Oncogene (2016) **35,** 2542–2546
© 2016 Macmillan Publishers Limited All rights reserved 0950-9232/16

www.nature.com/onc

SHORT COMMUNICATION

Onco-exaptation of an endogenous retroviral LTR drives *IRF5* expression in Hodgkin lymphoma

A Babaian[1,2], MT Romanish[1,2], L Gagnier[1,2], LY Kuo[1], MM Karimi[2,3], C Steidl[4,5] and DL Mager[1,2]

The transcription factor interferon regulatory factor 5 (IRF5) is upregulated in Hodgkin lymphoma (HL) and is a key regulator of the aberrant transcriptome characteristic of this disease. Here we show that *IRF5* upregulation in HL is driven by transcriptional activation of a normally dormant endogenous retroviral LOR1a long terminal repeat (LTR) upstream of *IRF5*. Specifically, through screening of RNA-sequencing libraries, we detected LTR-IRF5 chimeric transcripts in multiple HL cell lines but not in normal B-cell controls. In HL, the LTR was in an open and hypomethylated epigenetic state, and we further show the LTR is the site of transcriptional initiation. Among HL cell lines, usage of the LTR promoter strongly correlates with overall levels of *IRF5* mRNA and protein, indicating that LTR transcriptional awakening is a major contributor to IRF5 upregulation in HL. Taken together, oncogenic *IRF5* overexpression in HL is the result of a specific LTR transcriptional activation. We propose that such LTR derepression is a distinct mechanism of oncogene activation ('onco-exaptation'), and that such a mechanism warrants further investigation in molecular and cancer research.

Oncogene (2016) **35,** 2542–2546; doi:10.1038/onc.2015.308; published online 17 August 2015

INTRODUCTION

A major fraction of the human genome is the accumulated remnants of transposable elements (TEs).[1] Over evolutionary time, these sequences can contribute to functional regulatory or coding components of genes, a phenomenon termed exaptation.[2–5] Notably, growing evidence indicates that TEs, in particular the endogenous retroviruses, donate promoters and enhancers naturally found in their long terminal repeats (LTRs) to genes and regulatory networks.[2,4,6] One mechanism believed to be a central repressor of TE function is targeted DNA methylation.[7] The genomes of cancerous cells are frequently hypomethylated, especially at TE sequences,[8] thus creating a state in which these sequences may become transcriptionally active. We postulate that a process analogous to exaptation occurs during cancer evolution in which TEs are exploited as a means to promote oncogenesis, a process we term onco-exaptation.

To date, the most well-characterized example of such a phenomenon with a clear oncogenic effect has been reported in Hodgkin lymphoma (HL).[9] The malignant cells of HL, the Hodgkin and Reed-Sternberg cells, are predominantly of B-cell origin,[10] yet Hodgkin and Reed-Sternberg cells lose their B-cell character and take on a more plastic expression profile over their evolutionary course.[11] The HL oncogene, colony-stimulating factor 1 receptor (*CSF1R*), is natively restricted to the myeloid lineage but this restriction is subverted in HL through the transcriptional activation of a normally dormant endogenous retrovirus LTR as an alternative promoter.[9] *CSF1R* overexpression in HL is oncogenic and correlates with poor patient outcome.[11] The 'exaptation,' an LTR promoter, provides the means with which an otherwise fate-restricted proto-oncogene may be accessed. Similarly, in some patients with diffuse large B-cell lymphoma, an endogenous retrovirus LTR initiates transcription of the normally brain-expressed fatty acid-binding protein 7 (*FABP7)*. The chimeric LTR2-*FABP7* isoform produces a novel protein with altered functional characteristics and a role in diffuse large B-cell lymphoma cell proliferation.[12] These distinct examples share a mechanism of oncogene activation, but the overall prevalence and significance of this mechanism remains unknown.

To look for instances of TE-mediated gene activation in HL, we screened[12,13] HL cell line RNA sequencing (RNAseq) libraries. We noted that the proinflammatory transcription factor (TF) interferon regulatory factor 5 (*IRF5)* is recurrently upregulated in HL-derived cell lines and this is associated with LTR-*IRF5* chimeric transcripts consistent with onco-exaptation. IRF5 belongs to a multimember family of TFs responsible for inducing transcription of cytokines and chemokines in response to interferon signaling,[14] but had not been implicated in HL before. Enticingly, during our investigation an independent study of genome-wide DNase hypersensitivity data by Kreher *et al.*,[15] identified *IRF5* as being a key TF upregulated specifically in HL cells and crucial for their survival. Further, IRF5 cooperates with nuclear factor-κB as a central regulator of the HL transcriptome. Here we show that transcriptional activation of a normally dormant LTR has a significant role in the upregulation of IRF5 in HL. Hence, the HL-associated deregulation of at least two genes with major roles in this disease, *CSF1R* and *IRF5*, is mediated through the awakening of ancient LTR promoters.

[1]Terry Fox Laboratory, British Columbia Cancer Agency, Vancouver, British Columbia, Canada; [2]Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada; [3]Biomedical Research Centre, University of British Columbia, Vancouver, British Columbia, Canada; [4]Department of Lymphoid Cancer Research, British Columbia Cancer Agency, Vancouver, British Columbia, Canada and [5]Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, British Columbia, Canada. Correspondence: Dr C Steidl, Centre for Lymphoid Cancer, BC Cancer Agency, 675 West 10th Avenue, Vancouver, British Columbia, Canada V5Z1L3 or Dr DL Mager, Terry Fox Laboratory, BC Cancer Agency, 675 West 10th Avenue, Vancouver, British Columbia, Canada V5Z1L3.
E-mail: CSteidl@bccancer.bc.ca or dmager@bccrc.ca

Received 20 February 2015; revised 21 May 2015; accepted 13 July 2015; published online 17 August 2015

RESULTS AND DISCUSSION

To screen for TE-gene chimeric transcripts in HL, paired-end RNAseq reads were analyzed as previously described,[12] in nine HL, three primary mediastinal large B-cell lymphoma-derived cell lines[16] and nine normal CD77+ centroblast B-cell controls[17] (Supplementary Table S1). The screen identified a chimeric transcript from an LOR1a LTR upstream of *IRF5*, which was present in 7/9 HL lines (not detected in UH-O1 and the NPL-HL line, DEV), 1/3 PMBCL (MEDB1) and 0/9 B-cell samples (Figure 1, Supplementary Figure S1 and Supplementary Table S1).

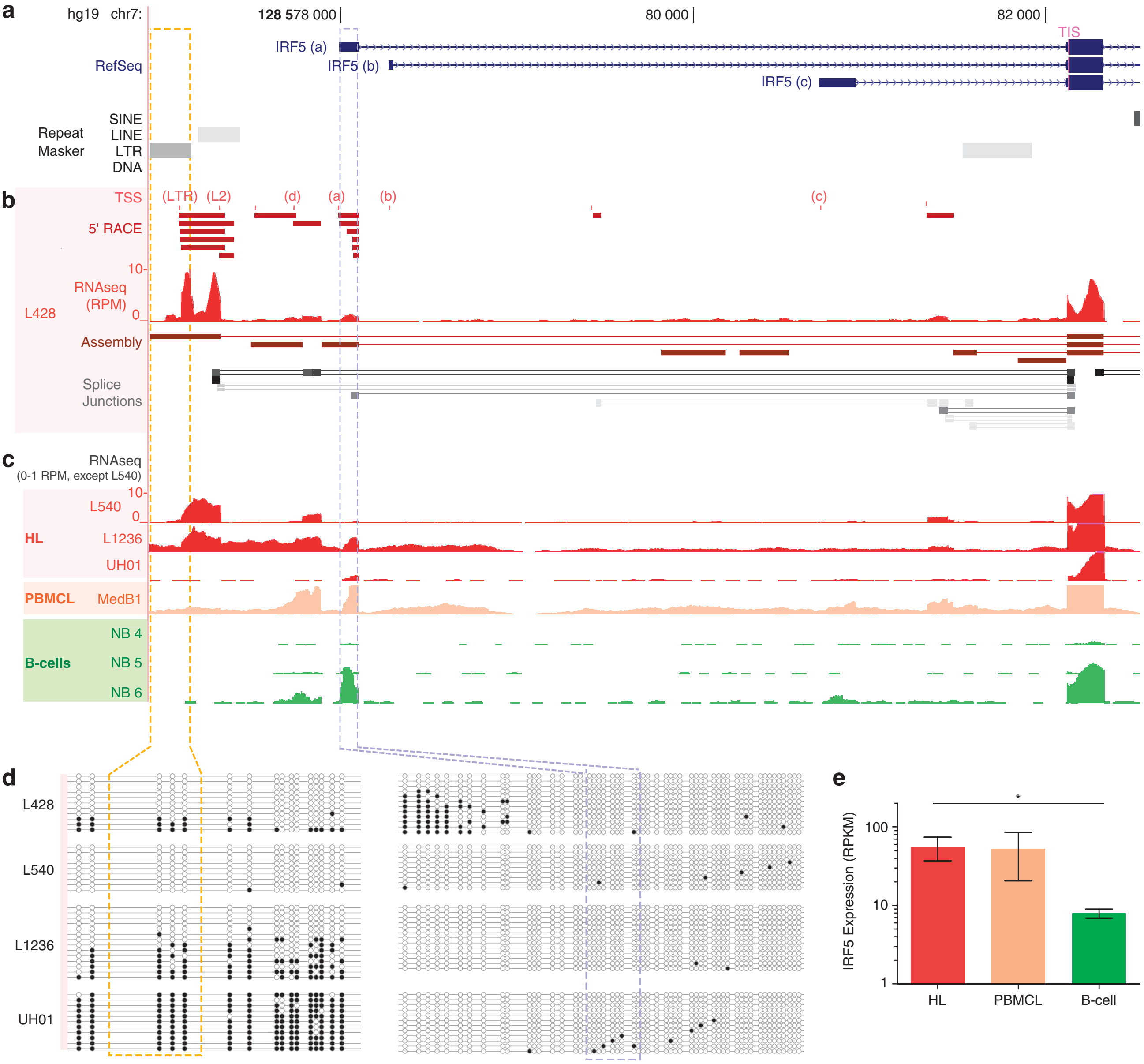

Figure 1. A LOR1a LTR element drives *IRF5* expression in HL. (**a**) A UCSC genome browser view of the 5′-end of RefSeq-annotated *IRF5*, RepeatMasker-defined TEs and *IRF5* transcription start sites (TSS) for native isoforms a–d[18] and LTR, and L2 isoforms described in this publication. The IRF5 translation initiation site (TIS) begins in the native exon 2. (**b**) RNAseq from the HL cell line L428,[16] aligned to the hg19 genome with Tophat 2.0.8 (Kim *et al.*[28]) using standard parameters. RNAseq coverage plot of uniquely mapped reads in reads per million (RPM) shows expression of upstream exons initiating within a LOR1a LTR, relative of the native *IRF5* transcripts. Unique first exons in L428 from 5′ rapid amplification of cDNA end (RACE) clones prepared using FirstChoice RLM kit (Ambion Life Technologies, Grand Island, NY, USA) kit with Superscript III (Invitrogen Life Technologies, Carlsbad, CA, USA) polymerase used for reverse transcription and Sanger sequencing (Eurofins MWG, Ebersberg, Germany). *Ab initio* RNA assembly track and splice-junction map created using Cufflinks v2.1.1.[29] Splice junctions are shaded by supporting reads from one (pale gray) to ⩾20 (black). (**c**) Representative HL RNAseq coverage scaled from 0 to 10 RPM for L540 and 0 to 1 RPM for L1236 and UH01,[16] showing a range of LTR promoter usage from high in L540 to absent in UH01. Representative primary mediastinal large B-cell lymphoma (PBMCL) line, MedB1 (orange) and normal B-cell transcriptomes (green)[17] predominantly transcribe *IRF5* from the native isoform a and d promoters but lack transcription from the LTR. Complete panel in Supplementary Figure S1. (**d**) Bisulphite sequencing (performed as previously described,[30]) of the LTR and native promoter regions with open circles showing unmethylated CpGs and solid circles showing methylated CpG sites. Cell lines with active LTRs are hypomethylated, whereas UH01 that uses the native promoter is hypermethylated. (**e**) Total expression of *IRF5* in HL ($n = 9$), PBMCL ($n = 3$) and B-cell ($n = 9$) RNAseq libraries calculated as reads per kilobase per million reads (RPKM).[29] Error bars are s.e.m. Two-tailed Welch's *t*-test was performed to test for difference in the means with unequal variance with *P-values = 0.0332 between HL and B cells. Primers for 5′ RACE and bisulphite sequencing in Supplementary Table S3.

To determine the tissue specificity of chimeric *IRF5*, we inspected ENCODE RNAseq data from 17 cell lines and 31 normal primary tissues, and no evidence for LOR1a-*IRF5* chimera was found, except for the three transformed B-cell lines GM12878, GM12891 and GM12892 (Supplementary Table S1). The absence of IRF5 chimera in primary tissues, in particular lymphocytes and leukocytes, suggests that the LOR1a LTR transcriptional activity is a transformed B-cell-specific and recurrently occurring phenomenon. Indeed, although several promoters for *IRF5* have been described in normal cells,[18] this LTR has not previously been detected as a promoter. The chimeric transcript contains the complete open reading frame for *IRF5*, which begins in native exon 2, and full-length chimeric *IRF5* cDNA could be PCR amplified (Figure 1a and Supplementary Text S1).

To determine whether the LTR element is truly the transcription initiation site of *IRF5* in HL, 5′ rapid amplification of cDNA ends was employed. In L428, transcription initiated from within the LOR1a LTR and at both the native 'a' and 'd' start sites,[18] as well as five other minor transcription start sites not previously characterized (Figure 1b and Supplementary Text S1). In UHO1, which is negative for the LTR chimera, only native start sites were found. The most prominent alternative transcription start site in L428 cells was in the LTR with the highest RNAseq coverage and split read (splice site) support (Figure 1a). In the normal B-cell RNAseq libraries, the LTR expression peaks were not detected (Figure 1c and Supplementary Figure S1), suggesting that transcription of the alternative chimeric isoform is acquired during the process of transformation from B-cell to HL.

To measure the relative contribution of the LTR and native promoters to total *IRF5* transcripts, two methods were employed. The proportion of reads aligning to the splice junctions of native exon 1 or LTR exon 1 and the second exon were measured (Figure 2a), the results of which closely matched quantitative reverse transcriptase–PCR (Supplementary Figure S2). The HL cell lines have 6.98-fold higher *IRF5* mRNA relative to B-cell controls (Figure 1e). Further, the presence of LTR transcript is associated with higher *IRF5* mRNA and IRF5 protein levels in HL (Figure 2). Together, our results show that this oncogene is overexpressed through the exploitation of a usually dormant LOR1a LTR.

To assess the epigenetic state of the LTR element between chimera-positive and -negative HL we investigated the methylation status of both the native and LTR promoters. In chimera-positive L428, L540 and L1236 cells the LOR1a LTR exists in a hypomethylated state, whereas in the chimera-negative UHO1 cells the LTR was hypermethylated (Figure 1d). The primary native promoter (start site 'a') exists within a CpG island and is unmethylated regardless of activity (Figure 1). By mapping the available DNase1 hypersensitivity data[15] of HL and non-HL cell lines, we observed that the hypomethylated LTR in L1236, L428 and L591 cells was within a DNase1 hypersensitive region, whereas it was not in the non-HL lines Namalwa and Reh (Figure 3a). Together, the absence of DNA methylation and open chromatin state suggests that this locus would be accessible to TFs and the transcriptional initiation machinery.

Little is known about the LOR1a family of LTRs beyond an entry in the repetitive sequence database, Repbase, that reports a consensus LTR sequence of 497 bp (Figure 3b).[19] The LOR1a LTR locus upstream of IRF5 is only 239 bp and the Repeatmasker annotation suggests it is missing the 5′-end. To investigate the LTR structure further, we retrieved the non-TE segment of 134 bp immediately 5′of the annotated LTR and looked for related sequences throughout the genome. Alignment of this upstream region to the hg19 human genome identified 34 homologous sequences of 69 bp upstream of other regions annotated as LOR1a (Supplementary Table S2), suggesting that the full LTR of this distinct subfamily is longer than annotated. Indeed, by examining the termini of these extended LTRs, we were able to identify putative 4 bp target site duplications, which are created on integration of retroviruses,[20] and therefore deduce the full length of these LOR1a subtypes, which is 308 bp for the copy upstream of *IRF5* (Figure 3b and Supplementary Table S2). Supplementary Figure S3 shows the full LTR sequence with transcription start sites defined by 5′ rapid amplification of cDNA ends and other features indicated. Evolutionary sequence comparisons indicate this LTR copy integrated at least 45–50 million years ago, as it is present in both New and Old World primates but is absent in non-primates (Figure 3d).

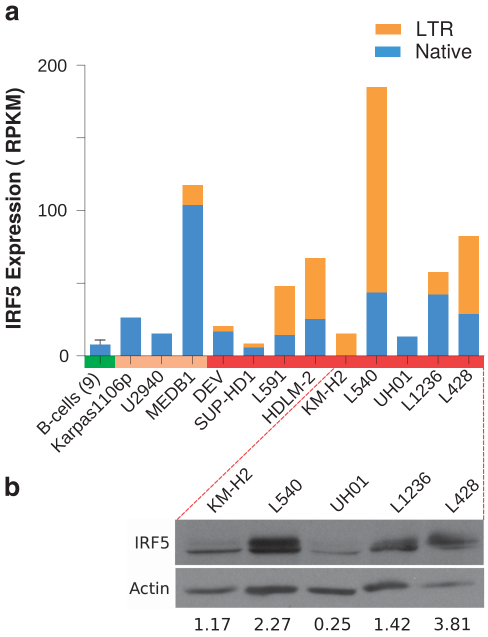

Figure 2. LTR contribution to *IRF5* mRNA levels and total protein. (**a**) Promoter contribution of LTR and native first exons to *IRF5* was calculated by defining all known *IRF5* exons from RefSeq, rapid amplification of cDNA ends and assembly, and creating a reference map of all possible splice junction combinations. RNAseq reads were then aligned using bowtie2[ref.31] to the splice junctions. The coverage at the splice junction for each promoter–exon pair was summed to measure the relative LTR:Native promoter contribution to overall expression measured in reads per kilobase per million (RPKM). Code is available at https://github.com/ababaian/Cypress. (**b**) Cell lysates (KM-H2, L540, UH01, L1236 and L428 all received from DSMZ, Braunschweig, Germany, and validated by karyotype and RNAseq) were prepared following RIPA (sodium deoxycholate 0.5%, Ipegal, 0.01%, SDS 10% in phosphate-buffered saline) lysis as previously described.[32] IRF5 and actin were detected with anti-IRF5 mouse monoclonal (Abnova, Taipei City, Taiwan; 2E3-1A11) and anti-actin rabbit polyclonal (Abcam, Cambridge, UK; ab8227) antibodies, respectively. Protein band-intensity quantification was performed with ImageJ software[33] and the ratio of IRF5 to actin is shown below each lane. Blotting was performed in duplicate.

Although no mention was made of the LTR, Mancl *et al.*[21] previously investigated the promoter activity of a region called 'P-V1' surrounding this LTR (Figure 3) and identified within it a critical interferon regulatory factor binding element (IRFE) that controls promoter activity in a luciferase reporter assay in response to various IRFs, in particular IRF5 itself. We identified the same IRFE using JaspScan[22] within this region and, intriguingly, found it to be located directly at the boundary of the LOR1a and the target site duplication, such that the IRFE site contains the target site duplication and first few bases of the LTR (Figure 3 and Supplementary Figure S3). This TF-binding site was therefore created serendipitously millions of years ago when the LOR1a element retrotransposed. Hence, the inherent core promoter motifs of an LTR plus the formation of an IRFE site

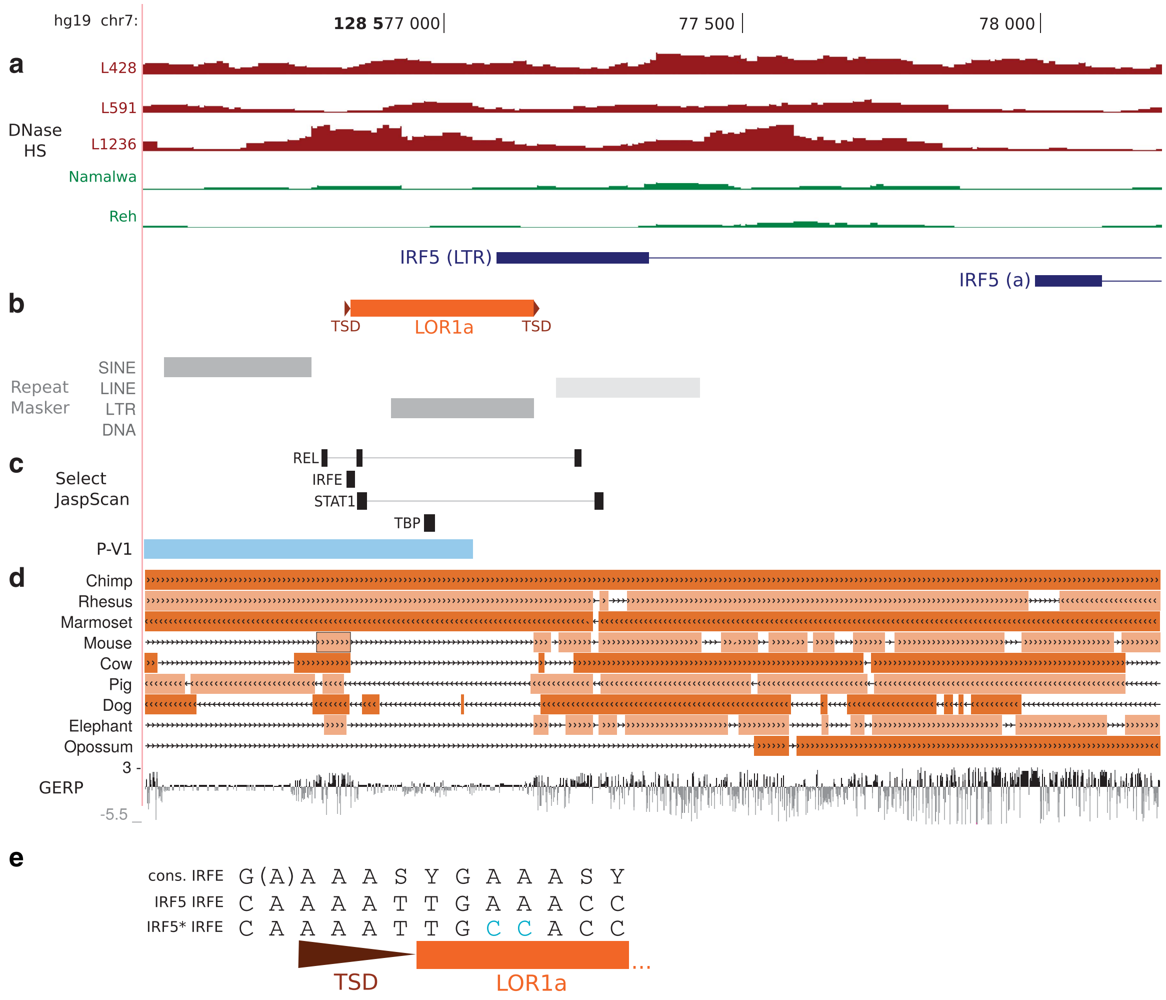

Figure 3. Features of the LOR1a LTR genomic region. (**a**) DNase hypersensitivity[15] tracks for three HL and two non-HL cell lines show open chromatin conformations over the LOR1a LTR in lines expressing chimeric *IRF5*. The 5′-ends of the LTR-initiated transcript and the native 'a' transcript[18] are shown in dark blue below the DNAse I tracks. (**b**) The inferred complete LOR1a LTR, shown as an orange bar above the Repeatmasker track, was identified by the tandem site duplications (TSDs, brown triangle) and homology of the upstream region to different LOR1a elements in hg19 found using BLAST alignment. The LOR1a extends past the RepeatMasker annotation. (**c**) Select JaspScan[22] motifs identified in the LOR1a include REL, IRF and STAT, and TATA-binding protein-binding sites. The 'P-V1' promoter region analyzed by Mancl *et al.*[21] is shown in light blue. (**d**) Multiple species alignments[34] and genomic evolutionary rate profiling (GERP) score[35] show that the LOR1a retrotransposition occurred in a common primate ancestor. (**e**) The consensus IRFE, the sequence found in the human genome upstream of (IRF5) and the inactive/mutated sequence (IRF5*) previously identified[21] are shown aligned to the 'AAAT' TSD sequence and beginning of the LOR1a LTR 'TGAAACC'.

unique to this integration event have combined to create this active promoter in HL.

In conclusion, we have shown that the LOR1a LTR upstream of *IRF5*, which is dormant in normal tissues, has been re-purposed in HL, resulting in LTR promoter activation and associating with overexpression of *IRF5*. Although IRF5 is oncogenic in HL,[15] the necessity and sufficiency of the LOR1a LTR-driven *IRF5* transcript to oncogenesis requires experimental validation via transcript isoform-specific knockdown of *IRF5* in HL cells. This onco-exaptation occurs recurrently in multiple independent HL lines, suggesting overexpression of IRF5 may be selected for and the LOR1a IRFE site provides an exploitable genetic circuit for this. *IRF5*, along with *CSF1R*[9] and *FABP7*,[12] are the best characterized examples of onco-exaptation of LTRs but this is likely to be a broadly occurring phenomenon in lymphomagenesis. In diffuse large B-cell lymphoma, at least 97 other chimeric transcripts were identified, many with unknown biological consequence.[12] In embryonic stem cells (with hypomethylated genomes), up to 15% of all transcript start sites fall within TEs[23] and LTRs are especially enriched in long non-coding RNA initiation.[4] Another recent study showed that LTR-promoted very long non-coding RNAs are more prevalent in cancer and embryonic stem cells compared with other tissues.[24] Furthermore, several studies have shown that the antisense promoter of L1 retroelements is hypomethylated and can drive gene expression, specifically in cancer cells.[25] One example of this is the L1-promoted transcript that produces a truncated c-MET protein in some cancers.[26,27] Taken together, these studies allude that cancer-specific transcription driven by activated LTRs or other TEs, namely onco-exaptation, is a distinct and underinvestigated mechanism for oncogene activation, with a unique etiology and, possibly, unique targets for therapy.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by a grant from the Leukemia and Lymphoma Society of Canada to DM and CS, with core support provided by the British Columbia Cancer Agency. AB is supported by a studentship award from the Natural Sciences and Engineering Research Council of Canada, MMK is supported by a postdoctoral fellowship from the Michael Smith Foundation for Health Research (MSFHR) and CS is supported by a scholarship award from MSFHR.

REFERENCES

1 Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J *et al.* Initial sequencing and analysis of the human genome. *Nature* 2001; **409**: 860–921.

2 Rebollo R, Romanish MT, Mager DL. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annu Rev Genet* 2012; **46**: 21–42.

3 Rebollo R, Farivar S, Mager DL. C-GATE - catalogue of genes affected by transposable elements. *Mob DNA* 2012; **3**: 9.

4 Kelley D, Rinn J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol* 2012; **13**: R107.

5 Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G *et al.* Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet* 2013; **9**: e1003470.

6 Xie M, Hong C, Zhang B, Lowdon RF, Xing X, Li D *et al.* DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. *Nat Genet* 2013; **45**: 836–841.

7 Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet* 2007; **8**: 272–285.

8 De Smet C, Loriot A. DNA hypomethylation in cancer: epigenetic scars of a neoplastic journey. *Epigenetics* 2010; **5**: 206–213.

9 Lamprecht B, Walter K, Kreher S, Kumar R, Hummel M, Lenze D *et al.* Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. *Nat Med* 2010; **16**: 571–579 1p following 579.

10 Roullet MR, Bagg A. Recent insights into the biology of Hodgkin lymphoma: unraveling the mysteries of the Reed-Sternberg cell. *Expert Rev Mol Diagn* 2007; **7**: 805–820.

11 Steidl C, Diepstra A, Lee T, Chan FC, Farinha P, Tan K *et al.* Gene expression profiling of microdissected Hodgkin Reed-Sternberg cells correlates with treatment outcome in classical Hodgkin lymphoma. *Blood* 2012; **120**: 3530–3540.

12 Lock FE, Rebollo R, Miceli-Royer K, Gagnier L, Kuah S, Babaian A *et al.* Distinct isoform of FABP7 revealed by screening for retroelement-activated genes in diffuse large B-cell lymphoma. *Proc Natl Acad Sci USA* 2014; **111**: E3534–E3543.

13 Karimi MM, Goyal P, Maksakova IA, Bilenky M, Leung D, Tang JX *et al.* DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs. *Cell Stem Cell* 2011; **8**: 676–687.

14 Lazzari E, Jefferies CA. IRF5-mediated signaling and implications for SLE. *Clin Immunol* 2014; **153**: 343–352.

15 Kreher S, Bouhlel MA, Cauchy P, Lamprecht B, Li S, Grau M *et al.* Mapping of transcription factor motifs in active chromatin identifies IRF5 as key regulator in classical Hodgkin lymphoma. *Proc Natl Acad Sci USA* 2014; **111**: E4513–E4522.

16 Liu Y, Abdul Razak FR, Terpstra M, Chan FC, Saber A, Nijland M *et al.* The mutational landscape of Hodgkin lymphoma cell lines determined by whole-exome sequencing. *Leukemia* 2014; **28**: 2248–2251.

17 Morin RD, Mendez-Lago M, Mungall AJ, Goya R, Mungall KL, Corbett RD *et al.* Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature* 2011; **476**: 298–303.

18 Clark DN, Read RD, Mayhew V, Petersen SC, Argueta LB, Stutz LA *et al.* Four promoters of IRF5 respond distinctly to stimuli and are affected by autoimmune-risk polymorphisms. *Front Immunol* 2013; **4**: 360.

19 Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 2005; **110**: 462–467.

20 Panganiban AT. Retroviral DNA integration. *Cell* 1985; **42**: 5–6.

21 Mancl ME, Hu G, Sangster-Guity N, Olshalsky SL, Hoops K, Fitzgerald-Bocarsly P *et al.* Two discrete promoters regulate the alternatively spliced human interferon regulatory factor-5 isoforms. Multiple isoforms with distinct cell type-specific expression, localization, regulation, and function. *J Biol Chem* 2005; **280**: 21078–21090.

22 Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 2000; **16**: 276–277.

23 Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM *et al.* The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet* 2009; **41**: 563–571.

24 St Laurent G, Shtokalo D, Dong B, Tackett MR, Fan X, Lazorthes S *et al.* VlincRNAs controlled by retroviral elements are a hallmark of pluripotency and cancer. *Genome Biol* 2013; **14**: R73.

25 Hur K, Cejas P, Feliu J, Moreno-Rubio J, Burgos E, Boland CR *et al.* Hypomethylation of long interspersed nuclear element-1 (LINE-1) leads to activation of proto-oncogenes in human colorectal cancer metastasis. *Gut* 2014; **63**: 635–646.

26 Wolff EM, Byun H-M, Han HF, Sharma S, Nichols PW, Siegmund KD *et al.* Hypomethylation of a LINE-1 promoter activates an alternate transcript of the MET oncogene in bladders with cancer. *PLoS Genet* 2010; **6**: e1000917.

27 Weber B, Kimhi S, Howard G, Eden A, Lyko F. Demethylation of a LINE-1 antisense promoter in the cMet locus impairs Met signalling through induction of illegitimate transcription. *Oncogene* 2010; **29**: 5775–5784.

28 Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 2013; **14**: R36.

29 Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010; **28**: 511–515.

30 Rebollo R, Karimi MM, Bilenky M, Gagnier L, Miceli-Royer K, Zhang Y *et al.* Retrotransposon-induced heterochromatin spreading in the mouse revealed by insertional polymorphisms. *PLoS Genet* 2011; **7**: e1002301.

31 Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012; **9**: 357–359.

32 Romanish MT, Nakamura H, Lai CB, Wang Y, Mager DL. A novel protein isoform of the multicopy human NAIP gene derives from intragenic Alu SINE promoters. *PLoS One* 2009; **4**: e5761.

33 Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. *Nat Methods* 2012; **9**: 671–675.

34 Chiaromonte F, Yap VB, Miller W. Scoring pairwise genomic sequence alignments. *Pac Symp Biocomput* 2002. 115–126.

35 Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* 2010; **6**: e1001025.

Supplementary Information accompanies this paper on the Oncogene website (http://www.nature.com/onc)