Molecular Genetic and DNA Methylation Profiling of Chronic Lymphocytic Leukaemia

A Focus on Divergent Prognostic Subgroups and Subsets

NICOLA CAHILL
Dissertation presented at Uppsala University to be publicly examined in Rudbecksalen, Rudbeck Laboratory, Dag Hammerskjöld’s väg 20, Uppsala. Friday, April 13, 2012 at 09:15 for the degree of Doctor of Philosophy (Faculty of Medicine). The examination will be conducted in English.

Abstract

Advancements in prognostication have improved the subdivision of chronic lymphocytic leukaemia (CLL) into diverse prognostic subgroups. In CLL, IGHV unmutated and IGHV3-21 genes are associated with a poor-prognosis, conversely, IGHV mutated genes with a favourable outcome. The finding of multiple CLL subsets expressing ‘stereotyped’ B-cell receptors (BCRs) has suggested a role for antigen(s) in leukemogenesis. Patients belonging to certain stereotyped subsets share clinical and biological characteristics, yet limited knowledge exists regarding the genetic and epigenetic events that may influence their clinical behaviour. This thesis aimed to, further investigate Swedish IGHV3-21-utilising patients, screen for genetic and DNA-methylation events in CLL subgroups/subsets and study DNA methylation over time and within different CLL compartments.

In paper I, IGHV gene sequencing of 337 CLL patients from a Swedish population-based cohort revealed a lower (6.5%) IGHV3-21 frequency relative to previous Swedish hospital-based studies (10.1-12.7%). Interestingly, this frequency remained higher compared to other Western CLL (2.6-4.1%) hospital-based cohorts. Furthermore, we confirmed the poor-outcome for IGHV3-21 patients to be independent of mutational and stereotypy status.

In paper II, genomic events in stereotyped IGHV3-21-subset #2, IGHV4-34-subset #4 and subset #16 and their non-stereotyped counterparts were investigated via SNP arrays (n=101). Subset #2 and non-subset #2 carried a higher frequency of events compared to subset #4. A high frequency of del(11q) was evident in IGHV3-21 patients particularly subset #2 cases, which may partially explain their poor-prognosis. In contrast, the lower prevalence of aberrations and absence of poor-prognostic alterations may reflect the inherent low-proliferative disease seen in subset #4 cases.

In papers III and IV, differential methylation profiles in IGHV mutated and IGHV unmutated patients were identified using DNA-methylation microarrays. CLL prognostic genes (CLLU1, LPL), tumor-suppressor genes (TSGs) (ABI3, WISP3) and genes belonging to TGF-β and NF-kB/TNFFR1 pathways were differentially methylated between the subgroups. Additionally, the re-expression of methylated TSGs by use of methyl and deacetyl inhibitors was demonstrated. Interestingly, analysis of patient-paired diagnostic/follow-up samples and patient-matched lymph node (LN) and peripheral blood (PB) cases revealed global DNA methylation to be relatively stable over time and remarkably similar within the different compartments.

Altogether, this thesis provides insight into the aberrant genomic and DNA methylation events in divergent CLL subgroups. Moreover this thesis helps distinguish the extent to which DNA methylation changes with respect to time and microenvironment in CLL.

Keywords: DNA methylation, Chronic lymphocytic leukemia, SNP, array, IGHV3-21, IGHV4-34

Nicola Cahill, Uppsala University, Department of Immunology, Genetics and Pathology, Rudbecklaboratoriet, SE-751 85 Uppsala, Sweden.

© Nicola Cahill 2012

ISSN 1651-6206
urn:nbn:se:uu:diva-168945 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-168945)
Dedicated to my family
List of Papers

This thesis is based on the following Papers, which are referred to in the text by their Roman numerals.


* Equal first authors

Reprints were made with the permission of the publishers
Disruption of DNA methylation regulatory mechanisms in tumourigenesis ................................................................................. 44
Use of DNA methylation markers in clinical routine .................. 45
DNA methylation technologies ...................................................... 46
Epigenetic based treatment ............................................................ 49

Thesis Aims ......................................................................................... 50

Materials & Methods ........................................................................ 51
Patient samples and patient characteristics .................................... 51
IG gene sequencing .......................................................................... 52
250K Affymetrix SNP array ............................................................. 52
Illumina Infinium HumanMethylation27 and HumanMethylation450 BeadChip array ................................................................. 54
Bioinformatic analysis of the Illumina human27 and 450 DNA methylation array data .............................................................. 56
Methylation-specific PCR analysis .................................................. 57
Real-time quantitative PCR analysis ............................................... 57
Bisulfite sequencing ......................................................................... 57
5-aza-2’-deoxycytidine and Trichostatin A treatment ...................... 58
Pyrosequencing ............................................................................... 58
Statistical analysis ............................................................................ 58

Results & Discussion ....................................................................... 60
Paper I: IGHV3-21 gene frequency in a Swedish population-based cohort .................................................................................. 60
Paper II: High density array screening for genomic events in stereotyped subsets of CLL ........................................................ 64
Paper III: Array-based methylation profiling in prognostic subsets of CLL .................................................................................. 67
Paper IV: DNA methylation screening of CLL samples over time and within different CLL compartments ..................................... 71

Conclusions ...................................................................................... 77
Paper I ............................................................................................. 77
Paper II ........................................................................................... 77
Paper III .......................................................................................... 77
Paper IV .......................................................................................... 78

Acknowledgements ......................................................................... 79

References ....................................................................................... 81
Abbreviations

5’aza 5-azacytidine
ABI3 ABI family member 3
aCGH Array comparative genomic hybridisation
ADORA3 Adenosine A3 receptor
AID Activation induced deaminase
AKT V-akt murine thymoma viral oncogene homolog
AML Acute myeloid leukaemia
AMP Adenosine monophosphate
ANGPT2 Angiopoietin 2
ATM Ataxia telangiectasia mutated
BAC Bacterial artificial chromosome
BCL10 B-cell CLL/lymphoma 10
BCL2 B-cell CLL/lymphoma 2
BCR B cell receptor
BMSCs Bone marrow stromal cells
C Constant
CARD15 Caspase recruitment domain family, member 15
CCL22 C-C motif chemokine 22
CD Cluster of differentiation
CD40L CD40 ligand
CDK Cyclin dependent kinase
CDR Complementarity determining region
CGH Comparative genomic hybridization
CHARM Comprehensive high-throughput arrays for relative methylation
CLL Chronic lymphocytic leukaemia
CLLU1 Chronic lymphocytic leukaemia up-regulated gene 1
CMV Cytomegalovirus
CNA Copy number alteration
CNAT Copy number analysis tool
CNN-LOH Copy number neutral loss of heterozygosity
CNV Copy number variation
CpG Cytosine-phosphate-guanine
CXCR4 C-X-C chemokine receptor 4
D Diversity
DAC 5-aza 2’-deoxycytidine
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAPK1</td>
<td>Death-associated protein kinase 1</td>
</tr>
<tr>
<td>ddNTPs</td>
<td>Dideoxy nucleotide triphosphates</td>
</tr>
<tr>
<td>DLEU7</td>
<td>Deleted in lymphocytic leukaemia 7</td>
</tr>
<tr>
<td>DLEU1</td>
<td>Deleted in lymphocytic leukaemia 1</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>DNA-PK</td>
<td>DNA dependent protein kinase</td>
</tr>
<tr>
<td>DNMT</td>
<td>DNA methyltransferase</td>
</tr>
<tr>
<td>DNMT3B</td>
<td>DNA methyltransferase 3</td>
</tr>
<tr>
<td>DNMT3L</td>
<td>DNA methyltransferase 3 ligand</td>
</tr>
<tr>
<td>EBV</td>
<td>Epstein Barr virus</td>
</tr>
<tr>
<td>ERK</td>
<td>Extracellular signal related kinase</td>
</tr>
<tr>
<td>ESCs</td>
<td>Embryonic stem cells</td>
</tr>
<tr>
<td>EXO</td>
<td>Exonuclease</td>
</tr>
<tr>
<td>EZH2</td>
<td>Enhancer of zeste homolog 2</td>
</tr>
<tr>
<td>FAS</td>
<td>TNF receptor superfamily, member 6</td>
</tr>
<tr>
<td>FISH</td>
<td>Fluorescent in-situ hybridisation</td>
</tr>
<tr>
<td>FOXE1</td>
<td>Forkhead Box E1</td>
</tr>
<tr>
<td>FWR</td>
<td>Framework region</td>
</tr>
<tr>
<td>GC</td>
<td>Germinal center</td>
</tr>
<tr>
<td>GLRB</td>
<td>Glycine receptor Beta</td>
</tr>
<tr>
<td>GRM7</td>
<td>Glutamate receptor metabotropic 7</td>
</tr>
<tr>
<td>H3K27</td>
<td>Histone 3 lysine 27</td>
</tr>
<tr>
<td>H3K4</td>
<td>Histone 3 lysine 4</td>
</tr>
<tr>
<td>H3K9</td>
<td>Histone 3 lysine 9</td>
</tr>
<tr>
<td>HELP</td>
<td>HpaII tiny fragment Enrichment by Ligation-mediated PCR</td>
</tr>
<tr>
<td>HIP1R</td>
<td>Huntingtin interacting protein 1 related</td>
</tr>
<tr>
<td>HPLC</td>
<td>High pressure liquid chromatography</td>
</tr>
<tr>
<td>IBTK</td>
<td>Inhibitor of Bruton agammaglobulinemia tyrosine kinase</td>
</tr>
<tr>
<td>IG</td>
<td>Immunoglobulin</td>
</tr>
<tr>
<td>IGH</td>
<td>Immunoglobulin heavy</td>
</tr>
<tr>
<td>IGHD</td>
<td>Immunoglobulin heavy diversity</td>
</tr>
<tr>
<td>IGHJ</td>
<td>Immunoglobulin heavy joining</td>
</tr>
<tr>
<td>IGHV</td>
<td>Immunoglobulin heavy variable</td>
</tr>
<tr>
<td>IGK</td>
<td>Immunoglobulin kappa</td>
</tr>
<tr>
<td>IGL</td>
<td>Immunoglobulin lambda</td>
</tr>
<tr>
<td>IL17RC</td>
<td>Interleukin 17 receptor C</td>
</tr>
<tr>
<td>IL2</td>
<td>Interleukin 2</td>
</tr>
<tr>
<td>IPF1</td>
<td>Pancreatic and duodenal homeobox 1</td>
</tr>
<tr>
<td>IWCLL</td>
<td>International workshop of chronic lymphocytic leukaemia</td>
</tr>
<tr>
<td>J</td>
<td>Joining</td>
</tr>
<tr>
<td>JAIRD2</td>
<td>Jumonji- and ARID-domain-containing protein</td>
</tr>
<tr>
<td>LC</td>
<td>Light chain</td>
</tr>
</tbody>
</table>
LEF1  Lymphoid enhancer binding factor 1
LINES  Long Interspersed Nuclear Elements
LMO2  LIM domain only 2
LN  Lymph node
LPL  Lipoprotein lipase
LPS  Lipo-polysaccharide
MAPK  Mitogen activated protein kinase
MBD2  Methyl-CpG-binding domain protein 2
MBL  Monoclonal B lymphocytosis
Mbp  Mega basepair
MCL-1  Induced myeloid leukaemia cell differentiation protein
MDM2  Mdm2 p53 binding protein homolog (mouse)
MDR  Minimal deleted region
me/ me2/ me3  mono/di/tri methylation
MEACS  Non-muscle myosin heavy chain IIA exposed apoptotic cells
MECP2  Methyl CpG binding protein 2
MeDIP  Methyl DNA immunoprecipitation
MI  Methylation index
miRNA  MicroRNA
MLL  Mixed-lineage leukaemia
MRE  methyl sensitive restriction enzyme
MSP  Methylation specific PCR
MYCN  V-myc myelocytomatosis viral related oncogene
MYD88  Myeoid differentiation primary response gene
MYF6  Myogenic factor 6
MZ  Marginal zone
n  Non-templated
NFkB  Nuclear factor kappa B
NGFR  Neural growth factor receptor
NHL  Non-Hodgkin lymphoma
NK  Natural killer
NLCs  Nurse like cells
NOTCH1  Notch homologue 1
OS  Overall survival
p  Palidromic
P2RY14  Purinergic receptor P2Y
PAK5  p21 protein (Cdc42/Rac)-activated kinase 7
PB  Peripheral blood
PBMC  Peripheral blood mononuclear cells
PC  Proliferation center
PCG7  Primordial germ cell 7
PDI  Protein disulfide isomerase
PI3K  Phosphoinositide-3-kinase
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLD1</td>
<td>Phospholipase D1</td>
</tr>
<tr>
<td>PPP1R3A</td>
<td>Protein phosphatase 1 regulatory (inhibitor) subunit 3A</td>
</tr>
<tr>
<td>PRAME</td>
<td>Preferentially expressed antigen in melanoma</td>
</tr>
<tr>
<td>PREs</td>
<td>Polycomb response elements</td>
</tr>
<tr>
<td>PRF1</td>
<td>Perforin 1</td>
</tr>
<tr>
<td>RAG</td>
<td>Recombination activating gene</td>
</tr>
<tr>
<td>RASGRP3</td>
<td>RAS guanyl releasing protein 3</td>
</tr>
<tr>
<td>RE</td>
<td>Restriction enzyme</td>
</tr>
<tr>
<td>RLGS</td>
<td>Restriction landmark genome scanning</td>
</tr>
<tr>
<td>RQ-PCR</td>
<td>Real time quantitative PCR</td>
</tr>
<tr>
<td>RSS</td>
<td>Recombination signal sequences</td>
</tr>
<tr>
<td>SAHA</td>
<td>Suberoylanilide hydroxamic acid</td>
</tr>
<tr>
<td>SAP</td>
<td>Shrimp alkaline phosphatse</td>
</tr>
<tr>
<td>SDF1</td>
<td>Stromal cell derived factor 1</td>
</tr>
<tr>
<td>SF3B1</td>
<td>Splicing factor 3b, subunit 1</td>
</tr>
<tr>
<td>SHM</td>
<td>Somatic hypermutation</td>
</tr>
<tr>
<td>SLC22A18</td>
<td>Solute carrier family 22, member 18</td>
</tr>
<tr>
<td>SLL</td>
<td>Small lymphocytic lymphoma</td>
</tr>
<tr>
<td>SNP</td>
<td>Single nucleotide polymorphism</td>
</tr>
<tr>
<td>Sp1</td>
<td>Specificity factor 1</td>
</tr>
<tr>
<td>T</td>
<td>Thymine</td>
</tr>
<tr>
<td>TBR</td>
<td>Eomesodermin</td>
</tr>
<tr>
<td>TBX3</td>
<td>T-box 3</td>
</tr>
<tr>
<td>TdT</td>
<td>Terminal deoxynucleotidyl transferase</td>
</tr>
<tr>
<td>TNF</td>
<td>Tumour necrosis factor</td>
</tr>
<tr>
<td>TP53</td>
<td>Tumour protein p53 gene</td>
</tr>
<tr>
<td>Trx</td>
<td>Thioredoxin</td>
</tr>
<tr>
<td>TSA</td>
<td>Tricostatin A</td>
</tr>
<tr>
<td>TSG</td>
<td>Tumour suppressor gene</td>
</tr>
<tr>
<td>TSS</td>
<td>Transcription start site</td>
</tr>
<tr>
<td>TTT</td>
<td>Time to treatment</td>
</tr>
<tr>
<td>TWIST-2</td>
<td>Twist homolog 2</td>
</tr>
<tr>
<td>USP</td>
<td>Unmethylation specific PCR</td>
</tr>
<tr>
<td>UTR</td>
<td>Untranslated region</td>
</tr>
<tr>
<td>V</td>
<td>Variable</td>
</tr>
<tr>
<td>VCAM1</td>
<td>Vascular cell adhesion molecule 1</td>
</tr>
<tr>
<td>VHL</td>
<td>von Hippel-Lindau</td>
</tr>
<tr>
<td>VLA4</td>
<td>Very late antigen 4</td>
</tr>
<tr>
<td>WHO</td>
<td>World health organisation</td>
</tr>
<tr>
<td>WISP3</td>
<td>WNT1 inducible signaling pathway protein 3</td>
</tr>
<tr>
<td>XPO1</td>
<td>Exportin 1</td>
</tr>
<tr>
<td>ZAP-70</td>
<td>Zeta-chain (TCR) associated protein kinase 70kDa</td>
</tr>
<tr>
<td>ZNF 540</td>
<td>Zinc finger protein 540</td>
</tr>
</tbody>
</table>
Introduction

Chronic lymphocytic leukaemia (CLL) is a malignant clonal disorder characterised by the accumulation of mature, functionally impaired neoplastic CD5⁺ B cells. Originating in the bone marrow these cells slowly progress to infiltrate the peripheral blood, lymph nodes and spleen. At both the clinical and biological level, CLL is highly heterogeneous in nature. This heterogeneity gives rise to highly divergent disease courses, ranging from an asymptomatic disease in some patients to an aggressive and rapidly fatal condition in others.¹ Determination of the events contributing to CLL heterogeneity are key to identifying new prognostic markers that in turn will lead to better subdivision of CLL patients, allowing better implementation of optimal treatment regimes.

Over the past twenty years, studies of the B cell receptor (BCR) molecule on CLL cells have denoted this molecule as a key player in CLL leukemogenesis and prognostication. For instance, the mutation status of the immunoglobulin heavy chain variable region gene (IGHV) has been established as one of the most reliable prognostic markers in CLL.²,³ Moreover, studies characterising the molecular structure of immunoglobulin (IG) gene rearrangements support the notion that certain CLL subsets sharing a distinct clinical outcome present with highly similar IG rearrangements that may recognise a common antigen and confer a growth advantage to the CLL clone.⁴⁻⁶ In certain cases, other factors such as recurrent genomic alterations are known to contribute to CLL pathogenesis,⁷ however knowledge of aberrant DNA methylation events in this disease is limited. Hence, this thesis will focus on the relationship between the BCR structure, global genomic and DNA methylation events that may contribute to CLL leukemogenesis in different prognostic subgroups of CLL patients.

Chronic lymphocytic leukaemia

Epidemiology and aetiology

Today, CLL represents ~30-40% of all leukaemias and is the most commonly encountered adult leukaemia in the Western world.⁸ Demonstrating a
gender bias, the incidence of CLL is almost twice that in men as in women.\textsuperscript{9,10} In Sweden alone, \textasciitilde500 cases are diagnosed annually.\textsuperscript{11} The median age of CLL diagnosis is \textasciitilde70 years, however a third of patients are younger than 60 years of age at diagnosis.\textsuperscript{10,11} First-degree relatives of patients with CLL can have a 7-8.5 fold increased risk of developing CLL and a 2-fold increase for other lymphoproliferative diseases, suggesting a shared genetic component in familial CLL cases.\textsuperscript{12-15}

To date, the aetiology of CLL is still largely unknown, however a number of linkage studies have implicated some susceptibility loci in CLL. For instance, recent investigations have identified susceptibility loci at 2q13 2q37.1, 6p25.3, 11q24.1, 15q23 and 19q13.32 thus providing the first evidence for the existence of common, low-penetrance susceptibility loci in sporadic CLL.\textsuperscript{16-18} Monoclonal B lymphocytosis (MBL), a condition characterised by the presence of \textasciitilde5.0x10\textsuperscript{9} clonal B cells per litre of peripheral blood (PB), is now suspected to represent the pre-clinical stage of CLL. MBL is common among the general population, occurring in 3-5\% of individuals over the age of 50 years.\textsuperscript{19-22} MBL can be categorised in to three immunophenotypes; i) CLL-like MBL, CD5 positive with low CD20 expression, ii) double positive CD5/CD20 MBL, resembling atypical CLL and iii) CD5 negative MBL.\textsuperscript{20,22} CLL-like MBL is the most frequent MBL immunophenotype encountered. Nevertheless, it must be kept in mind, that in the vast majority of MBL cases CLL will not ultimately emerge.\textsuperscript{21} Unlike other forms of leukaemia, only suggestive relationships between exposure to pesticides and certain solvents such as benzene have been indicated in CLL aetiology, thus the effect of hazardous environmental substances warrants further investigation.\textsuperscript{23,24}

CLL is thought to be driven by a multistep pathogenic process. This involves the evolution of CLL cells over time, gaining genomic and epigenetic alterations along their evolutionary path to become increasingly tumourigenic. At each step, CLL is driven by the combined contribution of intrinsic CLL defects with tumour promoting extrinsic microenvironmental factors to finally form fully fledged CLL. However, the specific nature and exact timing of initiating/transforming events in CLL pathogenesis remain elusive.\textsuperscript{25-27} Some of these elements contributing to CLL pathogenesis will be discussed in detail within the following sections.

### Diagnosis

The most characteristic feature of CLL is a peripheral blood B lymphocytosis of \textasciitilde5.0 x10\textsuperscript{9}/L.\textsuperscript{28} Morphologically, CLL lymphocytes are smaller than normal B lymphocytes (Figure 1). Furthermore, CLL cells appear to be more fragile resulting in the formation of characteristic "smudge" cells on a blood
film preparation. Interestingly, some reports claim that the amount of smudge cells found in routine analysis may in fact be an independent prognostic factor for CLL. Immunophenotypically, CLL monoclonal B cells typically express CD5, CD19, CD20, CD23 and faint levels of surface Ig (sIg). The clinical features of CLL such as lymphadenopathy and splenomegaly are generally indicative of later disease stages where these symptoms arise due to the accumulation of lymphocytes in the bone marrow (BM), spleen, lymph nodes (LN) and liver. Hypogammaglobulinemia is also common in late CLL disease and is associated with an increased susceptibility to infection.

Figure 1. May-Grünwald-Giemsa staining of CLL cells in the bone marrow.

Staging
In clinical practice, two staging systems are used at present, Rai and Binet staging, however both of these systems have a limited ability to predict the clinical course at an early stage of the disease. This is due to the fact that these systems are dependent on the appearance of clinical symptoms which often occur late in disease. At diagnosis, the majority of CLL patients are asymptomatic and it is only through routine testing that CLL is identified in these cases. Approximately 30-50% of these patients with an indolent CLL at diagnosis will progress, however predicting which patients will progress is uncertain when using such clinical staging systems. Therefore, the need for new highly sensitive, specific and easily attainable prognostic markers is warranted.
Infectious complications play a major role in the mortality of CLL patients. These complications come secondary to the immune defects associated with primary CLL disease and secondary to immunosuppression caused by certain treatments. In fact, about 80% of CLL patients will experience infectious complications at some point during their disease that can severely reduce patient quality of life. Moreover, 50-60% of patients will die due to infection. Some examples of commonly contracted infections are bactremias and pneumonia caused by *Streptococcus pneumoniae*, *Haemophilus influenzae*, *Legionella pneumophila*, *Staphylococcus aureus*, *Salmonella*, and others. Herpes viruses are also commonly encountered in CLL patients. For instance, Epstein-Barr virus (EBV) can be a severe complication in some CLL cases, moreover it is thought to be involved in the transformation of CLL to aggressive Richters syndrome.

**Treatment**

Currently, no outright cure for CLL exists, however various treatment strategies are employed. Asymptomatic CLL patients undergo a watch and wait approach, where treatment is only administered upon evidence of progressive or symptomatic disease. On disease progression, cyclophosphamide in combination with fludarabine is the first line treatment for CLL with the goal of long-term remission. Fludarabine, a purine analogue and cyclophosphamide, an alkylating agent, work jointly by inhibiting DNA synthesis, inhibiting replication and initiating cell death. Despite the initial success of this combination treatment, some patients may become refractory to such treatment. In these relapsed refractory cases, use of monoclonal antibodies such as anti-CD20 (rituximab) have shown improved response rates. In patients with aggressive CLL with poor prognostic markers such as del(17p), the use of anti-CD52 (Alemtuzumab) antibodies can be effective. Today, the only available potentially curative treatment is the use of allogenic stem cell transplantation; however this is solely considered for younger patients with unfavourable prognostic markers.

**Normal B cells and B cell development**

In order to comprehend CLL pathogenesis, it is fundamental to understand the function of normal B cells and the processes of normal B cell development, IG gene rearrangement and B cell communication with antigen. B cells are the main cellular components of the adaptive immune response that augment humoral immunity through the production of antibodies. These cells function as key immune surveyors that recognize foreign antigens and
generate specific antibody responses that are adapted to aid the elimination of specific invading pathogens. In order to produce antibodies that specifically bind antigens, the B cell must undergo a process of differentiation, a hierarchical pathway of development where each step is characterised by a specific pattern of cell surface markers. B cell maturation occurs in the bone marrow where stimulation from stromal cells and cytokines induce the differentiation of pro B cells expressing CD43, CD19 and CD10 from stem cells. Further maturation to the pre B cell stage denotes the earliest cell type synthesising a detectable Ig, the cytoplasmic μ heavy chain. This μ heavy chain associates with surrogate light chains and Igα and Igβ signal transducer molecules to form the pre BCR.

The next stage involves the production of the kappa or lambda light chains. These light chains complex with μ heavy chains to produce IgM on the surface of the B cell at the immature B lymphocyte stage. At this stage, these cells do not respond to antigen, in fact on encountering antigen these cells may enter apoptosis or an anergic state. The transition from immature to mature cells is marked by the co-expression of μ and δ heavy chains producing membrane bound IgM and IgD, and this is accompanied by mature functional competence. Subsequently, fully mature B cells enter the circulation and lymphoid organs where they await antigen activation. Upon antigen encounter, the naïve B cell initiates cell signaling resulting in growth and proliferation of the B cell. This creates an amplified clone of plasma cells that secrete the antigen-specific Ig. Furthermore, following activation, B cells leave a lasting impression in the form of memory cells. These cells persist in circulation to produce a more rapid immune response should they encounter future challenges by the same antigen.

Structure of the immunoglobulin molecule
The IG molecule is a membrane protein complex that is composed of two IG heavy chains and two IG light chains (Figure 2). Fundamental to the generation of a diverse IG repertoire is the variable (V) region of the IG molecule. This region is generated through a distinct process of gene recombination events at both the IG heavy chain (IGH) and light chain (LC) loci known as VDJ recombination (see below). The V region is characterized by conserved framework regions (FWR) integrated among highly variable complementarity determining regions (CDRs), CDR1, CDR2 and CDR3. In particular, the CDR3 is the most hypervariable region of the IG molecule that is formed through the joining of three IG gene segments, the V, diversity (D) and joining (J) gene segments. Thus, the CDR3 serves as the main player in determining IG specificity at the antigen binding groove. As such, IG constant regions (C) do not partake in antigen specificity; instead they constitute the isotype of the IG and determine the effector function of the IG.
These regions may bind complement and receptor proteins to augment processes such as lysis, opsonisation and cell de-granulation.\textsuperscript{48}

\textbf{Figure 2.} Immunoglobulin structure and VDJ recombination of the IGH locus.

**VDJ recombination**

As mentioned, the intricate process of VDJ recombination is a complex mechanism involving a series of sequential rearrangements that play key roles in IG diversity.\textsuperscript{48,54} In short, the V region of the heavy chain is created by assembling distinct variable (IGHV), diversity (IGHD) and joining (IGHJ) gene segments. There are 123-129 IGHV gene segments available for recombination, however only 38-46 of these are functional gene segments. Moreover, 23 IGHD and 6 IGHJ functional gene segments as well as 9 constant (IGHC) gene segments are available for rearrangement (Figure 2).\textsuperscript{52,55,56} Rearrangement of the IGH locus commences at the pro B cell stage. On maturing to the pre B cell stage, the rearranged IGH complexes with a $\mu$ constant region to form IgM on the B cell surface.\textsuperscript{48}

In addition, rearrangement of the kappa (IGK) and lambda (IGL) light chain loci initiates during the pre-B cell stage, however only one specificity will be presented (kappa or lambda) on the cell surface of the mature B cell. Unlike rearrangement events at the IGH locus, only the joining of distinct V and J gene segments occurs at the IGK and IGL loci. In total, 36-40 functional
IGKV, 5 IGKJ and one IGKC genes are available for recombination. At the IGL locus, 29-33 functional IGLV genes, 4-5 IGLJ and 4-5 IGLC genes are available for rearrangement. Due to the lack of D genes at the light chain loci, the level of diversity is less than that observed at the IGH locus.\textsuperscript{48,52,56}

In short, the VDJ recombination mechanism is initiated by double stranded DNA breaks introduced by two lymphocyte specific recombination activating gene (RAG) enzymes, RAG 1 and RAG2.\textsuperscript{57-59} These enzymes work by forming a complex that specifically recognises recombination signal sequences (RSS) that flank the coding V, D and J segments. These RSSs are made up of 3 distinct elements; a conserved heptamer and a nonamer separated by a spacer element of 12bp or 23bp.\textsuperscript{55,57,58} The length of this spacer element plays a key role in driving recombination specificity.\textsuperscript{60} This is explained by the 12/23 spacer rule that states that only pairs of dissimilar spacer RSSs are efficiently recombined, so that spacers of 12 nucleotides will only be recombined with spacers containing 23 nucleotides.\textsuperscript{54,61} In accordance with the 12/23 spacer recombination rule, a D gene segment will first combine with a J gene segment forming a DJ complex. Following DJ joining a V gene will join to the DJ segment forming a VDJ complex.\textsuperscript{48}

After the above recombination events have taken place, DNA repair mechanisms are employed to mend DNA breaks. DNA-dependent protein kinase complexes (DNA-PK) are recruited to the broken DNA ends.\textsuperscript{60} Here, the DNA-PK recruits several other proteins and DNA polymerases to the free ends. On aligning of the two free DNA ends, the DNA-PK complex recruits the enzyme terminal deoxynucleotidyl transferase (TdT).\textsuperscript{62} TdT works to add random non-templated (n) nucleotides to the free DNA ends thus generating further IG diversity. Alternatively, deletion of nucleotides through exonuclease activity can provide a further layer of diversity at these junctional regions. Finally, DNA polymerases may insert additional nucleotides as needed to make the two ends compatible for joining. A ligase IV enzyme finally links DNA strands on opposite ends of the break to each other, completing the joining process.\textsuperscript{57,62}

The germinal center reaction and somatic hypermutation

Secondary lymphoid organs such as the lymph node and spleen provide unique micro-environmental niches to allow naïve B cells the chance to encounter antigen, proliferate and produce specific antibodies in a process termed the germinal centre (GC) reaction. Within these lymphoid organs are highly specialized GC follicles that provide niches encompassing a network of antigen presenting cells, co-stimulator cells and cytokines. The GC reaction is initiated when the naïve B cell encounters antigen in the extra follicular spaces. Upon activation of the cell it moves to T helper cell zones. On
recognizing the antigen, T helper cells produce CD40L which binds to the B cell CD40 receptor to promote B cell activation.\textsuperscript{63} This induces the B cell to transform into a highly proliferative centroblast within the dark zone of the GC follicle.\textsuperscript{63} In an effort to generate antibodies with a higher affinity to the initiating antigen, the process of somatic hypermutation (SHM) refines the sequences of the BCRs by introducing single nucleotide mutations randomly into the IG genes.\textsuperscript{64-66} In fact, SHM is a highly specialized process that occurs at a rate $10^6$ times higher than spontaneous mutation.\textsuperscript{64,67} At this stage, B cells acquiring disadvantageous SHM patterns will die by apoptosis.\textsuperscript{68} However, those that acquire favourable antibody mutations move to the light zone of the follicle and are termed noncycling centrocytes.\textsuperscript{69} These centrocytes bind their cognate antigen with the help of follicular dendritic cells and T cells to differentiate into plasma or memory B cells.\textsuperscript{63,70} Although GCs are considered the principal sites capable of sustaining SHM, they are not thought to be the sole sites of such activity. It is thought that SHM may occur outside of the GC and independently of T cell help.\textsuperscript{71-74} For instance, it has been suggested that B cells contained within the extra follicular marginal zone (MZ) of the spleen can also undergo SHM.\textsuperscript{71-74}

SHM represents a second mode of IG diversification after VDJ recombination. The process of SHM allows a further increase in IG diversity and the production of IGs with a higher specificity,\textsuperscript{75} however, knowledge of the exact mechanisms governing SHM are incomplete. SHM is known to involve the activation of induced cytidine deaminase (AID), an enzyme that works to deaminate cytosine bases to uracil in single stranded DNA creating a U:G mismatch.\textsuperscript{75} The mismatch is recognised by the cell’s DNA repair enzymes that excise the uracil base, however, error prone polymerases that work to fill in this gap result in the formation of mutations at this site.\textsuperscript{75} These mutations are mainly substitution mutations; however insertions and deletions also occur. Mutations are more commonly encountered in the CDR sequences where certain hotspots within these regions appear to be targeted more often to other regions.\textsuperscript{76-80}

**Immunoglobulin diversity**

As mentioned, the vast variety of unique IG molecules making up the human Ig repertoire is partially given by the intricate processes of VDJ recombination, SHM and non-templated (n) and palindromic (p) nucleotide addition/excision at junctional borders. In fact, due to these processes, the likelihood of finding two identical BCRs is negligible. More specifically, when considering all permutations of V, D and J genes at the IGH locus, the likelihood of the same VDJ recombination event presenting in normal healthy persons is 1 in 6348. Taking into account the chances that individuals carrying the same heavy chain carry the same kappa or lambda rearrangement the
probability further diminishes to approximately 1 in 2.3 million. Finally, SHM events and the enzymatic addition and or removal of n and p nucleotides at junctional regions further reduces the probability of two healthy persons expressing identical BCRs to $2.3 \times 10^{12}$.

The CLL cell and CLL pathogenesis

On a blood film, CLL cells appear as small mature B lymphocytes containing a dense nuclear structure, partially aggregated chromatin and a narrow rim of cytoplasm. However, CLL cells have intrinsic BCR signalling defects that are suggested to be characterized in some cases by truncated CD79b Ig domains, defective assembly, glycosylation and folding of $\mu$ and CD79a chains. In other cases, somatic mutations in the cytoplasmic domain of CD79b have also been described. These impairments may result in a low surface Ig expression and an inability to engage in effective signal transduction, jeopardising the ability of the CLL cell to undergo apoptosis. Moreover, the high expression of the anti-apoptotic BCL2 gene family described in CLL and the up-regulation of the cell cycle arrest protein p27KIP may aid the accumulation of the CLL clone. Further propagated by continuous CLL cell proliferation in lymph nodes, CLL cells thrive and out-compete normal B cells.

Once released to the peripheral blood, it is thought that CLL cells continually re-circulate to the secondary lymphoid tissues, to receive growth and survival promotional signals from accessory cells, cytokines and antigens contained within the microenvironment. One theory proposes genetic and epigenetic alterations to act as early initiating/driving events in CLL transformation. In this model, microenvironmental extrinsic signals merely provide a support mechanism to the CLL cells, whereby antigen(s) is proposed to administer a chronic stimulation to BCR receptive CLL cells, propelling the clone along its tumorigenic path. In an opposing theory, antigen is suggested to act as the key culprit in CLL initiation, whereby long lived CLL cells acquire pathogenic genomic and epigenetic alterations as secondary hits over time. Despite the lack of knowledge about the order and sequence of such tumorigenic events, the proposed interplay between genomics, epigenomics and microenvironment suggests a complex multistep tumorigeneis model in CLL. However, key to this model is assessing the CLL cell of origin which to date remains elusive.

CLL cell of origin

Over the past two decades the perception of the candidate cell(s) of origin for CLL has changed. Initially it was believed that a naive CD5$^+$ B cell was
the originator of CLL. However, the discovery of divergent CLL subgroups (see prognostic markers section), IGHVmutated CLL, characterised by SHM of their IGHV genes and IGHV unmutated CLL, characterised by the lack of SHM at their IGHV genes altered this opinion. This finding lead to the hypothesis that unmutated CLL cells were derived from CD5+ naïve pre GC B cells, whereas mutated CLL derived from antigen experienced post GC memory B cells. In the recent past, it has been suspected that both mutated and unmutated CLL cases derive from antigen experienced B cells, as CLL cells were shown to have a similar gene expression profile to that of antigen experienced memory B cells. The candidate single cell origin of CLL is considered to be the marginal zone (MZ) B cell. The MZ B cell shares many CLL cell features such as, an active membrane phenotype, and BCRs encoded by either mutated or unmutated genes. Moreover, MZ B cells are also shown to demonstrate both poly and autoreactivity. Finally like CLL cells, these MZ B cells can respond to T cell dependent and independent antigens.

Nowadays however, the notion of a single cell of origin is being challenged. Given the fact that IGHV mutated and unmutated CLL have previously been shown to have rather similar gene expression profiles, it is now suspected that perhaps the activated state of CLL cells may mask larger gene expression differences between these subgroups. If this is the case, then conceivably it is more likely that mutated and unmutated CLL derive from two independent cell origins. Since CLL leukemogenesis is considered to follow a long stepwise process perhaps spanning several steps in B cell development, the prospect arises that transformation can occur anytime during B cell maturation once IG heavy chain is expressed. Theoretically, this opens up the possibility that CLL could be derived from multiple cell precursors. Today, it has been suggested that the propensity to generate CLL clonal B cells is already decided at the hematopoietic stem cell stage.

The microenvironment in CLL

In vitro CLL cell culturing studies have long implicated the microenvironment to play a crucial role in CLL survival. CLL cells in culture rapidly undergo apoptosis in the absence of vital survival factors which are now known to occupy niches within complex in-vivo micro-environmental networks. These networks involve active molecule cross-talk between CLL cells, accessory cells, activated T cells and a medley of soluble pro-survival chemokines and cytokines. These molecules communicate through ligand-receptor and adhesion molecule interactions to activate autocrine and exocrine signaling in CLL cells. Overall, microenvironmental interactions of CLL cells seem to mainly mirror those of normal lymphocytes. However, certain distinct microenvironmental interactions controlling CLL
migration do exist that are not seen in normal B cells. In addition to providing support to the CLL cell, the microenvironment is said to provide physical protection from cytotoxic drug damage, thus contributing to treatment resistance. CLL cell interactions within the bone marrow, peripheral blood and lymph node microenvironment are discussed below.

**Bone marrow**

In the bone marrow, CLL cells come into contact with bone marrow stromal cells (BMSCs). In vitro, CLL cells co-cultured with BMSCs have been shown to live longer. This longevity is partially attributed to the interactions of BMSC CD54 and CD106 integrins with CLL cell CD49d/CD11a or CD11b/CD18 integrins. This interaction is suggested to bring about protection from spontaneous apoptosis by increasing anti-apoptotic signals such as BCL2. Accompanying survival promoting events include the release of stromal cell chemokines such as, stromal cell derived factor 1 (SDF-1). SDF-1 production is constitutively produced by BM stromal cells in CLL patients. In response to SDF-1 release, CLL cells migrate towards these growth supporting accessory cells via their highly expressed CXCLR4 receptors. Once recruited, the stromal cells adhere to CLL cells via VCAM-1 to VLA4 integrin interaction to subsequently activate anti-apoptosis pathways.

**Peripheral Blood**

Unlike mononuclear cells in the peripheral blood of normal individuals, some cells in CLL patients regularly morph into large nurse like cells (NLCs) that adhere to leukemic cells in vitro. Similar to BMSCs, in-vitro co-culturing experiments have indicated the importance of NLC derived SDF-1 in the prevention of apoptosis through its interaction with CXCLR4 on CLL cells. It is thought that perhaps NLCs help counteract apoptosis through the activation of MAP kinase ERK1/2 pathways that increase the expression of anti-apoptotic proteins like MCL-1.

**The lymph node**

The lymph node represents a specialized micro-environmental niche for CLL cells since it is the proposed main site of antigen encounter, BCR signaling and proliferation. In the past, CLL was solely considered to be an accumulative disease driven by defects in apoptosis. Nowadays it is seen as a disease afflicted by a concomitant increase in proliferation of the leukemic clone. Up to 1% of the CLL clone is now known to proliferate daily, within specialized pseudo-follicle structures known as proliferation centers (PCs). Large para-immunoblasts and pro-lymphocytes within these structures highly express the proliferation marker Ki-67 compared to smaller lymphocytes in the vicinity of the follicle. These large cells are also known to have constitutive activation of the NFκB pathway, thus promoting
cell proliferation through tight regulation of NFκB target genes such as BCL-2 and MCL-1.\textsuperscript{116-118} In support of these findings, CLL cells isolated from LNs have been recently shown to have an increased expression of active BCR signaling and proliferation genes relative to cells from the blood or bone marrow.\textsuperscript{115}

In PCs, subpopulations of T cells, stromal cells, macrophages and endothelial cells provide a cocktail of signaling mediators. Normally, the B cell receives survival signals via T cell CD40-CD40L stimulation. In CLL, increased numbers of T cells concentrate round PCs in response to B cell CCL22 release.\textsuperscript{116,119,120} Together, secretion of T cell derived cytokines like IL-4 and accessory cell chemokines such as SDF-1 support the expansion of CLL clones by up regulating anti-apoptotic SURVIVIN and BCL2 regulators.\textsuperscript{97,120}

LN derived CLL cells are known to harbor higher levels of CD38, compared to PB derived CLL cells.\textsuperscript{121,122} Additionally, in-vitro studies show an increase in CD38 expression upon CLL activation\textsuperscript{123} thus proposing CD38 positivity as a characteristic of the CLL cell-LN microenvironment interaction. CD38+ cells have been reported to interact with CD31 on stromal cells to promote CLL proliferation.\textsuperscript{124,125} In fact, CD38+ cells are said to proliferate ~2X as fast as CD38- cells \textit{in vivo}.\textsuperscript{126} It is thought that CD38+ cells gain an enhanced migratory ability to home to favourable microenvironments due to an elevated responsiveness to the CD31 stromal ligand. Alternatively, given the strong association between CD38 positivity and increased CD31+ vascularisation in lymph nodes, it is further proposed that the microenvironment itself may induce CD38 expression.\textsuperscript{127,128} Given its function as an accessory BCR signaling molecule, it is suggested that CD38+ cells can strongly transduce signals from their BCR, unlike their negative counterparts.\textsuperscript{129,130} Altogether these circumstances have culminated to suggest that CD38+ cells constitute the fraction of cells primed to proliferate in CLL.

Evidence of antigen(s) involvement in CLL pathogenesis

Currently, the exact role of antigen(s) in CLL leukemogenesis is unknown. Deciphering whether antigen(s) play vital a part in CLL initiation by providing a proliferative stress or whether they contribute to CLL progression through chronic stimulation of the CLL clone is an active area of current research. Despite these uncertainties pertaining to time of antigen engagement in CLL pathogenesis, a plethora of circumstantial evidence supports the involvement antigen at some stage of CLL development. This evidence stems from the finding that compared to the normal repertoire; CLL shows a biased use of certain IGHV genes, namely IGHV1-69, IGHV3-7, IGHV3-23, IGHV3-21 and IGHV4-34.\textsuperscript{3,131,132} Additionally, IG sequence analysis has
described the presence of multiple CLL subsets with closely homologous or stereotyped BCRs. More specifically, these stereotyped subsets are characterised by an amino acid identity of ≥60% in the heavy chain CDR3 and restricted usage of similar IGHV-D-J genes and light-chain genes. In fact, later reports have evidenced that ~30% of CLL patients belong to subsets (>100 defined today) carrying ‘stereotyped’ CDR3 sequences on their heavy and light chains. Considering the likelihood of finding two B cell clones with identical BCRs is almost negligible, these findings are highly suggestive of antigen(s) involvement in leukemogenesis.

The molecular nature of antigen(s) in CLL

The specific nature of the antigen(s) involved in CLL is largely unknown. It is suspected that foreign antigens, auto-antigens and super-antigens or combinations thereof play a role in CLL leukemogenesis. The reactivity of CLL cells to antigen stimulation has been shown to differ depending on the mutational status of the BCR. For instance, unmutated BCRs have been shown to be polyreactive against such molecules as bacterial lipopolysaccharides (LPS). Moreover, unmutated BCRs have demonstrated autoreactivity against insulin and DNA molecules. In contrast, mutated BCRs appear to lose their polyreactivity upon acquiring mutations through SHM.

CLL BCRs including several stereotyped cases, have demonstrated homology to anti-DNA, anti-rheumatoid factor, and anti-cardiolipin auto-antibodies. Moreover, BCR-stereotyped CLL cells have been shown to bind intracellular autoantigen such as cytoskeletal proteins vimentin and oxidised low density lipoproteins. More recently, auto-antigen in the form of antigenic motifs on apoptotic blebs from dead cells have been implicated in CLL leukemogenesis. It is thought that the BCR can recognise these motifs leading to self stimulation and expansion of the CLL clone. For example, unmutated CLL monoclonal antibodies have been shown to specifically bind non-muscle myosin heavy chain IIA (MYHIIA9) exposed apoptotic cells (MEACS). This event has been particularly evident in ‘subset #6’ patients expressing IGHV1-69/IGHD3-16/J3 genes. In light of the latter finding, it is has been suggested that CLL clones may arise from dual functional B cells that maintain their ability as scavengers of apoptotic residues, whilst sustaining their ability to bind conserved bacterial cell motifs.

Interestingly, self antigens have been noted to bind certain IGHV3 and IGHV4 BCRs outside of the CDR3 region, instead binding to IG FWRs. Similarly, bacterial super-antigens have been shown to bind FWR1, FWR3 and CDR2 regions of some IGHV3 BCRs. Details of these proposed
superantigens are discussed below in relation to the IGHV gene family in which they are suspected to have a role.

IGHV3-21 and IGHV4-34 CLL subgroups: IG sequences, antigens and clinical implications

IGHV3-21 utilising patients

IGHV3-21 patients have been shown to demonstrate a poor clinical outcome and aggressive disease regardless of their IGHV mutation status.\(^4,132,134,144\) In recent years, studies of IGHV3-21 CLL have reported an abundance of evidence linking antigen involvement to IGHV3-21 pathogenesis.\(^4,6,132,134,145\) For instance, approximately 50% of CLL patients using the IGHV3-21 gene demonstrate stereotyped BCRs and these are denoted as subset #2 patients. More specifically, subset #2 IGHV3-21 patients carry stereotyped BCRs consisting of a conserved 9 amino acid long HCDR3 sequence.\(^134,145-147\) In a large proportion of cases, the amino acid sequence of these stereotyped HCDR3 regions are identical (ARDANGMDV). However, in some cases sequences may differ merely by 1-3 amino acids at different positions. Furthermore, in the majority of subset #2 cases, it is not possible to assign any D gene usage.\(^134,144,145,147\) Interestingly, IGHV3-21 stereotypy has also been suggested to influence clinical behavior. For instance, patients carrying stereotyped CDR3s are proposed to have a more progressive disease compared to non-stereotyped patients.\(^5,146\) Supporting this proposal is the finding that IGHV3-21 patients with homologous HCDR3s more frequently express high levels of the poor prognostic markers CD38 and ZAP70.\(^5,144,147\)

IGHV3-21 patients with both stereotyped and non-stereotyped BCRs have been reported to share an equally poor overall survival similar to that displayed by other IGHV unmutated cases (see prognostic markers section). Nevertheless, IGHV3-21 stereotypy has also been suggested to influence clinical behavior. For instance, patients carrying stereotyped CDR3s are proposed to have a more progressive disease compared to non-stereotyped patients.\(^5,146\) Supporting this proposal is the finding that IGHV3-21 patients with homologous HCDR3s more frequently express high levels of the poor prognostic markers CD38 and ZAP70.\(^5,144,147\)

IGHV3-21 sequences are predominantly border-line mutated. This finding can be partially explained by the under-targeting of SHM across all regions of IGHV3-21 sequences compared to other IGHV3 subgroup genes.\(^6\) Interestingly, IGHV3-21 subset #2 patients have been shown to have lower targeting of mutations across all IG regions except the HCDR2 compared to non-subset 2 IGHV3-21 sequences. That said, several recurrent amino acid changes have been observed among subset #2 cases. Remarkably, a serine deletion at HCDR2 codon 59 is commonly detected in stereotyped IGHV3-
21 CLL sequences. This deletion appears to be both a CLL and subset biased mutation. The finding of specific recurrent mutations in IGHV3-21 CLL further strengthens the role of antigen selection in CLL.

Interestingly, IGHV3-21 patients and in particular stereotyped patients, showed a strong tendency to retain germline configuration in the binding motif for Staphylococcal protein A, a suspected superantigen in CLL. However, the biologic and clinical implications of this finding (if any) remain unknown. Additionally, mutated IGHV3-21 subset #2 cases have been shown to bind the intracellular auto-antigen coflin-1, an actin binding protein localised with molecular complexes on the cell surface at apoptosis.

Finally, a geographic bias has been attributed to this poor prognostic IGHV3-21 gene. For instance, Scandinavia has demonstrated the highest IGHV3-21 gene frequency (10-13%) compared to other Southern European countries and North America (~3-4%). These findings suggest that aside from geographical bias, differences in ethnicity and/or environment may account for the varying frequency of the IGHV3-21 gene seen globally. Nevertheless, it is also possible that biases in sample selection can contribute to the varying IGHV3-21 gene frequencies detected around the globe.

IGHV4-34 utilising patients
IGHV4-34 gene usage is seen in approximately 8-10% of CLL patients. Two CLL subsets carrying stereotyped IGHV4-34 BCRs namely subset #4 and subset #16 have been described in CLL. Subset #4 is the most common stereotyped subset occurring at an overall frequency of ~1%. This subset is characterised by a 20 amino acid long CDR3 sequence and restricted usage of the IGKV2-30 light chain. Furthermore, these patients almost exclusively carry IGHV mutated genes. Remarkably, these restricted biological features are associated with an indolent disease course compared to non-stereotyped IGHV4-34 CLL patients displaying non-restricted IG features. The indolent disease course noted for these patients, may in part be explained by the fact that these cases are associated with low expression of the poor prognostic CD38 marker. Interestingly, subset #4 patients have a low median age at diagnosis. Moreover, these patients demonstrate a potential association with persistent infection by common herpes viruses such as EBV and CMV. Recently, subset #4 cases have been shown to display extensive intraclonal diversification suggesting a role of ongoing active antigen stimulation in these cases.
The second IGHV4-34 stereotyped subset, subset #16 is present at an overall frequency of 0.3%. This subset is characterised by mutated IGHV genes, a restricted IGKV3-20 light chain usage and a 24 amino acid long homologous CDR3 sequence. However, as this subset is not commonly encountered little is known about the clinical outcome for these patients.5

In the germline state IGHV4-34 BCRs appear to be inherently autoreactive and may bind to auto-antigens. However, since IG sequencing studies have revealed IGHV4-34 BCRs to be mainly mutated,6 it has been suggested that through SHM, IGHV4-34 BCRs can alleviate their auto-reactivity.155 More specifically, stereotyped IGHV4-34 cases demonstrate stereotyped amino acid sequence changes at their IGHV genes and these changes are largely subset biased. Stereotyped IGHV4-34 subsets #4 and #16 have distinctive SHM distribution patterns in their HCDRs and HFRs compared to their non-stereotyped counterparts. For instance, subsets #4 and #16 carrying basic lysine residues within their HCDR3 regions also incur stereotyped mutations in the HCDR1 region resulting in the introduction of negatively charged residues. The latter mutations are said to eliminate the potential DNA binding properties given by the positively charged HCDR3 region. Furthermore, it is noted that in stereotyped IGHV4-34 subsets the HFR1 motif conferring anti I/i reactivity is rather conserved compared to non-subset cases. Hence, these subsets hold the potential to bind the I/i blood group antigen or the B cell isoform of CD45 containing the N-acetyllactosamine antigenic determinant. Hence, through SHM, IGHV4-34 utilising B cells lose their auto-reactivity transitioning to a more safe state so that they can persist within the functional IG repertoire.6,92

In summary, the proposed stepwise model of CLL leukemogenesis suggests that the survival of the preleukemic clone depends on the interaction of CLL cells with the pro-survival microenvironment. Antigen(s) residing in this microenvironment are thought to provide transient or chronic stimulation to the CLL clone. Under this model, antigen stimulation, aberrant genetic alterations and epimutations work together in a stepwise fashion to create a thriving CLL clone. Transforming and progressive events may occur at currently unknown discrete B cell developmental stages or randomly depending on the type, efficiency and potency of promotional microenvironmental signals. A summary of the genetic and epigenetic alterations seen in CLL will be described within the following sections.
Prognostic markers in CLL
IGHV gene mutational status

In 1999, two independent studies demonstrated that the IGHV mutational status could categorise CLL into divergent clinical subgroups. Remarkably, the IGHV mutation status allowed the subdivision of CLL into two disease entities where patients with unmutated IGHV genes were shown to experience an inferior prognosis while, patients with mutated IGHV genes demonstrated a superior prognosis.\textsuperscript{2,3} More specifically, patients with IGHV unmutated genes have a more aggressive condition, including evidence of advanced progressive disease, atypical peripheral blood cell morphology, adverse cytogenetic features, clonal evolution, and resistance to therapy, than those with mutated IGHV genes.\textsuperscript{2,3} As mentioned, an exception to this rule is the presence of the IGHV3-21 gene (see the IGHV3-21 utilising patients section).

Today it is widely accepted that a 98% identity cut-off is best at making the clinically relevant distinction between mutated and unmutated CLL. CLL patients who carry IGHV genes with $\geq 98\%$ identity to germline are considered unmutated whereas patients with $<98\%$ identity to germline are denoted as mutated.\textsuperscript{2,3,156,157} Nevertheless, applying this arbitrary mathematical cut-off to a biological event must be interpreted with caution, particularly when dealing with borderline cases. Despite the advantages of this robust prognostic marker, the methods used to obtain IGHV gene sequences can be laborious, thus the search for more accessible prognostic markers conducive to routine procedures are well sought after.

Protein based markers: CD38 and ZAP70 expression

CD38 expression was the first prognostic factor found to correlate with IGHV mutational status, where high expression was shown to be associated with unmutated genes and a poor prognosis.\textsuperscript{2,121,156,158} However, it has been evidenced that this relationship does not always hold true.\textsuperscript{156,159,160} Another issue preventing the use of CD38 expression as a substitute marker for mutational status is the lack of a definitive threshold delineating what percentage of the CLL clone should possess elevated CD38 expression in order to assign CD38 positivity.\textsuperscript{158,161} Moreover, CD38 has been reported to vary over time.\textsuperscript{121,159} Nevertheless, CD38 expression is considered an independent prognostic marker, albeit not the strongest.\textsuperscript{162}

Similarly, ZAP70 expression shows a good correlation to IGHV mutational status. In general, the majority of mutated cases are shown to be ZAP70 negative, whereas unmutated cases are mainly shown to be positive.\textsuperscript{163} Nev-
Nevertheless, Rassenti et al demonstrate that approximately 23% of patients show discrepant results that do not correlate with mutational status. Discordant cases may frequently present with other poor prognostic features such as del(17p), del(11q) and/or IGHV3-21 expression. For instance, 17p and 11q deleted patients have been found to have a low expression whereas IGHV3-21 patients have been shown to have a high expression. Interestingly, Rassenti et al have further shown ZAP70 expression to be a better predictor of time to progression compared to IGHV mutational status. Another advantage of ZAP70 measurement is the fact that ZAP-70 levels also appear to remain stable over time.

ZAP70 measurement holds many advantages in that it can be easily measured by readily available methods such as flow cytometry at the protein level and RQ-PCR at the gene level. ZAP70 is expressed by other immune cells such as natural killer (NK) and T cells which may influence ZAP70 expression analysis. Therefore, it is required that CLL tumour B cells be purified away from such cells prior to ZAP70 expression analysis. Hence, the inconvenience of having to perform cell purification prior to analysis makes measurement of ZAP70 expression non conducive for use in clinical routine.

RNA based markers: \(LPL\) and \(CLLU1\) expression

Today, Lipoprotein lipase \((LPL)\) gene expression serves as an independent prognostic marker in CLL. Over a decade ago, gene expression studies identified \(LPL\) to be differentially expressed in IGHV mutated and unmutated CLL. Although not highly expressed in CLL, \(LPL\) expression is higher in IGHV unmutated patients relative to IGHV mutated CLL patients. Expression further correlates with other negative prognosticators such as high CD38, ZAP-70 expression and poor genetic alterations. Predictive of a shorter treatment free survival and in most cases a shorter overall survival, \(LPL\) expression in some instances has been deemed as a superior independent prognosticator compared to IGHV mutational status. Moreover, its prognostic value is maintained independently of clinical stage. Very low \(LPL\) expression in normal peripheral blood mononuclear cells (PBMCs) permits the use of non-sorted PBMCs from CLL patients for analysis. Altogether, quantification of \(LPL\) expression by RQ-PCR may offer a cheaper, less labour intensive method to identify prognostic subgroups of CLL patients.

Differential display screening first identified CLL upregulated gene 1 \((CLLU1)\) to be differentially expressed in CLL. \(CLLU1\) is highly expressed in IGHV unmutated patients relative to mutated patients and is associated with other poor risk characteristics namely advanced clinical stage, poor cytogenetics and high CD38 and ZAP70 expression. Like \(LPL\), high
CLLU1 expression correlates to a shorter overall survival and treatment free survival.\textsuperscript{170,175,177} Recently, the value of CLLU1 expression in IGHV mutated patients has revealed important prognostic information, indicating those IGHV mutated patients with high CLLU1 expression to have a worse prognosis relative to mutated patients with low expression.\textsuperscript{176} Additionally, CLLU1 expression appears to be exclusive to CLL cells and is stable over time.\textsuperscript{177} Hence, determination of CLLU1 expression levels can be used as a reliable indicator of tumour burden and is suggested to complement currently used techniques for minimal residual disease (MRD) monitoring in CLL patients post therapy.\textsuperscript{178}

Genomic alterations

CLL is not defined by a single recurrent genomic alteration but rather by genomic instability that frequently culminates into chromosomal aberrations which are present in up to 80\% of patients. Today, fluorescent \textit{in-situ} hybridization (FISH) techniques are implemented to detect the most common chromosomal alterations namely del(13q), del(11q), del(17p) and trisomy 12, which are of important prognostic significance in CLL (Table 1).\textsuperscript{7}

Deletion 13q

Deletion at chromosome 13q14 is the most frequently encountered alteration affecting up to 55-60\% of CLL patients (Table 1).\textsuperscript{5} Deletion of 13q most often occurs as the sole alteration and is associated with a favourable prognosis and indolent disease course.\textsuperscript{7} However, the prognostic significance of mono-allelic versus bi-allelic deletions of 13q still remains controversial. Recently, a study revealed that a similar time to initial treatment and overall survival was attained in both mono- and bi-allelic del(13q) CLL patient groups.\textsuperscript{179} However, conflicting reports have suggested that bi-allelic disease is associated with a tendency for better overall survival compared to mono-allelic disease.\textsuperscript{180} That said, a study by Chena \textit{et al.}\textsuperscript{181} has shown bi-allelic del(13q) patients to have a shorter treatment free survival relative to heterozygous patients. Moreover, copy number neutral loss of heterozygosity (CNN-LOH) events involving the loss of one allele and the subsequent duplication of the alternative allele have been reported in CLL.\textsuperscript{180} Interestingly, these events have been reported on 13q, particularly in patients showing bi-allelic deletions of 13q.\textsuperscript{180} These observations suggest that the recurrence of CNN-LOH at chromosome 13 is an important mechanism in CLL biology.

It is further argued that detection of del(13q) is not fully sufficient to define a CLL with a good prognosis. For instance, the size of the clone carrying this alteration may influence the disease course. Recent reports reveal that patients harboring a higher percentage of del(13q) cells among their CLL clone have a worse clinical outcome compared to those with a lesser percentage of
Further, evidence is given by recent SNP profiling studies demonstrating the heterogeneity of 13q deletions in terms of alteration size, the genes they cover and the location of their breakpoints. For instance, in most 13q deletion cases a minimally deleted region (MDR) covering DLEU1, DLEU2, miRNA-15a and miRNA-16-1 genes is noted. Interestingly, a study by Ouillette et al found that patients with del(13q) covering the MDR and the RB gene positioned beyond the MDR region were associated with a higher Rai stage compared to del(13q) patients not including the RB gene. Moreover, patients with deletions covering the MDR in addition to the RB gene were more likely to be treated. Hence, it appears that genes outside of the MDR play key roles in hampering patient outcome.

In del(13q) CLL, it is suspected that deregulation of miRNA-15a/miRNA-16-1 and hence, deregulation of its downstream target the anti-apoptotic BCL2 gene, may confer a selective cumulative advantage. The roles of DLEU1 and 2 are less clear, however it is known that DLEU2 expression provides the primary transcript for production of miRNA-15a/miRNA-16-1. Like miRNA-15a/miRNA-16-1, the DLEU7 gene is also positioned within the MDR in 13q deleted patients. DLEU7 plays a key function as a potent NFkB inhibitor by interacting with members of the tumour necrosis factor (TNF) receptor family. Moreover, studies on lung cancer cell lines have shown DLEU7 expression to result in an increase in apoptosis. These observations suggest that loss of DLEU7 may work in an additive manner with the loss of miRNA-15a/miRNA-16-1 to augment the pathogenesis of del(13q) CLL.

A recent mouse study has provided in vivo evidence suggesting that the DLEU2/miRNA-15a/miRNA-16-1 locus controls B cell expansion by modulating proliferation rather than affecting survival via the anti-apoptotic BCL2 gene. This study suggests that the function of this locus may be to tightly regulate cell cycle entry in response to antigen. This is supported by the observation that DLEU2/miRNA-15a/miRNA-16-1 expression was down-regulated in proliferating B cells within germinal centers, thus establishing a tumor suppressor role for this locus. Moreover, deletion of the DLEU2/miRNA-15a/miRNA-16-1 cluster in mice was shown to cause the development of indolent B cell-autonomous, clonal lymphoproliferative disorders akin to the spectrum of CLL-associated phenotypes observed in humans. In addition, this study demonstrated that miRNA-15a/miRNA16-1-deletion in human and mouse leads to an accelerated B cell lymphoproliferative state that arises through modulating the expression of genes controlling cell-cycle progression.

Deletion 11q
Deletion of 11q is the second most common alteration in CLL (12-18%) and is associated with an aggressive disease course and a poor clinical outcome.
These patients are often characterized with abdominal or mediastinal lymphadenopathy. Molecular characterization of del(11q) has identified the ATM tumour suppressor gene as the most likely candidate gene in del(11q) CLL, since it is known to play a vital role in activating p53 and augmenting the DNA damage response. Due to the integral role of ATM in maintaining genome stability, mutation of ATM is strongly suggested to be involved in CLL pathogenesis. Furthermore, approximately one third of 11q deleted CLL patients display a somatic or germline ATM gene mutation. Similar to del(13q) alterations, a marked complexity in the size and structure of 11q deletions has been shown. For instance, a deletion of 11q can be discontinuous which is indicative of complex genomic events taking place. In some cases 11q deletions do not cover the ATM gene thus indicating the importance of alternative genes such as RBX, MLL and BCL1 in 11q deleted CLL cases.

**Trisomy 12**

Today trisomy 12 is considered the third most common genetic alteration in CLL occurring in 11-16% of patients (Table 1). Trisomy 12 is caused by the duplication of one of its chromosome pairs and is associated with an intermediate survival and shorter time to treatment. Little is known of the pathogenesis behind trisomy 12 CLL, however, chromosome banding studies have identified a partial trisomy 12q13-q15 in a minority of patients covering the MDM2 gene, a known gene deregulated in CLL and a negative regulator of p53. Another candidate gene known to be up regulated in CLL is the disease specific CLLU1 gene located at 12q22, however CLLU1 expression was not found to relate to trisomy 12. Nevertheless, new gene expression studies have provided further insight into trisomy 12 CLL pathogenesis. Recently, four genes have been shown to be significantly associated with trisomy 12 CLL, including overexpression of the HIP1R gene at 12q24 and the MYF6 gene at 12q21. In contrast, the P2RY14 gene at 3q21-q25 and CD200 at 3q12-q13 were shown to be under-expressed in trisomy 12 patients.

**Deletion 17p**

Deletion of 17p, covering the TP53 gene, occurs in up to 10% of CLL patients and is associated with a particularly poor survival and short time to treatment (Table 1). Moreover, deletion of 17p is commonly observed to co-exist with other unfavorable prognostic markers such as unmutated IGHV genes, high ZAP70 and CD38 expression. Additionally, del(17p) patients experience chemo-refractoriness as chemotherapeutic reagents often work by inducing apoptosis or impeding the cell cycle via the p53 pathway. Nevertheless, reagents which utilize p53 independent mechanisms such as alemtuzumab, an anti-CD52 monoclonal antibody can be implemented in these cases. Recent studies have shown that the majority of 17p deleted patients
present with TP53 mutations on their other allele, however it was also shown that a subset of CLL patients carry TP53 mutation without a deletion of 17p (Table 1).

In fact, patients solely carrying a 17p deletion or TP53 mutation alone have an equally poor overall survival similar to those patients carrying both del(17p) and TP53 mutation. Thus, p53 is an independent prognostic indicator that is important for steering treatment choices.

**Deletion 6q**

Today, approximately 6-7% of patients acquire del(6q) (Table 1). However, it is considered a progression marker as it is mainly detected in CLL patients at follow up. Deletions are most frequently encountered at 6q21, however deletions at 6q27 have also been detected. No definitive tumour suppressor genes have been accounted for in del(6q) CLL pathogenesis. The presence of del(6q) has been associated with an increased white blood cell count and in some cases a more extensive lymphadenopathy. Despite these associations, it does not appear to be a negative prognostic indicator in CLL. That said, some studies have associated 6q deletions with a shorter treatment free survival.

**Table 1. Summary of Genomic alterations in CLL.**

<table>
<thead>
<tr>
<th>Alteration</th>
<th>Del(17p)</th>
<th>TP53 mutation</th>
<th>del(11q)</th>
<th>Trisomy 12</th>
<th>del(13q)</th>
<th>Deletion 6q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>5-10%</td>
<td>1-5%</td>
<td>12-18%</td>
<td>11-16%</td>
<td>55%</td>
<td>6-7%</td>
</tr>
<tr>
<td>Candidate gene(s)</td>
<td>TP53</td>
<td>TP53</td>
<td>ATM</td>
<td>unknown</td>
<td>miR-15/16</td>
<td>unknown</td>
</tr>
<tr>
<td>Clinical outcome</td>
<td>poor</td>
<td>poor</td>
<td>poor</td>
<td>intermediate</td>
<td>good</td>
<td>progression</td>
</tr>
<tr>
<td>Use in clinical routine</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
</tbody>
</table>

**Translocations in CLL**

In the past, cumbersome classical cytogenetic methods used to detect translocations in CLL were often unsuccessful (see the Genomic profiling technologies section). Hence, the frequency of CLL translocations was grossly underestimated at this time. Today, new and improved cytogenetic analysis has allowed the detection of translocations in up to 30% of CLL patients, however these are largely non-recurrent events. The presence of balanced and unbalanced translocations are linked to an unfavourable clinical outcome in CLL. Interestingly, recurring breakpoints in CLL translocations are shown to cluster in regions affected in other lymphomas, such as 1(q21q25), 1(p32p36), 2(p11p13), 6(p11p12), 6(p21p25), 13q14, 14q32, and
18q21. These regions are suspected to contain genes involved in CLL leukemogenesis. For instance the t(14;19)(q32;q13) translocation involves the juxtaposition of the transcriptional co-activator BCL3 gene to the IGH locus causing an increase in BCL3 expression.\textsuperscript{215,216}

**Genomic complexity**

Recently, a correlation between genetic complexity and clinical outcome has been described in CLL.\textsuperscript{217} For instance, a recent SNP study showed CLL patients carrying more than two alterations to have a poorer overall survival.\textsuperscript{217} Similarly, patients carrying a higher number of larger sized alterations have been associated with a worse clinical outcome compared to those with smaller alterations.\textsuperscript{180} Additionally, the genomes of patients carrying deletions of 11q and 17p are shown to be more complex. For instance, a recent study demonstrated that patients with four or more alterations greater than 5Mbs carried del(17p) or del(11q). Hence, due to the loss of vital genes integral to genome stability, the genomes of 11q and particularly 17p deleted patients are more complex and thus more likely to acquire further alterations compared to other CLL patients.\textsuperscript{180}

**Novel alterations**

Reports have indicated del(14q) as a recurrent alteration in CLL occurring in ~1.5% of patients.\textsuperscript{218} It is has been shown that this aberration most often occurs in IGHV unmutated patients.\textsuperscript{218} Furthermore, del(14q) patients were shown to have a high occurrence of trisomy 12. Deletion of 14q was associated with an intermediate survival and shorter time to treatment although low sample numbers were investigated. Deletions of 14q are heterogeneous in size; however, recurrent breakpoints have been documented.\textsuperscript{218} Due to the frequency of these alterations and their association with disease outcome, it is widely debated whether or not to include deletions of 14q to the established FISH panel for detection of known recurrent alterations in CLL.

Through the use of new microarray studies, some novel alterations such as gain of 2p, gain of 8q and loss of 4p, 8p and 22q have been identified,\textsuperscript{180,219-221} however no consensus exists regarding the frequency of these novel alterations. This is likely due to discrepancies in the array platforms employed and the study populations investigated. Interestingly, gain of 2p covering the MYCN oncogene was recently found in up to 28% of untreated Binet stage B/C patients.\textsuperscript{222} These patients also experienced an increased MYCN mRNA expression, suggesting MYCN to have a pathogenic role in these CLL cases.\textsuperscript{222} Additionally, a recent SNP-array study showed this alteration to occur frequently in del(11q) patients.\textsuperscript{180} In patients with a loss at chromosome 22q, the PRAME gene has been implicated as the candidate gene of interest.\textsuperscript{220} Finally, dic(17;18)(p11.2;p11.2) a dicentric chromosome alteration has been described as a novel recurring abnormality in CLL. This aber-
ration is associated with a complex karyotype, progressive disease, unmutated IGHV genes and an early age at diagnosis relative to other CLL patients.223

Today, a small percentage of trisomy 12 patients 1.6-5% are known to carry a coexisting trisomy on chromosome 19.224-226 More recent studies have shown this genetic profile to occur exclusively in isotype switched sIg positive CLL. Interestingly, this profile has been associated with the use of IGHV mutated genes and high CD38 expression. Furthermore, it has recently been discovered that these cases are often found to carry trisomy 18.227 Notably, trisomy 12 has been reported to be associated with CLL patients expressing IgG-positive stereotyped unmutated IGHV4-39/IGKV1-39 (1D-39) BCRs.228 In some instances, these cases have been shown to be associated with the t(14;19)(q32;q13) translocation. As mentioned above (in the translocations in CLL section), t(14;19)(q32;q13) causes the BCL-3 proto-oncogene to become juxtaposed next to the IGH locus resulting in overexpression of the BCL-3 protein and deregulation of NF-κB.174,215,229

With the advent of whole genome sequencing technologies a number of novel mutations have recently been discovered in CLL. Studies have identified repeated mutations in XPO1, a nuclear transport factor gene in ~2.4% of patients230,231 and NOTCH1, a ligand activated transcription factor gene in ~12.2% of CLL cases. NOTCH1 mutations are more frequently found in IGHV-unmutated/ZAP70+ cases compared to IGHV-mutated/ZAP70- patients. More specifically, NOTCH1 mutations are reported to occur in 18-20% of IGHV-unmutated/ZAP70+ patients and 4-7% of mutated/ZAP70 cases.231,232 In contrast, a frequency of 4% in IGHV-unmutated/ZAP70+ and 1.5% in mutated/ZAP70 cases has also been reported. NOTCH1 mutations in CLL have been frequently found to occur in trisomy 12 patients. Altogether, these findings suggest NOTCH1 activation to play a role in IGHV-unmutated/ZAP70+ trisomy 12 CLL.230 In addition to these genes, recurrent mutations of MYD88 and SF3B1 have been identified in 10% and 15% of CLL cases respectively. Interestingly, SF3B1 mutations are shown to occur in CLL patients carrying deletion 11q.233

Genomic profiling technologies

Karyotyping ascertains gross chromosomal features such as chromosome number, size and structure. Conventional cytogenetic staining techniques such as G, Q and R-banding pinpoint the chromosomal regions affected by genomic alterations, however, only large alterations (approximately from 5-10 Mb) across chromosome metaphase spreads are analysed.234 This method requires cumbersome cell culturing which has often lead to a low percentage
of successful karyotypes for CLL cells, given their inherent inability to survive and proliferate *ex-vivo*. However, today the use of growth stimulants such as CpG oligonucleotides, IL-2 and CD40L, have allowed better detection rates in CLL.

Today, in clinical practice, interphase FISH is instead the gold standard method of choice for CLL, since it has the ability to detect smaller DNA target sequences (probe size starting at 100-1000bp) without cumbersome culturing methods. FISH involves the hybridization of a panel of fluorescently labelled region specific probes to complementary DNA sequences at specific chromosomal regions. In CLL, interphase FISH can detect clinically relevant and recurrent lesions in up to 80% of CLL patients. The advantage of FISH relies on its ability to detect alterations even in samples with a low percentage of tumour cells. Other advantages include its capacity to detect homo- and heterozygous deletions and gains, as well as complex chromosomal translocations and rearrangements. That said, FISH is limited in its ability to detect aberrations that lie outside of the probe specific regions, making global screening difficult. Techniques such as spectral karyotyping and multiplex FISH simultaneously employ multiple probes sets in order to assess all chromosomes. However, these technologies are costly and labour intensive and in some instances limited by difficulties in probe design to certain chromosomal regions.

Unlike FISH, comparative genomic hybridisation (CGH) has the ability to readily examine the entire genome for alterations (approximately 5-10 Mbp in size). CGH involves the quantitative hybridization of differentially labeled normal and tumour DNA to normal metaphase chromosomes. On the other hand, CGH has a low sensitivity, requiring a high percentage of cells to have a particular alteration in order to be detectable. Furthermore, such alterations as inversions and translocations can not be identified.

Today, the need for genome wide, high resolution scanning has culminated in the development of array technologies and sequencing platforms allowing broader characterization of known alterations and novel smaller genetic lesions contributing to complexity. Array-comparative genomic hybridization (aCGH) and single nucleotide polymorphism (SNP) array technologies investigate global genomic events with the exception of repetitive regions, without the requirement of cumbersome metaphase chromosome preparations. Using aCGH, copy number alterations (CNAs) at a level of <100 kb can be detected.

In general, aCGH utilises the same principals as those of conventional CGH except that instead of employing metaphase chromosomes as targets, thousands of short DNA probes representing the genome are arrayed onto a
aCGH designed using bacterial artificial chromosomes (BAC) tend to hold a lower power of resolution compared to aCGH oligonucleotide arrays. Oligo aCGH permit a higher resolution to BAC arrays since a higher number of probes made up of 60-mer oligonucleotides are hybridised to the chip. Although a reliable method, aCGH is limited by its inability to detect CNN-LOH events. In contrast, SNP array platforms allow for the simultaneous detection of CNN-LOH events and CNAs at high resolution down to ~10-25kb in size depending on the array.

SNP arrays are constructed using 25-30-mer oligonucleotide probes containing specific SNPs. Each SNP position across the genome possesses two allelic variants on the array, SNP A and SNP B. Today up to 10⁶ probes can be spotted on the array. After hybridising the sample of interest to the array, SNP copy number is determined from the fluorescent signals arising from SNP A and B alleles. Normalising the SNP data to a previously analysed healthy individual reference set determines sample/reference ratios indicating the presence of chromosome alterations. One crux of such SNP and array-CGH technologies is that they do not detect alterations that occur in a low percentage of tumour cells. Failure to detect alterations also depends on the platform and the algorithm used. For instance, some algorithms do not tolerate normal cell contamination >20%, whereas some algorithms are adjusted to account for normal cell interference.

In the case of CLL, alternative optimization measures such as B cell enrichment techniques help resolve the issue of normal cell contamination and low tumour percentage. Recently, the Affymetrix 250K SNP array technology used in this thesis, have been correlated to FISH analysis in CLL patients. Remarkably, this array showed a 98.5% correlation to gold standard FISH analysis. On this basis, this array was suggested for use in clinical routine. Furthermore, this array enabled detection of the same prognostically relevant lesions as higher resolution arrays such as the Affymetrix 6.0 at lower cost.

DNA methylation

Establishment and maintenance of DNA methylation

DNA methylation is a heritable epigenetic mark that involves the non-random addition of a methyl group to the fifth position of a cytosine ring within CpG dinucleotides. DNA methylation is accompanied by concomitant histone modifications that play key roles in the control of gene activity and nuclear structure. For instance, DNA methylation and repressive histone marks create a condensed genomic structure that impedes the access or in-
hibits the binding of the transcription machinery resulting in gene inactivation. Conversely, the presence of activating histone marks and the lack of DNA methylation promotes an open chromatin structure supporting transcription and gene activation. Together, these actions function to regulate normal cell differentiation, germline/tissue specific gene expression, genomic imprinting, embryogenesis and transposon gene silencing.248-252

During early mammalian development, a global de-methylation event occurs during pre-implantation. Once implanted, the embryo undergoes a wave of de novo methylation.253-255 De novo methylation patterns are established through the actions of DNA methyl transferases DNMT3A and B.256 The above actions lead to wide spread global DNA methylation with the exception of CpG islands. Once established, this global bimodal methylation pattern is faithfully maintained through the actions of the ‘maintenance’ DNA methyltransferase DNMT1 enzyme. This enzyme is recruited by the replication machinery to hemi-methylated DNA where it catalyses the transfer of methyl groups to newly synthesized DNA strands after DNA replication.248,257

During mammalian development, the bimodal DNA methylation pattern encounters further de novo DNA methylating and de-methylating events in order to bring about cell differentiation patterns. Unlike DNA methylation, mechanisms of DNA de-methylation in mammals remain largely unexplored. That said, observations of global de-methylation during embryogenesis have pointed to both passive and active modes of action. Passive de-methylation is said to occur when DNA methylation pathways are inactivated and DNA methylation is diluted after subsequent replication cycles.253 Active DNA demethylation in mammals is thought to involve AID induced deamination of methyl cytosine to thymine.258 This is followed by a base excision repair by DNA glycosylase enzymes of the T/G mismatch.259,260 Alternatively, in vitro studies have shown that DNMTs 3a and 3B can also deaminate 5-methyl cytosine.261 Imprinted regions are sheltered from active global demethylation as DNA methylation maintenance proteins such as PGC7 and DNMT1 protect them.250,262,263 Moreover, it is also hypothesized that de-methylation may occur in somatic cells at sequence specific sites.253,261

DNA methylation regulation

The success of epigenetic marks governing gene activity and nuclear structure depends on the strict regulation of the epigenetic machinery. DNMT enzymes, histone modification enzymes, DNA-methyl-binding proteins, non-coding RNAs, polycomb complexes amongst others work together to target specific genes for activation or repression. To date, the order and se-
sequence in which these regulators crosstalk to co-ordinate epigenetic marks at specific targets is not fully elucidated. Nevertheless, clues from normal cell development and differentiation provide some indication as to how DNA methylation may be regulated. In development, histone marks are thought to play a role in establishing DNA methylation. Here, they act as platforms permitting the recruitment, or inhibition of, the DNA methylation machinery. In contrast, DNA methylation is considered to play a role in maintaining histone modifications. Methylated DNA bound by methyl binding proteins (MECP2 and MBD2) is said to act as a platform guiding the reassembly of disrupted histone modifications post cell replication.

**Protection of CpG islands during development**
Bimodal DNA methylation patterns set up during development are thought to be directed through specific histone modifications. For instance, unmethylated H3K4 histones are associated with DNA methylation in many cell types. At unmethylated H3K4 histones, it is proposed that DNA methylation is permitted through DNMT3L which acts as a bridge to summon active DNMT3A and 3B enzymes to H3. It is thought that repressive methylation marks at H3K4 inhibit the recruitment of de novo DNA methylating enzymes, keeping CpG islands free of DNA methylation repression. Alternatively, mouse studies indicate that CpG islands may escape de novo methylation through the action of cis acting sequences such as Sp1 sites that bind the Sp1 transcription factors and through unknown mechanisms hinder DNA methylation at these sites.

**Epigenetic regulation of cell differentiation.**
In embryonic stem cells, differentiation requires pluripotency genes to be silenced. In their active pluripotent state, these genes are found within an open chromatin structure bound by acetylated H3 and H4, methylated H3K4 and unmethylated DNA marks. Upon cell differentiation, gene silencing is thought to involve prompt repression by protein repressor molecules. The actions of G9 results in the methylation of H3K9, however at the same time by unknown means, H3K4 is demethylated. H3K9 methylation acts as a platform allowing the binding of heterochromatin proteins such as HP1. Additionally, G9 summons de novo DNA methyltransferases DNMT3A and 3B to promote DNA methylation resulting in the subsequent formation of condensed chromatin at these pluripotency genes.
Polycomb target genes in cell type specific DNA methylation

Polycomb-multi-protein repressive complexes silence polycomb target genes vital for mammalian development. It is thought that polycomb silencing of these developmental genes may be overcome during germline differentiation. In embryonic stem cells (ESCs), it is proposed that histones carrying easily reversible, bivalent repressive (H3K27me3) and activating (H3K4me3) marks sit at these polycomb target genes, allowing them to actively switch between active and repressive states depending on the demands of the differentiating cell.

Interestingly, at the DNA level, the majority of these target genes stay unmethylated during development. This is thought to be partially attributed to the fact that these genes are mainly affiliated to CpG island promoters that are largely safeguarded from DNA methylation during development. In certain instances, such as the differentiation of ESCs to neural precursors, DNA methylation of polycomb target genes has been described. DNA methylation of these genes is thought to be mediated by polycomb complexes. Polycomb response elements (PREs), regulatory sequence specific elements, are proposed to enable the binding of polycomb (PRCs) subunits, marking the gene for repression. Binding of EZH2, the histone methyltransferase subunit of the PRC2 complex, is known to catalyse the methylation of H3K27.

In vitro studies indicate EZH2 to interact with the DNMTs to mediate DNA methylation of polycomb target genes.

DNA methylation in tumorigenesis

Today, it is widely accepted that alterations to the DNA methylome and changes to histone modification patterns contribute to cancer pathogenesis. In particular, DNA methylation of vital genes involved in key cellular pathways is now recognized as one of the hallmark mechanisms of aberrant gene silencing (Figure 3). Moreover, hypomethylation of instable repetitive genomic regions is further thought to contribute to cancer (Figure 3). Unlike genetic lesions, such as mutations and deletions, DNA methylation provokes gene silencing without changes to the DNA sequence. In some instances, aberrant DNA methylation is suggested as an early event in cancer occurring at the stem cell stage. However, other arguments support aberrant DNA methylation as a late event, in tumorigenesis. Many of the DNA methylation changes found in cancer are said to merely reflect the patho-biology of long lived progressive tumour cells and that these changes are mostly indicative of passenger rather than driver epigenetic events. Overall, altered DNA methylation is one component in a very complex epigenetic network that co-
operates with intrinsic cellular defects and tumour promoting extrinsic factors to aid carcinogenesis.

Figure 3. Illustration of DNA methylation in normal and cancer cells. (Adapted from Baylin 2005)²⁹¹

Hypermethylation

As previously described, globally, DNA is largely methylated with targeted regional hypomethylation of most CpG islands in normal cells (Figure 3). Vast DNA methylation covering repetitive genomic regions and transposable elements ensures the maintenance of genomic stability.²⁵² On the other hand, hypomethylation of CpG islands allows the transcription of vital genes governing normal cell growth, proliferation, differentiation and tissue specific gene expression.²⁴⁹,²⁸¹ In cancer, the opposite DNA methylation pattern ensues where targeted promoter hypermethylation and global hypomethylation work in synergy with anomalous histone modifications and interaction proteins to disrupt gene regulation.²⁴⁸,²⁷⁹,²⁸²-²⁸⁴ Pivotal to carcinogenesis is the role of promoter DNA methylation and transcriptional silencing of vital tumour suppressor genes.⁹⁰,²⁷⁹,²⁸³ This largely involves methylation of cytosines within densely populated CpG regions known as CpG islands spanning the 5’ end of many genes within their promoters. Nevertheless, new evidence suggests that methylation of CpG island shores and shelf regions of low CpG density located outside the vicinity of traditional CpG islands are fundamen-
tal to gene silencing. For instance, Doi et al have demonstrated a number of overlapping differentially methylated loci involved in normal tissue programming, epigenetic reprogramming to pluripotency and aberrant programming of cancer cells. Interestingly, these loci were shown to be predominantly contained within CpG island shores. These findings suggest that epigenetic changes encountered in cancer are likely due to the reprogramming of the normal pattern of tissue specific differentiation and highlight the importance of CpG island shores in regulating transcription of specific loci.

Hypomethylation

In cancer, aberrant global hypomethylation has been proposed to inappropriately active possible oncogenic genes, little evidence of this has been documented. More importantly, hypomethylation of gene poor repetitive DNA sequences that account for up to 30% of the genome are suspected to contribute to genome instability. Furthermore, it has recently been described in mature B cell lymphomas that unlike global hypo-methylation previously implicated in cancer, gene specific hypo-methylation can fraternize with local DNA hypermethylation during tumorigensis.

DNA methylation and mutagenesis

DNA methylation can further influence genome stability through the introduction of mutations at CpG sites. Methylated cytosines are subjected to a higher rate of mutagenesis relative to unmethylated cytosines. This is thought to occur due to an increased propensity of spontaneous deamination at methylated cytosines. These mutations account for a large number of known point mutations contributing to genetic disease and are said to account for up to 30% of all germline point mutations. Moreover CpG sites in the coding regions of TP53 are methylated in all human tissues. Approximately 25% of all inactivating mutations in cancer in general can be attributed to transitions at CpG dinucleotides.

Role of DNA methylation in CLL

Like other cancers, the genome of CLL patients is hypomethylated compared to PB mononuclear cells from healthy normal controls. Until recent times, epigenetics was explored solely on a gene to gene basis. Nevertheless, this candidate gene approach identified a number of DNA methylated promoters contributing to CLL pathogenesis. For instance, TWIST-2, a transcription factor and well known silencer of p53 function in other malignancies, was shown to be methylated predominantly in IGHV mutated patients. Similarly, silencing of ZAP70 through methylation has been
strongly correlated to IGHV mutational status. Additionally, loss or reduced expression of DAPK1 through promoter methylation has been shown in a large proportion of CLL. In mutated CLL, DNA methylation of CD38 and BTG4 has been correlated to a favourable outcome, whereas DNA methylation of HOXA4 has been affiliated to poor outcome.

More recently, a study by Rush et al has demonstrated the genome-wide methylation profiles of ten CLL patients. Using restriction landmark genome scanning (RLGS), a restriction enzyme, 2D gel electrophoresis method, only a limited number of CpG sites (~3,000) were analysed across the whole genome. Interestingly, between 2.5–8.1% of the CpG islands were shown to be aberrantly methylated in CLL samples compared with normal controls. Moreover, 193 aberrantly methylated sequences were identified, of which 93% had CpG island characteristics and 90% showed homology to genes or expressed sequences. Many of the genes identified were transcription factors such as DERMO1, FOXE1, TBX3, and IPF1. Other loci known to play roles in the nervous system such as TBR1, GLRB, and PAK5 were also identified, however their role in CLL if any is still unknown. Finally, the GRM7 gene, a gene functioning in the possible inhibition of cyclic AMP signaling in the induction of apoptosis was identified as methylated in CLL patients.

With the advent of higher resolution microarray and sequencing technologies, overall DNA methylation scores have been indicated as strong independent predictors in CLL progression. Patients with a higher genome-wide DNA methylation index relative to their age expected index have been associated with an earlier indication for therapy and Rai stage IV disease. On the other hand, patients displaying a lower index with respect to age have clinically stable, Rai stage 0-1 disease. Additionally these new technologies have allowed the interrogation of repetitive sequences in CLL patients. Here, aberrant hypomethylation at repeat sequences including ALU, LINES and SATα, is shown to be more marked in 17p deleted cases. In addition, a lower SATα methylation level has been shown to act as an independent marker predicting a shorter time to first treatment in CLL. Furthermore, the methylation profiles of CLL patients are shown to segregate according to CD38 expression.

Disruption of DNA methylation regulatory mechanisms in tumourigenesis

In cancer, the epigenetic machinery fall victim to destruction and disorganisation. The exact mechanisms involved in epigenetic deregulation are largely unknown. Genomic loss or mutation to the DNA methyltransferases and
other regulatory molecules are suspected to contribute to epigenomic disarray. For instance, somatic mutations decreasing de novo DNMT3A activity has been found in up to 20% of acute myeloid leukaemia (AML) patients.\textsuperscript{306,307} Moreover, mutations in polycomb repressive complex components have also been described. For example, mutations at tyrosine 641 in the SET methyltransferase domain of EZH2, are shown to change the substrate specificity of EZH2 for various methylation states at H3K27 in non-Hodgkin lymphoma (NHL).\textsuperscript{308} For instance, the Y641C mutation abrogates the enzymatic activity of EZH2 against unmethylated and mono methylated H3K27. Remarkably, this mutation allows this enzyme to carry out better catalysis and formation of trimethylated H3K27 compared to the wild type enzyme.\textsuperscript{308} Additionally, in myeloid disorders, inactivating mutations of EZH2 have also been reported.\textsuperscript{309}

Altered expression levels and improper activation of the epigenetic machinery are also considered to be attributed to epigenetic disorder. For instance high expression of DNA methyltransferases are thought to be involved in the pathogenesis of some cancer cell types. Other studies have noted the over-expression of EZH2 to be associated with cancer progression. One study by Chen et al suggests the over expression of cyclin dependent kinases (CDKs) in cancer to aberrantly phosphorylate and activate EZH2.\textsuperscript{310} Other in vitro prostate cell line studies suggest that increases in EZH2 expression are due to genomic loss of miRNA 101.\textsuperscript{311} Interestingly, in MYC induced lymphoma, MYC has been implicated to stimulate EZH2 expression by repression of its negative regulator miR-26a, thus contributing to lymphomagenesis, therefore a regulatory role for miRNAs in the epigenetic control of gene regulation has been proposed.\textsuperscript{312}

Use of DNA methylation markers in clinical routine

In recent years, the plethora of genome wide DNA methylation studies have noted differentially methylated gene patterns in different cancers.\textsuperscript{290,313-315} In some instances, these methylation changes constitute tumour and tissue-specific changes that contribute to tumorigenesis,\textsuperscript{285,286} thus marking these genes as possible candidate prognostic markers and therapeutic targets in cancer. Despite the volume of these differentially methylated candidates, very few have graduated to be true bona fide prognosticators or therapeutic targets. One reason for this is the inability to decipher true tumour specific differentially methylated genes from normal tissue and patient specific changes.\textsuperscript{316} For instance, changes attributed to such factors as age, sex,\textsuperscript{317} hormones (androgens)\textsuperscript{318} and diet complicate matters since increasing age and poor folate diets leading to insufficient in vivo methyl stores can aid hypomethylation events.\textsuperscript{319} Furthermore, changes attributed to patient treat-
ment regimes and disease progression may mask true tumour specific changes.

Another barrier preventing the possible use of such candidate markers in clinical routine is the question of methylation stability over time. To date, little evidence exists on this issue, since many studies are based on snap shot analysis of different individuals and diseases. One study investigating bladder cancer metachronous tumours from two individuals, has shown DNA methylation to be relatively stable over time both within the individual and between patients. In contrast, a study by Bjornsson et al focusing on absolute inter- and intra-individual DNA methylation changes have identified substantial inter-individual age variation and considerable differences occurring over time within individuals that would otherwise be missed by group averaging. Interestingly, these changes were shown to be clustered within family members. Hence, these findings support the age related loss of normal methylation patterns as a mechanism for the late onset of disease and further highlight the possible influence of genetic background on DNA methylation changes. That said, a study of DNA methylation in 40 monozygotic twins found 14 to have a discordant degree of total DNA methylation, particularly those over 28 years of age. Thus, the latter finding is indicative of environmental factors that are already known to alter DNA methylation. While issues of stability and specificity prevail and since the clinically relevant level of DNA methylation conducive to target gene inactivation is unknown, use of differential DNA methylation changes as candidate prognostic markers or treatment targets remain to be seen.

DNA methylation technologies

DNA methylation quantification methods have evolved rapidly over the past two decades. Initially, global DNA methylation levels were estimated using high performance liquid chromatography (HPLC). This method, although highly quantitative, failed to give information on the exact sequences afflicted with aberrant DNA methylation. The need for region specific analysis, provoked the development of cytosine methylation enrichment techniques employing methylation sensitive restriction enzyme (MRE) digestion, McrBC methylation dependent enzyme fractionation, 5′-methylcytosine antibody immuno-precipitation, methyl-binding protein capture methods and chemical bisulfite treatment approaches. Applying enriched methylated DNA to various microarray and sequencing technologies has allowed the interrogation of DNA methylation at individual regions, particularly within CpG islands and gene promoters. More recently, advancing technologies have enabled a more global view, uncovering the DNA methylation status of non CpG island sites and repetitive sequences throughout the genome.
Early sequence specific genome wide DNA methylation investigations utilised the (RLGS) platform, an MRE based 2D gel method.\textsuperscript{330,331} This technology explored the whole genome on a limited scale, deciphering methylation events contained only within MRE recognition sites within a limited number of genes as reviewed in Fouse et al.\textsuperscript{329} Plagued by false positive and negative results originating from incomplete digestion the true reflection of DNA methylation at these sites is often hampered, as reviewed in Esteller et al.\textsuperscript{288} Considering the disadvantages of RLGS and the fact that it only assess ~3,000 CpG sites,\textsuperscript{303,330} the need for more reliable, higher resolution analysis paved the way for the use of high fidelity bisulfite treatment in combination with microarray technology.

The onset of bisulfite conversion revolutionised DNA methylation analysis. This allowed the reproducible change of unmethylated cytosines to be deaminated to uracil, leaving methylated cytosines unchanged as they are refractory to treatment.\textsuperscript{328} However, bisulfite treatment reduces the sequence complexity of the genome complicating its application to oligonucleotide arrays. Nevertheless, today bisulfite treatment is considered the gold standard method of cytosine DNA methylation enrichment\textsuperscript{328} for many high throughput DNA methylation sequencing platforms such as shotgun bisulfite sequencing\textsuperscript{332} and DNA methylation microarray Illumina human27 and human450 technologies. The 27K array allows the investigation of ~28,000 CpG sites covering ~14,000 genes, whereas the 450K array covers ~485,000 sites across the whole genome. These arrays, particularly the latter permit, the interrogation of CpG sites both within gene promoters and outside of their coding regions.\textsuperscript{333}

In brief, microarray platforms interrogating DNA methylation at predetermined CpG sites\textsuperscript{334-336} generally offer less coverage (10\textsuperscript{3}-10\textsuperscript{6}) of the DNA methylome relative to DNA methylation sequencing methods and next generation technologies (10\textsuperscript{6}-10\textsuperscript{7}).\textsuperscript{329,337,338} Nevertheless, the microarray comprehensive high throughput arrays for relative methylation (CHARM) tiling array provides a high CpG coverage similar to that of some sequencing technologies ~10\textsuperscript{6}.\textsuperscript{326} In general, microarray platforms based on the hybridization of bisulfite treated and MRE prepared DNA can provide single CpG site resolution. In contrast, microarray platforms utilising immunoprecipitated or methyl binding protein enrichment, quantify DNA methylation across a region of ~1000 base pairs. Hence, the latter provide higher coverage of the genome yet at a poorer resolution (as reviewed in Fouse et al).\textsuperscript{329} Such techniques as MeDIP and MIRA-Chip rely on the faithful precipitation of DNA methylated sites using specific antibodies against 5’methylcytosine or methyl CpG binding proteins which bind methylated cytosines. Enriched DNA is then hybridized to a genomic microarray.\textsuperscript{289,327} A summary of the
advantages, disadvantages and characteristics of some of the available DNA methylation platforms is described below in Table 2a and 2b.

Table 2a: Summary of DNA methylation microarray platforms.

<table>
<thead>
<tr>
<th>Microarray Technology</th>
<th>Pre-Processing</th>
<th>CpG site Coverage/Resolution</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illumina Human27</td>
<td>Bisulfite conversion</td>
<td>~27,000/~1 bp</td>
<td>Cheap, Reliable, Single bp resolution</td>
<td>Low coverage, Promoter biased, No assessment of repeat regions</td>
</tr>
<tr>
<td>Illumina Human450</td>
<td>Bisulfite conversion</td>
<td>~450,000/~1 bp</td>
<td>Reliable, Single bp resolution</td>
<td>Low coverage, Inability to assess repeat regions</td>
</tr>
<tr>
<td>HELP-Chip Promoter array</td>
<td>MRE digestion</td>
<td>10^7/~1000 bp</td>
<td>High genome coverage</td>
<td>Low resolution, Limited to MRE sites</td>
</tr>
<tr>
<td>MIRA-Chip (CGI array) (Global tiling array)</td>
<td>Sonication or MRE/MB column</td>
<td>10^7/~1000 bp, 10^6/~1000 bp</td>
<td>High genome coverage, No need for ssDNA</td>
<td>Low resolution, Limited to MRE sites</td>
</tr>
<tr>
<td>MeDIP-Chip Promoter array Global tiling array</td>
<td>Sonication/IP</td>
<td>10^6/~1000 bp, 10^5/~1000 bp</td>
<td>High coverage, Not limited by restriction sites</td>
<td>Low resolution, Enrichment bias towards high density CpG sites, Needs ssDNA</td>
</tr>
<tr>
<td>CHARM tiling array</td>
<td>McrBC digestion</td>
<td>10^9/~1000 bp</td>
<td>High coverage, Less biased towards high density CpG sites compared to MeDIP</td>
<td>No single site resolution, Limited to McrBC sites, Can not interrogate low density CpG sites</td>
</tr>
</tbody>
</table>

IP: immunoprecipitation, MRE: methylation sensitive restriction enzyme, CGI: CpG island, MB column: methyl binding protein column, CHARM comprehensive high throughput arrays for relative methylation. HELP- HpaII tiny fragment enrichment by ligation mediated PCR. MeDIP- methylated DNA immunoprecipitation, MIRA: methylated CpG island recovery assay, ssDNA: single stranded DNA. Adapted from Fouse *et al* 2010.329
Table 2b: Summary of DNA methylation sequencing platforms.

<table>
<thead>
<tr>
<th>Sequencing Platforms</th>
<th>Pre-processing</th>
<th>CpG site Coverage / Resolution</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methyl-seq /Help-seq</td>
<td>MRE digest</td>
<td>$10^5$-$10^6$/~1 bp</td>
<td>High coverage Interrogates repeat sequences Single bp resolution</td>
<td>Limited to MRE sites</td>
</tr>
<tr>
<td>Shotgun Bisulfite sequencing</td>
<td>Bisulfite /sonication</td>
<td>$10^7$/~1 bp</td>
<td>High coverage Interrogates repeat sequences Single bp resolution</td>
<td>Difficulties in assembling bisulfite converted sequences</td>
</tr>
<tr>
<td>MeDIP-seq</td>
<td>Sonication/IP</td>
<td>200-300 bp</td>
<td>High coverage Interrogates repeat sequences</td>
<td>No single site resolution Complicated to analyse</td>
</tr>
</tbody>
</table>

IP: immunoprecipitation, MRE: methylation sensitive restriction enzyme, HELP- *Hpa*II tiny fragment enrichment by ligation mediated PCR. MeDIP- methylated DNA immunoprecipitation, MIRA Adapted from Fouse *et al* 2010.329

Although all the aforementioned technologies offer useful information, interpretation of observations from these methods often need to be independently reconfirmed with alternative methods. Semi-quantitative methods such as methylation specific PCR (MSP)339 and bisulfite sequencing are often employed.340 MSP, although easy to implement, is restricted in that it solely interrogates the CpG sites covered by the primer. Unlike MSP, bisulfite sequencing is a labour intensive technique. That said, bisulfite sequencing is much more advantageous than MSP since it readily interrogates CpG sites contained within the amplified region determined by the primers. Fully quantitative techniques such as pyrosequencing, can overcome some of the disadvantages of the former methods. Specifically, pyrosequencing easily allows the quantification of DNA methylation across a chosen site, often containing a number of CpG sites with minimal DNA input.341,342

Epigenetic based treatment

Epigenetic defects are heritable changes that do not affect DNA sequences; hence their effects are potentially reversible through the use of methyl inhibitors and histone deacetylation drugs. Today, the use of hypo-methylating drugs 5-azacytidine (5’aza) and its more potent counterpart 5-aza-2’-deoxycytidine (DAC) has been approved for the treatment of myelodysplastic syndrome patients. The use of histone deacetylase inhibitors also represents another group of potential drugs that can be used in epigenetic cancer therapy. Recently the first one of its type, suberoylanilide hydroxamic acid (SAHA), has been approved for the treatment of cutaneous T cell lymphoma.343
Thesis Aims

The overall aim of this thesis was to screen for genetic and DNA-methylation events in divergent CLL prognostic subgroups and investigate DNA methylation over time and within different CLL compartments. The specific aims were as follows:

- To determine the frequency of IGHV3-21 gene usage among a Swedish population-based cohort using Sanger sequencing and assess the impact of IGV3-21 stereotypy status on clinical outcome. (Paper I)

- To investigate genomic alterations in CLL patients expressing stereotyped and non-stereotyped IGHV3-21 and IGHV4-34 BCRs using 250K SNP arrays. (Paper II)

- To study DNA methylation profiling in prognostically diverse CLL subsets using 27K DNA methylation arrays. (Paper III)

- To evaluate the contribution of time and microenvironment on the global DNA methylation profiles of CLL patients, by assaying patient-matched diagnostic and follow-up samples and patient-matched LN and PB samples on 450K DNA methylation arrays. (Paper IV)
Materials & Methods

Patient samples and patient characteristics

In Paper I, peripheral blood (PB) was obtained from 337 patients at diagnosis including, 325 CLL and 12 CLL/small lymphocytic leukaemia (SLL) cases from the Swedish cohort of a Scandinavian population-based case-control study called SCALE (Scandinavian Lymphoma Etiology). In Paper II, tumour material from 101 CLL patients was obtained from four cooperating institutes in France (n=15), Greece (n=25), Denmark (n=4) and Sweden (n=57). Samples mainly originated from peripheral blood (n=88), although samples from spleen (n=5) and bone marrow (n=8) were also included. In total, 42 IGHV3-21 samples, encompassing 29 subset #2 and 13 non-subset #2 cases, and 59 IGHV4-34 samples, comprising of 17 subset #4, 8 subset #16 and 34 non-subset #4/16 cases were analysed. Subsets were defined according to Stamatopoulos et al and Murray et al. A CLL sample set (n=203) from a recent array study was used for comparison purposes.

In Paper III, 23 CLL samples (6 IGHV mutated, 7 IGHV unmutated and 10 IGHV3-21) were obtained from the Biobank at the Department of Pathology, Uppsala University Hospital, Sweden. All samples were derived from peripheral blood. For validation experiments, an independent set of 10 and 50 CLL samples were analysed by MSP-PCR and RQ-PCR respectively. For control purposes, one peripheral blood mononuclear cell (PBMC) sample, one CD19 sorted B-cell sample and one negative whole-genome amplified DNA control were additionally analysed. For comparison purposes, 2 CLL EBV transformed cell lines (I83 derived from a mutated CLL patient and HG3 derived from an unmutated CLL patient) were included for analysis.

In Paper IV, a total of 36 PB CLL samples obtained from the Swedish cohort of the SCALE study were chosen for analysis. To study the DNA methylation profiles of IGHV mutated and unmutated subgroups, sorted PB samples derived form 9 IGHV mutated and 9 unmutated patients at diagnosis were analysed. Sorted PB samples derived from these same patients were taken at follow-up to investigate DNA methylation over time. (Patient-paired samples: IGHV mutated cases n=9 cases at diagnosis and n=9 cases at follow-up and IGHV unmutated cases n=9 cases at diagnosis and n=9 cases at follow-up.
In addition, a separate set of 10 patient-matched PB and LN samples collected from biobanks at Uppsala University and Linköping University hospitals were analysed. PB from three age-matched normal individuals and one whole-genome amplified DNA derived from PB were included as controls.

All samples were diagnosed according to the WHO classification and the International Workshop on Chronic Lymphocytic Leukaemia criteria (IWCLL). In Paper I, samples with highly variable tumour load were included for analysis. In Papers II and III, only samples with ≥70% tumour cell content were included for analyses. In Paper IV, all samples were CD5+/CD19+ sorted. Clinical and molecular data characterising Binet stage, IGHV mutational status, treatment and survival status, and chromosomal alterations are detailed in Papers I-IV. Informed consent was obtained according to the Declaration of Helsinki.

IG gene sequencing

In all Papers, CLL patients were characterized according to IG gene usage, mutational status and the presence of stereotyped and non-stereotyped receptor sequences. These characteristics were detected through PCR amplification of gDNA or cDNA using IGHV gene family specific framework 1 or leader primers. Consensus primers for the kappa/lambda FR1 and IGKJ/IGLJ genes were used to amplify light chain IG rearrangements. Following amplification, most clonal products were sequenced using the Big Dye terminator cycle sequencing reaction kit (Applied Biosystems, Foster City, CA). Where direct sequencing was unsuccessful, gel purification and/or cloning prior to sequencing was performed. Sequences were then aligned in the IMGT database using the IMGT/V-QUEST online tool. IGHV genes were denoted unmutated where a ≥98% identity to the corresponding germline gene was determined. Sequences with <98% identity were considered mutated. In Paper I, IGHV unmutated cases were further classified into truly unmutated (100% identity to germline), minimally (99-99.9% identity) and borderline mutated (98-98.9% identity). Stereotypy status was determined based on established molecular IG sequence criteria.

250K Affymetrix SNP array

In Paper II, 250K NspI SNP-arrays were applied to study CNAs and CNN-LOH events in IGHV3-21 and IGHV4-34 stereotyped and non-stereotyped samples. Array experiments were performed according to the standard proto-
cols for Affymetrix GeneChip® Mapping NspI-250K arrays (Figure 4) (Gene Chip Mapping 500K Assay Manual (P/N 701930 Rev2), Affymetrix Inc, Santa Clara, CA, USA). Subsequent to array scanning using the GeneChip® Scanner 3000 7G, probe level normalisation and SNP calling were accomplished using Affymetrix GeneChip® Genotyping Analysis Software (GTYPE) 4.1.

**Figure 4.** The 250K Affymetrix SNP array contains 262,000 SNP probes covering the whole genome. In short, the tumor DNA is enzymatically fragmented using the NspI restriction enzyme and adapters are added allowing PCR amplification of the fragments. After amplification the fragments are further fragmented and end labeled. Following hybridization of these labeled fragments to the array, copy number is estimated based on the ratio of fluorescence signals detecting gains and losses. (Adapted from affymetrix.com)

Single sample QC was performed using the Dynamic model algorithm and genotype calls were made using the BRLMM algorithm. Quality control measures were given as a neighbor score. This score was calculated by averaging the euclidean distances between the log₂ ratios of 5 consecutive SNPs. A neighbor score of ≤0.4 was applied as cut-off for sample inclusion. Log transformation of the data producing log₂-ratios allowed for a more symmetrical data distribution. Subsequently, copy-number normalization was performed using the Copy Number Analysis Tool (CNAT) 4.0.1 and normal samples (n=82) previously run at the Uppsala Array Platform were used as a reference set for comparison.
Copy-number analysis using the BioDiscovery Nexus Copy Number 3.0 software (BioDiscovery, El Segundo CA, USA) operating the Rank Segmentation algorithm was employed. CNAs were defined on significance thresholds (p-value) of $1 \times 10^{-6}$ and a log$_2$ ratio cut-off at ±0.2 for regions sized 200-500 kbp and ±0.15 for regions >500 kbp, respectively. These settings were defined through previously described real-time quantitative PCR (RQ-PCR) validation experiments.$^{180}$ To exclude polymorphic regions, CNAs that overlapped with >50% of a known CNV region were not considered as true CNAs and excluded from analysis. CNN-LOH analysis was performed using SNP array-data, taking into account the fraction of normal cells indicated from the flow-cytometry data.$^{349}$ CNN-LOH regions larger than 3 Mbp with less than 50% overlap to CNVs were acknowledged as legitimate CNN-LOH events.

**Illumina Infinium HumanMethylation27 and HumanMethylation450 BeadChip array**

In Paper III, the HumanMethylation27 array (Illumina, San Diego, USA) was applied to study the differences in methylation profiles between mutated, unmutated and IGHV3-21 CLL patients. In Paper IV, the HumanMethylation450 array was employed to study genome wide methylation differences between 1) IGHV mutated and unmutated subgroups, 2) paired diagnostic and follow-up CLL samples and 3) paired matched PB and LN CLL samples. All samples underwent bi-sulfite conversion of genomic DNA according to the standard EZ DNA Methylation Kit protocol (Zymo Research, Orange, CA, USA).

Following conversion, the DNA was whole-genome amplified, enzymatically fragmented, precipitated, resuspended and hybridized over night to locus specific oligonucleotide primers. The 27K array based on the infinium I assay design employs locus specific oligonucleotide primers that are linked to individual bead types (Figure 5). Two bead types correspond to each CpG locus – one to the methylated cytosine (C) and the other to the unmethylated thymine (T) state. Following hybridization, single base primer extension allows the detection of C or T nucleotides through a fluorescent signal that is measured using the Illumina BeadStation GX scanner. The intensities of the unmethylated and methylated bead types are measured and a DNA methylation value, described as the Beta values or methylation index (MI), is recorded for each locus via BeadStudio software. Unlike the 27K array, the 450K array employs both infinium I and infinium II technologies (Figure 5). The infinium II design utilizes a single bead type. In this case, the methylation status is calculated after hybridisation at the single base extension phase.
The intensity of the fluorescent signal emitted from this extension is scanned and measured. Since the 450K array elicits the use of both infinium I and II assay designs, the array can increase genome wide coverage across CpG sites compared to the limited coverage of the 27K array. In both array types, the methylation status for each detected CpG can range between 0 (completely unmethylated) to 1 (completely methylated). MI cut-offs of ≥0.7 and ≤0.4 (average of all MI values in each group) were considered methylated and unmethylated, respectively.

**Figure 5.** Diagram representing the principals of the HumanMethylation 27 and 450 BeadChip arrays. The 27K array is based solely on the infinium I design employing two bead types per locus, one for the methylated (M) and one for the unmethylated (UM) state. The 450K array is based on both Infinium I and II design. The Infinium II design uses only a single bead type (S). The methylated state is determined at base pair extension. (Adapted from Illumina.com)
Bioinformatic analysis of the Illumina human27 and 450 DNA methylation array data

To systematically identify differentially DNA methylated genes in paper III, bioinformatic analysis of the methylation data was carried out using the statistical computing language R (http://www.r-project.org) software. Similarly for Paper IV, bioinformatic analysis of the methylation data was conducted using the R based illumina methylation analyzer (IMA) package. In both studies, raw data was quantile normalized, arcsin transformed and an empirical Bayes moderated t-test was then applied using the ‘limma’ package to identify differential methylation patterns between the various sample comparisons. The p-values were adjusted using the method of Benjamini and Hochberg, and a level of p<0.05 was used as a cutoff. In Paper III, to ensure only CpG sites with a large methylation differences were obtained, an average geometric difference in MI of 0.45 between the IGHV mutated and IGHV unmutated subgroups, 0.35 between the IGHV3-21 and IGHV mutated subgroups and 0.35 between IGHV3-21 and IGHV unmutated subgroups was applied.

In Paper IV, an average difference in MI of ≥0.40 was applied to decipher recurrent differentially methylated sites between the IGHV mutated and IGHV unmutated patients at diagnosis. To identify recurrent methylation differences between paired diagnostic and follow-up samples, and patient matched samples from different anatomical compartments, a paired t-test was adopted. Again, P values were adjusted according to the Benjamini and Hochberg method and a P value of <0.05 was applied as a cut-off. To determine recurrent differentially methylated sites across the paired over time and compartment samples, an average difference in MI of >0.20 was first applied. On recognising the low number of recurrent sites passing the >0.20 threshold an average difference ≤0.20 was employed. In order to assess the likelihood of small methylation differences representing technical variation, R² values were calculated for diagnostic/follow-up pairs and patient matched compartment pairs and compared to the R² value reported by Illumina for technical replicates assayed on the 450k array (R²>98).

To assess non-recurrent intra-individual methylation differences between patient-paired samples and patient-matched compartment cases that would otherwise be masked by averaging of the data, intra-individual analysis based on non transformed data of the paired samples was conducted. Specifically, the number of sites differing by ≥0.4 in beta value (MI) between the individual diagnostic and follow-up pairs was employed. Sites differing by ≥0.4 in MI between the LN and PB patient-matched pairs was first applied to determine non recurrent differences across the compartments. Subsequently a filter of ≥0.3 in beta value was assessed.
Methylation-specific PCR analysis

In Paper III, methylation array results were verified using methylation specific (MSP)-PCR for 4 differentially methylated genes. Methylated specific primers were designed according to Li and Dahiya selection criteria using the Methprimer design software (Bisearch, San Francisco, USA). Standard PCR protocol was applied and bands were visualised on a 2% agarose gel.

Real-time quantitative PCR analysis

In Paper III, to investigate the influence of methylation status on gene expression of specific genes, the expression of 4 differentially methylated genes were tested using RQ-PCR on 60 independent samples. Total RNA was extracted using the RNA extraction Kit (Qiagen, Venlo, Netherlands) and reverse transcription was performed using the MMLV-RT kit (Invitrogen, Carlsbad, CA) and random hexamers (Fermentas, Burlington, Canada) according to the manufacturer’s protocol. RQ-PCR primers were designed using the Primer3 software (Broad institute, Boston, USA) (www.wi.mit.edu/). RQ-PCR analysis was performed using 2X SYBR green master mix (Fermentas, Burlington, Canada) and expression was analyzed using the Stratagene Mx 3005p (Stratagene, La Jolla, CA, USA) detection system and Max Pro QPCR software (Stratagene).

Bisulfite sequencing

To confirm the methylation status of individual genes in Paper III, two genes underwent bi-sulfite sequencing. Bi-sulfite sequencing primers (BSP) were designed according to primer selection criteria as outlined by Li and Dahiya. A touch down PCR was applied as described in Paper III. The PCR product was cloned using the Invitrogen TOPO TA cloning kit-2.1 TOPO vector (Invitrogen, Carlsbad, CA, USA), and 10-12 positive clones per sample were selected. The product was subsequently exonuclease/shrimp alkaline phosphatase (Exol/SAP) purified and sequenced using the BigDye Terminator Cycle Sequencing Reaction Kit (Perkin-Elmer, Foster City, CA). Sequences were aligned and analyzed using the bisulfite sequencing web-based tool BiQ Analyzer software (Max-Planck Institut für informatik, Saarbrucken, Germany).
5-aza-2’-deoxycytidine and Trichostatin A treatment

To test for the re-expression of methylated genes after methyl inhibitor and/or histone deacetylase inhibitor treatment, primary CLL cell samples and CLL cell lines were cultured in supplemented RPMI-1640 media as outlined in Paper III. Confluent cells were subdivided and cultured over 3 days in supplemented RPMI media treated with one of the following treatments: 1) medium containing 5-aza-2’-deoxycytidine (DAC) (5µM/L, Sigma Aldrich, St. Louis, MO) for 72 hours (medium was changed every 24 hours), 2) medium containing Trichostatin A (TSA) (500nM/lit, Sigma Aldrich, St. Louis, MO) for the last 24 hours, and 3) medium containing DAC for 48 hrs followed by addition of TSA for 24 hrs. Control cells were cultured in similar way with no drugs added.

Pyrosequencing

In Paper IV, pyrosequencing was chosen to validate the methylation status of selected differentially methylated genes seen on the Illumina 450K methylation array. Pyrosequencing, a sequencing by synthesis method, gauges the primer directed and polymerase extended incorporation of specific nucleotides to a bisulfite converted DNA template. Pyrosequencing quantitatively measures the incorporation of nucleotides since the amount of pyrophosphate emitted is ultimately converted to a proportional light signal. The degree of methylation is determined at each CpG position in the analysed sequence through the ratio of T and C.342

First, bisulfite converted DNA was PCR amplified using a forward and 5’ biotin labeled reverse primer pair designed using the PyroMark™ software (Qiagen, Venlo, Netherlands). Primers were designed to cover the exact CpG sites identified on the 450k methylation array. Once amplified, PCR products were immobilized on strepavidin sepharose beads, denatured to form single strands and annealed to the sequencing primer (designed by PyroMark™ software). Separation and annealing steps were enabled by the PyroMark Q24 prep vacuum workstation. Following annealing, CpG site methylation analysis was performed on the PyroMark Q24 pyrosequencing instrument and interpreted on the PyroMark Q24 software (Qiagen, Venlo, Netherlands).

Statistical analysis

In Paper I, the statistical significance of IGHV3-21 gene frequency differences incurred across our study and others was determined by a chi-squared
test. Overall survival was measured from the diagnosis date until the date of death or last known follow-up. Time to treatment was calculated from the diagnosis date to the date of first known treatment. Kaplan-Meier analysis visualized OS and TTT within the different groups studied in the form of survival curves. A log rank test was employed to determine the statistical differences in OS and TTT between the studied groups.

In Paper II, statistical differences in the frequency and size of the alterations between the studied groups were calculated via the Chi-square test, one way ANOVA test and t-test where appropriate. In Paper III, statistical differences in expression as determined through RQ-PCR were calculated by a t-test and represented in box plot graphs. In Paper IV, statistical differences in the level of methylation determined by pyrosequencing were calculated using a t-test. All statistical analyses were performed using the Statistica 8.0 and 10.0 software (StatSoft, Tulsa, OK).
Results & Discussion

Paper I: IGHV3-21 gene frequency in a Swedish population-based cohort

A decade ago, the importance of IGHV3-21 gene usage as an independent prognostic marker was discovered. In light of IGHV3-21 as a poor prognosticator irrespective of mutational status, a myriad of studies were then dedicated to investigate the prevalence and clinical outcome of this gene. Across European and North American CLL cohorts, CLL was established to display a biased IGHV gene repertoire with an over-representation of IGHV1-69, IGHV4-34, IGHV3-7, IGHV3-23 and IGHV3-21 genes. Additionally, studies continually demonstrated both mutated and unmutated IGHV3-21 patients to experience a poor survival, similar to that of other unmutated CLL cases utilising alternative IGHV genes. However, on comparing the predominance of this poor prognostic IGHV3-21 gene across different geographical regions, large variances in IGHV3-21 gene frequency were continually noted. Remarkably, Scandinavia has been shown to present with a high IGHV3-21 gene frequency (10-13%) compared to other regions namely the USA and Southern Europe (3-4%) (Table 3).

On recognising the geographical bias of IGHV3-21 and the particularly poor clinical outcome of these patients, attentions turned to the biological reasons behind such findings. Sequencing of IGHV3-21 gene rearrangements noted a large proportion of cases to carry stereotyped HCDR3 sequences. Moreover these stereotyped cases were shown to frequently present with restricted IGLV3-21 gene usage. Stemming from the findings of IGHV3-21 over representation and IGHV3-21 BCR sequence restriction, the prospect of a specific antigen contributing to IGHV3-21 leukemogenesis was suggested. Interestingly, some studies have reported stereotyped IGHV3-21 (subset #2) patients to have a shorter time to progression relative to their non stereotyped IGHV3-21 counterparts. That said, other studies have not found such a correlation, thus no consensus exists regarding the clinical impact of IGHV3-21 stereotypy in CLL.

Since many of these above findings are based on hospital-based CLL cohorts and are subjected to large cohort selection bias, it was consequently of inter-
est to determine the frequency of IGHV3-21 gene usage among a CLL Swedish population based cohort (n=337). Subsequently, we attempted to address the clinical impact of IGHV3-21 stereotypy and mutational status on CLL patient survival and progression.

This population-based study reports the Swedish IGHV3-21 frequency to be 6.5%. Although this frequency is lower compared to previous Scandinavian rates constructed from university hospital data (10-13%), the frequency still remains higher relative to other large studies form the USA and the Mediterranean (3-4%) (Table 3). In light of this finding, the differences in frequency encountered between small hospital-based cohorts demonstrating an over-representation of IGHV3-21, such as Ireland (7.9%), Britain (9.6%), and the Ukraine (5.8%) must be interpreted with caution, since discrepancies can be introduced by such factors as sample size and cohort selection bias. That said, despite the increased tendency to select for more aggressive CLL cases in hospital-based studies, such reports from the Mediterranean and USA still find IGHV3-21 frequency to remain rather low (3-4%). Despite the different nature of the CLL cohorts studied, in this population-based report versus our previous hospital-based study we reconfirmed that a true IGHV3-21 geographical bias exists in Sweden.

Table 3. Comparison of IGHV3-21 gene frequency in large CLL cohorts from Scandinavia, Mediterranean and USA

<table>
<thead>
<tr>
<th>IGHV3-21 cohort</th>
<th>Current Swedish population-based study</th>
<th>Hospital-based Scandinavian study</th>
<th>Multi-centre Italian study</th>
<th>Multi-centre Mediterranean Study</th>
<th>USA Study 149</th>
</tr>
</thead>
<tbody>
<tr>
<td>IGHV3-21</td>
<td>6.5%</td>
<td>10.1%</td>
<td>4.1%</td>
<td>3.5%</td>
<td>2.6%</td>
</tr>
</tbody>
</table>

Other explanations for disparity in global IGHV3-21 frequency may be explained by differences in genetic background, influencing the expansion of the IGHV3-21 precursor CLL cells. Interestingly, Iranian CLL hospital based studies have shown no IGHV3-21 usage. Alternatively, it is plausible that the IGHV3-21 geographic bias may reflect differences in the patient environment and the type of antigen exposure encountered. Given that we find a similar proportion of stereotyped cases in this report (14/22, 63.6%) as in other western CLL cohorts from the Mediterranean area (16/32, 50%), Italy (32/58, 55.2%) and the US (25/63, 39.9%), the notion of a native antigen primarily encountered in Sweden impacting on IGHV3-21 gene frequency and leukemogenesis is less reasonable. On the other hand, non-
stereotyped IGHV3-21 BCRs, given their heterogeneous molecular structure, may enable the binding of diverse antigens during leukemogenesis.

This study, like many previous, has established IGHV3-21 patients to experience a very poor overall survival (Figure 6).\textsuperscript{134,144,146-148} However, the clinical impact of IGHV3-21 stereotypy is less clear. Given the suggestion of specific antigen involvement in stereotyped IGHV3-21 CLL and since IGHV3-21 stereotypy has been affiliated to disease progression,\textsuperscript{5,146,148} it was of particular interest to determine its clinical impact on the disease. In a large collaborative European study incorporating 90 IGHV3-21 cases collected mainly from Sweden and Germany, no differences in survival were encountered between stereotyped IGHV3-21 patients and their non-stereotyped counterparts.\textsuperscript{147} In support of these findings, studies from the US and Italy encompassing 63 and 37 IGHV3-21 cases respectively corroborated the disassociation between survival and stereotypy. As mentioned, smaller studies from the Mediterranean (n=16 and n=32) have shown IGHV3-21 cases with stereotyped receptors to undergo a shorter time to progression relative to non-stereotyped cases.\textsuperscript{5,146} Perhaps these findings imply that the nature of the antigen may influence patient clinical behavior. That said, our study could not corroborate this finding, since no difference in time to treatment was seen between stereotyped and non-stereotyped IGHV3-21 patients (p=0.68) (Figure 6).

Again, the possible reason for this disparity between our study and others is perhaps explained by differences in cohort type and size. Alternatively, biological reasons such as a higher expression of the negative prognosticator CD38 and the higher occurrence of del(11q) documented in stereotyped relative to non-stereotyped IGHV3-21 cases may partially explain the poor disease course seen in these earlier studies.\textsuperscript{144} Although this study showed a trend towards having a higher frequency of del(11q) in stereotyped IGHV3-21 patients, a higher expression of CD38 was not observed in these cases. Nevertheless, other studies have implicated a high WNT-16 expression and expression of alternative negative indicators ZAP70, CD49d, and CD49b to contribute to the poor prognosis indicated in stereotyped IGHV3-21 patients.\textsuperscript{144}
In summary, using a population-based cohort we reconfirm the geographic bias in IGHV3-21 gene frequency previously identified in Scandinavia. Moreover, we determined that IGHV3-21 patients experience an inferior outcome independently of stereotypy. Regarding the prognostic role of stereotypy in IGHV3-21 CLL, further large collaborative studies are needed before a definite conclusion can be made since a large heterogeneity among published studies exists.
Paper II: High density array screening for genomic events in stereotyped subsets of CLL

Although the advancement of new prognostic markers has allowed improved subdivision of CLL, the causes of CLL still remain unknown and there is yet no single genetic event indicated in all cases. Since the existence of multiple subsets of CLL expressing ‘stereotyped’ BCRs implies a role for antigen(s) in CLL,⁵,⁶ and considering ‘stereotypy’ may influence CLL clinical course, it was of interest to determine if biases existed in the genomic profiles of stereotyped patients. As mentioned, both the stereotyped IGHV3-21 subset #2 and IGHV3-21 non-stereotyped counterparts are associated with a poor overall survival.¹³⁴,¹⁴⁷,¹⁴⁹ Moreover, some studies report subset #2 to be associated with a more progressive disease⁵,⁶,¹⁴⁶ relative to non stereotyped IGHV3-21, a finding we did not find evidence of among our own cohort of Swedish stereotyped IGHV3-21 patients in Paper I. On the other hand, stereotyped IGHV4-34 subset #4 is associated with an indolent disease course compared to its non-stereotyped counterparts.⁵,⁶ Regarding IGHV4-34 subset #16, little is known about this less commonly encountered subset. Thus, the aim of Paper II was to perform a detailed genetic characterization of clinically divergent stereotyped CLL subsets using high-resolution SNP arrays, to identify genetic events that may give clues to their pathogenesis and diverse clinical outcomes. In particular this study focused on the comparison of poor prognostic IGHV3-21 subset #2 and favourable prognostic IGHV4-34 subset #4 patients. In an attempt to identify stereotyped specific events, the genetic profiles of stereotyped subsets were compared to their non-stereotyped counterparts and a general CLL cohort.

SNP analysis revealed a higher number of poor prognostic IGHV3-21 samples to carry alterations, i.e. 97% of subset #2 and 92% of non-subset #2 samples. In contrast, the proportion of subset #4 samples bearing alterations was considerably less (76%). Similarly, 75% of subset #16 samples carried aberrations, whereas 88% of the more clinically heterogeneous non-subset 4/16 cases displayed alterations. Besides the marked differences in the overall frequencies of samples affected with CNAs, differences in the genomic complexity regarding the number, type and size of alterations encountered were noted between the study groups. Overall, subset #2 and non-subset #2 displayed a higher average number of aberrations compared to subset #4. More specifically, 48% of subset #2 and 39% of non-subset #2 samples carried ≥3 alterations compared to 18% of subset #4 and 13% of subset #16 (Figure 7). In fact, the majority (53%) of subset #4 patients carried only one alteration. The increased genome complexity observed for IGHV3-21 patients, particularly for subset #2, can be partially accounted for by the high frequency of the known recurrent alterations observed. Interestingly, known recurrent alterations were detected in 90% of subset #2 cases compared to,
54% of non-subset #2 patients. More noticeably, a low frequency of the known recurrent alterations was seen among the favourable prognostic subset #4 patients (35%). Similarly, a low frequency was detected in subset #16 (25%). However, IGHV4-34 non-subset 4/16 patients displayed a similar frequency of known recurrent alterations compared to IGHV3-21 patients (59%).

Figure 7. Level of genomic complexity (CNA number and size).

Notably, poor prognostic deletions of 11q were more frequent in subset #2 and non-subset #2 (31% and 23%) relative to subset #4 and IGHV4-34 non-subset #4/16 patients (0% and 12%). Furthermore, deletions of 13q were particularly high in subset #2 cases and concurrent alterations at 11q and 13q were also found to be common in these patients. In contrast, deletion of the good prognostic marker del(13q) was the only known recurrent aberration.
detected in subset #4 (35%) (Table 4). On considering the larger number of alterations and the high frequency of poor genomic markers carried by subset #2 in comparison to subset #4 patients, these findings suggest that subset #2 are more likely to acquire genomic events relative to subset #4. These events may represent important genetic events during the pathogenesis of IGHV3-21 CLL particularly when considering the frequent finding of concurrent 11q and 13q deletions, especially in subset #2 cases (Table 4). Hypothetically, these aberrations may be acquired during the phase of active stimulation by (unknown) antigens, which trigger the IGHV3-21 precursor cell to undergo rapid cell division and hence, under such proliferative pressure acquire such clonal alterations. In addition, it is interesting to speculate that these observations may account for the adverse clinical outcome reported for IGHV3-21 CLL, particularly for subset #2 patients.

Table 4. Known genomic alterations in IGHV3-21 and IGHV4-34 subsets compared to a general CLL cohort.

<table>
<thead>
<tr>
<th>Subset/ non-subset</th>
<th>del(11q) (poor prognosis)</th>
<th>del(13q) (good prognosis)</th>
<th>trisomy 12 (intermediate prognosis)</th>
<th>del(17p) (poor prognosis)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subset #2</td>
<td>31%</td>
<td>79%</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Non-subset #2</td>
<td>23%</td>
<td>31%</td>
<td>23%</td>
<td>7.7%</td>
</tr>
<tr>
<td>Subset #4</td>
<td>0</td>
<td>35%</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Subset #16</td>
<td>13%</td>
<td>0</td>
<td>13%</td>
<td>0</td>
</tr>
<tr>
<td>Non-subset #4/16</td>
<td>12%</td>
<td>44%</td>
<td>6%</td>
<td>3%</td>
</tr>
<tr>
<td>General CLL cohort</td>
<td>13%</td>
<td>54%</td>
<td>11%</td>
<td>4%</td>
</tr>
</tbody>
</table>

Recent reports suggest that not only the type of alteration but the size of the alteration and the nature of their breakpoints influence clinical outcome. In this study, alteration size accounted for another layer of genomic complexity. Size distribution analysis showed subset #2 and non-subset #2 patients to have a higher number of alterations greater than 1Mb in size compared to the other subgroups, particularly subset #4. This can be accounted for by the high frequency of large known recurrent alterations seen in IGHV3-21 patients, particularly in subset #2. However, further investigations revealed that both subset #2 and non-subset #2 have similar average number of CNAs/sample comparable to the level observed in our recent array-based study on Swedish CLL. Hence, subset #2 and non-subset #2 cases do not appear to be more complex when accounting for all genomic alterations compared to other CLL patients in general. Their crux appears to come from
their susceptibility to acquire known recurrent alterations and larger sized alterations, which might be reflective of their more active proliferative status.

On the other hand, subset #4, #16 and non-subset #4/16 showed a higher prevalence of smaller alterations <1Mbp, owing to the lack of large recurrent alterations such as del(11q), trisomy 12 and del(17p) in these groups. Additionally, 13q deletions in subset #4 were shown to be of a smaller average size compared to those encountered in the IGHV3-21 groups. Overall, on considering the lower number and size of alterations in subset #4 and the lack of poor prognostic alterations compared to subset #2, these features suggest that subset #4 patients are less likely to accumulate genomic alterations. This is suspected to be due to a low proliferative drive stemming from the IGHV4-34 BCR. Altogether, the lack of such genomic aberrations may in part be responsible for the indolent disease course experienced by subset #4 CLL patients.

Novel events were identified on chromosomes 2q, 7q and 14q in IGHV4-34 patients; however these were not attributed to any particular subset. Recurrent gains at 2q and losses at 3p were demonstrated in two subset #2 cases. Recurrent CNN-LOH events were mainly detected on chromosome 13q, although independently of stereotypy status.

This study concludes that genomic aberrations were more common in subset #2 and non-subset #2 compared to subset #4. The particularly high frequency of del(11q) in subset #2 may be linked to the adverse outcome reported for these patients. In contrast, the lower prevalence and size of CNAs and the absence of poor-prognostic aberrations in subset #4 may reflect the indolent disease course seen in these patients.

Paper III: Array-based methylation profiling in prognostic subsets of CLL

Aberrant DNA methylation is a key component in augmenting tumorigenesis, where genome-wide hypomethylation and regional hypermethylation of tumor suppressor gene (TSG) promoters are signature characteristics of many cancers. Aberrant DNA methylation is a key component in augmenting tumorigenesis, where genome-wide hypomethylation and regional hypermethylation of tumor suppressor gene (TSG) promoters are signature characteristics of many cancers. As mentioned, in CLL, the candidate gene approach has revealed a significant association linking promoter methylation to transcriptional silencing for a small number of individual gene promoters, namely DAPKI, ZAP70 and TWIST-2. Since the global methylation profiles of CLL were yet to be discovered, Paper III sought to investigate genome-wide DNA methylation patterns in three divergent CLL prognostic sub-
groups; IGHV mutated, unmutated and IGHV3-21 patients. Considering the clinical and biological diversity of these subgroups, Paper III set out to perform a detailed genome wide, albeit somewhat promoter biased, epigenetic characterization using the Illumina 27K high-resolution DNA methylation microarray, as an important means to identify differentially and aberrantly methylated genes that may play a role in their pathogenesis and diverse clinical outcome. Moreover, as methylation changes are reversible with the use of de-methylating agents, whole genome scanning may help identify new therapeutic targets in CLL. Hence, Paper III investigated the significance of DNA methylation in regulating gene promoters by re-inducing the expression of methylated TSGs using the methyl-inhibitor 5-aza-2'-deoxycytidine and the histone deacetylase inhibitor TSA.

Corroborating other studies, Paper III demonstrated that CLL is hypomethylated compared to normal healthy individuals on the global level. Furthermore, by applying stringent bioinformatic approaches we identified for the first time distinct differential methylation profiles in divergent CLL prognostic subgroups. A number of significantly differentially methylated genes occupying key cellular pathways and functions were revealed between the studied subgroups. For instance, 7 TSGs were preferentially methylated and potentially silenced in the poor prognostic IGHV unmutated subgroup (e.g. $VHL, ABI3$) (Table 5). In contrast, these genes remained unmethylated and hence potentially expressed in the favourable prognostic IGHV mutated CLL. That said, the IGHV mutated subgroup did display methylation of 2 TSGs, $PPP1R3A$ and $WISP3$ that were unmethylated in IGHV unmutated patients (Table 5). Similarly, most methylated TSGs found in the IGHV unmutated subgroup were found to be unmethylated in IGHV3-21 patients, one exception was the preferential methylation of the $SLC22A18$ gene. These observations suggest that the specific methylation patterns of certain TSGs may play a role in the pathogenic pathways that lead to these clinically divergent subgroups. That notwithstanding, many of the TSGs identified in this study have been implicated in several other cancers, however, their role in CLL has yet to be determined (Table 5).
Table 5. A brief summary of the differentially methylated genes identified between IGHV mutated and IGHV unmutated CLL

<table>
<thead>
<tr>
<th>Gene function</th>
<th>IGHV Mutated-CLL</th>
<th>IGHV Unmutated-CLL</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aiding tumourigenesis</td>
<td>Methylated</td>
<td>Unmethylated</td>
<td>\textit{ADOR\textsubscript{A}}3\textsuperscript{360}, ANGPT2\textsuperscript{361}</td>
</tr>
<tr>
<td>Suppressing tumourigenesis</td>
<td>Unmethylated</td>
<td>Methylated</td>
<td>\textit{VHL}\textsuperscript{355}, \textit{ABI3}\textsuperscript{356}, \textit{LCOD1}\textsuperscript{362}</td>
</tr>
<tr>
<td>Anti-apoptotic</td>
<td>Methylated</td>
<td>Unmethylated</td>
<td>\textit{BCL2}\textsuperscript{85}</td>
</tr>
<tr>
<td>Pro-apoptotic genes</td>
<td>Unmethylated</td>
<td>Methylated</td>
<td>\textit{BCL10}\textsuperscript{365}</td>
</tr>
</tbody>
</table>

Certain cellular pathways, such as NFκB, PI3K/Akt and MEK/ERK, are dysregulated in CLL B cells leading to activation of anti-apoptotic pathways.\textsuperscript{113,114,117,364,365} Our data strengthens this idea and further indicates that certain genes involved in such ‘proliferative’ pathways may be regulated by DNA methylation in CLL. For example, genes activating the NFκB pathway, such as \textit{ADOR\textsubscript{A}}3\textsuperscript{360} and \textit{CARD15},\textsuperscript{366} were specifically unmethylated in the unmutated CLL subgroup. Similarly, the \textit{LOC340061}\textsuperscript{367} and \textit{PRF1}\textsuperscript{368} genes, which activate the NFκB and MAPKinase pathways, respectively, were unmethylated and therefore potentially expressed in poor-prognostic IGHV3-21 CLL compared to IGHV mutated CLL. In line with these findings, genes repressing the NFκB pathway (\textit{LCOD1} gene)\textsuperscript{362} and MAPKinase (\textit{ZNF540} gene)\textsuperscript{369} pathways were generally methylated in both IGHV unmutated and IGHV3-21 samples in contrast to IGHV mutated CLL. Again most genes have not been associated with CLL, while many are known to be involved in the tumourigenesis of several other cancers and leukaemias.

Genes known to facilitate tumourigenesis within other cancer types, such as \textit{IFNB1}\textsuperscript{370} and \textit{IL17RC},\textsuperscript{371} were noted to be preferentially unmethylated in the IGHV unmutated and IGHV3-21 CLL subgroups in comparison to IGHV mutated CLL. Similarly the \textit{LMO2} oncogene\textsuperscript{372} and \textit{RASGRP3}\textsuperscript{373} which regulate B cell proliferation by facilitating B-cell receptor-Ras signaling, were unmethylated in IGHV3-21 samples. Thus, the above data indicates that poor-prognostic CLL subsets favor expression of genes aiding tumor proliferation.

Interestingly, array analysis provided insights into the deregulation of anti- and pro-apoptotic genes in IGHV mutated CLL. For instance, \textit{PLD1}, a known anti-apoptotic gene,\textsuperscript{374} remained unmethylated in IGHV mutated CLL.
samples. On the other hand, pro-apoptotic genes such as *BCL10* and *TP53* were methylated in IGHV mutated samples unlike IGHV unmutated CLL.

In short, genes related to prognosis, for instance the *ANGPT2* gene, which has been implicated in progressive CLL disease states were shown to be preferentially methylated in the favourable prognostic mutated CLL cases while being unmethylated in unmutated CLL (Table 5). In addition, the *NGFR* gene which has been associated with a favorable prognosis in ALL, was shown to be unmethylated in IGHV mutated CLL and showed increased expression on the RNA level within this subset.

Independent verification of array results was performed for a select number of genes of different functionalities using an independent CLL sample set. MSP-PCR semi-quantitatively verified the methylation status of four methylated genes (*PRF1, ADORA3, IGSF4* and *BCL10*), RQ-PCR demonstrated the correlation between gene expression and methylation status for 4 genes (eg. *VHL, ABI3* and *NGFR*) in an independent CLL cohort and finally, bisulfite sequencing further quantified the degree of methylation for two genes (*ABI3* and *VHL*).

Finally, through the inhibition of DNA methylation using methyl- and HDAC inhibitors, this study evidenced the induced expression of methylated TSGs (e.g. *IGSF4, ABI3* and *VHL*) in unmutated CLL primary samples. In general, although a low induction of gene expression was observed with methyl inhibitor treatment only, a more pronounced effect was seen when the methyl inhibitor and histone acetyl inhibitor drugs were applied in combination. That said, the low induction can be partially explained by the fact that CLL primary cells do not proliferate at a high rate under *in vitro* conditions. These data indeed strengthen the role of DNA methylation in regulating promoter expression and open up the possibility of targeted therapy in CLL (Figure 8).
In conclusion, Paper III shows distinct methylation profiles encompassing a number of candidate genes involved in cellular pathways regulating proliferation and apoptosis in IGHV unmutated/mutated and IGHV3-21 CLL and thus underlines the critical role of epigenetic changes during leukemogenesis.

**Paper IV: DNA methylation screening of CLL samples over time and within different CLL compartments**

On identifying differential methylation profiles in IGHV mutated and unmutated CLL in Paper III, we set out to further investigate their methylation profiles using a higher resolution 450K global DNA methylation microarrays. More specifically, we aimed to determine whether the DNA methylation profiles of these subgroups changed over time during the course of disease and whether DNA methylation of CLL cells differed according to the anatomical site from which the cells resided. The first part of this study, in accordance with Paper III, again identified IGHV mutated and unmutated CLL to have differential DNA methylation profiles, albeit on a larger scale than previously reported. Specifically we found 2239 differentially methylated sites between these subgroups compared to a mere 99 sites identified previously. We found a relatively even distribution of differentially methylated sites across the whole genome. However, we noted that the largest proportion of differentially methylated sites was located outside of the CpG islands and within gene bodies. Only recently with the advent of whole ge-
nome analysis has the relevance of this latter finding come to light. Traditionally DNA methylation analysis has solely focused on CpG islands, particularly within gene promoters, since it is here that gene regulation is primarily regulated. In Paper III, we identified DNA methylation of ADORA3, positioned approximately 1kb away from the promoter, to have a regulatory function on gene expression. On a much larger scale Irizarry et al\textsuperscript{286} further corroborates this finding, showing the majority of altered DNA methylation sites in colon cancer to reside outside of CpG islands in regions up to 2kb distant from the CpG island in regions called CpG island shores. Moreover, these sites were shown to correlate to gene expression.

Gene ontology analysis identified a number of enriched physiological and molecular functionalities, most of which have not been implicated in CLL before. Many genes were known as members of established signaling pathways, whereas others were shown to be involved in other cancers. For instance, CpG sites within DNA repair genes, DNA methyltransferase genes $DNMT3B$,\textsuperscript{256} members of the polycomb complex $EED$,\textsuperscript{377} $JARID2$\textsuperscript{378} and polycomb target genes were detected as differentially methylated. Specifically 230 CpG sites covering 158 polycomb target genes previously characterized in embryonic fibroblasts\textsuperscript{379} presented as differentially methylated between IGHV mutated and unmutated CLL. Approximately half of the sites presented with a higher methylation level in unmutated CLL, whereas the other half were found to have a higher level within the mutated subgroup. Preferential methylation of developmental polycomb target genes has been frequently recognised in other forms of cancer.\textsuperscript{380,381} That said, the exact biological relevance of this finding is largely unexplored in cancer. What is known is that tumor-specific targeting of \textit{de novo} methylation at these targets is pre-programmed by an established epigenetic system namely H3K27me3 that normally has a role in marking embryonic genes for repression. Unlike normal cells, the EZH2 polycomb complex maintaining H3K27me3 recruits DNA methyltransferase to these sites in cancer cells.\textsuperscript{381}

A focused search for genes formerly implicated in CLL, found a number of survival related ($TCF3$,\textsuperscript{382} $LEF1$\textsuperscript{383}), proliferation ($CD80$,\textsuperscript{384} $CD86$\textsuperscript{385}) and apoptosis ($FAS$)\textsuperscript{386} and BCR signaling related genes ($IBTK$).\textsuperscript{387} Moreover, 12 differentially methylated genes identified in Paper III were found to overlap to this present study and were shown to be largely concordant in terms of methylation status. For example, the preferential methylation of tumour suppressor genes $ABI3$\textsuperscript{356} in IGHV unmutated and $WISP3$\textsuperscript{358} in IGHV mutated subgroups was re-confirmed. In support of Corcoran \textit{et al}.,\textsuperscript{300} we found the CLL prognostic gene $ZAP70$,\textsuperscript{163} to be preferentially methylated in IGHV mutated patients relative to unmutated cases. Furthermore, Corcoran \textit{et al}.,\textsuperscript{300} have shown that methylation of $ZAP70$ correlates well to $ZAP70$ expression. Recently, methylation analysis of $ZAP70$ intron 1 by pyrosequencing has
been shown to be a robust method correlating to ZAP70 and CD38 expression, IGHV mutation status, overall survival and time to treatment. Given the lack of sufficient standardisation with current flow cytometry quantitation methods, quantification by pyrosequencing can provide an alternative mode of analysis. For the first time, CLL prognostic genes LPL and CLLU1 were noted to be differentially methylated between IGHV mutated and unmutated CLL. More specifically, higher DNA methylation within the gene body of CLLU1 and the transcription start site TSS1500 and gene body region of LPL in IGHV mutated CLL relative to IGHV unmutated patients was demonstrated. In recent times, expression of both LPL and CLLU1 has been noted to be higher in IGHV unmutated patients compared to IGHV mutated cases. In CLL, high expression of these genes is associated with poor clinical outcome. In view of these observations, it is interesting to speculate that perhaps DNA methylation of these genes may contribute to their differential expression seen in CLL.

One of the enriched canonical pathways highlighted through IPA analysis was the molecular mechanisms of cancer pathway. Within this IPA derived pathway, we identified pathways previously implicated in CLL namely the NF-κB/TNFR1 and TGF-β pathways. Of particular interest is the TNF/NF-κB pathway. Here ten members of this pathway were found to be differentially methylated. For instance, a CLL autocrine growth factor TNF, TNFRSF8 (CD30), a receptor for thioredoxin (Trx) a survival factor for CLL cells and two TNF receptor genes, TNFSF1A and TNFRSF1B, recently shown to be redox-controlled by protein-disulfide isomerase (PDI) in complex with thioredoxin-1 (Trx1), were found to be differentially methylated in mutated and unmutated CLL. Although these findings are interesting, relating the role of DNA methylation to the functionalities of such a pathway in CLL is difficult. The variable incidence of multiple or single CpG sites across genes at variable locations along different gene regions (gene body, TSS1500 etc) makes interpretation challenging. Sometimes multiple CpG sites across the same gene at different regions will have variable methylation values. These challenges are further complicated by the fact that the biologically relevant level of DNA methylation needed to effect gene expression is currently unknown.

With the prospect of aberrant DNA methylation marks as candidate prognosticators, it is essential that such marks be stable in terms of DNA methylation level. To date, a limited number of studies exist documenting the stability of DNA methylation over time. Of the few studies available variable results ensue. This is perhaps in part reflective of the different diseases and nature of the cohort studied in these reports. However, more pressing inconsistencies stem from the alternative techniques employed and differences in the processing and interpretation of DNA methylation data. For example, a
report examining bladder cancer metachronous tumours from two individuals, has shown DNA methylation to be relatively stable over time both within the individual and between patients. In contrast, a study by Björnsson et al has identified inter-individual age variation and differences occurring over time within individuals, that would otherwise be over-looked by group averaging of the data. These changes were shown to be clustered within family members highlighting the possibility that genetic background may effect DNA methylation changes. Nevertheless, a study of DNA methylation in 40 monozygotic twins found 35% of twin couples to have a disconcordant level of total DNA methylation, particularly those over the age of 28 years. In this instance, different environmental influences and exposures were thought to play a role.

In light of the above findings we dedicated the second part of the study to examine DNA methylation over time in paired CLL diagnostic and follow-up patients. Based on average DNA methylation values across the groups, this study for the first time established that global DNA methylation is relatively stable over time. Both IGHV mutated and unmutated patients incurred only a few, small recurrent changes over time. Interestingly however, on analyzing the intra-individual changes a larger difference in the number of non recurrent differentially methylated sites were noted over time between IGHV mutated and unmutated cases, that would otherwise be missed on group averaging of the data. More specifically, intra-individual analysis of IGHV unmutated patients showed a large difference whereas intra-individual investigation of IGHV mutated patients showed no significant difference in the number of non-recurrent differentially methylated sites over time (Figure 9). That said, in global terms, these non-recurrent changes in IGHV-unmutated cases equated to a very small overall global change (<1%), when taking into account the 485,000 sites available on the array. Nevertheless, it is interesting to speculate that perhaps these findings are reflective of the more indolent and aggressive disease course associated with IGHV mutated and unmutated patients respectively. Notably, global DNA methylation within IGHV unmutated patients over time was relatively stable despite the implementation of treatment between diagnosis and follow-up. Thus it appears that DNA methylation of CLL cells is not largely effected by therapy, at least not in respect to overall DNA methylation.

It is difficult to ascertain whether or not some of these non-recurrent DNA methylation changes act as driver epi-mutations causing the expansion of the CLL clone. A more likely explanation, given the lack of recurrent changes over time, indicates that the changes encountered merely act as passenger events arising from the higher proliferative nature in the preexisting unmutated clone. Nevertheless, it is interesting to speculate that perhaps these
changes may contribute to CLL leukemogenesis by contributing to overall epigenomic instability.

Figure 9. Box plot visualizing the difference in number of non recurrent differentially methylated sites overtime within and between patient matched diagnostic and follow-up IGHV mutated and IGHV unmutated CLL patients.

Resting PB CLL cells are known to communicate with a pro-survival microenvironment in proliferation centers of the LN resulting in proliferation (~1%/day) of the CLL clone. Recently, a higher expression of genes involved in proliferation and active BCR signaling were shown in the LN relative to the blood of patient-matched CLL samples. In light of the anti-correlation of DNA methylation to gene expression, we hypothesized that DNA methylation may partake in the mechanism of differential gene expression seen within the different compartments of CLL. Furthermore, in mind of the possibility of aberrant DNA methylation marks as candidate prognostic markers, it is important to establish whether such marks are stable over different sampling sites. Therefore, in the third part of the study, we examined the DNA methylation profiles of patient-matched CLL cells derived from the PB and LN, to ascertain the influence of the microenvironment on CLL patient DNA methylation profiles. This study for the first time found that global DNA methylation is rather similar in patient matched PB and LN samples. Only few, small recurrent changes were identified. Moreover, the number of non-recurrent changes although relatively low in global terms (<1%), were found to vary in number across the intra-individual PB and LN pairs. Therefore it seems that unlike gene expression, DNA methylation in CLL cells is not heavily influenced by the microenvironment.
In sum, on a larger scale to Paper III, this study reconfirms IGHV mutated and IGHV unmutated CLL to have differential methylation profiles. More specifically, CLL prognostic genes, polycomb target genes, and key members of TGF-β and NF-κB/TNFR1 pathways were noted to be differentially methylated. Finally, global DNA methylation was shown to be relatively stable over time and to be remarkably similar in CLL cells derived from different anatomical CLL compartments.
Conclusions

This thesis applied IGHV gene sequencing and high resolution SNP and DNA methylation microarray techniques in order to interrogate genetic and methylation events that may contribute to CLL leukemogenesis in diverse prognostic subgroups and subsets of CLL.

Paper I

- Within the Swedish population-based cohort studied, the geographic bias in IGHV3-21 gene frequency previously identified in Sweden was validated, albeit at a lower frequency than formerly reported in our hospital-based cohort study.

- IGHV3-21 patients were shown to have an inferior outcome independently of stereotypy. Hence, no prognostic value could be attributed to IGHV3-21 stereotypy status.

Paper II

- Poor prognostic IGHV3-21 subset #2 patients commonly carried more genomic aberrations and larger sized alterations compared to favourable prognostic IGHV4-34 subset #4 patients. Thus, it appears that subset #2 patients are more genomically complex relative to subset #4.

- IGHV4-34 subset #4 patients exclusively carried the favourable prognostic del(13q) marker, whereas poor prognostic subset #2 patients carried a high frequency of the poor prognostic del(11q) marker. Hence, these findings may in part be reflective of the favourable prognosis seen in subset #4 and the adverse outcome seen in subset #2 patients.

Paper III

- Distinct differential methylation profiles were detected in three divergent prognostic subsets of CLL, IGHV mutated, unmutated and IGHV3-21.
Differentially methylated genes, such as TSGs, were preferentially methylated in the poor prognostic IGHV unmutated subgroup compared to the favourable prognostic IGHV mutated CLL. Moreover, genes involved in cell proliferation were specifically unmethylated in the poor prognostic IGHV unmutated and IGHV3-21 subgroups. Hence, these findings suggest that the methylation status of certain key cellular genes may contribute to CLL pathogenesis and the divergent clinical outcome seen in certain CLL subsets.

- Through the use of methyl inhibitor and histone deacetylation drugs, the re-expression of methylated TSGs (ABI3, VHL) was demonstrated. Hence the reversible nature of DNA methylation was functionally established.

### Paper IV

- On a larger scale to Paper III, this study reconfirmed that IGHV mutated and IGHV unmutated CLL are characterized by differential methylation profiles. Differentially methylated sites between these subgroups primarily presented outside of CpG islands. Preferential DNA methylation of CpG sites within known CLL prognostic genes **CLL1** and **LPL** was notably higher in IGHV mutated cases relative to IGHV unmutated patients. Key members of TGF-β and NF-κB/TNFR1 pathways and large numbers of polycomb targets were also shown to be differentially methylated between IGHV mutated and unmutated CLL.

- Global DNA methylation in paired diagnostic and follow-up IGHV-UM (treated) CLL and IGHV-M (untreated) CLL patients was deemed to be relatively stable over time. Hence, this finding supports the notion that aberrant DNA methylation events are early events in CLL pathogenesis.

- DNA methylation changes in patient-matched peripheral blood and lymph node samples were found to be relatively similar. Hence, it seems the differences in the PB and LN micro-environment does not heavily influence the DNA methylation profiles of CLL cells.

Taken together, these studies provide a greater insight into the genetic alterations and DNA methylation events occurring in divergent CLL prognostic subgroups.
Acknowledgements

This work was carried out at the Department of Immunology Genetics and Pathology, Rudbeck Laboratory, Uppsala University in association with the School of Biological Sciences, Dublin Institute of Technology, Ireland. Financial support was provided by the Nordic Cancer Union, the Swedish Cancer Society, the Swedish Medical Council and ABBEST Ireland.

First of all I would like to sincerely thank my supervisor, Prof Richard Rosenquist. Thanks for taking a chance on me and for giving me the opportunity undertake such rewarding projects.

Thank you to my co-supervisor Fergus Ryan and my most recent co-supervisor Larry Mansouri. You have always encouraged me whenever I needed it. Thanks for all the assistance, words of advice and suggestions I really appreciate it!

To my lab lads and lassies, both past and present, Meena, thanks for allowing me to collaborate with you at the start of my studies, I will never forget our most interesting discussions!!!. Fiona, thanks for keeping me going, making me laugh albeit sometimes at my own expense!! You’re all right for a south-sider! Maria! Spiggelty wiggelty!!! Thanks for taking an interest in using Irish slang you are only ‘savage’ at it now! Marie, thanks for your good humour and sound advice!!! Rebeqa! I really enjoyed working with you on the SNP Paper, thank you for keeping me focused! Lesley! Thanks for everything!! You truly are a great friend and a great agony aunt!!. Thanks for taking me in when I first came to Uppsala and keeping me forever entertained Mi, thanks a million for all the fun times and laughs!! To Arifin, Ja, Ingrid, Anna, Djana, Cecilia and Andreas thanks for making the lab such a great place to work.

I would also like to thank Christer Sundstrom for all his help in finding samples and for “donating” the occasional few millilitres of blood!!! Special thanks to our Swedish collaborators Lotta Bergh and Anders Rosen!! Thanks to our other European collaborators. Kostas Stamatopolous, Fred Davi, Christian Geisler and Gunnar Juliusson for all their contributions! Sincere thanks to Hanna Göransson Kultima and Isaksson for all their hard bioinformatic work.
To all the Rudbeck/Akademiiska/Ångstrom/BMC gang!! Where do I start…. you know who all you trouble makers are….but just in case others don’t know, I feel its my duty to warn them…Jelena, Lucy, Johan S, Chris, Johan K, Sara K, Jess, Ammar, Paul, Demet, and Daniel and Tatjana. Thanks for keeping me sane!!!! More importantly thanks for making my 4 years here enjoyable. I will miss all the parties, laughs and friendship!

To Antonia, Charlotte, Prasoon, Lotta S, Pernilla, Helena JW, Fredrik Ö, Chandra’s army- (Tanmoy, Deb, Matthieu, Sanhita Gaurav), Tobias B, Grzegorz, Viktoria A Spyros, Xiang, Umash, Jimmy, Sara L,Yanara, Lothar, Marie H, Sonia, the dutch dudes, Eva H, Annika H, the genome centre girlies, the teaching crew-Viktoria, Angie, Madhu, Rachel, Lena Å and Hamid … big thanks for making Rudbeck a cool place to work!!!

I would like to thank Joe Vaughan, Brid Ann Ryan and Paddy Mc Hale (DIT) for helping me and encouraging me over my undergraduate and post graduate years!

To my Irish mates, Dave, Eileen, Sarah, Gav, Lynne, Cliona and Fats!! Thanks for coming over to visit me and staying friends with the NERD!!!

Big thanks to all my Swedish buddies!!!

To my Swedish family the “Janssons”. Thanks for taking me in as one of your own. You have truly looked after me!!! Special thanks to pappa. Tack för alla sms.!!!

To my family, thanks for always being there and never doubting me! Thanks for your never ending support even if you still don’t understand what exactly what I have been doing over here!!!!!!! “NO… dad I am not still writing my essay for school… this is it… what you think?”

Last but by no means least…Mattias! I know I might regret putting this in writing…but you have the patience of a saint. I can simply say that without you I wouldn’t have made it.!! Thanks for everything xxx
References

56. Lefranc MP, Giudicelli V, Kaas Q, et al. IMGT, the international ImMunoGeneTics information system. Nucleic Acids Res. 2005;33:D593-597.


235. Buhmann R, Kurzeder C, Rehklau J, et al. CD40L stimulation enhances the ability of conventional metaphase cytogenetics to detect chromosome aberra-


353. Matthews C, Catherwood MA, Morris TC, Alexander HD. V(H)3-48 and V(H)3-53, as well as V(H)3-21, gene rearrangements define unique subgroups in CLL and are associated with biased lambda light chain restriction, homologous LCDR3 sequences and poor prognosis. Leuk Res. 2007;31:231-234.


Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 745

Editor: The Dean of the Faculty of Medicine

A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine.