Activity of endogenous L1 retrotransposons in human embryonal cells

Thesis submitted for the degree of

Doctor of Philosophy

at the University of Leicester

by

Raheleh Rahbari

Department of Genetics

University of Leicester

2011

Abstract

Activity of endogenous L1 retrotransposons in human embryonal cells
Raheleh Rahbari, 2011

Recent high throughput studies have led to the discovery of *de novo* L1 retrotransposition in malignant somatic cells, as well as large numbers of novel insertions, many of which are highly active in cell culture assays. These data suggest that L1 elements are robustly active, undergoing extensive diversification in contemporary human genomes. Despite this there is little direct evidence of endogenous L1 retrotransposition in the human germline or early embryogenesis: data from very rare disease causing insertions is indirect, subject to strong acquisition bias, and is often equivocal with respect to the origin of the insertions. For L1s to be evolutionarily successful they must retrotranspose during early human development or in the germline, in order to transmit copies to the next generation. The purpose of this thesis was to develop sensitive and yet robust methods to screen human embryos and embryonic cell models for *de novo* full-length endogenous L1 insertions. We developed a new high throughput sequencing technique, which was able to recover single molecule retrotransposition events. Based on this technique we identified 172 candidate novel L1 insertions in a total of three human embryos, represented by whole-genome amplified DNA of individually dissected blastomeres and the remaining blastocyst tissue. 57 of these insertions are potentially genuine *de novo* endogenous L1 insertions. Moreover, we have identified a candidate germline specific L1 insertion from a healthy adult donor. Therefore, this study has detected candidate *de novo* L1 retrotransposition events in human embryos and germlines, using an approach that enables complete validation and characterization of the insertions, despite operating at the single molecule and single cell level. We consider this technical innovation will be most significant in the ongoing dissection of how L1, the dominant human transposon, is actively driving the evolution of modern human genomes.

Acknowledgements

My greatest appreciation goes to Dr. Richard Badge for welcoming me into his lab four years ago. I can't put into words how much I have learnt from him, in particular his great passion for science always kept me motivated. I would like to thank him for being a great and supportive supervisor, and for being a friend who always returned my silly mistakes with a kind smile.

Also, I would like to thank all the members of the Badge group over the last four years, especially many thanks to Dr. Catriona Macfarlane for her help and support in the lab and making the lab an enjoyable place to work. Also many thanks to Pam Collier, whose lab solutions I was still using right up until the end of my PhD, despite her leaving nearly three years ago! I would like to thank Dr. Daniel Zadik for teaching me Perl scripting and thanks for his bioinformatics advice for analysing my high throughput data.

I would like to thank all the great people (past and present) in the lab G18/G19. My appreciation also extends to Rita Neumann, who always helped me with the technical problems during the lab work and for her homemade PCR buffers, which were magic! Many thanks to my fellow PhD students: Jon, who always kindly listened to my moans and groans. Thanks for your big and warm hugs, I wish you and your beautiful family all the best in future. Also Tina who made the early hours in the lab so fun; I enjoyed having the whole lab to ourselves to work at our regular 7:00am morning slot. Thanks to Linda for our coffee breaks and wish you all the best in future. Thanks to Dr. Ingrid Berg who always made me laugh and cheered me up with her kind and generous gifts.

My appreciation goes to Dr. Robert Hardwick. Many thanks for being the kindest and smartest person that you are. Thanks for cheering me up in all the times when I was down and tired of writing. A big thank to you for proof reading this thesis at the busiest time of your career. I wish you all the best in your new job.

Last but definitely not least, my greatest appreciation goes to my beautiful mum Mahnaz, my dearest dad Mehrdad and my little brother Rahi for being the wonderful people that you are. Many thanks for your help and support, both emotionally and financially, all through my life. Without your help and support my dreams would not come true at all stages of my life. Whatever I am now is all because of you and I hope that one day I can make you proud. Also thanks to my dearest family in Iran and my friends that I had to sacrifice seeing to pursue my PhD. Indeed, this thesis is dedicated to my beautiful grandma (Tajmah) and my dearest grandparents who passed away while I was here doing my PhD.

Table of Contents

Biological Abbreviations

Abbreviation	Full term
aa	amino acid
AP	apurinic/apyrimidinic
ASP	Anti Sense Promoter
ATLAS	Amplification Typing of L1 active Subfamilies
BAC	Bacterial Artificial Chromosome
bp	base pair
BLAST	Basic Local Alignment Search Tool
cpm	counts per minute
C-domain	Cysteine rich C terminal domain
CGD	Chronic Granulomatous Disease
cDNA	complementary DNA
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease
dNTP	Deoxynucleoside triphosphate
dbRIP	Database of Retrotransposon Insertion Polymorphisms
EDTA	Ethyleneodiamine tetraacetate
EN	Endonuclease
FCS	Fetal calf serum
G418	Geniticin sulfate
gag	Group-specific antigen
GFP	Green Fluorescent Protein
Hg19	University of California Santa Cruz human genome build 19
HGWD	Human Genome Working Draft
hESC	Human Embryonic Stem cells
hNPC	Human Neuronal Progenitor Cells
kb	Kilobasepairs
kDa	Kilodalton
L1	LINE-1 element
LINE	Long interspersed nuclear element (autonomous non-LTR retrotransposon)

LTR	Long Terminal Repeat
MBD	Methyl-CpG-Binding Domain
MBq	Megabecquerel
mRNA	messenger RNA
Myrs	Million years
OD	Optical Density
ORF	Open Reading Frame
PBS	Phosphate Buffered Saline
PCR	Polymerase chain reaction
Perl	Practical extraction and report language
rpm	revolutions per minute
RT	Reverse Transcriptase
RC	Retrotransposition Competent
RNP	Ribonucleoprotein
RNA	Ribonucleic Acid
SV40	Simian Virus 40
SINE	Short Interspersed Nuclear Elements
SRY	Sex-determining Region Y
SA	Splice Acceptor
SD	Splice Donor
SNP	Single Nucleotide Polymorphism
SVA	SINE-R, VNTR and Alu
SDS	Sodium Dodecyl Sulphate
TE	Transposable Element
Tris	Tris (hydroxymethyl) aminomethane
tRNA	Transfer RNA
TSD	Target Site Duplications
TPRT	Target Primed Reverse Transcription
TE	Transposable Element
UCSC	University of California at Santa Cruz
U	Units
UTR	Un-Translated Region
VNTR	Variable Number Tandem Repeat
YY1	Ying Yang 1

Chapter 1

Introduction

1.1 Historical background of mobile elements

For a long time it was assumed that a genome was an assembly of genes interrupted by their regulatory elements. However, it was soon recognised that the morphological complexity of an organism does not necessarily directly correlate with its genome size (Thomas, 1971; Gregory and Hebert, 1999).

Progressive development in sequencing technologies and their large-scale application culminated in the elucidation of the first draft of the human genome (Lander *et al*., 2001). Although suspected for some time, the draft sequence established that the human genome contains a very substantial amount of non-coding and repetitive sequence. The non-coding portion forms more than 95% of the human genome, ~50% of which is repetitive sequence (Thurston *et al*., 2007; Zingler *et al*., 2001). These repetitive sequences were often described as "junk DNA" as there was not any evidence of beneficial function for the host (Ohno, 1972; Pagel and Johnstone, 1992). A small percentage of repetitive sequences are comprised of simple tandem repeats-

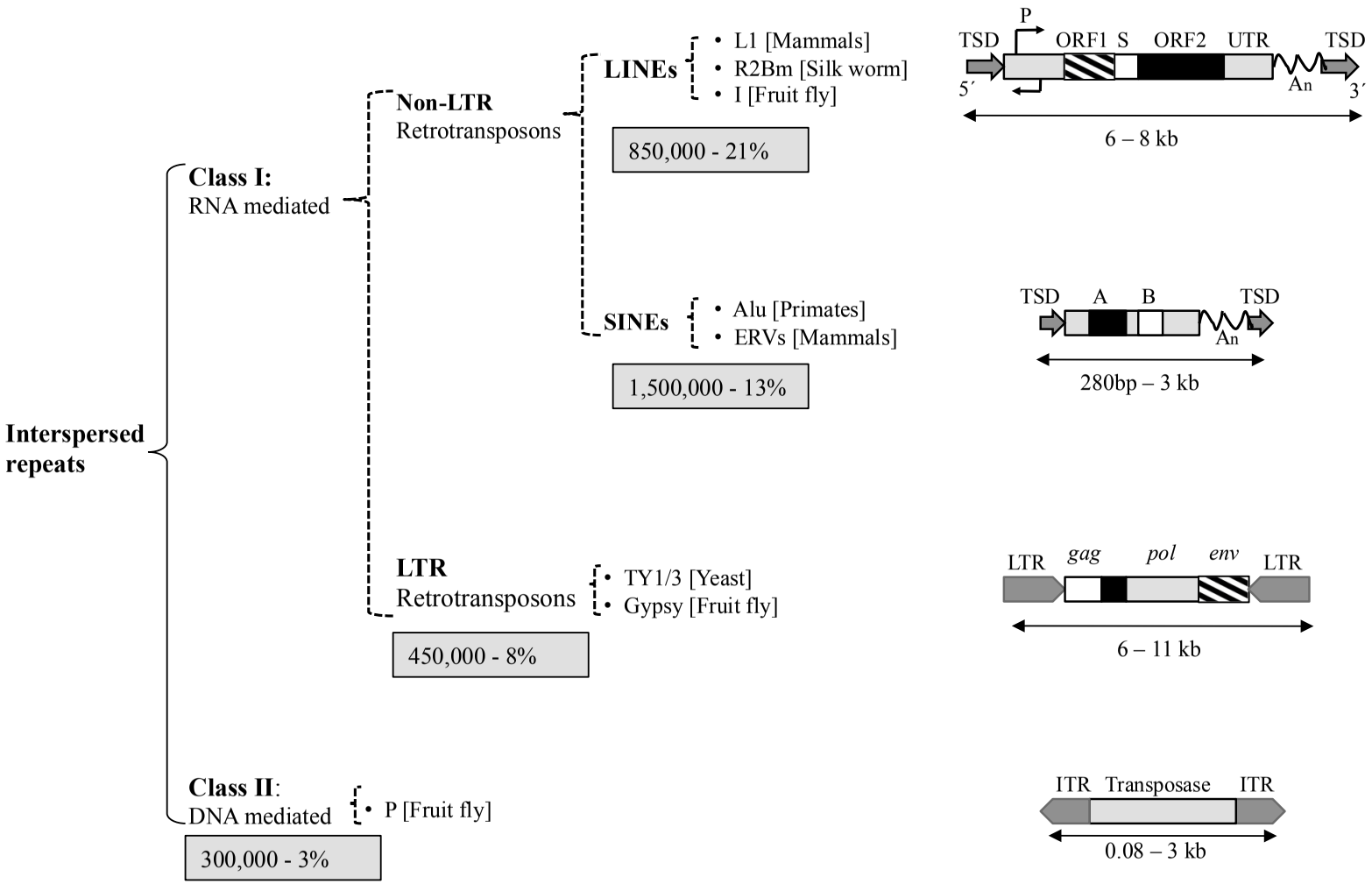

Figure 1.1 Classes of interspersed repeats with their examples (termini of the branches, with their *taxa* of origin in square brackets). For each class their copy number and contribution (%) to the genome summarised in the grey boxes. TSD, Target Site Duplication; P, Promoter; ORF1/2, Open Reading Frame; S, Spacer An, Poly Adenosine nucleotides; A and B, RNA Polymerase III conserved regions, LTR, Long Terminal Repeat; *gag*, group specific antigen; *pol,* polymerase; *env,* envelope; ITR, Inverted Terminal Repeats,. Adapted from Lander *et al.,* 2001.

-like microsatellites or telomeric repeats, but the vast majority derives from transposable elements (TEs).

TEs were first identified in the late 1940s by Barbara McClintock, in maize, *Zea mays* (McClintock, 1950). Today, many different kinds of mobile DNA have been identified in virtually all species ranging from bacteria and yeast to plants and mammals, as illustrated in Figure1.1.

The question of why TEs have been so successful throughout evolution is the subject of ongoing discussion. Transposable elements have been called "selfish genes" (Dawkins, 1976) and "genomic parasites" (Yoder *et al*., 1997) in relation to their host genome, but evidence has accumulated over the last several decades demonstrating that, despite their disease-causing potential (reviewed in Kazazian, 1998), TEs might have some overall beneficial effect. For example, TEs can increase genomic diversity and consequently drive genome evolution within a species (Boeke and Pickeral, 1999; Nekrutenko and Li, 2001; Seleme *et al*., 2006); they can play a role in the stress response of the host cell (Li and Schmid, 2001); and they can take over vital cellular functions, such as telomere function (Pardue *et al.*, 1996).

TEs can also have practical uses. For example, human specific mobile element insertions (mostly L1 and Alu) have the potential to be used for inferring human geographical origin, sex identification, DNA identification and quantification (Xing *et al*., 2007). However, while the contribution of mobile elements to host genomic architecture and fluidity is undeniable, little is currently known about the evolutionary dynamics of their mobilisation in humans.

1.2 Human Transposable elements

In *Homo sapiens*, transposable elements are responsible for the formation of at least 45% of the genome (Lander *et al*., 2001). Figure 1.1 illustrates the different types of mobile elements that have been involved in mammalian and human genome expansion.

TEs can be classified into two groups based upon their genomic integration method (Pace and Fechotte, 2008). Class I elements transpose via an RNA intermediate,

utilising reverse transcriptase and include long and short interspersed nuclear elements, and long terminal repeat elements. The Class I transposition mechanism can be thought of as a 'copy and paste' method and as such is inherently replicative. Class II mobile DNA integrates into the human genome, using a DNA intermediate, through a 'cut and paste' mechanism (Pace and Fechotte, 2008; Kazazian *et al*., 2002).

1.2.1 DNA transposons; class II transposable elements

The mechanism of DNA transposition is a 'cut and paste' mechanism that is not inherently replicative. DNA transposons mobilise via a DNA intermediate, which is mediated by a transposase. Only about 3% of the human genome is derived from DNA transposons (Fig. 1.1) (Lander *et al*., 2001).

The structure of DNA transposons (Class II TEs) generally consists of the coding sequence of a transposase enzyme flanked by inverted terminal repeats. The transposase enzyme usually binds near the inverted repeat termini to hydrolyse DNA phosphodiester bonds and excise the transposon, exposing 3′ hydroxyl groups (OH). The exposed 3′ OH group allows insertion (after target site cleavage) into a new site prior to gap filling by host DNA repair proteins. This process leads to the formation of direct terminal repeats at the target site, which are known as target site duplications (TSDs) (Moran and Gilbert, 2002). The DNA transposition machinery acts in *trans* and mobilises any elements with transposase recognition signals and so is not specific for the encoding elements. Competitive parasitism of the active DNA transposon machinery by inactive elements is likely to have led to the extinction of active DNA transposons in the human genome (Lander *el al*., 2001).

The evolutionary history and genomic impact of transposons have been well studied in mammals. It is known that all ~300,000 DNA transposons identified in the human genome are genomic fossils that have been inactive for at least 50 Myr (Lander *et al*., 2001; Pace and Feschotte, 2007; Smit and Riggs, 1996), and therefore any effects of transposition in the human genome must come from a different class of transposable element. The active transposable elements in humans are retrotransposons and mainly its youngest subfamily, the L1Hs-retrotransposons. Comparative genomic analysis between human genome reference and the draft chimpanzee genome showed that

there are 1,174 L1 specific to humans (Mills *et al*., 2006). It is this group of retrotransposons that are the subject of this thesis.

1.2.2 DNA retrotransposons; class I mobile elements

By far the largest portion of mobile DNA in humans originates from retrotransposons. In contrast to DNA transposition, DNA retrotransposition is inherently replicative and functions via a 'copy-and-paste' mechanism, involving transcription of the complete element, reverse transcription of the RNA into a cDNA, and integration of the cDNA into a new locus in the genome. Thus, one functional retrotransposon can generate multiple copies at new genomic locations. This circumstance, and the fact that there is at least one family of retrotransposons still active in humans (the L1Hs family), may account for the excess of retroelements in the human genome. Retrotransposons can be divided into two major classes that are phylogenetically and structurally unrelated (Craig, 2002). The long terminal repeat (LTR) retrotransposons account for 8% of the human genome, and are characterised by direct LTRs flanking the element's coding regions (Figure 1.1). LTR and non-LTR retrotransposons do share some important functional characteristics. They each have a robust and functional 5′ promoter (Hata and Sakaki 1997), which is responsible for transcription of full-length RNA, and they each encode a reverse transcriptase enzyme in order to produce a cDNA copy of this RNA. However, there are also important differences: in the autonomous elements (non-LTR retrotransposons), the cDNA integrates into new genomic loci using its own unique protein machinery (Curcio and Derbyshire 2003) and the integration process is initiated by an element-encoded endonuclease (EN).

1.2.2.a Long Terminal Repeat (LTR) Retrotransposons

LTR retrotransposons are also called 'retrovirus-like elements' or 'endogenous retroviruses' because their replication pathway is similar to that of retroviruses. They are thought to originate from retroviruses that have lost a functional *env*-gene, confining them to strictly intracellular replication (Esnault *et al*., 2008). Thus, endogenous retroviruses cannot infect other cells, and are forced to go through their

replicative cycle within a single cellular lineage. With the possible exception of HERV-K, which is a putatively active human endogenous retrovirus, all known human LTR-retrotransposons are genomic fossils that have not been active for the last 40 Myr (Costas and Naveira, 2000; Lander *et al*., 2001).

1.2.2.b Non-Long Terminal Repeat (non-LTR) Retrotransposons

Non-LTR retrotransposons are evolutionarily more ancient than LTR retrotransposons (Furano, 2000). Protein sequence comparisons indicate that they share a common origin with RT-bearing group II introns of bacteria and mitochondria (Yang *et al*., 1999). Comprising more than one third of human DNA (32%), non-LTR retrotransposons clearly have had a great impact on the human genome.

Based on the structure of their coding regions, the autonomous non-LTR elements are further subdivided into the restriction enzyme (RE) type and the apurinic/apyrimidinic endonuclease (APE) type. The RE-type non-LTR retrotransposons are characterised by a single open reading frame (ORF) with a RE-like EN domain following the C-terminal end of the RT domain. RE-type elements represent the oldest lineage of non-LTR retrotransposons (Malik, 1999).

Most retrotransposons discovered so far are APE-type non-LTR retrotransposons. They are recognised by having two ORFs and the existence of an EN domain that is distantly related in sequence to the apurinic/apyrimidinic (AP) endonucleases (Martín *et al*., 1995; Feng *et al*., 1996). The EN domain is localised at the N-terminal end of ORF2p, upstream of the RT domain. Based on the elements' structures, and on phylogenetic analyses of their RT domains, we can currently distinguish four groups of APE-type non-LTR retrotransposons, and these can further be subdivided into a further 11 clades (Burke *et al*., 1999; Malik and Eickbush 1998; Lovsin *et al*., 2001).

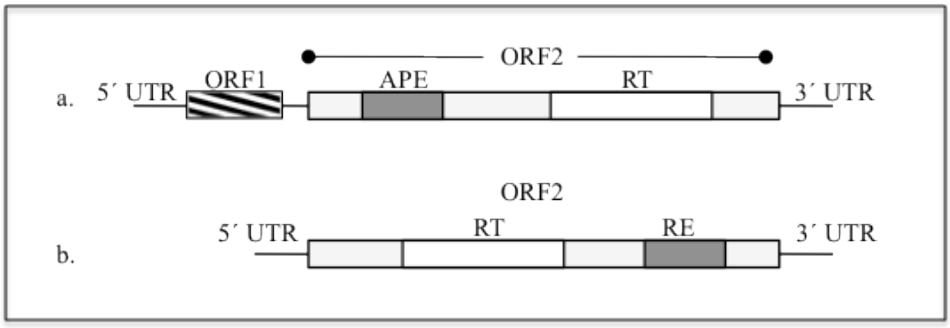

Figure 1.2 schematic diagrams of RE-type and APE-type non-LTR retrotransposons. They differ in their structural organisation and in their coding capacity, **a.** APE-type non-LTR retrotransposons, **b.** RE-type non LTR retrotransposons, UTR: untranslated region, ORF: open reading frame, APE: apurinic/apyrimidinic, RT: reverse transcriptase, RE: restriction enzyme-like endonuclease (Craig, 2002).

1.3. Autonomous and non-autonomous non-LTR retrotransposons

The non-LTR retrotransposons can also be categorised as either autonomous or non-autonomous retrotransposons. Autonomous retrotransposons are able to encode the required proteins for their own retrotransposition. However, non-autonomous elements are unable to retrotranspose without hijacking the retrotransposition machinery of autonomous elements (Lander *et al*., 2001).

1.3.1. Human Short Interspersed Nuclear Elements (SINES)

SINEs are non-autonomous, non-LTR retrotransposons. These elements exist in the human genome at a very high copy number of around 1.5 million, which comprises about 13% of the genome (Lander *et al*., 2001). SINEs vary in length between 100-300 bp (fig. 1.1) and contain internal promoters for RNA polymerase III (Singer, 1982; Okada and Ohshima, 1995; Schmid, 1996), which shows that they originated from functional non-coding RNAs (tRNA, 7SL and 5S rRNA). In primates, SINEs consist of two or more modules: the 5′ UTR and a poly A tract at their 3′ UTR. Two major families of primate SINE elements are Alu and SVA (SINE-VNTR-Alu).

Alu elements are 300 base-pair DNA sequences that derive from the 7SL RNA gene, which is the RNA scaffold of the signal recognition particle (SRP) that binds to nascent signal peptide sequences and transiently arrests translation (Ullu and Tschudi, 1984; Siegel and Walter, 1988). Alu elements consist of two monomers: monomer A includes a Pol III promoter, which directs transcription from the first nucleotide of the element, and monomer B, which is separated from monomer A by an A-rich linker sequence (Ullu and Weiner, 1985). Alus have no protein-coding capacity and they can only ensure that their RNA is transcribed. For reverse transcription and integration they use L1 element proteins (Dewannieux *et al*., 2003). Indeed, the secondary structure of Alu RNA resembles ribosomal associated RNA. By mimicking functional ncRNAs Alu RNA may be able to associate with ribosomes in close physical proximity to nascent LINE proteins, and misappropriate them for their own replication (Boeke, 1997; Weichenrieder *et al*., 2000; Dewannieux *et al*., 2003). Alu elements are commonly found in the untranslated regions of genes, introns and intergenic regions of the genome (Batzer and Deinnger, 2002). Alu elements also contain a poly-A tail, which is necessary for its retrotransposition (Roy-Engel *et al*., 2002) Alu activity is estimated at 1 new insertion in 200 births (Batzer and Deinnger, 2002). Young and polymorphic Alu insertions are mainly derived from three Alu subfamilies, which are actively retrotransposing in contemporary human genomes. These are Ya, Yb and Yc (Batzer and Deinnger, 2002).

SVAs are the youngest active human retrotransposons, are hominid specific, and produce non-coding RNA (Wang *et al*., 2005). To date, several disease-causing insertions associated with SVA elements have been reported. Structurally, SVA elements consists of a 5-6 nt repeat (CCCTCT) at their 5′ region followed by an Alu-like domain, a GC-rich variable number tandem repeat (VNTR), and an *env* –like gene at their 3′ UTR which derives from HERV-K (Wang *et al*., 2005). Recently SVA elements have been demonstrated to utilise the L1 machinery for their replication, at least in cell culture based assays (Hancks *et al*., 2011).

1.3. 2 Human Long Interspersed Nuclear Elements (LINES)

Long interspersed nuclear elements (LINEs) are the only autonomous non-LTR retrotransposons, *i.e.* they encode the proteins required for their own

retrotransposition. LINE retrotransposons are further classified into three sub-groups in the human genome: LINE1 (L1), LINE2 (L2) and LINE3 (L3). LINE1 is the only active member of this family and it has a copy number of around 500,000, which represents about 17% of the genome. LINE2 and LINE3 are older lineages that together comprise less than 4% of the genome. They have accumulated numerous mutations during the course of evolution, and so they unlikely to be still retrotranspositionally active (Lander *et al.,* 2001).

99% of LINE1s are inactive due to 5´ truncation, internal rearrangements or deletions, but it has been estimated that in an average diploid human genome there are 80-100 full-length L1 with intact ORFs, which are likely to be retrotranspositionally competent L1s (RC-L1s) (Deininger *et al*., 2003 and Brouha *et al*., 2003).

LINE element proteins display a *cis* preference, *i.e.* they preferentially retrotranspose their encoding RNA, largely ensuring that only functional copies are propagated (Wei *et al*., 2001). This *cis* preference, from an evolutionary point of view, minimises the impact of the accumulation of mutated elements on active L1 retrotransposition. However, it is known that the LINE1 autonomous machinery can also act in *trans* to retrotranspose non-autonomous retrotransposons such as short Interspersed Elements (SINEs) and SVA (SINE/VNTR/Alu) elements (Callinan *et al*., 2006) and cellular transcripts (Esnault *et al*., 2000; Boeke, 1997). In rare cases the *cis* preference of LINEs is also circumvented by spliced mRNAs of cellular genes. This results in an intronless and promoterless retropseudogene copy of the original gene transcript, followed by a polyA tail flanked by target site duplications (Vanin, 1985). Therefore, these processed retropseudogenes are also a direct result of LINE activity (Esnault *et al*., 2000). Since LINE1s are the only active family of LINEs in the human genome and they are the subjects of this project, they are discussed in more detail below.

1.4 L1 retrotransposon structure and retrotransposition

To date, the human L1 is the most thoroughly characterised mammalian APE-type non-LTR retrotransposon (Ostertag and Kazazian, 2001a; Moran and Gilbert, 2002).

Human specific L1s are further divided into pre-Ta (Transcribed, subset a), Ta0, Ta1, Ta1nd, and Ta1d subfamilies.

The preTa subfamily of L1 is characterised by an ACG diagnostic sequence at its 3′ UTR. Based on the nucleotide at the position 6040 (compared to the reference element L1.3, Accession L19088) this subfamily can be further grouped into two types: ACG-G (younger lineage: ~1.92 Myrs old) or ACG-A types (~3.24 Myrs old) (Salem *et al*., 2003). It has been suggested that the preferential genomic sequence recognition site for integration of preTa L1s is TTTT/A and TCTT/A (Jurka, 1997). Moreover, it has been suggested by Salem *et al*., (2003) that the pre-Ta families preferentially integrate into low GC content (36%) genomic DNA. The majority of pre-Ta family elements are 5′ truncated, but 29 full-length preTa elements with intact ORFs have been reported. As a result they are considered retrotranspositionally active elements as well as having given rise to one case of human genetic disease (by integration into the factor VIII gene, resulting in hemophilia A) (Kazazian *et al*., 1988; Salem *et al.,* 2003). Sequence variation is also observed at the 3′ UTR, some of which is caused by 3′sequence transduction (Salem *et al*., 2003).

The Ta family (or transcribed subset A) is the youngest and most active L1 family, and has been associated with 16 of the 17 disease-causing insertions in humans (Kazazian, 2004). Over 50% of these elements show dimorphism (presence or absence) across human populations (Boissinot and Furano, 2001). These families of L1 emerged after the divergence of humans from chimpanzees about 6 Myrs ago and so are specific to humans. There are two main Ta subfamilies: L1Ta0 and L1Ta1 (Boissinot *et al*., 2000). The ACA nucleotides at position 5954 - 5956 of the 3′ UTR is diagnostic for this family. Based on the nucleotides at positions 5557 and 5560 in their ORF2 they are further divided into Ta1 and Ta0. The Ta1 elements have T and G nucleotides at these two positions and Ta0 have G and C at these positions (Boissinot *et al*., 2000). Ta0 has more sequence similarity to the non-Ta L1s, and therefore has been suggested to be an older family of elements. Their sequence is more diverged and so they are very unlikely to be highly active in the genome, whereas the Ta1 family is younger than pre-Ta and Ta0 families, and therefore have accumulated fewer inactivating mutations. Hence they still actively retrotranspose and are expanding their numbers in the human genome (Boissinot *et al*., 2000). It is estimated that the Ta1 family arose about 1.6 Myrs ago and can be further divided

into two subfamilies: the Ta1d and Ta1nd. The Ta1d (deleted) group are recognised by a deletion at position 74 in the 5′ UTR whilst the Ta1nd (non-deleted) group lacks this deletion. There are around 90 full-length human L1s with intact ORFs in the human genome reference sequence, which are potentially RC-L1s (Brouha *et al*., 2003). However, cell culture based retrotransposition assays demonstrated that only 6 of these elements account for 84% of the total retrotransposition activity (Brouha *et al*., 2003). This data suggests that these very active elements dominate retrotransposition activity in the human genome. Four of the "hot" L1 elements characterised by Brouha *et al*. (2003) belong to the Ta1d family, one belongs to the Ta1nd and one belongs to the Ta0 family (Brouha *et al*., 2003). Recent sequence-based studies have estimated the rate of L1 insertion into the human genome is around 1 in 212 live births (Xing *et al*., 2009) and 1 in 140 (Ewing and Kazazian, 2010). These estimates are much lower than was previously estimated (1 in 33 live births) for L1 insertions, based on the activity of disease-causing elements (Brouha *et al*., 2003).

1.4.1 L1 structure

A complete retrotransposition-competent (RC) L1 element is 6-8 kb in length and contains two non-overlapping open reading frames: ORF1 and ORF2 (Fig. 1.4). The nucleotide sequence of a representative functional member of the L1-family, L1.3 (Accession No L19088), is given in Appendix 1 and is the basis for sequence coordinates used throughout this thesis.

The 5′ untranslated region (UTR) of a RC-L1 is approximately 900 bp in length. A major polymorphism of L1 elements occurs within this region: the presence or absence of a 131-bp sequence (Minakami *et al*., 1992). The L1 sense promoter is also located within the 5′ UTR region and the first 155 bp have been demonstrated to be involved in L1 expression (Minakami *et al.,* 1992). The structure of each L1 component and their role in L1 retrotransposition (where known) are discussed in more detail in the following sections, and a schematic diagram of an intact L1 retrotransposon and its modules is presented in Figure 1.4.

1.4.1.2 The L1 promoter and transcription of the L1 element

The 5′ UTR of the L1, which is about 900 bp in length, accommodates two internal promoters (+1 to +670). The region between +1 to +100 shows the highest promoter activity, although no TATA box is present (Swergold, 1990). The L1 5′ UTR contains a sense promoter (SP), which initiates transcription at +1 of the L1 sequence and an antisense promoter (ASP), positioned between +399 to +467 bp of the L1 sequence. Both sense and antisense L1 promoter sites are highly conserved in human L1PA10 - L1PA1 families over 40 million years of evolution. It is suggested that over 1/3 of L1Hs elements contain highly active ASPs, which are capable of interfering with normal gene expression (Niguman *et al*., 2002; Speek, 2001) when located intragenically.

The L1 sense promoter possesses characteristics of both RNA polymerase II (PolII) promoters, which control transcription of all protein-coding genes, and RNA polymerase III (PolIII) promoters that are responsible for synthesis of tRNA, 5S RNA and several small and non-coding RNAs (Kurose *et al*., 1995). The L1 transcript is about 6 kb long and it has two protein-coding regions and a polyadenylated extension at the 3′ end of the transcript. These characteristics suggest this is a PolII dependent promoter. However inhibition studies on the L1 promoter have shown it is less sensitive to α-amanitin, a Pol II inhibitor, and is more sensitive to tagetitoxin, a specific PolIII inhibitor (Kurose *et al*., 1995). Moreover, the L1 transcript terminates at the T-tract (T=20 nt) on the non-template strand of L1, which is characteristic of

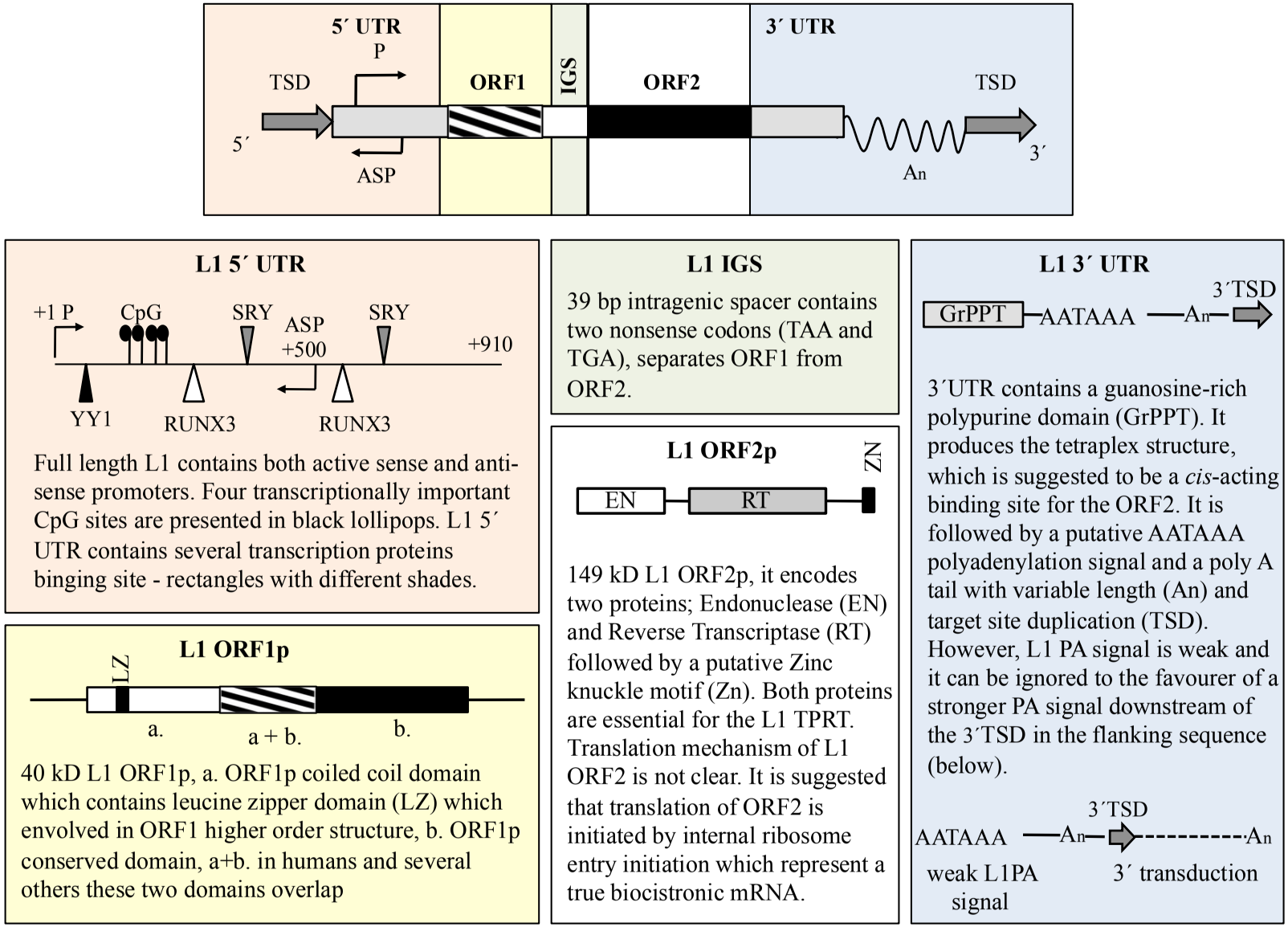

Figure 1.3 L1 structure and the summary details of its component including: L1 5′ and 3′ untranslated region (UTR), open reading frame 1 and 2 (ORF1/2) Intragenic spacer (IGS), poly A tail (An) and target site duplication (TSD). **L1 5′ UTR** structure is adapted from Badge, poster publication 2007, **L1ORF1P** adapted from Martin *et al*., 2000, **L1ORF2p** structure: Goodier *et al*., 2004, and **L1 3′ UTR** adapted from Craige *et al*., 2002.

PolIII transcripts. These data suggest that the L1 promoter is PolIII dependent, but produces transcripts more characteristic of Pol II.

This unusual sensitivity may be explained by the importance of YY1 transcription factor in L1 transcription initiation (Athanikar *et al.,* 2004), which is utilised at both PolII and PolIII promoters.

The L1 sense promoter creates a long, protein-encoding, polyadenylated transcript and the promoter is internal, such that it initiates transcription at position +1 of the L1 sequence but lacks features characteristic of PolII promoters such as upstream TATA and CAAT boxes (Swergold, 1990). Moreover, the frequency of an extended G nucleotide tract between the 5′ TSD and the L1 start site may be due to the L1 RT reverse transcribing the RNA 5′ 7-methyl G-cap structure that is added upon RNA PolII mediated transcription (Lavie *et al*., 2004). This poly-G tract upstream of the L1 5′ UTR is absent in Alu elements, which are transcribed by PolIII (Lavie *et al*., 2004). These observations suggest that it is more likely that the L1 promoter is PolII driven. The L1 5′ UTR also contains several PolII transcription factor binding-sites, which have been shown to be involved in the transcriptional regulation of L1s.

1.4.1.2.a YY1 binding site

The ubiquitous transcription factor YY1 (Ying Yan 1), which is a PolII and PolIII transcription factor, has been introduced as an important sequence in L1 transcription, and is located at +13 to +26 of the L1 5′ UTR sequence (Becker *et al*., 1993; Kurose *et al*., 1995). Since YY1 is capable of both activating and repressing transcription, this protein may play a role in down-regulating L1 transcription in some cell types, while activating it in others (Becker *et al*., 1993). YY1 regulates L1 transcription by enhancing accurate transcription initiation rather than initiating it, as even L1s, which lack the YY1 site have functional promoters (Anthanikar *et al*., 2004). It has been demonstrated that inhibition of the YY1 binding site in tissue culture assays has a minor effect on L1 transcription activation and retrotransposition (Anthanikar *et al*., 2004). However, it has been demonstrated that the deletion of the YY1 site in the first 20 bp significantly reduces (5 fold) L1 retrotransposition in cell culture assays (Singer *et al*., 1993).

Since deletion of the YY1 binding site does not inhibit L1 transcription, L1 must transcribe from upstream or downstream of this site. Transcription initiation from downstream of the YY1 binding site leads to 5′ truncated progeny, which may not be retrotranspositionally competent due to the 5′truncation. It has been shown that most RC-L1s are transcribed from +1 or nearby, such that their progeny are also full length and able to autonomously retrotranspose (Anthanikar *et al*., 2004).

1.4.1.2.b Other L1 transcription factor binding sites

Previous studies have demonstrated that the L1 5′UTR contains four methyl-CP2-responsive elements at the following positions: +36, +101, +304 and +481 (Hata and Sakaki, 1997). The C-methyl binding proteins bind to methylated DNA (Feng and Zhang, 2001). Based on the recognition-binding site these proteins are divided into two types: the MBP (Methyl Binding Proteins) group binds to the methylated DNA, while the second group, MeCPs (Methyl-CpG binding proteins) and MDBP (Methylated DNA Binding Proteins), has no sequence specificity to methylated DNA (Feng and Zhang, 2001). Among these the MeCP2 proteins are the most abundant methyl-Cytosine binding proteins and it has been demonstrated MeCP2 binds to methylated-DNA only in the context of chromatin, contributing to the long-term repression and nuclease-resistance of methyl-CpGs (Meehan *et al*., 1992, Hata and Sakaki, 1997).

Moreover, Tchenio *et al*., (2000) demonstrated that the human L1 promoter contains two functional sites for SRY (sex determining region Y) transcription factors. SRY transcription factors are members of the SOX protein family, and are expressed in the urogenital ridge of the embryo and in adult, testis, hypothalamus and midbrain (Tchenio *et al*., 2000). *In vitro* studies have shown that ectopic over-expression of one of the SRY families, Sox11, results in a 10 folds *trans*-activation of endogenous L1Hs (Tchenio *et al*., 2000). The two potential binding sites for SOX transcription factors are located in the first 670 nucleotides of the L1 promoter. The first site, SRY$_A$, is located between nucleotides 427-477, and SRY$_B$ is located between 572-577. An *in vivo* study has demonstrated that SRY transcription factor binding at the L1 promoter

can drive transcription in cell culture, and mutations at the SOX binding site can inhibit L1 transcription (Tchenio *et al*., 2000).

The RUNX3 family contains heterodimeric transcription factors, which can potentially bind to three regions in the L1 promoter: nucleotides +83 to +101 and +526–508 of the L1 5′ UTR, and potentially influence L1 transcription by regulating both sense and antisense promoters (Yang *et al.,* 2003). Mutation analysis on each of the three sites has demonstrated that mutation at the first binding site reduces L1 transcription. Mutations at the other two binding sites do not have any significant effect on L1 transcription, and this is perhaps due to the second and third binding sites being located outside the +100 of L1 5′ UTR, which is important for transcription initiation (Yang *et al.,* 2003).

1.4.1.2.c CpG modifications of L1 promoter and their effect on L1 activity

The 5′ UTR of L1 contains a CpG island that is usually heavily methylated in somatic cells (Woodcock *et al*., 1997). Thayer *et al*. (1993) studied eight cell lines, and observed an inverse correlation between ORF1 protein (ORF1p) expression and the methylation status of the 5′ end of L1 elements. This indicates that methylation of this region could play a role in L1 regulation. Also, a study by Hata and Sakaki (1997) on the L1 promoter showed that L1 retrotransposons are exquisitely sensitive to their methylation status: they demonstrated that methylation of four conserved CpG dinucleotides in the L1 promoter strongly represses its activity, implying that demethylation is required for L1 mobilisation (Hata and Sakaki, 1997). Interestingly, it has also been demonstrated that hESC lines frequently show methylation variation at CpG islands containing L1 sequences (Allegrucci *et al*., 2007). More recently studies have demonstrated a connection between L1 expression and another DNA CpG modification, 5-cytosine hydroxy methylation (5chm). It has been shown that hydroxymethylation of the L1 promoter can activate L1 transcription and expression in a mouse model (Ficz *et al*., 2011). The significance for this observation for human L1s remains to be established as mouse and human L1 promoters are not related in sequence, although both contain CpG islands. L1 DNA methylation is explored in more detail in the Introduction section of chapter 5 (5.1).

1.4.1.3 L1 ORF1 and ORF2 and translation of the L1 retrotransposition machinery

Despite host defence mechanisms acting against L1 retrotransposition, these potentially mutagenic insertions occur in germline and somatic tissues, as evidenced by disease causing insertions. Because the L1 translational machinery has a strong *cis*-preference functional protein crosstalk between individual elements is greatly reduced, and lack of competition from partially functional mutants may explain the longevity of L1 activity. However this phenomenon requires both ORFs to be expressed from the same transcript, so co-expression of the ORF encoded proteins is likely a marker of active L1 retrotransposition. Co-expression of the two L1-encoded proteins, ORF1p and ORF2p, has been detected by immunohistological analyses in pre-spermatogonia of human foetal testis and in germ cells of human adult testis (Ergün *et al*., 2004). Also, most of the disease-causing L1 insertions are germline in origin (Kazazian, 2004). These data and parallel observations of ORF1p expression in mouse pachytene spermatocytes (Martin and Bushman, 2001) fit with the expectation that potentially mutagenic transposable element confine their replication to germlines where they can maximise their probability of transmission, without compromising host viability. In the following section the structure and function of each ORF is explained in more detail.

1.4.1.3.a Translation and role of L1-ORF1 in L1 retrotransposition

The first open reading frame of L1 (L1 ORF1) is 1017 bp in length and encodes a 338 amino acid cytoplasmic protein also known as p40 (Hohjoh and Singer, 1997). The centrally located leucine zipper domain in human L1ORF1 is involved in formation of higher order ORF1p multimers (Craig, 2002). The carboxyl domain of ORF1 is basic and has several conserved amino acids, which are likely to play a role in RNA binding. However, this carboxyl domain lacks the common functional motifs, which are required for RNA binding proteins such as the RNP motif and the Arg-rich motif (Craig, 2002). The sequence of ORF1p is not closely related to any protein with

known function and its role in the L1 retrotransposition cycle is incompletely understood (Basame *et al*., 2006). It is suggested that L1ORF1p is translated by ribosomal initiation at the 5′UTR followed by ribosomal scanning at the 661 nt position (McMillan and Singer, 1993). Results of co-immunoprecipitation experiments demonstrate that ORF1p is a high affinity RNA binding protein with no sequence binding specificity (Kolosha and Martin, 2003). It has also been demonstrated that the nucleic acid chaperone activity of ORF1p is important for successful L1 retrotransposition (Martin *et al*., 2005). Also, *in vitro* and *in vivo* experiments have each demonstrated that L1ORF1p exists in many copies in the cytoplasm (Hohjoh and Singer, 1996).

Several roles have been proposed for ORF1p in the L1 retrotransposition process. It has been suggested that the L1 RNA is very unstable and therefore ORF1p with its RNA binding activity is required to coat and protect the L1 RNA intermediate in the cytoplasm before the target-primed reverse transcription (TPRT) process. It is thought that *cis* preference acts to ensure that the L1 proteins associate with their functional encoding RNA (Moran and Gilbert, 2002). Although ORF1p has only been detected in the cytoplasm it could still be involved in the later stages of L1 retrotransposition, such as TPRT (Martin and Bushman, 2001). The ORF1p nucleic acid chaperone activity is also likely involved in strand transfer (first and second strand), which allows the annealing of the DNA primer from the target site to the RNA primer during the process of reverse transcription (Martin and Bushman, 2001). It is also possible that ORF1p facilitates the reverse transcription process by enabling movement of polymerase during formation of the first and second cDNA strands through the RNA secondary structure (Martin and Bushman, 2001).

1.4.1.3.b Translation and role of L1-ORF2 in L1 retrotransposition

The second open reading frame (ORF2) of L1 encodes a protein of about 149 kDa containing 1275 amino acids (Scott *et al*., 1987). The initiator codon methionine of ORF2 in the human L1 element is separated from ORF1 by a 66-bp in-frame spacer region containing three stop codons. It is not clear how the separate translation of both ORFs from the bicistronic RNA is accomplished; this problem is made even more

intriguing by the fact that the spacer region is not conserved between L1 elements of different species (McMillan and Singer, 1993). It has been suggested that translation of ORF2 must be accomplished either by reinitiating translation (Kozak, 1987) or by internal initiation via an internal ribosomal entry site (IRES) (McMillan and Singer, 1993).

The ORF2 protein has proven to be very hard to detect, largely due to the lack of robust and specific ORF2p antibodies (Bradley *et al*., 2011). Thus, indirect methods, such as measuring enzymatic activity have been used to study the role of this protein in the L1 retrotransposition cycle. It has been demonstrated that ORF2p has two major activities, each of which can be assigned to specific domains. The N-terminal contains a conserved endonuclease activity domain. Its sequence and crystal structure is similar to the AP-like EN, which is involved in the base excision repair pathway (Ergun *et al*., 2004 and Weichenrieder *et al*., 2004). Despite its conservation, it has been demonstrated that L1s lacking an EN domain are still able to retrotranspose at a lower efficiency by likely using pre-existing DNA nicked sites for their integrations (Morrish *et al*., 2002). The central domain of ORF2p is responsible for the reverse transcriptase activity, and contains a conserved Z-motif (Mathias *et al*., 1991). The L1 RT domain is related to those in other non-LTR elements (Malik *et al*., 1999) and also shows some sequence similarity to LTR retrotransposons and retrovirouses (Xiong and Eickbush, 1990). At the C-terminal end, there is a conserved "C-domain" containing a cysteine-rich region whose function is not clear. It has been suggested that this region has evolved in response to interactions with other L1 sequences or host factors (Bradley *et al*., 2011). Also, it has been shown that mutations in this region abolish the ability of ORF2p to interact with L1 RNA and ultimately block L1 retrotransposition in cultured cells (Feng *et al*., 1996 and Doucet *et al*., 2010).

1.4.1.4 L1 3′ UTR and poly A tail

The 3′ UTR covers the terminal 205 bp of full-length elements, includes a polyadenylation (PA) signal and terminates in a poly (A) tail. One of the characteristics of the L1 PA signal is the ability to transduce genomic DNA (up to 1.6 kb *in vitro)* downstream of its 3′ UTR (Holmes *et al*., 1994). In the process of

polyadenylation the poly-A tail is added to the putative AAUAAA polyadenylation specificity-factor 1 (CPSF1) binding site. However, the L1 PA signal lacks the conserved elements that normally reside downstream of the poly-A site in canonical RNA polymerase II transcripts. Hence it has been suggested that the L1 PA site is weak and can be bypassed by the transcription machinery in favour of a stronger PA site in the 3′ flanking genomic sequence (Moran *et al*., 1999). L1's weak PA signal is suggested to be an evolutionary adaptation that allows L1 to reside within introns with minimum effect on gene expression through the induction of premature polyadenylation (Moran *et al*., 1999). Around a third of L1 elements carry a 3′ transduction and they are estimated to have contributed 33 Mb of DNA to the human genome (Moran *et al*., 1999; Pickeral *et al*., 2000; Goodier *et al*., 2000; Szak *et al*., 2003). The 3′sequence transduction process is discussed in more detail in Chapter 4.

The L1 3′ UTR also contains the sequence motif ($CACAN_5GGGA$) at position 5796 – 5884 nt, which has a high binding affinity for the nuclear export factor 1 (NXF1) (Lindtner *et al*., 2002). It has been suggested that its role is similar to the constitutive transport elements (CTE), which facilitate the nuclear transport of viral intronless mRNA, such as simian type D retroviruses (Lindtner *et al*., 2002).

The 3′ UTR of the L1 element is poorly conserved within and between species (Scott *et al*., 1987). Interruption of this region by additional nucleotides does not seem to have severe effects on retrotransposition. This is also demonstrated in reporter assays, where L1 tolerates marker genes of up to 3500 bp in length in its 3′ untranslated region (Moran *et al*., 1996; Ostertag *et al*., 2000; Gilbert *et al*., 2002

All the classifications above apply to full-length copies of L1. However, only 5 % of endogenous human L1 elements are full length (6 kb). The remaining 95% are 5′ truncated, internally rearranged or deleted (Szak *et al*., 2002). Some of this damage to L1 structure may be the result of mutations and genomic rearrangements after integration of the retrotransposon, but 5′ truncation and inversion most probably occur during the retrotransposition process (Ostertag and Kazazian, 2001a). The low processivity of the L1 reverse transcriptase might be one of the causes of L1 5′ truncation. If the RT and the RNA template dissociate before completion of reverse transcription, the resulting insertion will be truncated at the 5′end (Ostertag and Kazazian, 2001a). In inverted L1 elements the 5′ truncated region is orientated in an

antisense direction to its 3′ end. This structure is thought to be the consequence of a mechanism called 'twin priming' (Ostertag and Kazazian, 2001b). Inversions can be detected in about 25% of insertions in members of the Ta family (Ostertag and Kazazian, 2001a; Skowronski *et al*., 1988).

L1 integrants are usually flanked by variable TSDs with lengths of up to 60 bp (Szak *et al*., 2002). These TSDs are generated during the process of L1 replication. Some TSDs are difficult to identify due to statistical uncertainties about the occurrence of short duplications; the presence of multiple mutations in TSDs of ancient integrants; the presence of blunt end nicking sites (Van Arsdell and Weiner, 1984); or the presence of a staggered double strand break with a 5' overhang instead of a 3' overhang. The latter process causes a deletion of the target site instead of duplication (Gilbert *et al*., 2002). However the vast majority of L1 insertions have identifiable TSDs, suggesting they originate in an endonuclease dependent process.

1.4.2 Mechanism of retrotransposition

The mechanism of retrotransposition of non-LTR retrotransposons is not entirely understood. However, the first steps of integration of these elements have been elucidated by biochemical experiments on the site-specific RE-type retrotransposon R2BM from the silkworm *Bombyx mori* (Luan *et al*., 1993). These studies led to the model of L1 retrotransposition called 'target primed reverse transcription' (TPRT) (Cost *et al*., 2002), illustrated in Figure 1.5.

Although RE-type and APE-type elements belong to different families of non-LTR retrotransposons that share very few structural similarities, the basic mechanism of transposition initiation by TPRT is relatively conserved. This has been demonstrated by reconstitution of the initial steps of L1 element transposition *in vitro*, by providing only the complete L1 ORF2 protein, L1 RNA, and a target DNA in appropriate buffers (Cost *et al*., 2002). Also, further experiments have shown that the EN domains of the two types of retrotransposons (RE and APE) initiate the integration process by nicking the target DNA (Cost *et al*., 2002; Eickbush and Malik, 2002). The generated 3′hydroxyl group serves as a primer for reverse transcription of the element's RNA. It has been demonstrated that L1 integration can also occur at pre-formed nicks and

double strand breaks in the target DNA, which is known as endonuclease independent-TPRT (Morrish *et al*., 2002). Therefore, endonuclease-independent insertion provides an alternative pathway for L1 retrotransposition in the human genome (Sen *et al*., 2007). It has thus been concluded that nicking and reverse transcription are two independent steps in TPRT (Cost *et al*., 2002; Eickbush and Malik, 2002).

The EN domain, can also cleave the second strand of target DNA at a slower rate compared to the nicking of the first strand (Cost *et al*., 2002). Depending on the position of the second nicking site relative to the initial one, TPRT can generate a target site deletion, a simple 'blunt' integration, or a target site duplication (TSD) which flanks the inserted element (Cost *et al.,* 2002; Eickbush and Malik, 2002).

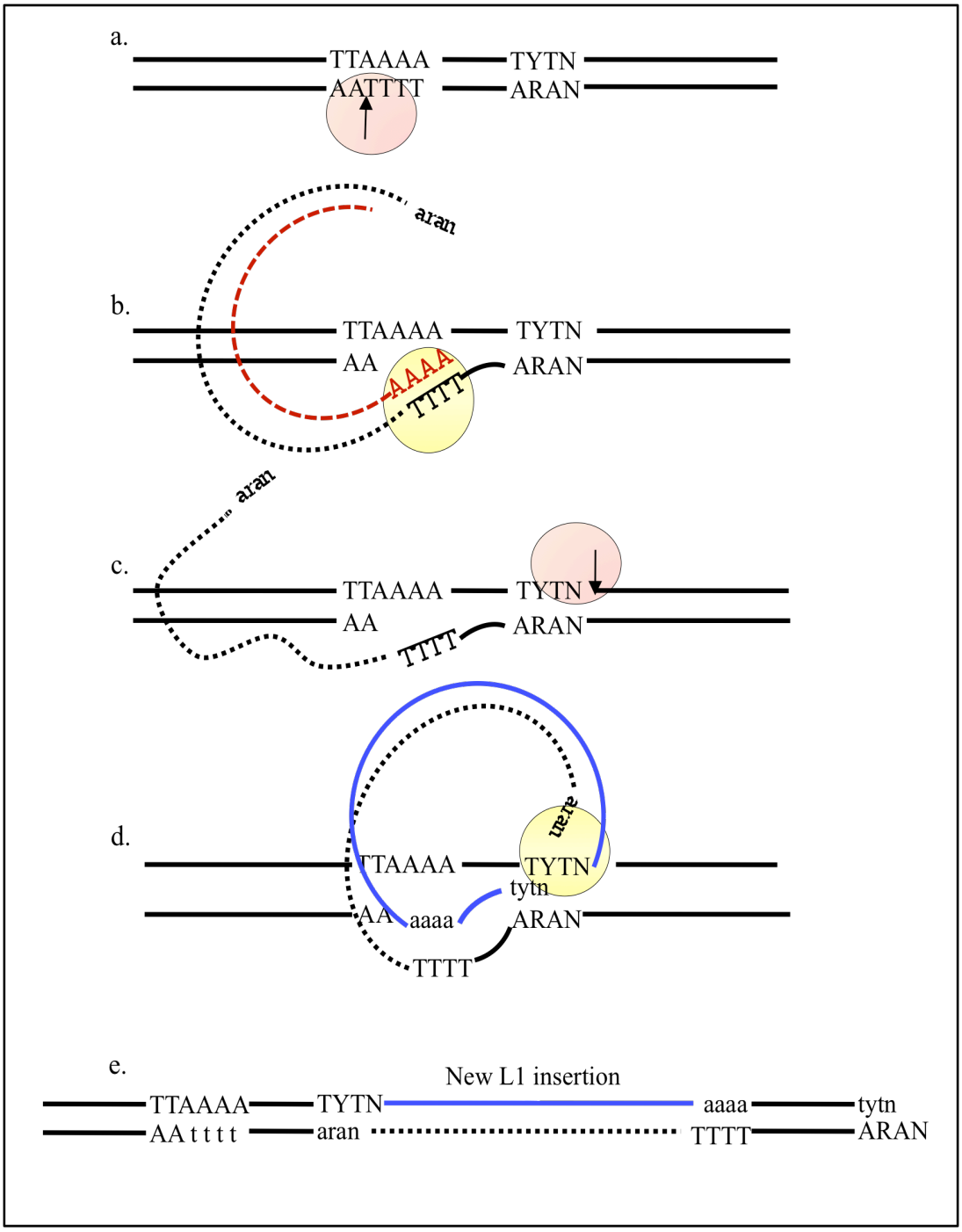

Figure 1.4 Schematic diagram of 'target primed reverse transcription' (TPRT) mechanism, **a.** First strand cleavage by the L1 endonuclease enzyme (pink oval) producing 3´ OH and 5´ POH, **b.** annealing of L1 RNA (red dotted line) to the poly T repeat and first cDNA strand synthesis (black dotted line) by the reverse transcriptase enzyme (yellow circle), **c.** Second strand cleavage by the L1-EN enzyme, **d.** Second strand cDNA (blue line) synthesis catalysed by RT, using the first strand cDNA as a template, **e.** Integration of new L1 insertion and production of target site duplications (the synthesised TSDs are shown in lower case letters). Adapted from Luan *et al.*, 1993; Martin and Bushman, 2001; Eickbush and Malik, 2002.

A major unresolved issue regarding the mechanism of LINE retrotransposition is what occurs after second-strand cleavage. Despite extensive efforts, *in vitro* experiments with R2 protein did not lead to the detection of intermediates expected for second-strand synthesis (Luan *et al*., 1993). In contrast, *in vitro* TPRT of L1 yielded 5′ junctions between the L1 sequence and the target DNA. This result indicates that the RT is able to accept cDNA as a template for second-strand synthesis, probably by a second round of TPRT (Cost *et al*., 2002; Eickbush and Malik, 2002).

However, this *in vitro* process is very inefficient and it does not necessarily reflect the natural mode of retrotransposon integration and still leaves open the question of how the damaged genomic DNA is repaired. It is generally assumed that cellular DNA repair pathways are involved in these final steps of integration and that these activities generate the observed TSDs.

1.5 Genomic distribution of human L1s

Human LINEs are distributed all across the genome, but not distributed evenly. There are some parts of the genome, which have very low repeat density. This could be because these regions cannot tolerate insertion of repeats due to essential *cis* regulatory architecture. An example of repeat poor regions are the homeobox (HOX) gene clusters, which contain the lowest reported density of interspersed repeats (Lander *et al*., 2001). In contrast to this some parts of the genome are very rich in repeats, such as chromosome Xp11, which contains a 525 kb region comprised of 89% repeats. Overall it is suggested that LINEs are more abundant in gene poor, and thus AT rich regions, which usually show low recombination rates (Lander *et al*., 2001). In comparison to Alu, LINEs have been reported to insert at a four fold higher density in GC poor regions, while Alus have a lower tendency (five fold lower) to insert in AT rich regions (Lander *et al*., 2001). One reason for this insertional bias of LINEs towards the AT rich regions is suggested to be due to the consensus L1 endonuclease target site TT/AAAA, which is more common in AT rich regions (Lander *et al*., 2001; Jurka, 1997; Cost and Boeke, 1998). However, Alu elements also use the L1 machinery in *trans* to integrate into the genome, but Alus have a high density in GC rich regions. Therefore, the biasing of L1 insertion in the AT rich

region may not be only due to endonuclease site selection but could also be a part of the hosts evolutionary response to this mutagenic elements. It has been suggested that L1 insertion occurs in AT and GC rich regions, but that insertions in GC-rich regions are lost through selection. It is clear that L1s inserted within genes can have a variety of negative effects on their host gene such as altered splicing, interference with gene regulation and level of expression, and premature polyadenylation (Cost and Boeke, 1998; Lander *et al*., 2001).

1.6 Impact of L1 integration on human genome plasticity

Recently, efforts have been directed at unveiling the molecular mechanisms by which L1 impacts gene expression and mammalian cell development, differentiation, and cancer. New L1 integrations have a great impact on host genome diversification and thus evolution. The ways that L1 retrotransposition can alter the host genome are discussed in detail below.

1.6.1 Increasing the size of the human genome

An orthologous sequence comparison of the human and chimpanzee genomes suggested that the human genome continues to expand, either because of inherent mutational mechanisms or through being less efficient at deleting such events, or perhaps because of shifts in retrotransposition activity (Liu *et al*., 2003). Therefore, one of the greatest impacts of L1 on the human genome is their contribution to expanding genome size through an ongoing accumulation process (Liu *et al*., 2003). Considering that L1 is also responsible for Alu retrotransposition in the genome, it contributes about 750 Mb to the human genome (Lander *et al*., 2001). Moreover, the ongoing expansion of L1 has also created significant inter- and intra-individual variation by introducing L1 insertional polymorphisms (presence / absence) at orthologous loci.

1.6.2 Disease causing L1 retrotransposition

There are 17 cases of human genetic diseases caused by L1 integration into genes, and they are estimated to account for approximately 1 in 1200 human pathogenic mutations (Kazazian *et al*., 2004). Based on L1 retrotransposition assays it has been suggested that about 10% of *de novo* L1 retrotransposition events occur in the introns of actively transcribed genes (Moran *et al*., 1999). In fact, recent studies have suggested that evolutionarily successful L1s (active L1s) preferentially insert into genes, which are transcriptionally active and therefore have an open chromatin configuration (Macia *et al*., 2011).

The first L1 disease-causing insertion was reported in two patients with haemophilia, where an L1 was integrated into exon 14 of the human factor eight gene (Kazazian *et al*., 1988). Subsequently cases of L1 disruption of the dystrophin gene have been reported to cause muscular dystrophy and cardiomyopathy in four unrelated individuals (Holmes *et al*., 1994; Matsuo *et al*., 1991 and Yoshida *et al*., 1998). It has also been shown that a heritable full length L1 insertion into intron two of the ß-globin gene (L1ß-thal) is responsible for some cases of ß-thalassemia (Divokey *et al*., 1996; Kimberland *et al*., 1999). Also insertion of a full length L1 into an intron of the X-linked RP2 gene is responsible for progressive retinal degeneration and ultimately retinitis pigmentosa (XLRP) (Schwahn *et al*., 1998). Moreover, a case of colon cancer has reported to be caused by somatic insertion of a truncated L1 into the APC gene (Miki *et al*., 1992). More recently it has been reported that somatic *de novo* L1 retrotransposition events are detectable in lung cancer cells (Iskow *et al*., 2011). Also, up regulation of L1 RNA and OPRF1p has been reported in several tumours including breast sarcomas and in 10% of tumours of germline origin, such as ovarian and testicular tumours (Asch *et al*., 1996; Bratthauer and Fanning, 1993).

1.6.3 Genome instability caused by L1 retrotransposition

In addition to mutagenic insertions, L1 retrotransposition can generate local genomic instability through several other mechanisms, which are explored in this section. DNA double strand breaks (DSBs) can be caused by endogenous L1ORF2p, which has an

endonuclease activity (Gasior *et al*., 2006). It is been shown that the number of DNA DSBs generated by L1ORF2p is much higher than the number of actual L1 insertions (Gasior *et al*., 2006). However, the extent of genome instability introduced by endogenous L1 retrotransposition is not clear due to a lack of sensitive antibodies to target ORF2p and also because the repair of L1-mediated DSBs does not leave any sign of L1ORF2p involvement. As a result the attribution of L1ORF2p to genomic DSBs, which are highly mutagenic and prone to induce recombination, is underestimated (Cordaux and Batzer, 2009). In addition, to generating genome instability L1 can also cause genomic rearrangements through insertion-mediated deletions. Studies on L1 retrotransposition in cell culture have demonstrated that about 20% of L1 insertions are associated with structural rearrangements, including flanking genomic deletions at the insertion site (Gilbert *et al*., 2002; Gilbert *et al*., 2005; Symer *et al*., 2002). Another study reported a lower frequency of deletion (2%) when compared to cell culture assays, with endogenous L1 retrotransposition causing deletions with an average size of 800 bp in the human genome (Han *et al*., 2005). Since L1-mediated insertion deletions are generally grouped into two sizes <100 bp and >1kb, it is suggested that each group is caused by a different mechanism. In general, small deletions may arise due to template switching with subsequent 5′to 3′exonuclease activity on both the exposed 5′ ends. Larger deletions can be mediated by non-homologous end joining when the nascent cDNA invades a double strand break with a 3′ overhang located upstream of the integration site. Subsequent gap repair will remove the cDNA and the adjacent segment to cause a large deletion (Han *et al*., 2005). A study by Mine *et al*. (2007) has demonstrated a 46 kb full length L1 insertion-mediated deletion event that possibly occurred through the template jumping process. This deletion results in removal of seven exons of the pyruvate dehydrogenase complex, component X (PDHX) gene, which causes pyruvate dehydrogenase complex deficiency.

1.6.4 Ectopic recombination upon L1 retrotransposition

Due to the high copy number of L1 in the human genome they can also create structural variation at the post-integration stage through non-allelic homologous recombination or ectopic recombination. Ectopic recombination events seem

relatively rare and are usually mediated by truncated elements (Boissinot *et al*., 2000). Indeed there is no evidence of polymorphic L1 associated ectopic recombination in humans. This can be explained by the low activity of retrotransposition competent L1 in the modern human genome (Boissinot *et al*., 2000), or perhaps by the frequency with such mutations are deleterious (Song *et al*., 2006). Ectopic recombination causes various types of genomic rearrangements, including duplications, deletions, and inversions.

It has been proposed that ectopic recombination between Alu elements is one mechanism for the generation of segmental duplications, which are duplicated blocks of sequences ranging from 1 kb to 300 bp in size (Bailey *et al*., 2003; Kazazian *et al*., 2004). Segmentally duplicated regions can contain paralogous copies of genes, promoters and other regulatory components (Samonte and Eichler, 2002). It has been suggested that segmentally duplicated regions are associated with the creation of novel genes and the formation of pseudogenes (Lynch and Conery, 2000). Moreover, ectopic recombination can cause recombination-associated deletion events (RADs). Genome-wide comparisons of the human and chimpanzee genomes have identified 73 human specific L1RADs events that occurred following the divergence of humans from chimpanzees (Han *et al*., 2008). Although L1RAD events are not very common, it has been suggested that they are responsible for the deletion of about 450 kb of the human genome (Han *et al*., 2008). This event is most frequent in heterochromatic regions, which suggests that there may be negative selection against L1RADs in euchromatin (Boissinot *et al*., 2006).

As mentioned earlier, L1-mediated ectopic recombination is also involved in gene inversion events. Comparisons of the inversions events that are present in the human genome but absent from the chimpanzee have demonstrated that nearly half of these inversions were associated with L1 and Alu elements (Lee *et al*., 2008). It is suggested that L1 contributes to genomic inversion possibly through the formation of secondary structures or by providing a target site for double strand breaks (Lee *et al*., 2008). Among the characterised inversions mediated by L1 insertions, some include the exonic regions of known genes, which suggest that L1-mediated inversions can generate alterations in gene function (Lee *et al*., 2008; Cordaux *et al*., 2009). Therefore, although this type of recombination does not affect the size of the genome it can produce genomic variation.

1.6.5 L1-mediated sequence transduction

In addition to duplicating themselves, L1s sometimes carry upstream or downstream flanking genomic sequences (termed 5′ and 3′ transduction, respectively) with them, providing a novel mechanism for genome evolution. L1-mediated sequence transduction occurs when L1 transcripts extend upstream or downstream of the genomic flank and then transduce these sequences into new genomic locations through the L1 retrotransposition process. L1 5´ sequence transduction is usually very short, ranging between 5-8 nt sequences and it is not a common process. This process occurs when L1 sequences are transcribed by a host promoter upstream of the L1 5´ terminus, and subsequently mobilised during the L1 retrotransposition cycle (Pavlicek *et al*., 2002a; Pickeral *et al.*, 2000; Szak *et al*., 2003). The 3´ sequence transduction process is more common, and occurs when transcription of the L1 bypasses the weak polyadenylation (Poly-A) signal in favour of a stronger canonical Poly-A signal in the 3´ genomic flank, followed by mobilisation of the genomic flanking DNA to a new location. It has been demonstrated that L1 is capable of mobilising up to 2kb of flanking sequences down stream of its 3´ site without perturbing the retrotransposition process (Moran *et al*., 1999). The sequence transduction process seems to be more common in active or recently active elements: it has been demonstrated, in cell culture assays that between 10%-20% of recent active human insertions contains sequence transductions (Goodier *et al*., 2000). Also, in some cases sequence transduction caused by L1 retrotransposition may not be identified as such, due to the extensive L1 5´ truncation at the site of integration (Pavlicek *et al*., 2001). During the process of sequence transduction, exons, promoters and other regulatory sequences upstream and downstream of the L1 can be transduced into the new genomic location and cause exon shuffling potentially altering the expression and or structure of the active gene. This process maintains genome plasticity and genome evolution (Goodier *et al*., 2000). It has been suggested that genome shuffling caused by retrotransposons has had a role in the divergence of humans from chimpanzee (Brosius, 1999).

1.6.6 Regulation of gene expression

As mentioned above, L1s can affect the genome at the DNA level. In this section more details of the effect of L1 at the RNA level are discussed. It has been demonstrated that L1 can affect transcription in several different ways. They can generate alternative splice sites, and intronic L1s may sometimes interfere with transcriptional elongation and produce different lengths of mRNA from a gene. If the L1 inserts in the antisense orientation relative to the native genes, it can potentially produce truncated cellular transcripts by premature polyadenylation (Han *et al*., 2004). Moreover, L1 can produce novel transcripts by the activity of its antisense promoter (ASP). Nearly 1/3 of the L1Hs studied contain active ASP (Speek *et al*., 2001). Therefore it is possible that some of the transcripts initiated from the L1 ASPs are translationally competent. In addition, insertion of L1 into an intronic region of a gene can potentially "break" a gene where an L1 inserted in the opposite orientation to a host gene can generate two novel partial transcripts: one from the endogenous promoter including exons upstream of the L1 insertion, and a second internal transcript driven by the L1 ASP. Indeed, bioinformatic analysis of the human genome sequence has highlighted 15 genes and transcription units that have potentially been affected by L1 insertions in this way (Wheelan *et al.*, 2005). Finally a recent study of intragenic L1s in lung cancer cells has shown that L1 pre-mRNA binds to the Ago2 complex to suppress the transcription of cancer genes (Aporntewan *et al*., 2011). Therefore, with interference of L1 endogenous sense and antisense promoters, polyadenylation signal, and L1 transcripts, L1 exhibits a great potential to impact human transcriptome diversification.

1.6.7 Epigenetic regulatory role of human L1s

Because L1 elements are frequently found in or near genes, heterochromatin formed at retrotransposons could spread and repress the transcription of nearby genes. One example of L1's epigenetic regulatory role is in X chromosome inactivation, likely to be mediated by L1 retrotransposons. X chromosome inactivation is a well-established mechanism of gene regulation that acts to achieve gene dosage compensation between male and female embryos. During female early embryogenesis, the majority of genes

on a randomly selected X chromosome are transcriptionally silenced to maintain X-linked gene dosage between female (XX) and male (XY) individuals. X inactivation initiates at the X inactivation centre (XIC) (Rastan, 1983), which contains several genes that produce non-coding RNAs (Chureau *et al*., 2002). The Xist RNA (~1000 bp) is uniquely expressed from the female inactivated X chromosome in somatic cells and acts to 'coat' the X chromosome, to silence it (Borsani *et al*., 1991). X inactivation has three stages, beginning with X inactivation initiation during early embryogenesis, followed by spreading from the XIC in *cis* along the 150 Mb of the X chromosome, and finally maintenance of imprinted X- linked genes during successive somatic cell divisions (Bailey *et al*., 2000). Little is known about how XIC spreads across the chromosome, although it has been proposed that L1s play a role in the *cis* spreading of X chromosome inactivation (Lyon, 1998). L1s are enriched on the X chromosome compared to autosomes, and significantly so at Xq13 where the XIC is located. To support this idea it has been demonstrated that genes on the X chromosome which escape X inactivation are generally located in L1 poor regions (Ross *et al*., 2005). In addition, the failure of chromosomal inactivation beyond the boundaries of X chromosomal sections in X: autosomal translocations are correlated with chromosome bands with low L1 density (Bailey *et al*., 2000). Recently it has been demonstrated that L1 participates in the generation of silent nuclear compartments where silenced genes become recruited (Chow *et al*., 2010). More importantly it has been shown that young L1s, which are more likely to be active, can escape X chromosome inactivation and are expressed on the silenced copy of the X chromosome (Chow *et al*., 2010). It is suggested that truncated L1s, which are silent on the X chromosome are involved in the assembly of heterochromatic nuclear compartments induced by Xist, while active L1s are involved in the local spreading of XCI into regions that would be prone to the escape of X inactivation (Chow *et al*., 2010). For young L1s the proposed involvement in X inactivation is also linked to methylation. Indeed, It has been shown that demethylation and activating of the L1 ASP can drive the transcription of neighboring genes; Weber *et al*. (2010) have shown that demethylation of L1ASP in colon cancer cell lines induces the expression of L1 and proto-oncogene cMet (L1-cMet) transcripts. This result showed the involvement of L1 in gene regulation and its link to methylation. However, the formal demonstration of direct retrotransposon-mediated epigenetic control of neighboring genes in humans and the evaluation of the extent of this phenomenon at a genome-

wide scale are active topics of investigation.

1.7 Host defence mechanisms against L1 retrotransposition

As well as the direct mutational effects of L1 insertion, L1 has also been associated with genetic instability in the human genome (Symer *et al.,* 2002). Various forms of genetic instability caused by L1 integration include the generation of L1 chimeras, intrachromosomal deletions (chromosomal deletions of >11 kb), intrachromosomal duplications, and chromosomal inversions (approximately 120 kb in length) (Gilbert *et al*., 2002; Han *et al*., 2005; Symer *et al*., 2002). It has been demonstrated by Gilbert *et al*. (2005) that the L1 reverse transcriptase can faithfully replicate its own transcript and has a base misincorporation error rate of ~1 in 7,000 bases. All these observations indicate that L1 retrotransposition can lead to a variety of genomic rearrangements and suggests that host processes should be under selection to restrict L1 activity, as integration of L1 and other retrotransposons poses a potential threat. As a result organisms have evolved diverse mechanisms to combat retrotransposon activity. Indeed, the initial step in L1 retrotransposition was described as a host/parasite "battleground" that serves to limit the number of active L1s in the genome (Gilbert *et al*., 2005). Since L1 has been actively mutating mammalian genomes for a long time, it is likely that the host has evolved multiple mechanisms to combat L1 mobility at discrete steps of the retrotransposition cycle. In the following sections the mechanistic strategies used by the host to restrict L1 retrotransposition are discussed in more detail.

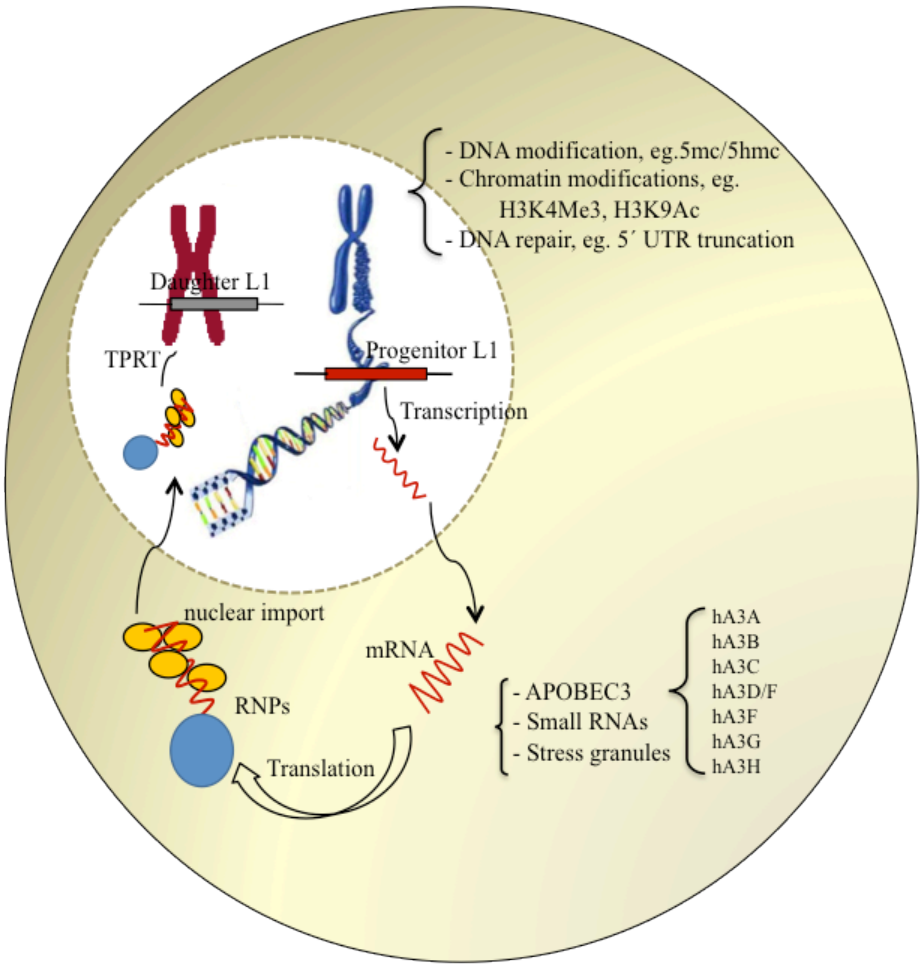

Figure 1.5 Schematic diagram of host defence mechanisms against endogenous L1 retrotransposition; defence mechanisms applied at different stages of L1 retrotransposition. Based on the location in the cell, the defence mechanisms against L1 retrotransposition can be divided into two categories: nuclear and cytoplasmic defences. Most of the defence mechanisms against endogenous L1 retrotransposition, and their timing are not understood in detail. Different studies have contributed to this diagram, and they are cited in section: 1.7.

1.7.1 Epigenetic modifications regulate L1 retrotransposition

Studies on 5-methylcytosine residues in the L1 promoter, especially at the four transcriptionally important CpG sites, show that DNA methylation can repress L1 activity both *in vivo* and *in vitro* (Hata and Sakaki 1997). In contrast to the suppressive effect of DNA methylation on L1 promoters, it has been demonstrated that 5-hydroxylation of the methylcytosine moiety (hm5c) can be an activating factor However, a study of hm5c protein interactions showed that it does not interact with the same proteins as the 5mc pathway, which suggests that hm5c must regulate the L1

promoter through other mechanisms (Williams *et al*., 2011). Indeed, Ficz *et al*. (2011) demonstrated that hm5c methylation modifications are enriched in euchromatic regions and show a positive correlation with L1 expression, at least in the mouse. Also, a recent study has demonstrated that the Tet protein can generate other cytosine modifications downstream of hm5c. These modifications are 5-formylcytosine (5fc) and 5-carboxylcytosine (5ca5) (Ito *et al*., 2011). Whether these newly discovered DNA cytosine modifications have any direct and controlling effect on L1 promoters and L1 expression remains to be investigated.

Many studies have shown that a variety of epigenetic modifications can regulate L1 activity, and these are not limited to DNA modifications. Chromatin modifications are also likely to have an important role in controlling L1 activity, for example Teneng *et al*. (2011) have recently demonstrated the direct association of H3K4 and H3K9 modifications with L1 activity. In fact they demonstrated that the exposure of HeLa cells to Benzo (a) pyrene (Bap) causes L1 reactivation in HeLa cells through induction of early enrichment of the transcriptionally active chromatin markers histone H3 trimethylation at lysine 4 (H3K4Me3) and histone H3 acetylation at lysine 9 (H3K9Ac), and also reduces association of DNMT1 with the L1 promoter. These processes cause depletion in cellular DNMT1 expression, which subsequently reduces cytosine methylation within the L1 promoter CpG island (Teneng *et al*., 2011).

Other evidence for chromatin modifications regulating L1 activity was uncovered in hippocampus neural (HCN) stem cells. Muotri *et al*. (2005) showed that histone deacetylase 1 (HDAC1) and methylation of H3 at Lys9 (K9), which both associate with transcriptional silencing in undifferentiated HCN cells, was directly correlated with L1 reporter construct activity in transgenic mice, whereas acetylation of H3K9 and methylation of H3K4 (associated with transcriptional activation) associated with high levels of L1 transcripts in HCN differentiated cells. This data supports the idea that chromatin remodelling during the early stages of neuronal cell differentiation allows transient stimulation of L1 retrotransposition (Muotri *et al*., 2005).

1.7.2 Role of small RNAs in regulation of L1 retrotransposition

Small RNAs inhibit retrotransposon proliferation in the host genome via two mechanisms, which are independently directed by either small RNA interference (siRNAs) or PIWI-interacting RNAs (piRNAs) (Meister *et al*., 2004; Soifer *et al*., 2006). The mechanisms by which these small RNAs are generated and how they inhibit retrotransposon mRNAs are still not fully understood, but there is strong evidence for a connection. It has been reported that host siRNA can repress retrotransposition through the post-transcriptional disruption of L1mRNA. In principal double stranded RNAs (dsRNAs) of 21-23 nt in length can be generated from simultaneous sense and antisense expression of L1s (Ketting *et al*., 1999). The dsRNAs bind to Dicer proteins and are subsequently processed and cleaved into a single stranded siRNA. The siRNAs, which are complementary to the L1 mRNA, are selectively incorporated into the RNA induced silencing complex (RISC). As a result the siRNA directs RISC to the L1 mRNA, which leads to L1 mRNA degradation (Wu-Scharf *et al*., 2000, Aravin *et al.,* 2007, Levin *et al*., 2011).

Another mechanism that has been suggested to suppress retrotransposon mRNA are piRNAs, which are generated from genomic loci that encode long precursor RNAs containing the remnants of different families of TE elements (Malone *et al*., 2009). Malone *et al*. (2009) demonstrated that in the *Drosophila* germline a premature piRNA transcript, which contains sequences derived from TEs is processed into a mature piRNA (24-35 nt). After this processing step, a subfamily of Argonaute proteins, known as the PIWI clade of proteins, bind to mature piRNAs and direct them to complementary sequences in TE mRNA. The mature antisense piRNA binds to PIWI proteins, and this directs the complementary TE mRNA to the complex, inducing the endonucleolytic cleavage of the mRNA and subsequently formation of the second sense piRNA. The sense piRNA then binds to one of the argonate 3 (AGO3) proteins, and this complex then directs the complementary sequence in the original precursor piRNA and causes the endonucleolytic cleavage and production of antisense piRNA. This cycle, which is known as the "ping-pong cycle", leads to the destruction of TE mRNA in the germline (Malone *et al*., 2009).

It can be speculated that small-RNA-based mechanisms like this may also play role in silencing mammalian L1 elements. It has been demonstrated that an antisense

promoter located within the human L1 5′ UTR allows the production of an antisense RNA transcript (Speek *et al*., 2001) that, in principle, could base pair with sense-strand L1 mRNA to establish a dsRNA substrate for Dicer protein (Levin *et al*., 2011). Furthermore, mouse mutants lacking the murine PIWI family proteins (MILI or MIWI2) exhibit a loss of methylation of L1 and IAP. This loss correlates with their transcriptional activation in male germ cells and suggests that MILI and MIWI2 play essential roles in establishing *de novo* DNA methylation of L1 retrotransposons in the fetal male germline (Kuramochi-Miyagawa *et al*., 2008).

1.7.3 RNA editing enzymes modulating the L1 retrotransposition

Members of the apolipoprotein B mRNA editing complex polypeptide 1-like (APOBEC) family of enzymes exhibit modulatory activity against variants of exogenous and endogenous retrovirus-like elements, including L1 retrotransposons. It has been demonstrated that APOBEC3A, APOBEC3B and APOBEC3F suppress L1 retrotransposition in humans and IAP elements in mouse (Lovsin and Peterlin 2009). However, an immunoprecipitation and immunofluorescence study on the interaction of APOBEC3 proteins 3A and 3B showed that there is no direct interaction of 3A with L1 proteins, and although the 3B protein binds to ORF1, it does not co-localise with ORF1p, which suggests that the APOBEC3 protein indirectly suppresses the activity of L1, possibly through interference RNA targeting (Lovsin and Peterlin 2009). Recently a knockdown study of APOBECs and their effect on L1 retrotransposition in hESC and iPS cells has shown that only knockdown of APOBEC3B enhances L1 retrotransposition in hESCs. Knockdown of other APOBEC3 family members does not have any effect on L1 retrotransposition (Wissing *et al.,* 2011).

Moreover, previous studies suggest that APOBEC3B and APBEC3F repress the L1 retrotransposition process in a deamination-independent pathway, and it is more likely that they repress L1 retrotransposition by producing L1 integration barriers (Stenglein *et al*., 2006). Recent studies on the activation-induced deaminase (AID)-like gene, which is the potential ancestral progenitor of APOBEC lineages in mammals, demonstrated that AID could inhibit the retrotransposition of L1 through a DNA

deamination-independent mechanism (MacDuff *et al*., 2009). This mechanism may manifest in the cytoplasmic compartment, co- or post-translationally, and suggests that APOBEC proteins might also exhibit similar inhibitory reactions in L1-mediated retrotransposition (MacDuff *et al*., 2009).

1.7.4 L1- ribonucleoprotein particles and host cellular defence

Despite long study the processes involved in the formation of L1 ribonucleoprotein (RNP) particles and their transportation to the cell remains unclear. Due to the suppression of L1 retrotransposon expression in most somatic cells and the association of L1 with many cellular mRNAs, it is difficult to directly detect endogenous L1RNPs. Recently Goodier *et al*. (2008) have demonstrated the subcellular co-localisation of L1 RNA and proteins (ORF1p and ORF2p), in cytoplasmic RNP foci. One of the suggested cellular host defence mechanisms to repress L1 retrotransposition is the transport of L1 RNPs to stress granules. It had previously been demonstrated that L1RNP foci also localise with nucleoli (Goodier *et al*., 2007). Indeed, the endogenous endonuclease of Ty3 and Gag proteins of Tf1 retrotransposons are localised with the nucleolus in yeast cells, and these proteins are possibly responsible for retrotransposon nuclear transport (Lin *et al*., 2001; Teyseet *et al*., 2003). The nucleolus has several different roles, which are mainly involved in protein assembly (ribosome biogenesis and RNP assembly) and RNA export for retroviruses and LTR retrotransposons (Goodier *et al*., 2007). The discovery of a new family of L1 chimeras in the genome showed that L1 RNAs frequently recombine with small RNA sequences such as U3, U5 and U6, which are transcribed and evenly expressed in the human genome (Buzdin *et al*., 2003). In addition it is known that these small RNA are frequently trafficked through the nucleolus, and so it is likely that L1 RNPs interact with small RNAs in the nucleolus (Goodier *et al*., 2007). Furthermore, mutation analysis has revealed that ORF1p can direct L1 RNP distribution in the cell, probably through its secondary structure and its affinity to bind macromolecules (Goodier *et al.,* 2008). However, non-autonomous non-LTRs such as Alu and SVA manifest subcellular colocalisation different from that of L1RNPs, despite the fact that they utilise the L1 enzymatic machinery (Goodier *et al*., 2008). This variability suggests differences in the retrotransposition cycle of these

elements and therefore host defence mechanisms may act differently to suppress them.

Cytoplasmic RNA granules in somatic cells, stress granules, and processing bodies, have emerged as important players in post-transcriptional regulation of gene expression. Stress granules are discrete cytoplasmic aggregates, which are induced by a range of stress conditions, such as viral infection and over expression of some cellular proteins (Anderson and Kedersha, 2006). Processing bodies are dynamic cytoplasmic compartments, which contain high concentrations of molecules that are involved in mRNA decay and translation inhibition (Goodier *et al*., 2007). In mammalian cells, the RNA induced silencing complex (RISC), which is the main pathway for degrading retrotransposon-derived mRNA, has been found in processing bodies, including Argonaute 2 (Ago2), an RNA binding protein necessary for miRNA silencing. Also, APOBEC3G exists in processing bodies and there is evidence of its trafficking from processing bodies to stress granules (Goodier *et al*., 2007; Gallois-Montbrun *et al*., 2007).

The processing bodies and stress granules are related compartments that overlap and share some components depending upon the nature of the cellular stress. It has been suggested that stress granules control whether the mRNA should be transferred into processing bodies for degradation or returned to polyribosomes for translation. Goodier *et al*., (2007) demonstrated that ORF1p foci co-localise with cytoplasmic stress granules in both stress and unstressed conditions. However, in unstressed conditions fewer ORF1p foci engaged with stress granules. The discovery of L1ORF1p and L1 polyadenylated RNA in stress granules suggests a mechanism for host defence against the potential mutagenic effects of retrotransposition by migrating L1RNPs to stress granules and further degradation of L1mRNA in processing bodies. However, this does not rule out the possibility that the stress granules may be involved in the retrotransposition life cycle rather than their degradation, *i.e*. in stress conditions they may stop ORF1p translation and after the stress has passed they may redirect the L1RNPs to the polyribosomes for translation.

1.7.5 L1 post-translational host defence mechanisms

It has been estimated that about 95% of the L1 retrotransposons are 5′ truncated in the human genome and therefore are not retrotranspositionally competent. L1 5′ truncation is perhaps a result of the low processivity of non-LTR endogenous reverse transcriptase, resulting in premature termination of reverse transcription. However, a study of the activity of the reverse transcriptase of R2 (a non-LTR retrotransposon) in *Bombyx mori* has demonstrated that while the non-LTR retrotransposon reverse transcriptases are very divergent, their functions are similar to the retroviral enzymes. However non-LTR RTs are more processive than the reverse transcriptases encoded by retroviruses (Eickbush and Jamburuthugoda, 2008). Hence it may be more likely, that L1 5′ truncation is a result of a host defence mechanism acting post-translationally. It is speculated that L1 5′ truncation occurs during TPRT through dissociation of the L1 reverse transcriptase from the cDNA or L1 mRNA degradation (Levin *et al*., 2011). Study of R2 elements suggests that for these elements cDNA synthesis is required for the second genomic nick. Therefore, the flap intermediate (L1 cDNA and L1 mRNA) could represent a stable intermediate during TPRT. Hence, DNA repair proteins that recognise this type of 3′ flap intermediate may be able to process the L1 integration intermediate (Gilbert *et al*., 2005). This leads to the idea that in order for full-length L1 integration to occur TPRT must be completed before the host defence can recognise the intermediate and act upon its integration (Gilbert *et al*., 2005). A DNA excision repair pathway protein complex that is known to recognise 3′ flap intermediates is ERCC1-XPF. Indeed, cell culture studies by Gasior *et al*. (2008) have demonstrated that the ERCC1 deficiency in hamster cells results in an increase in frequency of L1 retrotransposition, suggesting there may be a causal link.

1.8. Effects of reprogramming in early human development on L1 activity

In mammals, both parental genomes undergo dramatic epigenetic changes after fertilisation to form the diploid somatic genome. Epigenetic reprogramming, including demethylation of genomic DNA, occurs in mammalian primordial germ

cells (PGCs) and in early embryos and is important for the erasure of imprints and epimutations as well as the return to pluripotency (Reik *et al*., 2001). Reprogramming in early embryos occurs through both active and passive demethylation mechanisms (Young and Beaujean 2004). However, the exact demethylation process is not clear. It has been demonstrated that in mice the paternal genome undergoes a massive active demethylation process within 6 to 8 hours of fertilisation in the egg cytoplasm (Reik *et al*., 2001). In contrast the maternal genome undergoes a passive demethylation after several cleavage divisions (Mayer *et al.,* 2000). Moreover, it has been suggested that the active demethylation of the paternal genome may be associated with epigenetic chromatin remodelling in the sperm genome and in this way establishes parent-specific developmental programmes during early embryogenesis (Mayer *et al.,* 2000).

Studies of human embryonic stem cell lines and embryonal tumour cells (including those derived from germ line cells, NTera2D1, PA1) have detected L1 proteins in these cell lines, while L1 proteins are barely detectable in differentiated and normal somatic tissues and cells (Leibold *et al*., 1990; Bratthauer *et al*., 1994; Hata and Sakaki 1997; Woodcock *et al.,* 1997). Therefore, it is possible that L1s become active during early embryogenesis and could possibly play regulatory roles in embryonic development. Indeed, it has been demonstrated that *in vitro* RNAi targeting of L1 sequences in the male pronucleus leads to subsequent developmental arrest in mouse embryos at two and four-cell stages (Thurston *et al*., 2007; Beraldi *et al*., 2006). This result suggests the need for L1 regulation during embryonic development.

1.9 Ongoing L1 retrotransposition in different tissues

Due to the disease-causing potential of L1 retrotransposition, the host genome has an evolutionary advantage if transposition is down regulated in germline and somatic cells. However, since L1 can only propagate by vertical transmission, L1 expression and transposition must occur in cells contributing to the germline *(i.e.* germ cells or early embryonal cells) in order to proliferate (Ergün *et al*., 2004). Although it is estimated that up to 5% of newborns may contain a *de novo* L1-mediated retrotransposition event (Naas *et al*., 1998; Garcia-Perez *et al*., 2007), relatively little is known about the developmental timing or cell types that accommodate LINE-1

retrotransposition in humans. *In vivo* studies using mouse models indicate that LINE-1 expression or retrotransposition can occur in male and female germ cells during early development, and also in select somatic tissues (Kidwell and Lisch 2000; Brouha *et al*., 2003; Mine et al., 2007). A recent study has also demonstrated that *de novo* L1 insertions can occur in lung cancer (Iskow *et al*., 2011). In addition, L1 retrotransposition events must occur in the germline or in early human embryogenesis before germline differentiation in order to be evolutionary effective (Ergun *et al*., 2004). An *in vitro* retrotransposition assay has been used to demonstrate exogenous LINE-1 retrotransposition in a variety of human and rodent transformed cell lines (Ostertag *et al*., 2002; Ergun *et al*., 2004; Garcia-Perez *et al*., 2007), in rat neuronal progenitor cells (Ostertag *et al*., 2002), and at a low level in primary human fibroblasts (Bruke *et al*., 1998; Brouha *et al*., 2003). It has been shown that human embryonic stem cells can accommodate the retrotransposition of engineered LINE-1 elements *in vitro* (Garcia-Perez *et al*., 2007). These data suggest that LINE-1 retrotransposition events may occur at early stages in human embryogenesis and that some individuals in the population may be genetic mosaics with respect to their LINE-1 content (Van den Hurk *et al*., 2007). Below three potential environments for *de novo* L1 retrotransposition are discussed in more detail.

1.9.1 L1 retrotransposition in neuronal progenitor cells

The human nervous system is complex, containing a vast diversity of neuronal cell types and connections that are influenced by complex and incompletely understood environmental and genetic factors (Tang *et al*., 2001). As mentioned earlier, L1s must retrotranspose in germ cells or during early embryogenesis to be evolutionarily successful, but the activity of these elements during this period and their effect on other somatic cells is not clear yet. A study on neuronal cells in transgenic mice reported that L1 constructs can retrotranspose, and that the activity of endogenous L1 promoter is strongly correlated with expression of the Sox2 gene (Muotri *et al.,* 2005). Indeed, an *in vitro* study on transgenic mice has been demonstrated that the L1 promoter is repressed by the Sox2 gene in undifferentiated hippocampus neural cells (HCN cells) but in the early stages of HCN differentiation depletion in the level of Sox2 expression directly correlates with the L1 transcript (Muotri *et al.,* 2005). It is

speculated that some of the genomic changes necessary for the uniqueness of individuals within a population, as defined by their neural circuitry, might be driven, in part, by the activities of mobile elements (Muotri *et al.,* 2005). Further a recent study has demonstrated that neural progenitor cells isolated from human foetal brain and derived from human embryonic stem cells also support the retrotransposition of engineered L1s (Coufal *et al*., 2009). Moreover, a high level of endogenous L1 transcripts have been detected in the hippocampus and several regions of the human brain, but few L1 transcripts were detected in other somatic cells, such as heart and liver, from the same individuals (Coufal *et al*., 2009). These data suggest that *de novo* L1 retrotransposition events may occur in the human brain and can contribute to brain somatic mosaicism (Coufal *et al*., 2009; Singer *et al*., 2010). Finally it has been shown that the activity of L1 in human brain cells can vary due to environmental factors (Singer *et al*., 2010). Recent studies on neuronal progenitor cells derived from Rett syndrome (RTT) patients and human iPS cells have found that mutations in MeCP2 can influence the activity of L1 retrotransposition in human brain cells. Therefore, if MeCP2 regulates L1 retrotransposition in a tissue-specific manner in human neuronal cells, this could add to the plasticity of human neuronal cells (Muotri *et al.,* 2010). It is still not clear if L1 retrotransposons have any functional impact on neuronal cells and why neuronal cells might accommodate a high level of L1 retrotransposition, when compared to other somatic cells.

1.9.2 L1 retrotransposition in malignant derived cells

Several studies have suggested that L1 can become active in cancer cells. Many of these studies correlate genome-wide hypomethylation during cancer progression to L1 reactivation, due to an increase in ORF1p and ORF2p expression in several malignant derived cells (Belancio *et al*., 2010). For example Aleves *et al*. (1996) demonstrated the hypomethylation of the sequences flanking the 5′ ends of L1Hs elements in T-47D breast cancer cell line. Moreover, they compared the hypomethylated L1 loci in cancer cells and germline cells, revealing that different subset of L1Hs are hypomethylated in each of the cell types. This suggests that the subset of L1Hs, which become reactivated in malignant-derived cells, may not be a random sample.

A recent study using L1-specific high throughput sequencing of lung cancer DNA has reported nine *de novo* L1 retrotranspositions in 6 lung tumors (Iskow *et al*., 2010). In parallel with finding these *de novo* L1 insertions the tumors bearing them show genome-wide hypomethylation, which is consistent with previous speculations that epigenomic alterations can have an effect on L1 activity. However, it is not clear whether L1 activation during cancer progression is only a consequence of genome alteration during cancer cell growth or whether L1 has an active role in driving tumorigenesis.

1.9.3 L1 retrotransposition in the human germline

As mentioned previously, for *de novo* L1 insertions to be evolutionary successful they must occur in the germline or during early embryogenesis before germline differentiation (Ergun *et al*., 2004). To date, most of the discovered disease-causing insertions are thought to be germline in origin as deleterious embryonic mutations are likely to be lost during development (Freeman *et al*., 2011). Discovery of a *de novo* LRE3 element insertion in exon four of the CYBB gene of a chronic granulomatous disease (CGD) by Brouha *et al*. (2002) has suggested that the L1 insertion into the CYBB gene is most likely to be germline in origin and occurred during prophase of maternal meiosis I. This and other cases of L1 disease-causing insertions, suggest that L1 retrotransposition can occur early in female oogenesis and embryonic development. Although these findings suggest that L1s must actively retrotranspose in the female germline, direct study of the female germline is very limited due to the difficulty in obtaining oocytes (Freeman *et al*., 2011). Based on the studies of L1 disease-causing insertions there is no direct evidence of *de novo* L1 retrotransposition in the male germline, but sperm provide an accessible resource for screening for *bona fide de novo* L1 insertions in the germline (Freeman *et al*., 2011). The sperm nucleus is a highly compact structure, and studies in mice have demonstrated that basic DNA associated proteins called protamines are important for post-meiotic chromatin condensation (Lee *et al*., 1995). Protamines are histone H1-derived, sperm-specific histone variants which associate with sperm DNA, thus permitting tight chromatin packaging (Lewis *et al*., 2004; Wouters-Tyrou *et al*., 1998). The dense packaging of DNA in sperm renders it transcriptionally inactive and so unlikely to be a good

substrate for the L1 endonuclease, meaning it is very unlikely that L1 retrotransposition will occur in mature spermatozoa. From this logic retrotransposition must occur in the early stages of spermatogenesis rather than the later stages. However, in oocytes there is no evidence of such tight chromatin packaging, and so they may be a preferential substrate for L1 retrotransposition. Indeed, immunohistochemical analysis has demonstrated the co-expression of ORF1 and ORF2 together in pre-spermatogonia, spermatocytes and immature spermatids but this is not detected in spermatogonia, suggesting L1 activity before and after meiosis (Ergun *et al*., 2004). This observation suggests that L1 is able to be transcribed and translated and possibly able to retrotranspose during spermatogenesis. Besides the extensive studies on L1 retrotransposition in the male germline there is no evidence of *de novo* L1 insertions in these cells and so it is not clear at which stage of human spermatogenesis L1s can become retrotranspositionally active (Freeman *et al*., 2011).

1.9.4 L1 retrotransposition in early human embryogenesis

Previously L1 retrotransposition was thought to occur predominantly in the germline (Ostertag and Kazazian, 2001; Bourchis and Bestor, 2004). However, recent studies on transgenic mice have demonstrated that L1 retrotransposition in the germline is quite uncommon, and the bulk of engineered L1 retrotransposition occurs in early embryogenesis with only a fraction of these insertions partitioning into the germline and being transmitted to the progeny (Kano *et al*., 2009). Indeed, except in one reported case (Brouha *et al*., 2002), where L1 retrotransposition is more likely to have occurred during maternal meiosis I, the rest of the known disease causing *de novo* L1 insertions could have occurred in early human embryogenesis (Kano *et al*., 2009). In support of this possibility Garcia-Perez *et al.* (2007) showed that endogenous L1 elements are expressed during human embryogenesis. A study on isolated ribonucleoprotein particles (RNPs) from undifferentiated human embryonic stem cell lines revealed the presence of ORF1P and L1 mRNA, and subsequent L1 RT-PCR showed that RNAs belonged to both active and old (and largely inactive) L1 subfamilies (Garcia-Perez *et al.,* 2007). To investigate whether or not human embryonic stem cells (hESC) can support exogenous L1 retrotransposition, Garcia-Perez *et al.* (2007) set up a tissue culture retrotransposition assay in which they

transfected undifferentiated hESC cells with RC-L1s, driven by either the activity of their endogenous 5´promoter, or a cytomegalovirus immediate early promoter. Using the culture cell retrotransposition technique, Garcia Perez *et al*. (2007) found that human embryonic stem cells express endogenous L1 elements and can accommodate exogenous L1 retrotransposition *in vitro*.

In addition, Van den Hurk *et al*. (2007) studied a case of Choroideremia, which was caused by a L1 insertion. This case of Choroideremia, an X-linked progressive eye disease, is caused by the insertion of a full length L1 into the CHM gene. This L1CHM gene has two 3´ transductions, and its transposition path is from a precursor L1 on either Chromosome 10p15 or 18p11 that transposed to chromosome 6p21 and then to the CHM gene on Chromosome Xq21 (Van den Hurk *et al.,* 2007). Using a PCR-based assay, the mutant CHM allele containing the L1 CHM insertion was amplified from the patient's family using markers within the CHM gene, and this showed the presence of an L1 CHM insertion in the mother. The results indicated that the mother was a somatic mosaic for the L1 insertion, and since the patient's mother showed both somatic and germline mosaicism for the L1 insertion into the CHM gene, the L1 retrotransposition event must have occurred during early embryogenesis, prior to germline segregation from the somatic lineages.

Based on this evidence and the failure to find *de novo* L1 insertions in male germline cells (Freeman *et al*., 2011), it seems likely that *de novo* L1 retrotransposition is more frequent in early human embryogenesis.

1.10 Project overview

From the preceding literature review it can be seen that L1 retrotransposition has had a profound effect on genome evolution. To date, the direct detection of *de novo* L1 insertions has only been demonstrated in human somatic lung cancer tissue (Iskow *et al*., 2010) and besides extensive research in this area using different display techniques and genome sequences, finding a *de novo* L1 retrotransposition event in the human germline and embryonic cells has proven challenging (Freeman *et al*., 2011). Consequently, very little is known about the dynamics of L1 retrotransposition in the germline and early embryogenesis. Some of the reasons explaining this low rate of success in finding new L1 retrotranspositions could be that there are no human germline cell cultures to study, the fact that this phenomena occurs across the genome leaving no specific locations to focus on, and finally because the rate of L1 retrotransposition (based on recent publications) has been estimated to be even lower than previously contemplated. A recent estimate, 1 insertion in 400 individual (Freeman *et al.,* 2011), is very different from earlier estimates of 1 in 33 (Brouha *et al*., 2003).

The purpose of the current project is to investigate L1 retrotransposition activity in early human embryogenesis. To do this only a subset of full-length L1 elements, which are more likely to be retrotranspositionally active, were targeted for screening. To investigate the activity of L1 during human early embryogenesis a display assay was designed to isolate the potentially retrotransposition competent L1s in different embryonal samples. Using this technique allows us to address, in principle, the precise stage of human development at which L1s retrotranspose and can also allow us to estimate the rate of L1 retrotransposition in early human embryogenesis.

1.11 Experimental approaches

Initial attempts to analyse *de novo* L1 insertions were based on a display method for screening genomic DNA to identify rare L1 insertions with low frequency, which was developed by Badge *et al*. (2003). Amplification typing of L1 active subfamilies (ATLAS) is suppression PCR display technique, which allows a selective amplification of the active L1 subfamily (L1Ta) amongst a high copy number of older elements, which are likely to be fixed in the genome. This display technique directly displays polymorphic L1Ta insertions, which subsequently can be recovered and subjected to cloning and sequencing (Badge *et al*., 2003). The products of the ATLAS technique are variable in length and generally they are short PCR fragments < 1kb. Each PCR product contains part of the L1 structure (< 200 bp) and its junction with genomic flanking DNA, which can be from a few nucleotides up to a few hundred long, depending on the restriction enzyme site in the genomic flanking DNA. Therefore, the initial aim of this project was to adapt the display method to allow the detection of *de novo* insertions in human embryonal cell lines.

ATLAS can detect low frequency insertions by selectively analysing active (L1Ta subfamily) elements. However, depending on the frequency of the restriction enzyme used for library preparation this display technique makes available only a fraction of the genome Therefore, a fraction of L1 insertions are not accessible for analysis with this technique. Moreover, in the case of low frequency insertions or a single molecule event, the L1 insertion may fail to amplify in the PCR and novel junction fragments generated from *de novo* insertions cannot be validated as the molecule from which the fragment originated is destroyed during amplification. *De novo* insertions can be validated indirectly by comparing the structure of the junction fragments with known L1 insertions, and by genotyping other samples such as sperm and blood from the same individual to show the absence of the insertion prior to spermatogenesis, and to exclude germline mosaicism. However, the major difficulty in this technique is to find very low frequency or single molecule events across the whole genome. This thesis outlines the development of a method designed to recover junction fragment from single molecule *de novo* insertions using different combinations of suppression PCR, display techniques and source DNA.

1.11.1 Investigation of endogenous L1 retrotransposition using single cell derived clonal cell population

The cellular milieu in hESC is known to be amenable to endogenous L1 expression and exogenous L1 retrotransposition (Garcia-Perez *et al*., 2007). Based on this principle, human embryonic stem cells (hESCs) can be used as a model of active L1 retrotransposition during early human development. Extended culturing of hESCs will provide opportunities for endogenous L1 retrotransposition with each cell division event, and could thus lead to the accumulation of insertions, and the formation of a mosaic population of cells.

In the single cell clone technique we assessed L1 insertional mosaicisim by firstly performing a serial dilution to isolate single cells from a mosaic population and used these cells to generate separate clonal lines (Fig.1.8). Individual clonal lines represent samples of the insertion diversity in the progenitor population. Clonal line amplicons are then compared to DNA from the progenitor population with the ATLAS display technique (see section 1.8.2). There are two possibilities for the origin of clonal line amplicons (insertions). The insertion may be a pre-existing insertion, which is detectable in the progenitor DNA. However, if the amplicon is present in the clonal line but absent in the progenitor line, the insertion might be a *de novo* insertion or a very rare mosaic insertion (Fig.1.9).

De novo insertion events arise by mutation during the outgrowth of the isolated single cell. A sensitive PCR genotyping assay can be used to distinguish between these possibilities, and can also help to estimate the frequency of pre-existing mosaicism as well as that of *de novo* insertion. We proposed to apply this cloning technique to hESC lines, but prior to the experiment we applied the technique to several human tumour-derived cell lines, which have an embryonic characteristics such as the NTera2D (pluripotent embryonal carcinoma) and PA1 (ovarian carcinoma cells) cell lines.

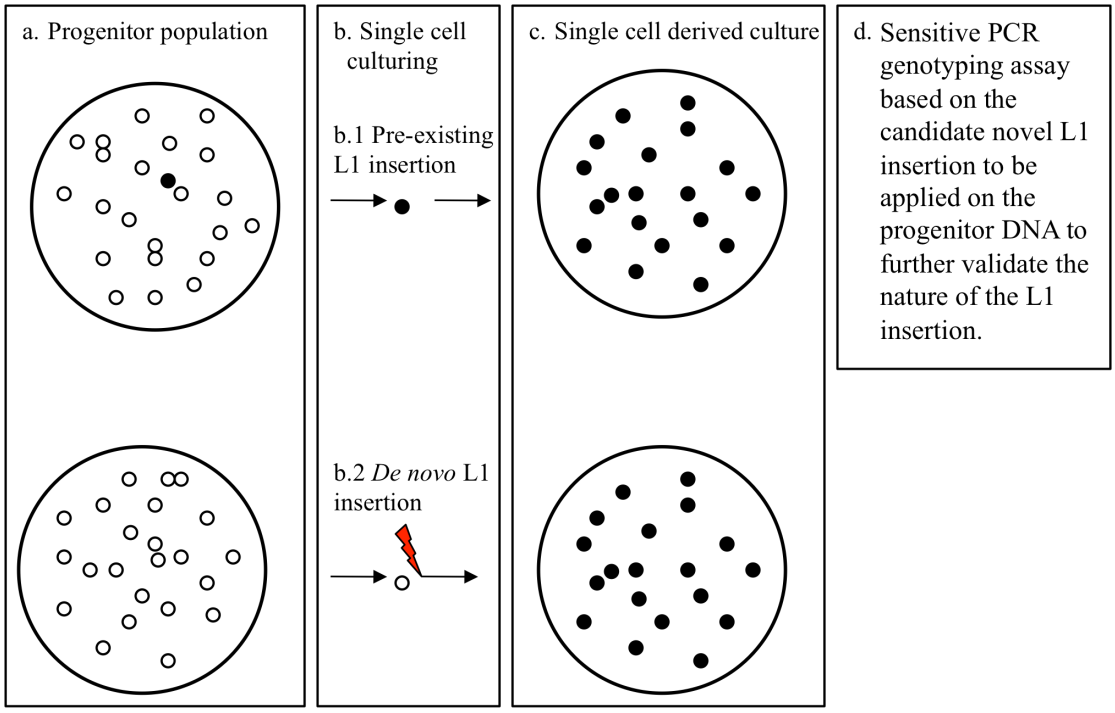

Figure 1.6 Proposed single-cell derived clonal population analysis of endogenous L1 retrotransposition activity. **a.** Progenitor population, **b.** Single cell extracted from the population and further cultured, b.1 a novel L1 insertion which was present in the progenitor population may have been selected, b.2 a single cell without a pre-existing insertion was further cultured and during the cell replication an endogenous *de novo* L1 insertion occurred. **c.** Single cell derived clonal population, **d.** To further validate the nature of the insertion, a PCR assay will used to amplify the novel L1 locus in the progenitor population. Black dot represents cells containing novel insertions.

1.11.2 ATLAS-based methylation sensitive differential digest to analyse the methylation status of young L1's promoter

As discussed earlier in this chapter, it is apparent that during early human development the genome undergoes numerous epigenetic alterations. Epigenetic alterations in early human embryogenesis are mainly required to maintain genome stability and direct chromatin remodelling (Magdinier *et al*., 2002; Suzuki and Bird, 2008). DNA methylation instability during embryogenesis is one characteristic of epigenetic regulation in early human development. L1 promoters are CpG-rich regions and have several binding sites for different transcription factors but L1 promoters are heavily methylated in normal somatic cells where they are not expressed. L1 promoters can become demethylated through a genome wide demathylation wave during early human embryogenesis. Thus, it can be speculated

that active L1s are likely to have a relaxed methylation status at their promoters during these events, so that they can be transcribed. This property can then be used as a method for screening for *de novo* L1s. Methylation of the CpG dinucleotides of the L1 promoter in several human embryonal cells have been studied in this thesis to verify the epigenetic dynamics of L1 in human embryogenesis and potentially to use it as a technique to target active L1 retrotransposons. To analyse the methylation status of L1 promoters, a combination of genome-wide and locus-specific methylation display techniques were used.

For the genome-wide methylation analysis the principle of the ATLAS technique was modified. Briefly, a genomic DNA library was constructed and following this the linkered fragments were subjected to a methylation sensitive differential digest (explained in more detail in the methods section 2.4). The resulting display pattern reflects the methylation status of CpGs present in the first +100 bp of the L1 promoter, including the four transcriptionally important binding sites of CpG dinucleotides. One of the major problems with this technique is the presence of CpG dinucleotides in the genomic flank. This technique may not always display the L1 CpG status and some of the fragments might represent the methylation status of CpGs upstream of the L1 in the genomic flank. However, data from the methylation status of the CpGs in the neighbouring genomic DNA can reveal a clearer picture of the epigenetic regulation occurring at the flanking genomic DNA surrounding the L1. Further validation can be applied to the recovered bands from this display system, such as cloning and sequencing to verify whether the corresponding methylation status reflects the L1 promoter or the junction genomic DNA. Further validation of the corresponding methylation status can assessed by designing a locus-specific bisulphite PCR-based assay.

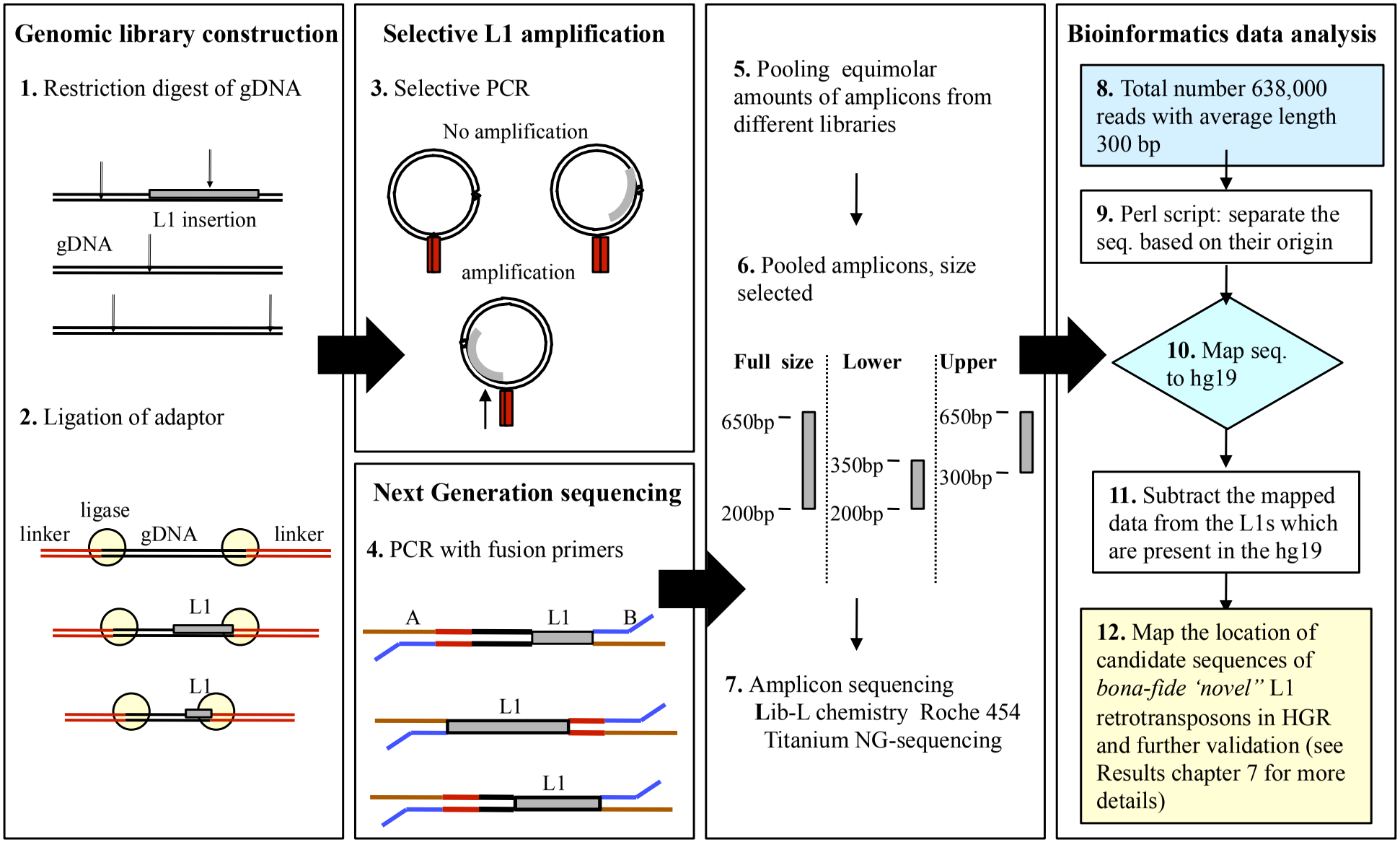

Figure 1.6 schematic diagram of combined ATLAS and NGS techniques to isolate *bona fide de novo* L1 insertion from human WGA-embryo. The ATLAS gDNA library construction is adapted from Badge *et al*., 2003.

To compare the methylation status of active L1s during human embryogenesis, the above genome-wide and locus-specific techniques were applied to hESCs as a model for human embryogenesis, and also to NTera2D1 (teratocarcinoma cells), where L1s are more likely to be hypomethylated, and germline (sperm DNA) or blood DNA from healthy volunteers, where L1s are more likely to be hypermethylated. These were also used as positive and negative controls, respectively.

1.11.3 Identification of human specific L1 mediated retrotransposition by NGS

As mentioned in the literature review of this chapter, the rate of *de novo* L1 insertion has been estimated to be very low due to the low activity of these elements in humans. Therefore, finding a *de novo* L1 insertion with a very low frequency or a single molecule event amongst the whole genome by using the current display techniques has proved to be challenging.

This thesis outlines the development of a method to screen for full length active L1s at single molecule resolution with a high sequence coverage. In chapter three, a higher resolution of ATLAS was introduced, which was able to detect rare insertions (single molecule events). However, due to the nature of the low sequence coverage of the rare events it is not known how many DNA molecules are required to be screened with this method to achieve characterisation of *de novo* insertions. Therefore, to improve the coverage of the single molecule events we have combined the display system with next generation sequencing. In this way detection of *de novo* L1 retrotransposons should be easier to achieve in a single experiment.

As mentioned earlier, one of the problems with using PCR-based display techniques such as ATLAS is that the single insertion molecule can be lost making it impossible to validate a *de novo* insertion. To solve this problem we applied whole genome amplification on the genomic DNA prior to using the ATLAS display system. This allows the preservation of copies of a single molecule insertion, and makes it possible to validate the *de novo* L1 retrotransposons by independent genotyping of the source genomic DNA. This display technique in principle is similar to the ATLAS technique. We have designed a new library construction procedure based on a more common restriction enzyme site than the one used in the original ATLAS. The advantages of

this restriction enzyme are that it produces a less sequence-biased library with higher genome coverage of nearly 80% (estimated by *in silico* genomic digestion). Following the selective amplification of the L1Ta subfamily, all the PCR products from each DNA sources were tagged by multiplex identifiers (MIDs, detailed in Chapter 6), and samples from different DNA sources were pooled for high throughput sequencing.

To investigate the activity of L1 during early human embryogenesis, the DNA of three human embryos including their total DNA and a sample of individual blastocysts, were screened using this novel technique (summarised in figure 1.6). To find *de novo* L1s, the L1 distribution in human embryos was compared with the germ line (sperm) and somatic tissue (blood) of a healthy adult individual. As is explained in more detail in Chapter six, we were able to recover and assign different amplicons to their DNA source using the barcoding system for high through put sequencing. Also, we were able to explore the intra-individual distribution of L1s in different cells such as sperm and blood from one individual. More importantly, we were able to sequence single molecule events, *i.e.* heterozygous insertions from single blastomeres, demonstrating the technical feasibility of our approach. We were able to propose a list of candidate *de novo* insertions purely based on sequence data analysis.

Chapter 2

Materials and Methods

2.1 Materials

2.1.1 Chemical reagents and laboratory equipment

All chemicals were supplied by one of the following suppliers: AB gene (Epsom, UK), Amersham Biosciences (Little Chalfont, UK), Applied Biosystems (Warrington, UK), Bio-Rad (Hemel Hempstead, UK), Boehringer, Cecil Instruments (Cambridge, UK), Clare Chemical research (Delores, USA), Clontech (Palo Alto, USA), Eppendorf Scientific (Hamburg, Germany), Fisher Scientific (Loughborough, UK), Fisons (Beverly, USA), Flowgen (Ashby-de-la-Zouch, UK), FMC Bioproducts (Rockland, USA), Hybaid (Teddington, UK), Invirogen (Paisley, UK), MJ Research (Waltham, USA), New England Biolabs (Hitchen, UK), Nagle Nune International (Hereford, UK), New Brunswick Scientific Co. (New Jersey, USA), Perkin Elmer (Cambridge, UK), Qiagen LTD (Crawley, UK), Serva, Sigma Aldrich (Pool, UK), Star lab (Milton Keynes, UK), Syngene Thermo Shandon (Pittsburg, USA), USB (Staufen, Germany), UVP Life Sciences (Cambridge, UK) and Zymo research (Cambridge, UK), KAPA Biosystems (Woburn, USA).

2.1.2 Enzymes

The restriction enzymes *EcoR*I, *Acc*I, *Mse*I, *Nla*III and *Vsp*I were supplied by New England Biolabs. *Taq*I, *Msp*I, *Tas*I and *Hpa*II were supplied by Fermentas (York, UK), Optikinase was supplied by USB (Staufen, Germany). *Taq* and *Pfu* DNA polymerases by KAPA Biosystems, (Woburn MA, USA).

2.1.3 Molecular weight markers

50 bp, 100 bp and 1 kb molecular weight markers were supplied by New England Biolabs, and λ DNA digested with *Hin*dIII was supplied by ABgene.

2.1.4 Standard solutions

Southern blot solutions (denaturing and neutralising), 20 × Sodium Chloride Sodium-Citrate (SSC) buffer and 10 × Tris-borate/EDTA (TBE) electrophoresis buffer, were made as described by Sambrook (Sambrook and Russell, 2001) and were supplied by the Media Kitchen at the Department of Genetics, University of Leicester.

2.1.5 Oligonucleotides

DNA oligonucleotides were synthesised by Sigma-Aldrich Company (Poole, UK).

2.1.6 Cell lines and embryo WGA samples

HeLa cell line; H1 was provide by Prof. Andrew Fry (University of Leicester, UK, Leicester); H2 was provided by Prof. Fred Gage (Salk Institute, La Jolla, USA, CA); H3 was provided by Dr. Raj Patel (University of Leicester, UK, Leicester); H4 was provided by Prof. John V. Moran (University of Michigan Medical School, US, Ann Arbor); H5 (Hep2) was provided by Dr. Simon Kilvington (University of Leicester,

UK). Placenta DNA was provided by Dr. Raymond Dalgleish (University of Leicester, UK). Human Embryonic Stem Cells (H1 and H9 and their clonal lines) were provided by Prof. John V. Moran (University of Michigan Medical School, US, Ann Arbor). Human embryonic stem cell clonal lines of the H9 progenitor (hESC1-hESC-20) and Whole Genome Amplified human blastocyst DNA (Embryo 3, 4 and 6) was provided by Dr. Jose Garcia-Perez (University of Granada, Spanish Stem Cell Bank, Spain, Granada). SW480 and SW620 were provided by Dr. Cristina Tufarelli (University of Nottingham, UK). NTera2D1 cell lines and HeLa cells were provide by Dr. Nicola J. Royle (University of Leicester, UK). CEPH family Lymphoblast genomic DNA and Zimbabwean genomic DNA were provided by Prof. Sir Alec J. Jeffreys (University of Leicester, UK).

2.1.7 Web services and software were used for data analysis

UCSC: The human genome browser at UCSC (Kent *et al.,* 2002), Repeat masker: http://repeatmasker.org (A.F.A. Smit, R. Hubley & P. Green RepeatMasker), BaseLine (Hasting and Badge, unpublished), dbRIP: http://dbrip.brocku.ca/citations.html (Wang *et al.,* 2006), primer3: Primer3 on the WWW for general users and for biologist programmers. (Rozen and Skaletsky, 2000), MethPrimer: http://www.urogene.org/methprimer/index1.html (Li and Dahiya, 2002), Galaxy: http://main.g2.bx.psu.edu/ (Blankenberg *et al*., 2010), Galview http://www.jalview.org/download.html (Waterhouse et al., 2009), Muscle http://www.ebi.ac.uk/Tools/msa/muscle/ (Edgar R.C., 2004), ImageJ software version 1.44 (available from the department of Genetics, University of Leicester), Image quant TL version 8.0

2.2 Methods

2.2.1 Tissue culture

Tissue culture was performed in a class II laminar flow hood in a designated tissue culture area. Cells were grown in 5% CO2 with at ~ 80% humidity at 37°C in an incubator (Sanyo). NTera2D1 cells were grown in Dulbecco's Modified Eagle Medium (D-MEM) supplemented with Glutamax-I (Gibo, Paisely, UK) and 10% foetal calf serum.

2.2.1.1 Dilution cloning of cultured human cells

Cloning experiments were performed as follow: cells were grown in 25cm^2 flasks until they reached 70% confluence. Trypsin (1ml per 20cm flask, with a concentration of 0.25% Trypsin, Gibco) was added to the flask for 5 minutes in order to detach the cells and to form a single cell suspension. The trypsinisation was stopped after minutes by adding D-MEM cell culture medium supplemented with foetal calf serum. In all cloning experiments, cells were centrifuged the pellet re-suspended in 10 ml of D-MEM media containing: 20% of Gold serum (PAA laboratories, Austria). The cells were counted and serial dilutions were made to give a concentration of ~8 cells / ml. Overall, 100 µl of the cell suspension was pipetted into each well of a 96-well plate and approximately 8 plates were seeded in this way for deriving NTera2D1 clonal cell lines.

The plates were observed periodically in order to determine the wells that contained colonies derived from single cells. Once the cells covered almost the whole base of the well, the medium was removed and the cells were washed using 100 µl of 1× PBS (137 mM sodium chloride, 2.7 mM potassium chloride, 4.3 mM disodium phosphate, 1mM potassium dihydrogen phosphate; MP biomedicals, Germany). The PBS was removed and 100 µl of trypsin was added per well until cells formed a single cell suspension. The cells were transferred to 25 cm^2 flask containing 8.5 ml of media and grown until they were harvested for DNA extraction.

2.2.1.2 DNA extraction from tissue culture cells

Cells were detached by trypsinisation and added to medium containing 10% foetal calf serum to stop the trypsin degrading the cells. The cells were centrifuged and the pellet re-suspended and washed twice in PBS. The final cell pellet was re-suspended in 250 µl (for 1 x 10^6 cells) of 1 x SSC (15 mM sodium citrate, 150 mM sodium chloride) buffer. Cells were lysed by adding 250 µl lysis buffer (100 mM Tris HCl pH 7.5, 100 mM NaCl, 10 mM EDTA, 1% Sarkosyl). RNase (final concentration 10 mg/ml) was added to the mixture for 20 minutes in order to degrade RNA. Proteins were digested by adding proteinase K to a final concentration of 20 mg/ml and incubated in a water bath at 55°C for 5 to 6 hrs. DNA was extracted from the solution with an equal volume of phenol: chloroform: isoamyl alchol (25:24:1) by gently mixing to form an emulsion. The organic and aqueous phases were separated by centrifugation at 13000 rpm (Eppendorf 5415 centrifuge) for 6 min at room temperature using phase-lock gel tubes (Eppendorf). The aqueous phase was removed and the DNA was precipitated using 1.0 volume of 2M sodium acetate (pH 5.6) and 2.5 volumes of 100% ethanol. The DNA was re-suspended in 100 to 500 µl of 1× TE (10 mM Tris, 1mM EDTA). The concentration of the DNA was estimated using a UV spectrophotometer (Eppendorf biophotometer).

2.2.2.1 Isolation of blastomeres from the inner cell mass of human embryos for deep sequencing

All the procedures for isolating blastomeres are performed by Dr. Jose Garcia-Perez and Dr. Jose Luis Cortes (embryologist) at University of Granada, Spanish Stem Cell Bank, Spain, Granada and the protocol provided through personal communication with Dr. Jose Garcia-Perez. The procedure approved by the local authorities and the Spanish national embryo steering committee.

Cryopreserved human embryos were donated to this study upon informed consent by couples that had already undergone an IVF cycle. All extractions were carried out in a GMP room, and the embryologist wore a lab suit aimed to prevent any cross-contamination. Cryopreserved human embryos were thawed using thawing specific

media (Vitrolife, Sweden), in the preimplantational stage (Day +1 - Day +6 after fecundation). In experiments aimed to isolate single blastomeres from human embryos, embryos were cultured until the 6-8 cells stage (Day +3 - Day +4 after fecundation) using G-1 v.5 and G-2 v.5 media (Vitrolife, Sweden). On the other hand, some embryos were cultured to the blastocyst stage after thawing (Day +5 – Day +6) using G-2 v.5 media (Vitrolife, Sweden). In the blastocyst stage, the trophoblast and the inner cell mass are easily distinguished.

The biopsy of the blastomeres was conducted using an inverted microscope (Olympus IX-71, Japan), using a micromanipulation system incorporated into the microscope (Eppendorf, Germany) and a laser drill (Octax, Germany). The embryo was held with the left micromanipulator using a holding capillary. Next, a single pulse with the laser drill was directed at the *zone pellucida* of the embryo. Following this, a biopsy capillary was attached with the right micromanipulator to isolate single blastomeres. Each blastomere was then introduced into an Eppendorf microcentrifuge tube (Low-bind) for further analyses.

In an alternative method, the human embryos were treated with Tyrode's Acid (Irvine, USA), to completely dissolve the *zone pellucida*. Then, single blastomeres were deposited in eppendorf tubes (Low-bind). In all samples, the rest of the human embryo was collected in a separate eppendorf tube (mass or bulk control).

2.2.2.2 Whole genome amplification (WGA) from human blastomeres

Whole genome amplification performed within 48 hours after blastomere extraction. The method described by Spits *et al*. (2006) was used without any additional modifications. Negative controls of all the reagents, which were used during the isolation of blastomeres were included (water, culture media, washing buffer, etc). Following the protocol first the gDNA of independent blastomeres was amplified. In the next step, samples with a DNA concentration of higher than 3ng/ul and whose negative controls gave a reading lower than 1ng/ul were selected (A Nanodrop ND-100 spectrophotometer was used for DNA quantification). The selected samples were split into three Eppendorf microcentrifuge tubes (low-bind), and an additional round

of WGA was performed. Finally, all three independent WGA gDNAs for each blastomere were pooled together.

2.2.3 Agarose gel electrophoresis

All PCR products were fractionated on low electroendosmosis (LE) agarose (Cambrex) gels. Fragments sized at <1 kb were fractionated on 2% weight-per-volume (w/v) agarose gels, 1 kb to 2 kb fragments on 1% w/v gels and >2 kb fragments on 0.8 % w/v gels. All gels were made using, and run in, 0.5 × Tris-Borate EDTA (TBE) (recipe in appendix II) buffer containing 0.5 μg/ml Ethidium Bromide (EtBr). The voltage at which samples were run and the running time varied depending on the samples loaded. For example, samples with a size of <1 kb were generally run at 4 V/cm on 2% w/v agarose gels for two hours. Amplicons of >1 kb were run at a lower voltage (~100 V) for longer (3-5 hours) on 0.8% w/v agarose gels. Ethidium bromide was used as a stain for nucleic acids to visualise the DNA band under UV light transillumination at 260 nm. Samples were loaded onto gels with 1 X Tris-Borate EDTA (TBE) loading buffer by adding the appropriate volume of the 6X concentrate. Fragment size was confirmed by comparing the samples against an appropriate DNA molecular weight marker; for example a 1 kb ladder or a 100 bp ladder, or *Hin*dIII-digested λ DNA. Electrophoresis tanks were manufactured by the university of Leicester workshop, and power packs supplied by Bio-Rad. To visualise DNA bands on the gel either a hand held UV wand (Chemical-vue UVM-57, UVP Life Science) or UV transilluminator, or a "Dark Reader" visible light hand lamp or transilluminator (Clare Chemical Research) was used. The "Dark Reader" instruments use visible light between 400-500 nm to avoid DNA degradation by UV.

Gel photographs were taken using a dark room cabinet with a CCD camera (Syngene), using the GeneSnap image acquisition software (Syngene).

2.2.4 Standard DNA digestion using restriction enzymes

Genomic DNA was quantified using UV spectrophotometry. The digestion mix consisted of 1× reaction buffer (supplied with the enzyme), and 5 units of enzyme per µg of DNA. The desired amount of DNA was added to the digestion mix and the whole reaction was brought to a final volume of 20 µl by adding ultra-pure water. Each aliquot was then incubated for 1 hour at 37°C. Heat sensitive enzymes were inactivated by incubation at 65°C for 20 minutes. The heat-insensitive enzymes were inactivated by an addition of equal volumes of stop solution (appendix I) into each reaction. A small amount of each sample (5 µl) of the digested DNA was then loaded onto an 8% w/v agarose gel. The over-digest control was set up by elongating the digestion time to overnight.

2.2.5 PCR-based methods

The primers sequences used in this project are listed in appendix II.

2.2.5.1 Primer dilutions

All the primers and their derived dilutions were kept in PCR-clean condition *i.e.* they were only opened in a class II laminar flow hood. PCR primers were diluted in UV-irradiated 5 mM Tris HCl (pH 7.5) to a stock concentration of 100 µM. For the purpose of PCR, 10 µM dilutions of the primers were used. Any primers that had been opened outside the class II laminar flow hood, at any point, were marked as 'non-PCR clean' and never returned to the flow hood.

2.2.5.2 Standard PCR conditions

All the PCR mixes were assembled in a class II laminar flow hood to ensure that all the reagents were kept PCR-clean. The standard PCR mix was generally made up to a final volume of 10 or 20 µl per reaction. A 20 µl PCR reaction contained 0.4 U/µl of *Taq* DNA polymerase, 1× PCR buffer (11.1 × buffer), 0.5 µM of each primer and 20

ng of human genomic DNA (gDNA). PCRs were carried out in an MJ tetrad PCR machine PCT250, with the following cycles: variant cycles of a denaturing cycle at 96°C for 1 min: 96°C for 30 sec; annealing temperature (TM) °C for 30 sec and 72 °C for 1min (This was varied relative to the size of the amplicon, allowing 1 minute extension per each kilo base pair), and a final extension step at 72°C. The annealing temperatures and the relevant PCR cycles for each primers combination are summarised in table 2.1.

Primer combination	**TM°C**	**Number of cycle**	**Type**
RBX1/CM5DP1T3	64	30	Sequencing
RBX4/RB5PA2	64	30	PCR
RB5PA2/RR0812A	64	30	PCR
RB5PA2/RR8633A	58	30	PCR
RB5PA2/RB696A	61	30	PCR
RB5PA2/VM164A	63	30	PCR
RR0812A/RR0812B	64	30	PCR
RB696A/RB696B	61	30	PCR
RR8633A/RR8633B	56	30	PCR
CM5DP1	68.8	60	PCR
VM164A/B/RB164K2	63	30	PCR
RVECPA1/RVS5A2	62	25	PCR
RVECPA2/RVS5B2	62	25	PCR
454-A/454-B	64	30	PCR
MID(*X)-A/MID(X)-B	75	25	PCR

Table 2.1 the primer combinations and their PCR annealing temperature and number of cycles for an optimised PCR. *X: are the MID tagged primers from Chapter six.

2.2.6 Southern blotting and Dot-blotting

2.2.6.1 Southern blotting protocol

The Southern blot technique is a means of transferring DNA bands from agarose gels onto a Magna Nylon transfer membrane (GRI). This technique allows the screening of DNA products by hybridising them with radio-labelled-probes. The Southern blot technique starts with pre-blotting. In order to distinguish the gel orientation after blotting, a corner of the membrane was cut off, and nicks were made in the edge of the gel corresponding to the molecular weight marker bands. This allowed hybridised band sizes to be estimated after autoradiography. The gel surface was made as flat as possible by slicing off any raised wells with a scalpel blade. The gel was then transferred onto a tray and covered with denaturing solution (0.5 M NaOH, 1M NaCl) for 20 min with gentle shaking. The denaturing solution was then replaced with neutralising solution (0.5 M Tris-HCl (pH 7.5), 3 M NaCl) for 20 min with gentle shaking. Once the 20 minutes had elapsed, the gel was rinsed in distilled water and placed onto a rig. The rig consists of a tray half-filled with 20× SSC, and a glass plate spanning the length of the tray with a wick made of Whatman 3MM paper covering the plate and reaching down into the 20× SSC. Ensuring that the wick was soaked with 20× SSC, the gel was inverted and laid flat on top of the wick. Air bubbles were smoothed out with a glass pipette. Saran wrap was laid around the gel to ensure that all the transfer solution ran through the gel. A sheet of Nylon transfer membrane (which had been soaked in 3× SSC prior to use) was laid on top of the gel and any air bubbles smoothed out. Two sheets of 3MM paper (pre-soaked in 3× SSC) were laid over the membrane and the air bubbles were removed. A stack of paper towels were placed on top of the 3MM paper followed by a glass plate and a weight to hold the paper towels in place. Complete DNA transfer from the gel to the membrane took at least 3 hours. After the blotting stage the Nylon membrane was washed briefly in 3× SSC and dried in a 3MM paper envelope at 80°C for 10 minutes. The blot was then transferred to a CL 1000 ultraviolet cross-linker (UVP) and exposed to 70,000 μJ/cm² of UV (254 nm). The blots were then stored in 3MM paper envelopes, wrapped in and Saran wrap (Dowe) in the dark at 4°C.

2.2.6.2 Dotblotting

The dotblot method is an alternative method to Southern blotting for the screening of PCR product with radiolabelled probes. For the dotblot procedure, two 13 × 9.5 cm sheets of Whatman 3 MM paper and one 12 × 8.5 sheet of nylon transfer membrane were cut and soaked in 3× SSC. The two pieces of 3 MM paper were placed on the bottom of a 96-well Hybri-blot manifold (BRL life technology INC, Gaithersburgh, USA), and the nylon transfer membrane placed on top of the 3 MM paper. The manifold was then assembled and the holding screws tightened evenly. For every kilobase of amplicon length, 30 to 100 ng of PCR product were required for blotting. 0.25× the volume of dotblot loading mix (30% v/v glycerol, 0.5× TBE, 0.025% bromophenol blue) was added to each sample along with 5 volumes of denaturing mix. Samples were mixed by pipetting several times. A vacuum was applied to the dotblot manifold using a dry vacuum pump/compressor (Fisherbrand), and the samples were loaded using a multichannel pipette. Once the samples had been pulled through, the wells were washed and neutralised with 150 µl 2× SSC. Once the SSC has been pulled through, the vacuum was released and the dotblot manifold disassembled. Dotblots prepared in this way were dried at 80°C in 3MM paper for 10 minutes, and then transferred to a CL 1000 ultraviolet cross-linker (UVP) and exposed to 70,000 µJ/cm² of UV. Blots were stored in the dark at room temperature (short term) or 4°C (long term), covered with a 3MM paper envelope and wrapped in Saran wrap (Dowe).

2.2.6.3 5′ end labelling of oligonucleotide probes with γ-^{32}P-ATP, using Optikinase (USB)

Oligonucleotide hybridisation is fairly fast and highly specific, and it was generally used to confirm the identity of bands on a Southern-blotted gel. This method has a limited sensitivity and it is not good for detecting extremely small amounts of DNA. 2 pmol of oligonucleotides were added to 1 µl of 10 × Optikinase buffer, 5 units of Optikinase buffer, and 0.5 µl γ-^{32}P-ATP and made up to a final volume of 10 µl with water. The probe was incubated for an hour at 37°C and then used for hybridisation immediately after. The modified Church solution (Sambrook and Russell, 2001) was used for hybridisation at 45-50°C. The blotted and cross-linked membranes were pre-

hybridised in Church buffer for at least 30 minutes. Hybridisation was first performed in 10 ml of Church buffer for 45 min and followed by 2 washes at 45°C. Following this pre-hybridisation step, the Church solution was discarded and replaced with 10 ml of fresh Church solution at 45°C for an hour. The blots were then washed in twice with 20 minutes between washed. In the washing time interval the old Church solution was replaced with 10 ml fresh Church solution. The blot was then washed three times in 3× SSC at room temperature and wrapped in Saran wrap. Blots were then either exposed to a piece of X-ray film or a phosphoimager screen (Amersham Biosciences).

2.2.6.4 Preparation of PCR amplified probes for random prime labelling with α-^{32} P-dCTP

For this labelling technique, probes of 1-5 kb in length were amplified using standard PCR conditions, and the samples were run out on agarose gels. If a single band of the correct size was present, the PCR amplicons were purified using a PCR purification kit. For fragments <4 kb, the MiniElute PCR clean up kit and the standard protocol (Qiagen) was used.

2.2.6.5 Random prime labelling

In this method, purified PCR generated probes were denatured by heating in order to produce single stranded DNA suitable for polymerase extension primed by random hexamer oligonucleotides.

2.2.6.6 α-^{32} P-dCTP probe labelling using Rediprime II random prime labelling system

The DNA to be labelled was first diluted to achieve a mass of 2.0-25 ng in 10mM TrisHCl, pH 8.0, 1mM EDTA at the final volume of 45 µl. The probe was then denatured at 100°C for 5 min in a water bath, followed by snap cooling on ice for a further 5 min. The tube was then pulse centrifuged and the contents transferred to a Rediprime II reaction tube (Amersham Bioscience). The reaction was thoroughly

mixed by pipetting until the reaction tube marker dye was evenly dispersed. 2.5 µl of α-^{32} P-dCTP was added and the tube sealed and tapped gently to mix the contents, prior to being incubated at 37°C for 60 min. The reaction mix was then boiled for 5 min and snap cooled for a further 5 min to denature the probe. The labelled probe was then used immediately for hybridisation.

2.2.6.7 Hybridisation of random prime labelled probes

Blots were pre-hybridised for at least 1 hour in 20 ml of pre-heated Church buffer (at 65°C). Hybridisation was then carried out in 20 ml Church buffer overnight at 65°C. Following overnight hybridisation two washes were carried out. The first wash was carried out in 0.2 x SSC and 0.5 % SDS for 15 min at 65°C and repeated once. The second wash was carried out once in 0.1× SSC and 0.1% SDS for 30 minutes. The blots were then washed 3 times in 3x SSC and wrapped in Saran wrap. Hybridised blots were then either exposed to an X-ray film or phosphoimager screen (Amersham Biocsiences).

2.2.6.8 Blot stripping

Southern blots and dotblots were stripped of probes by washing several times in boiling 1% SDS and monitored with a Geiger counter until the radiation levels dropped below 5 counts per second. The blots were then washed with distilled water, and then 3x SSC. Stripped blots were either re-hybridised, or stored at 4°C in the dark, wrapped in Saran wrap.

2.2.7 Cloning and sequencing PCR amplified samples

All media and containers were autoclaved prior to use. Standard sterile techniques were used throughout to prevent contamination of the bacterial samples.

2.2.7.1 Ligation of PCR generated DNA fragments into a plasmid vector

PCR amplified DNA fragments were cut from either agarose gels or polyacrylamide gels. Bands that were cut from polyacrylamide gels were initially soaked in PCR-clean 5MT and then frozen at -20°C overnight. 5 µl of the eluted amplicons extracted from the agarose gels were purified using a QIAquick gel extraction kit (Qiagen). Following DNA extraction, the reactions outlined in Table 2.2 were set up using the CloneJET-PCR cloning kit (Fermentas). Before the ligation, all the sticky end PCR products were made blunt ended, following the protocol in Table 2.2.

Component	Volume /µl
PJET1.2/blunt cloning vector (50 ng/µl)	1
T4 DNA ligase (5 U/µl)	0.5
5x ligase buffer (Promega)	1.5
Total volume	10

Table 2.2 the reaction used to blunt the sticky end of PCR products, from the sticky-end cloning protocol of the cloneJET-PCR cloning kit (Fermentas).

The blunting mixture reaction was incubated at 70°C for 5 min and chilled briefly on ice. Following the blunting reaction, ligation was carried out at room temperature (22°C) for 5-30 min, and the remaining steps were performed using the CloneJET-PCR cloning kit (Fermentas), as described in the manufacturer's instructions.

Component	Volume (µll)
2x reaction buffer	10
PCR products	1-2
Water (nuclease-free)	Up to 17
DNA blunting enzyme	1
Total reaction volume	18

Table 2.3 the reactions used to ligate purified blunt ended PCR amplified fragments into the pJET1.2 vector prior to transformation to *E.coli.* The ratio of the vector to the PCR products was 3:1.

2.2.7.2 Transformation of competent E. coli prepared with Transformed Aid bacterial transformation kit (#K2710)

200 µl competent cell aliquots were defrosted on ice, and 5 µl of the ligation mix (provided with the kit) added. Samples were mixed by flicking and left on ice for 30 minutes before being heat shocked in a water bath at 42°C for 30 sec, and then snap-cooled on ice for 2 min. 900 µl of SOC media (containing 4mg/ml glucose), pre-warmed to 37 °C, was added to each sample and incubated at 37°C for 45 min. LB agar plates containing 0.2 mg/ml Ampicillin (LB Amp plates) were pre-warmed to room temperature, and 40 µl 50 mg/ml Xgal and 20 µl 24 mg/ml IPTG spread onto each plate. 200 µl of each transformation was plated onto the LB Amp plates and incubated overnight at 37 °C. 200 µl of the positive control was plated onto LB Amp plates, and 200 µl of a transformation with no added plasmid DNA was spread onto both an LB plate and an LB Amp plate (these were viability and Ampicillin sensitivity controls). Colonies containing inserts were identified due to their white colour, resulting from disruption of the β-galactosidase gene.

2.2.7.3 Preparation of plasmid DNA

5 ml aliquots of LB liquid media containing 100 µg/ml Ampicillin were inoculated in triplicate with a single positive colony from each plate. The cultures were incubated overnight in a shaker at 37°C. Plasmid DNA was recovered from the cultures using a QIAprep spin mini prep kit (Qiagen). After extraction, 5 µl of plasmid DNA was digested with *Mse*I. Digested and undigested plasmid DNA both were run side by side on a 1% LE agarose gel to reveal a restriction fragment fingerprint. As the cultures were grown in triplicate when any set showed more than one restriction fragment fingerprint the one clone with each distinct restriction fragment fingerprint was sequenced.

2.2.7.4 Sequencing using Big Dye Version Terminator v3.1 (Applied Biosciences)

The sequencing reaction protocol is outlined in Table 2.4. The reactions were placed in an MJ tetrad PCR machine, and the following cycle performed: 96°C for 10 sec; 50°C for 5 sec; 60°C for 4 min. This was repeated for 25 more cycles.

Component	Quantity
Big Dye V3.1	1 µl
5×Big Dye Buffer	1.5 µl
3.3 µM primer	1 µl
DNA	20-13 ng / kb
H_2O	To volume of 10 µl

Table 2.4 The required components for DNA sequencing and their quantities (one sequencing reaction).

2.2.7.5 Clean-up of sequencing reaction

A master mix of 10µl distilled water and 2 µl 2.2% SDS per sequencing reaction was prepared. 12 µl of this mix was added to each reaction, mixed by pipetting, and then heated to 98°C for 5 min followed by 10 min at 25°C in an MJ Tetrad PCR machine.

Unincorporated Big Dye nucleotides were removed using PERFORMA DTR Gel filtration Cartridges (Edge BioSystems & VIt Bio Ltd). Finally, the samples were submitted to the Protein and Nucleic Acid Chemistry Laboratory (PNACL) at the University Leicester, UK for capillary electrophoresis and data collection.

2.2.8 Amplification typing of L1 active subfamilies (ATLAS)

All steps of the ATLAS procedure were performed in a class II laminar flow hood that had been decontaminated by UV exposure for at least 30 min prior to use. All reagents were PCR clean (i.e. opened only in the hood and used only for PCR), and 18 mΩ water, from a distiller (Pure1TE select, ONDEO, UK) or supplied by Sigma-Aldrich, was used.

2.2.8.1 Library construction

600 ng of gDNA was digested with 15 units of *Ase*I (NEB) for 3 hours or overnight at 37°C in a water bath, in a final reaction volume of 30 µl (the *Nla*III restriction enzyme based library construction mentioned in appendix II). Several controls were included in the digestion step: DNA negative (H_2O); digestion enzyme negative (replaced with 50% glycerol); and a DNA/reaction positive. After the digestion the reactions were heated at 65°C for 20 min to inactivate the digestion enzyme (digested DNA was stored at -80°C in a PCR clean condition). 20 µl of each linker primer [RBMSL2:(5′GTGGCGGCCAGTATTCGTAGGAGGGCGCGTAGCATAGAACG-3′) and RBD3 (5′-TACGTTCTATGCTAC-3′)] were annealed together by incubating them at 65°C for 10 min and then allowing them to cool down at room temperature over 30-69 minutes. In the standard ATLAS protocol (Badge *et al.,* 2003), 100 ng of the digested DNA was ligated to a 40-fold molar excess of the annealed suppression linker. The amount of linker was calculated by assuming the enzyme completely digested the genome into 'X' number of fragments with two ligatable ends, and 3 pg of DNA represents one haploid genome equivalent. X varies with respect to the enzyme's cutting frequency (but all calculations are necessarily approximate). For *Mse*I, 2.7 µl of annealed linker was used for each ligation in a final volume of 20 µl.

100 ng of genomic DNA were ligated with the annealed linker overnight at 15°C, in a final reaction volume of 20 µl. The linker negative (H_20), and enzyme negative (50% glycerol) and two reaction positives were included as ligation stage controls. A 20 µl ligation reaction final volume consisted of, 5 µl of 100 ng digested DNA, and 2.7 µl (10 µl) annealed linker, 1.34 µl (4 Weiss units) T4 ligase (Promega), 2 µl 10×ligase buffer and 8.96 µl H_2O. To inactivate the ligation and also to remove the 'dummy' RBD3 oligonucleotides, reactions were incubated at 70°C for 10 min. The excess of linkers and short DNA fragments (<100 bp) were removed by using the Qiaquick PCR purification system (Qiagen) according to the manufacturer's instructions. The purified linkered DNA was then eluted in PCR clean 5MT at a final volume of 30 µl. These eluates were aliquoted into three sets of 10 µl and stored at -80 °C.

2.2.8.2 Primary PCR

Standard ATLAS primary PCR was carried out as follows: A 10 µl final reaction volume consisting of 9 µl PCR mix and 1 µl of constructed library DNA was assembled. The PCR mix was made of 1× buffer (11.1× PCR buffer) (recipe in appendix I), 0.5 µl of 50 µM RBX4 (Linker-specific primer) and RB5PA2 (L1 internal primer), appendix III, 0.4 units of *Taq* (0.08 µl). PCR was then performed under the following conditions: an initial denaturing step at 96°C for 30 s; followed by 32 cycles of 96°C for 30 s, 64°C for 30 s, 72°C for 1 min; and followed by an extension step at 72°C for 10 min.

2.2.8.3 Display PCR using γ^{33}P-ATP-labelled oligonucleotides

50 µM of the display primer (*e.g. CM5DP1*) was labelled with γ^{33}P to give 1.5 pM of Labelled-primer per 10 µl reaction. For example, for 100 reactions, 10 µl γ^{33}P-ATP, 4 µl Optikinase buffer (10x), 2 µl Optikinase (10 units), 3 µl 50 µM Cm5DPI, and 40 µl distilled water were added and incubated at 37°C water bath for an hour. Display PCR was carried out in a final volume of 10 µl (9 µl PCR mix and 1 µl of primary PCR product), with a final concentration 1x buffer C (recipe in appendix I), 0.5 Units *Taq* and 1.5pM of γ^{33}P-ATP-labelled display primer. Finally PCR (three steps) was carried out in PCR thermo-cycler (Gene Amp. PCR system 9600) under the following cycle conditions: denature at 96°C for 1 min, followed by 60 cycles of [96 °C for 30 s, 68.8°C 30 s, 72°C for 1 min], and a final extension at 72°C for 10 min.

2.2.8.4 Display Gel; Single 6% (w/v) gel

Gel mixes were prepared in advance and stored at 4°C. For 1× gel mix (100 ml), 50 g urea, 12 ml 50% long ranger gel solution (Lonza), 5 ml 20× Glycerol tolerant gel buffer (recipe in appendix I) and water were added to the final volume of 100 ml. The gel mixes were simultaneously degassed and filter sterilized using a 0.22 μm Millipore Express Plus vacuum filter and wrapped in foil keep out light. To polymerase the gel, 100 μl TEMED and 100 μl freshly made Ammonium Persulphate solution (25%) [50 mg in 200ul H_2O] were added and mixed thoroughly prior to pouring of the gel. The gels were left for a minimum 3 hours or overnight to completely set. Gels were then run at a constant power of 75-100W using a Biorad power pack (Power PAC basic model 300). Long runs took 4-6 hours to complete, while short runs lasted 3-4 hours.

2.2.9 Protocol for SMD/SP-ATLAS (Single molecule dilution / Small pool)

All the library construction section is similar to the general ATLAS protocol (section 2.2.8). After the gDNA library construction, assuming the 80% recovery of ligated DNA and the concentration of ligated DNA in eluate has calculated. Serial dilutions in a maximum of 1 in 10 dilution steps, were performed to yield a range of concentrations to sub-haploid genome equivalent levels- <3pg/ul. SMD diluents were freshly prepared (SMC 1M TrisHCl pH 7.5 and 2mg/ml sonicated *E. coli* genomic DNA) to give a solution that is 5mM TrisHCl pH 7.5 and 5ng/ul *E. coli* genomic DNA. The carrier DNA allows accurate dilution without adsorption. Primary PCR Planed to include at least 5-20 primary positive pools and at least 10 primary negative pools. Appropriate dilution for poisson analysis was determined with small pool numbers (5 replicates). Once poisson range determined higher replication numbers was used (10-20).

2.2.9.1 Primary PCR

Standard ATLAS primary PCR was carried out as follows: A 10 μl final reaction volume consisting of 9 μl PCR mix and 1 μl input DNA dilution (250ng-2.5pg/ul). The PCR mix was made of 1× buffer (11× PCR buffer, recipe in appendix I), 0.5 μl of 50 uM RVECPA1 (Linker-specific primer) and RV5SA2 (L1 internal primer) at the

final concentration of 250nM per primer (appendix III), 0.4 units of *Taq* (0.08 µl) and 0.1ul of100ng/ul sonicated *E.coli* genomic DNA (final concentration=1ng/ul). Replicate pools were prepared by adding DNA in bulk to PCR mix, mixed them and aliquoted to individual tubes. Standard primary cycle conditions (Tetrad) Step1: 96^{o}C 1min, Step 2: 96 ^{o}C 30s, Step 3: 62^{o}C 2 min, Step 4: return to Step 2 19-21 more times, Step 5: 62^{o}C 10 minutes, Step 6: 4 ^{o}C forever.

2.2.9.2 Dilution step

Sufficient Single Molecule Dilution Diluent (SMDD= 5mM Tris HCl pH7.5, 5ng/ul sonicated *E.coli* genomic DNA) was prepared for all primary PCRs and controls (including secondary bench negatives). 98ul SMDD aliquoted in thin wall PCR tubes in PCR clean racks. The secondary PCR mix prepared and aliquoted at same time. 2ul of each primary was added into each 98ul of SMDD.

2.2.9.3 Secondary PCR

Secondary PCR mix was prepared for all reactions (+10-20%). Standard secondary PCR is 10ul, 9ul mix + 1ul diluted (2+98ul) primary PCR (final primary primer concentration ~0.5nM). Per reaction, 1ul of AJJ 11X, 0.5ul of 50uM RVECPA2, 0.5ul of 50uM RV5SB2 (final concentration= 250nM) 0.4 units of Taq (0.08ul) were used. Standard secondary cycle was the same as the primary: Cycle conditions: Step1: 96^{o}C 1min, Step 2: 96 ^{o}C 30s, Step 3: 62^{o}C 2 min, Step 4: return to Step 2 19-21 more times, Step 5: 62^{o}C 10 minutes, Step 6: 4 ^{o}C forever.

The followings steps including the radioactive labelled display PCR was same as the above protocol (section 2.2.8).

2.2.10 Low Complexity-ATLAS (LC-ATLAS)

600 ng of genomic DNA was digested to completion with 20 units of *Nla*III (NEB) in the manufacturer's recommended buffer, at 37°C for 3 hours. After incubation reactions were heated to 65°C for 20 minutes to inactivate the restriction enzyme. Prior to setting up the ligation reaction, linker oligonucleotides were freshly annealed by mixing equal volumes of 20 µM RBMSL3 and RBD4, heating to 65°C for 10 minutes, and then cooling to room temperature over 30-60 minutes.

100 ng of the digested DNA was ligated to a 40-fold molar excess of the annealed suppression linker (2.7µl of 10uM annealed linker for *Nla*III libraries) with 4 Weiss units T4 DNA ligase (Promega) in 1× Ligase Buffer (Invitrogen) overnight (~16 hrs) at 15°C, in a final volume of 20 µl. After ligation the reaction was heated to 70°C for 10 minutes to inactivate the ligase. Excess linkers and short DNA fragments (*i.e.* <100 bp) were removed with the Qiaquick PCR purification system (Qiagen), following the manufacturers protocol, but eluting the DNA in 30 µl 5 mM TrisHCl pH 7.5. In our hands the purification is ~80% efficient, resulting in a purified library containing approximately 2.7ng /µl of genomic DNA. Libraries are sensitive to freeze/thaw, and so were aliquotted and stored frozen at -20°C. Genomic DNA was amplified in 10 µl PCR reactions containing 1 × PCR buffer (11.1× buffer) 1.25 µM RBX4, 1.25 µM RB5PA2 and 0.4 units *Taq* DNA polymerase (ABgene). Reactions were cycled in a Tetrad 2 Thermal Cycler (MJ Research / Biorad, Hercules, CA) using the following conditions: 96°C -1 min; 30 × [96°C -30 s; 63°C -30 s; 72°C -1 min]; 72°C -10 min. and 1 µl diluted PCR reaction was added to 9 µl secondary PCR reactions containing 1 × PCR buffer (11.1 × buffer), 0.625 mM RRNY1T/RRNY1A, RRY1C and RRYG 0.625 mM RB5G, 0.4 *Taq* DNA polymerase (ABgene) Reactions were cycled in a Tetrad 2 Thermal Cycler (MJ Research) using the following conditions: 96°C -1 min; 30 × [96°C -30 s; 64°C -30 s; 72°C -1 min]; 72°C -10 min.

2.2.11 Transduction Specific-ATLAS (TS-ATLAS)

TS-ATLAS is a modification of ATLAS (Badge *et al*., 2003) and relies on the use of transduction-specific PCR primers to selectively amplify L1 loci containing transduced sequence from oligonucleotide-linkered genomic libraries. All oligonucleotides were HPLC-purified by the manufacturer (Sigma) and re-suspended

at 50 µM in 5 mM TrisHCl, pH 7.5. All pre-PCR reactions were set up in a Class II laminar flow hood (Walker) decontaminated by UV exposure for at least 30 min prior to use.

2.2.11.1 Library Construction and Amplification

The method described below is the 980-transduction lineage specific protocol. Variations required for 011 and 958 lineages are described in appendix II. 600 ng of genomic DNA was digested to completion with 20 units of *Nla*III (NEB) in the manufacturer's recommended buffer at 37°C for 3 hours. After incubation reactions were heated to 65°C for 20 minutes to inactivate the restriction enzyme. Prior to setting up the ligation reaction, linker oligonucleotides were freshly annealed by mixing equal volumes of 20 µM RBMSL3 and RBD4, heating to 65°C for 10 minutes, and then slowly cooling to room temperature.

100 ng of the digested DNA was ligated to a 40-fold molar excess of the annealed suppression linker (2.7µl of 10 µM annealed linker for *Nla*III libraries) with 4 Weiss units of T4 DNA ligase (Promega) in 1× Ligase Buffer (Invitrogen) overnight (~16 hrs) at 15°C, in a final volume of 20 µl. After ligation the reaction was heated to 70°C for 10 minutes to inactivate the ligase. Excess linkers and short DNA fragments (i.e., < 100 bp) were removed with the Qiaquick PCR purification system (Qiagen), following the manufacturers protocol, but eluting the DNA in 30 µl 5 mM TrisHCl, pH 7.5. In our hands the purification is ~80% efficient, resulting in a purified library containing approximately 2.7 ng/µl of genomic DNA. Libraries are sensitive to freeze/thaw, and so were aliquoted and stored frozen at -20°C.

To suppress amplification of the AC002980 L1 insertion, 10 µl of the ligation reaction was incubated with 10 units *Mun*I (Roche) for 3 hours at 37°C, in a final reaction volume of 20 µl. After digestion reactions were heated to 65°C for 20 minutes to inactivate the restriction enzyme, cooled on ice, and centrifuged briefly. 1µl of ligated and *Mun*I digested linkered genomic DNA was amplified in 10 µl PCR reactions containing 1 × PCR buffer (11.1× PCR), 1.25 µM RBX4, 1.25 µM RB980TD2 and 0.4 units *Taq* DNA polymerase (ABgene). Reactions were cycled in a Tetrad 2 Thermal Cycler (MJ Research / Biorad, Hercules, CA) using the following conditions: 96°C -1 min; 30 × [96°C -30 s; 63°C -30 s; 72°C -1 min]; 72°C -10 min.

Primary suppression PCR reactions were diluted 1:50 in Single Molecule Dilution Diluent (SMDD: 5 mM Tris HCl pH 7.5, 5 ng/µl sonicated *E.coli* genomic DNA) and 1 µl of the diluted PCR reaction was added to a 9 µl secondary PCR reaction containing 1× PCR buffer (11.1× buffer), 0.625 mM RBY1, 0.625 mM RB980TD3, 0.4 *Taq* DNA polymerase (ABgene) Reactions were cycled in a Tetrad 2 Thermal Cycler (MJ Research) using the following conditions: 96°C -1 min; 30[96°C -30 s; 64°C -30 s; 72°C -1 min]; 72°C -10 min.

2.2.11.2 Recovery and Analysis of TS-ATLAS Products

10 µl of secondary TS-ATLAS PCR products were fractionated on 2% Seakem LE (Cambrex) 0.5× TBE agarose gels against a 100 bp ladder (NEB) size marker and visualised by ethidium bromide (0.5 µg/ml) staining. Novel PCR products (*i.e.* amplicons not corresponding in size to the suppressed progenitor or known transduction loci) were excised from the gel and purified using the Qiagen Mini-elute system (Qiagen) following the manufacturers protocol, but eluting the DNA in 10 µl of 5 mM TrisHCl, pH 7.5.

Purified PCR products were directly sequenced with ABI BigDye Ver. 3.1 Ready Reaction, using 3.3uM RBY1 as the sequencing primer. Sequencing reactions were purified using Performa DTR spin columns (Edge BioSystems & Vlt Bio Ltd) and the sequencing data collected using an ABI 3730 capillary sequencer by the PNACL core DNA sequencing service (University of Leicester).

The sequences of the TS-ATLAS amplicons were imported into the CHROMAS sequence viewer and the L1 transduction flanking sequences mapped to the Human Genome Reference (HGR) Sequence (hg19) assembly using BLAT (University of California, Santa Cruz; (http://genome.ucsc.edu). The accession number of the genomic location was determined using the National Centre for Biotechnology Information (NCBI) BLASTN program (http://www.ncbi.nih.gov).

2.2.11.3 Primer design for novel L1insertions related to AC002980, LRE3 and RP

Upon identification of the insertion points of novel insertions, the flanking DNA sequence was downloaded from the UCSC genome database (hg19) and repeats were masked (http://www.repeatmasker.org/), prior to the design of PCR primers using Primer 3 (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi). Primers were placed such that the 3′ flanking primer (downstream of the L1 poly A tail) lay 3′ of the restriction site to which the library linker was ligated, to enable independent verification of the ligation point. Where flanking sequence was highly repetitive, primers were positioned across the junction of repeats to yield locus specific primers.

2.2.11.4 Verification of novel L1s containing 3′transductions

The presence of L1 sequences upstream of 3′ transductions related to AC002980, LRE3 or RP was verified by PCR amplification using a locus specific 3′ flanking primer and the primer RP3PA2, which is specific for the 3′ end of human specific L1s. PCR products were separated on 2% agarose gels, purified using the Qiagen Mini-elute system (Qiagen), cloned using the pGEM-T easy kit (Promega) and transformed into ultra competent DH5α *E. coli* cells. Plasmid DNA was recovered using a QIAprep Spin mini prep kit (Qiagen). 20-30 ng / kb of plasmid DNA was sequenced using the Big Dye Terminator v3.1 as above with 3.3 µM sequencing primer (M13F or M13R).

2.2.11.5 Presence/absence polymorphism

The dimorphism of L1 insertions related to AC002980, LRE3 or RP was determined using a three primer-two PCR assay, which amplified the 3′ end of a L1 and its flanking DNA as described previously (Sheen *et al*. 2000; Badge *et al*. 2003). A panel of unrelated Northern European CEPH genomic DNAs (n=129) was used to estimate allele frequency.

2.2.12 Methylation Sensitive-ATLAS (MS-ATLAS)

2.2.12.1 Library construction

All the steps of library construction were the same as for the standard ATLAS technique explained in 2.2.8.

2.2.12.2 Differential Methylation Digest

22 ng of DNA (~8 µl of the eluate) was digested with either 10 units of *Hpa*II (Fermentas), *Msp*I (Fermentas) or a mock digest containing 1µl 50% glycerol in 1× NEB at 37°C overnight and a final reaction volume of 10 µl. The enzyme was heat inactivated at 65°C for 20 minutes. Following the differential digest, all the steps (including the suppression and amplification PCR, labelling and linear amplification PCR) were as described in the ATLAS technique described in section 2.2.8.

2.2.13 Sodium bisulphite treatment

2.2.13.1 Bisulphite Conversion

500 ng of each DNA sample was converted using the EZ DNA Methylation Gold kit (Zymo Research). The kit uses a coupled heat-denaturation/conversion step to convert unmethylated cytosines into uracil. DNA is then purified and desulphonated using column chromatography. The manufacturer's protocol was followed with this exception: as the volume of the DNA input was higher than 20 µl, CT conversion reagent was prepared using 850 µl of H_2O (Sigma) and each reaction contained 25 µl of DNA and 125 µl of CT conversion reagent. Incubation conditions were 98^{o}C for 10 min followed by 64^{o}C for 2.5 hrs.

2.2.13.2 Combine Bisulphite Restriction Analysis of L1 (COBRA L1)

Bisulphite treated DNA (2.2.13.1) was subjected to 35 cycles of PCR with two primers, RRCOBRAF: 5′-CCGTAAGGGGTTAGGGAGTTTTT-3′ and RRCOBRAR: 5′-RTAAAACCCTCCRAACCAAATATAAA-3′, using an annealing temperature of 50°C. The amplicons were digested in a 10 µl reaction volume with 2 U of *Taq*I or 8 U of *Tas*I in 1 X *Taq*I buffer (MBI Fermentas) at 65°C overnight and were then electrophoresed on 12% non-denaturing polyacrylamide gels. The intensities of the DNA fragments were measured using a PhosphorImager and the Image Quant software (Amersham Bioscience). LINE-1 methylation levels were calculated as a percentage of the intensity of *Taq*I divided by the sum of *Taq*I- and *Tas*I-positive amplicons. The LINE-1 amplicon size is 160 bp. Methylated amplicons (*Taq*I positive) yield two 80 bp DNA fragments, whereas unmethylated amplicons (*Tas*I positive), yield 63 and 97 bp fragments.

2.2.13.3 Direct Bisµlphite Sequence analysis for different L1 loci

Bisulphite specific primers were designed using the Methyl Primer Express Software v1.0 (Applied Biosystems). All reactions were prepared in a total volume of 20 µl. 1 µl of the converted genomic DNA (~50 ng) was used as template and each reaction contained 1× PCR buffer (11.1× buffer). The primer final concentration was 0.625 µM, and each reaction contained 0.02 units/µl *Taq* polymerase (ABGene). All the locus specific bisulphite PCR reactions are summarised in Table 2.5.

L1 Locus	Strand	Primers	PCR condition
AC005885	Sense	VMB885E1/ VMB885F1	96°C for 30 sec followed by 40 cycles [96°C for 30 sec, 56°C for 30 sec, 72°C for 1 min] followed by 72°C for 10 min
AC114499	Sense	VMB499D1/ VMB499E1	96°C for 30 sec followed by 40 cycles (96°C for 30 sec, 56°C for 30 sec, 65°C for 2 min) followed by 65°C for 10 min.
AC069384	Sense	RBB384E/ RBB384F	96°C for 30 sec followed by 40 cycles (96°C for 30 sec, 55°C for 30 sec, 65°C for 2 min) followed by 65°C for 10 min.
AC002980	Sense	RR980F/ RR980R RRL980A/ RRL980B	Reaction cycle: 96˚C 1min, followed by 40 cycles of 96˚C 30s, 63.7˚C 30s, 68˚C 1 min, final cycle of 68˚C 10 min

Table 2.5 List of 4 L1 loci at which methylation was analysed using direct bisulphite methylation analysis. The primers and the PCR conditions for each locus are listed above. Primers sequences are presented in appendix II.

2.2.13.4 Sequencing

Sequencing reactions for the sodium bisulphite treated DNA were carried out using the protocol described earlier with the following modifications: each reaction included 1µl of Big Dye v3.1 Terminator Ready Reaction mix, instead of 4µl, plus 1.5µl of 5x Big Dye Terminator Buffer (Applied Biosystems).

2.2.14 High throughput L1 amplicon sequencing

All the sequencing preparation steps that involved library construction and PCR were performed in a class II laminar flow hood that had been decontaminated by UV exposure for at least 30 min prior to use. All reagents were PCR clean (*i.e.* opened only in the hood and used only for PCR) also H2O for the PCR was HPLC-purified by the manufacturer (Sigma).

2.2.14.1 Library construction for amplicon sequencing PCR

200 ng of WGA DNA was digested with *Nla*III (NEB) for 3 hours at 37°C in a water bath, in a final reaction volume of 20 µl and an enzyme concentration of 20U/ µl in 1× buffer (NEB). Several controls were included in the digestion step: DNA negative (H_2O); digestion enzyme negative (replaced with 50% glycerol); and a DNA / reaction positive. Following digestion the reactions were heated at 65°C for 20 min to inactivate the digestion enzyme (digested DNA was stored at -80°C in PCR clean conditions). An equal volume of each linker primer RRNBOT2: 5′-ACTGGTCTAGAGGGTTAGGTTCCTGCTACATCTCCAGCCTCATG-3′ and RRNDUP1: 5′-AGGCTGGAGATGTAGCAG-3′) 50 µmol were mixed. The mixed adapters were then denatured and annealed by heating to 65°C for 10 min and then cooling to room temperature at the rate of 1°C every 15 s. In the standard ATLAS protocol (Badge *et al.,* 2003), 100 ng of the digested DNA was ligated to a 40-fold molar excess of the annealed suppression linker. The amount of linker is calculated by assuming the enzyme completely digested the genome into 'X' number of fragments with two ligatable ends, and 3 pg of DNA represents one haploid genome equivalent. 'X' varies with respect to the enzyme's cut frequency (but all calculations are necessarily approximate). For *Nla*III, 2.7 µl of the 50 µmol annealed-linker was used for each ligation in a final volume of 20 µl. 100 ng of genomic DNA was ligated with the annealed linker overnight at 15°C, in a final reaction volume of 20 µl. The linker negative (H_2O), and enzyme negative (50% glycerol) and two reaction positives were included as ligation stage controls. A 20 µl ligation reaction final volume consisted of, 100 ng digested DNA, and 2.7 µl annealed linker, 1.34 µl (4 Weiss units) T4 ligase (Promega), 2 µl 10× ligase buffer and 8.96 µl H_2O. To further inactivate the ligation and also to remove the 'dummy' RRNDUP1 oligonucleotides, reactions were incubated at 70°C for 10 min. The excess of linkers and short DNA fragments (<100 bp) was removed using the Qiaquick PCR purification system (Qiagen) according to the manufacturer's instructions. The purified DNA was then eluted in PCR clean 5MT (5mM TrisHCl, pH 7.5) to a final volume of 30 µl. This was aliquoted into three sets of 10 µl and stored at -80°C.

2.2.14.2 Primary PCR

Standard primary PCR was carried out. A 15 µl final reaction volume, consisting of 13 µl PCR mix and 1 µl of constructed library DNA was made. The PCR mix was made out of 1 × PCR buffer (11.1× buffer), 0.5 µl of 50 µM RVECPA1 (L1-specific linker primer) and RV5SA2 (L1 internal primer), appendix III, 0.4 units of *Taq* DNA Polymerase (0.08 µl). PCR was then performed under the following conditions: an initial denaturing step at 96°C for 30s, followed by 25 cycles of 96°C for 30s, 62°C for 2 min, and then an extension step at 72°C for 10 min.

2.2.14.3 secondary PCR

Standard primary PCR was carried out. A 50 µl final reaction volume consisting of 45 µl PCR mix and 5 µl of constructed library DNA was made. The PCR mix was made out of 1 × PCR buffer (11.1× buffer), 0.125 µM fusion primer A (containing an L1-specific linker primer) and 0.125 µM fusion primer B (containing an L1 internal primer), and 0.4 units of *Taq* DNA Polymerase (0.08µl). All of the fusion primers are listed in table (Table 2.2.6). PCR was then performed under the following thermo cycling conditions: an initial denaturing step at 96°C for 30 s; followed by 25 cycles of 96°C for 30 s, 75°C for 2 min; and followed by an extension step at 72 °C for 10 min.

2.2.14.4 Fusion primers design

In order to be able to separate the DNA from different samples, we designed the fusion primers. Each pair of the fusion primers consisted of forward and reverse primers. All the forward fusion primers were constructed of the following: Roche Lib L primer A: (5′CGTATCGCCTCCCTCGCGCCATCAG3′), a 10 nucleotide MID (Multiplex identifier), and an L1-specific linker primer RVECPA2 5′CCTGCTACATCTCCAGCC3′. All the reverse fusion primers were constructed as follows: Roche Lib L primer B: 5′-CTATGCGCCTTGCCAGCCCGCTCAG-3′, a 10 nucleotide MID (Multiplex identifier) and L1-specific primer RV5SB2 5′- CTTCTGCGTCGCTCACGCT- 3′. A list of all the fusion primers is presented in the appendix III.

2.2.14.4 Pooling the PCR products and product size separation

To determine DNA concentration prior to the pooling of libraries, samples were analysed on a 2100 BioAnalyzer (Agilent) and Pico-Green Analyser (Invitrogen). Equimolar concentrations of each sample were pooled together. Three equal volumes (100 µl) of the pooled samples were loaded on a 2.5% agarose gel and run at 120 V for 2 hrs. The gel was then transferred onto a dark reader and was cut to divide the pooled products into three different size ranges: 200-450 bp, 350-600 bp, and 200-600bp. DNA was extracted from the gel using the Qiaquick gel extraction system (Qiagen) according to the manufacturer's instructions. The purified DNA was then eluted in PCR clean 5MT (5mM Tris-HCl, pH 7.5) to a final volume of 30 µl. This was aliquoted into three sets of 10 µl and stored at -80°C. Samples were sequenced on the 454 Roche GS-FLX sequencer using LibL Titanium chemistry, by the University of Leicester NUCLEUS genomics service.

2.2.14.5 Computational analysis

It was necessary to develop a novel computational pipeline since our technique does not correspond to whole-genome resequencing, or any other existing application of next-generation sequencing. Firstly, Perl scripts (appendix VI) were written to sort all the sequences into the 14 libraries according to their MID's. In the next step sequence reads were trimmed by removing 53 bp from the 5′ end and 35 bp from the 3′ end. Following trimming, the reads were mapped to the reference genome (hg19), using the LastZ tool on Galaxy (Goecks *et al*., 2010), and only matches with 95% or more identity were reported (all the steps and their parameters are presented in appendix V). In the next step we looked for all the intersections between sequence-reads and our L1-oligo-specific-data set (this dataset was made by mapping the L1-specific primer, RV5SB2, in hg19). In this way all those L1 sequences which were present in the reference genome were seperated from those which were not present in the reference. For those which were present in the reference genome, 10 sequences from each library were further analysed manually using the UCSC genome browser assembly hg19 (http://genome.ucsc.edu) and RepeatMasker (http://www.repeatmasker.org/), to make sure that all the sequence reads were corresponding L1 sequences which are present in the genome.

Any insertions, which did not map to the reference genome was further trimmed by removing the L1 sequences. Following the trimming the remaining sequence reads were mapped back to the hg19 reference sequence to locate their genomic location (empty site).

2.2.14.6 Site-specific PCR

The presence of non-reference insertions was verified via site-specific PCR. The 5′ ends and flanking regions of non-reference L1s were amplified using the L1-specific primer and a 5′ flanking region determined using the reference genome sequence. The "empty" site, that is, the allele that does not contain an L1 insertion, was also amplified from the genome using primers flanking the suspected site of insertion on the 5′ and 3′ ends.

Chapter 3

Activity of intact endogenous L1 retrotransposons in human embryonal cell lines

3.1 Introduction

L1s are important genome modifiers, altering mammalian genomes in many ways, both constructively and destructively (Kazazian, 2004). Our genome has accumulated several hundred thousand L1 copies over evolutionary time. Based on disease-causing insertions 17 human *de novo* L1 retrotransposition insertions have been characterised to date (Kazazian 2004). However, since in all cases there is no definitive evidence as to when many of these retrotransposition events occurred, some of these *de novo* L1 retrotranspositions may have occurred early in embryogenesis. Studies of such heritable disease-causing L1 insertions have shown that L1s are still accumulating in the contemporary human genome. Therefore, they must either retrotranspose in germ cells during gametogenesis or during early embryogenesis prior to germline partitioning, in order to be incorporated into germ cells. L1 somatic retrotransposition events that do not get incorporated into germ cells are not heritable and will not accumulate in the genome. Moreover, L1 RNA and proteins have been found predominantly in germ cells and infrequently in differentiated tissues (Branciforte and Martin 1994; Trelogan and Martin 1995). Therefore, the prevailing view has been that

the bulk of L1 retrotransposition occurs in germ cells. Studies of human L1 elements in transgenic mice have demonstrated direct germline retrotransposition occurred only when the L1 transgene was driven by a heterologous germline-specific promoter (Ostertag *et al*., 2002).

Despite the tremendous impact of L1 on the human genome, much of the process of L1 retrotransposition *in vivo* remains unexplored and it is unclear whether new insertions are produced in the germline as frequently as it is estimated from the prevalence of disease causing mutations, due to the strong bias in favour of X-linked insertions (X or Y). Also due to the lack of effective assays to capture very rare or single insertion molecules in bulk DNA, studying *de novo* insertions has been very difficult. Despite frequent attempts to find new L1 insertions and success in identifying somatic insertions in cancer cells, germline *de novo* L1 insertion have not yet been found except for the disease causing mutations discussed above. As a consequence, most of the transposon insertions that have been detected to date are common or fixed insertions that have been discovered by genome sequencing projects (Lander *et al*., 2001; Venter *et al*., 2001), and many younger and thus rarer insertions have been discovered by display techniques (Badge *et al*., 2003, Brouha *et al*., 2004, Wang *et al*., 2006, Kano *et al*., 2009, Ewing *et al.,* 2010). Although discoveries of polymorphic insertions are very informative towards our understanding of the biology and evolution of these endogenous elements, new, young insertions are much more interesting than common alleles, as they are less likely to have been removed by purifying selection and so will more faithfully represent the spectrum of mutations. Thus, the full extent of human germline mutagenesis by endogenous retrotranspons remains relatively uncharacterized.

A recent study by Iskow *et al* (2010) demonstrated that *de novo* somatic L1 insertions occur at detectable frequencies in the human lung cancer genome. Their data suggest that transposon-mediated mutagenesis could be extensive in human genomes of both germline and somatic cells, through the mobilisation of highly active (Brouha *et al.,* 2003) endogenous L1 retrotransposons. Among all of the insertions discovered, those in introns outnumber exonic insertions, and many of the polymorphic insertions have a minor allele frequency of less than 5%, which suggests that they are very young and arise from recent cell divisions (Iskow *et al*., 2010).

Moreover a recent study of active L1s using fosmid-based, paired-end DNA sequencing in five unrelated human genomes from geographically diverse populations, revealed 68 novel L1 insertions (*i.e.* that were absent from the reference human genome) (Beck *et al*, 2010). Interestingly, *in vitro* analysis using a retrotransposition assay (Moran *et al*., 1996) revealed that more than half of these novel L1 insertions were retrotranspositionally active. This data has suggests that active L1s are more abundant in the human population than previously appreciated, and therefore ongoing L1 retrotransposition is more actively contributing to the diversity of the human genome than previously suspected.

It is estimated that up to 5% of newborns may contain a *de novo* L1-mediated retrotransposition event (Garcia-Perez *et al*., 2007). However, little is known about the developmental timing or cell types that accommodate L1 retrotransposition in humans. The discovery of ongoing L1 retrotransposition in somatic human lung cancer revealed that this can occur in malignant cells, but ongoing L1 retrotransposition in human germ cells and embryonic cells has yet to be experimentally demonstrated.

In vitro studies using mouse models indicate that L1 expression and retrotransposition can occur in germ cells, during early development, and also in selected somatic tissues (Brouha *et al*., 2003, Mine *et al*., 2007). Moreover, *in vitro* retrotransposition assays have demonstrated L1 retrotransposition in a variety of human and rodent transformed cell lines (Ergun *et al.,* 2004, Garcia-Perez *et al.,* 2007), in rat neuronal transformed cells (Muotri *et al*., 2005), and at a relatively low level in primary human fibroblasts (Bruke *et al*., 1998, Brouha *et al*., 2003). However, a recent study by Freeman *et al.* (2011) showed that despite extensive screening of sperm DNA of a male donor for *de novo* L1 insertions, the experiment could not isolate any *bone fide de novo* L1 insertions. Based on this observation the rate of L1 retrotransposition in the male germline is less (>1 in 400 haploid genomes) than was previously estimated, (1 in 33) by Brouha *et al.,* (2003).

Although the low activity of L1s in the male germline is paradoxical with the fact of ongoing L1 retrotransposition, it does suggest that L1 retrotransposition could occur pre-mitotically, during the early stages of human embryogenesis. Also, it has been demonstrated that human embryonic stem cells can accommodate the

retrotransposition of engineered L1s *in vitro* (Garcia-Perez *et al*., 2007). These data suggest that L1 retrotransposition events may occur at early stages in human embryogenesis and that some individuals in the population may be genetically mosaic with respect to their L1 content (Van den Hurk *et al*., 2007).

Based on these findings concerning L1 retrotransposition during human embryogenesis, this chapter aimed to directly elucidate the activity of L1 retrotransposons during the early stages of human development. For this purpose, several different cell lines with embryonic properties (NTera2D1, PA1 and hESC) were chosen as potential models of human embryogenesis.

Embryonal carcinoma cells are pluripotent stem cells derived from teratocarcinoma tumours. Their biochemical and immunological characteristics resemble early embryonic cells (Fukuda *et al.,* 1985). As a result embryonal carcinoma cells (*e.g*. the Ntera2DI and PA1 cell lines) have been useful models for studying early embryonic development (Fukuda *et al*., 1985; Andrews *et al*., 1984). Moreover, it has been asserted that the NTera2D1 cell line supports endogenous L1 retrotransposition, based on the expression of the L1 ORF1 protein (Gilbert *et al*., 2005).

Additionally, the cellular milieu in hESCs is known to be amenable to the expression of endogenous L1 retrotransposons (Garcia-Perez *et al.,* 2007) and so they can be used as a model to investigate the activity of L1 retrotransposons in early human development.

Here a genome-wide approach has been used to try to study the activity of L1 elements by finding *de novo* L1 insertions segregating or fixed within clonal cellular populations. We have used the ATLAS technique developed by Badge *et al.* (2003) to specifically screen for *de novo* L1 insertions (details of library construction are explained in Chapter 2). We used the 5′-ATLAS technique to isolate full length L1 retrotransposons as these are the types of elements responsible for most ongoing activity: knowing the rate at which truncated and therefore "dead on arrival" insertions occur is only relevant to estimating mutational load, not rates of productive retrotransposition. It has been estimated that only 5% of genomic L1s are full length and the majority 95% are 5′ truncated, so in this study we are biasing towards a fraction of L1s. However it is clear that the proportion of full length elements in young, human specific families is much higher (Boissinot and Furano, 1999), (~30%)

indicating that the level of bias is less than the length distribution of fixed genomic elements would suggest. Cell culture based retrotransposition assays indicate that even reporter constructs carrying selectable markers generate full length insertions at a frequency of around 6% suggesting the true proportion of endogenous full length insertions lies somewhere between these bounds. As a result the 5′ sites of full-length L1 retrotransposons that carry sequence variants associated with active families (L1Ta1d) were studied in single cell derived clonal lines of NTera2D1, PA1 and hESC. The extended culturing of single cell clonal lines can provide an opportunity for endogenous L1 retrotransposition during each cell division and insertions can accumulate leading to the formation of a somatic mosaic population. In addition screening clonal isolates necessarily samples the standing variation (if any) within the progenitor population, without the need to achieve single molecule sensitivity.

De novo L1 insertions can be further distinguished from pre-existing insertions using PCR to screen the progenitor cell line for the insertion. This technique can lead to the identification of young insertions which have a low allele frequency (<0.05% in the population) as well as private insertions which have been generated so recently that they are found in only single individuals (Mills *et al*., 2007 and Badge *et al.,* 2003). These insertions have the potential to be used as individual-specific genetic markers (Rahbari *et al*., 2009).

Following the screening of embryonal clonal lines for new L1 insertions using the ATLAS technique (Badge *et al*., 2003), we have developed a modified version of the ATLAS technique that has a higher genome coverage with less display gel resolution complexity, which is more sensitive for identifying low allele frequency and perhaps younger L1 insertions within populations.

3.2 Results

3.2.1 Genome-wide comparative analysis of L1 activity in hESC vs. PA1 clonal cell lines

Clones from two independently derived human embryonic stem cell clonal lines (WA01 and WA09), were compared to clones of the ovarian carcinoma cell line PA1 (both cell lines were provided by Prof. John Moran, Dept. of Human Genetics, University of Michigan) for their L1 insertion variation. WA01 is the progenitor of the H1-H4 and H6 clonal cell lines and WA09 is the progenitor of the H7 and H9 clonal lines. Also, nine PA1 clonal lines were used for this experiment (P1-P9). Unfortunately the original progenitor cells for these clonal lines were not available. All of these clonal lines were expanded from single cells transfected with a L1 retrotransposition vector that conferred G418 resistance. These lines were generated during experiments to demonstrate L1 retrotransposition in hESC and PAI cells (Garcia-Perez *et al*., 2007).

Restricted genomic DNA libraries were constructed according to the standard ATLAS protocol (Badge *et al.,* 2003) using the *Mse*I restriction enzyme to digest the genomic DNA of each clonal cell line (methodology explained in section 2.2.3). As is demonstrated in Figure 3.1, various bands were generated over the length of the ATLAS display gel. Each band on the gel represents the 5′ end of an L1 with its flanking genomic DNA, terminated by a *Mse*I site. The control lanes 1-6 show that the elimination of one or two components (for example: restriction digest enzyme –ve, ligase –ve, gDNA –ve and others) cannot generate the same pattern as when all components are present. As can be seen in figure 3.1, some variations were observed between the different clonal lines. Some of the bands were found in both the clonal lines as well as their progenitor (blue boxes), whereas others were sporadic bands that only appeared in one of the clonal lines (red boxes). Some of the variations were observed in inter-clonal lines that were derived from the same progenitor, and some of the bands exhibited intra-clonal line variation between different cell lines. All the observed variations on the display gel were subjected to re-amplification, cloning and sequencing.

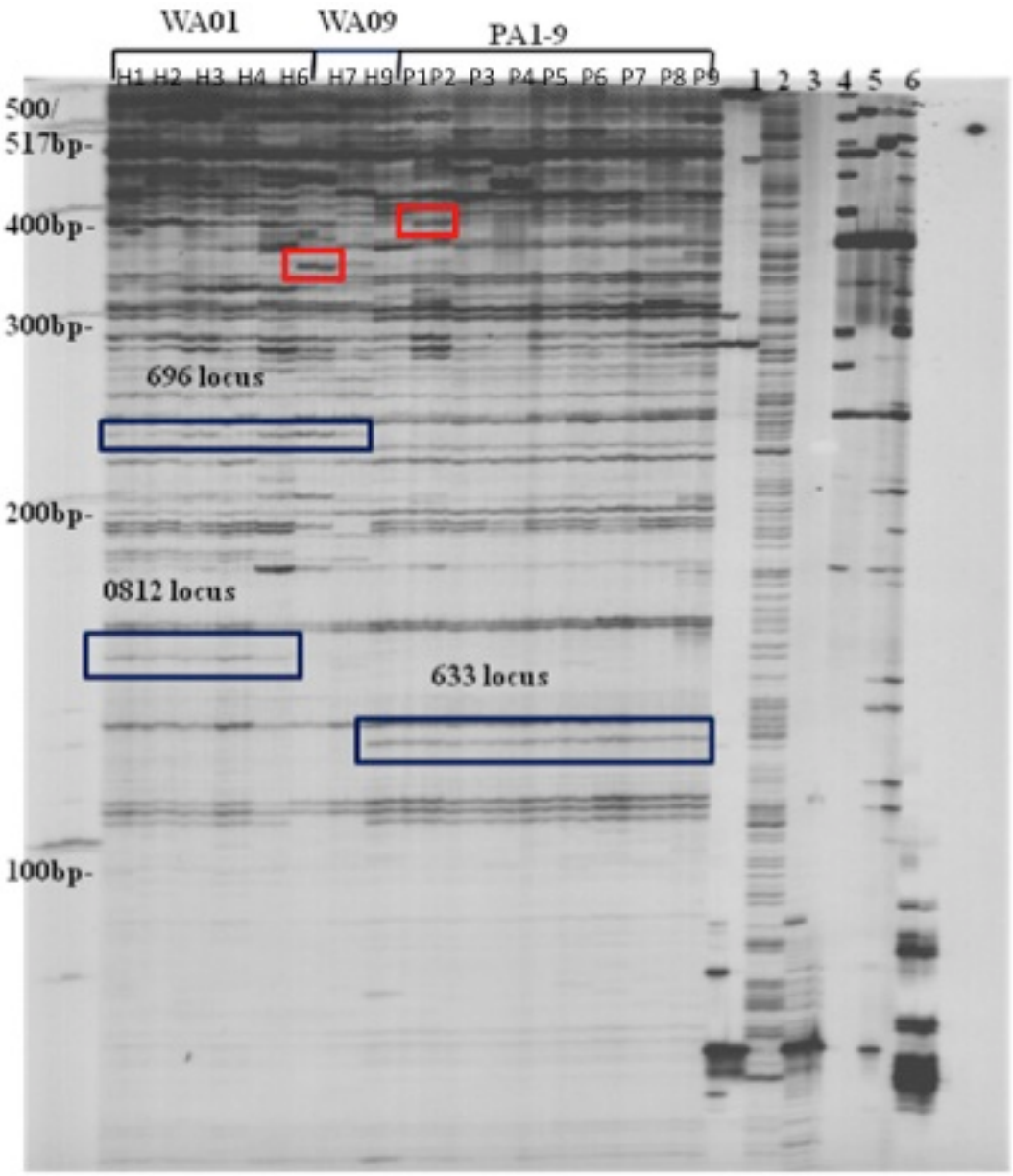

Figure. 3.1 5′- ATLAS display gel (*Mse*I library) of human embryonic stem cell clonal lines (hESCs) derived from WA01 (H1-H4 and H6) and WA09 (H7 and H9) as well as nine human ovarian carcinoma clonal cell lines (PA1). Controls for library construction: 1. Digest DNA –ve, 2. Digest Enzyme –ve, 3. Ligation DNA –ve, 4. Ligation linker –ve, 5. Ligation ligase –ve, 6. Primary PCR –ve.

3.2.1.1 Experimental validation and L1 insertion analysis

One of the variant bands, L1-AC090633, was present in all the PA1 derived clonal cell lines but absent from the hESCs, suggesting it was characteristic of the individual from which the cell line was derived. Sequence analysis revealed a polymorphic novel insertion that was absent from the reference human genome (hg19), and instead had been previously acquired by ATLAS population screening experiments. The AC090633-L1 insertion is a full length L1Hs, Ta1d subfamily with an allele frequency of 0.03.

The second band, AC108696, was present in the hESC clonal lines WA01 and WA09 but absent from the PA1 clonal lines. This AC108696-L1 is also a polymorphic, full-length novel L1 insertion that is absent from hg19. The Locus-696 band was recovered from the display gel and the insertion site was genotyped with the flanking DNA primers and the L1 5′-specific primer. The result is shown in figure 3.2, confirming that the AC108696 L1 insertion is specific to the hESC clonal lines.

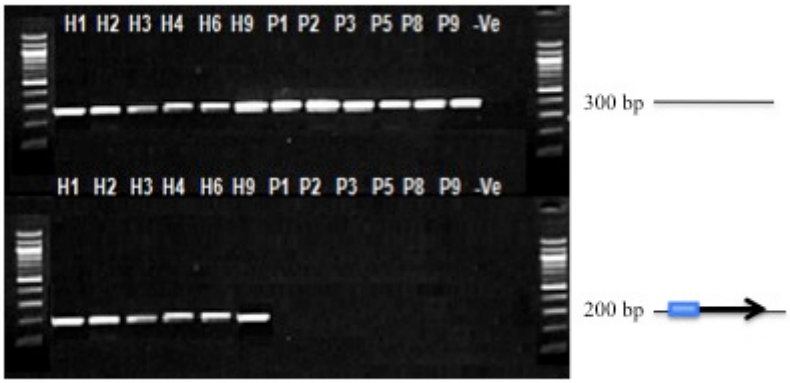

Figure. 3.2 hESC clonal cell lines (H1-H4, H6 and H9) and PA1 clonal cell lines (P1-P3, P5, P8 and P9) were genotyped for the AC108696 L1 insertion. A 300bp amplicon derived from the empty site (upper panel) is present in all the samples. 5′L1 amplification showed the presence of a 200bp product only in hESC clonal lines and not in the PA1 lines.

The next variant band, AC090812 (chromosome 18), was present only in WA01 clonal lines. Sequence analysis of the recovered band from the display gel revealed that the insertion was a novel L1Hs Ta1d not previously described in the reference

human genome. A genotyping assay was developed to amplify the empty site (RR0812A/RR0812B primers) and the filled site (RR0812A/RB5PA2), and the results confirmed a full-length insertion. Genotyping of 30 unrelated individuals from the CEPH family panel for this locus showed that four of these individuals were heterozygous for the insertion, giving an allele frequency of 0.06.

3.2.1.2 Display gel- sporadic bands

In addition to the display pattern variations that were observed between independently derived sets of clonal lines, there were several variants that were restricted to individual clonal cell lines, as illustrated in Figure 3.1 (see red boxes). A selection of these variable bands were excised and sequenced, and analysis of the sequences from these individual bands revealed that some of them belong to old L1 subfamilies, which are not expected to be active, but fixed in the human population. Other bands from the sequencing data were shown to be chimeras, resulting from chance ligation of DNA from different loci. None of the sporadic bands characterised (six bands) originated from novel full-length L1 insertions restricted to individual clonal lines.

3.2.2 Characterising variant bands in NTera2D1 clonal cell lines (*Ase*I library)

The same library construction procedures described in section 3.2 were applied to 30 NTera2D1 clonal cell lines (provided by Dr. N. J. Royle, University of Leicester). Some bands were only present in an individual clonal line (red boxes, Fig 3.3) and were absent from the progenitor line (Lane 1) as well as other clonal lines. The three variable bands (red boxes, Fig 3.3) were further re-amplified and sequenced. The result revealed that they belonged to the older L1 families such as L1PA2 that appeared on the display system due to mutations at the L1 primer site used in ATLAS (Figure 3.5). One of the variable bands was a concatamer, apparently resulting from Linker-Linker dimers formed during ATLAS library preparation.

To further investigate the L1 insertional variation in NTera2D lines we increased the sample size of the clonal lines and generated 110 single cell derived NTera2D1 clonal lines (details in section 2.2.1). Following DNA extraction from the clonal lines, an

*Ase*I library was constructed for 5′-ATLAS screening (section 3.2) and variable bands were observed in all of the five ATLAS display gels (data not shown). In total, 40 bands from the display gel were further characterised by cloning and sequencing. All the sequenced bands belonged to one of the following groups: older L1 families such as L1PA2; known L1HS insertions that were present in the human genome (hg19); concatamer products made during library construction. The results are summarised in figure 3.4.

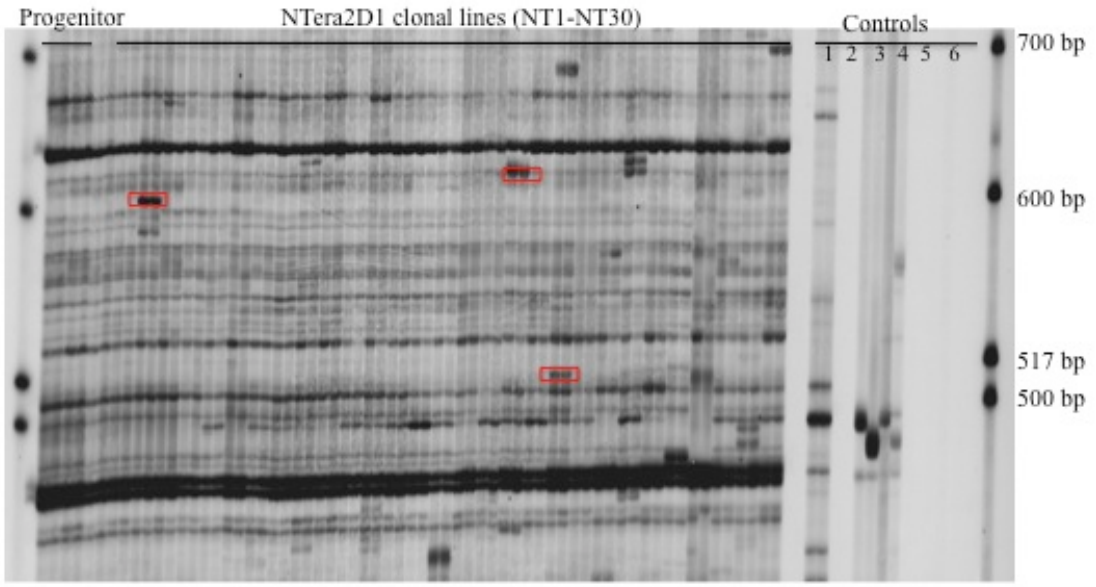

Figure. 3.3 ATLAS display gel for NTera2D1 cell lines (human tetracarcinoma cells). The *Ase*I restriction enzyme was used to make the library. Lane 1: progenitor line; Lane 2-31: clonal cell lines. The red boxes indicate sporadic variant bands that are only present in one clonal cell line and absent in the rest of the clonal lines.1: Digest DNA –ve; 2: Digest Enzyme –ve; 3: Ligation DNA –ve; 4: Ligation linker –ve; 5: Ligation control: ligase –ve; 6: Primary PCR –ve. Eliminating one or two components in the control samples show that it cannot generate the same pattern as when all the components are present.

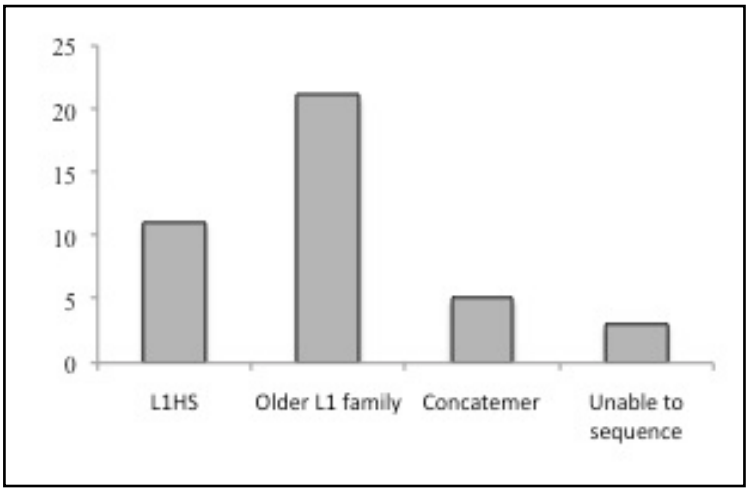

Figure 3.4 Summary of the 40 characterised bands from the ATLAS display gel of 110 NTera2D1 clonal cell lines. 27.5% of the sequenced bands belonged to known L1HS families (present in the hg19), 52.5% of the characterised bands were older L1 families such as L1PA2 and LMA8, 12.5% were L1 Linker-Linker dimers produced during library construction, and 7.5% of the bands had a poor sequence quality.

NTera2D1-clonal line 31: L1PA2; NT_006316; AC096719

ACGAGTGCGTCCTGCTACATCTCCAGCCTCATGATTTTAAAAAAACACACAGAAATC

L1 insertion site

ATTCTACTGGGGAGGAGCCAAGATGGCCGAATAGAAACAGCTCCGGTCTACAGCTCC

CAGCGTGAGCGACGCAGAAGACGCACTCGT

AGCGTGAGCGACGCAGAAGACGGTGATTTCTG (Primer sequence)

AGCGTGAGCGATGCAGAAGACGGTGATTTCTG (Sequence of clonal line 31)

Figure 3.5 Example of sequence mispriming in the primary PCR stage of the 5′-ATLAS preparation. An L1PA2 with a mutated C>T has amplified.

3.2.3 Characterising variant bands in hESC clonal cell lines 5´-ATLAS

3.2.3.1 hESC clonal cell lines - *Ase*I library

The same library construction procedures described in section 3.2 were applied to 20 human embryonic stem cell clonal lines (provided by Dr. Jose Garcia-Perez). The display gel is presented in Figure 3.7. As shown on the hESC display gel, no variation was observed amongst the clonal lines with respect to their L1 insertions *i.e.* all the loci were constitutive amongst the clonal lines and no presence/absence variation of any locus was observed.

The only observed variations were between the hEF control and the hESC clonal lines. This eliminates the possibility of hEF contaminating hESC during the culturing of the clonal lines with human embryonic fibroblast as a feeder. Fourteen different amplicons were randomly selected from the display gel and characterised. All the results are summarised in Figure 3.6.

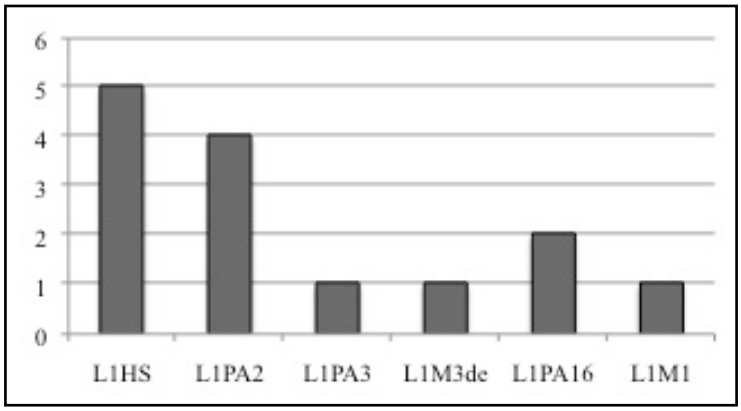

Figure 3.6 Summary of the characterised bands from the *Ase*I ATLAS display gel of hESC clonal lines 1-20 and their progenitor H9. 36% of the analysed bands were L1HS, and were present in all the clonal lines as well as the progenitor line.

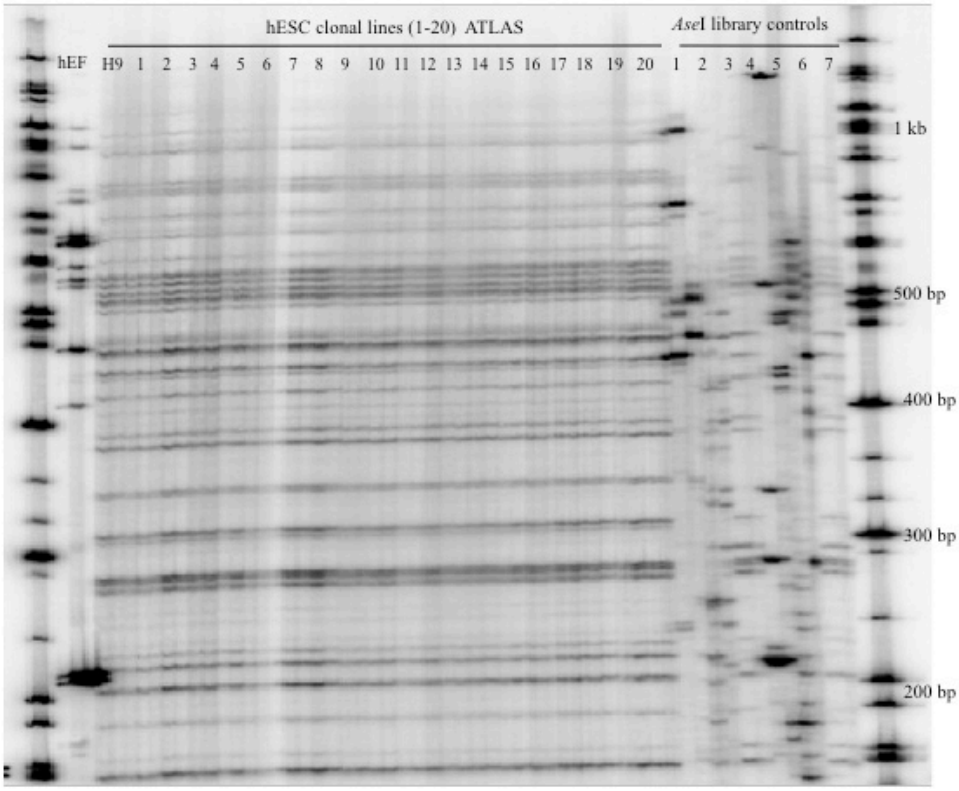

Figure. 3.7 ATLAS display gel for hESC clonal lines (*Ase*I library). hEF (human embryonic fibroblast) used as an initial feeder of hESC clonal lines, H9, progenitor line; 1-20: clonal lines; Controls: 1: Digest DNA –ve; 2: Digest Enzyme –ve; 3: Ligation

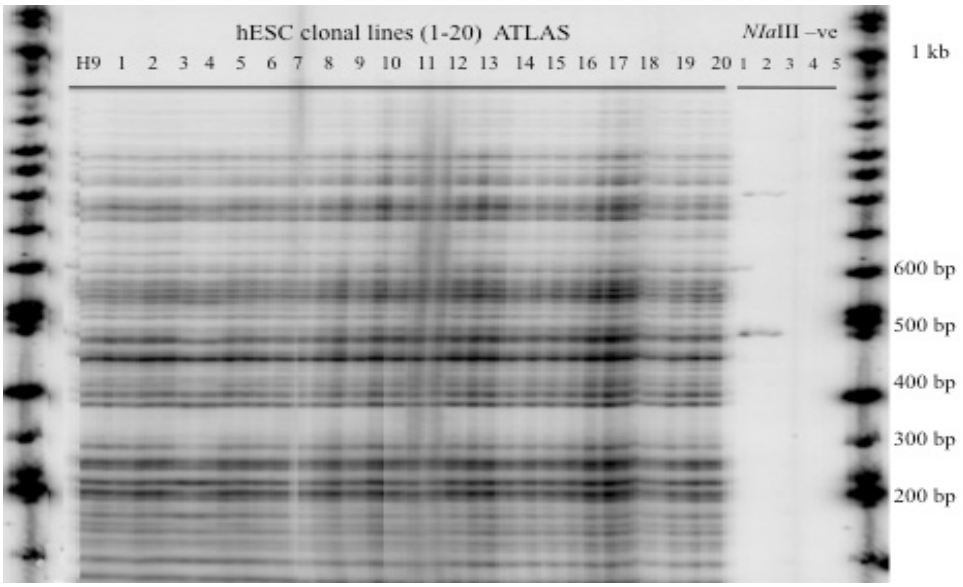

Figure. 3.8 ATLAS display gel for hESC clonal lines (*Nla*III library). H9, progenitor line; 1-20: clonal lines; Controls: 1: Digest DNA –ve; 2: Ligation linker –ve; 3: Ligation control: ligase –ve; 4: Primary PCR –ve. 5: display gel background control. As presented on the display gel the banding patterns are too complex for definitive analysis.

3.2.3.2 hESC clonal cell lines - *Nla*III library

In order to increase the genome coverage of ATLAS, we used the higher frequency cutting restriction enzyme *Nla*III (that recognises the 4 base pair sequence, 5′-CATG-3′). In contrast to the *Ase*I restriction enzyme, the *Nla*III restriction enzyme recognition site is not biased towards AT rich sequences and it gives up to 80% accessibility to the genome (the library construction is explained in section 2.2.7). *Nla*III-libraries display gels showed a greatly increased number of amplified L1 loci across the gel compared to the same sample analysed with *Ase*I, figures 3.7 and 3.8 respectively.

No clear intra-clonal variation was observed in the *Nla*III- display gel of the hESC clonal lines. However, due to the high background the display gel had a much lower resolution. This only allowed us to identify more pronounced variations (reflecting efficiently amplifying amplicons) and we could not detect any rare variants in this display gel. In order to validate that the display bands belonged to L1 loci, ten randomly selected bands were excised from the display gel and sequenced. However, the sequences were of poor quality and had a high background so we were unable to further characterise them. This was likely because the *Nla*III library had a higher genome coverage, which resulted in amplification of many more L1 loci simultaneously. Hence it was likely that each individual band on the display gel may represent more than one L1 locus, which would produce high background noise during direct sequencing.

3.2.3.3 Characterising variant bands in hESC clonal cell lines –*Nla*III differentiating Y1_primer library

Because the *Nla*III constructed library display gel was very complex, unequivocally distinguishing variability was impossible. In order to make this library less complex we used secondary primers (Y1A, Y1T, Y1C and Y1G) that discriminate amplicons on the basis of the nucleotide immediately 3′ of the *Nla*III site. The methodology for constructing LC-ATLAS (Lower Complexity ATLAS) is explained in section 2.2.8. A display gel of hESC LC-ATLAS is presented in figure 3.10.

Figure 3.10 shows that several variable bands were observed amongst the clonal lines, which could not be detected by using the high-complexity ATLAS display gel (non-differential *Nla*III ATLAS) or low-coverage ATLAS display gel (*Ase*I library). 40 of these variable bands were excised from the display gel and they were characterised by sequencing. All the sequenced bands are summarised in figure 3.9. Further investigation was carried out for three loci (bands 1, 3 and 25). In all these three loci L1.3 similar sequence was detected, however no L1 repeats were detected at these loci in hg19, and therefore they looked like suitable candidates for *de novo* L1 insertions. Genotyping assays were designed and optimised for all three loci. The genotyping result of L1 -AC068631 (band 1) and L1- AL133402 (band 3) showed that they were present in all the hESC clonal lines as well as the progenitor (H9) and so showed a false variability (dimorphism) amongst the clonal lines on the display gel. Both bands 1 and 3 were novel L1 insertions (absent from the hg19) and they both were full-length L1HS with allele frequencies of 0.03% and 0.01% respectively (genotyping panel = 100 unrelated CEPH individuals) Since allele frequencies of both insertions were low (< 0.05%) it is suggested that they are relatively young L1s with low prevalence in the population.

Sequence analysis of the L1-AJ510022 locus (figure 3.11) showed that it is a putative L1Hs insertion on chromosome X (chrX: 61,720,513-61,720,721). However, further PCR-based genotyping of the L1 insertion failed as the L1 had inserted into satellite DNA, making the PCR assay non-specific. Therefore, we cannot further discuss the status of this insertion; it could be a genuine *de novo* L1 insertion or a false positive, but these possibilities cannot be distinguished.

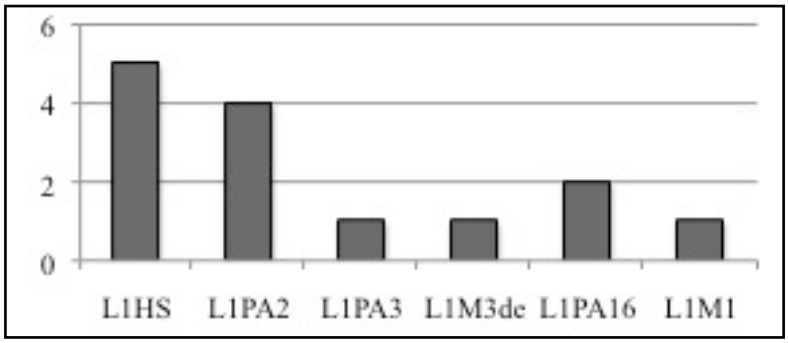

Figure 3.9 Summary of the characterised bands extracted from LC_ATLAS display gel of hESC clonal lines 1-7 and their progenitor H9. The majority (46%) of the analysed sequences belong to the L1HS family.

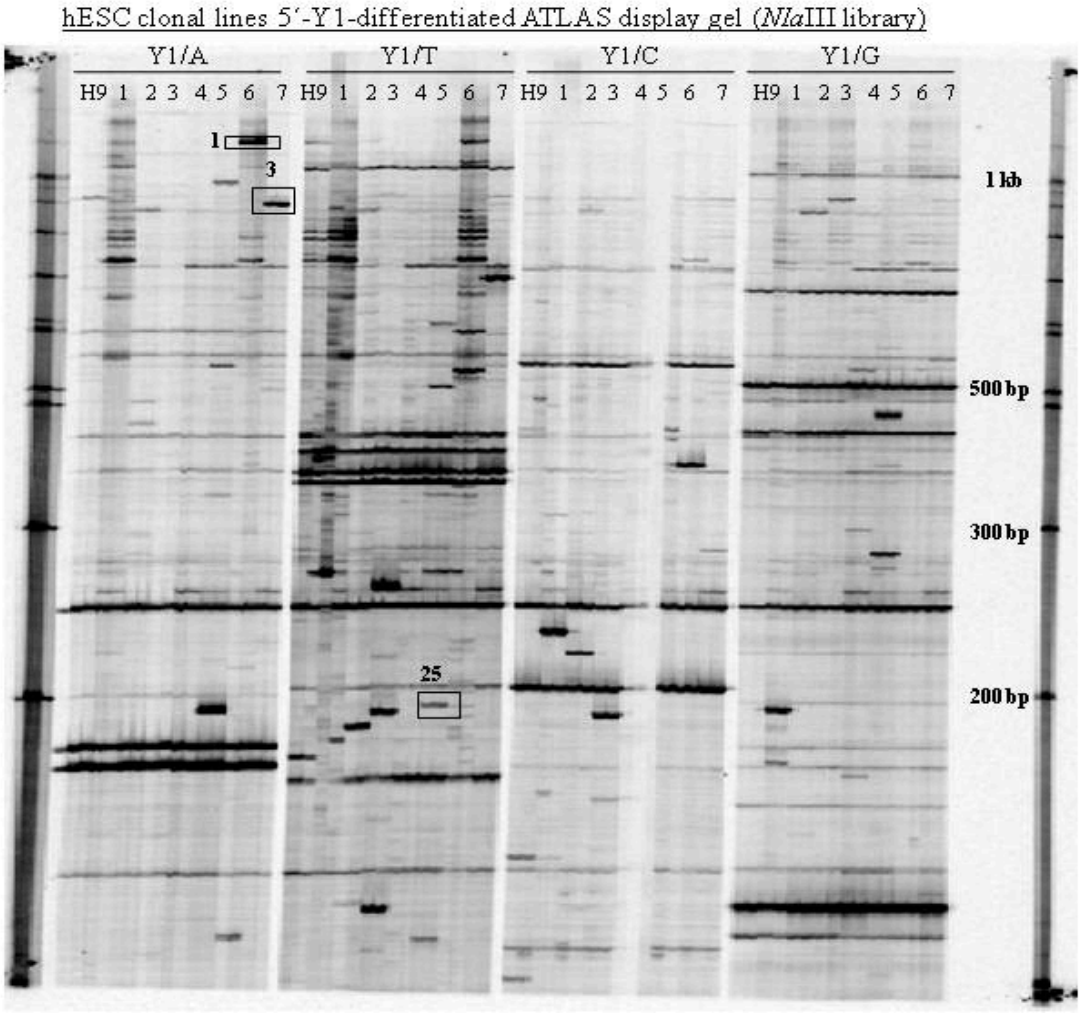

Figure 3.10 LC-ATLAS display gel for hESC clonal lines (*Nla*III library). H9, progenitor line; 1-7: clonal lines; Controls: 1: Primary PCR –ve, 2: secondary PCR –ve, 3: display PCR –ve control. More L1 loci variability can be observed on this gel due to its higher sensitivity and lower complexity. Bands 1, 3 and 25 were novel L1Hs insertions (absent from hg19).

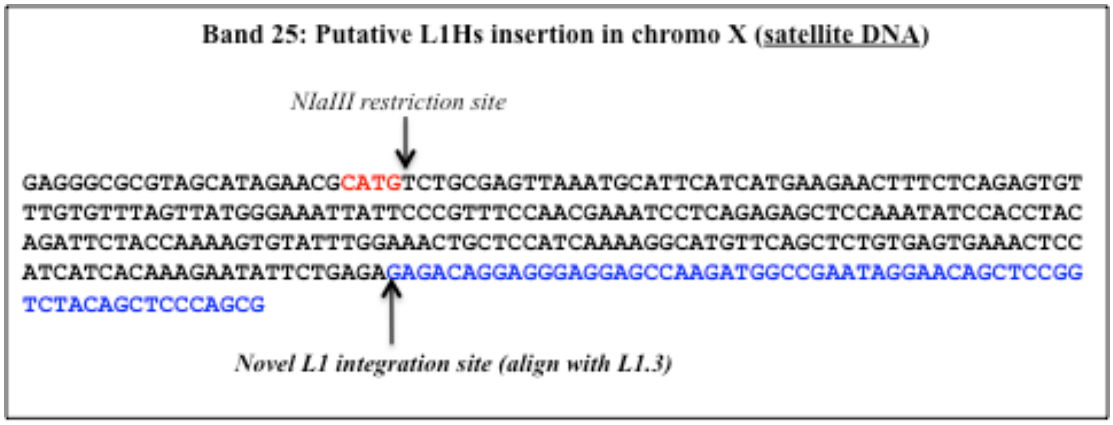

Figure 3.11 Band 25 from display gel (figure 3.10) extracted, cloned and sequenced. The *Nla*III restriction site is shown in red, the sequence in black (after the restriction site) is part of the satellite region in chromosome X, the blue sequence is absent from the hg19 sequence and it aligned with L1.3 sequences.

3.2.4 Diagnostic L1 insertion, characteristic for the HeLa cell line

In Using 5′-ATLAS to conduct comparative screening of HeLa and other human cell lines, an L1 insertion which appeared to be private to HeLa clonal lines and absent from other cell lines was revealed (the insertion was isolated from a display gel made by V. Modes). This full-length L1 insertion, AL137164, is found on chr14q12 in HeLa cells within intron 2 of the *STXBP6* (*amisyn*) gene. This L1 element is inserted in the same transcriptional orientation as the *amisyn* gene, has intact open reading frames (ORFs), and thus may be active in cell culture retrotransposition assays (V. Modes and R. Badge, unpublished data).

3.2.4.1 Genotyping of the AL137164 L1 insertion

An initial population screening experiment, using a standard PCR genotyping assay developed previously by Modes *et al*. (unpublished data) for this locus was carried out. The flanking genomic DNA of the empty site was amplified using the primers VM164A and VM164B. The 5′ upstream region of the L1 insertion (the filled site) was amplified using VM164A and the L1 internal primer RB5PA2. A schematic diagram of the PCR genotyping assay and the size of the products are shown in Figure 3.12.

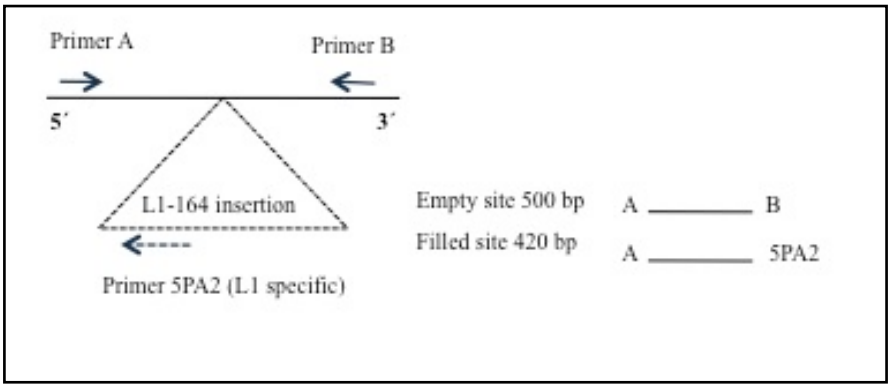

Figure 3.12 Schematic diagram of the PCR-based genotyping assay for Locus-164 the flanking site primers A and B were used to amplify the 500 bp empty site. Primer A and the L1 internal primer (5PA2) were used to amplify the 420 bp filled site.

This genotyping assay was applied to 5 independently sourced HeLa cell lines, and 5 non HeLa cell lines. Various laboratories in the USA and the UK donated the HeLa lines and all had been independently sourced from cell line stock centres (listed in Table 3.1).

Culture	Culture information	Culture source
H1	HeLa	Prof Andrew Fry (University of Leicester, Leicester, UK)
H2	HeLa	Prof Fred Gage (Salk Institute, La Jolla, CA, USA)
H3	HeLa	Dr Raj Patel (University of Leicester)
H4	HeLa	Prof John Moran (University of Michigan, Ann Arbour, MI, USA)
H5	Hep2 [morphologically identical to HeLa (Moore *et al*., 1995) and by STR profiling]	Dr Simon Kilvington (University of Leicester)
H6	HeLa S3 [clonal sub-line derived from the original HeLa culture (Puck *et al*., 1956 and Chen *et al*., 1988)]	ECACC/HPA (Salisbury, UK)
N1	AJ (patient lymphoblastoid cell line) (Varley *et al*., 2000)	Dr Nicola Royle (University of Leicester)
N2	BJ (primary foreskin fibroblast cell line)	Dr Nicola Royle; ATCC USA (Manassas, VA, USA)
N3	GM03798 (lymphoblastoid cell line)	Dr Nicola Royle; Coriell Cell Repositories (Camden, NJ, USA)
N4	NTera2D1 teratocarcinoma	Dr Nicola Royle; ATCC USA
N5	HepG2	Dr Fred Tata (University of Leicester); ECACC/HPA

Table 3.1. Cell Lines Used in This Study, Rahbari *et al*., 2009.

Preliminary genotyping results suggested that the L1-164 insertion could potentially be an insertion private to HeLa cell lines To confirm this hypothesis the population frequency of this L1 insertion was investigated using a sensitive dot blot assay on 72 DNA samples from unrelated black Zimbabwean individuals, as this sample set is most closely related to the HeLa donor's likely population of origin. The genotyping

result revealed that all the Zimbabwean samples were homozygous for the insertion empty site (figure 3.13). Genotyping 364 unrelated individuals from geographically diverse populations (Africans, African-Americans, Northern Europeans, German whilte-origin, Asians, and South Americans) failed to identify any other carriers of the AL137164 insertion.

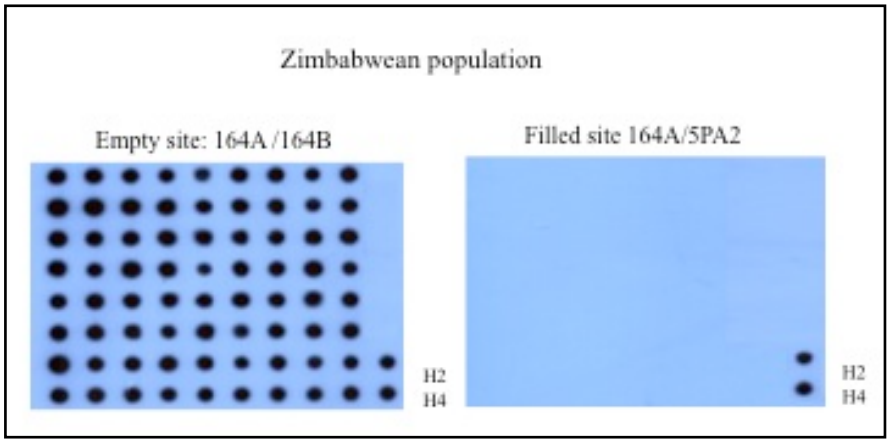

Figure 3.13 Dot-blots of empty and filled sites for Locus-164 in a Zimbabwean population. H2 and H4 are two independently sourced HeLa cell lines, which were used as positive controls for the presence of the Locus-164 insertion. The Locus-164 insertion was absent in the Zimbabwean population, but present in the H2 and H4 positive controls.

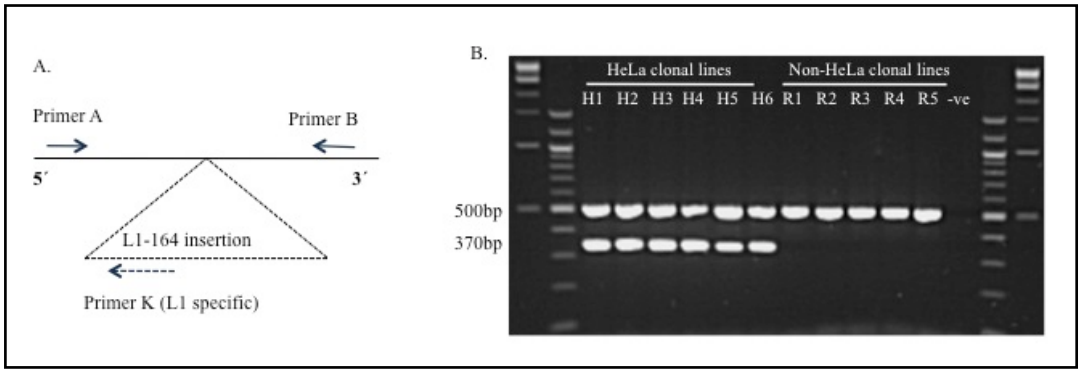

Figure 3.14 A. The three-primer-based PCR assay is a diagnostic test for HeLa contamination. Primers A and B are the universal primers responsible for amplification of the flanking genomic DNA. In the presence of a Locus-164 insertion, Primers A and K will only amplify the insertion. **B.** Three-primer PCR-based assay on H1-H5 (independently sourced HeLa cell lines) and R1-R5 (non-HeLa cell lines). The insertion site (370 bp) was only amplified in HeLa cell lines.

3.2.4.2 Diagnostic single duplex PCR for HeLa contamination

The original genotyping assay was modified into a single duplex PCR format (figure 3.14 A), providing a simple way to discriminate HeLa and non-HeLa cell lines. As shown in figure 3.14 B, all tested HeLa isolates, including the HeLa S3 sub-line (H6), were positive using this assay (*i.e.* they all generated a 370-bp insertion–specific amplicon). A DNA mixing experiment (figure 3.15) using HeLa DNA and NTera2D1 genomic DNA indicate that the single-tube PCR assay is highly sensitive, and can detect 1% of HeLa mixed with non-HeLa DNA. The assay is also robust with respect to the input material: even genomic DNA present in un-purified frozen cell pellets (figure 3.16) can be genotyped using this assay.

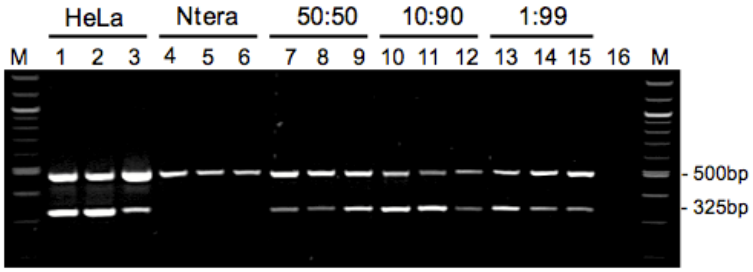

Figure 3.15 Sensitivity test for the PCR-based assay. Samples are all in triplicates. Sample 1-3: 100% HeLa cell line; Sample 4-6: 100% Ntera2D1 cell line; Sample 7-9: 1:1 ratio of HeLa: Ntera2D1 cell lines; Sample 10-12: 9:1 ratio of Ntera2D: HeLa cell lines; Sample 13-15: 99:1 ratio of Ntera2D: HeLa cell lines; Sample 6: 0.1 dilution of the HeLa cell line; and the DNA –ve control for the PCR assay.

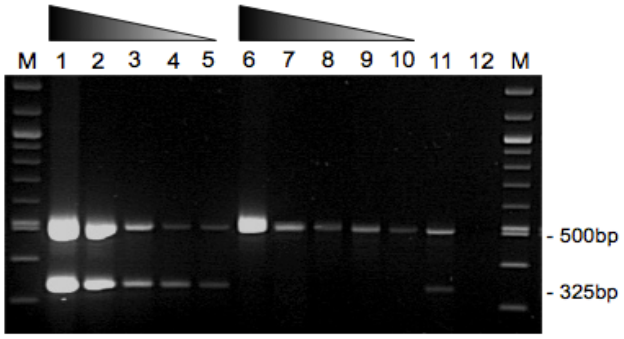

Figure 3.16 a PCR-based assay for detecting HeLa cross-contamination using cell pellets (CP). Samples are in triplicates. Sample 1: 25µl CP+ 25µl 5MT; Sample 2: 5.0µl CP+ 45µl 5MT; Sample 3: 25.0µl 10^{-1} dilution of CP + 25.0µl 5MT; Sample H4: 10µl of 0.1µl 10^{-1} CP dilution + 40.0µl 5MT; Sample 5: 5µl 10^{-1} dilution of CP+45.0µl 5MT; Samples 6-10-non-HeLa cell pellet (N1-N5 have the same dilutions as H1-H5) were used as a control for this experiment. +ve control is a HeLa DNA and the–ve PCR control lacks any DNA input.

3.3 Discussion

As mentioned in the introduction section of this chapter, previous studies suggested that L1 retrotransposition is likely to occur during early human embryogenesis (Freeman *et al.,* 2011, Garcia-Perez *et al*., 2007 and Van den Hurk *et al*., 2007). Therefore this chapter aimed to try to demonstrate ongoing endogenous L1 retrotransposition during human embryonal cells. To do this we used the display method to screen for *de novo* L1 retrotransposition in human embryonal cell lines.

Embryonal carcinoma cells are undifferentiated stem cells derived from teratocarcinoma tumours (Michiko *et al.,* 1984). They are pluripotent stem cells with developmental equivalence to normal early embryonic stem cells. Isolated cell lines show a remarkable biochemical and immunological resemblance to early embryonic cells (Michiko *et al.,* 1984). Several human embryonal cells have been used as a model for human embryogenesis, including: PAI, NTera2D1 and hESC. The PAI cell line is derived from human ovarian teratocarcinoma cells. The late-passages of these cells represent a fairly homogenous population of malignant cells similar to embryonal carcinoma cells (Zeuthen *et al*., 1980). The NTera2D1 cell line is a clonal subline of Tera2, which is a human teratocarcinoma cell line; its pluripotent character and identity as an ES cell line (Andrews, 2002) have made it a suitable model for studying human embryogenesis. In addition, it is known that the NTera2D1 line expresses the full-length coding strand of genomic LINE-1 (L1) elements (Skowronski *et al*., 1988). HESCs are cells derived from the early embryo and can be propagated indefinitely in a primitive undifferentiated state while remaining pluripotent. They share these properties with embryonic germ (EG) cells (Pera M.F. *et al*., 2000). HESCs are known to be amenable to the expression of endogenous L1 retrotransposition (Garcia-Perez *et al.,* 2007) and therefore, they can be used as a model to investigate the activity of L1 retrotransposons in early human development. Here we have demonstrated that single cell clonal lines can in principle be used to study endogenous L1 insertions in clonal human embryonic cell lines. The technique is based on the fact that extended culturing of these cells may allow ongoing endogenous L1 retrotransposition, forming a mosaic population of cells. In theory these insertional mosaics could then be directly identified by isolating single cells from mosaic populations of hESC, NTera2D1 and PA I clonal cell lines.

Individual clonal lines represent samples of the extent of the insertional diversity of the progenitor population. As is discussed in more detail in the Introduction section (section 3.1) of this chapter, 5′ ATLAS is preferable to 3′ ATLAS as it can selectively amplify full length L1-elements (Badge *et al.,* 2003), whereas 3′ ATLAS amplifies all the L1-Ta subfamilies regardless of whether their 5′ends are truncated or not. Also, 5′ ATLAS produces a higher resolution display gel compared to 3′ ATLAS, due to the varying lengths of poly-A tails at the 3′ end of L1, which can result in an uneven size distribution of products. Hence, using these amplicons with variable sizes can obscure the display gel image, making it harder to detect lower frequency insertion events.

3.3.1 Comparative 5′ ATLAS of hESC Vs PA1 clonal cell lines

For this comparative study the 5′ATLAS technique (Badge *et al*., 2003) was optimised in order to analyse L1 insertional mosaicism within the clonal lines. This optimised 5′ATLAS allows comparison of the insertional diversity of the clonal lines with the bulk DNA from progenitor populations in order to detect full-length *de novo* insertions.

The comparative 5′ ATLAS display gel between hESC and PA1 clonal cell lines is presented in section 3.2.1, with different lines on the gel representing different L1 insertion loci. Constitutive L1 insertions appeared as common bands between all the cell lines on the display gel. However, some intra- and inter-cell line variation can be observed, as well as sporadic variation which was only present in one or two of the clonal lines and absent from the rest. Further characterisation of bands from the display gel resulted in the identification of two full-length polymorphic L1HS: AC108696 (allele freq. 0.03) and AC090633 (allele freq. 0.06). Both insertions represent L1 inter-lines variations, as they are novel insertions with respect to hESCs and PAI cell lines respectively. Another recovered polymorphic full-length L1HS, AC090812 L1 (allele freq. 0.03), represents an L1 intra-cell line variation as it is restricted to the WA01 hESC clonal lines and absent from the WA09 and PAI clonal lines. Besides this intra and inter clonal cell line variation, several sporadic variations were also observed, *i.e.* bands which were present in one or two of the clonal lines and absent from the rest of the clones. It is important to mention that analysing the sporadic bands is of great interest to find *de novo* retrotransposition events, *i.e.* one

possible explanation for the sporadic bands on the display gel is *de novo* L1 retrotransposition events during the culturing of the clonal lines. Analysis of recovered sporadic bands from the display gel showed that they were chimeric products, which can be either introduced during the ATLAS library construction or can be a result of the stochastic nature of PCR amplification. However, since in all cases these sporadic bands were observed in duplicated samples rather than being a singleton event it is more likely that they are products of the library construction rather than stochastic PCR amplification. Since the progenitor lines of all the three cell lines (WA01, WA09, and PAI) were not available we were not able to further compare the clonal lines for novel L1 insertions during the process of making the clonal lines. However, since all the clonal lines share the same insertions, it is less likely that the L1 retrotransposition happened during the cell culturing process and it is more likely that they would also be present in the progenitor lines. All the observed L1 insertional variation between different clonal lines on the display gel has been validated by genotyping the genomic DNA for each insertion. This shows that the 5′ ATLAS technique is a powerful and sensitive approach to study L1 insertional variation in human embryonal cells.

3.3.2 Activity of L1 retrotransposons in NTera2D1 clonal cell lines (*Ase*I library)

The optimised 5′ ATLAS method was applied to 30 NTera2D1 clonal cell lines (provided by Dr. N. J. Royle, University of Leicester). As is shown in figure 3.3, variation was detectable in clonal cell lines. This variation was absent in the progenitor cell line, and therefore it was likely to have arisen from a *de novo* insertion. Further characterisation of the recovered bands from the gel revealed that the majority of these insertions belonged to older L1 families such as L1PA2, which appeared in the display system due to a mutation in the L1 primer site. An example of this mis-priming effect is demonstrated in figure 3.5. Also, some of the observed variability was due to the formation of chimeric products (linker-linker fragments) during library construction. Overall, amongst the 30 clonal lines and their progenitors, no *de novo* L1 retrotransposition was characterised. Since recent studies have estimated the rate of the L1 retrotransposition to be 1 in 95-270 individuals (Ewing and Kazazian, 2010), one of the explanations for not observing any *de novo* L1 insertions was that not enough clonal lines were screened. Hence a larger set of

NTera2D1 clonal lines was produced (n=110) for more 5´ATLAS screening. Forty different variable bands across five display gels were recovered and further characterised by cloning and sequencing, and this revealed 27.5% of the recovered bands to be polymorphic L1HS (present in the hg19), and 52.5% of the characterised bands to be older L1 families such as L1PA2 and LMA8. A fraction of the sequenced bands (12.5%) were L1 Linker-Linker products produced during the Library construction.

Overall we have analysed 110 NTera2D1 single-cell derived clonal lines using *Ase*I to screen for *de novo* L1 insertions, but we could not isolate any genuine *de novo* insertions. This could be explained by having a low number of screened clonal lines, and we might have been able to find *de novo* insertions if we could have screened more lines. Also, low genome coverage could be another possible explanation. Since we have used *Ase*I for library construction of these clonal lines and knowing that this restriction enzyme gives only 11% accessibility to the whole genome (*i.e.* only 11% of the genome is within 1kb of a *Ase*I site, by *in silico* genomic digest), this constrained library coverage might have been a limitation in finding new insertions. Overall 11% of the genome of each of 110 genomic DNA samples of the clonal lines have been analysed and the number of amplified loci calculated, and it can be thus concluded that we have analysed 726 pg of the genomic DNA in the NTera2D1 cell line and we could not find any *de novo* L1 insertions. Based on this we can estimate the rate of L1 retrotransposition in these cell lines to be less than 1 in 110 cells. Although it has been shown that NTera2D1 supports L1 retrotransposition by expressing ORF1, it cannot be said with certainty that ongoing endogenous L1 retrotransposition occurs in these cell lines. Since they have been in tissue culture for a very long time it is possible that as a part of cell culture adaptation, L1 retrotransposition does not occur very often in these cell lines.

3.3.3 Activity of L1 retrotransposons in hESC clonal cell lines

Standard 5´ATLAS screening was applied to 20 clonally derived hESCs and their progenitor (H9) (provided by Jose Garcia-Perez, Granada, Spain) results are summarised in section 3.2.3. Display gels of these clonal lines revealed no variation amongst the clonal lines and their progenitors. The only observed L1-insertional variation was between H9-derived clonal lines and human fibroblasts on the gel. This

observation proves that the technique was successful in detecting insertional variation between independent cell lines. Further sequencing of several constitutive bands showed that a majority of the recovered bands (36%) belonged to the L1Hs Ta-1 family. As the results show, there is no L1 insertional variation in the *Ase*I-constructed hESC clonal lines. *Ase*I is an enzyme with a six base pairs recognition sequence and it is biased towards AT-rich regions. Since this enzyme cuts less frequently in the genome, the genome coverage of the library constructed by this enzyme is only 11%. Therefore, the *Ase*I constructed libraries have poor resolution compared to more frequently cutting enzymes and are biased towards AT rich genomic regions. Hence, this library may not be sensitive enough to screen for rare events *i.e. de nov*o insertions. In order to increase genome coverage the *Ase*I enzyme was replaced with the *Nla*III restriction enzyme. *Nla*III is a more frequent cutter and can cover almost 80% of the genome. Also, this enzyme does not have a sequence composition biased restriction site, and so can present a less biased picture with higher coverage of the genome than the *Ase*I analysis. All the methodology and linker sequences for construction of *Nla*III ATLAS libraries are in Chapter 2.

Following the optimisation of the modified-ATLAS, it was applied to the hESC clonal lines. As the display gel in figure 3.8 demonstrates, more L1 loci have been amplified by using the modified technique. Comparative analysis between the number of amplified bands on the display gel from *Nla*III and *Ase*I libraries showed that more visible L1 loci were amplified by using the more frequently cutting restriction enzyme. However, the exact increase in number cannot be verified, since some of the bands on the display gel may represent more than one L1 locus and therefore the total number of amplified loci could be more than what is observable on the display gel. Using the *Nla*III library increased genome coverage and increased the number of amplified L1 loci on the display gel, which resulted in higher display gel complexity. With the limited resolution and more complex nature of the gel, verification of rare insertions proved to be harder. Further characterisation of 10 randomly recovered bands from the modified-ATLAS display gel failed, due to the high sequence background noise for each band. In most of the cases each band represented more than one L1 locus and therefore it was not possible to fully characterise each amplicon.

This result showed that although the *Nla*III ATLAS has a higher coverage than *Ase*I the higher complexity of the *Nla*III display gel made screening for *de novo* L1 insertions very difficult. To lower the complexity of the display gel and make it more accessible for sequence analysis we developed a lower complexity *Nla*III ATLAS by using a differential linker primer (Y1 primer) for the secondary PCR. The method for constructing a LC-ATLAS (low complexity ATLAS gel) is explained in detail in Chapter 2.

Following the optimisation of the LC-ATLAS method it was applied to the hESC clonal lines. Interestingly the display gel of hESC clonal lines (figure 3.10) revealed L1 insertional intra-variation between the clonal lines, which in the standard *Nla*III or *Ase*I ATLAS was not observed due to the high complexity and low coverage respectively. Forty of these variable bands were extracted from the display gel and further characterised by cloning and sequencing. The result revealed that the majority of the analysed bands (46%) belong to the L1Hs subfamily. These are polymorphic elements and showed some level of dimorphism amongst the clonal lines on the display gel. However, it is more likely that the polymorphic L1s are present in all the clonal lines and the false insertional variation can all be explained by the high sensitivity of this display gel, *i.e.* some of the loci may not have amplified with the same efficiency across clonal lines, and this can result in observations of false variation on a sensitive and less complex display gel. Characterisation of three of these bands (1, 3 and 25) suggested that they could potentially be *de novo* insertion events as all three insertions were absent from other clonal lines and the progenitor, and they were also absent from the hg19. Genotyping assays were developed and optimised for these three loci and all the hESC clonal line gDNAs were genotyped. Genotyping results for bands 1 and 3 revealed that they were present in the genomic DNA of the rest of the clonal lines and the progenitor. Further genotyping of these loci in the CEPH panel showed that they have allele frequencies of 0.03 (L1-AC068631) and 0.01 (L1-AL133402) respectively. Since the allele frequencies of these two insertions are relatively low, the efficiency of PCR amplification may vary. This means that great care should be taken when interpreting the display gels; for example, the image of these two loci on display gels in figure 3.10 shows singleton bands that could be interpreted as *de novo* insertions. In fact, genotyping of genomic DNA showed that they were present in all the clonal lines and so are constitutive and

not *de novo*. This problem may not have been encountered previously with the ATLAS display technique. In the modified version (LC-ATLAS) the level of display gel complexity is significantly reduced to make it a more sensitive display system and hence able to detect more PCR products from low frequency loci. These low allele frequency insertions are more likely to be recent events, and therefore they are worth pursuing for characterisation.

Genotyping the third L1 insertion (AJ510022) failed due to its genomic location. The sequence obtained from the recovered band on the display gel clearly showed an L1.3-like sequence, which was inserted in the satellite DNA on the X chromosome. This insertion was not present in the hg19 and by its appearance on the display gel, it is also absent from the progenitor and the rest of the clonal lines. Therefore it potentially could be a genuine *de novo* L1 insertion in this clonal line. However, due to the high rate of false insertion variation of this gel this cannot be confirmed without further genotyping. Although we could not detect or characterise any *de novo* insertions among the hESC clonal lines (n=6) using the LC ATLAS it does not mean that *de novo* L1 retrotransposition does not occur in these clonal lines. Perhaps one explanation for this observation could be the limited number of clonal lines that have been analysed for this experiment (n=6). Also the existence of barriers to further amplification of *de novo* insertions at their genomic inserted sites could be another factor in not being able to report any L1 *de novo* insertions among these clonal lines. Perhaps screening more clonal lines using the LC-ATLAS could lead to finding new insertions in these clonal lines, but this was out of the timeframe of the current research. However, we have demonstrated that the LC-ATLAS has the potential to find new insertions due to its high sensitivity as well as low complexity.

3.3.4 A Diagnostic L1 insertion to identify HeLa cells

HeLa cell lines were first grown in cell culture more than 50 years ago (1951) by George Gey (Masters, 2002). Since then, HeLa-derived cell lines have been the first line subjects for study in cancer research. Following the establishment of HeLa cells as the first cancer cell line which could be grown in cell culture, thousands of continuous cell lines from almost every type of human cancer have been established

during the 1970s and 1980s (Masters, 2002). Due to the long period of culturing HeLa, today they are highly adapted to the tissue culture environment and therefore they have great potential to cross-contaminate other cell lines. There is evidence to suggest that a number of publications have been based on cell lines, which are known to be contaminated with HeLa cells (Drexler *et al*., 2003; Lacroix *et al*., 2008; Gartler, 1967). Table 3.4 summarises some of these HeLa-contaminated cell lines. STR-based DNA fingerprinting technology is now commonly used to check cell cultures for cross-contamination (Masters, 2001). However, the expense of the STR-based technique is a problem for many research laboratories. Therefore, the STR-based technique is beneficial when used on an industrial scale by cell culture vendors, but less preferable when used in research laboratories for the monitoring of cell cultures for contamination.

In section 3.2.4 we have demonstrated how a private L1 insertion can be used as a diagnostic tool for HeLa cell contamination. In the process of screening for new L1 insertions in HeLa cell clonal lines (Mode V. and Badge R unpublished data) the L1 164–locus was isolated from the display gel. This insertion appeared as a novel L1 insertion in HeLa cells, and is absent from hg19. This observation was further validated by a PCR-based genotyping assay for 72 Zimbabwean individuals (Zimbabwean individuals were used because the HeLa cell line is African-American in origin). The genotyping result revealed that the insertion is absent from the entire genotyped population and therefore larger (364 individuals) and geographically diverse populations (including African, Asian, African-American, South American and German European-origin) were also genotyped for this locus, showing the insertion to be absent from the whole panel.

Since it appeared that the L1-164 insertion is only present in HeLa cells, this property makes the insertion valuable as a genetic marker for the detection of HeLa cell lines. Based on these observations a basic PCR-based assay that can be easily used for the monitoring of HeLa contamination in the research lab was designed. To make the method more applicable to cell culture research labs the sensitivity of the PCR-based assay to HeLa contamination was tested. As is shown in the results section (Figure 3.15), this PCR-based assay is highly sensitive and can detect low levels of HeLa cell cross-contamination within cell cultures. The developed PCR assay can use unpurified cell pellets as its source of input DNA. This alternative eases the workload

and expenditure in cell culture labs for the routine monitoring of their cell cultures for HeLa cross-contamination (Rahbari *et al*., 2009).

Cell line	*Putative origin*	*True identity*
Chang liver	Liver cells	HeLa cells
Girardi heart	Atrial myoblast cells	HeLa cells
Hep-2	Larynx carcinoma cells	HeLa cells
INT407	Embryonic intestine cells	HeLa cells
J111	Monocytic leukemia cells	HeLa cells
KB	Oral epidermoid carcinoma cells	HeLa cells
L132	Embryonic lung epithelium cells	HeLa cells
MT-1	Breast cancer cells	HeLa cells
NCTC2544	Skin epithelium cells	HeLa cells
WISH	Amnion cells	HeLa cells

Table 3.4 List of cross-contaminated cell lines. The above cell lines have still cited up to 2007 by scientists who were not aware of the true identity of these cell lines (reviewed in Lacroix *et al.,* 2007; also cited in Nelson-Rees *et al.,* 1976; Chen TR 1988; Macleod RA, 1999; Gartler 1967; Ogura *et al.,* 1993; Lavappa *et al.,* 1976; Nelson-Rees *et al.,* 1981), Rahbari *et al*., 2009.

Overall although we could not isolate any *de novo* L1 insertions from the screened embryonal clonal cell lines we have developed a more sensitive technique, which is more compatible with rare insertions, including *de novo* insertions. Also, we have demonstrated one potential *de novo* L1 insertion in hESC clonal lines as well as characterised novel L1 insertions in different clonal lines. Some of these are private to a specific cell line and therefore they can potentially be used as a marker in tissue culture. Finally, it can be concluded that combining a high coverage ATLAS variant (*Nla*III library) with a high throughput sequencing technique would yield a more optimal approach for the screening cellular genomes for endogenous *de novo* L1 insertions. This approach is discussed in more detail in Chapter six.

Chapter 4

Tracing active L1 lineages using 3´-transductions

4.1 Introduction

The role of L1 in shaping our genome is undeniable. In addition to self-mobilisation, L1 elements can also retrotranspose their flanking genomic DNA to another location in a process known as transduction. If this process occurs due to transcriptional initiation upstream of the L1 5´ terminus it is known as 5´ transduction (Pickeral *et al*., 2000) While 5´ transduction is not uncommon, transduction of 3´ flanking genomic DNA is much more frequent (Szak *et al.,* 2003), and is not dependent upon fortuitous activation of upstream cellular promoters. 3´ transduction is thought to occur when transcription reads into the genomic flanking sequence downstream of the L1, such that is incorporated into the L1 transcript and hence the ribonucleoprotein particle (RNP) retrotransposition intermediate. Ultimately this incorporation results in the movement of the 3´ flanking sequence to another genomic location. It has been suggested that a possible reason for the imprecision of formation of the 3´end of the L1 transcript is due to a weak polyadenylation signal which on occasions can be bypassed in favor of a stronger polyadenylation signal downstream of the L1 3´ end, in the flanking genomic sequence (Moran *et al*., 1996 and Boeke *et al*., 1999). The L1 3´ transduction process has also been observed in cases where L1 elements have

integrated into genes: insertions in *APC* (Miki *et al*., 1992), Dystrophin (Holmes *et al*., 1994), *CYBB* (Meischl, *et al*., 2000), *RP2* (Schwahn, 1998) and *CHM* (van den Hurk 2003; 2007) all contain non-L1 inserted sequences 3′ of the canonical 3′ terminus. The 3′ transduction of genomic sequences by L1s is a common event and it has been estimated that this process is responsible for a 6% increase in the size of the human genome (Pickeral *et al*., 2000).

Besides expanding the human genome, L1-mediated 3′ sequence transduction can be an important source of genome diversification, by causing exon shuffling and potentially the evolution of new genes (figure 4.1) (Moran *et al*., 1999, Eickbush 1999 and Boeke *et al*., 1999). The exon shuffling ability of L1 was demonstrated, *in vivo,* by Moran *et al* (1999) using a retrotransposition assay involving a reporter cassette containing a splice acceptor site downstream from the polyadenylation signal of an intact L1. This study showed that L1s can retrotranspose into transcriptionally active regions and co-mobilise the reporter cassette, which can be expressed after splicing. Therefore, L1 3′ sequence transduction can transduce exons, promoters and enhancers and potentially lead to the production of new genes and changes in gene expression of existing genes. Moran *et al.,* (1999) also noted that the magnitude of L1-mediated transduction depends on the number of active L1s in the genome, their genomic location and also their rate retrotransposition.

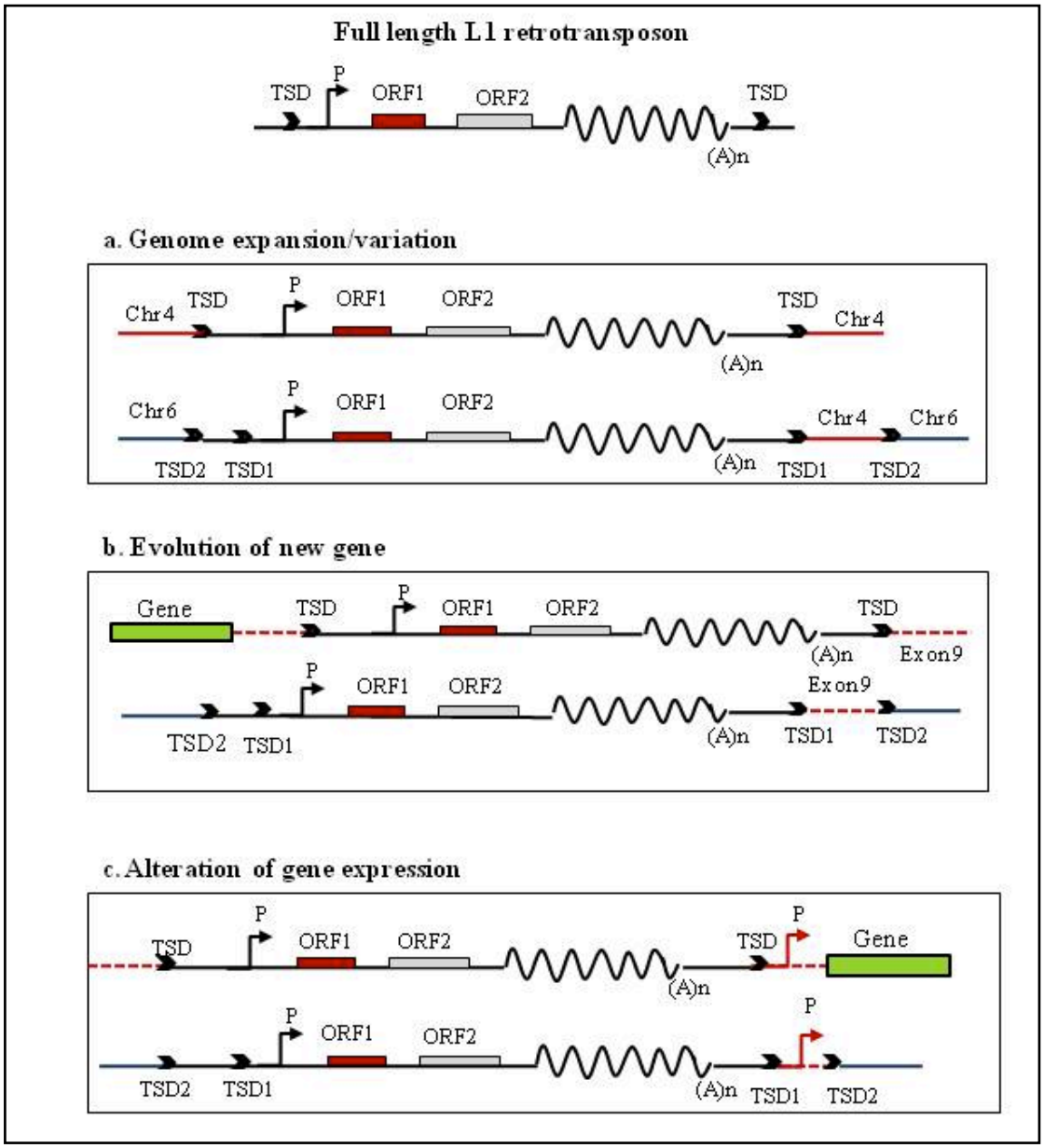

Figure 4.1 L1 3′ transduction alters the genome in different ways: **a.** 3′ transduced- sequence causes genome expansion, as well as structural variation. **b.** New genes can be generated through exon transduction (exon shuffling). **c.** L1 mobility can alter gene expansion by transducing promoters or enhancers to a different locus. Adapted from Goodier *et al*., 2002 and Ejima *et al*., 2003.

As explained earlier, this project aimed to investigate the activity of L1 retrotransposons during early human development, by detecting *de novo* endogenous L1 retrotransposition. It has been demonstrated that young lineages of L1 elements show high sequence similarity and hence the process of looking for putative novel L1 insertions is technically challenging. About 9% of hot L1s

contain 3´transduced sequences (Goodier *et al*., 2000), and their transductions enable us to distinguish even very closely related sequences. This chapter described the results of investigating three highly active L1 lineages using the Transduction Specific - Amplification Typing of L1 Active Subfamilies (TS-ATLAS) technique, as described by Macfarlane *et al*. 2011, *In preparation.* TS-ATLAS enables specific amplification of L1 insertions carrying a particular lineage-specific sequence tag. One of the advantages of using this technique for this experiment is the much lower complexity of the PCR amplicon libraries generated from linkered genomic DNA as compared to the ATLAS technique (Badge *et al*., 2003) (Chapter three). This results from the fact that only amplicons carrying the transduction lineage specific sequence amplify exponentially. Therefore, it should be easier to detect single molecule / *de novo* L1 insertion (low copy number) events by screening for insertions generated by these three very active, young, L1 transduction lineages.

In TS-ATLAS common 3´ transduced sequences are used to identify any related insertions, particularly siblings of the "founder" element and to establish likely progenitors of particular insertions. This technique has also been used *in silico* to find the likely progenitor of disease causing insertions. In this chapter the TS-ATLAS method was applied to three highly active L1 sub-families: AC002980, LRE3, and RP. Of these the LRE3 and RP lineages are disease causing and contain very active members (Brouha *et al*., 2002 and Kimberland 1999).

The L1 AC002980 (chr Xp22.2) comes from the youngest Ta-1d group, L1PA1 subfamily (Brouha *et al*., 2003). It was demonstrated that it is one of the hottest L1 elements in a cell culture based retrotransposition assay (Brouha *et al*., 2003). The likely progenitor of this lineage, AL118519, has given rise to three sub-lineages with long, intermediate and short transductions through the use of alternative poly adenylation sites. Of these the AC002980 L1 belongs to the short transduction subdividision of this subfamily (Macfarlane *et al.,* 2011, *In preparation*).

Investigating the 3′ transduction sequence of $L1_{CYBB}$, which caused a case of chronic granulatomatous disease by inserting into the *CYBB* gene $L1_{CYBB,}$ (figure 4.2) revealed that the L1 LRE3 is its potential progenitor and has clearly been recently retrotranspositionally active in the human genome (Brouha *et al*., 2002).

LRE3 (L1 Retrotransposable Element-3) is the most active human L1 in cell culture assay (Brouha *et al*., 2002). LRE3 and RP differ by one silent change in ORF1, four changes in noncoding regions and one amino acid (Thr to Asn) change in ORF2, and it has been demonstrated that LRE3 is 50% more active than RP (Brouha *et al*., 2002). L1RP was first identified in a patient with X-lined retinitis primentosa. The L1 retrotransposition occurred in an intron of a novel gene, which is responsible for progressive retinal degeneration (Schwahn *et al*., 1998).

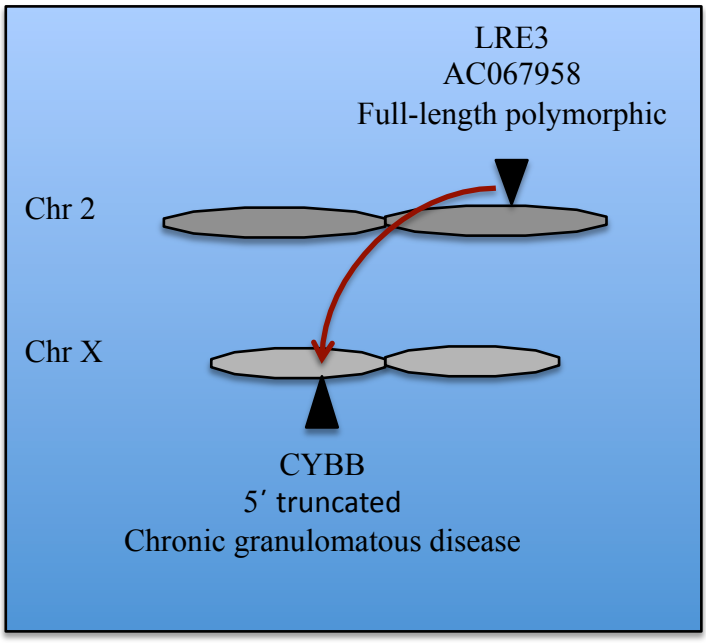

Figure 4.2. The progenitor locus, LRE3 caused a case of Chronic Granulatomatous Disease (CGD)

TS-ATLAS is based on similar principles to ATLAS (Badge *et al*., 2003) and relies on the use of transduction-specific PCR primers to selectively amplify L1 loci containing transduced sequence from oligonucleotide linkered genomic libraries (Macfarlane e*t al.,* 2011 *In preparation*). To verify a novel / *de novo* insertion, following the sequencing of TS-ATLAS products, an independent L1 locus-specific PCR amplification and sequencing experiment is required to verify the L1 structure, its 3′ transduction sequence, any target site duplications (TSDs), and the segregation and dimorphism status of each prospective L1 locus. In this chapter we have applied TS-ATLAS specific for three highly active L1 loci to

human embryonal cells and malignancy derived cell lines, to investigate the activity of L1 retrotransposition in these three young L1 loci.

4.2 Results

4.2.1 L1 transduction family: AC002980

As mentioned in the introduction section L1-AC002980 is a highly active L1 lineage and carries a 174 bp 3´transduction (Brouha *et al*., 2003). Primers were designed for the L1-AC002980 transduced sequence and the AC002080 specific TS-ATLAS method was applied to embryonal cell lines to examine the diversity of this very active L1 subfamily in these cell lines. Both *Nla*III and *Ase*I libraries were constructed for each clonal cell line. The clonal cell lines used for this experiment were: two independently derived human Embryonic Stem Cell (hESC) clonal lines, H1-derived clonal lines (n=5, Moran J), H9 progenitor and its clonal lines (n=20, Garcia-Perez J), and HeLa-AJ clonal lines (n=5) and HeLa-RP clonal lines (n=2). In addition, a small panel (n=4) of CEPH pedigree lymphoblastoid cell line DNAs were used as controls.

The result of L1-AC002980 TS-ATLAS on hESC (H1 progenitor) clonal lines revealed three L1 insertions: L1-AC004740, L1-AC048382, and L1-AP001029. All these three L1 AC002980 related elements are novel full-length L1 retrotransposons that are absent from the human genome reference sequence. All of these three insertions are known to be polymorphic as they have been previously isolated from the sperm and blood genomic DNA of a panel of anonymous Caucasian healthy volunteers (Macfarlane *et al*., unpublished data). Their allele frequencies were determined by genotyping 192 CEPH DNAs for each of these insertions (summarised in table 4.1). All of these three insertions were full length with allele frequencies of 0.016 to 0.485.

These insertions were further validated by genotyping the hESC clonal lines for all the three loci, which were recovered from the TS-ATLAS display gel (above). The genotyping results revealed that the L1-AC002980 and L1-AC004740 were present in all the genotyped clonal lines. However, the L1-AC048382 and L1-AP001029 loci appeared to be dimorphic (presence / absence) in the clonal lines derived from

the same progenitor (figure 4.3). Since the progenitor hESC (H1) DNA was not available, it was not clear whether this observation was due to somatic mosaicism in the H1 progenitor or other possibilities (explored in more detail in the discussion section of this chapter). The same assay was applied to hESC (H9 progenitor) clonal lines (n=20), which is presented in Figure 4.3. To verify the transduced sequence, all the variable bands amongst the hESC clonal lines in the gel were recovered and sequenced. The result revealed that the majority of observed variability amongst the clonal lines was from the older L1 insertions such as L1PA2. Due to mutated primers site at the TS-ATLAS priming sites, they were able to amplify sporadically across the clonal line. Indeed, sequence analysis showed these insertions to be very similar to the L1 primers used, but their sporadic appearance is not easily explained. The remaining sequences were belonged to the AC002980 subfamilies including the L1-AC004740, L1-AC048382 and L1-AL118519 (figure 4.3). As mentioned earlier, it has been demonstrated by Macfarlane *et al.* (unpublished data, 2009) that the L1-AL118519 insertion is the progenitor of the three L1 sub-lineages including the L1-AC002980. One of characterised sequences belonged to the L1Hs family and did not have a poly-A tail present in the HGR. Further characterisation showed that it was AC048382, an AC002980 lineage element. To confirm the observed variation from the display gel, all the clonal lines were genotyped for this locus. The genotyping result also confirmed that the AC048382 was only present in the progenitor (H9), hESC clonal lines 3, 14 and the positive control (figure 4.3).

The L1-AC002980-specific TS-ATLAS was also applied to seven HeLa clonal lines. The HeLa TS-ATLAS revealed that the L1-AC010749 and L1-AP001029 as well as their progenitor L1 AC002980 were present in HeLa clonal lines (figure 4.4). Although L1- AC002980 was present in all the clonal lines, L1-AP001029 and L1-AC010749 showed to be variably present amongst the clonal lines (figure 4.4). Further genotyping of HeLa clonal lines also confirmed that these two L1 loci are polymorphic. HeLa cells all originated from the same progenitor (Henrietta Lacks) so they are expected to have all been identical at the start, but have had many generations in culture to diverge and therefore may show polymorphism with respect to young L1 loci due to loss of heterozygous chromosomal regions. Since the original progenitor of the HeLa cell lines are not available it is not clear whether

this variability is due to genomic instability of the clonal lines and their subsequent loss of the insertion.

Accession/ L1	Library	Location	TSDs (5′-3′)	§Activity	†Allele Freq
AL118519	*Vsp*I	6q13	CAAAACAAAACAAAGCAAAC	-	FIXED
AC002980	*Nla*III	Xp22.2	AAAAAAAATCACCA	132%	0.485
AC004740*	*Nla*III	7p21.3	AACAATATGTA	-	0.064
AC048382	*Nla*III	15q25.2	AAGATGTAAGTAGAAA	-	0.034
AP001029*	*Nla*III	18p11.21	AAGAAAATCCT	-	0.016
AC010749	*Nla*III	7p21.1	-	173%	0.053

Table 4.1. Loci recovered from hESC and HeLa clonal lines using AC002980 specific TS-ATLAS *Reported in (Beck *et al*., 2010), TSDs determined in (Macfarlane *et al*., 2011 *In preparation*), †Allele frequency determined in (Myers *et al.,* 2002). Loci are described as polymorphic if they were present and absent in more than one individual in the blood donor panel (n=9). Allele frequencies are based upon the genotyping of 129 unrelated CEPH DNAs. §Activity indicates the rate of retrotransposition in a cell culture-based retrotransposition assay, relative to the reference element L1.3 (L19088) ND = not determined (Brouha *et al*., 2002; Brouha *et al*., 2003; Beck *et al*., 2010).

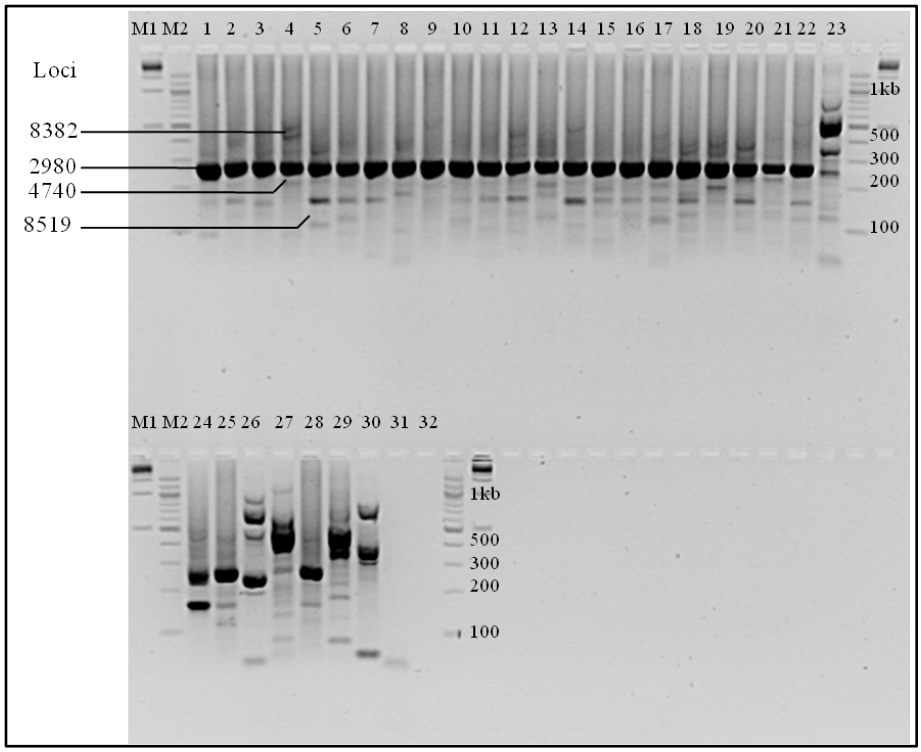

Figure 4.3. A. Representative TS-ATLAS display gel showing the results of applying the AC002980-specific (*Ase*I) assay in H9-derived hESC clonal lines. Lane 1: human fibroblast DNA, lane 2: H9-hESC progenitor Lanes 2-22: H9 clonal cell lines. Control reactions for setting up the library in lanes 23-30. M1 and M2 = molecular weight marker. As demonstrated on the gel the L1- AC002980 progenitor present in all the clonal lines as well as its other subfamilies: AL118519, AC048382, and AC004740.

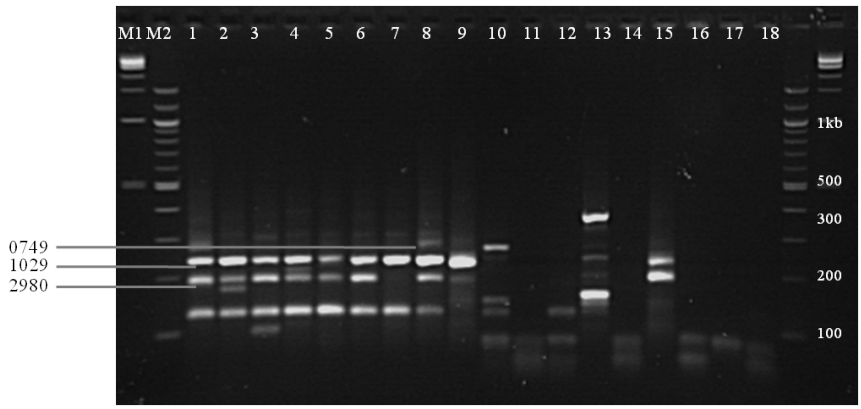

Figure 4.4. A. Representative TS-ATLAS display gel showing the results of applying the AC002980-specific (*Nla*III) assay to HeLa clonal lines. Lanes 1-5: AJ HeLa clonal lines, lane 6 and 7: RP HeLa clonal lines. Control reactions for setting up the library in lanes 8-15, Lanes 16-18, -ve PCR controls. M1 and M2, molecular weight markers. It demonstrates the L1- AC002980 and its subfamilies: AC004740 and AP001029.

4.2.2 L1 transduction family: LRE3

The L1-AC067958-specific TS-ATLAS protocol (Macfarlane *et al*., 2011*, In preparation*) was used to examine the diversity and segregation pattern of this active subfamily in human embryonal cell lines (H1 and H9 progenitor) and HeLa clonal lines. The results revealed that two subfamilies of L1-AC067958 were present in hESC and HeLa clonal lines. Both of the insertions: L1-BX927359 and L1-AL592182 were novel L1s (absent from the human genome reference sequence). L1-AC067958-specific TS-ATLAS for hESC clonal lines (H9) is presented in figure 4.5. Genotyping showed that these two insertions (L1-BX927359 and L1-AL592182) were present at very low allele frequencies of 0.0-0.015 (table 4.2). To confirm the presence of the observed LRE3 subfamilies on a display gel, all the clonal lines were genotyped for each of the insertions. The genotyping results also confirmed that these insertions are stable and present in all the samples. Both of these LRE3 subfamilies are full length and thus potentially active (Macfarlane *et al*., 2011*, In preparation*).

Accession L1	Library	Location	TSDs (5´-3´)	Activity	Allele Freq
AC067958	*Mse*I	2q24.1	GAAAGAAAGAAAGAA	50%	0.315*
BX927359	*Mse*I	14q32.23	AAAATGAAATAAAAT	ND	0.015
AL592182	*Mse*I	1p33	AGAAAACAACAGAGGGG	ND	0.0

Table 4.2. Loci recovered from hESC and HeLa clonal lines using LRE3 specific TS-ATLAS *Reported in (Brouha *et al*., 2002), TSDs determined in (Macfarlane *et al*., 2011*, In preparation*), ND = not determined (Brouha *et al*., 2002; Brouha *et al*., 2003; Beck *et al*., 2010).

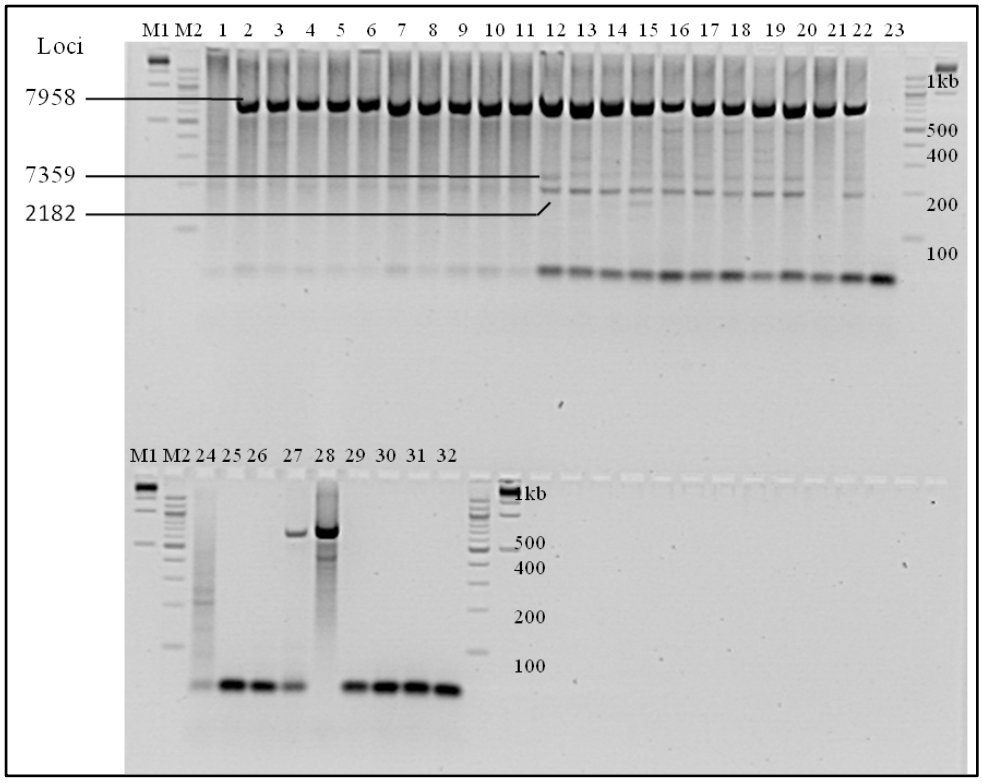

Figure 4.5. A. Representative TS-ATLAS display gel showing the results of applying the LRE3-specific (*Ase*I) assay in H9-derived hESC clonal lines. Lanes 1 HFB DNA, 1- H9-hESC progenitor 2-22- hESC clonal lines. Control reactions for setting up the library are in lanes 23-30. M1 and M2 = molecular weight marker. As demonstrated on the gel the L1-AC067958 progenitor present in all the clonal lines as well as its two subfamilies L1-BX927359 and L1-AL592182.

4.2.3 L1 transduction family: RP

The disease-causing insertion in the *RP2* gene (Kimberland *et al*., 1999) carries a short 11 nucleotides transduction. The RP-specific TS-ATLAS protocol (Macfarlane *et al*., 2011*, in preparation*) was applied to human embryonic stem cell lines (H1 and H9 progenitor) and HeLa clonal lines. Screening of the human embryonal cell lines (H1- and H9-derived hESC clonal lines and HeLa clonal lines) was performed using the RP-specific TS-ATLAS. The result showed that there were two polymorphic RP-related elements, L1-AL050308 and L1-AC005939, in both hESC clonal lines and HeLa clonal lines. Figure 4.7 shows the RP-specific TS-ATLAS on hESC clonal lines and its two related subfamilies. The L1-AL050308 insertion is not present in the HGR and has also previously been isolated from blood donor DNA samples by Macfarlane *et al.* (2011, *in preparation*). However, the latter RP-related, L1-AC005939 is present in the HGR. AC005939 is a sibling element of $L1_{RP}$ having the same transduction

sequence but unique TSDs (table 4.3). A variable size band (AL365508) on a display gel of H9-derived hESC clonal lines was recovered from clonal line six and was also present in some of the clonal lines with the exception of the progenitor and clonal lines two and three (Figure 4.7). L1Hs sequence is present in the sequence and consequently the mapping of the sequence to the HGR did not indicate the presence of any poly-A tail. However, by aligning the transduced sequence of AL365508 with the rest of the RP-related transduction sequence it was revealed that this band was probably a missprimed insertion that appeared on the display gel (figure 4.6).

Table 4.3. Loci recovered from hESC and HeLa clonal lines using RP specific TS-ATLAS TSDs determined in (Macfarlane *et al*., unpublished data), ND = not determined.

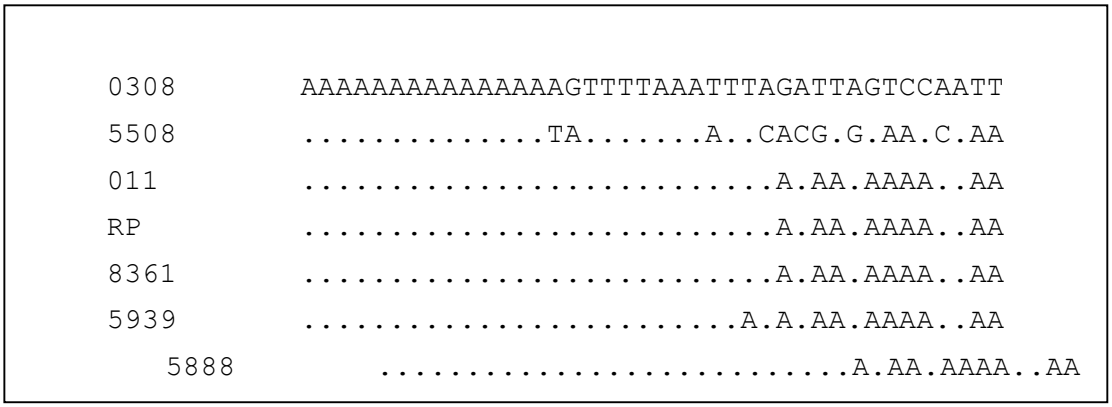

Figure 4.6. Alignment of all known 011 transductions, the 5508 loci transduction does not align with the rest of the RP related transductions.

Accession/ L1	Library	Location	TSDs (5´-3´)	Activity	Allele Freq
AL050308	*Vsp*I	Xq27.2	AAAAAGTTTTAAATTT	ND	0.149
AC005939	*Vsp*I	17q24.3	AAGATTTTGTG	ND	-

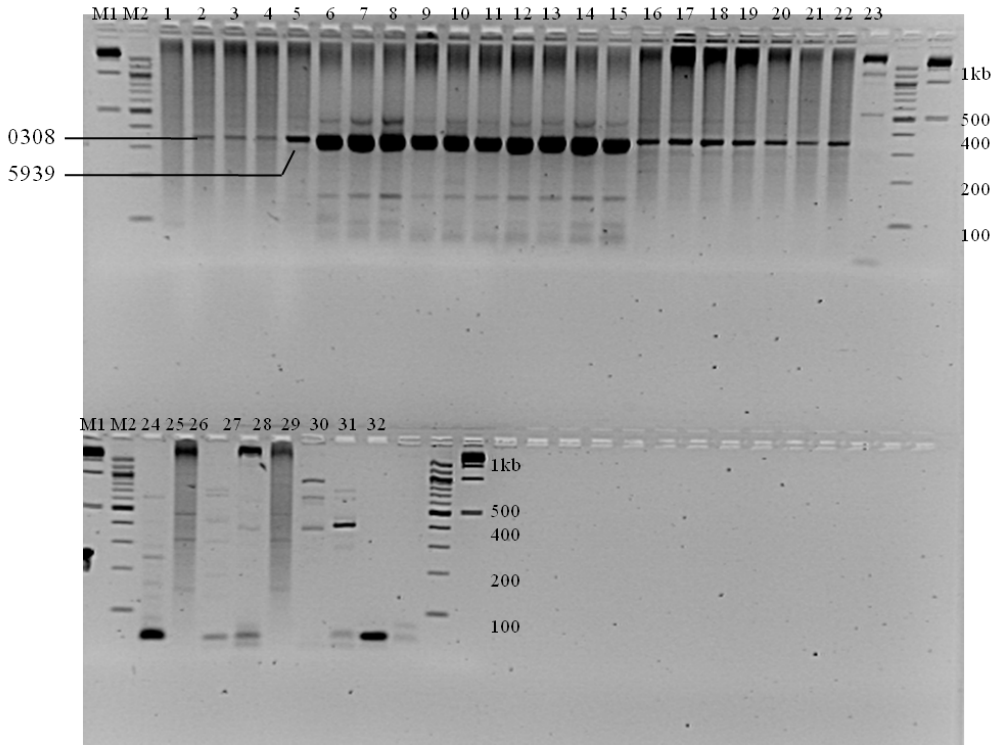

Figure 4.7. A. Representative TS-ATLAS display gel showing the results of applying the LRE3-specific (*Ase*I) assay in H9-derived hESC clonal lines. Lane 1: human fibroblast DNA, lane 2: H9 progenitor lanes: 2-22 hESC clonal cell lines. Control reactions for setting up the library in lanes 23-30. M1 and M2 = molecular weight marker. The LRE3 and its subfamily AC005939 demonstrated on the gel.

4.3 Discussion

L1 retrotransposons are the only active autonomous retroelements in the human genome. However, it has been demonstrated using *in vitro* retrotransposition assays that different alleles of active L1s show variation in their activity (Brouha *et al*., 2002, Seleme *et al*., 2006). Therefore, it is important to find out which lineages of L1 are active and make the biggest contribution to the mass of L1 retrotransposons in the genome.

Due to the vertical evolution of L1 retrotransposons the active lineages of L1 are replaced by younger newly evolved elements (Boissinot *et al*., 1999). As a result, the young and active L1 retrotransposon sequences are highly similar. This sequence similarity facilitates the genome-wide targeting of hot L1 retrotransposons in order to discover the rate of L1 retrotransposition in these cell lines.

Several *in vivo* and *in vitro* techniques have been developed to isolate active L1 retrotransposons such as the ATLAS technique (Badge *et al*., 2003, Buzdin *et al*., 2003, Roy *et al*., 1999; reviewed in Beck *et al*., 2011). However, using these genome-wide approaches to find *de novo* L1 retrotransposition can be challenging due to the high sequence similarity of the active L1s and the low copy number of *de novo* insertions.

In this chapter the TS-ATLAS technique (Macfarlane *et al*., unpublished data) was used to investigate the rate of L1 retrotransposition in human embryonal cells. This technique allows tracking of particular lineages of L1 in a genome-wide context and can rapidly identify young and novel offspring insertions. The significance of using this method as a tool to investigate the activity of L1s in human embryonal cells is that by selectively amplifying only related active L1 elements it is possible to simultaneously reduce the complexity of the linkered library and increase the sensitivity of the amplification reaction. Consequently, it increases the possibility of amplifying fragments present at less than constitutional levels. As a result using this technique should be a good approach to address the question of the rate of endogenous human L1 retrotransposition - at least for particular lineages.

Since this technique utilises frequently cutting restriction enzymes, L1s within a much greater proportion of the genome are accessible. Whole genome *in silico* restriction

analyses showed that only 10% of the sequenced human genome is within suppression PCR range (<~1000bp) when libraries are constructed with *Vsp*I, but 80% is accessible using *Nla*III (Dr. Richard Badge, *pers. comm*.). The small number of co-amplifying fragments derived from transduction-specific suppression PCR primers makes fractionation and isolation of novel loci by agarose gel electrophoresis feasible.

All three L1 subfamilies (AC002980, LRE3 and RP), which had been investigated for the TS-ATLAS, belong to highly active ("hot") L1 groups based on their disease-causing properties (Brouha *et al*., 2002 and Kimberland 1999). As determined by cell culture-based retrotransposition assays, both L1 LRE3 and L1 AC002980 were consecutively the most active known L1 retrotransposons (Brouha *et al*., 2002; Seleme *et al*., 2006) and they both carry extensive 3′ transductions (Goodier *et al.,* 2000; Brouha 2002).

The results presented revealed the diversity of each of these lineages amongst the human embryonal clonal lines tested.

The result of screening about 37000 molecules of DNA (hESC and HeLa clonal lines DNA) using the AC002980-specific TS-ATLAS assay showed that 60% of the reported AC002980-related families (Macfarlane *et al*., unpublished data) are present in the hESC and HeLa clonal lines. All of these L1s are full-length insertions. Two of these insertions, AC002980 and AL118519 (putative progenitor of the AC002980 lineage), are present in the HGR. The remaining three insertions (AC004740, AC048382 and AP001029) are novel insertions (absent from the HGR) that were previously discovered in human germline DNA by Macfarlane *et al*. 2011, *in preparation*)

Using approximately the same number of molecules of DNA as above, the activity of the LRE3-related transduction family was investigated. The results revealed that 50% of the reported LRE3 families (as reported by Brouha *et al*., 2002 and Macfarlane *et al*., unpublished data) are present in human embryonal genomic DNA and all of these insertions are full length and likely to be active. The AC067958 insertion has been reported through its mutagenic effect on the *CYBB* gene (Brouha *et al*., 2002). The BX927359 and AL592182 insertions are novel (absent from the HGR) and have also been found in germline DNA (Macfarlane *et al*., 2011, *in preparation*).

The result of RP-specific TS-ATLAS on the same genomic DNA samples showed two full length insertions: AL050308 and AC005939 - the former element is a novel insertion which has been discovered in blood DNA (Macfarlane *et al*., 2011, *in preparation*).

Lineage-specific TS-ATLAS screening on hESC and HeLa clonal lines showed some level of L1 complement variation amongst the clonal lines. As has been presented in this chapter, the TS-ATLAS display gel for hESC and HeLa cell lines showed size variability amongst the clonal cell lines.

One of the possible explanations for the observed variation is the fact that human embryonic stem cells and induced pluripotent stem cells that have cultured for a period of time are susceptible to genomic instability and which decreases their reliability for therapeutic purposes (Elliott *et al*., 2010). The only way to monitor genomic stability is by karyotype analysis. Because of the low resolution of this technique, only large-scale chromosomal abnormalities such as aneuploidies can be screened in this way (Elliott *et al*., 2010). The karyotypic analysis of hESC (H9-derived) clonal lines did not show any aneuploidy and they were reported as being karyotypically very stable (Garcia-Perez, *pers. comm*.). Therefore, this observation is likely due to other factors than chromosomal instability.

Another possible explanation could be a low level of somatic mosaicism in the progenitor. Such mosaicism could be common if endogenous elements retrotranspose in early embryogenesis as frequently as active L1 constructs in transgenic mice and rats (Kano *et al*., 2009). This can result from stochastic amplification amongst the clonal lines and can be confirmed by pool TS-ATLAS analysis on the progenitor DNA to discover the level of mosaicism for each locus.

Another possible factor for this observation might be genome instability amongst clonally derived cells as this can cause loss or gain of genomic DNA at certain loci. It has previously been demonstrated that in HeLa clonal lines genomic DNA apparently becomes unstable indicated by the loss of several L1 loci (Badge *et al*., unpublished data). To confirm whether gain or loss of genomic DNA is responsible for such a stochastic observation among the single-cell-derived clonal lines, the presence and absence of genomic DNA at each specific locus could be characterised by using array-CGH or other such techniques.

In conclusion, by utilising the distinctive sequences that some L1s mobilise it is possible to directly capture L1 sequences in embryonal-derived genomic DNA samples to capture ongoing endogenous L1 retrotransposition events in a lineage-specific manner.

Chapter 5

Methylation instability of the L1 promoter in human embryonal cells

5.1 Introduction

The human genome harbours around 3,000 copies of full or nearly full-length L1 elements (Kazazian, 2004) of which approximately 80-100 are retrotranspositionally active (Brouha *et al.*, 2003). Despite the presence of these active elements the rate of L1 insertion, as ascertained by screening human mutation databases, is very low: only ~ 0.02% of catalogued mutations are caused by L1s (Chen *et al.,* 2006). Thus in humans L1 retrotransposition is apparently repressed, perhaps as a defence against the mutagenic effects of *de novo* L1 insertions.

A possible mechanism by which the activity of the many potentially active human L1s could be suppressed is methylation of cytosine bases in their promoters, some of which are critical for promoter activity (Hata and Sakaki 1997). Interestingly, human embryonic stem cells, HeLa and NTera2D1 cell lines are characterised by global hypomethylation (Hoffmann and Schulz 2005), which could potentially activate L1 expression and induce genomic instability via retrotransposition events. It is an open question as to whether this genome-wide hypomethylation is reflected in the methylation status of potentially active L1 promoters. Therefore in this chapter we

report the results of applying a number of different techniques to study the genome-wide and locus-specific methylation of active L1 retrotransposons in these cell lines.

5.1.1 L1 promoter

The 5′ UTR of full length L1s harbours a strong internal sense RNA polymerase II promoter, which is not dependent on flanking sequences for its activity, as well as an antisense promoter that together give rise to overlapping sense and antisense transcripts (Swergold, 1990, Ostertag and Kazazian, 2001a). Macia *et al.* (2010) demonstrated that both of these promoters are robustly conserved throughout L1 evolution, at least as far back as L1PA10 (~ 60 Mya, Walser *et al*., 2008). The first 100bp of the 5′ UTR are recognised as the most critical region for initiation of L1 transcription, assisted to a lesser extent by additional sequences within the first 668 bp (Swergold, 1990). It has been shown that L1 transcription is not restricted to the nucleotide at position +1 and can start upstream or downstream of the start site, ranging between positions -9 to +4. As a result, some sequence variation, presumably arising from this variable initiation, is observed in the 5′ UTR of L1s (Lavie *et al.,* 2004).

The L1 sense promoter (SP) contains several transcription factors binding sites including: RUNX3, SRY family transcription factors and YY-1 (Yang *et al.,* 2003, Tchenio *et al.,* 2000, Becker *et al.,* 1993). The YY1 transcription factor has only a modest effect on the level of L1 expression but it is important for accurate transcription initiation (Athanikar *et al.* 2004). The 5′ UTR of the L1 also contains an antisense promoter (ASP) located around +500 nucleotides, which can affect the transcription of upstream genes (Speek, 2001). A schematic diagram of the L1 promoter and its transcription binding sites is shown in Figure 5.1.

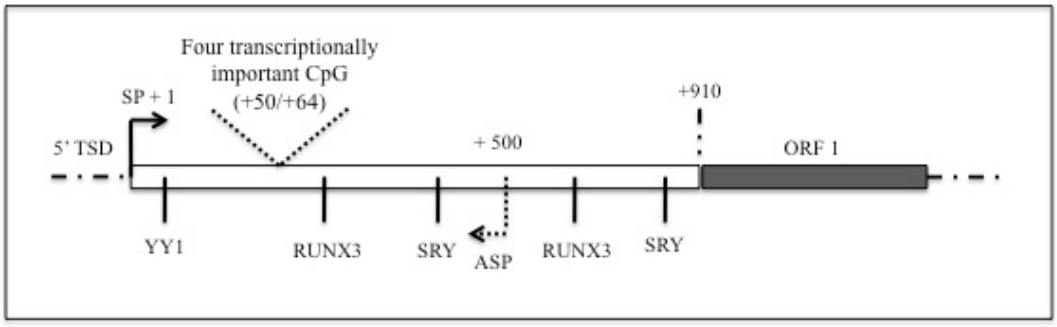

Figure 5.1 Schematic diagram of the L1 promoter and all the factors binding sites. L1 Sense promoter (SP): +1, Antisense Promoter (ASP): +500, four transcriptionally important CpG sites: +52, +58, +61 and +70. Coordinates refer to the reference sequence: L19088 (L1.3).

5.1.2 Methylation of cytosine residues in CpG dinucleotides

Methylation involves the addition of a methyl-group (CH_3-) to the cytosine moiety, which will change it into 5-methylcytosine. This happens predominantly in the cytosine of CpG dinucleotides in mammals (Goll and Bestor, 2005). The modified base 5-methylcytosine (m5C) is present in the DNA of all vertebrates, flowering plants and invertebrates to varying degrees. However, the biological functions of methylation are fundamentally different in prokaryotes and eukaryotes. Cytosine methylation is mediated by a conserved group of proteins called DNA methyltransferases (Dnmts). There are three groups of DNA methyltransferases: Dnmt1, Dnmt2 and Dnmt3. Dnmt1 maintains the methylation status of the cytosine residue, *i.e.* it methylates hemi-methylated CpGs sites resulting from DNA replication to ensure that methylation patterns will be preserved after each cell division (Bestor *et al*., 1998, Stein *et al.,* 1988). Dnmt2 is the most strongly conserved, most widely distributed, but a Dnmt2 knock out study in mouse ES cells indicated that this enzyme is not essential for the cell survival *in vitro* (Okano *et al*., 1998) and therefore, the role of this enzyme is not well defined. The mammalian genome encodes two functional cytosine methylatransferases of the Dnmt3 family. Knock out studies in mice demonstrated that Dnmt3A and Dnmt3B are essential. This highlights the importance of CpG methylation (Li *et al.,* 1992, Okano *et al.,* 1999). Dnmt3A and Dnmt3B are responsible for *de novo* methylation, *i.e.* both transfer methyl groups to

hemimethylated and unmethylated substrates at equal rates and without any sequence specificity beyond CpG dinucleotides (Okano *et al.,* 1998). Another homologous gene in the Dnmt3 group, Dnmt3L, lacks the methyltransferase activity and functions as a regulatory factor, influencing DNA methylation especially in germ cells. It interacts with histone deacetylases and activates *de novo* methylation enzymes and suppresses L1 expression (Aapola *et al.,* 2002). It is demonstrated that the loss of Dnmt3L does not interfere with oogenesis or early development in heterozygous embryos derived from homozygous Dnmt3L mutant oocytes, but Dnmt3L-deficient male germ cells display meitotic catastrophe with non-homologous synapsis and accumulation of highly abnormal synaptonemal complexes (Bourc'his and Bestor, 2004).

Due to the spontaneous hydrolytic deamination, which results in 5-methylcytosine being converted to thymine (Bird, 1986), CpG dinucleotides are under-represented in mammalian genomes compared to their expected frequency based on the (G+C) fraction of the genome (Swartz *et al.,* 1962). This conversion process is quite frequent and is apparently responsible for 35% of disease-related point mutations in humans (Cooper and Youssoufian, 1988). CpG methylation is involved in a variety of biological processes, such as gene transcription, defence against retroelements, X chromosome inactivation, genomic imprinting and carcinogenesis.

Methylated promoters are usually inactive, but the factors that mediate this suppression are unknown. It has been proposed that methylation can repress gene expression by two mechanisms: the inhibition of transcription factor binding due to chemical modification of their recognition sequence (Watt and Molloy 1988), and the attraction of the methyl-CpG binding domain (MBD) protein family (Hendrich and Bird, 1998). MBDs can associate with histone deacetylases and remodel chromatin into a transcription-inhibiting form (Nan *et al.,* 1998). Similar to genes, methylation has an inhibitory effect on L1 retrotransposons. It has shown that the L1 promoter has a methylcytosine binding protein 2 (MeCP2) binding site, which suppress methylated promoters (Nan *et al.,* 1998). *In vitro* expression experiments showed that four CpGs in the L1 promoter are critical for transcription: +52, +58, +61, +70 and successful inhibition requires all four CpGs to be methylated (Hata and Sakaki, 1997). Since most 5-methylcytosines (90%) reside within transposable elements it has been proposed that DNA methylation is a defensive mechanism against transposable elements (Yoder *et al.*, 1997).

5.1.3 Cytosine methylation in host defence and genome instability

The majority of cytosine methylation in plants and mammals resides in repetitive elements and a large proportion of this lies in retrotransposons, which constitute more than 42% of the human genome. Transposons can only proliferate in genomes where the fitness of transposons is greater than that of the host. Therefore, host defence mechanisms are under selective pressure to suppress these elements (Bestor, 2003), and DNA methylation is primarily a mechanism of transposon suppression in the genome. Also, in somatic cells L1 promoters are generally hypermethylated, but in malignancy-derived cells, the global hypomethylation of CpG dinucleotides is correlated with L1 activity. This correlation was revealed by the recent identification of several *de novo* L1 insertions in a cohort of lung tumours (Iskow *et al*., 2010) with an enrichment of insertions being observed in tumours showing significant genomic hypomethylation.

As previously mentioned, a variety of studies have suggested that *de novo* L1 retrotransposition is more likely to occur in germ cells and / or during early embryonic development (Garcia-Perez *et al*., 2007 and Van den Hurk *et al*., 2007), where a pair of global hypomethylation events occur at the genome reprogramming stages. Although it has been frequently suggested that methylation of CpGs has a regulatory role, especially in suppressing repetitive elements, there is evidence against this hypothesis (Walsh and Bestor, 1999), such as the somatic inheritance of genomic methylation patterns in mammals (Riggs, 1975). Therefore, chromatin modifications such as DNA methylation could be a consequence of active transcription rather than being an intiating factor, but this remains to be fully elucidated.

HESCs and human embryonic carcinoma such as NTera2D1 are good models to study L1 methylation during early human development, as they mimic the pluripotent cells of human embryos (Michiko *et al*., 1984; Peter *et al*., 1984). Up to 20% of the L1s expressed in hESCs belong to potentially retrotranspositionally active subfamilies (Macia *et al*., 2011). Since L1 expression is directly related to L1 methylation, the study of L1 methylation status in hESCs can be a *bona fide* model to represent these elements' activity during the development.

This chapter involves the investigation of global- and locus-specific methylation variability of the L1 promoter in a series of clonally derived human embryonal cell lines (including human embryonic stem cells and NTera2D1). We show that four loci bearing full-length L1 insertions exhibit bidirectional clonal variation in their methylation status in cultured human cells. Also, the methylation status of intact L1 insertion loci is robustly stable in primary somatic DNA. Interestingly, hESC DNA shows a methylation status typical of somatic cells with respect to L1 methylation, at the four studied loci. However, L1 global methylation analysis shows that hESC are significantly hypomethylated and more similar to placental DNA.

5.2 Results

5.2.1 Comparative genome-wide methylation analysis of L1 retrotransposons in human embryonal cells using COBRA (Combined Bisulphite Restriction Analysis)

The genomes of mature sperm and egg in mammals are highly methylated and are comparable in methylation level to somatic cells, although there may be differences in specific patterns (Bestor, 2000). However, in the development of mammalian embryos there are two stages of genome-wide demethylation which each last a few days. It is likely that during these demethylation phases L1 promoters could become active and be able to initiate retrotransposition in the genome. In this experiment the modified COBRA assay (Chalitchagorn *et al*., 2004) was used to evaluate LINE-1 methylation status in the human embryonal cell lines NTera2D1, PA1 and hESC. The global methylation of LINE-1 in embryonal cell lines was compared with two extreme groups; group one contained germ cells (sperm DNA) and somatic cells (blood DNA) where global hypermethylation of LINE-1 is expected, and in group two were the carcinoma cell lines HeLa, SW (620 and 480) where LINE-1 is expected to be globally hypomethylated. As detailed in chapter 2 (Methods and Materials), unmethylated cytosines in DNA samples are converted to uracil by treatment with sodium bisulphite. The modified DNA was subjected to PCR amplification with 5′ UTR LINE-1 specific primers and then digested with *Tas*I and *Taq*I restriction

enzymes that recognise unmethylated and methylated CpG dinucleotides respectively as illustrated in Figure 5.2.

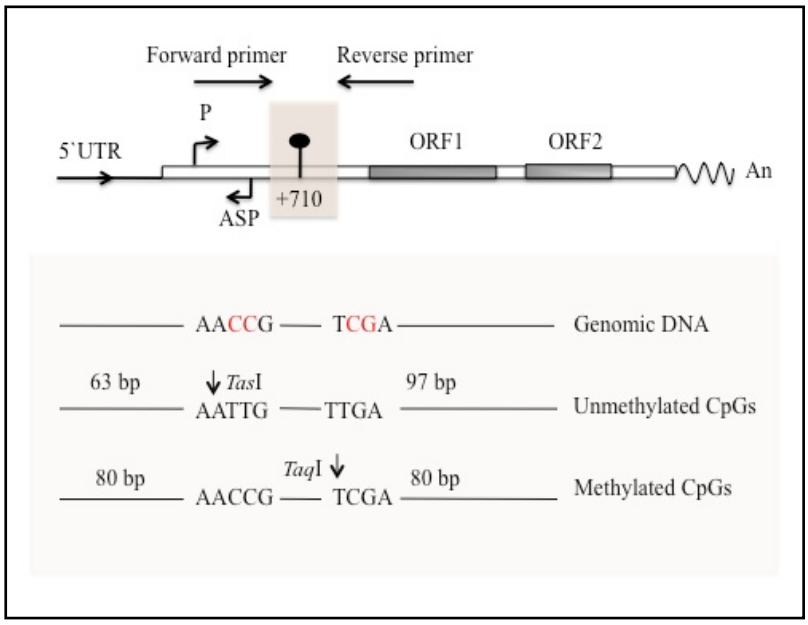

Figure 5.2 L1-COBRA assay, schematic diagram of the L1 showing the relative location of the bisulphite primers. Following the PCR, products were separately digested with *Tas*I (methylation independent site), which produces products of 63 and 97 bp. *Taq*I (methylation dependent site) produces two 80 bp products (adapted from Chalitchagorn *et al*., 2004).

The digested product of *Tas*I produces two fragments of 97 and 63 bp, but this site is only created when the CpG at position +701 is unmethylated and converted to TpG by bisulphite treatment. Conversely *Taq*I (which is insensitive to DNA methylation can only cut its recognition site if the methylated CpG dinucleotides is unconverted to produce two 80bp fragments. An example of a double digest with *Tas*I and *Taq*I on the bisulphate-treated PCR products is shown in figure 5.3. The digested products were further characterised by cloning and sequencing one of the samples (sperm DNA) for all three digested products to check if they had the correct corresponding methylation status. The sequences were mapped correctly to the site of the L1.3 sequence, and they had the correct restriction site. As is demonstrated in figure 5.3, an extra band of 190 bp was also observed in all the samples. This was also cloned and sequenced. The sequence for this band contained mixed signals and is probably degraded DNA; this could have resulted from the damage of the DNA through bisulphate treatment and the restriction enzyme double digest, therefore it is not

possible to establish if the undigested DNA is derived from sequences that are resistant to digestion, or that lack the restriction sites.

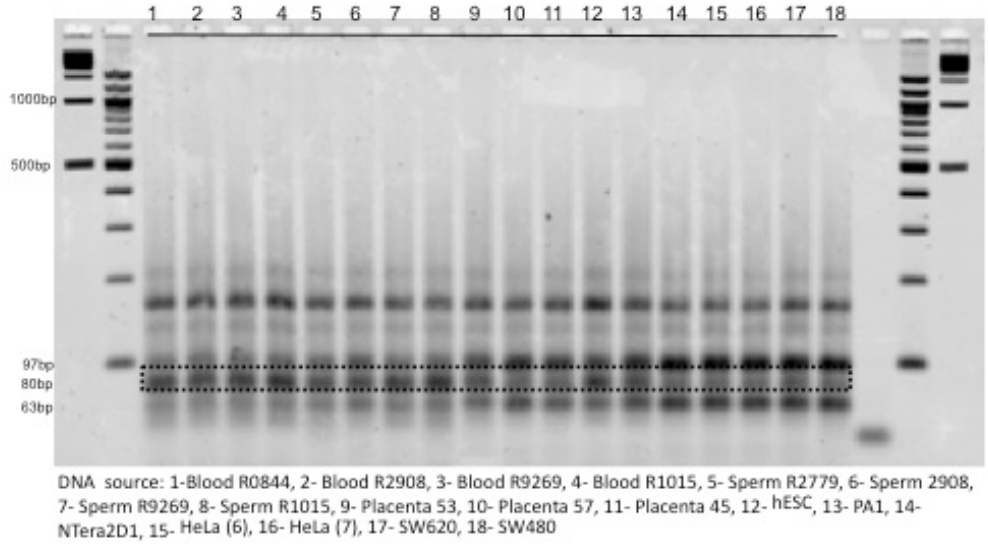

Figure 5.3 COBRA-L1 assay (9): The genome-wide L1Hs methylation status of the various human tissues and cell lines was analysed by using restriction digests of a 160bp L1-specific PCR product amplified from bisulphite-modified DNA. In this assay, an unmethylated CpG dinucleotide is bisulphite modified to produce a site for the restriction enzyme *Tas*I, producing two fragments of 97 and 63 bp. If a nearby CpG dinucleotide is methylated, the unconverted cytosine provides a target site for the restriction enzyme *Taq*I, producing two 80 bp fragments (boxed). To determine the relative genome-wide methylation status of L1Hs, the ratio of the intensity of the methylated bands (80bp) to the sum of the methylated and unmethylated bands was calculated. An example assay is shown above.

The level of global LINE-1 hypomethylation in each sample was calculated by dividing the measured intensity (using the Image J analysis software) of *Tas*I-digested amplicons with the sum of *Tas*I and *Taq*I products. The average hypomethylation intensities (n=10) for each of the samples are presented in table 5.1. The COBRA-L1 analysis on theses samples revealed three significant clusters of global LINE-1 hypomethylation: germ cells and somatic DNA (sperm and blood), malignant cell (HeLa and SWs) and embryonal cell (hESC, PA1 and NTera2D1) DNA, and finally placenta DNA made one cluster. In sperm and blood DNA 80% of L1 had methylated CpG dinucleotides. Also, there was a significant difference in the global methylation status of hESC and malignant cells when compared to somatic and germ cells ($p=0.000251$). Analysis showed that hESC and placental DNA show a distinctive

intermediate level of methylation when compared to other tissues (p=0.00404) and cultured cells (p=0.0095).

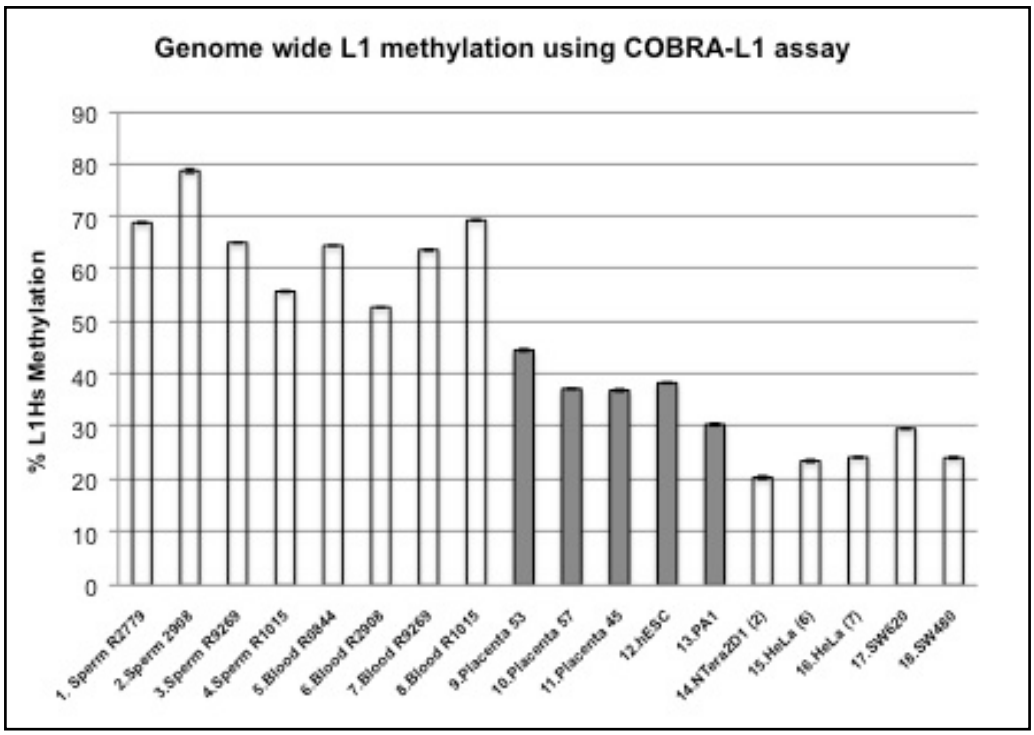

Table 5.1 Bisulphite-modified gDNA samples were amplified in replicate sets (n=10) and digested and quantified by the intensity of EtBr staining (using Image J). The mean methylation levels of each DNA sample are plotted above (±S.E.M). Human somatic and germline tissues show high L1 methylation, cultured human cells show low L1 methylation, and placenta DNA and hESC show intermediate levels. The Exact Wilcoxon Rank Sum Test shows a significant difference between global methylation of tissues and cultured cell DNA (P= 0.000251 at 0.1% level).

5.2.2 L1 locus specific methylation analysis

Methylation of a critical set of four CpG sites (shown in Figure 5.1) can substantially repress the activity of the L1 promoter in reporter assays, and thus modulate L1 expression in cell culture (Hata and Sakaki, 1997). Following the global methylation analysis (above), four loci harbouring full-length L1 insertions were analysed to establish their L1 methylation status. For this study four different L1 insertions were chosen based on their retrotransposition activity in cell culture assays (Moran *et al.*, 1999). We have studied, by direct bisulphite sequencing, the methylation status of L1s

with 0% activity, medium activity of 2.3%, and very active (130% of L1.3) and one locus of unknown activity. Analysing the methylation status of different L1 loci with a variable *in vitro* retrotransposition activity could show if there is any correlation between L1 activity and the amount of promoter methylation. Direct bisulphite L1 sequence methylation analysis on AC005885 locus is summarised in figure 5.4. The L1- AC005885 insertion has an allele frequency of 0.71 and has an activity of 2.3% in cell culture based retrotransposition assays (Brouha *et al*., 2003). All the sperm, blood and hESC DNAs were hypermethylated for the first 16 CpG trinucleotide as well as the four transcriptionally important CpG sites (C, D, E, and F). However, the malignant cells (HeLa and SWs), NTera2D1 and placenta DNA showed methylation variation – especially at the four transcriptionally important CpG sites. The second L1 locus analysed, AC069384, has an allele frequency of 0.51 and 0% retrotransposition activity (Brouha *et al*., 2003). Direct bisulphite methylation analysis of this locus revealed that the hESC line was heavily methylated and had a similar methylation pattern as the germ line DNA. However, methylation variation – especially at the four transcriptionaly-critical sites (C-F) – was observed amongst the malignant cells and placenta (figure 5.4). The next studied locus, AC114499, had an allele frequency of 1.00 with an unknown retrotransposition activity. The hESC line was heavily methylated at this locus similar to the previous locus, AC069384. However, placental DNA, NTera2D1 and malignant cell DNA (HeLa and SW480 and SW620) had a significantly variable CpG dinucleotide at the beginning of the L1 promoter (+1 - +93). 62% of the CpGs in the transcriptionally important regions (+50/+64) were partially methylated or unmethylated in these DNA samples. The last studied locus, AC002980, was one of the very active L1 loci compared to the L1 RP, with an activity of 132% (Brouha *et al*., 2003) and an allele frequency of 0.485. As was discussed earlier in Chapter 4, the AC002980 belongs to a transduction family and has several offspring and sibling elements (Macfarlane *et al*., 2011, *In preparation*). For this locus the first +467 of the L1 promoter plus up to -500 upstream of the L1 (genomic flank) were analysed using two overlapping bisulphite-modified PCR assays (Chapter 2).

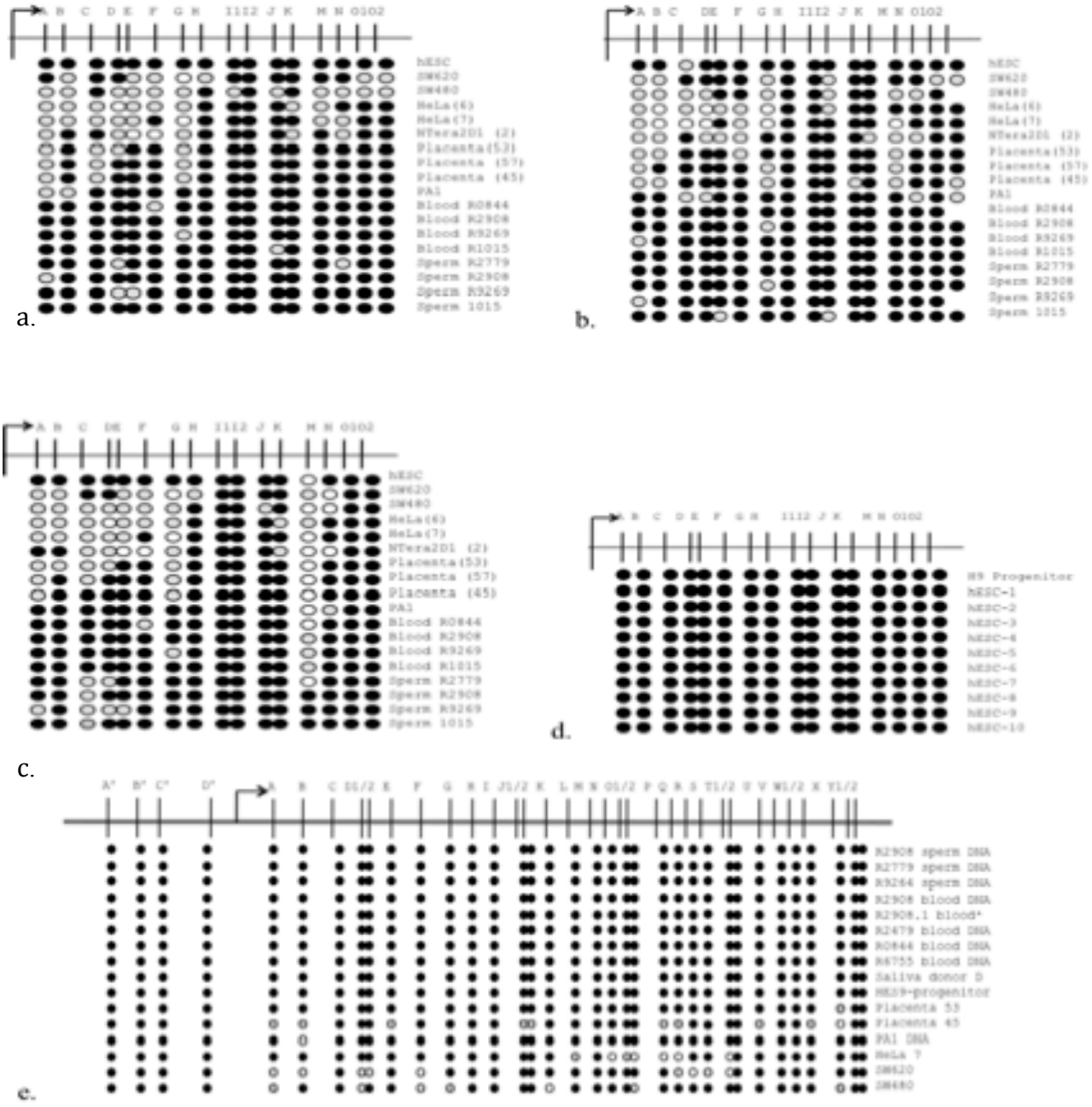

Figure. 5.4 Direct bisulphite L1 sequence analysis, (↱) L1 promoter +1, A- Y2 CPG dinucleotides of the 5′-L1 promoter, A′-D′: CpG of the flanking DNA upstream of the L1 promoter. C–F: transcriptionally important binding sites. Black lollipops: hypermethylation, white lollipops: hypomethylation, grey lollipops: partially methylated CpG. **a.** AC005885 locus direct bisulphite sequence analysis, 78% of the CpGs were methylated at this locus **b.** AC069384 locus direct bisulphite sequence analysis, 79% of the CpG were methylated, **c.** AC114499 locus direct bisulphite sequence analysis showed that 74% of the CpG were methylated, **d.** AC005885 locus direct bisulphite sequence analysis in hESC clonal lines, no methylation dimorphisim observed amongst the clonal lines at this locus **e.** AC002980 locus direct bisulphite sequence analysis, the majority (94%) of the CpG dinucleotides were hypermethylated at this locus.

The result showed that all of the four CpG dinucleotides at the 500bp proximity of the L1 promoter were heavily methylated. Also, 71% of the CpG dinucleotides within the first 467 bp of the L1 promoter were heavily methylated (figure 5.4).

To investigate the L1 methylation instability amongst the hESC clonal lines more fully, the AC005885-bisulphite PCR assay was applied to the bisulphite chemically treated DNA of hESC clonal lines (n=10). The sequence analysis showed that 100% of the CpG were methylated and no L1 methylation dimorphism was observed in these clonal lines (figure 5.4(d)).

5.2.3 Identifying L1 Loci With Variable Methylation Status in hESC

In order to screen for potentially active L1 elements showing variable methylation, the methylation-sensitive ATLAS method was used on single cell-derived human embryonic clonal lines (hESC). This technique is a modified version of ATLAS (Amplifying and Typing L1 Active Subfamilies) developed by Badge *et al*., 2003.

The samples used for this experiment included the hESC_H9 progenitor and its nine single cell-derived clonal lines (H9.1-H9.9). Since these clonal cell lines were first grown on HEF (Human Embryo Fibroblast) cells and then transferred onto feeder-free Matrigel medium, the HEF-cell DNA was used as an internal control for DNA cross-contamination. Fractions of linker ligated DNA from each clonal line were digested with the methylation insensitive (*Msp*I, Fermentas) and methylation sensitive (*Hpa*II, Fermentas) restriction enzymes (chapter 2). This technique targets the CpGs at positions in the +1 to +70 of the L1 promoter (figure 5.4).

As mentioned in the introduction to this chapter, these sites are critical for L1 mRNA expression. Following the differential digest, a semi-nested radioactive-labeled PCR was carried out and products were fractionated on polyacrylamide gels (Figure 5.7). As explained in more detail in chapter 2, to analyse the methylation status of each loci it is necessary to compare the digestion patterns of the Mock, *Hpa*II and *Msp*I digests for each single locus.

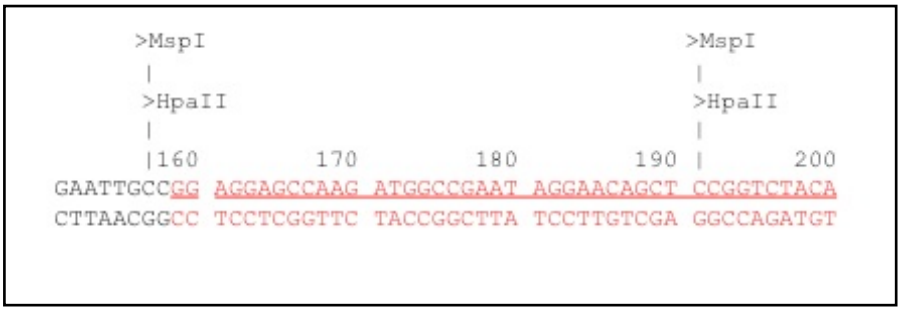

Figure 5.5 MS-ATLAS targets the critical transcription site CpG dinucleotide in the L1 promoter. The underlined red sequence shows the 5′ of the L1Hs sequence. The black sequence is the L1 junction sequence with the flanking genomic DNA. Also the *Hpa*II and *Msp*I restriction sites are presented in the sequence.

Twenty loci were extracted from the MS-ATLAS display gel (figure 5.7) and were re-amplified, cloned and sequenced. The analysed loci were grouped into three categories (figure 5.6): The first category of loci showed continuous bands across all three differential digests (Mock, *Hpa*II and *Msp*I) for each locus (for example bands 2 and 3 in figure 5.7). Since the *Msp*I restriction enzyme is methylation insensitive, it was expected to cut all the CCGG sites containing CpG dinucleotides regardless of their methylation status. The sequences of bands 2 and 3 revealed that these loci belonged to older L1 subfamilies such as L1PA2 with mutated *Msp*I restriction sites, and this has resulted in undigested products in the *Msp*I digest lane.

The second pattern is the appearance of bands in mock and *Hpa*II lanes and no bands in the *Msp*I digest bands 5 and 9 (figure 5.7). This pattern shows that the CpG dinucleotides within the CCGG site targeted by the assay were methylated. Therefore, it was not digested by *Hpa*II, but was digested by *Msp*I. The sequencing results from bands 5, 6, 14, 17 and 18 were L1Hs elements, which if the targeted *Hpa*II site is indicative, likely have methylated promoters (table 5.2).

The third category pattern observed on the display gel included bands 1-3, 15 and 16 (figure 5.7), with one band in their mock digest and no bands in either their *Hpa*II and *Msp*I lanes. This appearance is consistent with the unmethylated status of the targeted CpG dinucleotides. However, the sequences of unmethylated extracted bands all belonged to the older subsets of L1 such as L1PA2 and L1PA5. All the analysed sequences and their inferred methylation status are summarised in table 5.2.

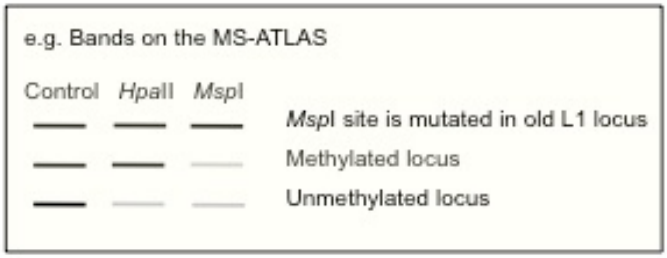

Figure 5.6 Schematic diagrams of different methylation patterns observed on the MS-ATLAS display gel.

The HEF DNA cross-contamination control in the hESC-MS-ATLAS display gel (figure 5.7) shows different patterns of methylation compared to hESC clonal lines. As is shown on the gel several bands were only present in the HEF cell lines and not in the hESC, while other bands were only present in the hESC clonal lines and not in HEF cell. This result suggested that there is no detectable level of hESC clonal lines gDNA contamination with the human fibroblasts feeder cells.

Methylation patterns of hESC clonal lines and their progenitor H9 cell revealed a stable and consistent L1 methylation pattern across the progenitor and its clonally derived cell lines (figure 5.7).

Bands	Locus	Repeat	Methylation status
1	AC087714	L1PA2	Unmethylated
2	AC130893	L1PA3	Unmethylated
3	AC024198	L1PA2	Unmethylated
4	Poor seq quality	NA	NA
5	AL133320	L1HS	Methylated
6	AC009902	L1PA3	Methylated
7	Poor seq quality	NA	NA
8	AC040934	L1PA	Mutated *Msp*I site
9	AC109822	L1HS	Methylated
10	AC066611	L1PA3	Mutated *Msp*I site
11	AC066611	L1PA3	Methylated
12	Poor seq quality	NA	NA
13	AC066611	L1PA3	Mutated *Msp*I site
14	AC069384	L1HS	Methylated
15	AC087783	L1PA2	Unmethylated
16	AC087783	L1PA2	Unmethylated
17	AC006269	L1HS	Methylated
18	AC005885	L1HS	Methylated

Table 5.2 Summary of the analysed sequences from the bands excised from the MS-ATLAS display gel (Figure 5.7)

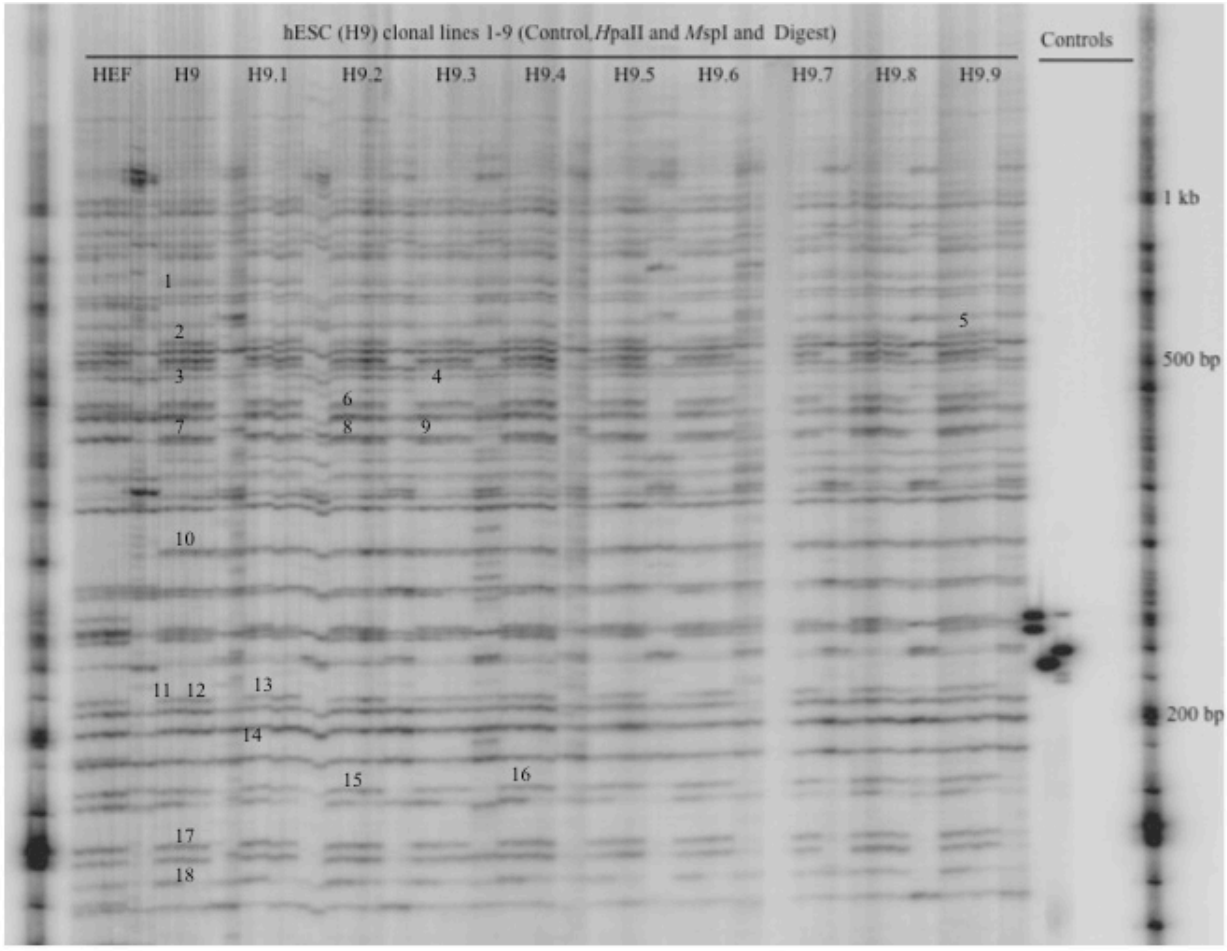

Figure 5.7 MS-ATLAS display gel for HES_H9 clonal lines (H9.1-H9.9), samples were run on the gel in duplicate: first two lanes for each sample are the mock digests, lanes 3 and 4 are for the *Hpa*II (methylation sensitive) digest, lanes 5 and 6 are the *Msp*I digest (methylation insensitive). The first six lanes are from the HEF gDNA; lanes 7-12 are HES_H9 gDNAs (progenitor).

5.2.4 Genome-wide L1 methylation in hESC, somatic and teratocarcinoma cells

In order to compare the global methylation of the subset of active L1 loci in hESCs with other cell lines such as NTera2D1 and sperm and blood DNA, the MS-ATLAS method was applied independently to all the samples, and their products were fractionated on the denaturing polyacrylamide gel electrophoresis (DPAGE) (figure 5.11). According to the standard protocol (chapter 2), the genomic DNA was digested with *Ase*I to make a linkered library. Following library construction the ligated gDNA was subjected to the three methylation-differential digests described above.

Following the methylation differential digest all the digested gDNA was subjected to suppression and exponential PCR, as explained in more detail in Chapter 2.

The MS-ATLAS display gel of these samples is presented in figure 5.8. The gel showed consistent patterns amongst the hESC clonal lines and their progenitor. In blood and sperm samples for each donor it was expected that there would be variation within each donor arising because of L1 tissue-specific methylation patterns between blood and sperm DNA. However, the MS-ATLAS on NTera2D1 clonal lines showed variation amongst the clonal lines and their progenitor. The display patterns of these samples revealed losses and gains of methylation at several (60% of the L1 loci) different loci in different clonal lines. Again 15 different L1 loci from the polyacrylamide display gel (figure 5.8) were characterised by sequencing and the results are summarised in table 5.3.

Sequence	Accession No.	Repeat	Methylation status
1	Unmappable	NA	Mutated *Msp*I site
2	AL606752	LTR	Methylated
3	AL353719	L1PA5	Mutated *Msp*I site
4	AC127380	C ALR/Alpha Satellite	Mutated *Msp*I site
5	Z98754	L1MC	Mutated *Msp*I site
6	Unmappable	L1PA6	Mutated *Msp*I site
7	AL139288	L1P	Mutated *Msp*I site
8	AC008943	L1HS	Methylated
9	AL035459	No repeat	NA
10	AP000652	L1HS	Methylated
11	AL035459	No repeat	Methylated
12	AL669984	L1PA2	Methylated
13	AC104454	L1HS	Unmethylated
14	AL669984	No repeat	NA
15	AC007556	L1HS	Methylated

Table 5.3 Summary of the analysed sequences from the bands excised from the MS-ATLAS display gel (figure 5.8)

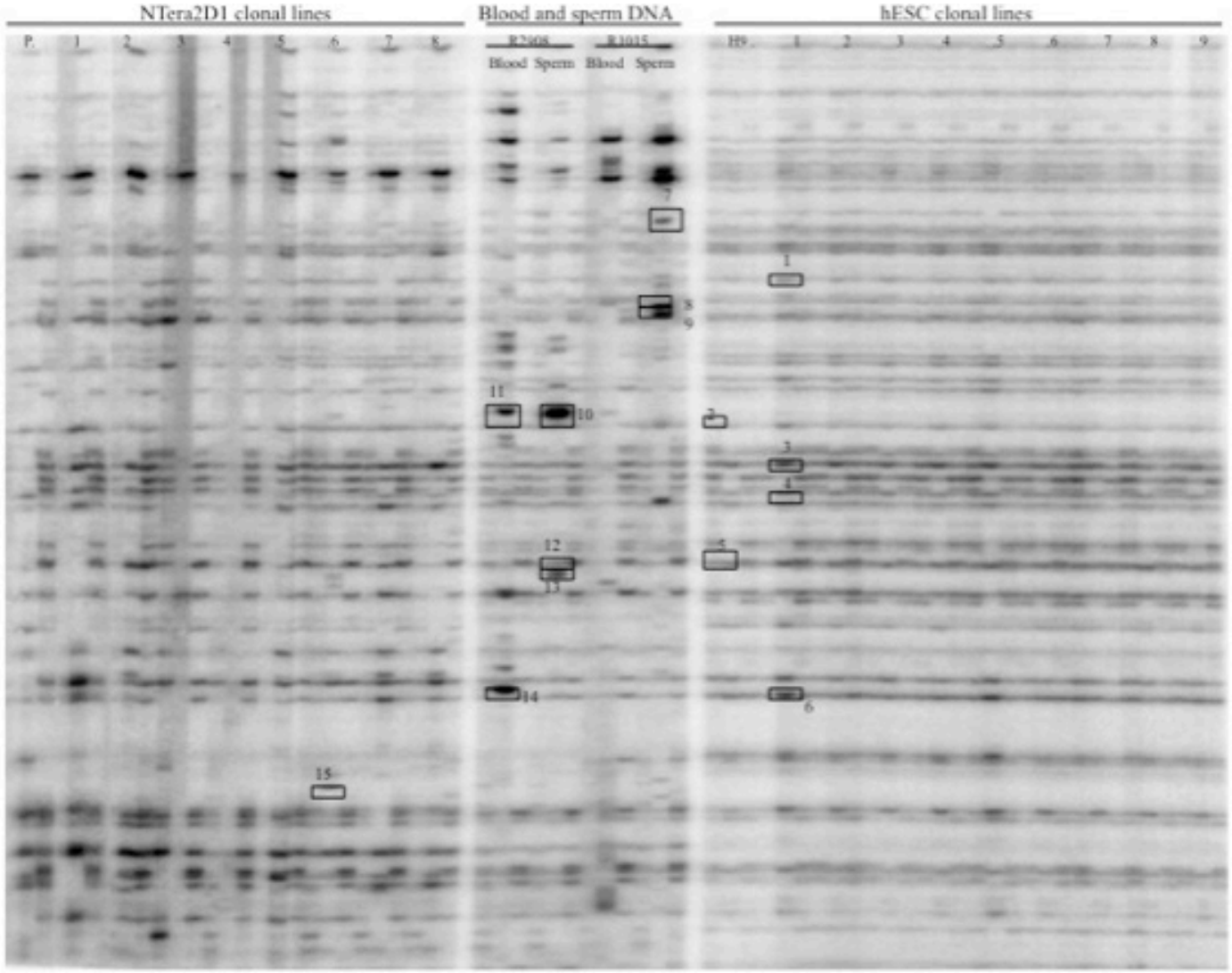

Figure 5.8 MS-ATLAS genome wide L1 L1 methylation comparison of L1-hESC clonal lines with somatic (blood and sperm) cell and NTera2D1. Differential digest order for each samples was as follow: Mock (-ve) digest, *Hpa*II (methylation sensitive), *Msp*I (methylation insensitive). Samples are duplicated for each digest. Columns from left to right: NTera2D1 progenitor, and its 8 clonal lines, second column: blood and sperm DNA form donors:r2908 and R1015, the third column (right): H9 progenitor and its 9 hESC clonal lines. Boxed bands were extracted from the display gel for further analysis.

5.2.5 MS-ATLAS display gel quantification analysis

To find out if there was any variation in the methylation of different L1 loci for different clonal lines (NTera2D1 and hESCs) we selected seven L1 loci randomly across the polyacrylamide gel (figure 5.8) at the following positions: 1kb, 600 bp, 500bp, 450 bp, 400 bp, 250 bp and 100 bp. The level of methylated L1 promoters at different loci on the MS-ATLAS display gel was quantified using the ImageJ software. The intensity of undigested bands *Hpa*II and *Msp*I restriction digests was measured together for each clonal line at each selected locus with ImageJ for both NTera2D1 and hESC clonal lines. The T-test analysis for each locus in hESCs revealed no significant variation in methylation loss/gain across clonal lines (P=0.47). However, the results for the NTera2D1 showed a significant ($p<0.0092$) methylation variation across the clonal lines for each of the selected locus. The two-way with replicates Analysis Of Variance (ANOVA) of each of the cell lines (hESC, NTera2D1) showed that there is a significant variation in L1 methylation between the NTera2D1 and hESC clonal lines (p=0.038). In other word the L1 methylation status is significantly correlated to its gDNA source. Also the result showed that there is a significant correlation between L1 methylation and the L1 locus (p=9.8639E-22), *i.e.* some L1 loci are significantly methylated/ unmethylated compared to the other loci in each cell lines (table 5.4).

Two way ANOVA with replicates

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
L1-Loci	25231.30159	6	4205.216931	31.05027005	9.8639E-22	2.180563624
Cell line	595.8412698	1	595.8412698	4.399542911	0.038198286	3.925834021
Interaction	8877.714286	6	1479.619048	10.92513698	1.48913E-09	2.180563624
Within	15168.44444	112	135.4325397			
Total	49873.30159	125				

Table 5.4 Two-way ANOVA of each of the cell lines: NTera2D1 progenitor and 8 clonal lines, hESC (H9) progenitor and 9 clonal lines, the source of variation for each DNA is presented in the table. As it is demonstrated there is a significant difference between the L1 methylation in hESCs and NTera2D1 clonal lines (P=0.038). Also the methylation of L1 is significantly related to the locus (P=9.8639E-22).

5.2.6 MS-ATLAS display gel replication analysis

In order to test whether the MS-ATLAS display gel technique is reproducible, the libraries of hESC clonal lines were treated independently to the display PCR and two separate display gels were produced (figures 5.7, 5.8). Intensity of the methylated bands (*Hpa*II) and unmethylated bands (*Msp*I) for the seven randomly selected loci (mentioned above in section 5.2.5) from both gels were measured using Image J software. The average intensities of all seven loci were compared between the first and second replicates for each of the clonal lines. The result is demonstrated in figure 5.9 and it shows that there is no significant (p=0.127) difference between the two replicates.

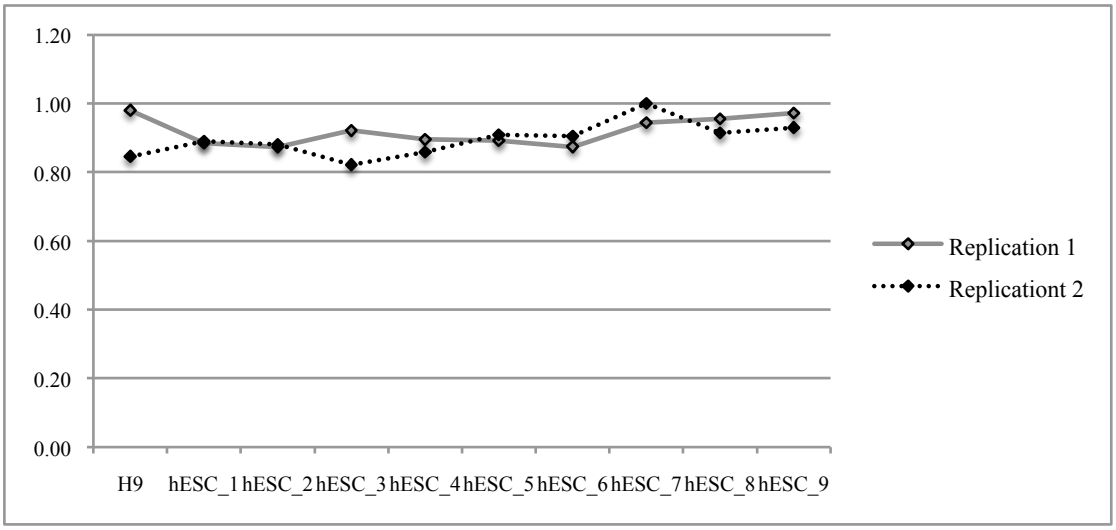

Figure 5.9 Comparison between the average intensities of *Hpa*II and *Msp*I digests of seven L1 loci of hESC clonal lines MS-ATLAS display gels (replicates 1 and 2). The replicate 1 (in grey): display PCR carried out according to the protocol (chapter 2) on hESC clonal lines differentiated libraries. Experiment 2 (in dotted black) second set of display PCR on the same differential libraries of hESC clonal lines. There is no significant difference between replications (p=0.127).

Referring to figure 5.9, it can be seen that both experiments are following the same pattern. Some level of variability was expected between the replicates. This variability could be due to different yield in the PCR steps or differential efficiency of the enzymatic reactions. However, statistical analysis showed that there was no significant variation (p=0.127) between the replicates and so MS-ATLAS was concluded to be reproducible.

5.3 Discussion

Methylation of cytosine residues of DNA, as a mark of transcriptional repression seems likely to have primarily developed as a host defence mechanism against mobile elements, most obviously in mammals. Therefore, the study of the methylation status of active L1 retrotransposons could provide insights into how host defence mechanisms function to suppress these elements activities in remodelling the genome, and also how active subfamilies escape these epigenetic marks to be able to retrotranspose freely, which they must do to remain active.

According to the host defence model it is more likely for the younger L1s to be targeted by methylation, since RC-L1s are more of a threat to genome integrity. On the other hand old L1 subfamilies such as L1PA2, which were once a threat to the genome and consequently heavily methylated, have acquired inactivating mutations such that they can no longer be active. During the *de novo* methylation of the genome during gametogenesis and post implantation development, these old L1 loci can remain unmethylated and may not be re-methylated, but this has no deleterious consequences as they can no longer encode the L1 proteins. However, in the case of young L1s, these modifications may have not enough time to act on young L1s and keep their activity in check, therefore some methylation dimorphism might be observed amongst younger L1 loci.

Conversely, the situation may be that the methylation pattern of L1s is irrelevant in somatic cells as these pose little threat to the host and therefore methylation patterns are not stably maintained.

It has been argued that embryogenesis and the cancer phenotype may lead to a breakdown in the methylation machinery, although whether this involves the *de novo* or maintenance methylation machinery is not known. It is known that in malignancy-derived cells all repetitive elements (especially L1s) undergo global hypomethylation, which could be an important cause of genome instability in cancer cells. However, whether this global hypomethylation of L1s only effects older elements rather than RC-L1s remains unknown. The results presented here involve close observation of the methylation status of full-length L1 loci. A variety of embryonal cell lines such as

hESCs and malignant derived cells have been used for this study to give insight into the methylation status of the young and potentially active L1 elements during embryogenesis and different stages of cancer progression.

5.3.1 Global L1 methylation analysis

Prior to this study, the methylation status of full-length L1 loci, both young and older L1 families have been analysed. The result of applying L1-COBRA (Chalitchagorn *et al*., 2004) to different cell lines revealed that there is variability in L1 global methylation in different cell lines. L1 global methylation was grouped into three distinct clusters amongst cell lines. 75% of the gDNA associated with the L1 that amplify with COBRA primers in somatic cells (blood DNA) and germ cells (sperm DNA) was hypermethylated, while in malignant-derived cell lines (HeLa, SW480, SW620 and NTera2D1) only 25% of the gDNA associated with the L1 were methylated. However, on average 50% the gDNA associated with the L1s in human embryonal cells was methylated. Also, analysis showed that hESC and placental DNA show a distinctive intermediate level of methylation when compared to other tissues ($p=0.00404$) and cultured cells ($p=0.0095$).

The global L1 methylation analysis showed hypermethylation of L1s in sperm and blood DNA and global hypomethylation of L1s in malignancy derived cells. These results were similar to other studies (Pornthanakasem *et al*., 2008). However, in embryonal cells, an intermediate level of L1 methylation was observed which was significantly different from the L1 global methylation of germ cells as well as malignant derived cell lines. We had thought that embryonal cells, especially hESCs, would have a L1 global methylation pattern more similar to that of germ cells. One explanation for this observation could be that the unmethylated loci in embryonal cells might be from the older L1 families such as L1PA2, which are mostly not retrotransposition competent due to mutation accumulation and therefore they are no longer counted as a threat to the genome. Hence during the *de novo* methylation process in the imprinting stage of embryo development they may remain unmethylated. Yet despite this it is not clear why less of the L1 CpG dinucleotides are methylated in human embryos when compared to the germ cells and somatic cells.

However, due to the limited resolution of the L1-COBRA assay it is not possible to further investigate the reason for an intermediate methylation level of L1 loci in human embryonal cell lines, as the assay sums the status of many individual loci. Therefore, other techniques are required to investigate L1 methylation status in human embryonal cells in more depth, and at higher resolution. Results from methylation locus-specific analysis and genome-wide methylation analysis of active L1 subfamilies are discussed in the following parts of this discussion.

5.3.2 Locus-specific L1 methylation analysis

As discussed above, because of the poor resolution of the L1-COBRA assay the global methylation analysis of L1 is not informative with respect to which families of L1 were methylated and which loci remained unmethylated in different cells. To discover the methylation status of active L1 loci in different cell lines, four active L1 loci were selected for direct bisulphite methylation analysis. The four young L1HS loci were as follows: AC005885, AC069384, AC114499 and AC002980. Each locus had different retrotransposition activities (section 5.2.2). This enabled the study of the correlation between the L1 activity and its promoter methylation status, but the results showed no significant correlation. However, the data suggest that there is a correlation between L1 methylation loci and cell type.

Statistical analysis showed that the hESC cells were more similar to tissue DNA with respect to their L1 methylation status at these four loci ($p=0.03479$), unlike the results from the L1-COBRA assay. However, the placenta DNA remained significantly hypomethylated compared to adult tissue DNA ($p=0.02424$): this finding is supported by Cotton *et al.* (2009), as they observed hypomethylation of LINE1 elements in both male and female placentas.

5.3.2.1 L1 methylation stability of hESC-derived cells

In order to determine if L1 methylation remains stable in clonally derived human embryonic cells, the methylation status of ten, single-cell derived (H9) clonal lines

(hESC 1-10) was studied for the AC005885 insertion. The result of the bisulphite sequence analysis suggested that the methylation status in the hESCs was highly stable and 100% of the CpGs had identical methylation status to their progenitor cell (H9). However, this finding is opposite to our previous observation of intact L1 insertion loci showing bidirectional clonal variation in their methylation status in HeLa cells (Modes and Badge, unpublished data, 2006). This data also suggests that the L1 methylation status is specific to the cell type. The hESC clonal lines are apparently very stable with respect to their L1 methylation status.

Overall, the L1 locus-specific methylation analysis has suggested that it is more likely that young and active L1 subfamilies become methylated as part of a host defence mechanism. Therefore, the majority of globally umethylated L1 loci observed in L1-COBRA analysis may derive from these numerous older elements. However, the question still remains: can a small subset of RC-L1 escape the epigenetic mark to actively retrotranspose in the genome? To answer to this question would require more information on the genome-wide methylation status of L1 in hESCs and other cell lines, which is discussed below.

5.3.3 Genome-wide methylation analysis of active L1 subfamilies

Although direct bisulphite methylation analysis has high resolution for studying the methylation status of individual L1 loci, there are a number of limitations with this technique. Firstly, this technique is expensive and consumes large amounts of gDNA during bisulphite conversion. Secondly, it is not always possible to develop a bisulphite PCR assay, depending on the genomic location of the L1. For instance, an attempt to develop a bisulphite-based PCR assay to study the methylation status of the AC067958 insertion failed, due to the nature of the insertion, which is located within another repeat (an Alu repeat upstream of the L1-AC067958). As a result it was not possible to study the L1 methylation of this locus specifically. Hence, there are often limitations with designing locus specific bisulphate assay due to the nature of the locus. To solve this problem we have developed the methylation sensitive ATLAS technique to analyse the genome-wide methylation status of young L1 promoters. The steps for this technique are detailed in Chapter 2 and section 2 of this chapter.

5.3.3.1 L1 methylation analysis in human embryonal clonal lines with MS-ATLAS

The results derived from the sequences of the analysed bands from the MS-ATLAS in hESC clonal lines are summarised in table 5.2. 27% of the characterised bands belonged to the L1Hs family and their methylation was verified by sequencing all three bands from mock, *Hpa*II, and *Msp*I digests. Two of the L1HS elements isolated from the display gel belonged to the L1-AC005885 and L1-AC069384 loci, whose L1 methylation statuses were verified earlier by direct bisulphite sequence analysis. The results of L1 methylation status from both techniques were essentially the same, *i.e.* the bisulphate technique could validate the result of MS-ATLAS in respect to these two L1 loci.

Moreover, the MS-ATLAS display patterns of hESC clonal lines showed that 43% of the displayed L1 loci were unmethylated. This result supports the intermediate level of L1 methylation in hESC using the COBRA-L1 assay. However, the MS-ATLAS data suggest that more of the active L1 loci must contribute to the unmethylated portion of L1s in the genome of hESC than expected, and therefore the active L1s as well as the older families are likely to be responsible for the bulk of L1 unmethylated loci in hESC genome.

A comparison of the results of the genome-wide L1 methylation status between different cell lines (figure 5.8) showed that there is significant L1 methylation variation between the NTera2D1 clonal lines ($P<0.0092$). Interestingly, there are no significant methylation variations ($P=0.47$) between the hESCs clonal lines. Also, the genome-wide methylation analysis on HeLa cell clonal lines (using MS-ATLAS) showed loss and gain of L1 methylation in different clonal DNAs (Modes and Badge *et al*., 2006 unpublished data). It can be concluded that there is a high L1 methylation variability amongst the malignancy-derived clonal lines such as HeLa and NTera2D1. This is likely to be due to the global genome instability in these cell lines. In contrast, L1 methylation in hESCs clonal lines remains stable due to their robust genome stability. Moreover, data has suggested that there is a significant variation in L1

methylation between the NTera2D1 and hESC clonal lines (P=0.038). In other words the L1 methylation status is significantly correlated with its gDNA source.

It is still unclear whether L1 methylation is a targeted or non-targeted process in the genome, and if methylation of each locus is defined by the methylation status of the surrounding genomic DNA. However the statistical analysis on genome-wide L1 methylation data of display gel 5.8 showed that the level of L1 methylation is significantly correlated with the cell type and therefore that L1 expression might be regulated differently in different cell types.

Also statistical analysis revealed that there is a significant difference in L1 methylation of different loci in hESCs, this observation supports the hypothesis that in the case of young L1s *de novo* methylation may have not enough time to take hold and therefore some methylation dimorphism might be observed amongst the younger loci. It can be suggested that L1 self-regulates its rate of replication, by maintaining a low transcription rate with little genomic impact and hence little negative selection. However, using the MS-ATLAS technique, there is an increased probability of finding these rare, but active elements.

5.3.3.2 L1 replication test for the MS-ATLAS technique

To assess the MS-ATLAS techniques reliability and robustness, it was applied to two separate display PCRs of the hESC clonal lines and these were run separately on two separate display gels (figure 5.7 and 5.8). Measuring the intensity of the methylated bands for each clonal line and comparing the two sets of replicates showed there is no significant difference between the methylation patterns observed for each replication figure 5.9. This data suggests that the MS-ATLAS technique is reproducible and the data can be reproduced in the same pattern regardless of other variable factors.

Finally, besides a few limitations with the MS-ATLAS technique, such as the fact that it relies on enzymatic reactions and complex display gels, this technique has a high coverage of the genome with good resolution and it can analyse a defined subset of active L1 subfamilies simultaneously. Sequence analysis of the extracted bands from the MS-ATLAS display gel showed the presence of some of the older subfamilies

such as L1PA2 as well as the L1HS family. Although ATLAS is designed to only capture young L1 families (Badge *et al*., 2003) some elements from older families can also be amplified due to mutations in the primer site that "phenocopy" younger elements. Sequence analysis of captured older L1s revealed that all of these elements were amplified due to a single base mutation at the +74bp (G>T) of the L1, and therefore, they can be amplified during the suppression PCR. However, amplification of the older families is also informative, as it will give a clearer idea of L1 global methylation by including a proportion of "control" loci.

Chapter 6

Investigation of L1 retrotransposition in human embryos

6.1 Introduction

In 2003, Brouha *et al*. suggested that although there are about 90 full-length L1s with intact ORFs in the reference human genome, only six "hot-L1s" are responsible for the majority (86%) of the total L1 retrotransposition activity. Since then studies on human-specific L1 retrotransposition using many different methodologies such as comparative bioinformatics analyses, or transposon display techniques such as ATLAS (Badge *et al*., 2003**,** Buzdin *et al*., 2003, Roy *et al*., 1999; reviewed in Beck *et al*., 2011) have revealed many more active elements segregating in human populations. However it has proven very difficult to find *de novo* L1 retrotransposition events, largely due to the low copy number of active L1s, the low frequency of such events and the lack of high-resolution and high coverage techniques. Recently, high throughput sequencing approaches using Roche 454, Illumina, as well as array-based systems have began to facilitate the study of L1s as genomic structural variants. Currently, there are several approaches to study L1

retrotransposition in the genome. One approach is the PCR-based transposon display techniques used to characterise polymorphic human L1s in the human genome (Badge *et al*., 2003, Mathews *et al*., 2003). Generally these techniques rely on the selective amplification of groups of retrotransposons based on diagnostic nucleotide polymorphisms specific for each subfamily (*e.g.,* the trinucleotide sequence ACA discriminates the Ta subfamily from older subfamilies). In these methods, selective / suppressive PCR is usually applied to constructed gDNA libraries, and polymorphic elements that show presence / absence variation between individuals are isolated for characterization by sequencing. A good example of a PCR-based display technique used to study active and polymorphic L1s in the genome is ATLAS (Badge *et al*., 2003). Using the ATLAS technique, Badge *et al*. (2003) identified nine full-length L1 insertions in an individual genome, of which three were counted as hot L1s in cell culture retrotransposition assays. Moreover, as has been presented in the three earlier result chapters of this thesis, modified ATLAS-based procedures have been used to study L1 retrotransposition (chapter three), the diversity of active L1 3′ transduction families (chapter four), and the genome-wide methylation status of young L1 promoters (chapter five). As discussed in Chapter 3, we have isolated two novel polymorphic L1 insertions and one potential *de novo* insertion by applying a modified version of the ATLAS technique to one human embryonic stem cell line and six of the clonal lines derived from this cell line. These data suggests that young HsL1s are under-represented in the human genome reference sequence (reviewed in more detail in Beck *et al*., 2011).

Another technique to investigate polymorphic retrotransposon insertions is microarray profiling. This technique uses the same strategy as the display techniques to selectively amplify specific retrotransposon subfamilies and their flanking genomic DNA, followed by hybridisation of the amplicons to the arrays, for characterization of the retrotransposon flanking genomic DNA without sequencing it. Application of microarray-based techniques has resulted in the identification of novel L1, Alu and HERV-K insertions (Gresham *et al*., 2011). Moreover, Huang *et al*. (2010) performed transposon insertion profiling by microarray (TIP-chip) to map human L1Ta elements genome-wide: they identified 111 novel (*i.e.* absent from hg19) human L1Ta insertions with highly variable allelic frequencies (0.013 and 0.987) Based on this, Huang *et al*. (2010) estimated the occurrence of new L1-mediated retrotranspositions

to be twice as high as previously estimated: 1 insertion in every 108 birth, and described these repeats as "under-recognised" as a source of human genomic diversity.

Other recent studies have used next generation sequencing approaches such as 454 and Illumina DNA sequencing to identify numerous novel and polymorphic L1 and Alu insertions in the human genome (Ewing *et al*., 2010; Iskow *et al*., 2010; Hormozdiari *et al*., 2011). Indeed, Ewing *et al*. (2010) identified 367 novel L1 insertions, which were absent from the human genome reference sequence by using PCR-based capture and Illumina high throughput sequencing. Iskow *et al*. (2011) characterised nine somatic *de novo* L1 insertions in lung cancer tumours by using restriction enzyme profiling followed by Sanger and 454 next generation sequencing. Hormozdiari *et al*. (2011) performed computational analysis using paired-end mapping sequences from the Illumina platform and identified 4342 Alu insertions, of which eighty percent of the insertions were reported as being novel. Finally, Witherspoon *et al*. (2011) characterised 487 novel Alus in four unrelated individuals by using Illumina high throughput sequencing.

Another method to detect retrotransposon polymorphisms is paired-end fosmid sequencing. In this technique the strict packaging limit of these vectors enables the identification of indels due to the inconsistent placement of sequence traces from the ends of the insert (Kidd *et al*., 2008, 2010; Beck *et al*., 2010). Indeed, using the above technique, Beck *et al*. (2010) identified 68 novel full length L1 insertions from six individual human genomic libraries, and they demonstrated that 37 of these insertions were highly active L1s, by using cell culture-based retrotransposition assays.

All of the above studies are significant to our understanding of the degree to which mobile DNAs contribute to genetic diversity, heritable disease, and perhaps oncogenesis. Yet with the advent of new generations of high throughput technologies, this information is rapidly expanding the catalogue of genomic variants created by retrotransposon activity. However, despite all these advances in L1 diversity and the fact that L1 must retrotranspose in the germline to be evolutionarily successful, direct assessment of *de novo* L1 retrotransposition in the germline and, or early embryogenesis has not been achieved for endogenous L1 elements. A recent study by Ewing *et al*. (2011) has suggested that the frequency of polymorphic L1 insertion is

very low, varies between different populations and is more likely to occur in the human germline. A direct study of *de novo* L1 retrotransposition in sperm DNA by Freeman *et al*. (2011) suggested that the rate of L1 retrotransposition in the germline is much lower than previously estimated (1 in 400 individuals). Based on these downwards, revised estimates of the L1 retrotransposition rate we attempted to directly access L1 retrotransposition in hESC and several human embryonal cell line by using modified display ATLAS techniques (Chapter three). However, as discussed in Chapter 3, the low level of L1 activity and the complexity of the display patterns generated established that to investigate *de novo* L1 retrotransposition in human embryogenesis would require higher genome coverage. Therefore in this chapter we have combined a high resolution ATLAS technique with 454 high throughput sequencing at high sequence depth (11-25 X per single amplicon) to study L1 retrotransposition in whole genome amplified (WGA) DNA from three human embryos, not only using the total DNA from blastocysts but also from single blastomeres (single cells). The advantage of using WGA gDNA is that allows us to characterise single molecule *de novo* L1 retrotransposition events, without destroying the single molecule in the characterization process. Therefore real single molecule event can clearly be distinguished from rare somatic L1 mosaicism by bioinformatics analysis, and further validated by PCR in the source WGA gDNA.

The current chapter follows on from chapter three to produce a sensitive, genome-wide, high throughput assay based on the ATLAS display technique, to characterise *de novo* L1 retrotransposition. For this purpose we used human embryos and individual blastomeres to investigate *de novo* L1 retrotransposition in early human embryogenesis. By screening single cells derived from blastocysts for their L1 complement and comparing this with the total embryonic DNA prepared from the remaining blastocysts it is possible in principle to unequivocally identify *de novo*. In this chapter, we demonstrate that our proposed high throughput display technique is able to recover single molecule L1 retrotransposition events from the genome and identified several novel L1 insertions, which have not previously been reported. Also, by using this technique we have discovered candidate *de novo* endogenous L1 - retrotransposition events in human embryos as well as germline specific L1 insertions in sperm genomic DNA.

6.2 Result

The work in this chapter is divided into three stages: genomic DNA preparation and quantification for NGS-library construction, NGS-experimental design, and finally high throughput data analysis. These stages are explained in more detail below.

6.2.1 Preparation and quantification of human embryo WGA

Whole-genome amplified gDNA from embryos and blastomeres, was provided by Dr. Jose Garcia-Perez (University of Granada, Spanish Stem Cell Bank, Spain, Granada); the procedure for WGA of the single cell blastomere is explained in more detail in section 2.2.2. Human embryos were at the pre-implantation stage (Day +1 - Day +6 after fertilisation), and single blastomeres were isolated from 6-8 cell stage embryos (Day +3 – Day +6). Table 6.1 summarises all the stages of the embryos and their blastomeres.

Embryo/Sex	**Total DNA**	**Blastomere code**	**454-experiment**	**454-diluted libraries**
E3/F*	E3T1-3	Stage +5, Blastocyst	454-library 1	-
E3/F		E3B2-1	454-library 2	-
E3/F		E3B2-2	454-library 3	-
E3/F		E3B2-4	454-library 4	-
E3/F		E3B2-5	454-library 5	-
E3/F		E3B2-6	454-library 6	-
E4/F	E4T1-4	Stage +6, late Blastocyst	454-library 7	-
E4/F		E4B3-1	454-library 8	100 fold diluted
E4/F		E4B3-2	-	-
E4/F		E4B3-3	454-library 9	100 fold diluted
E4/F		E4B3-4	-	-
E4/F		E4B4-1	-	-
E6/M*	E6T1-6	Stage +5, Blastocyst	454-library 10	-
E6/M		E6B5-1	454-library 11	-
E6/M		E6B5-2	-	-
E6/M		E6B5-3	454-library 12	-
E6/M		E6B5-5	-	-
Adult gDNA	**Source**	**-**		
R2908B	Blood	NA	454-library 13	-
R2908S	Sperm	NA	454-library 14	-

Table 6.1 Summary of all the analysed embryos and their blastomeres, *F/M is the summary of the result from genetic sex identification (section 6.2.1.1). (Female/Male). Samples were selected based on their quality for 454-sequencing (library 1-14), Libraries 8 and 9 were diluted 100-fold prior to the emPCR, since they showed very biased amplification of a small number of distinctively sized amplicons.

Diluting these libraries meant that their high representation of just a few sequences did not result in their being sequencing at wastefully high coverage.

6.2.1.1 Sex determination of the embryos DNA

A ZFX/Y test was carried out on total WGA gDNA from each embryo. The genotyping results are presented below in Figure 6.1, showing embryos three and four to be female and embryo six to be a male. This result is also summarised in Table 6.1.

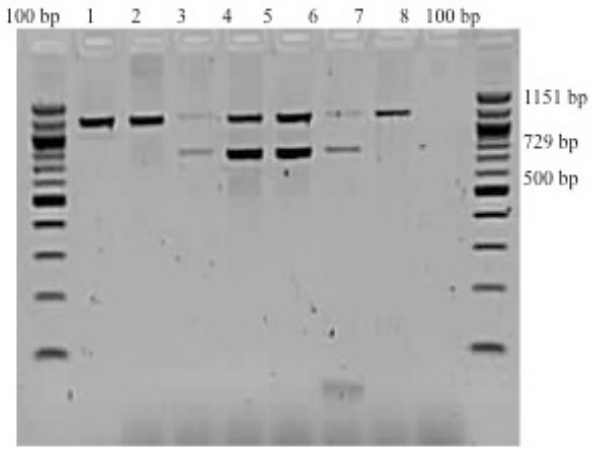

Figure 6.1 Result of sex determination of total DNA of each embryo using the ZFX (1151bp) and ZFY (729bp) PCR test. Lane 1: E3T1-3, lane 2: E4T1-4, lane 3: E6T6-1, lane 4: Sperm DNA, 5: CEPH panel male, 6: CEPH panel (male), 7: CEPH panel (female), 8: PCR –ve control. As the result shows, embryos three and four are females (one amplified band at the ZFX location 1151 bp) and embryo six is a male as it presented with two bands (ZFY= 729bp and ZFX= 1151 bp).

6.2.1.2 WGA embryos qualification

In order to check the quality of the WGA gDNA we have genotyped all the three human embryos total DNA for ten randomly selected polymorphic L1HS loci (Figure 6.2). The reason for choosing L1 related genomic-loci was to be able to distinguish between gDNA locus drop out, from the homozygote L1 insertion in the designated locus.

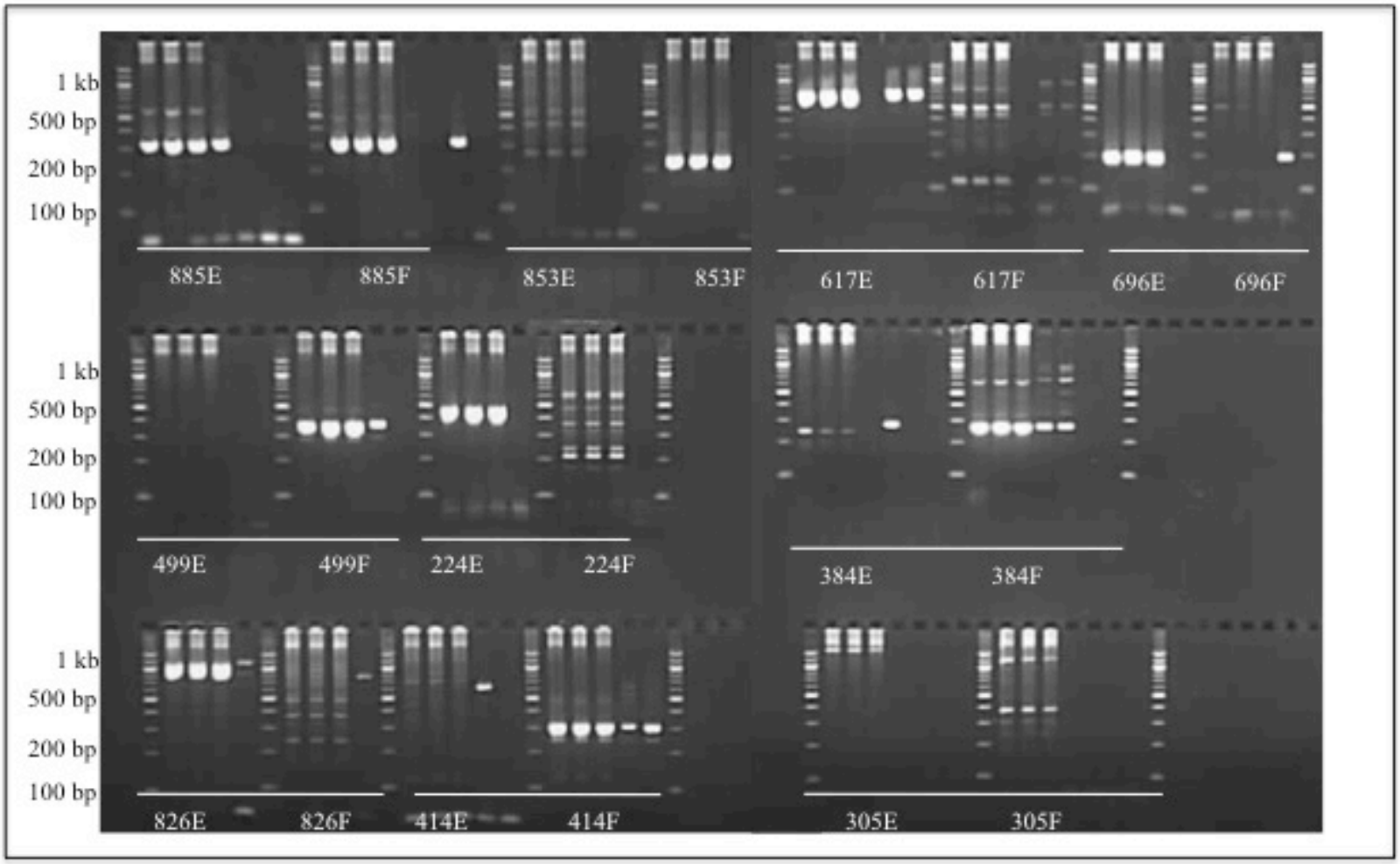

Figure 6.2 Genotyping the three embryos total DNA for ten random L1HS polymorphic insertion to check the L1 loci representation in whole genome amplicon gDNA, E: empty site (no L1 insertion only the genomic flank), F: filled site (L1 insertion), For each genotyped loci: lane 1: Embryo 3 Total DNA (WGA), lane 2: Embryo 4 Total DNA (WGA) and lane 3: Embryo 6 Total DNA (WGA), the rest of lanes are –ve and +ve controls were they are available. AC005**885**: E= 346 bp, F= 283 bp; AL583**853**: E= 445 bp, F= 225 bp; AC015**617**: E= 726bp, F= 241 bp; AC108**696**: E= 200 bp, F= 200bp; AC114**499**: E= 643 bp, F= 375 bp; AC087**224**: E= 472 bp, F= 491 bp; AC069**384**: E= 281, F= 246 bp; AC008**826**: E= 1139, F= 775 bp; AC009**414**: E= 641 bp, F= 310bp; AL050**305**: E= 1810, F= 1300 bp

All the ten loci were scored one if a band was present in empty site or filled site or both. The result showed that 10 out of 10 of these gDNA loci were present in all the WGA embryos.

6.2.2 Single molecule (*Vsp*I library) ATLAS

Prior to next generation sequencing, the quality of the WGA gDNA of the embryos was further tested by carrying out single molecule ATLAS (section 2.7, Macfarlane *et al.*, unpublished data). This technique is able to check the L1 copy number representation for a L1 locus of known genotype at the single molecule level, in each individual library. Libraries were constructed (Using the *Vsp*I library protocol) for both embryos 3 and 4. Libraries utilised DNA (WGA) amplified from individual blastomeres (*e.g.* E3B2-1) and the total DNA derived from the remainder of the blastocyst after blastomere dissections (*e.g.* E3T1-3). Also, non-WGA amplified blood and sperm DNA samples from two healthy adult donors were used as positive controls. Following the single molecule ATLAS protocol (section 2.7), several dilutions of the source DNA (WGA and non-WGA DNA) were made (Figure 6.3), and primary PCRs were carried out on this dilution series. The PCR products were blotted followed by hybridisation to a radiolabelled probe specific for the L1 locus AC005885. The same procedure was carried out for two more L1 loci: AC069384 and AC114499 (from the earlier genotyping we know that these insertions are present in all three embryos, figure 6.2). By comparing the intensity (ImageJ) between the embryos and blood and sperm gDNA, the number of amplifiable molecules from each L1 locus present in the original source DNA was calculated (Jeffreys *et al.,* 1994, Monckton *et al*., 1994).

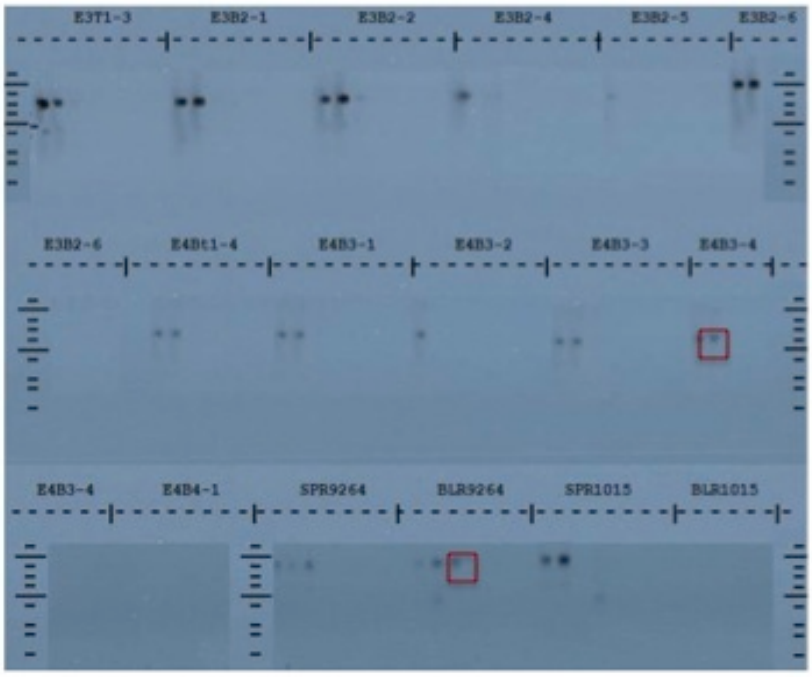

Figure 6.3 Single molecule ATLAS primary PCR (dilution series) probed with the L1-885 locus for embryos three and four as well as blood and sperm DNA from two different individuals. The dilutions were prepared at the following concentrations for each sample (ng/ul): 5.32, 3.99, 2.66, 0.26, 0.026, 0.0026 and 0.00026. e. g. Comparison the intensity (red boxed) from the amplified L1-885 locus molecules in blood with embryo four blastomere 3-4 shows that the L1 representation at the 885-locus is same as in the blood gDNA.

Based on the L1 locus-specific hybridisation results on the different dilutions the correct dilution for the secondary PCR was calculated. The L1 locus hybridisation result showed variable single molecule dilution levels for some of the blastomeres at all three loci. As in the example of the L1-AC005885 locus (Figure 6.3), specific hybridisation showed that for the majority of the libraries the second dilution (3.99ng / ul) was suitable as an equivalent for single molecule level. However, most of the blastomeres (E4B3-2, E4B3-4 and E4B4-1) of embryo four showed a first dilution equivalent to single molecule level (Figure 6.3). Based on this, the secondary PCR was carried out at this level, and the result is presented in Figure 6.4. Most of the blastomeres in embryo four only had very few amplified L1 loci, shown by the presence of discrete bands rather than a smear of products on the gel (Figure 6.4.) In contrast, the L1 loci representation at single molecule level was good for the total gDNA from embryo three and in the blastomeres, except for E3B2-5, which had few distinct bands, and the gDNA controls. However, several blastomeres of embryo four,

as mentioned above, showed many L1 loci to have dropped out during the WGA process.

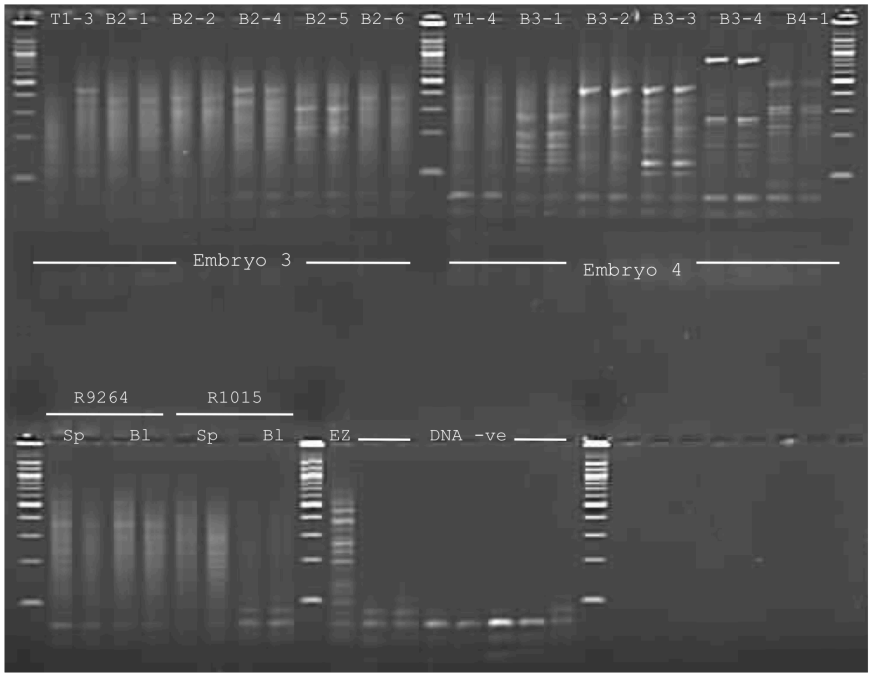

Figure 6.4 Secondary PCR (RVECPA2+RV5SB2) of Embryos 3 and 4 (*Vsp*I libraries), top right set of wells: Embryo 3 total DNA (T1-3), and its five blastomeres (E3B2-1 – E3B2-2 and E3B2-4 – E3B2-6), in duplicate, Top left set of wells: Embryo 4 total DNA and its five blastomeres (E4B3-1 – E4B3-4 and E4B4-1) in duplicate. Bottom left set of wells: R2908 and R1015 (healthy donors) sperm (Sp) and Blood (Bl) DNA in duplicates, last part; controls: lane 1: ligase –ve control, and lane 2-7: DNA –ve and PCR –ve controls.

Following the secondary PCR, the display PCR was carried out and the result is presented in Figure 6.5. The display gel also confirmed the frequently poor coverage and L1 locus drop out during WGA for the embryo four derived libraries. Five different L1 loci from the polyacrylamide display gel autoradiograph were randomly recovered from the gel and cloned and sequenced. All five sequenced loci belonged to groups of polymorphic L1 loci, which were also present in the hg19. Since the L1 loci were poorly represented in several blastomeres (mainly blastomeres derived from embryo four), it was decided not to use this embryo for NGS library construction. Based on this, 12 samples from three embryos were selected for the high throughput sequencing, as summarised in table 6.1.

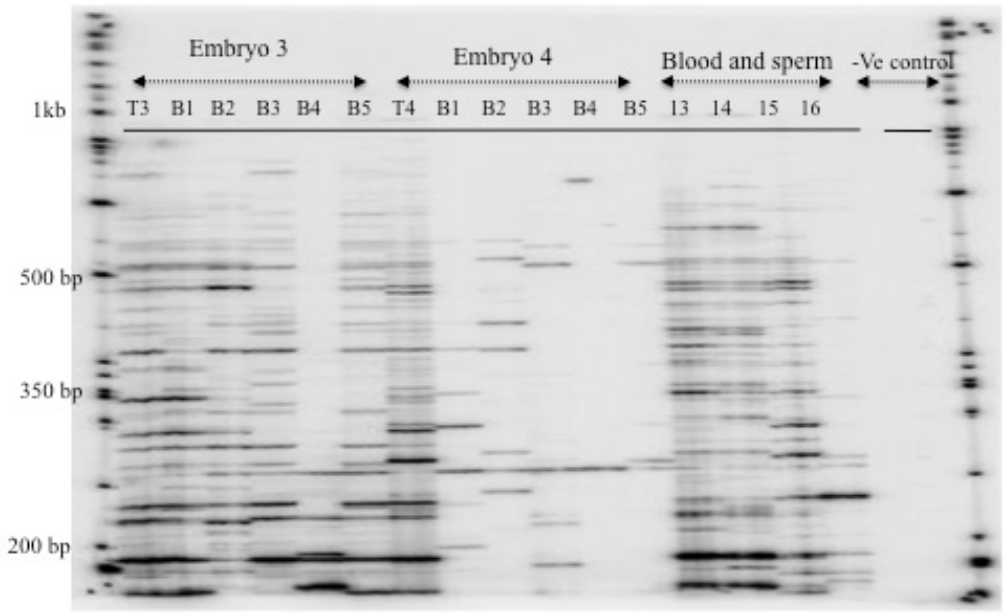

Figure 6.5 ATLAS display gel (*Vsp*I) library from embryo three total DNA (lane 1) and its blastomeres: lane 2: E3B2-1, Lane 3: E3B2-2, lane 4: E3B2-4, lane 5: E3B2-5, lane 6: E3B2-6, lane 7: embryo four total DNA (E4T1-4), lane 8: E4B3-1, lane 9: E4B3-2, lane 10: E4B3-3, lane 11: E4B3-4, lane 12: E4B4-1, lane 13: sperm (donor R2908), lane 14: blood gDNA (donor R2908), lane 15: sperm (donor R1015), lane 16: blood gDNA (donor R1015), -ve controls including secondary PCR –ve control and display PCR –ve control, gamma 32 labelled-100 bp ladder (Promega). As it is presented on the display gel some of the samples such as blastomere five of the embryo three (lane 4) and majority of embryo 4 blastomeres (lane 8-12) have poor L1 locus representation. This is probably due to locus drop out during the whole genome amplification process (section 2.2).

6.2.3. High throughput ATLAS experimental design

Prior to sequencing, the required sequence coverage was calculated to generate enough reads to be able to confidently characterise single molecule events, and a target sequence coverage of 11 reads per amplicon was identified as achievable.

To arrive at this number, the L1Hs Ta-specific oligonucleotides used in the primary suppression PCR were initially mapped to hg19 to identify ~3,000 L1 loci. Data from exhaustive fosmid sequencing (Beck *et al*. 2010) enables an estimate of the number of novel (*i.e.* not previously characterised) L1s per screened genome as between 4 and 6 per individual. Since this is a small fraction of the ~3000 oligo binding sites shared by the majority of human genomes (determined by *in silico* mapping) failing to account

for these in the coverage estimates will only result in a very small overestimation. By contrast the total number of polymorphic L1Ta elements (in any genome) is about 30% (Boissinot and Furano, 2000), making ~3000 L1 amplicons per average genome a substantial overestimate (as many insertions will be absent from a given genome). By this logic these assumptions can only lead to an underestimate of the coverage achieved. In the current protocol, the *Nla*III restriction enzyme was used to construct the genomic DNA library (more details on library construction in section 2.13). Knowing (from *in silico* digestion RM Badge, JM Rouillard *pers comm*) that about 80% of the human genome is within 1kb of a *Nla*III site, the number of accessible L1 loci for this experiment would be 80% of ~3,000, i.e. ~2,400 L1 loci, assuming an even distribution of L1 loci and restriction sites. Since for this study ¾ of a picotiter plate was used and the number of beads per region should be around 160,000 (according to the manufacturers data), the total number of the reads expected from all three regions is 160,000 × 3 = 480,000 reads. Fourteen libraries were sequenced, and so the number of expected reads per amplicon would be 480,000 / 2400 X 14, or ~11 reads per amplicon. Therefore, it was estimated that for a single molecule present in one library we should detect about 11 reads, but the coverage would be much higher for constitutive L1 loci present in all libraries.

Based on the calculation above, WGA gDNA libraries of the embryos and donor R2908 blood and sperm gDNA were constructed according to the protocol (section 2.13) and the amplicons were prepared for sequencing. To use the maximum capacity of the picotiter plate, all the secondary PCR products were pooled prior to sequencing. In order to separate the sequences according into their original gDNA sources, all the amplicons were tagged with multiplex identifiers (MIDs). The fusion primer experimental design used to incorporate these MIDs is explained in more detail below.

6.2.3.1 Designing the fusion primers

For high throughput sequencing experiments the amplicon length needs to be given careful consideration: although the 454 technology used can achieve much longer read lengths than competing technologies such as SOLiD and Illumina, this is

sequence dependent and highly variable. To extract maximum information the amplicons would need to be fully sequenced to enable their accurate mapping to the reference genome, and the determination of their structure. The estimated sequence read length of the 454 Sequencing System with the GS FLX Titanium chemistry used is about 450 nucleotides (from key sequence to key sequence), but some of the amplicons generated by the full-length L1 specific suppression PCR were bigger than 450 bp (up to 750 bp). We reasoned that 450bp bidirectional reads using primer A and Primer B should have produced sufficient overlap in the middle of the larger amplicon sequences such that entire amplicons were reliably covered. The sequencing keys and the template-specific primers (and the MID sequences) were included in the read length, but they were not part of the target sequence available for analysis. This had direct implications for the amplicon length. For example, for a read to cover an amplicon entirely, it must traverse its key (4 nucleotides) at the proximal end and the template-specific primers (20 nt) and MID sequences (10 nucleotides each) at both ends. This adds up to ~50 nucleotides from each side of the amplicon, and was considered during the experimental design. Hence for each gDNA source library a pair of forward and reverse fusion primers was designed, with the forward fusion primer (5′ to 3′ direction) consisting of a Lib-L chemistry A primer (Roche-454-NGS) followed by a CATG key, and a 10 bp barcode unique to the 5′ end of each of the samples, and finally the linker primer. Similarly, the structure of the reverse primer comprised a Lib-L chemistry B primer (Roche-454-NGS) followed by a CATG key and a 10 bp barcode unique to the 3′ end of each of the samples, and a L1 primer at its 3′ end (the sequences of the fusion primers and their melting temperature are listed in Appendix III, Table 4). All barcodes were chosen from the Roche-454-barcode list (Meyer *et al*., 2007). After the fusion primers were designed, each pair of forward and reverse primers was checked for their possible secondary and complementary structures using the OligoCalc software (an online oligonucleotide properties calculator) (Kibbe, 2007). Post-sequencing analysis showed that on average 95% of the reads were full length and had tagged MIDs on both sides. Reads were assigned to libraries by analysing the combination of MIDs in the read using Perl script 1 (Appendix VI).

6.2.3.2 Secondary PCR optimisation and pooling

Following from the protocol in section 2.13, a secondary PCR was carried out and optimised using the fusion primers. Details of the optimised PCR conditions are explained in section 2.13. In order to make sure that the fusion primers are intact and that they can recall the correct samples, secondary PCR products from each sample were cloned, colonies were randomly selected, and these were sequenced using the Sanger method. All the recovered sequences had intact forward and reverse fusion primers, and their MID sequence correctly identified the gDNA source for each tagged sample. 60% of the analysed sequences belonged to the L1Hs subfamily, and 30% of them were novel sequences in respect to their presence/absence in the hg19. All the analysed sequences and their accession numbers are listed in Table 6.2.

Sample	5′-MID	3′-MID	Correct seq. Id call	Sequence analysis result	Representation in hg19
E3T1-3	√	√	√	AC096710 (L1PA2)	Present
E3T1-3	√	√	√	AC010245 (L1PA2)	Present
E3B2-1	√	√	√	AL133320 (L1HS)	Present
E3B2-2	√	√	√	AC104689 (L1PA2)	Present
E3B2-3	√	√	√	AC004976 (L1HS)	Present
E3B2-4	√	√	√	AL354976 (L1HS)	Present
E3B2-6	√	√	√	AL035459 (L1HS)	Absent
E6T1-6	√	√	√	AP001876 (L1HS)	Present
E6B5-1	√	X	√	AL008729 (L1HS)	Absent
E6B5-1	√	√	√	AC140658 (L1HS)	Absent
E6B5-2	√	√	√	-	No trace of L1 sequence
E6B5-3	√	√	√	AC004976 (L1HS)	Absent
E6B5-4	√	√	√	-	Poor sequence quality
E6B5-5	√	√	√	-	Unmmapable
R2908-B	√	√	√	AC09733 (L1HS)	Present
R2908-S	√	√	√	-	Unmappable

Table 6.2 Sequence analysis of embryo - WGA DNA amplicons. 454-secondary PCR amplicons were re-amplified with A and B 454 sequencing primer and cloned and sequenced with M13F. As shown in the table, except for sample E6B5-1, which was 3′ truncated, all the rest of the sequences were full length (had both primers A and B), and all the MIDs were correctly called relative to their original DNA sources for each sample.

Prior to sequencing an equi-molar concentration of the secondary PCR products from each library were pooled together. For secondary PCR quantification all the individual samples were subjected to picogreen analysis (Ahn *et al*., 1996) as well as

run on an Agilent Bioanalyzer (Invitrogen). Then an equal molarity of each sample was pooled except for the two samples E4B3-1 and E4B3-3, for which a 100 folds dilution was used for pooling. The reason for pooling many fewer molecules of these two libraries was that during quantification of their secondary PCR products using the Agilent system it was noticed that there were not many L1 loci contributing to these libraries and therefore by lowering the number of molecules 100-fold it would still provide enough coverage to analyse the existing L1 loci in these two samples, without sequencing them at wastefully high coverage.

6.2.3.3 Pooling the libraries

The ATLAS secondary products contain a range of PCR products with variable lengths of 200-800 bp. To minimise length biasing during the emulsion phase PCR (emPCR), pooled libraries were divided into three size-fractionated batches, with different ranges of amplicon length. Each batch was sequenced separately on three physically separate regions of a picotiter plate. The first batch (pool number one) contained the original full range (200-750 bp) of products to investigate if there is any amplicon length biasing during sequencing. The second batch (pool number two) contained smaller amplicons ranging from 200-350 bp. The third and final batch contained the longer length amplicons ranging from 300 to 750 bp. The lower and upper range products were made by fractionating the pooled secondary PCR products on an agarose gel and extracting the lower range sizes >350 bp and the upper range products <300 bp. 50bp overlaps were allowed between the upper and lower ranges to avoid losing any products of 300-350 bp (the procedure is explained in more detail in section 2.13). The three sets of size ranges are presented in Figure 6.6. Sequencing was carried out by the NUCLEUS sequencing service (Department of Genetics, University of Leicester) following the manufacturer's protocol, with the emPCR Lib-L SV kit.

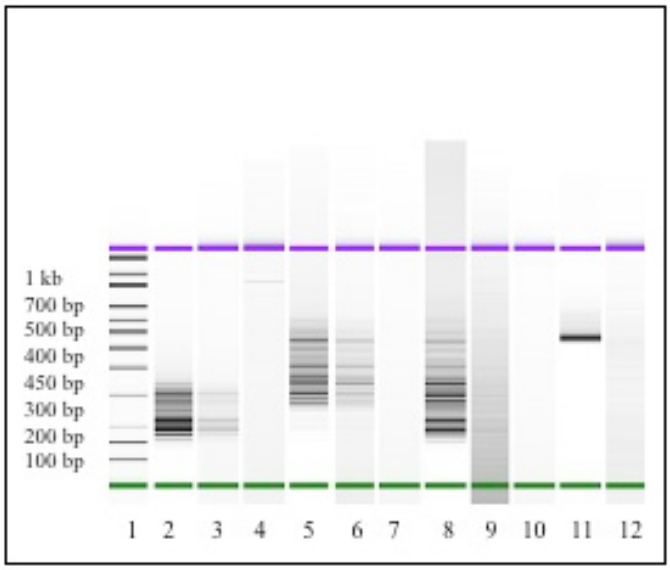

Figure 6.6 The Agilent bioanalyser (high sensitivity DNA kit, Invitrogen) for three sets of pooled amplicons with different size ranges. Lane 1: Marker, lane 2: lower range product size (200-350 bp), lane 3 and 4: controls, lane 5: upper range product size (300-750 bp), lane 6 and 7: controls, lane 8: whole range of product size (200-750 bp), and lane 9,10, 11, and 12: controls.

6.2.4. Post-sequencing data processing

The post-sequencing image data was processed using two independent pieces of software: the Roche 454 amplicon processing software (version 2.3), and the shotgun data processing software (version 2.3.1). Following this, the data from the two different software packages were compared to select the data set with the higher number of full-length amplicons (the raw numbers are compared in Appendix IV). Overall the amplicon processing software yielded 438,287 reads with an average length of 191 bp, whilst the shotgun data resulted in 685,444 reads with an average length of 200 bp. This suggested that our amplicon library constructs were best suited for analysis with the shotgun data processor, with the amplicon processing method's filters being too stringent for this study. Indeed, comparing the number of full-length sequences processed independently by the amplicon and shotgun data processing software showed a significant increase in the number of the full-length amplicons (key to key). The result of this comparison is summarised in Figure 6.7.

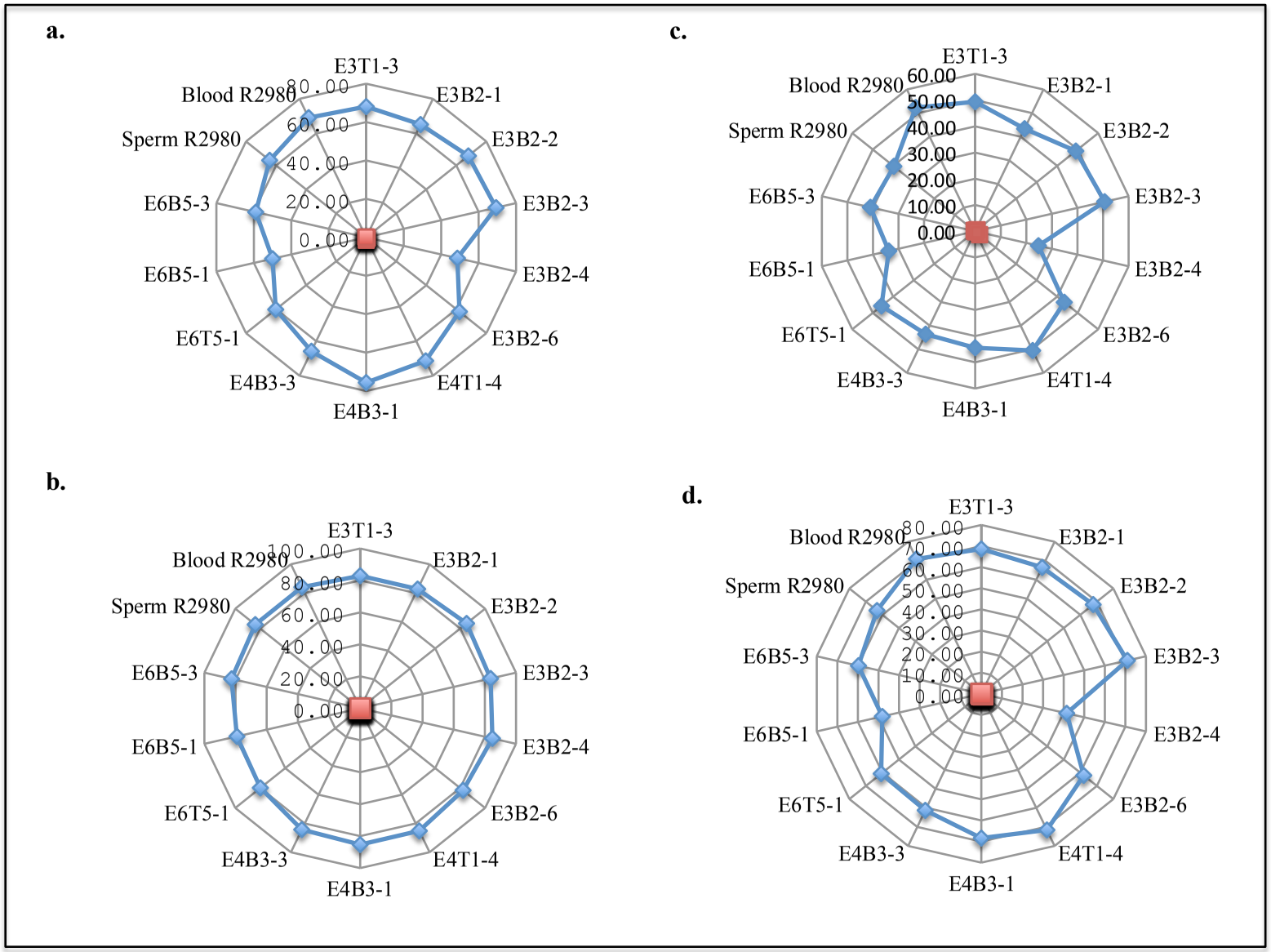

Figure 6.7 Comparison (of the number of the full length amplicons filtered through using the Roche version 2.3 post analysis) between Amplicon Processing (red) and shotgun data processing (blue) software, **a. picotiter plate-region one:** Pooled amplicons with size range of 200-750 bp, **b. picotiter plate-region two:** lower range of amplicons 200-350 bp, **c. picotiter plate-region three:** upper size range of amplicons 300-750 bp, **d.** all three regions together.

Further to this, some of the parameters and filters for the shotgun processing were adjusted, for example the Valley fillter was relaxed from default (eight) to four. This filters or trims reads with many off-peak signal intensities. A Valley flow is defined as an intermediate signal intensity, i.e., a signal intensity occurring in the valley between the peaks for 1-mer and 2-mer incorporations, or the 2-mer and 3- mer, etc. The signal distribution of all reads of the Run is used to define the peaks of the homopolymer incorporations relative to these, the valleys or borderline zones for classification of intermediate signals (Genome Sequencer FLX System Software Manual, version 2.3). However, no significant differences were observed in the number of reads surviving filtering. This suggests that the shotgun processing parameters were set at their optimum setting to filter the sequences for this project and that no further adjustments were required.

6.2.5. High throughput sequencing length biasing

As explained earlier, amplicons were sequenced according to their size ranges in three physically separated regions on a picotiter plate. Region one comprised the whole size-range of amplicons (200-700 bp), the second region comprised the smaller range of amplicons (200-350 bp), and finally the third region comprised larger sized products (300-750 bp). The post-sequencing read length of amplicons from across all three regions were measured for each library, the resulting number and the average read length for each region is summarised in figure 6.8. As demonstrated in the figure the average read length of each region corresponded to the range of products length in each region, *i.e.* region one has the intermediate length of products, while the average read length for the second region is the smallest among the three regions and the average read length for region three is higher than the other two regions for all the libraries. The result suggests that there is length biasing, as expected, in favour of the smaller size amplicons. If there was no biasing the same average length for the fractionated and non-fractionated sets was expected, as this was a function of the length distribution of the amplicons alone. As a longer average read length for the larger size fraction (region 3) was observed, despite the fact that the whole range fraction (region 1) contains these same amplicons, this suggests that the full range fraction is suffering a differential sequencing bias towards smaller amplicons.

6.2.6. Homopolymeric G tract length analysis

Sequence traces from all the reads of all three regions were analysed to reveal the length distribution of the homopolymeric G nucleotide tract present at the 5´ terminus of full length L1s. A Perl script (script 2, Appendix VI) was used to classify each read containing the terminus of an L1 by its G tract length, between 5-10 G nucleotides. The length of the poly-G tracts in the 454 reads was compared to that of a database of full length human specific L1 sequences mined from online sequence databases (baseLINE Hastings 2009 PhD thesis) and is presented in Figure 6.9.

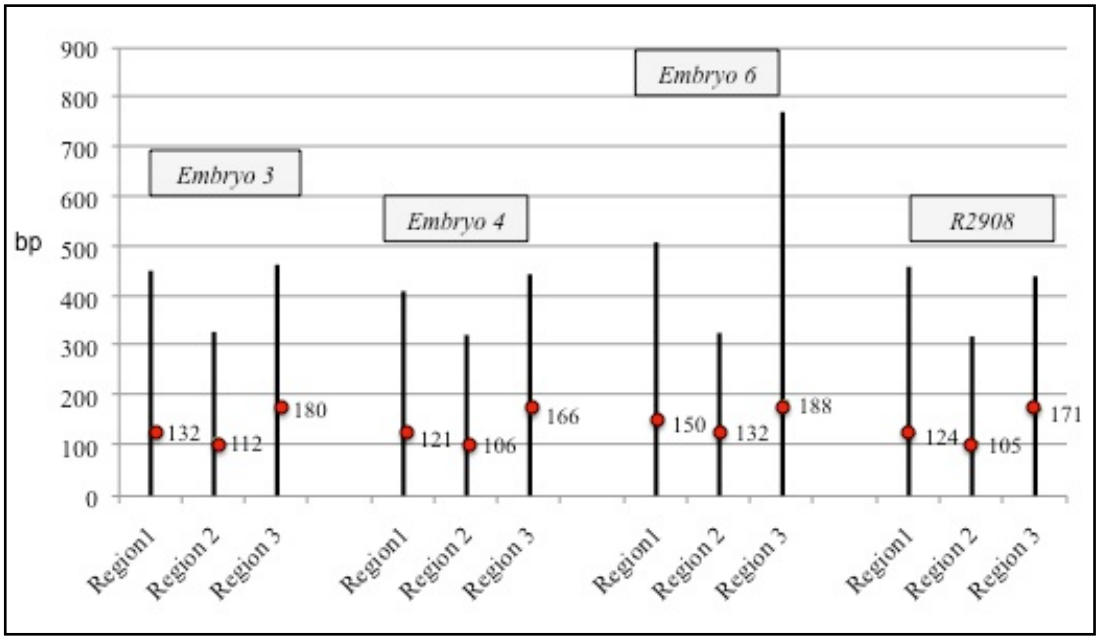

Figure 6.8 Length of the full-length amplicons in bp (from key to key) (y axis) from pooled products on each region of the picotiter plate. Region 1: whole range of amplicons (200 -750 bp), Region 2: lower range of amplicons (200 – 350 bp), Region 3: upper range amplicons (300 – 750 bp). The red dotes demonstrate the average read length for each region.

In figure 6.9 the Y-axis plots the proportion of sequences in the dataset with G tract lengths of 5-10bp, shown on the X-axis. Overall the proportion of sequences at each G tract length is quite similar for the two datasets, suggesting that the known inaccuracy of the 454-technology when sequencing homopolymeric tracts is not

biasing the data. Indeed five random sequences of each of the poly-G repeat tract lengths (5-10 bp) were manually mapped to hg19 and checked to see if the 454 read faithfully represented the genomic sequence determined by Sanger sequencing. The result revealed that for the longer repeat (ten and nine poly G), most of the G tracts were participating in L1 5′ TSDs and 74% of the total analysed sequences were novel L1s (absent from hg19) with long TSDs. One case in the ten poly G track was a simple tandem G repeat which was not present in the hg19, located 10 bp upstream of them L1. For the poly G tracts ≤ 8 most of the G tracts were either part of the novel L1s with long G tracks (8-7 G repeats) or for some of the G tract it was partly from L1 and the genomic flank sequences. Interestingly 100% of analysed sequences, associated with long poly G tracks belonged to L1Hs (young L1s).

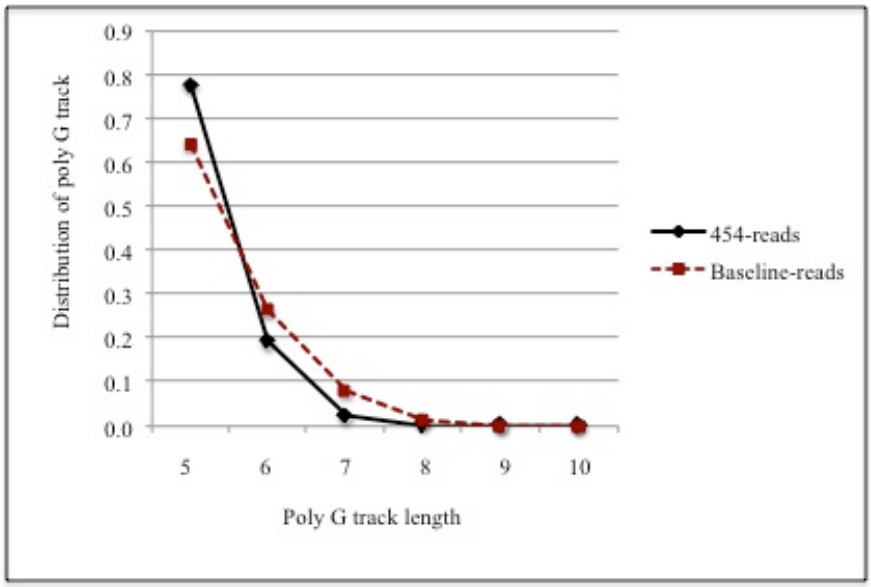

Figure 6.9 Poly G tract length distribution in 454 reads (black) compared to the baseLINE (baseLINE Hastings 2009 PhD thesis)(red). Both data sets have similar proportions of elements with each length of poly G tract analysed.

It should be noted that the baseLINE sequences comprise full-length L1 sequences that were extracted from the human hg18 reference and Celera assemblies and the 2008 version of GenBanks nr nucleotide division. All elements show ≥ 98.5% identity at the nucleotide level to L1.3 (L19088) and are at least 5900bp long. The results are presented in Figure 6.9. The result for the unbiased and probably active L1 data set also showed a wide range of poly-G tracts upstream of the L1-5′ UTR.

6.2.7. Processed data analysis

The result of post-sequencing signal processing with the shotgun data software pipeline was 685,444 reads with an average length of 200 bp. Following this, the sequence reads were analysed to identify novel polymorphic and *de novo* L1 sequences, using the approach summarised in Table 6.4. Briefly, the bulk reads from all three picotitre plate regions were separated according to their MIDs into their libraries using Perl script one (Appendix VI), and then all the sequences introduced by PCR (linkers, fusion primers) were trimmed off each read from both sides except for the sequence of the L1 specific primer (RV5BS2), which was retained in the sequence structure. To map the sequences to the human genome reference sequence, all the reads, which belonged to each individual were analysed using the public instance of the Galaxy web service: http://main.g2.bx.psu.edu/ and several Perl scripts (Appendix VI). All the reads from the different libraries for each individual DNA source (embryo or R2908) were combined and filtered to remove sequences less than 30 bp. These filtered sequences were subsequently mapped to hg19 using the LastZ tool at Galaxy. The number of unmappable reads was not high (average < 5% for each independent library). After mapping all the sequences to hg19 to obtain genomic coordinates for each read, all the sequences belonging to each individual were compared to the L1 oligo data set (this data set was generated by mapping the L1 oligo primer RV5BS2 in hg19) to see if their map coordinates overlapped. This analysis showed that 67% of reads corresponded to L1 loci already present in the hg19. The 23% of sequences, which were absent from hg19, were further compared to the Badge lab oligo dataset to determine if they corresponded to previously isolated, but unpublished L1s. The Badgelab oligo dataset is derived from mapping all the L1 loci specific primers in the lab database to hg19 and includes novel L1 insertions that have been previously characterised by members of Richard Badge's group at the University of Leicester and which are generally absent from hg19. The result revealed that about 6% of traces fell into genomic intervals covered by the Badge lab oligo dataset intervals. Therefore, to validate the *in silico* genotypes of the 454 sequences, seven L1 loci from the Badge lab oligo dataset which appeared to intersect with 454 sequences were randomly selected, to compare their *in silico* genotypes with

experimental genotyping, as illustrated in Figure 6.10. If an embryo is +ve by *in silico* genotyping it can only be Homozygous Present or Heterozygous for the insertion: this was consistently observed for all of the +ve *in silico* genotypes (16/16) and is summarised in figure 6.10. Moreover, two of the –ve *in silico* genotypes (E4-AC004051 and E3-AL050308) were Homozygous Present and Heterozygous by experimental genotyping respectively. This result could still be a consistent genotype, since it is possible that the insertion simply failed to amplify prior to the sequencing. Moreover, the *in silico* and the experimental genotyping data showed that we could recover single molecule events. For example, in one case in embryo four the genotyping showed that this individual is heterozygote for AL050308 insertion. *In silico* genotyping result revealed that this sequence can be recovered from the total blastocyst DNA as well as the E4B3-3 blastomere. Since we could recover a heterozygote insertion from a blastomere (a single cell) this means that our technique has the capacity to amplify single DNA molecules.

L1 -loci	*Novel insertion*	*In silico genotyping*				*In vitro genotyping*				*Consistency between In vitro and In silico data*
		E3	*E4*	*E6*	*R2908*	*E3*	*E4*	*E6*	*R2908*	
AC067958	N	-	+	-	+	HmA	Hmp	HmA	Hmp	Consistent
AL050305	N	+	+	-	-	Het	Het	HmA	NA	Consistent
AC009413	N	-	-	+	-	HmA	HmA	Het	HmA	Consistent
AC004051	N	+	-	-	+	Hmp	Hmp	NA	Hmp	Consistent
AL050308	Y	-	+	-	+	Het	Het	HmA	Hmp	Consistent
AL008729	N	+	+	+	+	Het	Het	Het	Het	Consistent
AL005935	Y	+	+	+	-	Het	Hmp	Hmp	HmA	Consistent

Figure 6.10 comparison between the *in silico* and *in vitro* genotyping of seven different L1 loci, in silico typing presented by +/- showing the present/absent of the genomic flank sequence for each L1 loci in each individual embryo and R2908 gDNA, the in vitro typing, involved PCR amplification of both empty site and filled site: HmA/P: Homozygote Absent/Present, Het: Heterozygote, comparing the result of *in silico* with the *in vitro* genotyping shows consistency for all of the genotyped L1 loci.

6.2.8. Characterising candidate novel and *de novo* L1-mediated retrotransposition events

Following on from the analysis above (section 6.2.7), all the L1-related sequences that were not present in hg19 or in the Badge lab dataset (22% of the total sequences) were analysed further. Intially, all sequences derived from individual embryos (i.e. their total DNA and their blastomeres) and the donor R2908 (blood and sperm) were aligned using the MUSCLE multiple alignment software (Edgar, 2004). The clusters of sequences resulting from automated alignment were manually analysed using the Jalview software to identify groups of traces with an average number of reads per cluster of 10. This threshold was used as it approximates the expected coverage of single molecule derived amplicons, based on the predicted number of individual amplicons (~2400) and the number of 454 traces generated. Each cluster was further checked manually for the presence of a *Nla*III site and a correct L1 primer site. After this check an average of 10% of the sequences were discarded from each set of sequences. The clusters, identifying distinct amplicons were compared to a range of published L1 data sources and software, including hg19 using BLAT at UCSC, Repeatmasker, fosmid L1 sequences (Beck *et al*., 2010), dbRIP (Wang *et al*., 2006) novel L1 sequences discovered in HTS data from the 1000 genomes project (Ewing and Kazazian (2010), baseLine (Hastings and Badge, unpublished annotation database). In total, 97 novel sequences were verified from the previous studies above, and all were confirmed to be absent from the human genome reference sequence. Figure 6.11 shows a 454 L1 sequence trace, which co-locates with the 5′ flanking DNA of a known polymorphic L1 element that is absent from hg19 but previously reported by Kidd *et al.* (2010) through end sequencing of fosmids. This example illustrates how the combination of the ATLAS technique and high throughput sequencing can selectively amplify full length L1Hs (containing L1 Ta1d specific sequence variants).

In total 172 sequences (sequence coordinates are listed in Appendix VII) across all the libraries were absent from all the available L1 databases as well as the human genome reference sequence, and therefore they are introduced as candidate novel L1 retrotransposons, which have been discovered through this study.

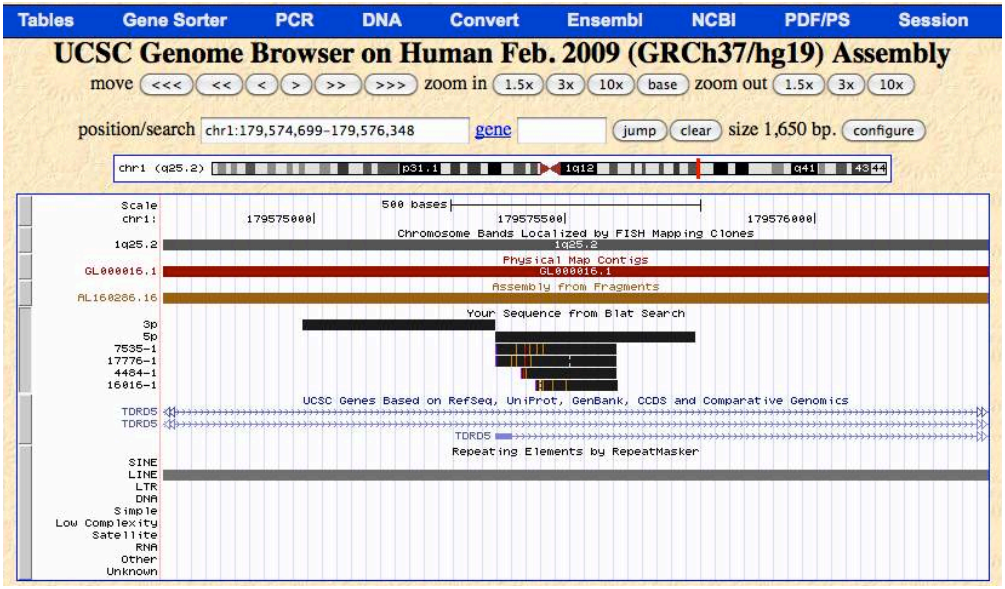

Figure 6.11 ATLAS and high throughput sequencing can capture known polymorphic L1 insertions The screen shot above shows the result of a BLAT search using 454 traces (numbered black rectangles) that co-locate with the 5´ flanking (black rectangle labeled "5p") DNA of a known polymorphic L1 element that is absent from hg19. This novel L1 insertion previously reported by Kidd *et al*. (2010), and it is recorded as RIP: 2000532 in dbRIP.

The traces of all candidate novel L1 insertions from above were further investigated by *in silico* genotyping of each insertion in BLAST against other blastomeres or DNA sources. The bioinformatics steps for analysing these sequences are summarised in figure 6.13. Through this investigation, 47 of the sequences were revealed to be present in the blastomeres as well as the remaining blastocyst DNA of the same embryos. Therefore these insertions are candidate novel L1 insertions discovered in this analysis but they are not *de novo* insertions since they are present in the whole blastocyst gDNA of the embryo. From the remaining 125 sequences, 57 sequences occurred only in blastocyst gDNA. These sequences are also candidates for novel L1 retrotransposition events, as they apparently are absent from all the individual blastomeres of an embryo, but present in the remaining blastocyst DNA. While we cannot exclude the possibility that these sequences results from locus drop out in the blastomere libraries, sampling of a somatically mosaic embryo is another feasible explanation.

Interestingly 14 of the 125 sequences derive from bulk DNA from blood and sperm gDNA of a single anonymous male donor (R2908), and therefore are less likely to be *de novo* insertions, but this suggests that this donor contains a significant number of

previously uncharacterized full length L1 insertions. Overall we conclude that 54 L1 related sequences, which found in a single blastomere and for which similar traces were absent from the remaining blastocyst DNA of the same embryos are candidate *de novo* L1 insertions. However, these must be validated by PCR to be able to confirm the *de novo* L1 retrotransposons. Figure 6.12 demonstrate an example of our candidate *de novo* L1 retrotransposition, which occurred in a single cell of embryo six and we could not find any similar sequence trace in neither the other blastomere nor the remaining blastocyst DNA of embryo six.

Figure 6.12 query sequence 32739-1 from embryo six blastomere E6B5-1; the blat search result at UCSC genome browser showed that the genomic part of the sequence matches with the genomic flank in chromosome eight, but part of the sequence does not match. The result of the repeat masker analysis assigns the unmappable part to a L1HS. The *Nla*III site is presented in blue capitals in the sequence. This insertion is absent from other L1 datasets, and so could be *de novo* L1 retrotransposition in a single cell of this embryo.

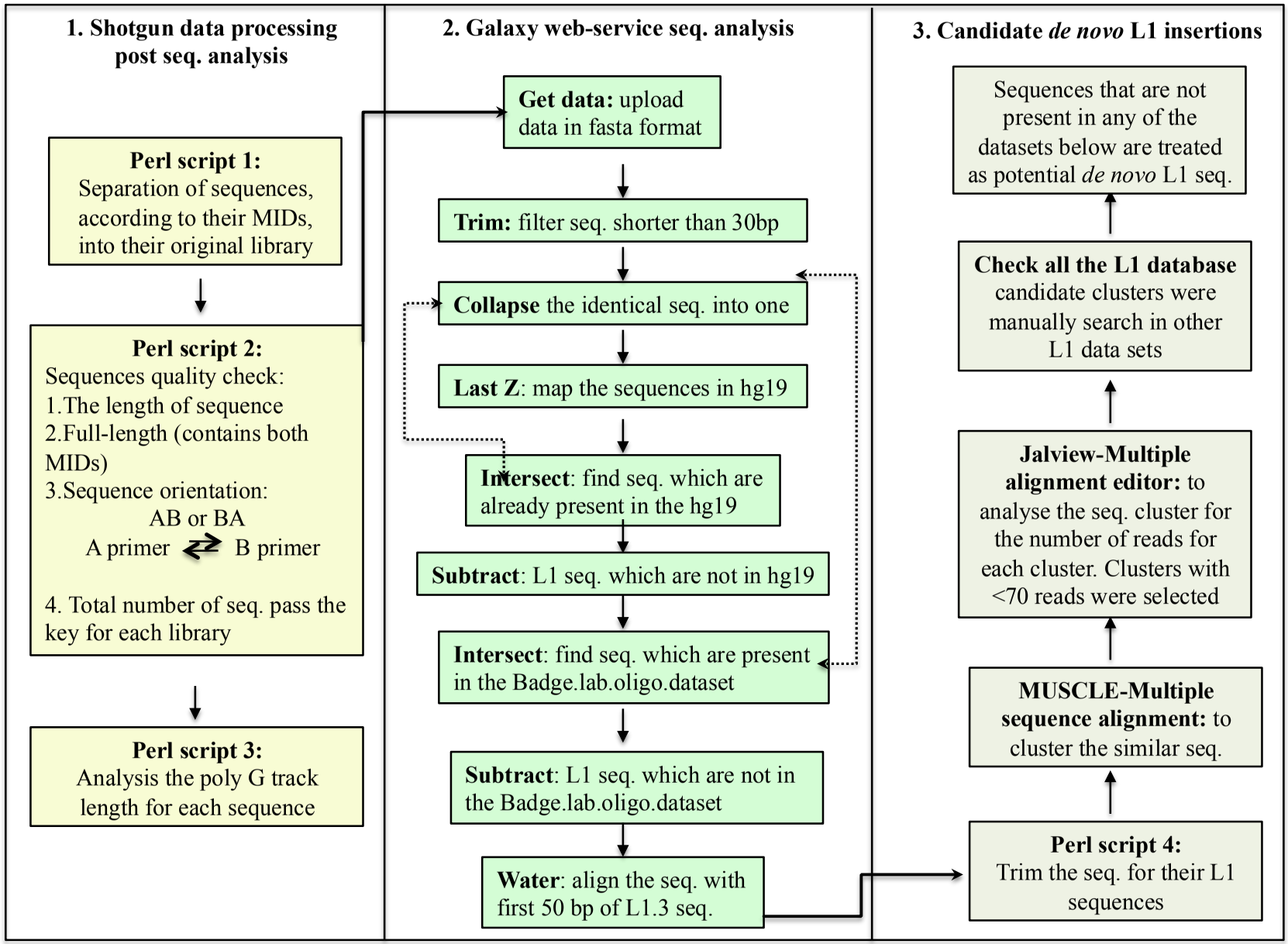

Figure 6.13 flow charts of the bioinformatics steps, which were used to analyse NGS data for the current experiment

For R2908, each cluster was queried individually in sperm-derived sequences versus blood-derived sequences to check for insertions, which might be specific to the germline (sperm) or the somatic (blood). One sequence in R2908 was identified which is specific to sperm DNA, and similar traces could not be found in the blood DNA of the same individual. Therefore, it is likely that this insertion is germline specific and could be the result of a *de novo* insertion in a sperm cell, or low-level germline mosaicism.

6.3 Discussion

The advent of high throughput sequencing technologies has lead to a higher resolution view of human genomic structural variation. However, the activity of endogenous L1 retrotransposition is still poorly understood, and only very recent studies leveraging these new technologies have begun to shed some light. As mentioned in the introduction section of this chapter, several studies such as those by Beck *et al*. (2010), Ewing and Kazazian, (2010), Kidd *et al*. (2010) and Iskow *et al*. (2010) have discovered a considerable number of novel L1 retrotransposons in the human genome that are absent from the human genome reference sequence. Moreover, Beck *et al*. (2010) demonstrated that more than half of the novel L1s they discovered were highly active in cell culture based retrotransposition assays. Therefore it can be concluded that many novel L1s are currently active in the genome and that their frequency and activity has been systematically underestimated. Despite all the evidence that L1s are still active and that they must actively transmit new copies to future generations, there is as yet no direct evidence of endogenous L1-mediated retrotransposition events either in gametogenesis or during human embryogenesis, apart from exquisitely rare disease-causing insertions, such as those at the CYBB and CHM loci (Brouha *et al*., 2002, van den Hurk *et al*., 2007). One of the possible explanations for failing to observe such events is the low rate of L1 retrotransposition at the early stages of human development, which might be due to stringent suppression host defence systems. Another reason is the lack of robust and effective techniques to isolate these rare insertions from the human genome. In chapter three of this thesis I have introduced a LC-ATLAS technique, which is sensitive and therefore suitable for detection of rare/single molecule L1 retrotransposition events. However, due to the

low rate of L1 retrotransposition, all the display techniques for characterising *de novo* L1 insertions require intensive screening with unlimited genomic DNA resources, which is not available for many samples, such as human embryos. In this chapter we described the development of a combined ATLAS display and high throughput sequencing technique, which can selectively amplify and sequence full length and likely active L1 elements in depth. It has been demonstrated that this combined technique is able to isolate rare/single molecule insertions, and we have identified candidate germline specific L1 insertions as well as potential *de novo* endogenous L1 insertions, for the first time in human embryos. This is discussed in more detail in the following sections.

6.3.1 L1 locus representation in human embryos and blastomeric WGA DNA

As is represented in the results section 6.2.1, embryos three and four were female and embryo six was male. Before any high throughput experiment were started, the quality of the WGA was checked to make sure that the genomic DNA was at a satisfactory level for next generation sequencing. Since during the WGA process (and especially from single cells) there is a high chance of genomic locus drop out, the embryonic WGA DNAs were tested by genotyping ten randomly selected L1Hs loci using the remaining blastocyst DNA from each embryo. All the genotyped loci were presented in all the individual genomes, and several L1s were dimorphic (present/absent) at several loci. Prior to NGS we further tested the L1 loci found at the single molecule level in all the human embryo libraries, by carrying out single molecule ATLAS (section 2.7, Macfarlane *et al*., unpublished data) and we compared this with genomic DNA from sperm and blood. The results of the single molecule dilution PCR for four independent L1 loci revealed that some of the blastomeres, especially for embryo four, showed poor L1 locus representation at the single molecule level. A likely explanation for poor L1 loci representation in several blastomeres is that these loci had failed to amplify during the WGA process.

6.3.2 High throughput ATLAS

Intially it was decided to sequence all the blastomeres and the total DNA from the female embryos (three and four) to avoid contamination between the male and female gDNA during the sequencing. However, as explained above, the result of the single molecule ATLAS revealed that some of the blastomeres in embryo four had a considerable amount of L1 locus drop out, and therefore two of the blastomeres and the remaining blastomere DNA of embryo four were selected for NGS as well as the remaining blastomere DNA and the two blastomeres of embryo six (male).

As discussed in chapter three the genomic coverage of the ATLAS technique can be improved to 80% of the genome by using the *Nla*III restriction enzyme for library construction. However, a problem arises, since higher genome coverage increases the display gel complexity and consequently it becomes harder to characterise single molecule L1 retrotransposon insertions. This problem was resolved in the current chapter by combining this high coverage display technique with high throughput sequencing. Based on our experimental design, for each single molecule event, an average of 11 reads was expected to be observed, in the NGS data. However, this estimation is based on our best assumptions, and could vary due to several reasons, such as a biasing towards the smaller amplicons (discussed later) at the emPCR stage. Indeed, as mentioned earlier we have recovered the L1-050308 from blastomere E4B3-3, for which the embryo (embryo four) is a heterozygote, demonstrating that this technique can recover rare/single molecule events.

6.3.3 Data processing

The amplicons were sequenced with Roche 454 Lib-L chemistry but the recommended amplicon data processing software was not suitable for this experiment. We have shown in the results section that the amplicon processing software was too stringent for amplicons with variable sizes, and tended to retain only sequences of intermediate length while filtering out the smaller and longer products. Thus it can be concluded that for similar studies the shotgun data processing method (Roche, version 2.3) is a better option, as the number of full-length sequences that could pass through

the filters was significantly higher than that seen with the amplicon processing software.

6.3.3 Amplicon length biasing during NGS

In this chapter we evaluated the significance of length biasing during sequencing As presented in the results section, there was a strong biasing towards the shorter products, and this biasing is more likely to occur during the emPCR phase since the smaller fragments can amplify more efficiently during emPCR. Based on this observation, to reduce amplicon length biasing during the sequencing for this technique, it is recommended to divide the amplicons into overlapping size ranges as explained in the results section.

6.3.4 Long homopolymeric-G tract associated with L1s

Investigation of the observed homopolymeric-G tract lengths demonstrated that it is more likely that these long poly-G tracks (up to ten G nucleotides long) are biologically real and do not represent systematic sequencing errors. One of the reasons for this conclusion is that the sequences did not terminate after the poly-G tract, and all the sequences containing long poly-G tracts had high sequence quality (Q=40). To validate this we have also analysed 500 non-redundant L1Hs sequences, which were extracted from the Celera and GenBank databases (BaseLINE by Hastings and Badge, 2009, unpublished annotation database). The result showed the same pattern of long poly-G tracts, although the maximum observed G repeats amongst these sequences was eight G nucleotides. Although the length of the poly-G tracts for 454 reads was more skewed towards five and six poly-G tracts than the database sequences (these had a higher distribution towards six and seven), the Chi square test between the two datasets showed that there is no significant ($P > 0.05$) difference in the distribution of poly-G between these two datasets. Although the biological effect of these long poly-G tracts on L1 retrotransposition is not known, from our data we can observe that the younger L1s tend to have longer poly-G tracts.

6.3.5 Candidate novel / *de novo* L1 mediated retrotransposition

All the reads were checked to see if they corresponded to any known L1 in hg19 in UCSC or other L1 datasets. The checking revealed 172 candidate novel L1 insertions, which were extracted from four human genomes (three human embryos and one adult donor) that had not been previously identified. All these insertions were absent from all the L1 datasets examined and all had a sequence depth of greater than six reads, suggesting they are unlikely to be artifacts.

One of the candidate L1s appeared to be a germline-specific insertion, as it could not be recovered from the blood sequences of the same individual (Chr12: 118,608,935-118,610,304). Therefore, it is likely that this insertion occurred after germ cell partitioning and perhaps during gametogenesis. Although it has been shown that germline specific insertions into specific target loci are very uncommon (Freeman *et al.,* 2011), this process is important for the transmission of L1 into the progeny.

As mentioned earlier 47 of the sequences were present in the blastomeres as well as the remaining blastocyst DNA of the same embryos. Therefore these insertions are candidate novel L1 insertions but likely not *de novo* insertions. From the remaining 125 sequences 57 sequences occurred only in the remaining blastocyst DNA of the embryos. These sequences are perhaps candidates for *de novo* L1 retrotransposition events, but we cannot exclude the possibility of widespread locus specific drop out in WGA blastomere libraries.

Moreover, 14 of the 125 sequences belong to bulk DNA from donor R2908, so are less likely to be *de novo* insertions, but could be novel insertions specific to this individual. Overall we suggest 54 L1 related sequences (each of which are present in single blastomeres and could not be identified in the remaining blastocyst DNA) as candidate *bona fide de novo* L1 retrotransposition events in single blastomeres, that occurred during human embryogenesis. However it should be mentioned that other possibilities cannot be ruled out at this time, as this analysis is only based on bioinformatics analysis which requires experimental validation but which is beyond the limited time of this PhD project. One of the possibilities is that rare somatic insertions were not sequenced by chance in the total DNA and dropped out during

WGA amplification of blastomeres. These possibilities can be addressed by locus specific genotyping of the remaining WGA DNA amplified from blastocysts and blastomeres that has been retained by our collaborator Dr Jose Garcia-Perez (University of Granada).

Chapter 7

Discussion

Most of our knowledge of the biology of L1 retrotransposons is based on the characterisation of disease-causing L1 insertions and *in vitro* analysis using cell culture-based retrotransposition assays. Consequently, until recently little was known directly of the endogenous activity of L1 retrotransposons in our genome. One of the main reasons is the lack of robust and sensitive techniques to study these elements mobilisation *in vivo*. However, recent advances in the application of high-throughput sequencing to the study of L1 retrotransposons have provided a foundation for our knowledge of endogenous L1 diversity and activity. A recent fosmid-based cloning study, by enabling exhaustive recovery of full length L1s elements from individual human genomes has revealed a large number of novel L1 retrotransposons, of which more than half were very active in cell culture-based retrotransposition assays (Beck *et al*., 2010). Although this study suggests that many L1s should be retrotranspositionally competent, a recent study by Ewing *et al*. (2011) concluded that the frequency of polymorphic L1 insertions is very low across different populations (1 in 200 individuals). Until very recently there was no direct evidence of endogenous L1 retrotransposition in the human genome, apart from disease causing insertions. However, Iskow *et al*. (2011) demonstrated that *de novo* somatic endogenous L1 retrotransposition occurs in human lung tumours. This evidence points to L1s being active during tumorigenesis, but it is not known whether the activation of L1s in these

cells are a cause or effect of their malignant phenotype. Moreover, a study by the research group of Geoff Faulkner recently validated 14 *de novo* L1 retrotransposition events in human brains, using L1 selective hybridisation and high-throughput sequencing (paper in press). Despite these findings it is still not known why L1 insertions accumulate in human neuronal cells, or what pathological effects (if any) these insertions cause. Nor has there been a direct demonstration of *de novo* retrotransposition in the germline or early human embryogenesis, with the exception of very rare insertions detected due to their overt pathology.

It is clear that L1 retrotransposition in early human development or in germlines is necessary for the evolutionary persistence of these elements. As selfish genetic elements with no known function, individual copies of L1 elements decay as pseudogenes, requiring the steady production of new retrotransposition competent elements at a rate greater than their loss. The extraordinary success of L1 elements in mammalian genomes, in terms of their census and longevity, indicates that their strategy is robust, but the details remain to be elucidated. Previously it was considered self evident that the potentially deleterious effects of transposable element (TE) activity should be confined to germlines, as somatic activity risked compromising host viability with no chance of benefiting the TE. Consistent with this idea, of the 19 disease-causing L1 insertion reported to date 18 were originally characterized as being of germline in origin. The one exception, an insertion in the APC gene that was associated with a case of colon cancer (Miki *et al.,* 1992) must have been somatic, although most likely occurred in a mucosal epithelial stem cell. Subsequent studies have shown that some apparently germline insertions may have resulted from cryptic somatic mosaicism. This possibility was confirmed by the re-analysis of an apparently germline insertion into the CHM gene (L1CHM) that led to a case of choroideremia. This case illustrated how mobilization during early embryogenesis or gametogenesis is an effective transmission route for L1 elements (van den Hurk *et al*., 2003 and 2007). Indeed only one of the disease-causing insertions has been unequivocally established as being germline in origin – the L1CYBB insertion that occurred during meiotic prophase I (Brouha *et al.,* 2002).

Freeman *et al*. (2011) suggested that L1s are more likely to be active during early embryogenesis after failing to discover any *de novo* L1 retrotransposition events in sperm DNA. In this study they estimated the rate of L1 retrotransposition in germ

cells to be very low (1 in 400 individuals) by exhaustive screening of sperm DNA at loci amenable to L1 insertion. Based on this estimation, investigating *de novo* endogenous L1 retrotransposition in early human embryogenesis requires a robust and sensitive technique to capture these rare insertions into the genome. Thus the main aim of this study was to investigate the activity of L1 retrotransposons in human embryogenesis and design experimental approaches to capture *de novo* endogenous L1 insertions in early human embryos, and tractable models.

As a model for human embryogenesis we used several different human cell lines that have embryonal characteristics, including hESCs, embryonal teratocarcinoma cells and germline tumours (such as NTeraD1 and PA1, testicular and ovarian embryonal teratomas), and human embryos. However, due to the limited resources of the latter samples (human embryos) the first two samples were mainly used to study different aspects of L1s such as their promoter methylation status, their sequence transduction capability, and for the design of a robust and sensitive High Throughput sequencing-based technique to capture rare insertions in limited gDNA resources. Following this it was hypothesised that if L1 retrotransposition is an active process during early embryogenesis then extended culturing of the above embryonal cell lines could potentially lead to accumulation of *de novo* L1 retrotransposition events in single-cell-derived clonal lines. Therefore, throughout this thesis we have screened embryonal progenitor cells as well as their derived clonal lines for *de novo* L1 retrotransposition. Initially the 5´-ATLAS technique (Badge *et al*., 2003) was used to capture full-length elements, which are more likely to be retrotransositionally active (as explained in Chapter Three). However, screening of the NTera2D1, PA1 and hESC clonal cell lines failed to find any variation amongst the single-cell-derived clonal lines. The genome coverage of the original ATLAS technique is about 11% (based on *in silico* *Ase*I restriction digest), but this is clearly biased towards AT-rich regions due to the recognition site (5′-ATTAAT-3′) of the *Ase*I restriction enzyme. Several studies have suggested that L1s predominate in AT-rich heterochromatic regions (Korenberg and Rykowski, 1988), but Moran *et al*. (1999) demonstrated that there is little, if any bias against genes (generally located in GC-rich DNA) as a site of L1 retrotransposition in cultured cells. It seems likely therefore that previous studies of biased L1 distribution reflected selective pressures that have affected L1 accumulation during genome evolution. In contrast cell culture-based retrotransposition assays can detect novel,

minimally selected L1 insertions and therefore reflect more accurately L1 integration sites in the genome. Therefore we have to consider the possibility that failing to observe a *de novo* insertion could be due to the restriction site bias of *Ase*I, and its low genome coverage.

To improve the coverage, genomic libraries were constructed with a restriction enzyme with more sites in the genome (*Nla*III) and a compositionally unbiased recognition site (5′-CATG-3′). In principle this increased genome coverage of the ATLAS display technique to ~80%. However, this modification also increased the complexity of the display, making the observation of low copy number/single molecule insertions less likely. To address this issue we used display primers that differentiated different subsets of amplicons to lower the display gel complexity, and in Chapter 3 we demonstrate that this low complexity display technique, with much higher genome coverage is suitable for screening for young L1 insertions. However, this approach failed to detect any *de novo* L1 retrotransposition events when applied to six human embryonic stem cell clonal lines and their progenitors. This observation could be due to the limited number of clonal lines analysed. Success perhaps would require the screening of at least another 160 clonal lines, based on recent estimates of L1 activity (1 in every 200 births, Ewing and Kazazian, 2010). However, this would a large number of clonal lines, which considering both the screening time of the technique and the available resources, was out of the scope of this thesis.

Based on the observations in Chapter 4, I decided to use another aspect of L1 biology to look for *de novo* L1 retrotransposition. The ability of L1 elements to transduce 3′ downstream flanking genomic DNA was described in detail in Chapters 1 and 4, and a number of highly active L1 lineages are characterized by this process. Importantly the use of PCR schemes that focus analysis on sequences associated with L1 activity dramatically reduces the complexity and presumably increases the sensitivity of display procedures. Therefore we used L1 transduction ATLAS (TS-ATLAS - Macfarlane *et al*., 2011, in preperation) to screen for *de novo* L1 retrotransposition events in embryonal cells and their clonal lines. We screened our genomic DNA resources for insertions from three young L1 loci: AC002980, LRE3, and RP. All three L1 lineages are reported as containing " hot" (highly active) members according to cell culture-based retrotransposition assays (Brouha *et al*., 2003). It was considered that since these three L1 lineages are polymorphic in human populations and have

highly active members it is more likely that they actively produce offspring insertions in the genomes that contain them. Using the TS-ATLAS technique would allow the detection of novel L1 offspring, which is more likely to be seen given these lineages can actively generate *de novo* insertion. This study led to the discovery of many novel L1 insertions, which were absent from the human genome reference sequence, but once again it was not possible to identify any *de novo* L1 retrotransposition events. Since *de novo* L1 retrotransposition could not be demonstrated with this technique, we suggest that display-based techniques require such extensive screening, especially for models of embryogenesis models, and the rate of L1 retrotransposition in gametogenesis (possibly early human embryogenesis) is so low (*e.g.* 1 in 400 cells (Freeman *et al*., 2011) that these approaches are impractical, as implemented.

To overcome this problem, in Chapter Five it was decided to approach the targeting of active L1 loci by studying their promoter status. As explained in the Introduction, the L1 promoter is robustly active in many tissues types, and rich in CpG dinucleotides. It was previously demonstrated that the 5mC modification of CpG can suppress L1 activity substantially (Hata and Skaki, 1997). Therefore, in Chapter Five I report the results of my studies of the methylation status of the L1 promoter in human embryonal cells. I hypothesised that if L1 loci are retrotranspositionally active then some of these should show correlated changes in the methylation status of their promoter, *i.e.* for the retrotranspositionally active L1 loci it was expected that they would have some degree of hypomethylation. If these potentially active promoters could be found it would be possible to target a smaller subset of L1s for screening again reducing the amount of material required to be screened.

As presented in Chapter Five, a genome-wide and locus-specific methylation analysis of L1 promoters was conducted. The result of the global methylation analysis using the COBRA-L1 assay (Chalitchagorn *et al*., 2004) showed that both placenta and hESCs had an intermediate state regarding their L1 promoter methylation status when compared to tissue DNA (>70% hypermethylated L1 loci) and malignancy-derived cells (<20% hypomethylated L1 loci). This experiment demonstrated that half of the L1 loci in the hESCs have a hypomethylated/semimethylated status.

In order to find out which L1 loci are hypomethylated on a genome-wide scale the MS-ATLAS technique was developed. This genome-wide methylation analysis

revealed that some states of hypomethylation in hESC clonal lines remained stable across all the clonal lines, in comparison to NTera2D1 clonal lines. It was concluded that the variation in genome-wide methylation in the clonal lines derived from NTera2D1 cell lines could be due to genome instability in these carcinoma cell lines. By contrast hESC clonal lines were stable with respect to their L1 promoter methylation status. The sequence of bands recovered as differentially methylated by MS-ATLAS of hESC clonal lines revealed that all the captured loci originated from young L1 families, supporting the logic of the approach.

To explore these results further bisulphite-sequencing assays were designed for four polymorphic L1 loci to assess the methylation status of the first 16 CpGs in the L1 promoter, across a variety of samples. Locus-specific methylation analysis revealed that the methylation status of L1s at the studied loci were hypermethylated. This was similar to the tissue DNA and all the +20 CpGs were heavily methylated. In contrast, malignancy-derived cells and the placenta showed a reduction in the methylation of CpGs, especially at the four transcriptionally important CpG dinucleotides. A comparison between the genome-wide and locus-specific L1 methylation status showed that somatic DNA is hypermethylated with respect to the L1 promoter at all times. This observation is unsurprising since L1s are expected to be retrotranspositionally inactive in differentiated cells. Also, there was a consistent reduction in methylation of CpGs in malignancy-derived cell lines, which would be predicted due to the general genome instability and global hypomethylation seen in malignant cells. Moreover, the placenta also showed a somewhat hypomethylated CpG status in both genome-wide and locus-specific methylation. This could also be due to the general genome hypomethylation in placenta, which apparently affects the L1 promoter. However, locus-specific and global methylation analysis of hESCs revealed hypermethylated L1 loci and an intermediate level of methylation respectively. One explanation for this observation is that this study reflects two L1 classes of promoters: one is associated with young and active L1s which are heavily methylated, and the second class of promoters occur in older elements which are fixed in the genome. As a result it is hypothesized that host defence systems effectively ignore ancient and inactive elements so they remain unmethylated. However, some of the loci studied, such as AC002980, have been demonstrated to be retrotranspositionally very active (Brouha *et al*., 2003), and it is very likely that these

loci are also active *in vivo*. In this case the methylation status may not correctly reflect the epigenetic modification of these potentially active loci. Recently it was demonstrated that 5mC modification can be further hydroxylated through the TET proteins resulting in 5hmC (5′ hydroxy-methylcytosine, Williams *et al*., 2011). It was demonstrated that the 5hmC and other downstream modifications are associated with transcriptionally-active regions and the level of 5hmC correlates with the level of expression, even for L1 in the mouse model (Ficz *et al*., 2011). Based on this, it is very tempting to speculate that these potentially active L1 loci are likely to be 5hmC modified, and able to bind to transcriptional factors that drive their expression. Techniques such as bisulphite modification and methylation-sensitive differential digests do not measure the 5hmC status of the CpGs. This requires further investigation using ChIP assays to pull down the 5hmC with the help of antibodies or by using deamination treatment that makes *Hpa*II sensitive to hydroxymethyl cytosine (Kinney *et al*., 2011). Therefore, one future development for this project could be to deduce if there are any epigenetic modifications that specifically regulate active L1 promoters.

Since in Chapter Five we could not isolate active L1s and potential *de novo* L1s by targeting their promoters, in Chapter Six it was decided to use the full potential of our modified ATLAS technique (Chapter Three) with high genomic coverage by combining it with high throughput sequencing. For this experiment embryos at developmental stages of +5 and +6 of embryogenesis were used. We sequenced WGA DNA from single blastomeres and the remaining DNA of the blastocysts of three unrelated embryos. As discussed in the results section of Chapter Six, different parameters, including the amplicon size range, sequencing direction, tagging primers for accurate DNA source identification and the post-sequencing software were tested to elucidate the optimum approach. In total we sequenced 12 different DNA samples from a total of three embryos, as well as the sperm and blood of a healthy adult individual as a control. 638,000 sequences passed the quality standards with an average read length of 190bp (key to key). We have applied different sequence analyses, which are explained in detail in Chapter Six. Through this study we found that young L1s are associated with rather long poly-G tracts of >5 nucleotides, and this was confirmed by analysing 500 young human specific L1s from baseLINE annotation database (Hastings and Badge, PhD thesis 2009) which also showed that

long poly-G tracts associate with L1 sequences. However, the impact that these long poly-G tracts have on L1 retrotransposition is not known. Moreover, we suggest that this technique (a combination of ATLAS with high-throughput sequencing) is a robust and sensitive technique to identify single molecule / *de novo* L1 retrotransposition. This was demonstrated by the frequent recovery of known heterozygote L1 insertions from human blastomeres, which must result from the isolation and amplification of single DNA molecules.

In addition we have identified a novel insertion, which only occurred in a library derived from the sperm gDNA of healthy adult donor, and was absent from a library constructed using his blood gDNA. Based on this observation it can be suggested that this is a germline-specific full length L1 insertion, which occurred after germline partitioning in the early stages of embryogenesis. Moreover, we have discovered 172 candidate novel L1 retrotransposition events, which are absent from the human genome reference sequence and also have not been reported previously. Most importantly, 54 out of these novel insertions are suggested to be candidate sequences for *de novo* L1 retrotransposons. This is potentially a direct demonstration of *de novo* endogenous L1 retrotransposition in single cells of human blastomeres at an early stage of embryogenesis (stage +5 to +6). Thus we may have demonstrated for the first time that *bona fide de novo* endogenous L1 retrotransposition is a feature of normal early human embryogenesis. (*i.e.* is not just associated with disease-causing insertions).

All the above findings are based on sequence analysis and bioinformatic data validation. However, experimental validation including the design of PCR assays to amplify both ends of candidate *de novo* insertion from the identified blastomere source is required to reveal the insertions target-site duplications. Only this evidence will enable us to conclude that these insertions are the result of genuine TPRT based L1 retrotransposition. These assays can also show that the insertion is not amplified from other blastomeres within the detection limits imposed by the amount of DNA available (WGA of single cells). In order to eliminate contamination issues during the PCR validation of the candidate *de novo* L1 insertions it is necessary to perform this genotyping remotely in the laboratory of our collaborator, Dr. Jose Garcia-Perez, in Spain. This PCR validation will be carried out using an aliquot of the WGA gDNA

from the same embryos, retained by Dr Garcia-Perez prior to genomic library construction.

Therefore, at this stage we cannot exclude possibilities other than *de novo* L1 retrotransposition. One possible explanation is that the candidate insertion exists in all or a subset of the blastomeres but was only recovered by chance from one of them. This will easily be established by PCR genotyping although *a priori* the probability of such extreme amplification bias is low (assuming a Poisson process, with an average amplicon coverage of 11X, $p=e^{-11} = 1.7 \text{ X } 10^{-5}$). Another possible cause is that the candidate *de novo* L1 insertions are germline insertions transmitted by the parental gametes, but which confer some selective disadvantage during cell division resulting in their low representation the embryo. Low but non-unique representation could be investigated using single-molecule PCR techniques (Chapter 2.) to determine accurately the representation of the filled and empty site insertion chromosomes. However as the DNA of the embryos's parents is not available (embryos were donated anonymously for research use, on this basis) this possibility cannot be distinguished from insertion during embryogenesis.

The transcriptome analysis of L1ASP and Alu elements in hESCs by Macia *et al*. (2011) has revealed that most of the expressed Alus are from the youngest subset of Alus and interestingly most of the expressed L1s are found in genes. This study suggests that there is an element of epigenetic regulation acting on L1 elements to control their activity. However, potentially we have demonstrated that endogenous L1 retrotransposition is ongoing during human embryogenesis, which raises questions about the role of host defence mechanisms in keeping these elements silenced, especially in the important stages of embryogenesis. Activity at this stage would appear to be doubly dangerous for the host: not only risking damage to the embryo, but greatly increasing the chance of transmission of deleterious insertions. Several possibilities can be explored to try to explain this paradox. One is that L1s are mutagenic in the short term, but in the course of evolution can become useful. However this explanation does not fit a logical framework; if they are mutagenic in the short term, then how can they survive to be useful? Another speculation is that since L1s can evolve faster than the host genome due to their high copy number, it is possible that they have gained the ability to escape host genome defence mechanisms, or alternatively can exploit the host genome to enhance their survival chances. Indeed,

it has been demonstrated that the L1ASP pre-mRNA can bind to AGO2 proteins, and the protein complex is able to suppress an oncogene upstream of the L1 insertion (Aporntewan *et al*., 2011). Moreover, it has been suggested by Singer *et al*. (2010) that new somatic L1 insertions in human neuronal cells could affect neuronal plasticity and behavior, and hence they have speculated that L1-induced neuronal diversity could extend the spectrum of behavioral phenotypes. Based on these findings it could be speculated that L1s, or a subset of them, have gained regulatory roles in the genome and therefore either they are not being repressed, or host defence systems regulate them in a different way.

There are many aspects about the biology of L1 and its relationship with the host genome, which are yet to be discovered. Recent research has increased our understanding of L1s and their interactions with the genome but there are still many questions that remain to be answered. Some of these questions are: why do L1s apparently accumulate in human brain cells, and what are the pathological impacts of L1s in human neuronal disorders? What is the impact of L1 in tumourigenesis, and is re-activation of the L1 in malignancy-derived cells a cause or effect of instability?

Perhaps recent technological advances in capturing *de novo* endogenous L1 retrotransposition events, such as our combined ATLAS and high-throughput sequencing technique, can help to answer some of these questions and allow us to better understand biology of L1 and the dynamics of its interaction with its host genome.

Bibliography

Aapola, U., Liiv, I., and Peterson, P. (2002). Imprinting regulator DNMT3L is a transcriptional repressor associated with histone deacetylase activity. *Nucleic Acids Research* **30**, 3602-3608.

Allegrucci, C., Wu, Y.Z., Thurston, A., Denning, C.N., Priddle, H., Mummery, C.L., Ward-van Oostwaard, D., Andrews, P.W., Stojkovic, M., Smith, N., Parkin T., Jones, M.E., Warren, G., Yu, L., Brena, R.M., Plass, C., and Young, L.E. (2007). Restriction landmark genome scanning identifies culture-induced DNA methylation instability in the human embryonic stem cell epigenome. *Human Molecular Genetics* **16**, 1253-1268.

Alves, G., Tatro, A, and Fanning, T. (1996). Differential methylation of human LINE 1 retrotransposons in malignant cells. *Gene* **176**, 39-44.

Anderson, P., and Kedersha, N. (2006). RNA granules. *Journal of Cell Biology* **172**, 803-808.

Andrews, P.W. (2002). From teratocarcinomas to embryonic stem cells. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **357**, 405-417.

Andrews, P.W., Damjanov, I., Simon, D., Banting, G.S., Carlin, C., Dracopoli, N.C., and Føgh, J. (1984). Pluripotent embryonal carcinoma clones derived from the human teratocarcinoma cell line Tera-2. Differentiation in vivo and in vitro. *Laboratory Investigation* **50**, 147-162.

Aporntewan, C., Phokaew, C., Piriyapongsa, J., Ngamphiw, C., Ittiwut, C., Tongsima, S., and Mutirangura, A. (2011). Hypomethylation of intragenic LINE-1 represses transcription in cancer cells through AGO2. *Public Library of Science One* **6**, e17934.

Aravin, A.A., Hannon, G.J., and Brennecke, J. (2007). The Piwi−piRNA pathway provides an adaptive defense in the transposon arms race. *Science,* **318**, 761-764.

Asch, H.L., Eliacin, E., Fanning, T.G., Connolly, J.L., Bratthauer, G., and Asch, B.B. (1996). Comparative expression of the LINE-1 p40 protein in human breast carcinomas and normal breast tissues. *Oncology Research* **8**, 239-247.

Athanikar J.N., Badge R.M., and Moran J.V. (2004). A YY1-binding site is required for accurate human LINE-1 transcription initiation. *Nucleic Acids Research* **32**, 3846-3855.

Badge, R.M., Alisch, R.S. and Moran, J.V. (2003). ATLAS: a system to selectively identify human-specific L1 insertions. *American Journal of Human Genetics* **72**, 823-838.

Bailey, J.A., Carrel, L., Chakravarti, A., and Eichler, E.E. (2000). Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 6634-6639.

Bailey, J.A., Liu, G., and Eichler, E.E. (2003). An Alu transposition model for the origin and expansion of human segmental duplications. *American Journal of Human Genetics* **73**, 823-834.

Basame, S., Wai-lun Li, P., Howard, G., Branciforte, D., Keller, D., Martin, S.L. (2006). Spatial assembly and RNA binding stoichiometry of a LINE-1 protein essential for retrotransposition. *Molecular Biology* **357**, 351-357.

Batzer M.A., and Deininger P.L. (2002). Alu repeats and human genomic diversity. *Nature Reviews Genetics* **3**, 370-379.

Beck, C.R., Collier, P., Macfarlane, C., Malig, M., Kidd, J.M., Eichler, E.E., Badge, R.M., and Moran, J.V. (2010). LINE-1 retrotransposition activity in human genomes. *Cell* **141**, 1159–1170.

Beck, C.R., Garcia-Perez, J.L., Badge, R.M., and Moran, J.V. (2011). LINE-1 Elements in Structural Variation and Disease. *Annual Review of Genomics and Human Genetics* **12**, 187-215.

Becker, K.G., Swergold, G.D., Ozato, K. and Thayer, R.E. (1993). Binding of the ubiquitous nuclear transcription factor YY1 to a cis regulatory sequence in the human LINE-1 transposable element. *Human Molecular Genetics* **2,** 1697-702.

Belancio, V.P., Whelton, M. and Deininger, P., (2007). Requirements for polyadenylation at the 3' end of LINE-1 elements. *Gene* **390** 98-107.

Belancio, V.P., Roy-Engel, A.M., Pochampally, R.R., and Deininger, P. (2010). Somatic expression of LINE-1 elements in human tissues. *Nucleic Acids Research* **38**, 3902-3922.

Beraldi, R, Pittoggi C, Sciamanna I, Mattei E, and Spadafora C. (2006). Expression ofLINE-1 retroposons is essential for murine preimplantation development. *Molecular Reproduction and Development* **73,** 279-287.

Bestor, T.H. (1998). The host defense function of genomic methylation patterns. *Novartis Foundation.Symposium*. **214**, 187–195.

Bestor, T.H. (2000). The DNA methyltransferases of mammals. *Human Molecular Genetics* **9**, 2395-2402.

Bestor, T.H., (2003). Cytosine methylation mediates sexual conflict. *Trends in Genetics* **19**, 185-190.

Bestor, T. H., and Bourc'his, D. (2004). Transposon silencing and imprint establishment in mammalian germ cells. *Cold Spring Harbor Symposia on Quantitative Biology* **69**, 381–387.

Bird, A.P., (1986). CpG-rich islands and the function of DNA methylation. *Nature,* **321**, 209-213.

Blankenberg, D., Gordon, A., Von Kuster, G., Coraor, N., Taylor, J., Nekrutenko, A.; Galaxy Team. (2010). Manipulation of FASTQ data with Galaxy. *Bioinformatics***26**, 1783-1785.

Boeke, J.D. and Pickeral, O.K. (1999). Retroshuffling the genomic deck. *Nature* **398,** 108-111.

Boeke, J.D. (1997). LINEs and Alus - the polyA connection. *Nature Genetics* **16,** 6-7.

Boissinot, S., Chevret, P., and Furano, A.V. (2000). L1 (LINE-1) retrotransposon evolution and amplification in recent human history. *Molecular Biology Evolution*, **17**, 915–928.

Boissinot S., and Furano A.V. (2001). Adaptive evolution in LINE-1 retrotransposons. *Molecular Biology and Evolution* **18**, 2186-2194.

Boissinot, S., Entezam A., Young L., Munson P.J., and Furano A.V. (2004). The insertional history of an active family of L1 retrotransposons in humans. *Genome Research* **14**, 1221-1231.

Boissinot S., Davis J., Entezam A., Petrov D., and Furano A.V. (2006). Fitness cost of LINE-1 (L1) activity in humans. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 9590-9594.

Borc'his, D., and Bestor, T.H. (2004). Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature* **431**, 96–99.

Borsani, G., Tonlorenzi, R., Simmler, M.C., Dandolo, L., Arnaud, D., Capra, V., Grompe, M., Pizzuti, A., Muzny, D., Lawrence, C., Willard, H.F., Avner, P., and Ballabio, A. (1991). Characterization of a murine gene expressed from the inactive X chromosome. *Nature* **351**, 325-329.

Branciforte, D., and Martin, S.L. (1994). Developmental and cell type specificity of LINE-1 expression in mouse testis: implications for transposition. *Molecular and Cellular Biology* **14**, 2584-2592.

Bratthauer, G.L. and Fanning, T.G. (1993). LINE-1 retrotransposon expression in pediatric germ cell tumors. *Cancer* **71**, 2383-2386.

Bratthauer, G.L., Cardiff, R.D., and Fanning, T.G. (1994). Expression of LINE-1 retrotransposons in human breast cancer. *Cancer* **73**, 2333-2336.

Brouha B., Meischl C., Ostertag E., de Boer M., Zhang Y., Neijens H., Roos D., and Kazazian H.H. Jr. (2002). Evidence consistent with human L1

retrotransposition in maternal meiosis I. *American Journal of Human Genetics* **71**: 327-336.

Brouha, B., Schustak, J., Badge, R.M., Lutz-Prigge, S., Farley, A.H., Moran, J.V. and Kazazian, H.H., Jr. (2003). Hot L1s account for the bulk of retrotransposition in the human population. *Proceedings of the National Academy of Sciences of the United States of America* **100,** 5280-5.

Brosius, J. (1999). Genomes were forged by massive bombardments with retroelements and retrosequences. *Genetica* **107**, 209-238.

Burke, W.D., Malik, H.S., Jones, J.P. and Eickbush, T.H. (1999). The domain structure and retrotransposition mechanism of R2 elements are conserved throughout arthropods. *Molecular Biology and Evolution* **16,** 502-511.

Burke, W.D., Malik, H.S., Lathe, W.C., 3rd and Eickbush, T.H. (1998). Are retrotransposons long-term hitchhikers? *Nature* **392,** 141-142.

Bushman, F. D. (2003). Targeting survival: integration site selection by retroviruses and LTR−retrotransposons. *Cell* **115**, 135–138.

Buzdin, A., Gogvadze, E., Kovalskaya, E., Volchkov, P., Ustyugova, S., Illarionova, A., Fushan, A., Vinogradova, T., and Sverdlov, E. (2003). The human genome contains many types of chimeric retrogenes generated through in vivo RNA recombination. *Nucleic Acids Research* **31**, 4385-4390.

Callinan, P.A., and Batzer, M.A. (2006). Retrotransposition of marked SVA elements by human L1s in cultured cells. *Genome Dynamics* **1**, 104-115.

Chalitchagorn, K., Shuangshoti, S., Hourpai, N., Kongruttanachok, N., Tangkijvanich, P., Thong-ngam, D., Voravud, N., Sriuranpong, V., and Mutirangura, A. (2004). Distinctive pattern of LINE-1 methylation level in normal tissues and the association with carcinogenesis. *Oncogene* **23**, 8841-8846.

Chen, J.M., Férec, C., and Cooper, D.N. (2006). LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease: mutation detection bias and multiple mechanisms of target gene disruption. *Journal of Biomedicine and Biotechnology* **2006**, 56182.

Chen, T.R. (1988). Re-evaluation of HeLa, HeLa S3, and HEp-2 karyotypes. *Cytogenetics and Cell Research* **48**, 19-24.

Chow, J.C., Ciaudo, C., Fazzari, M.J., Mise, N., Servant, N., Glass JL, Attreed, M., Avner, P., Wutz, A., Barillot, E., Greally, J.M., Voinnet, O., and Heard, E. (2010). LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. *Cell* **141**, 956-969.

Chureau, C., Prissette, M., Bourdet, A., Barbe, V., Cattolico, L., Jones, L., Eggen, A., Avner, P., and Duret, L. (2002). Comparative sequence analysis of the X-

inactivation center region in mouse, human, and bovine. *Genome Research* **12**, 894-908.

Cooper, D.N., and Youssoufian, H. (1988). The CpG dinucleotide and human genetic disease. *Human Genetics* **78**, 151-155.

Cotton, A.M., Avila, L., Penaherrera, M.S., Affleck, J.G., Robinson, W.P., Brown, C.J. (2009). Inactive X chromosome-specific reduction in placental DNA methylation. *Human Molecular Genetics* **18**, 3544-52.

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

Cordaux, R. and Batzer, M.A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews Genetics* **10**, 691-703.

Cordaux, R., Hedges, D.J., Herke, S.W., and Batzer, M.A. (2006). Estimating the retrotransposition rate of human Alu elements. *Gene,* **373**, 134–137.

Cost, G.J., Feng, Q., Jacquier, A. and Boeke, J.D. (2002). Human L1 element target-primed reverse transcription in vitro. *EMBO Journal* **21,** 5899-5910.

Cost, G. J. & Boeke, J. D. (1998). Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. *Biochemistry* **37**, 18081–18093.

Costas, J., and Naveira, H. (2000). Evolutionary history of the human endogenous retrovirus family ERV9. *Molecular Biology and Evolution* **17**, 320-330.

Curcio, M.J. and Derbyshire, K.M. (2003). The outs and ins of transposition: from mu to kangaroo. *Nature Reviews Molecular and Cellular Biology* **4,** 865-77.

Craig , N.L. (2002). Mobile DNA: an introduction. In Mobile DNA II, Craig, N.L., Craigie, R., Gellert, M. and Lambowitz, A.M. (eds). 3-12 American Society for Microbiology, Washington, D.C.

Dawkins, R. (1976). The Selfish Gene. *Oxford University Press*, Oxford.

Deininger, P.L Moran J.V., Batzer M.A., and Kazazian H.H. Jr. (2003). Mobile elements and mammalian genome evolution. *Current Opinion in Genetics and Development* **13**, 651-658.

Dewannieux, M., Esnault, C. and Heidmann, T. (2003). LINE-mediated retrotransposition of marked Alu sequences. *Nature Genetics* **35,** 41-8.

Divoky, V., Indrak, K., Mrug, M., Brabec, V., Huisman, T.H.J. & Prchal, J.T. (1996). A novel mechanism of ß thalassemia: The insertion of L1 retrotransposable element into ß globin IVS II. *Blood* **88**, 148a.

Doucet, A.J., Hulme, A.E., Sahinovic, E., Kulpa, D.A., Moldovan, J.B., Kopera, H.C., Athanikar, J.N., Hasnaoui, M., Bucheton, A, Moran, J.V., and Gilbert, N. (2010). Characterization of LINE-1 ribonucleoprotein particles. *Public Library of Science Genetics* **6,** e1001150.

Drexler, H.G., Matsuo, Y., and MacLeod, R.A. (2003). Persistent use of false myeloma cell lines. *Human Cell* **16**, 101-105.

Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **19**, 113.

Eickbush, T. (1999). Exon shuffling in retrospect. *Science* **283**, 1465-1467.

Eickbush, T.H., and Jamburuthugoda, V.K. (2008). The diversity of retrotransposons and the properties of their reverse transcriptases. *Virus Research* **134**, 221-234.

Eickbush, T.H. and Malik, H.S. (2002). Origins and evolution of retrotransposons. In Mobile DNA II, Craig, N.L., Craigie, R., Gellert, M. and Lambowitz, A.M. (eds) pp. 1111-1144. *American Society for Microbiology*, Washington, D.C.

Ejima, Y., and Yang, L. (2003). Trans mobilization of genomic DNA as a mechanism for retrotransposon-mediated exon shuffling. *Human Molecular Genetics* **12**, 1321-1328.

Elliott, A.M., Elliott, K.A., and Kammesheidt, A. (2010). High resolution array-CGH characterization of human stem cells using a stem cell focused microarray. *Molecular Biotechnology* **46**, 234-242.

Esnault, C., Maestre, J. and Heidmann, T. (2000). Human LINE retrotransposons generate processed pseudogenes. *Nature Genetics* **24,** 363-367.

Esnault, C., Priet, S., Ribet, D., Heidmann, O., and Heidmann, T. (2008). Restriction by APOBEC3 proteins of endogenous retroviruses with an extracellular life cycle: ex vivo effects and in vivo "traces" on the murine IAPE and human HERV-K elements. *Retrovirology* **5**, 75.

Ergün, S., Buschmann, C., Heukeshoven, J., Dammann, K., Schnieders, F., Lauke, H., Chalajour, F., Kilic, N., Strätling, W.H. and Schumann, G.G. (2004). Cell type- specific expression of LINE-1 ORF1 and ORF2 in fetal and adult human tissues. *Journal of Biological Chemistry* **279**, 27753-27763.

Ewing, A.D., and Kazazian, H.H. Jr. (2011). Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. *Genome Research* **21**, 983-990.

Ewing, A.D., and Kazazian, H.H. Jr. (2010). High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome Research* **20**, 1262-1270.

Feng, Q., Moran, J., Kazazian, H.H., Jr. and Boeke, J.D. (1996). Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87,** 905-916.

Feng, Q., and Zhang, Y. (2001). The MeCP1 complex represses transcription through preferential binding, remodeling, and deacetylating methylated nucleosomes. *Genes and Development* **15**, 827-832.

Ficz, G., Branco, M.R., Seisenberger, S., Santos, F., Krueger, F., Hore, T.A., Marques, C.J., Andrews, S., and Reik, W. (2011). Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature* **473**, 398-402.

Freeman, P., Macfarlane, C., Collier, P., Jeffreys, A.J., and Badge, R.M. (2011). L1 hybridization enrichment: a method for directly accessing de novo L1 insertions in the human germline. *Human Mutation* **32**, 978-988.

Fukuda, M.N., Dell, A., Oates, J.E., and Fukuda, M. (1985). Embryonal lactosaminoglycan. The structure of branched lactosaminoglycans with novel disialosyl (sialyl alpha 2----9 sialyl) terminals isolated from PA1 human embryonal carcinoma cells. *Journal of Biological Chemistry* **260**, 6623-6631.

Furano, A.V. (2000). The biological properties and evolutionary dynamics of mammalian LINE-1 retrotransposons. *Progress in Nucleic Acid Research and Molecular Biology* **64**, 255-294.

Gallois-Montbrun, S., Kramer, B., Swanson, C.M., Byers, H., Lynham, S., Ward, M., and Malim, M.H. (2007). Antiviral protein APOBEC3G localizes to ribonucleoprotein complexes found in P bodies and stress granules. *Journal of Virology* **81**, 2165-2178.

Garcia-Perez, J.L,, Marchetto, M.C., Muotri, A.R., Coufal, N.G., Gage, F.H., O'Shea K.S., and Moran, J.V. (2007). LINE-1 retrotransposition in human embryonic stem cells. *Human Molecular Genetics* **16**, 1569-1577.

Gartler, S.M. (1968). Apparent Hela cell contamination of human heteroploid cell Lines. *Nature* **217**, 750-751.

Gasior, S.L., Wakeman, T.P., Xu, B., and Deininger, P.L. (2006). The human LINE-1 retrotransposon creates DNA double-strand breaks. *Journal of Molecular Biology* **357**, 1383-1393.

Gasior, S.L,. Roy-Engel, A.M., and Deininger, P.L. (2008). ERCC1/XPF limits L1 retrotransposition. *DNA Repair* **7**, 983-989.

Goecks, J., Nekrutenko, A., Taylor, J.; Galaxy Team. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology* **11**, R86.

Goll, M.G., and Bestor, T.H. (2005). Eukaryotic cytosine methyltransferases. *Annual Review of Biochemistry* **74**, 481–514.

Goodier, J.L., Ostertag, E.M. and Kazazian, H.H., Jr. (2000). Transduction of 3'-flanking sequences is common in L1 retrotransposition. *Human Molecular Genetics* **9,** 653-657.

Goodier, J.L., Zhang, L., Vetter, M.R., and Kazazian, H.H. Jr. (2007). LINE-1 ORF1 protein localizes in stress granules with other RNA-binding proteins, including components of RNA interference RNA-induced silencing complex. *Molecular and Cellular Biology* **27**, 6469-6483.

Goodier, J. L. and Kazazian, H. H. Jr. (2008). Retrotransposons revisited: the restraint and rehabilitation of parasites. *Cell* **135**, 23–35.

Gilbert, N., Lutz-Prigge, S. and Moran, J.V. (2002). Genomic deletions created upon LINE-1 retrotransposition. *Cell* **110,** 315-325.

Gilbert, N., Lutz, S., Morrish, T.A., and Moran, J.V. (2005). Multiple fates of L1 retrotransposition intermediates in cultured human cells. *Molecular and Cellular Biology* **25**, 7780-7795.

Gregory, T.R. and Hebert, P.D. (1999). The modulation of DNA content: proximate causes and ultimate consequences. *Genome Research* **9,** 317-324.

Han, J.S., Szak, S.T., and Boeke, J.D. (2004). Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature* **429**, 268-274.

Han, K., Sen, S.K., Wang, J., Callinan, P.A., Lee, J., Cordaux, R., Liang, P., and Batzer, M.A. (2005). Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages. *Nucleic Acids Research* **33**, 4040-4052.

Han, K., Lee, J., Meyer, T.J, Remedios, P., Goodwin, L., and Batzer, M.A. (2008). L1 recombination-associated deletions generate human genomic variation. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 19366-19371.

Hancks, D.C., Goodier, J.L., Mandal, P.K., Cheung, L.E., and Kazazian, H.H. Jr. (2011). Retrotransposition of marked SVA elements by human L1s in cultured cells. *Human Molecular Genetics* **20**, 3386-3400.

Hata, K., and Sakaki, Y. (1997). Identification of critical CpG sites for repression of L1 transcription by DNA methylation. *Gene* **189**, 227-234.

Hendrich, B., and Bird, A. (1998). Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Molecular.and Cellular.Biology,* **18**, 6538-6547.

Hoffmann, M.J., and Schulz, W.A. (2005). Causes and consequences of DNA hypomethylation in human cancer. *Biochemistry and Cell Biology* **83**, 296-321.

Hohjoh, H. and Singer, M.F. (1996). Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. *EMBO Journal* **15,** 630-639.

Hohjoh, H. and Singer, M.F. (1997). Sequence-specific single-strand RNA binding protein encoded by the human LINE-1 retrotransposon. *EMBO Journal* **16,** 6034-6043.

Holmes, S.E., Dombroski, B.A., Krebs, C.M., Boehm, C.D., and Kazazian, H.H. Jr. (1994). A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion. *Nature Genetics* **7**, 143-148.

Iskow, R.C., McCabe, M., Mills, R.E., Torene, S., Pittard, S., Neuwaid, A.F., Van Meir, E.G., Vertino, P.M., Devine, S.E. (2010). Natural Mutagenesis of Human Genomes by Endogenous Retrotransposons. *Cell* **141**, 1253-1261.

Ito, S., Shen, L., Dai, Q., Wu, S.C., Collins, L.B., Swenberg, J.A., He, C., and Zhang, Y. (2011). Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**, 1300-1303.

Jurka, J. (1997). Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 1872–1877.

Kano, H., Godoy, I., Courtney, C., Vetter, M.R., Gerton, G.L., Ostertag, E.M., and Kazazian, H.H. (2009). L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. *Genes and Development* **23**, 1303–1312.

Kazazian, H. H. Jr, Wong, C., Youssoufian, H., Scott, A.F., Phillips, D.G., Antonarakis S.E. (1988). Haemophilia A resulting from *de novo* insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* **332**,164–166.

Kazazian, H.H., Jr. (1998). Mobile elements and disease. *Current Opinion in Genetics and Development* **8,** 343-50.

Kazazian, H.H.,Jr and Goodier, J.L. (2002). LINE drive. retrotransposition and genome instability. *Cell* **110**, 277-280.

Kazazian, H.H. Jr. (2004). Mobile elements: drivers of genome evolution. *Science* **303**, 1626-1632.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Hausler, D. (2002). The human genome browser at UCSC. *Genome Research* **12**, 996–1006.

Ketting, R.F., Haverkamp, T.H., van Luenen, H.G., and Plasterk, R.H. (1999). Mut-7 of C. elegans, required for transposon silencing and RNA interference, is a homolog of Werner syndrome helicase and RNaseD. *Cell* **99**, 133-141.

Kidd, J.M., Graves, T., Newman, T., Fulton, R., Hayden, H.S., Malig, M., Kallicki, J., Kaul, R., Wilson, R.k., and Eichler, E.E. (2011). A Human Genome Structural Variation Sequencing Resource Reveals Insights into Mutational Mechanisms. *Cell* **143**, 837-847.

Kidwell, M.G., and Lisch, D.R. (2000). Transposable elements and host genome evolution. *Trends in Ecology and Evolution* **15**, 95-99.

Kimberland, M.L., Divoky, V., Prchal, J., Schwahn, U., Berger, W., and Kazazian, H.H. Jr. (1999). Full-length human L1 insertions retain the capacity for high frequency retrotransposition in cultured cells. *Human Molecular Genetics* **8**, 1557-1560.

Kolosha, V.O., and Martin, S.L. (2003). High-affinity, non-sequence-specific RNA binding by the open reading frame 1 (ORF1) protein from long interspersed nuclear element 1 (LINE-1). *Journal of Biological Chemistry* **278**, 8112-8117.

Kozak, M. (1987). Effects of intercistronic length on the efficiency of reinitiation by eucaryotic ribosomes. *Molecular and Cellular Biology* **7,** 3438-3445.

Kuramochi-Miyagawa, S., Watanabe, T., Gotoh, K., Totoki, Y., Toyoda, A., Ikawa, M., Asada, N., Kojima, K., Yamaguchi, Y., Ijiri, T.W., Hata, K., Li, E., Matsuda, Y., Kimura, T., Okabe, M., Sakaki, Y., Sasaki, H., and Nakano, T. (2008). DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. *Genes and Development* **22**, 908-917.

Kurose, K., Hata, K., Hattori, M. and Sakaki, Y. (1995). RNA polymerase III dependence of the human L1 promoter and possible participation of the RNA polymerase II factor YY1 in the RNA polymerase III transcription system. *Nucleic Acids Research* **23,** 3704-9.

Lacroix, M. (2008). Persistent use of "false" cell lines. *International Journal of Cancer* **122**, 1-4.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., *et al*. (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921.

Lavappa, K.S., Macy, M.L., and Shannon, J.E. (1976). Examination of ATCC stocks for HeLa marker chromosomes in human cell lines. *Nature* **259**, 211-213.

Lavie, L., Maldener, E., Brouha, B., Meese, E.U., and Mayer, J. (2004). The human L1 promoter: variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity. *Genome Research* **14**, 2253-2260.

Lee, J., Han, K., Meyer, T.J., Kim, H.S., and Batzer, M.A. (2008). Chromosomal inversions between human and chimpanzee lineages caused by retrotransposons. *Public Library of Science One* **3**, e4047.

Lee, K., Haugen, H.S., Clegg, C.H., and Braun, R.E. (1995). Premature translation of protamine 1 mRNA causes precocious nuclear condensation and arrests spermatid differentiation in mice. *Proceedings of the National Academy of Sciences of the United States of America* **92**, 12451-12455.

Leibold, D.M., Swergold, G.D., Singer, M.F., Thayer, R.E., Dombroski, B.A., and Fanning, T.G. (1990). Translation of LINE-1 DNA elements in vitro and in human cells. *Proceedings of the National Academy of Sciences of the United States of America* **87**, 6990-6994.

Levin, H.L., and Moran, J.V. (2011). Dynamic intractions between transposable elements and their hosts. *Nature Reviews Genetics* **12**, 615-627**.**

Lewis, A., Mitsuya, K., Umlauf, D., Smith, P., Dean, W., Walter, J., Higgins, M., Feil, R., and Reik, W. (2004). Imprinting on distal chromosome 7 in the placenta involves repressive histone methylation independent of DNA methylation. *Nature Genetics* **36**, 1291-1295.

Lewis, J.D., Saperas, N., Song, Y., Zamora, M.J., Chiva, M., and Ausió, J. (2004). Histone H1 and the origin of protamines. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 4148-4152.

Li, E., Bestor, T.H. and Jaenisch, R., (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69**, 915-926.

Li, L.C., and Dahiya, R. (2002). MethPrimer: designing primers for methylation PCRs. *Bioinformatics***18**, 1427-1431.

Li, T.H. and Schmid, C.W. (2001). Differential stress induction of individual Alu loci: implications for transcription and retrotransposition *Gene*, **276,** 135-41.

Liu, G.E., Alkan, C., Jiang, L., Zhao, S., Eichler, E.E. (2009). Comparative analysis of Alu repeats in primate genomes. *Genome Research* **19**, 876-885.

Lin, S.S., Nymark-McMahon, M.H., Yieh, L., and Sandmeyer, S.B. (2001). Integrase mediates nuclear localization of Ty3. *Molecular and Cellular Biology* **21**, 7826-7838.

Lindtner, S., Felber, B.K., and Kjems, J. (2002). An element in the 3' untranslated region of human LINE-1 retrotransposon mRNA binds NXF1 (TAP) and can function as a nuclear export element. *RNA* **8**, 345-356.

Lovsin, N., and Peterlin, B.M. (2009). APOBEC3 proteins inhibit LINE-1 retrotransposition in the absence of ORF1p binding. *Annals of the New York Academy of Sciences***1178**, 268-275.

Lovsin, N., Gubensek, F. and Kordi, D. (2001). Evolutionary dynamics in a novel L2 clade of non-LTR retrotransposons in Deuterostomia. *Molecular Biology and Evolution* **18,** 2213-24.

Luan, D.D., Korman, M.H., Jakubczak, J.L. and Eickbush, T.H. (1993). Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* **72,** 595-605.

Lynch, M., and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151-1155.

Lyon, M.F. (1998). X-chromosome inactivation: a repeat hypothesis. *Cytogenetics and Cell Genetics* **80**, 133-137.

MacDuff, D.A., Demorest, Z.L., and Harris, R.S. (2009). AID can restrict L1 retrotransposition suggesting a dual role in innate and adaptive immunity. *Nucleic Acids Research* **37**, 1854-1867.

Macia, A., Muñoz-Lopez, M., Cortes, J.L., Hastings, R.K., Morell, S., Lucena-Aguilar, G., Marchal, J.A., Badge, R.M., and Garcia-Perez, J.L. (2011). Epigenetic control of retrotransposon expression in human embryonic stem cells. *Molecular and Cellular Biology* **31**, 300-316.

MacLeod, R.A., Dirks, W.G., Matsuo, Y., Kaufmann, M., Milch, H., and Drexler, H.G. (1999). Widespread intraspecies cross-contamination of human tumor cell lines arising at source. *International Journal of Cancer* **83**, 555-563.

Magdinier, F., D'Estaing, S.G., Peinado, C., Demirci, B., Berthet, C., Guérin, J.F., and Dante, R. (2002). Epigenetic marks at BRCA1 and p53 coding sequences in early human embryogenesis. *Molecular Human Reproduction* **8**, 630-635.

Malik, H.S., and Eickbush, T.H. (1998). The RTE class of non-LTR retrotransposons is widely distributed in animals and is the origin of many SINEs. *Molecular Biology and Evolution* **15**, 1123-1134.

Malik, H.S., Burke, W.D., and Eickbush, T.H. (1999). The age and evolution of non-LTR retrotransposable elements. *Molecular Biology and Evolution* **16**, 793-805.

Malone, C.D., Brennecke, J., Dus, M., Stark, A., McCombie, W.R., Sachidanandam, R., and Hannon, G.J. (2009). Specialized piRNA pathways act in germline and somatic tissues of the Drosophila ovary. *Cell* **137**, 522-535.

Martin, S.L. (2010). Nucleic acid chaperone properties of ORF1p from the non-LTR retrotransposons, LINE-1. *RNA Biology* **7**, 706-11.

Martin, S.L. and Bushman, F.D. (2001). Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. *Molecular and Cellular Biology* **21**, 467-475.

Martin, S.L., Cruceanu, M., Branciforte, D., Wai-Lun Li, P., Kwok, S.C., Hodges, R.S., Williams, M.C. (2005). LINE-1 retrotransposition requires the nucleic acid chaperone activity of the ORF1 protein. *Molecular Biology* **348**, 549-61.

Martín, F., Maranon, C., Olivares, M., Alonso, C. and Lopez, M.C. (1995). Characterization of a non-long terminal repeat retrotransposon cDNA (L1Tc) from Trypanosoma cruzi: homology of the first ORF with the ape family of DNA repair enzymes. *Journal of Molecular Biology* **247,** 49-59.

Masters, J. (2002)**.** False cell lines. *Carcinogenesis* **23**, 371.

Matsuo, M., Masumura, T., Nishio, H., Nakajima, T., Kitoh, Y., Takumi, T., Koga, J., and Nakamura, H. (1991). Exon skipping during splicing of dystrophin mRNA precursor due to an intraexon deletion in the dystrophin gene of Duchenne muscular dystrophy kobe. *Journal of Clinical Investigation* **87**, 2127-2131.

Mathias, S.L., Scott, A.F., Kazazian, H.H.,Jr, Boeke, J.D. and Gabriel, A. (1991). Reverse transcriptase encoded by a human transposable element. *Science* **254**,1808-1810.

Mayer ,W., Niveleau, A., Walter, J., Fundele, R., and Haaf, T. (2000). Demethylation of the zygotic paternal genome. *Nature* **403**, 501-502.

McClintock, B. (1950). The Origin and Behavior of Mutable Loci in Maize. *Proceedings of the National Academy of Sciences of the United States of America* **36,** 344-355.

McMillan, J.P. and Singer, M.F. (1993). Translation of the human LINE-1 element, L1Hs. *Proceedings of the National Academy of Sciences of the United States of America*, **90,** 11533-11537.

Meehan, R.R., Lewis, J.D., and Bird, A.P. (1992). Characterization of MeCP2, a vertebrate DNA binding protein with affinity for methylated DNA. *Nucleic Acids Research* **20**, 5085-5092.

Meischl, C., Boer, M., Ahlin, A., and Roos D. (2000). A new exon created by intronic insertion of a rearranged LINE-1 element as the cause of chronic granulomatous disease. *European Journal of Human Genetics* **8**, 697-703.

Meister, G., Landthaler, M., Patkaniowska, A., Dorsett, Y, Teng, G., and Tuschl, T. (2004). Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Molecular Cell* **15**, 185-197.

Miki, Y., Nishisho, I., Horii, A., Miyoshi, Y., Utsunomiya, J., Kinzler, K.W., Vogelstein, B., and Nakamura, Y. (1992). Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer Research* **52**, 643-645.

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

Mills, R.E., Walter, K., Stewart, C., Handaker, R.E., Chen, K., Alkan, C., Abyzov, A., Yoon, S.C., Ye, K., Cheetham, K., Chinwalla A., Conrad D.F., Fu, Y., Grubert, F., Hajirasouliha, I., Hormozdiari, F., Iakoucheva, L.M., Iqbal, Z., Kang, S., Kidd, J.M., Konkel, M.K., Korn, J., Khurana, E., Kural, D., Lam, H.Y.K., *et al.* (2011). Mapping copy number variation by population−scale genome sequencing. *Nature,* **470**, 59–65.

Minakami, R., Kurose, K., Etoh, K., Furuhata, Y., Hattori, M. and Sakaki, Y. (1992). Identification of an internal cis-element essential for the human L1 transcription and a nuclear factor(s) binding to the element. *Nucleic Acids Research* **20,** 3139-45.

Miné, M., Chen, J.M., Brivet, M., Desguerre, I., Marchant, D., de Lonlay, P., Bernard, A., Férec, C., Abitbol, M., Ricquier, D., and Marsac, C. (2007). A large genomic deletion in the PDHX gene caused by the retrotranspositional insertion of a full-length LINE-1 element. *Human Mutation* **28**, 137-142.

Moran, J.V. and Gilbert, N. (2002). Mammalian LINE-1 retrotransposons and related elements. In Mobile DNA II, Craig, N.L., Craigie, R., Gellert, M. and Lambowitz, A.M. (eds). 836-869. American Society for Microbiology, Washington, D.C.

Moran, J.V., DeBerardinis, R.J. and Kazazian, H.H., Jr. (1999). Exon shuffling by L1 retrotransposition. *Science* **283,** 1530-1534.

Moran, J.V., Holmes, S.E., Naas, T.P., DeBerardinis, R.J., Boeke, J.D. and Kazazian, H.H., Jr. (1996). High-frequency retrotransposition in cultured mammalian cells. *Cell* **87,** 917-927.

Moran, J.V., Mecklenburg, K.L., Sass, P., Belcher, S.M., Mahnke, D., Lewin, A. and Perlman, P.S. (1994). Splicing defective mutants of the COX1 gene of yeast mitochondrial DNA: initial definition of the maturase domain of the group II intron aI2. *Nucleic Acids Research* **22,** 2057-2064.

Morrish, T.A., Gilbert, N., Myers, J.S., Vincent, B.J., Stamato, T.D., Taccioli, G.E., Batzer, M.A. & Moran, J.V. (2002). DNA repair mediated by endonuclease-independent LINE-1 retrotransposition, *Nature Genetics* **31,** 159-165.

Muotri, A.R., Chu, V.T., Marchetto, M.C., Deng, W., Moran, J.V., and Gage, F.H. (2005). Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* **435**, 903-910.

Muotri, A.R., Marchetto, M.C., Coufal, N.G., Oefner, R., Yeo, G., Nakashima, K., and Gage, F.H. (2010). L1 retrotransposition in neurons is modulated by MeCP2. *Nature* **468**, 443-446.

Myers, J.S., Vincent, B.J., Udall, H., Watkins, W.S., Morrish, T.A., Kilroy, G.E., Swergold, G.D., Henke, J., Henke, L., Moran, J.V., Jorde, L.B., and Batzer, M.A. (2002). A comprehensive analysis of recently integrated human Ta L1 elements. *American Journal of Human Genetics* **71**, 312-326.

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

Naas, T.P., DeBerardinis, R.J., Moran, J.V., Ostertag, E.M., Kingsmore, S.F., Seldin, M.F., Hayashizaki, Y., Martin, S.L., and Kazazian, H.H. (1998). An actively retrotransposing, novel subfamily of mouse L1 elements. *EMBO Journal* **17**, 590-597.

Nekrutenko, A. and Li, W.H. (2001). Transposable elements are found in a large number of human protein-coding genes. *Trends in Genetics* **17,** 619-21.

Nelson-Rees, W.A., Daniels, D.W. and Flandermeyer, R.R. (1981) Cross-contamination of cells in culture. *Science* **212**, 446-452.

Nelson-Rees, W.A. and Flandermeyer, R.R. (1976). HeLa cultures defined. *Science* **191**, 96-98.

Nigumann, P., Redik, K., Mätlik, K, and Speek, M. (2002). Many human genes are transcribed from the antisense promoter of L1 retrotransposon. *Genomics* **79**, 628-634.

Ogura, H., Yoshinouchi, M., Kudo, T., Imura, M., Fujiwara, T. and Yabe, Y. (1993). Human papillomavirus type 18 DNA in so-called HEP-2, KB and FL cells--further evidence that these cells are HeLa cell derivatives, *Cellular and Molecular Biology* **39**, 463-467.

Ohno, S. (1972). So much "junk" DNA in our genome. In Evolution of genetic systems, Smith, H.H. (ed) 366-370. Gordon and Breach, New York.

Ohshima, K. and Okada, N. (1994). Generality of the tRNA origin of short interspersed repetitive elements (SINEs). Characterization of three different tRNA-derived retroposons in the octopus. *Journal of Molecular Biology* **243**, 25-37.

Okano, M., Xie, S., and Li, E. (1998). Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nature Genetics* **19**, 219–20

Okano, M., Bell, D.W., Haber, D.A. and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**, 247-257.

Ostertag, E.M., Prak, E.T., DeBerardinis, R.J., Moran, J.V., and Kazazian, H.H. Jr. (2000). Determination of L1 retrotransposition kinetics in cultured cells. *Nucleic Acids Research* **28**, 1418-1423.

Ostertag, E.M. and Kazazian, H.H., Jr. (2001a). Biology of Mammalian L1 retrotransposons. *Annual Review of Genetics* **35,** 501-538.

Ostertag, E.M. and Kazazian, H.H., Jr. (2001b). Twin priming: A proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Research* **11,** 2059-2065.

Ostertag, E.M., DeBerardinis, R.J., Goodier, J.L., Zhang, Y., Yang, N., Gerton, G.L., and Kazazian, H.H. Jr. (2002). A mouse model of human L1 retrotransposition. *Nature Genetics* **32**, 655-660.

Pace, J.K. 2nd, and Feschotte, C. (2007). The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. *Genome Research* **17**, 422-432.

Pagel, M. and Johnstone, R.A. (1992). Variation across species in the size of the nuclear genome supports the junk-DNA explanation for the C-value paradox. *Proceedings of the Royal Society London B: Biological Sc*iences **249,** 119-24.

Pardue, M.L., Danilevskaya, O.N., Lowenhaupt, K., Slot, F. and Traverse, K.L. (1996). Drosophila telomeres: new views on chromosome evolution. *Trends in Genetics* **12,** 48-52.

Pavlícek, A., Jabbari, K., Paces, J., Paces, V., Hejnar, J.V., and Bernardi, G. (2001). Similar integration but different stability of Alus and LINEs in the human genome. *Gene* **276**, 39-45.

Pavlícek, A., Paces, J., Elleder, D., and Hejnar, J. (2002). Processed pseudogenes of human endogenous retroviruses generated by LINEs: their integration, stability, and distribution. *Genome Research* **12**, 391-399.

Pera, M.F., Reubinoff, B., and Trounson, A. (2000). Human embryonic stem cells. *Journal of Cell Science* **113**, 5-10.

Pickeral, O.K., Makalowski, W., Boguski, M.S. and Boeke, J.D. (2000). Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome Research* **10,** 411-415.

Popp, C., Dean, W., Feng, S., Cokus, S.J., Andrews, S., Pellegrini, M., Jacobsen S.E., and Reik, W. (2010). Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. *Nature* **463**, 1101-1105.

Pornthanakasem, W., Kongruttanachok, N., Phuangphairoj, C., Suyarnsestakorn, C., Sanghangthum, T., Oonsiri, S., Ponyeam, W., Thanasupawat, T., Matangkasombut, O, and Mutirangura, A. (2008). LINE-1 methylation status of endogenous DNA double-strand breaks. *Nucleic Acids Research* **36**, 3667-3675.

Rahbari, R., Sheahan, T., Modes, V., Collier, P., Macfarlane, C., and Badge RM. (2009). A novel L1 retrotransposon marker for HeLa cell line identification. *Biotechniques* **46**, 277-284.

Rastan, S. (1983). Non-random X-chromosome inactivation in mouse X-autosome translocation embryos--location of the inactivation centre. *Journal of Embryology and Experimental Morphology* **78**, 1-22.

Reik, W., Dean, W., Walter J. (2001). Epigenetic reprogramming in mammalian development. *Science* **293**, 1089-1093.

Riggs, A.D. (1975). X inactivation, differentiation, and DNA methylation. *Cytogenetics and Cell Genetics* **14**, 9-25.

Ross, M.T., Grafham, D.V., Coffey, A.J., Scherer, S., McLay, K., Muzny, D., Platzer, M., Howell, G.R., Burrows, C., Bird, C.P., Frankish, A., Lovell, F.L., Howe, K.L., Ashurst, J.L., Fulton, R.S., Sudbrak, R., Wen, G., Jones, M.C., Hurles, M.E., Andrews, T.D., Scott, C.E., Searle, S., Ramser, J., Whittaker, A., Deadman, R., Carter, N.P., Hunt, S.E., Chen, R., Cree, A., Gunaratne, P., Havlak, P., Hodgson, A., Metzker, M.L., Richards, S., Scott, G., Steffen, D., Sodergren, E., Wheeler, D.A., Worley, K.C., *et al*. (2005). The DNA sequence of the human X chromosome. *Nature* **434**, 325-337.

Rozen, S., and Skaletsky, H. (2000). Primer3 on the WWW for general users and for biologist programmers. *Methods in Molecular Biology* **132**, 365-386.

Salem A.H., Myers J.S., Otieno A.C., Watkins W.S., Jorde L.B., and Batzer M.A. (2003). LINE-1 preTa Elements in the Human Genome. *Molecular Biology*, **326**, 1127-1146.

Sambrook, J., and Russell, D. W. (2001). Molecular cloning, a laboratory manual, **1**, 3rd edn (New York: cold Spring Harbour Laboratory press).

Samonte, R.V., and Eichler, E.E. (2002). Segmental duplications and the evolution of the primate genome. *Nature Reviews Genetics* **3**, 65-72.

Schmid CW. (1996). Alu: structure, origin, evolution, significance and function of one-tenth of human DNA. *Progress in Nucleic Acid Research and Molecular Biology* **53**, 283-319.

Schwahn, U., Lenzner, S., Dong, J., Feil, S, Hinzmann, B., van Duijnhoven, G., Kirschner, R., Hemberger, M., Bergen, A.A., Rosenberg, T., Pinckers, A.J., Fundele, R., Rosenthal, A., Cremers, F.P., Ropers, H.H., and Berger, W.

(1998). Positional cloning of the gene for X-linked retinitis pigmentosa 2. *Nature Genetics* **19**, 327-332.

Scott, A.F., Schmeckpeper, B.J., Abdelrazik, M., Comey, C.T., O'Hara, B., Rossiter, J.P., Cooley, T., Heath, P., Smith, K.D. and Margolet, L. (1987). Origin of the human L1 elements: proposed progenitor genes deduced from a consensus DNA sequence. *Genomics* **1,** 113-25.

Seleme, M.C., Vetter M.R., Cordaux R., Bastone L., Batzer M.A., and Kazazian H.H. Jr. (2006). Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity. *Proceedings of the National Academy of Sciences of the United States of America,* **103**, 6611-6616.

Sen, S.K., Huang, C.T., Han, K., and Batzer, M.A. (2007). Endonuclease-independent insertion provides an alternative pathway for L1 retrotransposition in the human genome. *Nucleic Acids Research* **35**, 3741-3751.

Siegel, V. and Walter, P. (1988). Each of the activities of signal recognition particle (SRP) is contained within a distinct domain: analysis of biochemical mutants of SRP. *Cell* **52,** 39-49.

Sheen, F.M., Sherry, S.T., Risch, G.M., Robichaux, M., Nasidze, I., Stoneking, M., Batzer, M.A., and Swergold, G.D. (2000). Reading between the LINEs: human genomic variation induced by LINE−1 retrotransposition. *Genome Research* **10**, 1496–1508.

Singer, M.F. (1982). SINEs and LINEs: highly repeated short and long interspersed sequences in mammalian genomes. *Cell* **28**, 433-434.

Singer, M.F., Krek, V., McMillan, J.P., Swergold, G.D., and Thayer, R.E. (1993). LINE-1: a human transposable element. *Gene* **135**, 183-188.

Singer, T., McConnell, M.J., Marchetto, M.C., Coufal, N.G., and Gage, F.H. (2010). LINE-1 retrotransposons: mediators of somatic variation in neuronal genomes? *Trends in Neuroscience* **33**, 345-354.

Skowronski, J., Fanning, T.G. and Singer, M.F. (1988). Unit-length LINE-1 transcripts in human teratocarcinoma cells. *Molecular and Cellular Biology* **8,** 1385-1397.

Smit, A.F. and Riggs, A.D. (1996). Tiggers and DNA transposon fossils in the human genome. *Proceedings of the National Academy of Sciences of the United States of America*, **93,** 1443-8.

Song M., and Boissinot S. (2007). Selection against LINE-1 retrotransposons results principally from their ability to mediate ectopic recombination. *Gene* **390**, 206-213.

Soifer, H.S., and Rossi, J.J. (2006). Small interfering RNAs to the rescue: blocking L1 retrotransposition. *Nature Structural Molecular Biology* **13**, 758-759.

Spits, C., Le Caignec, C., De Rycke, M., Van Haute, L., Van Steirteghem, A., Liebaers, I., and Sermon, K. Whole-genome multiple displacement amplification from single cells. (2006). *Nature Protocols* **1**, 1965-1970.

Speek, M. (2001). Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. *Molecular and Cellular Biology* **21**, 1973-1985.

Stein, R., Gruenbaum, Y., Pollack, Y., Razin, A. and Cedar, H. (1982). Clonal inheritance of the pattern of DNA methylation in mouse cells. *Proceedings of the National Academy of Sciences of the United States of America,* **79**, 61-65.

Stenglein, M.D., and Harris, R.S. (2006). APOBEC3B and APOBEC3F inhibit L1 retrotransposition by a DNA deamination-independent mechanism. *Journal of Biological Chemistry* **281**, 16837-16841.

Suzuki, M.M. and Bird, A.(2008). DNA methylation landscapes: provocative insights from epigenomics. *Nature Reviews Genetics* **9**, 465-476.

Swartz, M.N., Trautner, T.A., and Kornberg, A. (1962). Enzymatic synthesis of deoxyribonucleic acid. XI. Further studies on nearest neighbor base sequences in deoxyribonucleic acids. *Journal of Biological Chemistry* **237**, 1961-1967.

Swergold, G.D. (1990). Identification, characterization, and cell specificity of a human LINE-1 promoter. *Molecular and Cellular Biology* **10,** 6718-6729.

Symer, D.E., Connelly, C., Szak, S.T., Caputo, E.M., Cost, G.J, Parmigiani, G., Boeke, J.D. (2002). Human L1 retrotransposition is associated with genetic instability in vivo. *Cell* **110**, 327–338.

Szak, S.T., Pickeral, O.K., Makalowski, W., Boguski, M.S., Landsman, D. and Boeke, J.D. (2002). Molecular archeology of L1 insertions in the human genome. *Genome Biology* **3,** research0052.

Szak, S.T., Pickeral, O.K., Landsman, D. and Boeke, J.D. (2003). Identifying related L1 retrotransposons by analyzing 3' transduced sequences. *Genome Biology* **4**. R30.

Tang, Y., Lopez, I., and Baloh, R.W. (2001). Age-related change of the neuronal number in the human medial vestibular nucleus: a stereological investigation. *Journal of Vestibular Research* **11**, 357-363.

Tchenio, T., Casella, J.F. and Heidmann, T. (2000). Members of the SRY family regulate the human LINE retrotransposons. *Nucleic Acids Research* **28,** 411-415.

Teneng, I., Montoya-Durango, D.E., Quertermous, J.L., Lacy, M.E., and Ramos, K.S. (2011). Reactivation of L1 retrotransposon by benzo(a)pyrene involves complex genetic and epigenetic regulation. *Epigenetics* **6**, 355-367.

Teysset, L., Dang, V.D., Kim, M.K., and Levin, H.L. (2003). A long terminal repeat-containing retrotransposon of Schizosaccharomyces pombe expresses a Gag-like protein that assembles into virus-like particles which mediate reverse transcription. *Journal of Virology* **77**, 5451-5463.

Thayer, R.E., Singer, M.F. and Fanning, T.G. (1993). Undermethylation of specific LINE-1 sequences in human cells producing a LINE-1-encoded protein. *Gene* **133,** 273-7.

Thomas, C.A. (1971). The Genetic Organization of Chromosomes. *Annual Reviews in Genetics* **5,** 237-256.

Thurston, A., Lucas, E.S., Allegrucci, C, Steele, W., Young, L.E. (2007). Region-specific DNA methylation in the preimplantation embryo as a target for genomic plasticity. *Theriogenology* **68,** 98-106.

Trelogan, S.A., and Martin, S.L. (1995). Tightly regulated, developmentally specific expression of the first open reading frame from LINE-1 during mouse embryogenesis. *Proceedings of the National Academy of Sciences of the United States of America* **92**, 1520-1524.

Ullu, E. and Tschudi, C. (1984). Alu sequences are processed 7SL RNA genes. *Nature* **312,** 171-2.

Ullu E., and Weiner A.M. (1985). Upstream sequences modulate the internal promoter of the human 7SL RNA gene. *Nature* **318**, 371-374.

Van Arsdell, S.W. and Weiner, A.M. (1984). Pseudogenes for human U2 small nuclear RNA do not have a fixed site of 3' truncation. *Nucleic Acids Research* **12,** 1463-1471.

van den Hurk, J.A., van de Pol, D.J., Wissinger, B., van Driel, M.A., Hoefsloot, L.H., de Wijs, I.J., van den Born, L.I., Heckenlively, J.R., Brunner, H.G., Zrenner, E., Ropers, H.H., and Cremers, F.P. (2003). Novel types of mutation in the choroideremia (CHM) gene: a full-length L1 insertion and an intronic mutation activating a cryptic exon. *Human Genetics* **113**, 268-275.

van den Hurk, J.A., Meij, I.C., Seleme, M.C., Kano, H., Nikopoulos, K., Hoefsloot, L.H., Sistermans, E.A., de Wijs, I.J., Mukhopadhyay, A., Plomp, A.S., de Jong, P.T., Kazazian, H.H., and Cremers, F.P. (2007). L1 retrotransposition can occur early in human embryonic development. *Human Molecular Genetics* **16**, 1587-1592.

Vanin, E.F. (1985). Processed pseudogenes: characteristics and evolution. *Annual Review of Genetics* **19,** 253-72.

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

Wagstaff, B.J., Barnerssoi, M., and Roy-Engel, A.M. (2011). Evolutionary conservation of the functional modularity of primate and murine LINE-1 elements. *Public Library of Science One* **10**, e19672.

Walser, J.C., Ponger, L., and Furano, A.V. (2008). CpG dinucleotides and the mutation rate of non-CpG DNA. *Genome Research* **18**, 1403-1414.

Walsh, C.P., and Bestor, T.H. (1999). Cytosine methylation and mammalian development. *Genes and Development* **13**, 26-34.

Wang, H., Xing, J., Grover, D., Hedges, D.J., Han, K., Walker, J.A., and Batzer, M.A. (2005). SVA elements: a hominid-specific retroposon family. *Journal of Molecular Biology* **354**: 994-1007.

Wang, J., Song, L., Grover, D., Azrak, S., Batzer, M.A., and Liang, P. (2006). dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. *Human Mutation* **27**, 323-329.

Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., and Barton, G.J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189-1191.

Watt, F., and Molloy, P.L. (1988). Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. *Genes and Development* **2**, 1136-1143.

Weber, B., Kimhi, S., Howard, G., Eden, A., and Lyko, F. (2010). Demethylation of a LINE-1 antisense promoter in the cMet locus impairs Met signalling through induction of illegitimate transcription. *Oncogene* **29**, 5775-5784.

Weichenrieder, O., Wild, K., Strub, K. and Cusack, S. (2000). Structure and assembly of the Alu domain of the mammalian signal recognition particle. *Nature*, **408,** 167-73.

Weichenrieder, O., Repanas, K., and Perrakis, A. (2004). Crystal structure of the targeting endonuclease of the human LINE-1 retrotransposon. *Structure* **12**, 975-986.

Wei, W., Gilbert, N., Ooi, S.L., Lawler, J.F., Ostertag, E.M., Kazazian, H.H., Jr., Boeke, J.D. and Moran, J.V. (2001). Human L1 retrotransposition: cis preference versus trans complementation. *Molecular Cell Biology* **21,** 1429-1439.

Wei, W., Morrish, T.A., Alisch, R. and Moran, J.V. (2000). A transient assay reveals that cultured human cells can accomodate multiple LINE-1 retrotransposition events. *Analytical Biochemistry* **284,** 435-438.

Wheelan, S.J., Aizawa, Y., Han, J.S., and Boeke, J.D. (2005). Gene-breaking: a new paradigm for human retrotransposon-mediated gene evolution. *Genome Research* **15**, 1073-1078.

Williams, K., Christensen, J., Pedersen, M.T., Johansen, J.V., Cloos, P.A., Rappsilber, J., and Helin, K. (2011). TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* **473**, 343-348.

Witherspoon, D.J., Xing, J., Zhang, Y., Watkins, W.S., Batzer, M.A., and Jorde, L.B. (2010). Mobile element scanning (ME−Scan) by targeted high−throughput sequencing. *BMC Genomics* **11**, 410.

Woodcock, D.M., Lawler, C.B., Linsenmeyer, M.E., Doherty, J.P. and Warren, W.D. (1997). Asymmetric methylation in the hypermethylated CpG promoter region of the human L1 retrotransposon. *Journal of Biological Chemistry* **272,** 7810-6.

Wouters-Tyrou, D., Martinage, A., Chevaillier, P., and Sautière, P. (1998). Nuclear basic proteins in spermiogenesis. *Biochimie* **80**, 117-128.

Wu-Scharf, D., Jeong, B., Zhang, C., and Cerutti, H. (2000). Transgene and transposon silencing in Chlamydomonas reinhardtii by a DEAH-box RNA helicase. *Science* **290**, 1159-1162.

Xing, J., Witherspoon D.J., Ray D.A., Batzer M.A., and Jorde L.B. (2007). Mobile DNA elements in primate and human evolution. *American Journal of Physical Anthropology,* **Suppl 45,** 2-19.

Xing, J., Zhang, Y., Han, K., Salem, A.H., Sen, S.K., Huff, C.D., Zhou, G., Kirkness, E.F., Levy, S., Batzer, M.A., and Jorde, L.B. (2009). Mobile elements create structural variation: analysis of a complete human genome. *Genome Research* **19**, 1516-26.

Yang, J., Malik, H.S. and Eickbush, T.H. (1999). Identification of the endonuclease domain encoded by R2 and other site- specific, non-long terminal repeat retrotransposable elements. *Proceedings of the National Academy of Sciences of the United States of America*, **96,** 7847-52.

Yang, N., Zhang, L., Zhang, Y. and Kazazian, H.H., Jr. (2003). An important role for RUNX3 in human L1 transcription and retrotransposition. *Nucleic Acids Research*, **31,** 4929-40.

Yoder, J.A., Walsh, C.P. and Bestor, T.H. (1997). Cytosine methylation and the ecology of intragenomic parasites. *Trends in Genetics*, **13,** 335-40.

Yoshida, K., Nakamura, A., Yazaki, M., Ikeda, S., and Takeda, S. (1998). Insertional mutation by transposable element, L1, in the DMD gene results in X-linked dilated cardiomyopathy. *Human Molecular Genetics* **7**, 1129-1132.

Zeuthen, J., Nørgaard, J.O., Avner, P., Fellous, M., Wartiovaara, J., Vaheri, A., Rosén, A., and Giovanella, B.C. (1980). Characterization of a human ovarian teratocarcinoma-derived cell line. *International Journal of Cancer* **25**, 19-32.

Zingler, N., Weichenrieder, O. and Schumann, G. (2005). APE-type non-LTR retrotransposons: determinants involved in target site recognition. *Cytogenetic and Genome Research* **110**, 250-268.

Appendices

Appendix I:

Materials and method solutions

Recipe 1: *Denaturing solution*

1.5 M NaCl

0.5 NaOH

Recipe 2: *Depurination solution*

0.2 HCl

Recipe 3: *10x Denaturing and hybridising buffer*

450 mM Tris-HCl PH 8.8

110 mM ammonium sulphate

45 mM MgCl2

67 mM 2-mercaptoethanol

44 mM EDTA

Recipe 4: *Dot-blot denaturing solution*

0.5 M NaOH

2 M NaCl

25 mM EDTA

Recipe 5: *IPTG*

0.48 g IPTG top up to 20 ml with 18 MΩ distilled water

Recipe 6: *LB plates*

400 ml Luria Agar – Boil until all the agar melted, cool in a 37°C water bath for 15 minutes, pour plates

For LB-Ampicillin plates- same as for LB plates, but prior to pour plates, add 800 µl 100 mg/ml Ampicillin

Recipe 7: *Xgal (bromo-chloro-indolyl-galactopyranoside)*

20 ml Dimethylformamide

1 g Xgal

Make aliqouts- store at -20°C (light- sensitive)

Recipe 8: *SOB media*

To make 1 L

20 g Tryptone

5 g Yeast extract

0.5 g NaCl

625 µl 4M KCl

10 g Glycine

Make up to 800 ml, and bring to pH 7 with 1 M NaOH- top up to 990 ml

Autoclave and top up to 1 L with 1 M MgCl2

Store at 4°C

Recipe 9: *SOC media*

To make 1 L

20 g Tryptone

5 g Yeast extract

0.5 g NaCl

625 µl 4M KCl

Make up to 800 ml, and bring to pH 7 with 1 M NaOH- top up to 990 ml

Autoclave and top up to 1 L with 1 M MgCl2

Store at 4°C

Recipe 10: *Modified CHURCH buffer*

Neutralising solution

1 M Tris pH 7.5

1.5 M NaCl (0.5 M Tris pH 7.2, 1M NaCl)

Phosphate wash

Recipe 11: *20x SSC*

3.0 M NaCl

0.3 M Sodium Acetate

Adjusted to pH 7 with 14 N HCl

Recipe 12: *TAE buffer (1x Tris-Acetate EDTA)*

40 mM Tris-acetate

1 mM EDTA

Recipe 13: *TBE (0.5x Tris-Borate EDTA)*

45 mM Tris-borate

1 mM EDTA

Recipe 14: *TBE loading buffer*

0.5x TBE

12.5% Ficole

Bromophenol blue

Recipe 15: *TB buffer*

To make 0.5 L

1.5 g 10 mM PIPES

1.1 g 15 mM CaCl2.H2O

7.3 g 250 mM KCl

pH to 6.7 with 10 M KOH

5.45 g MnCl2.4H2O

Make up to 500 ml then filter sterilis- store at 4°C

Recipe 16: *TE buffer*

10 mM Tris-HCl pH 8.0

1 mM EDTA

Recipe 17: *11 x PCR buffer*

45 mM Tris-HCl pH 8.8

11 mM Ammonium Sulphate

4.5 mM Magnesium Chloride

6.7 mM 2-Mercaptoethanol

4.4 mM EDTA

1 mM dATP, 1 mM dCTP, 1 mM dGTP, 1 mM dTTP

113 µg/ml BSA

Recipe 18: *10x (A, B and C) PCR buffer*

10x PCR buffer are used in three different concentrations of dNTPs. Buffer A: 1mM dNTPs, B (0.5mM dNTPs), C (0.2mM dNTPs)

Component (Stock concentration)	**A (µl)**	**B (µl)**	**C (µl)**
2 mM Tris-HCl pH 8.8	222	222	222
1 M NH4SO4	112	112	112
ß-Mercaptoethano (100%)	4.8	4.8	4.8
BSA (50 mg/ml)	22	22	22
1 M MgCl2	42	42	42
dATP (100 mM)	90	44	44
dGTP (100 mM)	90	44	44
dCTP (100 mM)	90	44	44
dTTP (100 mM)	90	44	44
PCR clean H2O	137.7	339.8	459.4

- Buffers were aliqouted to 50 µl in PCR clean screw cap tubes and stored at -20°C

Recipe 19: *single molecule diluent (SMD)*

5 ng/µl E.coli DNA

5 mM Tris-HCl (pH 7.5)

Appendix II:

Protocols for different restriction enzymes -ATLAS libraries constructions

Protocol 1: *ATLAS NlaIII Library Construction and Amplification*

600 ng of gDNA was digested with 15 units of *Nla*III (NEB) for 3 hours or overnight at 37°C in a water bath, in a final reaction volume of 30 µl (the *Nla*III restriction enzyme based library construction mentioned in Appendix II). Several controls were included in the digestion step: DNA negative (H_2O); digestion enzyme negative (replaced with 50% glycerol); and a DNA/reaction positive. After the digestion the reactions were heated at 65°C for 20 min to inactivate the digestion enzyme (digested DNA was stored at -80°C in a PCR clean condition). 20 µl of each linker primer RBMSL3and RBD4 were annealed together by incubating them at 65°C for 10 min and then allowing them to cool down at room temperature over 30-69 minutes. In the standard ATLAS protocol (Badge *et al.,* 2003), 100 ng of the digested DNA was ligated to a 40-fold molar excess of the annealed suppression linker. The amount of linker was calculated by assuming the enzyme completely digested the genome into 'X' number of fragments with two ligatable ends, and 3 pg of DNA represents one haploid genome equivalent. X varies with respect to the enzyme's cutting frequency (but all calculations are necessarily approximate). For *Nla*III, 2.7 µl of annealed linker was used for each ligation in a final volume of 20 µl. 100 ng of genomic DNA were ligated with the annealed linker overnight at 15°C, in a final reaction volume of 20 µl. The linker negative (H_20), and enzyme negative (50% glycerol) and two reaction positives were included as ligation stage controls. A 20 µl ligation reaction final volume consisted of, 5µl of 100 ng digested DNA, and 2.7 µl (10 µl) annealed linker, 1.34 µl (4 Weiss units) T4 ligase (Promega), 2 µl 10×ligase buffer and 8.96 µl H_2O. To inactivate the ligation and also to remove the 'dummy' RBD4 oligonucleotides, reactions were incubated at 70°C for 10 min. The excess of linkers and short DNA fragments (<100 bp) were removed by using the Qiaquick PCR purification system (Qiagen) according to the manufacturer's instructions. The purified linkered DNA was then eluted in PCR clean 5MT at a final volume of 30 µl. These eluates were aliquoted into three sets of 10 µl and stored at -80 °C.

Protocol 2: *TS-ATLAS LRE3-specific MspI Library Construction and Amplification*

600ng of genomic DNA was digested to completion with 15 units of *MseI* (NEB) in the manufacturer's recommended buffer at 37°C for 3 hours. After incubation reactions were heated to 65°C for 20 minutes to inactivate the restriction enzyme. Prior to setting up the ligation reaction, linker oligonucleotides were freshly annealed by mixing equal volumes of 20 mM RBMSL2 and RBD3, heating to 65°C for 10 minutes, and then slowly cooling to room temperature. 100ng of the digested DNA was ligated to a molar excess of the annealed suppression linker (2.7ul of 10uM annealed linker for *MseI* libraries) with 4 weiss units T4 DNA ligase (Promega) in 1 X Ligase Buffer (Invitrogen) overnight (~16hrs) at 15°C, in a final volume of 20µl. After ligation the reaction was heated to 70°C for 10 minutes to inactivate the ligase. Excess linkers and short DNA fragments (i.e., < 100 bp) were removed with the Qiaquick PCR purification system (Qiagen), following the manufacturers protocol, but eluting the DNA in 30ul 5mM TrisHCl pH7.5. To suppress amplification of known transduction locus 10µl of the ligation reaction was incubated with 10 units *Bbs* I (NEB) for 3 hours at 37°C**,** in a final reaction volume of 20ul. Reactions were heated to 65^{o}C for 20 minutes to inactivate the enzyme, cooled on ice, and centrifuged briefly. 1ul of ligated and *Bbs* I digested genomic DNA was amplified in 10ml PCR reactions containing 1 X PCR buffer (45mM Tris HCl pH 8.8, 11mM NH_4SO_4, 0.9mM $MgCl_2$, 6.7mM b-mercaptoethanol, 113 µg /ml BSA, 1mM dNTPs,),1.25 µM RBX4, 1.25µM RB3PA1 and 0.4 units *Taq* DNA polymerase (ABgene). Reactions were cycled in a Tetrad 2 Thermal Cycler (MJ Research / Biorad, Hercules, CA) using the following conditions: 96°C -1min; 30 X [96°C -30s; 59.6°C -30s; 72°C -1min]; 72°C -2min. Primary suppression PCR reactions were diluted 1:50 in Single Molecule Dilution Diluent (SMDD: 5mM Tris HCl pH7.5, 5ng/ul sonicated *E.coli* genomic DNA) and 1µl diluted PCR reaction was added to 9µl secondary PCR reactions containing 1 X PCR buffer (45mM Tris HCl pH 8.8, 11mM NH_4SO_4, 0.9mM $MgCl_2$, 6.7mM b-mercaptoethanol, 113µg/ml BSA, 1mM dNTPs,), 0.625mM RBY1, 0.625mM CM958TD1, 0.4 *Taq* DNA polymerase (ABgene) Reactions were cycled in a Tetrad 2 Thermal Cycler (MJ Research / Biorad, Hercules, CA) using the following conditions: 96°C -1min; 30[96°C -30s; 59.6°C -30s; 72°C -1min]; 72°C -2min. 10ul of secondary PCR products were fractionated on 2% Seakem LE (Cambrex, get location) 0.5X TBE agarose gels against the 100bp ladder (NEB) size marker and visualised by ethidium bromide (0.5 µg/ml) staining. Novel PCR products (i.e. amplicons not corresponding in size to the suppressed known transduction locus) were excised from the gel and purified using the Qiagen Minelute system (Qiagen) following the manufacturers protocol, but eluting the DNA in 10ul of 5mM TrisHCl pH7.5. Purified PCR products were directly sequenced with ABI BigDye Ver. 3.0 ReadyReaction, using 3.3uM RBY1 as the primer. Sequencing reactions were purified using Performa DTR spin columns (Edge BioSystems & Vlt Bio Ltd) and the sequencing data collected using an ABI 3730 capillary sequencer by the PNACL core DNA sequencing service (University of Leicester).

Protocol 3: *TS-ATLAS RP-specific VspI Library Construction and Amplification*

600ng of genomic DNA was digested to completion with 20 units of *VspI* (Promega**)** in the manufacturer's recommended buffer at 37°C for 3 hours. After incubation reactions were heated to 65°C for 20 minutes to inactivate the restriction enzyme. Prior to setting up the ligation reaction, linker oligonucleotides were freshly annealed by mixing equal volumes of 20 mM RBMSL2 and RBD3, heating to 65°C for 10 minutes, and then slowly cooling to room temperature. 100ng of the digested DNA was ligated to a molar excess of the annealed suppression linker (2.7ul of 10uM annealed linker for *VspI* libraries) with 4 weiss units T4 DNA ligase (Promega) in 1X Ligase Buffer (Invitrogen) overnight (~16hrs) at 15°C, in a final volume of 20µl. After ligation the reaction was heated to 70°C for 10 minutes to inactivate the ligase. Excess linkers and short DNA fragments (i.e., < 100 bp) were removed with the Qiaquick PCR purification system (Qiagen), following the manufacturers protocol, but eluting the DNA in 30ul 5mM TrisHCl pH7.5. 1ul of ligated genomic DNA was amplified in 10ml PCR reactions containing 1 X PCR buffer (45mM Tris HCl pH 8.8, 11mM NH_4SO_4, 0.9mM $MgCl_2$, 6.7mM b-mercaptoethanol, 113 µg /ml BSA, 1mM dNTPs,), 1.25µM RBX4, 1.25µM RB3PA1 and 0.4 units *Taq* DNA polymerase (ABgene). Reactions were cycled in a Tetrad 2 Thermal Cycler (MJ Research / Biorad, Hercules, CA) using the following conditions: 96°C -1min; 30 X [96°C -30s; 59.6°C -30s; 72°C -1min]; 72°C -10min. Primary suppression PCR reactions were diluted 1:50 in Single Molecule Dilution Diluent (SMDD: 5mM Tris HCl pH7.5, 5ng/ul sonicated *E.coli* genomic DNA) and 1µl diluted PCR reaction was added to 9µl secondary PCR reactions containing 1 X PCR buffer (45mM Tris HCl pH 8.8, 11mM NH_4SO_4, 0.9mM $MgCl_2$, 6.7mM b-mercaptoethanol, 113µg/ml BSA, 1mM dNTPs,), 0.625mM RBY1, 0.625mM RB011TD1, 0.4 *Taq* DNA polymerase (ABgene) Reactions were cycled in a Tetrad 2 Thermal Cycler (MJ Research / Biorad, Hercules, CA) using the following conditions: 96°C -1min; 30[96°C -30s; 58°C -30s; 72°C -1min]; 72°C -10min. 10ul of secondary PCR products were fractionated on 2% Seakem LE (Cambrex) 0.5X TBE agarose gels against the 100bp ladder (NEB) size marker and visualised by ethidium bromide (0.5 µg/ml) staining. Novel PCR products (i.e. amplicons not corresponding in size to the suppressed known transduction locus) were excised from the gel and purified using the Qiagen Minelute system (Qiagen) following the manufacturers protocol, but eluting the DNA in 10ul of 5mM TrisHCl pH7.5. Purified PCR products were directly sequenced with ABI BigDye Ver. 3.0 ReadyReaction, using 3.3uM RBY1 as the primer. Sequencing reactions were purified using Performa DTR spin columns (Edge BioSystems & Vlt Bio Ltd) and the sequencing data collected using an ABI 3730 capillary sequencer by the PNACL core DNA sequencing service (University of Leicester).

Appendix III:

List of all the oligos and adaptors, which have used for this thesis

Table 1: *list of adaptors used for various library constructions*

Library Construction	Sequence (5′-3′)
RBMSL2	GTGGCGGCCAGTATTCGTAGGAGGGCGCGTAGCATAGAACG
RBMSL3	GTGGCGGCCAGTATTCGTAGGAGGGCGCGTAGCATAGAACGCATG
RBD3	TACGTTCTATGCTAC
RBD4	CGTTCTATGCTACG
RBD5	CGCGTTCTATGCTAC
RBX4	GTGGCGGCCAGTATTC
RBY1	GAGGGCGCGTAGCATAGAAC
RB980TDA2	CAAATTTGTGTACGTTAAATATGTGAAAC
RB980TDA3	TGCTGGTTACACCTCAATAAAGC
CM958TD1	AGAAAAGCAAAATGTCTATTCCG
RB011TD1	AAAAAAAAAAAAAAAAAAAAAGTTTTAAATTT
RRNBOT2	ACTGGTCTAGAGGGTTAGGTTCCTGCTACATCTCCAGCCTCATG
RRNDUP1	AGGCTGGAGATGTAGCAG
RBX1	GTGGCGGCCAGTATTCGTAG

Table 2: *Primers for direct Bisµlphite Sequence analysis for different L1 loci*

Bisulphite-modified primers	Sequence (5′-3′)
VMB885E1	AATTTGATTTTAATGTGGAGGT
VMB885F1	AAATAACCCAATTTTCCAAATA
VMB499D1	TTGAAATTTGAGGTGATTAGAATTT
VMB499E1	AAAAAAAAACTCCCTAACCCC
RBB384E	TTTTGAGTTGTTAAATATGTTTTTTGT
RBB384F	AAAAAAAACTCCTTAACCCCTT
RRB980F	GGGGGTTGTGGAGAATGTAAT
RRB980R	ACCTCGTTACCGCCTTACAA
RRBL980A	TAATTTGTAGAGTAGTAAAATTGT
RRBL980B	CGAAACTATTCCTATTCGAC

Table 3*: Primers sequences used for various PCR*

Primer	Sequence (5′-3′)
RB5PA2	TGGAAATGCAGAAATCACCG
RB3PA2	ACCTAATGCTAGATGACACA
RB3PB	GCACATGTACCCTAAAACTTAG
RBM13F	GTTTTCCCAGTCACGAC
RBM13R	CAGGAAACAGCTATGAC
CM5DP1	ACGCTGGGAGCTGTAGACCG
CM5DP1T1	ACGCTGGGAGCTGTAGACC
CM5DP1T2	ACGCTGGGAGCTGTAGAC
RR0812A	AGACCAGTGATGGAAGACTTGTGC
RR0812B	CTGAGAAATACGCAGTGAGCGAAT

RR0812C	GCTTGATTTAATCTTTCAACAAC
VM164A	TGCCTCCTAGATCGTATTCCC
VM164B	GCACTCTGTGGCATGAAGGT
RB164K	GCTCCTCCCTCTATTATCG
RB164K1	GGCTCCTCCCTCTATTATCG
RB164K2	TGGCTCCTCCCTCTATTATCG
CM1029A	CAGCTCAATTCTGGTGGTTG
CM1029B	TTTCTGGTGACAAAGCTTCAGA
RB696A	CGAGACTGAGCTTTGTAACTC
RB696B	TGCATAGAGTCCACATGAAACC
RR8633A	TAAAGCTAAACAATTATCTAAATCTG
RR8633B	ACTAATCCTATAACCGTTTATTTTC
RB980A	GGCTGTGGAGAATGCAATTGTAAG
RB980B	GCTCTATTCCCAAGGCCTAGAACA
CM011A	TCTGCGGCTTCCTGATTGAG
CM011B	TGGAATGCCCCTCAAAACAA
CM0308A	GACTCTTTCAGTTGCCAGATGC
CM0308B	CCAGTGTAAAAAGATGCGGCT
CM5939A	CTGGAGAGCACGTTCAAACA
CM5939B	GTGCAGGTGTGTAGGTGTGG
CM5888A	TCTGCTGTGCTTTTGCATTC
CM5888B	TCAATGAGCCTCTCCCATTC
CM958A	GAGGCCATAAATCCCCACAT
CM958B	TGTGGAGTGTTTCTCAAACTTTTT
CM8382A	ACCTCTCACCACTCCACCAC
CM8382B	CACTGGACAGGCAGAAACAA
PC4740A	CACACCACTGGAGAGATACGCTTT
PC4740B	CACTTGACTTCTCCCAGCTTTCTG
CM286A	TCCTGAACAACTAATGGGTCAAT
CM286B	CTTGCTCTACCTCTCAACTTTATTGAA

CM7359A	TCCTCACGCACCACACAC
CM7359B	TGCTGTCCTTCTCCTCCTTC
CM1584C	CACACACGCACAGAGGAAAC
CM1584D	TCATTTCCCGTTAAGAACTGTC
CM2182A	CAGATTGTGATAAGGGATAAGAAAAA
CM2182B	GTCAGAGGATGGGGATAGAATG
CM0387A	TTGCATTACTTGCTTGAAATTGA
CM0387B	TGCAGAAGGCCTTACGTTTT
CM0308A	GACTCTTTCAGTTGCCAGATGC
JM0308D	TTTGGATTAAAAAGTTTTAAATTGGGGG

Table 4*: 14 pairs of 454-fusion primers: the linker-specific and L1-specific primers seuqneces are shown in bold, with MIDs shown in blue. Sequences in normal font belong to the 454 A and B primers.*

Primer	Primer sequence 5′-3′
LibA-1A	CGTATCGCCTCCCTCGCGCCATCAGAGTACGCTAT**CCTGCTACATCTCCAGCC**
LibA-1B	CTATGCGCCTTGCCAGCCCGCTCAGACGCTCGACA**CTTCTGCGTCGCTCACGCT**
LibA-2A	CGTATCGCCTCCCTCGCGCCATCAGAGACGCACTC**CCTGCTACATCTCCAGCC**
LibA-2B	CTATGCGCCTTGCCAGCCCGCTCAGAGCACTGTAG**CTTCTGCGTCGCTCACGCT**
LibA-3A	CGTATCGCCTCCCTCGCGCCATCAGATCAGACACG**CCTGCTACATCTCCAGCC**
LibA-3B	CTATGCGCCTTGCCAGCCCGCTCAGATATCGCGAG**CTTCTGCGTCGCTCACGCT**
LibA-4A	CGTATCGCCTCCCTCGCGCCATCAGCGTGTCTCTA**CCTGCTACATCTCCAGCC**
LibA-4B	CTATGCGCCTTGCCAGCCCGCTCAGCTCGCGTGTC**CTTCTGCGTCGCTCACGCT**
LibA-7A	CGTATCGCCTCCCTCGCGCCATCAGATACGACGTA**CCTGCTACATCTCCAGCC**
LibA-7B	CTATGCGCCTTGCCAGCCCGCTCAGTCACGTACTA**CTTCTGCGTCGCTCACGCT**
LibA-8A	CGTATCGCCTCCCTCGCGCCATCAGCGTCTAGTAC**CCTGCTACATCTCCAGCC**
LibA-8B	CTATGCGCCTTGCCAGCCCGCTCAGTCTACGTAGC**CTTCTGCGTCGCTCACGCT**
LibA-9A	CGTATCGCCTCCCTCGCGCCATCAGTGTACTACTC**CCTGCTACATCTCCAGCC**
LibA-9B	CTATGCGCCTTGCCAGCCCGCTCAGACGACTACAG**CTTCTGCGTCGCTCACGCT**
LibA-10A	CGTATCGCCTCCCTCGCGCCATCAGCGTAGACTAG**CCTGCTACATCTCCAGCC**
LibA-10B	CTATGCGCCTTGCCAGCCCGCTCAGTACGAGTATG**CTTCTGCGTCGCTCACGCT**
LibA-11A	CGTATCGCCTCCCTCGCGCCATCAGTACTCTCGTG**CCTGCTACATCTCCAGCC**
LibA-11B	CTATGCGCCTTGCCAGCCCGCTCAGTAGAGACGAG**CTTCTGCGTCGCTCACGCT**
LibA-12A	CGTATCGCCTCCCTCGCGCCATCAGTCGTCGCTCG**CCTGCTACATCTCCAGCC**
LibA-12B	CTATGCGCCTTGCCAGCCCGCTCAGACATACGCGT**CTTCTGCGTCGCTCACGCT**
LibA-13A	CGTATCGCCTCCCTCGCGCCATCAGACGCGAGTAT**CCTGCTACATCTCCAGCC**
LibA-13B	CTATGCGCCTTGCCAGCCCGCTCAGACTACTATGT**CTTCTGCGTCGCTCACGCT**
LibA-14A	CGTATCGCCTCCCTCGCGCCATCAGACTGTACAGT**CCTGCTACATCTCCAGCC**
LibA-14B	CTATGCGCCTTGCCAGCCCGCTCAGAGCGTCGTCT**CTTCTGCGTCGCTCACGCT**

Appendix IV:

Raw numbers of reads for amplicon processing vs shotgun processing software (Roche version 2.3)

All three regions	Amplicon	Shutgun
Number of Reads	438287	685443
Average Length	191.7	200.3
std Dev	**68.2**	**83.4**
Region 1		
Number of Reads	162443	251865
num reads with intact MIDs	161912	241636
% mappable	**99.7**	**95.9**
Region 2		
Number of Reads	173520	212618
num reads with intact MIDs	173082	205135
% mappable	**99.7**	**96.5**
Region 3		
Number of Reads	102324	220960
num reads with intact MIDs	101981	210228
% mappable	**99.7**	**95.1**

Appendix V:

Galaxy Workflow constructed from history 'E3T1-3 Total DNA"

Step 1: *Input dataset: E3T1-3.refion-1.fa*

Step 2: *Input dataset: E3T1-3.refion-2.fa*

Step 3: *Input dataset: E3T1-3.refion-3.fa*

Step 4: *Input dataset: Badge_lab_oligo.fa*

Step 5: *Concatenate datasets: Dataset 1, Dataset 2, Dataset 3*

Step 6: *Collapse: Library to collapse output dataset 'output' from step 5*

Step 7: *Filter sequences from step 6 by length:*

Fasta file

Minimal length

30

Maximum length

0

Step 8: *LastZ- align Badge_lab_oligo dataset from step 4 against reference sequences that are locally cached*

Using reference genome

/galaxy/data/hg19/seq/hg19.2bit (value not yet validated)

Output format

Intervals

Lastz settings to use

Commonly used

Select mapping mode

Roche-454 98% identity

Do you want to modify the reference name?

No

Do not report matches below this identity (%)

95

Do not report matches above this identity (%)

100

Do not report matches that cover less than this percentage of each read

0

Step 9: *Get flanks for mapped oligos from step 8*

Region

Whole feature

Location of the flanking region/s

Both

Offset

0

Length of the flanking region(s)

500

Step 10*: Lastz- Align sequencing reads in Output dataset 'output' from step 7 against reference sequences that are locally cached*

Using reference genome

/galaxy/data/hg19/seq/hg19.2bit (value not yet validated)

Output format

Tabular

Lastz settings to use

Commonly used

Select mapping mode

Roche-454 98% identity

Do you want to modify the reference name?

No

Do not report matches below this identity (%)

95

Do not report matches above this identity (%)

100

Do not report matches that cover less than this percentage of each read

0

Convert lowercase bases to uppercase

Yes

Step 11: *Intersect*

Return

Overlapping Intervals of

Output dataset 'output1' from step 10

that intersect

flank oligo_dataset sequences from step 9

for at least

1

Step 12: *Cut*

Cut columns

c10

Delimited by

Tab

From

Output dataset 'output' from step 11

Step 13: *Subtract Whole Dataset*

Subtract

Output dataset 'out_file1' from step 12

from

Output dataset 'output' from step 7

Restrict subtraction between 'begin column'

1 (value not yet validated)

and 'end column'

1 (value not yet validated)

Step 14: *Join two Datasets*

Join

Output dataset 'output' from step 13

using column

1 (value not yet validated)

with

Output dataset 'output' from step 7

and column

1 (value not yet validated)

Keep lines of first input that do not join with second input

No

Keep lines of first input that are incomplete

No

Fill empty columns

No

Step 15: *Cut*

Cut columns

c2,c3

Delimited by

Tab

From

Output dataset 'out_file1' from step 14

Step 16: *Tabular-to-FASTA*

Tab-delimited file

Output dataset 'out_file1' from step 15

Title column(s)

1 (value not yet validated)

Sequence column

2 (value not yet validated)

Step 17: *Input dataset: L1.3_first_50bp_seq.fa*

Step 18: *water*

Sequence 1

Output dataset 'output' from step 17

Sequence 2

Output dataset 'output' from step 16

Gap open penalty

15.0

Gap extension penalty

1.0

Brief identity and similarity

Yes

Output Alignment File Format

FASTA (m)

Step 19: *FASTA-to-Tabular*

Convert these sequences

Output dataset 'out_file1' from step 18

How many columns to divide title string into?

1

How many title characters to keep?

0

Appendix VI:

Perl scripts used to analyse 454-reads

Perl Script 1*: to separate the sequences according to their MIDs*

```
#!/usr/bin/perl
use strict;
use warnings;

my $wanted =$ARGV[0];

################################ loop thro MIDs

my $MID_file = "/Users/rahelehrahbari/Desktop/MIDlist.text";
#my @MID_seqs;
my %lib_by_MIDs;
my %AB_by_MIDs;
my %BMID_by_lib;
open (MID_IN, "<$MID_file") or die "\n\n\n".__LINE__." - cant open $MID_file !!!!!!!!!!!!!!!!!!!!!\n\n\n";
while( <MID_IN> ){
    my $row = $_;
    #print "$row";
    chomp ($row);
    my @row_array = split(/\t/, $row); # splits the row into cols , whenever it sees a tab (\t)
    my $lib = $row_array[2];
    my $A_or_B= $lib;
    $A_or_B=~ s/LibA\-\d+//g;
    #print $A_or_B;

    $lib =~ s/.$//g;  # substitude last single character with nothing
    my $seq = $row_array[3];
    #print "$MID_name\n";
    my @seq_array = split(//, $seq); # make array of bases in seq
    #print "$seq_array[0]$seq_array[1]$seq_array[2]\n";
    my @MID_seq = "@seq_array[25..34]"; # the mid seq if bases 25-34 of the primer-plus-mid seq
    my $MID_seq = join ('',@MID_seq);
    $MID_seq =~ s/ //g; #sub any space with nothing globaly
    #print "$MID_seq\n";
    #push (@MID_seqs, $MID_seq);
    $lib_by_MIDs{$MID_seq}=$lib;
```

```
    #print "$lib $MID_seq\n";
    $AB_by_MIDs{$MID_seq}=$A_or_B;
    $BMID_by_lib{$lib}=$MID_seq if ($A_or_B eq "B");       # save seq for b mid for this lib

    my $RC_seq = &RC ($MID_seq);


     $lib_by_MIDs{$RC_seq}=$lib;

    $AB_by_MIDs{$RC_seq}=$A_or_B;

}
close MID_IN;

############################## loop thru seqs

my %seq_number_by_lib;
my %MIDs_in_no_order_by_seq_number;
my %seq_by_number;
my $seq_file = "//Users/rahelehrahbari/Desktop/shotgun processing/Copy of 3.TCA.454Reads.txt";
open (SEQ_IN, "<$seq_file") or die "\n\n\n".__LINE__." - cant open $seq_file !!!!!!!!!!!!!!!!!!!!!\n\n\n";

my $seq_name =0;
my $count_all_seqs =0;
my $count_lib_seqs =0;
$seq_by_number{$seq_name} ="";
while( <SEQ_IN> ){
    my $row = $_;
    #print "row in\n$row\n\n----------\n\n";
    $row =~ s/\r/\n/g;
    chomp ($row);


    if ($row =~ m/\>(.+) rank/){          # for each title
            $count_all_seqs ++;
            ########## the last sequence is complete so process it

                    my $sequence = $seq_by_number{$seq_name};
                    my $MIDs_in_NO_order = "";
            foreach my $MID_seq ( keys %lib_by_MIDs ) {
                                my $lib =$lib_by_MIDs{$MID_seq};
                                my $A_or_B =$AB_by_MIDs{$MID_seq};
```

```
                        #print "$MID_seq $lib\n";
                        if ($sequence =~ m/$MID_seq/) {
                        $MIDs_in_NO_order .= "$A_or_B";
                        #print "$MID_seq $lib is in $seq_name\n";
                                if ($seq_number_by_lib{$lib}){
                                        if ($seq_number_by_lib{$lib} !~ m/$seq_name/){
                                                $seq_number_by_lib{$lib} .= " $seq_name";
                                                }
                                  }else{
                                        $seq_number_by_lib{$lib} = "$seq_name";
                                }
                        }
                }
                #print "$seq_name `\n";
                $MIDs_in_no_order_by_seq_number{$seq_name} = "$MIDs_in_NO_order";

        ############ and this line is the title row

        $seq_name = $1;
  }else{

        ########  this line is a seq row

        my $sequence_row = $row;
        if ($seq_by_number{$seq_name}){
                $seq_by_number{$seq_name} .= "$sequence_row";
        }else{
                $seq_by_number{$seq_name} = "$sequence_row";
        }
  }
}

close SEQ_IN;

############################### print out

my $AB_count =0;
my $Aonly_count =0;
my $Bonly_count =0;
my $BA_count =0;

open (OUT, ">/Users/rahelehrahbari/Desktop/out$wanted.fa");
```

```
open (A, ">/Users/rahelehrahbari/Desktop/A_$wanted.fa");
open (B, ">/Users/rahelehrahbari/Desktop/B_$wanted.fa");
open (AB, ">/Users/rahelehrahbari/Desktop/AB_$wanted.fa");
open (BA, ">/Users/rahelehrahbari/Desktop/BA_$wanted.fa");
foreach my $lib ( keys %seq_number_by_lib ) {      # only sequence seq_names in each library
        my $list_of_seq =$seq_number_by_lib{$lib};
        if ($lib eq "LibA-$wanted") {
                #print "$lib \n\n";

    my @seq_names = split(' ', $list_of_seq);
                foreach my $seq_name (@seq_names) {


                        $count_lib_seqs ++;
                        my $MIDs_in_no_order =$MIDs_in_no_order_by_seq_number{$seq_name};  #
                        my $sequence =$seq_by_number{$seq_name};
                        my $sequence_without_primers = $sequence;



                        my $linker_of_A_primer = "CCTGCTACATCTCCAGCC";
                        my $BMID = $BMID_by_lib{$lib};
                        my $RC_A = &RC($linker_of_A_primer);
                        my $RC_B = &RC($BMID);


                        #print "\$linker_of_A_primer = CCTGCTACATCTCCAGCC\n";
                        #print "\$BMID = $BMID\n";
                        #print "\$RC_A = $RC_A\n";
                        #print "\$RC_B = $RC_B\n\n";


                        #my @left_ends = ($linker_of_A_primer,$RC_B);
                        #my @right_ends = ($BMID,$RC_A);
                        #my @right_ends = ($linker_of_A_primer,$RC_B);
                        #my @left_ends = ($BMID,$RC_A);
                        #my @right_ends = ($linker_of_A_primer,$RC_A);
                        #my @left_ends = ($BMID,$RC_B);
                        #my @right_ends = ($linker_of_A_primer,$BMID);
                        #my @left_ends = ($RC_A,$RC_B);
                        my @left_ends = ($linker_of_A_primer,$BMID);
                        my @right_ends = ($RC_A,$RC_B);
```

```
#print "$sequence_without_primers\n";
#print "right:\n";
foreach my $linker_or_MID_shortening (@right_ends){

my $regex = "$linker_or_MID_shortening.*\$";
#print "before\n";
($linker_or_MID_shortening, $sequence_without_primers) =
&substitution($linker_or_MID_shortening, $regex, $sequence_without_primers);
#print "after\n";

while ($linker_or_MID_shortening  =~ m/[ACGT][ACGT]/){
my $regex = "$linker_or_MID_shortening\$";
($linker_or_MID_shortening,
$sequence_without_primers) = &substitution($linker_or_MID_shortening, $regex,
$sequence_without_primers);
$linker_or_MID_shortening =~ s/.$//;
}
}

#print "left:\n";
foreach my $linker_or_MID_shortening (@left_ends){

my $regex = "^.*$linker_or_MID_shortening";
($linker_or_MID_shortening, $sequence_without_primers) =
&substitution($linker_or_MID_shortening, $regex, $sequence_without_primers);

while ($linker_or_MID_shortening  =~ m/[ACGT][ACGT]/){
my $regex = "^$linker_or_MID_shortening";
($linker_or_MID_shortening, $sequence_without_primers) =
&substitution($linker_or_MID_shortening, $regex, $sequence_without_primers);
$linker_or_MID_shortening =~ s/^.//;
}

}

#print "\n----------
\n$seq_name\t$MIDs_in_no_order\n$sequence_without_primers\n\n----------\n";
if ($MIDs_in_no_order =~ m/AB/){
$AB_count++;
print AB "$seq_name $sequence_without_primers
$MIDs_in_no_order\n";
}
if ($MIDs_in_no_order eq "A"){
$Aonly_count++;
```

```
                                        print A "$seq_name $sequence_without_primers
        $MIDs_in_no_order\n";
                                }
                                if ($MIDs_in_no_order eq "B"){
                                        $Bonly_count++;
                                        print B "$seq_name $sequence_without_primers
        $MIDs_in_no_order\n";
                                }
                                if ($MIDs_in_no_order =~ m/BA/){
                                        $BA_count++ if ($seq_name !~ m/poscont/);
                                        print BA "$seq_name $sequence_without_primers
        $MIDs_in_no_order\n";
                                }




                                print OUT
">$seq_name"."_$MIDs_in_no_order\n$sequence_without_primers\n";
                        }
        }
}


print "A  - $Aonly_count\nB  - $Bonly_count\nAB - $AB_count\nBA - $BA_count\n...but we're not yet sure
that AB/BA is really telling us the order in the seq\n";


close OUT;

close A;

close B;

close AB;

close BA;


my $percent = 100*$count_lib_seqs/$count_all_seqs;

#print
"\n=================\n\$count_lib_seqs\t$count_lib_seqs\n\$count_all_seqs\t$count_all_seqs\n\$perc
ent\t$percent\n";


############## sub routines


sub RC{
                                                my $seq = shift;
                                                # Reverse and ...
        my @seq = split('', $seq);
        @seq = reverse(@seq);
        my $RC_seq = join('',@seq);
```

```
        #....complement
        my %replace = (
          A  => "T",
          a  => "t",
          T  => "A",
          t  => "a",
          C  => "G",
          c  => "g",
          G  => "C",
          g  => "c",
        );
        my $regex = join "|", keys %replace;
        $regex = qr/$regex/;
        $RC_seq =~ s/($regex)/$replace{$1}/g;
        return $RC_seq;
}


sub substitution{                                       # to remove primers
                                #print "start\n";
                                my ($linker_or_MID_shortening, $linker_or_MID_regex,
$sequence_without_primers) = @_;
                                if ($linker_or_MID_shortening  =~ m/[ACGT][ACGT]/){ # if its a seq of
two+ bases
                                        #print "is $linker_or_MID_shortening something\n";
                                        if ($sequence_without_primers =~
m/$linker_or_MID_regex/){
                                                $sequence_without_primers =~
s/$linker_or_MID_regex//;
                                                #print "$linker_or_MID_regex            >
        $sequence_without_primers\n";
                                                $linker_or_MID_shortening = "";
                                        }
                                }
                                #print "end\n";
         return ($linker_or_MID_shortening, $sequence_without_primers);
}
```

Perl Script 2: *To calculate the length of the poly G nucleotide track in the sequences*

```
#!/usr/bin/perl
use strict;
use warnings;

#pooling sequences with G track >= 4 bp

my %query_seq;
my $dir= "/Users/rahelehrahbari/Desktop/Amplicon_processing";

my $dir1="$dir/454Reg1fastafiles";
my $dir2="$dir/454Reg2fastafiles";
my $dir3="$dir/454Reg3fastafiles";
my $data_FILES_and_paths ="";
opendir DIR1, $dir1 or die "cannot open dir $dir1: $!";
my @data_set1= readdir DIR1;
closedir DIR1;
foreach my $data_FILE (@data_set1){
	if ($data_FILE !~ m/^\./){
		$data_FILES_and_paths =$data_FILES_and_paths." ".$dir1."/".$data_FILE;
	}
}
opendir DIR2, $dir2 or die "cannot open dir $dir2: $!";
my @data_set2= readdir DIR2;
closedir DIR2;
foreach my $data_FILE (@data_set2){
	if ($data_FILE !~ m/^\./){
		$data_FILES_and_paths =$data_FILES_and_paths." ".$dir2."/".$data_FILE;
	}
}
opendir DIR3, $dir3 or die "cannot open dir $dir3: $!";
my @data_set3= readdir DIR3;
closedir DIR3;
foreach my $data_FILE (@data_set3){
	if ($data_FILE !~ m/^\./){
		$data_FILES_and_paths =$data_FILES_and_paths." ".$dir3."/".$data_FILE;
	}
}
```

```
my $data_set= "$dir/data_set.fasta";

#print "cat $data_FILES_and_paths > $data_set";
system "cat $data_FILES_and_paths > $data_set";

open (DATA_SET, "<$data_set") or die "\n\n\n".__LINE__." - cant open $data_set !!!!!!!!!!!!!!!!!!!!!!\n\n\n";
my %data_seq;

my $title;
my $seq;
my $has_4_gs;
my $polyg_track_sequence= "/Users/rahelehrahbari/Desktop/polyg_track_sequence.fasta";
open (POLYG, ">$polyg_track_sequence") or die "\n\n\n".__LINE__." - cant open $polyg_track_sequence
!!!!!!!!!!!!!!!!!!!!!\n\n\n";

while( <DATA_SET> ){
        my $row = $_;
        #print "$row";
        chomp ($row);
        $row=~ s/\r//;

        if ($row =~ m/\>(.+)/){
                $title=$1;
                $seq="";
                $has_4_gs = "no";
        }else{
                $seq= $seq.$row;
                if ($seq =~ m/ggggggggg/i){

                        $has_4_gs = "yes";
                }
                $data_seq{$title}= $seq;
        }

       if ($has_4_gs eq "yes"){

                my $length = length($seq);
                if ($length >= 15){
                        print POLYG ">$title\n$seq\n";
                }
       }
}
```

```
close POLYG;
close DATA_SET;
```

Perl Script 3*: To trim the sequences for their L1 sequences, which are with aligned first 50bp of L1.3 sequence*

```
#!/usr/bin/perl
use strict;
use warnings;

print "start\n";

my %intersect_list;
my $water_data = "/Users/rahelehrahbari/Desktop/Amplicon processing/Water data tabular format/GalaxywaterReg3Tabular/GalaxywaterReg3lib11.txt";
open (water_data_IN, "<$water_data") or die "\n\n\n".__LINE__." - cant open $water_data !!!!!!!!!!!!!!!!!!!!\n\n\n";
while( <water_data_IN> ){
    my $row = $_;
    print "$row";
    chomp ($row);
    my @row_array = split(/\t/, $row); # splits the row into cols , whenever it sees a tab (\t)
    my $name = $row_array[0];
    my $L1_seq = $row_array[1];
    $L1_seq =~s/\-//;
    $name =~s/\>//;
    $intersect_list{$name}=$L1_seq;
}
#close intersect_IN;
close water_data_IN;

my $fasta_file = "/Users/rahelehrahbari/Desktop/Amplicon processing/Intersect/Intervals_Reg3/int.reg3.lib11.fa";

open (fasta_IN, "<$fasta_file") or die "\n\n\n".__LINE__." - cant open $fasta_file !!!!!!!!!!!!!!!!!!!!\n\n\n";
my $out_file= "/Users/rahelehrahbari/Desktop/Amplicon processing/unmappable seq. without L1 seq (final results)/Reg1/WATER_file.Reg3.lib11.fa";
open (OUT, ">$out_file") or die "\n\n\n".__LINE__." - cant make $out_file !!!!!!!!!!!!!!!!!!!!\n\n\n";

#my $print_switch = "default";
```

```
my $name;
while( <fasta_IN> ){
   my $row = $_;
   #print "$row";
   chomp ($row);
   if ($row =~ m/^\>(.+)$/){
                        $name = $1;
#                       if (exists $intersect_list{$name}){
#                               #print "dont want $row\n";
#                               $print_switch = "off";
#                       }else{
#                               #print "$row\n";
#                               $print_switch = "on";
#                       }
            }else{
                        my $seq = $row;
                        print "|$name|\n";
                        my $L1_seq = $intersect_list{$name};
                        $seq =~ s/$L1_seq//;
                        print OUT ">$name\n$seq\n";
            }
}

close fasta_IN;
close OUT;
```

Appendix VII:

Genomic locations of all candidate novel L1 sequences

Sequence ID	***Genomic location***
>12405-1	chrX:118,770,697-118,770,845
>9312-1	chr1:75,665,828-75,665,883
>21474-1	chr11:71,460,788-71,460,878
>20324-1	chr9:89,891,676-89,891,758
>2778-1	chr12:9,680,619-9,680,741
>14921-1	chr1:191,060,592-191,060,699
>14561-1	chr14:71,197,761-71,197,79
>16880-1	chrX:71,359,579-71,359,847
>5217-1	chr4:132,181,662-132,181,784
>14267-1	chr6:100,802,476-100,802,577
>10474-1	chr3:24,136,912-24,137,245
>20884-1	chr9:128,349,258-128,349,337
>1505-2	chr9:15,873,025-15,873,163
>2925-1	chr2:86190756-86190852
>13587-1	chr4:159467742-159467811
>13229-1	chr12:66,016,049-66,016,141
>16465-1	chr17:64,744,646-64,744,773
>19912-1	chr7:100,142,209-100,142,289
>20630-1	chr3:85,576,497-85,576,566
>17927-1	chr4:132,181,648-132,181,784
>13305-1	chr7:147,978,093-147,978,143
>14482-1	chr7:2,804,232-2,804,429
>10949-1	chr3:186,371,013-186,371,178
>2447-2	chr7:144,856,013-144,856,080
>8643-1	chr3:186,371,042-186,371,177

>17080-1 chr8:62,035,243-62,035,519

>7879-1 chr4:104,817,111-104,817,277

>2505-2 chr3:81,616,259-81,616,348

>27891-1 chr16:47,277,448-47,277,591

>34267-1 chr16:63,315,530-63,315,703

>13358-1 chr1:171,253,495-171,253,572

>1684-2 chrX:76,809,050-76,809,117

>18060-1 chr3:20,750,431-20,750,628

>2871-2 chr1:38,141,872-38,141,966

>15182-1 chr20:15,011,472-15,011,631

>24670-1 chr9:13,807,548-13,807,744

>24422-1 chr9:106,985,465-106,985,543

>9351-1 chr14:59,424,306-59,424,533

>4454-1 chr12:25,991,340-25,991,426

>9508-1 chr20:12,854,888-12,855,045

>4492-1 chr1:49,891,598-49,891,867

>17068-1 chr4:13,895,910-13,896,209

>18174-1 chrX:143,087,825-143,088,018

>12000-1 chrX:68,595,469-68,595,694

>23313-1 chr11:61,303,848-61,304,083

>5465-1 chrX:132,820,022-132,820,347

>25139-1 chr15:92,043,093-92,043,419

>27287-1 chr16:56,442,982-56,443,412

>34297-1 chr8:40,516,138-40,516,441

>20372-1 chr13:108,373,917-108,374,186

>12706-1 chr9:104,829,543-104,829,765

>27349-1 chrX:129,982,472-129,982,742

>2241-2 chr7:151,690,748-151,690,990

>7799-1 chr20:45,444,601-45,444,844

>7888-1 chr22:28,035,493-28,035,728

>19433-1 chr11:129,981,821-129,982,008

>5773-1 chr20:41,010,147-41,010,279

>21733-1 chr9:103,553,550-103,553,689

>33469-1 chr4:162,209,409-162,209,482

>17241-1 chrX:12,087,989-12,088,297

>22338-1 chr14:22,628,314-22,628,491

>24033-1 chr20:15,011,472-15,011,665

>9822-1 chrX:49,492,261-49,492,581

>32681-1 chr15:98,705,608-98,705,876

>33781-1 chr4:162,209,392-162,209,473

>21548-1 chr9:76,879,070-76,879,342

>19236-1 chrX:68,552,698-68,552,870

>19338-1 chrX:69,156,737-69,157,015

>3002-1 chr12:97,956,987-97,957,181

>8050-1 chr8:40,516,001-40,516,166

>6358-1 chr9:16,161,460-16,161,540

>17188-1 chr15:98,705,608-98,705,876

>1318-3 chr3:60,397,878-60,397,917

>1367-3 chr2:76,542,471-76,542,549

>34441-1 chr22:30,026,964-30,027,132

>30485-1 chr16:65,537,639-65,537,875

>13819-1 chr16:56,442,979-56,443,101

>19489-1 chr15:98,711,534-98,711,638

>20334-1 chr12:52,169,637-52,169,679

>13730-1 chr22:24,761,019-24,761,103

>11276-1 chr14:99,448,071-99,448,357

>30308-1 chr14:35,319,619-35,319,728

>36208-1 chr13:108,373,991-108,374,236

>25318-1 chr12:59,733,825-59,734,109

>2335-2 chr1:193,207,703-193,208,104

>35381-1 chr12:47,937,498-47,937,719

>4942-1 chr12:47,937,717-47,937,889

>35502-1 chrX:136,163,531-136,163,750

>608-5 chr1:207,125,544-207,125,622

>30835-1 chr1:193,207,837-193,207,959

>11455-1 chr11:42,457,119-42,457,375

>34055-1 chr4:149,246,217-149,246,292

>32773-1 chr8:40,515,067-40,515,406

>35541-1 chr14:60,894,365-60,894,550

>15271-1 chr7:127,185,717-127,185,885

>6814-1 chr9:14,391,858-14,392,056

>30809-1 chr9:76,880,797-76,880,877

>19735-1 chr12:25,991,340-25,991,410

>34001-1 chrX:143,087,828-143,088,018

>5094-1 chr15:44,246,449-44,246,570

>14419-1 chr8:40,516,172-40,516,454

>20976-1 chr2:31,123,693-31,123,820

>10476-1 chr4:132,181,648-132,181,784

>18965-1 chr20:40,834,790-40,834,859

>27725-1 chr4:121,569,648-121,569,734

>13519-1 chr3:85,576,497-85,576,558

>19885-1 chr3:186,371,705-186,372,006

>2280-2 chr8:126,626,462-126,626,544

>2158-2 chr15:28,461,705-28,461,762

>1172-3 chr4:188,651,021-188,651,062

>16382-1 chr1:214,078,689-214,078,876

>2357-2 chr3:126272496-126272568

>11865-1 chr1:230,850,014-230,850,245

>7278-1 chr1:179,575,408-179,575,606

>22317-1 chr10:131,467,241-131,469,860

>12658-1 chr3:84,455,786-84,458,795

>13319-1 chrX:22,309,578-22,311,747

>10216-1 chr14:31,149,715-31,152,144

>18121-1 chr20:40,833,261-40,835,160

>25417-1 chr11:114,894,558-114,896,307

>6200-1 chr12:43,429,358-43,429,569

>9418-1 chr3:80,590,163-80,590,294

>20364-1 chr3:80,590,163-80,590,294

>5091-1 chr20:40,834,296-40,834,475

>71458-1 chr16:32,911,882-32,911,974

>29707-1 chr16:33,761,192-33,764,761

>39969-1 chr16:33,761,663-33,762,151

>48201-1 chr10:131,468,454-131,468,680

>28263-1 chr4:25,383,048-25,383,230

>83224-1 chr21:35,307,127-35,307,312

>78149-1 chr16:33,764,089-33,764,255

>44757-1 chr21:35,308,285-35,308,531

>65200-1 chr10:84,109,843-84,109,879

>27542-1 chr22:19,210,829-19,210,913

>53160-1 chrY:7,796,391-7,796,595

>45696-1 chr1:220,289,017-220,289,258

>38151-1 chr17:79,332,165-79,332,441

>63588-1 chr8:40,433,179-40,433,273

>4896-2 chr2:12,667,885-12,667,935

>70549-1 chr22:22,714,729-22,715,020

>80037-1 chr10:50,450,354-50,450,508

>62729-1 chr2:12,667,886-12,667,935

>64987-1 chr6:100,802,476-100,802,572

>74370-1 chr11:114,895,346-114,895,521

>39231-1 chr16:32,909,976-32,910,091

>27572-1 chr2:198,560,203-198,560,369

>48064-1 chr7:130,573,546-130,573,639

>4549-3 chr20:51,780,144-51,780,263

>65414-1 chr2:242,263,550-242,263,674

>3627-3 chr1:174,586,912-174,586,941

>64243-1 chr21:35,304,649-35,305,026

>3741-3 chr4:94,567,925-94,568,058

>41554-1 chr21:35,307,127-35,307,312

>21088-1 chr12:9,681,228-9,681,369

>27250-1 chr6:66,567,219-66,567,356

>54324-1 chr4:131,568,009-131,568,282

>78301-1 chr2:99,887,419-99,887,522

>81296-1 chr19:54,063,202-54,063,285

>61858-1 chrX:78,092,155-78,092,412

>73914-1 chr22:22,715,184-22,715,416

>21962-1 Chr12:118,608,935-118,610,304

>36416-1 chr1:225,614,384-225,615,943

>32739-1 chr10:124,454,401-124,455,910

>29329-1 chr1:121,143,073-121,143,562

>38613-1 chr14:31,149,726-31,152,135

>12782-1 chr1:110,405,527-110,408,156

>42159-1 chr6:123,852,344-123,855,903

>25625-1 Chr3:85,576,518-85,576,568

>26012-1 chr5:103,055,215-103,055,377

>25426-1 chr10:116,329,838-116,331,757

>10795-1 chr18:33,362,242-33,364,381

>30772-1 chr21:34,974,844-34,976,623