Developing informatic resources for LINE-1 retrotransposons
Thesis submitted for the degree of Doctor of Philosophy at the University of Leicester

Robert K. Hastings
Department of Genetics
University of Leicester
June 5, 2009
# Contents

## Abstract

## Acknowledgements

## Abbreviations

## Literature Review

1. **Genome complexity** ................................................. 6

1.1 Simple Sequence Repeats (SSRs) ................................. 10

1.2 Repetitive sequences in the human genome ................. 9

- Segmental Duplications ............................................. 10
- Pseudogenes .......................................................... 11
- Transposon-Derived Repeats ....................................... 11
  - DNA transposons (Class II transposable elements) ......... 13
  - Retrotransposons (Class I Transposable elements) ....... 13
  - Long Terminal Repeat (LTR) Retrotransposons .......... 13
  - Non-Long Terminal Repeat (non-LTR) Retrotransposons 14
  - SINEs .............................................................. 14
  - Long Interspersed Nuclear Elements (LINEs) ............ 17

1.3 L1 families ........................................................ 18

- L1 structure ......................................................... 18
- PreTa sub-family .................................................. 19
# CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ta L1 family</td>
<td>21</td>
</tr>
<tr>
<td>Ta-L1 polymorphism</td>
<td>21</td>
</tr>
<tr>
<td>L1 activity</td>
<td>22</td>
</tr>
<tr>
<td>1.4 Retrotransposition</td>
<td>23</td>
</tr>
<tr>
<td>Transcription</td>
<td>23</td>
</tr>
<tr>
<td>Role of the sense promoter in transcription</td>
<td>24</td>
</tr>
<tr>
<td>YY1 binding sites</td>
<td>25</td>
</tr>
<tr>
<td>RUNX3 binding sites</td>
<td>26</td>
</tr>
<tr>
<td>SOX/SRY binding sites</td>
<td>26</td>
</tr>
<tr>
<td>Antisense promoter</td>
<td>27</td>
</tr>
<tr>
<td>Interaction between L1 and the splicosome</td>
<td>27</td>
</tr>
<tr>
<td>Translation of L1</td>
<td>28</td>
</tr>
<tr>
<td>Open Reading Frame 1 protein (ORF1p)</td>
<td>29</td>
</tr>
<tr>
<td>Open Reading Frame 2 protein (ORF2p)</td>
<td>30</td>
</tr>
<tr>
<td>Endonuclease domain</td>
<td>30</td>
</tr>
<tr>
<td>Reverse transcriptase domain</td>
<td>31</td>
</tr>
<tr>
<td>Target primed reverse transcription (TPRT)</td>
<td>32</td>
</tr>
<tr>
<td>How L1 EN dependent retrotransposition events shape the human genome</td>
<td>34</td>
</tr>
<tr>
<td>Three prime transduction</td>
<td>34</td>
</tr>
<tr>
<td>Five prime transduction</td>
<td>37</td>
</tr>
<tr>
<td>Inversion</td>
<td>37</td>
</tr>
<tr>
<td>Exon shuffling by retrotransposition</td>
<td>39</td>
</tr>
<tr>
<td>L1 trans-driven retrotransposition</td>
<td>39</td>
</tr>
<tr>
<td>Endonuclease-independent L1 insertion</td>
<td>40</td>
</tr>
<tr>
<td>1.5 L1 polymorphism classification and detection</td>
<td>41</td>
</tr>
<tr>
<td>ATLAS (amplification typing of L1 active subfamilies)</td>
<td>42</td>
</tr>
<tr>
<td>L1 polymorphism in human genome assemblies</td>
<td>42</td>
</tr>
<tr>
<td>L1 polymorphism in trace archive sequences</td>
<td>43</td>
</tr>
<tr>
<td>1.6 Genomic distribution of L1 in the human genome</td>
<td>44</td>
</tr>
<tr>
<td>Insertion bias of L1 and Alu elements in the human genome</td>
<td>45</td>
</tr>
</tbody>
</table>
## CONTENTS

Ta-1 L1 distribution ................................................. 46  
Identity by descent ............................................... 47  

1.7 Effects of L1 sequences in the human genome ............. 48  
   L1 insertions are deleterious ................................ 48  
   The effects of L1 insertions on gene expression .......... 49  
   The effect of the antisense promoter (ASP) in the human genome .... 49  
   Ectopic recombination ...................................... 50  
   L1 disease-causing insertions ............................... 51  
   The role of L1 sequences in X-inactivation ............... 53  

1.8 Sequencing, Assemblies and Databases ...................... 54  
   Genome sequencing and assembly ........................... 55  
   Existing sources of FL-L1 sequences ...................... 58  
   Trace archives ............................................... 59  
   Short-read archives (SRA) .................................. 61  
   Nucleotide accession databases ............................ 61  
   dbSNP ..................................................... 62  
   L1Base - Annotation of full-length, intact L1 elements .... 62  
   dbRIP .................................................... 63  
   Line fusion genes .......................................... 64  

1.9 Software tools for L1 analysis ................................ 65  
   TSDfinder .................................................. 65  
   L1Xplorer .................................................... 66  

2 Materials and methods ........................................... 68  

2.1 Introduction .................................................. 68  
2.2 Hardware ...................................................... 68  
2.3 Software ....................................................... 69  
   Practical Extraction and Report Language (PERL) ........ 69  
   Comprehensive Perl Archive Network (CPAN) ............ 70  
   Bioperl ..................................................... 70  
   Common Gateway Interface (CGI) .......................... 71
CGI-FormBuilder .................................................. 72
MySQL .............................................................. 72
Perl DataBase Interface (DBI) ................................. 72
Template Toolkit (TT2) ........................................... 73
DAS ................................................................. 74
ProServer ......................................................... 75
BLAST-like Alignment Tool (BLAT) ......................... 76
Searching BLAT sequence databases ....................... 76
BLAST ............................................................. 77
FASTA2 sequence suite ......................................... 78
RepeatMasker ....................................................... 78
Repbase libraries .................................................. 79
cross_match ....................................................... 79
BLAST database construction ................................. 79
Trace archive construction ...................................... 80
BLAST database searches ........................................ 81
GBrowse Karyotype viewer ...................................... 81
  Data for GC content XY plot .................................. 81
  Data for gene content XY plot ................................ 82
Perl code .......................................................... 82
  Baseline::Util .................................................. 83
  Baseline::Database ........................................... 83
  Baseline::Parse::Blast ........................................ 84
  Baseline::Run::SeqSearch .................................... 84
  Baseline::Run::SeqExtract ................................... 85
  Baseline::Run::TSDmapper .................................... 85
  Baseline::Extract::Sequence ................................ 85
Baseline perl scripts .......................................... 85
  blast.pl ....................................................... 86
  blast_parser.pl .............................................. 86
  TSDmapper.pl ................................................ 86
CONTENTS

3 Full-length L1 sequence discovery

3.1 Introduction

3.2 Searching of resources containing FL-L1HS sequences

3.3 Number of FL-L1HS sequences in different human genome assemblies

3.4 Distribution of FL-L1HS

3.5 FL-L1HS sequences from other sources

3.6 Resolution of duplicated FL-L1HS sequences from nt database

3.7 Resolution of duplicated hits between sources
### CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.8 FL-L1HS sequence contribution to the human genome</td>
<td>109</td>
</tr>
<tr>
<td>3.9 In-silico screening for polymorphic FL-L1HS sequences in the nt database</td>
<td>111</td>
</tr>
<tr>
<td>3.10 Trace Archives</td>
<td>113</td>
</tr>
<tr>
<td>3.11 FL-L1HS sequence annotation and analysis</td>
<td>118</td>
</tr>
<tr>
<td>Sub-family distribution of FL-L1HS sequences</td>
<td>120</td>
</tr>
<tr>
<td>Computational prediction of active FL-L1HS sequences</td>
<td>121</td>
</tr>
<tr>
<td>Sub-family classification: Intact ORFs vs. non-intact ORFs</td>
<td>122</td>
</tr>
<tr>
<td>FL-L1HS sequences with intact ORFs</td>
<td>123</td>
</tr>
<tr>
<td>FL-L1HS sequences with non-intact ORFs</td>
<td>124</td>
</tr>
<tr>
<td>Comparison of predicted FL-L1HS sequences with experimental methods</td>
<td>125</td>
</tr>
<tr>
<td>FL-L1HS sequences in genes</td>
<td>126</td>
</tr>
<tr>
<td>Do FL-L1HS show an insertion orientation bias when inserting into genes?</td>
<td>128</td>
</tr>
<tr>
<td>Into what type of gene do FL-L1HS sequences insert?</td>
<td>131</td>
</tr>
<tr>
<td>Do FL-L1HS sequences insert into genes of certain sizes?</td>
<td>131</td>
</tr>
<tr>
<td>Gene Ontology</td>
<td>132</td>
</tr>
<tr>
<td>3.12 Discussion</td>
<td>136</td>
</tr>
<tr>
<td>4 TSD detection and analysis</td>
<td>147</td>
</tr>
<tr>
<td>4.1 Introduction</td>
<td>147</td>
</tr>
<tr>
<td>4.2 Design and implementation of TSDmapper</td>
<td>148</td>
</tr>
<tr>
<td>4.3 TSDmapper sequence validation</td>
<td>149</td>
</tr>
<tr>
<td>4.4 PolyA tail discovery</td>
<td>150</td>
</tr>
<tr>
<td>4.5 Finding potential 5' and 3' TSDs</td>
<td>152</td>
</tr>
<tr>
<td>4.6 Development of a TSD scoring matrix from endonuclease cut sites</td>
<td>154</td>
</tr>
<tr>
<td>4.7 Scoring TSDs in TSDmapper</td>
<td>158</td>
</tr>
<tr>
<td>4.8 Transductions</td>
<td>162</td>
</tr>
<tr>
<td>4.9 Empty Sites</td>
<td>163</td>
</tr>
</tbody>
</table>
CONTENTS

4.10 Comparison of TSDmapper and TSDfinder performance relative to manually called TSDs ............................... 164
4.11 TSD analysis of FL-L1HS using TSDmapper ............... 165
    PolyA tail length ............................................. 166
    Characteristics of TSDs indentified by TSDmapper ........ 169
    Endonuclease recognition site consensus .................... 170
    Transductions ................................................ 173
    Empty site validation ....................................... 181
    TSDmapper - other applications ............................ 182
4.12 Discussion .................................................. 183

5 baseLINE - A FL-L1HS database ................................. 191
5.1 Introduction ................................................... 191
5.2 Database Design .............................................. 192
    Database tables ............................................... 193
        baseLINE L1s ............................................. 195
        Duplicates ............................................. 195
        TSDmapper ............................................. 196
        L1Xplorer ............................................. 196
        Genes ................................................ 197
        Sequences ............................................ 197
        Blast ................................................ 197
5.3 Website ....................................................... 198
    Website navigation .......................................... 198
    Home Page ................................................ 198
    Search ..................................................... 200
        Accession search .................................... 200
        Chromosome search ................................ 202
        Gene search ........................................ 203
        TSD searching ...................................... 204
    View ..................................................... 205
<table>
<thead>
<tr>
<th>CONTENTS</th>
<th>ix</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1 record</td>
<td>207</td>
</tr>
<tr>
<td>TSDmapper</td>
<td>219</td>
</tr>
<tr>
<td>TSDmapper output</td>
<td>221</td>
</tr>
<tr>
<td>5.4 Discussion</td>
<td>227</td>
</tr>
<tr>
<td>6 Final summary and future directions</td>
<td>233</td>
</tr>
<tr>
<td>Appendices</td>
<td>235</td>
</tr>
<tr>
<td>A Literature review : Appendix</td>
<td>236</td>
</tr>
<tr>
<td>A.1 Alu-disease causing insertions</td>
<td>236</td>
</tr>
<tr>
<td>A.2 L1 disease insertions</td>
<td>236</td>
</tr>
<tr>
<td>B Full-length L1 sequence discovery: Appendix</td>
<td>239</td>
</tr>
<tr>
<td>B.1 Number of FL-L1HS in different human genome assemblies</td>
<td>239</td>
</tr>
<tr>
<td>B.2 Trace Archives</td>
<td>239</td>
</tr>
<tr>
<td>B.3 FL-L1HS in genes</td>
<td>239</td>
</tr>
<tr>
<td>B.4 Over-represented Gene Ontology terms</td>
<td>239</td>
</tr>
<tr>
<td>C TSD detection and analysis : Appendix</td>
<td>246</td>
</tr>
<tr>
<td>C.1 Development of a TSD scoring matrix from endonuclease cut sites</td>
<td>246</td>
</tr>
<tr>
<td>References</td>
<td>249</td>
</tr>
</tbody>
</table>
Abstract

The human genome contains vast numbers of sequences that have copied themselves to new genomic locations by retrotransposition. Long Interspersed Nuclear Element-1 (LINE-1 or L1) is the only sequence in the human genome still capable of autonomous retrotransposition. L1 elements have contributed to the evolution of the human genome via insertional mutagenesis, pseudogene formation, sequence transduction, and recombination events (producing insertions, deletions and inversions). Currently general and L1-specific sequence databases do not reflect the true level of Full Length Human Specific L1 (FL-L1HS) variation, due to the polymorphic nature of these elements and the way the databases were compiled.

Methods to identify FL-L1HS were applied to three sequence assemblies (Reference, Celera and HuRef) and the nucleotide accession database from NCBI. A non-redundant set of 533 FL-L1HS was discovered in these four sources, of which 164 resided in genes. The trace archives from Ensembl were also searched and a further 48 potential FL-L1HS were found. Computational analyses showed 154 FL-L1HS were potentially capable of retrotransposition, including 54 that resided in genes. Alongside these analyses a Target Site Duplication (TSD) detection and analysis tool, TSDmapper was developed to automatically detect TSDs in FL-L1HS sequences and provide annotation on sequence transduction. TSDmapper was used to predict the pre-insertion sequence of all 533 unique L1s, which facilitated in-silico genotyping.

A new informatic resource, baseLINE (http://baseline.gene.le.ac.uk), was created to display and enable searching of all the L1 annotation information generated. Data can be viewed in a genomic context in chromosome ideograms or can be exported via the Distributed Annotation Service (DAS) on to the Ensembl genome browser. TSDmapper is also provided as a web application at baseLINE for users to perform TSD annotation of their sequences of interest.
Acknowledgements

Firstly I would like to thank Richard for all his advice and guidance throughout the project, its been a fun few years (hope you agree too!). Also I would like to thank Tony for keeping me in food and beer tokens when the PhD money ran out and allowing me to become a HGVbase monkey with all the other guys.

Special thanks to Owen, Rob Free, Mummi and Adam who provided great technical advice and support on all types of bioinformatics related matters, and for putting up with me and my grumpy moods whilst writing!

Thanks to members of Richard’s group, past and present for there support, especially Pete, Pam, and Catriona. Also thanks to all members of Tony’s group who made me feel welcome and gave me great support, especially Reshma and Colin.

Thanks to the Old Horse crew, Tall Rob (I will never forget the mixing bowl!), George (fancy a green tea?), Aline, Christine, Emma, Katherine, Jon, Mahmut, Aaron and many others I have probably forgotten, for many a karaoke fun filled night! Also for the beers when things were not going so well. Also thanks to Jacqui for the support and the gluten-free cakes!

Special thanks to my folks, Phil and Jane who have always supported me in what ever I have done. Also to my two little sisters (who are not so little now), Karen (‘what do you do again Robert?’) and Rachel, along with Kasey, Freddie and Terry.
Abbreviations

Biological abbreviations

A list of biological abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full term</th>
</tr>
</thead>
<tbody>
<tr>
<td>aa</td>
<td>amino acid</td>
</tr>
<tr>
<td>ASP</td>
<td>Anti-Sense Promoter</td>
</tr>
<tr>
<td>BAC</td>
<td>Bacterial Artificial Chromosome</td>
</tr>
<tr>
<td>bp</td>
<td>base pairs</td>
</tr>
<tr>
<td>C-domain</td>
<td>Cysteine rich C terminal domain</td>
</tr>
<tr>
<td>C-value</td>
<td>$C_{ot}$ value</td>
</tr>
<tr>
<td>cDNA</td>
<td>Complementary DNA</td>
</tr>
<tr>
<td>CGD</td>
<td>Chronic Granulomatous Disease</td>
</tr>
<tr>
<td>DMD</td>
<td>Duchenne Muscular Dystrophy</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribo-Nucleic Acid</td>
</tr>
<tr>
<td>EN</td>
<td>Endonuclease</td>
</tr>
<tr>
<td>FL-L1HS</td>
<td>Full Length Human Specific LINE-1 element $(\geq 5922 \text{ bp and } \geq 98% \text{ identity})$</td>
</tr>
<tr>
<td>kb</td>
<td>Kilobase pairs</td>
</tr>
<tr>
<td>LINE</td>
<td>Long Interspersed Nuclear Element</td>
</tr>
<tr>
<td>LTR</td>
<td>Long Terminal Repeat</td>
</tr>
<tr>
<td>M</td>
<td>Million</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger RNA</td>
</tr>
<tr>
<td>MYA</td>
<td>Million Years Ago</td>
</tr>
<tr>
<td>MY</td>
<td>Million Years</td>
</tr>
<tr>
<td>ORF</td>
<td>Open Reading Frame</td>
</tr>
<tr>
<td>ORF1p</td>
<td>Open Reading Frame encoded protein 1</td>
</tr>
<tr>
<td>ORF2p</td>
<td>Open Reading Frame encoded protein 2</td>
</tr>
</tbody>
</table>
### Biological terms

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full term</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCR</td>
<td>Polymerase Chain Reaction</td>
</tr>
<tr>
<td>Pol</td>
<td>Polymerase</td>
</tr>
<tr>
<td>RC</td>
<td>Retrotransposition Competent</td>
</tr>
<tr>
<td>RNA</td>
<td>Ribo-Nucleic Acid</td>
</tr>
<tr>
<td>RNP</td>
<td>Ribonucleoprotein</td>
</tr>
<tr>
<td>rRNA</td>
<td>Ribosomal RNA</td>
</tr>
<tr>
<td>RT</td>
<td>Reverse Transcriptase</td>
</tr>
<tr>
<td>RUNX3</td>
<td>Runt domain transcription factor</td>
</tr>
<tr>
<td>SINE</td>
<td>Short Interspersed Nuclear Element</td>
</tr>
<tr>
<td>SNP</td>
<td>Single Nucleotide Polymorphism</td>
</tr>
<tr>
<td>SRY</td>
<td>Sex-determining Region Y</td>
</tr>
<tr>
<td>snRNA</td>
<td>small nuclear RNA</td>
</tr>
<tr>
<td>SVA</td>
<td>SINE-R, VNTR and Alu</td>
</tr>
<tr>
<td>TE</td>
<td>Transposable Element</td>
</tr>
<tr>
<td>TPRT</td>
<td>Target-site Primed Reverse Transcription</td>
</tr>
<tr>
<td>tRNA</td>
<td>Transfer RNA</td>
</tr>
<tr>
<td>TSD</td>
<td>Target Site Duplications</td>
</tr>
<tr>
<td>UTR</td>
<td>Un-Translated Region</td>
</tr>
<tr>
<td>VNTR</td>
<td>Variable Number Tandem Repeat</td>
</tr>
<tr>
<td>YY1</td>
<td>Ying Yang 1</td>
</tr>
<tr>
<td>(+/+)</td>
<td>Homozgous present - Only L1 filled site is reported (assumes diploidy)</td>
</tr>
<tr>
<td>(+/-)</td>
<td>Heterozygous - L1 filled site and empty site reported</td>
</tr>
<tr>
<td>(-/-)</td>
<td>Homozgous absent - Only L1 empty site reported</td>
</tr>
</tbody>
</table>
Software abbreviations

A list of software abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full term</th>
</tr>
</thead>
<tbody>
<tr>
<td>API</td>
<td>Aplication Programming Interface</td>
</tr>
<tr>
<td>BLAST</td>
<td>Basic Local Alignment Search Tool</td>
</tr>
<tr>
<td>blastn</td>
<td>Basic Local Alignment Search Tool for nucleotide searches</td>
</tr>
<tr>
<td>BLAT</td>
<td>BLAST Like Alignment Tool</td>
</tr>
<tr>
<td>CPAN</td>
<td>Comprehensive Perl Archive Network</td>
</tr>
<tr>
<td>DDBJ</td>
<td>DNA Data Bank of Japan</td>
</tr>
<tr>
<td>dbRIP</td>
<td>Database of Retrotransposon Insertion Polymorphisms in Humans</td>
</tr>
<tr>
<td>dbSNP</td>
<td>The Single Nucleotide Polymorphism database</td>
</tr>
<tr>
<td>EBI</td>
<td>European Bioinformatics Institute</td>
</tr>
<tr>
<td>GFF3</td>
<td>Gene Finding Format</td>
</tr>
<tr>
<td>GMOD</td>
<td>Generic Model Organism Database</td>
</tr>
<tr>
<td>HGVbase</td>
<td>Human Genome Variation Database</td>
</tr>
<tr>
<td>hg18</td>
<td>University of California Santa Cruz human genome build 18</td>
</tr>
<tr>
<td>MGI</td>
<td>Mouse Genome Informatics</td>
</tr>
<tr>
<td>NCBI</td>
<td>The National Center for Biotechnology Information nucleotide accession database</td>
</tr>
<tr>
<td>Perl</td>
<td>Practical extraction and report languague</td>
</tr>
<tr>
<td>SQL</td>
<td>Structure Query Language</td>
</tr>
<tr>
<td>UCSC</td>
<td>University California Santa Cruz</td>
</tr>
<tr>
<td>XML</td>
<td>Extensible Markup Language</td>
</tr>
</tbody>
</table>
Chapter 1

Literature Review

1.1 Genome complexity

Currently available, completed genome sequences (as of April 2008) comprise 762 organism genomes (53 Archaeal, 626 Bacterial, and 83 Eukaryotic), with a further 2773 genomes currently being sequenced (89 Archaeal, 1749 Bacterial, and 935 Eukaryotic, taken from http://www.genomesonline.org/). This vast amount of genetic data, generated by the combined efforts of many sequencing projects continues to provide opportunities to explore variability both within and between species. At the same time it has become increasingly important to discover the different types of genetic variations that exist between and within individuals, as these variations may cause or contribute to the risk of disease. Tools that can automatically detect genetic sequence variation between and within species will provide valuable information on population genetics, population history as well as insights into the evolution of genomes.

Even before the large-scale sequencing of genomes, biologists noticed a puzzling observation that genome size does not correlate well with organismal complexity. This observation was termed the C-value paradox (Swift, 1950b). A number of experiments led to the idea that a species could be classified by the amount of DNA in its nuclei,
which turned out to be relatively constant (Pagel and Johnstone, 1992; Swift, 1950a,b) between individuals of the same species.

The amount of DNA in a nuclear genome of an organism’s gametes is called the DNA C-value (Constant-value), which refers to the haploid DNA amount (Swift, 1950b). In diploid organisms, which include most animals and a small number of plants, the genome size and the DNA C-value are identical. However if the organism is polyploid then the C-value is not simply related to the haploid DNA content as the C-value will comprise more than one genome (Gregory, 2005a).

Large variation of the DNA C-value exists across species even within the same taxon. For example, in Eukaryotes this can encompass 4 orders of magnitude: assuming that 1 Pico-gram (pg) of DNA equates to approximately $10^9$ bases). The genome sizes of the largest known Eukaryote genome (Amoeba dubia C-value=700 pg) and one of the smallest (Saccharomyces cerevisiae C-value=0.009 pg) vary by 80,000 fold (Gregory, 2005b). Clearly at the extremes of the distribution there is not a direct relationship between genome size and complexity: the genome size of Homo sapiens (C-value=3.5 pg), a multi-cellular metazoan is 200 times smaller than the unicellular protozoan, Amoeba dubia.

A list of species and their DNA C-value genome sizes are summarized in a table of genome sizes (Table 1.1). This table shows that genome size is clearly not related to organism complexity, which is known as the “C-value paradox”. Even amongst multi-cellular animals there is great variation, as the smallest known animal genome is that of Pratylenchus coffeae (a parasitic nematode) (0.02 pg) (Gregory, 2005b) which is approximately 6,600 fold smaller than the largest known animal genome of the Protopterus aethiopicus (Marbled lungfish, C-value 132.83 pg). In mammals, genome size variation is less pronounced with the largest genome size belonging to Tympanoctomys barrerae (Red viscacha rat, C-value 8.4 pg) and the smallest being that of Miniopterus schreibersi (Bent-winged bat, C-value 1.73 pg) a difference of approximately 5 fold. The average mammalian genome size is 3.49pg ±0.04 , suggesting humans (3.5pg) are typical mammals in this respect (Gregory, 2005a).
Table 1.1: Genome sizes of Eukaryotes. All values are given as 1C in pg. Key: *= Mean C value calculated based on 4 values for *Pan troglodytes*

<table>
<thead>
<tr>
<th>Species</th>
<th>Common name</th>
<th>C-value (pg)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Homo sapiens</strong></td>
<td>Human</td>
<td>3.5</td>
<td>Gregory (2005b)</td>
</tr>
<tr>
<td><strong>Pan troglodytes</strong></td>
<td>Chimpanzee</td>
<td>3.68*</td>
<td>Gregory (2005b)</td>
</tr>
<tr>
<td><strong>Mus musculus</strong></td>
<td>House mouse</td>
<td>3.29</td>
<td>Gregory (2005b)</td>
</tr>
<tr>
<td><strong>Tympanoctomys barrerae</strong></td>
<td>Red viscacha rat</td>
<td>8.40</td>
<td>Gregory (2005b)</td>
</tr>
<tr>
<td><strong>Miniopterus schreibersi</strong></td>
<td>Bent-winged bat</td>
<td>1.73</td>
<td>Gregory (2005b)</td>
</tr>
<tr>
<td><strong>Prototoperus aethiopicus</strong></td>
<td>Marbled lungfish</td>
<td>132.83</td>
<td>Gregory (2005b)</td>
</tr>
<tr>
<td><strong>Drosophila melanogaster</strong></td>
<td>Fruit fly</td>
<td>0.18</td>
<td>Gregory (2005b)</td>
</tr>
<tr>
<td><strong>Pratylenchus caffeae</strong></td>
<td>Plant-parasitic nematode</td>
<td>0.02</td>
<td>Gregory (2005b)</td>
</tr>
<tr>
<td><strong>Caenorhabditis elegans</strong></td>
<td>Nematode</td>
<td>0.09</td>
<td>Gregory (2005b)</td>
</tr>
<tr>
<td><strong>Fritillaria assyriaca</strong></td>
<td>Angiosperm - Lily</td>
<td>127.40</td>
<td>Bennett and Leitch (2004)</td>
</tr>
<tr>
<td><strong>Arabidopsis thaliana</strong></td>
<td>Thale cress</td>
<td>0.16</td>
<td>Bennett and Leitch (2004)</td>
</tr>
<tr>
<td><strong>Ostreococcus tauri</strong></td>
<td>Algae</td>
<td>0.01</td>
<td>Bennett and Leitch (2004)</td>
</tr>
<tr>
<td><strong>Saccharomyces cerevisiae</strong></td>
<td>Yeast</td>
<td>0.009</td>
<td>Kullman <em>et al.</em> (2005)</td>
</tr>
<tr>
<td><strong>Encephalitozoon cuniculi</strong></td>
<td>Microsporidium</td>
<td>0.0029</td>
<td>Kullman <em>et al.</em> (2005)</td>
</tr>
<tr>
<td><strong>Amoeba dubia</strong></td>
<td>Amoeba</td>
<td>700</td>
<td>Bennett and Leitch (2004)</td>
</tr>
</tbody>
</table>

In some species where polyploid takes place, this only contributes to a small amount of genome size variation and so differences in genome size must be due to other causes. For example, the fact that wheat (*Triticum aestivum*) is hexaploid (genome size 16,000 Mb) accounts for only 8% of its genome relative to rice (*Oryza sativa* genome size 430 Mb). This difference can be explained by the presence, in wheat, of large amounts of repetitive sequences that are absent from the rice genome (*Hartl*, 2000).

The C-value paradox has been partially resolved by the discovery that genomes can contain large amounts of non-coding DNA, largely repetitive sequences, rather than large differences in unique and coding DNA. A good example of this is the human genome, which contains a relatively low number of protein-coding genes (20,000-25,000 *IHGSC (2004)*), only twice as many as the worm (*Caenorhabditis elegans*) and the fly (*Drosophila melanogaster*) (*Lander *et al.*, 2001), which equates to less than 5% of the human genome being occupied by coding sequence (*Lander *et al.*, 2001). This is in stark contrast to the estimated 70% of the yeast genome (*Saccharomyces cerevisiae*...
C-value 0.09 pg) that is coding DNA. Thus variation in the amount of non-coding DNA sequences would seem to explain the C-value paradox (Pagel and Johnstone, 1992).

The accumulation of non-coding repetitive sequences as an explanation of the C-value paradox is consistent with ‘Junk DNA’ and selfish DNA theories (Orgel and Crick, 1980; Orgel et al., 1980). In these theories an organism’s DNA can tolerate the accumulation of repetitive sequences until the cost of replicating it becomes deleterious (Pagel and Johnstone, 1992). Organisms could also use the accumulation of repetitive DNA as a mechanism for slowing the rate of development by increasing cell-cycle length (Pagel and Johnstone, 1992). By contrast, if an organism is selected to attain a faster rate of development, then having efficient mechanisms for deleting the non-coding DNA could be an advantage (Pagel and Johnstone, 1992).

Having considered how genome size in a number of organisms can be influenced by repetitive DNA, which partly resolves the issue of the C-value paradox, the focus of this chapter will be repetitive sequences in the Human genome, which constitute around half of our genome (Lander et al., 2001).

### 1.2 Repetitive sequences in the human genome

The initial analysis of the draft human genome sequence, produced in 2001 by the International Human Genome Sequencing Consortium (Lander et al., 2001) revealed many interesting insights into the sequence composition of the human genome. Repetitive sequences, mainly transposable elements, comprise almost 50% of the genome. This is in contrast to the predicted 5% coding DNA (Lander et al., 2001). These repetitive sequences fall into four main classes. 1) simple sequence repeats 2) segmental duplications 3) processed and non-processed pseudo-genes, and 4) transposon-derived repeats (interspersed repeats).
Simple Sequence Repeats (SSRs)

Simple Sequence Repeats (SSRs) are perfect or slightly imperfect direct tandem repeats of certain $k$-mers. SSRs with repeat units of 1-13 nt are often referred to as micro-satellites. SSRs with longer repeat units of 14-500 bases are often called mini-satellites (Lander et al., 2001). SSRs make a small contribution to genome composition of approximately 3%, with the greatest single contribution coming from dinucleotide repeats (0.5%) (Lander et al., 2001). AC and AT di-nucleotides (50 and 35% of all di-nucleotide repeats), are the most frequent while AG and GC di-nucleotides (15 and 0.1% of all di-nucleotide repeats) being the least common (Lander et al., 2001).

Tri-nucleotide repeats are also present in the human genome but are found less frequently than di-nucleotides (Lander et al., 2001). SSRs can be highly variable between individuals, displaying length polymorphisms that can be used in human disease mapping studies. In the United Kingdom 10 micro-satellite loci are currently used to generate genetic profiles for use in forensic analysis enabling genetic individual identification, more commonly known as “DNA fingerprinting” (Tamaki and Jeffreys, 2005).

Segmental Duplications

Segmental duplications are regions of genomic sequence that have been duplicated from one genomic location to another. In the human genome these can vary between 1-200 kb (Lander et al., 2001). To be classed as segmental duplications regions must be $\geq 1$ kb in length and show at least 90% sequence identity to each other (Zhang et al., 2005). Many segmental duplications appear to have arisen recently in evolutionary time as they are absent from closely related species and also because they share very high sequence identity to each other (Lander et al., 2001).

Zhang et al. (2005) reported that around 4% of the genome is contained within segmental duplications and observed that the level of segmental duplication varies from 1% to 14% among the 24 chromosomes. Segmental duplications can be divided into two categories, intra-chromosomal and inter-chromosomal. Intra-chromosomal
duplications occur within a chromosome or chromosome arm and tend to be larger and more frequent in 15 of the human chromosomes than inter-chromosomal duplications (Zhang et al., 2005). Inter-chromosomal duplications are segments that are duplicated among non-homologous chromosomes (Lander et al., 2001).

Segmental duplications tend to be enriched at peri-centromeric and sub-telomeric regions with duplication frequencies in these regions 3-4 fold above the genome average. The proportion of segmental duplications containing complete genes is reported to be 3.4% (for known genes) and 10.4% for Ensembl genes (Zhang et al., 2005), suggesting segmental duplications account for a significant level of gene copy number change.

**Pseudogenes**

Pseudogenes fall into two main types, processed (retrotransposed) and non-processed. Processed pseudogenes are inactive copies of genes that arise when an mRNA copy of a gene is reverse transcribed into a cDNA copy and integrated into a new genomic location. Pseudogenes are usually inactive copies of genes that have been generated during segmental duplication formation (Lander et al., 2001).

**Transposon-Derived Repeats**

Lander et al. (2001) estimates that approximately 45% of the human genome is derived of transposon derived repeats, making these repeats the most important sequence class by mass in the genome. Human transposon-derived repeats can be further divided into four types: DNA transposons which transpose directly via DNA, and three others, Short Interspersed Nuclear Elements (SINEs), Long Terminal Repeat (LTR) retrotransposons and Long Interspersed Nuclear Elements (LINEs) which use an RNA intermediate to transpose. These types of repeats are often referred to as mobile DNA elements as some types have the ability to replicate and integrate into new genomic locations, either autonomously or non-autonomously. The types of transposable elements in the human genome are illustrated in Figure 1.1.
Figure 1.1: Classes of transposable elements in the human genome. Shown are the classes of repetitive elements with their approximate contribution to genome size and copy number (CN) broken down into the sub classes. Adapted from Lander et al. (2001)
CHAPTER 1. LITERATURE REVIEW

DNA transposons (Class II transposable elements)

DNA transposons are characterized by short inverted terminal repeats and encode a transposase activity that allows these transposons to move by a so called ‘cut and paste’ mechanism. The transposase binds to DNA near the terminal inverted repeats and cuts the DNA at sequence-specific sites (e.g. for the Tc1/mariner class, TA dinucleotides) removing the DNA transposon from its current site and pasting it into a new location that is usually in close proximity to the original insertion. This process is known as ‘local hopping’ (Kazazian, 2004). DNA transposons comprise approximately 3% of the human genome draft sequence and although their mechanism of transposition is known from other species they are unlikely to be still capable of transposition in the human genome and thus can be regarded as evolutionary fossils (Lander et al., 2001). The youngest DNA transposons in the human genome are estimated to have transposed 30-40 million years ago (MYA) (Pace and Feschotte, 2007).

Retrotransposons (Class I Transposable elements)

Retrotransposons can be classified as autonomous or non-autonomous. Elements are autonomous if they encode the proteins required for their mobilization. Autonomous retrotransposons can be further divided into Long Terminal Repeat (LTR) Retrotransposons and non-Long Terminal Repeat (non-LTR) Retrotransposons. These elements are discussed in the following sections.

Long Terminal Repeat (LTR) Retrotransposons

Retrovirus-like elements (RLEs) are retrotransposons characterized by two Long Terminal Repeats (LTRs) which contain regulatory sequences required for transcription of an internal region containing viral-like gag and pol genes. LTR elements can either be autonomous (they encode the proteins required for mobilization) or non-autonomous. Autonomous LTR retrotransposons may contain three ORFs coding for gag, pol and env proteins. The gag protein has a structural function, forming the virus-like particle (VLP) in which reverse transcription takes place. The pol protein has
various enzymatic properties that include a protease that cleaves the pol poly-protein, a reverse transcriptase that copies RNA of the retrotransposon into cDNA and an integrase that integrates the cDNA into the new genomic location. LTR elements in the human genome comprise up 8% of our DNA (Lander et al., 2001). Autonomous LTR elements vary in length from ~6-11kb, whilst non-autonomous elements are smaller, ranging from 1.5-3 kb (Lander et al., 2001).

Non-Long Terminal Repeat (non-LTR) Retrotransposons

Non-LTR elements (illustrated in Figure 1.1) do not have flanking LTRs. This class of retrotransposon contains elements that are autonomous and non-autonomous. Autonomous retrotransposons encode the proteins required for their mobilization, whilst non-autonomous retrotransposons are able to parasitise the retrotransposition machinery of the autonomous elements to maintain their mobility (Lander et al., 2001).

SINEs

SINE elements contribute ~13% of the human genome sequence. SINEs vary in length from 100-300 bp and have a copy number of ~1.5 Million (M) in the human genome (Lander et al., 2001). Alu SINE elements are the most numerous of the SINEs with a copy number of ~1 M and contribute ~10% of the human genome sequence. SINE elements harbour an internal polymerase III promoter that is located within a tRNA-derived region present in all SINEs except Alu elements which are derived from a 7SL RNA (Dewannieux et al., 2003). SINEs do not encode proteins required for their mobilization and have been shown to use the Long Interspersed Nuclear Elements’ (LINEs) transcriptional machinery as a method of transposition (Dewannieux et al., 2003).

A full-length Alu element consists of two monomers, the left monomer containing the RNA polymerase III promoter which directs transcription from the first nucleotide of the element and the right monomer which is separated from the left by an A-rich linker sequence and lacks the promoter, but has an additional 31 bp of sequence (Ullu and
Weiner, 1985). Alu elements are \( \sim 300 \) bp in length, depending on the length of the terminal polyA tail (Batzler and Deininger, 2002), and are flanked by variable length Target Site Duplications (TSDs) (Batzler and Deininger, 2002).

The process of retrotransposition in Alu involves an Alu derived RNA polymerase III transcript being reverse transcribed. As Alu sequences encode no proteins for transposition they were long thought to borrow the LINE-1 retrotransposition machinery (Mathias et al., 1991). Intact LINE-1 elements encode a reverse transcriptase and an endonuclease activity (see section 1.4) that provide the necessary enzymatic functions for their transposition. Alu uses target site-primed reverse transcription as the mechanism of insertion using the ORF2 protein endonuclease function to nick genomic DNA at an endonuclease consensus site TT↓AAAA (where the down arrow indicates the point of cleavage). Dewannieux et al. (2003) demonstrated that Alu RNA is retrotransposed by the L1 proteins, and that retrotransposition of Alu can occur even when the open reading frame 1 (ORF1) protein of the LINE element is not functional. Therefore the minimum requirement for an Alu element to undergo retrotransposition is a functional open reading frame 2 (ORF2) LINE protein. This means that L1 sequences with a disrupted ORF1 and a functional ORF2 can promote de novo Alu insertions in the human genome.

Previous to the demonstration that L1 proteins retrotranspose Alu elements, an insertion pattern analysis of human Alu (and Rodent ID) elements was carried out to determine if they use the same endonuclease-cutting site as L1 elements (Jurka, 1997). This analysis used 344 human Alu sequences with \( \geq 10 \) bp or more of upstream flanking TSDs and 10 bp or more of downstream flanking TSDs. Only those Alu sequences that had TSDs immediately flanking the 5’ end of the sequence were selected for, and the base compositions at the different positions of the TSDs were recorded for all 344 Alu elements.

This analysis showed that the nick in the genomic DNA was non-random and showed a consensus site of 5’-TT↓AAAA-3’ suggesting the L1 endonuclease is the best candidate for the generation of Alu insertions (Jurka, 1997). Figure 1.2 a) illustrates that positions -1 and -2 (TT) and positions +1 to +4 (AAAA) are the most significant. At the 3’ end the analysis also demonstrated that the 3’ ends flanking the TSDs also show non-random
Figure 1.2: a) $\chi^2$ values for individual positions of Alu and ID TSDs with flanking DNA indicating that at the 5' end positions -2 to +4 are significant at the 0.001 levels. b) $\chi^2$ values for individual positions of Alu and ID TSDs with flanking DNA indicating that at the 3' end there is not a strong consensus. Taken from Jurka (1997).

base occurrences at positions -4, -3 and -2 (AA|TTTT) (Jurka, 1997). This is illustrated in Figure 1.2 (b).

Alu elements are commonly found in un-translated regions in genes, introns and intergenic regions of the genome (Batzer and Deininger, 2002) and are dispersed throughout the human genome, although they are most commonly found in GC rich genomic regions (Bailey et al., 2003). Alu elements are the only known active member of the SINE class and increase in copy number by retrotransposition. Alu elements also have long PolyA tails (>50 bp) the removal of which abolishes Alu retrotransposition.
De novo Alu insertions have caused 20 cases of human genetic diseases, including breast cancer, Huntington’s disease and Haemophilia A and B. See Appendix section A.1 for a full list of Alu disease causing insertions. The majority of Alu disease-causing insertions come from the the Alu Y subfamily. It has been estimated that Alu insertions potentially account for 0.3% of human genetic disease (Deininger and Batzer, 1999). However only a small fraction of Alu insertions have been reported to cause disease with the vast majority having apparently little or no impact on the human genome behaving like neutral alleles and subject to genetic drift (i.e. A change in allele frequency from generation to generation) (Batzer and Deininger, 2002). Alu retrotransposition is estimated to create 1 new Alu insertion in the human populations every 200 births (Deininger and Batzer, 1999).

Almost all of the recently integrated Alu elements in the human genome (∼5,000 or 0.5% of the total number of Alu elements in the human genome) belong to closely related, young sub-families (Y, Yc1, Yc2, Ya5, Ya5a2, Ya8 and Yb9) derived from three Y-subfamily lineages (Ya, Yb, Yc). Many of these sequences have been shown to be polymorphic with respect to their presence or absence at particular loci in the human genome, i.e. they show dimorphism. Around ∼25% of young Alu repeats are dimorphic (Batzer and Deininger, 2002). These large non-revertible genetic markers may provide insights into human population genetics, as individuals who share an insertion are most likely to have inherited it from a common ancestor. As an Alu insertion is an independent event it is unlikely to have occurred at exactly the position in the genome in different lineages, allowing Alu elements to be used in phylogenetic studies as it is expected these insertions should show low levels of homoplasy. Thus SINEs (and LINEs) can used as useful genetic markers that show identity by descent (Salem et al., 2005).

**Long Interspersed Nuclear Elements (LINEs)**

In the human genome Long Interspersed Nuclear Elements (LINEs) are the only autonomous non-LTR elements, with LINE-1 being the only currently active known member. LINEs as a family comprise ∼21% of the human genome sequence (Lander et al., 2001) and have three main sub-families. LINE-1 is by far the most numerous
of the families with a copy number of \( \sim 500,000 \) equating to \( \sim 17\% \) to the genome. LINE-2 contributes only \( \sim 3\% \) of the genome sequence and LINE-3 makes a smaller contribution (<1%). LINEs also vary in length (6-8 kb) with the known active families of LINE-1 elements being \( \sim 6 \) kb long.

### 1.3 L1 families

LINE-1 elements can be considered to be the most important repetitive element in the human genome, comprising 17% of the DNA by mass (Lander et al., 2001). There are sub-families of LINE-1 elements that are still able to undergo retrotransposition, but the vast majority have been inactivated by mutation or rearrangement. LINE-2 and LINE-3 elements only contribute a small amount to the genome size (<4% combined) and are no longer capable of retrotransposition. For this reason LINE-2 and LINE-3 elements will not be discussed any further. Elements belonging to LINE-1 sub families, particularly the youngest members known as L1PA1 or L1HS are discussed in detail below.

### L1 structure

The structure of a typical retrotransposition competent (RC) human (Homo sapiens) specific full-length L1 (L1HS) is illustrated in Figure 1.3. L1HS elements are approximately 6 kb in length, with a 910 bp 5' un-translated region (UTR). Within the first 100-150 bp of the 5' UTR is an internal sense RNA polymerase II (RNA Pol II) promoter (Kurose et al., 1995). Also contained within the 5' UTR is an anti-sense promoter (ASP) with a predicted Transcription Start Site (TSS) between nucleotides +400 and +600 bp (Speek, 2001). Following the 5' UTR there are two non-overlapping open reading frames (ORFs) separated by a 63 bp intergenic spacer containing two stop codons (Dombroski et al., 1991), and a 205 bp 3' UTR containing a polypurine-rich tract and ending in a poly-A tail. L1 elements that are inserted into the genome contain an unconventional poly-A tail and are typically flanked in the genome by short variable
length direct repeats known as target site duplications (TSDs) (Dombroski et al., 1991; Szak et al., 2002).

PreTa sub-family

The preTa subfamily of L1 elements is characterized by a three base-pair ACG sequence in the 3’ UTR which separates them from other L1 sub-families. The preTa subfamily can be further divided into two subgroups, ACG/A and ACG/G as illustrated in Table 1.2 based on there being an A or G nucleotide at position 6040 relative to accession M80343 (L1.2) (Salem et al., 2003).

The average age of these two subgroups has been estimated to be 1.92 Myrs for ACG/G and 3.24 Myrs for ACG/A and it is thought that the L1 Ta subfamily originated from one of these preTa subsets (Salem et al., 2003).

Integration sites for preTa elements have been studied by Salem et al. (2003) and they observed a pattern where some sites are preferred for integration over others. In particular TTTT/A and TCTT/A were preferred conforming to the endonuclease
recognition site suggested by Jurka (1997). Analysis of these elements suggests that they insert preferentially into genomic sequences where the GC content is less than 36%, however they have been reported to have inserted into sites with a high GC content (56%) as well as low GC content (26%) (Salem et al., 2003).

Table 1.2: Subclass defining nucleotide changes at specific positions in L1PA1 and L1PA2 families. Adapted from Brouha et al. (2003)

<table>
<thead>
<tr>
<th>Subclass defining nucleotides of L1 families</th>
<th>L1 Family</th>
<th>74</th>
<th>711</th>
<th>1820</th>
<th>5557</th>
<th>5560</th>
<th>5954</th>
<th>5955</th>
<th>5956</th>
<th>6040</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ta-1d</td>
<td></td>
<td>-</td>
<td>t</td>
<td>t</td>
<td>t</td>
<td>g</td>
<td>a</td>
<td>c</td>
<td>a</td>
<td>g</td>
</tr>
<tr>
<td>Ta-1nd</td>
<td></td>
<td>g</td>
<td>c</td>
<td>c</td>
<td>t</td>
<td>g/c</td>
<td>a</td>
<td>c</td>
<td>a</td>
<td>g</td>
</tr>
<tr>
<td>Ta-0</td>
<td></td>
<td>g</td>
<td>c</td>
<td>c</td>
<td>g</td>
<td>c</td>
<td>a</td>
<td>c</td>
<td>a</td>
<td>g</td>
</tr>
<tr>
<td>Pre-Ta (ACG/G)</td>
<td></td>
<td>g</td>
<td>c</td>
<td>c</td>
<td>g</td>
<td>c</td>
<td>a</td>
<td>c</td>
<td>g</td>
<td>g</td>
</tr>
<tr>
<td>Pre-Ta (ACG/A)</td>
<td></td>
<td>g</td>
<td>c</td>
<td>c</td>
<td>g</td>
<td>c</td>
<td>a</td>
<td>c</td>
<td>g</td>
<td>a</td>
</tr>
<tr>
<td>L1PA2</td>
<td></td>
<td>g</td>
<td>c</td>
<td>c</td>
<td>g</td>
<td>c</td>
<td>g</td>
<td>a</td>
<td>g</td>
<td>a</td>
</tr>
</tbody>
</table>

The sequence diversity of preTa elements is largely generated by 5’ truncation resulting in some elements being only a few hundred base pairs long while others remain full length (6 kb) (Salem et al., 2003). Sequence variation is also observed in the 3’ tails of the L1 elements and 3’ transduction events have been attributed to the sequence variations at this end of the elements (Salem et al., 2003).

Although preTa family members are considered to be slightly older than Ta-1 L1 family members there are still some L1 elements within this family that have intact open reading frames. Salem et al. (2003) identified 105 preTa elements that are full length in the human genome and 29 of these are reported to have intact open reading frames and therefore considered by a computational means to be putatively active or retrotransposition competent (RC). The L1s that do not have intact open reading frames may have in the past acted as the source for the growth of the preTa elements but now the majority have become inactive due to an accumulation of mutations over a period of time (Salem et al., 2003).
Ta L1 family

The L1PA1 family is also known as Ta L1 or Transcribed subset A (Skowronski et al., 1988) and is the only L1 family that is human specific (L1Hs) and currently active (i.e. capable of retrotransposition) in the human genome. It has been estimated that the Ta family emerged approximately 5 Mya after the divergence between humans and chimpanzees (∼6 Mya) and has diverged into two distinct subfamilies called Ta-0 and Ta-1 (Boissinot et al., 2000).

The Ta family can be defined by an ACA 3 nt character in the 3' UTR (5954-5956 see Table 1.2) and can be divided into the Ta-0 and Ta-1 subgroups by the nucleotides at positions 5557 and 5560 in ORF2. Ta-1 subfamily elements have a T and G at these two positions whereas Ta-0 elements have a G and C respectively, like the ancestral L1 element families, LPA2 and LPA3 (both non-Ta families (Boissinot et al., 2000)). Ta-0 is therefore apparently the older and more divergent of the Ta subfamilies and because of its age contains fewer active elements than the Ta-1 subfamily, which has replaced Ta-0 as the dominant Ta subfamily in human populations (Boissinot et al., 2000).

The Ta-1 subfamily is estimated to have arisen ∼2.5 Mya, with the majority of Ta-1 elements being generated in the last ∼1.6 Mya. It has been estimated that over 80% of Ta-0 elements inserted before 1.6 Mya (Boissinot et al., 2000). Ta-1 can be further divided into subsets; these are Ta-1d (deletion), which is the younger of the two and contains a deletion at nucleotide position 74 in the 5' UTR and Ta-1nd (no deletion) that does not contain this deletion.

Ta-L1 polymorphism

Over 50% of the Ta family show presence or absence polymorphism across human populations (Boissinot and Furano, 2001). Presence or absence polymorphism is where one individual will have a copy of a Ta L1 element at a certain location within their genome whilst other individuals will display the pre-insertion site. The Ta L1 subfamily is the youngest and most active of the L1 families and has been associated with 16 of the 17 cases of L1-induced human disease.
Consistent with this high level of insertional polymorphism, the Ta-1 subfamily has more members that can actively retrotranspose than the Ta-0 or PreTa subfamilies. This is because the Ta-1 family is younger than the Ta-0 and pre-Ta families and thus has had less time to accumulate inactivating mutations.

**L1 activity**

To examine the level of L1 activity in the human genome a study was carried out by Brouha *et al.* (2003). This involved determining the activity of full-length L1 elements in the December 2001 freeze of the human genome working draft sequence (HGWD). It had been determined that there were 89 L1s with intact ORFs (and one L1 with intact ORFs but with a 400-bp deletion in the 5' UTR). Eighty two of these L1 elements were cloned and assayed for retrotransposition activity using a cell-culture-based retrotransposition assay (Moran *et al.*, 1996). Forty L1s were determined to be active, with these elements being further divided into 6 highly active or ‘hot’ L1s, and 34 weakly active. The remaining 42 elements were inactive (Brouha *et al.*, 2003). Out of the 40 active elements, the 6 “hot” L1 elements accounted for approximately 84% of the total measured activity in the HGWD sequence. These highly active L1s are frequently polymorphic and it has been estimated that there are approximately 80 to 100 retrotransposition competent L1s in the average human genome (Brouha *et al.*, 2003). The majority of highly active L1s are present at low frequencies in the human genome and may be the progenitors of L1s that cause disease-producing insertions (Deininger *et al.*, 2003).

Three of the highly active L1s identified in the Brouha *et al.* (2003) study belong to the youngest Ta family, Ta-1d, and these are known as AC002980, AC004200 and AL356438 (these are the accession numbers of the genomic sequences within which the L1s reside). Another one of the L1s, AL512428, is classified as being a member of the Ta-1nd Ta subfamily, while AL137845 has been classified as being a member of the Ta-0 family. AC021017 is also highly active and polymorphic but it is described as non-canonical. AL356438 which is a hot L1, and AC093886, which has been classified
as inactive, were not found in any of the 46 genomes used in allele-frequency studies by Brouha et al. (2003).

Retrotransposition frequency in humans has been estimated to occur frequently, with 1 in 2 or 1 in 33 individuals carrying a *de-novo* insertion (Deininger et al., 2003). Another estimate of retrotransposition frequency is that 1 in 10 individuals carry a novel insertion (Kazazian, 1999). These estimates are based on disease-causing mutations and so may be an underestimate and may not reflect the true retrotransposition frequency in the human genome. The process of L1 retrotransposition is discussed in more detail in section 1.4.

### 1.4 Retrotransposition

The retrotransposition of a L1Hs element requires a number of different steps. These are transcription, RNA processing, mRNA export, translation, post-transcriptional modification and ribonucleoparticle (RNP) formation, nuclear import, and reverse transcription. The L1 then integrates into its new genomic location (Ostertag and Kazazian, 2001a). The process of retrotransposition is illustrated in Figure 1.4.

**Transcription**

Important sequences for L1 transcription include a Ying Yang 1 (YY1) binding site (Athanikar et al., 2004; Becker et al., 1993) that appears to direct transcription from the first L1 nucleotide, a runt-domain transcription factor 3 (RUNX3) binding site (Yang et al., 2003), two SRY binding sites for transcription factors from the family of the testis-determining factor gene SRY (the SOX family) (Tchenio et al., 2000) and an Antisense Promoter (ASP) (Speek, 2001). Each of these features will be briefly considered in the following sections.
Role of the sense promoter in transcription

The first 670 bp of the 5′ UTR region in L1 serves as a promoter for L1 transcription. The region between position +1 to +100 displays the most promoter activity even though no TATA-box is present (Swergold, 1990). L1 transcription was initially reported to initiate at or near +1 of the element (Swergold, 1990) although transcription initiation sites have been reported to start at positions upstream and downstream of the L1 element (-9 to +4) with only a minority of L1s initiating transcription at position +1 (Lavie et al., 2004). L1s have also been shown to display extra nucleotides between the 5′ TSD and the start of the L1 (based on the L1 5′ consensus starting sequence of 5′-GGAGGAGCC…-3′). These sequences do not belong to either the L1 or the TSD (Lavie et al., 2004). A high reported frequency of G nucleotides between the 5′ TSD and the L1 start may possibly be due to L1 RT reverse transcribing the RNA 5′-
7-methylG CAP structure that is added upon RNA Pol II-mediated transcription. This phenomena is absent in Alu retrotransposition intermediates which are transcribed by RNA Pol III and to which no cap is added (Lavie et al., 2004). This phenomenon is another indication that L1 uses RNA Pol-II for transcription. If several upstream G nucleotides are observed between the 5′ TSD and the L1 start site this could be due to repeated L1 retrotransposition cycles (Boeke, 2003; Lavie et al., 2004).

**YY1 binding sites**

A binding site for YY1, a ubiquitously expressed RNA Pol II transcription factor (Kurose et al., 1995), has been mapped to positions +13 to +21 in the 5′ UTR, and is thought to play an important role in L1 transcription. This role involves enhancing accurate transcription initiation rather than initiating it (Athanikar et al., 2004) as elements lacking the YY1 site have functional promoters.

In addition mutations in the YY1 protein binding site have only minor affects on transcription rates and the YY1 site is not essential for transcription to occur (Athanikar et al., 2004), although deletion of the YY1-binding site in the first 20 bp is reported to reduce transcription by 5-fold (Singer et al., 1993). However if transcription cannot start at the +1 nucleotide in the 5′UTR, it must initiate from another position either upstream or downstream of the +1 site. Transcription downstream of the +1 site leads to a reduction in length of the element’s progeny and so ultimately reduces the number of L1s that are able to retrotranspose autonomously by allowing progressive truncation of the 5′ UTR (Athanikar et al., 2004). Most full-length elements with an intact 5′ UTR initiate transcription at or near the +1 nucleotide site and this ensures that the internal promoter has every chance of being intact in the progeny. This also means that the progenitor and the progeny are then able to retrotranspose autonomously (Athanikar et al., 2004).
CHAPTER 1. LITERATURE REVIEW

RUNX3 binding sites

The RUNX family contains heterodimeric transcription factors that are composed of $\alpha$ and $\beta$ subunits (Yang et al., 2003). The $\alpha$ subunit is a homologue of the *Drosophila melanogaster* segmentation gene *runt*. It contains a conserved region of 128 amino acids which is needed for DNA binding and heterodimerisation with the $\beta$ subunit (Yang et al., 2003). There are three potential binding sites in the human L1 5' UTR for the RUNX3 transcription factor. Two of these RUNX3 sites occur on the sense strand and one is found on the anti-sense strand. The sites fall between +83 to +100 and between +389 to +407 on the sense strand and at +526 to +508 on the anti-sense strand (Yang et al., 2003). Mutagenesis studies have been performed involving the RUNX3 sites located in the 5' of transfected L1 constructs. Mutations in the first RUNX site decrease retrotransposition activity: a single nucleotide mutation (84G>A) decreases retrotransposition by 85% (Yang et al., 2003).

By contrast, mutations in the second and third RUNX3 sites have no significant effect on retrotransposition activity probably due to the fact that these sites are outside the first 100 bp of the 5' UTR the region that has been determined to be the most important for the transcription initiation (Yang et al., 2003).

SOX/SRY binding sites

The SRY (Sex-determining Region Y) gene (a member of the SOX family) has been demonstrated to regulate L1 promoter activity (Tchenio et al., 2000). Human L1s contain two functional binding sites for transcription factors of the SRY family, namely SOX factors. These binding sites are responsible for efficient trans-activation of the L1 promoter by the SOX family (Tchenio et al., 2000). The two potential binding sites for SOX transcription factors are located in the first 670 nucleotides of the L1 sequence. The first site, SRY$_A$, is located between nucleotides 472-477 and SRY$_B$ is located between positions 572-577. Both of these SRY binding sites have the sequence AACAAA and interact with the DNA-binding domain of SRY (Tchenio et al., 2000). Binding of the SRY sites in the L1 5' UTR can drive transcription of the L1 promoter in
cell culture and point mutations abolish SOX protein binding, preventing transcription (Tchenio et al., 2000).

Antisense promoter

L1 elements also have another transcriptional region, in addition to the promoter located on the sense strand. The antisense promoter (ASP) drives transcription in the anti-sense direction relative to normal L1 transcription. The ASP allows for co-transcription of the 5’ UTR and the genomic regions upstream of the L1 element (Speek, 2001). The minimal promoter region of the ASP is from +399 bp to +467 bp with additional nucleotides 3’ also contributing to activity. A deletion in this region reduces ASP activity by 4-5 fold (Speek, 2001). In addition it has been shown that over 1/3 of L1HS 5’ UTRs show high ASP levels in cell culture (Nigumann et al., 2002; Speek, 2001).

Interaction between L1 and the splicosome

It is currently unknown how the L1 mRNA is exported from the nucleus but the L1 transcript is not ordinarily spliced. However there is evidence that it can interact with the spliceosome, as cell-culture retrotransposition assays require splicing of a reporter construct intron (Buzdin et al., 2003). In addition L1 machinery has been proposed to be involved in the creation of new families of chimeric retrotranscripts and pseudogenes in the human genome through a process called template switching (Buzdin et al., 2003, 2002; Esnault et al., 2000).

One of the most abundant families of these chimeric pseudogenes or retrotranscripts contains a full copy of a U6 small nuclear RNA (snRNA) fused at the 3’ end with a truncated 5’ L1 that occurs in at least 56 families (Buzdin et al., 2002). These retrotranscripts are flanked by TSDs containing the preferred L1 EN nicking site of TTAAAA suggesting that these chimeras utilise the L1’s integration machinery (Buzdin et al., 2002; Jurka, 1997). The structural characteristics of U6-L1 chimeras suggest that they are integrated into genomic DNA as pre-formed units rather than occurring after genomic integration (Buzdin et al., 2002).
The creation of U6-L1 chimeras is proposed to occur by the mechanism of template switching from L1 mRNA to cellular RNA during L1 reverse transcription (Buzdin et al., 2002) and other similar chimeric retrogenes have likely been created this way (Buzdin et al., 2003).

There is a strong correlation between the presence of chimeric transcripts and the U6, U3, and U5 snRNAs involved in the spliceosome. Ninety three percent of the 5' ends of the chimeras identified by Buzdin et al. (2003) were DNA copies of U6, U3, and U5 (82, 10 and 1 %). A high frequency of template switching to spliceosomal RNAs suggest a relationship between the L1 machinery and the spliceosomal RNAs (Buzdin et al., 2003) and the close spatial localization of L1 reverse transcription/integration complex and spliceosomes. Template switching has also been suggested as a way that new L1 families evolve (Buzdin et al., 2003) due to the fact that mammalian L1s have different promoter types (Furano, 2000).

### Translation of L1

Once the L1 mRNA has been exported to the cytoplasm it is translated to produce two proteins designated ORF1p and ORF2p. ORF1p and ORF2p associate with the L1 RNA from which they were translated to form a RNP (see Figure 1.4). This process is termed cis preference (Wei et al., 2001). The mechanism of cis preference is currently unknown but two main suggestions exist. The first suggestion is that the close proximity of the L1 proteins to the L1 RNA at the ribosome mediates this effect (Moran and Gilbert, 2002) and the second suggestion is that L1 proteins have a short half-life in the absence of L1 RNA and thus must bind the RNA that encoded them in order to be stabilised (Moran and Gilbert, 2002). It is thought that cis preference acts to ensure that the L1 proteins associate with the L1 functional mRNA and also so that the proteins interact with the L1 RNA rather than any cellular RNAs that may be present (Moran and Gilbert, 2002). It has been predicted that as few as two ORF2p molecules are generated per L1 transcript (Moran and Gilbert, 2002; Wei et al., 2001) while multiple copies of ORF1p have been detected in vitro and in vivo (Hohjoh and Singer, 1996).
Experiments using L1 constructs in cell culture have suggested that ORF1p is translated by ribosomal initiation in the 5’ UTR followed by ribosomal scanning around nt 661 (McMillan and Singer, 1993) Recently, Dmitriev et al. (2007) have suggested that the translation activity of the 5’ UTR of L1 mRNA is dependent on a cap structure (m7GpppN) which supports previous claims by Moran et al. (1996). Uncapping is reported to reduce the initiating activity of the L1 5’ UTR. Translation can also be inhibited by AUG codons in the 5’ UTR (Dmitriev et al., 2007).

One suggestion for ORF2 translation is that it generally initiates from the first in-frame methionine codon of ORF2, and that neither ORF1 nor the intergenic spacer are needed for ORF2 translation (Alisch et al., 2006). It has also been suggested that ORF2 translation initiation is not directed by internal initiation at all but by a complex mechanism of ribosomal retention (Dmitriev et al., 2007) but definitive evidence is lacking.

Open Reading Frame 1 protein (ORF1p)

ORF1 encodes a 40-kDa RNA-binding protein designated ORF1p that has a high affinity RNA-binding activity and which forms a Ribo-Nucleoprotein Particle (RNP) complex together with L1 RNA and ORF2p (Hohjoh and Singer, 1997).

The understanding of human ORF1p is rather incomplete but some insight has been gained from studies in mouse. Murine L1 ORF1p binds to single-stranded nucleic acids and acts as a nucleic acid chaperone that is required for retrotransposition (Martin et al., 2003, 2005b). The binding activity is located in a basic domain in the C-terminus of the protein which is generally well conserved across species (Martin, 2006). The chaperone activity of ORF1p is thought to direct the rearrangement of nucleic acids to the most stable conformation (Martin et al., 2005b). During target-site-primed reverse transcription (TPRT) ORF1p is thought to have a role in melting and displacement of the DNA strands while also holding the L1 RNA in place, to aid the priming of reverse transcription (Martin et al., 2005b).
ORF1p contains several blocks of conserved amino acids that are present in human, mouse, rat and rabbit L1s (Hohjoh and Singer, 1996). The blocks of amino acids that are located near the C-terminus of ORF1p are REKG 235-238, ARR 260-262, and YPSKLS 282-287. These residues have been shown to decrease retrotransposition frequency when mutated, compared to wild type L1’s retrotransposition activity (Moran et al., 1996). Insertion of a stop codon at S119 in ORF1p and the replacement of RR 261-262 with AA (Alanine, Alanine) eliminates retrotransposition completely in a cultured cell assay (Moran et al., 1996). It has also been shown by Martin et al. (2005b) that mutations of RR 261-262 to AA alter the two known functions of mouse ORF1p, (binding to single-stranded nucleic acids and nucleic acid chaperone activity) providing evidence that these functions are conserved and required for retrotransposition.

Open Reading Frame 2 protein (ORF2p)

ORF2 has the potential to encode a 150-kDa protein, designated ORF2p. ORF2p has three conserved domains, one encoding an endonuclease (EN) domain (Feng et al., 1996), another encoding a reverse transcriptase (RT) domain (Mathias et al., 1991) and a carboxy-terminal cysteine rich domain (C) of unknown function (Mathias et al., 1991). The endonuclease and reverse transcriptase are considered in more detail below.

Endonuclease domain

Despite the conservation of this domain, L1s lacking a functional EN domain are still able to integrate into non-typical DNA sites utilizing pre-existing DNA nicks for integration, through a process termed endonuclease-independent integration. This process is significantly less efficient than L1 integration in the presence of a functional EN domain (Morrish et al., 2002). ORF2p mutations in the endonuclease domain (N14A, E43A, D145A, D205G and H230A) have been shown to decrease retrotransposition frequency (Feng et al., 1996).
CHAPTER 1. LITERATURE REVIEW

Reverse transcriptase domain

The L1 RT domain is related to those in all non-LTR retrotransposons (Malik et al., 1999) and shows some sequence similarity with RTs from LTR retrotransposons and retro-viruses (Xiong and Eickbush, 1990). Despite this sequence similarity, the fact that ORF2p functions in the nucleus using genomic DNA as a primer during the process of TPRT, rather than in the cytoplasm using a tRNA primer like retroviral and LTR-retrotransposons, suggests a major functional divergence.

ORF2p contains a number of conserved amino acids that have been demonstrated experimentally to be important for retrotransposition. These are N5, R363, K485, Q689, T760, D795, I1220 and S1259. Amino acid changes in these functional domains adversely affect retrotransposition of the L1 constructs in cell culture. The amino acid changes of I1220M and S1259L, which are located downstream of the conserved C domain of ORF2 (Lutz et al., 2003) account for 20% and 80%, respectively, of the difference in retrotransposition activity between L1.2A and L1.2B, alleles of the active L1.2 element (Lutz et al., 2003). These changes deviate from the highly active L1 consensus derived by Brouha et al. (2003) and it seems that amino acid differences, not nucleotide changes most profoundly affect retrotransposition activity (Lutz et al., 2003). It has been suggested that I1220M and S1259L affect retrotransposition activity by altering template binding, RNP transport and TPRT (Lutz et al., 2003).

Other important conserved amino acids identified by mutation analysis of ORF2p are located between the EN and the RT domains. The mutation R363G accounts for <10% of the difference in retrotransposition between the L1.2A and L1.2B. Within the RT domain another mutation, Q689R, also deviates from the highly active consensus but this has relatively little affect on retrotransposition (Lutz et al., 2003).

Finally a set of conserved amino acid motifs in ORF2p, Y115, T192, SDH 228-230, FADD 700-703, HMKK 1091-1094 and SSS 1096-1098 have also been shown to reduce retrotransposition activity when mutated (Feng et al., 1996; Moran et al., 1996).
Target primed reverse transcription (TPRT)

Once the RNP particle has entered the nucleus, L1 reverse transcription and integration is initiated, by TPRT (see Figure 1.5). The first step of TPRT is indicated by A) in Figure 1.5. The EN domain typically cleaves the DNA target site at a consensus sequence of 5'-TTTT↓AA-3' (Feng et al., 1996; Jurka, 1997) exposing a 5' PO₄ (phosphate, indicated by the yellow circle in Figure 1.5) and 3' OH termini. The L1 RNA (the red line in Figure 1.5) inserts at the cleavage point shown at B) in Figure 1.5 and uses the 3'OH as a primer for reverse transcription producing a cDNA templated by the L1 RNA, shown by the green line.

The EN domain then cleaves the second DNA strand of the target site to produce a double stranded break shown at C) in Figure 1.5. The cDNA then inserts into the break shown at D) in Figure 1.5 and the break is filled by host DNA repair mechanisms producing short direct repeats known as target site duplications (TSDs) shown by purple triangles at E) in Figure 1.5. TSDs are usually between 7-20 bp in length and are usually found immediately flanking the 5' and 3' ends of the L1 although they can be found relatively far away from the ends of the L1 element if the element has undergone 5' or 3' transduction (see sections on 5' or 3' transduction for more details).
Figure 1.5: Target-site primed reverse transcription. A) Endonuclease (EN) cuts the DNA B) 3’OH acts as a primer for reverse transcription producing a cDNA templated by the L1 RNA. C) EN domain then cleaves the second DNA strand of the target site to produce a double stranded break D) the break is filled by host DNA repair mechanisms producing short direct repeats known as target site duplications (TSDs) shown by purple triangles.
CHAPTER 1. LITERATURE REVIEW

How L1 EN dependent retrotransposition events shape the human genome

L1 EN-dependent retrotransposition can alter the genomic landscape and structure of the genome by mobilising upstream and downstream flanking DNA in cis through a process called L1-mediated transduction (Goodier et al., 2000; Lander et al., 2001; Moran, 1999; Pickeral et al., 2000; Symer et al., 2002; Szak et al., 2003). L1s therefore have the potential to mobilise exons and other regulatory sequences to other sites in the genome enabling them to alter the function of genes in a process called exon shuffling (Moran et al., 1999). EN-dependent retrotransposition can also generate truncated 5' insertions and inversions (Ostertag and Kazazian, 2001b), and drive trans-retrotransposition of Alu and SVA (SVA (SINE-R, VNTR and Alu) insertions, all of which help to shape the human genome. Also it has been estimated that ~6% of EN-dependent L1 retrotranspositions result in a significant deletion of genomic DNA sequence (Gilbert et al., 2002; Symer et al., 2002) at the insertion site.

Three prime transduction

The process of 3' transduction involves the mobilisation of downstream L1 flanking sequence due to the L1’s weak polyadenylation signal being by-passed in favour of a stronger polyadenylation signal in the 3' flanking DNA (Moran et al., 1999). The weak polyadenylation signal may allow L1s to reside within introns and not cause disruption to gene expression by inducing premature polyadenylation (Moran et al., 1999). Three-prime transduction is more common than 5' transduction (Goodier et al., 2000), although both forms may occur in association with a single L1 element insertion.

The first recognised 3' transduction event was associated with a 2-kb L1 insertion into the Dystrophin gene of a muscular dystrophy patient (Holmes et al., 1994). The insertion carried at its 3’ end a 500-bp piece of sequence that was retrotransposed along with one 5’ truncated L1. Another two instances of L1 with 3' transductions are known to cause disease in humans. A 538-bp L1 insertion inserted into the APC gene that was responsible for a case of colon cancer (familial adenomatous polyposis) has a 180-
Figure 1.6: A typical 3' transduction event showing the mobilisation of the three prime TSD (purple arrow head) along with genomic DNA (yellow block) 3' of the L1 (green block arrow) when the L1 has undergone retrotransposition and inserted into a new genomic location (blue blocks). The pink arrow heads indicate the new TSDs of the L1 insertion.

bp 3' transduction (Miki et al., 1992). In addition a 1.7-kb L1 insert into the CYBB gene causing chronic granulomatous disease has a 280-bp transduction that showed no homology to L1 or CYBB sequences (Brouha et al., 2002).

An example of a 3' transduction event is shown in Figure 1.6. Typical hallmarks of 3' transduction are a consensus polyadenylation sequence of AATAAA or ATTAAA and downstream segments of transduced sequence ranging from 89-975 bp that are located up to 10-35 bp upstream of the preceding polyA tail (Pickeral et al., 2000). The process of 3' transduction allows for the 3' ends of the L1 to evolve rapidly and create a set of L1s whose 3' ends have become structurally diverse (Pickeral et al., 2000).

Goodier et al. (2000) predicted that the insertion of transduced sequences has contributed approximately 19 Mb or 0.65% to the human genome based on there being around 400,000 L1s in the haploid human genome, the mean length for 3' transduction being 207 bp and the frequency of 3' transduction being around 23%. This estimate however assumes that transduction frequency has remained constant throughout evolution and that the selected elements are typical. Goodier et al. (2000), also state that 3' transductions are associated more frequently with full-length L1s and that younger elements may have a higher transduction frequency that older elements. Another study using 129 full-length elements with high similarity (94%) to a known
CHAPTER 1. LITERATURE REVIEW

active full-length L1 element, L1.2 by Pickeral et al. (2000), used computational analyses to reveal that the mean 3’ transduced sequence length was 231 bp. These authors estimated that the contribution of 3’ transduced sequence to the human genome is between 25.2 and 30.5 Mb or ~1% of the human genome, based on there being ~600,000 L1s in the human genome.

As described above, a number of estimates based on genome-wide analyses have been produced for the frequency of L1-mediated 3’ transduction, Lander et al. (2001) estimated the frequency to be ~21%, Goodier et al. (2000) 23%, Pickeral et al. (2000) 24-32%, and Myers et al. (2002) ~26%. Based on these various estimates Chen et al. (2005) suggested that the frequency in vivo of L1-associated 3’ transduction is ~25%, indicating that this phenomenon is a significant component of L1 transposition biology.

In addition to determining the frequency with which L1s mobilise their flanking DNA, identifying L1s that have undergone 3’ transduction also allows the potential discovery of progenitor and progeny elements. L1s with similar transduced sequences could be either offspring of elements carrying the same transduction that have undergone polyadenylation at the same or very similar sites or siblings originating from a common progenitor element. The average size of an L1 transduction family (i.e. elements carrying the same or very similar transduced sequence) is 2.5 L1s and there are 25 families derived from 63 elements in human genome NCBI build 25 (2001), as determined by the 3’ transduction analysis of Szak et al. (2003). L1s that have undergone transduction will have new TSDs and the distinction of these new TSDs from the old TSDs that have been carried along with the genomic flanking DNA, allows the determination of the L1 boundary. The transduced sequence at either the 5’ or 3’ end of an element can then be used to relate L1 families or lineages. For an L1 element to be a likely progenitor the element must be full length having an internal promoter (assuming that 5’ truncated elements do not retrotranspose). If two elements are not full length but share the same transduced sequence they are likely siblings of the same progenitor (Szak et al., 2003).

With the goal of automating the analysis of L1 transduction identification, Szak et al. (2003) developed an algorithm for the detection of 3’ transduction in L1s based on what they consider are four important characteristics. These are, 1) A sequence must not
appear on the same chromosome, 2) The downstream sequence must be greater than or equal to 90% identical to the transduced sequence 3) The length of the alignment must be greater or equal to 30% of the putative transduced length 4) The orientation of the matching downstream sequence must be the same as the L1 upstream sequence.

The vast majority of 3’ transduction events identified by this algorithm have transduced sequences of less than 500 nucleotides with a median length of 290 nucleotides (Szak et al., 2003).

**Five prime transduction**

The process of 5’ transduction involves the mobilisation of upstream flanking sequence by an adjacent non-L1 promoter transcription initiation in the 5’ flanking sequence (Lander et al., 2001; Symer et al., 2002). A typical example of a 5’ transduction event is shown in Figure 1.7. Transductions at the 5’ end of an L1 are not as common as transductions at the 3’ end and are typically smaller, although there have been two documented examples of 5’ transduction with sizes of 143 bp and 215 bp (Lander et al., 2001). The addition of transduced sequence at the 5’ end of the L1 element could also possibly lead to the acquisition of new L1 promoters, if they utilize a mechanism to ensure transcription initiation upstream of the cis acting promoter sequences.

**Inversion**

Inversion involves 5’ truncation and inversion of the 5’ end of an L1 element. Inversion may have no effect on the sequence content of the element or it may cause deletions or duplications. Ostertag and Kazazian (2001b) have proposed a model for L1 inversion called twin priming.

Twin priming is based on the L1 mechanism of TPRT (see Figure 1.5), but with a few major differences. As in TPRT the L1 EN domain causes a nick in the genomic DNA, making available a T rich tail, or polyT primer. The polyA tail of the L1 RNA anneals to the polyT primer and the L1 RNA is used as a template by L1 RT
to start reverse transcription. This process is the same as normal TPRT. The differences between TPRT and twin priming are that the L1 EN cleaves the second strand of genomic DNA before the reverse transcription of the RNA has been completed and so produces another internal primer (Ostertag and Kazazian, 2001b). This new internal primer site invades the L1 RNA and primes reverse transcription. The RNA is removed from the RNA/cDNA structure and the single stranded cDNAs anneal at regions of complementarity. DNA synthesis is then completed and this produces an inverted L1 with perfect TSDs. If the L1 RNA is represented by 5’-A-B-C-D-E-3’ before inversion the L1 element after inversion could be represented, 5’-C-B-D-E-3’ where the 5’ end of the L1 is truncated (i.e. ‘A’ is removed from the L1 structure) (Ostertag and Kazazian, 2001b).

The points of inversion may cause a deletion or duplication in the L1. Inversion can be divided into three main categories, those inversion events that have large deletions (>50 nt), small deletions (1-50 nt) and those with duplications (1-50 nt) (Ostertag and Kazazian, 2001b). It has been suggested that these variable inversions are caused by the disassociation of the RT from the L1 RNA template occurring at variable sites (Ostertag and Kazazian, 2001b). The inversion points also seem to be clustered towards the 3’ end of the L1.
CHAPTER 1. LITERATURE REVIEW

39

Exon shuffling by retrotransposition

L1s can retrotranspose into genes and can transduce 3’ flanking genomic DNA to new genomic locations. If this 3’ transduced sequence includes exons, this process can lead to the formation of new genes by exon shuffling. L1s can retrotranspose into either the sense or antisense strand of a gene and there is apparently little or no bias for or against genes as sites for L1 retrotransposition (Moran et al., 1999).

While Moran et al. (1999) suggested the possibility of exon shuffling, verified examples are rare. A chimaeric gene has been generated by exon shuffling in the new world primate the owl monkey (Aotus trivigatus) (Sayah et al., 2004). In old world primates TRIM5-α blocks human immunodeficiency type 1 virus (HIV-1) infection after the virus has entered the cells (Sayah et al., 2004). Cyclophilin A (CypA) binding to the viral caspid performs a similar activity in human cells. In owl monkeys, members of the new world monkeys, formation of the cyclophilinA/TRIM5 fusion gene prevents HIV entry into owl monkey cells (Sayah et al., 2004).

L1 trans-driven retrotransposition

L1 3’ transduction is a potential pathway for shuffling exons (Moran et al., 1999) however, this method is limited to regions 3’ of the L1 (Ejima and Yang, 2003). Genomic DNA that is far away from the L1 may also be mobilised by the trans action of L1 (Esnault et al., 2000), illustrated in Figure 1.8. This is where LINE proteins retrotranspose, with a much reduced efficiency, transcribed DNA by acting in trans on a cellular RNA. L1 retrotransposition in trans results in the formation of processed pseudogenes that are generated by reverse transcription of intronless mRNAs. The cDNA is integrated into its new genomic location presumably by the L1 machinery via the TPRT process (Esnault et al., 2000). L1 elements that can induce the retrotransposition of cellular mRNA indicates that they are not limited to reverse transcription of their own mRNA (Dewannieux et al., 2003).

Alu elements undergo a similar mechanism of amplification. L1 proteins are involved in Alu retrotransposition as human Alu inserts have TSDs and endonuclease cleavage sites
that closely resemble those generated by the L1 EN (Dewannieux et al., 2003). ORF2p is involved in this process, but ORF1 it is not required in cell culture assays. In principle this means that those L1 elements with a disrupted ORF1 but intact and functional ORF2p could mobilise Alu sequences even if they cannot retrotranspose themselves (Dewannieux et al., 2003).

**Endonuclease-independent L1 insertion**

L1 insertions may also occur that have no TSDs flanking the 5′ and the 3′ of the L1. This process is termed endonuclease-independent insertion (Morrish et al., 2002) (see Figure 1.9).

It is thought that this process occurs when an L1 inserts into a double stranded DNA break that has not been caused by the L1 endonuclease. L1s that have inserted by this alternative method often show a truncation at the 5′ end of the L1, lack a polyA tail

---

**Figure 1.8:** L1 trans-driven retrotransposition showing the ORF proteins acting in trans on cellular RNA to produce a processed pseudogenes.
and show some 3′ truncation, as well as lacking TSDs (Morrish et al., 2002). They are also often inserted at sites that are not preferred substrates for L1 endonuclease cutting. Some L1s inserted by this method also show target-site deletions ranging from 11 bp to 1.5 kb while some also display additional nucleotides at their 5′ ends that have been suggested to arise from non-templated nucleotide additions (Morrish et al., 2002).

mRNA fragments have also been shown to be able to accompany endonuclease-independent retrotranspositions in cell culture but it is not clear if the sequences are reverse transcribed by the L1 reverse transcriptase (Morrish et al., 2002).

1.5 L1 polymorphism classification and detection

Having considered the processes by which L1 elements move around and remodel the host genome the next section considers their classification.

The L1-Ta family have been showed to be polymorphic, with respect to presence or absence, in human populations (Badge et al., 2003; Brouha et al., 2003; Konkel et al., 2007). L1 insertions in autosomal loci may display 3 genotypes when analysed in human populations. L1 insertions may be homozygous filled (+/+ i.e. only the filled site of the L1 insertion is observed, heterozygous filled/empty (+/-) i.e. one copy of the
L1 element insertion and one copy of the empty site is observed, or homozygous empty (-/-), \textit{i.e.} no L1 insertion is observed in the screened individual, only the empty site is present. The following section discusses laboratory and computational methods that have been used to detect novel polymorphic insertions in human populations.

\textbf{ATLAS (amplification typing of L1 active subfamilies)}

Current DNA sequence databases cannot represent the entirety of human genome variation as they are derived from a small number of individual humans. A technique has been developed, Amplification Typing of L1 Active Subfamilies (ATLAS), that can selectively amplify and display DNA fragments from multiple individuals simultaneously that may be absent from DNA sequence databases \cite{Badge2003}. This technique is able to select loci in genomic DNA that contain young L1 insertions (specifically of the Ta family) and can be specific for the L1 5’ or 3’ termini. This technique has been shown to identify polymorphic elements that were not present in the HGWD (\textit{circa} 2001). The detection of polymorphic insertions can include L1 elements at low allele frequencies in different populations as well as the detection of putatively private insertions (\textit{i.e.} the L1 element is only seen in one individual screened). During the development of this technique a database of FL Ta insertions was constructed by mining nucleotide accession databases. This resource currently consists of 256 FL L1 elements that are 98\% identical to a known highly active L1 Ta, L1.3 (Accession L19088) and are \(\geq\)5922 bp in length. Manual curation was used to generate extensive annotation for 243 elements including the information that 96 elements contain intact open reading frames. Some of the annotated elements in this resource are known to be capable of retrotransposition in cell culture \cite{Brouha2003}. This resource can be found at \url{http://www.le.ac.uk/ge/ajj/LINE1/} and was last updated in 2002.

\textbf{L1 polymorphism in human genome assemblies}

In 2001 two human genome assemblies were published that had been derived from two distinct sequencing and assembly strategies. The International Human Genome
Sequencing Consortium (IHGSC) produced a publicly available human genome assembly (PHG), whilst a private venture initiated by Celera Genomics also produced its own version of the human genome sequence (CHG). This has enabled researchers to look for structural variation differences between individuals by using an assembly as a reference for experimental laboratory work and also by comparing the two assemblies computationally.

One such analysis conducted in 2002 used BLAST (Altschul et al., 1990) to search draft sequences of the human genome sequence (PHG) and identified 468 unique L1-Ta elements (Myers et al., 2002), 124 of which were full length, with 44 of these containing two intact open-reading frames, suggesting they may be capable of retrotransposition. The analysis of L1 polymorphism by Myers et al. (2002) estimated that L1-Ta elements show a high level polymorphism, in the region of 45%, (of 262 tested). This is considerably higher than the estimate of Bennett et al. (2004).

In 2007 Konkel et al. (2007) analyzed the two main human assemblies for L1 insertions that are specific to the each of the respective assemblies, i.e. they show a presence or absence between the two assemblies. This led to 34 loci being identified as polymorphic between the assemblies (28 PHG, and 6 CHG). Thirty two out of 34 potentially polymorphic loci were validated by PCR: 1 insertion was not found in any of the 80 individuals tested, and one other could not be validated because PCR primers could not be designed due to the repetitive nature of the flanking DNA. Of the 32 loci identified, only 25% of the polymorphic elements (8 elements) were full-length (FL) (≥5922 bp). This study found that polymorphic L1 elements are biased towards FL insertions or small insertions of ≤1000 bp, with 75% of insertions being found to belong to these two classes.

**L1 polymorphism in trace archive sequences**

L1 polymorphism data can also be discovered by mining raw DNA sequence reads, either from genome sequencing or re-sequencing projects. These pieces of sequence are often not assembled into contigs but are often derived from many unrelated individuals.
and so are an excellent source of information about DNA variation, which has led researchers to discover novel L1 polymorphic elements.

An analysis by Bennett et al. (2004) used a combination of custom perl scripts and publicly available programs such as BLAST (Altschul et al., 1990) and RepeatMasker (Smit et al., 1996-2004) applied to 16.4 M human trace archive sequences from the CSHL SNP consortium (TSC), NCBI Trace archive database (Whole Genome Shotgun sequences (WGS)) and whole chromosome shotgun sequence (WCS). This in-silico approach discovered 65 non-redundant L1 insertion deletion (indel) polymorphisms, all with polyA tails and flanked by TSDs. After Alu insertions, L1 indels were the second most abundant polymorphic transposon, comprising 12.6 and 11% of TSC and WCS datasets. The majority of indel polymorphisms observed of the TSC and WGS data sources were members of the young L1 Ta family (12.1 and 9.9% of the total number of polymorphisms detected). The number of polymorphic Ta-1 L1s detected was consistent with the fact that 16 of the 17 known disease-causing insertions have come from this L1 subfamily (Chen et al., 2005). If all families of L1 elements are taken into consideration then the L1 family has one of the lowest average polymorphism frequencies (0.00018) of all transposons in the human genome, with one polymorphic insertion per 5556 copies. However if the L1-Ta family are considered on their own, then the average polymorphism frequency increases dramatically. If the L1-Ta sub family has ~520 copies (1040 in a diploid genome) in the human genome (Myers et al., 2002) then the highest polymorphism frequency is 0.161, or one L1-Ta insertion per 6.2 copies. This is a strong indication that active L1 subfamilies and in particular L1-Ta are very frequently polymorphic.

1.6 Genomic distribution of L1 in the human genome

Having considered that the Ta-1 family is frequently polymorphic in human populations, the distribution of these sequences and their insertion preference is discussed in the following sections.
CHAPTER 1. LITERATURE REVIEW

Insertion bias of L1 and Alu elements in the human genome

LINEs can be found at all regions within the human genome, but they are more abundant in gene-poor, AT-rich (GC-poor) regions, with low recombination rates (Lander et al., 2001).

In the human genome L1s can be extremely abundant in certain areas, for example on chromosome Xp11 there is a 525-kb region that has an overall transposable element content of 89%, and within this region there is 100-kb of DNA in which L1 sequences comprise 89% (Lander et al., 2001). Also there are extremely high densities of Alu on chromosome 7q11, the ancient transposons MIR on chromosome 1q36, and LINE2 on chromosome 22q12 (Lander et al., 2001).

Other regions of the genome have low repeat density. This may be due to the regions in question being unable to tolerate insertions, due to the presence of essential cis-regulatory elements. The homeo-box gene clusters, HOXA, HOXB, HOXC and HOXD contain the lowest known density of interspersed repeats, with regions of 100 kb in each of the four genes containing <2% repeat sequence (Lander et al., 2001).

LINE elements have been reported to occur at an approximately four-fold higher density in AT-rich DNA, compared to SINEs such as Alu, which show a five-fold lower density in similarly AT-rich regions of the genome (Lander et al., 2001). The preference for L1s to insert in AT-rich DNA can be attributed to the L1 endonuclease targeting the consensus sequence of TT↓AAAA (Jurka, 1997; Lander et al., 2001) which is more frequent in AT-rich DNA (Cost and Boeke, 1998). It has also been suggested that L1s have adapted to the genome’s structure, like a parasite adapting to a host, becoming over represented in AT-rich DNA due to these regions being gene-poor and so able to tolerate insertions (Cost and Boeke, 1998; Lander et al., 2001). L1 elements that insert into coding regions of the genome could potentially have deleterious effects on the host by disrupting exons and gene transcriptional activity, altering splice sites and ultimately causing the gene to become inactive (discussed in more detail in section 1.7). However, older Alu elements have been discovered to be more abundant in GC-rich regions of the genome even though they use the same L1-encoded endonuclease as L1 elements, whilst younger Alu elements appear to be enriched in AT-rich regions in the genome.
(Lander et al., 2001). The latter point might suggest that older Alu in GC-rich DNA are under positive selection for retention in these genomic regions (Lander et al., 2001). However young, polymorphic and fixed Alu elements are found in genomic regions where the GC content is not statistically distinguishable from the human genome average, suggesting these elements are inserted randomly into the genome and are not subject to any form of selection, positive or negative (Cordaux et al., 2006).

**Ta-1 L1 distribution**

The active L1 family, L1 Ta, which is the youngest L1 family in human populations, is detected more frequently in AT-rich areas (GC-poor), regardless of whether fixed or polymorphic (Boissinot et al., 2004). The GC content of the flanking DNA of Ta-1 insertions is significantly lower than the average GC content of the human genome which suggests that Ta-1 elements are not randomly distributed with respect to GC content but prefer insertion sites in genomic regions were GC content is low (Boissinot et al., 2004). This distribution reflects the ancestral L1 families such as L1PA2 and L1PA5. The discovery that Ta-1 elements are more abundant in GC-poor regions contradicts analysis from Ovchinnikov et al. (2001) who reported that Ta L1 elements are inserted randomly regardless of the GC content.

Ta-1 L1s are found on all autosomal chromosomes although the number per chromosome ranges from 1 insertion on chromosomes 17 and 20 to 38 insertions on chromosome 4. Large chromosomes have more Ta-1 inserts than smaller ones and the size of the chromosome is nearly proportional to the number of insertions found on it (Boissinot et al., 2004). However the sex chromosomes are reported to not be as enriched in Ta-1 elements compared to older L1 elements (Boissinot et al., 2004; Lander et al., 2001). Chromosome 4 contains significantly more Ta-1 insertions that expected for its length and gene density and this is not observed when L1PA2 and L1PA5 are analysed (Boissinot et al., 2004).
Identity by descent

*Alu* and L1 insertions offer a unique advantage over other polymorphisms when studying population genetics as the presence of the same insertion in two individuals is thought to demonstrate identity by descent (IBD) ([Salem et al., 2005](#)). This is due to the idea that two young independent insertions would be very unlikely to insert independently into exactly the same genomic location ([Batzer and Deininger, 2002](#)). Using *Alu* and L1 elements for these types of studies also has an advantage, as the ancestral state is always known (i.e. the empty site with no insertion). The elements in the human genome that show polymorphism are younger, and therefore likely still active elements and can show different allele frequencies between different populations. Insertion homo-plasy can occur over long evolutionary periods between different organisms ([Vincent et al., 2003](#)), however in the human genome the use of young *Alu* and L1 elements in population studies can be considered to be homoplasy free due to the short evolutionary period these elements have been in the genome ([Salem et al., 2005](#); [Vincent et al., 2003](#)). When comparing human L1 insertion sites with other species, [Vincent et al. (2003)](#) observed a very low frequency (0.52%) of parallel insertions of any type at L1 insertion sites, and did not observe the insertion of another L1 element, suggesting that L1 insertion polymorphisms are homoplasy free.
1.7 Effects of L1 sequences in the human genome

The insertion of a L1 element can affect the human genome in a number of different ways. This section discusses the effects L1 insertions have on gene expression, how L1 sequences can be deleterious to the genome, how L1s cause disease, ectopic recombination and X-inactivation.

L1 insertions are deleterious

If an L1 element insertion reduces the fitness of the individual that possesses it, then it will be considered deleterious (Graham and Boissinot, 2006) and will be unlikely to reach high allele frequencies in human populations. Such deleterious insertions will probably be lost rapidly and be very rarely observed (Boissinot et al., 2006; Graham and Boissinot, 2006). Full length (FL) L1 Ta-1 elements are present in the human genome at a significantly lower frequency than Truncated (TR) Ta-1 insertions. This suggests that FL L1s impose a higher fitness cost than TR L1s and therefore are more deleterious (Boissinot et al., 2006). FL elements appear to be under different selective constraints than TR elements, with FL Ta-1 elements under negative selection and TR elements (and Alu) behaving essentially like neutral alleles (Boissinot et al., 2006). The fact that TR elements appear to be neutral suggests that the low frequency of the FL-containing alleles is due to purifying selection (Boissinot et al., 2006). If TR elements act as neutral alleles then insertional mutagenesis cannot be a major factor for the deleteriousness of FL elements (Boissinot et al., 2006).

Although it is not possible to determine the exact time at which Ta-1 L1s imposed a fitness cost on humans it is likely that L1s were deleterious in ancestral populations as the majority of Ta-1 L1s can be found in all human sub-populations, with only a few being population specific. It is likely that the fitness of human populations is going to continue to be reduced as Ta-1 amplification is still ongoing (Boissinot et al., 2006).
The effects of L1 insertions on gene expression

Gene function can be disrupted when L1 insertions occur in exons. When an L1 insertion occurs in an intron the element can affect gene expression rather than just being spliced out of mRNA. Han et al. (2004) have reported that L1 sequences were present in lower amounts in highly expressed genes (an average of 918 bp of sense L1 and 1,760 bp of antisense L1 per gene) whereas lower expressed genes were found to contain more L1 sequences (an average of 4,760 bp of sense L1 and 8,860 bp of antisense L1 per gene). Therefore these data suggest that the mutational impact of L1s is lower in genes that are lowly expressed than in genes that are highly expressed.

Han et al. (2004) also showed that the presence of the L1 ORF2 in the sense orientation of the gene affects transcription of the host to a greater extent than if ORF2 is inserted in the antisense orientation. The interference in the transcription levels of the gene by L1 is not due to premature transcription termination but due to the A-rich nature of the L1 sense strand. Intronic L1s have been proposed to be able to modulate gene expression (Han et al., 2004) because polymerases do not elongate efficiently through L1 ORF2 sequences and this reduces the amount of transcript produced.

Boissinot et al. (2004) have reported that Ta-1 L1 elements have inserted into the introns of 46 known genes, and the orientation of the L1 insert with respect to the gene was found to be in the antisense orientation twice as often as compared with the sense direction, suggesting that this bias could be due to the negative selection. L1 elements that insert in the same transcriptional orientation as the gene could potentially lead to an introduction of an L1 polyadenylation site and this would lead to premature transcription termination of the gene (Perepelitsa-Belancio and Deininger, 2003).

The effect of the antisense promoter (ASP) in the human genome

As L1 sequences are spread across in the genome in vast amounts, their presence gives rise to the possibility there may be a large number of potentially active ASPs in the human genome that may be competent (Nigumann et al., 2002; Speek, 2001)
CHAPTER 1. LITERATURE REVIEW

The ASP has been shown to drive transcription of a number of gene sequences and analysis of hundreds of ESTs (expressed-sequence tags) has revealed chimeric L1 transcripts that are known to be active in a number of tissues (Nigumann et al., 2002; Speek, 2001). Analysis of chimeric transcripts reveals that they originate from the ASPs of full-length L1s located in introns up to 60kb from protein coding sequences (Nigumann et al., 2002; Speek, 2001). When L1s are inserted into genes in the opposite transcriptional orientation they can produce transcripts that contain the anti-sense 5' UTR of a L1 and a mRNA tail (non L1 sequences) (Speek, 2001) and also have the ability to truncate cellular transcripts by premature polyadenylation (Han et al., 2004). This means that an L1 insertion into an intron of a cellular gene could in theory “break” the gene and produce multiple transcripts from a single transcriptional unit (Wheelan et al., 2005). Wheelan et al. (2005) identified 15 potential genes and transcriptional units which could be broken by an L1 insertion. Three of these transcripts were successfully amplified and analysed. Secernin 3, identified as a potential gene broken by L1 sequence insertion had its ASP-driven transcript isolated by RT-PCR. The ASP transcript generated revealed an intronic L1, followed by 25 bp of exon 5 skipping to exon 7. Exon 6 was presumed to be spliced out. This demonstrated that in HeLa cells the coding regions of Secernin 3 produced two transcripts split by the L1 sequence.

Ectopic recombination

Ectopic recombination refers to recombination between sites, usually containing identical or similar sequences at different locations in the genome resulting in deletions, insertions, or inversions.

The density of retrotransposons has been reported to be higher in regions of the genome where recombination is lower (Boissinot et al., 2001). Boissinot et al. (2001) showed that in the non-recombining region (NRY) of the Y-chromosome the proportion of FL-L1HS elements is higher than on the autosomes, and suggested that FL-L1HS elements have undergone purifying selection which had been much less efficient in the NRY. Song and Boissinot (2007) found that the NRY contained a larger fraction of FL L1 (based on L1 families L1PA2-L1PA6) than on the autosomes and on the X
chromosome. In addition these authors found that long elements and long truncated (TR) elements >1.2 kb tended to be found in non- and low-recombining areas of the genome and this agreed with a previous analysis by Myers et al. (2005). The difference in the distribution of FL and long TR L1 elements compared to short L1 elements must however occur after initial insertion as both types of elements use the same mechanism of replication and insertion (Boissinot et al., 2006; Martin et al., 2005a). Therefore it appears that L1 elements >1.2 kb length are more likely to be subject to negative selection than elements <1.2 kb in length.

The frequency of ectopic recombination exchange should in theory correlate with the recombination rate, and so L1s that insert into areas of the genome with a high recombination rate should have a greater deleterious effect than those residing in low recombining regions of the genome. L1s in high recombining regions of the genome should be selected against and eliminated from the genome (Graham and Boissinot, 2006; Song and Boissinot, 2007).

These distribution data are consistent with L1 sequences being efficient mediators of ectopic recombination. This process can lead to chromosomal rearrangements which are reported to be very deleterious. A documented example is the ectopic recombination event between two LINE-1 sequences in the human gene encoding the beta subunit of phosphorylase kinase (PHKB). This event resulted in a deletion of 7574 bp of genomic DNA including exon 8 of the gene, causing glycogen storage type disease through phosphorylase kinase deficiency (Burwinkel and Kilimann, 1998). Another example of an homologous recombination event between two L1 elements occurred in an individual giving rise to Alport syndrome (AS) and associated diffuse leiomyomatosis (DL) (Segal et al., 1999 Jan).

**L1 disease-causing insertions**

Having considered the indirect deleterious ways in which L1 sequences can interact with their host genome we will now consider the direct disruption of genes by L1 insertion.
L1 retrotransposons have the capability to insert into genes and to promote other retrotransposons to insert into genes. To date there have been 16 reported L1 insertions in humans that cause disease. See Appendix section A.2 for the types of disease, the gene the L1 inserted into and the length of the insertion.

The first cases of L1s causing diseases were reported in 1988 by Kazazian et al. (1988) who found two different 5′ truncated de novo L1 insertions in exon 14 of the human factor VIII gene, causing Haemophilia A in two patients. This demonstrated the fact that L1s were still active in the human genome and were capable of causing disease. Four L1 insertions (all truncated) have been reported in the Dystrophin gene resulting in two diseases, Muscular Dystrophy (Holmes et al., 1994; Narita et al., 1993) and X-linked dilated cardiomyopathy (XLDCM) (Yoshida et al., 1998).

To date, only four L1 disease causing insertions (L1βthal (β-thalassemia), L1RP (Retinitis Pigmentosa), L1CHM (Choroideremia) and L1PDHX (Pyruvate dehydrogenase complex deficiency)) have been found to be full length L1s (≥5922 in length), with the remaining 12 L1s either truncated and/or rearranged. L1βthal and L1RP have been tested in a retrotransposition assay and display an activity similar to a known highly active L1, L1.3 (L19088). The sequence identity between L1βthal and a consensus of active FL-L1HS is 99.4% and for L1RP 99.9% (Kimberland et al., 1999).

While it is estimated that the average human genome contains 80-100 active FL-L1HS sequences (Brouha et al., 2003), L1 insertions only account for approximately 1 in 1200 human mutations (Kazazian, 2004). It is estimated therefore, that 1 in 50 individuals harbour a de novo insertion occurring early in embryonic development or in the parental germline (Kazazian, 2004). In cultured retrotransposition assays 5-10% of de novo insertions occur within the introns of actively transcribed genes, and since approximately 15% of the human genome consists of genes, it has been suggested that there is no active mechanism to stop L1s inserting into genes (Moran, 1999). It remains unclear as to what factors maintain the mutation load generated by L1 at these low levels, despite the prevalence of active elements and their potential to directly disrupt genes.
The role of L1 sequences in X-inactivation

Having considered the mechanisms by which L1s unequivocally can influence the structure and function of their host genome, it is important to also consider roles for L1 sequences that are less well defined.

X chromosome inactivation is a mechanism of gene regulation and involves transcriptionally silencing the majority of the genes on one X chromosome during female embryogenesis. This process occurs to maintain the correct dosage of X-linked genes between XX cells and XY cells, i.e. so that the phenotypic expression of characteristics by genes on the X chromosome are equally expressed in both females who have two copies of the X chromosome (XX) and males who have one copy of the X and one copy of the Y chromosome (XY).

X-inactivation is thought to occur in three stages, the initiation of inactivation early in development, the spread of inactivation from the X-inactivation centre (a unique cis acting region located at Xq13) in cis along the length of the X chromosome, and the maintenance of X-inactivation in all successive somatic cell divisions (Bailey et al., 2000). X-inactivation is brought about by the X inactive-specific transcript (Xist) non-coding RNA, which coats the inactive X-chromosome.

L1 sequences were suggested by Mary Lyon as being candidates for the cis acting feature of X-inactivation in spreading along the X chromosome (Lyon, 1998). The Lyon hypothesis suggests that LINE elements on the X chromosome act as booster elements to promote the spread of Xist mRNA. It has been noted that the X chromosome is enriched with L1 sequences compared with autosomes, with an analysis by Bailey et al. (2000) finding a two-fold enrichment for L1 sequences on the X chromosome. The bulk of this enrichment on the X-chromosome has been found to be from younger L1 subfamilies, with a significant clustering occurring near the time of the eutherian (placental mammals) radiation, and which continued beyond the emergence of primates (Bailey et al., 2000).

Also there is a significant enrichment of L1 sequences at Xq13 where the X-inactivation centre is located (Bailey et al., 2000; Lander et al., 2001), while genomic loci that
escape X-inactivation have significantly reduced L1 density compared to those that are subjected to inactivation (Bailey et al., 2000). These authors propose that the distribution pattern of L1 sequences on the X chromosome plays an important function role in L1 X-inactivation. The clusters of L1 sequences have been proposed to serve, indirectly through a protein complex, as X chromosome binding sites for an Xist RNA/protein complex (Bailey et al., 2000). In this model the over-representation of L1 Sequence at Xq13-q21 ensures the appropriate initiation of X-inactivation and clusters of L1 sequences act to ensure X-inactivation throughout the entire chromosome by synergistic condensation (Bailey et al., 2000). The failure of chromosomal inactivation beyond x:autosomal translocations also correlates with low L1 sequence in those chromosome bands (Bailey et al., 2000).

The accumulation and clustering of L1 sequences of the X chromosome appears to support the Lyon hypothesis, however the enrichment of L1 sequence could also be due to the heterochromatic nature of the the inactive X chromosome simply facilitating L1 accumulation. Therefore it may be the case that L1 sequence has contributed to a more stable mechanism of X-inactivation, rather than being the initial mechanism by which X-inactivation occurs (Bailey et al., 2000).

1.8 Sequencing, Assemblies and Databases

Having discussed the literature relating to how active L1 retrotransposons damage, shape and perhaps cooperate with their host genome, attention in this section will turn to the technologies, data and informatic resources that enable genomic scale analyses of these elements.

DNA sequencing is an important tool that is used in many research areas from genetics, molecular biology, to archaeology and forensic science. One of the most important uses of DNA sequencing in genetics and molecular biology is the generation of complete genome sequences. The DNA sequencing of many genomes provides the opportunity to characterize and explore genetic variability of genes and genomes within
and between species. This can provide valuable genetic information that may indicate that a particular polymorphic variant may cause or increase the risk of a genetic disease.

With the increasing amounts of sequence data being generated, bioinformatics is an important tool for sequence-based life science research. Tools are needed that can automatically organise, search, and annotate genomic data to provide meaningful biological insights, that otherwise would be too cumbersome, costly and slow for manual annotation. Nucleotide sequence databanks, such as GenBank, EMBL, and DDBJ provide researchers with the opportunity to deposit and share their genomic data and allow other researchers to search and compare nucleotide sequences with bioinformatic tools such as BLAST, BLAT and CLUSTALW, or develop their own tools in the form of computer scripts and programs. The following section discusses some of the methods used to sequence, assemble and store genomic sequences containing a high density of repeats.

Genome sequencing and assembly

One of the first methods for sequencing DNA was described by Sanger and Coulson (1975). This technique, sequencing using DNA polymerases and dideoxy terminators, was used to determine the sequence of two parts of the bacteriophage ($\Theta$X174) genome. Developments in this technology led to more efficient and accurate sequencing methods, however these methods could still only be applied to small cloned genomic sequences and small genomes such as bacteriophages (e.g. $\lambda$) (Sanger et al., 1977).

Despite refinements in this technology, and the development of shotgun sequencing in 1982 by Sanger et al. (1982), manual intervention was still required to produce high quality DNA sequences and assemble them, and so this technique was thought to be impractical for sequencing large genomes. As a result, this method of sequencing was only applied to small genomes such as those of bacteria and viruses with small, repeat-poor genomes so that reads from overlapping sequences could be reliably assembled without risk of mis-assembly (Waterston et al., 2002). The genomes of higher animals were known to be much larger and contain a higher proportion of repeated sequences presenting challenges for accurate assembly. Thus new automated
sequencing technologies and computational algorithms were needed to overcome these issues.

The three main methods developed for sequencing large and complex genomes are the Hierarchical Shotgun Assembly (HSA) approach (Lander et al., 2001), the Whole Genome Shotgun Assembly approach (WGSA) (Fleischmann et al., 1995; Venter et al., 2001) and a hybrid approach called pooled genomic indexing which uses a combination of the HSA and WGSA (Gibbs et al. (2004); Milosavljevic et al. (2005)). The HSA and WGSA approach have both been used to sequence and assemble a genome sequence for Humans (Lander et al., 2001; Venter et al., 2001), while the hybrid approach has been used to assemble the genome of the brown Norway rat (Gibbs et al., 2004). The two main methods initially used to sequence the human genome from the International Human Genome Sequencing Consortium (IHGC) (Lander et al., 2001) and Celera (Venter et al., 2001) are illustrated in Figure 1.10.

Hierarchical shotgun assembly was the method applied to sequence the genome of a number of humans by the International Human Genome Sequencing Consortium (IHGC) (Lander et al., 2001). This method is illustrated in the left panel of Figure 1.10. The HSA approach allows the correct assembly of high copy number repeats, which can pose problems in the assembly process in genomes with high numbers of near identical copies. Shotgun sequencing of Bacterial Artificial Chromosomes (BACs) (length ∼200 kb) is one way to reduce the size of the repeat problem in complex genomes as this method will only sample a small region of the genome. As a result it will therefore reduce the number of reads that have to be assembled, making the computational problem of assembly with many interspersed repeats that are highly similar a lot simpler as they are more likely to be in different clones. This method though usually requires a high-resolution genetic map and low-resolution physical map. The major drawback to this method is the cost of genome sequencing and the time taken.

The method used by Celera for sequencing of the human genome attempted to resolve the repeat problem by double-barrelled whole-genome shotgun (WGS) sequencing (Venter et al., 2001) (see Figure 1.10 right panel). This method uses paired reads that are obtained from both ends of clone inserts of various sizes. Paired reads can help to resolve repeats by spanning across them. When paired reads are taken from plasmid
Figure 1.10: A comparison of the two main methods used to sequence the human genome, the HSA splits the genome into a tiling path of overlapping BAC clones, and shotgun sequencing is performed on each BAC which are then reassembled and then merged with adjacent clones. WGS involves performing shotgun sequencing of the whole genome which must then be anchored to the genome. Adapted from Waterston et al. (2002).

inserts longer than 5,000 bp they can for example resolve full-length LINE elements as these mate pair reads can be used by assembly algorithms to infer that mate pairs originate from the same genomic region.

Despite attempts to achieve a correct assembly of the human genome, sequence errors and mis-assemblies occurred in the original draft sequences of both human genome assemblies. In 2004 it was reported that finished human BAC sequences contained a single base pair error every 74 kb with mis-assemblies occurring every 2.6 Mb (Schmutz et al., 2004). In WGS assemblies repeats often contribute to errors because DNA reads originating from distinct copies of the repeat appear identical to the assembler program. This problem is confounded for near-identical repeats, as it is difficult to distinguish a sequencing error from a genuine polymorphism (SNP) between repeat copies. This can sometimes be avoided by identifying reads whose mate pairs
are anchored in unique genomic sequence. The incorrect placement of repetitive reads may lead to mis-assembly that is not a true reflection of the sequence at that region of the genome (Phillippy et al., 2008).

Mis-assemblies in a genome sequence can be caused by repeat collapse or repeat expansion (Phillippy et al., 2008). Repeat collapses are caused when the assembly program incorrectly joins reads originating from distinct repeat copies into a single pseudo-repeat unit. In a repeat expansion extra copies of a repeat are included in the assembly. The consequences of these mis-assemblies often result in a greater (or sometimes lesser) density of reads than is expected from the random shotgun process (Phillippy et al., 2008). A missing repeat copy increases read density in the remaining repeat copies, whereas in repeat expansion mis-assembly, the read density drops below normal (Phillippy et al., 2008). With these issues in mind care needs to be taken when considering whether a sequence is truly polymorphic or has been has just been excluded from or included in the genome due to the nature of the assembly process.

It is clear from the discussion above that high copy number, high identity repeats represent significant challenges for genome sequencing projects. As such the impact of these issues on informatic analyses of L1 sequences needs to be borne in mind when collating sources of L1 sequence data.

Existing sources of FL-L1 sequences

One of the simple ways to view LINE elements graphically is to use a genome browser such as the UCSC genome browser (University of Santa Cruz California http://genome.ucsc.edu), Ensembl genome browser (http://www.ensembl.org) or NCBI MapView genome browser (http://www.ncbi.nlm.nih.gov/projects/mapview/) using the appropriate repeat annotation track. This allows a visualization of the repeat of interest in a genomic context. However L1 elements found in human genome assemblies only represent a fraction of the true level of L1 diversity present in human populations as these elements are likely to be the ones that occur at high allele frequencies, due to the way the genome assemblies are constructed. The genome browser views allow one to search for a sequence in the genome (using BLAST (Altschul et al., 1990) or BLAT
(Kent, 2002)), and this may be helpful in identifying a novel polymorphism discovered from laboratory work (if it is not in the assembled sequence), but this method of L1 detection does not provide any information on insertion frequency or polymorphism status.

In response to the under-representation of L1 repeats in genome assemblies two types of databases have evolved. The first type are sequence databases or archives that contain sequence data that may or may not be present in genome assemblies *i.e.* sequence trace archives, and nucleotide accession databases, such as GenBank. The second type (the more specialist type of databases), only provide information about certain types of sequence and tend to add extra layers of annotation that are not always present in genome browsers or in sequence databases. These two types of databases, that contain L1 sequence data, are discussed below.

**Trace archives**

The trace archive at NCBI is a rapidly growing database comprising of over 1.8 billion sequencing trace files from 4400 different organisms see Figure 1.11 (Wheeler et al., 2008b).

Since 2001, re-sequencing of human genomes has continued to provide vast quantities of genomic data. The human specific trace archives illustrated in Figure 1.11 show rapid expansion in 2005 when the trace sequences from the Celera genome sequencing project were made available and also in 2007 when the diploid human (Levy et al., 2007) genome sequence of a single individual was published. The human trace archive as of April 2008 contained more than 210 million traces submitted by many different sequencing centers across the world. This resource provides an excellent source of sequence data that is not found in the various human genome assemblies, and may yield many L1 polymorphisms if mined with appropriate tools. Novel L1 insertions not found in genome assemblies as well as the empty sites of L1s that are represented can, in principle be discovered by mining this resource.
Figure 1.11: The trace archive statistics showing the growth of trace sequence files from 2001 available at NCBI. The black line indicates all the traces available from all species. The green line indicates human trace sequence file growth.
CHAPTER 1. LITERATURE REVIEW

Short-read archives (SRA)

The development of new sequencing platforms, which include the Roche-454 life sciences system (http://www.454.com), and the Illumina Solexa system (http://www.illumina.com), has lead to the creation of a central repository for short-read sequencing data, at NCBI called the Short-Read Archive (SRA) Wheeler et al. (2008b)(http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi). The resource’s aim is to try to meet the bioinformatic challenges concerning the large amount of data becoming available (Wheeler et al., 2008b). The sequence read lengths produced by these new technologies are a lot shorter (250bp for 454, 35bp for Illumina Solexa) than the Sanger method (600-900 bp) but sequence depth is much higher and so the quantity of data is vast. The SRA, as of April 2008, holds data from 44 different studies.

The first individual human genome to be sequenced using one of the new technologies was the diploid genome of James D. Watson, which was sequenced to 7.4-fold coverage using the Roche-454 life sciences system (Wheeler et al., 2008a). The raw sequence files are available at ftp://ftp.ncbi.nih.gov/pub/TraceDB/Personal_Genomics/Watson. These short-read sequences from the new sequencing technologies may assist in-silico genotyping experiments of known FL L1 elements and also could potentially be used to find new polymorphic FL L1 elements, but the tools for this analysis do not currently exist.

Nucleotide accession databases

At NCBI, in addition to the Trace and Short Read archives, there are many other sources of genomic sequence data. These resources can be divided into CoreNucleotide, Expressed Sequence Tags (ESTs) and Genome Survey Sequence (GSS) sequences (Wheeler et al., 2008b). Within the CoreNucleotide section there are three remaining source divisions. GenBank, NCBI Reference Sequences (RefSeq) and Whole Genome Shotgun Sequences (WGS) (Wheeler et al., 2008b). The GenBank division contains publicly available nucleotide sequences from 260,000 named organisms, obtained from submissions from laboratories throughout the world and batch submissions from large
scale genome projects (Benson et al., 2008). Around 12% (12% of 99.1Gb = 11.9Gb) of the sequences in the GenBank division are of human origin making this a valuable resource for searching for L1 sequences.

**dbSNP**

One particular genetic variation database that contains repetitive element sequence data is dbSNP – the NCBI database of genetic variation (Sherry et al., 2001). dbSNP mainly focuses on Single Nucleotide Polymorphisms (SNPs) and contains variation information for 44 organisms. The most common form of human genetic variation are SNPs and dbSNP comprises over 99.77% SNPs (Sherry et al., 2001). However, other sequence variations are present in dbSNP, including insertion/deletion polymorphisms (in-del)s, micro-satellite repeats, multi nucleotide polymorphisms (MNPs), named-locus variants (typically retrotransposons such as Alu and LINE elements) and sites which have been tested for variation but none has been detected, indicated by the “NO VARIATION” annotation. Although dbSNP accepts submissions from researchers for Alu and LINE elements no attempt is made to curate the elements further than indicating “LARGEINSERTION” or “LARGEDELETION”. In most cases no allele frequency data are available for these types of repeat variations. As a result, while a significant number of L1 polymorphisms may be deposited in dbSNP their utility is limited.

**L1Base - Annotation of full-length, intact L1 elements**

L1Base (http://l1base.molgen.mpg.de/) is a specialised resource containing putatively active LINE-1 (L1) insertions residing in human and rodent genomes (Penzkofer et al., 2005). The data are divided into insertions with two intact open-reading frames (ORFs), designated FLI-L1s, and those insertions that have intact ORF2 but have disrupted ORF1 designated ORF2-L1s. Also all those L1s that are >6000bp in length are included in the database due to their regulatory potential, and these are designated FLnI-L1s. The database focuses on full-length insertions, as this class of L1 are the most likely to still be capable of retrotransposition.
The L1Base dataset was derived from genome assemblies of the respective species (mouse and human). Data mining for the human section of L1Base was performed by a genomic BLAST search of the human genome assembly (*Homo sapiens* NCBI build 35) with a known active L1, L1.2 (M80343) as query, and analysis of repeat masker annotation provided from Ensembl ([http://www.ensembl.org](http://www.ensembl.org)). After the L1s are detected, the regions containing the L1s of interest are extracted from the genome assembly and L1Xplorer (see 5.2) is used to check the intactness of the ORFs and the features known to be important for L1 activity.

The L1Base dataset (derived via a genome assembly mining approach) can be queried using MySQL regular expressions and boolean queries. The L1s can be viewed in a graphical context with each sequence feature being displayed in different colours. L1Base user queries can be exported and L1Base data can be viewed on a genome browser via a DAS (Distributed Annotation Service) at Ensembl ([http://www.ensembl.org](http://www.ensembl.org)).

**dbRIP**

dbRIP is a resource that collates retrotransposon insertion polymorphism (RIP) data ([http://dbrip.brocku.ca](http://dbrip.brocku.ca)) ([Wang et al., 2006](http://dbrip.brocku.ca)). The resource is a non-redundant database totalling 2095 RIPS, comprising of 1625 *Alu* elements, 407 L1 elements, and 63 SVA elements.

dbRIP’s aim is to capture human-specific genomic variation derived from young retrotransposon subfamilies such as L1 Ta-1, Ta-1nd, by collating all the known polymorphic insertions from experimental and computational sources.

Experimentally discovered polymorphic L1s were collected from all published papers containing L1 data ([Wang et al., 2006](http://dbrip.brocku.ca)) and were compiled in XML (Extensible Markup Language). The data collected for each RIP consist of the original published or assigned identifier, the type of retrotransposon, its family and subfamily classification, any disease associations, the DNA sequence of the RIP, and its Target Site Duplications (TSDs). In addition 400bp regions of DNA flanking either side of the RIP are listed.
along with PCR primers, experimental PCR conditions, expected PCR sizes for the filled site and empty site and the chromosome position of the insertion relative to NCBI build 35 of the human genome. Genotype data for each RIP are reported as three genotypes: homozygous filled (+/+), where only the filled site of the RIP was observed, heterozygous (+/-) where the filled site and the empty site of the RIP was observed, and homozygous empty (-/-) where the RIP was not present in that population.

dbRIP displays the collated RIP data alongside other genome annotation information in a custom UCSC genome browser that allows users to view RIP data in a genomic context. The database offers the ability to search the RIP data by two methods, described as quick and advanced. The “quick” search enables the user to search dbRIP by a RIP identifier (a unique identifier assigned by dbRIP), or the original identifier, and chromosomal position of the insertion. These RIP features can also be searched through a custom UCSC genome browser. The “advanced” search offers a greater range of RIP features to search, including the chromosome the RIP is located on and how the RIP is represented in a gene context, to annotate those RIPs located in introns, exons and promoters.

This search functionality enables the user to find RIPs with particular properties and detailed information is represented on an individual RIP record page.

**Line fusion genes**

LINE FUSION GENES is a database developed for detecting LINE expression in known human genes (Kim *et al.*, 2006). The database compiles the structure and expression patterns of LINE elements along with their relative position in genes, and also provides annotation on the tissue distribution of gene expression and chromosomal location of the particular genes affected.

This resource classifies LINE FUSION GENES into three categories, Type 1-alternative promoter, Type 2-alternative polyadenylation signal and Type 3-exonization. This classification is based on the effect of the LINE insertion in genes.
LINE FUSION GENES of Type 1 involve an insertion of the LINE element near the 5’ UTR of a gene or first intron. Type 1 genes may be transcribed from the L1 promoter rather than the promoter of the gene.

Type 2 LINE FUSION GENES (LFGs) occur when the LINE element has a polyadenylation signal within the 3’ UTR of the gene which could potentially lead to a transduction event. The transcript from the gene promoter is captured by the LINE polyadenylation signal rather than one present within the gene. Transcription can also be halted when the LINE polyadenylation signal is incorporated into an intron 5’ of the 3’ UTR.

Type 3 LFGs occur when LINE element sequences are recognised as splice sites (AG-CT) or as intact exons by the spliceosome, meaning the LINE sequence could become fused to mRNA coding sequences, a process known as exonisation. In-silico analysis of the human genome reveals that 1329 genes may be affected by LINE elements during expression (Kim et al., 2006).

LINE FUSION GENES (as of 28/01/2008) is inaccessible (http://www.primate.or.kr/line/).

1.9 Software tools for L1 analysis

Currently there are a number of L1 specific software analysis tools that exist in the public domain. As the development of L1 analysis tools is an aim of this thesis these tools will be briefly described in the following section.

TSDfinder

TSDfinder is a computer program written in perl to find TSDs, polyA tails and inversion breakpoints associated with L1 sequences (Szak et al., 2002). This program uses the RepeatMasker program, and a custom L1 library to identify the beginnings and ends of
L1s in genomic sequence. Once the L1 boundaries have been located, polyA tails that meet a minimum length criteria downstream of the 3’ UTR are detected.

For the detection of TSDs, 100 bp upstream of the 5’ L1 boundary and 3 kb downstream of the 3’ L1 boundary, as defined by RepeatMasker, are used as search criteria to find matching TSDs at either end of the L1. The 3’ search sequence is longer to allow for 3’ transduction events. The first 15 bp of L1 at either end of the L1 boundaries are included to allow for imprecision in the RepeatMasker call. The program bl2seq (Tatusova and Madden, 1999) is used to locate matching TSDs. High-scoring pairs (HSPs) identified by bl2seq are then scored based on the HSP position in the 5’ flank, 3’ flank and a HSP quality score. The scoring algorithm gives higher scores to HSPs that are closer to the 5’ end of the L1, with those within 10 bp scoring highest.

The 3’ HSP position score is weighted to give higher scores to HSPs that are closer to the 3’ end of the L1, therefore discriminating against 3’ transduction events. The final score is a combination of the 5’ HSP score, 3’ HSP score and the HSP quality score.


**L1Xplorer**

A set of annotation tools developed by Penzkofer et al. (2005), collectively called L1Xplorer, was developed to assist the annotation of L1s with features that have been shown to be important for L1 activity. These tools are a collection of Perl scripts that detect sequence features, such as conserved amino acid residues and transcription-factor binding sites. The annotation tools also provide a system for automated L1 subfamily classification by analyzing diagnostic sequences in the 3’ UTR and 5’ UTR, as well as additional external annotation information such as coding gene annotation, known SNPs (from dbSNP) and RepeatMasker annotation (Penzkofer et al., 2005).

L1Xplorer can be run on genome assemblies or on individual L1 sequences and provides a graphical map of the annotation relative to the L1 query sequence. L1Xplorer
is part of L1Base (Penzkofer et al., 2005), which is an annotation database of Full-Length, Intact L1 Elements. L1Xplorer can be found at http://l1base.molgen.mpg.de/
Chapter 2

Materials and methods

2.1 Introduction

The methods used from third party software and those developed specifically for this thesis for the detection and annotation of FL-L1HS as well as those used in the construction of a FL-L1HS database resource are explained below.

2.2 Hardware

The initial development of the computational annotation tools and database were performed on an Intel Pentium 4 2.4 GHz PC with 1 GB of RAM running SUSE 10.0 and a Apple Macbook 2 GHz Intel core 2 Duo with 2 GB RAM running OS 10.4 (“Tiger”). The development work was moved to a Linux Bio-cluster (see Figure 2.1) hosted by University of Leicester IT Services. The majority of testing and analysis for the final datasets was performed on the “workhorse” node (viti.gene.le.ac.uk) under a specific project directory baseline (/srv/data/projects/baseline). The Bio-cluster runs the Linux operating system Debian 4.0.
CHAPTER 2. MATERIALS AND METHODS

69

CHAPTER 2. MATERIALS AND METHODS

69

Figure 2.1: Hardware specifications of the University of Leicester Bio-cluster

The baseLINE database and website is currently hosted on viti.gene.le.ac.uk and can be accessed at the Uniform Resource Locator (URL) http://baseline.gene.le.ac.uk.

2.3 Software

Below is a list of software programs used with a brief description, version details and URL, where appropriate.

Practical Extraction and Report Language (PERL)

Perl is a high-level programming language developed by Larry Wall in 1991. Perl provides the computer programmer with powerful text processing facilities allowing the manipulation of practically any type of text file, of any length. This makes Perl the ideal language for database access, software tools, graphical programming and World
Wide Web (WWW) programming. When referring to the language, “Perl” is used, and when referring to the implementation of the language “perl” is used. Perl was used in this thesis to develop custom computer programs to parse and manipulate large data files and develop algorithms to detect important biological characteristics such as Target Site Duplications (TSDs).

Version : v5.8.8

**Comprehensive Perl Archive Network (CPAN)**

CPAN is an archive of thousands of software modules and documentation written in the programming language Perl. CPAN is located at [http://www.cpan.org/](http://www.cpan.org/). These modules house commonly used pieces of code that can be used by various Perl programs. Please see the baseline script and module folder on the appendix CD for details of which modules were used. Detailed documentation for each module can be found on CPAN. Typical installation of a Perl module from CPAN on a UNIX file system is as follows

```
sudo perl -MCpan -e shellinstall 'module of choice'
```

Cpan script version 1.03
CPAN.pm version 1.9205

**Bioperl**

Bioperl is a collection of perl modules used in common bioinformatic applications, such as parsing output files from BLAST searches, performing alignments using ClustalW, parsing GenBank and SwissProt files, and providing a perl interface to bioinformatic resources, such as EMBOSS ([http://emboss.sourceforge.net/index.html](http://emboss.sourceforge.net/index.html)). The Bioperl project is flexible enough to support enterprise-level applications such as Ensembl, while maintaining an easy learning curve for beginners in the Perl programming language ([Stajich et al., 2002](#)).
Bioperl 1.5.1 was downloaded from http://www.bioperl.org/ and installed following the instructions available at http://www.bioperl.org/wiki/Installing_Bioperl_for_Unix. In the baseline code repository any module or script that uses a module with a prefix `Bio::'module name'` originates from the bioperl project. A good example of usage in the baseline code of bioperl is in the BLAST parsing module (Baseline::Parse::Blast.pm in the baseline code repository). This module uses the `Bio::SearchIO` module from bioperl which provides an efficient way of parsing multiple blast files. `Bio::SearchIO` is a driver for parsing sequence database searches (http://doc.bioperl.org/releases/bioperl-1.0/Bio/SearchIO.html). All documentation from the bioperl modules used in the baseline code can be found using perldoc (after bioperl is installed)

```
perldoc Bio::SearchIO or from CPAN (http://www.cpan.org/).
```

Version: 1.5.1

**Common Gateway Interface (CGI)**

Common Gateway Interface (CGI) is a specification for transferring information between a World Wide Web (WWW) server, and a CGI program. A CGI program may be written in any programming language, although for this thesis these programs were written in Perl. The main purpose of a CGI program is to accept and return data that conforms to the CGI specification. As the HTML pages that are generated by the template toolkit for TSDmapper and baseline explorer include forms, a user can submit sequence data, or search criteria through these forms. The CGI programs (i.e. perl scripts) for each of these web resources then process the data submitted by the user and return the results as HTML pages. Please see `from.pl` and `explore.pl` for the perl code using CGI available of the appendix CD.
CHAPTER 2. MATERIALS AND METHODS

CGI-FormBuilder

CGI-FormBuilder (http://formbuilder.org/) is an external Perl module that is used by the TSDmapper web application. This module is designed to handle all parts of a CGI form request - generation, submission, and validation. This module allows the use of a single script that, when invoked, controls the application based on CGI parameters. FormBuilder is an object-orientated Perl module and fits into the baseline perl module structure. This module allows the easy generation of a web form with minimal code. CGI-FormBuilder is available from CPAN and the author’s website.

Version: 3.0501

MySQL

MySQL is a Relational Database Management System (RDMS) (http://www.mysql.com) that provides a fast, reliable and easy to use open-source solution to storing data in separate but related tables rather than storing data in a single monolithic table. Tables containing data are linked by a set of defined relationships, which makes it possible to access, and combine data from a number of tables upon a user’s request. Data can be added, accessed and processed using the Structured Query Language (SQL) to enable the reliable construction of queries and execution of database updates.

Version: 5.0.51a-3-log (Debian)

Perl DataBase Interface (DBI)

Perl DBI (DataBase Interface) module is the most common database interface for the Perl programming language. DBI is the interface between the application and one or more Data BaseDrivers (DBDs) (Figure 2.2). DBDs exist for different types of databases, such as MySQL and Oracle. The DBI and DBD Perl packages allow Perl programmers to access many database environments in a uniform way.
CHAPTER 2. MATERIALS AND METHODS

Perl Script

MySQL Database

Oracle Database

DBI interface

DBD::Oracle

DBD::MySQL

Figure 2.2: The DBI interface dispatches the method calls from the script to the appropriate driver for execution of the SQL statements on the appropriate database.

The DBI and the DBD Perl modules for MySQL used in the analysis scripts were installed from CPAN. The code written for database connections used in this thesis can be found in the Baseline::Database.pm module, in the Baseline script repository.

DBI Version : 1.607
DBD Version : 4.007

Template Toolkit (TT2)

The Template Toolkit (TT2) is a fast and flexible template processing system that is written in the Perl programming language. TT2 allows a template processor to output changeable data surrounded by some invariant data. TT2 supports command-line, modular, CGI and mod_perl operation. This makes the TT2 an ideal templating system for a static and dynamic website generation and includes many pre-written modules to make this process simple. TT2 allows variable substitution, loops and conditionals as well as supporting complex data types such as arrays, hashes, objects and subroutine references. This allows the “back-end” of a website to be built in Perl. In TT2 the “logic” i.e. the perl scripts that bring data back and forth from
Figure 2.3: DAS client issuing a request in the form of a URL, the DAS server responding to the request and sending annotation back to the DAS Client by HTTP, in the form of XML. Adapted from Cerami (2005).

the database or other application to a user on a web-page is kept separate from the website site code (HTML). For example a website’s appearance or “skin” can be changed without affecting the scripts controlling the programming logic. TT2 is free and open-source. It is available from CPAN. Information on the TT2 can be found at http://template-toolkit.org/index.html.

Version: 2.0

**DAS**

The Distributed Annotation System (DAS) is an XML-based communication protocol used to exchange biological annotations, typically to enable the sharing of genome annotation data. DAS allows annotations to be spread over multiple databases instead of all being held in one central resource. DAS is specified by a client/server protocol and a small set of XML queries, with an associated set of XML Documents Type Definitions (DTDs). DAS client applications connect to DAS servers, send queries in the form of URLs and receive XML data. A DAS server can act as a reference server (i.e. to hold a reference map for a particular genome), an annotation server (containing the actual genomic annotation data) or both and issues responses to requests from the DAS client in the form of XML. Transport of these data is provided by HTTP (see Figure 2.3).

An example of a DAS client is the Ensembl genome browser (http://www.ensembl.org/). Examples of DAS Servers include ProServer and Lightweight DAS Server (LDAS). When using a DAS client, a user can request data from a number of different DAS servers to build up custom genome annotation views (see Figure 2.4).
ProServer

ProServer is a standalone, lightweight DAS server, written in Perl and designed to have low system requirements. It is simple to install and configure and has existing adaptors for a wide variety of data sources which makes it ideal for large or small bioinformatics projects. It is also easily extensible allowing adaptors to be written for other data sources. The DAS server can be configured for each resource provided by the server using a ‘.ini’-format configuration file. ProServer has been set up on the server viti following the instructions included with the source code. The configuration ‘.ini’ for baseline was set up in the general.ini file specifying the source adapter to use (Baseline.pm). A source adapter has been modified from the examples to serve baseLINE FL-data from the database to the Ensembl genome browser. The source adapter is called Baseline.pm. The source adapter builds the features for display via DAS by collecting the chromosomal coordinate start and stop positions of the L1 feature on a particular chromosome using a MySQL query to obtain the data from the database (see Baseline.pm for more details). The feature that is constructed also contains the type of feature (i.e. L1) and an id, which is the unique identifier assigned to that L1 based on its location within a particular source. ProServer can be started with the following command . /proserver -x -c general.ini ProServer information is available from Wellcome Trust Sanger Institute (WTSI) from
http://www.sanger.ac.uk/proserver/. The source code is available for download from Subversion using the following command.

```
svn co https://proserver.svn.sf.net/svnroot/proserver/trunkBio-Das-ProServer
```

Examples of baseLINE DAS usage are discussed in more detail in chapter 5.

**BLAST-like Alignment Tool (BLAT)**

The BLAST-like Alignment Tool was developed by Kent (2002) for fast mRNA, DNA and protein alignments. BLAT is reported to be 500 times quicker than BLAST for mRNA/DNA alignments and 50 times quicker for protein alignments. The program rapidly locates short matches and then extends these matches into high-scoring pairs. BLAT differs from BLAST by building an index of the database to be searched rather than building a index of the query sequence as BLAST does. BLAT triggers extensions on any number of near-perfect or perfect hits, whereas BLAST only triggers extensions when one or two hits occur in proximity. BLAT can cope with introns in RNA/DNA alignments.

The online version of BLAT at the University California Santa Cruz (UCSC) can be accessed at the following web address http://genome.ucsc.edu/. For local use the BLAT suite v. 34 was downloaded from http://www.soe.ucsc.edu/~kent/src/ and installed following the instructions included with the source files. Within the BLAT suite there are two programs, gfServer and gfClient that enable large scale usage of BLAT. These programs are particularly useful for mapping large numbers of sequence files to a BLAT-indexed database in the 2bit format.

**Searching BLAT sequence databases**

gfServer was set up to hold the human genome assembly (hg18.2bit) in the computer’s memory, allowing extremely rapid searching. A gfServer was set up on viti with the following commands.
CHAPTER 2. MATERIALS AND METHODS

gfServer start localhost 8888 hg18.2bit

Using the program gfClient, the sequence files can be then be mapped to the genome assembly using the following commands.

gfClient localhost 8888 'path to human genome' 'sequence files to be mapped' -out='BLAT output file'

However the searching of BLAT databases was performed via a script gfClient.pl to automate this process for many sequences.

Version: v. 34

BLAST

BLAST (Altschul et al., 1990) is a sequence-comparison algorithm that is used to search protein and nucleotide sequence databases for alignments to a query sequence. The blast archive suite can be downloaded from NCBI FTP site ftp.ncbi.nlm.nih.gov/blast/executables/. The blast program suite contains a number of utilities that allow sequence comparison on a local computer. The main programs used in this thesis are search utilities (blastall and bl2seq) and sequence manipulation utilities (fastacmd, formatdb). blastall is a program that, given sequences in FASTA format, will search and find similar sequences in a BLAST database. Basic command line usage is shown below. There are five BLAST programs that are available to use as a search. The ones used in the analysis are described below. blastn compares a nucleotide sequence against a nucleotide sequence database. bl2seq is a program that given two sequences in FASTA format will find regions of local similarity. Typically bl2seq is used to align two sequences without the need for a BLAST database. formatdb is a sequence manipulation program that creates BLAST databases from FASTA sequences. The fastacmd program retrieves the whole or sub-sections of FASTA sequences from BLAST databases.

Version: 2.2.17
CHAPTER 2. MATERIALS AND METHODS

FASTA2 sequence suite

The FASTA2 suite (http://fasta.bioch.virginia.edu/fasta_www2/fasta_down.shtml) is a collection of computer programs for comparisons of protein and DNA sequences (Pearson and Lipman, 1988). Within the FASTA2 suite is a local alignment program called lalign (Huang and Miller, 1991) that compares two sequences looking for local sequence similarities. lalign reports a specified number of alignments (the default is 10) between the two sequences and shows the actual local alignments between the two sequences and their scores. The lalign programs accepts standard sequence formats (e.g. FASTA). The FASTA 2 suite source code was downloaded (http://fasta.bioch.virginia.edu/fasta_www2/fasta_down.shtml) and built according to the installation instructions included with the download.

RepeatMasker

RepeatMasker (http://www.repeatmasker.org) (Smit et al., 1996-2004) screens DNA sequences for low-complexity and interspersed repeats. The program produces annotation detailing the nature of the repeat, giving the repeat class, family and positions of the repeats in the query sequence and where the repeat matches in the Repbase libraries of repeats. The query sequence is also masked by Ns to denote the repeat or repeats in the sequence. For this thesis RepeatMasker was set up to use the cross_match algorithm for sequence comparisons against the RepeatMasker libraries, downloaded from Repbase.

Version: open-3.2.1
CHAPTER 2. MATERIALS AND METHODS

Repbase libraries

Repbase is a database of repetitive elements (http://www.girinst.org/). The RepeatMasker versions were downloaded from the Repbase website and configured to run with RepeatMasker as per the installation instructions that accompany RepeatMasker.

Version: 20071204

cross_match

The program cross_match (http://www.phrap.org/) is a general-purpose utility developed for comparing two sets of DNA sequences. cross_match and SWAT are part of the Phrap package developed by Phil Green at the University of Washington. cross_match was obtained from Phil Green as a uuencoded .tar.Z file, and built using the installation instructions.

Version: 0.990329

BLAST database construction

Previous studies that have detected FL-L1HS sequences have used assembled genome sequences. The human genome assembly resources were downloaded from the NCBI website (ftp.ncbi.nih.gov). All sequence sources were downloaded from NCBI in April 2008 by File Transfer Protocol (FTP) onto the server viti and placed into the baseLINE directory. The data, if not already in a BLAST database format, were built into BLAST databases using formatdb and the following command

formatdb -i 'sequence_file.fa' -o T -p F

where -i is the input sequence file, -o is set to true (T) so sequence identifiers can be queried against this database and sequences extracted, and -p specifies the type of file F (nucleotide).

Celera: Celera Genomics whole genome shotgun assembly (AADB), November 2001. This assembly included both WGS and BAC sequence data.

HuRef: J Craig Venter Institute whole genome shotgun assembly (May 2007) (see http://www.jcvi.org/research/huref/). This assembly represents a composite haploid version of the diploid genome sequence from a single individual (J. Craig Venter).

The pre-formatted BLAST databases for nucleotide (nt April 2008) accession searches were downloaded from NCBI and so do not have to be formatted.

**Trace archive construction**

To make the trace archive sequences useful they first needed to be formatted into a database that could be searched using BLAST.

The trace archive databases used in this analysis were constructed using methods developed by Gudmundur Thorisson and Owen Lancaster of the Department of Genetics, University of Leicester.

The Fasta sequences and clipping information files were downloaded from the Ensembl Traceserver (http://trace.ensembl.org/). The quality clipping information contained the start and stop where each accession sequence should be clipped to removed poor quality sequence.

The process of creating the trace archives involved 3 main steps, detailed below:

1) Create a file from the sequence clipping information so the data can be loaded into a MySQL database using the perl script `traces_clipinfordb.pl`.

2) Use the perl script `traces_qualclip.pl` to clip the sequence files and write them to a new fasta-file.
3) Format clipped sequences as a BLAST database using `formatdb` which splits the database into manageable chunks of some default size.

These BLAST trace archive databases were then searched with the `blast.pl` script.

**BLAST database searches**

The highly active Human Specific L1 (L1-HS) L1.3 (Accession: L19088) was used as a query sequence for BLAST searches against the formatted databases. The polyA tail of 38 nt was removed as it is known that polyA tails can be highly variable between L1 sequences. The perl script `blast.pl` was used for the detection of FL-L1HS in the BLAST databases. All hits that matched the criteria of 5922 bp in length and ≥98% identity to L1.3 were selected from the various resources. All final BLAST searches were performed on sequences that were downloaded in April 2008, and were part of the NCBI build 36.3 release.

**GBrowse Karyotype viewer**

The GBrowse Karyotype viewer available from [http://gmod.org/wiki/Gbrowse](http://gmod.org/wiki/Gbrowse) enables a karyotype to be drawn in a web page. This was done so that the FL-L1HS matches detected in these analyses could be shown graphically in a genomic context.

The cytoband data for the human genome was obtained from the Ensembl database from Ensembl’s public ftp server at [ftp://ensembldb.ensembl.org](ftp://ensembldb.ensembl.org) and loaded into the ideogram provided with GBrowse karyotype, using the information from [http://gmod.org/wiki/GBrowse_karyotype_ideogram.pl](http://gmod.org/wiki/GBrowse_karyotype_ideogram.pl).

**Data for GC content XY plot**

Underneath each chromosome karyotype on the view web page of the baseLINE web site is a XY plot of percentage GC content in 1Mb bins. These data were mined from
the Redon et al. (2006) supplementary data and populated in a MySQL table. These data are then used to plot the graphs for each chromosome.

**Data for gene content XY plot**

Underneath each chromosome karyotype on the view web page of the baseLINE web site is a XY plot of gene density in 1 Mb bins. The RefSeq genes were downloaded from the UCSC genome table browser [http://genome.ucsc.edu/cgi-bin/hgTables](http://genome.ucsc.edu/cgi-bin/hgTables) from the March 2006 assembly (NCBI build 36.3, hg18). The script in the bioperl distribution ([generate_histogram.pl](http://search.cpan.org/dist/Getopt-Euclid/lib/Getopt/Euclid.pm)) was used to calculate the gene density in a 1 Mb bin which was then loaded into an MySQL database. These data are then used to plot the gene density graph for each chromosome.

**Perl code**

The Perl scripts written for L1 detection and annotation reported in this thesis were designed in a modular fashion to allow the maximal re-use of Perl code that performs a particular function. The Perl modules are self-contained pieces of code that can be used by different Perl scripts or other Perl modules to perform the desired task.

For more information on the usage instructions of each of the perl scripts contained on the appendix CD the following command can be used by a user to provide help documentation as well as detailing the required and optional input parameters (if any) providing Perl is installed.

```
perl 'path to perl folder/script name.pl' --help
```

Each perl script uses the perl module Getopt::Euclid which is available from CPAN ([http://search.cpan.org/dist/Getopt-Euclid/lib/Getopt/Euclid.pm](http://search.cpan.org/dist/Getopt-Euclid/lib/Getopt/Euclid.pm)). This module uses the perl script’s own documentation to create a command-line argument parser. This
ensures that the perl script’s documented interface and its actual interface always agree. The module names used for each Perl module are unique. Components of module’s names are separated by double colons (::). Each module is contained in a single file with the extension ‘.pm’ and stored in a hierarchical subdirectory that parallels the module’s hierarchy. This is illustrated in table baseline hierarchy (Table 2.1)

<table>
<thead>
<tr>
<th>Module name</th>
<th>Stored in</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline::Util</td>
<td>Baseline/Util.pm</td>
</tr>
<tr>
<td>Baseline::Database</td>
<td>Baseline/Database.pm</td>
</tr>
<tr>
<td>Baseline::Extract::Sequence</td>
<td>Baseline/Extract/Sequence.pm</td>
</tr>
<tr>
<td>Baseline::Parse::Blast</td>
<td>Baseline/Parse/Blast.pm</td>
</tr>
<tr>
<td>Baseline::Run::SeqExtract</td>
<td>Baseline/Run/SeqExtract.pm</td>
</tr>
<tr>
<td>Baseline::Run::SeqSearch</td>
<td>Baseline/Run/SeqSearch.pm</td>
</tr>
<tr>
<td>Baseline::TSDMapper</td>
<td>Baseline/TSDmapper/TSDmapper.pm</td>
</tr>
</tbody>
</table>

The general functions of the modules are described in the following sections.

**Baseline::Util**

The Perl module Baseline::Util contains the general Perl functions that are needed by all other module. These include common functions such as opening files for reading and writing data.

**Baseline::Database**

The Perl module Baseline::Database contains all the Perl code needed to make a connection to a MySQL database. To connect to the database the perl DBI module is used. The module also contains a number of methods for querying the database using SQL statements and returning the data to the desired script or subroutine for further manipulation.
Baseline::Parse::Blast

The Perl module Baseline::Parse::Blast contains a number of methods for parsing the different styles of BLAST output. The main method in this module, parse_blast, parses the output from several different blast output files, including tabular and pair-wise alignment versions. This subroutine can also parse BLAT output that is produced in the NCBI BLAST output format, or NCBI BLAST tabular format with or without a header. This parse_blast method uses the bioperl module Bio::SearchIO for the parsing of the BLAST files. The subroutine, when called from the appropriate script (e.g., blast.pl), takes the user’s desired parsing parameters (percentage identity to the query, the length of the hit) and returns all results that match these criteria. The parsed results are placed into an array of hashes which is returned to the script that called the subroutine. The script can then call any of the Baseline::Parse::Blast ‘get’ methods. For example, to get the hit start position in the alignment between the hit and the query sequence this method can be called from the perl script.

```perl
my $hit_hit_start = $importer->get_hit_start($record);
```

This returns the start position of the hit in the alignment. This allows many scripts to be written than can access the Baseline::Parse::Blast module and get only the desired results back. For example one particular script may need the hit query nucleotide string where as others do not, so only the results specifically needed by that particular perl script are returned.

Baseline::Run::SeqSearch

The Perl module Baseline::Run::SeqSearch contains all the methods needed to perform a sequence search. The blast run method in the module is called blast and when this is called from a script (e.g., blast.pl) or from another module (e.g., Baseline::Extract::Sequence) the method runs the blastall program with the parameters that are passed to it. The method returns the BLAST output file to the calling script or module for further processing (e.g., so the results can be parsed), or
if the BLAST was unsuccessful it will return an error. A similar method exists for the program BLAT called blat.

Baseline::Run::SeqExtract

The Perl module SeqExtract contains methods for calling the fastacmd BLAST suite program for extracting sequences from BLAST databases.

Baseline::Run::TSDmapper

The Baseline::Run::TSDmapper module is described in more detail in chapter 4. This module contains all the subroutines specific to the TSDmapper script, TSDmapper.pl. The module contains methods for sequence validation, polyA tail finding, endonuclease extraction, endonuclease scoring, distance scoring, transduction finding, transduction extraction, and empty site reconstruction.

Baseline::Extract::Sequence

The Baseline::Extract::Sequence module contains all the methods needed to extract flanking DNA, L1 sequence, and L1 and flanking DNA. This module is used by the extractseq.pl script.

Baseline perl scripts

Descriptions of each of the main perl scripts used in the analysis during this thesis are outlined below. All perl scripts are located in the scripts directory in the baseline folder on the appendix CD. Detailed information on the inputs the scripts accept can be found by running the scripts with the --help flag.
chapter 2. materials and methods

blast.pl

The blast.pl script performs a BLAST search with a user defined sequence against any BLAST formatted database and parses the results with a user-defined set of parameters for length of hit and percentage identity.

blast_parser.pl

The blast_parser.pl script is a parser that returns all hits from a BLAST output file that meet a user’s desired criteria of length and percentage identity. This script is used to refine BLAST results without having to re-run the BLAST search which would take additional time.

TSDmapper.pl

The TSDmapper.pl script is a perl script developed to detect TSDs in flanking DNA surrounding the 5’ and 3’ ends of an L1 element. The script requires the flanking DNA to be 5’ → 3’ and to have some L1 sequence following the 5’ flanking sequence and preceding the 3’ flanking sequence.

gfClient.pl

The gfClient.pl script performs a ‘blat’ search using the gfClient program and parses the results based on a user’s desired criteria for sequence length and percentage identity to the query sequence.

extractseq.pl

The extractseq.pl extracts from a BLAST database a whole sequence, a region of sequence, 5’ and 3’ flanking DNA (mainly for TSDmapper analysis), L1 DNA and L1 and 8-kb sequence.
Chapter 2. Materials and Methods

Gene Finder.pl

The script gene_finder.pl takes the position of the FL-L1HS that has been mapped to the genome and looks in the RefSeq gene co-ordinates to see if it falls between the transcriptional start and stop of a gene.

Tax Finder.pl

The tax_finder.pl script looks up the accessions from the nucleotide accession (nt) BLAST database searches and reports the taxonomy of the accession sequence. The taxonomy database was downloaded from the FTP site at NCBI the same time as the nt database (April 2008).

BaseLINE Website

The computer scripts that control the BaseLINE website are a set of CGI-Perl scripts. This section discusses the BaseLINE website. For screenshots and examples of the results that this code generates please see Chapter 5.

The main code for the BaseLINE website is contained within a folder called ttbook. In this folder there is a cgi-bin folder, which contains all the CGI-Perl scripts. The scripts that control the website use CGI to accept and return data to and from HTML pages that are generated by the Template ToolKit 2 (TT2). Each of the main sections on the BaseLINE website are explained below.

Header and Footer Templates

The header and footer templates are located in the directory templates/src. The header file (header.tt2) contains the paths to the css file and contains javascript that may be needed on each page. The footer file contains (footer.tt2) links to the “about”, “conditions of use”, and “disclaimer” pages.
These two files are processed by all other templates when generating the HTML to give a consistent look and feel to the website.

Search template

The search template (search.tt2) contains javascript to control the search options. This template prints all the dynamic values into HTML that the CGI perl script search.pl generates. All search results are returned to this page.

L1view template

The view template (view.tt2) contains all the HTML-generating code to display each L1 record with all annotation features (see chapter 5 for more details).

Explorer template

The explorer (exploreL1s.tt2) template contains all the HTML generating code to display the GBrowse karyotype program. This program displays the genomic position of the FL-L1HS on the Reference assembly. This template generates all of the options that lets users customise the look and feel of the website.

TSDmapper template

There are two TSDmapper templates, (form.tt2 and tsdmapper_results.tt2) described below.

form.tt2 generates the submission form using the Form::Builder perl module. The tsdmapper_results.tt2 template generates the HTML for the results generated by the TSDmapper script. The TSDmapper script is slightly modified version of the script discussed in chapter 4 as it draws a TSD graphic of TSD annotation features using the Bio::Graphics module that is part of the bioperl distribution.
Chapter 3

Full-length L1 sequence discovery

3.1 Introduction

The aim of this thesis was to develop informatic resources for LINE-1 retrotransposons. A decision was made to focus on developing these resources for young Full-Length (Ta) L1s. As outlined in the introduction, these are the family of elements in the human genome that are still capable of retrotransposition. They also have the potential to be polymorphic with respect to presence or absence between individuals (i.e. di-morphic (Sheen et al., 2000)) and so are still shaping the human genome. Current resources such as dbRIP (Wang et al., 2006), L1base (Penzkofer et al., 2005) and a spreadsheet constructed by Dr Richard Badge (Badge et al., 2003) only typically consider one resource (either experimental or computational) when representing L1s. Since 2001 the availability of human genomic sequences has increased dramatically, with the release into the public domain of the Celera human genome assembly (Venter et al., 2001), re-sequencing efforts that have increasingly populated the human trace archives, as well as advances in sequencing technologies that have made it cost efficient to sequence the genomes of individuals. As of 2008 the assembled genome sequences of two individual males, the first “personal genomes” of Dr Craig J. Venter (Levy et al., 2007) and Dr James D. Watson (Wheeler et al., 2008a) are available in the public domain. Thus with the wealth of human genomic sequence data available now an opportunity exists
to discover in a more detailed manner than previously possible, the contribution of FL-L1HS to the human genome and to represent this contribution in an accessible database.

To achieve the goal of building an accessible database describing FL-L1HS elements, these sequences first needed to be collected. This chapter focuses on the methods developed in detecting FL-L1HS sequences in a number of different sequence resources as well as assessing the contribution that FL-LHS elements make to the human genome. The detection stage of the project concentrated on the following approaches:

1) The identification and building of locally accessible human genome sequence databases from a number of different data sources

2) The searching of these databases and the identification of FL-L1HS sequences that met a particular set of criteria

3) The identification and resolution of allelic variants, exact sequence duplications and polymorphisms

4) Annotation of a key set of structural characteristics for each L1

5) Analysis of the FL-L1HS sequences in their genomic context

3.2 Searching of resources containing FL-L1HS sequences

This section discusses the searching of a number of genomic resources for the detection of FL-L1HS sequences.

Query sequence and parameters

The sequence data resources identified as containing FL-L1HS sequences were downloaded from the NCBI FTP site (ftp.ncbi.nih.gov/) and built into BLAST
databases (see chapter 2 for details). The three main assemblies searched were: the human genome reference assembly (Lander et al., 2001) from the international Human Genome Sequencing Consortium (IHGC) (referred to from this point as the Reference assembly), Celera human genome assembly (referred to hereafter as Celera) (Venter et al., 2001) and the diploid human genome assembly, (referred to as HuRef) (Levy et al., 2007). The nucleotide accession database (nt) in the form of a preformatted BLAST database (April 2008) was also downloaded.

To identify FL-L1HS sequences, a previously identified highly active L1HS, L1.3 (Accession L19088.1) was chosen as the query sequence for BLAST (Altschul et al., 1990) searches of the built sequence databases. The 37 bp polyA tail, located at 3′ prime end of the L1.3 element sequence was removed so that L1 hits would not missed due to variation in the length of this simple-sequence tract. This version of L1.3 was designated L1.3_mpa (minus poly A tail).

All L1s that met the criteria of ≥5922bp in length, and showing ≥98% identity to the L1.3_mpa sequence are classed as FL-L1HS sequences. These selection criteria were used because they were effective in previous successful in-silico experiments that used BLAST to identify FL-L1HS sequences that are polymorphic and capable of retrotransposition (i.e. active) by Badge et al. (2003) and Brouha et al. (2003).

The number of sequence resources available meant it was impractical to run each BLAST search on each resource manually using the command line options available from the BLAST suite. An automated search and parse script was therefore developed called blast.pl (see the directory baseline for perl code on the appendix CD). This script implements the sequence similarity search algorithm BLAST, using the program blastn designed to search for and identify nucleotide sequences. The script blast.pl searches one or more specified databases with the L1.3_mpa query and a set of user-defined BLAST parameters. The script also parses the results from the BLAST output file and returns a set of results based on the user’s desired input parameters (i.e. the selection criteria outlined above). Information on the parameters that the blast.pl script accepts can be found by running the script with the --help flag.
Hit inclusion parameters

Searching the sequence databases with blast.pl required the flanking DNA sequence size in the script to be set 1000 bp either side of the 5’ L1 start and the 3’ end stop as defined by BLAST. Hits that fell within these boundaries were given a unique identifier (see below) and placed into the results file. 1000 bp of DNA was chosen as the default flanking DNA length as this gave the maximum number of elements with enough flanking DNA for unique mapping to the human genome Reference assembly, while not excluding many hits that were close to the boundaries of an accession.

Those hits that did not have enough flanking DNA were excluded from the hit result file and were put aside for manual annotation. Figure 3.1 illustrates the inclusion and exclusion criteria based on flanking DNA positions within the accession.

For searches against sequence databases where the organism code was known at the point of download (e.g. Human genome assemblies such as Reference, Celera, HuRef), the taxonomy option was set to the default of false (F). However for the nucleotide accession database (nt) the organism code for each FL-L1HS element could not be determined at the point of download due to the fact that this database contains sequences from multiple species. Therefore the taxonomy database was queried with the hit accession and the corresponding species code was retrieved with the results. Those hits not classified as Homo sapiens were placed into a separate results file for later analysis.

L1 naming convention

Each FL-L1HS hit was assigned a unique identifier based on the accession in which it was located. Since there is the possibility that multiple FL-L1HS sequences meeting the selection criteria might be located in one accession, each hit was assigned a unique identifier for that source based on the position within that accession. For example, the first hit found in an accession that is closest to the start nucleotide of the accession is assigned a, the next b, etc.. If the number of hits in an accession exceeds 26 then those
hits would be assigned aa, ab, ac, etc. The naming convention for multiple hits in accessions is illustrated in Figure 3.2.

3.3 Number of FL-L1HS sequences in different human genome assemblies

The script `blast.pl` was run on all the three human genome assemblies with the parameters described above. The hits from the BLAST searches that met the FL-L1HS criteria (≥5922 bp in length and ≥98% identity to L1.3.mpa) are shown in
Figure 3.2: Naming convention for multiple hits in an accession: The first hit in the accession is defined by the BLAST hit start and stop positions in that accession. The first hit is named a, the next b and so on. The blue arrows indicate the 5’ and 3’ flanking DNA.

The reference assembly contained the most FL-L1HS sequences, with a total over all chromosomes of 462. The Celera assembly contained the second highest number of FL-L1HS sequences with 353 and the HuRef assembly contained 239 FL-L1HS sequences. The Reference assembly contains 109 more FL-L1HS sequences than the Celera assembly and 223 more than the HuRef assembly. No FL-L1HS sequences were excluded from the final counts of each assemblies due to insufficient flanking DNA, as these were assembled sequences and none of the FL-L1HS sequences were located less than 1 kb from a sequence gaps or regions not represented in assemblies (e.g. telomeres and centromeres).

Figure 3.3 illustrates that larger chromosomes (1, 2, 3, 4, 5, 6, 7, 8) in all assemblies have a higher number of FL-L1HS than smaller chromosomes (13, 14, 15, 16, 17, 20, 21, 22, Y) which might be expected. However there are exceptions. Chromosome 18, in all assemblies, seems to contain an elevated number of FL-L1HS (Reference: Total 18, size 76 Mb, Celera: Total 15, size 75 Mb, HuRef: Total 12, size 75 Mb) compared to Chromosome 17 which is of similar size (Reference: Total 4, size 79 Mb, Celera: Total 2, size 78 Mb, HuRef: Total 2, size 77 Mb). In the HuRef assembly there appears to be a very much reduced number of FL-L1HS on the X Chromosome (HuRef: 1) compared to the Reference and Celera assemblies (Reference: 38, Celera: 30). In the HuRef assembly no FL-L1HS were detected on chromosome 19 or on the Y chromosome.
3.4 Distribution of FL-L1HS

To investigate whether the numbers of FL-L1HS sequences in each assembly are distributed according to chromosome size, the density of FL-L1HS per Megabase haploid genome was calculated in each of the three assemblies. To do this the total genome size was calculated to the nearest Mb. The reference haploid genome was calculated to be the largest (3079 Mb) followed by the Celera haploid genome (2837 Mb) and the smallest, the HuRef assembly (2810 Mb). The density of FL-L1HS per haploid genome ($L_1D$) was calculated using the following formulae:

**L1 density per Haploid genome ($L_1D$)**

$L_1D = \frac{\text{total number of FL-L1HS sequences per haploid genome}}{\text{haploid genome size (Mb)}}$

Table 3.1 shows that the Reference genome has the highest FL-L1HS density per Mb, followed by the Celera genome and finally the HuRef. It is expected therefore that one
Table 3.1: The density of FL-L1HS sequences per Mb in the Reference, Celera and HuRef assemblies

<table>
<thead>
<tr>
<th>FL-L1HS element density per Mb for three genome assemblies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>Reference</td>
</tr>
<tr>
<td>Celera</td>
</tr>
<tr>
<td>HuRef</td>
</tr>
</tbody>
</table>

would find a FL-L1HS every 0.15 Mb in the Reference genome, 0.124 Mb in the Celera genome and every 0.085 Mb in the HuRef genome if FL-L1HS were distributed evenly.

Using the FL-L1HS element density per haploid genome, the expected number of FL-L1HS sequences was calculated based on the size of each chromosome, for each of the three assemblies.

**Reference Assembly**

The Figure 3.4 illustrates the observed number of FL-L1HS sequences from the reference assembly vs the expected number of FL-L1HS based on the reference genome FL-L1HS density of 0.15 FL-L1HS sequences per MB.

Figure 3.4 shows that, in the Reference Assembly, on chromosomes 1, 2, 3, 4, 5, 6, 8, 12, 18, and X the number of observed FL-L1HS hits was more than the expected number. However on chromosomes 7, 9, 10, 11, 13, 14, 15, 16, 17 19, 20, 21, 22 and Y the observed numbers of FL-L1HS were lower than expected.

**Celera Assembly**

Figure 3.5 shows that in the Celera assembly chromosomes 1, 2, 3, 4, 5, 6, 12, 18, X and Y the number of observed FL-L1HS sequences was more than the expected number for a chromosome of that size. However on chromosomes 7, 8, 9, 10, 11, 12, 13, 15, 16, 17,
19, 20, 21, 22 the observed numbers of FL-L1HS sequences are lower than expected for the size of the chromosome.

**HuRef Assembly**

Figure 3.6 shows that in the HuRef assembly chromosomes 1, 2, 3, 4, 5, 6, 10, 14, 18, the number of observed FL-L1HS sequences was more than the expected number for a chromosome of that size. However on chromosomes 7, 8, 9, 11, 12, 13, 15, 16, 17, 19, 20, 21, 22, X and Y the observed numbers of FL-L1HS sequences are lower than expected for the size of the chromosome. The number of FL-L1HS sequences on the X chromosome in the HuRef assembly (1) is considerably less than on the X chromosome in the reference assembly (38) and the X chromosome in the Celera assembly (30).
Figure 3.5: Observed vs expected number of FL-L1HS sequences in the Celera genome assembly.

Figure 3.6: Observed vs expected number of FL-L1HS sequences in the HuRef genome assembly.
Are FL-L1HS sequences distributed across the genome evenly?

To determine if FL-L1HS sequences were distributed evenly across the genome according to chromosome size, a χ² test for each of the three assemblies was performed.

Null Hypotheses

The null hypothesis for each assembly was that FL-L1HS sequences are distributed proportionately according to the size of the chromosome, such that larger chromosomes will have more FL-L1HS sequences than smaller chromosomes.

The χ² values for each assembly can be seen in Table 3.2. The number of degrees of freedom (df) is 23 (based on 24-1 for the number of chromosomes). For 23 df P=0.05 (35.17) and P=0.01 (41.64).

The values for χ² generated from the reference assembly (55.69) and the HuRef assembly exceed the critical value at both P=0.05 and P=0.01 while the χ² value for the Celera assembly (30.24) does not exceed the critical value P=0.05 and P=0.01. Therefore we can reject the null hypothesis that L1s are distributed according to the size of the chromosome in the Reference and HuRef assemblies. FL-L1HS sequences are not distributed evenly and so FL-L1HS chromosome density is not explained by the size of the chromosome for the Reference and HuRef assemblies. The null hypothesis is accepted for the Celera assembly, suggesting that in the Celera assembly FL-L1HS sequences are distributed according to the size of the chromosome.

Table 3.2: χ² values for the human Reference, Celera and HuRef assemblies

<table>
<thead>
<tr>
<th>Assembly</th>
<th>χ²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>55.69</td>
</tr>
<tr>
<td>Celera</td>
<td>30.24</td>
</tr>
<tr>
<td>HuRef</td>
<td>46.25</td>
</tr>
</tbody>
</table>
Relationship between chromosome size and the number of FL-L1HS sequences

To determine if the number of FL-L1HS sequences found on each chromosome is correlated with the size of the chromosome a Spearman Rank Correlation Coefficient analysis was performed for each of the three assemblies.

Null hypothesis for Spearman Rank Correlation

Null hypothesis states there is no relationship between the size of a chromosome and the number of FL-L1HS on each chromosome. i.e. an increase in chromosome size will have no effect on the number of FL-L1HS sequences detected in it.

Table 3.3: Spearman Rank Correlation values for the human Reference, Celera and HuRef assemblies.

<table>
<thead>
<tr>
<th>Assembly</th>
<th>$R^2$</th>
<th>Degrees of Freedom (df)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>0.9258</td>
<td>22</td>
<td>6.52×10^{-14}</td>
</tr>
<tr>
<td>Celera</td>
<td>0.9304</td>
<td>22</td>
<td>3.22×10^{-14}</td>
</tr>
<tr>
<td>HuRef</td>
<td>0.8109</td>
<td>22</td>
<td>2.04×10^{-14}</td>
</tr>
</tbody>
</table>

The chromosome size and the number of FL-L1HS sequences for each assembly were both ranked from the highest to the lowest. The values for $R^2$ for each assembly and the P values are shown in Table 3.3. These P values indicate that for each of three assemblies the null hypothesis that the chromosome size will have no effect on the number of FL-L1HS sequences found on each chromosome can be rejected ($P<0.01$). Therefore larger chromosomes have larger numbers of FL-L1HS sequences. The size of the chromosome is therefore positively correlated with the number of FL-L1HS sequences in all three assemblies.
Figure 3.7: Number of FL-L1HS sequences plotted against chromosome size for the Reference assembly.

Figure 3.8: Number of FL-L1HS sequences plotted against chromosome size for the Celera assembly.

To illustrate the Spearman Rank Correlation analysis, the number of FL-L1HS sequences vs the size of the chromosome for each the three assemblies was plotted (see Figure 3.7, Figure 3.8, and Figure 3.9). These figures show that as the size of the chromosome increases the number of FL-L1HS sequences found on the chromosome increases. However there are anomalies which are discussed in the following section.
Figure 3.9: Number of FL-L1HS sequences plotted against chromosome size for the HuRef assembly.

**Anomalies for number of FL-L1HS sequences Vs chromosome size**

Although the data suggest that in general the number of FL-L1HS sequences found on a chromosome increases with chromosome size as we might expect, there do appear to be a few exceptions to this explanation. One such example is chromosome 18. In all three of the human genome assemblies tested, this chromosome has more FL-L1HS sequences than expected for its size. Chromosome 18 is one of the smaller chromosomes in each of the three assemblies, yet it has more FL-L1HS sequences than chromosome 17, which is of a similar size. The number of FL-L1HS sequences for each of the three assemblies for chromosomes 17 and 18 differs on the reference assembly by a ratio of 1:4.5, Celera assembly 1:7.5 and on the HuRef assembly 1:6 (see Table 3.4).

Another anomaly in the data is the X chromosome. Table 3.5 illustrates the number of FL-L1HS sequences on the X chromosome compared to the similarly sized chromosome 7. Table 3.5 indicates that there is a distinct lack of FL-L1HS sequences present in the HuRef assembly on the X chromosome. The number of FL-L1HS sequences discovered on the X chromosome in the Reference and the Celera assemblies suggest a small enrichment for FL-L1HS sequences when compared to a chromosome of a similar size.
Table 3.4: Proportions of FL-L1HS sequences in different genome assemblies for similar sized chromosomes 17, and 18

<table>
<thead>
<tr>
<th>Assembly</th>
<th>Chr 17 (MB)</th>
<th>Chr 18 (MB)</th>
<th>Chr 17 $N^o$ FL-L1HS</th>
<th>Chr 18 $N^o$ FL-L1HS</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>79</td>
<td>76</td>
<td>4</td>
<td>18</td>
<td>1:4.5</td>
</tr>
<tr>
<td>Celera</td>
<td>78</td>
<td>75</td>
<td>2</td>
<td>15</td>
<td>1:7.5</td>
</tr>
<tr>
<td>HuRef</td>
<td>77</td>
<td>75</td>
<td>2</td>
<td>12</td>
<td>1:6</td>
</tr>
</tbody>
</table>

Table 3.5: Proportions of FL-L1HS sequences in different genome assemblies for similar sized chromosomes 7, and X

<table>
<thead>
<tr>
<th>Assembly</th>
<th>Chr 7 (MB)</th>
<th>Chr X (MB)</th>
<th>Chr 7 $N^o$ FL-L1HS</th>
<th>Chr X $N^o$ FL-L1HS</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>159</td>
<td>155</td>
<td>22</td>
<td>38</td>
<td>1:1.73</td>
</tr>
<tr>
<td>Celera</td>
<td>154</td>
<td>155</td>
<td>19</td>
<td>30</td>
<td>1:1.58</td>
</tr>
<tr>
<td>HuRef</td>
<td>153</td>
<td>144</td>
<td>12</td>
<td>1</td>
<td>1:0.08</td>
</tr>
</tbody>
</table>

3.5 FL-L1HS sequences from other sources

FL-L1HS sequences detected in human genome assemblies do not represent the true level of FL-L1HS sequence variation that is present in human populations, since the assemblies are generated from a small number of individuals (one in the case of HuRef). FL-L1HS sequences can be polymorphic with respect to presence or absence, and therefore only FL-L1HS sequences that are present at high allele frequencies are likely to be included in published assemblies.

The number of FL-L1HS sequences detected from the three assemblies varies sufficiently to indicate that there is FL-L1HS variation between individuals, which has been demonstrated experimentally by L1 insertion genotyping. In order to extend this “in-silico” genotyping approach other sources of genomic sequence data that may have contained FL-L1HS not present in human genome assemblies were downloaded and searched.
Figure 3.10: Distribution of FL-L1HS sequences detected from a search of the April 2008 nt BLAST database with L1.3_mpa

**Searching the nucleotide accession database**

The nucleotide accession database (nt) provides a set of redundant assembled nucleotide sequences from GenBank, EMBL, and DDBJ sequence repositories from different species (21800 according to the taxonomy database at NCBI). In April 2008 the nt pre-formatted database was downloaded from the NCBI FTP site (ftp.ncbi.nih.gov). Using the blast.pl Perl script, the nt blast database was searched for FL-L1HS sequences using a query sequence of L1.3_mpa, with the same parameters as the assembly searches using the taxonomy parameter so that the taxonomy report was also generated.

The number of hits detected from this source that met the criteria for FL-L1HS sequences (≥5922 bp; ≥98% identity to L1.3_mpa) was 765. There were 685 hits classed as human when searched against the taxonomy database, while 80 were classified as non-human. Figure 3.10 shows the four categories into which the 765 hits were split into.

Twenty eight human FL-L1HS sequences (3.7%) were removed from the automated analysis, as they did not have enough flanking DNA (a minimum 1 kb at the 5′ and 3′ end of the L1) to be accurately mapped to a unique genomic location on the Reference assembly.
Non-human FL-L1HS hits from nt database

Figure 3.11: Species distribution of non-human FL-L1HS detected from the April 2008 nt BLAST database

Figure 3.11 shows the species distribution of BLAST hits detected in various non-human species. Surprisingly 75 full length L1s that met the search criteria were detected in *Pan troglodytes*, 3 from *Pan paniscus* and 2 from *Gorilla gorilla*. These non-human L1s were not used in all further analyses.

3.6 Resolution of duplicated FL-L1HS sequences from nt database

The 657 FL-L1HS sequences were mapped to the reference assembly to determine their genomic location and to see if a particular FL-L1HS sequence had already been detected in the assembly searches. The FL-L1HS sequences, as defined by the BLAST start and stop coordinates, were extracted from the relevant nucleotide accession using the script `extracteq.pl` (see chapter 2), along with 1 kb of genomic flanking sequence 5’ of the start of the L1 and 1 kb 3’ of the L1s 3’ end. A local instance of a BLAT server using the programs `gfServer` and `gfClient` and the hg18.2bit BLAT database (downloaded from the UCSC FTP site) was set up. A perl script called `gfClient.pl` searched each of the 8-kb sequences against the hg18.2bit BLAT database and reported all hits that were ≥98% identical and ≥7500 bp in length. Hits that did not map to the reference assembly with these parameters, were then searched
against the hg18.2bit BLAT database using 2-kb of 5′ flanking sequence and 1-kb of FL-L1HS sequence. This was done to find potential empty-sites in the reference assembly. Those hits that mapped ≥98% and ≥1.5 kb and ≤2 kb were classified as having an empty site (i.e. there was no FL-L1HS sequence at that location in the genome), those hits that mapped with ≥98% identity and with >2 kb and ≤3 kb sequence were manually examined to determine if there was really a FL-L1HS sequence present in the assembly or if the FL-L1HS sequence was present in a segmental duplication. A FL-L1HS sequence was considered to be at the same location as in the reference assembly if it mapped to within 100 bp of the hit within the assembly.

The resolution of duplicates (i.e. multiple accessions containing the same FL-L1HS sequence) and manual curation of the dataset resulted in 504 unique FL-L1HS sequences being mapped to unique locations from the nt accession database. Figure 3.12 illustrates the chromosomal distribution of these hits. It is possible to see that within this dataset, chromosome 18 is enriched with FL-L1HS (22 FL-L1HS sequences), as observed for the assembly searches. One hundred and five duplicates detected within the nt dataset and 50 FL-L1HS sequences were not able to be mapped due to lack of flanking DNA, or the flanking DNA being too repeat rich to satisfy the criteria. These 50 FL-L1HS sequences (<10% of the dataset) were not further analysed as extensive manual curation would be required.

3.7 Resolution of duplicated hits between sources

The BLAST searches of the four main databases (Reference, Celera, HuRef and nt) are likely to have identified the same FL-L1HS sequence in several instances. Using the reference assembly as a coordinate system on which to uniquely locate sequences, FL-L1HS sequences from the two other human genome assemblies and the nt accession database were mapped to the assembly as described above. A non-redundant set of 533 FL-L1HS sequences from the four data sources was produced. The overlap from the three assemblies is illustrated in Figure 3.13. Figure 3.13 illustrates that the reference assembly contains 89 FL-L1HS sequences that are not found in any of the other two assemblies (A). The Celera assembly however only has 2 FL-L1HS sequences that are
Figure 3.12: Chromosomal distribution of FL-L1HS sequences found in April 2008 nt accession database.

not found in any other assembly (B). The HuRef assembly has 10 FL-L1HS sequences that are not found in the other two assemblies (C). The Reference assembly has the greatest overlap with the other two assemblies with 342 FL-L1HS sequences shared between the Reference and the Celera assemblies, and 223 being shared between the Reference and HuRef assemblies. The number of FL-L1HS sequences that are common between the three assemblies is 192.

Not all FL-L1HS sequences from the Celera and HuRef assemblies could be successfully mapped to the Reference assembly. Only 345/353 (97.7%) of FL-L1HS sequences detected in the Celera assembly were mapped to the Reference assembly successfully. Eight (3.3%) of the FL-L1HS sequences were unable to be mapped due to a number of reasons. These reasons include: repeat variation at one end of the flanking DNA, or that the L1 fell below the required cut-off due to SNPs between the L1 sequences, or that the L1 identified as a FL-L1HS sequence in the Celera search was not a FL-L1HS sequence in the reference assembly, but in fact an older L1PA2 element, which did not meet the criteria of \( \geq 98\% \) and 5922 bp in length. From the HuRef
assembly 234/239 (97.9%) were mapped successfully to the Reference assembly, with 5 (2.1%) FL-L1HS sequences not successfully mapped due to SNP variation within the FL-L1HS sequence and the L1 in the reference assembly.

The overlap from the three assemblies and the nt accession database is illustrated in Figure 3.14. The identification of FL-L1HS sequences from four different sources has identified 188 which are the same across the four data sources, 6 that are only present in the reference, 1 present only in the Celera assembly, 9 present only in the HuRef and
58 present only in nt database. The comparison of the four sources shows that there is polymorphism of FL-L1HS sequences between sources. Polymorphism of FL-L1HS sequences is further investigated in chapter 4 through in-silico reconstruction of empty sites for all 533 FL-L1HS sequences.

3.8 FL-L1HS sequence contribution to the human genome

Using the data collected on FL-L1HS sequences from the genome assemblies it was possible to calculate an estimate of the contribution that FL-L1HS sequences make to the human genome sequence. Using a haploid human genome size of 3079 Mb (based on the Reference assembly size), and the average size of human FL-L1HS sequence (0.006 Mb,) the total size of all 533 FL-L1HS sequences in the human genome is 3.198 Mb. Therefore FL-L1HS sequences contributes to \(~0.1\%\) of the haploid human genome size, excluding any 5’ or 3’ transductions that the FL-L1HS may carry.

However this estimate assumes that all individuals carry all 533 elements, which is unlikely since at least 125 FL-L1HS sequences are polymorphic to some degree (Badge et al., 2003; Brouha et al., 2003; Wang et al., 2006). If the numbers of polymorphic FL-L1HS sequence are taken into consideration, the minimum number of FL-L1HS sequences that a human individual is expected to have is 408 (533-125). Using this minimum L1 genotype the total size contribution to the genome from FL-L1HS sequences would be 2.448 Mb or \(~0.08\%\) of the haploid genome size. However this genotype is also highly unlikely as this assumes an individual will have none of the polymorphic FL-L1HS sequences. If we consider the numbers of FL-L1HS sequences identified in the three assemblies analysed (Reference: 462, Celera: 352, HuRef: 234) as independent samples of FL-L1HS composition (i.e. no assembly contains sequences derived from the same individual) the expected number of FL-L1HS sequences is likely to fall within the range of these values. However the number of individual chromosomes represented at any given locus is undefined for the Reference and Celera assemblies, and is 1 or 2 for the HuRef assembly (1 on the X and Y, 2 on the autosomes). Using
Figure 3.14: Overlap of FL-L1HS sequences between the three human genome assemblies: Reference, Celera, and HuRef and the nt accession database. Based on the online 4 way Venn diagram generator available at [http://www.pangloss.com/seidel/Protocols/venn.cgi](http://www.pangloss.com/seidel/Protocols/venn.cgi).
the total number of FL-L1HS sequences from each source (462, 352, 234) one measure of the expected number of FL-L1HS sequences in an individual is the median value of these samples, which is 352 (2.12 Mb, \(\sim 0.07\%\)). However this analysis only takes into account FL-L1HS sequences from assembled sequences, and not any FL-L1HS sequence that may only be represented in the nt accession database or other sequence databases. Although polymorphism status (i.e. fixed or polymorphic) is known for many elements (through genotyping studies) and can be inferred by in-silico analysis it will only be possible to calculate a plausible estimate of the number of FL-L1HS sequences in an “average” human genome with accurate estimates of allele frequencies for all elements. At the time of writing this is unfortunately not possible.

3.9 In-silico screening for polymorphic FL-L1HS sequences in the nt database

Initially the nt database was searched for the reconstructed empty-sites of identified FL-L1HS sequences, since it was felt that this database would contain the best quality assembled sequences available. The empty site search was performed using the blast.pl script on the nt database and each FL-L1HS reconstructed empty-site was used as the query sequence. For a FL-L1HS sequence to be classed as having a 5' empty-site, the identified sequence was required to be \(\geq 98\%\) identical to the reconstructed empty site and \(\geq 300\) bp in length. The length of 300 bp was chosen to remove all hits that would match either the 5' or 3' flanking DNA of a filled site FL-L1HS sequence. Of the 533 reconstructed empty sites analysed parsing of the BLAST output returned 206 results of which 117 were unique FL-L1HS empty site hits. Therefore using the nt April 2008 database to find polymorphic FL-L1HS sequences indicates that from the total of 533 FL-L1HS sequences identified during this investigation 117 (\(\sim 22\%\)) are potentially polymorphic. However since the nt database comprises DNA sequences from different species, the local identifier of each hit (i.e. the accession) was queried against the taxonomy database (April 2008) to identify from which organism the hit had come. Since the original sequences contained human specific FL-L1HS sequences and were identified as coming from human DNA, there
is the possibility that the empty site (i.e. flanking DNA with no FL-L1HS) may be found in humans that do not have the FL-L1HS sequence and also other species such as chimpanzee (Pan troglodytes) where there would be no human-specific L1.

Analysis of the 117 hits found that 55.6% (65 hits) of the empty sites only mapped to one known accession, whereas 44.4% (52 hits) of the empty sites mapped to more than one accession. Table 3.6 illustrates the empty site hits that have mapped to only one accession and have had their species of origin identified using the taxonomy database (April 2008) using the perl script tax_finder.pl. Of the 65 FL-L1HS empty sites that map to one accession three different species have been identified, Gorilla gorilla (Western Gorilla), Pan troglodytes (Chimpanzee) and Homo sapiens (Human). As expected the majority of FL-L1HS empty sites map to human accessions with 46 out of the 65 hits (70.8%) belonging to this category, with 17 (26.2%) found to map to accessions reported to be from chimpanzee DNA and 2 (3.1%) from Gorilla DNA.

Analysis of the 52 FL-L1HS empty sites that map to more than one accession in the nt database is discussed below. Table 3.7 illustrates the number of FL-L1HS empty sites that map to just one species (either Homo sapiens or Pan troglodytes) or both. Twenty seven FL-L1HS empty sites mapped to human accessions only, 10 mapped to Pan troglodytes and 0 mapped to Gorilla gorilla. Fifteen FL-L1HS empty sites mapped to at least one human accession and one accession identified to be from Pan troglodytes.

Table 3.6: Species classification of the accessions FL-L1HS empty sites map to in the nt April 2008 database

<table>
<thead>
<tr>
<th>FL-L1HS empty site clasification</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gorilla gorilla</td>
<td>2</td>
</tr>
<tr>
<td>Pan troglodytes</td>
<td>17</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>46</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>65</strong></td>
</tr>
</tbody>
</table>

The species distribution of all 117 empty sites is illustrated in Table 3.8. Seventy three sequences were identified as originating from human DNA sequences only with a further 15 empty site hits mapping to human and chimpanzee DNA. This indicates at
Table 3.7: Species classification of the FL-L1HS empty sites that map to the nt April 2008 database in multiple accessions (≥2)

<table>
<thead>
<tr>
<th>FL-L1HS empty site classification</th>
<th>Species Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens</td>
<td>27</td>
</tr>
<tr>
<td>Homo sapiens &amp; Pan troglodytes</td>
<td>15</td>
</tr>
<tr>
<td>Pan troglodytes</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>52</td>
</tr>
</tbody>
</table>

Table 3.8: Species classification of the FL-L1HS empty sites that map to the nt April 2008 database

<table>
<thead>
<tr>
<th>FL-L1HS empty site classification</th>
<th>Species Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens</td>
<td>73</td>
</tr>
<tr>
<td>Homo sapiens &amp; Pan troglodytes</td>
<td>15</td>
</tr>
<tr>
<td>Pan troglodytes</td>
<td>27</td>
</tr>
<tr>
<td>Gorilla gorilla</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>117</td>
</tr>
</tbody>
</table>

least 88 FL-L1HS sequences (75.2% from the 117 empty site hits or 14% of the total number of FL-L1HS sequences tested (522)) can be considered to be polymorphic from computational analysis of the nt April 2008 database.

Twenty seven FL-L1HS empty sites where identified as originating from Pan troglodytes DNA sequences confirming these FL-L1HS sequences are indeed human-specific, with a further 15 hits mapping to both human and chimpanzee DNA. Only two FL-L1HS hits mapped to an accession identified as Gorilla DNA.

3.10 Trace Archives

Identifying FL-L1HS sequences from genome assemblies and nt databases is limited by the nature of the construction of these resources. Genome assemblies such as the
CHAPTER 3. FULL-LENGTH L1 SEQUENCE DISCOVERY

Reference assembly and Celera assembly are made up of a small number of individuals and reflect only a small proportion of the true FL-L1HS variation present in human populations. The HuRef assembly is the genome sequence of one individual (Craig Venter) and only reflects the FL-L1HS variation of one individual’s genome. Thus genome assemblies are likely to only reflect FL-L1HS sequences present at high allele frequencies with FL-L1HS sequences present at lower allele frequencies not making it into the final assembly.

The trace archives present an opportunity to discover new FL-L1HS sequences not present in the genome assemblies. As the trace archive DNA sequences are much smaller than assembled genomic sequences, the approach taken to identify FL-L1HS sequences was modified. Instead of searching for a complete FL-L1HS sequence (~6 kb in length) the first 250 bp of the 5’ UTR of the L1.3 mpa was used to search the trace archives using the blast.pl script. The results of this search were parsed using the blast.parser.pl script using the criteria of ≥98% identity to the query sequence and ≥50 bp of L1 sequence, to capture all traces that may have a piece of FL-L1HS contained within them. By only using the 5’ end of the FL-L1HS sequence, it would favour the detection of FL-L1HS sequences in the trace archives.

The analysis yielded 4209 trace files meeting the search criteria. As many of these traces could be FL-L1HS sequences that were already known from assembly and nt database searches, these needed to be excluded from the final count. Also as each FL-L1HS sequence could be detected multiple times by being in different traces (due to shotgun coverage, different sequencing centres sequencing different individuals, or repeated sequencing of regions) the true number of additional FL-L1HS sequences found in the trace archives would have to be further refined.

The refinement of the trace archive search results involved 6 main steps: 1) the extraction of all the trace archive hits from the trace archive database (see chapter 2 for details on construction) using the extractseq.pl script using option 5 (return whole trace file sequence); 2) concatenation of all the trace files into one single file; 3) RepeatMasking of the trace files using RepeatMasker to mask out all of the L1 sequences; 4) “blat” searching of the repeat-masked trace sequence against the human genome blat database hg18.2bit using the perl script gfClient.pl. Using this script
CHAPTER 3. FULL-LENGTH L1 SEQUENCE DISCOVERY

1787 hits that were 50-1000 bp in length and showed 98% identity to the mapped genomic sequence were extracted; 5) comparison of mapped results to the known 533 FL-L1HS sequences start and stop positions using the baseLINE database. This analysis revealed 685 trace files that could potentially be new FL-L1HS sequences; 6) Manual curation of new traces to give an estimate of the number of FL-L1HS sequences in the trace archives that are not in the nt databases. This was done by ensuring that there was no L1 sequence in the assembly. In addition manual curation was used to resolve more than one trace mapping to the same genomic location. The results of this analysis are illustrated in Figure 3.15 (see Appendix Table B.2).

It is possible to see from Figure 3.15 that the trace archives provide an additional 48 potential FL1-L1HS sites that are not found in the assembled sequences. Chromosome 6 has the most FL-L1HS sequences detected that were not found in assembled sequences with eight sites on this chromosome potentially containing a FL-L1HS sequence. The trace analysis also reveals that chromosome 18 which showed an increased number of FL-L1HS sequences for its size (22 FL-L1HS) from the nt and genome assembly analysis contained no potential FL-L1HS sequences. Of 48 potentially novel FL-L1HS sequences that were found from the trace archives 20 trace files (41.7% ) were mapped to a known gene, suggesting that the trace archive could be a good source for finding polymorphic FL-L1HS sequences that have inserted into genes.

An example of a trace file (79800004_abc_a) from chromosome 6 that contains a potentially novel FL-L1HS sequences is illustrated in Figure 3.16. The figure shows that trace file 79800004_abc_a (from Agencourt Bioscience Corporation (abc)) is probably polymorphic as no LINE sequence is reported in the human genome assembly or the baseLINE database. However this region is reported to be a copy number polymorphic (CNP) (cnp 559 182645 bp) region by Redon et al. (2006) indicating that the authors either detected this L1 as a CNP due to its polymorphic nature or that the L1 has inserted into a region of the genome which is copy number variable. Searching the NCBI trace archive (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?TRACE_NAME = “79800004”) and selecting the retrieve info button, reveals the trace information. The individual identifier of this trace is NA19240 and when this is searched (http://ccr.coriell.org/Sections/Search/Search.aspx?PgId=165&amp;q=NA19240) it shows the
DNA is from a Female Yoruban from Nigeria. There was only one trace file that met the criteria outlined above and mapped to this region, suggesting that this L1 could either be a private insertion of the Nigerian individual genotyped or be population-specific to the Yoruba. However this is impossible to confirm computationally and further genotyping experiments would be needed to validate these hypotheses.
Figure 3.16: Genome view of trace file 79800004.abc_a that potentially contains a unique FL-L1HS, unique DNA sequence mapped to chromosome 6.
Also discovered were 4 potential FL-L1HS trace files 1) tjb51a01.b1.wugsc.a, 2) 6S17.2-490a11.q1ka.sc.a, 3) 19866805309593_cra.a, 4) 1099476344387.tcag.jcvijtc.a) that map to the gene ZDHHC14 on chromosome 6, as shown in Figure 3.17. These traces were mapped to the genome assembly and were not found in previous searches (human genome assemblies and nt searches). Since there are four trace files that have come from four different sequencing centres there is the possibility that they could be from different individuals. One trace file, 4) 1099476344387.tcag.jcvijtc.a is from the J. Craig Venter Institute (JCVI) originating from the sequencing of the diploid human genome sequence (HuRef) and is therefore known to come from J. Craig Venter. Trace file 1) tjb51a01.b1.wugsc although reported in the trace information to be from chromosome 7, in fact maps to chromosome 6 with 98.5% over 536 bp and is from the Washington University Genome Sequencing Center (WUGSC). The tracefile 19866805309593_cra.a is from Celera Genomics (CRA), and the trace file 6S17.2-490a11.q1ka.sc.a is from the Sanger Centre (SC). This L1 is probably not a private insertion (i.e. more than one person in the world has this insertion), but is probably polymorphic since it is not in any of the human genome assemblies.

Although this analysis cannot definitely confirm the presence of the whole 6-kb L1, it does provide enough data for researchers to design PCR experiments to a) prove that a FL-L1HS does really exist at this position in the genome and b) if a FL-L1HS is present, perform genotyping experiments on a number of different individuals to determine its polymorphism status and population frequency. It is evident that the trace archives provide useful information on FL-L1HS sequences despite the small size of the DNA sequences (250-1000 bp). All trace file names found in this analysis can be found in the Excel spreadsheet trace_results.xls in the electronic appendix.

3.11 FL-L1HS sequence annotation and analysis

All 533 unique FL-L1HS sequences and their 103 duplicates detected in the searches described previously were analysed using the L1Xplorer annotation tool (Penzkofer et al., 2005). Remote access to the results was provided at the following URL,
Figure 3.17: Genome view of 4 potential FL-L1HS unique sequences mapped to chromosome 6.
CHAPTER 3. FULL-LENGTH L1 SEQUENCE DISCOVERY

http://132.187.22.5/~penkofer/badge. Login details are available on request. The L1Xplorer analysis provided extensive annotation information on the FL-L1HS dataset which was used in the analyses performed in this chapter. The L1Xplorer analysis was reformatted, excluding the TSD, PolyA tail and CpG island analysis. The TSD analysis was discarded due to the analysis employed by L1Xplorer being based on an erroneous implementation of the TSDfinder algorithm (as judged by inspection of computer-predicted and manually identified TSDs). The polyA tail analysis was also unnecessary as the TSD prediction algorithm developed in this thesis includes a polyA tail locating program. The reformatted L1Xplorer results were then stored in the baseLINE database.

Sub-family distribution of FL-L1HS sequences

The sub-family distribution of 533 FL-L1HS sequences was obtained from the baseLINE database using the perl l1_xplorer_analysis.pl. The FL-L1HS sub-family distribution of the 533 FL-L1HS sequences is illustrated in Figure 3.18.

The subfamily distribution of the 533 FL-L1HS sequences in Figure 3.18 illustrates that the largest subfamily detected is Ta1-nd with 382 (∼72%) FL-LHS sequences belonging to this subfamily. The second largest subfamily is Ta1-d with 102 (∼19%) of the FL-L1HS sequences belonging to this subfamily. Forty one (∼8%) FL-L1HS
sequences are classified as non-canonical (i.e., they do not belong to either Ta1-nd or Ta1-d subfamilies) while 8 (~1.5%) FL-L1HS sequences are classed as having multiple subfamily classification, due to the fact that the subfamily classification varies between the same FL-LHS detected in different sources (i.e. Removing the FL-L1HS sequences that cannot be classified (non-canonical, and multiple) the ratio of FL-L1HS sequences that are Ta1-d to Ta1-nd is 1:3.75, which is consistent with the Ta1-d subfamily being the youngest.

**Computational prediction of active FL-L1HS sequences**

Using the information from the L1Xplorer analysis, computational predictions were used to determine if each of the acquired FL-L1HS sequence was potentially capable of retrotransposition. This was done using the perl script `l1_xplorer_analysis.pl`. A FL-L1HS sequence was initially classified as being capable of retrotransposition if it had two intact ORFs. An ORF was defined as intact when it contained no gaps, no frameshifts and no stop codons. Using these criteria 181/533 FL-L1HS (34% of the total FL-L1HS dataset) were deemed to have two intact ORFs and are therefore predicted to be capable of retrotransposition. Therefore 352 FL-L1HS (66%) have at least 1 gap, 1 stop codon or frameshift in either one or both of the ORFs making it unlikely that these elements are capable of retrotransposition.

However, within the 181 FL-L1HS sequences there may still be FL-L1HS that have intact ORFs but which have mutations at functional amino acid residues whose disruption is known to prevent retrotransposition. As a result, an additional analysis of the FL-L1HS sequence annotation using the known conserved functional motifs in ORF1 and ORF2 was performed. The conserved motifs considered were: ORF1 (REKG 235-238, ARR 260-262, and YPSKLS 282-287 (Moran et al., 1996)), and ORF2 (N14, E43, Y115, D145, N147, T192, D205, FADD700 (Feng et al., 1996), R363, I1220 (Lutz et al., 2003) and SSS1096 (Moran et al., 1996)). Requiring these sites to be conserved, combined with ORF intactness but allowing ≤1 in-frame deletions/insertions in either ORF, resulted in 154/533 (29% of total dataset) FL-L1HS sequences predicted to be capable of retrotransposition. The reason that in-
frame deletions and insertions were allowed, (but only if they did not disrupt conserved functional motifs) was that elements assayed by Brouha et al. (2003) contained in-frame insertions/deletions, but were nonetheless active in retrotransposition assays.

Using the 154 FL-L1HS sequences that were identified as potentially active, and the 379 that were considered to be incapable of retrotransposition, the sub-family classification was re-analysed to see how potentially retrotransposition-competent FL-L1HS sequences were distributed between the sub-families.

**Sub-family classification: Intact ORFs vs. non-intact ORFs**

Figure 3.19 illustrates that, when considering only FL-L1HS sequences that are likely to be retrotransposition competent (*i.e.* that have two intact ORFs and all conserved key amino acid residues), the number of elements that belong to the Ta1-d subfamily is 62 (∼40% (62/154) of the intact dataset, ∼61% (62/102) of all Ta1-d detected in database searches), and 79 belong to the Ta1-nd subfamily (∼51% (79/154) of the intact dataset, ∼21% (79/382) of all Ta1-nd detected in database searches).

Forty FL-L1HS sequences were classified as being members of the Ta1-d sub-family and determined to be non-intact (∼10% (40/379) of all non-intact FL-L1HS sequences, ∼39% (40/102) of all Ta1-d detected in database searches). Three hundred and three FL-L1HS sequences were classified as being members of the Ta1-nd sub-family (∼80%
(303/379) of all non-intact FL-L1HS sequences, \(~79\%~(40/382)\) of all Ta1-nd detected in database searches). Those elements that could not be classified as Ta1-d or Ta1-nd were analysed further. The sub-family distribution of non-intact FL-L1HS sequences is illustrated in Figure 3.20.

The observed FL-L1HS subfamily classification ratio in the total dataset of Ta1-d (102) and Ta1-nd (382) elements is 1:3.75. We therefore might expect a similar ratio regardless of whether L1 elements are classed as being intact or not. The following section reports the results of analysis to test this hypothesis.

**FL-L1HS sequences with intact ORFs**

For FL-L1HS sequences with intact ORFs, after removing the FL-L1HS that could not be classified (10 non-canonical, 3 multiple), the expected proportions based on the subfamily classification counts (Ta1-d and Ta1-nd) for the whole data were tested using the exact binomial test for goodness-of-fit. The results of this test are illustrated in Table 3.9. The p-value (two-tailed) for this test is \(1.55 \times 10^{-9}\) meaning the null hypothesis that Ta1-d and Ta1-nd proportions for FL-L1HS sequences that have intact ORFs should be in the proportion of 1:3.75 is rejected.
Table 3.9: Intact ORFs observed vs. expected for exact test for goodness-of-fit

<table>
<thead>
<tr>
<th>L1 sub-family</th>
<th>Intact ORFs observed</th>
<th>Intact ORFs proportions</th>
<th>Intact ORF expected numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ta1-nd</td>
<td>79</td>
<td>3.75</td>
<td>111</td>
</tr>
<tr>
<td>Ta1-d</td>
<td>62</td>
<td>1</td>
<td>30</td>
</tr>
</tbody>
</table>

FL-L1HS sequences with non-intact ORFs

For FL-L1HS sequences with non-intact ORFs, after removing the FL-L1HS sequences that cannot be classified (31 non-canonical, 5 multiple), the expected proportions based on the subfamily classification counts for the whole data were tested using the Exact binomial test for goodness-of-fit. The results of this test are illustrated in Table 3.10. The p-value (two-tailed) for this test is $4.62 \times 10^{-6}$ meaning the null hypothesis that, Ta1-d and Ta1-nd proportions for FL-L1HS with disrupted ORFs should be in the proportion of 1:3.75 is rejected.

Table 3.10: Non-intact ORFs observed vs. expected for exact test for goodness-of-fit

<table>
<thead>
<tr>
<th>L1 sub-family</th>
<th>Non-intact ORFs observed</th>
<th>Non-intact ORFs proportions</th>
<th>Non-intact ORF expected numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ta1-nd</td>
<td>303</td>
<td>3.7</td>
<td>270</td>
</tr>
<tr>
<td>Ta1-d</td>
<td>40</td>
<td>1</td>
<td>73</td>
</tr>
</tbody>
</table>

The results of the intact and non-intact analysis are consistent with the Ta1-d subfamily being younger and thus containing a higher proportion of intact elements. From the analysis the proportion of intact Ta1-d is 62/102 ($\sim$61%) and for Ta1-nd is 79/382 ($\sim$21%) i.e. a Ta1-d element is approximately three times as likely to be intact as a Ta1-nd element.
Comparison of predicted FL-L1HS sequences with experimental methods

The 154 FL-L1HS sequences that are predicted to be capable of retrotransposition were compared to a set of FL-L1HS sequences that were tested for retrotransposition activity (Brouha et al., 2003). The study by Brouha et al. (2003), tested 82 FL-L1HS sequences for retrotransposition activity and reported an activity for each, classifying elements as highly active, weakly active or inactive. Comparing the FL-L1HS sequences detected in this current study to those of the Brouha et al. (2003) study could provide validation of element activity. Figure 3.21 illustrates that only 57 (B) of the FL-L1HS sequences collected here are found in the Brouha et al. (2003) data set. 97 FL-L1HS sequences are found to have intact open reading frames and no mutations in key amino acid residues (A). 25 FL-L1HS from the Brouha study were not found in the FL-L1HS data set. The reasons for this are listed in Table 3.11.
Table 3.11: A list of explanations for why FL-L1HS sequences tested in the Brouha dataset are missing from this dataset.

<table>
<thead>
<tr>
<th>Reason why brouha FL-L1HS not in predicted active dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORF2 contains a stop codon</td>
</tr>
<tr>
<td>L1 falls below the length cut off (&lt;5922 bp)</td>
</tr>
<tr>
<td>ORF1 contains a stop codon</td>
</tr>
<tr>
<td>Mutation in a key amino acid</td>
</tr>
<tr>
<td>Accession is replaced</td>
</tr>
</tbody>
</table>

**FL-L1HS sequences in genes**

L1s have been reported to cause disease in 16 individuals by inserting into genes, with 4/16 (25%) being full-length. To assess how many FL-L1HS sequences out of the total dataset of 533 FL-L1HS are inserted into genes, the UCSC gene track RefSeq Genes (table refGene, containing 26,662 genes) was downloaded from the UCSC table browser (http://genome.ucsc.edu/). Each gene symbol was filtered to remove duplicated instances from different transcripts if they occurred within 10-kb of the transcriptional start and stop of the largest transcript size for that particular gene. This was done using the script `gene_unique_filter.pl`. This gave a filtered gene set of 20,304 genes.

To determine if an FL-L1HS sequence was present in a gene, the reference start and stop was required to fall within the gene’s largest transcriptional start and stop range. Using the perl script `gene_finder.pl`, 164 FL-L1HS sequences (30.4%) out of the 533 FL-L1HS dataset were found to reside within a gene, by these criteria. The 164 FL-L1HS sequences located in genes were then tested to see if they were overlapped with the 154 FL-L1HS sequences that are predicted to be active because they have intact ORFs and no mutations in key amino acid motifs. 54/164 (32%) FL-L1HS sequences were located in genes and had intact ORFs, while 110/164 (68%) FL-L1HS sequences were located in genes and had disrupted ORFs.
To see if FL-L1HS sequences that were found to be located within genes cluster on certain chromosomes, the numbers of intact, and non-intact FL-L1HS sequences that were found in genes were plotted against the chromosome on which they were discovered. Figure 3.22 shows the distribution of all FL-L1HS sequences located in genes throughout the genome, along with those that have intact ORFs and non-intact ORFs. This illustrates that larger chromosomes have more FL-L1HS sequences in genes than smaller chromosomes, displaying a similar distribution to the total number of FL-L1HS sequences illustrated in Figure 3.3. Also it is clear, that for most chromosomes, fewer intact FL-L1HS sequences are found in genes than non-intact FL-L1HS sequences. However there do appear to be exceptions. For example on chromosome 6/11/14 numbers of intact/nonintact are very similar or identical (see Appendix Table B.3). On chromosomes 19, 20 and 22 there are no FL-L1HS sequences located in genes. On chromosomes 18 and 21 there are no FL-L1HS sequences detected in genes with open reading frames. On the X chromosome only 1/8 (12.5%) of FL-L1HS sequences are intact and located in genes.
Do FL-L1HS show an insertion orientation bias when inserting into genes?

To address the issue of a possible insertion orientation bias of FL-L1HS sequences when present in genes, the insertion orientation of all 533 unique FL-L1HS sequences detected in the database searches was analysed. The insertion orientation, relative to the reference assembly, of all 533 FL-L1HS sequences was analysed with a MySQL query. The results are illustrated in Table 3.12. As the “plus” and “minus” strands of the assembled chromosome sequences are determined arbitrarily there is no reason to suppose that there should be a bias of L1 insertion for one strand or another.

Table 3.12: Insertion orientation of 533 FL-L1HS sequences

<table>
<thead>
<tr>
<th>Insertion orientation</th>
<th>N° FL-L1HS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sense (+)</td>
<td>254</td>
</tr>
<tr>
<td>Antisense (-)</td>
<td>279</td>
</tr>
</tbody>
</table>

Table 3.12 illustrates that FL-L1HS sequences appear to insert into the genome in roughly equal proportions relative to the Watson and Crick strands, as expected from models of the TPRT process. The ratio of FL-L1HS sequences in the sense orientation vs FL-L1HS sequences inserted in the antisense orientation in genomic sequence is 1:1.10. However it has been reported previously that L1 insertions of all sizes are more often inserted in the opposite orientation with respect to gene transcription (Roy-Engel et al., 2005). A further analysis was performed to discover if the unbiased orientation of all FL-L1HS sequences is different for those FL-L1HS sequences that have inserted in genes.

If a FL-L1HS inserts into a gene in the same transcriptional orientation as a gene the L1’s sense promoter could potentially interfere with gene transcription (Han et al., 2004). As there is no priori reason to think that the L1 retrotransposition machinery discriminates between transcribed and non-transcribed regions, any insertion bias observed for L1 in genes must arise subsequent to insertion, likely as a result of selection against sense insertions that disrupt gene transcription. To test whether FL-L1HS sense orientation insertions are selected against, the 164 FL-L1HS sequences
that were found to reside within the transcriptional start and stop coordinates of known genes, were classified as either inserting into the same transcriptional orientation as the gene (gene/FL-L1HS +/+ or -/-) or the opposite transcriptional orientation (gene/FL-L1HS +/- or -/+). Sixty one FL-LHS sequences were found to have inserted into the same transcriptional orientation as their host gene (+/+ or -/-) and 103 FL-L1HS sequences were found to have inserted into the opposite transcriptional orientation (+/- or -/+). Table 3.13 illustrates the FL-L1HS insertion orientation with respect to gene orientation. To determine if these observed numbers for insertion orientation in genes was significant a $\chi^2$ test of independence was performed. The expected numbers of FL-L1HS sequences in the same orientation as the gene it had inserted into was 82, while 82 FL-L1HS sequences were expected to insert in the opposite orientation to the gene (i.e. a 1:1 ratio).

**Null Hypotheses**

The null hypothesis is that FL-L1HS sequences should insert in equal proportions into genes in either orientation (same and opposite).

The $\chi^2$ value of 10.76 (1 d.f P=0.05(3.84), P=0.01(6.64)) means that the null hypothesis should be rejected, suggesting FL-L1HS sequences are not observed to insert into equal proportions into genes with the same and opposite orientation as the FL-L1HS sequence. Therefore there is a significant difference in the the number of FL-L1HS sequences that are observed in the same transcriptional orientation and opposite transcriptional orientations.

Table 3.13: Insertion orientation of 164 FL-L1HS sequences with respect to gene orientation

<table>
<thead>
<tr>
<th>Gene/FL-L1HS orientation</th>
<th>Observed $N^{o}$ FL-L1HS</th>
<th>Expected $N^{o}$ FL-L1HS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Same orientation (+/+,-/-)</td>
<td>61</td>
<td>82</td>
</tr>
<tr>
<td>Opposite orientation (+/-,-/+</td>
<td>103</td>
<td>82</td>
</tr>
</tbody>
</table>

FL-L1HS sequences that are able to undergo retrotransposition have the capability to initiate transcription from the internal sense promoter at the 5' end of the
element, therefore FL-L1HS sequences that are inserted into the same transcriptional orientation as a gene have the potential to disrupt or interfere with gene transcription as well as cause premature polyadenylation (endogenous L1s are known to be variably polyadenylated) and transcription of the sense strand could lead to premature polyadenylation of the gene transcript, which would cause at best transcript degradation through non-sense mediated decay or at worst truncation (Perepelitsa-Belancio and Deininger, 2003). To test whether FL-L1HS sequences that have two intact ORFs inserted independently of their host gene orientation, the 164 FL-L1HS sequences within the transcriptional start and stop of known genes were further analysed to see if they appeared in the 154 FL-L1HS sequences that have two intact open reading frames. Table 3.14 illustrates the results of this analysis. Elements with intact ORFs are inserted in the same and opposite transcriptional orientation with equal frequency (Ratio 0.97:1). By contrast, elements with disrupted ORFs are found much less frequently in the same transcriptional orientation (ratio 1:2.5). Twenty nine FL-L1HS sequences are found in the opposite transcriptional orientation with respect to gene orientation compared to 74 FL-L1HS sequences with non-intact ORFs.

Table 3.14: Insertion orientation of 164 FL-L1HS with intact ORFs and non-intact ORFs located in genes

<table>
<thead>
<tr>
<th>Transcriptional orientation</th>
<th>Intact ORFs</th>
<th>Non-intact ORFs</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>30</td>
<td>31</td>
<td>61</td>
</tr>
<tr>
<td>N°° FL-L1HS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(same transcriptional orientation)</td>
<td>15</td>
<td>19</td>
<td>34</td>
</tr>
<tr>
<td>Gene/FL-L1HS orientation (+/+)</td>
<td>15</td>
<td>12</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>74</td>
<td>103</td>
</tr>
<tr>
<td>N°° FL-L1HS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(opposite transcriptional orientation)</td>
<td>12</td>
<td>38</td>
<td>50</td>
</tr>
<tr>
<td>Gene/FL-L1HS orientation (+/-)</td>
<td>17</td>
<td>36</td>
<td>53</td>
</tr>
</tbody>
</table>

A Fisher’s exact test was used to test for independent insertion orientation of the two classes of FL-L1HS sequences into the same or opposite transcriptional orientation as their host gene.
Null hypothesis

The orientation of FL-L1HS sequences with intact ORFs or with disrupted ORFs should be in equal proportions regardless of the transcriptional orientation of the gene.

The result of the Fisher’s exact test (p=0.00758), indicates that the null hypothesis can be rejected; there is a significant difference in the ratio of intact ORFs/non-intact ORFs when the transcriptional orientation of the inserted FL-L1HS and gene orientation are taken into consideration. If we assume that due to the strong cis preference of L1 proteins for their encoding RNA, most newly inserted full length insertions will also be functionally intact (barring reverse transcriptase errors) and elements with intact ORFs will on average be younger than insertions with disrupted ORFs. As a result, the lack of orientation bias of intact FL-L1HS sequences in genes is consistent with an unbiased insertion mechanism and the subsequent loss of insertions in the same orientation is due to selection.

Into what type of gene do FL-L1HS sequences insert?

The annotation of FL-L1HS sequences collected in this analysis showed that 164 FL-L1HS sequences have inserted into genes. We next considered whether genes containing FL-L1HS sequences had particular properties, or were simply a random sample of the available gene targets. The genes harbouring FL-L1HS sequences were compared to all other known human genes with respect to their size and their associated Gene Ontology (GO) to see if there was over-representation or under-representation of certain sizes or classes of genes.

Do FL-L1HS sequences insert into genes of certain sizes?

To determine gene “size” it was necessary to simplify the complex transcriptional output typical of many human genes. For this analysis the genomic transcriptional start and stop coordinates of the largest transcript of a particular gene were defined as its genomic “size”. Under this assumption 20,304 genes in the RefSeq database have
an average genomic size of 56.2 kb in length with a mean exon number of ∼10. In stark contrast, the mean size of a gene that contains a FL-L1HS was ∼391 kb, with a mean exon number of ∼18. Removing the 164 FL-L1HS sequences found in genes from the whole RefSeq gene set reduces the mean genome size for all genes not containing a FL-L1HS to 53 kb. The difference in gene size between genes that have FL-L1HS sequences and those that do not is illustrated in Figure 3.23. To test whether this difference was significant, a paired t-test was performed.

**Null hypothesis**

There is one measurement variable, gene size, and one nominal variable, gene class (FL-L1HS present in a gene, no FL-L1HS in a gene). The null hypothesis is that the mean gene size in the two gene classes are the same.

The results of the paired t-test (163 d.f., $P = 8.01 \times 10^{-22}$) suggest that the null hypothesis should be rejected. Thus it appears that the size distributions of genes that contain a FL-L1HS and those that do not, are statistically different.

This analysis strongly suggests that FL-L1HS insertions occur most frequently within larger sized genes, which might not be expected under a model of random insertion.

**Gene Ontology**

The results from the gene-size test suggest that FL-L1HS sequences insert into genes that are larger than the average size of genes that do not contain FL-L1HS sequences. To determine if the 164 FL-L1HS sequences that have inserted into genes are located within specific classes of genes Gene-Ontology (GO) terms associated with these genes were investigated.

The GO database (http://www.geneontology.org) provides a useful tool to annotate and analyze the function of large numbers of genes. Using the 164 FL-L1HS in genes an online analysis was used to examine if these genes have particular shared properties. The program GOstat (http://gostat.wehi.edu.au/ (Beissbarth and Speed, 2004)) analyses lists of genes and provides statistics about the GO terms contained in the genelist. It
also sorts the GO annotations, giving the GO terms most representative of the gene list first. The gene symbol of the 164 genes containing FL-L1HS sequences were uploaded to the analysis web page. The GO annotation of the gene list was compared against the GO annotation of all human genes (as analysed at the EBI: goa_human). As some genes contain multiple FL-L1HS sequences the non-redundant list of genes was reduced to 150. Of these genes 136/150 (90.6%) have been annotated with GO terms. The number of GO terms found was 1014, with 432 being unique GO terms (i.e. occurring only once in the list). The p-value cutoff was set to 0.01 and the correction method used was the Benjamini algorithm. All other options were left at default. The top 10 over represented GO terms are listed in Table B.4, Table B.5 and Table B.6. There were no under-represented terms that met the p-value cutoff. The top 10 GO terms that returned significant p-values are listed in more detail in Table 3.15.
Table 3.15: 8-10 hits terms and definitions for the top 10 most significantly over-represented GO terms.

<table>
<thead>
<tr>
<th>GO Term</th>
<th>Ontology</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:004459</td>
<td>Cellular component</td>
<td>Plasma membrane part</td>
</tr>
<tr>
<td>GO:0031226</td>
<td>Cellular component</td>
<td>Intrinsic to plasma membrane</td>
</tr>
<tr>
<td>GO:0044425</td>
<td>Cellular component</td>
<td>Membrane part</td>
</tr>
<tr>
<td>GO:0005886</td>
<td>Cellular component</td>
<td>Plasma membrane</td>
</tr>
<tr>
<td>GO:0051179</td>
<td>Biological process</td>
<td>Localization</td>
</tr>
<tr>
<td>GO:0022610</td>
<td>Biological process</td>
<td>Biological adhesion</td>
</tr>
<tr>
<td>GO:0007155</td>
<td>Biological process</td>
<td>Biological adhesion</td>
</tr>
<tr>
<td>GO:0016020</td>
<td>Cellular component</td>
<td>Membrane</td>
</tr>
<tr>
<td>GO:0005887</td>
<td>Cellular component</td>
<td>Integral to plasma membrane</td>
</tr>
<tr>
<td>GO:0006810</td>
<td>Biological process</td>
<td>Transport</td>
</tr>
</tbody>
</table>

The results from the GO search suggest that FL-L1HS sequences are found preferentially inserted into genes annotated with “membrane” related terms, as 60% of top hits fall into this category.

The GO terms that were found to be significant were compared to the 150 unique genes to see how many of this set were within the unique gene set. The overlap between genes that have significant GO terms and the full gene list containing a FL-L1HS is shown in Figure 3.24.

The average size of the genes harbouring L1s with significant GO terms was then compared to those genes containing L1 that did not have significant GO terms. The average size of L1 containing genes with significant GO term was ∼398 kb. The average size of L1 containing genes that lacked a significant GO term was ∼327 kb. To test whether this difference is significant a paired t-test was performed.
Figure 3.24: Overlap of genes with significant GO terms and the unique gene dataset (150) containing a FL-L1HS.
Null hypothesis

There is one measurement variable, gene size, and one nominal variable GO class (significant GO term, no significant GO term). The null hypothesis is that the mean gene size in the two GO classes are the same.

The results of the t-test (66 d.f., P= 0.085) suggest that the null hypothesis should be accepted. Thus it appears that the size distributions of genes that contain FL-L1HS sequences and have a significant GO term and those that do not are not statistically different.

This analysis suggests that while 60% of L1 containing genes are associated with particular GO terms (membrane functions), these genes are no larger than those genes with different GO annotations.

3.12 Discussion

This chapter aimed to address the collation of FL-L1HS sequences that met the criteria of ≥ 98% identity to L1.3_mpa and ≥5922 bp in length. These criteria were selected, as they would likely yield L1s that were young (Ta-1), full-length, and therefore likely to be still retrotransposition competent.

These sequences were collected by searching for FL-L1HS in four main sources (Reference (Lander et al., 2001), Celera (Venter et al., 2001), HuRef (Levy et al., 2007) and the nt accession database), using a combination of pre-written sequence analysis programs and custom-built Perl scripts. The searches revealed that across all sources of potential FL-L1HS sequences, 533 distinct FL-L1HS sequences were detected. This is in contrast to previous searches for FL-L1HS sequences that found 124 (Myers et al., 2002) (draft Reference assembly) and 256 FL-L1HS sequences (Badge et al., 2003) (http://www.le.ac.uk/ge/ajj/LINE1/ nt accession). Therefore this more recent analysis has found 409 (76.7%) and 277 (52%) more than the analyses from (Myers et al., 2002) and (Badge et al., 2003) respectively. However the Myers et al. (2002) analysis was performed on the draft human genome assembly, that would have been of poorer quality
than the “finished” genome sequences available today (Schmutz et al., 2004) (miss-assemblies were reported to occur every 2.6 Mb, on average in these data (Schmutz et al., 2004)). The Badge et al. (2003) analysis used the nt databases, (circa 2001) that would have contained significantly fewer sequences than current nt databases due to continued re-sequencing efforts and the recent deposits of the Fosmid sequencing generated by Bovee et al. (2008). Therefore an increase in the number of FL-L1HS sequences detected is to be expected from mining higher quality genome assemblies and larger nt databases. However such a large increase suggests that previous searches either missed FL-L1HS sequences or, because of the continued sequencing from individuals, there is more FL-L1HS variation between individuals than was previously suspected.

The computational methods developed in this thesis for the detection of FL-L1HS sequences therefore appear to be highly efficient in detecting and characterising FL-L1HS from a number of different data sources and can be used on future builds of the current human genome assemblies with the minimum of effort. Also the methods developed could, in theory, be applied to any species genome that contains LINE-1 elements that have the potential to be polymorphic. There is however a caveat in using this method to detect FL-L1HS insertions. If a FL-L1HS sequence were not in a genome assembly or a genomic database then these computational methods would never be able to detect it. This method of detection also relies upon the accurate assembly of the sequences, due to the high sequence identity required to ensure detection of a FL-L1HS. The presence of poor quality sequences would strongly affect the numbers of FL-L1HS sequences found. Experimental techniques such ATLAS (Badge et al., 2003) however do not have these issues and can screen for polymorphic FL-L1HS insertions in as many individuals as a researcher desires, detecting FL-L1HS insertions at low allele frequencies, that are population-specific or private (i.e. only one copy has ever been detected).

The computational methods used to search genome assemblies will most likely detect FL-L1HS at high- allele-frequencies (with the exception of the nt database which may contain low-allele-frequency insertions that are not in the genome assemblies). There is also very little or no information on population or allele frequency available from these resources. The main disadvantage of this method though is the time and cost required to
find these FL-L1HS insertions, as they can be re-run every time the genome assembly and the genomic database are updated, which in some cases could be every week.

The results of searching the genome assemblies is that there is a large amount of FL-L1HS variation between the different assemblies (Reference: 462, Celera: 353, HuRef: 239). As each assembly was constructed and produced by different methods, and used different individuals, some variation in the number of FL-L1HS sequences detected was expected due to the polymorphic nature of FL-L1HS concordant with previously reported analyses (Konkel et al., 2007). However there does appear to be a large difference between the source of the largest number of FL-L1HS sequences detected in the Reference assembly and the other two assemblies, Celera and HuRef. Possible reasons for this include the methods used to sequence and construct the assemblies. The Reference assembly was generated by shotgun sequencing of BACs (~100-200 kb) (Lander et al., 2001), therefore reducing the problem of many near-identical copies of repeats during the assembly process. The reference Assembly has been reported by Istrail et al. (2004) to have better coverage of near and exact repeats. However the other two assemblies did not use this method, but used whole-genome shotgun sequencing using paired-read ends to span across repeats (Levy et al., 2007; Venter et al., 2001). If an L1 is located in a region of repeat-rich DNA there is the possibility that the assembly algorithm could become inaccurate (due to the high number of equally likely assemblies which are near identical). This can lead to certain regions being excluded in error. A possible future investigation could be to see where all of the inter-assembly polymorphisms occur and analyse the repeat content of these regions to see if this is indeed the case (i.e. that high densities of repeats surrounding FL-L1HS insertion points exclude them from the assemblies generated by WGS). However there is also the possibility that these could be genuine polymorphisms. However at this time there are no data available to support either theory until further investigations have been performed (computational and PCR genotyping).

As mentioned above, each of the three assemblies analysed display some degree of inter-assembly polymorphism (i.e. a FL-L1HS absent in one or more of the assemblies). The Reference assembly has 89 FL-L1HS sequences that are not present in any of the other assemblies displaying the largest amount of assembly polymorphism. The Celera assembly has only 2 FL-L1HS sequences that were not found in the reference
assembly and the HuRef had 10 that were not found in the Reference assembly. Only 192 FL-L1HS sequences were shared across the three assemblies (Reference, Celera and HuRef). A previous computational analysis between the Reference assembly and the Celera assembly Konkel et al. (2007) detected 34 polymorphic loci between the two assemblies (28 Reference, 6 CHG), with only 25% (8/34) being full-length. The searching of the Reference and Celera assemblies in this thesis had found 91 potentially polymorphic loci (89 Reference, 2 Celera). However these loci would have to be validated by PCR to confirm if they were genuine polymorphisms or just artefacts generated by the way the two assemblies were constructed.

As well as screening the genome assemblies, the nt accession database was also searched for FL-L1HS insertions. This analysis returned 765 hits (685 human and 80 non-human). Twenty eight FL-L1HS sequences were removed from the human analysis as they could not be mapped to unique positions leaving 657 human hits, 504 of which were unique. Of the 504 hits, 58 were unique to the nt database as they were not detected in any of the human genome assemblies. These sequences could have been heterozygous (-/+ in the individual(s) sequenced with the empty site being represented in the assembly or simply not present in the individuals sequenced for the genome assemblies.

These searches also retrieved non-human hits which met the criteria for a FL-L1HS (75 Pan troglodytes, 3 Pan paniscus and 2 Gorilla gorilla ) suggesting that the nt database could be a good resource of non-human L1s which display high similarity to known FL-L1HS and could thus be potentially capable of retrotransposition.

Searching for FL-L1HS sequences in this data source does have its pitfalls; for example if a L1 is split across two accessions (i.e. the L1 is located at the end of one accession and the start of another) these FL-L1HS sequences would be excluded as they would not meet the length cut off criteria. In this analysis 29 (29/533 5.4%) FL-L1HS sequences were not found in the nt database but were found in the genome assemblies. There is a possibility that these may be FL-L1HS sequences split between two accessions that are not found in any other resource apart from the nt accession database. Using this method therefore would miss these L1s. A possible solution could be to employ the same BLAST searching technique, but search with the 5’ and 3’ ends of L1.3_mpa and
select all hits which show $\geq 98\%$ identity. The 5$'$ and 3$'$ flanking DNA of the L1 could then be mapped to the genome. Even if the L1 is not in the assembly, the flanking DNA should map to the same location. Then the flanking DNA could be searched for matching TSDs using TSDmapper. Matching TSDs and mapping to the same locus would strongly suggest that these sequences are in fact the same L1. A disadvantage of this method would be that L1s that are truncated may also be detected.

The resolution of duplicate L1 sequences from each of the four sources was an important step to determine if the L1s found in different sources were exact copies or were alleles. This was done as there are reported cases of L1 alleles showing different retrotransposition activity. Seleme et al. (2006) analysed three hot L1s AL512428 (BL000229), AC002980 (BL000503), and AC021917 (not detected in this analysis) and found different degrees of activity existed between the individual alleles tested. The two FL-L1HS sequences detected in the baseLINE analysis were found to be in the Reference assembly and the nt database and were both found to have intact ORFs and no mutations in conserved motifs by L1Xplorer analysis (Penzkofer et al., 2005). This suggests that sequences from the assembly and the accession database are probably the same allele. However in this analysis there are examples of FL-L1HS alleles that are predicted computationally to be active and inactive (e.g. BL000060, NC_000002.10_aj: Active, AC_000045.1_y: Active, AC_000134.1_t: Inactive). The predicted inactive L1 allele AC_0000134.1_t is known to come from a single individual (Craig Venter, HuRef assembly) whilst the other two are from at least one other individual. This analysis could therefore be used to computationally predict active or inactive alleles that could then be tested in a retrotransposition assay (Moran et al., 1996). Even if the alleles are predicted to be inactive, the retrotransposition assay may provide information on the relative activity that each of the FL-L1HS alleles possesses. The computational prediction of activity relies on the quality of the sequence for each of the four data sources, so the computational predictions may be inaccurate if there are sequencing errors in any of the motifs tested. Also if the L1s found in this analysis were to be tested, the FL-L1HS sequences would have to be extracted from the clone or individual originally used because, as mentioned above, different people can have different L1 alleles displaying different activity status.
Distribution of FL-L1HS sequences

The analysis of the distribution of FL-L1HS sequences on each of the genome assemblies reveals that they are not distributed according to the size of the chromosome and the density of FL-L1HS per chromosome is not explained by chromosome size for the Reference and HuRef assemblies. However for the Celera assembly this was not the case as the data suggest that FL-L1HS sequences are indeed distributed according to the size of the chromosome. This may be artificial as the other two assemblies do not display this effect and it may be due to the way the Celera assembly was constructed.

An analysis of the relationship between chromosome size and the number of FL-L1HS sequences revealed that the size of the chromosome is positively correlated with the number of insertions which agrees with an analysis by Boissinot et al. (2004) who considered Ta-1 inserts. However there were exceptions, as chromosome 18 on all assemblies has more FL-L1HS sequences than expected for its size compared with a similarly sized chromosome, chromosome 17. Nusbaum et al. (2005) reported an extremely low density of protein-coding genes on chromosome 18, which may explain the high numbers of FL-L1HS sequences found on this chromosome. As FL-L1HS insertions are often deleterious, the lack of protein-coding genes on chromosome 18 could allow the accumulation of FL-L1HS insertions without affecting gene expression. Han et al. (2004) have also reported that L1 sequences were present in lower amounts in highly expressed genes and lower-expressed genes had more L1 sequences. Boissinot et al. (2004) in addition reported that the number of FL-L1HS sequences are related to gene size. The data from this analysis therefore fit with previous observations that an increased amount of FL-L1HS sequences should be found on chromosomes with a low amount of protein-coding genes or found in genes that are expressed at lower levels.

The X chromosome, from the Reference and the Celera assembly, also displays a larger amount of FL-L1HS sequences than expected for its size (Ref size: 155 Mb L1s: 38, Celera size: 155 Mb L1s: 30). This excess of full-length L1 sequences might reflect the reduced efficiency of selection on the X chromosome, as it spends 1/3 of its time in males, where only the pseudoautosomal regions can recombine. In principle this makes the X chromosome a good candidate for the accumulation of FL-L1HS sequences. The larger than expected number of FL-L1HS sequences found on the X chromosome from
these two sources agrees with previous study from Boissinot et al. (2001) who reported a higher density of retrotansposons in regions of the genome where the recombination rate is lower. They reported that FL and TR (≥1.2 kb) were also found in non- and low-recombining regions of the genome. L1s that insert into high-recombining regions of the genome should be selected against due to being efficient mediators of ectopic recombination, which can lead to chromosomal rearrangements (Burwinkel and Kilimann, 1998). This could therefore explain the increase in the number of FL-L1HS sequences on the X chromosome.

Other authors have proposed that X-inactivation, the process by which X chromosome dosage compensation is achieved in mammals, could be connected to the increased number of FL-L1HS sequences. L1 sequences were suggested by Lyon (1998) to act as booster elements to promote the spread of Xist, a non-coding RNA which coats the inactive X-chromosome. Bailey et al. (2000) and Lander et al. (2001) found that there was a significant enrichment at Xq13 where the X-inactivation centre is reported to occur and that there is a reduced L1 density in genomic loci that escape X-inactivation. Thus these FL-L1HS enrichment results could also support the theory that L1s play a role in L1 X-inactivation.

In contrast to the Reference and Celera assemblies, the HuRef assembly has significantly fewer FL-L1HS sequences on the X chromosome (HuRef size 144 Mb, L1s 1). A possible explanation for this exceptionally low number on this chromosome could be due to the source DNA that was used in the sequencing. As the DNA is known to come from a male donor (Craig Venter) there would only have been one copy of the X-chromosome available for sequencing thus reducing the coverage of DNA sequenced compared to autosomes. Also as the X chromosome is ~11 Mb smaller than those reported on the Reference and Celera assemblies it could suggest that repeat regions may have been excluded (assembly collapses).

Trace file screening

An additional source of FL-L1HS sequences that was examined were the trace archives. Searching for FL-L1HS sequences in this source was more difficult as it was impossible to extract the whole 6-kb L1 sequence due to the size of the sequence fragments in
this data source (∼200-1000 bp). Searching for the 5’ ends of FL-L1HS sequences in the trace archive and excluding the previously identified FL-L1HS sequences from other sources found an additional 48 potential FL-L1HS sequences that were not in the other sources. Mapping these traces to the human genome revealed 20 (41.7%) were found within the transcriptional start and stop boundaries of a RefSeq gene, suggesting that the trace archives are a good place to look for polymorphic L1s that have inserted into genes. This analysis concurs with previous work by Bennett et al. (2004) who also discovered rare polymorphic FL-L1HS loci that are not found in human genome assemblies or nt accession databases. However, the computational analysis alone cannot confirm that these 48 L1s are polymorphic FL-L1HS sequences as this would have to be confirmed by genotyping. In addition, if the insertions present only in the trace archives were very rare, the likelihood they could be verified is questionable. However the analysis has provided 48 loci to test that would not have been found if only assembled sequences were searched.

**In-silico genotyping**

The 533 FL-L1HS sequences detected from the searches of the four sequences databases provides a large sample set to perform in-silico genotyping. The empty sites of all 533 FL-L1HS sequences were generated using TSDmapper. These empty sites were screened against the nt databases revealing 117 empty site hits with 88 FL-L1HS empty sites mapping to human DNA. Therefore, that 14% of FL-L1HS sequences detected in this analysis are predicted to be polymorphic to some degree, which is consistent with the analysis by Bennett et al. (2004) that the active Ta-1 subfamily elements are frequently polymorphic. The advantage of this analysis is that it can be re-run every time the nt database is updated.

**Subfamily distribution Intact ORFs**

The analysis of subfamily distribution for all 533 FL-L1HS sequences found that that largest detected subfamily in this dataset was Ta1-nd (382), followed by Ta1-d (102) and non-canonical (41), with 8 FL-L1HS sequences having a multiple-subfamily classification. The ratio of Ta1-d (the youngest subfamily) to Ta1-nd (older subfamily)
was found to be 1:3.7, which is consistent with previous observations (Bennett et al., 2004).

The number of FL-L1HS sequences identified with intact ORFs (and conserved functional motifs) that were therefore predicted as potentially active (i.e. able to undergo retrotransposition) was 154 (29% of the total dataset) with 379 predicted to be inactive. The sub family classification of the predicted active and inactive elements revealed that 62 FL-L1HS sequences from the Ta1-d family (youngest sub-family) or 61% (62/102) of the all Ta1-d detected were potentially capable of retrotransposition. Seventy-nine FL-L1HS sequences classified as Ta1-nd or 21% (79/382) of all Ta1-nd detected were potentially active, indicating that the younger subfamily FL-L1HS sequences are more likely to be capable of retrotransposition. The 154 FL-L1HS sequences predicted to be capable of retrotransposition represents an increase of 72 FL-L1HS sequences from the analysis by Brouha et al. (2003). However the overlap between these two data sources was not as close as expected and reasons for this are listed in Table 3.11. However because of the increased number of sources searched, a higher number of predicted active FL-L1HS sequences was to be expected.

**Number FL-L1HS sequences in genes**

Out of the total 533 dataset of L1 insertions discovered, 164 FL-L1HS sequences (30% of the FL-L1HS dataset) were found to reside within the transcriptional start and stop of RefSeq genes. 54 of these elements were found to have intact ORFs and be potentially capable of retrotransposition. The insertional orientation ratio of all 533 FL-LHS sequences with respect to the genome was found to be 1:1 (+/- strands) as expected from models of TPRT. The number of FL-L1HS sequences that were found to have inserted into genes (164) was tested to see whether there was any bias in the insertion orientation. Sixty one were found to be in the same transcriptional orientation (+/+ or -/-) and 103 (+/- or -/+ in the opposite transcriptional orientation (ratio 1:1.7). These results suggest that FL-L1HS sequences are nearly 2 times more likely to insert in the opposite transcription of a gene that those in the same transcriptional orientation. These results agree with observations by Boissinot et al. (2004) who found that Ta-1 elements were twice as likely to be in the antisense orientation compared with the sense orientation, and the observations made by Roy-Engel et al. (2005) who found
that in the draft sequence of the human genome there was an orientation bias in genes where fewer L1s and Alu sequences were observed when the polyadenylation signal (AATAAA) was in the same transcriptional orientation as the gene. As an insertion of a FL-L1HS into the same transcriptional orientation of a gene may disrupt transcription (Han et al., 2004) and there is no reported evidence that the L1 retrotransposition machinery preferentially targets transcribed or non-transcribed regions, the selection against these types of insertions probably occurs after insertion. An insertion into the same transcriptional orientation of a gene in functional or regulatory sequences (promoters, exons and enhancers) could potentially interfere with gene transcription elongation (Han et al., 2004) or cause alternative splicing (Narita et al., 1993). These observations support the results that FL-L1HS sequences are more likely to be observed in the opposite transcriptional orientation of a gene due to the negative mutational impact FL-L1HS sequences have on the genome when in the same transcriptional orientation as a gene.

FL-L1HS also have a second promoter, located in the 5' UTR (their antisense promoter (ASP)). When a FL-L1HS is inserted into the opposite transcriptional orientation of a gene, the ASP can produce transcripts that contain part of the L1 5' UTR and part of the cellular mRNA tail (Speek, 2001). L1s can also, in theory, ‘break’ genes and produce multiple transcripts (Wheelan et al., 2005), but we did not analyse the potential for gene breaking by elements in our dataset. However the results of this analysis (that FL-L1HS occur twice as frequently in the opposite transcriptional orientation of genes) suggests that although FL-L1HS may still interfere with gene expression in the antisense orientation, the mutational effect of ASP generated transcripts are less severe than the effect of a FL-L1HS that has inserted into the same transcriptional orientation as the gene.

FL-L1HS sequences with intact ORFs were found to occur in the same and opposite transcriptional orientation of genes with equal frequency (0.97:1), but FL-L1HS sequences with disrupted ORFs were found less frequently in the same transcriptional orientation. The lack of orientation bias for intact elements also supports the hypothesis of an unbiased FL-L1HS insertion mechanism where insertions in the same orientation as their host gene are lost over time due to selection.
Gene ontologies

An analysis of the size of the genes in which FL-L1HS sequences occur revealed that they often insert into large genes (on average 391 kb in length). To see if FL-L1HS sequences insert into specific classes of genes the gene ontology of these genes was analysed. The GO analysis terms suggested an over representation of genes with GO terms that fall into the membrane class. Possible reasons for this over representation could be that because membrane proteins can be large, an L1 insertion will have less chance of hitting an important regulatory or coding sequence.
Chapter 4

TSD detection and analysis

4.1 Introduction

The aim of the research described in this chapter was to develop an informatic resource, in the form of a set of Perl scripts, for the automated detection and analysis of L1 retrotransposon Target Site Duplications (TSDs) and transductions.

L1 elements that have recently inserted in the human genome can be recognised using BLAST (Altschul et al., 1990), BLAT (Kent, 2002) and RepeatMasker (Smit et al., 1996-2004), but to automatically determine the actual insertion point and the true boundaries between L1 and genomic sequences at the 5′ and 3′ ends of the element, an algorithm needed to be developed that took into account the biological mechanism of L1 insertion, Target-Site Primed Reverse Transcription (TPRT).

TSDs are short direct repeats, typically 7-20 bp in length of variable sequence, that flank the 5′ and 3′ end of the L1. They are thought to be generated by the host DNA repair mechanisms during L1 integration. TSDs are just one of a number of variable features that can potentially make the detection of correct L1 boundaries difficult. Other variable features include the length of the L1 polyA tail, the position of the polyadenylation signal, and 5′ and 3′ transductions. Thus using algorithms that do not take into account these factors when analysing L1s, can lead to inaccurate identifications of 5′ and
3’ transduction events. For example, one important consequence of not recognising transductions is that primers for PCR-based genotyping assays can be placed within transduced sequences leading to erroneous genotype data. To solve this problem a Perl program called TSDmapper was developed and tested using a manually curated dataset of 184 FL-L1HS sequences, and the 533 FL-L1HS sequences discussed in chapter 3. The algorithm’s performance is compared to another TSD-finding algorithm developed by Szak et al. (2002) called TSDfinder (http://www.ncbi.nlm.nih.gov/CBBresearch/Landsman/TSDfinder/). This chapter discusses the TSDmapper algorithm’s design, its implementation and the analysis of the program’s output.

4.2 Design and implementation of TSDmapper

The development of the algorithm for TSDmapper involved six main steps

1) Sequence validation

2) Identification of polyA tails

3) Detection of potential 5’ and 3’ TSDs

4) Development of a scoring system

5) Identification of 5’ and 3’ transductions

6) Automated reconstruction of empty-sites

During the development and testing of the TSDmapper algorithm the FL-L1HS BL000503 (AC002980.1.a) was used, and in the following sections illustrative examples are given using this FL-L1HS. This element was chosen because it has intact open reading frames (ORFs) and is known to be active in cell culture (Brouha et al., 2003) indicating that it is a member of the currently active L1 cohort. The amount of flanking DNA used in the detection of TSDs was defined as 2 kb either side of the L1. It was felt that this amount of sequence should provide an effective search space for the identification of TSDs, and was the same as the flanking DNA distance
used in a database of manually identified and curated FL-L1HS TSDs created by Dr Richard Badge (http://www.le.ac.uk/ge/ajj/LINE1/FLDB.html). In this database the largest identified 5’ transduction was 577 bp and the largest identified 3’ transduction was 1589 bp. Therefore it was considered that 2 kb of flanking DNA would be of a suitable length for the accurate automated detection of TSDs.

4.3 TSDmapper sequence validation

The first part of the TSDmapper algorithm is the identification of the FL-L1HS 5’ start and 3’ end, along with the validation of the nucleotide sequence. The flanking DNA sequence, is required to be in a specific format with the 5’ flanking DNA preceding a FL-L1HS to be the first sequence in the file, and the 3’ flanking DNA following the 3’ end of the FL-L1HS to be the second sequence, example sequences can be found in the help section of the baseLINE web site (http://baseline.gene.le.ac.uk).

The 5’ start and 3’ ends of the FL-LHS are identified using the Blast2Seq algorithm with L1.3_mpa as the query sequence. At least 250 bp of FL-L1HS was required to be present at both the 5’ and 3’ end of the flanking DNA. The FL-L1HS hits were required to be $\geq 98\%$ identical to L1.3_mpa.

The methods written to validate that the ‘hit’ sequence is a piece of sequence that TSDmapper can accept are illustrated in Figure 4.1. This figure illustrates the ‘blasting’ of L1.3_mpa against the target sequence. Position (1) ensures that the sequence has L1 sequence at the 5’ and 3’ ends, position 2) shows the parsing of the results and the validation of the flanking sequence and position (3) validates the sequence to make sure that it conforms to the IUPAC nucleotide code. Any sequences that do not meet these criteria are excluded from further analyses.
4.4 PolyA tail discovery

The second part of the TSDmapper algorithm is the identification of polyA tails immediately following the 3' end of FL-L1HS. The polyA tail detection in TSDmapper implements the TSDfinder (Szak et al., 2002) algorithm for finding polyA tails. This part of the algorithm was implemented in Perl and the code from TSDfinder was rewritten to fit with the design of TSDmapper. The implementation of the TSDfinder polyA tail algorithm and the adaptation is available to be viewed in the TSDmapper perl module (TSDmapper.pm) in a method called FindPolyATails.

The FindPolyATails method produces a list of polyA tails in the 3' genomic flanking DNA, as illustrated in Table 4.1 for the example FL-L1HS, AC002980.1_a. This list was produced with the following parameters: the threshold for FL-L1HS,
CHAPTER 4. TSD DETECTION AND ANALYSIS

polyA tail detection was set to a minimum polyA tail length of 10 bp, a contamination threshold of 2 non-As, and utilised 3 kb of flanking DNA.

Table 4.1: PolyA tail results for AC002980.1. All start and stop coordinates are relative to the 3 kb of flanking DNA extracted from the L1 sequence using the 3’ L1 blast call

<table>
<thead>
<tr>
<th>polyA tail number</th>
<th>polyA tail start (bp)</th>
<th>polyA tail stop (bp)</th>
<th>polyA tail length</th>
<th>polyA tail</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>156</td>
<td>174</td>
<td>19</td>
<td>ACAAGAAAAAAACAAACAA</td>
</tr>
<tr>
<td>2</td>
<td>239</td>
<td>248</td>
<td>10</td>
<td>AAAAAACACA</td>
</tr>
<tr>
<td>3</td>
<td>753</td>
<td>764</td>
<td>12</td>
<td>AAGAACAAAAAA</td>
</tr>
<tr>
<td>4</td>
<td>1003</td>
<td>1038</td>
<td>36</td>
<td>ATAATAAAAAAAA (cont..)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>AAAAAAAAAAAAAACACA</td>
</tr>
<tr>
<td>5</td>
<td>1041</td>
<td>1066</td>
<td>26</td>
<td>AAAAAAAAAAAAAACAA (cont..)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CAAAACAAA</td>
</tr>
<tr>
<td>6</td>
<td>1209</td>
<td>1231</td>
<td>23</td>
<td>AAAAAAAAAAAAAAAACAA (cont..)</td>
</tr>
<tr>
<td>7</td>
<td>2077</td>
<td>2088</td>
<td>12</td>
<td>AACAGATAAAA</td>
</tr>
<tr>
<td>8</td>
<td>2388</td>
<td>2397</td>
<td>10</td>
<td>AAAAAGAAAA</td>
</tr>
<tr>
<td>9</td>
<td>2770</td>
<td>2781</td>
<td>12</td>
<td>AATAAAATAACA</td>
</tr>
</tbody>
</table>

Table 4.1: PolyA tail results for AC002980.1.a. All start and stop coordinates are relative to the 3 kb of flanking DNA extracted from the L1 sequence using the 3’ L1 blast call.

All methods described from this point forward are specific to TSDmapper and are not related to any found in the TSDfinder perl script. The polyA tails identified by the FindPolyATails method are then passed to another method called ExtractPolyAtails. This method requires as parameters a hash of arrays containing the polyA tail results, the 3’ flanking sequence, the length of the 3’ flanking sequence, and the directory in which the extracted sequences should be placed.

The ExtractPolyAtails method extracts a query sequence from the 3-kb flanking sequence based on the polyA tail stop position, derived from the output of FindPolyATails. The search query start is based on the polyA tail stop position, minus 35 bp and the search query end is based on the polyA stop plus 35 bp. The generation of the 71 bp query sequence is illustrated in Figure 4.2.

The generated query sequences are placed into output files by the method WritePolyASeq so they can be used in the identification of the TSDs in the 5’ flank. The query searches were selected to be 71 bp in length as it was felt that this length would be large enough to contain a potential 3’ TSD that followed a polyA tail and
CHAPTER 4. TSD DETECTION AND ANALYSIS

Figure 4.2: Generation of the query sequence for TSD detection from the polyA tail sequence. The green nucleotides indicate sequence -35 bp of the polyA tail stop, the red nucleotides indicate sequence +35 bp of the polyA tail stop. The example shown is the query sequence generated from polyA tail identifier 6 from AC002980.1_a.

would allow for some boundary imprecision resulting from the BLAST identification calls of the 5’ and 3’ ends of the element. It was proposed that this length of query sequence could be used to identify a TSD in the 5’ flanking sequence. The method RunLalign searches each query search against the 5’ flanking DNA.

4.5 Finding potential 5’ and 3’ TSDs

To find potential TSDs in the 5’ and 3’ flanks a method was developed using a program from the FASTA2 suite of sequence analysis utilities (Pearson and Lipman, 1988), LALIGN (Huang and Miller, 1991). LALIGN was chosen because it is able to detect the best local alignments between two sequences from 6 bp upwards. Since TSDs are reported to vary in size from 7-20 bp LALIGN can find similarities even at the smaller end of the TSD scale.

The method RunLalign in the Perl module TSDmapper takes a list of the query search files generated from the 3’ polyA tail search and the 5’ flanking sequence, and the Expectation value (E) threshold, which is user-specified. The E value default was
CHAPTER 4. TSD DETECTION AND ANALYSIS

lalign -n -f-14 -g-4 -r +4/-12 -E 10 -m 2 "search file" "five prime flanking sequence" >"outfile"

<table>
<thead>
<tr>
<th>lalign options</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>-n</td>
<td>Force the query sequence to be treated as a DNA sequence</td>
</tr>
<tr>
<td>-f</td>
<td>Penalty for the first residue in a gap</td>
</tr>
<tr>
<td>-g</td>
<td>Penalty for each additional residue in a gap</td>
</tr>
<tr>
<td>-r</td>
<td>Specifies a scoring matrix file</td>
</tr>
<tr>
<td>-E</td>
<td>Used to override the expectation value</td>
</tr>
<tr>
<td>-m</td>
<td>Alternate display of matches and mismatches in alignments</td>
</tr>
</tbody>
</table>

| five prime flanking sequence | The five prime flanking sequence file |
| outfile | The results file |

Figure 4.3: The RunLalign method executes the command line options of the lalign program for searching of potential TSDs in the 5’ flanking sequence.

set to 10 for the detection of small similar sequences. The method RunLalign runs the command-line code illustrated in Figure 4.3 for each query search against the 5’ flanking DNA.

All results from the LALIGN searches are placed into a named output file and the location of all the output files from successful searches are returned to the TSDmapper script for further analysis.

The method ParseLalign is used to parse all the lalign results from the LALIGN output file from all of the query searches. All hits are returned regardless of hit identity or length.
The parsed results from LALIGN that contain identical or near-identical matches occurring in both the 5’ and the 3’ flank were defined as potential TSDs. These potential TSDs were scored based on the endonuclease cut site sequence preceding the 5’ TSD and the position of the potential TSD in the 5’ flank, as detailed below.

4.6 Development of a TSD scoring matrix from endonuclease cut sites

To develop a scoring system for TSDs, the integration biology of mammalian elements was analysed. For Alu and ID sequences (mobilised by the L1 machinery) the preferred nicking site for the ORF2-encoded endonuclease domain has been reported to be the consensus 5’-TTAAAA-3′, with a strong preference for the first two bases immediately preceding the start of the TSD to be TT (Jurka, 1997). Using this endonuclease cut site consensus derived from of Alu and ID insertions, the nucleotide occurrence frequency preceding the 5’ TSD 5’ nucleotide and the 5’ TSD itself were analysed, using a dataset of 196 FL-L1HS sequences.

A manually curated dataset of TSDs from 196 FL-L1HS sequences was downloaded (http://www.le.ac.uk/ge/ajj/LINE1/FLDB.html) and 8 bp and 20 bp were extracted from the flanking DNA immediately preceding the 5’ TSD start position. The 8 bp and 20 bp immediately 3’ of the 5’ TSD start were also extracted (i.e. the TSD and L1 length (see Figure 4.4 for an illustration using AC002980.1_a). For each of the FL-L1HS in the manually curated set, the base occurrences frequency was recorded for the +8/-8 bp analysis and the +20/-20 bp analysis. Table C.1 and Table C.2 illustrate the base occurrences at different positions of TSDs and the surrounding regions.

To discover if any of the observed base pair occurrences at certain positions were different from that expected from genomic nucleotide composition, a $\chi^2$ test was performed on each nucleotide count at each of the base-pair positions. The expected base-pair compositions for the -8/+8 bp and -20/+20 bp are shown in Table 4.2 along with the expected based-pair compositions for the genome using the -20/+20 bp analysis.
Figure 4.4: Generation of the table of base occurrences immediately flanking the 5′ TSD start for manually curated TSDs. The example shown illustrated the 8 bp surrounding the 5′ TSD for AC002980.1_a.

Table 4.2: Expected base pair count for the -8/+8 bp, -20/+20 bp based on the total base pair count of the 16 bp and 40 bp analysis for 196 FL-L1HS sequences. Also shown is the base pair count for the -20/+20 bp analysis based on a genome average of 41% GC content.

<table>
<thead>
<tr>
<th>Analysis</th>
<th>T</th>
<th>C</th>
<th>A</th>
<th>G</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected bp counts -8/+8 bp</td>
<td>53.63</td>
<td>23.94</td>
<td>90.88</td>
<td>27.55</td>
<td>196</td>
</tr>
<tr>
<td>Expected bp counts (-20/+20 bp)</td>
<td>48.58</td>
<td>25.65</td>
<td>74.58</td>
<td>47.20</td>
<td>196</td>
</tr>
<tr>
<td>Expected bp counts (genome)</td>
<td>57.82</td>
<td>40.18</td>
<td>57.82</td>
<td>40.18</td>
<td>196</td>
</tr>
</tbody>
</table>
Figure 4.5 illustrates $\chi^2$ values for individual base positions surrounding the start position of the 5' TSD of 196 FL-L1HS from a manually curated dataset of TSDs. The $\chi^2$ values for -8/+8 bp (16 bp), -20/+20 (40 bp), and the genome average (based on 40 bp) above the purple line indicate base positions that are significant at $P < 0.001$ (3 df). There appears to be a cluster of $\chi^2$ values at positions -2 to +5 for all analyses surrounding the 5' TSD. This cluster increases to -2 to +6 bp if the -8/+8 bp analysis is not considered.

The data from this analysis largely agree with an analysis of Jurka (1997). These authors found a cluster of significant $\chi^2$ values surrounding the 5' end of the TSDs in Alu and ID elements from -2 to +4 bp. This analysis of FL-L1HS suggests that there is an additional base position that is significantly non-random and may indicate that an extra base position could be included in the endonuclease consensus cut site. The analysis of both data sets (-8/+8 bp and -20/+20 bp) suggests the endonuclease cut site for FL-L1HS is therefore best represented by a heptamer 5'-'TT↓'AAAAA-3'. If the -20/+20 bp data set is considered an octamer 5'-'TT↓'AAAAAA-3' consensus can be suggested.

Figure 4.5 also illustrates that there is a cluster of $\chi^2$ values at positions +13 to +20 that exceed the significant p value of $< 0.001$. The non-randomness of the base occurrences here could be explained by these base pairs belonging to the 5' end of the L1 poly-purine-rich tract (a run of G nucleotides), since base positions +13 to +20 bp are approaching the upper limits of known TSD lengths and may, by chance, include parts of the L1 5'UTR.

The base positions (-2 to +6 bp) where the $\chi^2$ values exceed the 0.001 P value thresholds were used as the basis of a scoring matrix to score TSDs based on the endonuclease recognition sites. The scoring matrix consists of the percentage counts from the -20/+20 bp analysis for positions -2 to +6 bp. This endonuclease-scoring matrix is shown in Table 4.3.

The development of the scoring matrix is one part of the TSD scoring system. The second part is to take account of how distant the TSD is from the 5' L1 start as defined by BLAST. Since 5' transductions are less common than 3' transductions, the closer the 5'
Figure 4.5: $\chi^2$ values for individual base positions surrounding the start position of the 5' TSD of 196 FL-L1HS from a manually curated dataset of TSDs.

Table 4.3: Percentage of nucleotide occurrences from the -20/+20 bp analysis for the 8 significant $\chi^2$ positions (-2 to +6 bp), consensus TT↓AAAAAA.
TSD start is to the 5’ start of the L1, the higher the score for distance is assigned, leading to a preference for 5’ TSDs located close to the 5’ end of the L1 element. The following section describes the implementation of the Perl methods for scoring potential TSDs and gives examples of scored TSDs using the endonuclease-scoring matrix, combined with a TSD distance score.

### 4.7 Scoring TSDs in TSDmapper

The scoring and validation of a pair of TSDs is initiated from the method `TsdValidation`, in the `TSDmapper.pl` script. The `TsdValidation` method takes as arguments the parsed LALIGN results file, containing the potential TSDs, the polyA tail result file, the directories of the 5’ and 3’ flanking DNA, the 5’ start of the L1 and the 3’ end of the L1 as defined by BLAST, and the flanking length of the 5’ flanking sequence and uses these parameters to filter the potential TSDs.

The LALIGN results need to be parsed and filtered due to the nature of the query search sequence. The query search sequences used to detect potential TSDs were generated from 3’ flanking DNA containing polyA tails, thus many hits detected by LALIGN are likely to be purely polyA tails and not TSDs at all. To remove these hits a method called `StatsTSDs` is called within the `TsdValidation` method. This method analyses the 3’ and 5’ TSDs for the percentage of A nucleotides and the number of mismatches between the putative 5’ and 3’ TSD. Hits that return a percentage mismatch of ≤10% and a percentage composition of A’s of ≤80% are then scored using the method `TSDScoring`.

The `TSDScoring` method accepts a number of different parameters, that include: the TSD hit name (based on the query search sequence), the 5’ L1 start and stop as defined by BLAST, the 3’ and 5’ TSD sequences that meet the TSD filter criteria for percentage ‘A’ identity and percentage mismatches, the length of the TSD, the name of the 5’ and 3’ flanking DNA sequences, the start and stop positions of the 3’ and 5’ TSDs, the search query start and stop, and the length of the 5’ flank.
The first step in this method is to extract the 8-bp endonuclease recognition site based on the start position of the 5' TSD. The endonuclease start position is calculated based on the 5' TSD start minus -2 bp and the endonuclease end position is calculated based on the 5' TSD start position +5 bp, giving a start and stop position for the endonuclease-recognition site in the 5' flanking DNA, spanning 8 bp. The endonuclease-recognition site is then cut out from the 5' flanking DNA using the method BioSeqExtract and returned for further analysis in FASTA format.

The endonuclease sequence is translated into a code based on each of the four possible nucleotides. TCAG becomes translated into the code 0123, so in an array structure (with an array starting at zero) all percentage scores for ‘T’ will be at position 0, ‘C’ position 1, ‘A’ position 2 and ‘G’ position 3. The percentage scores of each nucleotide can then be looked up based on the nucleotide position. For example if there is a T nucleotide at endonuclease position -2 in the endonuclease-recognition sequence, the T would become translated to 0 and the resulting code would be 0, -2 which would give a score for that nucleotide (T) of 56.63 from the scoring matrix. An example of endonuclease-recognition scoring is illustrated in Figure 4.6, based on the results from polyA tail 6 (see Table 4.1) for accession AC002980.1.a. This endonuclease score is a metric based on the endonuclease consensus of TT↓AAAAA as it achieves the highest possible total score of 554.09 when the individual position scores for each nucleotide are simply added together.

After the endonuclease-recognition site scoring, the proximity of the 5' TSD stop to the 5' start of the L1 is analysed. An example of the TSD 5' scoring based on distance from the 5' end of the L1 is illustrated in Figure 4.7 and shows two potential TSDs from the accession AC002980.1.a. The green TSD represents the furthest TSD from the 5' end of the L1 and the purple triangle represents the closest TSD to the 5' end of the L1. The score assigned to each TSD is a percentage based on the 5' stop position of the TSD within the 3 kb of 5' flanking DNA.

The distance scores for the two potential TSDs are 57.55 for the green TSD and 99.75 for the purple TSD, suggesting that the purple TSD is the more likely TSD, based largely on the distance score due to the rarity of large 5' transductions. However basing the prediction of a TSD purely on the distance score is not enough, as this would exclude
Figure 4.6: Endonuclease scoring example for polyA tail 6 from accession AC002980.1_a.
CHAPTER 4. TSD DETECTION AND ANALYSIS

Figure 4.7: Distance scoring of two potential TSDs. The green triangle represents the furthest TSD from the 5' start of the L1 and so gets a lower score. The purple triangle represents the closest TSD to the 5' end of the L1 and therefore gets a higher distance score.

those hits that do have large 5' transductions. Also, since TSDs are simple direct repeats, there is also the possibility that TSDs predicted solely by distance scores alone could have occurred by chance and in fact not be TSDs at all. Those TSDs predicted purely by endonuclease score, could be incorrect as the endonuclease preceding the TSD in question could be part of another repeat found in the flanking DNA and not part of the L1 in question. To solve these issues the TSD scoring system was further developed to combine endonuclease-recognition site and distance scoring.

The TSD score for the endonuclease-recognition site was weighted at 60%, while the scoring of the distance of the 5' TSD from the start of the L1 was weighted at 40% of the overall score. The TSD scoring was arbitrarily weighted this way as it was considered that having a good endonuclease recognition site was an important characteristic of L1 insertions, and this also reflects the observation that large 5' transductions are rare. In the manually curated dataset (from Dr Richard Badge) 5' transductions >50 bp in length were found in only 9/196 (4.6%) FL-L1HS used to derive the scoring matrix.

To illustrate this weighting bias, for the example FL-L1HS, AC002980.1-a, there are two potential TSDs that pass the filter for mismatches and percentage ‘A’ identity and so can be scored using the endonuclease and distance-scoring methods. Examples of scoring from the two TSDs generated from polyA tail sequence 1 and polyA tail...
sequence 6 can be seen in Table 4.4.

Table 4.4: Comparison of two TSDs that pass the mismatch and percentage ‘A’ cut off of ≤10% and ≤80%.

<table>
<thead>
<tr>
<th>TSD scoring examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotation</td>
</tr>
<tr>
<td>Endonuclease cut site</td>
</tr>
<tr>
<td>Endonuclease start</td>
</tr>
<tr>
<td>Endonuclease stop</td>
</tr>
<tr>
<td>Endonuclease score</td>
</tr>
<tr>
<td>Endonuclease score (60%)</td>
</tr>
<tr>
<td>TSD</td>
</tr>
<tr>
<td>TSD start</td>
</tr>
<tr>
<td>TSD stop</td>
</tr>
<tr>
<td>TSD distance score</td>
</tr>
<tr>
<td>TSD distance score (40%)</td>
</tr>
<tr>
<td>Total score</td>
</tr>
</tbody>
</table>

The TSD generated from polyA tail 6 (TSD: AAAAAAAATCACCA) gives the highest total score of 99.89 whilst the TSD generated from polyA tail 1 (TSD: AAACAAATTTAC) scores less with a total score of 67.60, suggesting that AAAAAAAATCACCA is the TSD of AC002980.1a. The TSD AAAAAAAATCACCA matches the manually called TSD by Dr Richard Badge in the manually curated dataset (http://www.le.ac.uk/ge/ajj/LINE1/FLDB.html) suggesting the scoring system identified the correct TSD for this element. There is the remote possibility that both sources had identified the incorrect TSD, but this is unlikely as these TSD predictions agree with the TSD suggested by the program TSDfinder (Szak et al., 2002) for same the L1 (AC002980.1a).

4.8 Transductions

An important part of L1 annotation is whether the element has any transduced sequence at its 5’ or 3’ end. As transduced sequences are derived from the DNA flanking the L1 insertion site the identification of transductions can lead to the construction
of L1 families and the identification of progenitor/offspring and sibling relationships. TSDmapper identifies 5’ transductions based on the position of the TSD in relation to the 5’ L1 start as defined by BLAST and 3’ transductions based on the position of the TSD in relation to the 3’ end of the L1 as defined by BLAST. The 5’ and 3’ transductions are identified by TSDmapper using the methods below, which are called by the TSDscoring method.

Using the L1 AC002980.1.a as an example TSDmapper identifies a 4 bp 5’ transduction and a 221bp 3’ transduction which compares favourably to the manually curated 5’ and 3’ transductions (0 bp 5’ and 223 bp 3’ including all polyA tails) identified by Dr Richard Badge (http://www.le.ac.uk/ge/ajj/LINE1/FLDB.html).

4.9 Empty Sites

The next step, after the TSD scoring method is applied to potential TSDs is the sorting of TSD scores such that the top-scoring TSD can be used for further analysis.

The top-scoring TSD hit was used to create empty sites. An empty site comprises the 5’ flanking DNA preceding the L1 start, one copy of a TSD and the 3’ flanking DNA after the 3’ end of the L1. The method CreateEmptySites in TSDmapper creates the empty sites for the top TSD hits but limits the 5’ and 3’ to 250 bp each so that the empty site sequence is small enough to be efficiently searched for in the trace archive. The creation of empty sites has two main purposes: the identification of polymorphic elements in the trace archives, and the validation of the TSDs identified from TSDmapper. The empty site for the top TSD created by TSDmapper for the example AC002980.1.a can be seen in Figure 4.8.
>AC002980.1_a_bothflanks_2kb:AC002980.1_a_bothflanks_2kb_polyA_6
ATTTTAGATAAGCCACTGTATATTTGGAGCTTTCCCTCCTGCAAGGGGACTGTGTAACACATCAAAGTGGCATACTGGCTAGGAGAAGGAGTAGATGTCTCTGTTATTTATCAGTGCTGGGGCTTGGGAAATTTGTAAGCAAATACGATTACAGCTATTACATAGGATTTCCTCATATAATACATCCCTCTTAACCTCCAACACATACTGTTAAAAAATATCAACAGTGAATTCGCTCTAGAGAAATCCAGTGTCAAAAAGGCCAAGTTATTTGCAAAATTTGGTTCTATGAAAGACCTCGACCTTCAACATTCAACAAGAAAAAGATCAATTGCATGTAATTGGCTTGCAGGGTAATTTGCTGTGTCACATAATATGGTGAAGATCCCTCTCTTCGCTACAACAAGAATTAATGAAACAACACAACAAACACGATCTTTGCTTTGAATAGCAATTGCTAGAACCCTCAGGACATCATTCAAG

Figure 4.8: Empty site created by TSDmapper based on the top TSD hit AAAAAAAAATCACCA for AC002980.1_a. The 5′ flanking DNA is highlighted in purple, the endonuclease nicking site (TT]AAAAAA) is illustrated by the rectangular box and the TSD is highlighted in red text. The 3′ flanking DNA is highlighted in blue.

4.10 Comparison of TSDmapper and TSDfinder performance relative to manually called TSDs

To investigate the accuracy of TSDmapper, compared to another TSD-identifying program, TSDfinder (Szak et al., 2002) 184 FL-L1HS sequences from the database http://www.le.ac.uk/ge/ajj/LINE1/FLDB.html were used. It was required that these FL-L1HS had 2 kb of flanking DNA and their TSDs had been manually called (by Dr Richard Badge). This data set was analysed with both TSDmapper and TSDfinder.

TSDmapper and TSDfinder were both run with default parameters. For TSDmapper this meant requiring 2 kb of flanking DNA with 1 kb of the 5′ and 3′ ends of the L1 present. The polyA tail length cut off was set to 10 bp, the non-A polyA tail contamination threshold to 2, the E value cut off for LALIGN was set to 10, the percentage weighting for the endonuclease site score and TSD distance score set to 60 % and 40 % respectively, and the LALIGN percentage identity score to 99%. The top hit (i.e. best-scoring hit) of TSDmapper was used for the analysis.
The TSDfinder script available from http://www.ncbi.nlm.nih.gov/CBBresearch/Landsman/TSDfinder/TSDfinder.pl.txt was run with default parameters. The default parameters for the 5' and 3' flank were not modified (100 bp and 3 kb respectively). TSDfinder was written to process chromosomes containing FL-L1HS sequences, one at a time and assumes that each sequence file contains the complete sequence of a chromosome. TSDfinder also requires a RepeatMasker file for each chromosome containing the L1 annotation. As each FL-L1HS that was tested was not necessarily assembled into a chromosome, each FL-L1HS and the flanking sequence was treated as an individual chromosome, meaning that effectively TSDfinder was run on 184 individual chromosomes.

The comparison of the TSDmapper and TSDfinder algorithms with the 184 manually curated TSDs involved a direct comparison of the TSD call in each of the three TSD datasets to each other. Figure 4.9 illustrates the number of TSDs that were an exact match between TSDmapper, TSDfinder and the manually curated TSDs in the form of a Venn diagram. From TSDmapper, 131 TSD calls (71.1%) were determined to be exact matches to the manually curated dataset, while 120 TSD calls from TSDfinder (65.2%) were determined to be an exact match to the manually curated dataset. The combined TSD calls of both TSDmapper and TSDfinder that were exact matches to the manually curated data set was 102 (55.4%). There were 29 TSD calls (15.8%) that were found by TSDmapper to match exactly the manually curated TSD dataset and not the TSD call from TSDfinder, while 18 TSD calls (9.8%) were found by TSDfinder to match exactly the manually curated TSD dataset and not the TSD call from TSDmapper.

Applying both algorithms to the manually called TSDs, 81% of the TSD calls that are detected are exact matches. However 19% of manually called TSDs still cannot be identified by TSDmapper or TSDfinder.

4.11 TSD analysis of FL-L1HS using TSDmapper

The perl script TSDmapper.pl was run with default parameters on the 533 FL-L1HS sequences detected from the searches of genome assemblies and nucleotide accession
databases described in chapter 3. The number of FL-L1HS sequences that had a TSD called by TSDmapper was 522, with 11 FL-L1HS sequences having no TSDs that met the polyA tail selection and scoring criteria. Using the top overall score for each TSD, 520 FL-L1HS were selected for further analysis. There were 2 FL-L1HS that had two TSD calls with the same score, and so were excluded from further TSD analysis.

**PolyA tail length**

When a copy of an L1 is inserted into the genome during TPRT, the polyA tail of the L1 transcript is used as a primer for reverse transcription (Feng *et al.*, 1996; Moran *et al.*, 1996). It is well known that *de-novo* disease-causing insertions have much longer polyA tails than L1s resident in the genome for some generations, suggesting that some process (replication slippage for example) causes the reduction of polyA tract length over time. To see if there was any pattern in the polyA tail length distribution, related to element age, for the 520 TSDs predicted by TSDmapper, the polyA tail lengths were divided into 9 bins of 10 bp each.
CHAPTER 4. TSD DETECTION AND ANALYSIS

167

Figure 4.10: PolyA tail distribution of 520 FL-L1HS sequences divided into length bins of 10bp.

The distribution of polyA tail length is shown in Figure 4.10 for the 520 FL-L1HS sequences. The 10-bp categories start from 1-10 bp and finish at 81-90 bp. The majority of polyA tail lengths fall in the bin 11-20 bp (156) closely followed by 21-30 bp (152). The mean length of the polyA tail in the 520 FL-L1HS data set is 27 bp.

To investigate whether there was a difference between the length of polyA tail sequences of those elements with intact ORFs (i.e. those elements that are likely to be younger and still capable of retrotransposition) and those with disrupted ORFs, the polyA tail lengths of FL-L1HS sequences with intact ORFs and non-intact ORFs were analysed.

Using the definition described in chapter 3 for determining if an FL-L1HS has intact ORFs, 152 FL-L1HS sequences were deemed to have intact ORFs (29.2%), with 368 having non-intact ORFs (70.8%) from the 522 FL-L1HS sequences analysed.

The distribution of polyA tail length appears to be different for FL-L1HS sequences that have intact ORFs from those with non-intact ORFs. Figure 4.11 illustrates the
CHAPTER 4. TSD DETECTION AND ANALYSIS

Distribution of polyA tails for 152 FL-L1HS sequences with intact ORFs and 368 FL-L1HS sequences with non-intact ORFs divided into length bins of 10 bp.

The distribution of FL-L1HS polyA tails with intact ORFs and non-intact ORFs. The number of FL-L1HS sequences appears to increase as the polyA increases (up to 40 bp) for elements with intact ORFs, while after 40 bp the number decreases. For those FL-L1HS sequences that have non-intact ORFs there is an initial increase in the number of FL-L1HS sequences as the polyA tail length increases from 1-10bp to 11-20bp. After this, the number of FL-L1HS sequences decreases as the length of the polyA tail decreases.

The mean polyA tail length for the 155 FL-L1HS with intact ORFs was 32 bp, while the mean length for the 368 FL-L1HS sequences with non-intact ORFs was 25 bp. To test whether this difference in means was significant, a t-test was performed.

**Null hypothesis**

*FL-L1HS sequences with intact ORFs and non-intact ORFs have identical mean polyA tail lengths. That is, the difference between the means is zero.*
t-test

Mean polyA tail length non-intact ORF = 25.1685 (stdv 11.6167) Mean polyA tail length intact ORF = 32.3421 (stdv 14.7989)

non-intact ORF count= 368, intact ORF count= 152, t = 5.892, Degrees of Freedom = 518, p ≤6.937x10^{-09}

The null hypothesis that intact ORFs and non-intact ORFs have identical mean polyA tail lengths can be rejected (p ≤6.937x10^{-09}). Therefore FL-L1HS sequences with intact ORFs have significantly larger polyA tails than those FL-L1HS sequences with non-intact ORFs.

The result that is consistent with the idea that those elements with disrupted ORFs are likely to have been present in the genome longer than L1s with intact ORFs. It appears therefore that the longer an L1 is resident in the genome the shorter its polyA tail.

Characteristics of TSDs indentified by TSDmapper

To determine the characteristics of TSDs identified by TSDmapper, an analysis of TSD length and composition was performed using the 520 FL-L1HS data set.

The minimum TSD length observed in the 520 FL-L1HS data was 10 bp occurring in 64 FL-L1HS sequences. The most frequent (modal) TSD length observed was 14 bp and this occurred in 98 FL-L1HS sequences. The longest TSD observed was 47 bp and occurred in 1 FL-L1HS element. The distribution of the length of TSD s identified by TSDmapper is shown in Figure 4.12. The bimodal distribution of TSD lengths observed might suggest that the large number of TSDs identified as being 10 bp in length may be artefactual, resulting from the parameters used by the algorithm (in particular the minimal TSD size cutoff). The average length of a TSD from the 520 FL-L1HS sequences was 15 bp. These data are comparable to the 14-bp average length detected in vitro (Szak et al., 2002).
CHAPTER 4. TSD DETECTION AND ANALYSIS

Figure 4.12: Distribution of TSD lengths from 520 FL-L1HS sequences identified by TSDmapper.

To see if the TSDs identified by TSDmapper conform to any consensus, a consensus TSD logo was generated by WebLogo (Crooks et al., 2004; Schneider and Stephens, 1990) (http://weblogo.berkeley.edu/). The consensus TSD sequence is shown in Figure 4.13 and illustrates that the first six base pairs (1-6) are non-random which is an overrepresentation of A nucleotides. This sequence conforms to the endonuclease cut site consensus (TT↓AAAA) at positions 1-6 bp.

**Endonuclease recognition site consensus**

The 5’ endonuclease recognition cut-site sequence used for the scoring matrix in TSDmapper suggested an octamer TT↓AAAAAA consensus. To confirm if this was indeed the case an analysis of the endonuclease cut-sites of 520 FL-L1HS sequences was performed to generate a sequence logo to represent the consensus. All 8 bp were processed through the nucleotide logo generator WebLogo (http://weblogo.berkeley.edu/).
edu/) to produce a graphical representation of the endonuclease cut-site nucleic acid consensus. The results from Weblogo are illustrated in Figure 4.14.

Figure 4.14 shows that the predicted octamer consensus of TT↓AAAAAA, is detected in the 520 FL-L1HS sequences (as expected) as this sequence was used to weight the scoring of the TSDs. Positions 1-2 in Figure 4.14 conform to the predicted consensus of positions -1 to -2 in the endonuclease recognition sites (TT), positions 3-8 conform to the predicted consensus of +1 to +6 in the consensus based on a smaller dataset of 197 FL-L1HS sequences. Positions +5 and +6 bp (7 and 8 in the figure) are the weakest positions in the consensus but these results still suggest that an octamer is the ideal consensus recognition site for the L1 endonuclease.

There is evidence that the L1 endonuclease makes a second cut in the genomic DNA during the process of TPRT. To see if there is a consensus sequence for the 3′ endonuclease cut site, 187 TSDs from the two most frequently observed TSD length occurrences (14 and 15 bp) were reverse complemented so that the TSD was in the orientation 3′→5′ to represent the TSD on the antisense strand. The reverse complemented TSD is illustrated in Figure 4.15.
CHAPTER 4. TSD DETECTION AND ANALYSIS

Figure 4.14: A sequence logo generated by WebLogo of 520 FL-L1HS endonuclease cut sites detected by TSDmapper

Figure 4.15: TSD consensus (3’ to 5’) for TSDs that are 14 and 15 bp in length. At the end of the 5' TSD the +6 bp endonuclease cut site is clearly visible (TTTTTT positions 10-15).
To get a better idea of the potential 3’ endonuclease consensus at the 3’ end, the final 6 bp of the reverse complemented TSD were removed from the analysis as these base positions represented the 5’ 1-6 bp positions of the 5’ endonuclease cut site and conform very strongly to the 5’ endonuclease cut site sequence, as previously reported. Figure 4.16 represents the revised sequence logo for the 3’ endonuclease cut site and illustrates that at the 3’ endonuclease does not appear to have a strong consensus, unlike the first 5’ endonuclease cut site, however a possible consensus site could be 3’-T[ACTG]↓A[ACGT]AAAA-5’.

Transductions

By accurately locating the most likely set of TSDs of an inserted L1 element, TSDmapper also has the ability to detect transductions (TDs) at the 5’ and 3’ end of an element i.e. to identify situations where the 5’ and 3’ TSDs are not immediately adjacent to the beginning or end of the L1 sequence.
The following sections discuss 5’ and 3’ transductions generated from the TSDmapper results.

**Five prime transduction analysis**

5’ transductions from FL-L1HS sequences can be caused by a number of different factors, such as upstream initiation of transcription and polyG tract variation due to reverse transcription of the L1 transcript’s 7-methylG cap.

Analysis of 520 FL-L1HS sequences revealed 29.4% (153) do not have any 5’ transductions. This means that 367 FL-L1HS could potentially have 5’ transductions. Two hundred and seventy-five FL-L1HS sequences (52.9%) have reported transductions varying in size from 1-50 bp. However 96.7% of the 275 FL-L1HS sequences from this category (266) are less than 20 bp in length and so cannot be validated as genuine 5’ transductions due to the inability to map sequences of this small size uniquely in the genome. Also 5’ transductions that fall into this bin size may also result from variation in the 5’ polyG tract at the start of FL-L1HS, very slight inaccurate initiation of transcription or be artefacts of the TSDmapper algorithm. Therefore only 9 FL-L1HS sequences with 5’ transductions of 21-50 bp (3.3%) could possibly be mapped to unique genomic location, and hence validated. TSDs ≥ 51 bp in length comprised 92 FL-L1HS sequences (17.7%) of the 520 FL-L1HS data set. The largest 5’ transduction observed was 1979 bp.

**Three prime transduction analysis**

Analysis of 520 FL-L1HS sequences for 3’ transductions revealed that 492 (94.6%) have some transduced sequence at the end the 3’ end of the L1 sequence. Twenty eight FL-L1HS sequences (9.4%) were found to have no detectable 3’ transductions. Three hundred and one (57.9%) FL-L1HS sequences were found to have 3’ transductions from 1-50 bp. Only 1 FL-L1HS had a 3’ transduction below 10bp. 56 FL-L1HS sequences had 3’ transductions of 7-20 bp. Two hundred and forty-four FL-L1HS sequences has 3’ transductions that between 21-50 bp that could be potentially mapped to the genome. One hundred and ninety one FL-L1HS sequences (36.7%) had a 3’ transduction ≥51bp in length. The largest 3’ transduction detected by TSDmapper was 1945 bp.
Origins of transduced sequences

To determine the possible origins of the transduced sequence for 5’ and 3’ transductions in the human genome, and to find relationships between FL-L1HS sequences, all transduced sequences ≥ 51 bp were RepeatMasked and mapped to the hg18.2bit BLAT database using gfClient and gfServer (see chapter 2). The number of hits that were ≥ 98% identity and ≥ 51 bp in length from the BLAT search was 147.

Each FL-L1HS was considered to be part of a family or lineage if the transduced sequence mapped to more than one location in the genome. Table 4.5 illustrates a selection of multiply located transduced sequences that have been manually verified by locating the transduced sequence using the UCSC genome browser indicating that these sequences are part of an L1 family/lineage.

Table 4.5 shows the positions of transduced sequences that map to a site in the genome where a known FL-L1HS is present. All sequences in Table 4.5 are 3’ transductions. There were 5 families identified by this semi-automated transduction mapping method. Two L1 families, 1 and 4 were previously reported by Szak et al. (2003) while families 2, 3, and 5 were not. In two of the L1 families it is possible to identify a parent or progenitor L1. An example of an L1 family (L1 family 2) is shown in Figure 4.17.

BL000190 from chromosome 5 is a possible progenitor of BL000051 located on chromosome 2 since analysis by TSDmapper revealed that the 232 bp from BL000051 contains the 34-bp TD from BL000190. Since the BL0000190 TSD ends after the 34-bp transduction it is predicted that this L1 retrotransposed and moved the 234 bp transduction to its new genomic location. An alignment of the two transductions is illustrated in Figure 4.18.

Transduction analysis has therefore revealed three potential relationships between L1s that were not previously reported. There were no 5’ transduction families identified in this analysis.

Analysis of transductions that do not map to a known FL-L1HS are shown in Table 4.6. The transduction analysis has identified 9 sites in the genome that could potentially contain a polymorphic FL-L1HS, or are from a FL-L1HS that is only present at low
Figure 4.17: A UCSC view of the 234 bp 3’ transduction from BL000051 mapping to two different locations in the human genome.
Figure 4.18: Alignment of the transductions from BL000190 and BL000051 showing that both L1 transduction sequences share a 34-bp piece of genomic DNA and that BL000051 has the 3’ TSD of BL000190 (green and red text). The 3’ transduction of BL000051 maps to a location immediately past the 3’ end of the TSD of BL000190 allele frequencies or no longer exist in human populations. Seven of the empty-site transductions are 3’ transductions and the other two are 5’ transduction empty-sites. Identification of these sites provides a potential starting point for the detection of novel FL-L1HS by prospective PCR assays. An example of an empty-site transduction is illustrated in Figure 4.19.
Table 4.5: 3' transduction families identified from the TSDmapper results. *-Truncated L1HS (656) * transduction start stop; \(^{a}\) Szak family 5, \(^{b}\) Szak family 6

<table>
<thead>
<tr>
<th>baseLINE family ID</th>
<th>baseLINE family ID</th>
<th>Chromosome location</th>
<th>Genome start</th>
<th>Genome stop</th>
<th>TD 3' length (bp)</th>
<th>Parent FL-L1HS</th>
<th>TSD sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1(^{a})</td>
<td>BL000075</td>
<td>2</td>
<td>86735033</td>
<td>86735890</td>
<td>858</td>
<td>AAAAAGCTGCTATGC</td>
<td></td>
</tr>
<tr>
<td>1(^{a})</td>
<td>BL000522</td>
<td>X</td>
<td>82684062</td>
<td>82690066</td>
<td>31</td>
<td>Y</td>
<td>GAAATCTATAACCTT</td>
</tr>
<tr>
<td>2</td>
<td>BL000051</td>
<td>2</td>
<td>167553212</td>
<td>167559226</td>
<td>31</td>
<td>Y</td>
<td>TTTAAAATT</td>
</tr>
<tr>
<td>2</td>
<td>BL000190</td>
<td>5</td>
<td>166326385</td>
<td>166332397</td>
<td>34</td>
<td>AAAAATTTGAAGGCTT</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>BL000321</td>
<td>9</td>
<td>112593200</td>
<td>112599210</td>
<td>187</td>
<td>AAAAGTTTCAG</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>BL000364</td>
<td>11</td>
<td>108553452</td>
<td>108559460</td>
<td>68</td>
<td>AAGAAGGGCATTTTCAG</td>
<td></td>
</tr>
<tr>
<td>4(^{b})</td>
<td>BL000266</td>
<td>7</td>
<td>69300255</td>
<td>69306259</td>
<td>638</td>
<td>AAAGACACTG</td>
<td></td>
</tr>
<tr>
<td>4(^{b})</td>
<td>BL000078</td>
<td>2</td>
<td>112977881</td>
<td>112983890</td>
<td>171</td>
<td>Y or sibling</td>
<td>AAGAGTTCAGACCA</td>
</tr>
<tr>
<td>4(^{b})</td>
<td>N/A</td>
<td>5</td>
<td>140466369*</td>
<td>140466972*</td>
<td>603</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>BL000265</td>
<td>7</td>
<td>65389296</td>
<td>65395302</td>
<td>117</td>
<td>TATTTTCCTGGGGGA</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>N/A</td>
<td>X~</td>
<td>94743770*</td>
<td>94743880*</td>
<td>117</td>
<td>N/A</td>
<td></td>
</tr>
</tbody>
</table>
Table 4.6: Transduction sites that do not map to other known FL-L1HS sequences in the human genome

<table>
<thead>
<tr>
<th>baseLINE Identifier</th>
<th>Chr location</th>
<th>TD 5' or 3'</th>
<th>Transduction (TD) Length</th>
<th>TD start</th>
<th>TD stop</th>
<th>TSD sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>BL000273</td>
<td>7</td>
<td>3'</td>
<td>484</td>
<td>111396496</td>
<td>111396979</td>
<td>AAAAATTACTGTCTA</td>
</tr>
<tr>
<td>N/A</td>
<td>18</td>
<td>3'</td>
<td>792</td>
<td>61596746</td>
<td>61597537</td>
<td>N/A</td>
</tr>
<tr>
<td>N/A</td>
<td>13</td>
<td>3'</td>
<td>227</td>
<td>59668226</td>
<td>59668452</td>
<td>N/A</td>
</tr>
<tr>
<td>BL000124</td>
<td>3</td>
<td>3'</td>
<td>303</td>
<td>90000604</td>
<td>90000906</td>
<td>AAAAGCATTTCTAC</td>
</tr>
<tr>
<td>N/A</td>
<td>15</td>
<td>3'</td>
<td>254</td>
<td>46034959</td>
<td>46035212</td>
<td>N/A</td>
</tr>
<tr>
<td>BL000297</td>
<td>8</td>
<td>3'</td>
<td>162</td>
<td>92646073</td>
<td>92646234</td>
<td>AAAAATGTCAGGCATC</td>
</tr>
<tr>
<td>N/A</td>
<td>4</td>
<td>3'</td>
<td>123</td>
<td>141311774</td>
<td>141311896</td>
<td>N/A</td>
</tr>
<tr>
<td>BL000359</td>
<td>11</td>
<td>3'</td>
<td>221</td>
<td>90345116</td>
<td>90345336</td>
<td>AAAAAAAATTCCA</td>
</tr>
<tr>
<td>N/A</td>
<td>4</td>
<td>3'</td>
<td>205</td>
<td>152549863</td>
<td>152550067</td>
<td>N/A</td>
</tr>
<tr>
<td>BL000497</td>
<td>X</td>
<td>3'</td>
<td>336</td>
<td>129926592</td>
<td>129926927</td>
<td>AAAGAAAAAAATG</td>
</tr>
<tr>
<td>N/A</td>
<td>5</td>
<td>3'</td>
<td>304</td>
<td>85660648</td>
<td>85660951</td>
<td>N/A</td>
</tr>
<tr>
<td>BL000493</td>
<td>X</td>
<td>5'</td>
<td>260</td>
<td>118459485</td>
<td>118459744</td>
<td>AAAACATTTT</td>
</tr>
<tr>
<td>N/A</td>
<td>6</td>
<td>5'</td>
<td>215</td>
<td>114153049</td>
<td>114153265</td>
<td>N/A</td>
</tr>
<tr>
<td>BL000269</td>
<td>7</td>
<td>5'</td>
<td>119</td>
<td>97080789</td>
<td>97080907</td>
<td>AAATCAATATTTGG</td>
</tr>
<tr>
<td>N/A</td>
<td>10</td>
<td>5'</td>
<td>116</td>
<td>1926088</td>
<td>1926203</td>
<td>N/A</td>
</tr>
<tr>
<td>BL000440</td>
<td>16</td>
<td>3'</td>
<td>163</td>
<td>52639520</td>
<td>52639683</td>
<td>AAAGAAAGCTAAGATTTG</td>
</tr>
<tr>
<td>N/A</td>
<td>12</td>
<td>3'</td>
<td>77</td>
<td>15584183</td>
<td>15584260</td>
<td>N/A</td>
</tr>
<tr>
<td>BL000012</td>
<td>1</td>
<td>3'</td>
<td>426</td>
<td>208153354</td>
<td>208153779</td>
<td>AAAGAAAGAAAAATGG</td>
</tr>
<tr>
<td>N/A</td>
<td>5</td>
<td>3'</td>
<td>378</td>
<td>78750973</td>
<td>78751350</td>
<td>N/A</td>
</tr>
</tbody>
</table>
Figure 4.19: A UCSC view of the BL000012 transduction mapping to a site in the human genome where no FL-L1HS is reported in the genome assembly.
Empty site validation

One unique feature of TSDmapper is the ability to perform computational reconstruction of FL-L1HS empty-sites. Constructing accurate empty-sites is a useful feature that serves two purposes. Firstly if an exact match to the reconstructed empty site is found in a nucleotide or trace database then the empty-site reconstruction is accurate and the TSD predicted by TSDmapper is likely to be the correct TSD. Secondly, identification of an empty-site sequence in a database may allow the identification of elements as being polymorphic, without the need for genotyping.

To validate the empty-sites generated by TSDmapper the *in-silico* empty-sites of 522 FL-L1HS that were identified as having TSDs were searched against the trace archive. The trace archives were chosen as they would provide potentially more empty sites than just searching the genome assemblies or the nt database (see ??).

Since the empty-site generated by TSDmapper consists of 250 bp of 5' flanking DNA, one copy of the predicted TSD and 250 bp of 3' flanking DNA, all matches were required to be $\geq 290$ bp in length and $\geq 98\%$ to the reconstructed empty-site. This was done so that one copy of the TSD would most likely be included in the resulting hit.

Analysis of the results from the trace archive empty-site search revealed that TSDmapper had identified 104 empty sites from this source. To compare the overlap of the trace archive search to the previous nt search, each FL-L1HS' empty site corresponding baseLINE identifier was analysed using the Venn diagram creator web program available at http://www.pangloss.com/seidel/Protocols/venn.cgi. The results are illustrated in Figure 4.20. The results suggest that TSDmapper is accurate at predicting the empty sites of FL-L1HS sequences with 104 empty sites being detected from the trace archives alone. When the trace archive and the nt search are compared to each other, 85 FL-L1HS empty sites are found to be common between the two datasets, with an additional 19 empty sites detected from the searching of the trace archive. Also, 3 empty sites were not found in the trace archive but only in the nt database. As a result of TSDmapper’s findings from the trace archive search, the number of FL-L1HS that are predicted to be polymorphic by computational methods that are present in assembled
sequence databases (Reference assembly, Celera, HuRef and nt database) has increased to 107, an increase of 19 L1s from the initial search of the nt database which found that 88 FL-L1HS sequences were polymorphic. Thus TSDmapper has helped to computationally predict that approximately 20% of all known FL-L1HS sequences from assembled sequence databases appear to be polymorphic. A list of polymorphic FL-L1HS sequences predicted by TSDmapper is available in the electronic appendix, in the Excel file empty sites.xls.

**TSDmapper - other applications**

Since TSDmapper was designed to identify TSDs by endonuclease and distance scores, it is possible to use TSDmapper to identify TSDs in other transposon sequences. As Alu (a non-autonomous retroelement) uses the ORF2 protein from L1HS to mobilise itself, the TSD endonuclease and distance scoring system could in principle be applied to finding Alu TSDs, as well as L1 TSDs.

TSDmapper was run on 377 AluYa5 sequences with manual TSD calls downloaded from dbRIP, using the same default parameters as used for finding FL-L1HS, with a few minor alterations. The minor alterations included only requiring 390 bp of 5′ and...
3’ flanking DNA with 125 bp of 5’ and 3’ Alu sequence. The scoring matrix used for endonuclease scoring was the same as the one used for FL-L1HS detection.

The Alu sequences were extracted from the dbRIP record XML files and the 5’ and 3’ sequences were converted into the flanking DNA sequence files recognised by TSDmapper. TSDmapper indentified 53.3% (201) that were an exact match to the manual call made in the dbRIP XML file. Therefore TSDmapper can computationally identify TSDs from other non-LTR elements found in the human genome using a human-specific TSD scoring matrix developed for FL-L1HS sequences.

4.12 Discussion

This chapter describes the design and implementation of an informatic resource, TSDmapper (a collection of Perl scripts) that was developed for the automated detection and analysis of L1 retrotransposon Target Site Duplications (TSDs) and transductions. The performance of TSDmapper and another TSD-finding algorithm, TSDfinder (Szak et al., 2002) were compared to each other, and also to a manually curated TSD dataset. The TSDs, polyA tails and transductions of the 533 distinct FL-L1HS sequences described in chapter 3 were also analysed using TSDmapper.

TSDmapper design

The algorithm for TSDmapper was developed to meet a number of requirements 1) sequence validation 2) the identification of polyA tails 3) detection of 5’ and 3’ TSDs 4) the scoring of TSDs based on knowledge of the biological process of Target Primed Reverse Transcription (TPRT) 5) the identification of 5’ and 3’ transductions 6) the automated reconstruction of empty sites.

The implementation of the TSDmapper algorithm requires the input sequences to be in a specific format, with the 5’ flanking DNA preceding the FL-L1HS sequence to be first in the file followed by the 3’ flanking DNA following the 3’ end the L1. The validation that these sequences conformed to this format was performed by the sequence comparison algorithm blast2seq using the sequence of a known active L1 - L1.3_mpa. This approach
was chosen because blast2seq is a widely used algorithm for sequence comparison, and was also used in TSDfinder. However the way this external program is implemented in TSDmapper (called by a system command in perl) could probably be done more effectively using a Bioperl module such as Bio::Tools::dbAlign.pm (http://search.cpan.org/~cjfields/BioPerl-1.6.0/Bio/Tools/Align.pm). This perl module is an extension to perform dynamic programming based pairwise comparisons (local and global alignments). Using this program would be quicker and more efficient than blast2seq as all results would be stored as bioperl objects, which would make the parsing and further sequence manipulation more rapid and straightforward. The current implementation requires having to open and parse the result files individually.

The polyA-tail discovery algorithm was based extensively on the Perl code for finding polyA tails used in TSDfinder, with only minor modifications made to the code so it would work with the TSDmapper code. This algorithm has proved to be very efficient at finding polyA tails and so could probably not be greatly improved by re-writing.

Although the detection of the polyA tails in TSDmapper uses the same algorithmic process as TSDfinder, the rest of the script uses a different approach to find TSDs. Once the polyA tail has been discovered, a polyA tail TSD search sequence is constructed from the end position of the polyA tail and extending back 35 bp and extending forward 35 bp, using TSDmapper. This was done to include all of the potential TSD. This sequence was then used to search for TSDs in the 5’à flanking DNA. Finally, the sequence is then used by the program LALIGN to detect potential TSDs in the 5’à flanking DNA.

**Comparison of TSDmapper and TSDfinder to manually called TSDs**

Analysis of 196 FL-L1HS sequences leads us to propose a larger endonuclease cut site sequence consensus. This consensus TT↓AAAAA (8 and 20 bp analysis) or TT↓AAAAAAA (20 bp analysis) agrees largely with the consensus proposed by Jurka (1997) for Alu of TT↓AAAA. It therefore appears for FL-L1HS that a heptamer or octamer is the size best representative of the endonuclease cut site sequence. The first two base-pairs immediately preceding that of the TSD sequence still have a very strong consensus for TT, as proposed by Jurka (1997), which our newer analysis confirms.
One of the major differences between the two algorithms is the scoring of TSDs. TSDmapper extracts the endonuclease cut site (8 bp) sequence from around the predicted TSD (-2 – +6 bp) and scores this sequence using an L1-specific matrix of the most commonly occurring base pairs at each of the eight positions.

The TSDmapper algorithm reflects the biological mechanism of L1 insertion compared to the TSDfinder program which does not use this method. The algorithms are similar in the way that they score the position of the TSD in relation to the start of the L1 sequence. Both algorithms give a greater score to TSDs that are closer to the 5’ of the L1. The TSDfinder algorithm also scores TSDs higher if they are closer to the 3’ end of the L1, despite the fact that this could discriminate against large 3’ transductions.

Another difference between TSDmapper and TSDfinder is the amount of 5’ flanking DNA used to find TSDs. TSDfinder only uses 100 bp whereas TSDmapper used 2 kb. TSDmapper therefore allows the detection of large 5’ transductions (≥100 bp) which would not be found using TSDfinder, when run with the default parameters. Lander et al. (2001) found two large transductions of 143 and 215 bp that TSDfinder would not have detected. The inability to detect these larger transduction events would mean that TSDfinder would probably either report the incorrect pair of TSDs or report no TSDs at all. Thus TSDmapper has the ability to detect 5’ transductions that TSDfinder would never find, increasing its usefulness when analysing FL-L1HS sequences.

TSDmapper also has the advantage over TSDfinder in that it can be run on individual L1 sequences, or on large batches. TSDmapper does not operate only on chromosomes from genome assemblies thus being more flexible since any sequence from many different sources can be submitted. TSDmapper can also output reconstructed empty-sites, which TSDfinder cannot, again making TSDmapper a useful tool in in-silico genotyping analysis. The reconstructed empty-site sequences can be searched against the trace archives and the nt databases to see if the FL-L1HS is indeed polymorphic in human populations. The reconstruction of these empty sites could also allow the confirmation that the FL-L1HS is indeed human-specific, as the Chimpanzee genome assembly and the trace files could also be searched with this sequence.
A direct comparison of TSDmapper to TSDfinder was performed on 184 FL-L1HS whose TSDs had been selected by Dr Richard Badge (Department of Genetics, University of Leicester) (http://www.le.ac.uk/ge/ajj/LINE1/). Both algorithms were run with default parameters. TSDmapper found 131 (71.1%) of TSDs (exact matches) called in the manually curated TSD dataset, whilst TSDfinder found 120 (65.2% exact matches) meaning that for this dataset TSDmapper found 11 (5.9%) more TSDs (exact matches) than TSDfinder. The overlap between the two algorithms was 102 (55.4%) TSDs that were exact matches to the manually called datasets. TSDmapper found 29 TSDs that TSDfinder did not find and TSDfinder found 18 TSDs not found by TSDmapper. Variation was to be expected in the datasets from TSDmapper and TSDfinder due to their different TSD detection and scoring algorithms. Overall both algorithms found 81% of all manually called TSDs, suggesting that both algorithms could be applied for accurate automated detection of TSDs.

An observation made when manually inspecting some of the TSD calls generated by TSDmapper was that some of the TSD called were ‘fuzzy’ at the end of the TSD by 1-2 nucleotides suggesting that TSDmapper had called the correct TSDs but, because it was not an exact match to the manually curated TSD, it was rejected as non-matching. Therefore it is possible that TSDmapper does in fact find more than 71.1% of all manually called TSDs. These ‘fuzzy’ nucleotides could be a result of the way the datasets were generated. The manually curated TSD dataset was generated by hand, with no computer scripts to aid detection of the TSD sequence. There is the possibility of human error in some of the TSD calls with the addition or deletion of one or two base pairs at the end. In the TSDmapper script the program that detects TSDs, LALIGN, allows mismatches and gaps in the TSDs (albeit small ones). This may have contributed to the ‘fuzzy’ TSD ends. A solution to this, if the analysis was re-run, could be to allow mismatches and gaps at the ends of the TSD (90% of the TSD length).

The TSDmapper program was run on 520/533 distinct FL-L1HS discovered during the searching and gathering analysis (see chapter 3). Eleven FL-L1HS were excluded because they had no TSDs detected by TSDmapper that met the minimum criteria. These 11 FL-L1HS need to analysed further to see if they are the result of endonuclease-independent events (Chen et al., 2005; Morrish et al., 2002; Sen et al., 2007). Two FL-L1HS sequences had two different copies of TSDs with the same score so were also
excluded. The polyA tails, endonuclease cut site, TSD lengths and transductions of the remaining elements were analysed in greater detail, and these are discussed below.

**PolyA tails**

PolyA tails are known to be used as a primer for reverse transcription (Feng et al., 1996; Moran et al., 1996) during TPRT, so the size distribution of the polyA tails detected in 520 FL-L1HS sequences were analysed. The mean length of the polyA tails detected in this data set was 27 bp. The majority of polyA tail lengths fell into the bin size of 11-20bp (156) and 21-30 (152). As polyA tail length can be an indication of activity in Alu (Roy-Engel et al., 2002) the polyA tail lengths of L1 elements with intact and non-intact were analysed (152 intact, 368 non-intact), revealing that there is a significant difference in the mean (32 bp intact, 25 bp non-intact) polyA tail lengths between the two classes of L1 elements. Having intact ORFs is a likely indication of retrotransposition and being in the genome for a shorter period of time, and so it appears that the polyA tail length gets shorter the longer an L1 is resident in the genome. These observations agree with an Alu analysis by Roy-Engel et al. (2002) and a L1 analysis by Ovchinnikov et al. (2001). Therefore longer polyA tails appear to be a hallmark of young FL-L1HS sequences that are potentially capable of retrotransposition.

**TSDs**

The bimodal length distribution of the TSDs, detected by TSDmapper suggests that the large number of TSDs around 10 bp in length may be artificial. This may be due to the minimum percentage cut off for TSD identity. However the mean TSD length for this dataset (15 bp) agrees strongly with the mean length found by Szak et al. (2002) (14 bp). Therefore it appears that the mean length of TSDs in the human genome is between 14-15 bp in length.

**Endonuclease cut site**

Analysis of the TSD’s sequence composition suggests that the first six base pairs conform to the endonuclease consensus proposed in chapter 4 and that a sequence composition analysis of the expected endonuclease cut sites reveals a consensus of TT↓AAAAA agreeing with the new proposed consensus for FL-L1HS. In addition,
the second target DNA cut required for L1 insertion and proposed by Jurka (1997) for Alu was unable to be detected in the analysis of FL-L1HS endonuclease sites. Szak et al. (2002) found a weak consensus site of TT/AAAA for L1 sequences which was not detected in this data set. The best possible consensus that could be proposed is 3’-T[ACTG]↓[ACGT]AAAA-5’. This consensus is not as strong as the consensus for the 5’ cut site. This may suggest that the second cut by the L1 endonuclease does not have to be so specific.

Transductions

The accurate detection of 5’ and 3’ TSDs allows the discovery of transductions at the 5’ and 3’ ends of L1s. The number of FL-L1HS sequences with 5’ transductions was found to be 367 (71%). However the majority of these were small transductions (1-50 bp), most of which will be impossible to validate as real transductions because their small sizes do not allow mapping to a unique location in the genome. Nine of the 5’ transductions could be potentially mapped to discrete genomic locations (21-50 bp) along with 92 ≥51 bp. The program used to map sequences to the genome in this analysis was BLAT which can only detect sequence matches of 95% or greater similarity with sequence length being 25 bases or more meaning that even if the 266 transductions ≤20 bp were real they could never be located in the genome using this algorithm.

Possible reasons for a high number of small transductions include upstream initiation of transcription, and polyG tract variation from the reverse transcription of the L1 transcript’s 7-methylG cap. Also the high number could be due to the TSD-finding algorithm not detecting the true start of the L1 correctly.

Analysis of FL-L1HS sequences revealed that 492/520 have 3’ transductions of various lengths suggesting that 3’ transductions are more common than 5’ transductions. In contrast to the 5’ transductions, 244 FL-L1HS sequences could potentially be mapped to the genome.

The transduced sequences were mapped to NCBIb36.3 to see if any progenitor/sibling relationships could be found. Each FL-L1HS was considered to be part of a family if the transduced sequence mapped to more than one location within the genome, revealing
potential transduction families. Two of these families have previously been found in an analysis by Szak et al. (2003). This new analysis has found 3 previously unreported transduction families, all of which were 3′ transductions.

If we consider the number of FL-L1HS sequences found and the number of transductions reported by TSDmapper, there does appear to be a relatively low number of transduction families detected along with only small transduction families. The largest 3′ transduction family detected was transduction family 4 (see chapter 4) with the transductions mapping to 3 sites in the genome. This transduction family was also detected by Szak et al. (2003). As only one transduction family was found to have 3 sites, it suggests that genome assemblies and sequence databases are not the ideal place to look for these transduction families. A possible explanation for this could be that the progenitor elements are active (and therefore are young elements) and may have been removed from the genome by recombination (i.e. they do not exist any more) due to the deleterious effects they would have (Boissinot et al., 2001). Another explanation could be that because the progenitor FL-L1HS is polymorphic it could be of low allele frequency and so does not exist in sequence databases available. It is therefore clear that to increase the number of FL-L1HS transduction families more FL-L1HS sequences are required (either through mining of trace archives, or experimental techniques such as ATLAS (Badge et al., 2003)) because of the low hit rate in finding transduction families from the sequences already analysed.

Empty site validation

Validation of TSD empty sites generated by TSDmapper was performed on the nt accession database and the trace archives. The number of empty sites returned from the trace archives (104) and the nt database (88) suggests that TSDmapper is accurate for reconstructing empty sites. The combination of the two different datasets suggests that 20% (107) of FL-L1HS sequences detected in these analyses are polymorphic from computational mining of sequence databases. Finding 107 FL-L1HS that are potentially polymorphic from mining of sequence databases is an increase in the number that are reported to be polymorphic at dbRIP (http://dbrip.brocku.ca/) (Wang et al., 2006). This resource reports 76 L1HS polymorphisms that are ≥5922 in length. Therefore using TSDmapper to create empty sites is a useful way of increasing the number of FL-
L1HS sequences that could potentially be polymorphic. These can then be validated in genotyping assays.
Chapter 5

baseLINE - A FL-L1HS database

5.1 Introduction

A new full-length human-specific L1 (FL-L1HS) database is needed to bring together all important FL-L1HS annotation data and integrate new annotations and discoveries. Current large genetic variation resources such as dbSNP, and HGVSbaseG2P provide little or no support for the annotation and curation of FL-L1HS. Therefore a smaller L1-specific database and web resource, sometimes termed a ‘boutique’ database, was needed to bring to the attention of the scientific community the level of human genetic variation caused by FL-L1HS sequences. This new resource, presented in the chapter, is baseLINE – A Full-length Human Specific LINE-1 database.

It is important that baseLINE is specifically a full-length human-specific LINE-1 database (FL-L1HS) as current L1 databases, dbRIP (Wang et al., 2006) and L1Base (Penzkofer et al., 2005), only reflect a small proportion of the FL-L1HS variation that is present in human populations. This is because of the way they search for and gather insertions. dbRIP collects L1 insertions from experimental studies, with the majority of the L1 insertions being truncated. L1Base collects FL-L1HS insertions from the Reference assembly which does not reflect the true level of FL-L1HS sequence variation present in human populations, due to the fact that some FL-L1HS insertions are variably present or absent in the assembly.
The FL-L1HS insertions gathered for baseLINE were computationally mined from three human genome assemblies, The Reference human genome assembly (NCBI build 36.3), the Celera genome, and the diploid sequence of a single individual (HuRef). Also, since the nucleotide accession database includes sequences not found in human genome assemblies the nucleotide accession database (April 2008 freeze) was also searched for FL-L1HS sequences (see chapter 2 and chapter 3).

The following sections describe the database and web site design, annotation features provided and the intended usage of the database.

### 5.2 Database Design

The baseLINE database is constructed using MySQL (http://www.mysql.com), a Relational data base Management System. MySQL provides a fast, reliable and easy to use, open-source solution to storing L1 data in separate tables rather than storing all data in a monolithic data table. These tables containing computationally mined L1 data are linked by a set of defined relationships, which makes it possible to access and combine data from a number of tables on a user’s request. Data can be added, accessed and processed using the Structured Query Language (SQL) to enable the easy construction, maintenance and use of baseLINE.

The design of the baseLINE database is represented as an Entity Relationship (ER) diagram in Figure 5.1. The ER model shown in Figure 5.1 describes visually the type of information that can be stored in baseLINE. Each of the Entities in the ER diagram is represented by a rectangle and can be thought of as nouns, e.g. baseLINE L1s, TSDs, L1Xplorer annotation, Sequence, Duplicates and Blast. The relationship between two entities can be described by how they are related to one another i.e. like verbs linking two or more nouns. For example a ‘has’ relationship could exist between a baseLINE L1 and Target Site Duplication annotation. The relationships between these two entities are represented as diamonds and are connected by lines. Figure 5.2 illustrates the relationship between the L1 and TSD entities. Since an L1 sequence may have zero
or more TSDs, Figure 5.2 indicates that one baseLINE L1 has zero, one or more TSDs associated with it.

The relationships between all entities are illustrated in Figure 5.1 and listed in Table 5.1. Every entity has a uniquely identifying attribute(s), known as a primary key. The primary key is used to uniquely identify each row in a data table (an entity).

**Database tables**

A description of each of the entities and the attributes associated with them are described in the following sections.
Table 5.1: A description of baseLINE relationships

<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>One baseLINE L1 may have zero, one or more Duplicates</td>
</tr>
<tr>
<td>B</td>
<td>One baseLINE L1 has one or many blast results (i.e. more one than due to multiple sources)</td>
</tr>
<tr>
<td>C</td>
<td>One baseLINE L1 has one or many L1Xplorer annotations (i.e. more one than due to multiple sources)</td>
</tr>
<tr>
<td>D</td>
<td>One baseLINE L1 has zero, one or many TSDmapper annotations</td>
</tr>
<tr>
<td>E</td>
<td>One L1 can be present in zero, one or more genes</td>
</tr>
<tr>
<td>F</td>
<td>One baseLINE id has one or more sequences (i.e. more one than due to multiple sources)</td>
</tr>
<tr>
<td>G</td>
<td>One duplicate identifier has one sequence</td>
</tr>
<tr>
<td>H</td>
<td>One duplicate identifier has one blast result</td>
</tr>
</tbody>
</table>

Figure 5.2: The baseLINE entity-relationship diagram. The baseLINE entity-relationship diagram. The green boxes indicate the entities and grey diamonds the relationships between entities.
baseLINE L1s

The baseLINE database consists of a central reference table of L1 annotation named baseLINE. This table contains 16 fields (attributes) providing information about each baseLINE L1. As each baseLINE L1 can potentially be found in 4 sources (Reference, Celera, HuRef assemblies and the April 2008 nt database) this table provides an automatically generated primary key and another unique identifier, the baseLINE identifier. The baseLINE identifier represents one or more L1s that have been mapped to a specific genomic location on NCBI build 36.3. The start and stop position of the 8-kb L1 sequence (1-kb of 5’ and 3’ flanking DNA and the ~6-kb L1 DNA) in NCBI build 36.3, the insertion length and the 1-kb flanks is also recorded. If the L1 is not present in NCBI Build 36.3, the expected insertion point is recorded as the start and stop coordinates are the same. As a result, the insertion length is recorded as zero. The orientation of the L1 and the chromosome to which it was mapped is also recorded. The unique identifier from each of the 4 sources is also recorded (if present). If the L1 in question is present in the reference assembly and then was detected in the other sources, its allelic status is also recorded (for Celera, HuRef and nt). For the nt-derived L1s there is a possibility that more than one instance of the L1 could exist in different accessions so there is a field in this table indicating if there are duplicate nt accessions.

Duplicates

The baseLINE identifier is used to locate the duplicates in the Duplicates table. A duplicate is defined as a piece of DNA sequence containing an L1 that maps to the same genomic location as an L1 in the reference assembly, or that has the same flanking DNA regions if the L1 is not present (i.e. absent from the Reference assembly but copies exist in other sources). There were no specific selection criteria for which nt L1 sequence should be classed as a duplicate. Arbitrarily the first sequence retrieved from the BLAST results was used as the nt identifier in the baseLINE table. All subsequent instances of the same L1 located in accessions in the nt database were placed in the duplicates table.
TSDmapper

The TSDmapper table contains all the annotation data generated by the perl script TSDmapper (see chapter 4). This table contains annotation from five main output areas of the script; identifiers, polyA tail search, predicted TSDs, endonuclease cut site and transductions.

The identifiers field contains an auto-generated identifier, so that each row is unique within the table and a link can be made back to the baseLINE or Duplicates table.

The polyA tail annotation has the following fields: a polyA tail name, polyA sequence start and stop (within the 3’ flanking DNA), length of the polyA tail, and polyA tail sequence. The table also contains information about the polyA tail search sequence that is generated to find the TSDs in the 5’ flanking DNA of the L1. This annotation includes the start and stop of the search sequence within the 3’ flanking DNA (from which it was extracted) and the polyA search sequence itself.

The TSD annotation part of this table includes the TSD start and stop within the 5’ and 3’ flanking DNA and the full sequence of the 3’ TSD. Any variations in the 5’ TSD sequence are stored as well as the TSD length, and the percentage identity of the 3’ TSD sequence to the 5’ TSD. The TSD start and stop positions within the 8-kb sequence (L1 and flanking sequence) are also stored.

The endonuclease part of this table contains information on the endonuclease cut site sequence, their start and stop position within the 5’ flanking DNA and the scores and grades generated by TSDmapper (see chapter 4).

The transduction part of this table stores the start and stop position of the 5’ and 3’ transductions, the length of each transduction and the transduction sequence.

L1Xplorer

The L1Xplorer table contains annotation generated from the program suite L1Xplorerer. This analysis was performed by Dr Tomasz Zemojtel of the Max Planck Institute
of Molecular Genetics, Berlin, on all FL-L1HS found in the BLAST searches. For a full list of L1Xplorer annotation please see http://l1xplorer.molgen.mpg.de/. This annotation was filtered to remove all TSD annotation and polyA annotation due to the fact that TSDmapper table already provides this and the TSD-finding program implemented in L1Xplorer does not always provide accurate results.

**Genes**

The genes table contains an auto-generated identifier so that each record is unique in the table. The baseLINE identifier, the insertion orientation of the L1 in NCBI build 36.3 of the human genome assembly, the gene common name, the gene orientation and the chromosome of the reported gene are also recorded.

**Sequences**

The Sequence table contains the L1 sequence, the 1-kb 5′ flank and the 1-kb 3′ flank for each accession.

**Blast**

The Blast table provides all of the Blast information parsed from the blast searches of the Reference, Celera, HuRef, and nt data sources. These sequences met the criteria of \( \geq 5922 \) bp in length and \( \geq 98\% \) identity (see chapter 3 for more details). This table contains the unique source accession for a particular L1, the L1 start and stop positions in the accession, the percentage identity to the query L1 (L1.3_mpa), the orientation, the common name of the species and the species code. This table also contains information on L1s that are not in the baseLINE reference table because they have been excluded because they do not have enough flanking DNA, or are not human L1s.
5.3 Website

To view FL-L1HS sequences annotation data in the baseLINE database a web resource has been constructed and is available at http://baseline.gene.le.ac.uk. The baseLINE web resource is a collection of CGI scripts that dynamically generate web pages using perl scripts and a webpage templating system called the Template Toolkit (see chapter 2 for more details). baseLINE provides access to each FL-L1HS record, and allows the visualisation of the FL-L1HS in its genomic context. In addition, baseLINE provides an online tool for the detection and annotation of Target Site Duplication (TSDs), called TSDmapper. This is a graphical web interface to the perl script TSDmapper tool described in chapter 4. The following sections describe the baseLINE web resource and gives examples of how to view and search for FL-L1HS sequences.

Website navigation

At the top of each page of baseLINE, five tabs are available for the navigation around the baseLINE web resource. Figure 5.3 shows the tabs, as they appear on the home/entry page. Clicking on the home tab takes the user to the home page, clicking the search tab provides a page of search options, clicking on the view tab provides a graphical way to search for and view FL-L1HS sequences. Clicking on the TSDmapper tab provides a way to search for TSDs in nucleotide sequences and clicking the Help tab provides information on how to use the site.

Home Page

The first page a user will see when accessing baseLINE (http://baseline.gene.le.ac.uk) is the baseLINE home page. This is illustrated in Figure 5.4. The baseLINE home page provides a welcome message and a short introduction to the resource. It also provides links to the search, view and TSDmapper pages.
CHAPTER 5. BASELINE - A FL-L1HS DATABASE

Figure 5.3: The five navigation tabs of the baseLINE website. The tab highlighted in grey is the currently selected tab.

Figure 5.4: The home page of the baseLINE website, displaying a welcome message and brief instructions on how to search and use the website.
Search

One of the main features of the baseLINE website is the ability to search for FL-L1HS sequences using a number of different characteristics. Clicking on the search tab presents the user with a search page enabling the user to search by accession, chromosome, gene, and TSDs. Examples of each of the different types of searches are described below. All results from the search are re-posted to the same webpage so the user has the search options at hand to perform another search.

Accession search

The baseLINE database can be searched using an accession from the four main FL-L1HS sources (Reference, Celera, HuRef, and nt2008) or by a baseLINE identifier. For example if a user wanted to obtain all FL-L1HS accessions from the Reference assembly a query such as ‘NC’ may be used, as all Reference assembly contigs start with the NC identifier. This returns 462 results, in a table with 10 results being displayed per page. The next set of results can be accessed by clicking on the page number illustrated in Figure 5.5. To view pages past the 10 pages available at the bottom of webpage the user can click the ‘next set’ link and the next 10 pages will be available for viewing. The results table generated from the accession search provides information on the following baseLINE annotations: the baseLINE identifier (with the prefix BL (Base Line) followed by six digits i.e. BL000001 with a link to the baseLINE record), the chromosome to which the FL-L1HS sequence has been mapped, the source accession identifier and the source of the DNA. Users can also search for a specific FL-L1HS if they know the accession of the L1 for which they are searching. Figure 5.6 illustrates the results of searching for the FL-L1HS with the accession identifier AC002980. If a user searches for an accession that is not in the database, then a message is posted to the webpage informing them that there are no results for that search term.
CHAPTER 5. BASELINE - A FL-L1HS DATABASE

Figure 5.5: Searching for FL-L1HS by L1 source. Shown are the results of searching the Reference assembly for the term ‘NC’.

Figure 5.6: Searching for FL-L1HS by L1 source. Shown are a search of the accession database for the L1 ‘AC002980’.
CHAPTER 5. BASELINE - A FL-L1HS DATABASE

Figure 5.7: Searching for FL-L1HS insertions by L1 by chromosome. Shown are the results of searching for L1s on chromosome 1.

Chromosome search

To search for FL-L1HS sequences by chromosome, the user clicks on the radio button at the top of the search page labelled chromosome. This will then display the options available for chromosome searching. Using the drop down box a user can select to search for FL-L1HS either by all available autosomes and sex chromosomes or by individual autosomes or sex chromosomes. Figure 5.7 displays the results from the chromosome FL-L1HS search. This search shows that there are 49 FL-L1HS insertions located on chromosome 1.
CHAPTER 5. BASELINE - A FL-L1HS DATABASE

Figure 5.8: The results of searching for FL-L1HS sequences in the BRCA1 gene returns no results from the database.

**Gene search**

The baseLINE website offers the user the opportunity to search for FL-L1HS sequences in a gene-centric manner. Clicking on the radio button ‘gene’ enables the gene search. A user can type a gene symbol into the gene search text box and the results of the search are displayed in a table on the same web page. If a gene is searched for that does not contain a FL-L1HS sequence, the web page displays no results found. For example, searching for the gene symbol BRCA1, where mutations in this gene are associated with a significant increase in the risk of breast cancer, finds no FL-L1HS sequences. The results of this search are illustrated in Figure 5.8.

However searching for FL-L1HS sequences that have inserted into genes returns a table of results, with the following FL-L1HS annotations; the baseLINE id with a link to the baseLINE record of the L1, the insertion orientation of the L1, the common gene name, the orientation of the gene the L1 has been reported to have inserted into, and finally the chromosome on which the L1 is located. Figure 5.9 illustrates a gene search using the gene symbol RABGAP1L. This search returns 4 FL-L1HS sequences that are reported to have inserted in the RABGAP1L gene. This result shows that all 4 are inserted in the opposite transcriptional orientation with respect to FL-L1HS/gene orientation (-/+).
Results of the gene search RABGAP1L

Figure 5.9: The results of searching for FL-L1HS sequences in the RABGAP1L gene returns 4 hits from the database.

**TSD searching**

To search for a FL-L1HS sequence by TSD length or polyA tail length the user clicks the ‘TSDs’ radio button. The user is presented with the search options for TSD and polyA tail searches. An example of a TSD length search is illustrated in Figure 5.10. This figure shows the results of a TSD length search for all FL-L1HS sequences that have a TSD length (generated from TSDmapper) \( \geq 14 \) and \( \leq 15 \) bp (i.e. 14 or 15). There are 192 FL-L1HS sequences that have TSDs of these lengths.

Also in this search option, the user has the ability to search for FL-L1HS sequences by polyA tail length. Figure 5.11 illustrates a search for FL-L1HS sequences with polyA tails \( \geq 50 \) and \( \leq 60 \) bp. The ability to search for by polyA tail length means the user can identify polyA tails that are unusual for young FL-L1HS. These FL-L1HS sequences are unusual since the mean polyA tail detected in chapter 4 was 27 bp and these polyA tails are almost double the mean length of the analysed dataset (520 elements), as it has been previously reported that the mean polyA tail length of Ta subfamilies was 41 bp...
Finally, the identification of FL-L1HS sequences with longer than average polyA tails can give an indication of the age of the FL-L1HS element, as younger elements have been reported to have longer polyA tails (Ovchinnikov et al., 2001).

**View**

The baseLINE website allows users to view FL-L1HS sequences from the different database sources in a graphical genomic context, using the GBrowse Karyotype viewer (http://gmod.org/wiki/GBrowse_karyotype). The karyotype viewer allows the user to view the locations and distribution of elements on a set of chromosome diagrams. Underneath the ideograms are XY plots of the gene density and GC content in 1-Mb windows. The gene density was calculated from the RefSeq genes (see chapter 2 for more details) and the GC content was taken from the supplementary data of the Cooper et al. (2007) analysis.
Figure 5.11: The results of the polyA tail length search. Shown are the number of FL-L1HS sequences in the database that have a polyA tail length between 50 and 60 bp.

Using the control options on the ‘View’ page it is possible to select different chromosomes and different FL-L1HS database sources. The karyotype view allows the quick visual identification of elements that vary between sources, suggesting that these could be real genetic polymorphisms that exist between human individuals. All possible FL-L1HS insertions on chromosome 1 from the four sources analysed in this thesis are illustrated in Figure 5.12. Also shown is the baseLINE FL-L1HS track (i.e. all FL-L1HS insertions regardless of source). This figure indicates that one of the FL-L1HS sequences varies between the four data sources. BL000001 is not found in the Celera assembly but is detected in the HuRef, nt 2008 and Reference databases. Also shown in this figure is the gene density divided into 1-Mb bins on an XY plot. The yellow star in Figure 5.12 shows an increase in gene density and a lack of FL-L1HS detected around this point. The GC percentage average in a 1-Mb bin is also shown on a separate XY plot and the red star indicates a drop in GC content at the centromere of the chromosome. The display of the GBrowse karyotype can also be changed to only display FL-L1HS insertions from one source. This is shown in Figure 5.13.
Another example of FL-L1HS variation between data sources is illustrated in Figure 5.14. This figure shows the karyotype view of the X chromosome showing FL-L1HS from all sources. This view illustrates the lack of FL-L1HS present on the X chromosome in the HuRef assembly.

Once a user has selected a view that they want to analyse further, they can do so by clicking on the triangle from any of the sources and this will then take the user to the annotation record of that FL-L1HS sequence.

**L1 record**

The L1 record page contains all the annotation details for each FL-L1HS sequence in the baseLINE database. This page can be accessed upon finding an L1 or can be accessed from the view pages. The following section discusses the annotation displayed on the L1 record page. The baseLINE L1, located on chromosome 2, and BL000139, located on chromosome 3, are used to illustrate the types of annotation information available on this page.

The first part of the L1 record page is the identifier section. This section contains the baseLINE identifier, the source accessions (*i.e.* which source accession the L1 was found). If the L1 is not present in a particular source then it is described as absent. Figure 5.15 illustrates the identifier section of the L1 record page. It is possible to see for BL000060 that this L1 is found in three out of the four L1 sources (Reference, Celera and and HuRef) but it was absent from the nt accession database. In contrast to BL000060, the L1 BL000139 is found in only one source, the nt accession database and is absent from all of the human genome assemblies. The identifier section also contains information on the insertion orientation of the L1 and its flanking DNA with respect to the NCBI b36.3 human genome Reference assembly, the genomic position of the 8-kb sequence (flanking DNA and L1) and the chromosome to which the L1 was mapped in the genome assembly. As BL000139 is absent from the Reference assembly the insertion position of the 5' end of the L1 is recorded instead (1 bp). Any duplicates (*i.e.* multiple copies of the same L1 found in the nt database) are also displayed for the L1 in question.
Figure 5.12: A GBrowse karyotype view of chromosome 1 displaying all baseLINE FL-L1HS sources. Shown by the purple triangles are FL-L1HS insertions detected in the Celera assembly, red triangles FL-L1HS insertions detected in the HuRef Assembly, blue triangles FL-L1HS insertions detected in nt 2008, and green triangles FL-L1HS insertions detected in the Reference assembly. The yellow triangles indicate all FL-L1HS insertions regardless of source i.e. the baseLINE FL-L1HS. The yellow star indicates an increase in gene density and lack of FL-L1HS, while the red star indicates a drop in GC content.
Figure 5.13: A GBrowse karyotype view of chromosome 1 displaying a single L1 source (baseLINE).
Figure 5.14: A GBrowse karyotype view of the X chromosome displaying all baseLINE FL-L1HS sources. Shown by the purple triangles are FL-L1HS insertions detected in the Celera assembly, red triangles FL-L1HS insertions detected in the HuRef Assembly, blue triangles FL-L1HS insertions detected in nt 2008, and green triangles insertions FL-L1HS detected in the Reference assembly. The yellow triangles indicate all FL-L1HS insertions regardless of source i.e. the baseLINE FL-L1HS. Highlighted by the circle is the only FL-L1HS hit detected in the HuRef assembly on the X chromosome.
A useful feature on the L1 record page is the ability to view L1 data on two of the most widely used genome browsers: UCSC genome browser (http://genome.ucsc.edu/) and the Ensembl genome browser (http://www.ensembl.org/). baseLINE provides a direct link to the UCSC browser for each mapped L1 sequence, displaying a view of the L1 (if present) and its flanking DNA with the RepeatMasker track turned on. Figure 5.16 illustrates a UCSC view for BL000060 showing the L1 sequence and 1 kb of flanking DNA at the 5' and 3' end of the L1, showing that this element is present in the Reference assembly. Figure 5.17 illustrates the UCSC view for the L1 BL0000139 that is only found in the nt assembly (i.e. it is highly likely that this element is polymorphic) and showing the insertion point of the L1 in the assembly (1 bp view).

The baseLINE annotation data can also be viewed on the Ensembl genome browser as a Distributed Annotation Service (DAS) track. The baseLINE data are available for export from a server and can be searched for in the DAS registry (http://www.dasregistry.org/) under the name ‘baseline - A collection of human-specific L1
sequences’. By allowing the baseLINE annotation data to be displayed on external websites that support DAS tracks (such as Ensembl genome browser) the data can be displayed together with other types of annotation data available on the genome browser. For example, baseLINE annotation can be put into a genomic context with standard tracks like gene information and EST annotation and then be compared to data from other external DAS sources such as the L1Base DAS track (http://line1.molgen.mpg.de/). Figure 5.18 illustrates BL000060 annotation data exported as a DAS track, with the DAS track from L1Base also shown. Figure 5.19 illustrates the annotation of BL000139 on the Ensembl genome browser even though there is no FL-L1HS sequence in the Reference assembly. It is also possible to see that L1Base does not have any
Figure 5.17: The UCSC genome browser view section of the L1 record page for BL000139 showing the insertion point of the L1 sequence that is not in the Reference assembly.

annotation information on BL000139. Viewing baseLINE data in this way gives a user a good indication that the L1 BL000139 is highly likely to be polymorphic and probably capable of retrotransposition.
Figure 5.18: The Ensembl genome browser view section of the L1 record page for BL000060 showing the baseLINE annotation data as a DAS track, with the L1Base DAS tracks, L1Base NCBI35 FLI and L1Base NCBI35 FLn and the RefSeq gene DAS track.
Figure 5.19: The Ensembl genome browser view section of the L1 record page for BL000139 showing the baseLINE annotation data as a DAS track, with the L1Base DAS tracks, L1Base NCBI35 FL1 and L1Base NCBI35 FLn and the RefSeq gene DAS track.
Information on TSD annotation, generated by TSDmapper, is also displayed on the L1 record page. Figure 5.20 illustrates the annotation generated from TSDmapper for BL000060 and BL000139. The annotation that is displayed in this section can be further subdivided into annotation on polyA tails, endonuclease recognition sites, TSD sequence and TSD scoring. The TSDmapper annotation was generated using the default TSDmapper options (minimum polyA tail length of 10 bp, a contamination threshold of 2 non-As, and utilised 3 kb of flanking DNA).

The polyA annotation contains information about the length of the polyA tail and the polyA tail sequence. The endonuclease cut site sequence is shown and a downwards arrow indicates the cut site. The score of the endonuclease cut site is also displayed.

The TSD annotation includes the best TSD prediction from TSDmapper and the sequences at the 5′ and 3′ end are displayed, along with any variations between the sequence at either end of the L1 sequence. The TSD distance score, which relates to how close the 5′ TSD is to the 5′ end of the L1, is also displayed. Other TSD annotation includes an overall TSD score which is a weighted score of the endonuclease-recognition score and the distance score (60/40%), and a TSD grade to give the user an indication of how good TSDmapper’s overall score is. Hits above 80 are graded excellent if they have good endonuclease recognition scores and the 5′ TSD is close to the 5′ end of the L1. Scores below 79-60% are graded good candidates and those below 60% are graded poor (as they usually have poor endonuclease cut site sequences and/or are far away from the 5′ end of the L1).

The transduction annotation consists of the lengths of the 5′ and 3′ transductions (if any have been detected) along with the transduced sequence. This sequence can potentially be used to determine progenitor/sibling relationships. Figure 5.21 illustrates the transduction annotation for BL000060 and BL000139.

An important piece of annotation is the retrotransposition capability prediction. The prediction of retrotransposition capability is based on the ORF status from the program L1Xplorer. An intact ORF is defined as having 0 stop codons, 0 gaps and \( \leq 1 \) in-frame deletions/insertions in either ORF, and no mutations in motifs that are known to affect retrotransposition (see chapter 1). A L1 element is predicted to be capable
 CHAPTER 5. BASELINE - A FL-L1HS DATABASE

Figure 5.20: The TSD section of the L1 record page for BL000060 and BL000139 showing the polyA tail, endonuclease cut site and TSD annotation.

Figure 5.21: The transduction section of the L1 record page for BL000060 and BL000139 showing the two sequences.
Figure 5.22: The retrotransposition and sub-family classification section of the L1 record page for BL000060 and BL000139. Highlighted by the blue box for the BL000060 element are alleles that are predicted to be potentially active FL-L1HS sequences while the predicted inactive allele is highlighted by the red box.

of retrotransposition if both ORFs are classed as being intact. Figure 5.22 illustrates the retrotransposition and sub-family classification annotation. BL000060 is present in 3 out of the 4 data sources mined for FL-L1HS and two of the L1 sequences have been predicted to be active (i.e. capable of retrotransposition). The 2 FL-L1HS sequences are from the Reference and Celera genome assemblies (NC_000002.10_aj, AC_000045.1_y) while the FL-L1HS from the HuRef assembly (AC_000134.1_t) is inactive. This FL-L1HS is an example of how allelic FL-L1HS sequences may have different retrotransposition activity. The single allele sequence of BL000139 (AC216905.3_a) is also predicted to be capable of retrotransposition.
The empty site annotation section contains the reconstructed empty site sequence generated by TSDmapper. This includes 250 bp of 5' flanking DNA (red sequence in Figure 5.23) one copy of the TSD (blue sequence Figure 5.23) and 250 bp of 3' flanking DNA (green sequence in Figure 5.23). This sequence contains no L1 sequence and represents the putative ancestral state, even though this sequence cannot be found in human populations any longer.

TSDmapper

The baseLINE website also contains a web interface to the TSDmapper program, based on the script described in chapter 4. This section describes the use of TSDmapper on the baseLINE website.

Clicking on the TSDmapper tab takes the user to the TSDmapper submission form. This is where a user will submit two sequences for analysis and set up the parameters...
Figure 5.24: The TSDmapper web submission form on the the baseLINE website, the sequence submission section.

for the TSD searching. Figure 5.24 illustrates the TSDmapper submission form, with the 5' and 3' sequences of BL000503. Using 2 kb of flanking DNA and 1 kb of 5' and 3' L1 from the sequence NC_000023.9_b (which is the L1 sequence from the Reference assembly for BL000503) the TSDmapper web program was run with default parameters (polyA minimum length 10 bp, polyA contamination threshold of 2 non-consecutive A residues, LALIGN E-value 10 and the percentage identity to the query sequences ≥ 99%). The configuration page is illustrated in Figure 5.25. The sequences that TSDmapper requires are; A 5' sequence containing a minimum of 250 bp of 5' flanking DNA with some 5' L1 sequence following it; the 3' sequence is required to contain 3' L1 sequence and at least 250 bp of 3' flanking DNA. To identify TSDs with greater confidence, larger flanking sizes are needed as the TSDmapper algorithm has proved most effective when both 5' and 3' flanking sizes are between 1-2 kb in length, as this enables the detection of large transductions.

TSDmapper can analyse sequences in a number of different formats (raw and fasta). If the sequence does not conform to these supported sequences then the TSDmapper...
Figure 5.25: The TSDmapper web submission form on the baseLINE website, the parameters section, allowing customisation of the polyA tail-finding algorithm and the sensitivity of the TSD searching program.

Once a user clicks on the submit button, and assuming the sequences submitted meet all minimum requirements, the results page will be loaded. For the example L1 sequence NC_000023.9_b (BL000503), the TSDmapper results page is described below.

**TSDmapper output**

The first section of the TSDmapper results with which the user is presented, is a graphical representation of the submitted sequences, illustrating where TSDmapper has found flanking sequence, 5’ and 3’ transductions, the TSDs at the 5’ and 3’ end and the 5’ and 3’ ends of the L1s. Figure 5.26 illustrates the graphic drawn by TSDmapper for
the submitted sequences from NC_000023.9_b. It is possible to see that in total \( \sim 6 \) kb of sequence was submitted to TSDmapper (2 kb of flanking DNA and 1 kb of L1 for both the 5' and 3'). TSDmapper is designed to be able to accept just the ends of the L1, to allow for situations where it is not possible for the user to have the whole L1 sequence (i.e. only having the 5' and 3' ends from trace files). The line joining the two parts of the L1 together in Figure 5.26 represents the unknown sequence of the middle part of the L1. The image TSDmapper gives the user a visual idea of where the important features are located such as TSDs and transductions in relation to the L1 sequence. The annotation key provided by TSDmapper describes the various annotation colours. For the example shown in Figure 5.26 it is possible to see that the L1 sequence has a large 3' transduction, however it is difficult to see the 5' transduction due to it being very small. As small transductions will not always show up on the image, the transduced sequence is given in the TSDmapper annotation table (Figure 5.30).

The TSDmapper summary table provides the user with the coordinates and sequence annotation of each of the annotation features detected by TSDmapper. Figure 5.27 shows the position of the L1 in each of the 5' and 3' sequences. Also shown in Figure 5.27 is the position of the polyA tail at the 3' end of the L1, along with its length and the polyA tail sequence.

The endonuclease annotation, illustrated in Figure 5.28 shows the coordinates of the endonuclease cut site in the 5' submitted sequence along with the endonuclease cut site sequence, with a downwards arrow indicating the likely point of cleavage by the L1 endonuclease. Also shown in this figure is the TSDmapper score for the quality of the detected endonuclease cut site, along with the distance score of how far away the 5' TSD is away from the L1. Both of these scores are expressed as a percentage of the maximum possible score.
Figure 5.26: A TSDmapper-generated graphic of the TSDs and transductions in the submitted sequence of NC_000023.9_b. The 5’ flanking sequence is represented by a blue rectangle, 5’ transductions by the green rectangles (not shown in this diagram), the L1 sequence is represented by the black rectangle, the TSDs by red lines, 3’ transductions by the orange rectangle and the 3’ flanking sequence by the brown rectangle. The two submitted sequences are joined by a line. The sequences are shown 5’ → 3’.
CHAPTER 5. BASELINE - A FL-L1HS DATABASE

Poistions of the L1 in the two submitted pieces of sequence

PolyA tail information in the 3' submitted sequence

Figure 5.27: The start and stop positions of the L1 sequences and polyA tails in the two submitted sequences.

Endonuclease cut site information from the 5' submitted sequence

Endonuclease cut site score and Distance score

Figure 5.28: The endonuclease cut site information generated from the 5' submitted sequence to TSDmapper. The cut site of the L1 endonuclease is indicated by ‘↓’.
The TSD annotation illustrated in Figure 5.29, shows the start and stop positions of both the 5' and 3' TSDs relative to the submitted sequence. Also illustrated in Figure 5.28 is an alignment of the 5' and 3' TSDs, with the 5' TSD shown in blue and the 3' TSD shown in red. The length of the TSD and the TSDmapper score (endonuclease score and the distance score combined) is displayed alongside the TSDmapper grade. For the L1 NC_000023.9_b TSDmapper prediction is classed as excellent as the TSDs are close to the 5' end of the L1 and the endonuclease cut site conforms to the endonuclease consensus (see chapter 1 and chapter 4).

The transduction annotation is illustrated in Figure 5.30 and shows the start and stop positions of the 5' and 3' transduction (if present) along with the transduction sequence. The empty-site sequence, generated by TSDmapper comprises 250 bp of 5' and 3' DNA joined by a single copy of a TSD with no L1 sequence as illustrated in Figure 5.31.
CHAPTER 5. BASELINE - A FL-L1HS DATABASE

Figure 5.30: The transduction annotation generated by TSDmapper from the two submitted sequences.

<table>
<thead>
<tr>
<th>5' transduction start</th>
<th>1996</th>
</tr>
</thead>
<tbody>
<tr>
<td>5' transduction stop</td>
<td>1999</td>
</tr>
<tr>
<td>5' transduction sequence</td>
<td>AGGGA</td>
</tr>
<tr>
<td>3' transduction start</td>
<td>1002</td>
</tr>
<tr>
<td>3' transduction stop</td>
<td>1223</td>
</tr>
<tr>
<td>3' transduction sequence</td>
<td>TATATATGAATAGCCTTGAGCCCTCGCTTCACGACAGAGGAGCTGATGGAAGAACAGCGCTTCAGAAGGACAGGAGACCCCTGAA</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 5.31: The TSDmapper generated empty site from the submitted sequences. The 5' flanking DNA is represented by red text, the 3' flanking DNA by the green text, and one copy of the TSD by the blue text.

BL000503 (NC_0000239_b)

| Empty site sequence | ATTTTGAATAGCCTTGAGCCCTCGCTTCACGACAGAGGAGCTGATGGAAGAACAGCGCTTCAGAAGGACAGGAGACCCCTGAA |
|                     |     |

BL000503 (NC_0000239_b)
5.4 Discussion

This chapter describes the design and implementation of a database and web resource, baseLINE (http://baseline.gene.le.ac.uk), which was developed for the storage and analysis of full length human specific L1 (FL-L1HS) retrotransposon annotation. baseLINE comprises a database whose content is made searchable and available for analysis through a freely accessible web site, as well as a web interface to the TSDmapper program which enables website visitors to analyse their own sequences.

Database design

The baseLINE database was designed to allow the collation and storage of L1 annotation from the Reference, Celera, and HuRef human genome sequence assemblies, as well as the nt accession database. The database stores the source FL-L1HS sequence identifier (i.e. the accession), the position of the FL-L1HS in the Reference assembly genome, the original source BLAST results, annotation from TSDmapper, annotation from L1Xplorer, information on insertions in genes, the L1 sequences themselves and the sequence of their flanking DNA. These data are made searchable by use of a website interface that enables exploration and display of individual L1s in their genomic context.

Database future

The database could be developed in the future to include novel (i.e. previously unreported) FL-L1HS insertions discovered using genome-wide display procedures such as ATLAS (Badge et al., 2003) as well as providing additional annotation in the form of allele frequencies and population distribution. Also, additional L1 sequences derived from computational “mining” of the trace archive database could be incorporated automatically to give a better representation of the number of FL-L1HS residing in human DNA sequence databases. In principle the analysis pipelines and database construction and publication could be fully automated to allow periodic updates to the database.
CHAPTER 5. BASELINE - A FL-L1HS DATABASE

Also, since this analysis was performed, the genome of James D Watson has been published (Wheeler et al., 2008a). This genome could also be searched for FL-L1HS sequences and included in the database, increasing the content of the baseLINE resource. It seems likely that more “personal genomes” will become available in the future and the automated processing of their L1 content would be a major aim of future work in this area.

**baseLINE website feature comparisons with other L1 resources**

The baseLINE website front end provides access to the FL-L1HS annotation contained in baseLINE via search options and graphical representations of the data. While some of the data are present in other resources, none have the same focus on human-specific full-length elements or suite of search and visualisation tools. However as baseLINE is not the only L1-specific web resource currently available, the differences between baseLINE and other L1 databases, such as Database of Retrotransposon Insertional Polymorphisms (dbRIP) (http://dbrip.brocku.ca/, (Wang et al., 2006)) and L1Base – A database of Full-Length, Intact L1 Elements (http://l1base.molgen.mpg.de/, (Penzkofer et al., 2005)), will be considered below.

**dbRIP**

dbRIP is a database of polymorphic human retrotransposons, including L1, *Alu* and SVA elements. Currently dbRIP stores polymorphic FL-L1HS insertions that have either been directly submitted by researchers or collated from published papers. This resource only contains 76 L1HS that are greater than ≥5922 bp in length. Although dbRIP is not specifically a full-length L1 resource it clearly contains fewer FL-L1HS insertions than baseLINE (533 FL-L1HS). In baseLINE, 107 FL-L1HS sequences are predicted to be polymorphic from the discovery of empty sites in the trace archives and/or the nt accession databases. Therefore even though baseLINE is not specifically a L1 polymorphism database the way the data are gathered (from multiple sources) and annotated (empty site *in-silico* genotyping) baseLINE potentially represents more polymorphic FL-L1HS sequences than dbRIP.

baseLINE summarises information on real and assembly FL-L1HS polymorphisms by providing annotation from 4 data sources. The variation between the 4 different sources
can be viewed graphically on a chromosome ideogram (Clicking the ‘View’ tab on the baseLINE website), providing the location of each of the 533 FL-L1HS on the reference assembly, along with options to view a single source or all sources. Accessing the underlying data concerning individual insertions is achieved by clicking on the relevant symbol (see Figure 5.12, Figure 5.13, Figure 5.14). The image also shows the percentage GC content and gene density in 1-Mb bins to illustrate the genomic context of the insertions.

An ideogram image is also provided by dbRIP, showing the location of polymorphic L1s on chromosomes. However this image does not allow one to identify specific L1s, and there is no direct link to the annotation information, unlike the FL-L1HS insertions represented on the ideograms in baseLINE. Also there is no direct indication as to in which source the polymorphic L1 was discovered. dbRIP does however implement a local instance of the UCSC genome browser to enable viewing of polymorphic L1s in their genomic context. Although baseLINE does not have its own local genome browser, it does have the capability to export annotation information to the Ensembl genome browser using DAS. This approach has an advantage over a local genome browser installation as the data can be viewed alongside other DAS tracks from other sources, allowing comparison with related sources. Also, as the DAS annotation is relatively lightweight compared to a full local UCSC genome browser installation, very little maintenance is required: Ensembl serves the genome browser functionality, ensuring high availability and responsiveness. Also as the baseLINE DAS annotation track is in the DAS registry (http://www.dasregistry.org/) this allows users to discover the baseLINE DAS annotation by searching for key terms. The baseLINE DAS annotation that is displayed on the Ensembl genome browser contains information on the source of the annotated L1, the insertion orientation and most importantly a direct link to the L1 record on the baseLINE website. Another advantage of using DAS annotation to display baseLINE data is that it could easily be adapted to run on other genome browser frameworks that support DAS, such as GBrowse (http://gmod.org/wiki/Gbrowse). This would allow the display of baseLINE annotation data on other types of variation databases such as HGVbaseG2P (http://www.hgvbaseg2p.org) (Thorisson et al., 2009) and the Database of Genomic Variants (http://projects.tcag.ca/variation/) (Iafrate et al., 2004) which both use Gbrowse to display human genome variation data.
The search criteria available on the dbRIP website are more detailed than currently available on baseLINE, with users being able to search by genomic regions, ethnicity, polymorphism frequency, disease type and authors, or being able to search for multiple identifiers or locations by uploading a file. However baseLINE allows users to search for FL-L1HS sequences by polyA tail length, TSD length or whether the insertion occurred within a gene. Searches for FL-L1HS having long polyA tails can help identify a FL-L1HS sequence that may be young and still capable of retrotransposition, while the gene search function is useful for researchers who study specific genes and may not be aware that a FL-L1HS sequence is present in some individuals (due to it not be represented in the human genome Reference assembly). Therefore both websites provide distinctive search functionalities for locating L1s in human genome sequences.

dbRIP and baseLINE both provide annotation information on TSDs. TSD annotation on baseLINE is generated by an automated detection and annotation tool, TSDmapper. This tool can also recreate the empty-site sequence and identify transductions. dbRIP currently provides no TSD annotation tools for users to analyse new polymorphic insertions. In contrast to this, baseLINE provides an online version of TSDmapper, to allow annotation of user submitted sequences.

Searching for L1 sequences present in dbRIP is possible via BLAT, allowing user-uploaded putative insertions to be compared to known insertions. This enables users to map new L1 insertion sequences to build 35 of the human genome reference sequence. At this time baseLINE does not offer any sequence search functionality.

Another feature that is available on baseLINE is gene annotation. These data show the gene symbol into which the L1 is inserted and the insertion orientation of the L1 with respect to the gene’s orientation. This feature may be particularly useful for researchers examining how L1 insertions can interact with the transcription of their host gene.

One set of features that is available in dbRIP, but not baseLINE, is the 5’ and 3’ primers used to genotype the L1 insertions reported. This feature would be very useful for researchers wanting to design genotyping assays. It is possible that this feature could be added to the baseLINE website by computationally generating genotyping primers using an automated primer design program. Such a feature could be useful to end-users
but ultimately this is an unverified “suggestion” rather than the established genotyping assays presented by dbRIP.

**L1Base**

L1Base is a Lin resource containing putatively active L1 insertions in human and rodent genomes, extracted exclusively from assembled genomic sequences. L1Base can be searched by chromosome, start and stop positions as well as a number of important sequence motifs that if mutated are known to affect retrotransposition activity.

baseLINE contains annotation generated from the L1Xplorer annotation tool (kindly supplied by Dr Tomasz Zemojtel), but does not display each individual annotation motif like L1Base. baseLINE instead collates the motif information and ORF status (intact or non-intact) to make a prediction to see if the L1 might be active or not (i.e. capable of retrotransposition). L1Base displays all motif annotation information on a custom-built browser to show where all the sites are relative to the sequence of interest. Finally, L1Base also displays annotation information for TSDs and polyA tails using the TSDfinder script, however on most occasions this implementation of TSDfinder does not find the correct TSDs.

L1Base offers the ability to search for the L1 loci in the database using BLAST. Unfortunately at the moment baseLINE does not offer any form of sequence search.

One of the major differences between L1Base and baseLINE is the way that L1s are gathered for inclusion into each resource. L1Base gathers all elements from human genome assembly build 35 NCBI (2004). baseLINE, by contrast, gathers L1s from 3 human genome assemblies and the nt accession database. Therefore baseLINE is most likely more representative of the true level of human genomic L1 variation present in human sequence databases, than L1Base.

Both baseLINE and L1Base export data via DAS, as discussed above with regard to dbRIP. As a result, the two resources can be visually compared with each other side-by-side on the Ensembl genome browser to examine overlap between the two datasets.
TSDmapper web version

One of the main features of the baseLINE website is the web version of TSDmapper. TSDmapper was developed to allow the automated detection of TSDs and transductions in L1 sequences. Although L1Base offers an implementation of TSDfinder online, as part of the L1Xplorer suite of tools, TSDmapper is the only online resource that provides annotation information on TSDs, endonuclease cut sites, 5' and 3' transductions as well as generating a reconstructed empty site that can be used for \emph{in-silico} genotyping. A graphical representation of where annotation features are located in submitted sequences is also displayed, to enable direct identification of elements carrying transductions. In the future, TSDmapper could be applied to \textit{Alus} /older L1 elements with specific scoring matrices to enable analysis of these sequences.

In conclusion, the baseLINE website provides a large number of FL-L1HS sequences that are not present in dbRIP and L1base as well as providing additional annotation information not available in these resources. The main feature lacking in baseLINE is sequence searching. This is needed so that users can see if a FL-L1HS sequence has been detected previously, which would be a very useful and should be added to future versions of the website.
Chapter 6

Final summary and future directions

There have been many advances since Sanger and Coulson (1975) described one of the first methods for sequencing DNA. Developments in this technology over the years have allowed analysis of multiple regions in genomes of many species by both small and large research groups.

Since the publication of the first two human genome assemblies in 2001 by Lander et al. (2001) and Venter et al. (2001) it has been possible to analyse computationally, variations between genome assemblies to get an insight into human genome variation. With the addition of the nucleotide accession databases and trace archives, the wealth of human DNA sequences available to the bioinformation has continued to expand, allowing many large-scale analyses of human genome variation that were never previously possible.

Further advances in sequencing technologies by large biotechnology companies have lead to a reduction in the cost of sequencing the DNA, so much so that ‘personal’ genomes of two individuals (J.Craig Venter (Levy et al., 2007) and James D Watson (Wheeler et al., 2008a) have been sequenced.

Using the two human genome assemblies published in 2001 (Reference and Celera), the HuRef genome (Levy et al., 2007), the nt database and the trace archives, a computational analysis was performed to try and find the true extent of full-length
LINE-1 variation in these sources. The results of the analysis found 533 distinct FL-L1HS, which was significantly greater than previous analysis, suggesting that the level of FL-L1HS variation is greater than previously thought. With an increase in the number of FL-L1HS sequences now available, an opportunity existed to develop informatic analysis tools and resources. TSDmapper was developed to provide annotation (TSDs, transduction, empty site reconstruction) that was previously not available as a collection in other annotation tools. With an increase in the number of polymorphic and potentially active FL-L1HS sequences a resource was developed to display this variation to the wider scientific community, this resource is called baseLINE.

In the future the techniques developed in this thesis would have to be adapted to search for FL-L1HS sequences in genomes sequenced by technologies other than Shotgun sequencing. An initial pilot study on the James Watson Genome (sequenced using Roache-454) revealed that around 250 FL-L1HS sequences could be found from the 533 distinct FL-L1HS sequences using *in-silico* genotyping, suggesting further refinements to the methods would be needed to yield more FL-L1HS sequences.

In the last year there have been at least three more human genomes sequenced, further increasing the number of sources that could be potentially mined for FL-L1HS variation. The first Asian genome ([Wang et al., 2008](http://www.1000genomes.org)), and the first African genome ([Bentley et al., 2008](http://www.1000genomes.org)), both of which were done on the Illumina platform were released along with the first female genome, (Marjolein Kriek), generated at the Lieden University Medical Center in the Netherlands.

However these genomes are generated by ‘short read’ technologies (typically producing 35-bp reads in length), and although they allow genomes to be sequenced at great depth, they would probably struggle with insertion/deletion polymorphisms and resolving regions that are highly repetitive. This would prove to be a major issue for finding new FL-L1HS sequences that are not already known, as any ‘new’ FL-L1HS would most likely be polymorphic.

Also announced within the last year (2008) was the 1000 genomes project ([http://www.1000genomes.org](http://www.1000genomes.org)) an international research consortium formed to enhance our current
knowledge of human genetic variation. This project will produce vast quantities of data that can be analysed for genetic variation. However to be able to handle these data, a large amount of computer disk space (petabytes) will be needed along with advanced algorithms to be able to search and make sense of the data. The techniques developed in the thesis would most probably struggle to mine and make sense of the data, since they were developed for finding FL-L1HS sequences in assembled sequences and trace files that are significantly larger than 35 bp. Therefore these methods would not be suitable for analysing this type of data without extensive modifications. Therefore as sequencing technologies rapidly advance and the data continues to pile up, a radical re-think on how to process these data for detection of FL-L1HS variation is needed for these new genome assemblies. However the methods developed in this thesis can still be of major use for mining the nt accession databases and the trace archives (both of which continue to rapidly grow), along with all three human genome assembly future builds.
Appendix A

Literature review : Appendix

A.1  Alu-disease causing insertions

Table A.1 represents Alu insertions that have been shown to cause genetic disease. This table was adapted from http://www.med.upenn.edu/genetics/kazazianlab_human.shtml

A.2  L1 disease insertions

Table A.2 represents L1 insertions that have been shown to cause genetic disease. This table was adapted from http://www.med.upenn.edu/genetics/kazazianlab_human.shtml
Table A.1: *Alu* disease-causing insertions. a: Wallace *et al.* (1991); b: Muratani *et al.* (1991); c: Vidaud *et al.* (1993); d: Janicic *et al.* (1995); e: Miki *et al.* (1996); f: Halling *et al.* (1999); g: Lester *et al.* (1997); h: Abdelhak *et al.* (1997); i: Oldridge *et al.* (1999); j: Goldberg *et al.* (1993); k: Zhang *et al.* (2000); l: Stoppa-Lyonnet *et al.* (1990); m: Mustajoki *et al.* (1999); n: Economou-Pachnis and Tsichlis (1985); o: Li *et al.* (2001); p: Wulff *et al.* (2000); q: Sukarova *et al.* (2001); r: Ganguly *et al.* (2003);

<table>
<thead>
<tr>
<th>Phenotype observed</th>
<th>Inserted element</th>
<th>Gene disrupted</th>
<th>Insertion site</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neurofibromatosis(^a)</td>
<td>Alu (Ya5)</td>
<td>NF1</td>
<td>Intron</td>
</tr>
<tr>
<td>Acholinesterasemia(^b)</td>
<td>Alu (Yb8)</td>
<td>BCHE</td>
<td>Exon</td>
</tr>
<tr>
<td>Haemophilia B(^c)</td>
<td>Alu (Ya5)</td>
<td>F9</td>
<td>Exon</td>
</tr>
<tr>
<td>Familial hypocalciuric hypercalcemia(^d)</td>
<td>Alu (Ya4)</td>
<td>CASR</td>
<td>Exon</td>
</tr>
<tr>
<td>Breast cancer(^e)</td>
<td>Alu (Y)</td>
<td>BRCA2</td>
<td>Exon</td>
</tr>
<tr>
<td>Hereditary desmoid disease(^f)</td>
<td>Alu (Yb8)</td>
<td>APC</td>
<td>Exon</td>
</tr>
<tr>
<td>X-linked agammaglobulinemia(^g)</td>
<td>Alu (Y)</td>
<td>BTK</td>
<td>Exon</td>
</tr>
<tr>
<td>X-linked severe combined immuno-deficiency(^g)</td>
<td>Alu (Ya5)</td>
<td>IL2RG</td>
<td>Intron</td>
</tr>
<tr>
<td>Branchio-oto-renal syndrome(^h)</td>
<td>Alu (Ya5)</td>
<td>EYA1</td>
<td>Exon</td>
</tr>
<tr>
<td>Apert syndrome(^i)</td>
<td>Alu (Ya5)</td>
<td>FGFR2</td>
<td>Exon</td>
</tr>
<tr>
<td>Apert syndrome(^i)</td>
<td>Alu (Yb8)</td>
<td>FGFR2</td>
<td>Exon</td>
</tr>
<tr>
<td>Huntington disease(^j)</td>
<td>Alu</td>
<td>ADD1</td>
<td>Intron</td>
</tr>
<tr>
<td>Glycerol kinase deficiency(^k)</td>
<td>Alu (Ya5)</td>
<td>GK</td>
<td>Intron</td>
</tr>
<tr>
<td>C1 inhibitor deficiency(^l)</td>
<td>Alu (Y)</td>
<td>SERPING1</td>
<td>Intron</td>
</tr>
<tr>
<td>Acute intermittent porphyria(^m)</td>
<td>Alu (Ya5)</td>
<td>MMBS</td>
<td>Exon</td>
</tr>
<tr>
<td>Associated with leukemia(^n)</td>
<td>Alu (Ya5)</td>
<td>MLV12</td>
<td>Unknown</td>
</tr>
<tr>
<td>Haemophilia B(^o)</td>
<td>Alu (Ya3al)</td>
<td>F9</td>
<td>Exon</td>
</tr>
<tr>
<td>Haemophilia B(^p)</td>
<td>Alu</td>
<td>F9</td>
<td>Exon</td>
</tr>
<tr>
<td>Haemophilia A(^q)</td>
<td>Alu (Yb8)</td>
<td>F8</td>
<td>Exon</td>
</tr>
<tr>
<td>Haemophilia A(^r)</td>
<td>Alu (Yb9)</td>
<td>F8</td>
<td>Exon</td>
</tr>
</tbody>
</table>
Table A.2: L1 disease causing insertions. a: Kazazian et al. (1988); b: Woods-Samuels et al. (1989); c: Miki et al. (1992); d: Narita et al. (1993); e: Bakker E, van Ommen GJ, pers comm; f: Holmes et al. (1994); g: Divoky et al. (1996); h: Schwahn et al. (1998); i: Yoshida et al. (1998); j: Meischl et al. (1998); k: Meischl et al. (2000); l: Kondo-Iida et al. (1999); m: Li et al. (2001); n: van den Hurk et al. (2003); o: Mine et al. (2007);

<table>
<thead>
<tr>
<th>Phenotype observed</th>
<th>Inserted element</th>
<th>Gene disrupted</th>
<th>size (bp)</th>
<th>3’ transduction</th>
<th>Insertion site</th>
<th>Orientation (wrt gene)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haemophilia A&lt;sup&gt;a&lt;/sup&gt;</td>
<td>L1(JH-27)</td>
<td>F8</td>
<td>3800</td>
<td>None</td>
<td>Exon</td>
<td>Sense and rearranged</td>
</tr>
<tr>
<td>Haemophilia A&lt;sup&gt;a&lt;/sup&gt;</td>
<td>L1(JH-28)</td>
<td>F8</td>
<td>2200</td>
<td>None</td>
<td>Exon</td>
<td>Sense</td>
</tr>
<tr>
<td>Haemophilia&lt;sup&gt;b&lt;/sup&gt;</td>
<td>L1(JH25)</td>
<td>F8</td>
<td>681</td>
<td>None</td>
<td>Intron</td>
<td>Sense</td>
</tr>
<tr>
<td>Colon cancer&lt;sup&gt;c&lt;/sup&gt;</td>
<td>L1(APC)</td>
<td>APC</td>
<td>538</td>
<td>Yes</td>
<td>Exon</td>
<td>Unknown</td>
</tr>
<tr>
<td>Muscular Dystrophy&lt;sup&gt;d&lt;/sup&gt;</td>
<td>L1(Dyst1)</td>
<td>DMD</td>
<td>608</td>
<td>None</td>
<td>Exon</td>
<td>Sense</td>
</tr>
<tr>
<td>Muscular Dystrophy&lt;sup&gt;e&lt;/sup&gt;</td>
<td>L1(Dyst2)</td>
<td>DMD</td>
<td>878</td>
<td>None</td>
<td>Exon</td>
<td>Sense</td>
</tr>
<tr>
<td>Muscular Dystrophy&lt;sup&gt;f&lt;/sup&gt;</td>
<td>L1(JH1001)</td>
<td>DMD</td>
<td>2000</td>
<td>Yes</td>
<td>Exon</td>
<td>Sense and rearranged</td>
</tr>
<tr>
<td>β-thalassemia&lt;sup&gt;g&lt;/sup&gt;</td>
<td>L1(β-thal)</td>
<td>HBB</td>
<td>6000</td>
<td>No</td>
<td>Intron</td>
<td>Antisense</td>
</tr>
<tr>
<td>Retinitis Pigmentosa (RP)&lt;sup&gt;h&lt;/sup&gt;</td>
<td>L1(RP)</td>
<td>RP2</td>
<td>6000</td>
<td>No</td>
<td>Intron</td>
<td>Antisense</td>
</tr>
<tr>
<td>X-linked dilated cardiomyopathy (XLDCM)&lt;sup&gt;i&lt;/sup&gt;</td>
<td>L1(XLDCM)</td>
<td>DMD</td>
<td>524</td>
<td>No</td>
<td>Exon</td>
<td>Antisense</td>
</tr>
<tr>
<td>Chronic granulomatous disease (CGD)&lt;sup&gt;j&lt;/sup&gt;</td>
<td>L1(CYB)</td>
<td>CYBB</td>
<td>1700</td>
<td>Yes</td>
<td>Exon</td>
<td>Unknown</td>
</tr>
<tr>
<td>Chronic granulomatous disease (CGD)&lt;sup&gt;k&lt;/sup&gt;</td>
<td>L1(CYB)</td>
<td>CYBB</td>
<td>940</td>
<td>No</td>
<td>Intron</td>
<td>Sense</td>
</tr>
<tr>
<td>Fukuyama type congenital muscular dystrophy (FCMD)&lt;sup&gt;l&lt;/sup&gt;</td>
<td>L1(FCMD)</td>
<td>FKTN</td>
<td>1100</td>
<td>No</td>
<td>Intron</td>
<td>Sense</td>
</tr>
<tr>
<td>Haemophilia B&lt;sup&gt;m&lt;/sup&gt;</td>
<td>L1(FIX)</td>
<td>F9</td>
<td>520</td>
<td>No</td>
<td>Exon</td>
<td>Sense</td>
</tr>
<tr>
<td>choroideremia&lt;sup&gt;n&lt;/sup&gt;</td>
<td>L1(CHM)</td>
<td>CHM</td>
<td>6000</td>
<td>No</td>
<td>Exon</td>
<td>Unknown</td>
</tr>
<tr>
<td>Pyruvate dehydrogenase complex deficiency (PDHc) 7&lt;sup&gt;o&lt;/sup&gt;</td>
<td>L1(PDHX)</td>
<td>PDHX</td>
<td>6086</td>
<td>No</td>
<td>Deletion</td>
<td>Sense</td>
</tr>
</tbody>
</table>
Appendix B

Full-length L1 sequence discovery: Appendix

B.1 Number of FL-L1HS in different human genome assemblies

A table to show the number of FL-L1HS per chromosome in the three human genome assemblies meeting the criteria of $\geq 5922$ bp in length and $\geq 98\%$ identity to L1.3.mpa

B.2 Trace Archives

A table to show the number of potential new FL-L1HS per detected from a search of the June 2008 trace archives

B.3 FL-L1HS in genes

A table to show the distribution of FL-L1HS with intact and non-intact ORFs found in genes

B.4 Over-represented Gene Ontology terms
Table B.1: Number of FL-L1HS per chromosome in the three human genome assemblies meeting the criteria of $\geq 5922$ bp in length and $\geq 98\%$ identity to L1.3_mpa.

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Reference</th>
<th>Celera</th>
<th>HuRef</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>45</td>
<td>32</td>
<td>24</td>
</tr>
<tr>
<td>2</td>
<td>42</td>
<td>30</td>
<td>23</td>
</tr>
<tr>
<td>3</td>
<td>36</td>
<td>28</td>
<td>24</td>
</tr>
<tr>
<td>4</td>
<td>42</td>
<td>28</td>
<td>25</td>
</tr>
<tr>
<td>5</td>
<td>32</td>
<td>25</td>
<td>21</td>
</tr>
<tr>
<td>6</td>
<td>27</td>
<td>22</td>
<td>18</td>
</tr>
<tr>
<td>7</td>
<td>22</td>
<td>19</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>25</td>
<td>16</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>13</td>
<td>11</td>
<td>7</td>
</tr>
<tr>
<td>10</td>
<td>18</td>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td>11</td>
<td>19</td>
<td>15</td>
<td>10</td>
</tr>
<tr>
<td>12</td>
<td>20</td>
<td>19</td>
<td>10</td>
</tr>
<tr>
<td>13</td>
<td>12</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>14</td>
<td>13</td>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>15</td>
<td>9</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>16</td>
<td>11</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>17</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>18</td>
<td>18</td>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>5</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>21</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>22</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>X</td>
<td>38</td>
<td>30</td>
<td>1</td>
</tr>
<tr>
<td>Y</td>
<td>5</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>462</strong></td>
<td><strong>353</strong></td>
<td><strong>239</strong></td>
</tr>
</tbody>
</table>
Table B.2: Number of potential new FL-L1HS per detected from a search of the June 2008 trace archives

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Number of traces</th>
<th>Number of traces in genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>X</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Y</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>48</strong></td>
<td><strong>20</strong></td>
</tr>
</tbody>
</table>
Table B.3: Distribution of FL-L1HS with intact and non-intact ORFs in genes

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>$N^{o}$ intact FL-L1HS in genes</th>
<th>$N^{o}$ non-intact FL-L1HS in genes</th>
<th>$N^{o}$ FL-L1HS in genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>13</td>
<td>18</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>10</td>
<td>17</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>19</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>22</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>X</td>
<td>1</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>Y</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>54</strong></td>
<td><strong>110</strong></td>
<td><strong>164</strong></td>
</tr>
</tbody>
</table>
Table B.4: Hits 1-3 over-represented Gene Ontology terms from the analysis performed at GOstat (http://gostat.wehi.edu.au/)

<table>
<thead>
<tr>
<th>GO Term</th>
<th>Known Gene Identifier</th>
<th>N° genes</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:004459</td>
<td>PARD3B, ACVR1C, MME, XPR1, ERC2, GRID2, CDH13, LRP2, PTPRM, PKP4, EPHA3, CD96, DLGAP2, MMP16, KNCB2, PHACTR1, APP, MUSK, CLCN5, GPC5, SCN5A, KCNQ5, RYR2, GRIK2, PKHD1, NRCAM, HOMER2</td>
<td>27</td>
<td>2.91×10⁻⁶</td>
</tr>
<tr>
<td>GO:0031226</td>
<td>ACVR1C, MME, XPR1, GRID2, PTPRM, EPHA3, CD96, MMP16, KNCB2, APP, CLCN5, MUSK, GPC5, SCN5A, RYR2, KCNQ5, GRIK2, PKHD1, NRCAM</td>
<td>19</td>
<td>4.23×10⁻⁶</td>
</tr>
<tr>
<td>GO:004425</td>
<td>PARD3B, TPST1, SLC25A43, ACVR1C, XPR1, GRID2, IL20RB, FER1L6, CDH13, PKP4, EPHA3, CD96, PCDH9, SDK1, DLGAP2, MMP16, TSPAN5, ANTXR2, FRAS1, PHACTR1, APP, HEPH, CLCN5, SORCS1, PDE3B, C14ORF37, GRIK2, ATP10B, TMEM117, IMMP2L, PCDH11Y, ODZ4, NUP62CL, NRCAM, UXS1, MME, PGAP1, CNTNAP2, ERC2, SLC10A7, LRP2, PTPRM, B3GALT1, SLC22A15, CD38, LPHN3, PLEKHH2, KNCB2, SCFD1, MUSK, GPC5, SCN5A, KCNQ5, RYR2, MS4A13, HCN1, SLC24A2, PKHD1, HOMER2</td>
<td>59</td>
<td>5.84×10⁻⁶</td>
</tr>
</tbody>
</table>
Table B.5: Hits 4-7 over-represented Gene Ontology terms from the analysis performed at GOstat (http://gostat.wehi.edu.au/)

<table>
<thead>
<tr>
<th>GO Term Identifier</th>
<th>Known Gene Identifier</th>
<th>N⁰genes</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0005886</td>
<td>PARD3B, ACVR1C, XPR1, GRID2, CDH13, PKP4, EPHA3, PCDH9, CD96, DLGAP2, MMP16, PRMT8, LSAMP, ANTXR2, PHACTR1, NEGR1, APP, CLCN5, GRIK2, PCDH11Y, NRCAM, MME, ERC2, LRP2, PTPRM, KCNIP4, CD38, LPHN3, KNCB2, MUSK, GPC5, SCN5A, RYR2, KCNQ5, PKHD1, HOMER2</td>
<td>36</td>
<td>$8.44 \times 10^{-6}$</td>
</tr>
<tr>
<td>GO:0051179</td>
<td>SLC25A43, GRID2, CDH13, SNX30, COL24A1, HEPH, APP, SPOCK1, CLCN5, PDE3B, RAB3IP, GRIK2, ATP10B, IMMP2L, RASEF, NUP62CL, SIL1, NRCAM, KIF6, LAMA2, SH3GL3, PGAP1, PRKCA, COL19A1, SLC10A7, LRP2, KNCIP4, SLC22A15, KNCB2, SCFD1, SPIRE1, CAMK4, EXOC4, SCN5A, KNCQ5, RYR2, KIF5C, SLC24A2, HCN1, PKHD1</td>
<td>40</td>
<td>$8.44 \times 10^{-6}$</td>
</tr>
<tr>
<td>GO:0022610</td>
<td>COL19A1, CNTNAP2, CNTNAP5, CDH13, PTPRM, COL24A1, PKP4, CD96, PCDH9, SDK1, LSAMP, NEGR1, APP, SPOCK1, PCDH11Y, PKHD1, NRCAM, LAMA2,</td>
<td>18</td>
<td>$8.44 \times 10^{-6}$</td>
</tr>
<tr>
<td>GO:0007155</td>
<td>COL19A1, CNTNAP2, CNTNAP5, CDH13, PTPRM, COL24A1, PKP4, CD96, PCDH9, SDK1, LSAMP, NEGR1, APP, SPOCK1, PCDH11Y, PKHD1, NRCAM, LAMA2,</td>
<td>18</td>
<td>$8.44 \times 10^{-6}$</td>
</tr>
</tbody>
</table>
Table B.6: Hits 8-10 over-represented Gene Ontology terms from the analysis performed at GOstat (http://gostat.wehi.edu.au/)

<p>| GO ontology and definitions for the top 10 most significantly over-represented terms |
|------------------------------------------|----------------|----------------|</p>
<table>
<thead>
<tr>
<th>GO Term</th>
<th>Known Gene Identifier</th>
<th>(N^0)genes</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0016020</td>
<td>PARD3B, TPST1, SLC25A43, XPR1, IL20RB, PKP4, EPHA3, HPSE2, CD96, SDK1, MMP16, PHLP, HCN1, HOMER2 TSPAN5, ANTXR2, PHACTR1, NEGR1, APP, HEPH, C14ORF37, PCDH11Y, ODZ4, NRCAM, SH3GL3, UXS1, ERC2, SLC10A7, LRP2, KCNIP4, LPHN3, SCFD1, KCNB2, MUSK, GPC5, MS4A13, SLC24A2, PKHD1, ACVR1C, GRID2, FER1L6, CDH13, PCDH9, DLGAP2, PRMT8, LSAMP, FRAS1, SBF2, CLCN5, SORCS1, PDE3B, GRIK2, ATP10B, TMEM117, IMMP2L, NUP62CL, MME, PGAP1, DOCK4, CNTNAP2, PTPRM, B3GALT1, SLC22A15, CD38, PLEKHH2, SCN5A, RYR2, KCNQ5,</td>
<td>68</td>
<td>(8.44 \times 10^{-6})</td>
</tr>
<tr>
<td>GO:0005887</td>
<td>ACVR1C, MME, XPR1, GRID2, PTPRM, EPHA3, CD96, MMP16, KCNB2, APP, CLCN5, MUSK, GPC5, SCN5A, RYR2, KCNQ5, GRIK2, NRCAM</td>
<td>18</td>
<td>(8.44 \times 10^{-6})</td>
</tr>
<tr>
<td>GO:0006810</td>
<td>SLC25A43, GRID2, CDH13, SNX30, COL24A1, HEPH, APP, CLCN5, PDE3B, RAB3IP, GRIK2, ATP10B, IMMP2L, RASEF, NUP62CL, SIL1, KIF6, SH3GL3, PGAP1, COL19A1, PRKCA, SLC10A7, LRP2, KCNIP4, SLC22A15, KCNB2, SCFD1, SPIRE1, CAMK4, SCN5A, EXOC4, RYR2, KCNQ5, KIF5C, SLC24A2, HCN1</td>
<td>36</td>
<td>(3.06 \times 10^{-6})</td>
</tr>
</tbody>
</table>
Appendix C

TSD detection and analysis : Appendix

C.1 Development of a TSD scoring matrix from endonuclease cut sites

Base occurrences at different positions of TSDs and the surrounding regions (8bp and 20 bp).
Table C.1: Observed base pair occurrences at positions -8 to +8 (16bp) flanking the 5’ TSD start. The most abundant nucleotide at each position is column P.

<table>
<thead>
<tr>
<th>Base position</th>
<th>P</th>
<th>T</th>
<th>C</th>
<th>A</th>
<th>G</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>-8</td>
<td>A</td>
<td>57</td>
<td>43</td>
<td>58</td>
<td>38</td>
<td>196</td>
</tr>
<tr>
<td>-7</td>
<td>A</td>
<td>56</td>
<td>45</td>
<td>69</td>
<td>26</td>
<td>196</td>
</tr>
<tr>
<td>-6</td>
<td>T</td>
<td>68</td>
<td>34</td>
<td>64</td>
<td>30</td>
<td>196</td>
</tr>
<tr>
<td>-5</td>
<td>A</td>
<td>59</td>
<td>37</td>
<td>61</td>
<td>39</td>
<td>196</td>
</tr>
<tr>
<td>-4</td>
<td>A</td>
<td>57</td>
<td>38</td>
<td>68</td>
<td>33</td>
<td>196</td>
</tr>
<tr>
<td>-3</td>
<td>A</td>
<td>56</td>
<td>41</td>
<td>63</td>
<td>36</td>
<td>196</td>
</tr>
<tr>
<td>-2</td>
<td>T</td>
<td>111</td>
<td>11</td>
<td>60</td>
<td>14</td>
<td>196</td>
</tr>
<tr>
<td>-1</td>
<td>T</td>
<td>135</td>
<td>38</td>
<td>8</td>
<td>15</td>
<td>196</td>
</tr>
<tr>
<td>1</td>
<td>A</td>
<td>13</td>
<td>2</td>
<td>171</td>
<td>10</td>
<td>196</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>8</td>
<td>0</td>
<td>160</td>
<td>28</td>
<td>196</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>11</td>
<td>2</td>
<td>138</td>
<td>45</td>
<td>196</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>13</td>
<td>0</td>
<td>159</td>
<td>24</td>
<td>196</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>33</td>
<td>15</td>
<td>119</td>
<td>29</td>
<td>196</td>
</tr>
<tr>
<td>6</td>
<td>A</td>
<td>66</td>
<td>20</td>
<td>93</td>
<td>17</td>
<td>196</td>
</tr>
<tr>
<td>7</td>
<td>A</td>
<td>58</td>
<td>28</td>
<td>77</td>
<td>33</td>
<td>196</td>
</tr>
<tr>
<td>8</td>
<td>A</td>
<td>57</td>
<td>29</td>
<td>86</td>
<td>24</td>
<td>196</td>
</tr>
<tr>
<td>Composition %</td>
<td>N/A</td>
<td>27.36</td>
<td>12.21</td>
<td>46.36</td>
<td>14.06</td>
<td>100</td>
</tr>
</tbody>
</table>

The table above shows the observed base pair occurrences at positions -8 to +8 (16bp) flanking the 5’ TSD start. The most abundant nucleotide at each position is column P.
Table C.2: Observed base pair occurrences at positions -20 to +20 (40bp) flanking the 5' TSD start. The most abundant nucleotide at each position is column P.

<table>
<thead>
<tr>
<th>Base position</th>
<th>P</th>
<th>T</th>
<th>C</th>
<th>A</th>
<th>G</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>-20</td>
<td>A</td>
<td>47</td>
<td>37</td>
<td>66</td>
<td>46</td>
<td>196</td>
</tr>
<tr>
<td>-19</td>
<td>A</td>
<td>54</td>
<td>37</td>
<td>63</td>
<td>42</td>
<td>196</td>
</tr>
<tr>
<td>-18</td>
<td>A</td>
<td>61</td>
<td>31</td>
<td>67</td>
<td>37</td>
<td>196</td>
</tr>
<tr>
<td>-17</td>
<td>A</td>
<td>53</td>
<td>30</td>
<td>77</td>
<td>36</td>
<td>196</td>
</tr>
<tr>
<td>-16</td>
<td>A</td>
<td>59</td>
<td>27</td>
<td>74</td>
<td>36</td>
<td>196</td>
</tr>
<tr>
<td>-15</td>
<td>A</td>
<td>53</td>
<td>26</td>
<td>69</td>
<td>48</td>
<td>196</td>
</tr>
<tr>
<td>-14</td>
<td>A</td>
<td>61</td>
<td>26</td>
<td>82</td>
<td>27</td>
<td>196</td>
</tr>
<tr>
<td>-13</td>
<td>A</td>
<td>66</td>
<td>30</td>
<td>75</td>
<td>25</td>
<td>196</td>
</tr>
<tr>
<td>-12 T/A</td>
<td>A</td>
<td>67</td>
<td>27</td>
<td>67</td>
<td>35</td>
<td>196</td>
</tr>
<tr>
<td>-11</td>
<td>A</td>
<td>54</td>
<td>36</td>
<td>77</td>
<td>29</td>
<td>196</td>
</tr>
<tr>
<td>-10</td>
<td>A</td>
<td>61</td>
<td>32</td>
<td>71</td>
<td>32</td>
<td>196</td>
</tr>
<tr>
<td>-9</td>
<td>A</td>
<td>53</td>
<td>37</td>
<td>60</td>
<td>46</td>
<td>196</td>
</tr>
<tr>
<td>-8</td>
<td>A</td>
<td>57</td>
<td>43</td>
<td>58</td>
<td>38</td>
<td>196</td>
</tr>
<tr>
<td>-7</td>
<td>T</td>
<td>56</td>
<td>45</td>
<td>69</td>
<td>26</td>
<td>196</td>
</tr>
<tr>
<td>-6</td>
<td>T</td>
<td>68</td>
<td>34</td>
<td>64</td>
<td>30</td>
<td>196</td>
</tr>
<tr>
<td>-5</td>
<td>A</td>
<td>59</td>
<td>37</td>
<td>61</td>
<td>39</td>
<td>196</td>
</tr>
<tr>
<td>-4</td>
<td>A</td>
<td>57</td>
<td>38</td>
<td>68</td>
<td>33</td>
<td>196</td>
</tr>
<tr>
<td>-3</td>
<td>A</td>
<td>56</td>
<td>41</td>
<td>63</td>
<td>36</td>
<td>196</td>
</tr>
<tr>
<td>-2</td>
<td>T</td>
<td>111</td>
<td>11</td>
<td>60</td>
<td>14</td>
<td>196</td>
</tr>
<tr>
<td>-1</td>
<td>T</td>
<td>135</td>
<td>38</td>
<td>8</td>
<td>15</td>
<td>196</td>
</tr>
<tr>
<td>1</td>
<td>A</td>
<td>13</td>
<td>2</td>
<td>171</td>
<td>10</td>
<td>196</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>8</td>
<td>0</td>
<td>160</td>
<td>28</td>
<td>196</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>11</td>
<td>2</td>
<td>138</td>
<td>45</td>
<td>196</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>13</td>
<td>0</td>
<td>159</td>
<td>24</td>
<td>196</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>33</td>
<td>15</td>
<td>119</td>
<td>29</td>
<td>196</td>
</tr>
<tr>
<td>6</td>
<td>A</td>
<td>66</td>
<td>20</td>
<td>93</td>
<td>17</td>
<td>196</td>
</tr>
<tr>
<td>7</td>
<td>A</td>
<td>58</td>
<td>28</td>
<td>77</td>
<td>33</td>
<td>196</td>
</tr>
<tr>
<td>8</td>
<td>A</td>
<td>57</td>
<td>29</td>
<td>86</td>
<td>24</td>
<td>196</td>
</tr>
<tr>
<td>9</td>
<td>A</td>
<td>41</td>
<td>26</td>
<td>80</td>
<td>49</td>
<td>196</td>
</tr>
<tr>
<td>10</td>
<td>A</td>
<td>53</td>
<td>34</td>
<td>66</td>
<td>43</td>
<td>196</td>
</tr>
<tr>
<td>11</td>
<td>A</td>
<td>60</td>
<td>27</td>
<td>61</td>
<td>48</td>
<td>196</td>
</tr>
<tr>
<td>12</td>
<td>A</td>
<td>43</td>
<td>40</td>
<td>59</td>
<td>54</td>
<td>196</td>
</tr>
<tr>
<td>13</td>
<td>G</td>
<td>53</td>
<td>23</td>
<td>50</td>
<td>70</td>
<td>196</td>
</tr>
<tr>
<td>14</td>
<td>G</td>
<td>32</td>
<td>36</td>
<td>59</td>
<td>69</td>
<td>196</td>
</tr>
<tr>
<td>15</td>
<td>G</td>
<td>34</td>
<td>19</td>
<td>45</td>
<td>98</td>
<td>196</td>
</tr>
<tr>
<td>16</td>
<td>G</td>
<td>24</td>
<td>16</td>
<td>51</td>
<td>105</td>
<td>196</td>
</tr>
<tr>
<td>17</td>
<td>G</td>
<td>21</td>
<td>13</td>
<td>47</td>
<td>115</td>
<td>196</td>
</tr>
<tr>
<td>18</td>
<td>G</td>
<td>13</td>
<td>9</td>
<td>52</td>
<td>122</td>
<td>196</td>
</tr>
<tr>
<td>19</td>
<td>G</td>
<td>12</td>
<td>13</td>
<td>50</td>
<td>121</td>
<td>196</td>
</tr>
<tr>
<td>20</td>
<td>G</td>
<td>10</td>
<td>11</td>
<td>61</td>
<td>114</td>
<td>196</td>
</tr>
</tbody>
</table>

Base composition %

| N/A | 24.78 | 13.09 | 38.05 | 24.08 | 100 |
References


REFERENCES


REFERENCES


REFERENCES


IHGSC (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011), 931–945. 8


REFERENCES


REFERENCES


REFERENCES


REFERENCES


C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J.,
Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A.,
Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz,
S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E.,
Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandraoumiswaran, I., Charlab,
R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C.,
Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins,
M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin,
X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A.,
Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J.,
A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong,
W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, G., Cravchik,
Barrow, I., Beeson, K., Busam, D., Carver, D., Center, A., Cheng, M.L., Curry, L.,
Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S.,
Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, F., Kline, L.,
Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen,
I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V.,
Qureshi, H., Readon, R., Rogers, Y.H., Romblad, D., Ruhfel, B.,
Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R.,
Tint, N.N., Tse, S., Veich, C., Wang, G., Wetter, J., Williams, S., Williams, M.,
Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J.F., Guigo,
R., Campbell, M.J., Sjolander, K.V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B.,
Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna,
V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu,
A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y.H.,
Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham,
S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M.,
Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings,
D., Jordan, C., Jordan, J., Kashy, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M.,
Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M.,
Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W.,
Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R.,

Vidaud, D., Vidaud, M., Bahnak, B.R., Siguret, V., Gispert Sanchez, S., Laurian, Y.,
novo insertion of a human-specific Alu subfamily member within the coding region

REFERENCES
REFERENCES


REFERENCES


