# A GENOMIC INVESTIGATION OF MAREK'S DISEASE LYMPHOMAS

By

Alexander Cordiner Steep

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Genetics—Doctor of Philosophy

2019

**ABSTRACT**

A GENOMIC INVESTIGATION OF MAREK'S DISEASE LYMPHOMAS

By

Alexander Cordiner Steep

Meq, a bZIP transcription factor and the viral oncogene for pathogenic strains of Marek's disease virus (MDV), is required to induce CD4 T cell lymphomas that characterize Marek's disease (MD) in chickens. However, Meq is not sufficient for neoplastic transformation as not all birds infected with pathogenic strains of MDV developed Marek's disease. We hypothesize that additional drivers or somatic mutations in the chicken genome are required for MDV-induced transformation. Using and integrating DNA and RNA genomic screens of Marek's disease tumors from genetically-defined experimental layers, our analyses reveal 0.3 somatic mutations per megabase consisting primarily of somatic single nucleotide variants (SNVs) and small insertions and deletions (Indels). Somatic deletions, insertions, and point mutations were enriched in *IKZF1* (Ikaros), the first driver gene of Marek's disease lymphomas. Ikaros, a Zn-finger transcription factor and the master regulator of lymphocyte development, is a known tumor suppressor in human and murine acute leukemias and lymphomas. In our surveyed Marek's disease tumors, 41% of the samples had somatic mutations in key N-terminal Zn-finger binding domains, strongly suggesting perturbed Ikaros function in its ability to bind DNA and regulate transcription. Somatic mutations in *IKZF1* were preferentially found in tumors of gonadal tissues as well as their metastatic clones. *IKZF1* mutant Marek's disease tumors revealed gene expression profiles indicative of Ikaros perturbation. In addition to *IKZF1*, other putative somatic mutations reside in *ZNF384, EFNA5, CLDND1*, *FOXD1, ROBO1, and ROBO2* and

warrant evaluation. Our results suggest MDV-induced tumors are driven by both Meq expression and *IKZF1* somatic mutations that in combination lead to unregulated proliferation, increased cell adhesion, increased migration, and dedifferentiation.

**ACKNOWLEDGEMENTS**

Hans Cheng took a chance when accepted my application into the Avian Disease and Oncology Lab. He knew I was unqualified for such an ambitious project. Instead, he invested in me. He provided a valuable opportunity and pushed me to fulfill it. He developed my thought process, promoting skepticism while protecting creativity. I have him to thank for an education both domestic and abroad. It has been a privilege to work under his supervision.

I would like to thank Prof. Dmitrij Frishman, who hosted me at the Technical University of Munich. Dr. Frishman provided a collaborative environment rich in inspiration and supported with the infrastructure to bring those ideas to fruition. In Weihenstephan, I collaborated most closely with Dr. Hongen Xu. Hongen introduced me to big data coding and management. His efforts propelled our work forward. I give a special thanks to Drazen Jalsovec for his camaraderie and IT management. And to the rest of the Frishman lab—I will miss our interactions, the lunches, the coffee, and the Helles. Linus Zimmermann and Maureen Wunderlich introduced me to Bavaria shared their kindness and hospitality.

Many other colleagues and collaborators helped to complete this research. Dr. Cari Hearn, Laurie Molitor, Dr. Andrey Grigoriev, Dr. John Dunn, Dr. Henry Hunt, Lonnie Milam, Sue Umthong, Dr. Bill Muir, and Dr. Mary Delany made substantial contributions to this work.

I would like to thank my committee: the late Dr. Michele Fluck, Dr. Ana Vazquez, Dr. Jerry Dodgson, and Dr. Eran Andrechek for their guidance. A special thanks to Dr. Andrechek for engaging the cancer community at MSU, an effort I hope he continues. The Institute for Cyber-Enable Research (iCER) was my “computer school.” I especially thank Dr. Matt Scholz and Pat Bills for teaching me to code and manage big data. Dr. Titus Brown taught me about

reproducible bioinformatics and the importance of documentation. The MSU RTSF Genomics Core helped design and implement sequencing studies, especially Dr. Kevin Childs and Kevin Carr.

Dr. Barbara Sears, Dr. Vilma Yuzbasiyan-Gurkan, and Dr. James W. Lloyd collectively inspired my candidacy for graduate school, a deed I will always be thankful for. Dr. Cathy Ernst leads the Genetics program at MSU and has created a rich and engaging environment that has benefitted us all.

David Glenn was the best teacher I ever had; he fostered an interest in science at the right time in life. My sister, Stacey, has always supported my ambitions and led by example. My mother and father have always encouraged me to pursue my interests without compromise. They have supported me in every way.

**TABLE OF CONTENTS**

**LIST OF TABLES**

**LIST OF FIGURES**

## KEY TO ABBREVIATIONS

| | |
|---|---|
| ALL | Acute lymphoblastic leukemia |
| AML | Acute myeloid leukemia |
| AP-1 | Activator protein 1 |
| APOBEC | Apolipoprotein B mRNA Editing Catalytic Polypeptide |
| AID | Activation Induced Deaminase |
| BAF | B allele frequency |
| B-ALL | B-cell acute lymphoblastic leukemia |
| BCR-ABL1 | Breakpoint cluster region-Abelson proto-oncogene fusion gene |
| bp | Base pairs |
| bZIP | Basic leucine zipper protein |
| CD133 | Cluster of differentiation 133 |
| CD3 | Cluster of differentiation 3 |
| CD4 | Cluster of differentiation 4 |
| CD8 | Cluster of differentiation 8 |
| CDK2 | Cyclin-dependent kinase 2 |
| CDK6 | Cyclin-dependent kinase 6 |
| Chip-Seq | Chromatin immunoprecipitation followed by sequencing |
| cIAP-1 | Cellular inhibitor of apoptosis protein-1 |
| CML | Chronic myelogenous leukemia |
| CNA | Copy number alteration |
| CNV | Copy number variant |

| | |
|---|---|
| COSMIC | Catalogue of Somatic Mutations in Cancer |
| CRE | Cyclic AMP response element |
| DBD | DNA-binding domain |
| DE | Differentially expressed |
| DLBCL | Diffuse large B-cell lymphoma |
| dpi | Days post infection |
| EBV | Epstein-Barr virus |
| Egr1 | Early growth response protein 1 |
| ErbB | Erythroblastic leukemia viral oncogene homolog (aka EGFR) |
| ERK | ELK-related tyrosine kinase (aka EPHB2) |
| FDR | False discovery rate |
| FPR | False positive rate |
| GATK | Genome Analysis Toolkit |
| GFP | Green fluorescent protein |
| GSEA | Gene set enrichment analysis |
| HDAC | Histone deacetylase |
| HHV6 | Human herpesvirus-6 |
| IGV | Integrative Genomics Viewer |
| IKDN | Ikaros dominant-negative |
| IKZF1 | Ikaros family zinc finger 1 |
| Indels | Insertions and deletions |
| JAK-STAT | Janus kinases and Signal Transducer and Activator of Transcription |

| | |
|---|---|
| LFC | $Log_2$ fold change |
| LOH | Loss of heterozygosity |
| LRR | Log R ratio |
| LSC | Leukemic stem cell |
| MACS | Magnetic-activated cell sorting |
| MAPK | Mitogen-activated protein kinase |
| MB | Megabase |
| MD | Marek's disease |
| MDV | Marek's disease virus |
| MEK1/2 | MAPK/ERK kinase 1 and 2 (aka MAP2K1 and MAP2K2) |
| MERE | Meq responsive elements |
| mRNA | Messenger RNA |
| MSU | Michigan State University |
| NCBI | National Center for Biotechnology Information |
| NGS | Next-generation sequencing |
| NURD | The nucleosome remodeling and deacetylase complex |
| PCA | Principle component analysis |
| PFB | Population B allele frequency |
| pfu | Plaque-forming units |
| PIAS | Protein inhibitor of activated STAT |
| PP1 | Protein phosphatase 1 |
| PP2 | Protein phosphatase 2 |

| | |
|---|---|
| DUSP4 | Dual specificity phosphatase 4 |
| RAG | Recombination activating gene |
| rlog | Regularized-log |
| RNA-seq | RNA sequencing |
| ROBO1 | Roundabout homolog 1 |
| ROBO2 | Roundabout homolog 2 |
| rRNA | ribosomal RNA |
| RTSF | Research technology support facility |
| SHP-1 | Src homology region 2 domain-containing phosphatase 1 |
| SNP | Single nucleotide polymorphism |
| SNV | Single nucleotide variant |
| SOCS2 | Suppressor of cytokine signaling 2 |
| Src | Proto-oncogene tyrosine-protein kinase |
| STAT3 | Signal transducer and activator of transcription 3 |
| SV | Structural variant |
| SVA | Surrogate variable analysis |
| SWI-SNF | Switch/sucrose non-fermentable nucleosome remodeling complex |
| T-ALL | T cell acute lymphoblastic leukemia |
| TCGA | The Cancer Genome Atlas |
| TCR | T cell receptor |
| TSG | Tumor suppressor gene |
| VAF | Variant allele frequency |

| | |
|---|---|
| VDJ | Variable, diversity and joining recombination |
| VEP | Variant Effect Predictor |
| VST | Variance stabilizing transformation |
| WGS | Whole genome sequencing |

# CHAPTER 1

# Introduction

Marek's disease (MD) is a lymphoproliferative disease in chickens caused by Marek's disease virus (MDV), a highly oncogenic alphaherpesvirus. Marek's disease is characterized by a chronic peripheral neuropathy, in which demyelination of nerve sheaths and paralysis occurs. Often these observations are associated with the most serious pathology that is characteristic of the disease, an acute lymphoma in visceral organs[1]. Marek's disease was first described by Jószef Marek in 1907 as polyneuritis, however, the disease has become increasingly neoplastic since the induction of high-density chicken rearing practices introduced in the 1960s.[2] Vaccine regimens have been proposed to promote MDV to become more virulent[3]. Although highly successful, the current control strategy for Marek's disease is not sustainable as vaccination does not prevent horizontal transmission of more virulent viral genotypes.[4] In high-density chicken rearing environments, MDV is ubiquitous. In such situations, chickens are immediately exposed after hatch, suggesting that unless vaccinated, the first pathogen to interact with their immune system is MDV. These series of events have generated one of the most rapid transformation processes known, which is estimated to occur between 2-4 weeks within the cell. Elucidating the ordered series of events surrounding Marek's disease transformation influence lymphomagenesis will be instrumental in controlling one of the world's most prevalent and acute cancers.

**Marek's disease lymphomas**

MDV is a highly contagious, oncogenic alpha-herpesvirus that elicits inflammation, immunosuppression, polyneuritis, and acute T cell lymphomas in infected chickens. The current model of pathogenesis and subsequent transformation begins in the respiratory tract, where

inhaled virus is engulfed by phagocytic cells, including macrophages, and is transported to the major lymphoid organs, such as the spleen, bursa, and thymus within 24 hours. In these visceral organs where lymphocyte populations are highly proliferative, MDV targets B- and activated CD4 T-lymphocytes for cytolytic infection. The infection peaks 3 to 6 days post infection (dpi) often causing atrophy to the bursa and thymus.

The CD4 T cell population is the primary target for viral latency and subsequent transformation. Latency is a hallmark of herpesviruses and this dormant period allows the virus to evade the immune system. MDV transitions to latency with some overlap between the reduction of cytolytic replication and the induction of latency, usually within 7 dpi. For example, latency has been observed as early as 24 hours post infection and peaked at 4 dpi under experimental conditions.[5] Transformation with gross lesions occurs as early as two to four weeks post infection.[6] The transition process that occurs in CD4 T lymphocytes is not well understood. However, the CD4 targets of MDV replication have been described as highly proliferative from the onset of latency to their transformation.[7] This environment, rich in highly proliferative lymphocytes and oncogenic virus, is ideal for the transformation process because somatic mutations accumulate with each cycle of cell division, especially under inflammatory conditions.

Cancer is a disease of the genome that nearly always results from somatic mutations or sometimes with the aid of viral infection. The key factors involved in Marek's disease transformation include: the pathogen MDV, its CD4 host cell genome, and the timing and context of these interactions. Past research has examined the MDV genome, including the necessary integration and expression of viral constituents These efforts have shed light on two

events known to be necessary for CD4 transformation:[8] integration of MDV into the chicken genome[5,9] and the activity of the primary viral oncogene, Meq.[10]

**Marek's disease virus integration drives Marek's disease lymphomas**

MDV enters latency, including integration, as a natural part of its herpesvirus life cycle preceding transformation. In 1993, Delecluse and Hammerschmidt demonstrated the first evidence that MDV integrates into the host genome[11] and subsequent efforts confirmed and further explained the mechanics of this process.[5,9,12] Similar events have also been observed in herpesvirus-associated malignancies in human; both Epstein-Barr Virus (EBV) and Human Herpesvirus-6 (HHV6) are capable of integration into their respective host genomes prior to transformation.[13] Integration is a necessary step that must occur prior to transformation[5,9] and studies have suggested that tumors are monoclonal in vivo[5,9] and in cell lines.[11] But these studies focused on either tumors from middle and late stages of Marek's disease or in vitro cell cultures. One particularly interesting experiment[5] examines the temporal series of events surrounding integration as the early stages and initiation of Marek's disease pathogenesis.

Integration events are key markers for Marek's disease lymphomagenesis and progression. They illustrate the microevolution of Marek's disease tumors from a heterogeneous mix of pre-transformed and transformed cells, to gross tumors with predominantly homogenous CD4 populations of transformed cells. Subsets of predominantly B and T cells from spleen, bursa, and thymus from birds inoculated at hatch show both cytolytic growth and integration in tandem.[5] These groups of cells demonstrate both integrated viral genomes and associated virus as early as 1 dpi; however, the transitions to cells demonstrating

only integrated virus—MDV-integrated-only cells demonstrating latency—occurs sometime between 2 and 3 weeks in these major lymphoid organs. This process is most pronounced in the spleen.[5] Robinson et al. observed that a rapid increase of MDV-integrated-only cells in the spleen coincided with tumor incidence. Although both B and T lymphocytes show MDV-integration in the major lymphoid organs, spleen showed the highest concentration of cells with integration-only profiles.[5] These observations are supported in part by Marek's disease studies revolving around the spleen. In addition to harboring MDV-integrated cells and gross tumors, the spleen is an important site in Marek's disease pathogenesis because infection leads to splenomegaly.[14] Collectively, these observations suggest that spleens may play a role in and harbor the transformation process. The spleen is not required for transformation, however, as birds having undergone splenectomy and subsequent infection with MDV still demonstrated gross tumors.[15]

Robinson et al.'s study of MDV-integrated lymphocytes in spleens indicated a pattern for the transformation process. A heterogeneous population of transformed cells and tumors characterized by their diverse integration signatures would eventually give way to a predominant, likely a clonal, population of transformed cells—based on identical integration profiles— that also constituted late-stage tumors.[5] Consistent with their prior observations,[9] late-stage tumors demonstrated uniformity in their integration profiles. Collectively, these observations suggest a series of events in which a diverse group of T and possibly B cells undergo the genetic and cellular reprogramming necessary for transformation. More than one cell is likely transformed and these cells likely emigrate to or from lymphoid organs, but especially to the spleen. Time series analysis of MDV integration profiles suggests that the

process that preferentially selects for a predominant—usually clonal—transformed cell type typically occurs within the spleen.[5,9]

**Meq drives Marek's disease lymphomas**

Although integration is a hallmark of Marek's disease lymphomas, it alone is not sufficient to transform the cell. Meq is the primary viral oncogene and required for transformation.[10,16] The instrumental mechanism of Meq was revealed when Lupiani et al. showed that a Meq-deleted Md5 (rMd5ΔMeq) recombinant virus failed to induce transformation in vivo, resulting in an attenuated virus that could still replicate.[10] Indeed, both integration and Meq activity are required for transformation.[8] Robinson et al. used a similar approach and demonstrated that cells with BACdelMEQ[17] demonstrated viral replication and even integration but did not enter latency or subsequent transformation, suggesting that Meq homodimers are not required for integration but are necessary for, or are thought to promote, the establishment of latency.[5] Although other viral proteins and products have been examined for their oncogenicity, such as pp38, Meq is widely accepted as the primary oncogene associated with Marek's disease lymphomagenesis.

Meq is the most well studied MDV protein. It has been shown to be expressed throughout the lytic and latent phases of the MDV life cycle[18] and during transformation.[7,10] Its functions are diverse. It is a transactivator and chromatin remodeler that can bind to itself, viral and cellular proteins, and DNA.[6] The influential function of Meq seems to be delivered through transcriptional regulation.[19] Both the chicken and integrated viral genomes are subject to regulation of gene expression.[1] Meq is able to bind to transcription factors associated with cell

cycle control, such as Rb, p53, and cyclin-dependent kinase 2 (CDK2).[18] As Meq is a bZIP protein it is thought to bind and dimerize most commonly and strongly to proteins within the bZIP family, such C-Jun, JunB, Fos, and itself.[20] Of all heterodimers, the interaction between Meq and Jun is considered both the most biochemically stable[20] and relevant to positive regulation of transcription.[19]

Although heterodimers of Meq and Jun have been shown to up-regulate genes, Meq homodimers have been shown to act on distinctly different sites to down-regulate genes and both of these mechanisms have been extensively characterized.[19] Meq is a multifunctional transcription factor that regulates many pathways involved in cellular proliferation, survival, and migration. Dimers bind to promoters containing MERE sites (Meq responsive elements that harbor CRE/TRE cores). Homodimers bind preferentially to MERE-II sites to repress transcription,[19,20] and heterodimers containing Jun bind MERE-I AP-1 like sites to enhance expression.[20,21] Both homo- and hetero-dimerization of Meq are necessary for transformation;[22] and their effects on gene expression and subsequent pathway enrichment have been examined.[19] The five most enriched pathways associated with Meq regulation involve the ERK/MAPK signaling, ErbB signaling, apoptotic and death receptor signaling, and JAK-STAT signaling.[19] In nearly all major pathways that are up-regulated by Meq dimers, in general, Meq homodimers down-regulate the cellular proteins tightly regulating pathways—often tumor suppressors—and Meq heterodimers up-regulate cellular proteins—often oncogenes—that activate their pathway.

An example of this is demonstrated in the ErbB/MAPK pathway in which heterodimers up-regulate kinases, such as ErbB and Src, and down-regulate phosphatases—such as PP1, PP2,

and DUSP4—known to target the MAPK signaling pathways. Furthermore, Meq up-regulates Ras and MEK1/2 activities. Subramaniam et al. used specific inhibitors of ErbB1/2 and MEK1/2 in cell proliferation assays demonstrating reduced proliferative effect of Meq and further supporting the role of Meq on MAPK activation. Meq was also suggested to positively influence proliferation via activation of the JAK-STAT pathway by down-regulating tumor suppressors—SHP-1, SOCS2, and PIAS, and up-regulating the oncogene, STAT3.[19] Meq aids to prevent cell death by up-regulation of the anti-apoptotic genes, Bcl-2, Bcl-XL,[26–28] and cIAP-1. In addition, Meq down-regulates pro-apoptotic proteins Bid, Caspase 3 and Caspase 6.[19] Collectively, Meq dimers work in tandem to utilize opposing functions of activation or repression of gene expression orchestrated on the same target pathways to achieve increased cell proliferation and resistance to apoptosis.

The expression of Meq in vivo has been assumed to occur in lytic and latent phases and in transformed cells, based on experiments in vitro.[23,24] However, a recent in vivo analysis of the temporal expression of Meq fused with Green Fluorescent Protein (GFP) challenges past assumptions. In vivo expression of Meq was not detected during the presumed peak of lytic growth. Instead, the first evidence of cells expressing Meq was in CD4 cells (CD4+Meq+) occurring at 7 dpi.[7] These observations are in tune with the onset of latency and the current understanding that Meq is the only protein consistently translated during latency.[16] A stark increase in CD4+Meq+ cells occurred from 7 to 14 dpi, indicating a rapid proliferative process. This observation is of interest because of its temporal coincidence with the expression of Meq. The authors could not deduce whether these cells were transformed or activated;[7] however, it

is possible that Meq could play a role in triggering a rapid proliferation response, especially considering its ability to activate pathways associated with proliferation.

Presumably, all transformed cells compete to emerge from the transformation process by expressing Meq and harnessing its ability to reprogram the cell. What gives the final monoclonal cell population strategic advantage during this process is unknown. Whether those advantages were somatically inherited prior to infection or acquired—perhaps in the setting of highly proliferative T and B lymphocytes infected with MDV and triggered by Meq—is of great interest. Regardless of this uncertainty, the most likely place for such a third party to reside is within the chicken genome in the form of somatically acquired mutations.

**Characteristics of cancer genomes**

Cancer is a genetic disease that arises from mutations in the genomes of cancer cells, most often as somatic mutations. Somatic mutations arise through cellular divisions, which may occur normally or are induced. In cancer cells, these mutations often represent the accumulation of mutational processes that have been working through every division of the cell—from the fertilized egg to the cancer cell.[25–27] Tumors often arise from a clone of cells in an unregulated manner, demonstrating an "uncontrolled growth" caused by somatically acquired mutations. Additional forces may work in tandem with somatically acquired mutations to influence the growth of the cell, including integration or expression of oncogenic proteins from viruses[28] as well as epigenetic changes.[29] Somatic mutations occur most commonly as single nucleotide variants (SNVs) representing a single nucleotide substitution, or less frequently as small insertions or deletions generally fewer than 50 base pairs in length (Indels), large Indels

often accompanied by changes in copy number of DNA segments, and rearrangements in DNA that sometimes express themselves as gene fusion events.

The acquisition of somatic mutations is a normal process that occurs in all cells during cellular division. This process occurs in ovo or in utero during development and is continued in postnatal life as self-renewing cells replenish themselves. Acquisition may be influenced by inherited "germline" mutations in the fertilized egg—thus in all somatic cells of the individual—that increase cellular susceptibility to mutation acquisition. Collectively, endogenous and exogenous forces act through cell division to drive the acquisition of somatic mutations. Endogenous mechanisms, including methylation-mediated spontaneous deamination of 5-methylcytosines, drive the acquisitions of C>T and G>A SNVs in CpG dinucleotides and CpNpG trinucleotides.[30] Exogenous influence of mutation acquisition may directly result from chemical influence—carcinogens from cigarette smoke[31]—or biological influence—virally induced inflammation and subsequent APOBEC/AID driven cytidine deaminase.[32] Germline mutations can further enhance susceptibly to somatic mutation acquisition by influencing endogenous somatic mutations rates, the severity of the inflammatory process, normal or cancer cell growth, or the metabolism of carcinogens.

A small subset of somatic mutations in cancer clones is referred to as "driver mutations." "Drivers" often reside in cancer genes and impart a distinct growth advantage to the cancer cell by reprogramming the cell to deregulate normal processes like cellular proliferation, differentiation, and cell death. These normal functions become deregulated and the homeostasis between the cancerous cell and its microenvironment shifts out of balance in favor of uncontrolled growth. The driver mutations that provide growth advantage allow the

cancer clone to outgrow all other cells from the same tissue, invade other surrounding tissues, or even metastasize.[26,33] Cancer cells are suggested to have between 1-10+ driver mutations depending on cancer type.[34] The remaining and vast majority of somatic mutations are classified as "passenger mutations." By definition, passenger mutations do not convey growth advantages to the cell. The number of passenger mutations reflects the number of mitotic cellular divisions between the fertilized egg and the cancer cell and the mutation rate at each division. This phenomenon can be seen in the disparity in somatic mutation quantity between two morphologically identical colorectal tumors from individuals generations apart.[33] Passenger mutations, however, provide useful clues into the mutational process; their somatic mutation signature and frequency help to elucidate mutational processes and pinpoint clusters of mutations in "hotspots" often associated with cancer drivers.

**Somatic mutations in cancer genomes**

Tumors from different cell and tissue types demonstrate stark differences in mutational frequency. Cancers may range in their mutational load by several orders of magnitude.[33,35] Typically, there are somewhere between 1,000 and 10,000 somatic substitutions in the genomes of most adult cancers[26] and the number of mutations in self-renewing tissue is often directly correlated with age.[36] Tumors in individuals in different generations and with distinct mutational mechanisms occur at each end of the spectrum of mutational frequency. For instance, tumors with high mutational loads such as lung, melanoma, stomach, colorectal, endometrial, and cervical cancers demonstrate "hypermutation."[37] These tumors may demonstrate scores of thousands of somatic mutations, which reflect their mutation processes.

Tumors harboring high mutation frequencies may result from carcinogenic processes such as exposure to mutagens[31] and defects in DNA-repair.[38] For instance, lung cancer tumors from smokers have nearly 10-fold more somatic mutations than tumors nonsmokers.[31] Tumors on the other end of the spectrum such as pediatric solid tumors arising from non-self renewing tissues and leukemias have a much lower mutational load.[33] For instance, glial cells of the brain are non-self renewing and normally do not demonstrate high somatic mutation frequencies in glioblastoma. The influence of exogenous factors however may also occur in non-self renewing tissues of young individuals. This is revealed in patients with glioblastoma treated with temozolomide or lomustine, which demonstrate significantly increased mutation rates and subsequent tumor mutation load.[39]

**Cancer genes and drivers**

Prior research supports the concept that that driver mutations grant a selective growth advantage to a cancer cells but passenger mutations do not.[40] Although driver mutations occur in driver genes, there is an important distinction between the two. A driver gene is a cancer gene, which contains function-altering driver mutations. However a driver gene may also contain passenger mutations. Therefore, not all mutations in driver genes act as cancer drivers.[33] In large cohorts of individuals with identical tumor types, the patterns associated with the mutations in cancer genes can aid in their discovery.

Driver mutations within driver genes may be recognized from distinct patterns, including, but not limited to: their mutation types, whether the mutations cluster in a hotspot of significantly enriched mutational frequency, the pattern of mutational clusters in context the

architecture of the gene, and in some cases the mutually exclusive nature of drivers in the same pathways in large sample cohorts. The majority of cancer drivers in common solid tumors of humans demonstrate about 95% of mutations as somatic SNVs. The remaining cancer drivers are usually small Indels, occurring within the gene in a non-synonymous fashion.[33] The distribution of these variant types and others is also of great importance as significantly mutated genes speckle the cancer genome and label drivers for discovery. In cancer genome landscapes, if the most mutated driver genes are the "mountains," then the remaining less, yet significantly, mutated genes are referred to as "hills." This common analogy explains the universal mutation patterns in cancers across the mutational spectrum. There is often a small number of mountains and a large collection of hills.[41]

The functional relevance of driver mutations within cancer genes—mountains and hills—may be revealed from their recurrent mutation patterns and clustering via the "20/20 rule."[33] Driver mutations usually influence gene products by granting a gain-of-function in oncogenes and a loss-of-function in tumor suppressor genes.[27,42] For a gene to be classified as an oncogene, >20% of missense mutations must tightly cluster at recurrent positions, usually within conserved/functionally relevant areas of the gene. To be classified as a tumor suppressor > 20% of mutations must be inactivating.[33] Although inactivating mutations may occur throughout most tumor suppressor genes, there are exceptions to this rule in which >20% of missense mutations will impart loss-of-function to a tumor suppressor, yet collectively cluster in a pattern indicative of drivers in oncogenes. An example of such an exception occurs in our investigation of Marek's disease lymphomas.

A typical tumor contains 1 to 10+ driver mutations and the remaining are classified as passenger mutations.[34] These estimates are derived from the number of coding substitutions under positive selection, which is possible because most cancer genomes receive limited impact from negative selection. On average tumors demonstrate approximately 4 coding variants under positive selection, which is commonly reflected by their mutational load. For instance, low mutation thyroid and testicular cancers demonstrated <1 coding variant under positive selection. Furthermore, in low-grade glioma, originating from non-self replicating glial cells, roughly 75% and about 15-20% of non-synonymous mutations in known cancer genes and the remaining protein-coding genes are predicted to be drivers, respectively. Whereas in melanoma, a cancer from self-renewing tissue often demonstrating higher mutational loads, only about 25% of non-synonymous variants are suspected to be drivers.[34]

# CHAPTER 2

## The Somatic Landscape of Marek's Disease Lymphomas

## Introduction

Marek's disease virus (MDV) is a highly oncogenic α-herpesvirus that drives Marek's disease (MD), a lymphoproliferative disease of chickens. Originally characterized by a chronic polyneuritis in 1907, Marek's disease has evolved into a more severe lymphoproliferative disease, demonstrating monoclonal invasion of peripheral nerves—resulting in paralysis—and the acute onset of metastatic T cell lymphoma in visceral organs.[1] Selective pressures first introduced in the 1960s—such as high-density poultry rearing practices, shorter growing periods, vaccination control strategies, and incorporated genetic resistance to Marek's disease—continue today and have increased the virulence of MDV field strains and the severity of the disease. Marek's disease is consistently listed as one of the top infectious poultry diseases of concern. The primary strategy of control is widespread vaccination, however, despite initial successes repeated vaccine breaks have occurred through the later half of the 20$^{th}$ century.[6,43] To develop more sustainable control strategies an understanding of the genetic etiology of lymphomagenesis is needed.[43]

Infection with Marek's disease virus alone is not sufficient to drive transformation; not every bird that is infected with Marek's disease virus develops Marek's tumors. We anticipated that a series of coordinated events were necessary to aid the oncogenic capacity of Marek's disease virus, most likely through the acquisition of somatic mutations. This hypothesis is consistent with the paradigm that cancer is a genetic disease that results from alterations of the genome.[25–27] A small minority of somatic mutations, referred to as "drivers," is predicted to

drive transformation by granting significant influence on cellular processes such as proliferation, dedifferentiation, and cell death.

The genesis of Marek's disease lymphomas is not well understood. Both integration of MDV into the chicken genome and the primary viral oncogene Meq[10,16] are necessary for transformation,[5,9,10] however, we suspect that additional somatically acquired mutations are required for transformation. The series of events that lead to Marek's disease lymphomas begins when MDV is inhaled and seeded in the respiratory tract where it is engulfed (presumably) by phagocytic cells and within 24 hours is transported to the spleen, bursa, and thymus. B and T lymphocytes are targeted by MDV for cytolytic infection, which peaks 3-6 dpi.[6] In B and T cells MDV usually transitions to latency within 7 dpi but may integrate into the genome as early as 1 dpi under experimental conditions.[5] CD4 T cells are the primary target for transformation and if transformed results in one or more monoclonal neoplasms detectable as gross lesions 2-4 weeks post infection.[6] CD4 T lymphocytes, are highly proliferative in the period from latency to their transformation.[7] This inflammatory environment, rich in cellular division—and presumably deregulated by MDV and Meq—is ideal for mutation and we suspect drives the transformation process. Somatic drivers and the mechanism(s) driving their accumulation in the Marek's disease cancer genome play a fundamental role in MDV-induced transformation.

To catalog somatic mutation types, frequencies, and mechanisms of accumulation, we performed a comprehensive investigation of the Marek's disease genome in metastatic tumors seeded in the gonad. To compare the genetic differences that were somatically unique to tumors, we incorporated whole genome sequencing (WGS), transcriptome sequencing (RNA-

Seq), custom Affymetrix SNP-arrays, and targeted genomic sequencing. This combination of technologies revealed an exceptionally low somatic mutation frequency. Tumors demonstrated between 0-3 non-synonymous mutations on average and approximately 0.3 mutations per MB in sufficiently powered samples. Recurrent non-synonymous mutations cluster in the DNA-binding domain (BDB) of *IKZF1*, which encodes Ikaros—a master regulator of T cell development and a known driver of acute leukemia in human. A high proportion of somatic C>T and G>A transitions suggests methylation-mediated spontaneous deamination of 5-methylcytocines at CpG dinucleotides is one likely endogenous mechanism contributing to somatic mutation accumulation. In this dissertation, we focus on mutations in *IKZF1*—our most confident candidates to drive Marek's disease lymphomas—and how they drive the progression of Marek's disease.

## Materials and Methods

### Experimental birds, tissue sampling, and extraction of DNA

The genetic backgrounds of birds used in these experiments were designed to reduce genetic variation between biological samples, while maintaining genetic backgrounds heterozygous for alleles associated with resistance and susceptibility to Marek's disease. Experimental birds were an $F_1$ cross between highly inbred parental lines 6 and 7, which were Marek's disease resistant and susceptible, respectively. $F_1$ progeny provided highly similar genetic backgrounds within progeny cohorts and served as biological replicates. Use of the

heterozygous 6x7 $F_1$ progeny allowed for genetic examination of the Marek's disease lymphoma genome landscapes in reference to both genetic backgrounds.

At hatch, 200 line 6x7 $F_1$ chicks were subcutaneously infected with 1,000 pfu of MDV strain JM/102W. Birds were cared for and monitored twice daily for evidence of moribundity. If birds became moribund or reached 8 weeks of age, they were euthanized and immediately necropsied. Tumors within each bird were counted and the largest were collected. Gonadal tumors were given priority because they were the largest and most homogenous. Tumor size was captured from pictures of tumors in reference to a scale. Images were processed with ImageJ[44] and an approximate estimate of tumor size was calculated. Tumors were collected from gonad, heart, spleen, thymus, pancreas, liver, proventriculus, kidney, and bursa. Normal tissue samples visibly absent of infiltrating lymphoma were also collected from liver and spleen. Tumor tissue was divided for different analyses: RNA-Seq samples were stored in RNAlater at -20°C and DNA-Seq and SNP-array samples were stored at -80°C. Additional controls for microarray analysis were extracted from six line 6x7 $F_1$ progeny (blood) that were not challenged with MDV.

RNA from CD4 T cells of unchallenged $F_1$ birds was used as matched normal in comparison to tumor RNA. CD4 T cells were extracted from spleens of uninfected birds via MACS CD4 T Cell Isolation Kit, human (Miltenyi Biotec, Bergisch Gladbach, Germany). A group of 16 uninfected birds from the same hatch were separately housed and were not challenged with MDV. Uninfected birds were necropsied for RNA in 4 cohorts of 4 birds at 2,4,6, and 8 weeks of age. Only birds from cohorts 2 and 6 weeks of age resulted in ample RNA of sufficient quality,

resulting in 8 samples. MACS extracted CD4 T cell samples were subjected to flow cytometry and measured for immune cell types.

Whole genomic DNA was extracted from frozen tumor and control tissue via the QIAamp DNA Blood Mini Kit (Qiagen; Germantown, MD). DNA integrity and quantity was measured via gel electrophoresis and Qubit. Whole genome sequencing (WGS) was performed in 3 different batches; 26 gonadal tumor samples from 22 tumors from 22 birds, 3 matched normal samples, and 19 matched normal samples. All samples underwent WGS 125 bp paired-end reads via the Illumina TruSeq Nano DNA Library Preparation Kit on Illumina HiSeq machines at the Michigan State University (MSU) RTSF Genomics Core.

RNA was extracted from tumor biopsies using the miTNeasy mini kit (Qiagen). RNA quantification and integrity was measured via the Agilent Bioanalyzer 2100 (Agilent, CA, USA). RNA sequencing was performed in 1 batch on 26 samples from 22 gonadal tumors (4 biological replicates) from 22 birds and 8 samples of isolated CD4 splenic T cells from 8 uninfected birds. RNA-Seq libraries were prepared using the NuGen Ovation Single Cell RNA-Seq System. All samples underwent sequencing to produce 125 bp paired-end reads on an Illumina HiSeq.

**Bioinformatic analysis of whole genome sequencing data**

***DNA-Seq processing and mapping***

The following analysis was designed following the Genome Analysis Toolkit (GATK) best practices pipeline.[45] Reads were inspected for quantity and quality before and after trimming with FastQC (v0.11.3).[46] Reads were trimmed of low quality bases and primers via Cutadapt

(v.1.14).[47] Trimmed reads were aligned to the Gallus gallus 5 reference genome[48] with BWA-MEM.[49] Read group annotation was added via Picard tools (v1.113).[50] Reads within each sample and sequencing lane were filtered of duplicate reads and were realigned around Indels via Picard tools (v1.113)[50] and GATK (v3.7.0).[51] Indel realignment was performed once more after samples were merged across lanes. Additional processing procedures were performed with SAMtools (v1.3.1).[52,53]

A rudimentary power analysis was performed to assess the power to detect heterozygous somatic point mutations within monoclonal tumor samples according to tumor purity. The coverage at each genomic loci, excluding micro chromosomes, was measured using SAMtools (v1.3.1)[52,53] across genomic positions with base and mapping qualities ≥ 20 in aligned files. A sequencing error rate of $10^{-3}$, a false positive rate of $5x10^{-7}$, and at least 3 supporting reads were required to predict the power to detect somatic variants based on coverage across the genome/exome using equations 9 and 10 from Carter et. al.[54]

***Detection of somatic variants***

Somatic copy number alterations

Somatic copy number alterations (CNAs) were detected using both whole genome sequencing data and Affymetrix SNP-arrays that were designed for our experimental model, $F_1$ chickens.[55] The log R ratio (LRR) and the B allele frequency (BAF) for each SNP-array was calculated using PennCNV-Affy.[56] The population B allele frequency (PFB) was generated from 6 uninfected $F_1$ birds using PennCNV.[56] LRR and BAF data was normalized for input into genoCN

(v1.26.0),[57] which is able to detect gain and loss of copy number as well as loss of heterozygosity (LOH). CNAs were also called in whole genome sequencing data via control FREEC (v9.8b)[58,59] (gains, losses, and LOH) and copycat (v1.6.11) (gains and losses), which is an extension of the ReadDepth algorithm.[60]

Somatic structural variants

Somatic structural variants (SVs) were called from WGS data using three callers: Breakdancer (v1.4.3),[61] Delly (v0.7.8),[62] and novoBreak (v1.1.3rc).[63] Breakdancer has the ability to detect multiple SV types including: deletions, insertions, inversions, intrachromosomal translocations, and interchromosomal translocations. Variants with a confidence score of ≥ 80 were called from genomic loci with a minimum mapping quality ≥ 20. Delly has the ability to detect copy-number variable deletions and duplications, inversions, and reciprocal translocations.[62] Variants ≥ 400 base pairs with at least 3 supporting variant reads at a minimum allele frequency of 0.10 were called from ≥ 10 reads of mapping quality ≥ 20. All putative candidates were further filtered and removed if found in any normal matched tissue. We used Novobreak to detect deletions, insertions, duplications, inversions, and translocations.[63] Variants ≥ 100 base pairs and mapping quality ≥ 20 were filtered using the empirically determined default filters of the algorithm[63] and a confidence score threshold ≥ 40.

Somatic gene fusions

Gene fusions were queried from paired-end RNA-Seq data mapped collectively to the Gallus gallus 5.0 reference genome[48] and the Gallid herpersvirus 2 genome (NC_002229.3)[64]

(detailed methods in Chapter 3). Both tumor and normal CD4 samples were queried via ChimPipe[65] with default parameters. Putative gene fusions in tumors were queried across the entire cohort of normal CD4 T cell samples to reduce false positive calls.

Somatic single nucleotide variants and small insertions and deletions

Somatic single nucleotide variants (SNVs) and small insertions and deletions (Indels) were called from WGS. Somatic SNVs were called from 6 algorithms: MuSE (v1.0rc_c039ffa),[66] MuTect2 (v1.1.7),[67] JointSNVMix2 (v0.7.5),[68] SomaticSniper (v1.0.5.0),[69] VarDict (v1.4.4),[70] and VarScan2 (v2.4.1).[71] Somatic Indels were called from 4 algorithms that were also used to call somatic SNVs (VarScan2, MuTect2, JointSNVMix2, and VarDict). The default hard filters were used for all algorithms.

Somatic SNVs and Indels were processed, annotated, and combined via SomaticSeq (v2.0.2),[72] SAMtools (v1.3.1),[52,53] GATK (v3.5),[51] Picard tools (v1.141),[50] and SnpEff (v4.0e)[73] with supporting annotation from the Gallus gallus dbSNP (build 145)[74] and SNPs from our Affymetrix SNP-arrays.[55] Additional annotation was applied to each putative variant to infer relevant functional and biological information associated with variants. To determine the effect of variants on genes, transcripts, and their protein products we generated a pipeline incorporating the Ensembl Variant Effect Predictor (VEP) (release 87),[75] SnpEff (v4.0e),[73] PROVEAN (v.1.1.5),[76] SIFT (v4.2)[77,78] to infer mutational impact in and around genes. To differentiate driver mutations from passenger mutations, all raw calls were fed into MuSiC (v0.0401),[79] oncodriveCLUST (v1.0.0)[80], and MUFFINN (v1.0.0).[81] Genes harboring and in proximity to putative variants were queried for high confidence orthologs in human via the

Ensembl database. Orthologs of all mutated genes were queried for cancer-specific annotations in the COSMIC database[82] (Figure 1. Somatic SNV and Indel detection pipeline).

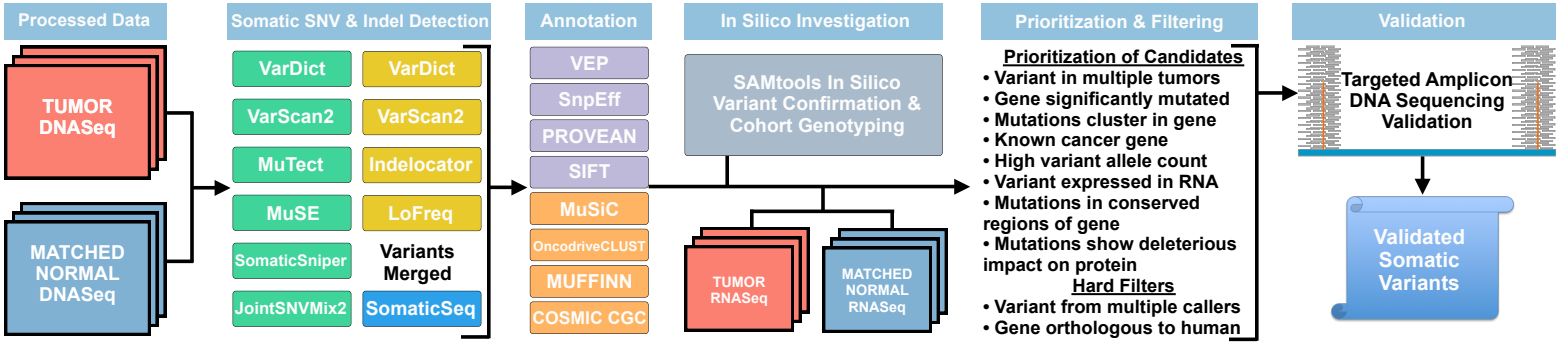



**Figure 1. Somatic SNV and Indel detection pipeline**: Six somatic SNV callers (green) and 4 somatic Indel callers (yellow) are run on tumor and matched normal pairs of aligned WGS BAM files. Resulting variants are annotated for functional impact on gene products (purple) and cancer-specific mutation and gene characteristics (orange). By means of SAMtools mpileup and custom investigation, variants are validated in silico; genotyped across the entire cohort of tumors and matched normal samples; and non-synonymous variants expressed in mRNA are annotated. After being prioritized by annotation and undergoing hard filters, variants are validated as true by targeted high coverage amplicon DNA sequencing.

### *Genomic survey of putative results*

Somatic variants were queried for overlaps (minimum 10% variant ≥ 100 base pairs) per sample within each respective variation type to distinguish the agreement between algorithms and technologies. Overlap between callers was visualized with UpSetR[83] and gridExtra.[84] Somatic variants were collectively visualized across samples as an "oncoprint" via GenVisR.[85]

## Validation of putative variants

### *In silico validation and prioritization*

The amount of calls from somatic SNV and Indel callers was very large suggesting a high number of false positive calls. To compensate, we applied additional filters with custom scripts. Genomic loci with reads of mapping and base qualities ≥ 20 were queried for in their respective BAM files with SAMtools (v1.3.1).[52,53] High quality reads were required to demonstrate a variant allele frequency ≥ 0.05 and coverage of ≥ 4x at loci in both the tumor and match normal samples. Each putative somatic variant was queried within tumor BAM files manually with Integrative Genomics Viewer (IGV)[86,87] and with a custom script using SAMtools mpileup.[52,53] Somatic non-synonymous SNVs and Indels variants were granted more confidence if they appeared in mRNA sequence reads.

Putative variants were considered for validation based on the following criteria in order of importance: variants called by multiple algorithms, variants in multiple tumors, variants in known cancer genes according to the literature and the COSMIC database,[82] variants with high

impact mutations, and variants supported by more 2 or more reads. In total, 733 putative variants met these criteria.

***Targeted DNA sequencing and variant validation***

Approximately 150 of the most high confidence variants (136 genomic loci) were tested for validation. In total, 188 samples were tested including tumors and matched normal samples of the gonad, heart, spleen, thymus, pancreas, liver, proventriculus, kidney, and bursa. Additionally, samples from uninfected lines 6 and 7 and Marek's disease cell lines MSB1, RP2, RP19 were tested. Targeted Illumina DNA sequencing (Illumina MiSeq 150 bp paired-end reads) was performed on each genomic region of interest across all samples via the Agriplex Genomics PlexSeq$^{(TM)}$ method.

Resulting amplified regions were analyzed via 2 independent methods and compared: an analysis from Agriplex via the PlexCall$^{(TM)}$ software and an in-house investigation with custom scripts utilizing SAMtools mpileup[52,53] and VarDict.[70] Reads were trimmed of low quality bases and primers via Cutadapt (v.1.14).[47] Trimmed reads were aligned to regions of interest from the Gallus gallus 5 reference genome[48] with BWA.[49] Further processing of aligned reads was performed with Picard tools (v1.113),[50] Bamtools (v2.2.3),[88] and GATK (v3.7.0).[51] Reads with a mapping quality ≥ 60 for somatic SNVs and ≥ 20 for somatic Indels were queried with SAMtools (v1.5)[52,53] and VarDict (v1.4.4)[70] and variant allele frequencies were calculated for all variants. To rule out germline variants putative somatic variants were filtered via SAMtools mpileup[52,53] if any matched normal sample from within the cohort demonstrated a variant allele frequency ≥ 0.1; normal samples with contaminating tumor tissue were omitted from this step. Variants

were called if there were ≥ 5 variant supporting reads, ≥ 500x high quality read coverage, and ≥ 0.01 variant allele frequencies.

**Annotation of *IKZF1* variants**

Ikaros protein isoforms were collected from the UniProt database[89] and referenced from IKZF1-201 and IKZF1-202 transcript sequences from both Ensembl[90] and RefSeq.[91] Ikaros isoforms were compared to proteins in the UniRef90 database[92] via pBlast (E-threshold of 0.001)[93,94] and resulting protein sequences were aligned via Clustal Omega.[95] Hierarchical analysis of amino acid residue conservation[96] was performed within JalView[97] and results were reduced to 5 species. A custom illustration of somatic mutations was performed on Ikaros alignments.

**Results and Discussion**

**Genomic datasets and sample cohorts**

Marek's disease lymphomas are known to be driven by the MDV bZip transcription factor, Meq. MDV-infected birds demonstrate genomic integration of MDV in CD4 T cells and latent expression of Meq, however, not all birds develop tumors. Therefore, we hypothesize that additional somatic mutations are required to promote tumorigenesis. The genomes of MDV-transformed CD4 T cells that constitute Marek's disease tumors have not been thoroughly examined for somatic mutations. To investigate the genomic landscape of Marek's disease tumors, we studied two cohorts (Hatch 1 & 2) each of 200 experimental white leghorn chickens,

challenged at one day of age with 1,000 plaque-forming units (pfu) of virulent MDV. All birds were $F_1$ progeny from highly inbred parental lines $6_3$ (MD resistant) and $7_2$ (MD susceptible), which have been used extensively for studies on genetic resistance to Marek's disease (e.g. Stone, H., 1975[98] and Cheng H. H. et al., 2015[55]). MDV strain JM/102W was chosen because it promotes the induction of large gonadal tumors. All birds were necropsied for gross tumors until moribund or until eight weeks of age. In total, 117 birds (29%) demonstrated one or more gross focal tumors across multiple tissues. Genomic material extracted from the largest gonadal tumors—22 tumors and biological replicates from 4 of the largest tumors—underwent whole genome sequencing (WGS), transcriptome sequencing (RNA-Seq), and custom Affymetrix SNP-arrays. In each of these 22 birds, DNA from normal matched tissue also underwent WGS and was collected from tissue free of visible lesions (Table 1. Summary of biological samples and genomic datasets).

**Table 1. Summary of biological samples and genomic datasets:** Birds (Line 6x7 $F_1$) from 2 hatches were subjected to multiple genomic screens.

| Sample | Hatch 1 | Hatch 2 | Total | |
|---|---|---|---|---|
| Challenged birds | 200 | 200 | 400 | |
| Unchallenged birds | 20 | 0 | 20 | |
| Tumor positive birds | 54 | 63 | 117 | |
| Tumors | 112 | 112 | 224 | |
| **Genomic Test** | **Normals** | **Tumors** | **Birds** | **Hatch** |
| Whole genome sequencing | 22 | 22 | 22 | 1 |
| Targetted DNA-Seq | 22 | 153 | 98 | 1 & 2 |
| RNA-Seq | 8 | 14 | 22 | 1 |
| Affymetrix SNP-arrays | 0 | 60 | 36 | 1 |

**Preliminary investigation of the somatic landscape**

In a preliminary investigation, we used multiple genomic screens (WGS, RNA-Seq, and SNP-arrays) and variant calling algorithms to query somatic variants from 22 gonadal tumors. Priority was given to variants called by multiple algorithms, which we refer to as candidate variants (Figure 2. Agreement between somatic mutation algorithms). The majority of variants were somatic SNVs and Indels, which we investigated in greater scrutiny and validated with high coverage targeted DNA sequencing. Somatic SVs, CNAs, and gene fusions remain candidates and share pathways with validated somatic SNVs and Indels (Tables S1-4).

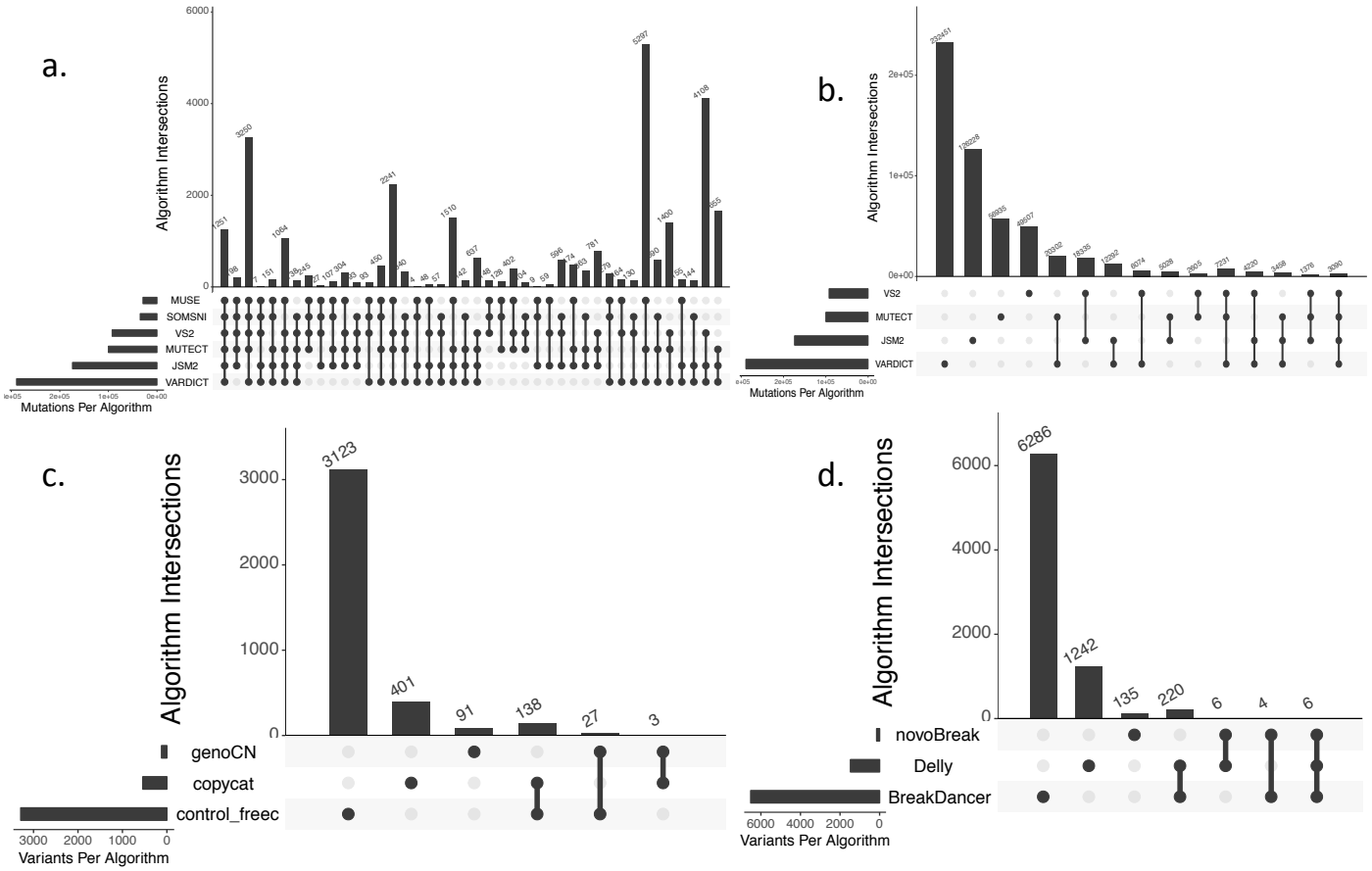



**Figure 2. Agreement between somatic mutation algorithms**: The agreement between somatic variant calling algorithms by variant type. Somatic SNVs (a), Indels (b), CNAs (c), and SVs (d). The height of bars and annotated values represent variant frequencies—the agreement—between algorithms. The horizontal bar plots to the left of algorithm names represent the total variants called per algorithm. Algorithms used include: MuSE "MUSE", SomaticSniper "SOMSNI", VarScan2 " VS2", MuTect "MUTECT", JoinSNVMix2 "JSM2", and VarDict "VARDICT", genoCN "genoCN", CopyCat "copycat", ControlFREEC "control_freec", novoBreak "novoBreak", Delly "Delly", BreakDancer "BreakDancer".

**Marek's disease lymphomas demonstrate low somatic mutation frequency**

The majority of well-annotated cancer driver mutations occur in coding regions of protein coding genes,[25] and somatic SNVs make up the majority of cancer driver mutations in common tumors in human.[33] Somatic SNV and Indel detection has not been comprehensively performed in Marek's disease, so we took inspiration from a successful machine learning pipeline that used SomaticSeq[72] and created a similar workflow that incorporated multiple callers,[66–71,99,100] disease-associated annotations, and multiple sequencing technologies: WGS, RNA-Seq, and amplicon DNA resequencing (Figure 1. Somatic SNV and Indel detection pipeline).

Somatic mutation rates in Marek's disease genomes were exceptionally low—approximately 0.30 mutations per MB in high and medium powered samples (Table 2. Somatic single nucleotide variant frequency in Marek's disease lymphomas). Tumors demonstrated between 0-3 non-synonymous mutations each.

**Table 2. Somatic single nucleotide variant frequency in Marek's disease lymphomas:** The Average somatic mutation frequencies across samples in high, medium, and low purity tumors.

| Estimated Purity[Ŧ] | Tumors | SNVs per Tumor* | Nonsynonymous SNVs per Tumor* |
|---|---|---|---|
| High | 9 | 329 +/- 67.6 | 2.44 +/- 1.53 |
| Medium | 8 | 263 +/- 81.5 | 2.00 +/- 1.36 |
| Low | 6 | 165 +/- 29.3 | 0.83 +/- 0.61 |

[Ŧ] Tumor purity estimated from CD4 mRNA expression: High (x>x), Medium (x>x), Low (x>x).
* Mean ± 2 standard errors.

Somatic mutation signatures across Marek's disease tumors are enriched for C>T and G>A transitions, of which many occur in CpG dinucleotides; (Figure 3. Mutation signature of somatic single nucleotide variants in Marek's disease lymphomas) it is possible that

methylation-mediated spontaneous deamination of 5-methylcytosines is helping to drive the acquisition of C>T and G>A SNVs.

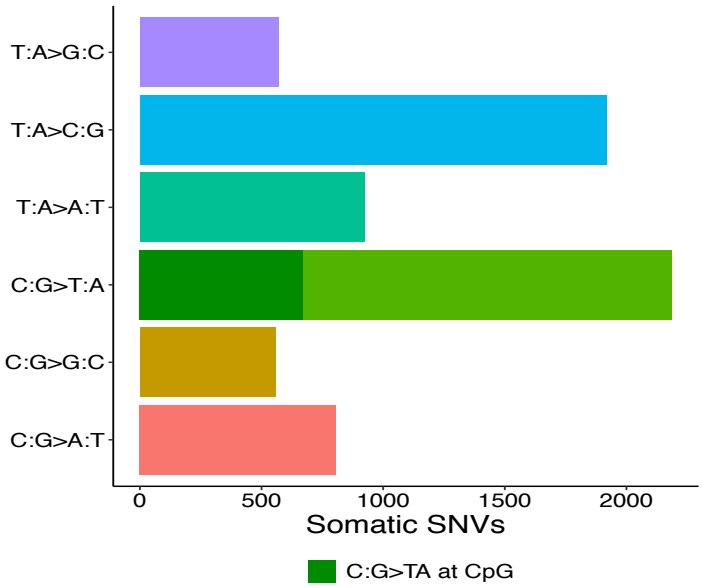



**Figure 3. Mutation signature of somatic single nucleotide variants in Marek's disease lymphomas:** The mutation signatures of high confidence and validated somatic single nucleotide variants.

**Whole genome sequencing identifies IKZF1 as frequently mutated**

The workflow above yielded seven validated, somatic non-synonymous mutated genes (*IKZF1, PRDM12, DHX35, DESI2, BAG1, PEX10,* and *ATP6V1C1*). The most frequent and recurrent somatic mutations across tumors occurred in *IKZF1*, the gene encoding the transcription factor Ikaros. Approximately 41% of Marek's disease gonadal tumors tested contained somatic non-synonymous mutations in *IKZF1*. Furthermore, 60% of somatic non-synonymous mutations targeted *IKZF1*—the only gene to demonstrate non-synonymous SNVs and Indels across tumors (Figure 4. Summary of somatically mutated genes in Marek's disease lymphomas).

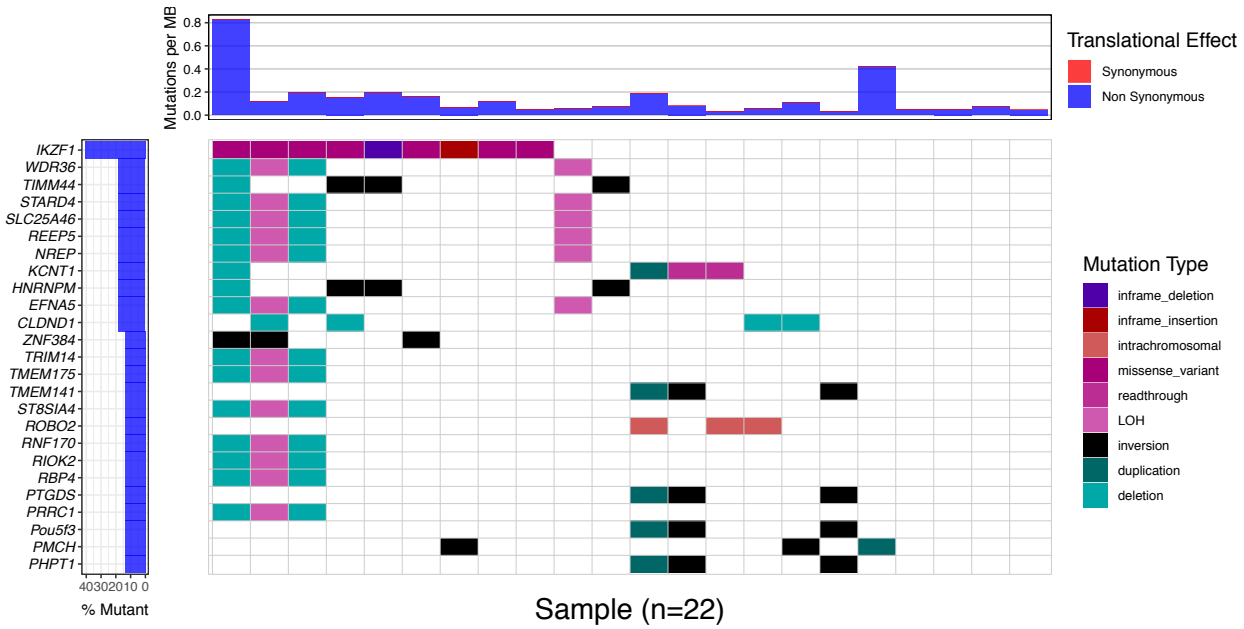



**Figure 4. Summary of somatically mutated genes in Marek's disease lymphomas**: Recurrently mutated genes in Marek's disease lymphomas harboring somatic variants. Genes with somatic non-synonymous mutations are listed in the Y-axis in decreasing order of occurrence. The column labeled "%Mutant" represents the percentage of samples with non-synonymous somatic mutations per gene across 22 tumors and 4 additional replicates sequenced.

**Genomic gains and losses from whole genome sequence and SNP-arrays**

We complemented our initial analysis of 22 tumors that underwent WGS to query large genomic variants with supporting Affymetrix 15K SNP-arrays. We detected deletions and amplifications called by some combination of Breakdancer,[61] Copycat,[60] genoCN,[57] novoBreak,[63] control FREEC,[58,59] and Delly.[62] In general, we measured more deletion events than amplifications—a trend which has been previously reported across malignancies.[101] We measured 11 candidate amplifications in 3 tumors. Amplifications were large, often, spanning entire chromosomes or chromosomal arms.

Amplifications occured sparsely. One sample demonstrated 3 large amplifications. The largest event occurred across 20% of chromosome 1 (p arm) spanning cancer genes *MDM2, LRIG3, WIF1, PDGFB, HMGA2, MET, BRAF, ST7, BTG1,* and *MYH9*. We also measured

amplifications in the entire q arm of chromosome 17 (*GARNL3*), across 40% of chromosome 23 (*RCAN3*), and in the entire the q arm of chromosome 26 (*MICAL1*). Additional tumors demonstrated a duplication of chromosome 17 (*NOTCH1, SET, PPP6C, CNTRL*, and *TNC*) and small amplifications that converged on a RFAM and miRBase predicted[102] miRNA (ENSGALG00000031911) in the middle of the q arm of chromosome 1. Cancer genes that spanned these amplifications included kinases and growth factors (*MET*, *PDGFB*, and *BRAF*), microtubules-associated genes (*MICAL1* and *CNTRL*), contained EGF motifs (*NOTCH* and *TNC*), and demonstrated significant differential gene expression across tumors (*MET, LRIG3*, and *ST7*).

Large genomic deletions and losses were more frequent across the tumor cohort. In total, we observed 178 losses across 18 tumors. There was a median of 6 deletions per tumor with one tumors demonstrating 46 deletions across major chromosomes, considerably more than any other tumor. The majority of deletions were either small and focal or the length of an entire chromosome or chromosome arm—also a common trend in human malignancies.[101] We detected 10 deletions of chromosome arms in 6 tumors, most commonly in chromosome 27 and the p arm of chromosome 3. One tumor demonstrated total deletion of chromosomes 24 and 28. In total, 50 genes from the COSMIC Cancer Genes Consensus were deleted, of which the most frequent were *CLDND1* and *FOXD1*. Focal deletions in 4 samples converge on *CLDND1* and a mix of large and small events converge on *FOXD1* in 3 samples. Recurrent deletion breakpoints were measured in *ADARB1*, *PDGFD*, and *TARP* and lone breakpoints were found in cancer genes *FAT3*, *GPHN*, and *XPO1*.

**Genomic structural rearrangements in Marek's disease lymphomas**

We detected somatic inversions and loss-of-heterozygosity (LOH) events from genomic DNA sequencing and SNP-arrays. LOH events were predicted using controlFREEC[58,59] and genoCN.[57] To our surprise, LOH events were infrequent across tumors; we chose highly heterozygous $F_1$ birds for this experiment to better detect LOH events. Only 3 candidate events were detected in 2 tumors. Two LOH events targeted chromosome Z in 2 tumors and converged on a small cluster of genes (*EFNA5, NREP, REEP5, SLC2546, STARD4*, and *WDR36*); a segment that was also targeted for deletion in 2 additional tumors. The other sizable event (37 MB) was observed in the p arm of chromosome 7 over a cluster of cancer genes (*ERBB4, COL3A1*, and *ITGAV*), of which *COL3A1* is the most recurrently mutated.

Intrachromosomal inversions were predicted using Breakdancer,[61] novoBreak,[63] and Delly[62] (interchromosomal events were not considered in this analysis). A total of 115 intrachromosomal inversion candidates were detected across 20 tumors. Samples demonstrated between 1 and 17 (median = 5) events per tumor. The largest event observed was 5.12 MB and the remaining were focal. Recurrent events targeted 26 genes; only one cancer gene—*ZNF384*—was recurrently targeted. The gene *IGF1* was the only gene to demonstrate recurrent breakpoints from inversions. Other targeted cancer genes include *OLIG2, KDSR, IRF4, ETV1*, and *BCL2*.

One algorithm was used to detect gene fusions from RNA-Seq. The roundabout protein *ROBO2* demonstrated candidate frameshift intrachromosomal chimeras with pre-mature stop codons in 4 tumors, respectively (Table S4. Somatic gene fusion candidates). *ROBO2* is a known driver gene in colorectal cancer, adenocarcinoma, and melanoma.[82]

These diverse mutation types converge to target common pathways in complementary fashion. Collectively, somatic driver candidates target pathways associated with immune response, chromatin modification, lymphocyte differentiation, and neural growth. There are many genes that warrant further investigation; in this dissertation we focused our analyses on the most (likely) influential driver—*IKZF1*.

**There is an association between *IKZF1* mutation status and gonadal tumors**

Empirical knowledge and prior studies have shown that the diversity and distribution of neoplastic lesions is contingent upon the strain of MDV and the genetic background of its host.[103–107] We chose a specific strain of MDV—JM/102 W[108]—for its consistent ability to produce large gonadal tumors.[103–105,107] Although we expected to observe an enrichment of gonadal tumors from all collected tumor tissue, we considered whether mutations in *IKZF1* were preferential to tumors of certain tissue types; alterations in certain driver genes appear tissue-specific.[109]

To determine the distribution of non-synonymous somatic mutations across tumors of different tissue types, we queried results from targeted deep sequencing (~1,500x) in 158 tumors (98 birds) of gonad, spleen, heart, proventriculus, liver, kidney, pancreas, thymus, and bursa (Table 1. Summary of biological samples and genomic datasets). Additionally, samples from uninfected lines 6 and 7 and Marek's disease cell lines MSB1, RP2, RP19 were tested. In total, 150 variants across 136 genomic loci (50 bp) were queried.

The only recurrent non-synonymous SNVs and Indels across tissues occurred in the DNA-binding domain of *IKZF1*. Twenty-four tumors from 19 birds demonstrated *IKZF1*

mutations in gonad, spleen, heart, liver, and proventriculus. The majority of mutations occurred in tumors seeded in gonad consisting of 31% of gonadal tumors tested across the entire cohort (Table 3. Tumors subjected to targeted DNA-Seq and somatic variant validation). Tumors harnessing mutations in *IKZF1* were also found in heart, spleen, proventriculus, and liver in 7 birds. Five of seven birds also harbored a gonadal tumor with the identical somatic *IKZF1* variant; gonads were not sequenced in the remaining two birds. To test if *IKZF1* mutations were associated with gonadal tissue, a stratified logistic regression was performed adjusting for splenic tumors (splenic tumors lacked *IKZF1* mutations). Among non-splenic tumors, gonadal tumors have 3.36 times the odds to contain an *IKZF1* mutation compared to other tumor types (95% C.I. = 1.47 - 8.35; p-value = 0.0208). In mice, heterozygous drivers in the DNA-binding domain of *IKZF1*, similar to those found in Marek's disease tumors, causes lymphoproliferative characteristics in which lymphocytes aggressively infiltrate nonlymphoid organs.[110] It is possible that putative driver mutations in the DNA binding domain of *IKZF1* may correlate with the ability of transformed CD4 cells of Marek's disease lymphomas to preferentially migrate and invade gonadal tissue.

**Table 3. Tumors subjected to targeted DNA-Seq and somatic variant validation:** Marek's disease tumors are grouped by tissue type.

| Tissue | Tumors | *IKZF1*-mutant tumors |
|---|---|---|
| Gonad | 55 (36%) | 17 |
| Spleen | 47 (31%) | 1 |
| Heart | 28 (18%) | 4 |
| Proventriculus | 9 (6%) | 1 |
| Liver | 6 (4%) | 1 |
| Kidney | 4 (3%) | 0 |
| Pancreas | 2 (1%) | 0 |
| Thymus | 1 (1%) | 0 |
| Bursa | 1 (1%) | 0 |

***IKZF1* mutations target the DNA-binding domains of Ikaros**

In total, 35 somatic non-synonymous *IKZF1* mutations were found across 24 tumors (Table 4. Validated somatic mutations in *IKZF1*). With respect to *IKZF1*, all somatic variants clustered in the critical $C_2H_2$ zinc finger binding domains in exon 4 (Figure 5. Validated focal somatic mutations in *IKZF1* mapped onto Ikaros). Ikaros isoforms 1 (Ensembl IKZF1-201; RefSeq IKZF1-X1) and 2 (Ensembl IKZF1-202; RefSeq IKZF1-X3) were chosen for this analysis because isoforms 1 and 2 are the most abundantly expressed isoforms throughout development of hematopoietic cells in both human and mouse.[111,112] Chicken Ikaros isoforms 1 and 2 closely resemble the human Ikaros isoforms 1 and 2 with 86% and 78% amino acid identity, respectively.

**Table 4. Validated somatic mutations in *IKZF1*:** A total of 35 Somatic mutations in *IKZF1* in 24 tumors from a cohort of 132 additional tumors. PROVEAN and SIFT predict deleterious impact of somatic mutations.

| Chrom | Pos | Ref | Alt | Variant | AA Change | PROVEAN | SIFT | Prediction | Samples (DNA) | VAF (DNA) |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 80972101 | C | T | missense variant | p.Arg162Cys | -6.945 | 0 | Deleterious | 777-1; 862-1; 901-2; 901-2-2; 841-3; 777-3 | 0.012; 0.136; 0.167; 0.140; 0.026; 0.522 |
| 2 | 80972102 | G | T | missense variant | p.Arg162Leu | -6.098 | 0 | Deleterious | 835-1 | 0.173 |
| 2 | 80972104 | C | T | missense variant | p.His163Tyr | -5.237 | 0 | Deleterious | 842-2-2; 927-2; 842-2 | 0.451; 0.042; 0.427 |
| 2 | 80972114 | T | TGCACTC | inframe insertion | p.His167 Ser168dup | -9.668 | NA | Deleterious | 43-G | 0.061 |
| 2 | 80972116 | C | T | missense variant | p.His167Tyr | -5.205 | 0 | Deleterious | 44-G; 901-2; 901-2-2 | 0.106; 0.011; 0.045 |
| 2 | 80972118 | C | G | missense variant | p.His167Gln | -6.817 | 0 | Deleterious | 112-G; 756-2; 756-3 | 0.020; 0.015; 0.022 |
| 2 | 80972141 | G | A | missense variant | p.Cys175Tyr | -9.652 | 0 | Deleterious | 36-H; 36-G; 901-2; 874-2 | 0.098; 0.157; 0.023; 0.055 |
| 2 | 80972141 | G | GCCA | inframe insertion | p.His176dup | -8.770 | NA | Deleterious | 901-2; 874-2 | 0.020; 0.054 |
| 2 | 80972149 | T | A | missense variant | p.Cys178Ser | -8.767 | 0 | Deleterious | 35-G | 0.165 |
| 2 | 80972149 | TGTAACTACGCCTGCCGGCGCA | T | inframe deletion | p.Cys178 Arg185delinsTrp | -56.780 | NA | Deleterious | 911-1; 911-2; 911-1-2 | 0.181; 0.117; 0.094 |
| 2 | 80972152 | A | AACT | inframe insertion | p.Tyr180dup | -8.854 | NA | Deleterious | 918-3 | 0.216 |
| 2 | 80972167 | C | T | missense variant | p.Arg184Cys | -7.028 | 0 | Deleterious | 927-2; 776-1 | 0.147; 0.146 |
| 2 | 80972168 | G | A | missense variant | p.Arg184His | -4.390 | 0 | Deleterious | 112-G; 26-H | 0.223; 0.010 |
| 2 | 80972173 | GA | TT | missense variant | p.Asp186Phe | -7.913 | 0 | Deleterious | 25-G | 0.071 |
| 2 | 80972174 | A | G | missense variant | p.Asp186Gly | -6.128 | 0.1 | Deleterious | 112-P; 112-G | 0.161; 0.059 |

In human and mouse Ikaros, zinc fingers 2 and 3 are necessary for DNA binding, which strongly suggests that MD tumor specific *IKZF1* mutations disrupt the ability of Ikaros to bind to DNA, resulting in loss of tumor suppressor function.[111] All somatic mutations in *IKZF1* are concentrated in a conserved functionally relevant DNA-binding domain**.**

At both the genome and protein levels, *IKZF1* and Ikaros are highly conserved. All of the amino acid residues targeted for substitution by non-synonymous mutations in Marek's disease tumors are conserved across species from sea lamprey—the most distant species with orthologous Ikaros—to human (Figure 5. Validated focal somatic mutations in *IKZF1* mapped onto Ikaros). Ikaros function is also highly conserved.[113] Furthermore, all residues targeted by missense mutations and in-frame deletions are considered essential for human Ikaros to bind to DNA (Figure 5. Validated focal somatic mutations in *IKZF1* mapped onto Ikaros).[114] In zinc finger 2, these targeted mutations include the two zinc chelating histidines (163H; 167H) as well as the arginine at position 6 of the alpha helices (162R). In zinc finger 3, essential residues targeted include the two zinc chelating cysteines (175C; 178C), as well as arginine (184R) and aspartic acid (186D) at positions -1 and 2 of the alpha helices, respectively. We also noticed in-frame insertions that do not remove essential residues. We speculate that in the alpha helix of zinc finger 2 and the beta-sheet of zinc finger 3, somatic in-frame insertions may elongate these essential structures in attempts to denature these zinc fingers (Figure 5. Validated focal somatic mutations in *IKZF1* mapped onto Ikaros). Consistent with our observations, PROVEAN and SIFT protein analysis both predicted a loss of protein function for all non-synonymous variants (Table 4. Validated somatic mutations in *IKZF1*).

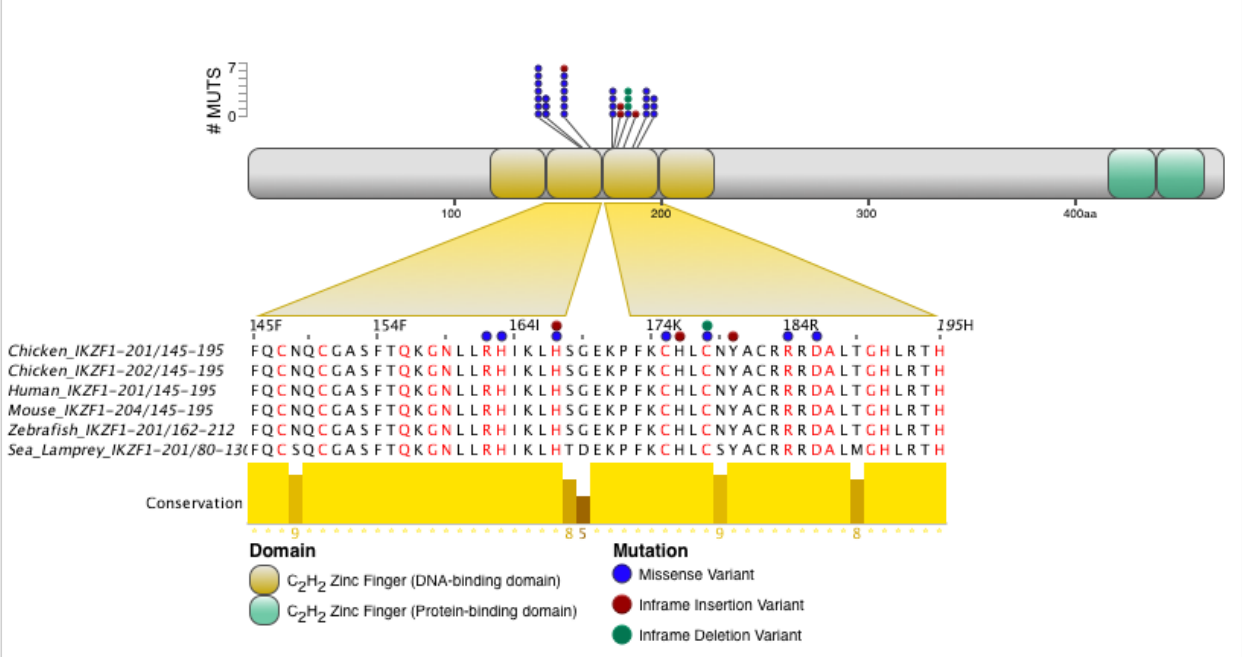



**Figure 5. Validated focal somatic mutations in *IKZF1* mapped onto Ikaros**: *IKZF1* alterations in Marek's disease lymphomas. A schematic presentation of the Ikaros protein showing the collection of all validated somatic non-synonymous mutations from Marek's disease tumors. Mutations cluster on essential amino acids (red) for DNA-binding in zinc fingers 2 and 3. Amino acid conservation scores (yellow bars under sequence) were calculated[96] from aligned consensus protein sequences.[95,97]

Ikaros is a known tumor suppressor gene (TSG) in human leukemia and lymphoma (e.g. T and B cell acute lymphoblastic leukemia [T-ALL; B-ALL], Acute myeloid leukemia [AML], Chronic myelogenous leukemia [CML], and Diffuse large B-cell lymphoma [DLBCL]).[115,116] The mutational signature more closely resembles an oncogene with a gain of function based on the 20/20 rule.[33] It is also known in human and mouse that the two C-terminal zinc fingers are necessary for dimerization with itself and other Ikaros family members,[110,117,118] providing a mechanism for how somatic mutations in the N-terminal zinc finger domains can act as dominant negatives. Our data suggest heterozygous mutations resulting in the same dominant negative behavior observed in human lymphoma and leukemia;[117] DNA (0.010 to 0.523) and mRNA (0.029 to 0.466) variant allele frequencies and estimated tumor (96% or less) collectively suggest

heterozygous mutations and stable copy number (Table 4. Validated somatic mutations in *IKZF1*). Furthermore, the diverse mutations are entirely in-frame, thus preserving the two C terminal zinc fingers required for protein dimerization. Therefore, we believe that mutated Ikaros loses DNA-binding function yet retains protein-binding ability to become a dominant negative driver of Marek's disease lymphomas.

# CHAPTER 3

## Mutations in *IKZF1* Driver Marek's Disease Lymphomas

**Introduction**

The somatic landscapes of Marek's disease (MD) lymphomas reflect it's unique nature: a young, virally induced, aggressive/metastatic, lymphoma originating from a differentiated CD4 T cell. Compared to other well-studied cancer types, Marek's disease lymphomas share similarities with pediatric leukemias in humans, especially acute leukemias such as acute myeloid leukemia (AML) and acute lymphocytic leukemia (ALL).[119–121] MD, AML, and ALL demonstrate mutations in the tumor suppressor gene, *IKZF1*. Although we previously demonstrated that Marek's tumors harbor somatic mutations in *IKZF1* indicative of drivers, we must also scrutinize the transcriptional landscape of Marek's tumors for the effects of (potentially) dominant negative Ikaros isoforms.

The most frequently mutated gene in Marek's disease lymphomas is *IKZF1,* which likely acts as a dominant negative driver. *IKZF1* and its collective isoforms—Ikaros—are the founding members of the zinc finger family of transcription factors.[122] Family members also include Ikaros homologs Helios, Aiolos, Eos, and Pegasus.[123] Each family member contains two C-terminal $C_2H_2$ zinc-fingers mediating protein-protein interactions and up to four N-terminal $C_2H_2$ zinc fingers for recognition of DNA-sequences. The number of DNA-binding zinc fingers may vary among isoforms of each family member due to deletions or alternative splicing. Isoforms lacking N-terminal zinc fingers often result in null or dominant-negative products.[110] For instance, small deletions over critical N-terminal DNA-binding zinc fingers in *IKZF1* produce dominant-negative Ikaros isoforms that may hetero-dimerize with wild type isoforms to drive human and murine hematopoietic malginacies.[115,118,124,125] Furthermore, isoforms may interact

with each other from across the family generating a large number of combinations.[123] Much like alternative splicing and deletions, point mutations have the capacity to generate isoforms with null or dominant-negative function.[114] Marek's lymphomas demonstrate point mutations and in frame insertions and deletions in the *IKZF1* N-terminal binding domains, suggesting that these mutations may generate null or dominant negative Ikaros transcripts.

Both the sequence and function of Ikaros are conserved across species[126–128] and the mutations we see in Marek's tumors are well studied for their functional impact in other systems.[129] To better characterize mutations in *IKZF1* across Marek's tumors we thoroughly examined sequencing data (DNA-Seq and RNA-Seq) and supporting metadata from our 22 gonadal tumor cohort. *IKZF1* mutations target residues critical for Ikaros function, strongly suggesting they are legitimate drivers. Further evidence was found in mutant transcripts of *IKZF1*; however, to our surprise, no alternatively spliced dominant negative transcripts lacking N-terminal zinc fingers were evident.

Marek's tumors demonstrated expression profiles indicative of genomic reprogramming by mutant Ikaros isoforms. Profiles were remarkably similar across tumors and differentially expressed genes were enriched for targets of Ikaros regulation. Differentially expressed genes, pathways, and ontologies were indicative of hematopoietic malignancies and developing neurons. The most enriched pathways and ontologies included: cellular adhesion, cellular migration, extracellular matrix organization, axon guidance, lymphocyte dedifferentiation, and a stem-cell state. We strongly suspect that tumors arise from differentiated T cells reprogrammed to a stem-cell state. Furthermore, it is possible that transformed stem-like T

cells elicit the highly proliferative attributes of developing neurons by awaking the, otherwise, dormant pathways of early neuron development.

## Materials and Methods

### RNA-Seq processing and mapping

RNA-Seq reads were assessed for quality before and after read trimming via FastQC (v0.11.3).[46] Reads were trimmed for adaptors and low quality reads via Trimmomatic (v0.33).[130] Reads were mapped to Gallus gallus ribosomal RNA (rRNA) obtained from GenBank[64] to determine the proportion of rRNA per sample and to compare to percentages of reads successfully mapped to the Gallus gallus 5 reference genome. Reads that mapped to rRNA and did not map to the reference genome were removed. Reads were mapped to both the Gallus gallus 5.0 reference genome[48] and the Gallid herpersvirus 2 genome (MDV genome) (NC_002229.3);[64] the MDV genome was added to the Gallus gallus 5 fasta file as an additional chromosome. MDV annotations were retrieved from GenBank[64] and adjusted with custom Python scripts to gff3 and gtf file formats. Reads were mapped to the combined chicken and MDV genomes via the spliced read mapper—TopHat (v2.1.0) of the Bowtie suite (v2.2.6)[131]. The total reads successfully mapped to the reference genome was measured and BAM files were generated and further processed via SAMTools (v1.2).[52,53]

**Power of RNA-Seq analysis**

Gene counts were investigated under an unsupervised model in R using the DESeq2 infrastructure.[132] Genes with low expression of < 10 reads per gene were removed.[133] Gene counts were normalized (regularized log).[132] Single value decomposition,[96] subsequence principle component analysis (PCA)[132] and complete linkage hierarchical clustering[135] were performed to assess potential factors influencing variance in the data, whether biological or technical. Factors influencing variance were explored under different models with aid from Surrogate Variable Analysis (SVA).[136] Assessment of RNA-Seq power was performed with datasets and interactive analyses from the PROPER package.[133]

**Differential gene expression analysis**

Differential expression analysis was performed with DESeq2.[132] Two comparisons were made under identical technical parameters to detect differentially expressed genes. To compare tumors against CD4 T cell samples, nine tumor samples were compared to 8 samples of uninfected CD4 T cells. To compare tumors with mutations in *IKZF1* from those without, 6 tumors with mutations and 3 without were compared. All tumors demonstrated estimated purity ≥ 50%. Genes were differentially expressed if they demonstrated a $\log_2$ fold change ≥ 0.50 and an FDR ≤ 0.05. Statistical models in DESeq2 were adjusted based on estimated tumor purity. Genomic annotation was incorporated from Ensembl[90] via GenomicFeatures[137] and from AnnotationDbi.[138]

**Pathway enrichment analysis**

The list of 1,598 differentially expressed genes between Marek's disease tumors and normal CD4 T cell samples was queried for enrichment in custom gene sets and databases associated with gene ontologies,[139,140] pathways (e.g. Kegg and Reactome),[141–148] gene sets,[149,150] and putative genes regulated by Ikaros.[151] Enrichment was performed with Enrichr,[152,153] gprofiler[154], and gene set enrichment analysis (GSEA).[149] Heatmaps were generated with pheatmap.[155]

**Differential exon expression in *IKZF1***

A candidate intragenic deletion in *IKZF1* was detected with Delly.[62] Although the deletion was only supported with 2 reads and did not pass our filtering parameters, it was investigated. WGS reads and supporting mRNA sequencing reads spanning the *IKZF1* locus were manually investigated and visualized via IGV (v2.3.91).[86] mRNA reads were quantified per exon via Subread (v1.6.1)[156,157] using the NCBI Gallus gallus 5.0.86 annotation.[91] Differential *IKZF1* exon expression was assessed and visualized via DEXSeq (v1.28.0).[158,159]

**Results and Discussion**

**Analysis of Marek's disease lymphomas with RNA-Seq**

To investigate the transcriptional landscape of Marek's disease tumors, RNA-Seq was performed on 14 of the 22 gonadal tumors that underwent WGS (Table 1. Summary of biological samples and genomic datasets). Most Marek's disease tumors originate form

transformed CD4 T cells,[160] therefore, transcriptome profiles of isolated CD4 cells from 8 uninfected $F_1$ 6x7 birds were compared to tumor transcriptome profiles. RNA-Seq analysis revealed that tumors were primarily monoclonal with an average estimated purity of 45% (33.2% – 56.88%, 95%C.I.) and demonstrated gene expression profiles indicative of their respective somatic mutation landscape suggesting T cell dedifferentiation, "stemness", increased adhesion, and expression of transcripts in neural growth and development pathways.

### *IKZF1* dominant negative transcripts and splice variants were not detected

Whole genome sequencing data suggested a low-confidence candidate deletion in the *IKZF1* DNA binding domain from exons 3 to 6 (*IKZF1Δ3-6*) (Figure 6. Putative deletion between exons 2 and 7 of *IKZF1* (*IKZF1*Δ3-6). Small deletions over the *IKZF1* DNA binding domain often produce null or dominant-negative Ikaros protein isoforms—detectable as dominant-negative transcripts—that drive many human and murine lymphoid malignancies.[115,118,124,125] We did not have supporting RNA-Seq data from this particular tumor; however, we suspected *IKZF1Δ3-6* deletions might have escaped out detection across other tumors. To determine if small deletions in *IKZF1* occurred in additional tumors, we performed exon-specific and isoform-specific expression analysis. Our results show that the relative usage of each exon is consistent between control samples and tumors (Figure 7. Normalized and relative expression of *IKZF1* exons). Dominant negative isoforms were not preferentially expressed in the tumors we measured.

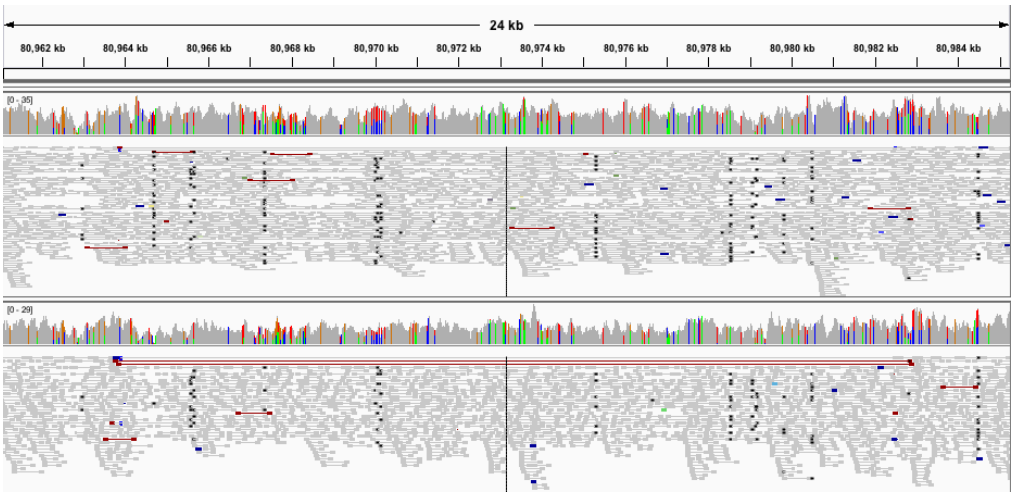



**Figure 6. Putative deletion between exons 2 and 7 of *IKZF1* (*IKZF1*Δ3-6)**: The top and bottom panels represent the matched normal and tumor samples from bird 017788, respectively. Reads are matched as pairs and colored by insert size. The two reads at the top of the bottom panel represent a putative deletion.

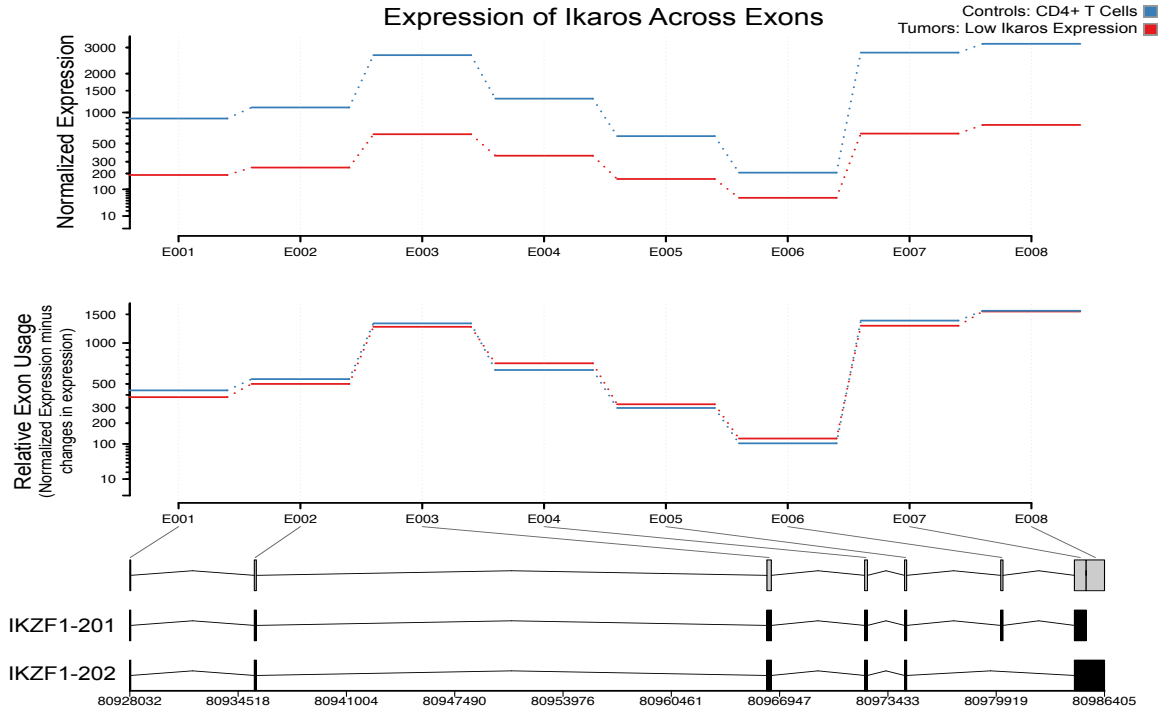



**Figure 7. Normalized and relative expression of *IKZF1* exons**: Comparative *IKZF1* exon expression and usage of Marek's disease low purity tumors (red) versus CD4 T cell control samples (blue). Top: the normalized (regularized log) expression per exon. Middle: the relative exon usage. Bottom: Isoforms of *IKZF1*.

**Marek's disease tumors were primarily monoclonal**

Gross and visible Marek's disease tumors have been suggested to be monoclonal based on MDV integration signatures and TCR profiling[5,9,161]. However, Marek's disease tumor clonality associated with transformation has not been measured in context with genetic drivers. To test if *IKZF1* mutation was present at different parts of tumors, sections from opposite ends of two tumors were investigated. In addition to *IKZF1*, pre-neoplastic markers were investigated in tumors. Mature CD4 T cells, the cells that typically undergo Marek's disease transformation, clonally express single or double T cell receptor (TCR) VBeta (VB) specificity.[162] In mice with germline dominant negative *IKZF1* mutations, the majority of T cells are clonal, highly lymphoproliferative, and demonstrate single or double VB specificity.[110] To determine if one or more predominant neoplastic origin was present in tumors, TCR spectratyping was performed in 13 tumors and 2 biological replicates (6 tumors and one replicate pair with mutant *IKZF1*). RNA-Seq of the TCR VB region was also examined in 10 of the same tumors. In 9 of 10 tumors and in all replicates, TCR VB spectratyping and RNA-Seq agreed; the majority of tumors were monoclonal or biclonal (Figure 8: T cell receptor Vbeta-1 and Vbeta-2 region spectratype signatures). Five of six tumors with *IKZF1* mutations demonstrated single or double clonality.

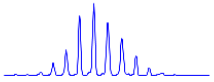
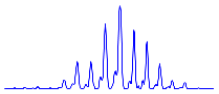
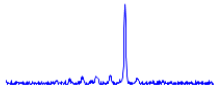
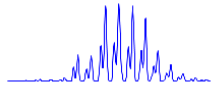
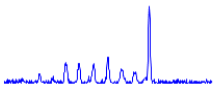
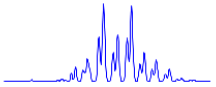
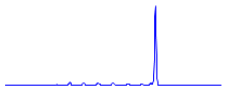
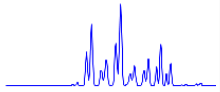
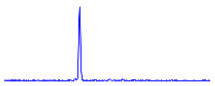
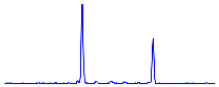
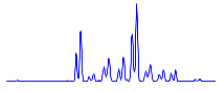
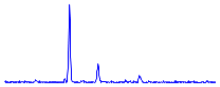
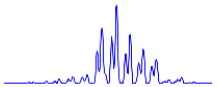
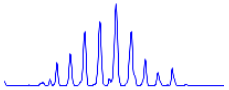
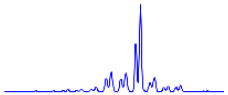
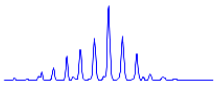
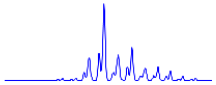
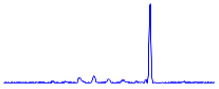
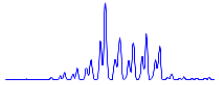
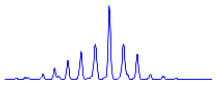
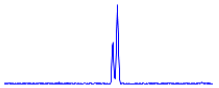
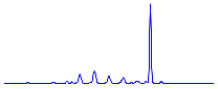
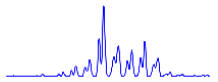
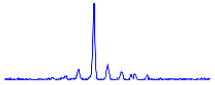
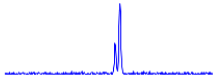
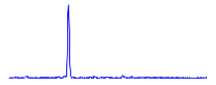
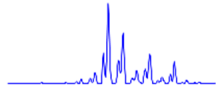
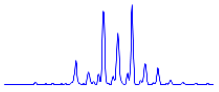
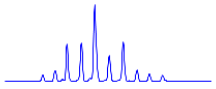
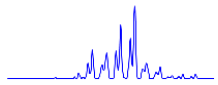
| Tumor | Vbeta-1 Spectratype | Vbeta-2 Spectratype | Tumor | Vbeta-1 Spectratype | Vbeta-2 Spectratype |
|---|---|---|---|---|---|
| Control |  |  | 017834-2 |  |  |
| 017741-1 |  |  | 017835-1* |  |  |
| 017766-1 |  |  | 017842-2_2* |  |  |
| 017777-3* |  |  | 017855-1 |  |  |
| 017787-2 |  |  | 017911-1* |  |  |
| 017798-1 |  |  | 017911-1_2* |  |  |
| 017798-1_2 |  |  | 017918-3* |  |  |
| 017833-1 |  |  | 017927-2* |  |  |

**Figure 8. T cell receptor Vbeta-1 and Vbeta-2 region spectratype signatures**: Biological replicates share the first 7 digits of sample ID listed. Spectratype signatures are not represented to scale. Tumors with *IKZF1* mutations denoted with an asterisk.

**Marek's disease tumor purity was estimated from CD4 expression**

Genetic material was collected from tumor cross sections, which we suspected were contaminated by surrounding tissue. Therefore, we assumed that the 14 tumors in our RNA-Seq analysis contained contaminating RNA from gonadal tissue and infiltrating cells. We noticed a correlation between the expression of mRNA unique to CD4-specific genes (e.g. CD4) and the allelic frequencies of *IKZF1* point mutations; a relationship that allowed us to predict tumor purity from CD4 expression.

The constituents of advanced gonadal tumors consist primarily of transformed CD4 T cells[163] with a minority of B-cells and other immune cell types.[103–105] If *IKZF1* maintained normal copy number and was mutated before tumorigenesis (i.e. a truncal mutation), then *IKZF1* variant allele frequencies would equal approximately one half of tumor purity in monoclonal tumors. *IKZF1* demonstrated the most candidate truncal mutations across tumors and acts as a truncal mutation in subsets of Acute Lymphoblastic Leukemia.[164] We queried normalized gene expression for genes that correlated with allele frequencies of truncal mutations. Genes with the strongest correlations were CD4-specific (CD47, ARHGAP15, and FLI1), especially CD4 itself (Adjusted R-squared: 0.8564, p-value = $6.11 \times 10^{-4}$). We constructed a simple linear model to estimate tumor purity from CD4 expression.

Tumor purity can be approximated by the equation

$y = 9.871e\text{-}05x - 1.152 \times 10^{-1}$,

where y = tumor purity and x = normalized CD4 expression (regularized log).

This relationship is illustrated in Figure 9 by comparing tumor CD4 expression, which we transformed in comparison to normal CD4 expression, to *IKZF1* variant allele frequencies multiplied by 2.

We defined the transformed CD4 expression (CD4adj) for each tumor by the equation

CD4adj = CD4/CD4mu

where CD4 = the expression of CD4 per tumor

and CD4mu = the average of CD4 expression of all normal CD4 T cell samples.

All expression values were normalized to transcripts per million (TPM).

Transformed CD4 expression and *IKZF1* variant allele frequencies (multiplied by two) demonstrated strong correlation (p-value = $1.61\times10^{-3}$; $R^2 = 0.8029$) (Figure 9. Transformed CD4 expression vs. *IKZF1* VAF*2).

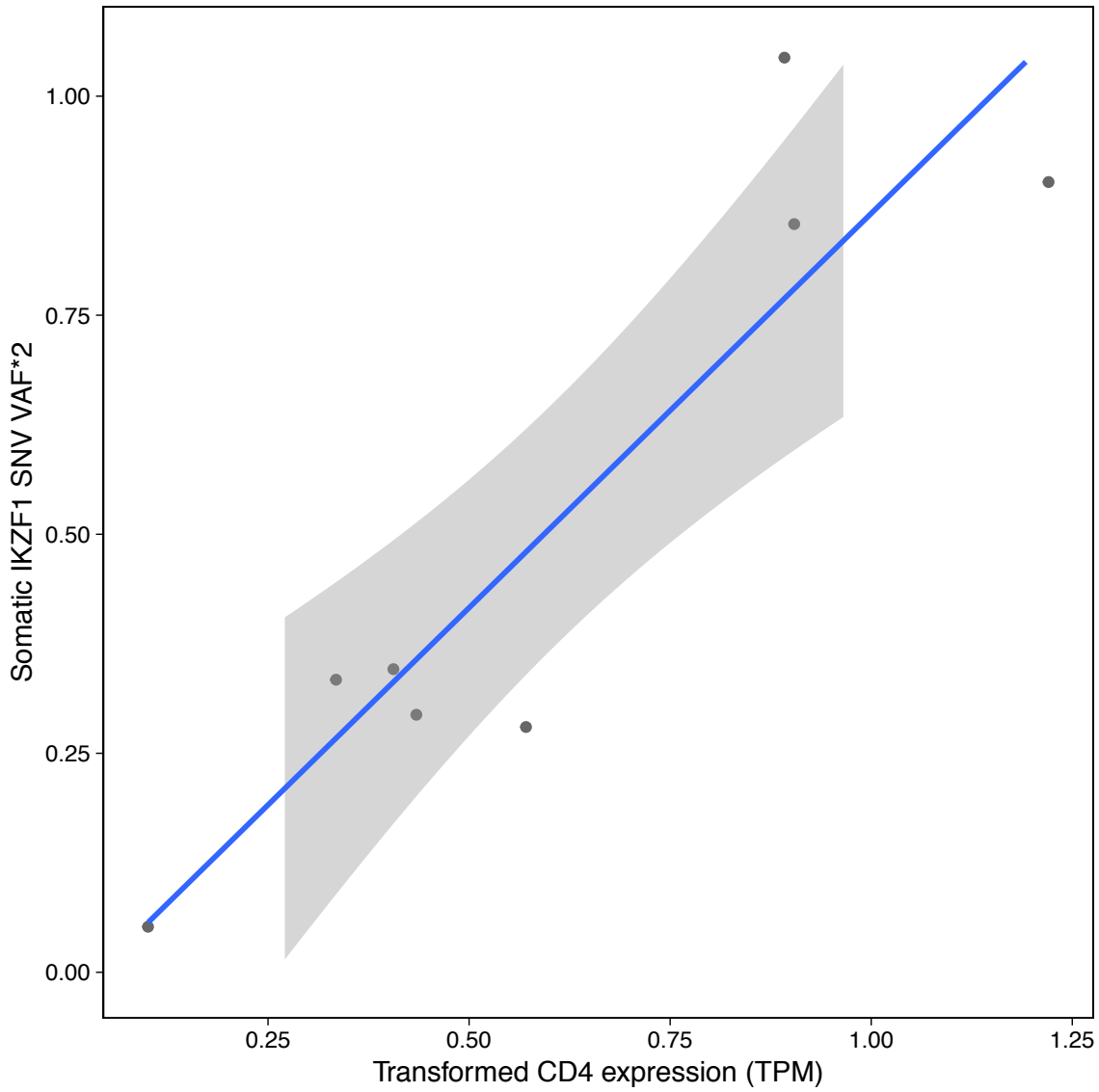



**Figure 9. Transformed CD4 expression vs. *IKZF1* VAF*2:** The linear relationship between tumor purity and CD4-associated genes is evident illustrated with the correlation between the variant allele frequencies (multiplied by 2) of somatic truncal single nucleotide variants in *IKZF1* and the transformed expression of CD4 in predominantly monoclonal tumors (p-value = $1.61\times10^{-3}$; $R^2$ = 0.8029).

***IKZF1* mutant and non-mutant tumors demonstrated similar expression profiles**

Marek's disease lymphomas have not yet been categorized into unique subtypes by their mutational, epigenetic, or expression profiles. Instead, they are often categorized by tissue tropism and the genetic backgrounds of their host and MDV strain. We suspected that *IKZF1* mutant tumors might represent a distinct subtype of Marek's disease lymphomas and

queried their expression profiles for unique characteristics. We were not able to determine if *IKZF1* mutant tumors demonstrated distinct gene expression profiles. However, we observed limited evidence that MDV expression may be influenced by *IKZF1* mutation status.

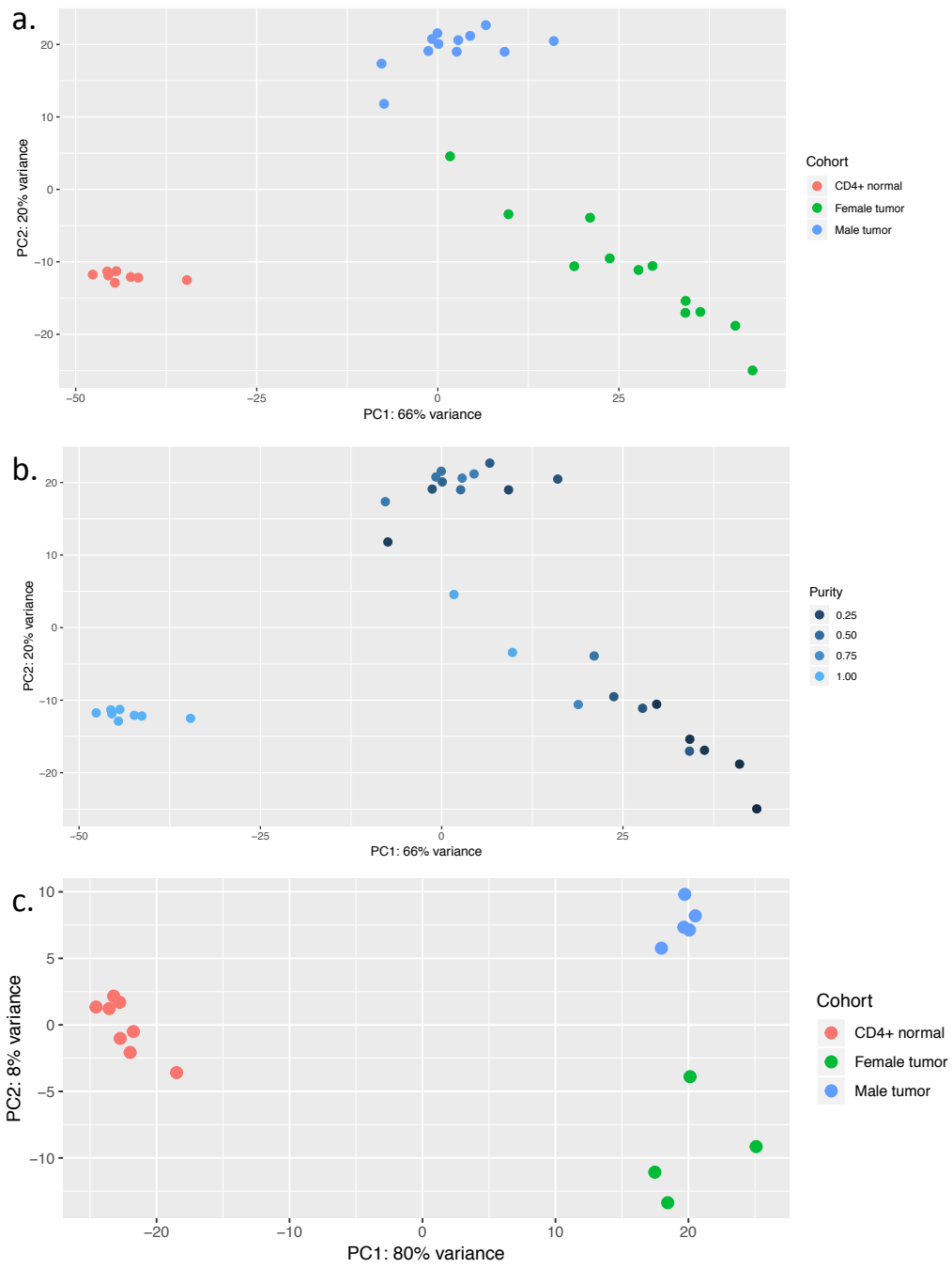



**Figure 10. Principal component analysis of RNA-Seq:** Principal component analysis before (a & b) and after (c) normalization for estimated tumor purity of tumors and removal of tumor samples with estimated purity < 50%. The variance between samples of different tumor purities has reduced, while the variance has increased between the primary dichotomous relationship of interest—Marek's disease tumors versus CD4 normal samples.

We performed a surrogate variable analysis (SVA)[136] on all tumors to test for infer the effects of contaminating gonadal tissue and test for batch effects. Although batch effect was not present, tumors expression profiles demonstrated significant influence from contaminating male and female gonadal tissue. We adjusted for tumor purity (Figure 10. Principal component analysis of RNA-Seq) and performed a supervised hierarchical clustering analysis to assess if expression profiles differed based on *IKZF1* mutation status. Under our adjusted model, the sex of the birds dominated clustering likely resulting from contamination of gonadal tissue rather than *IKZF1* mutation status. Gene expression profiles between *IKZF1* mutant and non-mutant tumors were remarkably similar across and between sexes; no host genes were differentially expressed between *IKZF1* mutant and non-mutant tumors (Figure 11. Clustering of tumor and normal gene expression profiles). We expected that the high dimensionality of gene expression would reveal signatures unique to *IKZF1* mutation status. However, it is possible that *IKZF1* mutant and non-mutant tumors demonstrate similar expression profiles; a phenomenon in cancer exists in which cancer cells converge upon prominent expression profiles across tissue types.[165]

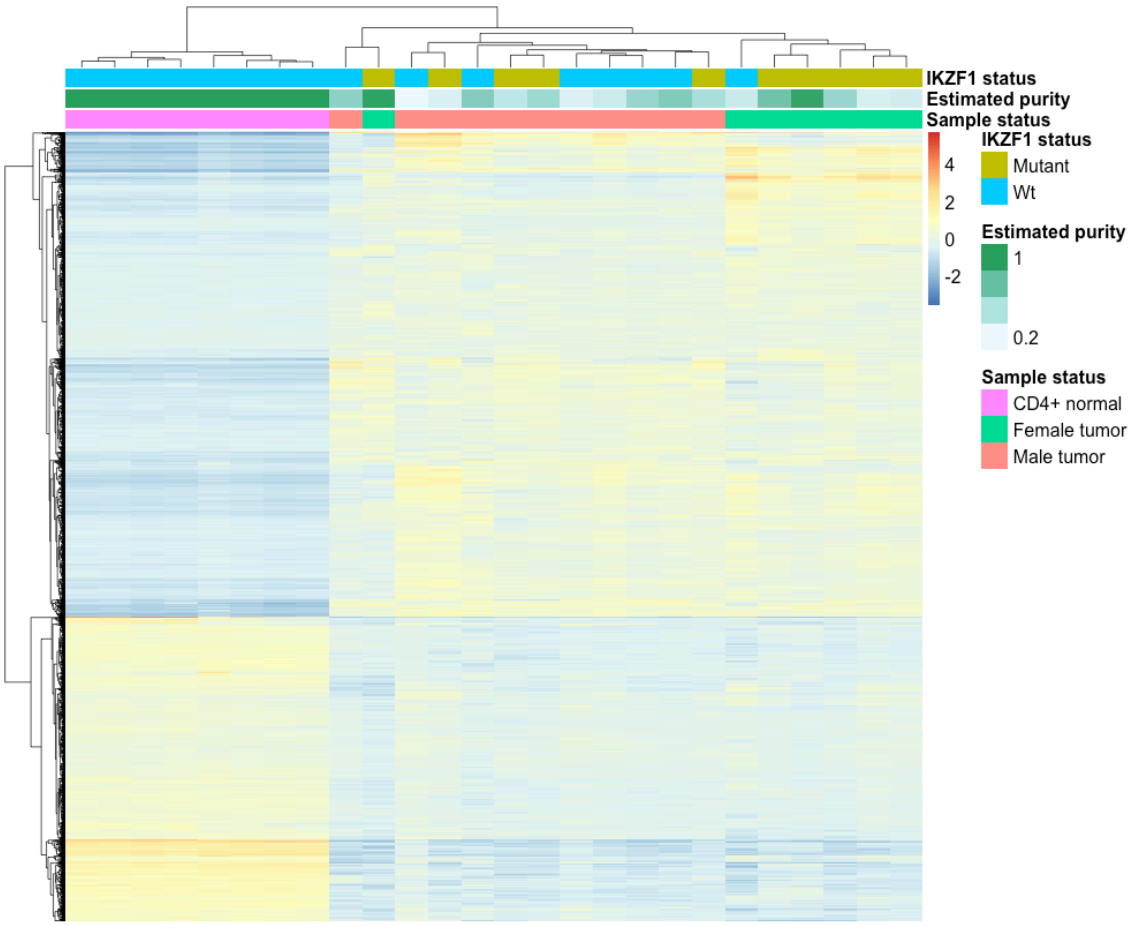



**Figure 11: Clustering of tumor and normal gene expression profiles**
**:** Hierarchical clustering across tumors and adjusted for tumor purity demonstrates that tumor expression profiles do not cluster predominantly by *IKZF1* mutation status. *IKZF1* status—*IKZF1* mutation status; Sample status—whether sample results from a male tumor, female tumor, or CD4 T cells; Estimated purity—estimated tumor purity on a scale of 0.2 to 1. CD4 normal samples were issued a purity of 0.95 arbitrarily because they represent 95% pure untransformed CD4 T cells.

We also compared MDV gene expression between *IKZF1* mutant and non-mutant tumors (estimated purity ≥ 40%). In our adjusted model we noticed significant differences between relative quantities of MDV transcripts; however, we cannot decipher whether MDV transcripts measured resulted from latent or cytolytic expression. In total, 13 MDV genes demonstrated significant difference in transcript quantities (LFC ≥ 0.5, adjusted p-value ≤ 0.1) resulting in 6 and 7 genes with comparatively low and high RNA read counts, respectively (Table

5. Marek's disease virus gene expression and *IKZF1* mutation status) (Figure 12. Marek's disease viral transcripts in Marek's disease tumors). Meq was expressed in all tumors, and in a separate cohort of advanced tumors we detected Meq via immunohistochemistry (Figure 13. Immunohistochemistry reveals Meq presence in gonadal tumors). However, in *IKZF1* mutant tumors both transcripts from both *Meq* loci demonstrated comparatively lower transcripts.

**Table 5. Marek's disease virus expression and *IKZF1* mutation status :** The differential gene expression ($Log_2$ Fold Change) of Marek's disease virus genes between tumors with and without *IKZF1* mutations.

| Gene Name | Gene ID | Log2FoldChange | Adjusted pval |
|---|---|---|---|
| MDV005:oncoprotein_MEQ | 4811550 | -5.210178769 | 2.21533E-06 |
| MDV097:protein_SORF2A | 4811456 | -4.273943055 | 0.005217402 |
| MDV076:oncoprotein_MEQ | 4811549 | -4.237186925 | 0.005733948 |
| MDV073:protein_pp38 | 4811534 | -3.027565428 | 0.000445262 |
| MDV072:protein_LORF5 | 4811533 | -2.460232485 | 0.042078235 |
| MDV010:lipase | 4811470 | -1.417083589 | 0.099367982 |
| MDV095:envelope_glycoprotein_I | 4811454 | 1.923555602 | 0.062440248 |
| MDV055:DNA_polymerase_processivity_subunit | 4811516 | 2.056389333 | 0.099367982 |
| MDV046:DNA_packaging_protein_UL32 | 4811506 | 2.234421506 | 0.062440248 |
| MDV050:tegument_protein_UL37 | 4811511 | 2.948403582 | 0.089389286 |
| MDV066:helicase-primase_primase_subunit | 4811527 | 3.298147945 | 0.071937315 |
| MDV061:transactivating_tegument_protein_VP16 | 4811522 | 3.598073831 | 0.004830689 |
| MDV051:capsid_triplex_subunit_1 | 4811512 | 4.377172609 | 0.007252397 |

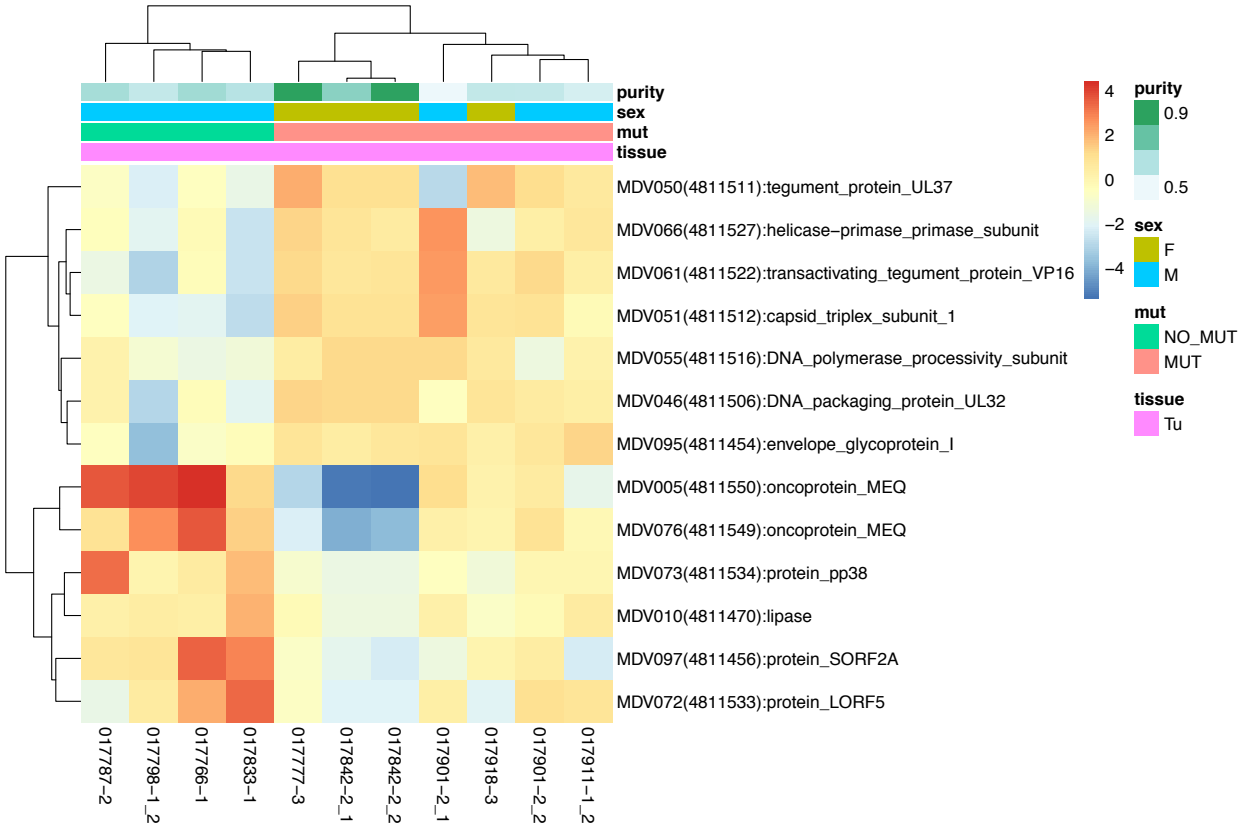



**Figure 12. Marek's disease viral transcripts in Marek's disease tumors:** Normalized expression (regularized log) of expressed genes in the MDV genome between tumors with and without mutations in *IKZF1*. mut—*IKZF1* mutation status; sex—the sex of host bird; purity—estimated tumor purity; and tissue—tumor tissue.

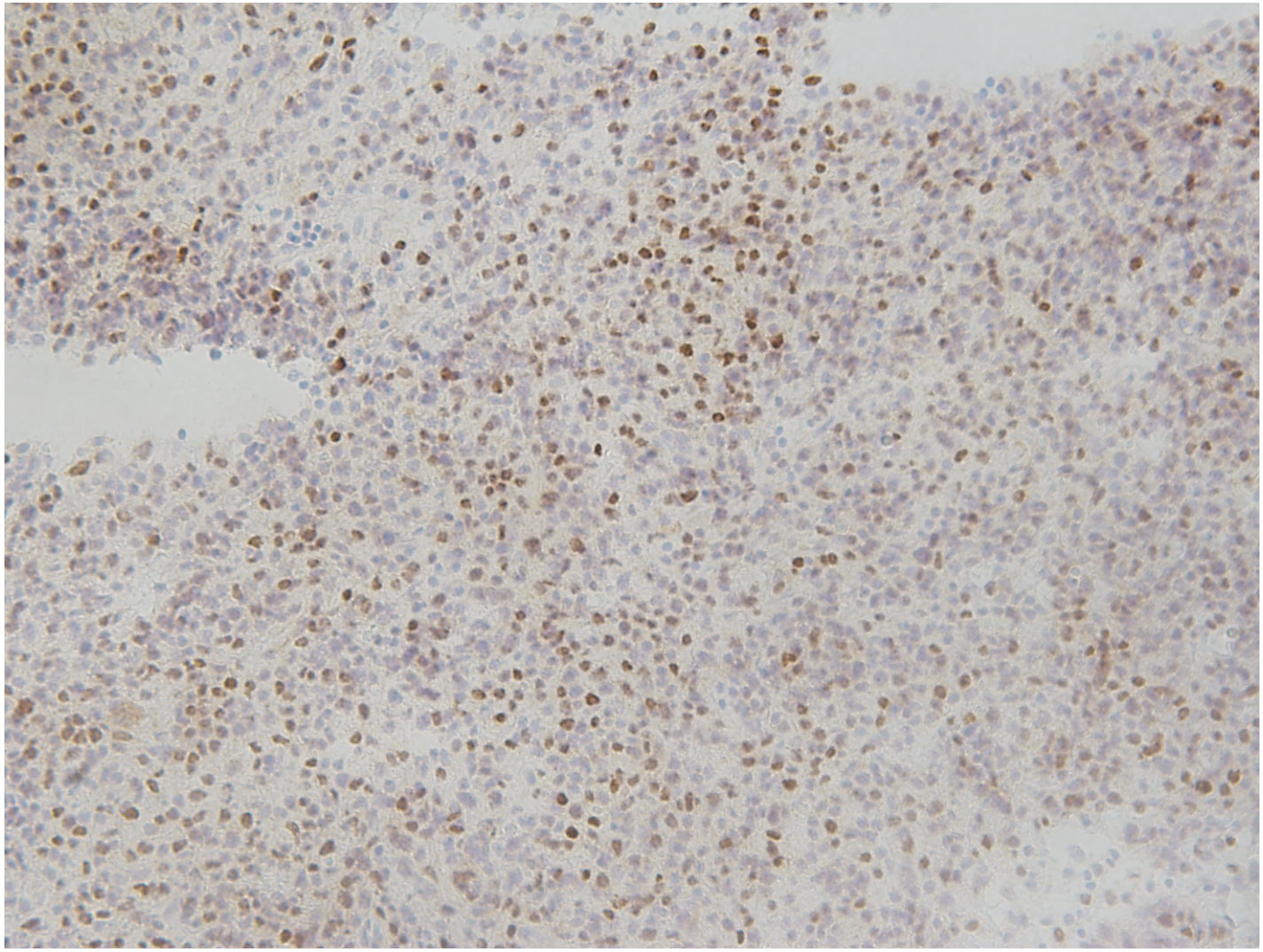


**Figure 13. Immunohistochemistry reveals Meq presence in gonadal tumors**: A late-stage Marek's disease gonadal tumor (*IKZF1*-mutant) cross-section illustrating the presence of Meq in neoplastic cell nuclei bound by Meq-specific antibody (brown stain).

**Differentially expressed genes associated with non-coding point mutations**

The expression profiles of tumor samples of sufficient purity (53 - 95%) were compared against profiles of 8 replicates of CD4 T cell isolated from uninfected birds. In total, 12,518 genes were queried for differential expression (LFC ≥ 0.5, adjusted p-value ≤ 0.05 using DESeq2) resulting in 1,253 significantly up regulated genes and 1,032 significantly down regulated genes. (Figure14. Top 50 differentially expressed genes in Marek's disease lymphomas).

As expected the vast majority of somatic point mutations in and around genes were non-coding, with the majority of variants residing in intronic regions of large genes. In total, 131 differentially expressed genes contained or were proximal to somatic non-coding mutations. We performed a logistic regression analysis to assess the association between genes proximal to non-coding mutations (intronic, upstream, and downstream) and differentially expressed genes. The odds that genes with non-coding mutations were differentially expressed were 1.697 times the odds of genes without non-coding mutations (95% CI: 1.43-2.01; p-value = 2.86e-07), suggesting that these often overlooked mutation types (non-coding mutations) influenced differential gene expression.

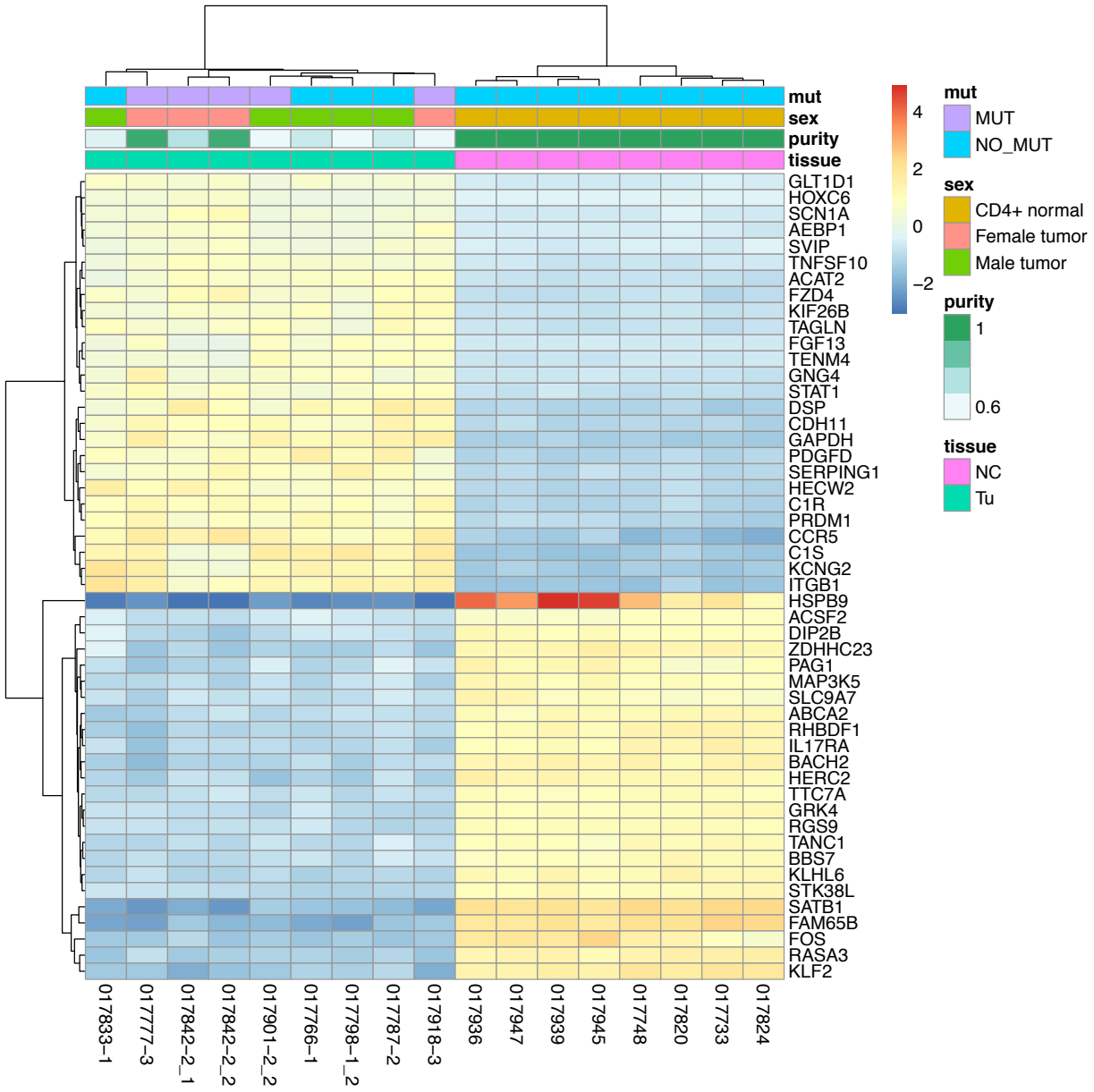



**Figure 14. Top 50 differentially expressed genes in Marek's disease lymphomas**: The normalized expression of the top The 50 most differentially expressed genes orthologous to human. Annotations: mut—mutational status; sex—bird status (male tumor, female tumor, or normal CD4 T cell sample); purity—estimated tumor purity; and tissue—status of sample (tumor or normal).

**Gene expression profiles of Marek's disease lymphomas**

In hematopoietic cells, Ikaros acts as a master regulator of gene expression and chromatin remodeling.[166] Reduction or even complete loss of Ikaros regulation has been shown to greatly alter gene expression resulting in cells with high-risk-gene expression signatures leading to lymphoma and leukemia.[167–173] Ikaros is essential for the highly permissive chromatin environment it regulates, and its function is so critical that it cannot be compensated for by other transcription factors.[174] Thus, the consequences of Ikaros loss impair tumors with specific hallmarks, and we believe that evidence for these hallmarks are expressed from Marek's disease tumor transcription profiles.

Six of nine tumors that underwent differential gene expression analysis demonstrated mutations in *IKZF1* and we anticipated that *IKZF1* mutations influenced tumor expression profiles. We queried a gene set of Ikaros targets—mouse leukemic T cells genetically deficient of *IKZF1* and exposed to retroviral induction of the Ikaros1 isoform[151]—and noticed enrichment for Ikaros targets in differentially expressed genes from Marek's disease tumors. Considering genes with basal expression (at least ten total reads per sample), differentially expressed genes were 34% ± 14% more likely to be Ikaros targets than genes not differentially expressed (p-value = $8.34 \times 10^{-6}$, 95% CI by logistic regression).

The list of differentially expressed genes between Marek's disease tumors and normal CD4 T cell samples was queried for pathway enrichment[141–143] and MSigDB gene sets[149,150] to infer phenotypic changes in Marek's disease tumors that differed from CD4 T cells. Furthermore, the list of 302 differentially expressed genes that also showed enrichment for putative Ikaros1 regulation was queried to infer the molecular mechanisms. Enriched terms and pathways show

striking similarities to those in leukemia with loss of Ikaros[174–177] demonstrating increased cellular adhesion, increased extracellular matrix organization, increased cell surface receptor signaling, increased cell migration and motility, decreased T cell receptor signaling, T cell dedifferentiation, and a neural stem cell-like phenotypes. A similar scenario has been described in BCR-ABL1$^{+}$ pre-B ALL cells with loss of Ikaros, which demonstrate a “neuro-epithelial” gene signature and epithelial cell-like phenotype with aberrant expression of genes normally expressed in neuro-epithelial cells.[174–177] Several cellular and molecular processes in pre-B ALL cells with loss of Ikaros are indicative of the enriched pathways and ontologies in Marek’s disease tumors including increased adhesion-mediated receptor signaling, actin filament-based processes, axon guidance, and integrin binding (Figure 15. Enriched pathways from differential gene expression). The 50 most differentially expressed genes in enriched pathways and enriched for Ikaros regulation demonstrate genes (potentially) driving these mechanisms (Figure 16. Differentially expressed genes in enriched pathways and Ikaros1 targets).

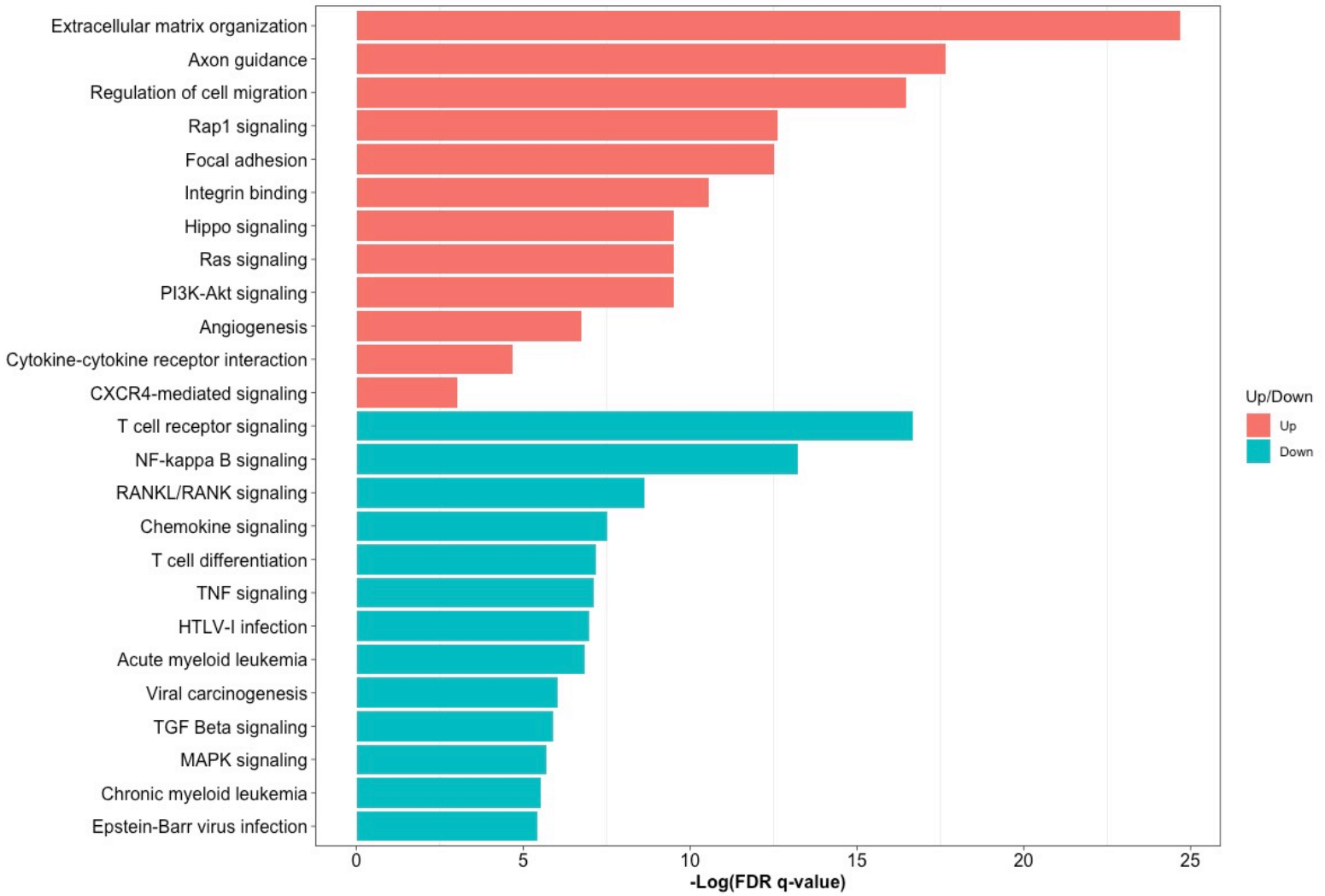



**Figure 15. Enriched pathways from differential gene expression:** Enriched pathways for differentially expressed genes between Marek's disease tumors and normal isolated CD4 T cell samples. Pathway enrichment was performed with gene ontologies[139,140] and pathways (e.g. Kegg and Reactome).[141–148] Enrichment was performed with Enrichr[152,153] and gprofiler.[154]

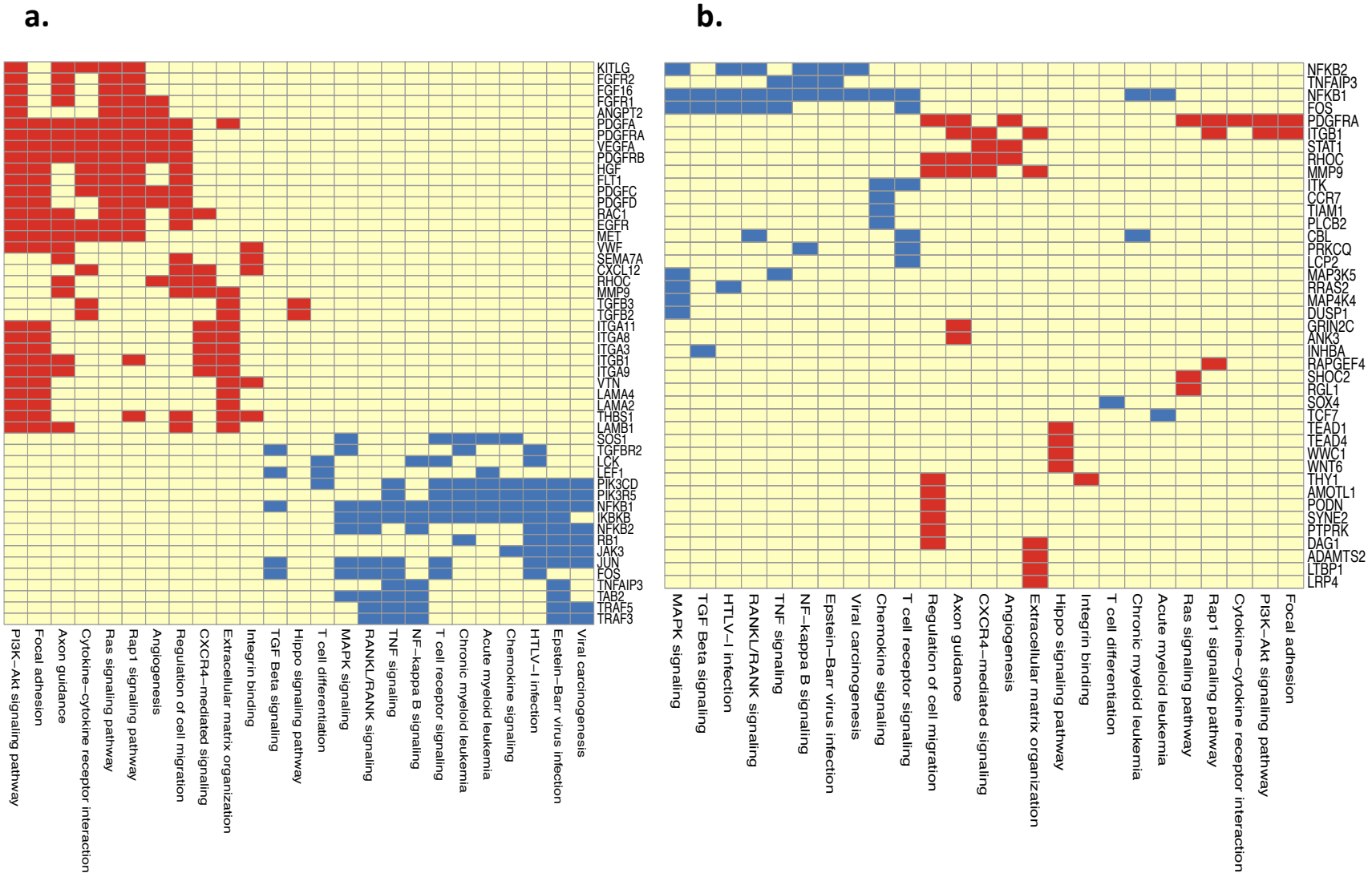



**Figure 16. Differentially expressed genes in enriched pathways and Ikaros1 targets:** The 50 most significantly differentially expressed genes in enriched pathways (a) and transcriptionally regulated by Ikaros[151] (b). Red and blue represent up and down regulation of gene expression, respectively.

Dedifferentiation is also a common theme shared between Ikaros Marek's disease tumors and pre-B ALL cells with a dominant negative Ikaros allele. In mature CD4 T cells, Ikaros mediates its tumor suppressor ability by repressing expression of lineage inappropriate genes, especially those found in hematopoietic stem cells and neural development pathways.[178] Loss of Ikaros has been shown to drive cells to dedifferentiate coupled with the acquisition of more stem-like and epithelial cell characteristics.[174,176] We noticed evidence of similar expression profiles in Marek's disease tumors; T cell receptor signaling and T cell differentiation were among the most significantly enriched pathways for down regulation of gene expression. Furthermore, the gene expression profiles of Marek's disease tumors were significantly enriched for gene sets up regulated in neural crest stem cells (Figure 17. Tumor expression profiles enriched for T cell dedifferentiation and stemness).[179] Among the most significantly up regulated genes was CDK6—a transcriptional target of Ikaros[151] and a master regulator of leukemic stem cell (LSC) activation.[180] CDK6 is frequently up regulated in tumors of hematopoietic origin including AML and ALL.[181–183] Specifically, CDK6 suppresses Egr1, which was also down regulated in Marek's disease tumors, allowing for LSC activation.[180]

Marek's disease tumor expression profiles reveal striking similarities to other well-documented cancers that lack Ikaros tumor suppressor ability. We strongly suspect that mutations in the DNA-binding domain of *IKZF1* drive Marek's disease tumors.

**a.** **b.**

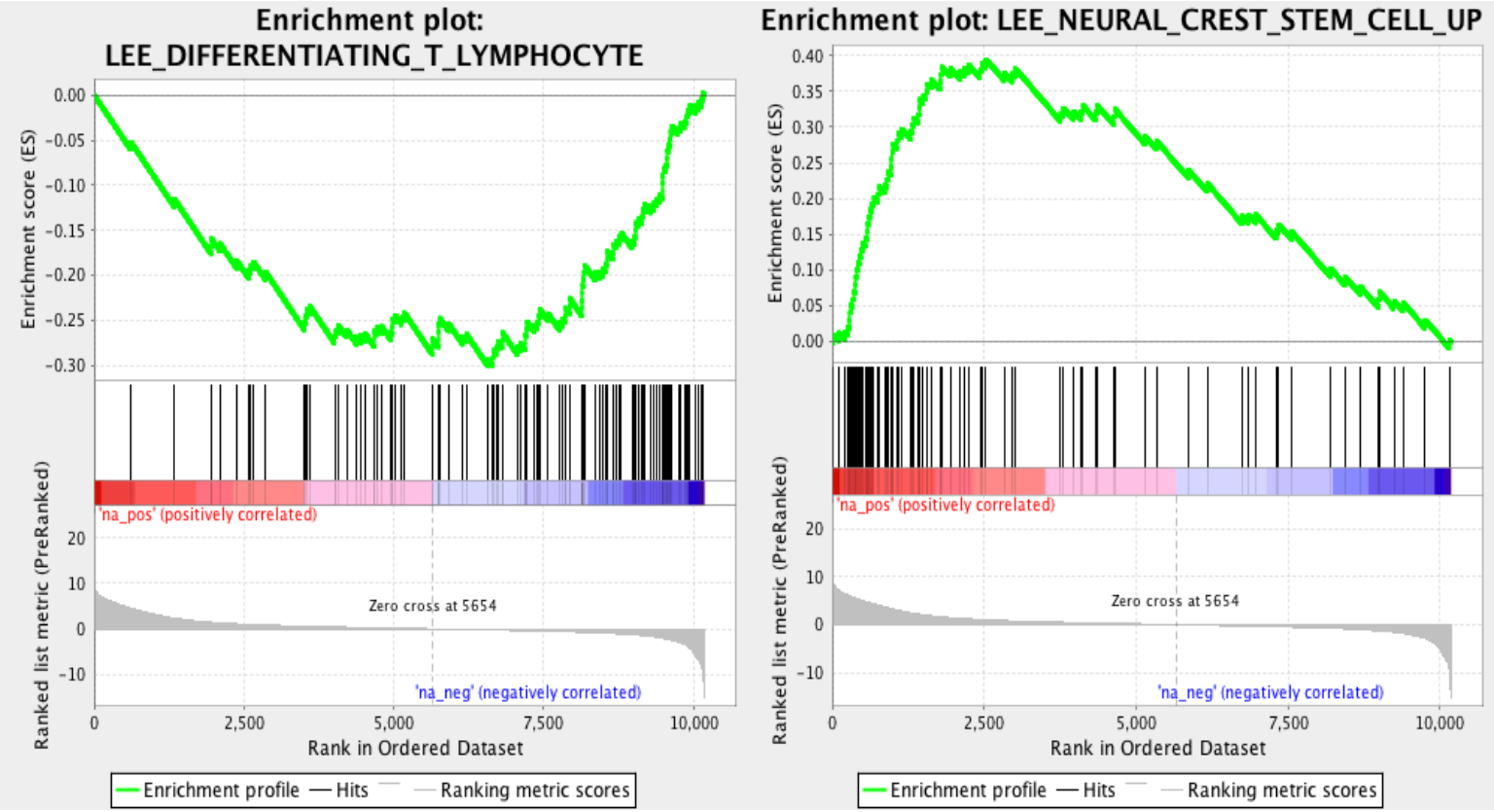



**Figure 17. Tumor expression profiles enriched for T cell dedifferentiation and stemness:** Gene set enrichment analysis (GSEA) of differentially expressed genes between Marek's disease tumors and normal CD4 T cell samples. Differentially down regulated genes in Marek's disease tumors (blue) are enriched for genes associated with T cell differentiation[184] (a). Normalized enrichment score: -3.72, nominal p-value: $< 1x10^{-5}$, FDR q-value: $< 1x10^{-5}$. Differentially up regulated genes in Marek's disease tumors (red) are enriched for genes associated with neural crest stem cells[185] (b). Normalized enrichment score: 4.237, nominal p-value: $< 1x10^{-5}$, FDR q-value: $< 1x10^{-5}$.

**SUMMARY**

Our observations expand upon the current model of Marek's disease lymphomagenesis. We observed that (1) MDV integration is necessary and must precede clonal expansion; and (2) Meq activity and expression is also necessary. We expand upon this model by comprehensively examining the somatic landscapes of Marek's tumors and reveal the first driver gene of Marek's disease lymphomas, *IKZF1*.

We demonstrated that the DNA-binding domain of *IKZF1* was frequently mutated in the largest gonadal tumors in our model. Mutations cluster across essential residues in the critical N-terminal zinc fingers of *IKZF1*, which is conserved across species. All considerations from the literature and bioinformatic algorithms suggest that mutated *IKZF1* is a driver gene and that somatic mutations are deleterious to the DNA-binding capacity of the mutated allele. Mutation of *IKZF1* is a driving event in Marek's disease lymphomas.

The relationship between mutant *IKZF1* and Meq has not been thoroughly examined, but we suspect that they work in tandem. These considerations require mutational analysis of *IKZF1* and Ikaros in chicken CD4 T cells in the context of MDV integration and Meq. The advent of CRISPR would allow for the precise functional analysis of the somatic *IKZF1* mutations in the Marek's disease context. Furthermore, the exact series of events surrounding driver acquisition could be investigated. We suspect that mutations in *IKZF1* are heterozygous and occur prior to neoplastic clonal expansion because *IKZF1* variants are consistently the dominant variant allele in Marek's tumors.

Pediatric malignancies in human are suggested to require at least two drivers and that initial driver mutations often occur spontaneously during development.[186] Bloodspots taken at birth from patients with acute lymphoblastic leukemia (ALL) demonstrated that the initial

mutation usually occurred in utero.[187] Similarly, blood recovered at hatch and prior to MDV infection for mutations in *IKZF1* could be queried to determine if *IKZF1* mutations spontaneously occur in ovo. On the other hand, mutations in *IKZF1* could arise from an MDV-induced inflammatory response. In ALL, mutations in *IKZF1* are consistently the secondary driver[164] and their acquisition have been associated with early viral infections.[120,188] Whether initiating or acquired, the etiology of *IKZF1* mutations and the mechanisms leading to mutagenesis may shed light on the mutational process and, by extension, the discrepancy of tumor incidence between Marek's disease resistant and susceptible birds.

Although we detected mutated *IKZF1*, it is plausible that additional drivers exist. We are confident that we successfully measured the majority of somatic variants—somatic SNVs and Indels—with adequate power and confirmed their presence and relative frequency, about 0.3 somatic SNVs and Indels per megabase. However, the average genomic coverage (14x) and tumor purity (45%) suggested that we queried early tumors with suboptimal power to comprehensively catalog complex somatic variants such as structural variants. However, whole genome sequencing suggested an additional deletion in the *IKZF1* DNA-binding domain; targeted sequencing demonstrated non-synonymous variants in additional genes, such as *FLT3*; multiple technologies suggested deletions and copy number loss in the *SOX1* and *VWF* regions of chromosome 1; and RNA-Seq generated candidate read-through fusion events in *ROBO1* and *ROBO2*. These candidates demonstrate sufficient evidence to warrant further bioinformatic interrogation and targeted sequencing to determine legitimacy.

*IKZF1* mutant Marek's disease tumors revealed gene expression profiles indicative of Ikaros perturbation including cellular adhesion, cellular migration, extracellular matrix

organization, axon guidance, lymphocyte dedifferentiation, and a stem-cell state. These characteristics are suggested through expression profiles; however, functional assays are required to determine if the *IKZF1* mutant cells legitimately exhibit these phenotypes. Furthermore, the targets of Ikaros transcriptional-regulation and accompanied motifs should be considered with chromatin immunoprecipitation sequencing and complimentary RNA-Seq. Although analyses of Marek's disease tumor transcriptomes may demonstrate enrichment of differentially expressed putative Ikaros targets, according to the literature, we may only speculate on the potential reprogramming to the Marek's disease lymphoma transcriptome by mutant Ikaros.

The investigations presented in this thesis resulted in a few robust claims and many more speculations. These experiments were hypothesis generating. We hope to inspire the Marek's disease community to further investigate Marek's disease lymphomagenesis and determine its major drivers and mutational mechanisms. These insights might aid in the development of better vaccines, genetic resistance, creative control strategies, and our knowledge of cancer biology.

**APPENDIX**

**Table S1. Somatic non-synonymous SNVs and Indels:** The collection of predicted somatic non-synonymous SNVs and Indels across 22 Marek's disease tumors.

| CHROM | POS | REF | ALT | VAR_TYPE | CON | TOOLS | SYMBOL | SAMPLE |
|---|---|---|---|---|---|---|---|---|
| 2 | 80972116 | C | T | SNV | missense_variant | 4 | IKZF1 | 901-2_2_S26, 901-2_S22 |
| 2 | 80972101 | C | T | SNV | missense_variant | 6 | IKZF1 | 777-3_S14, 901-2_S22 |
| 2 | 80972152 | A | AACT | INDEL | inframe_insertion | 4 | IKZF1 | 918-3_S10 |
| 2 | 80972118 | C | G | SNV | missense_variant | 4 | IKZF1 | 756-3_S3 |
| 2 | 80972104 | C | T | SNV | missense_variant | 5 | IKZF1 | 842-2_S20 |
| 1 | 179279863 | C | G | SNV | missense_variant | 5 | PARP4 | 842-2_S20 |
| 1 | 179279863 | C | G | SNV | missense_variant | 4 | PARP4 | 842-2_2_S25 |
| 12 | 7233917 | A | G | SNV | missense_variant | 4 | CHDH | 911-1_S24 |
| 12 | 7233915 | C | T | SNV | missense_variant | 4 | CHDH | 911-1_S24 |
| 1 | 176041476 | T | A | SNV | missense_variant | 5 | FLT3 | 794-1_S17 |
| 26 | 2618494 | C | CCGGCCG | INDEL | inframe_insertion | 4 | PFKFB2 | 798-1_S5 |
| 7 | 13771765 | T | G | SNV | missense_variant | 5 | NBEAL1 | 842-2_2_S25, 842-2_S20 |
| 4 | 81134701 | C | T | SNV | missense_variant | 6 | AFAP1 | 766-1_S4 |
| 3 | 34836601 | A | AT | INDEL | frameshift_variant | 4 | DESI2 | 834-2_2_S12 |
| 19 | 9001333 | G | A | SNV | missense_variant | 6 | AKAP1 | 901-2_2_S26, 901-2_S22 |
| 17 | 6421696 | TCTC | T | INDEL | inframe_deletion | 4 | PRDM12 | 834-2_2_S12 |
| 14 | 14979283 | T | C | SNV | missense_variant | 5 | ZP2 | 842-2_S20 |
| 3 | 48988967 | A | C | SNV | missense_variant | 5 | AKAP12 | 884-2_S21 |
| 26 | 2327458 | C | T | SNV | missense_variant | 6 | SLC26A9 | 833-1_S6 |
| 20 | 3990397 | G | A | SNV | missense_variant | 5 | DHX35 | 911-1_S24 |
| 2 | 104762631 | G | T | SNV | missense_variant | 4 | CHST9 | 841-3_S19 |
| 17 | 1129731 | C | T | SNV | missense_variant | 6 | AMBP | 766-1_S4 |
| 15 | 11190293 | A | T | SNV | missense_variant | 4 | NEFH | 927-2_S11 |
| 14 | 12504866 | C | A | SNV | stop_gained | 4 | SLX4 | 833-1_S6 |
| 12 | 11280899 | G | A | SNV | missense_variant | 5 | CCDC174 | 787-2_S15 |
| 1 | 49800139 | C | T | SNV | missense_variant | 6 | L3MBTL2 | 766-1_S4 |

**Table S2. Somatic structural variant candidates:** The collection of candidate somatic structural variants across 22 Marek's disease tumors.

| Chrom | Start | End | Type | Sample | Gene |
|---|---|---|---|---|---|
| chr1 | 24295614 | 60369698 | DUP | 017834-2 | PAH, TPH2, TEAD4, PP |
| chr1 | 52841831 | 52855913 | INV | 017788-1 | LARGE1 |
| chr1 | 53345388 | 53403253 | INV | 017788-1 | PAH, TPH2, TEAD4, PP |
| chr1 | 55360406 | 56151630 | INV | 017906-1 | IGF1, PARPBP, CHPT1, |
| chr1 | 55361021 | 56151506 | INV | 017918-3 | IGF1, PARPBP, CHPT1, |
| chr1 | 59617071 | 59978532 | DEL | 017756-3 | SINHCAF |
| chr1 | 63336212 | 63341684 | INV | 017777-3 | PAH, TPH2, TCP11L2, |
| chr1 | 77068345 | 77156789 | INV | 017756-3 | ING4, COPS7A, ZNF384 |
| chr1 | 77068527 | 77156914 | INV | 017842-2_2 | ING4, COPS7A, ZNF384 |
| chr1 | 77068652 | 77156788 | INV | 017901-2 | ING4, COPS7A, ZNF384 |
| chr1 | 79057362 | 79106476 | DEL | 017901-2_2 | ZYX, EPHB6, CD47, CC |
| chr1 | 82269182 | 82378762 | DEL | 017756-3 | GAP43 |
| chr1 | 102542273 | 102543130 | INV | 017906-1 | JAM2 |
| chr1 | 104554587 | 105212111 | INV | 017834-2 | CFAP298, SCAF4, OLIG |
| chr1 | 113131681 | 116830119 | INV | 017842-2_2 | XK, CYBB, DYNLT3, PR |
| chr1 | 122573459 | 123166687 | DEL | 017906-1 | TRAPPC2B, RAB9A, GPM |
| chr1 | 130103092 | 130118467 | INV | 017911-1 | EDNRB, PAH, TPH2, TE |
| chr1 | 140048249 | 140146532 | INV | 017901-2 | TNFSF13B |
| chr1 | 153077883 | 156257327 | DEL | 017855-1 | EDNRB, PAH, TPH2, TE |
| chr1 | 154413747 | 154423205 | INV | 017835-1 | EDNRB, PAH, TPH2, TE |
| chr1 | 164600039 | 164649334 | INV | 017756-3 | ELF1, SUGT1, OLFM4, |
| chr1 | 172050175 | 172050636 | INV | 017834-2_2 | EDNRB, TEAD4, PPARA, |
| chr1 | 178953267 | 178997740 | INV | 017901-2 | GJB2 |
| chr1 | 181040723 | 181100010 | INV | 017834-2 | EDNRB, TEAD4, PPARA, |
| chr1 | 182005407 | 182006079 | DEL | 017794-1 | PDGFD |
| chr1 | 182005458 | 182006071 | DEL | 017787-2 | PDGFD |
| chr1 | 186765604 | 186769576 | DEL | 017766-1 | FAT3 |
| chr2 | 361456 | 455848 | INV | 017884-2 | SSPO |
| chr2 | 26652916 | 27567220 | INV | 017842-2_2 | SCIN, ETV1 |
| chr2 | 42037543 | 42060179 | INV | 017911-1 | MRPL3, POMGNT2, SH3B |
| chr2 | 42037550 | 42060175 | INV | 017911-1 | MRPL3 |
| chr2 | 55293219 | 55330719 | INV | 017901-2_2 | TNS3 |
| chr2 | 66577991 | 67441089 | INV | 017911-1 | EXOC2, IRF4 |
| chr2 | 67884994 | 68793866 | INV | 017777-3 | SERPINB8, VPS4B, SER |
| chr2 | 67967886 | 67981148 | DEL | 017842-2 | SERPINB4 |
| chr2 | 105178187 | 106193560 | INV | 017777-3 | CDH2 |

**Table S2** (cont'd)

| Chrom | Start | End | Type | Sample | Gene |
|---|---|---|---|---|---|
| chr2 | 115789870 | 116358256 | DEL | 017901-2 | CPA6 |
| chr2 | 115790070 | 116358249 | DEL | 017901-2_2 | CPA6 |
| chr2 | 118357501 | 118358188 | INV | 017777-3 | STAU2 |
| chr2 | 121381554 | 121954081 | INV | 017756-3 | FABP5, TPD52 |
| chr2 | 127561287 | 127599604 | INV | 017901-2_2 | SDC2, RPL30, AZIN1, |
| chr3 | 4359738 | 4454621 | INV | 017766-1 | PTK7, SRF |
| chr3 | 6984643 | 7680185 | DEL | 017794-1 | NRXN1, STON1, GTF2A1 |
| chr3 | 7280868 | 7917090 | DEL | 017855-1 | EHD3, STON1, GTF2A1L |
| chr3 | 7917364 | 8345322 | INV | 017756-3 | YPEL5, LBH, LCLAT1 |
| chr3 | 60000033 | 60000390 | DEL | 017863-1 | NCOA7 |
| chr4 | 20178992 | 21968489 | INV | 017756-3 | GRIA2, NPY2R, CTSO, |
| chr4 | 46338551 | 46338722 | DEL | 017842-2 | MAPK10 |
| chr6 | 545588 | 9246463 | DEL | 017911-1_2 | CDHR1, SIRT1, NCOA4, |
| chr6 | 15082600 | 15734003 | INV | 017863-1 | ADK, AP3M1, VCL |
| chr6 | 19447558 | 19945624 | INV | 017901-2 | HTR7, PCGF5, TNKS2, |
| chr6 | 19479871 | 19954026 | INV | 017787-2 | HTR7, PCGF5, TNKS2, |
| chr11 | 14429252 | 14429681 | DEL | 017911-1_2 | WWOX |
| chr13 | 2676548 | 2995943 | INV | 017927-2 | HSPA9, FBXW11 |
| chr13 | 2677477 | 2995789 | INV | 017911-1_2 | HSPA9, FBXW11 |
| chr13 | 15938740 | 16142754 | INV | 017855-1 | PPP2CA, SKP1, C5orf1 |
| chr17 | 793963 | 1099834 | INV | 017738-1 | EDF1, FBXW5, PHPT1, |
| chr17 | 793973 | 1099796 | INV | 017884-2 | EDF1, FBXW5, PHPT1, |
| chr18 | 633069 | 668600 | DEL | 017901-2_2 | MYH2 |
| chr18 | 755781 | 1270831 | INV | 017927-2 | MYOCD |
| chr18 | 842373 | 1291110 | DEL | 017834-2_2 | MYOCD |
| chr19 | 4920321 | 4933412 | DEL | 017901-2_2 | ASL |
| chr20 | 835908 | 1215863 | INV | 017901-2_2 | NFS1, CPNE1, ROMO1, |
| chr20 | 836158 | 1215870 | INV | 017901-2 | NFS1, CPNE1, ROMO1, |
| chr28 | 87420 | 524183 | INV | 017911-1_2 | TIMM44 |
| chr28 | 87423 | 524195 | INV | 017766-1 | TIMM44 |
| chr28 | 87424 | 524093 | INV | 017835-1 | TIMM44 |
| chrZ | 21516654 | 22254222 | DEL | 017794-1 | SLC30A5, MTX3, CENPH |
| chrZ | 44670399 | 44883474 | INV | 017901-2 | FANCC, AUH, SYK, NAA |
| chrZ | 44670749 | 44884848 | INV | 017901-2_2 | CDC42SE2 |

**Table S3. Somatic copy number variant candidates:** The collection of candidate somatic copy number alterations across 22 Marek's disease tumors.

| Chrom | Start | End | Type | Sample | Gene |
|---|---|---|---|---|---|
| chr1 | 74374001 | 74414000 | loss | 017835-1 | VWF |
| chr1 | 83168001 | 83196000 | loss | 017756-3 | CLDND1 |
| chr1 | 83168001 | 83218000 | loss | 017906-1 | CLDND1 |
| chr1 | 83168001 | 83265098 | loss | 017835-1 | CLDND1 |
| chr1 | 83170001 | 83194000 | loss | 017863-1 | CLDND1 |
| chr1 | 139808080 | 139825940 | loss | 017842-2 | SOX1 |
| chr2 | 49280001 | 49345605 | loss | 017906-1 | AMPH |
| chr2 | 142699457 | 142842000 | loss | 017835-1 | MIR30D, MIR30B |
| chr2 | 142752001 | 143016000 | loss | 017906-1 | MIR30D, MIR30B |
| chr3 | 2364939 | 2601803 | loss | 017835-1 | RTN4, XPO1, RTN4 |
| chr3 | 2444001 | 2524000 | loss | 017906-1 | RTN4 |
| chr3 | 5662001 | 5692000 | loss | 017835-1 | OTOR |
| chr3 | 5674001 | 5756000 | loss | 017756-3 | OTOR |
| chr3 | 12222001 | 12278709 | loss | 017906-1 | CDC42BPA |
| chr3 | 108778001 | 108808000 | loss | 017906-1 | TFAP2D |
| chr11 | 3354001 | 3390116 | loss | 017835-1 | ESRP2 |
| chr17 | 10533828 | 10551688 | loss | 017842-2 | LMX1B |
| chr17 | 10838202 | 10943916 | gain | 017798-1 | FBXW2, GARNL3, SLC2A |
| chr17 | 10913448 | 10944219 | gain | 017834-2 | GARNL3 |
| chrZ | 48048264 | 48613911 | neutral | 017794-1 | EFNA5 |
| chrZ | 55146410 | 56245470 | neutral | 017756-3 | PHAX, LMNB1 |
| chr15 | 12632355 | 12746000 | loss | 017906-1 | SDSL, LHX5 |
| chr15 | 12636120 | 12746000 | loss | 017756-3 | SDSL, LHX5 |
| chr16 | 6001 | 254000 | loss | 017835-1 | RACK1, TRIM41, BRD2, |
| chr21 | 6776001 | 6862722 | loss | 017906-1 | DDOST |
| chr23 | 5751354 | 5773560 | loss | 017835-1 | RCAN3 |
| chr23 | 5761015 | 5772000 | gain | 017834-2 | RCAN3 |
| chr24 | 285760 | 303620 | loss | 017842-2 | MSANTD2, NRGN |
| chr27 | 4765048 | 4786480 | loss | 017842-2 | THRA |
| chr5 | 5206001 | 5224950 | loss | 017842-2_2 | PAX6 |
| chr5 | 31397310 | 31420368 | loss | 017788-1 | MEIS2 |

**Table S3** (cont'd)

| Chrom | Start | End | Type | Sample | Gene |
|---|---|---|---|---|---|
| chr5 | 55912516 | 55930376 | loss | 017842-2 | OTX2 |
| chr7 | 7253960 | 7300000 | loss | 017835-1 | ADARB1, ADARB1, ADAR |
| chr7 | 7274001 | 7314000 | loss | 017906-1 | ADARB1, ADARB1, ADAR |
| chr7 | 7300001 | 7334000 | loss | 017835-1 | ADARB1, ADARB1, ADAR |
| chr7 | 7846001 | 7908000 | loss | 017835-1 | GLS |
| chr7 | 16318001 | 16408000 | loss | 017911-1 | HOXD4, HOXD11, HOXD1 |
| chr8 | 27125768 | 27140056 | loss | 017842-2 | NFIA |
| chr9 | 11902001 | 11946000 | loss | 017842-2_2 | ZIC1 |
| chr9 | 16898001 | 16940000 | loss | 017911-1 | SOX2 |
| chr9 | 16906500 | 16921800 | loss | 017842-2_2 | SOX2 |

**Table S4. Somatic gene fusion candidates:** The collection of candidate somatic gene fusions across 22 Marek's disease tumors.

| Junction Coordinate | Fusion Type | Gene ID A | Gene ID B | Gene A | Gene B | Tumor |
|---|---|---|---|---|---|---|
| 1_73643375_+:1_73645365_+ | readthrough | ENSGALG00000036222 | ENSGALG00000017280 | na | KCNA1 | 017738-1 |
| 1_90844166_+:1_90966203_+ | intrachromosomal | ENSGALG00000040949 | ENSGALG00000015418 | na | EPHA6 | 017738-1 |
| 1_96340958_+:1_96490290_+ | intrachromosomal | ENSGALG00000042104 | ENSGALG00000015511 | na | ROBO1 | 017738-1 |
| 1_97305206_-:1_97052465_- | intrachromosomal | ENSGALG00000028863 | ENSGALG00000015519 | na | ROBO2 | 017741-1 |
| 17_8741205_-:17_8712864_- | readthrough | ENSGALG00000041608 | ENSGALG00000001645 | na | KCNT1 | 017741-1 |
| 10_20102972_-:10_20102901_- | readthrough | ENSGALG00000029714 | ENSGALG00000023237 | UNC45A | na | 017766-1 |
| 1_96340958_+:1_96490290_+ | intrachromosomal | ENSGALG00000042104 | ENSGALG00000015511 | na | ROBO1 | 017766-1 |
| 1_96340958_+:1_96490290_+ | intrachromosomal | ENSGALG00000042104 | ENSGALG00000015511 | na | ROBO1 | 017787-2 |
| Z_9146403_-:Z_9164929_+ | interstrand | ENSGALG00000026655 | ENSGALG00000023622 | na | AVD | 017794-1 |
| 1_96340958_+:1_96490290_+ | intrachromosomal | ENSGALG00000042104 | ENSGALG00000015511 | na | ROBO1 | 017794-1 |
| 7_6877361_+:7_6874546_+ | inverted | ENSGALG00000006141 | ENSGALG00000027664 | POFUT2 | YBEY | 017798-1 |
| 1_55614936_+:1_55616619_+ | readthrough | ENSGALG00000012763 | ENSGALG00000012766 | GNPTAB | SYCP3 | 017798-1 |
| 1_97305206_-:1_97052465_- | intrachromosomal | ENSGALG00000028863 | ENSGALG00000015519 | na | ROBO2 | 017798-1 |
| 3_26293731_+:3_26395483_+ | intrachromosomal | ENSGALG00000034313 | ENSGALG00000010000 | na | PRKCE | 017798-1 |
| 18_9346764_+:18_9344962_+ | inverted | ENSGALG00000042060 | ENSGALG00000006880 | na | SLC38A10 | 017798-1 |
| 1_96340958_+:1_96490290_+ | intrachromosomal | ENSGALG00000042104 | ENSGALG00000015511 | na | ROBO1 | 017798-1 |
| 2_60487565_-:2_60460152_- | readthrough | ENSGALG00000043391 | ENSGALG00000012683 | na | RNF144B | 017798-1 |
| 1_75650407_-:1_75636968_- | readthrough | ENSGALG00000013422 | ENSGALG00000013424 | RHNO1 | TULP3 | 017798-1_2 |
| 1_75650639_-:1_75636968_- | readthrough | ENSGALG00000013422 | ENSGALG00000013424 | RHNO1 | TULP3 | 017798-1_2 |
| 1_97305206_-:1_97052465_- | intrachromosomal | ENSGALG00000028863 | ENSGALG00000015519 | na | ROBO2 | 017798-1_2 |
| 10_20102968_-:10_20102687_- | readthrough | ENSGALG00000029714 | ENSGALG00000023237 | UNC45A | na | 017798-1_2 |
| 20_10248814_+:20_10239179_+ | inverted | ENSGALG00000038194 | ENSGALG00000006267 | na | TPX2 | 017798-1_2 |
| 1_96340958_+:1_96490290_+ | intrachromosomal | ENSGALG00000042104 | ENSGALG00000015511 | na | ROBO1 | 017798-1_2 |
| 1_96340958_+:1_96490290_+ | intrachromosomal | ENSGALG00000042104 | ENSGALG00000015511 | na | ROBO1 | 017833-1 |
| 11_20157498_+:11_20167292_+ | readthrough | ENSGALG00000003219 | ENSGALG00000000913 | TMEM231 | CHST6 | 017835-1 |
| 3_26556239_+:3_25402442_- | interstrand | ENSGALG00000010000 | ENSGALG00000009967 | PRKCE | LRPPRC | 017835-1 |
| 2_32690122_-:2_32678847_- | readthrough | ENSGALG00000028908 | ENSGALG00000027925, | HOXA6 | HOXA3,HOX/ | 017835-1 |
| 1_96340958_+:1_96490290_+ | intrachromosomal | ENSGALG00000042104 | ENSGALG00000015511 | na | ROBO1 | 017835-1 |
| 22_86148_+:22_96982_+ | readthrough | ENSGALG00000041911 | ENSGALG00000039261 | AAK1 | na | 017841-3 |
| 1_96340958_+:1_96490290_+ | intrachromosomal | ENSGALG00000042104 | ENSGALG00000015511 | na | ROBO1 | 017841-3 |

**Table S4** (cont'd)

| Junction Coordinate | Fusion Type | Gene ID A | Gene ID B | Gene A | Gene B | Tumor |
|---|---|---|---|---|---|---|
| AADN04008201.1_795_-:AADN0 | interchromosomal | ENSGALG00000041807 | ENSGALG00000042711 | PC | na | 017842-2 |
| 1_96340958_+:1_96490290_+ | intrachromosomal | ENSGALG00000042104 | ENSGALG00000015511 | na | ROBO1 | 017842-2 |
| AADN04008201.1_795_-:AADN0 | interchromosomal | ENSGALG00000041807 | ENSGALG00000042711 | PC | na | 017842-2_2 |
| 4_88306232_-:4_88095672_- | intrachromosomal | ENSGALG00000015966 | ENSGALG00000027853 | na | CTNNA2 | 017855-1 |
| 1_24736019_-:1_24735461_- | readthrough | ENSGALG00000028180 | ENSGALG00000037206 | RF02184 | ST7 | 017855-1 |
| AADN04008201.1_795_-:AADN0 | interchromosomal | ENSGALG00000041807 | ENSGALG00000042711 | PC | na | 017855-1 |
| 13_6705340_-:13_6706233_- | inverted | ENSGALG00000034774 | ENSGALG00000001750 | na | MAT2B | 017855-1_2 |
| 1_96340958_+:1_96490290_+ | intrachromosomal | ENSGALG00000042104 | ENSGALG00000015511 | na | ROBO1 | 017855-1_2 |
| 1_97305206_-:1_97052465_- | intrachromosomal | ENSGALG00000028863 | ENSGALG00000015519 | na | ROBO2 | 017863-1 |
| 33_584816_-:33_596443_+ | interstrand | ENSGALG00000037953 | ENSGALG00000040260 | TUBA1A | TUBA1C | 017863-1 |
| 1_96340958_+:1_96490290_+ | intrachromosomal | ENSGALG00000042104 | ENSGALG00000015511 | na | ROBO1 | 017863-1 |
| 17_9327965_+:17_9334340_+ | readthrough | ENSGALG00000001225 | ENSGALG00000046223 | RABGAP1 | na | 017884-2 |
| 5_12577450_-:5_12573281_- | readthrough | ENSGALG00000006342 | ENSGALG00000006312 | UEVLD | TSG101 | 017884-2 |
| 2_36715175_-:2_36520955_- | intrachromosomal | ENSGALG00000039873 | ENSGALG00000011283 | na | ZNF385D | 017884-2 |
| 5_16802494_-:5_16812988_- | inverted | ENSGALG00000041460 | ENSGALG00000007203 | na | RAB3IL1 | 017884-2 |
| 17_8741205_-:17_8712864_- | readthrough | ENSGALG00000041608 | ENSGALG00000001645 | na | KCNT1 | 017884-2 |
| 1_96340958_+:1_96490290_+ | intrachromosomal | ENSGALG00000042104 | ENSGALG00000015511 | na | ROBO1 | 017884-2 |
| 33_584816_-:33_596443_+ | interstrand | ENSGALG00000037953 | ENSGALG00000040260 | TUBA1A | TUBA1C | 017901-2_2 |
| 1_96340958_+:1_96490290_+ | intrachromosomal | ENSGALG00000042104 | ENSGALG00000015511 | na | ROBO1 | 017901-2_2 |
| 17_9327965_+:17_9334340_+ | readthrough | ENSGALG00000001225 | ENSGALG00000046223 | RABGAP1 | na | 017906-1 |
| 4_34582504_+:4_34589007_+ | readthrough | ENSGALG00000010185 | ENSGALG00000037040 | HSPA4L | na | 017906-1 |
| 1_96340958_+:1_96490290_+ | intrachromosomal | ENSGALG00000042104 | ENSGALG00000015511 | na | ROBO1 | 017906-1 |
| 1_55614936_+:1_55616619_+ | readthrough | ENSGALG00000012763 | ENSGALG00000012766 | GNPTAB | SYCP3 | 017911-1 |
| 22_2212167_+:22_2211334_+ | inverted | ENSGALG00000038826 | ENSGALG00000003092 | na | ERLIN2 | 017911-1 |
| 1_96340958_+:1_96490290_+ | intrachromosomal | ENSGALG00000042104 | ENSGALG00000015511 | na | ROBO1 | 017911-1 |
| Z_11948598_+:Z_11954617_+ | readthrough | ENSGALG00000032030 | ENSGALG00000003726 | na | EGFLAM | 017911-1_2 |
| 16_74803_+:16_88780_- | interstrand | ENSGALG00000041380 | ENSGALG00000033932 | BF2 | BF1 | 017911-1_2 |
| 1_96340958_+:1_96490290_+ | intrachromosomal | ENSGALG00000042104 | ENSGALG00000015511 | na | ROBO1 | 017911-1_2 |
| 20_9806274_+:20_9805703_+ | inverted | ENSGALG00000031365 | ENSGALG00000006062 | na | ZBTB46 | 017918-3 |
| 3_4395365_+:3_4412978_+ | readthrough | ENSGALG00000035047 | ENSGALG00000008609 | na | PTK7 | 017918-3 |
| 1_96340958_+:1_96490290_+ | intrachromosomal | ENSGALG00000042104 | ENSGALG00000015511 | na | ROBO1 | 017918-3 |
| 1_96340958_+:1_96490290_+ | intrachromosomal | ENSGALG00000042104 | ENSGALG00000015511 | na | ROBO1 | 017927-2 |

**REFERENCES**

## REFERENCES


1. Ross, N. L. J. T-cell transformation by Marek's disease virus. *Trends Microbiol.* **7,** 22–29 (1999).

2. *Marek's Disease: An Evolving Problem*. (Elsevier Academic Press, 2004).

3. Read, A. F. *et al.* Imperfect vaccination can enhance the transmission of highly virulent pathogens. *PLoS Biol.* **13,** e1002198 (2015).

4. Pastoret, P.-P. Introduction. in *Marek's Disease: An Evolving Problem* (eds. Davison, F. & Nair, V.) 1–7 (Elsevier Academic Press, 2004). doi:10.1016/B978-012088379-0/50005-0

5. Robinson, C. M., Cheng, H. H. & Delany, M. E. Temporal kinetics of Marek's disease herpesvirus: integration occurs early after infection in both B and T cells. *Cytogenet. Genome Res.* **144,** 142–154 (2014).

6. Osterrieder, N., Kamil, J. P., Schumacher, D., Tischer, B. K. & Trapp, S. Marek's disease virus: from miasma to model. *Nat. Rev. Microbiol.* **4,** 283–294 (2006).

7. Tai, S. H. S. *et al.* Expression of Marek's disease virus oncoprotein Meq during infection in the natural host. *Virology* **503,** 103–113 (2017).

8. McPherson, M. C., Cheng, H. H. & Delany, M. E. Marek's disease herpesvirus vaccines integrate into chicken host chromosomes yet lack a virus-host phenotype associated with oncogenic transformation. *Vaccine* **34,** 5554–5561 (2016).

9. Robinson, C. M., Hunt, H. D., Cheng, H. H. & Delany, M. E. Chromosomal integration of an avian oncogenic herpesvirus reveals telomeric preferences and evidence for lymphoma clonality. *Herpesviridae* **1,** 5 (2010).

10. Lupiani, B. *et al.* Marek's disease virus-encoded Meq gene is involved in transformation of lymphocytes but is dispensable for replication. *Proc. Natl. Acad. Sci.* **101,** 11815–11820 (2004).

11. Delecluse, H. J. & Hammerschmidt, W. Status of Marek's disease virus in established lymphoma cell lines: herpesvirus integration is common. *J. Virol.* **67,** 82–92 (1993).

12. Kaufer, B. B., Jarosinski, K. W. & Osterrieder, N. Herpesvirus telomeric repeats facilitate genomic integration into host telomeres and mobilization of viral DNA during reactivation. *J. Exp. Med.* **208,** 605–615 (2011).

13. Morissette, G. & Flamand, L. Herpesviruses and chromosomal integration. *J. Virol.* **84,** 12100–12109 (2010).

14. Calnek, B. W., Carlisle, J. C., Fabricant, J., Murthy, K. K. & Schat, K. A. Comparative pathogenesis studies with oncogenic and nononcogenic Marek's disease viruses and turkey herpesvirus. *Am. J. Vet. Res.* **40,** 541–548 (1979).

15. Schat, K. A. Role of the spleen in the pathogenesis of Marek's disease. *Avian Pathol.* **10,** 171–182 (1981).

16. Jones, D., Leet, L., Liu, J., Kung, H. & Tillotson, J. K. Marek disease virus encodes a basic-leucine zipper gene resembling the fos/jun oncogenes that is highly expressed in lymphoblastoid tumors. *Proc. Natl. Acad. Sci.* **89,** 4042–4046 (1992).

17. Silva, R. F., Dunn, J. R., Cheng, H. H. & Niikura, M. A MEQ-deleted Marek's disease virus cloned as a bacterial artificial chromosome is a highly efficacious vaccine. *Avian Dis.* **54,** 862–869 (2010).

18. Liu, J. L. *et al.* Functional interactions between herpesvirus oncoprotein MEQ and cell cycle regulator CDK2. *J. Virol.* **73,** 4208–4219 (1999).

19. Subramaniam, S. *et al.* Integrated Analyses of Genome-Wide DNA Occupancy and Expression Profiling Identify Key Genes and Pathways Involved in Cellular Transformation by a Marek's Disease Virus Oncoprotein, Meq. *J. Virol.* **87,** 9016–9029 (2013).

20. Qian, Z., Brunovskis, P., Lee, L., Vogt, P. K. & Kung, H. J. Novel DNA binding specificities of a putative herpesvirus bZIP oncoprotein. *J. Virol.* **70,** 7161–7170 (1996).

21. Levy, A. M. *et al.* Characterization of the chromosomal binding sites and dimerization partners of the viral oncoprotein Meq in Marek's disease virus-transformed T cells. *J. Virol.* **77,** 12841–12851 (2003).

22. Suchodolski, P. F. *et al.* Both homo and heterodimers of Marek's disease virus encoded Meq protein contribute to transformation of lymphocytes in chickens. *Virology* **399,** 312–321 (2010).

23. Gennart, I. *et al.* Marek's disease: Genetic regulation of gallid herpesvirus 2 infection and latency. *Vet. J.* **205,** 339–348 (2015).

24. Kung, H. J. *et al.* Meq: an MDV-specific bZIP transactivator with transforming properties. *Curr. Top. Microbiol. Immunol.* **255,** 245–260 (2001).

25. Alexandrov, L. B. & Stratton, M. R. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr. Opin. Genet. Dev.* **24,** 52–60 (2014).

26. Stratton, M. R. Exploring the genomes of cancer cells: progress and promise. *Science* **331,** 1553–1558 (2011).

27. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458,** 719–724 (2009).

28. Li, Y. *et al.* The MYC, TERT, and ZIC1 genes are common targets of viral integration and transcriptional deregulation in avian leukosis virus subgroup J-induced myeloid leukosis. *J. Virol.* **88,** 3182–3191 (2014).

29. Laird, P. W. Cancer epigenetics. *Hum. Mol. Genet.* **14,** R65–R76 (2005).

30. Cooper, D. N., Mort, M., Stenson, P. D., Ball, E. V & Chuzhanova, N. A. Methylation-mediated deamination of 5-methylcytosine appears to give rise to mutations causing human inherited disease in CpNpG trinucleotides, as well as in CpG dinucleotides. *Hum. Genomics* **4,** 406–410 (2010).

31. Govindan, R. *et al.* Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell* **150,** 1121–1134 (2012).

32. Ji, X. *et al.* Somatic mutations, viral integration and epigenetic modification in the evolution of hepatitis B virus-induced hepatocellular carcinoma. *Curr. Genomics* **15,** 469–480 (2015).

33. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339,** 1546–1558 (2013).

34. Martincorena, I. *et al.* Universal patterns of selection in cancer and somatic tissues. *Cell* **171,** 1029–1041 (2017).

35. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).

36. Tomasetti, C., Vogelstein, B. & Parmigiani, G. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proc. Natl. Acad. Sci.* **110,** 1999–2004 (2013).

37. Roberts, S. A. & Gordenin, D. A. Hypermutation in human cancer genomes: footprints and mechanisms. *Nat. Rev. Cancer* **14,** 786–800 (2014).

38. Parsons, R. *et al.* Hypermutability and mismatch repair deficiency in RER+ tumor cells. *Cell* **75,** 1227–1236 (1993).

39. McLendon, R. *et al.* Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455,** 1061–1068 (2008).

40. Thiagalingam, S. *et al.* Evaluation of candidate tumour suppressor genes on chromosome 18 in colorectal cancers. *Nat. Genet.* **13,** 343–346 (1996).

41. Wood, L. D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* **318,** 1108–1113 (2007).

42. Gonzalez-Perez, A. *et al.* Computational approaches to identify functional genetic variants in cancer genomes. *Nat. Methods* **10,** 723–729 (2013).

43. Nair, V. Evolution of Marek's disease - A paradigm for incessant race between the pathogen and the host. *Vet. J.* **170,** 175–183 (2005).

44. Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat. Methods* **9,** 671–675 (2012).

45. Van der Auwera, G. A. *et al.* From fastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. in *Current Protocols in Bioinformatics* 11–10 (2013). doi:10.1002/0471250953.bi1110s43

46. Andrews, S. (Babraham I. FastQC: A quality control tool for high throughput sequence data. *[Online] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/* (2013).

47. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10–12 (2013).

48. Hillier, L. W. *et al.* Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432,** 695–716 (2004).

49. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).

50. Broad Institute. Picard Tools.

51. Mckenna, A. *et al.* The Genome Analysis Toolkit : a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).

52. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

53. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27,** 2987–2993 (2011).

54. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30,** 413–421 (2012).

55. Cheng, H. H. *et al.* Fine mapping of QTL and genomic prediction using allele-specific expression SNPs demonstrates that the complex trait of genetic resistance to Marek's disease is predominantly determined by transcriptional regulation. *BMC Genomics* **16,** 816 (2015).

56. Wang, K. *et al.* PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17,** 1665–1674 (2007).

57. Sun, W. *et al.* Integrated study of copy number states and genotype calls using high-density SNP arrays. *Nucleic Acids Res.* **37,** 5365–5377 (2009).

58. Boeva, V. *et al.* Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *BIOINFORMATICS* **28,** 423–425 (2012).

59. Boeva, V. *et al.* Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics* **27,** 268–269 (2011).

60. Miller, C. A., Hampton, O., Coarfa, C. & Milosavljevic, A. ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLoS One* **6,** e16327 (2011).

61. Chen, K. *et al.* BreakDancer: an algorithm for high resolution mapping of genomic structural variation. *Nat. Methods* **6,** 677–681 (2009).

62. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28,** 333–339 (2012).

63. Chong, Z. *et al.* novoBreak: local assembly for breakpoint detection in cancer genomes. *Nat. Methods* **14,** 65–67 (2016).

64. Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res.* **44,** D67–D72 (2016).

65. Rodríguez-Martín, B. *et al.* ChimPipe: accurate detection of fusion genes and transcription-induced chimeras from RNA-seq data. *BMC Genomics* **18,** 7 (2017).

66. Fan, Y. *et al.* MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol.* **17,** 178 (2016).

67. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31,** 213–219 (2013).

68. Roth, A. *et al.* JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics* **28,** 907–913 (2012).

69. Larson, D. E. *et al.* Somaticsniper: Identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* **28,** 311–317 (2012).

70. Lai, Z. *et al.* VarDict: A novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* **44,** 1–11 (2016).

71. Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22,** 568–576 (2012).

72. Fang, L. T. *et al.* An ensemble approach to accurately detect somatic mutations using SomaticSeq. *Genome Biol.* **16,** 197 (2015).

73. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin).* **6,** 80–92 (2012).

74. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–311 (2001).

75. McLaren, W. *et al.* The Ensembl variant effect predictor. *Genome Biol.* **17,** 122 (2016).

76. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and Indels. *PLoS One* **7,** e46688 (2012).

77. Kaminker, J. S. *et al.* Distinguishing cancer-associated missense mutations from common polymorphisms. *Cancer Res.* **67,** 465–473 (2007).

78. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using SIFT algorithm. *Nat. Protoc.* **4,** 1073–1082 (2009).

79. Dees, N. D. *et al.* MuSiC: Identifying mutational significance in cancer genomes. *Genome Res.* **22,** 1589–1598 (2012).

80. Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29,** 2238–2244 (2013).

81. Cho, A. *et al.* MUFFINN: cancer gene discovery via network analysis of somatic mutation data. *Genome Biol.* **17,** 129 (2016).

82. Forbes, S. A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45,** D777–D783 (2017).

83. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: An R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33,** 2938–2940 (2017).

84. Murrell, P. *R Graphics*. (CRC Press, 2012).

85. Skidmore, Z. L. *et al.* GenVisR: genomic visualizations in R. *Bioinformatics* **32,** 3012–3014 (2016).

86. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14,** 178–192 (2013).

87. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29,** 24–26 (2011).

88. Barnett, D. W., Garrison, E. K., Quinlan, A. R., Stŕmberg, M. P. & Marth, G. T. Bamtools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27,** 1691–1692 (2011).

89. Bateman, A. *et al.* UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **45,** D158–D169 (2017).

90. Aken, B. L. *et al.* Ensembl 2017. *Nucleic Acids Res.* **45,** D635-642 (2017).

91. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44,** D733–D745 (2016).

92. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23,** 1282–1288 (2007).

93. Altschul, S. F. *et al.* Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).

94. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).

95. Remmert, M. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7,** 539–539 (2011).

96. Livingstone, C. D. & Barton, G. J. Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.* **9,** 745–756 (1993).

97. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2-a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25,** 1189–1191 (2009).

98. Stone, H. *USDA Tech. Bull. No. 1514*. (1975).

99. Banerji, S. *et al.* Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486,** 405–409 (2012).

100. Wilm, A. *et al.* LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.* **40,** 11189–11201 (2012).

101. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463,** 899–905 (2010).

102. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458,** 223–227 (2009).

103. Rouse, B. T., Wells, R. J. H. & Warner, N. L. Proportion of T and B lymphocytes in lesions of Marek's diseases: theoretical implications for pathogenesis. *J. Immunol.* **110,** 534–539 (1973).

104. Payne, L. N. & Rennie, M. The proportions of B and Y lymphocytes in lymphomas, peripheral nerves and lymphoid organs in Marek's disease. *Avian Pathol.* **5,** 147–154 (1976).

105. Hudson, L. & Payne, L. N. An analysis of the T and B cells of Marek's disease lymphomas of the chicken. *Nat. New Biol.* **241,** 52 (1973).

106. Payne, L. N. Pathology. in *Marek's Disease: Scientific Basis and Methods of Control* (ed. Payne, L. N.) 43–75 (Martinus Nijhoff, 1985).

107. Sharma, J. M. Laboratory diagnosis. in *Marek's Disease: Scientific Basis and Methods of Control* (ed. Payne, L.) 151–175 (Martinus Nijhoff, 1985).

108. Sevoian, M., Chamberlain, D. M. & Counter, F. T. Avian lymphomatosis. I. Experimental reproduction of the neural and visceral forms. *Vet. Med.* **57,** 500–501 (1962).

109. Schneider, G., Schmidt-supprian, M., Rad, R. & Saur, D. Tissue-specific tumorigenesis: context matters. *Nat. Rev. Cancer* **17,** 239 (2017).

110. Winandy, S., Wu, P. & Georgopoulos, K. A dominant mutation in the Ikaros gene leads to rapid development of leukemia and lymphoma. *Cell* **83,** 289–299 (1995).

111. Molnár, A. & Georgopoulos, K. The Ikaros gene encodes a family of functionally diverse zinc finger DNA-binding proteins. *Mol. Cell. Biol.* **14,** 8292–8303 (1994).

112. Morgan, B. *et al.* Aiolos, a lymphoid restricted transcription factor that interacts with Ikaros to regulate lymphocyte differentiation. *The EMBO Journal* **16,** 2004–2013 (1997).

113. Kuehn, H. S. *et al.* Loss of B cells in patients with heterozygous mutations in IKAROS. *N. Engl. J. Med.* **374,** 1032–1043 (2016).

114. Payne, M. A. Zinc finger structure-function in Ikaros. *World J. Biol. Chem.* **2,** 161–166 (2011).

115. Davis, K. L. Ikaros: master of hematopoiesis, agent of leukemia. *Ther. Adv. Hematol.* **2,** 359–368 (2011).

116. Hosokawa, Y. *et al.* The Ikaros gene, a central regulator of lymphoid differentiation, fuses to the BCL6 gene as a result of t(3;7)(q27;p12) translocation in a patient with diffuse large B-cell lymphoma. *Blood* **95,** 2719–2721 (2000).

117. Kastner, P. *et al.* Function of Ikaros as a tumor suppressor in B cell acute lymphoblastic leukemia. *Am J Blood Res* **3,** 1–13 (2013).

118. Sun, L. *et al.* Expression of dominant-negative and mutant isoforms of the antileukemic transcription factor Ikaros in infant acute lymphoblastic leukemia. *Proc. Natl. Acad. Sci.* **96,** 680–685 (1999).

119. Zhang, J. *et al.* The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature* **481,** 157–163 (2012).

120. Greaves, M. A causal mechanism for childhood acute lymphoblastic leukaemia. *Nat. Rev. Cancer* 1 (2018). doi:10.1038/s41568-018-0015-6

121. Welch, J. S. *et al.* The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150,** 264–278 (2012).

122. Georgopoulos, K., Moore, D. & Derfler, B. Ikaros, an early lymphoid-specific transcription factor and a putative mediator for T cell commitment. *Science* **258,** 808–812 (1992).

123. Fan, Y. & Lu, D. The Ikaros family of zinc-finger proteins. *Acta Pharm. Sin. B* **6,** 513–521 (2016).

124. Sun, L. *et al.* Expression of aberrantly spliced oncogenic Ikaros isoforms in childhood acute lymphoblastic leukemia. *J. Clin. Oncol.* **17,** 3753–3766 (1999).

125. Payne, K. J., Nicolas, J. H., Zhu, J. Y., Barsky, L. W. & Crooks, G. M. Cutting edge: predominant expression of a novel Ikaros isoform in normal human hemopoiesis. *J. Immunol.* **167,** 1867–1870 (2001).

126. Georgopoulos, K. Haematopoietic cell-fate decisions, chromatin regulation and Ikaros. *Nat. Rev. Immunol.* **2,** 162–174 (2002).

127. Haire, R. N., Miracle, A. L., Rast, J. P. & Litman, G. W. Members of the Ikaros gene family are present in early representative vertebrates. *J. Immunol.* **165,** 306–312 (2000).

128. Molnár, A. *et al.* The Ikaros gene encodes a family of lymphocyte-restricted zinc finger DNA binding proteins, highly conserved in human and mouse. *J. Immunol.* **156,** 585–592 (1996).

129. Cobb, B. S. *et al.* Targeting of Ikaros to pericentromeric heterochromatin by direct DNA binding Targeting of Ikaros to pericentromeric heterochromatin by direct DNA binding. *Genes Dev.* **14,** 2146–2160 (2000).

130. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).

131. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).

132. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).

133. Wu, H., Wang, C. & Wu, Z. PROPER: comprehensive power evaluation for differential expression using RNA-seq. *Bioinformatics* **31,** 233–241 (2015).

134. Anderson, E. *et al. LAPACK: Users' Guide*. (Society for Industrial and Applied Mathematics, 1999).

135. Becker, R. A., Chambers, J. M. & Wilks, A. R. *The New S Language*. (CRC Press Taylor & Francis Group, 1988).

136. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3,** 1724–1735 (2007).

137. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9,** e1003118 (2013).

138. Pagès, H., Carlson, M., Falcon, S. & Li, N. AnnotationDbi: annotation database interface. (2018).

139. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature* **25,** 25–29 (2000).

140. Carbon, S. *et al.* Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* **45,** D331–D338 (2017).

141. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28,** 27–30 (2000).

142. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44,** D457–D462 (2016).

143. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45,** D353–D361 (2017).

144. Fabregat, A. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res.* **46,** D649–D655 (2018).

145. Slenter, D. N. *et al.* WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.* **46,** D661–D667 (2018).

146. Nishimura, D. BioCarta. *Biotech Softw. Internet Rep. Comput. Softw. J. Sci.* **2,** 117–120 (2001).

147. Schaefer, C. F. *et al.* PID: the pathway interaction database. *Nucleic Acids Res.* **37,** D674–D679 (2009).

148. Thomas, P. D. *et al.* PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res.* **13,** 2129–2141 (2003).

149. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102,** 15545–15550 (2005).

150. Liberzon, A. *et al.* The molecular signatures database hallmark gene set collection. *Cell Syst.* **1,** 417–425 (2015).

151. Geimer Le Lay, A.-S. *et al.* The tumor suppressor Ikaros shapes the repertoire of notch target genes in T cells. *Sci. Signal.* **7,** ra28 (2014).

152. Kuleshov, M. V *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44,** W90–W97 (2016).

153. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14,** 128 (2013).

154. Reimand, J. *et al.* g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.* **8,** W83–W89 (2016).

155. Kolde, R. Pheatmap: pretty heatmaps. (2012).

156. Liao, Y., Smyth, G. K. & Shi, W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* **41,** e108 (2013).

157. Liao, Y., Smyth, G. K. & Shi, W. FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30,** 923–930 (2014).

158. Reyes, A. *et al.* Drift and conservation of differential exon usage across tissues in primate species. *Proc. Natl. Acad. Sci.* **110,** 15377–15382 (2013).

159. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22,** 2008–2017 (2012).

160. Schat, K. A., Chen, C.-L. L., Calnek, B. W. & Char, D. Transformation of T-lymphocyte subsets by Marek's disease herpesvirus. *J. Virol.* **65,** 1408–1413 (1991).

161. Mwangi, W. N. *et al.* Clonal Structure of Rapid-Onset MDV-Driven CD4+ Lymphomas and Responding CD8+ T Cells. *PLoS Pathog.* **7,** (2011).

162. Brady, B. L., Steinel, N. C. & Bassing, C. H. Update and Reappraisal Antigen Receptor Allelic Exclusion: An Antigen Receptor Allelic Exclusion: An Update and Reappraisal. *J Immunol Ref.* **185,** 3801–3808 (2010).

163. Calnek, B. W. Pathogeneis of Marek's Disease Virus Infection. in *Marek's Disease* (ed. Hirai, K.) 25–55 (Springer-Verlag, 2001).

164. Cazzaniga, G. *et al.* Developmental origins and impact of BCR-ABL1 fusion and IKZF1 deletions in monozygotic twins with Ph+ acute lymphoblastic leukemia. *Blood* **118,** 5559–64 (2011).

165. Dudley, J. T., Tibshirani, R., Deshpande, T. & Butte, A. J. Disease signatures are robust across tissues and experiments. *Mol. Syst. Biol.* **5,** 1–8 (2009).

166. Heizmann, B., Kastner, P. & Chan, S. The Ikaros family in lymphocyte development. *Curr. Opin. Immunol.* **51,** 14–23 (2018).

167. Georgopoulos, K. Acute lymphoblastic leukemia - on the wings of IKAROS. *N. Engl. J. Med.* **360,** 524–526 (2009).

168. Viale, A. *et al.* Cell-cycle restriction limits DNA damage and maintains self-renewal of leukaemia stem cells. *Nature* **457,** 51–56 (2009).

169. Mullighan, C. G. *et al.* Deletion of IKZF1 and prognosis in acute lymphoblastic leukemia. *N. Engl. J. Med.* **360,** 470–480 (2009).

170. Iacobucci, I. *et al.* Expression of spliced oncogenic Ikaros isoforms in Philadelphia-positive acute lymphoblastic leukemia patients treated with tyrosine kinase inhibitors: implications for a new mechanism of resistance. *Blood* **112,** 3847–3855 (2008).

171. Iacobucci, I. *et al.* Identification and molecular characterization of recurrent genomic deletions on 7p12 in the IKZF1 gene in a large cohort of BCR-ABL1 – positive acute lymphoblastic leukemia patients: on behalf of Gruppo Italiano Malattie Ematologiche dell'Adulto Acute Leu. *Blood* **114,** 2159–2167 (2009).

172. Harvey, R. C. *et al.* Identification of novel cluster groups in pediatric high-risk B-precursor acute lymphoblastic leukemia with gene expression profiling : correlation with genome-wide DNA copy number alterations , clinical characteristics , and outcome. *Blood* **116,** 4874–4884 (2013).

173. Mullighan, C. G. *et al.* BCR-ABL1 lymphoblastic leukaemia is characterized by the deletion of Ikaros. *Nature* **453,** 110–114 (2008).

174. Hu, Y. *et al.* Superenhancer reprogramming drives a B-cell-epithelial transition and high-risk leukemia. *Genes Dev.* **30,** 1971–1990 (2016).

175. Churchman, M. L. *et al.* Efficacy of retinoids in IKZF1-mutated BCR-ABL1 acute lymphoblastic leukemia. *Cancer Cell* **28,** 343–356 (2015).

176. Joshi, I. *et al.* Loss of Ikaros DNA-binding function confers integrin-dependent survival on pre-B cells and progression to acute lymphoblastic leukemia. *Nat. Immunol.* **15,** 294–304 (2014).

177. Schjerven, H. *et al.* Genetic analysis of Ikaros target genes and tumor suppressor function in BCR-ABL1+ pre–B ALL. *J. Exp. Med.* **214,** 793–814 (2017).

178. Iacobucci, I. *et al.* IKAROS deletions dictate a unique gene expression signature in patients with adult B-cell acute lymphoblastic leukemia. *PLoS One* **7,** e40934 (2012).

179. Jaatinen, T. *et al.* Global gene expression profile of human cord blood–derived CD133+ cells. *Stem Cells* **24,** 631–641 (2006).

180. Scheicher, R. *et al.* CDK6 as a key regulator of hematopoietic and leukemic stem cell activation. *Blood* **125,** 90–101 (2015).

181. Hu, M. G. *et al.* A requirement for cyclin-dependent kinase 6 in thymocyte development and tumorigenesis. *Cancer Res.* **69,** 810–818 (2009).

182. Van Der Linden, M. H. *et al.* MLL fusion-driven activation of CDK6 potentiates proliferation in MLL-rearranged infant ALL. *Cell Cycle* **13,** 834–844 (2014).

183. Placke, T. *et al.* Requirement for CDK6 in MLL-rearranged acute myeloid leukemia. *Blood* **124,** 13–23 (2014).

184. Lee, M. S., Hanspers, K., Barker, C. S., Korn, A. P. & McCune, J. M. Gene expression profiles during human CD4+ T cell differentiation. *Int. Immunol.* **16,** 1109–1124 (2004).

185. Lee, G. *et al.* Isolation and directed differentiation of neural crest stem cells derived from human embryonic stem cells. *Nat. Biotechnol.* **25,** 1468–1475 (2007).

186. Marshall, G. M. *et al.* The prenatal origins of cancer. *Nat. Rev. Cancer* **14,** 277–289 (2014).

187. Wiemels, J. L. *et al.* Prenatal origin of acute lymphoblastic leukaemia in children. *Lancet* **354,** 1499–1503 (1999).

188. Cazzaniga, G. *et al.* Possible role of pandemic AH1N1 swine flu virus in a childhood leukemia cluster. *Leukemia* **31,** 1819–1821 (2017).