CHARACTERISATION OF LENS EPITHELIUM DERIVED GROWTH FACTOR ISOFORMS IN CHRONIC LYMPHOCYTIC LEUKEMIA

STUDIES AND SIGNIFICANCE

Thesis submitted in accordance with the requirements of the

University of Liverpool for the degree of Doctor in Philosophy

by

Jemma Blocksidge

May 2014

Contents

Abstract

Over expression of lens epithelium derived growth factor (LEDGF) is described in a number of different tumours. In addition, self-antibodies against this protein are detectable in patients with lymphomas and chronic lymphocytic leukemia (CLL). Alternative splicing of LEDGF results in a number of different isoforms, and thus the initial aim of this thesis was to determine the variability in LEDGF isoform expression in a cohort of CLL cases. Utilising RT-PCR, it is shown that CLL cells variably express four different LEDGF isoforms. The study was then extended to quantitate the expression levels of these four isoforms in individual CLL cases, and to compare them to normal B cells. RT-qPCR analysis showed that the longest of these isoforms, p75, is significantly over expressed, whilst the shortest form, p52bΔE6 is significantly under expressed in CLL, compared to normal B cells. All four isoforms of LEDGF are generally over expressed in CD38 positive CLL cases compared to those that are deficient for surface expression of this molecule.

Alternative splicing occurs as a result of the inclusion or exclusion of different exons within the mRNA. This process is governed by the splicing machinery, of which Splicing Factor 3B, subunit 1 (SF3B1) is a critical component. Recently, it has been shown that mutations within SF3B1 are recurrent in CLL. Such mutations impact on the splicing of a significant array of genes within the cell. One of the genes predicted to be affected is LEDGF. Therefore, the mutational status of SF3B1 was examined in the same cohort of CLL cases and correlated with the expression of the LEDGF isoforms. As well as the commonly reported mutations in SF3B1, a novel nonsense mutation was identified. Although not statistically significant, cases with unmutated SF3B1 have higher levels of all isoforms of LEDGF and a possible explanation for this is discussed.

That two of the major isoforms of LEDGF (p75 and p52) have differing functions within the cell is well documented and can be explained by virtue of their participation in distinct protein complexes. In order to identify components of the complexes generated by the four LEDGF isoforms identified in this study, stable cell lines exogenously expressing these isoforms were generated. Each of the isoforms was ‘double-tagged’, and a purification protocol optimised that would allow successful characterisation of the complex associated with each isoform.

The results provide useful insights and provoke thought for the role of alternative splicing in the altered phenotype of cancer cells and its contribution as an additional epigenetic mechanism in normal and malignant gene regulation.

List of Figures

List of Tables

Acknowledgements

I would like to thank all the members of the Division of Haematology, Department of Molecular and Clinical Cancer Medicine, for their support over the last four years. In particular, I want to give appreciation to all the people who gave me their ideas and feedback for my work.

I would like express my great thanks to my primary supervisor Dr. Nagesh Kalakonda for his encouragement, ideas and enthusiasm throughout my PhD studies. Also, I would like to thank him for giving me the opportunity to perform my PhD research in his laboratory. I would also like to thank Dr. Joseph Slupsky for his help and feedback.

A special mention must be given to Dr. Mark Glenn. Without his constant help and support practically and academically, this thesis would not have been possible.

I would like to thank Dr. Mark Wilkinson and Dr. Fotis Polydoros for their help and input into my research.

Finally, I would like to thank my family and friends, especially Paul, for all of their support during my practical work and period of writing up.

Declaration

All practical laboratory work, except for the complex purification work using the low pressure chromatography system, was performed in the Cancer Epigenetics laboratory of Dr. Nagesh Kalakonda, Department of Molecular and Clinical Cancer Medicine, Institute of Translational Medicine, University of Liverpool.

The complex purification work using the low pressure chromatography system (Chapter 5) was performed under the supervision of Dr. Mark Wilkinson, Institute of Integrative Biology, University of Liverpool.

All of the work presented in this thesis is my own apart from the statistical analysis which was performed by Dr. Fotis Polydoros.

The practical laboratory work presented in this thesis was funded by The Royal Liverpool Hospital Leukemia Fund.

Abbreviations

ALL - Acute lymphocytic leukemia
AML - Acute myeloid leukemia
ANOVA - One-Way Analysis of Variance
APS - Ammonium persulfate
BB - Boiled beads
BCL-2 - B-cell lymphoma 2
BCR - B cell receptor
BH3 - Bcl-2 homology domain 3
bp - Base pairs
BSA - Bovine serum albumin
CBB - Coomassie Brilliant Blue
Cdc7-ASK - Cell division cycle 7-activator of S-phase kinase
cDNA - Complimentary deoxyribonucleic acid
CIP - Calf intestinal alkaline phosphatase
CLB - Clear lysis buffer
CLL - Chronic lymphocytic leukaemia
CML - Chronic myeloid leukemia
CO_2 - Carbon dioxide
COSMIC - Catalogue Of Somatic Mutations In Cancer
CR - Charged region
Ct - Threshold value
DMEM - Dulbecco's Modified Eagle's Medium
DMSO - Dimethyl sulfoxide
DNA - Deoxyribonucleic acid
DNMT - DNA methyltransferase
DTT - Dithiothreitol
E - Elution
EDTA - Ethylenediaminetetraacetic acid
F/T - Flow-through
FCS - Fetal calf serum
FGFR - Fibroblast growth factor receptor
G - Gravity
G1 - Gap 1 phase
G2 - Gap 2 phase
GLFG - Glycine/leucine/phenylalanine/glycine
Hr - Hour
H3K27me^3 - Histone 3 Lysine 27 trimethylation
H3K36me^3 - Histone 3 Lysine 36 trimethylation
H3K4 - Histone 3 Lysine 4
H3K4me^1 - Histone 3 Lysine 4 monomethylation

$H3K4me^3$	-	Histone 3 Lysine 4 trimethylation
H3K9me1	-	Histone 3 Lysine 9 monomethylation
HATH	-	Homologous to amino terminus of HDGF
HCl	-	Hydrochloric acid
HDGF	-	Hepatoma-derived growth factor
HEAT	-	Huntington, Elongation factor 3, protein phosphatase 2A, Targets of rapamycin 1
HEK	-	Human embryonic kidney cells
His	-	Histidine
HIV	-	Human immunodeficiency virus
hnRNP	-	Heterogeneous ribonucleoprotein
HOX	-	Homeobox
HRP	-	Horseradish peroxidise
HSE	-	Heat shock elements
HSP	-	Heat shock protein
IBD	-	Integrase binding domain
IGHV	-	Immunoglobulin heavy chain variable region genes
IN	-	Integrase
IP	-	Immunoprecipitated
kDa	-	Kilo daltons
L	-	Litre
LB	-	Luria-Bertani
LEDGF	-	Lens epithelium derived growth factor
Lip	-	Lipofectin
mA	-	Milliamps
MDS	-	Myelodysplastic syndrome
MeCP2	-	Methyl CpG binding protein 2
MEN1	-	Menin
mg	-	Milligrams
Min	-	Minute
miRNA	-	microRNA
mL	-	Millilitre
MLL	-	Mixed lineage leukaemia
mM	-	Millimole
mRNA	-	mRNA
NaCl	-	Sodium chloride
NaH_2PO_4	-	Sodium phosphate monobasic
NB	-	Normal B
ncRNA	-	Noncoding RNA
NFW	-	Nuclease-free water
ng	-	Nanogram
Ni	-	Nickel
NLS	-	Nuclear localisation signal
Nm	-	Nanometre

NMD	-	Nonsense mediated decay
NUP98	-	Nucleoporin 98
O/n	-	Overnight
OS	-	Overall survival
P	-	Purified
PAGE	-	Polyacrylamide gel electrophoresis
PBS	-	Phosphate buffered saline
PC4	-	Positive cofactor 4
PCR	-	Polymerase chain reaction
PHD	-	Plant homeodomain
PogZ	-	Pogo transposable element-derived protein with zinc finger
PSIP1	-	PC4- and SFRS1-interacting protein 1
PTB	-	Polypyrimidine tract–binding protein
PTD	-	Partial tandem duplication
PVDF	-	Polyvinylidene fluoride
PWWP	-	Proline-tryptophan-tryptophan-proline
RCF	-	Relative centrifugal force
RNA	-	Ribonucleic acid
RPL27	-	Ribosomal protein L27
rpm	-	Revolutions per minute
RT	-	Room Temperature
RT-PCR	-	Reverse transcription polymerase chain reaction
RT-qPCR	-	Quantitative reverse transcription polymerase chain reaction
SDS	-	Sodium dodecyl sulphate
SEREX	-	Serological identification by recombinant expression cloning
SET	-	Su(var)3-9 and 'Enhancer of zeste'
SF3b	-	Splicing factor 3b complex
SF3B1	-	Splicing Factor 3B, subunit 1
snRNP	-	Small nuclear ribonucleotide protein
SR	-	Serine/arginine
SRD	-	Supercoiled recognition domain
SRSF1	-	Serine/arginine-rich splicing factor 1
STRE	-	Stress response elements
T(Δ-S)	-	Time between diagnosis and sampling
TBE	-	Tris/Borate/EDTA
TBS	-	Tris buffered saline
TEMED	-	Tetramethylethylenediamine
TFT	-	Time to first treatment
TP53	-	Tumor protein 53
Trx-G	-	Trithorax group
TSG	-	Tumour suppressor gene
TSS	-	Transcriptional start sites
UTR	-	Untranslated region

UV - Ultraviolet
V - Volts
VEGF - Vascular endothelial growth factor
v.c - Viable cells
W - Wash
WBC - White blood cell
WCL - Whole cell lysate
WE - Wash end
WHSC1 - Wolf-Hirschhorn syndrome candidate 1
WT - Wild type
µg - Microgram
µL - Microlitre
µM - Micromolar

Cell line	Characteristics
HEK293	Human embryonic kidney cell line. Hypotriploid cell line.
293T	Highly transfectable derivative of HEK293 parental line. Contains SV40 T-antigen.
Hela	Epithelial cell line, contains telocentric chromosome.
KCL22	Cell line model of Chronic Myeloid Leukemia in blast crisis. Positive for Philadelphia chromosome.
K562	Cell line model of Chronic Myeloid Leukemia displaying morphology of lymphoblasts.
Jurkat	Cell line model of Acute T cell Leukemia displaying morphology of lymphoblasts.

Table - A Table of cell lines used for the investigations presented in this thesis.

The use of each cell line is described.

Chapter 1 : General Introduction

1.1. Lens Epithelium Derived Growth Factor

LEDGF is so called as it was isolated from a human lens epithelial library[1]. It is also known as PC4- and SFRS1-interacting protein 1 (*PSIP1*) by virtue of existing as a product of this gene located on chromosome 9p22.3[2]. LEDGF is primarily expressed in 2 alternative splice forms - LEDGF/p75 and LEDGF/p52. Both protein products have identical N-terminal sequences up to amino acid residue 325. The C terminal domain of LEDGF/p75 is 205 amino acids, resulting in a protein that is 530 amino acids in length. In contrast, LEDGF/p52 has a smaller C-terminal tail of 8 residues (333 amino acids).

1.1.1. Transcriptional regulation of LEDGF

It has been shown that the expression of LEDGF/p75 and p52 occurs from two separate transcriptional start sites (TSS) (Figure 1-1)[3]. Using Luciferase reporter experiments the promoter region for LEDGF/p75 (-723/+59) was shown to be ~12x more active than the region around the LEDGF/p52 TSS (+140/+781)[3]. The minimal promoter regions required to drive expression of p75 and p52 are located between -112/+59 and +609/+781 respectively. More interestingly, a region located between

+319/+397 appears to regulate the ratio between LEDGF/p75 and LEDGF/p52 expression.

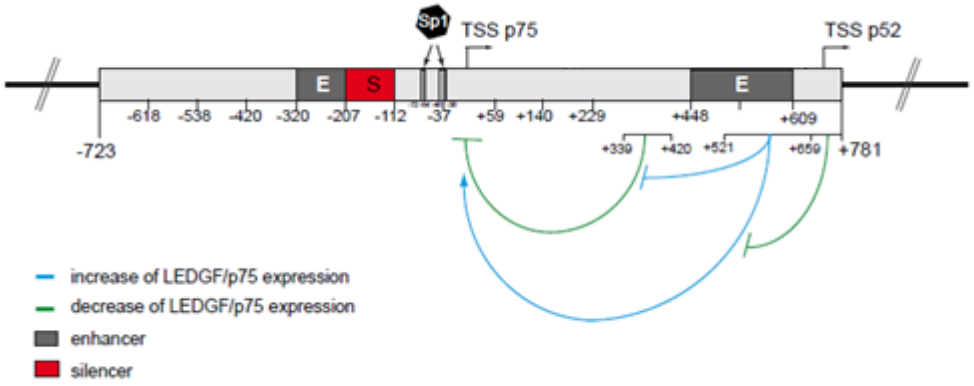


Figure 1-1: Model for transcriptional regulation of the *PSIP1* gene.

The promoter region of the gene is shown with the positions of the Enhancer (E) and Silencer (S) sequences highlighted. Expression of the two major isoforms of LEDGF, p75 and p52 are from separate TSS. Regions of the promoter that are involved in increasing or decreasing the transcription of LEDGF/p75 are shown. Stimulation is shown by arrows whereas repression is shown by a bar.

1.1.2. Structure of LEDGF

The common N-terminus of LEDGF/p75 and p52 encodes a PWWP domain (residues 1-93), a nuclear localisation signal (NLS) (residues 145-156), a pair of AT-hooks (178-198) and three charged regions (CR). The C-terminus of LEDGF/p75 contains a

further two charged regions as well as an integrase binding domain (IBD) (339-442) (Figure 1-2)[4].

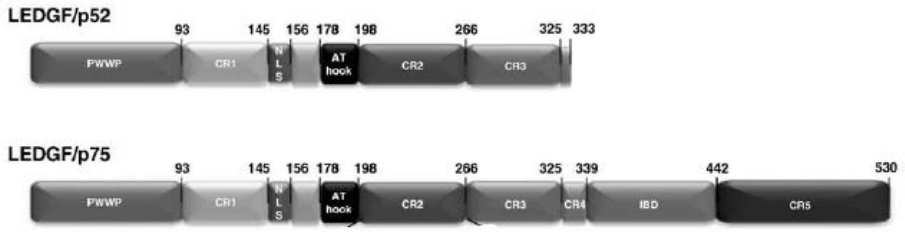


Figure 1-2: Representation of LEDGF/p75 and p52 protein structure.

The two isoforms are identical in sequences up to and including residue 325. Each isoform has a unique C-terminal tail. The common domains of the two isoforms are PWWP domain (PWWP), charged regions 1-3 (CR1-3), nuclear localisation signal (NLS) and AT hooks. The unique C-terminus of LEDGF/p52 is composed of 8 residues. The unique C-terminus of LEDGF/p75 is composed of charged regions 4 and 5 (CR4 and 5) and an integrase binding domain (IBD). The numbers designate the amino acids that form each domain.

Domain	Presence in LEDGF/p75 and p52	Function of domain
PWWP	Yes	Targets the protein to chromatin
Charged region 1	Yes	Aids the chromatin binding of LEDGF
Charged regions 2 and 3	Yes	Influence the activity of the AT hooks
Nuclear localization	Yes	Directs the proteins to the nucleus
AT hooks	Yes	Help protein dock onto minor groove of DNA at AT-rich region.
Integrase binding domain	LEDGF/p75 only	Docking platform for other proteins
Charged regions 4 and 5	LEDGF/p75 only	Role is unknown.

Table 1-1: Summary of the roles of the different domains in LEDGF.

The name of each domain in LEDGF is detailed on the left and the role of each domain is detailed on the right.

The PWWP domain, first identified in the Wolf-Hirschhorn syndrome candidate 1 protein (WHSC1)[5], is so called because of a conserved quartet of proline-tryptophan-tryptophan-proline residues. Although the 'PWWP' sequence is

generally conserved, there are critical variations (Figure 1-3)[6] that explain the ability of this family of proteins to interact with a variety of targets.

Initially the PWWP domain was predicted to be responsible for protein-protein interactions[7]. However, in 2002 it was demonstrated that mammalian DNMT3B binds to DNA via this domain[8]. Accumulating evidence suggests that as well as its' DNA binding ability, the PWWP domain can bind to methylated histones[9]. More recently, affinity for methylated lysines is described[10-12].

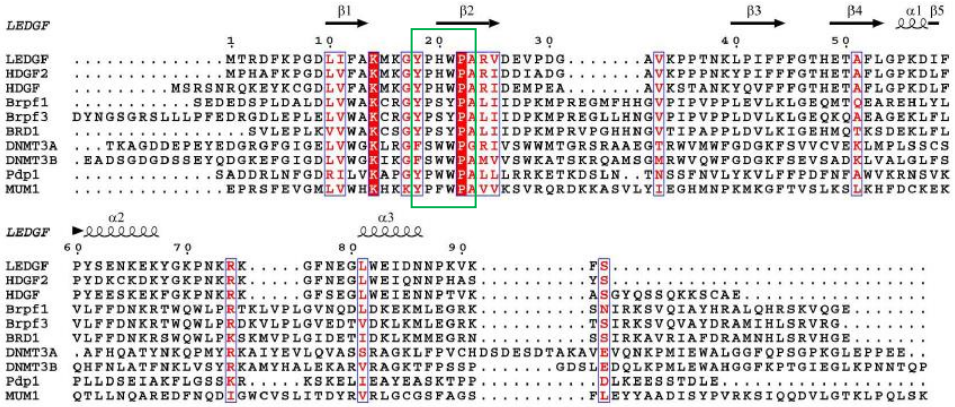


Figure 1-3: Structure-based sequence alignment of human PWWP domains.

The green box shows the conserved PWWP quartet of residues. The letters highlighted by a red box are invariant whilst other coloured residues (red in a blue outline box) have high global similarity scores.

Proteins containing the PWWP domain are histone modification 'readers'. That is, they are responsible for eliciting the correct response to the presence or absence of the histone marks[13]. Histones are primarily modified on their basic N-terminal tails

and these modifications affect the ability of the nucleosomes to interact with both DNA and other proteins[14]. Histone modifications read by domains such as the PWWP domain are acetylation of lysine residues and methylation of lysine and arginine residues.

Histone post-translational modification is a form of epigenetics. Epigenetics is an umbrella term encompassing molecular mechanisms that result in a change of phenotype without an alteration in DNA sequence. First proposed in 1939 by C.H Waddington[15], the definition has since been updated to include heritable changes in gene expression that are not due to any alteration in the DNA sequence[16]. Historically, epigenetics refers to DNA methylation, non-coding RNA (ncRNA) expression, histone post-translational modifications and alternative splicing. Epigenetic regulation is a fine tuned process and likely explains the variability in the phenotypes observed within populations. Dysregulation of the epigenome is now well documented in cancer and significantly contributes to neoplastic cell behaviour, variability in therapeutic responses and disease outcomes.

The Hepatoma-derived growth factor (HDGF) related proteins include HDGF-related proteins-1 to 4, LEDGF and HDGF. All members of this group share a common N-terminal PWWP domain-containing region (also called homologous to amino terminus of HDGF, (HATH) region) but their C-terminii vary in length and charge[17]. This group of proteins have functional roles inside and outside the plasma membrane. The PWWP domain of LEDGF, however, has a strong tethering effect that directs the protein to chromatin, and hence LEDGF is found to be constitutively localised in the nucleus[18].

The NLS sequence of LEDGF, described independently by Maertens *et al*. and Vanegas *et al*. [19, 20], conforms to the classical basic family. Lysine[150] was shown to be the critical residue and its mutation abolished nuclear localisation[19]. Interestingly, LEDGF with a mutant NLS was shown to be localised in the cytoplasm but upon passage, the mutant becomes captured and is then chromatin bound[20].

CR1 aids the chromatin-binding pattern of LEDGF, whereas CR2 and CR3 influence the AT hooks[21]. These are small DNA-binding motifs, centred around an arginine-glycine-arginine-proline sequence, that enable the protein to dock on the minor groove of DNA at AT-rich regions and alter the architecture of the DNA, thus allowing the access of transcription factors[22]. Turlure *et al.* illustrated that the two AT hooks and the NLS together mediate the binding of LEDGF/p75 to DNA, although the degree of contribution to this binding varied (AT2>NLS>AT1)[23]. The PWWP domain was also shown to contribute to the binding of LEDGF to chromatin, as a domain mutant was shown to affect the distribution of the protein in the nucleolus during cell division[23].

More recently, using a rat brain cDNA library, Tsutsui *et al*. illustrated that LEDGF/p75 recognises negatively supercoiled DNA via a novel DNA binding domain they termed supercoiled-DNA recognition domain (SRD) (residues 200-326) which is located within CR2 and 3[18]. The SRD was divided into a core region (200-274) which recognises the supercoiled DNA and an enhancer (275-336) which potentiates the interaction. It should be noted that whilst the work of both Turlure *et al*. and Tsutsui *et al*. focused on LEDGF/p75, Tsutsui *et al*. also show that both LEDGF/p75 and LEDGF/p52 are able to bind negatively supercoiled DNA[18, 23].

The PWWP domain, particularly that of LEDGF/p75, is important in targeting the protein (and associated binding partners) to specific sites on chromatin[6]. Thus, whilst other domains in the protein (the NLS, AT hooks etc.) are involved in aligning the protein with protein and DNA, the specificity of the binding location is determined by the PWWP domain. Pradeepa *et al*. suggest that the PWWP domain of LEDGF binds to the Histone 3 Lysine 36 trimethylation mark ($H3K36me^3$)[24]. This has since been confirmed by Eidahl *et al.* and van Nuland *et al*.[6, 25]. More importantly, Eidahl *et al*. demonstrated that the PWWP domain binds to purified mononucleosomes via interactions with core histones[6]. Using nuclear magnetic resonance, they identified two distinct regions of LEDGF PWWP domain structure, viz. a basic surface (made up of seven lysine residues and one arginine) that interacts non-specifically with DNA, and a hydrophobic cavity that interacts with the $H3K36me^3$ (Figure 1-4)[6].

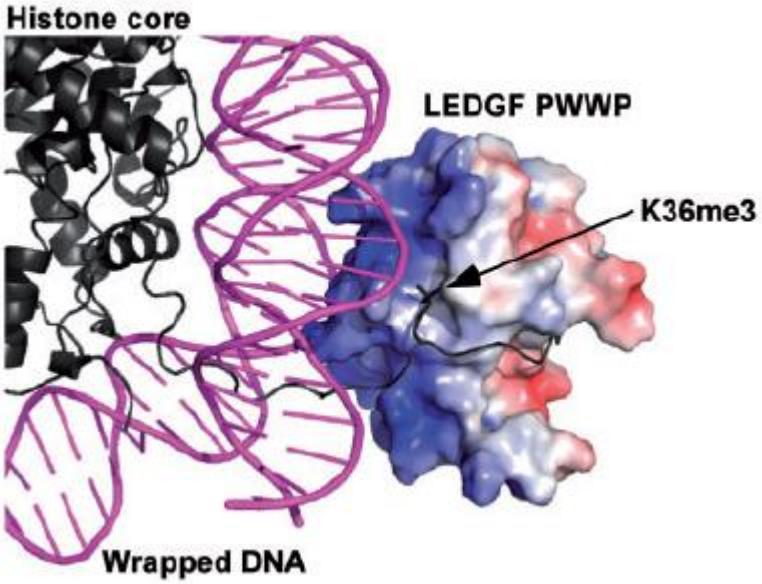


Figure 1-4: The binding of the LEDGF PWWP domain to a modified histone tail. The double stranded DNA molecule (purple) wrapped around the histone core (black) is shown on the left. The modified histone tail protruding from the histone core is shown binding to the PWWP domain of LEDGF (blue, right).

LEDGF/p75 and p52 are shown to have distinct localisation within the cell that is dependent on the cell cycle. Nishizawa and colleagues used green fluorescent tagged proteins transfected into live CHO-K1, MDCK and NRK cells to show that LEDGF/p75 and p52 were both predominantly in the nucleus[26]. Whereas LEDGF/p75 was uniformly distributed, LEDGF/p52 was segregated into distinct nuclear compartments. In Gap 1 phase (G1), LEDGF/p75 was present diffusely throughout the nucleoplasm, whereas LEDGF/p52 was located at the nuclear periphery and towards late G1 phase, formed a speckle-like pattern. As the cell entered S-phase, LEDGF/p75 began to cluster around the chromosomes. In contrast to this, LEDGF/p52 formed speckles at the nuclear periphery but these became larger in

size and progressed towards the nuclear centre. During Gap 2 phase (G2) when the chromosomes have started to condense and the centromeres are visible, the speckles formed by LEDGF/p52 coalesced into a cylindrical pattern but LEDGF/p75 segregated with chromatin. During prophase, the distribution pattern of LEDGF/p75 was retained, whereas for LEDGF/p52, the speckled pattern was lost, and the protein formed a complete cylindrical shape around the chromatin. In metaphase, when the chromosomes are aligned halfway between the two poles of the nucleus, the pattern of LEDGF/p52 remained but LEDGF/p75 developed a striped pattern. During telophase, both LEDGF/p75 and LEDGF/p52 moved to the poles of the nucleus. When the nucleus then divided (cytokinesis), equal amounts of each isoform were present in each daughter cell and LEDGF/p52 kept the cylindrical pattern but for LEDGF/p75, the nuclear positioning became diffuse.

1.2. Role of LEDGF

The unique C-terminal tails of the two isoforms helps mediate their differing interactions with cellular proteins (reviewed in chapter 5). It is this binding to distinct proteins that helps LEDGF/p75 and p52 perform their distinct and unique roles in the cell.

1.2.1. Role of LEDGF/p52

The list of described and potential binding partners of LEDGF/p52[24, 27] (discussed in chapter 5) suggests that this isoform has a role in the alternative splicing of class II

genes, and it has been proposed to coordinate activated transcription and pre-mRNA splicing by binding to, and modulating the activity of SRSF1 (formerly SFRS1 and ASF/SF2)[27]. By showing the interaction between LEDGF/p52 (via its' PWWP domain) and H3K36me^3, Pradeepa *et al.* provided evidence for a link between LEDGF/p52, chromatin and mRNA splicing[24]. By using a mutant LEDGF/p52 which lacked the C-terminal region, they showed that alternative splicing of target genes was affected, and that these effects were reversed by wild type LEDGF/p52. However, it should be noted that the mutant would fail to bind SRSF1 with implications for mRNA splicing.

1.2.2. Role of LEDGF/p75

Both LEDGF/p52 and p75 were shown to be transcriptional coactivators of the general transcription machinery following their purification with PC4[28]. However, in comparison to LEDGF/p52, LEDGF/p75 is a weaker coactivator. Whilst the downstream targets of the transcriptional activity of LEDGF/p52 are largely unknown (although Heat shock protein 27 (Hsp27) has been suggested as one)[29], those for LEDGF/p75 are discussed later.

LEDGF/p75 is suggested to have a role in growth and development, as highlighted by the lethality and abnormalities that result from its deletion as described by Sutherland *et al.*[30].

Integration of Lentiviruses, such as Human Immunodeficiency Virus type 1 (HIV-1), equine infectious anaemia virus and feline immunodeficiency virus into the host

genome is dependent on LEDGF/p75. Depletion of the protein renders cells immune to infection by these agents[31, 32]. Cuiffi *et al.* prepared cells depleted in LEDGF/p75 and showed that integration into optimal regions of the genome was reduced thereby suggesting that the choice of integration site is dictated by LEDGF/p75[33].

LEDGF/p75 acts as a dual tether by binding the integrase enzyme of HIV (HIV-IN) via its' C-terminal IBD, and chromatin/DNA via its N-terminal PWWP domain, and directs the HIV pre-integration complex to transcriptionally active regions of the host genome. By swapping the N-terminal region of LEDGF for other chromatin binding domains (namely the plant homeodomain (PHD) finger and chromodomain), Ferris *et al*. showed that the targeting of HIV-IN can be re-directed[34]. In addition to the tethering function, LEDGF/p75 protects HIV-IN from proteolytic degradation[35].

Mixed lineage leukemia (MLL) is a Histone 3 Lysine 4 (H3K4) trimethylase that, via LEDGF/p75, associates with the tumour suppressor menin. LEDGF/p75 functions to direct the MLL/menin complex allowing it to place the transcriptionally active Histone 3 Lysine 4 trimethylation mark ($H3K4me^3$) on chromatin (Figure 1-5)[36] . Therefore, the MLL/menin complex is a histone writer complex.

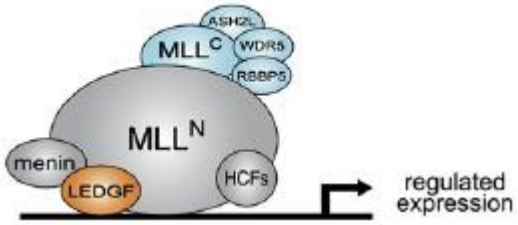


Figure 1-5: LEDGF/p75 binds to menin and MLL and directs the complex to chromatin.

Formation of the complex is necessary to allow the placement of the activation mark, H3K4me^3 on this chromatin. This results in regulated gene expression.

MLL binds, via its' LEDGF binding domain next to the binding site of menin, specifically to the IBD domain in the C-terminus of LEDGF/p75[37]. Menin is a ubiquitously expressed nuclear protein that acts as a TSG[38]. Menin binds to the very N-terminus of MLL and to the IBD of LEDGF/p75. Loss of function results in multiple endocrine neoplasia type 1 (MEN1)[39].

MLL is a member of the trithorax group (Trx-G) of transcriptional activators, and is one of three types of complexes that methylate H3K4[40]. The N-terminus is involved in directing the complex to specific sites, whereas the C-terminus encodes the enzymatic activity. By regulating expression of the *homeobox (HOX)* transcription factors, MLL plays an important role in the generation of the body plan and haematopoiesis[41, 42]. In addition MLL is involved in regulating cell cycle progression and is itself regulated by the cell cycle[42].

To date, more than 70 known translocations involving MLL have been described[40]. Such translocations are detected in ~10% of all leukemias, but the majority are found in infant patients (>70% infant Acute Lymphocytic Leukemia (ALL)), and are associated with a poor prognosis[43]. MLL translocations always involve fusion of the N-terminus of MLL (breakpoints occur between exons 8 and 12) to the C-terminal region of a partner gene such that MLL loses its PHD fingers, Taspase I cleavage site, transactivation domain and the *Su(var)3-9 and 'Enhancer of zeste'* (SET) methyltransferase domain[42]. However, with the retention of the N-terminus, the fusion protein can still bind LEDGF/p75 (and menin) and is therefore targeted to chromatin, whilst the C-terminal region of the fusion partner is still able to recruit their respective complexes (Figure 1-6)[36]. For example, fusions of MLL and AF9, ENL or AF10 are still able to interact with Dot1L, leading to up regulation of *HOX* genes and ultimately leukemogenesis[44].

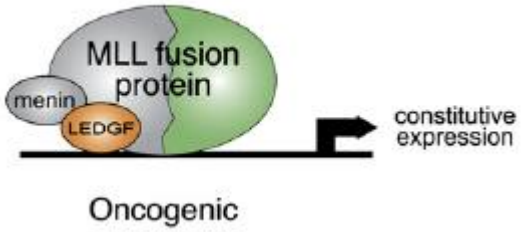


Figure 1-6: Translocations of MLL resulting in abnormal fusion proteins leads to abnormal gene expression.

The C-terminus of MLL is subject to translocations but retains its' N-terminus. Therefore, MLL can still form a complex with LEDGF/p75 and menin and is directed to chromatin. The fusion protein will recruit different binding partners compared to the complex containing wild type MLL and results in abnormal constitutive gene expression leading to leukemogenesis.

In addition to translocations, MLL can be disrupted by the inclusion of a partial tandem duplication (PTD) of N-terminal exons which causes repetition of the DNA binding domains and increased recruitment of binding partners to these domains. This abnormality has been described in acute myeloid leukemia (AML) and myelodysplastic syndrome (MDS)[41].

During micro-environmental stress (for example, serum starvation, thermal and oxidative stress), LEDGF/p75 has been shown to bind to stress response elements (STRE) and heat shock elements (HSE) in the promoter regions of target genes[45]. This results in the induced expression of downstream targets such as Hsp27 and the stress-related proteins αβ-crystallin, AOP2 and alcohol dehydrogenase[46]. Furthermore, LEDGF/p75 has been shown to transactivate the expression of anti-

apoptotic genes, thus enhancing cell survival[46]. A simple mechanism of the survival promoting action of LEDGF is shown in Figure 1-7[46].

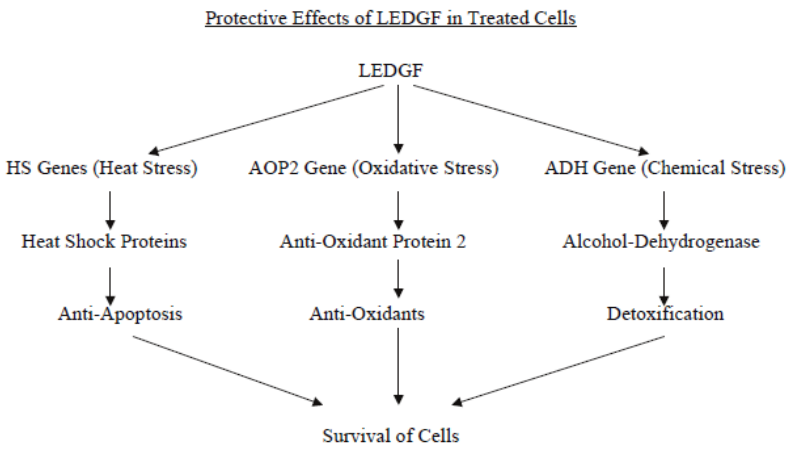


Figure 1-7: Mechanism of action of LEDGF following the encounter to environmental stress.

Upon stress, the expression of LEDGF/p75 is up regulated. This causes the activation of stress related proteins including heat shock proteins, anti-oxidant protein 2 and alcohol dehydrogenase. These proteins allow the cells to overcome the stress and survive.

More recently, it has been shown that following thermal and oxidative stress of H1299 cells, LEDGF/p75 binds to the STRE of Vascular endothelial growth factor (VEGF)-C thereby inducing its expression[47]. LEDGF/p75 has also been implicated in hormonal up-regulation of VEGF-C in tumours, leading to angiogenesis and lymphangiogenesis[48].

Isoform	Roles of each of the isoforms
LEDGF/p75	• Transcriptional co-activator • Growth and development • Molecular tether targeting the MLL complex and HIV-IN to chromatin • Promote cell survival against thermal and oxidative stress.
LEDGF/p52	• Alternative splicing of class II genes • Coordinate transcription and splicing • Transcriptional co-activator

Table 1-2: Summary of the roles of LEDGF/p75 and p52 within the cell.

LEDGF/p75 and p52 are two alternatively spliced products of the same gene yet they have differing roles within the cell. These roles are detailed above.

1.3. Cleavage of LEDGF by Caspases

LEDGF/p75 and p52 are both cleaved by caspase enzymes[29]. Initially, Wu *et al.* showed that LEDGF/p75 was cleaved by caspases 3 and 7 resulting in a disruption of the PWWP domain, thus diminishing its pro-survival function[49]. LEDGF/p52 is also cleaved by these caspases into a 38 kDa pro-apoptotic fragment which inhibits the ability of LEDGF/p75 to transactivate Hsp27[29,] a known downstream target of LEDGF/p75, as discussed in Chapter 1.1.2.

Disruption of the PWWP domain of the two isoforms has opposing affects in the cell. Thus, whilst deletion of an N-terminal portion of LEDGF/p75 increased its ability to transactivate Hsp27[45], disruption of the PWWP of LEDGF/p52 by caspase cleavage inhibits its ability to transactivate Hsp27[29]. Therefore, the PWWP domain may have a transcriptional repression function in LEDGF/p75, but may be essential for LEDGF/p52 mediated transcriptional activation function[29]. Additionally, LEDGF/p52 was shown to have a pro-apoptotic effect that contrasts to the pro-survival function of LEDGF/p75.

1.4. Role of LEDGF in Leukemia

Over expression of LEDGF is suggested to have a role in tumourigenesis. Thus, LEDGF/p75 is a cancer-associated protein and is over expressed in many cancers, whilst LEDGF/p52 over expression has been demonstrated in tumour cell lines[29] (reviewed in Chapter 3). Both LEDGF/p75 and p52 are involved in translocations with Nucleoporin 98 (NUP98) in AML, chronic myeloid leukemia (CML) and biphenotypic leukemia[50-56]. As a consequence of these translocations, the glycine/leucine/phenylalanine/glycine (GLFG) repeats of NUP98 fuse to the C-terminus of LEDGF which results in disruption or loss of the PWWP domain. It should be noted that the breakpoints in LEDGF in each case are different. Furthermore, four different isoforms of LEDGF have been described in AML[57] (reviewed in Chapter 3).

Due to the presence of antinuclear autoantibodies that target LEDGF/p75 (but not LEDGF/p52) it has been suggested this protein also plays a role in autoimmunity[58].

Antibodies against LEDGF/p75 have also been identified in patients with CLL and other lymphomas, suggesting this is a tumour associated antigen[59].

1.5. Chronic Lymphocytic Leukemia

CLL is a common leukemia characterised by an accumulation of long lived CD5+ B lymphocytes in the peripheral blood with involvement of the bone marrow and/or lymphoid organs. It can either be a rapidly progressing and aggressive disease that is often refractory to treatment, or it can follow a benign indolent course with patients requiring little or no intervention. To date, no single pathognomonic translocation or mutation that gives rise to the disease has been identified, but several cytogenetic abnormalities including deletion of 11q, 13q, 17p, trisomy 12, tumor protein 53 (TP53) mutations and aberrant microRNA (miRNA) expression are commonly described.

The disease occurs due to a failure of apoptosis which causes the accumulation of clonal mature lymphocytes. Important to this process are lymphoid organs such as the lymph nodes and spleen, where the CLL cells receive pro-survival signals from cytokines and growth factors within the micro-environment[60, 61]. In CLL cells, stimulation of the B-cell receptor (BCR) is important for the activation of several downstream kinase enzymes (Src family, PI3 Kinase etc.) which in turn lead to activation of transcription factors such as NF-κB resulting in malignant cell proliferation and survival[61].

Due to the heterogeneity of disease behaviour, several prognostic markers have been identified. Traditionally, patients were classified according to either the Rai or Binet staging system, which graded disease severity based on the absence or presence of organomegaly and suppression of normal bone marrow function. However, it is now routine for B cells to be analysed for their IGHV status (class and deviation from germline sequence), CD38% and ZAP-70 expression and presence of cytogenetic abnormalities such as those outlined earlier. Notably, certain prognostic factors can be associated with a long survival time. Namely, Low Rai and Binet staging (for example, 0 and A respectively) (12+ years), 13q- (133 months), mutated IGHV (293 months compared to 117 months for patients with unmutated IGHV), CD38- and ZAP70-[62].

CLL currently remains incurable, although recent advances in therapy allows for effective treatment in some patients. Treatment of the disease (reviewed by Hallek *et al*.[61]) was initially with the alkylating agent Chlorambucil (CLB). However, the introduction of newer therapies such as Fludarabine and monoclonal antibodies such as Rituximab (anti-CD20) or Campath/Alemtuzumab (anti-CD52) has resulted in more durable remissions and complete responses than has previously been observed. Notably, these different classes of drugs are increasingly administered in combination with each other (for example fludarabine combined with cyclophosphamide and fludarabine used in conjunction with rituximab). There is also increasing interest in inhibiting pathways important in the survival of CLL cells, for example, the use of kinase inhibitors.

As mentioned earlier, no single genetic abnormality has been identified that gives rise to CLL. Recently, however, whole genome and whole exome sequencing have pointed to the presence of recurring mutations in patients with CLL. An initial report published by Wang *et al*. described nine genes that were significantly and recurrently mutated in patients with CLL[63]. One of these, the SF3B1 gene, was the second most frequently mutated gene (15%) (after TP53) in the cohort investigated. All of the mutations described were missense, with K700E being the most common. Subsequent reports have confirmed that this site is one of the most commonly occurring in the gene and a change from lysine to asparagine (K700N) has also been demonstrated[64]. Interrogation of the Catalogue of Somatic Mutations in Cancer (COSMIC) database suggested that SF3B1 was recurrently mutated in cancer (for example, breast 1.52%, liver 1.82% and eye 17.86%) and that it was the most frequently mutated gene in CLL (Table 1-3).

Gene name	Mutation frequency in CLL
SF3B1	13%
NOTCH1	11%
ATM	11%
TP53	10%
MYD88	4%
POT1	3%
XPO1	2%
CDKN2A	2%
BCL11B	2%

Table 1-3: Summary of the ten most frequently mutated genes in CLL.

Data available in the COSMIC database describes genes recurrently mutated in CLL. The table above details the top ten genes that are recurrently mutated in CLL and their frequencies.

There has been a focussed effort to analyse different cohorts of patients with CLL to determine the mutational status of the SF3B1 gene and uncover any correlations with clinical variables. For example, Jeromin *et al*. (2013) published a comprehensive study of 1160 untreated CLL patients and identified several correlations between SF3B1 mutation and other prognostic factors[64]. Thus, mutated SF3B1 was highly correlated with unmutated IGHV. With regards cytogenetic abnormalities, SF3B1 mutation was associated with deletion of 11q. Mutated SF3B1

was shown to have an adverse effect on both time to first treatment and overall survival. Mutations in SF3B1 tended to be more common in males, in patients with high white cell counts, advanced Binet stage and CD38 positivity greater than 30%. It has been suggested that the degree of variability in the reported correlations between SF3B1 mutational status and clinical factors is due to the heterogeneity of CLL[65].

1.6. SF3B1

SF3B1 is a member of the splicesome which catalyses the removal of introns from pre-mRNA and is predicted to have an effect on the alternative splicing of the *PSIP1* gene. The spliceosome is a large multi-subunit complex, of which there are 2 in mammals: the major (U2-type) and the minor (U12-type)[66]. Both types of spliceosome contain the SF3b complex[67], of which SF3B1 is the largest component[68]. During the removal of introns by the U2 splicesosome, SF3B1 binds via its N terminus to the branch-point sequence at the 3´ end of the intron. This aids with the recruitment of the U2 small nuclear ribonucleotide protein (snRNP) to the branch point, one of the initiating steps in the assembly of the spliceosome[69].

The protein is composed of a hydrophilic N-terminal RNA binding domain and the C-terminal contains 22 Huntington, Elongation factor 3, protein phosphatase 2A, Targets of rapamycin 1 (HEAT) repeats (Figure 1-8)[70]. The HEAT repeats are predicted to wrap around and stabilise the U2snRNP[67] and act as a docking site for other components of the spliceosome to assemble[71].

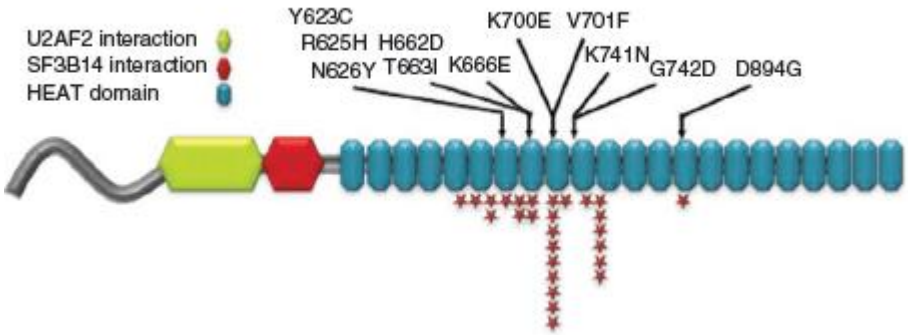


Figure 1-8: Structure of the human SF3B1 protein.

The structural domains of the protein are detailed. The protein has an N-terminal U2AF2 interaction domain followed by SF3B14 interaction domain. The C-terminus of the protein is composed of 22 HEAT repeats. As an example, mutations in the SF3B1 gene described by Quesada *et al*.[70] are noted. All mutations described are in the HEAT repeat region of the protein. The frequency of each mutation within the cohort is inferred by red stars. In this cohort, K700E is the most common mutation.

As discussed earlier, recurring mutations of SF3B1 have been described in CLL. The majority of these mutations are localised to the HEAT repeats, more specifically exons 14-16[70]. In this report, the authors also identified a number of genes with differential exon inclusion between mutated and wild type SF3B1 samples, one of these being LEDGF *(PSIP1*) with a p-value of 0.0001 suggesting the mutational status of the SF3B1 gene will have a direct effect on the splicing of LEDGF. The idea that SF3B1 mutations might alter the splicing of other genes that affect the pathobiology of CLL was also suggested by Wang *et al*.[63]. Further, Furney *et al*. showed that mutations within SF3B1 have a direct impact on differential alternative splicing of other genes in uveal melanoma[72].

Given that SF3B1 is one of the most recurrently mutated genes in CLL, and that such mutations are generally associated with a poor prognosis and shorter overall survival and time to treatment, there are suggestions to include the mutation status of this gene in a prognostic index[73]. This notion is still not fully accepted as different groups have described different criteria for their index[64, 73]. In addition, the number of patients investigated needs to be increased in order to ensure the index (and inclusion of the mutated genes) is suitable for use[74]. Nonetheless, with the correlations that have been suggested for SF3B1 in CLL[65], it is felt important to consider inclusion of SF3B1 (and other mutated genes) mutational status in a potential index[75].

1.7. Alternative Splicing

The mechanism of alternative splicing has been shown to be sensitive to chromatin architecture. Luco *et al*. (2010) examined alternative splicing of the human fibroblast growth factor receptor (*FGFR2*) and showed that in human mesenchymal stem cells, H3K36me^3 and Histone 3 Lysine 4 monomethylation (H3K4me^1) were enriched around the alternatively spliced region of *FGFR2* where exon IIIb is repressed, whereas the Histone 3 Lysine 27 trimethylation mark (H3K27me^3), H3K4me^3, and the Histone 3 Lysine 9 monomethylation mark (H3K9me^1) were reduced when compared to PNT2 cells, where the exon is included[76]. The exclusion of exon IIIb was shown to be dependent on the binding of polypyrimidine tract–binding protein (PTB) to silencing elements around the exon, resulting in its repression. In HEK293 cells, over expression of the SET2 enzyme responsible for the

H3K36me^3 mark resulted in reduced inclusion of this exon in *FGFR2*. The change in splicing pattern was not due to change in *FGFR2* or other factors illustrating it was a result of the histone modifications.

Alternative splicing is also regulated by the presence or absence of splicing enhancer and silencer sequences (*cis*-acting elements) present within the exons or introns that either promote or prevent the use of a particular splice site via the recruitment of serine/arginine (SR) proteins or heterogeneous nuclear ribonucleoproteins (hnRNP) (*trans*-acting splicing factors) respectively[77]. These splicing factors are themselves regulated by their expression levels and post-translational modification status (for example, methylation)[78].

1.7.1. Alternative Splicing as an Epigenetic Mechanism of Gene Regulation

Alternative splicing is a major mechanism of epigenetic regulation within the cell (Figure 1-9)[79]. The definition of an epigenetic mechanism is one which leads to a change in phenotype without changing the genomic sequence. By altering the inclusion or exclusion of exons, alternative splicing gives rise to different mRNA sequences, and hence, proteins with potentially different functions, from the same original coding DNA. Alternative splicing, like DNA methylation, histone modification and ncRNAs, is a fine-tuned mechanism that generates diversity within the cell. There is also increasing evidence of cross-talk between alternative splicing and the other accepted epigenetic mechanisms, further supporting that alternative splicing may be considered as an epigenetic phenomenon.

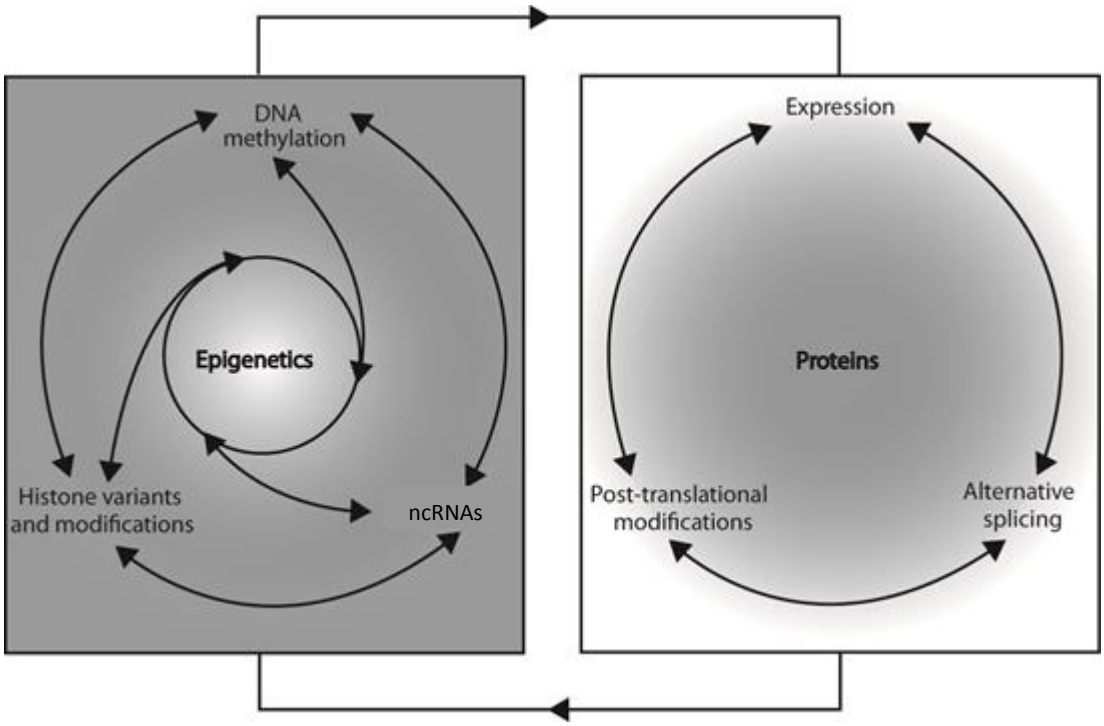


Figure 1-9: Epigenetic processes and regulation of gene expression.

The interplay of DNA methylation, post-translational modification of histones, and ncRNA expression are key for the integrity of the epigenome. Epigenetic processes regulate the expression of proteins (as well as their post-translational modifications and alternative splicing), which in turn affect the chromatin landscape of a cell.

1.8. Aims

The aims of my work were:

1. To determine the presence/absence of the different isoforms of LEDGF in CLL cells and normal B cells using RT-PCR.
2. To determine if the isoforms have different expression levels (using RT-qPCR) and correlate these with known clinical variables.
3. To assess the mutational status of the HEAT repeats in the SF3B1 gene in CLL samples and determine if this correlates with the expression pattern of the LEDGF isoforms.
4. To establish stable cell lines expressing tagged isoforms of LEDGF identified in Aim 1 above.
5. To develop a method that would enable each of the isoforms and their associated complexes to be purified thus allowing future identification of components of these complexes.

Chapter 2 : General Materials and Methods

This chapter is divided into 3 sections. Chapter 2.1 describes the cell culture techniques used throughout this study, Chapter 2.2 outlines the molecular techniques employed and Chapter 2.3 details the methods used for protein analysis. Details of techniques that are specific to each chapter, or where general techniques were modified slightly, are addressed in the relevant chapter.

Aqueous solutions were made using reverse osmosis purified water. Where required, solutions were sterilised either by autoclaving at 15 psi for 15 minutes or by filter sterilisation through 0.22 micron syringe filters.

2.1. Cell Culture

2.1.1. Culturing of cell lines

Cell lines were cultured in a sterile environment ready for use. Complete Dulbecco's Modified Eagle's Medium (DMEM) (Sigma Aldrich, UK) was prepared by the addition of FCS (final 10% v/v; Sigma Aldrich, UK), L-glutamine (2 mM; Hyclone, UK), penicillin and streptomycin (1000 units and 1 mg respectively; Sigma Aldrich, UK) to the DMEM base media.

2.1.2. Thawing of cryopreserved cell lines

Cells stored at -150 °C were revived by thawing and cultured ready for use. One vial of the cell line was removed from -150 °C storage and thawed in a waterbath set to 37 °C. The volume of cells was then washed in pre-warmed complete media and pelleted for 5 minutes at 500 x g, at room temperature (RT). Supernatant was discarded and the pellet resuspended in 5 mL complete media. The cell suspension was transferred to a T-25 cm^2 vented culture flask and incubated at 37 °C, 5% CO_2, in a humidified incubator.

2.1.3. Cell line passage

Cell lines were passaged bi-weekly to maintain their viability and suitability for use. Suspension cells were passaged by dilution. Cells were transferred to a suitable culture flask at a ratio of 1:5 (cells:media) of the required final volume and complete media was added. Adherent cells were passaged following trypsinisation. Briefly, media was discarded and the cell monolayer was washed with PBS. Trypsin-EDTA (Sigma Aldrich, UK) was added (1 mL to a T-25 cm^2 flask, 2 mL to a T-75 cm^2 flask and 3 mL to a T-175 cm^2 flask) and the cells incubated at room temperature for 5 minutes. Cells were dislodged by tapping the culture flask and the trypsin neutralised by the addition of complete media. The cell suspension was collected and pelleted for 5 minutes at 500 x g, RT. Supernatant was discarded and the pellet resuspended in complete media. Cells were transferred to a suitable culture flask and made up to the required volume, with complete media. All cultures (adherent and suspension) were incubated at 37 °C and 5% CO_2 in a humidified incubator. If

cell counts were required, they were performed manually using a Haemocytometer with Trypan Blue dye (Sigma Aldrich, UK) in order to determine the cell viability.

2.1.4. Cryopreservation of cells

A number of vials of cells were prepared and stored at -150 °C to provide a cell bank for future work. Cells were cultured as described. A manual cell count was performed and the required volume of cells to give a density of $1x10^7$ viable cells (v.c) per vial was calculated. The cells were pelleted for 5 minutes at 500 x g, RT, and supernatant was discarded. The cell pellet was washed in PBS and then resuspended in complete media containing 10% DMSO (Sigma Aldrich, UK). The cell suspension was aliquoted into cryovials (1 mL per vial) (Sigma Aldrich, UK) and stored at -150 °C.

2.2. Molecular Biology

2.2.1. RNA Extraction

RNA extraction is the purification of RNA from biological samples. Total RNA was extracted from cells using the RNeasy Mini Kit (Qiagen, UK) following the protocol 'Purification of Total RNA from Animal Cells Using Spin Technology'. Briefly, the cell pellet was resuspended in Buffer RLT supplemented with β-mercaptoethanol (Sigma Aldrich, UK) then transferred to a QIAshredder column within a 2 mL collection tube. This was centrifuged for 2 minutes at maximum speed. An equal volume of 70% ethanol (Sigma Aldrich, UK) was added to the flow-through and mixed. Seven

hundred microlitres of the lysate was transferred to an RNeasy spin column which had been placed within a 2 mL collection tube, and centrifuged for 15 seconds at ≥8000 x g. The flow-through was discarded and the step repeated for the remaining volume of lysate. Buffer RW1 (350 µL) was added to the RNeasy spin column and centrifuged for 15 seconds at ≥8000 x g. DNase I was prepared by mixing 10 µL DNase I stock solution and 70 µL Buffer RDD. This was transferred directly to the membrane of the RNeasy spin column and incubated for 15 minutes on the benchtop. After the incubation, 350 µL of Buffer RW1 was added to the RNeasy spin column. The tube was centrifuged for 15 seconds at ≥8000 x g and the flow-through discarded. The RNeasy spin column was washed by adding 500 µL buffer RPE and centrifuging for 15 seconds at ≥8000 x g and the flow through discarded. This was repeated but centrifuged for 2 minutes. The RNeasy spin column was carefully transferred to a clean 2 mL collection tube and centrifuged for 1 minute at full speed. The RNeasy spin column was then transferred to a clean 1.5 mL microcentrifuge tube and 30 µL of nuclease-free water (NFW) was added directly to the membrane. This was centrifuged for 1 minute at ≥8000 x g. The eluate was collected and replaced onto the membrane of the RNeasy spin column in the same 1.5 mL microcentrifuge tube and centrifuged for 1 minute at ≥8000 x g. The extracted RNA was stored at -80 °C ready for downstream applications. All volumes described were for the extraction of RNA from one cell pellet of $1x10^7$ viable cells.

2.2.2. cDNA synthesis

cDNA synthesis is the generation of DNA from a template of RNA. cDNA was prepared using 1 µg of total RNA to which was added 1 µL of Oligo(dT)$_{12-18}$ Primer (Life Technologies, UK), 1 µL 10 mM dNTP Mix (Life Technologies, UK) and NFW to give a final volume of 12 µL. The solution was mixed and heated for 5 minutes at 65 °C after which, 2 µL 0.1 M DTT (Life Technologies, UK) and 4 µL 5x First-Strand Buffer (Life Technologies, UK) was added. The solutions were mixed then heated for 2 minutes at 42 °C. After heating, 1 µL SuperScript II Reverse Transcriptase (Life Technologies, UK) enzyme was added. The solutions were mixed then incubated for 50 minutes at 42 °C followed by a further 15 minutes at 72 °C. The resultant cDNA was stored at -80 °C. All volumes described were for one reaction. A negative control was always prepared with 1 µL NFW added instead of 1 µL SuperScript II Reverse Transcriptase enzyme.

2.2.3. RT-PCR Amplification

RT-PCR is the amplification of a specific target of DNA. Primers for RT-PCR were designed using software available at http://eu.idtdna.com/Scitools/Applications/Primerquest/ such that at least one of the primers crossed exon-exon boundaries but still retained the required properties. All primers were purchased from Eurofins MWG Operon (Germany).

Master mixes for the required number of RT PCR reactions were always prepared. Each individual reaction required 22 µL Platinum PCR SuperMix High Fidelity (Life Technologies, UK) and 1 µL of each forward and reverse PCR primer (stock

concentration 2.5 µM, final concentration of 0.1 µM each). The mastermix was aliquoted into the number of tubes required and 1 µL of template was added. A negative control using 1 µL of NFW, instead of sample, was always prepared. Unless specified, samples were run on the following cycle:

94 °C 2 minutes (x1)

94 °C 30 seconds, Annealing temperature 30 seconds, 68 °C for 15 seconds (x35)

68 °C 5 minutes (x1)

Hold at 4 °C

The optimum annealing temperature of primer pairs was determined using cDNA prepared from cell lines and an annealing temperature gradient approximately 5 degrees below to 5 degrees above the annealing temperature recommended by the primer manufacturer (specific temperatures are given in the relevant sections). A negative control was always included.

2.2.4. Agarose Gel Electrophoresis

Agarose gel electrophoresis is a technique used to separate a DNA molecules based on their size. Agarose gels were prepared by melting the required amount of agarose (Web Scientific, UK) in 150 mL 0.5% TBE buffer (Sigma Aldrich, UK) using a microwave. The molten agarose was allowed to cool slightly then 8 µL of 10 mg/mL Ethidium Bromide (Promega, UK) was added. The agarose was poured into a

prepared gel tray and allowed to set. When solid, the gel was transferred to a running tray and overlayed with 0.5% TBE buffer. Samples were diluted in 6x loading buffer (10 mM Tris-HCl (Fisher Scientific, UK) pH 8.0, 50 mM EDTA (Sigma-Aldrich, UK), 15% Ficoll-400 (Sigma Aldrich, UK), 0.5% orange G, (Sigma Aldrich, UK)) then loaded. Gels were run for 1 hour (hr) at 120 V, then visualised using a manual gel documentation system (InGenius, Syngene, UK).

2.2.5. Gel and PCR Purification

DNA amplified in PCR reactions alone, or following electrophoresis on an agarose gel, is required to be purified from contaminants. Excised gel slices and PCR reactions were cleaned up using the Wizard SV Gel and PCR Clean Up Kit (Promega, UK) according to the manufacturer's instructions. Briefly, the weight of the gel slice was determined and Membrane Binding Solution was added at a ratio of 1 µL/1 mg. The mixture was vortexed then incubated for 10 minutes at 60 °C to allow the gel to dissolve. The gel mix was transferred to a SV Minicolumn in a 2 mL collection tube and incubated for 1 minute at room temperature. For a PCR product, an equal volume of Membrane Binding Solution was added to the product and mixed. This was transferred to a SV Minicolumn in a 2 mL collection tube. For both templates, the method was then the same. The SV Minicolumn (in a 2 mL collection tube) was centrifuged for 1 minute at 16000 x g and the flow-through discarded. To the SV Minicolumn, 700 µL of Membrane Wash Solution was added and centrifuged for 1 minute at 16000 x g. The flow-through was discarded and 500 µL of Membrane Wash Solution was added to the SV Minicolumn. This was centrifuged for 5 minutes

at 16000 x g and the flow-through was discarded. The SV Minicolumn within a 2 mL collection tube was centrifuged for 1 minute at 16000 x g, with the lid off. The SV Minicolumn was carefully transferred to a clean 1.5 mL microcentrifuge tube and 30 µL NFW was added directly to the column membrane. This was incubated for 1 minute on the benchtop then centrifuged for 1 minute at 16000 x g. The purified DNA was stored at -20 °C. All volumes described were for the purification of one sample.

2.2.6. Cloning of PCR products into pGEM-T vector

Cloning is used to create multiple identical copies of DNA fragments. pGEM-T is a linearised vector suitable for such purpose. Purified PCR products were ligated into pGEM-T (Promega, UK) at an insert:vector molar ratio of 3:1 as recommended by the manufacturer. The volume of DNA to achieve this ligation ratio is calculated using the formula:

$$\frac{\text{ng pGEM-T vector} \times \text{kb size of insert}}{\text{kb size of vector}} \times \text{insert:vector molar ratio} = \text{ng of insert}$$

Briefly, to a microfuge tube, 5 µL 2X Rapid Ligation Buffer, 1 µL pGEM-T (50 ng) vector and 1 µL T4 DNA Ligase was added (all from Promega, UK). The PCR product

was then added and the volume made up to 10 µL with NFW. The contents were mixed and incubated overnight at 4 °C.

2.2.7. Transformation

Transformation is the direct uptake of foreign DNA into a bacterial cell. As a result, the DNA will be propagated. Four microliters of the plasmid DNA to be transformed was mixed with 20 µL of competent cells and incubated for 30 minutes on ice. The cells were heat shocked for 20 seconds at 42 °C in a waterbath then returned to ice for 2 minutes after which 0.5 mL sterile SOC media (Sigma Aldrich, UK) was added. The cells were then incubated at 37 °C with shaking at 225 rpm for 2 hours. Following this, cells were pelleted for 30 seconds at 12500 x g and 380 µL of the supernatant removed. The pellet was resuspended in the remaining supernatant and 10% of the volume was spread onto an LB agar plate containing no antibiotic (control). The remaining volume was spread onto an LB agar plate containing the appropriate antibiotic and the plates were incubated overnight at 37 °C.

2.2.8. Plasmid DNA mini-prep

Plasmid DNA is purified from the bacterial cell following its' transformation, by a mini-prep procedure. Single positive colonies were picked and inoculated into 5 mL LB broth (Sigma Aldrich, UK) plus antibiotic and incubated overnight at 225 rpm, 37 °C. The following morning, 0.5 mL of culture was transferred to a cryovial, an equal volume of 50% glycerol (Sigma Aldrich, UK) was added and this glycerol stock stored

at -80 °C. Plasmid was purified from the remaining volume of culture using the Zyppy Plasmid Mini Prep Kit (Cambridge Bioscience, UK). Briefly, the culture was centrifuged for 30 seconds at maximum speed until the whole volume was pelleted in a 2 mL tube. The pellet was resuspended in 600 µL PBS and 100 µL 7x Lysis Buffer was added and mixed by inverting the tube. Following this, 350 µL cold Neutralization Buffer (supplemented with RNase A) was added and mixed by inverting the tube. Precipitated material was removed by centrifugation for 2 minutes at 16000 x g and the supernatant transferred to a Zymo-Spin IIN column in a 2 mL collection tube. This was centrifuged for 15 seconds at maximum speed and the flow-through discarded. Endo-Wash Buffer (200 µL) was added to the Zymo-Spin IIN column within a 2 mL collection tube and centrifuged for 15 seconds at maximum speed. Zyppy Wash Buffer (400 µL) was added to the Zymo-Spin IIN column and centrifuged for 30 seconds at maximum speed. The Zymo-Spin IIN column was carefully transferred to a clean 1.5 mL microcentrifuge tube and 30 µL NFW was added directly to the membrane of the column. This was incubated for 1 minute on the bench top followed by centrifugation for 15 seconds at 12000 x g. The concentration of the purified plasmid was determined using a NanoDrop 2000c analyser (Thermo Scientific, UK). Eluted DNA was stored at -20 °C.

2.2.9. Plasmid DNA midi-prep

A midi-prep procedure is used to purify out large quantities of DNA from transformed bacterial cells. Midi-prep plasmid purifications were performed using the Hi-Speed plasmid Midi kit (Qiagen, UK). The required glycerol stock was

removed from -80 °C and thawed on wet ice. From the cryovial, 50 µL was removed and inoculated into 5 mL LB broth plus antibiotic and incubated for 8 hours at 225 rpm, 37 °C. After 8 hours, 500 µL of the first culture was inoculated into 50 mL LB broth plus antibiotics and incubated overnight at 225 rpm, 37 °C. The next day, the cells were harvested by centrifugation for 15 minutes at 6000 x g, 4 °C. The supernatant was discarded and the pellet resuspended in 6 mL of Buffer P1 containing RNase A. When completely resuspended, 6 mL of Buffer P2 was added and the volume mixed by inverting the tube. The contents of the tube were incubated for 5 minutes at RT after which 6 mL of pre-chilled Buffer P3 was added and the lysate mixed by inverting. The lysate was poured into a QIAfilter cartridge with a cap screwed onto the outlet port and was incubated for ten minutes at RT. The cap was removed and, using the plunger, the lysate pushed into a HiSpeed Midi tip (pre-wetted with 4 mL Buffer QBT). The lysate was allowed to drain by gravity flow and 20 mL of Buffer QC was added and allowed to drain. DNA was eluted with 5 mL Buffer QF into a universal tube and 3.5 mL isopropanol was added. This was incubated at RT for 5 minutes to precipitate the DNA. A QIAprecipitator Midi Module was attached to a 20 mL syringe from which the plunger had been removed and the isopropanol mixture poured into the syringe. The plunger was inserted and the liquid pushed through the cartridge. The QIAprecipitator was taken off the syringe and the plunger was removed. The module was placed on the syringe barrel and 2 mL 70% ethanol was pushed through by re-inserting the plunger into the syringe. The QIAprecipitator was removed from the syringe and the plunger removed. The QIAprecipitator and the syringe were re-assembled and air was pushed through at a constant pressure. The module was blotted onto absorbent

tissue to ensure no carry-over of ethanol. A 5 mL syringe (no plunger) was attached to the QIAprecipitator and 600 µL NFW added to the syringe. The plunger was added to push this through into a 1.5 mL microcentrifuge tube. The QIAprecipitator was removed from the 5 mL syringe and the plunger removed. The previous step was repeated to ensure maximum recovery of the DNA. The concentration of the eluted DNA was determined by NanoDrop 2000c analyser and stored at -20 °C.

2.2.10. DNA sequencing

DNA sequencing is the elucidation of the nucleotide sequence of a specific DNA molecule. Cloned PCR products were submitted for Sanger sequencing by Source Bioscience Ltd (Nottingham, UK). PCR products cloned into the pGEM-T vector were sequenced using standard M13 forward and reverse sequencing primers. Whereas, clones generated in pcDNA3.1His vector were sequenced using standard T7 forward sequencing primer and BGH reverse sequencing primer. All sequencing primers were supplied by Source Bioscience Ltd. The DNA sequence that was supplied by the facility was analysed using Chromas Lite software and compared to reference sequences in the NCBI database.

2.3. Protein Analysis

2.3.1. Cell lysis and protein determination

Cell walls are disrupted to release the total protein of the cell and generate a whole cell lysate. The concentration of the total protein of the lysate is then determined.

Cell pellets were lysed in 200 µL clear lysis buffer (1% SDS (Sigma Aldrich, UK), 50 mM Tris (Fisher Scientific, UK) (pH 6.8), 5 mM EDTA (Sigma Aldrich, UK) and 10% glycerol). Lysates were sonicated on ice until clear and clarified at 10000 x g, 4 °C for 10 minutes. The protein concentration of the lysates was determined using the DC Protein Determination Kit (Bio-Rad). Protein standards were prepared by dissolving BSA (Sigma Aldrich, UK) in lysis buffer and diluting to 1.5, 1.0, 0.5 and 0.2 mg/mL. Each sample or standard was assayed in triplicate. Briefly, 5 µL of each sample (cell lysate or protein standard) was placed in one well of a 96 well plate. To each well, 25 µL of Reagent A' (prepared by adding 20 µL of reagent S to 1 mL of Reagent A) was added. Following this, 200 µL of Reagent B was then added and the plate gently agitated to ensure the volumes were mixed. The plate was incubated for 15 minutes and the absorbance read using the FLx800 Plate Reader (BioTek).

2.3.2. Western Blotting

Western blotting is a technique used to determine the presence or absence of a specific protein in a sample. Following separation and transfer of the proteins in the chosen sample, an antibody is used to detect the target. Unless otherwise stated, a 10% resolving gel (7.05 mL water, 4.25 mL resolving buffer (Geneflow, UK), 5.7 mL acrylamide (Sigma Aldrich, UK), 15 µL TEMED (Sigma Aldrich, UK) and 50 µL APS (Sigma Aldrich, UK)) and a 4% stacking gel (4.6 mL water, 1.9 mL stacking gel (Geneflow, UK), 1.0 mL acrylamide (Sigma Aldrich, UK), 15 µL TEMED (Sigma Aldrich, UK) and 50 µL APS (Sigma Aldrich, UK)) was prepared. To the cell lysates was added an appropriate volume of 4x SDS loading buffer (8% w/v SDS, 40% v/v glycerol, 20%

v/v β-mercaptoethanol, 0.008% w/v bromophenol blue (Sigma Aldrich, UK) and 250 mM Tris (pH 6.8) and the samples heated for 5 minutes at 95 °C. Samples were loaded onto the gel and run for approximately 1 hour at 30 mA per gel in 1x Running buffer (25 mM Tris, 190 mM glycine (Fisher Scientific, UK), 0.1% SDS). Separated proteins were electroblotted to Polyvinylidene fluoride (PVDF) membrane (pre-soaked in methanol (Sigma Aldrich, UK)) for 1 hour at 75 V in 1x transfer buffer (25 mM Tris, 190 mM glycine, Geneflow, UK). Membranes were blocked in 3% blocking solution (3% milk powder in 1x PBS-Tween (Sigma Aldrich, UK) (0.05%)) for 1 hour at RT with agitation. After blocking, the membrane was incubated in primary antibody (diluted according to the manufacturer's instructions in 3% blocking solution) overnight at 4 °C with agitation. The membrane was washed 3 x 5 minutes in 1x PBS-Tween (0.05%) with agitation then incubated for 1 hour with agitation with an appropriate HRP-conjugated secondary antibody which was diluted 1:5000 in 3% blocking solution. Following incubation, the membrane was washed again as described and then probed with SuperSignal West Femto Chemiluminescent Substrate (Thermo Scientific, UK) and visualised using the Luminescent Image Analyzer LAS-1000 (FujiFilm).

Chapter 3 : Characterisation of Variable LEDGF Isoform Expression in Primary CLL Cells

3.1. Introduction

CLL remains incurable with current therapies and so there remains a need to develop novel treatment strategies. It is suggested that antigens detected in patients may identify tumour markers that could be used to develop immunotherapeutic strategies. One method used to identify such antigens is 'serological identification by recombinant expression cloning' (SEREX). Briefly, cDNA expression libraries are constructed from tumour tissue, transformed into E *coli* and antigens expressed are detected using autologous sera. Positive clones that are reactive with the sera are isolated, purified and the DNA analysed to determine the sequence. The sequence was used to interrogate databanks to determine the identity of the potential antigen. Using this technique, Krackhardt *et al.* identified 14 antigens in CLL patients, 6 of which were found to be expressed exclusively in CLL by subsequent RT-qPCR analysis and 5 where present in CLL and other lymphomas[59]. One of these was found to have been previously described as LEDGF. As antibodies to this antigen have been detected in patients with other lymphomas[59] it was suggested that this might be a tumour associated antigen rather than one that is specific to CLL.

Indeed, previous studies have identified the over expression of LEDGF in a number of different tumour samples and cell lines. Datasets presented in the Oncomine Database (www.oncomine.org) indicate that LEDGF is overexpressed, compared to normal cells/tissues in as many as 12 different cancers including breast and cervical cancer and CLL (14 datasets deposited). More specifically, in 2005, a study by Daniels *et al*. showed that 18.4% of prostate cancer patients had autoantibodies to the p75 isoform of LEDGF (LEDGF/p75)[80]. This compared to only 5.5% of normal matched controls. Immunohistochemistry revealed that over expression of LEDGF/p75 was detected in 93% of prostate tumours, but not in normal prostate tissue.

In a second study, Daugaard *et al.* demonstrated that LEDGF/p75 mRNA levels were higher in primary and metastatic carcinoma compared with normal breast tissue[81]. They also present microarray data showing over expression of LEDGF/p75 in bladder carcinomas that had invaded muscle tissue as compared to the normal tissue.

Finally, a more recent study by Basu *et al.* examined LEDGF/p75 mRNA expression in 21 different human cancer tissues and compared this to the normal tissue counterpart using the Oncomine database[82]. All but bladder and uterine cancer showed a significant over expression of LEDGF/p75 compared to normal tissue counterparts. Further, the authors validated these results and used a qPCR array to demonstrate significant over expression of the LEDGF/p75 transcript in prostate, thyroid, breast and colon cancers. Using immunohistochemical methods, they were

able to demonstrate over expression of LEDGF protein in prostate, liver and thyroid tumours.

Whilst the over expression of LEDGF has been reported in a number of tumour types, there is also increasing evidence that translocations between LEDGF and NUP98 (t(9;11)) may facilitate development of leukemia. The first reported case was in a patient with acute biphenotypic leukemia[50], and it has since been reported that this t(9;11) translocation is a recurrent feature in patients with AML [51, 52, 55]. It has also been reported in a paediatric AML case[54], a patient with transformed CML[53] and in a patient with MDS[56].

Several isoforms of LEDGF have also been described in AML[57]. Of the isoforms identified, LEDGF/p75, p52 and p52b all share a common N-terminal region up to exon 9, but differ within the C terminal region. Three further isoforms were described that had varying deletions between exons 5 and 8 (Figure 3-1)[57]. Isoform LEDGF/p52bΔE6 had the complete exon 6 deleted which resulted in an altered reading frame within exon 7. Consequently, a premature stop codon is encoded which could result in the generation of a truncated protein of 167 amino acids.

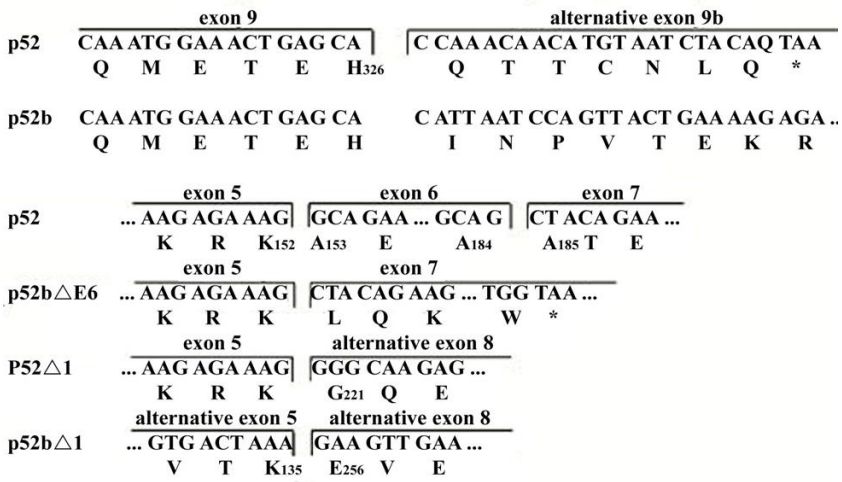


Figure 3-1: Alternative Isoforms of LEDGF described in AML.

In AML, four extra isoforms of LEDGF have been described, namely LEDGF/p52b, LEDGF/p52bΔE6, LEDGFp52Δ1 and LEDGF/p52bΔb1. The sequence of LEDGF/p52 is provided as a comparison. LEDGF/p52b differs from LEDGF/p52 due to its' unique exon 9b whilst the other three isoforms have varying deletions.

The aims of this chapter, therefore, were:

1. Determine which of the different reported isoforms of LEDGF are expressed in CLL cells by using RT-PCR.
2. Determine the relative expression levels of the different isoforms identified in aim 1.
3. Determine whether there are any significant differences in the expression levels of the different isoforms between CLL and normal B cells
4. Identify a suitable antibody in order to examine the protein expression of the different LEDGF isoforms in CLL cells
5. Correlate LEDGF isoform expression with clinical and prognostic factors

3.2. Methods

3.2.1. Sample Cohort

CLL cDNA samples were a kind gift from Dr. G. Johnson. Normal B cell cDNA samples were a kind gift from Dr. A. Duckworth. These were stored at -20 °C until ready for use.

The average age of the patients within the cohort was 62.7 years old and all patients had a white blood cell count $<5x10^9$ /L. The cohort included a variety of patients that covered the molecular abnormalities associated with CLL including deletions of chromosomes 13q, 11q and trisomy 12. A complete clinical data set for each patient is given in Appendix A.

3.2.2. RT-PCR reaction optimisation for LEDGF/p75 and p52

Determination of whether CLL cells expressed either of the 2 major isoforms of LEDGF was carried out by end point PCR using 2 sets of primers (Table 3-1). One pair was specific to the LEDGF/p75 isoform (LEDGFuniF and LEDGF/p75-R; amplicon size = 233 bp) whilst the second was specific to the LEDGF/p52 isoform (LEDGFuniF and LEDGF/p52-R; amplicon size = 198 bp). The primer pairs utilised a common forward primer that annealed in the shared N-terminal region whilst the reverse primer for each pair annealed to the respective unique C-terminal region.

The location of the PCR primers with respect to the sequences for LEDGF/p75 and LEDGF/p52 are shown in Figure 3-2. The forward primer (LEDGFuniF) common to each isoform is highlighted in yellow. The reverse primer specific for LEDGF/p75 is

highlighted in red and the corresponding primer for LEDGF/p52 is highlighted in green. The unique C-terminus tail of each isoform is underlined.

LEDGF isoform	Primer	Sequence (5' > 3')
p75	LEDGFuni-F	gttacttcaacctccgattc
	LEDGFp75-R	gtttctcgcttcttctccac
p52	LEDGFuni-F	gttacttcaacctccgattc
	LEDGFp52-R	gtagattacatgttgtttggtg

Table 3-1: Primer sequences for the PCR primers used for the screening of CLL samples.

The primers pairs are specific for either LEDGF/p75 (top) or LEDGF/p52 (bottom) and are for use in RT-PCR.

LEDGF/p75 (**NM_001128217)**

ATGACTCGCGATTTCAAACCTGGAGACCTCATCTTCGCCAAGATGAAAGGTTATCCC
CATTGGCCAGCTCGAGTAGACGAAGTTCCTGATGGAGCTGTAAAGCCACCCACAAAC
AAACTACCCATTTTCTTTTTTGGAACTCATGAGACTGCTTTTTTAGGACCAAAGGAT
ATATTTCCTTACTCAGAAAATAAGGAAAAGTATGGCAAACCAAATAAAAGAAAAGGT
TTTAATGAAGGTTTATGGGAGATAGATAACAATCCAAAAGTGAAATTTTCAAGTCAA
CAGGCAGCAACTAAACAATCAAATGCATCATCTGATGTTGAAGTTGAAGAAAAGGAA
ACTAGTGTTTCAAAGGAAGATACCGACCATGAAGAAAAAGCCAGCAATGAGGATGTG
ACTAAAGCAGTTGACATAACTACTCCAAAAGCTGCCAGAAGGGGGAGAAAGAGAAAG
GCAGAAAAACAAGTAGAAACTGAGGAGGCAGGAGTAGTGACAACAGCAACAGCATCT
GTTAATCTAAAAGTGAGTCCTAAAAGAGGACGACCTGCAGCTACAGAAGTCAAGATT
CCAAAACCAAGAGGCAGACCCAAAATGGTAAAACAGCCCTGTCCTTCAGAGAGTGAC
ATCATTACTGAAGAGGACAAAAGTAAGAAAAAGGGGCAAGAGGAAAAACAACCTAAA
AAGCAGCCTAAGAAGGATGAAGAGGGCCAGAAGGAAGAAGATAAGCCAAGAAAAGAG
CCGGATAAAAAAGAGGGGAAGAAAGAAGTTGAATCAAAAAGGAAAAATTTAGCTAAA
ACAGGGGTTACTTCAACCTCCGATTCTGAAGAAGAAGGAGATGATCAAGAAGGTGAA
AAGAAGAGAAAAGGTGGGAGGAACTTTCAGACTGCTCACAGAAGGAATATGCTGAAA
GGCCAACATGAGAAAGAAGCAGCAGATCGAAAACGCAAGCAAGAGGAACAAATGGAA
ACTGAGCAGCAGAATAAAGATGAAGGAAAGAAGCCAGAAGTTAAGAAAGTGGAGAAG
AAGCGAGAAACATCAATGGATTCTCGACTTCAAAGGATACATGCTGAGATTAAAAAT
TCACTCAAAATTGATAATCTTGATGTGAACAGATGCATTGAGGCCTTGGATGAACTT
GCTTCACTTCAGGTCACAATGCAACAAGCTCAGAAACACACAGAGATGATTACTACA
CTGAAAAAAATACGGCGATTCAAAGTTAGTCAGGTAATCATGGAAAAGTCTACAATG
TTGTATAACAAGTTTAAGAACATGTTCTTGGTTGGTGAAGGAGATTCCGTGATCACC
CAAGTGCTGAATAAATCTCTTGCTGAACAAAGACAGCATGAGGAAGCGAATAAAACC
AAAGATCAAGGGAAGAAAGGGCCAAACAAAAAGCTAGAGAAGGAACAAACAGGGTCA
AAGACTCTAAATGGAGGATCTGATGCTCAAGATGGTAATCAGCCACAACATAACGGG
GAGAGCAATGAAGACAGCAAAGACAACCATGAAGCCAGCACGAAGAAAAAGCCATCC
AGTGAAGAGAGAGAGACTGAAATATCTCTGAAGGATTCTACACTAGATAACTAG

LEDGF/p52 (**NM_021144)**

ATGACTCGCGATTTCAAACCTGGAGACCTCATCTTCGCCAAGATGAAAGGTTATCCC
CATTGGCCAGCTCGAGTAGACGAAGTTCCTGATGGAGCTGTAAAGCCACCCACAAAC
AAACTACCCATTTTCTTTTTTGGAACTCATGAGACTGCTTTTTTAGGACCAAAGGAT
ATATTTCCTTACTCAGAAAATAAGGAAAAGTATGGCAAACCAAATAAAAGAAAAGGT
TTTAATGAAGGTTTATGGGAGATAGATAACAATCCAAAAGTGAAATTTTCAAGTCAA
CAGGCAGCAACTAAACAATCAAATGCATCATCTGATGTTGAAGTTGAAGAAAAGGAA
ACTAGTGTTTCAAAGGAAGATACCGACCATGAAGAAAAAGCCAGCAATGAGGATGTG
ACTAAAGCAGTTGACATAACTACTCCAAAAGCTGCCAGAAGGGGGAGAAAGAGAAAG
GCAGAAAAACAAGTAGAAACTGAGGAGGCAGGAGTAGTGACAACAGCAACAGCATCT
GTTAATCTAAAAGTGAGTCCTAAAAGAGGACGACCTGCAGCTACAGAAGTCAAGATT
CCAAAACCAAGAGGCAGACCCAAAATGGTAAAACAGCCCTGTCCTTCAGAGAGTGAC
ATCATTACTGAAGAGGACAAAAGTAAGAAAAAGGGGCAAGAGGAAAAACAACCTAAA
AAGCAGCCTAAGAAGGATGAAGAGGGCCAGAAGGAAGAAGATAAGCCAAGAAAAGAG
CCGGATAAAAAAGAGGGGAAGAAAGAAGTTGAATCAAAAAGGAAAAATTTAGCTAAA
ACAGGGGTTACTTCAACCTCCGATTCTGAAGAAGAAGGAGATGATCAAGAAGGTGAA

AAGAAGAGAAAAGGTGGGAGGAACTTTCAGACTGCTCACAGAAGGAATATGCTGAAA
GGCCAACATGAGAAAGAAGCAGCAGATCGAAAACGCAAGCAAGAGGAACAAATGGAA
ACTGAGCACCAAACAACATGTAATCTACAGTAA

Figure 3-2: cDNA sequence of LEDGF/p75 (above) and LEDGF/p52 (below).

Sequences highlighted indicate the primers used in the PCR reactions whilst the underlined sequence identifies the unique C-terminal sequences in each of the isoforms. The common forward primer is highlighted in yellow, the reverse primer specific to LEDGF/p75 is highlighted in red and the reverse primer specific to LEDGF/p52 is highlighted in green. Each different exon is highlighted by a colour change from black to blue and vice versa.

Reaction parameters were optimised as described in Chapter 2.2.3 using the following annealing temperature gradient: - LEDGF/p75 52 - 64 °C; LEDGF/p52 51 - 63 °C. For all of these optimisations, cDNA from KCL22 and K562 cell lines was used as template. A negative control in the form of NFW was also included. Samples of each PCR were analysed by electrophoresis on 1% agarose gels as described in Chapter 2.2.4 and the chosen annealing temperature was determined by visualisation of the gel.

3.2.3. CLL Sample screening

CLL cDNA samples were initially screened for the presence or absence of the 2 major isoforms of LEDGF, namely LEDGF/p75 and p52 by RT-PCR as described in Chapter 2.2.3 using the optimised reaction parameters described above. For each

reaction, a negative control containing NFW and a positive control specific for that primer pair (commercially purchased cDNA clones) were included. The PCR reactions for each of the isoforms were performed separately but analysed on the same agarose gel so that the presence or absence of each isoform in each sample could be compared. PCR products with aberrant molecular weights that were generated during the screening procedure were purified from agarose gels as described in Chapter 2.2.5. These were then cloned into the pGEM-T vector, transformed into in JM109 competent cells (Promega, UK) and plasmid DNA purified using the mini-prep procedures outlined in Chapters 2.2.6, 2.2.7 and 2.2.8. Cloned inserts were sequenced as described in Chapter 2.2.10.

PCR primers that would allow specific amplification of LEDGF/p52b were designed to examine expression of this isoform in CLL (Table 3-2). The forward primer was designed in exon 6 of LEDGF as this was common to LEDGF/p75, p52 and p52b whilst the reverse primer was designed across the exon boundary of exon 9 and 10, as exon 10 is unique to LEDGF/p52b.

Primer	Sequence (5' > 3')
P52bscreeningF	gacctgcagctacagaag
P52bscreeningR	cagtaactggattaatgtgc

Table 3-2: Primer sequences for PCR primers used to screen CLL cDNA cells for the presence of LEDGF/p52b.

The primer pair is for use in RT-PCR.

The primer pair was optimised as described in Chapter 2.2.3 using a temperature gradient of 52 – 64 °C and KCL22 and K652 cell line cDNA as the template. PCR reactions were analysed by agarose gel electrophoresis. Bands were cut out of the gel, purified, cloned and sequences confirmed as described in Chapters 2.2.5- 2.2.8 and 2.2.10. These conditions were then used to screen the cohort of CLL samples for expression of this isoform. PCR reactions were analysed by agarose gel electrophoresis and gels visualised to determine if the isoform was present.

PCR primers (Table 3-3) were synthesised to screen the CLL cDNA samples for the other isoforms described by Huang *et al.* (see Figure 3-1)[57]. The forward primer (p52isoF) was designed such that it spanned the boundary of exon 4 and 5 whilst the reverse primer (p52isoR) was designed within exon 8.

Primer	Sequence (5' > 3')
P52isoF	gccagcaatgaggatgtgac
P52isoR	gaatcggaggttgaagtaaccc

Table 3-3: Primer sequences for the PCR primers used to screen CLL cDNA samples for the presence of LEDGF isoforms as described in Huang *et al*.[57].

The primer pair is designed in the regions common to all of the potential isoforms and therefore can only be used for RT-PCR analysis.

The primer pair was optimised as described in Chapter 2.2.3. A gradient of 57 – 67 °C was used with KCL22 and K652 cell line cDNA as the template. Commercially purchased clones of LEDGF/p75 and LEDGF/p52 were used as positive controls for the PCR amplification. This primer pair would amplify p75, p52 and p52b but would not discriminate between the 3 isoforms. The expression of other isoforms of LEDGF (namely p52bΔE6, p52Δ1 or p52bΔ1) as described by Huang *et al*.[57] would be indicated by multiple bands when the PCR products were examined by agarose gel electrophoresis.

PCR products from 5 CLL cases were analysed by 2% agarose gel electrophoresis and bands cut out of the gel and purified using the Wizard SV Gel and PCR Clean Up Kit as described in Chapter 2.2.5. Two samples were chosen and their purified products ligated into pGEM-T vector at a ratio of 3:1 (insert:vector) and transformed into JM109 competent cells by following the method in Chapters 2.2.6 and 2.2.7. Plasmid DNA was purified from selected clones using the mini-prep procedure described in Chapter 2.2.8. The sequences of the clones were verified externally at Source Bioscience Ltd (Chapter 2.2.10).

3.2.4. Quantitative RT-PCR analysis of LEDGF isoform expression in CLL

Quantitative RT-PCR (RT-qPCR) was performed using the HOT FIREPol EvaGreen qPCR Mix (Solis BioDyne) in 96 well PCR plates sealed with optical strip caps (Web Scientific, UK). Amplification reactions were carried out on a MX3005P qPCR machine and analysed using MxPro v4.1 software (Agilent Technologies, UK).

Four sets of primers, each designed to identify a specific LEDGF isoform, were used to investigate the relative levels of each of the isoforms in a cohort of CLL cDNA samples (including normal B cell cDNA) (Table 3-4). The primer pair specific for the LEDGF/p52b isoform was the same primer pair used for the end-point PCR screening already discussed (Table 3-2). Where possible, primers were designed such that each primer pair had similar annealing temperatures. The relative expression of each of the isoforms was determined by comparison to expression of a reference gene, ribosomal protein L27 (RPL27), using the ΔΔCt method.

Individual reactions required 5 µL of 5x HOT FIREPol EvaGreen qPCR Mix, 1 µL of each forward and reverse primer, (stock concentration 5 µM, final concentration of 0.2 µM each) and 16 µL of NFW. Each reaction was performed in triplicate. Thus, sufficient mastermix was always prepared for the required number of reactions for each sample. From this mastermix, 23 µL was aliquoted into the required number of wells on the plate and 2 µL of cDNA template (stock diluted 1:4 with NFW) was added to give a final volume of 25 µL. A negative control using 2 µL of NFW, instead of sample, was always prepared. Following completion of the reaction, the average Ct value for each sample was determined.

Target	Primer name	Primer sequence
LEDGF/p75	p75qpcrF p75qpcrR	ctgagcagcagaataaagatg gtgacctgaagtgaagcaag
LEDGF/p52	p52qpcrF p52qpcrR	agaagagaaaaggtgggagg gattacatgttgtttggtgctc
LEDGF/p52b	p52bscreeningF p52bscreeningR	gacctgcagctacagaag cagtaactggattaatgtgc
LEDGF/p52bΔE6	p52be6qpcrF p52be6qpcrR	gaaagagaaagctacagaag cttcatccttcttaggctg
RPL27	RPL27F RPL27R	caagttcatgaaacctgggaa gcagtttctggaagaaccactt

Table 3-4: Sequences of primers used for RT-qPCR analysis.

Each primer pair is specific for one target (left hand column), namely one of the four isoforms of LEDGF that has been shown to be present in CLL cases by RT-PCR and RPL27, the reference gene used for the expression analysis.

Cycling parameters were as follows:

95 °C 10 minutes (x1)

94 °C 10 seconds, annealing temperature for 15 seconds, 72 °C for 30 seconds, 80 °C for 7 seconds (fluorescence measurement) (x40)

To confirm the specificity of the primers and that only a single product was amplified, a sample of each PCR product from the reaction was analysed by electrophoresis on a 2% agarose gel (Chapter 2.2.4).

To determine if there was a correlation between the relative levels of LEDGF and any clinical prognostic factors, statistical analysis was performed by Dr. Fotis Polydoros, The Liverpool Clinical Trials Unit. One-Way Analysis of Variance (ANOVA) tests were performed using the STATA version 13 statistical software package. Data was presented in the form of boxplots.

3.2.5. Analysis of protein levels of LEDGF isoforms.

To investigate the levels of protein of each of the LEDGF isoforms present in CLL samples, an antibody to the N-terminal region common to all of the isoforms was required. To test the specificity of the antibody, a 10% resolving and 4% stacking gel was prepared. To the gel, 10, 20, 30 and 40 µg of 293T and Mec-1 (a CLL derived cell line) cell lysates where added and run as described in Chapter 2.3.1 and 2.3.2. The blots where probed with primary anti-PSIP1 antibody N-terminal region (Aviva, Rabbit polyclonal, 1.25:1000).

3.3. Results

3.3.1. Optimisation of RT-PCR conditions

Using cDNA generated from the cell lines K562 or KCL22, the annealing temperature for the primer pairs LEDGFuniF plus LEDGF/p75-R and LEDGFuniF plus LEDGF/p52R

was determined to be 57 °C and 54 °C for p75 and p52 respectively (Figure 3-3). The lack of any bands in the NFW and no enzyme control lanes confirm that there was no contamination of the PCR reaction.

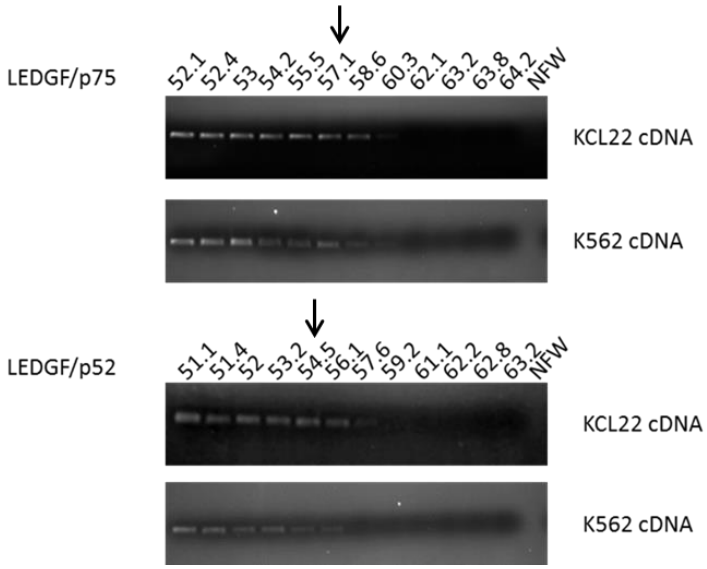


Figure 3-3: Optimisation of RT-PCR annealing temperatures.

RT-PCR reactions were performed using cDNA from either KCL22 or K562 cells with a range of annealing temperatures and a primer pair specific for LEDGF/p75 or LEDGF/p52. Reactions were analysed by agarose gel electrophoresis and the annealing temperature for the primer pairs determined visually (indicated by arrows). Specifically, the chosen annealing temperature for the primer pair specific to LEDGF/p75 was 57 °C (top) and for LEDGF/p52 was 54 °C (bottom).

3.3.2. Screening CLL cases for LEDGF/p75 or p52 isoform expression

Using the reaction conditions chosen above, cDNA samples from a cohort of 78 CLL cases (pre-screened for the presence of the reference gene to ensure integrity of

the template) were screened by PCR to determine whether or not they expressed either the p75 and/or p52 isoform of LEDGF. Each of the PCR reactions was performed separately and included positive controls in the form of commercially purchased cDNA clones for each of the isoforms. Negative controls in the form of NFW in place of cDNA were also included. Upon completion of the reaction, a sample of each PCR was analysed by agarose gel electrophoresis. In order to aid the analysis, PCR products for each isoform from the same sample were analysed in adjacent lanes. Figure 3-4 shows a representative gel for 6 of the CLL cases examined. Results of the entire cohort are summarised in Table 3-5.

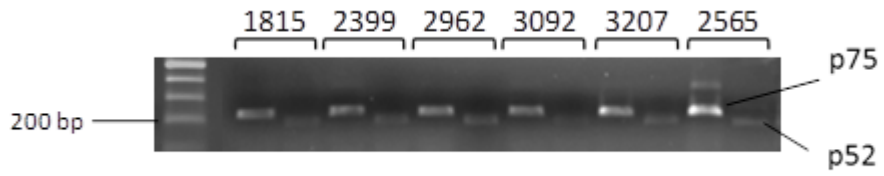


Figure 3-4: RT-PCR screening of CLL cases for expression of LEDGF isoforms p75 and p52.

cDNA from 78 CLL cases were screened by RT-PCR for expression of the p75 or p52 isoforms of LEDGF using primer pairs specific to each isoform at the chosen annealing temperatures described above. PCR products were analysed by agarose gel electrophoresis. This shows a representative gel for 6 of the cases (1815, 2399, 2962, 3092, 3207 and 2565). The bands representing LEDGF/p75 and LEDGF/p52 are highlighted.

The results of this screening show that 68 of the 78 cases examined expressed the p75 isoform, whilst 67 expressed the p52 isoform. With the exception of 2 cases (2499 and 3071) those that did not express the p75 isoform were not the same as those that did not express the p52 variant and vice versa. In some of those cases that did express LEDGF/p75, a faint band was also observed running at approximately 500 bp (for example sample 2565, Figure 3-4). As the nature of this band is uncertain and not present in all of the samples, it was further investigated.

3.3.3. Identification of a p52b isoform of LEDGF

During the process of screening CLL cases for the expression of the p52 and p75 isoforms of LEDGF, a band was identified on the agarose gel whose size did not correspond to the expected PCR products. Therefore this band was purified, cloned into pGEM-T and sequenced. The results from the sequencing are shown in Figure 3-5.

ATTTCCAATTAGGGGGCGATTGGGCCCGACGTCGCATGCTCCCGGCCGCCATGGCCG
CGGGATTGTTACTTCAACCTCCGATTCTGAAGAAGAAGGAGATGATCAAGAAGGTGA
AAAGAAGAGAAAAGGTGGGAGGAACTTTCAGACTGCTCACAGAAGGAATATGCTGAA
AGGCCAACATGAGAAAGAAGCAGCAGATCGAAAACGCAAGCAAGAGGAACAAATGGA
AACTGAGCACATTAATCCAGTTACTGAAAAGAGAATACAAGTGGAGCAAACAAGAGA
TGAAGATCTTGATACAGACTCATTGGACTGAATTTCCCCCTTCCCCCCATTGATGGA
AGAATGTTCCAGATTCTAAATTGAGGACTTCATTATTAATGGCATTACTGTGTTATG
ATTAACAAATTTCCTGTAAGGCAGAATAAAGATGAAGGAAAGAAGCCAGAAGTTAAG
AAAGTGGAGAAGAAGCGAGAAACAATCACTAGTGCGGCCGCCTGCAGGTCGACCATA
TGGGAGAGCTCCCAACGCGTTGGATG

H I N P V T E K R I Q V E Q T R D E D L D T D S L D **Stop** I S P F P P L **Met** E E C S R F

Stop I E D F I I N G I T V L **Stop** L T N F L **Stop**

Figure 3-5: Sequence of the alternative splice form of LEDGF.

(Top) The PCR product identified during the screening procedure was cloned into pGEM-T and sequenced. The primers used to generate the PCR product (LEDGFuni-F and LEDGF/p75-R) are highlighted in yellow. The red text signifies the inserted sequence identified in this study (182 nucleotides) which results in a larger product than LEDGF/p75 when analysed by agarose gel electrophoresis. (Bottom) The translate tool of Expasy was used to determine the amino acid sequence of the identified PCR product. The inserted sequence results in a premature stop codon.

This showed that the sequence immediately downstream of the universal forward primer was the expected sequence from exon 9 of LEDGF common to both the p52 and p75 isoform of LEDGF. However, the sequence then continued for 182 nucleotides into the intron following exon 9 and then spliced back into exon 10 of LEDGF/p75 ie an alternative splice donor site within intron 9 was used. In order to determine the effects of this extra nucleotide sequence on the amino acid

sequence, the Translate tool of Expasy (http://web.expasy.org/translate/) was used. The amino acid sequence returned is also shown in Figure 3-5.

The alternative splicing results in a disruption to the reading frame of LEDGF/p75 at the carboxyl region. The first His residue shown in Figure 3-5 is a result of the last 2 nucleotides from exon 9 (bases C and A) and the first nucleotide (C) from the intron. Following this the reading frame encodes 25 amino acids prior to a stop codon.

This result suggests there is a third isoform of LEDGF that is expressed in CLL that shares the same amino terminal sequence as LEDGF/p75 and p52, but which differs at the carboxy terminus. This would result in a protein that was bigger in size than LEDGF/p52 (351 residues compared to 333). Based on the results of the sequencing this isoform was subsequently identified as LEDGF/p52b, previously described by Huang *et al.* in AML[57], and the result suggests this isoform is also expressed in CLL.

3.3.4. Screening CLL cases for LEDGF/p52b isoform expression

Having shown that the p52b isoform of LEDGF was expressed in CLL cells, PCR primers (p52bscreeningF and p52bscreeningR) specific for this isoform (Table 3-2) were synthesised and used to screen the entire cohort by RT-PCR. PCR conditions were optimised by amplifying cDNA from KCL22 and K562 cells and the annealing temperature was determined to be 57 °C (Figure 3-6). Sequencing of the cloned PCR product confirmed its specificity.

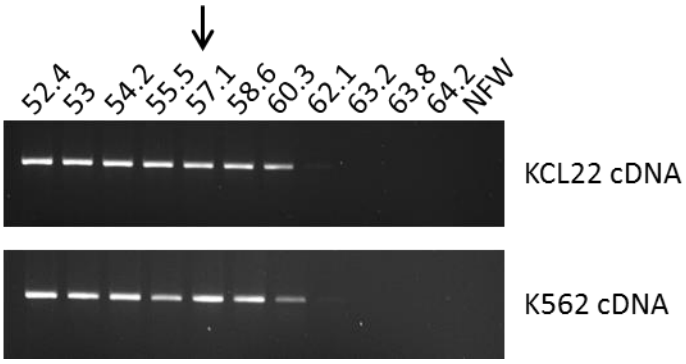


Figure 3-6: Optimisation of RT-PCR annealing temperature

RT-PCR reactions were performed with KCL22 cDNA (top) and K562 cDNA (bottom) with a range of annealing temperatures and the primer pair specific for LEDGF/p52b (p52bscreeningF and R). Reactions were analysed by agarose gel electrophoresis and the chosen annealing temperature determined visually (indicated by arrows) and shown to be 57 °C.

A representative gel from the screening of the CLL cohort using the PCR conditions optimised above is shown in Figure 3-7. The results demonstrated that this alternative splice form of LEDGF was expressed in 76 of the 78 samples analysed in the cohort (Table 3-5).

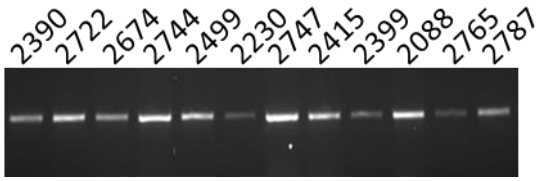


Figure 3-7: Representative image from the screening of CLL cDNA samples to examine the expression of LEDGF/p52b.

Primers p52bscreeningF and R were used in RT-PCR reactions at the chosen annealing temperature of 57 °C, to screen the CLL cDNA cohort for the presence of LEDGF/p52b. RT-PCR products were analysed by agarose gel electrophoresis and the presence or absence of the isoform determined visually. The CLL samples shown by this representative image are detailed at the top of the gel.

3.3.5. PCR screening of CLL cases for other isoforms of LEDGF

Three further isoforms of LEDGF were described by Huang *et al*.[57]. In order to screen CLL patients for evidence of expression of these isoforms, PCR primers (Table 3-3) were synthesised which would allow amplification of all three and PCR conditions optimised using cDNA from K562 or KCL22 as described earlier. As shown in Figure 3-8, the chosen annealing temperature for this primer pair was 64 °C.

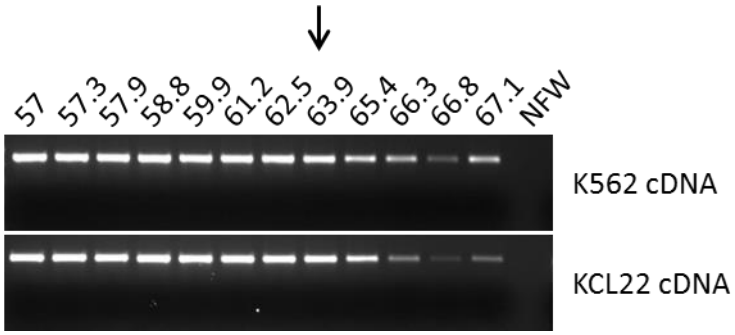


Figure 3-8: Optimisation of RT-PCR annealing temperatures.

RT-PCR reactions were performed on K562 (top) KCL22 (bottom) cell line cDNA with a range of annealing temperatures and the primer pair p52isoF and p52isoR. RT-PCR products were analysed by agarose gel electrophoresis and the annealing temperature determined visually to be 64 °C (indicated by arrows).

These chosen PCR conditions were then used to screen the 78 CLL cases for expression of these various isoforms. Figure 3-9 shows a representative example of the results obtained from the screening of the CLL cDNA samples. As the primers used cannot distinguish between LEDGF/p75, p52 and/or p52b, the major band at 443 bp could represent any of these isoforms. The band at approximately 350 bp would represent LEDGF/p52bΔE6. The lack of bands at 239 bp or 134 bp suggests that the 2 further isoforms described by Huang *et al.*[57] are not expressed in CLL.

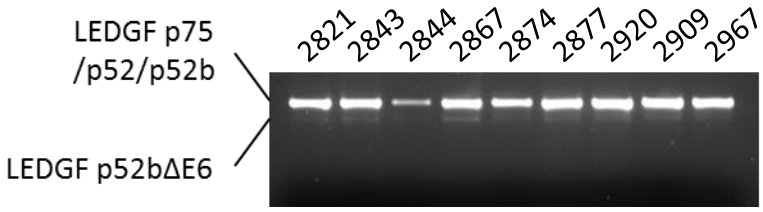


Figure 3-9: Representative image from RT-PCR screening of CLL cDNA to determine the presence of extra isoforms of LEDGF.

The primers, p52isoF and p52isoR, were used to screen CLL cDNA by RT-PCR at the chosen annealing temperature of 64 °C. The products of the reactions were analysed by agarose gel electrophoresis to determine the presence or absence of extra isoforms of LEDGF. The primer pair used did not discriminate between LEDGF/p75, p52 and p52b (top bright band). The faster migrating band represents LEDGF/p52bΔE6 and shows that this isoform I present in CLL.

In order to confirm the band at 350 bp was LEDGF/p52bΔE6, the PCR was repeated and the band was purified, cloned into pGEM-T and sequenced. The sequence returned confirmed this band was LEDGF/p52bΔE6.

Summary

As a summary of the PCR screening Table 3-5 shows the presence or absence of each isoform in the CLL samples used in this cohort.

Sample	p75	p52	p52b	p52bΔE6
2415	Y	Y	Y	N
2953	Y	N	Y	Y
2512	Y	N	Y	N
3047	Y	N	Y	Y
2902	Y	Y	Y	Y
2968	Y	Y	Y	Y
2950	Y	Y	Y	Y
2673	Y	Y	Y	Y
2979	Y	Y	Y	Y
2744	Y	N	Y	Y
2999	Y	Y	Y	Y
2248	Y	N	Y	Y
2565	Y	Y	Y	Y
2408	N	Y	Y	Y
1780	N	Y	Y	Y
2418	Y	Y	Y	Y
1800	Y	Y	Y	Y
2136	N	Y	Y	Y
2539	Y	Y	Y	Y
2533	Y	Y	Y	Y
1731	Y	Y	Y	Y
2472	Y	Y	Y	Y
1815	Y	Y	Y	N
2375	Y	N	Y	Y
2390	Y	N	Y	N
2722	Y	Y	Y	Y
2674	Y	N	Y	Y
2499	N	N	Y	Y
2230	Y	Y	Y	Y
2747	Y	Y	Y	Y
2399	Y	Y	Y	Y
2088	Y	Y	Y	N
2765	N	Y	Y	Y
2787	N	Y	Y	Y
2792	N	Y	Y	Y
2793	Y	Y	Y	N
2794	N	Y	N	Y
2809	Y	Y	N	N
2820	Y	Y	Y	N
2821	Y	Y	Y	Y
2843	Y	Y	Y	Y
2844	N	Y	Y	Y
2867	Y	Y	Y	Y
2874	Y	Y	Y	Y
2877	Y	Y	Y	Y

2909	Y	Y	Y	Y
2920	Y	Y	Y	Y
2924	Y	Y	Y	Y
2930	Y	Y	Y	Y
2937	Y	Y	Y	Y
2938	Y	Y	Y	Y
2949	Y	Y	Y	Y
2962	Y	Y	Y	Y
2967	Y	Y	Y	Y
2985	Y	Y	Y	Y
3016	Y	Y	Y	Y
3018	Y	Y	Y	Y
3035	Y	Y	Y	Y
3037	Y	Y	Y	Y
3038	Y	N	Y	Y
3068	Y	Y	Y	Y
3069	Y	Y	Y	N
3071	N	N	Y	Y
3076	Y	Y	Y	Y
3078	Y	Y	Y	Y
3089	Y	Y	Y	Y
3091	Y	Y	Y	Y
3092	Y	Y	Y	N
3139	Y	Y	Y	N
3151	Y	Y	Y	N
3144	Y	Y	Y	N
3187	Y	Y	Y	Y
3192	Y	Y	Y	Y
3160	Y	Y	Y	Y
3162	Y	Y	Y	Y
3193	Y	Y	Y	Y
3207	Y	Y	Y	Y
3210	Y	Y	Y	Y

Table 3-5: Summary of the RT-PCR screening results showing LEDGF isoform expression In CLL.

78 CLL samples were screened by RT-PCR for the expression of different isoforms of LEDGF. The sample numbers are detailed in the first column followed by separate columns specific for each isoform of LEDGF/p75, p52, p52b and p52bΔE6. Presence or absence of an isoform is indicated by “Y” or “N” respectively.

3.3.6. Quantitative RT-PCR analysis of LEDGF expression in CLL

Having identified which of the isoforms of LEDGF are expressed in CLL by RT-PCR, it was felt that it was important to determine their relative levels of expression. Therefore, RT-qPCR was performed on the same cohort of CLL samples using primer pairs that were specific to each of the identified isoforms.

3.3.7. Optimisation of RT-qPCR conditions

Using cDNA generated from HEK293 and Raji cell lines, the annealing temperature for the primer pairs specific for LEDGF/p75 and LEDGF/p52bΔE6 was determined using a standard PCR machine and shown to be 57 °C for LEDGF/p75 and 53 °C for LEDGF/p52bΔE6 (Figure 3-10). The lack of bands in the NFW control lanes confirms there was no contamination of the PCR reaction.

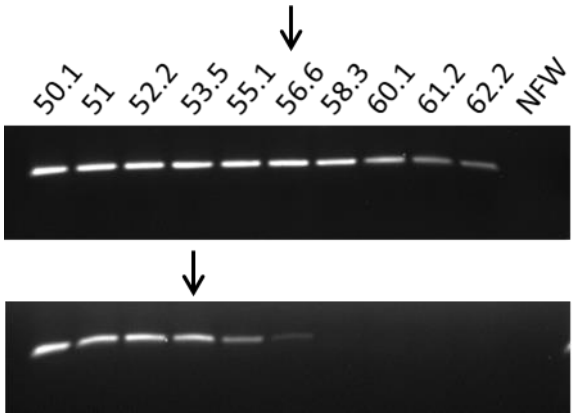


Figure 3-10: Optimisation of RT-qPCR annealing temperature

PCR reactions were performed using cDNA from either KCL22 or K562 cell lines with a range of annealing temperatures and primer pairs specific for LEDGF/p75 and LEDGF/p52bΔE6. PCR products for LEDGF/p75 (upper panel) and LEDGF/p52ΔE6 (lower panel) were analysed by agarose gel electrophoresis and the chosen annealing temperatures determined to be 57 °C and 54 °C respectively (indicated by arrows).

Using the chosen annealing temperature for the primer pair specific to LEDGF/p75, and with the knowledge that the annealing temperature for the primer pair specific for LEDGF/p52b is 57 °C (Figure 3-6), PCR reactions were performed on a qPCR machine in order to determine the optimal temperature at which to read the fluorescence signal. Figure 3-11 shows the melt curve obtained using PCR primers specific for LEDGF/p52 and p52b which were used to amplify cDNA from HEK293 and Raji cells. The cycling parameters were:

95 °C 10 minutes (x1)

94 °C 10 seconds, 57 °C 15 seconds, 72 °C for 30 seconds, 78 °C for 7 seconds (fluorescence measurement) (x40)

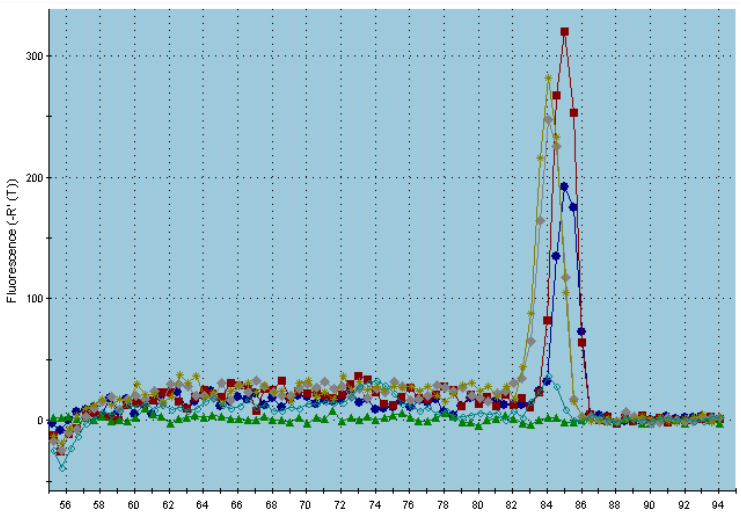


Figure 3-11: Melt curve analysis of LEDGF/p52 and p52b amplicons in cell line cDNA.

RT-qPCR reactions were performed with primers pairs specific to LEDGF/p52 and p52b and cell line cDNA as template. Following the amplification phase, the melt curve was produced and shown above. The yellow and red peaks indicate fluorescence obtained using Raji cDNA and primers specific for LEDGF/p52 or LEDGF/p52b respectively, whilst grey and blue peaks indicate fluorescence using the same primer pairs with HEK293 cDNA.

Analysis of the melt curve indicated that the read temperature could be increased to 80 °C and this was used in subsequent reactions.

3.3.8. RT-qPCR screening of CLL cases for the relative expression levels of the LEDGF isoforms and comparison to normal B cells

These optimised RT-qPCR conditions were then used to determine the expression levels of the various isoforms of LEDGF in CLL cells. Due to limited sample availability only a proportion (46) of the 78 cases already examined by RT-PCR, plus three normal B cell controls were used. The relative level of each isoform was determined by comparison to the level of expression of a reference gene, RPL27. On each separate RT-qPCR plate, a sample of HEK293 cell line cDNA was included. In addition to acting as a positive control, this allowed the normalisation of Ct values across the plates so they could be compared across the cohort. Figure 3-12 shows a representative example of the PCR products obtained from the analysis of the RT-qPCR performed on these cases.

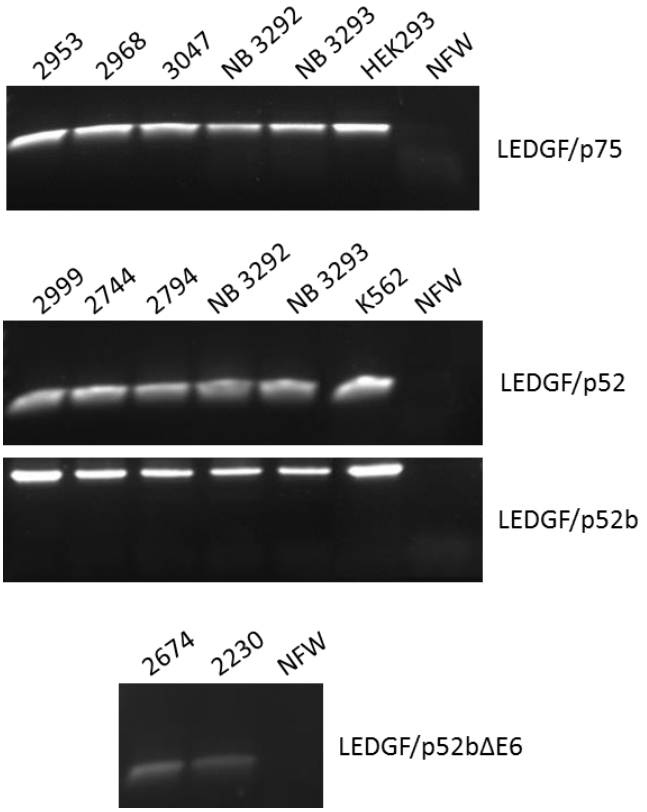


Figure 3-12: Representative image from the RT-qPCR screening of CLL cDNA samples.

CLL cDNA samples were examined for the expression of the four isoforms of LEDGF by RT-qPCR. Products were analysed by agarose gel electrophoresis to ensure a single product is amplified by each primer pair. The products amplified using primer specific for LEDGF/p75 are shown at the top, followed by those for LEDGF/p52, then LEDGF/p52b and finally LEDGF/p52bΔE6. The sample numbers (CLL samples and normal B (NB)) that had been analysed for the particular isoform are shown.

To determine the relative expression level of each isoform compared to the reference gene RPL27, the average Ct value of the three replicates was calculated. The ΔCt value was then calculated by subtracting the average Ct value of the

reference gene (RPL27), from the average Ct value of the gene of interest (isoform of LEDGF). The ΔCt values were calculated for all of the cDNA samples including the HEK293 cell line cDNA control.

$\Delta Ct_{(\text{isoform X})}$ in sample Y=$Ct_{(\text{LEDGF isoform})} - Ct_{(\text{reference gene})}$

The ΔΔCt value for each of the isoforms in each CLL sample was then calculated using the formula:

$\Delta\Delta Ct_{(\text{isoform X})}$ in sample Y= $\Delta Ct_{(\text{sample})} - \Delta Ct_{(\text{HEK293})}$

Relative expression of isoform X in sample Y=$2^{(-\Delta\Delta Ct)}$

An example of the results obtained for the LEDGF/p75 isoform is shown in Table 3-6.

Sample	Gene Ct	RPL27 Ct	ΔCt	Relative expression
HEK293 cDNA	22.47	19.2	3.27	
2408	24.76	19.84	4.92	0.3186
2539	26.96	20.4	6.56	0.1022
2472	25.08	19.54	5.54	0.2073
2744	25.48	20.38	5.1	0.2813

Table 3-6: Representative example of the calculation of ΔCt and relative expression values.

Using data from analysis of the expression of LEDGF/p75 in CLL cDNA samples and HEK293 cell line cDNA, the ΔCt value (column 4) is calculated by subtracting the Ct of the value of the RPL27 (reference gene, column 3) from the Ct value of the LEDGF/p75 (column 2). The relative expression (final column) is then determined by $\Delta\Delta Ct_{(isoform\ X)}$ in sample $Y= \Delta Ct_{(sample)} - \Delta Ct_{(HEK293)}$ and $Y=2^{(-\Delta\Delta Ct)}$. The relative expression of LEDGF/p75 was calculated for CLL samples detailed in column 1.

The ΔΔCt values for each isoform were then organised into ascending order and presented in a graphical form (Figure 3-13 - Figure 3-16).

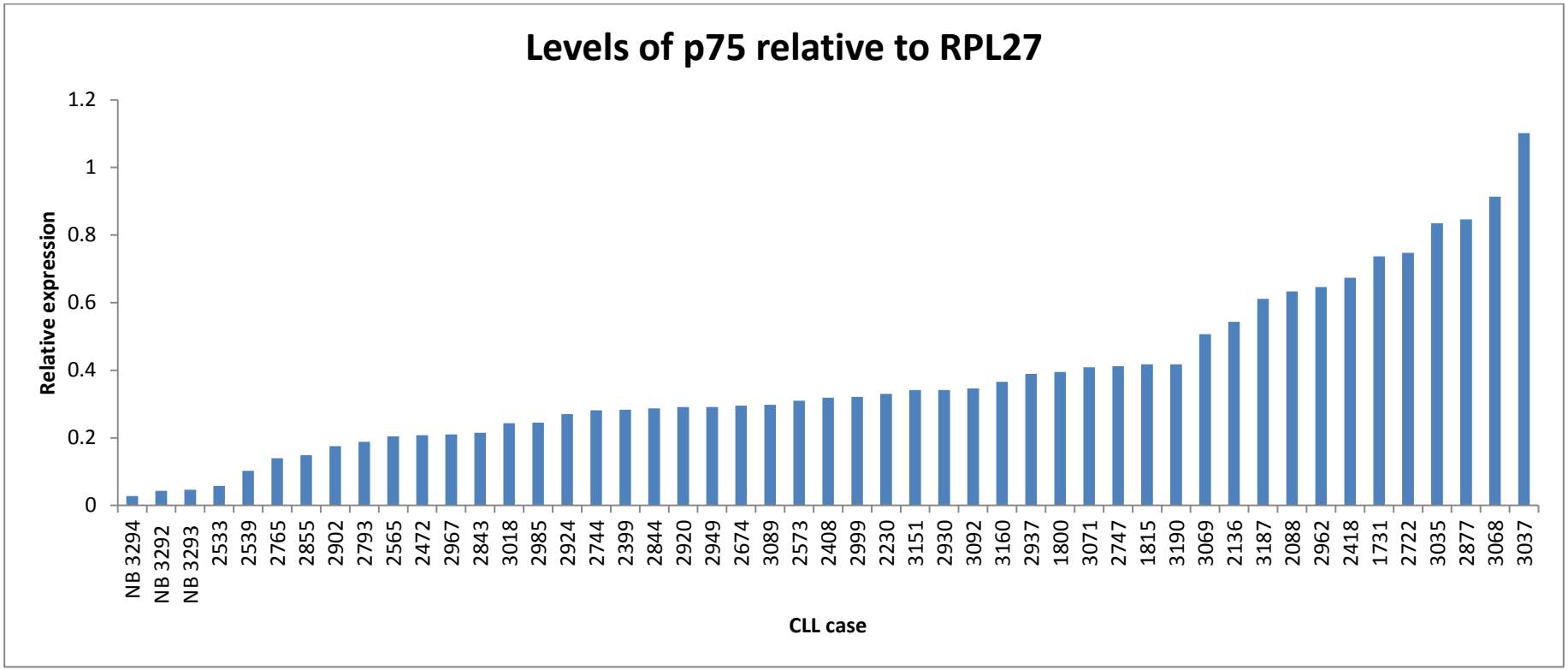


Figure 3-13: The expression levels of LEDGF/p75 in CLL cases and three normal B cell (NB) controls.

The relative expression of LEDGF/p75 in the cases was determined by RT-qPCR and is relative to the expression of the reference gene RPL27.

The expression is presented in increasing order.

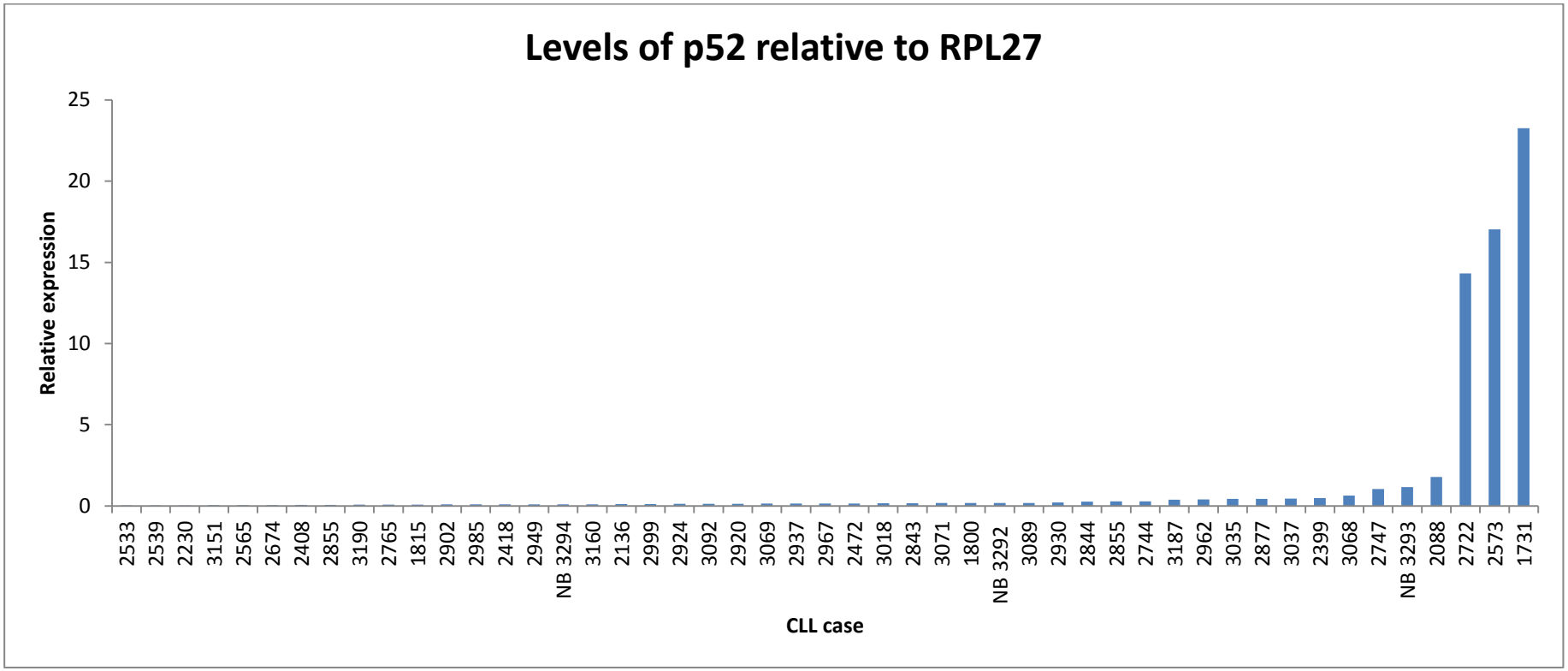


Figure 3-14: The expression levels of LEDGF/p52 in CLL cases and three normal B cell (NB) controls.

The relative expression of LEDGF/p52 in the cases was determined by RT-qPCR and is relative to the expression of the reference gene RPL27.

The expression is presented in increasing order.

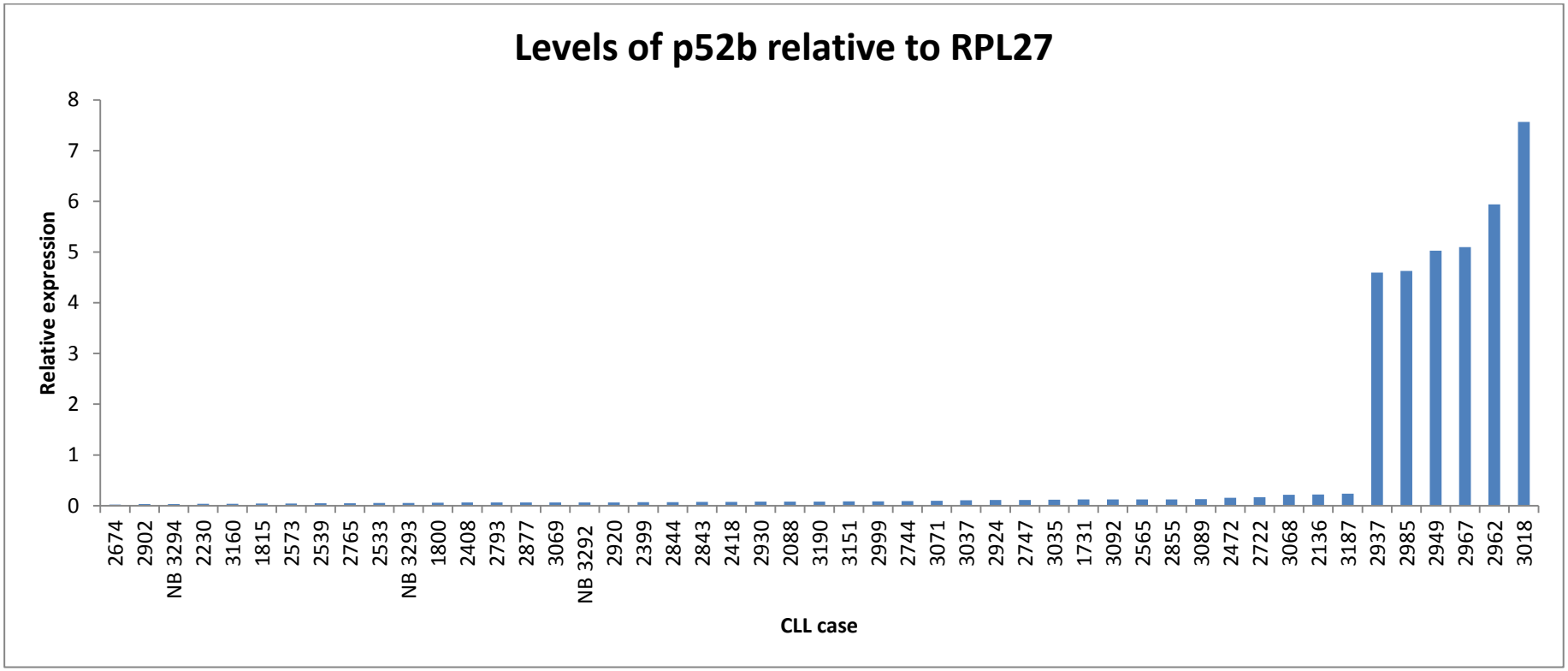


Figure 3-15: The expression levels of LEDGF/p52b in CLL cases and three normal B cell (NB) controls.

The relative expression of LEDGF/p52b in the cases was determined by RT-qPCR and is relative to the expression of the reference gene RPL27.

The expression is presented in increasing order.

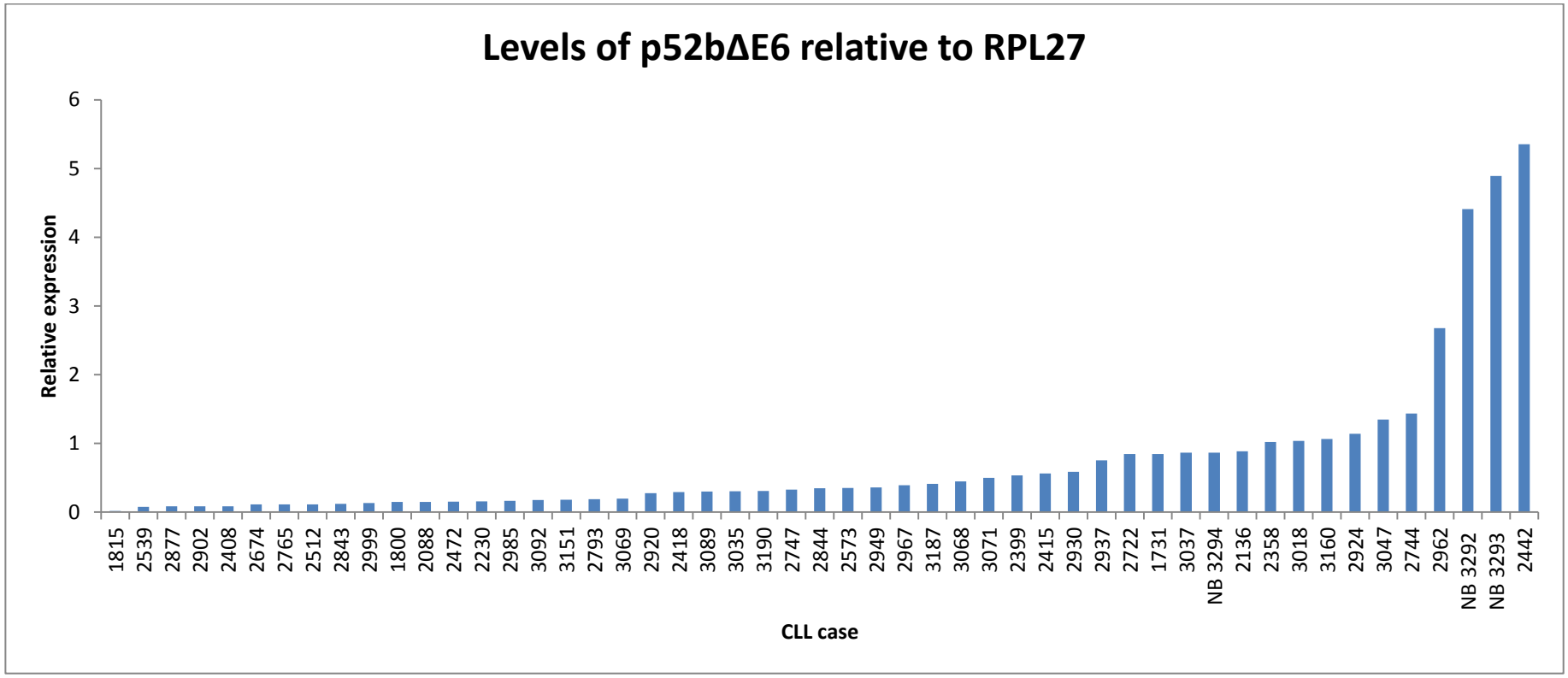


Figure 3-16: The expression levels of LEDGF/p52bΔE6 in CLL cases and three normal B cell controls.

The relative expression of LEDGF/p52bΔE6 in the cases was determined by RT-qPCR and is relative to the expression of the reference gene RPL27. The expression is presented in increasing order.

The results show that the range of the normalised ΔΔCt values for each isoform is variable. Thus, whilst the values for LEDGF/p75 show the narrowest range (0.028 to 1.101), the largest range comes from p52 at 0.0254 to 23.263. LEDGF/p52bΔE6 and LEDGF/p52b show intermediate ranges of expression (0.018 to 5.351 and 0.018 to 7.568 respectively).

The levels of each LEDGF isoform in the CLL cDNA cases were then compared to the levels in 'normal' B cells (3292, 3293 and 3294) using SPSS Statistics 20 software. The results obtained using a Non-parametric Mann-Whitney U Test are shown in Table 3-7.

Isoform	P-value	Significance
LEDGF/p75	0.000	Significant
LEDGF/p52	0.511	Not-significant
LEDGF/p52b	0.069	Not-significant
LEDGF/p52bΔE6	0.008	Significant

Confidence interval level=95%, α=0.05.

Table 3-7: Comparison of the differential expression of the various LEDGF isoforms between CLL and normal B cells.

Two of the four isoforms of LEDGF appear to be significantly differentially expressed between CLL cases and normal B cells. Specifically, LEDGF/p75 is significantly over expressed in CLL cases whilst LEDGF/p52bΔE6 is significantly under expressed in CLL cells compared to normal B cells.

The results show that the expression levels of the LEDGF isoforms differ between CLL and normal B cells. LEDGF/p75 appears to be the lowest expressed isoform in normal B cells (Figure 3-13), closely followed by LEDGF/p52b (Figure 3-15) then LEDGF/p52 (Figure 3-14). The highest expressed isoform in normal B cells was found to be LEDGF/p52bΔE6 (Figure 3-16).

When analysing the levels of the isoforms in CLL cells compared to their levels in the normal B cells, LEDGF/p75 is significantly over expressed in CLL cells compared to normal B cells and LEDGF/p52bΔE6 is significantly under expressed in CLL cells compared to normal B cells (Table 3-7). Although the difference between LEDGF/p52b expression in normal B cells and CLL cells isn't statistically significant (Table 3-7), there is a trend which suggests that this isoform is expressed at greater levels in CLL cases compared to normal B cells (Figure 3-15). In contrast, the results show that the level of expression of LEDGF/p52 observed in normal B cells is randomly distributed within the range seen in CLL cells (Figure 3-14).

3.3.9. Analysis of protein levels of LEDGF isoforms.

Having determined that LEDGF is differentially expressed at the mRNA level, it was felt that it was important to confirm protein translation of the different isoforms. Therefore, a western blot was performed on cell line lysates which was probed with an antibody that recognises residues 85 to 136 which is common to all of the isoforms discussed.

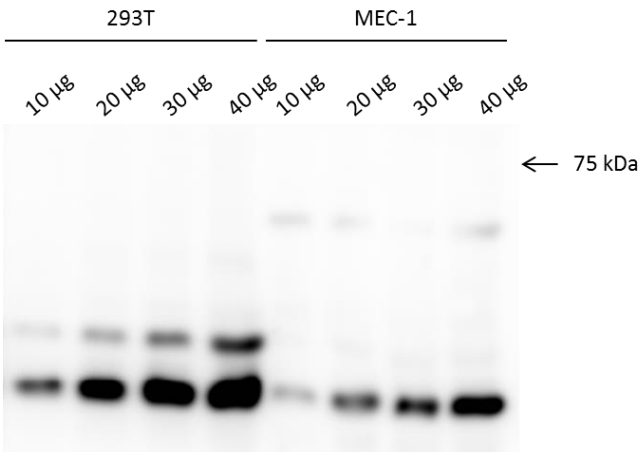


Figure 3-17: Western blot analysis of cell line lysates to examine the presence/absence of the four LEDGF isoforms.

Varying amounts (10-40 µg) of whole cell lysates from 293T and Mec-1 cells were run on a 10% resolving, 4% stacking gel. The proteins were transferred to PVDF membrane and the membrane probed using α-PSIP1 antibody (Aviva, Rabbit polyclonal, 1.25:1000). The levels at which a band of 75 kDa was expected, is highlighted.

As shown in Figure 3-17, a band of 75 kDa which would correspond to LEDGF/p75 was not detected, even when large amounts of lysates were analysed. In order to verify the presence of LEDGF protein, 10 µg of the same 293T lysate plus that from HeLa cells was analysed by western blotting and probed using an alternative antibody (α-LEDGF, BD Biosciences, Mouse monoclonal, 1:4000) that should detect LEDGF/p75. The result shows that LEDGF protein is present in these lysates (Figure

3-18) and suggests that the first antibody, which should detect all four isoforms of LEDGF, is not suitable for this study.

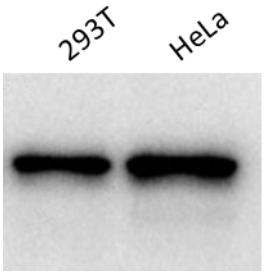


Figure 3-18: Western blot analysis of cell line lysates to examine the presence/absence of LEDGF isoforms.

10 µg of whole cell lysates from 293T and HeLa cells were run on a 10% resolving, 4% stacking gel. The proteins were transferred to PVDF membrane and the membrane probed using α-LEDGF antibody (BD Biosciences, mouse monoclonal, 1:4000). The bands in both lanes represent LEDGF/p75.

3.3.10. Correlation between the expression levels of the LEDGF isoforms and clinical and prognostic factors.

To help understand if there was a clinical significance to the different expression of the isoforms of LEDGF in CLL cells, the relationship between the expression of the isoforms and other clinical and prognostic factors was assessed (see Appendix B). Preliminary analysis suggested a possible link between the expression of LEDGF isoforms and the degree of CD38 expression. CLL cells are classed as CD38+ if over

30% of the total B-cells are CD38+, and negative if less than 30% are CD38+[83]. Therefore, the data relating to the expression of CD38 (a variable data set) was modified to be a categorical data set and samples were denoted positive or negative for CD38 expression. With this modification to the data, the relationship was re-evaluated and the results are shown in Figure 3-19.

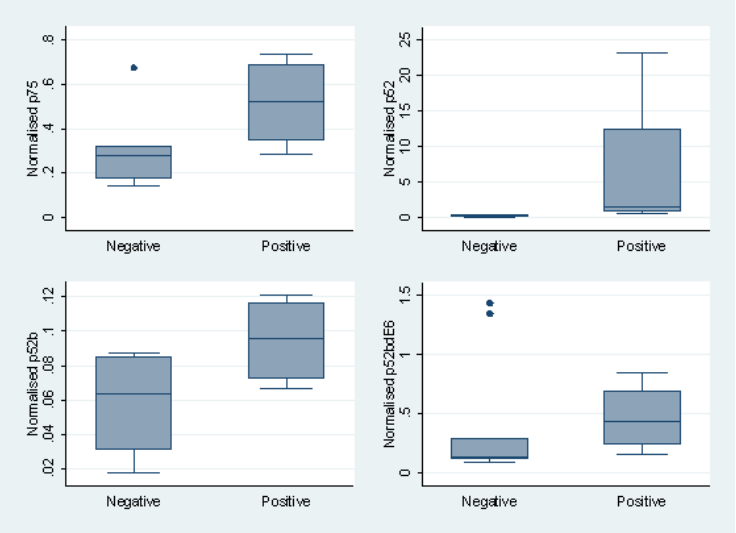


Figure 3-19: Correlation between the levels of expression of the LEDGF isoforms and CD38 expression.

ANOVA tests were employed to determine the correlation between the expression levels of the isoforms of LEDGF and CD38 expression.

The results show that in patients whose CLL cells are CD38 positive, although not significant, levels of mRNA for all LEDGF isoforms are generally higher compared to

those whose cells are CD38 negative (LEDGF/p75 (p=0.07), LEDGF/p52 p=0.06, LEDGF/p52b p=0.13 and LEDGF/p52bΔE6 p=0.40).

3.4. Discussion

With current treatment strategies, CLL remains incurable. Thus, there remains the need to identify potential new drug targets. One proposed target is LEDGF, of which several isoforms have been described. The aims of this chapter were five fold. Firstly, I set out to identify which of the reported isoforms of LEDGF are expressed in CLL cells using end point PCR. Secondly, having identified which of the isoforms are expressed, I used quantitative RT-qPCR to determine the relative levels of each isoform compared to the reference gene RPL27. Thirdly, the levels of the LEDGF isoforms expressed in CLL were compared to normal B cells. I then attempted to identify a suitable antibody that could be used to determine the protein levels for the different isoforms expressed in CLL. Finally, I analysed the data to see if there was any correlation between the expression of the different isoforms in CLL and any of the prognostic factors used in CLL.

The results demonstrated that of the 6 isoforms of LEDGF described by Huang *et al.*[57], only 4 were detected in CLL cells. Thus, LEDGF/p75, p52, p52b and p52bΔE6 were detected by RT-PCR. However, not all isoforms were detected in all of the CLL cases examined. Of the 78 samples included in the study using end-point PCR, 10

did not appear to express p75, 11 did not express p52, p52b was not detected in 2 cases whilst the p52Δe6 isoform was not detected in 13 cases (Table 3-5).

Whilst RT-PCR can determine the presence or absence of a particular LEDGF isoform, it does not provide a relative level of expression of one isoform compared to another or the level of expression of a particular isoform in different samples. Therefore, a quantitative PCR approach was employed. Results from this analysis showed that LEDGF/p75 has the smallest range of expression between the CLL cases examined (Figure 3-13). This is in contrast to LEDGF/p52 which had a wide range of expression levels (Figure 3-14). Using the NB4 AML cell line, Huang *et al*. suggested that the LEDGF/p52b isoform of LEDGF is more highly expressed than the LEDGF/p52bΔE6 isoform[57]. However, analysis of these 2 isoforms in CLL does not fully support this hypothesis. In CLL, the level of expression of LEDGF/p52b is more variable than seen for p52bΔE6, and is not simply expressed at a higher level.

The expression levels of the 4 different isoforms of LEDGF identified in CLL cells were also examined in normal B cells. Levels of the LEDGF/p75 isoform were very consistent and were lower than the levels seen in CLL cells, and this difference was statistically significant (p=0.000) (Figure 3-13, Table 3-7). As was seen for LEDGF/p75, the levels of LEDGF/p52b were fairly consistent between samples (Figure 3-15), unlike LEDGF/p52 which showed a wide variation in expression between the different samples (Figure 3-14). Within each normal B sample, levels of both LEDGF/p52 and p52b were higher than that seen for LEDGF/p75, although levels of LEDGF/p52b were lower than LEDGF/p52. However, this was not the case

for the CLL samples, as there was no correlation observed as regards levels of expression of each of the isoforms. Interestingly, LEDGF/p52bΔE6, which was reported by Huang *et al*. to have the lowest level of expression in AML[57], actually had the highest average level of expression amongst the normal B cell samples (Figure 3-16).

Analysis of the RT-qPCR results using SPSS Statistics 20 software showed that of the four isoforms, LEDGF/p75 was significantly over expressed in CLL cases compared to Normal B cells (p=0.000), whilst LEDGF/p52bΔE6 was significantly under expressed in CLL cases compared to Normal B cells (p=0.008) (Table 3-7). Even though the expression of LEDGF/p52b isn't significantly different between CLL and normal B cells, the general trend is that the levels are higher in CLL (Figure 3-15). In contrast to these isoforms, the expression level of LEDGF/p52 was shown to be highly variable in both normal B and CLL cells and there doesn't appear to be any trend/pattern associated with the levels of expression of this variant (Figure 3-14). These results are in agreement with the conclusion of both Basu *et al*. and Daugaard *et al*. that over expression of LEDGF/p75 mRNA may be a general feature of cancer[81, 82].

Even though the significant over expression of LEDGF/p75 in CLL is in accordance with the literature, specifically, that this isoform is over expressed in many cancers, caution should be employed with this study due to the low number of normal B cell control samples (n=3) included. The normal B cell control samples are obtained from healthy volunteers and are therefore in limited supply. However, it should be

noted that the normal B cell samples used within this study were all processed in the same way, as where the cohort of the CLL samples (a departmental standard operating procedure governs the processing of all blood samples). As there was not one single or group of samples who all exhibited the same pattern in isoform expression a batch effect is not considered a major variable.

Due to the exploratory nature of this study, no formal *a-priopi* power analysis was performed to determine the appropriate sample size; essential information such as population variance estimates and effect size could not be determined prior to the study. Instead, a sample cohort of convenience was used due to their availability. A retrospective or observed power analysis was not performed due to inappropriateness; post-hoc power procedures incorporate improper calculations of a probability and do not add to the interpretation of the results since the observed power is completely determined by the observed p-value.

Over expression of the p52bΔE6 isoform in HEK293 cells has been shown to increase apoptosis, whilst in contrast, over expression of the p52b variant in the same cells appeared to increase viability[57]. It would therefore appear that deletion of exon 6 imparts a pro-apoptotic phenotype on LEDGF. As the results presented in this thesis have shown that the expression levels of LEDGF/p52bΔE6 are generally lower in CLL compared to normal B cells, this would suggest that the pro-apoptotic function of this LEDGF variant is reduced in CLL cells, and hence provide CLL cells a mechanism to evade cell death.

LEDGF/p75 has been suggested to be a cancer associated protein. As discussed, it is involved in translocations involving the NUP98 protein and has an increased mRNA and protein expression in a variety of cancers. It is also part of the MLL complex (reviewed in chapter 1 and 5) that targets the complex to chromatin to regulate expression of *HOX* genes. The data presented in this thesis further supports the theory that LEDGF is a cancer associated protein.

Recent evidence indicates that augmented states of cellular oxidative stress contribute to the development of cancers[82, 84]. LEDGF/p75 is a known stress-response protein that binds to STRE and HSE and activates expression of a number of stress response genes, thus protecting mammalian cells against oxidative and environmental stressors and promoting cell survival. Thus, it has been shown LEDGF/p75 over expression protects cells against serum starvation[85] as well as thermal and oxidative stress[1]. The data presented by Singh *et al*. suggests that this resistance could be caused by the LEDGF/p75 dependent over expression of heat shock proteins, namely Hsp27 and αβ-crystallin[1]. An increased expression level of LEDGF/p75 mRNA has been demonstrated in murine lens epithelial cells and Cos7 cells following heat and oxidative stress[86]. Concomitant to this was an increase in the levels of LEDGF/p75 protein but not LEDGF/p52. LEDGF/p75 has also been shown to have a protective role in human retinal pigment epithelial cells exposed to UVB radiation via the up-regulation of Hsp27 expression[87]. Recent evidence suggests that CLL cells are under oxidative stress[88]. Given that it has been shown

that low/absent levels of LEDGF/p75 result in cell death[85] the data suggests that an over expression of p75 would provide an anti-apoptotic function to CLL cells.

Three of the isoforms of LEDGF that were identified as being expressed in CLL share a common amino terminal sequence. Thus, LEDGF/p75, p52 and p52b are identical up to and including exon 9. The C-terminal tails of both LEDGF/p52 and p52b are derived from within an intron of the LEDGF/p75 isoform. The predicted molecular mass of p75 is 60 kDa, whilst for p52b and p52 it is 40 and 38 kDa respectively. However, western blot analysis shows that LEDGF/p75 and LEDGF/p52 run with apparent molecular weights of 75 and 52 kDa[28]. This anomaly suggests the proteins are significantly modified post translationally.

The fourth isoform detected in CLL was the LEDGF/p52bΔE6. As this isoform has a complete deletion of exon 6, it alters the reading frame of the mRNA sequence and results in a premature stop codon after residue 167. The resulting truncated protein would have a predicted molecular mass of 19 kDa. The deletion of exon 6 and subsequent alteration of the reading frame disrupts the end of the NLS and abolishes the AT hooks and CR2 and 3 that is common to the other isoforms (including LEDGF/p52b). If this isoform undergoes significant post translational modification, it is difficult to predict what the apparent molecular weight of the protein would be on a western blot.

An attempt was made to analyse the expression of the 4 isoforms of LEDGF in CLL cells by western blot, but this proved unsuccessful. Western blot of LEDGF isoform

expression in CLL would be predicted to show a 'ladder' effect if each of the isoforms was present in the particular sample being analysed. LEDGF/p75 would be the largest isoform so would be the top of the lane, followed by p52b, then p52 and, finally, p52bΔE6 as this has the lowest molecular weight. The antibody used in this investigation was raised against an immunogen of residues 1-51 which was common to LEDGF/p75, p52, p52b and p52bΔE6. To confirm the specificity of the antibody, a western blot was performed using a lysate from 293T cells as LEDGF/p75 had been shown to be present in this cell line[29]. However, the western blot (Figure 3-17) failed to show the presence of a band at 75 kDa in either of the cell lysates tested, even when high levels of protein were loaded. A subsequent blot (Figure 3-18) using the same 293T lysate (plus an additional lysate from HeLa cells) that was probed with a different antibody (immunogen of residues 85-188) that would detect the p75 isoform (as well as LEDGF/p52 and p52b), showed the presence of a band that would represent LEDGF/p75 in these cell lines.

As a positive band at 75 kDa could not be identified in the 293T lysate using the antibody predicted to detect the four isoforms I have described in CLL, the double bands present in the 293T cell lysates at 37 and ~45 kDa cannot be assigned to any of the isoforms with any certainty (Figure 3-17).

There has been a flow cytometry and indirect intracellular staining method developed by Mous and colleagues to examine the expression levels of LEDGF[89]. However, the antibody used for the assay has an immunogen common to both LEDGF/p75 and p52 (and in theory p52b). Thus, in this assay it would be difficult to

distinguish between the level of expression of each individual isoform and indeed, analysis of cell lysates showed the presence of several bands on western blot[89]. Therefore, these methods would not be suitable to analyse the protein levels of each isoform in CLL cells due to their ability to detect at least two of the main isoforms of LEDGF.

It has been shown in this chapter that the levels of the LEDGF/p75 and p52bΔE6 are differentially expressed between CLL cells and normal B cells. Thus, CLL cells over express LEDGF/p75, whilst under expressing p52bΔE6, and although not significant, there is a trend to suggest that LEDGF/p52b is also over expressed in CLL cells. In addition to this, I have shown that there is a trend for patients who are CD38 positive to have a higher level of expression of each of the isoforms of LEDGF (Figure 3-19).

LEDGF/p75 has been shown to be over expressed in chemotherapy resistant and relapsed AML patients, and its effect on chemotherapy resistance was demonstrated by exogenous over expression in HEK293 cells which led to a reduction in daunorubicin induced apoptosis[57]. The report showed that expression of LEDGF/p52b also had a similar effect. Whilst this report demonstrated that expression of LEDGF/p52ΔE6 leads to an increase in spontaneous apoptosis, this effect was shown to be negated by co-expression of LEDGF/p75. The authors concluded that expression of LEDGF/p75 and p52b protects AML cells against spontaneous and drug induced cell death *in vitro*, and suggest that over expression

of these LEDGF isoforms in primary AML cells that are resistant to therapy is not coincidental.

The results from this thesis showing that CD38 positive CLL cells tend to have increased levels of all LEDGF isoforms support this suggestion. CD38 is a receptor that has been shown to cause an induction of proliferation and increased cell survival in CLL[90], and expression of this receptor has been associated with the aggressive form of CLL, a poor response to chemotherapy and a shorter survival time[90, 91]. Thus, those cases of CLL that are more likely to be resistant to chemotherapy are those that express the higher levels of LEDGF.

Future work

Even though differences in expression of two of the isoforms of LEDGF have been shown to be statistically significant between normal B cells and CLL cells (namely LEDGF/p75 and LEDGF/p52bΔE6), this was not the case for LEDGF/p52b. However, there is a definite trend in the data suggesting that LEDGF/p52b is over expressed in CLL (Figure 3-15). In order to see if this trend is true and does become significant, a greater number of normal and CLL cases could be screened.

It will be necessary to confirm whether the differences in the level of expression of the isoforms identified by RT-qPCR are mirrored by the levels protein expression. However, a suitable antibody that is capable of detecting all of the isoforms by western blot, or several antibodies capable of distinguishing each of the isoforms by

flow cytometry would be required. Suitable positive controls would also be required, although this issue could be overcome for western blotting through the use of purified over expressed recombinant proteins.

Chapter 4 : Mutational Analysis of SF3B1 and Correlation with LEDGF Isoform Expression

4.1. Introduction

In 2012, Quesada and colleagues published results of a study in which they performed whole-exome sequencing on matched tumour and normal samples from 105 CLL cases[70]. Of these, 60 were IGHV mutated cases whilst the remaining 45 were unmutated cases. They identified 78 genes that had mutations that could potentially lead to functional alterations which were present in more than one sample. One of the genes shown to be recurrently mutated in CLL is SF3B1, which forms part of the catalytic core of the splicing machinery.

Two further published reports confirmed that SF3B1 is recurrently mutated in CLL cells. Furthermore, it was demonstrated that SF3B1 mutations were frequently encountered in patients refractory to treatment with fludarabine or those with progressive disease[71], or in patients with deletions of 11q, who generally have a poor prognosis[63].

These mutations of SF3B1 in CLL patients are clustered within the HEAT repeats located within the C-terminal region of the protein[63, 70]. One very interesting point

from these reports was the link between the mutational status of SF3B1 with the splicing pattern of other mRNAs within the cell.

One of the genes predicted to be alternatively spliced between SF3B1 mutated and unmutated cases is LEDGF/p75 (NM_033222) ($p= 0.0001$)[70]. Having demonstrated that CLL cells express different isoforms of LEDGF, and that these were differentially expressed between CLL cases (see Chapter 3 for results), it was felt pertinent to examine the mutational status of the SF3B1 gene in the same cohort of patient samples.

The aims of this chapter, therefore, were:

1. Using RT-PCR and Sanger sequencing, determine the mutational status of the HEAT repeat region of SF3B1, in the cohort of CLL samples that were examined for LEDGF expression in Chapter 3.
2. To determine if there is any correlation between the mutational status of SF3B1 and the differential expression of the LEDGF isoforms identified in chapter 3.
3. To investigate the potential clinical implications of the mutational profiles in CLL.

4.2. Methods

4.2.1. Primer design

Given that the PCR products were going to be sequenced and the limitations with this method for long amplicons, it was deemed necessary that the region of interest be amplified as two separate products. Using this approach, the main recurring mutational hotspots that have been previously reported (namely nucleotides around residues 625 and 666, and 700[65]) would be in the centre of the regions to be sequenced allowing for greater confidence in the results. Therefore, 2 sets of PCR primers (Table 4-1) were designed around the HEAT regions based on the sequence NM_012433.2 (NCBI reference). The mRNA sequence for exons 11-17, which contains the HEAT repeat regions is shown below (Figure 4-1). This region contains the most commonly reported mutations.

Primer name	Primer sequence
SF3B1pcr507-693F	caccaccaatgagaaaggctg
SF3B1pcr507-693R	gctcatccacaagaccatgttc
SF3B1pcr600-823F	gcaaaggctgctggtctgg
SF3B1pcr600-823R	catcctgtgctgccagaagtg

Table 4-1: Sequences of RT-PCR primers used to screen the SF3B1 gene.

Both primer pairs are designed to the HEAT repeat region of the gene. Specifically, the top primer pair will amplify the region from amino acid 507 to 693 and the bottom pair will amplify the region from amino acid 600 to 823. The primers will be used in separate reactions.

GTTGATGTTGATGAATCAACACTTAGTCCAGAAGAGCAAAAAGAGAGAAAAATAATGAAGTTGCTTTT

AAAAATTAAGAATGGAACACCACCAATGAGAAAGGCTGCATTGCGTCAGATTACTGATAAAGCTCGTG

AATTTGGAGCTGGTCCTTTGTTTAATCAGATTCTTCCTCTGCTGATGTCTCCTACACTTGAGGATCAA

GAGCGTCATTTACTTGTGAAAGTTATTGATAGGATACTGTACAAACTTGATGACTTAGTTCGTCCATA

TGTGCATAAGATCCTCGTGGTCATTGAACCGCTATTGATTGATGAAGATTACTATGCTAGAGTGGAAG

GCCGAGAGATCATTTCTAATTTGGCAAAGGCTGCTGGTCTGGCTACTATGATCTCTACCATGAGACCT

GATATAGATAACATGGATGAGTATGTCCGTAACACAACAGCTAGAGCTTTTGCTGTTGTAGCCTCTGC

CCTGGGCATTCCTTCTTTATTGCCCTTCTTAAAAGCTGTGTGCAAAAGCAAGAAGTCCTGGCAAGCGA

GACACACTGGTATTAAGATTGTACAACAGATAGCTATTCTTATGGGCTGTGCCATCTTGCCACATCTT

AGAAGTTTAGTTGAAATCATTGAACATGGTCTTGTGGATGAGCAGCAGAAAGTTCGGACCATCAGTGC

TTTGGCCATTGCTGCCTTGGCTGAAGCAGCAACTCCTTATGGTATCGAATCTTTTGATTCTGTGTTAA

AGCCTTTATGGAAGGGTATCCGCCAACACAGAGGAAAGGGTTTGGCTGCTTTCTTGAAGGCTATTGGG

TATCTTATTCCTCTTATGGATGCAGAATATGCCAACTACTATACTAGAGAAGTGATGTTAATCCTTAT

TCGAGAATTCCAGTCTCCTGATGAGGAAATGAAAAAAATTGTGCTGAAGGTGGTAAAACAGTGTTGTG

GGACAGATGGTGTAGAAGCAAACTACATTAAAACAGAGATTCTTCCTCCCTTTTTTAAACACTTCTGG

CAGCACAGGATGGCTTTGGATAGAAGAAATTACCGACAG

Figure 4-1: mRNA sequence of SF3B1 exons 11-17.

The mRNA sequence is to the SF3B1 gene, NM_012433.2 (NCBI reference) and is part of the HEAT repeat region. Each exon is highlighted by a colour change (blue to black). Primers pair are highlighted in yellow (SF3B1pcr507-693F and R) or red (SF3B1pcr600-823F and R). The mutational hotspots at amino acids 625, 666 and 700, are underlined.

4.2.2. Optimisation of PCR reaction conditions

Reaction parameters where optimised as described in Chapter 2.2.3 using an annealing temperature gradient of 57 – 67 °C. For these optimisations, cDNA from HEK293 cells was used as template and a negative control in the form of NFW was included. PCR products were analysed by electrophoresis on 2% agarose gels as described in Chapter 2.2.4 and the annealing temperature determined by visualisation of the gel. A sample of the PCR product amplified at the chosen annealing temperature for each primer pair was purified from the agarose gel as described in Chapter 2.2.5 and sequenced as described in 2.2.10.

4.2.3. Analysis of the mutation status of the HEAT repeats of the SF3B1 gene

CLL cDNA samples that had been analysed for the presence or absence of the LEDGF isoforms (see Chapter 3) were screened to determine the mutational status of the HEAT repeats of the SF3B1 gene.

To determine if there was a correlation between the relative levels of LEDGF and the mutational status of the SF3B1 gene, statistical analysis was performed by Dr Fotis Polydoros. ANOVA tests were performed using the STATA version 13 statistical software package. Data was presented in the form of boxplots. In addition, the relationship between the mutational status of the SF3B1 gene and other clinical factors (for example age) was also investigated.

4.3. Results

4.3.1. Optimisation of RT-PCR conditions

Using cDNA generated from HEK293 cell line, the annealing temperatures for the primer pairs SF3B1pcr507-693F and SF3B1pcr507-693R; and SF3B1pcr600-823F and SF3B1pcr600-823R (Table 4-1) were determined and shown to be 64 °C and 67 °C respectively (Figure 4-2).

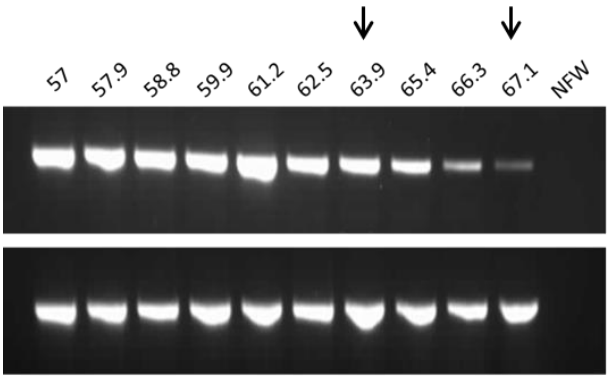


Figure 4-2: Optimisation of PCR annealing temperature.

PCR reactions were performed using cDNA from HEK293 cells with a range of annealing temperatures with primer pairs SF3B1pcr507-693F and SF3B1pcr507-693R (upper panel) and SF3B1pcr600-823F and SF3B1pcr600-823R (lower panel). Reactions were analysed by agarose gel electrophoresis and the annealing temperature determined visually (indicated by arrows). Specifically, the annealing temperature chosen for primer pair SF3B1pcr507-693F and SF3B1pcr507-693R was 64 °C and for SF3B1pcr600-823F and SF3B1pcr600-823R was 68 °C.

4.3.2. Determination of the mutational status of SF3B1

Utilising the chosen reaction conditions described above, the HEAT repeat region of interest of SF3B1 was amplified by PCR using cDNA samples from the same cohort of CLL cases used in Chapter 3. Figure 4-3 shows a representative gel of PCR products amplified with each primer set.

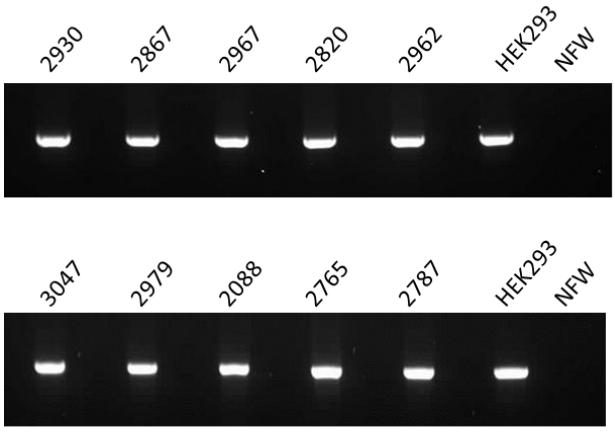


Figure 4-3: Representative image of the PCR amplification of the HEAT region of the SF3B1 gene in CLL cells.

Primer pairs SF3B1pcr507-693F and SF3B1pcr507-693R (upper panel) and SF3B1pcr600-823F and SF3B1pcr600-823R (lower panel) were used to amplify the HEAT repeat region from the SF3B1 gene in a cohort of CLL cDNA samples (numbered). RT-PCR reactions were analysed by agarose gel electrophoresis and bands purified and sequenced as described.

The sequences of the PCR products were analysed using Chromas Lite software and compared against the published sequence of the SF3B1 gene (Figure 4-1). Chromas traces for those samples in which mutations were identified are shown in Figure 4-4. The results identified SF3B1 gene mutations in 13 of the 75 CLL samples analysed (17%) which is a similar incidence as has been previously reported[65]. With the exception of one case, all of the mutations identified were missense mutations, with the most common resulting in changes to amino acids 663 (n=2), 636 (n=2), 666 (n=2), 684 (n=2) and 700 (n=2). In four patients within this cohort, two concurrent mutations within the HEAT repeat region of SF3B1 were identified (Figure 4-4 and Table 4-2). Unlike previous reports, deletion mutations were not identified[92]. In sample 2230, a nonsense mutation (Q659stop) was identified (Figure 4-4, Table 4-2). This mutation is a novel finding and has not been previously described. This mutation, if translated, would result in a truncated SF3B1 protein lacking 646 amino acids from the carboxy terminus.

A summary of the amino acid changes that would occur as a result of the mutations identified in this study is given in Table 4-2.

CLL case	Mutation status	Nucleotide change
2408	T663I	c.1988C>T
2358	T663A	c.1987A>G
2472	R630K, A636S	c.1889G>A, c.1906G>T
2230	Q659Stop, R661I	c.1975C>T, c.1982G>T
2789	N533Y, A660V	c.1597A>T, c.1979C>T
2793	K666N	c.1998G>T
2843	R684T	c.2051G>C
2877	R684I	c.2051G>T
3160	K666E	c.1996A>G
2999	K700E	c.2098A>G
3071	K700E	c.2098A>G
2674	A636T/P, P642S	c.1906G>C/A, c.1924C>T
2539	G742D	c.2225G>A

Table 4-2: Summary of the inferred amino acid change in the SF3B1 gene.

The CLL case shown to have the mutation in their SF3B1 gene is in the left hand column. The second column details the amino acid that is subject to the change and what the change is. More specifically, column three details the nucleotide that is changed that causes the change in amino acid.

Figure 4-4: Chromas traces of the SF3B1 gene sequences that contain point mutations.

In each trace, the amino acid sequence that is mutated is highlighted by a red box. If the peaks are a mixture of two, ie, the mutation is heterozygous, an 'N' is used. If there is one single peak representing one nucleotide, that peak is described specifically.

4.3.3. Analysis of the potential truncated protein

One of the SF3B1 mutations identified in this study was a previously unreported nonsense mutation that introduced a premature stop codon at nucleotide 1975. Translation of the mRNA containing this mutation would result in the expression of a truncated SF3B1 protein. To determine if this was the case, a western blot was performed (Chapter 2.3.1 and 2.3.2) using whole cell lysate of this sample, along with that from a CLL case shown to have wild type SF3B1 gene and a positive control (Jurkat cell lysate). The full length SF3B1 protein is 1304 amino acids long and runs with an apparent molecular weight of 155 kDa. The predicted truncated protein produced would be 658 amino acids and would therefore be predicted to have an apparent molecular weight of ~75 kDa. The results are given in Figure 4-5.

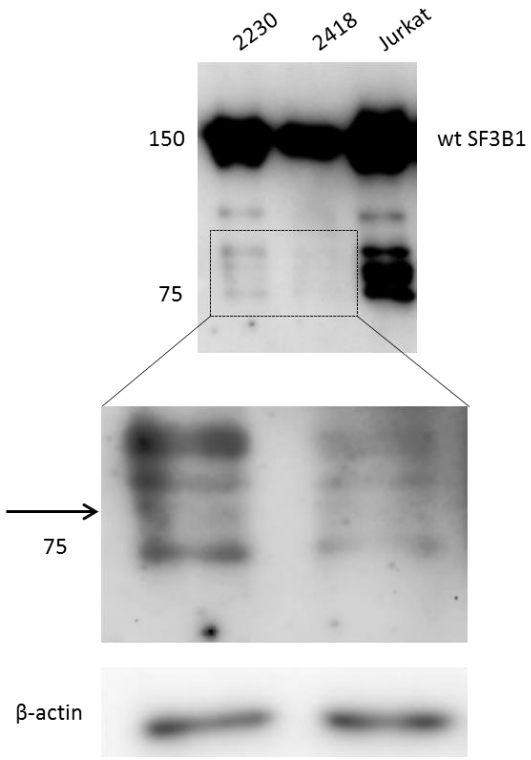


Figure 4-5: Western blot analysis of whole cell protein lysates of CLL cases 2230 and 2418 and the Jurkat cell line.

10 µg of whole cell lysate from two CLL cases (one with the potential nonsense mutation in the SF3B1 gene (2230) and one with wild type SF3B1 gene (2418)) and a jurkat cell line (positive for SF3B1) was run on a 10% resolving and 4% stacking gel. Proteins where transferred to PVDF membrane. The membrane was probed with anti-SF3B1 antibody (Abcam, Mouse monoclonal, 1:2500) (top). The highlighted area where the putative truncated protein is expected to run is shown in the enlargement (bottom). A band with the predicted molecular weight is highlighted by an arrow.

In all three cell lysates, there is a band at 155 kDa which represents the wild type SF3B1 protein (upper panel, Figure 4-5) showing that the western blot has worked successfully. The area around the 75 kDa mark has been enlarged for the two CLL cases (lower panel, Figure 4-5). The predicted molecular weight of the truncated protein would be approximately 75 kDa. There is a weak band (indicated by the arrow) running at this molecular weight that is present in sample 2230 (mutated SF3B1) that is much less apparent in sample 2418 (wild type SF3B1). This difference is not likely due to uneven loading of the sample, as the blot for β-actin indicates protein loading for each sample was similar. However, as can be seen in the upper blot, bands of a similar molecular weight are also seen in lysate from Jurkat cells, although due to the level of reactivity, visualisation of a discrete band is difficult. Therefore, it is difficult to confirm with any degree of certainty that a truncated SF3B1 protein is expressed in this CLL sample.

4.3.4. Correlation between SF3B1 mutation and LEDGF isoform expression

It has previously been suggested that mutations within the SF3B1 gene could affect splicing of the LEDGF pre-mRNA[70]. In Chapter 3, it was shown that CLL cells express different isoforms of LEDGF, and the results from Chapter 3 demonstrate that some of these same cases of CLL have mutations within the SF3B1 gene. The relationship between the presence of these mutations and the LEDGF isoform expression was

therefore investigated. In addition, the relationship between the mutation status of the SF3B1 gene and clinical factors (for example age) was also investigated.

The results show that samples with unmutated SF3B1 generally have higher levels of all isoforms of LEDGF (Figure 4-6). However, the differences did not reach statistical significance.

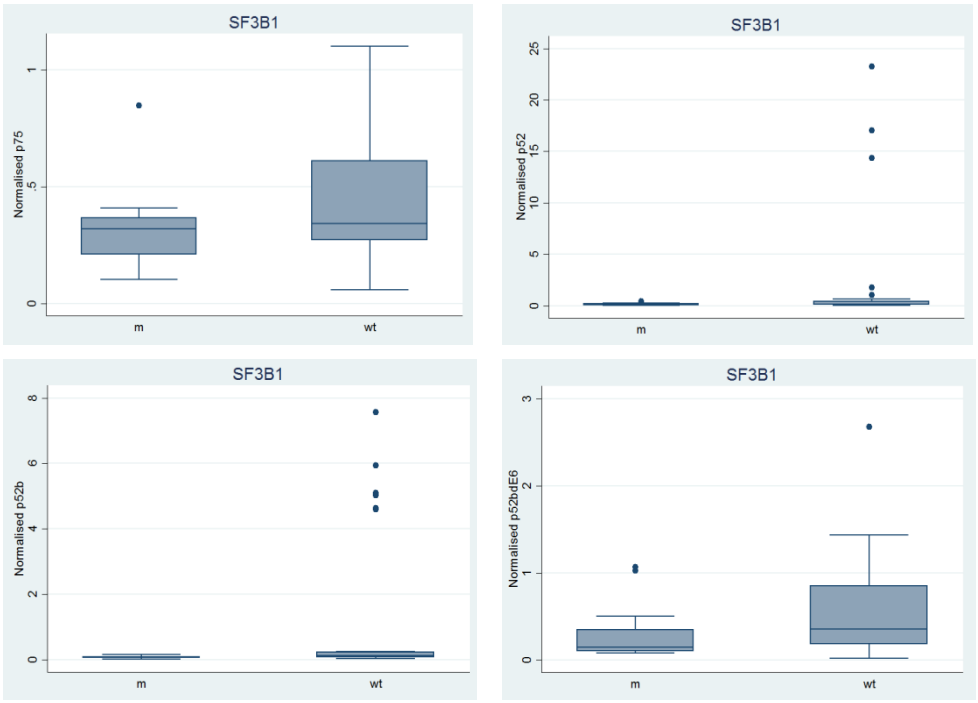


Figure 4-6: Correlation analysis of SF3B1 mutational status and LEDGF isoform expression.

ANOVA tests were employed to determine the correlation between the relative expression of the isoforms of LEDGF and the mutation status of the SF3B1 gene. Data was presented in the form of box plots. The correlation between CD38 expression and the relative expression of each isoform is considered individually.

The results, although not significant, suggest that mutations of SF3B1 do affect the expression of LEDGF. However, this effect is not specific to one isoform, as the levels of all isoforms are reduced in mutated SF3B1 samples. There remained the possibility that samples with extreme relative expression values for LEDGF/p52 and p52b (see Chapter 3.3.8) were skewing the result and should be considered. There was no relationship between the mutation status of the SF3B1 gene and clinical factors (Appendix C).

4.4. Discussion

SF3B1 forms part of the splicing machinery, a complex that is involved in the catalytic removal of introns from pre-mRNA. Recent studies have demonstrated that SF3B1 is frequently mutated in CLL[63, 70, 71, 93]. Moreover, alterations in pre-mRNA splicing in other genes has been suggested and identified in CLL cases harbouring SF3B1 mutations[63, 70]. One of the genes predicted to be affected by mutation of SF3B1 is LEDGF[70]. In the previous chapter, I showed that CLL cells differentially express four isoforms of LEDGF, including two novel isoforms not previously reported in CLL. It was therefore hypothesised that mutation of SF3B1 in CLL will affect the splicing, and hence isoform expression, of LEDGF in these cells. Therefore, cDNA samples that were used to examine the expression levels of the four isoforms of LEDGF (see Chapter 3) were screened for the mutational status of the HEAT repeats of SF3B1. The frequency of SF3B1 gene mutations within our cohort (17.3%) is in-line with published reports[65].

The most commonly reported mutation is K700E (50% of reported mutations), with additional common mutations occurring at G742 (19%) and K666 (12%)[94]. That mutations at these sites are also identified in the cohort examined in this study (Figure 4-4 and Table 4-2) instils confidence in the results. In addition, four patients in this cohort had two concurrent mutations in their SF3B1 gene, and this phenomenon has previously been described in CLL[64].

Given that the aim of this work was to determine whether mutation of the SF3B1 gene correlated with the isoform expression of LEDGF, it was sufficient to use RT-PCR amplification and Sanger Sequencing to determine the presence of mutations in the SF3B1 gene. Sanger Sequencing has been shown to have a detectable mutation load of 5-10% (confirmed by concurrent analysis with Next Generation Sequencing)[64]. However, there is some variability in the mutation load detectability based on the actual platform used[95]. Next generation sequencing is known to be more sensitive (~3% depending on the read count) but as the work did not require the detection of novel mutations (although some have been found), analysis by this method would have been costly both in expense and time, was and hence was not pursued.

It is expected that mutations within SF3B1 would impact on its normal function. In this study, it was found that threonine (a polar residue that can potentially be phosphorylated) at position 663 was changed to the hydrophobic residue isoleucine (sample 2408) or alanine (sample 2358). Thus, these mutations would not only

remove a potential phosphorylation site, but also alter the biochemical properties of the protein as a polar residue is replaced with a hydrophobic residue.

In case 2472, two mutations were identified. The first is an arginine to a lysine change and the second an alanine to a serine alteration. Arginine and lysine are both positively charged amino acids, and hence the mutation is conservative. Both of these amino acids are likely involved in interactions with negatively charged ligands, but as lysine is slightly smaller than arginine, the size difference brought about by the change may have a functional effect. The alanine to serine change, however, is likely more significant. As discussed above, alanine is a small hydrophobic residue, whereas serine is polar and has a hydroxyl group as part of its chain. This mutation not only alters the biochemical characteristics of the protein but also introduces a potential phosphorylation site. Sample 2674 also has two mutations. Firstly alanine is mutated to either a threonine or a proline at residue 636, whilst at residue 642 proline is mutated to a serine. The potential effects of a threonine/alanine alteration have already been discussed. Like alanine, proline is non-polar, and thus such a change will have no effect on charge. However, due to its ring like structure, proline can prevent some secondary structures (for example α-helices) from forming. Therefore, a mutation from alanine to proline will almost certainly have an effect on the tertiary structure of SF3B1. The second mutation in sample 2674 is the loss of a proline in favour of a serine at residue 642. Compared to proline, serine is polar, it is smaller and more flexible in its ability to interact with other residues and it also gives rise to an extra potential phosphorlyation site.

Like cases 2472 and 2674, case 2789 also has two mutations. At residue 533, an asparagine is replaced by a tyrosine. Asparagine and tyrosine are both polar. However, tyrosine is a bigger residue and contains an aromatic side chain with a joined hydroxyl group that is reactive. The change of asparagine to tyrosine would result in the loss of a potential site for N-linked glycosylation and introduce a potential phosphorylation site. The second mutation is at residue 660 and results in the change of an alanine to a valine. Valine is similar to alanine in that both are relatively small, have simple hydrocarbon side chains and are aliphatic and non-polar. However, of the two, valine is slightly larger.

The mutation described in cases 2793 and 3160 is a well-documented change. At residue 666, a lysine is substituted for asparagine or glutamic acid respectively. Samples 2999 and 3071 also have the same mutation (lysine is changed to a glutamic acid) but this is at residue 700. Lysine has a positively charged long, flexible side chain, whereas asparagine is a polar amino acid that has the potential to be modified by glycosylation. Glutamic acid, like lysine, does have a flexible side chain however it does carry a negative charge in comparison to lysine which is positive. Such mutations are likely to impact on both the structure and function of the protein.

In cases 2843 and 2877, an arginine at residue 684 is mutated to a threonine or isoleucine respectively. As discussed above, arginine is a positive amino acid with a long side chain capped by a unique guanidinium group. It is the most basic of the amino acids and is readily involved in interactions with negatively charged DNA. A

change from this to a polar residue like threonine, that also has the potential to be phosphorylated, would affect the function of the protein. Isoleucine is a large residue with a complex hydrophobic side chain and this mutation may have an impact on the tertiary structure of the protein.

In sample 2539 a glycine residue at 742 is mutated to an aspartic acid. Glycine and aspartic acid both have relatively simple side chains. Whilst glycine is a non-polar amino acid, aspartic acid is, in contrast, negatively charged. Thus, whilst such a mutation may not significantly affect the structure of the SF3B1 protein, it may affect its ability to interact with ligands.

Case 2230 is potentially interesting as this would be the first reported nonsense mutation in SF3B1. If translated, this sequence would result in the production of a truncated protein. This sample also contains a second mutation. However, this occurs downstream of the stop codon, and thus is unlikely to be translated.

To determine if this truncated protein is expressed in CLL, a Western blot was performed using whole cell lysate from this sample (Figure 4-5). However, there was no discernible band at the predicted molecular weight that could be confidently assigned to the truncated SF3B1 protein. The premature stop codon in the SF3B1 mRNA may be recognised and degraded by the nonsense mediated decay (NMD) pathway[96] such that the mRNA is not translated into a protein. If the species wasn't degraded by this process and the truncated protein was indeed translated, it raises

the possibility that it may have been degraded in the proteosome following its tagging with ubiquitin[97].

Determining the allele burden of the SF3B1 mutation isn't possible in this investigation as analysis is by end-point PCR and Sanger Sequencing of one PCR reaction. However, as the sequence traces show, the peaks relating to the mutated nucleotide are mixed. It can therefore be inferred that there is a mixture of wild-type and mutated mRNA species present. This is likely a result of a mixture of CLL cells expressing either wild type or mutated SF3B1 ie a heterozygous mutation. This observation is in-line with previous studies[98, 99].

Results presented in the previous chapter demonstrated that CLL cells express four different isoforms of LEDGF. In this chapter, I have shown that some of these CLL cases have mutations within the SF3B1 gene (Table 4-2). To determine if a correlation exists between the presence of these mutations and the levels of expression of LEDGF isoforms in CLL, normalised levels of LEDGF expression were compared between mutated and unmutated SF3B1 samples (Figure 4-6). Whilst the results suggest that unmutated samples generally have higher levels of expression of all LEDGF isoforms, the number of cases analysed meant that these differences did not reach significance. Although the cohort investigated for the mutational status of the SF3B1 gene contained 75 CLL cases, there was not sufficient material to allow the expression level of all isoforms to be determined in every sample by RT-qPCR.

Failure to correctly splice pre-mRNA would likely result in mRNA with an altered open reading frame, such that it would contain aberrant stop codons. As discussed earlier, such mRNA species may be degraded by the NMD pathway. Such an outcome would lead to down regulation of all LEDGF isoforms at the mRNA level. Indeed, the results presented in this thesis demonstrate a trend for CLL cases in the cohort with mutated SF3B1 to have lower levels of all four isoforms of LEDGF as compared to those with wild type SF3B1. The results presented in this thesis, although not significant, suggest that the mutations within SF3B1 identified in this study may impact on the ability of SF3B1 to correctly splice mRNA molecules such as LEDGF, leading to their degradation. However, to confirm this theory, further investigations with an increased cohort of samples will need to be investigated.

As yet, the functional significance of any of the mutations in SF3B1 has not been demonstrated[65]. However, it is thought that with the majority of them clustering in exons 14 – 16 of the protein[99], they may be involved in modifying the functional role of SF3B1[100], and the data presented in this thesis supports this notion.

Future Work

1. Extend the cohort of CLL patients for the study of SF3B1 mutational status.
2. Determine the expression levels of LEDGF isoforms in these patient samples.
3. Repeat the statistical analysis including this new data.

Chapter 5 : Cloning and Purification of LEDGF Isoforms

5.1. Introduction

Whilst purifying positive co-factor 4 (PC4), Ge *et al.* identified a co-purifying ~75 kDa protein which correspond to the full-length isoform of LEDGF[28, 101]. Using degenerate oligonucleotides, and a HeLa cDNA library, they identified an additional shorter p52 isoform. By virtue of differences in the C-terminii of these two known isoforms, LEDGF is likely to exhibit diverse protein interactions and participate in alternate complexes. The differences in protein complexes that LEDGF/p75 and p52 assemble are likely important for specialisation of function and transcriptional outcomes.

Cherepanov *et al.* demonstrated that LEDGF/p75 is a critical component of an HIV integrase complex[102], and subsequently showed that this requires the presence of residues 339 – 442 (the IBD)[103]. A number of other proteins have since been shown to interact with LEDGF/p75 through its' IBD motif. Maertens *et al.* demonstrated that the nuclear localisation of the Myc interacting protein JPO2 was dependent on its binding to residues 347-429 of the LEDGF/p75 IBD[104]. This interaction was subsequently confirmed in 2007 by Bartholomeeusen *et al.* who demonstrated that LEDGF/p75 also interacts with pogo transposable element-derived protein with zinc

finger (PogZ)[105, 106]. Cell division cycle 7-activator of S-phase kinase (Cdc7-ASK) also interacts with the IBD domain of LEDGF/p75[107]. Critically, this interaction requires the auto-phosphorylation of the kinase, and following such binding, LEDGF becomes primarily phosphorylated on Ser-206.

The role of LEDGF/p75 over expression in CLL, and translocations involving NUP98 in a variety of leukemias has been discussed (Chapters 1 and 3). Importance of its role in haematopoietic development and cancer is further inferred by its participation in the MLL/menin complex[36]. Within the complex, menin is required to directly link LEDGF/p75 to MLL and subsequently direct the complex to chromatin[36] and specifically to histone H3K36me[3 24]. More recently, Mereau *et al.* have confirmed that LEDGF/p75 is a critical component of menin/MLL complex[108].

With regards LEDGF/p52, this isoform is shown to specifically interact with SRSF1, an essential splicing factor[27]. Pradeepa *et al.* published a list of potential protein-partners of LEDGF/p52[24]. Of the proteins identified, ~95% have a role in pre-mRNA processing and the rest are involved in transcription. One of these proteins is indeed SRSF1 thus confirming the findings of Ge *et al*. Furthermore, the PWWP domain of LEDGF was not required for this interaction further validating the specificity of the interaction with LEDGF/p52.

To date, only one protein, MeCP2, has been shown to interact with both LEDGF/p75 and p52, via a shared amino acid sequence[109]. Importantly, the region of LEDGF/p75 and p52 responsible for the interaction with MeCP2 is preserved in the LEDGF/p52b

and p52bΔE6 isoforms, suggesting preservation of interaction. Given that LEDGF/p75 and p52 have unique and yet disparate interacting protein partners, and that two further isoforms (namely LEDGF/p52b and p52bΔE6) are now described, the pursuit of differences in the composition of each of the LEDGF isoform complexes was deemed worthwhile. Identification of the proteins that bind to the two newly described isoforms may also allow us to dissect their functional roles. In addition to this, the differences in complexes between the two well-known isoforms, LEDGF/p75 and p52 and their respective ΔPWWP domain mutants, may provide additional insights into the role of the PWWP domain, not only in the context of LEDGF, but also in terms of its role in binding modified histone or non-histone proteins and for chromatin localisation.

The aims of this chapter, therefore, were:

1. To prepare mammalian-expression plasmid constructs of all four (full length) isoforms of LEDGF and additionally ΔPWWP domain mutants of the two main isoforms LEDGF/p75 and p52.
2. To establish stable HEK293 cell lines each expressing a unique tagged protein.
3. To optimise methods for the future purification of the core-complex of each expressed isoform.

5.2. Methods

5.2.1. Construction of LEDGF Expression Plasmids

cDNA constructs for LEDGF/ p75 (cloned into pDNR-DUAL) and p52 (cloned into pCMV-Sport6) were purchased from OriGene Technologies and Open Biosystems respectively, and transformed into competent *E.coli* DH5α cells (Life Technologies, UK); and plasmid DNA prepared from 6 clones using the mini-prep procedure outlined in Chapters 2.2.7 and 2.2.8. The integrity of each isoform was confirmed by restriction-enzyme digestion of 200 ng of plasmid with *Not*I and *Sal*I (pCMV-Sport6-LEDGF/p52) or *Sal*I and *Hind*III (pDNR-DUAL-LEDGF/p75) (all restriction enzymes from New England Biolabs, UK). Digested plasmid was analysed by agarose gel electrophoresis as described in chapter 2.2.4, and the sequences of positive clones verified (see chapter 2.2.10).

In order to generate tagged cDNA clones for all isoforms identified in this study, PCR primers were designed to amplify the complete mRNA sequence of each variant isoform. In each case, the forward primer incorporated (from 5´ to 3´) an *EcoR*I restriction site, the sequence for a FLAG, HA or Myc tag and then specific sequence for each isoform whilst the reverse primers, in all cases, contained an *Xba*I restriction site (Figure 5-1). A stop codon was not included in the reverse primer. Primer sequences are shown in Table 5-1.

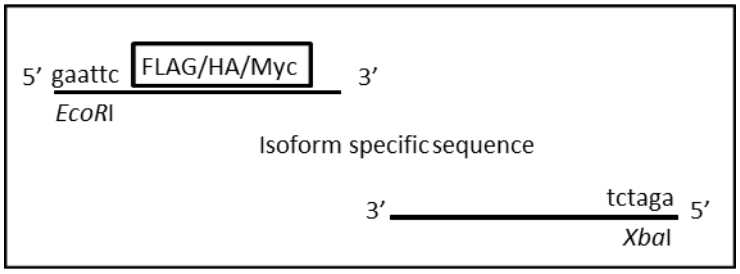


Figure 5-1: Organisation of primer sequences used to generate tagged cDNA clones of each isoform.

The forward primer is composed of an E*coR*I restriction site, a tag then the sequence specific to the N-terminus of the isoform that is to be amplified. The reverse primer is composed of a 5' X*ba*I restriction site and the 3' end of the sequence of the isoform that is to be amplified.

Primer	Primer sequence (5' > 3')
p75ecoRIHAtag	cctggaattctgtacccctacgacgtg cccgactacgccatgactcgcgatttcaaacctggag
p75ecoRIFLAGtag	cctggaattctggactacaaggac gacgacgacaagatgactcgcgatttcaaacctggag
p75notagXbaI	cctctagacgttatctagtgtagaatcc
p52notagXbaI-new	cctctagacctgtagattacatgttgtttggtgc
p52bc-mycF	cctggaattctggaacaaaaacttatttctgaa gaagatctgatgactcgcgatttcaaacctggag
p52bLinkerPrimerNEW	gaggaacaaatggaaactgagcacat taatccagttactgaaaagag
p52bmiddleR	tgctcagtttccatttgttcctcttgc
p52bRXbaI	gcctctagacgtccaatgagtctgtatcaagatc
p52bE6R	cctctagacccattttgggtctgcctcttgg
PWWPmutEcoMYC	cctggaattctggaacaaaaacttatttctgaagaagatctgatgag tcaacaggcagcaactaaacaatc

Table 5-1: Primer sequences used to generate tagged cDNA clones of each isoform.

All forward primers incorporate an E*coR*I restriction site, a tag and then the specific isoform sequence. All reverse primes incorporate the sequence specific to that isoform and an X*ba*I restriction site.

5.2.2. Generation of pcDNA3.1His-HA-LEDGF/p75 and pcDNA3.1His-FLAG-LEDGF/p75

Plasmids, pcDNA3.1His-HA-LEDGF/p75 (pcDNA3.1His mammalian expression vector + tag + isoform of LEDGF) and pcDNA3.1His-FLAG-LEDGF/p75, were generated by PCR amplification using pDNR-DUAL-LEDGF/p75 as template and primer pairs p75ecoRIFLAGtag/p75notagXbaI and p75ecoRIHAtag/p75notagXbaI (Table 5-1) (annealing temperature gradients of 66-68 °C and 59-69 °C respectively). Each PCR reaction contained 5 µL of 10x Expand High Fidelity buffer (Roche, UK), 1 µL PCR grade nucleotide mix (Roche, UK), 0.3 µM of each primer, 0.1 ng template and 0.75 µL Expand High Fidelity enzyme mix (Roche, UK), in a total reaction volume of 50 µL. Cycling parameters were as follows:

94 °C 2 minutes (x1)

94 °C 15 seconds, annealing temperature for 30 seconds, 72 °C for 1 minute (x10)

94°C 15 seconds, annealing temperature for 30 seconds, 72 °C for 2 minutes (x20)

72 °C for 7 minutes (x1)

Hold at 4 °C

Products generated at the chosen annealing temperature were excised from the gel and further purified using the Wizard SV Gel and PCR Clean-Up Kit, ligated into pGEM-T and transformed into competent JM109 cells as described in Chapters 2.2.4- 2.2.7. Plasmid DNA was purified from selected clones using the mini-prep procedure described in Chapter 2.2.8 and clone sequences verified (Chapter 2.2.10).

Tagged cDNA inserts were released by digesting two micrograms of purified plasmid (pGEMT-HA-LEDGF/p75 or pGEMT-FLAG-LEDGF/p75) with *Xba*I and *EcoR*I (New England Biolabs, UK). Empty Plasmid pcDNA3.1HisA was digested in the same manner and then dephosphorylated by treatment with 1 µL calf intestinal alkaline phosphatase (CIP) (Promega, UK) for 15 minutes at 37 °C. Digested DNA was separated by agarose gel electrophoresis (Chapter 2.2.4). Bands corresponding to the empty pcDNA3.1HisA (Life Technologies, UK), or tagged p75 PCR products were purified using the Wizard SV gel and PCR Clean Up Kit (Chapter 2.2.5). Purified products were ligated into linearised pcDNA3.1HisA at a vector to insert ratio of 1:5 using the Rapid Ligation Kit (Roche, UK). Briefly, the volume of vector plus insert was made up to 10 µL with 1x DNA dilution buffer. To this, 10 µL T4 DNA ligation buffer and 1 µL T4 DNA ligase were added. After vortexing the reaction was incubated for 30 minutes at RT. Ligations were transformed using Max Efficiency DH5α cells as described in Chapter 2.2.7.

5.2.3. Generation of pcDNA3.1His-FLAG-LEDGF/p52 and pcDNA3.1His-HA-LEDGF/p52

To avoid additional rounds of PCR amplification, HA and FLAG tagged constructs of LEDGF/p52 were generated by replacing the C-terminii of pcDNA3.1His-HA-LEDGF/p75 and pcDNA3.1His-FLAG-LEDGF/p75 with the C-terminus of p52. Two micrograms of plasmids pcDNA3.1His-HA-LEDGF/p75 and pcDNA3.1His-FLAG-LEDGF/p75 were digested with *Xba*I and *Xho*I (New England Biolabs, UK) then dephosphorylated by incubating with 1 µL CIP at 37 °C for 15 minutes. pCMV-SPORT-LEDGF/p52 was digested in the same manner but was not dephosphorylated. Digested DNA was separated by agarose gel electrophoresis (Chapter 2.2.4) and bands corresponding to the C-terminal *Xho*I-*Xba*I fragment of LEDGF/p52, pcDNA3.1His-FLAG-LEDGF/p75-*Xho*I-*Xba*I and pcDNA3.1His-HA-LEDGF/p75-*Xho*I-*Xba*I were excised and purified using the Wizard SV Gel and PCR Clean up Kit as detailed in (Chapter 2.2.5).

Two separate ligation reactions were performed as described in Chapter 5.2.2 which contained the C-terminal *Xho*I-*Xba*I fragment of LEDGF/p52 together with either pcDNA3.1His-FLAG-LEDGF/p75-*Xho*I-*Xba*I or pcDNA3.1His-HA-LEDGF/p75-*Xho*I-*Xba*I. Ligation products were transformed into Max Efficiency DH5α cells as described in Chapter 2.2.7.

5.2.4. Generation of pcDNA3.1His-MYC-LEDGF/p75ΔPWWP and pcDNA3.1His-MYC-LEDGF/p52ΔPWWP

Plasmids pcDNA3.1His-FLAG-LEDGF/p75 and pcDNA3.1His-FLAG-LEDGF/p52 were used as templates in a PCR reaction using primer pairs PWWPmutEcoMYC/p75notagXbaI and PWWPmutEcoMYC/p52notagXbaI-new (Table 5-1). Resulting amplicons were analysed by agarose gel electrophoresis (Chapter 2.2.4) and bands of interest purified using the Wizard SV Gel and PCR Clean Up Kit as described in Chapter 2.2.5. PCR products were ligated into pGEM-T vector to give pGEM-T-MYC-LEDGF/p75ΔPWWP and pGEM-T-MYC-LEDGF/p52ΔPWWP and transformed into JM109 competent cells (see sections 2.2.6 and 2.2.7). Plasmid DNA was purified using the mini-prep procedure and sequences verified externally at Source Bioscience Ltd (Chapter 2.2.10).

The resulting plasmids were digested with *Xba*I and *EcoR*I and separated by agarose gel electrophoresis. Bands of interest were purified (Chapter 2.2.5), ligated into pcDNA3.1HisA using the Rapid DNA Ligation Kit and transformed into Max Efficiency DH5α cells as described (Chapter 2.2.7).

5.2.5. Generation of pcDNA3.1His-MYC-LEDGF/p52b

Dual tagged LEDGF/p52b was generated by PCR amplification of the N- and C-terminal regions as 2 separate fragments which had complementary 3’ and 5’ sequences respectively, which were then joined together (Figure 5-2). Thus, the N-terminal region of LEDGF (amino acids 1-325) was generated by PCR amplification of

pDNR-DUAL-LEDGF/p75 using primer pair p52bc-mycF and p52bmiddleR (Table 5-1) whilst the C-terminal region of LEDGF/p52b was amplified with primer pair p52bLinkerPrimerNEW and p52bRXbaI (Table 5-1) using pcDNA3.1His-HA-LEDGF/p52 as template. PCR reactions were analysed by agarose gel electrophoresis (Chapter 2.2.4) and the bands of interest purified as described (Chapter 2.2.5).

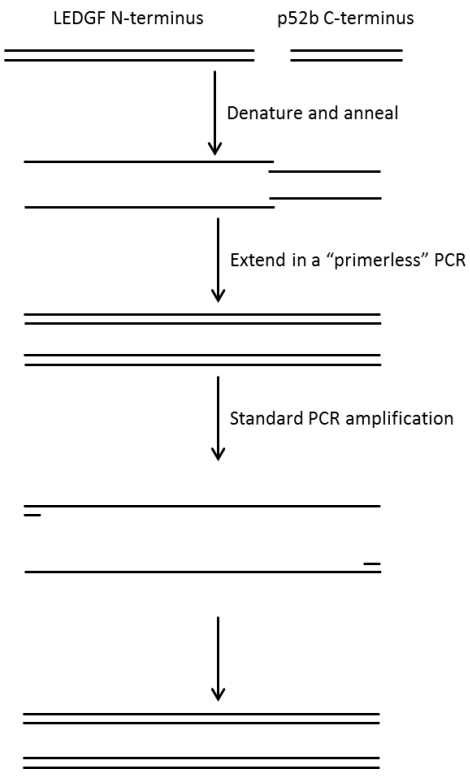


Figure 5-2: Diagram to show the amplification strategy used to generate cDNA for LEDGF/p52b.

Two separate PCR products (LEDGF N-terminus and p52b C-terminus) were joined together to give the full length tagged LEDGF/p52b isoform. The method first involved amplification in the absence of primers and then a standard RT-PCR amplification cycle.

The two products were then joined together to generate full length LEDGF/p52b by use of a "primer-less" PCR reaction (Figure 5-2). Thus, 5 µL each PCR product, 4 µL 10x Expand High Fidelity buffer, 1 µL nucleotide mix and 0.75 µL Expand High

Fidelity Enzyme mix were added to a tube and the volume made up to 40 µL with NFW. This reaction mix was subject to 5 cycles using the following parameters:

94 °C 2 minutes (x1)

94 °C 15 seconds, 40 °C 30 seconds, 72 °C 1 minute (x5)

Following this, 2.5 µL each of primers p52bc-mycF and p52bRXbaI (Table 5-1) (stock concentration 6 µM), 1 µL 10x Expand High Fidelity buffer and 4 µL NFW were added to the tube and the reaction mix amplified using the following PCR parameters:

94 °C 1 minute (x1)

94 °C 15 seconds, 40 °C 30 seconds, 72 °C 1 minute (x10)

94 °C 15 seconds, 45 °C 30 seconds, 72 °C 2 minutes 40 seconds (x20)

72 °C 7 minutes (x1)

Resulting amplicons were analysed by agarose gel electrophoresis and bands of interest purified and ligated into pGEM-T vector (Chapters 2.2.4-2.2.6) to give pGEM-T-MYC-LEDGF/p52b. The insert was released by digestion with *Xba*I and

*EcoR*I and separated by agarose gel electrophoresis (Chapter 2.2.4) and purified (Chapter 2.2.5). The purified band was ligated into pcDNA3.1HisA (enzyme digested and CIP prepared as described previously) using the Rapid DNA Ligation Kit. Ligations were transformed into Max Efficiency DH5α cells according to Chapter 2.2.7.

5.2.6. Generation of pcDNA3.1His-MYC-LEDGF/p52bΔE6

Primers p52bc-mycF and p52bE6R (Table 5-1) were used in a PCR reaction with an annealing temperature gradient as previously described using HEK293 cDNA as template. PCR products were purified by agarose gel electrophoresis, ligated into pGEM-T and transformed into *E.*coli as described previously (Chapters' 2.2.4-2.2.7) to generate pGEM-T-MYC-LEDGF/p52bΔE6. The resulting plasmid was digested with *Xba*I and *EcoR*I and separated by agarose gel electrophoresis (Chapter 2.2.4). Bands of interest were purified, ligated into linearised pcDNA3.1-HisA and transformed into Max Efficiency DH5α cells as described earlier.

5.2.7. Optimisation of transfection conditions

The optimum conditions for transfection were determined using the plasmid pcDNA3.1His-HA-LEDGF/p75 and 293T cells using a matrix (Figure 5-3) and the protocol outlined below.

Adequate plasmid DNA was obtained using a midi-prep kit as described in Chapter 2.2.9. Adherent 293T cells were passaged and plated as described in Chapter 2.1.3. Twenty four hours prior to transfection, two 6-well plates were seeded with 293T cells at the required densities shown in Figure 5-3 in 2 mL complete DMEM per well and incubated overnight at 37 °C and 5% CO_2 in a humidified incubator. The next day, for each transfection, 10 µL Lipofectin (Life Technologies, UK) was mixed with 90 µL serum-free DMEM media in a sterile microfuge tube and incubated at RT for 30-45 minutes.

1 µg	1 x 105 v.c/well	1.5 x 105 v.c/well	2 x 105 v.c/well
1.5 µg	1 x 105 v.c/well	1.5 x 105 v.c/well	2 x 105 v.c/well
2 µg	1 x 105 v.c/well	1.5 x 105 v.c/well	2 x 105 v.c/well
	2 x 105 v.c/well, pcDNA3.1His	2 x 105 v.c/well Lipofectin only	2 x 105 v.c/well Wild type

Figure 5-3: Matrix of different conditions to be used to determine the optimum viable cell density and amount of DNA for transfection using Lipofectin.

For the investigation, two 6-well plates were plated with 293T cells at the density stated in each well. Nine wells were then transfected with pcDNA3.1His-HA-LEDGF/p75 at the concentration detailed on the left of each row. Two other wells were transfected with either empty vector or the Lipofectin, the transfection reagent. The final well was a wild type control. The cells were incubated for a total of 48 hours.

The relevant amount of plasmid (according to Figure 5-3) was added and the volume made up to 100 µL with serum-free DMEM media. The DNA was mixed with the Lipofectin/DMEM and incubated for 15 minutes at RT. Media was removed from the cells and the adherent cells washed with 2 mL serum free DMEM. This was discarded, replaced with 800 µL of serum free DMEM media per well and the appropriate lipofectin/DNA mixture added (according to Figure 5-3). The plates were gently rocked to ensure the cell monolayer was evenly covered and transferred overnight to a humidified incubator (37 °C and 5% CO_2). The following morning the media was discarded and 2 mL complete DMEM media was added. The plate was returned to the incubator for a further 24 hours.

Forty-eight hours post-transfection, the adherent cells were harvested by Trypsin-EDTA digestion. Cell suspensions were collected and pelleted for 5 minutes at 500 x g, RT. Supernatant was discarded and the pellet washed in 1 mL PBS. Cells pellets were resuspended in 0.5 mL lysis buffer (50 mM NaH_2PO_4 (Sigma Aldrich, UK), 300 mM NaCl (Fisher Scientific, UK), 10 mM imidazole (Sigma Aldrich, UK), 0.05% Tween 20, pH 8) and 5 µL protease inhibitor cocktail (Merck Millipore, UK) and stored overnight at -80 °C. The following day, the cells were lysed by freeze/thawing (-80 °C/RT) three times followed by sonication on ice (continued until clear) and clarified by centrifugation at 10000 x g, 4 °C for 10 minutes. The supernatant was transferred to a fresh 1.5 mL microcentrifuge tube and stored at -20 °C until required.

The protein concentration of the lysates was determined using the RC DC Protein Determination Kit (Bio-Rad). Protein standards were prepared by dissolving BSA in

lysis buffer to yield 1.5, 1.0, 0.5 and 0.2 mg/mL. Briefly, 25 µL of each sample (cell lysate or protein standard) was placed in a 1.5 mL microcentrifuge tube and mixed with 125 µL Reagent I. The tubes were incubated at RT for 1 minute, and then 125 µL Reagent II added to each and mixed. After centrifugation for 3 minutes at 15000 x g at RT the supernatant was discarded. One hundred and twenty seven microliters of Reagent A' (prepared by adding 5 µL Reagent S to 250 µl Reagent A) was added to each tube and contents vortexed. The tubes were incubated for 5 minutes at RT before the addition of 1 mL Reagent B. Samples were then incubated for 15 minutes at RT and the absorbance at 750 nm determined using the NanoDrop 2000c.

Five microgram of protein from each transfection was analysed by Western Blotting with an anti-HA antibody (Covance, Mouse monoclonal, 1:1000) followed by a Goat-anti-mouse IgG-HRP (Santa Cruz, 1:5000) as described in Chapter 2.3.2. To assess equal loading, the membrane was probed with anti-β-actin antibody (Sigma Aldrich, Mouse monoclonal, 1:10000). Following visual capture of the gel, the blots were treated with 10 mL stripping buffer (1x TBS (Sigma Aldrich, UK), 2% SDS, 0.1 M β-mercaptoethanol) for 10 minutes at 55 °C. Membranes were washed 3 x 5 minutes in 1x PBS-Tween (0.05%) then probed with an anti-LEDGF primary antibody (BD Biosciences, Mouse monoclonal, 1:4000) followed by Goat-anti-mouse IgG-HRP secondary (Santa Cruz, 1:5000).

5.2.8. Generation of cell lines stably expressing different isoforms of LEDGF

Using optimal transfection conditions determined above (Chapter 5.2.7), stable HEK293 cell lines expressing the described LEDGF isoforms were generated. Forty-eight hours post-transfection, cells were selected with geneticin. The media in the wells was discarded and replaced with 2 mL of selective media (complete DMEM containing 400 µg/mL G418 (Merck Millipore, UK)) and the plates returned to the incubator. Every 2-3 days, selective media was removed and replaced with fresh selective media with antibiotic. When at a suitable density, the persistent cells were transferred to a T-25 cm^2 vented tissue culture flask and passaged (see Chapter 2.1.3) until there were sufficient cells to yield 5 vials for storage at -150 °C as detailed in Chapter 2.1.4.

5.2.9. Small scale purification of His tagged protein

A total of $1x10^7$ viable cells were lysed, sonicated and clarified as described in section 5.2.7. Thirty microlitres of Ni-NTA magnetic agarose beads (Qiagen, UK) were added to each tube and incubated overnight at 4 °C on a rotator. The following morning the tubes were centrifuged to ensure that the beads were at the bottom of the tube and then placed on a magnetic separator for one minute. The supernatant was discarded and after 3 further washes beads were suspended in wash buffer (50 mM NaH_2PO_4, 300 mM NaCl, 20 mM imidazole, 0.05% Tween 20, pH 8) and the tubes placed on the magnetic separator. Following a final wash,

bound His-tagged protein was eluted by incubating the beads with 50 µL elution buffer (50 mM NaH_2PO_4, 300 mM NaCl, 250 mM imidazole, 0.05% Tween 20, pH 8) for two minutes. The tubes were then placed on a magnetic separator for one minute and the elution transferred to a fresh tube and stored at -20 °C.

5.2.10. Large scale purification of His tagged proteins

Cell pellets (equivalent to $1x10^8$ viable cells) were resuspended in 0.5 mL lysis buffer (50 mM $NaHPO_4$, 300 mM NaCl, 10 mM imidazole, 0.05% Tween 20, pH 7.4), lysed, sonicated and clarified as described in Chapter 5.2.7. Supernatants from specific isoforms were pooled and transferred to a 25 mL universal tube and stored on wet ice. A small aliquot was sequestered to serve as non-purified control. Sample loading, washing and elution was continually monitored at 280 nm.

A 1 mL HisTrap HP column (GE Healthcare, UK) (connected to a peristaltic pump (P-1, Pharmacia LKB, Sweden) and UV monitor (Econosystem, Bio-Rad, UK)) was washed at 2 mL/min with elution buffer (50 mM $NaHPO_4$, 300 mM NaCl, 250 mM imidazole, 0.05% Tween 20, pH 7.4). The column was then equilibrated in lysis buffer until UV traces returned to baseline and stabilised, after which clarified lysate was loaded at a flow rate of 0.2 mL/min and the flow-through collected in a 25 mL universal tube held on ice. The column was washed with lysis buffer at a flow rate of 1 mL/min until UV traces returned to baseline. This was followed by a further wash (50 mM $NaHPO_4$, 300 mM NaCl, 50 mM imidazole, 0.05% Tween 20, pH 7.4) at a

flow rate of 0.5 mL/min. Wash buffer elutions were collected (2 mL for the first fraction and 1 mL for subsequent fractions) until UV traces returned to baseline. Bound proteins were then released by applying elution buffer (250 mM $NaHPO_4$, 300 mM NaCl, 250 mM imidazole, 0.05% Tween 20, pH 7.4) at a rate of 0.5 mL/min and 2 mL fractions were collected until UV traces returned to baseline. All samples were immediately transferred to -20 °C.

5.2.11. Immunoprecipitation of Nickel purified protein

To 250 µL of sample, 250 µL no-salt 1% Triton-X100 buffer (20 mM Tris-HCl pH 7.4, 1% Triton-X100, (Sigma Aldrich, UK)) plus 5 µL protease inhibitor cocktail, was added. Immunoprecipitation was carried out using anti-FLAG M2 Magnetic beads (Sigma Aldrich, UK). One hundred microlitres of the bead suspension (50 µL packed gel volume) was transferred to a 1.5 mL microfuge tube. The tube was placed on a magnetic separator for 1 minute and the supernatant removed and discarded. The beads were washed twice with 500 µL 1x TBS. To the washed beads, 500 µL of clarified lysate was added. To one set of beads, 500 µL Triton-X 100 buffer (supplemented with NaCl to give 150 mM) was added which acted as a 'beads only' negative control. The tubes were incubated overnight at 4 °C on a rotator. Following incubation, the tubes were placed on a magnetic separator for one minute. The supernatant was discarded and the beads washed three times with 1 mL 1x TBS. After the final wash, 40 µL 2x SDS loading buffer (125 mM Tris HCl pH 6.8, 4% SDS, 20% (v/v) glycerol and bromophenol blue) was added to each tube and the tubes

boiled for 3 minutes at 95 °C. After this, the tubes were placed on a magnetic separator for one minute and the supernatant transferred to a clean 1.5 mL centrifuge tube.

5.2.12. Coomassie Blue staining of acrylamide gels

Following electrophoresis, the gel was transferred to a clean plastic container. 100 mL of Coomassie Blue stain (0.25% CBB R250 (Sigma Aldrich, UK), 40% methanol and 10% glacial acetic acid (Fisher Scientific, UK)) was added and the gel incubated for one hour with gentle agitation at RT. The stain was then discarded and the gel incubated in de-stain buffer (30% methanol and 10% glacial acetic acid) with gentle agitation. The buffer was changed regularly until the bands reached a suitable intensity after which the gel was stored in ultrapure water.

5.2.13. Silver staining of acrylamide gels

Following electrophoresis, the gel was rinsed in ultrapure water then stained using the SilverQuest Silver Staining Kit (Life Technologies, UK). The water was discarded and the gel incubated in 100 mL fixative (40% ethanol, 10% acetic acid made with ultrapure water) for 1 hour followed by washing with 100 mL of 30% ethanol for 10 minutes with gentle agitation. The gel was then incubated for 10 minutes in 100 mL Sensitizing solution (10% sensitizer solution, 30% ethanol, in ultrapure water). The Sensitizer solution was discarded and the gel incubated in 100 mL 30% ethanol for

10 minutes followed by 10 minutes incubation in 100 mL ultrapure water. The water was discarded and replaced with 100 mL Staining Solution (1% Staining solution made up with ultrapure water). Following a 15 minute incubation the solution was decanted and the gel washed again with ultrapure water for 20-60 seconds. The gel was then incubated in 100 mL developing solution until the bands reached an acceptable intensity, at which point 10 mL Stopper solution was added and incubated for 10 minutes. Following this, the solution was discarded and the gel washed in 100 mL ultrapure water for 10 minutes. The gel was stored in fresh ultrapure water.

A summary of the experimental work performed in this chapter is detailed in Figure 5-4.

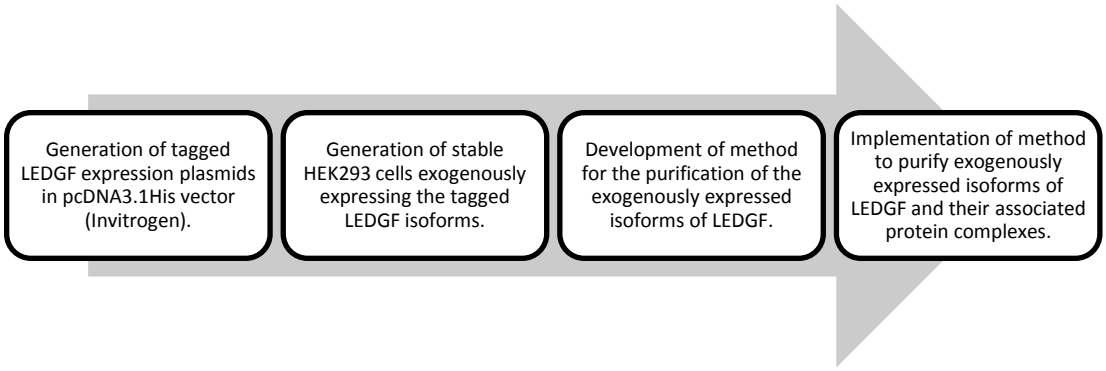


Figure 5-4: Summary of work described in this chapter.

The workflow summarises the steps taken to purify out the protein complex around each of the tagged isoforms of LEDGF.

5.3. Results

5.3.1. Verification of LEDGF Expression Plasmids

The integrity of commercially purchased clones of LEDGF/p52 and p75 was confirmed by restriction digestion and agarose gel electrophoresis. Briefly, pCMV-SPORT6-LEDGF/p52 was digested with *Not*I and *Sal*I, whilst pDNR-DUAL-LEDGF/p75 was digested with *Sal*I and *Hind*III. As seen in Figure 5-5, bands of 1593 bp and 1842 bp , corresponding to the p75 and p52 inserts were released by digestion, verifying the identity of the clones. Sanger sequencing confirmed clone integrity (data not shown).

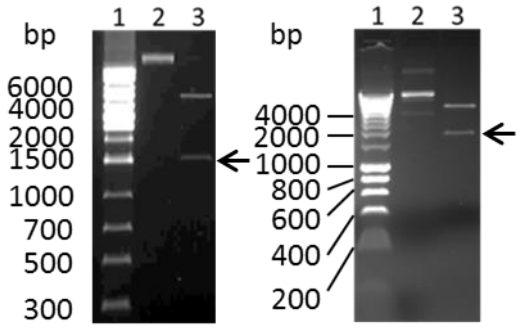


Figure 5-5: Restriction enzyme digestion of pDNR-DUAL-LEDGF/p75 and pCMV-SPORT6-LEDGF/p52.

200 ng pDNR-DUAL-LEDGF/p75 (left) or pCMV-SPORT6-LEDGF/p52 (right) was digested with either *Sal*I/*Hind*III or *Not*I/*Sal*I respectively. Resulting digests were analysed by agarose gel electrophoresis. Lane 1: molecular weight markers; Lane 2: undigested plasmid; Lane 3: digested plasmid. Arrows indicate the relevant inserts released following the double digestion.

5.3.2. Construction of tagged LEDGF cDNA expression clones

Using pDNR-DUAL-LEDGF/p75 as template, a PCR reaction was performed using primer pairs p75ecoRIHAtag/p75notagXbaI and p75ecoRIFLAGtag/ p75notagXbaI (Table 5-1) at different annealing temperatures. PCR reactions were analysed by agarose gel electrophoresis, and products at the chosen annealing temperature (determined to be 68 °C and 64 °C respectively; Figure 5-6) were purified.

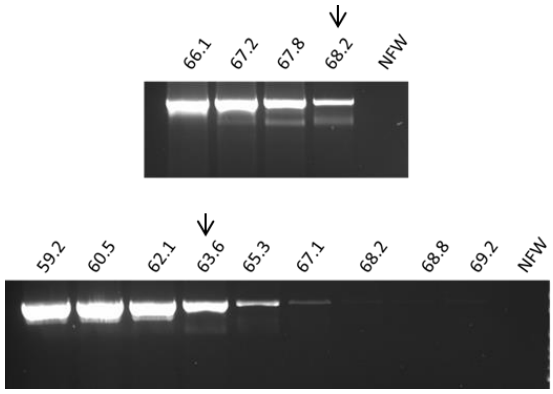


Figure 5-6: PCR amplification of the 75 kDa isoform of LEDGF.

PCR reactions were performed using pDNR-DUAL-LEDGF/p75 and primer pairs p75ecoRIHAtag and p75notagXbaI (upper panel) or p75ecoRIFLAGtag and p75notagXbaI (lower panel) with a range of annealing temperatures. The products were analysed by agarose gel electrophoresis and the annealing temperature determined visually (indicated by arrows).

The purified PCR products were cloned into pGEM-T to generate pGEM-T-HA-LEDGF/p75 and pGEM-T-FLAG-LEDGF/p75. Positive clones were subsequently digested with *Xba*I and *Eco*RI and the insert transferred into linearised pcDNA3.1HisA (digested with the same enzymes) to yield pcDNA3.1His-HA-LEDGF/p75 and pcDNA3.1His-FLAG-LEDGF/p75.

Expression constructs for LEDGF/p52 were generated by replacing the C-terminus of LEDGF in plasmids pcDNA3.1His-HA-LEDGF/p75 and pcDNA3.1His-FLAG-LEDGF/p75 with that of p52 (from plasmid pCMV-SPORT6-LEDGF/p52 as described in Chapter 5.2.3). Restriction digests were analysed by agarose gel electrophoresis (Figure 5-7).

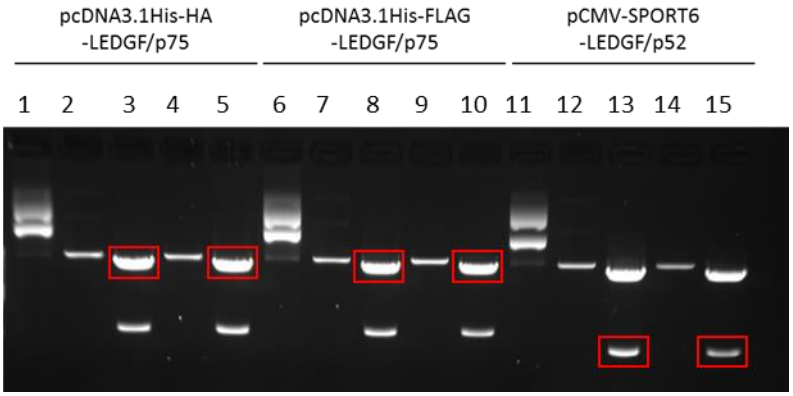


Figure 5-7: Restriction enzyme digestion of pcDNA3.1His-HA-LEDGF/p75, pcDNA3.1His-FLAG-LEDGF/p75 and pCMV-SPORT6-LEDGF/p52.

The plasmids were digested with X*ba*I and then X*ho*I in sequential reactions. At each stage, a sample was retained. Digests were analysed by agarose gel electrophoresis to follow the digestion and release of inserts. The plasmids digested are detailed at the top of the image. For each plasmid there is an undigested sample (lanes 1, 6 and 11), a single digested with *Xba*I (lanes 2, 4, 7, 9, 12 and 14) and a double digested with *Xba*I and *Xho*I (lanes 3, 5, 8, 10, 13 and 15) from constructs. Highlighted bands were excised from the gel.

Each digestion was performed in duplicate to ensure that sufficient vector and insert was recovered. For each construct that was double digested, the release of the insert and change in mobility of the larger band suggested the digestion was successful. Relevant bands were purified, and the LEDGF/p52 *Xho*I/*Xba*I fragment was ligated into the pcDNA3.1His-FLAG-LEDGF/p75-*Xho*I and pcDNA3.1His-HA-LEDGF/p75-*Xho*I vector backbones separately as described (Chapter 5.2.3) to form pcDNA3.1His-HA-LEDGF/p52 and pcDNA3.1His-FLAG-LEDGF/p52 respectively.

In order to study the role of the PWWP domain of LEDGF, expression constructs for the LEDGF/p75 and p52 isoforms were generated in which the PWWP domain was deleted. PCR primers that would introduce the deletion were synthesised (Table 5-1) and used in amplification reactions using pcDNA3.1His-FLAG-LEDGF/p75 and pcDNA3.1His-FLAG-LEDGF/p52 as templates. Reactions were performed using a range of annealing temperatures and the resulting amplicons analysed by agarose gel electrophoresis (Figure 5-8).

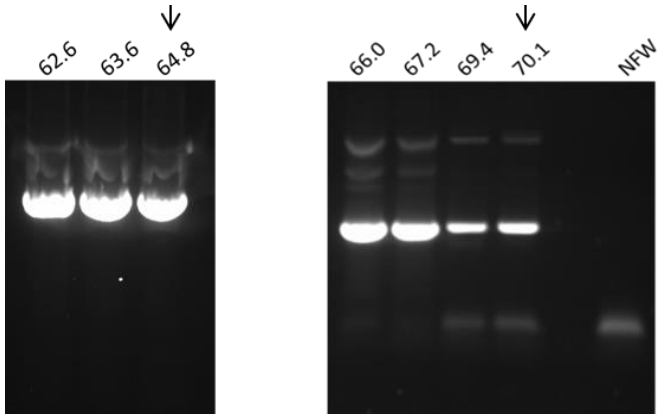


Figure 5-8: PCR amplifications of MYC-LEDGF/p75ΔPWWP (left panel) and MYC-LEDGF/p52ΔPWWP (right panel).

PCR reactions were performed using pcDNA3.1His-FLAG-LEDGF/p75 and pcDNA3.1His-FLAG-LEDGF/p52 and primers PWWPmutEcoMYC and p75notagXbaI and PWWPmutEcoMYC and p52notagXbaI-new with a range of annealing temperatures. PCR products were analysed by agarose gel electrophoresis and the chosen annealing temperature determined visually and indicated by the arrows.

Products generated at the chosen annealing temperature (determined to be 70 °C for primers PWWPmutEcoMYC/p52notagXbaI-new and 65 °C for PWWPmutEcoMYC/p75notagXbaI; Figure 5-8) were excised from the gel and purified. The PCR products were first cloned into pGEM-T as described in Chapter 5.2.4 and subsequently transferred to pcDNA3.1HisA to generate pcDNA3.1His-MYC-LEDGF/p75ΔPWWP and pcDNA3.1His-MYC-LEDGF/p52ΔPWWP.

To prepare pcDNA3.1His-MYC-LEDGF/p52b, a three-step PCR method was used. Using pDNR-DUAL-LEDGF/p75 as template, the N-terminal region of LEDGF corresponding to amino acids 1-325 was amplified in a PCR reaction using primers p52bc-mycF and p52bmiddleR (Table 5-1). Resulting amplicons were analysed and the band at the chosen annealing temperature (shown to be 70 °C; Figure 5-9) was purified.

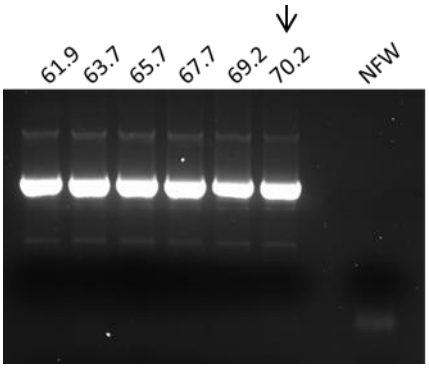


Figure 5-9: Optimisation of PCR annealing temperature.

PCR reactions were performed using p52bc-mycF and p52bmiddleR with a range of annealing temperatures. PCR products were analysed by agarose gel electrophoresis and the chosen annealing temperature determined visually and indicated by arrows. Specifically, the annealing temperature chosen was 70 °C.

The C-terminal region of LEDGF/p52b was amplified in a separate reaction with primers p52bLinkerPrimerNEW and p52bRXbaI (Table 5-1) using plasmid pcDNA3.1His-HA-LEDGF/p52 as template. The resulting amplicon was purified from an agarose gel (data not shown). The 2 amplicons have complementary 5´ and 3´ regions, such that following a denaturation step, they can anneal together. The annealed amplicons were allowed to extend in a “primerless” PCR (see Figure 5-2) which enriched for the full length cDNA for LEDGF/p52b. Following this, primers p52bc-mycF and p52bRXbaI (Table 5-1) were added and a normal PCR reaction was performed. Figure 5-10 shows the result of this PCR. The band corresponding to myc-tagged LEDGF/p52b was purified, cloned into pGEM-T then transferred to pcDNA3.1HisA to generate pcDNA3.1His-MYC-LEDGF/p52b.

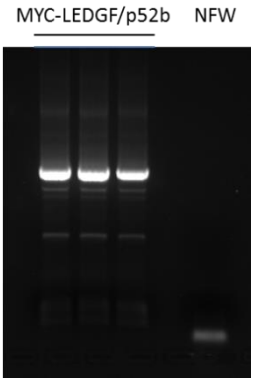


Figure 5-10: PCR amplification of full length MYC-tagged LEDGF/p52b.

The full length tagged isoform of LEDGF/p52b was generated by joining two separate PCR products representing the N-terminus and C-terminus of the isoform. The first step involved amplification in the absence of primers then the addition of primers and amplification as in a standard RT-PCR reaction.

The full length myc-tagged LEDGF/p52b was generated by the joining of the two separate N and C-terminal products using primers p52bc-mycF and p52bRXbaI. The product was run on a gel, excised and purified.

Expression plasmids for the LEDGF/p52b isoform lacking exon 6 were generated by PCR amplification of HEK293 cDNA using the primer pair p52bc-mycF and p52bE6R (Table 5-1). The optimal annealing temperature was determined to be 69 °C (Figure 5-11).

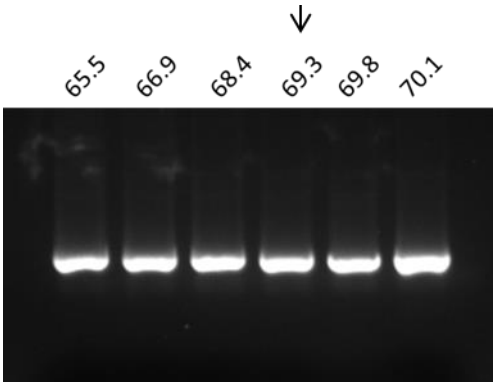


Figure 5-11: Optimisation of RT-PCR amplification of LEDGF/52bΔE6.

RT-PCR reactions were performed using p52bc-mycF and p52bE6R with a range of annealing temperatures. Products were analysed by agarose gel electrophoresis and the annealing temperature determined visually (indicated by arrows). Specifically, the chosen annealing temperature was 69 °C.

The band generated at 69 °C, the chosen annealing temperature, was excised from the gel and, by following the method described in Chapter 5.2.6, pGEM-T-MYC-LEDGF/p52bΔE6 was prepared and subsequently pcDNA3.1His-MYC-LEDGF/p52bΔE6 generated.

Plasmid maps for the constructs generated in pcDNA3.1HisA vector are shown in Appendix D.

5.3.3. Optimisation of transfection conditions

Having generated cDNA expression constructs for the 4 isoforms of LEDGF expressed in CLL, and the PWWP domain deletion mutants, the optimal transfection conditions that would yield the highest protein expression were determined using a transfection matrix (Figure 5-3) that comprised different cell numbers and amounts of DNA. Whole cell lysates were generated from transfected cells and analysed by western blotting. Results are shown in Figure 5-12.

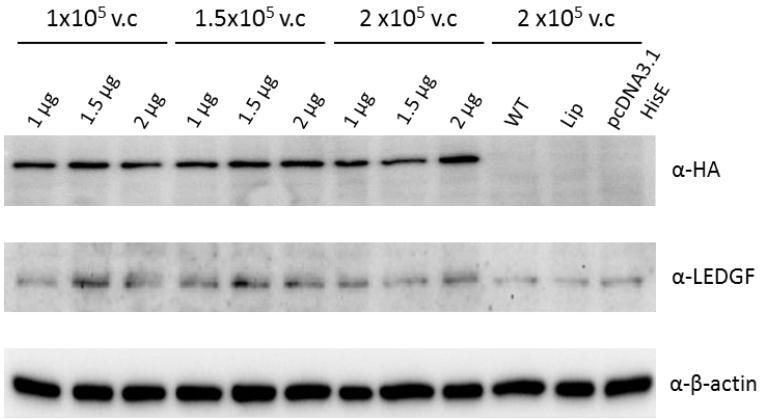


Figure 5-12: Western Blot analysis of whole cell lysates from the transfection of 293T cells at a range of viable cell densities and different DNA inputs.

Different cell densities of 293T cells were transfected with varying amounts of plasmid as per Figure 5-3. Cells were harvested 48 hours post-transfection and whole cell lysates analysed by western blotting using anti-HA (top), anti-LEDGF (middle) and anti-β-actin (bottom) antibodies. WT = wild type cells; Lip = cells treated with lipofectin but no plasmid; pcDNA3.1HisA = cells transfected with empty plasmid.

The first 9 lanes contain lysates of 293T cells all transfected with the pcDNA3.1His-HA-LEDGF/p75 plasmid whilst the last three depict controls. The presence of bands in the transfected lanes but not control lanes with anti-HA antibody demonstrates successful transfection. The blot probed with an anti-LEDGF antibody reveals endogenous LEDGF/p75 protein. In the lanes containing lysates from the cells transfected with the plasmid, there is evidence of a second slower migrating band indicating HA tagged exogenous LEDGF. The third blot was probed with anti-β-actin antibody to assess protein loading. It was deemed that a viable cell density of $1x10^5$ and 1.5 µg DNA would be most suitable for future transfections.

5.3.4. Verification of stable HEK293 cells expressing LEDGF isoforms

Having determined the optimal conditions, stable cell lines expressing each of the tagged LEDGF isoforms were generated by transfecting the relevant cDNA constructs into HEK293 cells and geneticin selection. Once stable lines were established, whole cell lysates were prepared from each and protein expression analysed by Western blotting.

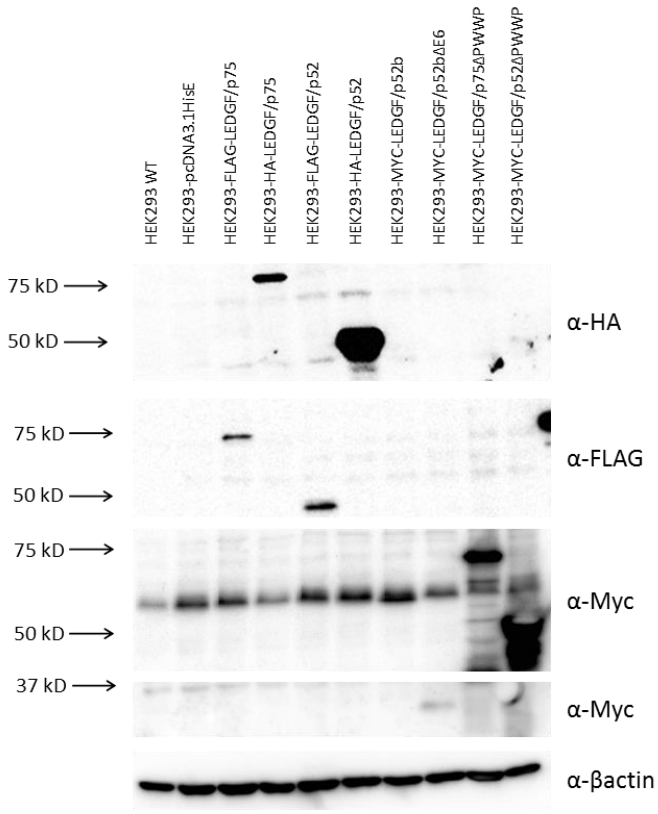


Figure 5-13: Western blot analysis of whole cell lysates of HEK293 cells stably expressing the tagged LEDGF isoforms.

10 µg of whole cell lysate from each stable cell line and wild type cells were analysed by western blotting and the membrane probed using antibodies to the tags. Namely, anti-HA (top), anti-flag (second panel) and anti-myc (third and fourth panel). Anti-β-actin was also included to show equal loading.

Equal amounts of protein from whole cell lysates were analysed using antibodies to the FLAG, Myc or HA tags (Figure 5-13). The result shows that the stable cell lines express the tagged isoforms of LEDGF. The anti-Myc antibody does identify endogenous c-Myc protein. The Myc tagged LEDGF/p52b is predicted to run at

approximately 54 kDa, and thus would run slightly faster than endogenous c-Myc. A doublet can be observed in the lane containing lysate from the HEK293 cells stably transfected with this isoform, corresponding to endogenous c-Myc (upper band) and the Myc-LEDGF/p52b (lower band). Visualisation of a band corresponding to Myc tagged LEDGF/p52bΔE6 required long exposure of the blot, and hence this portion of the blot is shown separately. In the lanes containing lysates of the ΔPWWP domain mutants, bands are seen at the predicted molecular weights.

5.3.5. Small scale purification of His-tagged LEDGF protein

Each of the exogenous LEDGF isoforms contains a plasmid encoded 6xHis tag that enables the protein to be purified using Nickel beads. Initially, this was carried out on a small-scale as detailed in Chapter 5.2.9 using lysate from cells transiently transfected with pcDNA3.1His-HA-LEDGF/p75 in preparation for a large-scale purification.

In the initial experiments, exogenous His-tagged protein was purified using bead manufacturer's instruction. Briefly, 500 µL lysate was incubated with 10 µL beads for 2 hours at 4 °C, and the beads washed and eluted as described in Chapter 5.2.9. However, results from the Western blot show that when the purified protein is compared to the input sample (10% of the crude lysate), only a small amount of protein was recovered (Figure 5-14).

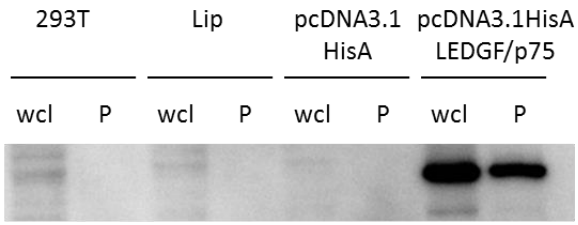


Figure 5-14: Western blot analysis of lysates from 293T cells transiently transfected with empty vector (pcDNA3.1HisA) or pcDNA3.1His-HA-LEDGF/p75 and purified with Ni-NTA beads.

Lysates from wild type (WT) cells and cells treated with Lipofectin only (Lip) served as controls. In each case, whole cell lysate (wcl) was loaded alongside the Ni-NTA beads purified material (P). The material was purified according to the manufacturers' instructions. The blot was probed with anti-HA antibody (Covance, Mouse monoclonal, 1:1000).

In an effort to maximise the amount of protein purified, the effects of increasing the volume of beads to 20 µL and incubating the beads with the lysate overnight was assessed. Cell pellets from HEK293 cells stably expressing the exogenous tagged LEDGF isoform or wild type cells were utilised. For comparison, lysates were purified with either 10 or 20 µL beads, and incubated with nickel beads for either 2 hours or overnight. Following incubation, beads were isolated by magnetic separation and the unbound lysate transferred to a clean tube. To determine if protein is lost in the washes, the first wash was retained and labelled "WE" (wash end), and the final elution from this purification was labelled "E1" (elution 1).

Proteins that remained bound to the beads were eluted by incubating the beads at 95 °C for 5 minutes in 20 µL of 2x SDS loading buffer and then placed on the magnet. The supernatant was transferred to a clean tube and labelled “beads”.

The unbound fraction, following overnight incubation, was incubated for a second time with fresh beads for 2 hours and the purification procedure repeated. The resultant elution was labelled “E2”(elution 2). The unbound fraction from this purification was incubated a third time with fresh beads and the elution labelled “E3”(elution 3). The unbound fraction at this stage was retained and labelled “End”. Aliquots of all fractions were then analysed by Western blotting and probed with an anti-FLAG antibody (Sigma Aldrich, Mouse monoclonal, 1:1000) and anti-β-actin antibody (Sigma Aldrich, Mouse monoclonal, 1:10000) (Figure 5-15).

As expected, there are no bands in the western blot of lysates from WT cells probed with the anti-FLAG antibody. Probing the blots for B-actin showed that the protein was only present in the whole cell lysate and “End” sample.

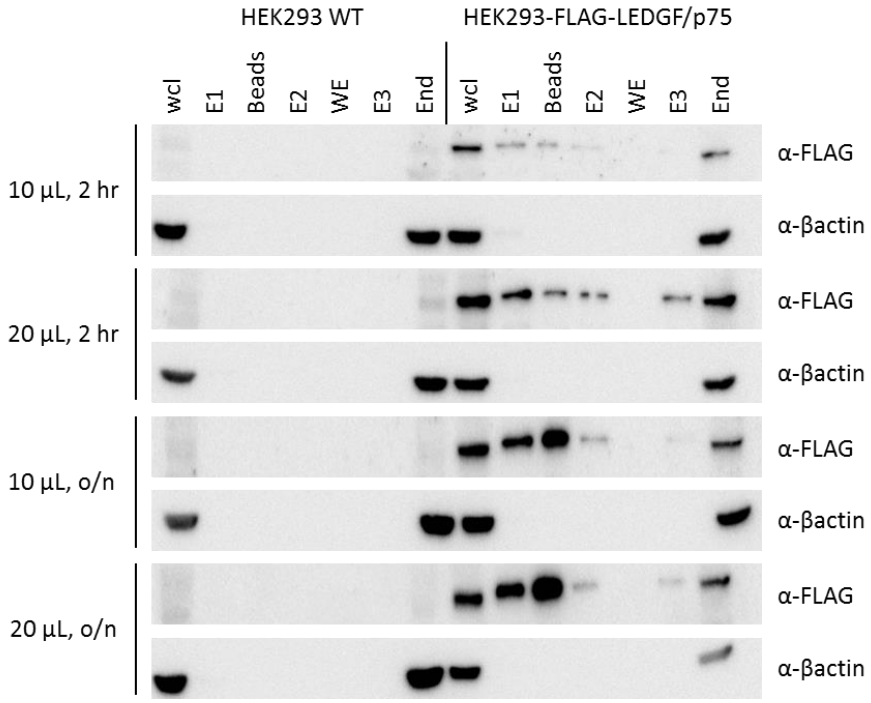


Figure 5-15: Western blot analysis of eluates from each step of the small scale purification of His-tagged FLAG-LEDGF/p75 using Ni-NTA beads depending on the volume of beads added (10 or 20 µL) and the incubation time (2 hour or overnight (o/n)).

Eluates collected at each step where analysed by western blotting and the membranes probed with anti-flag antibody. Membranes were also probed with anti-β-actin antibody as a control. WCL: whole cell lysate, E: elution from beads, WE: wash elution.

As for the transfected samples, there were no specific bands in the wash lane ("WE") showing that the over expressed protein is not eluted away during the wash procedure. Compared to the standard protocol when 10 µL beads are incubated for

2 hours with 500 µL lysate, increasing the volume of beads and/or increasing the initial incubation time resulted in a greater recovery of tagged protein ("E1"). However, this also resulted in an increased amount of protein that remained bound to the beads ("Beads"). This was particularly noticeable for overnight incubations. Furthermore, despite increasing the volume of beads, tagged protein was still present in the "End" sample.

Therefore, the method was further modified in two ways. Firstly, the length of time the beads were incubated in the elution buffer was increased, and secondly, following the initial magnetic isolation of the beads, the elution step was repeated a further two times (method summarised in Figure 5-16). Thus, wild type HEK293 and HEK293-FLAG-LEDGF/p75 cell lysates were incubated with 10 µL and 30 µL Ni-NTA Magnetic Agarose beads overnight at 4 °C on a rotator. The samples were then purified according to the method in section 5.2.9. Beads were isolated by magnetic separation and the unbound fraction transferred to a new tube. In addition, the first wash was collected and labelled "W"(wash). Bound protein was eluted from the beads and labelled E1 (1) (elution). Second and third elutions were labelled E1 (2) and E1 (3) respectively. Protein remaining bound to the beads was eluted by incubating at 95 °C for 5 minutes in 2x SDS buffer followed by magnetic separation. This elution was labelled E1BB. The unbound fraction that remained following the overnight incubation and the wash buffer supernatant ("W") was re-incubated with 10 or 30 µL beads (depending on the amount in the first overnight incubation) and incubated for 2 hours on a rotator at 4 °C. The beads were isolated from the

unbound fraction by magnetic separation and incubated in elution buffer to give fractions “E2” and “WE” respectively. Protein remaining bound to the beads was eluted by incubating at 95 °C for 5 minutes in 2x SDS buffer followed by magnetic separation to give fractions “E2BB” and “WEBB” respectively. Unbound proteins from the second incubation with beads was incubated for a third time with 10 or 30 µL beads for 2 hours on a rotator at 4 °C, and eluted protein labelled “E3”. The beads were incubated in 2x SDS buffer to elute remaining bound protein (“E3BB”). The remaining unbound fraction was labelled ‘End’. Aliquots of each sample were analysed by western blot and anti-FLAG antibody (Sigma Aldrich, Mouse monoclonal, 1:1000).

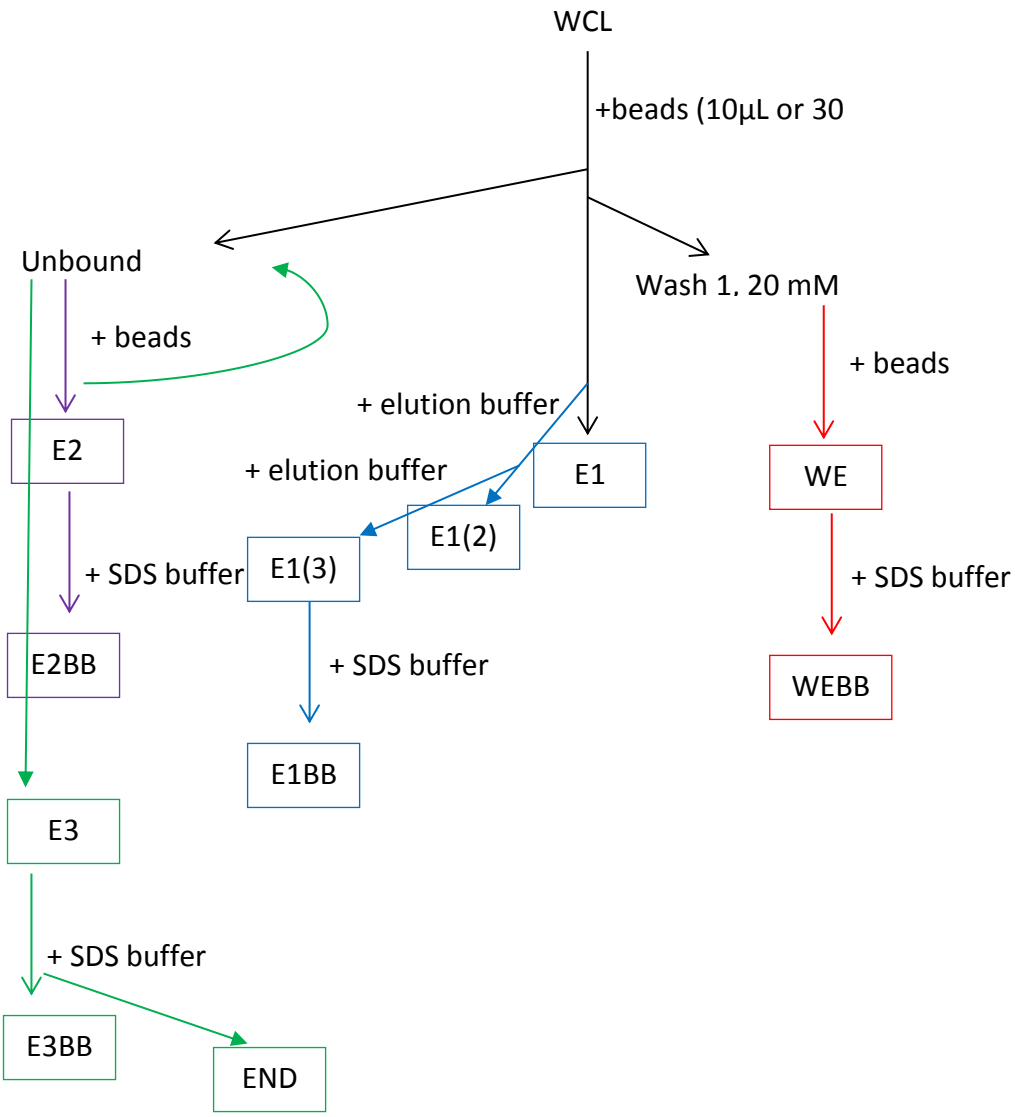


Figure 5-16: Summary of method to improve the recovery of the His-tagged protein during the small scale purification using the Ni-NTA beads.

A box indicates each sample taken to be analysed by western blotting.

As shown in Figure 5-17, and as with the previous experiment (Figure 5-15), increasing the volume of beads results in a greater yield of tagged protein (compare lanes E1 (1) between 10 and 30 µL bead volumes). However, even when the number

of elutions and the elution time was increased, tagged protein remained on the beads (lanes BB, E2BB and E3BB). This was more apparent when 30 µL of beads were used for the purification. Further, uneluted protein remained in the final supernatant following all of the isolations (lane "End").

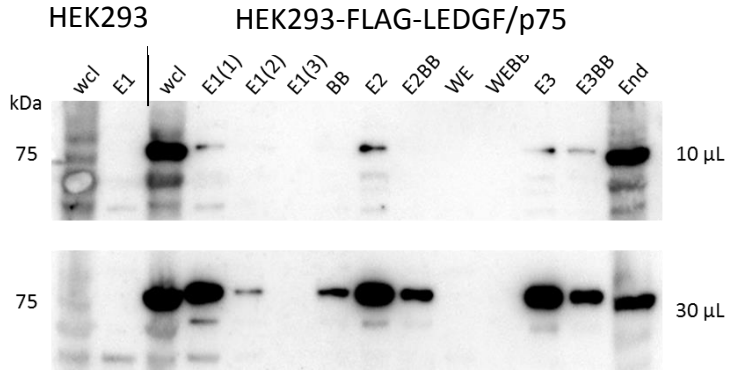


Figure 5-17: Analysis of eluates from each step of the small scale purification of His-tagged proteins using different volumes of Ni-NTA beads.

Eluates at each step of the small scale purification was collected and analysed by western blotting. Membranes were probed with α-flag antibody. Whole cell lysates (WCL) from wild type HEK293 cells and HEK293 cell exogenously expressing FLAG-LEDGF/p75 were incubated with 10 or 30 µL of Ni-NTA beads and purified as described. E: elution, BB: boiled beads, WE: wash elution.

Elutions E1-E3 from both HEK293 and HEK293-FLAG-LEDGF/p75 were separated by SDS-PAGE and the gel stained using the SilverQuest Silver Staining Kit. As seen in Figure 5-18, a band is present in elutions from the transfected cells corresponding

to FLAG-LEDGF/p75 which is not present in elutions from wild type HEK293 cells. As expected, the intensity of the band decreases with each subsequent elution.

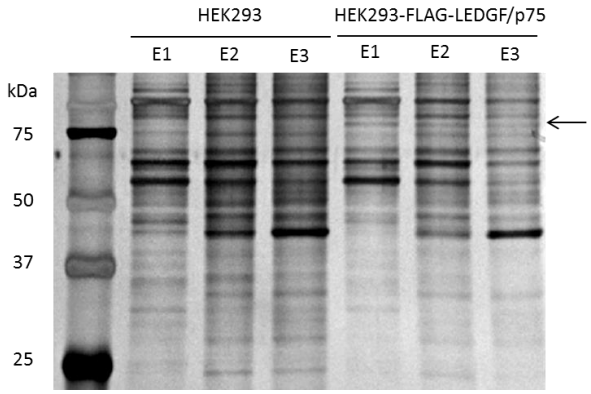


Figure 5-18: Silver staining of sequential elutions of FLAG-LEDGF/p75 from Ni-NTA beads.

Elutes collected from their elution from Ni-NTA beads during small scale purification of whole cell lysates from wild type HEK293 cells and HEK293-FLAG-LEDGF/p75 cells where separated by SDS-PAGE. The gel was silver stained. The arrow indicates the presence of a band in the transfected HEK293 cells that is not present in wild type HEK293, and which would correspond to purified FLAG-LEDGF/p75.

As well as a 6xHis tag, each of the different LEDGF constructs encoded a second tag viz. FLAG or Myc that would allow for double purification. Therefore, nickel bead purified protein samples from HEK293 and HEK293-FLAG-LEDGF/p75 cells were immunoprecipitated with magnetic anti-FLAG beads overnight according to the method described Chapter 5.2.11. Unbound protein was transferred to a clean tube

and labelled as IP waste. Aliquots of nickel bead purified, nickel bead plus immunoprecipitated sample (IP) (double purified), and IP waste samples were separated by SDS-PAGE and visualised using the SilverQuest Silver Stain Kit.

As shown in Figure 5-19, in lanes containing protein purified from HEK293-FLAG-LEDGF/p75 cells by either Nickel beads alone or double purified with Nickel beads and immunoprecipitation, there is a band that corresponds to FLAG-LEDGF/p75 which is absent in HEK293WT and beads-only lanes. This confirms that the double purification procedure was successful. The band (although very weak) is, however, present in the IP waste lane suggesting that not all of the protein is bound to the beads during the IP step. However, it should be noted that there are fewer bands present in the IP lane suggesting that this extra step removes non-specific proteins (for example, those that have bound to the Nickel beads via a naturally occurring run of His residues in their protein sequence).

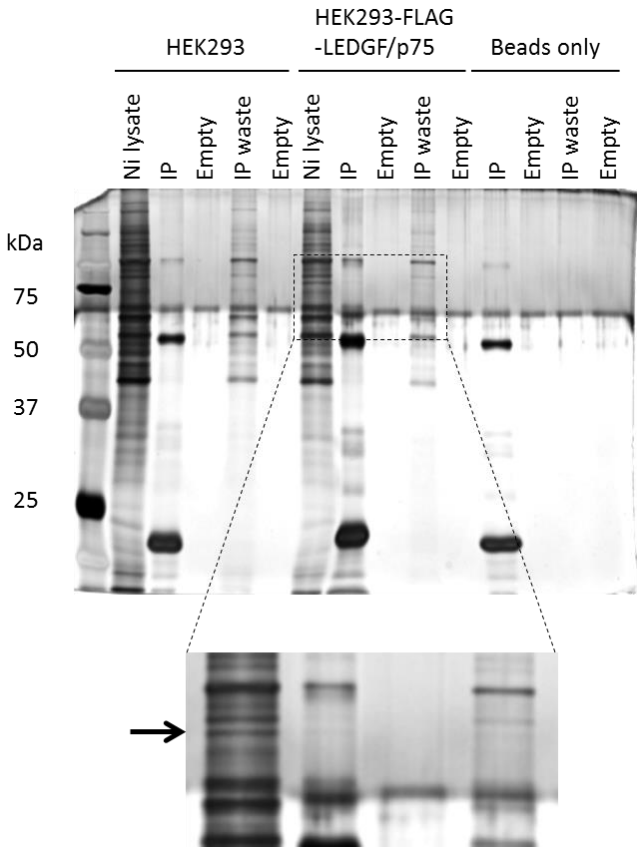


Figure 5-19: Silver staining of cell lysates from HEK293 and HEK293-FLAG-LEDGF/p75 cells purified by Ni-NTA beads alone or immunoprecipitated with FLAG beads following purification by Ni-NTA beads.

Eluates purified with Ni-NTA beads were subject to immunoprecipitation with anti-flag beads. Eluates from this (IP) were analysed alongside the respective Ni-NTA purified sample (Ni lysate). The remaining supernatant was also analysed (IP waste). The enlarged lanes show the region of the gel containing a band representing FLAG-LEDGF/p75 (indicated by the arrow).

5.3.6. Large scale purification of His-tagged LEDGF

The results so far demonstrate that the exogenous tagged LEDGF/p75 isoform can be purified on a small scale. However, the results also highlighted a number of limitations using this method (i.e. not all material binds to or elutes from the beads). It was therefore decided to use a larger scale purification system based on a low pressure chromatography system where the material was pumped through a 1 mL column containing the nickel beads in order to allow efficient binding (Chapter 5.2.10). Lysate from HEK293 cells stably expressing FLAG-LEDGF/p75 or pcDNA3.1HisA was re-cycled through the column five times to allow maximal binding. All washes and final elutions were collected and run on a 10% resolving gel, 4% stacking gel and stained with Coomassie Blue.

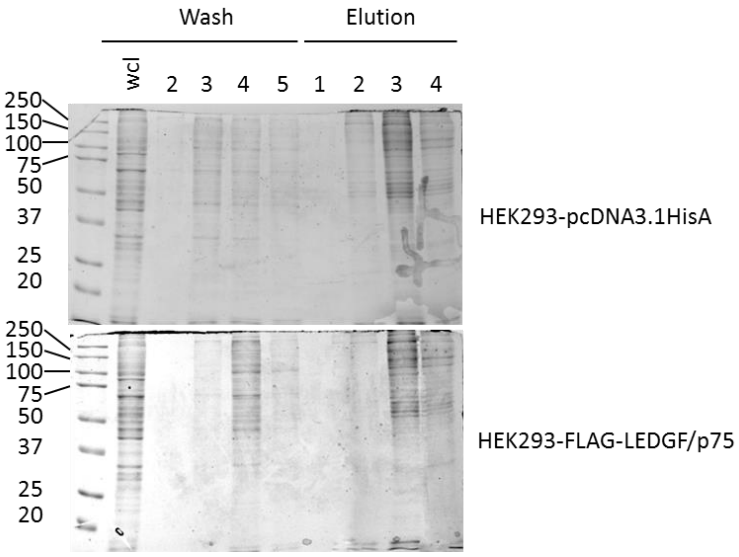


Figure 5-20: Analysis of wash and elution fractions collected from the purification of HEK293-pcDNA3.1HisA and HEK293-FLAG-LEDGF/p75.

A sample of the eluates collected from the washing of the column (wash buffer) and from the elution off the column in the elution buffer, were separated by SDS-PAGE. The gels were stained with Coomassie Blue. The top image is analysis of eluates from the purification of HEK293-pcDNA3.1HisA and the bottom is analysis of eluates from the purification of HEK293-FLAG-LEDGF/p75.

Elution 3, in each case, shows the most amount of relevant protein (Figure 5-20). To enable direct comparison between purified protein from wild type and transfected HEK293 cells, aliquots of elution 3 from both cell types were analysed next to each other on a polyacrylamide gel, stained with Coomassie Blue and examined for presence of LEDGF/p75. However, it was difficult to distinguish a band at 75 kDa

that would represent FLAG-LEDGF/p75 specifically in the transfected cell sample (Figure 5-21).

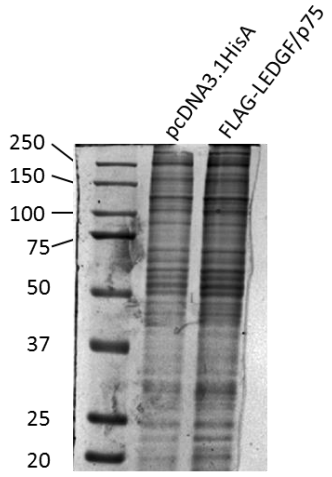


Figure 5-21: Analysis of the elution samples from the purification of HEK293 cells transfected with pcDNA3.1HisA empty vector and pcDNA3.1His-FLAG-LEDGF/p75. Corresponding elutes in elution buffer from the column purification of HEK293-pcDNA3.1HisA and HEK293-FLAG-LEDGF/p75 were separated by SDS-PAGE and the gels stained with Coomassie Blue.

To confirm that FLAG-LEDGF/p75 protein was purified, an aliquot (and dilutions of) of elution 3 from each cell line was separated by SDS-PAGE and probed by western blot using anti-FLAG antibody (Sigma Aldrich, Mouse monoclonal, 1:1000).

As shown in Figure 5-22, a band corresponding to FLAG-LEDGF/p75 is present in purified lysate from the transfected cells but not in wild type cells. This result

indicates that the purification procedure does work, and that FLAG-LEDGF/p75 protein can be purified using this method.

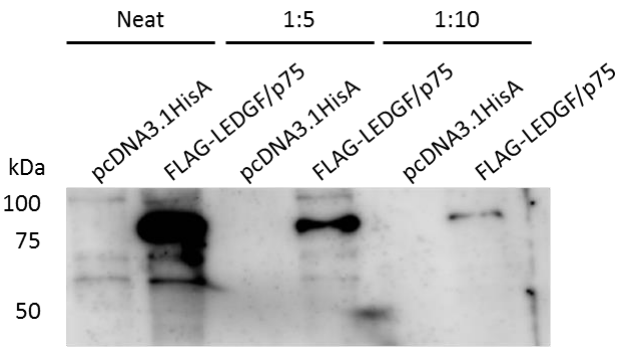


Figure 5-22: Western blot analysis of the purification of pcDNA3.1HisA and FLAG-LEDGF/p75.

Whole cell lysates from wild type and FLAG-LEDGF/p75 transfected HEK293 cells were purified using the low pressure chromatography system. Eluates in elution buffer were analysed by western blot using an anti-FLAG antibody. Samples were loaded neat, diluted 1:5 and 1:10.

Having demonstrated successful purification of FLAG-LEDGF/p75, it was important to show that proteins known to co-associate with LEDGF/p75 were also co-purified. One such protein is MeCP2[109]. To this end, an immunoprecipitation was performed using elution 3 (Figure 5-20) and anti-Flag magnetic beads. Equal volumes of precipitated proteins were separated on two SDS-PAGE gels and analysed by

western blot using anti-MeCP2 (Diagenode, Polyclonal, 1:1000) and α-FLAG antibodies. Results are illustrated in Figure 5-23.

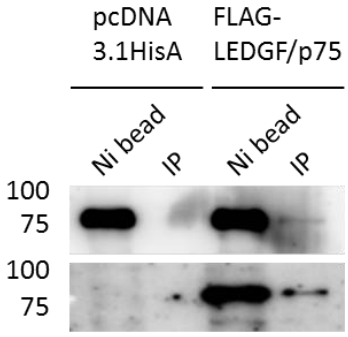


Figure 5-23: Western blot analysis of stable HEK293 whole cell lysates purified with nickel beads and then immunoprecipitation.

Samples of nickel purified lysates from HEK293 cells stably transfected with pcDNA3.1HisA or pcDNA3.1His-FLAG-LEDGF/p75 were subject to immunoprecipitation with anti-Flag beads. The precipitated proteins (IP) were analysed by Western blot alongside nickel purified input samples (Ni bead) as control. Blots were probed with anti-MeCP2 and anti-FLAG antibodies (upper and lower panel respectively).

MeCP2 protein contains a sequential run of 7 histidine residues which can bind to the nickel beads used in the purification. This is demonstrated by the presence of a band in the nickel purified sample from cells transfected with both the empty plasmid and FLAG-LEDGF/p75 (Figure 5-23, upper panel). However, following

immunoprecipitation with FLAG beads, this band is only present in the lane corresponding to cells transfected with the LEDGF construct. The results demonstrate the presence of MeCP2 in the purified tagged-LEDGF/p75 complex.

To further confirm the presence of FLAG-LEDGF/p75 in the immunoprecipitated complex, samples were analysed by Western blot and probed with an anti-FLAG antibody (Figure 5-23, lower panel). The presence of a band corresponding to FLAG-LEDGF/p75 in the IP lane confirms that the procedure was successful.

Western blot of elutions from either a nickel bead column or immunoprecipitated samples demonstrates that the method successfully purifies the LEDGF complex but sufficient material is not purified to allow visualisation by Coomassie Blue or Silver staining. There seem to be additional proteins that are common to the elution and control lanes (ie cells transfected with empty plasmid). It was therefore decided to add an extra wash step using 50 mM imidazole in the wash buffer in an effort to reduce the number of non-specifically purified proteins. This would mean the column was subjected to two washes with different concentrations of imidazole (20 and 50 mM) before elution at 250 mM imidazole. The column purification was repeated and all elutions collected. A sample of each eluate was analysed by western blot and anti-FLAG antibody (Sigma Aldrich, Mouse monoclonal, 1:1000).

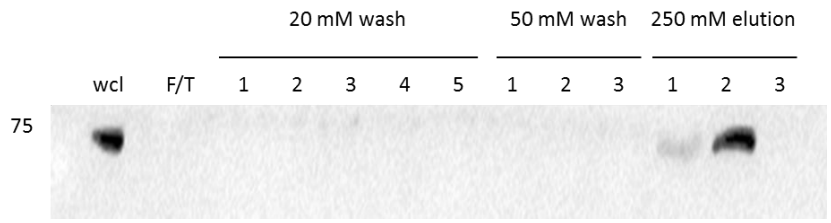


Figure 5-24: Western Blot analysis of nickel purified lysates from HEK293 cells transfected with FLAG-LEDGF/p75.

Whole cell lysates of HEK293-FLAG-LEDGF/p75 stable cells were purified using the low pressure chromatography system attached to a nickel column. Samples from all wash (50 mM wash) and elution (250 mM elution) steps were analysed by Western blot using an anti-FLAG antibody. As controls, samples of whole cell lysate (wcl) and flow through (F/T) were included.

A band representing FLAG-LEDGF/p75 was present in the unpurified whole cell lysate and elution 2 (Figure 5-24). There were no bands samples from cells transfected with empty plasmid (data not shown). To confirm that an LEDGF complex was purified, an immunoprecipitation was performed as described earlier using elution 2 (Figure 5-24) and anti-Flag beads and the gels probed with anti-FLAG and MeCP2 antibodies previously described.

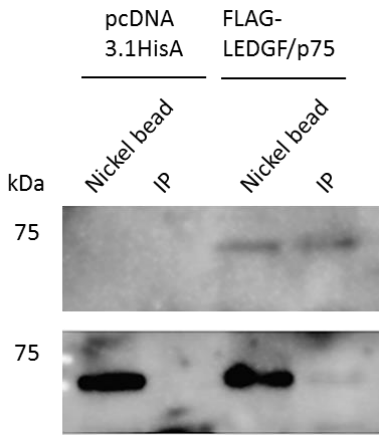


Figure 5-25: Western blot analysis of transfected HEK293 cell lysates purified with nickel beads and immunoprecipitation.

Samples of nickel purified lysates from HEK293 cells stably transfected with pcDNA3.1HisA or pcDNA3.1His-FLAG-LEDGF/p75 were subject to immunoprecipitation with anti-Flag beads. The precipitated proteins (IP) were analysed by Western blot alongside nickel purified input samples (Nickel bead) as a control. Blots were probed with an anti-FLAG antibody (upper panel) or an anti-MeCP2 antibody (lower panel).

As shown in Figure 5-25, there is a band in the lanes containing the immunoprecipitated sample on the blots probed with both the anti-FLAG and MeCP2 antibodies showing that the immunoprecipitation (and hence double purification) worked. Therefore, samples of elution 2 from both pcDNA3.1HisA and FLAG-LEDGF/p75 transfected HEK293 cells were run on a gel and probed using the SilverQuest Silver Staining Kit. To enable a direct comparison between each of the

transfections, the immunoprecipitated samples were loaded alongside each other on the gel. Figure 5-26 shows that there were no discernible bands specific to the lane containing immunoprecipitated sample from the FLAG-LEDGF/p75 transfected cells. As the previous result (Figure 5-25) had demonstrated the presence of FLAG-LEDGF/p75 in the immunoprecipitated sample, this result suggests that the amount of purified protein is below the detection limit of silver staining.

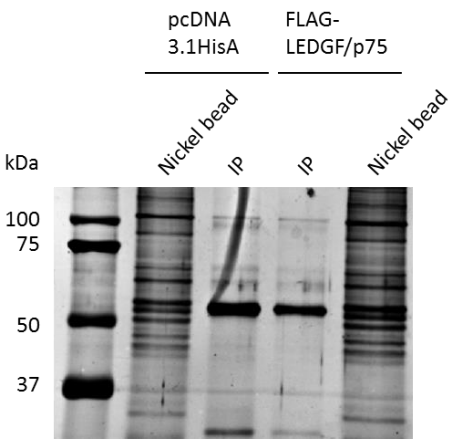


Figure 5-26: Analysis of nickel bead column and immunoprecipitated samples by silver staining.

Samples analysed by western blotting (Figure 5-25) were separated by SDS-PAGE and the gel visualised by silver staining. The immunoprecipitated samples (IP) from each of the samples were run adjacent to each other to allow for direct comparison. The respective column purified material (Nickel bead) was run alongside the IP samples.

As the results suggest that there is insufficient protein being purified to be visible by Coomassie Blue or silver staining, the amount of starting crude lysate was increased. Thus, lysate from $1x10^8$ total cells was applied to the nickel column and purified as previously described. Aliquots from all elution steps were analysed by Western blot with an anti-FLAG antibody (Sigma Aldrich, Mouse monoclonal, 1:1000) (Figure 5-27).

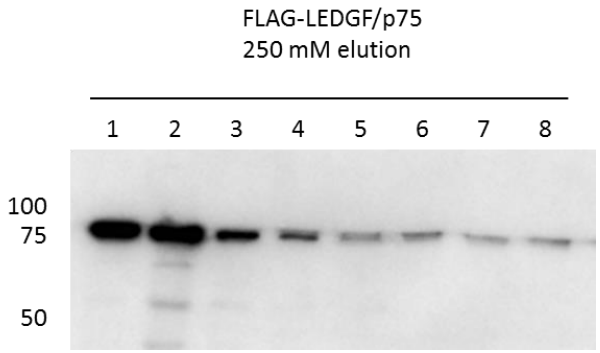


Figure 5-27: Western Blot analysis of eluates of the column purification of HEK293-FLAG-LEDGF/p75.

Whole cell lysates of HEK293-FLAG-LEDGF/p75 cells were subject to purification using the nickel column attached to the low pressure chromatography system. Each eluate in elution buffer was analysed by western blotting and the membranes probed with anti-flag antibody.

As expected, there were no bands present in any of the wash steps or elutions of lysate from cells transfected with empty plasmid nor were there any bands in

samples of the wash steps from cells transfected with FLAG-LEDGF/p75 (data not shown). The result also suggested that the purified protein was eluted from the column in the first four elutions (ie within 4 mL). It is therefore surmised that the final protocol should consist of the following steps:

1) Binding of the crude lysate to the column should be performed by passing the lysate through the column once at a very low flow rate.

2) Washes should be performed with 50 mM imidazole.

3) Purified protein should be recovered from the column with 2 x 2 mL elutions in 250 mM imidazole.

To this end, lysates from $1x10^8$ cells stably transfected with each of the LEDGF isoforms were purified and samples of the wash and final elutions were analysed by Western blot probed with either an anti-FLAG or anti-Myc antibody.

As can be seen in Figure 5-28, each of the isoforms was successfully purified.

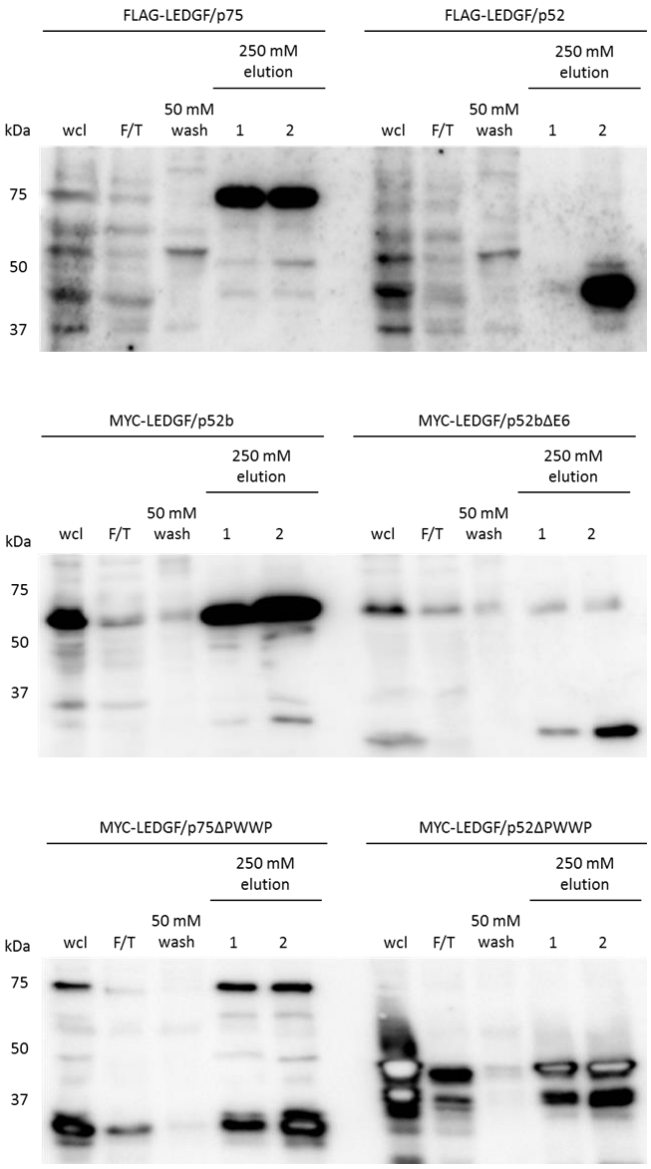


Figure 5-28: Western blot analysis of purified LEDGF isoforms.

Lysates from $1x10^8$ stably transfected HEK293 cells expressing each of the six isoforms of LEDGF were subject to nickel column purification as described. Samples of whole cell lysate (wcl), flow through (F/T), 50 mM wash and 250 mM elutions were analysed by western blot and the membranes probed with antibodies to the tags. Specifically anti-flag (top) and anti-myc (middle and bottom).

Samples of each of the elutions were run alongside on the same gel and silver stained as described in section 5.2.13. The result is shown in Figure 5-29.

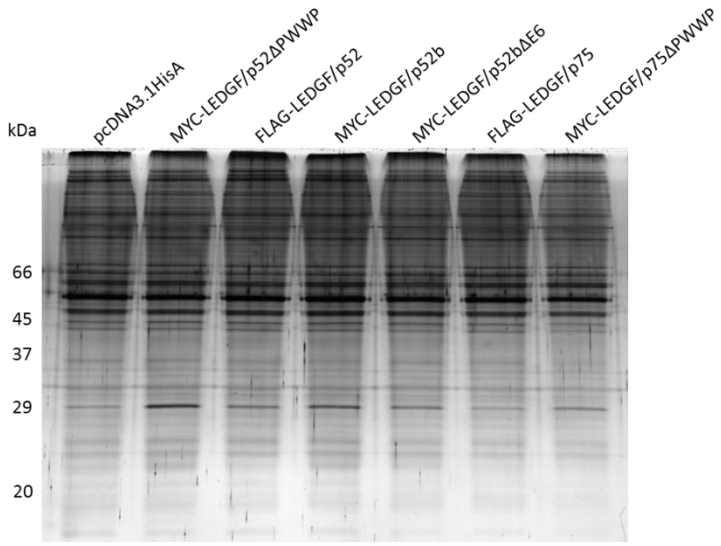


Figure 5-29: Analysis of eluates in elution buffer from the large scale purification of each of the tagged isoforms of LEDGF.

Whole cell lysates of each of the cell lines exogenously expressing a tagged isoform of LEDGF, were subject to purification on a nickel column attached to a low pressure chromatography system. Eluates in elution buffer were separated on a resolving gel and visualised by silver staining.

As in previous results, whilst the western blot clearly shows that the purification was successful, silver staining again failed to identify bands that are unique for each isoform.

5.4. Discussion

The aim of this chapter was to develop a method that could be used to examine how alternative splicing affects the ability of the different isoforms of LEDGF to participate in unique and distinct complexes, rather than in a disease-specific setting. A pitfall of such an investigation is the possibility of identifying proteins that are non-specifically bound to the molecule under investigation. In order to minimise this in my study, cDNA constructs of the different isoforms of LEDGF, identified in CLL cases, were generated to contain 2 unique tags. This would therefore allow LEDGF containing complexes to undergo 2 rounds of serial purification methods. Thus, constructs were cloned in to the pcDNA3.1HisA vector, such that recombinant proteins when expressed in cells would contain a 6xHis tag at the amino terminus. This would allow the complexes to be initially purified under native conditions. Secondly, the PCR primers used to generate the cDNA incorporated either a Myc, HA or FLAG tag which allowed for further enrichment by immunoprecipitation using antibodies specific for the tag.

The precise method used to generate the tagged constructs varied depending on the nature of the constructs. The cDNA molecules (with the exception of LEDGF/p52) that were generated were initially cloned into the pGEM-T cloning vector. This enabled the cDNA to be sequenced to ensure its integrity, and also allowed for easier digestion with the restriction enzymes (*Eco*RI and *Xba*I) used to generate the final pcDNA3.1HisA clones.

A modified technique was used to generate pcDNA3.1His-HA-LEDGF/p52 and pcDNA3.1His-FLAG-LEDGF/p52. A “cut and paste” approach was used that exploited the unique *Xho*I restriction site present within the N-terminus common to both the LEDGF/p75 and p52 isoforms, which obviated the use of a further PCR reaction and cloning into pGEM-T.

Due to the nature of the p52b isoform, a modified PCR technique was employed to generate pcDNA3.1His-MYC-LEDGF/p52b (Figure 5-2). This isoform is thought to arise from either the retention of part of the intron of LEDGF/p75 or the 3’-UTR of LEDGF/p52. If a reverse primer is designed to anneal to the end of the LEDGF/p52b C-terminal sequences, there is the potential to amplify the correct sequence at the very C-terminus, but then to diverge into the LEDGF/p52 sequence. Therefore, the N-terminus (amino acids 1 – 325) was amplified from the verified clone of pDNR-DUAL-LEDGF/p75. This ensured the sequence was the same in all constructs. A second “Touchdown” PCR reaction was used to allow specific amplification of the C-terminal region of LEDGF/p52b using the primers p52bLinkerPrimerNEW and p52bRXbaI (Table 5-1) and pcDNA3.1His-HA-LEDGF/p52 as template. After purification, the two resulting amplicons (N-terminus and C-terminus) were added to a 10-cycle primer-less PCR reaction. This allowed the two overlapping ends of the amplicons with complimentary sequences to anneal and be extended by the polymerase in the PCR reaction. Thus, at each cycle, the number of LEDGF/p52b full-length amplicons increased in a linear fashion. Following these 10 cycles, the

primers p52bc-mycF and p52bRXbaI were added which allowed for amplification of the full length LEDGF/p52b and also for the incorporation of the Myc tag.

Finally, PWWP domain mutants of both LEDGF/p75 and p52 were generated. The PWWP domain is found in members of the "royal superfamily" of proteins which have a role, amongst other functions, in epigenetic regulation as the domains recognise methylated lysine or arginine residues[110]. However, little else is known about the PWWP domain in the context of LEDGF, and thus generating deletion mutants would help dissect the molecular functions of proteins that contain the domain.

Initial studies to optimise the transfection conditions in order to obtain the maximum levels of expression were performed using pcDNA3.1His-HA-LEDGF/p75 and the 293T cell line. This cell line was chosen as it is highly receptive to transfection. Transfected cells were lysed in the buffer that would be used for nickel purification (Ni-purification was not carried out at this initial stage), and lysates examined by western blot using an anti-HA antibody (Figure 5-12). The results clearly demonstrate that the level of HA-LEDGF/p75 was similar in each of the conditions used. Bands were not detected in control transfection lanes (cells transfected with empty plasmid), confirming that the bands present in the positive transfection lysates were indeed the transfected protein. To further support this, when the blots were probed with an antibody against LEDGF bands were present in both the control transfection and transfected cell lanes that corresponds to endogenous LEDGF. In the lanes containing lysates from transfected cells, a second,

slower migrating band was observed that corresponds to the tagged LEDGF. Whilst the results suggested transfecting $2x10^5$ cells with 2 µg plasmid yielded slightly more protein, the increased numbers of cells and amount of plasmid required, especially if the transfection was to be scaled up, meant future transfections would be performed with $1x10^5$ viable cells with 1.5 µg plasmid.

Having established optimal transfection parameters, the integrity of each of the constructs was verified by western blot using whole cell lysates (Figure 5-13). Use of a β-actin antibody as a control meant it was possible to confirm protein had been loaded in the lanes containing whole cell lysate as bands were not visible using antibodies to the tags. As expected, bands were not detected in lanes containing lysates from mock or control transfected cells confirming the specificity of the antibodies that recognise the tags.

When the blots were probed with an anti-Myc antibody, a band of 60 kDa corresponding to endogenous C-myc protein acted as an internal positive control in the whole cell lysate lanes as well as a negative control in the lanes that contained purified protein that wasn't Myc-tagged. In the lane containing whole cell lysate from HEK293 cells over expressing MYC-LEDGF/p52bΔE6, a very faint band was visualised after longer exposure suggesting that the expression of this isoform in the HEK293 cells was low. When whole cell lysates from cells transfected with the ΔPWWP constructs were analysed, a number of bands with lower molecular weights were seen and not observed in lanes containing lysate from control transfections or purified protein. Deletion of the PWWP domain may result protein

instability, and hence these bands may represent breakdown products that are not purified with the Ni beads.

In order to examine the components of isoform specific LEDGF complexes, it was decided to establish stable cell lines expressing each of the splice forms. Several previous reports investigating the general role of LEDGF have used transiently transfected 293T cells[36, 102, 104, 107]. The 293T cell line is a variant of the HEK293 cell line which has been transformed to stably express the SV40 large T antigen. However, it is also highly resistant to neomycin and therefore stable cell lines cannot be established using G418. Therefore, stable cell lines were established using the parental HEK293 line.

There remained the possibility that the different isoforms of LEDGF could hetero- or homo-dimerise. This would confound the results if the over expressed protein dimerised with endogenous LEDGF. Thus, any binding partners identified in subsequent analyses would be specific for the isoform in question.

Previous studies that have identified binding partners associated with LEDGF have restricted the analysis to just one part of the protein. For example, it has been shown that the C-terminal region of LEDGF/p75 interacts with the HIV-1 IN, JPO2 and pogZ [102, 105, 106] and the Cdc7-ASK[107]. Other studies have restricted the analysis to a set of proteins with a specific function. Leoh *et al.* used transcription factor protein arrays to identify MeCP2 as a binding partner of LEDGF[109]. Martaens *et al.* did use a full length HA tagged LEDGF/p75 when they described JPO2 as an

interactor with the C-terminus of LEDGF[104]. However, very strict screening criteria were employed such that identified proteins were only considered if the peptides identified by mass spectrometry were present only in 293T-HAp75 samples and were completely absent from 293T samples, plus the proteins were identified by at least 2 unambiguous peptides. Using such an approach, several dozen potential interacting proteins were disregarded. It is interesting to note that Cdc7-ASK, MeCP2, PogZ and menin/MLL were all identified as binding partners of LEDGF/p75 following publication of this work[36, 106, 107, 109]. My work to identify the binding partners of LEDGF isoforms utilised full-length proteins. The ΔPWWP domain mutants of the two well documented LEDGF isoforms were also included in the protein complex investigation, as it is of interest to highlight any proteins that may solely bind to this important region.

The initial experiments to purify the constructs using Ni-NTA beads were carried out using the protocol described by the manufacturer. However, analysis of the elutions by western blot suggested the yields were not optimal. Therefore, it was necessary to optimise the amount of beads as well as incubation times. Whilst this led to an increase in amount of protein purified (Figure 5-15 and Figure 5-17), it gave rise to the problem that a lot of protein remained bound to the beads. Even when the number of elutions was increased, not all of the tagged protein was eluted from the beads (Figure 5-17). However, silver staining did show that as the number of elutions increased, the protein concentration in each elution decreased (Figure 5-18). As multiple elution sample analysis isn't practical; it would be better if

complexes could be purified in the lowest possible number of elutions. Even though it was very encouraging to visualise the over expressed protein on a gel that was silver stained, on closer inspection, it was noted that in the lanes containing "purified" wild type HEK293 lysates, there are multiple bands that would represent endogenous proteins in the cells that bind non-specifically to the nickel beads. These bands are also present in the lanes of the purified FLAG-LEDGF/p75 and it is therefore very difficult to highlight any bands that are present only in the purified over expressed protein lane and not in the wild type lane that may correspond to a unique protein that is part of a complex assembled around LEDGF/p75. This suggests that washing of the beads isn't efficient and/or the concentration of imidazole in the wash buffer should be increased from 20 mM. Furthermore, it should be noted that proteins with a natural run of histidine residues will non-specifically bind to the nickel beads.

As was described earlier, from the outset of this project, it was decided that the constructs should contain two different tags. As the results have so far demonstrated, a single purification using just nickel beads resulted in elution of proteins that were non-specifically purified from control cell lysates. Therefore, following nickel purification, an additional immunoprecipitation step was performed exploiting the FLAG tag in the protein to improve the purity of the proteins. This second step did purify the FLAG tagged LEDGF/p75 and reduced the number of the non-specific proteins (Figure 5-19) thus confirming the benefits of the extra purification strategy.

In order to perform mass spectrophotometric analysis, large amounts of starting material are required to purify the protein/complex of interest. As such, the small scale purification of the nickel tagged proteins using the Ni-NTA agarose beads isn't practical due to time, labour, amount of starting material required and inconsistencies in recovery and washing. To overcome this, a large scale purification method was developed using a low pressure chromatography system connected to a HisTrap HP column to achieve the maximum recovery of the tagged protein. HEK293 cells transfected with the empty pcDNA3.1His vector would allow for determination of proteins eluted in a non-specific manner.

Initially, the whole cell lysate was passed through the HisTrap HP column five times to give the over expressed protein maximum opportunity to bind. Eluates from each of the wash and elution steps were analysed by SDS-PAGE and Coomassie blue staining (Figure 5-20). Coomassie Blue staining was chosen as it was felt that if the over expressed protein purified by small-scale nickel beads could be visualised by silver staining, then the amounts of protein from a large-scale purification should be visible by this method.

The results demonstrated that the majority of the protein was eluted in the third elution with 250 mM imidazole. However, there were no bands apparent that were specifically found in the lane containing purified lysate from cells transfected with pcDNA3.1His-FLAG-LEDGF/p75. However, a subsequent western blot of the elutions demonstrated that the over expressed protein was present (Figure 5-22) and that the purification strategy was successful. Furthermore, an immunoprecipitation of

these same elutions with anti-FLAG magnetic beads and analysis by western blot illustrated that a known binding partner of LEDGF/p75, namely MeCP2[109] was purified using this method (Figure 5-23). Therefore, it would appear that this method was successful in purifying FLAG tagged LEDGF/p75 and at least one of the known interacting proteins. However, further optimisation was required to ensure maximum recovery of the tagged protein and the associated complex.

To this end, the first aspect to be considered was to improve the efficiency/stringency of the wash step. A second wash step was included with an increased imidazole concentration of 50 mM. Western blot analysis showed that this increase in imidazole did not result in any of the FLAG-LEDGF/p75 being displaced from the column (Figure 5-24). This extra wash step did not have any adverse effect on the subsequent immunoprecipitation following elution of bound proteins with 250 mM imidazole (Figure 5-25). Although it was anticipated that this extra wash step would aid in identifying bound proteins by silver staining, this was found not to be the case (Figure 5-26).

Therefore, the purification was repeated, but this time the lysate from ten times the number of cells was applied to the column. The method was once again successful as the release of the tagged protein could be tracked by western blot (Figure 5-27).

This method, however, was used to purify all of the tagged proteins with minor modifications. Firstly, the lysate was applied to the column just once, but with a low flow rate to allow as much of the tagged protein to bind. Secondly, the 20 mM wash

step was excluded, and washing was instead performed with 50 mM imidazole as it had been demonstrated that this concentration had no impact on the recovery of the tagged protein. Finally, the volumes of the 250 mM imidazole eluates collected were increased so as to reduce the number of fractions collected. The success of the purification was tracked by western blot (Figure 5-28) and the 250 mM eluate fractions analysed by silver staining (Figure 5-29).

Future work

As the large-scale purification method is now complete and material from the HisTrap HP column successfully purified and stored at -20 °C, the next steps are to investigate the options available to analyse the elutions to determine unique binding partners of the different isoforms of LEDGF.

Chapter 6 : Conclusions and Future Work

The novel findings resulting from this thesis are:

1. Both CLL and normal B cells express four isoforms of LEDGF, namely LEDGF/p75, LEDGF/p52, LEDGF/p52b and LEDGF/p52bΔE6.
2. These four isoforms of LEDGF are differentially expressed in CLL. More specifically, LEDGF/p75 is significantly over expressed in CLL and LEDGF/p52bΔE6 is significantly under expressed compared to normal B cells.
3. Mutations of the SF3B1 gene result in lower levels of expression of LEDGF.
4. The identification of a novel nonsense mutation in SF3B1.
5. Development of a method that allows for over expression of tagged LEDGF isoforms identified in CLL and normal B cells, and the subsequent purification of protein complexes involving these isoforms. This should allow for the determination of the components within the complexes associated with each of the LEDGF isoforms.

In this thesis, I have demonstrated that both normal B cells and B cells isolated from CLL patients express four isoforms of LEDGF. Given the sequence similarities between these isoforms, it is likely they all arise from alternative splicing of the *PSIP1* gene. The amino acid sequences of, LEDGF/p75, LEDGF/p52 and LEDGF/p52b

are identical up to residue 325 but then diverge into unique C-terminal sequences. The primary sequence of LEDGF/p52bΔE6 is identical to the three other isoforms up until residue 152. The whole of exon 6 is then deleted which leads to an altered open reading frame resulting in a premature stop codon. Ultimately, like the other isoforms, this isoform possesses a unique C-terminal tail[57]. It has been shown that LEDGF/p75 and LEDGF/p52 are associated with different protein complexes based on the ability of the unique C-terminal domains to bind to different proteins. Therefore, it is predicted that LEDGF/p52b and LEDGF/p52bΔE6 will also be associated with complexes that are unique to each isoform. It should be noted that there is one protein, namely MeCP2, which has been shown to bind to both LEDGF/p75 and p52 in their common N-terminal region[109]. Therefore, based on this knowledge, it is likely that each isoform may also share some common components within their respective complexes.

It would be of interest to determine the nuclear localisation of LEDGF/p52b and LEDGF/p52bΔE6 shown to be expressed in CLL and Normal B cells. Deletion of exon 6 in LEDGF/p52bΔE6 disrupts the end of the NLS and abolishes the AT hooks, although the critical lysine[150] residue, essential for localisation within the nucleus (discussed in Chapter 1) is retained. It could be predicted, therefore, that this isoform may be localised within the nucleus but it may not bind to the DNA as the AT hooks are absent.

By using RT-qPCR, it is shown that each of the four isoforms of LEDGF is differentially expressed in CLL compared to normal B cells. In particular, the data

shows that LEDGF/p75 is significantly over expressed in CLL cells whereas LEDGF/p52bΔE6 is significantly under expressed. The fact that LEDGF/p75 is over expressed in CLL is consistent with previous reports showing that this isoform is over expressed in a variety of other cancers[81, 82] and that it can be considered a cancer associated protein. The over expression of LEDGF/p75, which has been shown to protect cells from spontaneous apoptosis coupled with the under expression of pro-apoptotic LEDGF/p52bΔE6[57], may provide CLL cells with a mechanism to avoid programmed cell death. The opposing function of these two isoforms implies that one gene can undergo alternative splicing and yield separate protein products that have contrasting effects in the cell.

Statistical examination of the data generated from the RT-qPCR studies provides evidence for a relationship between the degree of CD38 positivity in CLL and the level of expression of LEDGF. Thus it was shown that those patients who are classified as CD38 positive have higher levels of each of the LEDGF isoforms. Expression of CD38 in CLL cells is associated with a poor prognosis and poor response to chemotherapy[90, 91], and it has been shown that high levels of expression of LEDGF/p75 renders AML cells resistant to daunorubicin induced cell death[57]. Taken together with the fact that LEDGF/p75 protects cells from apoptosis, it is not difficult to surmise that those patients with high levels of expression of LEDGF/p75 are more likely to be those with a poorer prognosis.

Given that LEDGF/p75 has been shown to be significantly over expressed in CLL and that a trend between CD38 expression and isoform expression has been suggested,

it may be possible to investigate LEDGF/p75 as a potential biomarker in CLL. Detection of LEDGF/p75 overexpression would imply a poor prognosis in individual cases. Hence, LEDGF/p75 could be explored as a therapeutic target in CLL. Previous data have suggested that low/absent levels of LEDGF/p75 results in cell death and thus over-expression could provide cells with an anti-apoptotic advantage (discussed in chapter 3). Therefore, targeting this isoform for degradation in CLL may tilt the balance to apoptosis in the cell. Feng *et al*., showed that BCL-2 regulates the transcriptional activity of a panel of proteins, including LEDGF[111]. BH3 mimetics are a class of drugs that target the BH3 domain in BCL-2 (and other family members) and inhibit its' activity thus affecting the pro-survival function of the protein[112]. Therefore, administration of this class of drugs to CLL patients would have a negative effect on the role of BCL-2, in-turn affecting the transcription of LEDGF/p75. Reduced levels of LEDGF/p75 could result in cell death.

There is growing interest in the role splicing factors play in the pathobiology of haematological malignancies, including CLL, after it was shown that SF3B1, a prominent member of the major U2 and minor U12 spliceosome complexes[67], is recurrently mutated in these diseases[63, 70]. These mutations were predicted to, and subsequently shown to, alter the function of SF3B1 by affecting the splicing patterns of other mRNA species within the cell[63, 70, 72]. Although the data did not reach statistical significance, in CLL patients with a mutated SF3B1 gene, there are lower levels of all four LEDGF isoforms as compared with the levels in patients with wild type SF3B1. This would suggest that the mutated SF3B1 gene does have an effect

on the expression and splicing of LEDGF. As with previous reports, the mutations are all clustered in the HEAT repeats. Mutations in this region are predicted to alter the function of SF3B1[100] and result in incorrect splicing.

All of the mutations within SF3B1 that have been described to date have been missense, silent or deletions. One of the novel findings of this thesis is the identification of a nonsense mutation. Although we have shown this mutation is present at the mRNA, we have not been able to say with conviction that this results in the expression of a truncated protein. Two explanations for this could be due to the mRNA being degraded by NMD or the protein itself is degraded following ubiquitination. Either way, this would prevent detection of the truncated protein by western blot. However, this result does raise the possibilities that nonsense mutations do occur in SF3B1.

As it has been shown that LEDGF/p75 and LEDGF/p52 are involved in different complexes by virtue of their divergent C-terminal domains, it is postulated that complexes involving LEDGF/p52b and LEDGF/p52bΔE6 will also be different. To determine the components of protein complexes formed around each isoform, mammalian expression constructs were generated that contained two different tags. This would allow for purification of LEDGF containing complexes by sequential methods, thus ensuring a high degree of purification and reliability of results. These constructs were transfected into HEK293 cells, and stable cell lines expressing each of the isoforms were established. Further, stable cell lines were established that express LEDGF/p75 and LEDGF/p52 with the PWWP domain deleted. A protocol was

developed that involves a double purification of each of the exogenously expressed isoforms, such that components of the associated complexes can be identified through mass spectrometry. Identification of these components may give an indication of the roles of the different LEDGF isoforms in the cell.

In conclusion, I aimed to provide further evidence that alternative splicing is a mechanism of epigenetics. An epigenetic mechanism is one in which there is change in phenotype that isn't caused by an alteration of DNA sequence. The *PSIP1* gene is known to be alternatively spliced into two well characterised isoforms: LEDGF/p75 and LEDGF/p52[2]. These are known to have differing nuclear localisations[26] as well as differing roles in the cell. Thus, whilst, LEDGF/p52 has been postulated to have a role in alternative splicing via modulating the inclusion and exclusion of exons[24, 27], LEDGF/p75 has several roles as a molecular tether[36, 102], involvement in the stress response[46] and a role in transcription[28]. The ability to perform these roles is dependent on the different complexes these isoforms are associated with. Thus, one gene gives rise to two proteins with different functions, which in turn is dependent on the complexes they form. Thus, the proteins are phenotypically different but are expressed from the same genetic sequence. Therefore, although not investigated to exhaustion, there is evidence to support that alternative splicing is a form of epigenetic regulation.

In order to carry this work forward, it will be important to determine the cellular localisation of LEDGF/p52b and LEDGF/p52bΔE6. As discussed, the cellular location of LEDGF/p75 and LEDGF/p52 is unique to each isoform and is specific to their

respective roles in the cell. The localisation of LEDGF/p52bΔE6 would be most interesting as the end of the NLS is disrupted in this isoform but the critical lysine[150] residue is still present. For such a systematic study it is essential that specific antibodies that each recognises the endogenous variant proteins are generated for the data to be meaningful. The generation of specific antibodies to each of the isoforms would involve them being raised against the respective unique C-terminus. As discussed, the C-terminal tails of LEDGF/p75, p52 and p52b all differ after residue 325. In the case of LEDGF/p52bΔE6, deletion of exon 6 yields a unique tail from residue 152 to 167. However, the unique part of the immunogen used to generate the antibodies for LEDGF/p52 and LEDGF/p52bΔE6 is relatively small and this should be noted when analysing the specificity of the antibody.

It is also important to determine the composition of the protein complexes associated with each isoform, as it has been shown that the complexes generated with LEDGF/p75 and LEDGF/p52 are unique to the specific roles of each of these isoforms. For example, Pradeepa *et al*. have shown that ~95% of proteins that bind to LEDGF/p52 have a role in alternative splicing[24]. To this end, it will be necessary to perform mass spectrophotometric analysis of the purified LEDGF isoform complexes, obtained from the stable cell lines.

Finally, by use of the PWWP domain mutants that were generated as part of these studies, the proteins that specifically bind to the PWWP domain should be determined. As discussed, this domain is an important reader of histone modifications. As it is present in each of the LEDGF isoforms that are products of

alternative splicing, it highlights this domain may be important, both structurally and mechanistically.

Appendix A:

Clinical data of all CLL patients used within this study to examine the expression level of each isoform of LEDGF and the mutation status of their SF3B1 gene.

Sample	WBC (10^9/L)	Rai	Binet	Status	CD38 (%)	Ig Isotype	VH segment useage	p53	Karyotype	Age at diag	Treated at/prior sampling?	Treated	T(Δ-S)	TFT	OS	status	SF3B1 status
2953	126	IV	C	u	98	M	3-09		11q-	.	N	Y	50.92	51.15	66.14	1	wt
2512	257	I	B	u	1	M	3-15			54	N	Y	15.78	15.98	60.36	0	wt
3047	141	IV	C	u	4	M	3-15		bi-13q-	59	Y	Y	82.74	13.38	82.74	0	wt
2902	40	0	A	m	5	M	3-30			71	N	N	78.3	.	90.99	0	wt
2968	87	II	B	m	.	G	3-30			.	N	N	155.13	.	155.13	0	wt
2950	150			m	.	M	3-48			78	N	N	10.06	.	22.09	0	wt
2673	65	0	A	u	9	M	3-49	wt	mon-13q-	.	N	Y	1.87	24.06	54.67	0	wt
2979	122			u	.	G	3-49			.	N	Y	83.86	83.96	89.15	1	wt
2744	29	0	A	m	1	M	3-7			66	N	Y	4.34	35.07	47.96	0	wt
2999	119	I	B	m	7	M	3-73		Clonal evolution	58	N	Y	65.52	82.51	106.18	0	m
2248	36	IV	C	m	.	M	4-34	C	bi-13q-	.	N	Y	0	37.08	63.61	0	wt
2565	122.9				.			WT	13q- (mono and bi)	.	Y	Y	116.04	0	116.04		wt
2408	77.9			u	.	M	1 69	WT	TRI-12	62	N	Y	165.94	176.92	117.91	1	m
2418	5704			u	24	M	1 69			74	N	Y	3.52	15.06	59.43	1	wt
1800	21.8			m	.	M	3 72	WT		72	Y	y	40.66	.	.	0	wt
2136	66			m	.	G	4 34	WT	13q- (mono and bi)	51	Y	y	33.96	0	94.08	0	wt
2539	64.4			m	.	M	3 74	WT	11q- 13q- (mono)	58	Y	Y	48.16	44.9	88.59		m
2533	116.1			u	.	M	3 7	WT	mon-13q-	.	N	Y	12.13	30.08	47.01	0	wt
1731	66.9				87			WT		.	N	N	0	.	0		wt
2472	21			u	.	M	3 74	WT		61	N		18.74	.	42.37	1	m
2573	76.1				.			WT		77	N		25.87	.	78.8	0	wt
1815	66.9	0	A	m	.	M	2 26	WT		40	N	Y	10.16	66.47	112.72	0	wt
2722	16.9			u	.	M	1 14		11q-	.	N	N	5.1	.	11.08	0	wt
2674	42.2	III	B	u	5	M	3 64	B	17p-	47	N	Y	2.76	9.43	24.1	1	m
2230	127.9			u	.	M	1 69	WT		.	Y	y	0	0	0		m
2358	146.7	IV	C	u	.	M	2 5	B		69	N	Y	10.52	10.52	15.38	1	m
2747	147.4		A	u	60	M	4 34			72	N	N	58.78	.	69.07	0	wt
2399	240			u	97	A	5 51	B		.	N	N	0	.	0	0	wt
2088	35		A	u	78	M	3 48	WT	13q- (mono), biclonal.	50	Y	y	2.43	.	4.04	0	wt
2765	29.6		C	u	25	M	1 69		17p- 13q- (mono and bi)	43	y	Y	1.25	0	42.97	0	wt
2787	41.4	0	A	m	6	M	3 7		13q-	60		y	0.49	.	1.64	0	wt
2789	12.2	0	A	m	2	M	3 7		mon-13q-	56	N		0.56	.	34.85	0	m
2792	24.4			m	.	M	3 30			.	N	y	156.44	.	200.49	0	wt
2793	14.9			m	7	M	3-30*3/3-33*01			57	N	Y	26.53	42.04	45.04	0	m
2794	16	II	B	u	2	M	1 2			69	N	Y	23.44	43.56	43.56	0	wt

Sample	WBC (10^9/L)	Rai	Binet	Status	CD38 (%)	Ig Isotype	VH segment useage	p53	Karyotype	Age at diag	Treated at/prior sampling?	Treated	T(Δ-S)	TFT	OS	status	SF3B1 status
2804	47	0	A	m	.	M	3 15			.	N	N	0	.	9.63	0	wt
2809	14.5			u	3	M	1 69	A	13q-	54	Y	Y	.	.	25.12	1	wt
2820	17.9		A	u	.	M	3 23	A		62	N	N	38.13	.	71.04	0	wt
2821	35.8		A	u	.	M	1 03			68	N		5.56	.	36.23	0	wt
2843	21.9			m	.	M	1 69	WT		.	N	N	0	.	12.98	0	m
2844	71.6		A	m	.	M	1 03	WT		.	N	N	87.48	.	87.48	0	wt
2855	15			m	.	G	3 53			61	N		3.98	.	6.31	0	wt
2867	9			m	.	M	3 23			62	Y	Y	43.26	.	56.71	1	wt
2874	20.9			m	.	M	3 9			64	N		74	.	74.75	0	wt
2876	63.1			m	.	M	3 33			.	N	N	75	.	35.14	0	wt
2877	160.8			m	.	M	3 49			76			159.04	.	159.04	0	m
2920	16.1			m	.	M	3 7			80	N	N	25.41	.	25.41	0	wt
2924	101.5			u	.	M	4 39		TRI-12	.	Y	Y	48.06	.	64.99	0	wt
2930	62.7		A	m	.	M	3 21			77			49.44	.	78.47	0	wt
2937	31.8		A	u	.	M	4 59			.	N	N	40.3	.	66.07	0	wt
2938	9.8				.				Multiclonal	.	N	N	1.97	.	11.67	0	wt
2949	38.2			m	.	M	3 48			64	N		0	.	28.6	0	wt
2962	25.3			u	.	M	3 20			78	N	N	0	.	11.77	0	wt
2967	22.3		A	m	.	M	2 5			59	N	N	0	.	0	0	wt
2985	76.7	0?	A	m	.	M	3 20			.	N	y	0	.	30.18	0	wt
3016	14.7				.				Multiclonal	.	N	N	0	.	0	0	wt
3018	16.4			m	.	M	1 24			.			29.82	.	29.82	0	wt
3035	282.1			u	.	M	3 23			61	N	y	17.13	.	44.05	0	wt
3037	78.9		B	m	.	M	3 23			70	Y	Y	46.02	42.31	63.45		wt
3068	32.3		B	u	.	M	3 11			55	Y	y	88.17	.	109.07	0	wt

Sample	WBC (10^9/L)	Rai	Binet	Status	CD38 (%)	Ig Isotype	VH segment useage	p53	Karyotype	Age at diag	Treated at/prior sampling?	Treated	T(Δ-S)	TFT	OS	status	SF3B1 status
3069	15.9		C	u	.	M	3 48			.	Y	Y	65.75	39.32	72.19	0	wt
3071				u	.	M	3 66			.	N	y	0	.	25.02	0	m
3076	40.8			u	.	M	3 33			.	N	Y	0	0	0	0	wt
3078	118.5		A	u	.	M	5 51			.	N	N	126.46	.	126.46	0	wt
3089	161.1			m	.	M	3 23			.	N	N	0	.	0.99	1	wt
3091	41.7				.					.	N	Y	0	.	6.57	0	wt
3092	29.9			m	.	M	3 9			.	N	N	0	.	0	0	wt
3125	86			u	.	M	1 8			.	N	Y	129.68	.	129.78	0	wt
3139	112.4			u	.	M	3 15			.		Y	36.79	.	36.79	0	wt
3151	14.6			m	.	M	1 69			.			0	.	34.42	0	wt
3144	114.8		C	m	.	G	4 34			51	Y	Y	143.26	101.12	158.61	1	wt
3187	16.3		A	m	.	M	1 2			.	N	Y	8.65	.	15.02	0	wt
3190				u	.	M	3 30		13q-	.	Y	y	0	.	36.92	0	wt
3192	70.8		A	m	.	M	3 15			.	N	N	0	.	0	0	wt
3160	184.4			u	.	M	1 69			.			0	.	0	0	m

Appendix A-Table 1: Clinical data of CLL cases used within this thesis.

The CLL case number is given in column 1 followed by the white blood cell count (wbc), the Rai and Binet stage, the status of the Ig gene (mutated or unmutated, m or u), CD38 expression, the Ig isotype, VH segment useage, the p53 status, karyotype data, the age at diagnosis, treatment prior to sampling, if the patient was treated at any time, the time between diagnosis and sampling (T(Δ-S)), time to first treatment (TFT), overall survival (OS), status of the patient (dead or alive) and the mutational status of the SF3B1 gene.

Expression data of each isoform of LEDGF in each CLL sample.

Sample	Normalised p75	Normalised p52	Normalised p52b	Normalised p52bΔE6
2953	.	.	.	.
2512	.	.	.	0.1134
3047	.	.	.	1.3472
2902	0.1756	0.0928	0.0317	0.0842
2968	.	.	.	.
2950	.	.	.	.
2673	.	.	.	.
2979	.	.	.	.
2744	0.2813	0.2912	0.0878	1.4340
2999	0.3209	0.1207	0.0854	0.1321
2248	.	.	.	.
2565	0.2045	0.0461	0.1216	.
2408	0.3186	0.0600	0.0621	0.0860
2418	0.6736	0.0988	0.0748	0.2912
1800	0.3950	0.1856	0.0552	0.1476
2136	0.5434	0.1096	0.2192	0.8827
2539	0.1022	0.0313	0.0448	0.0748
2533	0.0579	0.0254	0.0548	.
1731	0.7371	23.2636	0.1207	0.8467
2472	0.2073	0.1560	0.1571	0.1507
2573	0.3099	17.0299	0.0409	0.3511
1815	0.4175	0.0884	0.0409	0.0182
2722	0.7474	14.3204	0.1684	0.8467
2674	0.2952	0.0522	0.0177	0.1127
2230	0.3299	0.0349	0.0361	0.1550
2358	.	.	.	1.0210
2747	0.4118	1.0425	0.1127	0.3253
2399	0.2832	0.4965	0.0670	0.5322
2088	0.6329	1.7901	0.0791	0.1487
2765	0.1397	0.0802	0.0451	0.1127
2787	.	.	.	.
2789	.	.	.	.
2792	.	.	.	.
2793	0.1882	0.2774	0.0638	0.1895
2794	.	.	.	.
2804	.	.	.	.
2809	.	.	.	.
2820	.	.	.	.
2821	.	.	.	.
2843	0.2146	0.1672	0.0718	0.1191
2844	0.2872	0.2661	0.0693	0.3487
2855	0.1487	0.0621	0.1250	.
2867	.	.	.	.
2874	.	.	.	.
2876	.	.	.	.
2877	0.8467	0.4414	0.0643	0.0836
2920	0.2912	0.1358	0.0652	0.2755
2924	0.2698	0.1321	0.1119	1.1408

2920	0.2912	0.1358	0.0652	0.2755
2924	0.2698	0.1321	0.1119	1.1408
2930	0.3415	0.2161	0.0769	0.5864
2937	0.3896	0.1487	4.5948	0.7526
2938	.	.	.	.
2949	0.2912	0.1008	5.0281	0.3585
2962	0.6462	0.4118	5.9381	2.6759
2967	0.2102	0.1487	5.0982	0.3923
2985	0.2449	0.0941	4.6268	0.1627
3016	.	.	.	.
3018	0.2432	0.1627	7.5685	1.0353
3035	0.8351	0.4323	0.1183	0.3035
3037	1.1019	0.4506	0.1051	0.8645
3068	0.9138	0.6417	0.2161	0.4475
3069	0.5070	0.1436	0.0643	0.1975
3071	0.4090	0.1768	0.0934	0.5000
3076	.	.	.	.
3078	.	.	.	.
3089	0.2973	0.1908	0.1276	0.2973
3091	.	.	.	.
3092	0.3463	0.1349	0.1207	0.1743
3125	.	.	.	.
3139	.	.	.	.
3151	0.3415	0.0415	0.0830	0.1792
3144	.	.	.	.
3187	0.6113	0.3950	0.2365	0.4118
3190	0.4175	0.0759	0.0819	0.3057
3192	.	.	.	.
3160	0.3660	0.1081	0.0364	1.0644

Appendix A-Table 2: The expression of each isoform of LEDGF in each CLL sample.

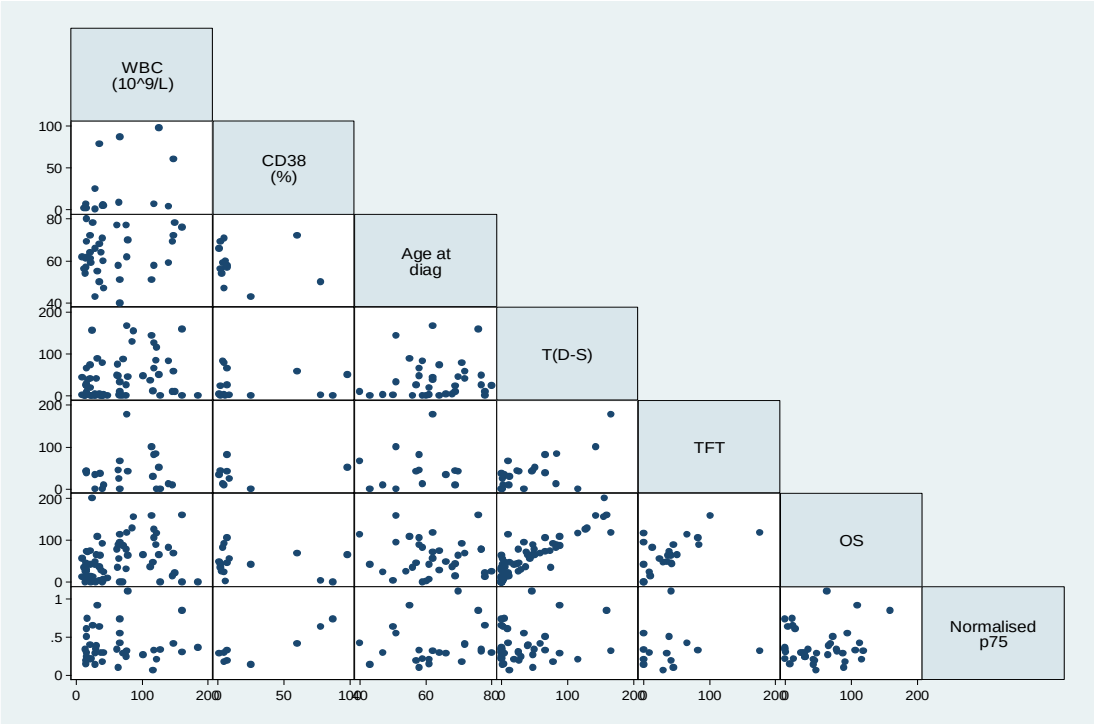
The expression level relative to RPL27 reference gene is given for each CLL patient.

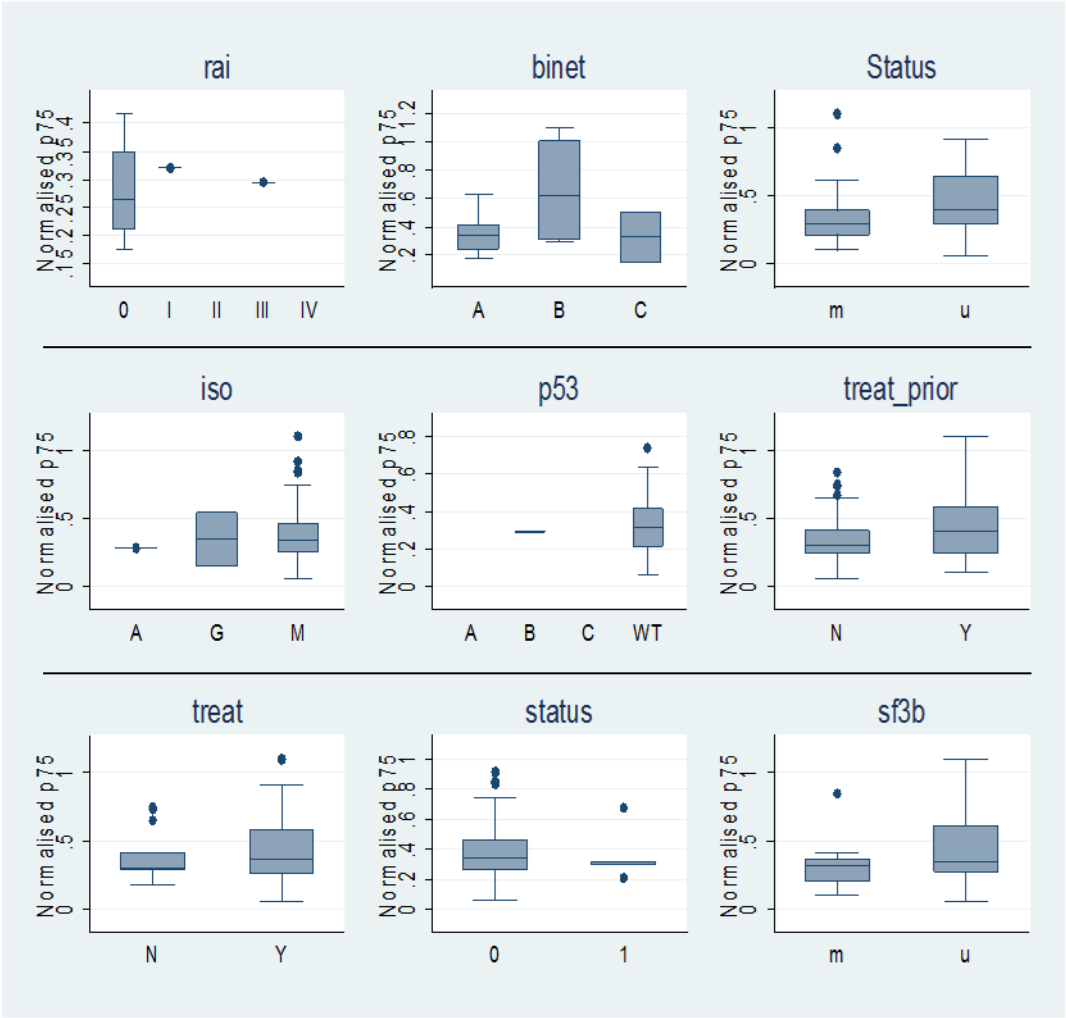
Appendix B

Exploratory analysis of potential relationships between the expression levels of each of the LEDGF isoforms in CLL samples and clinical factors associated with CLL.

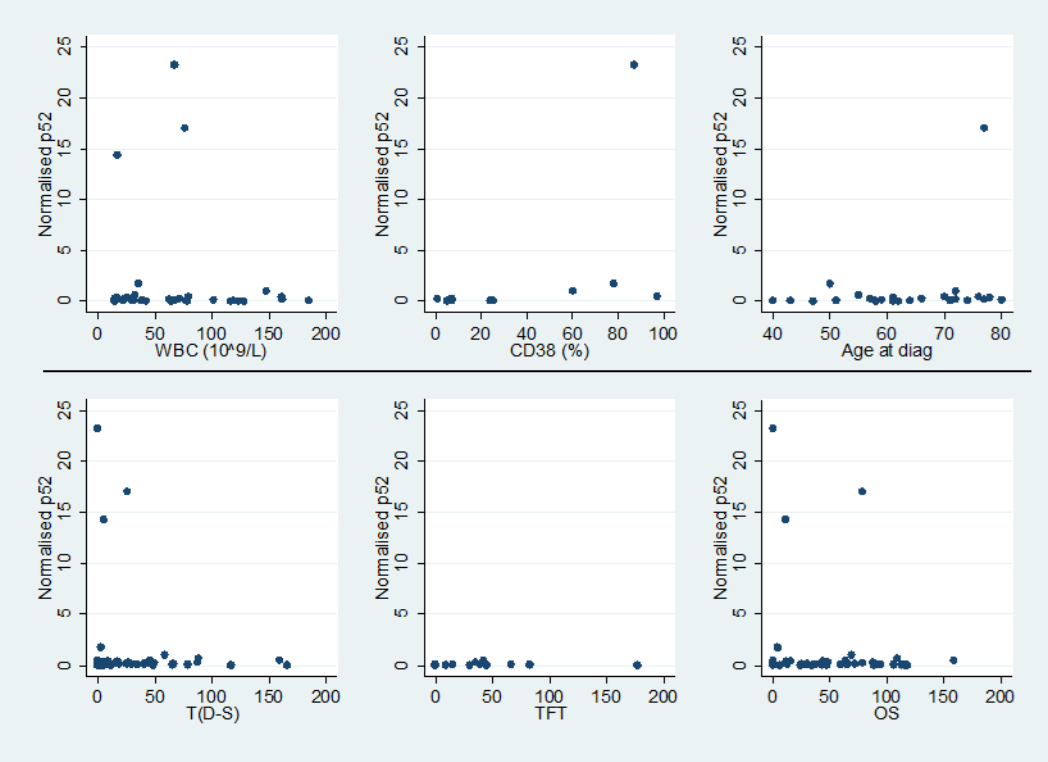
Clinical factors used for this investigation were white blood cell count (wbc), CD38 expression, age at diagnosis, time between diagnosis and sampling (T(D-S)), time-to-first-treatment (TFT), overall survival (OS), Rai and Binet status, IGHV gene status (unmutated or mutated, u or m), Ig isotype (iso A, G, M), p53 status (A, B, C or wild-type), treated prior to sampling (treat_prior), treated while alive (treat), status of patient (dead or alive, 1 or 0) and SF3B1 status (unmutated or mutated).

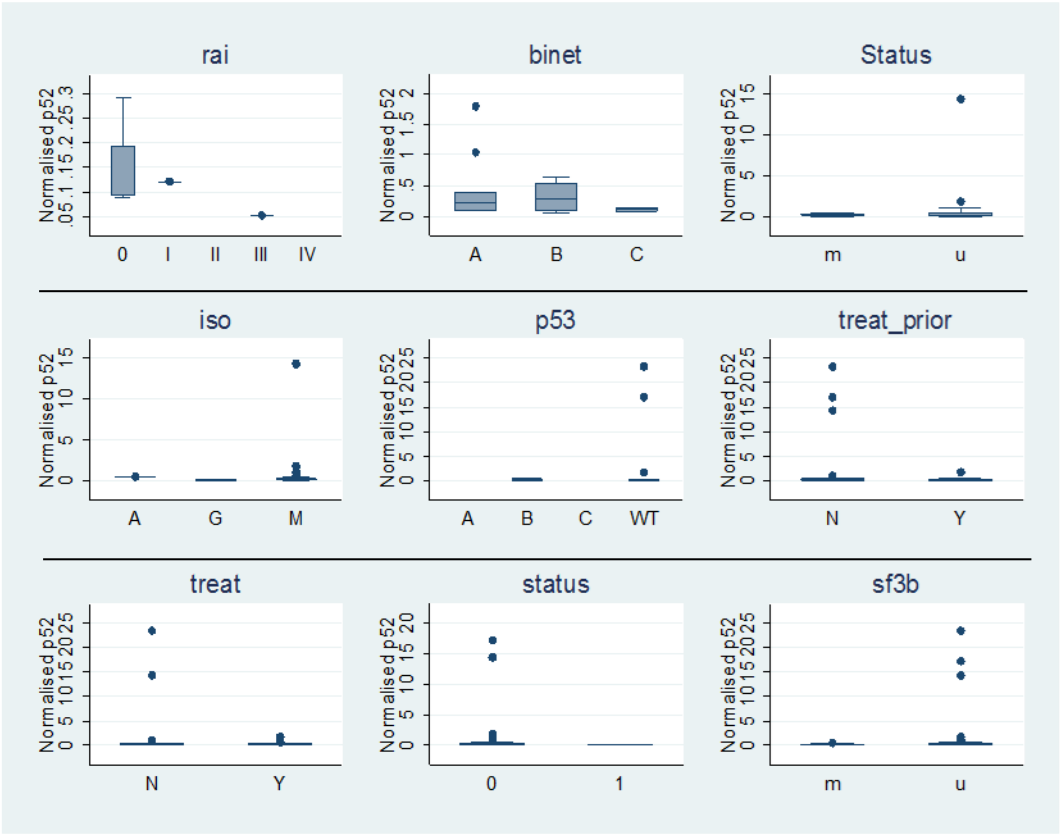
1. Relationship between LEDGF/p75 expression and clinical factors.



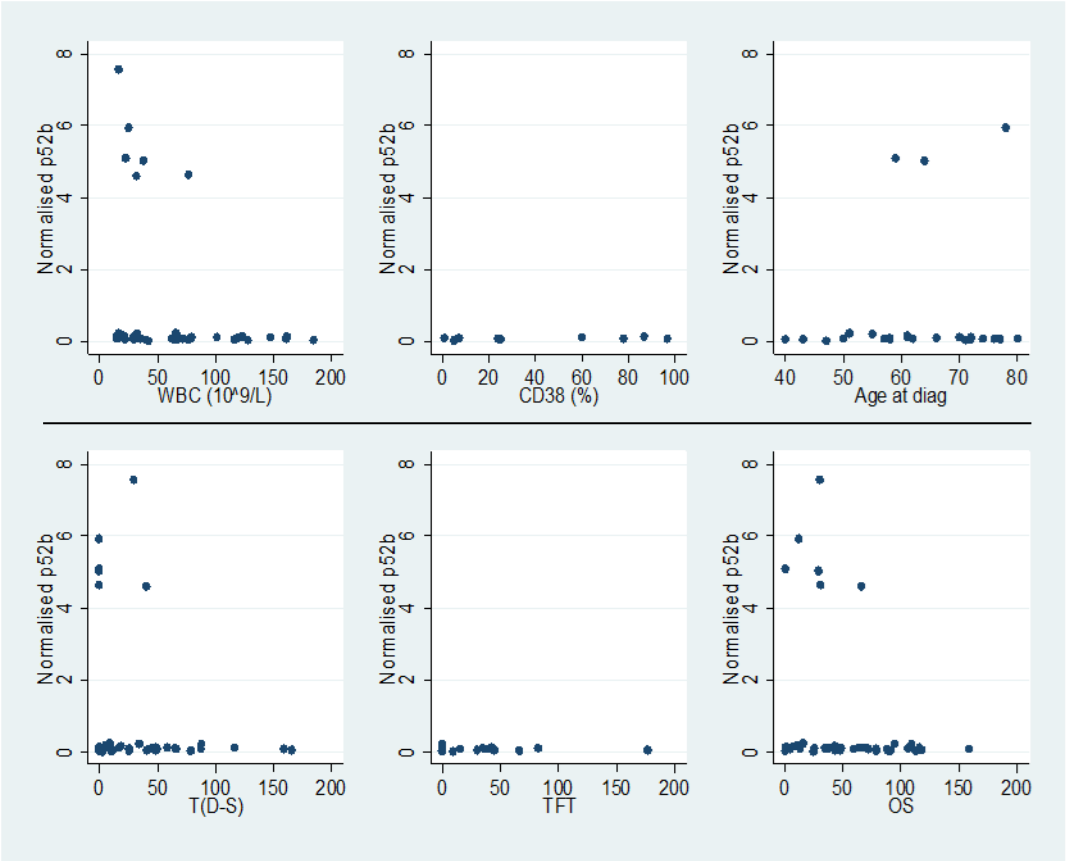

rai
Normalised p75
0 I II III IV
binet
Normalised p75
A B C
Status
Normalised p75
m u
iso
Normalised p75
A G M
p53
Normalised p75
A B C WT
treat_prior
Normalised p75
N Y
treat
Normalised p75
N Y
status
Normalised p75
0 1
sf3b
Normalised p75
m u

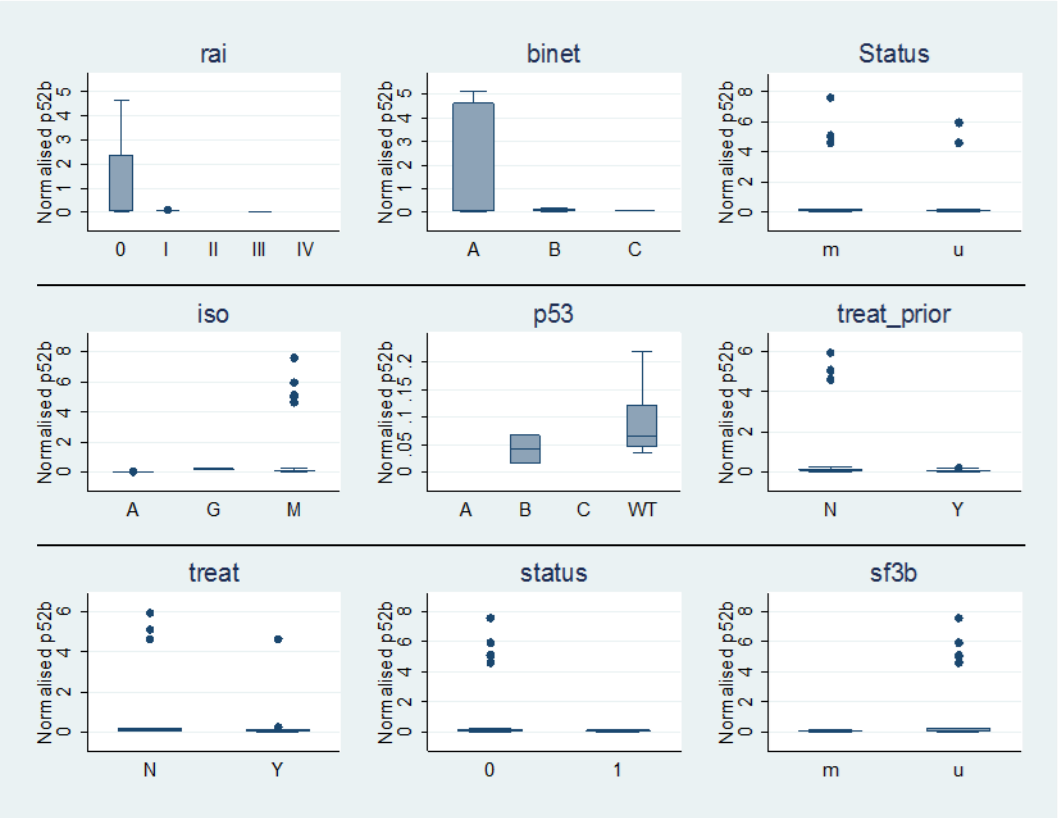
2. Relationship between LEDGF/p52 expression and clinical factors.



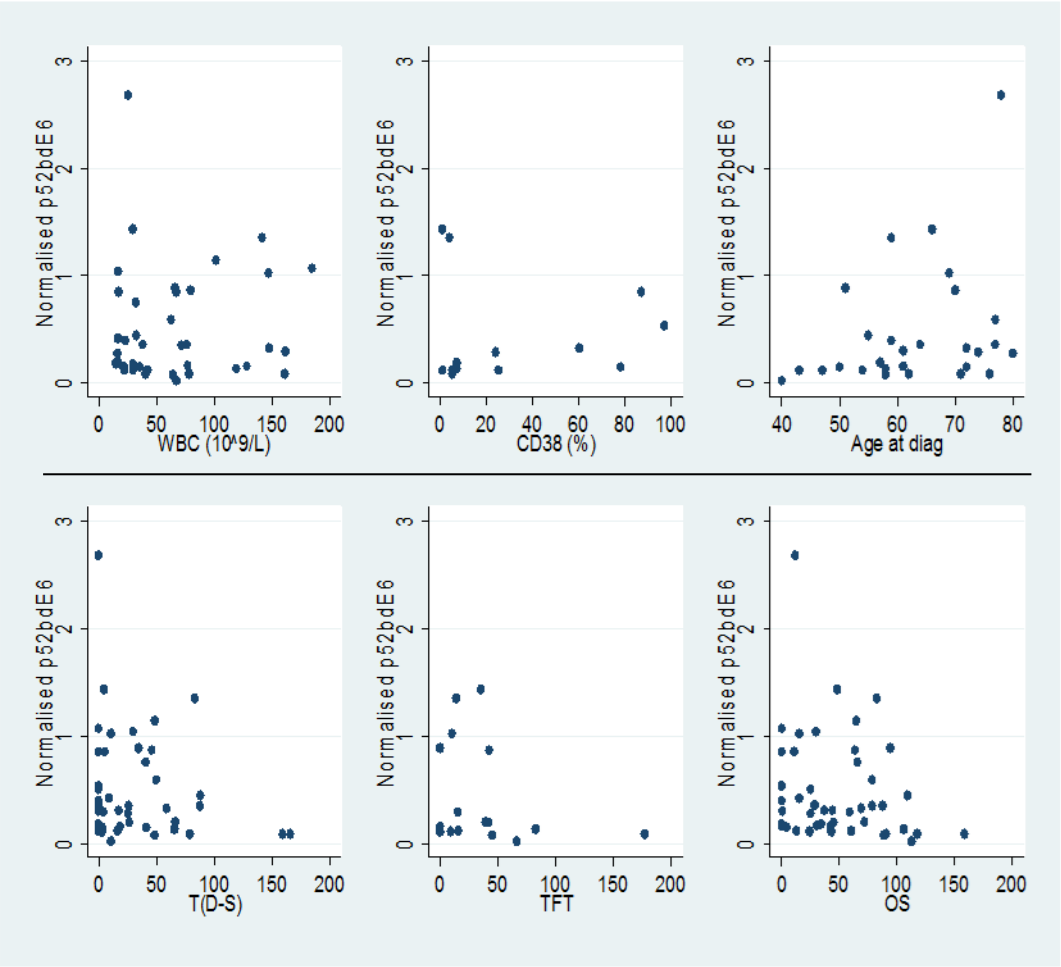

rai
Normalised p52
0 I II III IV
binet
Normalised p52
A B C
Status
Normalised p52
m u
iso
Normalised p52
A G M
p53
Normalised p52
A B C WT
treat_prior
Normalised p52
N Y
treat
Normalised p52
N Y
status
Normalised p52
0 1
sf3b
Normalised p52
m u

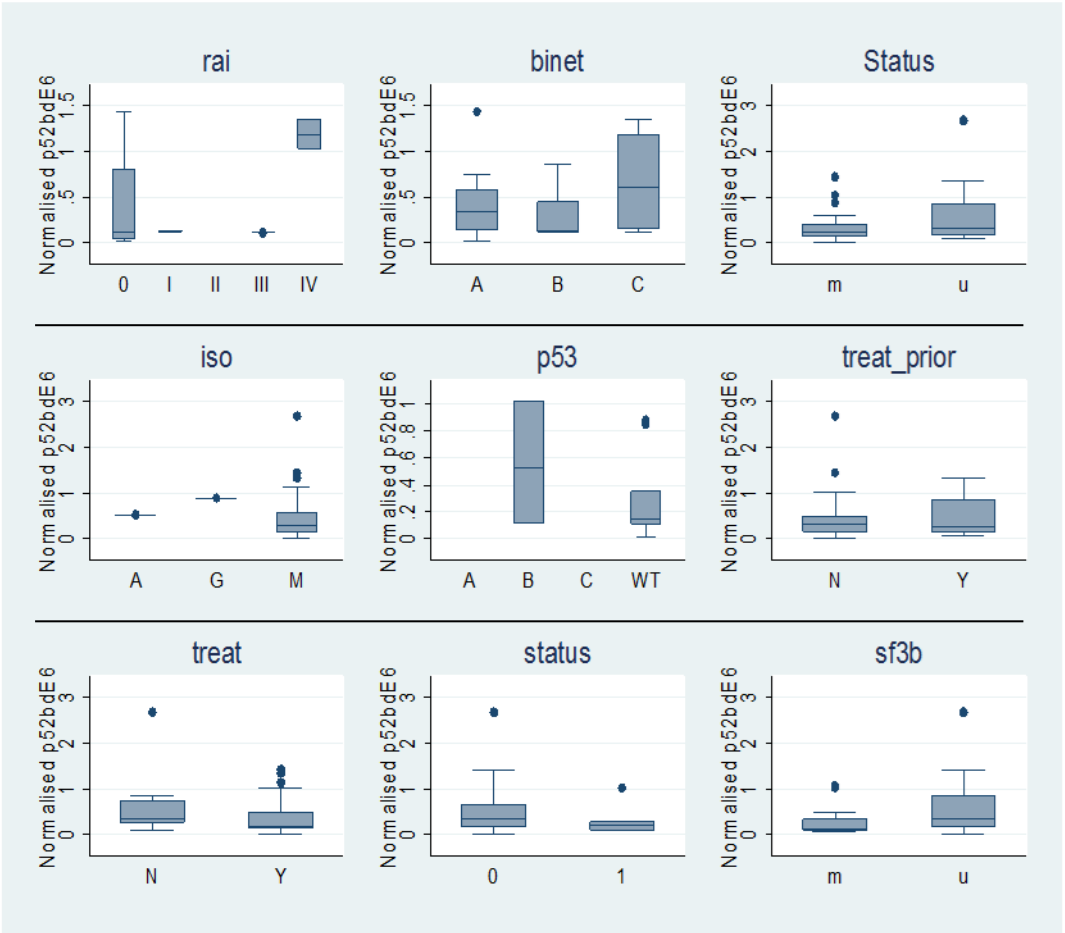
3. Relationship between LEDGF/p52b expression and clinical factors.




rai
Normalised p52b
0 1 2 3 4 5
0 I II III IV
binet
Normalised p52b
0 1 2 3 4 5
A B C
Status
Normalised p52b
0 2 4 6 8
m u
iso
Normalised p52b
0 2 4 6 8
A G M
p53
Normalised p52b
0 .05 .1 .15 .2
A B C WT
treat_prior
Normalised p52b
0 2 4 6
N Y
treat
Normalised p52b
0 2 4 6
N Y
status
Normalised p52b
0 2 4 6 8
0 1
sf3b
Normalised p52b
0 2 4 6 8
m u

4. Relationship between LEDGF/p52bΔE6 expression and clinical factors.




rai
Normalised p52bdE6
0
.5
1
1.5
0
I
II
III
IV
binet
Normalised p52bdE6
0
.5
1
1.5
A
B
C
Status
Normalised p52bdE6
0
1
2
3
m
u
iso
Normalised p52bdE6
0
1
2
3
A
G
M
p53
Normalised p52bdE6
0
.2
.4
.6
.8
1
A
B
C
WT
treat_prior
Normalised p52bdE6
0
1
2
3
N
Y
treat
Normalised p52bdE6
0
1
2
3
N
Y
status
Normalised p52bdE6
0
1
2
3
0
1
sf3b
Normalised p52bdE6
0
1
2
3
m
u

Appendix C

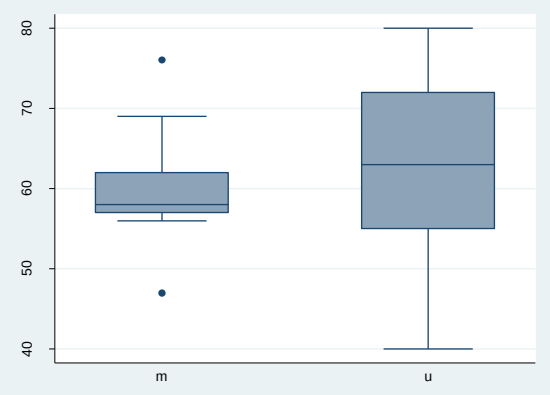
Exploratory analysis of potential relationships between the mutation status of SF3B1 gene and clinical factors associated with CLL.

1. The potential relationship between CD38 expression and SF3B1 mutation was investigated using a Fisher's Exact Test. There was no relationship between the two variables.

CD38 category	Status m	u	Total
Negative	6	8	14
Positive	0	4	4
Total	6	12	18

Fisher's exact = 0.245

2. The potential relationship between age and SF3B1 mutation status was investigated using a two-sided t-test. There was no relationship between the variables.



SF3B1 status	Age		
	Observations	Mean	Standard Deviation
mutated	9	60.44	8.23
unmutated	30	63.4	10.59
two sided t –test p-value = 0.39			

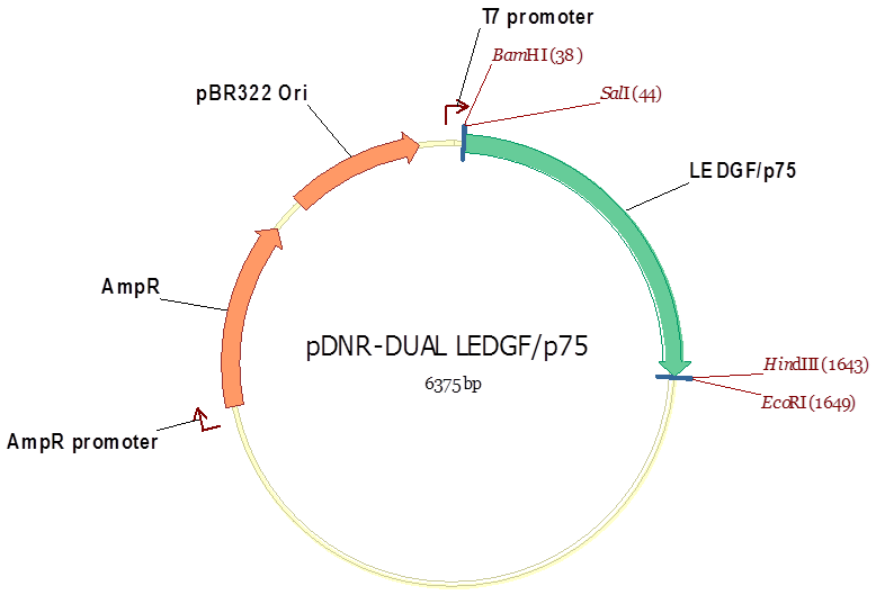
3. The potential relationship between the mutation status of the patient (dead or alive) and the mutation status of the SF3B1 gene was also investigated. There was no relationship between the variables.

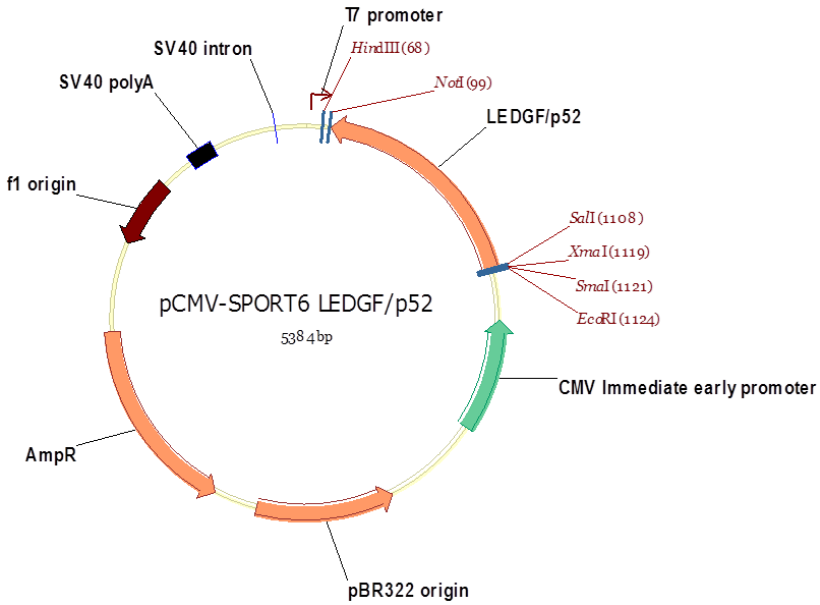
SF3B1 status	status Alive	Dead	Total
m	7	4	11
u	52	7	59
Total	59	11	70

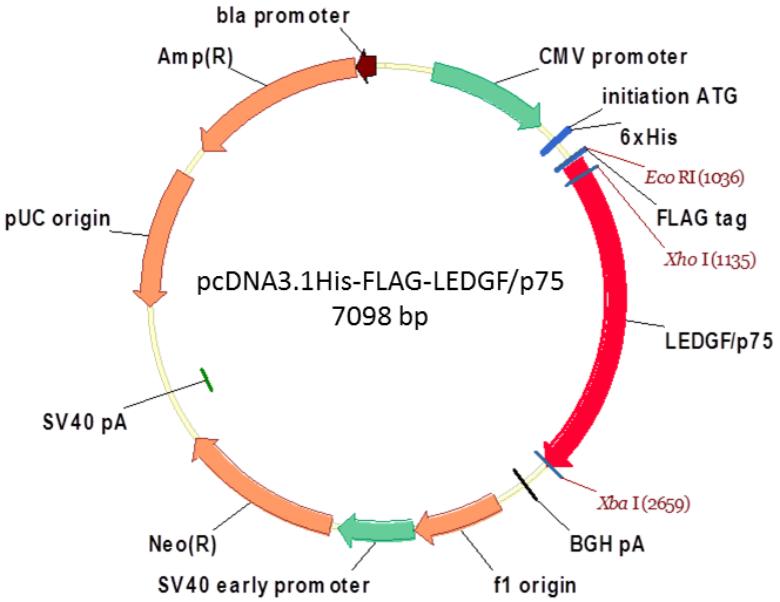
Fisher's exact = 0.063

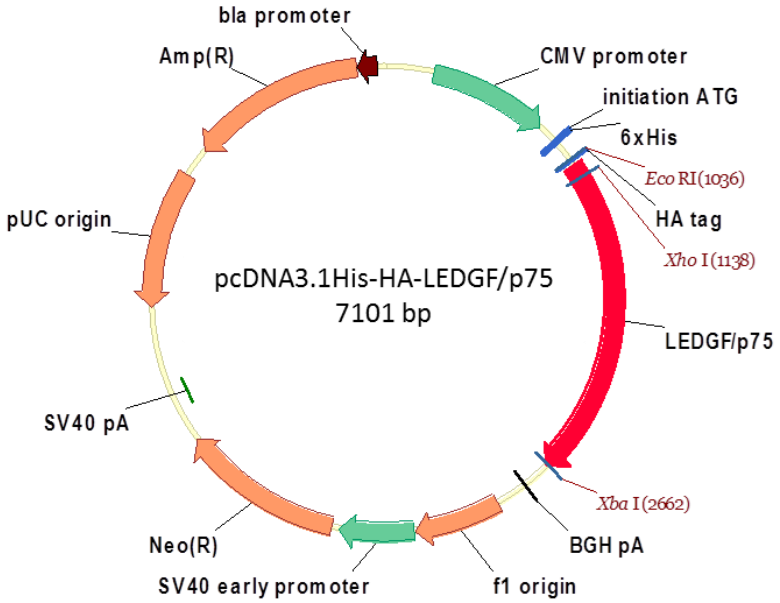
4. The relationship between other clinical variables and SF3B1 status was also examined but no relationship was detected in these cases. Factors investigated where white blood cell count, rai and binet stage, overall survival, if the patients had been treated, time-to-first-treatment and p53 status.

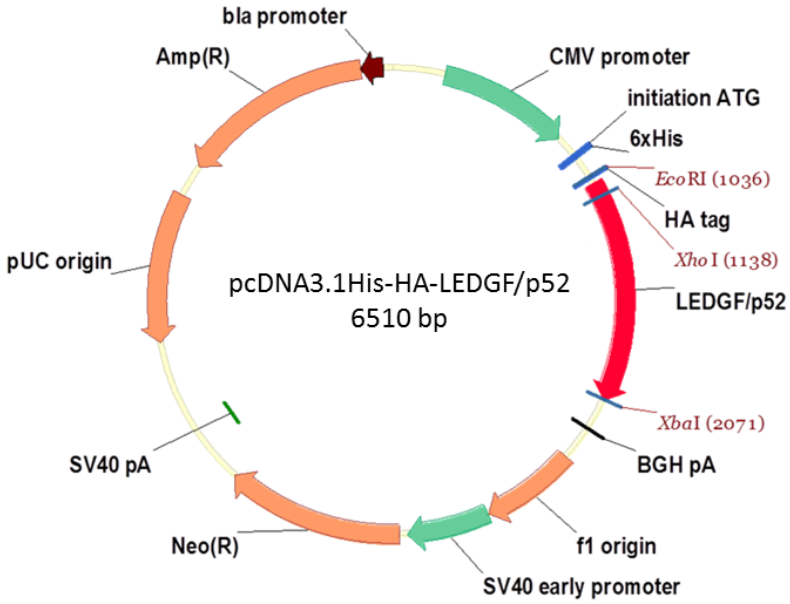
Appendix D

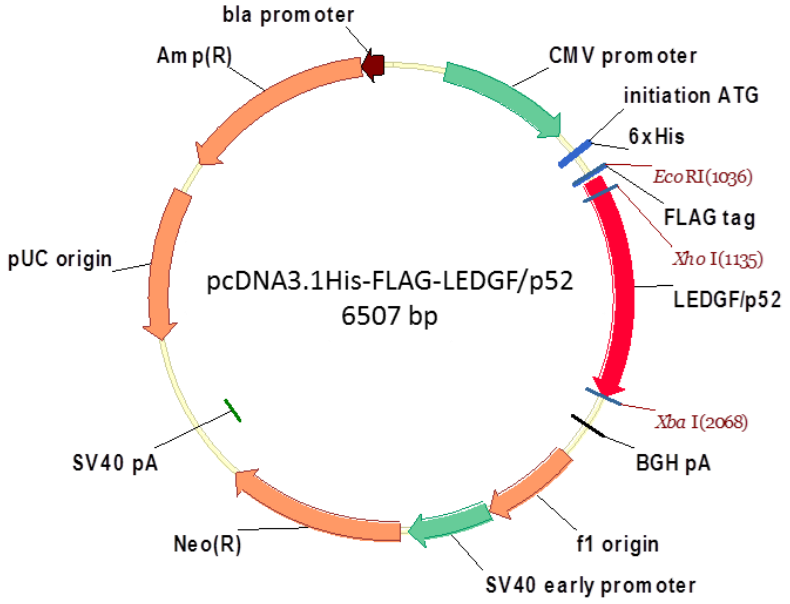

T7 promoter
BamHI(38)
SalI(44)
pBR322 Ori
LEDGF/p75
AmpR
pDNR-DUAL LEDGF/p75
6375bp
HindIII(1643)
EcoRI(1649)
AmpR promoter

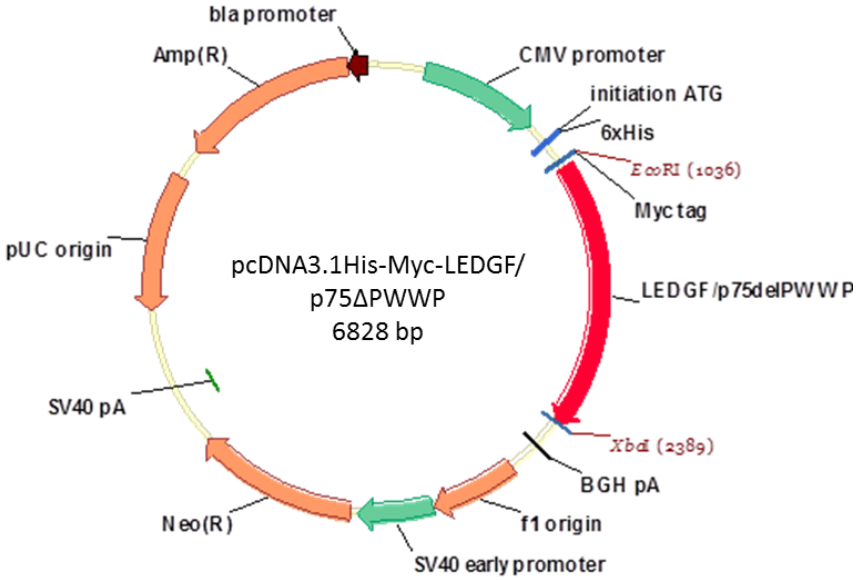

T7 promoter
SV40 intron
HindIII(68)
SV40 polyA
NotI(99)
LEDGF/p52
f1 origin
SalI(1108)
XmaI(1119)
SmaI(1121)
EcoRI(1124)
pCMV-SPORT6 LEDGF/p52
5384bp
CMV Immediate early promoter
AmpR
pBR322 origin

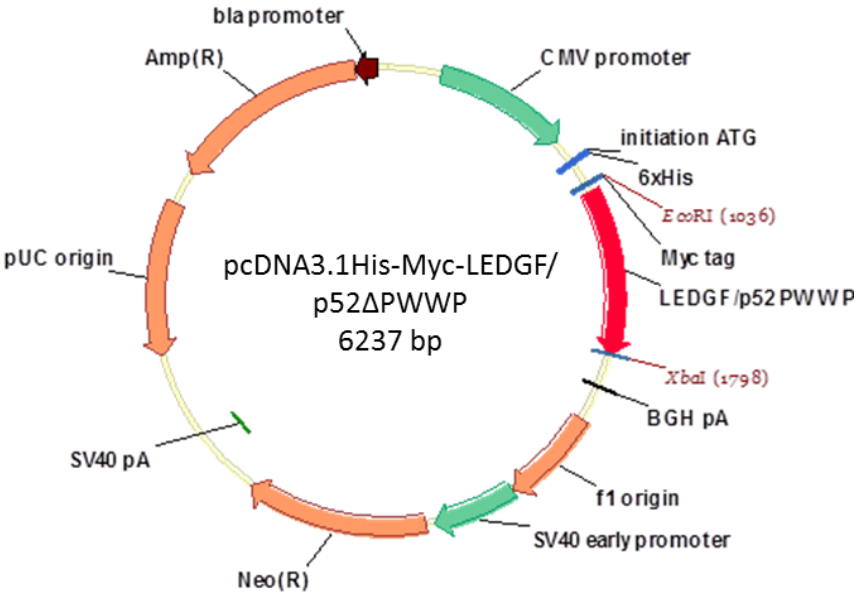

bla promoter
Amp(R)
CMV promoter
initiation ATG
6xHis
Eco RI(1036)
FLAG tag
Xho I(1135)
pUC origin
pcDNA3.1His-FLAG-LEDGF/p75
7098 bp
LEDGF/p75
SV40 pA
Xba I(2659)
Neo(R)
BGH pA
SV40 early promoter
f1 origin

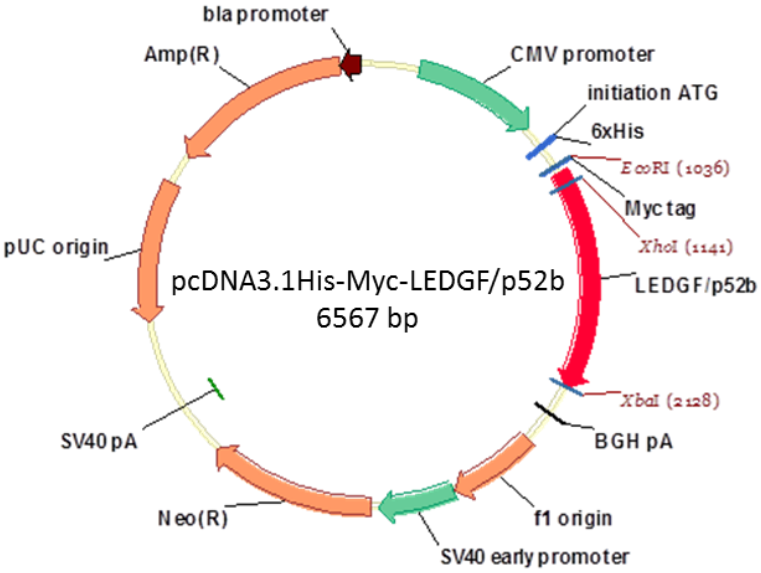

bla promoter
Amp(R)
CMV promoter
initiation ATG
6xHis
Eco RI(1036)
HA tag
Xho I(1138)
pUC origin
pcDNA3.1His-HA-LEDGF/p75
7101 bp
LEDGF/p75
SV40 pA
Xba I(2662)
Neo(R)
BGH pA
SV40 early promoter
f1 origin

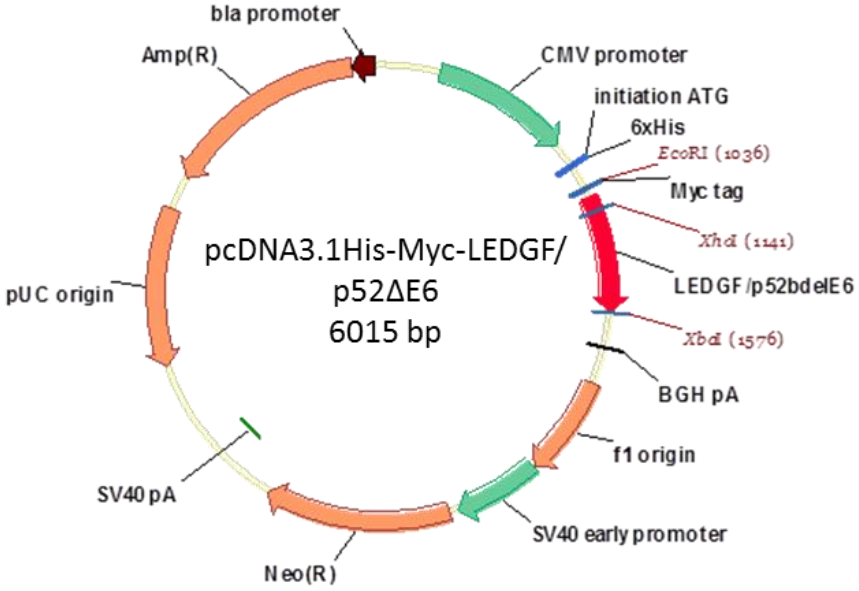

bla promoter
Amp(R)
CMV promoter
initiation ATG
6xHis
EcoRI (1036)
HA tag
XhoI (1138)
LEDGF/p52
XbaI (2071)
BGH pA
f1 origin
SV40 early promoter
Neo(R)
SV40 pA
pUC origin
pcDNA3.1His-HA-LEDGF/p52
6510 bp


bla promoter
Amp(R)
CMV promoter
initiation ATG
6xHis
EcoRI(1036)
FLAG tag
XhoI(1135)
LEDGF/p52
Xba I(2068)
BGH pA
f1 origin
SV40 early promoter
Neo(R)
SV40 pA
pUC origin
pcDNA3.1His-FLAG-LEDGF/p52
6507 bp


bla promoter
Amp(R)
CMV promoter
initiation ATG
6xHis
EcoRI (1036)
Myc tag
pUC origin
pcDNA3.1His-Myc-LEDGF/
p75ΔPWWP
6828 bp
LEDGF/p75delPWWP
SV40 pA
XbaI (2389)
BGH pA
Neo(R)
f1 origin
SV40 early promoter


bla promoter
Amp(R)
CMV promoter
initiation ATG
6xHis
EcoRI (1036)
pUC origin
pcDNA3.1His-Myc-LEDGF/
p52ΔPWWP
6237 bp
Myc tag
LEDGF/p52PWWP
XbaI (1798)
BGH pA
SV40 pA
f1 origin
SV40 early promoter
Neo(R)





Maps of plasmids that were utilised or generated in this study.

Bibliography

1. Singh DP, Ohguro N, Chylack LT, Jr., Shinohara T. Lens epithelium-derived growth factor: increased resistance to thermal and oxidative stresses. *Investigative ophthalmology & visual science* 1999 Jun; **40**(7)**:** 1444-1451.

2. Singh DP, Kimura A, Chylack LT, Jr., Shinohara T. Lens epithelium-derived growth factor (LEDGF/p75) and p52 are derived from a single gene by alternative splicing. *Gene* 2000 Jan 25; **242**(1-2)**:** 265-273.

3. Desfarges S, Hernàndez-Novoa B, Munoz M, Ciuffi A, Abderrahmani A. LEDGF/p75 TATA-less promoter is driven by the transcription factor Sp1. *Journal of molecular biology* 2011; **414**(2)**:** 177-193.

4. Bueno MT, Garcia-Rivera JA, Kugelman JR, Morales E, Rosas-Acosta G, Llano M. SUMOylation of the lens epithelium-derived growth factor/p75 attenuates its transcriptional activity on the heat shock protein 27 promoter. *Journal of molecular biology* 2010 Jun 4; **399**(2)**:** 221-239.

5. Stec I, Wright TJ, van Ommen GJ, de Boer PA, van Haeringen A, Moorman AF*, et al.* WHSC1, a 90 kb SET domain-containing gene, expressed in early development and homologous to a Drosophila dysmorphy gene maps in the Wolf-Hirschhorn syndrome critical region and is fused to IgH in t(4;14) multiple myeloma. *Human molecular genetics* 1998 Jul; **7**(7)**:** 1071-1082.

6. Eidahl JO, McKee CJ, Shkriabai N, Feng L, Plumb M, Kvaratskhelia M*, et al.* Structural basis for high-affinity binding of LEDGF PWWP to mononucleosomes. *Nucleic Acids Research* 2013; **41**(6)**:** 3924-3936.

7. Stec I, Nagl SB, van Ommen GJ, den Dunnen JT. The PWWP domain: a potential protein-protein interaction domain in nuclear proteins influencing differentiation? *FEBS letters* 2000 May 4; **473**(1)**:** 1-5.

8. Qiu C, Sawada K, Zhang X, Cheng X. The PWWP domain of mammalian DNA methyltransferase Dnmt3b defines a new family of DNA-binding folds. *Nature structural biology* 2002 Mar; **9**(3)**:** 217-224.

9. Shun MC, Botbol Y, Li X, Di Nunzio F, Daigle JE, Yan N*, et al.* Identification and characterization of PWWP domain residues critical for LEDGF/p75 chromatin binding and human immunodeficiency virus type 1 infectivity. *Journal of virology* 2008 Dec; **82**(23)**:** 11555-11567.

10. Dhayalan A, Rajavelu A, Rathert P, Tamas R, Jurkowska RZ, Ragozin S*, et al.* The Dnmt3a PWWP domain reads histone 3 lysine 36 trimethylation and guides DNA methylation. *The Journal of biological chemistry* 2010 Aug 20; **285**(34)**:** 26114-26120.

11. Vezzoli A, Bonadies N, Allen MD, Freund SM, Santiveri CM, Kvinlaug BT*, et al.* Molecular basis of histone H3K36me3 recognition by the PWWP domain of Brpf1. *Nature structural & molecular biology* 2010 May; **17**(5)**:** 617-619.

12. Wu H, Zeng H, Lam R, Tempel W, Amaya MF, Xu C*, et al.* Structural and histone binding ability characterizations of human PWWP domains. *PloS one* 2011; **6**(6)**:** e18919.

13. Musselman CA, Lalonde ME, Cote J, Kutateladze TG. Perceiving the epigenetic landscape through histone readers. *Nature Structural and Molecular Biology* 2012 Dec; **19**(12)**:** 1218-1227.

14. Zentner GE, Henikoff S. Regulation of nucleosome dynamics by histone modifications. *Nature Structural and Molecular Biology* 2013 Mar; **20**(3)**:** 259-266.

15. Waddington CH. Preliminary Notes on the Development of the Wings in Normal and Mutant Strains of Drosophila. *Proceedings of the National Academy of Sciences of the United States of America* 1939 Jul; **25**(7)**:** 299-307.

16. Holliday R. The inheritance of epigenetic defects. *Science (New York, NY)* 1987 Oct 9; **238**(4824)**:** 163-170.

17. Dietz F, Franken S, Yoshida K, Nakamura H, Kappler J, Gieselmann V. The family of hepatoma-derived growth factor proteins: characterization of a new member HRP-4 and classification of its subfamilies. *The Biochemical journal* 2002 Sep 1; **366**(Pt 2)**:** 491-500.

18. Tsutsui KM, Sano K, Hosoya O, Miyamoto T, Tsutsui K. Nuclear protein LEDGF/p75 recognizes supercoiled DNA by a novel DNA-binding domain. *Nucleic acids research* 2011 Jul; **39**(12)**:** 5067-5081.

19. Maertens G, Cherepanov P, Debyser Z, Engelborghs Y, Engelman A. Identification and characterization of a functional nuclear localization signal in the HIV-1 integrase interactor LEDGF/p75. *The Journal of biological chemistry* 2004 Aug 6; **279**(32)**:** 33421-33429.

20. Vanegas M, Llano M, Delgado S, Thompson D, Peretz M, Poeschla E. Identification of the LEDGF/p75 HIV-1 integrase-interaction domain and NLS reveals NLS-independent chromatin tethering. *Journal of cell science* 2005 Apr 15; **118**(Pt 8)**:** 1733-1743.

21. Llano M, Vanegas M, Hutchins N, Thompson D, Delgado S, Poeschla EM. Identification and characterization of the chromatin-binding domains of the HIV-1 integrase interactor LEDGF/p75. *Journal of molecular biology* 2006 Jul 21; **360**(4)**:** 760-773.

22. Aravind L, Landsman D. AT-hook motifs identified in a wide variety of DNA-binding proteins. *Nucleic acids research* 1998; **26**(19)**:** 4413-4421.

23. Turlure F, Maertens G, Rahman S, Cherepanov P, Engelman A. A tripartite DNA-binding element, comprised of the nuclear localization signal and two AT-hook motifs, mediates the association of LEDGF/p75 with chromatin in vivo. *Nucleic acids research* 2006; **34**(5)**:** 1653-1665.

24. Pradeepa MM, Sutherland HG, Grimes GR, Bickmore WA, Ule J. Psip1/Ledgf p52 binds methylated histone H3K36 and splicing factors and contributes to the regulation of alternative splicing. *PLoS Genetics* 2012; **8**(5).

25. van Nuland R, van Schaik FMA, Simonis M, van Heesch S, Cuppen E, Boelens R*, et al.* Nucleosomal DNA binding drives the recognition of H3K36-methylated nucleosomes by the PSIP1-PWWP domain. *Epigenetics & Chromatin* 2013; **6**.

26. Nishizawa Y, Usukura J, Singh DP, Chylack LT, Jr., Shinohara T. Spatial and temporal dynamics of two alternatively spliced regulatory factors, lens

epithelium-derived growth factor (ledgf/p75) and p52, in the nucleus. *Cell and tissue research* 2001 Jul; **305**(1)**:** 107-114.

27. Ge H, Si Y, Wolffe AP. A novel transcriptional coactivator, p52, functionally interacts with the essential splicing factor ASF/SF2. *Molecular cell* 1998; **2**(6)**:** 751-759.

28. Ge H, Si Y, Roeder RG. Isolation of cDNAs encoding novel transcription coactivators p52 and p75 reveals an alternate regulatory mechanism of transcriptional activation. *EMBO Journal* 1998; **17**(22)**:** 6723-6729.

29. Brown-Bryan TA, Leoh LS, Ganapathy V, Pacheco FJ, Mediavilla-Varela M, Filippova M*, et al.* Alternative splicing and caspase-mediated cleavage generate antagonistic variants of the stress oncoprotein LEDGF/p75. *Molecular cancer research : MCR* 2008 Aug; **6**(8)**:** 1293-1307.

30. Sutherland HG, Newton K, Brownstein DG, Holmes MC, Kress C, Semple CA*, et al.* Disruption of Ledgf/Psip1 results in perinatal mortality and homeotic skeletal transformations. *Molecular and cellular biology* 2006 Oct; **26**(19)**:** 7201-7210.

31. Llano M, Morrison J, Poeschla EM. Virological and cellular roles of the transcriptional coactivator LEDGF/p75. *Current topics in microbiology and immunology* 2009; **339:** 125-146.

32. Llano M, Saenz DT, Meehan A, Wongthida P, Peretz M, Walker WH*, et al.* An essential role for LEDGF/p75 in HIV integration. *Science (New York, NY)* 2006 Oct 20; **314**(5798)**:** 461-464.

33. Ciuffi A, Llano M, Poeschla E, Hoffmann C, Leipzig J, Shinn P*, et al.* A role for LEDGF/p75 in targeting HIV DNA integration. *Nature medicine* 2005 Dec; **11**(12)**:** 1287-1289.

34. Ferris AL, Wu X, Hughes CM, Stewart C, Smith SJ, Milne TA*, et al.* Lens epithelium-derived growth factor fusion proteins redirect HIV-1 DNA integration. *Proceedings of the National Academy of Sciences of the United States of America* 2010 Feb 16; **107**(7)**:** 3135-3140.

35. Christ F, Debyser Z. The LEDGF/p75 integrase interaction, a novel target for anti-HIV therapy. *Virology* 2013; **435**(1)**:** 102-109.

36. Yokoyama A, Cleary ML. Menin critically links MLL proteins with LEDGF on cancer-associated target genes. *Cancer cell* 2008 Jul 8; **14**(1)**:** 36-46.

37. Huang J, Gurung B, Wan B, Matkar S, Veniaminova NA, Wan K*, et al.* The same pocket in menin binds both MLL and JUND but has opposite effects on transcription. *Nature* 2012 Feb 23; **482**(7386)**:** 542-546.

38. Guru SC, Collins FS, Chandrasekharappa SC, Goldsmith PK, Lee Burns A, Marx SJ*, et al.* Menin, the product of the MEN1 gene, is a nuclear protein. *Proceedings of the National Academy of Sciences of the United States of America* 1998; **95**(4)**:** 1630-1634.

39. Chandrasekharappa SC, Guru SC, Manickam P, Olufemi SE, Collins FS, Emmert-Buck MR*, et al.* Positional cloning of the gene for multiple endocrine neoplasia-type 1. *Science (New York, NY)* 1997; **276**(5311)**:** 404-406.

40. Neff T, Armstrong SA. Recent progress toward epigenetic therapies: the example of mixed lineage leukemia. *Blood* 2013; **121**(24)**:** 4847-4853.

41. Zhang Y, Chen A, Yan XM, Huang G. Disordered epigenetic regulation in MLL-related leukemia. *International Journal of Hematology* 2012; **96**(4)**:** 428-437.

42. Liu H, Cheng EH, Hsieh JJ. MLL fusions: pathways to leukemia. *Cancer biology & therapy* 2009 Jul; **8**(13)**:** 1204-1211.

43. Thiel AT, Hua X, Huang J, Lei M. Menin as a hub controlling mixed lineage leukemia. *BioEssays* 2012; **34**(9)**:** 771-780.

44. Krivtsov AV, Armstrong SA. MLL translocations, histone modifications and leukaemia stem-cell development. *Nature Reviews Cancer* 2007; **7**(11)**:** 823-833.

45. Singh DP, Kubo E, Takamura Y, Shinohara T, Kumar A, Chylack LT, Jr.*, et al.* DNA binding domains and nuclear localization signal of LEDGF: contribution of two helix-turn-helix (HTH)-like domains and a stretch of 58 amino acids of

the N-terminal to the trans-activation potential of LEDGF. *Journal of molecular biology* 2006 Jan 20; **355**(3)**:** 379-394.

46. Shinohara T, Singh DP, Fatma N. LEDGF, a survival factor, activates stress-related genes. *Progress in retinal and eye research* 2002 May; **21**(3)**:** 341-358.

47. Cohen B, Addadi Y, Sapoznik S, Meir G, Kalchenko V, Harmelin A*, et al.* Transcriptional regulation of vascular endothelial growth factor C by oxidative and thermal stress is mediated by lens epithelium-derived growth factor/p75. *Neoplasia (New York, NY)* 2009 Sep; **11**(9)**:** 921-933.

48. Sapoznik S, Cohen B, Tzuman Y, Meir G, Ben-Dor S, Harmelin A*, et al.* Gonadotropin-regulated lymphangiogenesis in ovarian cancer is mediated by LEDGF-induced expression of VEGF-C. *Cancer research* 2009 Dec 15; **69**(24)**:** 9306-9314.

49. Wu X, Daniels T, Molinaro C, Lilly MB, Casiano CA. Caspase cleavage of the nuclear autoantigen LEDGF/p75 abrogates its pro-survival function: implications for autoimmunity in atopic disorders. *Cell death and differentiation* 2002 Sep; **9**(9)**:** 915-925.

50. Ha SY, Chan LC. Biphenotypic leukemia with t(9;11)(p22;p15). *Cancer genetics and cytogenetics* 1994 Sep; **76**(2)**:** 116-117.

51. Ahuja HG, Hong J, Aplan PD, Tcheurekdjian L, Forman SJ, Slovak ML. t(9;11)(p22;p15) in acute myeloid leukemia results in a fusion between NUP98 and the gene encoding transcriptional coactivators p52 and p75-lens epithelium-derived growth factor (LEDGF). *Cancer research* 2000 Nov 15; **60**(22)**:** 6227-6229.

52. Hussey DJ, Moore S, Nicola M, Dobrovic A. Fusion of the NUP98 gene with the LEDGF/p52 gene defines a recurrent acute myeloid leukemia translocation. *BMC genetics* 2001; **2:** 20.

53. Grand FH, Koduru P, Cross NC, Allen SL. NUP98-LEDGF fusion and t(9;11) in transformed chronic myeloid leukemia. *Leukemia research* 2005 Dec; **29**(12)**:** 1469-1472.

54. Morerio C, Acquila M, Rosanda C, Rapella A, Tassano E, Micalizzi C*, et al.* t(9;11)(p22;p15) with NUP98-LEDGF fusion gene in pediatric acute myeloid leukemia. *Leukemia research* 2005 Apr; **29**(4)**:** 467-470.

55. Lundin C, Horvat A, Paulsson K, Johansson B, Karlsson K, Olofsson T. t(9;11)(p22;p15) [NUP98/PSIP1] is a poor prognostic marker associated with de novo acute myeloid leukaemia expressing both mature and immature surface antigens. *Leukemia research* 2011; **35**(6)**:** e75-e76.

56. Yamamoto K, Yakushijin K, Funakoshi Y, Okamura A, Matsuoka H, Minami H*, et al.* Expression of the novel NUP98/PSIP1 fusion transcripts in myelodysplastic syndrome with t(9;11)(p22;p15). *European Journal of Haematology* 2012; **88**(3)**:** 244-248.

57. Huang TS, Myklebust LM, Kjarland E, Gjertsen BT, Pendino F, Bruserud O*, et al.* LEDGF/p75 has increased expression in blasts from chemotherapy-resistant human acute myelogenic leukemia patients and protects leukemia cells from apoptosis in vitro. *Molecular cancer* 2007; **6:** 31.

58. Ochs RL, Muro Y, Si Y, Ge H, Chan EKL, Tan EM. Autoantibodies to DFS 70 kd/transcription coactivator p75 in atopic dermatitis and other conditions. *Journal of Allergy and Clinical Immunology* 2000; **105**(6 I)**:** 1211-1220.

59. Krackhardt AM, Witzens M, Harig S, Stephen Hodi F, Jason Zauls A, Chessia M*, et al.* Identification of tumor-associated antigens in chronic lymphocytic leukemia by SEREX. *Blood* 2002; **100**(6)**:** 2123-2131.

60. Malavasi F, Deaglio S, Damle R, Chiorazzi N, Cutrona G, Ferrarini M. CD38 and chronic lymphocytic leukemia: A decade later. *Blood* 2011; **118**(13)**:** 3470-3478.

61. Hallek M. Signaling the end of chronic lymphocytic leukemia: new frontline treatment strategies. *Blood* 2013; **122**(23)**:** 3723-3734.

62. Bazargan A, Tam CS, Keating MJ. Predicting survival in chronic lymphocytic leukemia. *Expert review of anticancer therapy* 2012 Mar; **12**(3)**:** 393-403.

63. Wang L, Wan Y, Werner L, Zhang L, Zhang W, Sievers QL*, et al.* SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. *New England Journal of Medicine* 2011; **365**(26)**:** 2497-2506.

64. Jeromin S, Weissmann S, Haferlach C, Dicker F, Bayer K, Grossmann V*, et al.* SF3B1 mutations correlated to cytogenetics and mutations in NOTCH1, FBXW7, MYD88, XPO1 and TP53 in 1160 untreated CLL patients. *Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, UK* 2013.

65. Scott LM, Rebel VI. Acquired mutations that affect pre-mRNA splicing in hematologic malignancies and solid tumors. *Journal of the National Cancer Institute* 2013; **105**(20)**:** 1540-1549.

66. Tarn WY, Steitz JA. Pre-mRNA splicing: The discovery of a new spliceosome doubles the challenge. *Trends in biochemical sciences* 1997; **22**(4)**:** 132-137.

67. Golas MM, Sander B, Will CL, Lührmann R, Stark H. Molecular Architecture of the Multiprotein Splicing Factor SF3b. *Science (New York, NY)* 2003; **300**(5621)**:** 980-984.

68. Wang C, Chua K, Gozani O, Reed R, Seghezzi W, Lees E. Phosphorylation of spliceosomal protein SAP 155 coupled with splicing catalysis. *Genes and Development* 1998; **12**(10)**:** 1409-1414.

69. Gozani O, Reed R, Potashkin J. A potential role for U2AF-SAP 155 interactions in recruiting U2 snRNP to the branch site. *Molecular and cellular biology* 1998; **18**(8)**:** 4752-4760.

70. Quesada V, Ordóñez GR, Ramsay AJ, Puente DA, Velasco G, Freije JMP*, et al.* Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nature Genetics* 2012; **44**(1)**:** 47-52.

71. Rossi D, Bruscaggin A, Spina V, Rasi S, Fangazio M, Monti S*, et al.* Mutations of the SF3B1 splicing factor in chronic lymphocytic leukemia: Association with progression and fludarabine-refractoriness. *Blood* 2011; **118**(26)**:** 6904-6908.

72. Furney SJ, Marais R, Pedersen M, Turajlic S, Gentien D, Rapinat A*, et al.* SF3B1 mutations are associated with alternative splicing in uveal melanoma. *Cancer Discovery* 2013; **3**(10)**:** 1122-1129.

73. Rossi D, Rasi S, Spina V, Bruscaggin A, Monti S, Ciardullo C*, et al.* Integrated mutational and cytogenetic analysis identifies new prognostic subgroups in chronic lymphocytic leukemia. *Blood* 2013; **121**(8)**:** 1403-1412.

74. Cortese D, Sutton LA, Cahill N, Smedby KE, Geisler C, Gunnarsson R*, et al.* On the way towards a 'CLL prognostic index': focus on TP53, BIRC3, SF3B1, NOTCH1 and MYD88 in a population-based cohort. *Leukemia* 2013.

75. Rosenquist R, Cortese D, Bhoi S, Mansouri L, Gunnarsson R. Prognostic markers and their clinical applicability in chronic lymphocytic leukemia: Where do we stand? *Leukemia and Lymphoma* 2013; **54**(11)**:** 2351-2364.

76. Luco RF, Misteli T, Pan Q, Blencowe BJ, Tominaga K, Pereira-Smith OM. Regulation of alternative splicing by histone modifications. *Science (New York, NY)* 2010; **327**(5968)**:** 996-1000.

77. Luco RF, Misteli T, Allo M, Schor IE, Kornblihtt AR. Epigenetics in alternative pre-mRNA splicing. *Cell* 2011; **144**(1)**:** 16-26.

78. Luco RF, Misteli T. More than a splicing code: Integrating the role of RNA, chromatin and non-coding RNA in alternative splicing regulation. *Current Opinion in Genetics and Development* 2011; **21**(4)**:** 366-372.

79. Glenn M, Duckworth A, Kalakonda N. DNA Methylation Abnormalities in Mature B-cell Lymphoid Neoplasms. *Current Medical Literature: Leukemia & Lymphoma* 2012; **20**(4)**:** 109-118.

80. Daniels T, Zhang J, Gutierrez I, Elliot ML, Yamada B, Heeb MJ*, et al.* Antinuclear autoantibodies in prostate cancer: immunity to LEDGF/p75, a survival protein highly expressed in prostate tumors and cleaved during apoptosis. *The Prostate* 2005 Jan 1; **62**(1)**:** 14-26.

81. Daugaard M, Kirkegaard-Sorensen T, Ostenfeld MS, Aaboe M, Hoyer-Hansen M, Orntoft TF*, et al.* Lens epithelium-derived growth factor is an Hsp70-2 regulated guardian of lysosomal stability in human cancer. *Cancer research* 2007 Mar 15; **67**(6)**:** 2559-2567.

82. Basu A, Rojas H, Banerjee H, Cabrera IB, Perez KY, De Leon M*, et al.* Expression of the stress response oncoprotein LEDGF/p75 in human cancer: a study of 21 tumor types. *PloS one* 2012; **7**(1)**:** e30132.

83. Niu S, Chan R, Zou S, Berini P, Wang C. Morphology and expression status investigations of specific surface markers on B-cell chronic lymphocytic leukemia cells. *Microscopy Research and Technique* 2013; **76**(11)**:** 1147-1153.

84. Basu A, Drame A, Munoz R, Gijsbers R, Debyser Z, De Leon M*, et al.* Pathway specific gene expression profiling reveals oxidative stress genes potentially regulated by transcription co-activator LEDGF/p75 in prostate cancer cells. *The Prostate* 2011 Jul 27.

85. Singh DP, Ohguro N, Kikuchi T, Sueno T, Reddy VN, Yuge K*, et al.* Lens epithelium-derived growth factor: effects on growth and survival of lens epithelial cells, keratinocytes, and fibroblasts. *Biochemical and biophysical research communications* 2000 Jan 7; **267**(1)**:** 373-381.

86. Sharma P, Singh DP, Fatma N, Chylack LT, Jr., Shinohara T. Activation of LEDGF gene by thermal-and oxidative-stresses. *Biochemical and biophysical research communications* 2000 Oct 5; **276**(3)**:** 1320-1324.

87. Matsui H, Lin LR, Singh DP, Shinohara T, Reddy VN. Lens epithelium-derived growth factor: increased survival and decreased DNA breakage of human RPE cells induced by oxidative stress. *Investigative ophthalmology & visual science* 2001 Nov; **42**(12)**:** 2935-2941.

88. Gangemi S, Allegra A, Alonci A, Cannav A, Russo S, Musolino C*, et al.* Relationship between advanced oxidation protein products, advanced glycation end products, and S-nitrosylated proteins with biological risk and MDR-1 polymorphisms in patients affected by B-Chronic lymphocytic leukemia. *Cancer Investigation* 2012; **30**(1)**:** 20-26.

89. Mous K, Jennes W, De Roo A, Pintelon I, Kestens L, Van Ostade X. Intracellular detection of differential APOBEC3G, TRIM5alpha, and LEDGF/p75 protein expression in peripheral blood by flow cytometry. *Journal of immunological methods* 2011 Sep 30; **372**(1-2)**:** 52-64.

90. Deaglio S, Vaisitti T, Aydin S, Ferrero E, Malavasi F. In-tandem insight from basic science combined with clinical research: CD38 as both marker and key component of the pathogenetic network underlying chronic lymphocytic leukemia. *Blood* 2006; **108**(4)**:** 1135-1144.

91. Hamblin TJ, Orchard JA, Gardiner A, Oscier DG, Davis Z, Stevenson FK. Immunoglobulin V genes and CD38 expression in CLL [3] (multiple letters). *Blood* 2000; **95**(7)**:** 2455-2457.

92. Hahn CN, Scott HS. Spliceosome mutations in hematopoietic malignancies. *Nature Genetics* 2012; **44**(1)**:** 9-10.

93. Strefford JC, Sutton LA, Baliakas P, Cortese D, Cahill N, Mansouri L*, et al.* Distinct patterns of novel gene mutations in poor-prognostic stereotyped subsets of chronic lymphocytic leukemia: The case of SF3B1 and subset #2. *Leukemia* 2013; **27**(11)**:** 2196-2199.

94. Wan Y, Wu CJ. SF3B1 mutations in chronic lymphocytic leukemia. *Blood* 2013; **121**(23)**:** 4627-4634.

95. Martín-Subero JI, Campo E, López-Otín C. Genetic and epigenetic basis of chronic lymphocytic leukemia. *Current Opinion in Hematology* 2013; **20**(4)**:** 362-368.

96. Schoenberg DR, Maquat LE. Regulation of cytoplasmic mRNA decay. *Nature Reviews Genetics* 2012; **13**(4)**:** 246-259.

97. Alonso V, Friedman PA. Minireview: Ubiquitination-regulated G Protein-Coupled Receptor Signaling and Trafficking. *Molecular Endocrinology* 2013; **27**(4)**:** 558-572.

98. Mian SA, Smith AE, Kulasekararaj AG, Mohamedali AM, Ford K, Nasser E*, et al.* Spliceosome mutations exhibit specific associations with epigenetic

modifiers and proto-oncogenes mutated in myelodysplastic syndrome. *Haematologica* 2013; **98**(7)**:** 1058-1066.

99. Schnaiter A, Paschka P, Rossi M, Zenz T, Bühler A, Winkler D*, et al.* NOTCH1, SF3B1, and TP53 mutations in fludarabine-refractory CLL patients treated with alemtuzumab: results from the CLL2H trial of the GCLLSG. *Blood* 2013; **122**(7)**:** 1266-1270.

100. Foà R, Del Giudice I, Guarini A, Rossi D, Gaidano G. Clinical implications of the molecular genetics of chronic lymphocytic leukemia. *Haematologica* 2013; **98**(5)**:** 675-685.

101. Ge H, Roeder RG. Purification, cloning, and characterization of a human coactivator, PC4, that mediates transcriptional activation of class II genes. *Cell* 1994 Aug 12; **78**(3)**:** 513-523.

102. Cherepanov P, Maertens G, Proost P, Devreese B, Van Beeumen J, Engelborghs Y*, et al.* HIV-1 integrase forms stable tetramers and associates with LEDGF/p75 protein in human cells. *The Journal of biological chemistry* 2003 Jan 3; **278**(1)**:** 372-381.

103. Cherepanov P, Devroe E, Silver PA, Engelman A. Identification of an evolutionarily conserved domain in human lens epithelium-derived growth factor/transcriptional co-activator p75 (LEDGF/p75) that binds HIV-1 integrase. *The Journal of biological chemistry* 2004 Nov 19; **279**(47)**:** 48883-48892.

104. Maertens GN, Cherepanov P, Engelman A. Transcriptional co-activator p75 binds and tethers the Myc-interacting protein JPO2 to chromatin. *Journal of cell science* 2006 Jun 15; **119**(Pt 12)**:** 2563-2571.

105. Bartholomeeusen K, De Rijck J, Busschots K, Desender L, Gijsbers R, Emiliani S*, et al.* Differential interaction of HIV-1 integrase and JPO2 with the C terminus of LEDGF/p75. *Journal of molecular biology* 2007 Sep 14; **372**(2)**:** 407-421.

106. Bartholomeeusen K, Christ F, Hendrix J, Rain JC, Emiliani S, Benarous R*, et al.* Lens epithelium-derived growth factor/p75 interacts with the transposase-

derived DDE domain of PogZ. *The Journal of biological chemistry* 2009 Apr 24; **284**(17)**:** 11467-11477.

107. Hughes S, Jenkins V, Dar MJ, Engelman A, Cherepanov P. Transcriptional co-activator LEDGF interacts with Cdc7-activator of S-phase kinase (ASK) and stimulates its enzymatic activity. *The Journal of biological chemistry* 2010 Jan 1; **285**(1)**:** 541-554.

108. Méreau H, Kutz A, Juge S, Schwaller J, De Rijck J, Čermáková K*, et al.* Impairing MLL-fusion gene-mediated transformation by dissecting critical interactions with the lens epithelium-derived growth factor (LEDGF/p75). *Leukemia* 2013; **27**(6)**:** 1245-1253.

109. Leoh LS, Filippova M, Rios-Colon L, Basu A, Martinez SR, Tungteakkhun SS*, et al.* The stress oncoprotein LEDGF/p75 interacts with the methyl CpG binding protein MeCP2 and influences its transcriptional activity. *Molecular Cancer Research* 2012; **10**(3)**:** 378-391.

110. Chen C, Nott TJ, Jin J, Pawson T. Deciphering arginine methylation: Tudor tells the tale. *Nature Reviews Molecular Cell Biology* 2011; **12**(10)**:** 629-642.

111. Feng H, Xiang H, Mao YW, Wang J, Liu JP, Huang XQ*, et al.* Human Bcl-2 activates ERK signaling pathway to regulate activating protein-1, lens epithelium-derived growth factor and downstream genes. *Oncogene* 2004 Sep 23; **23**(44)**:** 7310-7321.

112. Billard C. BH3 mimetics: status of the field and new developments. *Molecular cancer therapeutics* 2013 Sep; **12**(9)**:** 1691-1700.