VIRAL-INDUCED TUMORIGENESIS: THE CHARACTERIZATION OF AVIAN LEUKOSIS VIRUS-INDUCED NEOPLASMS BY HIGH-THROUGHPUT SEQUENCING

by

James Franklin Justice IV

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

October, 2015
AVIAN LEUKOSIS VIRUS (ALV) is a simple retrovirus that infects chicken and integrates its DNA provirus into the host genome. Chickens exposed to either ALV subgroup-A (ALV-A) or ALV subgroup-J (ALV-J) during embryogenesis often develop tumors within a few months after hatching. Those exposed to ALV-A typically develop B-cell lymphomas, while birds exposed to ALV-J develop tumors of myeloid origin and hemangiomas. ALV-induced tumors have been the subject of numerous studies over the last thirty years. These neoplasms have been shown to develop via insertional mutagenesis, where the virus acts as a mutagen by integrating into the chicken genome and perturbing the expression of genes near the integration site. Early studies have located several common proviral integration sites which harbor genes thought to drive these tumors.

In this thesis, I analyze ALV-induced tumors with high-throughput sequencing, which enables a much more detailed view of the integration landscape than was possible with previous techniques. First, we infect 5- and 10-day chicken embryos with ALV-A to induce rapid-onset B-cell lymphomas, and then locate viral integration sites with the Illumina Hi-Seq. Four genes—MYC, MYB, Mir-155 and TERT—have previously been identified as common integration sites in ALV-A induced tumors. We identify these genes, and implicate additional genes as drivers of tumorigenesis in these neoplasms, including TNFRSF1A, MEF2C, CTDSPL, TAB2, RUNX1, MLL5, CXorf57, and BACH2. We also observe increased frequency of ALV integration near transcription start sites and within transcripts, as well as a weak integration site consensus sequence.
In addition, we studied ALV-J induced hemangiomas. We induced tumors by infecting 5-day chicken embryos with ALV-J strain PDRC-59831, a newly studied US isolate of ALV-J, which induced myeloid leukosis and hemangiomas. We sequenced integration sites and found expanded clones with integrations in the MET gene in two of the five hemangiomas studied, a gene which has not been implicated in ALV-J induced neoplasms to date. MET is a known proto-oncogene that acts through a diverse set of signaling pathways and is involved in many types of neoplasms. We show that tumors harboring MET integrations exhibit strong overexpression of the MET mRNA.
ACKNOWLEDGEMENTS

First, I would first like to thank Dr. Karen Beemon for her continued support and mentorship. Over the last 6 years I feel like I’ve gotten to know Karen very well and am so grateful for all of the discussions, both scientific and non-scientific, that we’ve shared over that time. I appreciate that regardless of whether the scientific method was serving up success or failure on a given week, I could count on her have a positive attitude and a calm knowledgeable perspective every time we talked. I also appreciate that Karen has always encouraged me to bring my own ideas to my work. I always knew that if I wanted to take my project in any alternate direction she would be supportive and eager to help in every way possible. Karen also has an eagerness to explore the newest, most cutting edge technology to answer scientific questions, which undoubtedly benefited my project immensely. Karen, I feel very fortunate to have spent the last 6 years with you as my mentor, thank you. I would also like to thank my thesis committee members, past and present, for their helpful suggestions and valuable input, Dr. Jef Boeke, Dr. David Zapulla, Dr. James Taylor, and Dr. Haig Kazazian.

I am also thankful for the contributions of my collaborators, without whom, my project would not have been possible. For their help generating chicken tumors, as well as helpful discussions and manuscript edits, I would like to thank Dr. Paul Neiman (Fred Hutchinson Cancer Center) and Dr. Robin Morgan (University of Delaware). For work with chickens I’d like to thank Miguel Ruano (University of Delaware), Erin Bernberg (University of Delaware), Amy Anderson (University of Delaware), Grace Isaacs (University of Delaware), Milos Markis (University of Delaware),
Delaware), and Grace Lagasse (University of Delaware). For help isolating virus I’d like to thank Dr. Guillermo Zavala (University of Georgia).

I am also thankful for all of the great people I have been fortunate to work with in Karen’s Lab. First I’d like to thank Mohan Bolisetty. When I joined the lab you took me under your wing and helped me learn the ropes. You were a great sounding board for ideas and always generous with your time and willing to help with any problem I encountered. I’d also like to thank Tom O’Hare, Gary Lam, and Sanandan Malhotra, the friendships we developed over the years have meant the world to me. Many thanks to our lab technician Yingying Li, for everything you do daily to make things run smoothly, I can’t imagine what it would have been like without you. I’d also like to thank Lin Quek, Shelby Winans, Johanna Withers, for your friendship and all of the great conversations over morning coffee.

Lastly, I’d like to thank my father and mother, Jay and Amy Justice. You’ve always stood by me with unwavering support and unconditional love. I hope I am making you proud.
# TABLE OF CONTENTS

ABSTRACT .................................................................................................................................................. ii
ACKNOWLEDGEMENTS .............................................................................................................................. iv

TABLE OF CONTENTS .............................................................................................................................. vi

LIST OF TABLES ......................................................................................................................................... ix

LIST OF FIGURES ....................................................................................................................................... x

LIST OF ABBREVIATIONS ........................................................................................................................ xi

Chapter 1. Introduction to avian retroviruses ......................................................................................... 1
  Introduction ............................................................................................................................................... 2
  ASLV Genome / Subgroups .................................................................................................................... 2
  Pathology ................................................................................................................................................ 3
  Viral entry .............................................................................................................................................. 10
  Nuclear Entry/Integration ..................................................................................................................... 11
  Transcription/Splicing/Nuclear Export ............................................................................................... 13
  Translation ............................................................................................................................................ 15
  Virion Assembly/Budding ....................................................................................................................... 17
  Conclusion ............................................................................................................................................ 17

Chapter 2. Characterization of ALV-A integration sites in B-cell lymphomas by high-throughput sequencing ........................................................................................................................................ 19
  Abstract ............................................................................................................................................... 20
  Introduction ........................................................................................................................................... 20
  Materials and Methods ........................................................................................................................ 25
    Tumor induction ................................................................................................................................. 25
    DNA extraction and deep sequencing ............................................................................................... 26
    Sequencing Analysis .......................................................................................................................... 29
Consensus Sequence and Feature and Gene Ontology analysis.......................... 30

Results .................................................................................................................. 31

Increased clonality in metastatic tumors vs bursal tumors ..................................... 31

Common integration sites .................................................................................. 34

Functional annotation enrichment analysis of ALV common integration sites. 44

ALV integration has a weak palindromic consensus sequence in vivo.................. 48

ALV prefers integration near promoters and within genes in vivo....................... 51

Discussion .......................................................................................................... 56

Chapter 3. Characterization of ALV-J integration sites in hemangiomas by high-throughput sequencing .......................................................... 62

Abstract ............................................................................................................... 63

Introduction ......................................................................................................... 63

Materials and Methods ..................................................................................... 67

Tumor induction .................................................................................................. 67

Phylogenetic analysis of ALV-J isolate PDRC-59831 env gene ............................ 69

DNA extraction and deep sequencing ................................................................. 71

Sequencing Analysis .......................................................................................... 72

MET gene expression ........................................................................................ 73

Results ............................................................................................................... 74

Characterization of ALV-J isolate PDRC-59831 .............................................. 74

Sequencing of ALV-J/cellular integration junctions .......................................... 75

Some ALV-J integrations are present in expanded clones .................................. 82

MET intron 1 is an expanded common integration site ....................................... 82

MET mRNA expression is elevated in tumors containing MET proviral integrations ........................................................................................................... 83
Discussion ................................................................. 86
Chapter 4. Conclusion and future directions ................................ 89
Introduction ........................................................................................................... 90

Is the preference for integration near transcription start sites due to selection?. 90

Could host cell factors be directing ALV integration?................................. 91

Does ALV prefer integration near CpG islands in vivo?.............................. 92

Does ALV prefer integration near highly expressed genes in vivo?.............. 93

Does nuclear architecture affect ALV integration?.......................................... 93

What is crosstalk and how can we work to eliminate it?................................. 94

Do the genes identified as common integration sites drive tumorigenesis?...... 97

References ........................................................................................................... 98

Curriculum Vitae .................................................................................................. 125
LIST OF TABLES

Table 2.1: Birds infected and neoplasms observed .......................................................... 24
Table 2.2: Expanded statistics for common integration sites .......................................... 39
Table 3.1: ALV-J tissues collected and neoplasms observed ........................................ 68
Table 3.2: Top 20 common integration sites in ALV-J induced hemangiomas ............ 81
# LIST OF FIGURES

- Figure 1.1: ASLV proviruses and transcripts .......................................................... 5
- Figure 2.1: ALV-A high-throughput integration site sequencing library preparation ..................................................................................................................... 28
- Figure 2.2: Metastatic tumors contain integrations within clonally expanded cells. 33
- Figure 2.3: Common sites of ALV proviral integration ............................................. 35
- Figure 2.4: Selected common ALV-A integration sites ........................................... 37
- Figure 2.5: Two liver tumors exhibited multiple unique integrations within CTDSPL1 and CTDSPL2 ................................................................................................................. 43
- Figure 2.6: ALV-A KEGG pathway analysis .............................................................. 46
- Figure 2.7: Gene ontology analysis .......................................................................... 47
- Figure 2.8: Consensus ALV-A target integration site .............................................. 50
- Figure 2.9: Preference for ALV-A integration near genomic features ..................... 52
- Figure 2.10: ALV-A integrations with respect to transcription start sites ............... 53
- Figure 3.1: Sequence comparison and phylogenetic analysis of env genes of ALV-J isolates ...................................................................................................................... 70
- Figure 3.2: ALV-J high-throughput sequencing library preparation ....................... 76
- Figure 3.3: Distribution of integration sites in ALV-J induced hemangiomas.......... 78
- Figure 3.4: ALV-J integration in the MET gene in chicken hemangiomas and non-tumor tissues .............................................................................................................. 79
- Figure 3.5: MET mRNA expression measured by qRT-PCR .................................. 85
- Figure 4.1: Dual-index library preparation .................................................................. 96
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALV</td>
<td>avian leukosis virus</td>
</tr>
<tr>
<td>ASLV</td>
<td>avian sarcoma/leukosis viruses</td>
</tr>
<tr>
<td>ASV</td>
<td>avian sarcoma virus</td>
</tr>
<tr>
<td>DR</td>
<td>direct repeat</td>
</tr>
<tr>
<td>ERV</td>
<td>endogenous retrovirus</td>
</tr>
<tr>
<td>HIV</td>
<td>human immunodeficiency virus</td>
</tr>
<tr>
<td>LPDV</td>
<td>lymphoproliferative disease virus of turkeys</td>
</tr>
<tr>
<td>LTR</td>
<td>long terminal repeat</td>
</tr>
<tr>
<td>MLV</td>
<td>murine leukemia virus</td>
</tr>
<tr>
<td>NMD</td>
<td>nonsense mediated decay</td>
</tr>
<tr>
<td>NRS</td>
<td>negative regulator of splicing</td>
</tr>
<tr>
<td>NLS</td>
<td>nuclear localization signal</td>
</tr>
<tr>
<td>PIC</td>
<td>preintegration complex</td>
</tr>
<tr>
<td>REV</td>
<td>reticuloendotheliosis virus</td>
</tr>
<tr>
<td>RSV</td>
<td>Rous sarcoma virus</td>
</tr>
<tr>
<td>TSS</td>
<td>transcription start site</td>
</tr>
</tbody>
</table>
Chapter 1. Introduction to avian retroviruses

Introduction

The study of avian retroviruses began with the work of two Danish researchers, Vilhelm Ellerman and Oluf Bang, more than a century ago. In a seminal 1908 publication, they showed that a form of leukemia and lymphoma that afflicts chickens, called avian leukosis, could be transferred by cell-free filtrates (Ellermann and Bang 1908). This was the first evidence that viruses could cause cancer, but because leukemia was not recognized as cancer, the significance of their work was not appreciated at the time. Three years later, however, the American scientist Peyton Rous discovered a second virus that induced sarcomas in chickens (Rous 1911). Rous sarcoma virus (RSV) has been studied extensively and has greatly aided our current understanding of both retroviral replication and cancer. We now know that RSV is derived from the more common avian leukosis virus (ALV). These, and related viruses, are collectively referred to as the avian sarcoma/leukosis viruses (ASLV) and are the most common and well-studied of the avian retroviruses. Other types of retroviruses discovered in birds include the gammaretrovirus, reticuloendotheliosis virus (REV) (Theilen, Zeigel, and Twiehaus 1966), and lymphoproliferative disease virus of turkeys (LPDV) (Biggs et al. 1978; L. N. Payne 1998).

ASLV Genome / Subgroups

ASLVs are members of the Alpharetrovirus genus of the family Retroviridae (King et al. 2011). The gag-pol-env gene order first determined for these simple retroviruses has proven to be common to all retroviruses. However, the complex
retroviruses also encode accessory proteins in overlapping reading frames. The ASLV \textit{gag} gene encodes the structural proteins capsid (CA), matrix (MA), and nucleocapsid (NC), as well as the viral protease (PR), which in most retroviruses is part of \textit{pol}. Thus, ASLVs synthesize relatively more PR than other retroviruses. The ASLV \textit{pol} gene encodes the reverse transcriptase (RT) and integrase (IN) enzymes, and the \textit{env} gene encodes the transmembrane (TM) and surface (SU) envelope glycoproteins. The viral genome is flanked on either side by long terminal repeats (LTRs). The most commonly studied ASLVs are classified into 6 distinct subgroups (A, B, C, D, E, and J) based on differences in their envelope glycoproteins, interference patterns, and host range in chicken cells of varying phenotypes. The chicken genome also encodes many endogenous retroviruses (ERVs), a minority of which are related to exogenous ASLVs (Bolisetty et al. 2012). Surprisingly, the majority of the ERVs are related to beta- and gammaretroviruses, suggesting they were once the dominant type of exogenous avian retrovirus (Bolisetty et al. 2012).

**Pathology**

Chickens are thought to be the natural hosts of ASLVs (L. N. Payne 1987), although some wild-bird species can also be infected (D. Li et al. 2013). ALV-A is the most common ASLV found in commercial flocks. It most often induces lymphoid leukemia, a B-cell lymphoma that begins in the bursa of Fabricius (an avian organ where B-cells mature) and can metastasize to the spleen, liver, and occasionally other organs. Although the virus is common, only a small fraction of infected birds in commercial flocks develop neoplasms. In addition, selective breeding has had some
success limiting viral spread, reducing the economic impact of the virus (Fadly and Nair 2008). The virus predominantly spreads horizontally by either direct or indirect contact, but can also be transmitted vertically from hen to egg.

ASLVs are divided broadly into two classes: slowly transforming and acutely transforming. ALVs (Figure 1.1A) make up the slow transforming class, causing tumors within a relatively long time frame of several weeks to months. Slowly transforming ALVs lack a viral oncogene and induce tumors by integrating within or near cancer-related host-cell genes, perturbing their expression or function. This process, first discovered with ALV, is called viral insertional mutagenesis and is common to other oncogenic simple retroviruses (Hayward, Neel, and Astrin 1981; G. S. Payne, Bishop, and Varmus 1982).
Figure 1.1: ASLV proviruses and transcripts

The ASLV proviruses and the transcripts of ALV (A), RSV (B), and MC29 (C) are shown. The location of viral oncogenes within the proviral genome are indicated when they exist. Protein coding portions of the transcripts are represented by colored rectangles. Locations of primer binding sites (PBS), RNA stability elements (RSE), polypurine tracts (PPT), and -1 framshift sites (-1 FS) are indicated. Polyprotein cleavage sites are not shown. This figure was adapted from (Petropoulos 1997).
ALV can act as a potent insertional mutagen when it integrates into the chicken genome. This is due to the fact that the provirus contains strong promoter and enhancer sequences in its viral long terminal repeats (LTRs). Thus, when the virus integrates, it is able to perturb the expression of nearby genes. It's still unclear exactly how far the virus has to be from a gene in order to perturb expression, although due to DNA looping and intrachromosomal interactions, this distance could be substantial.

Previous work has shown that if the virus integrates in the right spot (near an oncogene for example), it could potentially induce expression of this nearby gene and drive the cell towards tumorigenesis. Depending on where it integrates and its relation to the nearby genes, the virus can have other effects as well. For example, the virus could potentially reduce or eliminate the expression of a gene, it could induce expression of a truncated gene product (W. Jiang et al. 1997), or even perturb splicing or polyadenylation of a host transcript (G. S. Payne, Bishop, and Varmus 1982).

Experimentally infecting chicken embryos with ALV has been shown to induce metastatic B-cell lymphoma and occasionally other types of neoplasms. Like classic lymphoid leukosis, these lymphomas typically begin in the bursa and then metastasize to distant organs such as the liver, kidney, and spleen (Fadly and Nair 2008).

The latency of experimentally-induced tumors can vary between 1.5 and 6 months, and is dependent on the strain of ALV used and the age of the bird at the time of infection. Many studies have been conducted in the past using low-
throughput techniques to locate clusters of proviral integration in an attempt to identify genes that drive ALV-induced tumors. MYC was the first gene shown to be affected by ALV integrations in long latency B-cell lymphomas. These birds were infected 2–7 days after hatching and developed tumors by 4–6 months of age. This work showed that ALV integrations in the MYC gene locus caused deregulated expression of normal MYC protein from the ALV promoter (Hayward, Neel, and Astrin 1981; G. S. Payne, Bishop, and Varmus 1982). Similar integrations in the MYC locus were seen with REV-induced lymphomas around this time (Noori-Daloii et al. 1981). Shortly thereafter, another study showed that ALV insertion into the EGF receptor (c-erbB) locus can induce erythroblastosis in certain chickens lines (Nilsen et al. 1985).

Later, c-bic was shown to be a common integration site, and c-bic integrations often occurred in the same tumors as MYC integrations (Clurman and Hayward 1989). It turns out the c-bic is not protein coding, and instead codes for an oncogenic microRNA that was later given the name Mir-155 (Tam, Ben-Yehuda, and Hayward 1997). Later work showed that infecting 10-day embryos with a different strain of ALV, strain EU-8, resulted in tumors which developed more quickly and harbored integrations at the MYB locus (Kanter, Smith, and Hayward 1988).

More recently, clonal ALV integrations into the telomerase reverse transcriptase (TERT) promoter were seen in rapid-onset B-cell lymphomas (Yang et al. 2007). These tumors were shown to harbor clonal or oligoclonal integrations in the TERT promoter and are re-analyzed by high-throughput sequencing in Chapter 2 of this thesis.
It is now understood that both the viral strain and the time of infection are important in determining how quickly tumors develop and what genes are affected. EU-8, the strain that caused rapid onset B-cell lymphomas, is a recombinant strain of ALV that contains parts of ALV strain UR2AV and also ring-necked pheasant virus. Importantly, only 10-14 day embryonic EU-8 infections produced rapid-onset B-cell lymphomas. Infecting birds early with a different virus (UR2AV) produced only long-latency MYC tumors, as was the case if birds were infected after-hatching with EU-8. This experiment showed that not only is the virus an important factor in determining the latency of the tumors, but so is the time of infection.

Follow-up studies showed that EU-8 is able to rapidly induce tumors because it contains a 42-nucleotide deletion that disrupts the viral negative regulator of splicing (NRS). This NRS disruption reduces the efficiency of polyadenylation, increases the rate of viral readthrough, and increases the efficiency of splicing to downstream genes. These are factors that are thought to enable the virus to induce tumors so rapidly (M. R. Smith et al. 1997; O’Sullivan et al. 2002; Polony et al. 2003; Wilusz and Beemon 2006). Later, several modifications were made to ALV strain LR-9, a strain of ALV incapable of inducing rapid-onset B-cell tumors, and these changes were able to mimic the NRS deficiency of EU-8. These LR-9 mutant strains, LR9-Δ42 and LR9-G919A, were able to rapidly induce B-cell tumors, and were used to generate the TERT-integration-containing tumors mentioned previously (Polony et al. 2003; P. E. Neiman et al. 2003; Yang et al. 2007).

A new subgroup of ALV, ALV-J, was discovered in the late 1980’s. It has since spread rapidly resulting in great economic loss, especially in China. ALV-J is believed
to have originated from a recombination event between ALV and an endogenous retroviral element (Venugopal 1999; L. M. Smith et al. 1999; Sacco et al. 2000). The virus induces a different spectrum of tumors than ALV-A, primarily myeloid leukosis and hemangiomas (L. N. Payne, Gillespie, and Howes 1992; Cheng et al. 2010). The molecular basis of oncogenesis in these tumors was until recently not well understood. Recent work has shed some light on ALV-J-induced neoplasms, showing that MYC, TERT, and ZIC1 are targets of integration in ALV-J induced myeloid leukosis, and later in this thesis I show that MET is a common target in ALV-J induced hemangiomas (see Chapter 3) (Y. Li et al. 2014; Justice IV et al. 2015).

Acutely transforming viruses are distinct from the slow-transforming ALVs, in that they contain one or two viral oncogenes in their genome, and induce neoplastic transformation more rapidly, within days or weeks. These viral oncogenes are derived from host cellular genes that have been incorporated into the virus by gene capture and are over-expressed by the strong viral promoters. For example, RSV picked up the v-src oncogene, a mutated version of the c-src tyrosine kinase gene found in host cells (Figure 1.1B) (Stehelin et al. 1976; Ronald Swanstrom et al. 1983). Many acutely transforming ASLV strains have been isolated, carrying a wide variety of viral oncogenes, including myc, myb, fps, yes, jun, ets and erbB. With the exception of some strains of RSV, acutely transforming avian retroviruses are replication deficient (Figure 1.1C), requiring a helper virus to replicate (Vogt 1997).
**Viral entry**

The surface of all retroviral virions are studded with Env proteins, which bind to specific receptors on the target cell surface, beginning the infection cycle. Four distinct cellular receptors have been identified that mediate ASLV entry. The *tva*, *tvc*, and *chNHE1* (Chai and Bates 2006) genes confer susceptibility to ASLV of subgroup A, C, and J respectively, while *tvb* encodes the host cell receptor for subgroups B, D, and E. Two different alleles of *tvb* have been identified, *tvbs1* and *tvbs3*, which confer susceptibility to ASLV subgroups B and D, and B, D, and E respectively (Brojatsch et al. 1996; Adkins, Brojatsch, and Young 2000; Barnard and Young 2003). ALV has a broad cellular tropism, able to replicate in a range of tissues and organs in the chicken (Dougherty and Di Stefano 1967; Williams et al. 2004).

Once bound to its receptor on the host cell, virion-cell fusion is activated through a two-step mechanism. First, the envelope glycoprotein undergoes a conformational change at the cell surface that mediates viral uptake and endosomal trafficking. Then, the acidic environment of the endosome activates hemifusion and release of the capsid into the cytoplasm (Barnard and Young 2003). Once in the cytoplasm, reverse transcription of the viral RNA can begin. Reverse transcription is an intricate multi-step process in which cellular dNTPs and packaged tryptophan tRNAs are utilized by the reverse transcriptase to convert viral RNA into a double-stranded DNA provirus. Detailed reviews of the reverse transcription process are available elsewhere (Telesnitsky and Goff 1997).
**Nuclear Entry/Integration**

Prior to nuclear entry, the DNA copy of the viral genome associates with integrase and other viral and cellular proteins to form the preintegration complex (PIC). The ability of the PIC to gain access to nuclear DNA varies among retroviruses. Gammaretroviruses can only integrate after nuclear envelope disassembly in mitosis, while lentiviruses are able to infect non-dividing cells via active transport into the nucleus (Lewis and Emerman 1994). For decades it was thought that alpharetroviruses, like gammaretroviruses, were only able to infect dividing cells (Temin 1967). Contrary to this idea, it was observed that cells infected with avian sarcoma virus (ASV), upon the release of serum starvation, could be stably integrated into host genomic DNA prior to mitosis (Humphries, Glover, and Reichmann 1981), suggesting that nuclear envelope breakdown is not necessary for ASV integration. It is now thought that ASV can infect a variety of non-cycling cells (Hatzioannou and Goff 2001; Katz et al. 2002; Greger et al. 2004; Katz, Greger, and Skalka 2005), and the ASV integrase contains a nuclear localization signal (NLS) that can mediate active transport of the ASV integrase through the nuclear pore (Andrake et al. 2008). This work, together with earlier observations (Fritsch and Temin 1977), suggests a block at reverse transcription rather than nuclear import may be responsible for the low infectivity of ASLV in non-dividing cells (Katz, Greger, and Skalka 2005).

Once in the nucleus, the PIC mediates integration of the provirus into the host genomic DNA. Integration site selection varies among retroviruses, with some
showing strong selection for certain areas of the genome. Murine leukemia virus (MLV), for example, has a strong preference for integration near transcription start sites (TSSs) and CpG islands, while HIV shows decreased integration frequency near TSSs (Wu et al. 2003; Cattoglio et al. 2010). MLV’s preference for the TSS is mediated by the binding of the host bromodomain and extraterminal domain (BET) proteins to the MLV integrase (Sharma et al. 2013). Targeting of HIV integration has been shown to involve host protein LEDGF/p75 which interacts with integrase and mediates target site selection (Maertens et al. 2003; Cherepanov et al. 2003; Llano et al. 2004; Emiliani et al. 2005).

ASLV integration is more indiscriminant than that of MLV and HIV. Several studies have shown ALV integration occurs in a quasi-random fashion in cells grown in culture, with only slight preference for active transcription units (Mitchell et al. 2004; Narezkina et al. 2004; Barr et al. 2005). This was shown to be true regardless of whether ALV is infecting chicken or human cells (Barr et al. 2005). The reason for this pattern of integration targeting has not been elucidated (Craigie and Bushman 2012). Interestingly, a weak consensus sequence of ALV integration has also been observed. Other retroviruses such as MLV and HIV also exhibit integration consensus sequences. These consensus sequences are all unique to the virus, although in each case the sequence forms a palindrome. (Wu et al. 2005; Holman and Coffin 2005).

In chapter 2 of this thesis I will describe recent work which shows that in vivo, in tumors, we observe increased integration frequency of ALV-A near
transcription start sites (TSS). The reason for the discrepancy between earlier work in cell culture and our work may be that the integrations we map have undergone months of purifying selection. Hence, the preference for integration near transcription start sites may be due to selection for integrations near TSS by the tumor cells and not an actual preference of the PICs for TSSs.

**Transcription/Splicing/Nuclear Export**

Transcription of the ASLV provirus is directed by the viral LTRs, which contain strong promoter and enhancer sequences. Because ASLVs do not encode transcriptional transactivators, they rely entirely on host transcription factors, which bind to the U3 region of the LTR and drive RNA Polymerase II transcription of the provirus. All viral transcripts have an m7G cap at the 5′ end and undergo 3′ end cleavage and polyadenylation by cellular machinery prior to export from the nucleus.

All replication-competent ASLVs produce a single primary RNA transcript. This full-length, unspliced viral RNA serves as mRNA for translation of *gag* and *pol* genes, as well as the genomic material that is packaged into new virions. In addition, a fraction of these primary transcripts are spliced to generate the subgenomic *env* mRNA. In the case of RSV, a second spliced transcript is generated which encodes the *v-src* oncogene (Figure 1.1B). Replication-deficient ASVs such as myelocytomatosis virus (MC29) (Figure 1.1C) typically produce only a single unspliced *gag-one* fusion transcript and require a helper virus to replicate.
The full-length viral RNA transcript (~7-9 kb) is an aberration in the cellular context, in which intron-containing transcripts are not usually exported from the nucleus for translation. To ensure successful replication, ASLV has evolved a variety of RNA elements that aid in the efficient transcription, export, and translation of these long unspliced viral RNAs.

For example, ASLV employs several mechanisms to protect these transcripts from splicing, such as suboptimal 3’ splice sites (McNally and Beemon 1992), and cis-acting RNA elements. One cis-acting element is the negative regulator of splicing (NRS). The NRS acts as a faux 5’ splice site by interacting with the 3’ splice site and recruiting components of the spliceosome. This sequesters the 3’ splice site away from the authentic 5’ splice site and further reduces the efficiency of splicing (Giles and Beemon 2005).

Unlike complex retroviruses, which encode accessory proteins that mediate nuclear export of unspliced and incompletely spliced viral RNAs, ASLV relies on a 100 nt direct repeat (DR) RNA sequence for export. ALV employs a single DR in its 3’ untranslated region (UTR), while RSV has two sequences, located on either side of the src oncogene. The DRs must form a highly stable stem loop structure to mediate nuclear export, which is dependent on the nuclear export factor Tap (a.k.a Nxf1), although the DR does not appear to bind Tap directly (LeBlanc et al. 2007; Paca et al. 2000).

The full length viral RNA faces yet another obstacle during translation. Because the viral RNA has a stop codon at the end of the gag gene, the region
downstream appears to be a long 3′UTR, which often targets cellular transcripts for degradation by the nonsense mediated decay (NMD) machinery. In order to avoid degradation, ASLV has evolved a 400 nt element, named the RNA stability element (RSE) (Weil, Hadjithomas, and Beemon 2009). Positioned immediately downstream of the gag termination codon, the RSE protects the full length viral RNA from NMD-mediated degradation. More work is currently underway to determine the mechanism by which the RSE functions.

**Translation**

The full length viral transcript acts as template for synthesis of two different polyproteins, Gag-Pro and Gag-Pro-Pol. Because the virion is composed of more Gag structural proteins than Pol proteins, all retroviruses have developed means to synthesize more Gag than Pol, despite being encoded on the same transcript. ASLV accomplishes this by frameshifting. Frameshifting occurs on a short A-U rich “slippery sequence” just upstream of the gag termination codon that is followed by an RNA pseudoknot (Jacks et al. 1988). This pseudoknot pauses the ribosome over the slippery sequence, which occasionally (~5% of the time) causes it to slip backwards a single nucleotide before continuing forward. This “-1 frameshift” places the gag termination codon out of frame and allows the ribosome to read through to the pol termination codon, generating the Gag-Pro-Pol polyprotein.

During synthesis, ASLV Gag undergoes further modification. Most retroviral Gag proteins undergo low levels of phosphorylation and myristylation at their N-termini. ASLV Gag does undergo low level phosphorylation (R Swanstrom and Wills
1997), but it is not myristylated. Instead ASLV Gag is acetylated at its N-terminus (Palmiter et al. 1978). The function of this acetylation is currently unknown.

The Env polyprotein is synthesized from a separate spliced transcript. Unique among retroviruses, the ASLV env splice donor resides within gag, just downstream of the gag start codon. This appends the first six amino acids of Gag onto the beginning of each Env polyprotein (Ficht, Chang, and Stoltzfus 1984; Hunter et al. 1983; R Swanstrom and Wills 1997). The Env polyprotein undergoes additional processing and modification in the endoplasmic reticulum. During this process, the Env polyprotein is proteolytically cleaved into three fragments, is glycosylated, undergoes folding, forms a trimer (Einfeld and Hunter 1988), and is then exported to the cell surface. Once the trimer has formed, Env is competent to bind its cognate receptor which is also processed through the secretory pathway. This forms the basis of superinfection resistance. By binding the receptor in the ER and remaining bound as it is presented at the cell surface, infection of the cell with any virus that uses that receptor is inhibited.

In addition to the canonical translation products, three small peptides are synthesized from upstream open reading frames (uORFs) adjacent to the gag gene. These uORFs are conserved among all ASLVs and may play a role in translation and packaging (Donzé, Damay, and Spahr 1995; Moustakas, Sonstegard, and Hackett 1993).
Virion Assembly/Budding

The assembly and budding process is mediated in large part by Gag. Following synthesis, the RSV Gag protein is transiently imported to the nucleus where the Gag NC domain interacts with a packaging sequence (Ψ) on the viral genomic RNA (Scheifele et al. 2002; Gudleski et al. 2010). After Gag dimerization, a nuclear export signal within the p10 domain of Gag mediates export of the ribonucleoprotein (RNP) complex by the CRM1 Pathway (Gudleski et al. 2010; Scheifele, Ryan, and Parent 2005).

The RNP complex then undergoes phosphoinositide-dependent (Nadaraia-Hoke et al. 2013) trafficking from the nucleus and stably associates with the plasma membrane via a membrane binding domain (MBD) at the N-terminus of Gag (Verderame, Nelle, and Wills 1996). The degree to which assembly occurs prior to plasma membrane localization is not well understood. At the plasma membrane the viral polyproteins, processed Env proteins, two covalently linked viral genomic RNAs and tryptophan tRNAs somehow coalesce, and the virion buds from the cell surface. Various domains of Gag, and multiple host proteins have been shown to be important in this process (R Swanstrom and Wills 1997; Pincetic and Leis 2009). The viral PR mediates cleavage of the polyproteins, and the virion obtains a mature morphology capable of infection shortly after budding.

Conclusion

Since the discovery of avian retroviruses more than a century ago, much has been learned about their replication and life cycle. ASLV has proven to be very
adaptable. Its ability to capture and use cellular genes, evolve to use a variety of cellular receptors, and recombine to form ALV-J attests to this. Though our understanding of the virus has increased, there are still many aspects of ASLV replication that warrant further study.
Chapter 2. Characterization of ALV-A integration sites in B-cell lymphomas by high-throughput sequencing

Adapted from Justice IV, James, Robin Morgan, and Karen Beemon. 2015. Common viral integration sites identified in avian leukosis virus-induced B-cell lymphoma. [In review]
Abstract

Avian Leukosis Virus (ALV) induces B-cell lymphoma and other neoplasms in chickens by integrating within or near cancer genes and perturbing their expression. Four genes—MYC, MYB, Mir-155 and TERT—have previously been identified as common integration sites in these viral induced lymphomas and are thought to play a causal role in tumorigenesis. In this study, we employ high-throughput sequencing to identify additional genes driving tumorigenesis in ALV-induced B-cell lymphomas. In addition to the four genes implicated previously, we identify other genes as common integration sites, including TNFRSF1A, MEF2C, CTDSPL, TAB2, RUNX1, MLL5, CXorf57, and BACH2. We also analyze the genome-wide ALV integration landscape in vivo, and find increased frequency of ALV integration near transcriptional start sites and within transcripts. Previous work has shown ALV prefers a weak consensus sequence for integration in cultured human cells. We confirm this consensus sequence for ALV integration in vivo in the chicken genome.

Introduction

Avian leukemia virus (ALV) is a simple retrovirus that infects chicken and some other avian species (Justice IV and Beemon 2013). Like all retroviruses, ALV RNA is reversed-transcribed and the proviral DNA enters the nucleus, where it integrates into the genomic DNA of the host cell. Several studies have shown ALV integration occurs in a quasi-random fashion in human and chicken cells grown in culture, with only slight preference for active transcription units (Mitchell et al.
Infecting chicken embryos with ALV has been shown to induce metastatic B-cell lymphoma and occasionally other types of neoplasms. The latency of these tumors can vary between 1.5 and 6 months, and is dependent on the strain of ALV injected and the age of the bird at the time of infection. The lymphomas typically begin in the bursa (an avian organ in which B-cells mature), and then metastasize to distant organs such as the liver, kidney, and spleen (Fadly and Nair 2008).

Unlike the closely related Rous sarcoma virus (RSV), ALV does not carry a transforming oncogene. Instead, ALV induces tumors by insertional mutagenesis (Hayward, Neel, and Astrin 1981; G. S. Payne, Bishop, and Varmus 1982). ALV is a potent insertional mutagen because the provirus contains strong promoter and enhancer sequences in its viral long terminal repeats (LTRs). This means that when ALV randomly integrates into the genome, it can perturb the expression of genes in the vicinity of the proviral integration site. Hence, if the virus integrates near a cancer gene, the ALV-induced misexpression of that gene may contribute to the transformation of the cell and potentially tumorigenesis. Depending on where ALV integrates and its relation to the nearby genes, the virus can have other effects as well. For example, the virus could potentially reduce or eliminate the expression of a gene, it could induce expression of a truncated gene product (W. Jiang et al. 1997), or potentially perturb splicing or polyadenylation of a host transcript (G. S. Payne, Bishop, and Varmus 1982).
Much work has been done to identify genes that drive ALV-induced oncogenesis by locating clusters of proviral integration in these tumors. MYC was the first gene shown to be affected by ALV integrations in long latency B-cell lymphomas (Hayward, Neel, and Astrin 1981; G. S. Payne, Bishop, and Varmus 1982). These birds were infected 2–7 days after hatching and developed tumors by 4–6 months of age. Later c-bic was shown to be a common integration site, and c-bic integrations often occurred in the same tumors as MYC integrations (Clurman and Hayward 1989). It turns out the c-bic is not protein coding, and instead codes for an oncogenic microRNA that was later given the name Mir-155 (Tam, Ben-Yehuda, and Hayward 1997). Later work showed that infecting 10-day embryos with a different strain of ALV, strain EU-8, resulted in short latency tumors harboring integrations at the MYB locus (Kanter, Smith, and Hayward 1988). Recent work studying ALV subgroup-J has shown that MYC, TERT, and ZIC1 are targets of integration in ALV-J induced myeloid leukemia, and MET is a common target in ALV-J induced hemangiomas (Y. Li et al. 2014; Justice IV et al. 2015).

It is now understood that both the viral strain and the time of infection are important in determining how quickly tumors develop and what genes are affected. EU-8, the strain that caused rapid onset B-cell lymphomas, is a recombinant strain of ALV that contains parts of ALV strain UR2AV and ring-necked pheasant virus. Importantly, only embryonic EU-8 infections produced rapid-onset B-cell lymphomas. Infecting birds early with a different virus (UR2AV) produced only long-latency MYC tumors, as was the case if birds were infected after-hatching with EU-8. This experiment showed that not only is the virus an important factor in
determining the latency of the tumors, but so is the time of infection.

Follow-up studies showed that EU-8 is able to rapidly induce tumors because it contains a 42-nucleotide deletion that disrupts the viral negative regulator of splicing (NRS). This NRS disruption reduces the efficiency of polyadenylation, increases the rate of viral readthrough, and increases the efficiency of splicing to downstream genes. These factors that are thought to enable the virus to induce tumors so rapidly (M. R. Smith et al. 1997; O'Sullivan et al. 2002; Polony et al. 2003; Wilusz and Beemon 2006). Later, two modifications were made to ALV strain LR-9, a strain of ALV incapable of inducing rapid-onset B-cell tumors, which were able to mimic the NRS deficiency of EU-8. These LR-9 mutant strains, LR9-Δ42 and LR9-G919A, were able to rapidly induce B-cell tumors (Polony et al. 2003; P. E. Neiman et al. 2003).

In this work, we generated rapid onset B-cell lymphomas by infecting 5- and 10-day embryos with either ALV-A viral strain LR-9, LR9-Δ42, LR9-U916A, or LR9-G919A (Table 2.1).
Table 2.1: Birds infected and neoplasms observed

<table>
<thead>
<tr>
<th>Chicken ID</th>
<th>Inf. Day</th>
<th>Virus</th>
<th>Burza</th>
<th>Liver</th>
<th>Spleen</th>
<th>Brain</th>
<th>Kidney</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-G919-4</td>
<td>5</td>
<td>G919A</td>
<td>T</td>
<td>T</td>
<td>X</td>
<td>X</td>
<td>T</td>
</tr>
<tr>
<td>2-G919-4</td>
<td>5</td>
<td>G919A</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>T</td>
</tr>
<tr>
<td>2-G919-6</td>
<td>5</td>
<td>G919A</td>
<td>T</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>3-WLR9-4</td>
<td>5</td>
<td>LR-9</td>
<td>T</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>3-WLR9-5</td>
<td>5</td>
<td>LR-9</td>
<td>T</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>3-WLR9-7</td>
<td>5</td>
<td>LR-9</td>
<td>T</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>4-WLR9-3</td>
<td>5</td>
<td>LR-9</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>T</td>
</tr>
<tr>
<td>4-WLR9-7</td>
<td>5</td>
<td>LR-9</td>
<td>T</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>4-WLR9-8</td>
<td>5</td>
<td>LR-9</td>
<td>X</td>
<td>H</td>
<td>T</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>4-WLR9-9</td>
<td>5</td>
<td>LR-9</td>
<td>X</td>
<td>T</td>
<td>X</td>
<td>X</td>
<td>H</td>
</tr>
<tr>
<td>7-Con-1</td>
<td>none</td>
<td>none</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>A1-R580</td>
<td>10</td>
<td>ΔLR-9</td>
<td>T</td>
<td>I</td>
<td>X</td>
<td>X</td>
<td>T</td>
</tr>
<tr>
<td>A2-R588</td>
<td>10</td>
<td>ΔLR-9</td>
<td>T¹</td>
<td>T¹</td>
<td>T</td>
<td>X</td>
<td>T</td>
</tr>
<tr>
<td>A6-R794</td>
<td>10</td>
<td>ΔLR-9</td>
<td>NF³</td>
<td>X</td>
<td>T</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>A9-R896</td>
<td>10</td>
<td>ΔLR-9</td>
<td>T</td>
<td>X</td>
<td>I</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>B8-B975</td>
<td>10</td>
<td>G916A</td>
<td>T</td>
<td>I</td>
<td>X</td>
<td>X</td>
<td>I</td>
</tr>
<tr>
<td>C2-B253</td>
<td>10</td>
<td>G919A</td>
<td>T¹</td>
<td>T²</td>
<td>T</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>C3-B256</td>
<td>10</td>
<td>G919A</td>
<td>T¹</td>
<td>T¹</td>
<td>X</td>
<td>X</td>
<td>T</td>
</tr>
<tr>
<td>C4-B259</td>
<td>10</td>
<td>G919A</td>
<td>T</td>
<td>T¹</td>
<td>T</td>
<td>X</td>
<td>T</td>
</tr>
<tr>
<td>C6-B268</td>
<td>10</td>
<td>G919A</td>
<td>NF³</td>
<td>T¹,²</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>C7-B271</td>
<td>10</td>
<td>G919A</td>
<td>T</td>
<td>T²</td>
<td>X</td>
<td>X</td>
<td>T</td>
</tr>
<tr>
<td>D1-G157</td>
<td>10</td>
<td>LR-9</td>
<td>NF</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>T</td>
</tr>
<tr>
<td>D2-G158</td>
<td>10</td>
<td>LR-9</td>
<td>NF</td>
<td>T²</td>
<td>X</td>
<td>X</td>
<td>T</td>
</tr>
<tr>
<td>D4-G163</td>
<td>10</td>
<td>LR-9</td>
<td>T</td>
<td>I</td>
<td>X</td>
<td>X</td>
<td>I</td>
</tr>
<tr>
<td>D5-G169</td>
<td>10</td>
<td>LR-9</td>
<td>NF</td>
<td>T</td>
<td>T</td>
<td>X</td>
<td>T</td>
</tr>
<tr>
<td>D8-G172</td>
<td>10</td>
<td>LR-9</td>
<td>T</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Normal #1</td>
<td>none</td>
<td>none</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

T = B-cell lymphoma, NF = Neoplastic Follicles, I = Inflammation, H = Hemangioma, X = Tissue collected, no neoplasm observed. Shaded boxes represent samples which underwent high-throughput sequencing. 1. MYB integration identified by nested-PCR. R794 Bursa MYB rearrangement was confirmed by southern blot (P. E. Neiman et al. 2003). 2. TERT integration observed via inverse-PCR and confirmed via Southern blot (Yang et al. 2007).
A subset of these tumors were analyzed previously by lower throughput methods (P. E. Neiman et al. 2003; Polony et al. 2003; Yang et al. 2007). Some tumors were shown to harbor MYB integrations via locus-specific nested PCR, and inverse-PCR identified TERT as common integration sites in some tumors. Southern blot analysis showed several tumors appeared to be clonal or oligoclonal for TERT integrations, while others were clonal for MYB (Yang et al. 2007). In this study we use high throughput sequencing to locate proviral integrations to get a more complete understanding of the integration landscape in these tumors and the genes that are perturbed by ALV integration.

**Materials and Methods**

**Tumor induction**

5- and 10-day-old chicken embryos were injected with either ALV-LR9, ALV-ΔLR9, ALV-G919A, or ALV-U916A. Chickens injected at 5 days were SPAFAS embryos (Charles River) and were injected via the yolk sac route. The 10-day injections were inbred SC White Leghorn line embryos (Hyline International, Dallas, Iowa) and were injected into the chorioallantoic veins as described previously (Polony et al. 2003). A total of 10 birds were infected on embryonic day 5, and 15 birds were infected on day 10. Chickens were observed daily and were euthanized when apparently ill or at 12 weeks (for the 5-day injected cohort) or 10 weeks (for the 10-day injected cohort).

37 tissues were selected for characterization by high-throughput sequencing. Two uninfected tissues and several non-tumor tissues from infected birds were
sequenced to serve as controls (Table 2.1). Additional birds were infected but not all birds were analyzed in this study. Birds A2-R588, A6-R794, C2-B253, C3-B256, C4-B259, and C6-B268 underwent some early characterization and were found to harbor viral integrations in the MYB locus by nested PCR (P. E. Neiman et al. 2003). Likewise, birds C2-B253, C6-B268, C7-B271, and D2-G158 were previously shown to harbor clonal or oligoclonal integrations in the TERT promoter (Yang et al. 2007).

**DNA extraction and deep sequencing**

DNA was isolated as described previously (Justice IV et al. 2015). 5 μg of purified genomic DNA was sonicated with a Bioruptor™ UCD-200. End repair, A-tailing, and adapter ligation were performed as described by Gillet et al. (Gillet et al. 2011a) (adapter short arm: p-gatcgaagagcaaaaaaaaaaaaa, adapter long arm: caagcagaagacggcatacgagatxxxxxgtgactggagttcagacgtgtgctcttccgatc*t where the x’s denote the barcode sequence, P denotes phosphorylation, and * denotes a phosphorothioate bond). Nested PCR was performed to enrich the library for proviral junctions. The first PCR reaction was 23 cycles, and employed an ALV-A specific primer (cgcgaggacgtaagaaatttcagg) between the viral 3′LTR and env, and a primer (caagcagaagacggcatacgagat) within the adapter attached by ligation in the previous step. In the second round of PCR, a primer (aatgatacggcaccacggagatctacactcgacgactacgagcatgcatgaag) at the 3′ end of the LTR was used. This primer ended twelve nucleotides short of the junction between virus and genomic DNA. This primer was paired with a primer within the adapter on the opposite side of the fragment, which overlapped the adaptor’s barcode.
sequence (caagcagaacggaacggtacagagatxxxxxx). Libraries were quantified by qPCR, and then underwent single-end 75 bp or 100 bp multiplexed sequencing on the Illumina Hi-Seq 2000. A custom sequencing primer (acgactacgacatgcatgataacgagacgga) was used which hybridized near the end of the viral 3′ LTR, five nucleotides short of the proviral/genomic DNA junction. This resulted in reads that could be validated as genuine integrations by verifying that the read began with the last 5 nucleotides of the proviral DNA, “CTTCA”. The last two nucleotides of the unintegrated proviral DNA, “TT”, are cleaved by viral integrase upon integration, so the lack of these two nucleotides in the read acted as further validation of a true viral integration. Figure XX shows a diagram of primer placement during library preparation.
Figure 2.1: ALV-A high-throughput integration site sequencing library preparation

All primers used to prepare libraries for high throughput sequencing are shown. Barcode sequences are in red font and host genomic DNA is shown in a green font. Nucleotides highlighted in Grey at the final product step are sequences which bind to the flow cell.
Sequencing Analysis

Reads were first filtered with a custom Python script to remove sequences that did not begin with the last five nucleotides of viral DNA, “CTTCA.” Files were then uploaded to Galaxy (Giardine et al. 2005; Blankenberg, Von Kuster, et al. 2010; Goecks et al. 2010), which was used to perform some downstream analyses. In Galaxy, the quality scores were first converted to Sanger format with FastQ Groomer v1.0.4 (Blankenberg, Gordon, et al. 2010). Adapters were trimmed using the Galaxy Clip tool v1.0.1. This tool also removed reads containing an N, and reads less than 20 nucleotides in length after adapter removal. The remaining reads, were mapped with Bowtie (Langmead et al. 2009) using the Gallus gallus 4.0 genome (Nov. 2011). 100,000 random mapped reads were selected from each sample to be used for further analysis. If less than 100,000 reads were present for a sample, all available reads were used.

A custom Perl pipeline was developed to analyze the aligned reads output from Bowtie. Briefly, reads containing sequencing errors were filtered, and read counts and sonication breakpoints were quantified. Integrations found in multiple samples were assigned to the sample with the highest number of breakpoints. Files were annotated with refseq features, and the orientation and distance to the nearest gene were calculated for each integration.

Following pipeline analysis, integrations into repetitive regions were manually removed from the dataset. In all, 32,050 unique ALV integrations were obtained. Integration clusters were identified via a sliding window approach. If
twelve or more integrations were observed within a 50 kb window they were considered a cluster of viral integration. If the cluster was located within or near a gene, all additional integrations in that gene were also counted, as were any integrations within 10 kb upstream or downstream of that gene. If the integration cluster encompassed two genes, both genes were recorded and any integrations between the two genes and within 10 kb of either end were included in the cluster. The source code for this pipeline is available upon request.

**Consensus Sequence and Feature and Gene Ontology analysis**

Reads were mapped with Bowtie (Langmead et al. 2009). Only reads which mapped uniquely to the genome were kept, and any reads that mapped equally well to two locations were discarded. This step filtered out reads that originate from repetitive elements. Mapped bowtie reads from all samples were then combined into a single file and analyzed with HOMER (Heinz et al. 2010). HOMER calculates the nucleotide composition and enriched features at each integration loci.

A random integration control dataset was generated with Bedtools Random (Quinlan and Hall 2010). The genomic DNA sequences corresponding to the genomic coordinates obtained from Bedtools Random were extracted from the Gallus gallus 4 genome using the Galaxy tool Extract Genomic DNA (Goecks et al. 2010; Blankenberg, Von Kuster, et al. 2010; Giardine et al. 2005). Control sequences were mapped with Bowtie and analyzed with HOMER using the same conditions as above. A consensus Logo plot was constructed with Seq2Logo (Thomsen and Nielsen 2012). Gene ontology analysis for the top 48 common integration sites was
conducted with DAVID (Huang, Sherman, and Lempicki 2009a; Huang, Sherman, and Lempicki 2009b).

Results

We sequenced 37 tissue samples from 25 different birds (Table 2.1) and obtained approximately 2.39 million reads originating from viral integrations in tumor and non-tumor tissues. These reads mapped to 32,050 unique viral integration sites. Among these unique integration sites, we identified exactly 43,000 unique sonication breakpoints. The average number of breakpoints per integration was 1.342, with the vast majority of integrations (86.8%) showing only a single sonication breakpoint and therefore no evidence of clonal expansion.

Increased clonality in metastatic tumors vs bursal tumors

The bursa (an avian organ in which B-cells mature) is believed to act as the primary organ of transformation in cases of ALV induced B-cell lymphoma. Laboratory infected chickens typically develop multiple primary neoplastic follicles in the bursa, some of which may eventually form primary tumors within the bursa. Secondary tumors are also commonly found in the liver, spleen, kidneys, and sometimes other organs. These tumors are believed to arise when a single cell within the bursa acquires a combination of integrations and possibly other mutations that enable the cell to proliferate and then metastasize to a distant organ. Once at the distant location, the progenitor cell is thought to clonally expand and form a tumor, which typically presents as a nodular or diffuse tumor in the distant organ (Fadly and Nair 2008).
The extent to which the progenitor cell has clonally expanded can be measured by determining the number of different sonication breakpoints observed for an integration (Gillet et al. 2011a; Berry et al. 2012). Sonication breakpoints are generated during library preparation by the shearing of genomic DNA followed by ligation of adapters onto the sheared ends. When an integration occurs in a cell that later divides by clonal expansion, multiple sonication breakpoints can potentially be observed for that integration. In this way, it is possible to obtain a metric of relative clonal expansion for each integration in a given sample.

Consistent with the clonal expansion hypothesis, we observed that metastatic tumors often contained one or more integrations that have a high number of breakpoints, whereas bursal tumors only occasionally exhibited highly expanded integrations. This can be visualized via a pie chart, where the pie represents a tumor, each slice represents a specific integration, and the size of the slice corresponds to the number of sonication breakpoints observed for that integration. Pie charts for a typical metastatic liver tumor, bursa with neoplastic follicles, and liver exhibiting no tumor are shown in Figure 2.2. The liver tumor contains several integrations that show a high level of clonal expansion. The bursa contains many different neoplastic follicles, each with a unique complement of integrations and all with low levels of clonal expansion. A non-tumor liver is also shown for comparison that, as expected, exhibits almost no clonally expanded integrations.
Figure 2.2: Metastatic tumors contain integrations within clonally expanded cells

Each pie represents a specific tissue that underwent high-throughput integration-site sequencing. Each slice represents a specific integration and the size of each slice corresponds to the number of breakpoints observed for that integration. The integrations that exhibit the greatest clonal expansions are shown. A total of 200 breakpoints are shown for each sample. Left, C3-B256 metastatic liver tumor exhibits extensive clonal expansion. Middle, D1-G157 bursa with neoplastic follicles contains some integrations in moderately expanded clones. Right, D4-G163 non-tumor liver, exhibits very few integrations in expanded clones.
Common integration sites

A total of 37 tissues including 13 primary neoplasms and 17 metastatic tumors were sequenced. Analysis of the resulting integrations identified a diverse array of genes as targets of ALV integration. A list of the top 48 targets of integration is shown (Figure 2.3). All of these common integration sites exhibited at least 12 unique integrations within a single 50-kb sliding window. Several of the most targeted genes have been identified in previous ALV insertional mutagenesis screens. For example, the first gene identified as a common integration site in long latency ALV-induced lymphomas was MYC in 1981 (Hayward, Neel, and Astrin 1981). Although MYC is not among the top 50 common targets of integration, we did identify nine unique integrations into the MYC gene. In addition, the MYC integration cluster was among the most clonally expanded clusters in our study with 8.44 breakpoints per integration, second only to TERT (Figure 2.3). MYB, first seen as a common integration site in rapid onset lymphomas in 1988 (Kanter, Smith, and Hayward 1988) is tied for the fifth most targeted gene with 28 unique integrations. Likewise, Mir-155 was first seen as an ALV common integration site in 1989 (Clurman and Hayward 1989), and we observe it in our tumors as well with 12 unique integrations, making it the 40th most common target of integration.
Figure 2.3: Common sites of ALV proviral integration

The top 48 clusters of integration are shown. Clusters were defined as any 50-kb region that harbors 12 or more unique ALV integrations. If a cluster is within or near a gene, all integrations within that gene and +/- 10kb from the gene transcript were also included. Density represents the number of integrations per kb in a given cluster. The average number of breakpoints per integration is shown for each gene. A higher number of breakpoints indicates increased clonal expansion of the cells carrying that integration. MYC did not penetrate the 12 integration threshold but is shown for comparison purposes.
TERT had the most clonal integrations identified in our study, with an average of 19.19 breakpoints per integration. This is consistent with earlier work analyzing a subset of these same tumors that identified 5 clonal or oligoclonal integrations upstream of the TERT transcription start site by inverse-PCR (Yang et al. 2007). The position and orientation of each these previously characterized integrations was successfully verified by high-throughput sequencing. In addition, 21 integrations upstream or within the TERT promoter were identified that had not been seen previously (Figure 2A). Like the integrations identified earlier, most of the novel TERT integrations (16/21) were in the opposite orientation of the TERT gene, and all but one occurred in birds infected at embryonic day 10 (Table 2.2).
Figure 2.4: Selected common ALV-A integration sites
Integration clusters for TERT and TNFRSF1 are shown. The orientation of each integration is indicated by the direction of the triangle and the tip of the triangle corresponds to the exact location of integration – integrations with 1 breakpoint are gray, 2 to 5 breakpoints are orange, and greater than 5 breakpoints are red. The extent of clonal expansion is indicated by the color of the integration marker. TERT integrations marked with an asterisk (*) are the same integrations identified previously (Yang et al. 2007) via inverse PCR. (Continued on next page)
Figure 2.4 (Cont.): Selected common ALV-A integration sites
Integration clusters for CTDSPL, CTDSPL2, and CXorf57 are shown. The orientation of each integration is indicated by the direction of the triangle and the tip of the triangle corresponds to the exact location of integration – integrations with 1 breakpoint are gray, 2 to 5 breakpoints are orange, and greater than 5 breakpoints are red. The extent of clonal expansion is indicated by the color of the integration marker. TERT integrations marked with an asterisk (*) are the same integrations identified previously (Yang et al. 2007) via inverse PCR.
Table 2.2: Expanded statistics for common integration sites

For each common integration site, the number of integrations is shown, followed by the nearest gene. If the integration cluster encompassed two genes they are both listed. Integration density is the number of integrations observed per kb for the area of the integration cluster. Avg. BP is the mean number of breakpoints observed per integration in the cluster. Tissue refers to the tissue which harbored the integration. Injection refers to the age of the chicken embryo when it was infected with ALV, 5 or 10 day embryos were used in this experiment. Virus refers to the strain of virus that produced each integration (see Materials and Methods for details). Orientation refers to the orientation of the provirus with respect to the nearest gene. If the cluster encompasses two genes and they are in opposite orientations, the orientation with respect to each gene in the cluster is shown.
Although MYC, MYB, Mir-155 and TERT have been seen in previous ALV insertional mutagenesis screens, most of the top targets of integration that we identified have not been identified in similar lower-throughput studies conducted previously. One such gene is TNFRSF1a – it was the most frequent target of integration that we observed, with a total of 117 unique viral integrations at this locus. TNFRSF1a is a member of the TNF-receptor superfamily and is one of the major receptors for tumor necrosis factor–alpha (TNF-alpha). TNFRSF1a can activate NF-kappaB and has known roles mediating apoptosis and regulating inflammation and cell proliferation (Wertz 2014). The vast majority of the integrations (82.9%) are within TNFRSF1a intron 1, and most are in the same orientation as the gene (92.3%) (Figure 2.4B, Table 2.2). The location and orientation of these integrations suggest that the virus is promoting the transcription of a TNFRSF1a transcript lacking exon 1. Exon 1 encodes part of the protein’s extracellular domain, which is crucial for the binding to its ligand TNF-alpha (Balkwill 2006).

Although this is a frequent target for ALV integration, it was only identified in two highly expanded clones (>10 breakpoints) and was almost never found outside of the bursa (113/117, 96.6% bursa, Table 2.2). These results suggest that ALV may be inducing a truncated receptor that is unable to bind TNF-alpha and mediate apoptosis. The fact that this integration is rarely found outside of the bursa suggests that this truncated gene product does not contribute to metastasis of the neoplasm to distant organs.
MEF2C was the second most targeted gene for ALV integrations, with a total of 43 unique integrations within 10 kb of this gene. MEF2C belongs to a family of transcription factors that have been shown to be important regulators of apoptosis, proliferation, survival, differentiation, and cancer (M. Zhang, Zhu, and Davie 2015). MEF2C has been observed as a common integration site in other retroviral insertional mutagenesis screens conducted in mice. This work has observed integrations most often within introns 1 and 2 and in the same orientation as the gene (Sørensen et al. 1996; Suzuki et al. 2002; Suzuki et al. 2006; Akagi et al. 2004). We observe a similar pattern of MEF2C integrations, with 21 of the 43 MEF2C integrations occurring in introns 1 or 2, although we observed no preference for integration in the same orientation as the gene (Table 2.2).

Two related phosphatases, CTDSPL (also known as RBSP3 or HYA22) and CTDSPL2 were also common integration sites with 30 and 21 unique integrations, respectively. Both genes belong to a gene family of C-terminal domain phosphatases and contain a conserved Dullard-like phosphatase domain (Marchler-Bauer et al. 2015). CTDSPL is a known tumor suppressor that can dephosphorylate RB1 and affect cell cycle progression (Kashuba et al. 2004). It is down-regulated in primary non-small cell lung cancer and has been shown to promote proliferation by modulating pRB/E2F1 in acute myeloid leukemia (Senchenko et al. 2010; Zheng et al. 2012). CTDSPL2 is less studied and has never been definitively linked to cancer. Recent work has shown that CTDSPL2 directly interacts with and dephosphorylates SMAD 1/5/8, which negatively regulates BMP signaling (Zhao et al. 2014). We observed a strong cluster of integrations for both genes. Integrations were clustered...
within intron 2 in CTDSPL and within introns 1 and 2 for CTDSPL2. A strong preference for integration in the forward orientation was observed for both genes (Figure 2.4C, 2.4D). This pattern suggests the virus may be producing a truncated protein product in both cases. The relatively high number of breakpoints, 5.87 on average for CTDSPL and 2.95 for CTDSPL2, indicates that the cells harboring these integrations experienced a moderate level of clonal expansion. Interestingly, liver tumors from 2 different birds accounted for 16/30 of the CTDSPL integrations and 17/21 of the CTDSPL2 integrations (Figure 2.5). This suggests that these two genes may cooperate in ALV-induced lymphomagenesis.
Figure 2.5: Two liver tumors exhibited multiple unique integrations within CTDSPL1 and CTDSPL2

The tumors G-158-L and B-256-L are each represented by a bar chart. Each bar represents a unique viral integration and the height of the bar corresponds to the number of sonication breakpoints observed for that integration (which is a measure of clonal expansion). Multiple high breakpoint CTDSPL (red) and CTDSPL2 (blue) integrations can be seen in these tumors. In addition, 5 CTDSPL and 8 CTDSPL2 one or two breakpoint integrations were observed for G-158-L and 1 CTDSPL and 1 CTDSPL2 single breakpoint integrations were observed for B-256-L (not shown).
CXorf57 was the 10th most frequently targeted common integration site, and is among the most enigmatic genes that we identified. CXorf57 is conserved in humans but has never been characterized and hence has no known function. CXorf57 encodes a protein that has a conserved putative replication factor-a protein 1 domain. Genes with this domain that have been characterized have been shown to be involved in recognition of DNA damage for nucleotide excision repair (Brill and Stillman 1991; Marchler-Bauer et al. 2015). CXorf57 contains 24 unique integrations that are spaced throughout the gene and in no preferred orientation (Figure 2.4E). This integration pattern indicates that these proviral integrations may be disrupting the normal transcription of this gene, suggesting that it could be a novel tumor suppressor. Interestingly, a strong preference for integration in liver tumors was observed (18 of 24 integrations, Table 2.2).

**Functional annotation enrichment analysis of ALV common integration sites**

To determine whether these 48 major common integration sites (Figure 2.3) are enriched for genes of specific functions or involved in specific pathways, we conducted gene-annotation enrichment analysis with DAVID (Huang, Sherman, and Lempicki 2009a). We identify six enriched KEGG pathways and processes, most of which are related to cancer or are pathways active in immune cells (Figure 2.6). GO term analysis revealed strong enrichment (p < 0.005) for a number of different gene ontologies (Figure 2.7). The most significant enrichment was seen for regulators of transcription (both positive and negative). Additionally, strong enrichment was
observed for several types of positive regulators of metabolic and biosynthetic processes as well as several anti-apoptotic functional terms.
Figure 2.6: ALV-A KEGG pathway analysis

KEGG pathways enriched among the top 48 common integration sites.
Figure 2.7: Gene ontology analysis
Enriched gene ontology terms identified by DAVID for the top 48 common integration sites.
**ALV integration has a weak palindromic consensus sequence *in vivo***

It was shown in earlier work that ALV integration sites have a weak palindromic consensus sequence when integrating into human DNA (Wu et al. 2005; Holman and Coffin 2005). These analyses were performed in human cells in culture that had been engineered to express the TVA receptor, enabling them to be infected with ALV. To determine whether ALV exhibits a similar preference in its canonical host *in vivo*, we performed a similar analysis on our full dataset of integrations in chicken. We observed very similar results to those seen in human cell culture (Figure 2.8). For example, a strong preference for a T, –3 nucleotides from the viral integration site was observed. In addition, a strong preference for G/C at position 1 and A at position 9 was also observed. Notably, the nucleotide frequencies that we observe are nearly exactly what was seen in cultured human cells. For example, we calculated the frequency of T at position -3 to be 47%, which is exactly the same frequency reported in human cells (Wu et al. 2005). The preference for G/C at position 1 was 68.8% and was 71% in human cells, and the preference for A at position 9 was 39.8% in our study and 43% in human cells. These results show that the consensus sequence observed in human cells infected with ALV is also true *in vivo* in the virus’ natural host.

Interestingly, as with previous studies, we observed that the ALV consensus sequence is slightly asymmetric. This contrasts with other retroviruses such as HIV and MLV that have symmetric consensus sequences (Wu et al. 2005; Holman and Coffin 2005). One explanation may be the presence of 5-base duplications. Although
it has been shown that ALV integrase typically generates 6-base duplications, there are indications that 5-base duplications are possible under certain circumstances (Oh et al. 2006; Oh, Chang, and Hughes 2006). If a 5-base duplication is generated by ALV integrase at sufficient frequency, this could reduce the nucleotide preferences that we observe to the right of the duplication (Figure 2.8) but not on the left, which could explain the asymmetry that we observe. Further work characterizing ALV integrase duplications may be able to shed light on this issue in the future.
Figure 2.8: Consensus ALV-A target integration site

A. A sequence logo displaying the consensus sequence surrounding ALV integration sites in this study. The vertical black line represents the viral integration site and the 6 nucleotides of host sequence duplicated during viral integration are boxed. The arrow indicates the axis of symmetry. B. Base frequencies in the chicken genome at ALV integration sites are shown.
**ALV prefers integration near promoters and within genes* in vivo*

To determine whether ALV prefers integration near certain features* in vivo*, we employed the HOMER software suite (Heinz et al. 2010). 27,770 unique ALV integrations and an equal number of random, computer-generated integrations were annotated with the nearest genomic feature. This analysis revealed a preference for integration near promoters (Figure 2.9). To better understand the pattern of integrations surrounding the transcription start sites (TSSs), we plotted all integrations with respect to the nearest TSS (Figure 2.10). We observed enrichment for ALV integration within 30 kb of the TSS, with a stronger preference for integration upstream than downstream of the TSS. In addition, we observed a sharp drop in integration frequency within 50 bp of the TSS (Figure 2.10B). This pattern is similar to that seen in murine leukemia virus (MLV), and is believed to be due to the occupancy of this area by basal transcriptional machinery such as transcription factor IID (TFIID) (Cattoglio et al. 2010).
Enrichment for integration near genomic features was calculated with HOMER (Heinz et al. 2010). Fold enrichment was calculated by comparing ALV integrations to a randomly generated integration dataset. Promoters are defined as the region -1kb to +100bp from transcription start sites, while transcription termination sites (TTS) are defined as the region -100bp to +1kb flanking the transcription termination site.
Figure 2.10: ALV-A integrations with respect to transcription start sites

A. Integrations within 10 kb of transcription start sites are shown, placed into 100-bp bins. The red line represents ALV-A integrations, the black line represents randomly simulated integrations. A preference for integration is observed immediately flanking the TSS. B. Integrations within 1-kb of TSSs are shown in 10-bp bins. A striking lack of integrations was observed near the TSS. (Continued on next page)
Figure 2.9 (cont.) : ALV-A integrations with respect to transcription start sites

C. Integration frequency was calculated for expanded clones (red), non-expanded clones (blue), and randomly generated integrations (black), and are presented in 500 bp bins. Integration frequency is the fraction of total integrations which falls into each 500 bp bin. Integrations near the TSS are shown to be slightly more likely to result in clonal expansion.
Earlier work on ALV integration in cell culture has showed that the virus has a slight preference for integration near transcribed elements, but a strong preference for integration centered on transcription start sites was not reported in earlier studies (Narezkina et al. 2004; Mitchell et al. 2004; Barr et al. 2005). It’s possible that this inconsistency with earlier reports may be explained by the fact that we sequenced integrations that occurred in vivo. Hence, many of the integrations have been subject to selection, especially those found in clonally expanded cells. To determine the extent to which integrations in clonally expanded cells are affecting observed enrichment for integrations near TSSs, integrations that show evidence of clonal expansion were analyzed separately from those for which only a single sonication breakpoint was observed. This analysis shows that even integrations that show no evidence of clonal expansion show enrichment for integration near TSSs. It is possible that selection is still at work in the cases of integrations that are not clonally expanded, if for example, the gene near the integration promotes cell survival, but not proliferation.

This analysis also revealed preference for integration near other genomic features as well (Figure 2.9). Integration near promoters (-1kb to +100bp from transcription start sites) was the most enriched compared to the control, with a 1.75 fold increase. Other features for which enrichment was observed include exons (1.72 fold), 3′ UTRs (1.57 fold), transcription termination sites (-100bp to +1kb, 1.55 fold), and introns (1.36 fold). 5′ UTRs exhibited no increase in ALV integration vs control while intergenic regions were less likely to harbor ALV integrations than random (0.91 fold).
Discussion

In this study, we characterize the integration of proviruses in ALV-A-induced B-cell lymphomas with high-throughput sequencing. This method allows for a much more detailed characterization of integration sites than was possible in earlier studies of these types of neoplasms.

We observed that promoters and TSSs are preferred sites of ALV integration in vivo (Figure 2.9, Figure 2.10). This preference has not been seen in previous studies of ALV integration. Analyses of other retroviruses such as HIV and murine leukemia virus (MLV) have shown that MLV but not HIV prefers integration near TSSs (Cattoglio et al. 2010). MLV’s preference for the TSS is mediated by the binding of bromodomain and extraterminal domain (BET) proteins to the MLV integrase (Sharma et al. 2013). MLV also is known to prefer integration within 2.5 kb of the TSS, with greater preference upstream of the TSS than downstream. In addition, a strong decrease in MLV integration frequency has been shown within 100 bp of the TSS (Cattoglio et al. 2010).

The pattern of ALV integration that we report is very similar to MLV, but not identical. For example, while we observed a greater preference for integration upstream of the TSS and a sharp dropoff within 100 bp of the TSS (Figure 2.10), we do not observe a narrow peak of increased integration frequency +/- 2.5 kb from the TSS. Instead, we see a broader peak of elevated integration frequency that stretches as far as 20 kb on either side of the TSS (Figure 2.10C). Also, we observed a weaker preference for ALV integration in the immediate vicinity of the TSS than has been
published for MLV. Previous work has calculated a 4.7 fold increase in the frequency of MLV integrations within 5 kb of the TSS, although recent work has shown this can vary by cell type (Wu et al. 2003; LaFave et al. 2014). In contrast, we observed only a 2.3-fold increase for ALV over that range. Because our experiments were conducted in vivo where cells are subject to selection and clonal expansion, the preference for ALV integration that we observe may be partially due to these additional variables. This may explain why the TSS was not a preferred site of integration in earlier studies in cell culture.

To date, only four genes have been shown to be common integration sites in ALV-A induced B-cell lymphoma - MYC, MYB, Mir-155 and TERT. Here we identify all four of these genes as common integration sites, as well as a host of new genes that have not been previously implicated in ALV-induced lymphomagenesis.

Three reports have been published previously that partially characterize 8 of the 28 tumors that we analyze in this study (Table 2.1). Two of these publications utilized nested-PCR to map proviral integrations at the MYB promoter and showed some tumors contained one or more integrations into the MYB locus (P. E. Neiman et al. 2003; Polony et al. 2003). A third report used inverse-PCR to map proviral integrations (which is not biased to a specific locus), and showed multiple integrations in the TERT promoter in the opposite orientation (Yang et al. 2007).

By re-analyzing these tumors, we were able to verify many of the integrations seen in previous studies. First with regards to TERT, we verified by deep sequencing all 5 TERT promoter integrations that were described previously,
and identified an additional 21 integrations at the TERT locus in both newly analyzed and re-analyzed tumors. Previous work also showed that these integrations were clonal or oligoclonal by Southern blot, meaning that the integrations were present in a large fraction of cells in the tumor (Yang et al. 2007). Deep sequencing results confirm this finding - all of these integrations exhibited extensive clonal expansion by breakpoint analysis. Overall, TERT was the 8th most frequent target of integration with 26 unique integration sites identified by deep sequencing. Although it was not the most frequent target of integration, TERT integrations were often highly expanded, with an average of 19.19 sonication breakpoints observed per integration, which may explain why it was identified so readily by inverse-PCR in previous work. The extensive expansion of clones containing TERT integrations is consistent with the hypotheses that TERT activation is an early event in tumorigenesis.

MYB was the 5th most targeted gene with 28 unique integrations. Only one of these integrations was described in previous work (A2-R588 Liver, Table 2.1), suggesting that many of the MYB integrations identified in earlier work were not clonal, and possibly only present in a small number of cells.

Historically, MYC and Mir-155 were often seen in ALV-induced B-cell lymphomas. Both genes were prominent integration clusters in this study (Figure 2.3). As for Mir-155, we identified 12 unique Mir-155 integrations, making it tied for the 40th most targeted gene for proviral integration. Earlier studies have shown that Mir-155 integrations are often seen in metastatic tumors, which led to the
hypothesis that that Mir-155 is a late event in ALV tumor induction and may play a role in metastasis (Hayward, Neel, and Astrin 1981). 11 of the 12 Mir-155 integrations we observed occurred in metastatic liver tumors, with only one seen in the primary bursa in our study (Table 2.2), which is consistent with this hypothesis.

In this study, we identified only 9 integrations in the MYC locus (Figure 2.3). MYC was the first gene ever identified as a common integration site in ALV-induced lymphomas, and MYC integrations have since been seen in many studies of these neoplasms (Hayward, Neel, and Astrin 1981; G. S. Payne, Bishop, and Varmus 1982; Baba and Humphries 1985). The time of infection is thought to be an important factor in the development of MYC-associated tumors, with later infections (especially after hatching) more likely to induce tumors with MYC integrations. In contrast, we infected birds much earlier, at embryonic day 5 or day 10. Interestingly, all 9 of the MYC integrations occurred in birds that were infected at day 10, while no MYC integrations were observed at day 5 (Table 2.2). This supports the idea that the early timing of injections may explain why we see fewer MYC integrations than earlier work.

Interestingly, the most frequent target of integration was TNFRSF1a. This gene is a receptor for tumor necrosis factor–alpha (TNF-alpha). TNFRSF1a can activate NF-kappaB and has known roles mediating apoptosis and regulating inflammation and cell proliferation (Wertz 2014). Although TNFRSF1a harbored 117 unique integrations, it was only highly clonally expanded (>10 breakpoints) in two cases. This lack of highly expanded clones may explain why this gene was not
identified in previous experiments mapping ALV integration sites. The vast majority of the integrations occurred in the first intron of the gene and in the same orientation as the gene, and integrations were almost exclusively found in bursal tissues and not in metastatic tumors (Table 2.2). These data suggest that the integration may be producing a truncated protein product and that this product does not contribute to metastasis or proliferation, but gives the cell a survival advantage in the bursa.

While ALV-A induces lymphoid neoplasms, ALV-J is known to induce myeloid neoplasms and hemangiomas. We recently reported integrations in ALV-J induced hemangiomas, and interestingly we see very little overlap between the common integration sites in ALV-A induced lymphoid tumors and ALV-J induced hemangiomas. The only gene that appears to be shared as a common integration site between the two studies is ELF1, which was the second most frequently targeted gene in ALV-J hemangiomas and the 13th most frequent target of integration in ALV-A lymphoid tumors. The striking lack of overlap between these datasets is likely due to the biological differences between the types of cells affected and the genes involved in inducing lymphomas vs. hemangiomas.

Recent work characterizing HIV integrations identified BACH2 and MKL2 as common integration sites in individuals on suppressive combination antiretroviral therapy (cART) (Ikeda et al. 2007; Maldarelli et al. 2014). We identify BACH2 but not MKL2 as a common integration in this study. In one earlier study, BACH2 integrations showed a strong preference for integration in the forward orientation
(15/15 integrations), and 6 of 15 integrations were found in expanded clones. In ALV lymphomas, we see a weaker preference for integration in the forward orientation (17/24, 70.8%) with 5 of 24 present in clonally expanded cells. Although MKL2 was not a common integration site in our study, we did identify the related gene MKL1 as a common integration site. Both MKL1 and 2 are coactivators of the transcription factor SRF, which regulates genes involved in many biological processes including cell growth and migration (Pipes, Creemers, and Olson 2006).

In conclusion, this study greatly expands the number of genes known to be common integration sites in ALV induced B-cell lymphoma. As one might expect, many of the genes we identify have well characterized roles in cancer and related processes. These genes include RUNX1, Mir-221, Mir-222, IKZF1, CCNA2, ZEB1, CBLB, HMGB1 and many others. In addition to canonical cancer genes, we identified a number of genes as common integration sites that are conserved in human but have never been linked to cancer. These include CXorf57, CTDSPL2, TMEM135, ZCCHC10, FAM49B, and MGARP. In fact, three of these six genes, CXorf57, ZCCHC10, and FAM49B, not only have never been linked to cancer, but have never undergone any characterization and have no known functions. We think these genes, as well as others that we identify in this study, are interesting targets for further research.
Chapter 3. Characterization of ALV-J integration sites in hemangiomas by high-throughput sequencing

Adapted from: Justice IV, James, Sanandan Malhotra, Miguel Ruano, Yingying Li, Guillermo Zavala, Nathan Lee, Robin Morgan, Karen Beemon. 2015). The MET gene is a common integration target in avian leukosis virus subgroup J-induced chicken hemangiomas. Journal of Virology. 89(9):4712-9
Abstract

Avian leukosis virus, subgroup J (ALV-J), is a simple retrovirus that can cause hemangiomas and myeloid tumors in chickens and is currently a major economic problem in Asia. Here we characterize ALV-J strain PDRC-59831, a newly studied US isolate of ALV-J. Five-day-old chicken embryos were infected with this virus, and the chickens developed myeloid leukosis and hemangiomas within two months after hatching. To investigate the mechanism of pathogenesis, we employed high-throughput sequencing to analyze proviral integration sites in these tumors. We found expanded clones with integrations in the MET gene in two of the five hemangiomas studied. This integration locus was not seen in earlier work characterizing ALV-J-induced myeloid leukosis. MET is a known proto-oncogene that acts through a diverse set of signaling pathways and is involved in many neoplasms. We show that tumors harboring MET integrations exhibit strong overexpression of the MET mRNA.

Introduction

Avian leukosis viruses (ALVs) are classified into subgroups based on their envelope gp85 surface glycoprotein (SU), viral cross-neutralization patterns, and host range. Most avian retroviruses are classified as subgroups A, B, C, D, or E. ALV of subgroup J (ALV-J) was first isolated in 1988 in the United Kingdom; the prototype strain HPRS-103 causes primarily myeloid leukosis, but can induce other tumor types at low incidence (L. N. Payne et al. 1991; L. N. Payne, Gillespie, and Howes 1992; Justice IV and Beemon 2013). ALV-J is thought to have originated
from a recombination event between an exogenous ALV and an ancient endogenous avian (EAV) retroviral element (Bai, Payne, and Skinner 1995; L. M. Smith et al. 1999). This recombination event incorporated the endogenous retroviral env into ALV-J.

Since its discovery in the UK, a variety of ALV-J strains have been characterized in diverse geographical areas, including North America, Europe, East Asia, Australia, and the Middle East (Benson et al. 1998; Sung et al. 2002; Malkinson et al. 2004; Landman et al. 2002; Fenton, Reddy, and Bagust 2005; Fadly and Smith 1999). It is believed that these isolates derive from a single common ancestor, and are not the result of independent recombination events (Benson et al. 1998). The types of neoplasms caused by ALV-J vary and can be influenced by the specific strain of ALV-J that has infected the bird. Most often, the virus induces tumors of myeloid origin (as with HPRS-103), but some strains induce primarily hemangiomas. These are vascular tumors found in the skin or visceral organs that originate from the endothelial cells that line blood vessels (Fadly and Nair 2008). Other ALV-J strains are capable of inducing both myeloid tumors and hemangiomas. We found that the strain used in this study falls into the third category, inducing both hemangiomas and myeloid tumors at high incidence.

Besides myeloid tumors and hemangiomas, ALV-J has been shown to induce other types of tumors at low frequency such as skeletal myelocytomas, renal tumors, histiocytic sarcomas, and others (Venugopal 1999; L. N. Payne, Gillespie, and Howes 1992). This pathology contrasts sharply with that of the more studied
ALV-A, which induces mainly B-cell lymphomas but also erythroblastomas (P. Neiman, Payne, and Weiss 1980; Kanter, Smith, and Hayward 1988; Nilsen et al. 1985).

ALV-J infection can cause significant economic loss due to reduced egg production, stunted growth, and early death. The economic losses have been particularly extensive in China, where the virus commonly infects poultry (Gao et al. 2010). It has been recently shown that ALV-J infection is not limited to domesticated chickens. In fact, infection with ALV subgroups A, B, and J appears to be widespread in wild fowl throughout China (D. Li et al. 2013; L. Jiang et al. 2014).

ALVs do not carry a viral oncogene and instead cause neoplasia through insertional mutagenesis (Beemon and Rosenberg 2012). In order to complete the viral life cycle, all retroviruses must integrate into the genomic DNA of the infected cell. Thus, the provirus can act as a mutagen, landing within a gene and ablating its function. Alternatively, because the virus has potent enhancers and promoters in its long terminal repeats (LTRs), ALV can induce the expression of genes located near an insertion site. This process can drive tumor formation if the provirus integrates near and perturbs expression of cancer-related genes.

Virus-induced mutagenesis can be exploited to identify genes that may play a role in driving development of neoplasms. For example, the virus can be used to induce tumor formation, and common integration loci can be identified. These common integration sites (CISs) flag a genomic locus as potentially harboring an oncogene or tumor suppressor. Viral insertional mutagenesis screens have been
fruitful in identifying cancer genes in several model systems (Beemon and Rosenberg 2012). ALV is an especially useful virus for such a screen because it integrates in a largely random fashion, with only slight preference for active transcriptional units (Barr et al. 2005; Mitchell et al. 2004). This ensures that as many genomic loci as possible are probed for oncogenic potential by the virus.

Earlier studies implicated several genes as drivers of tumorigenesis in ALV-induced neoplasms. In ALV-A-induced B-cell lymphomas, common integration sites were identified near or within MYC, MYB, MIR-155, and TERT genes (Hayward, Neel, and Astrin 1981; G. S. Payne, Bishop, and Varmus 1982; Clurman and Hayward 1989; Yang et al. 2007). The study of proviral integrations within ALV-J-induced neoplasms has only recently begun. Early work has shown MYC, TERT, and ZIC1 to be targets of proviral integration in ALV-J-induced myeloid leukemia (ML) (Y. Li et al. 2014).

In this study, we conducted an insertional mutagenesis screen to identify the genes involved in ALV-J-induced tumors. To identify viral integration sites in these tumors, we employed high-throughput sequencing on the Illumina platform. We identified intron 1 of the MET gene as a common integration site in hemangiomas. MET encodes a well-studied receptor tyrosine kinase that binds hepatocyte growth factor/scatter factor and plays important roles in normal development and a wide range of human cancers (C. R. Graveel, Tolbert, and Vande Woude 2013). Because we observed integrations near a known oncogene in multiple tumors, we
hypothesize that ALV-J tumors, like those induced with ALV-A, are generated by insertional mutagenesis.

**Materials and Methods**

**Tumor induction**

ALV-J strain PDRC-59831 was isolated from a 38-week-old broiler breeder chicken. The case was recorded on 5/30/2007 on a farm near Danielsville, GA (Malhotra et al. 2015). In this study, ALV-J strain PDRC-59831 was inoculated into thirty 5-day-old SPAFAS embryos (Charles River) via the yolk sac route. Four embryos died at day 10 of embryogenesis, leaving twenty six viable embryos. Of these, eleven hatched (11/26; 42%), five pipped but could not complete hatching, and ten embryos did not hatch and died. In comparison, 90% (19/21) of the uninoculated control chickens hatched. Chickens were observed daily and euthanized when apparently ill or at 12 weeks of age. Of the eleven ALV-J-infected chickens, one was euthanized at week 2 and three died at week 11 for reasons unrelated to infection. One died at week 5 and one was euthanized at week 7, both had tumors. Five were euthanized at 12 weeks of age, all of which had tumors. In total, tumor tissue was obtained from six ALV-infected chickens (Table 3.1). The early death of bird #7 prevented the collection of useful tissue. Tumors were classified by gross examination, tissue samples were collected and flash frozen, and then stored at -80°C.
Table 3.1: ALV-J tissues collected and neoplasms observed

X = no neoplasm, H = hemangioma, M = myeloid tumor, - = tissue was not collected and no neoplasm was observed during dissection. The tissues from Bird #7 (B7) were not analyzed in this study.

<table>
<thead>
<tr>
<th>Bird</th>
<th>Bursa</th>
<th>Liver</th>
<th>Spleen</th>
<th>Brain</th>
<th>Kidney</th>
<th>Breast</th>
<th>Leg</th>
<th>Intestine</th>
<th>Lung</th>
<th>Head</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>M</td>
<td>H</td>
<td>-</td>
<td>-</td>
<td>H</td>
</tr>
<tr>
<td>B2</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>H</td>
</tr>
<tr>
<td>B3</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>H</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B4</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>H</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B5</td>
<td>X</td>
<td>M</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B6</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>M</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>B7</td>
<td>X</td>
<td>H</td>
<td>X</td>
<td>X</td>
<td>M</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C2</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Phylogenetic analysis of ALV-J isolate PDRC-59831 env gene

The sequence of the ALV-J env gene was determined by Sanger sequencing (NCBI accession number KP284572). Phylogenetic analysis was conducted using the Neighbor-Joining method (Saitou and Nei 1987). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches (Figure 3.1) (Felsenstein 1985). The evolutionary distances were computed using the Maximum Composite Likelihood method (Tamura, Nei, and Kumar 2004) and are in the units of the number of base substitutions per site. The analysis involved 16 nucleotide sequences. Accession numbers are as follows: HUB09JY03 (HQ634811), HPRS-103 (Z46390), UD4 (AF307951), UD5 (AF307592), PDRC-59831 (KP284572), YZ9902 (HM235670), NX0101 (DQ115805), NM2002-1 (HM235669), JS-nt (HM235667), SD07LK1 (FJ216405), SCDY1 (HQ425636), NHH (HM235668), JL093-1 (JN624878), JS09GY6 (GU982310), JS09GY3 (GU982308), ADOL-7501 (AY027920). Codon partitions included were 1\textsuperscript{st}+2\textsuperscript{nd}+3\textsuperscript{rd}+noncoding. All positions containing gaps and missing data were eliminated. These analyses were conducted in MEGA6 (Tamura, Nei, and Kumar 2004).
Figure 3.1: Sequence comparison and phylogenetic analysis of env genes of ALV-J isolates

Phylogenetic analysis was conducted using the Neighbor-Joining method. The optimal tree with the sum of branch length = 0.37621 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. Strain PDRC-59831 (bold) is shown to cluster closest to UD4, UD5, and YZ9902 – and on the same branch as prototype ALV-J isolate HPRS-103.
DNA extraction and deep sequencing

50 mg of tissue was homogenized with Kimble-Chase Kontes™ Pellet Pestle™, and then digested with Proteinase K at 50°C for 15 hours. The sample was then phenol extracted, put through a 25 gauge needle ten times, and ethanol precipitated; this procedure was repeated. The sample was then treated with 2 μg RNase A for 1 hour at 37°C. Phenol extraction, shearing, and ethanol precipitation were repeated, and DNA concentration was measured with a Thermo Scientific Nanodrop 2000c. 5 μg of purified genomic DNA was sonicated with a Bioruptor™ UCD-200. End repair, A-tailing, and adapter ligation were performed as described by Gillet et al. (Gillet et al. 2011b) (adapter short arm: p-gatcggaagagacaaaaaaaaaaaaaa, adapter long arm: caagcagaagacggcatacgagatxxxxxxgtgactggagttcagacgtgtgctcttccgatc*t where the x's denote the barcode sequence). Multiplexed barcodes each differed by at least two nucleotides. Two rounds of PCR (nested) were conducted to enrich the library for proviral junctions. The first PCR reaction had 23 cycles, and employed one ALV-J specific primer (gggactgtagcatgtataggcgctgag) between 3’ LTR and env, and a second primer within the adapter ligated on earlier (caagcagaagacggcatacgagat). The second round of PCR employed a primer (aatgatacgccgaccacccgatatcactcgacgtggcagacatgaag) at the 3’ end of the LTR, twelve nucleotides short of the junction between virus and genomic DNA, as well as a nested adapter primer overlapping the barcode sequence within the adapter (caagcagaagacggcatacgagatxxxxxx) (Figure 3.2). Libraries were quantified by qPCR, and then underwent single-end 100 bp multiplexed sequencing on the Illumina Hi-Seq 2000. A custom sequencing primer
(acgattgcagcacctgaatgaagtgaaagg) was used that hybridized near the end of the viral 3’ LTR, five nucleotides short of the junction between viral and genomic DNA. This allowed for reads to be validated as genuine integrations, by verifying that each read begins with the last 5 nucleotides of the proviral DNA. In all, 9,759,304 junction reads were obtained.

**Sequencing Analysis**

Reads were first filtered with a custom python script to remove sequences that did not begin with the last five nucleotides of viral DNA, “CTTCA.” A total of 1,352,053 reads (13.85%) were discarded because they did not begin with this verification sequence. The files were then uploaded to Galaxy (Giardine et al. 2005; Blankenberg, Von Kuster, et al. 2010; Goecks et al. 2010), which was used to perform downstream analyses. In Galaxy, the quality scores were first converted to Sanger format with FastQ Groomer v1.0.4 (Blankenberg, Gordon, et al. 2010). Adapters were then trimmed using the Galaxy Clip tool v1.0.1. This tool also removed reads containing an N (14,871, 0.18% of reads removed at this step) and reads less than 20 nucleotides in length after adapter removal (4,289,006, 51.02%). The remaining reads (4,103,370, 48.81%), were mapped with bowtie (Langmead et al. 2009) using the Gallus gallus 4.0 genome (Nov. 2011). 100,000 random mapped reads were selected from each sample to be used for further analysis. If less than 100,000 reads were present for a sample, all available reads were used. A custom Perl pipeline was developed to analyze the aligned reads output from Bowtie. Briefly, reads containing sequencing errors were filtered, and read counts and
breakpoints were quantified. Integrations found in multiple samples were assigned to the sample with the highest number of breakpoints. Integration clusters were identified by walking the genome. If two integrations were within 5 kb of each other they were placed into an integration group. Additional integrations were included in the group until an integration-free span of 5 kb was reached. Files were annotated with refseq features and the orientation and distance to the nearest gene were calculated for each integration. The source code is available upon request.

**MET gene expression**

RNA transcripts were isolated from all five hemangiomas, three myeloid tumors, and four non-tumor controls with RNA-Bee (Tel-Test). First-strand synthesis was performed with a poly-dT primer and Superscript III reverse transcriptase (Invitrogen) following the manufacturer’s protocol. Quantitative reverse transcription PCR (qRT-PCR) was conducted with iQ SYBR Green Supermix (Biorad) according to manufacturer’s protocol. MET primers (aggacattttgggtgtgtgt, aactgagccacttcttccag) were designed using primer3plus software (www.primer3plus.com). Thermal cycling was conducted on a Biorad C1000 Thermal Cycler/CFX96 Real-time System. MET expression was normalized to GAPDH using primers published previously (Heidari, Zhang, and Sharif 2008). PCR was repeated four times and each sample was present in duplicate during each run. Results were normalized to B3 Kidney Hemangioma using the comparative C_T method. Melt curves were conducted to verify specificity of the primers.
Results

Characterization of ALV-J isolate PDRC-59831

The tumors in this study were induced by ALV-J strain PDRC-59831, an American isolate of ALV-J that has not been previously described in the literature. Strain PDRC-59831 was isolated in 2007 from a meat-type chicken. The provirus was isolated from infected cells by PCR, four isolates were sequenced, and the consensus sequence was submitted to Genbank (KP284572). The env gene sequence was compared to env from other strains of ALV-J isolated from around the world. Phylogenetic analysis of PDRC-59831 env (Figure 3.1) shows a close relationship with two other American ALV-J isolates, ALV-UD4 (96.2% similarity) and ALV-UD5 (96.3%), as well as Chinese isolate YZ9902 (95.3%). These isolates along with PDRC-59831 cluster with the original UK isolate HPRS-103 (95.0% similarity).

Several studies have reported certain genetic alterations in ALV-J strains that have a propensity for generating hemangiomas. For example, an 11-nucleotide deletion was observed in the U3 region of the LTR of ALV-J strain SCDY1 and NHH (Shi et al. 2011). Also, two different 19-nucleotide insertions—one in the 5′ untranslated region (5′UTR) and another in the U3—were identified in hemangioma-inducing strains JL093-1, SD09DP03, and HLJ09MDJ-1 (Pan et al. 2011). Sequence alignments were carried out to determine if PDRC-59831 contained any of these alterations, and it does not. Instead, it shares similarity with the ML-inducing prototype strain HPRS-103 in these regions (data not shown). This suggests that these genetic alterations are not required to induce hemangioma, but
it does not rule out their possible contribution. Recently, a 205-nucleotide deletion in the 3′ UTR has been identified in some strains of ALV-J, that leads to higher oncogenicity and lethality rates in infected chickens (Wang et al. 2012). We determined by alignment that PDRC-59831 contains a 216-nucleotide deletion in this region. This deletion overlaps 187 of the 205 nucleotides described previously.

**Sequencing of ALV-J/cellular integration junctions**

All 7 birds infected with PDRC-59831 developed tumors (hemangiomas, myeloid tumors, or both) by twelve weeks of age. A total of six hemangiomas and four myeloid tumors were observed in total (Table 3.1). Genomic DNA was isolated from these tumors, randomly sheared by sonication, and proviral junctions were amplified in two successive rounds of nested PCR as outlined in Figure 3.2. Proviral integration/genomic DNA junctions were deep sequenced on the Illumina Hi-seq 2000 and mapped onto the Gallus gallus genome.
Genomic DNA was isolated from tumor and non-tumor samples. DNA was sonicated, end repaired, and A-tailed. Adapters containing barcodes were ligated onto the DNA fragments and they underwent two rounds of PCR (nested). The final product underwent multiplexed sequencing on the Illumina Hi-seq 2000.
We chose to sonicate the DNA to induce random fragmentation. This introduces different sonication breakpoints to unique DNA molecules across multiple cells. Sometimes multiple cells in the sample carry the same viral integration site due to clonal expansion. When this is the case, shearing the DNA from these cells can produce multiple fragments that share the same integration site but have different sonication breakpoints. By quantifying these breakpoints after deep sequencing, we were able to determine the relative abundance of an integration with respect to the other integrations in the sample. We refer to integrations that exhibit more than one sonication breakpoint as “expanded clones” (Figure 3.3A, Figure 3.4).
Figure 3.3: Distribution of integration sites in ALV-J induced hemangiomas

(A) A total of 30,850 breakpoints were identified in the ALV-J induced hemangiomas. Genes which had 10 or more breakpoints are highlighted in a separate pie. Each slice represents a unique integration site, and the size of the slice represents the number of breakpoints for that site. The table shows integrations with the highest number of breakpoints, along with the tumor in which that integration was identified. (B) The top ten most abundant integrations identified in the head and leg hemangiomas are shown. Integrations in the MET gene represent the most abundant integration sites in these tumors.

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Hemangioma</th>
<th># Breakpoints</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRIO</td>
<td>Lung</td>
<td>27</td>
</tr>
<tr>
<td>EYA4</td>
<td>Kidney</td>
<td>24</td>
</tr>
<tr>
<td>MET</td>
<td>Head</td>
<td>23</td>
</tr>
<tr>
<td>MET</td>
<td>Head</td>
<td>21</td>
</tr>
<tr>
<td>FBXW7</td>
<td>Head</td>
<td>16</td>
</tr>
<tr>
<td>RASA1</td>
<td>Lung</td>
<td>14</td>
</tr>
<tr>
<td>GOLGA1</td>
<td>Lung</td>
<td>12</td>
</tr>
<tr>
<td>CDC42BPB</td>
<td>Kidney</td>
<td>12</td>
</tr>
<tr>
<td>GTF3C6</td>
<td>Lung</td>
<td>11</td>
</tr>
<tr>
<td>KPNAA4</td>
<td>Lung</td>
<td>11</td>
</tr>
<tr>
<td>MXRA8</td>
<td>Kidney</td>
<td>11</td>
</tr>
<tr>
<td>Unplaced Contig</td>
<td>Lung</td>
<td>10</td>
</tr>
<tr>
<td>FERIL6</td>
<td>Lung</td>
<td>10</td>
</tr>
<tr>
<td>PLEKHC1</td>
<td>Lung</td>
<td>10</td>
</tr>
<tr>
<td>ncRNA LOC101750336</td>
<td>Kidney</td>
<td>10</td>
</tr>
<tr>
<td>ncRNA LOC101751731</td>
<td>Lung</td>
<td>10</td>
</tr>
<tr>
<td>ACSL6</td>
<td>Lung</td>
<td>10</td>
</tr>
</tbody>
</table>

Head Hemangioma

- MET
- MET
- FBXW7
- FIM1
- KIAA03930
- SLC39A8
- HS110P3
- ALG13
- ITPK1
- FOXG1

Leg Hemangioma

- MET
- SEMA3B
- GRIP1
- DOCK1
- GHT3
- RNAS2
- DOCK1
- SNX27
- FURIN
- LGR5

A total of 30,850 breakpoints were identified in the ALV-J induced hemangiomas. Genes which had 10 or more breakpoints are highlighted in a separate pie. Each slice represents a unique integration site, and the size of the slice represents the number of breakpoints for that site. The table shows integrations with the highest number of breakpoints, along with the tumor in which that integration was identified. The top ten most abundant integrations identified in the head and leg hemangiomas are shown. Integrations in the MET gene represent the most abundant integration sites in these tumors.
There were 9 integration sites at the MET locus observed in 5 hemangiomas (above) as well as 10 integration sites in the 24 non-tumor tissues (below) from infected birds. No MET integrations were observed in myeloid tumors. Integrations where multiple breakpoints were observed are represented as solid arrows, and the corresponding number of breakpoints is noted. Arrowheads represent the genomic location and the transcriptional orientation of each provirus. Most of the multi-breakpoint integrations are observed in intron 1 region of the MET gene.

**Figure 3.4: ALV-J integration in the MET gene in chicken hemangiomas and non-tumor tissues**
A total of 25,961 integrations were identified among five ALV-J-induced hemangiomas analyzed. Integrations were mapped and quantified with a custom data analysis workflow (see Methods). 15,738 integration “clusters” were identified among all myeloid tumor, hemangioma, and non-tumor tissues. These clusters were defined as any genomic loci containing at least two integrations where each integration is separated by no more than 5 kb. The top 20 clusters are shown in Table 3.2. Although these loci are sites of frequent integration, breakpoint analysis showed that these integrations are not found in highly expanded clones.

Interestingly, endogenous retrovirus LOC101750146 is tied for the most frequent target of integration among the hemangiomas we analyzed, with a total of 29 integrations (Table 3.2). However, none of these clones were highly expanded and many integrations were also seen in non-tumor tissues, so these integrations are unlikely to be relevant to tumor induction.
Table 3.2: Top 20 common integration sites in ALV-J induced hemangiomas

Clusters were defined as any genomic locus containing at least two integrations within 5 kb of each other. By definition, clusters are flanked on either side by at least 5 kb of sequence lacking an integration. The number of integrations observed in hemangiomas and the average number of breakpoints are listed for each cluster.

<table>
<thead>
<tr>
<th>Gene Name</th>
<th># Common Integrations</th>
<th>Average # Breakpoints</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOC101750148</td>
<td>29</td>
<td>1.0</td>
</tr>
<tr>
<td>ELF1</td>
<td>29</td>
<td>1.3</td>
</tr>
<tr>
<td>LAPTM4A/WDR35</td>
<td>19</td>
<td>1.0</td>
</tr>
<tr>
<td>MAML2</td>
<td>17</td>
<td>1.1</td>
</tr>
<tr>
<td>SPRY1/ANKRD90</td>
<td>17</td>
<td>1.5</td>
</tr>
<tr>
<td>NFKBIA</td>
<td>17</td>
<td>1.1</td>
</tr>
<tr>
<td>MAP4K4</td>
<td>16</td>
<td>1.1</td>
</tr>
<tr>
<td>ADARB1/GLS1/STAT1</td>
<td>16</td>
<td>1.2</td>
</tr>
<tr>
<td>SSBP3</td>
<td>15</td>
<td>1.1</td>
</tr>
<tr>
<td>GSTA3/CK</td>
<td>15</td>
<td>1.1</td>
</tr>
<tr>
<td>ATAD2/FBXO32</td>
<td>15</td>
<td>1.2</td>
</tr>
<tr>
<td>ADD3</td>
<td>14</td>
<td>1.4</td>
</tr>
<tr>
<td>RUNX3</td>
<td>14</td>
<td>1.1</td>
</tr>
<tr>
<td>SPI1R</td>
<td>14</td>
<td>1.1</td>
</tr>
<tr>
<td>RREB1</td>
<td>14</td>
<td>1.1</td>
</tr>
<tr>
<td>SRGAP3</td>
<td>14</td>
<td>1.2</td>
</tr>
<tr>
<td>INPP5A</td>
<td>13</td>
<td>1.2</td>
</tr>
<tr>
<td>ANKRD9</td>
<td>13</td>
<td>1.1</td>
</tr>
<tr>
<td>TNFRSF21</td>
<td>13</td>
<td>1.0</td>
</tr>
<tr>
<td>GRHL1</td>
<td>13</td>
<td>1.0</td>
</tr>
<tr>
<td>MET</td>
<td>3</td>
<td>16.0</td>
</tr>
</tbody>
</table>
Some ALV-J integrations are present in expanded clones

In order to measure the relative abundance of each integration, we quantified the number of sonication breakpoints for each site. Previous studies have shown that sonication breakpoints can be used as a measure of clonal expansion; if an integration has more than one breakpoint, the cell carrying that integration has undergone clonal expansion (Berry et al. 2012).

88.5% of the integrations we identified exhibited only a single sonication breakpoint, suggesting that these integrations were not clonally expanded, while 11.5% had two or more breakpoints, evidence of clonal expansion. The 17 most highly expanded clones are shown in Figure 3.3A. Each of these integrations had 10 or more breakpoints. The tumor which we refer to as a “head hemangioma” is a hemangioma that developed subcutaneously outside the skull.

MET intron 1 is an expanded common integration site

Breakpoint analysis revealed three multi-breakpoint integrations within the first intron of the MET gene. These integrations occurred in a head (2 integrations) and leg hemangioma. Strikingly, these MET integrations represented the most highly expanded clones in both tumors (Figure 3.3B). To confirm the deep sequencing results, genomic DNA/proviral junctions from this locus were PCR-amplified and sequenced. This verified that these integrations were present only in the expected tumors, and in the orientation and location as predicted by deep sequencing (data not shown). Interestingly, the two MET integrations in the head hemangioma occurred in different orientations and were determined by deep sequencing (of the
3’ ends) to be offset by only six nucleotides. We were able to verify the 3’ end of each of the proviral junctions by PCR and sequencing, but were unable to amplify the 5’ junction in both cases. This suggests that these two proviruses exist adjacent to each other on the same strand and in opposite orientations. The six nucleotide offset observed by deep sequencing could be the result of the characteristic 6 nucleotide duplication that occurs upon ALV integration (Hughes et al. 1981).

Several lower breakpoint MET integrations were observed in infected non-tumor controls (Figure 3.4), but none were seen in myeloid tumors.

Of the three multi-breakpoint integrations identified in MET hemangiomas, all existed in a tight cluster within intron 1, a region spanning 1599 bp. This suggests that ALV-J integration within this specific region of MET gave these tumor cells a selective advantage and may have driven hemangiomagenesis in these birds. Both of these tumors had at least one highly abundant integration in the same orientation as the MET gene. Since the first exon of MET is non-coding, this integration pattern suggests that the virus may be inducing over-expression of full-length MET protein.

**MET mRNA expression is elevated in tumors containing MET proviral integrations**

Next we wanted to determine the effects of MET intron 1 integrations on gene expression in these tumors. ALV LTRs are known to harbor strong promoter and enhancer sequences that promote gene expression near viral integration sites (Beemon and Rosenberg 2012). MET mRNA expression was analyzed by qRT-PCR
(Figure 3.5). As expected, we found that MET mRNA expression correlated closely with the abundance of the MET integration sites in the sample as measured by breakpoint analysis. For example, the head hemangioma which has two integrations (21 and 23 breakpoints), exhibited the highest MET expression of any tumor. Expression was a full 131-fold higher than the kidney hemangioma which lacked any high abundance MET integrations. The same is true for hemangioma of the leg which also contains a MET expanded clone and shows higher levels of MET mRNA expression compared to those samples which lack MET integrations. This suggests that MET expression is induced by viral integration and is playing a causal role in hemangioma development.
Figure 3.5: MET mRNA expression measured by qRT-PCR

Hemangiomas (H) and myeloid tumors (M) as well as non-tumor and uninfected controls (labeled control) are shown. MET expression has been normalized to GAPDH and is relative to the hemangioma with the lowest MET expression, B3 Kidney H. Error bars represent standard error of the mean (SEM). Those tumors containing highly abundant MET integrations are noted (*). These tumors exhibit markedly increased levels of MET mRNA expression compared to other tumors and non-tumor controls.
Discussion

In this study we present the first integration profiles of ALV-J-induced hemangiomas. Hemangiomas are vascular tumors characterized by uncontrolled angiogenesis (Mabeta and Pepper 2011). They can occur in many species, including humans. In fact, infantile hemangioma is the most common tumor of early childhood, estimated to develop in 7-10% of infants (Tan et al. 2000). They present as benign tumors that grow postnatally for 8-12 months then typically undergo a slow process of self-involution that can last several years (Blei 2005; Takahashi et al. 1994).

Previous studies analyzing ALV-induced neoplasms mostly focused on ALV-A-induced B-cell lymphoma and identified MYC, MYB, MIR-155, and TERT as common integration sites in these tumors (Hayward, Neel, and Astrin 1981; G. S. Payne, Bishop, and Varmus 1982; Clurman and Hayward 1989; Yang et al. 2007). Only recently have ALV-J-induced neoplasms been the subject of similar work. A recent study of ALV-J-induced myeloid leukemia showed that MYC, TERT, and ZIC1 are targets of proviral integration in myeloid tumors (Y. Li et al. 2014). Interestingly, we did not identify these genes as CISs in this study. This may be due to the fact that only three myeloid tumors were analyzed. It is conceivable that with a larger cohort of tumors these CISs may have been observed as well.

In this study we implicate MET overexpression as a causal agent in the development of ALV-J-induced hemangiomas in chickens. Interestingly, although viral-induced tumors have been the subject of much study, to the best of our
knowledge no previous work has implicated MET in tumor induction by insertional mutagenesis. This suggests that MET may play a unique role in inducing hemangiomas rather than other types of viral-induced tumors.

MET has been studied extensively, and has important roles in both development and cancer. The MET protein is a receptor tyrosine kinase that binds hepatocyte growth factor/scatter factor (HGF/SF), and can activate an array of downstream signaling pathways including PI3K-AKT, RAC1-CDC42, RAP1 and RAS-MAPK (Gherardi et al. 2012). Previous work has shown that MET is activated in many types of human cancer via mutation, amplification, and protein overexpression, and MET activation correlates with poor prognosis in cancer patients (Abounader and Laterra 2005; Garcia et al. 2007; Peghini et al. 2002).

It has also been shown that MET plays an important role in angiogenesis, a process that is crucial to the development of hemangioma. For example, activation of the HGF/SF-MET pathway is now understood to be a potent inducer of angiogenesis, specifically in endothelial cells—the same type of cells that give rise to hemangiomas (Birchmeier et al. 2003; Abounader and Laterra 2005; Bussolino et al. 1992; Grant et al. 1993). In addition, the HGF/SF-MET pathway can suppress TSP1, a negative regulator of angiogenesis, and can induce expression of VEGFA, a pro-angiogenic gene. In this way the HGF/SF-MET pathway controls an “angiogenic switch”, turning on angiogenesis (Y.-W. Zhang et al. 2003). Additionally, it has been shown that activating mutations in MET can induce hemangiosarcomas in mice (C. Graveel et al. 2004). Viewed in this context it does not seem surprising that ALV-J-
induced $MET$ overexpression can lead to hemangiomagenesis in chicken. Whether or not $MET$ can act as a causal agent in human hemangioma has yet to be determined, but this may be an avenue for further research.

In addition to $MET$, other CISs were identified as hotspots of integration (Table 3.2). Notably, none of these integrations were seen in highly expanded clones and they were also observed in non-tumor tissue. This suggests that these CISs are unlikely to be relevant to tumor induction. In contrast, some tumors exhibited non-$MET$ integrations in highly expanded clones. For example $TRIO$ and $EYA4$ both contained integrations with many breakpoints. $TRIO$ has guanyl exchange factor (GEF) activity and regulates Rho family GTPases, which coordinate cytoskeletal rearrangement and cell migration (Bellanger et al. 1998; Blangy et al. 2000). $EYA4$ is a member of the eyes absent (EYA) family of proteins, it has phosphatase activity and may function in eye development as a transcriptional activator (Borsani et al. 1999). Our data suggest that these genes may also be capable of contributing to hemangioma formation, but because similar integrations were not observed in multiple tumors we cannot establish with certainty that they are drivers of oncogenesis and not merely passenger integrations. Further studies involving a larger cohort of birds may help identify other common integration sites and genes that drive ALV-J-induced hemangiomagenesis.
Chapter 4. Conclusion and future directions
Introduction

In this thesis, I describe my work to sequence and map integration sites in ALV-A induced B-cell lymphomas and ALV-J induced hemangiomas. In both studies (Chapters 2 & 3) I identified genes which are common sites of ALV integration. We believe these genes, or a subset of these genes, may be playing a causal role in tumor development. In addition, I showed that ALV-A integrations in these tumors are observed at a higher frequency near transcription start sites and within transcripts. I also show that ALV-A has a consensus site for integration in vivo. This work greatly expands our understanding of ALV induced tumors, and provides several avenues for future research which I will describe in this chapter.

Is the preference for integration near transcription start sites due to selection?

In chapter 2, I describe the integration profile of ALV-A induced B-cell lymphoma. We observed that promoters and TSSs are preferred sites of ALV integration in vivo. Oddly, similar studies conducted in cell culture have shown that the virus has a slight preference for integration within and near transcribed elements and CpG islands, but a strong preference for integration centered on transcription start sites has not been observed (Narezkina et al. 2004; Mitchell et al. 2004; Barr et al. 2005). As I describe in chapter 2, this inconsistency with earlier reports may be the result of selection occurring in vivo. I presented evidence in chapter 2 that clonally expanded cells (integrations in which two or more sonication breakpoints were observed) exhibited a higher frequency of integration near TSSs
than those which were not clonally expanded (Figure 2.10). This analysis suggests that indeed, the enhanced frequency of integration near TSSs is due at least in part to selection. Interestingly though, this analysis shows that even integrations that show no evidence of clonal expansion show enrichment for integration near TSSs. So the question remains, is the increased frequency of integration near TSSs due entirely to selection or is it at least partially due to an innate preference of ALV integrase? One possible way to answer this question would be to conduct a deep-sequencing experiment in cell culture. We can infect chicken embryo fibroblasts (CEFs) or similar cells with ALV and sequence integration sites just as we have done with tumors. By sequencing integrations immediately without allowing the cells to proliferate, we should be able to determine the selection-free integration site preference of ALV. In fact, these experiments are already in progress by other members of the lab and should help address this issue.

**Could host cell factors be directing ALV integration?**

In this study, I observed a preference for integration of ALV-A near TSSs. Interestingly, a preference for integration near TSSs is not unique to ALV. Another retrovirus, murine leukemia virus (MLV), has been the subject of similar studies. This work has shown a preference for MLV integration near TSSs as well as CpG islands (Cattoglio et al. 2010). MLV’s preference for integration near TSSs has been shown to be mediated by the binding of host cell factors, which influence integration site selection. Specifically, the binding of bromodomain and extraterminal domain (BET) proteins to the MLV integrase is thought to target MLV integration (Sharma et
al. 2013). As I describe in Chapter 2, the integration pattern we observe with ALV near TSSs is similar to MLV but not identical. Specifically, MLV prefers integration within a tight region of 2.5 kb flanking the TSS, whereas we see increased ALV integration frequency as far as 30 kb from the TSS. One characteristic that they both share is a sharp drop in integration frequency within 100 bp of the TSS (Cattoglio et al. 2010). It is also possible that ALV, like MLV, relies on a host cell factor to target integrations. This is something that we've also begun to study in the lab by conducting pull-down experiments with ALV integrase, and the early results of these experiments look promising.

**Does ALV prefer integration near CpG islands *in vivo***?

As was mentioned above, MLV prefers integration near CpG islands in addition to TSSs. Interestingly, a very slight preference for ALV integration has been observed near CpG islands as well (Mitchell et al. 2004). It may also be fruitful to conduct an analysis to determine whether ALV has similar preference for integration near CpG islands *in vivo*. I have not tested this hypothesis in this thesis. This analysis requires a comprehensive list of the locations of CpG islands in the chicken genome, which is not readily available. If we generate such a list, we could compare the integrations that we observe to the location of CpG islands. Such an analysis should be possible with the data set I've generated and may yield important information about ALV integration site selection.
Does ALV prefer integration near highly expressed genes *in vivo*?

HIV has been studied in a similar manner to the MLV and ALV experiments that I talk about above, but a much different pattern of integration was observed. Not only were integrations near TSSs shown to be disfavored, HIV has also been shown to prefer integration within genes that are highly expressed (Cattoglio et al. 2010). There is some indication that ALV has a similar preference for expressed genes. A slight but statistically significant preference for integration was observed near transcribed genes in gene-dense regions in cell culture (Mitchell et al. 2004). We could attempt to confirm this hypothesis for ALV integration *in vivo* by performing RNAseq on chicken tumors. We could compare gene expression in each tumor to the locations of the integrations in that tumor, and determine if integration favors highly expressed genes. One drawback of such an experiment is that we may observe correlation between highly expressed genes and integrations because ALV itself is overexpressing nearby genes. To get around this issue we could instead conduct the experiment in cell culture, comparing the locations of ALV integrations in infected cells to the expression of genes in uninfected cells of the same type. This would ensure that the correlation between expression and integration exists independent of ALV-induced expression of nearby genes.

Does nuclear architecture affect ALV integration?

Another possible route of investigation is to determine the effect of nuclear architecture on ALV integration site preference. Recent work has shown that nuclear architecture is a strong determinant of HIV integration. Genes targeted by
HIV integration are often situated at the periphery of the nuclear envelope and are often associated with the nuclear pore and contain active transcription chromatin marks before viral integration (Marini et al. 2015). Additionally, HIV disfavors integration in nuclear lamin-associated domains and even disfavors transcriptionally active loci in regions not at the nuclear periphery (Guelen et al. 2008). It’s possible that the location of ALV integration is influenced by the location of the gene in the nucleus in a similar fashion and is an interesting avenue for future research.

What is crosstalk and how can we work to eliminate it?

In the first steps of preparing a library for high-throughput sequencing, a DNA adapter is typically ligated onto sample DNA. This adapter contains sequences which anchor the fragment to the flow-cell of the sequencing machine. The adapter also often contains an index (a.k.a. barcode) sequence. This index sequence allows multiple samples to be sequenced in the same sequencing run. The barcode is sequenced along with the rest of the read and is used to determine which sample each read belongs to at the conclusion of the run.

Sample mis-identification (calling the wrong barcode for a given read) is a known issue with high-throughput sequencing. There are several causes of sample mis-identification, ranging from mis-sequencing of the barcodes to the cross-contamination of indexed oligos by sequential purification on the same HPLC columns shortly after synthesis.
I was able to identify sample mis-identification in my experiments, which I called “crosstalk”. Crosstalk was most noticeable in my data in cases of a very highly abundant integration (many identical reads) in a given sample. In these scenarios, it was often the case that the abundant integration would appear in different samples at lower frequency (usually <1% of the number of reads) than the “true” sample in which the integration belonged. In cases where the same integration was observed in two or more samples, I was able to remedy the issue within my bioinformatic pipeline. I did this by comparing the integration among all samples and selectively keeping the integration only in the sample in which it exhibited the highest number of sonication breakpoints. All other instances of the integration were thrown out.

This issue has been described in-depth in a recent publication (Kircher, Sawyer, and Meyer 2012), which suggests a dual-indexing strategy to aid in correctly identifying which sample each read belongs to. I adapted this approach to my protocol (Figure 4.1). In this new library preparation procedure each sample is assigned two barcode sequences, which are embedded in the adapters on either end of the product to be sequenced. These new adapters are being used in all further high-throughput integration studies in the lab and should largely ablate the crosstalk issue that I observed.
Adapters:
as seen on the LEFT
5' CAAGCAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAGA
Do the genes identified as common integration sites drive tumorigenesis?

Lastly, in this thesis I identify a number of genes as common integration sites in ALV-A induced B-cell lymphoma and ALV-J induced hemangioma. As one might expect, many of these genes have well-characterized roles in cancer and related processes. These genes include RUNX1, Mir-221, Mir-222, IKZF1, CCNA2, ZEB1, CBLB, HMGB1, and MET. In addition to these canonical cancer genes, I identified a number of genes as common integration sites that are conserved in human but have never been linked to cancer. These include CXorf57, CTDSPL2, TMEM135, ZCCHC10, FAM49B, and MGARP. In fact, three of these six genes, CXorf57, ZCCHC10, and FAM49B, not only have never been linked to cancer, but have never undergone any characterization of any kind and have no known functions. We think these genes are interesting targets for further research. Specifically, we would be interested to test whether any of these genes are perturbed in human cancer and if so, how they may be contributing to carcinogenesis.
References


Balkwill, Frances. 2006. “TNF-Alpha in Promotion and Progression of Cancer.”


Interactions.” *Current Topics in Microbiology and Immunology* 281: 107–36.

Immunodeficiency Virus in the Chicken Genome.” *Journal of Virology* 79 (18):

and Murine Retroviruses.” In *Cancer Associated Viruses*, edited by Erle S.


Bellanger, Jean-Michel, Jean-Bernard Lazaro, Sylvie Diriong, Anne Fernandez, Ned
Factor Domains of Trio Link the Rac1 and the RhoA Pathways in Vivo.”
*Oncogene* 16 (2): 147–52. doi:10.1038/sj.onc.1201532.


Berry, Charles C., Nicolas A. Gillet, Anat Melamed, Niall Gormley, Charles R. M.


Donzé, Olivier, Pascal Damay, and Pierre-Francois Spahr. 1995. "The First and Third uORFs in RSV Leader RNA Are Efficiently Translated: Implications for


Goecks, Jeremy, Anton Nekrutenko, James Taylor, and Galaxy Team. 2010. “Galaxy: A Comprehensive Approach for Supporting Accessible, Reproducible, and


Ikeda, Terumasa, Junji Shibata, Kazuhisa Yoshimura, Atsushi Koito, and Shuzo Matsushita. 2007. “Recurrent HIV-1 Integration at the BACH2 Locus in Resting CD4+ T Cell Populations during Effective Highly Active Antiretroviral


LaFave, Matthew C., Gaurav K. Varshney, Derek E. Gildea, Tyra G. Wolfsberg, Andreas D. Baxevanis, and Shawn M. Burgess. 2014. “MLV Integration Site


Li, Yuhao, Xuemei Liu, Zhen Yang, Chenggang Xu, Di Liu, Jianru Qin, Manman Dai, et al. 2014. “The MYC, TERT, and ZIC1 Genes Are Common Targets of Viral...


Growth Factor and Hepatocyte Growth Factor Receptors in a Proportion of Gastrinomas Correlates with Aggressive Growth and Lower Curability."


Curriculum Vitae
James Franklin Justice IV

Johns Hopkins University
Department of Biology
3400 N. Charles St., Mudd Hall
Baltimore MD, 21218

Telephone (lab): 410-516-6571
Telephone (mobile): 216-375-2251
Fax (lab): 410-516-7292
E-mail: jjustic3@jhu.edu

EDUCATION

2009 - present
Ph.D. Cellular, Molecular, Developmental Biology and Biophysics
The Johns Hopkins University, Baltimore, MD
Graduation date: October 2015
Supervisor: Dr. Karen Beemon

2005 – 2008
Bachelor of Science, Molecular Genetics cum laude
The Ohio State University, Columbus, OH
Graduation date: December 2008
Research Supervisor: Dr. Gustavo Leone

RESEARCH EXPERIENCE

2009 – present
Identified the MET gene as common integration target in avian leukosis virus subgroup J-induced chicken hemangiomas
Characterization of avian leukosis virus subgroup-A induced B-cell lymphomas identifies a preference for integration near transcription start sites and implicates new genes in the development of ALV induced B-cell lymphoma

2006 – 2008
Undergraduate Research, The Ohio State University
Characterization of hepatocellular carcinoma in an E2F transgenic mouse model

PRESENTATIONS

Sept. 2015
JHU-MedImmune Science Day
Presented a poster entitled: High-throughput proviral integration site sequencing reveals genes driving avian leukosis virus induced neoplasms

May 2014
Carnegie Bioinformatics: Nitty Gritty Workflows,
Carnegie Institution of Washington, Department of Embryology
Baltimore, Maryland
Oral presentation entitled: Perl workflow to analyze viral integrations in avian leukosis virus induced B-cell lymphomas and hemangiomas

April 2014 17th Annual HIV Drug Resistance Program Think Tank Meeting, Frederick, Maryland
Oral presentation entitled: High-throughput sequencing of integration sites to identify genes involved in ALV-induced lymphomas

August 2013 25th Annual Workshop on Retroviral Pathogenesis, Reykjavik, Iceland
Oral presentation entitled: High-throughput sequencing of integration sites to identify genes involved in ALV-induced lymphomas

May 2012 Cold Spring Harbor Laboratory Meeting: Retroviruses, Cold Spring Harbor, New York
Presented a poster entitled: High-throughput sequencing of integration sites to identify genes involved in ALV-induced tumors

October 2011 13th Annual Department of Biology Retreat, St. Michaels, Maryland
Presented a poster entitled: High-throughput sequencing of avian leukosis virus integration sites to identify genes involved in B-cell lymphoma

August 2011 Workshop on Genomics and Proteomics of Oncogenic Viruses, Syria, Virginia
Presented a poster entitled: High-throughput sequencing of avian leukosis virus integration sites to identify genes involved in B-cell lymphoma

PUBLICATIONS


**RESEARCH INTERESTS**

Cancer  
Evolution  
Gene Therapy  
Genome Editing  
Genomics  
High-throughput sequencing technologies

**TEACHING EXPERIENCE**

- 2012  Molecular Biology Teaching Assistant  Johns Hopkins  
- 2011  Cell Biology Laboratory Assistant  Johns Hopkins  
- 2010  Genetics Laboratory Assistant  Johns Hopkins

**REFERENCES**

Dr. Karen Beemon  klb@ju.edu  410-516-6571  PhD. Supervisor  
Dr. David Zapulla  zappulla@jhu.edu  410-516-8749  Thesis Committee  
Dr. James Taylor  jxtx@jhu.edu  410-516-0152  Thesis Committee  
Dr. Haig Kazazian  hkazazi1@jhmi.edu  215-898-3582  Thesis Committee