Directly accessing *de novo* L1 retrotransposition in the human germline

Thesis submitted for the degree of
Doctor of Philosophy
at the University of Leicester

by

Peter J. Freeman
Department of Genetics
University of Leicester

October 2007

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Table of contents

Directly accessing *de novo* L1 retrotransposition in the human germline

Peter J. Freeman

Abstract

Human chromosomes are riddled with sequence elements that have in our evolutionary past reproduced and "jumped" to new locations. Long Interspersed Nuclear Element 1 (L1) is the only human autonomous retrotransposon thought to be currently active. L1 retrotransposition has significantly shaped our genome via insertional mutagenesis, sequence transduction, pseudogene formation, and ectopic recombination. However, L1 retrotransposition dynamics are little understood in the germline since *de novo* insertions occur infrequently (one event genome-wide per 16 - 500 haploid genomes).

The method developed in this investigation was capable of recovering full-length L1 insertions, from the single-molecule level, using hybridisation enrichment to capture rare L1-containing DNA from an aliquot of multiplex amplified DNA. Enrichment was achieved by the binding of biotinylated oligonucleotides (bio-oligos) to L1 sequences and their physical retrieval using streptavidin-coated paramagnetic beads. Using bio-oligos on total sperm DNA would be futile due to L1 comprising ~17 % by mass of the genome, and so the search area was limited to genomic sections (~5 kb) devoid of matches to the bio-oligo sequences. PCR had to be extremely efficient as a *de novo* insertion would appear once in a pool of sperm DNA. Multiplex PCR could amplify full-length L1 insertions into ~5 kb target sites to levels recoverable by hybridisation enrichment from single-molecules, despite the presence of overwhelming masses of empty target sites.

After screening ~600 μg of sperm DNA, no *de novo* L1 insertions were recovered. I propose a rate (within previous estimates) of < 3 insertions in 290 haploid genomes. This low rate may not be inconsistent with ongoing L1 proliferation. Rather the data supports the idea that L1s may retrotranspose rarely during embryogenesis before germline differentiation. Such "jackpot" insertions colonise numerous gametes, so have a greater chance of being inherited, even if this occurs in very few individuals.

...

Acknowledgements

First and foremost I would like to thank Alec and Richard for their guidance throughout this project. Alec said from day one that it would be a difficult (perhaps impossible) project, so thanks to both of my supervisors for their support.

I would also like to thank the people in Richard's empire: Pam for keeping the group running and organised (she is the real group leader Richard!); Catriona for assisting me with numerous extractions, and finally descending to my level of humour; and finally to Rob who provided me with various boring bio-informatics information which made the project possible (good luck with your thesis mate). There are also a few other people in lab G18/G19 who I'd like to mention: Thanks to Rita for providing PCR buffers by the pint; Esther for her enthusiastic advice and for teaching me the basics of DEASH; Foxy for keeping the lab (and its members) in working order; Matt and Jennie J who were my recreational gurus (teaching me the joys of Bloons and Celebdaq); Vic for teaching me several words that I haven't even heard on the rugby pitch; and to the staff in the media kitchen. I would like to thank all the other members of labs G18/G19 (past and present), including Yuri and his minions. Also, thanks to Pat for teaching me to play squash and then telling everyone that I get beaten by a lady in her 60s.

Thanks to Tony and Colin, for giving me a job even though I hadn't finished my thesis. It was a life-saver!

I'd like to acknowledge everyone at the mighty Aylestone Athletic RUFC who kept me going by making me roll about in the mud to fight for an odd shaped ball. Thanks to Lowry and Fester for ensuring I always had a beer, even when my funding ran out. Also to Pig, Steve O and old man Watson for mindless chatter every Saturday after games.

A special thank you to my parents and grandparents who have supported me since the start of my PhD, and also to Kimberley's parents who have encouraged me over the last year. Thanks also Jo for being such a great sister, but still sniggers every time she says the word thesis.

Finally I want to thank Kimberley. She has been a continuous support since we met, and has got me through the last year. I know I would not have got to this stage without your help, so it's a really good job that Tim introduced us (thanks mate).

Abbreviations

Standard terms

Aa	Amino acid
APC	Adenomatosis polyposis coli
AP	Apurinic/apyrimidinic
ASP	Antisense promoter
bp	Base pairs
C	Cysteine rich C terminal domain
cDNA	Complementary DNA
CGD	Chronic granulomatous disease
CHM	Choroideremia
DMD	Duchenne muscular dystrophy
EN	Endonuclease
FCMD	Fukuyama-type congenital muscular dystrophy
FIX	Factor 9
gDNA	Genomic DNA
HBB	Haemoglobin B
HOXD	Homeobox D
HS	Human-specific
IRES	Internal ribosome entry site
IS	Insertion sequences
LINE	Long interspearsed nuclear element
LTR	Long terminal repeat
mRNA	Messenger RNA
rRNA	Ribosomal RNA
tRNA	Transfer RNA
MPAS	Major polyadenylation signal
MYA	Million years ago
MY	Million years

ORF	Open reading frame
ORF1p	Open reading frame 1 protein
ORF2p	Open reading frame 2 protein
PDH	Pyruvate dehydrogenase
Pol	Polymerase
RC	Retrotransposition Competant
RNA Pol	RNA polymerase
RNP	Ribonucleoprotein
RP2	Retinitis Pigmentosa 2
RT	Reverse transcriptase
SINE	Short interspearsed nuclear element
SMC	Single-molecule clean
SRP	Signal recognition particle
SS	Single-stranded
TPRT	Target-primed reverse transcription
tRNA	Transfer RNA
TSD	Target site duplication
UTR	Untranslated region

Chemicals

BSA	Bovine Serum Albumen
DHB	Denaturing and Hybridiastion buffer
ED	Elution mix
EtBr	Ethidium bromide
NaAc	Sodium Acetate
SSC	Salt sodium citrate
SDS	Sodium dodecyl sulphate

Processes

ATLAS	Amplification Typing of L1 Active Sub-families

HE	Hybridisation Enrichment
PCR	Polymerase Chain Reaction

Chapter 1: Literature Review

1.1 Repetitive elements in the human genome

The initial analysis of the draft human genome sequence (Lander *et al.*, 2001) revealed a number of surprising facts with respect to the genome composition. Possibly the most surprising fact was the retention of an overwhelming mass of repetitive sequences, in comparison to genic sequence. Initial analysis of the draft human genome sequence (Lander *et al.*, 2001) suggested that transcribed sequence corresponded to less than 5 %, with 0.5 to 1 % being protein-coding (exonic) sequence. This is in stark contrast to the approximately 50 % of the genome accounted for by repetitive elements (Lander *et al.*, 2001).

1.1.1 Classes of repetitive elements in the human genome

Five major classes of repetitive elements have been identified in the human genome (summarised below).

1.1.1.a Tandem repeat blocks

Tandem repeats consist of direct repeats of units larger than 5 nucleotides (nt), account for approximately 3 % of the genome, are generally highly localised and often found near telomeres and centromeres (Lander *et al.*, 2001). Tandem repeat blocks are often seen at regions displaying copy number variation through ectopic recombination, i.e. recombinational crossover between non-allelic but homologous DNA sequences (Feuk *et al.*, 2006; Shaw and Lupski, 2004), for example the α-globin and β-globin gene regions (Holloway *et al.*, 2006; Lam and Jeffreys, 2006). They also occur within gene families which show high rates of evolution, for example the Major Histocompatability Complex class II (MHC class II) gene family (Jeffreys *et al.*, 1998).

Minisatellites generally have GC-rich variant repeats ranging in length from 10 to > 100 bp. Minisatellites are common in the genome, and are highly variable between individuals (Tamaki and Jeffreys, 2005). Restriction fragment length polymorphism analysis of multiple minisatellite loci, by Southern blotting and hybridisation of radio-labelled probes specific for a core sequence, was used to develop the first genetic fingerprint (Tamaki and Jeffreys, 2005).

Also, within the human genome, ribosomal RNA (rRNA) arrays are usually tandem arrays, as are some transfer RNA (tRNA) genes. Both rRNA arrays and tRNA genes show dosage repetition (Kedes, 1979; Long and Dawid, 1980), where numerous repeated copies of a gene are contained within the genome to allow the generation of a large amount of the particular product in an appropriate amount of time (Rubin *et al.*, 1976) in the absence of translational amplification.

1.1.1.b Microsatellites

Microsatellites are simple sequence tandem repeats with a repeat unit of 1 to 5 nt, typically repeated 5 – 30 times that are dispersed throughout the genome (Lander *et al.*, 2001; Tamaki and Jeffreys, 2005). Microsatellites are common in the genome and highly variable between individuals. As a result, multiplex PCR across ten tetranucleotide microsatellite loci, along with the amelogenin sex test, is currently used in the UK to generate genetic profiles in forensic analysis (Jeffreys et al., 1985; Tamaki and Jeffreys, 2005).

1.1.1.c Transposon-derived repeats

Transposon-derived repeats are sequences derived from mobile genetic elements, which show a dispersed pattern throughout the genome. Transposon-derived repeats are covered in more detail on page 5.

1.1.1.a Pseudogenes

Processed pseudogenes are inactive copies of cellular RNA species which have been dispersed throughout the genome, most likely by retrotransposition (Lander *et al.*, 2001). Processed pseudogenes lack promoters due to mRNA processing prior to retrotransposition. Pseudogenes are (usually) inactivate copies of genes generated by duplication of genomic regions, for example through segmental duplication formation (Lander *et al.*, 2001). Pseudogenes are usually incomplete, re-arranged or have been otherwise inactivated.

1.1.1.b Segmental duplications

Segmental duplications are blocks of duplicated genomic sequence of ≥ 1 kb, showing ≥ 90 % sequence identity, which appear to have been copied from one genomic location to another (Lander *et al.*, 2001; Zhang *et al.*, 2005). 4 % - 5 % of the genome is contained within segmental duplications, with the extent of duplication varying from 1 % to 14 % among the 24 chromosomes. Intrachromosomal duplications are larger and more frequent than interchromosomal duplications, and duplications tend to be enriched in pericentromeric and subtelomeric regions (3 to 4 times the genome average) (Zhang *et al.*, 2005).

1.1.2 Is repetitive DNA just genetic "junk"?

Despite their overwhelming abundance, except in the case of ribosomal and tRNA gene clusters, repetitive elements are often dismissed as non-functional and uninteresting "Junk" DNA (Lander *et al.*, 2001).

The most highly represented class of repetitive element in the human genome are transposon-derived repeats. Transposable elements account for > 50 % of the human genome sequence. This is likely to be an under estimate as aligned sequences with a sequence identity of less than 80 % are often considered too degenerate to be reliably recognised (Lander *et al.*, 2001). Of these elements, Short Interspersed Nuclear Elements (SINEs) account for 13 % of the genome sequence; Long Interspersed Nuclear Elements (LINEs) 20 %; Long Terminal Repeat (LTR) retrotransposons 8 %; and DNA transposons 3 % (Lander *et al.*, 2001).

As opposed to the initial view that transposable elements are merely junk DNA, it is now known that they have had an important impact in shaping our genome, for example as expressed by Wheelan *et al.*, in their paper 'Gene-breaking: a new paradigm for human retrotransposon-mediated gene evolution' (Wheelan *et al.*, 2005), "transposable elements are neither Junk DNA nor mere curiosities to be categorised taxonomically and relegated to dusty catalogs; rather they can affect gene expression in important ways and are a dynamic and significant part of our evolutionary history. "

1.1.3 Transposable elements and transposon-derived repeats

It is clear that transposable elements play a significant role in genome fluidity. The human genome apparently has no active mechanism to remove transposable elements resulting in the accumulation of transposable elements to > 50 % of the human genome sequence (Lander *et al.*, 2001). This accumulation has occurred even though the most active known human transposable element (L1: page 13) is relatively inactive when compared to the equivalent element in the mouse genome, where ~8 % of spontaneous mutations are attributed to retrotransposition (Ostertag and Kazazian, 2001a). This has resulted in very few transposon-derived spontaneous mutations in humans (Kazazian and Moran, 1998).

In contrast, more than 50 families of transposable elements comprise approximately 20 % of the *Drosophila melanogaster* genome. The activity of these elements causes more than 50 % of the spontaneous mutations observed in *D. melanogaster* (Charlesworth and Langley, 1989). Despite the frequent activity of *D. melanogaster* transposons, the accumulation of elements is restricted by deletion mechanisms which gradually remove insertions from the genome (Charlesworth and Langley, 1989; Petrov and Hartl, 1997; Petrov *et al.*, 1996).

Despite having very different levels of activity, numerous classes and families of transposable elements are found within the genomes of many taxonomic groups, from bacteria to complex higher eukaryotes such as mammals, and so transposable elements are ubiquitous (i.e. have been identified in all studied taxa). Transposable elements can be divided into two major classes, class I (DNA transposons) and class II (Retrotransposons).

1.1.3.a DNA transposons (Class I transposable elements)

In general DNA transposons consist of the coding sequence of a transposase enzyme flanked by inverted terminal repeats. The transposase enzyme binds near the inverted repeat termini and performs phosphodiester bond hydrolysis, excising the transposon, and leaving exposed 3′ hydroxyl groups (OH). The exposed 3′ OH group allows insertion into a new site, prior to gap filling by endogenous DNA repair proteins. This process leads to direct terminal repeats of the target site being formed known as Target Site Duplications (TSDs) (Moran and Gilbert, 2002). The DNA transposition machinery acts in *trans* and mobilises any element with transposase recognition signals and is not specific for the encoding element. Competitive

parasitism of the active transposon machinery by inactive elements has likely led to the extinction of active DNA transposons in the human genome (Lander *et al.*, 2001).

The existence of transposable elements was first suggested in the late 1940s, before the discovery of the structure of DNA (Watson and Crick, 1953). Barbara McClintock described the effects of the dissociation locus (*Ds*) discovered on the short arm of chromosome 9 in *Zea mays* (McClintock, 1950; McClintock, 1951; McClintock, 1956). Dissociation activity resulted in chromosomal breaking, and McClintock also described the effects of the insertion of *Ds* locus chromatin elements on the expression of genes adjacent to their insertion points (McClintock, 1950). Transposition of *Ds* was found to be dependent on the presence of activator (*Ac*), another element capable of transposition, often between different chromosomes (McClintock, 1950). Although it was unknown at the time, the elements described by McClintock were DNA transposons.

DNA transposons are prevalent in bacteria and are referred to as insertion sequences (IS) (Hirsch *et al.*, 1972; Jordan *et al.*, 1968; Saedler *et al.*, 1972; Shapiro, 1969). They replicate via a "cut and paste" method involving the excision of transposon DNA from one genomic location and its re-integration into another, using the proteins encoded by the IS. They have very limited target sequence preference, which can comprise a small number of nucleotides, (e.g. TA dinucleotides) allowing them to integrate into many genomic locations.

The P elements of *D. melanogaster* are one of the best studied of eukaryotic transposable elements. Discovered in the late 1960's germline transposition of P elements leads to a syndrome of traits termed hybrid dysgenesis (Hiraizumi, 1971). The symptoms of hybrid dysgenesis are a collection of traits including sterility, mutation induction, male recombination (normally repressed in *D. melanogaster*), and chromosomal abnormalities (Engels and Preston, 1979; Kidwell *et al.*, 1973; Kidwell *et al.*, 1977; Kidwell and Novy, 1979). All these dysgenic traits are restricted to the germline. Interestingly, P element containing populations (P strains) are generally found in the wild while many laboratory strains, collected around the turn of the previous century, (M strains) do not contain any P elements.

Being active in the germline, the P elements of *D. melanogaster* are passed to subsequent generations, and have spread into all wild *Drosophila* populations in less than a century. They therefore have the potential to be of evolutionary consequence. By contrast, DNA transposons in humans are extinct, and so are of no further evolutionary consequence. 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.

1.1.3.b Retrotransposons (Class II transposable elements)

Retrotransposition involves a "copy and paste" method of transposition by which the parental element is copied by transcription, the element RNA is processed internally within the cell, and then reverse transcribed into a cDNA copy that is re-integrated into a new genomic location. Retrotransposition is therefore an intrinsically replicative process in which new element copies are created. The reverse transcriptase enzyme (RT), which converts element-encoded mRNA into a cDNA copy, was first discovered in 1970 as a retroviral-encoded enzyme which catalysed DNA replication from an RNA template (Baltimore, 1970; Temin and Mizutani, 1970). Since its discovery, numerous RT-containing elements have been discovered in plants, animals and fungi. Although these elements differ structurally, the amino acid sequence similarity in the RT-encoding domains of these elements suggests a common evolutionary origin (Xiong and Eickbush, 1990).

1.1.3.b.i LTR retrotransposons

Long terminal repeats (LTRs) define the LTR class of retrotransposons, which are related to endogenous retroviruses (Malik *et al.*, 2000). The LTRs flank the transposon at the 5′ and 3′ ends and both have strong promoter activity. The coding sequence of LTR retroelements has a very similar structure to retroviruses, but LTR retroelements harbour an inactive *gag* gene, or lack a recognisable *gag* gene, which encodes the viral particle coat (GAG) of retroviruses. This prevents LTR retroelements from leaving the host cell, making them strictly intracellular parasites. LTR elements also carry a *pol* gene which codes for a reverse transcriptase (RT), and a gene for the Ribonuclease H enzyme.

The Ribonuclease H enzyme activity hydrolyses the RNA in the cDNA/RNA duplex after first-strand synthesis, allowing second-strand synthesis of the cDNA. Finally an element-encoded integrase enzyme is required for integration of the new double-stranded cDNA into the genome (Kazazian, 2004). Reverse transcription occurs within a virus-like particle (VLP) in the

cytoplasm prior to integration via a complex multi-step process (Kazazian, 2004). Many LTR retrotransposons have specific target sites, and often require cellular co-factors to aid their integration. For example the Ty3 elements of *Saccharomyces cerevisiae* require transcription factors TFIIIB and TFIIIB to specifically integrate in close proximity to an RNA polymerase III transcription site (Chalker and Sandmeyer, 1992), and Ty5 elements require Sir4 to target telomeric heterochromatin (Devine and Boeke, 1996).

1.1.3.b.ii Non-LTR retrotransposons

As their name suggests, Non-LTR retrotransposons do not have flanking LTRs. Autonomous retrotransposons encode the enzymes required for their mobilisation, and families of non-autonomous elements parasitise the retrotransposition machinery of the autonomous elements. The only autonomous retrotransposon apparently active in the human genome, Long Interspersed Nuclear Element-1 (LINE-1 or L1), is a Non-LTR retrotransposon. L1 elements are mammalian-specific retrotransposons, but related LINE-like elements can be identified in organisms as distant as *Neurospora crassa* (the Tad element) (Xiong and Eickbush, 1990).

L1 evolved approx 170 million years ago (MYA) at approximately the time of the marsupial/placental mammal divergence (Boissinot et al., 2000b). With approximately 500,000 copies per human haploid genome, and encompassing approximately 17 % of human genomic DNA, they are the most prominent transposable element in humans, as well as in many other mammals (Lander *et al.*, 2001; Moran and Gilbert, 2002). It has also been asserted that, by mass, L1 is the most influential single entity affecting the structure of the human genome (Boissinot et al., 2000b; Moran and Gilbert, 2002). Although the human genome contains many L1 elements, 99.9 % are not able to retrotranspose (Moran *et al.*, 1996), i.e. are not retrotransposition competent (RC). This is due to 5′ truncation or internal rearrangements that occur during retrotransposition. Also, once inserted, subsequent mutation of the functional regions of the element can also lead to inactivation (Boissinot et al., 2000b; Moran and Gilbert, 2002). There are at least 90 full-length human L1s in the human genome which have intact open reading frames (ORFs), and are potentially RC (Brouha *et al.*, 2003). Of the 90, the majority have been shown to be weakly active in cell culture, but six have been shown to be highly active (Brouha *et al.*, 2003). The details of the retrotransposition cycle of L1 remain poorly defined at

best, and very little is known of the overall dynamics of L1 retrotransposition in the human population.

1.1.3.c Vertical inheritance and evolution of non-LTR retrotransposons

Diverse families of non-LTR retroelements can be identified in the vast majority of eukaryotic organisms. Phylogenetic analysis of the RT domain (conserved between all classes of non-LTR retrotransposon) suggests that non-LTR retrotransposons originated over 600 million years ago (MYA) (Malik *et al.*, 1999).

Non-LTR elements sharing the same structural features can be grouped together into clades, each clade being named after one of the best characterised elements within the clade (Malik *et al.*, 1999). In 1999, phylogenetic analysis of the RT domain had identified eleven distinct clades of non-LTR retrotransposon, CRE, R2, R4, L1, RTE, Tad1, R1, LOA, Jockey, CR1 and I. By June 2000, this had been expanded to fourteen clades (Eickbush and Malik, 2002) by addition of the NeSL1 (Malik and Eickbush, 2000) and Rex1 (Volff *et al.*, 2000), and the division of the I clade into I and Ingi. Four of the clades (L1, RTE, CR1 and I) show wide distribution among eukaryotes, while the others are limited to one or two taxonomic groups. For example four of the remaining clades (R2, LOA, R1 and Jockey) are arthropod-specific. Figure 1.1 shows a phylogenetic tree of non-LTR retroelements, and a grouping of the clades based on structure (summarised in figure 1.1 b).

Phylogenetic analysis of the non-LTR retrotransposons has been simplified as, for the most part, they are restricted to vertical transmission (Malik *et al.*, 1999). Vertical transmission follows the identity by descent (IBD) model of inheritance in which elements are passed through the generations in a lineage-specific manner, via sexual reproduction. Horizontal transmission, however, allows mobile elements to be transmitted between different species independent of lineage and reproduction. As both DNA transposons and LTR retrotransposons generate extrachromosomal DNA elements, prior to re-integration, they have the capacity to be moved between species. Non-LTR retrotransposons by contrast do not form extrachromosomal DNA species as reverse transcription is carried out at the site of integration into the genome, using genomic DNA as a primer for cDNA synthesis (Malik *et al.*, 1999). This is known as target-primed reverse transcription (TPRT).

The evolution of retrotransposable elements in a vertical manner is dependent upon the generation of new RC elements in the germline. Over time retrotransposons will acquire mutations that either increase their activity, or reduce/eliminate their activity. A RC L1 will generate daughter elements, through retrotransposition. Over time the parent element will be rendered inactive by the acquisition of mutations. In order to be of evolutionary significance, the daughter element must have completed retrotransposition in the germline. Somatic retrotransposition is of no value to the retrotransposon as it results in loss of the elements as they will not be passed on to subsequent generations, and may cause diseases in the host. Non-LTR elements have been shown to retrotranspose in the germline, for example the I element of *D. melanogaster* (Bucheton *et al.*, 2002; Picard and L'Heritier, 1978). I elements were the first transposable elements found in *Drosophila*, and were named I factors as they induced hybrid dysgenesis (Bucheton *et al.*, 2002; Bucheton *et al.*, 1976; Picard, 1976).

The *D. melanogaster* population can be divided into two strains, inducer (I), and reactive (R) strains (Bucheton *et al.*, 1976). I strains contain several RC I elements which are absent from R strains (Bucheton *et al.*, 1984; Picard, 1976; Picard and Pelisson, 1979). In wild type I strains, I elements are repressed and therefore do not retrotranspose at readily detectable levels, but crossing I type males with R type females leads to hybrid dysgenesis resulting in severe infertility in F_1 females (Picard, 1976). I elements are able to retrotranspose in the female offspring of crosses between I type females and R type males, but at a reduced rate seen in the reciprocal cross. The reduced rate of retrotransposition allows fertility of the F_1 females with no evidence of hybrid dysgenesis (Picard, 1976). Thus the I element's retrotransposition is sex-specific (Picard, 1976), and germline-specific (Pelisson and Bregliano, 1987).

By contrast L1 retrotransposition in the human genome does not appear to be limited to the germline; for example an insertion into the adenomatosis polyposis coli (APC) gene has been linked with the development of colorectal cancer (Miki *et al.*, 1992). In order to have been propagated this insertion must have occurred in a colonic mucosal stem cell. However, there is evidence of L1 retrotransposition in the human female germline. A *de novo* insertion of an L1 (LRE3) into exon 4 of the CYBB gene, of a male patient, was identified as the cause of chronic granulomatous disease (CGD). As the disease is X-linked, the explanation for the insertion was retrotransposition during the first division of maternal meiosis (Brouha *et al.*, 2002).

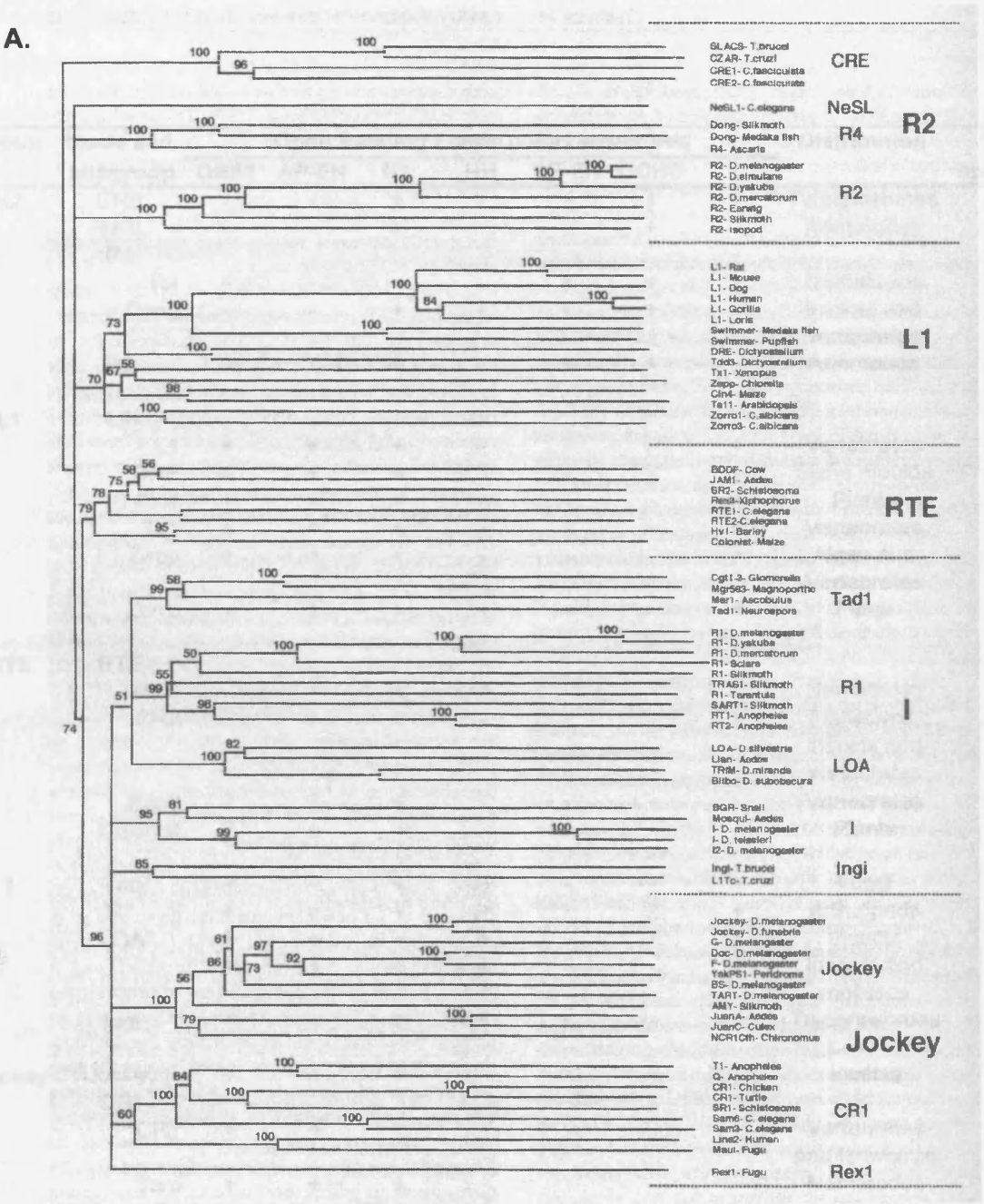


Figure 1.1: Part A, taken from Eickbush and Malik. (2002) (figure 2, pg 1115), shows a phylogenetic tree of non-LTR retrotransposons based on their reverse transcriptase domain (RT). The numbers at each node indicate bootstrap values. The name of each element and the host organism is given to the right. The elements are divided into clades, and the clades further divided into groups (larger font size divided by dotted lines). Each group contains elements with similar open reading frame (ORF) structure as outlined in table B, which was updated from Malik (1999) using additional information given in Eickbush and Malik (2002).

B.

Group	Clade and subgroup	Open Reading Frame (ORF) structures							Distribution	Site specific
		ORF1	AP-EN	RT	RH	RE-EN	CCHC	C		
R2	CRE	-	-	+	-	+	+	-	Tyrpanosomes	+
	NeSL	-	-	+	-	+	+	-	Nematodes	+
	R4							-		
	R4	-	-	+	-	+	+	-	Nematodes	+
	Dong	-	-	+	-	+	+	-	Insects and vertebrates	+
	R2	-	-	+	-	+	+	-	Arthropods	+
L1	L1									
	L1	+	+	+	-	-	-	+	Vertebrates	-/+
	DRE	+	+	+	-	-	-	+	Slime moulds	+
	Cin4	+	+	+	-	-	-	+	Plants	-
	Tx1	+	+	+	-	-	-	+	Vertebrates	+
	Zepp	+	+	+	-	-	-	+	Algae and Vertebrates	+
	Zorro	+	+	+	-	-	-	+	Fungi	-
RTE	RTE									
	RTE	-	+	+	-	-	-	-	Nematodes	-
	SR2	?	+	+	-	-	-	-	Flatworms, insects and vertebrates	-
	Rex3	?	+	+	-	-	-	-	Vertebrates	-
	Colonist	?	+	+	-	-	-	-	Plants	-
I	Tad	+	+	+	-/+	-	-	+	Fungi	-
	R1	+	+	+	-/+	-	-	+	Arthropods	+
	LOA	+	+	+	+	-	-	+	Insects	-
	I	+	+	+	+	-	-	+	Insects and molluscs	-
	Ingi	+	+	+	+	-	-	+	Trypanosomes	-
Jockey	Jockey	+	+	+	-	-	-	-	Insects	-/+
	CR1									
	CR1	+	+	+	-	-	-	-	Vertebrates and flatworms	-
	Sam	+	+	+	-	-	-	-	Nematodes	-
	T1	+	+	+	-	-	-	-	Insects	-
	L2	+	+	+	-	-	-	-	Vertebrates	-
	Rex1	+	+	+	-	-	-	-	Vertebrates	-

However, this paper failed to fully address the possibility of retrotransposition occurring prior to meiosis, for example during the formation of primordial germ cells. Five years later it has been shown that L1 does indeed retrotranspose during early embryogenesis and is not restricted to the meiotic stages of germ cell development (van den Hurk *et al.*, 2007). It has also been

confirmed that the mother of a patient with choroideremia shows both somatic, and germline, mosaicism for the disease-causing L1 insertion ($L1_{CHM}$) (van den Hurk *et al.*, 2007; van den Hurk *et al.*, 2003).

1.2 L1 structure and retrotransposition

1.2.1 L1 structure

The typical structure of a full-length human-specific RC L1 is shown in Figure 1.2. Full-length human L1s are approximately 6 kb in length. They have a 910 bp 5′ untranslated region (UTR) that harbours an internal RNA polymerase II (RNA pol II) promoter.

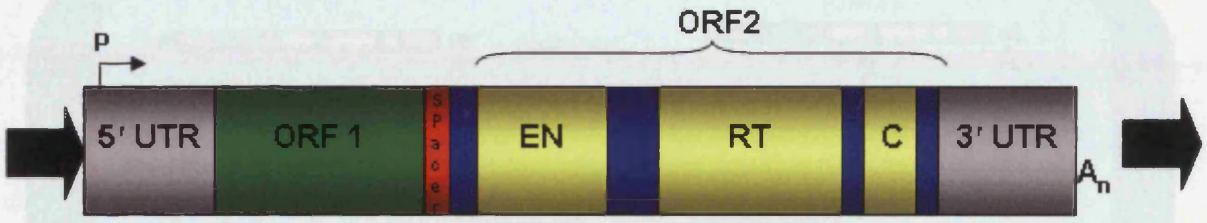


Figure 1.2: The structure of a typical full-length L1 element (not to scale). An internal RNA pol II promoter (P) is located in the 5′UTR. The Blue boxes indicate open reading frames, with separate functional units in yellow. The intergenic spacer is represented in red. The L1 poly A tail is represented (A_n), and is flanked in the genome by TSDs represented as block arrows.

The L1 retrotransposition machinery is encoded by two non-overlapping open reading frames (ORFs), ORF1 and ORF2, separated by an intergenic spacer. L1ORF1 protein (ORF1p) exhibits strong nucleic acid binding activity for both RNA and DNA, and is of essential but largely unknown function (Martin *et al.*, 2003). ORF2p contains an endonuclease domain (EN) with limited homology to apurinic/apyrimidinic (AP) endonucleases (Weichenrieder *et al.*, 2004), and contains the reverse transcriptase domain (RT) as well as a highly conserved cysteine-rich C terminal domain (C) of unknown function.

The $L1_{HS}$ 3′ UTR is 205 bp long and contains a G-rich poly-purine tract of unknown function. L1 copies in the genome contain an unconventional poly A tail which immediately

follows the CPSF1 binding site poly-adenylation signal. The element is flanked by short direct sequence repeats (7 to 20 bp) of the insertion site called target site duplications (TSDs) (Moran and Gilbert, 2002).

1.2.2 L1 Retrotransposition

The L1 retrotransposition cycle is inherently replicative since the parental L1 is copied by transcription and the transcript processed and integrated into a new genomic location. Many of the aspects of L1 retrotransposition were identified in cultured cells. Figure 1.3 outlines the retrotransposition cycle, but many of the details of the mechanisms (indicated by ?: Figure 1.3) are unknown.

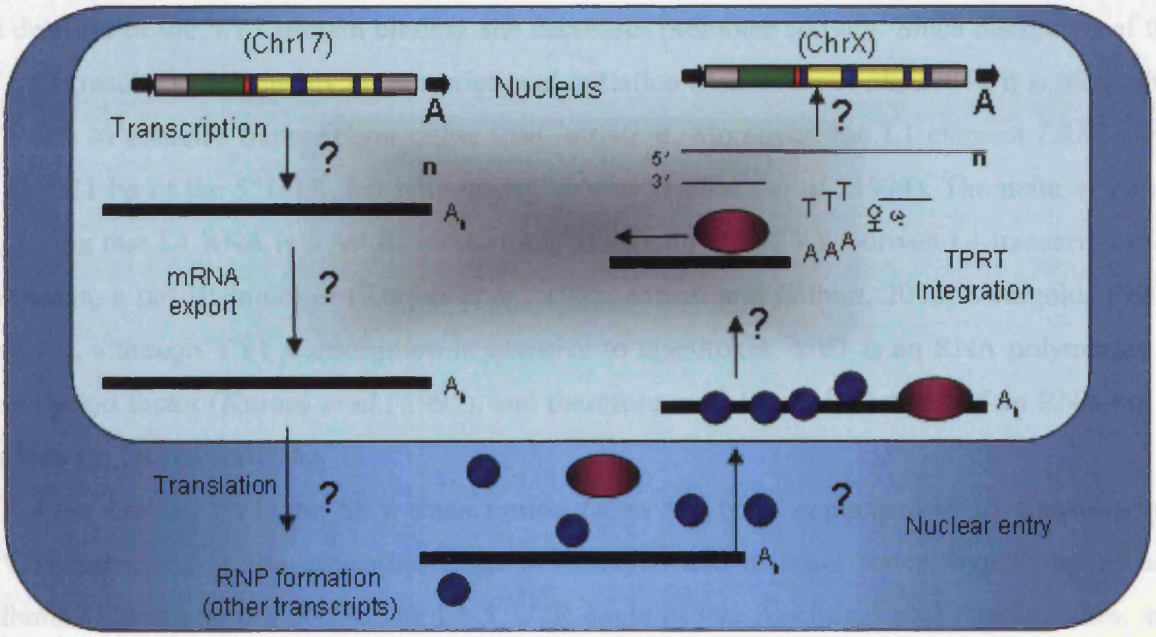


Figure 1.3: The progression of L1 through a replicative cycle from transcription to translation and re-integration. In this case the L1 has been transcribed from chromosome 17, and a *de novo* insertion into the X chromosome followed. The question marks show areas where the mechanistic details are unknown

1.2.2.a L1 transcription

To initiate retrotransposition, the parental L1 is transcribed from its internal promoter and polyadenylated. The 5′ UTR of L1 contains an internal promoter that directs transcription from the first nucleotide of the L1 sequence (Swergold, 1990). There has been much debate as to whether L1 is transcribed by RNA pol II or pol III; however, the bulk of the evidence points towards L1 RNA being a pol II transcript. For example, alpha amanitin, an RNA pol II inhibitor, can inhibit transcription of full-length sense strand L1 RNA in teratocarcinoma cells (Moran and Gilbert, 2002). Furthermore, deletion studies of sections of the 5′ UTR showed that the majority of promoter activity is located within the first 670 bp of the 5′ UTR (Moran and Gilbert, 2002; Swergold, 1990). YY1, a ubiquitously expressed RNA pol II transcription factor (Kurose *et al.*, 1995) thought to play a role in L1 transcription, binds within the first 20 bp of the L1 element, and deletion of the YY1 protein binding site decreases promoter activity. Since disruption of the YY1 site results in inaccuracy in transcriptional initiation (Athanikar *et al.*, 2004), it is likely that YY1 acts to enhance transcription rather than initiate it. Moreover, the L1 element *LRE2* lacks the first 21 bp of the 5′ UTR, but remains RC *in vivo* (Holmes *et al.*, 1994). The main evidence suggesting that L1 RNA is a pol III transcript is the sensitivity of YY1-driven L1 transcription to tagetitoxin, a pol III inhibitor (Kurose *et al.*, 1995; Moran and Gilbert, 2002; Swergold, 1990). However, although YY1 transcription is sensitive to tagetitoxin, YY1 is an RNA polymerase II transcription factor (Kurose *et al.*, 1995), and therefore most likely forms part of an RNA pol II complex for L1 transcription.

As well as YY1, the SRY transcription factor has been implicated in L1 transcription. SRY is expressed in the urogenital ridge of embryos and in adult testes, hypothalamus and midbrain (Tchenio *et al.*, 2000). The L1 5′ UTR contains two functional SRY binding sites, and binding can drive transcription of an L1 promoter fused to a *luciferase* reporter gene in human cultured embryonic carcinoma and testicular tumour cells (Tchenio *et al.*, 2000). Point mutations at these sites, that abolish SOX protein binding, prevent transcription. In addition ectopic expression of SOX11 induces retrotransposition of endogenous Human-specific L1s ($L1_{HS}$) (Tchenio *et al.*, 2000). It is therefore most likely that L1 is controlled by multiple *cis*-acting sequences which recruit RNA pol II transcription factors, such as YY1 and SOX, which act together to recruit endogenous transcriptional machinery (Moran and Gilbert, 2002; Tchenio *et al.*, 2000).

1.2.2.b Interaction between L1 and the spliceosome

The L1 transcript is exported from the nucleus, by an unknown mechanism, prior to translation. Although the L1 transcript itself does not get spliced during its export from the nucleus, there is some evidence that it does interact with the spliceosome.

The human genome contains numerous chimeric products proposed to have arisen from template switching, a process which is believed to have been occurring for at least the last 47 million years (MY) (Buzdin *et al.*, 2002). Chimeric RNA products of various cellular transcripts have subsequently been integrated into the genome by L1 reverse transcription (Buzdin *et al.*, 2003; Buzdin *et al.*, 2002). This phenomenon is not just specific to human L1s as it has also been observed in an R2 Non-LTR retrotransposon (Bibillo and Eickbush, 2002; Bibillo and Eickbush, 2004).

There is a significant correlation between the presence of chimeric transcripts and RNA species involved in the spliceosome pathway. 93 % of the 5′ parts of the chimeras connected to 3′ L1 sequences were copies of snRNAs U6, U3 and U5 involved in spliceosomes (82, 10 and 1 % respectively) (Buzdin *et al.*, 2003). It is possible that the high frequency of switching to spliceosomal templates might imply an association between the L1 retrotransposition machinery and spliceosomal RNAs. This association may involve specific co-localisation of the L1 transcription/integration complexes and the spliceosome (Buzdin *et al.*, 2003). Template switching has also been proposed as a model for generating new L1 families. One piece of evidence is that mammalian L1s have different promoter types (Furano, 2000), which may have resulted from L1 template switching (Buzdin *et al.*, 2003).

1.2.2.c Translation of the L1 retrotransposition machinery

The L1 transcript differs from standard mRNA species as it consists of two ORFs on a single transcript. L1 ORF1 (1014 bp) and ORF2 (3825 bp) are in the same reading frame separated by a 63 bp intergenic spacer containing two conserved in-frame translational stop codons (Leibold *et al.*, 1990; McMillan and Singer, 1993). The in-frame stop codons suggest that translation of the ORF2 protein (ORF2p) involves a complex mechanism.

It has been predicted that as few as one ORF2p molecule is generated per L1 RNA transcript (Moran and Gilbert, 2002; Wei *et al.*, 2001). This is supported by the inherent difficulty in detecting ORF2p using immunoassays, whereas ORF1p has been readily identified by use of antibodies, both *in vivo* and *in vitro* (Bratthauer and Fanning, 1992; Bratthauer and Fanning, 1993; Goodier *et al.*, 2004; Hohjoh and Singer, 1996; Holmes *et al.*, 1992; Leibold *et al.*, 1990; McMillan and Singer, 1993). ORF2p has since been detected by immunoprecipitation, and, as previously predicted, is present at very low levels (Ergun *et al.*, 2004).

The translation of L1 ORF1 is thought to conform to the scanning model for translation (Kozak, 1989). This involves recognition and binding of a cap structure (m7GpppN) by translation initiation factor eIF4F. The scanning model for translation initiation of L1 has, however, been disputed (Li *et al.*, 2006). Work carried out on mouse L1s suggests that Internal Ribosome Entry Sites (IRES) allow ribosomes to be recruited upstream of both ORF1 and ORF2, thus avoiding scanning of the highly structured 5′ UTR. However, an investigation which dispensed with the IRES regions outlined, by Li *et al*, still showed efficient retrotransposition of a tagged human L1 (Alisch *et al.*, 2006). Most interestingly the intergenic spacer could be replaced with an in-frame stop codon (ORF1-stop-ORF2), and still allow 24 % retrotransposition efficiency (compared to wild type L1.3). This casts doubt on the conclusion that an IRES necessary for ORF2 translation is contained within the intergenic spacer.

Another recent paper has cast further doubt on the theory that L1 is translated by IRES-mediated initiation (Dmitriev *et al.*, 2007). This paper supported the findings of Alisch *et al.*, suggesting that translation of L1 ORF1p is cap-dependent rather than IRES-dependent. This lends support to Alisch's theory that a novel complex mechanism of ribosomal retention allows ORF2p translation (Dmitriev *et al.*, 2007).

1.2.2.d The L1 retrotransposition machinery

L1 ORF1p is a 338 amino acid protein (40 kDa) which has been isolated both *in vivo* and *in vitro* as a multimeric 200 kDa protein complex bound to L1 RNA, and can be localised in the cytoplasm in *vivo* (Moran *et al.*, 1996). This ribonucleoprotein (RNP) complex has been suggested to be an intermediate step in the L1 retrotransposition cycle.

1.2.2.d.i ORF1p

Much of the structural analysis of ORF1p has been carried out on mouse ORF1p, although the exact relevance of the mouse L1 function in comparison to humans is not fully understood. The study of mouse ORF1p should however give some insight into the function and structure of human L1. Mouse ORF1p forms a homotrimer through a coiled coil domain located within the N-terminal half of each monomer (Martin *et al.*, 2003). The protein has a strong nucleic acid binding capacity through a basic domain, located C terminally. It is thought that ORF1p can act as a chaperone by directing the rearrangement of nucleic acids into their most stable conformation (Martin *et al.*, 2005). During TPRT, L1 ORF1p has a potential role in melting the target DNA and displacing the strands, followed by holding the L1 RNA in place to aid priming of reverse transcription after nicking of the DNA by EN (Martin *et al.*, 2005). This would require the ability to bind three nucleotide strands, the two strands of the target DNA duplex and also the L1 RNA. Being a homotrimer, it is hypothesised that ORF1p could fulfil this role (Martin *et al.*, 2003). Human ORF1p also has strong nucleic acid binding ability, showing strong preference for its encoding RNA (Hohjoh and Singer, 1997).

1.2.2.d.ii ORF2p

The L1 endonuclease (EN) domain is located in the amino terminus of the multi functional ORF2p. EN function has been shown to reside between positions 1 to 239 of ORF2p (Feng *et al.*, 1996). L1 EN has limited sequence similarity to apurinic/apyrimidinic (AP) endonucleases such as the APE1 DNA repair protein (Weichenrieder *et al.*, 2004); although L1 EN is distinct from AP endonucleases.

The RT domain of L1 ORF2p is a reverse transcriptase containing many of the functional elements found in other reverse transcriptases such as telomerase and retroviral reverse transcriptases (Dhellin *et al.*, 1997; Xiong and Eickbush, 1990). Seven domains of sequence similarity are shared between LTR retrotransposon-encoded RTs and the L1-encoded RT (Xiong and Eickbush, 1990). Also the RT domain can function as a reverse transcriptase in the absence of EN function (Dhellin *et al.*, 1997; Dombroski *et al.*, 1994; Mathias *et al.*, 1991; Moran *et al.*, 1996), but retrotransposition is less efficient. Unlike retroviral RT, the L1 RT domain can reverse transcribe cellular mRNAs including sequences with no homology to L1 transcripts *in vitro*

(Brosius and Tiedge, 1995; Buzdin *et al.*, 2003; Buzdin *et al.*, 2002; Goncalves *et al.*, 2000; Pavlicek *et al.*, 2006; Pavlicek *et al.*, 2002a; Pavlicek *et al.*, 2002b; Zhang *et al.*, 2003). Retroviral protein complexes require specific tRNA primers for reverse transcription to occur (Temin and Baltimore, 1972), whereas the L1 machinery can prime from and retrotranspose almost any cellular RNA (Dhellin *et al.*, 1997).

1.2.2.e Import of the L1 RNP into the nucleus

After translation, L1 ORF1p and ORF2p assemble on their encoding L1 RNA in the cytoplasm, to form a RNP complex. This strong *cis* preference of the proteins for their encoding RNA ensures that the coding RNA completes the retrotransposition cycle rather than a defective L1 RNA or cellular RNA. *Trans* mobilisation has been observed, but only rarely (Goodier *et al.*, 2004; Lander *et al.*, 2001; Moran and Gilbert, 2002). The L1 RNP is subsequently transported to the nucleus. A chaperoning role for ORF1p at this stage has been suggested, involving the nucleotide binding capabilities of the protein (Martin *et al.*, 2003).

The L1RNP is able to access the nucleus even in G1/S arrested cells (Kubo *et al.*, 2006). However, passive diffusion across the nuclear membrane seems unlikely due to the size of the RNP (Ergun *et al.,* 2004; Kubo *et al.,* 2006; Kulpa and Moran, 2006; Moran and Gilbert, 2002; Wei *et al.,* 2001). This suggests an active uptake of the L1 RNP reminiscent of the Tad Non-LTR retrotransposon of *Neurospora crassa* (Kinsey, 1993). The Tad element undergoes nuclear entry even though meiotic nuclear membrane degradation does not occur in filamentous fungi, strongly suggesting the existence of an active uptake mechanism, but the details are unknown.

A potential role for the nucleolus in nuclear import of L1 RNP has been suggested based on localisation studies of L1 proteins (Goodier *et al.*, 2004). ORF2p has two separate sub-cellular nucleolar localisation signals, and the nucleolus is a strong candidate for the localisation of cellular RNAs, ribosomal assembly and the L1 machinery. After nuclear entry, nicking of genomic DNA by the EN domain of ORF2p exposes a 3′ OH. The 3′ OH primes reverse transcription, from the 3′ end of the L1 transcript, by the ORF2p RT domain (Wei *et al.*, 2001). The reverse transcribed element is also simultaneously integrated and the second-strand synthesised in *situ*. This process is referred to as target-primed reverse transcription (TPRT) (Bibillo and Eickbush, 2002; Luan *et al.*, 1993). ORF2p can initiate this process in *vitro* but other

element-encoded activities are likely to be required, for example nucleic acid binding (Cost *et al.*, 2002) to complete the process.

1.2.2.f Target primed reverse transcription

TPRT was first identified in the R2Bm element of *Bombyx mori* (Luan *et al.*, 1993). Although both L1 and R2Bm are non-LTR poly A transposons, they share little structural similarity (Malik *et al.*, 1999). For example, L1 has an apurinic/apyrimidinic (AP) EN domain with a loose recognition site while R2Bm has a CCHC motif which targets a specific sequence in insect rDNA (Yang *et al.*, 1999). R2Bm also lacks an ORF1 protein (Cost *et al.*, 2002).

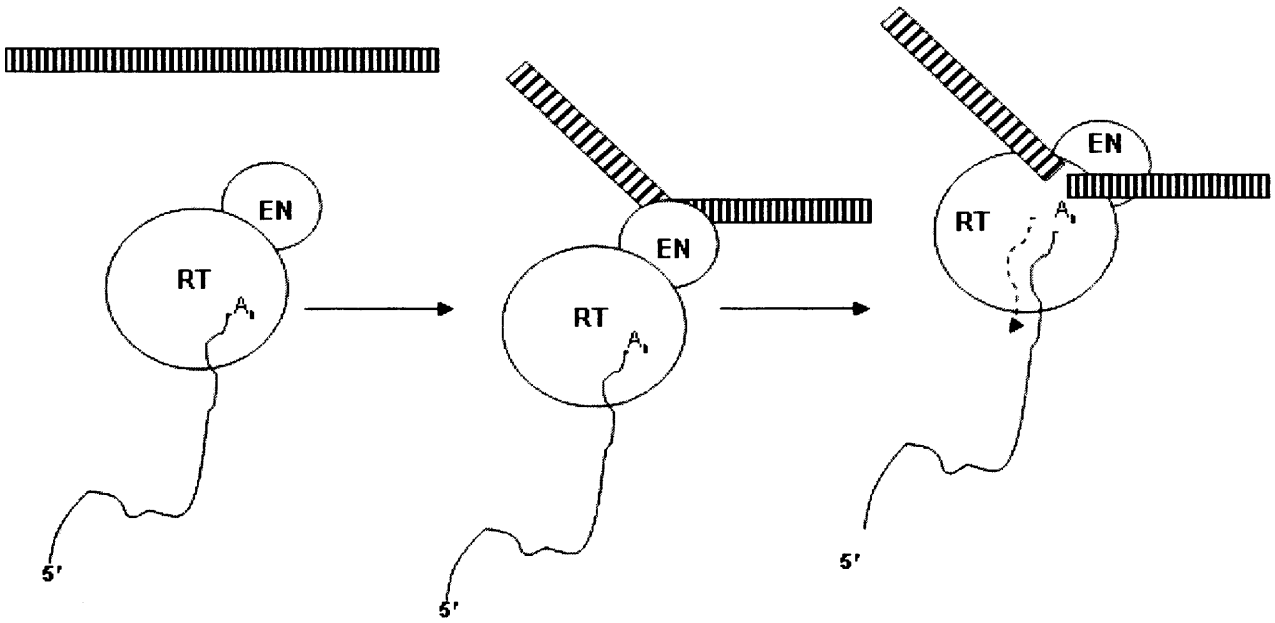


Figure 1.4: The L1 TPRT model, based on figure 8 in Cost *et al* (2002) (pg 5907). Following nicking of the target DNA by the L1 ORF2p EN domain, a 3′ OH is left exposed. The L1 RT domain extends a cDNA from this 3′ OH using the L1 transcript as a template

Despite these differences, the basic mechanism of retrotransposition initiation is conserved between ORF2p and R2Bm (Cost *et al.*, 2002). Incubation of purified ORF2p with a suitable DNA leads to the generation of branched molecules detectable by PCR. Figure 1.4 indicates how ORF2p nicks the target DNA and uses the 3′ OH overhang to prime cDNA

synthesis. In cell culture it has been demonstrated that the L1 machinery can retrotranspose short interspearsed nuclear elements (SINEs for example *Alu*), and cellular transcripts (resulting in processed pseudogenes) (Dewannieux *et al.*, 2003). This shows that L1 can initiate RT on almost any poly A species, but L1 poly A tails and 3′ UTRs most likely contain internal *cis* elements that are specifically recognised by L1ORF2p (Cost *et al.*, 2002).

1.2.2.g Targeting of TPRT

Initial analysis of the human genome showed a roughly four-fold enrichment for L1s within AT-rich DNA (Lander *et al.*, 2001). This trend suggests that L1 is less likely to insert into genes, which are GC- rich, and may be assumed to lower the mutational burden of L1 (Boissinot *et al.*, 2001). There have been reports of two insertions into alpha satellite DNA, one at the centromere of chromosome 3 which is comprised of tightly packaged constitutive heterochromatin, and also an insertion into the centromere of the Y chromosome (Santos *et al.*, 2000). It is difficult to estimate the numbers of L1s which have inserted into these sequences due to the difficulty in assembling the sequences of alpha satellite DNA (Lander *et al.*, 2001). Also a modest bias has been noted for insertion of L1s into pre-existing repetitive elements (Symer *et al.*, 2002; Szak *et al.*, 2002). Although these trends have been noted, the targeting of pre-existing elements is not significantly high when compared to the genome composition (Lander *et al.*, 2001).

It has been suggested that the EN domain targets L1 to AT-rich regions of the genome resulting in the trend of L1 distribution. However this trend is only true of old L1s (Feng *et al.*, 1996; Moran, 1999). 5 to 10 % of new insertions characterised in a cultured cell retrotransposition assay inserted into the introns of actively transcribed genes suggesting that there is no active mechanism preventing L1 from inserting into expressed sequences (Moran, 1999). However, it is unknown how closely the insertion dynamics of cultured cell retrotransposition systems resemble germline insertion and L1 distribution in the genome. Further investigations carried out by Symer *et al.* (2002) reported as many as 50 % of recovered insertions inserted into predicted genes, 62 % of which inserted in the reverse orientation relative to gene transcription. The average GC content of *de novo* insertion sites, identified in cultured cell assays, is 40.6 % +/- 5.4 % (Symer *et al.*, 2002). This is comparable to the average GC

content of the genome (Lander *et al.*, 2001; Symer *et al.*, 2002). In contrast, computational analysis of older L1s investigated by Szak *et al.* (2002) showed an average of 35 % GC content at their insertion sites (Szak *et al.*, 2002). Overall the patterns of insertion suggest a random distribution rather than targeting to AT-rich (GC-poor) genomic locations. The L1 EN consensus cleavage sequence has been relaxed to $(Pyrimidine)_4$ / purine where the / is the EN cleavage site (Feng *et al.*, 1996; Moran, 1999). There are very few genomic locations which do not contain this degenerate sequence, meaning that L1 can potentially insert anywhere in the genome.

1.3 Functional L1s in the human genome

1.3.1 The Ta L1 subfamily

16 out of 17 known insertions implicated in causing human disease are members of the same L1 sub-family, known as Ta. The Ta sub-family of L1 is therefore likely to be the most active autonomous retrotransposon family in the human genome. Analysis of genomic sequence data has concluded that the vast majority of L1s are no longer RC, and that almost all full-length L1s with intact ORFs are members of the Ta sub-family. Being the most active L1 sub-family, it also stands to reason that Ta is also phylogenetically the youngest member of the L1 family (Lander *et al.*, 2001) (Figure 1.6). The Ta subfamily emerged approximately 4 MY after the human and chimpanzee divergence (approximately 5 MYA) and is therefore human-specific (Boissinot et al., 2000b). The human-specific L1 ($L1_{HS}$) sub-family contains the Pre-Ta as well as Ta sub-family of L1s (Lander *et al.*, 2001). Ta elements were identified as a subset of the most abundant L1 full-length transcripts isolated from teratocarcinoma cells and named transcribed subset a, hence Ta (Skowronski *et al.*, 1988). All active L1s are thought to be either from the Pre-Ta or Ta sub-family as together these families account for all the insertions implicated in human diseases (Boissinot et al., 2000b; Kimberland et al., 1999; Lander et al., 2001).

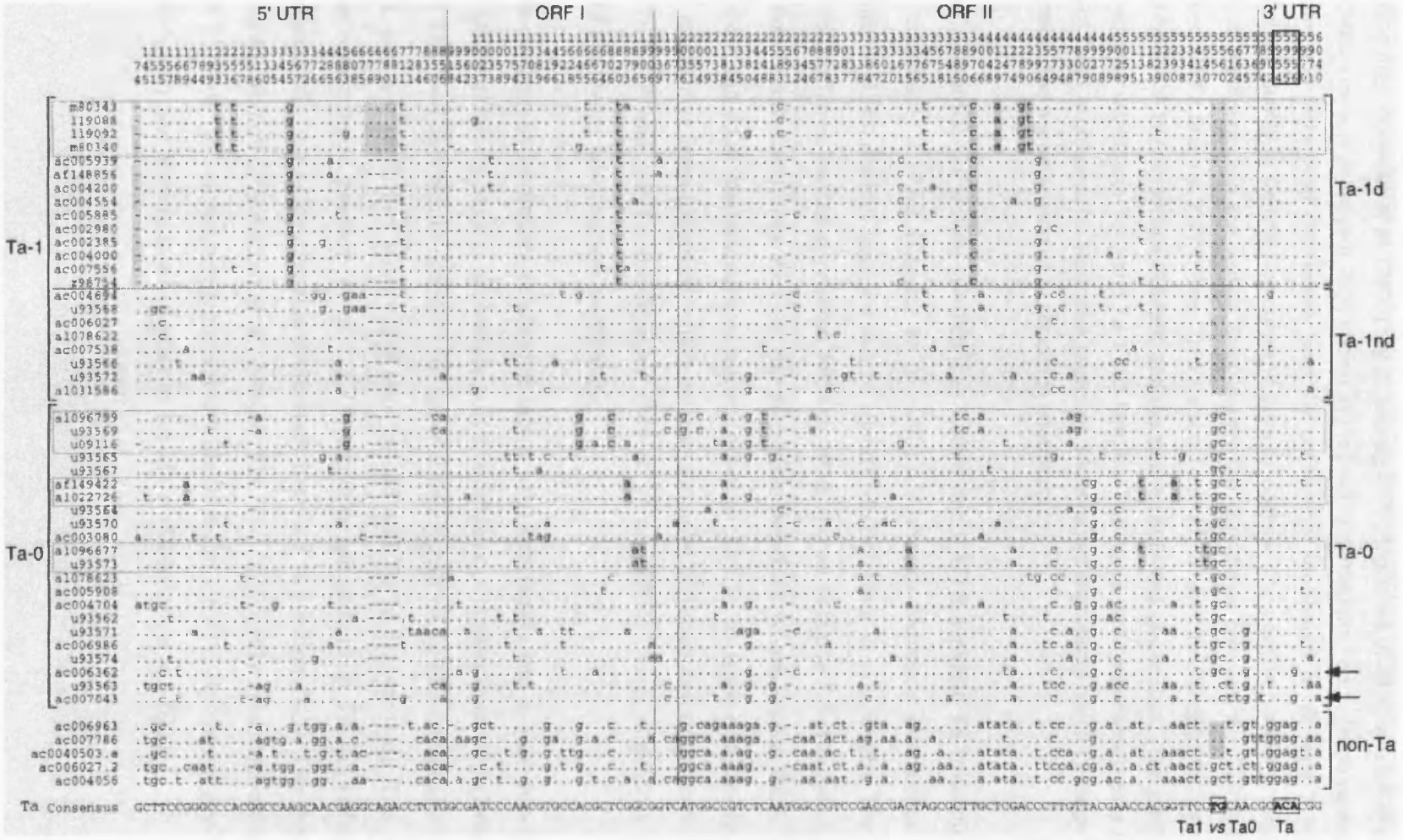

5' UTR
ORF I
ORF II
3' UTR
Ta-1
Ta-0
Ta-1d
Ta-1nd
non-Ta
Ta Consensus
Ta1 vs Ta0
Ta

Figure 1.5: Taken from Boissinot *et al.* (2000) (fig 2, and pg 917). Alignment of all known full-length Ta L1s from GenEMBL available in the year 2000 shown aligned with a consensus Ta L1. The consensus Ta is at the bottom, and highlighted differences from the consensus are shared by at least 3 different elements. Sequence identity is identified by dots and gaps by dashes. The position number is annotated across the top of the figure, and the vertical lines identify the boundaries between the 5'UTR, ORF1, Intergenic spacer, ORF2 and the 3'UTR. The "Ta-1 versus Ta-0" (T, G) and the Ta-defining ACA in the 3' UTR are highlighted in the consensus element, and the location indicated above (box). The grey boxes define distinct subsets within both Ta-0 and Ta-1 sub-families. The figure has been chosen as a simple representation of Ta L1 alignments for illustrative purposes. More recent alignments contain a much larger number of elements and a greater amount of sequence information.

Twelve diagnostic nucleotide substitutions within the 3′ UTR separate Ta from the next youngest L1 sub-family L1PA2 (Lander *et al.*, 2001). Although sequence similarities exist between L1 sub-families, distinct lineages can be identified (Boissinot et al., 2000b). The main Ta-defining substitution is located at position 5954 to 5956, shown in figure 1.5 (Boissinot et al., 2000b). The Pre-Ta sub-family has ACG in the respective positions where Ta has ACA in the 3′ UTR. However, the ACA characteristic is not restricted to Ta L1s, as some non-Ta elements with an ACA motif have been identified which harbour the most common defining substitutions of ancestral L1PA10 and pre-Ta sub-families (Myers *et al.*, 2002). Such sequences make alignments and phylogeny construction problematic.

The Ta family can be further subdivided into a Ta-0 sub-family and a Ta-1 sub-family (Figure 1.6), differentiated by several nucleotide polymorphisms outlined in figure 1.5 (Boissinot et al., 2000b). The Ta-1 sub-family accounts for approximately half of Ta sub-family (Boissinot et al., 2000b), and can be further subdivided into Ta-1nd and Ta-1d (Figure 1.6). The Ta-1d sub-family has a deletion at position 74 of 5′ UTR which is not present in Ta-1nd (where nd stands for no deletion). Ta-1d elements also exclusively have a T at position 1820. Two thirds of the Ta-1 sub-family is Ta1-d (Boissinot et al., 2000b). The Ta-1 sub-family arose approximately 2.5 MYA with 75 % of Ta-1 copies being generated within the last 1.6 MY. By comparison 80 % of the Ta-0 sub-family had already inserted prior to 1.6 MYA (Boissinot et al., 2000b).

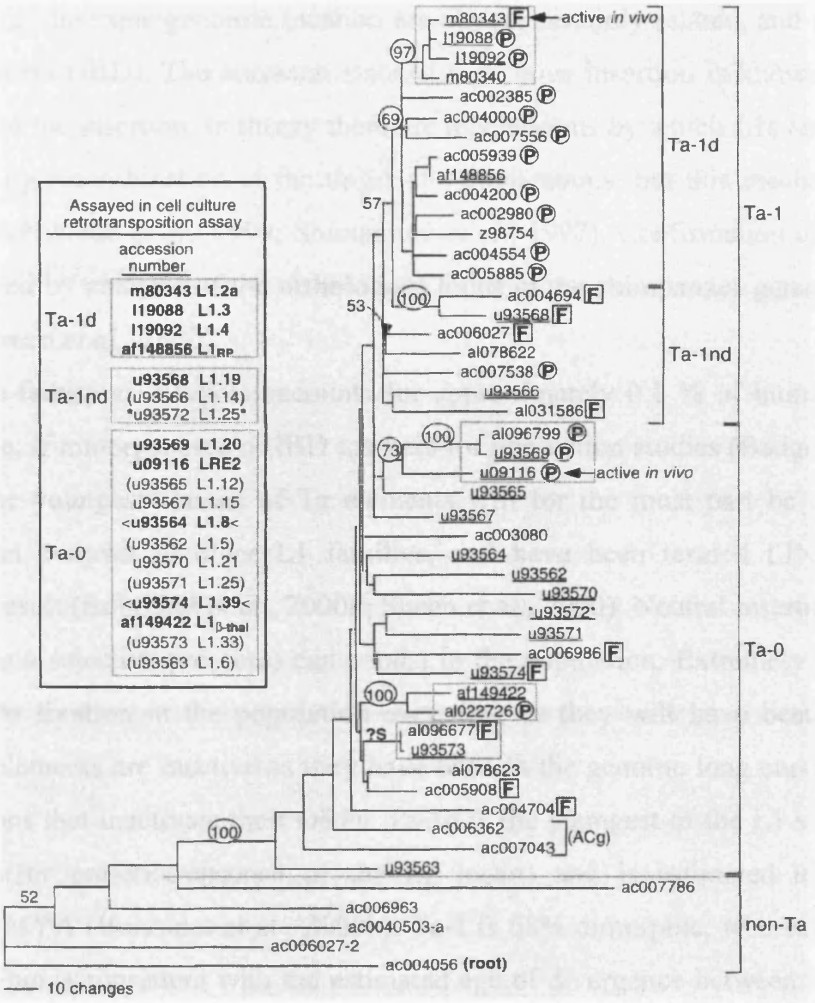


Figure 1.6: Taken from Boissinot *et al.* (2000) (fig 8, pg 923). The diagram shows a phylogenetic tree of full-length L1 elements. The tree is built on the entire L1 sequence exclusive of the 3' untranslated region (UTR). The node number indicates the percentage of appearance in 1,000 bootstrap replicates, where only values > 50% are indicated. The tree is rooted using a non-Ta L1 element. "F" represents elements fixed in human populations where elements labelled "P" show presence/absence polymorphism. The pre-Ta elements are indicated. Elements analyzed in earlier studies (Kimberland *et al.*, 1999; Sassaman *et al.*, 1997) are listed in the box, and underlined in the tree. Elements highlighted in bold type are active in a cell culture retrotransposition assay. The elements flanked by "<" are weakly active (retrotransposition frequencies of ≤ 10 % that of controls) and the bracketed elements contain interrupted open reading frames.

Another method for determining age relationships between L1 sub-families is a measure of their presence/absence polymorphic status in different individuals. The polymorphic nature of *de novo* L1 insertion gives a potential tool for population genetic analysis as individuals bearing

the same insertion in the same genomic location are almost certainly related, and the L1 shared via identity by descent (IBD). The ancestral state of a *de novo* insertion is known in principle, being the absence of the insertion. In theory there are mechanisms by which L1s can be removed from the genome, by recombination of the target site duplications, but this mechanism has not been demonstrated (Nikaido *et al.*, 1999; Shimamura *et al.*, 1997). Confirmation of the ancestral state can be achieved by analysis of the orthologous locus in the chimpanzee genome, and other higher primates (Sheen *et al.*, 2000).

The Ta sub-family as a whole accounts for approximately 0.1 % of human L1s and is therefore a valuable, if minor, source of IBD markers for population studies (Badge *et al.*, 2003). Loci containing the youngest classes of Ta elements will for the most part be dimorphic for presence/absence in contrast to older L1 families, and have been termed LINE-1 insertion dimorphisms as a result (Boissinot et al., 2000b; Sheen et al., 2000). Neutral insertions (i.e. those which do not cause a selective pressure) can persist in the population. Extremely old insertions will have moved to fixation in the population over time or they will have been removed by genetic drift. Old elements are inactive as they have been in the genome long enough to acquire deleterious mutations that inactivate their ORFs. Ta-1d is the youngest of the L1 sub-families, is 90 % dimorphic (for presence/absence of the L1 insert) and is estimated to have arisen approximately 1.4 MYA (Boissinot et al., 2000b). Ta-1 is 68% dimorphic, whereas Ta-0 is only 22 % dimorphic. This is consistent with the estimated age of divergence between the older Ta-0 element and the younger Ta-1 (Boissinot et al., 2000b).

1.3.2 The effects of Allelic variation on Retrotransposition efficiency

The existence of L1 is dependent on maintaining efficient retrotransposition, and generating numerous RC offspring. Being the youngest and most active L1 sub-family, Ta 1d is the most retrotranspositionally active sub-family; therefore it is likely that it will be the sire of the next most active sub-family that will succeed it. Less active families are more likely to become extinct as they will generate fewer RC offspring. Nucleotide substitutions leading to a higher degree of retrotranspositional activity will ultimately generate new L1 sub-families.

Interestingly, investigations into the activity of L1 has revealed a set of six elements account for the majority of L1 retrotransposition, referred to as "hot" L1s (Brouha *et al.*, 2003).

These L1s are categorised as elements that have a retrotransposition activity of at least 1/3 the activity of a RC L1 isolated from a patient with retinitis pigmentosa, called $L1_{RP}$ (Brouha *et al.*, 2003; Schwahn *et al.*, 1998). Using the cultured cell retrotransposition assay, the activity of five full-length elements associated with 14 known disease-causing insertions was assayed (Brouha *et al.*, 2003). $L1_{RP}$ (Schwahn *et al.*, 1998) and $L1_{\beta\text{-thal}}$ (Divoky *et al.*, 1996), are full-length insertions which have caused disease. L1.2 (Dombroski *et al.*, 1991) LRE2 (Holmes *et al.*, 1994) and LRE 3 (Brouha *et al.*, 2002) were identified as likely progenitors of the disease-causing insertions (Kimberland *et al.*, 1999; Ostertag *et al.*, 2000; Wei *et al.*, 2000). Where $L1_{RP}$, $L1_{\beta\text{-thal}}$ and LRE 3 were identified as "hot" elements, L1.2 and LRE2 were only weakly active by comparison. Further investigation into the activities of 82 elements isolated by long-range PCR amplification and cloning showed a range of activity of between 0.1 to 130 % the activity of $L1_{RP}$. By summing the total activity of all the elements, it was evident that 6 elements accounted for 84 % of the total activity. When the activity of the elements was compared to their estimated ages (based upon their divergence from a consensus element), it became evident that younger elements were more active, as had been previously predicted. The occurrence of disease-causing insertions such as $L1_{RP}$ and $L1_{\beta\text{-thal}}$ are unique in the population, as they are in effect *de novo* insertions. By contrast the "hot" elements, identified in this study, that are not disease-causing (LRE3, AC004200, AC002980, AL356438, AL512428, AC021017, and AL137845) will be more prevalent in the population as they have potentially no mutational burden. Of the 80 to 100 estimated potentially RC L1s in the genome (Brouha *et al.*, 2003), these "hot" elements are believed to be responsible for the majority of L1 retrotransposition in humans.

Further investigations into allelic variants of AL512428 (L1A), AC002980 (L1B), and AC021017 (L1C), isolated from 6 different ethnic populations, have been performed (Seleme Mdel *et al.*, 2006). Allelic variants showed retrotransposition activities of < 25 % the activity of $L1_{RP}$, and up to 175 % the activity of $L1_{RP}$, indicating that there are rare and extremely active L1s within the human population not found in the human genome sequence (which has only 6 hot L1s). For example a highly active element, not accounted for in the human genome sequence, was located in accession AL356438 (Brouha *et al.*, 2003), and has 73.8 % the activity of $L1_{RP}$. Acquisition of L1 elements from the human genome sequence is biased towards the identification of high frequency L1 alleles, as they are more likely to occur in the small pool of people sampled for the construction of the human genome sequence (Badge *et al.*, 2003; Myers *et al.*, 2002;

Ovchinnikov *et al.*, 2001; Seleme Mdel *et al.*, 2006; Sheen *et al.*, 2000). If a novel hot element existed in 1 in 1000 people, millions of copies would be present in the entire human population (Brouha *et al.*, 2003), so it is highly likely that numerous highly active elements exist in the human population that we do not know about. Although the Ta subfamily is believed to be the youngest and most active subfamily, other elements have the potential to form different subgroups (Casavant and Hardies, 1994). It is therefore possible that there could be groups of very rare and extremely active elements which are unknown at present, but in time could become the predominant active L1 subfamily, replacing Ta.

The 5′ end of most L1s is either truncated or inverted and truncated (Boissinot et al., 2000b; Lander et al., 2001; Ostertag and Kazazian, 2001b; Smit, 1999). It has been suggested that truncation arises due to low processivity of L1 RT (Ostertag and Kazazian, 2001a). Dissociation of the RT from L1 RNA during TPRT results in integration of a truncated element. The frequency with which an L1 element truncates during replication limits its ability to generate RC copies and consequently colonise a host genome (Farley *et al.*, 2004). If an L1 generates a full-length copy, and replication by the L1 RT is faithful, there is a strong chance that the offspring element will be RC. The generation of numerous active copies is essential to all mobile elements to prevent their extinction through mutation and genetic drift (Farley *et al.*, 2004). Mutations resulting in more processive RT variants have a higher chance of evading truncation, thus giving a higher proportion of full-length offspring elements (Farley *et al.*, 2004). Mutations preventing efficient reverse transcription will thus result in the loss of the mutated element, while those able to successfully complete reverse transcription will continue producing full-length copies. This would support the theory that younger presently active elements produce more full-length copies than older active sub-families, as suggested by Farley *et al* (2004).

Host cell factors which inhibit successful TPRT may also result in truncation. As L1 is, in most cases, a large insertion with a high mutational burden, it is likely that such insertions exert a selective pressure acting to reduce the accumulation of full-length elements (Farley *et al.*, 2004). If truncation was random, it would be expected that the lengths of truncated L1s would be evenly distributed. Instead it would appear that truncation is most likely to occur at the beginning of TPRT. If the first 2 to 3 kb of TPRT is successful the element is more likely to be full-length when reintegrated rather than truncated (Pavlicek *et al.*, 2002b). This suggests that cellular factors destabilise RT activity at the beginning of TPRT in order to prevent insertion of

full-length L1 elements. Such cellular inhibitors of the L1 RT have been identified (Bogerd *et al.*, 2006a; Bogerd *et al.*, 2006b). This opposes the theory that active L1 elements have a low RT processivity (Ostertag and Kazazian, 2001a). Elements encoding an RT with a low processivity will ultimately fail to generate RC offspring, as a result of inhibition by cellular factors, ultimately leading to extinction of that particular RT variant.

1.4 Non-autonomous retrotransposons

Non-LTR retrotransposons have the ability to mobilise other RNA transcripts in *trans* at low levels (Wei *et al.*, 2001). A distinct class of non-autonomous retrotransposons has evolved during LINE evolution that parasitises the LINE retrotransposition machinery (Dewannieux *et al.*, 2003; Schmid, 2003).

1.4.1 *Alu* elements

The *Alu* class of SINE retrotransposons was discovered as rapidly re-annealing primate DNA fragments that shared an *Alu*I restriction site (Houck *et al.*, 1979; Schmid and Deininger, 1975). *Alu* is the most abundant primate-specific repetitive element in the human genome (Bailey *et al.*, 2003). *Alu* has a copy number of approximately 1.1 M elements with each copy being around 300 bp in length, representing 11 % of the human genome by mass (Schmid, 2003). *De novo Alu* insertions have caused 20 known cases of human genetic disease (Kazazian, 2004). It has been estimated that approximately 1 in 30 individuals harbour a *de novo Alu* insertion (Kazazian, 2004).

Alu elements have a 282 nt consensus sequence derived from the 7SL signal recognition particle (SRP) RNA gene (Dewannieux *et al.*, 2003). *Alu* RNA is transcribed from an internal RNA pol III promoter (Schmid, 2003), unlike L1 which is thought to use RNA pol II (Kurose *et al.*, 1995; Moran and Gilbert, 2002; Swergold, 1990; Tchenio *et al.*, 2000). *Alu* has a long poly A tail (> 50 bp) (Roy-Engel *et al.*, 2002), removal of which abolishes *Alu* retrotransposition (Schmid, 2003). Although *Alu* is ubiquitously dispersed throughout the genome it is most commonly found in GC-rich genomic environments (Bailey *et al.*, 2003) which gives rise to its over-representation in gene-rich regions (Schmid, 2003). Within the human genome, three *Alu* families are present: the oldest *Alu* J is 65 to 40 MY old; *Alu* S is 45 to 25 MY old; and the

youngest presently active sub-family, *Alu* Y, originated around 30 MYA (Bailey *et al.*, 2003). *Alu* is a non-autonomous retrotransposon and relies on the L1 retrotransposition machinery for its mobilisation, and this has been demonstrated in cell culture using marked *Alu* elements (Dewannieux *et al.*, 2003).

The age and prevalence of *Alu*-S in the human genome correlates to a large-scale expansion of *Alu* copy number approx 35 to 40 MYA. Diagnostic analysis of aligned *Alu* sub-families suggests that *Alu*-S was the prominent sub-family at that time (Bailey *et al.*, 2003). It has been inferred that *Alu*-S contained sequences similar to pol III transcripts with strong SRP binding ability (Fan *et al.*, 1998). This may have brought about excessive clustering of *Alu* transcripts in the RER, increasing the level of co-localisation with the nascent proteins of L1s active at that time. The increased co-localisation would have led to a surge in *Alu*-S copy number.

The parasitic nature of *Alu* would have caused a subsequent decrease in L1 activity due to the level of L1 ORF2p recruited by *Alu*. However L1 is still active, therefore the surge in *Alu* copy number did not damage L1 fatally, most likely due to the *cis* preference of L1 proteins. It is therefore unlikely that the surge of *Alu*-S was purged by alterations in L1 structure, rather that a high number of *Alu* transcripts in the RER prevented normal processing of other SRP tagged cellular transcripts. Such a selective pressure would likely have caused divergence of *Alu* away from its strong SRP 9/14 binding ability to a less strongly SRP associating sequence. Sequence evolution to the presently active *Alu*-Y is proposed to have caused a decline in SRP 9/14 affinity during primate evolution (Fan *et al.*, 1998; Sarrowa *et al.*, 1997). Although such elements have evolved specifically to be able to retrotranspose at the expense of LINE elements they are not the only RNA species that the LINE retrotransposition machinery can mobilise (Pavlicek *et al.*, 2002a; Wei *et al.*, 2001).

1.4.2 SVA elements and HERV-W retrotransposition

SVA elements are composite retrotransposons comprised of SINE-derived sequences from human endogenous retrovirus (SINE-R); a variable number tandem repeats (VntR) segment, and a partial *Alu* sequence. Although only a few thousand SVA elements have been identified in the human genome, three cases of human disease have been linked to their insertion (Ostertag *et al.*, 2003). This suggests that SVA may show a particularly high frequency of mobilisation when

compared to the number of insertions which have caused disease from *Alu* (~1.1 M copies, 20 diseases) and L1 elements (~500 k copies, 17 diseases), despite their much smaller copy number (Ostertag *et al.*, 2003).

The HERV-W family of retroelements is the youngest of the human endogenous retroviral families, and includes the potentially active HML-2 subfamily of HERV-K. These elements may be the only HERVs active in the human genome. Unlike the HML-2 subfamily of HERV-K, which are complete proviruses (Macfarlane and Simmonds, 2004), HERV-W is unusual since it contains structures similar to retroviral mRNA but lacks complete LTRs and ORFs (Costas, 2002; Pavlicek *et al.*, 2002a), and so cannot be autonomous. The HERV-W family is also unusual since its copy number is not comparable to other HERV class retroelements. Unusually, HERV-W elements are commonly followed by a poly A tail and flanked by TSDs, a typical hallmark of L1 mobilisation. L1 is the only known retroelement that can mobilise RNA species in this way (Pavlicek *et al.*, 2002a).

HERV-W insertion has been linked to several human diseases including multiple sclerosis (Komurian-Pradel *et al.*, 1999), rheumatoid arthritis (Gaudin *et al.*, 2000), and upregulated expression of HERV-W has been linked to schizophrenia (Karlsson *et al.*, 2001). The prototype HERV-W *env* gene is located on chromosome 7 and has been suggested to be the gene which encodes syncytin. This protein is responsible for cell fusion during syncytiotrophoblast differentiation in human placental development (Mi *et al.*, 2000; Pavlicek *et al.*, 2002a).

1.4.3 Processed Pseudogenes

Pseudogenes are duplicated non-functional copies of genes with sequence similarity to the original gene derived from processed mRNA species. They often harbour multiple mutations and frequent stop codons (Pavlicek *et al.*, 2002a). The structure of processed pseudogenes is reminiscent of cDNAs which are the products of reverse transcription. As processed pseudogenes lack promoters, they are not considered to be retroelements; rather they are unique *de novo* events incapable of further proliferation.

Younger processed pseudogenes sometimes retain poly A tails and TSDs arising from L1 mobilisation of processed mRNA (Pavlicek *et al.*, 2002a), but most processed pseudogenes have lost these structures through sequence divergence. The estimated frequency of cellular mRNA

mobilisation by LINE is 0.01 to 0.05 % per LINE transposition (Pavlicek *et al.*, 2002a; Wei *et al.*, 2001). Bioinformatic analysis estimates a processed pseudogene copy number of 23,000 to 33,000 (Goncalves *et al.*, 2000) in the human genome, accounting for approximately 0.5 % of the genome by mass, equivalent to exon coverage (Lander *et al.*, 2001). This estimation however only accounts for processed pseudogenes with homology to protein-coding sequence. This figure should be increased by a factor of at least three when taking into account retrotransposition of all cellular mRNAs (Pavlicek *et al.*, 2002b). In view of the strong *cis* preference of the L1 retrotransposition machinery, in cell culture processed pseudogene formation seems to be rather common (Sassaman *et al.*, 1997; Wei *et al.*, 2001).

Processed pseudogenes are not usually active, but infrequently become activated (Pavlicek *et al.*, 2002a). Processed pseudogenes have on occasions inserted downstream of endogenous promoters or enhancer elements and become transcriptionally active. As the promoter is different to the promoter of the founder gene, intronless genes are often expressed at different times to their founder genes. This can include modification of cell specificity offering novel gene functions (Brosius, 1999). One such example is the testis-specific pyruvate dehydrogenase (PDH) subunit gene. This gene is derived from a mature mRNA transcript of the PDH-A1 subunit gene located on the X chromosome. PDH-E1 alpha is located on chromosome 4 and exhibits testis-specific expression. PDH-E1 alpha completely lacks introns but differs by only 26 nucleotides from the exonic sequence of the PHD-A1 subunit gene. Expression from this intronless gene gives rise to a testis-specific PDH (Dahl *et al.*, 1990). As L1 is the only known element that can mobilise cellular RNA species, it is likely that this processed pseudogene resulted from mobilisation by the L1 retrotransposition machinery. There are numerous intronless genes in the human genome of which a number of examples are tabulated in Brosius (1999). However it should be noted that distinctions between activated processed pseudogenes and genes which have apparently always been devoid of introns are often not clear cut.

1.5 The effects of L1 retrotransposition on genome plasticity

1.5.1 Direct L1 mutagenesis

L1 can cause disease by directly inserting into genes. A table containing information on known L1 and non autonomous retrotransposon insertions which have caused human disease can be found in appendix i. L1 insertions account for approximately 1 in 1200 human pathogenic mutations and an estimated 1 in 50 humans (Kazazian, 2004) harbour a *de novo* insertion occurring in the parental germline or early in embryonic development. Interestingly in a cultured cell retrotransposition assay, 5 to 10 % of *de novo* L1 insertions occur within the introns of actively transcribed genes. As the gene coverage of the human genome is approximately 15 % this suggests that there is no active mechanism preventing L1 from inserting into genes (Moran, 1999).

In 1988, two separate 5′ truncated *de novo* insertions into exon 14 of the human factor 8 gene were shown to have been the cause of haemophilia A in two separate patients (Kazazian *et al.*, 1988). This demonstrated that not only was L1 a common repetitive element but also an active transposon capable of (very occasionally) causing disease. Four truncated insertions into the dystrophin gene have been shown to cause Duchenne muscular dystrophy (Holmes et al., 1994; Narita et al., 1993) or cardiomyopathy (Yoshida *et al.*, 1998); a full-length (potentially RC) L1 into β-globin gene has been shown to cause β-thalassemia (Kimberland *et al.*, 1999); and the retinitis pigmentosa (RP2) gene was discovered due to the insertion of a full-length and potentially RC L1 (Schwahn *et al.*, 1998). A truncated somatic insertion into the APC gene has been linked with the development of colon cancer (Miki *et al.*, 1992). This insertion is important as it suggests that unchecked retrotransposition could result in tumorigenesis. Indeed, up-regulation of L1 RNA and L1ORF1p has been noted in certain epithelially derived tumours including breast sarcomas (Asch *et al.*, 1996), and germline tumours in both adults and children (Bratthauer and Fanning, 1992; Bratthauer and Fanning, 1993).

1.5.2 L1-mediated sequence transduction

L1-mediated transduction occurs when the L1 transcript extends into surrounding genomic DNA, retaining the flanking sequence throughout the retrotransposition cycle. 5′ transduction occurs through initiation of transcription from a promoter upstream of a full-length L1 followed by co-mobilisation of the flanking sequence (Pavlicek *et al.*, 2002a; Pickeral *et al.*, 2000; Szak *et al.*, 2003). L1 3′ transduction, which seems to be much more common, is associated with read-through of the L1 transcript into 3′ flanking DNA, when the weak L1 poly A signal is located 5′ of a stronger poly A signal located in the genomic DNA. 10 % to 20 % of recent human insertions characterised in cultured cell retrotransposition assays contain 3′ transductions (Pickeral *et al.*, 2000). Transduced sequence is estimated to account for 0.5 % to 1 % of the genome by mass, the equivalent of exon sequence (Pavlicek *et al.*, 2002a; Pickeral *et al.*, 2000). This may be an under-estimation as 3′ transduction followed by severe 5′ truncation may result in transduced sequences that lack any L1 sequence, and so are unlikely to be recognised as L1-mediated transduction events.

Transduction can also cause shuffling of exons and regulatory regions to new genomic locations thus potentially activating pseudogenes or altering the characteristics of active genes. Shuffling of genomic DNA maintains genome plasticity and may promote genome evolution (Pickeral *et al.*, 2000). The effect of retrotransposon shuffled functional elements and genomic sequence may have played a role in the chimpanzee/human divergence (Brosius, 1999).

Exon shuffling has been implicated in the formation of novel genes, for example the SCAN domain containing 2 gene (SCAND2) (Dupuy *et al.*, 2002). The SCAN domain is most commonly associated with the N terminus of the Krüppel-like zinc finger protein family (Bellefroid *et al.*, 1989; Bellefroid *et al.*, 1991; Williams *et al.*, 1995). SCAND2 however does not belong to the Krüppel-like zinc finger protein family (Dupuy *et al.*, 2000). The C terminal 167 amino acids of SCAND2p is a unique protein sequence with no significant similarity to any other known protein (due to out of frame translation of transduced sequence). The region of the mRNA which encodes the C terminus of SCAND2p domain shows 91% identity to a section of the *Clorf12* gene DNA sequence minus its introns (Dupuy *et al.*, 2000), reminiscent of retrotransposition of a cellular RNA species. Read-through of an L1 transcript into the flanking *Clorf12* gene is likely to have resulted in 3′ transduction. The transduction most likely inserted into intron 1 of a member of the SCAN family. The SCAND2 protein is believed to play a role in

transcriptional regulation by modulating SCAN protein transcriptional repression (Dupuy *et al.*, 2002).

1.5.3 The effects of L1 insertion on gene expression

L1 insertions into exons can clearly disrupt gene function, and L1 insertion into introns can also influence gene expression rather than merely being spliced out of mRNAs.

One observation is that highly expressed genes exhibit a lower percentage of L1 sequence when compared to less highly expressed genes (Han *et al.*, 2004). Genic L1 insertions are mostly intronic as they have a much lower mutational impact than insertion into exons. It had previously been assumed that genic transcribed L1 sequence would be simply spliced out and therefore have little effect on gene expression. However, given that L1 insertion is not tightly constrained (Feng *et al.*, 1996; Moran, 1999), and that the distribution of old L1 elements shows a strong bias for AT-rich DNA (Lander *et al.*, 2001), there would appear to be a selective pressure which removes L1 sequence from expressed regions of the genome (Boissinot *et al.*, 2001). This pressure could result from the inhibitory effect of L1 sequence on transcription.

Although the survival of retrotransposons is dependent on their efficient transcription, the presence of L1 ORF2 in the sense orientation has been shown to result in transcriptional inhibition on the host gene (due to the A-rich nature of L1s sense strand), with the effect being more potent than the presence of L1 in the antisense orientation (Han *et al.*, 2004). This potent interference in the transcription levels of genes is however not due to premature termination of transcription. No specific region resulting in the loss of transcription efficiency has been identified, but it has been demonstrated that the A rich bias in the sense strand is responsible (Han *et al.*, 2004). A highly active synthetic element that reduces the A richness of the sense strand of a mouse L1 by up to 40 % shows a marked reduction in transcriptional interference (Han and Boeke, 2004). Nuclear run-on assays have demonstrated that transcription initiation occurs efficiently, but engaged RNA polymerase density increases when L1 sequence is reached. This suggests that the polymerases do not elongate efficiently through L1 ORF2 sequence, thus reducing the overall quantity of the transcript (Han *et al.*, 2004). It has therefore been proposed that intronic L1s can modulate gene expression (Han *et al.*, 2004).

L1 inserted in the antisense orientation relative to a transcribed gene presents numerous polyadenylation sequences (Han *et al.*, 2004). Premature polyadenylation of cellular transcripts by L1 has become beneficial to certain genes by generating shortened stable isoforms of cellular mRNAs (Han *et al.*, 2004). For example, in human lymphoid tissue there are two prominent Attractin isoforms; one isoform is a transmembrane protein and the other is a secreted protein (Tang *et al.*, 2000). The protein sequence of the secreted isoform contains a short C terminal domain encoded by an exon generated by a 212 bp L1 insertion. The insertion also contains the stop codon, 3′ UTR and polyadenylation signal for the transcript, believed to be the native poly A signal of the inserted L1. The expression of two isoforms of the attractin gene is conserved in the mouse genome. This may be an example of how an L1 insertion has generated a novel mechanism for regulating the inflammatory response (Tang *et al.*, 2000). Another gene which uses a polyadenylation signal generated by L1 is the kinetochore protein Spc25. This gene is normally polyadenylated at the Major Poly-adenylation Signal (MPAS) of an L1 insertion (Wheelan *et al.*, 2005).

As described above, L1 sequence has an effect on the normal transcription of genes. The effects on transcription are the result of extension of the transcription unit by insertion, by premature polyadenylation (when the L1 sequence is in the antisense orientation), and through the inhibition of transcriptional elongation due to the A-rich nature of the L1 element (when the L1 sequence is in the sense orientation). Premature polyadenylation caused by the presence of L1 in the antisense orientation can also result in the phenomenon of gene breaking.

1.5.4 Transcription of human genes from an L1 antisense promoter (ASP) and gene breaking

L1 contains an antisense promoter (ASP) at position 400 to 600 of the 5′ UTR (Speek, 2001), which operates in the opposite orientation to the regular L1 promoter. Approximately 1/3 of randomly selected genomic $L1_{HS}$ show high level ASP activity in transfected Ntera2D1, *HeLa* and JEG3 cells (Speek, 2001). Given the genomic density of L1 sequences, there may be a large number of potentially active ASPs in the human genome (Nigumann *et al.*, 2002; Speek, 2001), and it is possible that some transcripts initiated from L1 ASPs might be translationally competent.

L1, when inserted into the antisense orientation to native genes, has the ability to truncate cellular transcripts by premature polyadenylation (discussed above) (Han *et al.*, 2004), as well as generate new transcription units from its ASP (Speek, 2001). Consequently, L1 insertion into an intron of a cellular gene could theoretically act to break the gene and produce two or three transcripts from a single native transcription unit (Wheelan *et al.*, 2005). There is evidence that gene breaking exists in the human genome from bioinformatic identification of 15 potential genes and transcription units which could potentially exhibit gene breaking by L1 insertion (Wheelan *et al.*, 2005). Twelve of the fifteen predicted events pre-date the human/chimpanzee divergence suggesting they have been retained as functional variants.

1.5.5 Genome instability caused by 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). Large scale instability attributed to L1 insertion including chromosomal inversions of approximately 120 kb in length, and chromosomal deletions of > 11 kb in length have been observed (Gilbert *et al.*, 2002; Han *et al.*, 2005; Symer *et al.*, 2002) in cultured cells.

Such phenomena are known to have occurred during the evolution of the human genome, but rearrangements of this scale, linked to L1 insertion, had not been observed during analysis of the human genome sequence (Han *et al.*, 2005). However in 2007 an approximately 46 kb genomic deletion in the PDHX gene was directly linked to insertion of a full-length L1. This pathological insertion resulted in a patient with pyruvate dehydrogenase complex (PDHc) deficiency (Mine *et al.*, 2007). Although the exact mechanism resulting in L1-derived large scale deletions and inversions is unknown, it has been hypothesised that erroneous second-strand nicking by the EN domain of ORF2p may be involved (Han *et al.*, 2005). A potential mechanism for large scale deletion associated with L1 is shown in figure 1.7.

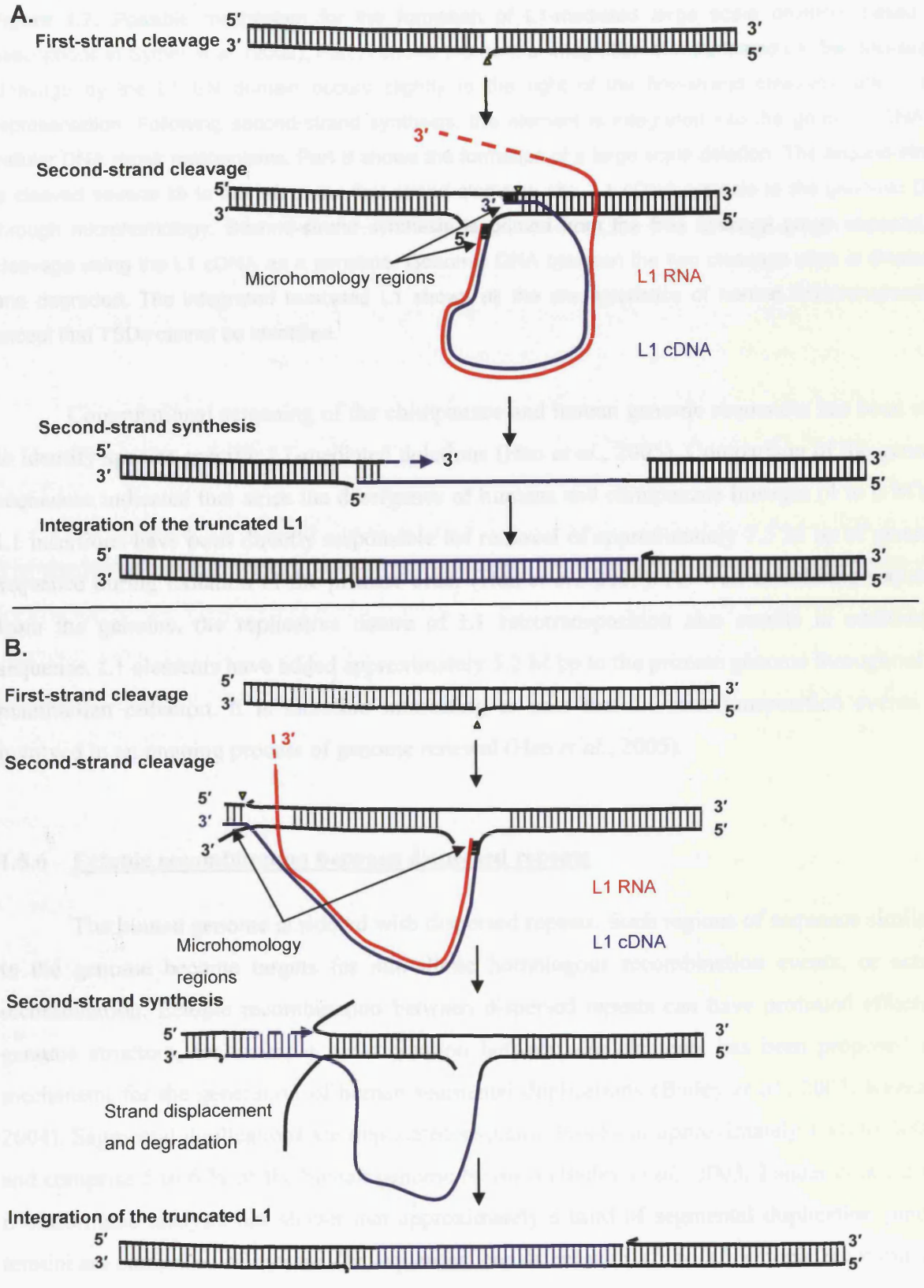

A.
First-strand cleavage
5′
3′
3′
5′
Second-strand cleavage
3′
5′
3′
3′
5′
5′
Microhomology regions
L1 RNA
L1 cDNA
Second-strand synthesis
5′
3′
3′
5′
3′
5′
Integration of the truncated L1
5′
3′
3′
5′
B.
First-strand cleavage
5′
3′
3′
5′
Second-strand cleavage
3′
5′
3′
3′
5′
L1 RNA
L1 cDNA
Microhomology regions
Second-strand synthesis
5′
3′
3′
5′
Strand displacement and degradation
Integration of the truncated L1
5′
3′
3′
5′

Figure 1.7: Possible mechanism for the formation of L1-mediated large scale deletion, based on descriptions in Symer *et al.* (2002). Part A shows the normal integration of a truncated L1. Second-strand cleavage by the L1 EN domain occurs slightly to the right of the first-strand cleavage site in this representation. Following second-strand synthesis, the element is integrated into the genomic DNA by cellular DNA repair mechanisms. Part B shows the formation of a large scale deletion. The second-strand is cleaved several kb to the left of the first-strand cleavage site. L1 cDNA anneals to the genomic DNA through microhomology. Second-strand synthesis is primed from the free hydroxyl group exposed by cleavage using the L1 cDNA as a template. Genomic DNA between the two cleavage sites is displaced and degraded. The integrated truncated L1 shows all the characteristics of normal retrotransposition, except that TSDs cannot be identified.

Computational screening of the chimpanzee and human genome sequences has been used to identify species-specific L1-mediated deletions (Han *et al.*, 2005). Comparison of the genome sequences indicated that since the divergence of humans and chimpanzee lineages (4 to 6 MYA) L1 insertions have been directly responsible for removal of approximately 7.5 M bp of genomic sequence during radiation of the primate order (Han *et al.*, 2005). As well as deleting sequence from the genome, the replicative nature of L1 retrotransposition also results in addition of sequence. L1 elements have added approximately 5.2 M bp to the primate genome throughout the mammalian radiation. It is therefore interesting to note that L1 retrotransposition events are involved in an ongoing process of genome renewal (Han *et al.*, 2005).

1.5.6 Ectopic recombination between dispersed repeats

The human genome is riddled with dispersed repeats. Such regions of sequence similarity in the genome become targets for non-allelic homologous recombination events, or ectopic recombination. Ectopic recombination between dispersed repeats can have profound effects on genome structure, for example recombination between *Alu* elements has been proposed as a mechanism for the generation of human segmental duplications (Bailey *et al.*, 2003; Kazazian, 2004). Segmental duplications are duplicated sequence blocks of approximately 1 kb to 300 kb, and comprise 5 to 6 % of the human genome by mass (Bailey *et al.*, 2003; Lander *et al.*, 2001). Bioinformatic analysis has shown that approximately a third of segmental duplication junction termini are comprised of mosaic *Alu* sequences, consistent with *Alu* to *Alu* ectopic recombination events (Bailey *et al.*, 2003).

A burst of segmental duplication formation also correlates with a surge in *Alu*-S copy number approximately 40 MYA which would have presented many targets for ectopic recombination (Bailey *et al.*, 2003). This *Alu*-S surge occurred during primate evolution, and thus humans and other primates have a much greater proportion of segmental duplications compared to other mammals (Bailey *et al.*, 2003). An estimate of the current rate of *Alu* insertion is approx 1 in every 16 to 200 individuals (1 per 32 to 400 genomes) (Kazazian, 1999). This is predicted to have been between 30 and 200 times higher 40 MYA (Bailey *et al.*, 2003; Ohshima *et al.*, 2003), and thus exerting a huge impact on genomic structure. The human genome underwent a secondary *Alu* burst on a much smaller scale between 2.5 and 3.5 MYA (Carroll *et al.*, 2001), subsequent to human / chimpanzee divergence. This suggests the possibility that there may be a higher percentage of segmental duplication in the human genome compared to the chimpanzee genome.

The rearrangement of the human genome by segmental duplications has resulted in novel changes at the genetic level as well as the structural level. Duplicated regions are littered with paralagous copies of genes and partial copies of genes, promoters and other regulatory elements (Samonte and Eichler, 2002). Through genomic rearrangement, segmental duplications harbouring non-functional intronic and exonic sequences have resulted in the creation of novel fusion/fission genes (Lynch and Conery, 2000). Although there is evidence that segmental duplications can result in the formation of pseudogenes, it is an open question as to whether this process has resulted in any evolutionary consequences. There is however some evidence to support this theory. One example is a testis-expressed creatine transporter gene (Iyer *et al.*, 1996). Creatine and creatine phosphate are essential for regeneration of ATP in tissue with fluctuating energy demands, for example muscle. Another cell type with a high demand for ATP is spermatozoa. The main creatine transporter gene, CT1, is located on the X chromosome, a location which is maintained in rabbit and mouse genomes. A duplicated section of the Xq28 region has been identified near the 16p11.1/16p11.2 boundary (Eichler *et al.*, 1996). Within this duplicated section a secondary creatine transfer gene, CT2, has been identified and mapped to 16p11.2. This gene is exclusively expressed in the testis (Iyer *et al.*, 1996), and since the X chromosome is inactivated during spermatogenesis, this gene is essential.

Considering that all chromosomes in the human genome harbour segmental duplications (Lander *et al.*, 2001), it is likely that numerous novel genes have arisen form the formation of

segmental duplications. Although the majority of transcripts originating in segmental duplications frequently contain premature stop codons, and have low coding potential (Lynch and Conery, 2000), novel genes are still generated.

1.5.7 The potential role of L1 in X chromosome inactivation

Inactivation of the X chromosome is a well established mechanism of gene regulation. During female early embryogenesis, the majority of genes on the X chromosome are transcriptionally silenced to maintain dosage of X linked genes between XX cells and XY cells.

X inactivation initiates at a unique *cis* acting region of the X chromosome named the X inactivation centre (Rastan, 1983). The X inactivation centre was mapped to the central region of the human X chromosome (Xq13) by analysis of the spread of chromosomal inactivation in X:autosomal translocations. The X inactivation centre is composed of an abundance of genes for non-coding RNAs (Chureau *et al.*, 2002). The *Xist* RNA is uniquely only expressed from the inactivated X chromosome in female somatic cells (Borsani *et al.*, 1991), and inactivated X chromosomes are coated with *Xist* RNA suggesting that it acts as a structural RNA (Brown *et al.*, 1992; Clemson *et al.*, 1996). X inactivation initiates early in development, and spreads from the X inactivation centre in *cis* along the length of the X chromosome, and is maintained through successive somatic cell divisions (Bailey *et al.*, 2000). As a result, it has been suggested that features required for *cis* inactivation spreading must be enriched on the X chromosome, but not exclusive to it (Bailey *et al.*, 2000; Gartler and Riggs, 1983; Lyon, 2003).

Mary Lyon's repeat hypothesis suggests L1 as a candidate for a *cis* acting feature in X inactivation spreading (Lyon, 1998). The X chromosome is nearly two fold enriched for L1 compared to the autosomes, and is significantly enriched at Xq13 where the X inactivation centre is located (Bailey *et al.*, 2000; Lander *et al.*, 2001; Ross *et al.*, 2005). If L1 sequences act as way stations for X inactivation, the attenuated spread of X inactivation from the X inactivation centre can be accounted for (Popova *et al.*, 2006). This is supported by the fact that genomic loci on the X chromosome which escape inactivation show significantly lower L1 density, closely resembling the average genome L1 density (Bailey *et al.*, 2000; Ross *et al.*, 2005). Also, the failure of chromosomal inactivation beyond the boundaries of X-chromosomal sections of X:

autosomal translocations has been correlated with chromosome bands poor in L1 sequence (Bailey *et al.*, 2000).

Although these seemingly unconnected events support Lyon's hypothesis, it is possible that L1 enrichment is simply a consequence of the heterochromatic nature of the inactive X chromosome. It could be possible that rather than being the initial mechanism by which X inactivation occurs, L1 has contributed to a more stable mechanism for X inactivation (Bailey *et al.*, 2000).

1.6 L1 proliferation is dependent on retrotransposition in the germline

In order for L1 retrotransposition to be of evolutionary consequence, it must occur in the germline or during the early stages of embryogenesis prior to germline differentiation (Ergun *et al.*, 2004). Molecular parasites such as L1 are selfish in that their whole structure is geared towards generating offspring elements (Bestor, 1999; Hickey, 1982). As the majority of LINE elements are contained in inactive chromatin, and are densely methylated in normal somatic cells, methylation is thought to be a cellular defence against the harmful effects of L1 retrotransposition (Hata and Sakaki, 1997).

1.6.1 Methylation prevents L1 retrotransposition

Hypomethylation in primordial germ cells removes transcriptional blocks, prior to *de novo* sex-specific imprinting during gametogenesis (Kierszenbaum, 2002; Li, 2002; Mann, 2001). The majority of spermatid-specific and oocyte-specific methylation modifications are removed post fertilisation prior to *de novo* methylation during early embryogenesis (Brandeis *et al.*, 1993). This provides two potential "windows of opportunity" for L1 retrotransposition (Schulz *et al.*, 2006). It is also possible that differential methylation during oocyte and spermatid development may cause differential dynamics of L1 retrotransposition in the male and female germlines. Maternal methylation imprints occur in the absence of DNA replication during oocyte growth (Lucifero *et al.*, 2004), while paternal-specific methylation occurs continuously throughout both

the mitotic and meiotic stages of spermatogenesis (Davis *et al.*, 1999; Davis *et al.*, 2000; Kerjean *et al.*, 2000; Ueda *et al.*, 2000).

1.6.2 Certain cellular factors modulate L1 retrotransposition

It has been theorised that there is an evolutionary arms race resulting in the rapid evolution of the APOBEC3 protein family, which are involved in the inhibition of viral and retroviral reverse transcription (Sawyer *et al.*, 2004; Zhang and Webb, 2004). The APOBEC3 proteins can inhibit endogenous LTR retroviruses as well as exogenous retroviruses in murine cells (Bogerd *et al.*, 2006a; Cullen, 2006; Dutko *et al.*, 2005; Esnault *et al.*, 2005), but in humans LTR retroelements are practically extinct. LTR retroelement activity can therefore not exert selective pressures that would cause a protein family to evolve at an increased rate (Bogerd *et al.*, 2006b), but the APOBEC3 protein family is still rapidly evolving. A3C shows nuclear localisation, and therefore has the potential to inhibit TPRT-driven L1 retrotransposition. The cultured cell retrotransposition assay was used to determine whether any of the APOBEC3 family members that enter the nucleus can inhibit L1 retrotransposition (Bogerd *et al.*, 2006b). The retrotransposition rate of an engineered indicator L1 (Wei *et al.*, 2001; Wei *et al.*, 2000), was assayed by co transfecting cells with plasmids containing APOBEC3 family genes. Expression of A3A and A3B caused significantly reduced levels of retrotransposition from the engineered L1 (Bogerd *et al.*, 2006b), suggesting these proteins can inhibit L1 retrotransposition.

Interestingly, *Alu* mobilisation does not require L1 ORF1p, and is solely modulated by L1 ORF2p (Dewannieux *et al.*, 2003). Co-transfection of an *Alu* reporter plasmid (Dewannieux *et al.*, 2003) with A3A or A3B showed inhibition of *Alu* mobilisation, leading to the conclusion that A3A and A3B inhibit *Alu* and L1 retrotransposition through interaction with ORF2p. The expression of A3B, a potent inhibitor of L1 retrotransposition, can be readily detected in human ovaries, testes and embryonic stem cells. This suggests that the A3B protein may be expressed to exert an inhibitory force against the mutagenic effects of L1, and consequently *Alu*, insertions at these crucial stages of development (Bogerd *et al.*, 2006b).

1.6.3 L1 retrotransposition in the human germline

Evidence of L1 retrotransposition in the female germline has also been presented: A *de novo* insertion of LRE3 into exon 4 of the CYBB gene was identified in a male patient diagnosed with chronic granulomatous disease (CGD). As the disease is X-linked, the simplest explanation for the insertion was retrotransposition during maternal meiosis (Brouha *et al.*, 2002). However this study failed to fully address the possibility that the *de novo* L1 insertion in the mother had occurred prior to the meiotic stages of oogenesis. Subsequent investigation of a full-length L1 insertion into the CHM gene of a Dutch patient ($L1_{CHM}$) indicated somatic, and germ-line, mosaicism for the L1 insertion in the patient's mother. Further investigation provided evidence that L1 retrotransposition does occur very early in human female embryonic development and is not limited to the meiotic stages of germ cell development (van den Hurk *et al.*, 2007; van den Hurk *et al.*, 2003).

It is unknown whether L1 retrotransposes during male early embryogenesis, although there is no reason to suspect that this is not the case. However, it is not clear when L1 retrotransposes during spermatogenesis. The sperm nucleus is packaged into a highly compact structure to create a mobile cell for male genome delivery. Studies in the mouse have shown that basic DNA associated proteins, protamines, are necessary for post-meiotic chromatin condensation (Lee *et al.*, 1995). Protamines are histone H1-derived, sperm-specific histone variants which associate with sperm DNA, allowing tight chromatin packaging (Lewis *et al.*, 2004; Lewis *et al.*, 2003; Wouters-Tyrou *et al.*, 1998). The dense packaging of DNA in sperm chromatin is likely to be a poor substrate for the L1 EN domain, and it is widely accepted that the sperm nucleus is transcriptionally inactive after nuclear condensation. It is therefore unlikely that retrotransposition occurs while the chromatin is tightly packaged. By contrast, the oocyte does not show similar chromatin packaging, so is likely to be more permissive to retrotransposition.

In order to retrotranspose *in vivo* both ORF1p and ORF2p are required (Moran *et al.*, 1996). Immunohistochemical analysis had previously only been able to detect ORF1p in testicular carcinoma cell lines. Using a monoclonal antibody directed against the ORF2p EN domain in conjunction with an antibody directed against ORF1p, cell type limited co-expression of ORF1p and ORF2p has been demonstrated (Ergun *et al.*, 2004). Co expression of ORF1p and ORF2p was detected in foetal pre-spermatogonia, spermatocytes, and immature spermatids. The detection of ORF1p and ORF2p in pre-spermatogonia is of interest as it could potentially lead to

germline mosaicism within individuals. Interestingly ORF1p and ORF2p are co-expressed in adult secondary spermatocytes, and immature spermatids, but not in spermatogonia. This suggests that L1 is active during and after meiosis but not during mitotic spermatogonia differentiation. L1 ORFp co-expression was also found to occur in some somatic cells such as Leydig, Sertoli and vascular endothelial cells (Ergun *et al.*, 2004). The presence of L1 ORF1 and ORF2 proteins during spermatogenesis suggests that not only is L1 being transcribed, but is also being translated. It is likely therefore that L1 is able to retrotranspose during spermatogenesis.

As sperm DNA is the only readily accessible source of human germline DNA, investigations into the rate of L1 retrotransposition in the germline are dependent on L1 retrotransposition occurring in the male germline. The identification of L1 ORF1p, ORF2p, L1 RNA and APOBEC3B in the human testes suggests that not only are all the components of active L1 elements present, but that there is also an inhibitory response to L1 retrotransposition. These factors all suggest that L1 retrotransposition should occur in the male germline, although to date *de novo* L1 retrotransposition in the male germline has not been demonstrated. Methods developed to detect *de novo* insertions directly in the male germline must bear in mind that while all necessary the components for L1 retrotransposition are present during spermatogenesis, this does not mean that retrotransposition occurs.

1.7 Project overview

From the preceding literature review it can be seen that L1 retrotransposition has had a profound effect on the evolution of the genome. To date, the direct detection of *de novo* L1 insertions in the human germline has not been achieved, except by chance in the case of disease-causing L1 insertions. Very little is therefore known about the dynamics of L1 retrotransposition. Three main factors have hampered attempts to access *de novo* L1 insertions: firstly there are no human germline cell cultures; secondly L1 is a relatively small insertion that can insert anywhere within an extremely large genome; finally the rate of *de novo* insertion is apparently extremely low (based on current estimations). My project was therefore to develop hybridisation selection as a mechanism to physically recover complete L1 insertions into target amplicons of limited size, to provide information on local insertion rates, the structure of insertions, the nature of the source element, and the presence of transduced sequence. Such

information would give a broad picture of L1 dynamics in the human male germline, the rates and structure of insertions into the genome of a single individual, and provide a method by which potential differences in the dynamics of insertion between different individuals could be investigated.

1.7.1 The challenges of analysing *de novo* L1 insertions

In order to analyse *de novo* L1 insertions into the human genome, two major challenges had to be overcome. Firstly the rate of L1 retrotransposition is very low; and the insertion pattern of L1 is essentially random in nature. Various estimations of the rate of L1 retrotransposition have been made previously. The most optimistic estimation was made by Haig Kazazian in 1999. At the time of his estimation the human mutation database contained 28 mutations resulting from Non-LTR retrotransposon insertions, 15 of which were L1 insertions (Kazazian, 1999). Thus of the 16,650 independent mutations in 860 genes recorded in the database at that time, 1 in 600 were Non-LTR retrotransposon-derived, and 1 in 1200 L1-derived. Assuming that the human nucleotide substitution rate is 1.3×10^{-9} per nucleotide per year, the size of the haploid genome is $3X10^{9}$ bp, and the human generation time is 25 years, every individual inherits 150 mutations within their genome, 75 from each parent. As L1 accounts for at most 1 in 1200 of reported mutation, 1 in 8 humans ($75 \times 2 \times 8 = 1200$) have a *de novo* L1 insertion somewhere in their genome (Kazazian, 1999). However subsequent estimates have been very diverse, for example from 1 in 33 humans harbouring a *de novo* L1 insertion (Brouha *et al.*, 2003), to 1 in 120 humans harbouring a *de novo* L1 insertion (Li *et al.*, 2001). At these rates we would expect a single insertion per 4.8×10^{7} kb to 1.5×10^{9} kb of germline DNA.

The only practical source of human germline DNA, at the quantities required for this project, is sperm DNA. A donor panel of normal volunteers was available in the Department of Genetics, University of Leicester. From the more optimistic estimation of L1 retrotransposition rate, 1 in 16 sperm will harbour a single *de novo* insertion somewhere within their genome. This presents the formidable challenge of locating 6 kb or less of *de novo* L1 sequence somewhere within 4.8×10^{7} kb of genomic DNA.

1.7.2 Initial attempts to analyse *de novo* L1 retrotransposition

A method for screening the genomic DNA (gDNA) to identify low frequency L1 insertion has been developed (Badge *et al.*, 2003). Amplification Typing of L1 Active Sub-families (ATLAS) is a suppression Polymerase Chain Reaction (PCR) based approach which can discriminate between the full-length Ta L1 sub-family and the overwhelming majority of old L1s in the genome. Figure 1.8 outlines the methodology behind the ATLAS approach. ATLAS directly displays polymorphic (in terms of presence / absence) Ta L1 insertions which can then be cloned and sequenced (Badge *et al.*, 2003).

ATLAS is by nature designed to be specific for the identification of Ta L1s. Although Ta is the most active L1 sub-family, ATLAS cannot identify other potentially RC L1 sub-families. The sequence information provided by ATLAS represents L1 / genomic sequence junction fragments. These fragments are short and therefore contain little information as to the structure of the insertion. Validation of polymorphic Ta insertions identified by ATLAS is dependent on designing PCR primers close to the L1 insertion site, and amplifying the insertion from gDNA of the individual from which the insertion was identified. Genotyping a population for newly identified insertions can be used to determine the presence and rate of occurrence of such insertions.

Work is presently being undertaken to adapt the display method to allow the detection of *de novo* insertions in the human germline (Macfarlane, unpublished data). This will involve screening of small pools of sperm DNA. As the pools of sperm in a particular experiment come from one individual, novel bands in the display autoradiograph could potentially arise from *de novo* insertion. However, the ATLAS method has inherent problems when attempting to identify *de novo* insertions.

ATLAS is capable of identifying low frequency insertions due to discrimination between the Ta L1 subfamily, and by recovering short L1 junction fragments. Although this allows the screening of a large area of the genome in a single method, the majority of the L1 sequence is lost in the PCR. Novel junction fragments generated from *de novo* insertions cannot be validated as the molecule from which the fragments originated is lost. Sequencing can confirm the presence of L1 sequence and flanking genomic sequence. A certain level of validation can be achieved by comparing the structure of sequenced junction fragments to known young L1 insertions and

identifying distinct features, for example an intact YY1 binding site and insertion at a consensus EN recognition site.

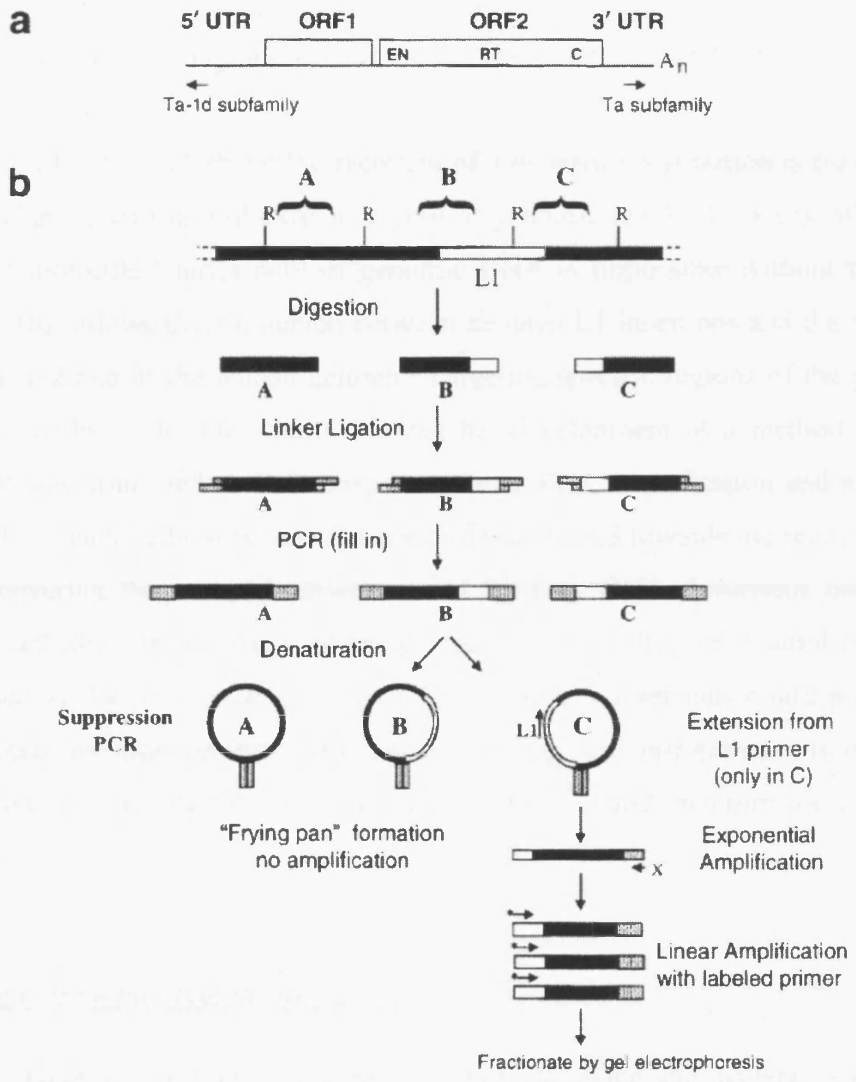


Figure 1.8: Taken from Badge *et al.* 2003 (figure 1, pg 824). a) An overview of L1 structure indicating the location of Ta-1d specific and Ta specific primers. b) The principles of ATLAS. Firstly genomic DNA is digested and ligated to a GC-rich oligonucleotide linker. This creates a fragment library flanked by defined oligonucleotide sequences. PCR amplification using linker specific primers is followed by denaturation that allows frying pan shaped loops to form through self-annealing of the linker sequences. Suppression PCR involves amplification using a linker-specific primer and an L1-specific primer. Extension will only commence if the loop structure contains an L1. Linear amplification using an element-specific radio-labelled primer in the absence of a second primer allows autoradiography after the fragments are fractionated by gel electrophoresis. The white bars indicate L1 sequence, the black bars indicate genomic non-L1 sequence, and the speckled bars represent linker sequence.

Amplification across the proposed insertion site from blood and sperm DNA of the same donor can also be used to demonstrate the absence of the insertion prior to spermatogenesis and show that the insertion did not originate from germline mosaicism. However the method is vulnerable to the possibility that junction fragments could be generated through PCR artefacts, rather than genuine *de novo* insertion.

The major difficulty underlying the recovery of a *de novo* L1 insertion is that, initially, a single-molecule of the insertion will exist in a pool of genomic DNA. Recovery of a single *de novo* L1 insertion molecule from a pool of genomic DNA is impossible without the use of a targeted approach that allows discrimination between *de novo* L1 insertions and the vast number of L1 insertions contained in the human genome. Targeting specific regions of the genome can however be achieved by PCR. This thesis outlines the development of a method designed to recover full-length and truncated *de novo* insertions using PCR amplification and hybridisation recovery of the PCR-generated molecules. The method was geared towards the recovery of entire insertions, by recovering the complete insertion and flanking DNA. Insertions could then be tracked back through the various stages of the protocol, to the initial PCR amplification, thus allowing validation at the level of the initial PCR. Full-length insertions could potentially be assessed for retrotransposition competence using the cultured cell retrotransposition assay. The method also did not discriminate between young L1 subfamilies, and therefore was not restricted to the Ta subfamily.

1.7.3 Targeting of genomic regions using PCR

The first challenge of such a method is where to place the targets. As previously discussed, the insertion of L1 is essentially random throughout a vast genome. Target locations were selected for different reasons. L1 disease-causing insertions are in effect *de novo* events, therefore it is reasonable to assume that if a region of the genome has been amenable to L1 insertion once, it remains amenable to L1 insertion. This makes such sequences good regions to begin screening for *de novo* insertions. Identification of a 6 to 7 kb region, devoid of $L1_{HS}$, at the chosen locus was required. While the target site sequences were required to be free from $L1_{HS}$ sequence, they also had to contain potential EN recognition sequences.

Following target site selection and determination of the suitability of each target, a set of primary target site primers and nested secondary target site primers were designed for each target site. In order to ensure the target site primers could efficiently amplify over the final 5 kb the target site, plus flanking sequence of comparable length to the approximate size of a full-length L1 and any transduced sequence (~ 7 kb), exterior control primers were also required (Figure 1.9). This allowed the Target site primers to be tested without constructing artificial target loci, which would have presented a potential source of contamination.

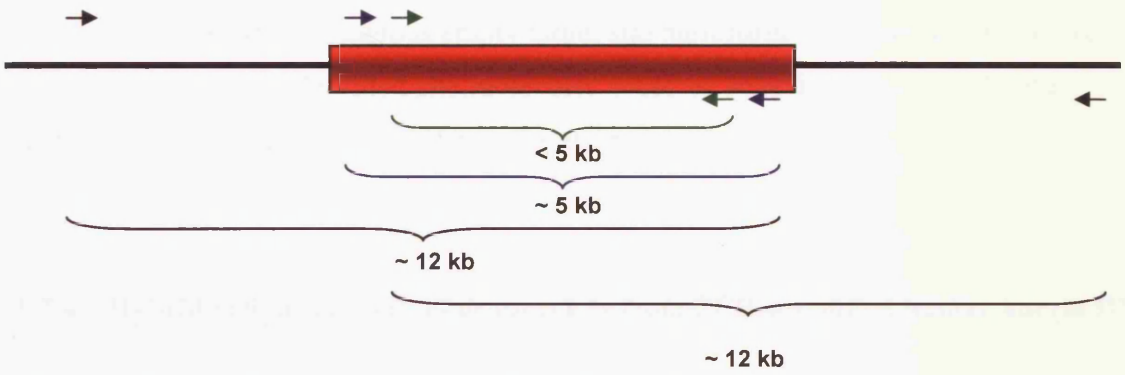


Figure 1.9: A representation of a target site. The target site is 6 kb in length (Red area). Within the target site is a set of primary target site primers (blue) giving an ~ 5 kb amplicon, and a set of nested secondary target site primers (green). ~ 9 kb either side of the target site is a control primer (red) which when amplified with an opposing target site primer will give an amplicon of approximately 12 kb.

At the most optimistic estimated rate of L1 insertion (Kazazian, 1999) approximately 1 in 8 humans have a *de novo* insertion somewhere within their genome. This equates to a *de novo* L1 insertion per 16 haploid genomes. Given that the human haploid genome size is approximately 3 pg in mass, and consists of approximately 3 x 10^9 bp, it was expected that there would be approximately 1 *de novo* L1 insertion in 0.024 ng of sperm DNA. However at the most pessimistic estimated rate of 1 in 120 humans harbouring a *de novo* L1 insertion (Li *et al.*, 2001), it would be expected that a *de novo* insertion would occur in approximately 0.72 ng of sperm DNA. The major drawback was that a single insertion could occur anywhere within a single sperm genome. Given that selected target sites were approximately 5 kb in length, and was around 1 / 600,000th the size of a haploid genome, this would equate to 1 insertion into a single 5 kb target site in 14.4 µg to 432 µg of sperm DNA. Given that the average human ejaculate yields around 600 µg of DNA, between 1 and 42 L1s per ejaculate would be expected to have

inserted into a single 5 kb target site. High quality high molecular weight sperm DNA is time consuming to extract from donated semen. Thus by amplifying multiple targets in a single PCR reaction, the overall target area for L1 insertion increased. For example, by simultaneously amplifying ten targets, the recovery of insertions should theoretically have increased to approximately 10 to 420 insertions per ejaculate.

In principle efficient multiplex amplification of the target sites will generate thousands of filled site molecules (molecules containing *de novo* L1 insertions), originating from a *de novo* insertion into a single empty target site molecule. However, the filled site molecules will be vastly outnumbered by numerous empty target site molecules. The approach therefore required a mechanism by which the filled site molecules could be selectively recovered, and enriched to a level where they could be detected and characterized.

1.7.4 Hybridisation recovery of *de novo* L1s from PCR-amplified human sperm DNA

Hybridisation enrichment was the tool chosen for recovery of target sites containing an L1 insertion. The allele-specific fractionation of sequences from pools of genomic DNA has, until recently, represented a major challenge in molecular genetics. DNA enrichment by allele-specific hybridisation (DEASH) (Jeffreys and May, 2003) is a technique for fractionation of rare variants from such pools of genomic DNA. Biotinylated oligonucleotides (bio-oligos) can be used to enrich specific DNA sequences that are under-represented in an aliquot of DNA. Enrichment is achieved by the binding of a bio-oligo to heat-denatured single-stranded selected site DNA sequences, and physical retrieval of the bio-oligo and the bound single-stranded DNA using streptavidin-coated paramagnetic beads. Successive rounds of PCR and enrichment can be used to increase yield (Jeffreys and May, 2003).

In order to identify *de novo* L1 insertions in sperm DNA, a modified DEASH-based method (hybridisation enrichment) was used. Figure 1.10 outlines the hybridisation enrichment methodology. Hybridisation enrichment required bio-oligos directed against the most conserved sequences within the 3′ 1.5 kb of young L1 subfamilies. As L1 is reverse transcribed from its 3′ end, restricting the bio-oligo sites to the 3′ 1.5 kb of L1 elements minimised the potential loss of insertions due to 5′ truncation. The use of the bio-oligos directly on pools of sperm DNA

would be futile due to the vast amount of L1 sequence dispersed throughout the genome. In order to detect *de novo* insertions in sperm DNA, the search area had to be limited to small genomic sections (~ 5 kb) devoid of young L1 subfamily sequence. This was achieved by PCR amplification. It is important to note that the PCR amplification of *de novo* insertions had to be efficient at the single-molecule level. A *de novo* insertion will appear once in a pool of sperm DNA, and had to be amplifiable by PCR even in the presence of an overwhelming majority of empty target sites, which being smaller, inevitably amplify more efficiently.

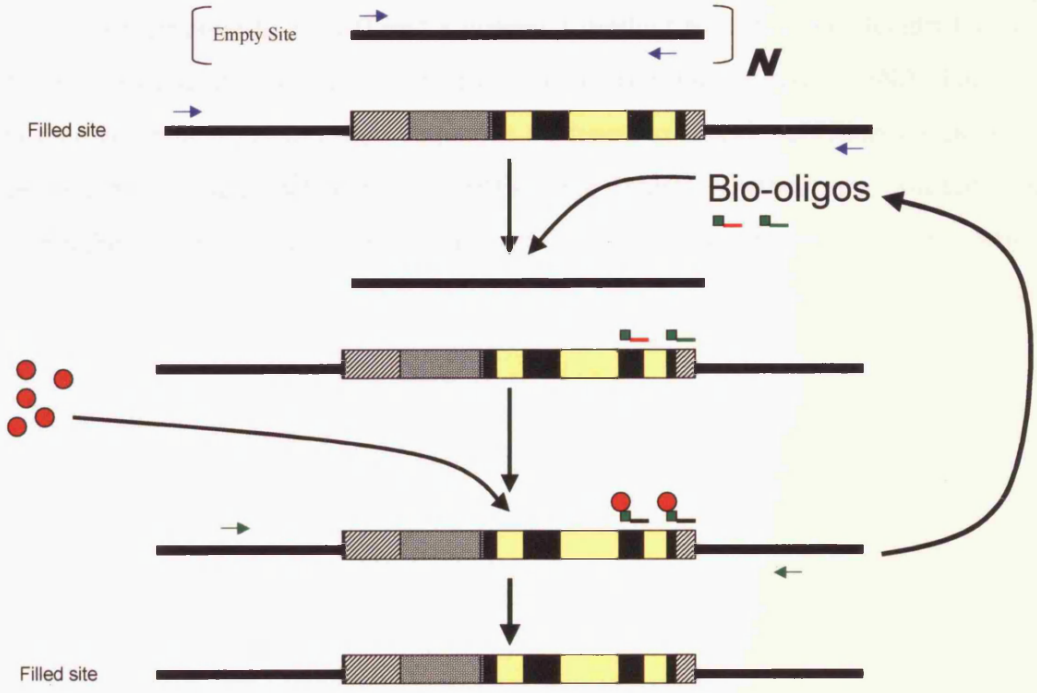


Figure 1.10: The principle of hybridisation enrichment, adapted from Jeffreys and May 2003 (figure 1, pg 2317). In an aliquot of sperm DNA, there will be millions of empty target sites and upon *de novo* insertion, one filled site. Amplification from the primary target site primers (blue) will increase the copy number of both the filled and empty site. The primary PCR products are denatured, and the bio-oligos are put into the mix and will specifically bind the L1 and not the empty site. The magnet to find the needle in the genomic haystack, streptavidin-coated paramagnetic beads specifically associate with the bio-oligos and can be recovered, hence enriching the filled site. Single-stranded filled site DNA enriched for by hybridisation is recovered from the bead/bio-oligo complex by thermal elution. Subsequent rounds of re-amplification using the nested secondary target site primers (green) followed by re-enrichment can potentially bring the filled site to totality prior to sequencing. Refer to figure 1.9 for target site design.

In order to ensure the method was capable of amplifying a single full-length L1 to recoverable levels, a target designed around a known polymorphic L1 was used. Dilution experiments were performed in which DNA containing a single-molecule of a known L1 insertion was diluted into 48 μg of DNA from an individual homozygous for the corresponding empty site. As a *de novo* insertion is in essence a single filled site molecule, this experiment closely replicated the recovery of a *de novo* insertion from a pool of DNA. This system allowed the method to be tested whilst remaining isolated from the target loci, thus preventing cross contamination between the experimental and control systems. Although hybridisation enrichment is limited to small genomic loci it offered a potential method by which full-length L1 insertions along with any transduced sequence could be recovered from human sperm DNA. The recovered L1s could then be sequenced and their sequence analysed to reveal: sub-family type; target site preference; insertion length; whether the ORFs were intact; whether any unusual structures existed within the L1 for example 5′ or 3′ transductions, internal inversions and / or deletions.

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: ABgene (Epsom, UK), Amersham Biosciences (Little Chalfont, UK), Applied Biosystems (Warrington, UK), Bio-Rad (Hemel Hempstead, UK), Boehringer, Mannheim, Cecil Instruments (Cambridge, UK), Clare Chemical Research (Delores, USA), Clontech (Palo Alto, USA), Eppendorf Scientific (Hamburg, Germany), Fisher Scientific (Loughborough, UK), Fisons (Beverley, USA), Flogen (Ashby-de-la-Zouch, UK), FMC Bioproducts (Rockland, USA), Hybaid (Teddington, UK), Invitrogen (Paisley, UK), MJ research (Waltham, USA), New England Biolabs (Hitchen, UK), Nalge Nunc International (Hereford, UK), New Brunswick Scientific Co. (New Jersey, USA), Perkin Elmer (Cambridge, UK), Qiagen LTD (Crawley, UK), Serva, Sigma Aldrich (Poole, UK), Starlab (Milton Keynes, UK) Syngene, Thermo Shandon (Pittsburgh, USA) and UVP Life Sciences (Cambridge, UK).

2.1.2 Oligonucleotides

Unmodified oligonucleotides were supplied by Protein and Nucleic Acids Chemistry Laboratory (University of Leicester) until 2004. Subsequently unmodified oligonucleotides were supplied by Sigma-Genosys (Haverhill, UK). HPLC purified oligonucleotides were supplied by Thermo Electron Corporation (Ulm, Germany). Biotinylated oligonucleotides (bio-oligos) were supplied by Eurogentec S.A (Seraing, Belgium).

2.1.3 Enzymes

Restriction enzymes *Sna*BI, *Xho*I *and Bss*SI were supplied by New England Biolabs. Shrimp Alkaline Phosphatase (SAP) was supplied by Roche (Lewes, UK). Optikinase was supplied by USB (Staufen, Germany). *Pfu* DNA polymerase was supplied by Stratagene. *Taq*

DNA polymerase was supplied by ABgene (Epsom, UK). Proteinase K was supplied by Sigma Aldrich, (Poole, UK).

2.1.4 Molecular weight markers

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

2.1.5 Standard solutions

A list and description of all solutions used during this investigation and mentioned in this chapter can be found in appendix i. All bench solutions were prepared using dH_2O, but PCR clean solutions were prepared using ultra pure 18 MΩ ddH_2O

2.1.6 Computers and software

This thesis was written on a PC using Windows XP. Word-processing was performed in Microsoft Office Word 2003, and referenced with Endnote 9. Data were presented and analysed using the following packages: Microsoft Office Powerpoint, Microsoft Office Excel, Photoshop Elements V2 and GCG. The Poisson confidence (Siméon-Denis Poisson, 1781 - 1840) interval program was written by A.J. Jeffreys in True Basic (True Basic INC, Hartford, USA).

2.2 Methods

2.2.1 Agarose gel electrophoresis

All PCR products were fractionated on low electroendosmosis (LE) agarose (Cambrex) gels. Fragments sized at <1 kb were fractionated on 2 % (w/v) gels, 1 kb to 2 kb on 1 % (w/v) gels and > 2 kb on 0.8 % (w/v) gels. All gels were made using, and run in, 0.5 x Tris–Borate

EDTA (TBE) buffer containing 0.5 μg/ml Ethidium Bromide (EtBr). The voltage at which gels were run and the running time were dependent on the sample being loaded. For example, samples containing fragments approximately 10 kb in length were run at around 100 to 150 Volts (V) for approximately 3 to 5 hours in a 0.8 % gel (w/v). Samples < 1 kb in length were run at 150 to 200 V for 1 to 2 hours in a 2 % (w/v) gel. Samples that contained several fragments approximately the same length (for example multiplex PCRs) were run at the highest suitable voltage to ensure the gel ran quickly and keep the individual bands tightly compacted. Ethidium bromide is used as a nucleic acid stain. It intercalates into the nucleic acid structure, and fluoresces with a red/orange colour when exposed to UV light at 260 nm making it possible to visualise bands on an agarose gel. Samples were loaded onto gels with Tris-Borate EDTA (TBE) loading buffer. Samples were loaded using 2.5 to 3 x loading buffer. Fragment size was confirmed by running against an appropriate DNA molecular weight marker, for example 1 kb molecular weight marker (NEB), 100 bp molecular weight marker (NEB) or *Hind*III-digested λ DNA.

Electrophoresis tanks were manufactured by the University of Leicester workshop, and the powerpacks supplied by Bio-Rad. DNA visualisation was performed using either a hand held UV wand (Chromato-vue UVM-57, UVP Life Sciences), or on a Dark Reader Transilluminator (Clare Chemical Research). This uses visible light between 400 and 500nm and avoids DNA degradation by UV. Gels were photographed in a darkroom cabinet using a CCD camera (Syngene). The photographs were subsequently analysed and presented using Gene Snap software (Syngene) on a PC and printed on photographic paper using a Sony digital graphic printer (Syngene).

2.2.2 Standard DNA digests using restriction enzymes

Genomic DNA concentration was estimated by UV spectrophotometry (see below). The protocol for estimating the concentration of PCR-amplified DNA can be found in section 2.2.5.e.

Once the desired mass of DNA was calculated, two aliquots of the following digestion mix were prepared: 1 x reaction buffer (supplied with the enzyme), 5 units of enzyme per μg of DNA being digested and the sample DNA. The reaction solution was brought to a final volume of 20 μl with water. Each aliquot was then incubated for 30 minutes and 60 minutes respectively

at 37 °C, where the 60 minute sample serves as an over-digest control. Running the equivalent volume of both the sample and over-digest control on a 0.8 % (w/v) LE agarose gel was used to determine whether complete digestion of the sample had occurred.

2.2.3 DNA preparations

2.2.3.a Standard practice

Unless otherwise stated, all solutions required for the preparation of genomic DNA (gDNA) were made in a vented class II containment hood which had been UV irradiated for 15 min to reduce the risk of contaminating with exogenous DNA molecules. The reagents were UV irradiated for a further 15 min, in the hood, prior to use to ensure they were clean and sterile. Extraction of gDNA was also carried out in the hood.

All pipettes were washed before and after use in 1 % (w/v) Virkon (Antec Intrenational), and rinsed with water. All liquids and plastic ware were discarded into 1 % (w/v) Virkon and left for at least 1 hour prior to disposal through the laboratory clinical waste procedure. Phenol waste was disposed of through the laboratory hazardous waste procedure. After DNA extraction, all pipettes and surfaces were swabbed with 1 % (w/v) Virkon and then with Industrial Methylated Spirits (IMS). Finally the hood was sterilised by being UV irradiated for 30 min.

2.2.3.b Human DNA collection

Sperm, blood and buccal swab DNAs were prepared from samples collected within the Department of Genetics, University of Leicester. Donors were approached to contribute on a voluntary basis. Donors were required to sign a donor consent form which complied with the appropriate ethical approval of the university. Volunteers' samples were immediately given coded identities. The form also enabled the donors to give consent for the collection of saliva samples.

2.2.3.c Preparation of single-molecule clean (SMC) human sperm DNA

In order to reduce the potential risk of contamination of sperm DNA with partner's DNA, donors were asked to wash their penis prior to ejaculation. Ejaculates were collected in 5 ml of UV irradiated 1 x Salt-Sodium Citrate solution (SSC) in a 50 ml screw cap centrifuge tube, and stored at -80 °C.

Fresh 50 µl stocks of 20 mg/ml proteinase K (Sigma) were prepared in water for each extraction and kept on ice. Fresh aliquots of 2-mercaptoethanol were also taken for each extraction and kept in the hood. These solutions were not UV irradiated.

2.2.3.d DNA preparation

The semen sample was thawed as rapidly as possible at room temperature and mixed by shaking. For samples known to have a relatively high sperm count (> 20 million sperm per ml), 0.2 ml of semen was added to 1 ml of 1 x SSC in a 1.5 ml eppendorf tube. 0.5 ml of semen was used for samples with a lower sperm count (< 20 million sperm per ml). Any remaining semen was immediately refrozen and stored at -80 °C. The tubes were centrifuged in an eppendorf 5415D centrifuge at 15700 x g for 1 minute, rotated through 180°, and re-centrifuged for 1 minute at the same speed.

After centrifugation and removal of the supernatant, the pellets were re-suspended in 0.8 ml 1 x SSC with vigorous shaking. A Sodium Dodecyl Sulphate (SDS) pre-lysis step followed in order to lyse any non-sperm cells, thus removing any contaminating cells and partner's cells. 16 µl 10 % (w/v) SDS (0.2 % (w/v) SDS final concentration) was added to the solution and mixed for 30 sec prior to centrifugation at 15700 x g for 1 minute, rotated through 180° and re-centrifuged for a further minute. The pellets were re-suspended in 0.8 ml 1 x SSC plus 16 µl 10 % (w/v) SDS (0.2 % (w/v) SDS final concentration.) with vigorous shaking and re-centrifuged for 1 min, rotated through 180° and centrifuged for a further minute.

After removing the supernatant, the pellets were re-suspended in 0.45 ml 0.2 x SSC prior to the addition of 50 µl 10 % (w/v) SDS (1 % (w/v) SDS final concentration), 35 µl 2-mercaptoethanol (1 M final concentration) and 4 µl 20 mg/ml proteinase K. This was followed by incubation at 37 °C for 45 - 60 min with gentle shaking.

After incubation, 0.35 ml of liquid phenol (general purpose grade, Fisher Scientific, stored at -20 °C) was added. The tubes were shaken gently until a stable emulsion formed, and were centrifuged at 15700 x g for 4 min. Yellow 200µl pipette tips with the end 3 mm cut off were used to transfer the viscous DNA-containing aqueous phase to a second eppendorf tube containing 0.35 ml phenol. This secondary phenol extraction ensures thorough removal of any protein contaminants. These second tubes were again shaken until a stable emulsion formed, and re-centrifuged at 15700 x g for 4 min. At this stage the supernatants were transferred to clean 1.5 ml eppendorf tubes. (For samples which contained high quantities of viscous DNA, a further phenol extraction was carried out in order to ensure the removal of all contaminants).

To the first set of tubes, 0.3 ml 1 x SSC, 1 % (w/v) SDS was added to maximise the recovery of any remaining DNA. This was mixed into a stable emulsion and centrifuged at 15700 x g for 4 min. The aqueous phase was transferred to the second tube, mixed and re-centrifuged at 15700 x g for 4 min. Finally, the supernatant was pooled in the clean 1.5 ml eppendorf tubes with the initial supernatants.

To precipitate the DNA, 2 volumes of 100 % ethanol were added to the pooled aqueous phases and mixed gently by inverting. The tubes were centrifuged at 15700 x g for 1 minute to pellet the DNA. The supernatants were discarded, and the pellets washed by gently inverting with 1 ml 80 % (v/v) ethanol. The tubes were centrifuged at 15700 x g for 1 minute and the supernatants discarded. The pelleted DNA was dissolved in 0.5 ml water. 50 µl of 2 M sodium acetate (NaAc) pH 8.0 and 2 volumes 100 % ethanol were added and mixed gently by inverting. The tubes were centrifuged at 15700 x g for 1 minute, and the DNA pellets washed with 1 ml 80 % (v/v) ethanol by inverting gently. After centrifugation for 1 minute at 15700 x g, all the liquid was removed and the pellets vacuum dried in a speed vac concentrator (Savant). The pellets were dissolved in 25 – 50 µl of 5mM Tris pH 7.5 overnight in a 50 °C water bath. The following day the samples were pooled and again placed at 50 °C for 3 – 4 hours to ensure the DNA was evenly dispersed.

DNA aliquots were stored in a clean box at 4 °C. DNA concentration was assessed by UV spectrophotometry using an eppendorf BioPhotometer. The DNA was diluted into three aliquots at a dilution of 1 in 100, three aliquots at 1 in 200 and three aliquots at 1 in 400 and the estimated concentrations were averaged. The OD 260:280 ratio was also recorded and averaged to assess the purity of the DNA in terms of protein content. 20ng of the DNA was then run on a 0.8 %

(w/v) LE agarose gel versus 20ng of λ DNA (NEB: 500 μg/ml) to ensure that the DNA was of high quality and free from degradation. Degraded samples gave a smear rather than a tight band when visualised by agarose gel electrophoresis.

2.2.3.e Extraction of single-molecule clean (SMC) Human DNA from blood samples

Venous blood samples (approximately 7 ml) were collected in an anti-coagulating agent, 1.2 ml of 15 % K3 EDTA (BD Vacutainer Systems, Plymouth, UK), and stored at -80 °C. None of the samples used during the course of this investigation were from known infected sources, but all samples were handled according to the laboratory safety procedure, and disposed of correctly through the laboratory clinical waste procedure (Category II level).

Fresh 50 μl stocks of 20 mg/ml proteinase K (Sigma) were prepared in water for each extraction and kept on ice. These solutions were not UV irradiated.

2.2.3.f DNA Preparation

Blood samples were thawed at room temperature, and 500 μl aliquots were placed into screw-top 1.5ml eppendorf tubes. Blue 1 ml pipette tips used for aliquoting had the end 3 mm cut off to allow the blood to be pipetted more easily. Any remaining blood was immediately re-frozen, and stored at – 80 °C To each aliquot, 800 μl 1x SSC was added, and the tubes centrifuged at 15700 x g in an eppendorf 5415D centrifuge for 2 min. The supernatants were discarded into 1 % (w/v) Virkon, and the samples re-suspended in 1ml 1x SSC. The tubes were again centrifuged for 2 min at 15700 x g, and the supernatants discarded into Virkon.

The pellets were re-suspended in 300 μl 0.2 x SSC and to lyse the cells, 30 μl of 10 % (w/v) SDS added. 3 μl 20 mg/ml proteinase K was also added prior to an incubation at 37 °C for 30 min. After incubation, 200 μl 25:24:1 phenol/chloroform/iso-amyl-alcohol was added, and the tubes shaken to form a stable emulsion. The tubes were then centrifuged at 15700 x g for 5 min. The aqueous phase of each tube was removed into a 2ml Phase Lock tube (eppendorf). To the original tubes, 100 μl 1x SSC / 1 % (w/v) SDS was added, and the tubes shaken to form a stable emulsion, and centrifuged at 15700 x g for 5 min. The aqueous layer was pooled with the previous aqueous layer in the Phase Lock tubes. 400 μl 25:24:1

phenol/chloroform/iso-amyl-alcohol was added, and the tube shaken to form a stable emulsion. The tubes were then centrifuged at 15000 x g for 5 min, followed by the addition of 400 µl chloroform and a further shake and spin at 15000 x g. The aqueous layers were subsequently transferred to fresh 1.5ml eppendorf tubes.

Two volumes of 100 % ethanol were added and mixed by gently inverting. The tubes were centrifuged for 30 sec to pellet the DNA prior to the removal of the ethanol. The pellets were washed in 1 ml 80 % (v/v) ethanol by gentle inversion, and centrifuged at 15700 x g for a further 30 sec to ensure the DNA remained in the bottom of the tubes. After removal of the ethanol, 90 µl of 18MΩ water was added to dissolve the DNA. Once the DNA was completely in solution, 10 µl 2M NaAc pH8 and 2 volumes of 100 % ethanol were added, and inverted gently to mix. The tubes were centrifuged at 15700 x g for 30 sec to pellet the DNA and the ethanol removed. Another 80 % (v/v) ethanol wash was carried out and after centrifugation the tubes briefly at 15700 x g, the ethanol was removed. The pellets were vacuum dried in a speed vac concentrator (Savant) and disolved in 20 µl 5 mM Tris-HCl (pH7.5). DNA was stored in a clean box at -20 °C, and working stocks stored at 4 °C. DNA concentration was assessed by use of a Cecil Instruments CE2040 Ultra Violet spectrophotometer. The DNA was diluted to 3 x 1 in 50, 3 x 1 in 100, and 3 x 1 in 200, and the optical density at 260 nm measured (OD_{260}). The following equation was used to determine the DNA concentration:

$$\text{DNA concentration } (\mu g/ml) = (\text{dilution factor} \times A_{260}) \times 50$$

The measured concentrations were averaged. 20 ng of the DNA was then run on a 0.8 % (w/v) LE agarose gel against 20 ng of λ DNA to ensure that the DNA was of high quality and free from degradation.

2.2.3.g Extraction of DNA from buccal swabs

Buccal samples were collected by moving a cheek swab around the inside of both cheeks for 30 sec. The swabs were supplied with Buccal DNA Isolation kits (Isohelix), and the DNA extracted from the swabs using the standard protocol outlined in the Buccal DNA Isolation kit manual (Isohelix).

2.2.3.h Phenol Chloroform purification of PCR products

The following protocol was carried out in a fume hood to prevent the escape of phenol and chloroform fumes. Phenol waste was disposed of through the laboratory hazardous waste procedure. The hood was cleaned before each procedure by wiping down with depurinating solution (0.25 M HCl) and IMS. 50 µl of 20 mg/ml fresh proteinase K was prepared prior to each clean up with water and kept on ice. The solutions required for the procedure were also prepared prior to purification and sterilised by being placed on a UV trans-illuminator for 10 min. All pipettes used for the procedure were also UV irradiated for 10 min.

PCRs were pooled as required. 1 µl 20 x SSC, 10 µl 10 % (w/v) SDS and 1 µl 20 mg/ml fresh proteinase K were added per 100 µl of the total volume of the PCR requiring clean up. This was then incubated at 37 °C for 30 min with occasional mixing

Approximately two volumes of Phenol/chloroform/isoamyl alcohol (25:24:1) were added to the mix, and shaken gently until a stable emulsion formed. The tubes were then centrifuged at 15700 x g in an eppendorf 5415D centrifuge, in a 1.5 ml microcentrifuge tube if the total mix was less than 1.5ml, or at 2400 x g in an eppendorf 5804 centrifuge for larger volumes in 15 ml centrifuge tubes. Centrifugation was carried out for 5 min. The aqueous layer was removed into a Phase Lock tube (eppendorf) of an appropriate size. To the original tubes which had contained solutions of < 0.5 ml (in a 1.5 ml eppendorf tubes) 0.5 µl 20 x SSC, 10 µl 10 %SDS and 89.5 µl H_2O was added. To tubes which had contained larger volumes (in 15 ml eppendorf tubes) 2 µl 20 x SSC, 20 µl 10 % (w/v) SDS and 179 µl H_2O were added. The original tubes were then shaken to form a stable emulsion, and centrifuged for 5 min at either 15700 or 2400 x g. The aqueous phases were pooled in the corresponding Phase Lock tube. Approximately double the original PCR volume of chloroform was added to the Phase Lock tubes. The tubes were then shaken to form a stable emulsion and centrifuged at either 15000 or 1500 x g for 5 min. The upper layer was then removed to a clean collection tube.

Two volumes of ice cold 100 % ethanol were added and mixed gently by inverting. The tubes were then put into a -20 °C freezer for at least 60 min. The tubes were then centrifuged at either 15700 x g in an eppendorf 5415D centrifuge or at 2400x g in an eppendorf 5804, depending on the size of the tube, for 25 min. The ethanol was then gently removed, and the

pellets washed in 80 % (v/v) ethanol by gentle inversion. The tubes were then centrifuged for 5 min at either 15700 or 2400 x g. The ethanol was removed, and the pellets were disolved in 500 µl water. 50 µl 2M NaAc (pH7) were added, followed by addition of 1 ml of ice cold 100 % ethanol. The samples were mixed by inverting gently, and the tubes were put at -20 °C for at least 1 hour prior to being centrifuged for 25 min at either 15700 or 2400 x g. The ethanol was removed and the pellets washed in 1ml 80 % (v/v) ethanol. The tubes were centrifuged for 5 min, and the ethanol removed. The smaller tubes were vacuum dried in a speed vac concentrator (Savant), but the larger tubes were dried in a hybridisation oven at 50 °C with a loose covering of saran wrap to protect them. Finally, the dried pellets were disolved in the desired volume of 5 mM Tris-HCl (pH 7.5).

2.2.4 PCR based Methods

A table of additional PCR primers (not listed in this chapter) can be found in the Appendix iii.

2.2.4.a Primer stock and working strength dilutions

All dilutions were carried out in a category II PCR hood which had been UV irradiated for 20 min to prevent contamination of the primers. PCR primers were diluted in UV-irradiated 5 mM Tris-HCl (pH 7.5) to a stock concentration of 100 µM. Working strength dilutions were also made to a concentration of 10 µM.

PCR primers were then categorised as PCR clean or Single-molecule Clean (SMC). PCR clean primers were only ever opened in a PCR hood, and fresh gloves would be worn prior to handling them. SMC gloves had an extra level of cleanliness in that they were only ever opened in a UV irradiated category II PCR hood, and fresh gloves were put on every time the hood was left and re-entered. This minimised the risk of contamination from the environment.

Primers opened outside the hood at any point were marked as non-PCR clean. These primers were no longer allowed to enter the PCR hood.

All primer stocks were kept in separate boxes depending on their "status" i.e. PCR clean, SMC, non-PCR clean.

2.2.4.b Standard PCR conditions

The Polymerase Chain Reaction (PCR) is a technique used for *in vitro* amplification of specific DNA sequences by the simultaneous primer extension of complementary strands of DNA. All PCR mixes were made up in a category II PCR hood to ensure all reagents were kept PCR clean. Extra precautions such as UV irradiating the hood for 20 min prior to use, and putting on fresh clean gloves before entering and re-entering the hood kept the required primers single-molecule clean (SMC). For PCR amplifications less than 2 kb in length standard PCR conditions were used. A standard PCR mix of 20 µl contained 0.4 U/µl of *Taq* DNA polymerase , 1.1 x PCR buffer, 0.5 µM primer concentration and 50 ng of human gDNA.

The PCRs were cycled in an MJ tetrad PCR machine PCT250. The following cycle was applied: 96 °C for 2 min, (96 °C for 30 sec; optimal annealing temperature (TM °C) for 30 sec, 72 °C for 2 min) for 30 cycles, TM °C for 10 min.

2.2.4.c Nested PCR

PCR amplification from nested primers allowed the amplification of specific molecules, contained within a PCR-amplified sample, by priming internally to the primary PCR primers. This allowed confirmation of the identity of PCR-amplified molecules. Nesting also prevents DNA produced in a secondary PCR from becoming contaminants to primary PCRs, as the secondary amplicons lack the primer sequences required for primer extension. This allowed rare molecules to be amplified to levels that could be visualised on a gel, while isolating them from the initial primary PCRs.

2.2.4.d Standard long-range PCR conditions

Long-range PCR was used for amplifications > 2 kb in length. Long-range PCR was optimised using a mixture of *Taq* and *Pfu* polymerases. *Pfu* was added as it has a proofreading 3′ to 5′ exonuclease activity, and removes base mismatches which stall *Taq* DNA polymerase activity. Initially four sets of target site primers (from the β globin (HBB), Duchenne Muscular

Dystrophy (DMD), Factor IX (FIX) & Major histocompatability complex class II (MH2) loci) were amplified as follows: 0.04 units per µl of a 20:1 mix of *Taq/Pfu*, 0.5 µM of each primer and 1.1 x PCR buffer (Jeffreys *et al.*, 1988) were used in the reaction mix. Water was added to a final volume of 20 µl. Cycling was optimised on the longest control amplicon of 13 kb. Two-step PCR was performed: 96 °C for 1 min; (96 °C for 20 sec; 62 °C for 14 min) for 30 cycles (note: 1 min per kb plus 1 minute); 61 °C for 30 min then held at 15 °C. Amplification was carried out using 50 ng of high quality human genomic DNA. Cycling was performed in an MJ tetrad PCR machine PCT250.

2.2.4.d.i Amendments

Following the optimisation of the multiplex PCR protocol (see chapter 5), the standard long-range PCR conditions were amended. The final conditions were amended to: 0.025 units per µl of a 20:1 mix of *Taq/Pfu*, 0.05 µM of each primer. The concentration of PCR buffer and the cycling conditions were not altered.

2.2.4.e Determining the number of amplifiable molecules in a sample by PCR

Samples analysed by this method were usually high molecular weight human genomic DNA (approximately 10 to 100 ng/µl in concentration), or multiplex PCR-amplified samples. The DNA sample being analysed was serially diluted by a factor of one in ten in single-molecule diluent (SMD). Dilution was carried out until a chosen target was present at an estimated 1 molecule per µl. For example 0.003 ng of gDNA is approximately equal to 1 haploid genome mass; therefore at 0.003 ng/µl one would expect 1 molecule of a given genomic locus per µl of the sample.

2.2.4.e.i Primary PCR

Primary PCRs were carried out in 96-well PCR plates (ABgene). Figure 2.1 outlines a typical primary PCR. Each well in a single row contained equivalent samples. In most cases the primary PCR took the form of a 12 kb long-range PCR. The amended standard long-range PCR protocol was followed, and cycled 30 times. A single amplification did not generate enough

PCR-amplified DNA to allow visualisation on an agarose gel, therefore a secondary PCR was required.

2.2.4.e.ii Secondary PCR

After the primary amplification, the samples were diluted by a factor of one in ten, in 5 mM Tris-HCl (pH 7.5) and 2 µl of the diluted sample used to seed a secondary PCR. This PCR gave a binary result. If the well from which the primary PCR was taken had contained an amplifiable molecule, amplification would occur in the secondary PCR. If no amplifiable molecule was present in the primary PCR the secondary PCR would not produce any amplified product. The secondary PCR generally took the form of a 3 kb long-range PCR using nested primers. Amended long-range PCR conditions were used, and 35 cycles of PCR performed. Following PCR, the samples were run out in their groups of eight on a 0.8 % (w/v) LE agarose gel alongside 1 kb molecular weight marker. Samples could then be scored for the presence/absence of secondary PCR products. The data generated were assessed using the Poisson confidence interval program written by A.J. Jeffreys, which generated a maximum likelihood estimate plus 95 % confidence intervals for the number of molecules contained in a µl of the original query sample.

2.2.4.f Multiplex PCR

The primers used for multiplex PCR are outlined in tables 2.2 and 2.3 (these primers were also used in the optimisation of long-range PCR). Multiplex PCR involved the amplification of multiple pairs of primers simultaneously in the same PCR reaction. The maximum number of primer pairs amplified simultaneously in this investigation was eleven. This took the form of all ten primary target site primer sets, with amplification product sizes of 4 to 5 kb in length, and a pair of control primers generating amplification sizes of 12 to 13 kb in length.

All multiplex PCR mixes were made in a category II PCR hood which had been UV irradiated for 20 min prior to use. Clean gloves were put on before entering and re-entering the hood. These precautions ensured all the primers were kept free from contamination and therefore were kept single-molecule clean.

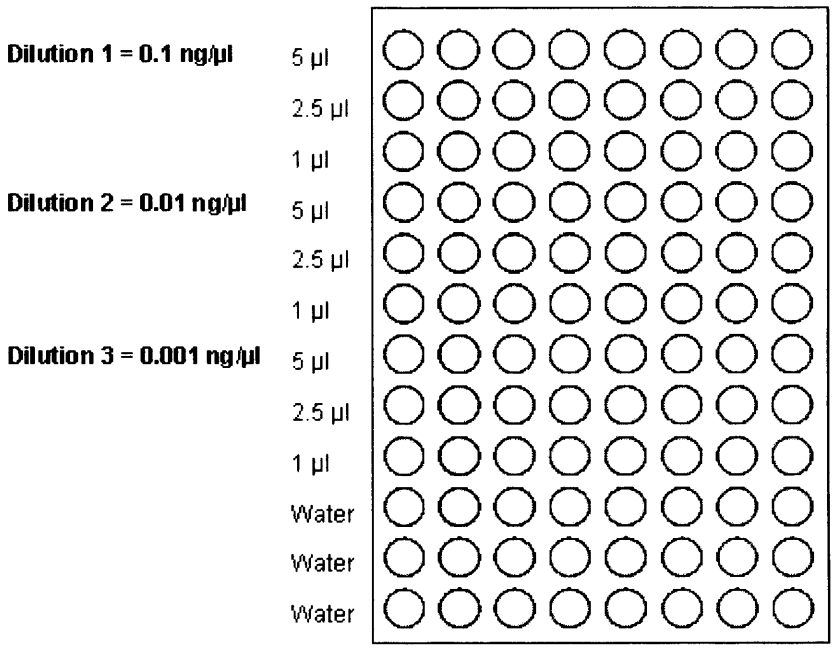


Figure 2.1: A diagrammatical representation of a 96-well plate used for primary PCR amplification. The dilution series can be seen to the left of the plate, and the volume of each dilution placed into a row of eight wells is also outlined.

Multiplex PCR using the primary target site primers required 1.1 x PCR buffer, 0.025 U/µl *Taq*/*Pfu* 20:1. The concentrations of the primary Target site primers, used in multiplex PCR, were adjusted to give equal amounts of PCR product when viewed under UV light on a 0.8 % (w/v) LE agarose gel (see table). Additional primers were added to this standard mix as required (al121819 locus control primers and Long-range control primers).

500 ng of high quality SMC sperm or blood DNA was used for amplification in 50 µl reactions. The cycle conditions were 96 °C for 1 minute, (96 °C for 20 sec; 62 °C for 14 min) for 20 cycles, 61 °C for 30 min. Cycling was performed in an MJ tetrad PCR machine PCT250.

2.2.5 Southern blotting and Dotblotting

2.2.5.a Standard Southern blotting protocol (Southern, 1974)

Once an agarose gel had been run, it was occasionally necessary to transfer the DNA from the gel onto a Nylon transfer membrane (Magna Nylon transfer membrane (GRI)) to allow screening using radiolabelled probes. Initially, all imperfections on the gel surface, for example raised wells, were removed using a clean scalpel blade. Blots were either hybridised with 5′ end labelled oligonucleotide probes, or with random prime labelled probes. If the blot was to be labelled using end-labelled probes, the gel was transferred to a Dark Reader prior to blotting, and the positions of the molecular weight marker bands highlighted by cutting slits in the gel. Following blotting, the position of the molecular weight marker bands was transferred onto the nylon transfer membrane using a pencil. This step was not required for internally labelled probe hybridisations as molecular weight marker DNA was added to labelling mixes to highlight the location of the marker bands.

Following trimming, the gel was transferred into a tray containing depurinating solution (0.25 M HCl) for 7.5 min with gentle shaking. Depurinating solution was then replaced with denaturing solution for 20 min with gentle shaking. Denaturing solution was replaced with neutralising solution for 20 min with gentle shaking. Once the 20 min had elapsed, the gel was rinsed in distilled water and placed onto Southern blotting apparatus. The Southern blotting apparatus consisted of tray half-filled with 20 x 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 x SSC. Ensuring that the wick was soaked with 20 x 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 (20 x SSC) ran through the gel. A sheet of nylon transfer membrane, which had been soaked for 5 min in 3 x SSC, was laid on top of the gel and the air bubbles smoothed out. Two sheets of Whatman 3MM paper, pre soaked in 3 x SSC, were laid over the nylon transfer membrane and the air bubbles were removed.

Primer ID	Sequence	Concentration (μM)	Accession Number	Location	Amplicon (bp)
PFHBBTF1	ATTTGCCTGGTATGCCTGGG	0.07	ac104389 (complement)	175485	
PFHBBTR1	CCTTGAAGCCAGGATGATGG	0.07	ac104389 (complement)	180490	5005
PFDMDTF1	GTGCTTTAGACATTACCCAGG	0.12	al031643 (complement)	111421	
PFDMDTR1	CAACACATCTCTTCATACAGAGGT	0.12	al031643 (complement)	116329	4908
PFFIXTF1	CACTGAGACCCCCTTCGG	0.08	al033433 (complement)	17106	
PFFIXTR1	CTGGTATAGTGCTGAGACAGG	0.08	al033433 (complement)	22657	5551
PFMH2TF1	TGCAGGGGCAGAAGAGGG	0.15	al662845 (complement)	32532	
PFMH2TR1	CCCCCTTGTCTTGCTAAGGG	0.15	al662845 (complement)	37549	5017
PFCHMTF1	TGAATGCTGGTTGTGGGAGG	0.05	al009175 (complement)	57913	
PFCHMTR1	ACCATCTATGGTGCTGCTGG	0.05	al009175 (complement)	63590	5677
PFAPCTF1	GGAAGCATTATGGGACATGG	0.025	ac008575	127059	
PFAPCTR1	CCTGAACAGACGAATGTGTGG	0.025	ac008575	132475	5416
PFCGDTF1	GTTGGCTAACCATCAGAGGG	0.07	al627245	38073	
PFCGDTR1	CAGCAAACTGAGGGATTGGG	0.07	al627245	43232	5159
PFRP2TF1	CACAGAAGAGGATTGGGAGG	0.03	al050307	86064	
PFRP2TR1	TCTGTCATGCCCAACCTCTG	0.03	al050307	90837	4773
PFFCMTF1	CCTTTCGGAAGAGTGCAAAGG	0.1	al158070	94766	
PFFCMTR1	TGGAGCTAAGGTTTCCCAGG	0.1	al158070	99844	5078
PFHODTF1	TTCGGTTCCTTCTCTGCTGG	0.1	ac009336	84862	
PFHODTR1	CTCCTAAAATGGCTCCCTGG	0.1	ac009336	89126	4264
PFLRCtrlA	TTAGGACGCCCACTACTGTG	0.1	ac008706	73112	
PFLRCtrlB	GCCAATCCTGTAAGGCAAAGGG	0.1	ac008706	85917	12805
PF819LRA	CCTGCTTTCACTTCACAGGG	0.05	al121819	105131	
PF819LRB	CAGGTAGAAACTTCCCAGGG	0.05	al121819	111585	6454

Table 2.2: Primers used for balanced primary target site primer multiplex PCR and the concentrations at which they were used. Multiplex PCR that did not require balancing used the primers at 0.05 μM. The table also gives the accession number of the sequence from which the primers were designed from and the location of the primer with respect to that particular accession. The amplicon length of each PCR is also shown. All primers were HPLC purified and supplied by Thermo Electron, and were pepared under single molecule clean conditions.

Primer ID	Sequence	Concentration (μM)	Accession Number	Location	Amplicon (bp)
PFHBBTF2	CCTTGAAGCCAGGATGATGG	0.05	ac104389 (complement)	175627	
PFHBBTR2	AGCCAGAAGCACCATAAGGG	0.05	ac104389 (complement)	180351	4724
PFDMDTF2	GTACCTCACAGCATAGAGGG	0.05	al031643 (complement)	111738	
PFDMDTR2	GTTCTGTGTTGTCAGACAGGG	0.05	al031643 (complement)	116167	4429
PFFIXTF2	TCTATGGAACTCTCCCCTGG	0.05	al033433 (complement)	17217	
PFFIXTR2	GCCATACGAACATGGAGTGG	0.05	al033433 (complement)	22443	5226
PFMH2TF2	TAGCAACTGACTCCATGAGG	0.05	al662845 (complement)	32604	
PFMH2TR2	GAAACCTGGATAGAGACGTGG	0.05	al662845 (complement)	37418	4814
PFCHMTF2	ACCCTGAAGGAGACTTCTGG	0.05	al009175 (complement)	57972	
PFCHMTR2	CCTCACAAAGGACATAGGTGG	0.05	al009175 (complement)	63376	5404
PFAP2TR2	GACCTAGTGGGAGAAGCTGG	0.05	ac008575	127170	
PFAPCTF2	TAGGCCTGCGAAGTACAAGG	0.05	ac008575	132404	5234
PFCGDTF2	TCCTCACCTCATGTGTGGAG	0.05	al627245	38145	
PFCGDTR2	CACGTACAATTCGTCTGGGTG	0.05	al627245	43130	4985
PFRP2TF2	GGTACAATTCTTGAGGGGGTG	0.05	al050307	86090	
PFRP2TR2	TCTGTCATGCCCAACCTCTG	0.05	al050307	90787	4697
PFFCMTF2	TGGCTGAGCAGTGGAACTTG	0.05	al158070	94837	
PFFCMTR2	TTGGACTGACAAACCCAGGG	0.05	al158070	99737	4900
PFHODTF2	CCTTAAGGCTTCCACGTTGG	0.05	ac009336	84995	
PFHODTR2	CGGGTCTTTATGTGTCTGGG	0.05	ac009336	89047	4052
PFLRCtrlC	GAGTCTAAAGAATCCTCCGGG	0.05	ac008706	73230	
PFLRCtrlD	TCAGGTAGTCAGTCACGTGG	0.05	ac008706	85868	12638
PFLRCtrlE	TAGCCCTGCTCTGTCTATGG	0.05	al121819	73268	
PFLRCtrlF	CCTTCAAAACTAGATGCCAGTG	0.05	al121819	85846	12578
PF819LRC	TGTTCTCACAGCCTGACAGG	0.05	al121819	105208	
PF819LRD	CTGGCTTGTAGGTACCAAGG	0.05	al121819	111536	6328

Table 2.3: Primers used for unbalanced secondary TSP multiplex PCRs. All primers were HPLC purified and supplied by Thermo Electron, and were pepared under single molecule clean conditions.

A stack of paper towels (approximately 30) were placed on top of the 3MM paper followed by a glass plate and a weight (approximately 1 kg) to compress the apparatus. All gels were blotted for a minimum of 5 hours.

After blotting and apparatus disassembly, the nylon transfer membrane was washed gently in 3 x SSC and dried in a 3MM paper envelope at 80 °C for approximately 15 min. The blot was then transferred to a CL 1000 ultraviolet crosslinker (UVP) and exposed to 70,000 $\mu J/cm^2$ of UV (254 nm). Blots were stored in 3MM paper and saran wrap, in the dark at room temperature.

2.2.5.b Dotblotting

Occasionally it was necessary to transfer PCR-amplified DNA to a nylon membrane without size fractionation. To carry out a dotblot, two 13 x 9.5 cm sheets of Whatman 3MM paper and one 12 x 8.5 cm sheet of nylon transfer membrane were cut. These were soaked in 3 x SSC. The 2 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 3MM paper. The manifold was then assembled and tightened.

For every kilobase of amplicon length, 3 to 100 ng of PCR product were required for blotting. One quarter the volume of dotblot loading mix (30 % v/v glycerol, 0.5 x TBE, 0.025 % bromophenol blue) was added to each sample along with 5 volumes of denaturing mix. The samples were mixed by pipetting. A vacuum was applied to the dotblot manifold using a Fisher brand dry vacuum pump/compressor, and the samples were loaded. Once the samples had been pulled through, the wells were washed and neutralised with 150 µl 2 x SSC. Once the SSC had been pulled through, the vacuum was released and the dotblot manifold disassembled. Dotblots prepared in this way were dried for 10 min at 80 °C in 3MM paper, and then transferred to a CL 1000 ultraviolet crosslinker (UVP) and exposed to 70,000 $\mu J/cm^2$ of UV. Blots were stored in the dark at room temperature covered with 3MM paper pockets, and wrapped in saran wrap.

2.2.5.c 5′ end labelling of oligonucleotide probes with γ -^{32}P –ATP using OptiKinase (USB)

The primers used for ologonucleotide hybridisations are shown in appendix iii.

Hybridisation with 5′ end labelled oligonucleotides had the advantage of being very quick, and very specific. This method was generally used to confirm the identity of bands on a southern blotted gel. The method however sometimes lacked sensitivity when extremely faint bands were being blotted.

2 pmol of oligonucleotide were added to 1 µl 10 x OptiKinase buffer, 5 units of OptiKinase, 0.5 µl γ -^{32}P –ATP (10 mCi / ml, 3000 Ci / mmol) and made up to a final volume of 10 µl with water. Mixes were incubated at 37 °C for 30 min and used immediately for hybridisation.

Hybridisations were carried out in modified Church buffer (Sambrook and Russell, 2001) at 45 °C. The transfer membranes were pre-hybridised in Church buffer for at least 30 min. Hybridisation was performed in 10ml Church buffer for 45 min prior to 2 washes in 10ml Church buffer at 45 °C. The blots were then washed three times in 3 x SSC and wrapped in saran wrap. Blots were either exposed to x-ray film or a phosphoimager screen (Amersham Biosciences).

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

Probes of 1 kb to 5 kb in length were amplified using standard long-range PCR conditions. Samples were run out on a 0.8 % (w/v) LE agarose gel, to assess the success of amplification. If a single band at the correct size was present, the PCR mixture was run through a PCR purification kit. For fragments < 4 kb, the MinElute PCR clean up kit and the standard protocol (Qiagen) was used. For fragments of 4 to 10 kb, the QIAquick PCR clean up kit and the standard protocol (Qiagen) was used.

Where more than one band was seen in PCR products, the rest of the sample was run on another 0.8 % gel and the band at the correct size cut from the gel on a Dark Reader using a clean scalpel. The DNA was purified using a QIAquick gel extraction kit and the standard protocol (Qiagen). Standard spin columns were used for fragments > 4 kb, and MinElute columns for fragments < 4 kb (Qiagen). DNA fragment concentrations were estimated by running 1, 2 and 4 µl against 125, 250 and 500ng of 1 kb molecular weight marker (NEB), and identifying the band closest in intensity to the query DNA fragment (Figure 2.4).

2.2.5.e Random prime labelling of probes

Random prime labelling involves heat-denaturing a PCR-based probe, and the generation of numerous short strands of radio labelled DNA using polymerases that require random hexamer oligonucleotides to prime, and have DNA double helix strand displacement capability. Hybridisation with these probes was very sensitive but more time consuming than 5′ end labelling.

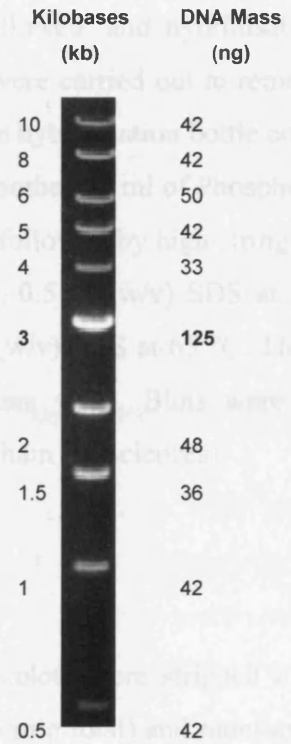


Figure 2.4: A photograph of 0.5 µg of Ethidium bromide stained1 kb DNA on a 0.8 % TAE agarose gel (Taken and adapted from http://www.neb.com/nebecomm/products/productN3232.asp). The mass of each fragment (ng) in 0.5 µg (1 µl) of molecular weight marker is shown.

2.2.5.f α -^{32}P -dCTP probe labelling using Rediprime II random prime labelling system (Amersham Biosciences)

Approximately 20ng of PCR-amplified probe DNA and 0.1ng 1 kb DNA Molecular weight marker (NEB) were diluted in TE buffer to a final volume of 45 µl. The probe was than 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 centrifuged briefly, and the contents added to a Rediprime II reaction tube and thoroughly mixed by pipetting until the blue marker dye was thoroughly

dispersed. 2.5 µl of α-^{32}P dCTP (10 mCi / ml, 3000 Ci / mmol) was added and the tube sealed and tapped gently to mix prior to being incubated at 37 °C for 60 min. The reaction was then boiled for 5 min and snap cooled for a further 5 min to denature the probe, and used immediately for hybridisation.

2.2.5.g Hybridisation of random prime labelled probes

Hybridisation was carried out in Church buffer at 65 °C. Pre-hybridisation of at least 1 hour in Church buffer was allowed, and hybridisation carried out in 20ml Church buffer over-night at 65 °C. Washes were carried out to remove unbound probe. 50 ml of Phosphate wash solution was added to the hybridisation bottle containing the blot, and shaken for 30 sec. This was then replaced with another 50 ml of Phosphate wash solution, and a 10 minute wash at 65 °C performed. This was followed by high stringency washing. This consisted of a single 30 minute wash in 2 % SSC, 0.5 % (w/v) SDS at 65 °C, followed by two fifteen minute washes of 0.2 % SSC, 0.2 % (w/v) SDS at 65 °C. The blots were then washed three times in 3 x SSC and wrapped in saran wrap. Blots were either exposed to x-ray film or to a phosphoimager screen (Amersham Biosciences).

2.2.5.h Blot stripping

Southern blots and dotblots were stripped of probes by washeing several times in boiling 1 % (w/v) SDS (2 to 4 min total) and monitored with a Geiger counter until radiation levels dropped below 5 counts per second. The blots were then rinsed in distilled water, and then in 3 x SSC. Stripped blots were either re-hybridised, or stored damp at 4 °C wrapped in Saran wrap.

2.2.6 Hybridisation Enrichment Protocol

5' biotinylated oligonucleotides (bio-oligos) were 18mer oligonucleotides (Table 2.5). The bio-oligos were ordered from Eurogentec S.A and were HPLC purified to ensure the removal of free biotin. The bio-oligos were disolved to a 100 µm concentrated stock. A mixture of all 4 bio-oligos was diluted to a 1.25 µM (total 5 µM) working stock. All stocks were stored in the dark at -20 °C.

Denaturing/hybridising/binding buffer (DHB buffer) was made in a category II PCR hood to a 10 x concentration (note this buffer is effectively PCR buffer without dNTPs or BSA). 10 x DHB was also diluted to a 1 x DHB stock. A separate stock of 1 x DHB 10 μg/ml BSA (DHB+BSA) was also prepared.

Elution mix (ED), 0.14 x DHB, 4.7 μg/ml single-stranded (heat-denatured) high molecular weight *E. coli* DNA, was prepared. This buffer preserves single-stranded DNA, especially at high temperatures. All buffers were stored at -20 °C.

The hybridisation enrichment protocol required siliconised 1.5ml eppendorf tubes, because single-stranded DNA has an affinity for plastic that reduces the DNA yield.

Primer ID	Sequence (5'- 3')	Information	Position wrt L1.3
bio-L1U1	bio-GGCACATGTATACATATG	universal for all L1 3' ends	5956
bio-L1U2	bio-GAAATACCATTTGACCCA	universal for all L1 3' ends	5489
bio-L1U3	bio-TGACAAAGGGCTAATATC	universal for all L1 3' ends	5124
bio-L1U4	bio-ACTACCTGACTTCAAACT	universal for all L1 3' ends	4602

Table 2.5: Biotinylated oligonucleotide sequences, the type of L1 they will hybridise to, and their location with respect to L1.3 (accession L19088).

Commercially available siliconised tubes appeared to result in a loss of yield, so manual siliconising was preferred. Siliconisation was carried out in a category II PCR hood with an external venting system. Tubes were soaked in dimethyldichlorosilane solution (BDH) for 1min, drained and left to air dry in the hood for 10 min. The tubes were then rinsed thoroughly with water and air dried in the hood. Tubes were then UV irradiated for 15 min and stored at room temp.

Dynabeads M-280 Streptavidin (Invitrogen Dynal) were stored in the dark at 4 °C as stated on the manufacturers' guidelines. Prior to enrichment the Dynabeads were prepared as follows: An aliquot of beads was put into a siliconised 1.5 ml eppendorf tube in a category II PCR hood. The aliquot was placed in a Dynal MPC -S magnetic particle concentrator and left for 10 sec. The tube was then twisted 90° either way from its original position several times to concentrate the beads into a small volume prior to the removal of the liquid by pipetting. To wash the beads, they were re-suspended in 100 μl 1 x DHB by tapping the tube, separated, and re-washed twice more. After the final separation, the particle separator was tapped on the bench top to ensure all the buffer was removed, then the beads were re-suspended to 1/12th of

their original volume in 1x DHB. The working concentration of beads was kept on ice in a covered ice bucket to limit exposure to light.

2.2.6.a Hybridisation Enrichment

For the enrichment protocol, an optimal annealing temperature (A°) was established (Chapter 6). Tubes were placed on ice along with 1 x DHB, DHB+BSA and DHB. Aliquots of DHB+BSA were also put into the water bath set at the optimal annealing temperature (A°) of 48 °C.

Annealing (Figure 2.6) was carried out in 0.2 ml PCR tubes in a thermal cycler. These contained 33 µl of purified PCR product, 4 µl 10 x DHB, 3 µl 5 µM bio-oligo mix (0.375 µM final concentration) giving a total volume of 40 µl. The mix was denatured at 96 °C for 75 sec followed by stepdown annealing, by 1 °C in 20 steps, from A + 9 °C to A + 1 °C. Finally annealing was completed by a final step of A° for 2 min. Annealing was carried out in a GeneAmp PCR system 9600 (Perkin Elmer Cetus).

Binding (Figure 2.6) of the bio-oligos was performed in pre-warmed siliconised tubes in the A° water bath. Annealed DNA was transferred into the siliconised tubes, and 3.6 µl of working stock Dynabeads was added. The contents of the siliconised tubes were mixed extremely gently by tapping the side of the tube. Mixing was repeated every 2 min for 10 min.

Washing (Figure 2.6) stages were carried out as follows. Beads were separated on the particle concentrator, and the unbound DNA was transferred into a fresh 0.2 ml PCR tube containing 0.7 µl of 4 x DHB and 3 µl of 5 µM bio-oligo mix for re-extraction (see below). The beads were washed by very gentle tapping in 100 µl of DHB+BSA on ice. The mix was then transferred to a fresh siliconised 1.5 ml eppendorf tube on ice. The beads were separated again, and the washings retained for analysis. The beads were then washed in 100 µl pre warmed DHB/BSA, and incubated at A° for 2.5 min. The beads were again separated, and the washings retained. The beads were washed in 100 µl of ED prior to transferring to a fresh siliconised eppendorf. The beads were separated, and the washings retained. The beads were re-suspended in 50 µl ED prior to thermal elution.

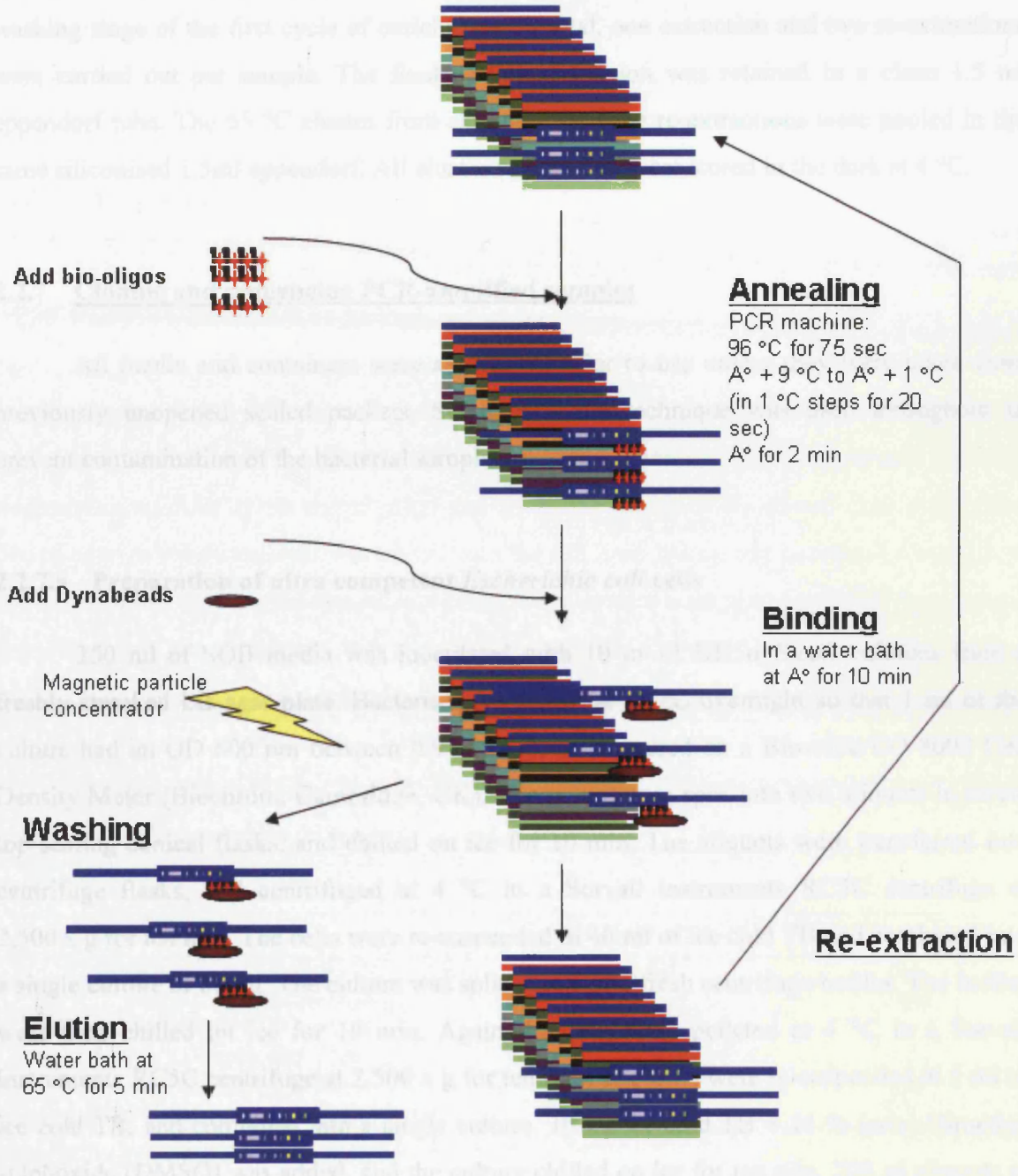


Figure 2.6: The process of recovering L1 insertions from multiplex PCR-amplified DNA.

Thermal Elution (Figure 2.6) of single-stranded DNA from the bead-bound bio-oligos was performed in a 65 °C water bath for 5 min. The beads were separated and the eluate recovered into a fresh 1.5ml eppendorf on ice.

Re-extraction (Figure 2.6) was performed on the un-bound fraction collected at the washing stage of the first cycle of enrichment. In total, one extraction and two re-extractions were carried out per sample. The final un-bound fraction was retained in a clean 1.5 ml eppendorf tube. The 65 °C eluates from extraction and the re-extractions were pooled in the same siliconised 1.5ml eppendorf. All eluates and washes were stored in the dark at 4 °C.

2.2.7 Cloning and sequencing PCR-amplified samples

All media and containers were autoclaved prior to use unless they were taken from previously unopened sealed packets. Standard sterile technique was used throughout to prevent contamination of the bacterial samples.

2.2.7.a Preparation of ultra competent *Escherichia coli* cells

250 ml of SOB media was inoculated with 10 to 12 DH5α *E.coli* colonies from a freshly streaked LB agar plate. Bacteria were grown at 18 °C overnight so that 1 ml of the culture had an OD 600 nm between 0.9 and 1.0, as measured on a Biowave CO 8000 Cell Density Meter (Biochrom, Cambridge, UK). The culture was split into two aliquots in screw top sealing conical flasks, and chilled on ice for 10 min. The aliquots were transferred into centrifuge flasks, and centrifuged at 4 °C in a Sorvall instruments RC5C centrifuge at 2,500 x g for ten min. The cells were re-suspended in 40 ml of ice cold TB and combined into a single culture of 80 ml. The culture was split in two into fresh centrifuge bottles. The bottles were then chilled on ice for 10 min. Again the cells were pelleted at 4 °C in a Sorvall instruments RC5C centrifuge at 2,500 x g for ten min. The cells were re-suspended in 5 ml of ice cold TB, and combined into a single culture. 10 ml ice cold TB + 14 % (w/v) Dimethyl sulphoxide (DMSO) was added, and the culture chilled on ice for ten min. 200 µl aliquots of the colony were taken into clean pre chilled 1.5 ml micro-centrifuge tubes. The aliquots were snap frozen in dry ice and industrial methylated spirits, and stored at – 80 °C.

2.2.7.b Ligation of PCR-generated DNA fragments into a plasmid vector

PCR-amplified DNA fragments were cut from agarose gels, and purified using a QIAquick gel extraction kit (Qiagen). Following DNA extraction, the reactions outlined in

table 2.7 were set up using the pGEM®-T Easy Vector System I kit (Promega). Ligation was carried out overnight at 4 °C. All other details for the protocol can be found in the protocol manual supplied with the pGEM®-T Easy Vector System I kit.

2.2.7.c Transformations

200 µl competent cell aliquots were defrosted on ice, and 5 µl of the ligation mix added. Samples were mixed by flicking, left on ice for 30 min, heat shocked in a water bath at 42 °C for 30 sec, and snap cooled on ice for 2 min. 900 µl of pre-warmed SOC media, 4 mg / ml glucose (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 and background control were plated onto LB Amp plates, and 200 µl of a transformation with no added plasmid DNA spread onto an LB only plate and an LB Amp plate (viability control). Colonies containing inserts were white as opposed to blue. The positive control gave a lawn of mostly white colonies, the background control a lawn of mostly blue colonies. The viability control gave a lawn of white colonies on the LB only plate, and no colonies on the LB Amp plate.

Reagent	Standard Reaction	Background Control
2X Rapid Ligation Buffer, T4 DNA ligase	5µl	5µl
pGEM-T easy vector (50ng)	0.5µl	0.5µl
Purified PCR product	3.5µl	-
Control Insert DNA	-	-
T4 DNA Ligase	1µl	1µl
Deionised water to a volume of 10µl	-	3.5µl

Table 2.7: The reactions used to ligate putified PCR-amplified fragments into the pGEM®-T Easy Vector prior to transformation of *E.coli.*

2.2.7.d Recovery of plasmid DNA

5 ml aliquots of LB plus 1 mg/ml Ampicillin were inoculated in triplicate with a single positive colony from each plate. The colonies were grown overnight in a shaker at 37 °C. DNA was recovered from the colonies using a QIAprep Spin miniprep kit (Qiagen).

Following extraction, *Mse*I digestion was carried out on 5 µl of the extracted plasmid DNA. Digested and undigested samples were run side by side on a 0.8 % (w/v) LE agarose gel to give a restriction fragment fingerprint. As cultures had been grown in triplicate, plasmid DNA isolated from cultures which showed differences in banding pattern were sequenced.

2.2.7.e Sequencing Using Big Dye Version Terminator v3.1 (Applied Biosystems)

Each sequencing reaction contained: 1µl of Big Dye terminator v3.1; 1.5µl 5 x Big Dye Buffer; 1µl of 3.3 µM sequencing primer (M13F or M13R); 20 to 30 ng DNA per kb of recovered plasmid; and water to a final volume of 10 µl. The reactions were placed in an MJ tetrad PCR machine PTC 225, and the following cycle performed: 96 °C for 10 sec; 50 °C for 5 sec; 60 °C for 4 min; and repeated for a total of 25 cycles.

2.2.7.f Clean up of the sequencing reactions

A master mix of 10 µl water and 2 µl 2.2 % (w/v) SDS per sequencing reaction was prepared. 12 µl were added to each reaction, mixed by pipetting then heated to 98 °C for 5 mins, and cooled to 25 °C for 10 mins, in an MJ tetrad PCR machine PTC 225. Traces of Big Dye were removed using PERFORMA DTR Gel Filtration Cartridges (Edge BioSystems & Vlt Bio Ltd). Finally the samples were submitted for sequencing by The Protein and Nucleic Acid Chemistry Laboratory (PNACL) at the University of Leicester, UK.

Chapter 3: Background work and experimental design

3.1 Results

Before any bench work was carried out various design aspects of the project were completed. The first task was to design the selector biotinylated oligonucleotides (bio-oligos) for use in hybridisation enrichment. These sequences were required to allow the selection of target sites, such that the target sites lacked sequences similar to the selector bio-oligos.

3.1.1 Bio-oligo design

The selector bio-oligos for use in hybridisation enrichment were designed to be complementary to the most conserved regions within the 3′ 1.5 kb of 90 aligned full-length L1s with intact open reading frames (ORFs; appendix iv on the accompanying CD-ROM). The full alignment can be found on the appendix CD-ROM that accompanies this thesis. Figure 2.1 shows the position and sequence of each bio-oligo within the sequence of the active human-specific L1, L1.3 (Accession L19088) (Dombroski *et al.,* 1991).

Each bio oligo was designed to be 18 nt in length and approximately 40 % GC. Each bio-oligo was modified with a 5′ mono biotin group, plus a phosphorothionate reverse linkage between nucleotides 17 and 18 to render it resistant to 3′ - 5′ exonuclease activity. The bio-oligos were synthesised by biomers.net (Ulm, Germany). The bio-oligos were HPLC purified to remove free biotin. Subsequently, bio-oligos were 5′ mono biotinylated only and were synthesised and HPLC purified by Eurogentec S.A. (chapter 6).

```
     ORF2>                                          bio-L1U4>
4561 aagtcaatcctaagccaaaagaacaaagctggaggcatcacactacctgacttcaaacta
4621 tactacaaggctacagtaaccaaaacagcatggtactggtaccaaaacagagatatagat
4681 caatggaacagaacagagccctcagaaataatgccgcatatctacaactatctgatcttt
4741 gacaaacctgagaaaaacaagcaatggggaaaggattccctatttaataaatggtgctgg
4801 gaaaactggctagccatatgtagaaagctgaaactggatcccttccttacaccttataca
4861 aaaatcaattcaagatggattaaagatttaaacgttaaacctaaaaccataaaaacccta
4921 gaagaaaacctaggcattaccattcaggacataggcgtgggcaaggacttcatgtccaaa
4981 acaccaaaagcaatggcaacaaaagacaaaattgacaaatgggatctaattaaactaaag
5041 agcttctgcacagcaaaagaaactaccatcagagtgaacaggcaacctacaacatgggag
                                bio-L1U3>
5101 aaaattttcgcaacctactcatctgacaaagggctaatatccagaatctacaatgaactt
5161 aaacaaatttacaagaaaaaaacaaacaaccccatcaaaaagtgggcgaaggacatgaac
5221 agacacttctcaaaagaagacatttatgcagccaaaaaacacatgaagaaatgctcatca
5281 tcactggccatcagagaaatgcaaatcaaaaccactatgagatatcatctcacaccagtt
5341 agaatggcaatcattaaaaagtcaggaaacaacaggtgctggagaggatgcggagaaata
```

5401 ggaacacttttacactgttggtgggactgtaaactagttcaaccattgtggaagtcagtg

bio-L1U2>

5461 tggcgattcctcagggatctagaactagaaataccatttgacccagccatcccattactg

5521 ggtatatacccaaatgagtataaatcatgctgctataaagacacatgcacacgtatgttt

5581 attgcggcactattcacaatagcaaagacttggaaccaacccaaatgtccaacaatgata

5641 gactggattaagaaaatgtggcacatatacaccatggaatactatgcagccataaaaaat

5701 gatgagttcatatcctttgtagggacatggatgaaattggaaaccatcattctcagtaaa

3•UTR>

5761 ctatcgcaagaacaaaaaaccaaacaccgcatattctcactcataggtgggaattgaaca

5821 atgagatcacatggacacaggaaggggaatatcacactctggggactgtggtggggtcgg

5881 gggaggggggagggatagcattgggagatatacctaatgctagatgacacattagtgggt

bio-L1U1>

5941 gcagcgcaccagcatggcacatgtatacatatgtaactaacctgcacaatgtgcacatgt

Poly-A Tail>

6001 accctaaaacttagagtataataaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa

Primer ID	Sequence (5'- 3')	Selectivity	Position wrt L1.3
bio-L1U1	bio-GGCACATGTATACATATG	universal for all L1 3' ends	5956
bio-L1U2	bio-GAAATACCATTTGACCCA	universal for all L1 3' ends	5489
bio-L1U3	bio-TGACAAAGGGCTAATATC	universal for all L1 3' ends	5124
bio-L1U4	bio-ACTACCTGACTTCAAACT	universal for all L1 3' ends	4602

Figure 3.1: The sequence shows the co-ordinates of each bio-oligo location with respect to human L1 insertion L1.3 (accession L19088). Bio-L1U1 is located in the 3' UTR and bio-L1U2 to bio-L1U4 are located within ORF2. The table also shows the sequence of each bio-oligo, and its selectivity.

Having designed and synthesised the bio-oligos, their sequences were used during the target locus selection and amplicon design procedure.

3.1.2 Target locus selection and design

Initially ten genomic regions were selected as candidate loci for target site design. Eight loci were selected because previously L1 insertions into these loci had been identified as causing human genetic disease, suggesting they were amenable to L1 insertion. In addition, two loci were selected that met the criteria outlined in target site design (chapter 3), but that had not been associated with previous L1 insertion. HOXD is a locus largely devoid of repetitive sequences for reasons that are not well established; and the MHC Class II (MHC2) region had been extremely well characterised for sequence diversity. Table 3.2 shows the selected target loci.

After the target genomic loci were selected, 6 kb to 7 kb candidate target regions were identified around the sites of disease-causing insertions, outlined in table 3.2. The HOXD candidate target region was selected by locating an area of the gene cluster that was completely devoid of repetitive elements. The MHC2 candidate region was selected from an area within the BRD 2 gene that showed a low density of L1 sequence. In order for a region to be accepted, the following criteria had to be met: 1. the region had to be devoid of human-specific L1 sequence and any close matches to the bio-oligo sequences; 2. the candidate sequence was required to contain numerous L1 EN recognition sites to allow L1 insertion (which are extremely common, so this was not a strict criterion).

Chromosome	Gene	Accession number	Significance	Reference
X	DMD	AL031643	X-linked dilated cardiomyopathy. Insertion into the 5'UTR of the DMD gene.	(Yoshida et al., 1998)
X	FIX	AL033403	Haemophilia B. Insertion into exon 5 of the factor 9 gene.	(Li et al., 2001)
X	CHM	AL009175	Choroideremia (CHM). Insertion into exon 6 of the CHM gene.	(van den Hurk et al., 2003)
X	CYBB	AL627245	Chronic granulomatous disease (CGD). Insertion into intron 5 of the CYBB gene.	(Meischl et al., 2000)
X	RP2	AL050307	Retinitis Pigmentosa 2 (RP2). Insertion into intron 1 of the RP2 gene.	(Schwahn et al., 1998)
11	HBB	AC104389	β-Thalassaemia. Insertion into intron 2 of the Haemoglobin B gene.	(Kimberland et al., 1999)
9	FCMD	AL158070	Fukuyama-type congenital muscular dystrophy (FCMD). Insertion into intron 7 of the FCMD gene.	(Kondo-Iida et al., 1999)
5	APC	AC008575	Colon Cancer Susceptibility (APPCC). Potential disease causing Insertion into exon 15 of the APC gene.	(Miki et al., 1992)
2	HOXD gene cluster	AC009336	Homeo box D gene cluster. Highly repeat deficient.	
6	MHC2 region	AL662845	MHC Class II region. Extremely well characterized in the lab.	

Figure 3.2: Selected target loci. The chromosome, gene name, and accession number of each target locus is shown. The significance column identifies the reason for selecting the loci. The reference column identifies the citation reporting the disease causing insertion.

The RepeatMasker algorithm (Smit *et al.,* unpublished data) was used to detect repetitive elements, including $L1_{HS}$, within the candidate sequence, and the candidate sequences were annotated with the results of the RepeatMasker analysis. Close matches to the bio-oligo sequences were identified using the Findpatterns program (Accelrys Inc) in GCG. A close match, defined as a contiguous sequence alignment with fewer than three nucleotides mismatched, to any of the bio-oligos resulted in rejection of the candidate region. Again the

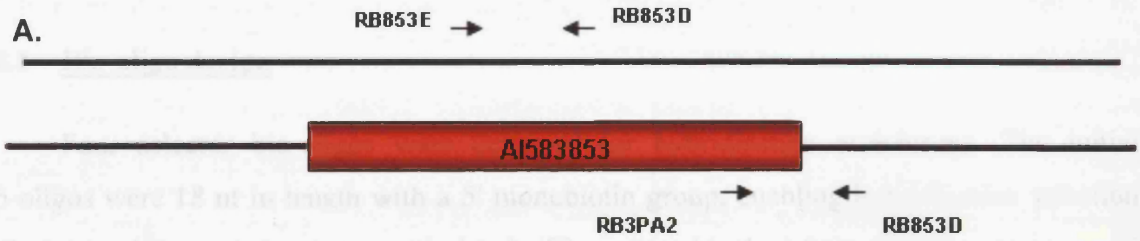
Findpatterns tool was used to identify EN recognition sites using a degenerate sequence based on the consensus endonuclease cleavage sequence, $(Pyr)_4$/Pur (Feng *et al.*, 1996; Moran, 1999).

If all criteria were met, the candidate region was accepted. Finally the RepeatMasker annotation was used to identify single-copy DNA for primer placement. Primary and nested secondary target site primers were designed within the single-copy DNA. In order to ensure that the target site primers could efficiently prime amplification of a 12 kb genomic DNA target (necessary if a 5 kb target site receives a full-length L1 insertion), control primers were also designed to generate an amplicon of approximately 12 kb when used in conjunction with the primary target site primers. Regions 12 kb 3′ of the 5′ primary Target site primers, and 12 kb 5′ of the 3′ primary Target site primers were scanned for single-copy sequence using the RepeatMasker algorithm. Control primers were placed within single-copy sequence DNA. All primers were approximately 20 bp in length, approximately 55% GC with a 3′ GG terminal dinucleotide to provide stable 3′ nucleotides for efficient primer extension.

3.1.3 Donor panel genotyping

To generate control insertion loci for assessing hybridisation enrichment efficiency, presence/absence genotype data for known polymorphic L1 insertions was required. Genotyping of the laboratory semen donor panel was carried out using samples collected by buccal swab (Materials and Methods). Presence/absence genotypes for the polymorphic L1 in accession AL583853 were determined by PCR using the primer combinations shown in figure 3.3 A. Presence of the insertion was determined using primers RB3PA2 (internal to the L1) and RB853D (located in the 3′ flank), giving a 225 bp amplicon. Absence of the insertion was determined using primers RB853D and RB853E, giving a 445 bp amplicon. Standard PCR amplification was used to determine genotypes (Materials and Methods). Following PCR amplification, samples were fractionated on 2 % (w/v) LE agarose gels, and genotypes were assigned (Figure 3.3 B).

A.



B.

Donor	Empty site	Filled site	Genotype
R1	0	1	+/+
R2	1	1	+/-
R3	0	1	+/+
R4	0	1	+/+
R5	0	1	+/+
R6	0	1	+/+
R7	0	1	+/+
R8	0	1	+/+
R9	1	1	+/-
R10	0	1	+/+
R11	0	1	+/+

Figure 3.3: A. A diagrammatic representation of the primer placement used in the genotyping of the AL583853 polymorphic L1 in the semen donor panel. B. The table summarises the genotyping results where the empty site does not contain the L1 insertion, and the filled site does contain the L1 insertion.

3.2 Discussion

3.2.1 Bio-oligo design

Four selector bio-oligos were designed for hybridisation enrichment. The initial bio-oligos were 18 nt in length with a 5′ monobiotin group, enabling hybridisation selection with streptavidin-coated paramagnetic beads. They were blocked from primer extension and hydrolysis at the 3′ end by synthetic modifications. This was necessary as the bio-oligos would be placed directly into tubes of un-purified PCR product potentially resulting in primer extension, which could interfere with the dynamics of hybridisation enrichment. A reverse linkage, at nucleotides 17 to 18, was added to prevent *Taq* DNA polymerase from extending the oligo, and a phosphorothioate group was added to prevent the 3′ to 5′ exonuclease activity of *Pfu* from removing the terminal linkage. Each bio-oligo was designed to be approximately 40 % GC.

The bio-oligos were placed in the most conserved regions in the 3′ 2 kb of an alignment of 90 full-length L1 elements with intact open reading frames (R. Badge, unpublished data, shown in appendix iv on the accompanying CD-ROM). For illustrative purposes the sequence of L1.3 (accession L19088) is used to show the location of the bio-oligos (Figure 3.1). Bio-oligos were targeted to the 3′ 2 kb of the L1 consensus sequence, because premature termination of reverse transcription frequently results in L1 insertions being 5′ truncated (Boissinot et al., 2000a; Ovchinnikov et al., 2002; Pavlicek et al., 2002b; Sassaman et al., 1997). Also, 13 out of the 17 known L1 insertions implicated in having caused human genetic disease were 5′ truncated (Appendix i). This strategy maximises the likelihood of recovering both truncated and full-length elements. The expectation was that *de novo* L1 insertions would most likely be generated by retrotransposition of evolutionary young Ta elements, as 16 out of 17 L1 insertions associated with human genetic disease are Ta elements (Ostertag and Kazazian, 2001a). Also the majority of retrotransposition competent (RC) L1 elements (71 %) identified by Brouha *et al.,* (2003) are Ta elements. This means that the young human-specific Ta sub-family was highly represented in the alignment of 90 full-length elements (65 %). The bio-oligos were therefore more likely to recover young L1 elements. However, any L1s inserting into the target loci (including older L1 families) which contain matches to the bio-oligo sequences should potentially also be recovered during enrichment. Due to the overall mass of L1 in the genome, very few loci are completely devoid

of L1 sequence. However, the vast majority of L1s in the genome are old (as shown by RepeatMasker analysis), and thus do not contain close matches to the bio-oligo sequences. This meant that target loci with no close matches to the bio-oligo sequences could be identified.

The initial principle behind bio-oligo design was to individually target L1s in two specific groups: Ta elements versus pre Ta elements, and universally for all L1s with matches to the bio-oligo sequences. The reason for not exclusively discriminating against non-Ta elements was that one case of human disease had been attributed to insertion of a pre-Ta L1 (Kazazian *et al.,* 1988). It would therefore have been short sighted to exclude pre Ta elements from the experiment. The Ta-specific bio-oligos (data not shown) were designed to discriminate between the Ta and pre Ta sub-family elements, dependant on an ACA or ACG, at the main Ta sub-family defining polymorphism located at position 5954 to 5956 (Boissinot and Furano, 2001) (See diagnostic polymorphisms of the Ta sub-family: chapter 1). Although these bio-oligos were designed, they were not used during the investigation. The universal oligos were designed to selectively anneal to young L1 elements. The method was therefore designed to be biased towards, but not exclusive for, the recovery of *de novo* $L1_{HS}$ insertions into the selected target sites.

3.2.2 Target locus selection and target site design

Initially ten loci were selected as candidate target sites for hybridisation enrichment (Figure 3.2). Eight of the selected loci had previously suffered an L1 insertion that had caused a genetic disease. These sequences must have been amenable for L1 insertion in the past, and it is likely that they will be amenable to L1 insertion in the selected semen donors. A locus devoid of repetitive sequences (HOXD) was also selected. This region either harbours a mechanism preventing mobile element insertion, or is subject to strong selection against the insertion of mobile elements (Greally, 2002). The final locus (MHC2) has been extremely well characterised for DNA sequence variation in the local semen donor panel and was subsequently ideal for robust primer design. Once target loci had been selected a target site was designed within each locus.

Each target site was designed to be approximately 6 kb in size and where possible was centred on the insertion site of the disease-causing insertion reported in the literature. Chosen sites were scanned for close matches to the bio-oligo sequences (defined as a sequence match

between any 15 out of the 18 nucleotides). Such matches would have resulted in target site rejection, unless the site could be shifted to avoid these close matches, but still incorporate the L1 insertion site. The bio-oligos used in the original DEASH protocol could discriminate between single nucleotide variants when used in conjunction with a non-biotinylated allele specific competitor oligonucleotide (ASO) directed against the unselected variant (Jeffreys and May, 2003). However hybridisation enrichment could not use a competitor ASO, so it was necessary to increase the level of discrimination. As a result, matches of 15 or more out of 18 nucleotides were considered close matches and avoided.

The target sites were scanned for matches to the EN recognition sequence, using a degenerate sequence and the Findpatterns tool in GCG, necessary for normal L1 integration. This sequence was, unsurprisingly, extremely common. Finally single-copy DNA, identified by RepeatMasker analysis, was used for the placement of primary target site primers, secondary target site primers and control primers. The control primers were used in conjunction with primary and secondary target site primers to ensure they were capable of priming amplicons of 12 to 13 kb in length. Another way to ensure that the primers could prime amplification through 12 to 13 kb of sequence would have been to generate an artificial target site containing an extra 6 kb of sequence, but this was avoided on the grounds of possibility of generating potentially contaminating molecules.

3.2.3 Genotyping the donor panel for the AL583853 insertion

The small donor panel used to supply this and other L1-related projects with DNA was set up to remove the burden of providing large amounts of DNA from the existing voluntary Departmental semen donor panel. As this project required large quantities of genomic DNA, individuals based within the Genetics department at the University of Leicester were asked to donate buccal, blood and semen samples as required. Each individual was given a unique identity code, and signed a consent form that complied with the existing ethical approval, granted by the Leicestershire and Rutland Ethical Review Committee (LeRC), allowing the collection of anonymised samples for use in genetic analysis. Genotyping was carried out on buccal DNA as part of an L1 genotyping project (Collier and Badge, unpublished data). The AL583853 genotyping data enabled the selection of a donor who was homozygous for the insertion.

Chapter 4: Long-range PCR optimisation

4.1 Results

In order to amplify a full-length L1 insertion into one of the target loci (~12 – 13 kb), long-range PCR had to be extremely efficient. Also, to amplify a *de novo* L1 insertion, which would exist as a single-molecule prior to PCR, amplification had to be efficient from the single-molecule level. This chapter details the optimisation of long-range PCR protocol capable of amplifying the target loci, and through a full-length L1 insertion.

4.1.1 Initial long-range PCR conditions

Two temperature PCR was performed throughout the investigation such that primer annealing and extension were carried out at the same temperature. This was routinely used for long-range PCR as it appeared to increase the overall efficiency of long-range PCR compared to three step PCR. Long PCRs were usually carried out in 10 or 20 µl reactions containing: 1.1 x PCR buffer; primers at a concentration of 0.5 µM; 50 ng of genomic DNA (gDNA) (increased to 100 ng for poor quality samples which had been frequently frozen and thawed); and 0.04 U/µl *Taq/Pfu* (20 units :1 unit).

Samples were heated to 96 °C for 1 min. Following the initial denaturation step, samples were heated to 96 °C for 20 sec, and then held at the optimal annealing temperature (66°C to 68°C) for 1 min per kb of the amplicon length, plus 1 extra min. This was repeated for 25 to 30 cycles. Finally the samples were held at the optimal annealing temperature minus 1 °C for 30 min.

This chapter details the experiments used to optimise the long-range PCR protocol. The finalised long-range PCR protocol can be found in the Materials and Methods.

4.1.2 Maximum range PCR

To generate sufficient L1-containing molecules for hybridisation enrichment recovery, efficient PCR amplification of 12 to 13 kb of target DNA was required. To ensure that this size of amplicon was within the capabilities of long-range PCR, amplification was performed using the control primers from the DMD and the HBB loci (PFDMDCF, PFDMDCR, PFHBBCF and PFHBBCR (Figure 4.1 A).

Standard long-range PCR conditions were used (as described above) except that a 21 min extension replaced the standard 14 min to allow for an additional 5 to 12 kb generated

by these amplifications. Annealing and extension were carried out at 68 °C. PCR was performed on 100 ng of gDNA from individual 133111 of the CEPH panel.

Figure 4.1 A shows the placement of the long-range control primers with respect to the target site, and the expected length of the amplicons. Figure 4.1 B shows that fragments consistent with expected amplicon sizes (approximately 18.2 kb and 24.5 kb) were generated.

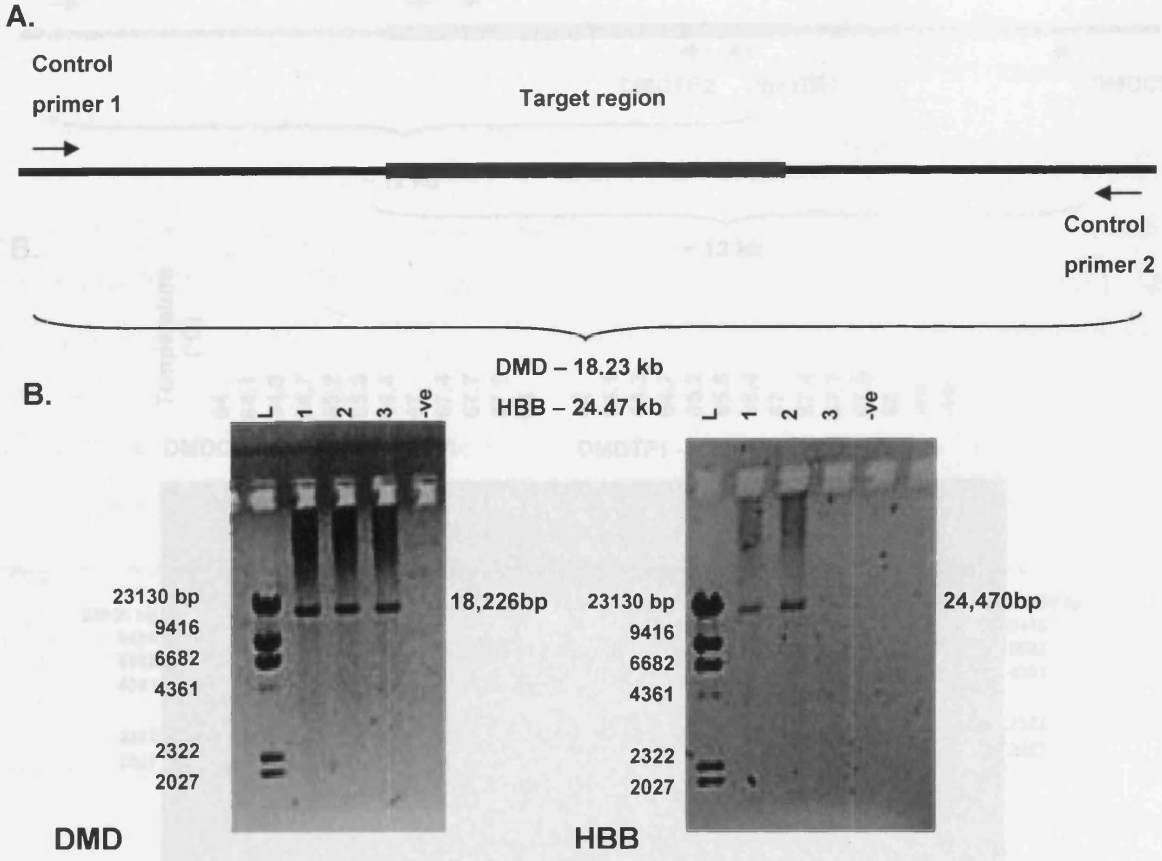


Figure 4.1: A. Diagram showing primer placement and expected amplicon length. B. Left, Ethidium Bromide (EtBr) stained gel of PCR products from the DMD locus after amplification with primers PFDMDCF and PFDMDCR. The lanes are triplicate samples from the same CEPH DNA. Right, HBB locus amplification showing the same triplicate, except one reaction failed to amplify. The primers used were PFHBBCF and PFHBBCR. The negative controls are PCR reactions containing water rather than genomic DNA. Molecular weight marker is 250 ng *Hind*III digested λ DNA

4.1.3 Optimisation of PCR

Temperature titration PCR experiments were used to identify the optimum annealing/extension temperature for long-range PCR. This involved setting the PCR machine (MJ tetrad PCR machine PCT225) to generate a range of annealing/extension temperatures

along the length of its thermal cycling block. Figure 4.2 A shows the primers used in a typical temperature titration PCR.

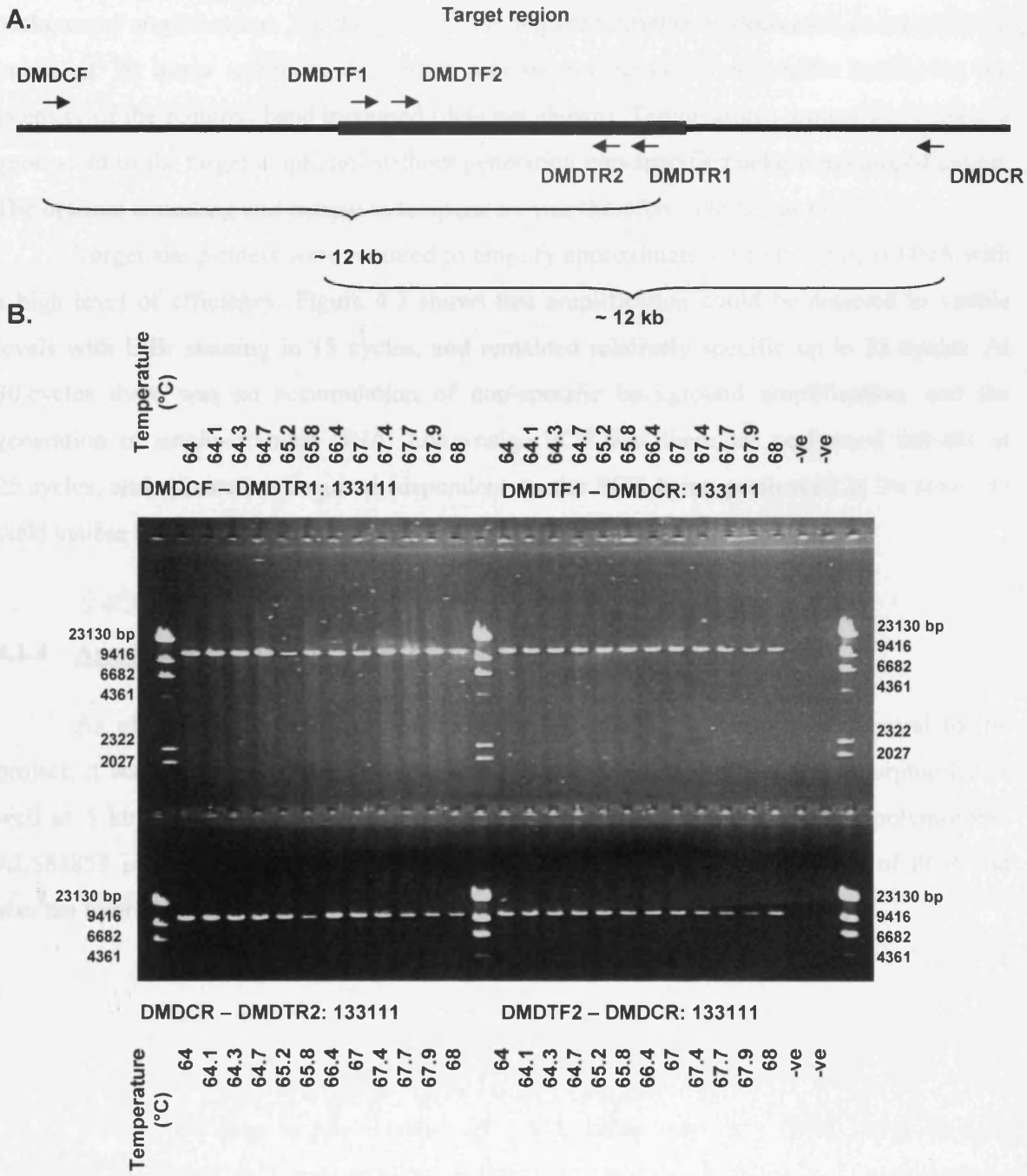


Figure 4.2: A. Primer placement used in temperature titration experiments (shown is the DMD target). Approximate amplicon sizes are shown. B. EtBr-stained gel from the DMD target temperature titration experiment. Amplification was performed on DNA from two separate members of the CEPH panel. The primer combinations used are shown, and also the identity of the CEPH family member used (133111). The negative controls are PCR reactions containing water. Molecular weight marker is 250 ng *Hind*III digested λ DNA.

The total range of temperatures tested was 60 °C to 80 °C, and figure 4.2 C shows a typical temperature titration experiment from 64 °C to 68 °C.

At higher annealing temperatures amplified samples showed very little non-specific background amplification, but the yield of the required amplicons decreased as temperature increased. At lower temperatures, certain loci showed several non-specific bands, but the intensity of the required band increased (data not shown). Temperatures around 66 °C gave a good yield of the target amplicons without generating non-specific background amplification. The optimal annealing and extension temperature was therefore selected as 66 °C.

Target site primers were required to amplify approximately 12 kb of target DNA with a high level of efficiency. Figure 4.3 shows that amplification could be detected to visible levels with EtBr staining in 15 cycles, and remained relatively specific up to 25 cycles. At 30 cycles there was an accumulation of non-specific background amplification, and the generation of single-stranded DNA. Long-range PCR was therefore performed initially at 25 cycles, and adjusted as required (dependent on the PCR being performed at the time), to yield visible bands using EtBr staining.

4.1.4 Amplification through a polymorphic L1

As efficient PCR amplification across a full-length L1 element was central to the project, it was necessary to demonstrate amplification through a full-length dimorphic L1 as well as 5 kb of flanking DNA (the size of a selected empty target site). The polymorphic AL583853 insertion, genotyped in chapter 3 was selected for the optimisation of PCR and also the hybridisation enrichment protocol.

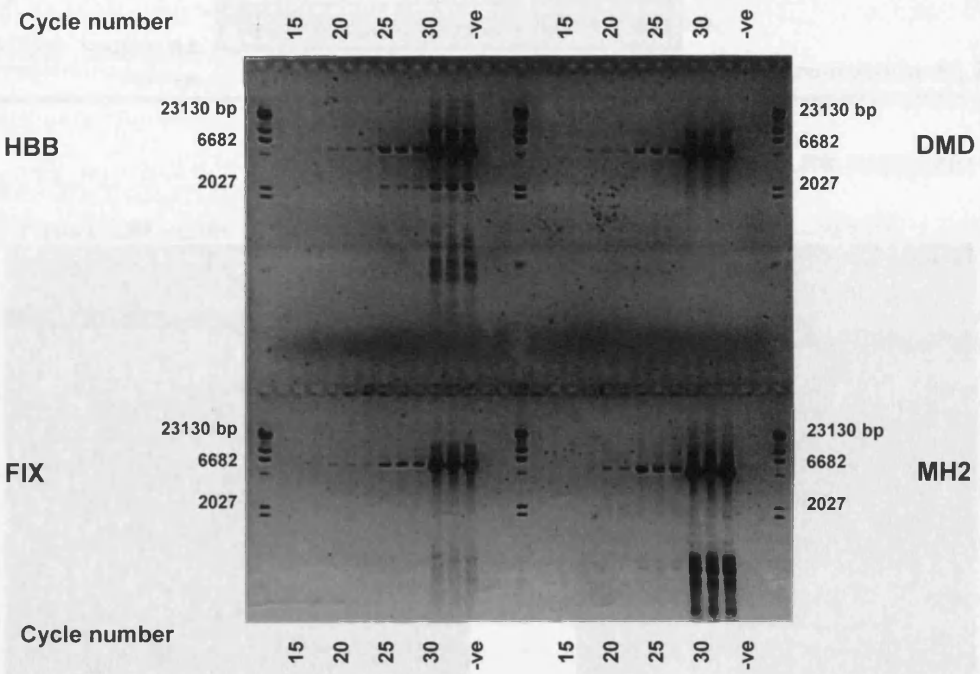


***Figure 4.3**:* Cycle titration from four target loci (HBB, DMD, FIX and MH2). Negatives are PCR reactions with water added. Molecular weight marker is 250 ng *Hind*III digested λ DNA.

Figure 4.4 A shows the location of the primers used to amplify AL583853, and 4.4 B shows the results obtained from PCR experiments. PCR of the AL583853 locus showed that amplification of the full-length element and its flanking DNA could not be achieved from the designed long-range primers. However further analysis showed that it was possible to amplify both through the L1 individually (4.4 C), and also through both the L1 plus 5′ flank (4.4 B) and L1 plus the 3′ flank (4.4 C). By comparison of gel photographs 4.4 B (primers PF853LRF to RB853D) and 4.4 C (primers PF853LRR1 to RB853E) an imbalance in the amplification efficiency between amplicons containing the 5′ and 3′ flanks was observed, despite both PCRs using the same mass of input gDNA.

The 3′ flanking amplicon amplified much more efficiently than the 5′ flanking amplicon in a donor who was homozygous for the absence of the insertion. This suggested a potential obstruction to efficient amplification 5′ of the L1 insertion site. Gel photograph 4.4 D also shows the much higher efficiency with which a smaller empty site amplicon amplified when compared to the larger filled site amplicon when PCR was carried out using DNA from an individual heterozygous for the AL583853 insertion.

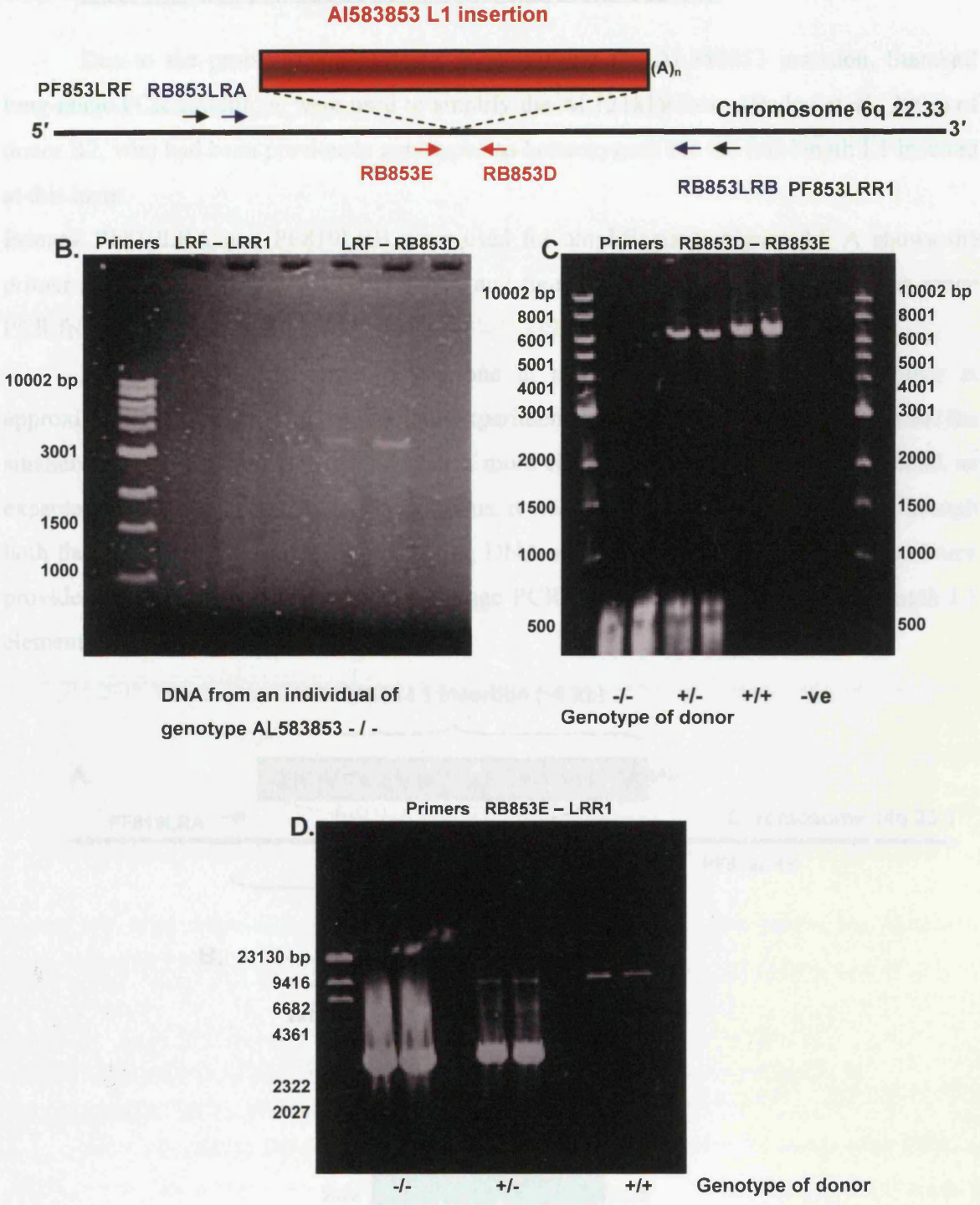


Figure 4.4: A. Primers designed around the AL583853 insertion site. B. Gel photograph showing amplification from the outlined primer sets (250 ng of NEB 1 kb molecular weight marker). EtBr-stained gel C confirms that amplification across the L1 is possible (250 ng of NEB 1 kb molecular weight marker), and photograph D shows efficient amplification across the L1 plus the 3' flank (250 ng of *Hind*III digested λ DNA).

4.1.5 Amplification across the AL121819 polymorphic insertion

Due to the problems encountered in amplifying the AL583853 insertion, Standard long-range PCR conditions were used to amplify the AL121819 locus (Badge *et al.*, 2003) of donor R2, who had been previously genotyped as heterozygous for the full-length L1 inserted at this locus.

Primers PF819LRA and PF819LRB were used for amplification. Figure 4.5 A shows the primer arrangement at the AL121819 locus, and figure 4.5 B shows the result of long-range PCR from these primers.

Figure 4.5 B shows two bands, one at approximately 6 kb and the other at approximately 12 kb. As with the previous experiment carried out at the AL583853 locus, the smaller band appeared to be amplified much more efficiently than the larger 12 kb band, as expected. However unlike the AL583853 locus, it was possible to amplify efficiently through both the full-length L1, and 6 kb of flanking DNA at the AL121819 locus. This experiment provided evidence that the standard long-range PCR system could amplify a full-length L1 element and its flanking target site.

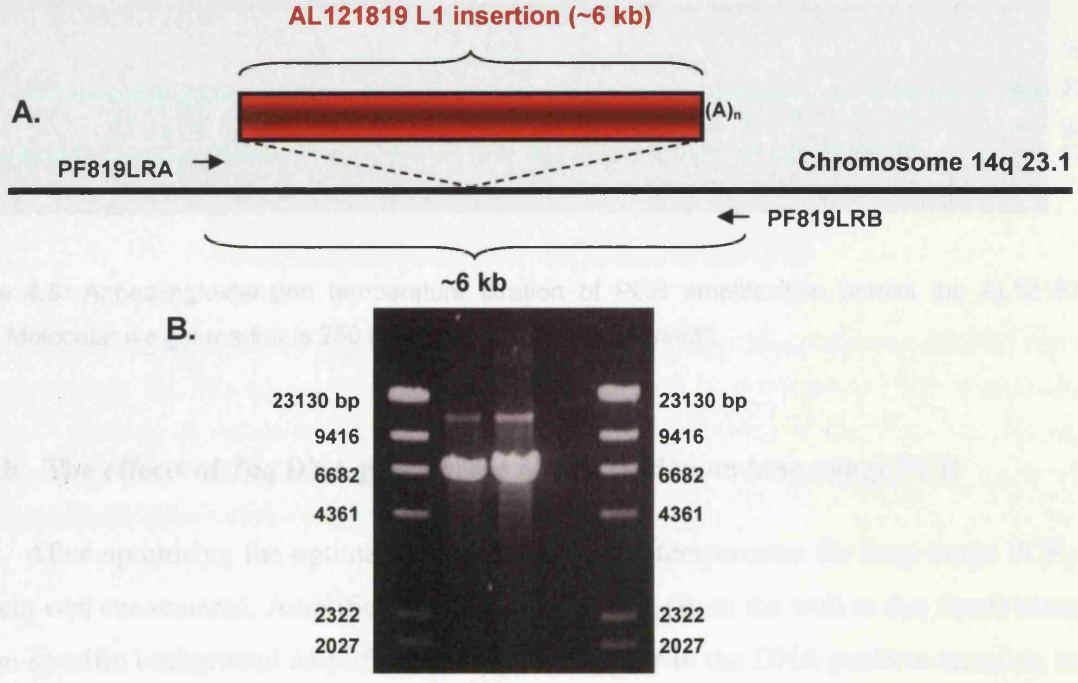


Figure 4.5: A. Primers used for long-range PCR amplification across the AL121819 locus. B. EtBr-stained gel showing the two fragments generated by amplification across the AL121819 target locus in a heterozygous male (donor R2). The larger band is ~ 12 kb and the smaller ~ 6 kb. Molecular weight marker is 250 ng λ DNA digested with *Hind*III.

4.1.5.a Temperature titration

To maximise the likelihood of recovering *de novo* L1 insertions, the efficiency of PCR needed to be increased. To maximise amplification of the 12 kb AL121819 filled site in the presence of a more efficiently amplifying empty site, a temperature titration PCR was performed. Again DNA from the heterozygous donor (R2) was amplified using PF819LRA and PF819LRB following the standard long-range PCR protocol.

Figure 4.6 shows the result of the annealing/extension temperature titration experiment. Amplification at 62 °C gave stable production of the 12 kb band to a relatively high yield without the generation of non-specific background amplification. Below 62 °C the level of non-specific amplification increased; however, above 62 °C the amount of the 12 kb AL121819 filled site amplicon began to decrease. 62 °C was therefore chosen as the optimal temperature for amplification across a full-length L1 plus 6 kb of its flanking DNA.

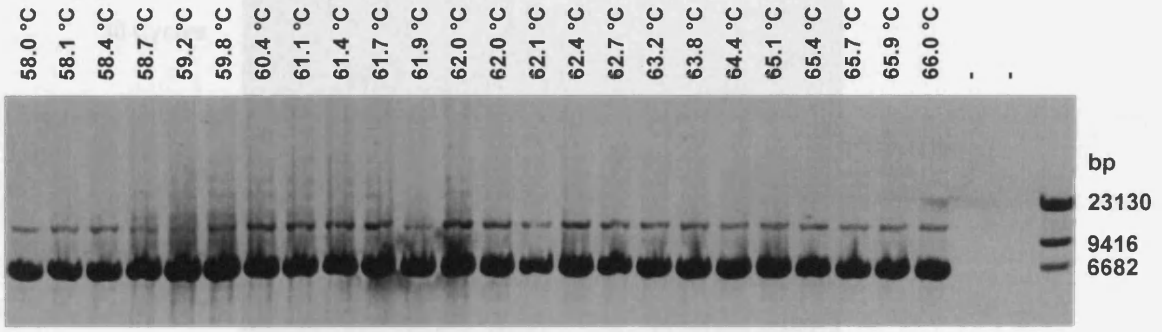


Figure 4.6: Annealing/extension temperature titration of PCR amplification across the AL121819 locus. Molecular weight marker is 250 ng λ DNA digested with *Hind*III.

4.1.5.b The effects of *Taq* DNA polymerase concentration on long-range PCR

After optimising the optimal annealing/extension temperature for long-range PCR, a problem was encountered. Amplified DNA of varying size (from the well to dye front) caused by non-specific background amplification began to appear in the DNA positive samples, and also in the DNA free control PCRs. This was possibly due to the high concentrations of *Taq* DNA polymerase in the PCR reactions causing erroneous primer extension. This had implications for long multiplex PCR due to the large numbers of primers present in such reactions.

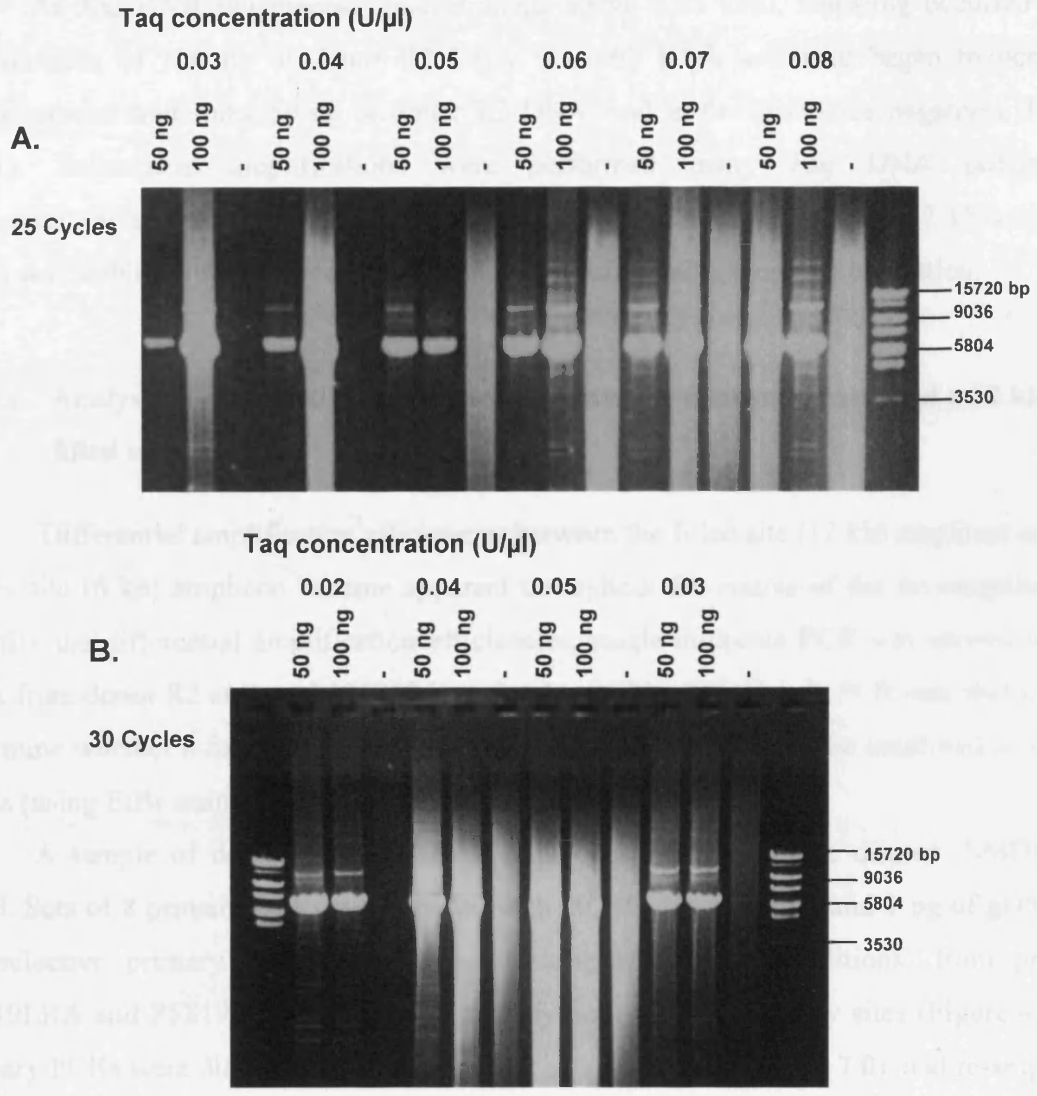


Figure 4.6: A. *Taq* concentration titration using the 819 locus in a heterozygous donor. 25 PCR cycles were performed. As *Taq* concentration increases, the amount of non-specific PCR amplification increases, resulting in smearing on the EtBr-stained gel. B. Repeat of the *Taq* concentration experiment using 30 PCR cycles. Molecular weight marker is 250 ng λ DNA double digested with *Eco*RI and *Sna*BI. (The 15720 bp band of the marker was generated by cohesive end joining in the λ digest).

To determine the optimal concentration of *Taq* DNA polymerase to be added to subsequent PCRs, a 20:1 unit mix of *Taq*:*Pfu* (5 U/µl : 0.25 U/µl) was added at different concentrations to PCR reactions. The AL121819 locus was amplified from both 50 and 100 ng of donor R2 DNA (heterozygous for the AL121819 insertion). Figure 4.6 shows a gel photograph of the results of a *Taq* DNA polymerase concentration experiment, using 25 (Figure 4.6 A) and 30 cycles (Figure 4.6 B) of PCR, but some non-specific bands are visible.

At *Taq* DNA polymerase concentrations above 0.03 U/µl, smearing occurred from amplification of 100 ng of donor R2 DNA. At 0.05 U/µl, smearing began to occur in amplifications containing 50 ng of donor R2 DNA, and in the DNA-free negatives (Figure 4.6 A). Subsequent amplifications were performed using *Taq* DNA polymerase concentrations of 0.025 U/µl. At this concentration generation of both the 12 kb and 6 kb bands was stable, with little smearing due to non-specific background amplification.

4.1.5.c Analysis of differential amplification between a 6 kb empty site and a 12 kb filled site

Differential amplification efficiencies between the filled site (12 kb) amplicon and the empty site (6 kb) amplicon became apparent throughout the course of the investigation. To quantify the differential amplification efficiencies, single-molecule PCR was carried out on DNA from donor R2 at the AL121819 insertion locus. Single-molecule PCR was also used to determine whether a full-length L1, plus 6 kb of flanking DNA, could be amplified to visible levels (using EtBr staining) from the single-molecule level.

A sample of donor R2 gDNA was diluted in single-molecule diluent (SMD) to 1 pg/µl. Sets of 8 primary PCRs were seeded with 20, 15, 10, 5, 4, 3, 2 and 1 pg of gDNA. A non-selective primary PCR (using standard long-range PCR conditions) from primers PF819LRA and PF819LRB was used to amplify both filled and empty sites (Figure 4.7 A). Primary PCRs were diluted at a ratio of 1 in 10 in 5 mM Tris-HCl (pH 7.0), and re-amplified in a secondary PCR.

The re-amplifications comprised either a non-selective secondary PCR (PF819LRC to PF819LRD) which amplified both the filled and empty molecules, or a selective PCR (PF819LRC to RBLR4015) which only amplified the L1-containing molecules due to the use of a primer internal to the L1 sequence (Figure 4.7 A). Non-selective amplification generated bands from the filled site (~ 12 kb), and empty site (~ 6 kb) (Figure 4.7 B). The selective secondary amplification only generated bands from the filled site (~ 8 kb) (Figure 4.7 C). In the non-selective secondary PCR, following a total of 60 cycles of PCR (30 in the primary PCR and 30 in the secondary PCR), the filled site (~ 12 kb) competes well with the empty site (~ 6 kb) at the 4 pg input level (Figure 4.7 B). This is shown by the presence of both the filled and empty site bands in the same lane on an EtBr-stained gel.

The data from the gels was input into the Poisson confidence interval program (A.J. Jeffreys) to estimate the mass of DNA containing a single amplifiable molecule of the 12 kb

filled site, and a single amplifiable molecule of the 6 kb empty site, plus 95 % confidence intervals for both estimates. The confidence interval program estimated that one amplifiable molecule of the empty site was present in 5.6 pg (95 % ci: 3.58 pg to 9.08 pg) of donor R2 DNA, and a single amplifiable molecule of the filled site was present in 11.5 pg (95 % ci: 7.33 pg to 18.89 pg) of donor R2 DNA.

Since the diploid genome size is approximately 6 pg, and the donor was heterozygous for the AL121819 insertion, this gives single-molecule amplification efficiency of 107 % for the empty site, and 52 % for the filled site. This trend was expected, as being double the size the filled site molecules were more likely to be lost due to DNA degradation and also amplified less efficiently due to the size of the amplicon.

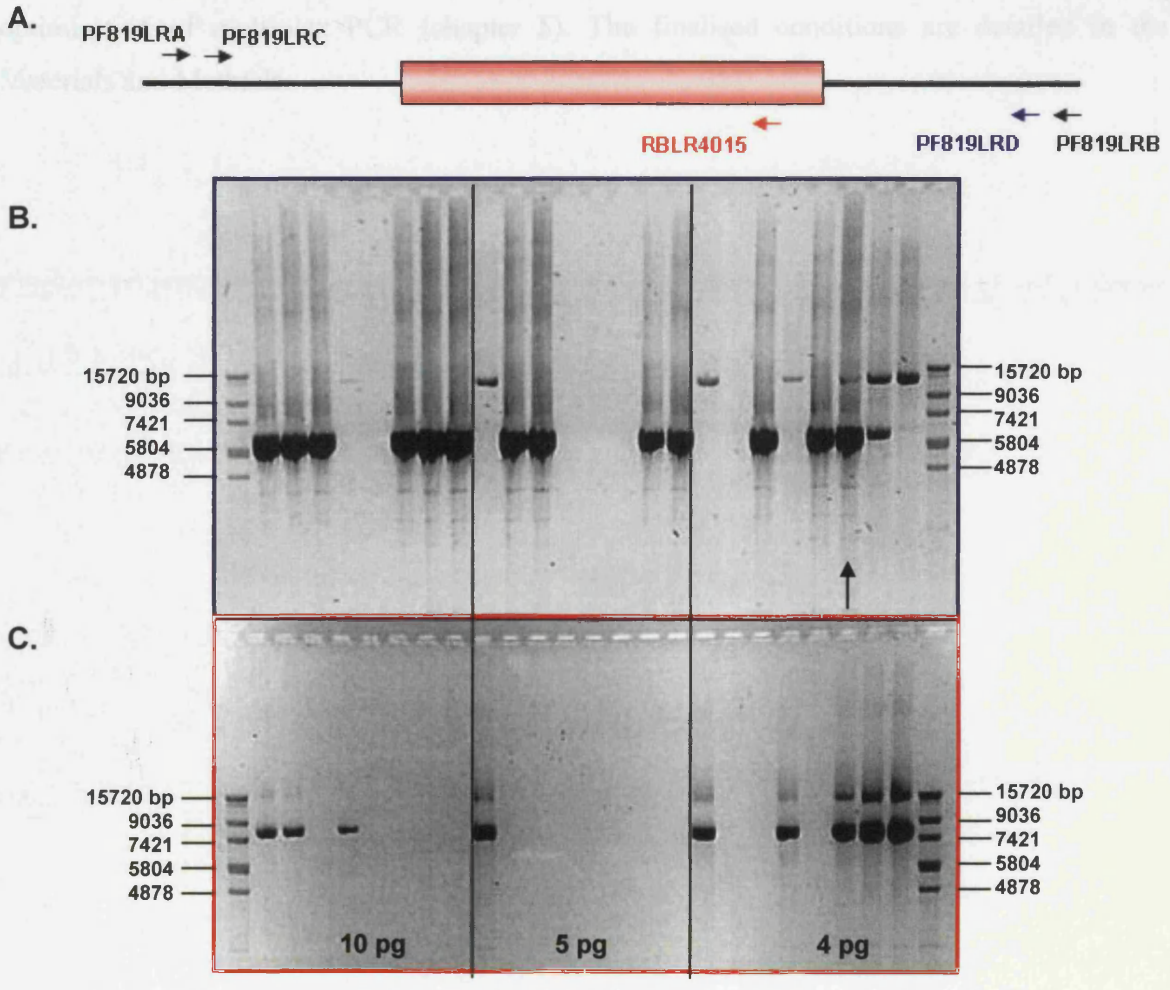


Figure 4.7: A. Primers used in single DNA molecule analysis. EtBr-stained gel showing non-specific amplification of both empty and filled sites. C. EtBr-stained gel showing specific amplification of the filled site. The arrow highlights a lane referred to in the discussion section (page 109) Molecular weight marker is 250 ng λ DNA double digested with *Sna*BI and *Eco*RI.

4.1.6 Optimised long-range PCR conditions

Following optimisation of the long-range PCR protocol, PCRs were carried out in 10 or 20 µl reactions containing: a primer concentration of 0.5 µM; PCR buffer at 1.1 x; 50 ng of gDNA (increased to 100 ng for poor quality samples); and 0.025 U/µl *Taq/Pfu* (20:1).

Samples were heated to 96 °C for 1 min. Following the initial denaturation step, samples were heated to 96 °C for 20 sec, and then held at the optimal annealing temperature of 62 °C for 1 min per kb of the amplicon length, plus 1 extra min. This was repeated for 25 to 30 cycles. Finally the samples were held at the optimal annealing temperature minus 1 °C for 30 min.

The primer concentration was subsequently reduced to 0.05 µM following optimisation of multiplex PCR (chapter 5). The finalised conditions are detailed in the Materials and Methods.

4.2 Discussion

4.2.1 Maximum range PCR

In order for hybridisation enrichment of *de novo* L1 insertion amplicons to be possible, it was necessary to ensure that the maximum required amplification distance, of 12 to 13 kb, did not exceed the upper limits of long-range PCR. The maximum required amplification distance was calculated as a 5 kb target site, 6 kb of L1 sequence and up to 2 kb of 3′ transduced sequence. To test the capabilities of the PCR protocol, control primers from the DMD and HBB loci were used in PCRs to generate amplicons of 18,226 bp and 24,470 bp respectively. The conditions under which the 24,470 bp HBB amplicon was amplified were sub optimal for an amplicon of its size: A 21 min extension was used during the PCR, but for an amplicon of approximately 24.5 kb we would usually use an extension time of 25.5 min (1 min per kb plus 1 min). Given that this PCR was successful for two out of three duplicate samples, it was concluded that 12 to 13 kb amplicons were well within the capabilities of the long-range PCR protocol. The size of the amplicons achieved in these maximum range experiments suggested that potentially the selected target sites could have been doubled in size, to allow the screening of twice the amount of DNA for insertions. This point however was not pursued due to the likely difficulties in efficient amplification of such large molecules, and the possibility that hybridisation enrichment would not be able to recover such large fragments. The maximum size of fragments recovered in the original DEASH protocol was 10.2 kb (30 % of selected target recovered with 100-fold enrichment efficiency), and 12.5 kb (< 8% of selected target site recovery with 200-fold enrichment efficiency) (Jeffreys and May, 2003). Thus the target amplicons were limited to 5 kb, giving a potential maximum fragment length of 12 to 13 kb following L1 insertion. Such target lengths were predicted to be within the capabilities of enrichment recovery from earlier DEASH data (Jeffreys and May, 2003).

4.2.2 Long-range amplification from the target site primers

Having shown that PCR could amplify the desired length of target gDNA (and longer amplicons), steps were taken to optimise the long-range PCR protocol. Control primers were designed such that PCR with an opposing target site primer would result in an amplicon of 12

to 13 kb in length. This avoided the requirement to construct artificial target sites that could become contaminants in subsequent experiments. The use of flanking control primers completely eliminated the risk of incorporating bio-oligo recognition sequence into target sites since the targets were selected to be free of bio-oligo recognition sites. Flanking DNA amplification was completely isolated from the target site amplicon by nesting relative to the control primers. The chosen strategy allowed optimisation of long-range PCR by amplification of gDNA rather than artificial targets constructed by PCR. It could however be argued that optimisation should have been carried out on a construct containing a full-length L1. This was not considered a viable strategy since it would have introduced a high risk of contamination.

The experiments performed to optimise the target site primers took the form of temperature titration amplifications. The experiments ensured the target site primers could amplify the required 12 to 13 kb of target from gDNA. Efficient amplification of all target site primers from all ten loci was required, using the same PCR conditions, to enable multiplexing of all target loci in the same reaction. For all target loci, primer extension through the control amplicons was extremely efficient at 66 °C with minimal background. This was true of both the primary target site primers and nested secondary target site primers.

As well as being able to amplify 12 kb to 13 kb of target from gDNA, the target site primers were also required to amplify the 5 kb target amplicon efficiently. This amplification had to be efficient using a 14 min extension time, as it would be used during the final experiment to allow amplification following L1 insertion. The result of the cycle titration showed that the primers amplify efficiently across the target site (figure 4.3).

As expected, all primers amplified under similar conditions since they had all been designed to meet a set of stringent criteria (chapter 3). Although it had been shown that these primers were capable of amplifying the required target length, amplification through a full-length L1 element had not been achieved. Potentially PCR amplification across a full-length L1 could have proved difficult due to its A T-rich nature (Han *et al.*, 2004). In addition, *de novo* L1 insertions often have very long poly A tails, for example 180 nucleotides (nt) (Miki *et al.*, 1992), which can be detrimental to PCR.

Amplification across a full-length L1 insertion in accession AL583853

To test amplification through a full-length L1, the known L1 insertion into AL583853 was chosen as it is prevalent in human populations with an allele frequency of 0.68 in North Western Europeans (Johns, T; unpublished data). Primers used to amplify the AL583853

insertion were designed using the same stringent criteria as the target site primers, and placed 3 kb 5′ and 3′ of the insertion in order to mimic a target site plus an L1 insertion. In theory, if the primers designed to amplify across the Al583853 insertion could do so efficiently, then the similarly designed target site primers would be able to amplify through the target site and a full-length L1 insertion.

Figure 4.4 A shows the AL583853 insertion locus, in intron 1 of the LAMA 2 gene (Chromosome 6q 22.33), and the placement of primers. The RepeatMasker algorithm showed that the region surrounding the AL583853 insertion site was free from $L1_{HS}$ sequences, and contained sufficient single-copy sequence for primer placement. GCG Findpatterns analysis showed that all four bio-oligo sequences were present in the AL583853 insertion, confirming that they could potentially hybridise to the L1. This suggested this insertion would be amenable to recovery by hybridisation enrichment. Unfortunately close matches (15 out of any 18 nt) to two of the bio-oligo sequences were detected within 1.5 kb 5′ of the insertion site. Analysis of the region using the RepeatMasker algorithm showed the presence of an L1PA7 element mapping to the same region as the closely matched bio-oligo sequences. This L1 was not expected to contain matches to the bio-oligo sequences, thus the primers had to be moved in order to generate an amplicon for optimisation of the hybridisation enrichment protocol (chapter 6).

Initially no combination of primers could amplify across either the filled or empty target site. Nested primers RB853E and RB853D, shown in figure 4.4 A, were used to dissect the AL583853 locus to identify the location of problematic DNA sequences. These primers had been designed previously (by Dr Richard Badge), and did not meet the stringent criteria used in target site primer design. However, they were still efficient in long-range PCR conditions. Figure 4.4 B shows a marked difference in amplification across the 5′ flank compared to the 3′ flank (Figure 4.4 C). This suggested the presence of a PCR blocking sequence in this region. The tandem repeat finder program (http://tandem.bu.edu/trf/trf.html) identified a purine rich microsatellite sequence, comprising an AG rich degenerate microsatellite sequence (AGAA ~24 copies), followed by a G rich tract 5′ of the AL583853 insertion. High GA content (98%) can potentially disrupt PCR due to G.A base pairing resulting in the formation of stable secondary structures (Seela *et al.,* 2005). Also microsatellites vary in size in the population, so the allele found in donor R2 may have been too large to amplify (*pers. comm.* Dr Richard Badge).

To achieve amplification through this L1, two separate PCRs were set up, the results of which are shown in figure 4.4 C and D. The first PCR used primers RB853D and RB853E,

and showed efficient amplification of the AL583853 L1 insertion and approximately 200 bp of flanking sequence (Figure 4.4 C). This confirmed that a full-length L1 could be amplified using long-range PCR. The second PCR, using primers RB853E and PFLRR1, showed amplification across the full-length element plus 3 kb of flanking sequence (Figure 4.4 D). Although this region was amplifiable, it did not amplify as efficiently as the 3 kb empty site (Figure 4.4 D). This could have been an intrinsic problem with the PCR primers, but it may also have been caused by using relatively poor quality CEPH gDNA. The laboratory stocks of these DNAs are frequently frozen and thawed, potentially causing DNA damage which may have contributed to a reduction in amplification efficiency. A certain imbalance in amplification efficiency between a 3 kb amplicon and 9 kb amplicon was expected.

In order to obtain an amplicon containing the AL583853 insertion plus 6 kb of flanking sequence, RB853E (located downstream of the identified purine-rich microsatellite sequence) was used in conjunction with PFENopF_PVU2D (11.89 kb amplicon). RB853E did not meet the design criteria used for target site primer design, but efficient PCR across the AL518853 insertion was still achieved. These primers were subsequently used to generate DNA for the optimisation of hybridisation enrichment (chapter 6).

Overall the AL583853 insertion had proved problematic for PCR amplification in terms of its sequence context, and although it successfully demonstrated amplification through a polymorphic full-length L1 insertion, it was decided that a more amenable L1 insertion was required.

4.2.3 Amplification across a full-length L1 insertion in accession AL121819

A number of young full-length L1 elements had been characterised in the group, including the polymorphic active L1 contained in accession AL121819 (Chromosome 14q 23.1) (Badge *et al.*, 2003). The sequence flanking the AL121819 insertion was scrutinised using Findpatterns in GCG and the RepeatMasker algorithm. No close matches to the bio-oligo sequences, and no young L1 insertions were identified in these flanking regions. The flanking sequence also contained sufficient single-copy sequences for primer placement.

Primer PF819LRA was designed 4247 bp 5′ of the insertion and PF819LRB 2205 bp 3′ of the insertion (figure 4.5 A). Amplification through a full-length element plus the flanking sequence (12425 bp based on a 6 kb insertion) was shown to be very efficient, even in the presence of the competing empty site locus which contained no L1 sequence (6452 bp).

As previously discussed, smaller amplicons amplify more efficiently than larger amplicons. To try to maximise the yield of the 12 kb filled site under competition from the 6 kb empty site, a temperature titration experiment was performed (Figure 4.5). Given an extension time of 14 min, the empty site should theoretically have completed primer extension in a high percentage of cycles. Thus in theory the number of molecules produced by the empty site should remain approximately the same throughout the temperature titration. However, by lowering the temperature annealing and primer extension becomes more efficient. This leads to a higher proportion of filled site molecules completing primer extension, thus increasing their yield.

In a standard PCR, the incorporation rate of *Taq* DNA polymerase is optimal at 75 °C to 80 °C (150 nt per second) (Gelfand and White, 1990). However, in a 2-step PCR, as used in this investigation, the optimal annealing/extension temperature is lower (40 °C to 66 °C) (Lopez and Prezioso, 2001). *Taq* DNA polymerase is error prone and mis-incorporates nucleotides at a rate of approximately one in every two thousand to twenty thousand bases (Smith *et al.*, 1997). *Taq* DNA polymerase becomes more error prone as its activity increases, and thus although it is more active at higher temperatures (> 75 °C), it is less productive when performing 2-step PCR. Long-range PCR requires proof reading activity of *Pfu*, thus the optimal temperature range increased to between 58 °C and 66 °C (Figure 4.5). Selecting the optimal annealing/extension temperature was difficult since the level of amplification was similar across the temperature range. 62 °C was selected to be the optimum temperature for annealing/extension because it showed a good level of filled site amplification but showed little non-specific amplification (Figure 4.6).

At this stage of the investigation, primer concentrations of 0.4 µM were routinely being used for long-range amplification. At these high primer concentrations, excess *Taq* activity can cause excessive non-specific amplification. This could also occur when performing multiplex amplification or secondary PCR amplifications on diluted primary amplifications. On two occasions during this investigation, new batches of *Taq* DNA polymerase showed greater activity than previous batches and had to be re-optimised. The second occasion is discussed in chapter 8. The first occasion was encountered during optimisation of long-range amplification through the AL121819 insertion. A *Taq* concentration titration experiment was performed to reduce non-specific amplification. At lower *Taq* concentrations (0.01 U/µl) amplification was very inefficient yielding sub-visible product if any (data not shown). At high concentrations (above 0.03 U/µl) non-specific amplification increased resulting in smearing when samples were fractionated by gel

electrophoresis (Figure 4.6). At this stage of the investigation, non-specific amplification was minimised by adopting 0.025 U/μl as the optimal *Taq* concentration.

4.2.4 Single-molecule analysis of differential amplification between a 6 kb empty site and a 12 kb filled site

Differential amplification occurred between the 6 kb empty site and the corresponding 12 kb filled site, when PCR was carried out from DNA of a donor heterozygous for the AL121819 insertion (R2). As a *de novo* L1 insertion into a target locus will be present as a single-molecule in a pool of sperm gDNA, hybridisation enrichment is dependent on the generation of numerous copies of the *de novo* L1 insertion by PCR. Subsequently hybridisation enrichment recovery is dependent on the ability of the bio-oligos to discriminate between L1-containing filled sites and non-L1-containing empty sites. However, the level of discrimination between filled and empty sites is finite, thus it was crucial that the L1-containing target could be amplified efficiently from the single-molecule level in the presence of a competing empty site. This was addressed through single-molecule PCR experiments and Poisson analysis.

From the results shown in figure 4.7, a single amplifiable 12 kb L1-containing filled site was estimated to be present in 11.46 pg of donor R2 DNA, and a single amplifiable molecule of the empty site in 5.62 pg of donor R2 DNA. The result suggested that in an individual heterozygous for the AL121819 insertion, there were approximately twice as many amplifiable empty site molecules as amplifiable filled site molecules in a sample of donor R2 DNA. Figure 4.7 also shows an alignment of gel photographs generated by amplification of a non-specific secondary PCR and an L1 specific secondary PCR. The L1 specific PCR was required because at higher DNA concentrations the larger L1-containing fragment was not visible on the gel photographs (see figure 4.7 10 pg samples). However at lower concentrations (4 pg), the L1-containing site competes well against the empty site and lanes appear that contain either the empty site, the filled site, or both sites simultaneously. Although there were twice as many amplifiable empty site molecules prior to PCR, after sixty cycles of PCR the presence of the empty site compared to the filled site (in the arrowed lane of figure 4.7 B) was, at worst, 1000 to 1 (estimated visually).

An estimate of the relative amplification efficiencies of the empty and filled sites was made. The amplification difference could be halved since the DNA sample contained twice as

many amplifiable empty site molecules. By taking the 60^{th} root (60 cycles) of the 500 fold difference between empty site molecules and filled site molecules, a relative amplification efficiency ratio of 1:1.11, between the filled and empty sites can be extrapolated.

$\sqrt[60]{500} = 1.11$ giving a relative efficiency of 1 (filled site) : 1.11 (empty site)

As Multiplex PCRs were optimised at 20 cycles (see chapter 4), the ratio of empty site to filled site molecules after 20 cycles of PCR was estimated at $1.11^{20} = 8$ fold. This implies that after 20 cycles of PCR, at worst, 8 fold more empty site molecules than filled site molecules will be generated. To recover *de novo* L1 insertions, enrichment efficiency must be much higher than 8 fold. This was required to ensure a net gain of filled site molecules following enrichment, despite their under-representation following PCR. This level of enrichment is well within the capabilities of hybridisation enrichment (Jeffreys and May, 2003), and as shown in chapter 6.

Chapter 5: Multiplex PCR Optimisation

5.1 Results

5.1.1 Tetraplex PCR

Initially, four loci were chosen for optimisation of multiplex PCR. The primary target site primers from four selected loci (HBB, DMD, FIX and MH2) were used simultaneously in tetraplex reactions. Standard long-range PCR conditions (see Materials and Methods) were used with the following amendments: reactions contained 100 ng genomic DNA (gDNA) in 20 µl reactions, and a 14 min annealing/extension time was used.

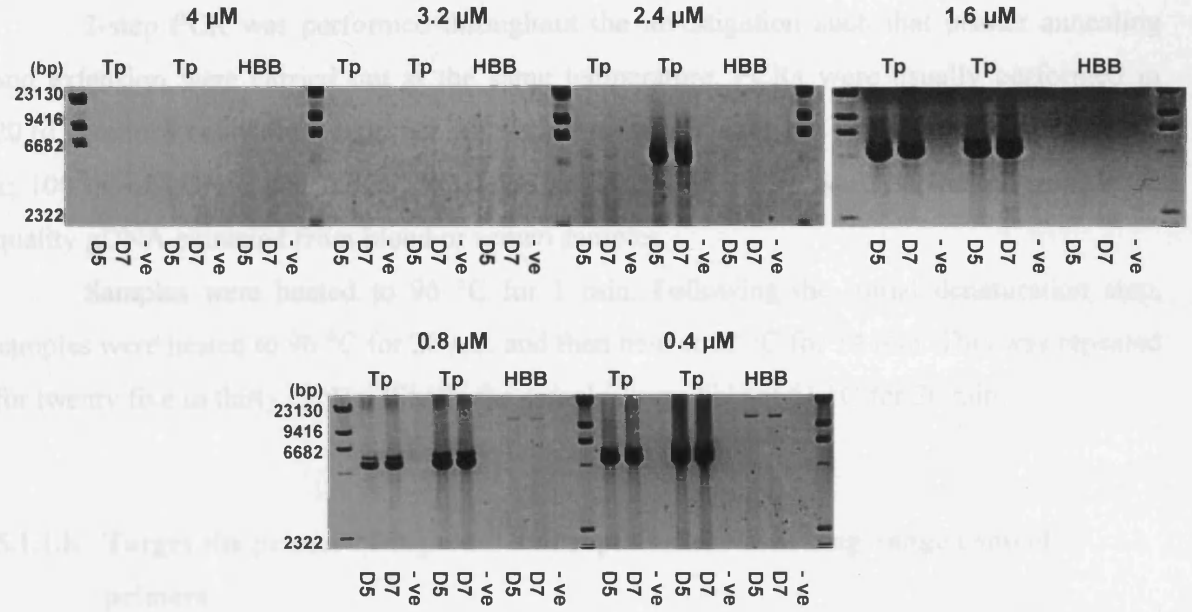


Figure 5.1: EtBr-stained tetraplex PCR (Labelled Tp). Every third lane is a negative control (PCR containing water instead of input DNA). Two different gDNAs were tested in duplicate (donor 5 and donor 7 of the departmental donor panel); also a long-range amplicon (HBBTF1 to HBBCR; labelled HBB) was amplified under the same conditions to assess long-range PCR at the lower primer concentrations. Negative control PCRs contained water rather than gDNA. Molecular weight marker is 250 ng of λ DNA digested with *Hind*III.

Initial attempts of this multiplex PCR failed to show amplification when visualised by Ethidium Bromide (EtBr) staining. There were no traces of smearing caused by non-specific amplification, suggesting that amplification had failed completely. Standard long-range PCR conditions contained a total primer concentration of 1 µM, but these initial multiplex reactions contained a total primer concentration of 4 µM. A primer concentration titration experiment

was performed (Figure 5.1). Separate PCRs containing total primer concentrations of 4 μM, 3.2 μM, 2.4 μM, 1.6 μM, 0.8 μM and 0.4 μM of were performed. All other PCR conditions were maintained.

Optimal results were obtained using a total primer concentration of 0.4 μM (0.05 μM of each primer), giving a final primer concentration of 0.4 μM. The presence of all four loci was confirmed by Southern blotting followed by hybridisation using relevant ^{32}P-5′-end labelled secondary target site primers (data not shown).

5.1.1.a Initial standard multiplex PCR conditions

2-step PCR was performed throughout the investigation such that primer annealing and extension were carried out at the same temperature. PCRs were usually performed in 20 μl reactions containing: a primer concentration of 0.05 μM per primer; PCR buffer at 1.1 x; 100 ng of gDNA; and 0.025 U/μl *Taq/Pfu* (20:1). All PCRs were performed using high quality gDNA extracted from blood or semen samples.

Samples were heated to 96 °C for 1 min. Following the initial denaturation step, samples were heated to 96 °C for 20 *sec*, and then held at 62 °C for 14 min. This was repeated for twenty five to thirty cycles. Finally the samples were held at 61 °C for 30 min.

5.1.1.b Target site primer "Drop out" and replacement with long-range control primers

To determine whether amplification of 12 kb of target DNA could be achieved in tetraplex reactions, "drop out" experiments were performed. One primary target site primer from one of the tetraplex targets was removed, and replaced with the corresponding control primer. Figure 5.2 A and 5.2 B illustrate this approach. The experiment was repeated so that all four loci were subject to a primer "drop out". PCRs were performed under standard multiplex conditions (figure 5.2 C).

Following PCR, bands could be visualised by EtBr staining at approximately 5 kb as expected for tetraplex amplification of target site primers. Bands could also be seen at approximately 12 kb which had not been seen in standard tetraplex PCR (Figure 5.1). These should have been bands generated by amplification between the remaining target site primer, and the replacement long-range control primer.

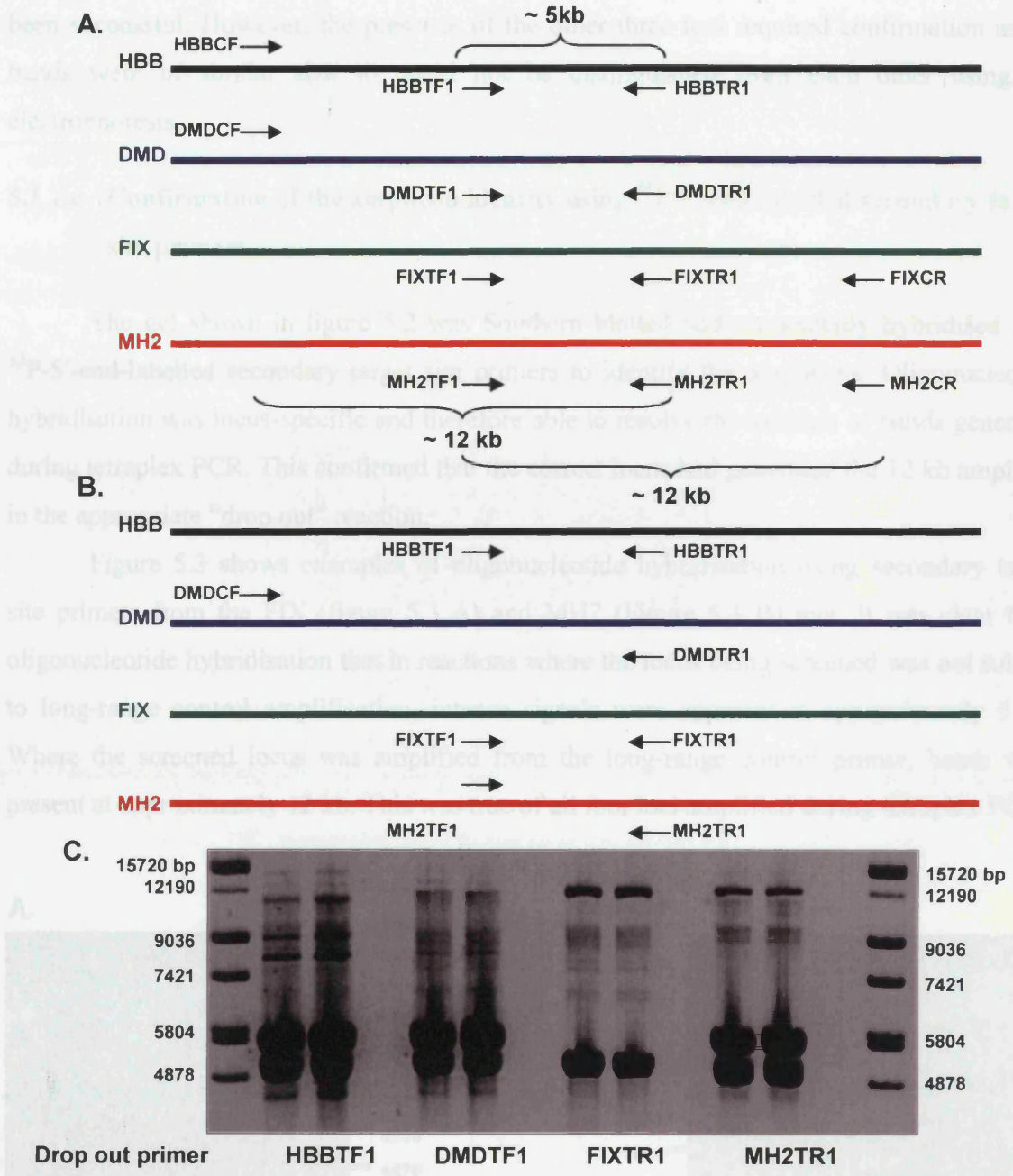


Figure 5.2: A. Diagram showing the position of the primers used in the "drop out" experiment. B. Diagram showing an example of the primers used in a "drop out" experiment, in this case for the DMD target site. C. EtBr-stained gel showing "drop out" tetraplex PCRs, and the identity of the dropped primer. The presence of amplicons at ~ 4.5 to 5.5 kb can be seen as well as control amplicons at ~ 12 kb. Some non-specific amplification occurred giving smearing and erroneous bands from ~ 5.8 kb to ~ 12 kb. Negative control PCRs contained water rather than gDNA. Molecular weight marker is 250 ng of λ DNA digested with *Eco*RI and *Sna*BI.

The FIX "drop out" experiment showed absence of the band at approximately 5.6 kb, and the presence of a band at approximately 12 kb. This suggested that the drop out experiment had

been successful. However, the presence of the other three loci required confirmation as the bands were of similar size so could not be distinguished from each other using gel electrophoresis.

5.1.1.c Confirmation of the amplicon identity using ^{32}P 5′ end labelled secondary target site primers

The gel shown in figure 5.2 was Southern blotted and sequentially hybridised with ^{32}P-5′-end-labelled secondary target site primers to identify the amplicons. Oligonucleotide hybridisation was locus-specific and therefore able to resolve the location of bands generated during tetraplex PCR. This confirmed that the correct locus had generated the 12 kb amplicon in the appropriate "drop out" reaction.

Figure 5.3 shows examples of oligonucleotide hybridisation using secondary target site primers from the FIX (figure 5.3 A) and MH2 (Figure 5.3 B) loci. It was clear from oligonucleotide hybridisation that in reactions where the locus being screened was not subject to long-range control amplification, intense signals were apparent at approximately 5 kb. Where the screened locus was amplified from the long-range control primer, bands were present at approximately 12 kb. This was true of all four loci amplified during tetraplex PCR.

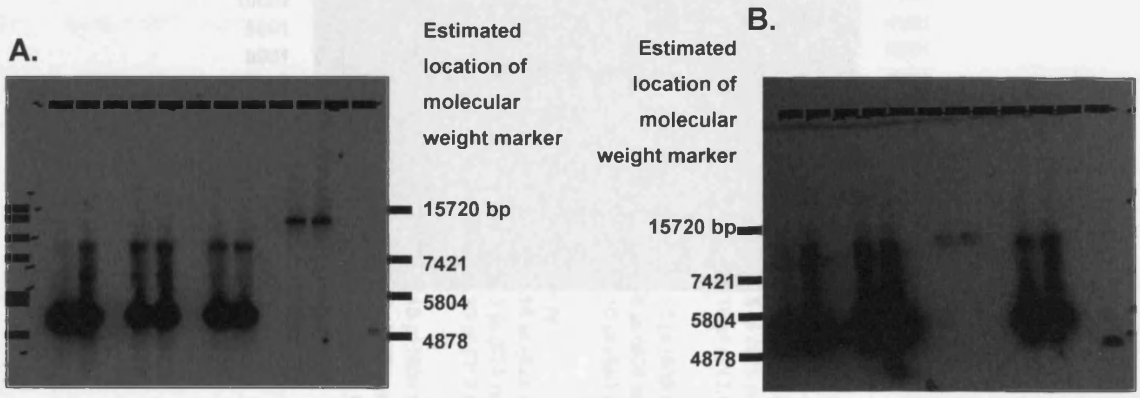


Figure 5.3: A. Autoradiograph showing hybridisation with primer PFMH2TR2. B. Autoradiograph showing hybridisation with primer PFFIXTF2. The autoradiograph was compared to the gel photograph (fig 5.2) to estimate the location of the molecular weight marker.

These drop out experiments demonstrated that a 12 kb amplicon could compete with three 5 kb amplicons when amplified in a tetraplex PCR, in the absence of a competing 5 kb amplicon generated from the same target locus.

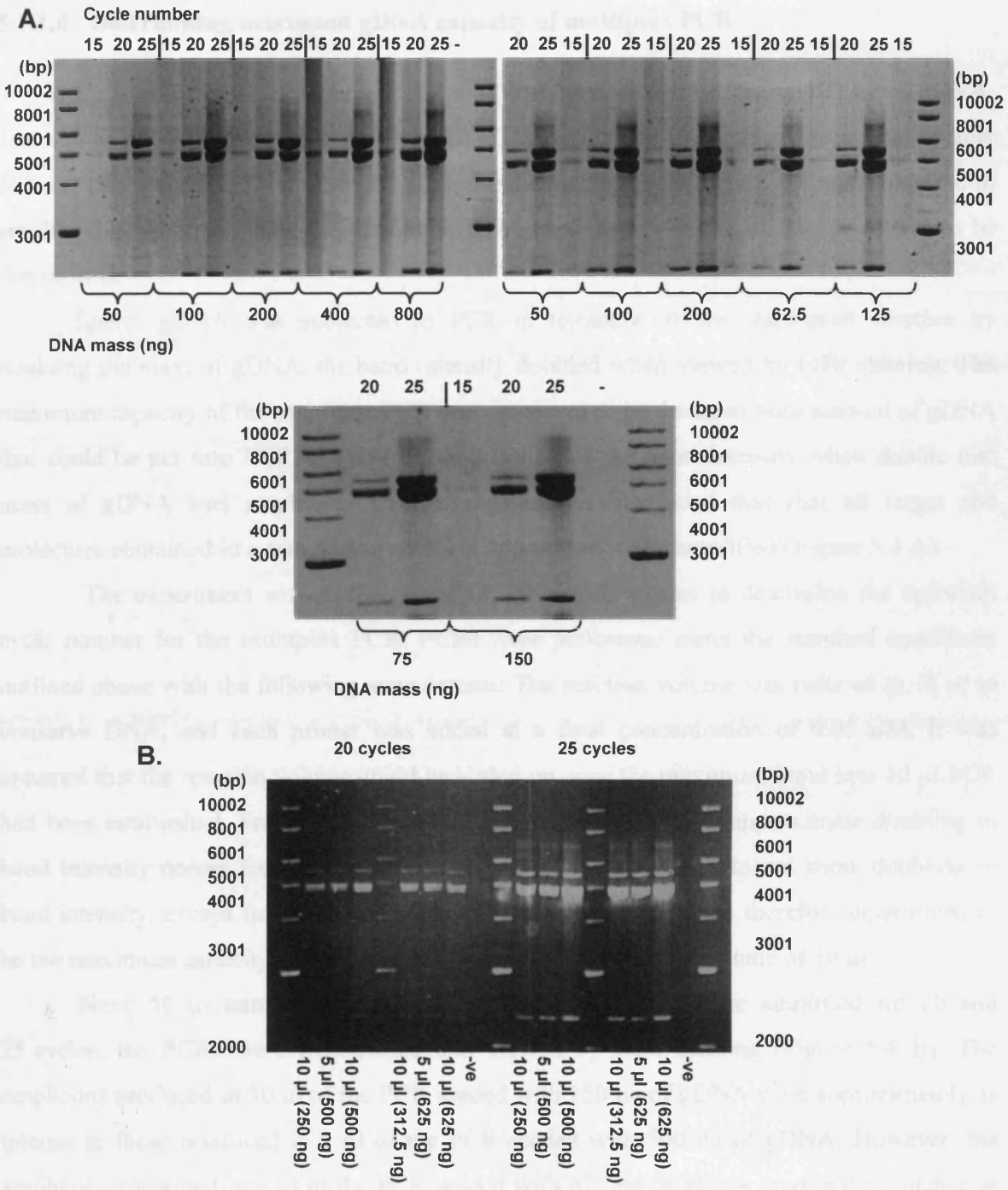


Figure 5.4: A. DNA input mass titrations of genomic DNA into 10 µl tetraplex PCRs. Samples were amplified for 15, 20 and 25 cycles. By comparing neighbouring samples at equivalent cycle number, an estimation of yield doubling can be made. B. EtBr-stained scaled tetraplex PCRs. Negative control PCRs contained water rather than gDNA (Donor 7). Samples were fractionated alongside 250 ng 1kb molecular weight marker.

5.1.1.d Determining maximum gDNA capacity of multiplex PCR

Up to this point, tetraplex amplification had been carried out using 100 ng of gDNA. The aim of the experiment was to screen all gDNA contained in an ejaculate (approximately 600 µg) for *de novo* L1 insertions. To minimise the number of multiplex reactions required to amplify this mass of DNA, the maximum DNA capacity for the multiplex PCR had to be determined.

Sperm gDNA was subjected to PCR in tetraplex. It was estimated whether by doubling the mass of gDNA, the band intensity doubled when viewed by EtBr staining. The maximum capacity of the multiplex PCR was estimated to be the maximum amount of gDNA that could be put into PCR, but still allow doubling of the band intensity when double that mass of gDNA was amplified. This maximised the likelihood that that all target site molecules contained in a sample had an equal opportunity to be amplified (Figure 5.4 A).

The experiment was performed at 15, 20 and 25 cycles to determine the optimum cycle number for the multiplex PCR. PCRs were performed using the standard conditions outlined above with the following amendments: The reaction volume was reduced to 10 µl to conserve DNA, and each primer was added at a final concentration of 0.05 µM. It was assumed that the reaction volume could be scaled up once the maximum input into 10 µl PCR had been established. From figure 5.4 it can be seen that although approximate doubling in band intensity occurs from 50 ng to 100 ng inputs, higher inputs do not show doubling in band intensity, except for possibly 62.5 to 125 ng. 100 ng input was therefore determined to be the maximum capacity of the tetraplex PCR, in a final reaction volume of 10 µl.

Next, 50 µl tetraplex PCRs were performed. Samples were amplified for 20 and 25 cycles, the PCRs were fractionated and viewed by EtBr staining (Figure 5.4 B). The amplicons produced in 10 µl of the PCR seeded with 250 ng of gDNA were approximately as intense as those produced in 5 µl of the PCR seeded with 500 ng of gDNA. However, the amplicons contained in 5 µl of the PCR seeded with 625 ng of gDNA were not as intense as the amplicons contained in 10 µl of the PCR seeded with 312.5 ng of gDNA. This confirmed that the maximum capacity of a 50 µl multiplex PCR was approximately 500 ng of gDNA.

The bands on the gel photographs (Figure 5.4 A and B) were clearest at 20 cycles. At 25 cycles PCR yield appeared approximately even at all inputs of DNA, thus it is likely that the reaction had reached saturation levels. At 15 cycles very little product was made so 20 cycles was selected as the optimum cycle number for multiplex PCR.

5.1.1.e Ensuring that all loci in tetraplex PCR amplify with approximately equal efficiency

De novo L1 insertion could occur into any of the target sites; thus it was necessary to ensure that all loci amplified with approximately equal efficiency. Agarose gel electrophoresis could not resolve the individual bands when stained with EtBr, so tetraplex PCRs were digested with *Bss*SI under the conditions outlined in the Materials and Methods. Primer concentrations were altered ("balanced"), and PCR repeated until all four bands appeared at approximately equal intensity following gel electrophoresis (Figure 5.5 B).

The "balanced" concentrations for each primer were: PFHBBTF1 and PFHBBTR1 at 0.07 μM; PFDMDTF1 and PFDMDTR1 at 0.12 μM, PFFIXTF1 and PFFIXTR1 at 0.08 μM; PFMH2TF1 and PFMH2TR1 at 0.15 μM.

A

Locus	Amplicon size (undigested)	Amplicon size (digested)
FIX	5552 bp	5552 bp
MH2	5018 bp	3699 + 1206 + 94 +18 bp
HBB	5006 bp	5006 bp
DMD	4909 bp	4746 + 162 bp

B

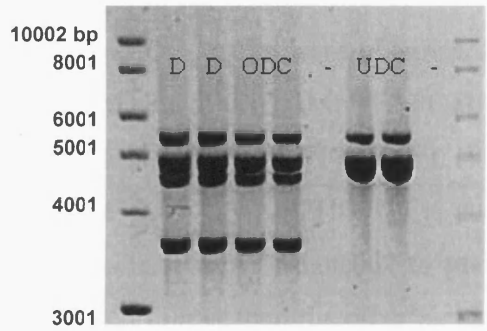


Figure 5.5: A. Table showing the lengths of tetraplex amplicons before and after *Bss*SI digestion. B. Fractionation of tetraplex PCRs on an EtBr-stained gel. *Bss*SI-digested and over digest control (ODC) (Sample digested for 1 hour rather than 30 min to ensure complete digestion had occurred) samples are shown along with undigested control (UDC) samples. Negative control PCRs contained water rather than gDNA. Samples were fractionated alongside 250 ng 1 kb molecular weight marker.

5.1.2 Hexaplex PCR

Following successful tetraplex PCR optimisation, the remaining six sets of primary target site primers (CHM, APC, CGD, RP2, FCMD and HOXD) were used simultaneously in a hexaplex reaction, under the standard multiplex PCR conditions.

Similarly to the tetraplex PCR, primer concentration balancing was performed. No restriction enzyme digestion was required as the bands could be resolved by agarose gel electrophoresis and EtBr staining.

A.

Locus	Amplicon Length
CHM	5678 bp
APC	5417 bp
CGD	5160 bp
FCM	5019 bp
RP2	4774 bp
HOD	4284 bp

B.

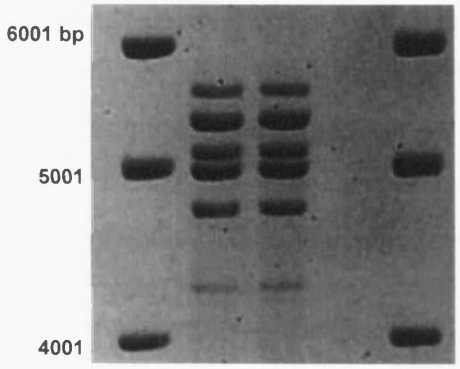


Figure 5.6: A. Table showing the lengths of the hexaplex amplicons. B. Hexaplex PCR visualised by EtBr staining. Negative control PCRs contained water rather than gDNA. Samples were fractionated alongside 250 ng 1kb molecular weight marker.

Figure 5.6 B shows the "balanced" hexaplex PCR. Optimal results were obtained at the following primer concentrations: PFCHMTF1 and PFCHMTR1 at 0.10 μM; PFAPCTF1 and PFAPCTR1 at 0.025 μM; PFCGDTF1 and PFCGDTR1 at 0.07 μM; PFRP2TF1 and PFRP2TR1 at 0.03 μM; PFFCMTF1 and PFCHMTR1 at 0.05 μM; PFHODTF1 and PFHODTR1 at 0.10 μM. Although the primer concentrations had been "balanced" to give approximately equal band intensities, the HOXD locus remained fainter than the other bands. The result, shown in figure 5.6, shows the best achieved amplification of the HOXD locus in a hexaplex PCR.

5.1.3 Decaplex PCR

The "balanced" tetraplex and hexaplex PCRs were combined to give decaplex PCRs. Decaplex PCRs were digested with *Bss*SI to separate the bands previously amplified in tetraplex, thus allowing resolution of the bands by EtBr staining. None of the amplicons amplified in the hexaplex contained *Bss*SI restriction sites. Both digested and undigested decaplex PCRs were fractionated alongside *Bss*SI-digested tetraplex PCRs, undigested tetraplex PCRs and hexaplex PCRs (Figure 5.7 B). The decaplex PCR was fractionated sufficiently so that all the bands could be resolved. Figure 5.7 shows the presence of all ten loci following gel electrophoresis. The presence of each band was subsequently confirmed by

Southern blotting followed by hybridisation using ^{32}P 5′ end labelled secondary target site primers (data not shown).

The EtBr-stained gel (Figure 5.7 B) suggested that the decaplex PCR was "balanced" such that all bands appeared with approximately equal intensity. The bands appearing in both the tetraplex and hexaplex PCRs could be identified in the decaplex lanes, except for four bands. The aforementioned bands manifested as two doublet bands in the digested decaplex samples, labelled 5 & 6 (containing the FCM and HBB amplicons), and 7 & 8 (containing the RP2 and DMD loci; figure 5.7 A). These doublet bands were, however, approximately double the intensity of the single bands shown in the decaplex samples (Figure 5.7 B). The finalised decaplex conditions are detailed in the Materials and Methods.

A.

Band	Locus	Amplicon size (undigested)	Amlicon size (digested)
1	CHM	5678 bp	NA
2	FIX	5552 bp	NA
3	APC	5417 bp	NA
4	CGD	5160 bp	NA
5 and 6	FCM	5019 bp	NA
	HBB	5006 bp	NA
7 and 8	RP2	4774 bp	NA
	DMD	4909 bp	4746 + 162 bp
9	MH2	5018 bp	3699 + 1206 + 94 +18 bp
10	HOD	4286 bp	NA

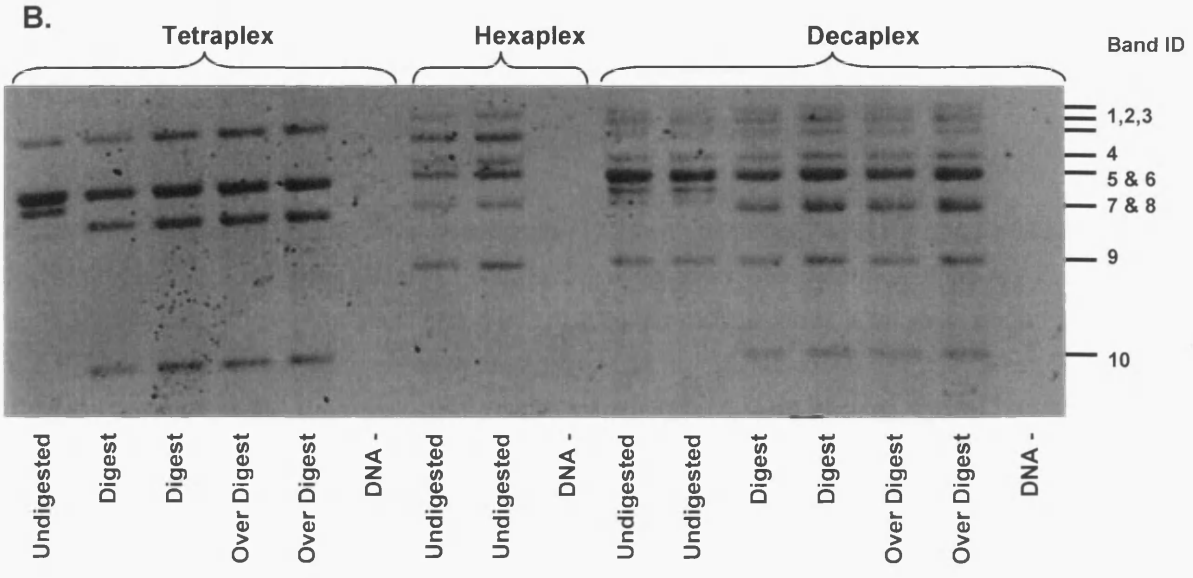


Figure 5.7: A. EtBr-stained gel showing tetraplex, hexaplex and decaplex PCR. Band identity and size of each amplicon both pre and post *Bss*SI-digestion. B. Tetraplex, hexaplex and decaplex PCRs visualised on a 0.8 % agarose gel. DNA negative controls contained water.

5.1.3.a Amplification of the AL121819 insertion in multiplex PCR

Primers PF819LRA and PF819LRB were added to the "balanced" decaplex PCR to amplify the Al121819 L1 insertion (selected site), at 0.05 μM. All other PCR conditions were conserved.

The resulting hendecaplex PCR-generated eleven amplicons when seeded with donor R2 gDNA (heterozygous for the AL121819 insertion). As well as the ten target site amplicons, a 6.2 kb empty site and 12 kb filled site were visible using EtBr staining (Figure 5.7). The filled and empty site bands were not visible in decaplex PCR (Figure 5.7), thus the bands must have been generated as a result of amplification from the AL121819 control primers. As expected, the 12 kb filled site amplicon was less intense than the 6 kb empty site amplicon; however it was visible despite competing with the 11 smaller amplicons. This experiment therefore showed that the AL121819 insertion could be amplified in multiplex PCR, thus generating recoverable selected site molecules.

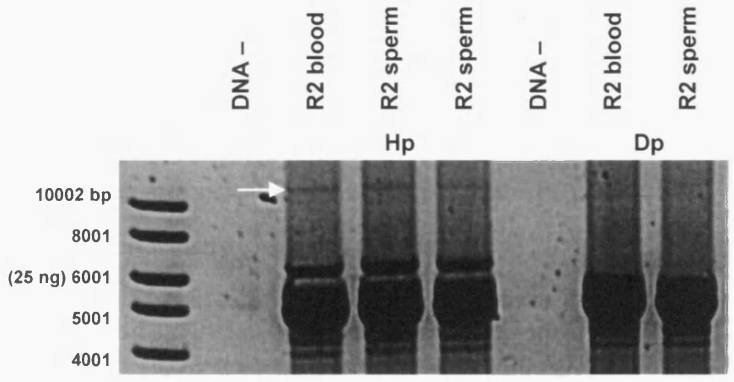


Figure 5.8: EtBr-stained gel showing hendecaplex PCR with primers PF819LRA and PF819LRB (Hp), and decaplex PCR (Dp). DNA negative controls contained water. The presence of the AL121819 filled (> 10 kb; marked with an arrow) and empty site (6.2 kb) amplicons can only be seen when PF819LRA and PF819LRB are used. Samples were run against 250 ng 1kb molecular weight marker. The mass of DNA contained in the 6 kb molecular weight marker band is also shown. This is required in the discussion.

5.2 Discussion

Multiplex PCR was essential to this investigation, as a single 5 kb target locus would only be subject to between 1 and 42 L1 insertions in a single ejaculate (chapter 1). Multiplex PCR allowed multiplication of the overall target length (10 x 5 kb), increasing the likelihood of an L1 inserting into a target region. For example, by amplifying all ten target loci simultaneously, between 10 and 420 insertions would be expected. Without multiplex PCR, numerous ejaculates would have to be screened for *de novo* insertions in each locus individually, which would have been extremely time consuming.

5.2.1 First attempts at multiplex PCR

The failure of the initial tetraplex PCR to simultaneously amplify four target loci (HBB, DMD, FIX and MH2) was due to an excessive overall concentration of primers in the PCR. As the ultimate goal of the hybridisation enrichment protocol was to allow decaplex (ten-plex) amplification, the lowest possible concentration had to be determined by titration. In PCRs containing primer concentrations of 2.4 μM or less, target locus amplicons were visible with EtBr staining. Southern blotting followed by hybridisation with-^{32}P 5′-end labelled secondary target site primers, from each locus, confirmed the presence of all four loci following tetraplex PCR. Amplification was most efficient at a primer concentration of 0.05 μM; as a result subsequent multiplex reactions were carried out at this primer concentration.

The primer titration experiment suggested an upper limit of 2.4 μM total primer concentration before PCR failed. 20 primers would be contained in a single decaplex PCR, therefore each primer had to be added at a concentration of 0.12 μM or less, so as not to inhibit PCR. It was encouraging that efficient tetraplex PCR could be obtained using primer concentrations as low as 0.05 μM. However, efficient amplification of long amplicons (12 to 13 kb) had to be efficient in the presence of numerous competing target site amplicons (5 to 6 kb).

5.2.2 Amplification of 12 kb amplicons in multiplex PCR

To test the capacity of tetraplex PCR to amplify 12 kb amplicons in the presence of competing 5 kb amplicons, a series of primer "drop out" experiments were performed (Figure

5.2 A). This experiment was designed to mimic the amplification of a target site containing a 6 kb insertion, without the need to generate an L1-containing PCR product (a serious contamination risk), under multiplex conditions.

"Drop out" PCR was performed on all four loci, and showed that all the tetraplex primers were capable of generating control amplicons in the presence of competing shorter amplicons though at a lower yield. Hybridisation with labelled secondary target site primers confirmed the presence of 5 kb amplicons where a target site primer had not been replaced, and a 12 kb amplicon when a target site primer had been replaced. This confirmed that extended amplicons were being amplified, in the presence of three other competing target site loci, though at a lower yield.

5.2.3 The maximum capacity of multiplex PCR

Between 10 and 420 *de novo* L1 insertions were expected in the target loci in a single human ejaculate, if all ten loci could be amplified simultaneously. Screening such a huge mass of DNA (~ 600 μg) would be extremely time-consuming, therefore to minimise the overall number of PCRs, the maximum mass of DNA amplifiable in multiplex PCR was determined.

PCR involves the rapid amplification of molecules, and thus if the initial number of target molecules is extremely high, as was the case for excessive masses of gDNA, limiting factors (such as insufficient dNTPs) quickly become problematic. By limiting the mass of gDNA used to seed a PCR, more cycles can be performed before limiting factors become scarce. This ensures that a single L1-containing target locus can be amplified numerous times during the exponential phase, where long (12 kb) molecules theoretically have an equal probably of amplifying compared to short (5 kb) molecules. Although it was necessary to maximise the amount of gDNA in a PCR and reduce the number of PCRs required to screen 600 μg of gDNA, it was necessary to avoid overloading the amplification system.

The maximum capacity of a 10 μl tetraplex PCR was determined as 100 ng gDNA (extrapolated from the 20 cycle data in figure 5.4). For the final protocol, gDNA would be amplified in 50 μl PCRs, so by extrapolation from the 10 μl reaction, the maximum capacity for multiplex PCR was theoretically 500 ng of gDNA in a 50 μl PCR reaction. This was confirmed by the experiment shown in figure 5.4 B.

5.2.4 Balancing the multiplex PCRs

The target site amplicons generated in tetraplex PCR were all of similar size (~5 kb). To allow the amplicons to be resolved by gel electrophoresis, restriction enzyme recognition sites were identified. *Bss* SI was used to digest tetraplex PCR prior to gel electrophoresis and allow resolution of the bands. Subsequent tetraplex PCRs were performed, digested with *Bss*SI and fractionated until the amplicons were of approximately equal intensity.

The hexaplex PCR was also "balanced" by performing successive amplifications and altering primer concentrations until the amplicons appeared at approximately equal intensities following gel electrophoresis. The only problematic amplicon was HOXD. This amplicon is extremely GC-rich (50 %) compared to the average base composition of the other five amplicons (34 % to 43 %), and amplified less efficiently.

The ultimate goal of multiplex PCR was to amplify all ten loci simultaneously. Fortunately the hexaplex amplicons all lacked *Bss*SI restriction sites. Once "balanced", the tetraplex and hexaplex optimal primer concentrations were combined to perform decaplex PCRs. The overall primer concentration of decaplex PCR came to 1.59 μM, well within the maximum primer concentration of 2.4 μM for multiplex PCR.

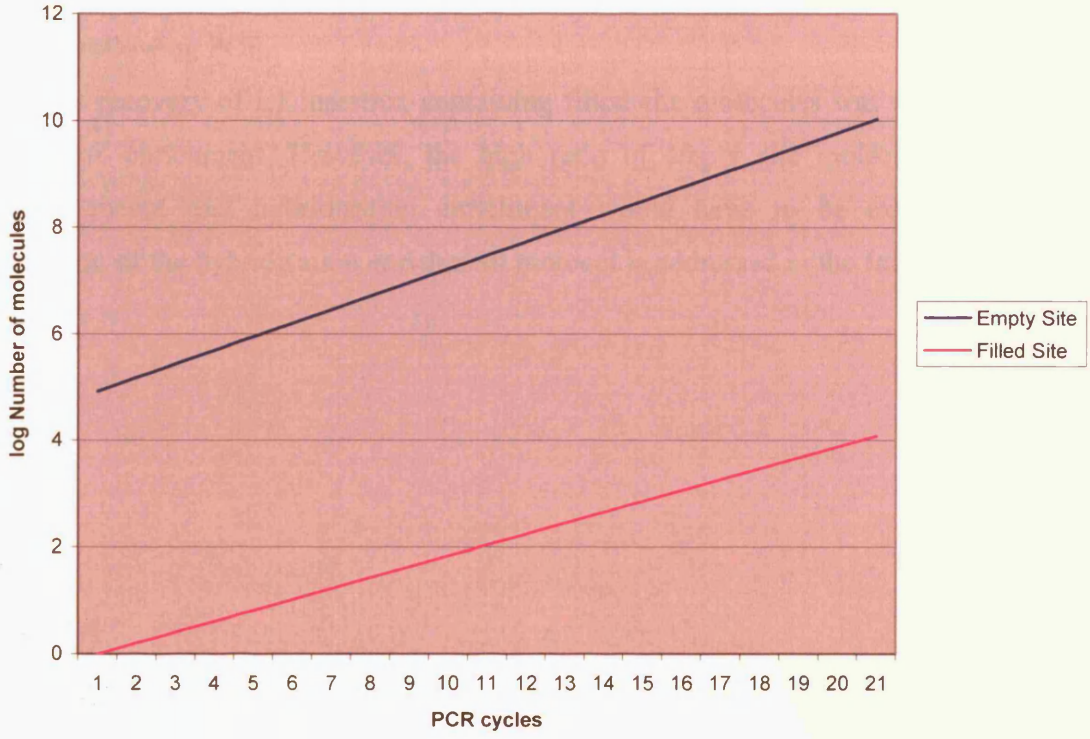
The experiments detailed in this chapter have demonstrated that not only was multiplex PCR possible, but also that efficient decaplex amplification could be performed. It was also established that a single 50 μl decaplex PCR reaction could amplify 500 ng of gDNA. This meant that to amplify 600 μg of gDNA (one human ejaculates worth of DNA), twelve and a half 96-well plates would be required (1200 individual reactions). However, a potential problem remained and had to be addressed.

5.2.5 Amplification of a full-length L1 in multiplex PCR

A *de novo* L1 insertion into one of the target site loci will be a single-molecule contained in millions of competing empty site target sites. Recovery of the L1 insertion by hybridisation enrichment was dependent on generating numerous L1-containing molecules during multiplex PCR. It was therefore essential that a full-length L1 insertion could be efficiently amplified in multiplex PCR.

The hendecaplex PCR (Figure 5.8) was seeded with 500 ng of donor R2 gDNA, or 166,667 sperm genomes worth of DNA. Being heterozygous for the AL121819 insertion, half (83,334) of the sperm contained the filled site and half contained the empty site. The human

haploid genome is comprised of approximately 3 x 10^9 bp, so there were 5 x 10^{14} bp of DNA in the 500 ng of sperm gDNA used to seed the hendecaplex PCR. Each bp therefore had a mass of 1 x 10^{-9} pg. Therefore, the 6455 bp empty site had a mass of 6.46 x 10^{-6} pg, and the12,455 bp filled site had a mass of 1.25 x 10^{-5} pg. The single-molecule amplification efficiency of the AL121819 insertion locus was determined in chapter 2. Since the diploid genome size is approximately 6 pg, so the single-molecule amplification efficiency for the empty site was 107 % (1 molecule in 5.6 pg of donor R2 DNA). This meant that all 83,334 empty site molecules of the empty site contained in 500 ng of sperm gDNA were amplifiable. However, the single-molecule amplification efficiency for the filled site was 52 % (1 molecule in 11.5 pg of donor R2 DNA), so only 43,334 molecules of the filled site were amplifiable.



Graph 5.9: A graph showing the log of the increase in empty site and filled site molecules during 20 cycles of PCR. The value of g_e was 1.8 and g_f was 1.6.

Approximately 50,000 pg (7.74 x 10^9 molecules) of the AL121819 insertion empty site were generated during 20 cycles of PCR (25 ng x 2 since 1/2 of the sample was loaded on the gel) (Figure 5.8). The per cycle gain of the empty site (g_e) was approximately 1.8 ($\sqrt[20]{7.74 \times 10^9}$ / 83,334). It was estimated that after 20 cycles of PCR, at worst 8-fold more empty site molecules were generated than filled site (chapter 2). Therefore an estimated

6,250 pg (5.00 x 10^8 molecules) of the AL121819 insertion filled site was generated. The gain per cycle of the filed site (g_f) was at worst 1.6 ($\sqrt[20]{5.00 \times 10^8 / 43{,}334}$).

In the actual experiment, a single L1-containing filled site molecule will be contained in approximately 83,300 molecules of the corresponding empty site. Following 20 cycles of PCR, at worst 12000 filled site molecules would be generated. It should also be noted that the AL121819 insertion filled site molecules have competed with ten other empty site loci. The result is encouraging as it shows efficient filled site amplification in hendecaplex PCR. However, the imbalance in g_e and g_f means that the empty site amplifies more efficiently that the filled site (Graph 5.9). Following PCR, the 12,000 filled site molecules will be contained in 1 x 10^{10} corresponding empty site molecules. It can be seen in graph 5.9 that the ratio between filled and empty site molecules gradually increases. However, the predicted value for g_f is likely to be an under estimate, thus decreasing the ratio between the empty and filled site molecules following PCR.

The recovery of L1 insertion containing filled site molecules was to be achieved by hybridisation enrichment. However, the high ratio of empty site molecules to filled site molecules meant that hybridisation enrichment would have to be extremely efficient. Optimisation of the hybridisation enrichment protocol is addressed in the following chapter.

Chapter 6: Optimisation of L1 hybridisation enrichment

6.1 Results

Despite the success of optimising multiplex PCR, L1-containing amplicons generated from novel insertions will still be massively under-represented compared to empty target site amplicons. The power of hybridisation enrichment lies in its ability to discriminate between the selected L1-containing molecules, and unselected non-L1-containing molecules. However, before efficient enrichment could be achieved, numerous conditions had to be optimised.

6.1.1 Assessment of L1 annealing by L1 selector oligonucleotides

Prior to enrichment, it was necessary to determine whether the bio-oligos were capable of binding specifically to L1 sequences. An L1-containing amplicon was generated by PCR across the AL583853 L1 insertion using primers RB853E and RB853D. The FIX locus was amplified using its primary target site primers, to generate an amplicon lacking matches to the bio-oligo sequences (chapter 3).

Decreasing masses of L1-containing DNA, and 375 ng of FIX DNA were dot-blotted (Materials and Methods) and sequentially hybridised with ^{32}P-5′-end labelled non-biotinylated versions of the bio-oligos (L1U1, L1U2, L1U3 and L1U4: Figure 6.1).

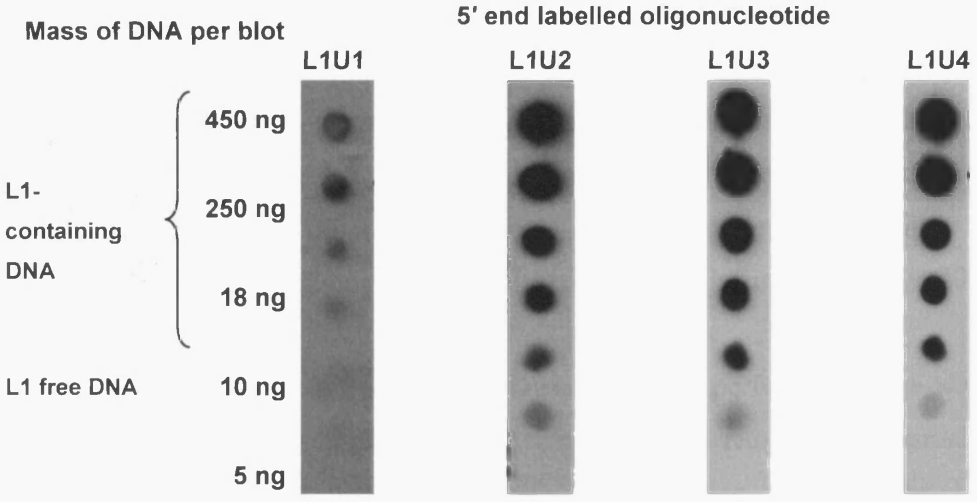


Figure 6.1: Autoradiograph of dot-blotted L1-containing DNA, amplified from the AL853853 L1 insertion, hybridised with ^{32}P 5′ end labelled L1U1 to L1U4.

All four oligonucleotide probes hybridised to the L1-containing dots, with a very faint signal on the FIX dots, establishing selective hybridisation to the desired target. Since the

bio-oligo sequences were capable of annealing to L1 DNA, optimisation of the hybridisation enrichment protocol could commence.

6.1.2 Introduction to hybridisation enrichment

6.1.2.a Summary of hybridisation enrichment

The hybridisation enrichment method is described here in general terms to make the optimisation process easier to understand. The method was derived from the DEASH protocol (Jeffreys and May, 2003). The stages of the method are described in the text, and illustrated in figure 6.2. The final detailed protocol for hybridisation enrichment is given in the Materials and Methods.

Annealing was carried out using a GeneAmp PCR system 9600 thermal cycler (Perkin Elmer Cetus). Annealing reactions contained the sample DNA, bio-oligos and 1 x denaturing/hybridisation buffer (DHB). DHB is a modified buffer based on a standard PCR buffer (Materials and Methods), used to minimise damage to target DNA when denaturing at high temperatures (Jeffreys and May, 2003).

Annealing mixes were denatured at 96 °C for 75 sec, and then the reaction was cooled to 9 °C above the annealing temperature and held at this temperature for 20 sec. The temperature was then reduced in 1 °C increments and held at each temperature for 20 sec until the annealing temperature was reached. Samples were held for two min at the annealing temperature. This wide "annealing window" allowed time for the bio-oligos to anneal to any single-stranded L1-containing (selected site) molecules (Jeffreys and May, 2003).

Binding of the bio-oligos to the Dynabeads was performed in pre-warmed siliconised tubes in a water bath set to the annealing temperature, to reduce the potential for non-specific bio-oligo annealing to unselected sites. After binding, the Dynabeads plus captured bio-oligos, were separated using a magnetic particle concentrator, and the remaining solution (containing unbound DNA) transferred to a fresh 0.2 ml PCR tube for re-extraction.

Washing steps were performed to remove unbound and loosely bound DNA. Following each wash, the Dynabeads were separated from the wash solution using a magnetic particle concentrator and the wash solution retained for further analysis.

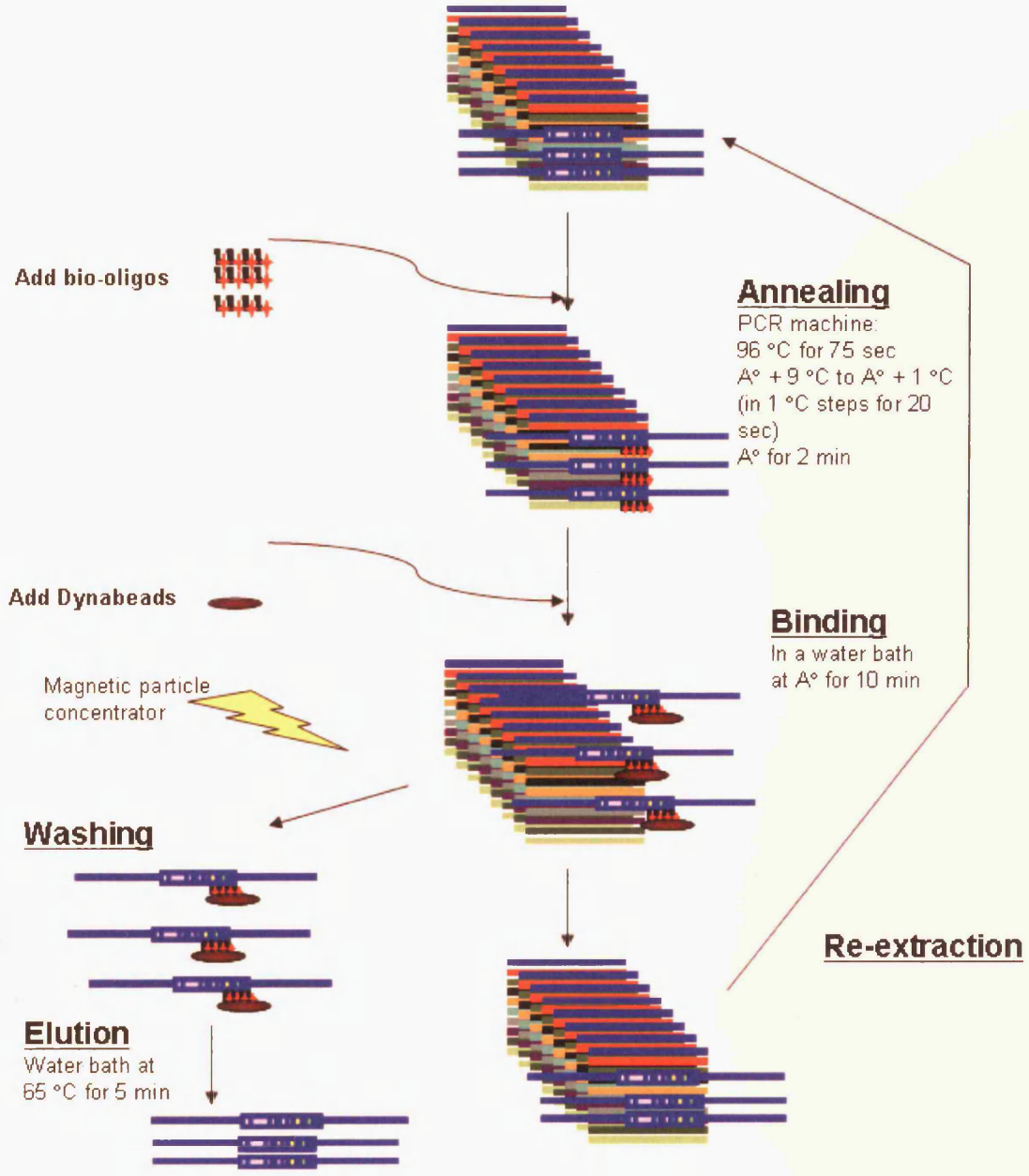


Figure 6.2: The process of recovering L1 insertions from multiplex PCR-amplified DNA. For details see the text (pages 129 - 131).

The Dynabeads were first washed in DHB/BSA on ice, then at the annealing temperature for 1 min, to remove loosely annealed target molecules. The Dynabeads were then washed on ice in elution mix (ED), and re-suspended in ED prior to thermal elution.

Thermal Elution of single-stranded DNA from the bead-bound bio-oligos was performed at 60 - 65 °C for 5 min. The Dynabeads were removed from solution using a magnetic particle concentrator, and the eluate recovered.

Re-extraction. Only a fraction of the L1-containing molecules are recovered in a single extraction, thus the remaining unbound fraction could be subjected to re-extraction. Samples were subjected to repeated denaturing, annealing, washing and thermal elution. In total, 1 extraction and 2 re-extractions were carried out per sample. The final un-bound fraction was retained in a 1.5 ml micro-centrifuge tube. The 65 °C eluate from the extraction and re-extractions were pooled in the same siliconised 1.5 ml micro-centrifuge tube.

6.1.2.b Amplicons used during hybridisation enrichment optimisation

Optimisation of the hybridisation enrichment protocol used two interchangeable full-length L1 insertion containing selected sites (as indicated for each experiment) which were generated by PCR, and purified as indicated.

6.1.2.b.i AL583853 selected site

The AL583853 insertion selected site was generated using primers RB853E and PFENopF_PVU2D (12.04 kb). Recovery of the AL583853 insertion selected site was assessed by PCR using primers PFOP_ENfA and PFOP_ENfB (~2.2 kb).

6.1.2.b.ii AL121819 selected site

The AL121819 selected site was generated by PCR using primers PF819LRA and PF819LRB (~12.45 kb). Recovery of the AL121819 insertion was assessed by PCR using primers RB819C and RBLR2519 (2.5 kb). The AL121819 insertion selected site was more

amenable to optimisation of the enrichment protocol, so was predominantly used for the latter stages of optimisation.

6.1.2.b.iii Unselected sites

Two unselected sites, devoid of close matches to the bio-oligo sequences (chapter 3), were also employed during optimisation of the enrichment protocol, both generated from the same locus (Accession AC008575; chromosome 5 22.2). A 12 kb unselected site (E12) was generated by PCR using primers PFENopE_PVU2A and PFENopE_PVU2B (11.97 kb: Figure 6.3). This unselected site amplicon was designed to be approximately the same size as the selected site loci (AL583863 and AL121819). Also, a 5 kb unselected site amplicon (E5) was generated by PCR using primers PFOPenE_PVU2A and PFOP_ENeE (4.97 kb: Figure 6.11). Empty target loci generated during multiplex PCR varied between 4.5 kb and 6 kb in length, and so the 5 kb empty site was more representative of standard unselected sites. Recovery of both E5 and E12 was assessed by PCR using primers PFOP_ENeC and PFOP_ENeD (2.94 kb) (Figure 6.3).

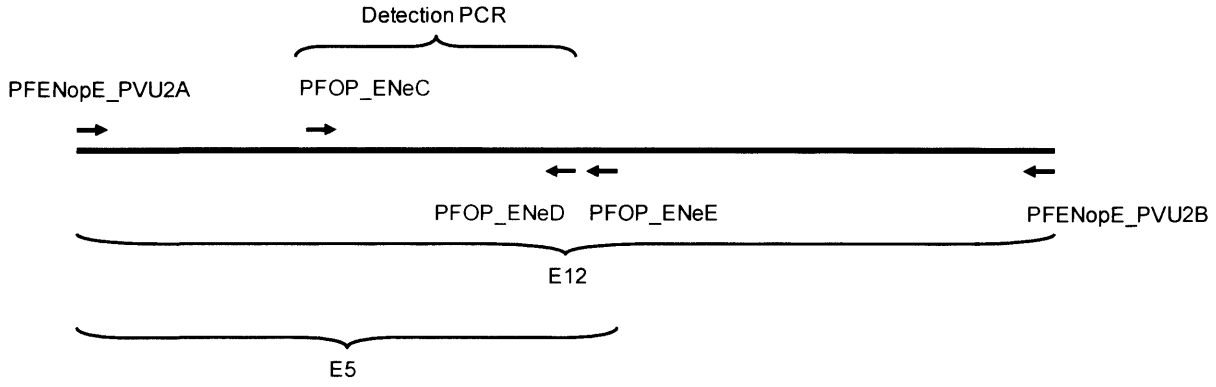


Figure 6.3: Diagram showing the locations the primers used to generate unselected sites E5 and E12. The location of the unselected site detection PCR is also shown.

6.1.2.c Initial enrichment conditions and estimation of selected site recovery

To test the initial enrichment conditions, enrichment was performed using a mixture of the AL583853 selected site, and the E12 unselected site. 20 µl enrichment reactions contained: 5 pg of the AL583853 selected site; 5 pg of the E12 unselected site; 1 x DHB; and one of the four bio-oligos (0.375 µM). Annealing and binding were performed at 35 °C, and

binding was performed using 7.2 μg/μl of Dynabeads. The annealing temperature wash was performed for 1 minute, and recovered DNA was eluted at 60 °C and 80 °C.

Following enrichment, detection PCRs were performed on 5 μl of the room temperature wash, the annealing temperature wash, the 60 °C eluate and the 80 °C eluate. For the L1 selected site only single-stranded molecules are bound by bio-oligos, so a 100 % yield from a sample containing 10 pg of the selected site equated to the recovery of 5 pg (assuming 0 % recovery of the complementary strand by non specific annealing). Since unselected site recovery was not strand-specific the recoverable yield from an input of 10 pg was 10 pg. Detection PCR using a sample of the unbound fraction acted as a positive PCR control due to the presence of the un-recovered complementary L1 strand. PCR products were visualised by Ethidium bromide (EtBr) staining and the percentage recovery estimated by comparison of the 60 °C eluate amplicon, and the amplicons in a control dilution series (indicated using boxes and lines, for an example see figure 6.4).

6.1.3 Hybridisation enrichment of L1-containing molecules with 3′ modified bio-oligos

Using the above conditions, enrichment was performed with 3′ modified bio-oligos. These were designed with protected 3′ ends, intended to prevent primer extension by *Taq* DNA polymerase (chapter 3). The 3′ modifications meant that in principle bio-oligos could be used directly in post PCR mixes without *Taq* DNA polymerase affecting the dynamics of hybridisation enrichment.

However, detection PCR performed on the 60 °C eluates and 80 °C eluates showed no amplification from any enriched sample, indicating a yield of < 1 % (data not shown).

6.1.4 Are non-3′ modified bio-oligos capable of recovering L1-containing DNA?

As indicated above, the 3′ modified bio-oligos were not capable of enriching L1-containing selected site DNA. Enrichment was therefore attempted using un-modified bio-oligos.

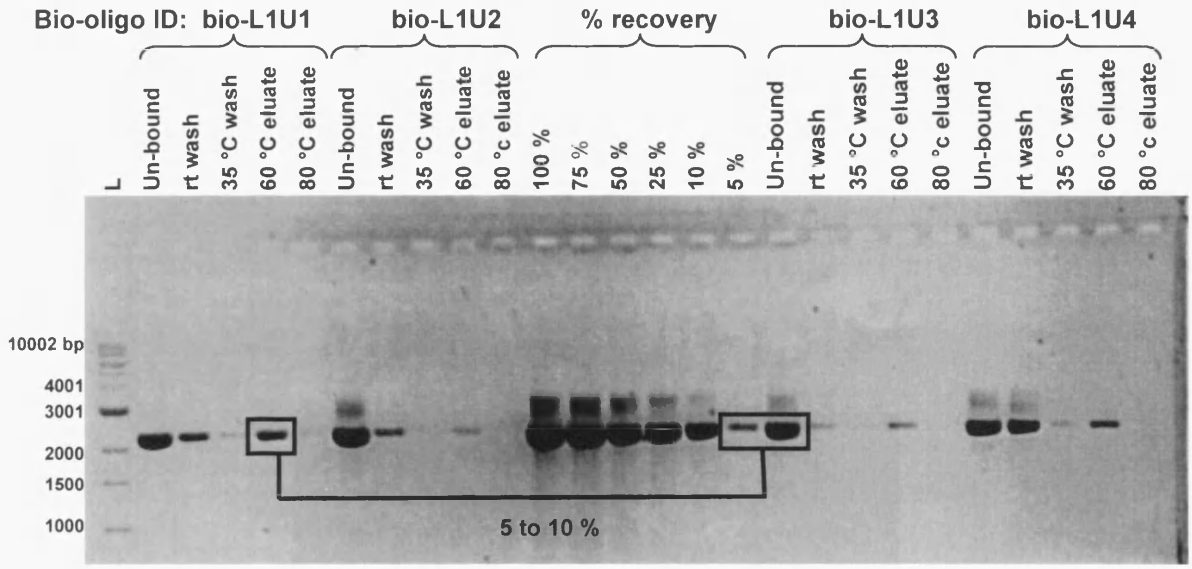


Figure 6.4: EtBr-stained gel showing filled site recovery following enrichment with bio-L1U1b, bio-L1U2b, bio-L1U3b and bio-L1U4b. The un-bound DNA acts as a positive control for detection PCR. Room temperature (rt), and annealing temperature (35 °C) washes are shown. Samples were run alongside 250 ng of 1 kb molecular weight marker.

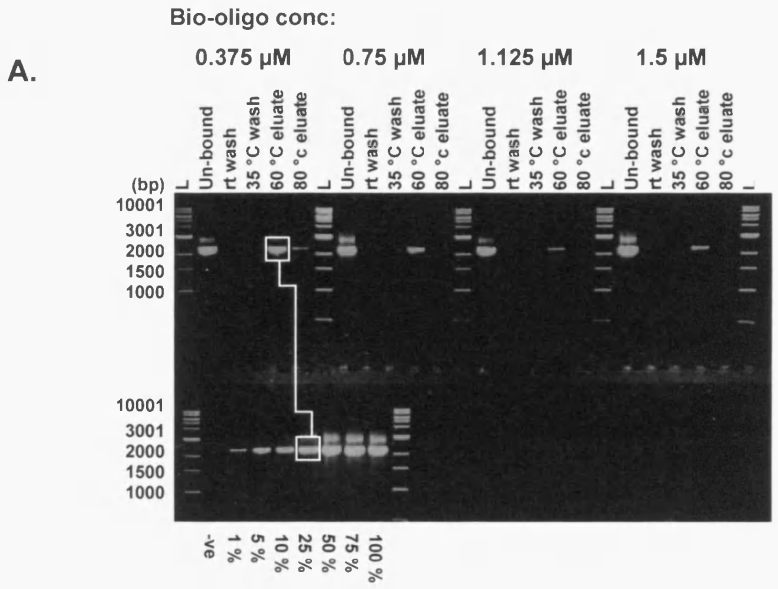
The AL583853 selected site and E12 unselected site were generated by PCR and subsequently digested with *Pvu* II. This resulted in 9.2 kb and 11.4 kb restriction fragments, respectively. The digested DNAs were then purified using a QIAquick PCR clean up kit (Qiagen).

Each bio-oligo was used individually to enrich samples containing 10 pg of the AL583853 insertion selected site and 10 pg of the E12 unselected site (using the initial enrichment conditions, pg 132). Between 5 and 10 % of the selected site was recovered using bio-L1U1b and bio-L1U4b but less than 5 % was recovered using bio-L1U2b and bio-L1U3b (Figure 6.4). These results suggested that the previous failure to enrich selected site DNA was likely due to a fault with the 3′ modified bio-oligos.

6.1.5 Increasing selected site recovery by bio-oligo mixing

The above experiment showed that although selected site DNA could be recovered by hybridisation enrichment, the level of recovery was low. Also, there was a marked discrepancy in the level of recovery achieved by the four bio-oligos (> 5 % by bio-L1U1 &

bio-L1U4, and < 5 % by bio-L1U2 and bio-L1U3: Figure 6.4). This discrepancy was unexpected as all four bio-oligos were designed to be similar in terms of base composition and structure. Long single-stranded molecules (for example the 12 kb selected site) can form complex secondary structures, and this may have reduced the availability of the bio-L1U2 and bio-L1U3 annealing sites.



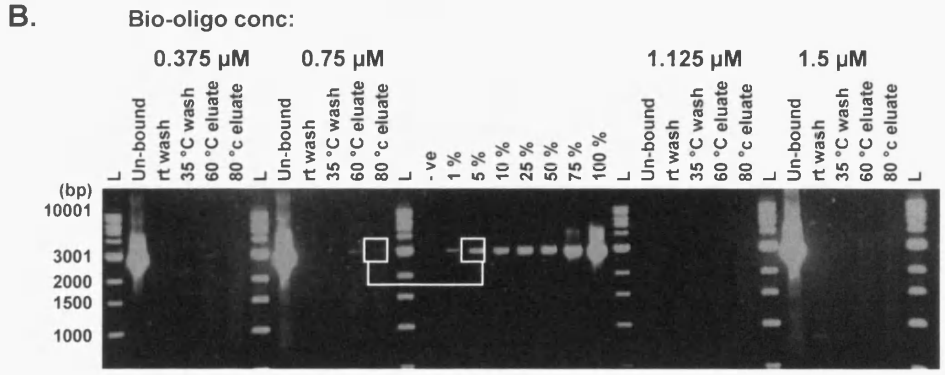


Figure 6.5: A. EtBr-stained gel showing recovery obtained by enrichment using all four bio-oligos mixed at a 1:1:1:1 ratio at final concentrations of 0.375 µM, 0.75 µM, 1.125 µM and 1.5 µM. B. EtBr-stained gel showing the recovery of unselected site recovery. Molecular weight marker is 250 ng of 1 kb molecular weight marker.

6.1.6 Investigation of the effects of primer extension on enrichment dynamics

The previous enrichments were performed used column-purified PCR products to avoid issues with primer extension. However column purification was not a viable option for

large scale recovery of amplified DNA, which would be required for the full scale experiment, so an alternative approach was required.

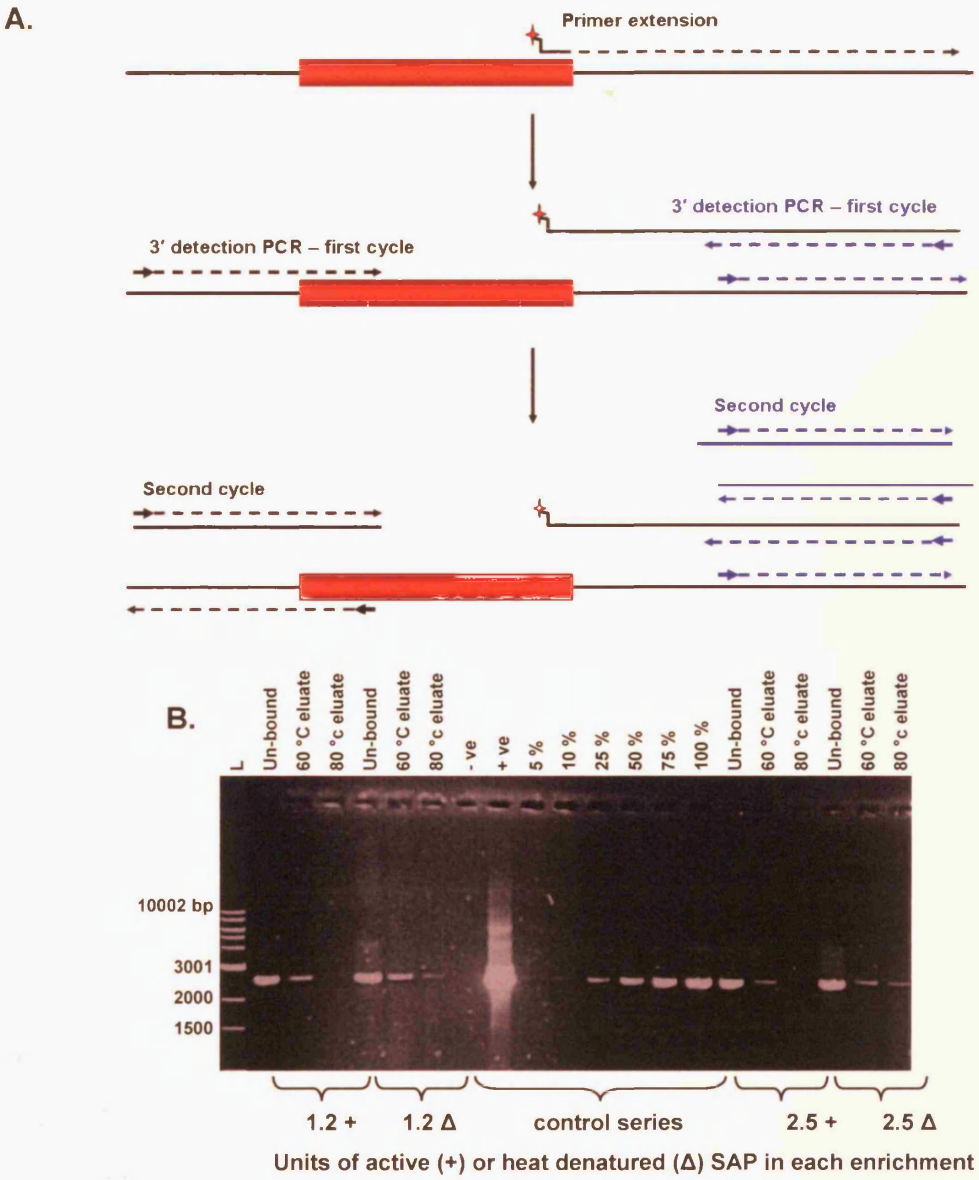


Figure 6.6: A. Diagram showing the effects of primer extension on detection PCR. 3' extension potentially contributes molecules amplifiable by the detection PCR resulting in overestimation of selected site recovery, so a 5' located detection PCR was used. B. EtBr gel showing side by side comparison of enrichment samples incubated with either SAP (+) or heat denatured SAP (Δ) prior to enrichment. 5' detection PCR was used to generate the amplicons. Samples were run alongside 250 ng of 1 kb molecular weight marker.

It was anticipated that primer extension, during the annealing stage of enrichment, would cause the bio-oligos to bind more tightly to annealed molecules. This would act to

increase the temperature required to elute these molecules, resulting in inefficient recovery of DNA in the 60 °C eluate, and significant recovery in the 80 °C eluate. High temperature elution also ran the risk of denaturing the streptavidin coating the Dynabeads, and releasing bio-oligos into the eluates.

To test whether primer extension did affect hybridisation enrichment efficiency, PCR-amplified samples were treated with either active or heat inactivated Shrimp Alkaline Phosphatase (SAP) (Roche), prior to enrichment. SAP dephosphorylates dNTPs, thus preventing their incorporation in DNA (Manual provided with the enzyme: Roche, UK). Duplicate reaction mixes were prepared containing: 1 x SAP reaction buffer (Roche) 1 x DHB, 1 µl of the PCR-amplified DNA mix and either 1.2 U, or 2.5 U of SAP. One of the two duplicate samples was incubated at 65 °C for 25 min to denature the SAP (as recommended by the manufacturer), and the other kept on ice. Both samples were then incubated at 37 °C for 30 min. 0.375 µM of mixed bio-oligos were then added prior to enrichment. The detection PCR used primers RB853H and RBLR2519 (~2.5 kb), which were located 5′ of the bio-oligo sequences to avoid interference by primer extension from the bio-oligos (Figure 6.6 A).

Samples which had been incubated with heat denatured SAP showed amplification from the 80 °C eluate, but samples incubated with active SAP showed very little amplification from the 80 °C eluate (Figure 6.6 B). This suggested that the SAP treatment prior to enrichment was acting to reduce 3′ extension of the bio-oliogos, and so confirmed that 3′ extension of the bio-oligos by *Taq* DNA polymerase does influence hybridisation enrichment kinetics.

It was therefore deemed necessary to remove *Taq* DNA polymerase from the samples generated by PCR. This was ultimately achieved by extracting the PCRs using phenol chloroform, concentrating the DNA using ethanol precipitation, and dissolving in 5 mM Tris-HCl (pH 7.5) (Materials and Methods). All amplicons used in subsequent enrichment experiments were purified in this way.

6.1.7 Refining the optimal annealing temperature

Having achieved hybridisation enrichment of purified selected site molecules, it was necessary to determine a set of optimal conditions that gave a high recovery of the selected

site, but minimise unselected site recovery. The first variable addressed was the optimal annealing temperature.

Enrichment was performed using various temperatures on samples containing 10 pg of the AL121819 insertion selected site and 4.1 ng of the E5 unselected site (a ratio of ~1 molecule: 10,000 molecules). When 48 °C was the annealing temperature, the selected site recovery was 15 % - 25 % (Figure 6.7 A), and the recovery of the E5 unselected site was 0.0025 % to 0.01 % (Figure 6.7 B). This gave an enrichment efficiency of 150 to 6,000 fold. Using 51 °C as the annealing temperature, selected site recovery was ~15 % (Figure 6.7 A), and E5 unselected site recovery was ~0.005 % (Figure 6.7 B). This gave an enrichment efficiency of ~3,000 fold.

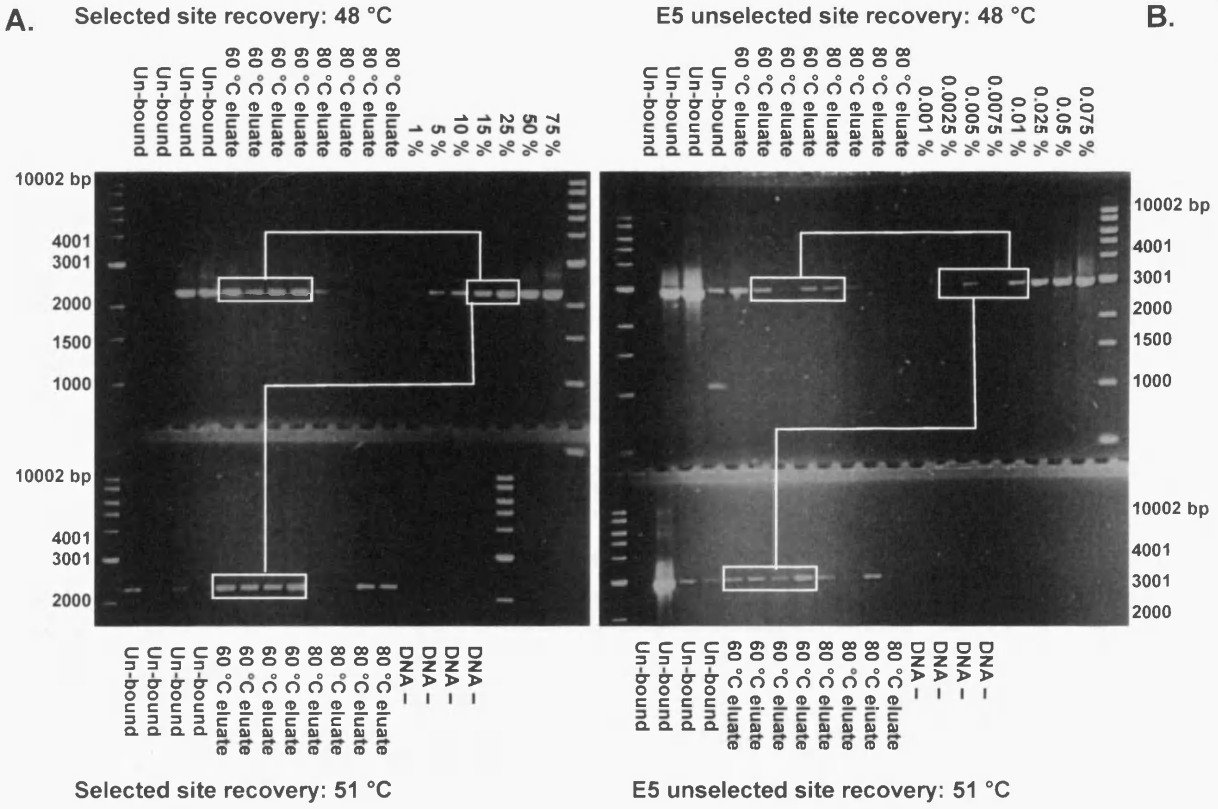


Figure 6.7: A. EtBr-stained gel showing selected site recovery when annealing was performed at 48 °C (top) and 51 °C (bottom). B. EtBr-stained gel showing E5 unselected site recovery when annealing was performed at 48 °C (top) and 51 °C (bottom). Samples were run alongside 250 ng of 1 kb molecular weight marker.

Annealing at 51 °C gave stable and high enrichment efficiency due to unselected site recovery remaining consistently low across all four enrichments (Figure 6.7 B). However, the

selected site recovery was a little higher when annealing at 48 °C. Since it was necessary to maximise selected site recovery, 48 °C was adopted as the optimal annealing temperature, since it was possible that enrichment efficiency could be stabilised by consistently minimising unselected site recovery.

6.1.8 Stabilising enrichment efficiency by increased washing

To minimise unselected site recovery, the length of time for which the annealing temperature wash was performed was investigated. The majority of non-specifically annealed unselected site molecules were removed during the annealing temperature wash, so enrichment efficiency could potentially be increased by increasing the duration of the annealing temperature wash. Enrichment was therefore performed using different durations of washes at the annealing temperature of 48 °C.

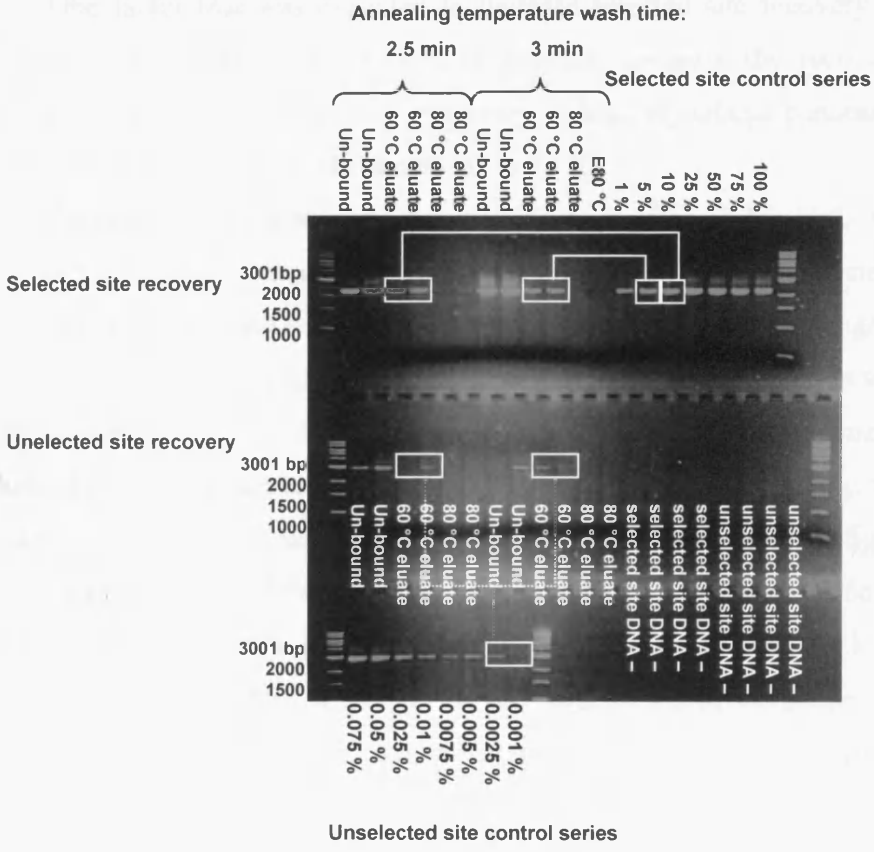


Figure 6.8: EtBr-stained gel showing selected site recovery and E5 unselected site recovery using annealing temperature washes of 2.5 min and 3 min. Samples were run alongside 250 of 1kb molecular weight marker.

By washing at the annealing temperature for 2.5 min, selected site recovery was ~10 %, and unselected site recovery 0.001 % to 0.0025 % (Figure 6.8). Enrichment efficiency was therefore between 4,000 and 10,000 fold (Figure 6.8). Although selected site recovery had decreased from ~15 % (Figure 6.7 A) to 10 % (Figure 6.8), suggesting that some recovered selected site molecules were loosely bound, enrichment efficiency had increased to > 4,000 fold.

Washing at the annealing temperature for 3 min made no difference to selected site recovery or enrichment efficiency. This suggested that washing at the annealing temperature for 2.5 min had maximised the removal of the unselected site. Subsequent enrichments used a 2.5 min annealing temperature wash.

6.1.9 Assessing the effect of increased Dynabead concentration

One factor that was expected to increase selected site recovery was increasing the Dynabead concentration. This should in principle increase the recovery of L1-annealed bio-oligos. To increase the elution efficiency at high Dynabead concentrations, the elution temperature was also increased from 60 °C to 65 °C.

Previous enrichments had been performed using 7.2 μg/μl, so enrichment was performed on samples containing 10 pg of the AL121819 insertion selected site and 4.1 ng of the E5 unselected site, at Dynabead concentrations of 9 μg/μl and 10.8 μg/μl.

Increasing Dynabead concentration to 10.8 μg/μl, and the elution temperature to 65 °C slightly increased the selected site recovery when compared to enrichment using 9 μg/μl of Dynabeads. Selected site recovery was > 10 % (Figure 6.9). Surprisingly, the alterations made no obvious difference to unselected site recovery when compared to figure 6.8. Unselected site recovery was ~0.001 % to 0.0025 % (Figure 6.9). Enrichment efficiency was therefore 4000 to 10,000 fold. Subsequent enrichments were performed using 10.8 μg/μl of Dynabeads and an elution temperature of 65 °C. At this stage of the investigation, a set of the optimal conditions for hybridisation enrichment had been selected. However there was one remaining adaptation of the protocol that could increase the selected site recovery without altering any of the previously optimised selected enrichment conditions.

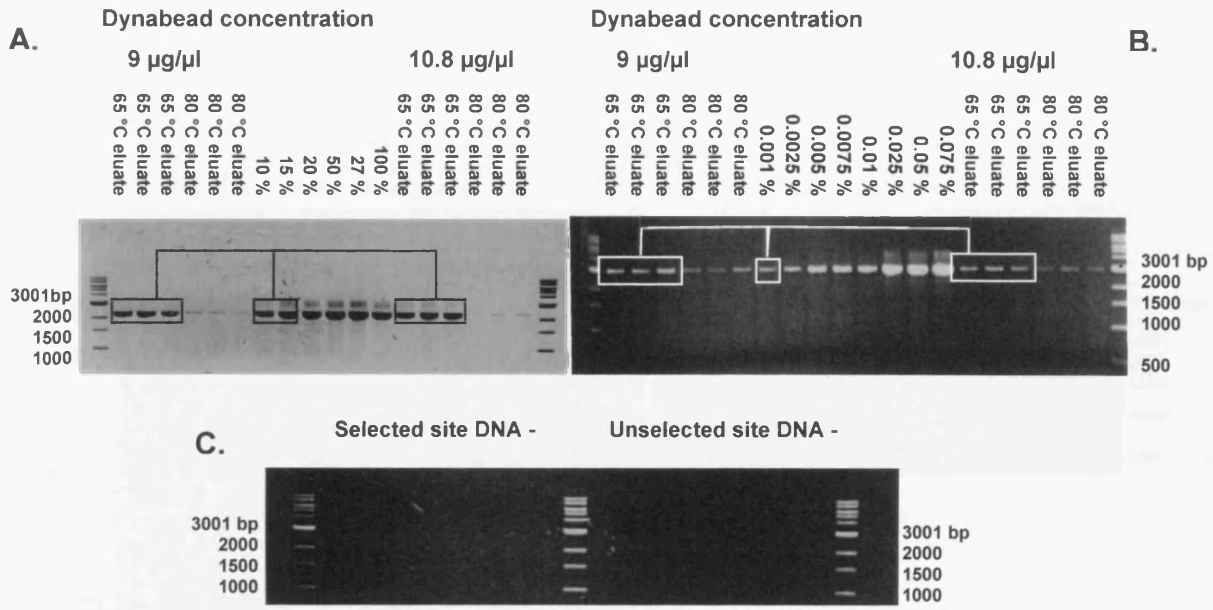


Figure 6.9: A. EtBr-stained gel showing selected site recovery at 9 µg/µl and 10.8 µg/µl Dynabead concentrations. B. EtBr-stained gel showing unselected site recovery at 9 µg/µl and 10.8 µg/µl. C. DNA free negatives relating to A and B. Samples were run alongside 250 ng of 1 kb molecular weight marker.

6.1.10 Increasing selected site yield using multiple rounds of enrichment

Under the optimised conditions, a single round of enrichment recovered ~10 % of the selected site molecules. Enrichment could therefore be re-performed on the unbound fraction that contained the remaining (~90 %) selected site molecules. This re-extraction would most likely increase the yield of the selected site.

To assess the effects of multiple rounds of re-extraction on selected site recovery and enrichment efficiency, samples containing 10 pg of the AL121819 insertion selected site and 4.1 ng of the E5 unselected site were extracted three times. The un-bound fraction was put into a fresh 0.2 ml PCR tube containing 1.5 µl 5 µM bio-oligo mix (1:1:1:1 ratio) and 0.7 µl of 4 x DHB. The eluates were not pooled so that selected site and unselected site recovery could be assessed following all three extractions.

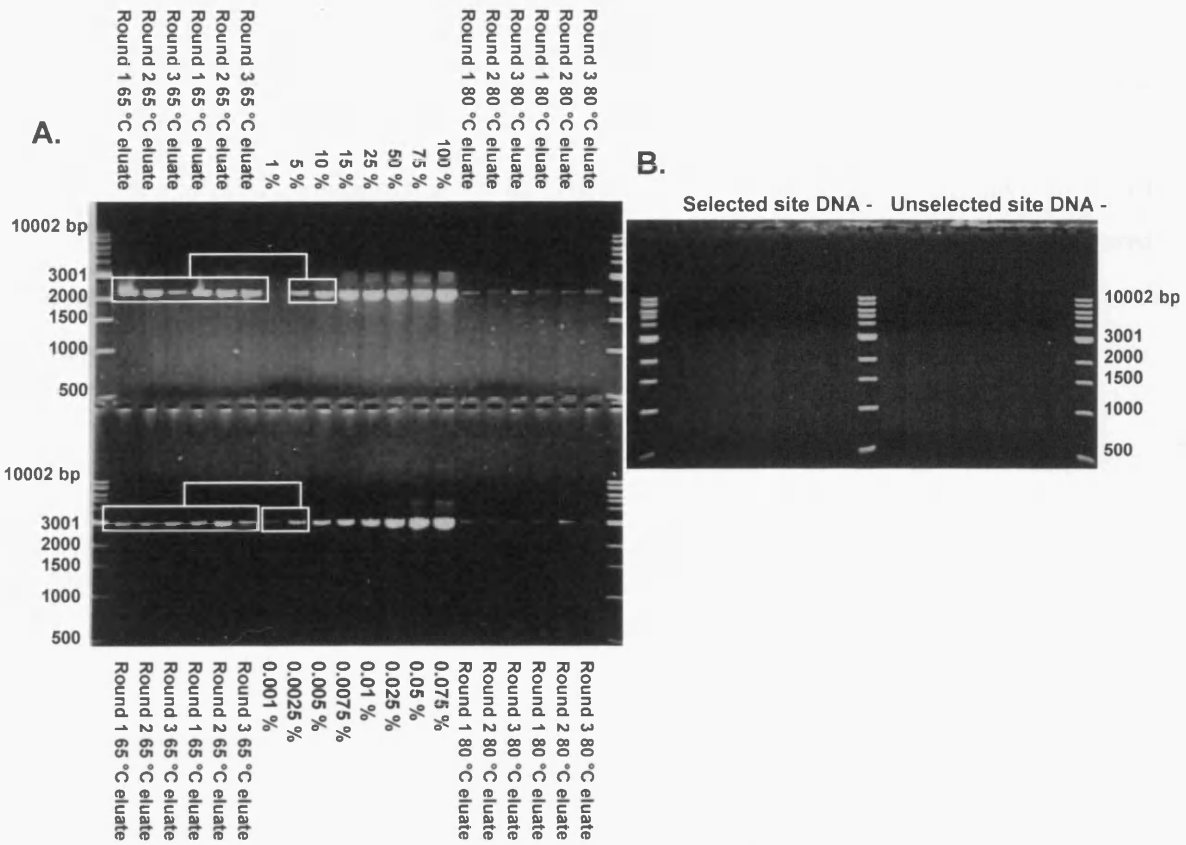


Figure 6.10: A. EtBr-stained gel showing selected site recovery (top) unselected site recovery (bottom), from duplicate enrichments, at 1, 2 and 3 rounds of extraction. B. DNA free negative controls relating to A. Samples were run alongside 250 ng of 1 kb molecular weight marker.

Figure 6.10 shows that for the first two extractions selected site recovery was ~10 %, and only dropped to between 5 % and 10 % on the third extraction. Overall, total selected site recovery was between 25 % and 30 % Unselected site recovery was 0.001 % to 0.0025 % for the first extraction, 0.001 % to 0.005 % for the second extraction, and 0.001 % to 0.0025 % for the third extraction. Overall enrichment efficiency was therefore between 12,000 and 35,000 fold.

6.1.11 Optimised hybridisation enrichment conditions

Following optimisation of the protocol, enrichment was performed in 20 µl reactions containing target DNA, 1 x DHB, and all four bio-oligos mixed in an equal ratio at a final

concentration of 0.375 μM. The mix was denatured at 96 °C for 75 sec followed by step-down annealing, by 1 °C steps, from 57 °C to 49 °C. Annealing was completed in a final step of 48 °C for 2 min.

The binding stage of the experiment was carried out at 48 °C in pre warmed siliconised micro-centrifuge tubes. Binding used Dynabeads at a final concentration of 10.8 μg/μl. Samples were washed in 50 μl DHB+BSA at room temperature, then in 50 μl DHB+BSA at 48 °C for 2.5 min, and finally in 50 μl ED at room temperature. Recovered single-stranded L1 DNA was eluted in 50 μl ED at 65 °C for 5 min. Two further extractions were subsequently performed for each enrichment.

6.2 Discussion

An initial set of enrichment conditions was adapted from the DEASH protocol (Jeffreys and May, 2003). By altering individual conditions in successive enrichment experiments, and choosing the best altered variable, an optimal set of enrichment conditions were determined. The goal was to efficiently recover a high yield of L1-containing selected site molecules while limiting the amount of unselected site molecules recovered. This led to a high enrichment efficiency, i.e. a high ratio of recovery of the selected site compared to the recovery of the unselected site.

6.2.1 Site specificity of the L1 selector oligonucleotides

To achieve selective recovery of L1-containing selected site molecules at the expense of non-L1-containing unselected site molecules, the L1 selector oligonucleotides needed to show specificity for L1 DNA. By 5′ end labelling non-biotinylated versions of the bio-oligos, and hybridising them sequentially to a dot blot, specific annealing to L1-containing DNA was demonstrated. The signal obtained by hybridisation to a small mass of L1-containing DNA (5 ng) was far more intense than the signal obtained by non-specific annealing to DNA lacking L1 sequences (375 ng: Figure 6.1). This indicated that the oligos could bind DNA non-specifically which might result in the recovery of unselected site molecules during enrichment. As a result, efforts were made to minimise unselected site recovery during optimisation of the enrichment protocol.

6.2.2 Limitations of 3′ modified bio-oligos

The success of the above experiment suggested the bio-oligos should be able to recover L1 DNA during hybridisation enrichment. Although the bio-oligos were stringently designed to be similar to the bio-oligos developed in the original DEASH protocol (Jeffreys and May, 2003), in terms of base composition and structure, initial attempts at hybridisation enrichment yielded no recovery in either the 60 °C eluate °C or 80 °C eluates. The positive controls demonstrated that amplification to visible levels (stained with EtBr) could be obtained from very small quantities of input DNA (< 0.005 pg of PCR product), indicating that the PCR-based detection method was extremely sensative. Since no visible band was

observed, recovery must have been less than 0.005 pg (< 1 %) used to seed the 1 % control PCR.

These initial attempts at hybridisation enrichment were performed using 5′ biotinylated oligonucleotides that had been 3′ modified to prevent primer extension by *Taq* DNA polymerase (chapter 3). The two most likely explanations for the inability of the modified bio-oligos to recover L1 DNA were that either the 3′ modifications prevented efficient annealing to the L1 target DNA (steric hindrance), or that the 5′ biotinylation had been lost. Steric hindrance by 3′ modifications would likely have reduced selected site recovery by preventing efficient bio-oligo annealing to the L1-containing molecules; however loss of the 5′ biotin would likely have completely prevented selected site recovery due to the inability to recover the bio-oligos with streptavidin-coated Dynabeads. Since the recovery of the selected site was practically zero, of the two explanations, the loss of the 5′ biotin group was most likely.

As a consequence of the above experiment, a set of bio-oligos with no 3′ modifications were ordered from another supplier. It was decided not to include the 3′ modifications as there was still a possibility that they would reduce selected site recovery by steric hindrance. However, since the bio-oligos were not 3′ modified, they were vunerable to 3′ primer extension by *Taq* DNA polymerase activity.

6.2.3 Hybridisation enrichment using non-3′ modified bio-oligos

The un-modified bio-oligos were tested similarly to the 3′ modified bio-oligos, but prior to enrichment the PCR-generated amplicons were purified using a QIAquick PCR clean up kit (Qiagen). This eliminated the possibility of 3′ extension of the bio-oligos. Using conditions similar to those used in the DEASH experiments (Jeffreys and May, 2003), selected site recovery of 5 % - 10% was achieved using bio-oligos bio-L1U1b and bio-L1U2b (Figure 6.2). However, recovery using the remaining bio-oligos was less than 5 %.

Long single-stranded DNA molecules tend to form complex secondary structures in solution. It was therefore possible that the single-stranded selected site DNA was forming secondary structures that prevented bio-L1U2b and bioL1U3b from accessing their annealing sites. Mixing the bio-oligos increased the probability that L1-containing DNA would be bound by the bio-oligo. For instance, if one complementary site was folded into a secondary structure, thus preventing a bio-oligo from reaching it, one of the other sites may have been

exposed, enabling capture. The experiment shown in figure 6.4 demonstrated that a single bio-oligo was sufficient for recovery. Also, an individual single-stranded target containing L1 sequence could be bound by between one and four bio-oligos (assuming the complementary sequences were present). It seemed likely that a single-stranded L1-containing target held by more than one bio-oligo would be more stably bound than the same target held by a single bio-oligo. If this was the case then less selected site target would be lost during the washing stages of enrichment, and this would lead to a higher yield in the eluates.

A 1:1:1:1 mixture of all four bio-oligos recovered between 10 % and 25 % of selected site DNA when used in enrichments at a final concentration of 0.375 μM. Increasing the bio-oligo concentration further was inhibitory to selected site recovery. The optimal bio-oligo concentration used in the original DEASH protocol was also 0.375 μM (Jeffreys and May, 2003), so this result was not unexpected. Since optimal recovery had been achieved using all four bio-oligos simultaneously, the bio-oligos were mixed for subsequent enrichments.

6.2.4 The effects of primer extension by *Taq* DNA polymerase on enrichment dynamics

The use of QIAquick PCR clean-up would not be suitable for purification of the amplified DNA during scaled up experiments. The volume of the PCR product would be too high for such a purification method, and the molecules generated by a full-length insertion (12 – 13 kb) could be above the size exclusion limit of the QIAquick columns. Since the protocol, as planned, was complex, it was necessary to simplify it whenever possible. If primer extension of the bio-oligos did not affect the dynamics of hybridisation enrichment, it may have been possible to add the unmodified bio-oligos directly to PCR-amplified DNA. As a result, before determining a suitable large-scale method to purify PCR-amplified DNA, experiments were performed to determine whether 3′ extension of the bio-oligos would affect the dynamics of hybridisation enrichment.

Primer extension from the un-modified bio-oligos should stabilise the annealed bio-oligos, thus requiring a higher elution temperature to recover the DNA. The excess of streptavidin to biotin in the enrichment reactions was estimated to be ~5.76 times. According to Gonzalez et al. (1997), the denaturation midpoint of streptavidin at this ratio of ligand to streptavidin is ~75 °C. It is therefore likely that some of streptavidin will have denatured whilst eluting at 80 °C, releasing bio-oligos into the eluate. At the ratio of biotin to streptavidin used in enrichment, denaturation of streptavidin begins at ~67 °C (Gonzalez *et*

al., 1997), so it was necessary to elute below 67 °C to avoid this. Inhibition of *Taq* DNA polymerase activity by dNTP dephosphorylation significantly reduced recovery of L1 DNA in the 80 °C eluate (Figure 6.6 B).

As the aim of the project was to recover complete L1 insertions, high elution temperatures could not be used for fear of thermally damaging the DNA, so the effects of primer extension had to be eliminated. Also, denaturation of the streptavidin during elution could potentially have resulted in the release of bio-oligos into the eluates. Being essentially modified PCR primers, the bio-oligos could have potentially interfered with subsequent PCR analysis. However, SAP treatment, while being effective in demonstration the effects of primer extension on hybridisation enrichment dynamics, could not be used in the scaled experiment. As an enzymatic process, SAP treatment could not be 100 % efficient in inhibiting *Taq* DNA polymerase activity. Also, SAP activity required alteration of the enrichment conditions. *Taq* DNA polymerase was therefore removed from post-PCR samples by proteinase K digestion and phenol chloroform extraction. The DNA was subsequently concentrated by ethanol precipitation, and then dissolved in 5 mM Tris-HCl (pH 7) prior to enrichment.

6.2.5 Optimisation of the enrichment protocol

In theory, the simplest way to obtain optimal enrichment efficiency was to adjust the annealing temperature such that a high recovery of the selected site was achieved while achieving minimal recovery of the unselected site. This was a "balancing act," as low annealing temperatures (38 °C to 42 °C) give a high recovery of both selected and unselected sites (data not shown) and consequently low enrichment efficiency. Increasing the annealing temperature decreased both selected and unselected site recovery. As expected, annealing at 48 °C recovered more of the selected site (~16 %) than when annealing was performed at 51 °C (~13 %) (Figure 6.7). Also, unselected site recovery decreased as annealing temperature increased from 48 °C (~0.01 %) to 51 °C (~0.006 %). The decrease in unselected site recovery, however, was significant enough to result in an increase in enrichment efficiency. At 48 °C enrichment efficiency was ~1,600 fold, but at 51 °C this increased to ~2166 fold.

Although higher enrichment efficiency was achieved by enriching at 51 °C, 48 °C was selected as the optimal annealing temperature as it gave a higher selected site yield. It was

predicted that the enrichment efficiency could be increased by increasing the length of time for which samples were washed at the annealing temperature, maximising the removal of loosely bound unselected site molecules. At the annealing temperature, the bio-oligos should be bound tightly to the complementary sites on single-stranded L1 DNA. By contrast any bio-oligos annealed non-specifically to the unselected sites would be mismatched to their annealing site. Washing the Dynabead bound bio-oligos at the annealing temperature for longer than 1 min should consequently remove more of the loosely bound unselected site with little effect on selected site recovery. Increasing the annealing temperature wash from 1 min to 2.5 min significantly reduced the recovery of the unselected site, thus increasing the enrichment efficiency to between 4,000 and 10,000 fold (Figure 6.8).

In an attempt to further increase selected site recovery, Dynabead concentration was increased to 10.8 μg/μl and the elution temperature increased to 65 °C. Selected site recovery increased marginally to > 10 % with no effect on unselected site recovery (Figure 6.9). Enrichment efficiency had therefore increased marginally to between 4,000 fold and 10,000 fold. At this stage optimal enrichment conditions had been determined, giving at least 10 % recovery of the selected site at an enrichment efficiency of around 4 to 10 thousand fold. It was extremely important that high enrichment efficiency was achieved in order to remove the overwhelming mass of unselected site molecules following PCR; however, this could not be at the expense of recovering sufficient selected site molecules. A higher selected site yield was therefore desirable.

Another approach to maximise recovery of the selected site was also investigated, which involved the use of multiple rounds of enrichment. In a single round of enrichment, just over 10 % of the recoverable selected site molecules were recovered, with a small percentage lost in the washes. Therefore a little less than 90 % of the recoverable selected site molecules remained in the un-bound fraction after enrichment. Re-extraction was performed on the remaining un-bound fraction. The optimised enrichment conditions were used. Selected site recovery was ~10 % for the first two extractions, dropping marginally in the third extraction (5 % to 10 %). Overall, selected site recovery was therefore between 25 % and 30 %. Unselected site recovery remained between 0.001 % 0.005 % throughout, and thus enrichment efficiency was therefore between 12,000 to 35,000 fold.

This extraction strategy achieved the goal of optimisation of the enrichment protocol, i.e. to achieve a sustainable and high yield of the selected site while maintaining high enrichment efficiency. Chapter 7 details the steps taken to combine the optimised multiplex

PCR conditions, and the optimised enrichment conditions, into a single method with the potential to recover *de novo* L1 insertions from human sperm gDNA.

Chapter 7: Combining multiplex PCR and hybridisation enrichment

7.1 Results

Previous enrichments had been carried out on small masses of DNA (maximum 10 ng). However, in order to screen 600 µg of sperm genomic DNA (gDNA), large scale multiplex PCR would be required. This produced extremely large volumes of PCR product requiring hybridisation enrichment. Phenol chloroform extraction followed by ethanol precipitation was used to purify and concentrate the amplified DNA, which was subsequently dissolved into a small volume (50 to 100 µl) of 5 mM Tris-HCl (pH7). However, the concentrated samples contained vast quantities of empty (unselected) target amplicons, so the recovery of L1-containing target amplicons would be extremely challenging.

This chapter deals with the steps taken to develop a method capable of recovering small masses of L1-containing DNA, through hybridisation enrichment, in the presence of extremely large amounts of empty site DNA.

7.1.1 PCR and enrichment recovery of the AL121819 insertion from a heterozygous donor

7.1.1.a Preparing DNA for hybridisation enrichment

Since it was essential to concentrate PCR-amplified DNA into a volume suitable for hybridisation enrichment, it was necessary to perform phenol chloroform extraction and ethanol precipitation of amplified samples.

Hendecaplex PCR (the ten target site loci and the AL121819 insertion locus) was performed using sperm gDNA from Donor R2 (AL1218189 insertion +/-) in 96-well PCR plates. Each 50 µl PCR contained 0.5 µg of gDNA. To ensure amplification had occurred, 2 µl samples were taken from the 12 plates, and fractionated on an agarose gel (a representative experiment is shown in figure 7.1 A).

Following PCR, a third of each reaction from the whole plate was pooled, and 5 % put aside for further analysis. Phenol chloroform extraction and ethanol precipitation was performed on the remaining sample (Materials and Methods). The purified DNA was subsequently dissolved in 50 µl 5 mM Tris-HCl (pH 7) prior to enrichment.

To determine the efficiency of the phenol chloroform extraction, a small sample of the purified DNA was fractionated on an agarose gel alongside a control series generated from

the set aside un-purified sample (Figure 7.1 B). Unlike in figure 7.1 A, 7.1 B was deliberately not fractionated sufficiently to resolve individual amplicons. This enabled visual approximation of DNA recovery following extraction, and showed that approximately 75 % of the DNA was recovered (Figure 7.1 B).

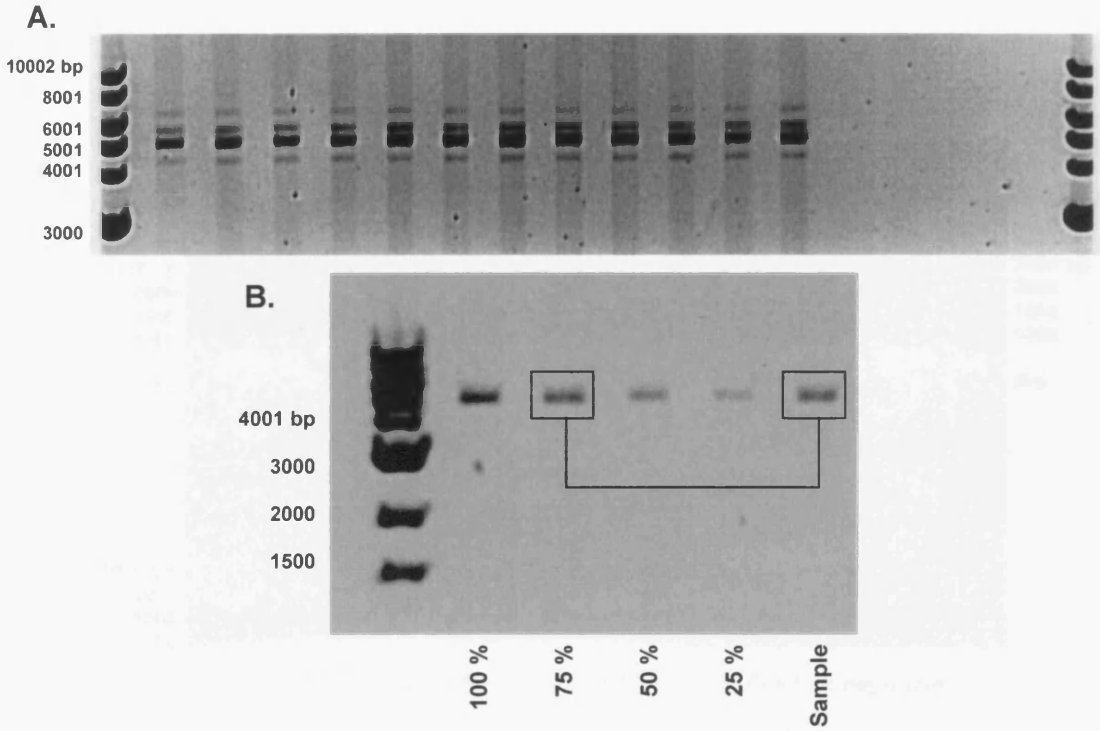


Figure 7.1: A. EtBr-stained gel showing Hendecaplex PCR of donor R2 DNA. The AI121819 insertion empty site can be seen at ~6.2 kb. The DNA negative controls, containing water, can be seen to the right hand side of the amplified samples. B. EtBr-stained gel showing the efficiency of phenol chloroform extraction. Samples were run alongside 250 ng 1 kb molecular weight marker.

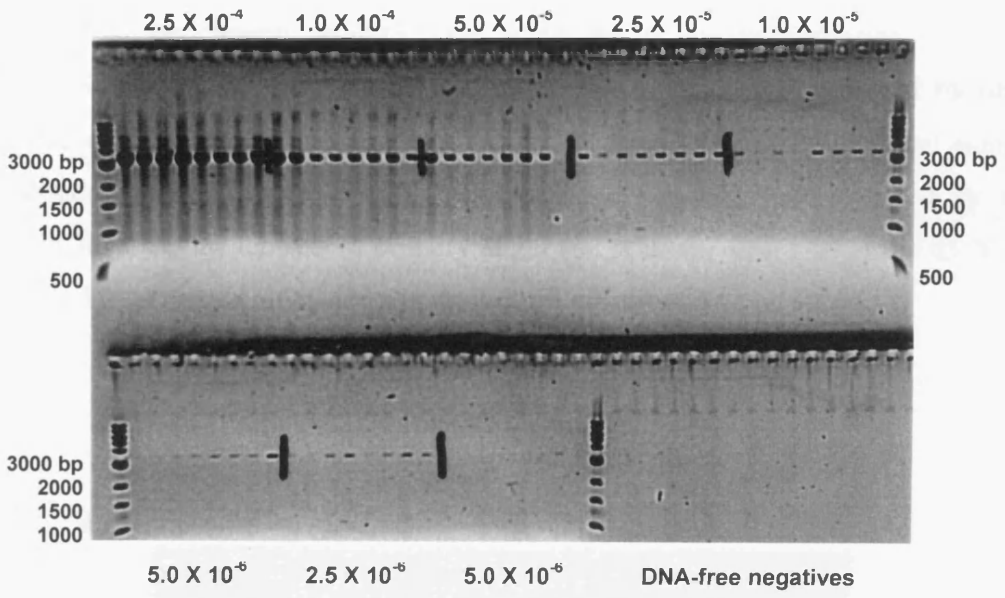
7.1.1.b Single-molecule quantification of AL121819 prior to enrichment

In order to assess the recovery of the AL121819 DNA by hybridisation enrichment, the number of molecules contained in the pre-enrichment sample had to be determined. This was achieved by single-molecule PCR (Materials and Methods).

Primary PCRs using primers PF819LRB and PF819LRC were performed on dilutions of the phenol chloroform extracted DNA. Secondary PCR used primers RB819A and RBLR2519 (3335 bp) and were seeded with 0.2 µl of the primary PCRs. The secondary PCRs were fractionated on a 0.8 % (w/v) LE agarose gel and visualised by EtBr staining (Figure 7.6 A). Amplification positive and negative results were counted and are tabulated in figure 7.6 B.

The results were input into the Poisson confidence interval program (A.J. Jeffreys) to estimate the number of AL121819 insertion molecules present per µl of the pre-enrichment sample. The Poisson confidence interval programme estimated that there were 256,385 molecules containing the AL121819 insertion per µl of the phenol chloroform extracted multiplex PCR, with 95 % confidence intervals of 133,803 to 450,615 molecules per µl.

A.



B.

Dilution (10^{-x})	Sample volume (µl)	Number of positive PCRs	Number of negative PCRs
x = 4	2.5	8	0
	1.0	8	0
5	5.0	8	0
	2.5	8	0
	1.0	6	2
6	5.0	5	3
	2.5	5	3
	1.0	1	7

Figure 7.6: A. Secondary amplification used to estimate the number of AL121819 insertion containing molecules contained in the phenol chloroform extracted multiplex PCR. B. Summary table.

These results suggested that the AL121819 insertion, plus 6 kb of flanking gDNA was amplifying efficiently in the hendecaplex PCR. Also, since the number of AL121819 insertion molecules per µl of the pre-enrichment sample had been determined, a control dilution series could be made to estimate the recovery of AL121819 insertion DNA during enrichment.

7.1.1.c Hybridisation enrichment recovery of AL121819-containing molecules

Single-molecule analysis had shown that numerous copies of the AL121819 insertion selected site were present in the pre-enrichment sample. However, there would still be an overwhelming mass of empty target site loci present in the pre-enrichment sample, which could prevent recovery of the selected site DNA.

To determine whether recovery of L1-containing selected site DNA was possible, 33 µl of the pre-enrichment samples were subjected to enrichment using 3 extractions (Materials and Methods). Recovery of the AL121819 insertion was assessed by detection PCR seeded with 0.5 µl and 1 µl of the 65 °C eluate, amplified and fractionated alongside a control dilution series (Figure 7.7 A). Unselected site recovery was assessed by seeding decaplex PCR (using the ten sets of primary target site primers) with 1 µl of the 65 °C eluate, amplified and fractionated alongside a control dilution series.

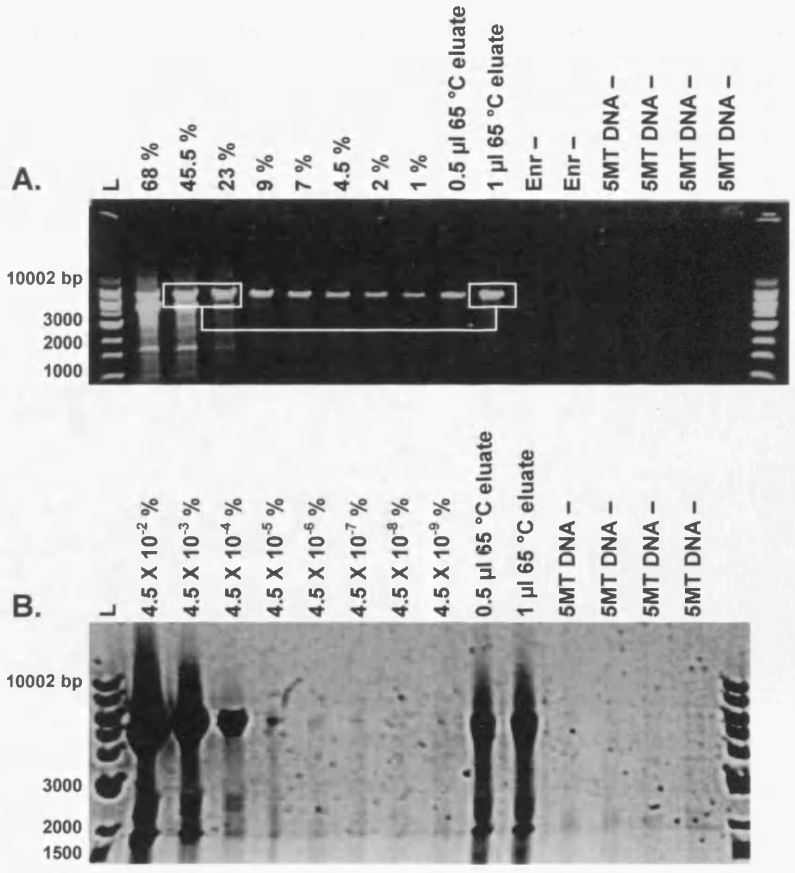


Figure 7.7: A. EtBr-stained gel showing selected site recovery (AL121819 insertion) following enrichment. B. EtBr-stained gel showing the unselected (empty target) site recovery following enrichment. The boxes indicate comparisons to estimated yield. Enrichment negative (Enr -: performed using no bio-oligos) and DNA free negative controls containing 5mM Tris Hcl (5MT -) were performed. Samples were run alongside 250 ng of 1 kb molecular weight marker.

~8.46 x 10^6 molecules of the AL121819 insertion selected site were input into each enrichment (33 µl of the pre-enrichment sample), of which between 23 % (1.95 x 10^6 molecules) and 45.5 % (3.85 x 10^6 molecules) was recovered, in the 150 µl eluate (Figure 7.7 A). However, recovery of the unselected site was between 4.5 x 10^{-3} % and 4.5 x 10^{-4} % (Figure 7.7 B). Enrichment efficiency was therefore between 5,100 and 101,000 fold.

7.1.1.d Increasing enrichment efficiency with an additional round of enrichment

Given that the pre-enrichment sample contained an extremely large number of empty target site amplicons, enrichment efficiency would have had to be extremely high to remove all empty site molecules.

To test whether multiple rounds of enrichment could increase enrichment efficiency, 33 µl of the previous 65 °C eluate was enriched a second time. The eluate contained at least ~428,000 (~13,000 molecules / µl) molecules, assuming 23 % recovery in the initial enrichment. Detection PCRs were performed on the secondary eluates alongside appropriate control series as above (Figure 7.8 A and B).

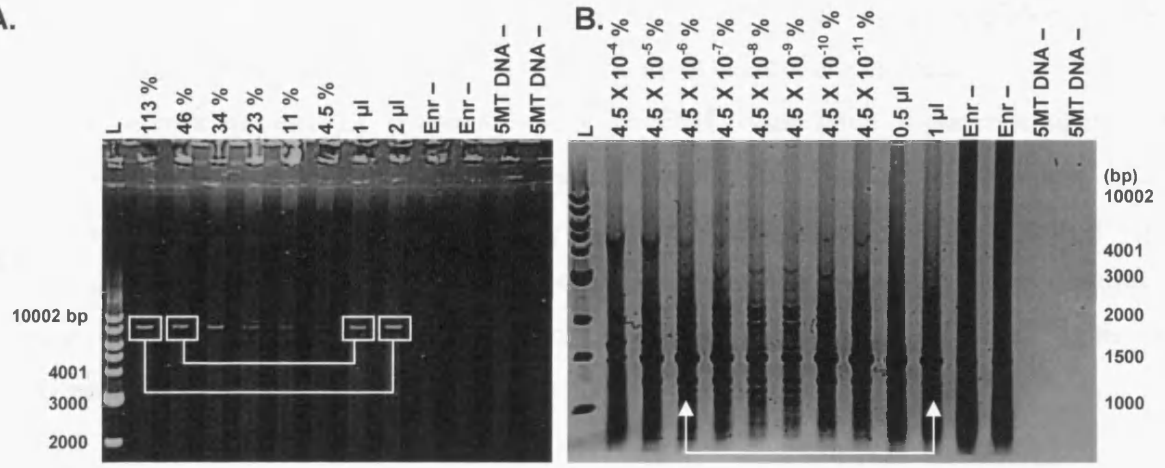


Figure 7.8: A. EtBr-stained gel showing selected site recovery in the secondary eluate. B. EtBr-stained gel showing unselected site recovery in the secondary eluate. Samples were run against 1 kb molecular weight marker. The boxes indicate comparisons to estimated yield. Samples were run alongside 250 ng of 1 kb molecular weight marker.

Since the molecules input into the secondary enrichment had been recovered in the primary enrichment, all the molecules were recoverable since they were single-stranded. Between 46 % (196,880 molecules) and 56.5 % (113 % / 2 µl: ~241,820 molecules) of the

L1-containing DNA was recovered (Figure 7.8 A). Therefore, of the ~8,460,000 molecules input into the primary enrichment, at worst 2.3 % (196,800 molecules) was recovered.

The unselected site control series was seeded with dilutions of the pre-enrichment sample, so after two rounds of enrichment < 4.5 $x10^{-6}$ % (Figure 7.8 B) of the empty target site loci were recovered. Enrichment efficiency was therefore at worst 510,000 fold.

Figure 7.8 B also shows non-specific amplification in all lanes except for the DNA negative control. The negative control containing elution buffer (ED) showed non-specific amplification. This suggested that the multiplex primers were priming non-specific amplification, perhaps as a result of interaction with the sonicated *E. coli* carrier DNA in the ED. This was corrected by increasing the *Pfu* DNA polymerase concentration used in each reaction such that there were 10 units of *Taq* to 1 unit of *Pfu*, rather than 20 to 1.

7.1.2 Amplification and recovery from a single-molecule of an L1-containing target

The above enrichment experiments showed that recovery of L1-containing molecules was possible following large-scale multiplex amplification, using gDNA from a donor heterozygous for the AL121819 insertion. 100 ng of the gDNA contained ~1600 amplifiable molecules of the AL121819 insertion molecules (Figure 7.9 A), so the 48 µg of gDNA used to seed the multiplex PCR contained ~768,000 amplifiable molecules of this locus.

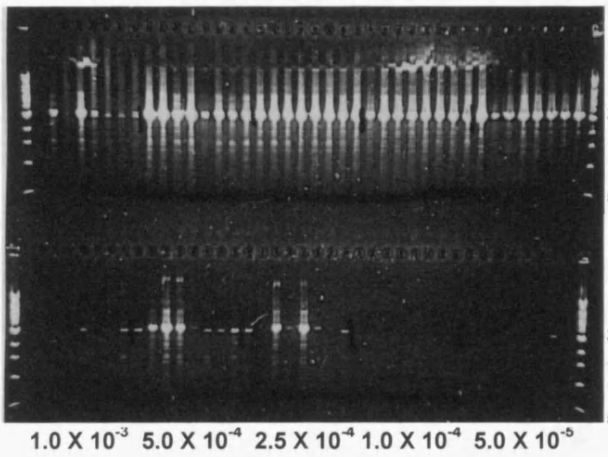
To recover *de novo* L1 insertions, multiplex PCR would have to generate sufficient molecules for hybridisation enrichment from a single-molecule. Although an extremely high enrichment efficiency could be achieved using two rounds of hybridisation enrichment, it was possible that as few as 2 % of L1-containing molecules would be recovered. It was therefore necessary to determine whether sufficient molecules could be generated from a single-molecule of a full-length L1 insertion to permit hybridisation enrichment.

7.1.2.a Single-molecule analysis of the AL121819 insertion from gDNA

Before PCR using a single-molecule of the AL121819 insertion could be performed, it was necessary to determine the number of these molecules in an aliquot of gDNA (AL121819 insertion +/-). Single-molecule PCR analysis was therefore performed on blood gDNA taken from donor R2. The secondary amplifications were fractionated by gel electrophoresis, and visualised by EtBr staining (figure 7.9 A).

Positive and negative lanes were counted (Figure 7.9 B) and entered into the Poisson confidence interval programme. From the data shown in figure 7.9 the 80 ng / µl aliquot of donor R2 blood gDNA contained an estimated 1623 molecules of the AL121819 insertion per µl. Assuming the DNA was high quality, the concentration was accurate and that the AL121819 insertion should be present in every 6 pg of DNA taken from a heterozygote, amplification efficiency of the AL121819 insertion was ~12 %. This was lower than the efficiency gained from single-molecule analysis of the AL583853 insertion (52 %; figure 4.7, section 4.1.5.c). The 95 % confidence intervals were 879 to 2771 molecules per µl.

A.



B.

Dilution (10^{-x})	Sample volume (µl)	Number of positive PCRs	Number of negative PCRs
x = 2	5.0	6	2
	2.5	8	0
	1.0	8	0
3	5.0	8	0
	2.5	8	0
	1.0	2	6
4	5.0	8	0
	2.5	6	2
	1.0	0	8
5	5.0	1	7

Figure 7.9: A. EtBr-stained gel showing the result of the single-molecule PCR used to determine the number of AL121819 insertion containing molecules in an aliquot of donor R2 blood DNA Samples were run alongside 250 ng 1 kb molecular weight marker. B. Summary of the number of positive and negative samples determined from photograph A.

7.1.2.b Mixing DNA to amplify single-molecules of the AL121819 insertion in multiplex

Once the number of molecules of the AL121819 insertion contained in the aliquot of donor R2 DNA had been established, a DNA mixing experiment could be performed. The experiment, illustrated in figure 7.10, was performed to replicate the recovery of molecules generated by PCR across a full-length L1 insertion plus the flanking target site, from the single-molecule level. This can be directly compared to recovery of a *de novo*L1 insertion in the overall experiment.

In total, one pre-enrichment sample contained amplicons generated from gDNA representing ~100 molecules of the AL121819 insertion (the positive control); 10

pre-enrichment samples contained amplicons generated from gDNA representing ~2 molecules of the AL121819 insertion; and one pre-enrichment sample contained no additional gDNA (the negative control). Phenol chloroform extraction was between 50 % and 75 % efficient for all but sample 2.1 (Figure 7.11 B), which showed ~ 10 % recovery (data not shown).

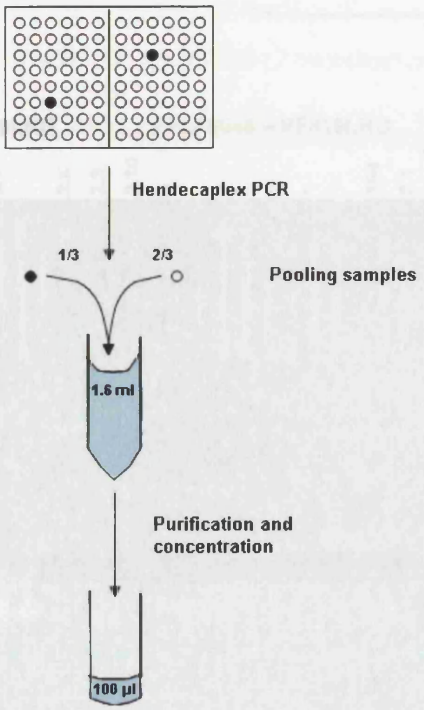


Figure 7.10: Diagram illustrating DNA mixing experiments. 96-well PCR plates were divided into two. Every well contained a 50 µl hendecaplex PCR seeded with 500 ng of DNA from a donor showing homozygous absence for the AL121819 insertion. One well (black spot) in each half plate was also seeded with a volume of donor R2 blood DNA (heterozygous for the AL121819 insertion), equivalent to either 100 molecules or 2 molecules. One plate was not seeded with donor R2 DNA and acted as a negative control. Post PCR, 2/3 of the white wells and 1/3 of the black wells were pooled, making it equivalent to pooling a third of one whole plate. The DNA was then purified by phenol chloroform extraction and concentrated by ethanol precipitation. The DNA was dissolved in 100 µl of 5 mM Tris-HCl (pH 7). Finally the efficiency of phenol chloroform extraction was estimated.

7.1.2.c Enrichment recovery of extracted samples

After sample purification, hybridisation enrichment was performed. 33 µl of each purified pre-enrichment sample was enriched using standard hybridisation enrichment protocol (Materials and Methods). After an initial round of enrichment, 33 µl of each primary eluate was enriched for a second time (Materials and Methods). Following enrichment, 40 µl of each

eluate was amplified in a 51 µl hendecaplex PCR or single-plex PCR (primers PF819LRA and PF819LRB).

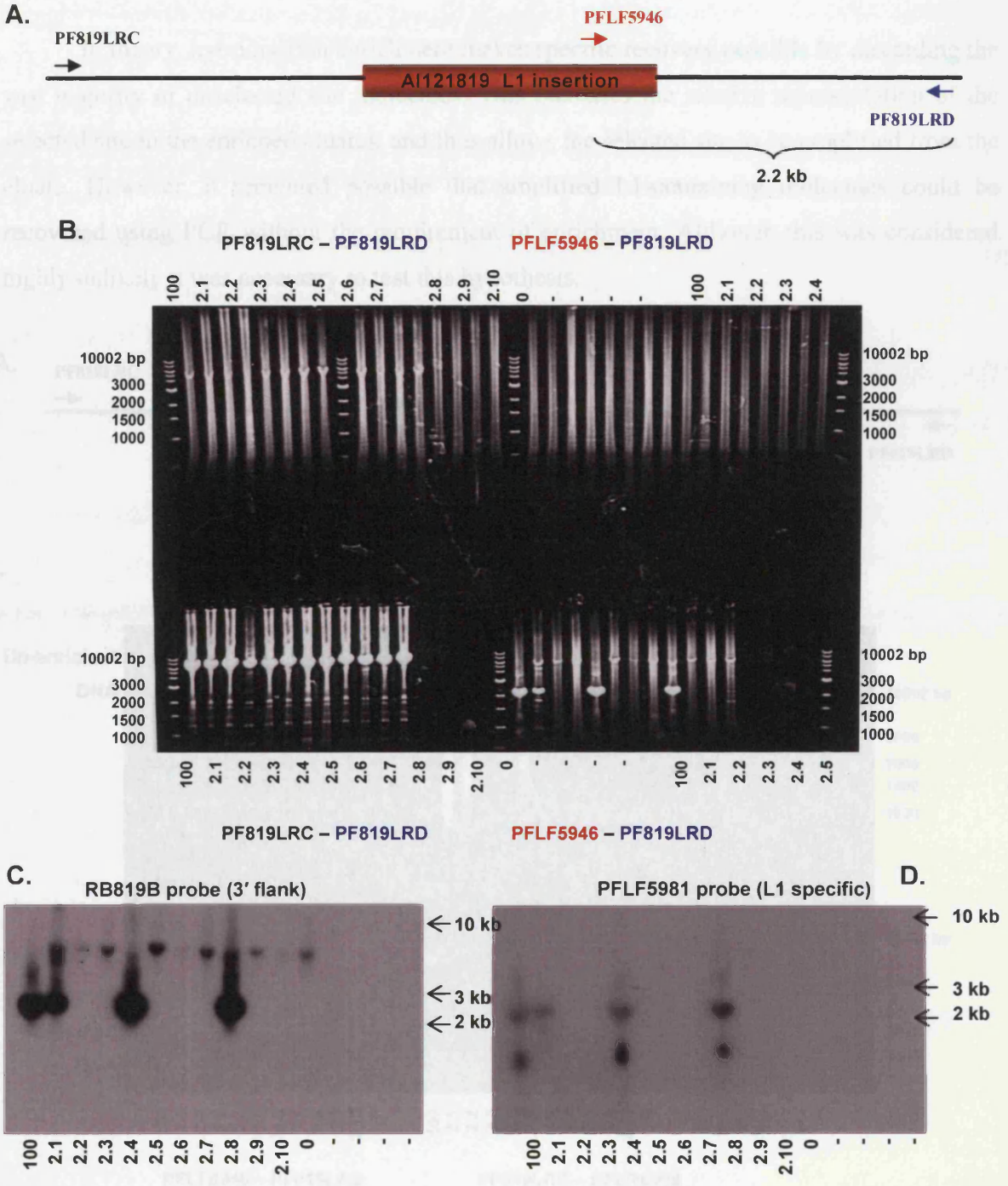


Figure 7.11: A. Diagram showing the placement of AL121819 insertion amplifying primers. B. EtBr-stained gel showing the result of AI121819 insertion detection PCRs. The primers used in each secondary PCR are shown. Samples were run alongside 250 ng 1kb molecular weight marker. C and D show the result of hybridisation with RB819B (C) and PFLF5981 (D).

7.1.2.d Is enrichment actually required to recover L1-containing amplicons?

In theory, hybridisation enrichment makes specific recovery possible by discarding the vast majority of unselected site molecules. This increases the relative representation of the selected site in the enriched eluates, and thus allows the selected site to be amplified from the eluate. However, it remained possible that amplified L1-containing molecules could be recovered using PCR without the requirement of enrichment. Although this was considered highly unlikely it was necessary to test this hypothesis.

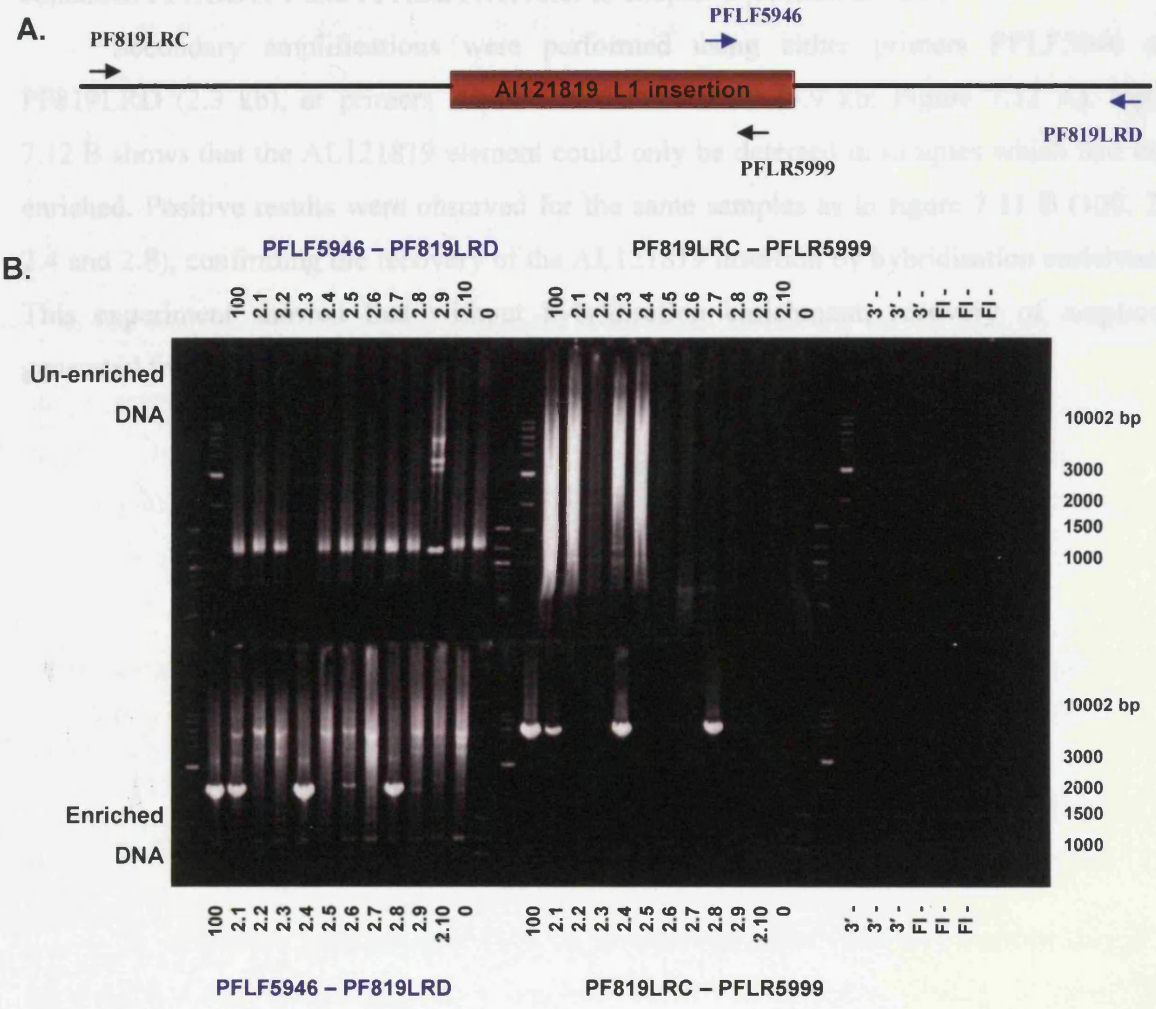


Figure 7.12: A. Placement of the primers used to amplify the AL121819 insertion. B. The result of secondary PCR amplification. Bands of the correct size were only generated when amplifications were performed on the enriched samples. The bands in the 3' flanking PCR correspond with the bands in the full-length element PCR. Samples were run against 250 ng NEB 1 kb molecular weight marker.

During enrichment, the DNA contained in 33 µl of purified PCR-amplified DNA was diluted into a final volume of 150 µl (3 x 50 µl eluates) of ED. In addition the DNA contained in 33 µl of the primary eluate was again diluted into a final volume of 150 µl of ED during secondary enrichment. This was mimicked by diluting 33 µl of the pre-enrichment sample to a final volume of 150 µl in ED. 33 µl of this dilution was further diluted to a final volume of 150 µl in ED.

40 µl of the secondary eluates, and 40 µl of the un-enriched samples were subject to PCR using primers PF819LRA to PF819LRB in 50 µl reactions (this amplification also contained PFHBBTF1 and PFHBBTR1; refer to chapter 8 (section 8.1.2.b).

Secondary amplifications were performed using either primers PFLF5946 and PF819LRD (2.3 kb), or primers RB819C and PFLR5999 (5.9 kb: Figure 7.12 A). Figure 7.12 B shows that the AL121819 element could only be detected in samples which had been enriched. Positive results were observed for the same samples as in figure 7.11 B (100, 2.1, 2.4 and 2.8), confirming the recovery of the AL121819 insertion by hybridisation enrichment. This experiment showed that without hybridisation enrichment, recovery of amplicons generated from a single-molecule of an L1 insertion would not be possible.

7.2 Discussion

Chapters three to six detailed the design and optimisation of separate protocols capable of amplifying and recovering a full-length L1 element, and up to 6 kb of flanking sequence. This chapter details the steps taken to combine the optimised protocols into a single method capable of capturing these molecules even if only one or a few molecules were present prior to multiplex PCR.

7.2.1 Amplification of mass genomic DNA

The ultimate goal of this investigation was to screen gDNA from a whole human ejaculate (approximately 600 µg). To make screening of this mass of DNA feasible, the maximum capacity of the multiplex PCR had previously been determined (chapter 5), as approximately 500 ng (0.5 µg) of DNA in a 50 µl multiplex PCR. PCR of ~600 µg of genomic DNA would thus require 1200 individual PCR reactions, which was feasible using 96-well PCR plates.

Similarly, it was impossible in terms of time and expense to perform 1200 individual enrichments, and the necessary controls. This problem was tackled by pooling a third of each reaction from a 96-well PCR plate, prior to phenol chloroform extraction and ethanol precipitation of the DNA. This concentrated the DNA so it could be dissolved in a volume of suitable for enrichment (50 to 100 µl). However, in a pre-enrichment sample estimated to contain 7.3 µg to 9.8 µg of amplified DNA (data not shown), the overwhelming majority was empty target site DNA. It was therefore possible that hybridisation enrichment might simply not be able to recover sufficient L1-containing DNA from such a concentrated sample.

7.2.2 Enrichment recovery of the AL121819 insertion from a heterozygous donor

To determine whether hybridisation enrichment could recover L1-containing DNA from purified and concentrated DNA, hendecaplex PCR was used to amplify 48 µg of DNA from a donor heterozygous for the AL121819 insertion. This PCR-amplified all ten target site loci, the Al121819 insertion empty site, and the AL121819 insertion filled site.

Following hendecaplex PCR, purification and concentration of the amplified DNA, single-molecule analysis of the pre-enrichment sample was performed to estimate the number of AL121819 insertion contained in the sample prior to enrichment. The pre-enrichment

sample contained ~256,000 amplifiable AL121819 insertion-containing molecules per µl (Figure 7.6). Of the ~8,448,000 molecules put into enrichment, between 23 % and 45.5 % were recovered (Figure 7.7 A). This suggested that although the pre-enrichment sample was predominantly empty target loci, a high yield of L1-containing DNA could still be recovered. Also, recovery of the unselected site was very low, at between 4.5 x 10^{-3} % and 4.5 x 10^{-4} % (Figure 7.7 B), showing enrichment efficiency of 5,100 – 101,000 fold. This suggested that hybridisation enrichment was extremely efficient in discarding empty target site loci.

This high yield of L1-containing DNA presented an opportunity to perform a secondary enrichment on the primary eluate. Although this would lead to a loss of selected site molecules, it should also remove far more unselected site molecules and further purify the L1-containing DNA. To determine whether this hypothesis was correct, 33 µl of the 65 °C eluate from the primary enrichment was subjected to a further round of enrichment, and between 46 % and 56.5 % of the L1-containing DNA was recovered (Figure 7.8 A). After two rounds of enrichment, of the ~8,460,000 L1-containing molecules input into the primary enrichment, at worst 2.3 % (page 155) were recovered. < 4.5 x10^{-6} % (Figure 7.8 B) of the empty target site loci were recovered, indicating that enrichment efficiency was at worst 510,000 fold.

Figure 7.8 B. shows non-specific background amplification from the well to the dye front in all lanes except for the 5 mM Tris-HCl (pH 7) DNA negative control. The enrichment negative control, containing ED, showed non-specific amplification resulting in smearing from the well to the dye front. This may have been due to multiplex primers amplifying non-specifically, perhaps as a result of interaction with the sonicated *E. coli* carrier gDNA contained in the ED. This occurred at a lower level in the enrichment positive samples and control series samples. This suggests that an absence of amplifiable molecules led to an increase in non-specific amplification. This problem was corrected by doubling the concentration of *Pfu* DNA polymerase contained in each PCR such that it was at a ratio 10:1 rather than 20:1 in relation to *Taq* DNA polymerase.

Although the results obtained in these initial scaling-up experiments were extremely encouraging, it was still necessary to demonstrate amplification and recovery from a single-molecule of a full-length L1 insertion. This would closely mimic the actual experiment in which a single genomic target site molecule accommodates a *de novo* L1 insertion. This molecule would then have to be efficiently amplified in the presence of millions of competing empty target sites, prior to pooling with another 95 insertion negative reactions followed by purification and subsequent enrichment. The 48 µg of gDNA amplified in the initial scaled

experiment contained ~672,000 amplifiable molecules of the AL121819 insertion locus. Thus although ~98 % of PCR-generated AL121819 insertion molecules were lost after two rounds of enrichment, there were still sufficient molecules to allow PCR recovery of the L1 insertion from the secondary eluate. However, the question remained as to whether it would be possible to achieve this from a single-molecule of the AL121819 insertion.

7.2.3 Amplification of single L1-containing molecule by multiplex PCR and its recovery using hybridisation enrichment

This experiment was similar to the previous experiment in that bulk gDNA was amplified and enriched. However this experiment amplified selected site molecules from a very low concentration of the AL121819 insertion by mixing bulk gDNA from a donor homozygous absent for the insertion, with DNA from a heterozygous donor.

Following two rounds of enrichment, the secondary eluates were screened for the presence of the AL121819 insertion. Since it was possible that up to 98 % of the molecules put into the primary enrichment would be lost, the secondary eluate likely contained a small number of L1-containing molecules. It was therefore deemed necessary to use two successive PCR reactions to recover these molecules from the secondary eluate. Following detection PCRs, the AL121819 insertion was recovered from the 100 molecule control, and samples 2.1, 2.4 and 2.8. However, this could only be achieved when the primary and secondary amplifications were specific for the AL121819 insertion.

This experiment demonstrated enrichment recovery of selected site molecules generated from close to single-molecules of the AL121819 insertion. This suggested that the multiplex PCR must have been extremely efficient, especially considering the loss of selected site molecules during two rounds of enrichment. However, it was possible that since the pre enriched sample contained sufficient AL121819 insertion molecules for enrichment recovery, it may have been possible to recover, and analyse, these molecules without the requirement for hybridisation enrichment.

7.2.4 Can amplified selected site molecules be recovered without enrichment?

To test this hypothesis, 33 µl of the pre-enrichment sample was diluted to a similar level as expected following two rounds of enrichment. Amplification of the AL121819

insertion was detected in lanes 100, 2.1, 2.4 and 2.8 (Figure 7.12 B), which directly corresponded to the pattern seen in figure 7.11 B. Figure 7.12 B also showed that the AL121819 insertion could only be detected in the samples which had been enriched, so PCR alone was not sufficient to recover the selected site molecules. It should also be noted that the enriched sample contained single-stranded DNA only, so automatically reducing the number of molecules entering the detection PCR by 50 %. Of the molecules input into enrichment, up to 98 % would have been lost during two rounds of enrichment. Therefore the un-enriched sample contained far more AL121819 insertion DNA, but detection of this DNA still required enrichment. This suggested that although the enrichment process discards a large percentage of selected site molecules, the vast level of purification achieved by efficient removal of unselected site molecules is key to recovering *de novo* L1 insertions from the human germline.

7.2.5 Conclusions derived from the control experiment

These control experiments allowed optimisation of the method without using an artificial control construct that could become a contaminant. The control experiments were run under the same conditions as the overall experiment, and these conditions were sufficiently sensitive to amplify and recover near to a single L1-containing target molecule. As the control experiment therefore closely resembled the overall experiment, this suggested that the designed method was sufficiently sensitive to amplify and recover L1-containing DNA from the single-molecule level.

Chapter 8: Hybridisation enrichment recovery of *de novo* L1 retrotransposon insertions

8.1 Results

The previous chapter demonstrated that combined multiplex PCR and hybridisation enrichment had the potential to recover *de novo* L1 insertions from targeted regions of the genome. A control experiment, designed to closely mimic recovery of a *de novo* insertion, showed amplification and recovery of amplified molecules from close to a single-molecule of the control L1 insertion. This chapter details how this developed protocol was executed in an attempt to recover *de novo* L1 insertions from human sperm genomic DNA.

8.1.1 Preparation of DNA for hybridisation enrichment

Screening for *de novo* L1 insertions was performed on 576 µg of sperm gDNA, extracted from donor 7 of the Department of Genetics donor panel (University of Leicester). Hendecaplex PCR, containing the ten primary target site loci and long-range control primers PFLRCtrlA and PFLRCtrlB (AC008706: Materials and Methods), was performed in twelve 96-well PCR plates, and each 50 µl PCR was seeded with 0.5 µg of gDNA (Materials and Methods). Post PCR, DNA was purified by phenol chloroform extraction, concentrated by ethanol precipitation and dissolved in 100 µl of 5 mM Tris-HCl (pH 7). Approximately 50 % of the amplified DNA was recovered during purification (Data not shown).

In order to amplify full-length *de novo* L1 insertions into the target loci, it was essential that 12 – 13 kb fragments could be amplified efficiently. The control primers PFLRCtrlA and PFLRCtrlB generate a 12,806 bp amplicon with no close matches to the bio-oligo sequences. Following purification, single-molecule PCR was performed on the pre-enrichment samples (Materials and Methods). Primary PCR used primers PFLRCtrlA and PFLRCtrlB, and secondary PCRs used primers PFOP_ENeA and PFOP_ENeB. Between 120,000 and 700,000 amplifiable long-range control molecules were present per µl in the pre-enrichment samples (data not shown), suggesting that large amplicons were amplifying efficiently in the initial multiplex PCR.

33 µl of each pre-enrichment sample was subsequently subjected to two rounds of hybridisation enrichment. The secondary eluates were then screened for the presence of L1-containing target locus molecules.

8.1.2 Screening eluates for *de novo* insertions

In chapter 7, a control experiment demonstrated hybridisation enrichment recovery of amplified L1-containing DNA that had been generated by PCR from near single-molecule amounts of the AL121819 insertion. The control experiment suggested that recovery of L1-containing DNA ideally required locus-specific primary PCR followed by locus-specific secondary amplification. However, in the control experiment the location of the L1 insertion was known, but in the overall experiment L1 insertions could be present in any of ten target site loci. Therefore the first screening experiment used multiplex primary PCR followed by L1 insertion-specific secondary PCR.

8.1.2.a Screening eluates using multiplex PCR followed by L1 insertion specific re-amplification

The first screening strategy used primary screening PCRs containing the ten sets of primary target site primers, seeded with 40 µl of the secondary eluates (total reaction volume 51 µl). The primary PCR was performed to amplify any L1 insertions and their flanking target site DNA (Figure 8.1).

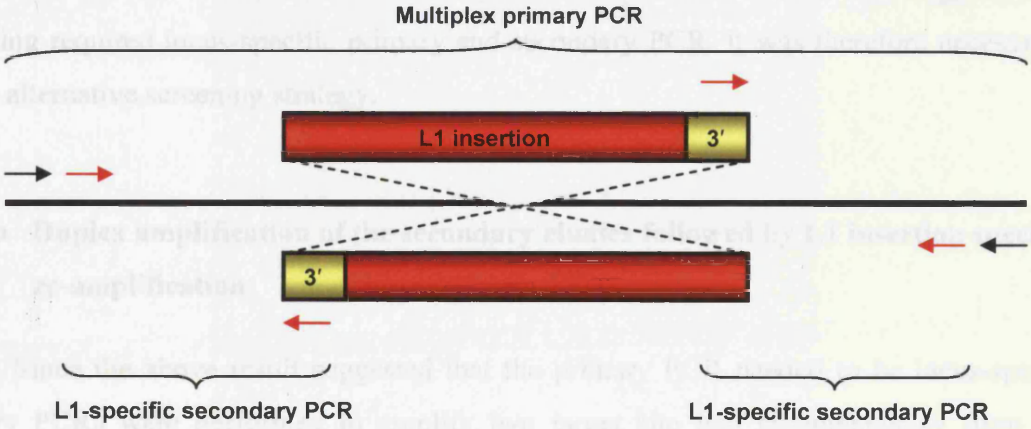


Figure 8.1: Diagram showing the PCRs used to screen the enrichment eluates for L1-containing target site loci. The locations of primers used in both primary and secondary PCRs are shown.

The secondary PCRs were L1 insertion-specific and locus-specific. The nested secondary target site primers from each locus were used separately along with an L1 specific primer (PFLF5946: Figure 8.1) which was located 3′ of the bio-oligo sequences. This nesting

strategy prevented the amplicons generated from becoming contaminants that were recoverable by hybridisation enrichment. Both secondary target site primers were used in PCR since *de novo* L1 insertions could insert in either orientation (Figure 8.1).

The secondary PCRs were fractionated by gel electrophoresis and stained with EtBr. Visible bands, viewed using a Dark Reader transilluminator (Clare Chemical Research), were excised and the DNA extracted using a QIAquick gel extraction kit. Recovered DNAs were ligated into a plasmid vector (pGEM®-T Easy Vector System; Promega), cloned and sequenced (Materials and Methods).

The sequences were analysed as follows. Firstly the Align tool of the NCBI (www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi) website was used to identify a consensus sequence for each sequence. The sequences were then converted into GCG format using seqed. The accessions containing the target site sequences were imported into GCG using the fetch tool, and consensus sequences were aligned with their target sequence using the fasta algorithm, in GCG. The consensus sequences were also aligned with human L1 element L1.3 (accession L19088). Consensus sequences showing strong identity to both the target site and L1.3 were exported from GCG and annotated (appendix v on the accompanying CD-ROM).

Of 10 recovered amplicons analysed, all but one of the recovered sequences resembled "jumping PCR" artefacts (section 8.1.3) rather than genuine L1 insertions (as detailed on pg 170). This was not unexpected since the control experiment shown had suggested that screening required locus-specific primary and secondary PCR. It was therefore necessary to use an alternative screening strategy.

8.1.2.b Duplex amplification of the secondary eluates followed by L1 insertion specific re-amplification

Since the above result suggested that the primary PCR needed to be locus-specific, primary PCRs were performed to amplify two target site loci simultaneously from their primary target site primers. These duplex primary PCRs were then re-amplified similarly to the secondary PCR used in section 8.1.2.a. The control experiment eluates (containing the AL121819 insertion) were also amplified in this way.

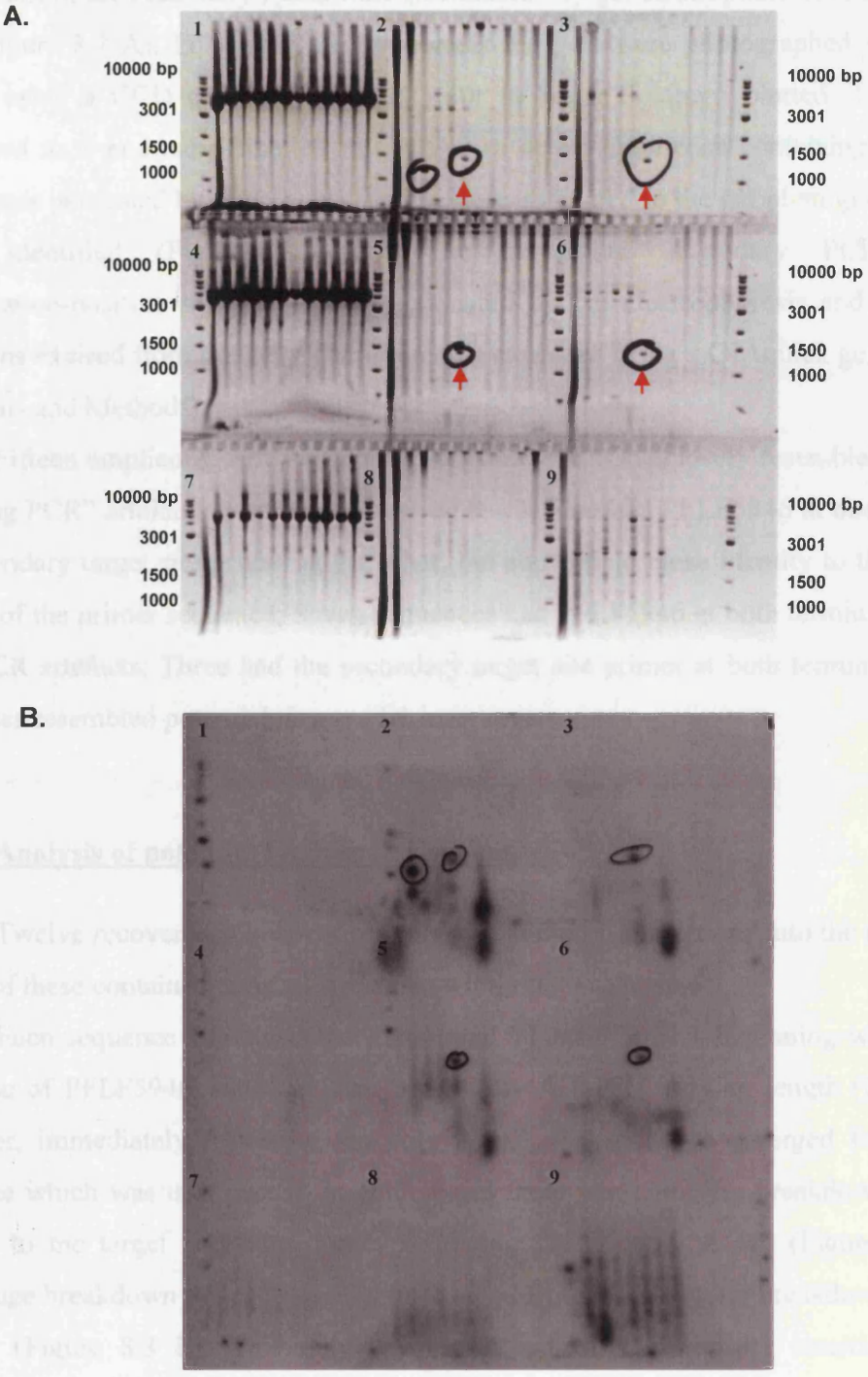


Figure 8.2: A. EtBr-stained gel showing secondary detection PCR. PCR 1 used primers HBBTF2 – HBBTR2, 2 used HBBTF1 – PFLF5946, 3 used HBBTR2 – PFLF5946, 4 used DMDTF2 – DMDTR2, 5 used DMDTF2 – PFLF5946, 6 used DMDTR2 – PFLF5946, 7 used FIXTF2 – FIXTR2, 8 used FIXTF2 – PFLF5946 and 9 used FIXTR2 – PFLF5946. B. Southern blot of A hybridised with 5' end labelled PFLR5999. Bands detected in B were identified in A. The red arrows show bands appearing at approximately 1.5 kb in lane 6 of each set. These were discarded as PCR artefacts.

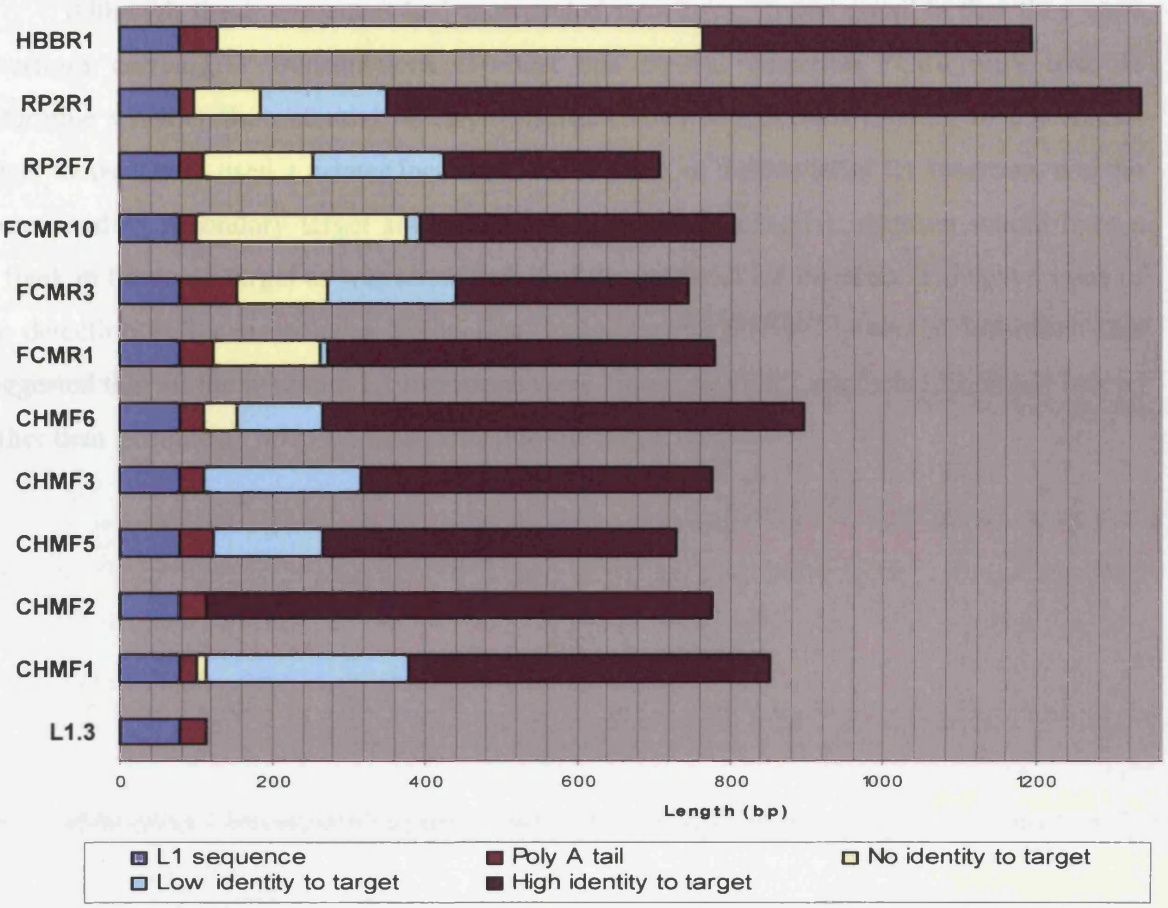
Half of the secondary PCRs were fractionated by gel electrophoresis and stained using EtBr (Figure 8.2 A). Following electrophoresis the gels were photographed in a darkroom cabinet using a CCD camera (Syngene) prior to being Southern blotted. The blots were hybridised to 5′-end-radio-labelled PFLR5999 to detect amplicons containing L1 sequence. The images generated by oligo hybridisation were compared to the gel photograph and visible bands identified (Figure 8.2 B). The remaining secondary PCR from the hybridisation-positive samples, were fractionated by gel electrophoresis and the identified amplicons excised from the gels. The DNA was extracted using a QIAquick gel extraction kit (Materials and Methods).

Fifteen amplicons were sequenced, and analysed. One closely resembled the potential "jumping PCR" artefacts described in section 8.1.3; Five had PFLF5946 at one terminus and the secondary target site primer at the other, but showed no close identity to the target locus outside of the primer sequence; Seven sequences had PFLF5946 at both termini and therefore were PCR artefacts; Three had the secondary target site primer at both termini. None of the sequences resembled potential *de novo* L1 insertions.

8.1.3 Analysis of potential L1 insertion sequences

Twelve recovered sequences resembled potential L1 insertions into the target loci, but eleben of these contained unusual structures within their sequences.

Each sequence contained the 3′-terminal 77 bp of an L1 beginning with the primer sequence of PFLF5946. Each L1 also had a poly A tail of varying length (Figure 8.3 A). However, immediately following the poly A tail, the sequence diverged from the target sequence which was unexpected. In come cases there was complete breakdown in sequence identity to the target site immediately following the L1 poly A tail (Figure 8.3 A). The percentage breakdown in sequence identity ranged from 12 % to total breakdown of sequence identity (Figure 8.3 B). This may have been indicative of an L1 insertion carrying a 3′ transduction. Interestingly, the breakdown in sequence identity usually correlated with the presence of an *Alu* element (Appendix v on the accompanying CD-ROM). Following the region of sequence identity breakdown, the sequences became > 99 % identical to the target sequences. The 12th sequence (CHMF2: Figure 8.3 A) showed > 99 % sequence identity to the target site from the L1 to the secondary target site primer sequence. This therefore resembled the expected structure of a potential *de novo* L1 insertion.



Sequence ID	length of high identity to the target (bp)	Length of low identity to the target (bp)	Length of no identity to the target (bp)	Breakdown in sequence identity (%)	Poly A tail length (bp)	L1 sequence length (bp)	5' mis-prime sequence (bp)
L1.3					37	77	
CHMF1	471	268	12	18.9	23	77	0
CHMF2	663	0	0	0.0	37	77	53
CHMF5A	463	143	0	14.0	46	77	0
CHMF3A	463	205	0	12.2	33	77	0
CHMF6A	629	112	45	40.8	32	77	0
FCMR1F	507	9	141	97.3	45	77	0
FCMR3F	302	170	120	52.1	75	77	0
FCMR10F	411	17	278	95.6	22	77	0
RP2F7F	280	172	146	59.7	32	77	0
RP2R1	987	166	88	45.3	19	77	0
HBBR1	429	0	635	100.0	52	77	0

Figure 8.3: A. A chart showing the basic structures of 11 potential *de novo* L1 insertions recovered by hybridisation enrichment, aligned with the human-specific L1 insertion L1.3 (L19088). The names of the sequences (x axis) refer to the sequences in appendix v and vi on the accompanying CD-ROM. B. A summary table which accompanies to the above chart. This table provides information on sequence identity breakdown, as well as the length of the individual structures shown on the chart. Also additional information such as the presence of mis-priming is also shown.

Although these sequences had unexpected structures, it was possible that they were insertions carrying 3′ transductions. To test this theory, detection PCRs were used to determine whether the potential *de novo* insertions had 5′ junctions relating to the correct target locus. These used a primer located in the 3′ flank of the potential L1 insertion, and the corresponding secondary target site primer. A genuine *de novo* L1 insertion would have a 5′ flank in the same target as was identified 3′ of the potential L1 insertion. However, none of the detection PCRs identified a 5′ junction, including the CHMF2 potential insertion. This suggested that all the potential L1 insertions were "jumping PCR" artefacts (discussed below) rather than genuine *de novo* L1 insertions into the target loci.

8.2 Discussion

The previous results detail the application of the method developed during this investigation. Human sperm gDNA was subjected to multiplex PCR, phenol-chloroform extracted and ethanol-precipitated and concentrated. Molecules containing bio-oligo recognition sequences were then recovered by hybridisation enrichment. The multiplex PCR was capable of amplifying 12 to 13 kb molecules, and thus should be able to amplify full-length L1 insertions contained in the target sites.

At this stage of the investigation, multiplex amplification of gDNA and subsequent enrichment recovery of amplified L1-containing molecules were standard procedures. The focus of this chapter was the design of a screening strategy capable of resolving recovered L1-containing target site molecules from non-specifically recovered empty target site loci. The information obtained in the control experiments (chapter 7) were used during the development of the screening strategy.

8.2.1 Multiplex amplification of the secondary eluates followed by L1 insertion specific re-amplification

It was determined in chapter 7 that up to 98 % of the L1-containing molecules generated by PCR would be lost after two rounds of hybridisation enrichment. This would mean that the concentration of recovered L1-containing molecules would be relatively low in the enriched eluates. Since primary PCR followed by secondary re-amplification was capable of amplifying single-molecules of an L1 insertion to visible levels, visualised using EtBr staining, it was decided to use a two-stage PCR screening strategy.

The control experiment suggested that locus-specific primary PCR and secondary PCR would be required to amplify L1-containing molecules from the eluates. However, a *de novo* L1 insertion could have inserted into any one of the ten target site loci, so was decided that decaplex PCR should be used in the primary screening PCR. Insertion-specific secondary amplification was used following the primary PCR.

Ten amplicons showing the characteristics of an L1 insertion were identified by the selected screening strategy. The sequences of all 10 amplicons showed high sequence similarity to L1.3 between the PFLF5946 primer site and the start of the poly A tail, so suggesting the potential L1 insertions may have originated from young L1s. However, six out

of ten had poly A tails shorter than that of L1.3 (Figure 8.3 A). *De novo* insertions are expected to have long poly A tails (for example 180 nucleotides (nt) (Miki *et al.*, 1992)). The length of *de novo* poly A tails in cultured cell assays have been estimated to be approximately 60 residues (Gilbert *et al.*, 2002) to 88 (Symer *et al.*, 2002) residues on average, but have been noted up to 115 residues (Symer *et al.*, 2002). Computational analysis of the human genome give an estimated average of Ta element poly A length at 23 residues, and 16 residues for L1pa2 (Han *et al.*, 2005). This began to cast doubt as to whether the sequences had been generated form genuine *de novo* L1 insertions. However, it was also possible that the long poly A tails may have collapsed during PCR, resulting in the observed reduced AT tract length.

Further doubt as to the origins of the potential L1 insertions was revealed by analysis of the flanking sequences. Sequences derived from *de novo* L1 insertions were expected to exhibit close to 100 % identity to the target site. ten of the eleven sequences displayed breakdown in sequence identity immediately flanking the L1 poly A tail (Figure 8.3 A). In all 10 cases breakdown in sequence identity began within 6 to 23 bases of a poly AT-rich simple repeat, commonly an *Alu* poly A tail (7/10).

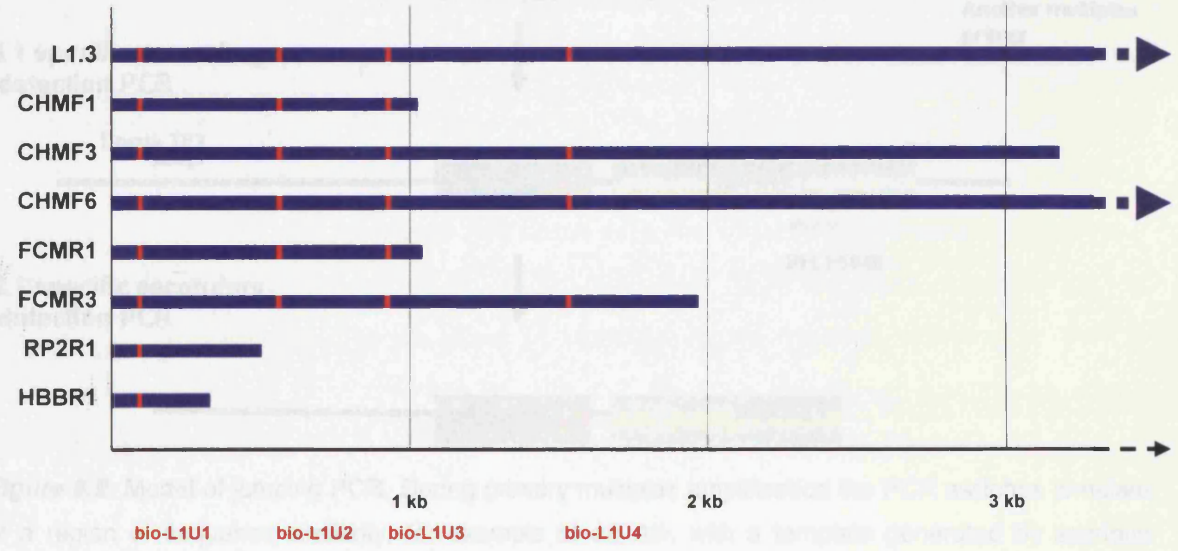


Figure 8.4: Chart showing the presence of bio-oligo recognition sequences in L1s identified by blat analysis. The L1s were identified by submitting the regions of poor sequence identity, described in the text, into BLAT searches, and identifying L1s in the RepeatMasker track. All the identified L1s had at least 1 bio-oligo recognition sequence (red).

The most plausible explanation for the breakdown of sequence identity was strand jumping during PCR at regions showing sequence similarity due to the presence of dispersed repeats, for example *Alu* elements. The strongest evidence supporting this theory was that in 6 of the 9 cases the L1 and flanking diverged sequence mapped to existing regions of the genome with sequence identity of > 99 % (Appendix v and vi). Five of the sequences aligned to a region of the genome containing $L1_{HS}$ elements and the remaining two mapped to regions of the genome containing L1PA2 elements (RepeatMasker track in the UCSC genome browser).

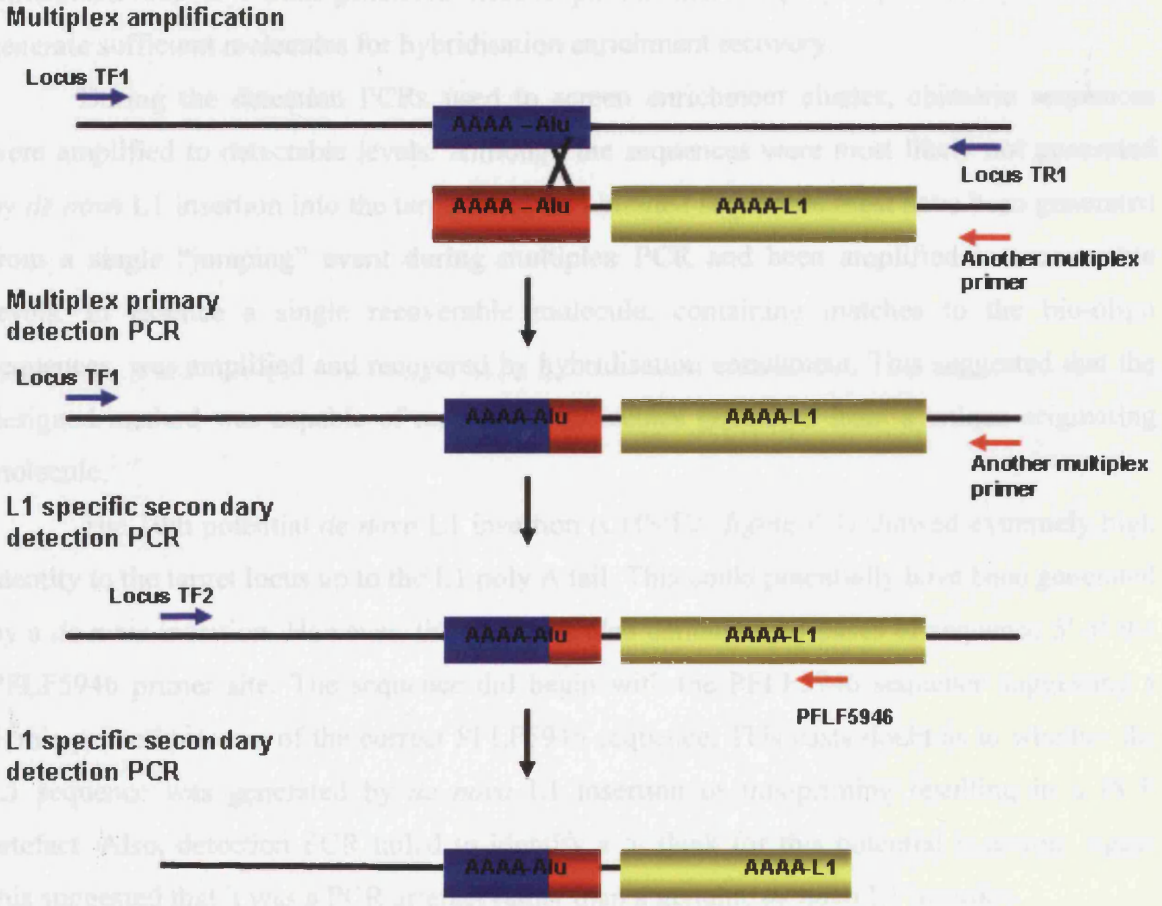


Figure 8.5: Model of jumping PCR. During primary multiplex amplification the PCR switches template at a region of sequence similarity, for example at an *Alu*, with a template generated by spurious priming of another primary target site primer. The erroneous template contains a truncated L1 and a sequence to which one of the twenty primary target site primers can anneal. If this occurred early in PCR it would generate sufficient molecules to be recovered by hybridisation enrichment. Following detection PCR, bands were excised from agarose gels, purified, cloned and the DNA sequenced. The sequences, described above, therefore contain L1, a region of sequence divergence, a chimeric *Alu* and a region of high identity to the target locus.

These could have been progenitor elements resulting in *de novo* L1 insertions carrying 3′ transductions. All the L1s identified by the RepeatMasker track searches had exact matches to at least one of the bio-oligo sequences, dependent on the level of 5′ truncation (Figure 8.4).

Although "jumping PCR" was the most likely explanation for these structures, the mechanism by which the sequences originate remains unknown. In theory, the templates for the chimeric jumping PCR products must have originated during primary multiplex PCR. Figure 8.4 shows a possible model of "jumping PCR". The jumping event would have to occur sufficient early during PCR to generate sufficient molecules for hybridisation enrichment recovery. Once generated these amplicons must amplify exponentially in order to generate sufficient molecules for hybridisation enrichment recovery.

During the detection PCRs used to screen enrichment eluates, chimeric sequences were amplified to detectable levels. Although the sequences were most likely not generated by *de novo* L1 insertion into the target loci, the obtained sequences must have been generated from a single "jumping" event during multiplex PCR and been amplified to recoverable levels. In essence a single recoverable molecule, containing matches to the bio-oligo sequences, was amplified and recovered by hybridisation enrichment. This suggested that the designed method was capable of recovering molecules generated from a unique originating molecule.

The 10th potential *de novo* L1 insertion (CHMF2; figure 8.3) showed extremely high identity to the target locus up to the L1 poly A tail. This could potentially have been generated by a *de novo* insertion. However, the sequence also contained 53 bases of sequence 5′ of the PFLF5946 primer site. The sequence did begin with the PFLF5946 sequence suggesting a 5′ mis-prime upstream of the correct PFLF5946 sequence. This casts doubt as to whether the L1 sequence was generated by *de novo* L1 insertion or mis-priming resulting in a PCR artefact. Also, detection PCR failed to identify a 5′ flank for this potential insertion. Again this suggested that it was a PCR artefact rather than a genuine *de novo* L1 insertion.

8.2.2 Duplex PCR of the secondary eluates followed by L1 insertion specific re-amplification

The results obtained in the control experiment suggested that an alternative screening strategy was required. However unlike the control experiment it was unknown in which loci any *de novo* insertions would be present. To reduce the number of detection PCRs, primary

PCRs were performed in duplex (two sets of primary target site primers in each reaction). The primary PCRs were then re-amplified in an L1 insertion specific manner using PFLF5946 and the relevant secondary target site primer.

Following detection, sixteen amplicons were recovered and sequenced. Of the sixteen, only one resembled a potential *de novo* L1 insertion; however it was similar in structure to the "jumping PCR" artefacts identified in section 8.1.2.a. The remaining 15 sequences were PCR artefacts generated by amplification from a single primer.

8.2.3 Experimental analysis of recovered sequences

Although there was doubt as to whether any of the L1 sequences were derived from *de novo* insertion, it was necessary to eliminate the possibility that this could be the case. A real *de novo* insertion into a target locus would result in an L1 element flanked on both sides by the target same site locus. Therefore the designed PCR would only amplify an insertion that had a true 5′ flank in the correct target site (Figure 8.6).

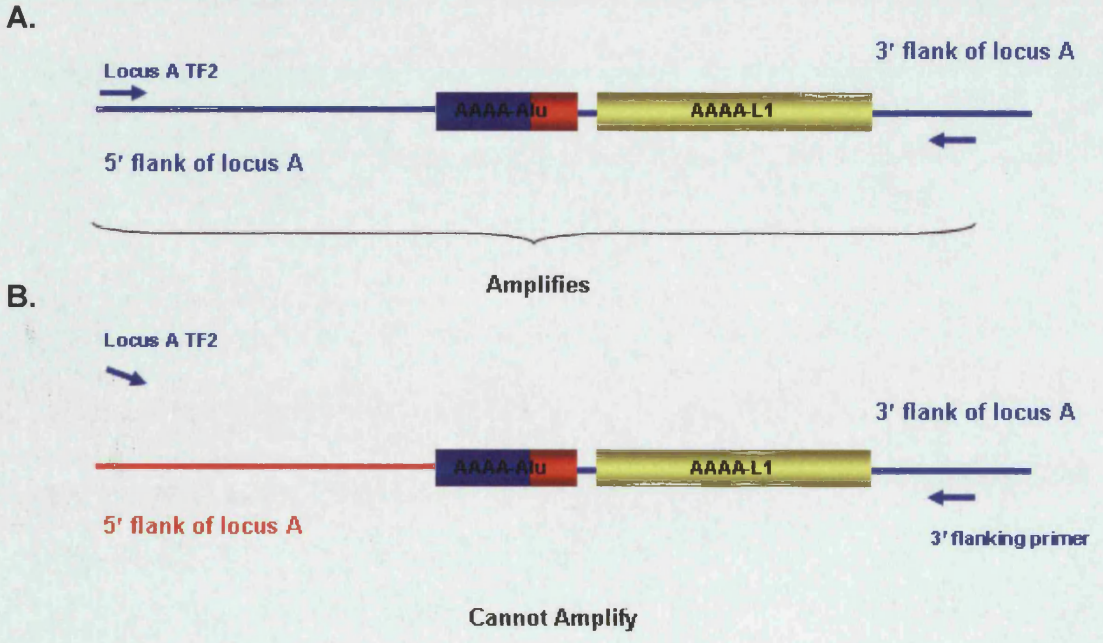


Figure 8.6: Diagram showing the PCR screening strategy used to determine whether the potential L1 insertions had 5′ flanks in the correct target locus. A genuine *de novo* L1 insertion would resemble A, and would amplify using this detection PCR. A spurious L1 generated by "jumping PCR" would have an incorrect 5′ flank, thus would not amplify using this detection PCR.

As expected, none of the PCRs generated bands that could have been indicative of L1 insertions (data not shown), suggesting that the sequences had originated from "jumping PCR" rather than *de novo* L1 insertion. It was however possible that PCR had generated insufficient product to be visible by EtBr staining. Southern blotting was performed and the blot hybridised with 5′ radio-labelled RB3PA2 a Ta specific oligonucleotide. No specific L1-containing bands were identified using hybridisation; therefore it was unlikely that the Southern blots contained L1 DNA. This evidence suggested that all the potential insertions were generated by "jumping PCR" rather than genuine *de novo* L1 insertion.

8.2.4 The capabilities of the designed method

Both the control experiment and the recovery of "jumping PCR" amplicons suggested that the designed method is sufficiently sensitive to amplify and subsequently recover L1-containing DNA at the single-molecule level. Also, recovered molecules could be recovered from enrichment eluates using a two stage PCR screening strategy.

Although the method should have been capable of recovering *de novo* L1 insertions, none were actually recovered from 576 μg of sperm gDNA. Chapter 9 discusses the potential reasons for the lack of recovered *de novo* L1 insertions in this experiment.

Chapter 9: Concluding remarks and future direction

9.1 Concluding remarks

9.1.1 Summary of the investigation

The overall aim of this investigation was to develop a method capable of recovering *de novo* L1 insertions from the human germline. Direct detection of *de novo* L1 insertions in the human germline has only been achieved by chance, for example in disease-causing L1 insertions. Very little is therefore known about the dynamics of L1 retrotransposition *in vivo*. Three main factors have hampered previous attempts to access *de novo* L1 insertions: firstly there are no human germline cell cultures; secondly L1 is a relatively small insertion that can insert anywhere within an extremely large genome (Feng *et al.*, 1996; Moran, 1999); finally, the rate of *de novo* insertion is apparently extremely low.

Transposable elements, for example L1, must retrotranspose in the germline, or sufficiently early during embryogenesis (prior to germline differentiation), in order to be transferred to subsequent generations. Since L1 can theoretically insert anywhere in the genome (Feng *et al.*, 1996; Moran, 1999), attempts to recover L1 insertions have been directed towards maximising the amount of the genome screened during the experiment. Such experiments however have to compromise on the recoverable data to maximise the amount of the genome that could be screened. For example the Amplification Typing of L1 Active Sub-families (ATLAS) method (Badge *et al.*, 2003) is restricted to recovery of L1 junction fragments specific to the Ta L1 subfamily. As a method for detecting *de novo* L1 insertions in the germline, ATLAS suffers from lack of validation since full-length *de novo* insertions are destroyed by the screening method.

An alternative approach, developed during this investigation, was to recover complete L1 insertions. Such a method allows a higher level of validation in that the whole insertion should be maintained, and insertions could be tracked through the individual stages of the experiment. However, it was necessary to limit the area of the genome that could be screened. Limiting the search area meant that the vast majority of *de novo* L1 insertions into the germline of the individual screened would be missed, since > 99.998 % of the genome was not included in the screening. The screening method did, however, include ~41.5 kb of genomic sequence (8 target loci) that has been amenable to L1 insertion, as these loci have historically harboured disease causing insertions in patients with genetic diseases.

The method developed during this investigation was to use multiplex PCR to amplify target DNA from 576 µg of sperm gDNA. Ten target loci were amplified simultaneously (total 50,817 bp) equating to approximately 1/59035 (0.0017 %) the size of the human sperm genome. The amplified DNA was purified by phenol chloroform extraction, recovered and concentrated by ethanol precipitation, and then dissolved in 100 µl of 5 mM Tris-HCl (pH 7). L1-containing molecules were then recovered using hybridisation enrichment. Hybridisation enrichment selectively recovered L1-containing molecules, and removed molecules that did not contain L1 sequence. Recovered molecules could then be amplified to detectable levels by long-range duplex PCR followed by L1 insertion specific secondary PCR.

Throughout the investigation conditions were optimised to ensure that: 1) a full-length L1 insertion into an target site molecule could be efficiently amplified; 2) the full-length L1 insertion would be efficiently amplified in competition with an overwhelming mass of empty target site molecules; 3) the amplified L1-containing molecules could be purified by phenol chloroform extraction and ethanol precipitation; 4) The L1-containing molecules could be efficiently recovered by hybridisation enrichment; 5) finally the L1-containing molecules could be efficiently detected by PCR amplification of the enrichment eluates.

A control experiment was performed that closely mimicked the main experiment. This experiment involved amplification of the AL121819 L1 insertion from the single-molecule level to recoverable amounts using hybridisation enrichment. The control experiment used exactly the same conditions as the main experiment for multiplex PCR, DNA purification, hybridisation enrichment and screening. Since a *de novo* L1 insertion into one of the target loci would exist initially as a single-molecule, the control experiment suggested that the method would be capable of amplifying and recovering the insertion from that single-molecule. However, following completion of the main experiment, no *de novo* L1 insertions were recovered.

9.1.2 Prior expectations

Previously, it was estimated that numerous *de novo* L1 insertions should be recovered from the DNA from a single human ejaculate (~576 µg), but this was not the case. Given that the sperm genome has a mass of 3 pg, approximately 1.92×10^8 individual sperm genomes were screened during this investigation. Although the control experiment showed that the method was capable of recovering *de novo* L1 insertions into the target loci, none were

actually recovered. Assuming that the selected loci were representative of any randomly selected genetic locus, and that L1 insertion into the genome is essentially random (Feng *et al.,* 1996; Moran, 1999), and that amplification of L1 insertions from gDNA is ~32 % efficient (52 % in section 4.1.5.c and 12 % in section 7.1.2.a) the insertion rate of L1 was less than 3 insertions in 4.35 x 10^8 kb of gDNA (note: half the target loci are on the X chromosome so are only present in 50 % of sperm). Thus the observed estimate of the insertion rate is within the previously calculated estimates, albeit at the lower end of the estimate.

These estimates of the rate of L1 insertion into the human genome suggest that between 1 in 8 (Kazazian, 1999) and 1 in 250 (Li *et al.,* 2001) humans harbour a *de novo* insertion. Given that the approximate size of the haploid human genome is 3 x 10^9 bp (Lander *et al.,* 2001; Venter *et al.,* 2001), from the estimates above, an L1 inserts into the human genome every 4.8 x 10^7 kb to 1.5 x 10^9 kb. Thus in the 9.8 x 10^9 kb of gDNA screened, an estimated 7 – 204 *de novo* L1 insertions should have been recovered according to these predicted rates of L1 insertion. However, these predicted rates were not generated by experimental investigation; rather they were mathematical estimations that used a number of generalisations and assumptions such that the accuracy of these estimates may be questionable.

Estimated rate of L1 insertion into the genome by Kazazian (1999)

Kazazian, for example, estimated an optimistic rate of L1 insertion into the genome by comparing the number of human diseases caused by insertion of retrotransposons (twelve L1, fourteen Alu and two others (March 1999)) to the number of independent mutations reported in the Gene Mutation Database (http://www.uwcm.ac.uk/uwcm/mg/hgmd0.html). 1 / 600^{th} (28) of the 16,650 independent mutations were thus retrotransposon-derived, but there are problems with this estimate. Firstly, disease causing retrotransposon insertions are often overlooked by methods used for mutation detection, and secondly recurrent mutations in the same gene may only be counted once (Kazazian, 1999).

In his estimate, Kazazian then used an estimated frequency of mutation in the human genome as 1 x 10^{-9} per nucleotide per year (10^{-6} per kb per year) to make his estimation of L1 retrotransposition frequency. However, the figure for the frequency of human mutation per nucleotide per year is based on an average of two estimates, 1.8 x 10^{-9} per nucleotide (nt) per year (Eyre-Walker and Keightley, 1999), and 4 x 10^{-10} (Neel *et al.,* 1986). These calculations were also based on assumptions, for example Eyre-Walker and Keightley determined the

number of nucleotide substitutions that caused amino acid changes in 46 orthologous human and chimpanzee genes, but synonymous mutations were ignored. They also assumed that the average length of human protein-coding sequence was approximately 1,523 bp per gene, and that there are 60,000 genes in the human genome (Eyre-Walker and Keightley, 1999; Perriere *et al.*, 2000). However, subsequent to this calculation the number of human genes was revised to between 20 to 25 thousand genes (Collins et al., 2004).

To complete his calculation, Kazazian then took 3×10^6 kb to be the size of the haploid human genome, and a human generation time of approximately 25 years, to estimate an average 75 mutations per haploid genome per year. 1 in 600 of these mutations were caused by retrotransposon insertion, thus 0.125 haploid genomes contained a *de novo* retrotransposon insertion. Here Kazazian made an error in his calculation by not doubling the haploid number, so estimated that 1 in 8 ($0.125 \times 8 = 1$) rather than 1 in 4 humans harbour a *de novo* retrotransposon insertion. However, only 43 % (12 out of 28 retroelement-derived mutations) were actually L1 insertions, thus the rate of L1 insertion was actually close to 1 in 9 people having a *de novo* L1 insertion. So, after correction of Kazazian's calculation, an L1 inserts into the human genome once every 5.4×10^7 kb of human gDNA.

Given that the data for estimates of the rate of L1 insertion into the human genome were likely to be affected by the assumptions made, they cannot be used to discount the observed frequency of insertion made in this investigation (< 3 insertion in 4.35×10^8 kb of gDNA). However the observed very low rate of L1 insertion seems at odds with the ongoing expansion and evolution of L1 in the human genome.

9.1.3 Reasons for the low observed rate of L1 insertion into the human genome

Three possible scenarios that could lead to the low observed rate of L1 insertion into the genome are addressed. The first explanation is that the wrong target sites may have been used during this investigation. The second explanation is that the selected donor may have exhibited an abnormally low rate of L1 retrotransposition. The final explanation is that although L1 must retrotranspose in the germline in order to be of evolutionary significance, it may retrotranspose prior to germline differentiation, and so will colonise numerous germ cells. All three explanations are addressed below.

9.1.3.a Were the correct target loci selected for this investigation?

The first explanation for the low observed rate of L1 retrotransposition could be that the selected targets were inappropriate. Eight of the target loci were selected because L1 insertions into them had caused genetic disease. The reason for selecting these targets was simply that since the sequences had been amenable to L1 insertion in the affected patients, it was likely that the sequences would be amenable to insertion in the future.

It is unlikely that selecting target sites within important genes would be detrimental to the experiment. Hypothetically, if an L1 had inserted into one of the target loci during early embryogenesis it could not have had a major detrimental effect as the selected donor was healthy. Therefore had the insertion occurred prior to germline differentiation, or during the early stages of spermatogenesis, it would have been represented in the donor's sperm DNA. However, its representation in the germline may have been limited by negative selection. A common gain of function mutation which results in Apert's syndrome confers a selective advantage to spermatogonial cells (Goriely et al., 2005), thus increasing their frequency in the germline. Since positive selection occurs in the germline, it is reasonable to suspect that negative selection against deleterious mutations may also exist to reduce their representation. However, although negative selection may have occurred, it is unlikely to have removed all the copies of the L1 insertion from the genome, and the method was efficient from the single-molecule level. Also, had an L1 inserted into one of the target loci during spermatogenesis, it could have had no detrimental effect on the selected donor and should have been represented in the donor's sperm DNA. The eight selected genic targets were therefore representative of any randomly selected 5 kb of gDNA.

The major exception was the HOXD target. A mechanism appears to exist that insulates the HOXD gene cluster from the effects of mobile element insertion (Lander *et al.*, 2001). We cannot exclude the possibility that there is a cellular mechanism that prevents mobile element insertion into the HOXD gene cluster. However, it is more likely that mobile element insertion into the gene cluster results in embryonic lethality due to the importance of the HOXD genes during early development (Greally, 2002). There is evidence that L1 retrotransposition may be restricted to as few as two "windows of opportunity" when L1 promoters are not heavily methylated, one during early embryogenesis and one in the primordial germ cells (Kierszenbaum, 2002; Li, 2002; Mann, 2001). Embryonic lethality caused by insertion into the HOXD gene cluster means that one of the "windows of

opportunity" must have been missed in a living donor, thus reducing the probability of recovering *de novo* insertion into the HOXD target.

The MHC2 target was a randomly selected 5 kb region of gDNA, chosen due to a high level of understanding of genetic variation in this region. Therefore, with the possible exception of the HOXD target, the selected target loci were essentially surrogates for any other locus. Although it could be argued that selecting ten autosomal loci may have maximised the screening area (given that the X chromosome is only present in 50 % of sperm), the X chromosome is nearly two-fold enriched for L1 compared to the autosomes, (Bailey *et al.*, 2000; Lander *et al.*, 2001; Ross *et al.*, 2005), so may be a preferential target. As a result we assert that the selected target loci are in essence as good as any other randomly selected target site, so it is unlikely that the low observed rate of insertion can be attributed to poor target site selection.

9.1.3.b Did the selected donor exhibit an abnormally low rate of L1 insertion in the germline?

The donor selected for this investigation was a healthy North Western European male in his mid to late 50s. Given that there have been no investigations carried out to determine different rates of L1 insertion between different ethnic groups, a North Western European male was used due to the availability of such men in the UK. The main reason for selecting this individual donor from the departmental donor panel was that he had a high sperm count, so large quantities of sperm gDNA could be recovered from a single ejaculate. However, it was possible that the donor had an unusually low rate of L1 retrotransposition.

The average human genome is predicted to harbour between 80 and 100 potentially retrotransposition competent (RC) L1s (Brouha *et al.*, 2003). It is therefore likely that the selected donor also had numerous RC L1s in his genome. Although the selected donor does not have the most active L1 insertion identified in the human genome sequence AC002980 (Brouha *et al.*, 2003), five hot L1s remain that account for 63 % of the summed activity of all L1s identified in the human genome sequence (Brouha *et al.*, 2003). Given the estimated allele frequencies of the remaining "hot" L1s, from a sample of 23 individuals of mixed ethnic groups (European, Chinese, Indio-Pakistani, Pacific and Sub-Saharan African) (Brouha *et al.*, 2003), it is likely that the donor had at least one of these "hot" L1s as well as between 80 to 100 other potentially active L1s in his genome. The donor also showed the presence of

two active L1s not contained in the human genome sequence (AL121819 +/-, and AL358779 +/-). Thus it seems unlikely, in terms of active L1 representation, that the donor would have exhibited an abnormally low level of L1 retrotransposition assuming his germline was amenable to L1 retrotransposition.

9.1.3.c Was sperm a suitable source of DNA for this investigation?

9.1.3.c.i Mobile elements must transpose in the germline

Molecular parasites like L1 are selfish in that their whole structure is geared towards generating offspring elements (Bestor, 1999; Hickey, 1982). Such elements rely on mobilisation and insertion into the germline of sexual species to colonise numerous genomes, and increase their prevalence in a population (Cavalier-Smith, 1980).

Throughout history, non-LTR retroelements have retrotransposed in the germlines of sexually reproducing organisms. Non-LTR retrotransposons originated over 600 million years ago (MYA), and their ability to retrotranspose in the germline has meant that numerous diverse families of non-LTR retroelements can be identified in the vast majority of eukaryotic organisms (Malik *et al.*, 1999) (chapter 1). However non-LTR elements such as L1 sometimes have serious deleterious consequences for the genome into which they insert (Cavalier-Smith, 1980). Thus, if the *de novo* insertion does not eliminate the infected genome, perhaps through embryonic lethality, then it has a chance of being passed onto the following generation through sexual reproduction. However insertions with a low mutational burden, or insertions that by chance confer a benefit to the organism, are more likely to be successful in moving to fixation within a population (Bestor, 1999; Hickey, 1982).

In a sexually reproducing organism, mobile element insertion in somatic cells can at most lead to a high level of somatic mosaicism, and the insertion is removed from the population following the death of the individual in which the insertion occurred. Somatic retrotransposition of L1 is therefore of no benefit to the element, and can cause damage to the host. For example a somatic cell (most likely a stem cell) L1 insertion into the Adenomatosis polyposis coli (APC) gene has previously caused colorectal cancer (Miki *et al.*, 1992). Therefore experimental investigations must concentrate on the germline in order to recover *de novo* L1 insertions.

9.1.3.c.ii The germline is a cellular "battle-ground"

Since parasitic mobile elements, such as L1, are geared towards germline mobilisation, the germline has become a "battle-ground." Host cell mechanisms have evolved to suppress mobile elements and protect subsequent generations from the deleterious effects of insertion (Bestor, 1999). The majority of LINE elements are densely methylated in normal somatic cells which inactivates their promoters, and this is thought to be a cellular defence against the harmful effects of L1 retrotransposition (Hata and Sakaki, 1997). Mammalian genomes are heavily methylated in germ cells of both sexes (Kierszenbaum, 2002; Li, 2002; Mann, 2001; Walsh *et al.,* 1998) which represses L1 promoter activity (Hata and Sakaki, 1997). Also, the APOBEC3B protein, a potent inhibitor of L1 retrotransposition, can be readily detected in human ovaries, testes and embryonic stem cells (Bogerd *et al.*, 2006b). This suggests that an L1 inhibitory force is expressed throughout germ cell development and early embryogenesis (Bogerd *et al.,* 2006b). The combination of heavy methylation of L1 promoters, and the expression of APOBEC3B during gametogenesis would seem to be a formidible host defence against L1 retrotransposition. In contrast, the identification of L1 ORF1p, ORF2p, L1 RNA in the human testes suggest that not only is L1 being transcribed in the male germline, but the element encoded proteins are also being translated (Ergun *et al.*, 2004). However, protein translation does not necessarily lead to retrotransposition, as demonstrated by APOBEC3B inhibition of L1 in cell culture (Bogerd *et al.,* 2006b).

9.1.3.c.iii L1 retrotransposition may be limited to distinct "windows of opportunity"

Hypomethylation in primordial germ cells removes transcriptional blocks, prior to *de novo* sex-specific imprinting during gametogenesis (Kierszenbaum, 2002; Li, 2002; Mann, 2001). Numerous spermatid-specific and oocyte-specific methylation modifications are also removed post fertilisation prior to *de novo* methylation during early embryogenesis (Brandeis *et al.*, 1993). These de-methylation events provide two potential "windows of opportunity" for the translation of mobile elements (Bestor, 1999; Schulz *et al.,* 2006). Interestingly both potential "windows of opportunity" occur extremely early in terms of gametogenesis.

Retrotransposition of L1 during the removal of methylation in early embryogenesis will lead to somatic and germline mosaicism. So long as the insertion does not cause

embryonic lethality, it likely partitions into the germline and thus can be transmitted to subsequent generations. Also the removal of methylation in primordial germ cells means that L1 may be able to retrotranspose at the very beginning of gametogenesis. Such "jackpot" insertions of L1 during these very early stages of germline development may be extremely advantageous, allowing a single *de novo* L1 insertion to colonise numerous germ cells.

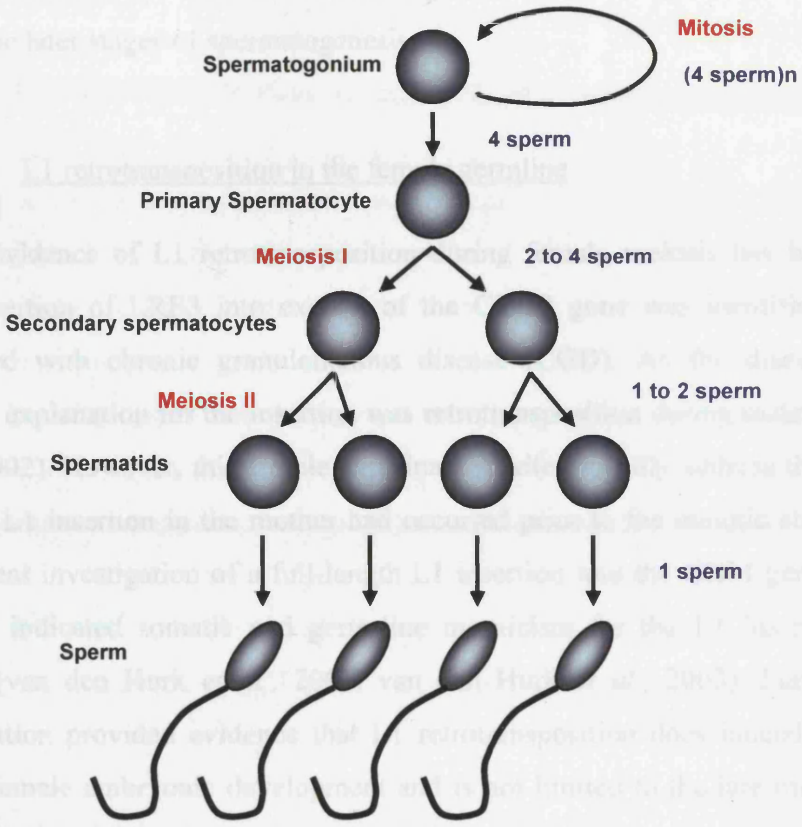


Figure 9.1: A simple representation of spermatogenesis. The blue text shows the number of sperm containing a *de novo* L1 insertion if retrotransposition occurred at that stage of spermatogenesis.

9.1.3.c.iv L1 retrotransposition in the male germline

Two germ cells are required to make an individual human each of which has "competed" with millions of other germ cells. If *de novo* L1 insertions are not represented in a high percentage of germ cells, there is a low probability that it will be passed to subsequent generations.

Previous estimations of the rate of L1 insertion into the genome assumed that the progenitor *de novo* insertion occurred during the meiotic stages of germ cell development. If retrotransposition was limited to the later stages of germ cell development, meiotic stages, it would have to retrotranspose at a much higher rate to ensure the maximum number of germ

cells possible carried a *de novo* L1 insertion (Figure 9.1). However, the sperm genome is heavily methylated (Kierszenbaum, 2002; Li, 2002; Mann, 2001; Walsh *et al.*, 1998) and the nucleus is packaged into a highly compact structure to create a mobile cell for male genome delivery (Lee *et al.*, 1995; Lewis *et al.*, 2004; Lewis *et al.*, 2003; Wouters-Tyrou *et al.*, 1998). The densely packaged sperm chromatin is likely to be a very poor substrate for the L1 endonuclease activity, and thus it is unlikely that L1 retrotransposition would be able to occur during the later stages of spermatogenesis.

9.1.3.c.v L1 retrotransposition in the female germline

Evidence of L1 retrotransposition during female meiosis has been presented. A *de novo* insertion of LRE3 into exon 4 of the CYBB gene was identified in a male patient diagnosed with chronic granulomatous disease (CGD). As the disease is X-linked, the simplest explanation for the insertion was retrotransposition during maternal meiosis (Brouha *et al.*, 2002). However, this simple explaination failed to fully address the possibility that the *de novo* L1 insertion in the mother had occurred prior to the meiotic stages of oogenesis. A subsequent investigation of a full-length L1 insertion into the CHM gene of a Dutch patient ($L1_{CHM}$) indicated somatic and germ-line mosaicism for the L1 insertion in the patient's mother (van den Hurk *et al.*, 2007; van den Hurk *et al.*, 2003). Further, the subsequent investigation provided evidence that L1 retrotransposition does indeed occur very early in human female embryonic development and is not limited to the late meiotic stages of germ cell development (van den Hurk *et al.*, 2007; van den Hurk *et al.*, 2003).

9.1.3.c.vi L1 retrotransposition in human embryonic stem cells

Using human embryonic stem cells as a model of embryogenesis, more evidence consistent with L1 retrotransposition during early embryogenesis has recently been provided by Garcia-Perez *et al.*, (2007). This group demonstrated the presence of endogenous L1 RNA, ORF1p and ORF 2p in undifferentiated human embryonic stem cells, suggesting that L1 elements may be being transcribed and translated during early embryogenesis.

In this study, undifferentiated embryonic stem cells were transfected with a marked retrotransposition competent L1, reminiscent of marked L1 elements used during the cultured cell retrotransposition assay (chapter 1), whose expression was driven from its own promoter (Garcia-Perez *et al.*, 2007). The marked L1s showed retrotransposition in the cells, but the

estimated efficiency of retrotransposition was approximately an order of magnitude lower than observed in transformed cell lines (Garcia-Perez et al., 2007). The low efficiency of retrotransposition may have been due to integration of the L1 into heterochromatin and silencing of the reporter cassette (Muotri *et al.,* 2005), or the presence of endogenous retroelement repressor proteins (Garcia-Perez *et al.,* 2007) for example APOBEC3B. However it also may have been due to the titration of cellular factors required for efficient L1 retrotransposition. Unlike *HeLa* cells, these embryonic stem cells express their endogenous L1s at a high level (Garcia-Perez *et al.,* 2007; Han and Boeke, 2004; Moran *et al.,* 1996; Morrish *et al.,* 2002; Wei *et al.,* 2000), and so the tagged L1 may have had to compete with cellular L1s for cellular factors that assist retrotransposition. Further, it was demonstrated that the embryonic stem cells were then capable of differentiating into embryoid bodies that expressed genes diagnostic for the three germ layers, showing that the marked L1s had retrotransposed in undifferentiated embryonic stem cells (Garcia-Perez *et al.,* 2007).

The most direct way to prevent a mobile element from transposing is to prevent expression. It is likely that L1 retrotransposition is almost completely repressed when the genome is methylated due to promoter inactivation; however, methylation has to be removed during certain stages of embryogenesis and gametogenesis. During these periods, it may be that APOBEC3B and other cellular repressor proteins act to restrict the level of L1 retrotransposition of expressed elements, thus maintaining a low level of retrotransposition.

9.1.3.c.vii How do these findings fit with this investigation?

The data presented in this investigation could be consistent with rare early embryonic retrotransposition of L1. Previous estimates of rates of L1 insertion into the genome predict a single insertion in between 4.8 x 10^7 kb (1 in 8 people) and 1.5 x 10^9 kb (1 in 250 people) of gDNA (Kazazian, 1999; Li *et al.,* 2001). As 9.8 x 10^9 kb of sperm gDNA was screened in this experiment, and assuming the target loci were representative of randomly selected genomic loci, and that the developed method was 32 % (thus 3.14 x 10^9 kb was in effect screened) efficient at amplifying full-length L1 insertions, between 2 and 65 *de novo* L1 insertions were expected to be recovered in to the ten target site loci. However, no insertions were recovered, so the rate of L1 insertion into the genome of the selected donor at these loci was < 3 in 4.35 x 10^8 kb of gDNA (< 3 in 145 people).

If L1 retrotransposition were limited to the meiotic stages of gametogenesis, such a low rate of L1 insertion would likely be inconsistent with the proliferation and survival of L1. By retrotransposition in primordial germ cells, or prior to differentiation of the germline, "jackpot" *de novo* L1 insertion can colonise numerous gametes, and thus has a greater chance of being passed to the next generation, even if this only occurs in very few individuals. It is therefore likely that a low proportion of males in the population have a high L1 load (high proportion of germ cells carrying a *de novo* insertion).

9.1.3.c.viii Was sperm a suitable source of DNA for this investigation?

Since it is possible that differential methylation during oocyte and spermatid development may cause different L1 retrotransposition dynamics in the male and female germlines (Davis *et al.*, 1999; Davis *et al.*, 2000; Kerjean *et al.*, 2000; Lucifero *et al.*, 2004; Ueda *et al.*, 2000), and there is no direct evidence linking L1 retrotransposition to the male germline (for example a male transmitted disease-causing insertion), it could be suggested that sperm gDNA was the wrong tissue in which to look for *de novo* L1 retrotransposition.

There is circumstantial evidence suggesting that L1 could retrotranspose in the male germline. For example the presence of L1 ORF1p and ORF2p, in the human testes (Ergun *et al.*, 2004). Further, a polymorphic L1 insertion into the centromere of the Y chromosome is difficult to explain if L1 does not retrotranspose in the male germline (Santos et al., 2000). Also, as L1 has to retrotranspose in the germline in order to proliferate, and sperm is the only readily accessible source of human germline gDNA, it was the only tissue available to screen for *de novo* L1 insertions.

If the rate of L1 retrotransposition were actually very low (due to retrotransposition occurring early in spermatogenesis), and L1 insertion into the genome is random, there is a low probability that a *de novo* L1 insertion will insert into the 0.0017 % of the genome screened during the investigation. Therefore in order to recover a *de novo* L1 insertion into one of the target loci, it is likely that numerous donors would have to be screened since the selected donor displayed no insertions into the ten target loci. Alternatively the same donor could be screened using different target loci.

The developed method has the potential to identify germline mosaicism, as would result from *de novo* L1 insertion prior to spermatogonial differentiation. This would also make the insertions easier to recover by hybridisation enrichment since they would be present

as numerous copies prior to multiplex PCR, and the method was developed to recover DNA generated from a single-molecule. If L1 retrotransposition occurs during early embryogenesis, but very rarely (i.e. 1 in 1,000 people), individuals who carry a *de novo* insertion are likely to pass it on, so maintaining L1 population diversity.

9.2 Future directions

Plans to continue the investigation are being considered, which address some of the potential reasons for the low observed rate of L1 retrotransposition.

To avoid the possibility that the selected donor exhibited an abnormally low rate of L1 retrotransposition, one approach is to screen between 48 μg and 96 μg of sperm gDNA from several different individuals (possibly from different ethnic origins) simultaneously. Multiplex PCR will be performed using the same target loci, and perhaps other target loci. It is also possible that the range of multiplex PCR can be increased by adding more target site loci. The theoretical maximum primer concentration for PCR to work efficiently was estimated to be 2.4 μM (chapter 5). Decaplex PCR contained a total primer concentration of approximately 1.6 μM (chapter 5). This gave an average of 0.16 μM per target site. Thus multiplex PCR could theoretically be expanded to simultaneously amplify 15 loci.

This alteration to the method also addresses the likelihood that L1 retrotransposition is limited to very early stages of gametogenesis and embryogenesis. This investigation showed that *de novo* L1 insertion into the selected target loci had not occurred in the selected donor. However, by screening numerous donors, the probability of a donor having an L1 insertion in one of the target loci increases. Interestingly, since early embryonic insertion will result in germline mosaicism, *de novo* L1 insertions will be easier to recover than if they insert during meiosis, providing the correct male was selected. The developed method was designed to recover a single-molecule of a *de novo* L1 insertion, but if an individual were a germline mosaic for the insertion, sperm gDNA would contain far more than a single-molecule of a *de novo* L1 insertion. Also, if a *de novo* L1 insertion occurred during early embryogenesis insertion, prior to differentiation of the soma and the germline, and was neutral or advantageous (i.e. did not result in cells that could not grow competitively during development); it would likely be detectable in other tissues such as blood or foetal DNA. This would give an additional level of validation to the method.

During this investigation I have developed a method capable of amplifying and recovering L1-containing DNA from the single-molecule level. I have provided evidence that

the method is capable of recovering *de novo* L1 insertions from human sperm gDNA. Although I did not achieve recovery of a *de novo* L1 insertion during the course of the investigation, I have laid the foundations of a continuing investigation. I believe this investigation will eventually result in recovery of *de novo* L1 insertions that will most likely show germline and possibly somatic mosaicism in the donor.

Appendix

Appendix i

Tables showing known retroelement insertions into the human genome which have linked to caused genetic disease. The tables were taken from http://www.med.upenn.edu/genetics/kazazianlab_human.shtml, and adapted to show the most recent disease causing L1 insertion.

L1 insertions

Inserted element	Disrupted gene	Insertion size (kb)	3' Transduction (y/n)	Insertion site	Orientation of insertion	Reference
JH-27	Factor viii	3.8	n	Exon	Sense	(Kazazian *et al.*, 1988)
JH-28	Factor viii	2.2	n	Exon	Sense and rearranged	(Kazazian *et al.*, 1988)
JH-25	Factor viii	0.681	n	Intron	Sense	(Woods-Samuels *et al.*, 1989)
APC	APC	0.538	y	Exon	Unknown	(Miki *et al.*, 1992)
Dystrophin	Dystrophin	0.608	n	Exon	Sense	(Narita *et al.*, 1993a)
Dystrophin	Dystrophin	0.878	n	Exon	Sense	(Bakker and van Ommen)
JH-1001	Dystrophin	2	y	Exon	Sense and rearranged	(Holmes *et al.*, 1994)
L1b-thal	β-globin	6	n	Intron	Antisense	(Divoky *et al.*, 1996)
L1XLCDM	Dystrophin	0.524	n	Exon	Antisense	(Yoshida *et al.*, 1998)
L1RP	RP2	6	n	Intron	Antisense	(Schwahn *et al.*, 1998)
L1CYB	CYBB	1.7	y	Exon	Unknown	(Brouha *et al.*, 2002; Meischl *et al.*, 1998)
L1CYB	CYBB	0.94	n	Intron	Sense	(Meischl *et al.*, 2000)
L1FCMD	Fukutin	1.1	n	Intron	Sense (no TSD - 7 nt deletion)	(Kondo-Iida *et al.*, 1999)
L1FIX	Factor ix	0.52	n	Exon	Sense	(Li *et al.*, 2001)
L1PDHX	PDHX	6	n	Intron (large genomic deletion)	Sense	(Mine *et al.*, 2007)
L1CHM	CHM	6	n	Exon	Antisense	(van den Hurk *et al.*, 2007; van den Hurk *et al.*, 2003)

SVA related insertions

Disrupted gene	Insertion type	Reference
FcMB	Full length	(Kobayashi *et al.*, 1998)
BTK	5p truncated (SINE R)	(Rohrer *et al.*, 1999)
α-spectrin	SVA mediated transduction	(Hassoun *et al.*, 1994)

Alu insertions

Gene	Disorder	Alu subfamily	Insertion site	Orientation	*De novo* (y/n)	Reference
NF1	Neurofibromatosis	Ya5	Intron	Antisense	y	(Wallace *et al.*, 1991)
BCHE	Acholinesterasemia	Yb8	Exon	Sense	n	(Muratani *et al.*, 1991)
FIX	Haemophilia B	Ya5	Exon	Sense	y	(Vidaud *et al.*, 1993)
CASR	Familial hypocalciuric hypercalcemia	Ya4	Exon	Antisense	n	(Janicic *et al.*, 1995)
BRCA2	Breast cancer	Y	Exon	Sense	unknown	(Miki *et al.*, 1996)
APC	Hereditary desmoid disease	Yb8	Exon	Sense	n	(Halling *et al.*, 1999)
BTK	X-linked agammaglobulinemia	Y	Exon	Antisense	y	(Lester *et al.*, 1997)
IL2RG	X-linked severe combined immunodeficiency	Ya5	Intron	Antisense	n	(Lester *et al.*, 1997)
EYA1	Branchio-oto-renal syndrome	Ya5	Exon	Antisense	y	(Abdelhak *et al.*, 1997)
FGFR2	Apert syndrome	Ya5	Intron	Antisense	y	(Oldridge *et al.*, 1999)
FGFR2	Apert syndrome	Yb8	Exon	Antisense	y	(Oldridge *et al.*, 1999)
ADD1	Huntington disease	Unknown	Intron	Sense	n	(Goldberg *et al.*, 1993)
GK	Glycerol kinase deficiency	Ya5	Intron	Antisense	unknown	(Zhang *et al.*, 2000)
C1NH	C1 inhibitor deficiency	Y	Intron	Sense	n	(Stoppa-Lyonnet *et al.*, 1990)
PBGD	Acute intermittent porphyria	Ya5	Exon	Antisense	n	(Mustajoki *et al.*, 1999)
MIVI-2	Associated with leukemia	Ya5	Unknown	Unknown	y (Somatic ?)	(Economou-Pachnis and Tsichlis, 1985)
FIX	Haemophilia B	Ya3al	Exon	Antisense	n	(Li *et al.*, 2001)
FIX	Haemophilia B	Unknown	Exon	Sense	n	(Wulff *et al.*, 2000)
FVIII	Haemophilia A	Yb8	Exon	Antisense	n	(Swergold, 1990)

Appendix ii

Materials and Methods solutions

The solutions used during the investigation are listed below.

Denaturing solution

1.5 M NaCl

0.5 M NaOH

Depurinating solution

(0.2N HCl)

10 x DHB

450 mM Tris-HCl pH 8.8

110 mM ammonium sulphate

45 mM MgCl2

67 mM 2-mercaptoethanol

44mM EDTA

20mg/ml single-stranded (heat denatured) high M.W. herring sperm DNA

Dotblot denaturing solution

0.5 M NaOH

2 M NaCl

25 mM EDTA

IPTG

0.48 g IPTG

Top up to 20 ml with 18 MΩ distilled water

(Makes 20 aliquots)

LB plates

400 ml Luria Agar

Boil until clear (all agar melted)

Cool in a 37 °C water bath for 15 minutes

Pour plates (approximately 20)

LB Ampicillin plates

As for the LB plates, but add 800 µl 100 mg/ml Ampicillin prior to pouring the plates.

Modified CHURCH buffer

Neutralising solution

1 M Tris pH 7.5

1.5M NaCl (0.5M Tris pH 7.2, 1M NaCl))

Phosphate Wash Solution

0.04 M Na_2HPO_4

0.5 % SDS

Single-molecule diluent (SMD)

5 ng/µl *E.coli* DNA

5 mM Tris-HCl (pH 7.5)

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 1M $MgCl_2$

Store at 4 °C

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

pH to 7 with 1 M NaOH

bring up to 990 ml with 18 MΩ distilled water

Autoclave

Top up to 1 L with 1M $MgCl_2$

Store at 4 °C

Add glucose to 4 mg/ml prior to use

20 x SSC

3.0 M NaCl

0.3 M Sodium Acetate

Adjusted to pH 7 with 14 N HCl

TAE (1 x Tris-Acetate EDTA)

40 mM Tris-acetate

1 mM EDTA

TBE Loading buffer

0.5X TBE

12.5% Ficoll,

Bromophenol blue added to the desired tone

TBE (0.5 x Tris-Borate EDTA)

45 mM Tris-borate

1 mM EDTA

TB

To make 0.5 L

1.5 g 10 mM PIPES

1.1 g 15 mM $CaCl_2.H_2O$

7.3 g 250 mM KCl

pH to 6.7 with 10 M KOH

5.45 g $MnCl_2.4H_2O$

Make up to 500 ml then filter sterilise

Store at 4 °C

TE buffer

10mM Tris-HCl pH 8.0

1mM EDTA

Xgal

20 ml Dimethylformamide

1 g Xgal

(Makes 20 aliquots)

Keep at – 20 °C (light sensitive)

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 μM EDTA (pH 8)

1 mM dATP

1 mM dCTP

1 mM dGTP

1 mM dTTP

113 μg/ml BSA

Appendix iii

Additional primers used for PCR and Oligo-hybridisation

Primer ID	Sequence	Accession
PFENopE_PVU2A	TGAAACAAGAACGGCCCTGG	AC008575
PFENopE_PVU2B	TTTTCCTGGCCCTTTCCAAGGG	AC008575
PFENopE_PVU2D	CAGCTGACTTTTTGGGGAAGG	AC008575
PFOP_ENeA	TCCCTATAGTCCCCTTCTGG	AC008575
PFOP_ENeB	TGCCAATGAGGGCACATGG	AC008575
PFOP_ENeC	GCTTACTACTGGGATGTGGG	AC008575
PFOP_EneD	ATGTCCAGTGAGGTGGATGG	AC008575
PFOP_ENeE	GCTCCCTTCCATCAATATAGG	AC008575
RB853LRR1	CTGCCAAGAGGTGCTTAGGG	AL583853
RB853LRF	ACACCCTATATACCTCGGTG	AL583853
RB853E	AGTCCAAGAGTCAGAGAGCTGCAC	AL583853
RB853H	AAGTATGATGACAGGGGG	AL583853
RB853D	TGCTGGCTAAATTTCCCAGAAGTA	AL583853
PFOP_ENfA	CAAAGCCGTCTCTACTTGGG	AL583853
PFOP_ENfB	AGCCCACATACCATAGCAGG	AL583853
RB819C	GAGAACAGGATAGAGCCAAGAT	AL121819
RB819B	ATCCCCATTACCACATCTCATTT	AL121819
RBLR2519	ATTAGGTCTGCTTGGTGC	L1 specific
RBLR4015	TGGCCAGAACTTCCAACA	L1 specific
PFLF5946	GCACCAGCATGGCACATG	L1 specific
PFLF5981	CCTGCACAATGTGCACATG	L1 specific
PFLR5999	CATGTGCACATTGTGCAGG	L1 specific
RB3PA2	ACCTAATGCTAGATGACACA	L1 specific

References

Alisch, R. S., Garcia-Perez, J. L., Muotri, A. R., Gage, F. H., and Moran, J. V. (2006). Unconventional translation of mammalian LINE-1 retrotransposons. Genes Dev *20*, 210-224.

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. Oncol Res *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 Res *32*, 3846-3855.

Badge, R. M., Alisch, R. S., and Moran, J. V. (2003). ATLAS: a system to selectively identify human-specific L1 insertions. Am J Hum Genet *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. Proc Natl Acad Sci U S A *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. Am J Hum Genet *73*, 823-834.

Baltimore, D. (1970). RNA-dependent DNA polymerase in virions of RNA tumour viruses. Nature *226*, 1209-1211.

Bellefroid, E. J., Lecocq, P. J., Benhida, A., Poncelet, D. A., Belayew, A., and Martial, J. A. (1989). The human genome contains hundreds of genes coding for finger proteins of the Kruppel type. DNA *8*, 377-387.

Bellefroid, E. J., Poncelet, D. A., Lecocq, P. J., Revelant, O., and Martial, J. A. (1991). The evolutionarily conserved Kruppel-associated box domain defines a subfamily of eukaryotic multifingered proteins. Proc Natl Acad Sci U S A *88*, 3608-3612.

Bestor, T. H. (1999). Sex brings transposons and genomes into conflict. Genetica *107*, 289-295.

Bibillo, A., and Eickbush, T. H. (2002). The reverse transcriptase of the R2 non-LTR retrotransposon: continuous synthesis of cDNA on non-continuous RNA templates. J Mol Biol *316*, 459-473.

Bibillo, A., and Eickbush, T. H. (2004). End-to-end template jumping by the reverse transcriptase encoded by the R2 retrotransposon. J Biol Chem *279*, 14945-14953.

Bogerd, H. P., Wiegand, H. L., Doehle, B. P., Lueders, K. K., and Cullen, B. R. (2006a). APOBEC3A and APOBEC3B are potent inhibitors of LTR-retrotransposon function in human cells. Nucleic Acids Res *34*, 89-95.

Bogerd, H. P., Wiegand, H. L., Hulme, A. E., Garcia-Perez, J. L., O'Shea, K. S., Moran, J. V., and Cullen, B. R. (2006b). Cellular inhibitors of long interspersed element 1 and *Alu* retrotransposition. Proc Natl Acad Sci U S A *103*, 8780-8785.

Boissinot, S., Chevret, P., and Furano, A. V. (2000a). L1 (LINE-1) retrotransposon evolution and amplification in recent human history. Mol Biol Evol *17*, 915-928.

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

Boissinot, S., Entezam, A., and Furano, A. V. (2001). Selection against deleterious LINE-1-containing loci in the human lineage. Molecular Biology & Evolution *18*, 926-935.

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

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

Brandeis, M., Kafri, T., Ariel, M., Chaillet, J. R., McCarrey, J., Razin, A., and Cedar, H. (1993). The ontogeny of allele-specific methylation associated with imprinted genes in the mouse. Embo J *12*, 3669-3677.

Bratthauer, G. L., and Fanning, T. G. (1992). Active LINE-1 retrotransposons in human testicular cancer. Oncogene *7*, 507-510.

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

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

Brosius, J., and Tiedge, H. (1995). Reverse transcriptase: mediator of genomic plasticity. Virus Genes *11*, 163-179.

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. Am J Hum Genet *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. Proc Natl Acad Sci U S A *100*, 5280-5285.

Brown, C. J., Hendrich, B. D., Rupert, J. L., Lafreniere, R. G., Xing, Y., Lawrence, J., and Willard, H. F. (1992). The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. Cell *71*, 527-542.

Bucheton, A., Busseau, I., and Teninges, D. (2002). I transposable elements in *Drosophila melanogaster*, In Mobile DNA II, N. L. Craig, R. Craigie, M. Gellert, and A. M. Lambowitz, eds. (Washington, D.C.: ASM Press), pp. 796 - 835.

Bucheton, A., Lavige, J. M., Picard, G., and L'Heritier, P. (1976). Non-mendelian female sterility in Drosophila melanogaster: quantitative variations in the efficiency of inducer and reactive strains. Heredity *36*, 305-314.

Bucheton, A., Paro, R., Sang, H. M., Pelisson, A., and Finnegan, D. J. (1984). The molecular basis of I-R hybrid dysgenesis in Drosophila melanogaster: identification, cloning, and properties of the I factor. Cell *38*, 153-163.

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 Res *31*, 4385-4390.

Buzdin, A., Ustyugova, S., Gogvadze, E., Vinogradova, T., Lebedev, Y., and Sverdlov, E. (2002). A new family of chimeric retrotranscripts formed by a full copy of U6 small nuclear RNA fused to the 3' terminus of I1. Genomics *80*, 402-406.

Carroll, M. L., Roy-Engel, A. M., Nguyen, S. V., Salem, A. H., Vogel, E., Vincent, B., Myers, J., Ahmad, Z., Nguyen, L., Sammarco, M., *et al.* (2001). Large-scale analysis of the *Alu* Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. J Mol Biol *311*, 17-40.

Casavant, N. C., and Hardies, S. C. (1994). Shared sequence variants of Mus spretus LINE-1 elements tracing dispersal to within the last 1 million years. Genetics *137*, 565-572.

Cavalier-Smith, T. (1980). How selfish is DNA? Nature *285*, 617-618.

Chalker, D. L., and Sandmeyer, S. B. (1992). Ty3 integrates within the region of RNA polymerase III transcription initiation. Genes Dev *6*, 117-128.

Charlesworth, B., and Langley, C. H. (1989). The population genetics of Drosophila transposable elements. Annu Rev Genet *23*, 251-287.

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 Res *12*, 894-908.

Clemson, C. M., McNeil, J. A., Willard, H. F., and Lawrence, J. B. (1996). XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure. J Cell Biol *132*, 259-275.

Collins, F. S., Lander, E. S., Rogers, j., Waterston, R. H., and the International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931-945.

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

Costas, J. (2002). Characterization of the intragenomic spread of the human endogenous retrovirus family HERV-W. Mol Biol Evol *19*, 526-533.

Cullen, B. R. (2006). Role and mechanism of action of the APOBEC3 family of antiretroviral resistance factors. J Virol *80*, 1067-1076.

Dahl, H. H., Brown, R. M., Hutchison, W. M., Maragos, C., and Brown, G. K. (1990). A testis-specific form of the human pyruvate dehydrogenase E1 α subunit is coded for by an intronless gene on chromosome 4. Genomics *8*, 225-232.

Davis, T. L., Trasler, J. M., Moss, S. B., Yang, G. J., and Bartolomei, M. S. (1999). Acquisition of the H19 methylation imprint occurs differentially on the parental alleles during spermatogenesis. Genomics *58*, 18-28.

Davis, T. L., Yang, G. J., McCarrey, J. R., and Bartolomei, M. S. (2000). The H19 methylation imprint is erased and re-established differentially on the parental alleles during male germ cell development. Hum Mol Genet *9*, 2885-2894.

Devine, S. E., and Boeke, J. D. (1996). Integration of the yeast retrotransposon Ty1 is targeted to regions upstream of genes transcribed by RNA polymerase III. Genes Dev *10*, 620-633.

Dewannieux, M., Esnault, C., and Heidmann, T. (2003). LINE-mediated retrotransposition of marked *Alu* sequences. Nat Genet *35*, 41-48.

Dhellin, O., Maestre, J., and Heidmann, T. (1997). Functional differences between the human LINE retrotransposon and retroviral reverse transcriptases for in vivo mRNA reverse transcription. Embo J *16*, 6590-6602.

Divoky, V., Indrak, K., Mrug, M., Brabec, V., Huisman, T. H. J., and Prchal, J. T. (1996). A novel mechanism of beta thalassemia: The insertion of L1 retrotransposable element into beta globin IVS II. Blood *88*, 580-580.

Dmitriev, S. E., Andreev, D. E., Terenin, I. M., Olovnikov, I. A., Prassolov, V. S., Merrick, W. C., and Shatsky, I. N. (2007). Efficient translation initiation directed by the 900-nucleotide-long and GC-rich 5' untranslated region of the human retrotransposon LINE-1 mRNA is strictly cap dependent rather than internal ribosome entry site mediated. Mol Cell Biol *27*, 4685-4697.

Dombroski, B. A., Feng, Q., Mathias, S. L., Sassaman, D. M., Scott, A. F., Kazazian, H. H., Jr., and Boeke, J. D. (1994). An in vivo assay for the reverse transcriptase of human retrotransposon L1 in Saccharomyces cerevisiae. Mol Cell Biol *14*, 4485-4492.

Dombroski, B. A., Mathias, S. L., Nanthakumar, E., Scott, A. F., and Kazazian, H. H., Jr. (1991). Isolation of an active human transposable element. Science *254*, 1805-1808.

Dupuy, D., Aubert, I., Duperat, V. G., Petit, J., Taine, L., Stef, M., Bloch, B., and Arveiler, B. (2000). Mapping, characterization, and expression analysis of the SM-20 human homologue, c1orf12, and identification of a novel related gene, SCAND2. Genomics *69*, 348-354.

Dupuy, D., Duperat, V. G., and Arveiler, B. (2002). SCAN domain-containing 2 gene (SCAND2) is a novel nuclear protein derived from the zinc finger family by exon shuffling. Gene *289*, 1-6.

Dutko, J. A., Schafer, A., Kenny, A. E., Cullen, B. R., and Curcio, M. J. (2005). Inhibition of a yeast LTR retrotransposon by human APOBEC3 cytidine deaminases. Curr Biol *15*, 661-666.

Eichler, E. E., Lu, F., Shen, Y., Antonacci, R., Jurecic, V., Doggett, N. A., Moyzis, R. K., Baldini, A., Gibbs, R. A., and Nelson, D. L. (1996). Duplication of a gene-rich cluster between 16p11.1 and Xq28: a novel pericentromeric-directed mechanism for paralogous genome evolution. Hum Mol Genet *5*, 899-912.

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

Engels, W. R., and Preston, C. R. (1979). Hybrid dysgenesis in Drosophila melanogaster: the biology of female and male sterility. Genetics *92*, 161-174.

Ergun, S., Buschmann, C., Heukeshoven, J., Dammann, K., Schnieders, F., Lauke, H., Chalajour, F., Kilic, N., Stratling, W. H., and Schumann, G. G. (2004). Cell type-specific expression of LINE-1 open reading frames 1 and 2 in fetal and adult human tissues. J Biol Chem *279*, 27753-27763.

Esnault, C., Heidmann, O., Delebecque, F., Dewannieux, M., Ribet, D., Hance, A. J., Heidmann, T., and Schwartz, O. (2005). APOBEC3G cytidine deaminase inhibits retrotransposition of endogenous retroviruses. Nature *433*, 430-433.

Eyre-Walker, A., and Keightley, P. D. (1999). High genomic deleterious mutation rates in hominids. Nature *397*, 344-347.

Fan, H., Goodier, J. L., Chamberlain, J. R., Engelke, D. R., and Maraia, R. J. (1998). 5' processing of tRNA precursors can Be modulated by the human La antigen phosphoprotein. Mol Cell Biol *18*, 3201-3211.

Farley, A. H., Luning Prak, E. T., and Kazazian, H. H., Jr. (2004). More active human L1 retrotransposons produce longer insertions. Nucleic Acids Res *32*, 502-510.

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

Feuk, L., Carson, A. R., and Scherer, S. W. (2006). Structural variation in the human genome. Nat Rev Genet *7*, 85-97.

Furano, A. V. (2000). The biological properties and evolutionary dynamics of mammalian LINE-1 retrotransposons. Prog Nucleic Acid Res Mol Biol *64*, 255-294.

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. Hum Mol Genet *16*, 1569-1577.

Gartler, S. M., and Riggs, A. D. (1983). Mammalian X-chromosome inactivation. Annu Rev Genet *17*, 155-190.

Gaudin, P., Ijaz, S., Tuke, P. W., Marcel, F., Paraz, A., Seigneurin, J. M., Mandrand, B., Perron, H., and Garson, J. A. (2000). Infrequency of detection of

particle-associated MSRV/HERV-W RNA in the synovial fluid of patients with rheumatoid arthritis. Rheumatology (Oxford) *39*, 950-954.

Gelfand, D. H., and White, T. J. (1990). Thermostable DNA polymerase, In PCR protocols, a guide to methods and applications, M. A. G. Innis, D. H. Sninsky, J.J. White, T.J., ed. (San Diago: Academic Press), pp. 129 - 141.

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

Goncalves, I., Duret, L., and Mouchiroud, D. (2000). Nature and structure of human genes that generate retropseudogenes. Genome Res *10*, 672-678.

Gonzalez, M., Bagatolli, L. A., Echabe, I., Arrondo, J. L., Argarana, C. E., Cantor, C. R., and Fidelio, G. D. (1997). Interaction of biotin with streptavidin. Thermostability and conformational changes upon binding. J Biol Chem *272*, 11288-11294.

Goodier, J. L., Ostertag, E. M., Engleka, K. A., Seleme, M. C., and Kazazian, H. H., Jr. (2004). A potential role for the nucleolus in L1 retrotransposition. Hum Mol Genet *13*, 1041-1048.

Goriely, A., McVean, G. A., van Pelt, A. M., O'Rourke, A. W., Wall, S. A., de Rooij, D. G., and Wilkie, A. O. (2005). Gain-of-function amino acid substitutions drive positive selection of FGFR2 mutations in human spermatogonia. Proc Natl Acad Sci U S A *102*, 6051-6056.

Greally, J. M. (2002). Short interspersed transposable elements (SINEs) are excluded from imprinted regions in the human genome. Proc Natl Acad Sci U S A *99*, 327-332.

Han, J. S., and Boeke, J. D. (2004). A highly active synthetic mammalian retrotransposon. Nature *429*, 314-318.

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 Res *33*, 4040-4052.

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

Hickey, D. A. (1982). Selfish DNA: a sexually-transmitted nuclear parasite. Genetics *101*, 519-531.

Hiraizumi, Y. (1971). Spontaneous recombination in Drosophila melanogaster males. Proc Natl Acad Sci U S A *68*, 268-270.

Hirsch, H. J., Saedler, H., and Starlinger, P. (1972). Insertion mutations in the control region of the galactose operon of E. coli. II. Physical characterization of the mutations. Mol Gen Genet *115*, 266-276.

Hohjoh, H., and Singer, M. F. (1996). Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. Embo J *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 J *16*, 6034-6043.

Holloway, K., Lawson, V. E., and Jeffreys, A. J. (2006). Allelic recombination and de novo deletions in sperm in the human beta-globin gene region. Hum Mol Genet *15*, 1099-1111.

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. Nat Genet *7*, 143-148.

Holmes, S. E., Singer, M. F., and Swergold, G. D. (1992). studies on p40, the Leucine zipper motif-containing protein encoded by the first open reading frame of an active human LINE-1 transposable element. The Journal Of Biological Chemistry *267*, 19765-19768.

Houck, C. M., Rinehart, F. P., and Schmid, C. W. (1979). A ubiquitous family of repeated DNA sequences in the human genome. J Mol Biol *132*, 289-306.

Iyer, G. S., Krahe, R., Goodwin, L. A., Doggett, N. A., Siciliano, M. J., Funanage, V. L., and Proujansky, R. (1996). Identification of a testis-expressed creatine transporter gene at 16p11.2 and confirmation of the X-linked locus to Xq28. Genomics *34*, 143-146.

Jeffreys, A. J., and May, C. A. (2003). DNA enrichment by allele-specific hybridization (DEASH): a novel method for haplotyping and for detecting low-frequency base substitutional variants and recombinant DNA molecules. Genome Res *13*, 2316-2324.

Jeffreys, A. J., Murray, J., and Neumann, R. (1998). High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot. Mol Cell *2*, 267-273.

Jeffreys, A. J., Wilson, V., Neumann, R., and Keyte, J. (1988). Amplification of human minisatellites by the polymerase chain reaction: towards DNA fingerprinting of single cells. Nucleic Acids Res *16*, 10953-10971.

Jeffreys, A. J., Wilson, V., and Thein, S. L. (1985). Hypervariable 'minisatellite' regions in human DNA. Nature *314*, 67-73.

Jordan, E., Saedler, H., and Starlinger, P. (1968). O0 and strong-polar mutations in the gal operon are insertions. Mol Gen Genet *102*, 353-363.

Karlsson, H., Bachmann, S., Schroder, J., McArthur, J., Torrey, E. F., and Yolken, R. H. (2001). Retroviral RNA identified in the cerebrospinal fluids and brains of individuals with schizophrenia. Proc Natl Acad Sci U S A *98*, 4634-4639.

Kazazian, H. H., Jr. (1999). An estimated frequency of endogenous insertional mutations in humans. Nat Genet *22*, 130.

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

Kazazian, H. H., Jr., and Moran, J. V. (1998). The impact of L1 retrotransposons on the human genome. Nat Genet *19*, 19-24.

Kazazian, H. H., Jr., Wong, C., Youssoufian, H., Scott, A. F., Phillips, D. G., and 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.

Kedes, L. H. (1979). Histone genes and histone messengers. Annu Rev Biochem *48*, 837-870.

Kerjean, A., Dupont, J. M., Vasseur, C., Le Tessier, D., Cuisset, L., Paldi, A., Jouannet, P., and Jeanpierre, M. (2000). Establishment of the paternal methylation imprint of the human H19 and MEST/PEG1 genes during spermatogenesis. Hum Mol Genet *9*, 2183-2187.

Kidwell, M. G., Kidwell, J. F., and Nei, M. (1973). A case of high rate of spontaneous mutation affecting viability in Drosophila melanogaster. Genetics *75*, 133-153.

Kidwell, M. G., Kidwell, J. F., and Sved, J. A. (1977). Hybrid Dysgenesis in DROSOPHILA MELANOGASTER: A Syndrome of Aberrant Traits Including Mutation, Sterility and Male Recombination. Genetics *86*, 813-833.

Kidwell, M. G., and Novy, J. B. (1979). Hybrid Dysgenesis in DROSOPHILA MELANOGASTER: Sterility Resulting from Gonadal Dysgenesis in the P-M System. Genetics *92*, 1127-1140.

Kierszenbaum, A. L. (2002). Genomic imprinting and epigenetic reprogramming: unearthing the garden of forking paths. Mol Reprod Dev *63*, 269-272.

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. Hum Mol Genet *8*, 1557-1560.

Kinsey, J. A. (1993). Transnuclear retrotransposition of the Tad element of Neurospora. Proc Natl Acad Sci U S A *90*, 9384-9387.

Komurian-Pradel, F., Paranhos-Baccala, G., Bedin, F., Ounanian-Paraz, A., Sodoyer, M., Ott, C., Rajoharison, A., Garcia, E., Mallet, F., Mandrand, B., and Perron, H. (1999). Molecular cloning and characterization of MSRV-related sequences associated with retrovirus-like particles. Virology *260*, 1-9.

Kondo-Iida, E., Kobayashi, K., Watanabe, M., Sasaki, J., Kumagai, T., Koide, H., Saito, K., Osawa, M., Nakamura, Y., and Toda, T. (1999). Novel mutations and genotype-phenotype relationships in 107 families with Fukuyama-type congenital muscular dystrophy (FCMD). Hum Mol Genet *8*, 2303-2309.

Kozak, M. (1989). The scanning model for translation: an update. J Cell Biol *108*, 229-241.

Kubo, S., Seleme Mdel, C., Soifer, H. S., Perez, J. L., Moran, J. V., Kazazian, H. H., Jr., and Kasahara, N. (2006). L1 retrotransposition in nondividing and primary human somatic cells. Proc Natl Acad Sci U S A *103*, 8036-8041.

Kulpa, D. A., and Moran, J. V. (2006). Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. Nat Struct Mol Biol *13*, 655-660.

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 Res *23*, 3704-3709.

Lam, K. W., and Jeffreys, A. J. (2006). Processes of copy-number change in human DNA: the dynamics of α-globin gene deletion. Proc Natl Acad Sci U S A *103*, 8921-8927.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome.[comment][erratum appears in Nature 2001 Aug 2;412(6846):565 Note: Szustakowki J [corrected to Szustakowski J]]. Nature *409*, 860-921.

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. Proc Natl Acad Sci U S A *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. Proc Natl Acad Sci U S A *87*, 6990-6994.

Lewis, J. D., Saperas, N., Song, Y., Zamora, M. J., Chiva, M., and Ausio, J. (2004). Histone H1 and the origin of protamines. Proc Natl Acad Sci U S A *101*, 4148-4152.

Lewis, J. D., Song, Y., de Jong, M. E., Bagha, S. M., and Ausio, J. (2003). A walk though vertebrate and invertebrate protamines. Chromosoma *111*, 473-482.

Li, E. (2002). Chromatin modification and epigenetic reprogramming in mammalian development. Nat Rev Genet *3*, 662-673.

Li, P. W., Li, J., Timmerman, S. L., Krushel, L. A., and Martin, S. L. (2006). The dicistronic RNA from the mouse LINE-1 retrotransposon contains an internal ribosome entry site upstream of each ORF: implications for retrotransposition. Nucleic Acids Res *34*, 853-864.

Li, X., Scaringe, W. A., Hill, K. A., Roberts, S., Mengos, A., Careri, D., Pinto, M. T., Kasper, C. K., and Sommer, S. S. (2001). Frequency of recent retrotransposition events in the human factor IX gene. Hum Mutat *17*, 511-519.

Long, E. O., and Dawid, I. B. (1980). Repeated genes in eukaryotes. Annu Rev Biochem *49*, 727-764.

Lopez, J., and Prezioso, V. (2001). A better way to optimize: Two-step Gradient PCR. Eppendorf BioNews Application Notes *September*, 3 - 4.

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.

Lucifero, D., Mann, M. R., Bartolomei, M. S., and Trasler, J. M. (2004). Gene-specific timing and epigenetic memory in oocyte imprinting. Hum Mol Genet *13*, 839-849.

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. Cytogenet Cell Genet *80*, 133-137.

Lyon, M. F. (2003). The Lyon and the LINE hypothesis. Semin Cell Dev Biol *14*, 313-318.

Macfarlane, C., and Simmonds, P. (2004). Allelic variation of HERV-K(HML-2) endogenous retroviral elements in human populations. J Mol Evol *59*, 642-656.

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

Malik, H. S., and Eickbush, T. H. (2000). NeSL-1, an ancient lineage of site-specific non-LTR retrotransposons from Caenorhabditis elegans. Genetics *154*, 193-203.

Malik, H. S., Henikoff, S., and Eickbush, T. H. (2000). Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res *10*, 1307-1318.

Mann, J. R. (2001). Imprinting in the germ line. Stem Cells *19*, 287-294.

Martin, S. L., Branciforte, D., Keller, D., and Bain, D. L. (2003). Trimeric structure for an essential protein in L1 retrotransposition. Proc Natl Acad Sci U S A *100*, 13815-13820.

Martin, S. L., Cruceanu, M., Branciforte, D., Wai-Lun Li, P., Kwok, S. C., Hodges, R. S., and Williams, M. C. (2005). LINE-1 retrotransposition requires the nucleic acid chaperone activity of the ORF1 protein. J Mol Biol *348*, 549-561.

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.

McClintock, B. (1950). The origin and behavior of mutable loci in maize. Proc Natl Acad Sci U S A *36*, 344-355.

McClintock, B. (1951). Chromosome organization and genic expression. Cold Spring Harb Symp Quant Biol *16*, 13-47.

McClintock, B. (1956). Controlling elements and the gene. Cold Spring Harb Symp Quant Biol *21*, 197-216.

McMillan, J. P., and Singer, M. F. (1993). Translation of the human LINE-1 element, L1Hs. Proc Natl Acad Sci U S A *90*, 11533-11537.

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. Eur J Hum Genet *8*, 697-703.

Mi, S., Lee, X., Li, X., Veldman, G. M., Finnerty, H., Racie, L., LaVallie, E., Tang, X. Y., Edouard, P., Howes, S., *et al.* (2000). Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature *403*, 785-789.

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 Res *52*, 643-645.

Mine, M., Chen, J. M., Brivet, M., Desguerre, I., Marchant, D., de Lonlay, P., Bernard, A., Ferec, 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. Hum Mutat *28*, 137-142.

Moran, J. V. (1999). Human L1 retrotransposition: insights and peculiarities learned from a cultured cell retrotransposition assay. Genetica *107*, 39-51.

Moran, J. V., and Gilbert, N. (2002). Mammalian LINE-1 Retrotransposons and Related Elements, In Mobile DNA II, N. L. Craig, R. Craigie, M. Gellert, and A. M. Lambowitz, eds. (Washington, D.C.: ASM Press), pp. 836 - 869.

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.

Morrish, T. A., Gilbert, N., Myers, J. S., Vincent, B. J., Stamato, T. D., Taccioli, G. E., Batzer, M. A., and Moran, J. V. (2002). DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nat Genet *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.

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., *et al.* (2002). A

comprehensive analysis of recently integrated human Ta L1 elements. Am J Hum Genet *71*, 312-326.

Narita, N., Nishio, H., Kitoh, Y., Ishikawa, Y., Minami, R., Nakamura, H., and Matsuo, M. (1993). Insertion of a 5' truncated L1 element into the 3' end of exon 44 of the dystrophin gene resulted in skipping of the exon during splicing in a case of Duchenne muscular dystrophy. J Clin Invest *91*, 1862-1867.

Neel, J. V., Satoh, C., Goriki, K., Fujita, M., Takahashi, N., Asakawa, J., and Hazama, R. (1986). The rate with which spontaneous mutation alters the electrophoretic mobility of polypeptides. Proc Natl Acad Sci U S A *83*, 389-393.

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

Nikaido, M., Rooney, A. P., and Okada, N. (1999). Phylogenetic relationships among cetartiodactyls based on insertions of short and long interpersed elements: hippopotamuses are the closest extant relatives of whales. Proc Natl Acad Sci U S A *96*, 10261-10266.

Ohshima, K., Hattori, M., Yada, T., Gojobori, T., Sakaki, Y., and Okada, N. (2003). Whole-genome screening indicates a possible burst of formation of processed pseudogenes and *Alu* repeats by particular L1 subfamilies in ancestral primates. Genome Biol *4*, R74.

Ostertag, E. M., Goodier, J. L., Zhang, Y., and Kazazian, H. H., Jr. (2003). SVA elements are nonautonomous retrotransposons that cause disease in humans. Am J Hum Genet *73*, 1444-1451.

Ostertag, E. M., and Kazazian, H. H., Jr. (2001a). Biology of mammalian L1 retrotransposons. Annu Rev Genet *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 Res *11*, 2059-2065.

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 Res *28*, 1418-1423.

Ovchinnikov, I., Rubin, A., and Swergold, G. D. (2002). Tracing the LINEs of human evolution. Proc Natl Acad Sci U S A *99*, 10522-10527.

Ovchinnikov, I., Troxel, A. B., and Swergold, G. D. (2001). Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion. Genome Research *11*, 2050-2058.

Pavlicek, A., Gentles, A. J., Paces, J., Paces, V., and Jurka, J. (2006). Retroposition of processed pseudogenes: the impact of RNA stability and translational control. Trends Genet *22*, 69-73.

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

Pavlicek, A., Paces, J., Zika, R., and Hejnar, J. (2002b). Length distribution of long interspersed nucleotide elements (LINEs) and processed pseudogenes of human endogenous retroviruses: implications for retrotransposition and pseudogene detection. Gene *300*, 189-194.

Pelisson, A., and Bregliano, J.-C. (1987). Evidence for rapidlimitation of the I element copy number in a genome submitted to several generations of I-R hybrid dysgenesis in *Drosophila melanogaster*. Molecular and general genetics *207*, 306 - 313.

Perriere, G., Duret, L., and Gouy, M. (2000). HOBACGEN: database system for comparative genomics in bacteria. Genome Res *10*, 379-385.

Petrov, D. A., and Hartl, D. L. (1997). Trash DNA is what gets thrown away: high rate of DNA loss in Drosophila. Gene *205*, 279-289.

Petrov, D. A., Lozovskaya, E. R., and Hartl, D. L. (1996). High intrinsic rate of DNA loss in Drosophila. Nature *384*, 346-349.

Picard, G. (1976). Non-mendelian female sterility in Drosophila melanogaster: hereditary transmission of I factor. Genetics *83*, 107-123.

Picard, G., and L'Heritier, P. (1978). A maternally inherited factor inducing sterility in *Drosophila melanogaster*. Drosophila Information Service *46*, 54.

Picard, G., and Pelisson, A. (1979). Non-Mendelian Female Sterility in DROSOPHILA MELANOGASTER: Characterization of the Noninducer Chromosomes of Inducer Strains. Genetics *91*, 473-489.

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.

Popova, B. C., Tada, T., Takagi, N., Brockdorff, N., and Nesterova, T. B. (2006). Attenuated spread of X-inactivation in an X;autosome translocation. Proc Natl Acad Sci U S A *103*, 7706-7711.

Rastan, S. (1983). Non-random X-chromosome inactivation in mouse X-autosome translocation embryos--location of the inactivation centre. J Embryol Exp Morphol *78*, 1-22.

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., *et al.* (2005). The DNA sequence of the human X chromosome. Nature *434*, 325-337.

Roy-Engel, A. M., Salem, A. H., Oyeniran, O. O., Deininger, L., Hedges, D. J., Kilroy, G. E., Batzer, M. A., and Deininger, P. L. (2002). Active *Alu* element "A-tails": size does matter. Genome Res *12*, 1333-1344.

Rubin, G. M., Finnegan, D. J., and Hogness, D. S. (1976). The chromosomal arrangement of coding sequences in a family of repeated genes. Prog Nucleic Acid Res Mol Biol *19*, 221-226.

Saedler, H., Besemer, J., Kemper, B., Rosenwirth, B., and Starlinger, P. (1972). Insertion mutations in the control region of the Gal operon of E. coli. I. Biological characterization of the mutations. Mol Gen Genet *115*, 258-265.

Sambrook, J., and Russell, D. W. (2001). Molecular cloning, a laboratory manual, Vol 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. Nat Rev Genet *3*, 65-72.

Santos, F. R., Pandya, A., Kayser, M., Mitchell, R. J., Liu, A., Singh, L., Destro-Bisol, G., Novelletto, A., Qamar, R., Mehdi, S. Q., *et al.* (2000). A polymorphic L1 retroposon insertion in the centromere of the human Y chromosome. Hum Mol Genet *9*, 421-430.

Sarrowa, J., Chang, D. Y., and Maraia, R. J. (1997). The decline in human *Alu* retroposition was accompanied by an asymmetric decrease in SRP9/14 binding to dimeric *Alu* RNA and increased expression of small cytoplasmic *Alu* RNA. Mol Cell Biol *17*, 1144-1151.

Sassaman, D. M., Dombroski, B. A., Moran, J. V., Kimberland, M. L., Naas, T. P., DeBerardinis, R. J., Gabriel, A., Swergold, G. D., and Kazazian, H. H., Jr. (1997). Many human L1 elements are capable of retrotransposition. Nat Genet *16*, 37-43.

Sawyer, S. L., Emerman, M., and Malik, H. S. (2004). Ancient adaptive evolution of the primate antiviral DNA-editing enzyme APOBEC3G. PLoS Biol *2*, E275.

Schmid, C. W. (2003). *Alu*: a parasite's parasite? Nat Genet *35*, 15-16.

Schmid, C. W., and Deininger, P. L. (1975). Sequence organization of the human genome. Cell *6*, 345-358.

Schulz, W. A., Steinhoff, C., and Florl, A. R. (2006). Methylation of endogenous human retroelements in health and disease. Curr Top Microbiol Immunol *310*, 211-250.

Schwahn, U., Lenzner, S., Dong, J., Feil, S., Hinzmann, B., van Duijnhoven, G., Kirschner, R., Hemberger, M., Bergen, A. A., Rosenberg, T., *et al.* (1998). Positional cloning of the gene for X-linked retinitis pigmentosa 2. Nat Genet *19*, 327-332.

Seela, F., Budow, S., Shaikh, K. I., and Jawalekar, A. M. (2005). Stabilization of tandem dG-dA base pairs in DNA-hairpins: replacement of the canonical bases by 7-deaza-7-propynylpurines. Org Biomol Chem *3*, 4221-4226.

Seleme Mdel, 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. Proc Natl Acad Sci U S A *103*, 6611-6616.

Shapiro, J. A. (1969). Mutations caused by the insertion of genetic material into the galactose operon of Escherichia coli. J Mol Biol *40*, 93-105.

Shaw, C. J., and Lupski, J. R. (2004). Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. Hum Mol Genet *13 Spec No 1*, R57-64.

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 Res *10*, 1496-1508.

Shimamura, M., Yasue, H., Ohshima, K., Abe, H., Kato, H., Kishiro, T., Goto, M., Munechika, I., and Okada, N. (1997). Molecular evidence from retroposons that whales form a clade within even-toed ungulates. Nature *388*, 666-670.

Skowronski, J., Fanning, T. G., and Singer, M. F. (1988). Unit-length line-1 transcripts in human teratocarcinoma cells. Mol Cell Biol *8*, 1385-1397.

Smit, A. F. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev *9*, 657-663.

Smit, A. F. A., Hubley, R., and Green, P. (unpublished data). RepeatMasker Open-3.0. 1996-2004.

Smith, D. B., McAllister, J., Casino, C., and Simmonds, P. (1997). Virus 'quasispecies': making a mountain out of a molehill? J Gen Virol *78 (Pt 7)*, 1511-1519.

Southern, E. M. (1974). An improved method for transferring nucleotides from electrophoresis strips to thin layers of ion-exchange cellulose. Anal Biochem *62*, 317-318.

Speek, M. (2001). Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. Mol Cell Biol *21*, 1973-1985.

Swergold, G. D. (1990). Identification, characterization, and cell specificity of a human LINE-1 promoter. Mol Cell Biol *10*, 6718-6729.

Symer, D. E., Connelly, C., Szak, S. T., Caputo, E. M., Cost, G. J., Parmigiani, G., and Boeke, J. D. (2002). Human l1 retrotransposition is associated with genetic instability in vivo. Cell *110*, 327-338.

Szak, S. T., Pickeral, O. K., Landsman, D., and Boeke, J. D. (2003). Identifying related L1 retrotransposons by analyzing 3' transduced sequences. Genome Biol *4*, R30.

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 Biol *3*, research0052.

Tamaki, K., and Jeffreys, A. J. (2005). Human tandem repeat sequences in forensic DNA typing. Leg Med (Tokyo) *7*, 244-250.

Tang, W., Gunn, T. M., McLaughlin, D. F., Barsh, G. S., Schlossman, S. F., and Duke-Cohan, J. S. (2000). Secreted and membrane attractin result from alternative splicing of the human ATRN gene. Proc Natl Acad Sci U S A *97*, 6025-6030.

Tchenio, T., Casella, J. F., and Heidmann, T. (2000). Members of the SRY family regulate the human LINE retrotransposons. Nucleic Acids Res *28*, 411-415.

Temin, H., and Baltimore, D. (1972). RNA-directed DNA synthesis and RNA tumour viruses. Advances In Virus Research *17*, 129-186.

Temin, H. M., and Mizutani, S. (1970). RNA-dependent DNA polymerase in virions of Rous sarcoma virus. Nature *226*, 1211-1213.

Ueda, T., Abe, K., Miura, A., Yuzuriha, M., Zubair, M., Noguchi, M., Niwa, K., Kawase, Y., Kono, T., Matsuda, Y., *et al.* (2000). The paternal methylation imprint of the mouse H19 locus is acquired in the gonocyte stage during foetal testis development. Genes Cells *5*, 649-659.

van den Hurk, J. A., Meij, I. C., del Carmen Seleme, M., Kano, H., Nikopoulos, K., Hoefsloot, L. H., Sistermans, E. A., de Wijs, I. J., Mukhopadhyay, A., Plomp, A. S., *et al.* (2007). L1 retrotransposition can occur early in human embryonic development. Hum Mol Genet *16*, 1587-1592.

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., *et al.* (2003). Novel types of mutation in the choroideremia (CHM) gene: a full-length L1 insertion and an intronic mutation activating a cryptic exon. Hum Genet *113*, 268-275.

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.

Volff, J. N., Korting, C., and Schartl, M. (2000). Multiple lineages of the non-LTR retrotransposon Rex1 with varying success in invading fish genomes. Mol Biol Evol *17*, 1673-1684.

Walsh, C. P., Chaillet, J. R., and Bestor, T. H. (1998). Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. Nat Genet *20*, 116-117.

Watson, J. D., and Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature *171*, 737-738.

Wei, W., Gilbert, N., Ooi, S. L., Lawler, J. F., Ostertag, E. M., Kazazian, H. H., Boeke, J. D., and Moran, J. V. (2001). Human L1 retrotransposition: cis preference versus trans complementation. Mol Cell Biol *21*, 1429-1439.

Wei, W., Morrish, T. A., Alisch, R. S., and Moran, J. V. (2000). A transient assay reveals that cultured human cells can accommodate multiple LINE-1 retrotransposition events. Anal Biochem *284*, 435-438.

Weichenrieder, O., Repanas, K., and Perrakis, A. (2004). Crystal structure of the targeting endonuclease of the human LINE-1 retrotransposon. Structure *12*, 975-986.

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 Res *15*, 1073-1078.

Williams, A. J., Khachigian, L. M., Shows, T., and Collins, T. (1995). Isolation and characterization of a novel zinc-finger protein with transcription repressor activity. J Biol Chem *270*, 22143-22152.

Wouters-Tyrou, D., Martinage, A., Chevaillier, P., and Sautiere, P. (1998). Nuclear basic proteins in spermiogenesis. Biochimie *80*, 117-128.

Xiong, Y., and Eickbush, T. H. (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. Embo J *9*, 3353-3362.

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. Proc Natl Acad Sci U S A *96*, 7847-7852.

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. Hum Mol Genet *7*, 1129-1132.

Zhang, J., and Webb, D. M. (2004). Rapid evolution of primate antiviral enzyme APOBEC3G. Hum Mol Genet *13*, 1785-1791.

Zhang, L., Lu, H. H., Chung, W. Y., Yang, J., and Li, W. H. (2005). Patterns of segmental duplication in the human genome. Mol Biol Evol *22*, 135-141.

Zhang, Z., Harrison, P. M., Liu, Y., and Gerstein, M. (2003). Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome Res *13*, 2541-2558.