Spring 5-15-2020

Epigenetic Dynamics in Normal and Disease Models

Hyo Sik Jang
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Molecular Genetics & Genomics

Dissertation Examination Committee:
Ting Wang, Chair
Charles K. Kaufman
Albert H. Kim
Rob Mitra
Tim Schedl

Epigenetic Dynamics in Normal and Disease Models

by

Hyo Sik (Josh) Jang

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2020

St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgements

These past six years have been the most enriching experiences in my life, both scientifically and personally. First, all the glory of my successes goes to God. There have been many dark moments throughout my PhD career, and I can confidently attest that faith was essential in helping me overcome the adversities to flourish. Second, I want to thank my family for supporting my decision to pursue a PhD and providing prayers of encouragement and health. Although they were on the other side of the planet, I could constantly feel their warmth through loving phone calls and emojis. Saranghae!

Also, I want to thank from the bottom of my heart, my thesis mentor and advisor, Ting Wang. What a crazy roller coaster ride it has been and I am so glad that I had you in the seat next to me. Ting has been a fantastic leader, inspiring role model, passionate scientist, and an incredible friend. There is no one like him in the world and I feel so honored and blessed to have him in my life. Much of my success is due to Ting's unwavering support and guidance to chase after novel biology and high-risk experiments. Ting's contribution to my scientific growth is indisputable. However, more importantly, Ting has taught me to become a better person. He leads by example with his calm demeanor, infinite patience and abundant generosity. He taught me to manage expectations and to really value living life over just performing science. And for that, I am eternally grateful. Although my PhD adventure with Ting has ended, I look forward to all the fun shenanigans we will continue to have as I move on to the next steps of my life.

Next, I want to thank my collaborators and mentors for deepening my scientific intellect and critically challenging my scientific progress. The members in my thesis committee, Drs. Albert Kim, Tim Schedl, Rob Mitra, Chuck Kaufman and Ting Wang, provided critical criticism, interdisciplinary expertise, and realistic expectations to minimize the stress and maximize the value of my PhD training. I thank you all for your wisdom, enthusiasm and sincerity to make my PhD experience so joyful.

Lot of the scientific advances would not have been possible without the help of essential personnel at Wash U. First, I want to thank those part of the CGS-SB: Jessica Hoisington-Lopez and MariaLynn Jaegar for invaluable assistance with sequencing, Brian Koebbe and Eric Martin for HTCF access, and David O'Donnell for mouse husbandry. I want to thank Catrina Fronick at MGI for Nova-Seq expertise. Majority of the cell sorting was performed at Siteman Flow Cytometry Core and I am very grateful to Dan Schweppe, Matt Patana and Michael Savio for their help. I want to thank everyone in the Zebrafish Facility, such as Zak Kupchinsky, Brian Stephens, John Engelhard, Megan Barbier and Sainan Li. When I came to Wash U, I did not think that I would be working on the zebrafish model system. My heart aches to remember and honor Stephen L. Johnson, a valuable mentor and friend, who introduced me to the beauty of zebrafish genetics. Also, I want to thank MGG DBBS program and leaders, Jim Skeath, Tim Schedl, John Edwards and Melanie Relich for their sincere love for the graduate students

Lastly, I want to acknowledge my lab and friends who brought joy to my life throughout my PhD career. First, I want to thank everyone in the lab for providing constructive criticism, establishing a positive environment and putting up with my terrible/inappropriate jokes. I want to spotlight Nakul Shah, my partner-in-crime, in all the cancer work in the lab. He is truly a hardworking, intelligent, and empathetic scientist who I am grateful to have in my life. Also, a sincere thank you to all the mentees who helped with my scientific progress. I am so proud of each and every one of you! I am looking forward to all the great things you will accomplish in the future!

I thank my funding source: NIGMS T32 GM007067.

Hyo Sik Jang

Washington University in St. Louis

May 2020

Dedicated to God and family, both in blood and in spirit.

ABSTACT OF THE DISSERTATION

Epigenetic Dynamics in Normal and Disease Models

by

Hyo Sik (Josh) Jang

Doctor of Philosophy in Biology and Biomedical Sciences

Molecular Genetics & Genomics

Washington University in St. Louis, 2020

Professor Ting Wang, Chair

Deciphering how epigenetic factors, such as DNA methylation and chromatin accessibility, shape normal development and disease progression has been an outstanding goal in developmental biology. Here, I present multiple branches of my thesis to elucidate the epigenetic controls that direct aging in brain, regulate cell fate decision of zebrafish iridophore in pigment differentiation, and dysregulate transposable elements (TEs) in cancer. The first branch focuses on benchmarking a computational statistic tool to characterize DNA methylation dynamics of aging in mouse prefrontal cortex by combining WGBS and TAB-seq to dissect the contribution of CpG methylation and hydroxymethylation. For the second branch, we take advantage of the elegant zebrafish model system to answer how epigenetic dynamics shape pigment development. We developed conditional CRISPR knockout method, which if combined with clonal analysis, can provide temporal and cell lineage-specific resolution. Furthermore, we profiled DNA methylation, chromatin accessibility, and gene expression across various biological timepoints of neural crest differentiation to pigment cells in zebrafish. Here, I focus on exploring the genetic and epigenetic dynamics that drive iridophore cell fate. In the third branch, TEs are an

underexplored genetic resource that impact both normal development and disease. Especially in the context of cancer, recent discoveries exemplify how particular TEs are epigenetically reactivated to provide enhancer or promoter regulatory roles, known as onco-exaptation, that contribute to oncogenesis. One example is the reactivation of cryptic promoters in TEs that provide alternative transcription start sites (TSS) for oncogenes. These alternative TSSs can generate chimeric or truncated oncogene transcripts that could accelerate tumorigenesis. However, TEs may be a double-edged sword for cancer, as aberrant TE activation can provide additional sequences to be translated into novel peptides that can be used as biomarkers or targets for immunotherapy through cancer vaccines or enhanced T cell therapy. Recent work has revealed that epigenetic therapy (epitherapy) can preferentially activate epigenetically silenced TEs, which generates epitherapy-specific transcripts and potential novel cancer-specific antigens that can be exploited as therapeutic targets for immunotherapy. I aim to study the prevalence of onco-exaptation events across numerous cancer types and explore potential immunotherapeutic approaches by exploiting TE-specific transcripts in the glioblastoma in the presence of epitherapy.

Chapter 1: Introduction

Humans share 99.9% of the same genetic code yet are phenotypically diverse in many ways, such skin color, body size and facial structure. The genetic code consists of letters (A, C, G, T) that represent the four nucleotides, which create the DNA. The context and order of these letters is responsible for the species diversity, ranging from simple virus to complex multicellular organisms, in the world. Rapid advances in DNA sequencing technologies unraveled the complex variations in DNA sequences that differentiate one organism from another. Multicellular organisms often arise from a single embryo with one genetic code. For example, in humans, once a sperm fertilizes the egg to generate a single cell embryo, the embryo faithfully and rapidly divides into three germ layers to further differentiate into various tissue types, such as skin, bone and nerves. How a single cell with static genetic code could give rise to morphologically and functionally complex cell types has been a question much sought after in the field of developmental biology.

1.1 Early theory on cell fate decision

Within the human genome, there are stretches of DNA regions called genes that are functional units of heredity and responsible for producing cellular traits. Around 1.5% of the human genome has been identified to protein-coding genes. The process of generating protein from DNA involves an intermediary RNA transcript that hold information on what amino acids should be attached together to generate a peptide. In 1957, Conrad Waddington, a developmental biologist and geneticist, proposed a controversial theory that cell fate is determined by series of gene expression modules and decisions[1–3]. Early in development, a cell has the potential to become various cell types, a cellular state called pluripotency. However, once a pluripotent cell chooses a specific gene expression pathway, it commits to a certain fate and no longer retains its ability to become any cell type. He represented this concept in his famous "epigenetic landscape" portrait where a pluripotent cell is depicted as a ball on top of a landscape with multiple paths that it can choose (**Fig. 1**). The valleys are divided with hills that make it impossible for the ball to cross over into a different valley once commits to a certain path. This interaction across genes, genome and development is what Waddington coined as "epigenetic" forces that regulate cell fate decisions.

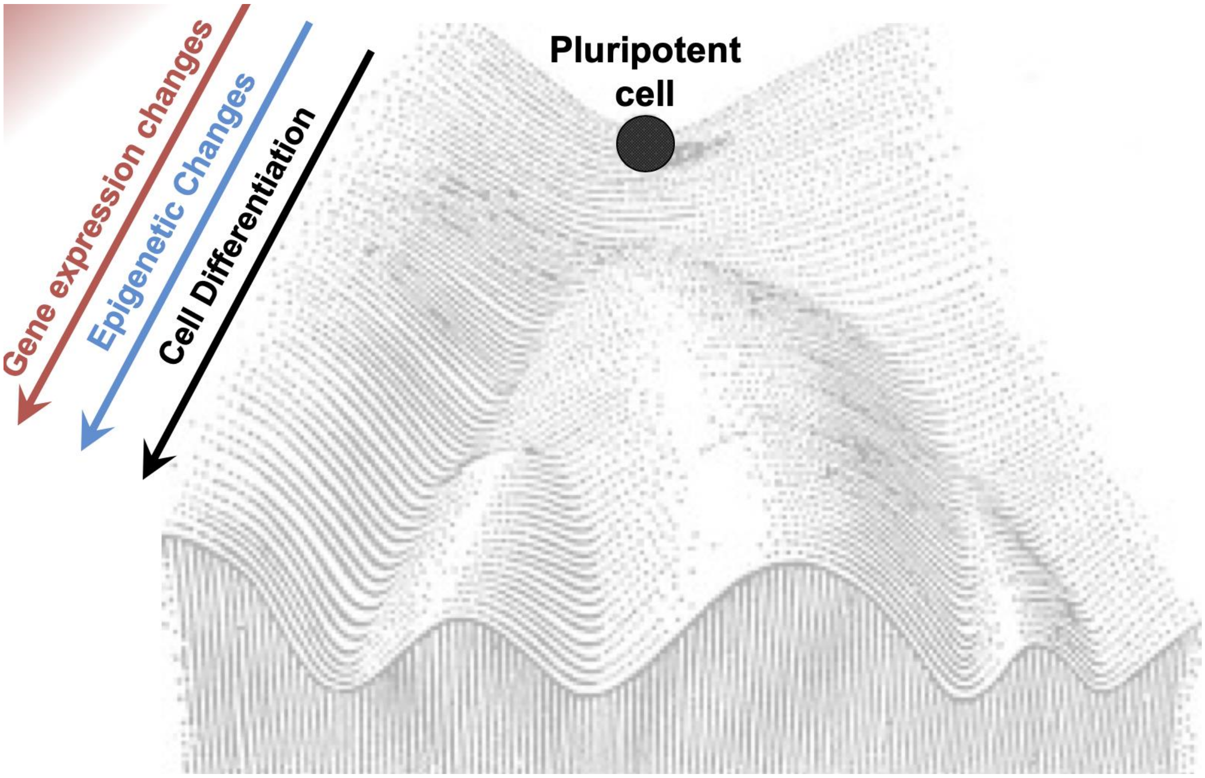


Figure 1. Waddington's epigenetic landscape. The path that the pluripotent cell chooses is defined by gene expression changes, which is controlled by epigenetic mechanism, to ultimately determine cell fate.

1.2 Transcription factors and epigenetic mechanisms

Unbeknownst to Waddington at the time, there were various mechanisms of epigenetic control discovered, beginning from late 1980s, that regulated gene expression. Now, epigenetics has been redefined to include any molecular or biochemical modifications that regulate genome activity, sans DNA sequence alterations[4]. These epigenetic mechanisms activate gene regulatory networks, specific to various cell types, through control of transcriptions factors[5]. Transcription factors (TFs) are DNA-binding proteins that recognize particular DNA sequences or motifs to bind and recruit transcriptional machinery to create RNA from genes. A single TF can bind to multiple genomic regions, including promoters or enhancers, to regulate various genes in a gene regulatory network that is essential for cell identity[6–10]. To minimize promiscuous TF binding in genomic regions that shouldn't be transcribed, the eukaryote genome evolved epigenetic mechanisms, such as DNA methylation[11], histone modifications[12,13] and nuclear organization[14], to repress TF binding. Therefore, to better understand how cell fate is defined, much effort went into defining the epigenetic landscape that determines various tissue and cell fate. Within the past decade, huge consortium efforts, such as ENCODE[15], Roadmap Epigenome[16] and 4DN[17], have epigenetically profiled numerous tissues in human and mouse. Leveraging epigenetic data provided plethora of monumental insights into cell fate decision and the essential TFs that are responsible for cell differentiation.

1.3 Transposable elements in cancer

A large portion of eukaryotic genomes, including at least 50% of the human genome, is derived from TEs[18–20]. TEs are often deemed parasitic DNA[21,22] and can be deleterious when transposition events disrupt protein coding sequences or gene regulatory elements[23–30]. To counteract the deleterious effects of TEs, cells use epigenetic mechanisms, including DNA methylation and repressive histone modifications, to silence transposon-derived sequences in somatic tissues[31–33]. Recently, a wave of discoveries has demonstrated how TEs alter the gene expression landscape during evolution, development, and disease[34–39]. Although epigenetically silenced in somatic tissue, TEs can become transcriptionally active in cancer due to the loss of epigenetic constraint via global DNA hypomethylation or other epigenetic deregulation, which can expose regulatory sequences within TEs and lead to functional consequences[35–38]. Further discussions on both how epigenetic mechanisms and TEs impact cancer progression and oncogenicity is discussed in Chapter 2 of this dissertation.

1.4 The rise of antigen-based immunotherapy

Cancer vaccine therapies are well-studied paradigms of cancer immunotherapy in numerous cancer types[40–43]. Somatic cells express major histocompatibility complex (MHC) proteins on the cell surface. These MHC molecules present various antigens (short peptide fragments) to the immune system, which allows immune cells to distinguish host cells from foreign cells[44]. Cancer neoantigens are defined as a class of human leukocyte antigen (HLA)-bound peptides that arise from tumor-specific mutations. They hold promise as the optimal targets for an anti-tumor immune response[45–49]. An entire research field of cancer neoantigen discovery has been born, which takes advantage of the recent availability of next-generation sequencing-based coding mutation discovery and machine learning approaches to predict mutated peptides with high-affinity binding of HLA molecules[50–52]. However, cancers differ drastically in their mutation rate, and only a handful of neoantigens have been identified to exist in more than 5% of patient samples[53], limiting their universal applicability. The entire expressed coding sequence space, albeit vast, still only makes up less than 1% of the genome. Studies have suggested that chromosomal rearrangement can result in novel, immunogenic protein junctions[54,55], but this mechanism has only limited impact on the space of targeted therapy or neoantigen discovery. Considering the promising potential of immunotherapy in the war against cancer, a push for discovery of novel sources of antigens is necessary to maximize therapeutic potential of antigen-based immunotherapy.

1.5 Crossroad between epigenetic therapy and immunotherapy in glioblastoma

Recent work suggests that transcription of TEs can be modified epigenetically[34]. The effect can be global, as via chemical-based epigenetic therapy; or local, via new technologies of site-specific epigenetic engineering. Epigenetic therapies have long been used to treat cancer as well as other diseases[56–58]. These therapies include inhibitors of DNA methyltransferases (DNMTi) and histone deacetylases (HDACi). In treating cancer, they reactivate epigenetically silenced tumor suppressor genes. We and others have shown that in response to DNMTi and HDACi, certain TEs become transcriptionally active in cancer cells and result in dsRNA, which triggers an anti-viral response[34,59,60]. Importantly, we also discovered that DNMTi and HDACi activate several thousand novel promoters, most of which derive from TEs[34]. Transcripts initiated from these TEs can readthrough and splice into downstream genes, resulting in their overexpression and, sometimes, the formation of a chimeric protein product that can be further exploited as a source for antigens. Thus, epigenetic therapy might be able to dramatically increase cancer type-specific production of antigens. Furthermore, recent evidence emerged that viral infections can also induce TE expression[61,62]. In parallel, we explore the possibility of adapting CRISPR/Cas9-based epigenetic engineering systems to modify epigenetic control of TEs. In our tested cases, we have successfully demethylated a candidate TE and generated a TE-gene fusion protein product with precision[39]. These results suggest that TEs, the vast sequences in our genome deemed "junk DNA", might provide an unprecedented opportunity for the discovery and manipulation of cancer-specific antigens.

We chose glioblastoma (GBM) as our model system to explore epigenetic regulation of TEs in producing cancer antigens. GBM remains a disease with a poor prognosis[63]. Recently, advancements in the field reinvigorated the promise of immunotherapy in treating GBM[64,65]. The systemic immune system can attack multiple targets and has the capacity to penetrate the blood-brain barrier. Several neoantigens have been characterized as targets in GBM[66]; however, the relatively low mutation rate of GBM combined with its immunosuppressive tumor environment have made the discovery of efficacious therapies difficult. Targeting recurrent antigens originating from missense/indel mutation, exon skipping, and cancer-enriched genes have been shown to increase immune activation and lymphocyte infiltration, but this has not been translated over to consistent clinical benefit[48,49,66,67]. Thus, the discovery and validation of a new source of recurrent tumor-specific antigens both before and after epigenetic therapy and virotherapy could greatly enhance the current repertoire of immunotherapy targets and lead to the development of clinically effective combinatorial therapies.

Chapter 2: Epigenetic Alterations in Cancer

Erica C. Pehrsson, PhD[1,2ǂ]*, Hyo Sik Jang[1,2ǂ], Ting Wang, PhD[1,2]

[1]Department of Genetics, Washington University in St. Louis, St Louis, Missouri 63108, USA

[2]Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St Louis, Missouri 63108, USA

[ǂ]Co-first authors

*Corresponding author

Keywords: epigenetics, cancer, DNA methylation, histone modification, chromatin accessibility, topologically-associating domain, epigenetic therapy, transposable element

Author contributions: All authors contributed to writing the manuscript.

This review chapter has been published as a book chapter in Cancer Genomics for the Clinician.

ISBN: 9780826168672

2.1 Introduction to epigenetics

Epigenetics investigates how the static genetic code of a single organism can be regulated in a diverse and dynamic manner to generate distinct phenotypes that can be heritable. In a common analogy, the genomic DNA sequence is the computer hardware, while epigenetic mechanisms are the software that interprets and displays what the hardware provides. Epigenetics involves biochemical modifications to DNA or chromatin, which consequently regulate gene expression through the spatial organization of the genome. In this chapter, we briefly introduce well-established epigenetic marks, illustrated in **Figure 1**, and how they canonically influence normal development as well as cancer initiation and progression. Additionally, we briefly discuss epigenetic therapies and their application in cancer.

2.1.1 DNA methylation

DNA methylation is one of the most well-studied epigenetic modifications associated with gene regulation. In the mammalian genome, DNA methylation is defined by the biochemical addition of a methyl group to the 5-carbon position of the cytosine nitrogenous base. The majority of DNA methylation occurs in the context of CpG dinucleotides, whose symmetrical arrangement preserves the methylation signature on both strands during DNA replication. Recent investigations have identified non-CpG methylation in embryonic stem cells and neuronal cell types, but the functional consequences of these marks have yet to be fully understood[68,69]. CpGs are disproportionally underrepresented (fewer than expected by chance) in the mammalian genome, and most are highly methylated[70,71]. Methylated CpGs are often associated with a repressive role in regulating gene expression, especially in the context of promoters and enhancers. At promoters that contain sparse CpGs, gene transcription inversely correlates with DNA methylation level, with low methylation levels over active promoters. However, roughly 70% of gene promoters have regions of high CpG density that span over 1kb[71]. These CpG-dense regions, known as CpG islands (CGIs), are often unmethylated throughout cell development and differentiation regardless of the expression level of the adjacent gene[69,70]. Notable exceptions include various genes essential for early embryonic development, which are silenced through promoter methylation during subsequent differentiation stages, reaffirming the importance of DNA methylation for proper cellular differentiation[68–71]. Such examples involve X-chromosome inactivation, imprinting, and germ cell-specific pathways.

Past array-based technology limited the number of genomic regions whose methylation could be interrogated, focusing primarily on CGI promoters. However, with the introduction of affordable whole genome sequencing technologies and more comprehensive methylation arrays, we can now profile DNA methylation across the whole genome. Whole genome methylation analysis revealed numerous CGIs and CpGs outside of genes that are differentially methylated depending on tissue type, such that methylation pattern alone can serve as a tissue biomarker. These intergenic variably methylated regions are predicted to be alternative promoters, enhancers, or insulators that regulate gene expression in a tissue-specific manner[69–71].

How DNA methylation regulates gene expression is thought to be two-pronged[68,71]. First, methylation of CpGs in transcription factor (TF) binding sites can directly impede the binding of TFs to the DNA in enhancer or promoter regions. Second, methyl-CpG binding domain proteins (MBDs), such as CXXC1 and MeCP2, can sterically block TFs from binding to the methylated region or can recruit histone modifying enzymes to compact the chromatin[68,69].

In contrast to cis-regulatory regions, CpGs within gene bodies are usually highly methylated, with evidence surfacing that higher gene body methylation correlates with higher expression[71]. Furthermore, gene body methylation can suppress intragenic promoters[70]. Also, exons exhibit slightly higher methylation levels than introns, suggesting that DNA methylation might also play a role in gene splicing events[70,71].

DNA methylation is most prevalent over repeat-rich regions of the genome consisting of transposable elements (TEs), centromeres, and other features. Some TEs, or “jumping” genes, have the ability to copy and reinsert their sequences into the genome, similar to viruses. As these insertion events can be deleterious, the host developed epigenetic mechanisms, such as DNA methylation, to silence and suppress these TEs. DNA methylation of TEs prevents transposition and recombination events that can lead to large deletions or chromosomal translocations, preserving chromosomal integrity[69–71].

Epigenetic modifier proteins can be classified into three functional groups: readers, writers and erasers. Three core enzymes, called DNA methyltransferases (DNMTs), are responsible for biochemically writing DNA methylation throughout the genome. DNMT1 is considered to be the maintenance DNA methyltransferase, as it preferentially binds hemimethylated CpGs during DNA replication and methylates the unmethylated daughter strand[69]. DNMT3A and DNMT3B are *de novo* methyltransferases that deposit methylated CpGs in novel locations. These *de novo* DNMTs are responsible for establishing tissue-specific methylation and are critical for differentiation[69,71]. Ten-eleven translocation family proteins (TET1, TET2, TET3) are responsible for active demethylation by biochemically oxidizing 5-methyl-cytosine to 5-hydroxymethyl-cytosine, which is further oxidized by TET enzymes to substrates that are removed through base-excision repair by the thymine-DNA glycosylase enzyme[70]. The role of hydroxymethylation in normal development is still a developing field of epigenetic research, with implications in cancer progression[72]. Lastly, there are numerous MBD proteins (e.g., MB1, MBD2, MeCP1/2) that act as readers of DNA methylation. These MBD proteins aid in suppressing gene expression as previously described and can also recruit histone-modifying enzymes that transform the chromatin configuration surrounding the methylated region[70].

2.1.2 Histone modifications

Within the nucleus, DNA is organized as chromatin, a complex of DNA and proteins that controls the accessibility and activity of the genome. The nucleosome is the fundamental unit of chromatin, consisting of 147bp of DNA wrapped around a protein octamer with two copies each of the histone proteins H2A, H2B, H3, and H4. The amino-terminal tails of the histones protrude from the nucleosome, and post-translational modifications to the tails, including acetylation, methylation, and ubiquitylation, influence DNA accessibility and serve as markers of chromatin activity that can be read by other enzymes[73]. Histone variants incorporated into nucleosomes at specific genomic locations provide additional functional control.

Characteristic histone modifications are found over active regulatory elements, transcribed genes, and repressed regions[74]. CGI promoters and active non-CGI promoters are flanked by nucleosomes modified with three methyl groups on the fourth lysine of the histone 3 tail (H3K4me3)[74], which is necessary for promoter activation. Enhancers are demarcated by mono-methylation of the same residue (H3K4me1). Nucleosomes flanking active regulatory regions also include the histone variants H2A.Z and H3.3, which may influence chromatin accessibility[74,75]. In contrast, actively transcribed gene bodies exhibit H3K36 tri-methylation (H3K36me3). Acetylation (ac) of histone 3 and 4 lysines activates regulatory elements: the negatively charged acetyl groups neutralize the positively charged histone tails, weakening their electrostatic interaction with the DNA phosphate backbone and increasing DNA accessibility[76].

Genes important for development are repressed in embryonic stem cells by the Polycomb complex PRC2, which deposits H3K27me3 modifications. This histone mark is recognized by the PRC1 complex, which mono-ubiquitylates histone H2A lysine 119, preventing RNA polymerase II transcript elongation[74]. Regulatory elements that exhibit both active histone modifications (H3K4me1 or H3K4me3) and the repressive H3K27me3 mark are considered "bivalent" or "poised" for activation. Prior to differentiation, the Trithorax complex removes the H3K27me3 mark from genes necessary for that lineage and mono-methylates H3K4 to activate those genes[74].

In contrast, heterochromatin is a stably repressed chromatin state characterized by H3K9 methylation[73]. It is mostly found over centromeres and silenced genes and is a primary mechanism of TE repression in somatic cells[77].

Similar to DNA methylation, histone modifications are regulated by reader, writer and eraser enzymes. Histone methyltransferases (writers) and demethylases (erasers) add and remove methyl groups from histone tails, respectively, and are generally specific to a particular lysine residue[73]. Histone acetyltransferases and deacetylases (HDACs) add and remove acetyl groups, respectively, and have broader specificity compared to histone methylation writers and erasers, including non-histone targets[74]. Many proteins are considered chromatin readers, interpreting existing modifications. Mutations in all three classes are prevalent in cancer and are discussed below (**Table 1**).

2.1.3 Chromatin accessibility

The accessibility of DNA to transcription factors and other transcriptional machinery is another indicator of regulatory activity. Active regulatory regions, including enhancers and promoters, have a lower nucleosome density, referred to as "nucleosome-depleted/free regions"[74]. In contrast, repressed regions and transcribed exons have a higher density of nucleosomes. Chromatin remodeling complexes (remodelers), including the SWI/SNF, ISWI, INO80, and

CHD complex families, are responsible for reorganizing nucleosomes to alter chromatin accessibility[74].

DNA methylation, histone modifications, and chromatin accessibility (summarized in **Figure 1**) act cooperatively to define active and repressed genome states**.** Many epigenetic modifiers include domains that recognize their own or correlated epigenetic features, allowing the maintenance or spreading of chromatin states through the recruitment of additional epigenetic modifications.

In addition to the epigenetic mechanisms discussed above, the quantity and localization of gene transcripts are influenced by a variety of post-transcriptional mechanisms, including non-coding RNA interactions and epigenetic modification of RNA. However, a detailed discussion of these phenomena is beyond the scope of this chapter and is reviewed elsewhere.

2.1.4 3D genome organization

To induce or constrain cooperativity of functional DNA sequences, eukaryotic cells evolved epigenetic mechanisms to tightly coil and organize DNA into specific configurations that occupy separate physical spaces within the nucleus. At the global level, chromosomes segregate into unique compartments. Inter-chromosomal contact has been documented, but the biological consequence of these interactions is still largely undefined. These large chromosomal compartments are divided into two categories based on activity. The “A” compartment represents active domains that have high transcription rates, while the “B” compartment represents silenced regions, such as heterochromatin[78]. The B compartment is primarily located at the nuclear lamina and overlaps with lamina-associated domains.

Within these large domains, chromosomes are subdivided into highly-conserved topologically associating domains (TADs), which shape the 3D chromatin structure to dictate which sequences interact with or are insulated from each other[79]. TADs range from 10kb to several hundred kb in length and are highly conserved across cell types and species[80]. DNA sequences within TADs form DNA loops (**Figure 1**), such that enhancers and promoters in loops have high contact frequency with each other but rarely interact with sequences outside the TAD. Furthermore, genes within the same TAD display coordinated transcription patterns, implying that these higher-order chromatin structures also define co-regulated transcriptional neighborhoods. Within these neighborhoods, an enhancer element >500kb upstream of a promoter can physically interact with the promoter to initiate transcription by coming into close proximity via loop formation[81]. Cell type-specific promoter-promoter interactions are also thought to create a promiscuous transcriptional hub that leads to increased gene expression[82].

DNA loop boundaries within TADs are enriched for binding sites of the insulator protein CTCF and cohesin complex proteins, which interact to create physical anchors that insulate one DNA loop from another[78,79]. Genetic experiments have revealed that loss of CTCF binding at the border of loops leads to disruption of loop formation and can alter the expression of genes within the loops[83,84]. Although the essential components for TAD boundaries have been identified, the mechanism that directs TAD formation and maintenance during differentiation is still under debate.

2.2 Epigenetic dysregulation in cancer

2.2.1 The cancer epigenome

In comparison to normal tissue, cancer tissue exhibits severe epigenetic dysregulation that influences the initiation and progression of the disease. In a normal somatic cell, ~80% of CpGs outside CGIs are methylated, but in many cancer types, this proportion falls to 40-60%[75]. Global levels of histone acetylation can also decrease[85]. Widespread DNA hypomethylation results in increased genomic instability, potentially through destabilization of pericentromeric heterochromatin[75]. DNA hypomethylation in cancer occurs in large blocks (0.05-10Mb) covering approximately one third of the genome, which may also exhibit histone acetylation and open chromatin[75]. These regions form early in cancer progression[75], apparently due to the dysregulation of large repressed regions that are typically sequestered at the nuclear lamina and are partially methylated in somatic cells[86]. In some cases, these lamina-associated regions overlap with gene-poor heterochromatin regions ("LOCKs", large organized chromatin K9 modifications), and they lose the heterochromatin marks H3K9me2 and H3K9me3 along with DNA methylation[86]. However, in other cases, LOCKs form over the hypomethylated blocks[74,75].

Widespread loss of DNA methylation and histone modification alterations can lead to the de-repression of silenced genes through promoter or enhancer reactivation (**Figure 2a**). A hallmark of carcinogenesis is the activation of oncogenes that give cancer its proliferative and stem cell-like characteristics. Epigenetic aberrations, or epimutations, can complement genetic mutations to establish a permissive epigenetic state that promotes cancer initiation[87,88]. Reactivated genes, such as cancer-testis genes whose expression is normally restricted to the germline, are often associated with pluripotency, proliferation, or germ cell development[68]. Examples of epigenetic reactivation of oncogenes are characterized in detail in numerous reviews[68,87,89].

Approximately 40% of differentially methylated regions in cancer include TEs[90], and one consequence of global DNA hypomethylation is the revival of TEs' inherent regulatory abilities. Many TEs are rich with transcription factor binding sites that serve as a template for novel transcription start sites or enhancers when epigenetically reactivated. Indeed, numerous TEs are exapted in cancer cells to provide alternative promoters for oncogenes, a process called "onco-exaptation"[91]. Furthermore, some TEs have the ability to "jump" or transpose when epigenetically reactivated. One example is LINE-1 elements, which are epigenetically silenced through DNA methylation in normal somatic cells. In multiple cancer types, epigenetically reactivated LINE-1 copies retrotranspose into novel locations in the genome, which can potentially lead to gene activation/disruption, splicing defects, and genome instability[92]. LINE-1 TEs also encode two open reading frames that are translated into proteins responsible for reverse transcription and transposition[92]. Whether these proteins impact carcinogenesis or can function as biomarkers of cancer is being extensively studied.

In addition to genome-wide hypomethylation, cancer exhibits focal hypermethylation of the CpG islands and shores of ~5-10% of CGI gene promoters[75,85], which are constitutively unmethylated in normal somatic cells (**Figure 2b**). This alteration is accompanied by a loss of active histone modifications and nucleosome positioning over the transcription start site, which reduces or eliminates expression from the allele[75,85]. Epigenetic silencing of tumor suppressor genes, such as p16^{ink4a} in lung cancer[75], can serve as one of two hits to the gene under Knudsen's two-hit hypothesis, complementing a genetic mutation or deletion that knocks out the other allele. This is further supported by the observation that aberrant promoter methylation of tumor suppressor genes is mutually exclusive with deactivating mutations of the same gene[74]. Promoter methylation can also silence DNA repair genes, leading to a drastic increase in the number of mutations in the genome. For instance, loss of expression of the DNA repair genes MLH1 or MGMT through promoter methylation in colorectal cancer causes a microsatellite instability phenotype and a greater incidence of G-to-A mutation, respectively[74,75]. Promoter methylation can also promote tumor metastasis, for instance, through reducing the expression of CDH1, which encodes E-cadherin[74].

Genes that undergo promoter hypermethylation in cancer are biased towards those under Polycomb repression in stem cells. These genes are important for differentiation and have constitutively unmethylated promoters in almost all cell types, even when they are included in repressed regions[75]. Indeed, while some of the aberrantly methylated genes are required in the tumor cell-of-origin, most are not[74], and they are frequently found in the DNA hypomethylated blocks that emerge during the breakdown of repressed domains in cancer[74,75]. In contrast to this seemingly stochastic dysregulation of the epigenome, mutations impacting cellular and signaling pathways can also epigenetically reprogram the cell to a more pluripotent state in a controlled manner. For instance, KRAS mutations downregulate TET enzymes, increasing methylation at tumor suppressor promoters[86].

Epigenetic dysregulation in cancer may also result from a loss of imprinting (LOI) due to DNA methylation changes. Imprinted genes (~1% of all genes) have different expression levels based on their parent-of-origin, which is typically mediated by DNA methylation. Many imprinted genes are involved in growth and metabolism, and changes in their expression level can result in uncontrolled growth and proliferation. Loss of methylation over the promoter or enhancer of an imprinted gene could increase its expression level through re-activation of expression. In the case of the imprinted gene IGF2 (insulin-like growth factor 2), DNA hypermethylation of the maternal allele of the nearby insulator H19-ICR blocks binding of CTCF, allowing the IGF2 promoter to aberrantly contact an upstream enhancer and doubling the expression level of the gene, resulting in increased cellular proliferation[75,86].

2.2.2 Aberrant chromatin structure

Disruption of normal TAD formation via deletion or insertion of CTCF binding sites can result in a disease phenotype[84,93]. For example, a subset of gliomas is characterized by gain-of-function mutations in the IDH gene, which converts production of α-keto-glutarate, a metabolite essential for TET2 function, to 2-hydroxyglutarate, a competitive inhibitor of TET2[74]. This ultimately leads to high global levels of CpG methylation due to the suppression of TET enzyme function. CTCF is a methylation-sensitive transcription factor, and IDH-mutated glioma cells display various CTCF binding abnormalities and deregulated TADs. One particular disruption abnormally couples the oncogene PDGFRA's promoter and a rogue FIP1L1 enhancer (located >500kb away) into the same TAD, resulting in a novel promoter-enhancer interaction that up-regulates PDGFRA expression and consequently increases cell proliferation (**Figure 2c)**[84]. This provides an example of how genetic and epigenetic alterations can interact to affect higher-order chromatin interactions that accelerate the tumorigenic phenotype in cancer.

2.2.3 Mutations in epigenetic regulators

Mutations in epigenetic regulators, the readers, writers, erasers, and remodelers discussed above, are extremely common across cancers[75] and are summarized in **Table 1**. In many cases, these alterations occur early in tumor development and may contribute to disease initiation[86]. The epigenetic implications of the mutation depend on the affected enzyme and the tumor type, as the tumor's cell-of-origin influences its initial epigenetic landscape.

Mutations to readers and writers of DNA methylation are frequent in hematological malignancies. DNMTA mutation leads to widespread DNA hypomethylation[94], repeat destabilization, and telomere lengthening[74], while TET2 mutation leads to hypermethylation of lineage-specific enhancers[94]. Several cancer types, including colorectal cancer and glioma, have well-characterized CIMP (CpG island methylator phenotype) subtypes that exhibit hypermethylation of promoter CGIs. In glioma, this phenotype is the result of the IDH1 mutation described above. Although the genes affected in colorectal cancer and glioma are different, they are frequently involved in developmental regulation[74,75]. IDH1 and TET mutations are mutually exclusive, and in AML, they are sufficient to drive cancer progression[75].

Histone modifying enzymes are also the targets of mutation or chromosomal rearrangements in several cancer types. Gain-of-function translocations that fuse histone modifiers to other proteins are particularly common in AML and include the histone methyltransferases MLL and NSD1 and the histone acetyltransferases CREBBP and EP300[74,75]. Fusion of MLL to recruitment proteins incorrectly targets H3K4 methylation to HOX gene promoters and upregulates their expression in cancer, particularly HOXA9[74]. NSD1 fused to NUP98 similarly activates the HOXA gene cluster via increased gene body H3K36me3[74]. The H3K27 methyltransferase EZH2 is one of the most commonly disrupted genes in cancer, reflecting the important role dysregulation of H3K27me3 plays in tumor progression. In addition to amplification and mutation, EZH2 expression can be upregulated by miR101 deletion[74].

Histones themselves can be mutated in a way that prevents or mimics methylation[74,75]. G34 mutations in H3F3A, which encodes the histone variant H3.3, are loss-of-function and result in DNA hypomethylation, genomic instability, and telomere lengthening[74]. In contrast, K27M mutations are gain-of-function and may mimic H3K27 di-methylation, a mark of Polycomb repression[74]. Mutations to H3F3A are mutually exclusive with mutations in ATRX and DAXX, which load H3.3 into nucleosomes[74].

Epigenetic modifications can in turn influence the tumor mutation rate. LOCKs are associated with a higher single nucleotide variant frequency in cancer compared to regions of open chromatin[74], potentially due to their epigenetic dysregulation. Hydrolytic de-amination of methylated cytosines results in a C-to-T mutation[75], and this conversion rate is especially high in rapidly proliferating tissues, where DNA repair enzymes are unable to keep pace with the rate of mutation. In fact, this mechanism results in a quarter of all TP53 mutations[74]. Methylated cytosines are also more likely to form adducts with carcinogens in cigarette smoke and pyrimidine dimers in response to UV exposure[75], increasing the mutation rate over methylated regions.

While many epigenetic programs are established during embryogenesis, the epigenetic landscape continues to be shaped by aging and dynamic interactions between the host and its environment throughout our lifespan. Many of the global epigenetic changes observed in cancer, including large hypomethylated blocks, are also observed in healthy tissue from elderly individuals, suggesting that aging predisposes cells to epigenetic dysregulation[86]. Environmental exposures such as diet, chronic inflammation, and smoking can also influence disease progression through epigenetic mechanisms[86]. Thus, genetic and epigenetic mechanisms interact with each other and the environment to promote oncogenic cellular states.

2.3 Epigenetic cancer therapy

Epigenetic therapy is currently being explored as a treatment option for a variety of cancers. Although only a few compounds are currently FDA-approved as cancer therapeutics (**Table 2**), numerous compounds are undergoing clinical trials. The three major classes of epigenetic therapy discussed here are all non-specific, targeting ubiquitous epigenetic pathways to reverse cancer-specific epigenetic alterations. However, targeted epigenetic therapies, which are specific to a particular gene or genomic location, are gaining interest as well.

2.3.1 DNA methyltransferase inhibitors

DNA methyltransferase inhibitors (DNMTi) decrease the genome-wide level of cytosine methylation, reversing the DNA methylation gains that occur at promoters in cancer. The currently approved compounds, 5-azacytidine (Vidaza/Mylosar) and decitabine (Dacogen), are nucleoside analogues with a modified cytosine C5 ring. They are incorporated into DNA and/or RNA and covalently bind all three DNMTs, blocking their methyltransferase activity and targeting them for proteasomal degradation[73,75]. DNMT inhibitors are non-specific, so while they reverse promoter hypermethylation, they impact normally methylated and hypomethylated regions as well. While this effect could further destabilize the cancer genome, it may also be crucial to the drugs' efficacy, as discussed below.

Dacogen and Vidaza are currently FDA-approved for the treatment of myelodysplastic syndrome[95,96]. In contrast to many compounds, the drugs are least toxic *and* maximally effective at low doses[73,85]. Other classes of DNMTi, including non-nucleoside analogues, are currently being explored in clinical trials.

2.3.2 Histone modifier inhibitors

The discovery that HDACs are overexpressed in cancer and that genetic knockdown of HDAC proteins led to decreased viability and proliferation in cancer cells has ignited an extensive search for HDAC inhibitors (HDACi) that could be potential cancer therapeutic drugs[97–99]. HDACs in humans are classified into four groups based on homology and molecular mechanism of action: zinc-dependent HDACs (Class I, Class II, Class IV) and NAD-dependent HDACs (Class III). Several HDACis can act either globally, impacting multiple classes of HDACs, or specifically, suppressing a single class or a particular HDAC enzyme. Only four HDACis are FDA-approved: vorinostat (SAHA), belinostat, and romidepsin (depsipeptide) for the treatment of refractory cutaneous and peripheral T-cell lymphomas[99], and panobinostat for multiple myeloma[100]. Numerous other HDACi agents are currently in Phase II and Phase III trials for various hematological cancer and solid tumors and are showing encouraging results of superior clinical activity, lower toxicity, and better prognosis relative to conventional HDACi treatments[99,101]. For example, entinostat, a Class I HDACi, is in a Phase III trial for hormone-receptor-positive breast cancer, while pracinostat, a pan-HDACi, is under Phase III trial for both AML and myelodysplastic syndrome[102].

The counterpart of HDACs, histone methyltransferases, have garnered much attention recently as a potential therapeutic target in cancer. Histone methyltransferase genes, such as EZH2 and MLL, are often hit with mutations or translocations to create fusion proteins that are associated with irregular histone methylation levels in cancer[68,101]. To modulate increases in histone methylation, multiple histone methyltransferase inhibitors are in early-stage clinical trials to measure efficacy and viability. Currently, molecular inhibitors targeting the DOT1L, EZH2, and LSD1 histone methyltransferases are in Phase I and II trials for numerous hematological cancers, such as AML, non-Hodgkin lymphoma, and multiple myeloma[101]. Furthermore, there is an ongoing search for viable molecular inhibitors for histone demethylases, primarily LSD-1 and JMJD, but no preclinical trials are currently underway. For more information, molecular inhibitors targeting histone modifications is extensively reviewed in Shortt et al. 2017[101].

2.3.3 Bromodomain inhibitors

Bromodomain inhibitors (BETi) target the BET (bromodomain and extraterminal domain) protein family of histone acetyl lysine readers. The BET family proteins, BRD2, BRD3, BRD4, and BRDT, each contain two bromodomains that recognize the acetylated lysines of histones H3 and H4[76]. The original class of BET inhibitors are the thienodiazepine compounds, which are acetyl-lysine mimetics that bind the bromodomains of all BET proteins. JQ1 is the most well-studied member of this class, targeting both BRD4 and the NMC-specific fusion BRD4-NUT[76]. Although JQ1 is not being investigated in clinical trials, several members of the class are.

BETi are well-tolerated due to their surprisingly cancer-specific effects[76]. BETi inhibit the activity of BRD4, which is found at active promoters and enhancers and is essential for transcriptional elongation[76]. In cancer, BRD4 is particularly enriched at tumor-specific super-enhancers that drive oncogene expression, and BETi appears to have the largest effect on the oncogenes associated with those enhancers[76]. Thus, although BETi target a ubiquitously expressed protein, they have the largest impact on cancer cells.

2.3.4 Mechanism of action of epigenetic therapies

Epigenetic therapies may exert anti-neoplastic properties through several mechanisms. For years, the primary mechanism of DNMTi was thought to be the re-activation of epigenetically silenced tumor suppressor genes, such as MLH1, RB, and p16, restoring expression to these silenced alleles by removing repressive epigenetic marks from their promoters (**Figure 3a**)[73]. In contrast, HDACi acts through diverse molecular mechanisms to either suppress or activate gene expression in multiple biological pathways, which is described in detail in other reviews[99,103,104]. In brief, the addition of HDACi tilts the gene expression pattern to favor expression of pro-apoptotic genes while suppressing proliferative genes, leading to tumor apoptosis. HDACi treatment has also been associated with upregulation of immunomodulatory genes, such as expression of MHC class I and II, to make cancer cells more immunogenic[105]. Furthermore, cancer cells treated with HDACis show an accumulation of acetylated non-histone proteins, such as Hsp90, which can impact gene regulation through de/stabilization and de/activation of certain proteins[104].

Recently, however, it has been recognized that epigenetic therapies also have epigenetic effects outside of silenced promoters. Cryptic or non-canonical promoters can be relieved of repressive marks, leading to the upregulation of chimeric or otherwise immune-privileged transcript isoforms, such as cancer-testis antigens[106]. If these transcripts are translated, they can form immunogenic neoantigens that trigger an immune response against the tumor[107]. Many TEs harbor cryptic regulatory elements that are upregulated upon combinatorial DNMTi and HDACi treatment, leading to the formation of thousands of previously non-annotated transcripts through splicing into downstream genes (**Figure 3b**)[108]. These novel transcripts can be translated into chimeric proteins or completely new peptides. Additionally, loss of epigenetic repression over endogenous retroviruses can lead to the production of double-stranded RNA, which can trigger a type I interferon anti-viral response against the cell (**Figure 3c**)[59]. Large-scale DNA hypomethylation also increases genomic instability[74], and further reduction of DNA hypomethylation by DNMT inhibition may be detrimental to the cell. Thus, rather than directly altering the expression of canonical tumor suppressor genes and oncogenes, epigenetic therapies may direct the immune system against the tumor.

2.3.5 Combination therapies

Although epigenetic therapies are approved for the treatment of hematological malignancies as single agents, they have shown lower efficacy in patients with solid tumors[75,85]. However, more recent trials have combined epigenetic therapy with other cancer therapies such as cytotoxic chemotherapeutics and immune checkpoint inhibitors in an attempt to harness their synergistic effects. Epigenetic therapy and checkpoint inhibitor immunotherapy show encouraging potential. A clinical trial of advanced, pre-treated non-small cell lung cancer patients suggested that epigenetic therapy primes tumors for additional treatment: although few patients responded to epigenetic therapy alone (2 of 65), it increased the response rate to subsequent anti-PD-L1 immunotherapy (3 of 6 responded vs. 16-17% with PD-L1 alone)[106]. BET inhibitors have also been shown to synergize with HDACis in *in vivo* models[76]. It is possible that novel transcripts, induced by epigenetic therapy, can be translated into peptides to act as neoantigens, which increase the immunogenicity of cancer cells.

In conclusion, dysregulation of the normal epigenetic landscape is a critical step in carcinogenesis, influencing and being influenced in turn by genetic abnormalities and the environment. Further investigation of the cancer epigenome is being enabled by advances in next-generation sequencing technologies, which will enhance our understanding of cancer and potentially open new avenues of treatment through epigenetic modification of tumor cells.

2.4 Figures and tables

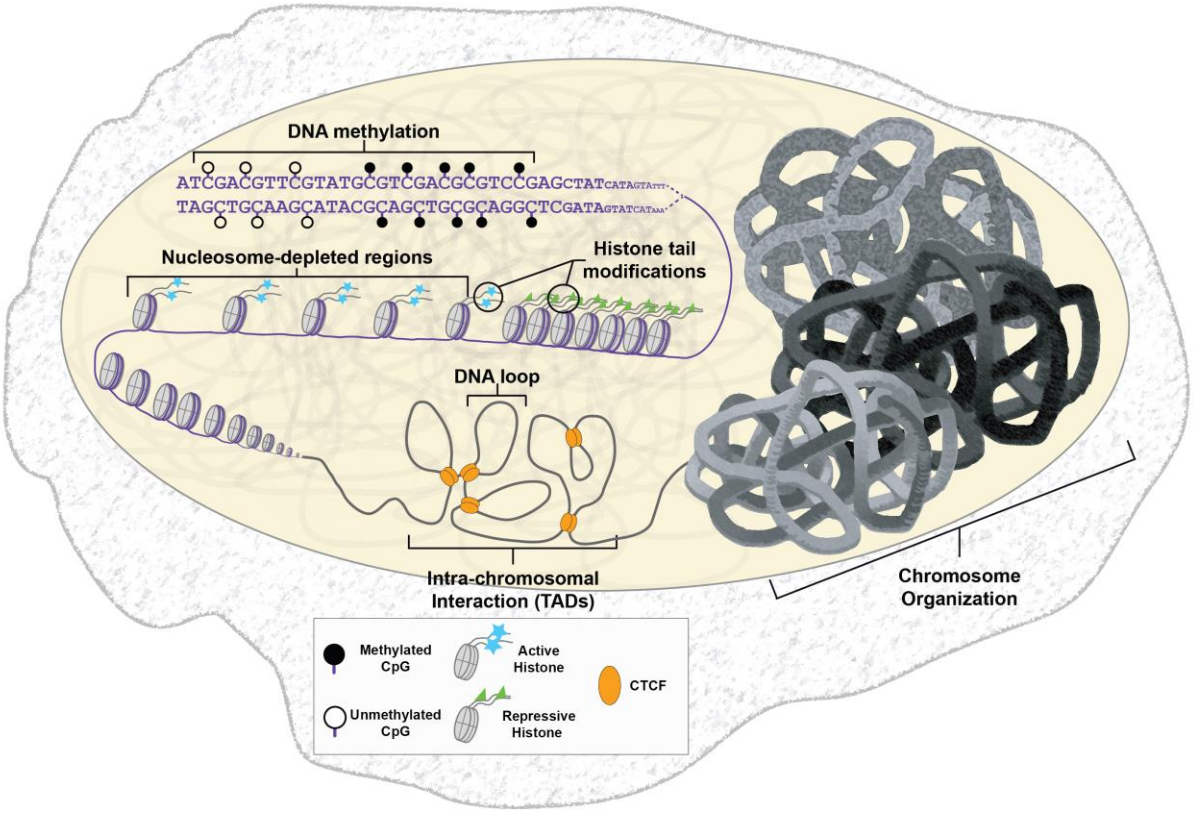

DNA methylation
ATCGACGTTCGTATGCGTCGACGCGTCCGAGCTATCATAGTATTT
TAGCTGCAAGCATACGCAGCTGCGCAGGCTCGATAGTATCATAAA
Nucleosome-depleted regions
Histone tail modifications
DNA loop
Intra-chromosomal Interaction (TADs)
Chromosome Organization
Methylated CpG
Active Histone
CTCF
Unmethylated CpG
Repressive Histone

Figure 1: Common epigenetic modifications in mammalian cells

A schematic representation of epigenetic control in mammalian cells, with emphasis on DNA methylation, histone post-translational modifications, chromatin accessibility, and higher-order 3D chromatin strucutre.

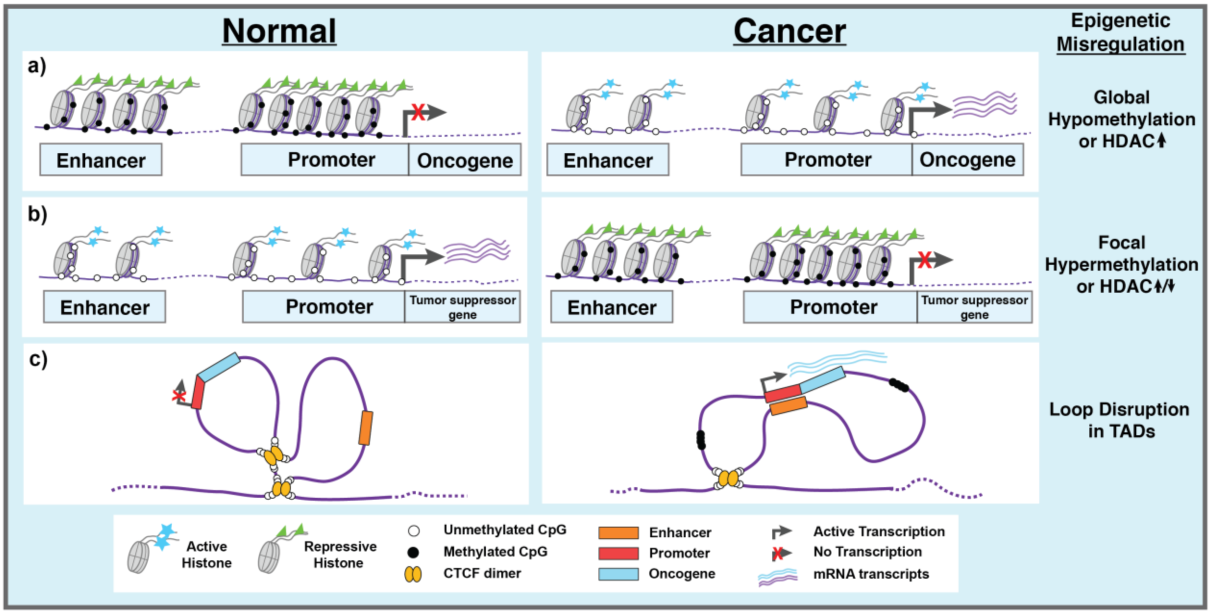


Figure 2: Epigenetic aberrations in cancer

a) Global hypomethylation and/or loss of repressive histone methylation in regulatory regions leads to activation of oncogenes in cancer. **b)** Focal hypermethylation and/or misregulation of histone acetylation represses tumor suppressor gene expression in cancer. **c)** Methylation of CTCF loop anchor sites disrupts proper DNA loop formation to activate oncogene expression through rogue enhancer recruitment to oncogene promoters.

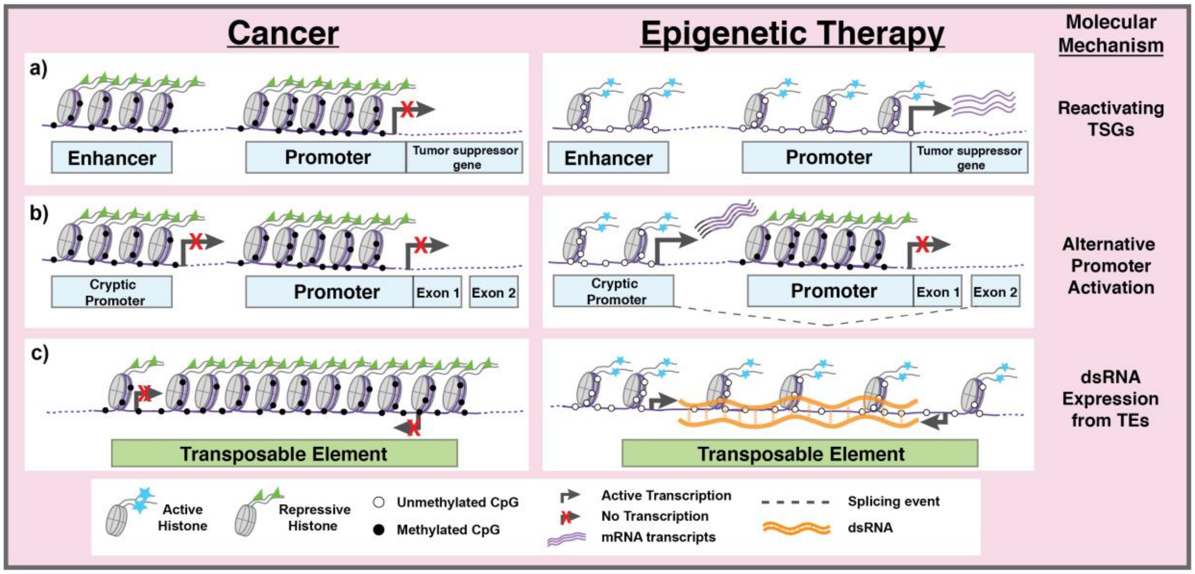


Figure 3: Molecular mechansims of epigenetic therapy in cancer

a) Global hypomethylation or gain of active histone modifications can reactivate pro-apoptotic or tumor suppressor genes in cancer. **b)** Revival of cryptic promoter activity generates novel or chimeric transcripts that can translate into neoantigens and trigger an immune response. **c)** Epigenetically reactivated transposable elements produce dsRNA, which induces an anti-viral pathway that slows proliferation and increases the immunogenicity of cancer cells.

Table 1. Epigenetic regulators mutated in cancer

<u>Category</u>	<u>Gene</u>	<u>Cancer type</u>	<u>Ref</u>
DNA methyltransferase	DNMT1	Colorectal	2,3
	DNMT3A	T-cell lymphoma, AML, myeloid malignancies	1,2,3*
DNA demethylase	TET1	Colorectal, AML	1,2,3
	TET2	Colorectal, bladder, B-lymphoma (FL), T-cell lymphoma, AML, myeloid malignancies	1,2,3*
	TET3		2
DNA methylation reader	MDB1/MBD2/MBD4	Colorectal, lung adenocarcinoma, breast, melanoma	2,3
Histone methyltransferase	MLL (KMT2A) *(H3K4)*	Gastric, bladder, lung, liver, colorectal, breast, ALL, AML	1,2,3
	MLL2 (KMT2B) *(H3K4)*	Breast, kidney (clear cell), lung, prostate, head and neck, B-lymphoma (DLBCL, FL), non-Hodgkin lymphoma, medulloblastoma	1,2,3
	MLL3 (KMT2C) *(H3K4)*	Gastric, breast, bladder, liver/hepatocellular, pancreas, medulloblastoma	1,2,3*
	MLL4 (KMT2D) *(H3K4)*		2*
	SETD1A (KMT2F) *(H3K4)*	Gastric adenocarcinoma, breast, CLL	2
	PRDM9 *(H3K4)*	Head and neck squamous cell carcinoma	1,2
	MEN1 *(MLL complex)*	Pancreatic neuroendocrine	1*
	EZH2 *(H3K27)*	Colorectal, gastric, breast, bladder, lung, medulloblastoma, melanoma, B-lymphoma (FL, DLCBL), Burkitt lymphoma, T-cell leukemia, head and neck squamous cell carcinoma, T-ALL, AML, myeloid malignancies	1,2,3*
	SUZ12/EED/ JARID2 *(PRC2 complex)*	T-ALL, myeloid malignancies, prostate, meningioma	1
	NSD1 *(H3K36)*	AML, head and neck squamous cell carcinoma, endometrial carcinoma, melanoma, colorectal, multiple myeloma	1,2
	NSD2 (WHSC1/ MMSET) *(H3K36)*	Multiple myeloma, pediatric ALL, colorectal, melanoma	1,2
	SETD2 *(H3K36)*	Clear cell renal cell carcinoma, T-ALL, high-grade glioma	1,2*

Histone demethylase	KDM1A (LSD1) *(H3K4/K9)*	Prostate	3
	KDM2B *(H3K4/K36)*	B-lymphoma (DLBCL)	2
	KDM5C (JARID1B/C) *(H3K4)*	Breast, kidney (clear cell), meningioma	1,2*
	KDM6A (UTX) *(H3K27)*	Kidney (renal cell carcinoma), bladder (transitional cell), esophageal squamous cell carcinoma, multiple myeloma, myeloid malignancies, meningioma, medulloblastoma, prostate, breast, lung, pancreas, colon, uterus, brain	1,2,3*
Histone methyl reader	ING1 *(H3K4me3)*	Melanoma, esophageal squamous cell, ALL	2
	PHF6	AML, T-ALL	1,2
Histone acetyltransferase	CREBBP (CBP)	Bladder (transitional cell), lung (SCLC), B-lymphoma (DLBCL, FL), Burkitt lymphoma, ovarian, relapsed ALL, medulloblastoma, AML	1,2,3*
	EP300	Endometrial serous, bladder (transitional cell), lung (SCLC), B-lymphoma (DLBCL, FL), T-ALL, pancreatic, breast, colorectal, AML	1,2,3*
	PCAF *(P300/CREBBP partner)*	Epithelial	3
Histone deacetylase	HDAC2	Colorectal, gastric, endometrial	2,3
	HDAC4	Breast adenocarcinoma	2
	HDAC9	Prostate adenocarcinoma	2
	SIRD1, HDAC5/7a		3
	P400 *(NuA4 complex)*		3
Histone lysine acetyl readers	BRD3/4	NMC	1,2
	BRD4	Burkitt lymphoma	1
	BRD8	Liver/hepatocellular	1,2
Histone deubiquitinase	BAP1 *(H2AK119)*	Kidney (clear cell), myeloid malignancies, mesothelioma, melanoma	1*
	ASXL1 *(PR-DUB component)*	Prostate, AML, myeloid malignancies	1*
Histone	HIST1H1B *(H1)*	CLL, B-lymphoma (FL), colorectal	2
	HIST1H1C *(H1)*	B-lymphoma (DLBCL, FL)	1
	HIST1H1E *(H1)*	CLL	1

	HIST1H3B *(H3.1)*	B-lymphoma (DLBCL), glioma (DIPG), GM, pediatric glioblastoma	1,2*
	H3F3A *(H3.3)*	Pediatric glioblastoma, GBM, glioma (DIPG), CNS primary neuroendocrine, giant cell tumor of bone	1,2*
	H3F3B *(H3.3)*	Chondroblastoma	2*
	HIST1H4B *(H4)*	Liver	1
Chromatin remodeler	ATRX	Pancreatic neuroendocrine, GBM (pediatric glioblastoma), medulloblastoma, neuroblastoma	1,2*
	DAXX	Pancreatic neuroendocrine, GBM (pediatric glioblastoma)	1,2*
	SMARCA2		3
	SMARCA4	Melanoma, Burkitt lymphoma, lung adenocarcinoma, medulloblastoma	1,2,3*
	SMARCB1	Pediatric malignant rhabdoid tumor, mesothelioma, medulloblastoma, CNS primitive, meningioma	1,2,3*
	SMARCD1	Breast	1,2
	SMARCE1	Clear cell meningioma	1,2
	ARID1A	Numerous epithelial, Burkitt lymphoma, ovarian (clear cell carcinoma), melanoma, medulloblastoma, neuroblastoma, hepatocellular carcinoma, breast, lung adenocarcinoma, colorectal	1,2,3*
	ARID1B	Breast, liver (hepatocellular carcinoma), melanoma, medulloblastoma, neuroblastoma	1,2
	ARID2	Breast, lung, liver (hepatocellular carcinoma), pancreatic adenocarcinoma, melanoma	1,2,3*
	BRD7	Bladder TCC	3
	PBRM1	Clear cell renal carcinoma, breast	1,2,3*
	CHD1	Breast, lung, prostate (ETS-negative)	1
	CHD2	CLL	1,2
	CHD4	Serous endometrial	1,3
	CHD5	Neuroblastoma, glioma, breast, lung, colon, ovary, prostate	2,3
	CHD6	Bladder	1
	CHD7	Medulloblastoma, gastric, colorectal	1,3
	CHD8	Lung	1
	CHD1/CHD3/CHD4/ CHD6/CHD7/CHD8	Gastric, colorectal, prostate, breast, bladder, serous endometrial	2
	PHF23	AML	2
Insulators	CTCF	Breast, T-ALL	1
	RAD21		*

[1], Reference 7, Figure 3; [2], Reference 17, Table 2; [3], Reference 8, Table 2 (mutations only); *Confirmed as causal somatic mutation in the COSMIC Cancer Gene Census (http://cancer.sanger.ac.uk/census)

Target or complex is specified in parentheses in italics in the Gene column

Table 2. FDA-approved epigenetic therapies for cancer treatment

Category	Compound	US brand name	Approved indications
DNA methyltransferase inhibitor (DNMTi)	Azacitidine (5-azacytidine)	Vidaza/Mylosar	Myelodysplastic syndrome
	Decitabine (5-aza-2'-deoxycytidine)	Dacogen	Myelodysplastic syndrome
Histone deacetylase inhibitor (HDACi)	Belinostat	Beleodaq	Relapsed or refractory peripheral T-cell lymphoma
	Panobinostat (with bortezomib and dexamethasone)	Farydak	Multiple myeloma, at least two prior treatments
	Romidepsin (depsipeptide)	Istodax	Cutaneous T-cell lymphoma, at least one prior systemic therapy
	Vorinostat (suberoylanilide hydroxamic acid)	Zolinza	Relapsed or refractory cutaneous T-cell lymphoma

From the NCI Drug Dictionary[95], Human Epigenetic Drug Database, Disease list (http://hedds.org/index.jsp)[96]

Chapter 3: mLRT: detecting differentially methylated regions by integrating 5hmC and 5mC signals from WGBS and TAB-seq data

Wei Wang[1*], Nan Lin[1,2†], Hyo Sik Jang[3,4*], Ting Wang[3,4]

[1]Department of Mathematics and Statistics, Washington University in St. Louis, St. louis, 63130, USA.

[2]Division of Biostatistics, Washington University School of Medicine, St. Louis, 63110, USA.

[3]Department of Genetics, Washington University School of Medicine, St. Louis, 63110, USA.

[4]The Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, 63110, USA.

[*]Contributed equally.

[†] Corresponding author

Author Contribution: HSJ and TW generated WGBS and TAB-seq data. WW and NL designed and performed all the statistical tests to call differential methylation. HSJ and TW performed functional characterization and analysis of DMRs. NL and TW supervised the work for this project. WW and HSJ wrote the manuscript with feedback from NL and TW. All authors have read the manuscript and agree with its contents.

This manuscript is currently in preparation for submission.

3.1 Abstract

Motivation: DNA methylation is an epigenetic mechanism that occurs by adding methyl groups to the DNA molecule. It is known to play a critical role in gene regulation, development, and tumorigenesis. It is thereby very important to study differential methylation patterns between two targeted samples for comparison, e.g., adult vs. fetal tissues. Recent advances in next generation sequencing technologies make it possible to distinguish between 5-methylcytosine (5mC) and 5-hydroxymethylcytosine (5hmC). Considering that 5hmC levels are usually very small and hence the difference regarding 5hmC between two samples could rarely be detected, most methods identify differentially methylated regions (DMRs) by comparing sum of 5hmC and 5mC levels between two samples, e.g., methods only using WGBS (Whole Genome Bisulfite Sequencing) data. However, recent study shows the necessity of integrating 5mC and 5hmC signals in differential analyses.

Results: We combine WGBS and TAB-seq (Tet-Assisted Bisulfite Sequencing) data to investigate if jointly testing the differences in 5hmC and 5mC levels would be more powerful regarding differential methylation analysis. Simulation studies show that our method of jointly testing 5hmC and 5mC levels using a likelihood ratio test (mLRT) gains more power to detect DMRs compared to methods only using WGBS data. We also compare our mLRT to natural alternatives for jointly testing, i.e., mFET (Fisher exact test based on maximum likelihood estimators for 5hmC and 5mC levels) and nFET (naive Fisher exact test), at CpG level by simulation. It shows that Type I error is controlled by mLRT while mFET could not, though mFET gains more power of detection at the CpG level. mLRT also outperforms nFET in terms of power. The application to adult and fetal mouse cortex data shows that mLRT from the combination of WGBS and TAB-seq data gives much more detections at the region level than

Fisher exact test (FET) from WGBS data only. Furthermore, we applied mLRT to young and old mouse frontal cortex samples and report novel DMRs with potential biological implications related to aging in the brain.

Availability: mLRT is freely available on the website at https://github.com/nihonoui/mLRT

3.2 Introduction

DNA methylation is a common mechanism of epigenetic regulation in eukaryotic organisms ranging from fungi to mammals. Numerous studies have been carried out to locate CpG sites where DNA methylation plays a role in gene regulation, development, and tumorigenesis[109]. Differentially methylated regions (DMRs) are groups of adjacent CpG sites that are mostly differentially methylated. In mammals, the bulk of DNA methylation in CpG context occurs as 5-methylcytosine (5mC). The other major epigenetic modification of cytosines is the oxidation product of 5mC, 5-hydroxymethylcytosine (5hmC). 5hmC coexists with 5mC in a range of mammalian cell populations and was also found to be involved in gene regulation[110]. Understanding the differential methylation patterns requires information of both 5mC and 5hmC in the genome.

Bisulfite sequencing (BS-seq) is a widely used sequencing technology for genome-wide DNA methylation profiling. The most popular protocols for BS-seq include Reduced Representation Bisulfite Sequencing (RRBS)[111] and Whole Genome Bisulfite Sequencing (WGBS)[112]. The yield of methylation from BS-seq is the sum of 5hmC and 5mC levels, hence BS-seq cannot distinguish between 5mC and 5hmC[113]. With recent innovations in next generation sequencing technologies, Oxidative Bisulfite Sequencing (oxBS-seq)[114] and Tet-Assisted Bisulfite Sequencing (TAB-seq)[115] could provide high-throughput single-base resolution measurements of 5mC and 5hmC, respectively. Methods for estimating 5hmC and 5mC levels at CpG sites from different combinations of next generation sequencing data are proposed in Quy *et al.* 2013[116] (combing any two of BS-seq, TAB-seq or oxBS-seq, or all three when available) and Xu *et al.* 2016[117] (combing BS-seq and oxBS-seq). After obtaining the methylation information of CpG sites, the typical downstream analysis is differential methylation analysis. For DMR detection, Sun *et al.* 2014[118] proposed model-based analysis of bisulfite sequencing data, MOABS, in which the significance of the differential methylation between two samples is represented by a metric named 'credible methylation difference'. Äijö *et al.* 2016[119] presented an integrative hierarchical model (Lux) from various combinations of sequencing data for detection of differential methylation based on Bayes factors. Shafi *et al.* 2017[120] provided a survey of the approaches for identifying differential methylation using bisulfite sequencing data.

In this paper, we integrate WGBS and TAB-seq data to investigate if jointly testing the differences in 5hmC and 5mC levels would gain more power in the analysis of differential methylation. To the best of our knowledge, there is no existing literature analyzing differential methylation by jointly testing 5hmC and 5mC levels. When inputs are count values, naive ways to jointly test 5hmC and 5mC levels involve Fisher exact test (FET)[121]. The challenge of performing FET in this scenario arises from missing count values in the contingency table observed from the mixture of WGBS and TAB-seq data. In order to perform FET, we need two steps: imputation for missing counts, and then applying FET to the complete contingency table. However, data for methylation analysis usually contain a very small number of replicates, and thereby the imputation for missing counts can get highly distorted. Instead, we present a likelihood ratio test (mLRT) for simultaneously comparing 5hmC and 5mC levels between two samples in CpG sites from WGBS and TAB-seq data. We compare our mLRT at the CpG level by simulation to two types of two-step FETs for jointly testing 5hmC and 5mC levels. We also compare mLRT with FET using WGBS data only. The simulation results indicate that mLRT outperforms the two two-step FETs at CpG level in terms of size and power. We also apply mLRT to a real dataset, i.e., adult and fetal mouse cortex data, to investigate biological interpretation of differential methylation derived from the four statistical methods.

3.3 Methods

WGBS and TAB-seq were constructed from genomic DNA of frontal cortex tissue of 7-week-old and 79-week-old male C57BL/6J mouse (Jackson Laboratory, 000664) using 5hmC TAB-seq Kit (WiseGene, K001), EZ DNA Methylation-Gold Kit (Zymo, D5005) and Accel-NGS Methyl-Seq DNA Library Kit (Swift Biosciences, 30024). WGBS and TAB-seq libraries were sequenced on Illumina NovaSeq 6000 platform and aligned to the mm9 reference genome using Bismark[122]. 5hmC values were adjusted based on the glucosylation protection rate and TET oxidation rate as previously described[123].

We assume that WGBS and TAB-seq data at a CpG site are independent and follow binomial distributions. Our model at the CpG level is then given by, for the *j*th replicate,

$$M_W^{g,j} \sim Bin(N_W^g, p_g) \text{ and } M_T^{g,j} \sim Bin(N_T^g, p_{g1}),$$

for $j = 1, \ldots, n_g$.

Notations used in this model are given as follows.

- The index $g \in \{A,B\}$ denotes two biological conditions, *A* and *B*. In the analysis of differential methylation, the two samples for comparison usually correspond to two different biological conditions, e.g., adult vs fetal.
- n_g is the number of replicates under condition *g*.
- Proportions of two different cytosine methylations, i.e., 5hmC and 5mC, are denoted by p_{g1} and p_{g2} for $g \in \{A, B\}$, respectively, and $p_g = p_{g1} + p_{g2}$.
- The index k denotes the choice of two next generation sequencing technologies WGBS and TAB-seq, and $k \in \{W, T\}$ with 'W' for WGBS and 'T' for TAB-seq.
- For the *j*th replicate, $M_k^{g,j}$ and $N_k^{g,j}$ are the count of 'C' read-outs and the count of total 'C' and 'T' read-outs from sequencing technology *k* under condition *g*, respectively.

We need to point out that for data with applications, our current model does not consider biological variations, which are commonly considered in Sun *et al.* 2014[118] and Äijö *et al.* 2016[119].

To study differential methylation patterns between Conditions A and B, our approach starts with jointly testing 5hmC and 5mC levels, i.e., testing the hypotheses,

$$H0 : (pA1, pA2) = (pB1, pB2) \ vs. H1 : (pA1, pA2) \neq (pB1, pB2). (1)$$

Since the inputs are counts, the natural way of testing (1) is performing FET on a contingency table. The underlying methylation data for the *j*th replicate at a CpG site can be expressed as the following 2 × 3 contingency table (Table 1), however, combining WGBS and TAB-seq data, we are not able to observe all cells in Table 1. For WGBS, we observe the sum of 5hmC and 5mC counts, i.e., for $g \in \{A, B\}, m_{g1}^{j} + m_{g2}^{j} = M_{W}^{g,j}$. For TAB-seq, we observe 5hmC counts, i.e., $m_{g1}^{j} = M_{T}^{g,j}$.

Table 1. The underlying methylation data

	5hmC	5mC	Unmethylated
Condition *A*	m_{A1}^{j}	m_{A2}^{j}	m_{A3}^{j}
Condition *B*	m_{B1}^{j}	m_{B2}^{j}	m_{B3}^{j}

Instead of performing FET by imputation for missing counts in Table 1, we derive the likelihood ratio test[124] for testing (1), see Section of Supplementary Data for detailed derivation. For convenience, we call this test ‘mLRT’, and will use this name throughout the remaining of this paper. The test statistic of mLRT is given by

$$T = \prod_{g\in\{A,B\}} \left(\frac{\hat{p}_0}{\hat{p}_{g,1}}\right)^{\sum_{j=1}^{ng} M_W^{g,j}} \left(\frac{1-\hat{p}_0}{1-\hat{p}_{g,1}}\right)^{\sum_{j=1}^{ng} N_W^{g,j}-M_W^{g,j}}$$

$$\times \left(\frac{\hat{p}_{1,0}}{\hat{p}_{g1,1}}\right)^{\sum_{j=1}^{ng} M_T^{g,j}} \left(\frac{1-\hat{p}_{1,0}}{1-\hat{p}_{g1,1}}\right)^{\sum_{j=1}^{ng} N_T^{g,j}-M_T^{g,j}} ,$$

where the estimates are maximum likelihood estimators (MLEs) as follow,

$$\hat{p}_0 = \begin{cases} \tilde{p}_0, & \text{if } \tilde{p}_0 \geq \tilde{p}_{1,0}, \\ \frac{\sum_g \sum_{j=1}^{ng} M_W^{g,j}+M_T^{g,j}}{\sum_g \sum_{j=1}^{ng} N_W^{g,j}+N_T^{g,j}}, & \text{otherwise;} \end{cases}$$

$$\hat{p}_{1,0} = \begin{cases} \tilde{p}_{1,0}, & \text{if } \tilde{p}_0 \geq \tilde{p}_{1,0}, \\ \hat{p}_0, & \text{otherwise;} \end{cases}$$

$$\hat{p}_{g,1} = \begin{cases} \tilde{p}_{g,1}, & \text{if } p_{g,1} \geq \tilde{p}_{g1,1}, \\ \frac{\sum_{j=1}^{ng} M_W^{g,j}+M_T^{g,j}}{\sum_{j=1}^{ng} N_W^{g,j}+N_T^{g,j}}, & \text{otherwise;} \end{cases}$$

$$\hat{p}_{g1,1} = \begin{cases} \tilde{p}_{g1,1}, & \text{if } \tilde{p}_{g,1} \geq \tilde{p}_{g1,1}, \\ \hat{p}_{g,1}, & \text{otherwise,} \end{cases}$$

$$\text{where } \tilde{p}_0 = \frac{\sum_g \sum_{j=1}^{ng} M_W^{g,j}}{\sum_g \sum_{j=1}^{ng} N_W^{g,j}}, \quad \tilde{p}_{1,0} = \frac{\sum_g \sum_{j=1}^{ng} M_T^{g,j}}{\sum_g \sum_{j=1}^{ng} N_T^{g,j}},$$

$$\tilde{p}_{g,1} = \frac{\sum_{j=1}^{ng} M_W^{g,j}}{\sum_{j=1}^{ng} N_W^{g,j}} \quad \text{and} \quad \tilde{p}_{g1,1} = \frac{\sum_{j=1}^{ng} M_T^{g,j}}{\sum_{j=1}^{ng} N_T^{g,j}}.$$

Under the null H_0 in (1),

$\Lambda = -2\log T \sim x_2^2$,

which is a Chi-square distribution with the degree of freedom of 2. We reject H_0 at level α if $\Lambda > x_{2,\alpha}^2$, where $x_{2,\alpha}^2$ is the $100(1-\alpha)$ percentile point of a Chi-Square distribution with the degree of freedom of 2.

To apply mLRT to the analysis of differential methylation at the region level, we start with applying mLRT to all CpG sites, and thereby obtaining a p-value for each CpG site. It commonly occurs that the number of CpG sites of interest is very large. In such a case, the analysis of DMR detection leads to a high-dimensional multiple testing problem. We adopt the widely used q-value method[125] under the FDR (false discovery rate) level α.

3.4 Results

To evaluate the performance of mLRT in terms of size and power, we first compare it at a CpG site with two-step FET tests involving a step of imputation for missing counts in the contingency table. We consider two types of such FET tests, mFET (Fisher exact test mFET based on maximum likelihood estimators) and nFET (naive Fisher exact test) at the CpG level. We also compare mLRT with the usual Fisher exact test (FET) only using WGBS data to illustrate the necessity of integrating 5hmC and 5mC information in differential methylation analysis. Finally, an illustration at the region level involves the application to a real dataset, i.e., the adult and fetal mouse cortex data.

Note that we do not consider biological variation. Therefore, at a CpG ng j site, data for replicates could be aggregated. Define $m_{gl} = \sum_{j=1}^{n_g} m_{gl}^{j}$, for g ∈ *{A,B}* and *l* = 1,2,3. Data at a CpG site can be expressed in the following 2 × 3 contingency table (Table 2). To carry out mFET and nFET, we need to first fill in unobserved counts m_{gl} in Table 2.

For mFET, the estimates for cells are given by the total count of each category (5hmC, 5mC or unmethylated) multiplied by its corresponding proportion derived from maximum likelihood estimation. Since we have two types of sequencing data, WGBS and TAB-seq, the total count is defined as the average of their observed coverage. Explicitly, the estimated cells in Table 2 are given by

$$m_{gl} = \hat{q}_{gl} \sum_{j=1}^{n_g} \left(N_W^{g,j} + N_T^{g,j} \right) /2, \quad \text{for} \quad g \in \{A, B\}, l = 1, 2, 3,$$

where $\hat{q}_{g3} = 1 - \hat{q}_{g1} - \hat{q}_{g2}$, and

$$(\hat{q}_{g1}, \hat{q}_{g2}) = \begin{cases} \left(\frac{\sum_{j=1}^{n_g} M_T^{g,j}}{\sum_{j=1}^{n_g} N_T^{g,j}}, \frac{\sum_{j=1}^{n_g} M_W^{g,j}}{\sum_{j=1}^{n_g} N_W^{g,j}} - \frac{\sum_{j=1}^{n_g} M_T^{g,j}}{\sum_{j=1}^{n_g} N_T^{g,j}} \right), \\ \qquad \text{if } \frac{\sum_{j=1}^{n_g} M_W^{g,j}}{\sum_{j=1}^{n_g} N_W^{g,j}} - \frac{\sum_{j=1}^{n_g} M_T^{g,j}}{\sum_{j=1}^{n_g} N_T^{g,j}} \geq 0; \\ \left(\frac{\sum_{j=1}^{n_g} M_W^{g,j} + M_T^{g,j}}{\sum_{j=1}^{n_g} N_W^{g,j} + N_T^{g,j}}, 0 \right), \quad \text{otherwise.} \end{cases}$$

The indices *g* and *j* denote the condition and the replicate, respectively.

Table 2. The underlying methylation data

	5hmC	5mC	Unmethylated
Condition *A*	m_{A1}	m_{A2}	m_{A3}
Condition *B*	m_{B1}	m_{B2}	m_{B3}

For nFET, we first estimate the counts of three categories (5hmC, 5mC or unmethylated) for each replicate. For each replicate, the proportion of 5hmC is simply estimated by the ratio of 'C' readouts to the coverage from TAB-seq. This estimated proportion of 5hmC is then applied to WGBS data to fill in unobserved counts in Table 2. The estimators of these counts are given by, for g ∈ *{A, B},*

$$\tilde{m}_{g1} = \sum_{j=1}^{n_g} \tilde{m}_{g1,j}, \quad \tilde{m}_{g2} = \sum_{j=1}^{n_g} \tilde{m}_{g2,j} \text{ and } \tilde{m}_{g3} = \sum_{j=1}^{n_g} \tilde{m}_{g3,j},$$

where

$$\tilde{m}_{g1,j} = \frac{M_T^{g,j}}{N_T^{g,j}} \cdot N_W^{g,j},$$

$$\tilde{m}_{g2,j} = \max\left(M_W^{g,j} - \tilde{m}_{g1,j}\ ,\ 0\right),$$

$$\tilde{m}_{g3,j} = N_W^{g,j} - M_W^{g,j}.$$

For FET using WGBS data only, the data observed at a CpG site can be expressed as the following 2 × 2 contingency table. Then Fisher exact test is carried out directly based on Table 3.

Table 3. The underlying methylation data

	Methylated	Unmethylated
Condition *A*	$\sum_{j=1}^{n_A} M_W^{A,j}$	$\sum_{j=1}^{n_A} N_W^{A,j} - M_W^{A,j}$
Condition *B*	$\sum_{j=1}^{n_B} M_W^{B,j}$	$\sum_{j=1}^{n_B} N_W^{B,j} - M_W^{B,j}$

3.4.1 Power comparison at a single CpG site

To mimic the counts of methylation in a real-world situation, all the CpG sites in the simulated data are uniformly drawn with replacement from the data of Chromosome 1 in the adult and fetal mouse cortex dataset. Figure 1 shows the distributions of 5hmC and 5mC levels from Chromosome 1 in the adult and fetal mouse cortex dataset. Two cases with the number of replicates are considered, nA = nB = 2 and nA = nB = 5. We set $p_{A1} + p_{A2} = p_{B1} + p_{B2} = 0.8$, in order to demonstrate the performance of jointly testing 5hmC and 5mC levels. Joint tests are also compared to FET using WGBS data only, which does not have the ability to identify differential methylation in such a setting of methylation proportions. We fix $p_{A1}$and p_{A2} at 0.1 and 0.7, respectively.

Table 4 shows the power comparison at significance level 0.05 based on 1000 replicates at the CpG level. When (p_{B1}, p_{B2}) = (0.1, 0.7), results are sizes of tests. As we can see from Table 4, the more replicates data have, the higher power all the tests could achieve. Among the three tests combining WGBS and TAB-seq data, mFET has the highest power, however, its sizes are 0.084 and 0.096, which indicates that mFET cannot control Type I error at level 0.05. mLRT can control Type I error since its sizes are around 0.05, and meanwhile outperforms nFET in terms of power. As expected, the power of FET using WGBS data only is around the nominal level 0.05 in any setup of proportions in Table 4. This demonstrates the incapability of FET from WGBS data in the analysis of differential methylation when the sum of 5hmC and 5mC levels remains the same across conditions.

Table 4. Power at significance level 0.5 with 1000 replicates in the setting of $\boldsymbol{p_{A1} + p_{A2} = p_{B1} + p_{B2} = 0.8}$, and $\boldsymbol{(p_{A1}, p_{A2})}$ fixed at (0.1, 0.7).

	$\boldsymbol{n_A = n_B = 2}$				$\boldsymbol{n_A = n_B = 5}$			
$\boldsymbol{(p_{B1}, p_{B2})}$	**mLRT**	**mFET**	**nFET**	**FET**	**mLRT**	**mFET**	**nFET**	**FET**
(0.1,0.7)	0.05	0.084	0.047	0.037	0.052	0.096	0.063	0.045
(0.12,0.68)	0.061	0.084	0.052	0.026	0.073	0.115	0.08	0.05
(0.14,0.66)	0.08	0.125	0.076	0.034	0.124	0.174	0.121	0.053
(0.16,0.64)	0.115	0.165	0.105	0.038	0.189	0.271	0.195	0.04
(0.18,0.62)	0.174	0.242	0.16	0.035	0.324	0.409	0.312	0.048
(0.2,0.6)	0.216	0.303	0.19	0.039	0.452	0.535	0.432	0.05
(0.22,0.58)	0.293	0.384	0.268	0.033	0.561	0.646	0.503	0.043
(0.24,0.56)	0.358	0.446	0.328	0.042	0.651	0.733	0.599	0.045

(0.26,0.54)	0.459	0.559	0.431	0.035	0.777	0.839	0.73	0.039
(0.28,0.52)	0.523	0.625	0.496	0.039	0.871	0.911	0.816	0.041
(0.3,0.5)	0.614	0.687	0.564	0.037	0.914	0.948	0.897	0.038

3.4.2 Real data: DNA methylation dynamics of aging in frontal cortex

To detect DMRs, each chromosome is divided into regions of 500 bp with read counts calculated over each region. We apply the four methods considered in the simulation study to each 500 bp region and compute the p-values. DMRs are then identified using q-values method after adjusting for multiple testing. Figure 2 shows the proportion of identified DMRs by each method for each chromosome. While we may further merge these 500 bp regions, we do not perform this step here for the ease of evaluating method performance.

Figure 2 shows the proportion of regions detected as differentially methylated to all regions of 500 bp in the genome-wide analysis. It indicates that as compared to the mixture of 5mC and 5hmC signals in WGBS data, jointly testing of 5mC level and 5hmC level from WGBS and TAB-seq data are much more capable of capturing differential methylation events. Although mFET overwhelms all the other tests in terms of power, our simulation study in previous section shows mFET cannot control Type I error. Across all chromosomes, mLRT uniformly outperforms nFET.

Figure 3 illustrates the overlap among DMRs detected by mLRT, mFET, nFET and FET in the genome-wide analysis. We observed that mLRT can detect most of the DMRs identified by FET, missing only 131 regions (about 2%) over all chromosomes. Next, to more confidently identify functionally relevant DMRs, we filtered for DMRs with minimum of 3 CpGs and methylation change (WGBS (mC+hmC), mC and hmC) of at least 0.2 (20%) or greater. After filtering, mLRT detected 95% and 99% of nFET DMRs and FET DMRs (Fig. 4A), respectively, highlighting the comphrehensive nature of mLRT in detecting differential methylation. mLRT method also identified almost double the number of DMRs as nFET method. Since these extra DMRs identified by mLRT could be spurious, we performed separate analysis on shared DMRs and mLRT-only DMRs to compare. We report that the mLRT-only DMRs share similar distribution and characteristics as DMRs identified in both nFET and mLRT methods (shared DMRs) (Fig. 4B,C). Furthermore, we report that 97.4% of hmC DMRs and 80.4% of mC DMRs detected by mLRT did not pass 20% difference threshold when only analyzing WGBS data due to concomitant decrease of mC levels in regions with gain of hmC (Fig. 4C). In traditional WGBS studies, these regions would not be detected as DMRs, emphasizing the importance of our tool for detecting novel methylation dynamics, especially in tissues with substantial hmC levels.

Next, we evaluated whether the DMRs identified by mLRT are biologically relevant and can provide novel insights into the DNA methylation changes that occur during aging in the frontal cortex. Since majority of the DMRs are present in intergenic or intronic regions, we utilized Genomic Regions Enrichment of Annotations Tool (GREAT[131]) to discover if DMRs are near genes related to brain maturation. Indeed, both shared and mLRT-only mC hypoDMRs and hmC hyperDMRs are present near genes responsible for biological processes such as cell shape regulation, DNA damage response, synapse maturation and neurodevelopment (Fig. 4D). We also report that mC hyperDMRs occur near blood-related genes, which could reflect the age-associated increase of global mC levels in blood circulating in the frontal cortex tissue[132]. Surprisingly, aging-related mC hypoDMRs and hmC hyperDMRs strongly enrich for exon and 3’ UTRs (Fig. 5A). In fact, 27-31% of exonic DMRs are located in the last exon of the gene further suggesting that 3’ end of the gene might be regulated by DNA mC and hmC in an age-related manner. Furthermore, we performed gene ontology enrichment on genes with DMRs in 3’ UTR using Metascape[133] to check if 3’ UTR regulation is occurring in biologically relevant genes related to aging in brain. All four categories of 3’ UTR DMRs enrich in genes responsible for modulation of chemical synaptic transmission or vesicle mediated transport in synapse biological processes (Fig. 5B). The functional mechanism of methylation or hydroxymethylation in 3’ UTR is still unclear, especially in context of neurodevelopment or diseases associated with aging of the brain. Whether these methylation dynamics directly impact epigenetic and gene regulation or is just a bio-marker for aging would be an exciting and unprecedented future direction in the field of epigenetics and neuroscience. In conclusion, mLRT is a powerful statistical tool that detects biologically meaningful methylation changes and can be source of novel discoveries that reveal the intricate dynamic between hydroxymethylation and methylation levels.

3.5 Discussion

There is a growing appreciation that DNA methylation regulates cell fate decisions and demarcates proper neurodevelopment and aging in the brain[115,123,126–128]. Indeed, aberrant DNA methylation is associated with neurological disorders, highlighting the importance of studying how the canonical DNA methylation dynamic during aging is disrupted in disease models. Monumental studies reported that aging in brain is associated with relatively stable global methylation with differential methylation near neurodevelopment-related genes[127]. The discovery of hydroxymethylation reinvigorated the effort of charting both mC and hmC dynamics in developing brain, which revealed that global hmC levels positively correlate with aging and that regions that gain hmC are often coupled with loss of mC thus these dynamics would not be detected using traditional WGBS[123,126,128–130]. However, due to lack of statistical tools, the differential 5hmC analysis was often performed by classifying large genomic regions as high or low 5hmC levels instead of comparing absolute 5hmC levels at single-base resolution. To address this issue, we present the mLRT method to detect differentially methylated regions based on the integration of WGBS and TAB-seq data. At the CpG level, our method of choice is a likelihood ratio test, jointly testing hmC and mC levels. Compared to FET and nFET, mLRT outperforms in both size and power at predicting methylation and hydroxymethylation for each CpG. Although mFET had higher power than mLRT, mFET suffered from higher Type I error in our simulated tests thus potentially introducing more false positive methylation dynamics. In conclusion, mLRT provides the highest power with adequate Type I error control compared to the other Fisher exact test statistical methods mentioned here.

One question that arose was whether the new methylation dynamism detected by mLRT was meaningful. Here, we evaluated the improved performance and the biological relevance of DMRs detected by mLRT from WGBS and TAB-seq data generated from young and old mouse frontal cortex tissue. mLRT was comprehensive in discovering differentially hydroxy/methylated regions as >95% of DMRs identified by both FET and nFET were also detected by mLRT. Furthermore, mLRT doubled the number of DMRs detected compared to nFET. Many of the DMRs would not have been detected by standard WGBS as the change in 5mC counter-balanced change in 5hmC. Also, these novel DMRs enriched for biologically relevant processes related to brain development, showcasing the power and sensitivity of mLRT tool.

However, there are some caveats to the current mLRT tool. First, these predictions are made by low coverage WGBS and TAB-seq data. Whether mLRT improves with deeper coverage could be interesting future direction to pursue. Considering the cost-prohibitive nature of generating high coverage WGBS and TAB-seq data, we are encouraged by the fact that mLRT is sensitive enough to detect DMRs with low coverage data and hope this tool can provide a financially viable way to study DNA methylation and hydroxymethylation dynamics to the scientific community. Second, mLRT does not account for biological variation. By incorporating biological replicates into mLRT, we can more accurately distinguish biologically meaningful changes from technical noise that might be introduced. In conclusion, the application of our novel method to the real data of front cortex demonstrated that mLRT is a powerful statistical tool that can detect biologically meaningful methylation dynamics.

3.6 Acknowledgements

We thank Joseph Dougherty (Washington University in St. Louis) and Michael Vasek (Washington University in St. Louis) for help with tissue collection.

3.7 Funding

This work has been supported by grants from National Institutes of Health (R01HG007175, U01CA200060, U24ES026699, U01HG009391) and American Cancer Society Research Scholar grant (RSG-14-049-01- DMC). H.J. has been supported by NIGMS T32 training grant (GM00767).

Conflict of Interest: None declared.

3.8 Data availability

The dataset generated in this study will be available in NCBI’s Gene Expression Omnibus (GEO) repository.

3.9 Figures

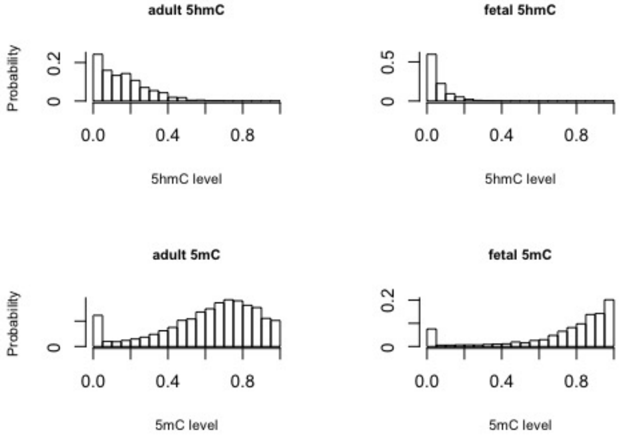


Figure 1. 5hmC and 5mC levels from Chromosome 1 in the adult and fetal mouse cortex dataset.

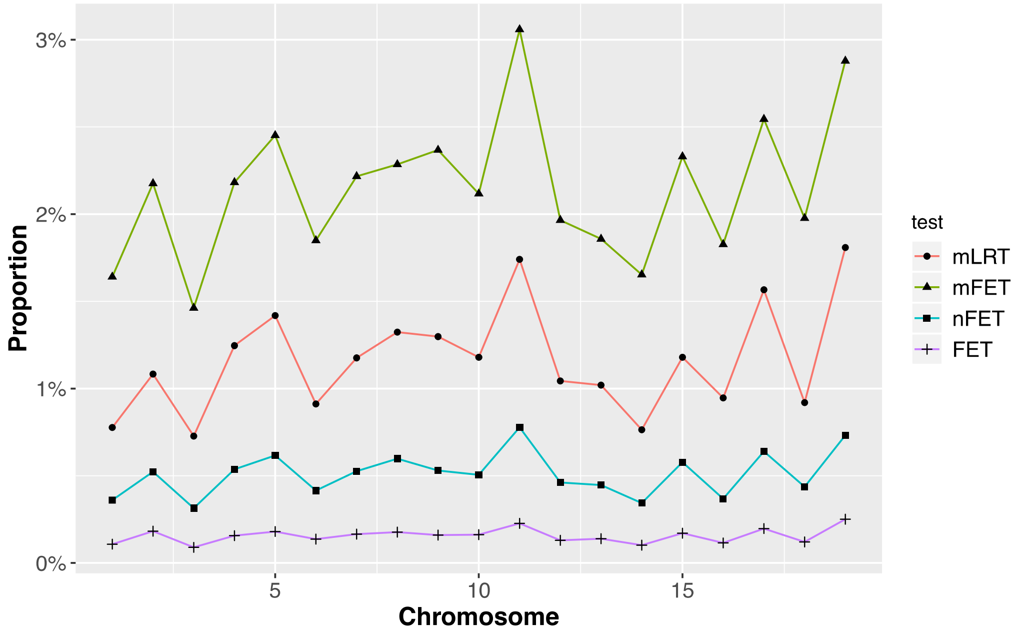


Figure 2. Proportion of DMRs to all regions of 500 bp in the genome-wide analysis.

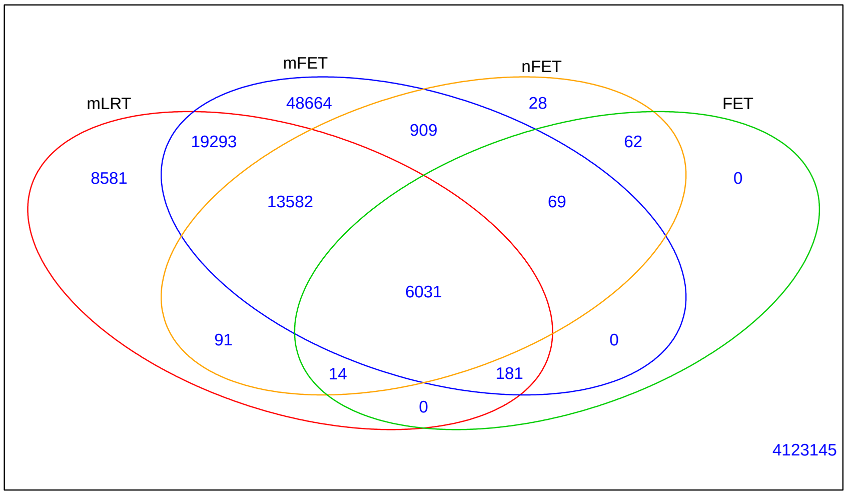
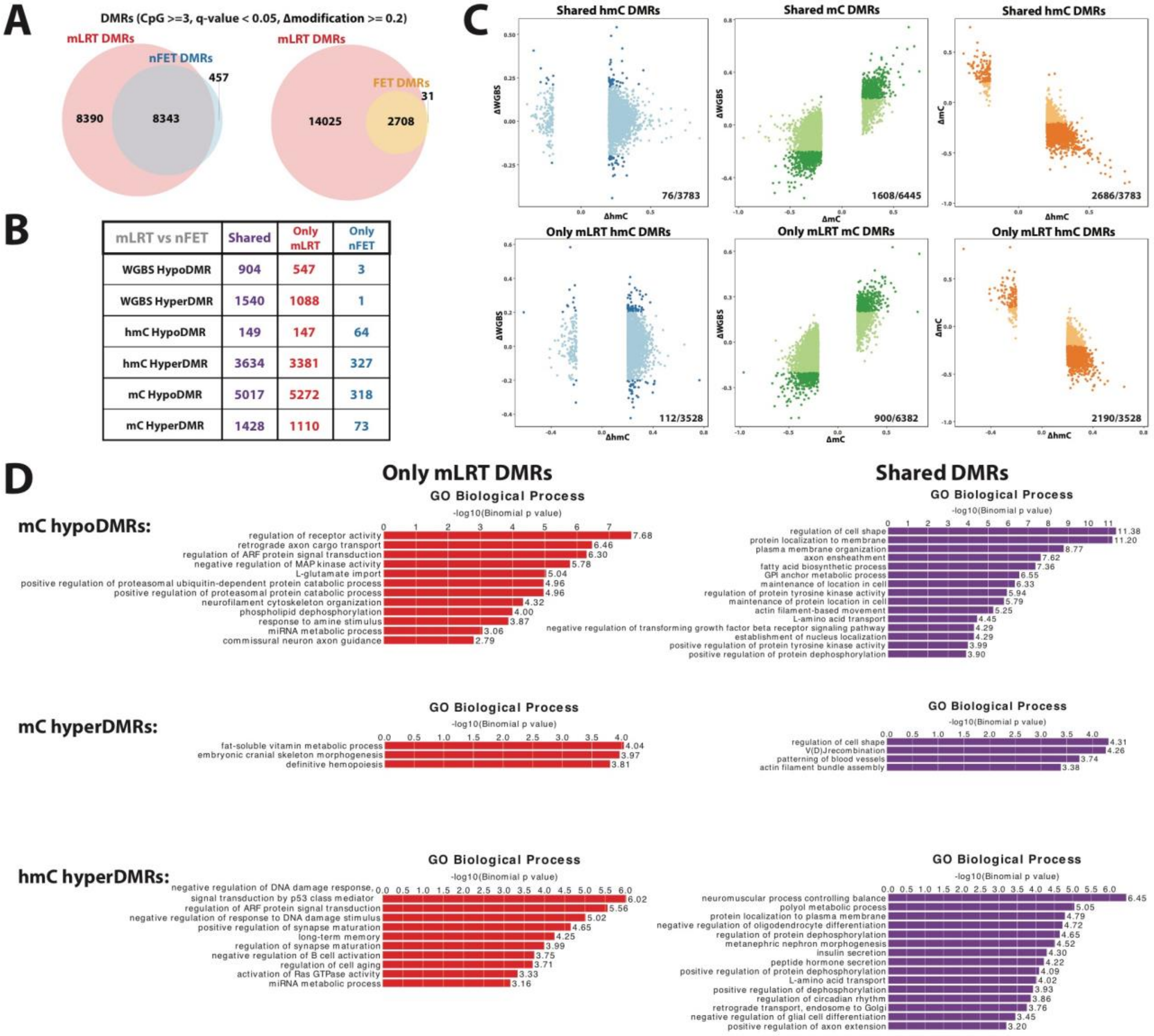


Figure 3. Overlap of DMRs detected by mLRT, mFET, nFET and FET.



mLRT vs nFET	Shared	Only mLRT	Only nFET
WGBS HypoDMR	904	547	3
WGBS HyperDMR	1540	1088	1
hmC HypoDMR	149	147	64
hmC HyperDMR	3634	3381	327
mC HypoDMR	5017	5272	318
mC HyperDMR	1428	1110	73

Figure 4. Biologically relevant DMRs detected by FET, nFET and mLRT. (A) Overlap of filtered DMRs detected by mLRT, nFET and FET. (B) Number of DMR types detected by each method. (C) Distribution of DNA modification of each DMR. Darker colored points represent DMRs that pass 0.2 methylation difference threshold (Number of shared DMRs/total DMRs). (D) Functional annotation of DMRs predicted by GREAT.

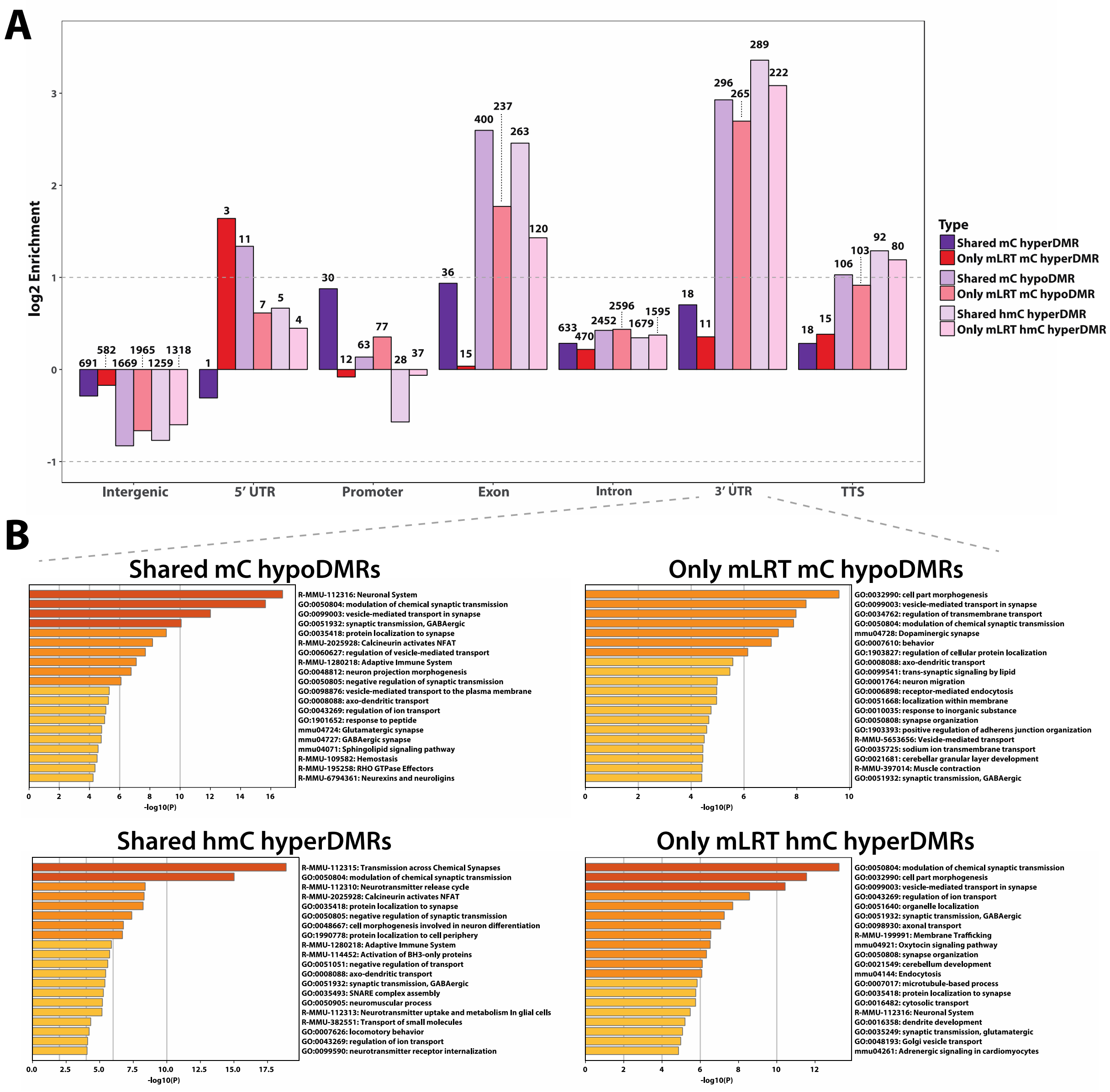


Figure 5. Characterizing DMRs from mLRT. (A) Genomic annotation enrichment of DMR types. (B) Gene ontology term enrichment of genes with 3' UTR DMRs.

3.10 Supplementary data

Derivation of mLRT:

Our goal is to test $H_0 : (p_{A1}, p_{A2}) = \ (p_{B1}, p_{B2})$ vs. $H_0 : (p_{A1}, p_{A2}) \neq \ (p_{B1}, p_{B2})$

which is equivalent to $H_0 : (p_{A1}, p_{A1} + p_{A2}) = (p_{B1}, p_{B1} + p_{B2})$ vs. $H_0 : (p_{A1}, p_{A1} + p_{A2}) \neq (p_{B1}, p_{B1} + p_{B2})$.

Let $g \in \{A,B\}$, and $p_g = p_{g1} + p_{g2}$. Let $j = 1, \cdots, n_g$ denote replicates under Condition g. The overall likelihood with $\Theta_1 = (p_{A1}, p_A, p_{B1}, p_B)$ is

$$f_1 = f_{A,1} \times f_{B,1}$$
$$= \prod_{g \in \{A,B\}} \prod_{j=1}^{n_g} \binom{N_W^{g,j}}{M_W^{g,j}} p_g^{M_W^{g,j}} (1 - p_g)^{N_W^{g,j} - M_W^{g,j}}$$
$$\times \binom{N_T^{g,j}}{M_T^{g,j}} p_{g1}^{M_T^{g,j}} (1 - p_{g1})^{N_T^{g,j} - M_M^{g,j}}.$$

The maximum likelihood estimator for p_{A1}, p_A, p_{B1}, p_B are given by

- (under H_0, pooled)

$$\hat{p}_A = \hat{p}_B = \hat{p}_0 = \begin{cases} \tilde{p}_0, & \text{if } \tilde{p}_0 \geq \tilde{p}_{1,0}; \\ \frac{\sum_g \sum_{j=1}^{n_g} M_W^{g,j} + M_T^{g,j}}{\sum_g \sum_{j=1}^{n_g} N_W^{g,j} + N_T^{g,j}}, & \text{otherwise;} \end{cases}$$

$$\hat{p}_{A1} = \hat{p}_{B1} = \hat{p}_{1,0} = \begin{cases} \tilde{p}_{1,0}, & \text{if } \tilde{p}_0 \geq \tilde{p}_{1,0}; \\ \hat{p}_A, & \text{otherwise,} \end{cases}$$

where $\tilde{p}_0 = \frac{\sum_g \sum_{j=1}^{n_g} M_W^{g,j}}{\sum_g \sum_{j=1}^{n_g} N_W^{g,j}}$ and $\tilde{p}_{1,0} = \frac{\sum_g \sum_{j=1}^{n_g} M_T^{g,j}}{\sum_g \sum_{j=1}^{n_g} N_T^{g,j}}$.

- (under H_1, unpooled)

$$\hat{p}_g := \hat{p}_{g,1} = \begin{cases} \tilde{p}_{g,1}, & \text{if } p_{g,1} \geq \tilde{p}_{g1,1}, \\ \frac{\sum_{j=1}^{n_g} M_W^{g,j} + M_T^{g,j}}{\sum_{j=1}^{n_g} N_W^{g,j} + N_T^{g,j}}, & \text{otherwise;} \end{cases}$$

$$\hat{p}_{g1} := \hat{p}_{g1,1} = \begin{cases} \tilde{p}_{g1,1}, & \text{if } \tilde{p}_{g,1} \geq \tilde{p}_{g1,1}, \\ \hat{p}_A, & \text{otherwise,} \end{cases}$$

where $\tilde{p}_{g,1} = \frac{\sum_{j=1}^{n_g} M_W^{g,j}}{\sum_{j=1}^{n_g} N_W^{g,j}}$ and $\tilde{p}_{g1,1} = \frac{\sum_{j=1}^{n_g} M_T^{g,j}}{\sum_{j=1}^{n_g} N_T^{g,j}}$.

Then, the likelihood ratio is given by

$$
\begin{aligned}
T &= \frac{\sup\limits_{\Theta_0} f_0}{\sup\limits_{\Theta_1} f_1} = \prod_{g\in\{A,B\}} \prod_{j=1}^{n_g} \left(\frac{\hat{p}_0}{\hat{p}_{g,1}}\right)^{M_W^{g,j}} \left(\frac{1-\hat{p}_0}{1-\hat{p}_{g,1}}\right)^{N_W^{g,j}-M_W^{g,j}} \\
&\qquad \times \left(\frac{\hat{p}_{1,0}}{\hat{p}_{g1,1}}\right)^{M_T^{g,j}} \left(\frac{1-\hat{p}_{1,0}}{1-\hat{p}_{g1,1}}\right)^{N_T^{g,j}-M_T^{g,j}} \\
&= \prod_{g\in\{A,B\}} \left(\frac{\hat{p}_0}{\hat{p}_{g,1}}\right)^{\sum_{j=1}^{n_g} M_W^{g,j}} \left(\frac{1-\hat{p}_0}{1-\hat{p}_{g,1}}\right)^{\sum_{j=1}^{n_g} N_W^{g,j}-M_W^{g,j}} \\
&\qquad \times \left(\frac{\hat{p}_{1,0}}{\hat{p}_{g1,1}}\right)^{\sum_{j=1}^{n_g} M_T^{g,j}} \left(\frac{1-\hat{p}_{1,0}}{1-\hat{p}_{g1,1}}\right)^{\sum_{j=1}^{n_g} N_T^{g,j}-M_T^{g,j}},
\end{aligned}
$$

where

$$
\begin{aligned}
\Theta_1 = \{(p_{A1}, p_A, p_{B1}, p_B) : \quad & 0 \le p_{A1} \le p_A \le 1 \\
& \text{and} \;\; 0 \le p_{B1} \le p_B \le 1\}
\end{aligned}
$$

and

$$
\begin{aligned}
\Theta_0 = \{(p_{A1}, p_A, p_{B1}, p_B) : \; & 0 \le p_{g1} \le p_g \le 1, \text{ for } g \in \{A, B\}; \\
& p_A = p_B \;\; \text{and} \;\; p_{A1} = p_{B1}\}.
\end{aligned}
$$

Finally, the likelihood ratio test is given by

$\Lambda = -2\log\mathrm{T}^{\mathrm{H}_0} \sim x_2^2,$

Which is a Chi-square distribution with the degree of freedom of 2.

Chapter 4: Single and multiplex “carpet bomb” gRNA vectors for somatic mutagenesis in zebrafish

Hyo Sik Jang[1,2], Robert C. Tryon[1], Thomas O’Reilly-Pol[1], You Rim Choi[1,2], Alicia N. Wilkening[1,2], Hyung Joo Lee[1,2], Yiran Hou[1,2], Diana Jiang[1,2], Ting Wang[1,2], Stephen L. Johnson[1,3]

[1]Department of Genetics, Washington University Medical School, St Louis, MO 63130, USA
[2]Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St Louis, Missouri 63108, USA
[3]Deceased.

Author Contributions: HSJ, SLJ and TW conceived and designed the study. HSJ, RCT, YRC, ANW, and DJ performed CRISPR experiments such as micro-injections, sample collection, genomic DNA extraction and library construction. HSJ performed computational analysis of indel frequency. RCT, HSJL and YH provided reagents and fish husbandry. TOP designed gRNA sequences provided in this manuscript. HSJ wrote the manuscript with input from all authors.

Work presented here is unpublished.

4.1 Abstract

The monumental advances in CRISPR/Cas9 technology in zebrafish model system have provided valuable insights on cell developmental pathways at a gene-specific resolution through reverse genetic approaches. Here, we introduce a simple and inexpensive alternative to the popular Gateway assembly to generate a multiplexed CRISPR guide RNA (gRNA) vector: a modified Tol2 transposon vector that includes four paralogous U6 promoters expressing unique gRNAs. One potential caveat of CRISPR/Cas9 system is the variable bi-allelic gene inactivation frequency of various gene targets, especially in a conditional mutagenesis system. To maximize conditional bi-allelic mutations, we target a single gene or exon with multiple gRNAs to improve mutagenesis rate, similar to how "carpet bombing" focuses firepower in a designated region. Here, we targeted two well-known pigment genes, *tyr* and *slc45a2*, with single gRNA or carpet bomb vectors to quantify gene inactivation frequency in a conditional mutagenesis system. We performed two complementary assays, targeted sequencing and haploid screening, to illustrate that carpet bombing generated up to a modest 1.5-fold to 7-fold higher frequency of frame-shift mutations and null phenotypes than conventional single gRNA targeting. We report that in carpet bomb-mediated mutagenesis, typical small indels do occur independently from each other, but large deletions spanning the distance of two gRNA cut sites are abundantly present. Collectively, we present a streamlined alternative method to constructing a "carpet bomb" vector that can potentially maximize conditional null phenotypes in zebrafish.

4.2 Introduction

Forward genetic screens in zebrafish have contributed an invaluable role in identifying candidate genes that direct certain developmental pathways [134,135]. However, it wasn't until the recent adaptation of morpholino oligonucleotides (MO), zinc finger nucleases (ZFN), transcription activator-like effector nucleases (TALEN), and CRISPR/Cas9 technologies that a reverse genetic technique became available to validate and assess those candidate genes in the zebrafish model system [136–142]. Out of these gene-editing technologies, CRISPR/Cas9 quickly became the method of choice in the zebrafish community for its ease of use, simplicity in design, and cost scalability for high-throughput screening experiments [143,144]. To generate somatic and germline mutations in zebrafish, gRNA constructs and Cas9 mRNA or protein are directly injected into the yolk of one cell-stage embryos [138,139,145–147]. The CRISPR-mediated gene editing can occur as early as at two-cell stage to generate stable alleles that are propagated through growth and ensuing selective breeding [148]. However, a limitation to this system is that certain essential genes could not be assessed since null phenotype leads to embryonic lethality. Therefore, a need emerged for a conditional CRISPR/Cas9 model with spatial and temporal control.

Currently, limited number of conditional CRISPR techniques are currently available for the zebrafish model system. Two aspects of conditional CRISPR design include: 1) presence of gRNA in target cells and 2) temporal or spatial control of Cas9 expression. In traditional CRISPR knockout design, a synthesized gRNA was directly delivered into embryo where, in the presence of Cas9, can be immediately utilized. However, whether the injected gRNA was stable enough to propagate through zebrafish's lifespan was questionable. Instead, Tol2-based transgenesis vectors were used to insert gRNA-expressing DNA components into the zebrafish genome [147,149]. Temporal or spatial control of when the mutagenesis occurred have been achieved through the use of heat-shock promoters or tissue-specific promoters that express Cas9 transgene [150,151]. For its ubiquitous use in clonal analysis experiments, we generated a stable transgenic line (j940) with Cas9 expressed behind a heatshock promoter (hsp70>Cas9). However, one potential challenge is that gRNAs have wide range of efficiency in generating null phenotypes, especially in a conditional model [139,147]. With the recent discovery of paralogous U6 promoters, a pol III promoter that constitutively transcribes small RNAs, in zebrafish, we have the ability to reliably express four to five gRNAs behind tandem U6 promoters [151]. We decided to target a single gene with four gRNAs to "carpet bomb" the gene with multiple mutagenesis events to maximize CRISPR/Cas9-mediated knockout efficiency. Currently, Gateway assembly is indeed the method of choice for engineering multiplex CRISPR vectors for its accurate integration of large DNA fragments into vectors through the use of site-specific recombination [147,151–155]. Here, we present a quick and cost-efficient alternative method for generating multi-gRNA targeting vectors, through the use of standard PCR amplification and Gibson assembly, for zebrafish genetic analysis. For their clear visible null phenotypes (lack of dark pigments), we focused our effort on two essential pigment genes, *slc45a2* (also known as *albino*) and *tyr*, to quantify CRISPR-mediated gene inactivation frequency [147,156,157]. We utilized next-generation sequencing (NGS) and haploid screens to show that "carpet bomb" transposon can improve CRISPR-mediated gene inactivation frequency up to 7-fold higher than conventional single gRNA system in our conditional model.

4.3 METHODS

4.3.1 Zebrafish Care

All zebrafish were used in accordance with the protocols approved by the Washington University Animal Studies Committee (Protocol 20140195) and maintained under standard conditions as dictated in The Zebrafish Book [158].

4.3.2 Designing gRNA sequences

We developed an algorithm that ranks candidate gRNAs generated by the E-CRISP tool in each gene of the Zv9 reference genome [159]. Our algorithm first requires that the candidate gRNA targeting sequence falls within an exon, and that the 20 base targeting sequence begins with a G nucleotide, consistent with the G at the transcription initiation site for zebrafish U6 promoters [160]. Candidate gRNAs for each gene were then placed in 1 of three bins, according to the number of potential off-target binding sites in the zebrafish genome (0, 1 or 2+). They were then ranked within each bin using the following formula: 60*[fraction GC content] + 10*[fraction of gene transcripts containing gRNA site] - 30*[relative position of target sequence from the transcription start site within the gene along the chromosome] + 2 (if position 20 is G) - 3(if position 20 is A). We then selected the best 10 candidate gRNAs, starting with high-scoring gRNAs from the "0" off-target bin, following with candidates from the "1" off-target bin, and lastly using gRNAs from the "2 or more" off-target bin. We note that this strategy for gRNA rankings is based on the investigators' intuition of how to balance gene-inactivating gRNAs against gRNAs that might result in off-target lesions in other genes. An excel file with candidate gRNAs for each zv9 gene is available at http://genetics.wustl.edu/sjlab/public-data/u6-grna-database/.

4.3.2 Construction of CRISPR transgene plasmid

Here, we present a streamlined method of constructing Tol2-based CRISPR vectors through the use of PCR extension and Gibson assembly [161]. We provide a user-friendly excel template for primer design and also sequences for all plasmids utilized in this experiment at http://genetics.wustl.edu/sjlab/lab-protocols/carpet-bombs. The carpet bomb construction involves four plasmids: U6-21 (U6 promoter located in chromosome 21) precursor transposon, U6-9 Template A plasmid, U6-11 Template B plasmid, and U6-6 Template C plasmid. The U6-21 gRNA precursor transposon contains a U6 Chr21 promoter followed by gRNA panhandle and a *Xenopus EF1α* promoter driving GFP expression in between Tol2 sequences. We linearized U6-21 gRNA precursor transposon by co-digestion with NruI restriction enzyme (NEB R3192S) and AClI restriction enzyme (NEB R0598S). For single gRNA vector construction, primers SF and SR were annealed and extended to generate a product with flanking sequences that overlap the ends of linearized U6-21 gRNA precursor transposon (**Fig. 1A**). The PCR product was cloned into precursor transposon by following standard Gibson Assembly Mastermix protocol (NEB, E2611S). The construction of carpet bomb vector includes the assembly of three separate PCR products. (**Fig. 1B)**. To minimize potential non-specific amplification during PCR, we provide template A, template B, and template C that has U6 Chr9 promoter, U6 Chr11 promoter, and U6 Chr6 promoter preceded by gRNA panhandle respectively. The PCR extended products will contain different overlapping sequences, derived from unique gRNA sequences, that can be properly oriented and inserted into the precursor transposon via Gibson assembly. The carpet bomb vector was constructed by following standard Gibson Assembly protocol. In brief, linearized U6-21 gRNA precursor transposon was incubated with PCR product A, PCR product B, and PCR product C at a molar ratio of 1:3:3:3 for 30 minutes. Gibson-assembled vectors were transformed into competent Top10 cells and extracted using High-Speed Plasmid Mini Kit (IBI Scientific, IB47102). Since carpet bomb vectors contain novel ApaI and NdeI restriction sites that are not present in “U6-21 gRNA precursor transposon”, we performed restriction digest screens to identify properly assembled candidates, which were further validated with Sanger sequencing. The primers used to generate *slc45a2* and *tyr* CRISPR vectors are listed in **Figure 2.**

4.3.3 Microinjection of CRISPR vectors

All injections were performed in 1 to 2-cell stage zebrafish j940 embryos (**Fig. 3A**). We injected approximately 1nl of 100ng/ul CRISPR transgene vector and 15ng/ul Tol2 capped transposase mRNA cocktail into embryos.

4.3.4 Heat-shock for induction of Cas9

To induce expression of Cas9, we first screened for normally developed transgenic embryos with GFP expression at 1dpf (**Fig. 3A**). Approximately 100 embryos placed in 15ml of egg water in 50ml conical tubes. We then added 15ml of pre-warmed (37°C) egg water. The tubes were immediately placed in 37°C water bath for 30 minutes following which embryos and 37°C egg water were transferred into petri dishes and allowed to cool to 28°C in incubator. Dead embryos were removed the next day.

4.3.5 Single-cell dissociation and FACS

We adapted single-cell dissociation protocol to digest 2dpf heat-shocked embryos into single cells for FACS [162]. First, the embryos were dechorionated in Pronase (Sigma, 10165921001). Approximately 150 dechorionated embryos were collected in 1.5ml eppendorf tube. We then replaced egg water with 1mL of TrypLE Express (ThermoFisher Scientific, 12605021) and incubated the embryos at room temperature for 10 minutes on a rotator. After incubation, embryos were mechanically dissociated by triturating with a 1ml pipette tip. Once in single cell suspension, the sample was centrifuged at 300g for 8 minutes at 4°C to pellet the cells. We re-suspended the pellet in 800ul of cold PBS+2%FBS. Resuspended cells were then filtered through a 100uM cell filter (Partec, 04-0042-2318) to remove clumped cells. Dissociated single cells were analyzed using flow-cytometry (Beckman Coulter MoFlo) and cells positive for GFP expression were collected for DNA extraction.

4.3.6 CRISPR sequencing library generation

We used targeted PCR and next-generation sequencing to quantitatively calculate mutagenesis rates at targeted sites. We first lysed GFP-positive cells with DNA extraction buffer (50mM Tris, 1mM EDTA, 0.5% SDS, 1mg/ml Proteinase K) by incubating overnight at 55°C. DNA was purified by phenol-chloroform extraction followed by ethanol precipitation. We amplified targeted regions of *tyr* exon 1 and *slc45a2* exon 1 using Phusion High-Fidelity Polymerase (NEB M0530) following manufacture's suggestions (**Fig. 2 & 3B**). Sequencing libraries were generated using standard Illumina library preparation as adapted previously [163]. Although *tyr* PCR product was ~800bp in size, all four gRNA targets are within 250bp of the PCR ends, which we were able to capture using the 250bp paired-end Illumina MiSeq platform.

4.3.7 Identifying CRISPR induced mutations

Sequencing reads were aligned to GRCz10 reference genome using BLAT [164]. Concordant paired reads that mapped to *tyr* and *slc45a2* were greater than 90% of total reads. Since all targeted sites were covered by our two paired reads, we can identify multiple combinations of indels. Additionally, we analyzed BLAT outputs in R to call insertions or deletions that occur within 30bp of predicted break site. For more accurate estimate of gRNA2/gRNA3 indel frequency (**Fig. S1**), we excluded large deletions (between gRNA1 and gRNA4) from our calculations since central gRNAs could have generated indels but are not captured. We note that this adjustment could slightly overestimate indel frequency to be higher at overlapping gRNA locations than what occurs.

4.3.8 Haploid screen analysis for functional inactivation

A subset of heat-shocked embryos was reared to maturity at 28°C. Once viable for breeding, eggs from founder female fish were in-vitro fertilized with UV-inactivated sperm following the published protocol [165]. Fertilized embryos were sorted for GFP expression at 1dpf stage and scored for loss of pigmentation at 3dpf.

4.4 RESULTS

4.4.1 Simple, cost-efficient approach for multiplex (carpet bomb) CRISPR vector construction

Here, we developed a method for constructing single and multiplex gRNA expressing vectors. We refer the multiplex vector, which expresses four gRNAs from four different U6 promoter, as carpet bomb. Our method here is a potential alternative to the previously published Golden Gate assembly for generating multiplex gRNA vectors. Our motivation for an alternative method stems from our experience that Golden Gate assembly could be a challenging and time-consuming task for the mass production of carpet bomb vectors targeting many genes. We highlight the simplicity of our approach, which utilizes commonplace techniques and requires minimal reagents and less time. Our approach is cost-efficient as only a pair or three pairs of primers are necessary to generate a single gRNA or carpet bomb vector, respectively. The inherent disadvantage in Gibson assembly is the need for 20-40bp overlap at ends of DNA fragments, which prevents systematic assembly of fragments with same homologous ends. However, in our method, we take advantage of the unique gRNA sequences and use them as anchors flanking U6 promoter and gRNA panhandle for subsequent Gibson assembly (**Fig. 1**). We designed primers that have terminal sequence of U6 promoters or initial sequence of gRNA panhandle, which we extended to include unique gRNA sequence (**Fig. 2**). This produced PCR fragments with unique flanking sequences that can be assembled via Gibson assembly. Furthermore, we can construct a carpet bomb vector with a single cloning step thus avoiding the time-consuming process of multiple cloning events necessitated in Golden Gate assembly.

4.4.2 Carpet bomb CRISPR mutagenesis from targeted sequencing

We sequenced CRISPR-targeted regions of *tyr* and *slc45a2* genes by generating sequence libraries for the 250bp paired-end MiSeq platform for each CRISPR experiments. The paired reads encompass all four gRNA target sites allowing us to measure co-occurrence of gRNA mutagenesis at individual molecule level (**Fig. 3B**). First, we asked whether carpet bomb technique outperform the conventional single gRNA vector. For each CRISPR vector, we quantified the frequency of reads that had an indel (including 1bp insertions and deletions) within 30bp of each gRNA target site (**Fig. 4**). Furthermore, we quantified how often frameshift occurred due to the indels generated by CRISPR activity to compare the rate of functional inactivation of the gene. For both gene examined, we observed an increase in the mutagenesis and frameshift events for carpet bomb vectors respective to the single gRNA counterparts (**Fig. 3C**). In *tyr* context, the carpet bomb vector produced 62.6% of reads (43.7% led to frameshift) while the single gRNA vector generated 10.7% of reads (6.2% led to frameshift) that had at least one indel in one of the target regions. In *slc45a2* context, we observe a modest increase of 33.6% to 39.9% of reads showing at least one indel (21.7% to 32.7% for frameshift indels) when comparing single gRNA vector to carpet bomb vector.

Second, with the rise of multiplexing and tiling CRISPR assays, we were curious whether overlapping gRNA have synergistic influence in CRISPR mutagenesis. In the carpet bomb design, second gRNA and third gRNA overlap where predicted cleavage sites are 1bp apart in *slc45a2* vector and 7bp apart in *tyr* vector. We quantified how often an indel was generated in these overlapping gRNA positions compared to first and fourth gRNA target regions in the carpet bomb condition. We observed no appreciable boost in CRISPR break frequency in the overlapping region, which suggests that there is no additive synergy in producing CRISPR-mediated breaks when gRNAs are tiled or overlapping (**Fig. S3**). In fact, interestingly, we note a slight decrease in indel frequency in the overlapping region, however, further experiments are necessary to substantiate the claim that overlapping gRNA are antagonistic.

Lastly, we asked what possible CRISPR mutagenesis patterns could be expected by the carpet bomb technique. We broadly categorized mutagenesis events of carpet bomb vectors based on: 1) “no indel” 2) “large deletion” 3) “discrete indel” 4) “complex rearrangement” (**Fig. 4**). Within “no indel” events, we discovered that extremely few reads had mismatch mutations in gRNA or PAM sequence, which might render that region untargetable through CRISPR-Cas9 mechanism. This could indicate that DNA repair is extremely precise during embryogenesis or that no CRISPR activity occurred in the cell. For carpet bomb vectors, deletions of large regions between gRNAs were most common; 30.5% and 50.1% of total reads in *slc45a2* carpet bomb and *tyr* carpet bomb respectively. These large deletions represent more than 70% of total CRISPR-generated mutations in both gene contexts. Collections of local small indels were also present at modest frequencies of 7.4% in *slc45a2* carpet bomb and 16.1% in *tyr* carpet bomb. Furthermore, we also captured a small fraction of CRISPR-induced breaks (~2%) that illustrated complex rearrangements and inversions, which have also been identified in other model systems [142,166–170].

4.4.3 CRISPR-mediated gene inactivation in haploid analysis

To more accurately quantify how well the carpet bomb CRISPR technique can generate a null phenotype, we scored haploids of CRISPR-modified embryos from founder females for loss of pigmentation. When characterizing haploids, we separated phenotype into two classes: complete loss of pigmentation and normal pigmentation (**Fig. 3A**). Correlating with targeted sequencing results, we observe an increase of pigmentation defect frequency in haploids from carpet bomb-induced founder than single gRNA-induced founders (65% vs 0% in *tyr* and 57% vs 25% in *slc45a2*) (**Fig. 3C**). We noticed that exon 1 in *tyr* gene encodes important signaling peptide and EGF (epidermal growth factor)-like domains that are crucial for the protein’s function [171].

Similarly, nonsense mutations in exon 1 of *slc45a2 gene* have been attributed to albinism in humans [172]. Large deletions of these functional domains can lead to null-phenotypes, which can explain why frequencies of unpigmented haploids were much higher than expected by frameshift mutations.

4.5 Discussion

In this study, we describe a simple and inexpensive alternative to Gateway assembly for constructing multiplex CRISPR vectors. Our lab has developed transposon-based clonal analysis in zebrafish to study fate restriction of cell lineages during development and fin regeneration [173–176]. Clonal analysis, using Tol2 constructs, generates a mosaic embryo where only one or two cell lineages are labeled with GFP in the caudal fin for lineage-tracing analysis [149,176]. We modified clonal analysis to incorporate CRISPR technology by designing the Tol2 vector to include U6-gRNA constructs. The individually labeled cell lineages provide spatial control on which cell populations experience CRISPR-Cas9 activity. For temporal control, the CRISPR vector was injected into a stable line with a heat-shock promoter driving Cas9 expression. By combining clonal analysis and CRISPR technology, we present a conditional knockout zebrafish model to analyze temporal requirements for genes' function and role in transposon generated somatic clones.

We sought to compare whether targeting a gene with more gRNAs would improve CRISPR efficiency in generating indels and in consequence, improve the chance of functionally inactivating the target gene. One limitation of single gRNA targeting is that we expect 1/3 of CRISPR-induced breaks will repair with no frameshift, and thus, unlikely to cease target gene's function. By targeting a gene with four gRNAs, four possible loci can independently experience mutagenesis events thus improving the overall frequency of mutations and the odds that a frameshift will occur in the gene. In both targeting sequence results and haploid results, we do observe an appreciable improvement in CRISPR-mediated indel frequency in the carpet bomb model. Furthermore, the haploid analysis supports our assumption that carpet bomb technique also increases the frequency of generating null alleles of the targeted gene. It was promising that a single heat shock induction produced such high rates of gene inactivation.

The results from these experiments provide a guideline in how to design a more effective carpet bomb vector. First, overlapping gRNAs do not provide a noticeable improvement in CRISPR-induced indel frequency in the target region. Therefore, gRNAs in carpet bomb vectors should be selected so that the gRNA target regions that are spaced apart. Further studies should be done to elucidate the optimal spacing among gRNA target sites before the gRNAs have redundant function. Second, carpet bomb CRISPR events frequently generate large deletions of regions between flanking gRNAs. Our NGS results at two albino loci suggest that close to 75% of mutations generated by carpet bombs are in fact deletions between targeting sites. These large deletions have the increased probability that the mutated chromosome or loci is inactivated for targeted gene function. Furthermore, if the function of the gene and the structure of the protein are known, one can choose gRNAs that flank important functional elements or structures and perform targeted deletion, likely rendering the gene product as nonfunctional. With these guidelines, we continue to optimize the carpet bomb conditional CRISPR technique to maximize CRISPR-mediated knockouts for transient analysis of gene function.

The interest in optimizing the carpet bomb conditional CRISPR technique spurs from the potential application of combining CRISPR-Cas9 screening and clonal analysis to perform reverse genetic screens of important developmental genes. Clonal analysis has been utilized in the zebrafish experiments for lineage-tracing and developmental pathway explorations.

However, “carpet bomb” CRISPR can be applied to any general conditional KO experiments to maximize genetic ablation potential. Coupling carpet bomb conditional CRISPR technology with clonal analysis provides numerous advantages when screening for essential developmental genes. First, many epigenetic related genes are critical for embryogenesis and are lethal when knocked out early in development. The conditional heat shock model allows us to control temporally when the CRISPR-mediated knockout is activated. Second, clonal analysis offers a cell-lineage specific resolution. Lastly, by maximizing CRISPR efficiency through the use of carpet bomb technique, we can perform mutagenesis screens without the need to generate homozygous mutant progenies for phenotypic analysis, which demands months of effort. However, it’s important to note that the current single heat shock model only increases CRISPR efficiency to <70%. We are pursuing conditions to reach close to 99% CRISPR efficiency through multi-heat shock conditions.

We recognize that the increase in gRNA number is accompanied by higher number of potential off-target effects. In the scope of this paper, we have not quantified the frequency of off-target mutations. We are encouraged by recent literatures revealing minimal off-target frequencies in zebrafish CRISPR experiments [138,145,147]. We value the importance of off-target effect’s potential contribution to CRISPR-mediated knockout phenotype. However, we will utilize this technology as a screening tool for important functional genes in fin regeneration, which we will further validate using alternative methods such as using other gRNAs or performing a rescue experiment. These complementary experiments can validate the target gene, not off-target effects, is causal for phenotype. Furthermore, rapid advances in CRISPR/Cas9 technology have optimized the system to minimize off-target efficient to undetectable levels [177–179]. We look forward to incorporating these exciting advances in CRISPR technology to continually improve conditional CRISPR-based assays in zebrafish.

4.6 Acknowledgements

This work has been supported by grants from National Institutes of Health (R01HG007175, U01CA200060, U24ES026699, U01HG009391) and American Cancer Society Research Scholar grant (RSG-14-049-01- DMC). H.J. has been supported by NIGMS T32 training grant (GM00767).

4.7 Figures

Figure 1

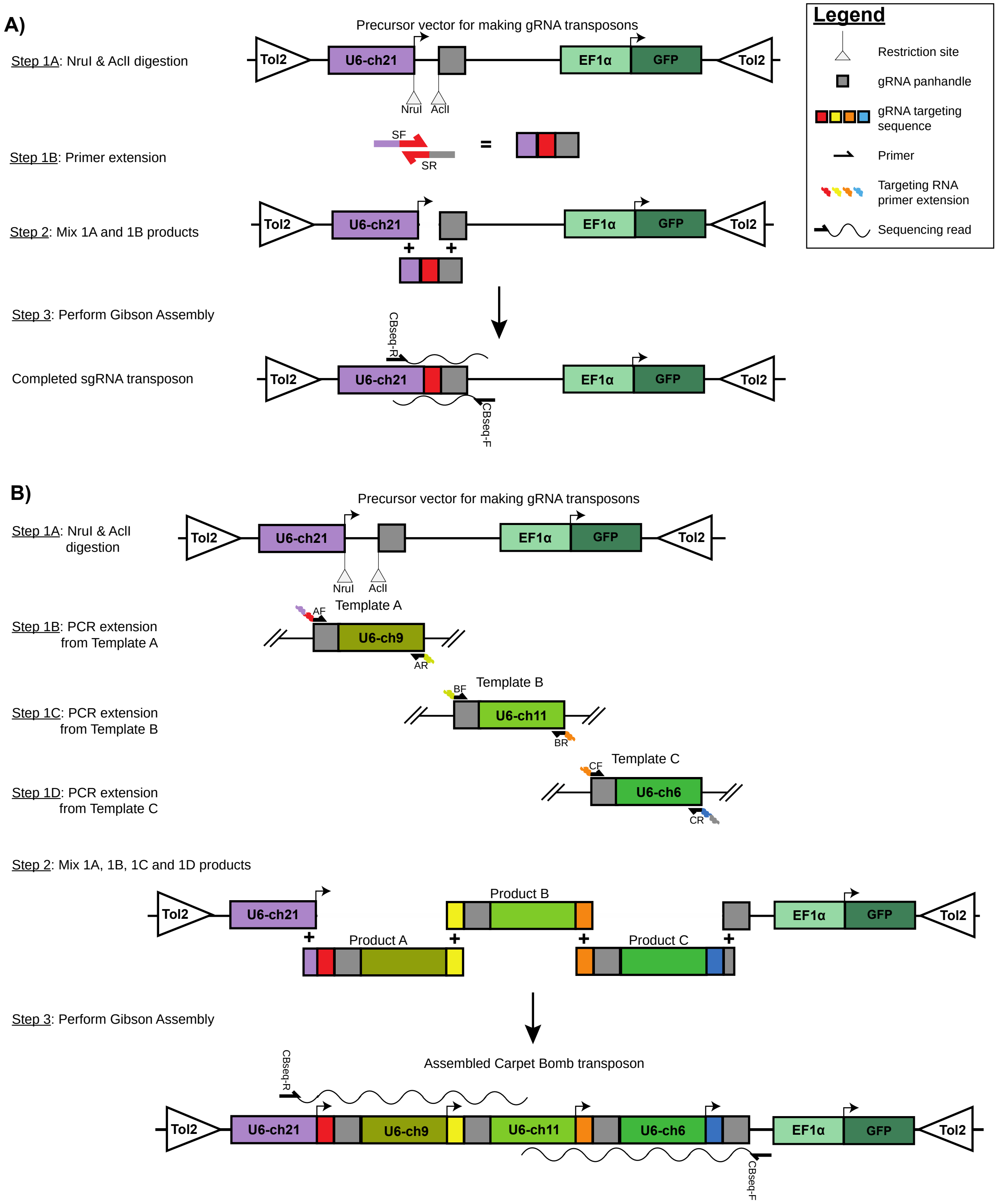
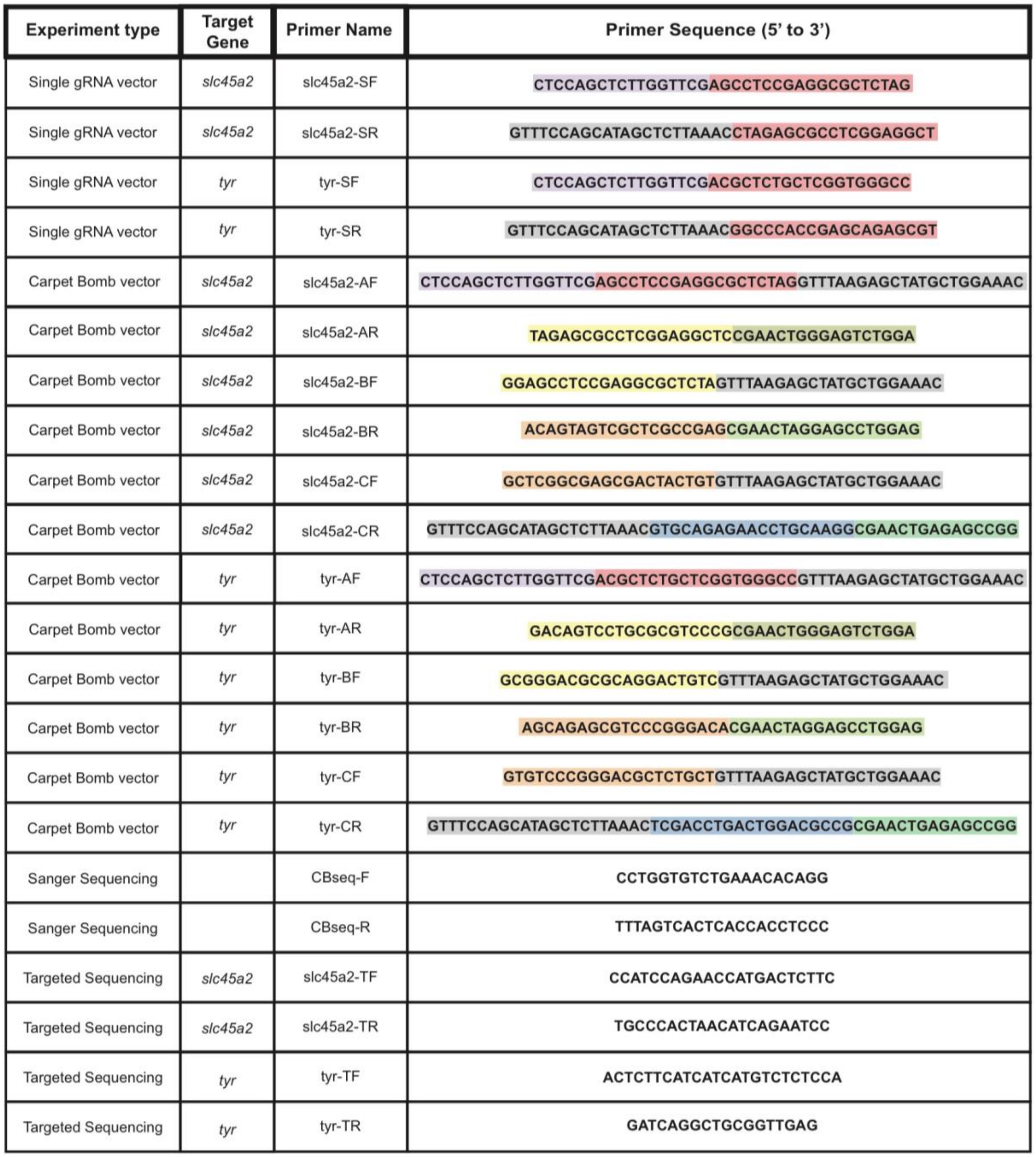

A)
Precursor vector for making gRNA transposons
Step 1A: NruI & AclI digestion
Tol2
U6-ch21
EF1α
GFP
Tol2
NruI
AclI
Step 1B: Primer extension
SF
SR
=
Step 2: Mix 1A and 1B products
+
+
Step 3: Perform Gibson Assembly
CBseq-R
Completed sgRNA transposon
CBseq-F
Legend
Restriction site
gRNA panhandle
gRNA targeting sequence
Primer
Targeting RNA primer extension
Sequencing read
B)
Precursor vector for making gRNA transposons
Step 1A: NruI & AclI digestion
NruI
AclI
Step 1B: PCR extension from Template A
AF
Template A
U6-ch9
AR
Step 1C: PCR extension from Template B
BF
Template B
U6-ch11
BR
Step 1D: PCR extension from Template C
CF
Template C
U6-ch6
CR
Step 2: Mix 1A, 1B, 1C and 1D products
Product A
Product B
Product C
Step 3: Perform Gibson Assembly
Assembled Carpet Bomb transposon
CBseq-R
CBseq-F

Figure 1: Schematic of CRISPR carpet-bomb vector construction. (A) The assembly of single gRNA vector includes two components: linearized precursor vector and insert template, which encodes the gRNA. The insert template is the product of annealing and extending primers SF and SR. The various colors represent unique sequences that are overlapping at the terminal ends of DNA fragments. These unique overlapping regions allow for proper assembly during Gibson Assembly. (B) Carpet bomb vector construction involves four components with unique flanking sequences that can be properly assembled. The primers are color coded where black indicates complementary sequence that the primer binds to during PCR while squiggly color lines indicate unique sequences that are extended at the ends of the product via primer extension.

Figure 2



Experiment type	Target Gene	Primer Name	Primer Sequence (5' to 3')
Single gRNA vector	*slc45a2*	slc45a2-SF	CTCCAGCTCTTGGTTCGAGCCTCCGAGGCGCTCTAG
Single gRNA vector	*slc45a2*	slc45a2-SR	GTTTCCAGCATAGCTCTTAAACCTAGAGCGCCTCGGAGGCT
Single gRNA vector	*tyr*	tyr-SF	CTCCAGCTCTTGGTTCGACGCTCTGCTCGGTGGGCC
Single gRNA vector	*tyr*	tyr-SR	GTTTCCAGCATAGCTCTTAAACGGCCCACCGAGCAGAGCGT
Carpet Bomb vector	*slc45a2*	slc45a2-AF	CTCCAGCTCTTGGTTCGAGCCTCCGAGGCGCTCTAGGTTTAAGAGCTATGCTGGAAAC
Carpet Bomb vector	*slc45a2*	slc45a2-AR	TAGAGCGCCTCGGAGGCTCCGAACTGGGAGTCTGGA
Carpet Bomb vector	*slc45a2*	slc45a2-BF	GGAGCCTCCGAGGCGCTCTAGTTTAAGAGCTATGCTGGAAAC
Carpet Bomb vector	*slc45a2*	slc45a2-BR	ACAGTAGTCGCTCGCCGAGCGAACTAGGAGCCTGGAG
Carpet Bomb vector	*slc45a2*	slc45a2-CF	GCTCGGCGAGCGACTACTGTGTTTAAGAGCTATGCTGGAAAC
Carpet Bomb vector	*slc45a2*	slc45a2-CR	GTTTCCAGCATAGCTCTTAAACGTGCAGAGAACCTGCAAGGCGAACTGAGAGCCGG
Carpet Bomb vector	*tyr*	tyr-AF	CTCCAGCTCTTGGTTCGACGCTCTGCTCGGTGGGCCGTTTAAGAGCTATGCTGGAAAC
Carpet Bomb vector	*tyr*	tyr-AR	GACAGTCCTGCGCGTCCCGCGAACTGGGAGTCTGGA
Carpet Bomb vector	*tyr*	tyr-BF	GCGGGACGCGCAGGACTGTCGTTTAAGAGCTATGCTGGAAAC
Carpet Bomb vector	*tyr*	tyr-BR	AGCAGAGCGTCCCGGGACACGAACTAGGAGCCTGGAG
Carpet Bomb vector	*tyr*	tyr-CF	GTGTCCCGGGACGCTCTGCTGTTTAAGAGCTATGCTGGAAAC
Carpet Bomb vector	*tyr*	tyr-CR	GTTTCCAGCATAGCTCTTAAACTCGACCTGACTGGACGCCGCGAACTGAGAGCCGG
Sanger Sequencing		CBseq-F	CCTGGTGTCTGAAACACAGG
Sanger Sequencing		CBseq-R	TTTAGTCACTCACCACCTCCC
Targeted Sequencing	*slc45a2*	slc45a2-TF	CCATCCAGAACCATGACTCTTC
Targeted Sequencing	*slc45a2*	slc45a2-TR	TGCCCACTAACATCAGAATCC
Targeted Sequencing	*tyr*	tyr-TF	ACTCTTCATCATCATGTCTCTCCA
Targeted Sequencing	*tyr*	tyr-TR	GATCAGGCTGCGGTTGAG

Figure 2: Primer sequences used in CRISPR vector construction. Here, we provide the sequences of all the primers that are used to generate the single gRNA vector and carpet bomb vector. We highlighted the parts of the primer sequences based on the sequence context shown in

Figure 1 to illustrate overlapping attributes. Furthermore, primers used for Sanger validation and targeted sequencing are also specified.

Figure 3

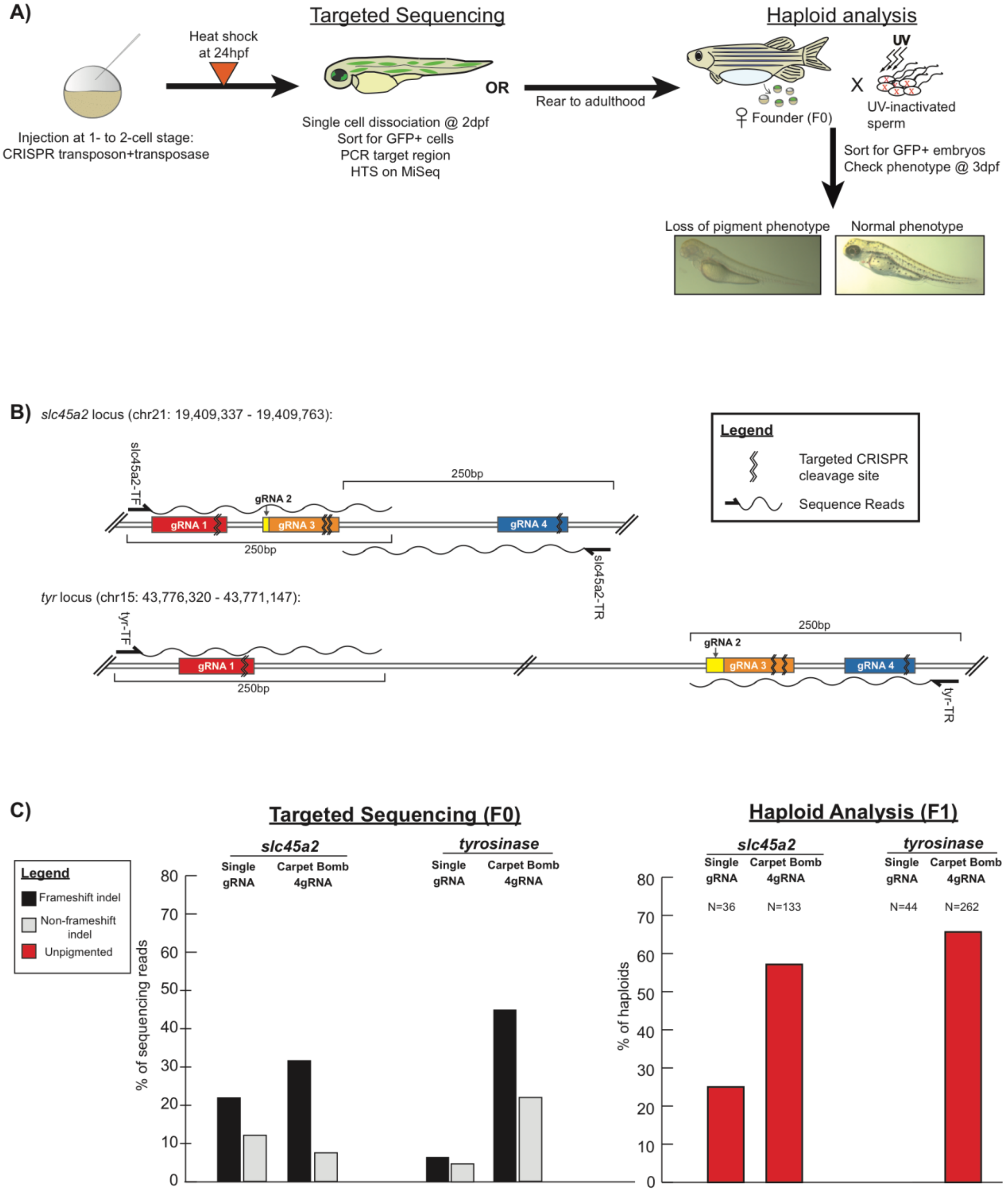
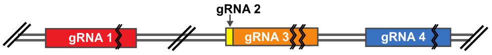
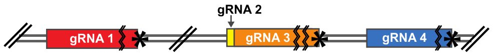
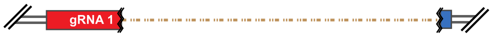
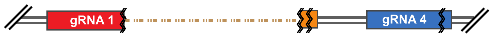
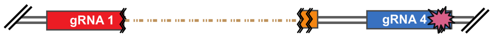
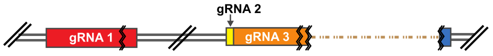
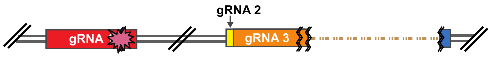
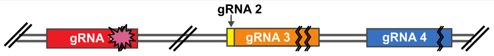
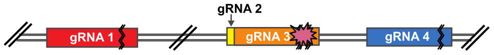
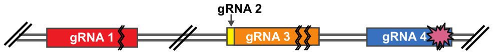
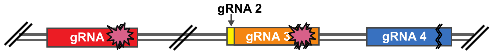
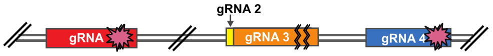
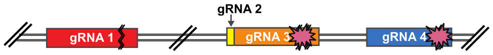
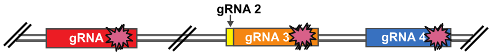
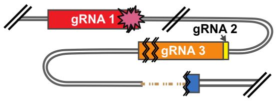

A)
Targeted Sequencing
Haploid analysis
Heat shock
at 24hpf
Injection at 1- to 2-cell stage:
CRISPR transposon+transposase
Single cell dissociation @ 2dpf
Sort for GFP+ cells
PCR target region
HTS on MiSeq
OR
Rear to adulthood
UV
X
♀ Founder (F0)
UV-inactivated
sperm
Sort for GFP+ embryos
Check phenotype @ 3dpf
Loss of pigment phenotype
Normal phenotype
B) slc45a2 locus (chr21: 19,409,337 - 19,409,763):
slc45a2-TF
gRNA 1
gRNA 2
gRNA 3
gRNA 4
250bp
250bp
slc45a2-TR
Legend
Targeted CRISPR
cleavage site
Sequence Reads
tyr locus (chr15: 43,776,320 - 43,771,147):
tyr-TF
gRNA 1
250bp
250bp
gRNA 2
gRNA 3
gRNA 4
tyr-TR
C)
Targeted Sequencing (F0)
Haploid Analysis (F1)
Legend
Frameshift indel
Non-frameshift
indel
Unpigmented
slc45a2
tyrosinase
Single
gRNA
Carpet Bomb
4gRNA
% of sequencing reads
0
10
20
30
40
50
60
70
80
slc45a2
tyrosinase
Single
gRNA
Carpet Bomb
4gRNA
N=36
N=133
N=44
N=262
% of haploids
0
10
20
30
40
50
60
70
80

Figure 3: Improved CRISPR-induced gene inactivation frequency in carpet bomb context. (A) A cocktail of CRISPR transposon vector and transposase is injected into 1- to 2-cell embryos and heat shocked at 24hpf to activate Cas9 expression. Mosaic GFP larvae were either dissociated to collect GFP-positive cells for targeted sequence analysis or grown up to adulthood for subsequent haploid analysis. The adult fish were screened to identify founder females, which produced GFP-positive embryos indicating germline transmission of the CRISPR transposon. The founders' eggs were fertilized with UV-inactivated sperm to generate haploid embryos. After sorting for GFP-positive haploid embryos, we counted the number of unpigmented haploids at 3dpf to quantify the frequency of gene inactivation caused by CRISPR activity. (B) Schematic of the locations of gRNA positions on target genes (not scaled). As shown, gRNA2 and gRNA3 are overlapping in both genes. The curvy lines indicate the targeted sequencing regions that are captured with 250bp PE MiSeq platform. (C) The results of both targeted sequencing analysis and haploid analysis show appreciable increase in gene inactivation frequencies indicated by the increase in frameshift indel frequencies and unpigmented haploid frequencies in both genes' carpet bomb conditions relative to single gRNA conditions.

Figure 4

Event type	Description	Frequency in *slc45a2* Carpet Bomb	Frequency in *tyr* Carpet Bomb
No Indel Events		**59.7%**	**31.8%**
No changes: 	No Break	59.7%	31.8%
Event 1: 	SNP in gRNA or PAM sequences	~0%	~0%
Large Deletion Events		**30.5%**	**50.1%**
Event 2: 	Deletion between gRNA1 & gRNA4	6.5%	24%
Event 3: 	Deletion between gRNA1 & gRNA2/3	15.3%	10.7%
Event 4: 	Deletion between gRNA1 & gRNA2/3 + indel near gRNA4	2.6%	7.9%
Event 5: 	Deletion between gRNA2/3 & gRNA4	2.9%	3.4%
Event 6: 	Deletion between gRNA2/3 & gRNA4 + indel near gRNA1	3.2%	4.1%
Discrete Indel Events		**7.4%**	**16.1%**
Event 7: 	Indel only near gRNA1	2.4%	3.6%
Event 8: 	Indel only near gRNA2/3	1.5%	2.4%
Event 9: 	Indel only near gRNA4	1.5%	3.1%
Event 10: 	Indel only near gRNA1 & gRNA2/3	0.9%	1.2%
Event 11: 	Indel only near gRNA1 & gRNA4	0.6%	3.6%
Event 12: 	Indel only near gRNA2/3 & gRNA4	0.2%	0.7%
Event 13: 	Indel near all gRNAs	0.3%	1.5%
Complex Rearrangement Events		**2.4%**	**2.0%**
Event 14: 	Inversions & DNA rearrangements	2.4%	2.0%

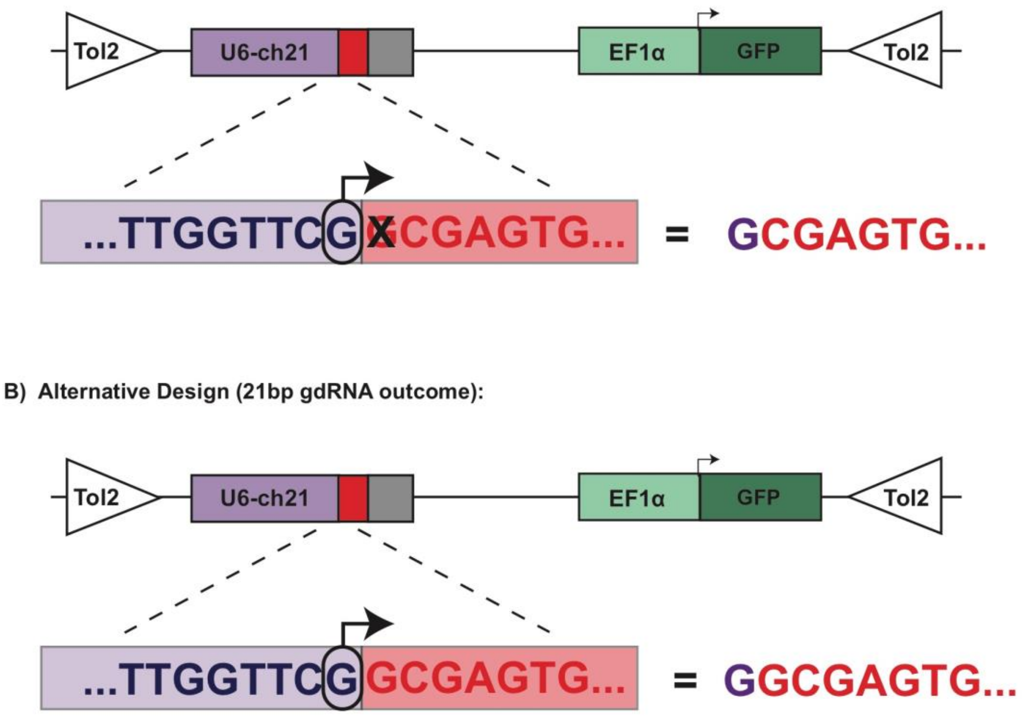
Legend

⌇┈┈⌇ = Deletion ✸ = Local Indel ✱ = SNP ⌇ = Targeted CRISPR cleavage site

Figure 4: Characterization of indel events generated by "carpet bomb" CRISPR vector. Using targeted sequencing results, we were able to identify unique combinations of indels generated by carpet bomb technique. We provide a simplified schematic of each possible indel permutations and present the percentage of total reads that fit within that category.

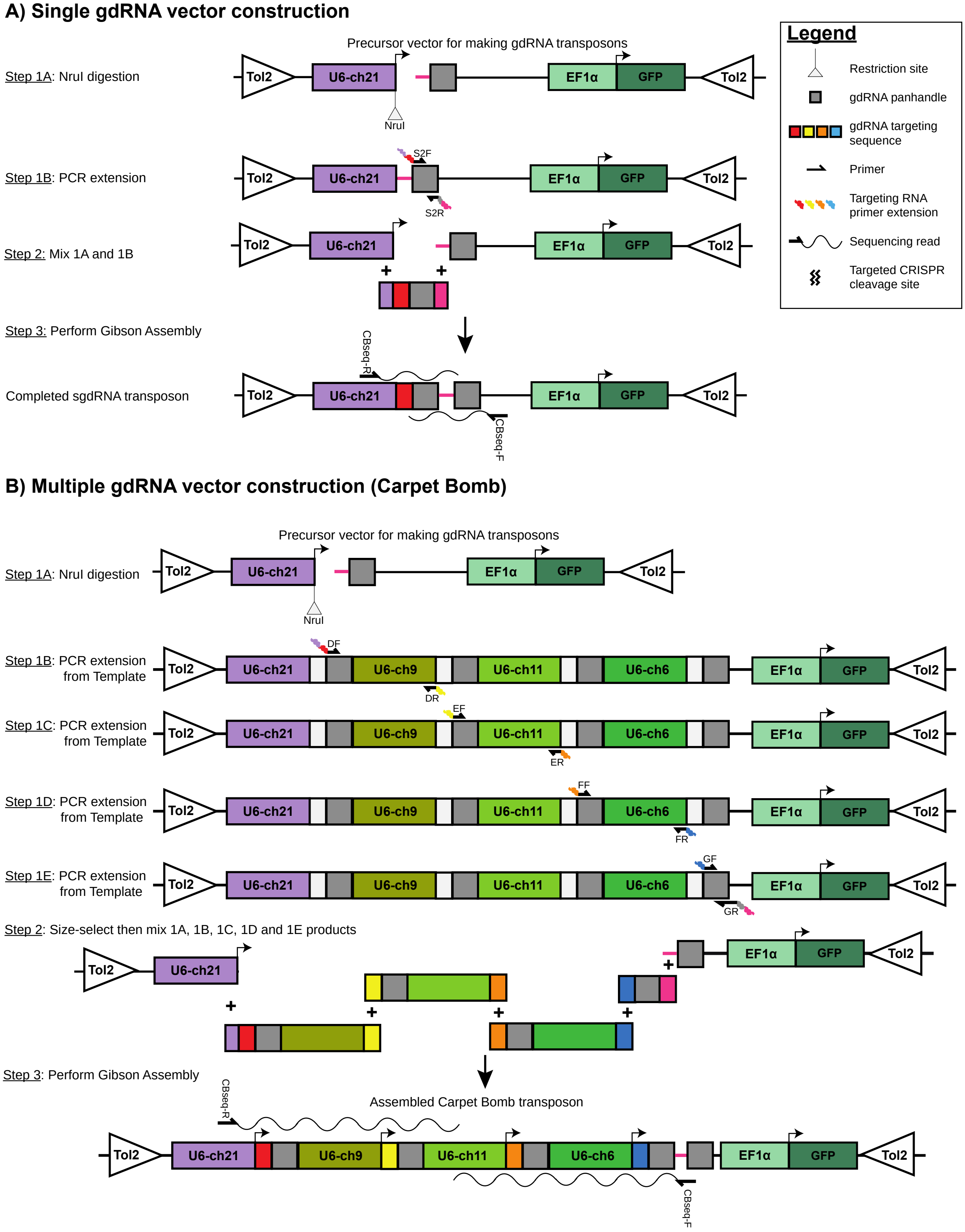
4.8 Supporting Information:

Supplental Figure 1: U6 Transcription Start Site hypothesis



Supplemental Figure 1: U6 transcription start site schematic. (A) If the terminal "G" from U6 promoter is transcribed, then not including the initial "G" in gRNA should lead to conventional 20bp gRNA from being transcribed.

Supplental Figure 2: Alternative CRISPR construction method

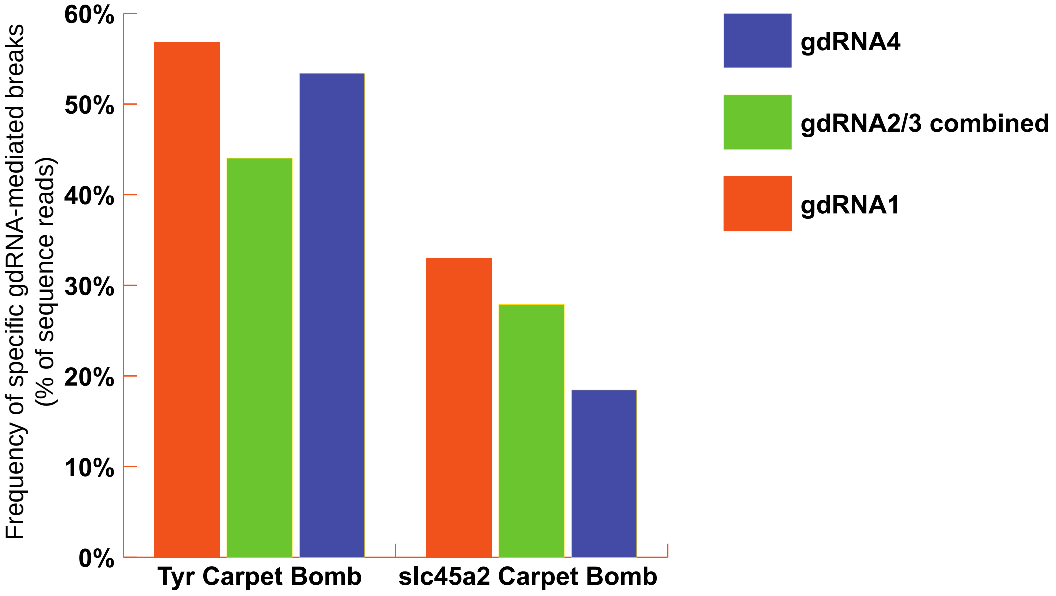


A) Single gdRNA vector construction

B) Multiple gdRNA vector construction (Carpet Bomb)

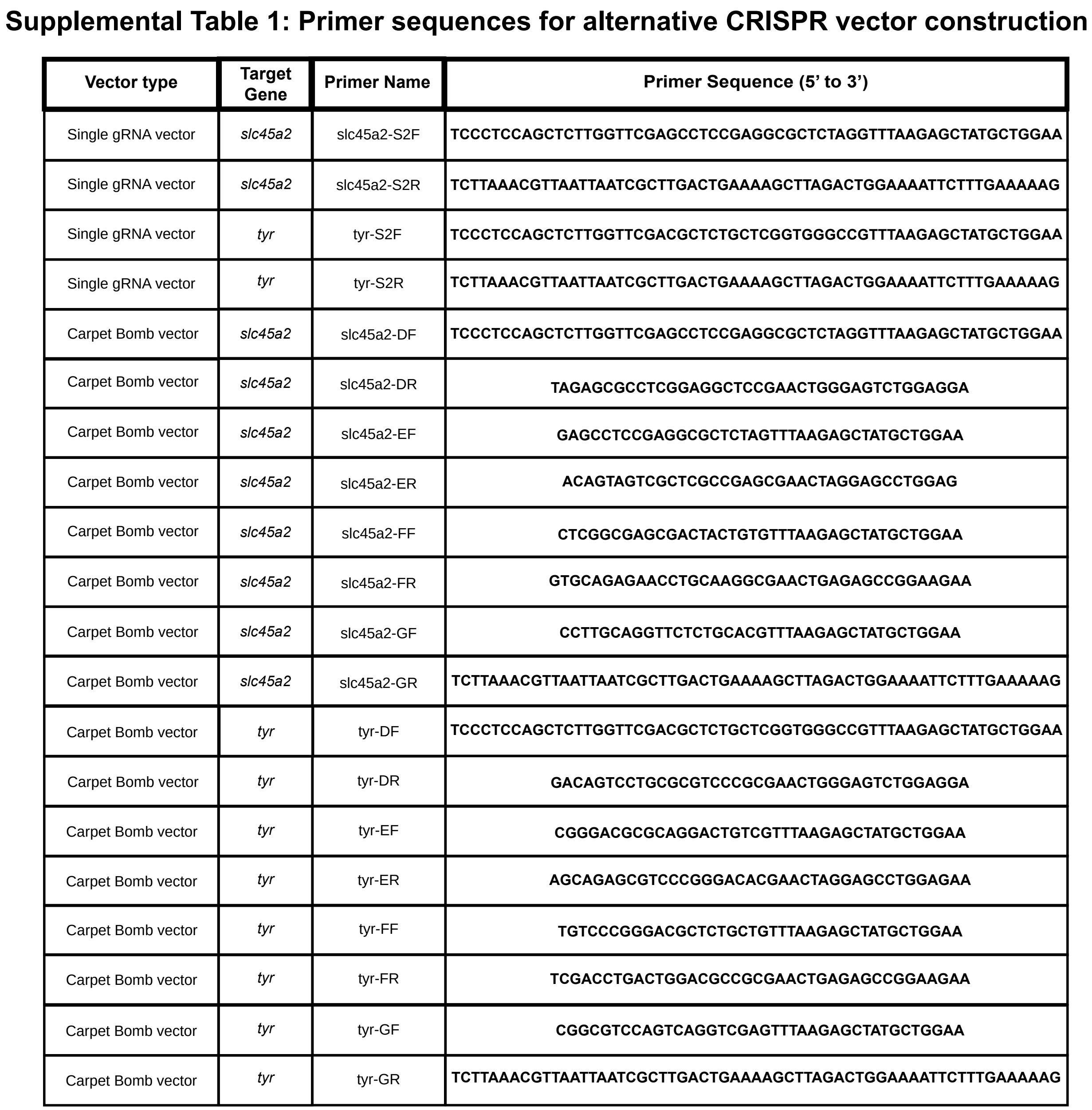
Precursor vector for making gdRNA transposons
Step 1A: NruI digestion
Tol2
U6-ch21
EF1α
GFP
Tol2
NruI
Step 1B: PCR extension
S2F
S2R
Step 2: Mix 1A and 1B
+
+
Step 3: Perform Gibson Assembly
CBseq-R
Completed sgdRNA transposon
CBseq-F
Legend
Restriction site
gdRNA panhandle
gdRNA targeting sequence
Primer
Targeting RNA primer extension
Sequencing read
Targeted CRISPR cleavage site
Precursor vector for making gdRNA transposons
Step 1A: NruI digestion
NruI
Step 1B: PCR extension from Template
DF
DR
U6-ch9
U6-ch11
U6-ch6
Step 1C: PCR extension from Template
EF
ER
Step 1D: PCR extension from Template
FF
FR
Step 1E: PCR extension from Template
GF
GR
Step 2: Size-select then mix 1A, 1B, 1C, 1D and 1E products
Step 3: Perform Gibson Assembly
Assembled Carpet Bomb transposon
CBseq-R
CBseq-F

Supplemental Figure 2: Alternative CRISPR construction method. (A) Instead of oligo annealing and extension, we provide alternative method where primers extend gRNA panhandle sequence from precursor vector, which can be combined with NruI-digested precursor vector for Gibson Assembly. (B) Here, a single template, any carpet bomb vector, can be used to generate PCR products that can be assembled into precursor vector. However, an extra step of size-selection is necessary to remove products from non-specific amplification. Although this is a viable strategy for generating carpet bomb vectors, we highly recommend and support the optimized protocol outlined in Figure 1.



Supplemental Figure 3: No synergistic effect between overlapping gdRNAs

Supplemental Figure 3: No synergistic effect between overlapping gRNAs.



Supplemental Table 1: Primer sequences for alternative CRISPR vector construction

Vector type	Target Gene	Primer Name	Primer Sequence (5’ to 3’)
Single gRNA vector	*slc45a2*	slc45a2-S2F	TCCCTCCAGCTCTTGGTTCGAGCCTCCGAGGCGCTCTAGGTTTAAGAGCTATGCTGGAA
Single gRNA vector	*slc45a2*	slc45a2-S2R	TCTTAAACGTTAATTAATCGCTTGACTGAAAAGCTTAGACTGGAAAATTCTTTGAAAAAG
Single gRNA vector	*tyr*	tyr-S2F	TCCCTCCAGCTCTTGGTTCGACGCTCTGCTCGGTGGGCCGTTTAAGAGCTATGCTGGAA
Single gRNA vector	*tyr*	tyr-S2R	TCTTAAACGTTAATTAATCGCTTGACTGAAAAGCTTAGACTGGAAAATTCTTTGAAAAAG
Carpet Bomb vector	*slc45a2*	slc45a2-DF	TCCCTCCAGCTCTTGGTTCGAGCCTCCGAGGCGCTCTAGGTTTAAGAGCTATGCTGGAA
Carpet Bomb vector	*slc45a2*	slc45a2-DR	TAGAGCGCCTCGGAGGCTCCGAACTGGGAGTCTGGAGGA
Carpet Bomb vector	*slc45a2*	slc45a2-EF	GAGCCTCCGAGGCGCTCTAGTTTAAGAGCTATGCTGGAA
Carpet Bomb vector	*slc45a2*	slc45a2-ER	ACAGTAGTCGCTCGCCGAGCGAACTAGGAGCCTGGAG
Carpet Bomb vector	*slc45a2*	slc45a2-FF	CTCGGCGAGCGACTACTGTGTTTAAGAGCTATGCTGGAA
Carpet Bomb vector	*slc45a2*	slc45a2-FR	GTGCAGAGAACCTGCAAGGCGAACTGAGAGCCGGAAGAA
Carpet Bomb vector	*slc45a2*	slc45a2-GF	CCTTGCAGGTTCTCTGCACGTTTAAGAGCTATGCTGGAA
Carpet Bomb vector	*slc45a2*	slc45a2-GR	TCTTAAACGTTAATTAATCGCTTGACTGAAAAGCTTAGACTGGAAAATTCTTTGAAAAAG
Carpet Bomb vector	*tyr*	tyr-DF	TCCCTCCAGCTCTTGGTTCGACGCTCTGCTCGGTGGGCCGTTTAAGAGCTATGCTGGAA
Carpet Bomb vector	*tyr*	tyr-DR	GACAGTCCTGCGCGTCCCGCGAACTGGGAGTCTGGAGGA
Carpet Bomb vector	*tyr*	tyr-EF	CGGGACGCGCAGGACTGTCGTTTAAGAGCTATGCTGGAA
Carpet Bomb vector	*tyr*	tyr-ER	AGCAGAGCGTCCCGGGACACGAACTAGGAGCCTGGAGAA
Carpet Bomb vector	*tyr*	tyr-FF	TGTCCCGGGACGCTCTGCTGTTTAAGAGCTATGCTGGAA
Carpet Bomb vector	*tyr*	tyr-FR	TCGACCTGACTGGACGCCGCGAACTGAGAGCCGGAAGAA
Carpet Bomb vector	*tyr*	tyr-GF	CGGCGTCCAGTCAGGTCGAGTTTAAGAGCTATGCTGGAA
Carpet Bomb vector	*tyr*	tyr-GR	TCTTAAACGTTAATTAATCGCTTGACTGAAAAGCTTAGACTGGAAAATTCTTTGAAAAAG

Supplemental Table 1: Primer sequences for alternative CRISPR vector construction.

Chapter 5: A Reflection on the Epigenome Dynamics in Zebrafish Pigment Cell Fate

Hyo Sik Jang[1,2]*, Yujie Chen[1,2]*, Alicia N. Wilkening[1,2], Jiaxin Ge[1,2], Jeffery W. McMillian[1,2], Fujr Ibrahim[1,2], Rebecca F. Lowden [1,2], Hyung Joo Lee[1,2], Yiran Hou[1,2], Xiaoyun Xing[1,2], Daofeng Li[1,2], Stephen L. Johnson[1,3], Ting Wang[1,2].

[1]Department of Genetics, Washington University School of Medicine, St Louis, MO 63130, USA.

[2]Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St Louis, MO63108, USA.

[3]Deceased.

*Equal contribution

Author Contributions: HSJ, RFL, SLJ and TW conceived and implemented the study; HSJ, YC and ANW contributed to the computational analysis; HSJ generated transcriptomic and epigenomic profiles of all samples; HSJ, YC, ANW and JG performed the CRISPR-mediated deletion experiments; HSJ, YC, ANW, JG and JWM performed miniCoopR experiments; HSJ and YC performed iridophore counting and picture taking for phenotypes; HJL, YH, XX and DL provided reagents and computational infrastructure; the manuscript was prepared by HSJ, YC, ANW and TW with input from all authors.

This manuscript is currently in preparation for submission.

5.1 Abstract

Resolving the genetic and epigenetic determinants that drive specific cell fate decision in complex organisms has been a long-standing goal in developmental biology. During zebrafish embryogenesis, a population of multipotent embryonic cells, called neural crest, is responsible for the migration and production of biologically unique cell types, such as neurons, bones and pigment cells. Zebrafish pigment cell differentiation, in particular, provides an attractive model for studying cell fate progression as a single neural crest progenitor engenders all three morphologically distinct pigment types: black melanophore (also called melanocytes), yellow xanthophores and reflective iridophores. Nontrivial classical genetic and transcriptomic approaches have revealed essential molecular mechanisms and gene-regulatory circuits that drive neural crest-derived cell fate decisions. However, how the epigenetic landscape contributes to pigment cell differentiation is poorly understood. Here, we chart the global changes in the epigenetic landscape during neural crest differentiation into melanocytes and iridophores to identify epigenetic determinants of pigment cell fate. Motif enrichment in the epigenetically dynamic regions, or potential cis-regulatory elements, revealed putative transcription factors that are responsible for driving pigment cell identity. Through this effort, in the relatively uncharacterized iridophores, we define a network of transcriptions factors that are predicted to bind to regulatory elements directly upstream of genes linked to guanine synthesis cycle, which are essential for iridophore function.

5.2 Introduction

Development of a multicellular organism is an intricate process of expansion and diversification of a pluripotent cell population. Rapidly following embryogenesis, the genome of stem cells experiences extensive biochemical and structural changes that allow these multipotent progenitor cells to faithfully commit and differentiate into various tissue and cell types. These decisions are often reflected by unique gene expression profiles and are shaped by epigenetic programs[180,181]. Although monumental consortium level efforts, such as ENCODE[182] and Roadmap Epigenomics[183], have significantly advanced the field of developmental epigenetics, these studies have mostly focused on profiling human and mouse model systems.

Zebrafish is one of the organism models that are widely used in biological field for various advantageous properties[184]. Zebrafish can rapidly mature into adulthood and need little space to propagate thus are cost-efficient. Numerous genetic-manipulation and cell-labeling technologies are available to interrogate how genetic elements impact development and disease. More importantly, zebrafish has a transparent embryo, which makes zebrafish an attractive model for embryonic development and cell fate dynamic studies. Zebrafish has three main pigment cell types[185], black melanocyte, reflective iridophore, and yellow xanthophore, which are all derived from a multipotent neural crest cell (NCC) population (**Fig. 1A**). Interestingly, NCCs can also differentiate into various morphologically and functionally distinct cell types, such as glia, neurons, cartilage, connective tissue and pigment cells[186]. How a single cell population with the same genetic content could generate such diverse cell types has been an active field of research in the developmental biology. In the case of pigment cell differentiation, previous work established that a subpopulation of NCCs commit to pigment cell fate (called pigment progenitor cells), then further differentiate into the mature iridophores or melanocytes[187,188].

Immense mutagenesis experiments in zebrafish provided insights into the genetic regulation and gene regulatory networks responsible for pigment cell differentiation[189–194]. Melanocyte development has been extensively studied for its translational potential in tackling melanoma cancer. In melanocytes, Sox10[195] and Wnt signaling[196] are required to activate and maintain *mitfa* transcription, which is an essential transcription factor regulating numerous melanocyte differentiation genes, including those controlling melanin synthesis[197]. Although relatively understudied, molecular mechanisms governing iridophore cell fate have been discovered in forward genetic screens. In iridophore development, *pnp4a*[198] was shown to encode an enzyme important in the biosynthesis of guanine, an important molecule responsible for the reflective characteristic in iridophores. PKA (Protein Kinase A) signaling[199] inhibits iridophore differentiation while promoting differentiation of melanophores in zebrafish larvae. Alk (Anaplastic lymphoma kinase) and Ltk (leucocyte tyrosine kinase) ligands[200] are essential for iridophore development. Sox10, which regulates the expression of transcription factor *mitfa* in melanocyte, is continuously expressed throughout development of the iridophore lineage[191]. Foxd3 transcription factor represses *mitfa* in the iridophore pigment progenitor cells to bias differentiation towards iridophore cell fate[201,202]. Recently, *tfec*[203,204] and *gbx2*[205] have also been implicated in iridophore cell fate.

The epigenetic dynamics that govern pigment cell fate is relatively unexplored in the zebrafish model. Here, we aim to fill this gap and provide high quality epigenetic landscape profiles of various stages of NCC differentiation into melanocytes and iridophores. Furthermore, Comparative epigenetics can be a powerful tool in deciphering both the genetic and epigenetic mechanisms that govern cell fate. Here, we characterize and leverage DNA methylation and chromatin accessibility dynamics to chart putative gene regulatory networks that govern pigment cell fate and reveal that *alx4a* is necessary and sufficient for iridophore development on the zebrafish body.

5.3 Results

5.3.1 Pigment cell differentiation is demarcated by cell-type specific loss of DNA methylation.

To characterize DNA methylation dynamics that govern NCC development and differentiation into pigment cells, we generated whole genome bisulfite sequencing (WGBS) libraries of early NCC (15-somite), fate-determined NCC (24hpf), and differentiated pigments cells, melanocytes and iridophores (**Fig. 1B**). We generated two biological replicates of each timepoint and sequenced to capture ~15 million CpGs with coverage ≥5 (**Fig. 2A**). Although there seem to be indiscernible difference in methylation distribution (**Fig. 2B**), global DNA methylation levels show slight decrease across NCC differentiation into pigment cells (~85% to ~81%, **Fig. 2C**). However, the variation in DNA methylation across the samples can separate samples based on cell identity as represented by CpG methylation correlation (**Fig. 2D**) and Principal component analysis (PCA) (**Fig. 2E**). To increase confidence in CpG methylation levels, biological replicates are combined so that almost 75% of CpGs have ≥5 coverage (**Fig. 3A**). The global DNA methylation levels remain similar in the combined samples (**Fig. 3B**) to show modest loss of DNA methylation throughout differentiation.

Since DNA methylation difference across samples can demarcate cell identity, the modest loss of DNA methylation could reflect abundant cell type-specific and local DNA methylation gain and loss. To better understand the DNA methylation dynamics that determine pigment cell fate, we identified differentially methylated regions (DMRs) using DSS[206]. We identified thousands of local DMRs (size ranging from 50 bp to 1000 bp, **Fig. 3C**) and found that pigment differentiation is accompanied with largely local loss of methylation and very minimal gain of methylation. In fact, >99% of DMRs between 24hpf NCC and differentiated pigment cells are hypoDMRs in differentiated pigment cells (**Fig. 3D,E**). We also note that melanocytes and iridophores share regions that undergo similar magnitude of methylation change (**Fig. 3F**) from 24hpf NCC. Considering the recent discovery of a bipotent pigment progenitor, melanoiridoblast[191,202], that can differentiate into both melanocytes and iridophores, these shared DMRs could predict the DNA methylation landscape of the intermediate pigment progenitor cell.

5.3.2 Dynamic transcriptomic landscapes reveal physiologically relevant genes and transcription factors responsible for pigment differentiation

Since loss of DNA methylation is often associated with gene activation[207], we asked whether gene expression dynamics during pigment differentiation reflected the epigenetic activation phenomenon. We have previously characterized the transcriptomic dynamics between 24hpf embryos vs melanocytes, retinal pigment cells and iridophores[208]. However, the transcriptomic landscape was generated from whole 24hpf so NCC gene expression could be masked by other cell types. To address this issue, we isolated NCC population and generated mRNA-seq libraries to provide better resolution in the gene expression dynamics during NCC to pigment cell differentiation. Furthermore, we characterized differentially expressed genes using DESeq2[209] to identify statistically significant differences. As expected, 15-somite and 24hpf NCC cluster closely together while the two pigment cell types are dispersed based on gene expression variation as represented by hierarchical clustering (**Fig. 4A**) and PCA analysis (**Fig. 4B**). Known gene markers are differentially expressed in appropriate cell types (**Fig. 4C**) reflecting robust quality of mRNA-seq libraries. Although pigment cell fate is coupled with loss of DNA methylation, we report relatively balanced gene expression dynamics where hundreds of genes are up- and down-regulated (**Fig. 5A**). We performed gene ontology (GO) enrichment of the DEGs by using Metascape[133] and report biologically relevant processes (**Fig. 5B**). For example, the top GO hit for genes downregulated from 15-somite to 24hpf NCC transition is tube development, which can represent neural tube formation that occurs 18-20hpf[210]. Furthermore, GO enrichment for melanocyte-specific genes reflect pigmentation and membrane-transport protein while iridophore-specific genes enrich for small molecule biosynthetic process and purine synthesis that could be responsible for guanine crystal stacks that give iridophore its reflective properties. We also note that genes that are specifically down-regulated specifically in iridophores, but not in melanocytes, enrich for neuronal GOs suggesting that iridophore progenitor cells might share more similarity to neuronal cells than melanocytes.

Epigenetic landscape is often intricately tied with transcription factor (TF) presence[211]. To better understand how transcription factors might influence the epigenetic dynamics that govern pigment cell fate decisions, we identified differentially expressed transcription factors for each cell type (**Fig. 6**). Since melanocyte development is relatively well-characterized, we focused on potential transcription factors that might drive iridophore differentiation. *sox10* and *tfec* have already been characterized to be important for iridophore differentiation[191,204]. As identified in previous efforts[208], we confirm that *alx1, alx3, alx4a, alx4b, ets1,* and *gbx2* are highly expressed in iridophore. Recently, morpholino knockdown of *gbx2* have been shown to diminish iridophore count in zebrafish larvae suggesting that *gbx2* is essential for iridophore differentiation[205]. We report other TFs, such as *hsf5, srebf1, foxi3b, nfkb2, tbx2a,* and *zbtb2a*, that are preferentially expressed in iridophores. Whether these TFs are important for determining iridophore identity warrants further investigation.

5.3.3 Chromatin accessibility potentially fine-tunes gene expression during pigment differentiation

Since DNA methylation cannot fully explain down-regulation of genes during differentiation, we hypothesized that chromatin accessibility must be playing an essential role in epigenetic suppression of gene activity. To explore this hypothesis, we generated Assay for Transposase-Accessible Chromatin with sequencing (ATAC-seq[212]) to identify chromatin accessibility dynamics across pigment development. We identified >100,000 ATAC peaks (**Fig. 7A**), which represent cell-type specificity as shown by hierarchical clustering (**Fig. 7B**) and PCA analysis (**Fig. 7C**). We identified differentially accessible regions (DARs) using DiffBind[213] and show that DARs are similar in size as DMRs (**Fig. 7D**). We report tens of thousands of regions that are closing, but only a few thousand regions opening, in both melanocyte and iridophore during pigment differentiation (**Fig. 7E**). Furthermore, thousands of regions are closing and opening in cell type-specific manner. These results suggest that although majority of the DNA methylation dynamics favor epigenetic activation, chromatin accessibility fine-tunes the gene regulatory network defining cell identity.

5.3.4 Dynamic DNA methylation and chromatin accessibility regions denote potential cis-regulatory element

DNA methylation and chromatin accessibility dynamics can co-exist to influence epigenetic control. Therefore, we characterized various dynamics of differentially methylated and accessible regions (DMARs) that can occur (**Fig. 8A**). If a DMR overlaps a DAR, we combined those two regions into one DMAR. There are thousands of regions that are both undergoing active DNA demethylation and opening (increasing in chromatin accessibility), which we classify as dynamic DMAR (**Fig. 8B**). The dynamic DMARs don’t increase in size relative to DMRs and DARs (**Fig. 9A**), which suggest that the overlapping DMRs and DARs are similar in size and co-occur in the same genomic vicinity. For DMARs that have both DNA methylation and chromatin accessibility associated with epigenetic activation (opening hypoDMARs), we hypothesize that these are more likely to function as regulatory elements than solo DMRs or

DARs, which might have antagonistic epigenetic marks. For example, we report that out of ~6,000 iridophore-specific solo hypoDMRs, ~4,000 hypoDMRs occur in closed chromatin regions (**Fig. 9B**). Even though the DNA methylation change correlates with epigenetic activation, there will be no functional consequence of the loss of DNA methylation since the chromatin is closed off and no transcription factor can access that region. Similarly, many solo opening DARs occur in regions with relatively high methylation (**Fig. 9C**). Therefore, we decided to focus on characterizing dynamic DMARs to decipher potential cis-regulatory roles that regulate cell fate decisions.

To better understand how epigenetic dynamics might shape gene expression, we explored promoter epigenetic status of DEGs. 88-90% of promoters of up-regulated genes are static in their epigenetic status from 24hpf NCC to pigment cell differentiation (**Fig. 10A**). Although a small fraction of down-regulated genes might be repressed by loss of promoter accessibility, majority of DEGs' promoters don't experience any epigenetic change. This result suggests that gene expression is more likely to be controlled by DMARs in enhancer context. Indeed, majority of the DMARs are present in intergenic or intronic regions (**Fig. 10B**), which if epigenetically active will likely provide a cis-regulatory role.

Transcription factors bind to enhancer regions to increase transcription of nearby genes[5]. To see if certain transcription factors might be binding to DMARs to provide enhancer-like function that regulate pigment differentiation, we identified motif modules that are enriched in melanocyte-specific opening DMARs and iridophore-specific opening DMARs (**Fig. 11A**). As expected, in melanocytes, we see an enrichment of TFAP-related motifs and MiT motifs, which correspond to *tfap2a* and *mitfa* TFs that regulate melanocyte differentiation[214]. Since genetic mechanisms that drive iridophore cell fate are relatively underexplored, we leveraged epigenetic information generated in this study to chart how the genetic factors intertwine with epigenetic dynamics to define iridophore cell fate. EN2 motif (homeobox-related) was highly enriched specifically in iridophore-specific opening DMRs, DARs and DMARs (**Fig. 11B, Fig. 12**). We asked which TFs within EN2 clusters were differentially expressed in iridophores and identified that the aristaless homeobox TFs (*alx1, alx3, alx4a*, *alx4b* and *gbx2)* and *pax7* paralogs were highly expressed. Motif footprinting analysis with CENTIPEDE[215] revealed strong footprinting signatures in iridophore ATAC peaks, indicative of TF binding (**Fig. 11C**). Further analysis into other motif cluster enriched in iridophores revealed known and novel TFs, such as *sox10, tfec, ets1,* and *hey1*, with positive motif footprinting signatures (**Fig. 13A,B**).

5.3.5 Iridophore-specific TFs putatively regulate genes in guanine synthesis cycle.

With epigenetic landscape data, we can start predicting how TFs might regulate gene regulatory networks crucial for cell biology and identity. In iridophores, we reveal that many genes in the guanine synthesis cycle are significantly up-regulated. When analyzing iridophore-specific DMARs within 50kb of guanine synthesis DEG promoters, we were surprised to find that almost all of these DMARs contain at least one instance of *alx, sox10,* and/or *tfec* motifs (**Fig. 14A**). Almost all iridophore-specific DEGs responsible for guanine generation and transport have at least one DMAR with *alx* motif (**Fig. 14B**), suggesting the putative regulatory potential of *alx* TFs for iridophore's reflective characteristic.

Furthermore, we asked how these iridophore TFs might be turned on and regulated during NCC differentiation into iridophores. When we scanned for activating differentially methylated and/or accessible regions (DM/ARs) near iridophore-specific TF promoters, we discovered that *alx4a* promoter had 15 DM/ARs with iridophore-related TF motifs (14 upstream intergenic of promoter and 1 intronic) (**Fig. 15A,B**). This result could represent the robust activation of *alx4a* expression that is critical for guanine production in iridophores but not melanocytes. By leveraging DMARs and motif presence near important iridophore TFs, we can construct a putative transcription factor network that drive iridophore cell fate (**Fig. 15C**), but further work must be done to validate which TFs are necessary for iridophore development.

5.3.6 *alx4a* is essential for iridophore differentiation in the body, but not the eye.

Since previous work has established that *gbx2* impacts iridophore development[205], we focused our attention on validating the necessity of *alx* TFs for iridophore differentiation. Currently, very little is known about the function of aristaless homeobox TFs in vertebrate development. Mutations in *Alx3* and *Alx4* is known to cause craniofacial abnormalities in humans and mice while mutations in *alx1* disrupts proper neural crest migration to cause frontonasal dysplasia in zebrafish[216–218]. Since *alx* TFs has only been studied in context of craniofacial development, we utilized CRISPR-Cas9 technology to introduce indels in exon 1 or 2 to create frameshift mutations in *alx1*, *alx3*, *alx4a*, *a*nd *alx4b* genes to investigate how knockout of these TFs impact iridophore differentiation. We report that iridophores develop normally in *alx1*, *alx3*, and *alx4b* KO fish with some instances of pigment pattern defect in the caudal fin (**Fig. 16A**). However, *alx4a* KO fish revealed complete ablation of iridophores in the body, but not the eye (**Fig. 16B**). We note rare instances of iridophore escape in the *alx4a* KO fish, but most fish result in complete loss of body iridophores in adults and 4dpf larvae (**Fig. 16C**). The *alx4a* mutant fish looks similar to *shady, rse,* and *tra* mutant fish with the exception of preserving eye iridophores[219,220]. The presences of iridophores in the eye suggests that an alternative gene regulatory network is responsible for eye iridophore differentiation, analogous to *otx* TFs' role in eye pigment development while *mitfa* regulates melanocyte differentiation in the body[221]. It's curious to ponder why *alx1, alx3*, and *alx4b* is highly expressed in iridophores, but have no functional consequence to iridophore development. One hypothesis could be that the trans-acting factors that activate *alx4a* or *alx4a* TF itself could also lead to transcription of *alx* genes, which can create a robust gene regulatory module that maintains high expression of iridophore-related TFs.

5.3.7 Ectopic expression of *alx4a* and *gbx2* biases pigment cell fate towards iridophores

Since *alx4a* and *gbx2* is necessary for proper iridophore development, we asked whether either TF was sufficient to push pigment cell fate towards iridophores. To ectopically express the TFs in early pigment progenitor cells, we took advantage of the miniCoopR transgenesis vector[222,223]. The miniCoopR vector consists of two *mitfa* promoters driving EGFP and *mitfa* minigene expression flanked by tol2 sequences, which allow transgene integration into the zebrafish genome. *Mitfa* is expressed as early as 18hpf and is expressed in bipotent pigment progenitor cells called melanoiridoblast[202]. The balance between *foxd3* and *mitfa* levels are responsible for bias towards iridophore or melanocyte cell fate[201,202]. Similarly, we asked whether early expression of *alx4a* and *gbx2* can bias the melanoiridoblast to differentiate into iridophores (**Fig. 17A**). Therefore, we replaced the *mitfa* minigene with *alx4a* and *gbx2* CDS in the miniCoopR vector and evaluate how pigment development is impacted (**Fig. 17B**). In both miniCoopR-*alx4a* and miniCoopR-*gbx2* transgenic fish, melanocyte differentiation and migration are diminished during embryo development (**Fig. 17C**). MiniCoopR-*alx4a* transgenic 3dpf larvae have increased number of iridophores than wild type (WT) larvae indicating that *alx4a* is sufficient to bias pigment cell fate towards iridophores. MiniCoopR-*gbx2* 3dpf larvae had significantly higher than WT, but less than miniCoopR-*alx4a*, iridophore counts suggesting that the role of *gbx2* is to suppress melanocyte development. Although embryonic melanocytes are present in 5dpf transgenic larvae, adult transgenic fish present almost complete ablation of melanocytes, reminiscent of *nacre/mitfa* mutant fish (**Fig. 17D**). Furthermore, we report varying levels of melanocyte ablation in various F1 adults from different founders (variable levels of integration of miniCoopR vector) suggesting that the TF expression levels could be intricately tied with melanocyte development in zebrafish. Considering that adult melanocytes are mostly derived from adult melanocyte stem cells, *alx4a* and *gxb2* could be repressing melanocyte differentiation or migration in the adult melanocyte stem cells, but have minimal impact on embryonic melanocyte development.

5.4 Discussion

In this study, we provide one of the first insights into the epigenetic dynamics that shape neural crest differentiation into pigment cells in zebrafish. By taking advantage of flow cytometry, we isolated enriched populations of NCC from 15somite and 24hpf embryos. We adapted the pigment isolation protocol from Higdon et al. 2013[208] to isolate melanocytes and iridophores. From these samples, we profiled DNA methylation, chromatin accessibility and transcriptomic landscapes to create comprehensive epigenetic maps that define pigment cell fate. Surprisingly, we found that cell differentiation in zebrafish is characterized by promiscuous loss of DNA methylation coupled with dynamic chromatin accessibility. We report that epigenetic status of DEG promoters are often static and majority of dynamic epigenetic changes occur in the intergenic or intronic regions. This suggests that gene regulatory networks that define pigment cell fate are mostly regulated by enhancer-like cis-regulatory elements rather than promoter dynamics. There are many shared DMRs, DARs and DMARs between melanocytes and iridophores. Recent clonal tracing studies revealed a bipotent pigment progenitor that have potential to differentiate into melanocytes and iridophores[188]. The shared epigenetic dynamics identified could represent the intermediate epigenetic landscape of the bipotent cells and warrants further investigation. By charting the intermediate epigenetic landscape, we can identify why these progenitor cells are restricted to two pigment cell fates and provide epigenetic building blocks behind cell fate logic.

Here, we also provide first insight into epigenetic dynamics that define iridophore development. Our efforts discovered that iridophore-specific DMARs enrich for motifs from homeobox-containing transcription factors. By pairing differential gene expression data, we provide putative gene regulatory network, potentially regulated by aristaless homeobox transcription factors, that is important for iridophore physiology. Indeed, loss of *alx4a* transcription factor ablated iridophore presence in zebrafish highlighting the strength that epigenetic-based analysis can provide in studying cell fate decisions. Surprisingly, we note that iridophores in the eye are not impacted by *alx4a* KO suggesting a separate differentiation pathway or potential functional redundancy for eye iridophore development. Furthermore, ectopic expression of *alx4a* and *gbx2* in early pigment progenitor cells biases cell fate against melanocyte differentiation and almost ablates melanocyte presence in adult transgenic fish. We report that miniCoopR-*alx4a* transgenic larvae have higher iridophore counts at 3dpf compare to wild type, while miniCoopR-*gbx2* transgenic larvae had higher, but less than miniCoopR-*alx4a,* number of iridophores. These results indicate that *alx4a* and *gbx2* are necessary and sufficient for iridophore cell fate.

It is intriguing that both loss of either *alx4a* or *gbx2* leads to preventing iridophore differentiation. This suggests that *alx4a* and *gbx2* have non-redundant function in regulating iridophore cell fate. However, both TFs have very similar DNA binding motifs, so investigating where each TFs binds within the iridophore genome would be the up most important next step in deciphering how these TFs differentially regulate the epigenome and transcriptome.

5.5 Materials and Methods

5.5.1 Zebrafish maintenance and strains.

All fish procedures for this study were carried out following strict guidelines outlined in protocol #20140195 and #20160109 approved by Washington University Animal Use Committee. The zebrafish strains utilized in this study were maintained according to standard conditions defined previously[224]. Neural crest cells were collected from *Tg(crestinA:EGFP)* line, in which 1,200 bp of *crestin* element (*crestinA*) was cloned upstream of EGFP and integrated into the genome via Tol2 transgenesis[149]. Differentiated melanocytes and iridophores were collected from *mlpha* strain[225], a *melanophilin* mutant strain that displays reduced dispersion of melanosomes in melanocytes. We chose *mlpha* to circumvent residual EGFP expression in *Tg(crestin:EGFP)* lines that might interfere with FACS isolation of pigment cells. For CRISPR and miniCoopR experiments, we utilized *AB** strain for its availability and wild type-like pigment characteristics and development.

5.5.2 Neural crest cell and pigment cell isolation

Tg(crestinA:EGFP) labels neural crest cells (NCCs) from 14-15 somite stage (neural crest formation) to differentiation into pigment cells.

For 15-somite and prim-5 (24 hpf) neural crest cell isolation, *Tg(crestinA:EGFP)* embryos at designated biological time points were dechorionated with 20mg/mL Pronase (Millipore Sigma, 10165921001), rinsed with egg water to remove chorion, and collected into 1.5ml Eppendorf tubes on ice. 15-somite embryos were dissociated into single cells with deyolking buffer (55mM NaCl, 1.8mM KCl, and 1.25mM $NaHCO_3$) and gentle pipetting. 24hpf embryos were single-cell dissociated by adding Gibco TrypLE Express enzyme solution (ThermoFisher Scientific, 12604021) and incubating at 37°C for 10 minutes followed by pipetting. To remove dissociation buffer, single-cell dissociated samples were pelleted by centrifugation at 300x g for 8 minutes at 4°C and the supernatant was discarded. The cell pellet was resuspended in 1× PBS + 2% FBS solution and filtered through 100μM CellTrics filters (Sysmex-Partec, 04-004-2328). Samples were pelleted and resuspended and kept on ice for subsequent FACS process. 7-AAD dye (ThermoFisher Scientific, A1310) was added to sample 10 minutes prior to flow cytometry to label dead cells. Neural crest GFP-positive cells were sorted and collected on Beckman Coulter MoFlo using 70μM nozzle.

For melanocyte and iridophore isolation, we adapted previously published protocol[208] developed by Johnson lab. In brief, 4-5dpf *mlpha* larvae were anesthetized with Tricane for 15 minutes and collected into 50ml conical tubes on ice. After removing egg water, the larvae were digested with Gibco TrypLE Express enzyme solution in 37°C shaking incubator (200rpm) for 30 minutes. The larvae solution was filtered with 120μM to collect dissociated cells. Melanocytes and iridophores were isolated via Percoll (Millipore Sigma, P1644) density centrifugation. Purified pigment cell solution was further processed on Beckman Coulter MoFlo (100μM nozzle) to separate melanocytes and iridophores as detailed previously[208].

5.5.3 Epigenome and transcriptome sequencing library construction

Genomic DNA (gDNA) for whole genome bisufilte sequencing (WGBS) was purified from NCCs and pigment cells via phenol-chloroform:isoamyl alcohol (PCI) extraction and ethanol precipitation method. 500ng of gDNA was bisulfite treated using EZ DNA Methylation-Direct kit (Zymo, D5020) and processed with TruSeq DNA Methylation Kit (Illumina, 15066014) to generate Illumina-compatible WGBS libraries.

Chromatin accessibility maps were generated from 15K-50K NCC and pigment cells by following previously published ATAC-seq method[212].

We isolated total RNA via TRIzol Reagent (ThermoFisher Scientific, 15596026) following manufacture's recommendation. Then total RNA was treated with TURBO DNase (ThermoFisher Scientific, AM2238) to remove any residual DNA contamination. mRNA-seq libraries were constructed with TruSeq RNA Library Prep Kit v2 (Illumina, RS-122-2001) following manufacturer's instructions.

All libraries were sequenced on the Illumina NextSeq 500 platform (75bp paired-end reads).

5.5.4 Identification of differentially methylated regions (DMRs)

Paired-end reads from WGBS libraries were trimmed for adapter sequences with Cutadapt[226] and mapped to danRer10 reference genome using Bismark[122] aligner with the following options: "-N 1 -L 28 –score_min L,0,-0.6". Redundant aligned reads were identified and removed using Picard[227] MarkDuplicates command (http://broadinstitute.github.io/picard/). Bismark_methylation_extractor command from Bismark and a custom script were used to calculate DNA methylation levels for each CpG.

To identify DMRs, biological replicates were combined to improve coverage of CpGs and then processed using DSS pipeline[206] with standard parameters plus "smoothing=TRUE, delta=0.30 (at least 30% methylation difference), and p.threshold=0.01". DNA methylation Pearson Correlation plot was generated using "corrplot" package in R while other figures were generated using custom R scripts.

5.5.5 Identification of differentially expressed genes (DEGs) and gene ontology enrichments

mRNA-seq libraries were adapter-trimmed and aligned to the danRer10 using STAR[228]. Gene transcript abundance (RPKM) was calculated with StringTie[229] using Danio_rerio.GRCz10.85.gtf as reference. Also, we processed aligned reads with HTSeq[230] to generate a count matrix for each gene, which was subsequently processed using DESeq2[209] to identify differentially expressed genes. More specifically in DESeq2, we identified significantly differentially expressed genes by filtering for only genes with counts >1, fold change >2 and p-value < 0.01. DEG expression plot was generated using Maplot function in DESeq2. Hierarchical clustering based on RNA expression was generated using "pheatmap" package[231] in R.

To identify which gene ontologies are enriched in DEGs across NCC and pigment cells, we further filtered the DEGs identified by DESeq2 for genes with RPKM > 5 to remove lowly expressed genes. The list of DEGs was processed by Metascape[133] for GO term enrichment.

Since no comprehensive zebrafish transcription factor (TF) list was available at the time of analysis, we manually curated a zebrafish TF list with AnimalTFDB 2.0[232]. Human TFs were converted into zebrafish orthologs using OrthoRetriever (http://lighthouse.ucsf.edu/orthoretriever/). Human TFs with no zebrafish orthologs detected by OrthoRetreiver were manually converted through literature search. Differentially expressed TF heatmaps were visualized using "ComplexHeatmap" package[233] in R.

5.5.6 Identification of ATAC peaks and differentially accessible regions (DARs)

ATAC-seq reads were trimmed for adapter sequences and aligned to danRer10 genome using bwa (bwa mem)[234]. Duplicate reads were removed with Picard MarkDuplicates. Then the libraries were downsampled to 35 million aligned reads to minimize artifacts introduced by library size difference for peak calling analysis. Since the ends of the reads represent Tn5 insertion locations, we processed the aligned reads by offsetting + strand reads by +4bp and – strand reads by -5bp. The offset position for each read was used as input for calling peaks with MACS2[235] using the following parameters: "-g 1.4e+9 -B –SPMR –keep-dup all –nomodel -s 75 –extsize 73 –shift -37 -p 0.01". With narrowPeak output from MACS2, we utilized irreproducible discover rate (IDR) framework[236] to generate a consensus peak file from each biological time point. To identify differentially accessible regions, we processed ATAC peaks with DiffBind[213] with a stringent cutoff of FDR <0.001.

5.5.7 Identification and characterization of differentially methylated and/or accessible regions (DMARs)

Differentially methylated and/or accessible regions were classified by identifying overlapping DMRs and DARs with BEDTools[237] intersect command. DMARs were annotated for genomic location using HOMER[238] annotatePeaks.pl. Furthermore, we performed BEDTools intersect to detect DMARs located within 50kb of DEG promoters,

ll DMRs, DARs and DMARs were processed with HOMER findMotifsGenome.pl to discover which known motifs are enriched in these epigenetically dynamic regions. Since Homer known motif database could be missing particular TFs expressed in zebrafish, we generalized the top 20 hits from HOMER by classifying each as a particular motif cluster/module defined by Roadmap Epigenomics Consortium (https://egg2.wustl.edu/roadmap/data/byDataType/motifanalysis/pouyak/viewByCluster/bycluster.html)[239]. For example, Phox2a motif is part of the EN2 module along with Lhx1, Lhx2, Lhx3 and Pax7. By partitioning motifs into modules, we can identify which particular cluster is specific to certain cell type and then analyze the expression pattern of TFs belonging to that cluster to potentially identify biologically relevant TFs. The motif enrichment plot was generated by averaging the p-values and % of target sequences with motif for each hits of a motif cluster. Motif footprinting in DMARs were generated by CENTIPEDE[215]. For each DMAR, we used FIMO[240] to scan and detect presence of particular motifs.

5.5.8 CRISPR-mediated knockout of *alx* transcription factors in zebrafish

To design gRNA sequences, we took advantage of CRISPOR[241] and CRISPRscan[242] algorithms to maximize specificity (CRISPOR) and efficacy (CRISPRscan). For each gRNA, a primer was ordered with the chosen gRNA sequence preceded by "aattaatacgactcactata" and followed by "gttttagagctagaaatagc." Each gRNA primer was then annealed to the universal primer scaffold, "ttttgcaccgactcggtgccactttttcaagttgataacggactagccttattttaacttgctatttctagctctaaaac". The sgRNAs were then transcribed *in vitro* using T7 RNA polymerase from the HiScribe™ T7 Quick High Yield RNA Synthesis Kit (New England Biolabs, E2050S). Cas9 mRNA was generated via *in vitro* RNA transcription of pCS2-nls-zCas9-nls plasmid (Addgene, 47929) with mMessage mMachine SP6 Transcription Kit (ThermoFisher Scientific, AM1340).

For each target gene, a 5ul injection cocktail was made with 2 µg of Cas9 mRNA, 0.5 µl of 1% or 0.5% phenol red dye, 400 ng of each of the two sgRNAs targeting a gene of interest. 0.5 nL of the CRISPR cocktail was injected directly into the cell of single cell embryo (AB*). To identify founders with indels in target genes, we pair-wise crossed CRISPR-injected adult fish and collected embryos to PCR amplify target gene locus and perform T7 endonuclease I (NEB, M0302S) assay. All homozygous indels were verified via Sanger sequencing.

5.5.9 Ectopic expression of *alx4a* and *gbx2* in pigment progenitor cells

To ectopically express transcription factors in pigment progenitor cells, we exploited the miniCoopR system[223,243]. We generated *alx4a* and *gbx2* CDS fragment from PCR amplifying cDNA from reverse transcribed 24hpf AB* mRNA. Since early pigment progenitor cells express *mitfa*, we cloned in candidate CDS in lieu of *mitfa* minigene via Gibson Assembly. We injected approximately 1nl of 100 ng/µl miniCoopR vector and 15 ng/µl Tol2 capped transposase mRNA cocktail into the yolk of single cell AB* embryos. All GFP+ F0 embryos were raised to adulthood and screened for founders. F1 stable lines were then established by crossing F0 fish.

5.6 Acknowledgements

We would like to thank J. Hoisington-López and M.L. Jaeger from The Edison Family Center for Genome Sciences & Systems Biology (CGSSB) for assistance with sequencing; B. Koebbe and E. Martin from CGSSB for data processing; M. Savio, M. Patana & D. Schweppe from the Siteman Flow Cytometry Core for FACS-related expertise; This work was funded by NIH grants 5R01HG007175, U24ES026699, and U01HG009391 and the American Cancer Society Research Scholar grant RSG-14-049-01-DMC.

5.7 Figures

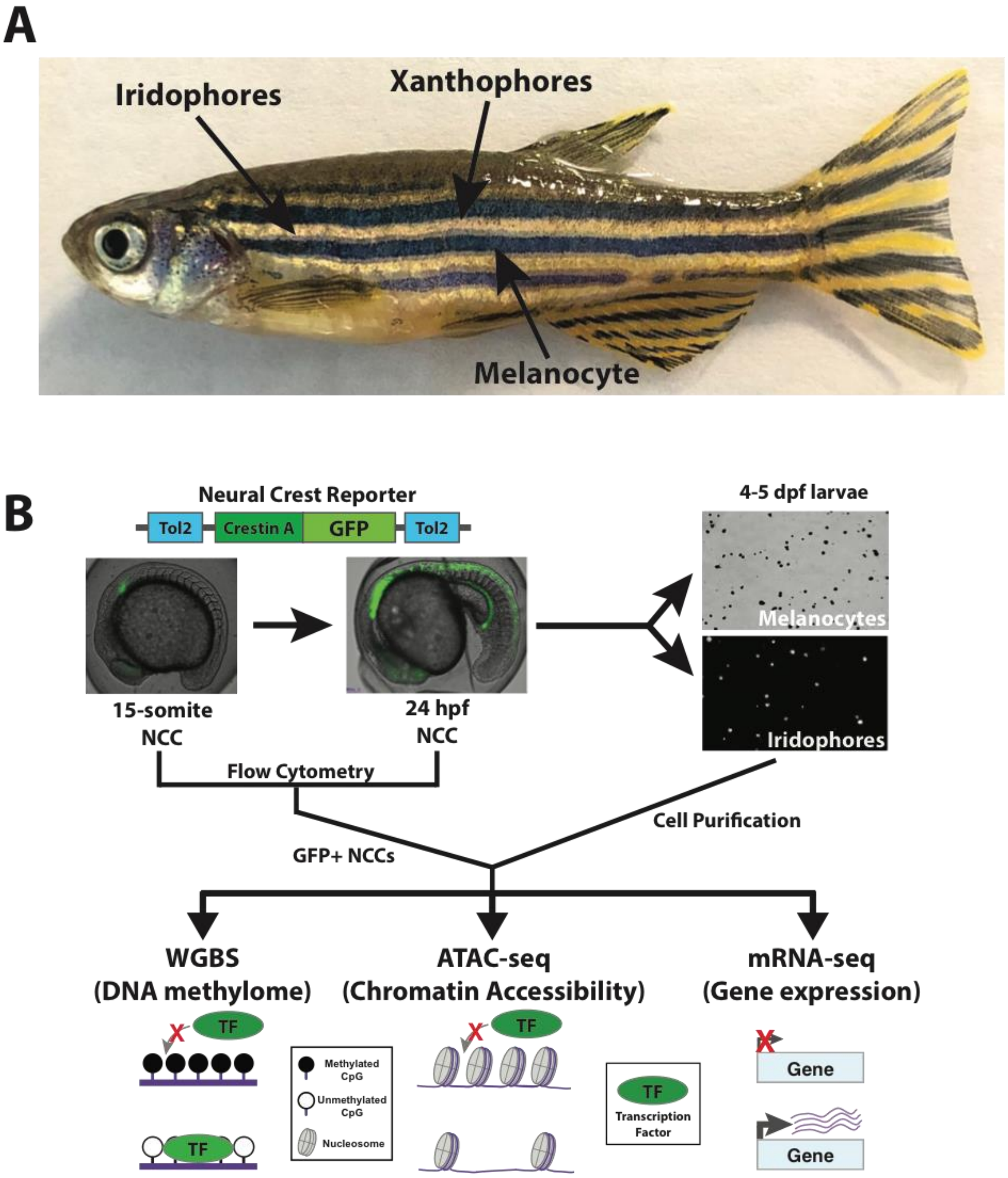


Figure 1. Zebrafish neural crest cell differentiation into pigment cells. A) Zebrafish pigment cell types. B) Visual schematic of experimental design for collecting NCC and pigment cells. TF, transcription factor.

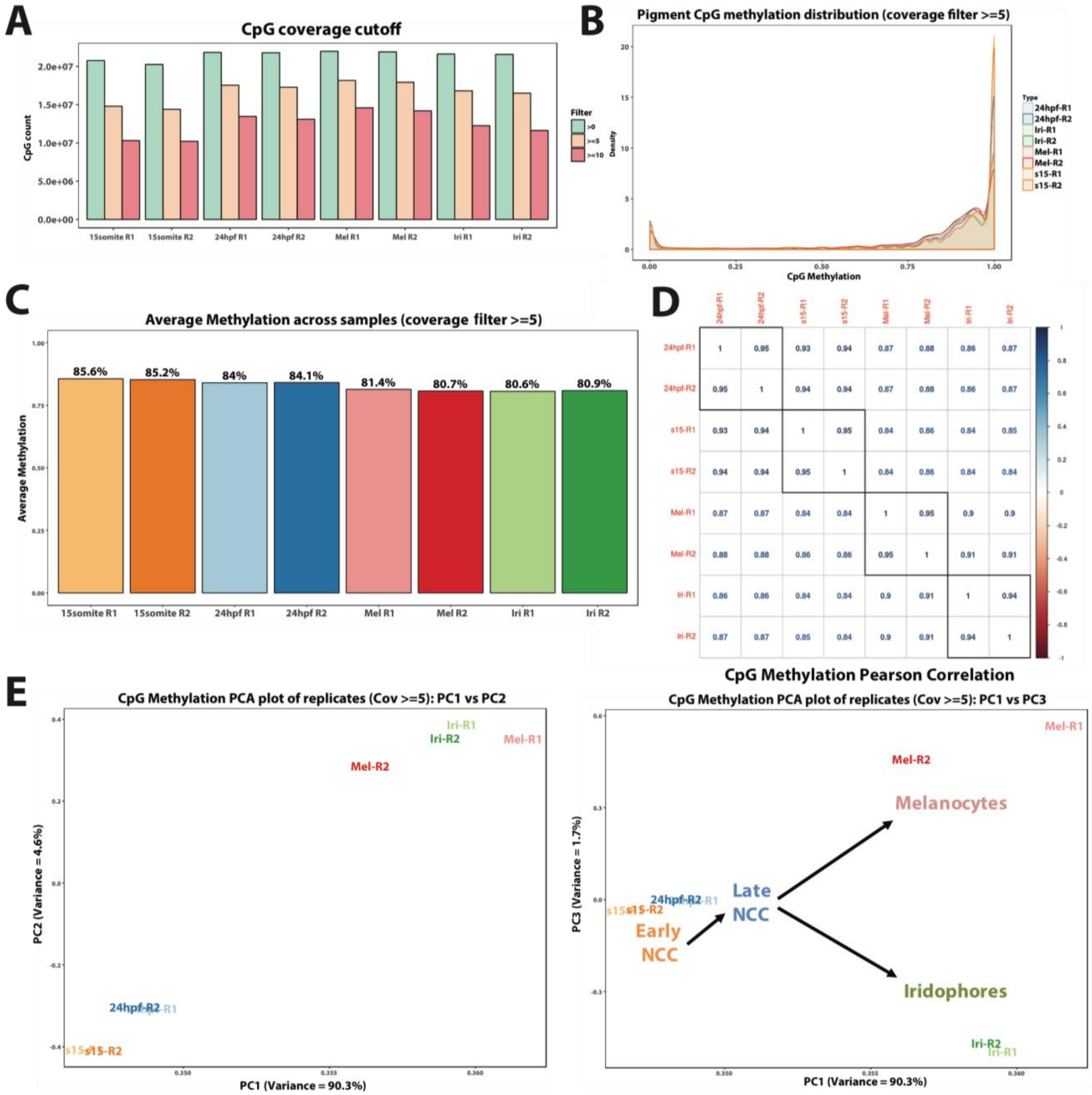


Figure 2. DNA methylation dynamics across zebrafish pigment development. A) Number of CpGs that pass coverage cutoff. B) Global DNA methylation distribution of NCC and pigment cells. C) Average global methylation levels. D) DNA methylation Pearson's correlation coefficients across samples. E) Principal component analysis (PCA) of DNA methylation.

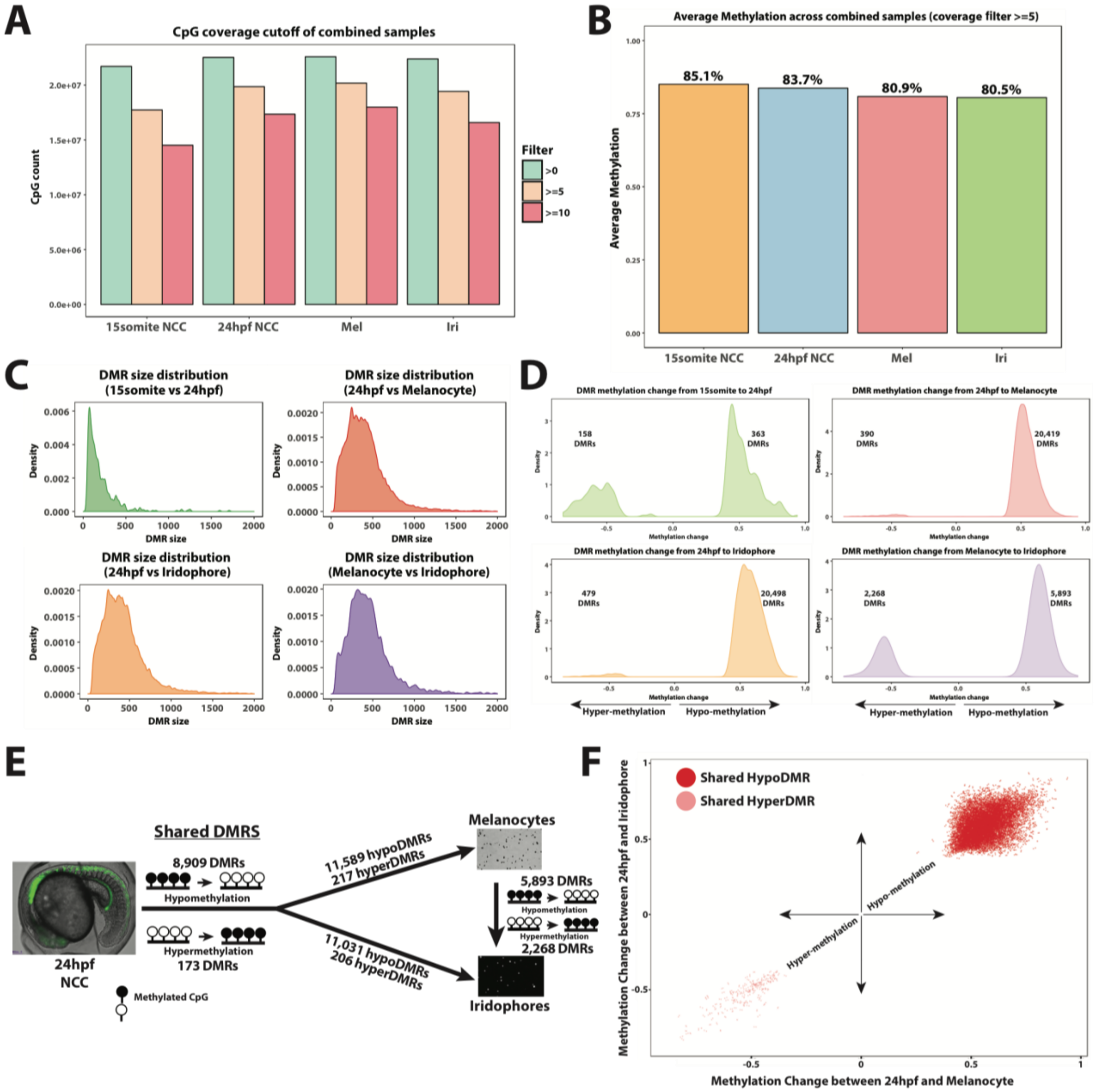


Figure 3. Differentially methylated region (DMR) analysis on combined biological replicates. A) Number of CpGs that pass coverage cutoff when biological replicates are combined. B) Average global methylation levels. C) DMR size distribution. D) Distribution of methylation changes in DMRs between 24 hpf NCC and differentiated pigment cells. E) Number of DMRs detected among 24 hpf NCC and pigment cells. F) DNA methylation differences between 24 hpf NCC and pigment cells in shared DMRs.

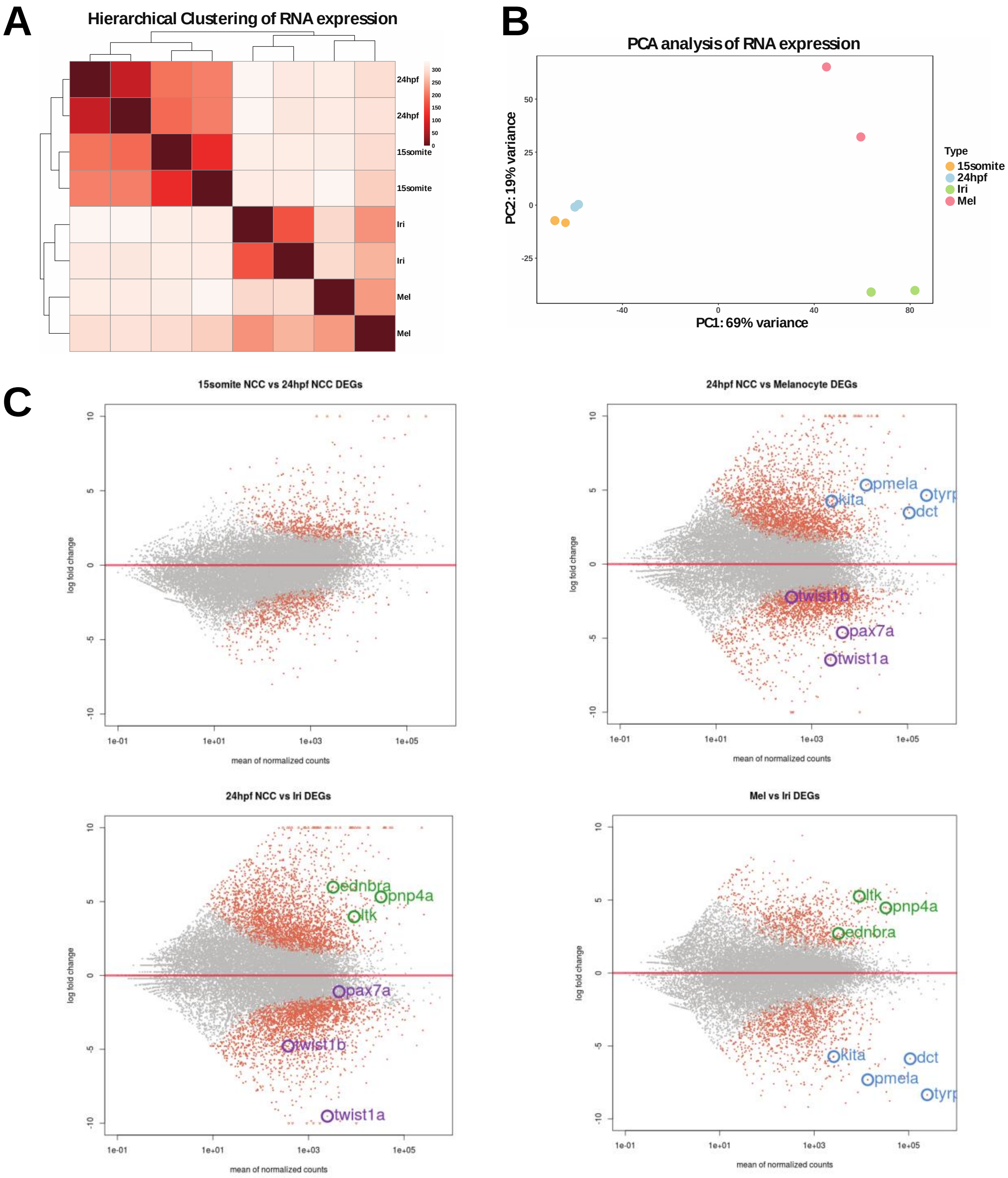


Figure 4. Validation of mRNA-seq library quality. A) Hierarchical clustering of samples based on gene expression. B) PCA analysis based on gene expression differences. C) MA plot of differentially expressed genes (DEGs) with cell type-specific marker genes highlighted.

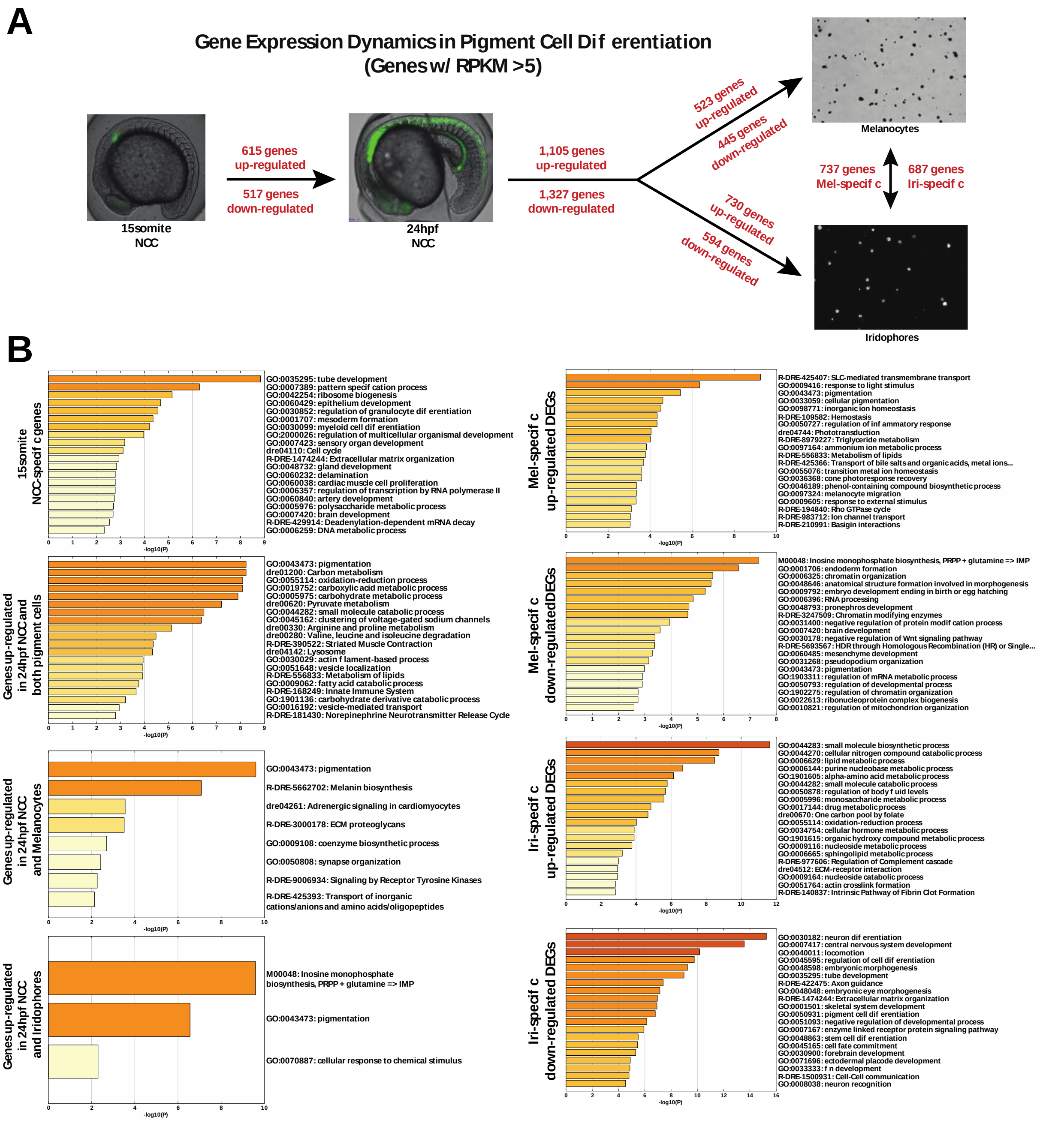


Figure 5. Gene expression dynamics across pigment cell differentiation. A) Number of DEGs in pair-wise comparisons among developmental time points. B) Gene ontology enrichment for various clusters of DEGs.

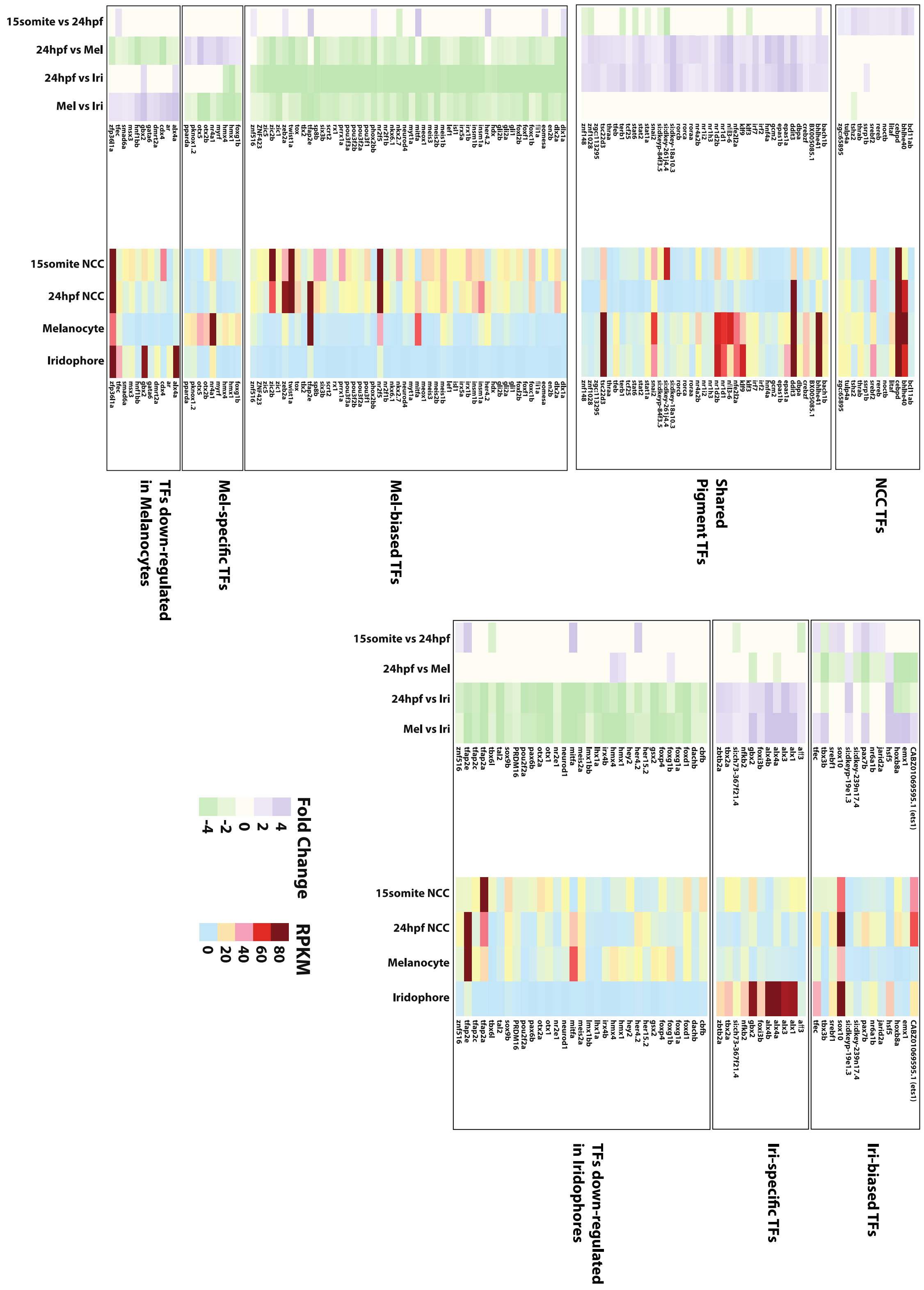


Figure 6. Differentially expressed transcription factors across pigment differentiation.

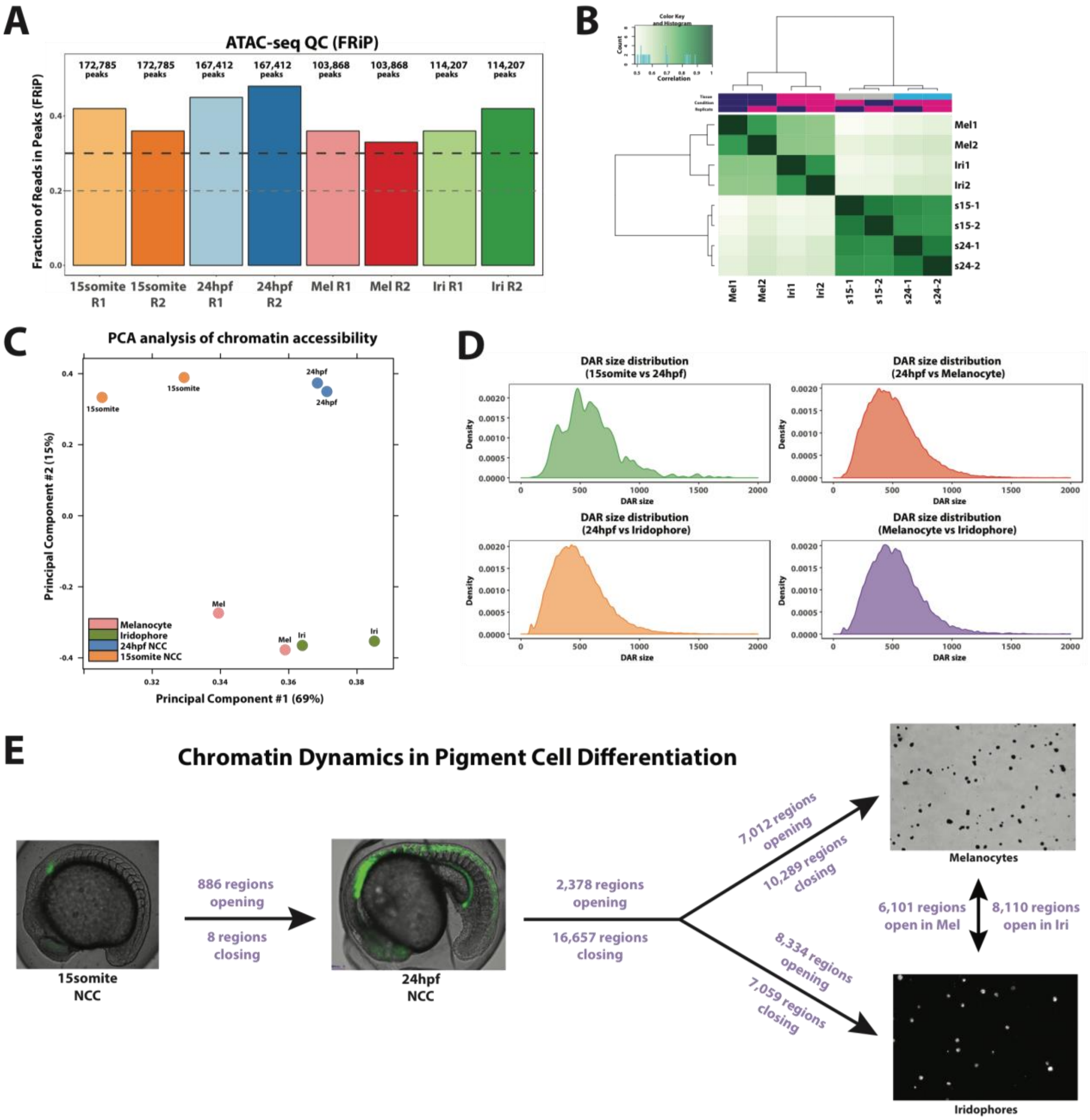


Figure 7. Chromatin accessibility dynamics captured by ATAC-seq. A) Number of peaks identified per sample with fraction of reads in peak (FRiP) value for quality check. B) Clustering of samples based on differential chromatin accessibility. C) PCA analysis of chromatin accessibility. D) Size distribution of differentially accessible regions (DARs). E) Number of DARs identified across pigment cell differentiation.

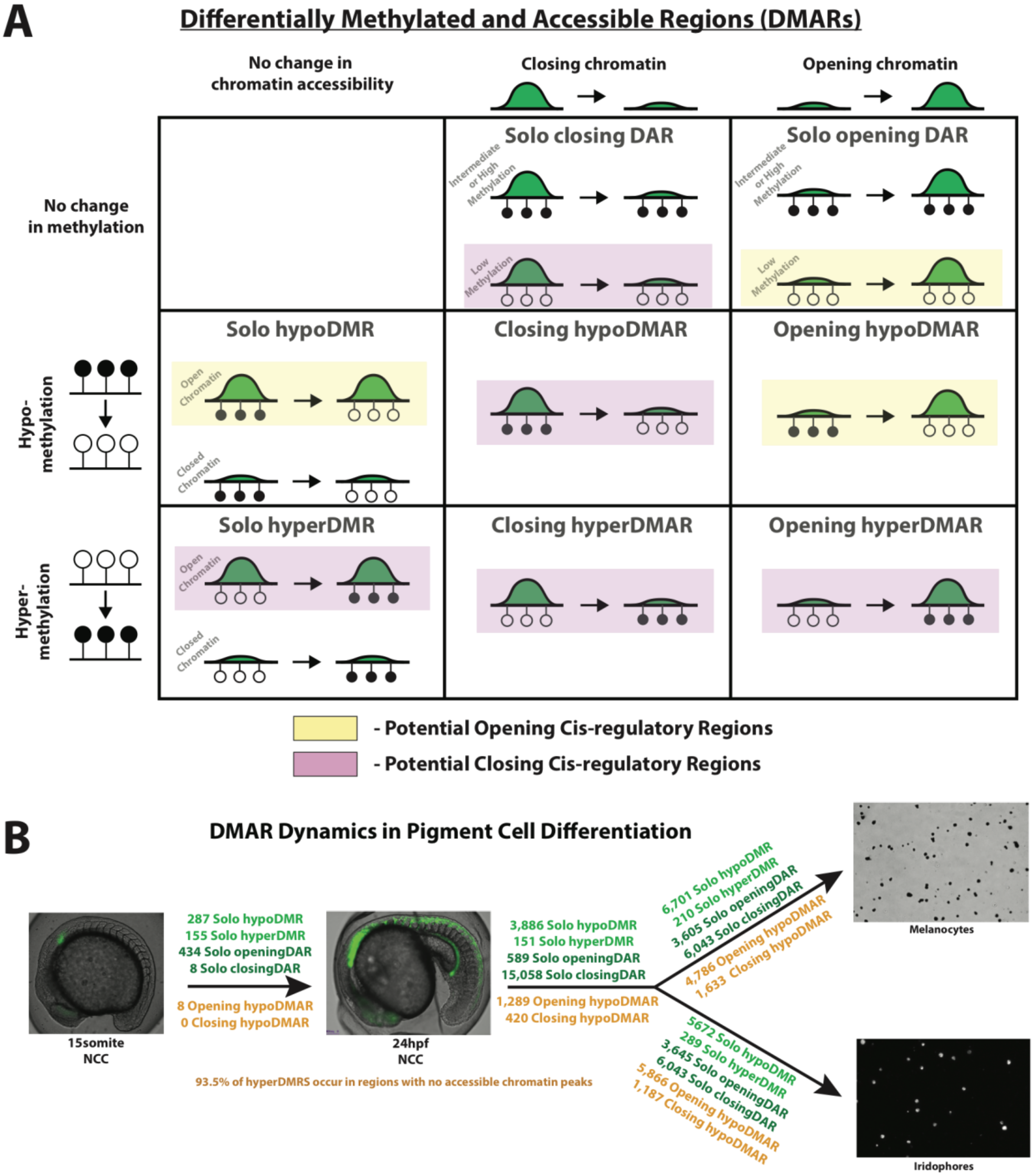


Figure 8. Defining differentially methylation and/or accessible regions. A) Schematic of possible types of DMARs. B) Number of DMARs across pigment cell differentiation.

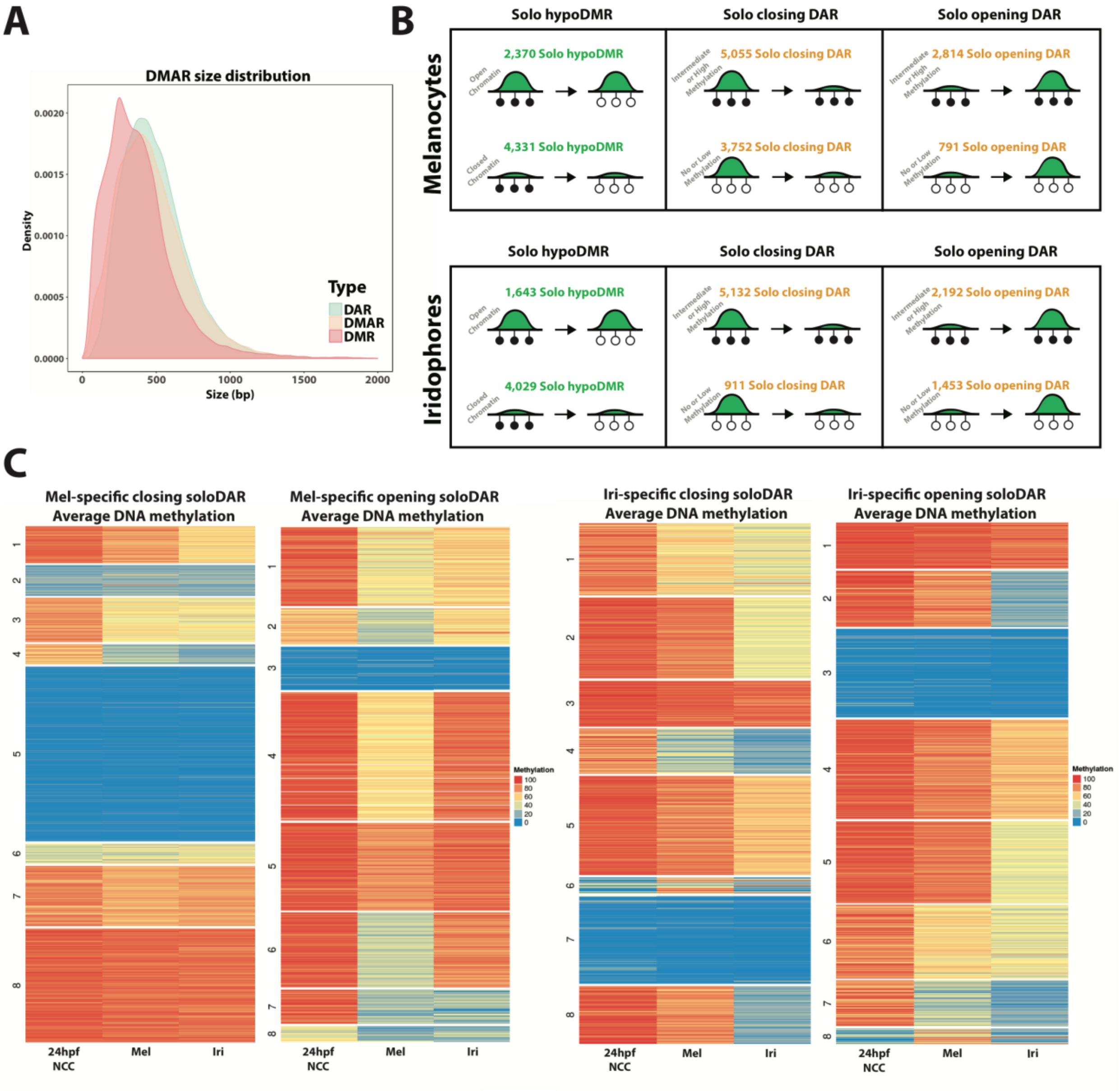


Figure 9. Characterization of methylation levels and chromatin accessibility in solo DM/ARs. A) Size distribution of DMRs, DARs and DMARs. B) Visualization and frequency of methylation or accessibility status in solo DMRs and DARs. C) Heatmap of methylation levels in soloDARs identified between 24hpf NCC and pigment cells.

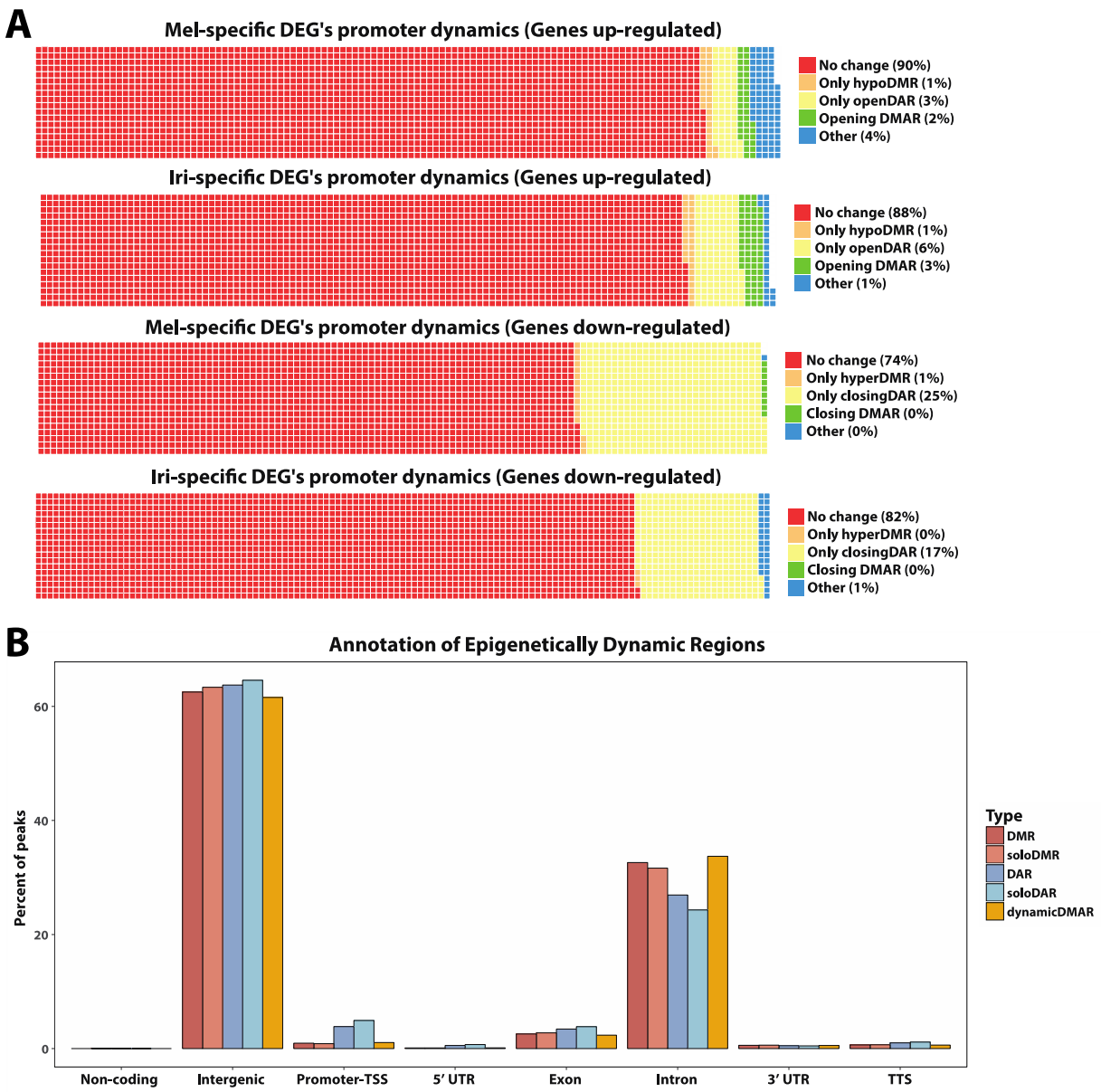


Figure 10. Potential role of cis-regulatory elements driving gene expression differences. A) Epigenetic dynamics of DEG promoters. B) Genome annotation of epigenetically dynamic regions.

A

Mel-specif c hypoDMR HOMER results

Rank	Motif	Name	P-value	log P-pvalue	q-value (Benjamini)	# Target Sequences with Motif	% of Targets Sequences with Motif	# Background Sequences with Motif	% of Background Sequences with Motif	Group
1	GTCATGTGAC	MITF(bHLH)/MastCells-MITF-ChIP-Seq(GSE48085)/Homer	1e-302	-6.971e+02	0.0000	2350.0	49.10%	10401.7	24.19%	MiT
2	GTCACGTGGT	Usf2(bHLH)/C2C12-Usf2-ChIP-Seq(GSE36030)/Homer	1e-170	-3.921e+02	0.0000	1128.0	23.57%	4189.5	9.74%	BHLHE41
3	GGTCACGTGA	USF1(bHLH)/GM12878-Usf1-ChIP-Seq(GSE32465)/Homer	1e-143	-3.308e+02	0.0000	1176.0	24.57%	4868.3	11.32%	BHLHE41
4	GCCTCAGGGCAT	AP-2gamma(AP2)/MCF7-TFAP2C-ChIP-Seq(GSE21234)/Homer	1e-116	-2.686e+02	0.0000	901.0	18.83%	3564.2	8.29%	TFAP2
5	GTCACGTG	BMAL1(bHLH)/Liver-Bmal1-ChIP-Seq(GSE39860)/Homer	1e-112	-2.589e+02	0.0000	3231.0	67.51%	22092.7	51.38%	BHLHE41
6	AGGGGCCCTTTGT	Sox9(HMG)/Limb-SOX9-ChIP-Seq(GSE73225)/Homer	1e-105	-2.424e+02	0.0000	1713.0	35.79%	9426.4	21.92%	SOX
7	ATGCCCTGAGGC	AP-2alpha(AP2)/Hela-AP2alpha-ChIP-Seq(GSE31477)/Homer	1e-97	-2.241e+02	0.0000	677.0	14.15%	2505.9	5.83%	TFAP2
8	GTCACGTG	CLOCK(bHLH)/Liver-Clock-ChIP-Seq(GSE39860)/Homer	1e-71	-1.638e+02	0.0000	1149.0	24.01%	6124.5	14.24%	BHLHE41
9	GTCACGTGACCC	TFE3(bHLH)/MEF-TFE3-ChIP-Seq(GSE75757)/Homer	1e-69	-1.599e+02	0.0000	437.0	9.13%	1512.0	3.52%	BHLHE41
10	CCTTTGTTCG	Sox10(HMG)/SciaticNerve-Sox3-ChIP-Seq(GSE35132)/Homer	1e-68	-1.577e+02	0.0000	2798.0	58.46%	19689.5	45.80%	SOX
11	GGATGACTCATC	Fra2(bZIP)/Striatum-Fra2-ChIP-Seq(GSE43429)/Homer	1e-56	-1.304e+02	0.0000	1109.0	23.17%	6231.2	14.49%	FRA
12	GGATGAGTCATC	Fra1(bZIP)/BT549-Fra1-ChIP-Seq(GSE46166)/Homer	1e-54	-1.265e+02	0.0000	1281.0	26.77%	7576.6	17.62%	FRA
13	GATGACTCATCC	Atf3(bZIP)/GBM-ATF3-ChIP-Seq(GSE33912)/Homer	1e-54	-1.263e+02	0.0000	1441.0	30.11%	8817.0	20.51%	JDP2
14	GCCACGTGAC	NPAS2(bHLH)/Liver-NPAS2-ChIP-Seq(GSE39860)/Homer	1e-54	-1.251e+02	0.0000	1825.0	38.13%	11914.8	27.71%	BHLHE41
15	ATGACTCATC	AP-1(bZIP)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer	1e-52	-1.200e+02	0.0000	1552.0	32.43%	9793.9	22.78%	KLF13
16	CCTTTGTTCC	Sox4(HMG)/proB-Sox4-ChIP-Seq(GSE50066)/Homer	1e-50	-1.169e+02	0.0000	1629.0	34.04%	10461.8	24.33%	SOX
17	TATGAGTCAT	BATF(bZIP)/Th17-BATF-ChIP-Seq(GSE39756)/Homer	1e-47	-1.093e+02	0.0000	1361.0	28.44%	8457.3	19.67%	JDP2
18	GATGACTCATCC	Jun-AP1(bZIP)/K562-cJun-ChIP-Seq(GSE31477)/Homer	1e-46	-1.069e+02	0.0000	498.0	10.41%	2233.0	5.19%	JUN
19	GATGAGTCATCC	Fosl2(bZIP)/3T3L1-Fosl2-ChIP-Seq(GSE56872)/Homer	1e-46	-1.063e+02	0.0000	682.0	14.25%	3462.1	8.05%	FRA
20	GATGACTCAT	JunB(bZIP)/DendriticCells-Junb-ChIP-Seq(GSE36099)/Homer	1e-45	-1.038e+02	0.0000	1169.0	24.43%	7054.2	16.41%	JUN

Iri-specif c hypoDMR HOMER results

Rank	Motif	Name	P-value	log P-pvalue	q-value (Benjamini)	# Target Sequences with Motif	% of Targets Sequences with Motif	# Background Sequences with Motif	% of Background Sequences with Motif	Group
1	CTAATTAAATTA	Phox2a(Homeobox)/Neuron-Phox2a-ChIP-Seq(GSE31456)/Homer	1e-351	-8.104e+02	0.0000	2738.0	46.68%	9598.8	22.78%	EN2
2	GTCATGTGAC	MITF(bHLH)/MastCells-MITF-ChIP-Seq(GSE48085)/Homer	1e-324	-7.471e+02	0.0000	2808.0	47.87%	10355.8	24.57%	MiT
3	CTAATTGAATTA	Prop1(Homeobox)/GHFT1-PROP1.biotin-ChIP-Seq(GSE77302)/Homer	1e-253	-5.832e+02	0.0000	3226.0	54.99%	14038.9	33.31%	EN2
4	GTCACGTGGT	Usf2(bHLH)/C2C12-Usf2-ChIP-Seq(GSE36030)/Homer	1e-196	-4.535e+02	0.0000	1363.0	23.24%	4134.9	9.81%	BHLHE41
5	GGTCACGTGA	USF1(bHLH)/GM12878-Usf1-ChIP-Seq(GSE32465)/Homer	1e-174	-4.011e+02	0.0000	1440.0	24.55%	4784.1	11.35%	BHLHE41
6	GTCACGTG	BMAL1(bHLH)/Liver-Bmal1-ChIP-Seq(GSE39860)/Homer	1e-142	-3.284e+02	0.0000	3947.0	67.29%	21426.3	50.84%	BHLHE41
7	GTCACGTG	CLOCK(bHLH)/Liver-Clock-ChIP-Seq(GSE39860)/Homer	1e-102	-2.368e+02	0.0000	1468.0	25.03%	6030.3	14.31%	BHLHE41
8	GTCACGTGACCC	TFE3(bHLH)/MEF-TFE3-ChIP-Seq(GSE75757)/Homer	1e-80	-1.862e+02	0.0000	530.0	9.04%	1496.4	3.55%	BHLHE41
9	AGGGGCCCTTTGT	Sox9(HMG)/Limb-SOX9-ChIP-Seq(GSE73225)/Homer	1e-79	-1.822e+02	0.0000	1910.0	32.56%	9211.5	21.86%	SOX
10	TAATCCGATTAC	Pax7(Paired,Homeobox),long/Myoblast-Pax7-ChIP-Seq(GSE25064)/Homer	1e-74	-1.711e+02	0.0000	355.0	6.05%	818.0	1.94%	EN2
11	GCCACGTGAC	NPAS2(bHLH)/Liver-NPAS2-ChIP-Seq(GSE39860)/Homer	1e-69	-1.602e+02	0.0000	2239.0	38.17%	11590.2	27.50%	BHLHE41
12	TACACGTG	NPAS(bHLH)/Liver-NPAS-ChIP-Seq(GSE39860)/Homer	1e-59	-1.360e+02	0.0000	3427.0	58.42%	20149.1	47.81%	BHLHE41
13	GATGACTCATCC	Atf3(bZIP)/GBM-ATF3-ChIP-Seq(GSE33912)/Homer	1e-50	-1.152e+02	0.0000	1687.0	28.76%	8653.3	20.53%	JDP2
14	AACTAGGTCA	RORgt(NR)/EL4-RORgt.Flag-ChIP-Seq(GSE56019)/Homer	1e-48	-1.127e+02	0.0000	445.0	7.59%	1477.5	3.51%	NR5A1
15	GATGACTCAT	JunB(bZIP)/DendriticCells-Junb-ChIP-Seq(GSE36099)/Homer	1e-47	-1.086e+02	0.0000	1382.0	23.56%	6830.8	16.21%	JUN
16	TAATTAGC	Lhx2(Homeobox)/HFSC-Lhx2-ChIP-Seq(GSE48068)/Homer	1e-46	-1.062e+02	0.0000	3213.0	54.77%	19144.1	45.43%	EN2
17	AGCTAATTAG	Lhx1(Homeobox)/EmbryoCarcinoma-Lhx1-ChIP-Seq(GSE70957)/Homer	1e-45	-1.047e+02	0.0000	3495.0	59.58%	21206.1	50.32%	EN2
18	TATGAGTCAT	BATF(bZIP)/Th17-BATF-ChIP-Seq(GSE39756)/Homer	1e-45	-1.037e+02	0.0000	1584.0	27.00%	8165.1	19.37%	JDP2
19	ATGACTCATC	AP-1(bZIP)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer	1e-44	-1.021e+02	0.0000	1783.0	30.40%	9464.6	22.46%	KLF13
20	GGATGAGTCATC	Fra1(bZIP)/BT549-Fra1-ChIP-Seq(GSE46166)/Homer	1e-40	-9.305e+01	0.0000	1435.0	24.46%	7379.2	17.51%	

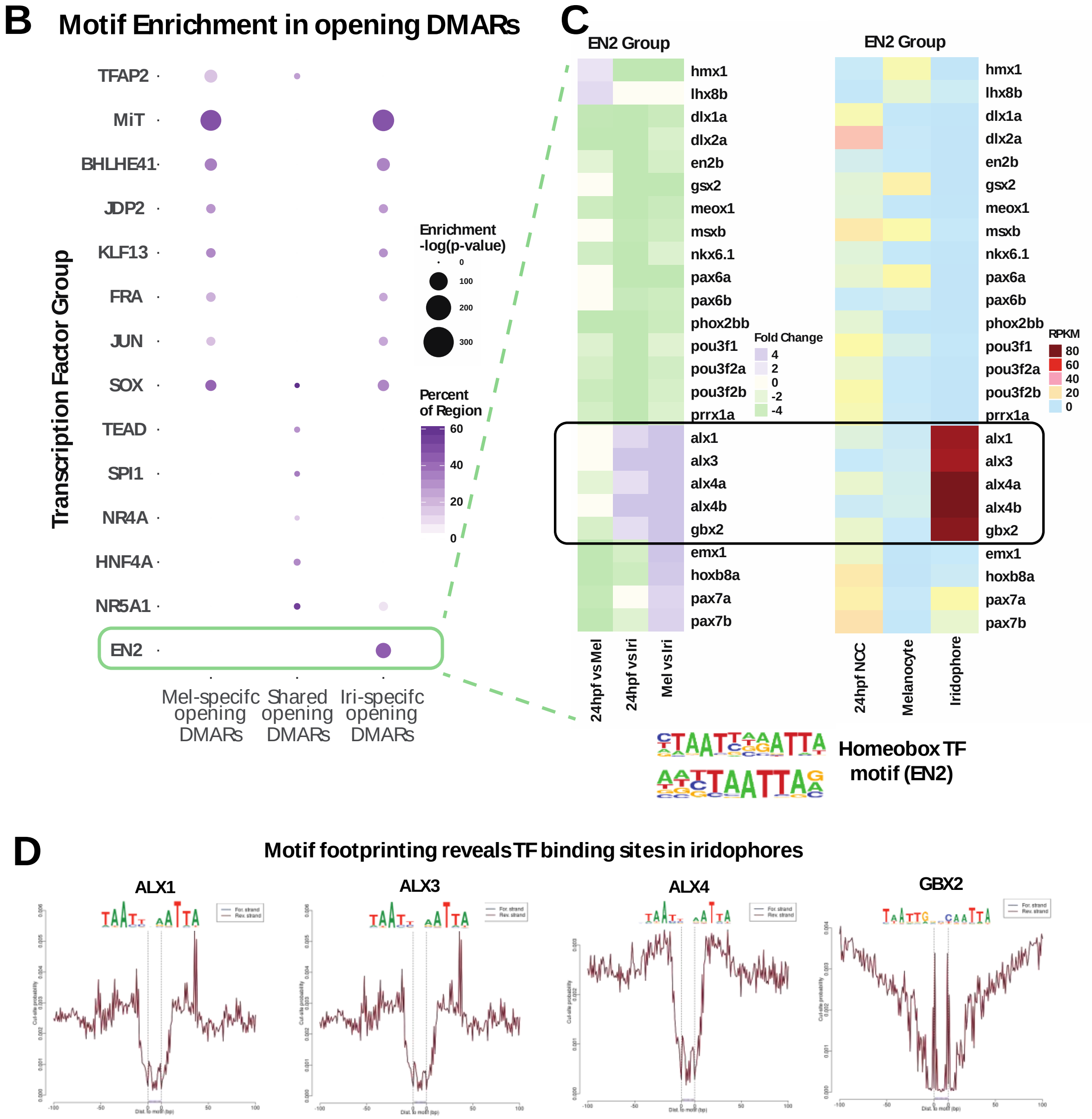
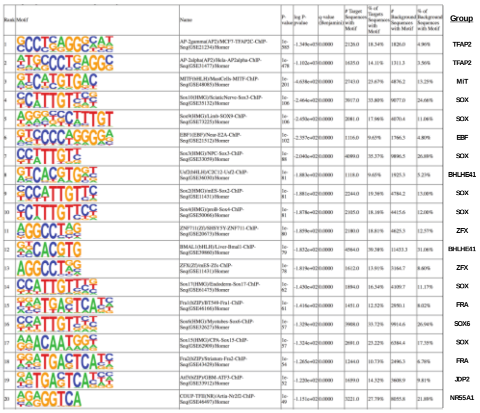


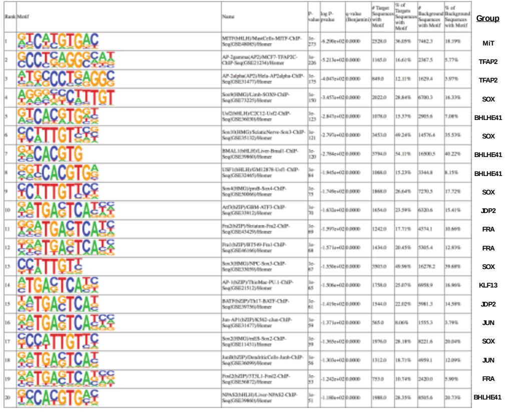
Figure 11. Discovery of transcription factors that shape epigenetic landscape during pigment cell differentiation. A) Partitioning HOMER motif enrichment results in dynamic DMARs to TF clusters. B) Enrichment of motif clusters in dynamic DMARs. C) Heatmap representation of differentially expressed transcription factors (TFs) present in EN2 group. D) Motif footprinting signatures of *alx* TFs and *gbx2* in iridophore ATAC-seq peaks.

Mel-specific hypoDMR HOMER results



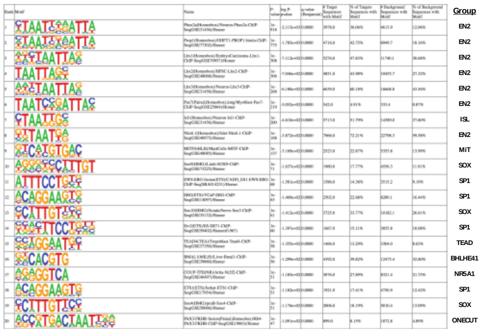
Rank	Motif	Name	P-value	log P-pvalue	q-value (Benjamini)	# Target Sequences with Motif	% of Targets Sequences with Motif	# Background Sequences with Motif	% of Background Sequences with Motif	Group
1		AP-2gamma(AP2)/MCF7-TFAP2C-ChIP-Seq(GSE21234)/Homer	1e-585	-1.349e+03	0.0000	2126.0	18.34%	1826.0	4.96%	TFAP2
2		AP-2alpha(AP2)/Hela-AP2alpha-ChIP-Seq(GSE31477)/Homer	1e-478	-1.102e+03	0.0000	1635.0	14.11%	1311.3	3.56%	TFAP2
3		MITF(bHLH)/MastCells-MITF-ChIP-Seq(GSE48085)/Homer	1e-201	-4.638e+02	0.0000	2743.0	23.67%	4876.2	13.25%	MiT
4		Sox10(HMG)/SciaticNerve-Sox3-ChIP-Seq(GSE35132)/Homer	1e-106	-2.464e+02	0.0000	3917.0	33.80%	9077.0	24.66%	SOX
5		Sox9(HMG)/Limb-SOX9-ChIP-Seq(GSE73225)/Homer	1e-106	-2.450e+02	0.0000	2081.0	17.96%	4070.4	11.06%	SOX
6		EBF1(EBF)/Near-E2A-ChIP-Seq(GSE21512)/Homer	1e-102	-2.357e+02	0.0000	1116.0	9.63%	1766.3	4.80%	EBF
7		Sox3(HMG)/NPC-Sox3-ChIP-Seq(GSE33059)/Homer	1e-88	-2.040e+02	0.0000	4099.0	35.37%	9896.5	26.89%	SOX
8		Usf2(bHLH)/C2C12-Usf2-ChIP-Seq(GSE36030)/Homer	1e-81	-1.883e+02	0.0000	1118.0	9.65%	1925.3	5.23%	BHLHE41
9		Sox2(HMG)/mES-Sox2-ChIP-Seq(GSE11431)/Homer	1e-81	-1.881e+02	0.0000	2244.0	19.36%	4784.2	13.00%	SOX
10		Sox4(HMG)/proB-Sox4-ChIP-Seq(GSE50066)/Homer	1e-81	-1.878e+02	0.0000	2105.0	18.16%	4415.6	12.00%	SOX
11		ZNF711(Zf)/SHSY5Y-ZNF711-ChIP-Seq(GSE20673)/Homer	1e-80	-1.859e+02	0.0000	2180.0	18.81%	4625.3	12.57%	ZFX
12		BMAL1(bHLH)/Liver-Bmal1-ChIP-Seq(GSE39860)/Homer	1e-79	-1.832e+02	0.0000	4564.0	39.38%	11433.3	31.06%	BHLHE41
13		ZFX(Zf)/mES-Zfx-ChIP-Seq(GSE11431)/Homer	1e-78	-1.819e+02	0.0000	1612.0	13.91%	3164.7	8.60%	ZFX
14		Sox17(HMG)/Endoderm-Sox17-ChIP-Seq(GSE61475)/Homer	1e-62	-1.430e+02	0.0000	1894.0	16.34%	4109.7	11.17%	SOX
15		Fra1(bZIP)/BT549-Fra1-ChIP-Seq(GSE46166)/Homer	1e-61	-1.416e+02	0.0000	1451.0	12.52%	2950.1	8.02%	FRA
16		Sox6(HMG)/Myotubes-Sox6-ChIP-Seq(GSE32627)/Homer	1e-57	-1.329e+02	0.0000	3908.0	33.72%	9914.6	26.94%	SOX6
17		Sox15(HMG)/CPA-Sox15-ChIP-Seq(GSE62909)/Homer	1e-57	-1.324e+02	0.0000	2691.0	23.22%	6384.4	17.35%	SOX
18		Fra2(bZIP)/Striatum-Fra2-ChIP-Seq(GSE43429)/Homer	1e-54	-1.265e+02	0.0000	1244.0	10.73%	2496.3	6.78%	FRA
19		Atf3(bZIP)/GBM-ATF3-ChIP-Seq(GSE33912)/Homer	1e-52	-1.220e+02	0.0000	1659.0	14.32%	3608.9	9.81%	JDP2
20		COUP-TFII(NR)/Artia-Nr2f2-ChIP-Seq(GSE46497)/Homer	1e-49	-1.151e+02	0.0000	3221.0	27.79%	8055.8	21.89%	NR55A1

Mel-specific opening DAR HOMER results



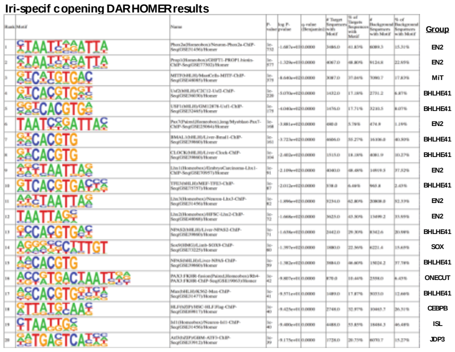
Rank	Motif	Name	P-value	log P-pvalue	q-value (Benjamini)	# Target Sequences with Motif	% of Targets Sequences with Motif	# Background Sequences with Motif	% of Background Sequences with Motif	Group
1		MITF(bHLH)/MastCells-MITF-ChIP-Seq(GSE48085)/Homer	1e-273	-6.290e+02	0.0000	2528.0	36.09%	7462.3	18.19%	MiT
2		AP-2gamma(AP2)/MCF7-TFAP2C-ChIP-Seq(GSE21234)/Homer	1e-226	-5.213e+02	0.0000	1165.0	16.61%	2367.5	5.77%	TFAP2
3		AP-2alpha(AP2)/Hela-AP2alpha-ChIP-Seq(GSE31477)/Homer	1e-175	-4.047e+02	0.0000	849.0	12.11%	1629.4	3.97%	TFAP2
4		Sox9(HMG)/Limb-SOX9-ChIP-Seq(GSE73225)/Homer	1e-150	-3.457e+02	0.0000	2022.0	28.84%	6700.3	16.33%	SOX
5		Usf2(bHLH)/C2C12-Usf2-ChIP-Seq(GSE36030)/Homer	1e-123	-2.847e+02	0.0000	1078.0	15.37%	2905.6	7.08%	BHLHE41
6		Sox10(HMG)/SciaticNerve-Sox3-ChIP-Seq(GSE35132)/Homer	1e-121	-2.797e+02	0.0000	3453.0	49.24%	14576.4	35.53%	SOX
7		BMAL1(bHLH)/Liver-Bmal1-ChIP-Seq(GSE39860)/Homer	1e-120	-2.784e+02	0.0000	3794.0	54.11%	16500.5	40.22%	BHLHE41
8		USF1(bHLH)/GM12878-Usf1-ChIP-Seq(GSE32465)/Homer	1e-84	-1.945e+02	0.0000	1068.0	15.23%	3344.8	8.15%	BHLHE41
9		Sox4(HMG)/proB-Sox4-ChIP-Seq(GSE50066)/Homer	1e-75	-1.749e+02	0.0000	1868.0	26.64%	7270.5	17.72%	SOX
10		Atf3(bZIP)/GBM-ATF3-ChIP-Seq(GSE33912)/Homer	1e-70	-1.632e+02	0.0000	1654.0	23.59%	6320.6	15.41%	JDP2
11		Fra2(bZIP)/Striatum-Fra2-ChIP-Seq(GSE43429)/Homer	1e-69	-1.597e+02	0.0000	1242.0	17.71%	4374.1	10.66%	FRA
12		Fra1(bZIP)/BT549-Fra1-ChIP-Seq(GSE46166)/Homer	1e-68	-1.571e+02	0.0000	1434.0	20.45%	5305.4	12.93%	FRA
13		Sox3(HMG)/NPC-Sox3-ChIP-Seq(GSE33059)/Homer	1e-67	-1.550e+02	0.0000	3503.0	49.96%	16278.2	39.68%	SOX
14		AP-1(bZIP)/ThioMac-PU.1-ChIP-Seq(GSE21512)/Homer	1e-65	-1.506e+02	0.0000	1758.0	25.07%	6958.9	16.96%	KLF13
15		BATF(bZIP)/Th17-BATF-ChIP-Seq(GSE39756)/Homer	1e-61	-1.419e+02	0.0000	1544.0	22.02%	5981.3	14.58%	JDP2
16		Jun-AP1(bZIP)/K562-cJun-ChIP-Seq(GSE31477)/Homer	1e-59	-1.371e+02	0.0000	565.0	8.06%	1555.3	3.79%	JUN
17		Sox2(HMG)/mES-Sox2-ChIP-Seq(GSE11431)/Homer	1e-59	-1.365e+02	0.0000	1976.0	28.18%	8221.6	20.04%	SOX
18		JunB(bZIP)/DendriticCells-Junb-ChIP-Seq(GSE36099)/Homer	1e-56	-1.303e+02	0.0000	1312.0	18.71%	4959.1	12.09%	JUN
19		Fosl2(bZIP)/3T3L1-Fosl2-ChIP-Seq(GSE56872)/Homer	1e-53	-1.242e+02	0.0000	753.0	10.74%	2420.0	5.90%	FRA
20		NPAS2(bHLH)/Liver-NPAS2-ChIP-Seq(GSE39860)/Homer	1e-51	-1.180e+02	0.0000	1988.0	28.35%	8505.6	20.73%	BHLHE41

Iri-specific hypoDMR HOMER results

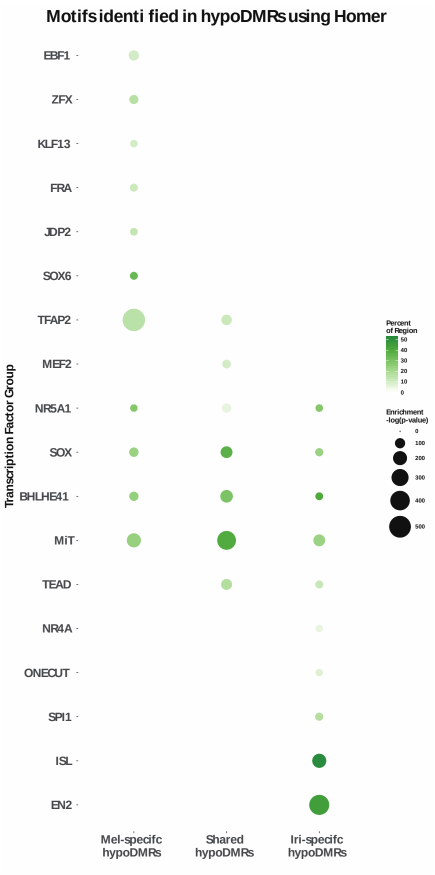


Rank	Motif	Name	P-value	log P-pvalue	q-value (Benjamini)	# Target Sequences with Motif	% of Targets Sequences with Motif	# Background Sequences with Motif	% of Background Sequences with Motif	Group
1		Phox2a(Homeobox)/Neuron-Phox2a-ChIP-Seq(GSE31456)/Homer	1e-918	-2.115e+03	0.0000	3970.0	36.06%	4613.9	12.06%	EN2
2		Prop1(Homeobox)/GHFT1-PROP1.biotin-ChIP-Seq(GSE77302)/Homer	1e-775	-1.785e+03	0.0000	4716.0	42.75%	6949.7	18.16%	EN2
3		Lhx1(Homeobox)/EmbryoCarcinoma-Lhx1-ChIP-Seq(GSE70957)/Homer	1e-308	-7.112e+02	0.0000	5276.0	47.83%	11740.1	30.68%	EN2
4		Lhx2(Homeobox)/HFSC-Lhx2-ChIP-Seq(GSE48068)/Homer	1e-306	-7.046e+02	0.0000	4851.0	43.98%	10455.7	27.32%	EN2
5		Lhx3(Homeobox)/Neuron-Lhx3-ChIP-Seq(GSE31456)/Homer	1e-268	-6.180e+02	0.0000	6639.0	60.18%	16668.8	43.56%	EN2
6		Pax7(Paired,Homeobox),long/Myoblast-Pax7-ChIP-Seq(GSE25064)/Homer	1e-219	-5.052e+02	0.0000	542.0	4.91%	333.4	0.87%	EN2
7		Isl1(Homeobox)/Neuron-Isl1-ChIP-Seq(GSE31456)/Homer	1e-200	-4.608e+02	0.0000	5713.0	51.79%	14389.0	37.60%	ISL
8		Nkx6.1(Homeobox)/Islet-Nkx6.1-ChIP-Seq(GSE40975)/Homer	1e-168	-3.872e+02	0.0000	7966.0	72.21%	22798.3	59.58%	EN2
9		MITF(bHLH)/MastCells-MITF-ChIP-Seq(GSE48085)/Homer	1e-137	-3.169e+02	0.0000	2523.0	22.87%	5353.6	13.99%	MiT
10		Sox9(HMG)/Limb-SOX9-ChIP-Seq(GSE73225)/Homer	1e-71	-1.637e+02	0.0000	1960.0	17.77%	4556.3	11.91%	SOX
11		EWS:ERG-fusion(ETS)/CADO_ES1-EWS:ERG-ChIP-Seq(SRA014231)/Homer	1e-68	-1.581e+02	0.0000	1586.0	14.38%	3515.2	9.19%	SP1
12		ERG(ETS)/VCaP-ERG-ChIP-Seq(GSE14097)/Homer	1e-63	-1.469e+02	0.0000	2502.0	22.68%	6289.1	16.44%	SP1
13		Sox10(HMG)/SciaticNerve-Sox3-ChIP-Seq(GSE35132)/Homer	1e-61	-1.412e+02	0.0000	3725.0	33.77%	10182.1	26.61%	SOX
14		Etv2(ETS)/ES-ER71-ChIP-Seq(GSE59402)/Homer(0.967)	1e-60	-1.387e+02	0.0000	1667.0	15.11%	3855.8	10.08%	SP1
15		TEAD4(TEA)/Tropoblast-Tead4-ChIP-Seq(GSE37350)/Homer	1e-58	-1.355e+02	0.0000	1466.0	13.29%	3304.0	8.63%	TEAD
16		BMAL1(bHLH)/Liver-Bmal1-ChIP-Seq(GSE39860)/Homer	1e-56	-1.299e+02	0.0000	4392.0	39.82%	12475.4	32.60%	BHLHE41
17		COUP-TFII(NR)/Artia-Nr2f2-ChIP-Seq(GSE46497)/Homer	1e-51	-1.183e+02	0.0000	3016.0	27.30%	8321.6	21.75%	NR5A1
18		ETS1(ETS)/Jurkat-ETS1-ChIP-Seq(GSE17954)/Homer	1e-51	-1.182e+02	0.0000	1921.0	17.41%	4790.9	12.42%	SP1
19		Sox4(HMG)/proB-Sox4-ChIP-Seq(GSE50066)/Homer	1e-51	-1.176e+02	0.0000	2006.0	18.19%	5010.4	13.09%	SOX
20		PAX3:FKHR-fusion(Paired,Homeobox)/Rh4-PAX3:FKHR-ChIP-Seq(GSE19063)/Homer	1e-47	-1.091e+02	0.0000	899.0	8.15%	1872.8	4.89%	ONECUT

Iri-specific opening DAR HOMER results



Rank	Motif	Name	P-value	log P-pvalue	q-value (Benjamini)	# Target Sequences with Motif	% of Targets Sequences with Motif	# Background Sequences with Motif	% of Background Sequences with Motif	Group
1		Phox2a(Homeobox)/Neuron-Phox2a-ChIP-Seq(GSE31456)/Homer	1e-732	-1.687e+03	0.0000	3486.0	41.83%	6089.3	15.31%	EN2
2		Prop1(Homeobox)/GHFT1-PROP1.biotin-ChIP-Seq(GSE77302)/Homer	1e-577	-1.329e+03	0.0000	4067.0	48.80%	9124.8	22.95%	EN2
3		MITF(bHLH)/MastCells-MITF-ChIP-Seq(GSE48085)/Homer	1e-375	-8.640e+02	0.0000	3087.0	37.04%	7090.7	17.83%	MiT
4		Usf2(bHLH)/C2C12-Usf2-ChIP-Seq(GSE36030)/Homer	1e-220	-5.070e+02	0.0000	1432.0	17.18%	2731.2	6.87%	BHLHE41
5		USF1(bHLH)/GM12878-Usf1-ChIP-Seq(GSE32465)/Homer	1e-175	-4.040e+02	0.0000	1476.0	17.71%	3210.5	8.07%	BHLHE41
6		Pax7(Paired,Homeobox),long/Myoblast-Pax7-ChIP-Seq(GSE25064)/Homer	1e-168	-3.881e+02	0.0000	480.0	5.76%	474.8	1.19%	EN2
7		BMAL1(bHLH)/Liver-Bmal1-ChIP-Seq(GSE39860)/Homer	1e-161	-3.723e+02	0.0000	4606.0	55.27%	16106.0	40.50%	BHLHE41
8		CLOCK(bHLH)/Liver-Clock-ChIP-Seq(GSE39860)/Homer	1e-104	-2.402e+02	0.0000	1515.0	18.18%	4081.9	10.27%	BHLHE41
9		Lhx1(Homeobox)/EmbryoCarcinoma-Lhx1-ChIP-Seq(GSE70957)/Homer	1e-91	-2.109e+02	0.0000	4040.0	48.48%	14919.5	37.52%	EN2
10		TFE3(bHLH)/MEF-TFE3-ChIP-Seq(GSE75757)/Homer	1e-87	-2.012e+02	0.0000	538.0	6.46%	965.8	2.43%	BHLHE41
11		Lhx3(Homeobox)/Neuron-Lhx3-ChIP-Seq(GSE31456)/Homer	1e-82	-1.896e+02	0.0000	5234.0	62.80%	20808.0	52.33%	EN2
12		Lhx2(Homeobox)/HFSC-Lhx2-ChIP-Seq(GSE48068)/Homer	1e-72	-1.668e+02	0.0000	3625.0	43.50%	13499.2	33.95%	EN2
13		NPAS2(bHLH)/Liver-NPAS2-ChIP-Seq(GSE39860)/Homer	1e-71	-1.638e+02	0.0000	2442.0	29.30%	8342.6	20.98%	BHLHE41
14		Sox9(HMG)/Limb-SOX9-ChIP-Seq(GSE73225)/Homer	1e-60	-1.397e+02	0.0000	1880.0	22.56%	6221.4	15.65%	SOX
15		NPAS(bHLH)/Liver-NPAS-ChIP-Seq(GSE39860)/Homer	1e-59	-1.382e+02	0.0000	3884.0	46.60%	15024.2	37.78%	BHLHE41
16		PAX3:FKHR-fusion(Paired,Homeobox)/Rh4-PAX3:FKHR-ChIP-Seq(GSE19063)/Homer	1e-42	-9.807e+01	0.0000	870.0	10.44%	2558.0	6.43%	ONECUT
17		Max(bHLH)/K562-Max-ChIP-Seq(GSE31477)/Homer	1e-41	-9.571e+01	0.0000	1489.0	17.87%	5033.0	12.66%	BHLHE41
18		HLF(bZIP)/HSC-HLF.Flag-ChIP-Seq(GSE69817)/Homer	1e-40	-9.425e+01	0.0000	2748.0	32.97%	10463.7	26.31%	CEBPB
19		Isl1(Homeobox)/Neuron-Isl1-ChIP-Seq(GSE31456)/Homer	1e-40	-9.400e+01	0.0000	4488.0	53.85%	18484.3	46.48%	ISL
20		Atf3(bZIP)/GBM-ATF3-ChIP-Seq(GSE33912)/Homer	1e-39	-9.175e+01	0.0000	1728.0	20.73%	6070.7	15.27%	JDP3



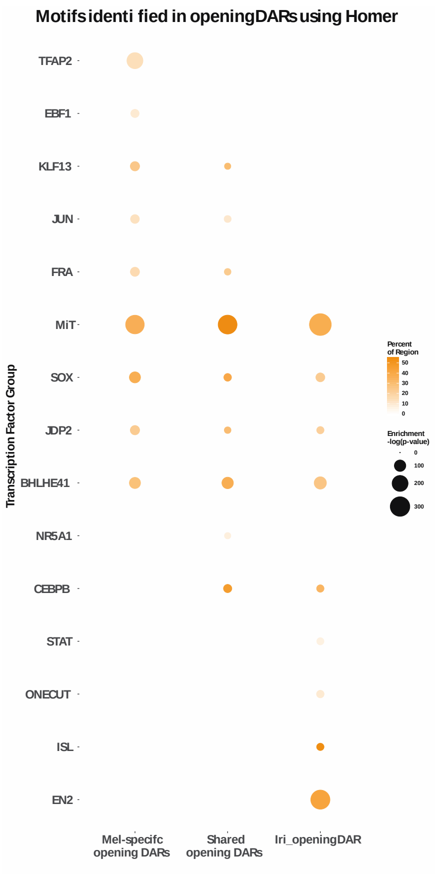


Figure 12. Homer motif enrichment results from dynamic DMR and DARs.

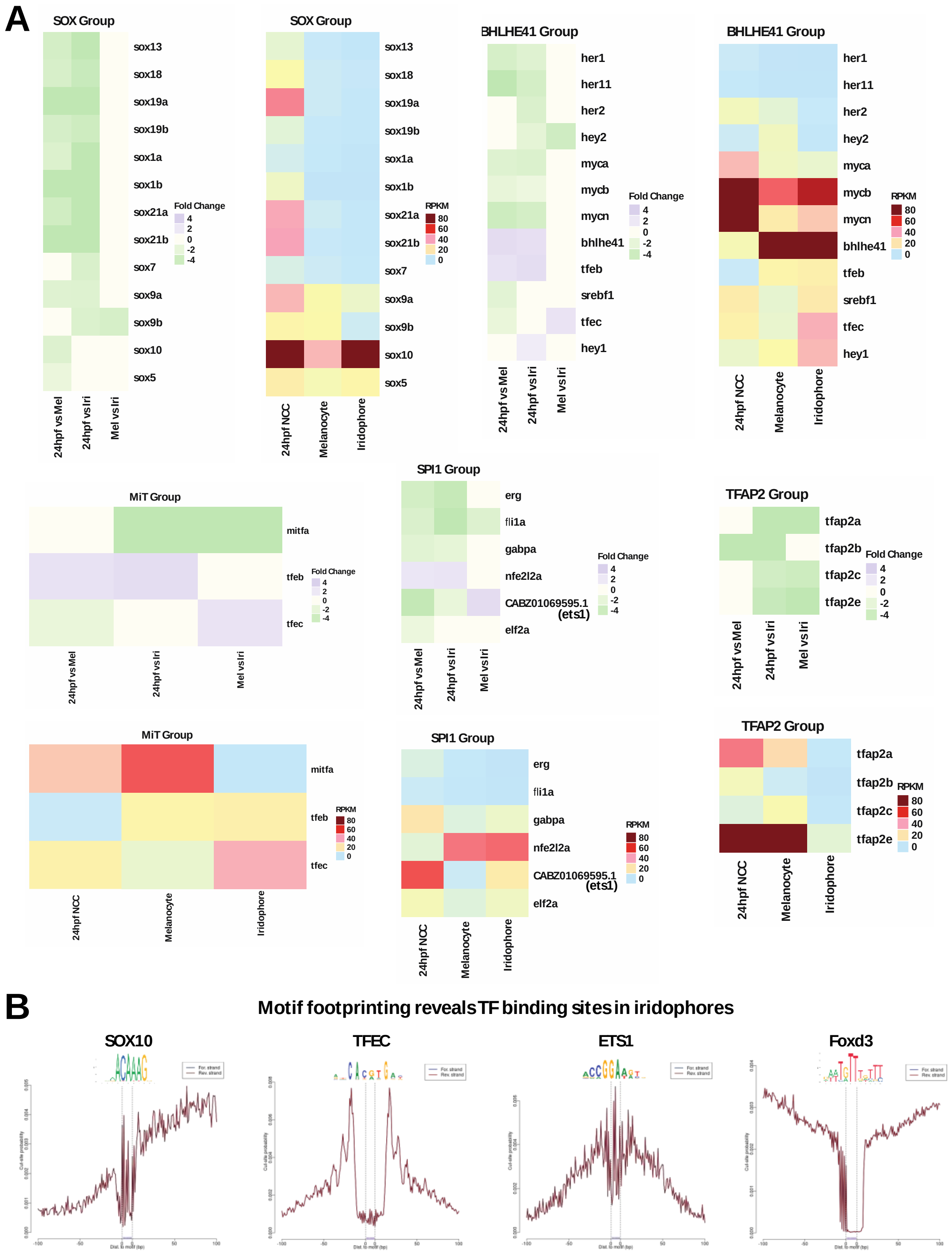


Figure 13. Pigment cell-specific TFs identified by motif enrichment and gene expression analysis. A) Gene expression heatmap of TFs predicted to bind to dynamic epigenetic regions.

B) Motif footprinting signatures of other TF candidates discovered by motif enrichment analysis in iridophores.

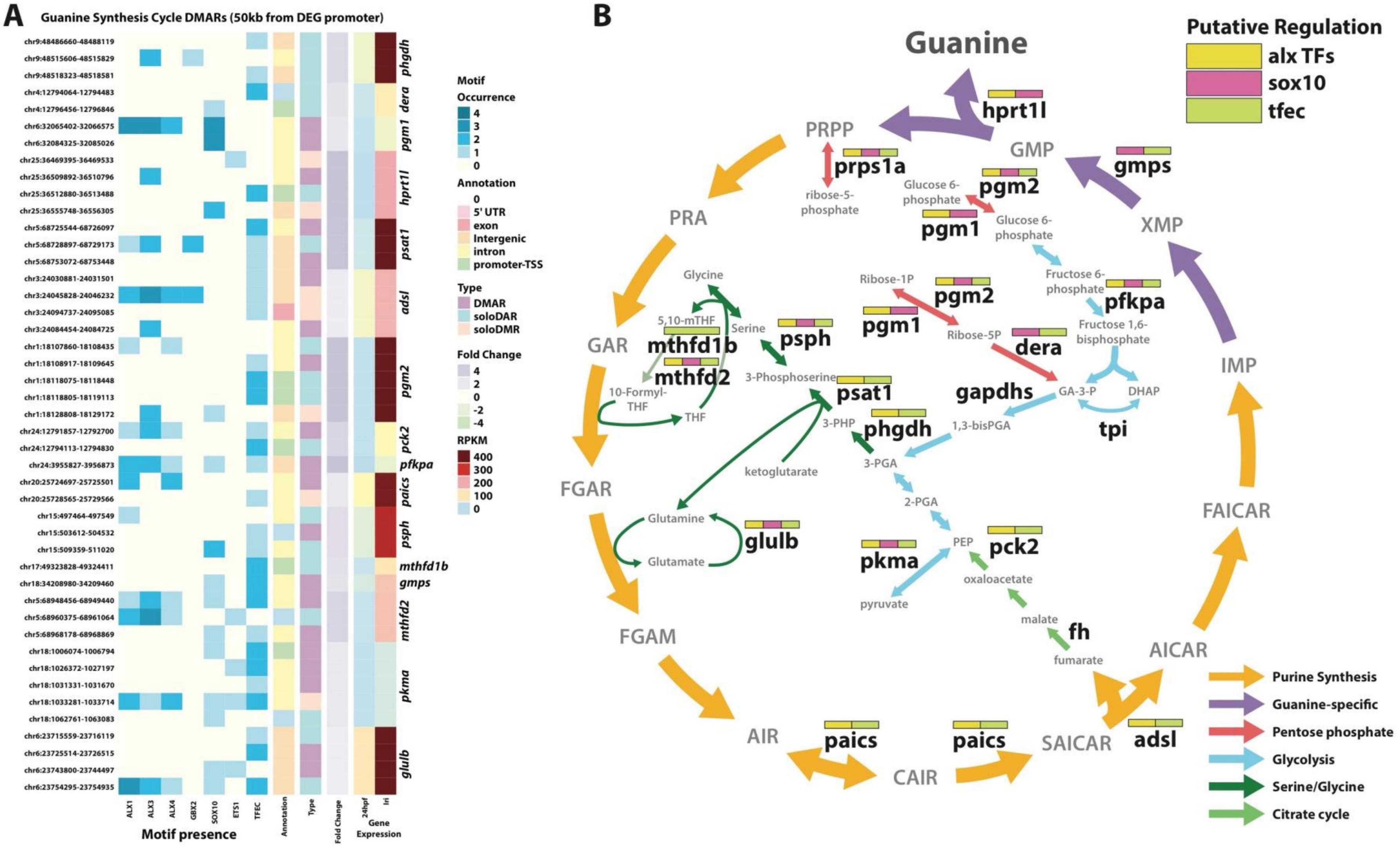

A
Guanine Synthesis Cycle DMARs (50kb from DEG promoter)
chr9:48486660-48488119
chr9:48515606-48515829
chr9:48518323-48518581
chr4:12794064-12794483
chr4:12796456-12796846
chr6:32065402-32066575
chr6:32084325-32085026
chr25:36469395-36469533
chr25:36509892-36510796
chr25:36512880-36513488
chr25:36555748-36556305
chr5:68725544-68726097
chr5:68728897-68729173
chr5:68753072-68753448
chr3:24030881-24031501
chr3:24045828-24046232
chr3:24094737-24095085
chr3:24084454-24084725
chr1:18107860-18108435
chr1:18108917-18109645
chr1:18118075-18118448
chr1:18118805-18119113
chr1:18128808-18129172
chr24:12791857-12792700
chr24:12794113-12794830
chr24:3955827-3956873
chr20:25724697-25725501
chr20:25728565-25729566
chr15:497464-497549
chr15:503612-504532
chr15:509359-511020
chr17:49323828-49324411
chr18:34208980-34209460
chr5:68948456-68949440
chr5:68960375-68961064
chr5:68968178-68968869
chr18:1006074-1006794
chr18:1026372-1027197
chr18:1031331-1031670
chr18:1033281-1033714
chr18:1062761-1063083
chr6:23715559-23716119
chr6:23725514-23726515
chr6:23743800-23744497
chr6:23754295-23754935
phgdh
dera
pgm1
hprt1l
psat1
adsl
pgm2
pck2
pfkpa
paics
psph
mthfd1b
gmps
mthfd2
pkma
glulb
ALX1
ALX3
ALX4
GBX2
SOX10
ETS1
TFEC
Motif presence
Annotation
Type
Fold Change
24hpf
Iri
Gene Expression
Motif
Occurrence
4
3
2
1
0
Annotation
0
5' UTR
exon
Intergenic
intron
promoter-TSS
Type
DMAR
soloDAR
soloDMR
Fold Change
4
2
0
-2
-4
RPKM
400
300
200
100
0
B
Guanine
Putative Regulation
alx TFs
sox10
tfec
hprt1l
PRPP
GMP
gmps
prps1a
pgm2
ribose-5-phosphate
Glucose 6-phosphate
pgm1
Glucose 6-phosphate
XMP
PRA
Fructose 6-phosphate
pfkpa
Glycine
Ribose-1P
pgm2
5,10-mTHF
Serine
psph
pgm1
Ribose-5P
dera
Fructose 1,6-bisphosphate
GAR
mthfd1b
IMP
mthfd2
3-Phosphoserine
psat1
gapdhs
GA-3-P
DHAP
10-Formyl-THF
THF
3-PHP
phgdh
1,3-bisPGA
tpi
3-PGA
ketoglutarate
FGAR
2-PGA
FAICAR
Glutamine
glulb
PEP
pck2
Glutamate
pkma
oxaloacetate
pyruvate
malate
fh
FGAM
fumarate
AICAR
Purine Synthesis
Guanine-specific
Pentose phosphate
Glycolysis
Serine/Glycine
Citrate cycle
AIR
paics
paics
SAICAR
adsl
CAIR

Figure 14. Iridophore-specific TFs predicted to regulate genes in guanine synthesis pathway. A) Heatmap profiling motif presence, genome annotation and epigenetic dynamics of DMARs within 50kb of DEGs in guanine synthesis pathway. B) Model of guanine synthesis cycle. Genes in bold are iridophore-specific DEGs. Color bars above DEGs represent presence of DMARs with TF motifs.

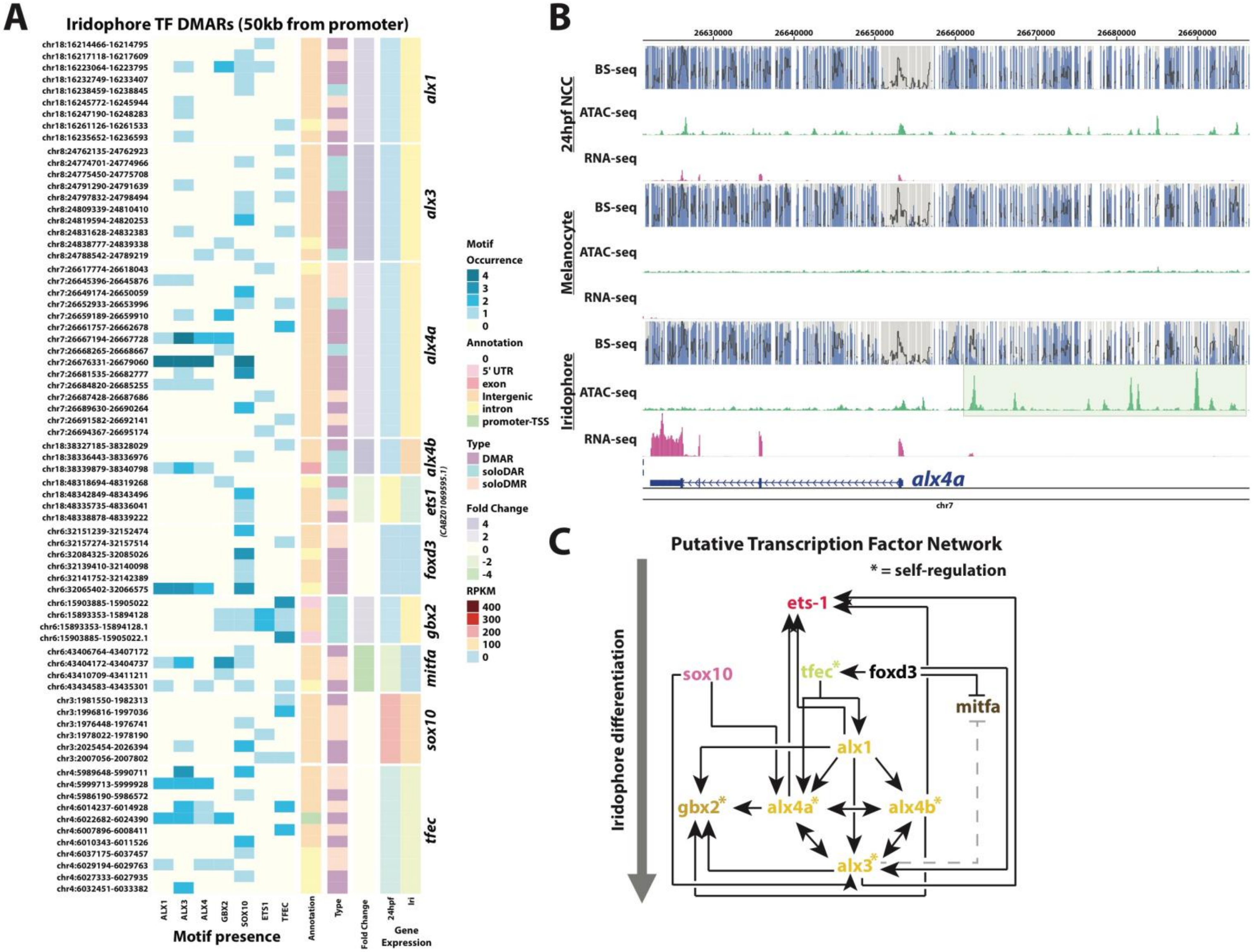

A
Iridophore TF DMARs (50kb from promoter)
chr18:16214466-16214795
chr18:16217118-16217609
chr18:16223064-16223795
chr18:16232749-16233407
chr18:16238459-16238845
chr18:16245772-16245944
chr18:16247190-16248283
chr18:16261126-16261533
chr18:16235652-16236593
chr8:24762135-24762923
chr8:24774701-24774966
chr8:24775450-24775708
chr8:24791290-24791639
chr8:24797832-24798494
chr8:24809339-24810410
chr8:24819594-24820253
chr8:24831628-24832383
chr8:24838777-24839338
chr8:24788542-24789219
chr7:26617774-26618043
chr7:26645396-26645876
chr7:26649174-26650059
chr7:26652933-26653996
chr7:26659189-26659910
chr7:26661757-26662678
chr7:26667194-26667728
chr7:26668265-26668667
chr7:26676331-26679060
chr7:26681535-26682777
chr7:26684820-26685255
chr7:26687428-26687686
chr7:26689630-26690264
chr7:26691582-26692141
chr7:26694367-26695174
chr18:38327185-38328029
chr18:38336443-38336976
chr18:38339879-38340798
chr18:48318694-48319268
chr18:48342849-48343496
chr18:48335735-48336041
chr18:48338878-48339222
chr6:32151239-32152474
chr6:32157274-32157514
chr6:32084325-32085026
chr6:32139410-32140098
chr6:32141752-32142389
chr6:32065402-32066575
chr6:15903885-15905022
chr6:15893353-15894128
chr6:15893353-15894128.1
chr6:15903885-15905022.1
chr6:43406764-43407172
chr6:43404172-43404737
chr6:43410709-43411211
chr6:43434583-43435301
chr3:1981550-1982313
chr3:1996816-1997036
chr3:1976448-1976741
chr3:1978022-1978190
chr3:2025454-2026394
chr3:2007056-2007802
chr4:5989648-5990711
chr4:5999713-5999928
chr4:5986190-5986572
chr4:6014237-6014928
chr4:6022682-6024390
chr4:6007896-6008411
chr4:6010343-6011526
chr4:6037175-6037457
chr4:6029194-6029763
chr4:6027333-6027935
chr4:6032451-6033382
ALX1
ALX3
ALX4
GBX2
SOX10
ETS1
TFEC
Motif presence
Annotation
Type
Fold Change
24hpf
Iri
Gene Expression
alx1
alx3
alx4a
alx4b
ets1
(CABZ01069595.1)
foxd3
gbx2
mitfa
sox10
tfec
Motif Occurrence
4
3
2
1
0
Annotation
0
5' UTR
exon
Intergenic
intron
promoter-TSS
Type
DMAR
soloDAR
soloDMR
Fold Change
4
2
0
-2
-4
RPKM
400
300
200
100
0
B
26630000
26640000
26650000
26660000
26670000
26680000
26690000
24hpf NCC
BS-seq
ATAC-seq
RNA-seq
Melanocyte
BS-seq
ATAC-seq
RNA-seq
Iridophore
BS-seq
ATAC-seq
RNA-seq
alx4a
chr7
C
Putative Transcription Factor Network
* = self-regulation
Iridophore differentiation
ets-1
sox10
tfec*
foxd3
mitfa
alx1
gbx2*
alx4a*
alx4b*
alx3*

Figure 15. Putative transcription factor network that drive iridophore cell fate. A) Heatmap profiling motif presence, genome annotation and epigenetic dynamics of DMARs within 50kb of iridophore-related transcription factors. B) WashU Epigenome browser view of DMAR clusters upstream of *alx4a* promoter. In BS-seq tracks, each bar represents presence of CpG and the blue color represents methylation level. Green peaks in ATAC-seq tracks represent accessible regions. Gene expression (RPKM) is represented by pink signal in RNA-seq tracks. C) Putative transcription factor network that drive iridophore cell fate.

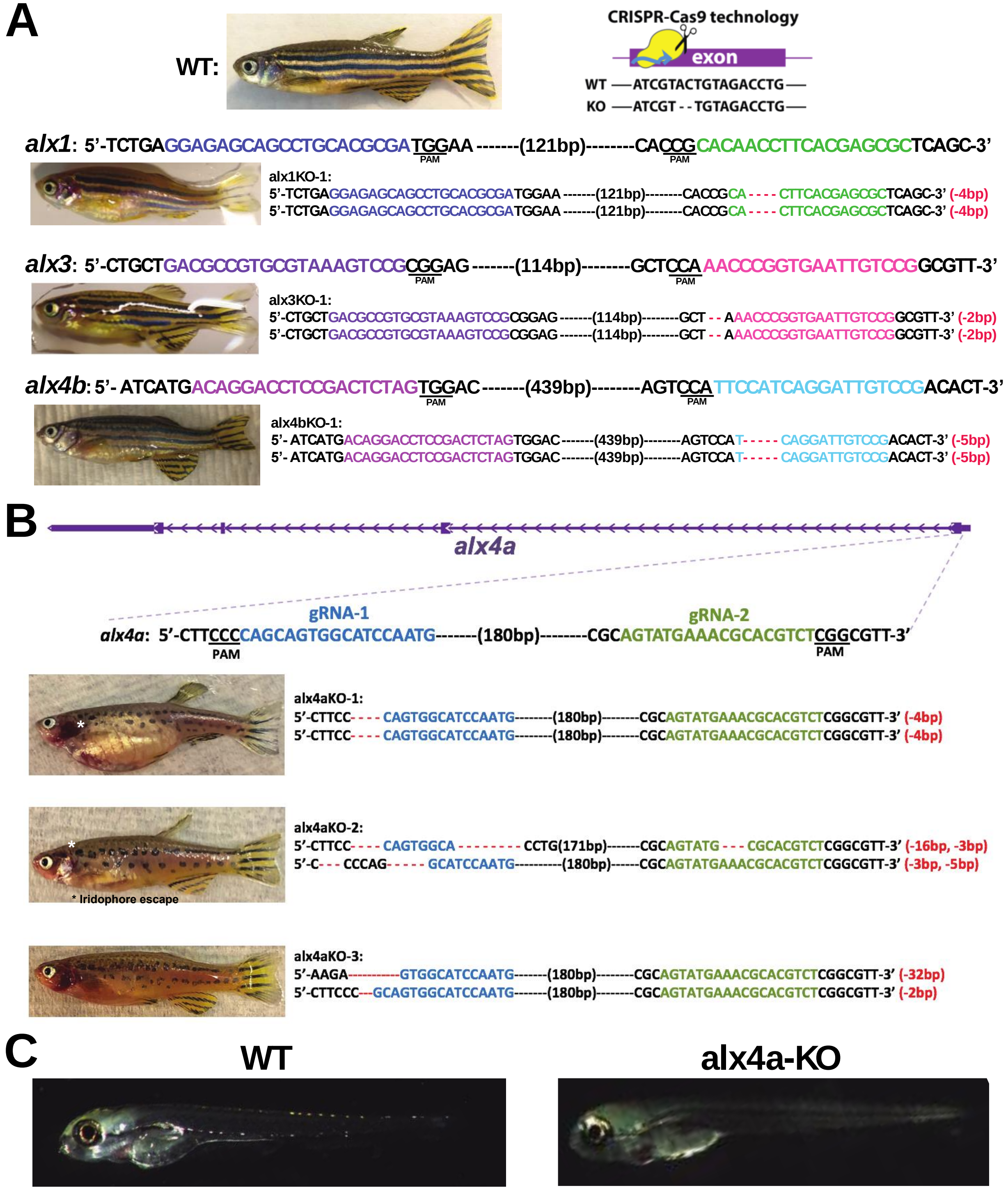

A
WT:
CRISPR-Cas9 technology
exon
WT —ATCGTACTGTAGACCTG—
KO —ATCGT - - TGTAGACCTG—
alx1: 5’-TCTGAGGAGAGCAGCCTGCACGCGATGGAA -------(121bp)--------CACCGCACAACCTTCACGAGCGCTCAGC-3’
PAM
PAM
alx1KO-1:
5’-TCTGAGGAGAGCAGCCTGCACGCGATGGAA -------(121bp)--------CACCGCA - - - - CTTCACGAGCGCTCAGC-3’ (-4bp)
5’-TCTGAGGAGAGCAGCCTGCACGCGATGGAA -------(121bp)--------CACCGCA - - - - CTTCACGAGCGCTCAGC-3’ (-4bp)
alx3: 5’-CTGCTGACGCCGTGCGTAAAGTCCGCGGAG -------(114bp)--------GCTCCA AACCCGGTGAATTGTCCGGCGTT-3’
PAM
PAM
alx3KO-1:
5’-CTGCTGACGCCGTGCGTAAAGTCCGCGGAG -------(114bp)--------GCT - - AAACCCGGTGAATTGTCCGGCGTT-3’ (-2bp)
5’-CTGCTGACGCCGTGCGTAAAGTCCGCGGAG -------(114bp)--------GCT - - AAACCCGGTGAATTGTCCGGCGTT-3’ (-2bp)
alx4b: 5’- ATCATGACAGGACCTCCGACTCTAGTGGAC -------(439bp)--------AGTCCATTCCATCAGGATTGTCCGACACT-3’
PAM
PAM
alx4bKO-1:
5’- ATCATGACAGGACCTCCGACTCTAGTGGAC-------(439bp)--------AGTCCAT- - - - - CAGGATTGTCCGACACT-3’ (-5bp)
5’- ATCATGACAGGACCTCCGACTCTAGTGGAC-------(439bp)--------AGTCCAT- - - - - CAGGATTGTCCGACACT-3’ (-5bp)
B
alx4a
gRNA-1
gRNA-2
alx4a: 5’-CTTCCCCAGCAGTGGCATCCAATG-------(180bp)--------CGCAGTATGAAACGCACGTCTCGGCGTT-3’
PAM
PAM
alx4aKO-1:
5’-CTTCC- - - - CAGTGGCATCCAATG--------(180bp)--------CGCAGTATGAAACGCACGTCTCGGCGTT-3’ (-4bp)
5’-CTTCC- - - - CAGTGGCATCCAATG--------(180bp)--------CGCAGTATGAAACGCACGTCTCGGCGTT-3’ (-4bp)
*
alx4aKO-2:
5’-CTTCC- - - - CAGTGGCA - - - - - - - - CCTG(171bp)-------CGCAGTATG - - - CGCACGTCTCGGCGTT-3’ (-16bp, -3bp)
5’-C- - - CCCAG- - - - - GCATCCAATG----------(180bp)------CGCAGTATGAAACGCACGTCTCGGCGTT-3’ (-3bp, -5bp)
*
* Iridophore escape
alx4aKO-3:
5’-AAGA-----------GTGGCATCCAATG-------(180bp)--------CGCAGTATGAAACGCACGTCTCGGCGTT-3’ (-32bp)
5’-CTTCCC---GCAGTGGCATCCAATG-------(180bp)--------CGCAGTATGAAACGCACGTCTCGGCGTT-3’ (-2bp)
C
WT
alx4a-KO

Figure 16. alx4a is necessary for iridophore development. A) Genotype and phenotype of *alx1*, *alx3* and *alx4b* KO adults generated via CRISPR-Cas9 technology. gRNA target sequences are labeled in color. B) Genotypes and phenotypes of *alx4a* KO adults. C) Representative picture of 4dpf larvae illuminated for iridophore detection.

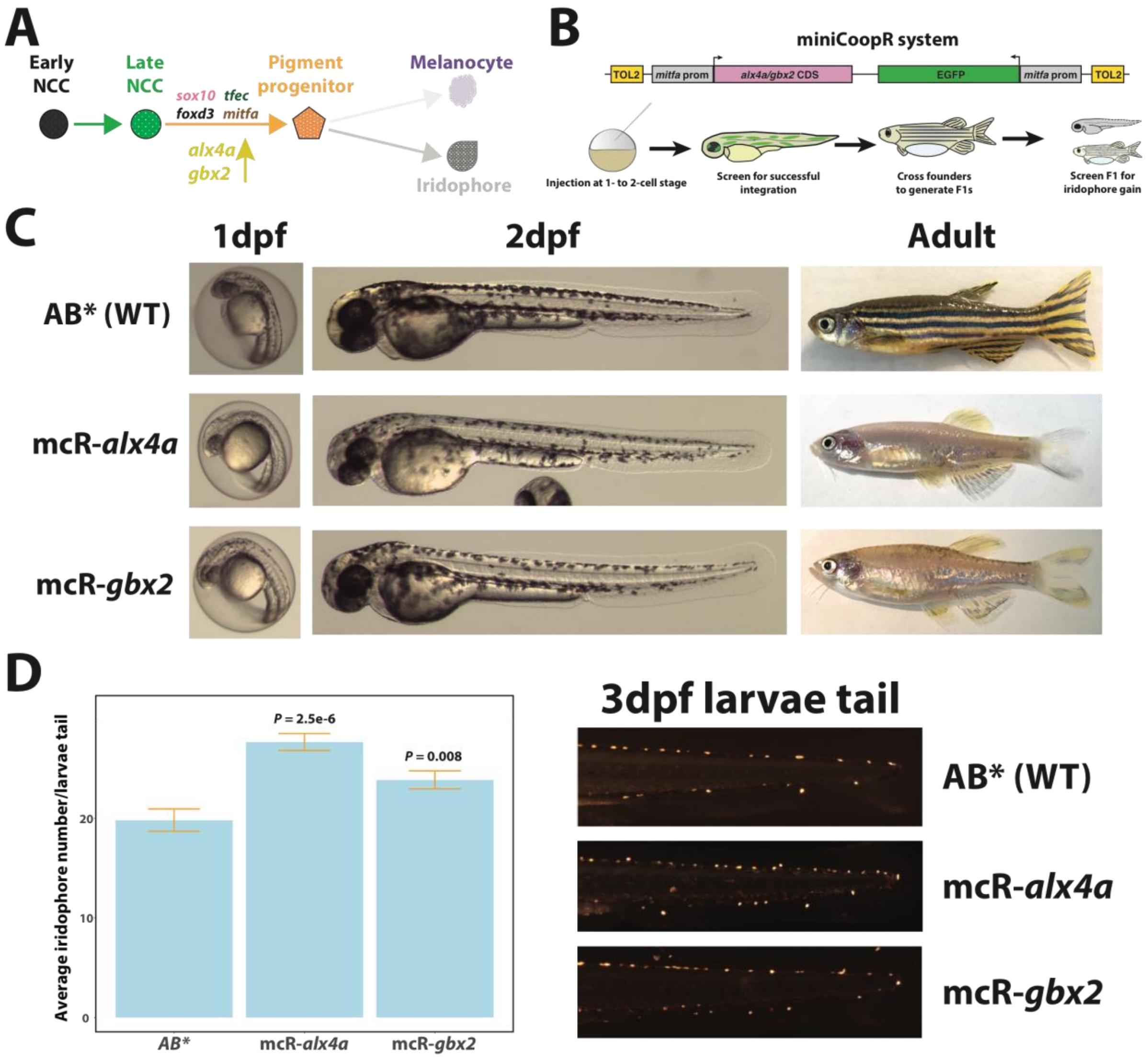


Figure 17. Ectopic expression of alx4a or gbx2 biases pigment progenitor cells towards iridophore cell fate. A) Time course representation of *alx4a* or *gbx2* ectopic expression using miniCoopR system. B) A graphical schematic of the miniCoopR system and experimental procedure. C) Delay of melanocyte formation and migration in 1dpf, 2dpf and adult miniCoopR transgenic fish. D) Mean number of tail iridophores in AB* (*n* = 21), miniCoopR-*alx4a* (*n* = 20)

and miniCoopR-*gbx2* (*n* = 20) at 3dpf larvae. Data is shown as mean ± s.e.m. P-values were derived from two-sided Welch's *t* test (compared to AB*). Representative pictures of illuminated tail iridophores of wild type and transgenic larvae at 3dpf.

Chapter 6: Transposable elements drive widespread expression of oncogenes in human cancers

Hyo Sik Jang[1,2], Nakul M. Shah[1,2], Alan Y. Du[1,2], Zea Z. Dailey[1,2], Erica C. Pehrsson[1,2], Paula M. Godoy[1,2], David Zhang[1,2], Daofeng Li[1,2], Xiaoyun Xing[1,2], Sungsu Kim[1,3], David O'Donnell[1,4], Jeffrey I. Gordon[1,4], Ting Wang[1,2]

[1]Department of Genetics, Washington University School of Medicine, St. Louis, Missouri, USA

[2]The Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, Missouri, USA

[3]Hope Center for Neurological Disease, Washington University School of Medicine, St. Louis, MO, USA

[4]Center for Gut Microbiome and Nutrition Research, Washington University School of Medicine, St. Louis, MO, USA

These authors contributed equally: Hyo Sik Jang and Nakul M. Shah.

Author Contributions: H.S.J., N.M.S. and T.W. conceived and implemented the study; N.M.S., H.S.J., E.C.P., D.L. and T.W. contributed to the computational analysis; H.S.J. generated transcriptomic and epigenomic profiles of cell lines; H.S.J., X.X. and D.Z. performed the CRISPR-mediated deletion experiments; H.S.J. and Z.Z.D. performed the promoter luciferase, motif mutagenesis, and let-7 qPCR experiments; H.S.J. and A.Y.D. performed the growth and migration assays; H.S.J., A.Y.D., D.O., and J.I.G. performed the xenograft experiments; H.S.J., P.G., and S.K. performed the targeted methylation experiments; H.S.J. performed the rescue experiments; and the manuscript was prepared and revised by H.S.J., N.M.S. and T.W. with input from all authors.

The work presented here is published in Nature Genetics. doi:10.1038/s41588-019-0373-3

6.1 Abstract

Transposable elements (TEs) are an abundant and rich genetic resource of regulatory sequences[244–246]. Cryptic regulatory elements within TEs can be epigenetically reactivated in cancer to influence oncogenesis in a process termed onco-exaptation[247]. However, the prevalence and impact of TE onco-exaptation events across cancer types are poorly characterized. Here, we analyzed 7,769 tumors and 625 normal datasets from 15 cancer types, identifying 129 TE cryptic promoter activation events involving 106 oncogenes across 3,864 tumors. Furthermore, we interrogated the AluJb-LIN28B candidate: the genetic deletion of the TE eliminated oncogene expression, while dynamic DNA methylation modulated promoter activity, illustrating the necessity and sufficiency of a TE for oncogene activation. Collectively, our results characterize the global profile of TE onco-exaptation and highlight this prevalent phenomenon as an important mechanism for promiscuous oncogene activation and ultimately tumorigenesis.

6.2 Main

The elucidation of mechanisms behind oncogene activation has been a long-standing goal in cancer biology. Genetic mutation, gene amplification, and chromosomal rearrangement are three classic genetic mechanisms that drive cancer progression and identity[248,249], but they provide an incomplete explanation for oncogene activation. Recently, a wave of discoveries has demonstrated how TEs change the gene expression landscape during evolution, development, and disease[244–246,250–252]. Although epigenetically silenced in somatic tissues, TEs can become active in cancer due to DNA hypomethylation, which can expose regulatory sequences and lead to functional consequences[253–255]. Indeed, some TEs are epigenetically reactivated as cryptic promoters to drive oncogene expression in cancer, a process known as onco-exaptation[247,256–261]. To our knowledge, no comprehensive study has investigated whether onco-exaptation is a widespread mechanism for oncogene activation across multiple cancer types.

To globally characterize onco-exaptation events, we canvassed RNA-seq data across 15 cancer types from the TCGA Research Network (http://cancergenome.nih.gov/) (Supplementary Fig. 1a). We constructed a computational pipeline that identifies TE-derived oncogene transcripts that are highly tumor-enriched (Supplementary Fig. 1b). A comprehensive list of 702 oncogenes was generated from previously annotated onco-exaptation examples[247,258] and ONGene[262] (Supplementary Table 1). Considering the technical limitations of RNA-seq data, we set stringent filters (Methods) to maximize the specificity for onco-exaptation events. In total, we analyzed 7,769 tumor samples and 625 tumor-matched normal samples (Supplementary Fig. 1b), which identified 625 TE-oncogene chimeric transcripts; this list includes five previously published onco-exaptation examples (Supplementary Table 2). After selecting further for high tumor-enrichment and expression contribution, we identified 129 high confidence onco-exaptation events across 106 oncogenes (Supplementary Table 3). In addition, we detected at least one onco-exaptation event in 49.7% of all tumors, with prevalence ranging from 12% to 87% across cancer types, indicating that onco-exaptation could be a promiscuous mechanism for oncogene activation (Fig. 1a). On average, each onco-exaptation event was discovered in 51 samples and often distributed across multiple cancer types. We report that the onco-exapted TEs strongly enrich for the long terminal repeat (LTR) class (Fig. 1b and Supplemental Fig. 2b). Examining the cancer-type distribution of onco-exaptation candidates (Fig. 1c) showed both cancer-type-specific events, such as THE1A-HMGA2 in SKCM[263], and highly prevalent events were present across multiple cancer types. Furthermore, for eight oncogenes, we observed various TEs activating an in-frame isoform of the same gene (Supplementary Table 4), a phenomenon that had only been described for one oncogene[259]. These additional examples support the cancer epigenetic evolution model as previously described[247]. In summary, we provide a global profile of onco-exaptation events across 15 cancer types and enumerate TEs' role in driving oncogene activation and upregulation.

Next, we examined transcript-level information for the top 10 most prevalent onco-exaptation candidates that on average accounted for greater than 50% of their target oncogenes expression (Fig. 1d). Eight of these candidates were predicted to form in-frame transcripts that conserve protein sequence, suggesting preservation of oncogene function. Onco-exaptation candidates include isoforms of genes such as *SALL4* and *LIN28B* that have recently emerged as potent cancer drivers[264–267]. Additionally, the L1PA2-derived isoform of *SYT1* occurs in more than 10% of all tumors, suggesting that it could be an important cancer marker. While investigating transcript-level abundance of candidates, we found that many of the onco-exaptation events were driving a significant fraction of oncogene expression; some greater than 90% (Fig. 1d & Supplementary Fig. 3). Furthermore, we report that half of the top candidates were associated with worse survival in at least 1 cancer type (Supplementary Fig. 4). For example, we show that the HERVH-SLCO1B3 transcript, a previously characterized onco-exaptation event, is abundant across various cancer types, highly expressed, and associated with worse prognosis[268]. These findings imply that TEs are not only associated with oncogene activation but also contribute significantly to overall oncogene expression and oncogenic potential.

For validation, we sought to confirm transcription initiation from a few exapted TEs. We queried the FANTOM5 promoter database[269] and discovered five out of the ten most prevalent onco-exaptation candidates show promoter signature. We validated a few FANTOM5 results by mapping transcription start sites (TSS) with Cap Analysis of Gene Expression (CAGE)-seq[269–271] in the H727 lung carcinoid cell line. Indeed, *SYT1* and *ARID3A* oncogenes are transcribed from alternative promoters located within TEs (Fig. 2a and Supplemental Fig. 5). In addition, we analyzed 27 RNA-seq datasets from lung cancer cell lines[272] and detected 5 of the 10 most prevalent onco-exaptation candidates (Supplementary Table 5). One of the most highly expressed candidates was an AluJb-LIN28B fusion transcript that is present in the H1299, RERF-LC-OK, and H838 cell lines. Considering that *LIN28B* is a well-characterized and potent oncogene[265,267,273–275], we pursued this candidate for further functional validation.

The AluJb TE is located 20 kb upstream of the canonical promoter of *LIN28B* and drives the majority of expression of LIN28B in a substantial number of tumors (Fig. 1d). To verify the existence of the AluJb-LIN28B isoform in lung cancer cell lines, we profiled TSSs in the H1299 and H838 cell lines using paired-end CAGE-seq. We confirmed a CAGE peak, composed of mate reads that align to *LIN28B*, that spans ~40 bp in the AluJb element in both cell lines (Fig. 2b). Next, we profiled DNA methylation levels and chromatin accessibility using WGBS-seq and ATAC-seq, respectively (Fig. 2b). The AluJb TE is completely methylated in somatic tissues profiled by Roadmap (http://www.roadmapepigenomics.org/) (Supplementary Fig. 6a). In H1299, the region surrounding the AluJb promoter (AluJb-P) is unmethylated, whereas in H838, it is ~50% methylated. In both cell lines, the region displayed accessibility, indicating an open chromatin state. Together, these findings suggest that an AluJb TE is epigenetically reactivated as an alternative promoter to drive LIN28B expression in lung cancer cell lines.

Next, we dissected the genetic determinants behind the AluJb-LIN28B onco-exaptation event. In H1299 and H838, we discovered that active epigenetic marks encompassed two TEs, a truncated AluJb and MLT1B, upstream of AluJb-P (Fig. 2b). Since various TEs are known to harbor transcription factor binding sites that could have *cis*-regulatory function[245], we tested whether these upstream TEs impact AluJb-P promoter strength. Luciferase assays using various combinations of TEs before a luciferase reporter showed that vectors without AluJb-P displayed minimal activity (Fig. 2c). Furthermore, the luciferase activity did not diminish in the solo AluJb-P vector relative to other vectors. These results illustrate that AluJb-P contains all the necessary sequences for strong promoter activity, and the upstream TEs have minimal *cis*-regulatory effect on AluJb-P transcription.

AluJb is a primate-specific subfamily within the short interspersed nuclear element (SINE) class of TEs. SINE elements are known to recruit RNA polymerase (RNAP) III to generate short transcripts that can potentially be retrotransposed[276]. However, majority of mRNAs are typically transcribed by RNAP II. We hypothesized that AluJb-P accumulated mutations through evolution that generated novel transcription factor binding sites that recruit RNAP II. To explore this hypothesis, we performed pair-wise sequence alignment using EMBOSS Needle[277] between the AluJb-P sequence and the AluJb consensus sequence from Dfam[278]. We then identified potential novel transcription factor motifs that were generated by mutations specific to AluJb-P with FIMO[279]. Previous work has demonstrated that NFYA binds to AluJb-P and knockdown of NFYA reduces promoter activity in Huh-7 cells[280]. However, the degree of NFYA's impact on AluJb promoter function is still unclear. Our analysis with FIMO detected four other transcription factor motifs that potentially arose from mutations: C/EBPD, SP1, SP4, and YY1 (Fig. 2d). To interrogate the functional importance of these motifs, we cloned AluJb-P sequences mutagenized for each motif into a luciferase reporter and assessed the change in promoter activity. In both H1299 and H838, mutating SP1, SP4, and YY1 sites significantly diminished relative luciferase expression, which is consistent with previous findings that SP transcription factors cooperate with YY1 to drive strong promoter expression (Fig. 2d)[281]. Furthermore, these results were recapitulated in the K562 leukemia cell line (Supplementary Fig. 8a,b), which does not express the AluJb-LIN28B transcript. This finding suggests that K562 cells have all the transcriptional machinery to transcribe from the AluJb-P, but DNA methylation might be suppressing the activity of the promoter (Supplementary Fig. 6a).

To evaluate the functional consequences of the AluJb-LIN28B onco-exaptation event, we first investigated whether the fusion transcript produces a protein product. Within the AluJb-P sequence, we detected a strong start codon 72 bp downstream of the TSS. This results in the addition of 22 amino acids at the N-terminus of exon 2 of LIN28B (Supplementary Fig. 6c), for a predicted protein size increase of 2.5 kDA compared to normal LIN28B. Western blots verified the expected size difference between the onco-exapted AluJB-LIN28B isoform present in H1299 and H838 cells compared to the canonical LIN28B protein present in K562 and HepG2 (Supplementary Fig. 6d). To confirm that the larger protein originated from AluJb-P, we performed CRISPR-Cas9-mediated deletion of AluJb-P in H1299 and H838 (Fig. 3a). In addition, we deleted a 1-kb sequence of the canonical *LIN28B* promoter (LIN28BP). The deletion of AluJb-P abolished the larger LIN28B protein, while the deletion of LIN28BP did not (Fig. 3b), verifying that AluJb-P produced the larger LIN28B isoform.

Since the AluJb-LIN28B protein is identical to canonical LIN28B, aside from the additional N-terminal amino acids, we examined whether AluJb-LIN28B retained normal LIN28B function. LIN28B represses let-7 miRNAs[273,274,282–284], ultimately contributing to oncogenesis through the upregulation of oncogenes such as MYC and RAS[265,267,275]. As anticipated, we observed an appreciable increase in the levels of let-7a, let-7b and let-7g in the AluJb-P knockout (KO) cells but not in LIN28BP KO cells of H1299 and H838 (Fig. 3c). We further assessed how the deletion of AluJb-P impacts cancer-specific attributes. In both H1299 and H838, AluJb-P KO cells show much slower growth (Fig. 3d) and migration (Fig. 3e) relative to the parental cell lines and LIN28BP KO cells. Also, parental H1299 and LIN28BP KO clone established rapidly growing tumors *in vivo,* whereas AluJb-P KO cells exhibited a marked defect in tumor growth during the time of inspection (Fig. 3f), consistent with the necessity of LIN28B for tumor growth in murine xenograft models[266,280]. In contrast, the deletion of AluJb-P in K562 cells did not result in elevated let-7 levels (Supplementary Fig. 8e) or loss of proliferation (Supplementary Fig. 8f), implying that the loss of AluJb-LIN28B was causal for the decreased oncogenic attributes in H1299 and H838 cells and not due to an off-target effect. Additionally, re-expression of canonical LIN28B and AluJb-LIN28B in H1299 and H838 AluJb-P KO cells reduced let-7 miRNA levels and modestly rescued proliferation (Fig. 3g). Altogether, these results indicate that TE-induced oncogene expression can retain its canonical function, which contributes to cell proliferation, migration, and tumor formation.

Most tumors exhibit global DNA hypomethylation, which provides cancer cells with an opportunity to exploit the regulatory potential of TEs. However, whether the loss of DNA methylation is causal for spurring TE's cryptic promoter activity has been underexplored due to a lack of efficient targeted methylation techniques. To directly assess how DNA methylation regulates AluJb-P activity, we utilized the CRISPR SUperNova tagging system (SunTag) to recruit either DNMT3A or TET1CD for targeted methylation or demethylation, respectively (Fig. 4a,b)[285–287]. This system allowed us to modestly increase DNA methylation of the AluJb TE by ~20-30% (Fig. 4c), which led to ~40% decrease in LIN28B expression in the H1299 (Fig. 4d), suggesting that DNA methylation of the TE is sufficient to decrease oncogene expression. Additionally, demethylation of the AluJb TE in K562 (Fig. 4e) led to the production of AluJb-LIN28B fusion protein (Fig. 4f). These results illustrate that dynamic DNA methylation is a driving epigenetic control that act as on-off switch for AluJb-P's activity and moreover suggests that TE onco-exaption events arise in tumors due to the unique epigenetic landscape.

6.3 Discussion:

In conclusion, TEs provide an additional means by which cancer can activate oncogenes. Stochastic, global DNA hypomethylation of cancer cells indiscriminately resurrects TEs of varying regulatory ability, which, if they confer a fitness advantage, can be epigenetically inherited and selectively propagated during tumor progression. Here, we present a global profile of tumor-enriched, TE-derived oncogene transcripts across 15 cancer types and show that onco-exaptation is a highly prevalent and promiscuous mechanism that contributes to oncogene activation in close to half of all tumors. By dissecting the mechanisms behind AluJb-derived LIN28B expression, we describe how TEs may be epigenetically and transcriptionally activated to drive oncogene expression. Recently, this tumor-specific *LIN28B* alternative promoter usage in liver cancer has also been characterized by Guo et al. (2018)[280], but not in an onco-exaptation context. Our concomitant findings in lung cancer cell lines provide cross-cancer support of the robust oncogenic potential of AluJb-LIN28B. Recognizing onco-exaptation events can provide additional insights into potential genetic and epigenetic mechanisms that drive promoter activity in cancer. For example, we were able to identify additional putative transcription factors that might be controlling AluJb promoter activity by exploring the evolution of the SINE element. Furthermore, we provide evidence that these onco-exaptation events are potentially reversible through targeted epigenetic alterations, which could present a translational avenue for personalized epigenetic oncotherapy. In summary, TEs act as double-edged swords for cancer by offering additional mechanisms for oncogene activation but also providing a potential target for therapeutics.

6.4 Acknowledgments

We would like to thank J. Hoisington-López and M.L. Jaeger from The Edison Family Center for Genome Sciences & Systems Biology (CGSSB) for assistance with sequencing; B. Koebbe and

E. Martin from CGSSB for data processing; M. Savio, M. Patana & D. Schweppe from the Siteman Flow Cytometry core for FACS-related expertise; Dr. M. Goodell and Y. Huang for valuable expertise with the SunTag-DNMT3A system; Dr. L. Maggi (Washington University School of Medicine) for generously gifting the H838 cell line. We would like to thank anonymous reviewers for constructive and insightful suggestions. This work was funded by NIH grants 5R01HG007175, U24ES026699, U01HG009391 and the American Cancer Society Research Scholar grant RSG-14-049-01-DMC. HSJ was supported by a grant from NIGMS (T32 GM007067). NMS is a Howard Hughes Medical Institute (H.H.M.I.) Medical Research Fellow, http://www.hhmi.org/. The xenograft work cited in this publication was performed in a facility supported by NCRR grant C06 RR015502.

Competing Interests:

Authors declare no competing interests.

6.5 Figures

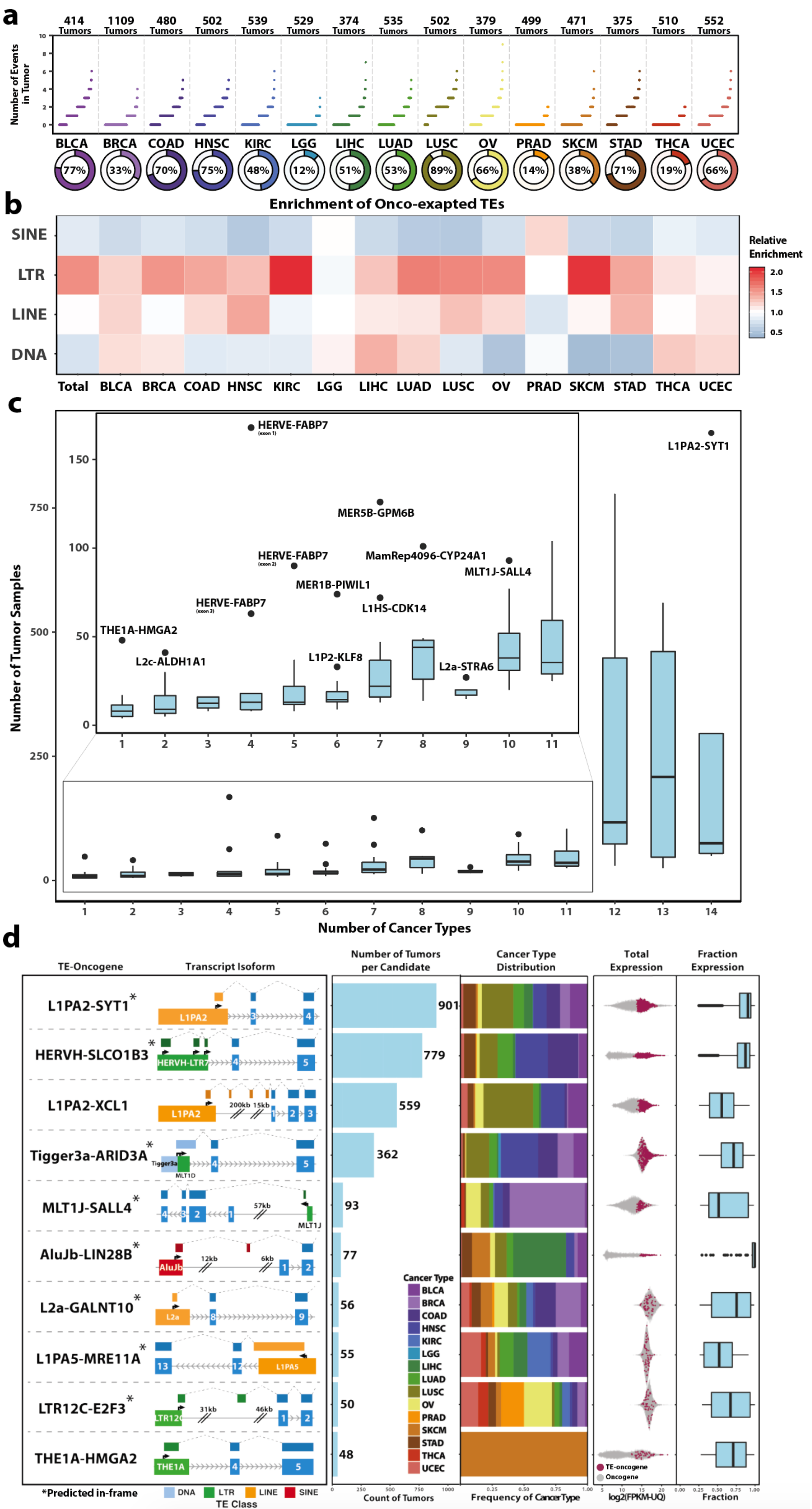
a
414 Tumors
1109 Tumors
480 Tumors
502 Tumors
539 Tumors
529 Tumors
374 Tumors
535 Tumors
502 Tumors
379 Tumors
499 Tumors
471 Tumors
375 Tumors
510 Tumors
552 Tumors
Number of Events in Tumor
0
2
4
6
8
10
BLCA 77%
BRCA 33%
COAD 70%
HNSC 75%
KIRC 48%
LGG 12%
LIHC 51%
LUAD 53%
LUSC 89%
OV 66%
PRAD 14%
SKCM 38%
STAD 71%
THCA 19%
UCEC 66%
b
Enrichment of Onco-exapted TEs
SINE
LTR
LINE
DNA
Total
BLCA
BRCA
COAD
HNSC
KIRC
LGG
LIHC
LUAD
LUSC
OV
PRAD
SKCM
STAD
THCA
UCEC
Relative Enrichment
2.0
1.5
1.0
0.5
c
Number of Tumor Samples
0
250
500
750
150
100
50
0
HERVE-FABP7 (exon 1)
MER5B-GPM6B
HERVE-FABP7 (exon 2)
MamRep4096-CYP24A1
MLT1J-SALL4
MER1B-PIWIL1
L1HS-CDK14
HERVE-FABP7 (exon 3)
THE1A-HMGA2
L2c-ALDH1A1
L1P2-KLF8
L2a-STRA6
L1PA2-SYT1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Number of Cancer Types
d
TE-Oncogene
Transcript Isoform
Number of Tumors per Candidate
Cancer Type Distribution
Total Expression
Fraction Expression
L1PA2-SYT1*
HERVH-SLCO1B3*
L1PA2-XCL1
Tigger3a-ARID3A*
MLT1J-SALL4*
AluJb-LIN28B*
L2a-GALNT10*
L1PA5-MRE11A*
LTR12C-E2F3*
THE1A-HMGA2
L1PA2
HERVH-LTR7
200kb
15kb
Tigger3a
MLT1D
57kb
MLT1J
AluJb
12kb
6kb
L2a
L1PA5
LTR12C
31kb
46kb
THE1A
901
779
559
362
93
77
56
55
50
48
Cancer Type
BLCA
BRCA
COAD
HNSC
KIRC
LGG
LIHC
LUAD
LUSC
OV
PRAD
SKCM
STAD
THCA
UCEC
TE-oncogene
Oncogene
*Predicted in-frame
DNA
LTR
LINE
SINE
TE Class
0
250
500
750
1000
Count of Tumors
0.25
0.50
0.75
1.00
Frequency of CancerType
5
10
15
20
log2(FPKM-UQ)
0.00
0.25
0.50
0.75
1.00
Fraction

Fig. 1: The TE onco-exaptation landscape across cancer types. a, Frequency of onco-exaptation events per tumor across cancer types. Donut plot reports the percent of tumor samples with at least one event. **b,** Enrichment of TE class in onco-exapted TEs across cancer types. **c,** A series of boxplots that highlight the distribution of the total number of tumor samples per candidate that is present in a certain number of cancer types. We have zoomed in on 1-11 so that the distribution can be more clearly seen. Each box represents the median and interquartile range, and the whiskers are 1.5× the IQR. Below each boxplot, we have labeled the number of candidates. We have also labeled all the outlier candidates. **d,** The top 10 most prevalent onco-exaptation candidates are presented. The left-most panel gives the TE-oncogene candidate label as well as a diagram of the transcript structure of the candidate. The next two panels display the number of tumor samples each candidate is present in as well as the distribution of the candidate across cancer types. The "Total Expression" panel displays the expression of the oncogene across all the tumor samples as grey dots, and the samples with the onco-exaptation candidate are highlighted in red. The "Fraction Expression" panel displays a boxplot of the percent of total expression of the oncogene contributed by the onco-exaptation candidate across the samples in which the candidate is present. Each box represents the median and interquartile range, and the whiskers are 1.5× the IQR.

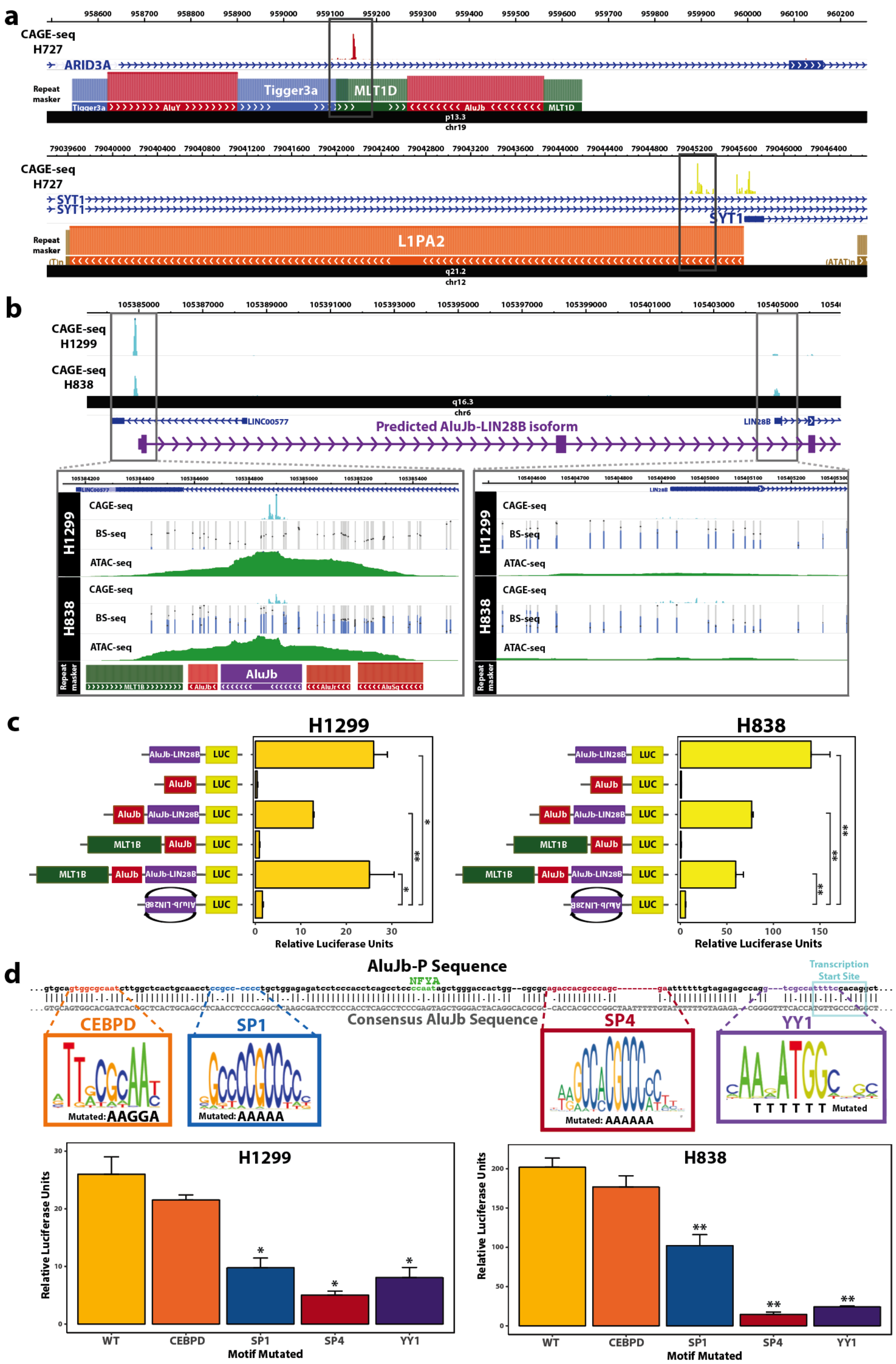

a
CAGE-seq H727
ARID3A
Repeat masker
Tigger3a
MLT1D
AluY
AluJb
p13.3
chr19
SYT1
L1PA2
(ATAT)n
q21.2
chr12
b
CAGE-seq H1299
CAGE-seq H838
q16.3
chr6
LINC00577
Predicted AluJb-LIN28B isoform
LIN28B
H1299
H838
CAGE-seq
BS-seq
ATAC-seq
Repeat masker
MLT1B
AluJb
AluJr
AluSq
c
H1299
H838
AluJb-LIN28B
AluJb
MLT1B
LUC
Relative Luciferase Units
d
AluJb-P Sequence
NFYA
Consensus AluJb Sequence
Transcription Start Site
CEBPD
Mutated: AAGGA
SP1
Mutated: AAAAA
SP4
Mutated: AAAAAA
YY1
T T T T T T Mutated
H1299
H838
Relative Luciferase Units
WT
CEBPD
SP1
SP4
YY1
Motif Mutated

Fig. 2: TEs provide bona fide promoters for oncogenes in lung cancer cell lines. a, CAGE-seq profile of H727 across onco-exaptation candidates (*ARID3A* & *SYT1*) visualized on WashU Epigenome Browser. Signals in CAGE-seq represent TSS locations. **b,** CAGE-seq and epigenetic profiles of the AluJb TE in the H1299 and H838. Signal in ATAC-seq represent open chromatin regions. Grey bars in the BS-seq track represent CpG locations while the height of blue bars indicate methylation %. **c,** Luciferase assays for transcriptional activity of various TE arrangements in H1299 (left) and H838 (right) (n = 3 independent experiments). **d,** Luciferase assays for promoter activity in H1299 (left) and H838 (right) with mutagenized transcription factor motifs in AluJb-P (n = 3 independent experiments). **c, d,** *P* values were derived from two-tailed Welch *t* test. All data are represented as means ± standard error (SE).

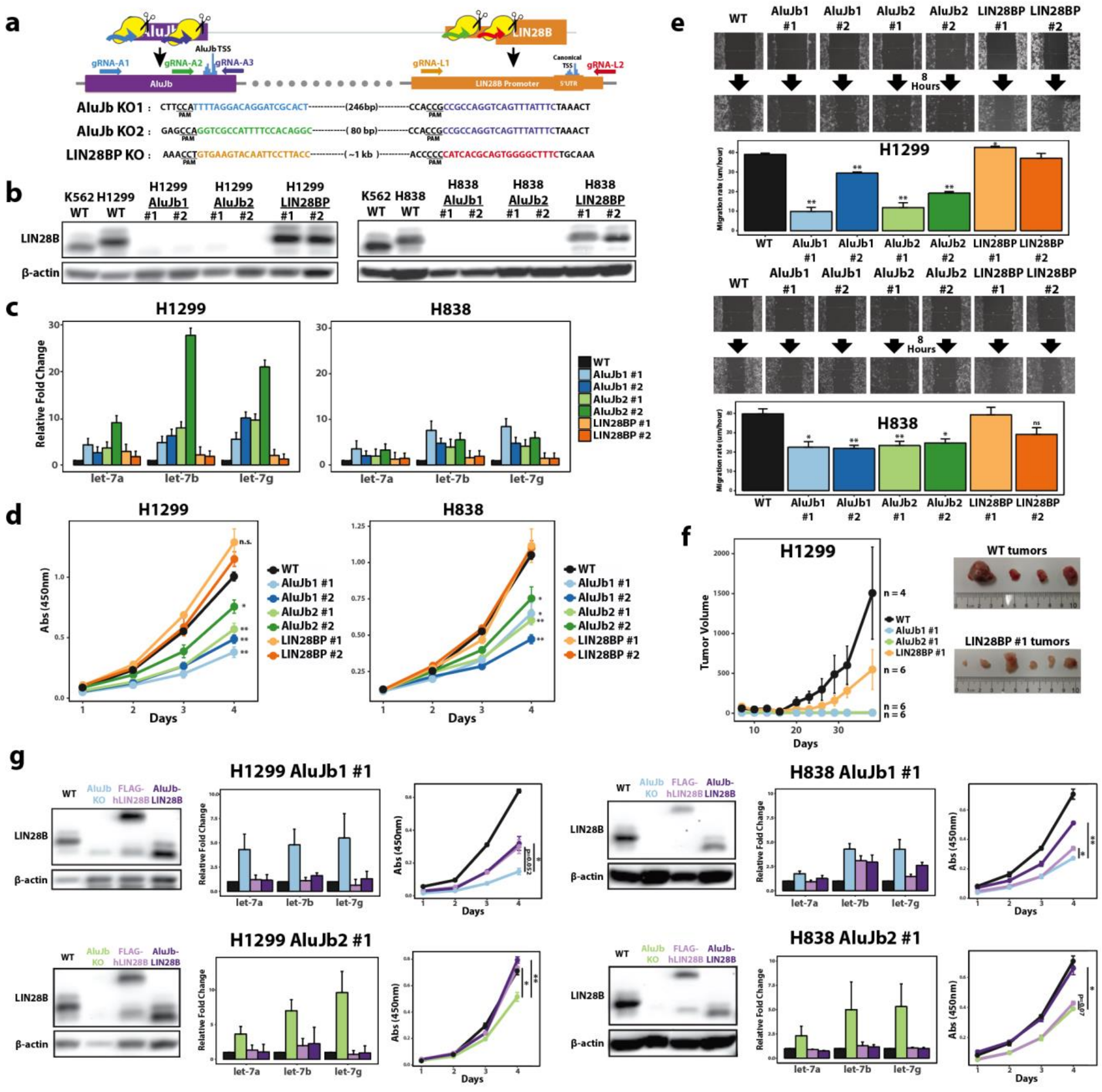


Fig. 3. AluJb drives LIN28B expression and contributes to oncogenesis in lung cancer cell lines. a, Schematic describing sgRNA locations and sequence targets within AluJb-P and LIN28BP. **b,** Cropped Western blot for LIN28B protein in H1299 (top) and H838 (bottom) CRISPR clones. This experiment was repeated twice with similar results. **c,** Relative let-7a, let-7b, and let-7g miRNA levels compared to WT in CRISPR-knockout clones of H1299 (n = 4 independent experiments) and H838 (n = 3 independent experiments) as measured by qPCR. **d,**

The effect of AluJb-P or LIN28BP deletion on cell growth rate as determined by CCK-8 assay in H1299 and H838 cells (n = 3 independent experiments). **e,** The effect of AluJb-P or LIN28BP deletion on cell migration in H1299 (top) and H838 (bottom) as measured by scratch migration assay (n = 3 independent experiments). **f,** Tumor growth of H1299 WT and H1299 CRISPR-knockout clones injected in nude mouse. Resected tumors of WT and LIN28BP #1 xenografts. **g,** Cropped Western blot (repeated twice with similar results) of re-expression of human FLAG-LIN28B or AluJb-LIN28B in AluJb KO clones and its effect on relative let-7 miRNA levels (number of independent experiments indicated in figure as n) and growth rate (n = 3 independent experiments). **d,e,g,** *P* values from CCK-8 growth assays and scratch migration assays were derived from comparing to WT with two-tailed Welch *t* test. All data are represented as means ± SE.

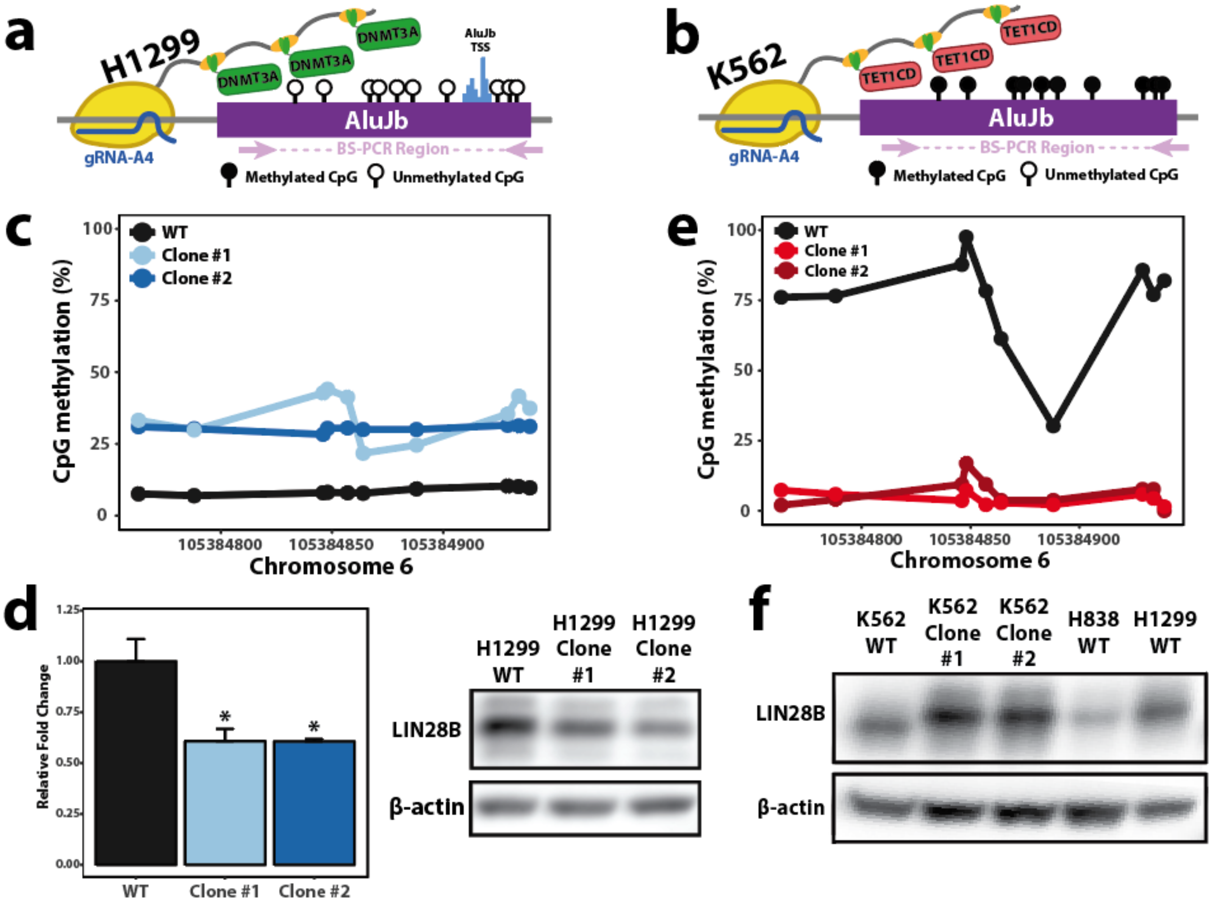


Fig. 4. Targeted DNA methylation dynamics uncover epigenetic control of AluJb promoter activity. **a** Schematics illustrating CRISPR-SunTag models for targeted de/methylation of AluJb. DNMT3A was recruited to AluJb loci in H1299 to increase methylation. **b,** TET1CD was recruited to AluJb in K562 to remove DNA methylation from the TE. **c,** Methylation levels of AluJb in WT and CRISPR-SunTag-DNMT3A clones of H1299 measured by BSPCR-seq. **d,** Relative abundance of LIN28B in H1299 CRISPR-SunTag-DNMT3A Clone #1 (left) and Clone #2 (right) compared to WT as measured by qPCR (n = 3 independent experiments) and cropped Western blot (repeated twice with similar results). *P* values were derived from two-tailed Welch *t* test. All data are represented as means ± SE. **e,** Methylation levels of AluJb in WT and CRISPR-SunTag-TET1CD clones of K562. **f,** Cropped Western blot (repeated twice with similar

results) illustrating the presence of larger LIN28B protein, similar size as AluJb-LIN28B in H1299 and H838, in K562 CRISPR-SunTag-TET1CD clones.

6.6 Methods:

Data download

All patient sample RNA-seq data analysis was done on the GDC Data Release 9.0 of TGCA data (10/24/17). Normal and tumor RNA-seq BAM files for the following 15 cancers were downloaded using the gdc-client version 1.3.0: bladder urothelial carincoma (BLCA), breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), low grade glioma (LGG), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian serous cystadenocarcinoma (OV), prostate adenocarcinoma (PRAD), skin cutaneous melanoma (SKCM), stomach adenocarcinoma (STAD), thyroid carcinoma (THCA), uterine corpus endometrial carcinoma (UCEC). In addition, normalized gene expression data (HTSeq-FPKM-Uq) and clinical metadata for all samples were downloaded using the gdc-client version 1.3.0. A total of 7,769 tumor samples and 625 matched-normal samples were used for analysis. 26 lung adenocarcinoma cancer cell line RNA-seq files were downloaded using sratools with the following accession: DRA001846. We included RNA-seq of the H838 lung cancer cell line, which has been previously generated in our laboratory and will be publicly available. GENCODE Version 25 was used as the transcript reference[288]. The GTF file of consensus transcripts was downloaded from https://www.gencodegenes.org/releases/25.html. Repeatmasker annotations were downloaded from the UCSC table browser for hg38[289,290]. FANTOM5 hg38-aligned peaks used for annotating the supplementary tables were downloaded from http://fantom.gsc.riken.jp/5/datafiles/reprocessed/hg38_latest/. 698 protein-coding oncogenes obtained from the ONGene database [262]. 4 genes from previous publications noting "onco-exaptation" were included in the list: IRF5, FABP7, SLCO1B3, IL33[247,258]. More details about the software used in our analysis can be found in the *Life Sciences Reporting Summary*.

Assembly and annotation of transcripts

BAM files were sorted and indexed and chr1-22, X, and Y were extracted. Stringtie version 1.3.3 was used to assemble the BAM files for all the RNA-seq samples (stringtie –m 100 –c 1)[291]. These transcripts were then annotated with features from GENCODE v25 with a custom script. Briefly, GENCODE v25 was first processed into a coordinate dictionary based on chromosome, start, and end location. Only the transcripts that were considered "appris_principal" were used so that alternative transcripts of the gene would not be excluded as potential TE-derived candidates. This set of principal transcripts as well as the Repeatmasker TE coordinates were used to annotate the transcripts generated from the stringtie assembly for each sample. The starting position of the transcript was annotated using the Repeatmasker table to find TE-derived transcription start sites. Then, the first exon of the transcript was annotated based on overlap with exonic or intronic features of GENCODE v25. If the exon overlapped both an exon and intron, then the exon was selected as the annotation for that element. Then, all subsequent exons in the transcript were annotated until one overlapped with a protein-coding gene exon; this exon of the protein-coding gene was selected as the "splice target" of that transcript. After all transcripts were annotated, candidate transcripts were selected based on the following criteria: the start site of the transcript being within a TE, the TE being intergenic or intronic, the starting exon not overlapping with exon 1 of the canonical gene, and the transcript splicing into a protein-coding gene. We further limited our analysis to only include a list of 702 oncogenes to increase likelihood of finding candidates with tumorigenic impact.

Generating a reference transcriptome including onco-exaptation candidates

Aggregating annotation data across all tumor and normal RNA-seq data sets, we constructed a list of unique onco-exaptation candidates based on the subfamily of the TE, the chromosomal coordinates of the TE, and the exon of the gene that the transcript spliced into. To remove potential assembly artifacts and genomic contamination, we removed candidates that had an average exon 1 length greater than the 99th percentile of all GENCODE v25 transcript first exons (2,588 bp). Furthermore, transcripts with first exons that retained an intron were also removed. Finally, we only included candidates that were present in at least 2 samples.

To further increase confidence of promoter activity, we interrogated all reads that uniquely mapped to each candidate TE. We subsequently annotated the mate pair of those reads to see if any overlapped directly with oncogene exons. For single-end reads, we annotated the portion of the read mapping outside the TE to see if it overlapped with an oncogene exon. First, we removed candidates that had zero files where there were at least 10 uniquely mapped reads that started within the TE. In addition, these events were required to have at least 1 sample with uniquely mapped paired-end reads where one of the pairs mapped to the TE and the other to the splice target of the candidate. For intronic onco-exaptation events, we also removed candidates that had evidence of exonization (there were reads mapping to both an upstream exon and the TE) in more than 15% of samples. Finally, candidates that were exclusively in single-end RNA-seq files were removed. The remaining candidate TE-derived transcripts were then merged with the reference GENCODE v25 annotation file using Cuffmerge to create a reference transcriptome inclusive of potential onco-exaptation events that have not been previously annotated.

Transcript-level quantification and candidate selection

To determine the contribution of candidates to overall gene expression, we used stringtie (-e -b) with the merged transcriptome as the reference. For each sample, we labeled a candidate as being present if it met the following criteria: (1) the transcript accounted for at least 25% of total gene expression, (2) there was at least one read covering the splice junction between the TE and the splice target (candidates without unique splice junctions were removed), and (3) the target gene had at least 1 FPKM expression. Next, we filtered for candidates that were highly tumor enriched (> 10× enrichment in the tumor samples) and present in at least 4 tumor samples. For the two cancers where there were no normal samples (OV and LGG), we removed candidates that had > 75% of their samples in these tumor types to avoid simply enriching for tissue-specific alternative promoters. This gave us a master list of 129 tumor-enriched onco-exaptation candidates involving 106 oncogenes. We then explored the abundance of these 129 candidates across the various cancer types to determine the prevalence of this phenomenon.

Open-reading-frame (ORF) prediction and FANTOM5 annotation

After determining the predicted transcript sequences of our candidates, we used CPC2 which predicted which candidates were coding or non-coding[292]. For coding transcripts, we subsequently used the start codon identified by CPC2 for the longest open reading frame and evaluated if it was in-frame or out-of-frame in relation to the canonical isoform. For FANTOM5 promoter annotation, we first filtered the FANTOM5 peaks in hg38 for samples that were not part of exposure or time-course experiments. Subsequently, we evaluated if there were any peaks that overlapped with the onco-exapted TE that were on the same strand as our candidate transcript.

Code Availability

All custom scripts are available from the authors upon request.

Cell culture methods

All cell lines were grown in a humidified incubator with 95% CO_2 at 37°C. H1299, H838, H727 and K562 cell lines were cultured in RPMI 1640 media (Gibco, 11875-085) supplemented with 10% fetal bovine serum (Corning, 35-011-CV) and 100U/ml penicillin-streptomycin (Gibco, 15140-122). HEK293T cell line was cultured in DMEM media (Gibco, 11965-084) supplemented with 10% fetal bovine serum and 100U/ml penicillin-streptomycin. Adherent cells were passaged at 70-90% confluency with 0.05% Trypsin-EDTA (Gibco, 25300-54).

Epigenome and transcriptome profiling

H1299 and K562 whole-genome bisulfite (WGBS)-seq and Cap Analysis of Gene Expression (CAGE)-seq were obtained from previously published results[270,293]. To generate WGBS-seq of H838 cell lines, we extracted genomic DNA with *Quick*-DNA Miniprep Kit (Zymo, D3024) and bisulfite converted 200 ng of DNA using EZ DNA Methylation-Direct kit (Zymo, D5020). For WGBS-seq, we processed the bisulfite-converted DNA with TruSeq DNA Methylation Kit (Illumina, 15066014). To evaluate DNA methylation of targeted regions, we performed bisulfite-PCR using ZymoTaq PreMix (Zymo, E2003) following manufacturer's protocol. Illumina adapters were ligated onto the BS-PCR product and amplified for sequencing. WGBS-seq and targeted BS-PCR libraries were sequenced on Illumina NextSeq and MiSeq platforms, respectively. The sequencing reads were aligned to hg19 genome with Bismark and CpG methylation values were calculated using bismark_methylation_extractor function[294].

To generate chromatin accessibility profiles for H1299 and H838, we followed the published Omni-ATAC-seq protocol[295]. Omni-ATAC-seq libraries were sequenced on Illumina NextSeq platform and reads were mapped to hg19 genome using bwa-mem[296].

Total RNA was extracted using TRIzol Reagent (ThermoFisher Scientific, 15596026) following manufacturer's protocol with few modifications. We performed an extra chloroform wash after transferring the aqueous phase. Furthermore, we added 5 µg of glycogen and 750 µl of isopropanol to the aqueous phase and incubated the solution overnight at -20°C to precipitate the RNA. Total RNA was treated with TURBO Dnase (ThermoFisher Scientific, AM2238). H838 RNA-seq library was generated using TruSeq RNA Library Prep Kit v2 (Illumina, RS-122-2001).

To annotate transcription start site locations, we generated CAGE-seq libraries using CAGE Preparation Kit (DNAFORM). In brief, 10 µg of total RNA was reverse transcribed using SuperScript III (ThermoFisher Scientific, 18080093) and 5' cap of mRNA was biotinylated. Biotinylated RNA/cDNA hybrid was purified using Dynabeads M-280 Streptavidine beads (ThermoFisher Scientific, 11205D) and processed to be sequenced on the Illumina sequencing platforms. For H727, we generated nanoCAGE-seq libraries[297]. In summary, polyA mRNA was extracted using Dynabeads™ mRNA DIRECT™ Purification Kit (ThermoFisher Scientific, 61011). The mRNA was enriched for 5' capped mRNA via Terminator exonuclease (Lucigen, TER51020) digestion. Then we followed standard nanoCAGE protocol to generate the cDNA via template-switching technology. H1299 and H838 CAGE-seq reads were aligned to the hg19 genome while H727 nanoCAGE-seq was aligned to hg38 genome with HISAT and processed using CAGEr package in R statistics[298]. All browser tracks are visualized with the WashU Epigenome Browser[299].

Quantitative PCR (qPCR) of let-7 miRNA and LIN28B

Let-7 miRNA levels were profiled using a published real-time PCR-based platform[300]. To summarize, 500 ng of total RNA was reverse transcribed using SuperScript IV First-Strand

Synthesis System (ThermoFisher Scientific, 18091050) with primers specific to let-7a, let-7b, let-7g and U6-snRNA transcripts. For *LIN28B, GAPDH* and *β-actin* qPCR, we processed 500 ng of total RNA with iScript Reverse Transcription Supermix (Bio-Rad, 1708840). Afterwards, we performed quantitative PCR on 1 µl of cDNA using PerfeCTa SYBR Green SuperMix (Quantabio, 95053-100). qPCR primers are listed in Supplemental Table 6. Results from qPCR were normalized to house-keeping gene to obtain ΔC_T and ΔC_T of samples were normalized to WT values to obtain $\Delta\Delta C_T$ values. Relative fold-change is calculated as $2^{-\Delta\Delta CT}$.

Western blot and antibodies

Whole-cell lysates for Western blots were extracted with Blue Loading Buffer Pack (Cell Signaling Technology, 7722S). Protein lysates were loaded into Novex 16% Tris-Glycine Mini Gels (Thermo Fisher Scientific, XP99165BOX) and separated by gel electrophoresis at 125V for 4 hours. LIN28B and β-actin were detected using an anti-LIN28B antibody (Cell Signaling Technology, #4196) and an anti-ACTB mouse monoclonal antibody (GenScript, A00702), respectively. More details about the antibodies can be found in the *Life Sciences Reporting Summary*. The Western blot was imaged with Thermo Scientific myECL Imager (Thermo Scientific, 62236).

Promoter and mutagenesis luciferase assay

Various promoter sequences derived from TEs were amplified and extended using primers listed in Supplementary Table 6 from H1299 genomic DNA. The minimal promoter sequence of pGL4.23 luciferase plasmid (Addgene, E8411) was removed with HindIII & NcoI restriction enzyme digest and the TE-derived promoters were cloned into pGL4.23 plasmid via Gibson Assembly following manufacture's protocol (New England Biolabs, E2661S). For mutagenesis assay, we mutated specific motifs within the AluJb promoter-luciferase vector with QuikChange Lightning Site-Directed Mutagenesis Kit (Agilent, 210518). We used the Neon transfection system (MPK5000) to deliver 400 ng of promoter-luciferase vector and 200 ng of pRL-TK *Renilla* vector (Addgene, E2241) into 3 × 10^4 H1299 cells, 3 × 10^4 H838 cells or 5 × 10^4 K562 cells. Luciferase levels were measured after 24 hours of incubation with Dual-Glo Luciferase Assay System (Promega, E2940).

CRISPR-Cas9-mediated Deletion of AluJb and LIN28B Promoter

We selected CRISPR-Cas9 sgRNAs by using both CRISPOR[301] and CRISPRscan[302] to identify sequences that have minimal off-targets and are highly efficient. We purchased pU6-(BbsI)_CBh-Cas9-T2A-BFP plasmid (Addgene, 64323) & pU6-(BbsI)_CBh-Cas9-T2A-mCherry plasmid (Addgene, 64324) as the CRISPR delivery vectors. For each sgRNA, we designed and annealed pairs of oligonucleotides that can be cloned into a BbsI-digested CRISPR vector through standard ligation techniques. We constructed BFP-CRISPR vectors that express sgRNAs targeting upstream and mCherry-CRISPR vectors that express sgRNAs targeting downstream of the region we want to delete. BFP-CRISPR vector and mCherry-CRISPR vector are co-transfected into H1299, H838 and K562 cells via Neon transfection system. After 24 hours of incubation, the transfected cells are analyzed by flow-cytometry (Beckman Coulter MoFlo) for BFP-positive and mCherry-positive fluorescence. We sorted double-positive fluorescent cells into 96-well plates for single-cell clone expansion. Genomic DNA from CRISPR clones was extracted using *Quick*-DNA Miniprep Kit for genotyping and validated with Sanger sequencing.

Cell proliferation assay

We seeded 2,500 wild-type cells or CRISPR-deletion clones in 100 µl of culture media into each well of 96-well plates. Ten µl of Cell Counting Kit-8 (Dojindo Molecular Technologies, CK04-01) were added to each well at appropriate time points. After 1 hour of incubation in humidified incubator with 95% CO_2 at 37°C, we recorded O.D. at 450 nm using BioTek Synergy H1 Hybrid Reader.

In vitro scratch migration assay

Wild-type cells and CRISPR-deletion clones were seeded into 6-well plates and grown to 100% confluency. We made straight scratches in middle of the well using 200-µl pipette tips and gently washed the well with culture media twice to remove free floating cells. Then, we imaged the scratch with Leica DMIL microscope and measured the width of the scratch using Leica Application Suite X software at the time of the scratch and 8 hours after the scratch.

Mouse xenograft experiment

All experiments were approved by the Institutional Animal Care and Use Committee of Washington University in St. Louis (Protocol #20170204) and conducted in accordance with the National Institutes of Health Guidelines for the Care and Use of Laboratory Animals. All experiments complied with the ethical regulations and considerations outlined in protocol. For H1299 xenografts, 3×10^6 wild-type cells or CRISPR-KO clones were collected and resuspended in 75 µl of chilled RPMI1640. Then, 75 µl of Matrigel Basement Membrane Matrix (Corning, 354234) was mixed into the cell solution and the sample was kept on ice until injection. The samples were injected subcutaneously into the right flank of four nude mice for WT and six nude mice for CRISPR KO clones (Jackson Lab, 002019, 4 weeks old homozygous NU/J females). Length (longer diameter) and width (shorter diameter) of the tumors were recorded and tumor volume was calculated by (Length x Width x Width)/2 equation.

CRISPR-SunTag vector construction

We obtained scFv-sfGFP-DNMT3A1 vector (Addgene, 102278) for targeted methylation vector. We purchased pHRdSV40-dCas9-10xGCN4_v4-P2A-BFP plasmid (Addgene, 60903) and pLKO5.sgRNA.EFS.tRFP657 plasmid (Addgene, 57824). For targeted demethylation, we replaced DNMT3A sequence with TET1 catalytic domain (CD) sequence, which was amplified from pPlatTET-gRNA2 plasmid (Addgene, 82559). Recent work revealed that dCas9-SunTag with 22aa linkers between GCN4 had higher demethylation efficiency[286]. In pHRdSV40-dCas9-10xGCN4_v4-P2A-BFP plasmid, we excised the 10xGCN4 sequence and cloned in GCN4-22aa sequence from pPlatTET-gRNA2 via Gibson Assembly. sgRNAs were cloned into pLKO5.sgRNA.EFS.tRFP657 plasmid.

Lentivirus production and transduction of CRISPR-SunTag vectors

HEK293T cells were seeded in 2 ml of DMEM complete media and grown to 50% confluency in 6-well plates. We co-transfected CRISPR-SunTag plasmids with pMD2.G and psPAX2 following polyethylenimine (PEI) transfection protocol. In brief, 6 µg of PEI and 2 µg of combined plasmids was added to 200 µl of Opti-MEM (ThermoFisher Scientific, 31985062) and incubated at room temperature for 30 minutes. The incubated PEI-vector mixture was added directly to HEK293T cells. After 48 hours, the viral supernatant was collected and filtered through 0.45-µm PES filter (Sigma-Aldrich, SLHV033RS). Then, polybrene (Sigma-Aldrich, TR-1003-G) was supplemented to the viral supernatant to a concentration of 5 µg/ml. The polybrene-viral supernatants of dCas9-SunTag-BFP, scFv-sfGFP-DNMT3A1/TET1CD and sgRNA.tRFP657 were added directly on top of H1299 and K562 cells in 6-well plates. After 2 days of incubation, the transduced cells were rinsed with PBS and analyzed by flow-cytometry (Beckman Coulter MoFlo) for BFP, GFP and farRFP657 fluorescence. Individual triple-positive fluorescent cells were sorted into 96-well plates and expanded. Once sufficiently expanded, the CRISPR-SunTag clones are resorted on the MoFlo for strong fluorescence and collected for downstream analysis of DNA methylation, gene expression and peptide expression.

Human LIN28B and AluJb-LIN28B rescue

We purchased pBABE-hLin28B plasmid (Addgene, 26358) that expresses FLAG-tagged human LIN28B protein[267]. We generated AluJb-LIN28B CDS from H1299 mRNA and cloned AluJb-LIN28B CDS in lieu of FLAG-hLIN28B sequence into the pBABE vector. We co-transfected pBABE plasmids with pMD2.G and pUMVC following polyethylenimine (PEI) transfection protocol into HEK293T cells. AluJb KO clones were transduced with viral supernatant supplemented with polybrene (5 µg/ml) for two days. Successfully infected cells were selected by 2 µg/ml puromycin (A.G. Scientific, P-1033-sol) treatment for 5 days before subsequent analysis.

Statistical analysis

Kaplan-Meier distributions between samples with or without candidate expression were compared using the logrank test. All statistics for in vitro experiments were performed using two-tailed Welch's *t* test. Enrichment for TE class was calculated with this formula: ((# of TE family onco-exapted / # of total TEs onco-exapted) / (# of total TE family / # of all TEs)).

6.7 Data Availability Statement

Datasets generated and analyzed in this study are available on Gene Expression Omnibus (GEO) under accession code GSE113946.

6.8 Supplementary Data Figures and Tables (tables will be available online)

Supplementary Table 1. Compiled oncogene and onco-exaptation list.

This table presents the comprehensive list of 702 oncogenes used in our analysis and their sources. Protein-coding oncogenes were procured from ONGene DB, a literature-based database of oncogenes (http://ongene.bioinfo-minzhao.org/). In addition, 4 oncogenes from previous publications focused on onco-exaptation were also included: IRF5, FABP7, SLCO1B3, and IL33.

Supplementary Table 2. All TE-derived alternative isoforms of oncogenes

For every alternative isoform beginning from a TE that was identified, the subfamily, family, and class of the TE as well as its coordinates in the genome are listed. The location of the TE is based on GENCODE v25 appris_principal labeled transcripts. If the TE is located in any of the introns, the label “intron” with the number of the intron is listed. If it is not found within the transcript coordinates of any gene, then it is listed as Intergenic. The Oncogene column has the symbol of the gene that the candidate splices into. The splice target is the exon number of the exon that is spliced into by the candidate. The Frame column has in-frame, out-of-frame, or noncoding based on the prediction from the coding potential calculator (CPC2). For all 15 cancers, there is a column for number of tumor samples and number of normal samples that each candidate is present in. Subsequently, the overall number of tumor samples is listed. The second to last column contains the fraction of total expression of the oncogene that the candidate accounts for on average, and the last column indicates whether the TE overlaps with an annotated FANTOM5 peak.

Supplementary Table 3. Tumor-enriched onco-exaptation events and their distribution across 15 TCGA cancers.

After filtering for 10x tumor-enrichment and presence in >=4 tumor samples, there were 129 onco-exaptation candidates found which were subsequently used in our analysis of this mechanism. These candidates have been detailed in this table. The format is the same as Supplementary Table 2, except the normal sample columns have been removed.

Supplementary Table 4. Multiple TE-derived oncogene activation.

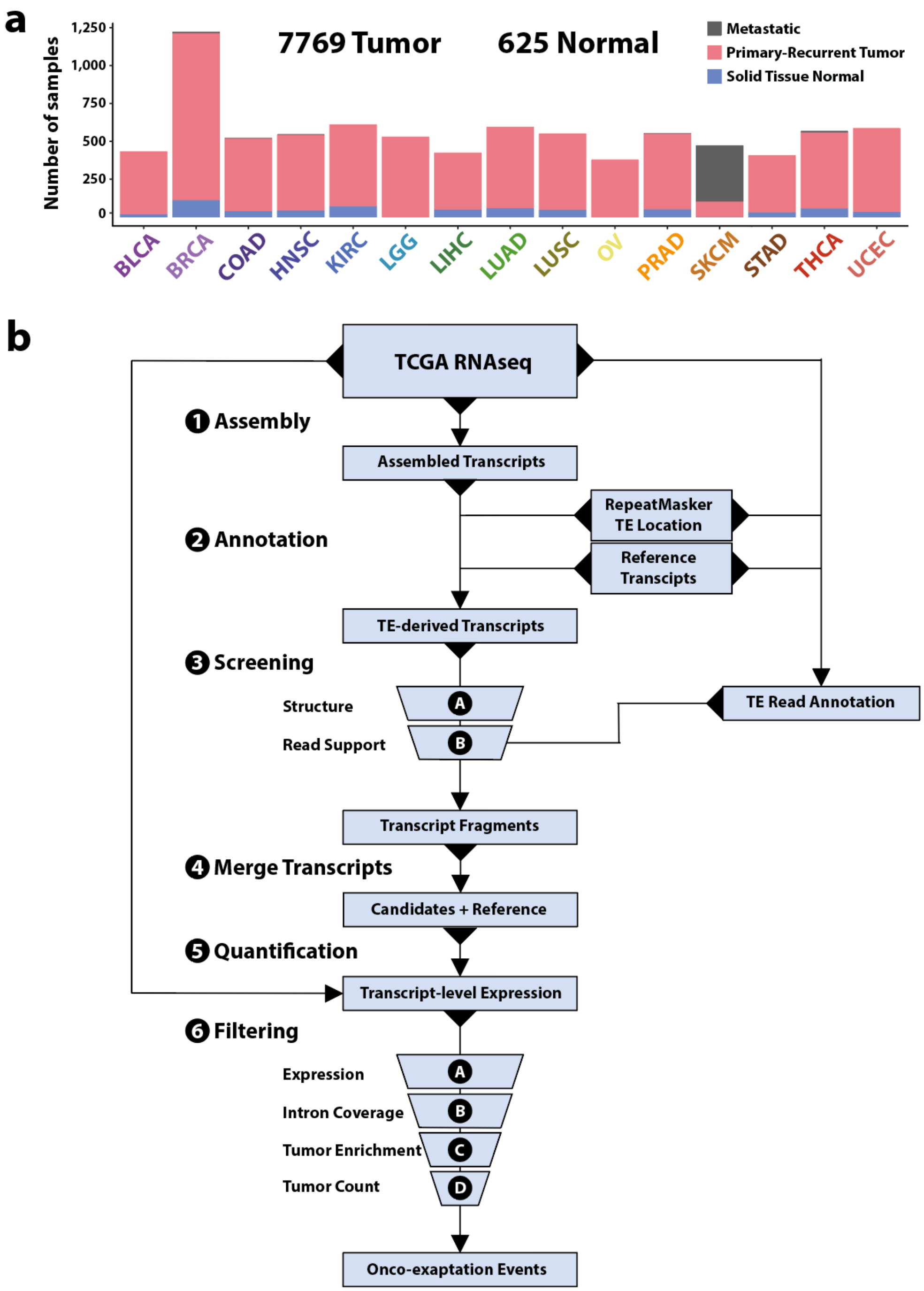
Oncogenes that had multiple in-frame isoforms originating from different transposable elements are listed in this table.

Supplementary Table 5. Top candidates in lung adenocarcinoma cell lines.

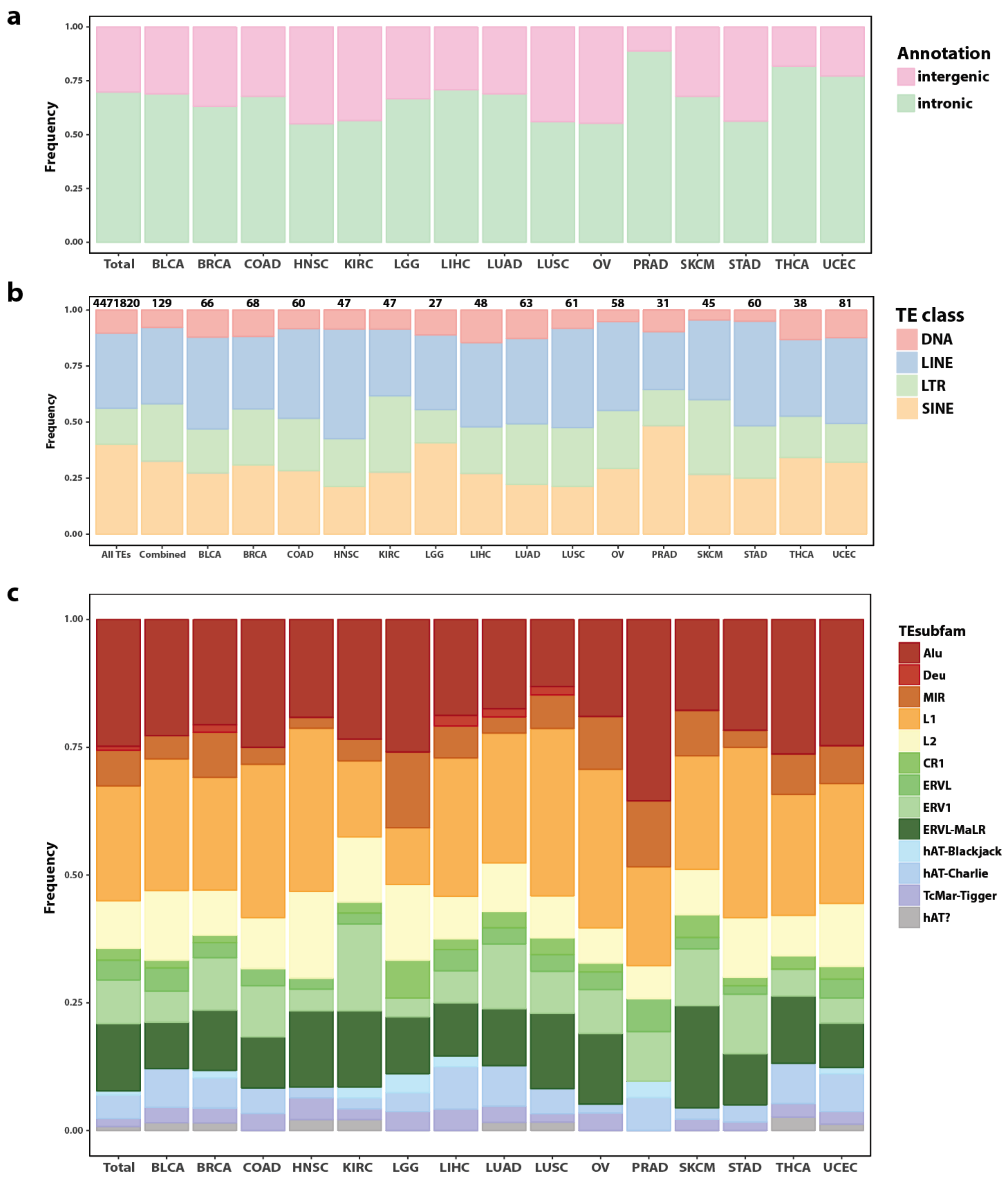
27 lung adenocarcinoma cell line RNA-seq datasets were analyzed for the presence of the most robust candidates. 26 of the cell lines were downloaded from a previous study, and one of the cell lines (H838) came from previously generated data from our own lab. 5 of the top 10 candidates were found in these cell lines. The tpm value of the transcript of each candidate is listed for each of the 27 cell line datasets.

Supplementary Table 6. Primer sequences.

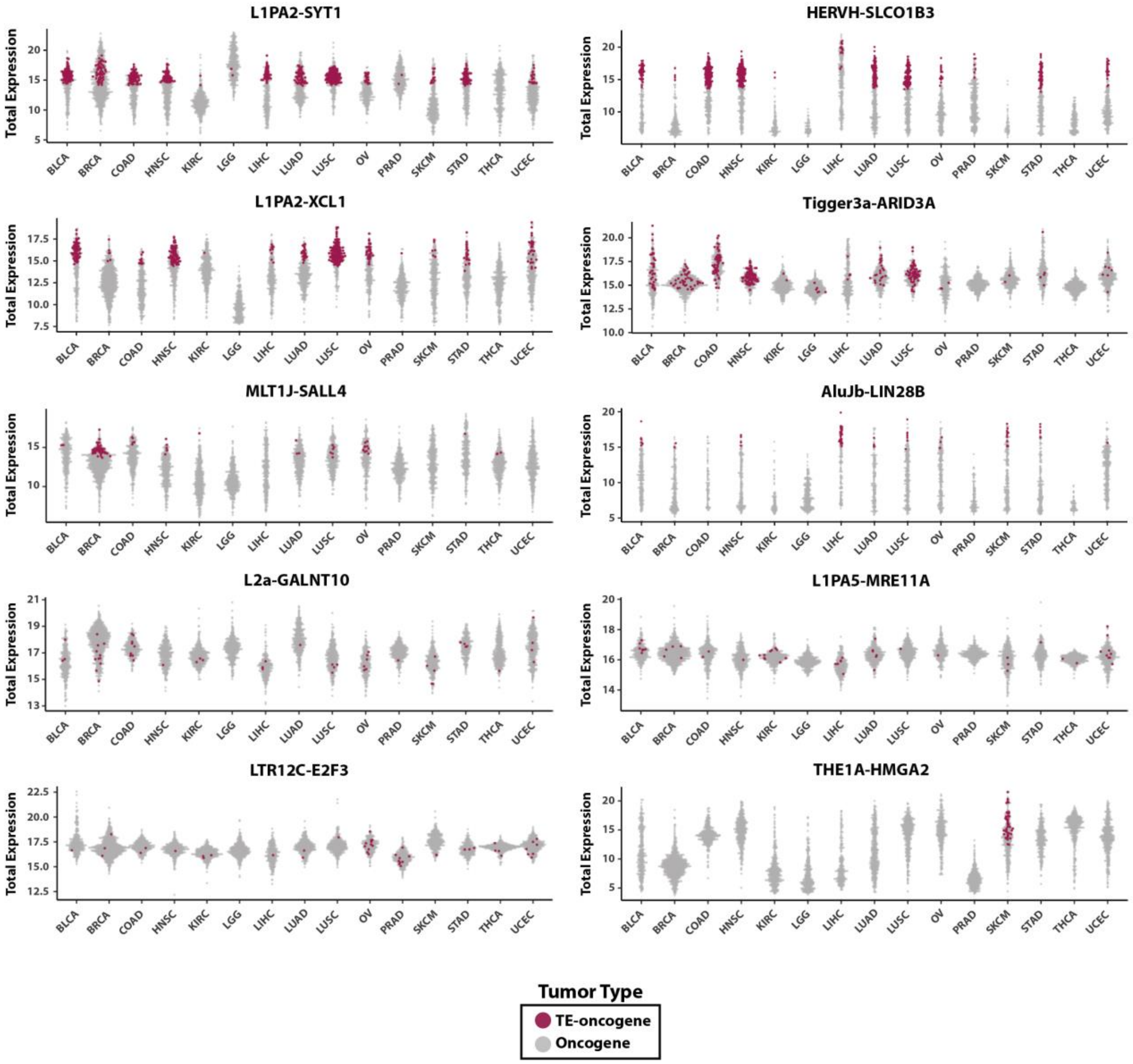
Sequences of primers used in experiments presented in the paper.



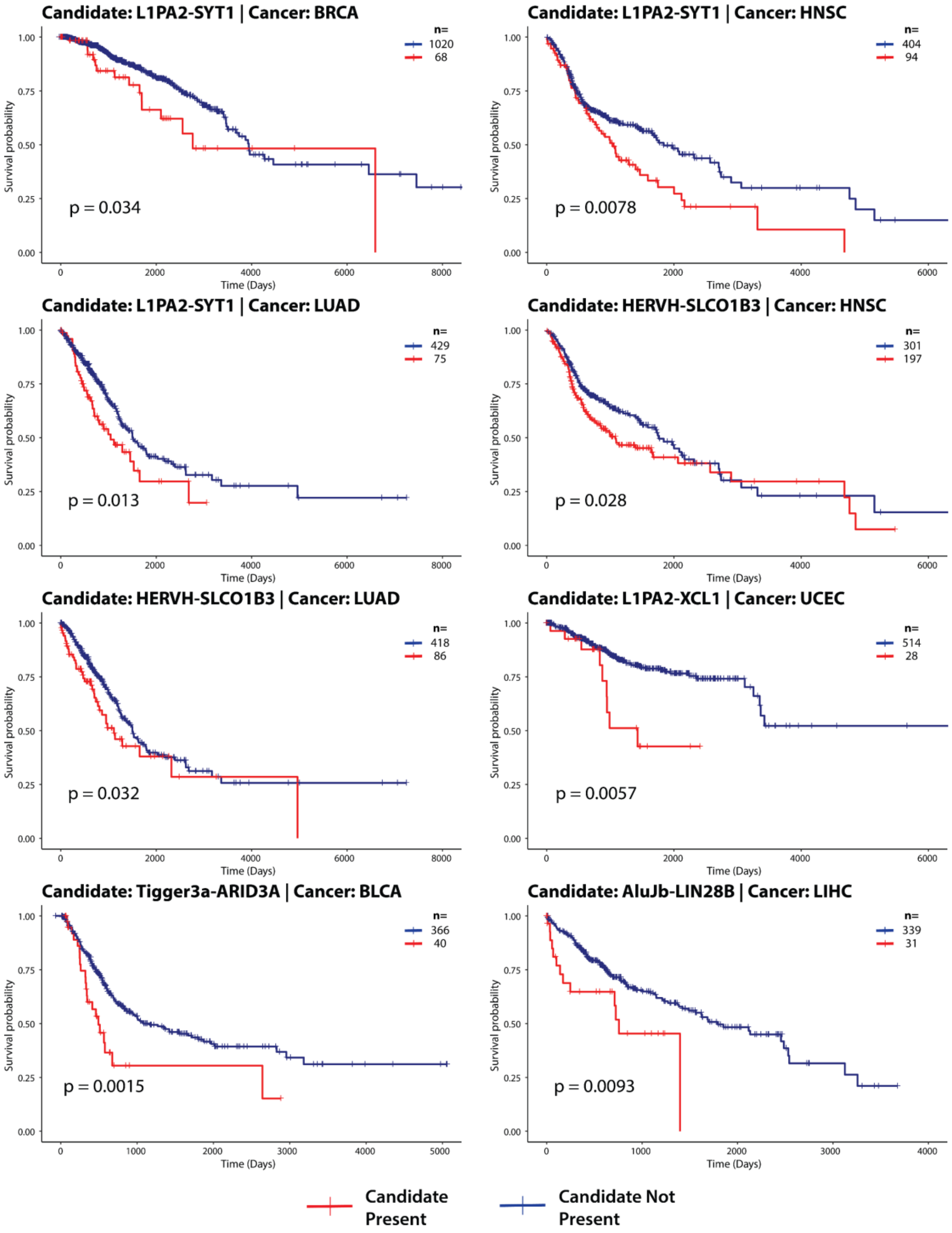
Supplementary Figure 1. RNA-seq computational pipeline detects numerous TE onco-exaptation events in 15 cancer types from TCGA. a, Number of cases from various cancer types that were analyzed**. b,** Schematic of the computational pipeline describing how RNA-seq from TCGA was processed to identify onco-exaptation candidates.



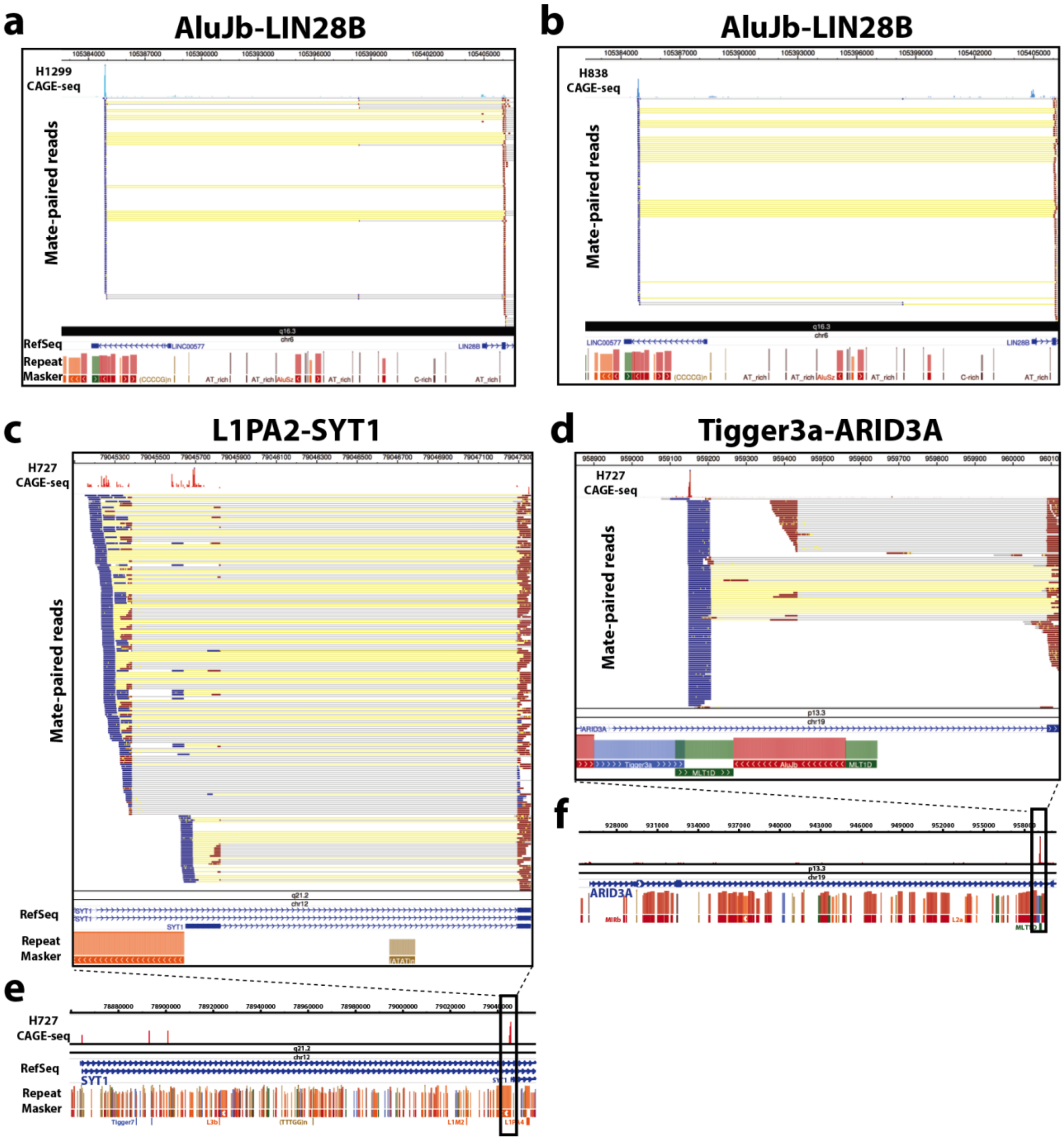
Supplementary Figure 2. TE locations and annotations that are implicated in onco-exaptation events. a, The genomic locations of TEs that act as cryptic promoters for oncogenes across different cancer types. **b,** Distribution of TE classes across cancer types. Total number of unique TEs that contribute to onco-exaptation events are labeled on top. **c,** Distribution of each TE family that contributed to onco-exaptation across 15 cancer types.



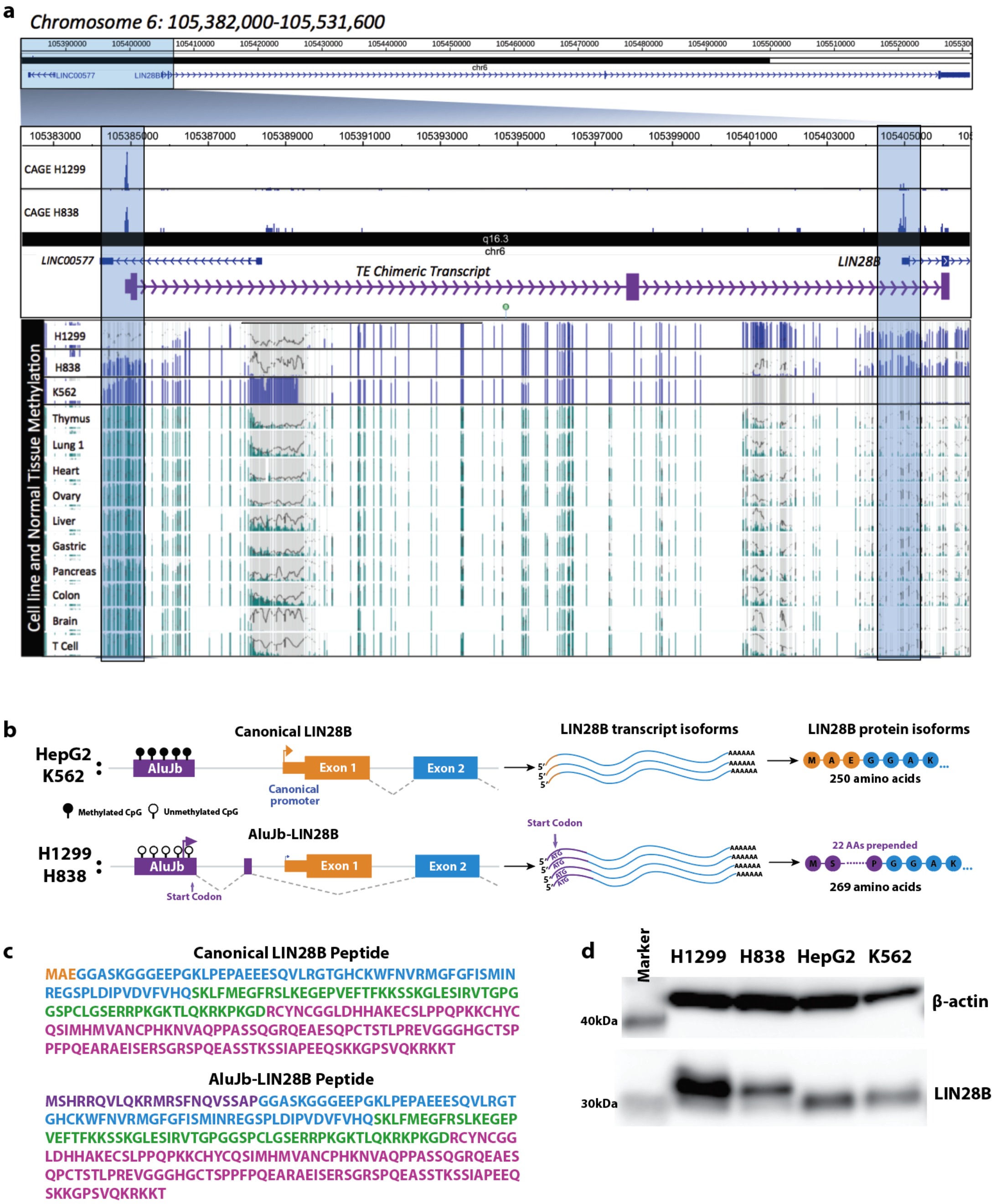
Supplementary Figure 3. Oncogene expression profiles of the top 10 onco-exaptation candidates across cancer types. Oncogene expression of each tumor with and without an onco-exaptation event. Each grey dot represents a tumor while the red dots reveal whether the tumor is predicted to have an onco-exaptation event. Total expression is represented as $\log_2$(FPKM-UQ) provided by TCGA GDC (https://portal.gdc.cancer.gov/).



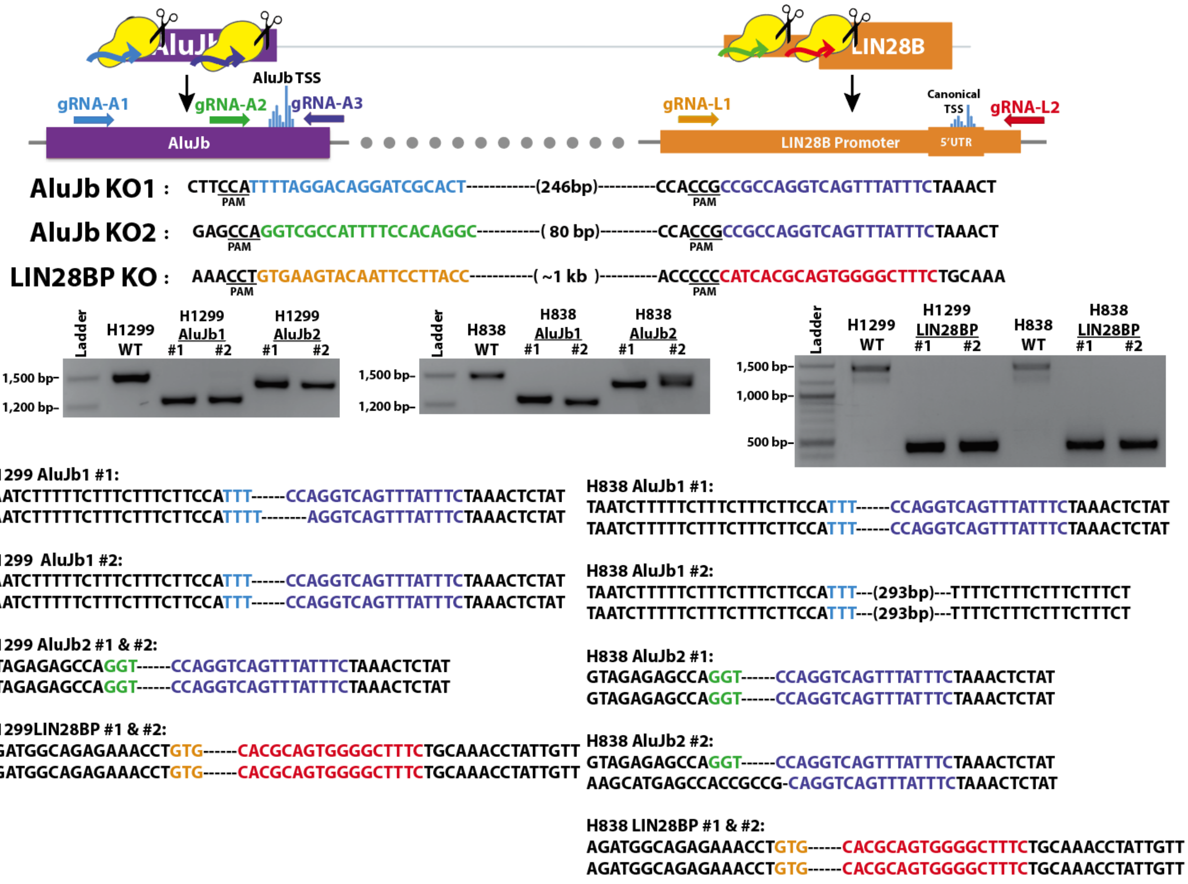
Supplementary Figure 4. Overall survival impact of oncoexaptation candidates. Kaplan-Meier Curves for the 8 examples of where a top 10 candidate was significantly prognostic in a cancer ($p<0.05$) based on log-rank statistical test (two-sided). The red line in each graph represents patients where the candidate was found to be present, and in blue line represents all the patients where the candidate was not detected. All were found to negatively impact overall survival. The number of biologically independent patients is listed within each plot.



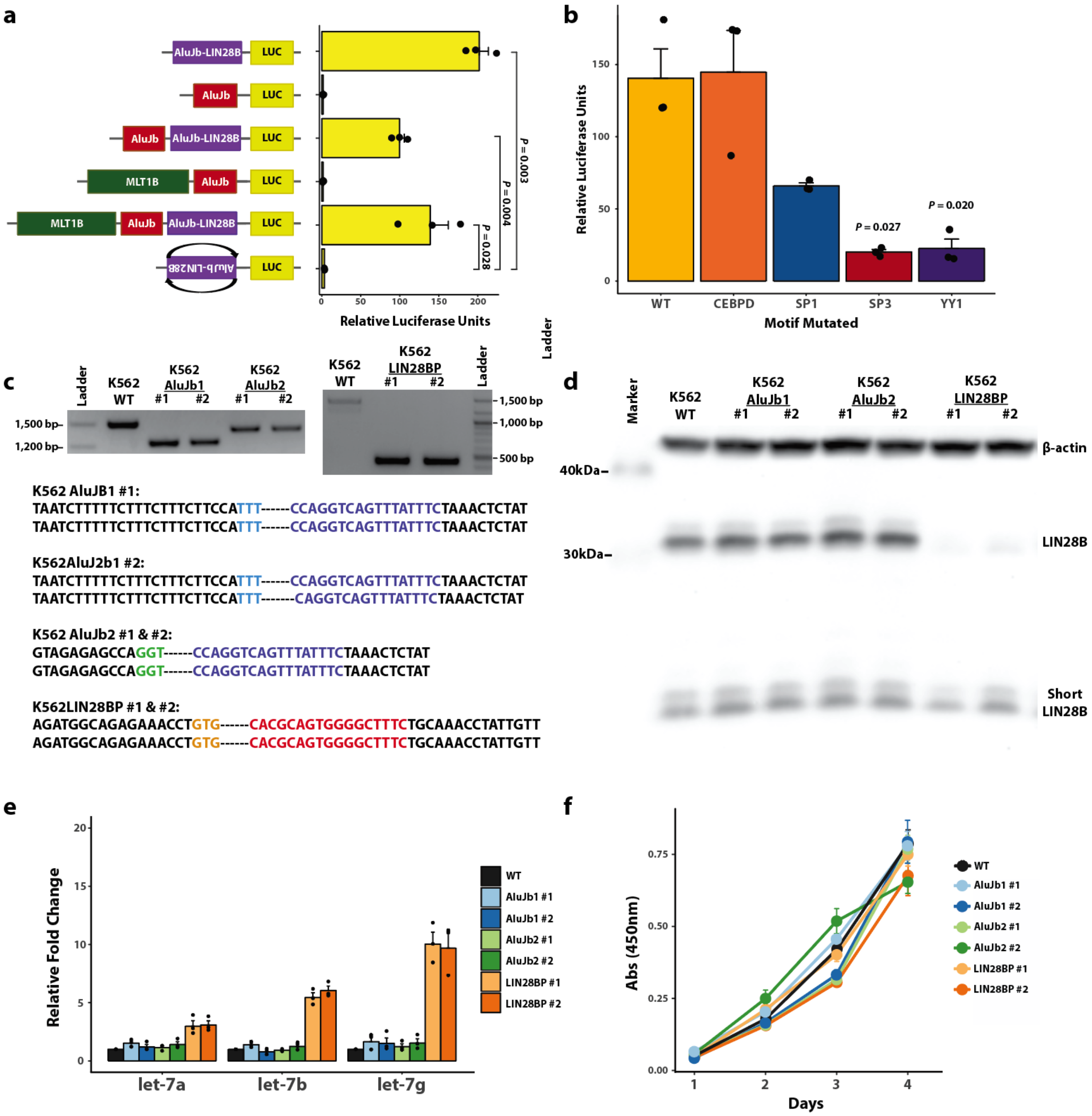
Supplementary Figure 5. Transcription start site characterization of onco-exaptation candidates in H727 lung cancer cell line. a, WashU Epigenome browser view of CAGE-seq and mate-paired reads where the forward read initiates from AluJb and the reverse read ends in the gene body of LIN28B in H1299 and **b,** H838. **c,** WashU Epigenome browser view of H727 CAGE-seq and mate-paired reads that where the forward read from L1PA2 and the reverse read ends in the gene body of SYT1. **d,** WashU Epigenome browser view of H727 CAGE-seq and mate-paired reads where the forward read initiates from Tigger3a/MLT1D and the reverse read ends in the gene body of ARID3A. **e,** WashU Epigenome browser view of H727 CAGE-seq over SYT1 promoters. **f,** WashU Epigenome browser view of H727 CAGE-seq over ARID3A promoter.



Supplementary Figure 6. The AluJb TE is methylated in somatic tissue and is epigenetically dysregulated in cancer. a, DNA methylome profiles of multiple somatic tissues from the Roadmap Epigenomics Project (http://www.roadmapepigenomics.org/) are displayed on the WashU Epigenome Browser (http://epigenomegateway.wustl.edu/browser/). **b,** Schematics showing DNA methylation levels of AluJb in different cancer cell lines. An alternative start codon present in AluJb generates a chimeric LIN28B peptide that lacks 3 amino acids contributed by exon 1, but has 22 novel amino acids prepended. **c,** Predicted amino acid sequence of AluJb-LIN28B protein. **d,** Cropped Western blot (repeated twice with similar results) representing the size difference between the AluJb-LIN28B protein and canonical LIN28B protein.


AluJb
LIN28B
AluJb TSS
gRNA-A1
gRNA-A2
gRNA-A3
gRNA-L1
Canonical TSS
gRNA-L2
AluJb
LIN28B Promoter
5'UTR
AluJb KO1 : CTTCCATTTTAGGACAGGATCGCACT------------(246bp)----------CCACCGCCGCCAGGTCAGTTTATTTCTAAACT
PAM
PAM
AluJb KO2 : GAGCCAGGTCGCCATTTTCCACAGGC-----------(80 bp)----------CCACCGCCGCCAGGTCAGTTTATTTCTAAACT
PAM
PAM
LIN28BP KO : AAACCTGTGAAGTACAATTCCTTACC-----------(~1 kb)----------ACCCCCCATCACGCAGTGGGGCTTTCTGCAAA
PAM
PAM
Ladder
H1299 WT
H1299 AluJb1 #1 #2
H1299 AluJb2 #1 #2
1,500 bp
1,200 bp
Ladder
H838 WT
H838 AluJb1 #1 #2
H838 AluJb2 #1 #2
1,500 bp
1,200 bp
Ladder
H1299 WT
H1299 LIN28BP #1 #2
H838 WT
H838 LIN28BP #1 #2
1,500 bp
1,000 bp
500 bp
H1299 AluJb1 #1:
TAATCTTTTTCTTTCTTTCTTCCATTT------CCAGGTCAGTTTATTTCTAAACTCTAT
TAATCTTTTTCTTTCTTTCTTCCATTTT--------AGGTCAGTTTATTTCTAAACTCTAT
H1299 AluJb1 #2:
TAATCTTTTTCTTTCTTTCTTCCATTT------CCAGGTCAGTTTATTTCTAAACTCTAT
TAATCTTTTTCTTTCTTTCTTCCATTT------CCAGGTCAGTTTATTTCTAAACTCTAT
H1299 AluJb2 #1 & #2:
GTAGAGAGCCAGGT------CCAGGTCAGTTTATTTCTAAACTCTAT
GTAGAGAGCCAGGT------CCAGGTCAGTTTATTTCTAAACTCTAT
H1299LIN28BP #1 & #2:
AGATGGCAGAGAAACCTGTG------CACGCAGTGGGGCTTTCTGCAAACCTATTGTT
AGATGGCAGAGAAACCTGTG------CACGCAGTGGGGCTTTCTGCAAACCTATTGTT
H838 AluJb1 #1:
TAATCTTTTTCTTTCTTTCTTCCATTT------CCAGGTCAGTTTATTTCTAAACTCTAT
TAATCTTTTTCTTTCTTTCTTCCATTT------CCAGGTCAGTTTATTTCTAAACTCTAT
H838 AluJb1 #2:
TAATCTTTTTCTTTCTTTCTTCCATTT---(293bp)---TTTTCTTTCTTTCTTTCT
TAATCTTTTTCTTTCTTTCTTCCATTT---(293bp)---TTTTCTTTCTTTCTTTCT
H838 AluJb2 #1:
GTAGAGAGCCAGGT------CCAGGTCAGTTTATTTCTAAACTCTAT
GTAGAGAGCCAGGT------CCAGGTCAGTTTATTTCTAAACTCTAT
H838 AluJb2 #2:
GTAGAGAGCCAGGT------CCAGGTCAGTTTATTTCTAAACTCTAT
AAGCATGAGCCACCGCCG-CAGGTCAGTTTATTTCTAAACTCTAT
H838 LIN28BP #1 & #2:
AGATGGCAGAGAAACCTGTG------CACGCAGTGGGGCTTTCTGCAAACCTATTGTT
AGATGGCAGAGAAACCTGTG------CACGCAGTGGGGCTTTCTGCAAACCTATTGTT

Supplementary Figure 7. Genotypes of AluJb-P and LIN28BP CRISPR-deleted clones. The CRISPR-mediated genetic breaks are illustrated for H1299 (left) and H838 (right) CRISPR clones. gRNA sequences are illustrated with various colors. AluJb1 KO clones were generated by deleted genomic region between gRNA-A1 and gRNA-A3. AluJb2 KO clones were generated by deleted genomic region between gRNA-A2 and gRNA-A3. LIN28BP KO clones were generated by deleted genomic region between gRNA-L1 and gRNA-L2. Two independent clones were profiled for each set of CRISPR KOs. The genotyping gel results were replicated twice in independent experiments.



K562 AluJB1 #1:
TAATCTTTTTCTTTCTTTCTTCCATTT------CCAGGTCAGTTTATTTCTAAACTCTAT
TAATCTTTTTCTTTCTTTCTTCCATTT------CCAGGTCAGTTTATTTCTAAACTCTAT

K562AluJ2b1 #2:
TAATCTTTTTCTTTCTTTCTTCCATTT------CCAGGTCAGTTTATTTCTAAACTCTAT
TAATCTTTTTCTTTCTTTCTTCCATTT-------CAGGTCAGTTTATTTCTAAACTCTAT

K562 AluJb2 #1 & #2:
GTAGAGAGCCAGGT------CCAGGTCAGTTTATTTCTAAACTCTAT
GTAGAGAGCCAGGT------CCAGGTCAGTTTATTTCTAAACTCTAT

K562LIN28BP #1 & #2:
AGATGGCAGAGAAACCTGTG------CACGCAGTGGGGCTTTCTGCAAACCTATTGTT
AGATGGCAGAGAAACCTGTG------CACGCAGTGGGGCTTTCTGCAAACCTATTGTT

Supplementary Figure 8. K562 control experiments for AluJb-LIN28B candidate validation. a, Promoter luciferase (n = 3 independent experiments) results illustrating transcriptional activity of various TE arrangements in K562. Welch's t-test was performed against reverse complement AluJb sequence (limited activity). **b,** Luciferase assays (n = 3 independent experiments) for mutagenized transcription factor motifs in K562. Welch's t-test against wild-type AluJb-P sequence. **c,** Genotypes of K562 CRISPR KO clones for AluJb-P and LIN28BP deletions. Genotype check was repeated twice with similar results. **d,** Cropped Western blot for LIN28B in K562 CRISPR KO clones. K562 also expresses a smaller isoform of LIN28B that is not present in H1299 and H838. This experiment was repeated twice with similar results. **e,** Relative let-7a, let-7b and let-7g miRNA levels compared to wild-type in CRISPR-knockout clones of K562 as measured by qPCR (n = 3 independent experiments). **f,** CCK-8 growth assay measuring cell growth rate of K562 WT and CRISPR clones (n = 3 independent experiments). **a,b,** P-values were calculated using two-tailed Welch t-test. **a,b,e,f,** All data are represented as means ± SE.

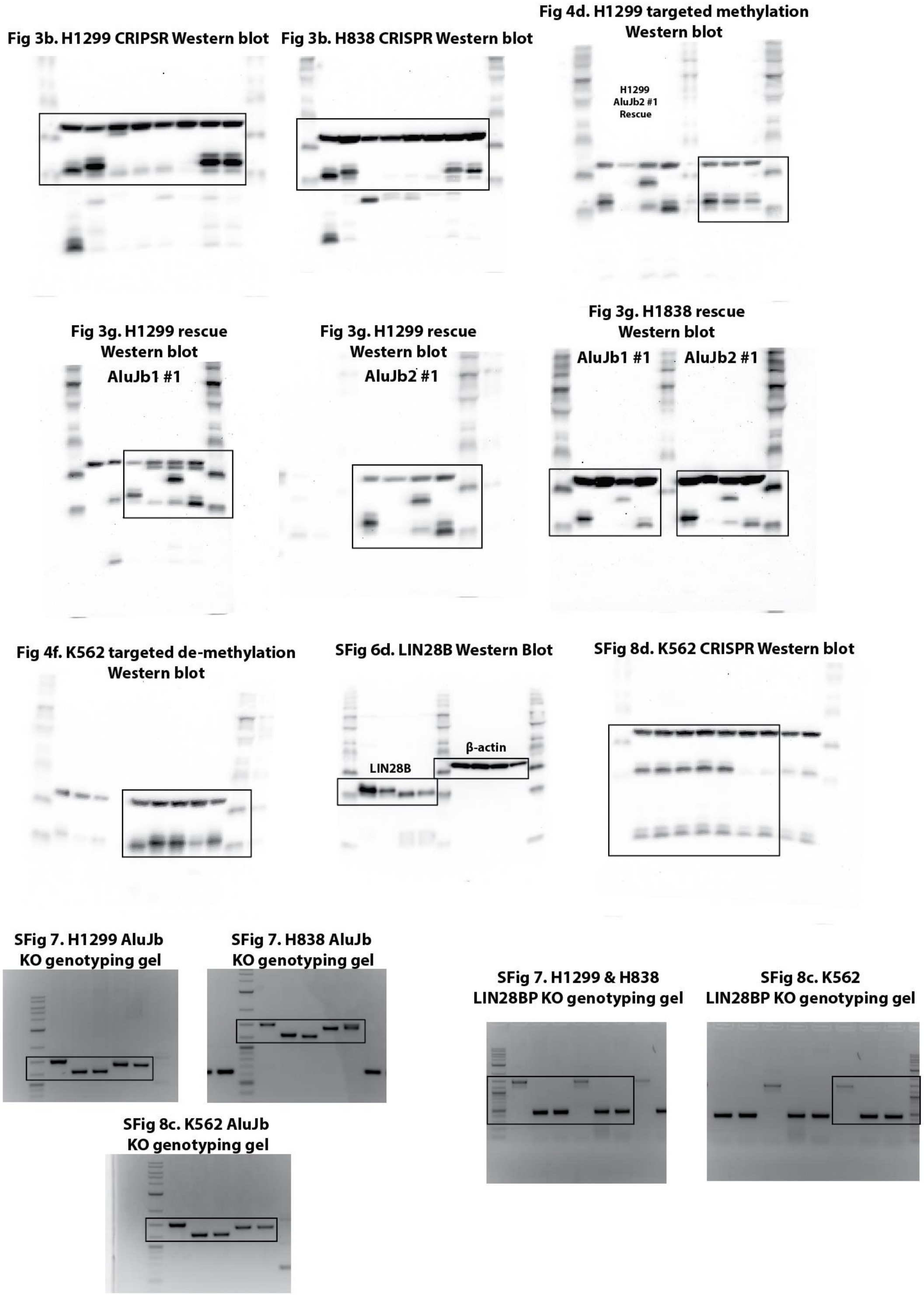


Fig 3b. H1299 CRIPSR Western blot
Fig 3b. H838 CRISPR Western blot
Fig 4d. H1299 targeted methylation Western blot
H1299 AluJb2 #1 Rescue
Fig 3g. H1299 rescue Western blot
AluJb1 #1
Fig 3g. H1299 rescue Western blot
AluJb2 #1
Fig 3g. H1838 rescue Western blot
AluJb1 #1
AluJb2 #1
Fig 4f. K562 targeted de-methylation Western blot
SFig 6d. LIN28B Western Blot
LIN28B
β-actin
SFig 8d. K562 CRISPR Western blot
SFig 7. H1299 AluJb KO genotyping gel
SFig 7. H838 AluJb KO genotyping gel
SFig 7. H1299 & H838 LIN28BP KO genotyping gel
SFig 8c. K562 LIN28BP KO genotyping gel
SFig 8c. K562 AluJb KO genotyping gel

Supplementary Figure 9. Uncropped Western blots and gel images. Black boxes denote cropped images that are presented in the manuscript.

Chapter 7: Prevalent tumor-specific transposable element-derived antigen detection in human cancers.

Nakul M. Shah[1,2*], Hyo Sik Jang[1,2*], Juheon Maeng[1,2], Changxu Fan[1,2], Noah L. Basri[1,2], Jiaxin Ge[1,2], Benjamin Katz[1,2], Daofeng Li[1,2], Xiaoyun Xing[1,2], Ting Wang[1,2]

[1]Department of Genetics, Washington University School of Medicine, St. Louis, Missouri, USA

[2]The Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, Missouri, USA

[*]Equal contribution

Author Contributions: N.M.S., H.S.J., and T.W. conceived and implemented the study; N.M.S., J.M, C.F, B.K, D.L. and T.W. contributed to the computational analysis; H.S.J., J.M. and X.X. generated transcriptomic profiles of cell lines; H.S.J. and J.G designed and implemented CRISPR experiments; H.S.J. performed Western blot; H.S.J. and N.L.B maintained and harvested cell lines for LC-MS ; H.S.J. performed the MHC-pulldown LC-MS; and the manuscript was prepared and revised by H.S.J., N.M.S. and T.W. with input from all authors.

Disclaimer: Various parts of the chapter written here will also be published in N.M.S. Master thesis dissertation.

This manuscript is currently in preparation for submission.

7.1 Abstract

Pervasive chimeric transcripts from transposable element (TE) exaptation events can provide immunogenic targets that we can exploit with antigen-based immunotherapies. We analyzed 33 TCGA cancer types and 675 cancer cell lines and performed comprehensively profiled all TE-gene fusion events. This effort discovered 2,461 TE-exaptation events that were tumor-specific and present in nearly all tumor samples. Computational prediction of coding potential and reading frames of these TE-chimeric reads discovered potential atypical tumor-specific TE-derived antigen (TS-TEA) candidates. Reexamining mass spectrometry data of published MHC-pulldown peptidomes from cell lines revealed that multiple novel antigens from these transcripts could be detected on MHC molecules. Furthermore, we perform our own MHC-pulldown and report discovery of one out-of-frame TS-TEA candidate in a cancer cell line. These preliminary results indicate that careful analysis of TE dysregulation in cancer can ultimately lead to novel protein products that have the potential to be targeted with immunotherapy. Furthermore, we highlight tumor-specific membrane proteins transcribed from TE-exapted promoters that potentially expose a novel epitope, which can be targets of CAR-T or antibody-based therapy. In conclusion, we showcase the high prevalence of TE-derived promoter activation in cancer and suggest multiple avenues by which this phenomenon can be targeted therapeutically.

7.2 Introduction

In this chapter, we hypothesize that epigenetic dysregulation in cancer leads to the activation of cryptic regulatory elements encoded by TEs, some of which may form cancer-specific products such as tumor-specific antigens or immunoreactive proteins. TEs make up 50% of the genome, and they locate in and around almost all genes. In our previous work, we found that TEs are widely used as cryptic promoters that drive abnormal gene expression in cancer cells[39]. Importantly, many well-known oncogenes are upregulated in this manner, a phenomenon that has been termed “onco-exaptation”. Transcripts initiated from epigenetically deregulated TEs promiscuously readthrough and splice into nearby protein coding genes, resulting in a chimeric RNA product that joins TE sequence with gene sequence. When such a chimeric RNA is translated, sometimes the TE sequence modifies the N-terminus of the protein by adding a few amino acids, as we illustrated in Jang et. al.[39] (**Fig. 1a**). These newly discovered and cancer-specific protein sequences represent a completely novel way to identify antigens.

TCGA has generated a rich resource of exome sequencing, RNA sequencing, and methylation data sets for hundreds of glioblastoma patients[303]. Recent studies have thoroughly characterized the landscape of cancer drivers, somatic mutations, and chromosomal rearrangements of these tumors and correlated these changes with immune phenotype[53,54,303,304]. However, these studies largely ignored the potential impact of transcriptional changes caused by epigenetic dysregulation, such as the activation of TE-derived promoters. Thus, we performed a thorough characterization of TE-derived alternative transcripts and assessed their potential as a significant source of tumor-specific antigens across all tumor types. Furthermore, we provide the first comprehensive TE-transcript map of normal tissues by processing TCGA, GTEx and FANTOM5 databases. This effort will assist in identifying candidate tumor-specific TE-derived antigens (TS-TEAs) that are not present in normal tissues to minimize possible autoimmune side-effects potentiated by vaccination strategies[305–307].

Though short-read RNA-sequencing technology is not optimized for accurate detection of 5’ ends of transcripts, we have developed a bioinformatics pipeline that can identify alternative TE-derived transcripts with a specificity >95% and assess their potential as a source of tumor-specific antigens (**Fig. 1b**). We use our pipeline to define the landscape of high-confidence TE-gene fusion events in cancer utilizing transcriptomic data of over 10,000 tumor samples across

33 tumor types available from TCGA. Then, candidates are filtered for tumor specificity. In addition to the matched-normal samples available from TCGA, we will also incorporate 15,000+ samples representing 54 tissue types from The Genotype-Tissue Expression (GTEx) project and promoter expression data from 1,800+ normal tissue samples from the FANTOM5 project. We can then stringently filter our candidates and confidently identify the candidates that are tumor-specific. Subsequently, we assess the coding potential of transcripts using *CPC2* and Kozak predictions and perform *in silico* translation to identify novel chimeric or out-of-frame isoforms[292]. Finally, with the patient transcriptome data, we determine their HLA class I and class II alleles and expression using *seq2HLA*[308]. Using *NetMHCpan-4.0* and *NetMHCIIpan-3.2,* to predict which novel peptide sequences from these isoforms can be presented on patient MHC molecules[51,309].

7.3 Results

The discovery of transposable element-derived transcripts also opens the door towards novel targeted therapies. In fact, recent publications have indicated that focusing on the non-coding regions of the genome and transposable element expression could significantly enhance the pool of tumor-specific antigens[310]. Chimeric peptides created through onco-exaptation or exaptation of other genes could be an additional, underexplored source of antigens to enhance immunotherapy or other targeted therapies[311,312]. For example, the AluJb-LIN28B candidate has the addition of 22 amino acids (AA) that are present in multiple tumor samples, absent in somatic tissues, and part of a highly expressed isoform (**Fig. 1a**); these 3 characteristics make it ideal as a target for immunotherapy[43,313]. Thus, we performed a comprehensive screen of TE-gene fusion events across 10,365 tumors and 675 cancer cell lines and evaluated them for tumor-specificity, antigenic potential, and MHC presentation.

We modified our previously developed computational framework to universally screen for TE-derived cryptic promoters across all genes (**Fig. 1b**). First, we screened RNA-sequencing data available across 33 tumor types from TCGA for these events and filtered them based on expression, intron structure, and tumor specificity (**Methods**). Considering the reported toxicity in off-target tissues with even basal low levels of target expression during immunotherapy trials[314,315], we further filtered candidates for expression in other adult tissues. First, we used the Snaptron[316] splice junction expression profiling data to remove candidates with significant expression of their unique splice junctions in any of the 31 adult tissues across 9,662 samples profiled in the Genotype-Tissue Expression (GTEx) project, excluding the testis (**Fig. 1c**). In addition, we removed candidates that had significant adult tissue expression according to the FANTOM5[317] promoter database (**Fig. 1d**). We arrived a final list of 2,461 tumor-specific TE-gene fusion events across all cancer types (**Fig. 1e**). Surprisingly, we found that nearly all tumors (97.9%) had at least one event, and the median level of events ranged from 2 in THCA to 57 in TGCT (**Fig. 2a**). Finding such a large number of highly tumor-specific TE-gene fusion examples indicates that this a promising resource of therapeutically targetable events. With solely the expression of the candidate TE-gene fusion transcripts, we performed a standard dimensional reduction technique, t-Distributed Stochastic Neighbor Embedding (t-SNE)[318], on all of our tumor samples to examine the similarity of across samples. As expected, we saw that clustering based on TE-gene fusion transcripts was mainly tumor-specific (**Fig. 2b**); however, there were certain cancers that were split based on subtype. For example, esophageal carcinoma (ESCA) was split between squamous carcinoma and adenocarcinoma, labeled “ESCA(1)” and “ESCA(2)” respectively. These clusters are also consistent with previous studies analyzing the data based on mRNA expression and chromatin accessibility[319], and our results indicate that the activation of these TE exaptation events may be related to differences in underlying mRNA expression, chromatin accessibility, and regulatory networks across these various tumor types.

To evaluate the presence of our candidates in cancer cell lines, we also processed RNA-sequencing data from a previous study and quantified the expression of our candidates in 675 cell-lines generated by Cancer Cell Line Encyclopedia (CCLE)[320]. We manually curated the cell lines with corresponding TCGA identifiers and also marked those identified as being commonly misidentified[321]. We were able to detect 56% of our candidates across the cell line catalogue that represented 26 of the 33 cancers profiled in TCGA as well as other cancer types such as Burkitt's lymphoma and multiple myeloma (**Fig. 2c**). With this analysis we have now created a resource of the ideal experimental system to interrogate a sizeable portion of our candidates.

Next, we predicted the most likely protein products of these transcripts and their antigenic potential. Two methods were used to predict the open reading frame: (1) the first was based off of using CPC2 to evaluate coding potential and (2) the second was based on finding the first start codon in a "strong" Kozak context (**Methods**). The location of the novel start codon was used to determine if this would create an in-frame or out-of-frame peptide. For in-frame candidates, we further evaluated if the candidates would be either of the following four categories: (1) normal, (2) truncated, (3) chimeric normal, or (4) chimeric truncated (**Fig. 3a**). The vast majority (90%) of our TE-gene fusion candidates were predicted to code for a protein in at least one of our two methods (**Fig. 3b**). Furthermore, considering a substantial portion of our candidates would be truncated or chimeric truncated, we evaluated how these new isoforms would affect the major protein domains and transmembrane domains using Pfam[322] and TMHMM[323]. Of the 1,500 truncated isoforms, 958 (63.9%) disrupt at least one Pfam domain and 239 (15.9%) disrupt at least one transmembrane domain. These TE-derived truncated isoforms could have alternative oncogenic functions from the canonical protein. For three promising candidates, we highlight the predicted protein structure, abundance across tumor types, and the cell lines with the highest levels of candidate expression (**Fig. 3c**). Strikingly, 763 (31%) of all the candidates were predicted to be out-of-frame, chimeric truncated, or chimeric normal and have the potential to be antigenic. Within tumor types, the presence of these antigenic TE-gene fusion events varied between 55% in THCA to 98.8% in ESCA, but overall, 84.6% of tumor samples had at least one antigenic candidate, further supporting this mechanism as a potential pan-cancer source of antigens (**Fig. 3d**).

Though our transcriptomic analyis allowed us to amass a valuable pool of potential cancer-specific targets, we sought to provide more evidence supporting the existence of these peptides as well as their presentation on MHC class I molecules. As a proof of principle, we utilized publicly available HLA-pulldown mass spectrometry data[324] for four of the cell lines that we profiled (SUPB15- acute lymphoblastic leukemia, HCT116- colon cancer, HCC1143- breast invasive ductal carcinoma, and HCC1937- breast ductal carcinoma) from the CCLE. The study also analyzed the EBV transformed JY cell-line that we did not have RNA-sequencing data for, but we also included that data to search for our protein products. In addition, we obtained SNV and HLA binding affinity predictions from the Tron Cell Line Portal[325] to predict neoantigen burden in the cancer cell lines. For our TE-gene fusion candidates, we used NetMHCPan-4.0 to predict the binding affinity of out-of-frame and chimeric peptides[51]. Then, we examined the relative amount of predicted strong (<=500 nM IC_{50}) HLA binders between the mutational neoantigens and the TE-gene fusion antigens (**Fig. 4a,b**). Across all the samples, it is apparent that HCT116 mutational neoantigens are a huge outlier with almost a magnitude more potential HLA-binders than the next closest sample. In addition, though the gap is much smaller for the other 3 cell lines, there are consistently more predicted mutational neoantigens than there are TE-gene fusion antigens with HCT116 having the least.

To compare the amount of each type of antigen that could be found in MHC-pulldown mass-spec experiments, we created a custom database composed of the Uniprot reference[326], all predicted neoantigens within the cell lines (regardless of predicted binding), and all TE-gene fusion candidates. We subsequently used Maxquant to search for these candidates in the HLA peptidomes of these cell lines[327]. For neoantigens, we were able to detect four for HCT116 and one for SUPB15 (**Fig. 4c**). This aligned with HCT116 having the largest number of potential candidates by a large margin. Shifting our focus to TE-gene fusion candidates, we were amazed to find a substantial number of antigens coming from these events: four in HCC1143, seven in SUPB15, and four in JY. TE-gene fusion peptides substantially increased the number of detected antigens for two cell lines when compared to neoantigen analysis alone. Though this was a small pilot for TE-gene fusion antigen discovery, the distribution of antigens show how these tumor-specific TE-derived antigens (TS-TEAs) could serve as complementary source of antigens to target therapeutically. In addition, our framework of identifying candidates could easily and quickly be incorporated into popular neoantigen detection pipelines that use transcriptomic data and could substantially increase the yield of protein targets[50,328].

We also personally validated the presence of TS-TEAs on MHC-molecules by performing MHC-pulldown LC-MS/MS experiments. We have successfully adopted a previously published pulldown methods[329] (**Fig. 5a**) to generate high-quality samples in accord of what is expected from MHC-pulldown. The peptides characterized by MaxQuant[327] from our MHC-pulldown samples represent 8-12 amino acid peptides with a mass/charge ratio from 250 to 1200 M/Z as expected of antigens presented on MHC-I molecules (**Fig. 5b**). Analysis of CCLE RNA-seq data provided candidate cancer cell lines with high prevalence of TE-derived transcripts that can ultimately be presented as antigens. To verify promoter activity of TEs identified to generate chimeric peptides, we generated nanoCAGE-seq libraries and mapped the transcription start sites in the candidate cancer cell lines (**Fig. 5c**). Furthermore, we set extremely stringent filters for processing antigens identified by MHC-pulldown experiments. These filters include RNA and CAGE support for TE promoter activity and BLAST-ing the peptide sequence against the current proteome database to determine peptide exclusivity to cancer samples. Furthermore, we reverse-translate the TS-TEA into all possible DNA sequences and BLAT the each sequence against the genome to determine if any other genomic regions can potentially transcribe a transcript that can be translated into the detected TS-TEA candidate. This often filters TE-derived peptides from young TEs, such as L1s and HERVs, due to conservation of sequence. For example, we detect numerous chimeric candidates transcribing from SVA elements. However, since there are multiple SVA elements that can generate the same peptide sequence, we filter out majority of SVA-derived antigen candidates due to lack of specificity. Even with these filters, we positively identified one TS-TEA antigen from an out-of-frame peptide from L1PA2-IBSP chimeric transcript in DMS53 cancer cell line (**Fig. 5c**). Indeed, L1PA2 is a bona-fide tumor-specific promoter in DMS53 further supporting that TE-chimeric transcript is expressed (**Fig. 5c**). However, synthetic peptide validation will further improve confidence that this is a real antigen.

Cell membrane proteins have been hot topic of discussion for immunotherapy-targets due to their intrinsic nature of being presented on the cell membrane surface. Often the case, in TE-chimeric transcripts, the TE sequences are prepended to a canonical transcript, which ultimately can be translated to add novel peptides to the N-terminus of the chimeric protein. Therefore, we hypothesized that TE-exapted membrane proteins have the potential of presenting a tumor-specific TE-derived epitope outside the cancer cells. Indeed, we identified numerous tumor-specific TE-exapted membrane protein events (data not shown). Here, we focus on two most prevalent candidates, L1PA6-STIM1 and SVA_F-GPR176, and attempt to validate its presence in cancer cell lines. L1PA2-STIM1 is a chimeric truncated transcript that retains its transmembrane domains and presents novel ~40 amino acid sequence in the N-terminus of STIM1, which is predicted to be outside the cell (**Fig. 6a**). L1PA6-STIM1 is present across numerous tumor types with highest incidence in BRCA (**Fig. 6b**). To validate this candidate, we profiled cancer cell line data and identified a cell line, H2110, that highly expresses the TE-exapted version of STIM1 (Fig. 6c). We profiled promoter activity in H2110 with nanocage-seq and verified that L1PA2 is actively transcribing (**Fig. 6d**). Since L1PA2-STIM1 is truncated version of STIM1(74kDA vs canonical 77.4kDa), we performed cytosol and membrane protein isolation and probed for STIM1 protein presence. We detect that smaller isoform of STIM1 present in the membrane, but not cytosol, of H2110 further supporting the presence of TE-exapted STIM1. Similarly, we validated the presence of SVA_F-GPR176 candidate in H1623 (**Fig. 6f,g**). Considering that the only isoform of GPR176 is transcribed from SVA_F TE in H1623 (**Fig. 6i**), we are confident that the SVA_F-GPR176 is present in the cell membrane (**Fig. 6j,k**). However, the size of the GPR176 is much larger than expected (TE-version: 54kDa vs. canonical version: 57kDa). We suspect that there could be post-translation modifications occurring on the SVA_F-GPR176 protein that could explain the size difference. However, further experiments, such as CRISPR deletion of SVA_F, will be necessary to validate that the GPR176 detected in H1623 is the TE-exapted version. If these TE-exapted membrane proteins are indeed real, it can open new therapeutic avenues of utilizing antibody-based or CAR-T therapies without the restrictive need for TE-chimeric peptides to be presented on MHC molecules.

7.4 Discussion

In cancer, transposable elements are promiscuously resurrected to drive expression of novel transcripts that are oftentimes highly expressed, widespread across tumor samples, and have the potential to produce novel peptides not predicted to be present in other adult tissues. Here, we provide a comprehensive analysis of TE-gene fusion events across 33 tumor types from TCGA. Even after stringent filtering for tumor specificity using both splice junction profiling in adult tissue and promoter expression, we find thousands of events present in the vast majority of tumors. In addition, a substantial portion of these events have unique peptide sequences, and our pilot MHC-pulldown mass spec data analysis shows that these events can be detected on HLA molecules and could be a promising complementary source of antigens for immunotherapy. Though this analysis was promising, there are many hurdles that our subset of TS-TEA face. They could be subject to some level of central tolerance as has been found for other tumor-specific antigens[43], and we will have the burden of proof in showing that these events are significantly immunogenic. In addition, even after stringent filtering, we are less confident in the tumor-specificity of our candidates when compared to neoantigens that result from genomic changes specific to the tumor since there could be rare subtypes of normal cells that have yet to be profiled. Nonetheless, considering that many of these candidates have disrupted epigenetic regulation, unique transcript sequences, and unique protein sequences, they offer a plethora of avenues to target them in cancer. Furthermore, for cancers with low levels of TE-exaptation events, previous work from our lab indicates that epigenetic therapy can be used to activate these transcripts and potentially lead to the production of antigenic peptides[330].

We plan to incorporate this rich database of tumor expression, cell-line expression, protein structure prediction, and antigenic peptide prediction generated in this study into an intuitive web interface. This way, researchers can easily locate promising candidates and identify the ideal experimental cell line model to validate and interrogate their chosen antigen. For our own future analysis, we have found the optimal combination of cell lines to validate half of the top candidates and increase likelihood of detecting candidates on subsequent MHC-pulldown experiments. Though our primary goal was to evaluate antigenicity of these candidates, there are many interesting biological questions that can be asked based on the various different isoforms detected. Does the absence of specific domains make a protein more oncogenic? Do any of the chimeric peptide sequences have a novel function in this tumor-specific protein-isoform? In addition, we are also working to package both our transcript annotation tools and our protein prediction algorithm to allow for easy incorporation of our methods into tumor sequencing and neoantigen detection workflows. We hope this will accelerate research on TS-TEA candidates and their incorporation into cancer therapies.

7.5 Methods

Data download

Normal and tumor RNA-seq BAM files from TCGA for the following 33 cancers were downloaded using the gdc-client version 1.3.0: adrenocortical carcinoma (ACC), bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), cervical carcinoma (CESC), cholangiocarcinoma (CHOL), colon adenocarcinoma (COAD), diffuse large b-cell lymphoma (DLBC), esophageal carcinoma (ESCA), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney chromophobe (KICH), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), acute myeloid leukemia

(LAML), low grade glioma (LGG), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), mesothelioma (MESO), ovarian serous cystadenocarcinoma (OV), pancreatic adenocarcinoma (PAAD), pheochromocytoma and paraganglioma (PCPG), prostate adenocarcinoma (PRAD), rectum adenocarcinoma (READ), sarcoma (SARC), skin cutaneous melanoma (SKCM), stomach adenocarcinoma (STAD), testicular germ cell tumors (TGCT), thymoma (THYM), thyroid carcinoma (THCA), uterine corpus endometrial carcinoma (UCEC), uterine carcinosarcoma (UCS), and uveal melanoma (UVM). In addition, normalized gene expression data (HTSeq-FPKM-Uq) and clinical metadata for all samples were downloaded using the gdc-client version 1.3.0. A total of 10,365 tumor samples and 729 matched-normal samples were used for analysis. The 675 cell line RNA-sequencing data were downloaded from the European Genome-phenome Archive (accession EGAD00001000725). GENCODE Version 25 was used as the transcript reference[288]. The GTF file of consensus transcripts was downloaded from https://www.gencodegenes.org/releases/25.html. Repeatmasker annotations were downloaded from the UCSC table browser for hg38 (refs. [331,332]). FANTOM5[317] hg38-aligned peaks used for annotating the supplementary tables were downloaded from http://fantom.gsc.riken.jp/5/datafiles/reprocessed/hg38_latest/. The genome reference used for RNA sequencing analysis (GRCh38.d1.vd1.fa.tar.gz) was the GDC reference files that were downloaded from https://gdc.cancer.gov/about-data/data-harmonization-and-generation/gdc-reference-files. Splice junction expression data (raw counts) for GTEX and TCGA were downloaded from Snaptron[316] using the provided command line client (https://github.com/ChristopherWilks/snaptron-experiments). HLA-pulldown mass spectrometry raw data from a previous study[324] for the HCT116, SUPB15-RT, SUPB15-WT, JY, HCC1143, and HCC1937 cell lines were downloaded from PRIDE using the following accession: PXD000394. Mass spectrometry mzML files for BRCA were downloaded from the CPTAC data portal (https://proteomics.cancer.gov/data-portal).

TE-gene fusion transcript pipeline:

This pipeline is available upon request, and the steps for candidate identification (1-4) have been described previously[39]. We further predict protein products from these transcripts and generates a fasta file with potential antigenic transcript products.

(1) Assembly and annotation of transcripts

BAM files were sorted and indexed and chr1-22, X, and Y were extracted. Stringtie version 1.3.3 was used to assemble the BAM files for all the RNA-seq samples (stringtie –m 100 –c 1). These transcripts were then annotated with features from GENCODE v.25 with a custom script. Briefly, GENCODE v.25 was first processed into a coordinate dictionary based on chromosome, start, and end location for the "appris_principal" transcription. This dictionary of principal transcripts as well as the Repeatmasker TE coordinates were used to annotate the transcripts generated from the stringtie assembly for each sample. The starting position of the transcript was annotated using the Repeatmasker table to find TE-derived transcription start sites. Then, the first exon of the transcript was annotated based on overlap with exonic or intronic features of GENCODE v.25. If the exon overlapped both an exon and intron, then the exon was selected as the annotation for that element. Then, all subsequent exons in the transcript were annotated until one overlapped with a protein-coding gene exon; this exon of the protein-coding gene was selected as the "splice target" of that transcript. After all transcripts were annotated, candidate transcripts were selected based on the following criteria: the start site of the transcript being within a TE, the TE being intergenic or intronic, the starting exon not overlapping with exon 1 of the canonical gene, and the transcript splicing into a protein-coding gene.

(2) Generating a reference transcriptome including onco-exaptation candidates

Aggregating annotation data across all tumor and normal RNA-seq data sets, we constructed a list of unique TE-gene fusion candidates based on the subfamily of the TE, the chromosomal coordinates of the TE, and the exon of the gene that the transcript spliced into. To remove potential assembly artifacts and genomic contamination, we removed candidates that had an average exon 1 length greater than the 99th percentile of all GENCODE v.25 transcript first exons (2588 bp). Furthermore, transcripts with first exons that retained an intron were also removed. Finally, we only included candidates that were present in at least 2 samples.

To further increase confidence of promoter activity, we interrogated all reads that uniquely mapped to each candidate TE. We subsequently annotated the mate pair of those reads to see if any overlapped directly with oncogene exons. For single-end reads, we annotated the portion of the read mapping outside the TE to see if it overlapped with an oncogene exon. First, we removed candidates that had zero files where there were at least 10 uniquely mapped reads that started within the TE. In addition, these events were required to have at least 1 sample with uniquely mapped paired-end reads where one of the pairs mapped to the TE and the other to the splice target of the candidate. For intronic onco-exaptation events, we also removed candidates that had evidence of exonization (there were reads mapping to both an upstream exon and the TE) in more than 15% of samples. Finally, candidates that were exclusively in single-end RNA-seq files were removed. The remaining candidate TE-derived transcripts were then merged with the reference GENCODE v.25 annotation file using Cuffmerge[333] to create a reference transcriptome inclusive of potential TE-gene fusion events.

(3) Transcript-level quantification and candidate selection

To determine the contribution of candidates to overall gene expression, we used stringtie quantification (-e -b) with the merged transcriptome as the reference. The FPKM values generated by this command were extracted from the ballgown output files to get transcript-level expression. In addition, intron read counts for unique splice junctions found in each transcript were also extracted from the stringtie ballgown output for further analysis.

(4) Candidate TE-gene fusion transcript filtering

For each sample, we labeled a candidate as being present if it met the following criteria: (1) the transcript accounted for at least 10% of total gene expression, (2) there was at least one read covering the closest unique splice junction to the splice target (candidates without unique splice junctions were removed), and (3) the target gene had at least 1 FPKM expression.

(5) Open-reading frame prediction and annotation

We utilized two strategies to define the coding reading frame of the transcripts. (1) First, we used CPC2 which predicted which candidates were coding or non-coding. For coding transcripts, we subsequently used the start codon identified by CPC2 for the longest open reading frame[292]. (2) Second, we searched for the first start codon (AUG) that met the following criteria: (A) the start codon matches the Kozak[334] sequence at the +4 or -3 positions and (B) the protein made had to be at least 50 AA in length. For each of these methods, we then determined if the start codon identified was in-frame or out-of-frame relative to the canonical start codon of the gene. If it was out-of-frame, we checked if it was present within the original transcript of the protein, and if it was, it was removed from consideration. For in-frame proteins, we determined if the start codon was predicted to make a protein in either of the following categories: original, truncated, chimeric original, and chimeric truncated. If both the selected start codon and the novel start codon were present in the canonical transcript, then the candidate was annotated as making the canonical isoform of the protein. In addition, we performed pairwise alignment of the original protein sequence and the predicted coding frames of the transcripts using the Biostrings R package[335], and conflicts between the nucleotide sequence-based annotation prediction and protein alignment were manually checked and corrected.

Protein and Transmembrane Domain Annotation

To explore the putative functional and transmembrane domains that candidates retained or lost from the canonical protein, we utilized the Pfam[322] domain annotations and TMHMM[323] transmembrane domain predictions available through Biomart[336]. Those proteins that retained at least one transmembrane domain were kept as potentially membrane-associated candidates. To further highlight proteins associated with the plasma membrane specifically, we queried the Uniprot[326] database with the gene names for the cellular localization and selected those candidates that had the term "plasma membrane" in their description.

Candidate filtering for tumor-specificity

We filtered for candidates that were highly tumor enriched within the TCGA samples (>8x enrichment in the tumor samples when compared to the normal samples in TCGA) and present in at least 5 tumor samples. To further filter for candidates restricted to tumors, we incorporated the Genotype-Tissue Expression (GTEx) project that transcriptionally profiled 9662 samples across 31 adult tissue types. We downloaded splice junction count information profiled by the Snaptron project for both the TCGA and GTEx projects. For the closest unique splice junction to the "splice target" of each candidate, we evaluated the expression across all TCGA and GTEx datasets and calculated a junction counts per million (jpm) value. We subsequently determined the maximum jpm values across TCGA and GTEx (excluding Testis) separately, and kept candidates with a maximum jpm in GTEx of 0 or that met the following criteria: (1) maximum jpm across TCGA samples > 0, (2) maximum jpm across GTEx samples <= 1, and (3) (maximum jpm across TCGA samples)/(maximum jpm across GTEx samples) >= 2.

For FANTOM5 promoter annotation, we first filtered the FANTOM5 peaks in hg38 for samples that were not part of exposure or time-course experiments. Subsequently, we evaluated if there were any peaks that overlapped with the onco-exapted TE that were on the same strand as our candidate transcript. If there were multiple peaks in the same transposable element, we combined them to get the amount of expression coming from the transposable element. We then calculated the mean expression level of the TE promoter (tpm) across all the adult tissues. We removed all candidates that had a mean tpm expression >=1 tpm in any adult tissue (exclusion Testis).

Raw RNA-sequencing data processing

For the 675 cell line rna-sequencing data, we first performed adapter trimming using cutadapt[226]. We subsequently aligned the reads using STAR v2.6.1b[337] (*STAR --runMode alignReads --runThreadN 6 --genomeDir <reference directory> --readFilesIn <R1.fastq.gz> <R2.fastq.gz> --readFilesCommand zcat --outFileNamePrefix <name.out> --outSAMtype BAM SortedByCoordinate --outSAMstrandField intronMotif --outSAMattributes NH HI NM MD AS XS --outSAMunmapped Within --outSAMheaderHD @HD VN:1.4 --outFilterMultimapNmax 20 --outFilterScoreMinOverLread 0.33 --outFilterMatchNminOverLread 0.33 --alignIntronMax 500000 --alignMatesGapMax 1000000 --twopassMode Basic*). We subsequently used stringtie to quantify the expression of TE-gene fusion candidates using the reference generated from the TCGA data (created in step 2 of the TE-gene fusion transcript pipeline*)*. We used the same parameters as we used for the TCGA data to identify if a candidate was present in a sample: (1) the transcript accounted for at least 10% of total gene expression, (2) there was at least one read covering the closest unique splice junction to the splice target, and (3) the target gene had at least 1 FPKM expression.

HLA-type Determination and Binding Prediction

For the TCGA samples, the HLA type has already been determined in a previous study[304], and the HLA allele information was downloaded using the gdc-client 1.3.0 using the manifest found at the following location: https://gdc.cancer.gov/about-data/publications/panimmune. For the 675 cell lines, the seq2HLA[308] program was run to predict the 4-digit HLA allele code for every cell line using their RNA-sequencing data. We then predicted the binding affinity of the unique peptide sequences of our TE-gene fusion candidates to the HLA alleles of samples with the candidate present using NetMHCPan-4.0[51]. For out-of-frame candidates, we included the entire predicted protein sequence. For in-frame chimeric candidates, we included the novel N-terminal AA + 10 AA from the original protein.

SNV Neoantigen Prediction for TCGA and Cell Line Data

SNV Neoantigens for TCGA and their predicted binding affinity to sample HLA alleles had been previously profiled, and we obtained the predicted neoantigens and their predicted binding affinities to samples of TCGA with the gdc-client 1.3.0 using the manifest found at the following location: https://gdc.cancer.gov/about-data/publications/panimmune. SNV neoantigens for all the cell lines had also been profiled by the TRON Cell Line Portal[325], and their sequences and binding affinity were obtained from the authors.

HLA Mass Spectrometry Analysis

Raw files our MHC-pulldown samples and publically available HLA-mass spectrometry experiments were analyzed using MaxQuant[327] Version 1.6.3.4. The parameters used in the proteomics search different from the default were the following: unspecific enzyme digestion, no protein-level FDR since we were interested in peptides, peptide FDR of 5%, peptide length limit between 8 and 15 AA, and maximum peptide mass of 1500 Da. In addition, a custom proteome database was used consisting of the Uniprot reference database, the sequences of potential antigenic peptides from our analysis, and the potential neoantigen sequences from the TCLP (JY was not available). Decoy sequences and contaminants were removed before performing subsequent identification analysis.

Detection of transcription start site locations

To detect transcription start site locations in promoters, we generated and processed nanoCAGE-seq libraries on cancer cell lines described in this paper by following protocol described previously[39].

MHC/HLA pull-down and LC-MS/MS procedure.

We adopted a published protocol[338,339] to generate HLA-I antigen pull-down samples with couple exceptions mentioned below. As per published protocol, anti-HLA-I antibodies were collected from W6/32 (ATCC HB-95) growth medium and were crosslinked to Protein A-sepharose 4B beads (ThermoFisher Scientific, 101041) with dimethylpimelimidate. We harvested and froze down cell line samples until time of lysis. Roughly 5x10^8 to 1x10^9 cells were lysed with ice-cold modified lysis buffer (0.3% sodium deoxycholate, 0.75% IGEPAL CA-630, 0.2mM iodoacetamide, 1mM EDTA, 1:200 Protease Inhibitors Cocktail, 1mM Phenyl-methylsulfonyl fluoride, 1% octyl- β-D glucopyranoside in PBS) on ice for 1 hour. The samples were slightly vortexed every 10 minutes to maximize lysis efficiency. The samples were centrifuged at 21,000x g at 4°C for 1 hour to pellet large cell debris and cell nuclei. The lysate supernatant was transferred to Poly-Prep chromatography columns (Bio-Rad, 7311550) with 500ul Protein A-sepharose 4B beads to remove endogenous antibodies. The endogenous antibody-depleted lysate was then transferred to Poly-Prep chromatography columns containing 1ml of crosslinked W6/32-proteinA beads. The flow-through was collected and loaded to the same column for a total of three flow throughs. Then the MHC-antigen bound beads were washed and MHC-I antigens were purified following the aforementioned protocol[338]. The antigens samples were then processed on nanoLC coupled to Orbitrap Fusion Lumos Mass Spectrometer. Mass spectrometry results were acquired following the "top10" method.

Western blot and immunofluorescent detection of TE-derived membrane proteins

All cell lines were grown in mediums and conditions designated by ATCC. Membrane-bound and cytosolic proteins were isolated from cell line samples with Mem-PER Plus Membrane Protein Extraction Kit (ThermoFisher Scientific, PI89842) following manufacturer's protocol. Extracted protein samples were denatured with Blue Loading Buffer Pack (Cell Signaling Technology, 7722S) and loaded into Novex 10% Tris-Glycine Mini Gels (Thermo Fisher Scientific, XP00100BOX) and separated by gel electrophoresis at 125V for 4 hours. STIM1 and GPR176 were detected using anti-STIM antibody (Cell Signaling Technology, #5668S) and anti-GPR176 antibody (abcam, ab122605) with anti-rabbit secondary antibody (Cell Signaling Technology, #7074). The Western blot was imaged with Thermo Scientific myECL Imager (Thermo Scientific, 62236).

For immunofluorescence detection of GPR176 in H1623 cell line, H1623 was seeded into 12-plate well culture-treated plates with 18mm glass coverslips. Once ~50% confluent, we aspirated the media and washed each well twice with room temperature (RT) PBS. Then the cells were fixed with 300µl of 4% paraformaldehyde (PFA) in PBS at RT for 10 minutes. After 3 washes with PBS, the cells were permeabilized with 0.01% Triton X-100 in PBS for 10 minutes at RT.

Then the fixed cells were treating with blocking buffer (5% bovine serum albumin in PBS) for 30 minutes at RT with slow shaking. For primary antibody incubation, anti-GPR176 antibody was diluted 1:100 in blocking buffer and added to cells for overnight incubation at RT in the dark. To label GPR176 with GFP immunofluorescence, we washed the cells three times with PBS and added Alexa Fluor 488 conjugated donkey anti-rabbit (ThermoFisher Scientific, A-21206) at dilution of 1:100 for 1-hour incubation at RT in the dark. For imaging, the cells were washed 3 times with PBS. For the second wash, DAPI was added to the PBS at 1:5000 dilution and incubated for 5 minutes to label nuclei. Then the coverslips were mounted on slides and cells were imaged with a Leica DM IL LED Fluorescence Inverted Microscope at 40x magnification.

7.6 Acknowledgments

We would like to thank J. Hoisington-López and M.L. Jaeger from The Edison Family Center for Genome Sciences & Systems Biology (CGSSB) for assistance with sequencing; B. Koebbe and E. Martin from CGSSB for data processing; This work was funded by NIH grants 5R01HG007175, U24ES026699, U01HG009391 and the American Cancer Society Research Scholar grant RSG-14-049-01-DMC. HSJ was supported by a grant from NIGMS (T32 GM007067). NMS is a Howard Hughes Medical Institute (H.H.M.I.) Medical Research Fellow, http://www.hhmi.org/. We thank Proteomics & Mass Spectrometry Facility at the Danforth Plant Science Center for its contribution and is supported by the National Science Foundation under Grant No. DBI-1827534 for acquisition of the Orbitrap Fusion Lumos LCMS/MS.

7.7 Figures

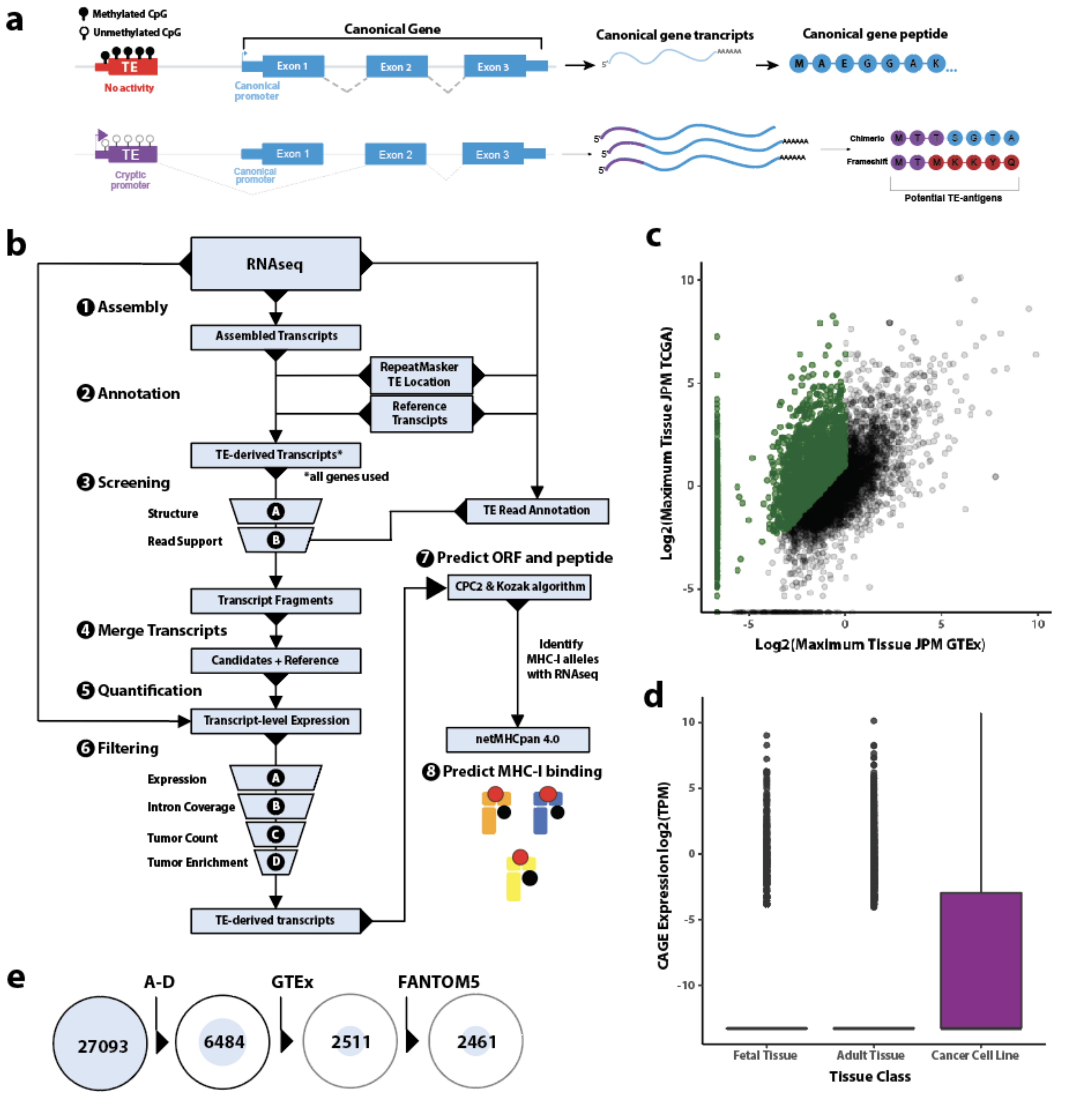


Figure 1: Tumor-specific TE-gene fusion candidate selection. a, Schematic of TE-gene fusion events having the potential to create novel chimeric peptides or to create out-of-frame proteins that are tumor specific **b,** Diagram of our computational framework for RNA-sequencing data to detect tumor-specific TE-gene fusion events, predict their reading frames, and predict potential antigens. **c,** Maximum unique splice junction expression in units of junctions per million (jpm) in

TCGA and GTEx data sets. Those labeled in green were kept as potential tumors-specific candidates. **d,** FANTOM5 CAGE expression (tpm) of candidates with FANTOM5 peaks in fetal tissue, adult tissue, and cancer cell lines.

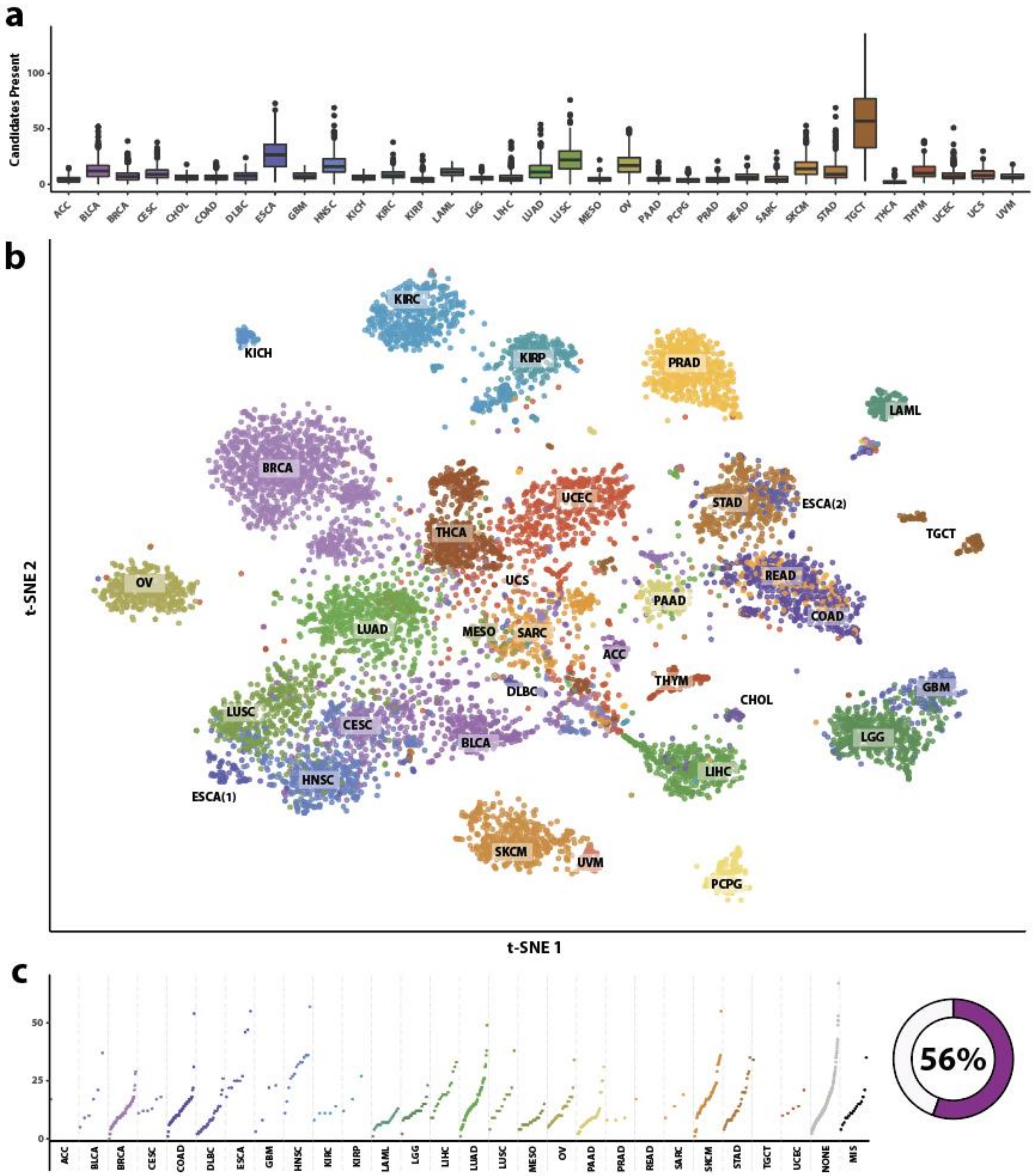


Figure 2: Landscape of TE-gene fusions in TCGA and in cancer cell lines. a, Series of boxplots for all 33 cancers in TCGA showing the distribution of the number of candidates detected per samples. **b,** t-SNE plot displaying the clustering of 10365 tumor samples based on the expression of the candidate TE-gene fusion transcripts. The dots are colored based on their tumor-type, and the clusters have been labeled with TCGA identifier corresponding to the cancer type that represents the majority of that cluster. ESCA has been labeled twice due to two large split clusters. **c,** Landscape of TE-gene fusion candidate presence across the 675 cancer cell lines profiled in this study. Cell lines not corresponding to any TCGA identifier are labeled as "NONE." Commonly misidentified cell lines are labeled as "MIS." The donut plot on the right represents the percentage of TCGA TE-gene fusion candidates that are detected in the cell lines.

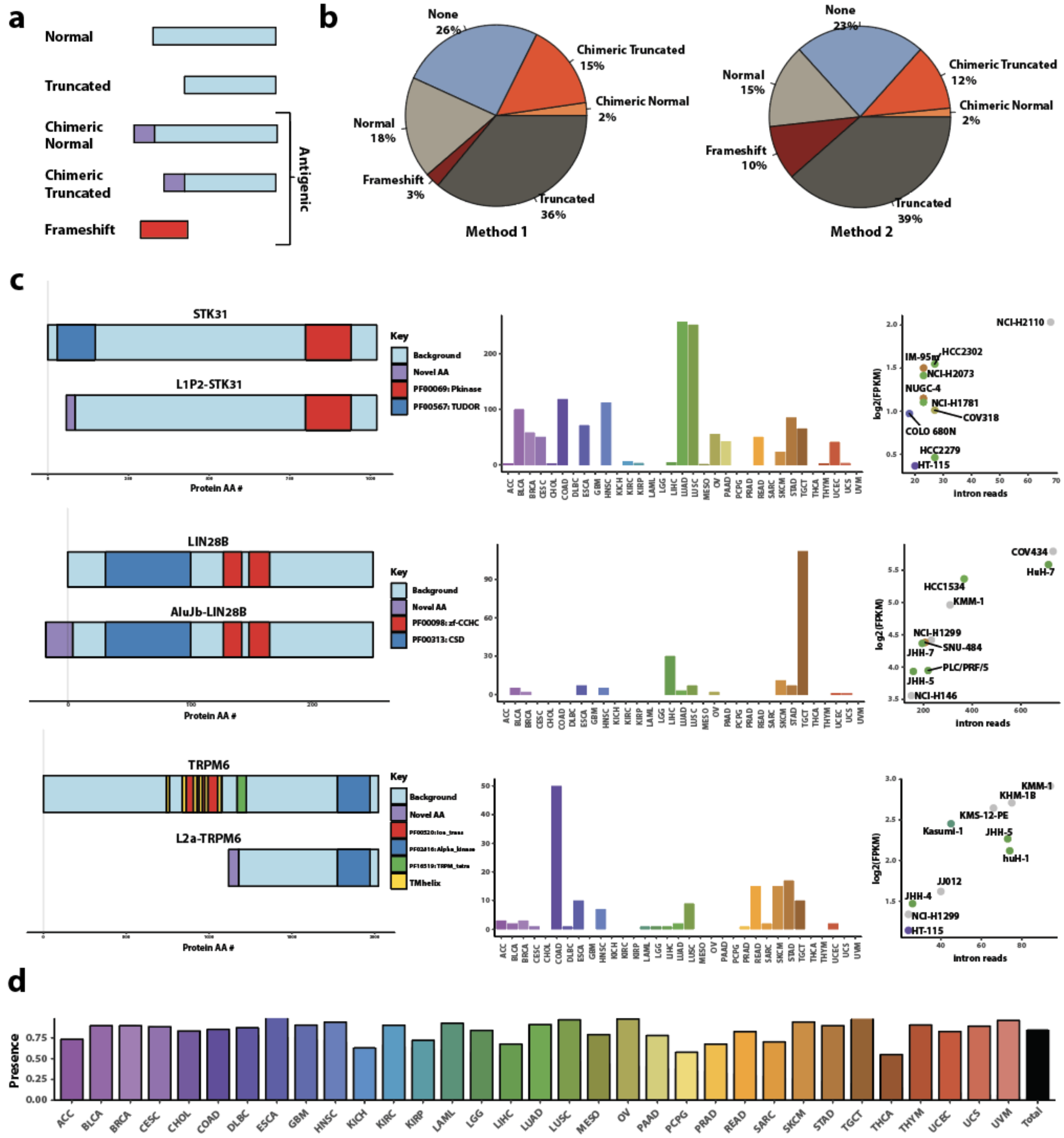


Figure 3: TE-gene fusion transcripts generate protein products that are potentially antigenic. a, Diagram of annotation possibilities of protein products from TE-gene fusion transcripts **b,** Pie graphs showing the distribution of protein product annotation using Method 1 (based on CPC2) and Method 2 (based on Kozak similarity). **c,** Three chimeric truncated candidates are highlighted. The left most column has the original protein structure (top) with the

candidate protein structure aligned (bottom). The middle column displays the presence of the candidate across the 33 TCGA cancer types. The right most column contains a plot showcasing the expression of the candidate in the top 10 cell lines where it is present. **d,** The proportion of samples in each cancer type that have at least 1 antigenic candidate.

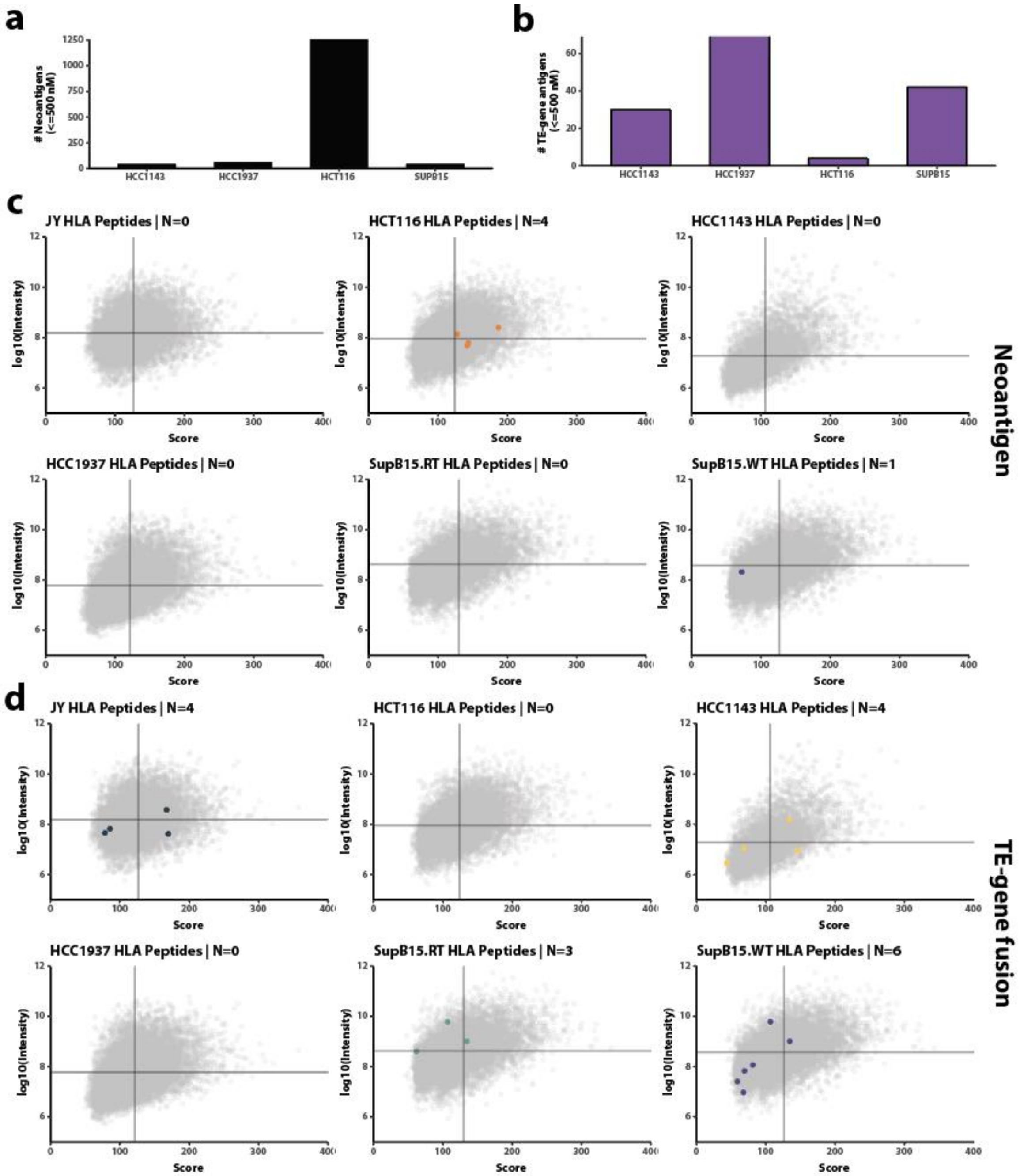


Figure 4: Detection of TE-gene fusion candidates in previously published HLA-pulldown LC-MS/MS data. a, Bar graph displaying the number of predicted neoantigen peptides with a predicted peptide affinity <= 500nM (IC_{50}). **b,** Bar graph displaying the number of predicted TE-gene fusion antigenic peptides with a predicted peptide affinity <= 500nM (IC_{50}). **c,** A series of scatter plots summarizing the MHC antigens detected based on spectral score (x-axis) and intensity (y-axis). Each grey dot represents a detected peptide (FDR<=.05). Colored dots highlight the neoantigen peptides that were detected. The horizontal and vertical lines represent the average intensity and score for the cell-line respectively. **d,** Same plots as **c** with the TE-gene fusion antigenic peptides highlighted instead of neoantigens.

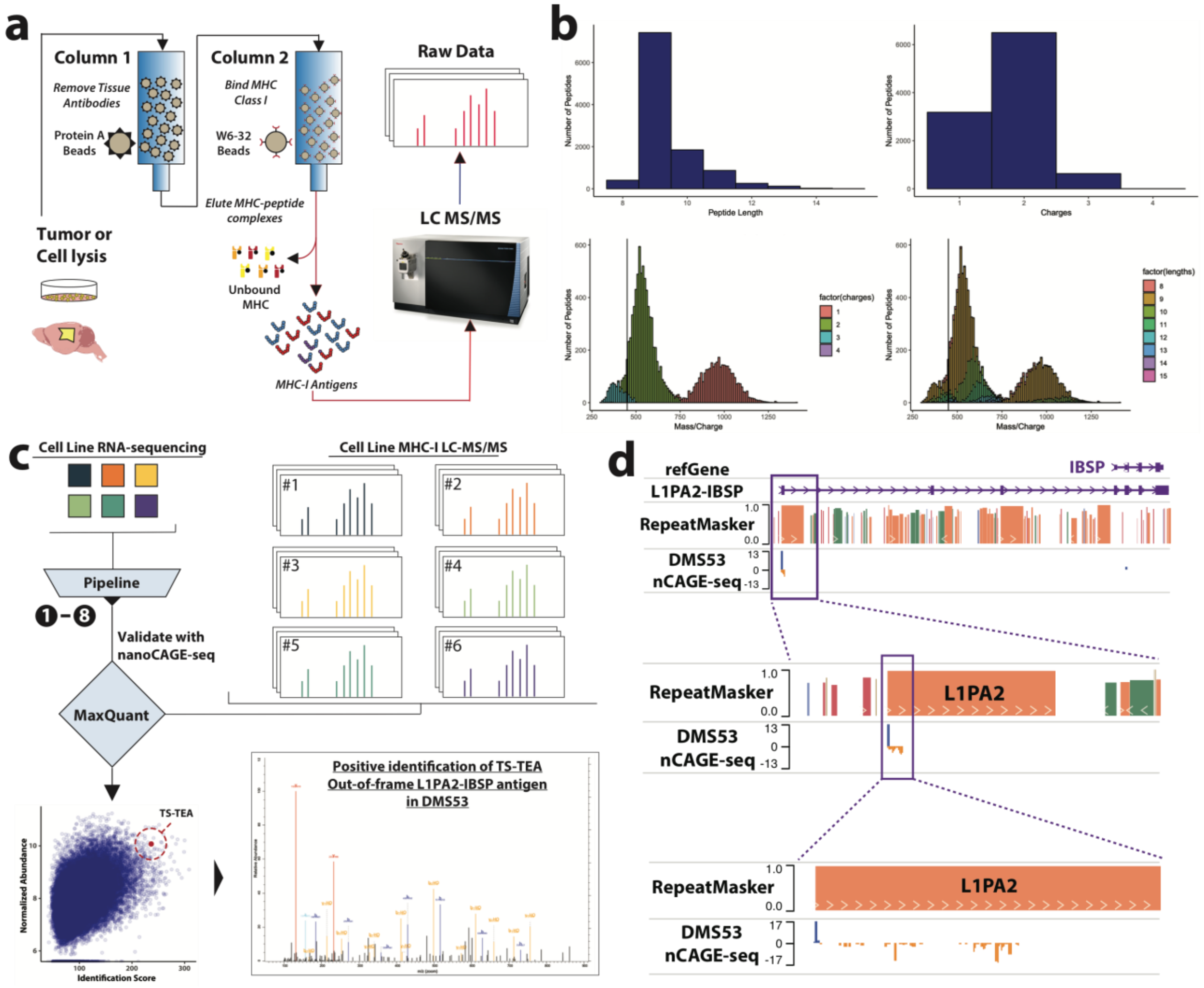


Figure 5: Identification of L1PA2-IBSP TS-TEA in DMS53. a, A schematic describing the process of MHC-I pulldown experiment. **b,** Characterization of peptide length, charge and mass/charge ratios from peptides detected by LC-MS/MS in MHC-I pulldown sample. **c,** Visual representation of validating presence of TS-TEAs in cancer cell lines. LC-MS/MS spectra result of L1PA2-IBSP TS-TEA. **d,** WashU Epigenome browser view of nanoCAGE-seq data in DMS53.

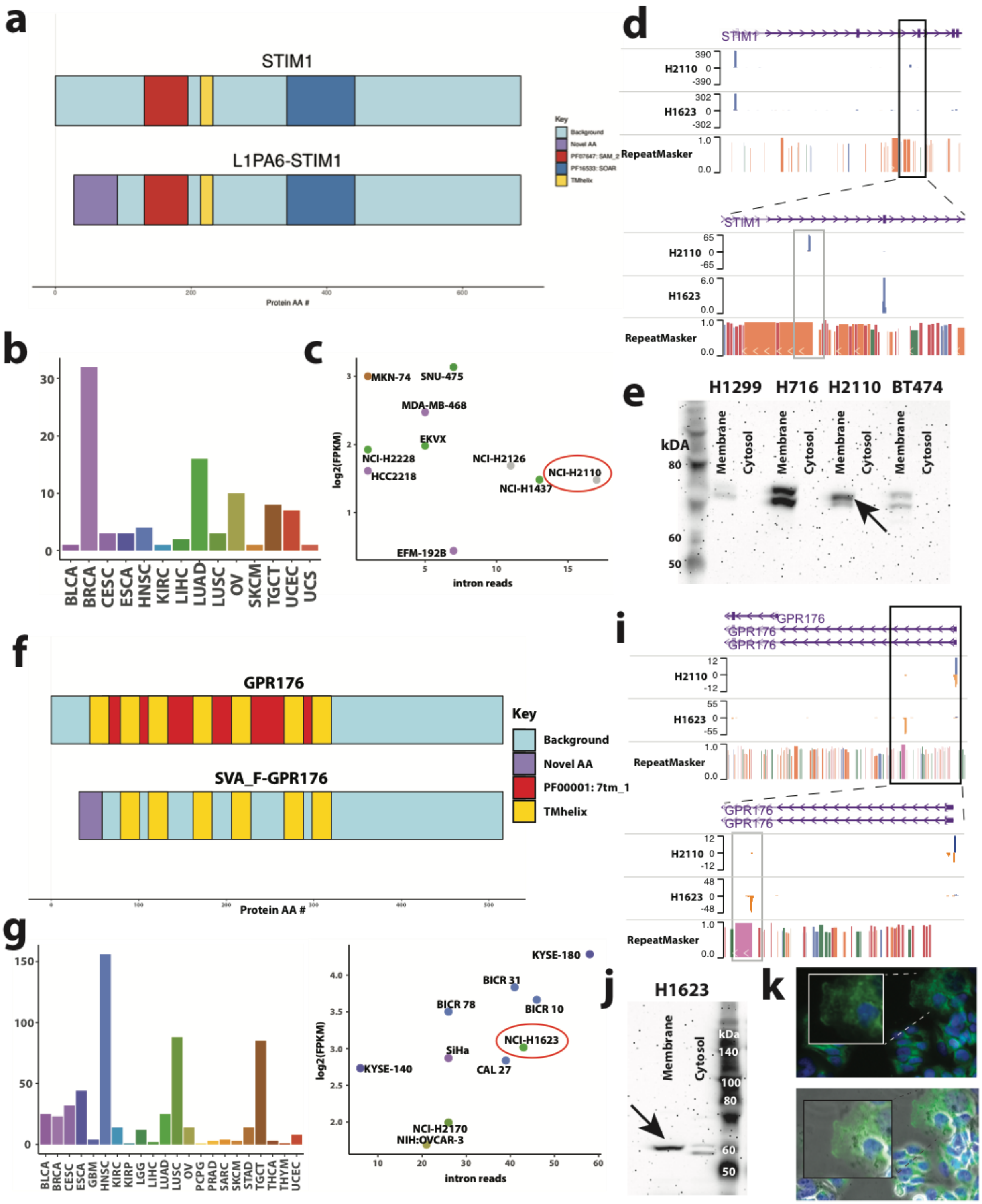


Figure 6: TE-derived membrane protein discovery and validation. a, A schematic representing canonical STIM1 and L1PA6-STIM1 protein structure. **b,** Frequency of L1PA2-

STIM1 candidate detected in various TCGA tumors. **c,** Expression profile of L1PA6-STIM1 in cancer cell lines. Intron reads represent number of reads that span the TE-gene junction. Red circle denotes cancer cell line used to validate L1PA6-STIM1. **d,** WashU Epigenome browser view of nanoCAGE-seq data of L1PA2 and STIM1 in H2110. **e,** Western blot of STIM1 in cell membrane and cytosol across cancer cell lines. Black arrow points to smaller STIM1 protein isoform in H2110. **f,** A schematic representing canonical GPR176 and SVA_F-GPR176 protein structure. **g,** Frequency of SVA_F-GPR176 candidate detected in various TCGA tumors. Expression profile of SVA_F-GPR176 in cancer cell lines. Intron reads represent number of reads that span the TE-gene junction. Red circle denotes cancer cell line used to validate SVA_F-GPR176. **i,** WashU Epigenome browser view of nanoCAGE-seq data of SVA_F and GPR176 in H1623. **j,** Western blot of GPR176 across cancer cell lines. Black arrow points to candidate SVA_F-GPR176 protein present on cell membrane. **k,** Immunofluorescence labeling of GPR176 in H1623. Blue represents DAPI. Green represents GPR176.

Chapter 8: Novel transposable element-derived antigens induced after epigenetic therapy in glioblastoma stem cells

Hyo Sik Jang[1,2*], Nakul M. Shah[1,2*], Juheon Maeng[1,2], Tatenda Mahlokozera[3,4,5,6], Devi Annamalai[3,4,5,6], Patrick DeSouza[3,4,5,6], Noah L. Basri[1,2], Justin Y. Chen[1,2], Jiaxin Ge[1,2], Hyung Joo Lee[1,2], Daofeng Li[1,2], Xiaoyun Xing[1,2], Albert H. Kim[3,4,5,6], Ting Wang[1,2]

[1]Department of Genetics, Washington University School of Medicine, St. Louis, Missouri

[2]The Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, Missouri

[3]Department of Neurological Surgery, Washington University School of Medicine, St Louis, Missouri.

[4]Siteman Cancer Center, Washington University School of Medicine, St Louis, Missouri.

[5]Department of Neurology, Washington University School of Medicine, St Louis, Missouri.

[6]Department of Developmental Biology, Washington University School of Medicine, St Louis, Missouri.

[*]Equal contribution

Author Contributions: H.S.J., N.M.S., and T.W. conceived and implemented the study; H.S.J., N.M.S., J.M, J.Y.C., D.L. and H.J.L contributed to the computational analysis; H.S.J., J.M. and X.X. generated epigenetic and transcriptomic profiles of cell lines; T.M., D.A., P.D. and A.H.K. generated GSCs and contributed reagents; H.S.J., N.L.B and J.G designed and implemented CRISPR-SunTag-MS2 experiments; H.S.J. performed the MHC-pulldown LC-MS; and the manuscript was prepared by H.S.J., N.M.S. and T.W.

This chapter is on-going, unpublished work.

8.1 Introduction

Transposable elements (TEs) make up 50% of the human genome but are referred to as “junk DNA” and often ignored in genomic medicine. However, our recent work reveals that TEs offer a unique opportunity to understand cancer-specific gene activities. We hypothesize that epigenetic dysregulation in cancer leads to the activation of cryptic regulatory elements encoded by TEs, some of which generate cancer-specific products including immunogenic antigens (**Fig. 1A**). Therefore, TEs are a rich and underexplored source of candidate targets for precision cancer therapy that deserve considerable attention.

The concept of vaccinations has been around for centuries, but it wasn’t until recently that this concept was systematically applied to cancer therapy. Monumental efforts have established that cancer-specific coding mutations create neoantigens that can be presented on the cell surface of tumors to trigger immunogenic clearance. Initial trials of neoantigen-based cancer vaccines showed therapeutic promise, as vaccines primed the immune system to better stimulate cytotoxic response against the tumor. However, current approaches to immunotherapy have not been universally effective, and this is especially true in tumors with a low mutational load which, in turn, carry a low conventional neoantigen burden. One example is glioblastoma, the most common and deadliest adult brain cancer. Neoantigen-vaccine clinical trials have validated the safety and immunogenicity of peptide-based vaccines in glioblastoma, and the evaluation of clinical response is ongoing. A major barrier to these approaches, however, is immunoescape in which primed T cell responses to a limited number of targetable immunogenic neoantigens eradicate only a subset of clones that make up the tumor. To overcome this limitation, there is a significant need to identify more tumor-specific antigens not expressed in normal tissue of the CNS.

Here, we explore how TE activation by epigenetic therapy can provide synergy with vaccine-based therapy. We develop and adapt a series of genomics and computational tools to identify and validate novel TE-derived antigens to demonstrate that these antigens can be presented on MHC molecules to activate T-cell responses. We pursued two FDA-approved epigenetic therapy drugs: Decitabine (DAC) and Panobinostat (LBH-589). DAC is a nucleoside analog that covalently binds to DNMT to incapacitate the enzyme. Pharmacokinetic and pharmacodynamic studies in murine models revealed that efficacious drug concentrations could be reached in the brain through both intraperitoneal and intravenous injections[340]. Panobinostat, a hydroxamic acid, is a pan-HDAC inhibitor commonly utilized in preclinical GBM trials for pre-sensitizing cancer cells to downstream therapy[341–343]. Although Panobinostat cannot cross the brain-blood barrier, various delivery mechanisms, such as nanoparticles, are currently being tested in glioblastoma models[344,345].

We focused on GSCs as these subpopulations in GBM are 1) responsible for the oncogenic potential of the tumor and 2) more generally, represent a tractable model system of human GBM tumors and stably recapitulate the genetic features of parent tumors in culture[346–348]. Our hypothesis argues that the connection between hypomethylation of transposable elements, epigenetic therapy, and cancer cells' immune response might give us a new way to understand and treat cancer. We profiled patient-derived glioblastoma stem cell (GSC) lines with state-of-the-art epigenomics technologies to identify novel sources of antigens originating from heretofore disregarded genomic regions. Our overarching goal is to characterize and verify the existence of TE-derived antigens in GBM and evaluate the potential of these antigens to stimulate an immune response for therapeutic outcome.

By comparing the epigenomic and transcriptomic profiles between epitherapy-treated samples and DMSO-treated (control) samples, we reveal the potential existence of treatment-induced transposable element-derived antigens (TI-TEAs) on HLA molecules. We distinguish cancer-specific reactivation of cryptic promoters after treatment by identifying those present in GSC samples but not in primary cell lines. Furthermore, by profiling numerous GSCs, we predict to find both shared and patient-specific examples, reinforcing the importance of both the global application and targeted strategy of this approach. We will pursue TI-TEAs by following the same strategy described in **Chapter 7** to verify that candidate antigens are presented on HLA-I. We will also improve epigenetic engineering strategies to target specific transposable elements, thereby selectively producing TI-TEAs for targeted therapy in GSCs.

Lastly, there is still an incomplete understanding of how epigenetic therapy impacts normal cells and tissues. Considering that autoimmune side-effects of immunotherapy can have potentially devastating consequences, it is imperative that we have a complete understanding of antigens' tumor-specificity. This is especially true if we aim to combine epigenetic therapy with immunotherapy to target induced TE-antigens. Although profiling the impact of epitherapy in all tissues and cell types is essential for translational progress of TI-TEAs, this effort can be overtly difficult due to the curation and treatment of proper "normal" samples and can be prohibitively expensive due to the cost of generating multi-omics libraries. Although not comprehensive, we provide the first look into how epitherapy might induce TE expression in normal cells by taking advantage of primary cell lines: adult human fibroblasts and normal human astrocytes. Normal cells in the body can be broadly categorized as dividing or non-dividing quiescent states, such as skin or brain cells respectively. To mimic these two cell states, we treated proliferating and contact-inhibition-induced[349–352] quiescent adult human fibroblasts and normal human astrocytes with epigenetic therapy (**Fig. 1B**). Our results present a cautionary tale for future efforts in harnessing TI-TEAs in the war against cancer.

8.2 Results

8.2.1 Epigenetic therapy results in promiscuous activation of the epigenome, with exception of quiescent cells.

The dose and length of treatment with epigenetic drugs were chosen to maximize activation of the epigenome through loss of DNA methylation and gain of histone acetylation without high cytotoxicity. Considering that decitabine's (DAC) efficacy is dependent on DNA replication[353] and GSC's doubling time is ~36-48 hours, we chose to treat the cells for 6 days to allow for a minimum of 3 cell replication cycles. Therefore, the epigenetic therapy regiment consisted of treating GSCs with 1uM DAC for 6 days and 100nM Panobinostat for the remaining last two days (**Fig. 1B**). Since Panobinostat was dissolved in DMSO, the control cell lines were treated with 0.05% DMSO for the last two days to control for cellular responses caused by DMSO treatment. Trypan blue staining assays revealed that cell death was similar to DMSO samples at the end of the treatment across all cell types assayed in this study.

To interrogate the epitherapy's efficacy and impact on DNA methylation, we generated

whole genome bisulfite sequencing (WGBS) library and calculated average CpG DNA methylation levels across all samples. We report variable, yet consistent, decreases of DNA methylation dependent on cell type (**Fig. 2A**). Epitherapy treatment reduced DNA methylation levels globally in GSC by 20-37% while primary adult human fibroblasts (hFB) and normal human astrocytes (NHA) displayed reduction of 10-13% DNA methylation. The discrepancy between GSC and primary cell lines could be explained by the differences in cell doubling time.

Indeed, in quiescent hFB (qhFB) and quiescent NHA (qNHA), there are minimal changes in the global DNA methylation levels reflecting DAC's dependency on cell replication and division for its activity **(Fig. 2A)**. Majority of the variation in DNA methylation dynamics is a factor of epitherapy treatment as illustrated by principal component analysis (PCA) **(Fig. 2B)**. Interestingly, although the loss of DNA methylation caused by DAC is considered to be a stochastic process, the GSC samples segregated into three clusters on PC2 based on cell identity suggesting that epitherapy could have cell line-specific DNA demethylation events. In summary, epigenetic treatment leads to global decrease in DNA methylation levels, but with varying magnitudes as the GSCs experienced the highest change and the quiescent cells showed almost no change.

Next, we asked how epitherapy impacted chromatin accessibility dynamics by comparing ATAC-seq results across the samples. As expected, Pearson's correlation clustering based on differentially accessible peaks revealed that GSCs clustered by epitherapy treatment condition. Proliferating epitherapy-treated primary cells also clustered with epitherapy-treated GSCs suggesting similar changes in chromatin accessibility (**Fig. 2C**). However, quiescent primary cells clustered with proliferating DMSO-treated primary cells regardless of epitherapy treatment status (**Fig. 2C)**. This result implied that epitherapy does not significantly alter the accessible chromatin landscape in quiescent cells, consistent with the finding that DNA methylation landscape also remains stable in quiescent cells after epitherapy. Furthermore, PCA reflected that the variance in chromatin accessibility can be mostly explained by epitherapy treatment with cell type-specific nuances (**Fig. 2D**). In conclusion, epigenetic therapy increased chromatin accessibility overall (**Fig. 2E**), which coupled with global DNA demethylation, could reactivate cryptic regulatory elements, such as promoters or enhancers.

8.2.2 Gene expression differences reflect epigenetic and viral mimicry activation.

We profiled the transcriptomic changes associated with epigenetic therapy in GSC and primary cell lines. Relative to epigenetic dynamics, samples are strongly demarcated by cell-type and then by epigenetic therapy as shown by sample-to-sample distance heatmap (**Fig. 3A**) and PCA of gene expression differences (**Fig. 3B**). Even within GSCs, the variance in PCA is equally distributed based on epitherapy condition and cell type. Many genes were up-regulated in epitherapy-treated samples compared to DMSO-treated samples (**Fig. 3C**) reflecting the potential consequence of epigenetic activation, rather than silencing, induced by epitherapy. Furthermore, gene ontology enrichment revealed that cytokine-related genes are activated in GSCs (**Fig. 3D**), consistent with previous findings that DAC can activate cell-intrinsic viral mimicry response[59,60,354]. Accordingly, we report that many viral mimicry-related genes and MHC/HLA genes are more highly up-regulated in the GSCs than primary cell lines, with quiescent primary samples showing the almost no fold change in gene expression (**Fig. 3E**). Here, we provide further support that epitherapy activates viral defense pathway in GSCs to potentially sensitize GSCs for subsequent immuno-based cancer therapies.

8.2.3 Epitherapy activates transposable elements to generate chimeric transcripts.

Across all treatment conditions and samples, we identified 464 TE promoter events involving the expression of 394 genes. Then, we asked how many TEs were specifically activated after epitherapy. In GSCs, we detected 154, 207, and 221 high-confidence TE-chimeric transcripts induced in B36, B49 and B66 respectively (**Fig. 4A**). DNA methylation levels surrounding these transcribing TEs in the epitherapy-treated samples are much lower than the levels in DMSO-treated samples (**Supplementary Fig. 1A**). Also, ATAC-seq signal focalized on reactivated TEs solely in epitherapy-treated samples, illustrating that epigenetic activation is a prerequisite for its transcriptional activity (**Supplementary Fig. 1C)**. However, majority of these TEs were also epigenetically activated in hFB and NHA after epitherapy (**Supplementary Fig. 2B**). Interestingly, 48-62% and 64-69% of epitherapy-induced TE-derived transcripts were also activated in hFB and NHA respectively. When we filtered for GSC-only activation events, the number of candidates drastically reduced to ~36 antigenic TE-chimeric transcripts (data not shown), of which most are too lowly expressed to be good sources for antigen processing and presentation on MHC molecules. In quiescent primary cells, many exapted TEs remained epigenetically dormant (**Supplementary Fig. 2B,C**) and did not generate any transcripts (**Fig. 4A**). We prioritized these quiescent-absent and epitherapy-induced TE-derived transcripts for further analysis.

We utilized two translation prediction tools, CPC2 and Kozak method, to predict the coding reading frame of the transcripts. Only candidates that generate chimeric normal, chimeric truncated, and out-of-frame peptides can provide novel amino acid sequences that can potentially be presented as immunogenic antigens. This narrowed down the candidate list to 33 to 45 TE-chimeric transcripts in GSCs, which are 1) induced by epitherapy, 2) not activated in quiescent cells, and 3) predicted to produce an antigenic peptide. Majority of these transcripts are predicted to be translated in-frame of the canonical protein based on CPC2 prediction while Kozak predicted more out-of-frame peptides (**Fig. 4B**). Although chimeric transcripts unique to each GSCs exist, the bulk of the TE-derived candidates are shared (**Fig. 4C**), highlighting the possibility of pan-GSC TI-TEA vaccine. However, it is important to note that these TI-TEAs, albeit not present in quiescent cells, are expressed in propagating NHA and hFB (**Fig. 4D**), thus can lead to devastating auto-immune consequences and must be followed up thoroughly before clinical use.

8.2.4 Treatment-induced chimeric transcripts are predicted to produce HLA-presented antigens.

With the final candidate list of treatment-induced TE-chimeric transcripts, we asked whether the translated peptide could be processed and presented on HLA molecules. Since each GSC has its own unique set of HLA alleles, seq2HLA[308] program was run to predict the 4-digit HLA allele code for every GSC line using their RNA-sequencing data. Then we utilized NetMHCPan-4.0[51] to predict the binding affinity of the TE-chimeric peptides to the GSC-specific HLA alleles. For each GSC, we identified over 20 treatment-induced TE-derived peptides that are predicted to bind strongly to various HLA alleles (**Fig. 5A,B,C**). We report numerous TI-TEAs that can be presented on HLA alleles of all three GSCs. We highlight two exceptional candidates, LTR12C-DENND3 and LTR12C-ACP6, that create in-frame chimeric peptides and are strongly expressed after epitherapy in all three GSCs (**Fig. 6A,B**). These LTR12C copies are not activated in quiescent cells thus can be promising candidates for future validation studies.

8.3 Discussion

Here, we explore the synergistic potential of combining TE biology and epigenetic therapy to generate novel antigens for immunotherapy applications in GSCs. We treated three GSC lines with DNMTi (DAC) and HDACi (Panobinostat) to investigate whether cryptic promoters in TEs can be epigenetically reactivated to express TE-chimeric transcripts. These chimeric transcripts could then be translated into peptides with novel amino acids derived from TE sequences. Indeed, we detected a couple hundred epitherapy-induced TE-derived transcripts in each GSC line. However, whether these TE resurrection events also occurred in normal tissue and cells is currently unknown. To address this issue, we used two primary cell lines, adult human fibroblasts and normal human astrocytes, as controls for this study. Considering that majority of the cells in the body are either quiescent or replicating, we profiled the impact of epitherapy on proliferating cells and contact-inhibited quiescent cells. Importantly, more than half of treatment-induced TE-derived transcripts were also expressed in proliferating primary cells after epitherapy. However, in quiescent cells, more than 70% of TE candidates identified above are transcriptionally silent, which can be attributed to inefficient epigenetic activation by DAC. Furthermore, we computationally predicted that many of the quiescent-absent, treatment-induced candidates could be translated into novel out-of-frame peptides or chimeric in-frame proteins and be processed into HLA-bound antigens. Currently, we are validating the presence of epitherapy-induced TE-derived peptide products through Western blot techniques. Furthermore, we will perform HLA-pulldown on epitherapy-treated and DMSO-treated GSC samples to prove that TI-TEAs can be presented on GSC HLA molecules.

To conclude, this work stresses the importance of having proper controls for testing novel antigen detection in cancer. Although TI-TEAs are attractive candidates, due to the promiscuous activation in primary cell lines after epitherapy, utilizing TI-TEAs in clinical settings could have devastating consequences from potential autoimmune side-effects. However, various strategies can be further investigated to overcome the activation of TEs in normal tissues to increase feasibility of TI-TEA application. In this study, we treated the samples with extremely high doses of epitherapy drugs to maximize TE reactivation events. It would be interesting to see if titrating the drug dose or modifying the treatment time could preferentially activate TEs in cancer cells, but not in normal cells. Another avenue could be to perform targeted delivery[344,345] of epitherapy drugs specifically to tumor to minimize TE activation in normal tissues. Another innovation involves developing precise epigenetic engineering tools to induce TE activation and TE-derived antigen production in diseased cells with precision. Currently, we are developing CRISPR-mediated targeted epigenetic technology by adapting the SUperNova tagging system (SunTag)[285–287,355] and MS2-loop system[356,357]. In the SunTag system, a dCas9 enzyme is modified to include a tail of peptide epitopes that can be recognized by single-chain variable fragment (scFv) antibodies. By fusing scFv antibodies to TET1 catalytic domain, dCas9 can now recruit TET1CD to specific locations in the presence of sgRNAs. Of the currently available DNA methylation engineering systems, SunTag outperforms others in its methylation/demethylation efficiency. For histone modifications, two MS2-stem loops were engineered into gRNA sequences. Proteins fused to MS2 bacteriophage coat protein (MCP) can be recruited to gRNA/Cas9 complex. By fusing MCP to histone modifiers, such as KRAB or p300, we can effectively modulate the histone methylation or acetylation of specific TEs[358,359]. By integrating SunTag and MS2 systems into one module (**Fig. 6**), we can robustly and specifically modulate the epigenetic landscape of candidate TEs.

8.4 Methods

8.4.1 Cell culture and chemicals

The GSC lines were established and cultured as previously described[360,361]. In brief, culture plates were treated with 0.01% poly-L-orinithine (Sigma-Aldrich, P2533) at 37°C for 20 minutes. The plates were washed twice with PBS and then treated with 1:200 diluted laminin solution (Sigma-Aldrich, L2020) at 37°C overnight. After incubation, laminin was apirated out and replaced with GSC media: Neurocult NSA media (STEMCELL Technologies, 05750), 1x Glutamax (ThermoFisher Scientific, 35050061), 0.25x Penicillin-Streptomycin (ThermoFisher Scientific, 15140122), 75ug/ml Bovine Serum Albumin (Sigma-Aldrich, A8412), 1x B-27 supplement (ThermoFisher Scientific, 17504001), 1x N-2 supplement (ThermoFisher Scientific, 17502001), 2ug/ml Heparin (Sigma-Aldrich, H3149), 20ng/ml FGF (PeproTech, 100-18B), and 20ng/ml EGF (PeproTech, 315-09). Half of media was exchanged with fresh media every two day. GSCs were harvested with Accutase (Sigma-Aldrich, A6964).

ATCC primary adult human fibroblasts (generous gift from Andrew Yoo, Washington University, St. Louis) were grown in media described previously[362]. Primary Normal Human Astrocytes (ScienCell, 1800) were grown in Astrocyte medium (ScienCell, 1801). 50% of culture media was refreshed every 2-3 days. Adult human fibroblasts were harvested with 0.25% Trypsin-EDTA (ThermoFisher Scientific, 25200056) while human astrocytes were harvested with Accutase. To induce quiescent states, primary cells were contact inhibited for 14 days before use.

For epigenetic therapy treatment, Decitabine (LC laboratories, D-3899) was dissolved in saline (0.9% NaCl) solution at 22mM concentration. Panobinostat (BioVision, 1612) was dissolved in DMSO at 2mM concentration. The epitherapy drugs were diluted in culture media and refreshed every two days.

8.4.2 Epigenomic profiling and data analysis

To generate WGBS of cell lines, we extracted genomic DNA with *Quick*-DNA Miniprep Kit (Zymo, D3024) and bisulfite converted 200-400 ng of DNA spiked with 0.5% lambda DNA using EZ DNA Methylation-Direct kit (Zymo, D5020). For WGBS, we processed the bisulfite-converted DNA with Accel-NGS Methyl-Seq DNA Library Kit (Swift Biosciences, 30024). WGBS libraries were sequenced on Illumina NovaSeq 6000 platform. The sequencing reads were aligned to hg38 genome with Bismark and CpG methylation values were calculated using bismark_methylation_extractor function[294].

omniATAC-seq libraries were generated as detailed previously[39,363]. omniATAC-seq reads were trimmed for adapter sequences and aligned to hg38 genome using bwa (bwa mem)[234]. Duplicate reads were removed with Picard MarkDuplicates. Since the ends of the reads represent Tn5 insertion locations, we processed the aligned reads by offsetting + strand reads by +4bp and – strand reads by -5bp. The offset position for each read was used as input for calling peaks with MACS2[235] using the following parameters: "-g 1.4e+9 -B –SPMR –keep-dup all –nomodel -s 75 –extsize 73 –shift -37 -p 0.01". With narrowPeak output from MACS2, we utilized irreproducible discover rate (IDR) framework[236] to generate a consensus peak file. To identify differentially accessible regions, we processed ATAC peaks with DiffBind[213] with a FDR cutoff of <0.01.

8.4.3 Transcriptomic profiling and data analysis.

mRNA-seq libraries were generated as previously described[39] and sequenced on the Illumina NextSeq platform. mRNA-seq libraries were adapter-trimmed and aligned to the hg38 genome using STAR[228]. We processed aligned reads with StringTie[229] to generate a count matrix for each gene, which was subsequently processed using DESeq2[209] to identify differentially expressed genes. DEG expression plot was generated using Volcanoplot function in DESeq2. To identify which gene ontologies are enriched in DEGs induced by epitherapy, the list of DEGs was processed by Metascape[133] for GO term enrichment.

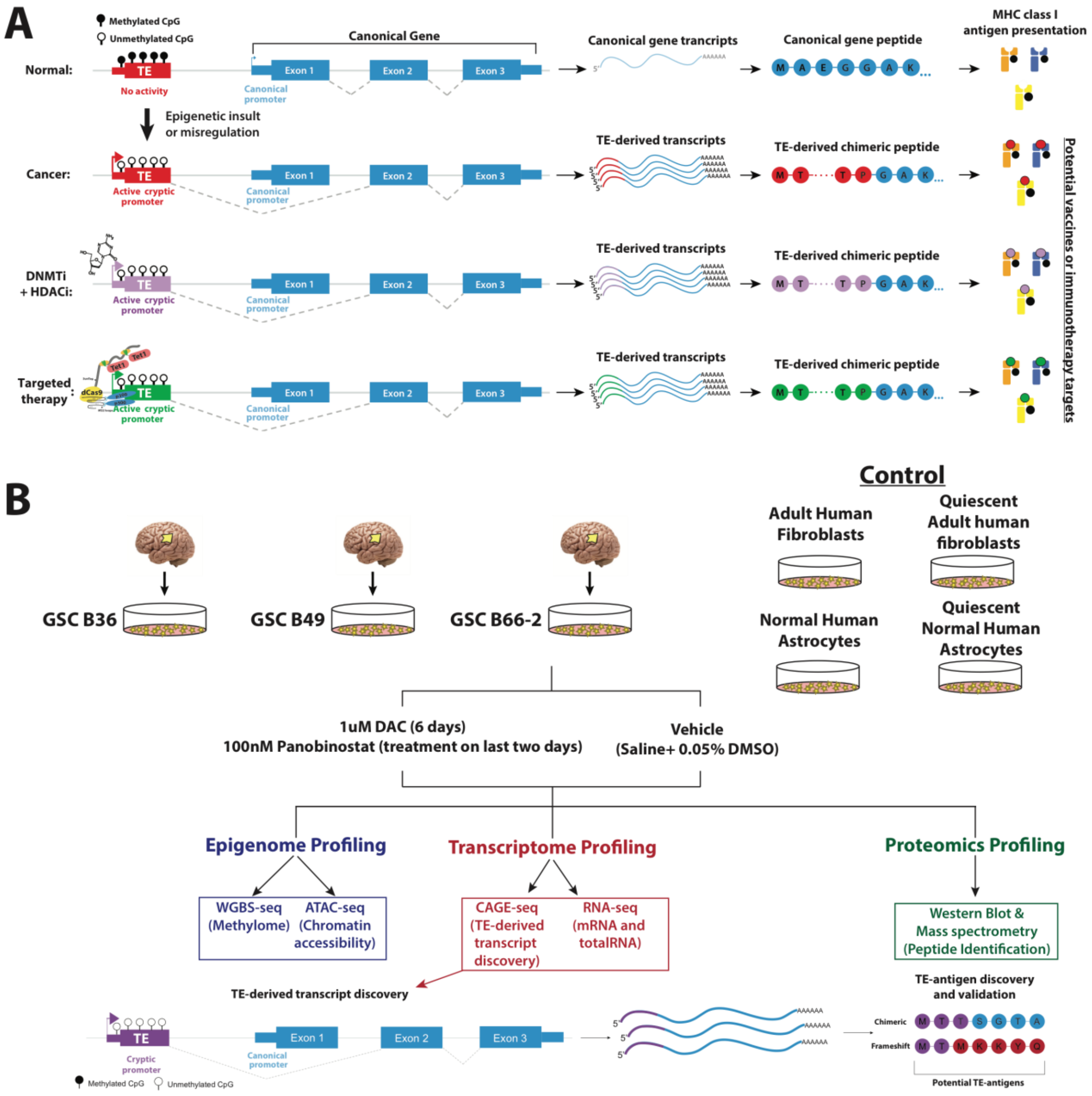
8.4.4 Detection of TE-derived transcripts and HLA-antigen presentation prediction

The pipeline for TE-chimeric transcript identification and prediction of antigen binding on HLA molecules is detailed extensively in **Chapter 7**: **Methods**.

8.5 Acknowledgments

We would like to thank J. Hoisington-López and M.L. Jaeger from The Edison Family Center for Genome Sciences & Systems Biology (CGSSB) for assistance with sequencing; B. Koebbe and E. Martin from CGSSB for data processing; This work was funded by NIH grants 5R01HG007175, U24ES026699, U01HG009391 and the American Cancer Society Research Scholar grant RSG-14-049-01-DMC. HSJ was supported by a grant from NIGMS (T32 GM007067). NMS is a Howard Hughes Medical Institute (H.H.M.I.) Medical Research Fellow, http://www.hhmi.org/. We thank Proteomics & Mass Spectrometry Facility at the Danforth Plant Science Center for its contribution and is supported by the National Science Foundation under Grant No. DBI-1827534 for acquisition of the Orbitrap Fusion Lumos LCMS/MS.

8.6 Figures



A
Methylated CpG
Unmethylated CpG
Normal:
TE
No activity
Canonical Gene
Exon 1
Exon 2
Exon 3
Canonical promoter
Canonical gene trancripts
Canonical gene peptide
M A E G G A K
MHC class I antigen presentation
Epigenetic insult or misregulation
Cancer:
Active cryptic promoter
TE-derived transcripts
TE-derived chimeric peptide
M T T P G A K
DNMTi + HDACi:
Targeted therapy
dCas9
Tet1
Potential vaccines or immunotherapy targets
B
Control
GSC B36
GSC B49
GSC B66-2
Adult Human Fibroblasts
Quiescent Adult human fibroblasts
Normal Human Astrocytes
Quiescent Normal Human Astrocytes
1uM DAC (6 days)
100nM Panobinostat (treatment on last two days)
Vehicle (Saline+ 0.05% DMSO)
Epigenome Profiling
Transcriptome Profiling
Proteomics Profiling
WGBS-seq (Methylome)
ATAC-seq (Chromatin accessibility)
CAGE-seq (TE-derived transcript discovery)
RNA-seq (mRNA and totalRNA)
Western Blot & Mass spectrometry (Peptide Identification)
TE-antigen discovery and validation
TE-derived transcript discovery
Cryptic promoter
Chimeric
M T T S G T A
Frameshift
M T M K K Y Q
Potential TE-antigens

Figure 1. Overview of TI-TEA study. A) Visual representations of various epigenetic aberrations that can activate cryptic TE promoters to generate novel immunogenic antigens. B) Experimental design for assaying the impact of epigenetic therapy (epitherapy) on the epigenetic, transcriptomic and peptidomic landscape in glioblastoma stem cells and control primary cell lines.

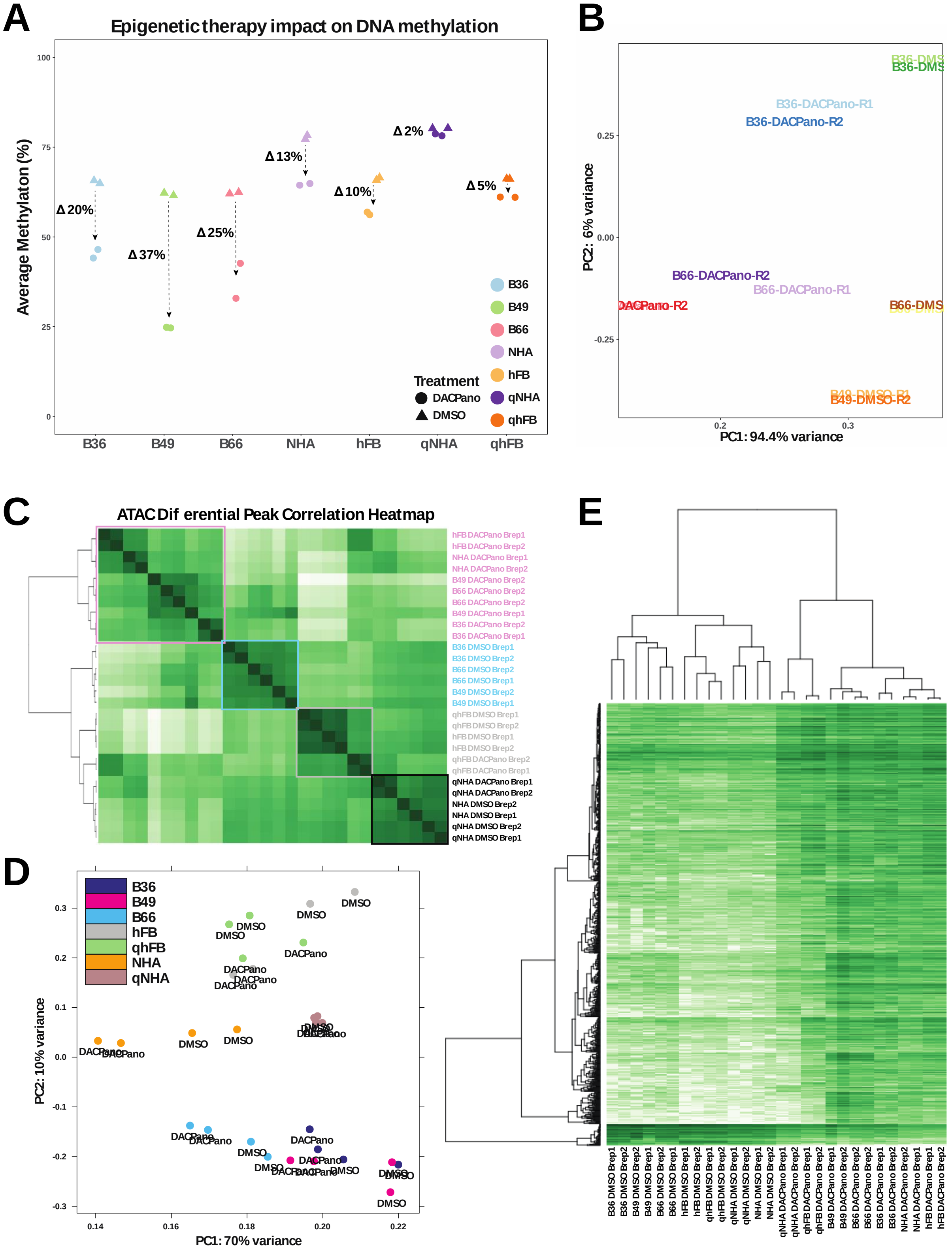

A
Epigenetic therapy impact on DNA methylation
Average Methylaton (%)
100
75
50
25
0
Δ 20%
Δ 37%
Δ 25%
Δ 13%
Δ 10%
Δ 2%
Δ 5%
B36
B49
B66
NHA
hFB
qNHA
qhFB
Treatment
● DACPano
▲ DMSO
B
PC2: 6% variance
PC1: 94.4% variance
0.25
0.00
-0.25
0.2
0.3
B36-DMS
B36-DMS
B36-DACPano-R1
B36-DACPano-R2
B66-DACPano-R2
B66-DACPano-R1
DACPano-R2
B66-DMS
B49-DMSO-R1
B49-DMSO-R2
C
ATAC Dif erential Peak Correlation Heatmap
hFB DACPano Brep1
hFB DACPano Brep2
NHA DACPano Brep1
NHA DACPano Brep2
B49 DACPano Brep2
B66 DACPano Brep2
B66 DACPano Brep2
B49 DACPano Brep1
B36 DACPano Brep2
B36 DACPano Brep1
B36 DMSO Brep1
B36 DMSO Brep2
B66 DMSO Brep2
B66 DMSO Brep1
B49 DMSO Brep2
B49 DMSO Brep1
qhFB DMSO Brep1
qhFB DMSO Brep2
hFB DMSO Brep1
hFB DMSO Brep2
qhFB DACPano Brep2
qhFB DACPano Brep1
qNHA DACPano Brep1
qNHA DACPano Brep2
NHA DMSO Brep2
NHA DMSO Brep1
qNHA DMSO Brep2
qNHA DMSO Brep1
D
B36
B49
B66
hFB
qhFB
NHA
qNHA
PC2: 10% variance
PC1: 70% variance
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
0.14
0.16
0.18
0.20
0.22
DMSO
DACPano
E
B36 DMSO Brep1
B36 DMSO Brep2
B49 DMSO Brep2
B49 DMSO Brep1
B66 DMSO Brep2
B66 DMSO Brep1
hFB DMSO Brep1
hFB DMSO Brep2
qhFB DMSO Brep1
qhFB DMSO Brep2
qNHA DMSO Brep1
qNHA DMSO Brep2
NHA DMSO Brep1
NHA DMSO Brep2
qNHA DACPano Brep1
qNHA DACPano Brep2
qhFB DACPano Brep1
qhFB DACPano Brep2
B49 DACPano Brep1
B49 DACPano Brep2
B66 DACPano Brep2
B66 DACPano Brep2
B36 DACPano Brep1
B36 DACPano Brep2
NHA DACPano Brep2
NHA DACPano Brep1
hFB DACPano Brep1
hFB DACPano Brep2

Figure 2: Epigenetic dynamics across epitherapy-treated and DMSO-treated samples. A) Average DNA methylation change between two treatments across all samples. B) Principal component (PC) analysis of DNA methylation differences in GSCs. C) Heatmap representing correlation of samples based on differential peaks identified by ATAC-seq. D) Principal component (PC) analysis of accessible chromatin dynamics across all samples. E) Heatmap representing the peak score and clustering of differential accessibility.

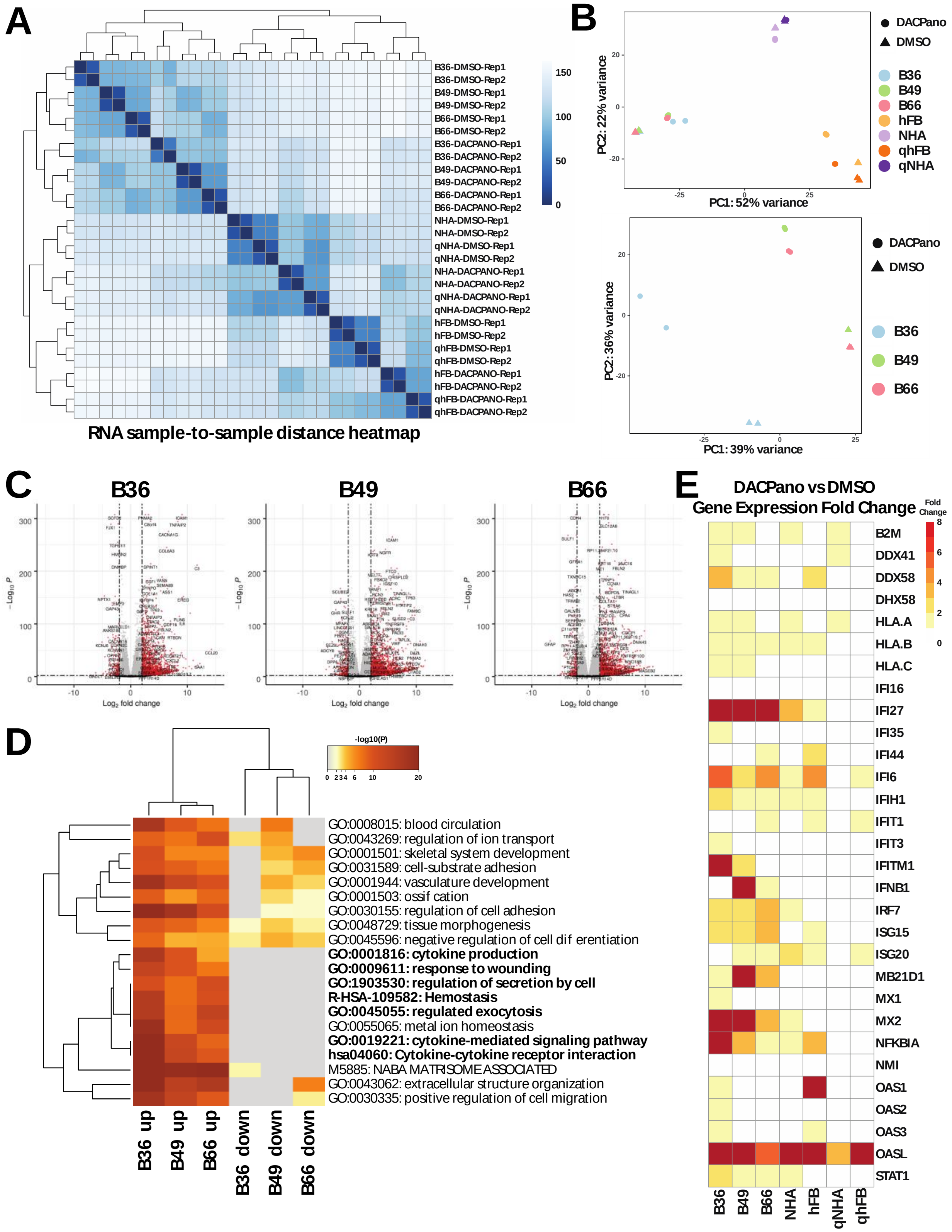

A
B36-DMSO-Rep1
B36-DMSO-Rep2
B49-DMSO-Rep1
B49-DMSO-Rep2
B66-DMSO-Rep1
B66-DMSO-Rep2
B36-DACPANO-Rep1
B36-DACPANO-Rep2
B49-DACPANO-Rep1
B49-DACPANO-Rep2
B66-DACPANO-Rep1
B66-DACPANO-Rep2
NHA-DMSO-Rep1
NHA-DMSO-Rep2
qNHA-DMSO-Rep1
qNHA-DMSO-Rep2
NHA-DACPANO-Rep1
NHA-DACPANO-Rep2
qNHA-DACPANO-Rep1
qNHA-DACPANO-Rep2
hFB-DMSO-Rep1
hFB-DMSO-Rep2
qhFB-DMSO-Rep1
qhFB-DMSO-Rep2
hFB-DACPANO-Rep1
hFB-DACPANO-Rep2
qhFB-DACPANO-Rep1
qhFB-DACPANO-Rep2
150
100
50
0
RNA sample-to-sample distance heatmap
B
PC2: 22% variance
PC1: 52% variance
20
0
-20
-25
0
25
● DACPano
▲ DMSO
B36
B49
B66
hFB
NHA
qhFB
qNHA
PC2: 36% variance
PC1: 39% variance
20
0
-20
-25
0
25
● DACPano
▲ DMSO
B36
B49
B66
C
B36
B49
B66
300
200
100
0
-10
0
10
−Log10 P
Log2 fold change
D
-log10(P)
0 2 3 4 6 10 20
GO:0008015: blood circulation
GO:0043269: regulation of ion transport
GO:0001501: skeletal system development
GO:0031589: cell-substrate adhesion
GO:0001944: vasculature development
GO:0001503: ossif cation
GO:0030155: regulation of cell adhesion
GO:0048729: tissue morphogenesis
GO:0045596: negative regulation of cell dif erentiation
GO:0001816: cytokine production
GO:0009611: response to wounding
GO:1903530: regulation of secretion by cell
R-HSA-109582: Hemostasis
GO:0045055: regulated exocytosis
GO:0055065: metal ion homeostasis
GO:0019221: cytokine-mediated signaling pathway
hsa04060: Cytokine-cytokine receptor interaction
M5885: NABA MATRISOME ASSOCIATED
GO:0043062: extracellular structure organization
GO:0030335: positive regulation of cell migration
B36 up
B49 up
B66 up
B36 down
B49 down
B66 down
E
DACPano vs DMSO
Gene Expression Fold Change
Fold Change
8
6
4
2
0
B2M
DDX41
DDX58
DHX58
HLA.A
HLA.B
HLA.C
IFI16
IFI27
IFI35
IFI44
IFI6
IFIH1
IFIT1
IFIT3
IFITM1
IFNB1
IRF7
ISG15
ISG20
MB21D1
MX1
MX2
NFKBIA
NMI
OAS1
OAS2
OAS3
OASL
STAT1
B36
B49
B66
NHA
hFB
qNHA
qhFB

Figure 3. Transcriptomic dynamics across epigenetic therapy-treated and DMSO-treated samples. A) Sample-to-sample distance heatmap based on gene expression. B) Principal components analysis based on gene expression differences for all samples and only GSC samples. C) Volcano plots representing statistically significant differentially expressed genes (DEGs). Red points indicate DEGs with $p < 0.01$ and expression fold change > 2. D) Gene Ontology (GO) enrichment of DEGs comparing epitherapy-treated and DMSO-treated GSC samples. Cytokine and immune-related GO terms are in bold. E) Heatmap displaying gene expression fold change of viral mimicry-related genes in all epitherapy-treated and DMSO-treated samples.

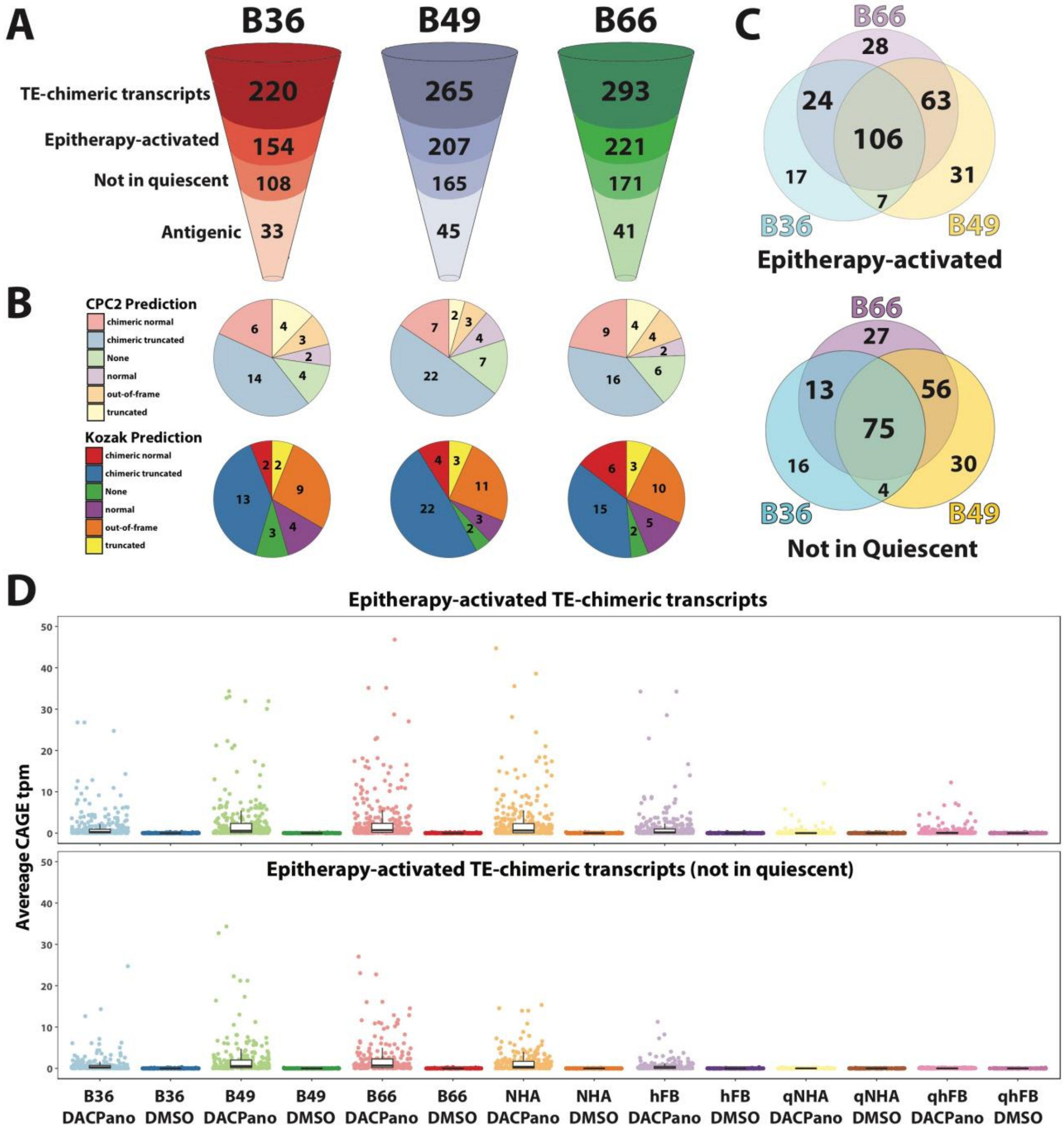


Figure 4: Detection of Epitherapy-induced TE-chimeric transcripts. A) Number of TE-chimeric transcripts after filtering. Epitherapy-activated transcripts are only present in epitherapy-treated samples. Not in quiescent filters for TE-derived transcripts that are not activated in quiescent control cells. Antigenic filters candidates that are predicted to generate chimeric peptide or out-of-frame peptides, which have potential to be presented as antigen on

MHC-I. B) Distribution of predicted translation of antigenic TE-derived transcripts using CPC2 method and Kozak method. C) Venn diagram showing number of filtered TE-derived transcripts shared across GSCs. D) Dot plot representing tpm expression of various filtered TE-derived transcripts calculated from CAGE data across all samples.

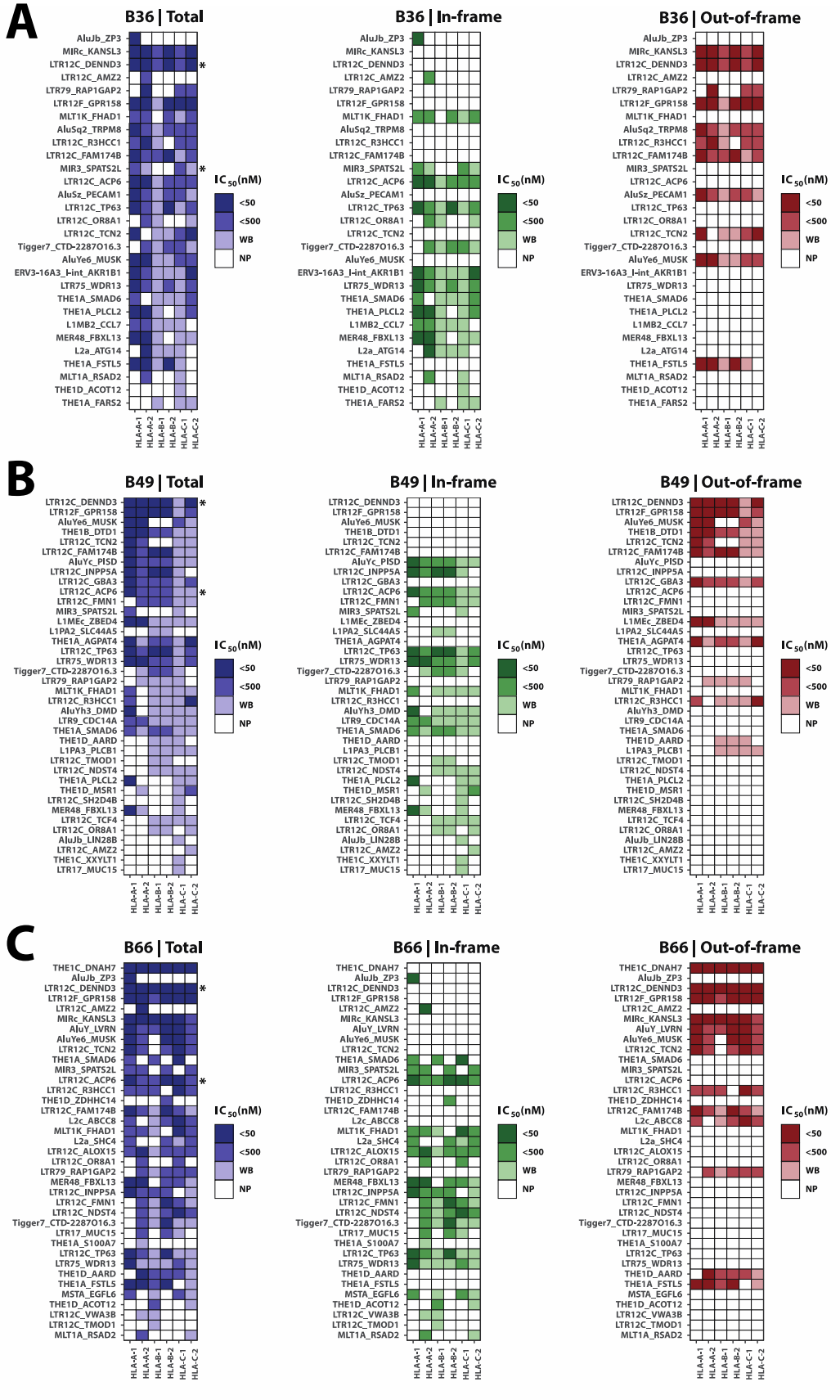

A
B36 | Total
B36 | In-frame
B36 | Out-of-frame
AluJb_ZP3
MIRc_KANSL3
LTR12C_DENND3
LTR12C_AMZ2
LTR79_RAP1GAP2
LTR12F_GPR158
MLT1K_FHAD1
AluSq2_TRPM8
LTR12C_R3HCC1
LTR12C_FAM174B
MIR3_SPATS2L
LTR12C_ACP6
AluSz_PECAM1
LTR12C_TP63
LTR12C_OR8A1
LTR12C_TCN2
Tigger7_CTD-2287O16.3
AluYe6_MUSK
ERV3-16A3_I-int_AKR1B1
LTR75_WDR13
THE1A_SMAD6
THE1A_PLCL2
L1MB2_CCL7
MER48_FBXL13
L2a_ATG14
THE1A_FSTL5
MLT1A_RSAD2
THE1D_ACOT12
THE1A_FARS2
*
*
HLA-A-1
HLA-A-2
HLA-B-1
HLA-B-2
HLA-C-1
HLA-C-2
IC50(nM)
<50
<500
WB
NP
B
B49 | Total
B49 | In-frame
B49 | Out-of-frame
LTR12C_DENND3
LTR12F_GPR158
AluYe6_MUSK
THE1B_DTD1
LTR12C_TCN2
LTR12C_FAM174B
AluYc_PISD
LTR12C_INPP5A
LTR12C_GBA3
LTR12C_ACP6
LTR12C_FMN1
MIR3_SPATS2L
L1MEc_ZBED4
L1PA2_SLC44A5
THE1A_AGPAT4
LTR12C_TP63
LTR75_WDR13
Tigger7_CTD-2287O16.3
LTR79_RAP1GAP2
MLT1K_FHAD1
LTR12C_R3HCC1
AluYh3_DMD
LTR9_CDC14A
THE1A_SMAD6
THE1D_AARD
L1PA3_PLCB1
LTR12C_TMOD1
LTR12C_NDST4
THE1A_PLCL2
THE1D_MSR1
LTR12C_SH2D4B
MER48_FBXL13
LTR12C_TCF4
LTR12C_OR8A1
AluJb_LIN28B
LTR12C_AMZ2
THE1C_XXYLT1
LTR17_MUC15
*
*
HLA-A-1
HLA-A-2
HLA-B-1
HLA-B-2
HLA-C-1
HLA-C-2
IC50(nM)
<50
<500
WB
NP
C
B66 | Total
B66 | In-frame
B66 | Out-of-frame
THE1C_DNAH7
AluJb_ZP3
LTR12C_DENND3
LTR12F_GPR158
LTR12C_AMZ2
MIRc_KANSL3
AluY_LVRN
AluYe6_MUSK
LTR12C_TCN2
THE1A_SMAD6
MIR3_SPATS2L
LTR12C_ACP6
LTR12C_R3HCC1
THE1D_ZDHHC14
LTR12C_FAM174B
L2c_ABCC8
MLT1K_FHAD1
L2a_SHC4
LTR12C_ALOX15
LTR12C_OR8A1
LTR79_RAP1GAP2
MER48_FBXL13
LTR12C_INPP5A
LTR12C_FMN1
LTR12C_NDST4
Tigger7_CTD-2287O16.3
LTR17_MUC15
THE1A_S100A7
LTR12C_TP63
LTR75_WDR13
THE1D_AARD
THE1A_FSTL5
MSTA_EGFL6
THE1D_ACOT12
LTR12C_VWA3B
LTR12C_TMOD1
MLT1A_RSAD2
*
*
HLA-A-1
HLA-A-2
HLA-B-1
HLA-B-2
HLA-C-1
HLA-C-2
IC50(nM)
<50
<500
WB
NP

Figure 5. MHC-binding prediction of epitherapy-induced, quiescent-absent and antigenic TE-derived peptides. Heatmap representation of predicted binding potential of TE-derived peptides to various MHC/HLA alleles in B36 GSC (A), B49 GSC (B) and B66 GSC (C). The heatmap is further partitioned into TE-derived peptides predicted to be translated in-frame (green) and out-of-frame (red). WB: Weak Binding, NP: Not Present.

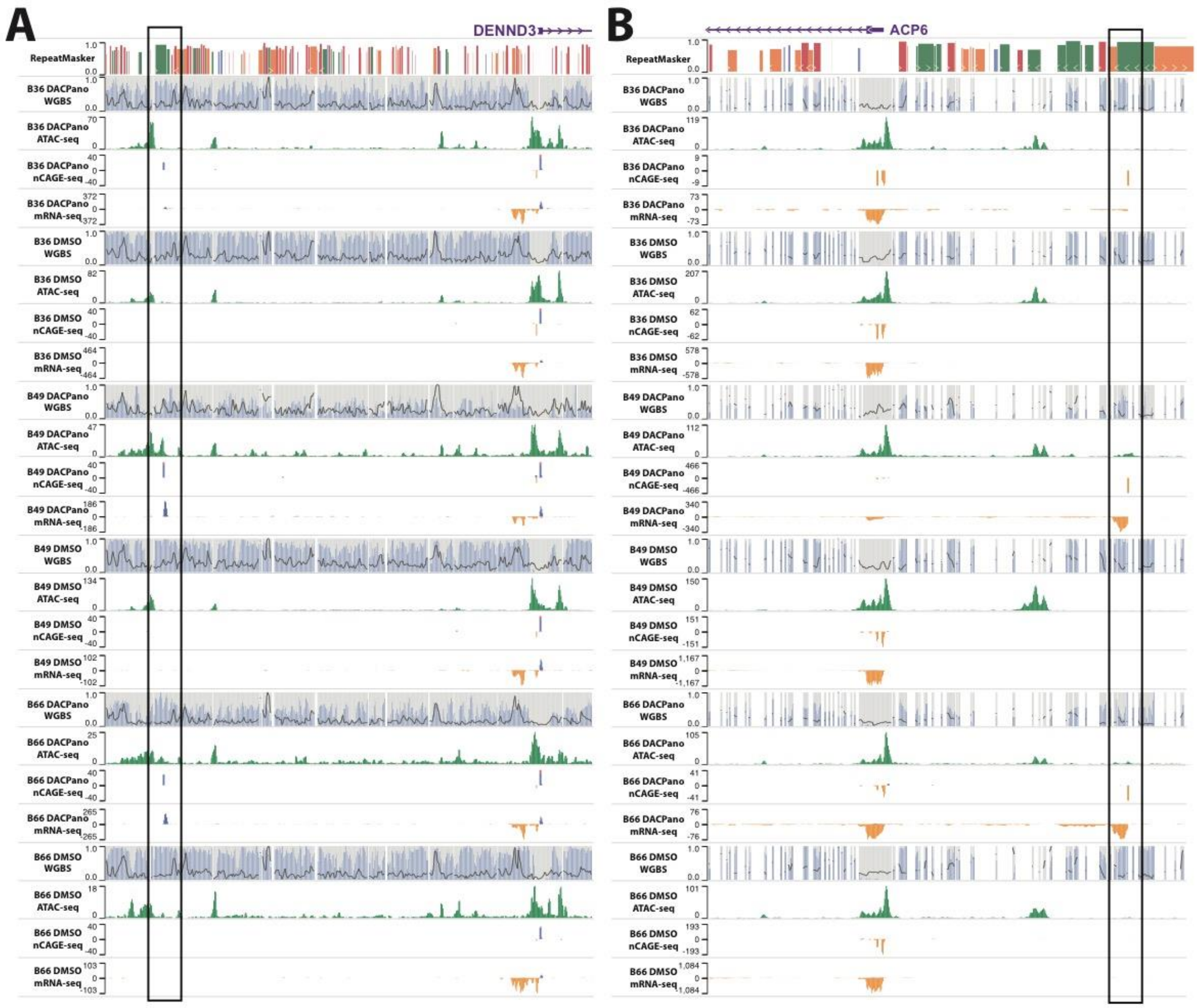


Figure 6. WashU Epigenome browser view of TE-derived transcripts predicted to be strong-binding TI-TEA in GSCs. A) WGBS, ATAC-seq, mRNA-seq and nanoCAGE-seq visualization of epitherapy-treated and DMSO-treated GSC samples of LTR12C-DENND3

candidate. Epitherapy-induced TE promoter is marked with black box. A) WGBS, ATAC-seq, mRNA-seq and nanoCAGE-seq visualization of epitherapy-treated and DMSO-treated GSC samples of LTR12C-ACP6 candidate. Epitherapy-induced TE promoter is marked with black box.

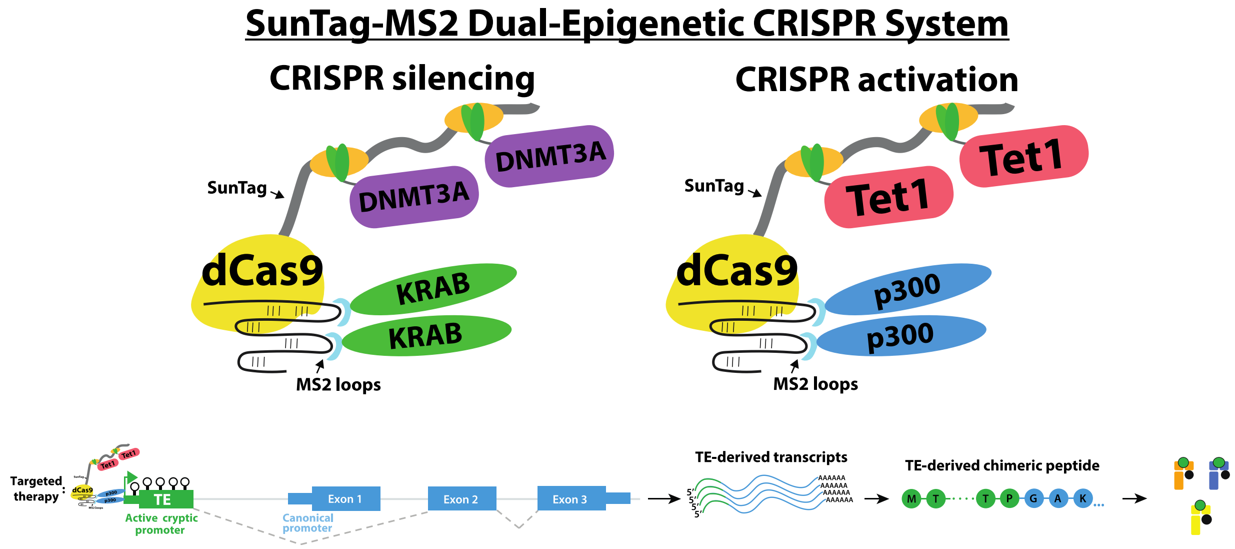
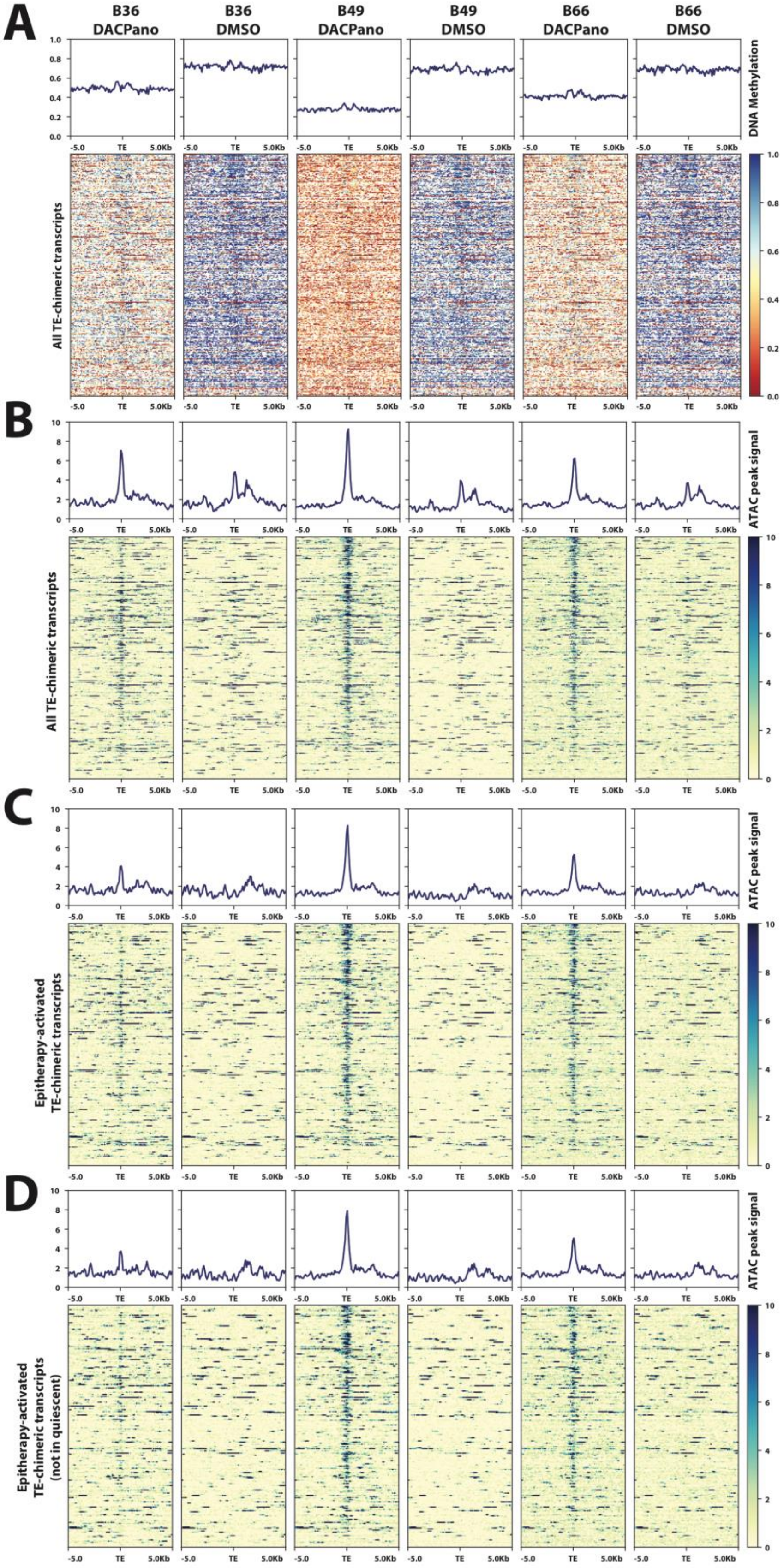
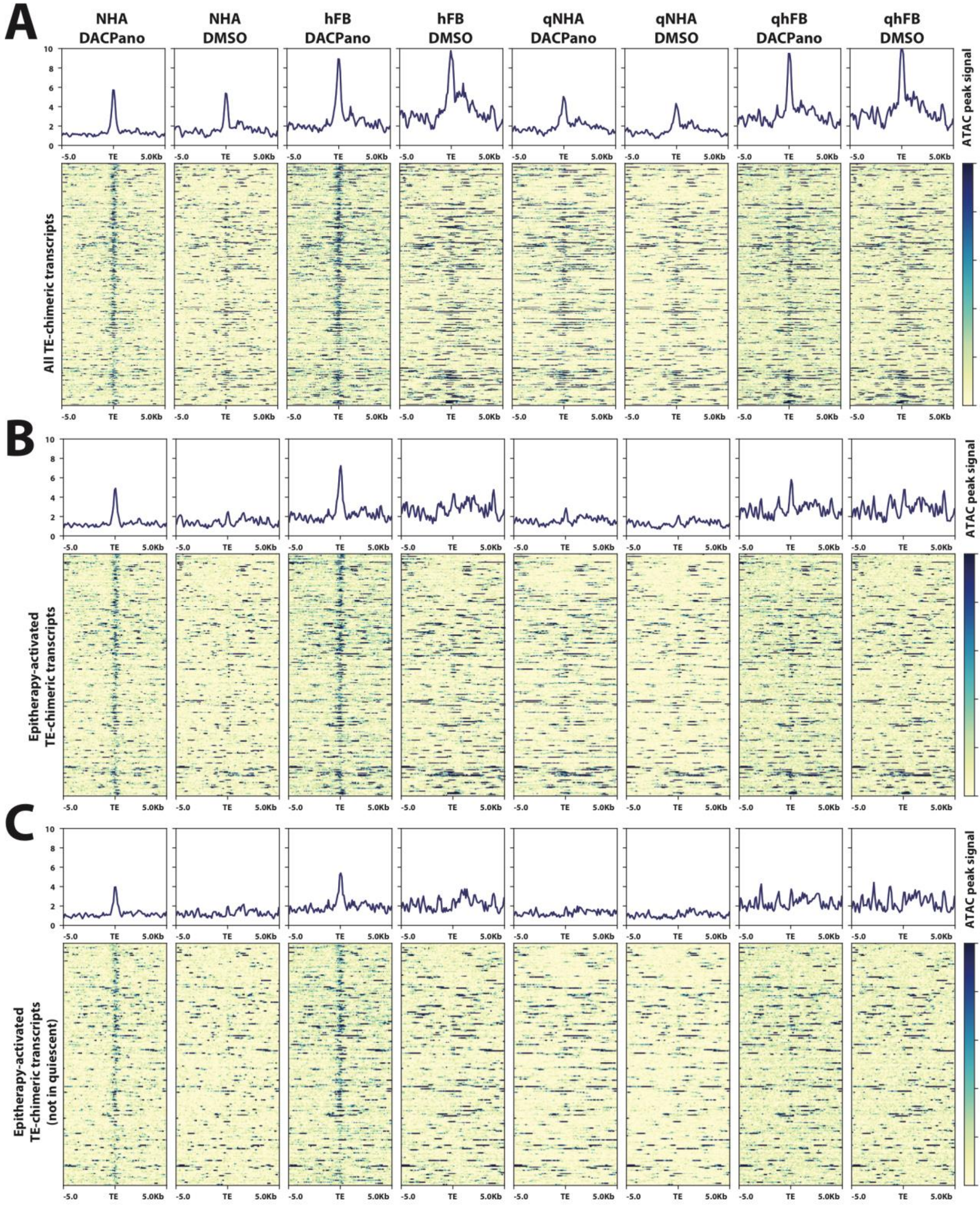


Figure 7. Schematic of dual-epigenetic CRISPR-dCas9 system utilizing SunTag and MS2 technology to perform targeted epigenetic manipulations.


A
B36 DACPano
B36 DMSO
B49 DACPano
B49 DMSO
B66 DACPano
B66 DMSO
DNA Methylation
All TE-chimeric transcripts
1.0
0.8
0.6
0.4
0.2
0.0
-5.0
TE
5.0Kb
B
ATAC peak signal
All TE-chimeric transcripts
10
8
6
4
2
0
C
ATAC peak signal
Epitherapy-activated TE-chimeric transcripts
D
ATAC peak signal
Epitherapy-activated TE-chimeric transcripts (not in quiescent)

Supplemental Figure 1: Epigenetic landscape of activated TE promoters in GSCs. A) Heatmap of DNA methylation levels over 10kb around TEs that generate chimeric transcript. Average DNA methylation levels are represented as line graphs. B) Heatmap of ATAC-seq signals representing chromatin accessibility over 10kb around TEs that generate chimeric transcript. Average ATAC-seq signals are represented as line graphs. C) Heatmap of ATAC-seq signals representing chromatin accessibility over 10kb around TEs that are induced by epitherapy. Average ATAC-seq signals are represented as line graphs. D) Heatmap of ATAC-seq signals representing chromatin accessibility over 10kb around TEs that induced by epitherapy, but not present in quiescent control cells. Average ATAC-seq signals are represented as line graphs.



Supplemental Figure 2: Chromatin landscape of activated TE promoters in control primary cell lines. A) Heatmap of ATAC-seq signals representing chromatin accessibility over

10kb around TEs that generate chimeric transcript. Average ATAC-seq signals are represented as line graphs. B) Heatmap of ATAC-seq signals representing chromatin accessibility over 10kb around TEs that are induced by epitherapy. Average ATAC-seq signals are represented as line graphs. C) Heatmap of ATAC-seq signals representing chromatin accessibility over 10kb around TEs that induced by epitherapy, but not present in quiescent control cells. Average ATAC-seq signals are represented as line graphs.

References

1. Creighton, H. & Waddington, C. H. The Strategy of the Genes. *AIBS Bull.* (1958) doi:10.2307/1291959.

2. Rajagopal, J. & Stanger, B. Z. Plasticity in the Adult: How Should the Waddington Diagram Be Applied to Regenerating Tissues? *Developmental Cell* (2016) doi:10.1016/j.devcel.2015.12.021.

3. Waddington, C. H. *The strategy of the genes: A discussion of some aspects of theoretical biology*. *The Strategy of the Genes: A Discussion of Some Aspects of Theoretical Biology* (2014). doi:10.4324/9781315765471.

4. Skinner, M. K. Role of epigenetics in developmental biology and transgenerational inheritance. *Birth Defects Research Part C - Embryo Today: Reviews* (2011) doi:10.1002/bdrc.20199.

5. Spitz, F. & Furlong, E. E. M. Transcription factors: From enhancer binding to developmental control. *Nature Reviews Genetics* (2012) doi:10.1038/nrg3207.

6. Bulger, M. & Groudine, M. Functional and mechanistic diversity of distal transcription enhancers. *Cell* (2011) doi:10.1016/j.cell.2011.01.024.

7. Levine, M. Transcriptional enhancers in animal development and evolution. *Current Biology* (2010) doi:10.1016/j.cub.2010.06.070.

8. Lenhard, B., Sandelin, A. & Carninci, P. Metazoan promoters: Emerging characteristics and insights into transcriptional regulation. *Nature Reviews Genetics* (2012) doi:10.1038/nrg3163.

9. Peter, I. S. & Davidson, E. H. Evolution of gene regulatory networks controlling body plan development. *Cell* (2011) doi:10.1016/j.cell.2011.02.017.

10. Davidson, E. H. Emerging properties of animal gene regulatory networks. *Nature* (2010) doi:10.1038/nature09645.

11. Wolffe, A. P. & Matzke, M. A. Epigenetics: Regulation through repression. *Science* (1999) doi:10.1126/science.286.5439.481.

12. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Research* (2011) doi:10.1038/cr.2011.22.

13. Berger, S. L. Histone modifications in transcriptional regulation. *Current Opinion in Genetics and Development* (2002) doi:10.1016/S0959-437X(02)00279-4.

14. Fedorova, E. & Zink, D. Nuclear architecture and gene regulation. *Biochimica et Biophysica Acta - Molecular Cell Research* (2008) doi:10.1016/j.bbamcr.2008.07.018.

15. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* (2012) doi:10.1038/nature11247.

16. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* (2015) doi:10.1038/nature14248.

17. Dekker, J. *et al.* The 4D nucleome project. *Nature* (2017) doi:10.1038/nature23884.

18. Mark D. Adams, 1 *et al.* The Genome Sequence of Drosophila melanogaster. *Genetics* **161**, 1507–1516 (2002).

19. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

20. Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).

21. Doolittle, W. F. & Sapienza, C. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**, 601–603 (1980).

22. Orgel, L. E. & Crick, F. H. C. Selfish DNA: the ultimate parasite. *Nature* **284**, 604 (1980).

23. Bennett, E. A. *et al.* Active Alu retrotransposons in the human genome. *Genome Res.* **18**, 1875–1883 (2008).

24. Callinan, a, Batzer, M. a. & Callinan, P. a. Retrotransposable Elements and Human Disease. *Genome Dyn.* **1**, 104–115 (2006).

25. Ferraccioli, G. F., Bianchi, G. & Savi, M. Similarity of the genetic background in rheumatic diseases between northern Italian and Israeli patients. *Annals of the rheumatic diseases* (1989) doi:10.1136/ard.48.1.83.

26. Hancks, D. C. & Kazazian, H. H. Active human retrotransposons: Variation and disease. *Curr. Opin. Genet. Dev.* **22**, 191–203 (2012).

27. Iskow, R. C. *et al.* Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell* **141**, 1253–1261 (2010).

28. Konkel, M. K. & Batzer, M. A. A mobile threat to genome stability: The impact of non-LTR retrotransposons upon the human genome. *Semin. Cancer Biol.* **20**, 211–221 (2010).

29. Vorechovsky, I. Transposable elements in disease-associated cryptic exons. *Hum. Genet.* **127**, 135–154 (2010).

30. Scott, E. C. *et al.* A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. *Genome Res.* **26**, 745–755 (2016).

31. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev* **16**, 6–21 (2002).

32. Morgan, H. D., Sutherland, H. G. E., Martin, D. I. K. & Whitelaw, E. Epigenetic inheritance at the agouti locus in the mouse. *Nat. Genet.* **23**, 314–318 (1999).

33. Slotkin, R. K. & Martienssen, R. Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.* **8**, 272–285 (2007).

34. Brocks, D. *et al.* DNMT and HDAC inhibitors induce cryptic transcription start sites encoded in long terminal repeats. *Nat. Genet.* **49**, 1052–1060 (2017).

35. Wiesner, T. *et al.* Alternative transcription initiation leads to expression of a novel ALK isoform in cancer. *Nature* **526**, 453–457 (2015).

36. Lamprecht, B. *et al.* Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. *Nat. Med.* **16**, 571–579 (2010).

37. Babaian, A. & Mager, D. L. Endogenous retroviral promoter exaptation in human cancer. *Mob. DNA* **7**, 24 (2016).

38. Scarfò, I. *et al.* Identification of a new subclass of ALK-negative ALCL expressing aberrant levels of ERBB4 transcripts. *Blood* **127**, 221–232 (2016).

39. Jang, H. S. *et al.* Transposable elements drive widespread expression of oncogenes in human cancers. *Nat. Genet.* **51**, 611–617 (2019).

40. Yarchoan, M., Johnson, B. A., Lutz, E. R., Laheru, D. A. & Jaffee, E. M. Targeting neoantigens to augment antitumour immunity. *Nat. Rev. Cancer* **17**, 209–222 (2017).

41. Sahin, U. & Türeci, Ö. Personalized vaccines for cancer immunotherapy. *Science (80-.).* **359**, 1355–1360 (2018).

42. Finn, O. J. The dawn of vaccines for cancer prevention. *Nat. Rev. Immunol.* **18**, 183–194 (2018).

43. Hu, Z., Ott, P. A. & Wu, C. J. Towards personalized, tumour-specific, therapeutic vaccines for cancer. *Nat. Rev. Immunol.* **18**, 168–182 (2018).

44. Rock, K. L., Reits, E. & Neefjes, J. Present Yourself! By MHC Class I and MHC Class II Molecules. *Trends Immunol.* **37**, 724–737 (2016).

45. Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. *Science (80-.).* **348**, 69–74 (2015).

46. Sahin, U. *et al.* Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature* **547**, 222–226 (2017).

47. Tanyi, J. L. *et al.* Personalized cancer vaccine effectively mobilizes antitumor T cell immunity in ovarian cancer. *Sci. Transl. Med.* **10**, 1–15 (2018).

48. Hilf, N. *et al.* Actively personalized vaccination trial for newly diagnosed glioblastoma. *Nature* **565**, 240–245 (2019).

49. Keskin, D. B. *et al.* Neoantigen vaccine generates intratumoral T cell responses in phase Ib glioblastoma trial. *Nature* **565**, 234–239 (2019).

50. Hundal, J. *et al.* pVAC-Seq: A genome-guided in silico approach to identifying tumor neoantigens. *Genome Med.* **8**, 11 (2016).

51. Jurtz, V. *et al.* NetMHCpan-4.0: Improved Peptide–MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J. Immunol.* ji1700893 (2017) doi:10.4049/jimmunol.1700893.

52. Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. *Science (80-.).* **348**, 69–74 (2015).

53. Charoentong, P. *et al.* Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell Rep.* **18**, 248–262 (2017).

54. Gao, Q. *et al.* Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. *Cell Rep.* **23**, 227-238.e3 (2018).

55. Kumar-Sinha, C., Kalyana-Sundaram, S. & Chinnaiyan, A. M. Landscape of gene fusions in epithelial cancers: seq and ye shall find. *Genome Med* **7**, 129 (2015).

56. Egger, G., Liang, G., Aparicio, A. & Jones, P. A. Epigenetics in human disease and prospects for epigenetic therapy. *Nature* **429**, 457–463 (2004).

57. Valdespino, V. & Valdespino, P. M. Potential of epigenetic therapies in the management of solid tumors. *Cancer Manag. Res.* **7**, 241–251 (2015).

58. Sharma, S., Kelly, T. K. & Jones, P. A. Epigenetics in cancer. *Carcinogenesis* vol. 31 27–36 (2009).

59. Chiappinelli, K. B. *et al.* Inhibiting DNA Methylation Causes an Interferon Response in Cancer via dsRNA Including Endogenous Retroviruses. *Cell* **162**, 974–986 (2015).

60. Roulois, D. *et al.* DNA-Demethylating Agents Target Colorectal Cancer Cells by Inducing Viral Mimicry by Endogenous Transcripts. *Cell* **162**, 961–973 (2015).

61. Nakayama, R., Ueno, Y., Ueda, K. & Honda, T. Latent infection with Kaposi's sarcoma-associated herpesvirus enhances retrotransposition of long interspersed element-1. *Oncogene* **38**, 4340–4351 (2019).

62. Leung, A. *et al.* LTRs activated by Epstein-Barr virus–induced transformation of B cells alter the transcriptome. *Genome Res.* **28**, 1791–1798 (2018).

63. Davis, M. E. Glioblastoma: Overview of disease and treatment. *Clin. J. Oncol. Nurs.* **20**, 1–8 (2016).

64. Tivnan, A., Heilinger, T., Lavelle, E. C. & Prehn, J. H. M. Advances in immunotherapy for the treatment of glioblastoma. *J. Neurooncol.* **131**, 1–9 (2017).

65. Brown, C. E. *et al.* Regression of glioblastoma after chimeric antigen receptor T-cell therapy. *N. Engl. J. Med.* **375**, 2561–2569 (2016).

66. Lim, M., Xia, Y., Bettegowda, C. & Weller, M. Current state of immunotherapy for glioblastoma. *Nat. Rev. Clin. Oncol.* **15**, 422–442 (2018).

67. Johanns, T. M. *et al.* Detection of neoantigen-specific T cells following a personalized vaccine in a patient with glioblastoma. *Oncoimmunology* **8**, 1–10 (2019).

68. Lübbert, M. & Jones, P. A. *Epigenetic therapy of cancer: Preclinical models and treatment approaches*. *Epigenetic Therapy of Cancer: Preclinical Models and Treatment Approaches* (2013). doi:10.1007/978-3-642-38404-2.

69. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14**, 204–220 (2013).

70. Schübeler, D. Function and information content of DNA methylation. *Nature* **517**, 321–326 (2015).

71. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–492 (2012).

72. Guibert, S. & Weber, M. *Functions of DNA Methylation and Hydroxymethylation in Mammalian Development*. *Current Topics in Developmental Biology* vol. 104 (Copyright &copy; 2013 Elsevier Inc. All rights reserved., 2013).

73. Yoo, C. B. & Jones, P. A. Epigenetic therapy of cancer: past, present and future. *Nat. Rev. Drug Discov.* **5**, 37–50 (2006).

74. Shen, H. & Laird, P. W. Interplay between the cancer genome and epigenome. *Cell* **153**, 38–55 (2013).

75. Baylin, S. B. & Jones, P. A. Epigenetic Determinants of Cancer. *Cold Spring Harb. Perspect. Biol.* **8**, 1–35 (2016).

76. French, C. A. *Small-Molecule Targeting of BET Proteins in Cancer*. *Advances in Cancer Research* vol. 131 (Elsevier Inc., 2016).

77. Wolf, G., Greenberg, D. & Macfarlan, T. S. Spotting the enemy within: Targeted silencing of foreign DNA in mammalian genomes by the Krüppel-associated box zinc finger protein family. *Mob. DNA* **6**, 17 (2015).

78. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).

79. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).

80. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).

81. Flavahan, W. A. *et al.* Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**, 110–114 (2015).

82. Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148**, 84–98 (2012).

83. Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci.* **112**, E6456–E6465 (2015).

84. Flavahan, W. A. *et al.* Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**, 110–114 (2015).

85. Azad, N., Zahnow, C. A., Rudin, C. M. & Baylin, S. B. The future of epigenetic therapy in solid tumours--lessons from the past. *Nat Rev Clin Oncol* **10**, 256–266 (2013).

86. Feinberg, A. P., Koldobskiy, M. A. & Göndör, A. Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. *Nat. Rev. Genet.* **17**, 284–299 (2016).

87. Baylin, S. B. & Jones, P. A. Epigenetic Determinants of Cancer. *Cold Spring Harb. Perspect. Biol.* **8**, 1–35 (2016).

88. Flavahan, W. A., Gaskell, E. & Bernstein, B. E. Epigenetic plasticity and the hallmarks of cancer. *Science (80-.).* **357**, eaal2380 (2017).

89. Flavahan, W. A., Gaskell, E. & Bernstein, B. E. Epigenetic plasticity and the hallmarks of cancer. *Science (80-.).* **357**, eaal2380 (2017).

90. Zhang, B. *et al.* Comparative DNA methylome analysis of endometrial carcinoma reveals complex and distinct deregulation of cancer promoters and enhancers. *BMC Genomics* **15**, 868 (2014).

91. Babaian, A. & Mager, D. L. Endogenous retroviral promoter exaptation in human cancer. *Mob. DNA* **7**, 24 (2016).

92. Burns, K. H. Transposable elements in cancer. *Nat. Publ. Gr.* (2017) doi:10.1038/nrc.2017.35.

93. Lupiáñez, D. G., Spielmann, M. & Mundlos, S. Breaking TADs: How Alterations of Chromatin Domains Result in Disease. *Trends Genet.* **32**, 225–237 (2016).

94. Jones, P. A., Issa, J.-P. J. & Baylin, S. Targeting the cancer epigenome for therapy. *Nat. Rev. Genet.* **17**, 630–641 (2016).

95. U.S. Department of Health and Human Services. NCI Drug Dictionary. https://www.cancer.gov/publications/dictionaries/cancer-drug.

96. Qi, Y. *et al.* HEDD: the human epigenetic drug database. *Database* **2016**, baw159 (2016).

97. Bolden, J. E., Peart, M. J. & Johnstone, R. W. Anticancer activities of histone deacetylase inhibitors. *Nat. Rev. Drug Discov.* **5**, 769–784 (2006).

98. Minucci, S. & Pelicci, P. G. Histone deacetylase inhibitors and the promise of epigenetic (and more) treatments for cancer. *Nat. Rev. Cancer* **6**, 38–51 (2006).

99. Mottamal, M., Zheng, S., Huang, T. L. & Wang, G. Histone deacetylase inhibitors in clinical studies as templates for new anticancer agents. *Molecules* **20**, 3898–3941 (2015).

100. Guha, M. HDAC inhibitors still need a home run, despite recent approval. *Nat. Rev. Drug Discov.* **14**, 225–226 (2015).

101. Shortt, J., Ott, C. J., Johnstone, R. W. & Bradner, J. E. A chemical probe toolbox for dissecting the cancer epigenome. *Nat. Rev. Cancer* **17**, 268–268 (2017).

102. Guha, M. HDAC inhibitors still need a home run, despite recent approval. *Nat. Rev. Drug Discov.* **14**, 225–226 (2015).

103. Shortt, J., Ott, C. J., Johnstone, R. W. & Bradner, J. E. Erratum: A chemical probe toolbox for dissecting the cancer epigenome (Nature reviews. Cancer (2017) 17 3 (160-183)). *Nature reviews. Cancer* vol. 17 268 (2017).

104. Minucci, S. & Pelicci, P. G. Histone deacetylase inhibitors and the promise of epigenetic (and more) treatments for cancer. *Nat. Rev. Cancer* **6**, 38–51 (2006).

105. Bolden, J. E., Peart, M. J. & Johnstone, R. W. Anticancer activities of histone deacetylase inhibitors. *Nat. Rev. Drug Discov.* **5**, 769–784 (2006).

106. Wrangle, J. *et al.* Alterations of immune response of Non-Small Cell Lung Cancer with Azacytidine. *Oncotarget* **4**, 2067–79 (2013).

107. Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. *Science (80-.).* **348**, 69–74 (2015).

108. Brocks, D. *et al.* DNMT and HDAC inhibitors induce cryptic transcription start sites encoded in long terminal repeats. *Nat. Genet.* **49**, 1052–1060 (2017).

109. Egger, G. & Arimondo, P. *Drug Discovery in Cancer Epigenetics*. (Academic Press, 2015). doi:https://doi.org/10.1016/C2014-0-02189-2.

110. Song, C. X., Yi, C. & He, C. Mapping recently identified nucleotide variants in the genome and transcriptome. *Nature Biotechnology* (2012) doi:10.1038/nbt.2398.

111. Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* (2008) doi:10.1038/nature07107.

112. Cokus, S. J. *et al.* Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* (2008) doi:10.1038/nature06745.

113. Huang, Y. *et al.* The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS One* (2010) doi:10.1371/journal.pone.0008888.

114. Booth, M. J. *et al.* Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. *Science (80-.).* (2012) doi:10.1126/science.1220671.

115. Yu, M. *et al.* Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* (2012) doi:10.1016/j.cell.2012.04.027.

116. Quy, J., Zhouy, M., Song, Q., Hong, E. E. & Smith, A. D. MLML: Consistent simultaneous estimates of DNA methylation and hydroxymethylation. *Bioinformatics* (2013) doi:10.1093/bioinformatics/btt459.

117. Xu, Z., Taylor, J. A., Leung, Y. K., Ho, S. M. & Niu, L. OxBS-MLE: An efficient method to estimate 5-methylcytosine and 5-hydroxymethylcytosine in paired bisulfite and oxidative bisulfite treated DNA. *Bioinformatics* (2016) doi:10.1093/bioinformatics/btw527.

118. Sun, D. *et al.* MOABS: Model based analysis of bisulfite sequencing data. *Genome Biol.* (2014) doi:10.1186/gb-2014-15-2-r38.

119. Äijö, T. *et al.* A probabilistic generative model for quantification of DNA modifications enables analysis of demethylation pathways. *Genome Biol.* (2016) doi:10.1186/s13059-016-0911-6.

120. Shafi, A., Mitrea, C., Nguyen, T. & Draghici, S. A survey of the approaches for identifying differential methylation using bisulfite sequencing data. *Briefings in bioinformatics* (2018) doi:10.1093/bib/bbx013.

121. Bibby, J. & Everitt, B. S. The Analysis of Contingency Tables. *Math. Gaz.* (1978) doi:10.2307/3617686.

122. Krueger, F. & Andrews, S. R. Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* (2011) doi:10.1093/bioinformatics/btr167.

123. Lister, R. *et al.* Global epigenomic reconfiguration during mammalian brain development. *Science (80-.).* **341**, (2013).

124. Hitchcock, S., Hogg, R. V. & Craig, A. T. *Introduction to Mathematical Statistics. Journal of the Royal Statistical Society. Series A (General)* vol. 129 (Pearson Education Limited, 1966).

125. Storey, J. D. A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.* (2002) doi:10.1111/1467-9868.00346.

126. Yu, M. *et al.* Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine. *Nat. Protoc.* (2012) doi:10.1038/nprot.2012.137.

127. Barter, J. D. & Foster, T. C. Aging in the Brain: New Roles of Epigenetics in Cognitive Decline. *Neuroscientist* **24**, 516–525 (2018).

128. Wen, L. *et al.* Whole-genome analysis of 5-hydroxymethylcytosine and 5-methylcytosine at base resolution in the human brain. *Genome Biol.* **15**, (2014).

129. Szulwach, K. E. *et al.* 5-hmC-mediated epigenetic dynamics during postnatal neurodevelopment and aging. *Nat. Neurosci.* **14**, 1607–1616 (2011).

130. Chen, H., Dzitoyeva, S. & Manev, H. Effect of aging on 5-hydroxymethylcytosine in the mouse hippocampus. *Restor. Neurol. Neurosci.* (2012) doi:10.3233/RNN-2012-110223.

131. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).

132. Kochmanski, J., Marchlewicz, E. H., Cavalcante, R. G., Sartor, M. A. & Dolinoy, D. C. Age-related epigenome-wide DNA methylation and hydroxymethylation in longitudinal mouse blood. *Epigenetics* **13**, 779–792 (2018).

133. Zhou, Y. *et al.* Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* **10**, (2019).

134. Lawson, N. D. & Wolfe, S. A. Forward and Reverse Genetic Approaches for the Analysis of Vertebrate Development in the Zebrafish. *Dev. Cell* **21**, 48–64 (2011).

135. Driever, W. *et al.* A genetic screen for mutations affecting embryogenesis in zebrafish. *Development* **123**, 37–46 (1996).

136. Auer, T. O. & Del Bene, F. CRISPR/Cas9 and TALEN-mediated knock-in approaches in zebrafish. *Methods* **69**, 142–150 (2014).

137. Bill, B. R., Petzold, A. M., Clark, K. J., Schimmenti, L. a & Ekker, S. C. A primer for morpholino use in zebrafish. *Zebrafish* **6**, 69–77 (2009).

138. Hruscha, A. *et al.* Efficient CRISPR/Cas9 genome editing with low off-target effects in zebrafish. *Development* **140**, 4982–7 (2013).

139. Hwang, W. Y. *et al.* Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nat Biotechnol* **31**, 227–229 (2013).

140. Lawson, N. D. Reverse Genetics in Zebrafish: Mutants, Morphants, and Moving Forward. *Trends Cell Biol.* **26**, 77–79 (2016).

141. Lossi, L. & Walker, J. M. *Neuronal Cell Death Series Editor*. (2015).

142. Xiao, A. *et al.* Chromosomal deletions and inversions mediated by TALENs and CRISPR/Cas in zebrafish. *Nucleic Acids Res.* **41**, 1–11 (2013).

143. Chang, N. *et al.* Genome editing with RNA-guided Cas9 nuclease in zebrafish embryos. *Cell Res.* **23**, 465–72 (2013).

144. Varshney, G. K. *et al.* High-throughput gene targeting and phenotyping in zebrafish using CRISPR/Cas9. *Genome Res.* **25**, 1030–1042 (2015).

145. Shah, A. N., Davey, C. F., Whitebirch, A. C., Miller, A. C. & Moens, C. B. Rapid Reverse Genetic Screening Using CRISPR in Zebrafish. *Zebrafish* **13**, 152–153 (2016).

146. Irion, U., Krauss, J. & Nusslein-Volhard, C. Precise and efficient genome editing in zebrafish using the CRISPR/Cas9 system. *Development* **141**, 4827–30 (2014).

147. Jao, L.-E., Wente, S. R. & Chen, W. Efficient multiplex biallelic zebrafish genome editing using a CRISPR nuclease system. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 13904–9 (2013).

148. Mckenna, A. *et al.* Whole organism lineage tracing by combinatorial and cumulative genome editing. **7907**, (2016).

149. Kawakami, K. Tol2: a versatile gene transfer vector in vertebrates. *Genome Biol.* **8 Suppl 1**, S7 (2007).

150. Ablain, J., Durand, E. M., Yang, S., Zhou, Y. & Zon, L. I. A CRISPR/Cas9 vector system for tissue-specific gene disruption in zebrafish. *Dev. Cell* **32**, 756–764 (2015).

151. Yin, L. *et al.* Multiplex Conditional Mutagenesis Using Transgenic Expression of Cas9 and sgRNAs. *Genetics* **200**, 431–441 (2015).

152. Hartley, J. L., Temple, G. F. & Brasch, M. A. DNA Cloning Using In Vitro Site-Specific Recombination. *Genome Res.* **10**, 1788–1795 (2000).

153. Kwan, K. M. *et al.* The Tol2kit: A multisite gateway-based construction Kit for Tol2 transposon transgenesis constructs. *Dev. Dyn.* **236**, 3088–3099 (2007).

154. Villefranc, J. A., Amigo, J. & Lawson, N. D. Gateway compatible vectors for analysis of gene function in the zebrafish. *Dev. Dyn.* **236**, 3077–3087 (2007).

155. Yin, L., Maddison, L. A. & Chen, W. *Multiplex conditional mutagenesis in zebrafish using the CRISPR/Cas system*. *Biophysical Methods in Cell Biology* vol. 135 (Elsevier Ltd).

156. Dooley, C. M. *et al.* Slc45a2 and V-ATPase are regulators of melanosomal pH homeostasis in zebrafish, providing a mechanism for human pigment evolution and disease. *Pigment Cell Melanoma Res.* **26**, 205–217 (2013).

157. Tsetskhladze, Z. R. *et al.* Functional Assessment of Human Coding Mutations Affecting Skin Pigmentation Using Zebrafish. *PLoS One* **7**, (2012).

158. Westerfield M. *The zebrafish book. A guide for the laboratory use of zebrafish (Dania rerio).* (Univ. of Oregon Press, 2007).

159. Heigwer, F., Kerr, G. & Boutros, M. E-CRISP: fast CRISPR target site identification. *Nat. Methods* **11**, 122–123 (2014).

160. Ma, H. *et al.* Pol III Promoters to Express Small RNAs: Delineation of Transcription Initiation. *Mol. Ther. Nucleic Acids* **3**, e161 (2014).

161. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–5 (2009).

162. Higdon, C. W., Mitra, R. D. & Johnson, S. L. Gene Expression Analysis of Zebrafish Melanocytes, Iridophores, and Retinal Pigmented Epithelium Reveals Indicators of Biological Function and Developmental Origin. *PLoS One* **8**, (2013).

163. Lee, H. J. *et al.* Developmental enhancers revealed by extensive DNA methylome maps of zebrafish early embryos. *Nat. Commun.* **6**, 1–13 (2015).

164. Kent, W. J. BLAT — The BLAST -Like Alignment Tool. *Genome Res.* **12**, 656–664 (2002).

165. Kroeger, P. T. *et al.* Production of haploid zebrafish embryos by in vitro fertilization. *J. Vis. Exp.* e51708 (2014) doi:10.3791/51708.

166. Choi, P. S. & Meyerson, M. Targeted genomic rearrangements using CRISPR/Cas technology. *Nat. Commun.* **5**, 3728 (2014).

167. Guo, Y. *et al.* CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell* **162**, 900–910 (2015).

168. Li, J. *et al.* Efficient inversions and duplications of mammalian regulatory DNA elements and gene clusters by CRISPR/Cas9. *J. Mol. Cell Biol.* **7**, 284–298 (2015).

169. Lee, H. J., Kweon, J., Kim, E., Kim, S. & Kim, J. Targeted chromosomal duplications and inversions in the human genome using zinc finger nucleases. 539–548 (2012) doi:10.1101/gr.129635.111.Freely.

170. Gupta, A. *et al.* Targeted chromosomal deletions and inversions in zebrafish. 1008–1017 (2013) doi:10.1101/gr.154070.112.1008.

171. Mendoza, E. & Burd, R. TYR (tyrosinase (oculocutaneous albinism IA)). *Atlas Genet. Cytogenet. Oncol. Haematol.* **16**, 918–920 (2012).

172. Simeonov, D. R. *et al.* NIH Public Access. **34**, 827–835 (2014).

173. Tryon, R. C. & Johnson, S. L. Clonal analysis of kit ligand a functional expression reveals lineage-specific competence to promote melanocyte rescue in the mutant regenerating caudal fin. *PLoS One* **9**, 1–9 (2014).

174. Tryon, R. C., Higdon, C. W. & Johnson, S. L. Lineage relationship of direct-developing melanocytes and melanocyte stem cells in the zebrafish. *PLoS One* **6**, (2011).

175. Tu, S. & Johnson, S. L. Clonal analyses reveal roles of organ founding stem cells, melanocyte stem cells and melanoblasts in establishment, growth and regeneration of the adult zebrafish fin. *Development* **137**, 3931–3939 (2010).

176. Tu, S. & Johnson, S. L. Fate restriction in the growing and regenerating zebrafish fin. *Dev. Cell* **20**, 725–732 (2011).

177. Wyvekens, N. *et al.* Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nat. Biotechnol.* **32**, 569–576 (2014).

178. Fu, Y., Sander, J. D., Reyon, D., Cascio, V. M. & Joung, J. K. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat Biotechnol* **32**, 279–284 (2014).

179. Kleinstiver, B. P. *et al.* High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490–495 (2016).

180. Levine, M. & Davidson, E. H. Gene regulatory networks for development. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 4936–4942 (2005).

181. Moris, N., Pina, C. & Arias, A. M. Transition states and cell fate decisions in epigenetic landscapes. *Nat. Rev. Genet.* **17**, 693–703 (2016).

182. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

183. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–329 (2015).

184. Meyers, J. R. Zebrafish: Development of a Vertebrate Model Organism. *Curr. Protoc. Essent. Lab. Tech.* **16**, (2018).

185. Hirata, M., Nakamura, K. I. & Hondo, S. Pigment cell distributions in different tissues of the zebrafish, with special reference to the striped pigment pattern. *Dev. Dyn.* **234**, 293–300 (2005).

186. Eisen, J. S. & Weston, J. A. Development of the neural crest in the Zebrafish. *Developmental Biology* vol. 159 50–59 (1993).

187. Lister, J. A. Development of pigment cells in the zebrafish embryo. *Microsc. Res. Tech.* **58**, 435–441 (2002).

188. Singh, A. P. *et al.* Pigment Cell Progenitors in Zebrafish Remain Multipotent through Metamorphosis. *Dev. Cell* **38**, 316–330 (2016).

189. Lister, J. A., Robertson, C. P., Lepage, T., Johnson, S. L. & Raible, D. W. Nacre Encodes a Zebrafish Microphthalmia-Related Protein That Regulates Neural-Crest-Derived Pigment Cell Fate. *Development* **126**, 3757–3767 (1999).

190. Johnson, S. L., Africa, D., Walker, C. & Weston, J. A. Genetic control of adult pigment stripe development in zebrafish. *Developmental Biology* vol. 167 27–33 (1995).

191. Petratou, K. *et al. A systems biology approach uncovers the core gene regulatory network governing iridophore fate choice from the neural crest*. *PLoS Genetics* vol. 14 (2018).

192. Rawls, J. F. *et al.* Coupled mutagenesis screens and genetic mapping in zebrafish. *Genetics* **163**, 997–1009 (2003).

193. Affecting, M. Genetic Screen for Postembryonic Development in the. **207**, 609–623 (2017).

194. Pickart, M. A. *et al.* Review : Innovative Technology Functional Genomics Tools for the Analysis of Zebrafish Pigment. *Microscopy* 461–470 (2004).

195. Elworthy, S., Lister, J. A., Carney, T. J., Raible, D. W. & Kelsh, R. N. Transcriptional regulation of mitfa accounts for the sox10 requirement in zebrafish melanophore development. *Development* **130**, 2809–2818 (2003).

196. Dorsky, R. I., Raible, D. W. & Moon, R. T. Direct regulation of nacre, a zebrafish MITF homolog required for pigment cell formation, by the Wnt pathway. *Genes Dev.* **14**, 158–162 (2000).

197. Widlund, H. R. & Fisher, D. E. Microphthalamia-associated transcription factor: A critical regulator of pigment cell development and survival. *Oncogene* **22**, 3035–3041 (2003).

198. Kimura, T., Takehana, Y. & Naruse, K. Pnp4a Is the causal gene of the medaka iridophore Mutant guanineless. *G3 Genes, Genomes, Genet.* **7**, 1357–1363 (2017).

199. Cooper, C. D. *et al.* Protein kinase A signaling inhibits iridophore differentiation in Zebrafish. *J. Dev. Biol.* **6**, (2018).

200. Mo, E. S., Cheng, Q., Reshetnyak, A. V., Schlessinger, J. & Nicoli, S. Alk and Ltk ligands are essential for iridophore development in zebrafish mediated by the receptor tyrosine kinase Ltk. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 12027–12032 (2017).

201. Curran, K., Raible, D. W. & Lister, J. A. Foxd3 controls melanophore specification in the zebrafish neural crest by regulation of Mitf. *Dev. Biol.* **332**, 408–417 (2009).

202. Curran, K. *et al.* Interplay between Foxd3 and Mitf regulates cell fate plasticity in the zebrafish neural crest. *Dev. Biol.* **344**, 107–118 (2010).

203. Spencer, S. A. The role of tfec in zebrafish neural crest cell and RPE development. *Theses Diss.* (2015).

204. Lister, J. A. The MITF paralog tfec is required in neural crest development for fate specification of the iridophore lineage from a multipotent pigment cell progenitor. (2019).

205. Hozumi, S., Shirai, M., Wang, J., Aoki, S. & Kikuchi, Y. The N-terminal domain of gastrulation brain homeobox 2 (Gbx2) is required for iridophore specification in zebrafish. *Biochem. Biophys. Res. Commun.* **502**, 104–109 (2018).

206. Wu, H. *et al.* Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic Acids Res.* **43**, 1–9 (2015).

207. Siegfried, Z. & Simon, I. DNA methylation and gene expression. *Wiley Interdiscip. Rev. Syst. Biol. Med.* (2010) doi:10.1002/wsbm.64.

208. Higdon, C. W., Mitra, R. D. & Johnson, S. L. Gene Expression Analysis of Zebrafish Melanocytes, Iridophores, and Retinal Pigmented Epithelium Reveals Indicators of Biological Function and Developmental Origin. *PLoS One* **8**, (2013).

209. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).

210. CB, K., WW, B., SR, K., B, U. & TF, S. Stages of embryonic development of the zebrafish. *Dev. Dyn.* **203**, 253–310 (1995).

211. Mayran, A. & Drouin, J. Pioneer transcription factors shape the epigenetic landscape. *Journal of Biological Chemistry* (2018) doi:10.1074/jbc.R117.001232.

212. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **2015**, 21.29.1-21.29.9 (2015).

213. Stark, R. & Brown, G. DiffBind : differential binding analysis of ChIP-Seq peak data. *Bioconductor* (2011).

214. van Otterloo, E. *et al.* Differentiation of zebrafish melanophores depends on transcription factors AP2 Alpha and AP2 Epsilon. *PLoS Genet.* **6**, (2010).

215. Pique-Regi, R. *et al.* Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.* **21**, 447–455 (2011).

216. Dee, C. T., Szymoniuk, C. R., Mills, P. E. D. & Takahashi, T. Defective neural crest migration revealed by a Zebrafish model of Alx1-related frontonasal dysplasia. *Hum. Mol. Genet.* **22**, 239–251 (2013).

217. McGonnell, I. M. *et al.* Evolution of the Alx homeobox gene family: Parallel retention and independent loss of the vertebrate Alx3 gene. *Evol. Dev.* **13**, 343–351 (2011).

218. Kayserili, H. *et al.* ALX4 dysfunction disrupts craniofacial and epidermal development. *Hum. Mol. Genet.* **18**, 4357–4366 (2009).

219. Frohnhöfer, H. G., Krauss, J., Maischein, H. M. & Nüsslein-Volhard, C. Iridophores and their interactions with other chromatophores are required for stripe formation in zebrafish. *Dev.* **140**, 2997–3007 (2013).

220. Krauss, J., Astrinides, P., Frohnhöfer, H. G., Walderich, B. & Nüsslein-Volhard, C. Transparent, a gene affecting stripe formation in Zebrafish, encodes the mitochondrial protein Mpv17 that is required for iridophore survival. *Biol. Open* **2**, 703–710 (2013).

221. Lane, B. M. & Lister, J. A. Otx but Not Mitf Transcription Factors Are Required for Zebrafish Retinal Pigment Epithelium Development. *PLoS One* **7**, (2012).

222. Iyengar, S., Houvras, Y. & Ceol, C. J. Screening for melanoma modifiers using a zebrafish autochthonous tumor model. *J. Vis. Exp.* (2012) doi:10.3791/50086.

223. Kaufman, C. K. *et al.* A zebrafish melanoma model reveals emergence of neural crest identity during melanoma initiation. *Science (80-.).* **351**, (2016).

224. Bradford, Y. *et al.* ZFIN: Enhancements and updates to the zebrafish model organism database. *Nucleic Acids Res.* **39**, 822–829 (2011).

225. Sheets, L., Ransom, D. G., Mellgren, E. M., Johnson, S. L. & Schnapp, B. J. Zebrafish Melanophilin Facilitates Melanosome Dispersion by Regulating Dynein. *Curr. Biol.* **17**, 1721–1734 (2007).

226. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* (2011) doi:10.14806/ej.17.1.200.

227. Thomer, A. K., Twidale, M. B., Guo, J. & Yoder, M. J. Picard Tools. in *Conference on Human Factors in Computing Systems - Proceedings* (2016).

228. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

229. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).

230. Anders, S., Pyl, P. T. & Huber, W. HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics* (2015) doi:10.1093/bioinformatics/btu638.

231. Kolde, R. Package `pheatmap’. *Bioconductor* (2012).

232. Zhang, H. M. *et al.* AnimalTFDB 2.0: A resource for expression, prediction and functional study of animal transcription factors. *Nucleic Acids Res.* **43**, D76–D81 (2015).

233. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* (2016) doi:10.1093/bioinformatics/btw313.

234. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. **00**, 1–3 (2013).

235. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* (2008) doi:10.1186/gb-2008-9-9-r137.

236. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* (2011) doi:10.1214/11-AOAS466.

237. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

238. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* (2010) doi:10.1016/j.molcel.2010.05.004.

239. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.* (2014) doi:10.1093/nar/gkt1249.

240. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: Scanning for occurrences of a given motif. *Bioinformatics* (2011) doi:10.1093/bioinformatics/btr064.

241. Concordet, J. P. & Haeussler, M. CRISPOR: Intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Res.* (2018) doi:10.1093/nar/gky354.

242. Moreno-Mateos, M. A. *et al.* CRISPRscan: Designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nat. Methods* **12**, 982–988 (2015).

243. Ablain, J. *et al.* Human tumor genomics and zebrafish modeling identify SPRED1 loss as a driver of mucosal melanoma. *Science (80-.).* **362**, 1055–1060 (2018).

244. Xie, M. *et al.* DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. *Nat. Genet.* **45**, 836–841 (2013).

245. Sundaram, V. *et al.* Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res.* **24**, 1963–1976 (2014).

246. Rebollo, R., Romanish, M. T. & Mager, D. L. Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes. *Annu. Rev. Genet.* **46**, 21–42 (2012).

247. Babaian, A. & Mager, D. L. Endogenous retroviral promoter exaptation in human cancer. *Mob. DNA* **7**, 1–21 (2016).

248. Botezatu, A. *et al.* Mechanisms of Oncogene Activation. in *New Aspects in Molecular and Cellular Mechanisms of Human Carcinogenesis* (ed. Bulgin, D.) 1–52 (InTech, 2016). doi:10.5772/61249.

249. Pierotti, M. A., Sozzi, G. & Croce, C. M. Mechanisms of oncogene activation. in *Holland-Frei Cancer Medicine* (ed. Kufe DW, Pollock RE, Weichselbaum RR, et al.) (BC Decker, 2003).

250. Batut, P., Dobin, A., Plessy, C., Carninci, P. & Gingeras, T. R. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Res.* **23**, 169–180 (2013).

251. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements: From conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017).

252. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science (80-.).* (2016) doi:10.1126/science.aad5497.

253. Hon, G. C. *et al.* Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome Res.* **22**, 246–258 (2012).

254. Baylin, S. B. & Jones, P. A. A decade of exploring the cancer epigenome — biological and translational implications. *Nat. Rev. Cancer* **11**, 726–734 (2011).

255. Esteller, M. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat. Rev. Genet.* **8**, 286–298 (2007).

256. Babaian, A. *et al.* Onco-exaptation of an endogenous retroviral LTR drives IRF5 expression in Hodgkin lymphoma. *Oncogene* **35**, 2542–2546 (2016).

257. Lamprecht, B. *et al.* Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. *Nat. Med.* **16**, 571–579 (2010).

258. Lock, F. E. *et al.* A novel isoform of IL-33 revealed by screening for transposable element promoted genes in human colorectal cancer. *PLoS One* **12**, 1–30 (2017).

259. Scarf, I. *et al.* Identification of a new subclass of ALK-negative ALCL expressing aberrant levels of ERBB4 transcripts. *Blood* **127**, 221–233 (2016).

260. Wiesner, T. *et al.* Alternative transcription initiation leads to expression of a novel ALK isoform in cancer. *Nature* **526**, 453–457 (2015).

261. Wolff, E. M. *et al.* Hypomethylation of a LINE-1 promoter activates an alternate transcript of the MET oncogene in bladders with cancer. *PLoS Genet.* **6**, (2010).

262. Liu, Y., Sun, J. & Zhao, M. ONGene: A literature-based database for human oncogenes. *J. Genet. Genomics* **44**, 119–121 (2017).

263. Raskin, L. *et al.* Transcriptome profiling identifies HMGA2 as a biomarker of melanoma progression and prognosis. *J. Invest. Dermatol.* **133**, 2585–2592 (2013).

264. Zhang, X., Yuan, X., Zhu, W., Qian, H. & Xu, W. SALL4: An emerging cancer biomarker and target. *Cancer Letters* (2015) doi:10.1016/j.canlet.2014.11.037.

265. Wang, T. *et al.* Aberrant regulation of the LIN28A/LIN28B and let-7 loop in human malignant tumors and its effects on the hallmarks of cancer. *Mol. Cancer* **14**, 125 (2015).

266. Nguyen, L. H. *et al.* Lin28b is sufficient to drive liver cancer and necessary for its maintenance in murine models. *Cancer Cell* **26**, 248–261 (2014).

267. Viswanathan, S. R. *et al.* Lin28 promotes transformation and is associated with advanced human malignancies. *Nat. Genet.* **41**, 843–848 (2009).

268. Babaian, A. *et al.* Onco-exaptation of an endogenous retroviral LTR drives IRF5 expression in Hodgkin lymphoma. *Oncogene* **35**, 2542–2546 (2016).

269. Forrest, A. R. R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).

270. Brocks, D. *et al.* DNMT and HDAC inhibitors induce cryptic transcription start sites encoded in long terminal repeats. *Nat. Genet.* **49**, 1052–1060 (2017).

271. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci.* **100**, 15776–15781 (2003).

272. Suzuki, A. *et al.* Aberrant transcriptional regulations in cancers: Genome, transcriptome and epigenome analysis of lung adenocarcinoma cell lines. *Nucleic Acids Res.* **42**, 13557–13572 (2014).

273. Johnson, C. D. *et al.* The let-7 microRNA represses cell proliferation pathways in human cells. *Cancer Res.* **67**, 7713–7722 (2007).

274. Newman, M. A., Thomson, J. M. & Hammond, S. M. Lin-28 interaction with the Let-7 precursor loop mediates regulated microRNA processing. *RNA* **14**, 1539–1549 (2008).

275. Zhou, J., Ng, S. B. & Chng, W. J. LIN28/LIN28B: An emerging oncogenic driver in cancer stem cells. *Int. J. Biochem. Cell Biol.* **45**, 973–978 (2013).

276. Moqtaderi, Z. *et al.* Genomic binding profiles of functionally distinct RNA polymerase III transcription complexes in human cells. *Nat. Struct. Mol. Biol.* **17**, 635–640 (2010).

277. Rice, P., Longden, L. & Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).

278. Hubley, R. *et al.* The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **44**, D81–D89 (2016).

279. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: Scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).

280. Guo, W. *et al.* A LIN28B Tumor-Specific Transcript in Cancer. *Cell Rep.* **22**, 2094–2106 (2018).

281. Beketaev, I. *et al.* cis-regulatory control of Mesp1 expression by YY1 and SP1 during mouse embryogenesis. *Dev. Dyn.* **245**, 379–387 (2016).

282. Heo, I. *et al.* Lin28 Mediates the Terminal Uridylation of let-7 Precursor MicroRNA. *Mol. Cell* **32**, 276–284 (2008).

283. Viswanathan, S. R., Daley, G. Q. & Gregory, R. I. Selective blockade of microRNA processing by Lin28. *Science (80-.).* **320**, 97–100 (2008).

284. Rybak, A. *et al.* A feedback loop comprising lin-28 and let-7 controls pre-let-7 maturation during neural stem-cell commitment. *Nat. Cell Biol.* (2008) doi:10.1038/ncb1759.

285. Tanenbaum, M. E., Gilbert, L. A., Qi, L. S., Weissman, J. S. & Vale, R. D. A protein-tagging system for signal amplification in gene expression and fluorescence imaging. *Cell* **159**, 635–646 (2014).

286. Morita, S. *et al.* Targeted DNA demethylation in vivo using dCas9-peptide repeat and scFv-TET1 catalytic domain fusions. *Nat. Biotechnol.* **34**, 1060–1065 (2016).

287. Huang, Y. H. *et al.* DNA epigenome editing using CRISPR-Cas SunTag-directed DNMT3A. *Genome Biol.* **18**, 1–11 (2017).

288. Harrow, J. *et al.* GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res.* **22**, 1760–1774 (2012).

289. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinforma.* **25**, 1–14 (2009).

290. Karolchik, D. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**, D493-496 (2004).

291. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).

292. Kang, Y. J. *et al.* CPC2: A fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.* **45**, W12–W16 (2017).

293. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

294. Krueger, F. & Andrews, S. R. Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).

295. Corces, M. R. *et al.* An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017).

296. Bayat, A., Gaëta, B., Ignjatovic, A. & Parameswaran, S. Improved VCF normalization for accurate VCF comparison. *Bioinformatics* **33**, 964–970 (2017).

297. Salimullah, M., Mizuho, S., Plessy, C. & Carninci, P. NanoCAGE: A high-resolution technique to discover and interrogate cell transcriptomes. *Cold Spring Harb. Protoc.* **6**, 96–111 (2011).

298. Haberle, V., Forrest, A. R. R., Hayashizaki, Y., Carninci, P. & Lenhard, B. CAGEr: Precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic Acids Res.* **43**, (2015).

299. Zhou, X. *et al.* The human epigenome browser at Washington University. *Nature Methods* (2011) doi:10.1038/nmeth.1772.

300. Wang, X. Primer sequences for 96 cancer-related miRNA assays. *RNA* **15**, 716–723 (2009).

301. Haeussler, M. *et al.* Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol.* **17**, 1–12 (2016).

302. Moreno-Mateos, M. A. *et al.* CRISPRscan: Designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nat. Methods* **12**, 982–988 (2015).

303. Brennan, C. W. *et al.* The somatic genomic landscape of glioblastoma. *Cell* **155**, 462–477 (2013).

304. Thorsson, V. *et al.* The Immune Landscape of Cancer. *Immunity* **48**, 812-830.e14 (2018).

305. Carithers, L. J. & Moore, H. M. The Genotype-Tissue Expression (GTEx) Project. *Biopreservation and Biobanking* (2015) doi:10.1089/bio.2015.29031.hmm.

306. Forrest, A. R. R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).

307. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385.e18 (2018).

308. Boegel, S. *et al.* HLA typing from RNA-Seq sequence reads. *Genome Med.* **4**, (2012).

309. Karosiene, E. *et al.* NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. *Immunogenetics* (2013) doi:10.1007/s00251-013-0720-y.

310. Laumont, C. M. *et al.* Noncoding regions are the main source of targetable tumor-specific antigens. *Sci. Transl. Med.* **10**, (2018).

311. Gu, S., Cui, D., Chen, X., Xiong, X. & Zhao, Y. PROTACs: An Emerging Targeting Technique for Protein Degradation in Drug Discovery. *BioEssays* (2018) doi:10.1002/bies.201700247.

312. Moreno, P. M. D. & PÃªgo, A. P. Therapeutic antisense oligonucleotides against cancer: hurdling to the clinic. *Front. Chem.* (2014) doi:10.3389/fchem.2014.00087.

313. Smith, E. L. *et al.* GPRC5D is a target for the immunotherapy of multiple myeloma with rationally designed CAR T cells. *Sci. Transl. Med.* (2019) doi:10.1126/scitranslmed.aau7746.

314. Knochelmann, H. M. *et al.* CAR T Cells in Solid Tumors: Blueprints for Building Effective Therapies. *Front. Immunol.* (2018) doi:10.3389/fimmu.2018.01740.

315. June, C. H., O’Connor, R. S., Kawalekar, O. U., Ghassemi, S. & Milone, M. C. CAR T cell immunotherapy for human cancer. *Science* (2018) doi:10.1126/science.aar6711.

316. Wilks, C., Gaddipati, P., Nellore, A. & Langmead, B. Snaptron: Querying splicing patterns across tens of thousands of RNA-seq samples. *Bioinformatics* (2018) doi:10.1093/bioinformatics/btx547.

317. Abugessaisa, I. *et al.* FANTOM5 CAGE profiles of human and mouse reprocessed for GRCh38 and GRCm38 genome assemblies. *Sci. Data* (2017) doi:10.1038/sdata.2017.107.

318. Maaten, L. van der & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* (2008) doi:10.1007/s10479-011-0841-3.

319. Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. *Science (80-.).* (2018) doi:10.1126/science.aav1898.

320. Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* (2019) doi:10.1038/s41586-019-1186-3.

321. Capes-Davis, A. *et al.* Check your cultures! A list of cross-contaminated or misidentified cell lines. *International Journal of Cancer* (2010) doi:10.1002/ijc.25242.

322. Finn, R. D. *et al.* Pfam: The protein families database. *Nucleic Acids Research* (2014) doi:10.1093/nar/gkt1223.

323. Krogh, A., Larsson, B., Von Heijne, G. & Sonnhammer, E. L. L. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.* (2001) doi:10.1006/jmbi.2000.4315.

324. Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L. J. & Mann, M. Mass Spectrometry of Human Leukocyte Antigen Class I Peptidomes Reveals Strong Effects of Protein Abundance and Turnover on Antigen Presentation. *Mol. Cell. Proteomics* (2015) doi:10.1074/mcp.m114.042812.

325. Scholtalbers, J. *et al.* TCLP: An online cancer cell line catalogue integrating HLA type, predicted neo-epitopes, virus and gene expression. *Genome Med.* (2015) doi:10.1186/s13073-015-0240-5.

326. UniProt Consortium, T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* (2018) doi:10.1093/nar/gky092.

327. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 (2016).

328. Hundal, J. *et al.* Accounting for proximal variants improves neoantigen prediction. *Nat. Genet.* (2019) doi:10.1038/s41588-018-0283-9.

329. Bassani-Sternberg, M. Mass spectrometry based immunopeptidomics for the discovery of cancer neoantigens. *Methods Mol. Biol.* **1719**, 209–221 (2018).

330. Brocks, D. *et al.* DNMT and HDAC inhibitors globally induce cryptic TSSs encoded in long terminal repeats. *Nat. Genet.* (2017).

331. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0. 2013-2015 . *http://www.repeatmasker.org* (2013).

332. Karolchik, D. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**, 493D – 496 (2004).

333. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* (2012) doi:10.1038/nprot.2012.016.

334. Kozak, M. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* (2005) doi:10.1016/j.gene.2005.06.037.

335. Pagès, H., Aboyoun, P., Gentleman, R. & DebRoy, S. Biostrings: Efficient manipulation of biological strings. *R package version 2.46.0* (2017) doi:10.1021/jm900485a.

336. Smedley, D. *et al.* BioMart - Biological queries made easy. *BMC Genomics* (2009) doi:10.1186/1471-2164-10-22.

337. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* (2013) doi:10.1093/bioinformatics/bts635.

338. Bassani-Sternberg, M. Mass spectrometry based immunopeptidomics for the discovery of cancer neoantigens. in *Methods in Molecular Biology* (2018). doi:10.1007/978-1-4939-7537-2_14.

339. Bassani-Sternberg, M. & Coukos, G. Mass spectrometry-based antigen discovery for cancer immunotherapy. *Curr. Opin. Immunol.* **41**, 9–17 (2016).

340. Morfouace, M. *et al.* Preclinical studies of 5-fluoro-2???-deoxycytidine and tetrahydrouridine in pediatric brain tumors. *J. Neurooncol.* **126**, 225–234 (2016).

341. Lee, D. H., Ryu, H. W., Won, H. R. & Kwon, S. H. Advances in epigenetic glioblastoma therapy. *Oncotarget* **8**, 18577–18589 (2017).

342. Berghauser Pont, L. M. E. *et al.* DNA damage response and anti-apoptotic proteins predict radiosensitization efficacy of HDAC inhibitors SAHA and LBH589 in patient-derived glioblastoma cells. *Cancer Lett.* **356**, 525–535 (2015).

343. Pont, L. M. E. B. *et al.* The HDAC inhibitors scriptaid and LBH589 combined with the oncolytic virus Delta24-RGD exert enhanced anti-tumor efficacy in patient-derived glioblastoma cells. *PLoS One* **10**, 1–20 (2015).

344. Singleton, W. G. *et al.* Convection enhanced delivery of panobinostat (LBH589)-loaded pluronic nano-micelles prolongs survival in the F98 rat glioma model. *Int. J. Nanomedicine* **12**, 1385–1399 (2017).

345. Yang, J., Li, Y., Zhang, T. & Zhang, X. Development of bioactive materials for glioblastoma therapy. *Bioact. Mater.* **1**, 29–38 (2016).

346. Pollard, S. M. *et al.* Glioma Stem Cell Lines Expanded in Adherent Culture Have Tumor-Specific Phenotypes and Are Suitable for Chemical and Genetic Screens. *Cell Stem Cell* **4**, 568–580 (2009).

347. Lee, J. *et al.* Tumor stem cells derived from glioblastomas cultured in bFGF and EGF more closely mirror the phenotype and genotype of primary tumors than do serum-cultured cell lines. *Cancer Cell* **9**, 391–403 (2006).

348. Aum, D. J. *et al.* Molecular and cellular heterogeneity: The hallmark of glioblastoma. *Neurosurg. Focus* **37**, 1–11 (2014).

349. Leontieva, O. V., Demidenko, Z. N. & Blagosklonny, M. V. Contact inhibition and high cell density deactivate the mammalian target of rapamycin pathway, thus suppressing the senescence program. *Proc. Natl. Acad. Sci. U. S. A.* (2014) doi:10.1073/pnas.1405723111.

350. Lemons, J. M. S. *et al.* Quiescent fibroblasts exhibit high metabolic activity. *PLoS Biol.* (2010) doi:10.1371/journal.pbio.1000514.

351. Mitra, M., Ho, L. D. & Coller, H. A. An in vitro model of cellular quiescence in primary human dermal fibroblasts. in *Methods in Molecular Biology* (2018). doi:10.1007/978-1-4939-7371-2_2.

352. Gos, M. *et al.* Cellular quiescence induced by contact inhibition or serum withdrawal in C3H10T1/2 cells. *Cell Prolif.* (2005) doi:10.1111/j.1365-2184.2005.00334.x.

353. Derissen, E. J. B., Beijnen, J. H. & Schellens, J. H. M. Concise Drug Review: Azacitidine and Decitabine. *Oncologist* (2013) doi:10.1634/theoncologist.2012-0465.

354. Liu, M. *et al.* Dual inhibition of DNA and histone methyltransferases increases viral mimicry in ovarian cancer cells. *Cancer Res.* (2018) doi:10.1158/0008-5472.CAN-17-3953.

355. Pflueger, C. *et al.* A modular dCas9-SunTag DNMT3A epigenome editing system overcomes pervasive off-target activity of direct fusion dCas9-DNMT3A constructs. *Genome Res.* (2018) doi:10.1101/gr.233049.117.

356. Konermann, S. *et al.* Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature* **517**, 583–588 (2015).

357. Fu, Y. *et al.* CRISPR-dCas9 and sgRNA scaffolds enable dual-colour live imaging of satellite sequences and repeat-enriched individual loci. *Nat. Commun.* **7**, 1–8 (2016).

358. Hilton, I. B. *et al.* Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat. Biotechnol.* **33**, 510–517 (2015).

359. Chavez, A. *et al.* Comparison of Cas9 activators in multiple species. *Nat. Methods* **13**, 563–567 (2016).

360. Gujar, A. D. *et al.* An NAD+-dependent transcriptional program governs self-renewal and radiation resistance in glioblastoma. *Proc. Natl. Acad. Sci. U. S. A.* (2016) doi:10.1073/pnas.1610921114.

361. Mao, D. D. *et al.* A CDC20-APC/SOX2 Signaling Axis Regulates Human Glioblastoma Stem-like Cells. *Cell Rep.* (2015) doi:10.1016/j.celrep.2015.05.027.

362. Richner, M., Victor, M. B., Liu, Y., Abernathy, D. & Yoo, A. S. MicroRNA-based conversion of human fibroblasts into striatal medium spiny neurons. *Nat. Protoc.* (2015) doi:10.1038/nprot.2015.102.

363. Corces, M. R. *et al.* An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* (2017) doi:10.1038/nmeth.4396.