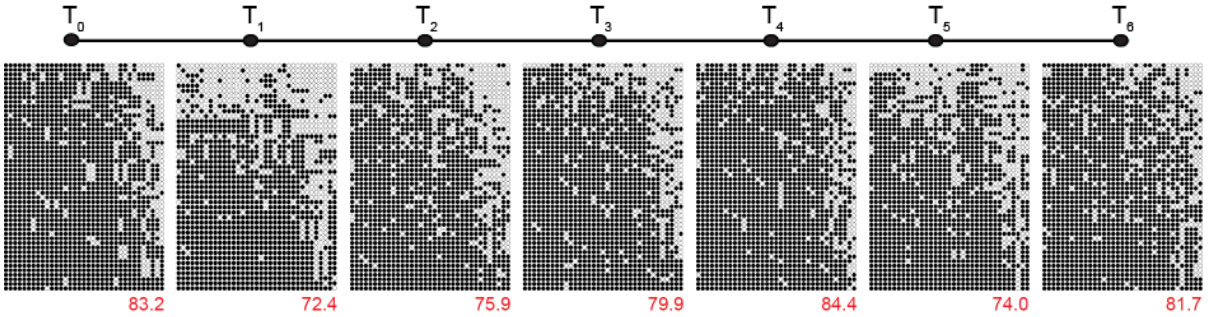
3.1.2: L1 promoter methylation assessment in the cell line hiPSC-CRL1502.

In the hiPSC-CRL1502 cell line, the donor L1, lineage progenitor L1 and L1-Ta subfamily population exhibited, in general, similar methylation patterns to those observed in the hiPSC-CRL2429 cell line (**Figure 34**). For instance, the donor L1 was less methylated in fibroblasts (59.5%) than in hiPSCs (45.1%). Overall, CpG dinucleotides in the donor L1 5'UTR were less methylated than those present in the lineage progenitor L1 or the broader L1-Ta subfamily (**Table 5**). In the case of the lineage progenitor L1, CpG methylation levels were higher than for the donor L1 at each time point and, again as for hiPSC-CRL2429, the L1-Ta subfamily was less methylated than the lineage progenitor L1, and more methylated than the donor L1 (**Table 5**). Again, across all samples tested, methylation levels were higher at T_0 than at T_1, and the overall methylation state of most L1 promoters decreased between T_4 to T_5 and then increased between T_5 to T_6.

A. Overall L1-Ta population, hiPSC-CRL1502 cell line



B. Lineage progenitor, hiPSC-CRL1502 cell line



C. Donor L1, hiPSC-CRL1502 cell line

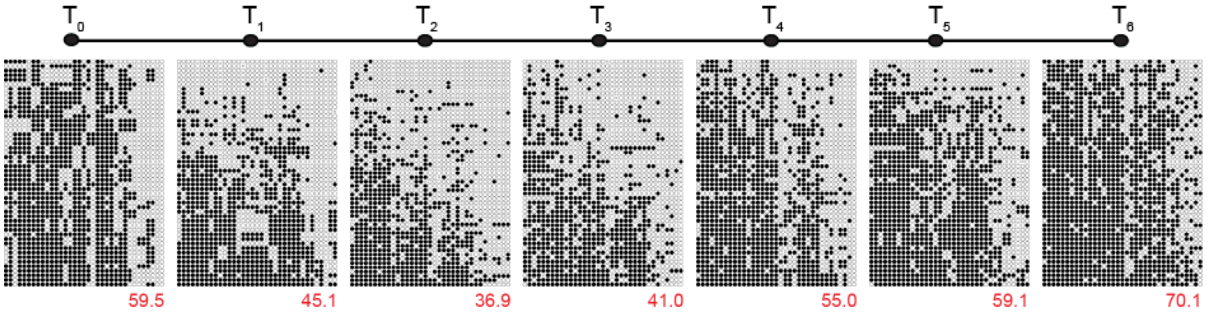


Figure 34: Methylation of L1_$_{1\text{-}14}$ family member 5'UTRs in the hiPSC-CRL1502 cell line. Cartoons represent the methylation status of 50 randomly selected bisulfite sequencing reads comprising locus-specific PCR products at the 5' junctions of L1 sequences. Transduction family members were analysed during hiPSC generation and neurodifferentiation in the hiPSC-CRL1502 cell line. Each column corresponds to a different time point (studied cell type, or time point, is indicated in the top of each column), (T_0= fibroblasts; T_1= hiPSCs; T_2= Neural epithelial; T_3= Immature neurons and neural rosettes; T_4= Neurons I and neural rosettes; T_5= Neurons II and neural rosettes; T_6= Neurons III and neural rosettes); each line contains the same set of CpG dinucleotides at L1 5'UTR region, each line contains a specific element of the family: **A.** L1-Ta subfamily population; **B.** Lineage progenitor; **C.** Donor L1; Red numbers below of each panel indicates the CpG percentage in each time point for each studied 5' UTR.

3.2: Overall L1 promoter CpG methylation levels during neurodifferentiation.

To facilitate comparisons amongst the various elements considered in the study, I generated average CpG methylation percentage values for each time point (**Figure 35**), Interestingly, in both cell line, the methylation status of the *de novo* L1 (hiPSC-CRL2429 only), donor L1 and lineage progenitor L1 appeared to increase with the relative age of each element.

A. **B.**

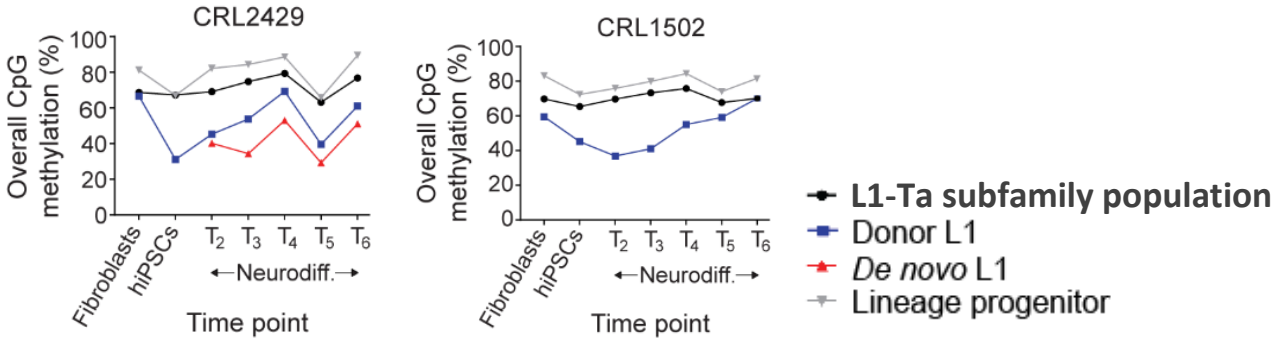


Figure 35: L1 promoter CpG methylation levels during neurodifferentiation time courses graphs II. (A and **B)** contain methylation level (y-axis) of the total CpG dinucleotides values in cell line hiPSC-CRL2429 **(A)** and in cell line hiPSC-CRL1502 **(B)** in each TP (T_0= fibroblasts; T_1= hiPSCs; T_2= Neural epithelial; T_3= Immature neurons and neural rosettes; T_4= Neurons I and neural rosettes; T_5= Neurons II and neural rosettes; T_6= Neurons III and neural rosettes). Graph legends: Overall L1 population, L1-Ta subfamily population (black line); Donor L1 (blue line); *de novo* L1 (red line); Lineage progenitor (grey line).

CONCLUSIONS CHAPTER 3:

In summary, all of the elements were relatively hypomethylated in hiPSCs, as expected, and were gradually methylated upon induction of somatic differentiation. The *de novo* L1 was hypomethylated shortly after retrotransposition in neural epithelial cells and it was gradually methylated during somatic differentiation, consistently presenting less methylation than the rest of the family members. In addition, the methylation status of the *de novo* L1, donor L1 and lineage progenitor L1s was correlated with their relative age in hiPSC-CRL2429 and hiPSC-CRL1502. Remarkably, shifts in methylation levels in some instances generated L1 promoters that were fully unmethylated in some cells, making their transcriptional activation plausible. These results reveal dynamic and distinct changes in methylation applied to germline and *de novo* L1 insertions during neurodifferentiation.

CHAPTER 4: DISCUSSION

This project has significantly elucidated how closely related L1 elements are distinctly repressed by their host genome during cellular reprogramming and differentiation. To this end, I first identified a *de novo* reprogramming-associated L1 insertion in hiPSCs. I used the 5' and 3' transduced sequences carried by this full-length L1 retrotransposition event as unique sequence tags to identify its immediate donor L1, and to then place these elements within an extended transduction family, composed of both reference and non-reference L1 elements. Although most of the family members described in this study have been identified in isolation by prior reports, my work is the first to describe these elements together as a 14 member family, which has expanded in the germline genomes of presumably healthy human individuals (**Table 3**).

The first evidence of extended L1 transduction families was provided by Goodier *et al*. [81]. In this study, the authors analysed L1 sequences present in the human and mouse reference genomes. They concluded that 23% of the analysed human L1s carried a 3' transduced sequence and, interestingly, they found a mouse L1 insertion bearing sequential transduction events. This finding pointed to the existence of L1 families where a daughter insertion retains the ability to mobilise again, resulting in more family members. Other transduction families have since been characterised by studying the L1 content of a subset of individuals [11, 12]. For example, by sequencing fosmid libraries generated from genomic DNA from 6 individuals, Beck *et al*., identified a cohort of polymorphic full-length L1s bearing 3' transductions, and proposed a model of L1 mobilisation where active L1s give rise to small subfamilies. A different study took advantage of the presence of 3' transductions to develop a new method for the identification of active L1 lineages, named Transduction-Specific Amplification Typing of L1 Active Subfamilies (TS-ATLAS) [12]. This method combined the enzymatic digestion of genomic DNA with the creation of plasmid libraries and subsequent L1-specific PCR to identify L1-genome junctions. The authors identified three separate polymorphic L1 transduction families using this approach. Despite the prominent role these previous studies have played in the discovery of L1 transduction families, the L1_$_{1\text{-}14}$ transduction family described in this thesis was not completely characterised previously, and it generated the first instance of an RC-L1 being associated with a *de novo* insertion in hiPSCs.

It is possible the L1_$_{1\text{-}14}$ transduction family could contain more members not identified in this study. Many of the family members identified here were polymorphic, and it

is unlikely that any individual would carry all of them. It follows that rarer polymorphic elements forming part of the transduction family are present in the global population. Also, to assign a particular L1 to this transduction family, I relied exclusively on the presence of transduced sequences. According to a prior estimate based on the presence of 3' genomic transduced sequences downstream of 129 referenced full-length elements, only around 10% of full-length L1s carry 3' transductions [82]. Therefore, there are almost certainly L1 insertions that originated from the family, but lacked transductions, and are therefore overlooked by my analysis. This highlights the importance of developing new methodologies to study the complete sequence of L1s, to identify internal nucleotide variants that allow classification of individual L1s into particular lineages [9, 87].

Remarkably, 3 of the 14 family members carried 5' transductions. This 5' transduction frequency (21.4%) is exceptionally high, given how rarely such events are found in the human germline (<0.1% of L1-Ta insertions [15]) and to my knowledge, a *de novo* L1 insertion with both 5' and 3' transduced sequences has never been reported before. It is presently unclear as to why 5' transductions are so frequent in this family. Two of these 5' transductions were relatively short (10nt, *de novo L1*; 18nt, Non-ref_ChrX_p11.4) and could be derived from the transcription initiating upstream of L1 position +1 directed by the L1 5' promoter region [209]. In contrast, the third 5' transduction identified in the element Ref_Chr1_p31.1_a was significantly longer, at 539 nucleotides, and is located ~126Kb upstream of the parental element. This suggests the transcription started at a distal promoter element and the regions between the transduced sequence and the L1 element were spliced out from the template mRNA.

Notably, certain chromosomes appear to have been targeted with disproportionate frequency by members of this transduction family. To elaborate, *de novo* L1, Ref_Chr1_p31.1_a, Ref_Chr1_p31.1 and Non-ref_Chr1_p22.2 were found on Chromosome 1, and the donor L1, Non-ref_Chr3_p24.3, Non-ref_Chr3_p12.2_a and Non-ref_Chr3_p12.2_b landed on Chromosome 3. Therefore, I speculate that certain L1 transduction families could have predisposition for inserting in genomic locations in physical proximity to their source element. Previous studies have described that proteins such as Hop1 and Red1, involved in resolving recombination intermediates during yeast meiosis, are influenced by chromosome structure [210]. Specific proteins and host factor with similar preferences could be directing nuclear L1-RNPs to certain chromosome locations with the appropriate structure, in this case the chromosomes 1 and 3.

Of the 14 family members identified here, 2 were already reported and tested for retrotransposition activity in previous studies; the donor L1, polymorphic in the human population and absent from the reference genome sequencing, [11] and the lineage progenitor of the L1 family, reported to be fixed in the human population and inactive [13]. Consistent with the results from Beck *et al.* [11], I found that the donor L1 (and *de novo* L1) were retrotransposition competent in a cultured cell retrotransposition assay. Interestingly, I also observed retrotransposition competency for two different alleles of the lineage progenitor L1 *in vitro*, demonstrating allelic heterogeneity for this element in the human population, and the potential consequences of this variation. Exhaustive sequence analysis to discard PCR induced mutations was not performed by Brouha *et al.*, where the lineage progenitor L1 was also assayed for retrotransposition [13]. In this thesis, I used a PCR based method, as for Brouha *et al.* [13], and by combining independent PCR amplification reactions to reconstruct elements. I performed Sanger sequencing on the entire length of the PCR fragment, identifying PCR induced mutations and reconstructing the L1s based on consensus sequences, where necessary. Via this method, I achieved similar levels of nucleotide sequence accuracy as the fosmid sequencing method used by Beck *et al*. [11]. Furthermore, I tested each of the two lineage progenitor L1 alleles *in vitro*. Brouha *et al.* did not publish the nucleotide sequences of the elements they tested [13], meaning I could not assess whether Brouha *et al.* tested one of the lineage progenitor L1 alleles considered here, or whether its previous assessment as an inactive element was due to point mutations introduced into the element's nucleotide sequence during PCR amplification during the cloning process. This discrepancy in the retrotransposition capability of the lineage progenitor L1 emphasizes the importance of allele specific analysis and demonstrates that some previously reported inactive L1s may in fact have active alleles in the human population.

As noted above, the lineage progenitor L1 alleles described in this study showed different retrotransposition capacities. Allele 1 carried a missense mutation (D523H) in the reverse transcriptase domain, which could explain its reduced retrotransposition activity compared to allele 2. The effect of this mutation has not been described in the literature but its location within the reverse transcriptase domain leads me to speculate that it could impact ORF2p reverse transcriptase activity directly, or interfere with the interaction of the L1 mRNA with ORF2p, or perhaps alter the interaction of the ribonucleoparticle with host factors involved in L1 retrotransposition [14, 167, 211, 212].

Once I had characterised the *de novo* L1 and its source transduction family, I

exploited a multiplexed targeted L1 bisulfite sequencing strategy [6, 7] to survey CpG methylation amongst these elements during reprogramming and neurodifferentiation. DNA methylation is a critical mechanism of defence against retrotransposons [213, 214]. L1 sequences are silenced via DNA methylation of the L1 5' UTR promoter region very early in mammalian embryogenesis [1, 4, 10, 116, 215, 216] and this repression is maintained in mature neurons. However, the dynamics of this methylation during neurodevelopment is relatively unexplored. Broadly, I observed that L1 promoters were methylated in fibroblasts, demethylated in hiPSCs, and gradually remethylated during neurodifferentiation. However, the magnitude of these changes varied greatly amongst the elements. I found that the donor L1 was recurrently hypomethylated when compared to the lineage progenitor L1 and the overall L1-Ta population in hiPSCs, and presented numerous cells where its promoter was fully unmethylated. This hypomethylated status suggests that the donor L1 could be transcriptionally active, and thus has the potential to generate mRNAs and new insertions in hiPSCs, as demonstrated by the discovery of a *de novo* L1 insertion generated by the donor L1. The potential for endogenous L1 mobilisation in hiPSCs should therefore be taken into consideration in discussing their use in medical therapies [217].

Notably, I observed a decrease in L1 methylation coincident with a potential gliogenic switch at T_5 [218, 219] that could potentially lead to the activation of L1. Although DNA methylation plays a prominent role in regulating this switch [220], the status of L1 promoter methylation has not been studied in gliogenesis. As L1 can be used as a surrogate of global DNA methylation [37, 221], I speculate the decrease in L1 promoter methylation relates to the genome-wide epigenetic changes associated with this gliogenic switch. Hypomethylation of L1 promoters suggests a window of opportunity may arise for L1 to mobilise in those cells entering the gliogenic switch. However, it is worth mentioning that the gliogenic switch observed in my neurodifferentiation does not necessarily resemble *in vivo* gliogenesis. In fact, Sun *et al.* described how older neural cultures enter glial differentiation processes in response to signals that induce neurodifferentiation in younger cultures [219]. Therefore, the gliogenic switch I observed in the final stages of the neurodifferentiation may not be the same as *in vivo* gliogenesis and thus, the window of opportunity for L1 retrotransposition may not lead to accumulation of L1 insertions in glial cells. This view is broadly supported by the literature evidence for L1 mobilisation [4, 104, 107, 143]. Single-cell *in vivo* studies have demonstrated that neurons accommodate more endogenous retrotransposition than glia [143].

Overall, I observed a correlation between the age and methylation level of the L1

promoters included here. While the *de novo* L1 was hypomethylated relatively shortly after its integration, the donor L1 was more methylated than this *de novo* L1 insertion. Interestingly, the lineage progenitor L1 was consistently more methylated than the donor L1, in hiPSCs and during neurodifferentiation. These findings suggest that younger L1s may be less methylated than older L1s during early embryogenesis, which hiPSCs are a model of. However, it is notable that I only interrogated the methylation state of these elements in 2 hiPSC cell lines and therefore it is unknown whether the same patterns are observed in hESCs or, for that matter, during human development *in vivo*. L1 promoter methylation has not been previously explored for the multipotent and immature neuronal cell types arising during neurogenesis. Interestingly, I found that *de novo* L1 remained quite hypomethylated during neural induction and neurodifferentiation. This result is consistent with previous publications that found sustained hypomethylation of *de novo* full-length L1 insertions from an engineered L1 reporter [222]. Therefore, I speculate that host genomes may need some rounds of cell cycling to recognize a *de novo* retrotransposition event and activate their defence mechanisms, in this case at the epigenetic level.

Given L1 hypomethylation was observed early in neuronal differentiation, it is perhaps surprising that I did not detect any additional *de novo* L1 insertions in mature neurons. However, this does not necessarily mean that L1 mobilisation did not take place, as any insertion would need to be present in enough cells to be detected at the RC-seq read count thresholds used here. Moreover, if a retrotransposition event took place during later neurodifferentiation, in immature or fully mature neurons, it would likely be present only in the cell where it arose because, at this stage of differentiation, the culture is composed of non-dividing cells. As I analysed genomic DNA from bulk cells, it is very difficult to achieve enough coverage to accurately identify an event which will only be present in one cell *and* PCR validate that event [127]. This could be partially solved by applying single-cell RC-seq [143] to a cohort of cells in different stages of neurodifferentiation. However, if a *de novo* L1 event was identified in a particular mature neuron, it would be very difficult to accurately infer when during differentiation that retrotransposition event took place. By using clonally expanded hiPSCs and analysing enough single-cells from each time point of neurodifferentiation it could nonetheless be possible to establish the pattern of when *de novo* L1 insertions predominantly accumulate. This approach could inform our general picture of how and when neural somatic mosaicism arises in the human brain [127].

The *de novo* L1 insertion most likely arose during hiPSC generation and was maintained during neural induction and neurodifferentiation, indicating that such events can

be present in differentiated cell lines derived from hiPSCs. However, despite sensitive PCR reactions showing an absence of the *de novo* L1 in DNA extracted from the matching fibroblasts, it is impossible to be sure that this L1 insertion was not in fact mosaic in the parental fibroblasts. To take this to an extreme, perhaps the L1 insertion was present in only one cell, and that cell was consumed by reprogramming, removing it from the parental fibroblast population. Nonetheless, we considered the L1 insertion to be "most likely" *de novo* because: i) L1-Ta family promoter sequences are heavily methylated in fibroblasts and hypomethylated in hiPSCs [1, 116], ii) L1 expression increases greatly during reprogramming [1, 223], iii) a prior WGS analysis of 10 human fibroblast populations expanded from single cells found no *de novo* L1 insertions [224]. For these reasons, the L1 insertion was very unlikely to occur in a fibroblast, supporting its annotation as being a *de novo* event during reprogramming.

The *de novo* L1 was a full-length insertion with no disabling mutations in its ORFs, and, as predicted, was retrotransposition competent in cultured cells. This insertion, like other *de novo* full-length events, could continue to impact the host genome, creating new daughter insertions and disrupting the genome with a cascade effect. In this case, the *de novo* L1 was intergenic and the accompanying transductions did not include protein-coding exons or regulatory elements [83], lessening the probability of a functional impact in neurons carrying the L1. Nonetheless, it is plausible that in the future by screening more hiPSC lines we could identify whether this *de novo* L1 insertion, or other similar insertions, impact the phenotype of hiPSC-derived cells.

Although DNA methylation strongly regulates the transcriptional activity of L1, there are other host factors able to restrict L1 activity. The adaptive methylation of retrotransposons seems to be influenced by the piRNA pathway [191, 192]. Piwi proteins act upstream of DNMT3L on repeats elements in mice [191, 192]. However, Piwi proteins are also known to regulate transposable elements (TEs) by targeting transcripts of active transposons [225]. In fact, a recent study showed the participation of Piwi proteins in controlling L1 retrotransposition in hiPSCs [159]. Piwi proteins also interact with the helicase MOV10-like-1 to silence retrotransposons in the mouse germline [193]. Thus, Piwi proteins could be regulating L1 activity in hiPSCs by targeting L1 mRNAs produced from the new L1 insertions which escaped restriction at the epigenetic level.

The findings described in this thesis support the hypothesis that L1 retrotransposition can have a multiplicative effect upon the genome, as a new L1 insertion can escape

repression and mobilise again and therefore have an exponential impact. Since the L1 family lineage described here is the largest transduction family discovered to date, it would be desirable to characterise it in additional hiPSC lines, and in hESCs. This approach could answer the following questions: i) Are there as-yet undescribed members of the L1_$_{1\text{-}14}$ transduction family in the human population? ii) Is the activity of the L1_$_{1\text{-}14}$ family more prevalent in hiPSCs than in hESCs or other cell types? iii) Is the L1_$_{1\text{-}14}$ family especially active in pluripotent cells compared to other L1 elements? And iv) Should L1 be considered a relevant mutagen when using hiPSCs in medical therapies? On this final point, the vast majority of L1 insertions found in hiPSCs and in hESCs are full length [1, 2, 115]. Do pluripotent cells hence provide a more “friendly” environment for L1 retrotransposition to generate full-length insertions. Although we lack a sufficient number of L1 insertions identified as *de novo* in hiPSCs to answer this question, it would nonetheless be sensible to further examine L1 mutagenesis during reprogramming as hiPSCs and their derivatives are being increasingly proposed as biomedical tools. It is also notable that other classes of mutation are encountered in hiPSCs, and that the epigenomic landscape can differ among hiPSC lines derived from the same cell type, leading to fluctuations in gene expression [226]. For these reasons, in the future it will be beneficial to fully characterise the genomes and epigenomes of numerous hiPSC lines and differentiate these into the broad range of mature cell types found in human organs. It would also be interesting to evaluate hiPSC genomes where retrotransposition is artificially bolstered and hindered, potentially using dCas9 fused to activator and repressor complexes [227].

In sum, this thesis identifies an L1 transduction family that is active during reprogramming, probably due to specific relaxation of DNA methylation via a mechanism that remains to be determined. More importantly, my results show how a new L1 insertion arising in a pluripotent stem cell can be recognized by the host genome and targeted for repression via DNA methylation, whilst still retaining capacity for further retrotransposition, even in mature neurons. This thesis therefore elucidates a key temporal aspect of the evolutionary “arms race” between mobile genetic elements and their host genome in somatic cells.

CONCLUSIONS

1. A *de novo,* full-length L1 insertion which most likely arose during reprogramming or early during the culture of hiPSCs was detected by RC-seq and PCR validated.

2. This is the first *de novo* L1 insertion described in hiPSCs carrying a 5' transduced sequence and a 3' transduced sequence.

3. The donor for the *de novo* L1 insertion was the first such element identified in hiPSCs and retained retrotransposition capability *in vitro*.

4. Further analysis revealed the donor L1 was related to numerous other L1s carrying a common 3' transduction, elucidating an L1 "transduction family" I named $L1\_{1\text{-}14}$.

5. The lineage progenitor L1 of the $L1\_{1\text{-}14}$ family was also active *in vitro*, and presented 2 allelic variants.

6. Bisulfite sequencing revealed that the donor L1 was hypomethylated in hiPSCs and was gradually remethylated during neurodifferentiation. The donor L1 was less methylated than the lineage progenitor L1, but more methylated than the *de novo* L1 insertion.

7. The *de novo* L1 insertion was hypomethylated relatively shortly after retrotransposition, during neural induction, and also gradually more methylated upon neuronal differentiation, suggesting common mechanisms methylated the *de novo* and donor L1s.

8. Cells carrying fully unmethylated L1-Ta sequences, including the *de novo* and donor L1s, were present throughout neurodifferentiation *in vitro*. This observation could explain somatic retrotransposition during human neurogenesis.

MATERIAL AND METHODS

hiPSC generation and neuronal differentiation.

The cell lines hiPSCs were episomally derived as described in [203]. as described by Shi et al., [203] Neuronal differentiation was performed with some minor changes in the protocol [203]. Feeder-free hiPSCs were cultured in MEF-conditioned KOSR medium (Gibco®) prior induction of neurodifferentiation, the cells were supplemented with 100ng/mL b-FGF. The addition of dual SMAD inhibitors SB431542 (10µM) and dorsomorphin (1µM) into the KOSR media initiated neurodifferentiation. The inhibitors were swapped for 3N medium (1:1 media mix of N2 and B27 containing media that contained of 1:1 neurobasal/DMEM-F12 media which was supplemented with 2% B27, 1% N2 and 2mM GlutaMax, 2.5µg/mL Insulin, 0.05mM NEAA, 0.05mM beta-mercaptoethanol (all from the company LifeTechnologies) in 25% incremental steps on days 4, 6, 8, 10. Matrigel-coated TC dishes were used to grow and harvest Neural rosettes , these cells were expanded in 3N medium adding the grow factor 20ng/mL b-FGF. Neuronal progenitors, were grown, approximately at day 30 using StemPro Accutasse Dissociation reagent (Thermo Fisher scientific) to obtain single cell solution to further seeding them in coated dishes. The dishes were coated with poly-L-ornithine (0.01% weight/volume and laminin (20µg/mL, respectively). For the indicated time points of this experiments, neural progenitors were feed in 3N medium.

Immunocytochemistry.

Neural cultures were fixed using 4% paraformaldehyde (Sigma) in PBS for 15min at room temperature and they were permeabilised in 0.01% Triton-X100 (Ajax Finechem) on Matrigel coated plastic cover slips in 3N media and were fixed in 4% paraformaldehyde (Sigma) in PBS for 15min at room temperature and permeabilised in 0.01% Triton-X100 (Ajax Finechem) in PBS for 15min at room temperature. All cells were blocked for 1hr with 10% goat serum (Invitrogen) in PBS. Primary antibodies used were OCT4 (1:100, Millipore), NANOG (1:100, Millipore), CUX1 (1:100, Abcam), GFAP (1:250, DAKO), TUBB3/TUJ1 (1:1000, Covance), BRN2 (1:100, Abcam), PAX6 (1:1000, DSHB), anti-phospho-histone H3 (Ser10) (1:200, Cell Signaling Technology) and were applied for 3-4hr at room temperature or overnight at 4°C. Isotype- and species-matched Alexa-Fluor conjugated secondary antibodies (1:1000, Invitrogen) were applied for 1hr at room temperature. Cells were washed in PBS and mounted on glass slides with prolong gold antifade containing DAPI (Invitrogen)

and imaged using an Olympus IX51 (Olympus) fluorescent microscope equipped with MicroPublisher 3.3 RTV CCD camera (QImaging) using Q-Capture Pro v6.0 software.

Nucleic acid extraction.

Per time point, 500,000 cells approximately were pelleted (1000 rpm, for 5 minutes), then washed with Dulbecco's Phospate Buffered Saline (DPBS) (-Calcium Chloride, -Magnesium Chloride) 1x (Gibco®). Then, the cells were pelleted another time with the same conditions and resuspended in UltraPure™ DNase/RNase-Free Distilled Water (Gibco®) in a volume of 100uL of Cells were lysed in 10mM Tris pH 9.0, 1mM EDTA, with 2% SDS and 100ug/mL proteinase K at 65°C. RNAse A was added to each sample in final concentration of 10ug/mL, samples were incubated at 37°C for 30 minutes. Using Phenol, DNA was extracted with phenol:chlorophorm:isoamyl alcohol (25:24:1) and chloroform:isoamyl alcohol (24:1). DNA was precipitated using with 0.1 volume of 3M sodium acetate and 2.5 volumes of 100% isopropanol. DNA was precipitated and then washed in 0.8 mL 75% EtOH, slightly air dried and resuspended in 50 µL of UltraPure™ DNase/RNase-Free Distilled Water (Gibco®). The quality and quantity of DNA were assessed by NanoDrop (Thermo Fisher Scientific) and Qubit using Qubit ® dsDNA HS Assay kit (Thermo Fisher Scientific), Qubit ® Assay Tubes and Qubit ® Fluorometer, following the manufacturer instructions.

Retrotransposon capture sequencing (RC-seq).

Extracted gDNA from the two hiPSC cell lines, hiPS-CRL2429 and hiPS-CRL1502 and from each time point, T_0, T_1, T_2, T_3, T_4, T_5 and T_6 was analysed by RC-seq, as described previously [143]. Briefly, the gDNA was sheared aiming to obtain ~450 nucleotides length fragments in 110uL of buffer TE: 10mM Tris-HCl, 1mM EDTA, pH8, using Covaris M220 Focused-ultrasonicator electronically controlled by Sonolab software with the protocol Covaris S220 (peak power, 50; duty factor, 10; cycles per burst, 200; duration, 90 seconds).

Concentrate DNA:

The sheared DNA was concentrated to the desired fragment size using 1.1 volumes of re-suspended room temperature Agentcourt® AMPure® XP beads (Beckman coulter), incubation was performed at room temperature for 15 minutes followed by separation of the beads with a DynaMag ™-2 Side magnetic rack (Thermo Fisher Scientific) for 2 minutes. Supernatant was transferred to new tubes and 400µL of 80% absolute ethanol (molecular grade) were added to wash the samples twice, supernatant was removed, and the tubes

were air dried removing them from the magnetic stand and allow them to dry for 15 minutes. 52uL of resuspension buffer (RB) were added to the tubes, incubate them at room temperature for 2 minutes, then the tubes were placed again for 2 minutes in the magnetic stand to remove the beads.

DNA quantification:

After shearing the gDNA to perform library preparation gDNA was measured to equalise all the samples to the lowest concentration obtained to use the same amount of DNA in the library preps using Pico Green Assay Kit (Sigma Aldrich).

Library Preparation:

Library preparation was then proceeded performing Illumina® TruSeq® Nano DNA Sample Prep Kit (Illumina®FC-121-4001/2) from 2µg input gDNA following the manufacturer instructions with some few modifications in the performance of the LM-PCR. The Illumina kit contains End Repair Mix, A-Tailing Mix, Stop Ligation Mix, Resuspension Buffer and a set of 16 Illumina Barcoded Library Adapters to its further use. Prior LM-PCR reaction agarose size selection was performed on High Resolution Agarose (Sigma-Aldrich). The gel electrophoresis was run at 120mA until the fragments were separated were enough. Bands were cut aiming to obtain fragments of 380-410 bp. The gel cuts were purified with MiniElute® Gel Extraction Kit (Quiagen). The LM-PCR was performed using 50µl Phusion High-Fidelity PCR Master Mix (2x), 18µl molecular grade water,1µl of each primer (TS-F Primer TS-R Primer) at 100µM, 30µl DNA library in a final volume of 100µl with the following conditions: 92°C for 45s; then 6 cycles at 98°C for 15s, 60°C for 30s, 72°C for 30s and 72°C for 5min. Once the PCR was complete the PCR reaction was cleaned with 1:1.1 ratio of Agentcourt® AMPure® XP beads (Beckman coulter) as described above: incubation at room temperature for 15 minutes followed by separation of the beads with a DynaMag ™-2 Side magnetic rack (Thermo Fisher Scientific) for 2 minutes. The supernatant was transferred to new tubes and washed twice with 400µL of 80% absolute ethanol (molecular grade), the supernatant was removed and the tubes were removed from the magnetic stand and air dried for 15min. 30uL of UltraPure™ DNase/RNase-Free Distilled Water (Gibco®) was added to the tubes and left toincubate at room temperature for 2 minutes. The tubes were then once again placed on the magnetic stand to separate the beads.

Confirmation of the libraries was performed running with 1µlof each library on an Agilent 2100 Bioanalyser instrument using Agilent DNA 1000 Reagents and DNA Chips (Agilent Technologies), to obtain approximately a library concentration of 20ng/µl.

Hybridisation of samples to capture probes:

The LNA capture probes (10 µM) were placed in 2 separated tubes with 4.5µl of LNA-5' and LNA-3' on each tube and they were incubated for 3 days at 47°C, with the lid set at 57°C in thermocycler block.

Recovery of capture sequences:

After the third day, Agentcourt® AMPure® XP beads (Beckman coulter) were prepared to perform another wash. Tubes with beads were incubated for 45 minutes, resuspending by flicking every 15 minutes. After the incubation, washes were performed at 47°C with 1x wash buffers (I, II, III and Stringent) using NimbleGen hybridization and wash kit (Roche) in the following order: 100µl Wash buffer I and incubation for 10s, 200µl Stringent buffer mix by pipetting 10 times for 5min twice. Another round of washes was performed at room temperature: 200µl Wash buffer I and incubation for 2min, 200µl Wash buffer II and incubation for 1min, 200µl Wash buffer III and incubation for 30s. After the final wash samples were resuspended in 50 µl of UltraPure™ DNase/RNase-Free Distilled Water (Gibco®).

Captured Sequences LM-PCR to amplify the post-hybridization library:

To the previous 50 µl sample Phusion® High-Fidelity PCR MasterMix LM-PCR was added with: 100µl (2x) Phusion® High-Fidelity PCR MasterMix, 46µl UltraPure™ DNase/RNase-Free Distilled Water (Gibco®), 1 µl of TS-F Primer (100 µM) to perform the LM-PCR with the following conditions: 92°C for 45s; then 6 cycles at 98°C for 15s, 60°C for 30s, 72°C for 30s and 72°C for 5min.Once the PCR finished the samples were cleaned using MiniElute® Gel extraction Kit (Qiagen) and quantified with Qubit ® dsDNA HS Assay kit (Thermo Fisher Scientific), Qubit ® Assay Tubes and Qubit® Fluorometer, following the manufacturer instructions. Capture samples were pooled 5' and 3' enrichments in a ratio 3:7 by molecular mass. Concentration and size distribution was measured again using Agilent DNA 1000 Assay (Sigma Aldrich). Enrichment of the samples was calculated by quantification by qPCR using Kapa Library Quantification qPCR kit Illumina®.

Sequencing:

The library was diluted and denaturalised following Illumina® manufacturer instructions for the specific run used in this study. Then, flow cell, sequencing cartridge and buffers for sequencing were prepared as Illumina® recommends.

Samples were sequenced in multiplex on an Illumina HiSeq2500 (Macrogen, South Korea). It was obtained a total of 726,181,832 paired-end 2×150mer reads across the 18 libraries (**Appendix 1**).

The identified (PRJEB27103) was deposited to name the RC-seq FASTQ files in the European Nucleotide Archive (ENA). TEBreak (https://github.com/adamewing/tebreak) pipe line was used to analyse RC-seq data. The HRG was used to align the reads using BWA-MEM [221] with parameters -Y –M. Those reads that were duplicates were indicated with Picard MarkDuplicates (http://broadinstitute.github.io/picard). In this analysis, TEBreak required a minimum read alignment of 30bp to the genome, and 30bp of L1 insertion, and in both cases at ≥95% identity.

To analyse the different L1 insertion candidates to be considered as putative *de novo* retrotransposition events (**Table 1**) the following different parameters were used: a) they were only present in one of the two hiPSC cell lines, whether in hiPS-CRL2429 or hiPS-CRL1502 or their derivatives but not in both cell lines at the same time, b) the insertions were absent from the T_0 or fibroblasts, c) they were not a previously described non-reference germline L1 insertion [8, 86, 91, 170, 171, 205, 228, 229] and d) they were supported by at least 2 RC-seq reads. Those insertions that were non-reference were annotated as polymorphic L1 insertions.

In silico analysis of consensus reads from RC-seq output data table.

1. From the sequencing output data, a study of the consensus sequences at both ends was performed; 5' end and 3' end, seeking the 3 mentioned hallmarks of retrotransposition (RTSN) [230]. DNA sequences were identified within the human genome working draft (HGWD) sequence using the BLAT server [199], at the UCSC Genome Bioinformatics website. Both consensus sequences should present a piece of the L1 sequence at its 5' or 3'end as well as a part of genomic DNA depending on where the insertion landed. In some cases, the input sequence can map to multiple genomic locations due to different repetitive elements. Aberrant sequences at junctions of the gDNA sequences were avoided.

For example, different pieces of the input sequence map to different genomic locations which are found when performing BLAT [207] on the genomic consensus using UCSC genome browser. An example of this is when the first 50bp portion of a 100bp genomic sequence maps to chromosome 1 and the last 50 bp maps to chromosome 2.

Another example is when microhomologies between the gDNA consensus sequence are found and the sequence of the L1 insertion (both sequences from the output table) that can be indicative of a PCR recombination event.

1. A. 5'end analysis: The insert consensus 5p was copied from each possible new L1 insertion and BLAST was performed using a sequence against an L1 reference sequence using Serial Cloner to check the quality of the alignment, and to know what part of the L1 sequence is present in the consensus read. Then, the UCSC genome browser was used to BLAT [207] the gDNA from the consensus sequence and double check that the same chromosomal coordinates as shown in the table were obtained. Many times, it is found that some L1 elements landed in repetitive sequences as old retrotransposons, LTRs [14] or centromeres, making the task of designing primers for PCR validations quite difficult. Hence, in many cases primers cannot be designed.

1. B. 3'end analysis: as for the 5' end analysis, the insert consensus 3p was copied from each possible new L1 insertion and BLAST was performed on this sequence against an L1 reference sequence to check the quality of the alignment. Then, the portion of the L1 sequence that is present within the consensus read was determined. A polyA tail should be found [74] followed by a TSD [64] and the gDNA where the insertion landed. As before, the UCSC database was used to BLAT [207] the gDNA from the consensus sequence.

Once the list of candidates was fully analysed, primers to PCR validate the insertions were designed in unique genomic DNA sequences flanking the 5' and the 3' ends. To be able to do this design, the gDNA was copied from each consensus, and BLAT was performed on the UCSC genome browser [207]. Most of the gDNA consensus sequences from the table are not very long, hindering the ability to design good primers, and raising the necessity to add more genomic DNA. To do so, it was necessary to copy the gDNA consensus sequence from the table in the UCSC Genome Browser, zoom out and then make a bigger nucleotide window to be able to find possible primers for PCR validations.

PCR validation of L1 insertions.

RC-seq reads indicating putative *de novo* L1 insertions were manually inspected and primers (**Appendix 2**) as previously described (**Figure 10**) were designed to PCR amplify integration sites and identify the hallmarks of bona-fide L1 retrotransposition events [231].

Empty/filled site, 5' L1-genome junction and 3' L1-genome junction PCRs were performed. Primers were situated within flanking genomic DNA sequences for empty/filled site PCRs. The same flanking primers were paired with appropriate L1-specific primers for L1-genome junction assays. Expand Long Range enzyme was used for empty/filled site PCRs using 1.75U Expand Long Template enzyme, 5µL of 5x Buffer with 12.5mM $MgCl_2$, 1.25µL DMSO 100%, 1.25µL 10mM dNTPs, 1µL primer mix (25µM each primer), 4ng genomic DNA template, and UltraPure™ DNase/RNase-Free Distilled Water (Gibco®) in a final volume of 25µL with the following PCR conditions: 92°C for 2min; then 10 cycles at 92°C for 10s, 59°C for 15s, 68°C for 6:30min, then 30 cycles at 92°C for 2min; 59°C for 15s, 68°C for 6:30min +20s/cycle and a single extension step of 68°C for 10min. 5' and 3' L1-genome junction PCR reactions were performed using 2U MyTaq hot-start DNA polymerase (Bioline #BIO-21112), 1x PCR buffer, 1µM of each primer and 5ng genomic DNA template, and molecular grade water in a final volume of 25 µL. Cycling conditions were as follows: 95°C for 2min; then 35 cycles at 95°C for 30s, 58°C for 30s, 72°C for 3min and a single extension step of 72°C for 5min. Amplified fragments were resolved on 1% and 2% agarose gels (1x TAE buffer) stained with SyberSafe (Life Technologies) for empty/filled site and 5' and 3' junction PCR assays, respectively, and imaged using a Typhoon FLA 9500 (GE Health-care life science, US). Amplicons of the expected size were excised from gels and DNA extracted using a QIAquick Gel Extraction Kit, followed by capillary sequencing to confirm and characterise L1 insertion structural features.

L1 genotyping and cloning.

To discard induced PCR mutations along the L1 sequences of interest and to distinguish possible allelic variants, four independent PCR reactions were performed using gDNA from the cell line hiPSC-CRL2429 to amplify the different family members assessed in this study used in retrotransposition assay. Expand long Range PCR were performed to amplify full-length L1s using 1.75U Expand Long Template enzyme, 5µL of 5x Buffer with 12.5mM $MgCl_2$, 1.25µL DMSO 100%, 1.25µL 10mM dNTPs, 1µL primer mix (25µM each primer), 4ng genomic DNA template, and UltraPure™ DNase/RNase-Free Distilled Water (Gibco®) in a final volume of 25µL with the following PCR conditions: 92°C for 2min; then 10 cycles at 92°C for 10s, 59°C for 15s, 68°C for 6:30min, then 30 cycles at 92°C for 2min; 59°C for 15s, 68°C for 6:30min +20s/cycle and a single extension step of 68°C for 10min. A NotI restriction enzyme sequence (5'-GC/GGCC) was introduced at the 5' end of each forward primer close to the L1-genome junction to facilitate cloning of full-length L1

insertions. Reverse primers were designed at each specific gDNA flanking regions. PCR amplification products were run in 1% agarose gel (1x TAE buffer) electrophoresis at 100mA for 1hour. Gel cutting of ~6Kb fragments were purified with Phenol, phenol:chlorophorm:isoamyl alcohol (25:24:1) and chlorophorm:isoamyl alcohol (24:1).

DNA was precipitated using with 0.1 volume of 3M sodium acetate and 2.5 volumes of 100% isopropanol. DNA was precipitated and then washed in 0.8 mL 75% EtOH, slightly air dried and resuspended in 20 µL of UltraPure™ DNase/RNase-Free Distilled Water (Gibco®). 500ng of purified PCR products were quantified by Nanodrop (Thermo Fisher Scientific) to perform NotI and Bstz17I (New England Biolabs) digestions in 1x CutSmart buffer at 37°C for 1hr. Digestions reactions were run in 2% agarose gels (1x TAE buffer) at 120mA for 45min and purified by Phenol, phenol:chlorophorm:isoamyl alcohol (25:24:1) and chlorophorm:isoamyl alcohol (24:1). DNA was precipitated using with 0.1 volume of 3M sodium acetate and 2.5 volumes of 100% isopropanol. DNA was precipitated and then washed in 0.8 mL 75% EtOH, slightly air dried and resuspended in 20µL of UltraPure™ DNase/RNase-Free Distilled Water (Gibco®).

The inserts containing L1 sequences were cloned into the vector TOPO®-XL PCR Cloning Kit (Life Technologies) according to the manufacturer's instructions. The ligation product was used to transform One ShotTOP10 electrocompetent bacteria as per the manufacturer's instructions using 5µL of ligation product. Competent bacteria were plated in LB agar containing 0.5 µg/mL of Kanamycin and incubated at 37°C overnight. Single colonies were picked and transferred to 5mL LB liquid containing 0.5 µg/mL of Kanamycin for Miniprep plasmid purification (QIAGEN). Stepping primers (L1_452_fwd, L1_1020_fwd, L1_1532_fwd, L1_1966_fwd, L1_2494_fwd, L1_3014_fwd, L1_3502_fwd, L1_4022_fwd, L1_4472_fwd, L1_4973_fwd and L1_5492_fwd (sequences are indicated in **Appendix 2**) were used to Sanger sequence the entire L1 sequences to reconstruct the consensus sequence from each element. Assemble of the 12 independent Sanger sequencing reactions was performed. The reactions were ~500 nucleotides in length and overlapping at both ends with the previous or the next sequencing read, excluding the first and the last sequencing reactions that only overlapped with the next and with the previous sequencing reaction respectively, to create the consensus of ~6 Kbs of each L1 element. The L1 sequences of the different elements were aligned to one another doing a multiple sequence alignment using Clustal omega software (https://www.ebi.ac.uk/Tools/msa/clustalo/) to create a consensus in each case. At least four clones were generated for each L1. Those unique mutations which were only present in one clone were considered induced PCR

mutations. The allelic variants were identified when particular nucleotides changes where shared in 2 or more independent clones.

Non-mutated nucleotide sequence fragments were chosen to reconstruct the original consensus for each element. The identical sequence of each element was rebuilt by paired strategic restriction digest of the PCR amplification products (all the restriction enzymes were from New England BioLabs®), those digestions were:

Lineage Progenitor_Allele 1: NotI-NheI; NheI-AgeI; AgeI-EcoRI; EcoRI-SpeI; SpeI-Bstz17I
Lineage Progenitor_Allele2: NotI-NheI; NheI-AgeI; AgeI-BamHI; BamHI-SpeI; SpeI-Bstz17I
Donor L1: NotI-NheI; NheI-AgeI; HindIII-Bstz17I
Non-Ref_Chr3_p24.3: NotI-SpeI; SpeI-Bstz17I

Each paired restriction enzyme digestion was run in 2% agarose gel (1x TAE buffer) for fragment separation to further ligate the resultant pertinent fragments from each L1 element and cleaned with MiniElute® Gel Extraction Kit (Quiagen). Ligation was realized into a pCEP4 vector using T4 ligase in the following (insert:vector) ratios.

Lineage Progenitor_Allele 1: ratio 5:1
Lineage Progenitor_Allele2: ratio 5:1
Donor L1: ratio 3:1
Non-Ref_Chr3_p24.3: ratio 2:1

In all the cases for each L1 element, 5µL of the ligation product was added to One ShotTOP10 chemically competent bacteria (Invitrogen) and transformations were performed as per manufacturer's instructions. 1µg/mL of ampicillin was added to LB agar were the bacteria were plated and incubated overnight at 37°C. Single colonies were grown in 5ml LB liquid for further Miniprep plasmid purification (QIAGEN). The elements were Sanger sequenced as previously described with stepping primers (L1_452_fwd, L1_1020_fwd, L1_1532_fwd, L1_1966_fwd, L1_2494_fwd, L1_3014_fwd, L1_3502_fwd, L1_4022_fwd, L1_4472_fwd, L1_4973_fwd and L1_5492_fwd, sequences are indicated in **Appendix 2**) to verify the re-built consensus sequences. The elements were digested using NotI and Bstz17I restriction enzymes (New England BioLab) and reconstructed into a retrotransposition indicator backbone. To verify resultant clones, these were capillary sequenced, as described above, using primers at both ends of the L1 sequence. While ~60 somatic mutations are known to occur per cell division in cultured fibroblasts [224], this equates to a probability of less than 1/1000 for such a mutation to have arisen in each L1

sequence in the parental fibroblasts prior to reprogramming (**Appendix 1**).

Retrotransposition cell-cultured assay

Prior to cell transfection, DNAs containing the different L1 of interest, were purified and resuspended in UltraPure™ DNase/RNase-Free Distilled Water (Gibco®) to 0.5µg/µL. High-glucose Dulbecco's modified Eagle's medium (DEMEM) without pyruvate (Gibco®), supplemented with 10% fetal bovine serum (Gibco®), 2mM L-Glutamine and 100U/mL penicillin, 100µg/mL streptomycin (Gibco®) (DMEM complete) media was used to grow Hela-JVM cells at 37°C in a humidified, 5%CO2 incubator at 37°C. Once obtained 80% of confluent cells, cells were dissociated with 0.05 % Trypsin (Gibco®) for 2min in at 37°C, neutralized with the same volume of media and counted with Neubauer cell counting chamber to plate $5x10^3$ cell/well in 6-well plates for the experiment and $4x10^4$ cell/well for transfection efficiency both performed in parallel. Cell transfections were performed after 12 hours at a ratio 4µL to 1µg plasmid DNA using FuGENE HD Transfection Reagent (Promega).

To perform transfection efficiency calculations, L1 reporter plasmids were co-transfected with 5µg of each construct and 0.5µg of pCAG-EGFP. After 48 hours cells were dissociated with 0.05 Trypsin (Gibco®) for 2min in at 37°C, neutralized with the DPBS Dubbecco's Phosphate Buffered Saline (-Calcium Chloride, -Magnesium Chloride) 1x of media and 10% Fetal Bovine Serum (FBS) (Gibco®). Cells were pelleted at 1000rpm for 4minutes and resuspended in 350µl of Dubbecco's Phosphate Buffered Saline (DPBS) (-Calcium Chloride, -Magnesium Chloride) 1x and taken to Cytoflex flow cytometer (Beckman-Coulter) at the Translational Research Institute Flow Cytometry Core to be analysed by Flow cytometry. The results obtained from each EGFP positive L1 reporter construct were used to normalize G418-resistant colony counts obtained in the colony forming assay or retrotransposition assay [232]. To perform the experiment of the retrotransposition assay, 72 hours after transfection, G418 (400µg/mL) (Sigma Aldrich) was added to the media to start the antibiotic selection [53], from that day, the antibiotic was added to the media every 48 hours for 14 days. The last day cells were fixed with Fixing solution (2% Formaldehyde, 0.2% Glutaraldehyde) 20minutes and stained with Crystal Violote Solution 1% 10min and washed with tap water and air dry over night to further manual colony count each well.

L1 CpG methylation analyses

Bisulfite sequencing of overall L1 population (L1-Ta subfamily-wide) and locus

specific was performed as previously described in Nguyen et al. [6]. Briefly, the gDNA on each time point, T_0, T_1, T_2, T_3, T_4, T_5 and $T_{6\ in}$ on each cell line hiPS-CRL2429 and hiPS-CRL1502 was bisulfite converted with EZ DNA Methylation-Lightning Kit (Zymo) following manufacturer instructions using 500ng of gDNA previously quantified by Nanodrop (Thermo Fisher Scientific) in a final volume of 20µl of UltraPure™ DNase/RNase-Free Distilled Water (Gibco®). A volume of 130µl of Lightening conversion reagent was used and samples were eluted in a final volume of 25µl of elution buffer. To PCR amplify the L1-Ta 5'UTR region containing a CpG island (**Figure 31**), I used L1_Bis-F and L1_Bis-R primers (**Appendix 2**). To amplify promoters in a locus-specific manner all the PCR amplifications were set up with the same reverse primer L1_Bis-R and a specific forward primer, designed in each case flanking the gDNA of each L1 promoter; Lineage progenitor, donor and *de novo* L1 insertions (L1_Bis-LP, L1_Bis-Donor, L1_Bis-DN, respectively). PCR reactions contained 1U of MyTaq hot-start DNA polymerase (Bioline), 2µl of bisulfite treated gDNA (1µg) from each sample, 1× reaction buffer and 2µM of each primer, in 20µL final volume. The cycling conditions used for PCR were the following: 95°C for 2min; then 40 cycles of 95°C for 30s, 54°C for 30s, 72°C for 30s, then a single extension step at 72°C for 5min. The purification of DNA from gel cuts was performed using phenol, phenol:chlorophorm:isoamyl alcohol (25:24:1) and chlorophorm:isoamyl alcohol (24:1). DNA was precipitated using with 0.1 volume of 3M sodium acetate and 2.5 volumes of 100% isopropanol. DNA was precipitated and then washed in 0.8 mL 75% EtOH, slightly air dried and resuspended in 30 µL of UltraPure™ DNase/RNase-Free Distilled Water (Gibco®). The quality and quantity of DNA were assessed by NanoDrop (Thermo Fisher Scientific). Library preparation was then performed with barcoded libraries following Illumina manufacturer instructions using TruSeq DNA PCR-free Library Preparation Kit (Illumina®) and subjected to multiplexed paired-end 2×300mer sequencing in Illumina MiSeq platform. Data were processed and analysed as described previously [6]. Briefly, this involved assembling read pairs into contigs and aligning them to mock converted target sequences (L1-Ta family or specific L1 loci) across their entire length using blastn. For each target sequence, QUMA software [233] was then used to analyse a total of 50 randomly selected, non-identical bisulfite converted sequences with default parameters (≤10% alignment mismatches and ≤5 unconverted CpH nucleotides) to generate methylation cartoons.

APPENDIX 1.

1. A Summary of RC-seq output read counts.

Cell line iPSC	Library DNA input	Time point	RC-seq reads (x2x150mer)	
			Count	Aligned %
CRL1502	Fibroblasts p4	T0	44033582	99.95
	hiPSCs p76	T1	42151994	99.74
	Neural epithelium	T2	33972001	99.77
	Immature neurons	T3	39766940	99.77
	Neurons I	T4	47155514	99.78
	Neurons II	T5	44381111	97.91
	Neurons III	T6	36222610	99.77
	hiPSCs p15	Earlier hiPSC passage	24385022	99.88
	hiPSCs p40	Earlier hiPSC passage	63130772	99.88
CRL2429	Fibroblasts p4	T0	24386590	99.91
	hiPSCs p70	T1	38460241	99.63
	Neural epithelium	T2	40174554	99.79
	Immature neurons	T3	46646999	99.78
	Neurons I	T4	27279492	99.79
	Neurons II	T5	46018310	99.77
	Neurons III	T6	36033944	99.54
	hiPSCs p11	Earlier hiPSC passage	64534189	99.40
	hiPSCs p40	Earlier hiPSC passage	27447967	99.39

APPENDIX 2.

Primer sequences (5'-3').

LM PCR primers

TS-F Primer	AATGATACGGCGACCACCGAGA
TS-R Primer	CAAGCAGAAGACGGCATACGAG

Genomic Primers for empty-filled validation of L1s

LP_Chr11_fwd	AGGAAACAGTGAGGGGAAGC
LP_Chr11_rev	TGAGGCCCAGGAGTCATATC
Donor_Chr3_fwd	TGTATGACAGTAAAATAATGGGTAGATGA
Donor_Chr3_rev	CTGGCCTCTTCACTGCATTT
DeNovo_Chr1_fwd	CTGGTAACCCCAGAATGACG
DeNovo_Chr1_rev	ATCCTGCCTCAGCGAACTTA
Non-ref_Chr3_fwd	TTGTGGGAAGGCAAAATGAT
Non-ref_Chr3_rev	TATTCAATCCCAACCCAGGA

L1-specific primers for validation of 5' and 3' L1-genome junctions

hL1_273_rev	ACCCGATTTTCCAGGTGCGT
hL1_ACshort_fwd	AGATATACCTAATGCTAGATGACAC

NotI/L1-genome junction spanning primers for cloning full-length L1s

LP_Chr11_NotI_fwd	CAAGCGGCCGCTTACATTTTTAAAGAATTGTAGGGGAG
Donor_Chr3_NotI_fwd	TAAAGCGGCCGCAACAGAATGAGTAAATAATGGAGGG
DeNovo_Chr1_NotI_fwd	TTCGCGGCCGCATTAAAGAAATGACATCTGAAATAATGGA
Non-ref_Chr3_NotI_fwd	CAACGCGGCCGCTTAAAGTTAAAGACACGG

L1-specific primers for sequencing full-length L1 elements

L1_452_fwd	GCCCAGGCTTGCTTAGGTA
L1_1020_fwd	TGATTTTGACGAGCTGAGAGAA
L1_1532_fwd	CCTCGAGAAGAGCAACTCCA
L1_1966_fwd	GCAAAATCACCAGCTAACATCA
L1_2494_fwd	AACTCAGCTCTGCACCAAGC

L1_3014_fwd	AAATCAGAGCAGAACTGAAGGAAA
L1_3502_fwd	GAGGCCAGCATCATTCTGATA
L1_4022_fwd	CAATCAGGCAGGAGAAGGAA
L1_4472_fwd	TCCCCATCAAGCTACCAATG
L1_4973_fwd	TGTCCAAAACACCAAAAGCA
L1_5492_fwd	TACCATTTGACCCAGCCATC

Primers for amplification of L1 promoters from bisulfite converted DNA

BiS_LP_Chr11_fwd	GATTTGTTTTTGGATTGTAAAATGGTT
BiS_Donor_Chr3_fwd	TGGGTAGATGAACAGATAAGTAAA
BiS_DeNovo_Chr1_fwd	GTTATTTGATAGTATTTTAATGAAGATT
BiS_OverallL1_fwd	TAGGGAGTGTTAGATAGTGG
BiS_hL1_rev	ACTATAATAAACTCCACCCAAT

APPENDIX 3. Validation by PCR and Sanger sequencing of the insertion TC_18 (*de novo* L1).

The amplification product of the insertion TC_18 at the 3'junction PCR was cloned and Sanger sequenced. Alignment with the 3'end of the L1 sequence was found as expected:

```
Seq_1  5878  cgggggaggggggagggatagcattgggagatatacctaatgctagatgacacattagtg  5937
                                      ||||||||||||||||||||||||||||||||||
Seq_2  115   ----------------------------AGATATACCTAATGCTAGATGACACATTAGTG  146

Seq_1  5938  ggtgcagcgcaccagcatggcacatgtatacatatgtaactaacctgcacaatgtgcaca  5997
             ||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||
Seq_2  147   GGTGCAGTGCACCAGCATGGCACATGTATACATATGTAACTAACCTGCACAATGTGCACA  206

Seq_1  5998  tgtaccctaaaacttagagtataataaaaaaaaaaaaaaaaa------------------  6039
             ||||||||||||||||||||||||||||||||||||||||
Seq_2  207   TGTACCCTAAAACTTAGAGTATAATAAAAAAAAAAAAAAAAGAATTGTAAAAAAAAAAA  266

Seq_1  6040  --------------------------aaaaaaaaaaaaa--------------------  6053
                                       |||||||||||||
Seq_2  267   TTATAAATAAAACAAAGAAGAATATGAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA  326
```

- Seq_1: L1 consensus sequence; Seq_2: Sanger sequencing data from one 3'junction TC_18 cloned product. In orange; a 3' transduced sequence of 44 nucleotides.

Alignment in between the genomic 3 consensus sequence and the Sanger sequencing results from one 3'junction TC_18 cloned product.

```
Seq_1  1     ----------------------------------------------------AAAAAAAAA  9
                                                                 |||||||||
Seq_2  181   TGTAACTAACCTGCACAATGTGCACATGTACCCTAAAACTTAGAGTATAATAAAAAAAAA  240

Seq_1  10    AAAAAAAA-------------------------------------------AAAAAAAA  25
             ||||||||                                           ||||||||
Seq_2  241   AAAAAAAAGAATTGTAAAAAAAAAAATTATAAATAAAACAAAGAAGAATATGAAAAAAAA  300

Seq_1  26    AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAGAAATGACATCTGTTAATCTTATTAA  85
             ||||||||||||||||||||||||||       ||||||||||||||||||||||||||
Seq_2  301   AAAAAAAAAAAAAAAAAAAAAAAAAA-------GAAATGACATCTGTTAATCTTATTAA  353

Seq_1  86    GTTCGCTGAGGCAGGATGAGGCTGGGTGCT-GGTAAATGGAGCAGATAAATGATCTGAGA  144
             ||||||||||||||||||              ||  |||   |||||||   |  |  |
Seq_2  354   GTTCGCTGAGGCAGGAT-----------AAGGGCGAATTCTGCAGATATCCGT-CACACT  401
```

- Seq_1: 3' end genomic consensus sequence; Seq_2: Sanger sequencing data from one 3'junction TC_18 cloned product; a 3' transduced sequence of 44 nucleotides.

After PCR validations the filled site amplification products were cloned in Topo-XL vector and Sanger-sequenced twice. The consensus sequence of the de novo insertion is:

Colour code of the sequence: 5'transduce sequence, green letters; L1 sequence of *de novo* L1, red letters; 3' transduced sequence, purple letters.

AAATAATGGAGGGGAGGAGCCAAGATGGCCGAATAGGAACAGCTCCGGTCTACAGCTCCCAGCGTGAGCGACGCAG
AAGACGGTGATTTCTGCATTTCCATCTGAGGTACCGGGTTCATCTCACTAGGGAGTGCCAGACAGTGGGCGCAGGCC
AGTGTGTGTGCGCACCGTGCGCGAGCCGAAGCAGGGCGAGGCATTGCCTCACCTGGGAAGCGCAAGGGGTCAGGG
AGTTCCCTTTCCGAGTCAAAGAAAGGGGTGACGGACGCACCTGGAAAATCGGGTCACTCCCACCCGAATATTGCGCTT
TTCAGACCGGCTTAAGAAACGGCGCACCACGAGACTATATCCCACACCTGGCTCGGAGGGTCCTACGCCCACGGAATC
TCGCTGATTGCTAGCACAGCAGTCTGAGATCAAACTGCAAGGCGGCAACGAGGCTGGGGGAGGGGCGCCCGCCATT
GCCCAGGCTTGCTTAGGTAAACAAAGCAGCCGGGAAGCTCGAACTGGGTGGAGCCCACCACAGCTCAAGGAGGCCT
GCCTGCCTCTGTAGGCTCCACCTCTGGGGGCAGGGCACAGACAAACAAAAAGACAGCAGTAACCTCTGCAGACTTAA
GTGTCCCTGTCTGACAGCTTTGAAGAGAGCAGTGGTTCTCCCAGCACGCAGCTGGAGATCTGAGAACGGGCAGACTG
CCTCCTCAAGTGGGTCCCTGACCCCTGACCCCCGAGCAGCCTAACTGGGAGGCACCCCCCAGCAGGGGCACACTGAC
ACCTCACACGGCAGGGTATTCCAACAGACCTGCAGCTGAGGGTCCTGTCTGTTAGAAGGAAAACTAACAACCAGAAA
GGACATCTACACCGAAAACCCATCTGTACATCACCATCATCAAAGACCAAAAGTAGATAAAACCACAAAGATGGGGA
AAAAACAGAACAGAAAAACTGGAAACTCTAAAACGCAGAGCGCCTCTCCTCCTCCAAAGGAACGCAGTTCCTCACCAG
CAACAGAACAAAGCTGGATGGAGAATGATTTTGACGAGCTGAGAGAAGAAGGCTTCAGACGATCAAATTACTCTGAG
CTACGGGAGGACATTCAAACCAAAGGCAAAGAAGTTGAAAACTTTGAAAAAAATTTAGAAGAATGTATAACTAGAAT
AACCAATACAGAGAAGTGCTTAAAGGAGCTGATGGAGCTGAAAACCAAGGCTCGAGAACTACGTGAAGAATGCAGA
AGCCTCAGGAGCCGATGCGATCAACTGGAAGAAAGGGTATCAGCAATGGAAGATGAAATGAATGAAATGAAGCGAG
AAGGGAAGTTTAGAGAAAAAAGAATAAAAAGAAATGAGCAAAGCCTCCAAGAAATATGGGACTATGTGAAAAGACC
AAATCTACGTCTGATTGGTGTACCTGAAAGTGATGTGGAGAATGGAACCAAGTTGGAAAACACTCTGCAGGATATTAT
CCAGGAGAACTTCCCCAATCTAGCAAGGCAGGCCAACGTTCAGATTCAGGAAATACAGAGAACGCCACAAAGATACT
CCTCGAGAAGAGCAACTCCAAGACACATAATTGTCAGATTCACCAAAGTTGAAATGAAGGAAAAAATGTTAAGGGCA
GCCAGAGAGAAAGGTCGGGTTACCCTCAAAGGAAAGCCCATCAGACTAACAGCGGATCTCTCGGCAGAAACCCTACA
AGCCAGAAGAGAGTGGGGGCCAATATTCAACATTCTTAAAGAAAAGAATTTTCAACCCAGAATTTCATATCCAGCCAA
ACTAAGCTTCATAAGTGAAGGAGAAATAAAATACTTTATAGACAAGCAAATGTTGAGAGATTTTGTCACCACCAGGCC
TGCCCTAAAAGAGCTCCTGAAGGAAGCGCTAAACATGGAAAGGAACAACCGGTACCAGCCGCTGCAAAATCATGCCA
AAATGTAAAGACCATCGAGACTAGGAAGAAACTGCATCAACTAATGAGCAAAATCACCAGCTAACATCATAATGACA
GGATCAAATTCACACATAACAATATTAACTTTAAATATAAATGGACTAAATTCTGCAATTAAAAGACACAGACTGGCAA
GTTGGATAAAGAGTCAAGACCCATCAGTGTGCTGTATTCAGGAAACCCATCTCACGTGCAGAGACACACATAGGCTCA
AAATAAAAGGATGGAGGAAGATCTACCAAGCCAATGGAAAACAAAAAAAGGCAGGGGTTGCAATCCTAGTCTCTGA
TAAAACAGACTTTAAACCAACAAAGATCAAAAGAGACAAAGAAGGCCATTACATAATGGTAAAGGGATCAATTCAAC
AAGAGGAGCTAACTATCCTAAATATTTATGCACCCAATACAGGAGCACCCAGATTCATAAAGCAAGTCCTCAGTGACC

TACAAAGAGACTTAGACTCCCACACATTAATAATGGGAGACTTTAACACCCCACTGTCAACATTAGACAGATCAACGA
GACAGAAAGTCAACAAGGATACCCAGGAATTGAACTCAGCTCTGCACCAAGCAGACCTAATAGACATCTACAGAACT
CTCCACCCCAAATCAACAGAATATACATTTTTTTCAGCACCACACCACACCTATTCCAAAATTGACCACATAGTTGGAAG
TAAAGCTCTCCTCAGCAAATGTAAAAGAACAGAAATTATAACAAACTATCTCTCAGACCACAGTGCAATCAAACTAGA
ACTCAGGATTAAGAATCTCACTCAAAGCCGCTCAACTACATGGAAACTGAACAACCTGCTCCTGAATGACTACTGGGT
ACATAACGAAATGAAGGCAGAAATAAAGATGTTCTTTGAAACCAACGAGAACAAAGACACCACATACCAGAATCTCT
GGGACGCACTCAAAGCAGTGTGTAGAGGGAAATTTATAGCACTAAATGCCTACAAGAGAAAGCAGGAAAGATCCAA
AATTGACACCCTAACATCACAATTAAAAGAACTAGAAAAGCAAGAGCAAACACATTCAAAAGCTAGCAGAAGGCAAG
AAATAACTAAAATCAGAGCAGAACTGAAGGAAATAGAGACACAAAAAACCCTTCAAAAAATCAATGAATCCAGGAGC
TGGTTTTTTGAAAGGATCAACAAAATTGATAGACCGCTAGCAAGACTAATAAAGAAAAAAAGAGAGAAGAATCAAAT
AGACACAATAAAAAATGATAAAGGGGATATCACCACCGATCCCACAGAAATACAAACTACCATCAGAGAATACTACA
AACACCTCTACGCAAATAAACTAGAAAATCTAGAAGAAATGGATACATTCCTCGACACATACACTCTCCCAAGACTAAA
CCAGGAAGAAGTTGAATCTCTGAATCGACCAATAACAGGCTCTGAAATTGTGGCAATAATCAATAGTTTACCAACCAA
AAAGAGTCCAGGACCAGATGGATTCACAGCCGAATTCTACCAGGGGTACAAGGAGGAACTGGTACCATTCCTTCTGA
AACTATTCCAATCAATAGAAAAAGAGGGAATCCTCCCTAACTCATTTTATGAGGCCAGCATCATACTGATACCAAAGCC
GGGCAGAGACACAACCAAAAAAGAGAATTTTAGACCAATATCCTTGATGAACATTGATGCAAAAATCCTCAATAAAAT
ACTGGCAAACCGAATCCAGCAGCACATCAAAAAGCTTATCCACCATGATCAAGTGGGCTTCATCCCTGGGATGCAAGG
CTGGTTCAATACACGCAAATCAATAAATGTAATCCAGCATATAAACAGAGCCAAAGACAAAAACCACATGATTATCTC
AATAGATGCAGAAAAAGCCTTTGACAAAATTCAACAACCCTTCATGCTAAAAACTCTCAATAAATTAGGTATTGGTGG
GACGTATTTCAAAATAATAAGAGCTATCTATGACAAACCCACAGCCAATATCATACTGAATGGGCAAAAACTGGAAGC
ATTCCCTTTGAAAACCGGCACAAGACAGGGATGCCCTCTCTCACCACTCCTATTCAACATAGTGTTGGAAGTTCTGGCC
AGGGCAATCAGGCAGGAGAAGGAAATAAAGGGTATTCAATTAGGAAAAGAGGAAGTCAAATTGTCCCTGTTTGCAG
ACGACATGATTGTTTATCTAGAAAACCCCATCGTCTCAGCCCAAAATCTCCTTAAGCTGATAAGCAACTTCAGCAAAGT
CTCAGGATACAAAATCAATGTACAAAAATCACAAGCATTCTTATACACCAACAACAGACAAACAGAGAGCCAAATCAT
GGGTGAACTCCCATTCACAATTGCTTCAAAGAGAATAAAATACCTAGGAATCCAACTTACAAGGGATGTGAAGGACCT
CTTCAAGGAGAACTACAAACCACTGCTCAAGGAAATAAAAGAGGAGACAAACAAATGGAAGAACATTCCATGCTCAT
GGGTAGGAAGAATCAATATCGTGAAAATGGCCATACTGCCCAAGGTAATTTACAGATTCAATGCCATCCCCATCAAGC
TACCAATGACTTTCTTCACAGAATTGGAAAAAACTACTTTAAAGTTCATATGGAACCAAAAAAGAGCCCGCATTGCCAA
GTCAATCCTAAGCCAAAAGAACAAAGCTGGAGGCATCACACTACCTGACTTCAAACTATACTACAAGGCTACAGTAAC
CAAAACAGCATGGTACTGGTACCAAAACAGAGATATAGATCAATGGAACAGAACAGAGCCCTCAGAAATAATGCCGC
ATATCTACAACTATCTGATCTTTGACAAACCTGAGAAAAACAAGCAATGGGGAAAGGATTCCCTATTTAATAAATGGT
GCTGGGAAAACTGGCTAGCCATATGTAGAAAGCTGAAACTGGATCCCTTCCTTACACCTTATACAAAAATCAATTCAA
GATGGATTAAAGATTTAAACGTTAAACCTAAAACCATAAAAACCCTAGAAGAAAACCTAGGCATTACCATTCAGGACA
TAGGCGTGGGCAAGGACTTCATGTCCAAAACACCAAAAGCAATGGCAACAAAAGACAAAATTGACAAATGGGATCTA
ATTAAACTAAAGAGCTTCTGCACAGCAAAAGAAACTACCATCAGAGTGAACAGGCAACCTACAACATGGGAGAAAAT
TTTTGCAACCTACTCATCTGACAAAGGGCTAATATCCAGAATCTACAATGAACTCAAACAAATTTACAAGAAAAAAACA

AACAACCCCATCAAAAAGTGGGCGAAGGACATGAACAGACACTTCTCAAAAGAAGACATTTATGCAGCCAAAAAACA
CATGAAGAAATGCTCATCATCACTGGCCATCAGAGAAATGCAAATCAAAACCACTATGAGATATCATCTCACACCAGT
TAGAATGGCAATCATTAAAAAGTCAGGAAACAACAGGTGCTGGAGAGGATGCGGAGAAATAGGAACACTTTTACACT
GTTGGTGGGACTGTAAACTAGTTCAACCATTGTGGAAGTCAGTGTGGCGATTCCTCAGGGATCTAGAACTAGAAATAC
CATTTGACCCAGCCATCCCATTACTGGGTATATACCCAAATGAGTATAAATCATGCTGCTATAAAGACACATGCACACG
TATGTTTATTGCGGCACTATTCACAATAGCAAAGACTTGGAACCAACCCAAATGCCCAACAATGATAGACTGGATTAA
GAAAATGTGGCACATATACACCATGGAATACTATGCAGCCATAAAAAATGATGAGTTCATATCCTTTGTAGGGACATG
GATGAAATTGGAAACCATCATTCTCAGTAAACTATCGCAAGAACAAAAAACCAAACACCGCATATTCTCACTCATAGG
TGGTAATTGAACAATGAGATCACATGGACACAGGAAGGGGAATATCACACTCTGGGGACTGTGGTGGGGTCGGGGG
AGGGGGGAGGGATAGCATTGGGAGATATACCTAATGCTAGATGACACATTAGTGGGTGCAGTGCACCAGCATGGCA
CATGTATACATATGTAACTAACCTGCACAATGTGCACATGTACCCTAAAACTTAGAGTATAATAAAAAAAAAAAAAAA
AAGAATTGTAAAAAAAAAAATTATAAATAAAACAAAGAAGAATATGAAAAAAAAAAAAAAATAAAAAAAAAAAAAA
AAAAAA

APPENDIX 4. Putative *de novo* L1 insertions data of: TSD_3prime, TSD_5prime, Genomic Consensus_5p and Genomic Consensus_3p RC-seq output data table.

Timecourse_1

TSD_5prime:

ATCTTGG

TSD_3prime:

ATCTTGG

Genomic_Consensus_5p:

AAAGAAACTACCATCAGAGTGAACAGACAACCTGCAGAATGAAAGAAAATATTTATAA
ACTATGCATCTGACAAAGGACTAATATCCAGAATCTGTAAGGAACTCACACATCTCAG
CAACAACAATGAAAAATAACCCCATTAAAAGGTGGGAAAGGGCCGTGCGCGGTGGCT
CACACCTGTAATCCCAGCACTTTGGGAGGCCAATGCGGGCGGATCACAAAGCCAAG
AGATCGAGACCATCTTGGAGA

Genomic_Consensus_3p:

TTCAGACGTGTGCTCTTCCGATCTTGGCTAACATGGTGAAACCCCGTCTCTACTAAAA
ATACAAAAAAAATTAGCTGGGCATGGTGGCAGATGCCCATAGTCCCAGCTACTCGGG
AGGCTGAGGCAGGAGAATGGCGTGAACCCGGGAGGC

Timecourse_2

TSD_5prime:

NA

TSD_3prime:

NA

Genomic_Consensus_5p:

AATATCCTTGATGAATATTGATGCAAAAATCCTCAATAAAATACTGGCAAACTGAATCC
AGCAGCACATCAAAAAGCTTATCCACCATGATCAAGTGGGCTTCATCCCTGGGATGC
AAGGCTGGTTCAATATACACAAATCAATAAATGTAATCCAGCATATAAACAGAACCAAA
GACAAAAACCACATGATTATCTCAATAGATGCAGAAAAGGCCTGTGACAAAATTCAAC
AACACTTCATGCTAAAAACTCTCAAGAAATTAGGAATAAGTTTACAAAGCTCATGGAAA
CCATGAACAAAGAGCTAAAGGACCATGAGAACAATGTCTCAATAAATAGAGTATATTG
ATAAAGAAACAGAACTTATATAAAG

Genomic_Consensus_3p:

GCAGATTTTACAAAAAAATAGTTTATAAAAAGCAATAAACAAAAACAACTACAACTGAT
GGCAAACAGCAAACACTCCTGAGAGGACAGAAAGGAAGAAATCTAATTTCTAGAGTT
GTTACATATAATATTCAAAATGTCCAGTATTTACCAAAAAATTATGTATCATGTGAAGAA
AAAATAAGTTATGGCCCATGGATAGGAGAAAGTAACAGAAATTGTGCCTGAGGAAGAT
CGGAAGAGCG

Timecourse_3

TSD_5prime:

AAGACA

TSD_3prime:

AAGACA

Genomic_Consensus_5p:

AAAGTTAATATTTCTCAGCCCAAAATCTCCTTAAGCTGATAAGCAACTTCAGCAAAGTC
TCAAGATACAAAATCAATGTACAAAAATCACAAGCATTCTTATACACCAACAACAGACA
AGGATCCCTCTCTCACCACTCCTATTCAACATAGTATTGGAAGTTCTGGCCAAGGCAA
TCAGTCAAGAGAAGGAAATAAAGCTATTCAAATAGGAAAAGA

Genomic_Consensus_3p:

GAGAATGGCGTGAACCTGAGAGGCGGAGCTTGCAGTGAGCCGAGATCCGCCACTGC
ACTCCAGCCTGGGCAACAGAGTGAGACTCCGTCTCAAAAAAAAAAAAAAAGAAAAAA
GAAAACTGGCACAAGACAAAGAACCCCCCCCCACCACC

Timecourse_4

TSD_5prime:

CAGGGGCAGTGACTCACATCTGTAATTCCAGCACACCTGTAATTCCAACACTTTGTGG
GGCCAAGATAGGAG

TSD_3prime:

CAGGGGCAGTGACTCACATCTGTAATTCCAGCACACCTGTAATTCCAACACTTTGTGG
GGCCAAGATAGGAG

Genomic_Consensus_5p:

GATCAGAATCACCTCAGAAGCTTTAAAAAACAATCTATAGGCCAGGGGCAGTGACTCA
CATCTGTAATTCCAGCACACCTGTAATTCCAACACTTTGTGGGGCCAAGATAGGAGTT
TTCCTTCTAACAGACAGGACCCTCAGCTGCAGGTC

Genomic_Consensus_3p:

CTACACGACGCTCTTCCGATCTCAGGGGCAGTGACTCACATCTGTAATTCCAGCACA
CCTGTAATTCCAACACTTTGTGGGGCCAAGATAGGAGGATCACTTGAGCCCAGGAGT
TCGAGACCAGCCTGGGTAACAGAGGGAGACTGCCCCC

Timecourse_5

TSD_5prime:

AAAAAACATGCCAAATTGTAAAGACCATCAAGGCTAGGAAGAAACTGCATCAA

TSD_3prime:

AAAAAACATGCCAAATTGTAAAGACCATCAAGGCTAGGAAGAAACTGCATCAA

Genomic_Consensus_5p:

GTCACCACCACGCCTGCCCTACAAGAGCTCCTGAAAGAAGCACTAAACATGGAAAGG
AACAACCGGTACCAGCCACTGAAAAAACATGCCAAATTGTAAAGACCATCAAGGCTA
GGAAGAAACTGCATCAATTAAAGATCGGAAGAGCGTC

Genomic_Consensus_3p:

GGGTGCAGCGCACCAGCATGGCACATGTATACATATGTAACTAACCTGCACAATGTG
CACATGTACCCTAAAACTTAGAGTATAATAAAAAAAACAAAAAAACAAAAAAAACAAAA
AAAAACATGCCAAATTGTAAAGACCATCAAGGCTAGGAAGAAACTGCATCAA

Timecourse_6

TSD_5prime:

NA

TSD_3prime:

NA

Genomic_Consensus_5p:

CCCATAGCTCGGAGTAATTTGATCGTCTGAAGCCTTCTTCTCTCAGCTCATCAAAGTC
ATTCTCCATCCAGCTTTGTTCCATTGCTGGTGAGGAACTGCGTTCCTTTGGAGGAAAA
GAGGCATTCTGGTTTTTGGAACTTTCCGCATTTTTGCACTGGGTTTTCCTCATCTTCAT
GGATTTATCTACCTTTGGTCTTTGATGTTGGTGACCTTTGGATGGGGTCTCTGAGTGG
AAGTGCT

Genomic_Consensus_3p:

CTTCCTTGCATTGAGTTAGAACATGCTCCTTTAGCTTGGAGGAGTTTGTTATTACCTAC
CTTCTGAAGCCTACTTCTGTCAATTCACCAAACTCATTCTCTATCCAGTTTTGTTCCCT
TGCTGTCAAGGAAGATCGGAAGAGCGTCGTGTA

Timecourse_7

TSD_5prime:

CATTCTTTTTTTTTTTTTTT

TSD_3prime:

CATTCTTTTTTTTTTTTTTT

Genomic_Consensus_5p:

AAAGTCACCCTAAATTCATTGGGAAAAGTTATACAAAAGAAGGGCCAATACCATTCTTT
TTTTTTTTTTTTTTTTTTTTTTTGAGACGGAGTCTCGCTCTGTGGCCCAGGCGGGACTG
CAGTGGCGCGATCTCGGCTCACTGCAAGCTCCG

Genomic_Consensus_3p:

ATCTCCTTAAGCTGATAAGCACTTCAGCAAAGTCTCAGGATACAAAATCAATGTGCAA
AAATCACAAGCATTCTTTTTTTTTTTTTTTTCCATTTTTTTTCTTTTTTATTGATCATTCTT
GGGTGTTTCTCGCAGAGGGGGATTTGGCAGGGTCACAGGACAATAGTGGAGGGAAG
G

Timecourse_8

TSD_5prime:

GGCGACAGAGCGA

TSD_3prime:

GGTGACAGCGCCA

Genomic_Consensus_5p:

GGGGAGTTTGCAGTGAGCCAAGATCGTACCTCCGCACTCCAGCCTGGGCGACAGAG
CGATACTCCATCTCAAAAAACAAAACGAAACAGGTCGGAAGATCACACGTCTTAAGTG
CACTGAGGAAGACAGCAGGAATGGCGACTTCGGCTTG

Genomic_Consensus_3p:

GATGACAGGATACTGGGTGCAGCGCACCAGCATGGCACGTGTATACATATGTAACTA
ACCTGCACAACGTGCACAGGTGACAGCGCCAGACTTGGTCTCAAGAAAAAAAGAAAT
AGAGGAGGGGGCTTCAGTCTGTGTAGGGGAGCTGGGG

Timecourse_9

TSD_5prime:

AAAAAAAAAAAAAAAA

TSD_3prime:

AAAAAAAAGAAAAAAAA

Genomic_Consensus_5p:

ACTGGATTTTTGGCCAGGTGCAGTGGCTCATGCCTGTAATCCCAGCATTTTGGGAGC
CTGAGGCGGGAGAATCACTTGAAGTCAGGAGTTTGAGACAAGCCTAGCCAACATAGT
GAAACTCTGTCTCTACAAAAAAAAAAAAAAAAACACATTAGCGAGGCAAAGAGGGGG
GGACCGGTGGACCCCGGCACGGGGGGGGGGGGAATGAGCAAAACACCCCACCCCAG
GCAGCAAGGGGTCGCG

Genomic_Consensus_3p:

GATGACACATTAGTGGGTGCAGCGCACCAGCATGGCACATGTATACATATGTAACTA
ACCTGCACAATGTGCACATGTACCCTAAAACTTAGAGTATAATAAAAAAAAAAAAAAGA
AAAAAAAAAAAAAACAAAACAAAACAAAAAAAAAAAAAAAA

Timecourse_10

TSD_5prime:

NA

TSD_3prime:

NA

Genomic_Consensus_5p:

ACAAAGACCCTGTCTAATAAAGAAAAACAAACCAAAAAACTACCGATACCTCTTGGCT
GGGCAGACTATAATATCCTAAAGAATTGCAGAGTCTGGGTAAATGCCTGATTTGGAG
GAAAGTGCCCCAAAGGGCAGCCTGTCCAACCTAAGGGGGTCCTTGTGACCCTTTGAT
GGAAATGCAGAAATCACCCGTCTTCTGCGTCGCTCACGCTGGGAGCTGTAGACCGG
AGCTGTTCCTATTCGGCCATCTTGGCTCC

Genomic_Consensus_3p:

AGTTCAGACGTGTGCTCTTCCGATCTAGTGGGAGCAACAACTACCTCTTCAGCTTCCA
GGGATTTGCTAGGGAGCTTGCTACAGCCATACATACGCCTGGTAGGTCTTGTCCCAG
AGGACATACGTTTTGAGATGAACTCAAACACTTGGA

Timecourse_11

TSD_5prime:

GAAAAT

TSD_3prime:

GAAAAT

Genomic_Consensus_5p:

ACTAGAAATGACTCATCCATTGAATTAAGTTAAGCATTTGACATAACATGATTTCATTAT
GTAGAATGATTTTAAATGCATTTGAGTCATATTTCATTATTAGAATTAAACTTATGTTTG
TGTTCTATTATATCAGTGATATTGTAGTAATATAGTAAAAAGAAAAGTTTTGGAATACTA
TTCCAGTCAACATAATTCTGAGGAAGTTGTTTGAAAATTTTTGCAACCTACTCATCTGA
CAAAGGGCTAATATCCAGAATCTACAATGAACTCAAACAAATTTATAAGAAAAAAACAA
ACAACCCCATCAAAAAGTGGGCGAAGGACATGAACAGACACTTCTCAAAAGAAGACA
TTTATGCAGCCAAAAAACACATGAAGAAATGCTCATCATCACTGGCCATCAGAGAAAT
GCAAATCAAAACCACTATGAGATATCATCTCACACCGGTTAGAATGGCAATCATTAAG
AAGTCAGGAAACAACAGGTGCTGGAGAGGATGCGGAGAAATAGGAACACTTTTACAC
TGTTGGTGGGACTGTAAACTAGT

Genomic_Consensus_3p:

AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAGAAAATGGCAGGTTACATTTC
CTCACAACAGTGTCACTACTGCTACACAAATAAAAAAAATCAATTTTTAAAAAGTGCTG
TAATCTGCTACTTCACACTAATGCCATATGGTCATCATTTTATAATTAATTGATAAATTA
GCA

Timecourse_12

TSD_5prime:

ATTTCCTTAGACATATTCTTGGAAATAGAAT

TSD_3prime:

ATTTCCTTAGACATATTCTTGGAAATAGAAT

Genomic_Consensus_5p:

TGGTGCGCTGCACCCACCAATGTGTCATCTAGCATTAGGTATATCAATAGCTTCAGAT
TTTCCATAATTTTTGATATAGCAATTCCATTTCCTTAGACATATTCTTGGAAATAGAATT
GCTATATCAAAAATTATGGAAAATCTGAAGCTA

Genomic_Consensus_3p:

ATCCATTCTGTGGATTCAAAATCATGTAGATCACCAGTCTCTAATGGTTGGACATTTAG
GATTTATCTGGTTTCTCATCTGAGTGTTTATTAGATGACTATTGGATTTCCTTAGACAT
ATTCTTGGAAATAGAATGCAATCATCAAATGGA

Timecourse_13

TSD_5prime:

AAAAAAACAAAACAACT

TSD_3prime:

AAAAAAACAAAACAACT

Genomic_Consensus_5p:

CATTGAATACACATGTTAAGATGGGAACAAGAGACACTGGGGACCACTAGATTGGGG
AGGAAAGGTAGGGGTTGTGGGCTGAAGAATTACCTGTTGGGTACTGTGTTTACTGCC
TGGGTGGTAGGATCACTGGGAGTCCAAGCCTCAGCATCACACAACTACTCATGTAAC
AGTCTTTCATTAACCTATAATAAAAGTTGAAATTAATTAAAAAAACAAAACAACTTTTGA
GAAGTGTCTGTTCATGTCCTTCGCCCACTTTTTGATGGGGTTGTTTGTTTTTTTCTTGT
AAATTTGTTTGAGTTCATTGTAGATTCTGGATATTAGCCCTTTGTCAGATGAGTAGGTT
GCAAAAATTTTCTCCCATGTTGTAGGTTGCCTGTTCACTCTGATGGAAGACATTTATG
CAGCCAAAAAACACATGAAGAAATGCTCATCATCACTGGCCAT

Genomic_Consensus_3p:

ACCGTGAAACTCAGAGAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACAAAACAA
CTACACTGTTTCATTGCTCTAGATTTCTTTTTGTCTCCATTTAATTATGGAGAGACTTG
CAGAGACAGAAATGACTGTCACGTGTTCTTATACACATAAAGCCCTAGAAACAGAAGC
CACAGCACAT

Timecourse_14

TSD_5prime:

NA

TSD_3prime:

NA

Genomic_Consensus_5p:

NA

Genomic_Consensus_3p:

AGATGACATGATTGTATATTTAGAAAACCCCATGTATACATATGTAACTAACCTGCACA
ATGTGCACATGTACCCTAAAACTTAGAGTATAATAAAAAAAAAAAATTAAAAAAAAAAA
AAAAAAAAAACTTAAAAAAAACTTTGGCAATTAAA

Timecourse_15

TSD_5prime:

AGTTCTTT

TSD_3prime:

AGTTCTTT

Genomic_Consensus_5p:

ACCAAGGTTGAAATGAAGGAAAAAATGTTAAGATCAGCCAGAGAGAAAGGTCGGGTT
ACCCACAGTGTTGTATTTTATAAAGAACTCTGCTCTATGCTTATTAAACTGGACATTAC
TAAAACATAATGACAGAGAAATGTTGAAAGTAAATAGATGAAAAAATAAACACCAGACA
AATATTAACTTAAAAATTTTGGTTTAGGCAATAGTTGACAAAATATATATTAATGACCAA
AGCATCATTAAAAATAAAGAAGGACATTACTTAATGATACAATGAATAATTTCATACAC
AAAATAA

Genomic_Consensus_3p:

TTGCAAAAATTTTCTTCCATTCTGTAGGTTGCCAGTTCACTCTGATGGTAGTTTTTTTTT
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTGTTGAAGTTCTTTTTTTTTCTTTTTTTTTTT
TTTTCTGTTTTTGGTTTTTTTTTTTTT

Timecourse_16

TSD_5prime:

AAAGAAAATGTG

TSD_3prime:

AAAGAAAATGTG

Genomic_Consensus_5p:

ATACCCAAAGATTTATAAATCATTCTGCTATAAAGACACATGCACATGTATGTGTATTG
CAGCACTGTTCGCAATAGCAAAGACTTGGAACCAACACAAATGCCCATCAATGACAG
ACTGGATAAAGAAAATGTGAGATCGGAAGAGCACA

Genomic_Consensus_3p:

AGATGACACATTAGTGGGTGCAGCGCACCAGCATGGCACATGTATACATATGTAACT
AACCTGCACAATGTGCACATGTACCCTAAAACTTAGAGTATAATAAAAAAAAAAAAAAA
AAAAAAGAAAATGTGGCACATAAACACCATGAAAA

Timecourse_17

TSD_5prime:

AAAGAAAAAGTG

TSD_3prime:

AAAGAAAAAGTG

Genomic_Consensus_5p:

CTAGCAATTACTAGGATTATAAATCATGCTACTATAAAGACACATGCACACATATGTTT
ATTGCAGCACTATTCACAATAGCAAAGACTTGCAACCAACCCAAATCCCCATCAATGA
TAGACTGGATAAAGAAAAAGTGAGATCGGAAGAG

Genomic_Consensus_3p:

AGATGACACATTAGTGGGTGCAGCGCACCAGCATGGCACATGTATACATATGTAACT
AACCTGCACAATGTGCACATGTACCCTAAAACTTAGAGTATAATAAAAAAAAAAAAAAA
AAAAAGAAAAAGTGGCACATATACACCATGAAATA

Timecourse_18

TSD_5prime:

CAGATGTCATTTCTTT

TSD_3prime:

CAGATGTCATTTCTTT

Genomic_Consensus_5p:

CCATATTTTACATAATTTATAATTTATGAAATAGAATATTTGAATAATAAACCAATTCCAT
TTTGAAGGCTTCTGATGTAGTACTCTGTTAAAAAAAAAAAAAAAAACTACTTGAGAAAA
GTATGGATTGACTATATTGGAAGTTGCAAGGCCTGAGGAATGTTTTCCCGTGATTTTA
GTCCCTCTCATCAGTGTTCTATGCCTCAGTTCTGGTAACCCCAGAATGACGCTGTTAC
CTGACAGTATTCTAATGAAGATTAAAGAAATGACATCTGAAATAATGGAGGGGAGGAG
CCAAGATGGCCGAATAGGAACAGCTCCGGTCTACAGCTCCCAGCGTGAGCGACGCA
GAAGACGGTGATTTCTGCATTTCCATCTGAGGTACCGGGTTCATCTCACTAGGGAGT
GCCAGACAGTGGGCGCAGGCCAGTGTGTGTGCGCACCGTGCGCGAGCCGAAGCAG
GGCGAGGCATTGCCTCACCTGGGAAGCGCAAGGGGTCAGGGAGTTCCCTTTCCGAG
TCAAAGAAAGGGGTGACGGACGCACCTGGAAAATCGGGTCACTCCCAC

Genomic_Consensus_3p:

ACCAAAGGAGAAAGTTTCTCTATGATGACTTTTCAACATGAAAAATAAAGTGTAGATGT
CTCAGATCATTTATCTGCTCCATTTACCAGCACCCAGCCTCATCCTGCCTCAGCGAAC
TTAATAAGATTAACAGATGTCATTTCTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
TTTTTTTTTTTTTTTTTTTTT

Timecourse_19

TSD_5prime:

NA

TSD_3prime:

NA

Genomic_Consensus_5p:

GCAGGCCAACATTCAAATTCAGGAAATTCAGAGAACACCACAAAGATACTCCTTGAGA
AGAGCAACCCCAAGACACATAATTGTCAGATTCACTAAGGTTGAAATAAGGAAAAAAT
GTTAAGGGCAAGATCGGAAGAGCGTCGTGTAGGGA

Genomic_Consensus_3p:

ACATATGCAACTAACCTGCACAATGTGCACATGTACCCTAAAACTTAGAGTATAATAAA
AAAAAAAAAAAAAAAAAAAAAACAAAATACAAGCCAGAAGAAGAAGGGGAACAATAC
TCAACATTCTTAAAGAAAAAAAATATCAACCCAG

Timecourse_20

TSD_5prime:

AAAAAAAAAAAAAAAGAAAAGAAAAGAAAAAAAA

TSD_3prime:

AAAAAAAAAAAAAAAAAAAGAAAAGAAAAAAAA

Genomic_Consensus_5p:

AATTGACAAATGGGATCTAATTAAACTAAAGAGCTTCTGCACAGCAAAAGAAACTACC
ATCAGAGTGAACAGGCAACCTACAACATGGGAGAAAATTTTCGCAACCTACTCATCTG
ACAAAGGGCTAATATCCAGAATCTACAATGAACTCAAACAAATCGGCAAGAGAAAAAC
AAGCATTCCTATCAAAACGTGTGCTAAGAACATGACTAAACAATTCTTTTTTTTCTTTTC
TTTTCTTTTTTTTTTTTTTTCTTTACGGGGCCTTGCCCTTCCGACCGGACTGG

Genomic_Consensus_3p:

CGGAACTTGCAGTGAGCCGAGAATGCGCCACTGCACTCCAGCCTGGTCGACAGAGC
AAGACTCCGTCAAGAAAAAAAAAAAAAAAAAAAAAGAAAAGAAAAAAAAAAAAAAAGT
GTTATTTCAATTTATTAAGACACATTTTAAAAAAAC

Timecourse_21

TSD_5prime:

TTTTTTTTTTTTT

TSD_3prime:

TTTTTTTTTTTTT

Genomic_Consensus_5p:

AGGGTAATGTTGGCTGGGTTTTTCTCCAAAAAATTCTTTTTTTCCTCTTTTTTTTTTTTT
ATTATACTTGAAGTCCTAGGGTACATGTGCACAATGTGCAGGTTTGTTACATATGAATA
CATGTGCCATGTTGGTGTGCTGCACCCATTAA

Genomic_Consensus_3p:

TGGCCAAGGTAATGTCTGCTAGGTTTCTCTACAGAAAATTACTATCTTCCTCTTTTTTT
TTTTTTTTTTTTTACTTTAAGTCCTGGGGCAATTGTGCCAAATTTGCGGGTTTTTTACAT
ATGATTAGAGGTCGCGTTTTGGTGGGGTGCGC

Timecourse_22

TSD_5prime:

NA

TSD_3prime:

NA

Genomic_Consensus_5p:

NA

Genomic_Consensus_3p:

GAGCCAATTTCTAGCATGGCTTATTAGGACAATGTTTTACTGTCAAGAAGTGCAGTGG
TCACTCAGTACTGCTTTTTCTAGGTGTATTTTTTTTTTTTTTTTTTCTTTTTTTTTTTTTAT
TATACTCTAAGTTTTAGGAAAAGTTCTTATTTTTAAGAGATATATACTGAGAAGTCTCAT
ATGTGCCTTTTATTGTCAATTTT

Timecourse_23

TSD_5prime:

AATTATTAATAAATAATTAAATTATTTATTAATTAATTT

TSD_3prime:

AATTATTAATAAATAATTAAATTATTTATTAATTAATTT

Genomic_Consensus_5p:

CCTTTATTTCCTTCGCCTGCCTGATTGCCCTGGCCAGAACTTCCAACACTATGTTGAA
TAGGAGTGGTGAGAGAGGGCATCCCTGTCTTGTGCCGGTTTTCAAAGGGAAATTATT
AATAAATAATTAAATTATTTATTAATTAATTTCAAGTTGACATTTAAGTTGAGGCGTGGA
GGTGAAGGAGACAGTTGGATAAATAAG

Genomic_Consensus_3p:

GACATTAATTATTAATAAATAATTAAATTATTTATTAATTAATTTATTTTTTTTTTTTTTTT
TTTTTTTTTTTTTATTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
TTTTTTTGTTATTTCTTTTTTTT

REFERENCES

1. Klawitter, S., et al., *Reprogramming triggers endogenous L1 and Alu retrotransposition in human induced pluripotent stem cells.* Nat Commun, 2016. **7**: p. 10286.
2. Wissing, S., et al., *Reprogramming somatic cells into iPS cells activates LINE-1 retroelement mobility.* Hum Mol Genet, 2012. **21**(1): p. 208-18.
3. Belancio, V.P., D.J. Hedges, and P. Deininger, *Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health.* Genome Res, 2008. **18**(3): p. 343-58.
4. Coufal, N.G., et al., *L1 retrotransposition in human neural progenitor cells.* Nature, 2009. **460**(7259): p. 1127-31.
5. Munoz-Lopez, M., et al., *Analysis of LINE-1 expression in human pluripotent cells.* Methods Mol Biol, 2012. **873**: p. 113-25.
6. Nguyen, T.H.M., et al., *L1 Retrotransposon Heterogeneity in Ovarian Tumor Cell Evolution.* Cell Rep, 2018. **23**(13): p. 3730-3740.
7. Schauer, S.N., et al., *L1 retrotransposition is a common feature of mammalian hepatocarcinogenesis.* Genome Res, 2018. **28**(5): p. 639-653.
8. Tubio, J.M.C., et al., *Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes.* Science, 2014. **345**(6196): p. 1251343.
9. Scott, E.C., et al., *A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer.* Genome Res, 2016. **26**(6): p. 745-55.
10. Macia, A., et al., *Engineered LINE-1 retrotransposition in nondividing human neurons.* Genome Res, 2017. **27**(3): p. 335-348.
11. Beck, C.R., et al., *LINE-1 retrotransposition activity in human genomes.* Cell, 2010. **141**(7): p. 1159-70.
12. Macfarlane, C.M., et al., *Transduction-specific ATLAS reveals a cohort of highly active L1 retrotransposons in human populations.* Hum Mutat, 2013. **34**(7): p. 974-85.
13. Brouha, B., et al., *Hot L1s account for the bulk of retrotransposition in the human population.* Proc Natl Acad Sci U S A, 2003. **100**(9): p. 5280-5.
14. Ahl, V., et al., *Retrotransposition and Crystal Structure of an Alu RNP in the Ribosome-Stalling Conformation.* Mol Cell, 2015. **60**(5): p. 715-727.
15. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
16. Ostertag, E.M. and H.H. Kazazian, Jr., *Biology of mammalian L1 retrotransposons.* Annu Rev Genet, 2001. **35**: p. 501-38.
17. Beck, C.R., et al., *LINE-1 elements in structural variation and disease.* Annu Rev Genomics Hum Genet, 2011. **12**: p. 187-215.
18. Bannert, N. and R. Kurth, *The evolutionary dynamics of human endogenous retroviral families.* Annu Rev Genomics Hum Genet, 2006. **7**: p. 149-73.
19. Boeke, J.D., *LINEs and Alus--the polyA connection.* Nat Genet, 1997. **16**(1): p. 6-7.
20. Ostertag, E.M., et al., *SVA elements are nonautonomous retrotransposons that cause disease in humans.* Am J Hum Genet, 2003. **73**(6): p. 1444-51.
21. Kajikawa, M. and N. Okada, *LINEs mobilize SINEs in the eel through a shared 3' sequence.* Cell, 2002. **111**(3): p. 433-44.
22. Moran, J.V., et al., *High frequency retrotransposition in cultured mammalian cells.* Cell, 1996. **87**(5): p. 917-27.
23. McCullers, T.J. and M. Steiniger, *Transposable elements in Drosophila.* Mob Genet Elements, 2017. **7**(3): p. 1-18.

24. Laricchia, K.M., et al., *Natural Variation in the Distribution and Abundance of Transposable Elements Across the Caenorhabditis elegans Species.* Mol Biol Evol, 2017. **34**(9): p. 2187-2202.
25. Lee, S.I. and N.S. Kim, *Transposable elements and genome size variations in plants.* Genomics Inform, 2014. **12**(3): p. 87-97.
26. Dombroski, B.A., et al., *Isolation of an active human transposable element.* Science, 1991. **254**(5039): p. 1805-8.
27. Dombroski, B.A., A.F. Scott, and H.H. Kazazian, Jr., *Two additional potential retrotransposons isolated from a human L1 subfamily that contains an active retrotransposable element.* Proc Natl Acad Sci U S A, 1993. **90**(14): p. 6513-7.
28. Sassaman, D.M., et al., *Many human L1 elements are capable of retrotransposition.* Nat Genet, 1997. **16**(1): p. 37-43.
29. Brouha, B., et al., *Evidence consistent with human L1 retrotransposition in maternal meiosis I.* Am J Hum Genet, 2002. **71**(2): p. 327-36.
30. Myers, J.S., et al., *A comprehensive analysis of recently integrated human Ta L1 elements.* Am J Hum Genet, 2002. **71**(2): p. 312-26.
31. Ostertag, E.M., et al., *Determination of L1 retrotransposition kinetics in cultured cells.* Nucleic Acids Res, 2000. **28**(6): p. 1418-23.
32. Lutz, S.M., et al., *Allelic heterogeneity in LINE-1 retrotransposition activity.* Am J Hum Genet, 2003. **73**(6): p. 1431-7.
33. Goodman, M., et al., *Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence.* Mol Phylogenet Evol, 1998. **9**(3): p. 585-98.
34. Skowronski, J., T.G. Fanning, and M.F. Singer, *Unit-length line-1 transcripts in human teratocarcinoma cells.* Mol Cell Biol, 1988. **8**(4): p. 1385-97.
35. Scott, A.F., et al., *Origin of the human L1 elements: proposed progenitor genes deduced from a consensus DNA sequence.* Genomics, 1987. **1**(2): p. 113-25.
36. Swergold, G.D., *Identification, characterization, and cell specificity of a human LINE-1 promoter.* Mol Cell Biol, 1990. **10**(12): p. 6718-29.
37. Yang, N., et al., *An important role for RUNX3 in human L1 transcription and retrotransposition.* Nucleic Acids Res, 2003. **31**(16): p. 4929-40.
38. Tchenio, T., J.F. Casella, and T. Heidmann, *Members of the SRY family regulate the human LINE retrotransposons.* Nucleic Acids Res, 2000. **28**(2): p. 411-5.
39. Minakami, R., et al., *Identification of an internal cis-element essential for the human L1 transcription and a nuclear factor(s) binding to the element.* Nucleic Acids Res, 1992. **20**(12): p. 3139-45.
40. Kuwabara, T., et al., *Wnt-mediated activation of NeuroD1 and retro-elements during adult neurogenesis.* Nat Neurosci, 2009. **12**(9): p. 1097-105.
41. Becker, K.G., et al., *Binding of the ubiquitous nuclear transcription factor YY1 to a cis regulatory sequence in the human LINE-1 transposable element.* Hum Mol Genet, 1993. **2**(10): p. 1697-702.
42. Athanikar, J.N., R.M. Badge, and J.V. Moran, *A YY1-binding site is required for accurate human LINE-1 transcription initiation.* Nucleic Acids Res, 2004. **32**(13): p. 3846-55.
43. Speek, M., *Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes.* Mol Cell Biol, 2001. **21**(6): p. 1973-85.
44. Nigumann, P., et al., *Many human genes are transcribed from the antisense promoter of L1 retrotransposon.* Genomics, 2002. **79**(5): p. 628-34.
45. Faulkner, G.J., et al., *The regulated retrotransposon transcriptome of mammalian cells.* Nat Genet, 2009. **41**(5): p. 563-71.
46. Denli, A.M., et al., *Primate-specific ORF0 contributes to retrotransposon-mediated diversity.* Cell, 2015. **163**(3): p. 583-93.

47. Martin, S.L. and F.D. Bushman, *Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon.* Mol Cell Biol, 2001. **21**(2): p. 467-75.
48. Hohjoh, H. and M.F. Singer, *Sequence-specific single-strand RNA binding protein encoded by the human LINE-1 retrotransposon.* EMBO J, 1997. **16**(19): p. 6034-43.
49. Feng, Q., et al., *Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition.* Cell, 1996. **87**(5): p. 905-16.
50. Mathias, S.L., et al., *Reverse transcriptase encoded by a human transposable element.* Science, 1991. **254**(5039): p. 1808-10.
51. Dmitriev, S.E., et al., *Efficient translation initiation directed by the 900-nucleotide-long and GC-rich 5' untranslated region of the human retrotransposon LINE-1 mRNA is strictly cap dependent rather than internal ribosome entry site mediated.* Mol Cell Biol, 2007. **27**(13): p. 4685-97.
52. Alisch, R.S., et al., *Unconventional translation of mammalian LINE-1 retrotransposons.* Genes Dev, 2006. **20**(2): p. 210-24.
53. Wei, W., et al., *Human L1 retrotransposition: cis preference versus trans complementation.* Mol Cell Biol, 2001. **21**(4): p. 1429-39.
54. Kazazian, H.H., Jr. and J.V. Moran, *The impact of L1 retrotransposons on the human genome.* Nat Genet, 1998. **19**(1): p. 19-24.
55. Kulpa, D.A. and J.V. Moran, *Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles.* Nat Struct Mol Biol, 2006. **13**(7): p. 655-60.
56. Kulpa, D.A. and J.V. Moran, *Ribonucleoprotein particle formation is necessary but not sufficient for LINE-1 retrotransposition.* Hum Mol Genet, 2005. **14**(21): p. 3237-48.
57. Hohjoh, H. and M.F. Singer, *Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA.* EMBO J, 1996. **15**(3): p. 630-9.
58. Esnault, C., J. Maestre, and T. Heidmann, *Human LINE retrotransposons generate processed pseudogenes.* Nat Genet, 2000. **24**(4): p. 363-7.
59. Kubo, S., et al., *L1 retrotransposition in nondividing and primary human somatic cells.* Proc Natl Acad Sci U S A, 2006. **103**(21): p. 8036-41.
60. Luan, D.D., et al., *Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition.* Cell, 1993. **72**(4): p. 595-605.
61. Jurka, J., *Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons.* Proc Natl Acad Sci U S A, 1997. **94**(5): p. 1872-7.
62. Cost, G.J., et al., *Human L1 element target-primed reverse transcription in vitro.* EMBO J, 2002. **21**(21): p. 5899-910.
63. Szak, S.T., et al., *Molecular archeology of L1 insertions in the human genome.* Genome Biol, 2002. **3**(10): p. research0052.
64. Gilbert, N., S. Lutz-Prigge, and J.V. Moran, *Genomic deletions created upon LINE-1 retrotransposition.* Cell, 2002. **110**(3): p. 315-25.
65. Morrish, T.A., et al., *DNA repair mediated by endonuclease-independent LINE-1 retrotransposition.* Nat Genet, 2002. **31**(2): p. 159-65.
66. Sen, S.K., et al., *Endonuclease-independent insertion provides an alternative pathway for L1 retrotransposition in the human genome.* Nucleic Acids Res, 2007. **35**(11): p. 3741-51.
67. Goodier, J.L. and H.H. Kazazian, Jr., *Retrotransposons revisited: the restraint and rehabilitation of parasites.* Cell, 2008. **135**(1): p. 23-35.
68. Hancks, D.C. and H.H. Kazazian, Jr., *Active human retrotransposons: variation and disease.* Curr Opin Genet Dev, 2012. **22**(3): p. 191-203.
69. Sayah, D.M., et al., *Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1.* Nature, 2004. **430**(6999): p. 569-73.

70. Kazazian, H.H., Jr., et al., *Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man.* Nature, 1988. **332**(6160): p. 164-6.
71. Hancks, D.C. and H.H. Kazazian, Jr., *Roles for retrotransposon insertions in human disease.* Mob DNA, 2016. **7**: p. 9.
72. Kazazian, H.H., Jr. and J.V. Moran, *Mobile DNA in Health and Disease.* N Engl J Med, 2017. **377**(4): p. 361-370.
73. Belancio, V.P., A.M. Roy-Engel, and P. Deininger, *The impact of multiple splice sites in human L1 elements.* Gene, 2008. **411**(1-2): p. 38-45.
74. Moran, J.V., R.J. DeBerardinis, and H.H. Kazazian, Jr., *Exon shuffling by L1 retrotransposition.* Science, 1999. **283**(5407): p. 1530-4.
75. Lin, C., et al., *Nuclear receptor-induced chromosomal proximity and DNA breaks underlie specific translocations in cancer.* Cell, 2009. **139**(6): p. 1069-83.
76. Gasior, S.L., et al., *The human LINE-1 retrotransposon creates DNA double-strand breaks.* J Mol Biol, 2006. **357**(5): p. 1383-93.
77. Perepelitsa-Belancio, V. and P. Deininger, *RNA truncation by premature polyadenylation attenuates human mobile element activity.* Nat Genet, 2003. **35**(4): p. 363-6.
78. Han, J.S., S.T. Szak, and J.D. Boeke, *Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes.* Nature, 2004. **429**(6989): p. 268-74.
79. Wheelan, S.J., et al., *Gene-breaking: a new paradigm for human retrotransposon-mediated gene evolution.* Genome Res, 2005. **15**(8): p. 1073-8.
80. Thomas, C.A., A.C. Paquola, and A.R. Muotri, *LINE-1 retrotransposition in the nervous system.* Annu Rev Cell Dev Biol, 2012. **28**: p. 555-73.
81. Goodier, J.L., E.M. Ostertag, and H.H. Kazazian, Jr., *Transduction of 3'-flanking sequences is common in L1 retrotransposition.* Hum Mol Genet, 2000. **9**(4): p. 653-7.
82. Pickeral, O.K., et al., *Frequent human genomic DNA transduction driven by LINE-1 retrotransposition.* Genome Res, 2000. **10**(4): p. 411-5.
83. Moran, J.V., *Human L1 retrotransposition: insights and peculiarities learned from a cultured cell retrotransposition assay.* Genetica, 1999. **107**(1-3): p. 39-51.
84. Holmes, S.E., et al., *A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion.* Nat Genet, 1994. **7**(2): p. 143-8.
85. Miki, Y., et al., *Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer.* Cancer Res, 1992. **52**(3): p. 643-5.
86. Lee, E., et al., *Landscape of somatic retrotransposition in human cancers.* Science, 2012. **337**(6097): p. 967-71.
87. Gardner, E.J., et al., *The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology.* Genome Res, 2017. **27**(11): p. 1916-1929.
88. Richardson, S.R., et al., *Heritable L1 retrotransposition in the mouse primordial germline and early embryo.* Genome Res, 2017. **27**(8): p. 1395-1405.
89. Evrony, G.D., et al., *Cell lineage analysis in human brain using endogenous retroelements.* Neuron, 2015. **85**(1): p. 49-59.
90. Iskow, R.C., et al., *Natural mutagenesis of human genomes by endogenous retrotransposons.* Cell, 2010. **141**(7): p. 1253-61.
91. Shukla, R., et al., *Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma.* Cell, 2013. **153**(1): p. 101-11.
92. Solyom, S., et al., *Extensive somatic L1 retrotransposition in colorectal tumors.* Genome Research, 2012. **22**(12): p. 2328-2338.

93. Gilbert, N., et al., *Multiple fates of L1 retrotransposition intermediates in cultured human cells.* Mol Cell Biol, 2005. **25**(17): p. 7780-95.
94. Boissinot, S., P. Chevret, and A.V. Furano, *L1 (LINE-1) retrotransposon evolution and amplification in recent human history.* Mol Biol Evol, 2000. **17**(6): p. 915-28.
95. Marchani, E.E., et al., *Estimating the age of retrotransposon subfamilies using maximum likelihood.* Genomics, 2009. **94**(1): p. 78-82.
96. Kinzler, K.W. and B. Vogelstein, *Lessons from hereditary colorectal cancer.* Cell, 1996. **87**(2): p. 159-70.
97. Fearon, E.R., *Molecular genetics of colorectal cancer.* Annu Rev Pathol, 2011. **6**: p. 479-507.
98. Zhang, W., et al., *Comparing genetic variants detected in the 1000 genomes project with SNPs determined by the International HapMap Consortium.* J Genet, 2015. **94**(4): p. 731-40.
99. Kimberland, M.L., et al., *Full-length human L1 insertions retain the capacity for high frequency retrotransposition in cultured cells.* Hum Mol Genet, 1999. **8**(8): p. 1557-60.
100. Badge, R.M., R.S. Alisch, and J.V. Moran, *ATLAS: a system to selectively identify human-specific L1 insertions.* Am J Hum Genet, 2003. **72**(4): p. 823-38.
101. Seleme, M.C., et al., *Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity.* Proc Natl Acad Sci U S A, 2006. **103**(17): p. 6611-6.
102. Schwahn, U., et al., *Positional cloning of the gene for X-linked retinitis pigmentosa 2.* Nat Genet, 1998. **19**(4): p. 327-32.
103. van den Hurk, J.A., et al., *L1 retrotransposition can occur early in human embryonic development.* Hum Mol Genet, 2007. **16**(13): p. 1587-92.
104. Evrony, G.D., et al., *Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain.* Cell, 2012. **151**(3): p. 483-96.
105. Philippe, C., et al., *Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci.* Elife, 2016. **5**.
106. Sanchez-Luque, F.J., et al., *LINE-1 Evasion of Epigenetic Repression in Humans.* Mol Cell, 2019. **75**(3): p. 590-604 e12.
107. Baillie, J.K., et al., *Somatic retrotransposition alters the genetic landscape of the human brain.* Nature, 2011. **479**(7374): p. 534-7.
108. Singer, T., et al., *LINE-1 retrotransposons: mediators of somatic variation in neuronal genomes?* Trends Neurosci, 2010. **33**(8): p. 345-54.
109. Levin, H.L. and J.V. Moran, *Dynamic interactions between transposable elements and their hosts.* Nat Rev Genet, 2011. **12**(9): p. 615-27.
110. Ostertag, E.M., et al., *A mouse model of human L1 retrotransposition.* Nat Genet, 2002. **32**(4): p. 655-60.
111. Kano, H., et al., *L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism.* Genes Dev, 2009. **23**(11): p. 1303-12.
112. Ewing, A.D. and H.H. Kazazian, Jr., *High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes.* Genome Res, 2010. **20**(9): p. 1262-70.
113. Ewing, A.D. and H.H. Kazazian, Jr., *Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans.* Genome Res, 2011. **21**(6): p. 985-90.
114. Thomson, J.A., et al., *Embryonic stem cell lines derived from human blastocysts.* Science, 1998. **282**(5391): p. 1145-7.
115. Garcia-Perez, J.L., et al., *LINE-1 retrotransposition in human embryonic stem cells.* Hum Mol Genet, 2007. **16**(13): p. 1569-77.

116. Wissing, S., et al., *Endogenous APOBEC3B restricts LINE-1 retrotransposition in transformed cells and human embryonic stem cells.* J Biol Chem, 2011. **286**(42): p. 36427-37.
117. Takahashi, K., et al., *Induction of pluripotent stem cells from adult human fibroblasts by defined factors.* Cell, 2007. **131**(5): p. 861-72.
118. Yamanaka, S., *Induced pluripotent stem cells: past, present, and future.* Cell Stem Cell, 2012. **10**(6): p. 678-84.
119. Gore, A., et al., *Somatic coding mutations in human induced pluripotent stem cells.* Nature, 2011. **471**(7336): p. 63-7.
120. Hussein, S.M., et al., *Copy number variation and selection during reprogramming to pluripotency.* Nature, 2011. **471**(7336): p. 58-62.
121. Laurent, L.C., et al., *Dynamic changes in the copy number of pluripotency and cell proliferation genes in human ESCs and iPSCs during reprogramming and time in culture.* Cell Stem Cell, 2011. **8**(1): p. 106-18.
122. Lister, R., et al., *Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells.* Nature, 2011. **471**(7336): p. 68-73.
123. Mayshar, Y., et al., *Identification and classification of chromosomal aberrations in human induced pluripotent stem cells.* Cell Stem Cell, 2010. **7**(4): p. 521-31.
124. Maherali, N., et al., *Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution.* Cell Stem Cell, 2007. **1**(1): p. 55-70.
125. Bhutani, K., et al., *Whole-genome mutational burden analysis of three pluripotency induction methods.* Nat Commun, 2016. **7**: p. 10536.
126. Arokium, H., et al., *Deep sequencing reveals low incidence of endogenous LINE-1 retrotransposition in human induced pluripotent stem cells.* PLoS One, 2014. **9**(10): p. e108682.
127. Faulkner, G.J. and J.L. Garcia-Perez, *L1 Mosaicism in Mammals: Extent, Effects, and Evolution.* Trends Genet, 2017. **33**(11): p. 802-816.
128. Obergasteiger, J., et al., *CADPS2 gene expression is oppositely regulated by LRRK2 and alpha-synuclein.* Biochem Biophys Res Commun, 2017. **490**(3): p. 876-881.
129. Li, Y., et al., *Personalized Medicine: Cell and Gene Therapy Based on Patient-Specific iPSC-Derived Retinal Pigment Epithelium Cells.* Adv Exp Med Biol, 2016. **854**: p. 549-55.
130. Seki, T. and K. Fukuda, *Methods of induced pluripotent stem cells for clinical application.* World J Stem Cells, 2015. **7**(1): p. 116-25.
131. Lin, K. and A.Z. Xiao, *Quality control towards the application of induced pluripotent stem cells.* Curr Opin Genet Dev, 2017. **46**: p. 164-169.
132. Nakagawa, M., et al., *Promotion of direct reprogramming by transformation-deficient Myc.* Proc Natl Acad Sci U S A, 2010. **107**(32): p. 14152-7.
133. Omole, A.E. and A.O.J. Fakoya, *Ten years of progress and promise of induced pluripotent stem cells: historical origins, characteristics, mechanisms, limitations, and potential applications.* PeerJ, 2018. **6**: p. e4370.
134. Mc, C.B., *The origin and behavior of mutable loci in maize.* Proc Natl Acad Sci U S A, 1950. **36**(6): p. 344-55.
135. An, W., et al., *Active retrotransposition by a synthetic L1 element in mice.* Proc Natl Acad Sci U S A, 2006. **103**(49): p. 18662-7.
136. Muotri, A.R., et al., *Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition.* Nature, 2005. **435**(7044): p. 903-10.
137. Gage, F.H., *Mammalian neural stem cells.* Science, 2000. **287**(5457): p. 1433-8.
138. Gotz, M. and W.B. Huttner, *The cell biology of neurogenesis.* Nat Rev Mol Cell Biol, 2005. **6**(10): p. 777-88.

139. Fasolino, M. and Z. Zhou, *The Crucial Role of DNA Methylation and MeCP2 in Neuronal Function.* Genes (Basel), 2017. **8**(5).
140. Amir, R.E., et al., *Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2.* Nat Genet, 1999. **23**(2): p. 185-8.
141. Muotri, A.R., et al., *L1 retrotransposition in neurons is modulated by MeCP2.* Nature, 2010. **468**(7322): p. 443-6.
142. Yu, F., et al., *Methyl-CpG-binding protein 2 represses LINE-1 expression and retrotransposition but not Alu transcription.* Nucleic Acids Res, 2001. **29**(21): p. 4493-501.
143. Upton, K.R., et al., *Ubiquitous l1 mosaicism in hippocampal neurons.* Cell, 2015. **161**(2): p. 228-39.
144. Erwin, J.A., et al., *L1-associated genomic regions are deleted in somatic cells of the healthy human brain.* Nat Neurosci, 2016. **19**(12): p. 1583-1591.
145. Flasch, D.A., et al., *Genome-wide de novo L1 Retrotransposition Connects Endonuclease Activity with Replication.* Cell, 2019. **177**(4): p. 837-851 e28.
146. Sultana, T., et al., *The Landscape of L1 Retrotransposons in the Human Genome Is Shaped by Pre-insertion Sequence Biases and Post-insertion Selection.* Mol Cell, 2019. **74**(3): p. 555-570 e7.
147. Hazen, J.L., et al., *The Complete Genome Sequences, Unique Mutational Spectra, and Developmental Potency of Adult Neurons Revealed by Cloning.* Neuron, 2016. **89**(6): p. 1223-1236.
148. Youssoufian, H. and R.E. Pyeritz, *Mechanisms and consequences of somatic mosaicism in humans.* Nat Rev Genet, 2002. **3**(10): p. 748-58.
149. Hozumi, N. and S. Tonegawa, *Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions.* Proc Natl Acad Sci U S A, 1976. **73**(10): p. 3628-32.
150. Market, E. and F.N. Papavasiliou, *V(D)J recombination and the evolution of the adaptive immune system.* PLoS Biol, 2003. **1**(1): p. E16.
151. Cai, X., et al., *Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain.* Cell Rep, 2014. **8**(5): p. 1280-9.
152. Gole, J., et al., *Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells.* Nat Biotechnol, 2013. **31**(12): p. 1126-32.
153. McConnell, M.J., et al., *Mosaic copy number variation in human neurons.* Science, 2013. **342**(6158): p. 632-7.
154. Lee, J., et al., *Different evolutionary fates of recently integrated human and chimpanzee LINE-1 retrotransposons.* Gene, 2007. **390**(1-2): p. 18-27.
155. Faulkner, G.J. and V. Billon, *L1 retrotransposition in the soma: a field jumping ahead.* Mob DNA, 2018. **9**: p. 22.
156. Jonsson, M.E., et al., *Activation of neuronal genes via LINE-1 elements upon global DNA demethylation in human neural progenitors.* Nat Commun, 2019. **10**(1): p. 3182.
157. Bundo, M., et al., *Increased l1 retrotransposition in the neuronal genome in schizophrenia.* Neuron, 2014. **81**(2): p. 306-13.
158. Bodea, G.O., E.G.Z. McKelvey, and G.J. Faulkner, *Retrotransposon-induced mosaicism in the neural genome.* Open Biol, 2018. **8**(7).
159. Marchetto, M.C., et al., *Differential L1 regulation in pluripotent stem cells of humans and apes.* Nature, 2013. **503**(7477): p. 525-9.
160. Prak, E.T., et al., *Tracking an embryonic L1 retrotransposition event.* Proc Natl Acad Sci U S A, 2003. **100**(4): p. 1832-7.
161. Babushok, D.V., et al., *L1 integration in a transgenic mouse model.* Genome Res, 2006. **16**(2): p. 240-50.

162. An, W., et al., *Conditional activation of a single-copy L1 transgene in mice by Cre.* Genesis, 2008. **46**(7): p. 373-83.
163. Newkirk, S.J., et al., *Intact piRNA pathway prevents L1 mobilization in male meiosis.* Proc Natl Acad Sci U S A, 2017. **114**(28): p. E5635-E5644.
164. Garcia-Perez, J.L., et al., *Epigenetic silencing of engineered L1 retrotransposition events in human embryonic carcinoma cells.* Nature, 2010. **466**(7307): p. 769-73.
165. Tang, Z., et al., *Human transposon insertion profiling: Analysis, visualization and identification of somatic LINE-1 insertions in ovarian cancer.* Proc Natl Acad Sci U S A, 2017. **114**(5): p. E733-E740.
166. Richardson, S.R., S. Morell, and G.J. Faulkner, *L1 retrotransposons and somatic mosaicism in the brain.* Annu Rev Genet, 2014. **48**: p. 1-27.
167. Goodier, J.L., *Restricting retrotransposons: a review.* Mob DNA, 2016. **7**: p. 16.
168. McConnell, M.J., et al., *Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network.* Science, 2017. **356**(6336).
169. Mir, A.A., C. Philippe, and G. Cristofari, *euL1db: the European database of L1HS retrotransposon insertions in humans.* Nucleic Acids Res, 2015. **43**(Database issue): p. D43-7.
170. Sudmant, P.H., et al., *An integrated map of structural variation in 2,504 human genomes.* Nature, 2015. **526**(7571): p. 75-81.
171. Wang, J., et al., *dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans.* Hum Mutat, 2006. **27**(4): p. 323-9.
172. Ostertag, E.M. and H.H. Kazazian, Jr., *Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition.* Genome Res, 2001. **11**(12): p. 2059-65.
173. Carreira, P.E., et al., *Evidence for L1-associated DNA rearrangements and negligible L1 retrotransposition in glioblastoma multiforme.* Mobile DNA, 2016. **7**.
174. Solyom, S., et al., *Pathogenic orphan transduction created by a nonreference LINE-1 retrotransposon.* Hum Mutat, 2012. **33**(2): p. 369-71.
175. Doucet-O'Hare, T.T., et al., *LINE-1 expression and retrotransposition in Barrett's esophagus and esophageal carcinoma.* Proc Natl Acad Sci U S A, 2015. **112**(35): p. E4894-900.
176. Helman, E., et al., *Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing.* Genome Res, 2014. **24**(7): p. 1053-63.
177. Rodic, N., et al., *Long interspersed element-1 protein expression is a hallmark of many human cancers.* Am J Pathol, 2014. **184**(5): p. 1280-6.
178. Hata, K. and Y. Sakaki, *Identification of critical CpG sites for repression of L1 transcription by DNA methylation.* Gene, 1997. **189**(2): p. 227-34.
179. Macia, A., et al., *Epigenetic control of retrotransposon expression in human embryonic stem cells.* Mol Cell Biol, 2011. **31**(2): p. 300-16.
180. Liang, G., et al., *Cooperativity between DNA methyltransferases in the maintenance methylation of repetitive elements.* Mol Cell Biol, 2002. **22**(2): p. 480-91.
181. Bojang, P., Jr. and K.S. Ramos, *Epigenetic reactivation of LINE-1 retrotransposon disrupts NuRD corepressor functions and induces oncogenic transformation in human bronchial epithelial cells.* Mol Oncol, 2018. **12**(8): p. 1342-1357.
182. Montoya-Durango, D.E., et al., *Epigenetic control of mammalian LINE-1 retrotransposon by retinoblastoma proteins.* Mutat Res, 2009. **665**(1-2): p. 20-8.
183. Montoya-Durango, D.E., et al., *LINE-1 silencing by retinoblastoma proteins is effected through the nucleosomal and remodeling deacetylase multiprotein complex.* BMC Cancer, 2016. **16**: p. 38.
184. Bourc'his, D. and T.H. Bestor, *Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L.* Nature, 2004. **431**(7004): p. 96-9.

185. Harris, C.R., et al., *p53 responsive elements in human retrotransposons.* Oncogene, 2009. **28**(44): p. 3857-65.
186. Trono, D., *Transposable Elements, Polydactyl Proteins, and the Genesis of Human-Specific Transcription Networks.* Cold Spring Harb Symp Quant Biol, 2015. **80**: p. 281-8.
187. Van Meter, M., et al., *SIRT6 represses LINE1 retrotransposons by ribosylating KAP1 but this repression fails with stress and age.* Nat Commun, 2014. **5**: p. 5011.
188. Puszyk, W., et al., *The epigenetic regulator PLZF represses L1 retrotransposition in germ and progenitor cells.* EMBO J, 2013. **32**(13): p. 1941-52.
189. Hamdorf, M., et al., *miR-128 represses L1 retrotransposition by binding directly to L1 RNA.* Nat Struct Mol Biol, 2015. **22**(10): p. 824-31.
190. Heras, S.R., et al., *The Microprocessor controls the activity of mammalian retrotransposons.* Nat Struct Mol Biol, 2013. **20**(10): p. 1173-81.
191. Aravin, A.A., G.J. Hannon, and J. Brennecke, *The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race.* Science, 2007. **318**(5851): p. 761-4.
192. Aravin, A.A., et al., *A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice.* Mol Cell, 2008. **31**(6): p. 785-99.
193. Frost, R.J., et al., *MOV10L1 is necessary for protection of spermatocytes against retrotransposons by Piwi-interacting RNAs.* Proc Natl Acad Sci U S A, 2010. **107**(26): p. 11847-52.
194. Cook, P.R., C.E. Jones, and A.V. Furano, *Phosphorylation of ORF1p is required for L1 retrotransposition.* Proc Natl Acad Sci U S A, 2015. **112**(14): p. 4298-303.
195. Chen, H., et al., *APOBEC3A is a potent inhibitor of adeno-associated virus and retrotransposons.* Curr Biol, 2006. **16**(5): p. 480-5.
196. Bogerd, H.P., et al., *Cellular inhibitors of long interspersed element 1 and Alu retrotransposition.* Proc Natl Acad Sci U S A, 2006. **103**(23): p. 8780-5.
197. Richardson, S.R., et al., *APOBEC3A deaminates transiently exposed single-strand DNA during LINE-1 retrotransposition.* Elife, 2014. **3**: p. e02008.
198. Benitez-Guijarro, M., et al., *RNase H2, mutated in Aicardi-Goutieres syndrome, promotes LINE-1 retrotransposition.* EMBO J, 2018. **37**(15).
199. Hu, S., et al., *SAMHD1 Inhibits LINE-1 Retrotransposition by Promoting Stress Granule Formation.* PLoS Genet, 2015. **11**(7): p. e1005367.
200. Richardson, S.R. and G.J. Faulkner, *Heritable L1 Retrotransposition Events During Development: Understanding Their Origins: Examination of heritable, endogenous L1 retrotransposition in mice opens up exciting new questions and research directions.* Bioessays, 2018. **40**(6): p. e1700189.
201. Cordaux, R. and M.A. Batzer, *The impact of retrotransposons on human genome evolution.* Nat Rev Genet, 2009. **10**(10): p. 691-703.
202. Georgiou, I., et al., *Retrotransposon RNA expression and evidence for retrotransposition events in human oocytes.* Hum Mol Genet, 2009. **18**(7): p. 1221-8.
203. Briggs, J.A., et al., *Integration-free induced pluripotent stem cells model genetic and neural developmental features of down syndrome etiology.* Stem Cells, 2013. **31**(3): p. 467-78.
204. Kuhn, A., et al., *Linkage disequilibrium and signatures of positive selection around LINE-1 retrotransposons in the human genome.* Proc Natl Acad Sci U S A, 2014. **111**(22): p. 8131-6.
205. Stewart, C., et al., *A comprehensive map of mobile element insertion polymorphisms in humans.* PLoS Genet, 2011. **7**(8): p. e1002236.
206. Grandi, F.C. and W. An, *Non-LTR retrotransposons and microsatellites: Partners in genomic variation.* Mob Genet Elements, 2013. **3**(4): p. e25674.

207. Kent, W.J., *BLAT--the BLAST-like alignment tool.* Genome Res, 2002. **12**(4): p. 656-64.
208. Pierrou, S., et al., *Cloning and characterization of seven human forkhead proteins: binding site specificity and DNA bending.* EMBO J, 1994. **13**(20): p. 5002-12.
209. Lavie, L., et al., *The human L1 promoter: variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity.* Genome Res, 2004. **14**(11): p. 2253-60.
210. Medhi, D., A.S. Goldman, and M. Lichten, *Local chromosome context is a major determinant of crossover pathway biochemistry during budding yeast meiosis.* Elife, 2016. **5**.
211. Khazina, E., et al., *Trimeric structure and flexibility of the L1ORF1 protein in human L1 retrotransposition.* Nat Struct Mol Biol, 2011. **18**(9): p. 1006-14.
212. Weichenrieder, O., K. Repanas, and A. Perrakis, *Crystal structure of the targeting endonuclease of the human LINE-1 retrotransposon.* Structure, 2004. **12**(6): p. 975-86.
213. Yoder, J.A., C.P. Walsh, and T.H. Bestor, *Cytosine methylation and the ecology of intragenomic parasites.* Trends Genet, 1997. **13**(8): p. 335-40.
214. Bestor, T.H., *Cytosine methylation mediates sexual conflict.* Trends Genet, 2003. **19**(4): p. 185-90.
215. de la Rica, L., et al., *TET-dependent regulation of retrotransposable elements in mouse embryonic stem cells.* Genome Biol, 2016. **17**(1): p. 234.
216. Castro-Diaz, N., et al., *Evolutionally dynamic L1 regulation in embryonic stem cells.* Genes Dev, 2014. **28**(13): p. 1397-409.
217. Macia, A., E. Blanco-Jimenez, and J.L. Garcia-Perez, *Retrotransposons in pluripotent cells: Impact and new roles in cellular plasticity.* Biochim Biophys Acta, 2015. **1849**(4): p. 417-426.
218. Rowitch, D.H. and A.R. Kriegstein, *Developmental genetics of vertebrate glial-cell specification.* Nature, 2010. **468**(7321): p. 214-22.
219. Sun, Y.E., K. Martinowich, and W. Ge, *Making and repairing the mammalian brain--signaling toward neurogenesis and gliogenesis.* Semin Cell Dev Biol, 2003. **14**(3): p. 161-8.
220. Fan, G., et al., *DNA methylation controls the timing of astrogliogenesis through regulation of JAK-STAT signaling.* Development, 2005. **132**(15): p. 3345-56.
221. Lisanti, S., et al., *Comparison of methods for quantification of global DNA methylation in human cells and tissues.* PLoS One, 2013. **8**(11): p. e79044.
222. Kannan, M., et al., *Dynamic silencing of somatic L1 retrotransposon insertions reflects the developmental and cellular contexts of their genomic integration.* Mob DNA, 2017. **8**: p. 8.
223. Friedli, M., et al., *Loss of transcriptional control over endogenous retroelements during reprogramming to pluripotency.* Genome Res, 2014. **24**(8): p. 1251-9.
224. Saini, N., et al., *The Impact of Environmental and Endogenous Damage on Somatic Mutation Load in Human Skin Fibroblasts.* PLoS Genet, 2016. **12**(10): p. e1006385.
225. Brennecke, J., et al., *Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila.* Cell, 2007. **128**(6): p. 1089-103.
226. Liu, X., et al., *Comprehensive characterization of distinct states of human naive pluripotency generated by reprogramming.* Nat Methods, 2017. **14**(11): p. 1055-1062.
227. Yeo, N.C., et al., *An enhanced CRISPR repressor for targeted mammalian gene regulation.* Nat Methods, 2018. **15**(8): p. 611-616.
228. Witherspoon, D.J., et al., *Mobile element scanning (ME-Scan) by targeted high-throughput sequencing.* BMC Genomics, 2010. **11**: p. 410.

229. Witherspoon, D.J., et al., *Mobile element scanning (ME-Scan) identifies thousands of novel Alu insertions in diverse human populations.* Genome Res, 2013. **23**(7): p. 1170-81.
230. Grimaldi, G., J. Skowronski, and M.F. Singer, *Defining the beginning and end of KpnI family segments.* EMBO J, 1984. **3**(8): p. 1753-9.
231. Carreira, P.E., S.R. Richardson, and G.J. Faulkner, *L1 retrotransposons, cancer stem cells and oncogenesis.* FEBS J, 2014. **281**(1): p. 63-73.
232. Uchida, N., et al., *Direct isolation of human central nervous system stem cells.* Proc Natl Acad Sci U S A, 2000. **97**(26): p. 14720-5.
233. Kumaki, Y., M. Oda, and M. Okano, *QUMA: quantification tool for methylation analysis.* Nucleic Acids Res, 2008. **36**(Web Server issue): p. W170-5.