Winter 12-15-2022

Functional Analysis of Recurrent Non-Coding Variants in Human Melanoma

Paula Maria Godoy
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Computational and Systems Biology

Dissertation Examination Committee:
Ting Wang, Chair
Charles Kaufman, Co-Chair
Christopher Maher
George Souroullas
Tychele Turner

Functional Analysis of Recurrent Non-Coding Variants in Human Melanoma

by
Paula M. Godoy

A dissertation presented to
Washington University in St. Louis
in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

December 2022
St. Louis, Missouri

Table of Contents

List of Figures

Chapter 2

Chapter 3

List of Tables

Chapter 2

Chapter 3

Acknowledgments

I am and will always be deeply grateful for the mentorship I have received from my thesis advisor, Charles Kaufman. Providing mentorship and guidance through a PhD journey is already a difficult task – remaining a stable and kind leader in a pandemic and throughout my pregnancy and postpartum journey, along with handling my other lab mates and his own family, is a testament to his patience and support. He models leadership through respect, honors individuality, and is a fountain of knowledge and wisdom. Without a doubt, I have become a better scientist and mentor because of him.

I would also like to thank Ting Wang, my thesis committee chair, who has supported my entire journey at Washington University, from interviews to graduation. He has always spoken truthfully, and I am grateful for his advice and mentorship. I'd like to thank Tychele Turner, Christopher Maher, and George Souroullas for being excellent committee members and guiding my research questions.

I am so grateful for the lab mates who have become a second family. To Rebecca, thank you for being a constant scientific and life advisor. To Eva, thank you for your partnership throughout motherhood and science. To Sophia, thank you for showing me how to be a better and more loving human. To Catie, thank you for your joyous spirit and your unrelenting support. To Megan, although I am sad our time overlapping was short, thank you for your positive friendship throughout. To Jonathon, thank you for your helpful, philosophical, and sometimes hilarious insight. I'm also grateful for the assistance provided by our lab technicians Amy White and Anna Zarov.

I am thankful for the Chancellor's Fellowship and National Science Fellowship Graduate Student Research Fellowship for supporting my Ph.D. stipend and providing supplemental funds that allowed me to go to many conferences. I would like to thank IMSD, especially Jim Skeath, for generating an inclusive environment and introducing me to some of the best scientists I know that have also become lifelong friends. Special thanks to Emilee Kotnik, Kiona Elliott, Kitra Cates, and Ryan Friedman who are fantastic human beings and scientists.

I am extremely grateful for my friends outside of graduate school: Anna Kobara, Ketty Blum, Synthia Wright, Monica Thunder, Tiffany Wu, Olivia Moralez, Kelly Scherer, Tia Ruggles, Raeann Alnas, Jocelyn Ferrara, and Casey Malone. From surprise bachelorette parties to raising children together, our friendship has remained a key source to my happiness and ability to endure.

I am so grateful for the support and love I receive every day from my parents, Tomas and Cecilia Godoy, and my sister, Sofia Godoy. To my dad who instilled in me the love of science. To my sister who lends me her confidence when I need it most. I feel especially grateful to my Mami who has visited St. Louis countless time for weeks on end to help me raise my daughter. To Hank and Walt who made the long journey by car with us to St. Louis and have been an unconditional source of joy and love. To my husband Josh for the infinite amount of support and for the partnership that makes me feel I can accomplish anything. To my baby Remi who has shifted my entire perspective, made all else feel insignificant, and has taken up so much of my time that it has also made me a much more efficient scientist. I love you all.

Paula Maria Godoy

Washington University in St. Louis

December 2022

Dedicated to my family and friends.

ABSTRACT OF THE DISSERTATION

Functional Analysis of Recurrent Non-Coding Variants in Human Melanoma

by

Paula Godoy

Doctor of Philosophy in Biology and Biomedical Sciences

Computational and Systems Biology

Washington University in St. Louis, 2022

Assistant Professor Charles Kaufman, Chair

Distinguished Professor Ting Wang, Co-Chair

Associate Professor Christopher Maher

Assistant Professor George Souroullas

Assistant Professor Tychele Turner

Worldwide incidence rates of cutaneous melanoma are increasing, and while survival rates for early stages of melanoma are high, rates drop precipitously for metastatic melanomas or those that are unable to be targeted by currently available treatments. As melanomas have a propensity to quickly metastasize, understanding the contributions of melanoma initiation remains critical for early intervention. Onset of melanoma is characterized most by mutations that stimulate mitogen-activated protein kinase (MAPK) signaling, disrupt DNA damage checkpoints, and trigger mechanisms to bypass senescence through elongation of telomeres. Additionally, in zebrafish melanoma models, the earliest cluster of melanoma-initiating cells activate expression of a neural crest reporter which remains on during the melanoma lifespan. Neural crest cells are highly multipotent and migratory stem cells that arise in early development. Bulk and single-cell RNA-sequencing have confirmed the prevalence of cells that are transcriptionally similar to neural crest

cells in human and mouse melanomas and established the importance of this lineage in initiation, metastasis, immune response, and drug evasion.

Cutaneous melanomas have one of the highest mutational loads of any cancer types. The most common protein-coding mutations occur in BRAF or NRAS, which activate MAPK signaling, and CDKN2A, PTEN, or TP53, which inactivate tumor suppressors. Even more common are non-coding mutations in the promoter of *TERT*. These non-coding variants, present in 80% of melanomas, create a novel GABPA binding site leading elevated *TERT* transcription. Emerging evidence suggests that up-regulation of TERT elongates telomeres and assists in bypassing senescence brought on by excessive MAPK signaling. Since the discovery of the TERT promoter mutation, several other functional non-coding variants have been identified in not only in melanoma but across other cancer types and diseases. As cutaneous melanomas have some of the highest mutation rates, the number of functional non-coding variants in melanoma likely remains largely uncharacterized.

Non-coding variants typically occur in or create cell-type specific enhancers. Under this assumption, we created a pipeline to identify recurrent variants in putative melanoma regulatory regions. Within these regions, we identified 140 statistically significant recurrently mutated regions, i.e. hotspots, that harbored ~2000 putative cis-regulatory variants. As we started with almost 21 million variants, the almost 10,000-fold reduction in the number of variants led to a pool of high-confidence variants for which to validate. Statistically significant variants were almost exclusively identified in promoters and more specifically at ETS transcription factor binding sites. Our pipeline identified the TERT promoter mutations as the 13th highest scoring hotspot. Through

several rounds of validation by luciferase assays and massively parallel reporter assays, we narrowed in on mutations in the promoter of CDC20 which is mutated in 39 out of 140 cutaneous melanoma samples, spanning different stages of the melanoma lifespan.

Almost all the promoter mutations in CDC20 reduced reporter activity significantly across seven melanoma cell lines, one primary melanocyte cell line, and a human embryonic kidney cell line, likely though the disruption of an ETS motif. As high CDC20 is prognostic for worse overall survival, we hypothesized that low levels of CDC20 were important for early stages of melanoma but disadvantageous at later stages. Inspecting variant allele frequencies across different stages of melanoma supported this hypothesis, demonstrating variant allele frequencies like that of BRAF and TERT, known early events, in primary melanomas. However, many of the CDC20 promoter variants were not detected in distant metastases, supporting the notion that CDC20 promoter variants are selected against in metastatic populations as high levels of CDC20 appear to be beneficial for migration. Furthermore, in samples with low levels of CDC20, we identified high levels of key neural crest transcription factors and known melanoma oncogenes including *SOX10*, an important neural crest and melanocyte lineage specifier. Previous studies have shown that increased expression of SOX10 is observed in melanoma and leads to faster melanoma formation while knock-down of SOX10 slows down melanoma onset.

We engineered a small indel in the promoter of CDC20 in a human melanoma cell line and observed decreased migration capabilities and increased expression of SOX10, among other key neural crest genes. Ongoing and future work will establish the mechanisms by which high and low CDC20 expression leads to metastasis and neural crest identity re-emergence, respectively. One

promising hypothesis is that changes in the length of the specific cell cycle stages could lead to changes in gene expression specific to differentiation pathways. For example, prolonging of the G2 phase in human embryonic stem cells upregulated pluripotency maintenance factors. Overall, we identify a non-coding variant that abolishes an ETS motif, leads to down-regulation of *CDC20*, and establishes a transcriptional program more reminiscent of early stages of melanoma as opposed to the metastatic stage. Therefore, we propose that CDC20 has a dosage-dependent effect on melanoma and that throughout the melanoma lifespan, CDC20 promoter variants confer different advantages at different stages of cancer progression.

Chapter 1: Introduction

1.1 Overview

Chapter 1 provides a summary of the molecular mechanisms driving melanoma. First, I describe the MAPK pathway and common genetic targets that lead to hyper-activity in melanoma. As lineage identity plays a key role in melanomagenesis, I describe the development and function of the melanocyte, the cell-of-origin of melanoma, and discuss aberrant genetic and epigenetic mechanisms that leverage certain attributes of the melanocyte in melanoma initiation and formation. I then discuss the transcriptional heterogeneity found in melanoma tumors and how certain signaling pathways are favorable at different stages. Lastly, as non-coding variants are the focus of this thesis, I describe the mechanisms by which non-coding variants alter transcription and provide relevant examples.

1.2 Melanoma – a deadly cancer of transformed melanocytes

Melanoma is a skin cancer that originates from pigmented skin cells called melanocytes. It is the 5th most common cancer type with almost 100,000 cases reported in 2022 (seer.cancer.gov). 90% of patients receive surgical treatment for localized or regional disease, and the 5-year relative survival is 99.5% if local and 70.6% if regional[1] (Figure 1, seer.cancer.gov). Strikingly, survival rates decrease to 31.9% if metastatic (seer.cancer.gov) which, while low, has drastically increased in the last decade due to the advent of targeted therapy and checkpoint combination therapy[1]. The targeted therapy available leverages hyperactivity of the MAPK pathway, which in half of all melanomas is caused by a mutation in BRAF, through BRAF inhibitor, dabrafenib, and MEK inhibitor, trametinib. The addition of MEK inhibitor vastly decreases the onset of acquired resistance which, when treated when dabrafenib alone, is around 5.3 months[1]. Checkpoint combination therapy improves T-cell performance by blocking pathways that negatively regulate T-cells and has led to a significant increase in the response rate. Despite advances in medical

therapy and surgical treatment, patients with wild-type *BRAF* (~50%) do not respond to BRAF inhibitors or immunotherapy and less than 18% of patients receiving one or multiple treatments are in complete remission. Therefore, there is a pressing need to further our understanding of the molecular mechanisms driving melanoma.

1.2.1 The subtypes of melanoma

Cutaneous melanoma is the most common type of melanoma in Caucasians, accounting for more than 85% of all diagnoses, and is associated with a high mutational burden due to ultraviolet radiation exposure. Cutaneous melanomas are typically found in sun-exposed skin. Acral melanomas account for 2-3% of cases in Caucasians but 50-80% of non-Caucasians and are found in sun-protected areas, such as palms and nail beds[2]. Mucosal melanomas account for 1% of all diagnoses and are typically located in the head and neck, vulva, or anus. Uveal melanomas are an ocular form of melanoma and account for 5% of diagnoses. Acral, mucosal, and uveal melanomas have much lower mutational burdens than cutaneous melanoma and, apart from iris uveal melanomas, are not associated with UV damage. Acral and mucosal melanomas also have a significantly higher number of structural variants than uveal and cutaneous melanoma, which is dominated by single point mutations[3,4].

1.3 The MAPK pathway is a common target in melanoma

1.3.1 The MAPK pathway in normal cells

The mitogen-activated protein kinase (MAPK) signaling pathway is a highly conserved signaling pathway broadly responsible for converting extracellular signals into a cellular response, often in the form of proliferation, growth, or differentiation[5]. The initial step in the MAPK pathway is binding of a growth factor to its membrane growth factor receptor. Binding activates and recruits several proteins to the plasma membrane. Activation of Ras through the exchange of GDP to GTP

initializes a phosphorylation cascade starting with phosphorylation of the serine/threonine kinase Raf, then Mek, and lastly Erk[6] Phosphorylated Erk translocates to the nucleus and activates transcription factors.

1.3.2 Genetic lesions in the MAPK pathway in melanoma

98% of melanomas have a mutation in the MAPK pathway[3] Almost half of all melanomas have mutations in *BRAF*. Of the BRAF mutations, 56% converts a valine to glutamate in the kinase domain of Braf, subsequently increasing Braf and Erk activity by 480-fold and 4.6-fold, respectively[7]. *NRAS* is the second most mutated gene in the MAPK pathway, detected in 28% of tumors[3]. Mutations have also been detected in *HRAS* and *KRAS* in less than 5% of tumors[3,8]. The third most mutated gene in the MAPK pathway is NF1, detected in 17% of melanomas[3,8,9]. As NF1 negatively regulates NRAS, loss of *NF1* leads to a similar behavior as BRAF and NRAS mutants.

In mucosal melanoma, inactivating mutations of another MAPK negative regulator SPRED1 was detected in 37% of samples[10]. KIT mutations are also found in other non-cutaneous melanoma subtypes such as acral (23%) and mucosal (15.6%)[11]. Uveal melanomas are predominantly driven by activating mutations in GNAQ/11 which indirectly activates the MAPK pathway through Protein Kinase C (PKC)[4]. Less frequent mutations are found in MAP2K1, MAP2K2, and RAC1[9,12]. The plethora of functional protein-coding mutations in the MAPK pathway underscores the importance of MAPK signaling in melanomagenesis.

1.3.3 Genetic lesions in the PI3K-Akt pathway

Ras genes can activate both the MAPK pathway and the PI3K-Akt pathway. Expression of phosphorylated Akt (p-Akt) increases throughout melanoma progression, observed in 49% of primary melanomas and 77% of metastases and is unexpectedly higher in BRAF mutants than NRAS mutants, suggesting a potential route by which BRAF activates the PI3K-Akt pathway[13,14].

Activation can occur through the MAPK signaling pathway via NRAS (~30% of melanoma cases), which activates PI3K, or KIT (~7%), which activates NRAS. Another common PI3K-Akt mutation in melanoma is loss of PTEN, which is mutated in 10-15% of melanomas and is often present at low protein levels[13]. PTEN expression is also believed to be silenced epigenetically via promoter methylation[15]. While PI3K phosphorylates PIP_2 to PIP_3, which subsequently adds a phosphate to Akt, Pten dephosphorylates this reaction, reducing PIP_3 levels and thereby decreasing Akt phosphorylation and activity. Depletion of PTEN in $BRAF^{V600E}$ mice and subsequent activation of Akt bypasses oncogene-induced senescence (reviewed below), dramatically decreasing time to melanoma metastasis to 25-50 days, unlike $BRAF^{V600E}$ mice which rarely develop melanomas[16,17]. Mutations in PREX2, a negative regulator of PTEN, were also found in 14% melanomas[18]. Because activation of NRAS already leads to PI3K signaling, NRAS and PTEN mutations are rarely found together and instead PTEN mutations most frequently co-occur with BRAF mutations[19]. Additionally, p-Akt inhibits Tsc, an mTOR inhibitor, and subsequently it was found that the mTOR signaling pathway is also activated in 73% in melanoma but only 4% of benign nevi[20].

1.3.4 Activation of the MAPK pathway leads to oncogene-induced senescence

The transformation rate of BRAF-mutated nevi to melanoma ranges from 0.0005% for men and women under 40 years of age to 0.003% for men older than 60[21], and although NRAS and BRAF mutations are found in most cutaneous melanomas, neither alone are sufficient to cause it. Additional alterations in tumor suppressors that are not detected in nevi are required for melanomagenesis[22]. In these nevi, increased activity of MAPK signaling does lead to temporary excessive growth but is arrested in part by oncogene-induced senescence (OIS). When an oncogene, such as BRAF, leads to excessive benign over-proliferation, it can be interrupted or

stopped by OIS[23]. OIS is still not completely understood but involves induction of the p53 signaling pathway by DNA damage[24]. Additionally, the proteins p14ARF and $p16^{INK4a}$, isoforms of the gene CDKN2A, are believed to play major roles in OIS by destabilizing p53-inhibiting Mdm2 and inhibiting cyclin-dependent kinases, respectively[25]. Interestingly, the most common familial melanoma syndrome results from a loss of CDKN2A[26]. Despite this, melanocytes with NRAS mutations can initiate OIS without these proteins, suggesting alternative mechanisms by which OIS is initiated in melanocytes[27]. In mice with $BRAF^{V600E}$ mutations, benign nevi emerged even in the absence of $p16^{INK4a}$,[28] In addition, one study found that it could not differentiate nevi and melanoma when comparing eight markers of senescence[29]. However, mutations in TP53 and CDKN2A are detected in 17% and 43% of samples, respectively (cbioportal.org), suggesting that in many melanomas OIS is bypassed through loss of at least one tumor suppressor, although this clearly does not explain all cases. Given the importance of controlling proliferation in an organism, it is perhaps not surprising that multiple overlapping and redundant mechanisms are in place to limit this process. For example, Type 1 interferons were shown to induce OIS in melanocytes[30]. Overall, OIS remains an important component of growth arrest in melanocytic lesions but the details of its underlying regulation are still debated as biological and technical inconsistences remain.

1.4 Melanocytes are the cell-of-origin of melanoma

Regardless of subtype or genetic archetype, all melanomas originate from the melanocyte. Melanomas are known to hijack cell-type specific transcriptional programs to initiate, progress, and metastasize. As the developmental precursor of the melanocyte is a multipotent stem cell with high rates of proliferation and migration, understanding what aspects of lineage identity are appropriated will contribute to the well-characterized mutational profiles of melanoma.

1.4.1 Melanocyte differentiation from the neural crest

The neural crest is an early and transient developmental structure made up of multipotent cells that will differentiate into a wide variety of cell types including neurons, glial cells, and melanocytes[31]. Specification of the neural crest begins during neurulation – expression of Msx1 and Pax3/7, which ensure neural crest formation, is turned on via input from BMP, WNT, and FGF signaling pathways at the neural plate border[32]. Establishment of pre-migratory neural crest cells, located at the dorsal neural tube, requires expression of master transcription factors of the neural crest lineage including Ets1, FoxD3, Myc, Tfap2a, and Tfap2b[31]. Wnt signaling and the expression of transcription factors Snai1/2, Twist, Sox9, and FoxD3 induces epithelial to mesenchymal transition (EMT) of neural crest cells allowing for migration throughout the embryo through downregulation of Type I epithelial cadherins and up-regulation of Type II mesenchymal cadherins[33]. Migrating neural crest cells express Sox10, a master transcription factor of the neural crest and melanocyte lineage.

Differentiation into the separate lineages is orchestrated by additional transcription factors and environmental cues before migration. There are five subtypes of neural crest cells based on migration patterns (cranial, vagal, trunk, cardiac, and sacral) and each give rise to specific cell lineages. Interestingly, the melanocyte lineage is unique in that it can be derived across any of the five neural crest subtypes. Melanocyte differentiation requires activation of microthphalima-associated transcription factor (MITF) by Sox10 and Pax3[34]. Sox10 is downregulated in differentiated melanocytes, as it can act unexpectedly as a repressor of Mitf targets[35]. Melanoblasts, melanocyte progenitors, express Mitf, Dct, and Pmel, melanocyte-specific markers, after delamination from the neural tube[36]. Establishment of the melanoblast requires downregulation of FoxD3 and/or Sox2 which bind to the promoter of Mitf and repress transcription[37,38]. Expression

of the receptor tyrosine kinase Kit and its ligand, Kitl, maintain melanoblast cells by ensuring proliferation and survival during migration[39].

1.4.2 Melanocyte function in the skin

Melanocytes are found sparsely throughout the basal layer of the epidermis, vastly outnumbered by keratinocytes at a ratio of 1:10, and at hair follicle and sebaceous glands in the skin, where they are more densely distributed[40]. Melanocyte stem cells located in the hair follicle will either differentiate and produce pigment or remain quiescent[41]. Differentiated melanocytes in the skin respond to UV exposure by producing melanin and distributing it to ~30 surrounding keratinocytes, where it is used as the nucleus' protective shield against UV radiation[42]. Melanin synthesis is regulated by DOPA-chrome tautomerase and tyrosinase and occurs in specialized organelles called melanosomes, which protect the cell from the toxic intermediates of the melanin synthesis pathway[43].

1.5 Re-emergence of neural crest identity in melanoma

The re-emergence of neural crest identity is seen not only across all subtypes of melanoma but also in neuroblastoma, which is believed to originate from the trunk neural crest lineage[44]. In melanoma, hijacking of lineage identity occurs, at least in part, via the melanocytic lineage factor MITF and a neural crest and melanocyte master transcription factor SOX10. MITF and SOX10 are at the center of a major transcriptional hub directing differentiation, proliferation, and growth[43]. SOX10 is the only transcription factor required to directly reprogram a fibroblast into a neural crest cell in vitro[45]. Mitf interacts with Sox10 to recruit chromatin remodeling complex Brg1 to multiple enhancers of genes regulated by MITF that induce cell proliferation of melanoma in vitro[46]. In support of this, overexpression of Sox10 decreases time to melanoma onset in zebrafish[47].

In a mouse model of giant congenital nevi, Sox10 was found to be expressed both in melanocytic lesions and melanomas and furthermore, silencing of Sox10 eradicated melanoma and downregulated expression of MITF[48]. Contrary to this, hypermethylation of the transcription start sites for SOX10 and downregulation of SOX10 and MITF were found in metastatic melanomas[49]. This may suggest that neural crest re-emergence is critical for melanoma initiation but not for progression and metastasis. In support of this, labelling zebrafish melanocytes with a neural crest marker led to the observation of the clonal expansion of a transplantable melanoma tumor[47]. However, it remains unclear how SOX10 or other neural crest factors upstream of MITF become dysregulated in melanoma as MITF mutations in human melanoma have been detected in 13% of melanomas and SOX10 mutations in 8% of samples[50].

MITF amplifications were first described in melanoma cell lines and are the most common genomic amplification found in all subtypes of melanoma detected in 10% of primary melanomas[3,8,51] and 21% of metastatic melanomas but not in benign nevi[52]. Recurrent germline mutations in MITF impair SUMOylating of MITF and downstream transcriptional activity, predisposing carriers to melanoma[53]. It was also shown that in tumors without the MITF amplification, super-enhancers led to increased MITF expression[54].

The role of MITF in melanoma may be dosage-dependent. One study using a temperature sensitive *mitfa* mutant in zebrafish described oncogenic behavior of *mitf* at low levels but melanoma regression when *mitf* was depleted[55]. This could suggest that while low levels of MITF are required for oncogenic activity, complete loss of MITF leads to OIS. Indeed, Guiliano et al. (2010) found that upon silencing of MITF, a DNA damage response pathway was activated leading to senescence[56]. Low levels of MITF were found to increase invasiveness while high levels of MITF led to activation of $p21^{Cip1}$ or INK4A, hypophosphorylation of Rb1, and consequently G1

cell cycle arrest[57,58]. However, in melanoma, the INK4A promoter is commonly methylated, preventing MITF binding and reducing cell cycle exit signals[59]. Additionally, MITF was found to be hypermethylated and downregulated in malignant melanoma, suggesting an important role in initiation and/or progression but not in invasion[60]. This is further supported by STAT3-mediated silencing of MITF leading to metastasis of melanoma[61], as well as by a study that found SOX10 and MITF to be regulators of the proliferative state but not of the invasive state[62].

MITF expression is also regulated by BRAF and is higher in primary normal human melanocytes than in melanocytes with the $BRAF^{V600E}$ mutation because mutant Braf activity signals Mitf for degradation[63,64]. Braf not only targets Mitf for degradation but uses transcription factors Pax3 and Brn2 to up-regulate expression of MITF[64,65]. Similarly, Kit activation of MAPK signaling was also found to affect Mitf in similar ways; phosphorylation at two sites both increase Mitf transactivation and mark it for degradation[66]. Overall, there is strong support that melanoma leverages aspects of its developmental precursor, the highly proliferative, migratory and multipotent neural crest cell, to initiate cancer formation.

1.6 Epigenetic dysregulation in melanoma

Melanoma is likely as heavily epigenetically altered as it is genetically altered and understanding the combined effect on pathway signaling will be key to understanding and treating melanoma. The most well-studied epigenetic phenomenon is melanoma is DNA methylation. Jin et al. (2015) found substantial hypermethylation of melanomas compared to normal human melanocytes and found that methylation of *SOX10*, *KIT*, and *PAX3* correlated with downregulation in metastatic tumors[49]. Hypermethylation in melanomas may be caused by upregulation of DNA methyltransferases, supported by the observation that DNMT3A inhibition by RNA interference reduced growth and invasiveness of melanoma in mouse models[67]. Indeed, $BRAF^{V600E}$ activity

leads to widespread DNA methylation not only in melanoma but in colorectal cancer as well via recruitment or activation of DNMT3B and DNMT1, respectively[68,69]. Enhancers were found to be the most differentially methylated regions in metastatic melanoma[70]. Importantly, several tumor suppressors are also hypermethylated and silenced in melanoma, including PTEN, $P14^{ARFa}$, $P16^{INK4a}$, and CDKN1B[67].

Chromatin remodelers are also implicated in melanoma. Mutations in ARID2 and gains in copy number of ARID2 and ARID1B, both of which are components of the SWI/SNF chromatin remodeling complex, are found in melanoma[3,8]. Overexpression of the H3K9 methyltransferase SETDB1, which is recurrently amplified in melanoma, and loss of KDM2A, a lysine demethylase, dramatically accelerates melanoma onset in zebrafish melanoma models, suggesting that histone methylation is an important factor in melanomagenesis[71,72]. Epigenetic regulation by microRNAs have also been implicated in melanoma via regulation of cell cycle genes and several neural crest lineage factors such as *MITF*, *TFAP2A*, *SNAI1*, and *ZEB1/2*[67].

1.7 Transcriptional heterogeneity in melanoma

Melanomas display a high degree of both inter-tumor and intra-tumor transcriptional heterogeneity. Transcriptional changes in the TGFβ and Wnt/β-catenin signaling pathways are characteristic of tumors with and without metastatic potential, respectively[73]. Single-cell RNA-sequencing (scRNA-seq) and bulk RNA-sequencing on melanoma tumors revealed extensive inter-tumor and intra-tumor heterogeneity with similar properties as those discovered earlier[74–76]. There were two consistent subpopulations found across two independent studies: a neural crest-like subpopulation with high levels of *SOX10* and *NGFR* and a pigmented/melanocytic subpopulation with high expression of *MITF*. Other sub-populations that did not directly overlap between studies were a *SOX10*-low/*SOX9*-high group that was classified as undifferentiated, an

intermediate group between neural crest-like and pigmented populations, and a group of cells with high activity of a gene regulatory network indicative of nutrient starvation[75,76].

scRNA-seq of an additional ten melanoma cultures indeed confirmed the existence of the neural crest-like, intermediate/transitory, and undifferentiated/mesenchymal subpopulations previously identified[77]. Migration analyses of single cells confirmed the high migratory capabilities of the undifferentiated/mesenchymal population compared to the other subpopulations. Moreover, the transcriptional identity of melanomas cells is flexible; knockdown of SOX10 led to a phenotype switch from neural crest-like to mesenchymal[77].

A parallel approach comparing primary melanocytes genetically engineered to contain a series of common oncogenic variants identified seven gene regulatory networks[78]. As expected, WT primary melanocytes and those harboring either only a CDKN2A knock-out or an additional $BRAF^{V600E}$ mutation had a high melanocytic signature and were unable to form tumors *in vivo*. Activation of TERT via introduction of the TERT promoter mutation shifted the gene regulatory program from a melanocytic signature to an EMT signature. However, these cells were also unable to form tumors. Engineering loss-of-function mutations in TP53 and/or PTEN and APC, which subsequently activates the Wnt signaling pathway, led to increased expression of genes in the Myc/mTORC1, S Phase, and G2/M Phase gene regulatory networks. These cells could form tumors but only those containing loss-of-function APC mutations formed metastases in the lung[78].

Recently, a study integrating scRNA-sequencing and spatially-resolved RNA-sequencing corroborated previously known subpopulations and specifically observed the neural crest-like subpopulation densely distributed around blood vessel. These cells supported primary tumor growth but did not initiate metastasis which was found to instead be driven by a small population of mesenchymal cells with high levels of Prxx1[79]

Overall, within a single melanoma tumor may exist a multitude of transcriptionally distinct cells each with their own tumor-driving specialties. Neural crest-like and melanocytic subpopulations appear to play important roles in tumor initiation and drug resistance, while mesenchymal/undifferentiated cells are highly migratory and are likely the source of metastasis-initiating cells. Importantly, cells are not hard coded to their transcriptional state and can switch phenotypes based on extrinsic cues[77].

1.8 Non-coding mutations as cis-regulatory variants

The source of transcriptional heterogeneity in melanoma can stem from epigenetic mechanisms as discussed above or mutations in the non-coding region of the genome. Cis-regulatory variants (CRV) are non-coding mutations that affect expression of a gene on the same allele as opposed to trans-regulatory variants which effect those on separate DNA molecule. CRVs, including single nucleotide variants and small inserts and deletions, are the focus of this thesis. CRVs can alter transcriptional activity if located in regulatory regions, such as promoters or enhancers, by destroying or creating transcription factor binding sites. CRVs in 5' UTRs and 3' UTRs can affect RNA processing, and intronic variants can either effect splicing or transcriptional regulation if the CRV targets an intronic enhancer. CRVs that affect binding sites for chromatin organizing proteins, such as CTCF, can also lead to changes in the three-dimensional structure of the genome which can broadly alter transcription or binding sites for DNA methyltransferases which epigenetically repress transcription. Additionally, CRVs in regulatory RNAs can lead to trans-effects through, for example, altering the target sequencing of a microRNA. As CRVs can alter transcription through multiple mechanisms, identifying and characterizing their role in melanoma can contribute to the understanding of altered gene regulatory networks. To date, few variants have been comprehensively validated.

1.8.1 Cis-regulatory variants in melanoma

A mutation in the promoter region of TERT was detected in a case of familial melanoma and sporadic melanoma[80,81]. A C to T mutation created a novel binding site for an Ets transcription factor, GABPA, and is associated with elevated mRNA expression of TERT[8,82]. This promoter mutation is found in BRAF-only (75%), NRAS-only (72%), or NF1-only (83%) melanomas and is found in 5% of triple wild-type melanomas[8]. When comparing the size of telomeres to the presence of the TERT promoter mutation, the mutation was associated with reduced telomeres as opposed to longer telomeres, suggesting a more complex interaction between TERT, telomere length, and melanoma progression[3]. Genome engineering of the single point mutation in primary melanocytes containing a CDKN2A deletion and the $BRAF^{V600E}$ mutation conferred replicative immortality, suggesting a potential mechanism by which elevated expression of TERT overcomes OIS[78].

Since the identification of the TERT promoter mutations, several non-coding variants in melanoma have been discovered. In a study of 183 whole-genome sequenced melanomas, recurrent non-coding mutations were detected in the promoters, 3' UTRs, and 5' UTRs of several genes, some of which were predicted to alter transcription factor binding. All but 3, TERT, RNF185, and RPS27, could not be associated with altered gene expression[3]. In a later study, RPS27 promoter variants were shown to decrease reporter activity, and analysis of RPS27 transcript levels in human melanomas demonstrated a bimodal distribution of RPS27, where high levels were indicative of a more proliferative and invasive state while low levels were important for survival in low-attachment states and drug resistance[83].

A recurrent mutation in the promoter of SDHD, a tumor suppressor, in 10% of cases is predicted to disrupt ETS binding motifs, has decreased expression of the gene compared to samples without the mutations, is associated with poor prognosis, and disrupts a GABPA binding site[84,85].

Zhang et al., (2018) discovered two somatic expression quantitative trait loci (eQTL) in the enhancer of HYI in ~20% of samples and in the promoter of DAAM1 in ~16% of samples, both of which were functionally validated by reporter assay. DAAM1 overexpression increased cell invasion although the overexpression was tested in a breast cancer cell line as opposed to a melanoma cell line[85]. A cluster of mutations located in the promoter of DPH3 and OXNAD3 were determined to increase luciferase activity and were also identified in basal cell and squamous cell carcinoma[86]. An intronic variant in MX2 led to increased expression of reporter activity likely due to creation of a YY1 transcription factor binding site and was found to accelerate melanoma onset in zebrafish[87].

While hundreds of recurrent non-coding mutations have been bioinformatically identified and scored, few have been validated. Only one of the studies listed above employed a massively parallel reporter assay and few others have validated across multiple loci. Moreover, aside from the TERT promoter mutation, none of the variants have been engineered into a melanoma cell line as either a single nucleotide variant or a small deletion. Therefore, there is a pressing need to more thoroughly characterize the effect of a variant on not only reporter activity but cellular biology.

1.8.2 Cis-regulatory variants in other cancers

Cis-regulatory variants have been identified in other cancers and diseases. In T-cell acute lymphoblastic leukemia (T-ALL), a small indel approximately 10kb from TAL1 leads to the creation of a super-enhancer through generation of a MYB binding site[88]. Binding of Myb drastically increases H3K27Ac and expression of *TAL1*. *LMO1* and *LMO2*, also T-ALL oncogenes, were enriched for non-coding variants, some of which have been validated[89]. In chronic lymphoblastic leukemia, a non-coding variant in the 3' UTR of *NOTCH1* led to differential splicing[90]. This is remarkable as most variants studied are expected to alter gene expression and

shows the wide spectrum of aberrations that non-coding variants can inflict. In diffuse large B cell lymphoma (DLBCL), mutations in super-enhancers are common in most patients and are linked to lineage-specific genes[91]. A recurrent mutation in a brain-specific enhancer breaks an Oct2/4 motif and subsequently increases expression of *MYC* to drive glioma[92]. In summary, non-coding variants are common across cancer types and appear to be located in cell-type specific regulatory regions or near cancer-specific oncogenes.

1.9 Objective of thesis

The objective of this thesis is to identify recurrent and putatively functional non-coding variants detected in human melanomas, validate the function of the variant via reporter assays, and characterize how the non-coding variant alters key melanoma phenotypes. Chapter 2 details the bioinformatic pipeline used to identify statistically significant hotspots, i.e. recurrently mutated regions, in putative melanoma regulatory regions (pMRRs). We investigate common attributes associated within statistically significant hotspots and compare expression levels of the genes associated with top-scoring hotspots across multiple melanoma cohorts. We validated selected mutations via luciferase assays and massively parallel reporter assays and identified dozens of functional variants across seven melanoma cell lines. Chapter 3 focuses in on the hotspot located in the promoter of CDC20. We perform bioinformatic analyses to understand the clonality of the CDC20 promoter variants and identify co-expression networks between CDC20 and certain neural crest transcription factors. Lastly, we engineered an indel in the promoter of CDC20 in a human melanoma cell line to characterize the effect of the variant on viability, migration, and transcription.

Chapter 2: Functional validation of recurrent non-coding variants in human melanoma

Preface

This chapter has been reproduced and adapted from the following preprint:

Godoy, P. M., Zarov, A. P. & Kaufman, C. K. Functional analysis of recurrent non-coding variants in human melanoma. *Biorxiv* 2022.06.30.498319 (2022) doi:10.1101/2022.06.30.498319.

2.1 Abstract

Small nucleotide variants in non-coding regions of the genome can alter transcriptional regulation, leading to changes in gene expression which can activate oncogenic gene regulatory networks. Melanoma is heavily burdened by non-coding variants, representing over 99% of total genetic variation, including the well-characterized TERT promoter mutation. However, the compendium of regulatory non-coding variants is likely still functionally under-characterized. We developed a pipeline to identify hotspots, i.e. recurrently mutated regions, in melanoma containing putatively functional non-coding somatic variants that are located within predicted melanoma-specific regulatory regions. We identified hundreds of statistically significant hotspots, including the hotspot containing the TERT promoter variants. Using a combination of massively parallel reporter assays and luciferase assays, we validated 35 variants that displayed statistically significant differences in reporter activity compared to their WT counterparts.

2.2 Introduction

With the widespread availability of whole-genome sequencing and fewer discoveries of novel functional coding mutations, recent efforts have increasingly focused on identification and characterization of variants in the non-coding space of cancer genomes. Cis-regulatory variants (CRV) modulate transcription by altering the regulatory landscape of a gene, which in turn can lead to dysregulation of genes involved in cancer-driving pathways[93]. Identifying CRVs of interest is therefore, generally, a three-step process: (1) identification of variants by whole-genome or targeted sequencing (Chapter 2), (2) validation of variants through reporter assays and/or precise genome editing (Chapter 2), (3) and characterization of the effect of the gene targeted by the CRV on tumorigenesis or cancer cell biology (Chapter 3). For example, TERT promoter mutations were one of the earliest highly recurrent non-coding mutations identified in melanoma and are

remarkable due to both a strong activating effect and prevalence in multiple cancers[80,81,94]. Present in ~80% of cutaneous melanomas, the TERT promoter mutation creates a novel ETS motif that leads to binding of GABPA and de-repression of TERT[82]. The full extent of TERT's influence on tumorigenesis, particularly via this regulatory variant, is still emerging, including its canonical role on telomere maintenance[78,94,95]. Beyond TERT promoter variants, few other CRVs have been identified and characterized in melanoma[83–85,87,95–97]. The next most common mutations in cutaneous melanoma are coding mutations in the MAPK pathway, predominantly $BRAF^{V600E/K}$ and $NRAS^{Q61K}$, as well as loss of key tumor suppressors like TP53, PTEN, and CDKN2A, all with relatively clear canonical growth regulatory and proliferative functions and discussed in Chapter 1[3,8].

2.2.1 Next generation sequencing in melanoma

Next generation sequencing (NGS) has revolutionized our understanding of melanoma. NGS has not only extended the catalog of protein-coding mutants but also our knowledge on the role of structural variants, the mutational signature associated with UV irradiation, and our awareness of the extraordinary number of variants in the non-coding genome. Most of the roughly 3,000 melanomas that have been sequenced are sequenced through targeted approaches, such as whole-exome sequencing (WES). Notably, the Cancer Genome Atlas (TCGA) has performed WES and WGS on 333 and 34 tumors, respectively (TCGA-SKCM[8]), and the International Cancer Genomics Consortium (ICGC) performed WGS on 183 melanomas (ICGC-MELA[3]).

Apart from DNA-sequencing, the transcriptional and epigenetic landscapes have been assayed via RNA-sequencing, ChIP-sequencing, ATAC-sequencing, and bisulfite sequencing. Regions that are bound by certain histone marks indicate regulatory regions such as enhancers marked by H3K27Ac and promoters marked by H3K4me3[98]. ATAC-seq specifically targets

accessible chromatin, i.e. able to be bound by a barcode-inserting transposase, which indicates putative regulatory regions as this means transcription factors would also be able to bind[99]). DNA methylation, which can alter the regulatory potential of an enhancer or promoter, is assessed via bisulfite conversion of unmethylated cytosines to uracil (Reinders et al., 2008).

2.2.2 Bioinformatic methods to detect putatively functional variants

Challenges

Predicting the impact of a non-coding variant remains a challenging task. Most non-coding variants have cell-type specific effects[100]. As melanoma is heterogenous within a single tumor and across the melanoma lifespan, variants may be functional at specific timepoints and within specific subpopulations of the tumor, as observed in lung cancer[101]. Most non-coding variants are in regions without any regulatory activity, as approximately 10% of the genome harbors regulatory potential[102] but determining the 10% that is functional will depend on the cell type and the transcription factors and/or chromatin modifiers that are expressed. Determining whether a variant affects binding of a transcription factor and/or chromatin modifier is based on position-weight-matrices (PWMs) which are experimentally determined through various assays including ChIP-seq. PWMs can be used to predict motif-breaking or motif-gaining but the majority of transcription factor binding sites remain unknown[103]. Additionally, PWMs for transcription factors within the same family are often similar and can lead to multiple predictions and inferences.

Bioinformatic pipelines that utilize RNA-sequencing

Many bioinformatic pipelines have been applied to detect and prioritize functional non-coding variants. Some of the most successful methods are those that pair a variant to a phenotype, often referred to as quantitative trait loci (QTL). Variants that alter expression of a nearby gene are referred to as expression QTLs (eQTLs). Similarly, QTLs can be associated with methylation,

accessibility, transcription factor binding, or any other paired phenotype. However, QTL analyses still require experimental validation to identify the causal SNP and are often underpowered to identify rare variants. Another analytical method that matches transcription to a variant is allele-specific expression (ASE) which is based on the idea that transcript levels between chromosomes can vary if a heterozygous non-coding variant affects expression on one chromosome but not the other. This method is powerful as it can test the effect of rare variants on expression but requires a transcribed heterozygous allele and phasing of the non-coding variant to the exonic allele[104].

These methods require matched whole genome sequencing (and in the case of ASE, long-read whole genome sequencing or otherwise inferred haplotyping) and RNA-sequencing. Unfortunately, of the 200 melanomas that have been whole genome sequenced to adequate read depth (>30X), the minority (~50) have also been analyzed by RNA-sequencing which renders eQTL and ASE analysis underpowered for most variants. Moreover, many tumors lack appropriate matched controls – blood is commonly used as a matched germline control but given the high mutational burden of melanomas, whole genome sequencing and RNA-sequencing of non-transformed melanocytes would provide a more useful control, although the logistics of such experiments render it almost impossible in living humans.

Summary of methods used to detect non-coding variants in melanoma

The TERT promoter mutation was simultaneously discovered through linkage analysis of a familial case of melanoma and through whole genome sequencing of 70 melanomas and 150 cell lines representing multiple cancer types[80,81]. No other variant in melanoma has been validated by reporter assay on the grounds of recurrence or inheritance alone. Instead, many studies have integrated available WGS and RNA-sequencing to identify cis-regulatory variants. Since so few melanomas have matched WGS and RNA-sequencing, studies often combine multiple cancer

types to identify significant events based on recurrence and change in expression[84,96] or use loci identified by GWAS which compares germline variants between cases and controls[87]. One study genotyped and RNA-sequenced 106 primary melanocytes to perform cis-eQTL analysis but were limited to common variants due to the sample size[105]. The ICGC-MELA cohort, which is the largest cohort of melanomas with WGS, detected almost 21 million variants, >99% of which were non-coding[3]. A preliminary analysis of recurrently mutated promoters, 5' UTRs, and 3' UTRs detected nine promoter variants that were statistically significant based on the OncoDriveFML algorithm[106]. Of these nine, only three displayed altered expression on the associated gene, one of which was the TERT promoter mutation.

Through these combined efforts, 44/837 variants have been validated via reporter assay: 831 variants detected by GWAS via MPRA, 4 detected in the TCGA-SKCM cohort via flow cytometry or luciferase assay, and 2 detected in the ICGC-MELA cohort by luciferase assay[83,87,96,105]. In this chapter, I describe a novel bioinformatic pipeline that bypasses the lack of RNA-sequencing to determine statistically significant recurrently mutated non-coding regions in putative melanoma-specific regulatory regions using as input the 21 million variants from the ICGC-MELA cohort and various ChIP-seq and ATAC-seq datasets and validate my findings via MPRA and luciferase assay.

The FunSeq2 algorithm

The bioinformatic pipeline described in this Chapter scores recurrently mutated regulatory regions based on two factors: recurrence and predicted impact of the variant. The FunSeq2 algorithm scores and prioritizes variants through integration of various features of the variant and the variant-associated gene[107]. Utilizing the comprehensive functional annotations generated by the Encode Project Consortium, variants within these putative regulatory regions or transcription

factor binding sites are scored higher. Scores are also dependent on changes to the PWM score indicating loss or gain of a motif and conservation of the nucleotide. The variant-associated gene is also scored higher if central to a gene regulatory network. FunSeq2 weighs each feature score with different weights based on significance to output a final score for every nucleotide.

2.2.5 Reporter Assays

Due to the large number of non-coding variants in melanoma, bioinformatic prioritization of variants is an essential first step. Once a panel of high-confident variants has been identified, experimental validation of the variant is key to filter out false positives. A major advantage of reporter assays is the ability to exogenously assay a variant compared to its wild-type (WT) counterpart without having to perform precise base editing which has low success rates and are low throughput. It is generally assumed that if the variant alters activity in a reporter assay it will also alter activity if endogenously engineered. However, this has yet to be comprehensively evaluated.

Luciferase assays are a medium-throughput reporter assay. A region-of-interest of variable length is cloned upstream of luciferase in a vector with either a minimal promoter, if the region-of-interest is in an enhancer, or in a promoter-less vector if the region-of-interest is in a promoter. The region-of-interest will either contain the WT or mutant allele and are transfected in separate wells. To control for transfection efficiency, a vector expressing renilla is co-transfected with each luciferase vector. Luciferase values are normalized by the renilla values and these normalized values are compared between WT and mutant luciferase vectors. The advantages of luciferase assays are their low cost, quick readout, and easy statistical analysis[108].

Massively parallel reporter assays

One of the major disadvantages of the luciferase assay is its throughput. For every WT and mutant allele, a separate well must be transfected. Therefore, it requires six wells to test one WT and one mutant allele in triplicate. To comprehensively assay the many mutations present in melanoma, a more high-throughput method is necessary. Massively parallel reporter assays (MPRAs) address this limitation. MPRAs rely on barcodes and next generation sequencing to assay many hundreds to thousands of variants and their WT counterparts in one single well. As with the luciferase vector, a region harboring either the WT or mutant allele is cloned upstream of a reporter, usually GFP which is used to qualitatively check transfection efficiencies, and a barcode[109]. After transfection, RNA is isolated, sequenced using universal primers targeting the MPRA cDNA, and read counts associated with each barcode are demultiplexed to allow comparison between the relevant WT and mutant alleles. The disadvantages associated with MPRAs are their high cost, as each region-of-interest needs to be synthesized, and the relatively more difficult analysis. Overall, both luciferase assays and MPRAs are excellent ways with their own advantages and disadvantages to ensure the predicted variant is indeed functional in one or more melanoma cell lines.

2.2.5 Melanoma cell lines

The need to validate variants in multiple cell lines is imperative, as multiple cell states exist within a melanoma tumor[74–77]. Generally, these cell states can be divided into 4 broad categories based on expression of 4 genes: “Melanocytic” (high levels of MITF and SOX10, low levels of NGFR and AXL), “Transitory” (high: SOX10 and NGFR, low: MITF and AXL), “Neural crest-like” (high: NGFR, SOX10, and AXL, low: MITF), and “Undifferentiated” (high: AXL, low: NGFR, SOX10, and MITF). These cell lines have varied transcriptional, protein, and epigenetic landscapes[110,111] .

In this chapter, we performed luciferase assays in nine cell lines. A375, LOX-IMVI, RPMI-7951, SK-MEL-28, SK-MEL-5, UACC-62 are BRAF-mutant melanoma cell lines that are commercially available. SK-MEL-2 contains the $NRAS^{Q61K}$. LOX-IMVI and RPMI-7951 have almost no detectable levels of SOX10 and are classified as "Undifferentiated", while A375 and SK-MEL-2 have relatively high levels of all four markers and are classified as both "Neural crest-like" and "Undifferentiated". SK-MEL-5 and UACC62 have low levels of NGFR and AXL but relatively higher levels of MITF and SOX10 and are thus classified as "Melanocytic". SK-MEL-28 are "Neural crest-like" due to low expression levels of AXL but high NGFR, SOX10, and MITF. In addition to melanoma cell lines, we validated results in 293FT cells, which are human embryonic kidney cells, and primary melanocytes derived from human foreskin melanocytes. These cell lines are used as non-melanoma comparators, in order to determine if there are any melanoma-specific effects throughout the work described in this dissertation.

2.2.6 Aim of Chapter 2

In this Chapter, I design a novel bioinformatic pipeline to identify recurrently mutated regions in putative melanoma regulatory regions (pMRRs). This pipeline utilizes previously published ChIP-seq and ATAC-seq datasets from relevant sample types to partition the genome. Regions not in pMRRs are used as an empirical null distribution in order to calculate statistical significance. As non-coding variants cannot be reliably predicted *in silico*, we use luciferase assays and MPRAs to experimentally validate statistically significant and high-scoring variants.

2.3 Methods and Materials

2.3.1 Calculating hotspot scores

Step 1: Merge mutations into hotpots. Mutation calls for SNVs and indels from the MELA-AU cohort were downloaded from dcc.icgc.org after receiving DACO approval[3]. Using a 25 bp

window, we merged mutation calls using bedtools intersect into hotspots based on the premise that highly recurrent variants may be under positive selection at some point during the melanoma life cycle (e.g. favor melanoma growth) and that a transcription factor binding site(s) (TFBSs) may be disrupted/created by modifying any of multiple nucleotides in this window[112].

Step 2: Filter hotspots not in putative enhancers/promoters. We downloaded processed peak calls from ChIP-seq (e.g. H3K27Ac, H3K4me3, CTCF) and ATAC-Seq (revealing accessible chromatin domains) data from 69 melanoma datasets to enrich for putative Melanoma Regulatory Regions (pMRRs) which we reasoned are more likely to bind transcription/chromatin factors (information available upon request). These are indicated by the blue "peaks" in the example Figure 1A. We excluded exons and those regions (e.g. highly repetitive) from Encode excluded regions list[113].

Step 3: Calculate Donor Score. The donor score for a given hotspot is represented as D^2/G, where D is the number of samples (donors) with the specific variant and G is the number of nucleotide locations with variants in the hotspot. For example, in Figure 1A, the purple hotspot shows D = 3 + 1 + 2 + 4 = 10 mutations, at G = 4 different locations, for Donor Score of $10^2/4$ = 25.

Step 4: Weight variants using FunSeq2 score. Each mutation is weighted for predicted functional significance by features including predicted TFBS motif creating/breaking effect and evolutionary conservation using pre-computed scores from the published FunSeq2 algorithm (http://funseq2.gersteinlab.org/downloads) with a higher score predicting higher likelihood of functional significance[107].

Step 5: Calculate Hotspot Score. Each hotspot is assigned a Hotspot Score as the product of the Donor Score (Step 3) and mean FunSeq2 score (Step 4) for all variants in the hotspot, to

weigh both the number of variants and their predicted functional consequence in one metric. For example, in Figure 1A, the purple box shows (Average FunSeq2 score)*(Donor Score) = 1.5*25 = 37.5

Step 6: Calculate p-value for each hotspot in MRRs relative to the empirical null distribution (non-pMRR regions from Step 2). For each hotspot score within pMRRs, we calculated a p-value by determining the proportion of null hotspots with hotspot scores greater than or equal to it. All p-values were adjusted for false discovery rate (FDR). Adjusted p-values equal to 0 are provided (Supplemental Table 2).

2.3.2 Genomic Analysis of Hotspots

For all pMRRs, statistically significant hotspots (FDR adjusted p-value < 0.05, 707 hotspots), and top-scoring hotspots outside of pMRRs (top 707 null hotspots by Hotspot Score), we annotated regions using the ChIPSeeker function *annotatePeak*[114] (Figure 1D*)*. For HOMER motif analysis, we ran findMotifsGenome.pl on BED files of all pMRRs and statistically significant hotspots to identify known motifs (Supplemental Figure 1A). For each variant within statistically significant hotspots, we made FASTA files with 20 bp sequences corresponding to either the WT or mutant sequence (variant at position 10). These were processed through HOMER using the findMotfs.pl function (Supplemental Figure 1A). A BED file containing only the CDC20 promoter variants were processed through motifBreakR[115] using the known and discovered motif information from transcription factor ChIP-seq datasets in Encode[116].

To calculate the ETS motif distribution, we first made FASTA files containing 11 bp sequences corresponding to either the WT or mutant sequence (variant at position 6) from the 707 statistically significant hotspots with FDR-adjusted p-values < 0.05. If a sequence contained the GGAA motif, we counted how far each variant within a statistically significant hotspot occurred

from the nearest GGAA (if more than one instance was detected). If the reverse complement, TTCC was identified, as the nearest ETS motif, we first rewrote the sequence as its reverse complement and then counted the distance. A consensus sequence was generated with Web Logo (https://weblogo.berkeley.edu/logo.cgi) using a re-oriented version of the 11 bp WT fasta file where the first G of the GGAA motif is always at position 5.

2.3.3 Selection of variants

Manual curation of 15 hotspots

As non-coding somatic variants are often heterozygous and their regulatory effects in cis, we reasoned that genes near hotspots displaying allele-specific expression (ASE) could be further evidence of a putative functional hotspot. In order to further investigate this, we downloaded the alignment files (BAM format) of RNA-seq performed on 56 tumor samples 46 donors from ICGC (EGAD00001003353). Following the GATK best practice guideline for short variant discovery in RNA-seq, we called variants and used the GATK ASEReadCounter tool to calculate wild-type and mutant alleles. To determine whether genes were displaying ASE, we first generated a normal distribution with parameters α=0.5 and β=0.1. This created a distribution with 1st and 3rd quartiles similar to the allelic ratios found in genes with low variance across melanoma and melanocyte samples[117] (0.4-0.6, data from GSE112509). Using this sample as our null distribution, we calculated p-values for each variant detected in a transcript. FDR-adjusted p-values were collapsed for each gene by calculating the average p-value.

To generate a list of high-confidence hotspot candidates, we considered each hotspot primarily by their test statistic. We then carefully considered the results of the ASE analysis, information from the literature, the gene expression pattern from an independent RNA-seq dataset, presence of a mutation within the genomic coordinates of candidate hotspots in a smaller

independent cohort, and whether it was considered a putative driver in a recent pan-cancer analysis[18,118] (Supplemental Table 3).

MPRA Variant Selection

Starting with all the variants present in hotspots with q-values less than 0.005, we looked for presence of the mutation in two independent cohorts: 25 metastatic melanomas and 18 melanoma cell lines from the Cancer Cell Line Encyclopedia (CCLE). 118 SNVs were detected in at least 4 samples from the secondary cohorts, corresponding to 108 unique hotspots.

Cohort comparison of Top 13 Genes

We downloaded DESeq2-normalized read counts from GSE112509 for the Kunz cohort and quantile-normalized read counts from Firehose (Broad GDAC) for the TCGA-SKCM cohort. The Kunz cohort is made of 23 laser-microdissected melanocytic nevi and 57 primary melanomas[117]. The TCGA cohort consists of 81 primary and 367 metastatic melanomas[8].

For ICGC-MELA, we downloaded BAM outputs from STAR from the European Genome-Phenome Archive (EGA) under Study ID EGAD00001003353. Gene counts were calculated using RSEM and normalized by DESeq2. This cohort comprises 56 melanomas from 46 donors and 25 metastatic melanomas, 17 primary melanomas, and 14 cell lines derived from tumors.

For the Baggiolini cohort, we obtained raw counts from the supplementary material of the corresponding publication and normalized counts by DESeq2[119]. This cohort is made up of human pluripotent stem cell derived cells that are engineered to contain doxycycline-inducible $BRAF^{V600E}$. KO lines contain deletions to RB1, TP53, and P16. These cells were then differentiated into neural crest cells, melanoblasts, and melanocytes. For our study, we only considered WT and KO melanoblast samples that had activated $BRAF^{V600E}$ expression. In line with the corresponding publication, we consider KO melanoblasts to be melanoma-like (based on the ability to form

tumors when subcutaneously injected into NSG mice) while WT melanoblasts were considered to be a non-tumorigenic precursor to melanocytes.

For each of the top 13 genes, we calculated the $\log_2$ fold-change between metastatic and primary melanomas (TCGA-SKCM and ICGC-MELA), primary melanoma and nevi (Kunz), and KO and WT melanoblasts. Survival rates and corresponding p-values for high and low expressing tumors were downloaded from cBioPortal (TCGA-SKCM) using the Onco Query Language (OQL): *GENE*: EXP < -0.5 and GENE: EXP > 0.5. Data was downloaded from cBioPortal.org and plotted with ggplot2.

2.3.4 Cell Culture

We obtained A375 (CRL-1619) and RPMI-7951 (HTB-66) cells from ATCC. SK-MEL-2, LOX-IMVI, SK-MEL-28, SK-MEL-5, UACC-62 cells were obtained directly from the NCI-60 collection following written request and approval and were grown in RPMI-1640 media with 2 mM L-Glutamine (Gibco, 11875) with 10% FBS and 1X Pen/Strep. Newborn foreskin melanocytes were ordered from the specimen research core at the SPORE in Skin Cancer at Yale University. HEK-293FT cells were obtained from Invitrogen (#R70007). Cells were grown in a dedicated incubator set to 37°C at 5% CO2. A375 and HEK 293FT cell lines were grown in DMEM media (Corning, 10-013-CV) with 10% Fetal Bovine Serum (Gibco, 261470) and 1X Penicillin/Streptavidin (Pen/Strep, Sigma-Aldrich, P4333). Primary melanocytes were grown in OPTI-MEM (Gibco, 31985) containing 5% FBS, 1X Pen/Strep, 10 ng bFGF (ConnStem, F1004), 4 mL of 5 mM IBMX (Sigma, #I-5879), 1 ng/mL Heparin (Sigma, #3393), and 200 μL of 0.1 M dbcAMP (Sigma, #D-0627). SK-MEL-5 and RPMI-7951 were grown in EMEM media (Corning, 10-009-CV) with 10% Fetal Bovine Serum (Gibco, 261470) and 1X Penicillin/Streptavidin (Pen/Strep, Sigma-Aldrich, P4333). UACC-62, LOX-IMVI, UACC-257, SK-MEL-28, and SK-

MEL-2 were grown in RPMI-1640 media (Corning, 10-040-CV) supplemented with 1X L-Glutamine (Gibco, # 25030081), 10% Fetal Bovine Serum (Gibco, 261470), and 1X Penicillin/Streptavidin (Pen/Strep, Sigma-Aldrich, P4333).

2.3.5 Luciferase Assays

For the first round of luciferase assays, we synthesized 300 bp sequences corresponding to WT and mutant hotspots with the variant centered at position 150 and sequenced into a luciferase vector with a minimal promoter (pGL3-Promoter, E1761). For the second round of luciferase assays, we synthesized a 170 bp sequence containing the WT CDC20 promoter sequence (chr1:43,824,464-43,824,633) (GenScript). From this template, we amplified a 150 bp sequence using primers pGL3-CDC20_F and pGL3-CDC20_R (Phusion High-Fidelity PCR Master Mix, NEB M0531, Supplemental Table 4) that added restriction sites for SacI and XhoI to the 150 bp sequence. Both the pGL3-Basic Luciferase vector (Promega, E1751) and the CDC20 promoter amplicon were digested using SacI-HF (NEB, R3156S) and XhoI (NEB, R0146S) at 37°C overnight, followed by heat inactivation at 65°C for 20 minutes. Digested vector and amplicon were ligated using T4 DNA Ligase (NEB, M0202S) and transformed into OneShot Top10 Chemically Competent Cells (ThermoFisher, C404010). Individual colonies were mini-prepped and confirmed by Sanger Sequencing (Azenta).

Using the Q5 Site-Directed Mutagenesis kit (NEB, E0554), we induced variants in the WT sequence using primers designed by NEBaseChanger (https://nebasechanger.neb.com/, Supplemental Table 4). Sequences that were successfully mutated, as well as the WT pGL3-Basic vector and pRL-TK (Promega, E2241), were midi-prepped (Qiagen, 12941).

For all transfections, 300,000 cells per well were seeded onto 6-well plates. All transfections were performed using 9 uL of Lipofectamine 2000 (Invitrogen, 11668), 1.5 μg of

luciferase vector, and 1.0 μg of control pRL-TK (renilla), following the manufacturer's protocol. All transfections were performed at minimum in duplicate.

The following day, luciferase and renilla luminescence were measured using the Dual-Luciferase Reporter Assay System (Promega, E1910) per manufacturer specifications. Cells were lysed using 500 μL of 1X Passive Lysis Buffer and incubated for 15 minutes on an orbital shaker. 20 μL of lysate were added to clear-bottom 96-well plates. We ran three technical replicates per sample. Luminescence was measured on a GloMax 96 Microplate Luminometer (Promega) using a standard Dual Reporter Assay program. All luciferase values were normalized to renilla, as the internal transfection control. We then normalized all variant ratios to the corresponding average WT value. p-values were calculated using Student's t-test.

2.3.5 Massively parallel reporter assay

Oligonucleotides were synthesized by IDT and were designed as described previously with several differences[120]. Each oligo contained an upstream universal primer binding site, a Step 1 restriction enzyme site, a 150 base pair regulatory region containing either the reference or mutated allele at the center, a Step 2 cloning site, a C spacer nucleotide, a second Step 2 cloning site, a unique 10-base pair barcode, a second Step 1 cloning site, and a universal downstream primer binding site. To create the Step 1 library, oligos were amplified with 6 cycles of PCR using the Phusion HiFi Master Mix (HF Buffer) using primers MPRA_Oligo_F and MPRA_Oligo_R (Supplemental Table 4). The reaction was cleaned up using the Qiagen PCR Clean Up Kit and digested with AseI and SalI. Reactions were stopped by heat inactivation at 80°C for 20 minutes. The backbone used for our MPRA assays had an EF1α promoter and an mCherry reporter gene. To obtain the Step 1 library, the vector was digested in the same manner as the oligos and ligated overnight at 16°C using a 1:5 molar ratio starting with 42 ng of vector. The step 1 library was

sequenced using a MiSeq to determine percent drop-out. The Step 1 library contains the vector backbone, the synthesized oligo containing the wild-type or mutant enhancer, and the barcode. To clone in the promoter and reporter, we digested the Step 1 library at the synthesized Step 2 cloning sites using PvuI and AatII restriction enzyme. We stopped the reaction by PCR clean-up. The EF1a-mCherry sequence was amplified with MPRA_pEF1a_mCh_6bp_ext_F and MPRA_pEF1a_mCh_6bp_ext_R that added the restriction enzyme sites (Supplemental Table 4). This amplicon was digested and ligated into the Step 2 vector. These counts are used to normalize the cDNA barcode counts. Libraries were sequenced on a MiSeq-v3 to a total of 2 million reads and at an average of 55 counts per barcodes. 114 of 118 variants were present in the final library.

2.5 μg of the Step 2 library was transfected into HEK 293FT, Primary Melanocytes, A375, SK-MEL-5, RPMI-7951, and UACC-62 using 10 μL of Lipofectamine 2000. RNA was isolated 24 hours post-transfection using the Qiagen RNeasy Mini Plus Kit. RNA was treated with TURBO DNase (following the 'Rigorous' protocol, ThermoFisher Scientific #AM2238) and converted into cDNA using the First Strand SuperScript III kit (Invitrogen, # 18080051). Barcodes were amplified using primers MiSeq_MPRA_Step2_Barcode_F and MiSeq_MPRA_Step2_Barcode_R (Supplemental Table 4). A second round of PCR is performed to add adapters and indices. The completed Step 2 library was sequenced on a MiSeq to a total of 2 million reads and 1000X coverage of barcodes. Barcodes were demultiplexed, counted, and divided by the tag counts obtain from the Step 2 library. For each corresponding WT and mutant hotspot, we calculated $\log_2$ fold-changes and p-values using the Student's t-test. All p-values were adjusted for false discovery rate.

2.4 Results

2.4.1 Putative regulatory regions in melanoma are enriched for hotspot mutations

To identify recurrent non-coding mutations in human melanoma, we used variants called from whole genome sequencing (WGS) data from the International Cancer Genome Consortium (ICGC), the largest collection of WGS for melanoma to our knowledge, including 183 melanoma samples made up of 75 primary tumors, 93 metastases, and 15 human melanoma cell lines, as exome sequencing does not include full promoters or distal regulatory elements. The bulk of these tumors are cutaneous (140) but includes 35 acral and 8 mucosal melanomas. A total of 20,894,255 substitutions and 96,467 indels were identified from the ICGC Melanoma cohort[3].

To refine our search space, we collated 69 previously published ChIP-seq and ATAC-seq datasets that were specifically performed on melanoma or melanocyte samples[121]. We reasoned these regions of the genome are more likely to bind transcription/chromatin factors and refer to them as putative melanoma regulatory regions (pMRRs). Genomic regions outside the pMRRs (red box, indicated by the lack of peak, Figure 1A) serve as an empirical null distribution but still have large numbers of recurrent mutations.

pMRRs account for only ~12% of the genome and harbor 2,142,063 variants (~10% of total variants detected in the ICGC cohort). Of these, 444,161 variants are merged into 118,741 hotspots (3 or more variants within 25 bp are merged). Our empirical null distribution accounts for 5,478,131 variants within 1,462,992 hotspots. The remaining variants are isolated (i.e. not within 25 bp of another variant) and thus were not designated as hotspots.

All hotspots are also scored based on recurrence (donor score) and the average predicted impact of all variants within a hotspot as computed by the FunSeq2 algorithm[107], which weighs attributes such as evolutionary conservation and likelihood of TF motif creation/destruction (Funseq2 score, Figure 1A'). Hotspots in pMRRs have higher hotspot scores (product of donor

score and FunSeq2 score) than those in null regions (Figure 1B). While donor scores are 4.9-fold higher in hotspots within pMRRs than those in null regions, FunSeq2 scores are 6.7-fold higher, drastically reducing the hotspot scores in regions outside of pMRRs and therefore potentially reducing false positives (Figure 1C).

Promoter regions are enriched in statistically significant test hotspots, while top-scoring null hotspots are commonly found in intergenic regions (Figure 1D). We identified 140 hotspots with FDR-adjusted p-values = 0 encompassing 2,631 mutations, notably including the known TERT promoter variant which has the 13th highest hotspot score (Supplemental Table 2).

In order to evaluate for enrichment of putative TF binding site motifs, we used Homer analysis of pMRRs which identified motifs for TFs known to play prominent roles in melanoma, including SOX10[47,48,122] (p-value = 1 x 10^{-472}) and ETS family factors[123] (Supplemental Figure 1A), as well the multifunctional chromatin regulator CTCF (p-value = 1 x 10^{-6092}). However, pMRRs that encompassed statistically significant hotspots are only enriched in ETS motifs, as previously observed (Supplemental Figure 1A). No ETS factor motifs are enriched in the mutant sequences, suggesting that most mutations break ETS transcription factor motifs (Supplemental Figure 1A). We found an almost identical distribution of mutations around the canonical GGAA ETS motif within the significant hotspots identified in our pipeline as previously reported[124,125] (Figure 1E).

2.4.2 An initial assessment of 15 hotspots by luciferase assay

Of a total of 26 variants corresponding to 15 hotspots, 19 altered reporter activities significantly in at least one melanoma cell line, 18 in at least two cell lines, and 3 in all three melanoma cell lines (Figure 2). These 3 included the two prominent TERT promoter mutations, G228A and G250A, which increased reporter activity and a mutation 2,200 bp upstream of SOX9 in the intron of SOX9-AS1 which decrease reporter activity compared to the WT sequence. We

detected two hotspots upstream of NRG1 (178 Kb and 555 Kb from TSS) that decreased reporter activity in 2 melanoma cell lines. We tested two variants 400 bp upstream of TCF3: C018T decreased reporter activity but C026T increased reporter activity, despite being 8 bp away. Similarly, variants 100 bp upstream of ASXL2 mostly led to an increase in reporter activity; C488T significantly increase reporter activity in SK-MEL-5 and A375 but C489T, the adjacent nucleotide, led to a 0.7 fold decrease in reporter activity. The same variant upstream of MCRS1 led to an increase in UACC-62 but a decrease in A375, suggesting cell-type specific effects. Two out of three variants in a hotspot downstream of CCNF led to a significant increase in reporter activity in UACC-62 and A375 but not SK-MEL-5. One hotspot in the intron of HDAC9 and downstream of TWIST1, an important EMT transcription factor, led to an increase in expression in UACC-62 and A375; another hotspot 100 kb away did not display any significant changes in reporter activity. A hotspot in the intron of ANGPT1 increased reporter activity in all three cell lines, two of which reached statistical significance and one which was just above at an FDR-adjusted p-value of 0.06.

2.4.3 A systematic assessment of 108 hotspots by MPRA

13 out of 118 variants reached statistical significance in 1 cell line, 5 in 2 cell lines, and 3 variants were significant in 3 cell lines. Fold changes correlated across similar cell types (Figure 3). One of the variants has previously been validated in a separate publication[83]. The other is upstream of SLC30A6; an adjacent variant decreased reporter activity in only one cell line. The GG528AA and G528A variants in the CDC20 promoter, which decreased reporter activity in the preliminary luciferase assay, were also observed to decrease reporter activity significantly in 3 and 1 of the 5 cell lines, respectively, by MPRA (Figure 4). Hotspot scores for variants that altered reporter activity by MPRA were high compared to variants not selected for validation (Figure 5).

2.4.4 A comprehensive assessment of CDC20 promoter variants across nine cell lines

To focus our efforts on a candidate(s) among the top scoring hotspots (i.e. those with scores higher than TERT, encompassing thirteen candidates), we looked for consistent changes in gene expression for the gene nearest the recurrent variants between different stages of melanomagenesis (Supplemental Figure 1B). We used RNA-sequencing from 4 studies to calculate the fold change of the genes nearest to the hotspot between primary and metastatic tumors (The Cancer Genome Atlas, TCGA-SKCM and ICGC-MELA), nevi and melanoma (Kunz), and hPSC-derived melanoblasts with (KO melanoblasts) and without (WT melanoblasts) deletions in key tumor suppressors (Baggiolini, see Methods for description of samples). *CDC20* (gene associated with the 8th highest-scoring hotspot) is consistently upregulated in expression between melanoma and nevi (Kunz) and the KO and WT melanoblasts (Baggiolini, Supplemental Figure 1B). We observe a small increase in metastatic tumors compared to primary tumors in the ICGC cohort and no change between primary and metastatic tumors in the TCGA. The only other $\log_2$ fold-change greater than 1 is seen in the ICGC cohort for *TERT* expression (increase in metastatic melanoma, Supplemental Figure 1B). Low levels of *RPL18A* (3rd highest-scoring hotspot), *HNRPNUL1* (6th), and *CDC20* (8th) tumors have higher survival rates than tumors with high expression of these genes (Supplemental Figure 1C). Taking both differential gene expression and association with survival rates for those with melanoma into consideration, we specifically focus on characterizing the *CDC20* promoter in melanoma.

The CDC20 promoter is mutated in 39 of 183 donors in the ICGC dataset, all of which are skin cutaneous melanomas (27.9% of cutaneous melanoma). The most common single-nucleotide variants (SNVs) are at adjacent positions chr1:43,824,528 (G>A, hereinafter termed G528A, mutated in 10 donors) and chr1:43,824,529 (G>A, G529A, 16 donors) as well as a SNV at position

chr1:43,824,525 (G>A, G525A, 4 donors) and a multi-nucleotide variant (MNV) at positions chr1:43,824,528-43,824,529 (GG>AA, GG528AA, 4 donors) and are located within an ETS motif (Figure 6A). While at adjacent positions, G528A and G529A have different FunSeq2 scores (second number) and Genomic Evolution Rate Profiling (GERP) scores (third number) reflecting different degrees of purifying selection[126]. G525A is located within the core ETS motif, at the position that is most often mutated when taking all variants within statistically significant hotspots into consideration (Figure 1E) but is not the most recurrent variant in the CDC20 promoter hotspot, occurring only in 4/39 donors. Like G528A, G525A has both a high FunSeq2 score and a high GERP score.

Overlaying chromatin-related assessments of the locus, the CDC20 promoter is accessible in 4/7 datasets that assay genome-wide chromatin accessibility (Supplemental Table 1). BRG1, CTCF, and TFAP2A are among the chromatin/transcription factors that have binding activity at the CDC20 promoter, as detected by ChIP-seq. ETV1, the only ETS factor with ChIP-seq data in our collation of melanoma-specific functional datasets, did not have binding activity at the CDC20 promoter in the 2 cell lines assayed (A375 and COLO-800).

To understand how the variants affect the regulatory activity of the CDC20 promoter, we performed luciferase assays using a 150 bp sequence length in a promoter-less luciferase vector (Figure 6). C520T reduced reporter activity in A375, SK-MEL-2, and primary melanocytes. G525A and C537T reduced reporter activity in all cell lines tested. G528A, the second most common variant, reduced expression in all but SK-MEL-2. G529A, the most common variant, reduced activity in A375, SK-MEL-2, RPMI-7951, HEK 293FT, and UACC-62. GG528AA only reduced activity in A375 and primary melanocytes but upregulation in HEK 293FT. CDC20

promoter hotspots are not more likely to co-occur with pathogenic *BRAF* mutations than *NRAS* (p-value = 0.67, Fisher's Exact Test, Supplemental Figure).

2.5 Discussion

Using the largest available cohort of melanoma whole-genome sequencing data and several dozen melanoma-specific functional genomics datasets, we have identified hundreds of mutational hotspots containing putatively functional non-coding somatic variants. Under the assumption that variants outside of pMRRs are not, or are less likely to be, functional, we generated an empirical null distribution with which to calculate significance. We chose to focus on characterizing variants in the promoter of CDC20, whose weighted rate of recurrence and predicted functional significance were greater than that of the well-studied TERT promoter variants and began to investigate how these variants alter melanoma behavior.

Our pipeline to identify putatively functional non-coding variants has important similarities and differences with other pipelines. Like many other studies, we selected recurrently mutated regions by performing a hotspot analysis[84,105,118]. In melanoma, recurrently mutated regions are significantly enriched at ETS transcription factor binding sites and at CTCF binding sites due to damage from UV irradiation[124,125,127]. These hotspots, frequently in promoters, likely lead to an increase in the number of false positives. However, mutations in transcription factor binding sites are one potential mechanism by which gene expression is altered and therefore should not be ignored simply due to mutational signature. Another major similarity is utilizing functional annotations of the genome to refine the search space. Unlike other methods, we specifically use ChIP-seq and ATAC-seq datasets from melanoma and melanocyte samples, enriching for hotspots within cell-type specific regulatory elements. The last important similarity is using a method by which to score each individual variant. In Weinhold et al. (2014), mutations were selected based

on location within an ETS motif. In Rheinbay et al. (2020), the most recent pan-cancecr analyses of non-coding variants, mutations were ranked based on a collective p-value determined from 12 methods to identify driver mutations[118]. These methods consider features such as the expected local mutation rate, the FunSeq2 score, algorithms specifically designed to look for driver mutations in lncRNAs, CADD scores, and conservation of a nucleotide. This study emphasized the lack of driver events despite a systematic and comprehensive analysis, suggesting an improvement in bioinformatic pipelines and/or larger numbers of samples processed both by RNA-sequencing and WGS to perform either eQTL analyses or ASE. Alternatively, experimentally validating variants and improving pipelines based on results could yield more accurate scoring algorithms.

Our pipeline builds on previous work by looking for functional non-coding variants specific to melanoma using cell-type specific functional annotations. We then validated our results by performing luciferase assays and MPRAs. In total, we assayed 25 variants as a 300 bp long enhancer by luciferase assay upstream of a minimal promoter, 6 variants as a 150 bp long promoter by luciferase assay, and 118 variants by MPRA as a 150 bp long enhancer upstream of the EF1α promoter. The CDC20 promoter variants, G528A, G529A, and GG528AA, were assayed across all three experiments. The results obtained from the luciferase assays were most similar except for the G529A variant which showed no change relative to WT. However, the MPRA results were discordant (Supplemental Figure 3). We concluded that the MPRA was faulty for three technical reasons: (1) the presence of the strong promoter, EF1α, likely overpowered the signal driven by the variant, (2) many of the variants were in promoters and thus were incorrectly assessed as an enhancer in this assay, and (3) we used an insufficient number of barcodes per variant to reliably determine statistical significance. While these technical limitations led to the inability to use the MPRA for a peer-reviewed publication, we were still able to use our MPRA to provide a

preliminary assessment of our pipeline. Overall, our hotspot scores correlated with our selection criteria and MPRA activity (Figure 5). Future work will go into validating more variants and improving our pipeline based on these results.

Many of the variants included in our assays were adjacent to one another, meaning one sample would have a variant at position *n*, another at *n + 1*, and in some cases both at *n + 1*. We expected these variants to have similar results in the reporter assay as they targeted almost the exact same nucleotide. However, we noted vastly different results for some of the hotspots (Supplemental Figure 4). This suggests a more complicated mechanism of altered transcription factor binding than a simple gain or loss.

Overall, we were most interested in three hotspots upstream of SOX9, in an ANGPT1 intron, and in the promoter of CDC20.The SOX9 hotspot is located ~2 kb away, is present in 8 donors, and leads to a decrease in reporter activity. In support of this, SOX9 has been to shown to repress SOX10 activity and delay onset of melanoma[128] (discussed further in Chapter 4). The ANGPT1 intronic hotspot is located > 5 kb from the TSS, is present in 9 donors, and has been shown to support angiogenesis in breast cancer[129].

The remainder of this thesis will focus on *CDC20* and a hotspot located in the promoter. At least half of the CDC20 promoter variants tested decreased reporter activity across all cell lines in this study. Four variants were within 2 bp of a core ETS motif but did not affect reporter activity to similar extents. G525A, located in the core GGAA ETS motif, reduced expression in every cell line tested, including non-melanoma cell lines such as HEK 293FT and primary melanocytes, suggesting its key location in a transcription factor binding site (Figure 6). Interestingly, this variant is only present in 4 donors, despite being at the position in the ETS motif most mutated across all significant hotspots, as opposed to G528A and G529A which led to a x-fold and y-fold

decrease (Figure 6). Most surprising was the double substitution, GG528AA, that had a less deleterious effect than G528A, and resembled G529A, suggesting a partial rescue of the transcription factor binding site. Overall, we see clear cell-type specific effects that does not seem to depend on the cell state (i.e. neural crest-like, undifferentiated, melanocytic, or transitory) but more likely on the expression levels of multiple transcription factors and their interactors, discussed more thoroughly in Chapter 3.

Future directions of the work performed in this chapter will aim to improve the bioinformatic pipeline through large-scale assessment of variants. Through the use of larger MPRAs and convolutional neural networks, we can learn sequence features that predict effect on reporter activity. Another limitation of this work is the assessment of the exogenous effect of the variant. To more accurately validate each variant, we can transduce pools of pegRNAs that, through prime editing, can introduce the actual variant to its endogenous location. However, this requires high efficiency, which is currently low for prime editing, single-cell RNA-sequencing, and single-cell genotyping.

2.6 Conclusions

CRVs in cancer genomes are emerging as a significant contribution to cancer onset and progression. Like protein-coding mutations, the challenge of identifying CRVs requires predicting the impact of the non-coding variant and then assessing how its impact contributes to tumorigenesis. However, unlike protein-coding mutations, the non-coding genome lacks the characterization of the amino acid code, making prediction, and therefore, prioritization difficult. We have developed a pipeline to confront these challenges. We use published datasets that assess accessibility and DNA binding specifically in melanoma and melanocyte samples to define regions of regulatory activity. We search for mutational hotspots, score them for recurrence and predicted

impact, and calculate statistical significance by generating an empirical null distribution of hotspots that are not predicted regulatory regions but still harbor many variants. Our pipeline identified hundreds of statistically significant hotspots, including the well-known TERT promoter, which has the 13th highest score. 170 variants were assayed either by luciferase assay or MPRA and 35 altered reporter activity significantly in at least one melanoma cell line. Variants in the CDC20 promoter region were validated by two separate rounds of luciferase assays and an MPRA. CDC20 stood out both in its up-regulation between early and later stages of melanoma and its association with overall worse survival.

2.7 Declarations

Ethics Approval and Consent to Participate

Not applicable.

Consent for Publication

Not applicable

Availability of Data and Materials

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE206639. Scrips for data visualization and the bioinformatic pipeline can be made available upon request.

Competing Interests

The authors have no competing interests.

2.8 Funding

This research was supported by the Melanoma Research Alliance Young Investigator Award #566840. P.G. was supported by NSF DGE-1745038.

2.9 Author Contributions

P.M.G. and C.K. conceived of the project. P.M.G. designed the bioinformatic pipeline. P.M.G. and A.Z. did the preliminary luciferase assay. P.M.G. designed and made the MPRA and performed all other validation experiments. P.M.G. and C.K. wrote the manuscript.

2.10 Acknowledgments

We thank members of the Kaufman laboratories for assistance, and my thesis committee and Ryan Friedman (Washington University) for helpful discussions.

2.11 Figures

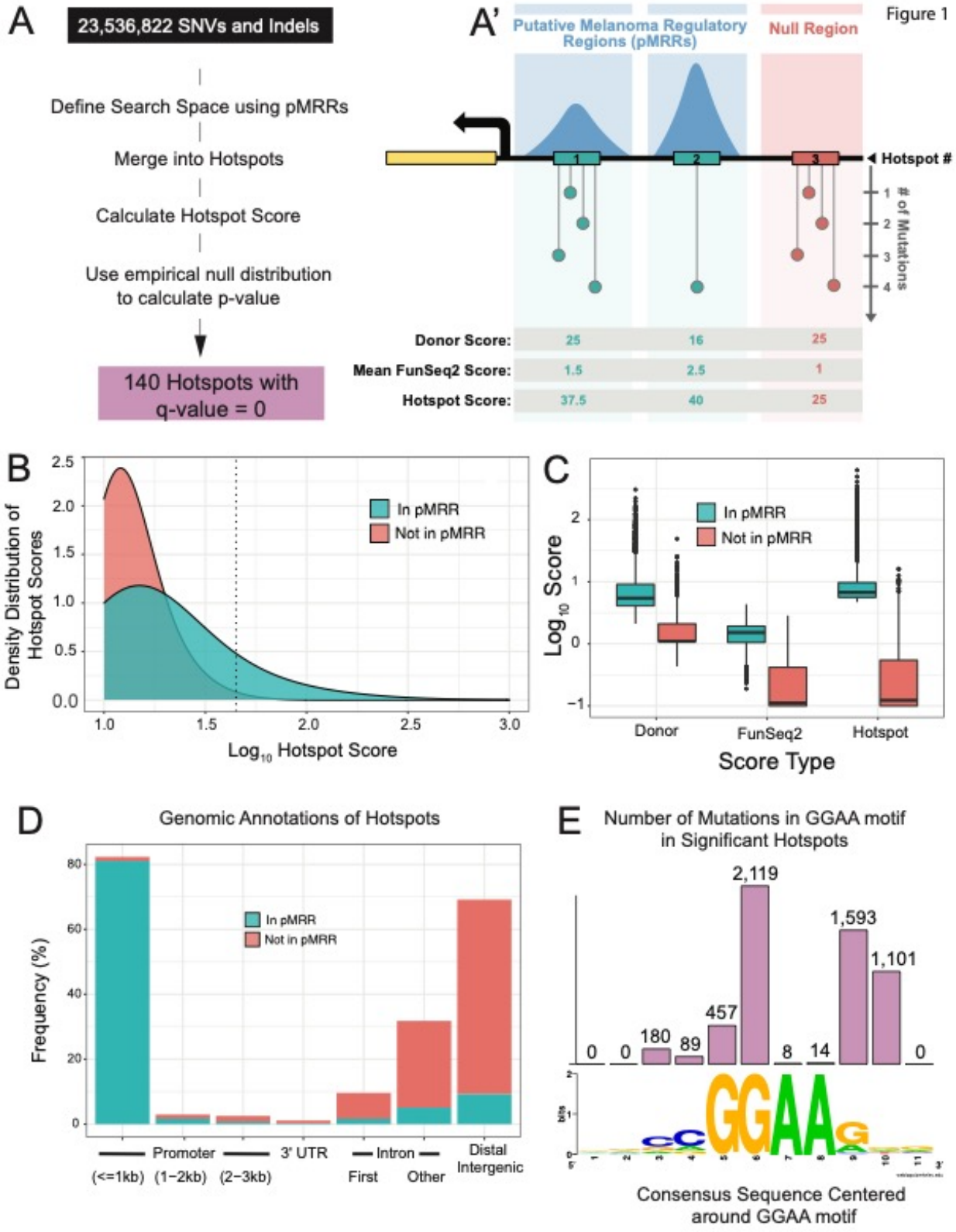

A
23,536,822 SNVs and Indels
Define Search Space using pMRRs
Merge into Hotspots
Calculate Hotspot Score
Use empirical null distribution to calculate p-value
140 Hotspots with q-value = 0
A'
Putative Melanoma Regulatory Regions (pMRRs)
Null Region
Hotspot #
# of Mutations
Donor Score: 25 16 25
Mean FunSeq2 Score: 1.5 2.5 1
Hotspot Score: 37.5 40 25
B
Density Distribution of Hotspot Scores
In pMRR
Not in pMRR
Log10 Hotspot Score
C
Log10 Score
In pMRR
Not in pMRR
Donor
FunSeq2
Hotspot
Score Type
D
Genomic Annotations of Hotspots
Frequency (%)
In pMRR
Not in pMRR
Promoter
(<=1kb)
(1−2kb)
(2−3kb)
3' UTR
Intron
First
Other
Distal Intergenic
E
Number of Mutations in GGAA motif in Significant Hotspots
0
0
180
89
457
2,119
8
14
1,593
1,101
0
bits
Consensus Sequence Centered around GGAA motif

Figure 1

Figure 1. A method to identify putative functional non-coding variants in human melanoma.

(A) Summary of pipeline to identify hotspots (A) with a generalized schematic of three theoretical hotspots (A'). Blue boxes indicate regions within putative Melanoma Regulatory Regions (pMRRs), and red box indicates null regions (i.e. those outside predicted regulatory regions). Numbered rectangles represent hotspots. Dot plots represent the number of variants within a given position. Donor score is equal to the square of the number of donors divided by the number of mutated positions, and FunSeq2 score is a weighting factor with higher values indicating higher conservation within regulatory regions and/or TF binding site motif altering. (B) Kernel density estimate of hotspot scores in pMRRs (blue) and not in pMRRs/in null regions (red). Hotspots with $\log_{10}$ scores lower than 1 are not shown. Dashed line depicts hotspot scores with a p-value = 1×10^{-6}, lower p-values are to the right (C) Boxplots showing the $\log_{10}$-transformed Donor, FunSeq2, and Hotspot (Donor x FunSeq2) for the Top 10,000 highest-scoring hotspots. (D) Bar chart demonstrating the frequency of genomic annotations for Top 10,000 null hotspots (red bars) and statistically significant hotspots (707 hotspots, FDR-adjusted p-value < 0.05, blue bars). (E) Bar chart of the total number of mutations in significant hotspots (707 hotspots) at each site within 4 bp of the core ETS motif, GGAA (top, represents 5,561 mutations out of a total of 8,514 mutations), and WebLogo of 11 bp WT sequence (bottom).

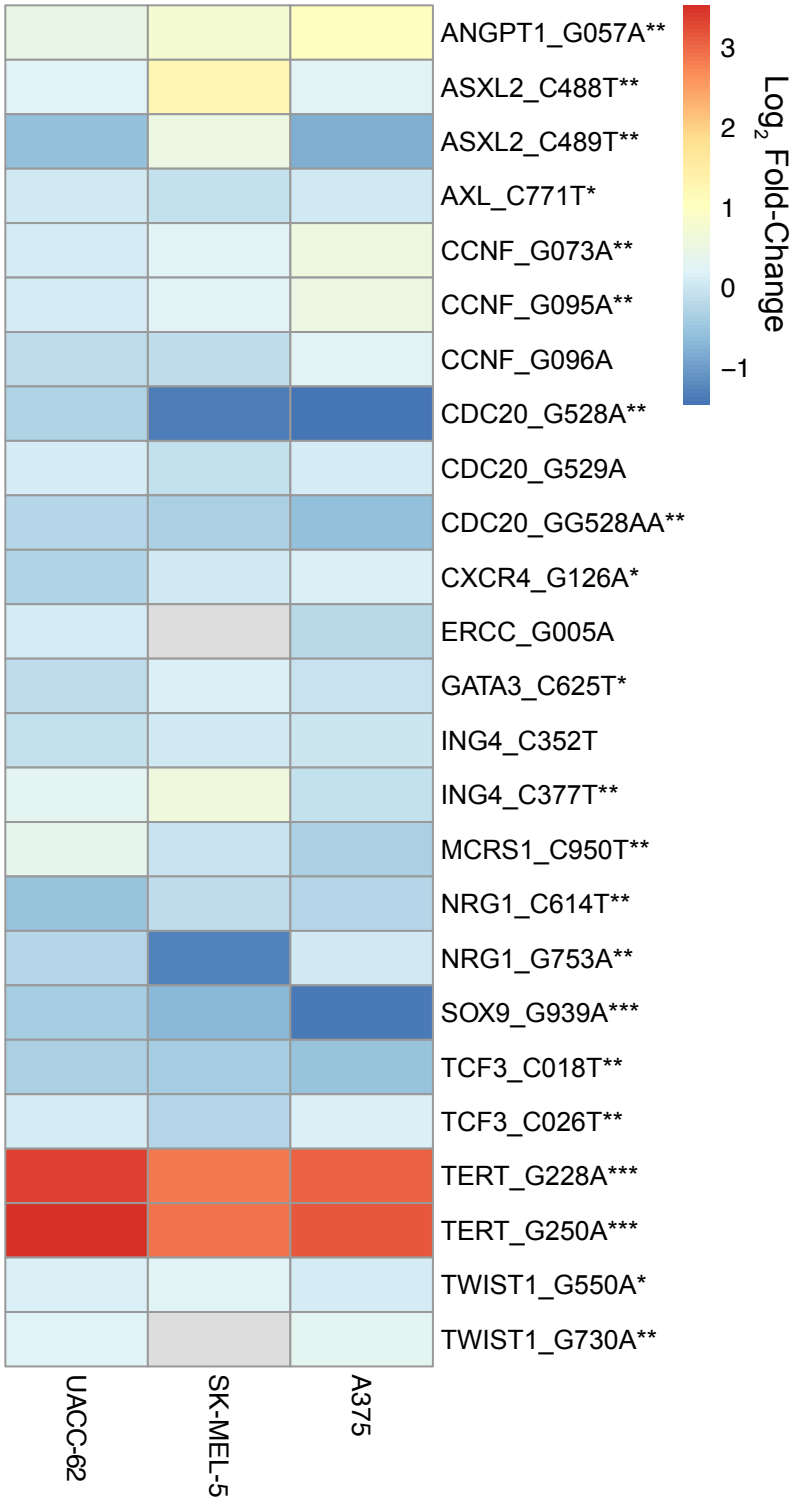


Figure 2. Validation of manually curated hotspots from preliminary analysis.

Heatmap depicting $\log_2$ fold-changes of variant reporter activity compared to WT reporter activity. Each row depicts a unique variant with the following notation: the variant-associated gene (by proximity) and the variant annotation (WT allele, the last 3 numbers of the nucleotide position, and the mutant allele). Stars depict statistical significance in one (*), two (**), or three (***) cell lines.

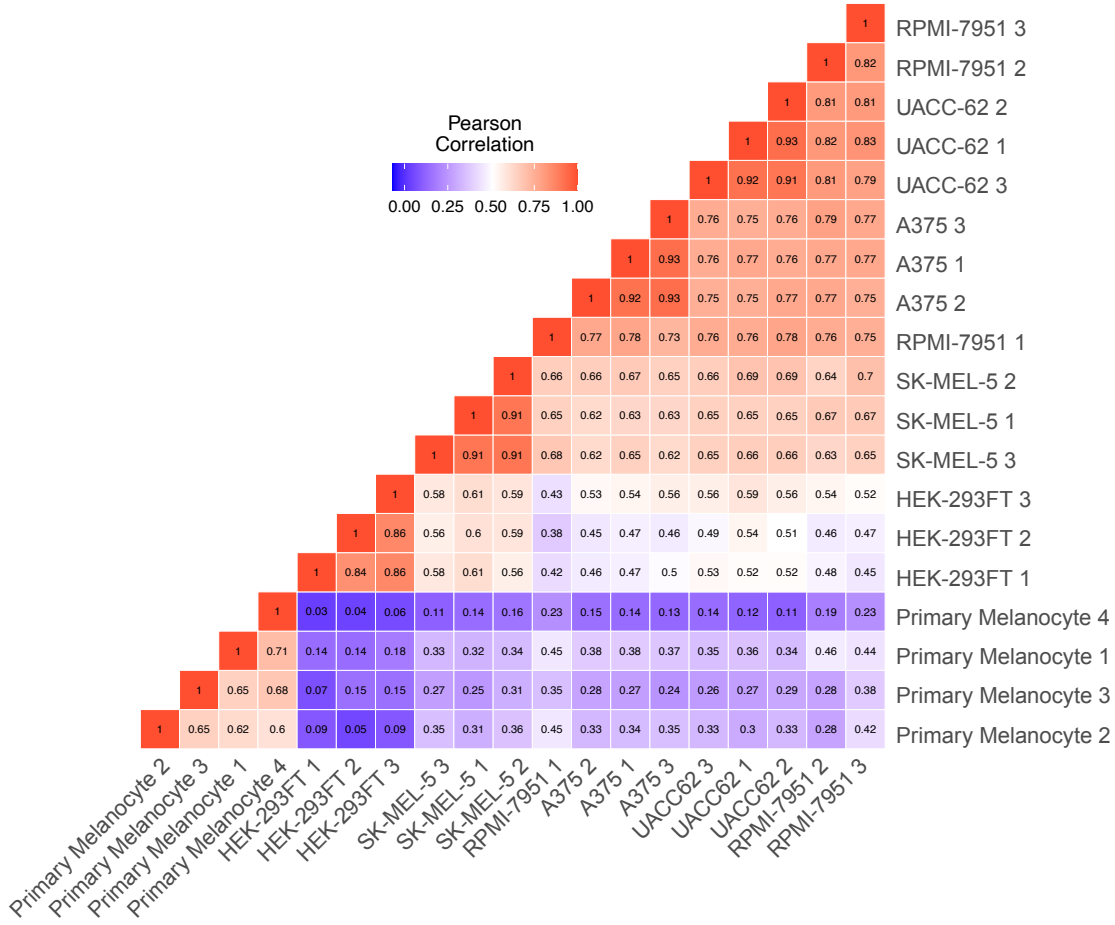


Figure 3. MPRA Tag Counts are correlated within biological replicates and within cell type.

Heatmap depicting Pearson correlation coefficient of tag counts (RNA/DNA) between each sample. Samples are clustered based on Euclidean distance and demonstrate highest correlation

within biological replicates, higher correlation within melanoma cell lines, and lowest correlation across sample types (i.e. between melanoma cell lines and HEK-293FT and primary melanocytes). Correlations between primary melanocyte replicates were low, suggesting low transfection efficiencies and were therefore discarded from the subsequent analyses.

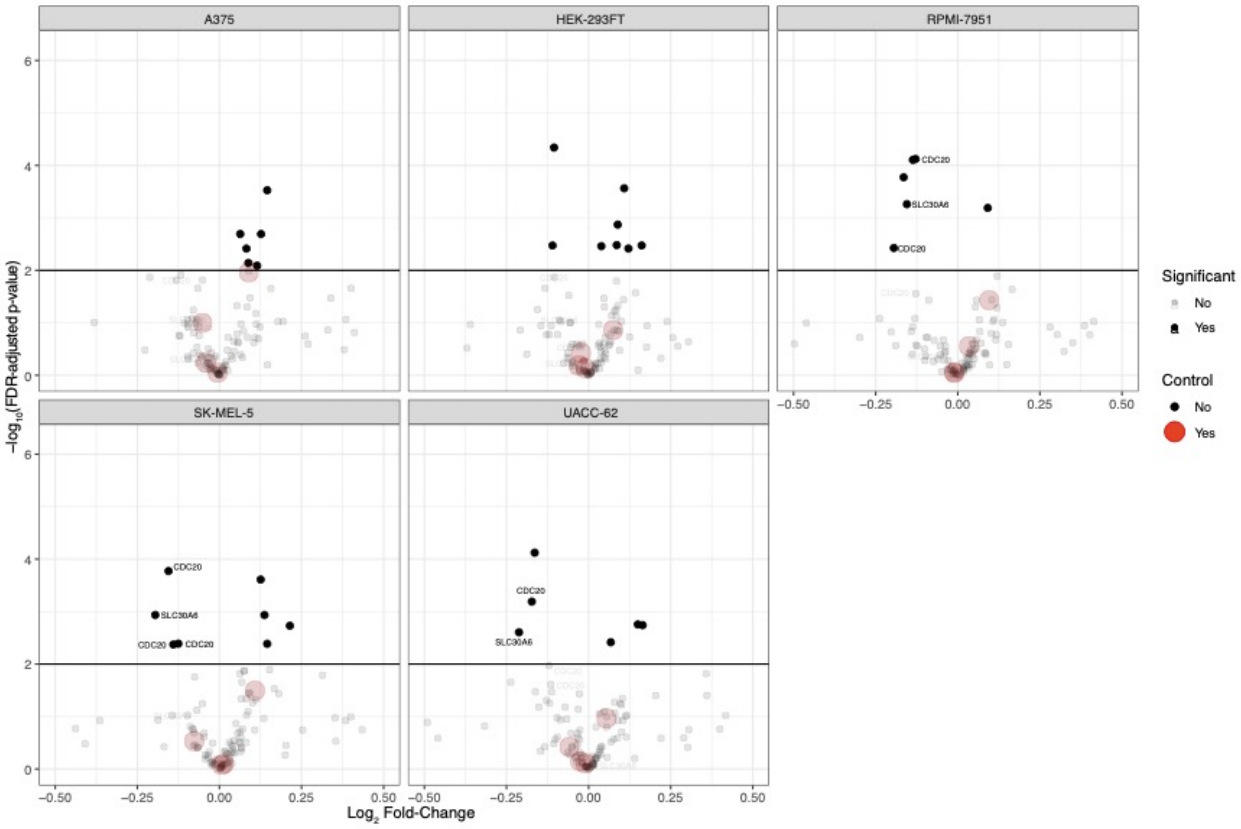


Figure 4. Functional analysis of 118 variants in statistically significant hotspots by massively parallel reporter assay.

Volcano plots of MPRA results from five different cell lines: A375 (BRAF mutant, neural crest-like), HEK-293FT (human embryonic kidney cells), RPMI-7951 (BRAF mutant, undifferentiated), SK-MEL-5 (BRAF mutant, melanocytic), UACC-62 (BRAF mutant, melanocytic/transitory). Black points are significant (FDR-adjusted p-value ≤ 0.05). Negative controls are depicted as large transparent red circles and are variants within hotspots that were

not statistically significant. Variants in the promoter of CDC20 and SLC30A6 are highlighted, as variants within these hotspots were significant in the majority of cell lines.

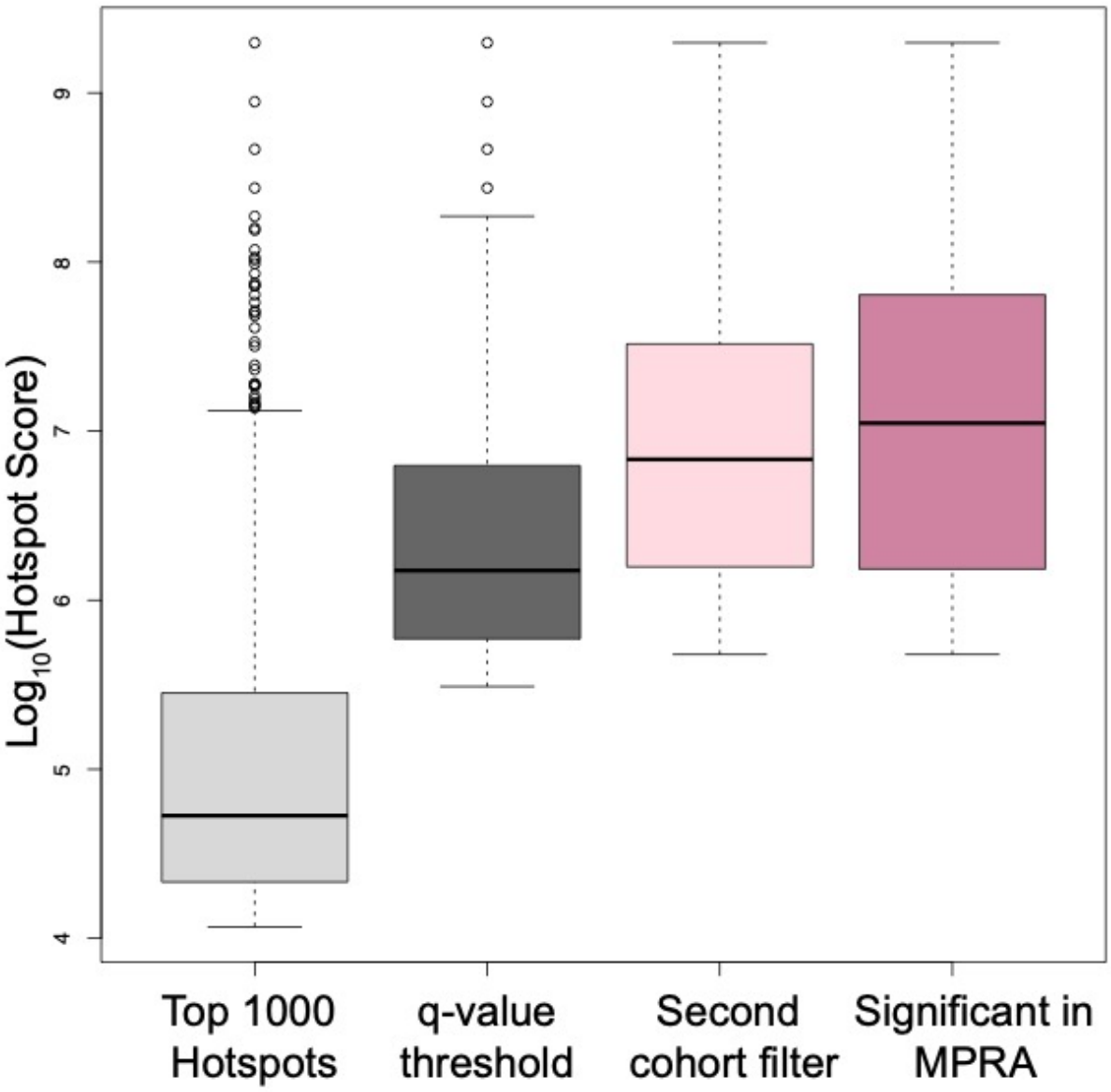


Figure 5. Validation of bioinformatic pipeline and variant selection criteria.

Variants to be validated by MPRA were selected on three criteria: (1) statistical significance, (2) presence in secondary cohort, and (3) detected in a total of 4 samples. Boxplots depict

distribution of $\log_{10}$ Hotspot Scores for top 1000 hotspots (grey), hotspots with a q-value = 0 (Top 140 hotspots, dark grey), hotspots containing variants detected in secondary cohorts (light pink), and hotspot scores of variants with statistically significant tag counts in the MPRA (dark pink).

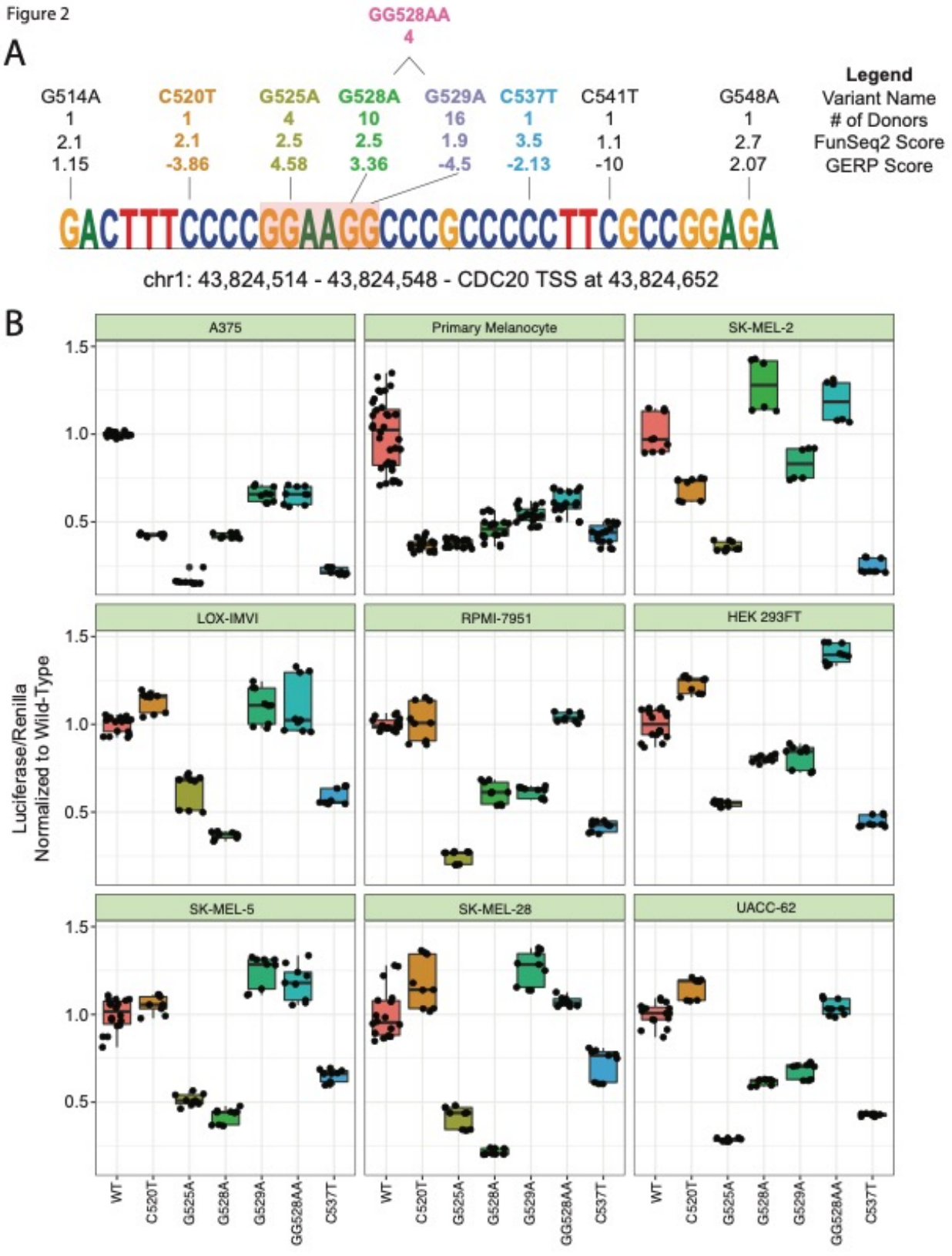


Figure 6. Functional analysis of recurrent CDC20 promoter variants.

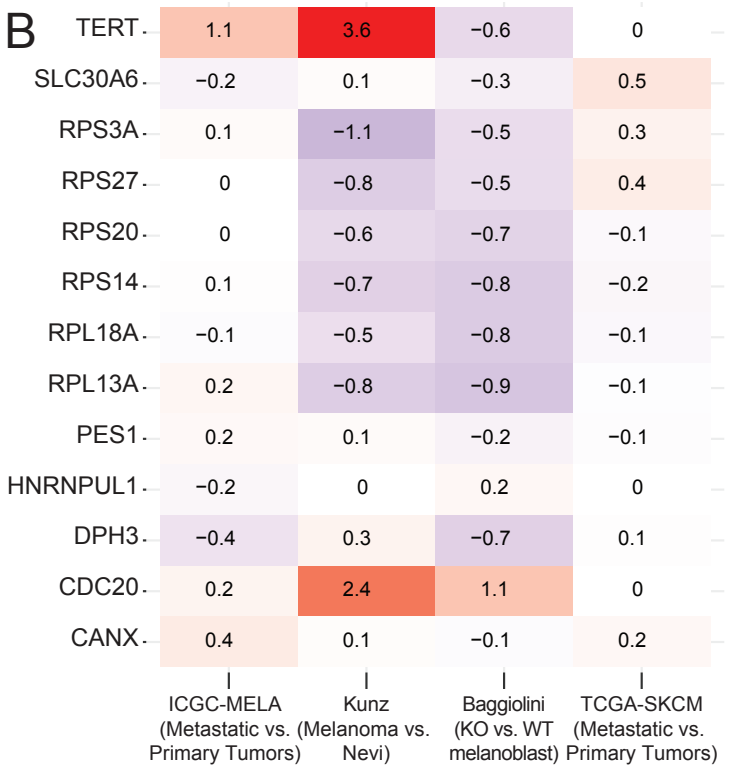
(A) The CDC20 promoter hotspot. All variants within the hotspot are denoted by name, # of donors with given mutation, FunSeq2 score, and GERP score. Co-occurring GG528AA double mutant is depicted above. Variants with colored text were validated by luciferase assay. (B) Altered CDC20 promoter activity for variants as assayed by luciferase reporter assays in melanoma (A375, SK-MEL-5, UACC-62, LOX-IMVI, RPMI-7951, SK-MEL-2, SK-MEL-28), primary melanocytes, and HEK 293FT cells. Boxplots depict normalized (to WT) luciferase assay results in these 3 different cell lines.

2.12 Supplemental Figures

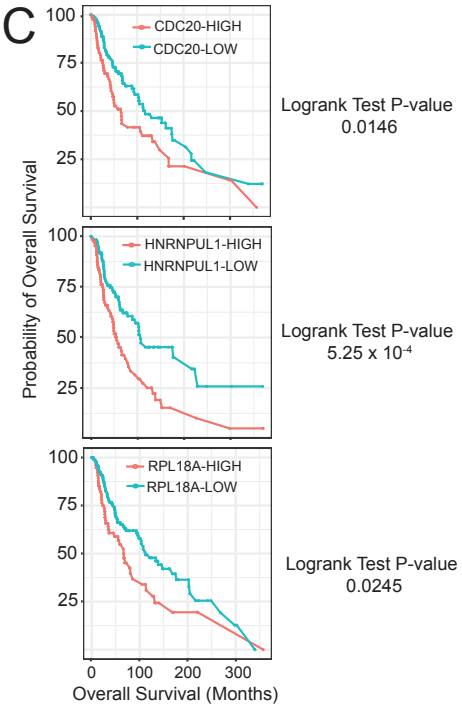
A

Name	Motif	p-value
Selected Motifs from distict families in pMRRs		
CTCF(Zf)/CD4+-CTCF-ChIP-Seq (Barski_et_al.)/Homer		1e-6092
Fra1(bZIP)/BT549-Fra1-ChIP-Seq (GSE46166)/Homer		1e-4839
Sox10(HMG)/SciaticNerve-Sox3-ChIP-Seq (GSE35132)/Homer		1e-472
Top 3 Motifs in Significant Hotspots		
Elk1(ETS)/Hela-Elk1-ChIP-Seq (GSE31477)/Homer		1e-178
Elk4(ETS)/Hela-Elk4-ChIP-Seq (GSE31477)/Homer		1e-164
ELF1(ETS)/Jurkat-ELF1-ChIP-Seq (SRA014231)/Homer		1e-161
Top 3 Motifs in 20 bp WT Sequence		
Elk1(ETS)/Hela-Elk1-ChIP-Seq (GSE31477)/Homer		1e-311
ELF1(ETS)/Jurkat-ELF1-ChIP-Seq (SRA014231)/Homer		1e-260
Elk4(ETS)/Hela-Elk4-ChIP-Seq (GSE31477)/Homer		1e-145
Motif in 20 bp Mutated Sequence		
Ik-1		1e-10

B



C



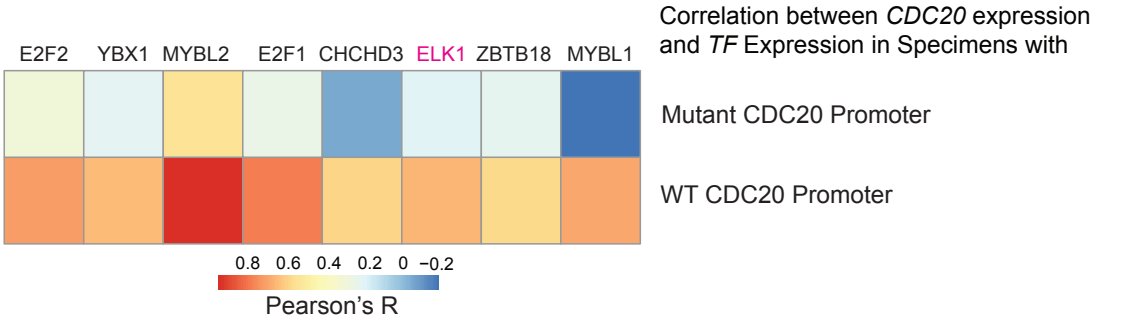
Supplemental Figure 1. Characterization of putative melanoma regulatory regions, hotspots, and associated genes.

(A) Table of selected motifs identified by Homer analysis. First section shows results for all pMRRs, regardless of whether region harbors a hotspot. To showcase diversity of transcription factors, we chose high-ranking motifs from three distinct transcriptional families. Second section shows top 3 motifs for pMRRs harboring statistically significant hotspots (707 hotspots, FDR-adjusted p-value < 0.05). Last two sections show top 3 motifs when input is a 20 bp sequence containing either the WT (top) or mutant (bottom) allele for all variants within statistically significant hotspots. (B) Log_2 Fold-Change for Top 13 genes in ICGC-MELA, TCGA-SKCM, Kunz, and Baggiolini. Order in which samples are written represents numerator and denominator (e.g. if higher in metastatic, positive fold-change). (C) Kaplan-Meier curves representing over-all survival rates for high (red) and low (blue) expressing tumors for the three genes listed). Data and p-values obtained from cBioPortal using OQL.

A

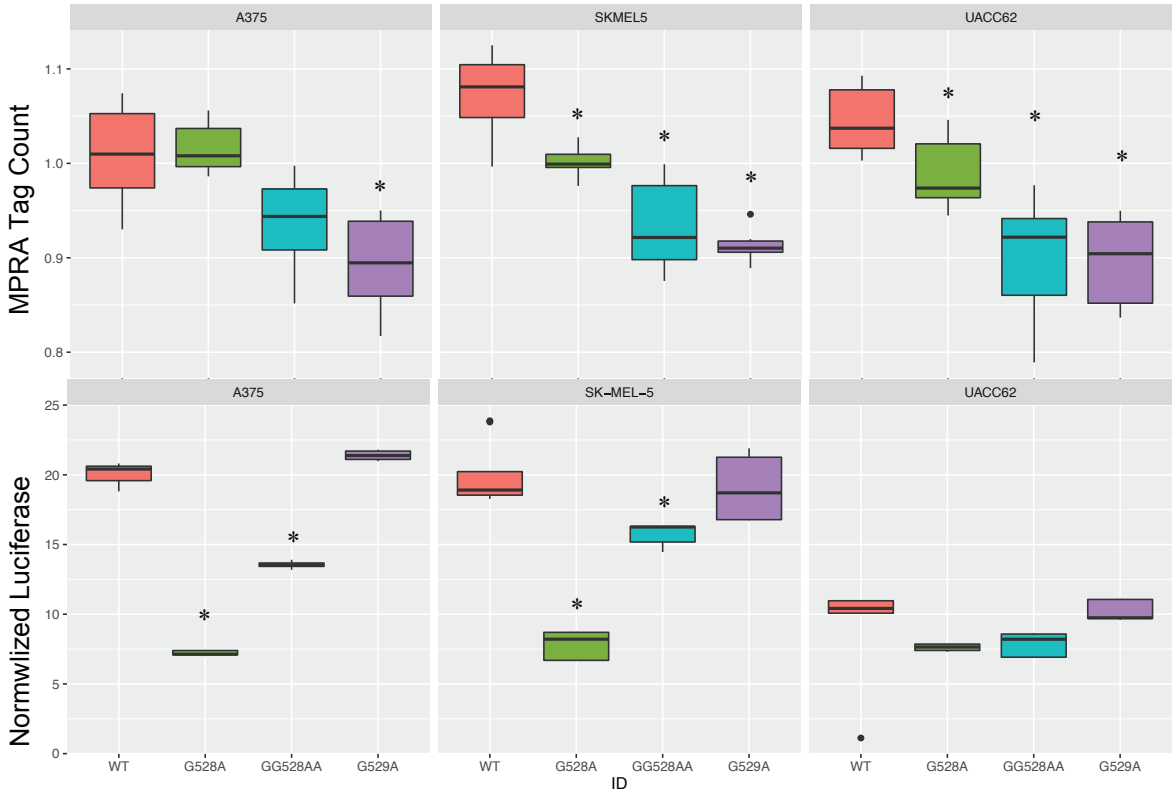
SNP ID	motifBreakR Results
G514A	GAIN: BATF, IRF4
C520T	GAIN: STAT, STAT1, STAT3, STAT5A, STAT5B LOSS: ELK1, PTF1A
G525A	GAIN: BCL6B, STAT, STAT1 , STAT5A, STAT5B LOSS: ELF1, ELF4, ELK1, ELK3, ELK4, ETS1, ETV1, ETV4, GABPA, NFE2L2, NR2C2, REL, TAF1
G528A	GAIN: FLI1, NFATC1, RELA LOSS: ELF1, ELF4, ELK3, ETS1, GABPA, HF1H3B, NFE2L2, SP4
C537T	GAIN: RXRA LOSS: HF1H3B, KLF13, SP2, SP4, ZBTB7A, ZNF219, ZNF281
C541T	LOSS: HF1H3B
G548A	GAIN: ELK1, FOXA1, REL, RELA, SP100

B



Supplemental Figure 2. Motif impact of recurrent CDC20 promoter variants.

(A) Table summarizing motifBreakR results. All transcription factors listed have strong and significant motif altering predictions. ETS transcription factors are colored in pink. (B) Heatmap of Pearson correlation values between TF (column) and samples with WT CDC20 promoters (top row) or mutant CDC20 promoters (bottom row.)



Supplemental Figure 3. Comparison of results obtained from luciferase assay and MPRA for the CDC20 promoter hotspot.

Top boxplot represents MPRA tag counts (RNA/DNA) and bottom boxplot represents normalized luciferase activity (relative to renilla) for plasmids containing the WT promoter, G528A, G529A, and GG528AA variants. Stars depict statistical significance (FDR-adjusted p-value < 0.05).

2.13 Supplemental Tables

Supplemental Table 1. List of datasets with peaks at CDC20 promoter hotspot.

Full dataset available upon request.

Accession	Cell Line	Experiment Type	CDC20 Promoter Hotspot
GSM3144751	MM032	H3K27Ac ChIP-seq	Yes

GSM1199827	Melanoma tumor tissue	H3K4me3 ChIP-seq	Yes
GSM1199828	Melanoma tumor tissue	H3K4me3 ChIP-seq	Yes
GSM1199830	Melanoma tumor tissue	H3K4me3 ChIP-seq	Yes
GSM1199831	Melanoma tumor tissue	H3K4me3 ChIP-seq	Yes
GSM1199832	Melanoma tumor tissue	H3K4me3 ChIP-seq	Yes
GSM3544062	MEL270	BRG1 ChIP-seq	Yes
GSM1953837	CJM	H3K27Ac ChIP-seq	Yes
GSM1953838	COLO679	H3K27Ac ChIP-seq	Yes
GSM1953839	LOX IMVI	H3K27Ac ChIP-seq	Yes
GSM1953840	SKMEL2	H3K27Ac ChIP-seq	Yes
GSM1953841	SKMEL30	H3K27Ac ChIP-seq	Yes
GSM1953842	UACC257	H3K27Ac ChIP-seq	Yes
GSM831874	MelJuSo	H3K4me3 ChIP-seq	Yes
GSM831875	MelJuSo doxorubicin	H3K4me3 ChIP-seq	Yes
GSM831876	MelJuSo etoposide	H3K4me3 ChIP-seq	Yes
GSM1024779	RPMI-7951	Dnase-Seq	Yes
GSM2178295	COLO829	CTCF ChIP-seq	Yes
GSM2178296	COLO829	CTCF ChIP-seq	Yes
GSM3664673	A375	DDX21 ChIP-seq	Yes
GSM1649543	neonatal foreskin	TFAP2A ChIP-seq	Yes
GSE102813	SKMEL-239	ATAC-seq	Yes
GSM1665991	SKmel147	H2A.Z ChIP-seq	Yes
GSM1665993	SKmel147	H2A.Z GFP ChIP-seq	Yes
GSM1665994	SKmel147	H2A.Z GFP ChIP-seq	Yes
GSM2199948	A375 shGFP	SMAD1/5/8 ChIP-seq	Yes
ENCFF862XVF	SK-MEL-5	Dnase-Seq	Yes
ENCFF600JNF	SK-MEL-5	Dnase-Seq	Yes

Supplemental Table 2. Test statistics and p-values of the top 13 hotspots.

Full list of hotspots available upon request.

	Donor Score	FunSeq2 Score	Hotspot Score	FDR-adjusted p-value	Nearest Gene
chr19:49990690-49990785	305.818182	2.05671532	628.980939	0	RPL13A (-73)
chr4:152020699-152020704	162	3.04812981	493.79703	0	RPS3A (-23)
chr19:17970680-17970687	196	2.07347159	406.400432	0	RPL18A (-1)

chr2:32390903-32390910	144	2.40809733	346.766015	0	SLC30A6 (-26)
chr8:56987107-56987150	105.125	2.93709915	308.762549	0	RPS20 (-61)
chr19:41769770-41769773	81	3.63082615	294.096918	0	HNRNPUL1 (-464)
chr1:153963167-153963241	126.5625	2.30430317	291.638371	0	RPS27 (-31)
chr1:43824513-43824549	152.1	1.76982066	269.189723	0	CDC20 (-153)
chr5:179125762-179125764	81	3.21893461	260.733704	0	CANX (-144)
chr22:30988168-30988213	122.5	2.10901816	258.354724	0	PES1 (-264)
chr3:16306501-16306534	152.818182	1.66400039	254.289515	0	DPH3 (-39)
chr5:149829293-149829346	75.5714286	3.23067364	244.146622	0	RPS14 (-10)
chr5:1295204-1295255	220.9	1.06309171	234.836958	0	TERT (-68)

Supplemental Table 3. List of manually-curated hotspots for preliminary analysis by luciferase assay.

Each row represents a hotspot designated by the most likely gene target. Subsequent information includes gene function, # of donors with mutations in ICGC dataset and Berger dataset, # of donors with significant ASE, gene expression pattern, and Brown's q-value.

Linked Gene	Gene Function	# of donors in ICGC (2017)	# of ICGC donors with sig. ASE	Gene Expression Up or down in melanoma vs melanocyte	# of donors in Berger (2012)	Brown's q-value (Rheinbay et al. 2020)
TERT	**Telomere Maintenance**	**49**	N/A	UP p=4.19E-11	17	9.97E-22
CDC20	**Cell Cycle Gene**	**39**	7	UP p=0.04	2	6.88E-12
ERCC1	**DNA Repair**	**13**	4	n.s.	2	NA
AXL	**Melanoma Progression and Resistance**	**22**	2	n.s.	0	NA
NRG1	**Nerve growth factor**	**8**	1	n.s.	1	NA
SOX9	**Neural crest factor**	**8**	2	DOWN p=0.03	0	NA
TWIST1	**EMT gene**	**8**	2	n.s.	0	NA
ANGPT1	**Angiogenesis**	**9**	1	DOWN p=0.01	1	NA
ASXL2	**Epigenetic Modifier**	**22**	7	n.s.	1	6.17E-09
MCRS1	**Several**	**18**	22	n.s.	3	NA
ING4	**Tumor Suppressor**	**40**	8	DOWN p=0.01	3	1.73E-06
CCNF, C16orf59	**DNA replication**	**39**	1	UP p=1.27E-07	8	9.00E-17
TCF3	**Neural crest factor**	**29**	1	UP p=0.03	1	NA

Supplemental Table 4. List of primers.

Pertains to entire dissertation.

Primer Name	Primer Sequence (5' – 3')
CDC20_genomicDNA_F	CCCCTCTCGTACCCTTCAA
CDC20_genomicDNA_R	GCTTTAACACGCCTGGCTTA
pGL3-CDC20_F	TGACTGGAGCTCCTCGTACCCTTCAAATCGCG
pGL3-CDC20_R	CAGTCACTCGAGGCTCCGAGCGCCTATTGG
WT_to_G525A_Q5_F	ACTTTCCCCGaAAGGCCCGCC
WT_to_G525A_Q5_R	CTCAGCTATCACGAGAGTCTAGCG
WT_to_G520A_Q5_F	CTGAGACTTTtCCCGGAAGGC
WT_to_G520A_Q5_R	CTATCACGAGAGTCTAGCG
WT_to_G537A_Q5_F	AGGCCCGCCCtCTTCGCCGGA
WT_to_G537A_Q5_R	TCCGGGGAAAGTCTCAGCTATCACGAGAGTCTAGC
XCC825c.h.CDC20.sp2	GGCGAAGGGGGCGGGCCTTCNGG
XCC825c.h.CDC20.sp4	CGAAGGGGGCGGGCCTTCCGNGG
XCC825c.hCDC20.DS.F	ttctgcaccgagttctgcat
XCC825c.h.CDC20.R	acgcctcttaaactctccgc
MPRA_Oligo_F	GTAGCGTCTGTCCGT
MPRA_Oligo_R	CTGTAGTAGTAGTTGGTCGAC
MiSeq_MPRA_Step2_Barcode_F	ACTGGAGTTCAGACGTGTGCTCTTCCGATCTgaggctgaagctgaaggac
MiSeq_MPRA_Step2_Barcode_R	TCTTTCCCTACACGACGCTCTTCCGATCTgatcagttatctagatccgg
MPRA_pEF1a_mCh_6bp_ext_F	AGCCATCGATCGgagtaattcatacaaaaggactcg
MPRA_pEF1a_mCh_6bp_ext_R	CCTGTCGACGTCctacttgtacagctcgtccat

Chapter 3: Towards understanding the role of CDC20 in melanoma

Preface

This chapter has been reproduced and adapted from the following preprint:

Godoy, P. M., Zarov, A. P. & Kaufman, C. K. Functional analysis of recurrent non-coding variants in human melanoma. *Biorxiv* 2022.06.30.498319 (2022) doi:10.1101/2022.06.30.498319.

3.1 Abstract

One of the most intriguing hotspots identified in Chapter 2 is upstream of the *CDC20* promoter. *CDC20* is up-regulated in many melanomas and is associated with overall worse survival. Unexpectedly, characterization of the variants in nine cell lines, including primary melanocytes, suggests that variants, which are present in 27% of cutaneous melanomas, decrease reporter activity. We went further to genome engineer a small indel in a human melanoma cell line and show decreased CDC20 expression levels. Additionally, most CDC20 variants are present in primary melanomas and lymph node metastases, which is often the first site of regional metastatic disease but are not as widely detected in distant metastatic tumors. Guided by the recently identified MITF-rheostat model, where high levels of MITF are associated with proliferative/melanocytic phenotype and low levels represent a more invasive state, we explored both CDC20-high and CDC20-low phenotypes using five RNA-sequencing melanoma cohorts, including our genome-engineered cell lines. We hypothesized that CDC20-high levels would be favored in later stages of melanoma, based on our results and previous work, and that CDC20-low would be favored in melanoma initiation. As expected, the CDC20-high phenotypes were ubiquitously enriched for pro-metastatic gene signatures. This was confirmed by migration assays on our WT and CDC20 promoter indel lines where we observed decreased migration capabilities in our mutant strains compared to WT. Most of the CDC20-low samples across the 5 cohorts were enriched for aneuploidy gene signatures, and interestingly, upregulation of certain key neural crest transcription factors, associating low levels of CDC20 with a more de-differentiated state.

3.2 Introduction

CDC20 is a highly conserved and essential regulator of the cell cycle. Deletion of CDC20 leads to arrest at metaphase in 2-day old embryos in mice[130]. It is a catalytic co-activator of the

Anaphase Promoting Complex/Cyclosome (APC/C) which is a large multi-subunit E3 ubiquitin ligase. CDC20 contains many degrons (i.e. protein 'motifs' that recognize and target substrates for degradation) for proteins including Cyclin B1 and Securin[131]. Degradation of Cyclin B leads to the inactivation of mitotic cyclin-dependent kinases and degradation of securin leads to activation of separase, causing separation of sister chromatids. The combined effect of the two mark the beginning of anaphase and end of mitosis[132]. The spindle assembly complex (SAC) binds to CDC20 and MAD2 (forming the mitotic checkpoint complex, MCC) at the kinetochore to inhibit APC-CDC20 until all sister chromatids are attached to the kinetochore[133]. Cyclin-dependent kinases also phosphorylate CDC20 which reduces the binding efficiency with APC/C, therefore preventing APC/C-CDC20 formation and G2-to-M transition[134]. This is an imperative function as the inability to inhibit APC-CDC20 leads to rapid tumorigenesis in mice[135].

CDC20 has also been implicated in functions outside of the SAC and MCC, including the mediation of chromatin loop formation through ubiquitylation of hnRNPU[136], maintaining stemness in human primary keratinocytes and human embryonic stem cells[137,138], attenuating cardiac hypertrophy[139], maintenance of the primary cilia[140], and regulating dendrite morphogenesis in neurons[141,142].

Expression of CDC20 is prognostic in many cancers[131]. High levels of CDC20 are associated with aggressive tumors in multiple cancer types and increased infiltration of immune cells and cancer-associated fibroblasts[143–145]. APC/C-CDC20 regulates SOX2 to enhance tumor migration and invasion in human glioblastoma[146]. One recent reporter identified germline mutations in CDC20 that accelerated tumor onset in conjunction with Myc overexpression and mitotic slippage[147]. Another paper identified another germline missense mutation that led to premature aging and cancer, likely due to an increase in the number of aneuploid cells[148]. siRNA-

mediated knockdown in multiple epithelial cell lines led to increased chromosomal instability and better survival in low pH culture conditions, a proxy for tumor-like conditions due to the Warburg effect[149].

Overall, CDC20 plays important and essential roles in diverse functions in cell biology. Support for CDC20 as an oncogene in more advanced tumors is emerging. However, because the CDC20 promoter mutations decrease reporter activity in many of the cell lines we assayed, we hypothesize that in certain contexts decreased expression of CDC20 could favor tumor formation. In this Chapter, we explore this hypothesis by investigating the epigenetics of the CDC20 promoter, the clonality and evolutionary history of the CDC20 promoter hotspots, and changes in cancer phenotypes upon genome engineering of an indel at the CDC20 promoter in a human melanoma cell line.

3.3 Materials and Methods

Correlation between TFs and CDC20

The ICGC RNA-sequencing cohort consists of 56 samples from 46 donors. 13 samples (from 10 donors) contained a variant in the CDC20 promoter. The remaining 43 samples were WT samples. We downloaded a list of all transcription factors from http://humantfs.ccbr.utoronto.ca. We calculated the Pearson correlation between *CDC20* and every TF in either the samples with WT or mutant CDC20 promoter variants. We selected those that had correlations between the TF and the WT CDC20 promoter samples greater than 0.5 and less than 0.2 for those between the TF and mutant CDC20 promoter samples.

Calculation of Variant Allele Frequencies

Mutation calls for SNVs and indels from the MELA-AU cohort were downloaded from dcc.icgc.org after receiving DACO approval. We calculated variant allele frequencies by dividing

the number of reads containing the variant divided by the total counts for each specimen. We then stratified specimens by tumor subtype (primary melanoma, lymph node metastasis, distant metastasis). VAFs from recurrent tumors or from cell lines derived from tumors were not considered. When plotting VAFs, the CDC20 promoter variants and the $BRAF^{V600E}$ variants were plotted separately. Variant allele frequencies for $NRAS^{Q61K}$ and $NRAS^{Q61R}$ and for TERT G228A and G250A were combined into one boxplot each.

Kunz RNA-sequencing analysis

We downloaded DESeq2-normalized read counts from GSE112509 for the Kunz cohort[117]. We classified each sample as CDC20-low, medium, or high based on *CDC20* expression. Samples with less than the 25th percentile of *CDC20* expression were classified as low, while samples with greater than the 75th percentile of *CDC20* expression were classified as high. All other samples are classified as having medium CDC20 expression.

We performed gene set enrichment analysis (GSEA) on the 20 CDC20-low and 20 CDC20-high samples using all gene sets in MSigDB (https://www.gsea-msigdb.org) that contained the keyword "melanoma". We used the following parameters: 1000 permutations, the phenotypes were always set as low versus high (ergo enrichment scores are positive for CDC20-low, negative for CDC20-high), and permutations were performed on the gene set.

We manually curated a list of 20 neural crest transcription factors from two previously published sources[31,150]. DESeq2-normalized read counts for these genes were used to construct the heatmap. Counts across every gene were scaled by setting the parameter "scale" to "row" in the heatmap plotting function *pheatmap*. Genes and samples were clustered using Euclidean distance.

Genome Engineering of A375

A375 cells were nucleofected on a Lonza 4D nucleofector according to manufacturer recommendations (P3 solution, nucleofection program EH-100). Each nucleofection was performed with 1 x 10^5 cells, 0.75 μL Cas9 Protein at 10 μg/μL (IDT v3 Cas9 protein, glycerol-free, # 10007806), and 0.75 μL of each sgRNA at 100uM (IDT) suspended in IDT Duplex Buffer (IDT, # 11-05-01-03) (Supplemental Table 5). Sham-nucleofections for WT A375 Cas9 controls were nucleofected with an equal volume of blank PBS. After nucleofection, cells were seeded into 500 μL of DMEM complete in a 24-well plate at standard incubator conditions.

72 hours post-nucleofection, cells were harvested, and split into 6-well culture for expansion and into lysis buffer for DNA extraction (homemade by GESC, formulation identical to Lucigen Quick-Extract buffer). PCRs were performed with Platinum Superfi II 2x master mix (Thermofisher, #12368010) and primers against the sgRNAs target site (Chapter 2, Supplemental Table 4). PCR products were sequenced by NGS using Illumina.

After confirmation of cutting activity at, the pools were single-cell sorted using a Sony SH800 cell sorter at 1 cell per well into 4 x 96-well plates with 100uL of DMEM, with 50% conditioned media, 5 μM Rock Inhibitor, and 100 μM sodium pyruvate. Plates were allowed to grow for ~10 days, then clones were harvested and re-screened using PCR primers against the targeted locus (Supplemental Table 5). Homozygous knockout clones were identified based on the presence of deletion junction and absence of the target locus. WT A375 Cas9 controls were sequenced at all gRNA target sites to confirm wild-type genotype. Homozygous knockout clones and wild-type Cas9 control clones were expanded, checked by STR profiling, tested for mycoplasma contamination, and used for subsequent experiments.

Cell Viability Assay of A375 CDC20 Promoter Knock-outs and Controls

For each strain (A3, A10, and the wild-type Cas9 control), we seeded 1500 cells per well in a clear-bottom 96-well plate (Corning, #3903) in DMEM media containing 10% fetal bovine serum and 1X Pennicilin/Streptavidin (DMEM complete), DMEM complete with 30 nM dabrafenib (Selleck Chemicals, S2807), or DMEM complete with 1% DMSO. To measure viability, we used CellTiterGlo (Promega, G7570) as per the manufacturer's protocol. Plates were read on a GloMax 96 Microplate Luminometer (Promega) using the standard CellTiterGlo program.

Cell Migration Assay of A375 CDC20 Promoter Knock-outs and Controls

Scratch assays were performed by seeding 1 million cells per well in a 6-well plate in DMEM complete media. Using a P200 pipette, we scratched the plate at indicated positions. Cells were washed with 1X PBS and imaged on a Nikon Eclipse Ts2. Cells were then plated with DMEM media with 1% FBS and 1X Pen/Strep. On the following day, cells were washed with 1X PBS and imaged.

RNA-sequencing of A375 CDC20 Promoter Knock-outs and Controls

300,000 cells of the parental A375 (in duplicate), two WT CRISPR/Cas9 clones (one replicate each), A3 (in duplicate), and A10 (in duplicate) were seeded on a 6-well plate. On the following day, we isolated RNA using the Qiagen RNeasy Plus Mini Kit (Qiagen, 74134). Samples were submitted to the Genome Technology Access Center at the McDonnell Genome Institute at Washington University School of Medicine for library preparation and sequencing.

Total RNA integrity was determined using Agilent Bioanalyzer or 4200 Tapestation. Library preparation was performed with 5 to 10ug of total RNA with a Bioanalyzer RIN score greater than 8.0. Ribosomal RNA was removed by poly-A selection using Oligo-dT beads (mRNA Direct kit, Life Technologies). mRNA was then fragmented in reverse transcriptase

buffer and heating to 94 degrees for 8 minutes. mRNA was reverse transcribed to yield cDNA using SuperScript III RT enzyme (Life Technologies, per manufacturer's instructions) and random hexamers. A second strand reaction was performed to yield ds-cDNA. cDNA was blunt ended, had an A base added to the 3' ends, and then had Illumina sequencing adapters ligated to the ends. Ligated fragments were then amplified for 12-15 cycles using primers incorporating unique dual index tags. Fragments were sequenced on an Illumina NovaSeq-6000 using paired end reads extending 150 bases. RNA-seq reads were then aligned and quantitated to the Ensembl release 101 primary assembly with an Illumina DRAGEN Bio-IT on-premise server running version 3.9.3-8 software.

Read counts were normalized using DESeq2, comparing WT to mutant strains[151]. Principal component analysis was performed using the *plotPCA* function in the DESeq2 package. The heatmap was generated with *pheatmap* using z-score normalized counts of the manually curated list of 20 neural crest transcription factors with FDR-adjusted p-values < 0.1 (between WT and mutant samples).

Gene set enrichment analysis was performed as previously described using the 25th and 75th quantile to establish CDC20-low and CDC20-high expression groups, respectively. To generate heatmaps of all 4 cohorts, we downloaded Kunz, TCGA, and ICGC RNA-sequencing datasets as previously described. For the Wouters cohort, we downloaded normalized counts from bulk RNA-sequencing of 33 melanoma cultures[77] (GSE134432). We calculated the mean across all samples classified as CDC20-low, medium, or high and plotted z-score normalized counts using the *pheatmap* function. Z-scores were calculated by scaling across rows, or genes.

Karyotyping of A375 CDC20 Promoter Knock-outs and Controls

Karyotyping and analysis was performed at the Cytogenetics and Molecular Pathology Laboratory at Washington University School of Medicine. The cytogenetic test/ karyotype analysis was performed to assess aneuploidy (gains and losses of whole chromosomes), structural changes (chromosomal translocations, inversions, segmental deletions and duplications). This assay involves growing of cells in appropriate culture medium, hypotonic treatment, fixing cells, staining cells with GTG banding and microscopic examination. Twenty cells are counted for enumerating the number of chromosomes in a metaphase spread. Three of these metaphase spreads are digitally processed to produce a detailed karyotype/karyogram to perform a detailed study (analysis) for variant counts and structural aberrations. Analyzing a metaphase is defined by band-by-band comparison between chromosome pairs.

3.4 Results

3.4.1 Motif analysis of the CDC20 promoter variants

We used motifBreakR to identify possible transcription factor binding motifs that are destroyed by the presence of the CDC20 promoter variants. As expected, the four variants closest to the core ETS motif are predicted to break sites for various ETS transcription factors, with G525A showing the largest reduction in ETS motifs of any variant (Supplemental Figure 1A).

We leveraged the RNA-sequencing data from a subset of the ICGC-MELA cohort to better predict the transcription factor (TF) that may have dysregulated binding in the mutated CDC20 promoter samples. We reason that if TF_i binds to the CDC20 promoter at the core ETS motif that is disrupted by G525A, G528A, G529A, and GG528AA, *CDC20* expression will correlate with TF_i expression in WT samples but not in samples with the disrupted core ETS motif. Therefore, we calculated the Pearson correlation between every TF[48] and *CDC20* in both WT and mutated samples. We identified 8 TFs that had high correlation of their expression (Pearson Correlation > 0.5) with

CDC20 expression in WT samples but low correlation in mutated samples (Supplemental Figure 1B). These include E2F1 and E2F2, which are known to regulate genes involved in cell cycle progression[152] and, interestingly, the ETS family TF ELK1.

3.4.2 CDC20-associated variants appear to be present as early clonal events but drop-out in distant metastatic melanomas

We next sought to understand the degree of clonality of the CDC20 promoter variants in sequenced tumors. Variant allele frequencies (VAF) can indicate how clonal a variant is by associating higher VAF with earlier appearance of the variant[153]. We compared the CDC20 promoter VAFs to those of the $BRAF^{V600E}$, $NRAS^{Q61K/R}$, and TERT G228A and G250A variants since these have all been reported to occur early[154] (Figure 1A, Supplemental Figure 1B). In primary tumors, the BRAF, NRAS, and TERT variants are detected at median frequencies around 0.30, 0.32, and 0.41, respectively (Figure 1A). The median VAFs for the two most common CDC20 promoter variants G528A (0.34) and G529A (0.33) are only slightly lower than TERT and slightly higher than BRAF and NRAS, suggesting G528A and G529A mutations as early occurring events in melanomagenesis.

Since CDC20 promoter variants led to a decrease in reporter activity and CDC20 has been shown to be essential for migration in melanoma mouse models[155], we hypothesized that promoter variants might decrease or disappear in later metastases. The G528A variant is detected mostly in lymph node metastases, often the first site of metastasis (n=6/11) and primary tumors (n=4/11). Only one distant metastatic sample out of a total of 51 had the G528A variant (Supplemental Figure 1C). Unlike G528A, G529A is detected across all stages and at median variant allele frequencies like those seen in earlier stages (Supplemental Figure 1B). Interestingly in A375 melanoma cells, G529A decreases reporter activity less than G528A, consistent with a model in which G529A,

which is less deleterious to *CDC20* expression, does not seem to drop out in distant metastases like G528A, which lowers reporter expression more and, in agreement with published work[155], thus would be disfavored in later metastases (Figure 1B).

3.4.3 Distinct transcriptional programs emerge in nevi and melanoma in a CDC20 dosage-associated manner

To begin to pry into the differences between a CDC20-low and CDC20-high phenotype and how these differences may drive or support cancer progression at different stages of melanoma, especially those representative of the earliest states of melanoma, we utilized the Kunz cohort of 23 nevi and 57 primary melanomas that were RNA-sequenced[117]. We stratified all samples by *CDC20* expression with CDC20-high and CDC20-low classifications based on the 75th and 25th percentile of CDC20 expression, respectively. Remaining samples were classified as medium expression (Figure 1B).

We performed gene set enrichment analysis (GSEA) on CDC20-low and CDC20-high samples. As expected, based on prior studies, genes in the CDC20-high samples are enriched for gene sets associated with metastasis[155] (Figure 1C, Supplemental Table 1). CDC20-low samples have an enrichment of genes expressed in uveal melanomas with high aneuploidy[156] (Figure 1D, Supplemental Table 1). This is in line with previous work that have shown increased aneuploidy in models with knockdown or mutated CDC20[135,149,157].

We next sought to understand whether CDC20-low samples were enriched for key neural crest TFs, some of which (e.g. *SOX10*) are known to play important roles in melanoma initiation[47,48,158]. Hierarchical clustering using 20 neural crest TFs clustered samples into 3 major groups: samples with mostly low *CDC20* (Group C, median $\log_2$ expression = 6.7), samples with mostly high *CDC20* (Group B, median $\log_2$ expression = 9.4), and samples with medium *CDC20*

expression (Group A, median $\log_2$ expression = 9.0, Figure 1E). This indicates differences in expression of these neural crest transcription factors is associated with *CDC20* expression. Surprisingly, CDC20-low samples cluster more closely with CDC20-high than CDC20-medium, despite having a larger difference in *CDC20* expression (Figure 3E).

Group C is made up of 13 nevi and 7 melanomas, 16 of which are classified as CDC20-low and four as CDC20-medium. This sample group has relatively high expression of genes prevalent in premigratory neural crest cells (*ETS1*, *SOX5*, *SOX9*, and *TFAP2B*) and melanocyte lineage specifiers (*SOX10* and *MITF*)[31]. Group A contains 6 nevi and 14 melanomas, 1 of which is classified as CDC20-low, 2 as CDC20-high, and 17 as CDC20-medium. This group has relatively high expression of *MYB* and *TFAP2B* which are prevalent in premigratory neural crest, *MSX1* (neural plate border), and *MAFB*, which is required for migrating cardiac neural crest cells[159]. Group B contains 32 melanomas and 4 nevi, 2 of which are classified as CDC20-low, 15 as CDC20-high, and 19 as medium. This group did not have relatively high expression across the group of any specific subset of transcription factors as seen with Group A and Group C. However, several isolated samples had relatively high expression of *SOX10*, *TEAD2*, and *RXRG*, as in Group C, and relatively high expression of *TFAP2A*, *MSX2*, and *HES4*, as in Group A. Notably, many of the neural crest transcription factors that are relatively higher in Group C are also known oncogenes in melanoma, particularly *SOX10*, *MITF*[55], *ETS1*[160], and *MYC*[161].

3.4.4. Genome-engineered CDC20 promoter mutants have altered phenotypes and transcriptional profiles

Thus far, we have identified variants prevalent in the CDC20 promoter in melanoma tumors that by luciferase reporter assay reduce transcriptional activity and see distinct profiles of neural crest transcription factors in naturally occurring human melanoma tumors and nevi

associated with high, medium, and low levels of *CDC20*. To determine the effect of CDC20 promoter mutations on key cancer phenotypes and gene expression programs, we generated two CRISPR/Cas9-engineered A375 melanoma cell lines termed A3 and A10 (Figure 2A). The A3 line contains an indel on both alleles, both of which have the G528 and G529 nucleotides deleted. One allele retains the core GGAA motif while the other does not. The A10 line contains a larger deletion that completely removes the G525, G528, and G529 mutations, as well as the core ETS motif in both alleles (Figure 2A).

Both mutations decrease *CDC20* expression by 2.0-fold on average as detected by RNA-sequencing (FDR-adjusted p-value = 1.8×10^{-40}, Figure 2B). The A3 strain has slightly lower *CDC20* expression than A10 despite having a smaller deletion and the retention of one core ETS motif (Figure 2B). Principal component analysis shows a separation along PC1 between the WT parental A375 line (high *CDC20*), the WT Cas9 control A375 line (high *CDC20*), and the mutant A3 and A10 line (low *CDC20*, Figure 2C).

Because CDC20 is an essential component of the cell cycle, we wondered if decreased *CDC20* levels would lead to decreases in cell viability. We assayed viability in the presence of media containing serum, media containing serum and DMSO, and media containing serum and 30 nM dabrafenib (MAPKi) daily over the span of 6 days (Supplemental Figure 3A). A10 grows slightly slower than A375 and A3, despite having slightly higher levels of *CDC20* than A3 (Supplemental Figure 3A). No change in growth rates between A3 and A375 were observed (Supplemental Figure 3A).

We performed GSEA on the A375 WT and CDC20 promoter indel lines A3 and A10 using the same gene sets as above (Figure 1C and Figure 1D). There is significant enrichment of genes upregulated in WT vs CDC20 promoter indel cells for genes in the Winnepenninckx Melanoma

Metastasis gene set[162] (Figure 2D). While not statistically significant, we did see slight enrichment of genes upregulated in the mutant A375 line in the Ehlers aneuploidy gene set[156] (Supplemental Table 1). To determine whether these gene sets are enriched in other melanoma cohorts, we performed GSEA on three other cohorts that underwent RNA-seq using the same *CDC20* stratification as before (Supplemental Table 1). All cohorts with high *CDC20* expression have statistically significant enrichment of genes associated with metastasis gene sets, while all CDC20-low cohorts except for TCGA-low have enrichment of genes in gene sets associated with low metastasis (Supplemental Table 1). The Kunz-low, TCGA-low, and Wouters-low samples have enrichment of genes in the Ehlers' Aneuploidy gene set (Supplemental Table 1). Together these analyses show association of CDC20-high states with metastasis and CDC20-low states with aneuploidy across multiple cohorts.

To see whether our A375 promoter indel lines have altered migration capabilities as suggested by the results of GSEA and the literature, we performed a scratch assay and observed decreased migration capabilities suggesting that, at least in this context, reduced levels of CDC20 affect migration more so than viability (Figure 2E). Because we see enrichment of an aneuploidy gene set in CDC20-low samples, we checked the A375 mutant lines for increased aneuploidy but did not observe any in a karyotyping analysis (Supplemental Figure 3B).

Finally, we performed hierarchical clustering using the 9 differentially expressed (FDR-adjusted p-value < 0.05) neural crest transcription factors from the 20 that were previously found to be correlated with differential CDC20 levels in naturally occurring nevi and melanomas (Figure 2E). We observe similar patterns of neural crest transcription factor levels in high and low CDC20 samples. *SOX10*, *SOX5*, *RXRG*, and *TFAP2B* are consistently high across the A375 mutant lines

and CDC20-low samples in the Kunz cohort. *TFAP2A*, *TFAP2C*, and *FOXD1* were upregulated in WT A375 cells as well as in CDC20-medium or CDC20-high samples in the Kunz cohort.

Because we observe a similar association between neural crest transcription factor expression and *CDC20* expression between our CDC20 promoter indel cell lines and a cohort of nevi and melanoma, we looked for such patterns in the ICGC-MELA, TCGA-SKCM, and Wouters samples as well. We stratified all samples by *CDC20* expression (low, medium, and high) and performed hierarchical clustering using the 20 neural crest transcription factors (Supplemental Figure 4A). We specifically looked at the 9 neural crest transcription factors that were differentially expressed between the CDC20 promoter indel and WT cell lines (Supplemental Figure 4B). *FOXD1* and *TFAP2C*, which are upregulated in WT lines compared to mutant, are also relatively highly expressed in CDC20-high or CDC20-medium samples in 3 out of the 4 cohorts. *TFAP2B*, *SOX5*, *RXRG*, and *MYC*, which are upregulated in the CDC20 promoter indel cells, are upregulated in CDC20-low or CDC20-medium samples in most or all cohorts. *TFAP2A*, *SOX10* and *ETS1* were relatively highly expressed in multiple CDC20-expression groups. We also observed an up-regulation of genes in the melanocytic and neural crest subpopulations as identified by scRNA-seq of a melanoma tumor in the CDC20 promoter indel lines[75]. In conclusion, we observe a consistent trend across 5 cohorts between certain neural crest transcription factors and *CDC20* expression, suggesting a phenotype switch from an "undifferentiated" state to a more "differentiated" state reminiscent of the melanocytic and neural crest-like subpopulations within a melanoma tumor.

3.5 Discussion

We have shown that seven of the most common variants in the CDC20 promoter reduce reporter activity in multiple melanoma cell lines, HEK 293FT, and in primary melanocytes. Most

of these variants are not detected in distant metastases, which in agreement with previous work, suggests that lower levels of *CDC20* may be disfavored in later metastases[155]. Therefore, we propose a dosage-dependent role of *CDC20* on melanoma onset and progression, in which low levels of *CDC20* are important in early stages of melanoma but higher levels may be important for later stages of melanoma.

Because we hypothesize that low *CDC20* levels are critical in early stages of melanoma, we looked for co-expression of CDC20 with neural crest transcription factors, a state observed in the first malignant cells of melanoma. We show both in our genome engineered cells and in four other naturally occurring human melanoma cohorts that samples with low *CDC20* can have relatively high expression of certain neural crest transcription factors, such as *SOX10*, *RXRG*, *SOX5*, *TFAP2B*, and *MYC*. Notably, SOX10 and MYC have already been established as oncogenes[47,48,158]. RXRG reportedly drives a neural crest stem cell state that is resistant to treatment in a subset of cells in melanoma minimal residual disease[76]. TFAP2B regulates the melanocyte stem cell lineage in adult zebrafish[163], and SOX5 has been shown to inhibit MITF. Taken together, we hypothesize that low levels of CDC20 in melanocytic nevi may support a transcriptional program associated with a partially de-differentiated, neural-crest like state.

Meanwhile, as *CDC20* levels increase, we observe an increase in a different subset of neural crest transcription factors, including *FOXD1*, which is known to impair migration and invasion in melanoma models when knocked-down[164]. Therefore, as *CDC20* levels increase, cells may gain migration capabilities. In conjunction, we did not detect some of the more deleterious CDC20 promoter variants (i.e. those leading to lower reporter expression) in distant metastases. Additionally, we found enrichment of metastatic gene signatures in the CDC20-high expressing

samples across all 5 melanoma cohorts (Supplemental Table 1) and observed loss of migratory capabilities in A3 and A10, the CDC20 promoter indel cell lines with lowered CDC20 levels.

Like TERT, CDC20 performs a variety of canonical and non-canonical functions, many of which can be implicated in cancer formation[131]. Most crucial is its role in the cell cycle, where it interacts with the anaphase-promoting complex to degrade cyclin B and signal the end of metaphase and the start of anaphase. Complete knock-down of *CDC20* is lethal but several studies have shown that partial knock-down or missense mutations that impair the ability of CDC20 to bind to other interacting proteins leads to aneuploidy, an important hallmark of cancer[147–149,157]. Although aneuploidy did not increase in the CDC20 promoter indel A375 cell lines, CDC20-low samples in 3/5 cohorts analyzed had enrichment of genes associated with increased aneuploidy (Supplemental Table 1). Additionally, we only observed a slight reduction in growth rates in one CDC20 promoter indel strain, A10, which has similar levels of CDC20 as the A3 strain despite a larger deletion in the CDC20 promoter, suggesting that at least in our model, a 2-fold reduction of CDC20 does not significantly alter cell growth rates.

3.6 Conclusion

The ever-expanding genomic data available for melanoma has been crucial in advancing our understanding of melanoma biology, but most of the largest datasets with publicly available clinical outcomes data (i.e. TCGA) overrepresent metastatic lesions, and even a subset of metastatic lesion types (i.e. lymph node metastases in TCGA). Thus, while *CDC20* has been implicated as a cancer-driving gene with higher levels often associated with melanoma metastases and poorer survival, we posit that specific levels of *CDC20* expression may be crucial to supporting or allowing passage of melanocytes through malignant transformation (*CDC20* low) to locally invasive cancer and then on to metastatic disease (*CDC20* high, Figure 2G). As in the case of

MITF, a rheostat model of *CDC20* may exist, whereby higher levels of *CDC20* drives metastasis and lower levels support a phenotype likely beneficial in earlier tumors[43].

3.7 Declarations

Ethics Approval and Consent to Participate

Not applicable.

Consent for Publication

Not applicable

Availability of Data and Materials

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE206639. Scrips for data visualization and the bioinformatic pipeline can be made available upon request.

Competing Interests

The authors have no competing interests.

3.8 Funding

This research was supported by the Melanoma Research Alliance Young Investigator Award #566840. P.G. was supported by NSF DGE-1745038.

3.9 Author Contributions

P.M.G and C.K. conceived of the project. P.M.G. analyzed all data and performed experiments. P.M.G. and C.K. wrote the manuscript

3.10 Acknowledgments

We thank the Genome Engineering and Stem Cell Center (GESC) at Washington University in St Louis for their Cell Line Engineering Services, the Cytogenetics and Molecular

Pathology Laboratory at Washington University School of Medicine, and the Genome Technology Access Center at the McDonnell Genome Institute at Washington University School of Medicine for help with genomic analysis. The Genome Technology Access Center is partially supported by NCI Cancer Center Support Grant #P30 CA91842 to the Siteman Cancer Center from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH), and NIH Roadmap for Medical Research. We thank Megan Glaeser and Rebecca Cunningham for critical reading of the manuscript and Catie Newsom-Stewart for help with figure conceptualization and design.

3.11 Figures

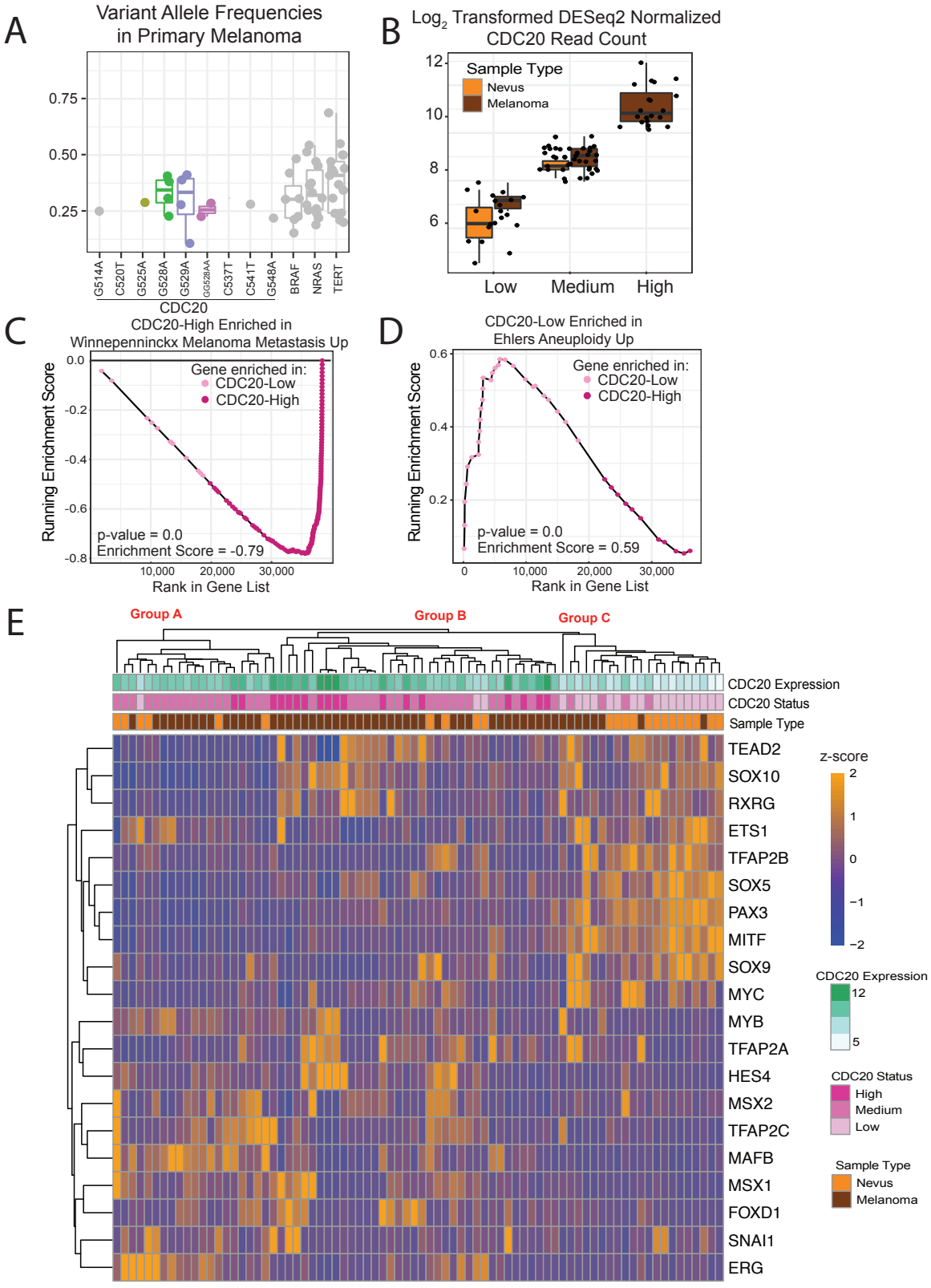

A
Variant Allele Frequencies in Primary Melanoma
0.75
0.50
0.25
G514A
C520T
G525A
G528A
G529A
GG528AA
C537T
C541T
G548A
BRAF
NRAS
TERT
CDC20
B
Log2 Transformed DESeq2 Normalized CDC20 Read Count
Sample Type
Nevus
Melanoma
12
10
8
6
Low
Medium
High
C
CDC20-High Enriched in Winnepenninckx Melanoma Metastasis Up
Gene enriched in:
CDC20-Low
CDC20-High
Running Enrichment Score
0.0
-0.2
-0.4
-0.6
-0.8
p-value = 0.0
Enrichment Score = -0.79
10,000
20,000
30,000
Rank in Gene List
D
CDC20-Low Enriched in Ehlers Aneuploidy Up
Gene enriched in:
CDC20-Low
CDC20-High
Running Enrichment Score
0.6
0.4
0.2
p-value = 0.0
Enrichment Score = 0.59
0
10,000
20,000
30,000
Rank in Gene List
E
Group A
Group B
Group C
CDC20 Expression
CDC20 Status
Sample Type
TEAD2
SOX10
RXRG
ETS1
TFAP2B
SOX5
PAX3
MITF
SOX9
MYC
MYB
TFAP2A
HES4
MSX2
TFAP2C
MAFB
MSX1
FOXD1
SNAI1
ERG
z-score
2
1
0
−1
−2
CDC20 Expression
12
5
CDC20 Status
High
Medium
Low
Sample Type
Nevus
Melanoma

Figure 1. Changes in CDC20 expression levels correlate with specific gene expression programs.

(A) Variant allele frequencies of the CDC20 promoter variants (each labelled), $BRAF^{V600E}$ (BRAF), $NRAS^{Q61K}$ and $NRAS^{Q61R}$ (NRAS), TERT G228A and G250A (TERT) that are detected in primary melanomas. C520T and C541T are not detected in primary melanomas and have no data points. There are no statistically significant differences between G528A, G529A and other variants. (B) *CDC20* expression in CDC20-Low, CDC20-Medium, and CDC20-High nevus or melanoma samples from Kunz et al. Each data point represents the $\log_2$ DESeq2-normalized read count of *CDC20*. No nevi are classified as CDC20-high. (C and D) Gene set enrichment analysis of results for the Winnepenninckx Melanoma Metastasis Up gene set (C) and the Ehlers Aneuploidy Up gene set (D). Each point represents a gene, ranked by expression, at the current running-sum statistic. Negative scores indicate enrichment in CDC20-high samples (as seen in C). Positive scores indicate enrichment in CDC20-low samples (as seen D). (E) Heatmap depicting z-score normalized expression patterns of 20 key neural crest transcription factors. Samples and genes are hierarchically clustered with orange and blue indicating relatively higher and lower gene expression, respectively, across samples. All columns are annotated by CDC20 expression (top row of boxes, $\log_2$ DESeq2-normalized read count), CDC20 expression group (second row, for low, medium, or high), and sample type (nevus in orange or melanoma in dark brown).

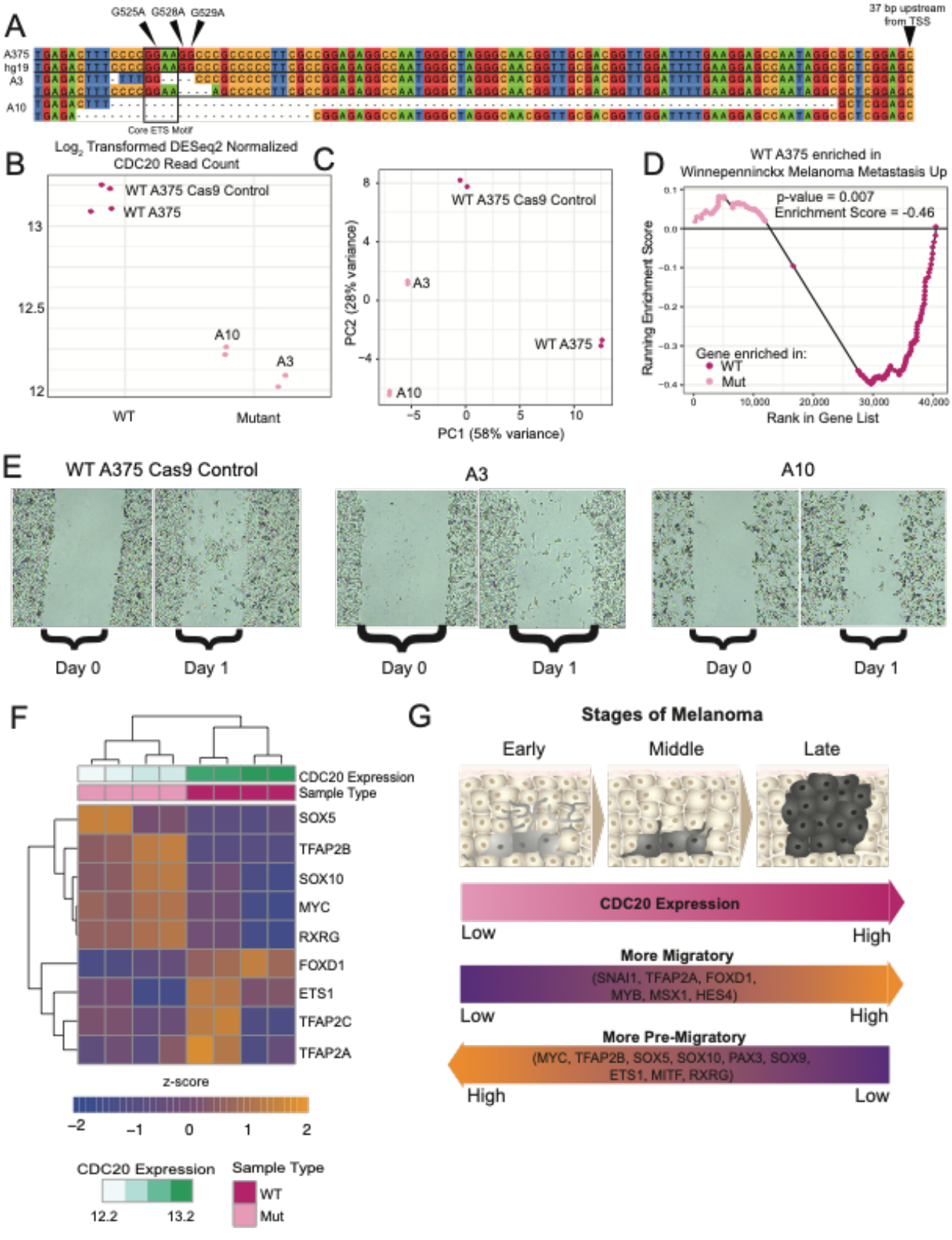

A
G525A G528A G529A
37 bp upstream from TSS
A375
hg19
A3
A10
Core ETS Motif
B
Log2 Transformed DESeq2 Normalized CDC20 Read Count
WT A375 Cas9 Control
WT A375
A10
A3
WT
Mutant
C
WT A375 Cas9 Control
A3
WT A375
A10
PC2 (28% variance)
PC1 (58% variance)
D
WT A375 enriched in Winnepenninckx Melanoma Metastasis Up
p-value = 0.007
Enrichment Score = -0.46
Running Enrichment Score
Gene enriched in:
WT
Mut
Rank in Gene List
E
WT A375 Cas9 Control
A3
A10
Day 0
Day 1
F
CDC20 Expression
Sample Type
SOX5
TFAP2B
SOX10
MYC
RXRG
FOXD1
ETS1
TFAP2C
TFAP2A
z-score
CDC20 Expression
12.2
13.2
Sample Type
WT
Mut
G
Stages of Melanoma
Early
Middle
Late
CDC20 Expression
Low
High
More Migratory
(SNAI1, TFAP2A, FOXD1, MYB, MSX1, HES4)
Low
High
More Pre-Migratory
(MYC, TFAP2B, SOX5, SOX10, PAX3, SOX9, ETS1, MITF, RXRG)
High
Low

Figure 2. Engineered indels at the recurrently mutated CDC20 promoter locus leads to decreased CDC20 expression and changes in melanoma behavior.

(A) Sequence alignment of the CDC20 promoter between hg19, WT A375, A3, and A10. Arrows denoting positions of G525A, G528A, and G529A. The ETS core motif is boxed. The last nucleotide of the sequence is 37 bp upstream of the TSS of *CDC20*. Nucleotides are color-coded and dashes indicate deletions. (B) Plot depicting $\log_2$ transformed DESeq2-normalized read counts of *CDC20* in WT A375 and CDC20 promoter indel strains, A3 and A10, with decreased CDC20 expression. Each point represents CDC20 expression in one sample. (C) Principal component analysis of read counts normalized by regularized log transformation using the top 500 most variable genes. The horizontal axis, PC1, explains 58% of the variance associated across all samples and separates out WT from CDC20 promoter indel cell lines. The vertical axis, PC2, explains 28% of the variance and separated A3 from A10. (D) Gene set enrichment analysis of results for the Winnepenninckx Melanoma Metastasis Up gene set. Each point represents a gene, ranked by expression, at the current running-sum statistic. Negative scores indicate enrichment in WT A375 samples as compared to the engineered indel lines A3 and A10. (E) CDC20 indel lines A3 and A10 show decreased migration capabilities compared to WT A375 cell lines. Images of scratch migration assay from day 0 (immediately after scratch) and day 1 (24 hours post-scratch). (F) Heatmap depicting z-score normalized expression patterns of 9 differentially expressed neural crest transcription factors. Samples and genes are hierarchically clustered with orange and blue indicating relatively higher or lower expression, respectively, of genes across samples. All columns are annotated by CDC20 expression ($\log_2$ DESeq2-normalized read count), and sample type (WT or mutant). *SOX5*, *TFAP2B*, *SOX10*, *MYC*, and *RXRG* are expressed at relatively higher levels in the CDC20 promoter indel-containing A3 and A10 cell lines. *FOXD1*, *ETS1*, *TFAP2C*,

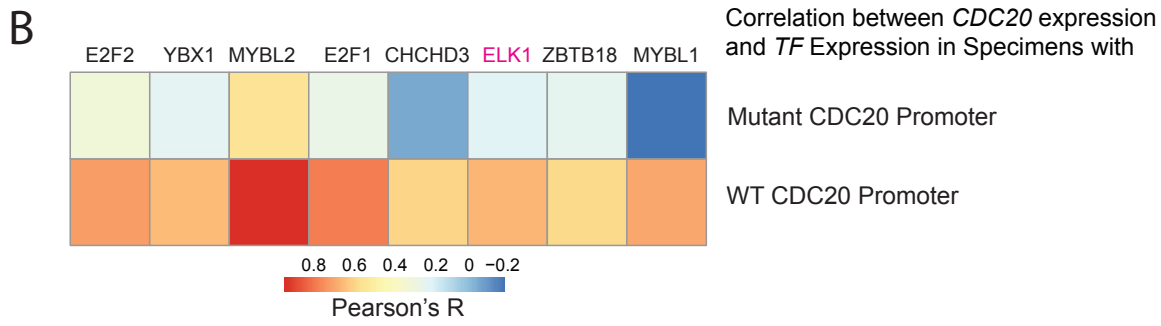
and *TFAP2A* are higher in WT (CDC20-high) A375 cell lines. (G) Model of *CDC20* expression and neural crest transcription factor signature over melanoma onset and progression. *CDC20* levels increase as melanoma progresses. Neural crest transcription factors that correlate with *CDC20* expression are more prevalent in migrating neural crest cells, whereas those that are relatively higher in CDC20-low settings are more prevalent in the melanocytic/pre-migratory neural crest states.

3.12 Supplemental Figures

A

SNP ID	motifBreakR Results
G514A	GAIN: BATF, IRF4
C520T	GAIN: STAT, STAT1, STAT3, STAT5A, STAT5B LOSS: ELK1, PTF1A
G525A	GAIN: BCL6B, STAT, STAT1 , STAT5A, STAT5B LOSS: ELF1, ELF4, ELK1, ELK3, ELK4, ETS1, ETV1, ETV4, GABPA, NFE2L2, NR2C2, REL, TAF1
G528A	GAIN: FLI1, NFATC1, RELA LOSS: ELF1, ELF4, ELK3, ETS1, GABPA, HF1H3B, NFE2L2, SP4
C537T	GAIN: RXRA LOSS: HF1H3B, KLF13, SP2, SP4, ZBTB7A, ZNF219, ZNF281
C541T	LOSS: HF1H3B
G548A	GAIN: ELK1, FOXA1, REL, RELA, SP100

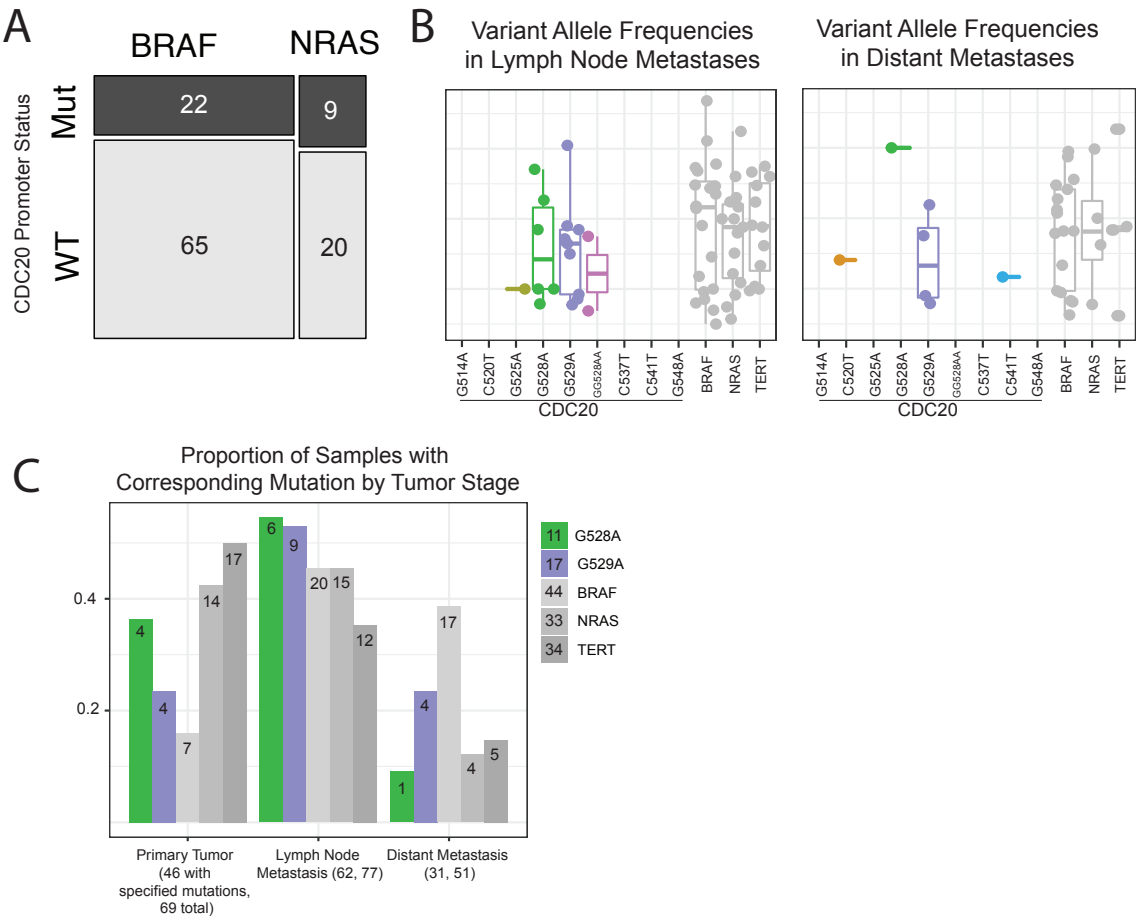
B



Supplemental Figure 1. Motif impact of recurrent CDC20 promoter variants.

(A) Table summarizing motifBreakR results. All transcription factors listed have strong and significant motif altering predictions. ETS transcription factors are colored in pink. **(B)** Heatmap

of Pearson correlation values between TF (column) and samples with WT CDC20 promoters (top row) or mutant CDC20 promoters (bottom row.)

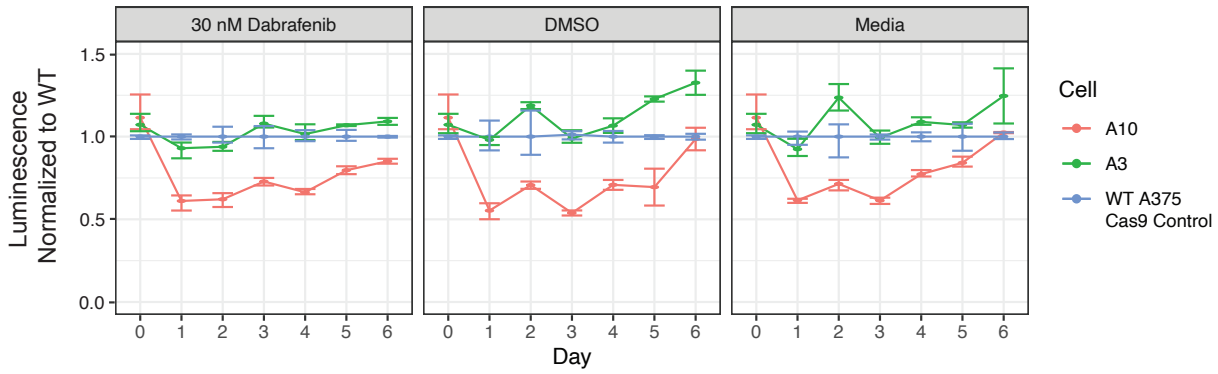


Supplemental Figure 2. Co-occurrence and clonality of recurrent CDC20 promoter variants.

A) Mosaic plot of the number of donors with either BRAF and/or NRAS mutations and WT and/or mutant CDC20 promoter. **(B)** Variant allele frequencies of the CDC20 promoter variants (each labelled), $BRAF^{V600E}$ (BRAF), $NRAS^{Q61K}$ and $NRAS^{Q61R}$ (NRAS), TERT G228A and G250A (TERT) that are detected in lymph node metastases or distant metastases. **(C)** Each bar represents the number of specimens with the corresponding variant within the corresponding tumor subtype divided by the total number of specimens with the corresponding variant across all subtypes.

Absolute counts for each subtype are labeled within the bar. Total counts across all subtypes are labeled within the legend. The total number of samples at each stage that contain one of the specified mutations are in parentheses below the row label; the number that follows is equal to the total number of samples in that tumor stage.

A Supplemental Figure 4



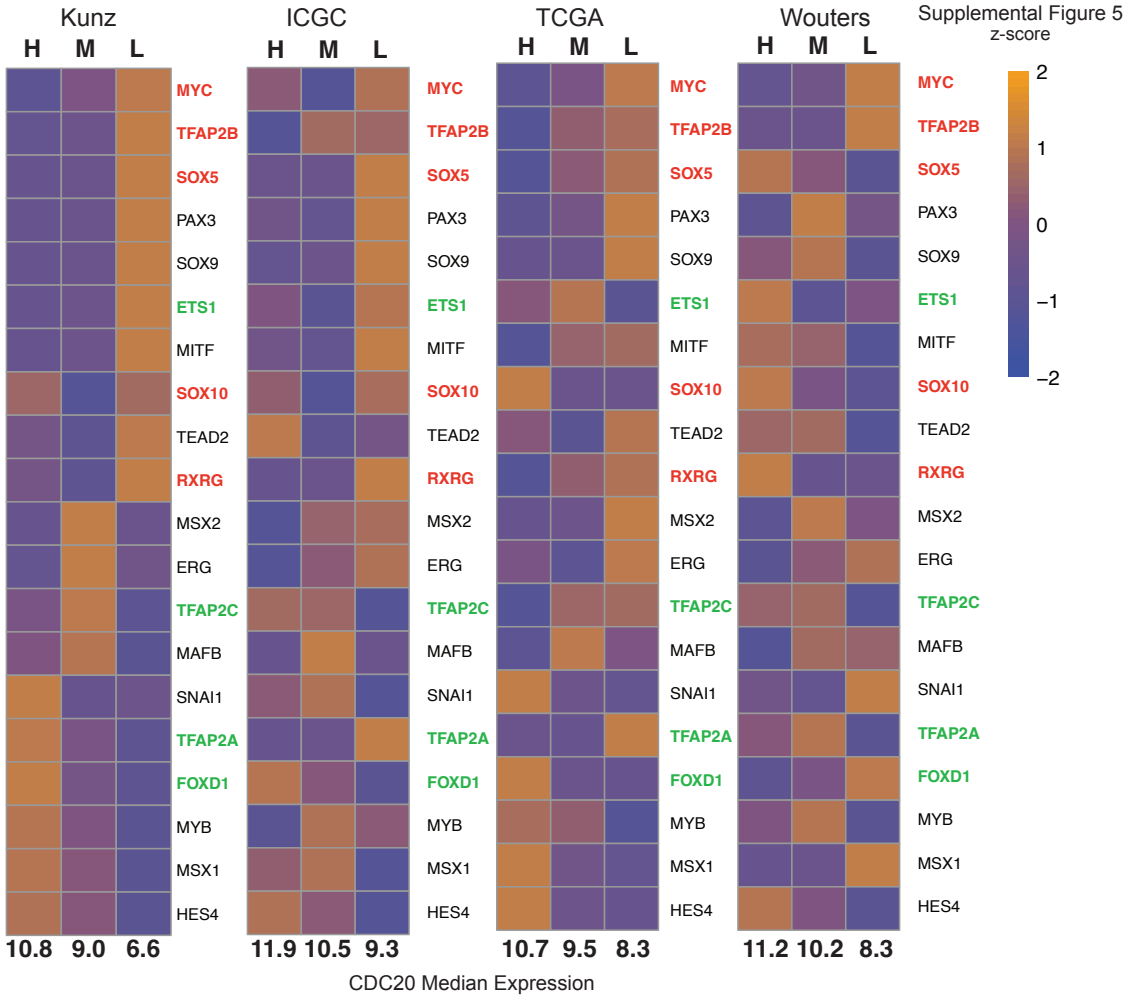
B

SNP ID	Karyotype/Nomenclature
A375	61~63<3N>,XX,-X/Y,+1,del(1)(p22p21),der(1)t(1;6)(p13,p12),-2,add(2)(q22),-4,-6,der(6)t(6;22)(q10;p10),der(8)(t(4;8)p10;q10),-10,del(11)(p15),del(13)(q33), add(15)(p11.2),del(17)(p11.2),-18,add(19)(p13),-21, -22x2,+2~4 mar[cp20]
A3	60~63<3N>,XX,-X/Y,+1,del(1)(p22p21),der(1)t(1;6)(p13,p12),-2,add(2)(q22),-4,-6,der(6)t(6;22)(q10;p10),der(8)t(4;8)(p10;q10),-10,del(11)(p15),-13, del(13)(q33), add(15)(p11.2),-18,add(19)(p13),-21,-22x2,+1~2 mar[cp20]
A10	61~63<3N>,XX,-X/Y,+1,del(1)(p22p21),der(1)t(1;6)(p13,p12),-2,add(2)(q22),-4,-6,der(6)t(1;6)(p34;p12),der(8)(t(4;8)p10;q10),-10,del(11)(p15),add(13)(p11.2), del(13)(q33), -18,add(19)(p13),-21,-22x2,+2~3 mar[cp20]

Supplemental Figure 3. Viability and aneuploidy of WT and CDC20 promoter indel cell lines.

(A) Proliferation rates are slightly lower in A10 but unchanged in A3. Plot shows luminescence values obtained from CellTiterGlo normalized to the average WT luminescence for each day and for each specific condition. Each point represents the average of three replicates. Confidence intervals are calculated using a nonparametric bootstrap method. **(B)** Table showing the karyotype/nomenclature of WT A375, A3, and A10. Differences across cell lines are color-coded.

A

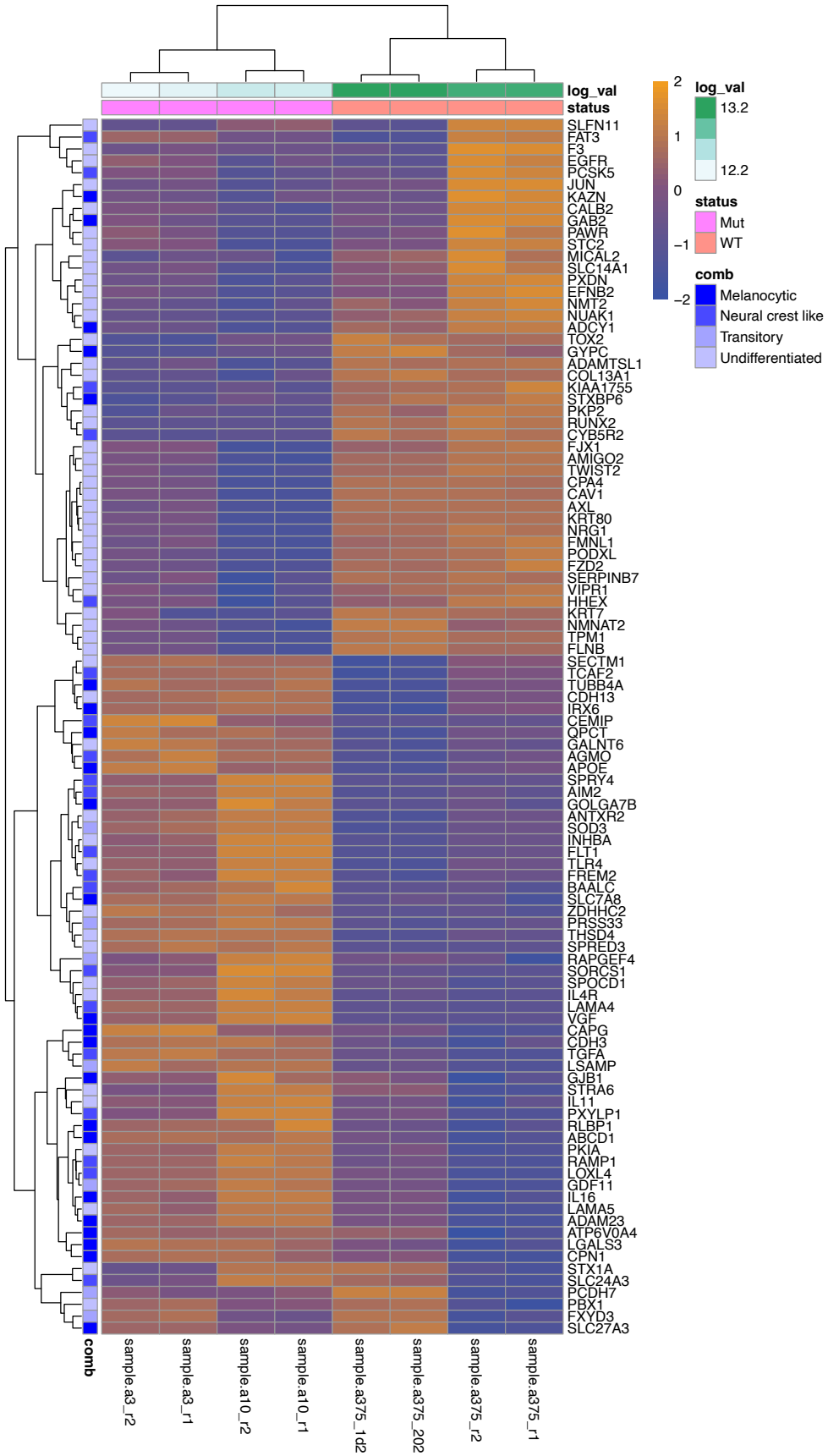


B

Gene	Agree	Disagree
FOXD1	Kunz, ICGC, TCGA	Wouters
TFAP2C	Kunz*, ICGC, Wouters	TCGA
TFAP2A	Kunz, Wouters*	ICGC, TCGA
ETS1	TCGA*, Wouters*	Kunz, ICGC
TFAP2B	Kunz, ICGC*, TCGA*, Wouters	
SOX5	Kunz, ICGC, TCGA*	Wouters
RXRG	Kunx, ICGC, TCGA*	Wouters
MYC	Kunz, ICGC*, TCGA, Wouters	
SOX10	Kunz*, ICGC*	TCGA, Wouters

Supplemental Figure 4. Neural crest transcription factor signature across 5 RNA-sequencing melanoma cohorts.

(A) Heatmap depicting relative expression of 20 neural crest transcription factors using the average of all samples classified as CDC20-high, medium, or low. The median log_2-normalized *CDC20* count is listed below each column of every heatmap. Orange indicates higher expression relative to other samples for the same gene. Genes in green are upregulated in WT A375 compared to CDC20 promoter indel cell lines. Genes in red are upregulated in CDC20 promoter indel cell lines. (B) Table summarizing whether a cohort has a relative gene level that matches or does not match the gene level seen in WT or CDC20 promoter indel cell lines. For a gene to agree, it needs to have relatively higher expression in the WT lines (green genes) or relatively higher expression in the CDC20 promoter indel lines (red genes). Cohorts that have an asterisk neither completely agree or disagree (e.g. relatively higher in CDC20-medium samples or relatively high in CDC20-low and CDC20-high, see SOX10 in TCGA).


log_val
status
SLFN11
FAT3
F3
EGFR
PCSK5
JUN
KAZN
CALB2
GAB2
PAWR
STC2
MICAL2
SLC14A1
PXDN
EFNB2
NMT2
NUAK1
ADCY1
TOX2
GYPC
ADAMTSL1
COL13A1
KIAA1755
STXBP6
PKP2
RUNX2
CYB5R2
FJX1
AMIGO2
TWIST2
CPA4
CAV1
AXL
KRT80
NRG1
FMNL1
PODXL
FZD2
SERPINB7
VIPR1
HHEX
KRT7
NMNAT2
TPM1
FLNB
SECTM1
TCAF2
TUBB4A
CDH13
IRX6
CEMIP
QPCT
GALNT6
AGMO
APOE
SPRY4
AIM2
GOLGA7B
ANTXR2
SOD3
INHBA
FLT1
TLR4
FREM2
BAALC
SLC7A8
ZDHHC2
PRSS33
THSD4
SPRED3
RAPGEF4
SORCS1
SPOCD1
IL4R
LAMA4
VGF
CAPG
CDH3
TGFA
LSAMP
GJB1
STRA6
IL11
PXYLP1
RLBP1
ABCD1
PKIA
RAMP1
LOXL4
GDF11
IL16
LAMA5
ADAM23
ATP6V0A4
LGALS3
CPN1
STX1A
SLC24A3
PCDH7
PBX1
FXYD3
SLC27A3
2
1
0
−1
−2
log_val
13.2
12.2
status
Mut
WT
comb
Melanocytic
Neural crest like
Transitory
Undifferentiated
comb
sample.a3_r2
sample.a3_r1
sample.a10_r2
sample.a10_r1
sample.a375_1d2
sample.a375_202
sample.a375_r2
sample.a375_r1

Supplemental Figure 5. CDC20 promoter indels recapitulate major subpopulations identified in scRNA-seq of melanoma.

Heatmap of genes expressed in melanocytic, neural crest-like, transitory, and undifferentiated subpopulations within a melanoma tumor that underwent single-cell RNA-sequencing. Samples and genes are clustered by Euclidean distance. WT and Mutant CDC20 Promoter lines cluster accordingly. Row annotations are colored based on the presence of the gene within a specific subtype. Gene annotations provided as a supplemental file from Tsoi et al. (2018).

3.13 Supplemental Tables

Supplemental Table 1. GSEA Results of CDC20-High and CDC20-Low Populations across 4 RNA-sequencing cohorts.

Melanoma Gene Set							
Kunz (CDC20-Low)							
NAME	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX
EHLERS_ANEUPLOIDY_UP	40	0.58	1.89	0.002	0.019	0.032	5823
WINNEPENNINCKX_MELANOMA_METASTASIS_DN	43	0.48	1.2	0.231	0.315	0.72	4657
Kunz (CDC20_high)							
NAME	SIZE	ES	NES	NOM p-val	FDR q-val	FWER	RANK AT MAX

						p-val	
WINNEPENNINCKX_MELANOMA_METASTASIS_UP	161	-0.79	-2.11	0	0.001	0.004	2869
JAEGER_METASTASIS_UP	44	-0.75	-2.04	0.002	0.001	0.008	3055
ICGC (CDC20-Low)							
NAME	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX
JAEGER_METASTASIS_DN	257	0.41860878	2.1427767	0	0	0	9890
WINNEPENNINCKX_MELANOMA_METASTASIS_DN	44	0.303456	1.137695	0.25985664	0.2978612	0.974	1321
ICGC (CDC20-High)							
NAME	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX
WINNEPENNINCKX_MELANOMA_METASTASIS_UP	161	-0.8027842	-3.8972762	0	0	0	2497
JAEGER_METASTASIS_UP	46	-0.6530079	-2.4758463	0	0	0	2219

EHLERS_ANEUPLOIDY_UP	40	-0.274367	-1.0184273	0.42729306	0.41150478	0.997	6014
TCGA (CDC20-Low)							
NAME	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX
EHLERS_ANEUPLOIDY_UP	40	0.5672705	1.9665768	0	0.0022359	0.002	5303
WINNEPENNINCKX_MELANOMA_METASTASIS_DN	45	0.33962148	1.2148776	0.16666667	0.2687455	0.831	6073
JAEGER_METASTASIS_DN	257	0.13631484	0.64607364	1	0.9795464	1	4919
TCGA (CDC20-High)							
NAME	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX
WINNEPENNINCKX_MELANOMA_METASTASIS_UP	162	-0.7958587	-3.3748016	0	0	0	1846
JAEGER_METASTASIS_UP	44	-0.6971762	-2.4024394	0	0	0	640

A375 Mutant (CDC20-Low)							
NAME	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX
JAEGER_METASTASIS_DN	258	0.38902858	1.321719	0.00851064	0.23987243	0.525	4387
WINNEPENNINCKX_MELANOMA_METASTASIS_DN	45	0.40882364	1.1028054	0.27348644	0.45067537	0.972	4162
EHLERS_ANEUPLOIDY_UP	40	0.38967028	1.0292898	0.3908524	0.5393845	0.997	9169
A375 WT (CDC20-high)							
NAME	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX
WINNEPENNINCKX_MELANOMA_METASTASIS_UP	162	-0.4002273	-1.2928551	0.03522505	0.37796304	0.649	11023
Wouters (CDC20-Low)							
NAME	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX
WINNEPENNINCKX_MELANOMA_METASTASIS_DN	44	0.61	1.84	0	0.001	0.001	3728

EHLERS_ANEUPLOIDY_UP	40	0.58	1.73	0	0.002	0.005	6259
JAEGER_METASTASIS_DN	257	0.43	1.69	0	0.004	0.017	6632
Wouters (CDC20-High)							
NAME	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX
WINNEPENNINCKX_MELANOMA_METASTASIS_UP	162	-0.74	-2.86	0	0	0	2358
JAEGER_METASTASIS_UP	45	-0.53	-1.66	0.002	0.008	0.049	804

Chapter 4: Discussion

4.1 Prediction and validation of non-coding variants

Chapter 2 presented a bioinformatic pipeline to predict putatively functional non-coding variants. This pipeline leveraged publicly accessible ChIP-seq and ATAC-seq datasets to partition the genome into test regions, i.e. putative melanoma regulatory regions, and null regions. We used our test and null regions to generate an empirical null distribution with which to calculate statistical significance. Non-coding variants were merged, i.e. hotspots, using a window of 25 bp. The Donor Score for each hotspot was calculated as a function of the number of donors with a mutation in that hotspot and the total number of unique sites. High-scoring hotspots were therefore those that were recurrently mutated across multiple samples with most mutations occurring at a few specific locations. This was meant to emulate the TERT promoter hotspot where most of the variants are at one of two positions[80,81]. We took the average FunSeq2 score of each variant within a hotspot. The FunSeq2 algorithm generates a score for each non-coding variant based on several features including conservation, impact on motif, and whether the variant-associated gene is a known cancer driver[107]. The fold-change between the FunSeq2 score in null and test hotspots was higher than the fold-change between the Donor Score (Chapter 2, Figure 1) leading to a larger separation between the Hotspot Scores corresponding to the null and test regions, suggesting that the addition of the FunSeq2 score further reduced the hotspot scores of potential false positives. Overall, we identified 140 hotspots with FDR-adjusted p-values = 0 harboring 2,631 mutations, which represents 0.01% of the total number of mutations used as input. The TERT promoter hotspot, the RPS27 promoter hotspot, the DPH3 promoter hotspot, and the CDC20 promoter hotspot, all within the top 13 highest-scoring hotspots, have been previously validated by luciferase assay[80,83,86,97].

Due to the high mutational load particularly at ETS motifs in promoters, we note that our bioinformatic pipeline almost exclusively identified hotspots within promoter regions at ETS

motifs. On the one hand, this selects for variants that are at actively bound transcription factor binding sites which is more likely to yield changes in gene expression and reporter activity. Moreover, ETS factors are known oncogenes in melanoma[160]. On the other hand, the alteration of DNA by bound ETS transcription factors is sensitive to UV irradiation and therefore leads to higher mutation rates at these specific locations[124,125]. Our bioinformatic pipeline is built is sensitive to this mutational signature and therefore not only increases the number of false positives but likely reduces the number of true positives. This brings to light an important observation – a variant at a transcription factor binding site may alter activity of a gene without significantly altering cancer biology. Therefore, assessment of variants needs to extend beyond reporter assays and into appropriate cancer models.

Systematic assessment of bioinformatic pipelines remain limited due to the small number of studies that used high-throughput reporter assays. While we understand that our MPRA requires technical improvements before thorough assessment, we did see a correlation between Hotspot Score and fold-change as calculated by the MPRA, suggesting that our method is able to prioritize to some degree functional variants (Chapter 2, Figure 5). However, this remains a preliminary assessment that requires follow-up.

Our bioinformatic pipeline can be improved in a variety of ways. For example, because our pipeline does not consider the combined regulatory space of a gene, a promoter hotspot will almost always score higher than a hotspot within an enhancer due to the aforementioned mutational processes. Genes are often regulated by dozens of enhancers extending hundreds of kilobases away. Combining variant scores from the entire regulatory space of a single gene may increase the likelihood of identifying functional variants in more distal regulatory elements. Understanding the cell-type specific regulatory process of a gene will also improve our search space. Indeed, various

studies have identified cis-acting regulatory variants via functional genomics datasets like ChIP-seq or ATAC-seq where the variant is observed to decrease or increase peak intensity akin to allele-specific expression[165]. A more systematic and accurate method to improve bioinformatic pipelines will be to perform high-throughput validation on as many variants as technically possible and learn through convolutional neural networks the sequence motifs that when altered lead to change in expression in a cell-type specific manner.

Validation of variants is another area for future improvement. The luciferase assays used here are a reliable and fast way to test a small number of variants (~dozens). For larger scale validations (~100s-1000s), exogenous MPRAs are currently the only pragmatic solution, although it may be important to assess different lengths as one study reported an increase in luciferase of a CDC20 promoter variant in HEK 293FT cells when using a ~700 bp sequence while we noticed a decrease in activity when using a 150 bp sequence[97]. Additionally, integration of the MPRA into the genome to test the endogenous effect of the variant may also lead to more accurate results, although there are many challenges to this method as well, including position-specific effects[166]. The most accurate method would involve introducing the exact variant into the genome and looking for corresponding changes in gene expression. At the time of writing, this would not be practical as methods are still inefficient and non-coding variants often fall in repetitive sequences.

4.2 The CDC20 promoter hotspot

The CDC20 promoter hotspot was the 8th highest scoring hotspot with an average FunSeq2 score of 1.77 (1.06 for the TERT promoter hotspot). The hotspot is located ~153 bp from the transcriptional start site of *CDC20* and is mutated in 39 donors. We validated several variants of the CDC20 promoter hotspot using a variety of methods: (1) luciferase assay using a 300 bp sequence upstream of a minimal promoter driving luciferase, (2) MPRA using a 150 bp sequence

upstream of EF1α, a strong promoter, driving GFP, and (3) luciferase assay using a 150 bp sequence in a vector with no promoter driving luciferase. In all three assays, variants either significantly decreased reporter activity or exacted no change. This is in direct contrast to a recent publication where CDC20 promoter variants led to an increase in luciferase activity[97]. However, there were major differences in other publications: (1) the other study used a 1,117 bp sequence was used that to the TSS and (2) only two cell lines were assayed, one of which was a melanoma cell line. A study on long and short enhancers in melanoma demonstrated that 190 bp regions were sufficient to capture the enhancer activity[167]. It may therefore be likely that in a 1,000 bp region, multiple enhancer units are captured and occlude the actual effect of the variant. Moreover, when aligning the sequence to the hg19 genome via BLAT, the sequence extends into the first exon.

Between our validations, we also noted discrepancies between the variants that reduced reporter activity. In the luciferase assay performed with a minimal promoter downstream of a 300 bp region harboring the WT or variant allele and the MPRA, G528A and GG528AA both led to a reduction in reporter activity while G529A exhibited no change (Chapter 2, Supplemental Figure 3). In the luciferase assay using a 150 bp sequence and no promoter, G528A strongly reduced reporter activity in all but one cell line while G529A and GG528AA varied in their response. This could represent transcriptional noise, the effect of additional sequences in the 300 bp luciferase assay, and cell-type specific effects. However, the evidence for downregulation, as opposed to upregulation, of the reporter via the CDC20 promoter hotspot is robust in our hands when tested across multiple pertinent (i.e. melanoma) cell lines.

One of the most intriguing observations across all three reporter assays is the drastic change in reporter activity when driven either by the G528A, G529A, or GG528AA variants. The G528A and the G529A variants are adjacent while the GG528AA variant is a multi-nucleotide variant.

Despite their proximity, the effects on reporter activity are dramatically different with G528A almost always leading to a strong decrease in promoter activity across multiple cell lines, while G529A leading to no change or a small decrease in reporter activity. Most unexpectedly is that GG528AA displayed effects on reporter activity most similar to G529A as opposed to G528A. A previous study demonstrated that mutating the nucleotides flanking the core ETS motif (i.e. mutations like G528A and G529A) could lead to either total loss of expression or even stronger and ectopic expression (Farley et al., 2016). Therefore, the G528A mutation, since it is closer to the core ETS motif, may directly impair binding whereas G529A may lead to altered binding of, for example, co-factors specific to the cell state. This may also depend on what other transcription factor binding sites are near the mutated ETS motifs.

It was also interesting to note that the G525A variant was not as common as the G528A, G529A, or GG528AA variants despite it being at the location in the core ETS motif most often mutated by variants in statistically significant hotspots (Chapter 2, Figure 1E). This may provide support for a dosage-dependent bias for CDC20 in melanoma, as G525A consistently led to large reductions in reporter activity, as opposed to G528A and G529A which are more common. Overall, the CDC20 promoter hotspot appears to alter a transcription factor binding site such that the variants at or near the core ETS motif decrease reporter activity, likely through disruption of an ETS transcription factor, while variants further away have cell-type specific effects.

Although it is predicted that an ETS transcription factor binds to the CDC20 promoter hotspot, motif analysis identified several different ETS family members along with other transcription factors. When defining our search space, we used ChIP-seq for ETV1 performed on two melanoma cell lines, neither of which had signal at the CDC20 promoter. Of the ChIP-seq datasets used, BRG1, CTCF, DDX21, TFAP2A, and SMAD1/5/8 were found to differentially bind

to the CDC20 promoter (compared to IgG control). It is likely that an ETS factor binds and recruits multiple co-factors, dependent on cell type, to regulate expression of CDC20. To provide support to the notion that an ETS family member binds to the promoter, we correlated expression of all TFs to CDC20 in samples with and without a mutation in the CDC20 promoter using the ICGC-MELA RNA-sequencing cohort and identified 9 transcription factors that had high correlations in WT samples but low correlations in mutant samples, suggesting impaired transcriptional regulation of that TF towards CDC20. These included ELK1, E2F1 and E2F2, and MYBL1 and MYBL2. E2F transcription factors regulate genes involved in cell proliferation. Interestingly, APC/C-CDC20 targets E2F1 for degradation in prometaphase[152]. However, CDC20 expression and translation is necessary in anaphase suggesting that E2F1 may inhibit CDC20 transcription. MYBL2 has also been associated with regulating not only cell cycle progression but embryonic stem cell fate[168]. Overall, the combined results of the luciferase assay and analysis of co-expressed transcription factors suggests a cell-type specific regulatory complex that likely involves members of the ETS transcription family, cell cycle regulators, and others. Advances in DNA binding assays such as multiCUT&TAG and proximity labelling can aid in understanding the full extent of the epigenetic landscape at the CDC20 promoter hotspot and how changes in the promoter lead to changes in DNA binding.

4.3 A CDC20-low population may be characterized by a more neural crest-like state

Our observation of a reduction in reporter activity was unexpected because high expression of CDC20 is prognostic of aggressive tumors in multiple cancer types[131]. Therefore, in order to generate a hypothesis as to why reduced transcriptional activity of CDC20 could contribute to melanoma onset, progression, metastasis, recurrence, and/or resistance, we looked at the variant

allele frequencies (VAFs) of the CDC20 promoter variants across melanoma types. We observed VAFs in primary melanomas that are similar to those of the TERT promoter mutations and BRAF mutations which are known to be early events[154]. We also observed a dramatic reduction in the presence of not only the VAFs but the variants themselves in later stages of melanoma. We therefore hypothesized that the promoter variants are beneficial to tumor formation in early stages of melanoma but not at later stages of melanoma, specifically the formation of metastases at distant locations (e.g. lungs, liver, or brain). Therefore, we stratified an RNA-seq dataset of primary melanomas and nevi by CDC20 expression and performed gene set enrichment analysis on a variety of melanoma and neural crest gene sets. As expected, we saw genes that co-expressed with high levels of CDC20 were enriched in migration and metastasis gene sets. While we did not observe gene set enrichment for any of the neural crest signatures, we did notice that genes co-expressing with low levels of CDC20 were enriched in an aneuploidy gene signature. Aberrant expression and germline mutations of CDC20 have been associated with aneuploidy[147,148,157].

We then looked at how CDC20 co-expressed with neural crest transcription factors, as re-emergence of the neural crest identity occurs in the very earliest cells of melanoma[47]. Interestingly, we noted that several neural crest transcription factors, including SOX10, correlated with low levels of CDC20, suggesting that reduced levels of CDC20 could lead to re-emergence of neural crest activity. To confirm this hypothesis, we genome engineered a small indel encompassing the mutated region at the CDC20 promoter in two cell lines. This allowed us to confirm the phenotypes that were associated with the gene expression signature of samples with low and high levels of *CDC20,* specifically the migratory and neural crest signature phenotypes. Moreover, we used the gene sets characteristics of the four cluster subtypes identified in Tsoi et al. (2018) to show that CDC20-knockdown converted A375 from a more “undifferentiated/neural crest-like” subtype to

a “neural crest-like/melanocytic” subtype[75]. In conclusion, by investigating the clonal history of the CDC20 promoter variants and generating a CDC20 promoter indel strain in a human melanoma cell line, we observed that low levels of CDC20, driven by mutations in the promoter, could contribute to the de-differentiation of melanocytes into a more neural crest-like state through up-regulation of SOX10 whereas high levels of CDC20 are more important for metastasis and migration.

4.4 Future Directions

4.4.1 How is CDC20 transcriptionally regulated and what effect do the CDC20 promoter variants have on its regulation?

The levels of CDC20 protein and mRNA oscillate throughout the cell cycle[169]. As expected, CDC20 protein levels are highest during mitosis and are promptly diminished after mitotic exit. CDC20 primarily targets itself for degradation through the APC/C-CDC20 complex. APC/C-CDC20 appears to be involved in a feedback mechanism where degradation of CDC20 also up-regulates *CDC20* transcription[133]. How APC/C-CDC20 leads to increased transcriptional up-regulation remains unknown. Additionally, the post-transcriptional regulation of CDC20 appears to play important roles as in plant stem cells, CDC20 mRNA remains in the nucleus until prometaphase where it is exported and translated[170]. This activity is mediated by the 5’UTR. As CDC20 plays important functions in and out of mitosis, transcriptional and post-transcriptional regulation is likely complex and mediated by multiple co-factors dependent on cell-type and context[171].

Based on our analysis of putative melanoma regulatory regions, we observed published instances of binding of BRG1, DDX21, TFAP2A, CTCF, and SMAD1/5/8 at the CDC20 promoter. SMAD1/5/8 represses MITF and SOX9 expression thereby preventing differentiation

and cell death[172]. When GDF6 is expressed, SMAD1/5/8 is detected at the CDC20 promoter. Knockdown of GDF6 causes loss of promoter binding activity[172]. TFAP2A binds to the CDC20 promoter in human primary melanocytes and is found at several other melanocyte-specific genes[173]. CTCF also binds to the CDC20 promoter, and mutations at CTCF binding sites are prevalent due to deficiencies in nucleotide excision repair due to active CTCF binding. Such changes in CTCF binding can lead to changes in 3D genome architecture but have not been shown to contribute to cancer[174]. Altogether, this suggests complex and dynamic promoter kinetics at the CDC20 promoter.

In order to better characterize the transcriptional regulation of CDC20, it will be imperative to perform assays such as multiCUT&TAG that can track binding of multiple DNA binding proteins at a time in multiple cell types to understand the transcription factors conserved throughout the cell cycle and those that could potentially be cell-type specific such as TFAP2A. Editing the promoter to include the common G528A and G529A single-nucleotide variants and the GG528AA multi-nucleotide variant as well as others, ideally in a genetically engineered tumor model such as mice, would elucidate the changes in promoter kinetics in mutant strains. Additionally, it will be important to understand the cyclical changes in the complement of transcription factors bidingn to the CDC20 promoter throughout the cell cycle and how CDC20 is post-transcriptionally regulated in mammalian systems.

4.4.2 Can we engineer CDC20 promoter variants in earlier models of melanoma?

The CDC20 promoter variants appear to contribute to melanomagenesis at early stages of melanoma. By generating CDC20 promoter mutations in primary melanocytes that harbor mutations in *BRAF* or *NRAS* and/or tumor suppressors like *TP43, PTEN,* or *CDKN2A*, we can

more accurately assess the stage at which low levels of CDC20 are important for up-regulating *SOX10*. Recently, researchers at the Broad Institute sequentially introduced mutations in human primary melanocytes[78]. Each mutation conferred growth advantages such that after several months, the allele would become the dominant clone in the primary melanocyte culture. First, CDKN2A was deleted, followed by introduction of $BRAF^{V600E}$. Interestingly, primary melanocytes were not immortalized until addition of the TERT promoter mutation. Subsequent mutations in TP53, PTEN, and/or APC led to cells that could form tumors *in vivo*. Engineering the CDC20 promoter variant in each of these engineered melanocytes would confirm that low levels of CDC20, particularly in early stages of melanoma, leads to an increase in expression of certain neural crest transcription factors, including *SOX10*. Additionally, we could better understand why it seems important that primary melanocytes first have $BRAF^{V600E}$ and/or the TERT promoter mutation. One possibility is that cells need to be cycling in a MAPK-dependent manner in order for a dosage-dependent effect to occur. Additionally, performing these experiments in a diploid model could lead to observations of aneuploidy, unlike the CDC20 promoter indel lines generated in this work. Because the A375 melanoma cell line is a hypotetraploid, we may not have detected additional chromosomal aberrations because there may be an upper limit of the amount of genomic instability a particular cell line can tolerate.

4.4.3 What does APC/C-CDC20 target?

One of the major questions remaining is how changes in expression of CDC20 leads to phenotype switching from "undifferentiated" to "neural crest-like" and why transcription factors such as SOX10 are up-regulated. During metaphase, CDC20 is phosphorylated and binds to the mitotic checkpoint complex until all sister chromatids have attached properly to the kinetochore[175]. This prevents binding of CDC20 to the Anaphase Promoting Complex/Cyclosome until the spindle

assembly checkpoint is satisfied, signaling the end of metaphase. APC/C-CDC20 ubiquitylates securin and Cyclin B1, causing the sister chromatids to separate from the kinetochore. Due to its essential role in the cell cycle, physical interactions between CDC20 and multiple cell cycle regulators have been observed and catalogued (BioGRID, https://thebiogrid.org). CDC20 has also been shown to interact with transcription factors, such as MYC and HOXD1, and chromatin remodelers like CTCF and HDAC1/2/6, suggesting that it can potentially mediate changes in gene expression through degradation of these proteins. Indeed, recruitment of APC/C-CDC20 to cell-type specific genes during mitosis leads to ubiquitylation of histones, allowing for rapid degradation and transcription during re-entry into interphase, establishing cell identity after transcription has been paused during mitosis[138]. Therefore, reduced levels of APC/C-CDC20 could lead to altered rates of histone ubiquitylation leading to changes in cell identity.

Characterizing interactors of APC/C-CDC20 through proximity-labelling techniques such as TurboID[176] or through ChIP-seq of ubiquitin tagged histones at K11 and K48 would elucidate the mechanisms by which reduced APC/C-CDC20 levels leads to up-regulation of SOX10. As MITF expression levels did not change in CDC20 promoter indel lines, perhaps MITF is a not a direct or indirect target of APC/C-CDC20. Interestingly, SOX5 has been to shown to repress SOX10 activity at melanocyte-specific loci; up-regulation of SOX5, which was seen in our CDC20 promoter indel lines, could therefore in principle contribute to a dedifferentiation from the melanocyte state[177].

4.4.4 How do changes in cell cycle length contribute to lineage identity in melanoma?

Another possible association between changes in CDC20 levels and changes in the gene expression program is through altering the length of the cell cycle. siRNA-mediated knockdown

of CDC20 increases the number of cells at the G2/M stage in hESCs, suggesting that decreasing levels of CDC20 could delay mitosis through a stoichiometric imbalance[138]. One study found that specifically in G1, H3K4me3 levels increase at certain developmental genes due to phosphorylation of a histone methyl-transferase by CDK1[178]. Up-regulation of these developmental genes allowed for pluripotency exit and differentiation. Decreasing or increasing this "window of opportunity" could alter differentiation rates. Through decreasing CDC20 and delaying G2/M, fewer cells may be at the G1 stage and paracrine signaling that would otherwise lead to differentiation could be missed by cells undergoing G2/M. Therefore, future work should go into quantifying the number of cells at each cell cycle stage and identifying differential changes in H3K27me3, a repressive marker, and H3K4me3, a marker of active transcription. Because H3K4me3 levels increase in developmental genes at G1, I hypothesize that in cells with lower levels of CDC20, transcription of developmental genes or cell-type specific genes would decrease due to a shorter G1 and/or longer G2/M stage. This would therefore delay pluripotency exit and differentiation. Indeed, genetic and chemical prolongation of the G2 stage in hESCs led to the up-regulation of pluripotency maintenance factors and inhibited pluripotent state dissolution[179].

How would altering the length of the cell cycle contribute to melanoma initiation? Initially, through oncogene activation and tumor suppressor loss, a population of rapidly dividing cells contributes to initial melanoma growth. This population is characterized by high levels of *MITF*[57]. Within this population, either through the CDC20 promoter mutations or a parallel epigenetic mechanism, cells that cycle slower may instead transcribe genes that would cause de-differentiation, such as *SOX10*. These cells would be more neural crest-like. Lineage analysis of single clones in a melanoma mouse model identified a population of stem-like cells that either duplicated or differentiated into progenitor-like cells[79]. The stem-like cells divided more frequently

than the progenitor-like cells, and the progenitor cells were more neural-crest like. Interestingly, neither of these populations appeared to contribute to metastasis; rather, a separate mesenchymal-like subpopulation were found to initiate metastases. Altogether, this supports the notion that low levels of CDC20 alter cell cycle kinetics such that a cell establishes a neural crest-like identity separate from the subpopulation responsible for migration and likely also separate from the stem-like subpopulation.

One lingering question remains – why CDC20? There are many important cell cycle regulators. Future work will have to go into characterizing the promoters of the other G2/M regulators – do they also have ETS motifs? Does reduction of other G2/M regulators also lead to an increase in SOX10 expression and a reduced capability to metastasize?

4.4.5 Why does CDC20 appear to be important for metastasis? Do CDC20 promoter indel lines inhibit melanoma in vivo?

Due to the promoter variants leading to a decrease in expression, the focus of Chapter 3 was on discovering why down-regulation of CDC20 would be beneficial in melanoma. However, it is clear from the migration assays and bioinformatic analysis that high levels of CDC20 are also important for later stages in melanoma, specifically metastasis. As our CDC20 promoter indel lines did not affect viability but did affect migration, it is unlikely that altered cell proliferation is the mechanism by which CDC20 contributes to metastasis. Instead, CDC20 could affect migration through potentially two pathways: (1) as in the case of low CDC20, by generating a change in lineage identity due to altered cycling dynamics or (2) by altered maintenance of the primary cilia.

As discussed in the previous section, down-regulation of CDC20 may delay the G2/M phase of the cell cycle. Therefore, it is plausible that up-regulation of CDC20 could also alter the length of particular cell cycle stages so that rather than inducing a more neural crest-like state, it

induces a more undifferentiated cell state. Genes that were down-regulated in the CDC20 promoter indel lines and up-regulated in the WT melanoma lines were more likely to be associated with the undifferentiated state. Therefore, while low levels of CDC20 seem to be important for establishing the neural crest-like population of cells, high levels of CDC20 may be present in either the recently-discovered stem-like subpopulation or the rapidly proliferating melanocytic subpopulation[79].

CDC20 has also been shown to regulate the length of the primary cilia and the disassembly of primary cilia[140]. Many receptors of major signaling pathways including Hedgehog, Wnt, and Notch signaling are located on the primary cilia[180]. Therefore, dysregulation of the length of the primary cilia or premature or delayed disassembly could lead to aberrant signaling of these pathways, all of which have been implicated in melanoma[181–183]. Moreover, a key step in melanoma metastasis is the deregulation of the primary cilia mediated by EZH2 which was downregulated in our CDC20 promoter indel cell lines, suggesting an association between EZH2 and CDC20[184].

Further research will need to address the mechanism by which CDC20 contributes to melanoma. Ongoing work will determine whether our CDC20 promoter indel lines fail to metastasize compared to their WT counterparts and analyses of these tumors may elucidate the missing features not discussed in this thesis.

4.4.6 Do other genes act in a dosage/time-dependent manner?

One of the major contributions of this thesis is providing further support that genes can contribute to cancer onset and progression in a dosage-dependent manner. This pattern has been observed with MITF where high levels appear to be important for and a marker of early melanoma growth while low levels of MITF are a marker of metastasis through MITF-dependent epigenetic regulation of Dia1[73,185]. MITF is the master melanocyte lineage factor and expression is essential

for melanocyte differentiation[186]. MITF is amplified in 6-10% of melanomas[3,8,187] and high expression is worse for overall survival. Like CDC20, MITF expression leads to different phenotypes based on expression level.

Another potential candidate that appears to act in a dosage- or temporal-dependent manner is *SOX9*. SOX9 acts antagonistically to SOX10 and is down-regulated in melanoma compared to melanocytes[128]. However, SOX9 also appears to be up-regulated in mesenchymal subpopulations in melanoma which are believed to be the population of cells that eventually will metastasize[77,79]. Moreover, SOX9 overexpression appears to be critical for metastasis in glioma and lung cancer[188,189].

4.5 Conclusions

In conclusion, this thesis identified and validated non-coding variants in human melanoma. We specifically focused on variants in the promoter of *CDC20* which reduced reporter activity and endogenous gene expression in a human melanoma cell line. We observed changes in migration capabilities and transcriptional state. Specifically, reducing the expression of *CDC20* led to an increase in certain neural crest transcription factors, most notably *SOX10*, and the down-regulation of genes associated with metastasis like *AXL*. Akin to the MITF rheostat model where high levels of MITF indicate rapid proliferation and low levels indicate metastasis, CDC20 also acts as a rheostat where low levels possibly contribute to neural crest-like state important for establishing early melanomas and high levels contribute to metastasis. Future work will go into further characterizing the dosage-dependent effect of *CDC20* at various stages in melanoma and will likely include identifying the targets of the APC/C-CDC20 complex and the changes in the epigenetic landscape when the length of the cell cycle is altered. Moreover, the knowledge that proteins can have a dosage-dependent effect in cancer onset and progression will be an important

consideration in drug development, as many reviews have posited that CDC20 is a potential therapeutic target.

References

1. Curti, B. D. & Faries, M. B. Recent Advances in the Treatment of Melanoma. *New Engl J Med* 384, 2229–2240 (2021).

2. Alicea, G. M. & Rebecca, V. W. Emerging strategies to treat rare and intractable subtypes of melanoma. *Pigm Cell Melanoma R* 34, 44–58 (2021).

3. Hayward, N. K. *et al.* Whole-genome landscapes of major melanoma subtypes. *Nature* 545, 175–180 (2017).

4. Johansson, P. A. *et al.* Whole genome landscapes of uveal melanoma show an ultraviolet radiation signature in iris tumours. *Nat Commun* 11, 2408 (2020).

5. Braicu, C. *et al.* A Comprehensive Review on MAPK: A Promising Therapeutic Target in Cancer. *Cancers* 11, 1618 (2019).

6. Wee, P. & Wang, Z. Epidermal Growth Factor Receptor Cell Proliferation Signaling Pathways. *Cancers* 9, 52 (2017).

7. Sharma, A. *et al.* Mutant V599EB-Raf Regulates Growth and Vascular Development of Malignant Melanoma Tumors. *Cancer Res* 65, 2412–2421 (2005).

8. Network, T. C. G. A. *et al.* Genomic Classification of Cutaneous Melanoma. *Cell* 161, 1681–96 (2015).

9. Krauthammer, M. *et al.* Exome sequencing identifies recurrent mutations in NF1 and RASopathy genes in sun-exposed melanomas. *Nat Genet* 47, 996–1002 (2015).

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

11. Beadling, C. *et al.* KIT Gene Mutations and Copy Number in Melanoma Subtypes. *Clin Cancer Res* 14, 6821–6828 (2008).

12. Nikolaev, S. I. *et al.* Exome sequencing identifies recurrent somatic MAP2K1 and MAP2K2 mutations in melanoma. *Nat Genet* 44, 133–139 (2012).

13. Davies, M. A. *et al.* Integrated Molecular and Clinical Analysis of AKT Activation in Metastatic Melanoma. *Clin Cancer Res* 15, 7538–7546 (2009).

14. Dai, D. L., Martinka, M. & Li, G. Prognostic Significance of Activated Akt Expression in Melanoma: A Clinicopathologic Study of 292 Cases. *J Clin Oncol* 23, 1473–1482 (2005).

15. Mirmohammadsadegh, A. *et al.* Epigenetic Silencing of the PTEN Gene in Melanoma. *Cancer Res* 66, 6546–6552 (2006).

16. Vredeveld, L. C. W. *et al.* Abrogation of BRAFV600E-induced senescence by PI3K pathway activation contributes to melanomagenesis. *Gene Dev* 26, 1055–1069 (2012).

17. Dankort, D. *et al.* BrafV600E cooperates with Pten loss to induce metastatic melanoma. *Nat Genet* 41, 544–552 (2009).

18. Berger, M. F. *et al.* Melanoma genome sequencing reveals frequent PREX2 mutations. *Nature* 485, 502–506 (2012).

19. Tsao, H., Yang, G., Goel, V., Wu, H. & Haluska, F. G. Genetic Interaction Between NRAS and BRAF Mutations and PTEN/MMAC1 Inactivation in Melanoma. *J Invest Dermatol* 122, 337–341 (2004).

20. Karbowniczek, M., Spittle, C. S., Morrison, T., Wu, H. & Henske, E. P. mTOR Is Activated in the Majority of Malignant Melanomas. *J Invest Dermatol* 128, 980–987 (2008).

21. Tsao, H., Bevona, C., Goggins, W. & Quinn, T. The Transformation Rate of Moles (Melanocytic Nevi) Into Cutaneous Melanoma: A Population-Based Estimate. *Arch Dermatol* 139, 282–288 (2003).

22. Melamed, R. D. *et al.* Genomic Characterization of Dysplastic Nevi Unveils Implications for Diagnosis of Melanoma. *J Invest Dermatol* 137, 905–909 (2017).

23. Michaloglou, C. *et al.* BRAFE600-associated senescence-like cell cycle arrest of human naevi. *Nature* 436, 720–724 (2005).

24. Mallette, F. A., Gaumont-Leclerc, M.-F. & Ferbeyre, G. The DNA damage signaling pathway is a critical mediator of oncogene-induced senescence. *Gene Dev* 21, 43–48 (2007).

25. Foulkes, W. D., Flanders, T. Y., Pollock, P. M. & Hayward, N. K. The CDKN2A (p16) Gene and Human Cancer. *Mol Med* 3, 5–20 (1997).

26. Jones, R. *et al.* A CDKN2A Mutation in Familial Melanoma that Abrogates Binding of p16INK4a to CDK4 but not CDK6. *Cancer Res* 67, 9134–9141 (2007).

27. Haferkamp, S. *et al.* Oncogene-Induced Senescence Does Not Require the p16INK4a or p14ARF Melanoma Tumor Suppressors. *J Invest Dermatol* 129, 1983–1991 (2009).

28. Dhomen, N. *et al.* Oncogenic Braf Induces Melanocyte Senescence and Melanoma in Mice. *Cancer Cell* 15, 294–303 (2009).

29. Tran, S. L. *et al.* Absence of Distinguishing Senescence Traits in Human Melanocytic Nevi. *J Invest Dermatol* 132, 2226–2234 (2012).

30. Katlinskaya, Y. V. *et al.* Suppression of Type I Interferon Signaling Overcomes Oncogene-Induced Senescence and Mediates Melanoma Development and Progression. *Cell Reports* 15, 171–180 (2016).

31. Martik, M. L. & Bronner, M. E. Regulatory Logic Underlying Diversification of the Neural Crest. *Trends Genet* 33, 715–727 (2017).

32. Roellig, D., Tan-Cabugao, J., Esaian, S. & Bronner, M. E. Dynamic transcriptional signature and cell fate analysis reveals plasticity of individual neural plate border cells. *Elife* 6, e21620 (2017).

33. Cano, A. *et al.* The transcription factor Snail controls epithelial–mesenchymal transitions by repressing E-cadherin expression. *Nat Cell Biol* 2, 76–83 (2000).

34. 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).

35. Greenhill, E. R., Rocco, A., Vibert, L., Nikaido, M. & Kelsh, R. N. An Iterative Genetic and Dynamical Modelling Approach Identifies Novel Features of the Gene Regulatory Network Underlying Melanocyte Development. *Plos Genet* 7, e1002265 (2011).

36. Baxter, L. L. & Pavan, W. J. Pmel17 expression is Mitf-dependent and reveals cranial melanoblast migration during murine development. *Gene Expr Patterns* 3, 703–707 (2003).

37. Thomas, A. J. & Erickson, C. A. FOXD3 regulates the lineage switch between neural crest-derived glial cells and pigment cells by repressing MITF through a non-canonical mechanism. *Development* 136, 1849–1858 (2009).

38. Adameyko, I. *et al.* Sox2 and Mitf cross-regulatory interactions consolidate progenitor and melanocyte lineages in the cranial neural crest. *Development* 139, 397–410 (2011).

39. Yoshida, H., Kunisada, T., Kusakabe, M., Nishikawa, S. & Nishikawa, S. I. Distinct stages of melanocyte differentiation revealed by anlaysis of nonuniform pigmentation patterns. *Development* 122, 1207–1214 (1996).

40. Cichorek, M., Wachulska, M., Stasiewicz, A. & Tymińska, A. Skin melanocytes: biology and development. *Adv Dermatology Allergology Postępy Dermatologii Alergologii* 30, 30–41 (2013).

41. Nishimura, E. K. Melanocyte stem cells: a melanocyte reservoir in hair follicles for hair and skin pigmentation. *Pigm Cell Melanoma R* 24, 401–410 (2011).

42. Gilchrest, B. A. & Eller, M. S. DNA Photodamage Stimulates Melanogenesis and Other Photoprotective Responses. *J Invest Derm Symp P* 4, 35–40 (1999).

43. Mort, R. L., Jackson, I. J. & Patton, E. E. The melanocyte lineage in development and disease. *Development* 142, 620–632 (2015).

44. Hu, Z. & Sauka-Spengler, T. Cellular plasticity in the neural crest and cancer. *Curr Opin Genet Dev* 75, 101928 (2022).

45. Kim, Y. J. *et al.* Generation of Multipotent Induced Neural Crest by Direct Reprogramming of Human Postnatal Fibroblasts with a Single Transcription Factor. *Cell Stem Cell* 15, 497–506 (2014).

46. Laurette, P. *et al.* Transcription factor MITF and remodeller BRG1 define chromatin organisation at regulatory elements in melanoma cells. *Elife* 4, e06857 (2015).

47. Kaufman, C. K. *et al.* A zebrafish melanoma model reveals emergence of neural crest identity during melanoma initiation. *Science* 351, aad2197 (2016).

48. Shakhova, O. *et al.* Sox10 promotes the formation and maintenance of giant congenital naevi and melanoma. *Nat Cell Biol* 14, 882–890 (2012).

49. Jin, S.-G., Xiong, W., Wu, X., Yang, L. & Pfeifer, G. P. The DNA methylation landscape of human melanoma. *Genomics* 106, 322–330 (2015).

50. Cronin, J. C. *et al.* Frequent mutations in the MITF pathway in melanoma. *Pigm Cell Melanoma R* 22, 435–444 (2009).

51. Guan, J., Gupta, R. & Filipp, F. V. Cancer systems biology of TCGA SKCM: Efficient detection of genomic drivers in melanoma. *Sci Rep-uk* 5, 7857 (2015).

52. Garraway, L. A. *et al.* Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature* 436, 117–122 (2005).

53. Yokoyama, S. *et al.* A novel recurrent mutation in MITF predisposes to familial and sporadic melanoma. *Nature* 480, 99–103 (2011).

54. Eliades, P. *et al.* High MITF Expression Is Associated with Super-Enhancers and Suppressed by CDK7 Inhibition in Melanoma. *J Invest Dermatol* 138, 1582–1590 (2018).

55. Lister, J. A. *et al.* A Conditional Zebrafish MITF Mutation Reveals MITF Levels Are Critical for Melanoma Promotion vs. Regression In Vivo. *J Invest Dermatol* 134, 133–140 (2014).

56. Giuliano, S. *et al.* Microphthalmia-Associated Transcription Factor Controls the DNA Damage Response and a Lineage-Specific Senescence Program in Melanomas. *Cancer Res* 70, 3813–3822 (2010).

57. Carreira, S. *et al.* Mitf cooperates with Rb1 and activates p21Cip1 expression to regulate cell cycle progression. *Nature* 433, 764–769 (2005).

58. Loercher, A. E., Tank, E. M. H., Delston, R. B. & Harbour, J. W. MITF links differentiation with cell cycle arrest in melanocytes by transcriptional activation of INK4A. *J Cell Biology* 168, 35–40 (2005).

59. Jonsson, A., Tuominen, R., Grafström, E., Hansson, J. & Egyhazi, S. High Frequency of p16INK4A Promoter Methylation in NRAS-Mutated Cutaneous Melanoma. *J Invest Dermatol* 130, 2809–2817 (2010).

60. Lauss, M. *et al.* Genome-Wide DNA Methylation Analysis in Melanoma Reveals the Importance of CpG Methylation in MITF Regulation. *J Invest Dermatol* 135, 1820–1828 (2015).

61. Swoboda, A. *et al.* STAT3 promotes melanoma metastasis by CEBP-induced repression of the MITF pathway. *Oncogene* 40, 1091–1105 (2021).

62. Verfaillie, A. *et al.* Decoding the regulatory landscape of melanoma reveals TEADS as regulators of the invasive cell state. *Nat Commun* 6, 6683 (2015).

63. Wellbrock, C. & Marais, R. Elevated expression of MITF counteracts B-RAF–stimulated melanocyte and melanoma cell proliferation. *J Cell Biology* 170, 703–708 (2005).

64. Wellbrock, C. *et al.* Oncogenic BRAF Regulates Melanoma Proliferation through the Lineage Specific Factor MITF. *Plos One* 3, e2734 (2008).

65. Smith, M. P. *et al.* A PAX3/BRN2 rheostat controls the dynamics of BRAF mediated MITF regulation in MITFhigh/AXLlow melanoma. *Pigm Cell Melanoma R* 32, 280–291 (2019).

66. Wu, M. *et al.* c-Kit triggers dual phosphorylations, which couple activation and degradation of the essential melanocyte factor Mi. *Gene Dev* 14, 301–312 (2000).

67. Hurk, K. van den *et al.* Genetics and epigenetics of cutaneous malignant melanoma: A concert out of tune. *Biochimica Et Biophysica Acta Bba - Rev Cancer* 1826, 89–102 (2012).

68. Fang, M., Hutchinson, L., Deng, A. & Green, M. R. Common BRAF(V600E)-directed pathway mediates widespread epigenetic silencing in colorectal cancer and melanoma. *Proc National Acad Sci* 113, 1250–1255 (2016).

69. Hou, P., Liu, D., Dong, J. & Xing, M. The BRAFV600E causes widespread alterations in gene methylation in the genome of melanoma cells. *Cell Cycle* 11, 286–295 (2012).

70. Bell, R. E. *et al.* Enhancer methylation dynamics contribute to cancer plasticity and patient mortality. *Genome Res* 26, 601–611 (2016).

71. Ceol, C. J. *et al.* The histone methyltransferase SETDB1 is recurrently amplified in melanoma and accelerates its onset. *Nature* 471, 513–517 (2011).

72. Scahill, C. M. *et al.* Loss of the chromatin modifier Kdm2aa causes BrafV600E-independent spontaneous melanoma in zebrafish. *Plos Genet* 13, e1006959 (2017).

73. Hoek, K. S. *et al.* Metastatic potential of melanomas defined by specific gene expression profiles with no BRAF signature. *Pigm Cell Res* 19, 290–302 (2006).

74. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* 352, 189–196 (2016).

75. Tsoi, J. *et al.* Multi-stage Differentiation Defines Melanoma Subtypes with Differential Vulnerability to Drug-Induced Iron-Dependent Oxidative Stress. *Cancer Cell* 33, 890-904.e5 (2018).

76. Rambow, F. *et al.* Toward Minimal Residual Disease-Directed Therapy in Melanoma. *Cell* 174, 843-855.e19 (2018).

77. Wouters, J. *et al.* Robust gene expression programs underlie recurrent cell states and phenotype switching in melanoma. *Nat Cell Biol* 22, 986–998 (2020).

78. Hodis, E. *et al.* Stepwise-edited, human melanoma models reveal mutations' effect on tumor and microenvironment. *Science* 376, eabi8175 (2022).

79. Karras, P. *et al.* A cellular hierarchy in melanoma uncouples growth and metastasis. *Nature* 1–9 (2022) doi:10.1038/s41586-022-05242-7.

80. Huang, F. W. *et al.* Highly Recurrent TERT Promoter Mutations in Human Melanoma. *Science* 339, 957–959 (2013).

81. Horn, S. *et al.* TERT Promoter Mutations in Familial and Sporadic Melanoma. *Science* 339, 959–961 (2013).

82. Bell, R. J. A. *et al.* The transcription factor GABP selectively binds and activates the mutant TERT promoter in cancer. *Science* 348, 1036–1039 (2015).

83. Floristán, A. *et al.* Functional analysis of RPS27 mutations and expression in melanoma. *Pigm Cell Melanoma R* 33, 466–479 (2020).

84. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat Genet* 46, 1160–1165 (2014).

85. Zhang, W. *et al.* A global transcriptional network connecting noncoding mutations to changes in tumor gene expression. *Nat Genet* 50, 613–620 (2018).

86. Denisova, E. *et al.* Frequent DPH3 promoter mutations in skin cancers. *Oncotarget* 6, 35922–35930 (2015).

87. Choi, J. *et al.* Massively parallel reporter assays of melanoma risk variants identify MX2 as a gene promoting melanoma. *Nat Commun* 11, 2718 (2020).

88. Mansour, M. R. *et al.* An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* 346, 1373–1377 (2014).

89. Hu, S. *et al.* Whole-genome noncoding sequence analysis in T-cell acute lymphoblastic leukemia identifies oncogene enhancer mutations. *Blood* 129, 3264–3268 (2017).

90. Puente, X. S. *et al.* Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* 526, 519–524 (2015).

91. Bal, E. *et al.* Super-enhancer hypermutation alters oncogene expression in B cell lymphoma. *Nature* 1–8 (2022) doi:10.1038/s41586-022-04906-8.

92. Yanchus, C. *et al.* A noncoding single-nucleotide polymorphism at 8q24 drives IDH1-mutant glioma formation. *Science* 378, 68–78 (2022).

93. Elliott, K. & Larsson, E. Non-coding driver mutations in human cancer. *Nat Rev Cancer* 21, 500–509 (2021).

94. Dratwa, M., Wysoczańska, B., Łacina, P., Kubik, T. & Bogunia-Kubik, K. TERT—Regulation and Roles in Cancer Formation. *Front Immunol* 11, 589929 (2020).

95. Chun-on, P. *et al.* TPP1 promoter mutations cooperate with TERT promoter mutations to lengthen telomeres in melanoma. *Science* 378, 664–668 (2022).

96. Zhang, T. *et al.* SDHD Promoter Mutations Ablate GABP Transcription Factor Binding in Melanoma. *Cancer Res* 77, 1649–1661 (2017).

97. He, Z. *et al.* Pan-cancer noncoding genomic analysis identifies functional CDC20 promoter mutation hotspots. *Iscience* 24, 102285 (2021).

98. Gates, L. A., Foulds, C. E. & O'Malley, B. W. Histone Marks in the 'Driver's Seat': Functional Roles in Steering the Transcription Cycle. *Trends Biochem Sci* 42, 977–989 (2017).

99. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biology* 109, 21.29.1-21.29.9 (2015).

100. Neavin, D. *et al.* Single cell eQTL analysis identifies cell type-specific genetic control of gene expression in fibroblasts and reprogrammed induced pluripotent stem cells. *Genome Biol* 22, 76 (2021).

101. Christensen, D. S. *et al.* Treatment Represents a Key Driver of Metastatic Cancer Evolution. *Cancer Res* 82, 2918–2927 (2022).

102. Li, Y., Shi, W. & Wasserman, W. W. Genome-wide prediction of cis-regulatory regions using supervised deep learning methods. *Bmc Bioinformatics* 19, 202 (2018).

103. Yu, C.-P. *et al.* Discovering unknown human and mouse transcription factor binding sites and their characteristics from ChIP-seq data. *Proc National Acad Sci* 118, e2026754118 (2021).

104. Robles-Espinoza, C. D., Mohammadi, P., Bonilla, X. & Gutierrez-Arcelus, M. Allele-specific expression: applications in cancer and technical considerations. *Curr Opin Genet Dev* 66, 10–19 (2021).

105. Zhang, T. *et al.* Cell-type-specific eQTL of primary melanocytes facilitates identification of melanoma susceptibility genes. *Genome Res* 28, 1621–1635 (2018).

106. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol* 17, 128 (2016).

107. Fu, Y. *et al.* FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol* 15, 480 (2014).

108. Smale, S. T. Luciferase Assay. *Cold Spring Harb Protoc* 2010, pdb.prot5421 (2010).

109. Inoue, F., Kreimer, A., Ashuach, T., Ahituv, N. & Yosef, N. Identification and Massively Parallel Characterization of Regulatory Elements Driving Neural Induction. *Cell Stem Cell* (2019) doi:10.1016/j.stem.2019.09.010.

110. Comandante-Lou, N., Baumann, D. G. & Fallahi-Sichani, M. AP-1 transcription factor network explains diverse patterns of cellular plasticity in melanoma cells. *Cell Reports* 40, 111147 (2022).

111. Khaliq, M., Manikkam, M., Martinez, E. D. & Fallahi-Sichani, M. Epigenetic modulation reveals differentiation state specificity of oncogene addiction. *Nat Commun* 12, 1536 (2021).

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

113. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep-uk* 9, 9354 (2019).

114. Yu, G., Wang, L.-G. & He, Q.-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* 31, 2382–2383 (2015).

115. Coetzee, S. G., Coetzee, G. A. & Hazelett, D. J. motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics* 31, 3847–3849 (2015).

116. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res* 42, 2976–2987 (2014).

117. Kunz, M. *et al.* RNA-seq analysis identifies different transcriptomic types and developmental trajectories of primary melanomas. *Oncogene* 37, 6136–6151 (2018).

118. Rheinbay, E. *et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* 578, 102–111 (2020).

119. Baggiolini, A. *et al.* Developmental chromatin programs determine oncogenic competence in melanoma. *Science* 373, (2021).

120. Friedman, R. Z. *et al.* Information content differentiates enhancers from silencers in mouse photoreceptors. *Elife* 10, e67403 (2021).

121. Godoy, P. M., Zarov, A. P. & Kaufman, C. K. Functional analysis of recurrent non-coding variants in human melanoma. *Biorxiv* 2022.06.30.498319 (2022) doi:10.1101/2022.06.30.498319.

122. Cunningham, R. L. *et al.* Functional in vivo characterization of sox10 enhancers in neural crest and melanoma development. *Commun Biology* 4, 695 (2021).

123. Rothhammer, T. *et al.* The Ets-1 transcription factor is involved in the development and invasion of malignant melanoma. *Cell Mol Life Sci Cmls* 61, 118–128 (2004).

124. Fredriksson, N. J. *et al.* Recurrent promoter mutations in melanoma are defined by an extended context-specific mutational signature. *Plos Genet* 13, e1006773 (2017).

125. Mao, P. *et al.* ETS transcription factors induce a unique UV damage signature that drives recurrent mutagenesis in melanoma. *Nat Commun* 9, 2626 (2018).

126. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* 15, 901–913 (2005).

127. Sivapragasam, S. *et al.* CTCF binding modulates UV damage formation to promote mutation hot spots in melanoma. *Embo J* e107795 (2021) doi:10.15252/embj.2021107795.

128. Shakhova, O. *et al.* Antagonistic Cross-Regulation between Sox9 and Sox10 Controls an Anti-tumorigenic Program in Melanoma. *Plos Genet* 11, e1004877 (2015).

129. Flores-Pérez, A. *et al.* Dual targeting of ANGPT1 and TGFBR2 genes by miR-204 controls angiogenesis in breast cancer. *Sci Rep-uk* 6, 34504 (2016).

130. Li, M., York, J. P. & Zhang, P. Loss of Cdc20 Causes a Securin-Dependent Metaphase Arrest in Two-Cell Mouse Embryos. *Mol Cell Biol* 27, 3481–3488 (2007).

131. Jeong, S. M., Bui, Q. T., Kwak, M., Lee, J. Y. & Lee, P. C.-W. Targeting Cdc20 for cancer therapy. *Biochimica Et Biophysica Acta Bba - Rev Cancer* 1877, 188824 (2022).

132. Sullivan, M. & Morgan, D. O. Finishing mitosis, one step at a time. *Nat Rev Mol Cell Bio* 8, 894–903 (2007).

133. Yu, H. Regulation of APC–Cdc20 by the spindle checkpoint. *Curr Opin Cell Biol* 14, 706–714 (2002).

134. Lara-Gonzalez, P. *et al.* The G2-to-M Transition Is Ensured by a Dual Mechanism that Protects Cyclin B from Degradation by Cdc20-Activated APC/C. *Dev Cell* 51, 313-325.e10 (2019).

135. Li, M., Fang, X., Wei, Z., York, J. P. & Zhang, P. Loss of spindle assembly checkpoint–mediated inhibition of Cdc20 promotes tumorigenesis in mice. *J Cell Biol* 185, 983–994 (2009).

136. Wavelet-Vermuse, C. *et al.* CDC20-Mediated hnRNPU Ubiquitination Regulates Chromatin Condensation and Anti-Cancer Drug Response. *Cancers* 14, 3732 (2022).

137. Quek, L. S., Grasset, N., Jasmen, J. B., Robinson, K. S. & Bellanger, S. Dual Role of the Anaphase Promoting Complex/Cyclosome in Regulating Stemness and Differentiation in Human Primary Keratinocytes. *J Invest Dermatol* 138, 1851–1861 (2018).

138. Oh, E. *et al.* Gene expression and cell identity controlled by anaphase-promoting complex. *Nature* 1–5 (2020) doi:10.1038/s41586-020-2034-1.

139. Xie, Y.-P. *et al.* CDC20 regulates cardiac hypertrophy via targeting LC3-dependent autophagy. *Theranostics* 8, 5995–6007 (2018).

140. Wang, W., Wu, T. & Kirschner, M. W. The master cell cycle regulator APC-Cdc20 regulates ciliary length and disassembly of the primary cilium. *Elife* 3, e03083 (2014).

141. Kim, A. H. *et al.* A Centrosomal Cdc20-APC Pathway Controls Dendrite Morphogenesis in Postmitotic Neurons. *Cell* 136, 322–336 (2009).

142. Yang, Y. *et al.* A Cdc20-APC Ubiquitin Signaling Pathway Regulates Presynaptic Differentiation. *Science* 326, 575–578 (2009).

143. Wu, F. *et al.* The Oncogenic Role of APC/C Activator Protein Cdc20 by an Integrated Pan-Cancer Analysis in Human Tumors. *Frontiers Oncol* 11, 721797 (2021).

144. Karra, H. *et al.* Cdc20 and securin overexpression predict short-term breast cancer survival. *Brit J Cancer* 110, 2905–2913 (2014).

145. Chang, D. Z. *et al.* Increased CDC20 expression is associated with pancreatic ductal adenocarcinoma differentiation and progression. *J Hematol Oncol* 5, 15 (2012).

146. Mao, D. D. *et al.* A CDC20-APC/SOX2 Signaling Axis Regulates Human Glioblastoma Stem-like Cells. *Cell Reports* 11, 1809–1821 (2015).

147. Chen, O. J. *et al.* Germline Missense Variants in CDC20 Result in Aberrant Mitotic Progression and Familial Cancer. *Cancer Res* 82, 3499–3515 (2022).

148. Fujita, H. *et al.* Premature aging syndrome showing random chromosome number instabilities with CDC20 mutation. *Aging Cell* 19, e13251 (2020).

149. Gu, Q. *et al.* CDC20 knockdown and acidic microenvironment collaboratively promote tumorigenesis through inhibiting autophagy and apoptosis. *Mol Ther - Oncolytics* 17, 94–106 (2020).

150. Kramer, E. T., Godoy, P. M. & Kaufman, C. K. Transcriptional profile and chromatin accessibility in zebrafish melanocytes and melanoma tumors. *G3 Genes Genomes Genetics* 12, jkab379 (2021).

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

152. Peart, M. J. *et al.* APC/CCdc20 targets E2F1 for degradation in prometaphase. *Cell Cycle* 9, 3956–3964 (2010).

153. Gopal, P. *et al.* Clonal selection confers distinct evolutionary trajectories in BRAF-driven cancers. *Nat Commun* 10, 5143 (2019).

154. Shain, A. H. *et al.* The Genetic Evolution of Melanoma from Precursor Lesions. *New Engl J Medicine* 373, 1926–1936 (2015).

155. Mukherjee, A., Bhattacharyya, J., Sagar, M. V. & Chaudhuri, A. Liposomally encapsulated CDC20 siRNA inhibits both solid melanoma tumor growth and spontaneous growth of intravenously injected melanoma cells on mouse lung. *Drug Deliv Transl Re* 3, 224–234 (2013).

156. Ehlers, J. P., Worley, L., Onken, M. D. & Harbour, J. W. Integrative Genomic Analysis of Aneuploidy in Uveal Melanoma. *Clin Cancer Res* 14, 115–122 (2008).

157. Malureanu, L. *et al.* Cdc20 hypomorphic mice fail to counteract de novo synthesis of cyclin B1 in mitosis. *J Cell Biol* 191, 313–329 (2010).

158. White, R. M. *et al.* DHODH modulates transcriptional elongation in the neural crest and melanoma. *Nature* 471, 518–522 (2011).

159. Tani-Matsuhana, S., Vieceli, F. M., Gandhi, S., Inoue, K. & Bronner, M. E. Transcriptome profiling of the cardiac neural crest reveals a critical role for MafB. *Dev Biol* 444, S209–S218 (2018).

160. Huang, L. *et al.* Targeting Pan-ETS Factors Inhibits Melanoma Progression. *Cancer Res* 81, 2071–2085 (2021).

161. Lin, X. *et al.* C-myc overexpression drives melanoma metastasis by promoting vasculogenic mimicry via c-myc/snail/Bax signaling. *J Mol Med* 95, 53–67 (2017).

162. Winnepenninckx, V. *et al.* Gene Expression Profiling of Primary Cutaneous Melanoma and Clinical Outcome. *Jnci J National Cancer Inst* 98, 472–482 (2006).

163. Brombin, A. *et al.* Tfap2b specifies an embryonic melanocyte stem cell that retains adult multifate potential. *Cell Reports* 38, 110234 (2022).

164. Wu, H. *et al.* Loss of neural crest-associated gene FOXD1 impairs melanoma invasion and migration via RAC1B downregulation. *Int J Cancer* 143, 2962–2972 (2018).

165. Liu, B. & Montgomery, S. B. Identifying causal variants and genes using functional genomics in specialized cell types and contexts. *Hum Genet* 139, 95–102 (2020).

166. McAfee, J. C. *et al.* Focus on your locus with a massively parallel reporter assay. *J Neurodev Disord* 14, 50 (2022).

167. Mauduit, D. *et al.* Analysis of long and short enhancers in melanoma cell states. doi:10.1101/2021.07.27.453936.

168. Zhan, M. *et al.* The B-MYB Transcriptional Network Guides Cell Cycle Progression and Fate Decisions to Sustain Self-Renewal and the Identity of Pluripotent Stem Cells. *Plos One* 7, e42350 (2012).

169. Wang, R., Burton, J. L. & Solomon, M. J. Transcriptional and post-transcriptional regulation of Cdc20 during the spindle assembly checkpoint in S. cerevisiae. *Cell Signal* 33, 41–48 (2017).

170. Yang, W., Wightman, R. & Meyerowitz, E. M. Cell Cycle Control by Nuclear Sequestration of CDC20 and CDH1 mRNA in Plant Stem Cells. *Mol Cell* 68, 1108-1119.e3 (2017).

171. Bruno, S. *et al.* CDC20 in and out of mitosis: a prognostic factor and therapeutic target in hematological malignancies. *J Exp Clin Cancer Res Cr* 41, 159 (2022).

172. Venkatesan, A. M. *et al.* Ligand-activated BMP signaling inhibits cell differentiation and death to promote melanoma. *J Clin Invest* 128, 294–308 (2017).

173. Seberg, H. E. *et al.* TFAP2 paralogs regulate melanocyte differentiation in parallel with MITF. *Plos Genet* 13, e1006636 (2017).

174. Poulos, R. C. *et al.* Functional Mutations Form at CTCF-Cohesin Binding Sites in Melanoma Due to Uneven Nucleotide Excision Repair across the Motif. *Cell Reports* 17, 2865–2872 (2016).

175. Jia, L., Li, B. & Yu, H. The Bub1–Plk1 kinase complex promotes spindle checkpoint signalling through Cdc20 phosphorylation. *Nat Commun* 7, 10818 (2016).

176. Cho, K. F. *et al.* Proximity labeling in mammalian cells with TurboID and split-TurboID. *Nat Protoc* 15, 3971–3999 (2020).

177. Kordaß, T. *et al.* SOX5 is involved in balanced MITF regulation in human melanoma cells. *Bmc Med Genomics* 9, 10 (2016).

178. Singh, A. M. *et al.* Cell-Cycle Control of Bivalent Epigenetic Domains Regulates the Exit from Pluripotency. *Stem Cell Rep* 5, 323–336 (2015).

179. Gonzales, K. A. U. *et al.* Deterministic Restriction on Pluripotent State Dissolution by Cell-Cycle Pathways. *Cell* 162, 564–579 (2015).

180. Wheway, G., Nazlamova, L. & Hancock, J. T. Signaling through the Primary Cilium. *Frontiers Cell Dev Biology* 6, 8 (2018).

181. Stecca, B. *et al.* Melanomas require HEDGEHOG-GLI signaling regulated by interactions between GLI1 and the RAS-MEK/AKT pathways. *Proc National Acad Sci* 104, 5895–5900 (2007).

182. Xue, G., Romano, E., Massi, D. & Mandalà, M. Wnt/β-catenin signaling in melanoma: Preclinical rationale and novel therapeutic insights. *Cancer Treat Rev* 49, 1–12 (2016).

183. Müller, C. S. L. Notch Signaling in Embryology and Cancer. *Adv Exp Med Biol* 727, 258–264 (2012).

184. Zingg, D. *et al.* EZH2-Mediated Primary Cilium Deconstruction Drives Metastatic Melanoma Formation. *Cancer Cell* 34, 69-84.e14 (2018).

185. Carreira, S. *et al.* Mitf regulation of Dia1 controls melanoma proliferation and invasiveness. *Gene Dev* 20, 3426–3439 (2006).

186. Levy, C., Khaled, M. & Fisher, D. E. MITF: master regulator of melanocyte development and melanoma oncogene. *Trends Mol Med* 12, 406–414 (2006).

187. Zehir, A. *et al.* Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat Med* 23, 703–713 (2017).

188. Ma, Y. *et al.* SOX9 Is Essential for Triple-Negative Breast Cancer Cell Survival and Metastasis. *Mol Cancer Res* 18, 1825–1838 (2020).

189. Liu, H. *et al.* SOX9 Overexpression Promotes Glioma Metastasis via Wnt/β-Catenin Signaling. *Cell Biochem Biophys* 73, 205–212 (2015).