Clonality in adult T-cell leukaemia/lymphoma

A thesis submitted to Imperial College London

for the degree of Doctor of Philosophy

by

Lucy B.M. Cook

Imperial College London

2013

Section of Immunology

Department of Medicine

Wright-Fleming Institute

Imperial College London

Norfolk Place

London W2 1PG

Declaration

All work presented in this thesis is the author's own other than where clearly stated in the 'Statement of Collaboration'.

Signature_____

Date_____

Summary

Human T-lymphotropic virus type 1 (HTLV-1) is a retrovirus that persists lifelong within the infected host by driving expansion of infected CD4+T-cells. It is the cause of adult T-cell leukaemia/lymphoma (ATL), an aggressive CD4+ T-cell malignancy, which arises in approximately 5% of individuals typically following decades of asymptomatic infection. The reasons why some individuals develop ATL remain unknown.

In this laboratory a novel customised high throughput sequencing and bioinformatic method has been developed in order to map and accurately quantify the proviral integration sites within each host genome in order to identify clonal populations within each host. In this study I aimed first to test the hypothesis that there is a single provirus integrated into each host genome, and secondly to test the hypothesis that the site of retroviral integration determines the risk of leukaemia.

In order to quantify the average number of proviral integration sites in each host cell, we isolated infected T-cells from the peripheral blood of infected individuals by limiting dilution cloning. Integration site analysis of these clones revealed that in natural infection each T-cell clone carries a single integrated provirus. This work formed the basis of a publication in the journal *Blood* (Cook et al 2012).

I describe the systematic analysis of the clonality, structure and the integrity of the proviral *tax* gene in a large cohort of ATL patients (n=197). I correlate these findings with the clinical subtype of ATL and the landscape of the host genome flanking the proviral integration site. Based upon our findings we conclude that the integration site *in cis* does not directly cause leukaemogenesis and hypothesise that the absolute number of infected clones within an individual, and not oligoclonal proliferation, predisposes to malignant transformation.

Acknowledgements

Firstly I would like to thank my supervisor Charles Bangham for his guidance, encouragement and patience over the last few years. I would like to extend this thanks to Graham Taylor and his clinical team who have welcomed me into the HTLV-1 service at St Mary's and supported my transition back to clinical medicine to develop a clinical ATL interest.

I would like to thank all former and current members of the Bangham and Taylor labs for their day-to-day good humour and for teaching me how to hold a pipette, not laughing too loudly at my mistakes and encouraging my transition from the Excel spreadsheet to R. In particular I would like to thank Anat Melamed, Aileen Rowan, Heather Niederer and Yorifumi Satou.

This work would not have been possible without collaboration from Professor Masao Matsuoka, Kyoto University who supported this project, provided patient material and clinical information to undertake the work and for helpful discussion with regards to the results. I would also like to thank Leukaemia and Lymphoma Research for funding my fellowship and for providing additional support to travel to Kyoto University.

Finally, love and thanks to my husband Stewart and the rest of the Cooks for their unswerving support and patience, and for not being too disappointed that I haven't found the cure for cancer.

Table of Contents

List of figures

List of tables

List of abbreviations

AC	Asymptomatic carrier
AP1	Activator protein 1
APOBEC3G	Apolipoprotein B mRNA-editing, enzyme-catalytic, polypeptide-like 3G
ATL	Adult T-cell leukaemia/lymphoma
BCP	Bayesian change point
BLV	Bovine leukaemia virus
BSP	Bisulfite sequencing PCR
CCL22	C-C motif chemokine 22
CCR4	C-C chemokine receptor type 4
CGH	Comparative genomic hybridisation
ChIP-seq	Chromatin immunoprecipitation-sequencing
CNS	Central nervous system
CRE	c-AMP responsive element
CREB/ATF	CRE binding/activating transcription factors
CSF	Cerebrospinal fluid
CTL	Cytotoxic T-lymphocyte
ELAND	Efficient Large-Scale Alignment of Nucleotide Databases
FBS	Fetal bovine serum
FOXP3	Forkhead box P3
GITR	Glucocorticoid-induced TNF receptor
GLUT-1	Glucose transporter 1
HAM/TSP	HTLV-1 associated myelopathy/tropical spastic paraparesis
HBZ	HTLV-1 bZIP factor
HDAC	Histone deacetylase
HIV	Human immunodeficiency virus

HLA	Human leucocyte antigen
HSPG	Heparan sulphate proteoglycan
HTLV	Human T-lymphotropic virus
HTS	High throughput sequencing
ICAM	Intercellular adhesion molecule
IFN	Interferon
IPA®	Ingenuity® pathway analysis
IPCR	Inverse PCR
IS	Integration/insertion site
KIR	Killer cell immunoglobulin-like receptor
LEDGF	Lens epithelium-derived growth factor
LMPCR	Linker-mediated PCR
LTR	Long terminal repeat
MAD1	Mitotic spindle assembly checkpoint protein
MSP	Methylation specific PCR
MLV	Murine leukemia virus
NFkB	Nuclear factor kappa-light-chain-enhancer of activated B cells
NRES	National research ethics service
NRP	Neuropilin
OCI	Oligoclonality index
ORF	Open reading frame
PBMC	Peripheral blood mononuclear cell
PCAF	P300/CBP-associated factor
PCR	Polymerase chain reaction
PCR ISH	PCR in situ hybridisation
PHA	Phytohaemagglutinin
PVL	Proviral load

RTPCR	Reverse transcription PCR
QPCR	Quantitative PCR
STAT	Signal transducer and activator of transcription
STLV	Simian T-lymphotropic virus
TCR	T-cell receptor
TFBS	Transcription factor binding site
TNFα	Tumour necrosis factor α
TRE	Tax-responsive element
Treg	Regulatory T-cell
TSS	Transcriptional start site
UCSC	University of California, Santa Cruz
UIS	Unique integration site
ZDV	Zidovudine (azidothymidine)

Glossary

- **Sisters** HTLV-1 infected cells which share a common integration site and are assumed to have arisen from a single infection event
- **Clone** The entire population of sisters for a given integration site
- **Clone abundance** The number of sisters within a clone
- **Relative abundance** The proportion of the proviral load occupied by each clone
- **Absolute abundance** The abundance of each clone per 10,000 PBMC
- **Small clones*** Relative abundance <1% PVL
- **Intermediate clones*** Relative abundance ≥1% PVL and <35% PVL
- **Large clones*** Relative abundance ≥ 35% PVL

***Details for abundance bins defined in section 4.3.8

Statement of collaboration

- Chapter 3: High throughput sequencing of naturally infected T-cell clones

Patient recruitment, consent for research and clinical diagnosis by Professor Graham Taylor, National Centre for Human Retrovirology, St Mary's Hospital, Imperial College Healthcare NHS Trust.

The protocol for isolating and expanding infected T-cell clones was established by Dr Aileen Rowan in this laboratory, who isolated many of the T-cell clones analysed here, and supervised me to isolate and expand further clones. The manuscript published in *Blood* based upon this work was co-authored by Cook and Rowan.

- Chapter 4: Characterisation of ATL cohort

Patient recruitment, consent for research and clinical diagnosis by Professor Masao Matsuoka, Kyoto University.

T-cell receptor gene rearrangement studies on 40 samples were undertaken by Mikel Valgannon, in the diagnostic Molecular Pathology Unit, Hammersmith Hospital, Imperial College Healthcare NHS Trust.

- Chapter 5: Genomic landscape of HTLV-1 proviral integration sites

Datasets: The asymptomatic carrier cohort dataset was provided to me by Dr Heather Niederer and the *in vitro* dataset by Dr Anat Melamed. Downstream bioinformatic and statistical analysis of these AC and *in vitro* datasets was undertaken by me. Dr Nirav Malani (FD Bushman group, University of Pennsylvania USA) developed the hiAnnotator R-package and generated *in silico* datasets.

Dr Anat Melamed developed the bioinformatic integration site data extraction pipeline (DEISA) and in conjunction with Daniel Laydon designed the multivariate analysis test.

Chapter 1. Introduction

Human T-Lymphotropic Virus Type-1 (HTLV-1) was first isolated in 1980 from cultured cells taken from a patient diagnosed with a cutaneous malignancy (Poiesz et al., 1980) and is the cause of adult T cell leukaemia/lymphoma (ATL). In 1985, HTLV-1 seropositive patients in French Martinique were diagnosed with a neurodegenerative disease known as tropical spastic paraparesis (TSP) and a similar disorder, HTLV-1 associated myelopathy (HAM) was later reported in Japan (Osame et al., 1986). Since no viral particles have been identified in the plasma HTLV-1 was thought for many years to be a latent virus. Now, it is understood that there is a complex and dynamic interaction between the virus and host immune system. Alterations in the balance of these interactions allow clonal expansion and subsequent transformation to malignant disease. The two broad major questions that remain in HTLV-1 research are how does the virus persist despite a strong, constitutive immune response, and, what factors determine the risk of clinical disease?

1.1. Human T-lymphotropic virus

HTLV-1 is a member of the deltaretrovirus family and so far, there have been four genotypes of human HTLVs identified (Wolfe et al., 2005) and four genotypes described which infect non-human primates (Slattery et al., 1999). There is high sequence identity between HTLV-1 and simian T-lymphotropic virus type 1 (STLV-1) suggesting a relatively recent common ancestor. HTLV-1 is of major clinical interest since it is the only HTLV to be definitely associated with clinical disease. Whilst HTLV-2 may be frequently observed within specific populations, especially amongst indigenous populations and intravenous drug users (Chang et al., 2013) it is not associated with specific clinical disease and is carried at an extremely low proviral burden (Murphy et al., 2004). The more recently described HTLV-3 and -4 have no known disease associations (Mahieux and Gessain, 2007).

There are 3 genotypes of HTLV-1: Melanesian, Central African and Cosmopolitan types, with four subsets of the Cosmopolitan group (Transcontinental, Japanese, Western African and North African). However, there does not appear to be any clear association between the viral genotype and disease which suggests that host specific factors must account for the remarkable variation observed in the clinical manifestations of HTLV-1 infection.

1.2. Epidemiology and Transmission

HTLV-1 has been recently estimated to infect over 10 -15 million people worldwide (Gessain and Cassar, 2012) and is endemic in regions of southern Japan, the Caribbean, sub-Saharan Africa, Brazil and northern Iran (de The and Bomford, 1993). Whilst over 90% of infected individuals remain asymptomatic, 1-2% will develop a debilitating progressive neurological condition known as HTLV-1 associated myelopathy/tropical spastic paraparesis (HAM/TSP) and approximately 5% will develop adult T cell leukaemia/lymphoma (ATL) (Yamaguchi and Watanabe, 2002). Prevalence of HTLV-1 in the Kyushu district of Japan has been reported at over 10% of the general population and the cumulative incidence of developing ATL among adult virus carriers is estimated at 6.6% males and 2.1% females (Arisawa et al., 2000). Epidemiological data in many other endemic regions, such as Sub-Saharan Africa, remain unknown. In Europe and North America HTLV-1 is mainly present among immigrant communities from endemic regions and within the UK it is estimated that there are approximately 22500 infected individuals (Tosswill et al., 2000).

Transmission of HTLV-1 requires cell to cell contact (Yamamoto et al., 1982) and cell-free blood products are non-infectious (Jason et al., 1985). HTLV-1 can be contracted through infected cellular blood product transfusions with a recent look-back by the UK Blood and Transplant Service reporting 29.4% transmission rate prior to leucodepletion of blood products and 3.7% following leucodepletion which was introduced in the UK in 1998 (Hewitt et al., 2013). HTLV-1 can also be

transmitted by sexual contact (Murphy et al., 1989) or by solid organ transplantation (Glowacka et al., 2013). In endemic countries, the major route of transmission is through breast feeding (Uchiyama, 1997; Verdonck et al., 2007), with an increased risk of transmission directly correlated with duration (Ureta-Vidal et al., 1999).

In the UK universal screening of blood donors for HTLV-1 antibody has been in practice since 2002 and diagnosis is based upon repeated reactive enzyme immunoassay confirmed with specific immunoblot (Dow et al., 2001). Seroprevalence among blood donors in England and Wales is estimated at 8 cases/million donations (Hewitt et al., 2013) and 470 cases per million pregnant women (Taylor et al., 2005). However the screening of blood, organ donors and pregnant women is still not undertaken in many regions of the world, including 'wealthy' nations.

1.3. HTLV-1 associated diseases

1.3.1. Adult T-cell leukemia/lymphoma

Adult T cell leukaemia is a neoplasm of mature, post thymic T-lymphocytes with a tumour that typically consists of an oligoclonal or monoclonal outgrowth of $CD4^+$ $CD25^+$ T lymphocytes carrying a complete or defective provirus of HTLV-1. Approximately 4% of cases are CD4-CD8+ and a similar proportion CD4+CD8+ or CD4- CD8- (Kamihira et al., 1992). The cells usually express the markers CD2 and CD5 whilst CD3 and TCRβ are frequently downregulated at the cell surface. The ATL cells frequently express molecules characteristic of activated and regulatory T-cells including CD25, CCR4 and the transcription factor FoxP3, making flow-cytometric cell identification of malignant, from non-malignant HTLV-1 infected cells difficult (Chen et al., 2006; Toulza et al., 2009). The molecular pathogenesis of ATL is discussed in more detail in chapters 4 and 5.

The median age of presentation is 40 years in Afro-Caribbean patients and 50 years in Japan, and the disease typically occurs decades after asymptomatic infection usually in individuals who are believed to have become infected during infancy. In contrast to HAM/TSP, ATL is not associated with recipients of HTLV-1 contaminated blood transfusions, the recipients of such transfusions typically being adults in the 6^{th} and 7^{th} decade.

Clinical manifestation of ATL is heterogeneous and may be characterised by lymphadenopathy, abnormal lymphocytosis, hepatosplenomegaly, skin lesions, pulmonary infiltrates and hypercalcaemia. Diagnosis is based upon the presence of morphologically abnormal lymphocytes ('flower cells' are typical but not necessarily present), presence of antibodies to HTLV-1 in the serum and the demonstration of monoclonal integration of HTLV-1 provirus in the tumour cells.

The disease is classified into 4 subgroups under the Shimoyama classification (Shimoyama, 1991) (classification detailed in Appendix 3); broadly they are considered as either 'aggressive' (acute and lymphoma subtypes) or 'indolent' (chronic and smouldering). The most common presentation is the acute form (60% of cases) and typically presents as an emergency with complications from a rapid leucocytosis and associated cytopenias, hypercalcaemia, bulky lymphadenopathy and infectious complications. In the chronic form, the lymphocytosis may be marked, but the patient may be asymptomatic for a number of years with the major clinical complications being skin lesions, opportunistic infection and the risk of transformation to acute disease. Approximately 20% of individuals present with a pure lymphoma with a normal circulating white cell count and normal peripheral blood film. There remain difficulties using the Shimoyama criteria in the classification of certain subtypes of disease e.g. (1) The acute subtype with bulky lymphadenopathy who behave clinically like the lymphoma subtype (2) The significance of the smouldering subtype, since a high percentage of abnormal lymphocytes may be present in the peripheral blood films of patients with non-malignant infection (Hodson et al., 2013) and (3) The Shimoyama classification excludes the recently recognised purely cutaneous subtype (Amano et al., 2008). However, with these caveats in

mind, it still remains useful for purposes of standardising clinical trials, comparing disease outcomes internationally, selecting therapies and for prognostic information.

Patients with aggressive ATL generally have extremely poor prognosis, mainly due to resistance to steroids and cytotoxic agents combined with significant immune suppression and susceptibility to opportunistic infection. In the aggressive subtypes, the best clinical trial results to date report complete response rates in only 25-40% patients, median progression free survival time of 5 to 7 months and median overall survival time of approximately 13 months (Tsukasaki et al., 2007; Yamada et al., 2001). Whilst the chronic and smouldering forms of ATL have a relatively indolent course they may transform into aggressive ATL and under a 'watchful waiting' strategy the median survival is approximately 5 years (Takasaki et al., 2010). By contrast, in Europe, Martinique and the USA where combination zidovudine and interferon (IFN) -α is the standard of care for chronic ATL, 100% overall survival has been reported and the median survival not reached (Bazarbachi et al., 2010). The risk of relapse with all subtypes remains extremely high and there are many unanswered clinical questions. For example, the role of zidovudine and IFN-α in long term consolidation treatment, the place of immunomodulatory therapies such as proteasome inhibitors, arsenic trioxide or the precise role of new monoclonal antibodies such as mogamulizumab which targets CCR4 expression on the tumour cells. The only truly curative approach remains allogeneic bone marrow transplantation, but this is only a therapeutic option for those individuals who are young-enough and fit-enough, have achieved a response to induction treatment , have a suitable HLA matched donor and in itself is a procedure associated with substantial morbidity and mortality.

1.3.2. HAM/TSP clinical manifestation

The lifetime risk for developing HAM/TSP is 0.25-4% (Kaplan et al., 1990) and is characterised by a gradual symmetric paraparesis of lower limbs with signs of pyramidal tract involvement, which progresses slowly and without remissions. Early in the disease the first symptoms are of weakness of the lower limbs and lumbar pain, although there may be sensory symptoms too. In many patients, urinary and sexual problems can be the first symptoms (De Castro-Costa et al., 2006). Dizziness is common (with normal clinical examination) and suggests disease in the vestibulospinal and motor tracts (Felipe et al., 2008).

The weakness in the lower limbs is associated with moderate to severe spasticity, up-going plantar reflex and hyperreflexia. Vibratory sense is frequently impaired with relatively preserved proprioception. In the upper limbs there is usually hyperreflexia without weakness. As the disease progresses, the weakness and the spasticity increase, and the gait deteriorates (Nagai and Osame, 2003). Neuropathic pain becomes common as the disease advances. Autonomic dysfunction of bladder and bowel are a common cause of morbidity resulting in recurrent urinary tract infections and resultant chronic renal failure.

1.3.3. HAM/TSP disease pathogenesis

The major histopathological changes found at post mortem in patients with HAM/TSP are long tract degeneration and demyelination affecting pyramidal, spino-cerebellar and spinothalamic tracts and hyalinoid thickening of media and adventitia of blood vessels in the brain, spinal cord and subarachnoid space (Akizuki et al., 1987; Cartier et al., 1997).

Immunostaining of post mortem biopsies shows that early in the disease process the leptomeninges, blood vessels and parenchyma are infiltrated with both CD4+ lymphocytes and CD8+ lymphocytes, B- lymphocytes and foamy macrophages whereas later in the disease CD8+ lymphocytes predominate with subsequent progression to a relatively acellular, atrophic pattern with axonal and myelin degeneration. The entire spinal cord can be affected, although the lower thoracic level is predominantly affected (Iwasaki et al., 1992).

Cerebrospinal fluid samples (CSF) may show mild pleomorphic lymphocytosis with mild to moderate increase in protein. Antibodies against HTLV-1 are present in the CSF and, in general, the HTLV-1 proviral load measured in the CSF of individuals with HAM/TSP are typically greater than twice their load in the peripheral blood (Takenouchi et al., 2003), whereas the ratio of CSF to peripheral blood HTLV-1 proviral loads in asymptomatic carriers are typically lower, reflecting either recruitment or expansion of HTLV-1 infected cells in the CNS (Oh and Jacobson, 2008). There is no evidence that HTLV-1 directly infects neuronal cells, astrocytes or microglia. However, HTLV-1 specific CD8+ lymphocytes that secrete the neurotoxic cytokines, IFN-γ and TNF-α are present (Greten et al., 1998) supporting a bystander damage effect (Ijichi et al., 1993). The proposed sequence of events for "bystander damage" may be: Activation of T cells following HTLV-1 infection allows the activation and migration of CD4+ and CD8+ T cells across the blood-brain barrier (from the peripheral blood into the central nervous system). The HTLV-1 specific CD8+ T cells are preferentially recruited or expanded within the CNS, and respond to HTLV-1 antigen expressing cells, most likely HTLV-1 infected CD4+ T cells. The HTLV-1 specific immune response that occurs in the CNS results in the release of neurotoxic cytokines and subsequent CNS damage (Nagai and Jacobson, 2001).

1.3.4. Associated diseases

HTLV-1 has been associated with a wide variety of other inflammatory diseases, such as uveitis (Mochizuki et al., 1992), polymyositis (Morgan et al., 1989), Sjogren's syndrome (Vernant et al., 1987) and bronchiectasis (Einsiedel et al., 2012). Typically these are diagnosed in known HTLV-1 carriers rather than following investigation for an underlying cause of these diseases. HTLV-1 carriers are susceptible to serious outcomes following secondary infection with pathogens such as varicella zoster, *Mycobacterium tuberculosis* (Pedral-Sampaio et al., 1997*)* and *Strongyloides stercoralis* (Marcos et al., 2008).

1.4. Cellular and Molecular biology

1.4.1. HTLV-1 receptor

Whilst HTLV-1 has been shown to infect a wide variety of cell types in vitro (Jones et al., 2008; Koyanagi et al., 1993), the major reservoir of infection in vivo appears to be the CD4+ T-cells and a smaller component in the CD8+ T-cells (Nagai et al., 2001; Richardson et al., 1990). The observation that HTLV-1 can infect several different cell types in vitro is consistent with the relatively recent identification that the widely expressed glucose transporter-1 (GLUT-1) (Manel et al., 2003), neuropilin-1(NRP) (Ghez et al., 2006), and heparan sulphate proteoglycans (HSPG) (Jones et al., 2005) receptors play an important role in HTLV-1 viral entry. The interplay between these molecules and the virus is not fully understood, although it is postulated that sequential binding of HSPGs, NRP1 and GLUT-1 allow a conformational change of viral Env and thus cell entry (Ghez et al., 2010).

1.4.2. Mechanism of cell-to-cell spread

HTLV-1 is highly cell associated (Yamamoto et al., 1982) requiring cell-to-cell contact for transmission. Several mechanisms of cell-to-cell spread of HTLV-1 have been proposed. Igakura (Igakura et al., 2003) first described the virological synapse through which HTLV-1 is transmitted from target cells to donor cells by polarisation of the cytoskeleton towards the uninfected cell, triggered by the viral *tax* gene in concert with stimulation of the intercellular adhesion molecule-1 (ICAM-1) (Nejmeddine et al., 2009). Electron tomography of the virological synapse demonstrates viral particles present in multiple clefts within the synapse, surrounded by tightly bound plasma membranes forming confined intercellular sites (Majorovits et al., 2008), consistent with the observation that cell free virus particles are not found in plasma thus allowing escape from immune surveillance by HTLV-1 specific neutralising antibody (Nejmeddine and Bangham, 2010).

More recently a biofilm-like extracellular structure has been described in which HTLV-1 infected cells covered with recently budded viral particles on the surface are rapidly transferred to the surface of target cells, resulting in new infection (Pais-Correia et al., 2010). In addition, HTLV-1 p8 protein, derived from processing of the $p12^{I}$ protein of the pX region, may induce the formation of cellular conduits among T-cells allowing HTLV-1 transmission (Van Prooyen et al., 2010). It is possible that HTLV-1 may combine multiple strategies to establish efficient viral transmission between cells and it is difficult to quantify the role played by each mechanism in vivo.

1.4.3. Mechanism of HTLV-1 proviral integration

The HTLV-1 viral replication life-cycle, as with other retroviruses, is characterised by two distinctive steps. Firstly, following viral entry into the cells, the viral genome is reverse transcribed in the cytoplasm using reverse transcriptase delivered to the host cell as part of the viral core, which generates a double stranded DNA copy. Secondly, after viral DNA synthesis the HTLV-1 genome is stably associated with the viral integrase, also carried in the viral core, and later transported to the nucleus for integration (Figure 1.1): The viral DNA is first processed by viral integrase at each 3' end whereby two nucleotides are removed resulting in new CA-3' ends ('3'-end processing'). These new ends attack a pair of phosphodiester bonds in the target DNA – in the case of HTLV-1 these lie 6 base pairs apart. The 3' end of the viral DNA joins to the 5'-end of the host genome, and to complete integration, the two unpaired bases left at the 5' ends of the viral genome are removed and the 5'-ends of the provirus can be ligated to host target DNA. Whilst the viral integrase is responsible for 3'-end processing, the remaining steps are thought to be catalysed by host genomic cellular enzymes (Craigie and Bushman, 2012). HTLV-1 viral targeting is not random and there is a preference for proviral integration in proximity to transcriptional units (Gillet et al., 2011; Meekings et al., 2008) and in proximity to specific transcription factor binding sites (TFBS) (Melamed et al., 2013).

The provirus is then replicated along with cellular DNA during cycles of cell division, using the highly-faithful host DNA polymerase. The provirus also serves as a template for transcription of viral RNAs – some of which are translated to yield viral proteins, whilst a proportion of full-length viral RNA serves as genomic RNA in progeny virions.

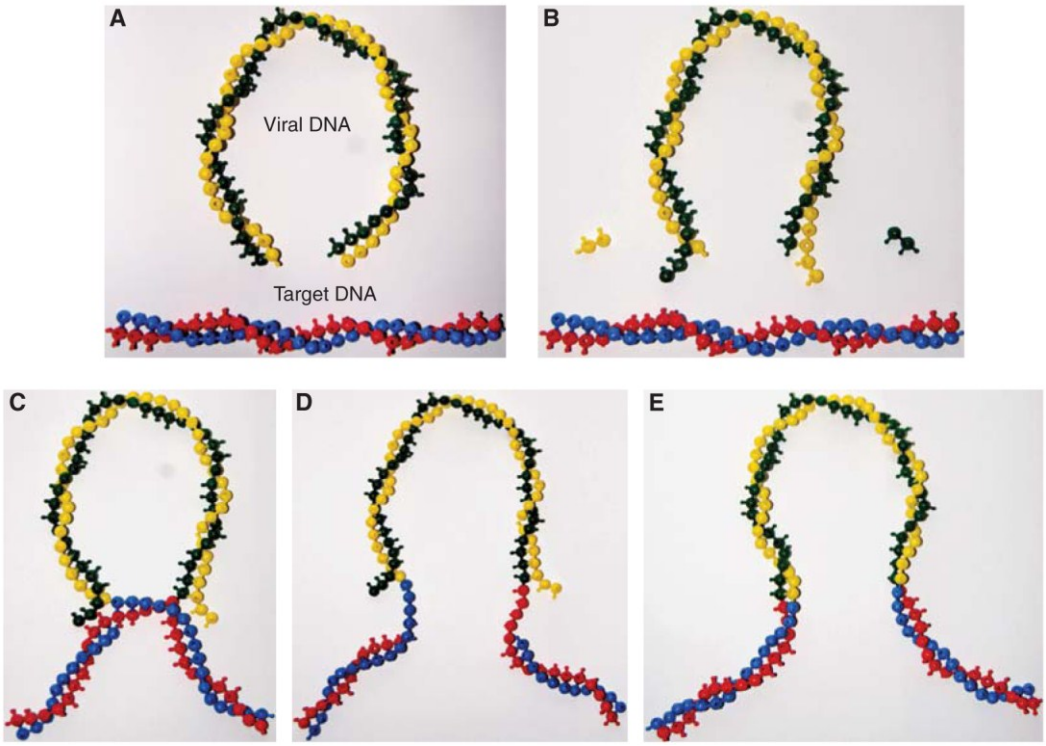


Figure 1.1: Retroviral integration (DNA breaking and joining reactions)

Figure adapted from Craigie and Bushman, © 2012, Cold Spring Harbour Lab Press (Cold Spring Harb Perspect Med 2012;2:a006890). DNA molecules shown by coloured balls. (A) Linear blunt-ended viral DNA (green and yellow) and the target host DNA (blue and red). (B) 3'end processing: Two nucleotides are removed from each of the 3' ends of the viral DNA. (C) The 3'ends of the viral DNA attack a pair of phosphodiester bonds in the target DNA at the integration site (6 base pairs apart in the case of HTLV-1). The 5'ends of the viral DNA are not joined.(D) Completion of integration requires removal of the two unpaired bases at the 5'end of the viral DNA and ligation of the 5'end of the viral DNA to the host genome.(E) Integrated provirus.

1.4.4. Mapping and quantification of proviral integration sites

There has been a longstanding interest in accurate mapping and quantification of integration sites in HTLV-1 and other retroviral infections e.g. MLV (Wu et al., 2003) and HIV (Schroder et al., 2002). The aim of this type of analysis is to answer questions such as whether a single locus of integration was responsible for adult T-cell leukaemia/lymphoma (Seiki et al., 1984); to understand whether integration is truly random or targeted and to determine if there are 'safe harbours' of the host genome which could be identified as potential targets for therapeutic gene therapy.

Detection of a single HTLV-1 integration site by Southern blot is characteristic of ATL (Yoshida et al., 1982) and monoclonal integration of HTLV-1 proviruses into tumour cells remains part of the diagnostic criteria for ATL (Tsukasaki et al., 2009). However, Southern blot lacks the sensitivity to detect minor populations of polyclonal HTLV-1 infected cells in both malignant and non-malignant infection. In the 1990s and 2000s, inverse PCR (IPCR, Takemoto et al., 1994) and linker mediated PCR (LMPCR, Derse et al., 2007; Meekings et al., 2008; Wattel et al., 1995) followed by cloning, transformation and Sanger sequencing were used to map integration sites in patients with malignant and non-malignant HTLV-1 infection. However, these PCR based techniques have two major limitations. First, the use of restriction enzymes leads to preferential detection of proviruses that lie near to the restriction site (30-60 base pairs for the restriction enzyme *MseI* (Wang et al., 2008). Second, PCR amplification preferentially amplifies short DNA fragments. These classical PCR techniques cause a systematic bias in the accurate quantification of infected CD4+ T-cell clones, recover relatively few integration sites per experiment, and cannot accurately estimate the abundance of each unique integration site within a host.

The observation of monoclonal integration of HTLV-1 provirus into ATL tumour cells, combined with the widespread use of Southern blot or IPCR /LMPCR integration site analysis has led to the widely held assumption that there is a single copy of HTLV-1 integrated into each infected CD4+ T-cell. To

overcome these technical limitations and for further hypothesis testing a customised high throughput sequencing method was developed in this laboratory to accurately map and calculate clonal abundance of each infected T-cell clone (Gillet et al., 2011)

1.4.5. HTLV-1 viral gene expression

The HTLV-1 proviral genome is nearly 9 Kb long and, similar to other retrovirus, the viral genome is flanked by two identical long terminal repeats (LTR). The virus encodes multiple proteins by using both sense and antisense RNA strands and accessing several open reading frames (ORFs). The provirus contains genes for structural proteins Gag and Env and enzymatic proteins (reverse transcriptase, integrase and protease). There is an additional pX region at the 3' end (Figure 1.2). The pX region, located between Env and the 3'LTR, contains genes for non-structural viral accessory proteins (p12, p13, p30 and p21) and regulatory genes *tax* and *rex*. In 2002 Gaudray and colleagues (Gaudray et al., 2002) identified an ORF on the complementary RNA strand of the pX region whose protein product comprises a C-terminal basic leucine zipper (bZIP) and therefore named HTLV-1 bZIP (HBZ). Whilst the viral proteins transcribed from the sense strand of the provirus are under the control of a promoter located at the 5'LTR, HBZ is controlled by a promoter within the 3'LTR and is therefore transcribed in an antisense direction. HTLV-1 proviral genes are differentially expressed by alternate splicing and may be unspliced, singly spiced or doubly spliced.

Structural genes

Gag, Pro and Pol are translated from a primary full length mRNA transcript, and ribosomal frameshifts result in the production of Gag-Pro and Gag-Pro-Pol fusion proteins, which are then subject to proteolytic cleavage (Nam et al., 1993). Gag encodes the structural proteins of the matrix, capsid and nucleocapsid forming the structural 'core' of the HTLV-1 virion. Env protein is transcribed as a singly spliced transcript and enhances viral entry and infectivity by mediating receptor binding and membrane fusion of virus particle with the cellular membrane (Delamarre et al., 1996).

The pX region

The *tax* gene encodes the viral transactivating protein Tax, the best studied of all the HTLV-1 proteins. Tax protein enhances viral gene expression through interactions with the Tax responsive elements (TRE) of the 5'LTR which consist of three repetitive 21 base pair sequences containing a core nucleotide sequence (TGACG) (Jeang et al., 1988), homologous to the cAMP response element binding factor (CREB). Tax does not bind to DNA directly, but the formation of Tax-CREB promoter complex serves as a high-affinity binding site for the recruitment of the cellular co-activators CBP, p300 and PCAF and Tax has been shown in vitro to transactivate many viral and host genes (Kashanchi and Brady, 2005) (Grassmann et al., 2005) (Matsuoka and Jeang, 2007). Host genes dysregulated by HTLV-1 Tax include genes involved in cell-cycle, apoptosis, cytokines and DNA repair. Several hundred Tax-binding partners have been reported in vitro and it has been demonstrated that Tax is subject to post-translational modifications including sumoylation, poly-ubiquination, phosphorylation and acetylation, each of which alters its binding affinities and functions (Shembade and Harhaj, 2010), but it is likely that there are many in vitro artefacts of reported Tax-protein-complexes that may never occur in vivo. The strongest evidence for the oncogenic potential of Tax has been demonstrated by its ability to immortalise cell lines in vitro (Grassmann et al., 2005) and to promote the growth of tumours in mice in vivo (Pozzatti et al.,

1990). Of note, HTLV-2 is not associated with aggressive leukaemia/lymphomas, but the HTLV-2 associated Tax2 protein is also capable of transforming rat fibroblasts in vitro, albeit less efficiently than Tax1 (Endo et al., 2002).

The regulatory protein Rex regulates proviral gene expression on a post transcriptional level and regulates the nuclear export of doubly spliced versus unspliced transcripts of HTLV-1. The Rex responsive element is a short sequence present in the U3-R region of the LTR and recruits Rex to the transcript via a stem loop structure (Seiki et al., 1988). Rex mediated regulation of RNA transport has been implicated in determining the kinetics of viral gene expression where initially Tax/Rex are expressed followed by other HTLV-1 transcripts (Rende et al., 2011).

The accessory protein p12 is expressed from ORFI of the pX and localises in the endoplasmic reticulum. It has been described to decrease surface expression of MHC-I (Johnson et al., 2001) and activates various pathways such as STAT5, resulting in reduced dependency on IL-2 for cell proliferation.

HTLV-1 p13 and p30 accessory proteins are expressed from ORF II of the pX region and share part of their sequence. Despite this their localisation and function is different: HTLV-1 p13 can be detected in the inner membrane of mitochondria and modulates mitochondrial morphology and metabolism (Ciminale et al., 1999) and it has been suggested that p13 has a role in apoptosis (D'Agostino et al., 2005; Hiraragi et al., 2005). HTLV-1 p30 regulates gene expression by sequestering Tax/Rex mRNA in the nucleus and by binding the Rex responsive element and it has been suggested that its function is dependent upon its concentration - at high concentrations p30 prevents viral replication by binding to Tax-Rex mRNAs and retaining them in the nucleus (Nicot et al., 2004) whilst at lower concentrations it functions as a transcription factor by modulating CREB-responsive promoters (Zhang et al., 2000).

HBZ is constitutively transcribed in all HTLV-1 infected cell lines and in primary ATL cells (Satou et al., 2006). Whilst experiments with *HBZ* mutants lacking functional domains have demonstrated that HBZ is dispensable for in vitro immortalisation of cell lines, cells lacking *HBZ* have a lower proliferation capacity (Arnold et al., 2006). Furthermore proliferation of infected T-cell lines could be inhibited by small interfering RNAs that suppress HBZ expression (Satou et al., 2006) and transgenic expression of HBZ in murine CD4+T-cells induces T-cell lymphomas (Satou et al., 2011). HBZ protein interacts with several transcription factors via its bZIP domain including JunB, Jun D and AP1 and, depending upon its binding partner, variably activates or inhibits host cellular gene transcription (Matsuoka and Jeang, 2011).

HBZ protein is difficult to detect in *ex vivo* samples with current laboratory techniques, even in the presence of high levels of HBZ mRNA (Suemori et al., 2009). Curiously, HBZ protein can be readily detected by western blot lysates from cells transfected with HBZ expression vectors which raises the question as to whether lack of detection of HBZ protein in patient derived *ex vivo* samples is indicative of a translation block *in vivo*. Site directed mutagenesis altering the ATG start codon of HBZ, reveals that the effects of HBZ on cellular proliferation is exerted at mRNA level (Satou et al., 2006). In addition, cells infected with an HTLV-1 molecular clone retain a large proportion of HBZ in the nucleus in a Rex-independent manner (Rende et al., 2011) suggesting that at least some of its function is exerted at nuclear RNA level rather than protein.

The strongest evidence for the expression of HBZ protein in vivo is the recent identification of an anti-HBZ immune response: Approximately 10% of HTLV-1 infected individuals have anti-HBZ antibody responses (Enose-Akahata et al., 2013) and HBZ specific cytotoxic T-lymphocytes are found in ~ 30% patients (Hilburn et al., 2011; Macnamara et al., 2010). Thus the difficulty in detecting HBZ protein in samples derived *ex vivo* suggests that either the HBZ protein levels are below the current limits of laboratory detection and/or that the translation of HBZ is tightly regulated by the provirus.

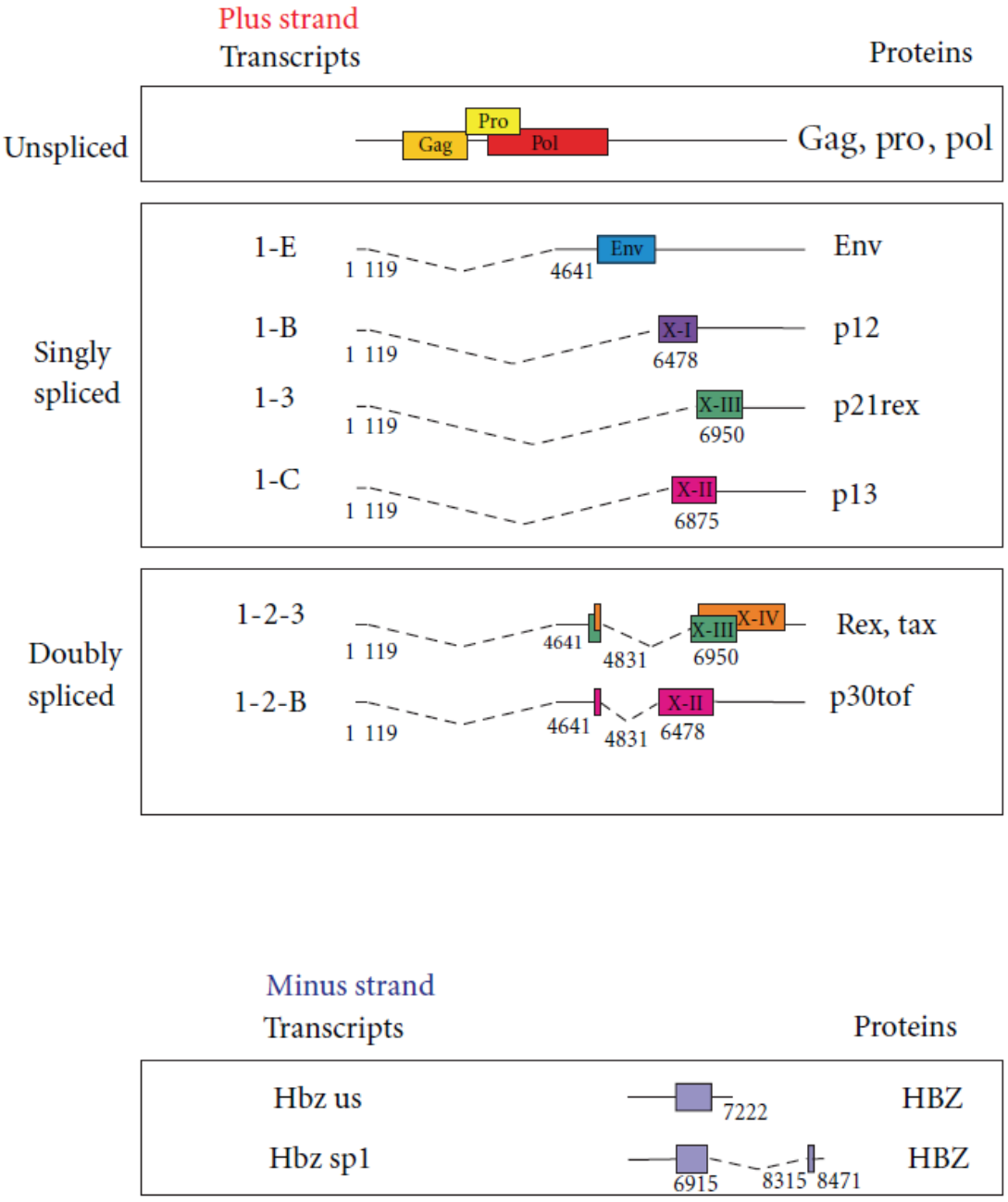


Figure 1.2: The HTLV-1 proviral genome

Figure adapted from Rende et al, Leukemia Research and Treatment, 2011. The organisation, alternate splicing and coding potential of HTLV-1 mRNAs is shown. ORFs are indicated by coloured boxes and the numbers correspond to nucleotide positions of HTLV-1 RNA sequence (Koralnik et al., 1992).

1.4.6. Mechanism of viral proliferation (mitotic versus infectious spread)

To persist within the host CD4+ T-cell compartment the virus pursues two strategies. Firstly via *de novo* infectious spread (as described) and secondly via mitotic proliferation of infected cells. During the chronic phase of infection mitotic proliferation appears to be dominant.

There is strong evidence to support the conclusion that mitotic proliferation is the dominant force in the maintenance of a steady state proviral load (PVL) of infected peripheral blood mononuclear cells (PBMC). Firstly, longitudinal studies of integration site identity demonstrated the existence of long-lived expanded clones, suggesting that each clone of cells originated from a common progenitor cell (Cavrois et al., 1996; Cavrois et al., 1995; Furukawa et al., 1992; Gillet et al., 2011). Secondly, there is remarkably low sequence diversity compared with other retroviruses (Overbaugh and Bangham, 2001) which is consistent with replication via the highly fidelity, proof reading host DNA polymerase rather than the error prone retroviral reverse transcriptase. In line with this are the results of a randomised, double-blind, placebo controlled study of six months' combination therapy with nucleoside analogues (zidovudine and lamivudine) in HAM/TSP: Inhibition of reverse transcriptase did not demonstrate a reduction in proviral load (Taylor et al., 2006). However, the relative contributions of infectious and mitotic spread remain unknown.

1.5. Immune determinants and response to HTLV-1

Life-long infection with HTLV-1 is frequently accompanied by a strong and constitutive immune response. It remains uncertain how the virus regulates the balance between proliferation and viral gene expression whilst escaping immune control.

The major predictor of disease is the proviral load: The proviral load within an individual remains relatively constant over years albeit with a slow steady rise over decades. In a large cohort study in Japan, the median PVL of those in the 40-60 year old age group was significantly greater than those younger than 40 years (Sasaki et al., 2010). However, the variation in proviral loads between different patients is great. Individuals may have a PVL that ranges between less than 0.001% PBMC to over 100% PBMC, with the risk of clinical disease rising in carriers with a PVL above 4% in Japan (Iwanaga et al., 2010) and a PVL > 10% in the UK (Demontis et al., 2013). The median proviral load in Japanese males has been reported as 2.1% and in females 1.4% (Sasaki et al., 2010), consistent with the observation that for unknown reasons more males develop ATL. Nonetheless, there is overlap in the range of proviral load seen between patients with disease and those that remain life-long asymptomatic carriers, making per-patient risk analysis extremely difficult.

The observation that asymptomatic family members of those with HAM/TSP or ATL carry higher proviral loads than those without a family history supports the hypothesis that there is a hereditary genetic determinant that predisposes to disease (Iwanaga et al., 2010): Class I HLA genotype plays a significant part in determining the risk of HAM/TSP and maintenance of the proviral load in individuals from the Kagoshima region of Japan (an HTLV-1 endemic population). The results showed a dominant protective effect of HLA-A*02 and HLA-Cw*08 in reducing the risk of HAM/TSP and in control of proviral load in asymptomatic carriers, whilst the presence of HLA-B*5401 or class II HLA-DRB1*0101 was associated with an increased risk of disease (Jeffery et al., 2000; Jeffery et al.,

1999). In the same cohort of individuals, it has been shown that the ability to present peptides from HBZ to the cytotoxic T-lymphocytes (CTLs), as determined by an individual's HLA genotype is associated with a low proviral load and reduced incidence of HAM/TSP (Macnamara et al., 2010). Recent work on same cohort has further shown that the protective effect of HLA- Cw*08 and detrimental effect of HLA- B54 was limited to those with a specific inhibitory molecule KIR2DL2 (Seich Al Basatena et al., 2011).

Analysis of non-HLA host genetic factors has revealed that specific polymorphisms also affect the risk of developing HAM/TSP e.g. the TNF-α promoter-863 A allele (Vine et al., 2002) promotes the risk of disease, whilst IL-10-592A (Sabouri et al., 2004), stromal-derived factor-1 +801A, and IL-15 +191C conferred protection against HAM/TSP (Vine et al., 2002), but the contributions of these non-HLA genes to the pathogenesis are largely unknown and requires validation in other ethnic populations in association with broader genome-wide studies.

HTLV-1 can elicit a strong antibody response and often include IgMs in both ACs and HAM/TSP suggesting persistent expression of HTLV-1 proteins (Nagasato et al., 1991). Antibodies to Gag, Env, Tax, Rex and HBZ have all been identified and the sera of the majority of patients contains antibodies directed against Tax (Souza et al., 2011). However, since HTLV-1 is highly cell-associated, it is not clear what role neutralising antibody may play in controlling infection.

There are several lines of evidence that HTLV-1 specific cytotoxic T-lymphocyte (CTL) responses have a significant impact upon infection: (1) The described observation of an association between specific HLA Class I alleles and protection from disease; (2) The existence of escape mutations in known CTL epitopes (Niewiesk et al., 1995); (3) The observation that CD8+ T-cells efficiently kill autologous Tax expressing lymphocytes in fresh PBMC from infected individuals (Hanon et al., 2000) and (4) The more recent identification of HBZ-specific CD4+ and CD8+ cells in vivo and a significant

association between the HBZ-specific CD8+ T-cell response and asymptomatic infection (Hilburn et al., 2011).

In recent years, more attention has been focused upon CTL quality rather than CTL frequency (Bangham, 2009) and that functional avidity, a metric of the ability of CTL to detect limiting quantities of antigen, is a better index (Kattan et al., 2009). High CTL avidity correlates with a lower proviral load and proviral gene expression which suggests that efficient control of HTLV-1 in vivo depends on the quality of the CTL response (Kattan et al., 2009) – although it is difficult to separate cause and effect when considering the host-viral dynamics in chronic phase.

The CTL response may be repressed by regulatory T-cells (Tregs): Expansion of a CD4+FoxP3+ subset is observed in HTLV-1 infection which could be explained by Tax-induced up regulation of CCL22 (Toulza et al., 2010), the natural ligand for C-C chemokine receptor type-4 (CCR4) on FoxP3+ cells, and there is an inverse correlation between CTL lysis efficiency and the % CD4+ FoxP3+Tax- CTLs (Toulza et al., 2008). These observations may, in part, contribute to the profound immune suppression observed clinically in HTLV-1 infection.

1.5.1. Immune response in ATL

Little is known about the role of HTLV-1 specific CTLs in the prevention of ATL, during transformation and following a therapeutic response to treatment. There have been case reports of ATL following solid organ transplantation in the context of immune suppressive therapies (Hoshida et al., 2001; Suzuki et al., 2006).

Despite the predominance of CD4+ malignant ATL cells, the absolute frequency of CD8+ T-cells remains approximately within the normal reference ranges (Arnulf et al., 2004). There are little data describing the functional properties of CTLs in ATL, but it is acknowledged that the *ex vivo* CTL response in ATL is weak (Shimizu et al., 2009). By comparison with ACs, the Tax-specific CTLs are directed at a narrower range of epitopes and are present at low or even absent frequencies with no detectable Env-specific CTLs in ATL subjects (Kozako et al., 2006). A subset of ATL patients have no detectable CTL when PBMCs are measured by functional assays, but after culture functional CTL can be occasionally observed (Arnulf et al., 2004; Kannagi et al., 1984).

Although the observed frequency of CTL response to HTLV-1 in ATL is reduced or absent, there is evidence for an efficient response in a subset of individuals: Furukawa observed amino acid change within the Tax 11-19 epitope which rendered Tax unrecognizable to CTLs and premature stop codons in the *tax* gene which prevent Tax transactivation (Furukawa et al, 2001). Loss of Tax expression or activity has the downstream effect of reduced expression of other viral genes that require Tax transactivation (*gag, pol* and *env)*. Therefore, it is implied that the loss of a single viral protein allows the infected cell to escape surveillance by CTL specific for other viral proteins.

1.6. Animal models of ATL

Since the majority of HTLV-1 infections occur during infancy via breast feeding, it is neither practical nor ethically possible to investigate early models of HTLV-1 infection in naturally infected human primate hosts. Therefore, animal models provide an excellent tool for understanding the biology of HTLV-1 driven ATL and particularly for the development of future vaccines and novel treatments. HTLV-1 animal models vary from those that are naturally infected hosts of similar viruses to engineered small animal models.

Bovine leukaemia virus (BLV) causes an aggressive B-cell lymphoma in <5% infected cattle aged over 5 years, and is a major economic problem in cattle-trading since infection may be spread through milk transmission. BLV is an appealing model for ATL as large animals tend to provide a more relevant model of human cancer development with closer physiology. The most widely used animal model of BLV is in sheep (Djilali and Parodi, 1989; Gillet et al., 2013) and provides some similarities with ATL models since the retrovirus contains both Tax and Rex viral genes, although no equivalent of the antisense *HBZ* gene has been identified (reviewed by (Hajj et al., 2012). Whilst interesting parallels may be drawn between the BLV model of B-cell lymphoma and ATL, these remain distinct viruses which infect different cell types and cause disease in hosts with differing immune responses.

'Old world' monkeys are frequently naturally infected with STLV-1, which has a nucleotide sequence closely related to HTLV-1 (Watanabe et al., 1985), in which clonal proliferation has also been identified by inverse PCR techniques (Gabet et al., 2003) and lymphoproliferative disorders have been observed (McCarthy et al., 1990; Tsujimoto et al., 1987; Voevodin et al., 1996). However, for logistic reasons, detailed characterisation of STLV-1 has not been achieved and there has been no analysis on the function of accessory or regulatory proteins. Recently Miura et al, identified that STLV-1 contains an STLV-1 bZIP factor (SBZ) on the antisense strand similar to *HBZ* and have reported

a high throughput sequencing technique to map STLV-1 integration sites which has shown asymptomatic STLV-1 infected macaques have proliferative features in common with those of HTLV-1 asymptomatic carriers, suggesting that these animals may serve as suitable models for analysis of HTL V-1 carriers (Miura et al., 2013).

Transgenic mouse models expressing Tax (Hasegawa et al., 2006) or HBZ (Satou et al., 2011) in T-cells drive tumour formation in vivo, but whether Tax or HBZ expression alone is sufficient for human T cell leukaemogenesis remains unclear since it is known that mice primary cells are substantially more easily transformed than human primary cells (Hahn et al., 1999). More recent mouse models of HTLV-1 include humanised mice in which human haemopoietic stem cells reconstitute a variety of immune deficient mice by infecting the stem cells in vitro prior to engraftment (Banerjee et al., 2010) or following intraperitoneal inoculation with irradiated HTLV-1 producing cells (Villaudy et al., 2011). Although these mice models allow for the study of early infection events and T-cell tumours, a crucial difference is that they lack a functional immune response.

1.7. Aims and Hypothesis

The central aim of this work was to test the hypothesis that the site of proviral integration determines the risk of developing ATL. To achieve this aim, I used a customised high-throughput sequencing technique previously developed in this laboratory to test the following hypotheses:-

1. There is a single copy of HTLV-1 integrated into each host genome
2. The genomic environment flanking the proviral integration site is associated with malignant transformation of HTLV-1 infected clones.

Chapter 2. Materials and Methods

2.1. Primary cells and cell lines

2.1.1. Patients samples

DNA samples from AC and ATL patient samples were provided by collaborators at the Institute for Viral Research, Kyoto University, Japan and were obtained with written consent in accordance with regulations defined by the Japanese Government and Kyoto University.

T-cell clones were derived from blood samples donated by patients attending the HTLV-1 clinic at the National Centre for Human Retrovirology, Imperial College Healthcare NHS Trust, with written informed consent in accordance with the UK National Research Ethics Service (NRES reference 09/H0606/106).

PBMC were immediately isolated from peripheral blood and layered over Histopaque-1077 (Sigma-Aldrich), centrifuged at 1700 rpm for 25 minutes with a slow brake setting, washed twice with PBS and cryopreserved in fetal bovine serum (Gibco-Life Technologies) containing 10% dimethylsulfoxide (Sigma-Aldrich) and stored in liquid nitrogen.

2.1.2. Cell lines

Jurkat E6.1 (JKT) was used as an HTLV-1 negative T cell line (Schneider et al., 1977) to identify potential contamination during LMPCR and high throughput sequencing (HTS). Tarl2 is a rat lymphoid cell line, containing one integrated copy HTLV-1 provirus per cell (Tateno et al., 1984) and was used for quantification of proviral loads. All cell lines were cultured in either containment level

2 (Jurkat) or containment level 3 (Tarl2) in complete medium: RPMI-1640 medium (Sigma-Aldrich) supplemented with 1% L-glutamine, 1% penicillin/streptomycin and 10% heat-inactivated FBS (Gibco).

Methylated and unmethylated DNA from ATL cell lines 43-T (methylated) and 48-T (unmethylated) for methylation-specific PCRs were provided by Dr Yorfumi Satou, Kumamoto University, Japan.

2.2. Isolation of CD4+ T-cell clones by limiting dilution

CD4+25+ T-cell clones were isolated from 10 patients (7 female, 3 male) with different clinical manifestations of HTLV-1: Two ACs (median PVL 6.2%), 5 HAM/TSP (median PVL 13.6%), 2 ATL (median PVL 57.5%) and 1 polymyositis (PVL 18.3%).

CD4+25+ cells were isolated by magnetic activated cell sorting ($CD4^+CD25^+$ isolation kit, Miltenyi Biotec). Cells were subsequently cloned by limiting dilution in RPMI containing 10% human AB serum (Invitrogen) in the presence of 50 IU/mL IL-2 (Promocell), 1 µg/mL PHA (Sigma-Aldrich), 10µM raltegravir (Selleck Chemicals), and 0.5×10^6/mL γ-irradiated feeder cells (mixed PBMCs from 3 uninfected donors). Clones were expanded with feeder cells and PHA every 14 days and fed with IL-2 and raltegravir twice weekly. Clones were cultured for 4-6 weeks before genomic DNA extraction (DNeasy Blood and Tissue Kit; QIAGEN, 69504.

2.3. Molecular biology methods

2.3.1. Proviral load measurements

The proviral load was measured using quantitative polymerase chain reaction (QPCR) using ABI7900HTFast (Applied Biosystems). PCR was carried out using ABI Fast SYBR green mastermix (Applied Biosystems) according to manufacturer's instructions. Proviral copies were quantified using *tax* specific primers (SK43, SK44)(Kwok et al., 1988) normalised to the number of β-actin copies (Actin-Fw, Actin-Rev). Primer sequences are shown in table 2.1.

Cycling conditions for *actin* and *tax* PCR consisted of: 95^{o}C for 20 seconds, 40 cycles at 95^{o}C for 1 second, 60^{o}C for 20 seconds.

Standard curves were generated using serial dilutions of Tarl2 DNA at six different concentrations (5ng/µl to 20 pg/µl). DNA samples from patients were measured at three different dilutions (5ng/µl to 0.56ng/µl). Proviral load was calculated as a ratio of *tax* to *β-actin* copies on the assumption of one copy of *tax* and two copies of *actin* per cell.

2.3.2. 5' long terminal repeat sequencing

Since the HTLV-1 LTR sequence varies between isolates, the LTR was sequenced prior to integration site library preparation to identify primer binding site polymorphisms before high throughput sequencing. The 3' and 5' LTR sequences are assumed to be identical (Seiki et al., 1983).

5'LTR sequencing was sequenced using a forward primer (5LTR-FW) in the LTR region and the reverse primer (5LTR-Rev) located in the viral *gag* gene sequence downstream of the LTR (primer

sequences are shown in table 2.1). Thermal protocol: 98°C for 3 minutes, 35 cycles 98°C for 10 seconds, 64°C for 20 seconds and 72°C for 20 seconds followed by 72°C for 10 minutes.

The amplified PCR product was sequenced using Sanger sequencing (MRC Clinical Sciences Centre, Core Genomic Laboratory).

2.3.3. Long range PCR to identify defective proviruses

Type 1 or type 2 defective proviruses are commonly identified in ATL samples and are unable to express Tax due to loss of promoter situated in either the 5LTR and/or *pol* regions. First an internal control region of the 3'end of the HTLV-1 genome is amplified for each case, followed by a long-range PCR to identify defective proviruses based upon the length of the long-range PCR product between 5'LTR and pX region (Figure 2.1). A product length of 6.5Kb defines a complete provirus, a product less than 6.5Kb as a type 1 provirus and failure to amplify any product, a type 2 defective (Tamiya et al., 1996).

DNA was amplified using KOD Hot Start DNA polymerase TB341 (Toyobo, Novagen). The primers for the control PCR (Control- Fw, Control- Rev) and long range PCR (LR- Fw, LR- Rev) are listed in table 2.1. Cycling conditions for the internal control PCR consisted of: 95°C for 2 minutes, 30 cycles: 95°C for 20 seconds, 59°C for 10 seconds and 70 °C for 48 seconds, followed by 70°C for 5 minutes. Cycling conditions for the long-range PCR consisted of: 95°C for 2 minutes, 30 cycles: 95°C for 20 seconds, 66°C for 175 seconds, followed by 72°C for 15 minutes.

The PCR products were electrophoresed on a 1% agarose gel with expected product size of 2.85Kb for the control PCR and 6.5Kb for a complete long-range PCR product.

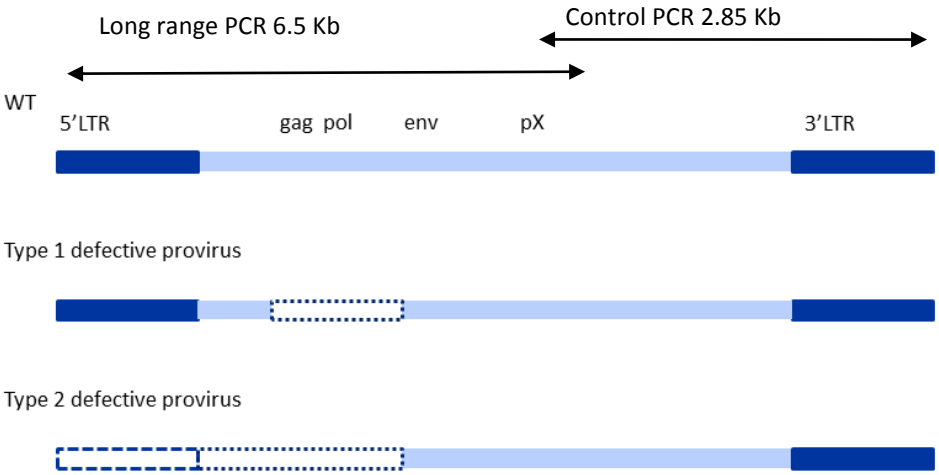


Figure 2.1: Schematic representation of wild type (WT) and defective proviruses

The complete (WT) structure is depicted in the top panel showing the presence of two LTRs and the internal gag, pol, env and pX regions of the provirus. A type 1 defective provirus is characterised by a deletion of the internal gag/pol regions (indicated by dashed box) and a type 2 defective provirus is characterised by absence of both the 5'LTR and variable lengths of the internal gag/pol regions. These subtypes of proviruses can be distinguished by PCR followed by examination of the length of the PCR product by agarose gel electrophoresis: the type 1 product shows a short band and a type 2 product shows absence of long-range PCR in the presence of a positive control PCR.

2.3.4. Identification of exon 2 and exon 3 tax gene mutations

Tax protein is 353 amino acids in length. Exon 2 provides just the start codon (methionine) whilst the remaining amino acids are transcribed and translated from exon 3. The reference sequence published by Fan et al was selected (GenBank BAH85788.1) since it was published on amino acid Tax sequences in a cohort of patients from a similar region of Japan and in an era of high-fidelity polymerases (Fan et al., 2010).

Exon 2

PCR products from complete proviruses identified by long-range PCR were purified using DNA PCR purification kit as per manufacturers protocol (QIAGEN). Exon 2 of *tax* was amplified using Phusion high fidelity DNA polymerase (NEB), inspected on 2% agarose gel for product length (318 base pairs) and sequenced by Sanger sequencing as before. Primers (Exon2-Fw, Exon2-Rev) listed in table 2.1. Cycling conditions as follows: 98 °C for 30 seconds, 20 cycles: 98°C for 5 seconds, 51.5°C for 20 seconds, 72°C for 10 seconds, followed by 72°C for 5 minutes.

Exon 3

The PCR products of the control long-range PCR (methods 2.3.4) were purified using DNA PCR purification kit as per manufacturers protocol (QIAGEN). Exon 3 of *tax* was amplified using Phusion high fidelity DNA polymerase (NEB), cycling conditions as follows: 98 °C for 30 seconds, 20 cycles: 98°C for 5 seconds, 51.5°C for 20 seconds, 72°C for 10 seconds, followed by 72°C for 5 minutes.

PCR products were inspected on 2% agarose gel for product length (1120 base pairs) and sequenced by Sanger sequencing as before using 6 different sequencing primers to capture the entire exon. Primers (Exon3-Fw, Exon3-Rev and sequencing primers I-VI) listed in table 2.1.

2.3.5. Identification of hypermethylated 5'LTR

Since hypermethylation of the 5'LTR of the provirus has been described in 10-15% ATL cases and has been shown to silence *tax* gene transcript expression, methylation specific PCR (MS-PCR) was undertaken as described by Takeda et al (Takeda et al., 2004). However, MS-PCR is an imperfect method to characterise precise methylation events because it can only determine methylation events of the CpGs of the primer binding sites only (and implies methylation of CpGs in the amplicon) but since the critical Tax response element-1 (TRE-1) lies proximal to the start of the provirus within a large CpG island, it is not possible to design good MSP primers in this region. By contrast, bisulfite sequencing PCR (BSP) is a method in which custom specific primers are designed with no bias towards methylated or unmethylated sequences and can be designed across the 5' proviral integration site (between the host genome and U3 region of 5'LTR, incorporating the critical TRE-1). The PCR products can then be subcloned into plasmid DNA and colonies sequenced. However, Takeda et al have demonstrated that MSP between the R-region of the 5'LTR and *gag* correlates extremely well with BSP sequencing results across the integration site and can be used for large numbers of samples (Takeda et al., 2004).

Methylation-specific PCR (MS-PCR)

DNA was treated with sodium bisulfite (Sigma) which converts unmethylated cytosine residues to uracil whilst methylated cytosine residues remain unchanged during the treatment. Bisulfite treated DNA was purified using Zymo EZ Bisulfite DNA clean-up as per manufacturer protocol.

Once converted, the methylation profile of the DNA could be determined by hemi-nested PCR amplification using specific primers for either methylated or unmethylated DNA. ATL control cell

lines T-43 (methylated) and T-48 (unmethylated) were used as both positive and negative controls for each PCR reaction. For the first PCR reaction, forward primer in the U3 region of the 5'LTR and reverse primer in gag and for the hemi-nested PCR, forward primer in the R region of the 5'LTR and the same reverse primer in gag (primer sequences listed in table 2.1). MS-PCR primers did not amplify unconverted HTLV-1 or host genomic DNA.

DNA was amplified using a non-proof reading enzyme, JumpStart RedTaq polymerase (Sigma). Cycling conditions for PCR1 consisted of: 94°C for 2 minutes, 35 cycles: 94 °C for 30 seconds, 53°C for 30 seconds and 72 °C for 2 minutes. Cycling conditions for PCR 2 consisted of: 94°C for 2 minutes, 30 cycles: 94 °C for 30 seconds, 57°C for 30 seconds, 72 °C for 2 minutes, followed by 72 °C for 5 minutes. The PCR product was inspected on a 2% agarose gel for length (428 base pairs).

Bisulfite Sequencing PCR (BSP)

BSP was undertaken with customised primers on 4 specific cases of ATL as proof of concept, to demonstrate that MSP correlates with BSP. DNA was bisulfite converted as before and amplified with custom specific primers from the host genome (upstream of the 5' LTR) to *gag* followed by a hemi-nested PCR in the U3 region, incorporating all three TREs. Primers are listed in table 2.2 (custom primers-forward, gag-rev and heminested- U3 rev). Cases were selected for proof of concept on the basis of similar melting temperatures. DNA was amplified using a non-proof reading enzyme, JumpStart RedTaq polymerase (Sigma). Thermal conditions for PCR 1 and PCR 2 as follows: 94°C for 2 minutes, 35 cycles: 94 °C for 30 seconds, 52°C for 30 seconds and 72 °C for 2 minutes.

PCR products were cloned and transformed into chemically competent *E.coli* TOPO TA cloning® kit for sequencing (Invitrogen). Twenty colonies were picked per plate and plasmid DNA was purified using QIAprep® Spin Miniprep kit (QIAGEN). PCR products were sequenced by Sanger sequencing as before.

2.3.6. T-cell receptor gene rearrangement studies

T-cell receptor γ gene rearrangement studies were undertaken by Mikel Valganon in the Molecular Pathology Diagnostic Unit, Hammersmith Hospital, Imperial College Healthcare NHS Trust using the established BIOMED-2 protocol (Langerak et al., 2012; van Dongen et al., 2003) followed by analysis using GeneMapper 4.1 software (Applied Biosystems). In brief, the BIOMED-2 protocol consists of a multiplex PCR of the T-cell receptor-γ (TCRG) locus at DNA level, selected due to the relative simplicity of the TCRG locus. The PCR products are then analysed for size and fluorescent intensity by heteroduplex analysis or genescanning.

2.4. Analysis and quantification of proviral integration sites

2.4.1. Preparation of integration site libraries for high-throughput sequencing

This protocol has been designed for amplification of integration sites defined as the junction of 3' end of the viral genome and the host genome (Figure 2.2)

Up to 10 µg DNA was sheared by focused ultrasonication (Covaris S2) , with the following protocol: Water bath at 6 to 8^{o}C, Quick burst step: 5 seconds at 20% duty cycle, intensity level 5 and 200 cycles per burst followed by 90 seconds at 5% duty cycle, intensity level 3 and 200 cycles per burst (Figure 2.2 B).

DNA ends were then end-repaired using 15U T4 DNA polymerase (New England Biolabs - NEB), 5 units of DNA polymerase I Klenow fragment (NEB), 50 units of T4 polynucleotide kinase (NEB) and

0.8 mM dNTP (Sigma) in T4 DNA ligase buffer (NEB) at 20^{o}C for 30 minutes. This is followed by the addition of adenosine to the 3' ends of the DNA using 30U Klenow Fragment 3' to 5' exo- (NEB) in NEB2 buffer (NEB) at 37^{o}C for 30 minutes (Figure 2.2 C/D).

DNA was ligated to 100pmol/µl of a partially double stranded DNA linker using a quick ligation kit (NEB). Each linker contains a sample-identifying 6 base barcode tag to allow multiplexing during sequencing (Figure 2.3).

Nested (two step) PCR selectively amplifies HTLV-1 at the junction of the 3'LTR. The first PCR was carried out between two primers Bio3 (viral LTR) and Bio 4 (linker) and the second PCR between Bio 5 in the viral LTR and P7 contained within the linker, (Figure 2.2 E/F). The second PCR also adds the necessary adapter molecules to the amplicons, required for tethering to the Illumina flow cell. Thermal conditions for PCR1 and PCR2 were as follows: 96^{o}C for 30 seconds, 7 cycles; 94^{o}C for 5 seconds, 68^{o}C for 1 minute, followed by 23 cycles; 94^{o}C for 5 seconds and 68 ^{o}C for 1 minute followed by 68 ^{o}C for 9 minutes. Primer sequences listed in table 2.1.

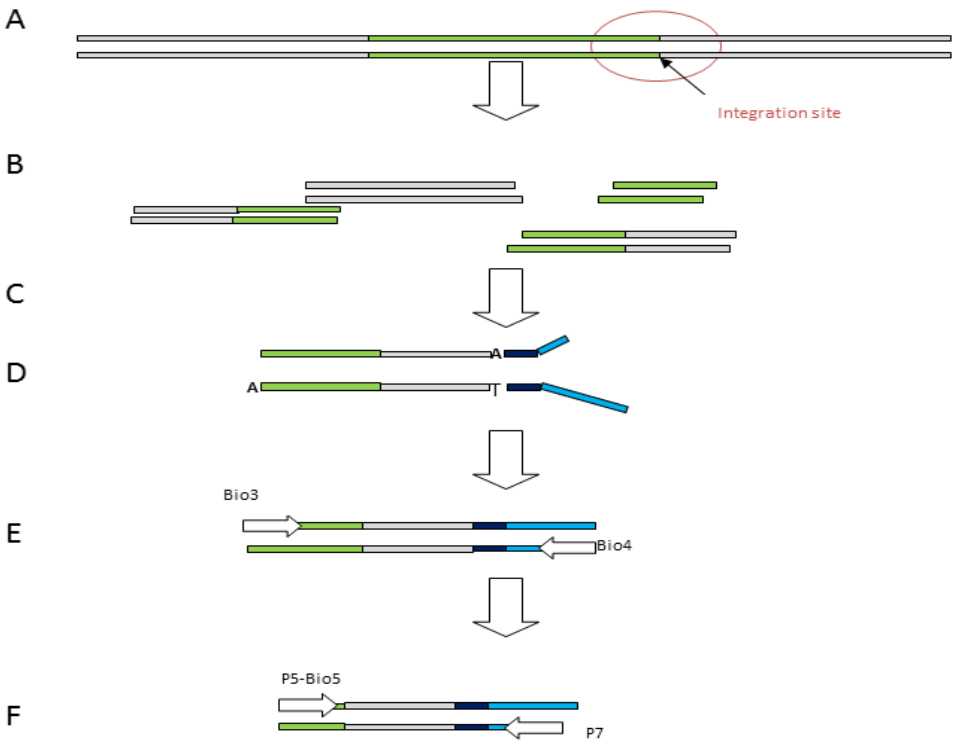


Figure 2.2: Pipeline for library preparation of HTLV-1 integration sites.

The grey bars represent host genome, green bars HTLV-1 genome and blue bars the linker. (A) The region of interest is circled in red and represents the junction of the 3'LTR and host genome.(B) DNA is randomly fragmented by sonication. (C)/(D) DNA end-repaired, adenosine molecules added to 3' ends prior to ligation of linker containing customised barcode. (E) First PCR amplifies between Bio3 (viral LTR) and Bio4 (linker). (F) Nested PCR amplifies between Bio5 (viral LTR) and P7 (linker). The P5 and P7 oligonucleotides are required for Illumina sequencing. P7 is incorporated within the design of the linker whilst the P5 oligonucleotide is added by incorporation into the PCR primer (P5-Bio5).

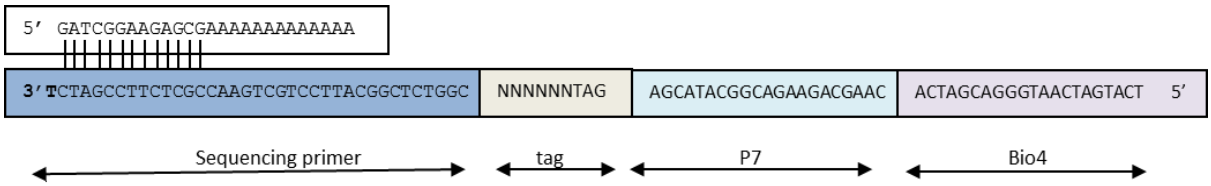


Figure 2.3: Basic linker structure

Schematic structure of the oligonucleotide sequence of the partially double-stranded linker which is ligated to DNA during library preparation and allows multiplexing of samples and incorporation of the P7 molecule required for Illumina high-throughput sequencing. The nested PCR primers for PCR1 (Bio4) and PCR2 (P7) are incorporated within the linker sequence. The linker does not contain a complementary sequence to the Bio4 primer: This increases PCR specificity since only DNA containing proviral sequences will be amplified. The 6 base bar-code tag denoted as 'NNNNNN' is located between the sequencing primer binding site and the P7 binding site.

2.4.2. Library quantification prior to sequencing

Libraries for sequencing were combined based upon DNA concentration. Libraries were then quantified by QPCR using primers specific to the amplicon structure (P5Bio5 and P7) with a standard curve based upon a previously quantified and sequenced reference library.

Each Illumina flow cell contains 8 lanes with a single DNA library per lane. One lane per flow cell is used as a control lane containing bacteriophage PhiX DNA.

The libraries were constructed such that there was only one DNA sample from an individual patient within a given lane and each linker/tag was used only once per lane.

2.4.3. High throughput sequencing and mapping using Illumina pipeline

Libraries were sequenced by Dr Laurence Game's laboratory, MRC Core Genomics Laboratory, Hammersmith Hospital Campus, London on an Illumina GA II or HiSeq2000 sequencer.

The standard Illumina analysis pipeline was used for image processing, base-calling and alignments, with default filter and quality settings. ELAND Paired algorithm (CASAVA) was used for read alignments, against the UCSC human genome build 18 and HTLV-1 sequences.

Each amplicon was sequenced on both sense and antisense strands (50 base pair reads) with a 6 bp barcode tag read (Figure 2.4).

Read 1: Uses the sequencing primer ('HTLVseq') and generates a sequence from the integration site.

Read 2: Uses the SBS8 sequencing primer and generates a sequence starting from the linker

Read 3: Uses the SBS8rev sequencing primer which generates a sequence from the linker to map the 6 base pair sample tag used for multiplexing.

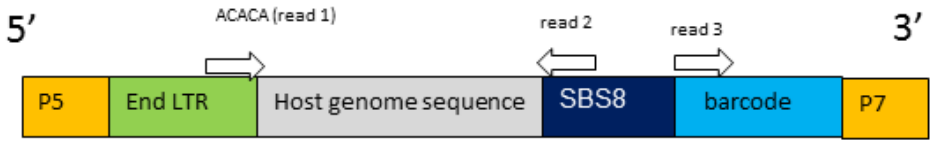


Figure 2.4: Basic structure of amplicon for high-throughput sequencing

P5 and p7 are the oligonucleotides added during LMPCR amplification required for illumina sequencing. Each amplicon is sequenced three times. Read 1 is sequenced from the end of the LTR (5' or 3' LTR), read 2 from the linker into either the host genome or the viral genome and read 3 which maps the 6 base pair barcode. Since the 3'LTR and 5'LTR are sequence identical, we would expect half the amplicons to contain HTLV-1 specific sequences and half from the host genome. These can be resolved by bioinformatic alignment.

2.4.4. Data extraction pipeline

The data extraction pipeline was customised and developed in this laboratory by Anat Melamed using Microsoft Access and designed such that other laboratory members can readily adapt it for their own usage.

The pipeline consisted of 3 key steps:-

1. Data filtering:
    i. Sequences must be specific and begin with the 3' terminal proviral LTR sequence 'ACACA'.
    ii. Sequences must pass Illumina sequencing quality scores and mapping criteria
2. Calculation of clonal abundance for each sample and each integration site:
    i. Identify distinct read 1 co-ordinates (integration site) and then, for each read 1, count the number of distinct read 2 co-ordinates (to distinguish distinct shear sites from PCR duplicates).
    ii. Sort different barcode tags (sample specific) to correctly attribute each integration to site to a particular sample
3. Data refinement:
    i. Remove artefacts e.g. frameshifts or other minor PCR or sequencing errors that cause mismapping.
    ii. Calibrate the number of shear-sites to overcome the likelihood that two samples belonging to the same clone have the same shear site by chance. Based upon a calibration experiment by Dr Nicolas Gillet, a spline function

was fitted to the data by Professor Charles Berry, UC San Diego and is applied to all datasets (Berry et al., 2012).

iii. To estimate sensitivity of the experiment by calculating an approximate input of viral copies (proviral load x input of DNA) and comparing to the proportion of proviruses recovered (observed number of copies/ input number copies)

4. Output

The output of the data-refinement pipeline is a large Microsoft Excel spread sheet containing lists of integration sites for each sample with a calculation of clonal abundance for each site. The lists of sites can then be annotated for genomic and epigenetic marks.

2.4.5. Calculation of abundance of each integration site

We define an infected clone as all the cells that share an integration site. Following calibration, the number of 'sister cells' of a clone can be calculated: this is defined by the different cells sharing an integration site (identical genomic location of read 1 but different read 2).

The relative abundance of each clone is calculated as the share of the proviral load occupied by each clone:-

Equation 2.1

$$\text{Relative abundance of a clone} = \frac{\text{(number of sisters in clone)}}{\text{(Total number of sisters in sample)}} \times 100$$

The absolute abundance of each clone is calculated as the number of sisters of each clone per 10,000 PBMC:-

Equation 2.2

$$\text{Absolute abundance of a clone} = \text{Relative abundance} \times \text{PVL}$$

2.5. Bioinformatic and statistical methods

2.5.1. Annotation of integration sites with genomic environment

Each individual integration site, from a random dataset or in vivo, was annotated for specific genomic or epigenetic elements. Detailed information on these annotations and source listed in Appendix 1.

Genomic co-ordinate data on transcriptional units, CpG islands and epigenetic marks was retrieved from publicly available NCBI ftp site (ftp.ncbi.nih.gov/gene/) and UCSC genome browser tables (http://genome.ucsc.edu/). Data on transcription factor binding site chromatin immunoprecipitation-sequencing (ChIP-seq) datasets were retrieved from published datasets (Appendix 1).

Comparison of integration sites and published annotations was carried out using a customised R-package developed and provided by Dr Nirav Malani from the Bushman group, University Pennsylvania, USA (http://malnirav.github.com/hiAnnotator).

2.5.2. Random sites (in silico datasets)

Much of the genomic integration site data analysis consisted of a comparison with an *in silico* dataset of random integration sites. An original list of 192000 genomic co-ordinates was randomly generated *in silico* by Dr Nirav Malani (FD Bushman laboratory, University Pennsylvania, USA) based upon the human genome hg18 reference build. Fifty base-pair DNA sequences at each genomic co-ordinate was generated using the Galaxy tool (http://galaxyproject.org/) and back-aligned to the human genome reference using the standard Illumina pipeline in order to generate integration sites consistent with in vivo datasets. These integration sites were then processed by Anat Melamed

(Bangham laboratory) following the same bioinformatic pipeline used to filter in vivo sequencing reads to generate 175505 integration sites for comparative analysis.

2.5.3. Calculation of Oligoclonality index

The oligoclonality index (OCI) is based upon the Gini index (Gini, 1912) and has been adapted by Gillet et al in order to quantify the relative dispersion of clonal abundances in a population of HTLV-1 infected clones (Gillet et al., 2011). The aim of the OCI was to objectively define clonality and to avoid the use of arbitrary terms such a polyclonality/ oligoclonality/ monoclonality. The Gini index for each sample was calculated using the 'reldist' R-package (http://cran.r-project.org/web/packages/reldist/index.html).

The OCI (Gini) varies between two extremes: A value of 0 denotes equal abundance of all clones, i.e. true polyclonality, whereas a value of 1 shows maximal inequality and indicates a monoclonal population of cells. This measure allows rigorous comparison of clonality between individuals, within the same individual over time and allows comparison of clonality between differing proviral loads and disease states.

2.5.4. Statistical analysis

Statistics on the large integration site datasets was carried out using R version 2.15.2 (http://www.R-project.org/). Where indicated, analysis in chapter 4 was carried out using GraphPad Prism 5. Non-parametric tests were used where appropriate and results were considered statistically significant when $p<0.05$. The Bonferroni correction for multiple comparison testing was applied where appropriate.

Table 2-1: Primer sequences

Primer name	PCR amplification	Forward or reverse (5' – 3')	Sequence
Actin-Fw	PVL	Forward	TCACCCACACTGTGCCCATCTATGA
Actin-Rev	PVL	Reverse	CATCGGAACCGCTCATTGCCGATAG
SK43	PVL	Forward	CGGATACCCAGTCTACGTGT
SK44	PVL	Reverse	GAGCCGATAACGCGTCCATCG
5LTR-Fw	5'LTR	Forward	CTCGCATCTCTCCTTCACG
5LTR-Rev	5'LTR	Reverse	CTGGTGGAAATCGTAACTGGA
LTR-seq	5'LTR sequencing	Forward	GGTTGAGTCGCGTTCT
Control-Fw	Long range PCR control	Forward	CTCTCACAGTGGGCTCGAGA
Control-Rev	Long range PCR control	Reverse	CAAAGACGTAGAGTTGAGCAAGC
LR-Fw	Long range PCR	Forward	CTTAGAGCCTCCCAGTGAAAAACATTTCC
LR-Rev	Long range PCR	Reverse	GATGCATGGTCCTGCAAGGATAACA
Exon2-Fw	Tax exon 2	Forward	CCTCAGCAATAAACAAACCC
Exon2-Rev	Tax exon 2	Reverse	CAATTGTGAGAGTACAGCAG
Exon 3-Fw	Tax exon 3	Forward	ATACAAAGTTAACCATGCTT
Exon 3-Rev	Tax exon 3	Reverse	AGACGTCAGAGCCTTAGTCT
Exon3seq-I	Tax exon 3 sequencing	Forward	ATACAAAGTTAACCATGCTT
Exon3seq-II	Tax exon 3 sequencing	Forward	CGTTATCGGCTCAGCTCTACA
Exon3seq-III	Tax exon 3 sequencing	Forward	TTCCGTTCCACTCAACCCTC
Exon3seq-IV	Tax exon 3 sequencing	Reverse	AGACGTCAGAGCCTTAGTCT
Exon3seq-V	Tax exon 3 sequencing	Reverse	GGGTTCCATGTATCCATTTC
Exon3seq-VI	Tax exon 3 sequencing	Reverse	GTCCAAATAAGGCCTGGAGT
U3 meth-Fw	MS-PCR PCR 1 methylated	Forward	TTAAGTCGTTTTTAGGCGTTGAC
U3 unmeth-Fw	MS-PCR PCR 1 unmethylated	Forward	TTAAGTTGTTTTTAGGTGTTGAT
Gag rev (bisulfite)	MS-PCR/ BSP	Reverse	AAAAAAATTTAACCCATTACC

R meth-Fw	MS-PCR PCR 2 methylated	Forward	GAGGTCGTTATTTACGTCGGTTGAGTC
R-unmeth-Fw	MS-PCR PCR 2 unmethylated	Forward	GAGGTTGTTATTTATGTTGGTTGAGTT
Heminested-Rev	BSP PCR 2 heminested	Reverse	ACCCCCTCCTAAACTATCCC
ATL 37 BSP Fw	BSP PCR1/2	Forward	TTTAGGGTAATTGATTTTTTGG
ATL 35 BSP Fw	BSP PCR1/2	Forward	TGTTTTTTTTGATTTTTGTTGG
ATL 164 BSP Fw	BSP PCR1/2	Forward	TTTTTTTGGTATTTGGAAGAAAA
ATL 38 BSP Fw	BSP PCR1/2	Forward	GTTTTTATGGGGTAGGGATAGA
ATL 52 BSP Fw	BSP PCR1/2	Forward	TTGTAGTTGAGAGGGTTGAAATT
Bio3	LMPCR PCR1	Forward	CCTTTCATTCACGACTGACTGCCG
Bio4	LMPCR PCR1	Reverse	TCATGATCAATGGGACGATCA
(P5)Bio5	LMPCR PCR2	Forward	(AATGATACGGCGACCACCGA)GATCTACA CTGGCTCGGAGCCAGCGACAGCCCAT
P7	LMPCR PCR2	Reverse	CAAGCAGAAGACGGCATACGA
HTLVseq	LMPCR illumina sequencing	Read1	CAGCCCATTCTATAGCACTCTCCAGGAGAGAAACTTAGT
SBS8	LMPCR illumina sequencing	Read2	CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT
SBS8rev	LMPCR illumina sequencing	tag read	GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCG

Chapter 3. Integration site analysis of naturally infected HTLV-1 $CD4^+CD25^+$ T cell clones

3.1. Introduction

3.1.1. High throughput sequencing in retroviral mapping

High throughput sequencing and sequencing of most of the human genome has revolutionised science by allowing users to acquire genome-wide data using massively parallel sequencing approaches. This opened the door to deeper analysis of integration site mapping of retroviruses (Brady et al., 2009; Bushman et al., 2008), vectors (Li et al., 2011; Ronen et al., 2011), retroelements (Williams-Carrier et al., 2010) and detailed evaluation of adverse events in gene therapy trials (Cavazzana-Calvo et al., 2010; Hacein-Bey-Abina et al., 2008; Wang et al., 2008).

All the commercial platforms share 3 critical steps: DNA-sample preparation which requires random DNA fragmentation and addition of linkers ('library preparation'), immobilisation of the library to a solid reaction chamber (e.g. a flow cell) and followed by sequencing.

The Illumina method utilises 'bridge amplification' to generate clusters for sequencing: Immobilised primers are present on the flow cell surface and contain sequences that correspond to the DNA adapters present in the prepared DNA library. Bridge-PCR initiates by hybridisation of the immobilised sequencing library fragment and a primer to form a surface-supported molecular bridge structure. The arched molecule is a template for a DNA polymerase-based extension reaction and the resulting bridged double-stranded DNA is freed using a denaturing reagent. Repeated cycles generate groups of thousands of molecules known as "clusters" on each flow cell lane. DNA clusters are finalised for sequencing by unbinding the complementary DNA strand to retain single molecular strands in each cluster. The prepared flow cell is then connected to a high-throughput imaging system. Illumina's 'sequencing by synthesis' technology utilises four fluorescently labelled nucleotides to sequence tens of millions of clusters on the flow cell in parallel. During each sequencing cycle, a single labelled dNTP is added to the nucleotide chain. The nucleotide label serves as a terminator for polymerisation, and so after each dNTP incorporation, the fluorescent dye

is imaged to identify the base and then enzymatically cleaved to allow incorporation of the next base (Bentley et al., 2008). At present, Illumina HiSeq 2000 sequencing can generate 3 billion reads per flow cell (Liu et al., 2012).

Following the rapid take-up of high throughput sequencing techniques in integration site mapping, many challenges remain. Principally to improve the recovery of each DNA sample, quality of sequencing and to quantitate the relative abundances of each integration site.

More recently, within this group, Gillet et al (Gillet et al., 2011) developed a customised novel high throughput sequencing approach using an Illumina platform to accurately map and quantify proviral integration sites in natural HTLV-1 infection. The integration site is defined as the junction between the 3'LTR of the provirus and the host genome. This method is based upon classical LMPCR and overcomes the described limitations of using restriction enzymes and Sanger sequencing and can accurately quantify the abundance of each detected integration site.

In this method there are 3 main improvements over the classical LMPCR and Sanger sequencing technique.

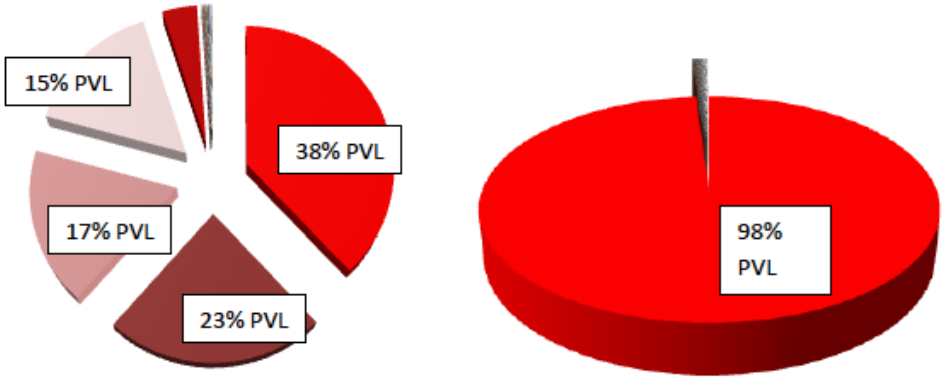
1. Restriction enzymes are replaced with sonication, which randomly shears the DNA and prevents systematic loss of integration sites, such as those that lie distant to a restriction enzyme cleavage site. Furthermore, sonication allows for accurate measurement of clonal abundance since PCR duplicates can be distinguished from true sister cells within a clone. That is to say, following sequencing, PCR duplicates are identified to have the same integration site (read1) and the same shear site (read2), whereas sister-cells within a clone have the same integration site (read1) but a different shear site (read 2) and can therefore be quantified programmatically (Gillet et al., 2011).
2. The DNA fragments generated by sonication are then ligated to partially double-stranded linkers, which are amplified by nested PCR to increase sensitivity and specificity. The

partially double-stranded linker results in increased specificity as selective amplification occurs between the conserved proviral LTR sequence and the linker. Each linker contains a 6 base pair 'tag' (or 'barcode') that allows for multiplexing of approximately 40 samples/lane of a HiSeq 2000 platform. Specificity is further increased by utilising a sequencing primer that binds 5 base-pairs short of the terminal LTR sequence. This means that all sequencing reads start with the same proviral sequence 'ACACA' and eliminates any mispriming events to focus upon true integration sites. Each amplicon is sequenced three times. The first sequences 50 base pairs complementary to the HTLV-1 sequence (read 1), the second sequences 50 base pairs complementary to the linker (read 2) and the third reads the barcode tag to allow multiplexing of samples.

3. High throughput sequencing of integration sites on ex-vivo PBMC generates a large number of reads originating from a high number of unique clones within each individual. Meekings et al, who utilised the classical LMPCR technique followed by cloning and Sanger sequencing, mapped a mean number of 13 unique HTLV-1 integration sites from 10ug DNA per asymptomatic carrier (range 1-24) (Meekings et al., 2008) whilst Gillet et al, who developed this high-throughput approach, identified a mean of 1489 integration sites per asymptomatic carrier (Gillet et al., 2011). Prior to the development of this technique it was estimated that each individual carried a total of approximately 100 clones. Mathematical modelling with high throughput data from the Bangham laboratory suggests that each asymptomatic carrier possess between 10^4-10^5 distinct clones (Laydon et al 2013, in submission).

3.1.2. Clonal distribution of HTLV-1 integration sites in PBMC ex vivo

Quantification of clonal abundance from the number of integration sites within an individual requires knowledge of the (average) number of proviruses integrated within each host genome. The evidence for a single provirus per infected cell is based upon the diagnostic criteria of monoclonal integration in ATL tumour biopsy tissue, which is typically demonstrated by Southern blot (Yoshida et al., 1982) or classical LMPCR or IPCR techniques (Takemoto et al., 1994; Wattel et al., 1995). However, these methods were not sufficiently powered to detect multiple proviruses within a single clone of cells since restriction enzymes and preferential PCR amplification of short products are likely to recover a single integration site, and to potentially miss any additional integration sites, leading to the false conclusion that there is a single proviral integration per cell. A clinical example is illustrated in Figure 3.1 showing the results from high throughput sequencing for two locally treated clinical cases of chronic ATL. Each slice of the pie chart shows the relative abundance of each clone within an individual sample. In one sample, patient LFK (right panel), there is a single dominant proviral integration site occupying 98% of the proviral load, whilst the second sample, from patient LGB, has the same clinical disease, but multiple 'expanded' integration sites. When more than 1 dominant integration site is identified, many questions arise e.g. how many of these integration sites might be malignant or whether there is one malignant clone containing multiple proviruses and whether these different clonal distributions carry any clinical significance e.g. in response to treatment or relapse.



Patient code: LGB	LFK
Disease: Chronic ATLL	Chronic ATLL
PVL: 140%	14%
No. UIS: 14	57

Figure 3.1: ATL is not necessarily caused by a monoclonal expansion of T-cell clones

An example of the relative abundance of integration sites detected in two patients with chronic ATL. The case on the right (patient LFK) shows that there is a single abundant integration site (constituting 98% of the proviral load) with the remaining 2% of the PVL occupied by a further 56 small abundance integration sites. However, in the case on the left (patient LGB), also with chronic ATL, it can be seen that there are 4 abundant integration sites, constituting between 15 to 38% of the proviral load.

3.1.3. Aim

The work within this chapter tests the hypothesis that during natural infection, each infected cell contains a single provirus. In order to test this hypothesis, clones of infected CD4+ CD25+ HTLV-1 infected T-cells were isolated by a method of limiting dilution and expanded in vitro. The T-cell clones were isolated from PBMC ex-vivo of 10 patients, with different clinical manifestations of HTLV-1 infection. HTLV-1 infection of expanded clones was confirmed by the presence of the *tax* gene by PCR and, if positive, followed by high throughput sequencing to determine the abundance of each unique integration site.

3.2. Results

3.2.1. Non-malignant infected CD4+ T cell clones contain a single integrated provirus.

Twenty-seven HTLV-1 infected T cell cultures were successfully expanded in vitro (*tax* gene present by PCR). To accurately quantify the abundance of unique integration sites within each culture, linker-mediated PR and high-throughput sequencing were undertaken.

We found that in 26 of these cultures, a single unique integration site constituted 99.9% of the proviruses detected (range 95.5-100%), with a median of 206197 sequence reads (range 6621-2599892 reads). This indicates that 26 clones had been cultured, each with a single dominant integration site. The remainder of integration sites detected within each clone were detected with low frequency (median 3 sequence reads) suggesting incomplete inhibition of HTLV-1 infectious spread during *in vitro* culture by raltegravir.

One HTLV-1 culture contained cells with two equally abundant integration sites. This observation suggested either the presence of two proviruses within each genome or that two clones had been cultured within the same well. Analysis of TCRG rearrangement revealed two distinct gene rearrangements in the Vγ1-8 region, consistent with the presence of two T cell clones. Results for each culture summarised in table 3.1. The genomic location for each of the dominant sites is reference to hg 18 reference genome (NCBI 36.1).

Clone	Clone derived from	Genomic location of dominant integration site	Proviral load composed of dominant integration site, %	Proviral structure
1	Asymptomatic carrier 1	Chr 6: 84613188	98.95	Complete
2	Asymptomatic carrier 1	Chr 1: 184745603	99.95	Type 2 defective
3	HAM/TSP patient 1	Chr 16: 52158560	99.9	Type 2 defective
4	HAM/TSP patient 1	Chr X: 114257179	99.9	Complete
5	HAM/TSP patient 1	Chr 10: 2277787	100	Complete
6	HAM/TSP patient 1	Chr 13: 74997821	98.88	Complete
7	HAM/TSP patient 2	Chr 10: 80828281	99.96	Complete
8	HAM/TSP patient 2	Chr 5: 1997980	99.9	Complete
9	HAM/TSP patient 2	Chr 4: 70601874	95.2	Complete
10	HAM/TSP patient 2	Chr X: 129048967	100	Complete
11	HAM/TSP patient 2	Chr 14: 45274700	100	Complete
12	HAM/TSP patient 3	Chr 4: 107219655	99.9	Complete
13	HAM/TSP patient 3	Chr 12: 40920384	100	Complete
14	HAM/TSP patient 4	Chr 5: 50609660	98.4	Complete
15	HAM/TSP patient 4	Chr 3: 76576960	99.9	Complete
16	HAM/TSP patient 4	Chr 19: 32974427	100	Complete
17	HAM/TSP patient 4	Chr 19: 38521388	100	Complete
18	HAM/TSP patient 4	Chr 14: 80728657	95.5	Complete
19	HAM/TSP patient 4	Chr 4: 169472575	100	Complete
20	HAM/TSP patient 5	Chr 3:75952018	100	Complete
21	Polymyositis patient 1	Chr 2: 214080658	99.95	Complete
22	Polymyositis patient 1	Chr 6: 167451858	99.9	Complete
23	Polymyositis patient 1	Chr 22: 31885379	100	Complete
24	Polymyositis patient 1	Chr 3: 32582626	98.6	Complete
25	ATL patient 1	Chr 22: 42654531	99.89	Complete
26	ATL patient 2	Chr 4: 9905297	99.8	Complete
27	Asymptomatic carrier 2	Chr 8: 87708389	52	Complete
27	Asymptomatic carrier 2	Chr 9: 11420980	42	Complete

Table 3-1: Summary of identified integration sites derived from 10 individuals with HTLV-1 infection

3.2.2. Confirmation of PCR bias to selectively amplify short PCR products

PCR amplification is biased toward short PCR products (Figure 3.2). The data within Figure 3.2 is taken from a typical sequencing flow cell (Genome Analyser II) and demonstrates that there are more PCR duplicates of short DNA fragments, and that the preferred amplicon length for PCR amplification and cluster generation is approximately 150-500 base pairs. There was a direct inverse correlation between amplicon size and the number of PCR duplicates (Spearman r -0.87, $p<0.0001$).

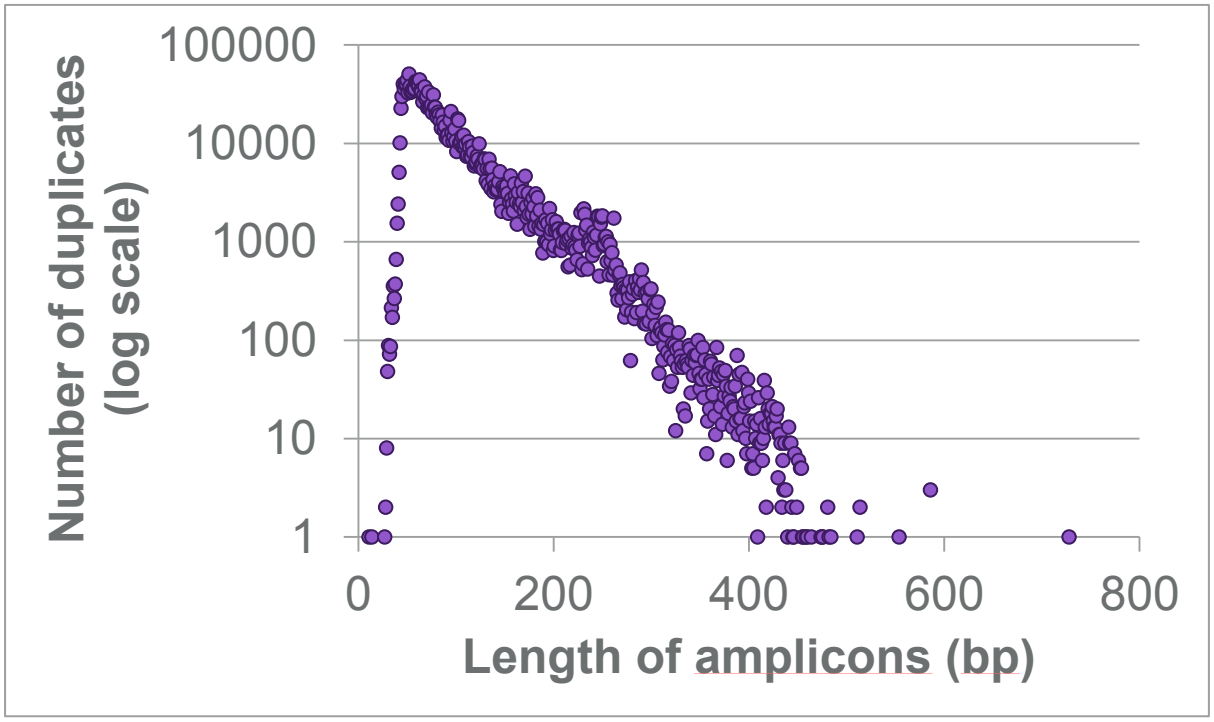


Figure 3.2: PCR most efficiently amplifies short amplicons

The number of PCR duplicates (on a log scale) is shown on the y-axis and the size of PCR amplicons (x-axis). Amplicons of approximately 150 base pairs were most efficiently amplified by PCR, whilst amplicons over 500 base pairs were infrequently amplified.

3.2.3. Clones with defective proviruses can be isolated ex-vivo

Long range PCR (Methods section 2.3.3) on DNA from each culture showed that 2 out of 27 clones (7.4%) contained a type 2 defective provirus and 25 clones contained a complete provirus. No type 1 defective proviruses were isolated.

Type 2 defective proviruses lack the 5'LTR-*gag* region containing the *tax* gene promoter/enhancer elements and are unable to express Tax protein. Miyazaki et al measured the frequencies of type 2 defective proviruses in two asymptomatic carriers by QPCR on DNA from a mixed population of PBMCs and reported frequencies of 3.9% and 0.6% respectively (Miyazaki et al., 2007), whilst Ramirez et al determined by a similar QPCR technique that no seropositive patients with HAM/TSP had type 2 defective proviruses (Ramirez et al., 2003). These studies were limited by the use of few patients, mixed PBMC populations and the relatively low sensitivity of QPCR to accurately quantify minor populations.

Our results suggest that defective proviruses may not be so rare in non-malignant disease, although our study was not powered to robustly quantify the absolute frequency of defective proviruses in different disease states – this would require the isolation and expansion of hundreds of unique clones.

3.2.4. T-cell clones from patients with acute ATL are difficult to expand in vitro

When we compared the integration sites of the clones that were isolated from patients with acute ATL with the putatively malignant clone (identified by high throughput sequencing of PBMC), these were not the large expanded clone but originated from one of the smaller background clones. In these two individuals, the abundant, putatively malignant clone made up 97% and 52% of the proviral load, whilst the isolated clones made up 0.12% and 1.6% of the proviral load respectively.

In fact, this preferential isolation of low-abundance clones was observed in the clones derived from all the non-malignant patients too: The T-cell clones that were isolated were of low abundance in mixed PBMCs (either not detected in the mixed PBMC or up to a maximum relative abundance 0.17%).

3.3. Discussion

3.3.1. Non-malignant clones of HTLV-1 contain a single provirus

In this study we found that all analysed cultures of HTLV-1 infected CD4+CD25+ T-cells contained a single provirus in the host genome, consistent with our hypothesis. This allows us to infer that each unique integration site mapped within the PBMCs of infected non-malignant cases may be called a clone, and allows accurate calculation of clonal abundance when analysing sequencing data on PBMCs. Since we did not isolate the abnormally expanded, putatively malignant clone from the ATL cases, we make no inference about the malignant clones. Whilst we undertook this study on patients with a broad spectrum of clinical manifestations of HTLV-1 infection, these were individually of small number (5 HAM/TSP, 2 AC, 1 polymyositis, 2 acute ATL). However, this study was not designed to investigate the mean proviral copy number between different disease states- this would require large numbers from each subtype of disease.

The approach using high throughput sequencing was justified, since we have observed preferential recovery of amplicons measuring between 150-500 base pairs. If we had used a classical LMPCR method that utilised restriction enzyme digestion, cloning and Sanger sequencing of individual T-cell cultures, we could not have been confident of the results since the finding of a single integration site could have arisen from a preferential restriction position within the host genome.

One possible mechanism for the observed findings of a single provirus in each T-cell clone is superinfection resistance (SIR), which is the capacity of a cell to prevent a second infection by a closely related virus. The mechanism of SIR in HTLV-1 is unknown: Studies of SIR mechanisms in simple retroviruses (e.g. Murine Leukaemia Virus, MLV) or complex retroviruses (e.g. HIV) do not reveal a common mechanism and do not appear to be mediated by neutralising antibodies and/or

virus specific CD8+T-cells (Allen and Altfeld, 2003). Viral recombination, an indicator of superinfection at a cellular level, is considered to be an important viral evolutionary strategy (Burke, 1997). Since recombinants have been identified in HIV patients it can be concluded that SIR is not absolute.

Recently, Josefsson et al quantified the number of HIV-1 proviruses from 9 HIV-infected individuals and identified that >85% of infected CD4+ T-cells contain a single copy of HIV DNA which was phylogenetically similar to plasma RNA, implying limited potential for recombination (Josefsson et al., 2011). Although it was widely thought that HIV infected CD4+ cells frequently contained multiple proviruses, to date they have only been identified in vivo in spleen tissue, which may act as a cellular reservoir (Gratton et al., 2000; Jung et al., 2002).

One of the major characteristics of HIV infected cells is down-modulation of the CD4 receptor (Levin et al., 2010) by viral Vpu, Env and Nef. As receptor down modulation is a simple way of preventing second viral infection, and a method that is successfully used by other retroviruses (e.g. Foamy virus), CD4 down-modulation was initially assumed to be the main SIR mechanism in HIV infection. However, SIR has been demonstrated to occur early (Volsky et al., 1996) (4-24 hours after primary HIV infection), whilst CD4 down-regulation occurs later (two days after infection) suggesting an alternative mechanisms. The mechanism responsible for SIR in HIV-1 infected cells is of interest particularly for the development of novel approaches to HIV therapy.

In HTLV-1 infection, the virus appears to persist chiefly by mitotic proliferation of infected CD4+ T-cells, rather than infectious spread. Since mitotic proliferation utilises the highly faithful host DNA polymerase II enzyme, rather than the error prone reverse transcriptase, the sequence of HTLV-1 is highly conserved. Utilisation of this strategy does not provide opportunity for productive viral recombination and thus no advantage to the HTLV-1 virus in allowing a multiply infected cell since this would not generate productive viral recombination. Indeed, it can be hypothesised that

multiple infection is more likely to result in increased viral gene expression and subsequent CTL killing. It is likely that HTLV-1 has evolved an as yet unidentified mechanism of superinfection resistance.

3.3.2. Clones carrying defective proviruses can be isolated from ACs and patients with HAM/TSP.

Defective proviruses are common in ATL and the frequency has been reported variably between 25.7 - 56% cases (Korber et al., 1991; Ohshima et al., 1991; Tamiya et al., 1996; Tsukasaki et al., 1997). The discrepancy between frequencies in the literature is in part due to the clinical phenotypes of ATL within each cohort, since acute ATL has a higher frequency of defective provirus than lymphoma subtype (31.9% versus 8.7%, (Kamihira et al., 2005). Secondly, there is variation in the method used to detect defective proviruses; both southern blot hybridisation and PCR-based techniques have been reported.

While defective proviruses have been frequently observed in ATL, it remains controversial as to whether a defective provirus is advantageous to the development of leukaemia, or has arisen as a consequence of the leukemic transformation secondary to widespread genomic instability. During the process of HTLV-1 integration, the viral integrase generates a short repetitive sequence of 6 base-pairs adjacent to both LTRs. Miyazaki et al sequenced across the 6 base pair repeats that lie adjacent to each LTR in 12 cases of ATL, and demonstrated that these repeat sequences were present in 8/12 cases and deleted in 4/12 cases. These findings show that deletion of the 5'LTR can occur pre-integration (Miyazaki et al., 2007). They quantified the frequency of type 2 defective proviruses in two asymptomatic carriers at <3.9%.

Our findings suggest that type 2 defective proviruses in non-malignant infection may be more frequent than these estimates. There have been no reports on the frequency of these in CD4+ clones previously, or on a large scale in mixed PBMC ex vivo. Whilst the cloning protocol was sufficient to clone many uninfected T-cells, it was an unexpected finding that HTLV-1 infected T-cell clones containing a type 2 defective provirus were successfully cloned, since expressing Tax might provide a proliferative advantage in culture. The proviral integration sites of both these clones were intergenic within the host genome and did not suggest the identity of the selection advantage. However, we selected our patients for cloning on the basis of a higher-than-average proviral load (in order to increase the likelihood of cloning infected CD4+ T-cells) and so perhaps the PBMC from these patients are biased to contain an increased number of defective proviruses.

3.3.3. The malignant clone from ATL cells is difficult to isolate and expand in vitro

The expanded ATL cells are notoriously difficult to isolate and culture ex vivo. In previous culture systems primary ATL cells were able to grow in liquid culture containing IL-2, although they only showed transient cytokine-dependent proliferation (Aboud et al., 1987; Lunardi-Iskandar et al., 1993; Uchiyama et al., 1985). Furthermore, it was shown that in culture ATL cells began to abundantly express the protein Tax, which interacts with a number of transcription factors resulting in the trans-activation or repression of many cellular genes involved in cell growth or apoptosis (Yoshida, 2001). However, ATL cells do not produce high levels of Tax protein in vivo, although *tax* gene expression has been reported using PCR ISH and RT-PCR ISH (Ohshima et al., 1996). Together these findings suggest that IL-2 supported growth in vitro does not necessarily reflect the growth mechanism of ATL cells in vivo. It has been speculated that there may be important interactions between the malignant cell and its micro-environment that may be replicated in vitro using co-

culture with a bone-marrow-derived stromal cell line (Bajenoff et al., 2006; Imura et al., 1997; Sasaki et al., 2005). Nagai (Nagai et al., 2008) published a novel technique that utilises a stromal cell layer co-culture system to provide various factors that support proliferation of ATL cells. Of note, whilst this group reported moderate success (growth in 4/8 acute ATL cases), they assumed monoclonal bands identified by southern blot hybridisation represented the malignant clone and did not specifically map the integration site to confirm.

There is currently an on-going project in this laboratory to modify and optimise the protocol published by Nagai et al in order to isolate and expand the malignant clones and to confirm the presence of the presumed malignant integration site by PCR.

3.3.4. Chapter summary

Modified linker-mediated PCR followed by high-throughput sequencing is an unbiased approach to quantify HTLV-1 integration sites. We have shown that, in natural infection, non-malignant HTVL-1 infected CD4+ T-cell clones contain a single provirus and suggest that clones carrying defective proviruses may be more frequent in ACs or HAM individuals than previously estimated.

3.3.5. Publication associated with this chapter

HTLV-1-infected T cells contain a single integrated provirus in natural infection,

Cook LB, Rowan AG, Melamed A, Taylor GP, Bangham CR

Blood 2012, 120(17):3488-3490

Chapter 4. Characterisation of ATL cohort

4.1. Chapter abstract and summary

The aim of this chapter is to characterise a large ATL cohort (n=197) of mixed clinical subtype in terms of PVL, capacity for Tax expression (defective provirus, tax gene mutation or hypermethylated 5'LTR) and to quantify the clonal distribution. This will allow us to determine that the samples are representative of published ATL cohorts, for the basis of a robust analysis of the influence of the host genomic landscape (presented in chapter 5).

4.2. Introduction

For several decades the molecular basis of cancer has been thought to be clonal expansion of a single common precursor cell that has undergone random genetic or epigenetic damage (mutational or epigenetic driver events) that result in a proliferative advantage, followed by subsequent genetic or epigenetic damage resulting in a malignant tumour (Aparicio and Caldas, 2013; Parsons, 2008) . It is thought that each mutation may individually only confer a small growth advantage but that cumulatively this advantage may result in massive expansion over many years (Vogelstein et al., 2013).

There is emerging evidence for a polyclonal origin of tumours where distinct small populations of tumour cells are present amongst a seemingly monoclonal tumour. The notion that cancers are ecosystems of evolving clones has implications for clinical practice in order to develop an accurate scientific understanding of tumourigenesis, to accurately model the mechanisms of tumour development and to improve understanding of different treatment approaches and to predict and manage relapsed disease (Aparicio and Caldas, 2013). These subclones might respond quite differently to the chemotherapy directed against the major tumour bulk. Initial response to therapy followed by relapse may result in the clonal outgrowth of the untargeted tumour population, particularly if therapy is focused upon the presence of a specific lesion, which may not be present in the subclones. This is of emerging interest in the HTLV-1 field since acute, bulky or lymphomatous ATL cases are typically associated with a brief clinical response when treated with chemotherapy, usually followed by a rapid and treatment-resistant relapse. Furthermore, a recent study by Seto et al demonstrated clonal evolution within individual tumour biopsies in ATL by using an array-comparative genomic hybridisation (CGH) method (Umino and Seto, 2013).

As discussed in chapter 3, ATL is regarded as a monoclonal tumour and the demonstration of monoclonal proviral integration is cited as part of the diagnostic criteria, although not uniformly

performed (Tsukasaki et al., 2009; Yoshida et al., 1984). One of the aims of this chapter is to quantify clonality at the proviral level in order to test the hypothesis that a single provirus is integrated within all cells (implying a common cell of origin) using our high-throughput mapping strategy.

4.2.1. Tax expression is not a requirement for ATL cells

Tax is thought to play a central role in ATL leukaemogenesis by its pleiotropic actions such as transactivation of NFκB and CREB (Franchini, 1995; Yoshida, 2001), and functional inactivation of p16, p53 and MAD1 (Ariumi et al., 2000; Jin et al., 1998; Suzuki et al., 1999). However, the enigma of Tax-induced leukaemogenesis is that Tax expression is detected in only 40% of human ATL cases (Akagi et al., 1995). There are three commonly observed mechanisms to disrupt Tax: Deletion of the 5'LTR which contains the promoter and enhancer elements for viral transcription from the sense strand (Tamiya et al., 1996); CpG hypermethylation of the promoter elements of the 5'LTR leading to transcriptional silencing of the sense strand (Taniguchi et al., 2005); and genetic mutations in the *tax* gene itself leading to silencing in approximately 10% cases (Takeda et al., 2004). It is noteworthy that these changes are predominantly observed in the aggressive forms of ATL (acute and lymphomatous subtypes). By contrast *HBZ* transcripts, encoded on the antisense strand, have been detected in 100% of cases of ATL (Satou and Matsuoka, 2007) and the promoter for *HBZ* is present in the U5 sequence of the 3'LTR. HBZ also possesses proliferative functions (Satou et al., 2006) and was originally reported to suppress Tax transcription (Gaudray et al., 2002). More recently HBZ transgenic mice have also been demonstrated to develop tumours (Satou et al., 2011). These findings have given rise to a widely held belief that Tax expression is important during the initiation of ATL and HBZ is required to maintain the transformed phenotype (Matsuoka and Jeang, 2011).

4.2.2. Subtypes of defective provirus in ATL

The 5'LTR is critical for transcription of viral genes encoded on the sense strand of the provirus and its loss results in the inability to express viral proteins including Tax. Even cells with a type 1 defect have been found not to produce Tax in vitro, presumed to be due to the loss of an internal promoter in the *pol* region (Hiramatsu and Yoshikura, 1986) (Figure 2.1).

A feature of retroviral integration is the presence of a short repeat sequence in the host genome flanking the 5' and 3'LTR of the provirus (Figure 1.1). HTLV-1 normally generates six base pair repeats which Miyazaki et al demonstrated were present at the junction of the host genome and a type 2 defective integrated provirus in 8/12 ATL cases (Miyazaki et al., 2007). This suggests that in these malignant cases Tax protein was never expressed during any stage of clonal expansion and is therefore contrary to the widely held opinion that Tax is important in early clonal proliferation.

Both type 1 and type 2 defective proviruses are reported in 30-50% of ATL cases, and type 2 proviruses are more frequently observed in aggressive subtypes rather than the chronic or smouldering ATL, suggesting a correlation between clinical subtype and defective provirus (Tamiya et al., 1996). The explanation for such a high frequency of defective provirus remains uncertain, although it has been suggested that cells infected with a defective provirus cannot express Tax and thus escape the immunodominant CTL response, resulting in a greater likelihood of leukemic transformation.

4.2.3. Tax mutations in ATL

Missense mutations in the *tax* gene have been widely reported and frequently associate with proviral subgroup based upon LTR sequence and are classified as 'Tax A' or 'Tax B' genotype. The widely used Seiki reference genome is Cosmopolitan B (Tax B) and it has been shown that specific nucleotide substitutions are associated with Type A, which are: C>T at position 7897, C>T at position 7959, G>A at position 8208 and A>C/G at position 8344. In addition, nucleotide alterations at positions 7720, 8120 and 8297 have been observed as recurrent mutations within the Tax A subtype (Furukawa et al., 2000). These Tax A nucleotide alterations are associated with amino acid change A>V at Tax position 221, S>N at Tax position 304 and Q>E at Tax position 334 and are not ATL disease specific (Furukawa et al., 2000). Genetic alterations in the *tax* gene that can escape the CTL response have been described (Furukawa et al., 2001). Furukawa et al firstly identified a stop codon in the 5' half of the tax gene that loses transactivation activity on the viral enhancer; second, a specific amino acid change that alters the immunodominant CTL Tax_{11-19} epitope was observed in the presence of HLA-A*02 rendering it undetectable by Tax_{11-19} -specific CTLs and third, large deletions of the tax gene were observed. Together, these findings suggest that at some stage HTLV-1 infected cells can escape the host immune system, gaining a selective advantage which perhaps allows these clones of cells to accumulate further genetic and epigenetic alterations, culminating in malignant transformation. Crucially, identical *tax* gene mutations were found within healthy siblings, suggesting that these mutations occurred early, in the mother, and were transmissible (Furukawa et al., 2001).

4.2.4. Hypermethylation of the 5'LTR in the ATL

Bisulfite-sequencing PCR (BSP), in which sequences from cloned PCR products generated by primers unbiased for either methylated or unmethylated DNA, has shown that complete methylation of the U3 region of the 5'LTR of HTLV-1 provirus results in loss of *tax* gene transcription (Takeda et al., 2004). Hypermethylation of the 5'LTR promoter/enhancer has been reported in 14% ATL samples (Takeda et al., 2004) whilst partial methylation of the 5'LTR is predominant and seen in 50% cases but does not silence *tax* gene transcription in cell lines or in fresh ATL cells (Takeda et al., 2004).

4.3. Results

4.3.1. Samples removed from analysis following unblinding of clinical diagnosis

The original cohort of ATL DNA samples provided by Professor Matsuoka's group in Japan consisted of 242 samples. The clinical subtype of ATL was unknown to me during both the laboratory work and bioinformatic analysis with unblinding at the final step. This blinding allowed unbiased processing and data analysis. However, some samples had to be disregarded either during the bioinformatic analysis or following unblinding, for the following reasons:-

Four cases were removed from the analysis as the provided tumour material did not show any evidence of HTLV-1 infection (undetectable long range PCR, undetectable 5'LTR by PCR, undetectable exon 2 or 3 tax gene by nested PCR, undetectable integration sites following LMPCR and high throughput sequencing).

Seventeen cases were removed following review of clinical results (undertaken in Japan) which suggested they had other haematological malignancies (3 cases with B-cell non Hodgkin lymphoma in HTLV-1 carriers), 3 cases were from asymptomatic carriers , 9 cases were from remission samples (not pre-treatment) and 2 were duplicate cases.

Following high throughput sequencing a further 24 samples were not analysed for a variety of reasons: 10 cases were not uniquely mapped to the host genome at both read 1 or read 2 (as required by the HTS protocol). Seven cases failed to amplify by LMPCR for reasons unknown, although 1 was found to have a primer binding site polymorphism. Seven cases amplified in the laboratory (as seen by gel smear) but failed to adequately sequence. Our collaborators confirmed that DNA samples were not always freshly extracted and some DNA samples were several (up to 20) years old and may have degraded during storage. There were no common features in terms of

clinical subtype or proviral load in the samples that were not suitable for bioinformatic analysis. This resulted in 197 pre-treatment diagnostic ATL samples, suitable for analysis.

A table detailing each analysed ATL case is listed in Appendix 2 and an attrition table which details where cases were removed from the analysis in Appendix 4.

4.3.2. Number of cases within each clinical subtype and proviral load.

ATL consists of 4 main clinical subtypes – acute, chronic, lymphoma and smouldering cases. Acute cases constitute approximately 60% of clinical presentations in Japan, lymphoma 20%, chronic 15% and smouldering 5%. The proportions of each subtype in the present cohort are summarised in Table 4.1.

Table 4-1: Proportion of each clinical subtype within cohort

	Acute	Chronic	Lymphoma	Smoldering	Unknown
Number cases	128	30	31	6	2
% of cases	65	15	16	3	1

4.3.3. Defective proviruses are detected in 39% of the ATL cohort

To identify the structure of the provirus, we amplified the provirus by long range PCR (Tamiya et al., 1996). We identified complete, type 1, type 2, multiple proviruses and a new entity, named 'indeterminate' defective proviruses in which both of the long range PCRA and PCR B products were absent (Table 4.2). Multiple proviruses were determined if there was both a complete provirus present and an additional shorter band, consistent with the presence of an additional type 1 provirus.

Table 4-2: Subtypes of provirus identified by long-range PCR

	Complete provirus only	Type-1 defective provirus	Type-2 defective provirus	Multiple provirues	Indeterminate defective proviruses
Number of cases	120	26	36	5	10

Our data identifies 39% cases with a defective provirus which is in keeping with previously reported frequencies (Tamiya et al., 1996). The indeterminate defective proviruses all contained a single dominant integration site, consistent with ATL and a median PVL 45% (range 0-154%). We therefore suggest that the proviruses in these samples contain a proviral genomic deletion or mutation in the overlapping primer binding sites regions of these two PCR reactions (Figure 2.1) between exon 2 and exon 3 of *tax* (Seiki nucleotides 6483-6570). Consistent with this hypothesis is that the coding region of *HBZ*, which is known to be expressed in ATL cells, would be unaffected by genomic deletions in

this region. The type 1, type 2, multiple and indeterminate defective subgroups are considered collectively as 'defective proviruses' for further downstream analysis.

4.3.4. Tax gene mutations are detected in 7% of the ATL cohort

The coding regions of the *tax* gene were amplified and sequenced (exon 2 and 3) in the 120 cases identified to contain complete proviruses. The methionine start codon is derived from the nucleotide sequence in exon 2, whilst the remainder of the protein is coded from the nucleotide sequence of exon 3 (amino acid 2-353). Figure 4.1 and table 4.3 summarise the following results:-

We amplified exon 2 by nested PCR using a high fidelity polymerase in 120 cases with complete proviruses and identified the normal start codon in 118 cases. In two cases, there were non-synonymous mutations in which the amino acid methionine was changed to isoleucine. Since these mutations removed the methionine start signal, we infer that these two tumour samples were unable to express Tax protein.

We sequenced exon 3 of *tax* in 120 cases and observed 97 distinct nucleotide alterations from 79 cases (reference genome NCBI GenBank J02029, (Seiki et al., 1983). These 97 alterations resulted in 48 distinct amino acid changes. Three of these amino acid changes were nonsense mutations at Tax amino acid positions 28, 56 and 248 and occurred in 11 cases. Additionally, in one sample there was a deletion of 249 base pairs (nucleotide positions 7434-7683).

In 24 samples, 4 specific nucleotide substitutions were observed which are known to represent the Tax A subtype and are strongly associated with the proviral subgroup Cosmopolitan A based upon LTR sequence (Furukawa et al., 2000) : C>T at position 7897, C>T at position 7959, G>A at position

8208 and A>C/G at position 8344. In addition, nucleotide alterations at positions 7720, 8120 and 8297 which have been observed as recurrent mutations within the Tax A subtype (Furukawa et al., 2000) were observed in 4/24 Tax A subtype samples and are known not to be ATL disease specific (Furukawa et al., 2000).

Two nucleotide changes have been reported to be specific to ATL and were observed in our cohort: A nucleotide change from G to A at position 7464 is a known hotspot of mutation and creates a premature stop codon at position 56 of the Tax protein and has been shown to be associated with loss of Tax protein function (Furukawa et al., 2001). We observed this mutation in 7 out of 120 samples (120 samples containing non-defective proviruses). One of these cases also carried the previously described missense mutation in the start codon of exon 2. The second reported ATL specific mutation is a missense mutation A>G at position 7337 resulting in an amino acid change from alanine to glycine at Tax position 14. The functional significance of this missense mutation has only been identified in the context of HLA-A*02 where it has been demonstrated to escape Tax-specific CTL killing (Furukawa et al., 2001). We observed this mutation in 3 cases, but in the absence of HLA-typing its precise significance cannot be assessed.

Two recurrent novel nonsense mutations were observed: Nucleotide substitutions of G>A at position 7380 occurred in 2 samples and results in a premature stop codon at Tax position 28. Nucleotide substitutions of G>A at position 8040 occurred in 3 samples and results in a premature stop codon at Tax position 248. Tax forms homodimers that contribute to the transcriptional activity from the HTLV-1 promoter (Jin and Jeang, 1997; Tie et al., 1996; Basbous et al., 2003) and Basbous et al have shown that the integrity of the Tax sequence is critical for Tax dimerization except for the last 16 residues (amino acids 338-353), and so we conclude that these two additional nonsense mutations will result in loss of Tax protein function (Figure 4.1). No splice site mutations were observed.

Additional missense mutations were found in 42 other amino acid positions, but, in the absence of HLA-class I genotype and functional studies we cannot infer that Tax function is diminished by the identified missense mutations. A summary of *tax* mutations is shown in Table 4.3 and Figure 4.1.

To summarise: Whilst missense mutations were frequently observed we have not inferred any functional consequence of these since we do not know the HLA type of each case. Known ATL specific nonsense mutations were seen in 7 cases with additional novel nonsense mutations observed at both Tax position 28 and Tax position 248 which we suggest would be critical to Tax expression.

Table 4-3: Summary table of amino acid alterations following tax sequencing

	Methionine start codon mutation only	**Nonsense mutations**	**Large Tax deletion**
Number of cases	1*	11	1

*Two cases with methionine start codon mutations were identified, but one case had an additional nonsense mutation in exon3 and has been counted there.

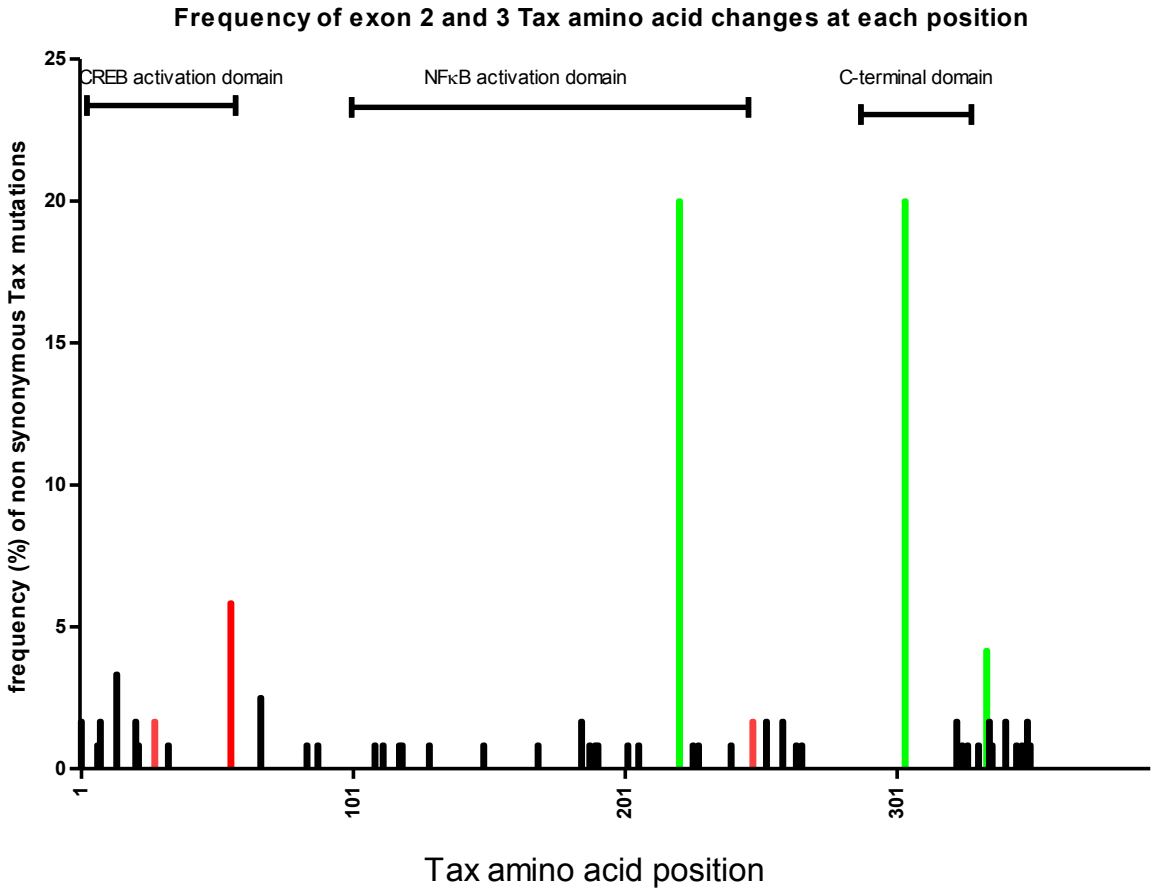


Figure 4.1: Frequency of predicted amino acid sequence changes in exon 2 and exon 3 of the tax gene

All non-synonymous alterations are shown with bars. Green bars represent amino acid changes characteristic of the Tax A proviral genotype (A>V at Tax 221, S>N at Tax 304, Q>R at Tax 334); red bars represent premature stop codons caused by nonsense mutations. The location of the functional CREB activation domains (amino acids 1-56), NFκB activation domain (amino acids 112-246) and C-terminal domains (amino acid 289-323) are highlighted.

4.3.5. Hypermethylated 5'LTR TRE causes tax silencing in approximately 8% of ATL cohort

Following *tax* gene amplification and sequencing, MS PCR was undertaken on 107 cases – those cases without a defective provirus or a critical *tax* gene alteration. The aim was to identify whether the *tax* promoter in the U3 region of the 5'LTR was silenced by methylation.

The samples were categorised as methylated, unmethylated or partially methylated.

Table 4-4: Summary table of methylation status of the 5'LTR (n=113)

	Fully methylated 5'LTR	Fully unmethylated 5'LTR	Partially Methylated 5'LTR	Not determined
Number of cases	16	24	64	3

Sixteen of the 107 cases (15%) of the samples contained a hypermethylated 5'LTR. This represents 8.1% of the cohort (16/197). Takeda et al, who first reported and demonstrated the utility of MS PCR and BSP techniques in DNA samples taken from ATL patients, identified hypermethylation in 11% ATL cases (14/41 samples).

As proof of principle, I confirmed hypermethylated 5'LTR MSP findings in 4 cases by integration site specific bisulfite PCR (BSP) followed by TOPO® cloning and sequencing of the PCR products. By sequencing between 10 and 20 colonies per sample, I demonstrated that in all cases the critical CpG sites covering the first promoter-proximal Tax responsive elements (TRE) were fully methylated. The second TRE was fully methylated in all cases and third TRE was fully methylated in 3/4 cases and 82% methylated (45/55 CpG methylated) in the fourth case.

Table 4-5:Summary of Tax silencing mechanisms for cohort

	Complete provirus (not defective, mutated or hypermethylated)	Defective provirus	Critical *tax* gene mutation	Hypermethylated TRE
Number of cases	90	78	13	16
% of cases	46%	39%	7%	8%

4.3.6. No difference in the median proviral load by clinical or proviral subtype

The proviral load (PVL) was measured by QPCR in all ATL and AC samples by calculating the ratio of *tax* copies to *actin*. The aim of measuring the PVL was to ensure that the ATL and control AC cohorts were similar to previously published cohorts, and for subsequent calculation of absolute abundance of clones following LMPCR and high throughput sequencing.

In the ATL group the median PVL measured 71.7% (range 0.00% - 700%). The median PVL for each clinical subgroup and each Tax-silenced subgroup is shown in Table 4.6 and Figure 4.2. The median proviral load of the defective proviruses was significantly lower than the complete proviruses (Wilcoxon test p=0.004) which could be explained by the defective cases containing deletions of the *tax* gene primer binding sites. The median PVL of ACs measured 1.8% (range 0.14 - 18.4%).

Table 4-6: Summary table of PVL and Tax status

Clinical Subgroup or tax subgroup	Median PVL (%)
AC	1.8
Acute ATL	64.7
Chronic ATL	62.8
Lymphoma ATL	43.8
Smoldering ATL	12.9
Complete provirus	79.4
Defective provirus	65.6
Tax mutated provirus	68.3
Hypermethylated 5'LTR	83.7

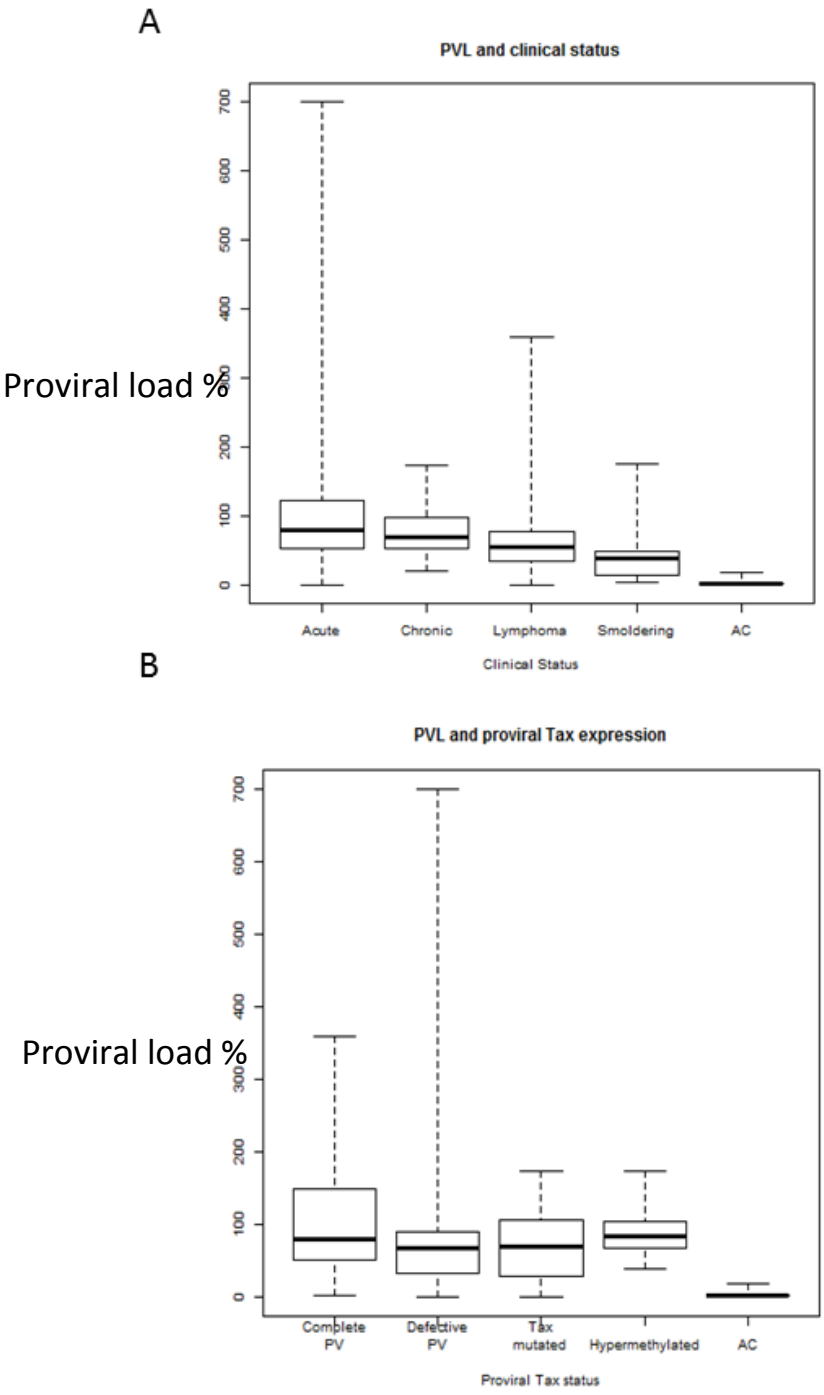


Figure 4.2: Proviral load (PVL) by clinical subtype (Panel A) and proviral *tax* gene status (Panel B)

PVL were measured by QPCR and quantified by measuring *tax* and *actin* gene copies. Box-and-whisker plot show PVL for all cases. The box displays the 25th, 50th and 75th centile whilst the whiskers show results that lie in the top and bottom quartiles. x-axis; ATL cohort split by disease status (Panel A) or *tax* gene status (Panel B). y-axis; PVL % (number of copies of *tax* per 100 PBMC).

4.3.7. No difference in the median oligoclonality index by clinical or proviral subtype

The oligoclonality index (OCI) was calculated for each sample (Methods section 2.4.3) to test the hypothesis that there was no difference in the clonal distribution between tumour samples that were capable of Tax expression and those that silenced Tax (Figure 4.3).

The median OCI of the ATL clinical subgroups was 0.91 (range 0.47-1.0) and there was no difference between ATL clinical subtype or between proviral *tax* subtypes. The median OCI of the ACs was 0.33 (0.14- 0.87).

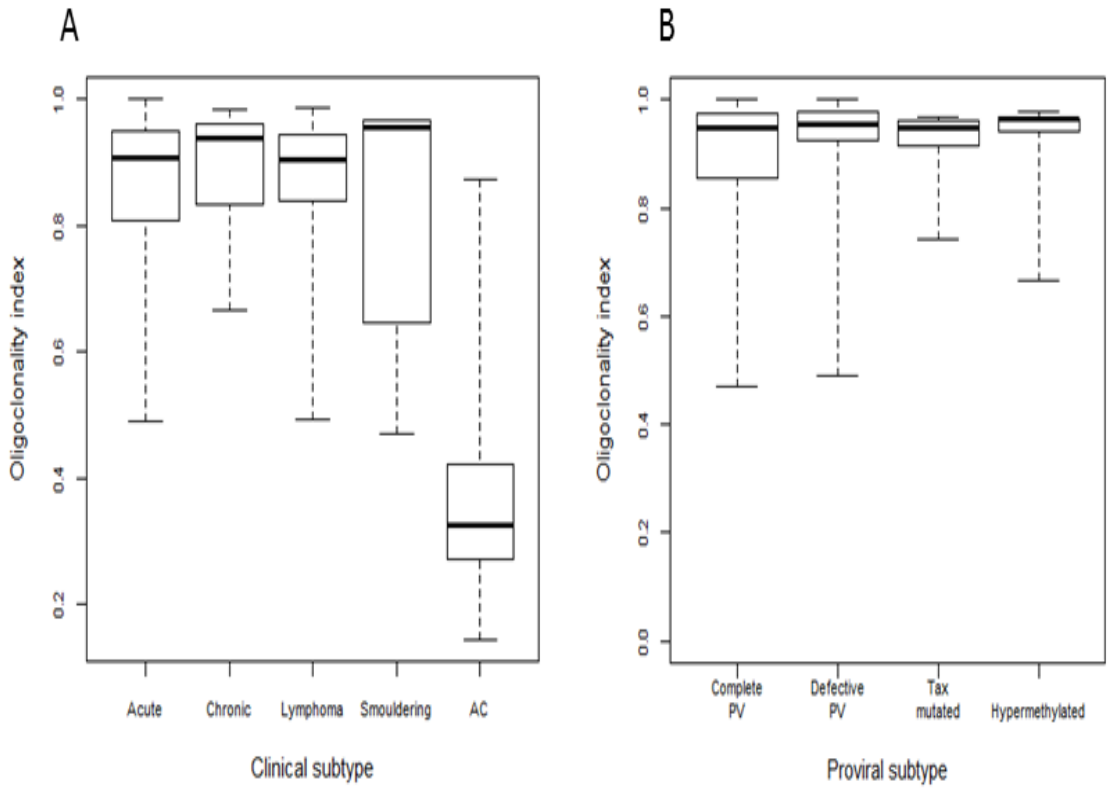


Figure 4.3: Oligoclonality index by clinical and proviral subtype

The OCI for the ATL and AC cohort split by clinical subtype (Panel A) shows median OCI of the ACs was 0.33 (range 0.14-0.87) and the median OCI for the ATL subtypes was 0.91 (range 0.47-1.0). There is no difference in OCI between ATL clinical subtypes. There was no difference in OCI by mechanism of proviral silencing (Panel B).

4.3.8. Definition and character of abundance bins

The ATL integration sites were subdivided into three abundance bins designated respectively as 'small', 'intermediate' and 'large'. Since malignant behaviour is defined by growth in vivo and not from any specific genomic characteristic, it remains an assumption, widely held, that the large clones that are observed in ATL are malignant.

Since the vast majority of total integration sites are small, with a relative abundance of <1%, we have called these 'small clones'. We have defined the large clones, presumed malignant in the ATL cases, as those integration sites with a relative abundance greater than or equal to 35% of the PVL. This threshold was determined after plotting the relative abundance of the largest integration site in each sample and observing a distribution in which all of the ATL samples have a dominant integration site that occupies >35% PVL (Figure 4.4). The intermediate sized clones are defined as those between the small and large clones with a relative abundance of greater than or equal to 1% but less than 35% (Table 4.7). The abundance bins are defined in the glossary for quick reference.

The relative abundance (rather than the absolute abundance) was determined to be the best measure for subgrouping of integration sites for analysis since the proviral load, which is required for the mathematical estimation of absolute abundance (Method section 2.4.5), varies significantly in the ATL samples due to, for example, *tax* gene deletions which would significantly underestimate the proviral load whilst the presence of multiple proviruses would overestimate the absolute abundance of the malignant clone within a cell population.

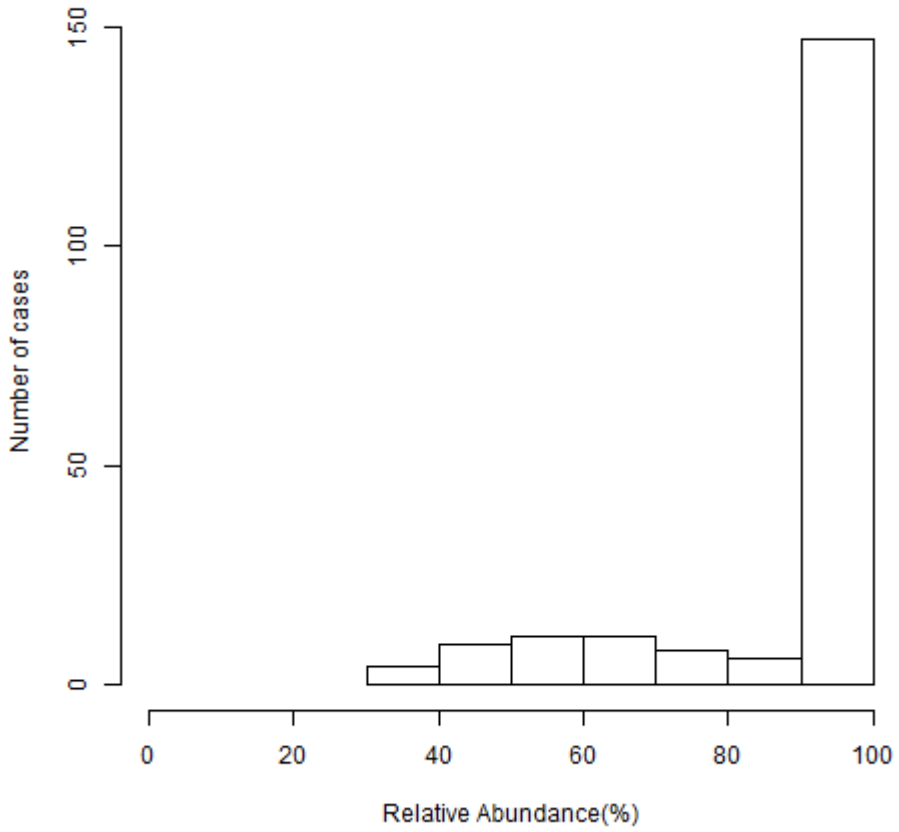


Figure 4.4: Relative abundance of the largest clone in each ATL patient is >35%

The relative abundance of the largest clone was plotted for each ATL case (relative abundance defined as the proportion of the proviral load occupied by an individual clone). The distribution of these integration sites suggested that the large presumed malignant clones typically occupied >35% relative abundance.

Table 4-7: The number of integration sites within each abundance bin

Clinical status	Number of samples	Small	Intermediate	Large	Total UIS
AC	75	16001	904	4	16909
ATL	197	5925	90	217	6232

4.3.9. Identification of multiple proviruses within a dominant clone

ATL is traditionally associated with monoclonal integration of an HTLV-1 provirus within the tumour cells, although there have been reports for many years of the presence of multiple proviruses, based upon the observation of multiple bands by Southern blot (Kamihira et al., 2005; Tamiya et al., 1996; Tsukasaki et al., 1997). Examples of the 3 ATL clonal distributions identified are shown in Figure 4.5.

Here, I found that in 157 cases (80% cases) a single dominant clone was present which occupies a median relative abundance of 99.4% (range 35-100%) of the proviral load (Figure 4.5, A). In 40 cases, monoclonal integration was less certain since more than one abundant integration site was observed. In each case, there was a one integration site with a relative abundance >35% (i.e. falls into the 'large ATL' abundance bin) but additionally a second site with a minimum relative abundance >10%. In many of these cases the two dominant sites were of near equal relative abundance i.e. 45% and 50% relative abundance, but in other cases this was less clear e.g. 65% contribution from one site and 30% for the second site (Figure 4.5 B,C). Therefore the question arises as to whether these represent one malignant clone containing 2 proviruses or if there are two distinct abnormally expanded clones. If a single malignant clone contains two proviruses, the cells should contain an equal abundance of those integration sites (50% contribution from each integration site), assuming a steady kinetic state and no recent re-infection with a second provirus. It is unlikely that a superinfected malignant clone would co-exist with its singly infected parent clone, since a second infection would confer either a proliferative advantage or a negative survival cost. Therefore, we assume that a single clone containing two proviruses would be expected to have near equal abundance of both integration sites and express a single TCR gene rearrangement. Alternatively, if there are two large independent clones proliferating, we would expect there to be a

measurable difference in the abundance of the detected two integration sites and two distinct TCR gene rearrangements.

To test the hypothesis that two observed HTLV-1 integration sites are present in one T-cell clone we used the Gaussian approximation to the binomial distribution, outlined below. The null hypothesis was that there is no difference in the number of unique shear sites (read 2) between the two integration sites.

Z-score calculated as:

$$z = \frac{|a-b| - 0.5}{(2ab/N)^{0.5}}$$ with correction for continuity.

Where,

a= Number shear sites for integration site 1

b= Number shear sites for integration site 2

N=Total number of shear sites (N=a+b)

The Z-score was converted into a 2 tailed p-value.

In 18/40 cases (9.1% of the total cohort) the null hypothesis was accepted, that is to say there was no difference in the number of shear sites between the abundant clones, suggesting multiple proviruses present in each T-cell clone. T-cell receptor gene rearrangement studies of DNA were

undertaken on these 40 cases and a single TCR gene rearrangement was detected in 7 of the 18 cases in the cohort containing two equally abundant proviruses (3.6% cohort).

Of the 18 cases in the cohort containing two proviruses per 'malignant' clone, 72% were of acute subtype, 11% chronic, 11% smouldering and 6% lymphoma (proportions not statistically different from representation within cohort).

In 22/40 cases, the null hypothesis was rejected. That is to say there was a significant difference in the abundance of the two dominant integration sites. In these cases there was typically a dominant site (occupying 40-80% PVL) with additional intermediate sized expanded clones (in these cases occupying 10-35% PVL).

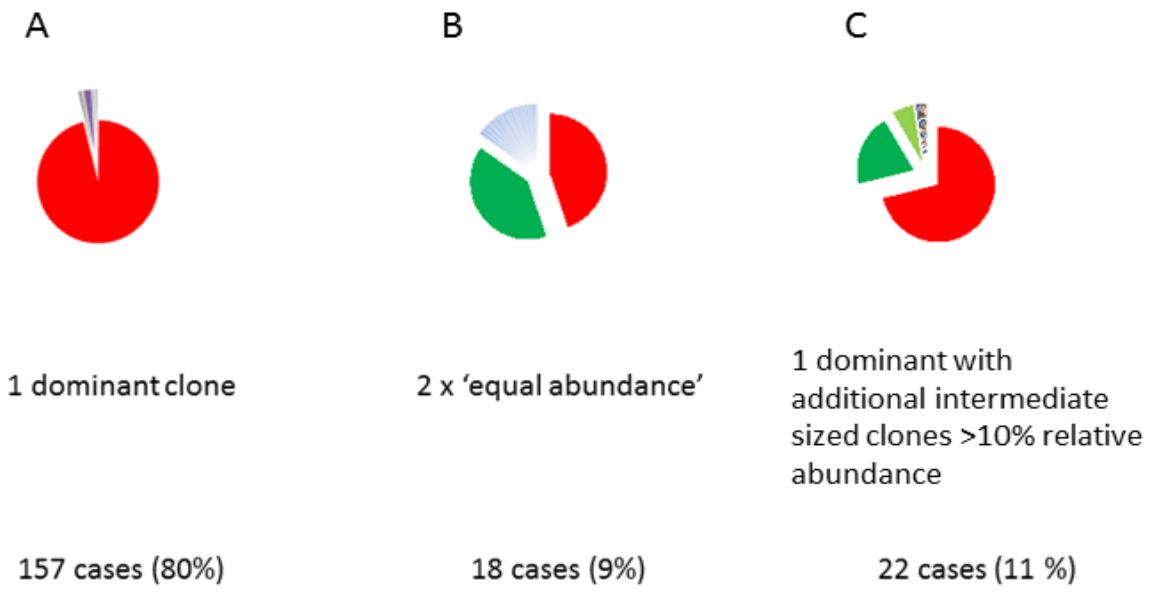


Figure 4.5: Clonal structure of the ATL cases

Examples from the typical clonal structures of ATL cases is shown in pie charts, where each sector depicts the relative abundance of the respective detected integration site following LMPCR and high throughput sequencing. (A) A typical ATL monoclonal tumour sample, PVL 107% (relative abundance dominant clone 97%). (B) Multiple copies per cell; two equally dominant integration sites, PVL 200.7%, (relative abundance 43% and 40% PVL). (C) ATL with dominant clone and additional intermediate sized clones, PVL 241% (relative abundance 67% and 23% PVL).

4.4. Discussion

I characterised the proviral structure in 197 cases of ATL and determined that our cohort was similar to previously published ATL cohorts. That is to say, 46% samples were inferred to be capable of Tax protein expression, and I observed the expected frequencies of defective proviruses, *tax* gene nonsense mutations and *tax* silencing by hypermethylation of the 5'LTR. Importantly, this similarity to previous data gives confidence that any observations or conclusions that are made in subsequent host genomic integration site analysis are representative. Furthermore, I made some novel observations during the characterisation of this cohort, as summarised below.

4.4.1. Identification of a novel category of defective provirus

The frequency of defective proviruses (42%) is consistent with the published literature (Kamihira et al., 2005; Tamiya et al., 1996). Here, I identified a category of defective provirus which I have termed 'indeterminate'. These are samples in which the proviruses fail to amplify by conventional long range PCR used for characterising defective proviruses, but there is evidence of HTLV-1 integration e.g. a monoclonal population of cells seen following HTS, suggesting at least the presence of the 3'LTR.

I hypothesise that the cause of the indeterminate provirus is a large deletion or mutations involving the primer binding regions of the long-range PCR. It has been reported that HBZ is detected in all HTLV-1 infected clinical states, including ATL (Usui et al., 2008). Although HBZ expression has not

been directly measured in these samples, the primer binding regions overlap with the intronic region of spliced HBZ, suggesting that a deletion in this region may not affect HBZ expression.

On the basis of these results there is a project underway in the local Molecular Diagnostic Unit to compare *HBZ* and *tax* qPCR results in ATL cases with an unusually low proviral load (<5%-10% PBMC) to identify if *HBZ* may be a more reliable measure of PVL to monitor certain ATL cases.

4.4.2. Identification of novel nonsense mutations in the *tax* gene

I found nonsense mutations that prohibit the expression of functional Tax protein expression in 13 cases (7 % cohort) which is consistent with the published literature (Takeda et al., 2004). There were two novel recurrent nonsense mutations that have not been previously reported (G>A at position 7380 and G>A at position 8040), which, based upon the amino acid position, I hypothesise will prevent functional Tax protein expression. A test of this hypothesis requires formal validation by measurement of *tax* mRNA and protein expression, although there were no cells available from these individuals to do so. Fan et al have described G-to-A base substitutions as the most frequent mutations in ATL cells, and responsible for nonsense mutations, which is in accordance with the target sequence of human APOBEC3G (Fan et al., 2010), the mammalian host defence against retroviruses. The novel nonsense mutations described here occur in the context of a TGG nucleotide sequence which is the target of APOBEC3G and results in TAG or TGA sequences and a stop codon. Since HTLV-1 proliferates mitotically, the provirus can proliferate with nonsense mutations providing it retains a minimum set of viral genes. Fan et al showed that the only viral gene that does not become mutated is *HBZ*, which is likely indispensable for ATL development (Fan et al., 2010). From these observations, together with the observation that APOBEC3G generates stop codons during

reverse transcription and that type 2 defective proviruses may also occur pre integration (Miyazaki et al., 2007), I conclude that at least in some ATL cases, Tax expression is not essential for immortalisation and expansion of malignant clones.

4.4.3. Wide variation in the PVL in ATL samples

The median proviral load of the ATL cases was 68.3% (range 0.0 -700%) and in the ACs 1.8% (range 0-18.5%). These results are comparable to data published on Japanese patients where a recent study on 1218 asymptomatic carriers identified a median PVL of 1.39% in females and 2.10% in males (Iwanaga et al., 2010). A recent study of proviral load measurements of ATL patients managed in London reported a median PVL of 50.3% in acute ATL, 7.9% in the peripheral blood of ATL lymphoma and 81.25% in the lymph node biopsy of lymphoma-type ATL (i.e. a 5:1 ratio of load in lymph node to peripheral blood) (Demontis et al., 2013).

In acute ATL, the PVL is not normally expected to exceed 100% (1 proviral copy in every PBMC), although a proviral load > 200% can result from the presence of multiple proviruses, or possibly due to genomic instability resulting in loss of an actin copy. PVL measurements by QPCR are known to fluctuate, although reproducibility is better at higher proviral loads (Demontis et al., 2013). Within an asymptomatic individual the proviral load remains constant over many years and a log-fold change in proviral load measurements is usually considered to be clinically relevant, although there may be little significance between proviral load measurements of 100% and 200%.

An interesting observation is that the PVL is frequently not 100%, even in cases where examination of a peripheral blood film suggests an entirely leukemic picture and heavy burden of disease. This raises the possibility that there is a polyclonal expansion of uninfected PBMCs. These polyclonally

expanded PBMCs are unlikely to be CD8+ T-cells since diagnostic flow cytometry of ATL cases typically shows predominantly CD4+CD25+ T-cells and the absolute number of CD8+ T cells is between 50% and 100% of the normal range (Arnulf et al., 2004; Rowan and Bangham, 2012; Shimizu et al., 2009). HTLV-1 infected CD4+ T-cells produce CCL22 (Toulza et al., 2010), one of the two ligands for the receptor CCR4 which is expressed on both tumour cells and on the non-malignant CD4+ FoxP3+ cells. Toulza et al showed that the level of CCL22 produced is sufficient to attract CD4+FoxP3+ cells and to enhance their viability, which may contribute to the persistence of HTLV-1 by suppressing the HTLV-1 specific CTL response (Toulza et al., 2010). It is perhaps the expansion of uninfected FoxP3+ Tregs that contributes to the expanded CD4+ compartment. Furthermore, polyclonal expansions of Treg cells have been reported in other haematological malignancies: For example, in Hodgkin's lymphoma the characteristic Reed-Sternberg cells constitute only a small proportion of the tumour, whilst infiltrating lymphocytes are highly enriched for Tregs, attracted by TARC and CCL22 (Marshall et al., 2004) and in B-CLL the presence of activated CD4+ T cells is required for tumour proliferation (Devereux, 2011; Patten et al., 2008). The question remains as to whether these reflect anti-tumour responses or are a bystander effect.

In HTLV-1 infection, separation of the infected and uninfected populations of T-cells by cell-sorting followed by high-throughput analysis of TCR gene rearrangements (e.g. ImmunoSeq™) would make it possible to quantify the immune repertoire of the infected and non-infected T-cell populations in different disease states.

4.4.4. Evidence of multiple proviruses in ATL cases.

I obtained evidence by HTS of 2 equally abundant proviruses within a single tumour sample in 18 cases (9% cohort) and used TCRG gene rearrangement studies on DNA to quantify the number of cases with a single TCR gene rearrangement. I found a single TCR gene rearrangement in 7/18 cases (3.6% of the whole cohort).

TCR gene rearrangement studies are complex to interpret. One of the major issues in the accurate identification of clonality is the occurrence of multiple clonal PCR products. Thymocytes undergo a hierarchical rearrangement in their TCR loci starting with D-D, D-J and V-DJ rearrangements in the TCRD locus, followed by V-J rearrangements in the TCRG locus (Langerak, 2012). Random nucleotides are often inserted into the VDJ genes but since an antigen receptor chain can only be formed by preserved triplet codons, many rearrangements are non-productive. For a given TCR locus, very frequently two rearrangements have taken place, one from each allele. The theoretical chance for a rearrangement to be in correct reading frame and lacking a premature stop codon is estimated between 20% and 30% (Langerak, 2012), implying that diallelic TCR gene rearrangements are more the rule than the exception. This means that monoclonal tumours will mostly contain diallelic rearrangements. This makes distinguishing a diclonal tumour from a monoclonal tumour with diallelic rearrangements difficult. The only definitive means of separating these two possibilities is by quantifying TCR gene expression using mRNA. However, within this cohort there were no available cells for RNA extraction.

On the basis of these data, I would therefore report the incidence of multiple proviruses within a single malignant clone at between 3 and 9 %. The presence of multiple proviruses in ATL patients has been widely reported although the incidence has not been previously defined. The mechanism by which the malignant clones have become superinfected is not known and raises the question as

to whether the mechanism by which the cells have become superinfected is a cause or a consequence of ATL.

4.4.5. The significance of intermediate sized clonal populations is uncertain

We have identified that whilst the majority of ATL tumour cell populations are monoclonal (containing either one or two proviruses), other tumours contain a dominant clone but with additional intermediate sized clones (figure 4.4, panel C). This raises several questions as to whether we are correct to assume that the most abundant clone is always malignant, whether the smaller clones contribute to the pathogenesis of ATL, and whether these intermediate sized clones contribute to clonal succession or chemotherapy resistance.

Malignancy is defined by the abnormal or uncontrolled growth of a cell population in vivo, rather than any specific genetic or epigenetic characteristic. Characteristics that must be acquired for a lesion to develop into cancer which include self-sufficiency in growth signals, insensitivity to anti-growth signals, limitless replicative potential, evasion of apoptosis, sustained angiogenesis, tissue invasion and metastatic spread (Hanahan and Weinberg, 2000, 2011). Here we have assumed that where there is a single dominant integration site (median relative abundance 99.2%, range 35.5-100%), this is the malignant clone. Where there are two proviruses within each clone, we do not know at what point during clonal expansion the superinfection arose and whether there was originally just a single provirus within the genome during malignant transformation, which later became superinfected; perhaps one of those proviruses is more important than the other in tumour pathogenesis? Since we cannot distinguish which of these integration sites might be more important, we consider them to share equal importance.

The significance of the intermediate sized clones is even less certain: There are a handful of studies describing the functional properties of CTL in ATL, which collectively report that the CTL response in ATL is reduced in frequency, or absent, and when present is functionally deficient with Tax-specific CTLs directed at a narrow range of epitopes (Arnulf et al., 2004; Kozako et al., 2006; Kozako et al., 2009; Shimizu et al., 2009). Perhaps, once transformation has occurred, the reduced or inefficient CTL response enables other clones to expand; these proliferative clones might have been formerly restricted by CTL killing. The host genomic landscape of these different clonal populations is further characterised in Chapter 5.

Chapter 5. Genomic landscape of HTLV-1 proviral integration sites

5.1. Chapter Abstract and summary

The aim of the work described in this chapter is to test the hypothesis that ATL malignant clones arise from the typical low abundance clones seen in asymptomatic carriers (rather than the higher abundance clones observed in some AC or HAMTSP patient samples) and to characterise and identify host genomic features that are unique to defined clonal populations seen within ATL samples and which may suggest an 'integration site determined predisposition' to ATL development. Furthermore, I aim to correlate host genomic characteristics with capacity for proviral Tax expression, in order to postulate mechanisms for aberrant clonal expansion. Where possible I will correlate findings with ATL clinical subtype.

The results described here are derived from a bioinformatic analysis of the high-throughput sequencing of ATL and AC samples compared with in silico (random) integration site datasets and, where indicated, from *in vitro* infection data. In summary the key findings of this chapter are:-

- AC clones, small ATL clones and large ATL clones show similar genomic characteristics with regards to preference for integration in proximity to transcriptional start sites (TSS), CpG islands, genes, same-sense relative transcriptional orientation and activatory epigenetic marks.
- Intermediate-sized ATL clones are abnormally abundant and show distinct genomic characteristics, not previously found in non-malignant HTLV-1 infection (AC or HAM/TSP).
- ACs and small ATL clones show a bias of integration into certain chromosomes (acrocentric)
- The ATL large clones (presumed malignant) do not show any recurrent sites ('hotspots') of genomic integration
- Gene ontology analysis of the nearest downstream TSS identifies recurrent cellular functions/pathways associated with the large ATL clones (5.6% of all ATL cases).

5.2. Introduction

5.2.1. Proviral integration site bias

The genomic site of retroviral integration into the host genome is important for both virus replication and pathogenesis of disease, since it may affect both viral gene expression (Shan et al., 2011) and host gene expression (Siliciano and Greene, 2011).

Some retrovirus-induced tumours are caused by insertional mutagenesis (Neel et al., 1981; Payne et al., 1982), whereby integration into a specific locus introduces a viral promoter or enhancer that *in cis* may activate an adjacent host cellular oncogene or, inactivate a gene through disruption of induction, expression or splicing (Uren et al., 2005). Seiki et al showed the absence of a common region of provirus integration in ATL leukaemia cells, which implied that insertional mutagenesis, was unlikely to be a mechanism of ATL leukaemogenesis (Seiki et al., 1984).

Interest in proviral integration sites has re-emerged in the last 10 years, in part due to improvements in the laboratory techniques which identify integration sites, and following publication of the annotated human genome. Additionally, interest has followed the high incidence of cases of T-acute lymphoblastic leukaemia (T-ALL) that arose in children treated with gene therapy for severe combined immunodeficiency (SCID), which was caused by insertional mutagenesis of the retroviral vector near the LM02, BMI1 and CCND2 proto-oncogenes (Hacein-Bey-Abina et al., 2008; Hacein-Bey-Abina et al., 2003). The risk of insertional mutagenesis has halted the use of gammaretroviruses as suitable vectors for gene therapy and there are active efforts to identify and validate 'safe harbours' of the genome - regions of the genome that allow predictable expression of the newly integrated DNA whilst minimising unwanted interactions between inserted genes and the

neighbouring host genome. Suzuki et al developed a definition for retroviral integration in cancer cells to discover potential cancer-related genes and defined a hot-spot as ≥ 4 integration sites within a 100 kb region, 3 integrations within 50 kb or 2 within 30 kb (Suzuki et al., 2002), and similar definitions have been adopted by others (Schroder et al., 2002,Cattoglio et al., 2007).

Retroviral integration sites can be found across the host genome, although they are not randomly distributed (Wu et al., 2005). Integration site selection differs between retroviruses but broadly all display a preference for integration at 3 distinct levels. First, integration is biased towards open conformation euchromatin which allows the retroviral preintegration complex access to the DNA. Second, the nucleotide sequence in the immediate vicinity of integration sites is typically a loose palindromic consensus sequence (Chou et al., 1996; Derse et al., 2007; Meekings et al., 2008) which is consistent with the two-fold symmetry of the retroviral integrase (Hare et al., 2010; Maertens et al., 2010). Third, there is an additional bias based upon interactions between the preintegration complex and specific cellular host factors, e.g. LEDGF/p75 in HIV infection which directs the preintegration complex into genes and away from intergenic regions. Similarly, the transcription factor YY1 plays a role in murine leukaemia virus (MLV) guiding integration. However, in most retroviral infections including HTLV-1 similar host co-factors have not been identified.

5.2.2. HTLV-1 integration site selection

Work undertaken within this laboratory over the last few years has focused on identifying the integration sites favoured in initial targeting of infection by using a model of short-term in vitro infection of human T-lymphocytes, and comparing the integration sites that are selected to survive in chronic infection in samples taken from patients with different clinical manifestations of HTLV-1 infection.

Gillet demonstrated the predominance of integration in transcriptionally active regions and reported a positive correlation between clonal abundance and proximity to genes, CpG islands, activatory epigenetic marks and same sense relative orientation (Gillet et al., 2011).

In addition Melamed et al have recently reported a remarkably strong bias toward integration ex vivo within 100 base-pairs of certain transcription factor binding sites, for example the tumour suppressor p53 and the transcriptional regulator of interferons STAT1, and in each case the integrated provirus was between 100 and 350-fold more likely to lie within 100 base-pairs of the respective binding site than by chance (Melamed et al., 2013). Flow cytometric sorting of cells into Tax-positive and Tax-negative cells followed by integration site analysis revealed that certain TFBS (e.g. STAT1) lying 10 or 100 base pairs upstream of the provirus were associated with spontaneous Tax expression, whilst at a similar distance downstream there was no effect. Conversely, the presence of BRG1, an ATPase that powers the chromatin-remodelling complex SWI/SNF, 10 to 100 base-pairs upstream was associated with the silencing of Tax whilst a BRG1 site at an equal distance downstream of the provirus was associated with spontaneous Tax expression. This asymmetry implies a mechanistic interaction occurs between transcription of the provirus and transcription of the host genome. This conclusion is reinforced by further observations that the relative orientation of the provirus and nearest host gene associate with frequency of spontaneous proviral expression; a same-sense host transcriptional start site upstream appears to suppress Tax expression whilst a same-sense host transcriptional start site downstream of the provirus associates with spontaneous Tax expression. These findings strongly suggest that the dominant interaction between the flanking host genome and the provirus is transcriptional interference; a nearby host promoter upstream of the provirus is able to silence the downstream 5'LTR HTLV-1 provirus (Melamed et al., 2013, reviewed Bangham C.R.M, 2013).

5.2.3. Determinants of clonal abundance in vivo

For many years it has been widely believed that oligoclonal expansion of HTLV-1 infected T-cells is responsible for both viral persistence of the infection and also maintenance of the proviral load which predisposes to both inflammatory and malignant disease.

With the development of HTS and the ability to quantify accurately clonal abundance, a strong determinant of clonal abundance was identified as proviral orientation: Proviruses that lie in the same sense transcriptional orientation as its nearest host gene or transcriptional start site (Gillet et al., 2011). Additionally, the recently published work by Melamed et al which flow cytometrically sorted Tax-positive and Tax-negative cell populations identified that small (low abundance) clones are more likely to express Tax *in vitro* than larger clones.

The understanding of clonality in ATL lags behind that of non-malignant HTLV-1 infection. For a variety of reasons, including the relative rarity of ATL and due to the labour intensive nature of the techniques required, the number of ATL cases that have been systematically analysed is relatively small: 23 ATL cases by Hanai (Hanai et al., 2004), 59 cases by Doi (Doi et al., 2005) and 33 cases by Ozawa (Ozawa et al., 2004). In addition to the limitations already described, integration site analysis in relatively few cases made correlation by clinical phenotype extremely difficult.

Whilst it is widely assumed that ATL is a monoclonal disease there are indications that clonality is more complex. Firstly, there are many HTLV-1 infected clones in addition to the large, presumed malignant clone which are not likely to be detected by classical sequencing methods and the identity and role of these clones in supporting the malignant clone is unknown. Secondly, it has long been observed that there may be more than one pathologically expanded provirus present. It has been previously untested as to whether these represent multiple malignant clones or superinfection of

one malignant clone with multiple proviruses. The possibility of more than one independently transformed malignant clone would be in keeping with the emerging understanding of the polyclonal origin of some solid tumours (reviewed Vogelstein et al., 2013). Third, the malignant clone does not necessarily develop from a large pre-existing infected T-cell clone, but may arise rapidly de novo from a clone of low abundance- Figure 5.1 demonstrates the emergence of a presumed malignant clone from a small clone detected 18 months before disease transformation and not from the large clone that was present when the patient was asymptomatic. Fourth, there have been case reports in the literature and anecdotally of 'clonal succession', in which the presumed malignant clone at a given time point may be replaced by a different clone at a later time point (e.g. at clinical relapse) and scenarios in which there are distinct dominant clones, both presumed malignant, present at the same time but in different compartments e.g. blood and lymph node (Tsukasaki et al., 1997).

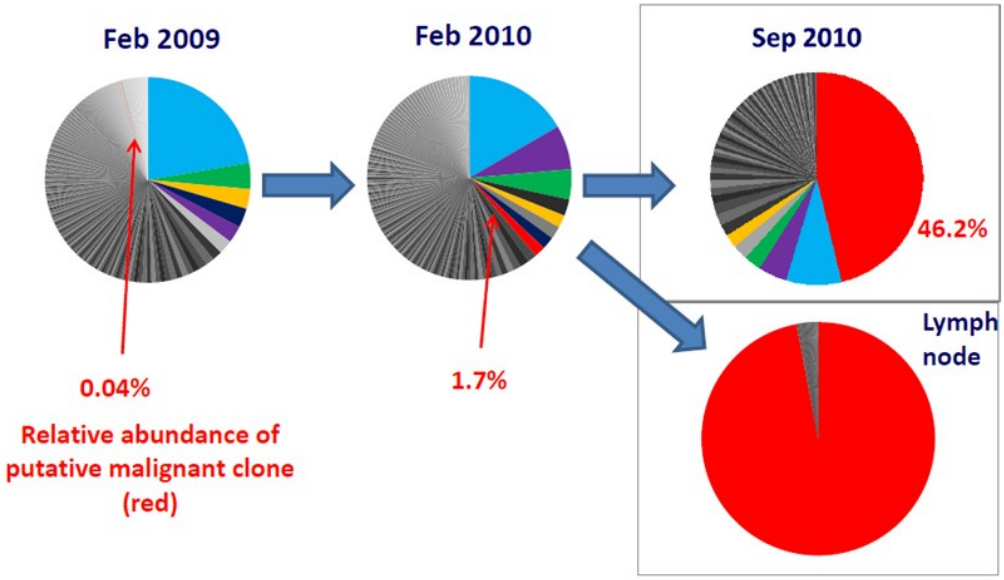


Figure 5.1: Clonal evolution in ATL

The pie charts represent the clonal distribution seen in a single patient at 3 different time points prior to and at the development of acute ATL. Each slice of the chart represents the relative abundance of an individual clone. The patient developed ATL in September 2010 and the clonal distribution in PBMC (top right) and lymph node (bottom right) on the day of diagnosis is illustrated. The presumed malignant clone is coloured red and when tracked back to February 2010 (7 months prior to diagnosis) represents 1.7% of the proviral load and in February 2009 (19 months prior to ATL diagnosis) is barely visible representing 0.04% of the proviral load.

5.3. Results

5.3.1. Definitions of the genomic environment flanking the host genome

The definitions of 'upstream', 'downstream', 'same-sense' and 'opposite sense' are all made with respect to the forward (plus/positive) strand of the proviral open reading frames. Therefore 'upstream' of the integrated provirus denotes genomic annotations closest to the 5' end of the provirus, whilst 'downstream' denotes genomic annotations in closer to proximity to the 3' end of the provirus. For each genomic feature analysed e.g. proximity to a gene, we defined the minimal distance to the integration site as the difference between the genomic co-ordinate (position) of the nearest respective feature, which may be upstream or downstream (Figure 5.2).

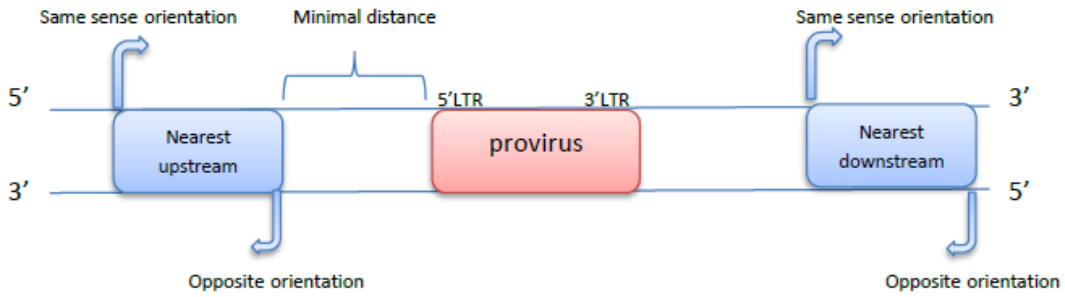


Figure 5.2: Definition of the host environment flanking the host genome

The red central block indicates the position of the HTLV-1 provirus integrated between two genomic elements. The definitions of 'upstream', 'downstream', 'same orientation' or 'opposite orientation' are defined with reference to the HTLV-1 provirus integrated onto the forward strand. The minimal distances are calculated separately for upstream and downstream elements.

5.3.2. Control datasets

DNA samples from 75 asymptomatic carriers originating from the Kumamoto region of Japan and identified as HTLV-1 seropositive by the Japanese blood service were prepared for high throughput sequencing in this laboratory by Dr Heather Niederer over the same time-course and often sharing the same sequencing flow cells as the investigated ATL cohort. Dr Niederer kindly provided the integration site dataset for use as a control dataset. Some experiments were validated bioinformatically with a second AC dataset originating from the Kagoshima region of Japan.

5.3.3. The absolute abundance of all ATL clones is larger than seen in ACs

Whilst absolute abundance is not optimal for allocating abundance bins within this cohort owing to the heterogeneity in proviral load and proviral structure, it remains important to compare the absolute abundances of small, intermediate and large clones of ACs with ATL in order to understand whether any observed differences could be attributed to clonal abundance *per se* or could be directly related to the clinical phenotype i.e. AC or ATL.

The median absolute abundance of the small ATL and AC clones (Figure 5.3) was numerically similar although statistically significantly different (0.588 v 0.478/10000 PBMC, Mann-Whitney test $p=7.13 \times 10^{-69}$), and there were also significant differences in the median absolute abundance of the intermediate AC and ATL clones (1.52 v 254/10,000 PBMC, Mann-Whitney test $p=1.53 \times 10^{-50}$) and in the median absolute abundance of the large AC clones and the presumed malignant ATL clones (156.9 v 6559/10,000 PBMC, Mann-Whitney test p=0.0014). Since the absolute abundance of AC

small and intermediate clones was significantly lower than the intermediate ATL clones, the AC abundance bins were all grouped together for further analysis.

The observation that in absolute terms the intermediate sized clones in ATL are significantly larger than the equivalent clones in ACs is important, since this implies that differences observed between the equivalent relative abundance bins in ACs or ATLs may be due to differences in absolute clone size, rather than associated with disease itself.

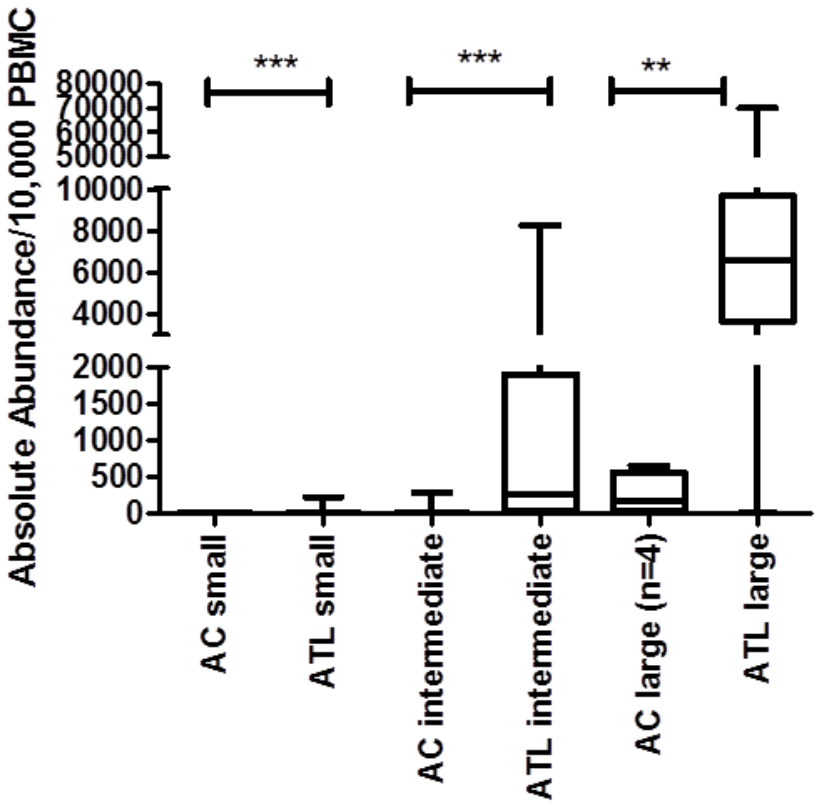


Figure 5.3: Absolute abundance of small, intermediate and large clones in AC and ATL cases

Absolute clonal abundance of ATL samples was significantly greater than ACs within each abundance category (small, intermediate and large). Note the breaks in the y-axis which highlights that there was a 4-5 log difference in size between the small AC/ATL clones and the large, presumed malignant ATL clones.

5.3.4. Preferential integration of small clones into chromosomes 13, 14, 15 and 21

In order to test for evidence of an in vivo selection on the distribution of integration sites the distribution of integration sites per chromosome was compared with both a random integration dataset and an in vitro dataset (in vitro dataset provided by Dr Anat Melamed) and statistically analysed by X^2 testing with correction for multiple comparison testing.

Meekings et al had previously analysed a total of 313 integration sites in vivo using the separate technique of classical LMPCR followed by cloning and Sanger sequencing and identified a significant excess of integration sites into chromosome 13 when compared with random sites (Meekings et al., 2008) but the biological significance of this observation was uncertain. Here, by a separate technique, I firstly tested whether there is evidence for chromosome targeting in vitro and found that there was no bias of integration into any particular chromosome within the in vitro dataset compared with random expectation.

When comparing the in vivo datasets against random, I first tested the control Kumamoto AC dataset and the ATL dataset and then validated the findings against a separate AC dataset arising from the Kagoshima region of Japan. Using X^2 tests followed by corrections for multiple comparison testing, I found a significant excess of integration sites in chromosomes 13 (Kumamoto AC X^2-test $p=8.7 \times 10^{-30}$, Kagoshima AC $p= 7.5 \times 10^{-15}$, ATL small X^2-test $p=4.5 \times 10^{-12}$) , chromosome 14 (Kumamoto AC X^2-test $p=8.35 \times 10^{-10}$, Kagoshima AC X^2-test $p=0.021$) and chromosome 21 (Kumamoto AC X^2-test $p=1.9 \times 10^{-17}$, Kagoshima AC X^2-test 2.3×10^{-19}, ATL small X^2-test $p=0.027$). There was a trend towards increased integrations in chromosome 15 (ATL small X^2-test $p=0.0025$ following correction) although the Kumamoto and Kagoshima cohorts lost significance following multiple correction comparison (Figure 5.4 and Table 5-1). These biases persist when considering both chromosome size and the number of genes per chromosome.

When comparing the in vivo datasets against in vitro (rather than random), the bias of integration into chromosome 13 and 15 remained.

Analysis of ATL intermediate and ATL large clones did not reveal any increased integrations across the chromosomes, but since their combined number of these sites was 307, this may reflect a lack of power.

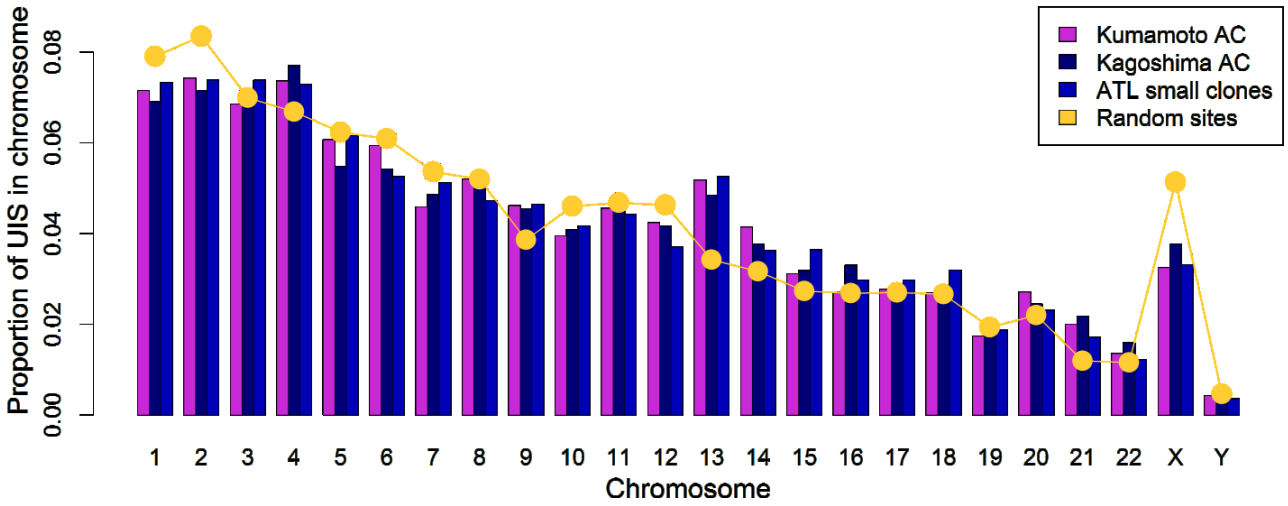


Figure 5.4: Preferential integration into chromosomes 13, 14, 15 and 21

The proportion of unique integration sites (UIS) per chromosome is shown for two independent AC datasets (Kumamoto and Kagoshima) and the small ATL clones. The yellow overlaid line represents the random dataset. There are an increased number of integrations within chromosomes 13, 14, 15 and 21 in the clones of asymptomatic carriers and small ATL clones when compared with random. The bias remains present in chromosome 13 and 15 when compared with control in vitro data (not shown). y-axis shows the proportion of unique integration sites (UIS) in each chromosome, x-axis chromosomes 1-22, X,Y.

Table 5-1: Preferential integration into chromosomes

Control data: / Subgroup:	Random	In vitro
In vitro	None	-
Kumamoto AC	Chromosome 13 Chromosome 14 Chromosome 21	Chromosome 13 Chromosome 15
Kagoshima AC	Chromosome 13 Chromosome 14 Chromosome 21	Chromosome 13 Chromosome 15
Small ATL clones	Chromosome 15 Chromosome 21	Chromosome 13 Chromosome 15

Table 5.1 demonstrates the significant associations of the different AC datasets and small ATL clones against both random datasets and the *in vitro* datasets

5.3.5. Integration within 10Kb of a TSS or CpG island is associated with acute subtypes of ATL and those carrying defective or tax mutated proviruses.

Gillet et al (2011) demonstrated that in vivo integration sites survive in a non-random fashion that favours insertion next to genes, CpG islands and epigenetic marks associated with the control of gene expression and postulated that this was due to a greater accessibility of these genomic regions to proviral integration. Here, I extend this analysis to consider ATL subgroups (both clonal abundance and clinical subtype) and investigate the proximity to transcriptional start sites (regardless of whether the integration site is within a gene or not), Figure 5.5.

When compared with random integration I observed in the Kumamoto ACs that there was a bias of integration within 10 Kb of a TSS of a gene (20.5% v 13.9%, X^2 test, $p<10^{-117}$) and within 10Kb of CpG island (20.9% v 14.2%, X^2 test, $p<10^{-118}$). This was also observed in the small ATL clones (TSS 20.0% v 13.9% X^2 test $p<10^{-61}$, CpG 20.9% v 14.2% X^2 test $p<10^{-43}$) and in the acute subtype of ATL (19.8% TSS X^2 test p=0.01, 18.9% CpG X^2 test p=0.04).

When the large ATL clones were analysed by clinical subtype and by capacity for proviral *tax* gene expression I found a bias towards integration in proximity to TSS and CpG islands in acute subtype (TSS 19.8% v 13.9%X^2 test p=0.01, CpG 18.9% v14.2% X^2 test p=0.04) and in those clones carrying a *tax* mutation (TSS odds ratio 5.3, X^2 test p=0.0031, CpG odds ratio 5.2, X^2 test p=0.00372).

There was no statistically significant enrichment in integration near TSS or CpG islands in the intermediate sized clones compared with random.

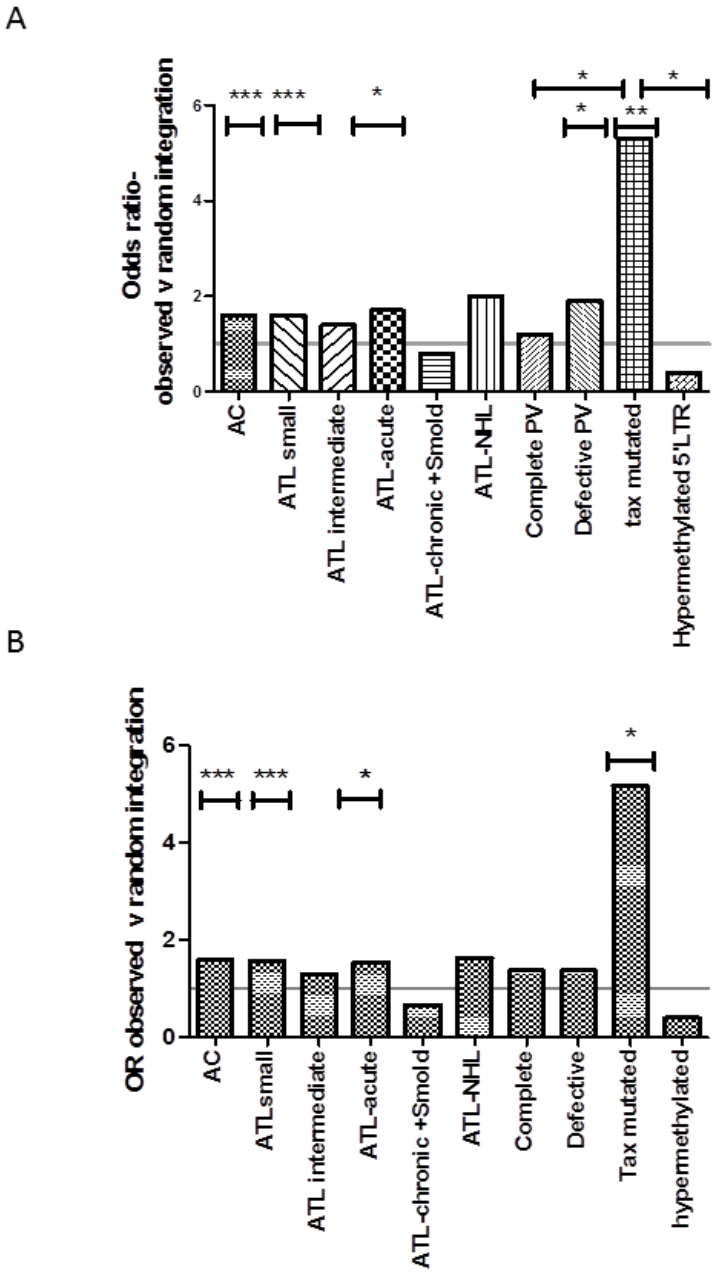


Figure 5.5: Preferential integration within 10Kb TSS and CpG islands

The proportion of integration site within 10Kb of a TSS (panel A) or CpG island (panel B) was compared between integration sites in AC, ATL cases versus random expectation. The large ATL clones were subdivided into clinical subtype and tax status. AC and small ATL clones show a preference for integration in proximity to TSS and CpG islands compared with random expectation. This is also seen in acute ATL cases and those with a *tax* mutated provirus.

5.3.6. Same sense transcriptional orientation and proximity to genes favours clonal expansion.

The transcriptional activity of the host genome in close proximity to an integrated provirus might influence the level of proviral transcription, either directly, through specific interactions between host promotor/enhancer elements and the provirus or indirectly, since unfolded chromatin becomes accessible to transcription factor complexes. Melamed et al reported that in non-malignant infection activatory transcriptional activity upstream appears to suppress Tax expression by a presumed mechanism of transcriptional interference, allowing CTL escape and clonal expansion (Melamed et al., 2013).

We report an increased frequency of integration within 50Kb of genes in ACs (54%, X^2 test $p<10^{-24}$) and ATL small clones (55%, X^2 test $p<10^{-16}$) when compared with random sites (50%). There is an increased frequency in the large ATL clones, although not statistically significant after correction for multiple testing (54%, X^2 test $p=0.07$). There was no bias towards integration in proximity to genes in the ATL intermediate sized clones (50%).

By chance, we would expect to see an equal number of integrations in the same or opposite transcriptional orientation to the provirus. When the analysis was extended to investigate preference for integration by relative transcriptional orientation, clinical subtype and *tax* gene structure within 50Kb of the nearest gene (Figure 5.6) we observed that there was a bias towards integration within the same transcriptional orientation in AC (X^2 test $p<10^{-15}$) , small ATL clones (X^2 test $p<10^{-4}$) and acute subtypes of ATL (X^2 test $p=0.009$) and in those with a defective provirus (X^2 test $p=0.02$).

Given the postulated mechanisms of transcriptional interference described by Melamed et al (Melamed et al., 2013), we considered whether in acute ATL the same or opposite orientation was associated with either upstream or downstream integration of the provirus relative to the nearest gene, but there were too few events within each category for meaningful analysis.

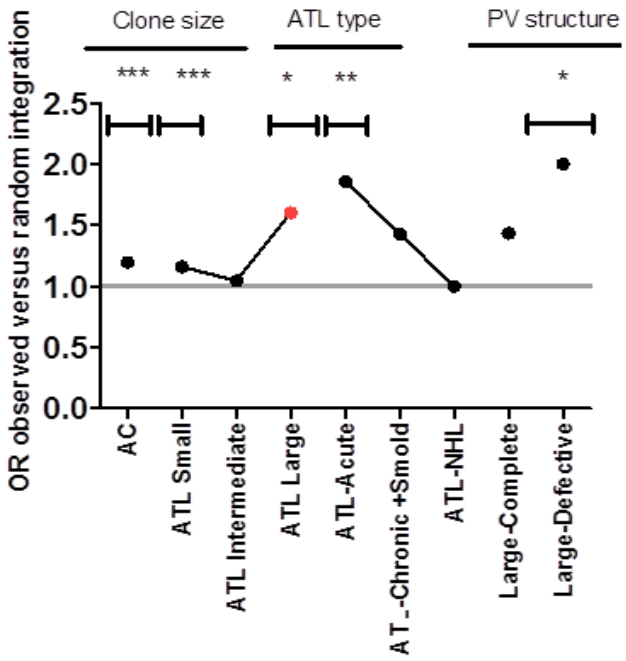


Figure 5.6: Integration in proximity to genes in the same sense transcriptional orientation

The proportion of integration sites within 50Kb of a gene and in the same transcriptional orientation was compared between random expectation and integration sites in AC or ATL cases. The large ATL clones (red dot) were further subdivided into clinical subtype and capacity for proviral *tax* gene expresssion (complete or defective). There were too few hypermethylated or *tax* gene mutated cases to subdivide by both orientation and proximity to genes. AC, small ATL and large ATL clones of acute subtype and defective proviruses showed a preference for integration in proximity to genes in the same transcriptional orientation.

5.3.7. Integration is favoured in proximity to activatory epigenetic marks

By comparison of *in vitro* and *in vivo* integration sites, Gillet et al reported that in chronic infection, clonal abundance was positively associated with integration in proximity to host activatory epigenetic marks and negatively associated with integration near gene-silencing marks. This observation suggested that ACs with low proviral loads and low risk of clinical disease, with efficient immune responses, counterselect integration sites in activatory regions of the genome.

Nine activatory and three inhibitory epigenetic marks were analysed and selected as marks defined by Barski (Barski et al., 2007), detailed in Appendix 1. The proximity to either activatory or inhibitory marks was defined by the proportion of integration sites with these marks above the 90^{th} centile when compared with a random dataset (within 10 Kb of each integration site).

Here, I observed an increase in number of activatory epigenetic marks when compared with random in both ACs and ATLs (all 9 activatory marks) and there was no significant difference between the ACs and large ATL clones. However, there were fewer activatory marks in proximity to the intermediate sized ATL clones (3/9 marks significant) (Figure 5.7, panel A).

By contrast, there was a striking bias towards integration in proximity to repressive epigenetic marks in the intermediate ATL clones compared with the ACs and both small and large ATL clones (Figure 5.7, panel B). There were no associations between proximity to activatory or inhibitory marks by clinical subtype or by the inferred capacity for proviral *tax* gene expression.

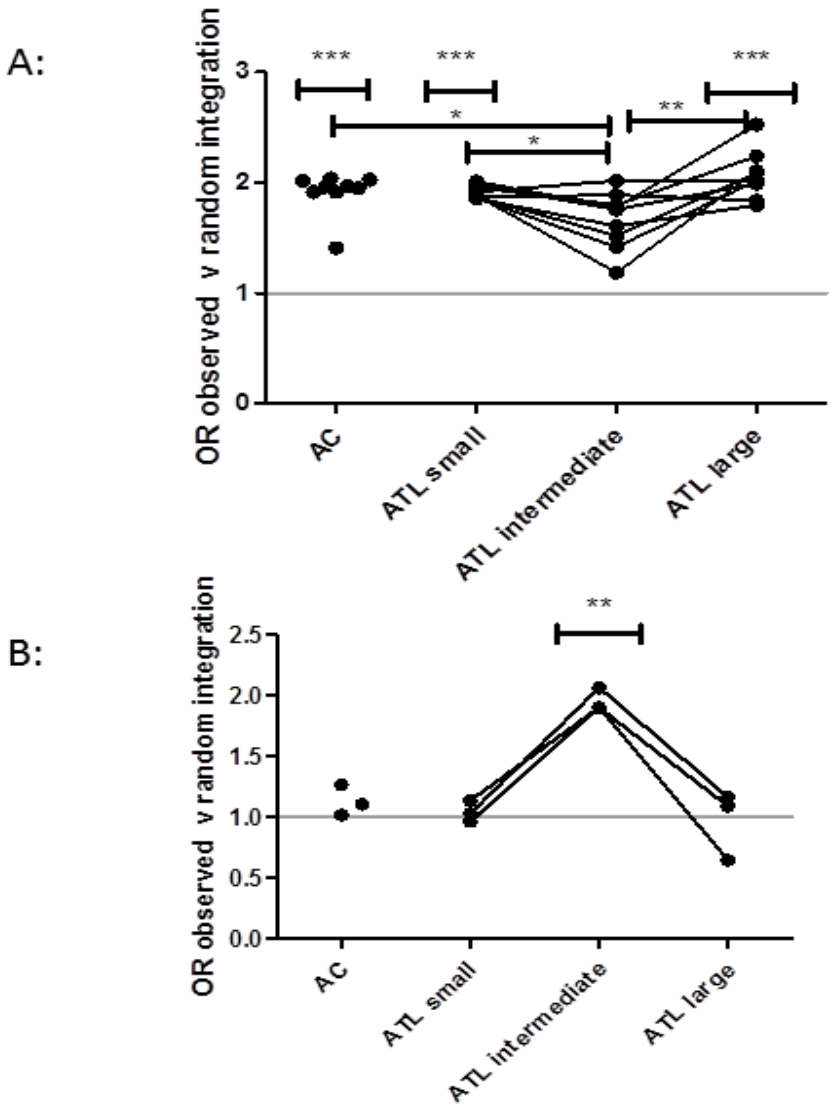


Figure 5.7: Small and large clones favour activatory epigenetic marks, whilst intermediate ATL clones favour inhibitory marks

The proportion of integration sites in proximity to 9 activatory (panel A) and 3 inhibitory (panel B) epigenetic marks was calculated, by comparison to a random dataset. The proportion of random sites in proximity to each mark was defined as the number of sites above the 90th centile and was the benchmark for comparison with the AC and ATL sites. ACs and both small and large ATL clones favour activatory epigenetic marks whilst the intermediate ATL clones show a preference for inhibitory marks

5.3.8. Integration in proximity to TFBS

We wished to test the hypothesis that proviral integration in proximity to transcription factor binding sites (TFBS) is associated with clonal expansion (intermediate sized clones) and malignant transformation (large ATL clones) within the ATL cases. The TFBS data used was downloaded from publicly available datasets of TFBS that have been identified by ChIP-seq experiments. Where possible we used datasets from primary CD4+ T-cells; otherwise data sets from only Jurkat or other human cell lines were analysed. Where primary data was available the SISSRs Perl script (Jothi et al., 2008) was used which is specifically designed to predict the sites of transcription factor binding for a given genomic location. A complete listing of all the datasets used is given in Appendix 1.

TFBSs have been reported to co-localise in the human genome (Dunham et al 2012), and so we used the approach previously described by Melamed (Melamed et al., 2013) to test which of the studied TFBS were independently associated with integration: The statistical approach used was first a likelihood ratio test to identify whether a particular integration site was selectively associated with either upstream or downstream integration. Then, each TFBS annotation (upstream and/or downstream) was subsequently tested separately in univariate analysis. Any significant p-value ($p<0.05$) after multiple comparison corrections, were combined into a multivariate model using a standard step-down general linear model regression approach, until only independent significant factors remained ($p<0.05$). Two separate models were run, to identify TFBS within 0- 100 nucleotides and TFBS within 0- 1000 nucleotides of the integration site.

There were no independent TFBS predictors for ACs, ATL small or ATL large clones and no predictors for clinical subtype or for proviral *tax* expression.

However, when comparing the intermediate sized clones of ATL with either the ACs or small ATL clones, four TFBS were identified to predict intermediate ATL clones: These factors were PCAF, Rad 21 and ZNF263 binding sites (upstream or downstream) at 100 base pairs from the integration site. With the same analysis at 1 Kb from the integration site, the observed TFBS that were associated with intermediate sized ATL clones were E2F4 upstream, Rad 21 downstream and ZNF 263 in either direction (Figure 5.8).

This data was validated against the second AC cohort from the Kagoshima region of Japan and the same predictive factors were identified in the ATL intermediate clones when compared with the Kagoshima ACs.

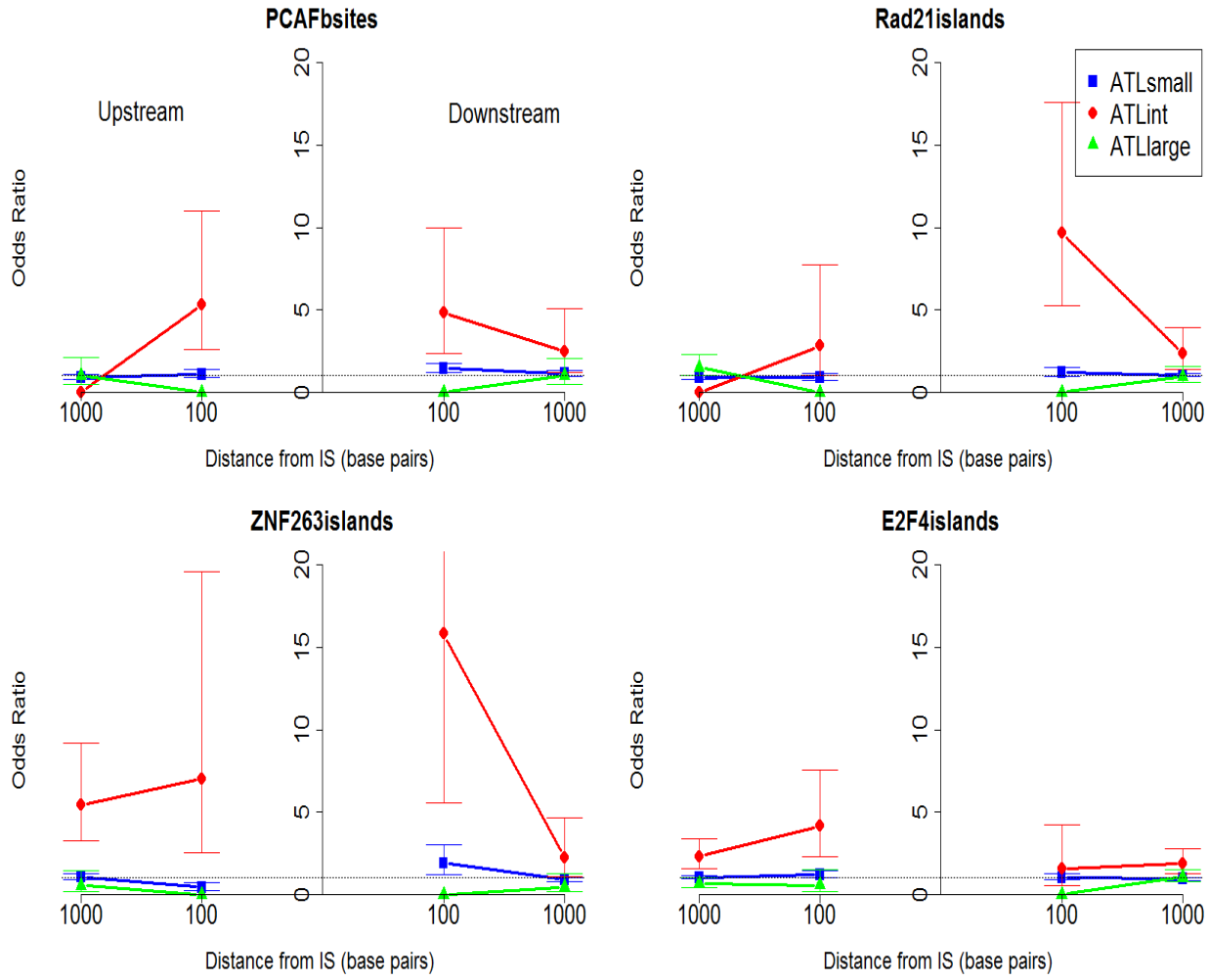


Figure 5.8: Intermediate sized ATL clones only are associated with integration in proximity to specific TFBS

Intermediate sized ATL clones were associated with close proximity to 4 specific TFBS when compared with either Kumamoto ACs, small or large ATL clones; y-axis shows the odds ratio when compared with Kumamoto ACs. The x-axis shows the distance in base pairs from the integration site upstream (left side of y-axis) or downstream (right side of y-axis). The junction of the y-axis with the x-axis represents the integration site. These results were reproducible when compared with an unrelated Kagoshima AC cohort. The definition of 'upstream' and 'downstream' are with respect to the sense strand of HTLV-1. See Appendix 1 for the full list of TFBS tested.

5.3.9. In vivo integration sites are associated with proximity to oncogenes, but are not associated with malignant transformation.

HTLV-1 has not been associated with any recurrent genetic lesion, whilst other closely allied retroviruses or gene therapy vectors have caused insertional mutagenesis and T cell malignancies. Furthermore other haematological malignancies are associated with over-expression of specific oncogenes for a variety of genetic or epigenetic reasons e.g. Burkitt lymphoma and over-expression of c-myc, and it was therefore important to test the hypothesis that HTLV-1 integration is associated with increased proximity to 'cancer associated genes'.

The list of 'cancer associated gene' databases is shown in Appendix 1. This is a list of genes known to be aberrantly expressed by any mechanism in any published cancer.

The datasets were annotated to investigate proximity to nearest cancer-associated genes and showed a significant association between integration near oncogenes in ACs (at 10Kb from insertion site X^2 test $p<10^{-12}$) and, ATL small clones (at 10Kb from insertion site X^2 test $p<10^{-6}$). The ATL large clones showed a bias within 150Kb of the insertion site (X^2 test $p=0.0085$) but was not sufficiently powered to investigate at closer proximities, whilst the intermediate clones did not show any bias compared with random expectation.

Since the AC and small ATL clones both showed a bias towards integration in proximity to oncogenes, this would suggest that this is a general feature of integration in vivo and unlikely to play a significant role in oncogenesis per se.

5.3.10. No hotspots of integration are associated with ATL

Whilst the analysis of integration into chromosomes did not identify any chromosomes with excessive number of integrations within the larger ATL clones (intermediate or large sized), one aim of this analysis is to identify whether there could be more subtle clustering or 'hotspots 'within any given chromosome.

In order to test the hypothesis that the large ATL clones are associated with hotspots or clusters of integration I undertook a bioinformatic analysis initially developed for the analysis of safety of gene therapy vectors (Presson et al., 2011) which was specifically designed to compare datasets of different sizes. I combined both the intermediate and large ATL clones on the basis that they are both abnormally expanded and this would increase the power of this analysis. Retroviral integration clusters in cancer cells have been variably defined as ≥3- 4 hotspots of integration within a 100 kb region, 3 integrations within 50 kb or 2 within 30 kb (Cattoglio et al., 2007; Schroder et al., 2002; Suzuki et al., 2002).

The methodology adopted here applied two methods of hotspot definitions based upon integration site densities. The first method, the 'z threshold', applies a threshold to z-transformed densities and the second method applies a Bayesian change-point analysis ('BCP'). These definitions operate by partitioning the genome into 1 Mb bins and sliding these bins across the genome every 0.25Mb in order to identify any hotspots at the 1Mb window cut-offs (Presson et al., 2011). As suggested by Wu (Wu et al., 2006) we used in vitro data (acute infection) as a control datatset for defining hot-spots in the corresponding ATL data. The acute infection data overcomes any natural virus-specific bias, rather than using a random dataset. Analysis of the in vitro dataset did not identify any significant hotspots in the genome.

We defined 'hotbins' as the 1Mb windows in which there appeared to be an excess of integrations and only further analysed hotbins when they were identified by both the z-threshold and BCP

methods. I then undertook both a ChiSq analysis with correction for multiple testing comparisons and calculated a Poisson distribution to confirm an excess of integrations in large clones compared with all others. The size of the hotbin could then be refined to a hotspot (the smallest window of excess integrations within that hotbin).

It has been previously reported by Doi (Doi et al., 2005) that alphoid repeats are favoured in both non-malignant and malignant infection. We did observe this effect in the cluster analysis but attributed these to mapping errors: Alphoid repeats (satellite repeats) are typically located in the centromeres and telomeres and represent short repetitive regions that are often displayed as gaps in the human genome sequencing project. There are significant efforts to close these gaps, but the nature of these repetitions makes it difficult to map a site uniquely and results in apparent clusters at the junctions of centromeres and the defined sequenced genome.

There were no identified hotbins by combined z-threshold/BCP methods in the in vitro, AC or small ATL clones. Within the ATL intermediate and large clones, there was only a single region of the genome that fulfilled the described definition in chromosome 14, at position 28230332-28258889 (human genome build 18). That is to say following multiple testing comparison there were integrations that are closer together than expected by comparison with the random dataset (X^2 test $p = 10^{-65}$), versus in vitro (X^2 test $p = 0.007$), versus AC (X^2 test $p = 0.001$), and versus small ATL clones (X^2 test $p = 10^{-9}$). However, this 28.6Kb region on chromosome 14, within a large intergenic region of the genome, comprises of integration sites from just two presumed malignant clones arising from two ATL cases.

I conclude that within this large cohort, using powerful bioinformatic tools, there is no evidence to support the idea that there are any clusters or hotspots of HTLV-1 integration in asymptomatic or malignant HTLV-1 infection.

5.3.11. Ingenuity Pathway Analysis (IPA®) suggests ontogeny of the nearest downstream gene may play a role in ATL proliferation.

Whilst I have previously demonstrated that there is no obvious integration in proximity to 'cancer associated genes' within the large ATL clones, this may be limited by the relatively few cancer associated genes compared with the size of the host genome and does not exclude a functional, indirect role for integration in proximity to a category of host genes that may confer a proliferative advantage.

The Ingenuity® platform is a gene ontology software that uses its own 'Ingenuity Knowledge Base' which contains biological and chemical interactions and functional annotations created from millions of individually modelled relationships between proteins, genes, tissues, cells and diseases. Like many gene ontology databases, is widely used by researchers investigating expression microarrays.

I used IPA to identify functional clustering of integration sites and the results indicated that there was an over-representation of genes involved in the cellular pathways associated with 'cell morphology', 'immune cell trafficking' and 'haematological system development and function' within the large ATL clones (Figure 5.9).

The genes associated with these pathways were only observed when looking at the nearest downstream gene (TSS) in the large ATL clones and were not observed in the asymptomatic carriers nor the small or intermediate sized clones. Furthermore, these effects were not present when looking at the nearest upstream gene in the large ATL clones.

These functional categories related to 11 specific genes (CD47, ITGA4, DPYSL2, RAP2A, CASP8, CDKN2A, GTF2I, TACR1, BCL2, IL6ST and HGF) arising from 11 different ATL cases (5.8% cases). These 11 cases were of mixed ATL subtype and of mixed proviral subtype (7 complete proviruses, 3

defective proviruses and 1 tax gene mutated provirus) and 10/11 of these genes are known to be dysregulated in cancers, particularly leukaemias. Of note, the median distance to the transcriptional start site of these 11 genes from the UIS was 13,733 nucleotides (range 628 – 294,863 nucleotides) compared with a median of 122,333 nucleotides when considering the proximity of all integration sites to the nearest oncogene (Mann Whitney, p=0.009). There were too few insertion sites to make any estimation of the contribution of relative transcriptional orientation.

I conclude that the nearest downstream host gene may include a functional category of genes that have a contributory role, not specifically causative, in the proliferation of ~6% ATL cases.

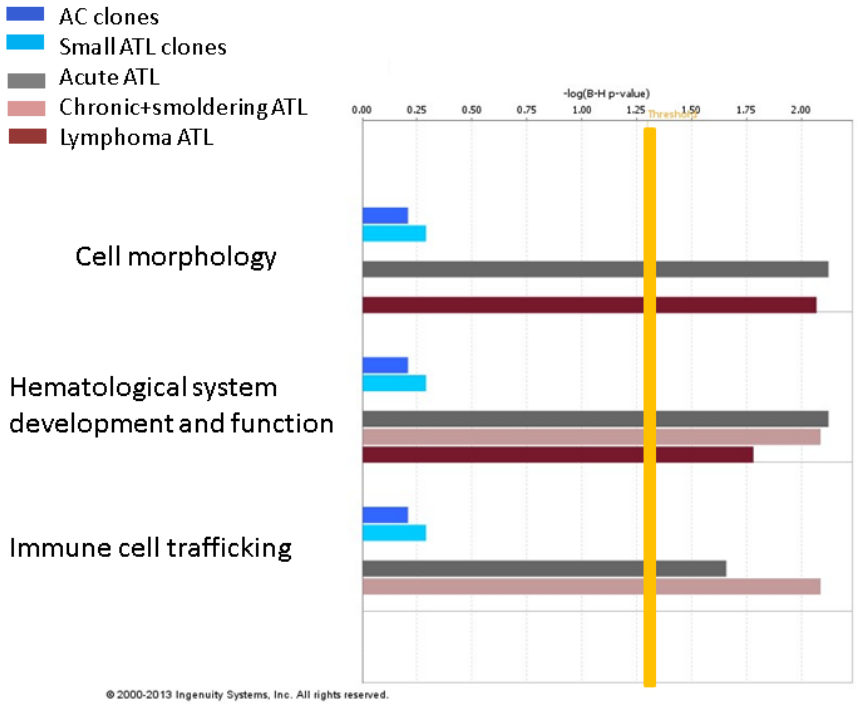


Figure 5.9: Functional classification of genes over-represented amongst the large ATL clones

Functional categories significantly over represented amongst the random, ACs, ATL small, intermediate and large ATL clones as analysed by the Ingenuity® software, using the Ingenuity Pathways Knowledge Base (IPKB) gene population as background. Bars are only visible where there is a statistical over representation compared with the IPKB. Since there are no over-represented pathways involving the random integration sites, or intermediate sized ATL clones the bars are not visible. The vertical yellow threshold represents the line of statistical significance ($p<0.05$) after correction for multiple testing comparison (B-H correction). The numbers of searchable genes for analysis against the IPKB were: random (n=96706), AC (n=5679), ATL small (n=1628), ATL intermediate (n=87), ATL large (acute n=141, lymphoma n=31, chronic & smoldering n=38).

5.4. Discussion

Understanding the mechanisms of HTLV-1 cellular transformation is of great importance both in ATL and within the wider field of retroviral oncogenesis. The factors that determine why only a small percentage of carriers develop ATL are unknown, why ATL typically takes decades of asymptomatic infection to emerge is unknown and predicting which individuals are at risk of disease remains difficult: Whilst patients with proviral load >10% are at high risk of disease (Demontis et al., 2013), there are many high-load ACs who never develop clinical disease.

The aim of the experimental and bioinformatic analysis within this chapter was to identify whether malignant ATL transformation occurs on a 'normal' polyclonal HTLV-1 background or whether there are specific host genomic differences within the proviral integration sites found in the large presumed malignant ATL clones that suggest a selection for certain clones to transform. Furthermore, by comparison of the small clones found in ATL samples with ACs, I was able to test whether the individuals who develop ATL have a chronic background selection for proviral integration sites with specific characteristics that may determine their risk of developing ATL. This has been previously untested, in part because previous techniques are too labour-intensive to investigate large numbers of patients and secondly, because only HTS methods are able to recover, map and quantify the abundance of smaller clones.

The AC cohort used as external controls originated from the same region of Japan and were experimentally prepared and analysed simultaneously using the same laboratory methods. Since the AC cohort originated anonymously from the Japanese blood service, I was not able to correlate my observations with factors such as age or gender. The age and gender of the ATL cases were not known in all samples (Appendix 2). Additionally, since there is a lifetime risk of developing ATL of approximately 5%, we might expect 3 or 4 of these ACs to develop ATL in the future. There is a small

bias in the ACs towards those with higher proviral loads, since we are not able to robustly undertake high-throughput sequencing on patients with very low proviral loads by QPCR (proviral loads <0.1%). This is a frequent problem within the HTLV-1 field when investigating expression of viral genes or proteins.

5.4.1. Preferential integration into acrocentric chromosomes in small clones

Meekings et al identified an excess of integrations into chromosome 13 in DNA samples derived from AC and HAM/TSP patients using classical LMPCR and Sanger sequencing, the biological significance of which was uncertain (Meekings et al., 2008). Here, by a separate technique I confirmed this observation in chromosome 13 and additionally identified excess integrations into chromosome 14, 15 and 21 which are all acrocentric chromosomes. This was initially observed in the two test datasets (Kumamoto ACs and ATL small clones) and independently observed in an unrelated dataset from Kagoshima ACs. There was no bias of integration when the in vitro dataset were compared with random suggesting that other factors are important in targeting of integration. However, the biases into acrocentric chromosomes within the patient derived integration sites remained evident when compared with in vitro (particularly in chromosomes 13 and 15). In addition to chromosomes 13, 14, 15 and 21, chromosome 22 is also acrocentric. However, whilst chromosome 22 was not statistically significantly overrepresented in this analysis, the trend for preferential integration was in the same direction. I did not observe these effects in the expanded ATL clones (ATL intermediate or large), possibly due to the smaller number of integration sites (total 317) across 24 different chromosomes of differing sizes. The increase in integration sites was evenly distributed along the q-arms of these chromosomes and not observed in the p-arms. However, the nucleotide sequences of these p-arms have not been well defined by the various human genome sequencing projects, largely due to the repetitive nature of the sequences, and are represented as

gap regions in the human sequence (Cole et al., 2008; Kidd et al., 2010). It has been recently estimated that 200 Mbp of the human genome remain unmapped (Genovese et al., 2013) which have been largely attributed to sequences within the centromeres and short arms of the acrocentric chromosomes. Next-generation sequence reads from these 'missing pieces' are typically disregarded or misaligned and it is currently estimated that whole genome sequencing projects using the NCBI GRch37 human reference genome build misalign 17Mbp of reads due to the fact that they arise from gaps or missing pieces (Genovese et al., 2013). It is therefore possible that if there are integrations that are associated directly with the ribosomal DNA genes, we cannot detect them by our method which principally relies upon uniquely mapped paired-end reads. The significance of integration into acrocentric chromosomes is uncertain.

An important feature of these chromosomes is their association with nucleoli. The very small p-arms contain approximately 400 rDNA genes (across all 5 chromosomes), also known as nucleolar organising regions, and after mitosis these regions form the nucleoli which act to concentrate the transcriptional and processing machinery required for ribosome formation. In addition to the main function of the nucleolus in ribosomal subunit biogenesis, additional functions have been recently described such as a role in cell-cycle progression and in the stress response (Boisvert et al., 2007; Pederson and Tsai, 2009). In addition, chromatin associated with nucleoli contains high-density AT-rich sequence elements, a low gene density, but a significant enrichment in transcriptionally repressed genes (van Koningsbruggen et al., 2010).

The observation of increased number of proviruses on these chromosomes suggests there is a selection advantage to integration on these chromosomes. There is less negative selection by the immune system and we hypothesise that the provirus is able to couple its transcription to that of the nearby transcriptionally repressed host genes which allows the provirus to persist without expressing Tax protein which allows CTL escape.

In order to validate these observations and mechanisms, it would be essential to first directly visualise these chromosomes and chromosome territories during the cell cycle, for example using single-cell FISH. It would then be useful to test for proviral expression of genes in conjunction with nearby host genes, for example using HTLV-1 T-cell clones derived as described in chapter 3.

5.4.2. Effects of the genomic environment on clonal expansion in vivo

The abundance of an HTLV-1 infected clone in vivo will be determined by the net effect of two opposing forces: Mitotic proliferation of an individual clone and its susceptibility to CTL killing. If these forces acted equally upon all clones we would expect clones in vivo to be of equal abundance. However, HTS data has shown that there is wide variation in clonal abundance both between infected individuals and within each individual and we hypothesise that the proviral integration site is a major determinant of this (Gillet et al., 2011; Melamed et al., 2013). The host genomic environment may determine the frequency and intensity of proviral gene expression which promotes a proliferative advantage.

Patient-derived integration sites (AC or small ATL clones) show identical host genomic characteristics in all the investigated host genomic characteristics, i.e. proximity to TSS, CpG islands, genes, same-sense transcriptional orientation, preference for activatory epigenetic marks, general bias in proximity toward oncogenes, lack of integration 'hotspots' and no preferential gene ontogeny of nearby host genes. This similarity suggests that there is no selection bias towards the survival of certain clonal populations specific to ATL individuals.

Interestingly the results here have shown that large ATL clones (putatively malignant) show very similar host genomic characteristics to the background polyclonal populations (both small ATL and

AC clones). When analysed by subtype, the bias was greatest in acute ATL and in those with either a defective provirus or containing a *tax* gene mutation that prevents Tax expression. Since it has been previously shown that defective proviruses may arise pre integration and *tax* gene mutations may occur early during asymptomatic infection and may be transmissible from mother-to-child (Furukawa et al., 2001; Miyazaki et al., 2007), this supports the hypothesis that by not expressing Tax, possibly for many years, clones have a selection advantage (presumably CTL escape) and that a nearby CpG island or host gene TSS that lies in the same transcriptional orientation confers a survival advantage.

Although this was a large ATL cohort, there were still not enough high-abundance ATL clones to statistically analyse the relative orientation of the provirus and nearby host genes by clinical ATL subtype and *tax* gene expression in order to test the notion that there is a mechanism of transcriptional interaction between the provirus and host. Nonetheless, the observation that non-expressing Tax clones integrated in same-sense transcriptional orientation become abundant remains consistent with the observation by Melamed et al (2013) in ACs and HAM/TSP.

5.4.3. Intermediate sized clones contain proviruses with unique host genomic characteristics

The most surprising observation was the consistent finding of a population of intermediate sized clones with unique genomic characteristics, not seen in the Kumamoto AC cohort and not previously observed by Gillet et al or Melamed et al. As previously described, the usual findings in non-malignant infection are linear correlations with clonal abundance and no intermediate sized population that appear to behave completely differently have been observed. The intermediate sized clones show no preference for proximity to CpG islands, TSS, genes, relative orientation or

activatory epigenetic marks. However, by contrast, there is a striking bias for integration in proximity to repressive epigenetic marks. Furthermore, the observation that specific TFBS are predicted to bind in proximity to these intermediate sized integration sites, whilst no bias is seen ACs, small or large ATL clones further supports the observation that they are a distinct clonal population. This is consistent with observations by Melamed et al who observed a TFBS bias that was greatest for targeting of in vitro integration, rather than for in vivo sites and suggested that proximity to TFBS did not confer an advantage on the infected clones during chronic infection.

Since TFBS can cluster or co-localise, I carried out a logistic regression analysis containing all analysed TFBS in order to ascertain which TFBS are independently associated with clonal expansion. The results (Figure 5.8) identify PCAF, Rad 21, ZNF 263 and E2F4 as independent correlates of intermediate sized clones when compared with ACs, small or large ATL clones. The most striking effects are observed with PCAF binding sites at 100 base pairs upstream or downstream of the provirus, Rad21 up or downstream and ZNF263 up or downstream, the effects of Rad21 and ZNF263 most prominent at 100bp downstream of the proviral integration site.

PCAF (P300/CBP-associated factor) is a transcriptional co-activator known to play a significant role in HTLV-1 infection. Tax mediated viral transcriptional activation occurs via complex interactions between Tax and CREB at the Tax-responsive element-1 (TRE-1) of the 5'LTR. The formation of Tax-CREB promoter complex serves as a high-affinity binding site for the recruitment of the multifunctional cellular co-activators such as CBP, p300 and PCAF (Kashanchi and Brady, 2005).

Rad21 is a component of the cohesin quaternary complex whose classical function is to prevent premature chromatin segregation during mitosis by tethering newly synthesised sister chromatids together. Aberrant cohesion expression has been identified as a significant pathogenic factor in the development of various tumour types associated with chromosomal instability including lymphoid malignancies (Sajesh et al., 2013) and it is known that tumours which display aneuploidy, such as

ATL, are associated with poor prognosis and the rapid acquisition of multi-drug resistance (Duesberg et al., 2001). The role of cohesin in HTLV-1 or other retroviruses, on higher order function is yet to be elucidated.

Intermediate sized clones occurred in 48 cases ATL, with a total of 90 integration sites (median 1 site per case, range 1-11 sites). Only a minority of integration sites were found in proximity to any particular TFBS and relate to 8 specific cases of ATL (17% of cases with intermediate sized clones). Therefore the existence of these TFBS in proximity to an integration site was not necessary for expansion of the intermediate abundance clones, but may have conferred a proliferative advantage on these clones, although the mechanism is not clear. Whilst silencing of Tax expression may confer a selection advantage in early infection, we hypothesise that following transformation, in the presence of an inefficient CTL response, the viral –host dynamics change and allow these abundant intermediate sized clones to express Tax and proliferate. To test this hypothesis would require cell-sorting of Tax positive and Tax-negative ATL PBMCs, followed by high-throughput sequencing and correlation with clonal abundance.

Importantly, I have shown that the intermediate sized population remains relatively large, by comparison with AC/HAMs (absolute abundance of ATL intermediate clones 254/10,000 PBMC versus 1.52/10,000 PBMC in AC/HAM intermediate clones) suggesting that these clones have arisen as a consequence of ATL, rather than causative. The presence of these intermediate sized clones also supports the hypothesis that malignant transformation occurs from a clone within the typical polyclonal background, since if it had arisen from the abnormal intermediate sized clones, we would expect the large ATL clones to have the same genomic characteristics as the intermediate sized clones.

5.4.4. ATL is not caused by integration in proximity to cancer-associated genes

Although it has been known for 30 years that ATL is not caused by a recurrent genomic lesion (Seiki et al., 1984), it remained important to systematically investigate whether there were targeted regions (clusters/hotspots) of the host genome that may contribute to a small percentage of ATL cases. These may only be detected by analysing large cohorts of patients. Here we observe, that all *in vivo* sites in the ACs, small and large ATL clones show a preference for integration in proximity to 'cancer associated genes' when compared with random (again not seen in the intermediate sized clones), but crucially there is no significant difference between the controls (AC, small ATLs) and the large presumed malignant ATL clones. Detailed cluster analysis using sensitive bioinformatic and statistical methods (Presson et al., 2011) does not convincingly suggest that any particular region of the genome is over-represented within the malignant ATL population.

Gene ontology using the IPA® platform suggests that in a small percentage of ATL cases (5.8%), recurrent functional categories were over-represented in the nearest TSS downstream within the large ATL clones. This was not seen in the other categories (random, AC, small ATL or intermediate ATL clones) and was not seen when looking at the nearest gene upstream in the large ATL clones. Furthermore, when these 11 cases of ATL were individually investigated, the implicated gene was close to the provirus (13,733 nucleotides) - significantly closer than the median distance of other larger clones to their respective nearest downstream TSS. This strengthens the case that a provirus upstream may interact transcriptionally with a downstream host gene, contributing to clonal proliferation (Melamed et al., 2013). All of these 11 genes are normally expressed in CD4+ T-lymphocytes and 10 of the 11 genes have been implicated in tumorigenesis, particularly of leukaemias (Chao et al., 2011; Berndt et al., 2013; Fabbri et al., 2013; Hsieh et al., 2013; Kentsis et al., 2012; Lauc et al., 2013; Minato, 2013; Munoz et al., 2012). These 11 cases were of mixed clinical

subtype, mixed proviral *tax* gene subtype and of mixed relative transcriptional orientation to the nearest downstream TSS, precluding inferences about mechanisms of transcriptional interference and the contribution of a nearby upstream HTLV-1 provirus.

Chapter 6. General discussion and summary of thesis work

6.1. Chapter Aim

The aim of this chapter is to summarise the key points of the thesis, which have been discussed in detail at the end of each chapter, and then to place these findings in context with current understanding of ATL pathogenesis and to suggest the implications of these findings in terms of future research work and treatment strategies.

6.2. Summary of major findings

The mechanism by which HTLV-1 transforms infected CD4+ T-cells is largely unknown and despite a wide variation in clinical presentation, few molecular determinants of different clinical presentations are known. Whilst a higher PVL may predispose to clinical disease, we are still unable to predict which patients will develop disease on a per-patient basis and there is no robust prognostic stratification.

Since ATL only arises in HTLV-1 infected individuals and is typically diagnosed 5 or 6 decades following infection, it may be considered that HTLV-1 per se is the 'first hit' to the host genome. On the basis that Tax is only expressed in 40% cases of ATL, the currently accepted model for ATL development is that Tax expression is required in the early stages of leukaemogenesis and allows immortalisation of a particular clone which gives that clone a survival advantage and the opportunity to acquire further genetic or epigenetic aberrations. Later in the disease the ATL clone then acquires a mechanism to down-regulate Tax and escape the immunodominant CTL response. Since it has been shown that *HBZ* sequence is remarkably conserved and mRNA is always detected in ATL, it is considered that HBZ is important in the maintenance of transformed clones (Matsuoka, 2005). However, there are some problems with this model: Firstly it has been shown that *tax* gene mutations and defective proviruses (Miyazaki et al., 2007) arise pre-integration suggesting that these clones would have never expressed Tax (Furukawa et al., 2001) and more recently, HBZ transgenic mice are capable of causing T-cell lymphoma (Satou et al., 2011), supporting the notion that HBZ may be the more dominant proviral gene in disease transformation.

The aim of this work was test the hypothesis that the site of retroviral integration contributes to malignant transformation. That is to say, the integration site, *in cis*, constitutes a 'viral factor' in disease pathogenesis.

The approach taken here was to firstly confirm the widely-held assumption that in non-malignant infection, each T-cell clone carried a single provirus, which allows a robust and simple interpretation of HTS data (i.e. that each detected integration site represented a distinct clone). As discussed in Chapter 3, this finding supports the hypothesis that there is an unknown mechanism of superinfection resistance and, additionally, validates the clinical utility of proviral load measurements. That is to say that the proviral load will accurately reflect the proportion of infected PBMCs.

We aimed to continue this approach with presumed malignant clones from ATL samples, but by comparison of isolated clones with HTS data from mixed PBMC recognised we found that we were not able to maintain these large clones in culture *ex-vivo*. A line of future work would be to develop methods to isolate and successfully expand these clones, since they would be a useful research tool – for example, to investigate the effect of an integrated provirus *in cis*, on higher-order chromatin structure and for potential whole genome sequencing projects.

The ATL integration site method described here allows the systematic evaluation of a large number of patients which was not previously possible using classical LMPCR/IPCR and Sanger sequencing. Previous studies of integration sites reported between 20-50 patients (Doi et al., 2005; Hanai et al., 2004; Ozawa et al., 2004). Furthermore, rapid advances in the annotation of the human genome such as the ENCODE project (Bernstein et al., 2012) and publicly available ChIP-seq datasets allows for a new approach to identifying characteristics of proviral integration in ATL. These new data also allows testing of dynamic changes in clonality over time, which can be applied not only in research, but also clinically in order to monitor the effects of therapy at clonal level (Hodson et al, in submission), to identify 'abnormal' clonal expansion in HTLV-1 infected siblings prior to allogeneic stem cell donation, and, in time, as part of clinical trials.

In chapters 4 and 5 regarding the analysis of the host genomic integration site, I present the main following findings:-

1. There is a preference for clonal survival in acrocentric chromosomes, indicating that integration into these chromosomes provides an advantage, perhaps to evade the CTL response.
2. The number of ATL cases with multiple proviruses has been quantified at 9.1% of the cohort – the majority in acute cases. Whilst multiple proviruses had been previously observed, the proportion of cases had not been previously determined.
3. By comparison of the presumed malignant ATL clones with smaller ATL clones and those from ACs, there were no discriminating genomic features that distinguished the transformed malignant clones from the polyclonal background. This conclusion supports the hypothesis that *trans*-acting factors, which are likely to include Tax and HBZ, are critical in ATL oncogenesis.
4. The presumed malignant clones do not show any recurrent hotspots or clusters of integration in the host genome, although gene ontogeny analysis supports the hypothesis that in a small number of cases (~5%), the nearest downstream gene may play a contributory role.
5. An 'intermediate' population of clones has been identified within the ATL patients that are not present in the Kumamoto control ACs or in other AC or HAM/TSP cohorts. These clones are substantially larger (in terms of absolute abundance) than the largest clones observed in ACs/HAM and show characteristics of the flanking host genome that are systematically different from those of other HTLV-1-infected clones, both the malignant clones in ATL and the clones from subjects with non-malignant infection.

The work in this chapter adds significantly to the previous ATL integration site work (Doi et al., 2005; Hanai et al., 2004), (Ozawa et al., 2004) and additionally to the HTS HTLV-1 integration site work of

(Gillet et al., 2011; Melamed et al., 2013) and Niederer et al (unpublished) to provide a growing picture of the factors associated with malignant transformation and how alterations in host dynamics shape clonal selection forces.

Since the abundance of an HTLV-1-infected clone *in vivo* is determined by the net effect of mitotic proliferation and its susceptibility to CTL killing, I tested here the hypothesis that the integration site itself contributed to one of the proviral proliferative factors. However, I conclude that malignant transformation may occur in any malignant clone, independently of the proviral integration site. This conclusion is consistent with the clinical example highlighted in Figure 5.1 (clonal evolution) in which longitudinal analysis of a single patient identified the future malignant clone in the minor population of clones prior to disease development. It would be important to determine the effect of the proviral integration site on transcriptional regulation of host genes nearby and more distantly (both *in cis* and *in trans*) to elucidate whether the provirus causes transcriptional interference and dysregulates nearby host gene expression. Although the ATL cohort reported here consisted of nearly 200 patients, this was still not sufficiently powered to analyse the effects of upstream or downstream integration and relative orientation effects on transcription of the provirus or nearby host genes. The T-cell clones isolated by limiting dilution as described in chapter 3 would be good model systems for these experiments.

The expanded intermediate clones observed in the ATL cases have not been previously seen. We suggest that these clones arise as a consequence of ATL development due to an inefficient CTL response: Whilst suppressing Tax might allow a clonal advantage early in ATL development (in contrast to the current model), once the malignant environment is established and the CTLs are inefficient then Tax-expressing clones may acquire a proliferative advantage i.e. the normal host dynamics are completely altered. The evidence that these clones emerge after ATL has developed, and do not cause the disease, is that the host genomic properties of these clones have no

resemblance to the large presumed malignant clones, which would be expected if the malignant population were derived from these intermediate-abundance clones.

It has been reported that approximately 5% CD8+ T-cells are infected with HTLV-1 and whilst CD8+-ATL has been described, these arise in αβ-T cells (and not from γδ T-cells). Furthermore, Melamed et al (in submission) have observed that in infection with HTLV-2 which naturally infects CD8+ T-cells, infected T-cell clones become abnormally abundant, although they do not undergo malignant transformation. Perhaps these intermediate-abundance clones observed in the ATL cases reflect abnormally expanded CD8+ infected T-cells in the face of abnormal host dynamics: This hypothesis needs to be tested. It is also possible that these clones support the growth or persistence of the malignant clone.

6.3. Absolute number of clones, rather than oligoclonal expansion, contributes to leukaemogenesis

It has been known for many years that a high proviral load is the main predictor of clinical disease, and it had be assumed that oligoclonal expansion was the cause of higher proviral loads seen in these individuals and that the monoclonal tumour population of ATL was an extension of this proliferation (polyclonal → oligoclonal → monoclonal). However, (Gillet et al., 2011) recently showed that at any given proviral load, individuals with HAM/TSP had a greater number of infected T-cell clones than ACs, and that the increased proviral load was due to the presence of an increased number of clones and not due to oligoclonal proliferation as previously assumed.

Here, I show that the presumed malignant clones in ATL have host genomic characteristics with similar properties to both the AC clones and the small ATL clones. This suggests that a small

proportion of individuals have a genetic predisposition towards widespread viral dissemination and an increased absolute numbers of clones. A strong candidate for this genetic factor is the presence of a detrimental HLA haplotype or, conversely, the absence of a protective haplotype which results in less efficient presentation of viral peptides to the immune system. The evidence for this genetic predisposition is based chiefly upon 3 main observations: Firstly, AC individuals with a proviral load >10% are at higher risk of ATL (>4% in Japan) and an individual's PVL set-point is determined early in the course of their infection. Secondly, specific HLA and KIR types are known to protect from HAM/TSP (as yet this has not been investigated in ATL). These protective HLA types are predicted to present HBZ peptides efficiently to CTLs (Macnamara et al., 2010). Thirdly, the observation that patients diagnosed with ATL often report 1st degree relatives that have died from leukaemia/lymphoma. In addition, there are a few families anecdotally reported where significant numbers of family members develop ATL suggesting a more critical host germline genetic susceptibility e.g. loss of heterozygosity of a tumour suppressor gene, although as yet none of these families have undergone whole genome sequencing to identify these potential lesions (Pombo-de-Oliveira et al., 2001; Kondo et al., 1985).

On a background of increased absolute numbers of clones, a series of step-wise 'bad luck' genetic or epigenetic events (reviewed recently by Vogelstein et al., 2013) may occur over decades, each individually contributing a small survival advantage (it is estimated that each driver mutation confers a 0.4% increase in difference between cell birth and cell death). Our Ingenuity Pathway Analysis® data suggests that in some of those cases (~6%) the ontogeny of the nearest downstream gene may play a contributory role. The net effect of these individually small advantages is to allow a clone to become immortalised and, with further genetic or epigenetic changes, to transform to the malignant phenotype. The number of genetic or epigenetic hits to the host genome is unknown – most whole genome/exome sequencing data suggest that leukaemias are associated with relatively few genetic hits (8-12 per case).

To test these hypotheses would ideally require longitudinal stored PBMC samples on reasonable numbers of individuals prior to and at the time of ATL development. This would allow for robust calculations of absolute clonal abundance whilst still in the asymptomatic phase and additionally to track the malignant clone at a 'pre-malignant' phase, isolate and compare with the same clone at disease transformation by whole genome/exome sequencing.

An interesting observation within this cohort, and in previously reported work, is that the proviral load of leukemic ATL patients is frequently not 100%, even when the blood film suggests large numbers of highly abnormal lymphocytes. This suggests that there is a large population of uninfected T-lymphocytes with a flow-cytometry phenotype resembling ATL cells. It is possible that these cells contributes to the profound immune suppression that is typical of ATL, particularly in cases where the cells share characteristics with the immune-suppressive regulatory T-cell such as expression of CCR4, FoxP3 and GITR. It would therefore be interesting to subgroup ATL cases by presenting proviral load, compare their response to treatment, and quantify the immune response at presentation and again when in remission.

6.4. Directions for future ATL research and direction of clinical management

By comparison with progress that has been made with other acute or chronic leukaemias there is urgent need for basic, translational and clinical research on ATL. The difficulties facing research in the ATL field include a high prevalence in parts of the world with limited resources and in wealthier nations limited access to high quality tumour material. Furthermore, in the acute phase, treatment

needs to be started urgently and even in countries such as Japan there is little time to collect untreated biopsy material for research prior to commencing treatment.

From a basic and clinical research perspective, there remain many important unanswered questions including:

- What precisely are the leukaemia initiating cells (LIC) and could they be therapeutically targeted?
- What, if any, are the molecular features that distinguish the clinical subtypes?
- Why do only a small proportion of individuals develop ATL?
- Who will develop ATL?
- Why does it inevitably take 5 or 6 decades to develop disease?
- Why does disease inevitably relapse?

Having investigated the role of the integration site *in cis* the natural next step would be to consider the role of the integrated provirus *in trans*. The integrated provirus may interact distally with important genes (e.g. oncogenes) on the same or other chromosomes. This could be initially investigated by a combined approach of genome-wide mRNA sequencing with whole exome sequencing to identify target genes, validated with methods such as whole genome sequencing to identify copy number aberrations, deep sequencing and methylation analysis of candidate genes. Following identification of target regions, the transcriptional relationship with the host provirus can then be hypothesised and experimentally tested. Experiments such as these are however costly, require large numbers of individuals to identify candidate regions (typical cancer exome sequencing projects investigate at least 100 samples) and require control DNA from the same individuals and purification of tumour material to exclude background noise in sequencing and to distinguish polymorphisms and de novo mutations.

The ‘first hit’ for ATL leukaemogenesis is thought be HTLV-1 proviral integration and most individuals are infected during infancy, but it remains unexplained as to why it takes several decades to develop disease. By comparison, paediatric precursor B-cell acute lymphoblastic leukaemia is associated with ETV6-AML1 fusion oncoprotein as an early initiating genetic lesion (identified on the heel-prick Guthrie card test in the first week of life) followed by a modest number of ‘driver’ alterations, and results in clinical disease in susceptible children within just a few years. Perhaps genetic lesions alone are not sufficient to drive ATL since one could assume that children and young adults would be observed to develop ATL more frequently. The impact of an integrated provirus (or transgene) upon the methylation of a host genome remains unknown but changes to methylation of the genome are known to occur over decades as part of the normal ageing process. Perhaps HTLV-1 either accelerates or interacts with this process resulting in tumours after several decades. This could be investigated by means of unbiased whole genome bisulfite sequencing of asymptomatic HTLV-1 carriers and matched uninfected controls in younger and older age groups, followed by comparison with ATL cases.

From a clinical perspective it has become clear that leukemic subtypes (acute or chronic) can respond to combination therapy with zidovudine and interferon whilst the lymphoma or ‘bulky acute’ subtypes still require induction with chemotherapy (Bazarbachi et al., 2010). One of the difficulties of utilising trial data, particularly the chemotherapy trial data from Japan, is that these trials often combine all of the clinical subgroups together and exclude patients with a poor performance status, hypercalcaemia or renal failure which makes extrapolation to ‘real life’ clinical scenarios difficult.

Despite the successes of ATL treatment in leukaemic cases with zidovudine and interferon, the mechanism of its action remains largely unknown, and only a proportion of individuals will tolerate long term treatment with almost inevitable disease progression on withdrawing treatment. A future aim of ATL treatment trials has to be twofold: Firstly better induction regimens that result in more

complete remissions and secondly, the eradication of 'leukemic clones' whilst in an apparent state of complete remission. There are emerging attempts to do both of these. Ramos et al have trialled a histone de-acetylase inhibitor (valproate) in patients in remission from chronic ATL (Ramos J.C., 2011) with the aim of inducing viral protein expression allowing autologous killing by the specific cytotoxic T-lymphocytes. Kchour (Kchour et al., 2013; Kchour et al., 2009) and Suarez (Suarez F, 2011) have both reported the successful clinical use of arsenic trioxide in combination with zidovudine and interferon to induce cell cycle arrest and apoptosis of HTLV-1 and ATL cells via a proposed mechanism of rapid shutdown of NFkB. Novel monoclonal antibodies are rapidly being developed, trialled and marketed for many cancers. Mogamulizumab, which targets CCR4 expression on ATL tumour cells has shown improved overall response rates, complete remissions and overall survival in Japan when used as monotherapy in relapsed disease (Ishida et al., 2012) and more recently improved response rates when combined with chemotherapy in first line treatment although it is too early to extrapolate to overall survival (interim analysis, verbal communication Dr Tsukasaki, Japan Lymphoma Study Group). Mogamulizumab is currently under clinical trial in Europe and the USA in patients that largely originate from Africa and the Caribbean. However, lessons from other malignant diseases would suggest that this (or any) targeted therapies are unlikely to be curative alone, but may be suitable for a specific subtype of disease or subgroup of patients or in combination with other agents; the purpose of these trials should be to identify where the successes will lie.

References

Aboud, M., Golde, D.W., Bersch, N., Rosenblatt, J.D., and Chen, I.S. (1987). A colony assay for in vitro transformation by human T cell leukemia viruses type I and type II. Blood *70*, 432-436.

Akagi, K., Suzuki, T., Stephens, R.M., Jenkins, N.A., and Copeland, N.G. (2004). RTCGD: retroviral tagged cancer gene database. Nucleic Acids Res *32*, D523-527.

Akagi, T., Ono, H., and Shimotohno, K. (1995). Characterization of T cells immortalized by Tax1 of human T-cell leukemia virus type 1. Blood *86*, 4243-4249.

Akizuki, S., Nakazato, O., Higuchi, Y., Tanabe, K., Setoguchi, M., Yoshida, S., Miyazaki, Y., Yamamoto, S., Sudou, S., Sannomiya, K., and et al. (1987). Necropsy findings in HTLV-I associated myelopathy. Lancet *1*, 156-157.

Allen, T.M., and Altfeld, M. (2003). HIV-1 superinfection. J Allergy Clin Immunol *112*, 829-835.

Amano, M., Kurokawa, M., Ogata, K., Itoh, H., Kataoka, H., and Setoyama, M. (2008). New entity, definition and diagnostic criteria of cutaneous adult T-cell leukemia/lymphoma: human T-lymphotropic virus type 1 proviral DNA load can distinguish between cutaneous and smoldering types. J Dermatol *35*, 270-275.

Aparicio, S., and Caldas, C. (2013). The implications of clonal genome evolution for cancer medicine. N Engl J Med *368*, 842-851.

Arisawa, K., Soda, M., Endo, S., Kurokawa, K., Katamine, S., Shimokawa, I., Koba, T., Takahashi, T., Saito, H., Doi, H., and Shirahama, S. (2000). Evaluation of adult T-cell leukemia/lymphoma incidence and its impact on non-Hodgkin lymphoma incidence in southwestern Japan. Int J Cancer *85*, 319-324.

Ariumi, Y., Kaida, A., Lin, J.Y., Hirota, M., Masui, O., Yamaoka, S., Taya, Y., and Shimotohno, K. (2000). HTLV-1 tax oncoprotein represses the p53-mediated trans-activation function through coactivator CBP sequestration. Oncogene *19*, 1491-1499.

Arnold, J., Yamamoto, B., Li, M., Phipps, A.J., Younis, I., Lairmore, M.D., and Green, P.L. (2006). Enhancement of infectivity and persistence in vivo by HBZ, a natural antisense coded protein of HTLV-1. Blood *107*, 3976-3982.

Arnulf, B., Thorel, M., Poirot, Y., Tamouza, R., Boulanger, E., Jaccard, A., Oksenhendler, E., Hermine, O., and Pique, C. (2004). Loss of the ex vivo but not the reinducible CD8+ T-cell response to Tax in human T-cell leukemia virus type 1-infected patients with adult T-cell leukemia/lymphoma. Leukemia *18*, 126-132.

Bajenoff, M., Egen, J.G., Koo, L.Y., Laugier, J.P., Brau, F., Glaichenhaus, N., and Germain, R.N. (2006). Stromal cell networks regulate lymphocyte entry, migration, and territoriality in lymph nodes. Immunity *25*, 989-1001.

Banerjee, P., Tripp, A., Lairmore, M.D., Crawford, L., Sieburg, M., Ramos, J.C., Harrington, W., Jr., Beilke, M.A., and Feuer, G. (2010). Adult T-cell leukemia/lymphoma development in HTLV-1-infected humanized SCID mice. Blood *115*, 2640-2648.

Bangham C.R.M, C.L.B., Melamed A. (2013). HTLV-1 clonality in adult T-cell leukaemia and non-malignant HTLV-1 infection. Seminars in Cancer Biology *in press*.

Bangham, C.R. (2009). CTL quality and the control of human retroviral infections. Eur J Immunol *39*, 1700-1712.

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823-837.

Basbous, J., Bazarbachi, A., Granier, C., Devaux, C., and Mesnard, J.M. (2003). The central region of human T-cell leukemia virus type 1 Tax protein contains distinct domains involved in subunit dimerization. J Virol *77*, 13028-13035.

Bazarbachi, A., Plumelle, Y., Carlos Ramos, J., Tortevoye, P., Otrock, Z., Taylor, G., Gessain, A., Harrington, W., Panelatti, G., and Hermine, O. (2010). Meta-analysis on the use of zidovudine and interferon-alfa in adult T-cell leukemia/lymphoma showing improved survival in the leukemic subtypes. J Clin Oncol *28*, 4177-4183.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R*., et al.* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. Nature *456*, 53-59.

Berndt, S.I., Gustafsson, S., Magi, R., Ganna, A., Wheeler, E., Feitosa, M.F., Justice, A.E., Monda, K.L., Croteau-Chonka, D.C., Day, F.R*., et al.* (2013). Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. Nat Genet *45*, 501-512.

Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., and Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

Berry, C.C., Gillet, N.A., Melamed, A., Gormley, N., Bangham, C.R., and Bushman, F.D. (2012). Estimating abundances of retroviral insertion sites from DNA fragment length data. Bioinformatics *28*, 755-762.

Birzele, F., Fauti, T., Stahl, H., Lenter, M.C., Simon, E., Knebel, D., Weith, A., Hildebrandt, T., and Mennerich, D. (2011). Next-generation insights into regulatory T cells: expression profiling and FoxP3 occupancy in Human. Nucleic Acids Res *39*, 7946-7960.

Boisvert, F.M., van Koningsbruggen, S., Navascues, J., and Lamond, A.I. (2007). The multifunctional nucleolus. Nat Rev Mol Cell Biol *8*, 574-585.

Botcheva, K., McCorkle, S.R., McCombie, W.R., Dunn, J.J., and Anderson, C.W. (2011). Distinct p53 genomic binding patterns in normal and cancer-derived human cells. Cell Cycle *10*, 4237-4249.

Brady, T., Agosto, L.M., Malani, N., Berry, C.C., O'Doherty, U., and Bushman, F. (2009). HIV integration site distributions in resting and activated CD4+ T cells infected in culture. AIDS *23*, 1461-1471.

Burke, D.S. (1997). Recombination in HIV: an important viral evolutionary strategy. Emerg Infect Dis *3*, 253-259.

Bushman, F.D., Hoffmann, C., Ronen, K., Malani, N., Minkah, N., Rose, H.M., Tebas, P., and Wang, G.P. (2008). Massively parallel pyrosequencing in HIV research. AIDS *22*, 1411-1415.

Cartier, L.M., Cea, J.G., Vergara, C., Araya, F., and Born, P. (1997). Clinical and neuropathological study of six patients with spastic paraparesis associated with HTLV-I: an axomyelinic degeneration of the central nervous system. J Neuropathol Exp Neurol *56*, 403-413.

Cattoglio, C., Facchini, G., Sartori, D., Antonelli, A., Miccio, A., Cassani, B., Schmidt, M., von Kalle, C., Howe, S., Thrasher, A.J.*, et al.* (2007). Hot spots of retroviral integration in human CD34+ hematopoietic cells. Blood *110*, 1770-1778.

Cavazzana-Calvo, M., Payen, E., Negre, O., Wang, G., Hehir, K., Fusil, F., Down, J., Denaro, M., Brady, T., Westerman, K.*, et al.* (2010). Transfusion independence and HMGA2 activation after gene therapy of human beta-thalassaemia. Nature *467*, 318-322.

Cavrois, M., Gessain, A., Wain-Hobson, S., and Wattel, E. (1996). Proliferation of HTLV-1 infected circulating cells in vivo in all asymptomatic carriers and patients with TSP/HAM. Oncogene *12*, 2419-2423.

Cavrois, M., Wain-Hobson, S., and Wattel, E. (1995). Stochastic events in the amplification of HTLV-I integration sites by linker-mediated PCR. Res Virol *146*, 179-184.

Chang, Y.B., Kaidarova, Z., Hindes, D., Bravo, M., Kiely, N., Kamel, H., Dubay, D., Hoose, B., and Murphy, E.L. (2013). Seroprevalence and Demographic Determinants of Human T-Lymphotropic Virus Type-1 and -2 Infections among First-time Blood Donors, U.S. 2000-2009. J Infect Dis. 2013 Oct 31 [epub ahead of print]

Chao, M.P., Alizadeh, A.A., Tang, C., Jan, M., Weissman-Tsukamoto, R., Zhao, F., Park, C.Y., Weissman, I.L., and Majeti, R. (2011). Therapeutic antibody targeting of CD47 eliminates human acute lymphoblastic leukemia. Cancer Res *71*, 1374-1384.

Chen, S., Ishii, N., Ine, S., Ikeda, S., Fujimura, T., Ndhlovu, L.C., Soroosh, P., Tada, K., Harigae, H., Kameoka, J.*, et al.* (2006). Regulatory T cell-like activity of Foxp3+ adult T cell leukemia cells. Int Immunol *18*, 269-277.

Chou, K.S., Okayama, A., Su, I.J., Lee, T.H., and Essex, M. (1996). Preferred nucleotide sequence at the integration target site of human T-cell leukemia virus type I from patients with adult T-cell leukemia. Int J Cancer *65*, 20-24.

Ciminale, V., Zotti, L., D'Agostino, D.M., Ferro, T., Casareto, L., Franchini, G., Bernardi, P., and Chieco-Bianchi, L. (1999). Mitochondrial targeting of the p13II protein coded by the x-II ORF of human T-cell leukemia/lymphotropic virus type I (HTLV-I). Oncogene *18*, 4505-4514.

Cole, C.G., McCann, O.T., Collins, J.E., Oliver, K., Willey, D., Gribble, S.M., Yang, F., McLaren, K., Rogers, J., Ning, Z.*, et al.* (2008). Finishing the finished human chromosome 22 sequence. Genome Biol *9*, R78.

Craigie, R., and Bushman, F.D. (2012). HIV DNA integration. Cold Spring Harb Perspect Med *2*, a006890.

D'Agostino, D.M., Silic-Benussi, M., Hiraragi, H., Lairmore, M.D., and Ciminale, V. (2005). The human T-cell leukemia virus type 1 p13II protein: effects on mitochondrial function and cell growth. Cell Death Differ *12 Suppl 1*, 905-915.

De Castro-Costa, C.M., Araujo, A.Q., Barreto, M.M., Takayanagui, O.M., Sohler, M.P., da Silva, E.L., de Paula, S.M., Ishak, R., Ribas, J.G., Rovirosa, L.C*., et al.* (2006). Proposal for diagnostic criteria of tropical spastic paraparesis/HTLV-I-associated myelopathy (TSP/HAM). AIDS Res Hum Retroviruses *22*, 931-935.

de The, G., and Bomford, R. (1993). An HTLV-I vaccine: why, how, for whom? AIDS Res Hum Retroviruses *9*, 381-386.

Delamarre, L., Rosenberg, A.R., Pique, C., Pham, D., Callebaut, I., and Dokhelar, M.C. (1996). The HTLV-I envelope glycoproteins: structure and functions. J Acquir Immune Defic Syndr Hum Retrovirol *13 Suppl 1*, S85-91.

Demontis, M.A., Hilburn, S., and Taylor, G.P. (2013). Human T cell lymphotropic virus type 1 viral load variability and long-term trends in asymptomatic carriers and in patients with human T cell lymphotropic virus type 1-related diseases. AIDS Res Hum Retroviruses *29*, 359-364.

Derse, D., Crise, B., Li, Y., Princler, G., Lum, N., Stewart, C., McGrath, C.F., Hughes, S.H., Munroe, D.J., and Wu, X. (2007). Human T-cell leukemia virus type 1 integration target sites in the human genome: comparison with those of other retroviruses. J Virol *81*, 6731-6741.

Devereux, S. (2011). Two-faced T cells in CLL. Blood *117*, 5273-5274.

Djilali, S., and Parodi, A.L. (1989). The BLV-induced leukemia--lymphosarcoma complex in sheep. Vet Immunol Immunopathol *22*, 233-244.

Doi, K., Wu, X., Taniguchi, Y., Yasunaga, J., Satou, Y., Okayama, A., Nosaka, K., and Matsuoka, M. (2005). Preferential selection of human T-cell leukemia virus type I provirus integration sites in leukemic versus carrier states. Blood *106*, 1048-1053.

Dow, B.C., Munro, H., Ferguson, K., Buchanan, I., Jarvis, L., Jordan, T., Franklin, I.M., and McClelland, M. (2001). HTLV antibody screening using mini-pools. Transfus Med *11*, 419-422.

Duesberg, P., Stindl, R., and Hehlmann, R. (2001). Origin of multidrug resistance in cells with and without multidrug resistance genes: chromosome reassortments catalyzed by aneuploidy. Proc Natl Acad Sci U S A *98*, 11283-11288.

Einsiedel, L., Fernandes, L., Spelman, T., Steinfort, D., and Gotuzzo, E. (2012). Bronchiectasis is associated with human T-lymphotropic virus 1 infection in an Indigenous Australian population. Clin Infect Dis *54*, 43-50.

Endo, K., Hirata, A., Iwai, K., Sakurai, M., Fukushi, M., Oie, M., Higuchi, M., Hall, W.W., Gejyo, F., and Fujii, M. (2002). Human T-cell leukemia virus type 2 (HTLV-2) Tax protein transforms a rat fibroblast cell line but less efficiently than HTLV-1 Tax. J Virol *76*, 2648-2653.

Enose-Akahata, Y., Abrams, A., Massoud, R., Bialuk, I., Johnson, K.R., Green, P.L., Maloney, E.M., and Jacobson, S. (2013). Humoral immune response to HTLV-1 basic leucine zipper factor (HBZ) in HTLV-1-infected individuals. Retrovirology *10*, 19.

Euskirchen, G.M., Auerbach, R.K., Davidov, E., Gianoulis, T.A., Zhong, G., Rozowsky, J., Bhardwaj, N., Gerstein, M.B., and Snyder, M. (2011). Diverse roles and interactions of the SWI/SNF chromatin remodeling complex revealed using global approaches. PLoS Genet *7*, e1002008.

Fabbri, G., Khiabanian, H., Holmes, A.B., Wang, J., Messina, M., Mullighan, C.G., Pasqualucci, L., Rabadan, R., and Dalla-Favera, R. (2013). Genetic lesions associated with chronic lymphocytic leukemia transformation to Richter syndrome. J Exp Med *210*, 2273-2288.

Fan, J., Ma, G., Nosaka, K., Tanabe, J., Satou, Y., Koito, A., Wain-Hobson, S., Vartanian, J.P., and Matsuoka, M. (2010). APOBEC3G generates nonsense mutations in human T-cell leukemia virus type 1 proviral genomes in vivo. J Virol *84*, 7278-7287.

Felipe, L., Goncalves, D.U., Santos, M.A., Proietti, F.A., Ribas, J.G., Carneiro-Proietti, A.B., and Lambertucci, J.R. (2008). Vestibular-evoked myogenic potential (VEMP) to evaluate cervical myelopathy in human T-cell lymphotropic virus type I infection. Spine (Phila Pa 1976) *33*, 1180-1184.

Franchini, G. (1995). Molecular mechanisms of human T-cell leukemia/lymphotropic virus type I infection. Blood *86*, 3619-3639.

Frietze, S., Lan, X., Jin, V.X., and Farnham, P.J. (2010). Genomic targets of the KRAB and SCAN domain-containing zinc finger protein 263. J Biol Chem *285*, 1393-1403.

Fujita, P.A., Rhead, B., Zweig, A.S., Hinrichs, A.S., Karolchik, D., Cline, M.S., Goldman, M., Barber, G.P., Clawson, H., Coelho, A.*, et al.* (2011). The UCSC Genome Browser database: update 2011. Nucleic Acids Res *39*, D876-882.

Fujiwara, T., O'Geen, H., Keles, S., Blahnik, K., Linnemann, A.K., Kang, Y.A., Choi, K., Farnham, P.J., and Bresnick, E.H. (2009). Discovering hematopoietic mechanisms through genome-wide analysis of GATA factor chromatin occupancy. Mol Cell *36*, 667-681.

Furukawa, Y., Fujisawa, J., Osame, M., Toita, M., Sonoda, S., Kubota, R., Ijichi, S., and Yoshida, M. (1992). Frequent clonal proliferation of human T-cell leukemia virus type 1 (HTLV-1)-infected T cells in HTLV-1-associated myelopathy (HAM-TSP). Blood *80*, 1012-1016.

Furukawa, Y., Kubota, R., Tara, M., Izumo, S., and Osame, M. (2001). Existence of escape mutant in HTLV-I tax during the development of adult T-cell leukemia. Blood *97*, 987-993.

Furukawa, Y., Yamashita, M., Usuku, K., Izumo, S., Nakagawa, M., and Osame, M. (2000). Phylogenetic subgroups of human T cell lymphotropic virus (HTLV) type I in the tax gene and their association with different risks for HTLV-I-associated myelopathy/tropical spastic paraparesis. J Infect Dis *182*, 1343-1349.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004). A census of human cancer genes. Nat Rev Cancer *4*, 177-183.

Gabet, A.S., Gessain, A., and Wattel, E. (2003). High simian T-cell leukemia virus type 1 proviral loads combined with genetic stability as a result of cell-associated provirus replication in naturally infected, asymptomatic monkeys. Int J Cancer *107*, 74-83.

Gaudray, G., Gachon, F., Basbous, J., Biard-Piechaczyk, M., Devaux, C., and Mesnard, J.M. (2002). The complementary strand of the human T-cell leukemia virus type 1 RNA genome encodes a bZIP transcription factor that down-regulates viral transcription. J Virol *76*, 12813-12822.

Genovese, G., Handsaker, R.E., Li, H., Kenny, E.E., and McCarroll, S.A. (2013). Mapping the human reference genome's missing sequence by three-way admixture in Latino genomes. Am J Hum Genet *93*, 411-421.

Gessain, A., and Cassar, O. (2012). Epidemiological Aspects and World Distribution of HTLV-1 Infection. Front Microbiol *3*, 388.

Ghez, D., Lepelletier, Y., Jones, K.S., Pique, C., and Hermine, O. (2010). Current concepts regarding the HTLV-1 receptor complex. Retrovirology *7*, 99.

Ghez, D., Lepelletier, Y., Lambert, S., Fourneau, J.M., Blot, V., Janvier, S., Arnulf, B., van Endert, P.M., Heveker, N., Pique, C., and Hermine, O. (2006). Neuropilin-1 is involved in human T-cell lymphotropic virus type 1 entry. J Virol *80*, 6844-6854.

Gillet, N.A., Gutierrez, G., Rodriguez, S.M., de Brogniez, A., Renotte, N., Alvarez, I., Trono, K., and Willems, L. (2013). Massive Depletion of Bovine Leukemia Virus Proviral Clones Located in Genomic Transcriptionally Active Sites during Primary Infection. PLoS Pathog *9*, e1003687.

Gillet, N.A., Malani, N., Melamed, A., Gormley, N., Carter, R., Bentley, D., Berry, C., Bushman, F.D., Taylor, G.P., and Bangham, C.R. (2011). The host genomic environment of the provirus determines the abundance of HTLV-1-infected T-cell clones. Blood *117*, 3113-3122.

Gini, C., Cuppini, C. (1912). Italian. Variabilita e Mutabilità, (Variability and Mutability), 156 pages.

Glowacka, I., Korn, K., Potthoff, S.A., Lehmann, U., Kreipe, H.H., Ivens, K., Barg-Hock, H., Schulz, T.F., and Heim, A. (2013). Delayed seroconversion and rapid onset of lymphoproliferative disease after transmission of human T-cell lymphotropic virus type 1 from a multiorgan donor. Clin Infect Dis *57*, 1417-1424.

Grassmann, R., Aboud, M., and Jeang, K.T. (2005). Molecular mechanisms of cellular transformation by HTLV-1 Tax. Oncogene *24*, 5976-5985.

Gratton, S., Cheynier, R., Dumaurier, M.J., Oksenhendler, E., and Wain-Hobson, S. (2000). Highly restricted spread of HIV-1 and multiply infected cells within splenic germinal centers. Proc Natl Acad Sci U S A *97*, 14566-14571.

Greten, T.F., Slansky, J.E., Kubota, R., Soldan, S.S., Jaffee, E.M., Leist, T.P., Pardoll, D.M., Jacobson, S., and Schneck, J.P. (1998). Direct visualization of antigen-specific T cells: HTLV-1 Tax11-19- specific CD8(+) T cells are activated in peripheral blood and accumulate in cerebrospinal fluid from HAM/TSP patients. Proc Natl Acad Sci U S A *95*, 7568-7573.

Hacein-Bey-Abina, S., Garrigue, A., Wang, G.P., Soulier, J., Lim, A., Morillon, E., Clappier, E., Caccavelli, L., Delabesse, E., Beldjord, K.*, et al.* (2008). Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. J Clin Invest *118*, 3132-3142.

Hacein-Bey-Abina, S., Von Kalle, C., Schmidt, M., McCormack, M.P., Wulffraat, N., Leboulch, P., Lim, A., Osborne, C.S., Pawliuk, R., Morillon, E.*, et al.* (2003). LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. Science *302*, 415-419.

Hahn, W.C., Counter, C.M., Lundberg, A.S., Beijersbergen, R.L., Brooks, M.W., and Weinberg, R.A. (1999). Creation of human tumour cells with defined genetic elements. Nature *400*, 464-468.

Hajj, H.E., Nasr, R., Kfoury, Y., Dassouki, Z., Nasser, R., Kchour, G., Hermine, O., de The, H., and Bazarbachi, A. (2012). Animal models on HTLV-1 and related viruses: what did we learn? Front Microbiol *3*, 333.

Hanahan, D., and Weinberg, R.A. (2000). The hallmarks of cancer. Cell *100*, 57-70.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646-674.

Hanai, S., Nitta, T., Shoda, M., Tanaka, M., Iso, N., Mizoguchi, I., Yashiki, S., Sonoda, S., Hasegawa, Y., Nagasawa, T., and Miwa, M. (2004). Integration of human T-cell leukemia virus type 1 in genes of leukemia cells of patients with adult T-cell leukemia. Cancer Sci *95*, 306-310.

Hanon, E., Hall, S., Taylor, G.P., Saito, M., Davis, R., Tanaka, Y., Usuku, K., Osame, M., Weber, J.N., and Bangham, C.R. (2000). Abundant tax protein expression in CD4+ T cells infected with human T-cell lymphotropic virus type I (HTLV-I) is prevented by cytotoxic T lymphocytes. Blood *95*, 1386-1392.

Hare, S., Gupta, S.S., Valkov, E., Engelman, A., and Cherepanov, P. (2010). Retroviral intasome assembly and inhibition of DNA strand transfer. Nature *464*, 232-236.

Hasegawa, H., Sawa, H., Lewis, M.J., Orba, Y., Sheehy, N., Yamamoto, Y., Ichinohe, T., Tsunetsugu-Yokota, Y., Katano, H., Takahashi, H.*, et al.* (2006). Thymus-derived leukemia-lymphoma in mice transgenic for the Tax gene of human T-lymphotropic virus type I. Nat Med *12*, 466-472.

Hewitt, P.E., Davison, K., Howell, D.R., and Taylor, G.P. (2013). Human T-lymphotropic virus lookback in NHS Blood and Transplant (England) reveals the efficacy of leukoreduction. Transfusion *53*, 2168-2175.

Hilburn, S., Rowan, A., Demontis, M.A., MacNamara, A., Asquith, B., Bangham, C.R., and Taylor, G.P. (2011). In vivo expression of human T-lymphotropic virus type 1 basic leucine-zipper protein generates specific CD8+ and CD4+ T-lymphocyte responses that correlate with clinical outcome. J Infect Dis *203*, 529-536.

Hiramatsu, K., and Yoshikura, H. (1986). Frequent partial deletion of human adult T-cell leukemia virus type I proviruses in experimental transmission: pattern and possible implication. J Virol *58*, 508-512.

Hiraragi, H., Michael, B., Nair, A., Silic-Benussi, M., Ciminale, V., and Lairmore, M. (2005). Human T-lymphotropic virus type 1 mitochondrion-localizing protein p13II sensitizes Jurkat T cells to Ras-mediated apoptosis. J Virol *79*, 9449-9457.

Hodson, A., Laydon, D.J., Bain, B.J., Fields, P.A., and Taylor, G.P. (2013). Pre-morbid human T-lymphotropic virus type I proviral load, rather than percentage of abnormal lymphocytes, is associated with an increased risk of aggressive adult T-cell leukemia/lymphoma. Haematologica *98*, 385-388.

Hoshida, Y., Li, T., Dong, Z., Tomita, Y., Yamauchi, A., Hanai, J., and Aozasa, K. (2001). Lymphoproliferative disorders in renal transplant patients in Japan. Int J Cancer *91*, 869-875.

Hsieh, Y.T., Gang, E.J., Geng, H., Park, E., Huantes, S., Chudziak, D., Dauber, K., Schaefer, P., Scharman, C., Shimada, H.*, et al.* (2013). Integrin alpha4 blockade sensitizes drug resistant pre-B acute lymphoblastic leukemia to chemotherapy. Blood *121*, 1814-1818.

Huret, J.L., Minor, S.L., Dorkeld, F., Dessen, P., and Bernheim, A. (2000). Atlas of genetics and cytogenetics in oncology and haematology, an interactive database. Nucleic Acids Res *28*, 349-351.

Igakura, T., Stinchcombe, J.C., Goon, P.K., Taylor, G.P., Weber, J.N., Griffiths, G.M., Tanaka, Y., Osame, M., and Bangham, C.R. (2003). Spread of HTLV-I between lymphocytes by virus-induced polarization of the cytoskeleton. Science *299*, 1713-1716.

Ijichi, S., Izumo, S., Eiraku, N., Machigashira, K., Kubota, R., Nagai, M., Ikegami, N., Kashio, N., Umehara, F., Maruyama, I., and et al. (1993). An autoaggressive process against bystander tissues in HTLV-I-infected individuals: a possible pathomechanism of HAM/TSP. Med Hypotheses *41*, 542-547.

Imura, A., Hori, T., Imada, K., Kawamata, S., Tanaka, Y., Imamura, S., and Uchiyama, T. (1997). OX40 expressed on fresh leukemic cells from adult T-cell leukemia patients mediates cell adhesion to vascular endothelial cells: implication for the possible involvement of OX40 in leukemic cell infiltration. Blood *89*, 2951-2958.

Ishida, T., Joh, T., Uike, N., Yamamoto, K., Utsunomiya, A., Yoshida, S., Saburi, Y., Miyamoto, T., Takemoto, S., Suzushima, H*., et al.* (2012). Defucosylated anti-CCR4 monoclonal antibody (KW-0761) for relapsed adult T-cell leukemia-lymphoma: a multicenter phase II study. J Clin Oncol *30*, 837-842.

Iwanaga, M., Watanabe, T., Utsunomiya, A., Okayama, A., Uchimaru, K., Koh, K.R., Ogata, M., Kikuchi, H., Sagara, Y., Uozumi, K*., et al.* (2010). Human T-cell leukemia virus type I (HTLV-1) proviral load and disease progression in asymptomatic HTLV-1 carriers: a nationwide prospective study in Japan. Blood *116*, 1211-1219.

Iwasaki, Y., Ohara, Y., Kobayashi, I., and Akizuki, S. (1992). Infiltration of helper/inducer T lymphocytes heralds central nervous system damage in human T-cell leukemia virus infection. Am J Pathol *140*, 1003-1008.

Jason, J.M., McDougal, J.S., Cabradilla, C., Kalyanaraman, V.S., and Evatt, B.L. (1985). Human T-cell leukemia virus (HTLV-I) p24 antibody in New York City blood product recipients. Am J Hematol *20*, 129-137.

Jeang, K.T., Boros, I., Brady, J., Radonovich, M., and Khoury, G. (1988). Characterization of cellular factors that interact with the human T-cell leukemia virus type I p40x-responsive 21-base-pair sequence. J Virol *62*, 4499-4509.

Jeffery, K.J., Siddiqui, A.A., Bunce, M., Lloyd, A.L., Vine, A.M., Witkover, A.D., Izumo, S., Usuku, K., Welsh, K.I., Osame, M., and Bangham, C.R. (2000). The influence of HLA class I alleles and heterozygosity on the outcome of human T cell lymphotropic virus type I infection. J Immunol *165*, 7278-7284.

Jeffery, K.J., Usuku, K., Hall, S.E., Matsumoto, W., Taylor, G.P., Procter, J., Bunce, M., Ogg, G.S., Welsh, K.I., Weber, J.N*., et al.* (1999). HLA alleles determine human T-lymphotropic virus-I (HTLV-I) proviral load and the risk of HTLV-I-associated myelopathy. Proc Natl Acad Sci U S A *96*, 3848-3853.

Jin, D.Y., and Jeang, K.T. (1997). HTLV-I Tax self-association in optimal trans-activation function. Nucleic Acids Res *25*, 379-387.

Jin, D.Y., Spencer, F., and Jeang, K.T. (1998). Human T cell leukemia virus type 1 oncoprotein Tax targets the human mitotic checkpoint protein MAD1. Cell *93*, 81-91.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science *316*, 1497-1502.

Johnson, J.M., Nicot, C., Fullen, J., Ciminale, V., Casareto, L., Mulloy, J.C., Jacobson, S., and Franchini, G. (2001). Free major histocompatibility complex class I heavy chain is preferentially targeted for degradation by human T-cell leukemia/lymphotropic virus type 1 p12(I) protein. J Virol *75*, 6086-6094.

Jones, K.S., Petrow-Sadowski, C., Bertolette, D.C., Huang, Y., and Ruscetti, F.W. (2005). Heparan sulfate proteoglycans mediate attachment and entry of human T-cell leukemia virus type 1 virions into CD4+ T cells. J Virol *79*, 12692-12702.

Jones, K.S., Petrow-Sadowski, C., Huang, Y.K., Bertolette, D.C., and Ruscetti, F.W. (2008). Cell-free HTLV-1 infects dendritic cells leading to transmission and transformation of CD4(+) T cells. Nat Med *14*, 429-436.

Josefsson, L., King, M.S., Makitalo, B., Brannstrom, J., Shao, W., Maldarelli, F., Kearney, M.F., Hu, W.S., Chen, J., Gaines, H.*, et al.* (2011). Majority of CD4+ T cells from peripheral blood of HIV-1-infected individuals contain only one HIV DNA molecule. Proc Natl Acad Sci U S A *108*, 11199-11204.

Jothi, R., Cuddapah, S., Barski, A., Cui, K., and Zhao, K. (2008). Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. Nucleic Acids Res *36*, 5221-5231.

Jung, A., Maier, R., Vartanian, J.P., Bocharov, G., Jung, V., Fischer, U., Meese, E., Wain-Hobson, S., and Meyerhans, A. (2002). Recombination: Multiply infected spleen cells in HIV patients. Nature *418*, 144.

Kamihira, S., Sohda, H., Atogami, S., Toriya, K., Yamada, Y., Tsukazaki, K., Momita, S., Ikeda, S., Kusano, M., Amagasaki, T., and et al. (1992). Phenotypic diversity and prognosis of adult T-cell leukemia. Leuk Res *16*, 435-441.

Kamihira, S., Sugahara, K., Tsuruda, K., Minami, S., Uemura, A., Akamatsu, N., Nagai, H., Murata, K., Hasegawa, H., Hirakata, Y.*, et al.* (2005). Proviral status of HTLV-1 integrated into the host genomic DNA of adult T-cell leukemia cells. Clin Lab Haematol *27*, 235-241.

Kannagi, M., Sugamura, K., Kinoshita, K., Uchino, H., and Hinuma, Y. (1984). Specific cytolysis of fresh tumor cells by an autologous killer T cell line derived from an adult T cell leukemia/lymphoma patient. J Immunol *133*, 1037-1041.

Kaplan, J.E., Osame, M., Kubota, H., Igata, A., Nishitani, H., Maeda, Y., Khabbaz, R.F., and Janssen, R.S. (1990). The risk of development of HTLV-I-associated myelopathy/tropical spastic paraparesis among persons infected with HTLV-I. J Acquir Immune Defic Syndr *3*, 1096-1101.

Kashanchi, F., and Brady, J.N. (2005). Transcriptional and post-transcriptional gene regulation of HTLV-1. Oncogene *24*, 5938-5951.

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S.M., Habegger, L., Rozowsky, J., Shi, M., Urban, A.E.*, et al.* (2010). Variation in transcription factor binding among humans. Science *328*, 232-235.

Kattan, T., MacNamara, A., Rowan, A.G., Nose, H., Mosley, A.J., Tanaka, Y., Taylor, G.P., Asquith, B., and Bangham, C.R. (2009). The avidity and lytic efficiency of the CTL response to HTLV-1. J Immunol *182*, 5723-5729.

Kchour, G., Rezaee, R., Farid, R., Ghantous, A., Rafatpanah, H., Tarhini, M., Kooshyar, M.M., El Hajj, H., Berry, F., Mortada, M.*, et al.* (2013). The combination of arsenic, interferon-alpha, and zidovudine restores an "immunocompetent-like" cytokine expression profile in patients with adult T-cell leukemia lymphoma. Retrovirology *10*, 91.

Kchour, G., Tarhini, M., Kooshyar, M.M., El Hajj, H., Wattel, E., Mahmoudi, M., Hatoum, H., Rahimi, H., Maleki, M., Rafatpanah, H.*, et al.* (2009). Phase 2 study of the efficacy and safety of the combination of arsenic trioxide, interferon alpha, and zidovudine in newly diagnosed chronic adult T-cell leukemia/lymphoma (ATL). Blood *113*, 6528-6532.

Kentsis, A., Reed, C., Rice, K.L., Sanda, T., Rodig, S.J., Tholouli, E., Christie, A., Valk, P.J., Delwel, R., Ngo, V.*, et al.* (2012). Autocrine activation of the MET receptor tyrosine kinase in acute myeloid leukemia. Nat Med *18*, 1118-1122.

Kidd, J.M., Sampas, N., Antonacci, F., Graves, T., Fulton, R., Hayden, H.S., Alkan, C., Malig, M., Ventura, M., Giannuzzi, G.*, et al.* (2010). Characterization of missing human genome sequences and copy-number polymorphic insertions. Nat Methods *7*, 365-371.

Kondo, T., Nonaka, H., Miyamoto, N., Yoshida, R., Matsue, Y., Ohguchi, Y., Inouye, H., Komoda, H., Hinuma, Y., and Hanaoka, M. (1985). Incidence of adult T-cell leukemia-lymphoma and its familial clustering. Int J Cancer *35*, 749-751.

Koralnik, I.J., Gessain, A., Klotman, M.E., Lo Monico, A., Berneman, Z.N., and Franchini, G. (1992). Protein isoforms encoded by the pX region of human T-cell leukemia/lymphotropic virus type I. Proc Natl Acad Sci U S A *89*, 8813-8817.

Korber, B., Okayama, A., Donnelly, R., Tachibana, N., and Essex, M. (1991). Polymerase chain reaction analysis of defective human T-cell leukemia virus type I proviral genomes in leukemic cells of patients with adult T-cell leukemia. J Virol *65*, 5471-5476.

Koyanagi, Y., Itoyama, Y., Nakamura, N., Takamatsu, K., Kira, J., Iwamasa, T., Goto, I., and Yamamoto, N. (1993). In vivo infection of human T-cell leukemia virus type I in non-T cells. Virology *196*, 25-33.

Kozako, T., Arima, N., Toji, S., Masamoto, I., Akimoto, M., Hamada, H., Che, X.F., Fujiwara, H., Matsushita, K., Tokunaga, M.*, et al.* (2006). Reduced frequency, diversity, and function of human T cell leukemia virus type 1-specific CD8+ T cell in adult T cell leukemia patients. J Immunol *177*, 5718-5726.

Kozako, T., Yoshimitsu, M., Fujiwara, H., Masamoto, I., Horai, S., White, Y., Akimoto, M., Suzuki, S., Matsushita, K., Uozumi, K.*, et al.* (2009). PD-1/PD-L1 expression in human T-cell leukemia virus type 1 carriers and adult T-cell leukemia/lymphoma patients. Leukemia *23*, 375-382.

Kwok, S., Ehrlich, G., Poiesz, B., Kalish, R., and Sninsky, J.J. (1988). Enzymatic amplification of HTLV-I viral sequences from peripheral blood mononuclear cells and infected tissues. Blood *72*, 1117-1123.

Langerak, A.W., Groenen, P.J., Bruggemann, M., Beldjord, K., Bellan, C., Bonello, L., Boone, E., Carter, G.I., Catherwood, M., Davi, F.*, et al.* (2012). EuroClonality/BIOMED-2 guidelines for interpretation and reporting of Ig/TCR clonality testing in suspected lymphoproliferations. Leukemia *26*, 2159-2171.

Langerak, A.W., van Dongen J.J.M (2012). Multiple clonal Ig/TCR products: implications for interpretation of clonality findings. J Hematolpathol *5*, 35-43.

Lauc, G., Huffman, J.E., Pucic, M., Zgaga, L., Adamczyk, B., Muzinic, A., Novokmet, M., Polasek, O., Gornik, O., Kristic, J.*, et al.* (2013). Loci associated with N-glycosylation of human immunoglobulin G show pleiotropy with autoimmune diseases and haematological cancers. PLoS Genet *9*, e1003225.

Lee, B.K., Bhinge, A.A., and Iyer, V.R. (2011). Wide-ranging functions of E2F4 in transcriptional activation and repression revealed by genome-wide analysis. Nucleic Acids Res *39*, 3558-3573.

Levin, A., Hayouka, Z., Friedler, A., Brack-Werner, R., Volsky, D.J., and Loyter, A. (2010). A novel role for the viral Rev protein in promoting resistance to superinfection by human immunodeficiency virus type 1. J Gen Virol *91*, 1503-1513.

Li, H., Malani, N., Hamilton, S.R., Schlachterman, A., Bussadori, G., Edmonson, S.E., Shah, R., Arruda, V.R., Mingozzi, F., Wright, J.F.*, et al.* (2011). Assessing the potential for AAV vector genotoxicity in a murine model. Blood *117*, 3311-3319.

Liao, W., Lin, J.X., Wang, L., Li, P., and Leonard, W.J. (2011). Modulation of cytokine receptors by IL-2 broadly regulates differentiation into helper T cell lineages. Nat Immunol *12*, 551-559.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L., and Law, M. (2012). Comparison of next-generation sequencing systems. J Biomed Biotechnol *2012*, 251364.

Lunardi-Iskandar, Y., Gessain, A., Lam, V.H., and Gallo, R.C. (1993). Abnormal in vitro proliferation and differentiation of T cell colony-forming cells in patients with tropical spastic paraparesis/human T lymphocyte virus type I (HTLV-I)-associated myeloencephalopathy and healthy HTLV-I carriers. The Journal of experimental medicine *177*, 741-750.

Macnamara, A., Rowan, A., Hilburn, S., Kadolsky, U., Fujiwara, H., Suemori, K., Yasukawa, M., Taylor, G., Bangham, C.R., and Asquith, B. (2010). HLA class I binding of HBZ determines outcome in HTLV-1 infection. PLoS Pathog *6* (9): e1001117. doi:10.1371/journal.ppat.1001117

Maertens, G.N., Hare, S., and Cherepanov, P. (2010). The mechanism of retroviral integration from X-ray structures of its key intermediates. Nature *468*, 326-329.

Mahieux, R., and Gessain, A. (2007). Adult T-cell leukemia/lymphoma and HTLV-1. Curr Hematol Malig Rep *2*, 257-264.

Majorovits, E., Nejmeddine, M., Tanaka, Y., Taylor, G.P., Fuller, S.D., and Bangham, C.R. (2008). Human T-lymphotropic virus-1 visualized at the virological synapse by electron tomography. PLoS One *3*, e2251.

Manel, N., Kim, F.J., Kinet, S., Taylor, N., Sitbon, M., and Battini, J.L. (2003). The ubiquitous glucose transporter GLUT-1 is a receptor for HTLV. Cell *115*, 449-459.

Marcos, L.A., Terashima, A., Dupont, H.L., and Gotuzzo, E. (2008). Strongyloides hyperinfection syndrome: an emerging global infectious disease. Trans R Soc Trop Med Hyg *102*, 314-318.

Marshall, N.A., Christie, L.E., Munro, L.R., Culligan, D.J., Johnston, P.W., Barker, R.N., and Vickers, M.A. (2004). Immunosuppressive regulatory T cells are abundant in the reactive lymphocytes of Hodgkin lymphoma. Blood *103*, 1755-1762.

Matsuoka, M. (2005). Human T-cell leukemia virus type I (HTLV-I) infection and the onset of adult T-cell leukemia (ATL). Retrovirology *2*, 27.

Matsuoka, M., and Jeang, K.T. (2007). Human T-cell leukaemia virus type 1 (HTLV-1) infectivity and cellular transformation. Nat Rev Cancer *7*, 270-280.

Matsuoka, M., and Jeang, K.T. (2011). Human T-cell leukemia virus type 1 (HTLV-1) and leukemic transformation: viral infectivity, Tax, HBZ and therapy. Oncogene *30*, 1379-1389.

McCarthy, T.J., Kennedy, J.L., Blakeslee, J.R., and Bennett, B.T. (1990). Spontaneous malignant lymphoma and leukemia in a simian T-lymphotropic virus type I (STLV-I) antibody positive olive baboon. Lab Anim Sci *40*, 79-81.

Meekings, K.N., Leipzig, J., Bushman, F.D., Taylor, G.P., and Bangham, C.R. (2008). HTLV-1 integration into transcriptionally active genomic regions is associated with proviral expression and with HAM/TSP. PLoS Pathog *4*, e1000027.

Melamed, A., Laydon, D.J., Gillet, N.A., Tanaka, Y., Taylor, G.P., and Bangham, C.R. (2013). Genome-wide determinants of proviral targeting, clonal abundance and expression in natural HTLV-1 infection. PLoS Pathog *9*, e1003271.

Minato, N. (2013). Rap G protein signal in normal and disordered lymphohematopoiesis. Exp Cell Res *319*, 2323-2328.

Miura, M., Yasunaga, J.I., Tanabe, J., Sugata, K., Zhao, T., Ma, G., Miyazato, P., Ohshima, K., Kaneko, A., Watanabe, A.*, et al.* (2013). Characterization of simian T-cell leukemia virus type 1 in naturally infected Japanese macaques as a model of HTLV-1 infection. Retrovirology *10*, 118.

Miyazaki, M., Yasunaga, J., Taniguchi, Y., Tamiya, S., Nakahata, T., and Matsuoka, M. (2007). Preferential selection of human T-cell leukemia virus type 1 provirus lacking the 5' long terminal repeat during oncogenesis. J Virol *81*, 5714-5723.

Mochizuki, M., Watanabe, T., Yamaguchi, K., Yoshimura, K., Nakashima, S., Shirao, M., Araki, S., Takatsuki, K., Mori, S., and Miyata, N. (1992). Uveitis associated with human T-cell lymphotropic virus type I. Am J Ophthalmol *114*, 123-129.

Morgan, O.S., Rodgers-Johnson, P., Mora, C., and Char, G. (1989). HTLV-1 and polymyositis in Jamaica. Lancet *2*, 1184-1187.

Munoz, M., Gonzalez-Ortega, A., and Covenas, R. (2012). The NK-1 receptor is expressed in human leukemia and is involved in the antitumor action of aprepitant and other NK-1 receptor antagonists on acute lymphoblastic leukemia cell lines. Invest New Drugs *30*, 529-540.

Murphy, E.L., Figueroa, J.P., Gibbs, W.N., Brathwaite, A., Holding-Cobham, M., Waters, D., Cranston, B., Hanchard, B., and Blattner, W.A. (1989). Sexual transmission of human T-lymphotropic virus type I (HTLV-I). Ann Intern Med *111*, 555-560.

Murphy, E.L., Lee, T.H., Chafets, D., Nass, C.C., Wang, B., Loughlin, K., and Smith, D. (2004). Higher human T lymphotropic virus (HTLV) provirus load is associated with HTLV-I versus HTLV-II, with HTLV-II subtype A versus B, and with male sex and a history of blood transfusion. J Infect Dis *190*, 504-510.

Nagai, K., Jinnai, I., Hata, T., Usui, T., Sasaki, D., Tsukasaki, K., Sugahara, K., Hishikawa, Y., Yamada, Y., Tanaka, Y.*, et al.* (2008). Adhesion-dependent growth of primary adult T cell leukemia cells with

down-regulation of HTLV-I p40Tax protein: a novel in vitro model of the growth of acute ATL cells. Int J Hematol *88*, 551-564.

Nagai, M., Brennan, M.B., Sakai, J.A., Mora, C.A., and Jacobson, S. (2001). CD8(+) T cells are an in vivo reservoir for human T-cell lymphotropic virus type I. Blood *98*, 1858-1861.

Nagai, M., and Jacobson, S. (2001). Immunopathogenesis of human T cell lymphotropic virus type I-associated myelopathy. Curr Opin Neurol *14*, 381-386.

Nagai, M., and Osame, M. (2003). Human T-cell lymphotropic virus type I and neurological diseases. J Neurovirol *9*, 228-235.

Nagasato, K., Nakamura, T., Ohishi, K., Shibayama, K., Motomura, M., Ichinose, K., Tsujihata, M., and Nagataki, S. (1991). Active production of anti-human T-lymphotropic virus type I (HTLV-I) IgM antibody in HTLV-I-associated myelopathy. J Neuroimmunol *32*, 105-109.

Nam, S.H., Copeland, T.D., Hatanaka, M., and Oroszlan, S. (1993). Characterization of ribosomal frameshifting for expression of pol gene products of human T-cell leukemia virus type I. J Virol *67*, 196-203.

Neel, B.G., Hayward, W.S., Robinson, H.L., Fang, J., and Astrin, S.M. (1981). Avian leukosis virus-induced tumors have common proviral integration sites and synthesize discrete new RNAs: oncogenesis by promoter insertion. Cell *23*, 323-334.

Nejmeddine, M., and Bangham, C.R. (2010). The HTLV-1 Virological Synapse. Viruses *2*, 1427-1447.

Nejmeddine, M., Negi, V.S., Mukherjee, S., Tanaka, Y., Orth, K., Taylor, G.P., and Bangham, C.R. (2009). HTLV-1-Tax and ICAM-1 act on T-cell signal pathways to polarize the microtubule-organizing center at the virological synapse. Blood *114*, 1016-1025.

Nicot, C., Dundr, M., Johnson, J.M., Fullen, J.R., Alonzo, N., Fukumoto, R., Princler, G.L., Derse, D., Misteli, T., and Franchini, G. (2004). HTLV-1-encoded p30II is a post-transcriptional negative regulator of viral replication. Nat Med *10*, 197-201.

Niewiesk, S., Daenke, S., Parker, C.E., Taylor, G., Weber, J., Nightingale, S., and Bangham, C.R. (1995). Naturally occurring variants of human T-cell leukemia virus type I Tax protein impair its recognition by cytotoxic T lymphocytes and the transactivation function of Tax. J Virol *69*, 2649-2653.

Oh, U., and Jacobson, S. (2008). Treatment of HTLV-I-associated myelopathy/tropical spastic paraparesis: toward rational targeted therapy. Neurol Clin *26*, 781-797, ix-x.

Ohshima, K., Hashimoto, K., Izumo, S., Suzumiya, J., and Kikuchi, M. (1996). Detection of human T lymphotrophic virus type I (HTLV-I) DNA and mRNA in individual cells by polymerase chain reaction (PCR) in situ hybridization (ISH) and reverse transcription (RT)-PCR ISH. Hematol Oncol *14*, 91-100.

Ohshima, K., Kikuchi, M., Masuda, Y., Kobari, S., Sumiyoshi, Y., Eguchi, F., Mohtai, H., Yoshida, T., Takeshita, M., and Kimura, N. (1991). Defective provirus form of human T-cell leukemia virus type I in adult T-cell leukemia/lymphoma: clinicopathological features. Cancer Res *51*, 4639-4642.

Osame, M., Usuku, K., Izumo, S., Ijichi, N., Amitani, H., Igata, A., Matsumoto, M., and Tara, M. (1986). HTLV-I associated myelopathy, a new clinical entity. Lancet *1*, 1031-1032.

Overbaugh, J., and Bangham, C.R. (2001). Selection forces and constraints on retroviral sequence variation. Science *292*, 1106-1109.

Ozawa, T., Itoyama, T., Sadamori, N., Yamada, Y., Hata, T., Tomonaga, M., and Isobe, M. (2004). Rapid isolation of viral integration site reveals frequent integration of HTLV-1 into expressed loci. J Hum Genet *49*, 154-165.

Pais-Correia, A.M., Sachse, M., Guadagnini, S., Robbiati, V., Lasserre, R., Gessain, A., Gout, O., Alcover, A., and Thoulouze, M.I. (2010). Biofilm-like extracellular viral assemblies mediate HTLV-1 cell-to-cell transmission at virological synapses. Nat Med *16*, 83-89.

Parsons, B.L. (2008). Many different tumor types have polyclonal tumor origin: evidence and implications. Mutat Res *659*, 232-247.

Patten, P.E., Buggins, A.G., Richards, J., Wotherspoon, A., Salisbury, J., Mufti, G.J., Hamblin, T.J., and Devereux, S. (2008). CD38 expression in chronic lymphocytic leukemia is regulated by the tumor microenvironment. Blood *111*, 5173-5181.

Payne, G.S., Bishop, J.M., and Varmus, H.E. (1982). Multiple arrangements of viral DNA and an activated host oncogene in bursal lymphomas. Nature *295*, 209-214.

Pederson, T., and Tsai, R.Y. (2009). In search of nonribosomal nucleolar protein function and regulation. J Cell Biol *184*, 771-776.

Pedral-Sampaio, D.B., Martins Netto, E., Pedrosa, C., Brites, C., Duarte, M., and Harrington, W., Jr. (1997). Co-Infection of Tuberculosis and HIV/HTLV Retroviruses: Frequency and Prognosis Among Patients Admitted in a Brazilian Hospital. Braz J Infect Dis *1*, 31-35.

Poiesz, B.J., Ruscetti, F.W., Gazdar, A.F., Bunn, P.A., Minna, J.D., and Gallo, R.C. (1980). Detection and isolation of type C retrovirus particles from fresh and cultured lymphocytes of a patient with cutaneous T-cell lymphoma. Proc Natl Acad Sci U S A *77*, 7415-7419.

Pombo-de-Oliveira, M.S., Carvalho, S.M., Borducchi, D., Dobbin, J., Salvador, J., Correa, R.B., Moellman, A., Loureiro, P., Chiattone, C., and Rios, M. (2001). Adult T-cell leukemia/lymphoma and cluster of HTLV-I associated diseases in Brazilian settings. Leuk Lymphoma *42*, 135-144.

Pozzatti, R., Vogel, J., and Jay, G. (1990). The human T-lymphotropic virus type I tax gene can cooperate with the ras oncogene to induce neoplastic transformation of cells. Mol Cell Biol *10*, 413-417.

Presson, A.P., Kim, N., Xiaofei, Y., Chen, I.S., and Kim, S. (2011). Methodology and software to detect viral integration site hot-spots. BMC Bioinformatics *12*, 367.

Raha, D., Wang, Z., Moqtaderi, Z., Wu, L., Zhong, G., Gerstein, M., Struhl, K., and Snyder, M. (2010). Close association of RNA polymerase II and many transcription factors with Pol III genes. Proc Natl Acad Sci U S A *107*, 3639-3644.

Ram, O., Goren, A., Amit, I., Shoresh, N., Yosef, N., Ernst, J., Kellis, M., Gymrek, M., Issner, R., Coyne, M.*, et al.* (2011). Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. Cell *147*, 1628-1639.

Ramirez, E., Fernandez, J., Cartier, L., Villota, C., and Rios, M. (2003). Defective human T-cell lymphotropic virus type I (HTLV-I) provirus in seronegative tropical spastic paraparesis/HTLV-I-associated myelopathy (TSP/HAM) patients. Virus Res *91*, 231-239.

Ramos J.C., T., N., Diaz L, Ruiz P., Barber G, and Harrington W (2011). Targeting HTLV-I latency in Adult T-cell Leukemia/Lymphoma. Retrovirology *8*, A48.

Rende, F., Cavallari, I., Corradin, A., Silic-Benussi, M., Toulza, F., Toffolo, G.M., Tanaka, Y., Jacobson, S., Taylor, G.P., D'Agostino, D.M.*, et al.* (2011). Kinetics and intracellular compartmentalization of HTLV-1 gene expression: nuclear retention of HBZ mRNAs. Blood *117*, 4855-4859.

Richardson, J.H., Edwards, A.J., Cruickshank, J.K., Rudge, P., and Dalgleish, A.G. (1990). In vivo cellular tropism of human T-cell leukemia virus type 1. J Virol *64*, 5682-5687.

Ronen, K., Negre, O., Roth, S., Colomb, C., Malani, N., Denaro, M., Brady, T., Fusil, F., Gillet-Legrand, B., Hehir, K.*, et al.* (2011). Distribution of lentiviral vector integration sites in mice following therapeutic gene transfer to treat beta-thalassemia. Mol Ther *19*, 1273-1286.

Rosenberg, N., and Jolicoeur, P. (1997). Retroviral Pathogenesis, in: Coffin, J., Hughes, S., Varmus, H., editors. Retroviruses, Cold Spring Harbor Lab Press, 1997.

Rowan, A.G., and Bangham, C.R. (2012). Is There a Role for HTLV-1-Specific CTL in Adult T-Cell Leukemia/Lymphoma? Leuk Res Treatment *2012*, 391953.

Rozowsky, J., Abyzov, A., Wang, J., Alves, P., Raha, D., Harmanci, A., Leng, J., Bjornson, R., Kong, Y., Kitabayashi, N.*, et al.* (2011). AlleleSeq: analysis of allele-specific expression and binding in a network framework. Mol Syst Biol *7*, 522.

Sabouri, A.H., Saito, M., Lloyd, A.L., Vine, A.M., Witkover, A.W., Furukawa, Y., Izumo, S., Arimura, K., Marshall, S.E., Usuku, K.*, et al.* (2004). Polymorphism in the interleukin-10 promoter affects both provirus load and the risk of human T lymphotropic virus type I-associated myelopathy/tropical spastic paraparesis. J Infect Dis *190*, 1279-1285.

Sadelain, M., Papapetrou, E.P., and Bushman, F.D. (2012). Safe harbours for the integration of new DNA in the human genome. Nat Rev Cancer *12*, 51-58.

Sajesh, B.V., Lichtensztejn, Z., and McManus, K.J. (2013). Sister chromatid cohesion defects are associated with chromosome instability in Hodgkin lymphoma cells. BMC Cancer *13*, 391.

Sasaki, D., Doi, Y., Hasegawa, H., Yanagihara, K., Tsukasaki, K., Iwanaga, M., Yamada, Y., Watanabe, T., and Kamihira, S. (2010). High human T cell leukemia virus type-1(HTLV-1) provirus load in patients with HTLV-1 carriers complicated with HTLV-1-unrelated disorders. Virol J *7*, 81.

Sasaki, H., Nishikata, I., Shiraga, T., Akamatsu, E., Fukami, T., Hidaka, T., Kubuki, Y., Okayama, A., Hamada, K., Okabe, H.*, et al.* (2005). Overexpression of a cell adhesion molecule, TSLC1, as a possible molecular marker for acute-type adult T-cell leukemia. Blood *105*, 1204-1213.

Satou, Y., and Matsuoka, M. (2007). Implication of the HTLV-I bZIP factor gene in the leukemogenesis of adult T-cell leukemia. Int J Hematol *86*, 107-112.

Satou, Y., Yasunaga, J., Yoshida, M., and Matsuoka, M. (2006). HTLV-I basic leucine zipper factor gene mRNA supports proliferation of adult T cell leukemia cells. Proc Natl Acad Sci U S A *103*, 720-725.

Satou, Y., Yasunaga, J., Zhao, T., Yoshida, M., Miyazato, P., Takai, K., Shimizu, K., Ohshima, K., Green, P.L., Ohkura, N.*, et al.* (2011). HTLV-1 bZIP factor induces T-cell lymphoma and systemic inflammation in vivo. PLoS Pathog *7*, e1001274.

Schneider, U., Schwenk, H.U., and Bornkamm, G. (1977). Characterization of EBV-genome negative "null" and "T" cell lines derived from children with acute lymphoblastic leukemia and leukemic transformed non-Hodgkin lymphoma. Int J Cancer *19*, 621-626.

Schroder, A.R., Shinn, P., Chen, H., Berry, C., Ecker, J.R., and Bushman, F. (2002). HIV-1 integration in the human genome favors active genes and local hotspots. Cell *110*, 521-529.

Seich Al Basatena, N.K., Macnamara, A., Vine, A.M., Thio, C.L., Astemborski, J., Usuku, K., Osame, M., Kirk, G.D., Donfield, S.M., Goedert, J.J.*, et al.* (2011). KIR2DL2 enhances protective and detrimental HLA class I-mediated immunity in chronic viral infection. PLoS Pathog *7*, e1002270.

Seiki, M., Eddy, R., Shows, T.B., and Yoshida, M. (1984). Nonspecific integration of the HTLV provirus genome into adult T-cell leukaemia cells. Nature *309*, 640-642.

Seiki, M., Hattori, S., Hirayama, Y., and Yoshida, M. (1983). Human adult T-cell leukemia virus: complete nucleotide sequence of the provirus genome integrated in leukemia cell DNA. Proc Natl Acad Sci U S A *80*, 3618-3622.

Seiki, M., Inoue, J., Hidaka, M., and Yoshida, M. (1988). Two cis-acting elements responsible for posttranscriptional trans-regulation of gene expression of human T-cell leukemia virus type I. Proc Natl Acad Sci U S A *85*, 7124-7128.

Seiki, M., Inoue, J., Takeda, T., and Yoshida, M. (1986). Direct evidence that p40x of human T-cell leukemia virus type I is a trans-acting transcriptional activator. EMBO J *5*, 561-565.

Shan, L., Yang, H.C., Rabi, S.A., Bravo, H.C., Shroff, N.S., Irizarry, R.A., Zhang, H., Margolick, J.B., Siliciano, J.D., and Siliciano, R.F. (2011). Influence of host gene transcription level and orientation on HIV-1 latency in a primary-cell model. J Virol *85*, 5384-5393.

Shembade, N., and Harhaj, E.W. (2010). Role of post-translational modifications of HTLV-1 Tax in NF-kappaB activation. World J Biol Chem *1*, 13-20.

Shimizu, Y., Takamori, A., Utsunomiya, A., Kurimura, M., Yamano, Y., Hishizawa, M., Hasegawa, A., Kondo, F., Kurihara, K., Harashima, N.*, et al.* (2009). Impaired Tax-specific T-cell responses with insufficient control of HTLV-1 in a subgroup of individuals at asymptomatic and smoldering stages. Cancer Sci *100*, 481-489.

Shimoyama, M. (1991). Diagnostic criteria and classification of clinical subtypes of adult T-cell leukaemia-lymphoma. A report from the Lymphoma Study Group (1984-87). Br J Haematol *79*, 428-437.

Siliciano, R.F., and Greene, W.C. (2011). HIV latency. Cold Spring Harb Perspect Med *1*, a007096.

Sjoblom, T., Jones, S., Wood, L.D., Parsons, D.W., Lin, J., Barber, T.D., Mandelker, D., Leary, R.J., Ptak, J., Silliman, N.*, et al.* (2006). The consensus coding sequences of human breast and colorectal cancers. Science *314*, 268-274.

Slattery, J.P., Franchini, G., and Gessain, A. (1999). Genomic evolution, patterns of global dissemination, and interspecies transmission of human and simian T-cell leukemia/lymphotropic viruses. Genome Res *9*, 525-540.

Souza, H.C., Carvalho, B.N., Morais, M.G., Monteiro, G.Z., Emori, F.T., and Latorre, L.C. (2011). Tropical pyomyositis in a patient with systemic lupus erythematosus and HTLV 1/2 infection. Rev Bras Reumatol *51*, 97-103.

Suarez F, M.A., Ghez D, Delarue R, Deau-Fischer B, Flore Sicre de Fontbrune C.A, Ysebaert L, Asnafi V, Canioni D, deThe H, Bazarbachi A and Hermine O (2011). Arsenic trioxyde in the treatment of HTLV1 associated ATLL. Retrovirology *8*, A59

Suemori, K., Fujiwara, H., Ochi, T., Ogawa, T., Matsuoka, M., Matsumoto, T., Mesnard, J.M., and Yasukawa, M. (2009). HBZ is an immunogenic protein, but not a target antigen for human T-cell leukemia virus type 1-specific cytotoxic T lymphocytes. J Gen Virol *90*, 1806-1811.

Suzuki, S., Uozumi, K., Maeda, M., Yamasuji, Y., Hashimoto, S., Komorizono, Y., Owatari, S., Tokunaga, M., Haraguchi, K., and Arima, N. (2006). Adult T-cell leukemia in a liver transplant recipient that did not progress after onset of graft rejection. Int J Hematol *83*, 429-432.

Suzuki, T., Shen, H., Akagi, K., Morse, H.C., Malley, J.D., Naiman, D.Q., Jenkins, N.A., and Copeland, N.G. (2002). New genes involved in cancer identified by retroviral tagging. Nat Genet *32*, 166-174.

Suzuki, T., Uchida-Toita, M., and Yoshida, M. (1999). Tax protein of HTLV-1 inhibits CBP/p300-mediated transcription by interfering with recruitment of CBP/p300 onto DNA element of E-box or p53 binding site. Oncogene *18*, 4137-4143.

Takasaki, Y., Iwanaga, M., Imaizumi, Y., Tawara, M., Joh, T., Kohno, T., Yamada, Y., Kamihira, S., Ikeda, S., Miyazaki, Y*., et al.* (2010). Long-term study of indolent adult T-cell leukemia-lymphoma. Blood *115*, 4337-4343.

Takeda, S., Maeda, M., Morikawa, S., Taniguchi, Y., Yasunaga, J., Nosaka, K., Tanaka, Y., and Matsuoka, M. (2004). Genetic and epigenetic inactivation of tax gene in adult T-cell leukemia cells. Int J Cancer *109*, 559-567.

Takemoto, S., Matsuoka, M., Yamaguchi, K., and Takatsuki, K. (1994). A novel diagnostic method of adult T-cell leukemia: monoclonal integration of human T-cell lymphotropic virus type I provirus DNA detected by inverse polymerase chain reaction. Blood *84*, 3080-3085.

Takenouchi, N., Yamano, Y., Usuku, K., Osame, M., and Izumo, S. (2003). Usefulness of proviral load measurement for monitoring of disease activity in individual patients with human T-lymphotropic virus type I-associated myelopathy/tropical spastic paraparesis. J Neurovirol *9*, 29-35.

Tamiya, S., Matsuoka, M., Etoh, K., Watanabe, T., Kamihira, S., Yamaguchi, K., and Takatsuki, K. (1996). Two types of defective human T-lymphotropic virus type I provirus in adult T-cell leukemia. Blood *88*, 3065-3073.

Taniguchi, Y., Nosaka, K., Yasunaga, J., Maeda, M., Mueller, N., Okayama, A., and Matsuoka, M. (2005). Silencing of human T-cell leukemia virus type I gene transcription by epigenetic mechanisms. Retrovirology *2*, 64.

Tateno, M., Kondo, N., Itoh, T., Chubachi, T., Togashi, T., and Yoshiki, T. (1984). Rat lymphoid cell lines with human T cell leukemia virus production. I. Biological and serological characterization. The Journal of experimental medicine *159*, 1105-1116.

Taylor, G.P., Bodeus, M., Courtois, F., Pauli, G., Del Mistro, A., Machuca, A., Padua, E., Andersson, S., Goubau, P., Chieco-Bianchi, L*., et al.* (2005). The seroepidemiology of human T-lymphotropic viruses: types I and II in Europe: a prospective study of pregnant women. J Acquir Immune Defic Syndr *38*, 104-109.

Taylor, G.P., Goon, P., Furukawa, Y., Green, H., Barfield, A., Mosley, A., Nose, H., Babiker, A., Rudge, P., Usuku, K*., et al.* (2006). Zidovudine plus lamivudine in Human T-Lymphotropic Virus type-I-associated myelopathy: a randomised trial. Retrovirology *3*, 63.

Tie, F., Adya, N., Greene, W.C., and Giam, C.Z. (1996). Interaction of the human T-lymphotropic virus type 1 Tax dimer with CREB and the viral 21-base-pair repeat. J Virol *70*, 8368-8374.

Tosswill, J.H., Taylor, G.P., Tedder, R.S., and Mortimer, P.P. (2000). HTLV-I/II associated disease in England and Wales, 1993-7: retrospective review of serology requests. BMJ *320*, 611-612.

Toulza, F., Heaps, A., Tanaka, Y., Taylor, G.P., and Bangham, C.R. (2008). High frequency of CD4+FoxP3+ cells in HTLV-1 infection: inverse correlation with HTLV-1-specific CTL response. Blood *111*, 5047-5053.

Toulza, F., Nosaka, K., Takiguchi, M., Pagliuca, T., Mitsuya, H., Tanaka, Y., Taylor, G.P., and Bangham, C.R. (2009). FoxP3+ regulatory T cells are distinct from leukemia cells in HTLV-1-associated adult T-cell leukemia. Int J Cancer *125*, 2375-2382.

Toulza, F., Nosaka, K., Tanaka, Y., Schioppa, T., Balkwill, F., Taylor, G.P., and Bangham, C.R. (2010). Human T-lymphotropic virus type 1-induced CC chemokine ligand 22 maintains a high frequency of functional FoxP3+ regulatory T cells. J Immunol *185*, 183-189.

Trojer, P., Cao, A.R., Gao, Z., Li, Y., Zhang, J., Xu, X., Li, G., Losson, R., Erdjument-Bromage, H., Tempst, P*., et al.* (2011). L3MBTL2 protein acts in concert with PcG protein-mediated monoubiquitination of H2A to establish a repressive chromatin structure. Mol Cell *42*, 438-450.

Tsujimoto, H., Noda, Y., Ishikawa, K., Nakamura, H., Fukasawa, M., Sakakibara, I., Sasagawa, A., Honjo, S., and Hayami, M. (1987). Development of adult T-cell leukemia-like disease in African green monkey associated with clonal integration of simian T-cell leukemia virus type I. Cancer Res *47*, 269-274.

Tsukasaki, K., Hermine, O., Bazarbachi, A., Ratner, L., Ramos, J.C., Harrington, W., Jr., O'Mahony, D., Janik, J.E., Bittencourt, A.L., Taylor, G.P*., et al.* (2009). Definition, prognostic factors, treatment, and response criteria of adult T-cell leukemia-lymphoma: a proposal from an international consensus meeting. J Clin Oncol *27*, 453-459.

Tsukasaki, K., Tsushima, H., Yamamura, M., Hata, T., Murata, K., Maeda, T., Atogami, S., Sohda, H., Momita, S., Ideda, S*., et al.* (1997). Integration patterns of HTLV-I provirus in relation to the clinical course of ATL: frequent clonal change at crisis from indolent disease. Blood *89*, 948-956.

Tsukasaki, K., Utsunomiya, A., Fukuda, H., Shibata, T., Fukushima, T., Takatsuka, Y., Ikeda, S., Masuda, M., Nagoshi, H., Ueda, R*., et al.* (2007). VCAP-AMP-VECP compared with biweekly CHOP for

adult T-cell leukemia-lymphoma: Japan Clinical Oncology Group Study JCOG9801. J Clin Oncol *25*, 5458-5464.

Uchiyama, T. (1997). Human T cell leukemia virus type I (HTLV-I) and human diseases. Annu Rev Immunol *15*, 15-37.

Uchiyama, T., Hori, T., Tsudo, M., Wano, Y., Umadome, H., Tamori, S., Yodoi, J., Maeda, M., Sawami, H., and Uchino, H. (1985). Interleukin-2 receptor (Tac antigen) expressed on adult T cell leukemia cells. J Clin Invest *76*, 446-453.

Umino, A., and Seto, M. (2013). Array CGH reveals clonal evolution of adult T-cell leukemia/lymphoma. Methods Mol Biol *973*, 189-196.

Uren, A.G., Kool, J., Berns, A., and van Lohuizen, M. (2005). Retroviral insertional mutagenesis: past, present and future. Oncogene *24*, 7656-7672.

Ureta-Vidal, A., Angelin-Duclos, C., Tortevoye, P., Murphy, E., Lepere, J.F., Buigues, R.P., Jolly, N., Joubert, M., Carles, G., Pouliquen, J.F*., et al.* (1999). Mother-to-child transmission of human T-cell-leukemia/lymphoma virus type I: implication of high antiviral antibody titer and high proviral load in carrier mothers. Int J Cancer *82*, 832-836.

Usui, T., Yanagihara, K., Tsukasaki, K., Murata, K., Hasegawa, H., Yamada, Y., and Kamihira, S. (2008). Characteristic expression of HTLV-1 basic zipper factor (HBZ) transcripts in HTLV-1 provirus-positive cells. Retrovirology *5*, 34.

Van Beveren, C., Rands, E., Chattopadhyay, S.K., Lowy, D.R., and Verma, I.M. (1982). Long terminal repeat of murine retroviral DNAs: sequence analysis, host-proviral junctions, and preintegration site. J Virol *41*, 542-556.

van Dongen, J.J., Langerak, A.W., Bruggemann, M., Evans, P.A., Hummel, M., Lavender, F.L., Delabesse, E., Davi, F., Schuuring, E., Garcia-Sanz, R*., et al.* (2003). Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. Leukemia *17*, 2257-2317.

van Koningsbruggen, S., Gierlinski, M., Schofield, P., Martin, D., Barton, G.J., Ariyurek, Y., den Dunnen, J.T., and Lamond, A.I. (2010). High-resolution whole-genome sequencing reveals that specific chromatin domains from most human chromosomes associate with nucleoli. Mol Biol Cell *21*, 3735-3748.

Van Prooyen, N., Gold, H., Andresen, V., Schwartz, O., Jones, K., Ruscetti, F., Lockett, S., Gudla, P., Venzon, D., and Franchini, G. (2010). Human T-cell leukemia virus type 1 p8 protein increases cellular conduits and virus transmission. Proc Natl Acad Sci U S A *107*, 20738-20743.

Verdonck, K., Gonzalez, E., Van Dooren, S., Vandamme, A.M., Vanham, G., and Gotuzzo, E. (2007). Human T-lymphotropic virus 1: recent knowledge about an ancient infection. Lancet Infect Dis *7*, 266-281.

Vernant, J.C., Buisson, G.G., Sobesky, G., Arfi, S., Gervaise, G., and Roman, G.C. (1987). Can HTLV-1 lead to immunological disease? Lancet *2*, 404.

Villaudy, J., Wencker, M., Gadot, N., Gillet, N.A., Scoazec, J.Y., Gazzolo, L., Manz, M.G., Bangham, C.R., and Dodon, M.D. (2011). HTLV-1 propels thymic human T cell development in "human immune system" Rag2(-)/(-) gamma c(-)/(-) mice. PLoS Pathog *7*, e1002231.

Vincent, K.A., York-Higgins, D., Quiroga, M., and Brown, P.O. (1990). Host sequences flanking the HIV provirus. Nucleic Acids Res *18*, 6045-6047.

Vine, A.M., Witkover, A.D., Lloyd, A.L., Jeffery, K.J., Siddiqui, A., Marshall, S.E., Bunce, M., Eiraku, N., Izumo, S., Usuku, K.*, et al.* (2002). Polygenic control of human T lymphotropic virus type I (HTLV-I) provirus load and the risk of HTLV-I-associated myelopathy/tropical spastic paraparesis. J Infect Dis *186*, 932-939.

Voevodin, A., Samilchuk, E., Schatzl, H., Boeri, E., and Franchini, G. (1996). Interspecies transmission of macaque simian T-cell leukemia/lymphoma virus type 1 in baboons resulted in an outbreak of malignant lymphoma. J Virol *70*, 1633-1639.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Jr., and Kinzler, K.W. (2013). Cancer genome landscapes. Science *339*, 1546-1558.

Volsky, D.J., Simm, M., Shahabuddin, M., Li, G., Chao, W., and Potash, M.J. (1996). Interference to human immunodeficiency virus type 1 infection in the absence of downmodulation of the principal virus receptor, CD4. J Virol *70*, 3823-3833.

Wang, G.P., Garrigue, A., Ciuffi, A., Ronen, K., Leipzig, J., Berry, C., Lagresle-Peyrou, C., Benjelloun, F., Hacein-Bey-Abina, S., Fischer, A.*, et al.* (2008). DNA bar coding and pyrosequencing to analyze adverse events in therapeutic gene transfer. Nucleic Acids Res *36*, e49.

Wang, Z., Zang, C., Cui, K., Schones, D.E., Barski, A., Peng, W., and Zhao, K. (2009). Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. Cell *138*, 1019-1031.

Watanabe, T., Seiki, M., Tsujimoto, H., Miyoshi, I., Hayami, M., and Yoshida, M. (1985). Sequence homology of the simian retrovirus genome with human T-cell leukemia virus type I. Virology *144*, 59-65.

Wattel, E., Vartanian, J.P., Pannetier, C., and Wain-Hobson, S. (1995). Clonal expansion of human T-cell leukemia virus type I-infected cells in asymptomatic and symptomatic carriers without malignancy. J Virol *69*, 2863-2868.

Williams-Carrier, R., Stiffler, N., Belcher, S., Kroeger, T., Stern, D.B., Monde, R.A., Coalter, R., and Barkan, A. (2010). Use of Illumina sequencing to identify transposon insertions underlying mutant phenotypes in high-copy Mutator lines of maize. Plant J *63*, 167-177.

Wolfe, N.D., Heneine, W., Carr, J.K., Garcia, A.D., Shanmugam, V., Tamoufe, U., Torimiro, J.N., Prosser, A.T., Lebreton, M., Mpoudi-Ngole, E.*, et al.* (2005). Emergence of unique primate T-lymphotropic viruses among central African bushmeat hunters. Proc Natl Acad Sci U S A *102*, 7994-7999.

Wu, X., Li, Y., Crise, B., and Burgess, S.M. (2003). Transcription start regions in the human genome are favored targets for MLV integration. Science *300*, 1749-1751.

Wu, X., Li, Y., Crise, B., Burgess, S.M., and Munroe, D.J. (2005). Weak palindromic consensus sequences are a common feature found at the integration target sites of many retroviruses. J Virol *79*, 5211-5214.

Wu, X., Luke, B.T., and Burgess, S.M. (2006). Redefining the common insertion site. Virology *344*, 292-295.

Yamada, Y., Tomonaga, M., Fukuda, H., Hanada, S., Utsunomiya, A., Tara, M., Sano, M., Ikeda, S., Takatsuki, K., Kozuru, M.*, et al.* (2001). A new G-CSF-supported combination chemotherapy, LSG15, for adult T-cell leukaemia-lymphoma: Japan Clinical Oncology Group Study 9303. Br J Haematol *113*, 375-382.

Yamaguchi, K., and Watanabe, T. (2002). Human T lymphotropic virus type-I and adult T-cell leukemia in Japan. Int J Hematol *76 Suppl 2*, 240-245.

Yamamoto, N., Okada, M., Koyanagi, Y., Kannagi, M., and Hinuma, Y. (1982). Transformation of human leukocytes by cocultivation with an adult T cell leukemia virus producer cell line. Science *217*, 737-739.

Yoshida, M. (2001). Multiple viral strategies of HTLV-1 for dysregulation of cell growth control. Annu Rev Immunol *19*, 475-496.

Yoshida, M., Miyoshi, I., and Hinuma, Y. (1982). Isolation and characterization of retrovirus from cell lines of human adult T-cell leukemia and its implication in the disease. Proc Natl Acad Sci U S A *79*, 2031-2035.

Yoshida, M., Seiki, M., Yamaguchi, K., and Takatsuki, K. (1984). Monoclonal integration of human T-cell leukemia provirus in all primary tumors of adult T-cell leukemia suggests causative role of human T-cell leukemia virus in the disease. Proc Natl Acad Sci U S A *81*, 2534-2537.

Zhang, W., Nisbet, J.W., Bartoe, J.T., Ding, W., and Lairmore, M.D. (2000). Human T-lymphotropic virus type 1 p30(II) functions as a transcription factor and differentially modulates CREB-responsive promoters. J Virol *74*, 11270-11277.

Appendices

Appendix 1: Datasets used in this work

Annotation	**Reference**
RefSeq genes	(Fujita et al., 2011)
CpG islands	(Fujita et al., 2011)

'AllOnco' Cancer gene database (Available from The Bushman Lab cancer gene list website),(Sadelain et al., 2012)

Gene set	Total number of Genes	7 Reference databases drawn together to compile **AllOnco** by Bushman lab
AllOnco	2070	(Huret et al., 2000)
		(Rosenberg and Jolicoeur, 1997)
		(Sjoblom et al., 2006)
		(Akagi et al., 2004)
		(Futreal et al., 2004)
		Cavazzana-Calvo and colleagues, Hopital Necker, Paris, France
		Waldman cancer gene database

Annotation	Mark	Cell type	Reference
Activatory epigenetic marks	H2BK120ac	Primary CD4+ T-cells	(Barski et al., 2007)
	H2BK5ac	Primary CD4+ T-cells	(Barski et al., 2007)
	H2BK20ac	Primary CD4+ T-cells	(Barski et al., 2007)
	H3K18ac	Primary CD4+ T-cells	(Barski et al., 2007)
	H3K27ac	Primary CD4+ T-cells	(Barski et al., 2007)
	H3K4ac	Primary CD4+ T-cells	(Barski et al., 2007)
	H4K5ac	Primary CD4+ T-cells	(Barski et al., 2007)
	H4K8ac	Primary CD4+ T-cells	(Barski et al., 2007)
	H4K91ac	Primary CD4+ T-cells	(Barski et al., 2007)
Inhibitory epigenetic marks	H3K9me2	Primary CD4+ T-cells	(Barski et al., 2007)
	H3K9me3	Primary CD4+ T-cells	(Barski et al., 2007)
	H4K20me3	Primary CD4+ T-cells	(Barski et al., 2007)

Annotation	Cell type	Reference
BAF155	HeLa	(Euskirchen et al., 2011)
BAF170	HeLa	(Euskirchen et al., 2011)
BRG1	HeLa	(Euskirchen et al., 2011)
CBP	CD4	(Wang et al., 2009)
cFos	GM12878	(Rozowsky et al., 2011)
cJun	GM12878	(Raha et al., 2010)
cMyc	GM12878	(Rozowsky et al., 2011)
CTCF	Primary CD4+ T-cells	(Barski et al., 2007; Jothi et al., 2008)
E2F4	GM06990	(Lee et al., 2011)
E2F6	K562	(Trojer et al., 2011)
EZH2	K562	Bernstein lab for the ENCODE project
FOXP3	Activated CD4	(Birzele et al., 2011)
FOXP3	Treg	(Birzele et al., 2011)
GATA1	K562	(Fujiwara et al., 2009)
GATA2	K562	(Fujiwara et al., 2009)
HDAC1	Primary CD4+ T-cells	(Wang et al., 2009)
HDAC2	Primary CD4+ T-cells	(Wang et al., 2009)
HDAC3	Primary CD4+ T-cells	(Wang et al., 2009)
HDAC6	Primary CD4+ T-cells	(Wang et al., 2009)
Ini1	HeLa	(Euskirchen et al., 2011)
IRF1	K562	Struhl lab for the ENCODE project
JunD	GM12878	(Rozowsky et al., 2011)
MEF2C	GM12878	Myers lab for the ENCODE project
MOF	Primary CD4+ T-cells	(Wang et al., 2009)
Myb	HepG2	Myers lab for the ENCODE project

NFkB	GM12878	(Kasowski et al., 2010)
NRSF	Jurkat	(Johnson et al., 2007; Jothi et al., 2008)
P300	Primary CD4+ T-cells	(Wang et al., 2009)
p53	IMR90	(Botcheva et al., 2011)
PCAF	Primary CD4+ T-cells	(Wang et al., 2009)
PML	K562	Myers lab for the ENCODE project
Rad21	GM12878	Myers lab for the ENCODE project
STAT1	Primary CD4+ T-cells	(Liao et al., 2011)
STAT1 IFN	Interferon stimulated CD4+ T-cells	(Liao et al., 2011)
STAT5a	K562	Myers lab for the ENCODE project
SUZ12	K562	(Ram et al., 2011)
TAL1	K562	Struhl lab for the ENCODE project
TCF7	HeLa	Struhl lab for the ENCODE project
Tip60	Primary CD4+ T-cells	(Wang et al., 2009)
Yy1	K562	Myers lab for the ENCODE project
ZNF263	K562	(Frietze et al., 2010)

Jurkat: T-lymphoblastic cell line; GM06990: B-lymphoblastoid cell line; GM12878: Lymphoblastoid cell line; K562: erythroleukaemia cell line; HeLa: Cervical cancer cell line; IMR90: fetal lung fibroblast cell line; HepG2: Hepatocellular carcinoma cell line

Appendix 2: Summary results for each patient sample

ATL	Subtype	Age*	Sex*	PVL %	Proviral tax status	Number of proviruses in presumed malignant clone
ATL1	acute		F	82	Type 1 defective	1
ATL3	acute		M	60	Complete	1
ATL4	acute		F	66	Complete	1
ATL5	chronic		M	101	Hypermethylated 5'LTR	1
ATL7	acute		M	112	Complete	1
ATL8	acute		F	287	Complete	2
ATL9	acute		F	71	Type 2 defective	1
ATL10	chronic		M	153	Complete	1
ATL11	chronic		F	173	Hypermethylated 5'LTR	1
ATL13	lymphoma		F	52	Type 2 defective	1
ATL14	acute		M	247	Complete	2
ATL15	chronic		M	40	nonsense *tax* exon 3	1
ATL16	acute		M	33	Complete	1
ATL17	acute	52	F	75	Type 2 defective	1
ATL18	acute		M	35	Complete	1
ATL20	lymphoma	67	M	144	Complete	1
ATL21	lymphoma		F	77	Complete	1
ATL22	lymphoma		M	27	Complete	1
ATL25	acute	66	F	163	Complete	1
ATL26	chronic		F	116	Indeterminate defective	1
ATL27	lymphoma		F	20	Type 1 defective	1
ATL28	chronic	71	F	54	Complete	1
ATL29	acute	77	M	151	Complete	1
ATL30	chronic		F	90	Type 1 defective	1
ATL31	acute	54	F	56	nonsense *tax* exon 3	1
ATL32	lymphoma	61	F	72	Type 2 defective	1
ATL33	acute	73	M	77	Type 2 defective	1
ATL34	chronic		F	97	Complete	1
ATL35	acute	55	M	68	Hypermethylated 5'LTR	1
ATL36	acute	64	M	73	Complete	2
ATL37	chronic		F	69	Complete	1
ATL38	acute	77	M	84	Hypermethylated 5'LTR	1
ATL39	chronic	46	M	54	Type 1 defective	1
ATL40	lymphoma	65	M	25	Type 1 defective	1
ATL41	lymphoma	44	F	50	Indeterminate defective	1
ATL42	acute	45	M	52	Type 2 defective	1
ATL43	acute	43	F	39	Complete	1
ATL45	lymphoma		M	55	Type 1 defective	1

ATL	Subtype	Age	Sex	PVL %	Proviral tax status	Number of proviruses in presumed malignant clone
ATL46	acute		M	63	Type 1 defective	1
ATL47	acute		M	65	Complete	1
ATL49	acute		F	38	Complete	1
ATL50	acute		M	110	Type 1 defective	1
ATL51	acute		F	24	Type 2 defective	1
ATL52	acute	77	F	54	Hypermethylated 5'LTR	1
ATL53	lymphoma		F	55	Complete	1
ATL54	acute		M	92	Type 1 defective	1
ATL55	acute	49	F	55	Complete	1
ATL56	acute	56	M	71	Type 2 defective	1
ATL57	chronic	47	F	83	Complete	1
ATL58	acute	54	F	87	nonsense *tax* exon 3	1
ATL60	chronic		M	19	Type 2 defective	1
ATL62	chronic		M	36	Type 2 defective	1
ATL64	acute	37	M	13	Type 1 defective	1
ATL65	acute		F	55	Complete	1
ATL66	chronic		F	54	Complete	1
ATL69	acute		M	49	Complete	1
ATL70	lymphoma		M	143	Complete	1
ATL71	acute	63	M	162	Complete	1
ATL73	acute	64	F	9	Complete	2
ATL74	acute	64	M	11	Complete	1
ATL75	acute	70	F	21	Complete	1
ATL78	acute		F	103	Type 1 defective	2
ATL79	lymphoma		M	41	Type 1 defective	1
ATL81	acute	49	F	17	Complete	1
ATL82	acute	62	F	79	Type 2 defective	1
ATL83	lymphoma		M	0	Tax exon 3 deletion	1
ATL84	acute	47	M	120	Complete	1
ATL85	chronic		F	68	Type 1 defective	2
ATL86	acute	75	F	109	nonsense *tax* exon 3	1
ATL87	acute		M	159	Complete	1
ATL88	acute		M	85	Complete	1
ATL89	acute	53	M	105	missense *tax* exon 2	1
ATL90	acute		M	106	Hypermethylated 5'LTR	1
ATL91	acute		M	85	Type 1 defective	1
ATL92	acute		F	101	Type 2 defective	1
ATL93	chronic		M	107	Hypermethylated 5'LTR	1
ATL94	acute	33	F	79	Complete	1
ATL95	lymphoma	41	M	50	Complete	1

ATL	Subtype	Age	Sex	PVL %	Proviral tax status	Number of proviruses in presumed malignant clone
ATL96	acute	74	F	180	Complete	1
ATL97	acute	51	M	201	Complete	2
ATL99	lymphoma		M	78	Type 1 defective	1
ATL100	acute		F	62	Complete	1
ATL102	acute	76	F	71	Type 2 defective	1
ATL103	smoldering	78	M	13	Complete	1
ATL104	acute		M	40	Type 1 defective	1
ATL105	acute			116	Type 1 defective	1
ATL107	acute			96	Complete	1
ATL110	lymphoma			80	Type 2 defective	1
ATL112	lymphoma			10	Complete	1
ATL114	acute	67	F	51	Complete	1
ATL115	acute			166	Complete	1
ATL116	acute			1	Indeterminate defective	1
ATL117	chronic			116	nonsense *tax* exon 3	1
ATL118	lymphoma		F	58	Type 2 defective	1
ATL119	chronic	40	F	52	Type 1 defective	1
ATL122	lymphoma	76	F	53	Type 2 defective	1
ATL123	lymphoma		M	359	Complete	2
ATL125	acute	73	M	20	nonsense *tax* exon 3	1
ATL128	acute			309	Complete	1
ATL129	chronic		F	22	Type 2 defective	1
ATL130	acute		M	12	Type 2 defective	1
ATL131	acute		F	700	Type 1 defective	1
ATL132	lymphoma		F	40	Type 2 defective	1
ATL133	acute			12	Complete	2
ATL135	acute	53		103	Type 2 defective	1
ATL136	acute	55	F	152	Complete	1
ATL137	acute			49	Complete	1
ATL138	chronic	38	F	39	Type 1 defective	1
ATL139	acute	56	F	146	Complete	1
ATL141	acute	52	F	12	Type 1 defective	1
ATL142	chronic		F	88	Type 1 defective	1
ATL143	acute		M	195	Complete	1
ATL144	chronic	64	F	156	Complete	2
ATL145	acute			98	Type 2 defective	1
ATL146	acute	64		147	Type 1 defective	2
ATL147	acute		M	118	Complete	1
ATL148	lymphoma		F	170	Type 2 defective	1
ATL149	chronic		M	96	Indeterminate defective	1
ATL150	acute	50	F	116	Complete	1

ATL	Subtype	Age	Sex	PVL %	Proviral tax status	Number of proviruses in presumed malignant clone
ATL151	chronic	61	F	63	Complete	1
ATL152	acute	66	F	242	Complete	1
ATL153	chronic		M	73	Complete	1
ATL154	acute	44	M	80	Type 1 defective	1
ATL155	acute			158	Complete	1
ATL156	acute			86	Complete	1
ATL157	acute	57		126	Complete	1
ATL158	acute	68	F	97	Hypermethylated 5'LTR	1
ATL159	acute	71		7	Indeterminate defective	1
ATL161	acute	54	M	38	Complete	1
ATL162	acute	65	M	266	Type 1 defective	1
ATL163	acute	35	F	34	Type 2 defective	1
ATL164	acute	58	M	100	Hypermethylated 5'LTR	1
ATL165	acute		F	99	Indeterminate defective	1
ATL166	chronic	52		70	Hypermethylated 5'LTR	1
ATL167	acute			74	Type 2 defective	1
ATL168	acute	54	M	114	Type 2 defective	1
ATL169	acute	71	M	154	Indeterminate defective	1
ATL171	acute	52	F	94	Complete	1
ATL172	lymphoma		M	40	Indeterminate defective	1
ATL173	acute	79		75	Complete	1
ATL174	acute	47	F	15	Type 1 defective	1
ATL175	acute		M	43	Complete	1
ATL176	lymphoma			163	Type 2 defective	1
ATL177	lymphoma	55	M	21	Type 2 defective	1
ATL180	acute	43	F	124	Complete	1
ATL182	acute			152	Complete	1
ATL183	lymphoma	69		2	Type 1 defective	1
ATL185	acute	64	M	96	Nonsense exon 3 and missense exon 2	1
ATL186	chronic			69	nonsense *tax* exon 3	1
ATL187	acute			152	Complete	1
ATL189	acute	46	F	77	Complete	1
ATL190	lymphoma	73	F	56	Complete	1
ATL191	acute		F	84	Hypermethylated 5'LTR	1
ATL192	acute	49	M	65	Type 1 defective	1
ATL193	acute		F	71	Type 2 defective	1
ATL195	lymphoma			44	Complete	1
ATL196	lymphoma	73	M	75	Complete	1
ATL197	acute		M	154	Complete	1
ATL198	acute	39	M	178	Complete	2

ATL	Subtype	Age	Sex	PVL %	Proviral tax status	Number of proviruses in presumed malignant clone
ATL199	smoldering	79	M	174	Complete	2
ATL200	lymphoma			67	Type 2 defective	1
ATL201	acute			174	nonsense *tax* exon 3	1
ATL202	acute			148	Hypermethylated 5'LTR	1
ATL204	acute			2	Complete	1
ATL205	acute			96	Complete	1
ATL206	acute			65	Hypermethylated 5'LTR	1
ATL207	acute			5	nonsense *tax* exon 3	1
ATL208	acute			137	Complete	2
ATL209	acute			19	Type 2 defective	1
ATL210	acute			0	Indeterminate defective	1
ATL211	acute			120	Complete	1
ATL212	lymphoma			3	Type 2 defective	1
ATL215	smoldering			38	Hypermethylated 5'LTR	1
ATL217	acute			122	Complete	1
ATL218	acute			7	Complete	2
ATL220	smoldering			3	Complete	1
ATL221	acute			56	Complete	1
ATL222	acute			243	Type 1 defective	1
ATL223	acute			133	Complete	2
ATL224	acute			52	Hypermethylated 5'LTR	1
ATL225	acute			56	Complete	1
ATL226	acute			71	Hypermethylated 5'LTR	1
ATL227	acute			21	Type 1 defective	1
ATL229	acute			3	Type 2 defective	1
ATL230	smoldering			37	Complete	2
ATL231	chronic			28	nonsense *tax* exon 3	1
ATL232	smoldering			48	nonsense *tax* exon 3	1
ATL233	acute			35	Complete	2
ATL234	acute			6	Indeterminate defective	1
ATL235	acute			66	Complete	1
ATL236	chronic			27	Type 2 defective	1
ATL237	acute			72	Type 1 defective	1
ATL238	acute			132	Complete	1
ATL239	acute			264	Complete	1
ATL240	ATL			111	Complete	1
ATL241	ATL			111	Type 2 defective	1
ATL242	chronic			68	Type 2 defective	1

*Age and gender provided where known.

Appendix 3: Shimoyama classification of ATL

Table adapted from (Shimoyama, 1991)

	Smoldering	**Chronic**	**Lymphoma**	**Acute**
Anti-HTLV-1 antibody	Present	Present	Present	Present
Lymphocyte ($x10^9$)	<4	≥4	<4	Usually ≥ 4
Abnormal T-lymphocytes	≥5%	≥5%	≤1%	Usually ≥ 5%
Flower cells	Occasionally	Occasionally	No	Present
LDH	≤ 1.5 normal	≤ 2 normal	Often high, not essential	Often high, not essential
Corrected Ca (mmol/L)	Normal	Normal	Often high, not essential	Often high, not essential
Histology proven lymphadenopathy	No	Yes, not essential	Yes	Yes, not essential
Tumour lesion				
Skin	Yes, not essential	Yes, not essential	Yes, not essential	Yes, not essential
Lung	Yes, not essential	Yes, not essential	Yes, not essential	Yes, not essential
Lymph node	No	Yes, not essential	Yes	Yes, not essential
Liver	No	Yes, not essential	Yes, not essential	Yes, not essential
Spleen	No	Yes, not essential	Yes, not essential	Yes, not essential
CNS	No	No	Yes, not essential	Yes, not essential
Bone	No	No	Yes, not essential	Yes, not essential
Ascites	No	No	Yes, not essential	Yes, not essential
Pleural effusion	No	No	Yes, not essential	Yes, not essential
GI tract	No	No	Yes, not essential	Yes, not essential

Appendix 4: Samples not suitable for bioinformatic analysis

ATL Samples provided by Kyoto University	**Number of cases =242**
No evidence of HTLV-1 provirus	4
Not ATL	6 (3 AC, 3 B-cell lymphoma in HTLV-1 carriers)
Post treatment remission samples	9
Duplicate samples	2
LMPCR failed to amplify (no gel smear)	6
Flow cell sequencing primer binding site polymorphism	1
Poor sequencing (defined by sisters < 100)	7
Not uniquely mapped to host genome at both read1 and read2	10
Total number samples suitable for analysis	**Number of cases=197**

Appendix 5: Abstract of publication associated with this thesis

HTLV-1-infected T cells contain a single integrated provirus in natural infection,

Cook LB, Rowan AG, Melamed A, Taylor GP, Bangham CR

Blood 2012, 120(17):3488-3490

Human T lymphotropic virus type 1 (HTLV-1) appears to persist in the chronic phase of infection by driving oligoclonal proliferation of infected T cells. Our recent high-throughput sequencing study revealed a large number (often > 10^4) of distinct proviral integration sites of HTLV-1 in each host that is greatly in excess of previous estimates. Here we use the highly sensitive, quantitative high-throughput sequencing protocol to show that circulating HTLV-1^+ clones in natural infection each contain a single integrated proviral copy. We conclude that a typical host possesses a large number of distinct HTLV-1–infected T-cell clones.

Appendix 6: Summary of permission for third party copyright works

Page number	Type of work	Source work	Copyright holder	Permission to reuse
P29	Figure 1.1	Leukemia Research and Treatment, Volume 2012 (2012), Article ID 876153, 14 pages	© 2012 Francesca Rende et al	P - under Creative Commons Attribution License
P35	Figure 1.2	Cold Spring Harb Perspect Med 2012; 2:a006890	© 2012 Cold Spring Harbor Laboratory Press	P - permission granted (see email copied beneath).

FW: CSHL Press Reprint Permission Request Form

1 message

Brown, Carol <brown@cshl.edu> Thu, Jan 23, 2014 at 8:49 PM

To: "l.cook@imperial.ac.uk" <l.cook@imperial.ac.uk>

You have our permission to use Fig 1 on p.3 of the article detailed below in your PhD thesis. Please cite this article as Cold Spring Harb Perspect Med 2012;2:a006890, copyright to Cold Spring Harbor Laboratory Press.

Regards,

Carol C. Brown

Books Development, Marketing and Sales

Cold Spring Harbor Laboratory Press

500 Sunnyside Blvd.

Woodbury, NY 11797-2924

Tel: 516-422-4038

Fax: 516-422-4095

E-mail: brown@cshl.edu