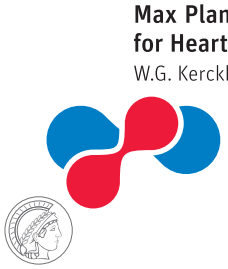


Max Planck Institute
for Heart and Lung Research
W.G. Kerckhoff Institute

Development and application of computational methods for cancer subtype detection from -omics data

Cumulative inaugural dissertation

in partial fulfillment of the requirments for the degree of
Doctor rerum naturalium (Dr. rer. nat.)

by

Jens Preussner

submitted to the
Faculty of Biology and Chemistry
Justus-Liebig-University
Giessen, Germany

prepared in the
Department of Cardiac Remodelling
Max Planck Insititute for Heart and Lung Research
Bad Nauheim, Germany

Bad Nauheim, 2020

Preface

Thesis reviewers

First reviewer Prof. Dr. Alexander Goesmann, *Bioinformatics and Systems Biology, Justus-Liebig-University, Giessen, Germany*

Second reviewer Prof. Dr. Dr. Thomas Braun, *Max Planck Institute for Heart and Lung Research, Bad Nauheim, Germany*

Examiner Prof. Dr. Reinhard Dammann, *Institute for Genetics, Justus-Liebig-University, Giessen, Germany*

Examiner Prof. Dr. Stefan Janssen, *Algorithmic Bioinformatics, Justus-Liebig-University, Giessen, Germany*

Date of defense

July 8th, 2020

Declaration

I declare that I have completed this dissertation single-handedly without the unauthorized help of a second party and only with the assistance acknowledged therein. I have appropriately acknowledged and cited all text passages that are derived verbatim from or are based on the content of published work of others, and all information relating to verbal communications. I consent to the use of an anti-plagiarism software to check my thesis. I have abided by the principles of good scientific conduct laid down in the charter of the Justus Liebig University Giessen "Satzung der Justus-Liebig-Universität Giessen zur Sicherung guter wissenschaftlicher Praxis" in carrying out the investigations described in the dissertation.

Abstract

Cancer is a complex and dynamic disease manifesting in ~100 distinct cancer types that arise in multiple cell types and organs due to different but related mechanisms. Research from the last decade has revealed vast heterogeneity within and between cancer types, hampering effective treatment and calling for more personalized treatment strategies. This thesis develops methodology for detection and molecular characterization of cancer subtypes by focusing on the analysis of experiments generating large datasets. The first objective was to provide algorithms for rapid detection and quantification of microRNAs and analysis and visualization of DNA methylation. The second objective was to investigate the cellular and molecular origin of embryonal rhabdomyosarcoma (ERMS), a rare and aggressive childhood cancer.

Two new computational methods were implemented and evaluated by comparison to previously published findings. DNA copy number alterations and gene expression estimates were obtained from a novel model system for ERMS and integrated with molecular data from cancer patients. Cell tracing experiments unambiguously demonstrated that ERMS is derived from tissue-resident muscle stem cells, at least in the model system used. In-depth data analysis revealed a diverse molecular basis of ERMS, confirming cancer heterogeneity. Surprisingly, activation of zygotic Dux factors identified a novel cancer subtype that is not limited to ERMS, but occurs in a broad range of different human cancer.

Based on the results, it can be concluded that computational methods and integrative data analysis are useful to delineate the origin of cancer subtypes and provide a valuable starting point for selection of relevant therapeutic targets. However, future research is needed to establish more holistic analysis approaches and transfer findings into existing clinical routines.

Contents

List of Figures

1 Introduction

1.1 Rationale

Cancer is one of the leading causes of death around the world, being responsible for nearly 10 million deaths in 2018 (Bray et al., 2018). Cancer is a complex disease that can arise in multiple tissues, originating in numerous cell types and by different albeit related mechanisms. Currently, ~100 distinct cancer types are known to emerge from interactions of hundreds to thousands of macromolecules. Recent decades of research have generated detailed insights into variations of cancer initiation, progression, severity and treatment resistance. However, a clear vision and path for the cure of cancer is still missing (Koutsogiannouli, Papavassiliou, & Papanikolaou, 2013; Nurse, 1997).

Efficiency of tumour treatment depends on and is affected by cellular and molecular tumour profiles. However, individual tumours from different patients exhibit different molecular profiles and properties like cellular morphology, gene expression, metabolism, proliferation or metastatic potential. Such intratumoural heterogeneity is caused by cancer subtypes and hampers effective design of treatment strategies. It is one of the biggest challenges for successful cancer treatment. A promising approach to overcome intratumoural heterogeneity aims to identify individual patients with similar cancer subtypes and to tailor specific treatments for those patients (Senft, Leiserson, Ruppin, & Ronai, 2017; Vincent, 2017). Termed *precision* medicine, it requires integration of patient data from multiple sources (e.g. genomics, epigenomics, clinical data, lifestyle and environment) to identify therapeutic targets that are essential for subtype-specific tumour initiation and progression. To fulfil those requirements and to support clinical decision making, appropriate computational methods for managing, integrating and analysing large and complex data sets are needed (Singer et al., 2017).

The initial sequencing of the human genome (International Human Genome Sequencing Consortium, 2001) has marked the beginning of cancer genomics and initiated a new era of modern biomedical research. Disruptive advances in DNA sequencing technology revolutionised not only cancer research, but also the way how genome-wide questions can be addressed (MacConaill, 2013). With the advent of *next generation sequencing* (NGS) technology, it became possible to profile cancer genomes (Pugh et al., 2012; Stephens et al., 2009), which significantly enhanced the ability to study neoplastic transformation based on changes in the genome sequence. Notwithstanding, enourmous amounts of data generated by NGS introduced new challenges in computational data analysis (Mardis, 2011; Wu, Rice, & Wang, 2012). Transformation of this data to gain a holistic understanding of the complex and dynamic systems of cancer is challenging (Grizzi et al., 2006; Sigston & Williams, 2017).

This thesis provides methodological development and application of software within the scope of computational cancer biology, focusing on the analysis of data from NGS experiments to support the characterization of the molecular basis of cancer subtypes and the investigation of mechanisms underlying cancer formation. The objectives are: to provide an algorithm for rapid detection and quantification of small, regulatory RNA; to develop a software pipeline for analysis and visualization of CpG-site methylation in a case-controlled setting; to investigate the cellular and molecular origin of a childhood cancer, embryonal rhabdomyosarcoma, by an integrative analysis of data generated from NGS experiments and to delineate potential cancer subtypes and the mechanisms of tumour formation.

1.2 Background

The origin of cancer and its development has been a subject of debate, covered by several theories. In the early nineteenth century, a professor of anatomy and pathology at the Royal Anatomical Museum in Berlin, Johannes Müller, recognised for the first time the cellular structure of cancer. Using microscopic pathology, he observed how *morbid growth* resembles the tissue from which the cancerous growth springs. Since then, modern oncology seeks the origin of cancer in a transformation of a healthy cell into a disease state characterized by uncoordinated and excessive cell growth.

The upcoming sections highlight several theories that provide explanation for different cancer-causing mechanisms with implications for cancer classification, diagnosis, therapy and research. Additionally, a brief introduction into integrated molecular data analysis is provided.

1.2.1 Genomic alterations transform cells into cancer cells

While normal cells retain their ability to control the production and release of growth-promoting signals and thereby provide tissue architecture and homeostasis of cell number, cancer cells are characterised by chronic proliferation and constant re-entry into the growth-and-division cycle (Hanahan & Weinberg, 2011). The somatic mutation theory postulates that molecular events such as genomic mutations precede cell transformation (Fig. 1a), black lightning and red cells) allowing cells to overcome cell control mechanisms. By such, transformed cancer cells are able to *e.g.* escape the control of growth suppressors, bypass mechanism of induced cell death, delay or avoid entry into cell senescence, induce angiogenesis and/or alter cell-to-cell contacts to activate invasion of surrounding tissue (Hanahan & Weinberg, 2011). At the core of this theory, mutations in so-called master genes, *i.e.* genes that have the potential to cause cancer (oncogenes) or genes that protect the cell from cancer progression (tumour suppressor genes), determine the onset of cancer.

The search for transformation-causing genome alterations accelerated through recent advances in NGS technologies which led to an impressive accumulation of data from large-scale genome sequencing projects like the Cancer Genome Project or The Cancer Genome Atlas. Collectively, both projects list 81 million simple somatic mutations across

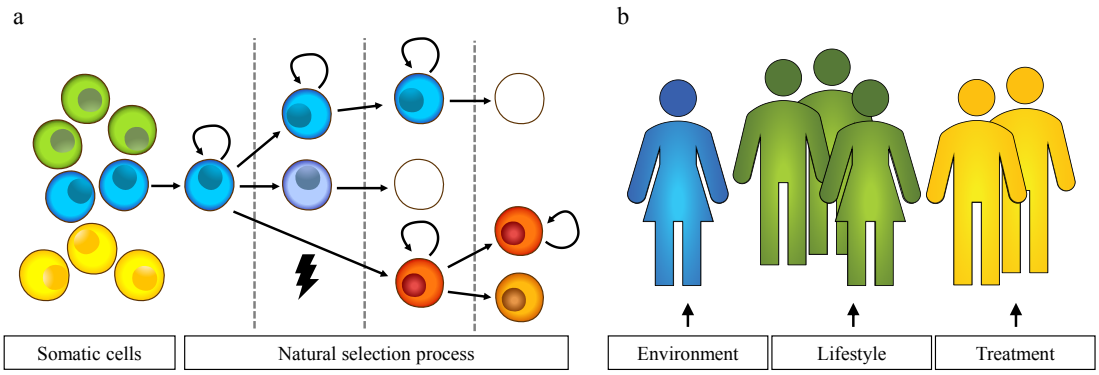


Figure 1: Heterogeneity in the process of cancer formation. (a) Molecular events (straight arrows, exemplified for one cell) increase genome context heterogeneity (colors) and confer differences in fitness between somatic cells. Oncogenic transformation (black lightning) enables positive natural selection and the evolution of cancer. (b) Cell population heterogeneity is additionally increased by external factors acting on the individuals genome context.

cancer genomes from nearly 25,000 individuals in the International Cancer Genome Consortium Data Portal Release 28. However, the search did not reveal a single pattern of genetic alterations that is universal to most cancers. Instead, a tremendous genetic heterogeneity in underlying mechanisms of cancer formation emerged, despite common features of cancer cells.

The observed genetic heterogeneity among cancer and even within similar cancer types suggests that most individual tumours exhibit altered genome contexts (genes, regulatory elements and genomic topology), with different genomic mutations. The Genome Theory is an extension of the somatic mutation theory and seeks to explain cancer heterogeneity by additionally including cell population heterogeneity and the process of natural selection (Heng et al. (2010), Fig. 1). Cancer formation is seen as an evolutionary process initiated through internal (*e.g.* somatic mutations) or external (*e.g.* environment, lifestlye, treatment) stress that results in genome context instability. Additional genetic or epigenetic mutations may occur in individual somatic cells with instable genomic contexts, increase the cell population heterogeneity and confer differences in fitness between somatic cells. Importantly, as genetic or epigenetic mutations are heritable and can be passed to a cells progeny, all requirements for natural selection are met and evolution of somatic cells within individual patients towards cancer is enabled.

The occurence of genome-level alterations in an instable genome context is a stochastic process and therefore the probability of successful progression towards cancer is highest through alterations significantly impacting the phenotype of a cell. However, the stochastic nature complicates the prediction of which pathway will become dominant prior to tumour formation and renders the characterization of individual cancer path-

ways less meaningful. Ultimately, it impedes the establishment of a general model of cancer origin and obscures understanding of how cancer can be managed in classical clinical treatment.

1.2.2 Disrupted context: Cancer as a tissue-based disease

Organs and higher-level structures are comprised of tissue with functional (parenchyma) and structural (stroma) parts. Information exchange between cells of a tissue via cell-to-cell contacts, cytokine signaling and the extracellular matrix, a macromolecular scaffold to support surrounding cells, enables maintenance of cell differentiation and tissue structure. Tissue-level interactions are important for embryonic development, regeneration and morphogenesis, *e.g.* to provide necessary mechanical forces for proper tissue formation (Hernández-Hernández, Rueda, Caballero, Alvarez-Buylla, & Benítez, 2014). Therefore, disruption of tissue organisation is thought to be carcinogenic through entailed disruption of tissue-level interactions.

The physiomitotic theory sees carcinogenesis as a problem of tissue organisation and relates the acquisition of mitotic activities to development of cancer by non-regulated cell turnover among normal tissue (Hirata & Hirata, 2002; Paduch, 2015). Two types of mitosis maintain tissue histology and continuity: duplicating mitosis regenerates a basal pool of undifferentiated cells in a space-restricted duplication area by creating two identical daughter cells. Those cells are constantly consumed in surrounding areas by maturating mitosis, which creates two daughter cells that are more mature than the parent cell and contribute to tissue diversity and functionality via cell differentiation (Fig. 2a). Disrupted tissue organisation and regulation may evoke duplicating mitosis at ectopic sites among normal tissue, thereby creating aberrant and undefined tissue identity. This might lead to non-regulated cell turnover and maturation into cancerous tissue (Fig. 2b).

Similarly, the tissue organisation field theory argues that interactions among different tissue components cannot be explained on a cellular level and that carcinogenesis cannot be reduced to cellular events (Paduch, 2015). Cancer initiation is preceded by a carcinogenic event and chronically disrupts reciprocal interactions between stroma and parenchyma of a tissue, but cannot be observed in individual, isolated cells.

A well-known example and important feature of interactions of stroma with surrounding cells is the maintenance of polarized epithelial sheets, a basic tissue type that lines outer surfaces of organs and surfaces of inner cavities. Cell polarity is established by interaction with the basement membrane and cell-to-cell contacts, like adherens junctions, gap junctions, tight junctions and desmosomes (Fig. 2c). Alterations of epithelial sheets, *e.g.* through wounding and subsequent activation of stromal fibroblasts, can lead to epithelial cell movement and proliferation. Similarly, sustained inflammation and continuous exposition to factors produced by invading immune cells and enzymes degrading the extracellular matrix (ECM, *e.g.* matrix metalloproteinases) stimulate proliferative and apoptotic mechanisms, which can lead to selection of apoptosis-resistant, premalignant cells and enhance formation of carcinoma (Fig. 2c, Bissell & Radisky (2001)). Loss or downregulation of E-cadherin, an important component of adherens

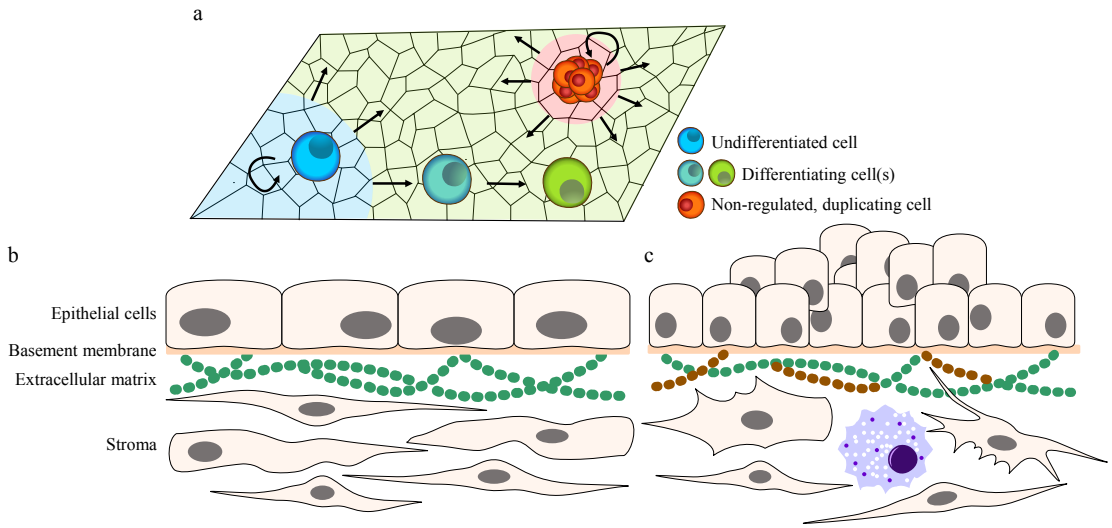


Figure 2: Cancer as a tissue-based disease. (a) Two types of mitosis maintain normal tissue histology and continuity under the physiomitotic theory: Duplicating mitosis (circular arrow) in a space-restricted duplication area (light blue shading) regenerates a basal pool of undifferentiated cells (blue), which are consumed by maturating mitosis (straight arrows) and contribute to more differentiated cells (green) to establish diverse and functional tissue (green shading). Duplicating mitosis at ectopic sites (red shading) among normal tissue might create cancerous tissue (red cells) and lead to non-regulated cell turnover (circular and straight arrows). (b) Schematic depiction of basic epithelial sheets. Epithelial cells are polarized by interaction with the basement membrane and underlying stroma. (c) Alteration of epithelial sheets through activation of stromal fibroblasts (jagged cells), degraded ECM (brown) or invading immune cells (blue) can stimulate epithelial cell movement and proliferation.

junctions, leads to a premalignant cell type that is prone to invasion and metastasis by passing through an epithelial-to-mesenchymal transition (Christofori & Semb, 1999). Accordingly, restoration of E-cadherin expression in such cells can suppress cellular transformation.

Experiments have shown that restoration of the cellular micro-environment can also lead to healthy phenotypes, *e.g.* normal differentiation is observed when teracarcinoma cells are injected into blastocysts, even after long passaging (Illmensee & Mintz, 1976). Similarly, experiments from 3D culturing with reconstituted basement membranes or co-culturing of malignant cells with normal stroma reverted their carcinogenic properties (Weaver et al., 1997). In summary, these experiments suggest that normal stroma provides contextual cues that promotes normal tissue identity and restricts proliferation of existing pre-malignant cells. In contrast, non-functional stroma releases this suppression and permits neoplastic transformation (Bissell & Radisky, 2001). These findings lead to the hypothesis that carcinogenesis might be reversed when neoplastic

tissue comes into contact with functional tissue or components thereof.

1.2.3 The stem cell as a cancer cell of origin

Similar to maintenance of tissue identity, the development of normal tissue requires complex crosstalk between cells, their local environment and the whole organ. Migration and proper localization of precursor cells is a prerequisite for formation of mature descendants that can carry out their tissue-specific function. The specific pattern of a cells tissue-forming division(s) has been termed its *lineage* (Chisholm, 2001). Furthermore, dissection of such cell lineages revealed a hierarchical organisation and helped to identify interactions and molecular signalling pathways that are important in tissue development and diseases. However, unidirectional division alongside the cell lineage would quickly lead to exhaustion of a cells tissue-generative potential and therefor calls for a mechanism such as duplicating mitosis as proposed by the physiomitotic theory: **Stem cells** are tissue-specific multipotent precursor cells residing at the apex of a lineage, and are capable of both (i), generation of common progenitor cells with increasing lineage commitment and (ii) self-renewal to regenerate and sustain the pool of stem cells. The inherent proliferative capacity and the ability to give rise to different, mature cell types renders stem cells particularly fascinating for the study of tissue development, regeneration and in the search for the cellular origin of cancer.

An important distinction has to be made between the origin of cancer cells (*i.e.* the normal cell that acquires the first cancer-promoting alteration (Creton et al., 2012)) and cancer stem cells, *i.e.* a cellular subset within a tumour that exclusively sustains malignant growth (Visvader, 2011). Intertumoural heterogeneity, *i.e.* the variability among discrete tumour types arising from the same tissue, has put forward two hypothesis how cancer stem cells are formed: (i) All tumours originate from common progenitor cells that accumulate different genetic or epigenetic mutations through their extended longevity and therefore result in different tumour types or (ii) different cells along the lineage hierarchy that still possess or can re-instigate proliferative capacity or prevent terminal differentiation (*e.g.* more restricted progenitor cells) constitute different cancer cell types including cancer stem cells upon oncogenic transformation (Perez-Losada & Balmain (2003), Visvader (2011) and Fig. 3a). Cells with self-renewal capacity are of paramount importance for tumour growth as they ensure long-term clonal growth. However, not all cancer cells possess self-renewal capacity and not all cells from which cancer origins are *bona fide* stem cells. So, how do cancer cells acquire their stemness, if not from normal tissue stem cells?

The *lineage-dependency* hypothesis suggests that many tumours might be dependent on (or *addicted to*) lineage-survival programmes that also operate during normal lineage development (Garraway & Sellers, 2006). In this view, cancer cells can aquire their stemness from lineage precursor cells, but rely on persistence and deregulation of lineage-specific proliferation and differentiation pathways (Fig. 3b). The *lineage-dependency* hypothesis inextricably associates lineage descendance and differentiation state of progenitor cells to cancer biology and complements *oncogene addiction*, in which

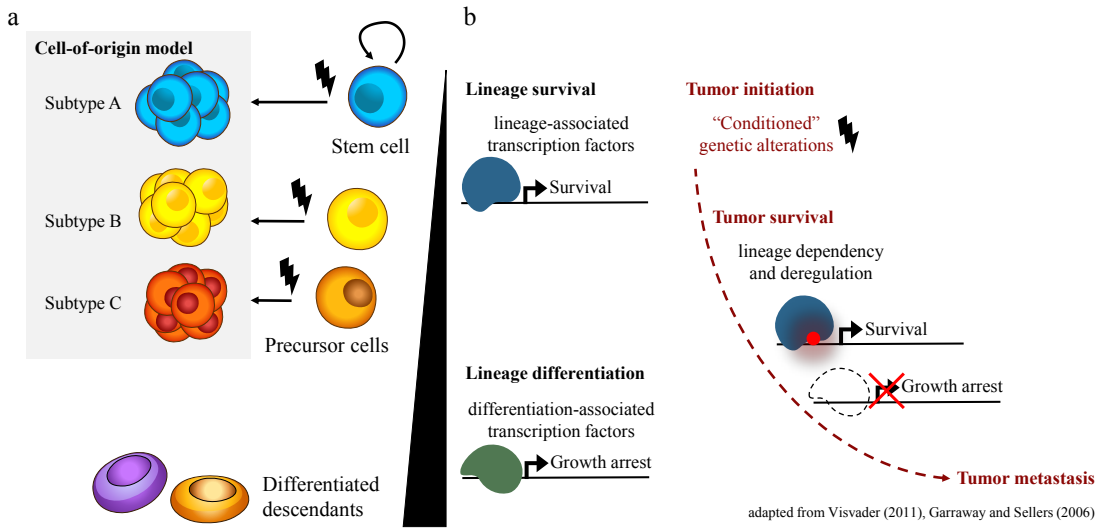


Figure 3: Cell lineages in tumour initiation and heterogeneity. (a) Cells along a lineage hierarchy that still possess proliferative capacity or can prevent terminal differentiation constitute cancer subtypes upon oncogenic transformation (black lightning). (b) Lineage survival and normal development are often dependent on lineage-associated transcription factors. Genetic alterations might be conditioned by lineage and subsequent tumour initiation crucially depends on persistence or deregulation of survival mechanisms programmed into precursor cell development: A mechanism termed lineage dependency.

tumour-specific gain-of-function events elicit a dependency on growth signalling that is absent in normal lineage development.

Activation of the same oncogenic pathway in tumours originating from different cell lineages may also profoundly influence tumour phenotype and degree of malignancy. For example, primary human melanocytes transformed with a set of genes form melanomas that frequently metastasize, while human fibroblasts or epithelial cells transformed with the identical set of genes rarely do (Gupta et al., 2005). Ultimately, therapeutic approaches might exploit *lineage dependency* for context-specific treatment, for example when *synthetic lethality* exists between two genetic factors (Kaelin, 2005).

A straightforward approach to evaluate the oncogenic capacity of different lineage stem and progenitor cell populations relies on reproducible separation of functionally defined subpopulations using *e.g.* cell sorting techniques. Relevant oncogenic lesions are introduced together into different precursor cell populations *ex vivo* with a fluorescent reporter, followed by orthotopical transplatation into immunocompromised mice. Emergence of pre-neoplastic or neoplastic tissue from transduced subpopulations serves as readout for evaluation of oncogenic potential for each subpopulation. Complimentary, and with sufficient knowledge about cell-specific promoters, *in vivo* conditional targeting of cell populations is also conceivable. This approach makes use of genetic mouse models to conditionally activate either an oncogene or inactivate a tumour suppressor

gene in different lineage subpopulations, *e.g.* by Cre-mediated deletion. Depending on the activated cell-specific promoter, different cancer subtypes might arise and reveal the cellular origin of the specific cancer subtype from within the cell lineage (Hayashi & McMahon, 2002; Visvader, 2011). However, established lineages and knowledge of cell specific promoters are missing for many tissues and organs and therefore hamper the approach described above.

1.2.4 Skeletal muscle regeneration as a model for stem cell tumours?

The ability of movement is an evolutionary advantage of all animals, and is powered by muscles. Vertebrate locomotion receives its power from striated, skeletal muscle, one of the three major muscle types in the body, that is composed of multiple bundled muscle fibres (fascicles). Each fibre is a multinucleated muscle cell formed by fusion of differentiated mononuclear muscle cells (myoblasts) and exhibits force and movement by coordinated activity of myosin II motor proteins within an actin filament scaffold. Skeletal muscle retains a remarkable ability to regenerate and adapt to changes in requirements, mediated by and dependent on muscle stem cells that reside in a niche between the muscle sarcolemma and the basal lamina of individual muscle fibres. Adult muscle stem cells are *bona fide* stem cells, being capable of both, self-renewal and myogenic differentiation, which ultimately leads to differentiated muscle cells (Almada & Wagers, 2016; Günther et al., 2013).

Muscle stem cells that are characterised by expression of Paired box protein 7 (Pax7), a transcriptional regulator, are mainly quiescent under homeostatic conditions. Upon muscle trauma, otherwise quiescent muscle stem cells become activated through exposure to extrinsic stimuli and switch to a highly proliferative state. Activation and proliferation of muscle stem cells depends on expression of two transcriptional regulators, myogenic factor 5 (Myf5) and myogenic determintation protein (Myod1), and precedes commitment to differentiation (Braun & Gautel, 2011). Downregulation of Pax7 and expression of myogenin (Myog) in a subset of activated muscle stem cells induces differentiation and ultimately leads to cell-cycle exit and formation of myoblasts that fuse with other myoblasts or existing muscle fibres to repair the muscle (Almada & Wagers, 2016; Braun & Gautel, 2011). Activated stem cells may also inhibit Myod1 expression and re-instating quiescence, thereby replenishing the pool of muscle stem cells for future rounds of muscle repair (Fig. 4).

Duchenne muscular dystrophy (DMD) is a genetic disorder leading to muscle weakness and decrease in the muscle mass (muscle wasting, atrophy). Dystrophin, the gene product causing DMD in affected individuals, is part of a larger complex that stabilizes the membrane of striated muscle cells. Dystrophic fibres are prone to get damaged by mechanical stress and die after repeated muscle contraction. Such fibres are often replaced by fibrotic, adipose or connective tissue that is not able to transmit sufficient muscular force (Almada & Wagers, 2016). Muscle degeneration elicits repair by expansion and differentiation of stem cells, but regenerated muscle fibres will also lack a functional dystrophin such that chronic cycles of degeneration and regeneration are passed through. Until now, the role of muscle stem cells in DMD remains elusive, with

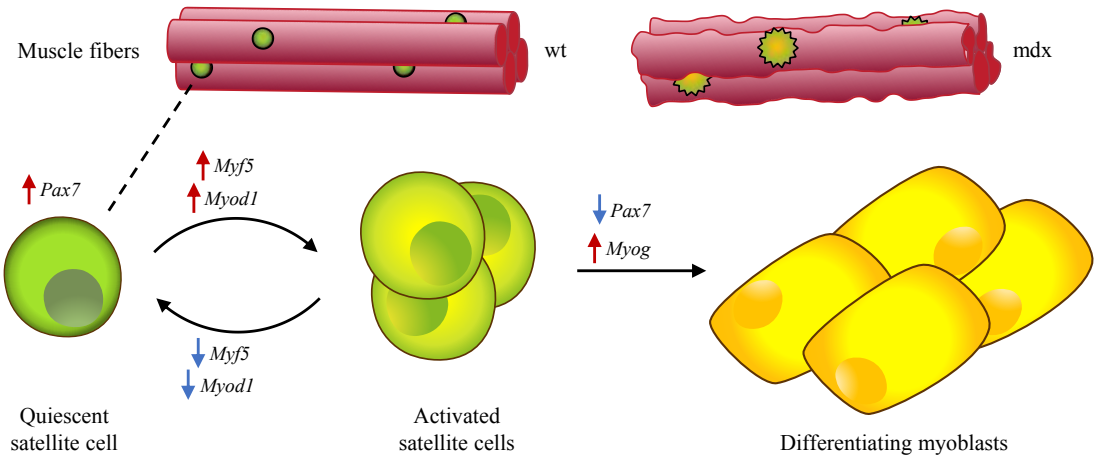


Figure 4: Stem-cell dependent regeneration of skeletal muscle fibres. In uninjured muscle (wt), quiescent muscle stem cells (left) reside between the muscle sarcolemma and the basal lamina of individual muscle fibres. Pro-myogenic stimuli from muscle trauma or under genetic disorders like DMD (mdx, right) activate muscle stem cells and lead to proliferation (middle). Differentiating myoblasts (right) arise through cell-cycle exit of a subset of activated muscle stem cells and constitute a basis for formation of novel fibres or repair of existing fibres.

only some evidence for a specialized role of dystrophin during stem cell division, but an important role of dystrophin for the pathological environment in disease progression (Almada & Wagers, 2016).

A genetic and experimental model of DMD is the *mdx* mouse, whose muscles retain a lifelong capacity to regenerate fibres and exhibits loss of muscle fibres and extensive fibrosis (Boldrin, Zammit, & Morgan (2015), Fig. 4). Recently, it was shown that germline inactivation of the tumour suppressor p53 in chronically regenerating mdx mice develop rhabdomyosarcoma (RMS) (Camboni, Hammond, Martin, & Martin, 2012; Chamberlain, Metzger, Reyes, Townsend, & Faulkner, 2007), a rare and aggressive childhood cancer and the most common soft-tissue sarcoma in children and adolescents. Histologically, RMS resembles developing skeletal muscle and is marked by expression of actin and myosin as well as myogenic factors (Drummond et al., 2018; El Demellawy, McGowan-Jordan, de Nanassy, Chernetsova, & Nasr, 2017). RMS is subdivided into four subgroups, with alveolar RMS (ARMS) and embryonal RMS (ERMS) being two major subgroups accounting for nearly all childhood cases of RMS, while spindle cell RMS and pleomorphic RMS occurring mostly in adolescents. A broad molecular basis has been identified in RMS, with interference of myogenic differentiation and emergence of chromosomal aberrations being main drivers of cancer progression. For example, aberrant expression of Notch2, Yap1, members of the Wnt gene family and Tgf-1 signalling have been implicated in disruption of myogenic differentiation (Chen et al., 2014; Judson et al., 2012; Schaaf et al., 2005; Wang et al., 2010). On the other hand, ex-

pression of Egr1, Met and signalling by the Fgf family seem to maintain proliferation of RMS cells (Sarver, Li, & Subramanian, 2010; Taulli et al., 2006; Wachtel et al., 2014). Genomic amplifications and translocations, as well as loss of heterozygosity from specific whole chromosomes have been reported for ERMS and ARMS subtypes. Prominent examples include interference of Pax3 and Pax7 expression levels in ERMS, promoting migration and invasiveness (Bridge et al., 2002; Chiappalupi, Riuzzi, Fulle, Donato, & Sorci, 2014), and Pax3-Foxo1 or Pax7-Foxo1 gene fusions in ARMS, leading to commitment of mesenchymal stem cells to the myogenic lineage by transactivation of Myod1 or Myog (Ren et al., 2008). In addition to genetic mechanisms, epigenetic and small RNA-mediated mechanisms have also been described to deregulate myogenic differentiation enabling escape of RMS cells from suppressive mechanisms (see also chapters 1.2.5 and 1.2.6).

RMS is marked by large heterogeneity that not only manifests in distinct subtypes, but also by different underlying genetic and epigenetic mechanisms. However, the cellular origin of RMS has remained elusive, despite large efforts to characterize the molecular basis of many RMS specimen in recent years. As introduced in chapter 1.2.3, high tumour heterogeneity can emerge from a cellular origin with stem cell-like properties. Subsequently, the cellular origin of RMS was claimed to reside in tissue-resident stem cells, *e.g.* muscle stem cells or mesenchymal stem cells (Hettmer & Wagers, 2010). As such and for RMS subtypes showing features of myogenic differentiation, the *mdx* mouse model with its constant regeneration of skeletal muscle could be used as a model for stem cell-dependent carcinogenesis and serve the discovery of the cellular origin of RMS.

1.2.5 Small RNA mediated carcinogenesis

Micro RNAs (miRNAs) are small, non-coding RNAs of ~22 nucleotides and serve numerous roles in negative gene regulation. In animals, most miRNAs exhibit their regulatory role through imperfect binding of a sequence in the 3' untranslated region (3'-UTR) of messenger RNA from target genes. Complementary binding can either repress translation of target gene(s) or mediate mRNA degradation, through a mechanism similar to RNA interference in plants (Jones-Rhoades, Bartel, & Bartel, 2006).

MiRNA biogenesis begins with the transcription of either independent miRNA genes or intronic regions from protein-coding genes into large precursor molecules (pri-miRNAs). Imperfect base-paring of folding pri-miRNAs results in hairpin structures that are further cleaved by an RNase III type endonuclease (Drosha, RN3) together with a double-stranded RNA binding domain (dsRBD) protein (Han, 2004) into ~70 nucleotide hairpins called pre-miRNAs, leaving a short characteristic single-stranded overhang at the 3'-end of pre-miRNAs. Exportin 5 recognizes such an overhang and arranges the transport of pre-miRNAs to the cytoplasm (Yi, 2003), where a second complex consisting of Dicer, a RNase III type enzyme, and TRBP, a dsRBD protein, cleave the pre-miRNA twice into a miRNA duplex (Chendrimada et al., 2005). One strand (the mature miRNA) preferentially enters the miRNA-induced silencing complex (miRISC), while the other is degraded, although the complementary miRNA is also competent for miRNA-mediated silencing (Schwarz et al., 2003). Imperfect double-strand pairing of pre-miRNAs as well

as imperfect digestion by Drosha and Dicer result in miRNAs with varying 3'- (silent modification, *isomiR*) or 5'-ends, that might affect complementary binding, representing a challenge for computational miRNA quantification following RNASeq.

Forward genetic experiments have revealed great importance of the miRNA machinery, exemplified by muscle-specific knockout of Dicer leading to complete embryonic development but perinatal death (Bernstein et al., 2003). Other experiments identified numerous individual miRNAs with roles in processes, like timing of development (Abrahante et al., 2003), differentiation (Chen, 2004) and growth control (Brennecke, Hipfner, Stark, Russell, & Cohen, 2003). Since phenotypic consequence remains elusive for the vast majority of miRNA, also computational approaches can and have been used to elucidate miRNA function (Liu & Wang, 2019; Ulitsky, Laurent, & Shamir, 2010).

Progression of cancer growth can also be altered by expression of certain miRNAs. The dedifferentiated phenotype of ERMS, for example, can result from downregulation of muscle-specific miRNAs (*myomiRs*, i.e. miR-1, miR-206 and miR-133a/b), that promote myogenic differentiation under normal conditions. Transfection of miR-206, a skeletal muscle-specific miRNA, induces cell differentiation in C2C12 myoblast cells (Kim, Lee, Sivaprasad, Malhotra, & Dutta, 2006). Additionally, expression of miR-1 and miR-206 are highly induced during muscle stem cell differentiation (Chen et al., 2010) and transfection of miR-206 into a RMS cell line notably decreased tumour cell migration and proliferation even more than switching to a differentiation medium (Taulli et al., 2009). Contrarily to those findings Boettger, Wüst, Nolte, & Braun (2014) report on miR-206/miR-133b dispensability for muscle stem cell differentiation, highlighting complex modulatory effects and overlapping functions of *myomiRs*. Non-myomiRs can as well promote myogenic differentiation, *e.g.* miR-26a mediates downregulation of cell-cycle progression by targeting Ezh2. Vice versa, downregulation of miR-26a in RMS results in upregulation of Ezh2 and therefore prevents myogenic differentiation (Ciarapica et al., 2009). Amplification of 13q31 in 25% of ARMS cases results in enhanced expression of the miR-17-92 cluster (*oncomiR-1*), a *bona fide* oncogene (Jin et al., 2013; Reichek et al., 2011; Sandhu et al., 2013), potentially targeting tumour suppressors like PTEN. Deregulation of another oncogene in RMS, miR-183, is reported to promote tumour cell migration, by targeting the transcription factor Egr1, a direct regulator upstream of other tumour suppressor genes (Mohamad, Kazim, Adhikari, & Davie, 2018; Sarver et al., 2010).

1.2.6 Epigenetic mechanisms in cancer initiation and progression

Development of normal tissue, as discussed in chapter 1.2.4, requires distinct cell types to arise during lineage-specification. Although equipped with identical genomic information, different cell types exhibit different gene expression programs and are able to pass such information on to their progeny (Margueron & Reinberg, 2010). How are such expression patterns specified and maintained? It is now accepted that not only information encoded as DNA in a cells genome, but also *epigenetic* information (*i.e.* the stable and heritable non-genetic counterpart to DNA) provides an important layer of regulation and plays pivotal roles in cell lineage specification and cell identity main-

tenance (Margueron & Reinberg, 2010). In eukaryotes, DNA is organized in a massive macromolecular complex called chromatin. Chromatin is formed by wrapping 147 base pairs of DNA around a histone octamer (nucleosomes), then compacting those further into topologically associated domains (TADs) separated by insulator proteins to allow independent and specific regulation. The activity of a genomic locus is controlled by its chromatin organisation. Accessible chromatin structures expose DNA elements, like proximal gene promoter sequences or distal enhancer sequences, to regulatory transcription factors and the transcriptional machinery to drive gene expression. Compact and inaccessible chromatin structures prevent such activity and render a locus inactive (Flavahan, Gaskell, & Bernstein, 2017).

Dynamic changes needed during tissue development call for mechanisms able to alter chromatin organisation in response to changed conditions (John & Rougeulle, 2018). Chromatin remodelling resulting in transcriptional repression is, for example, enforced by the Polycomb protein family, which can post-transcriptionally modify specific histone residues (*e.g.* trimethylation of histone H3, lysine 27 (H3K27me3)). Repressive chromatin states can be propagated through cell division by retention of catalytic enzymes on replicating DNA (Simon & Kingston, 2013) and functional interaction with DNA methylation and other regulatory proteins (Flavahan et al., 2017). Conversely, regulatory activity by *e.g.* binding of transcription factors and chromatin modifiers, seems to block repressive chromatin compaction (Zaret & Mango, 2016). Further, active loci that are marked by trimethylation of histone H3, lysine 4 (H3K4me3), in turn inhibit recruitment of DNA methyltransferases for *de novo* DNA methylation (Ooi et al., 2007), which is another potent epigenetic mechanism for stabilization of transcriptional repression (Jones, 2012). Methylated DNA functions as a *silencing* mark and is involved in processes like X-chromosome inactivation (Venolia & Gartler, 1983), repression of transposable and repetitive DNA elements (Yoder, Walsh, & Bestor, 1997) and might influence genome function when present at regulatory elements, like enhancers or chromatin insulators (Bell & Felsenfeld, 2000).

A compelling conceptualization of epigenetic regulation during cell lineage development has been postulated by developmental biologist Conrad H. Waddington, outlined in his assay entitled *The strategy of genes* more than 60 years ago (reprinted in Waddington (2014)). In his hypothesis, differentiating cells proceed downhill along branching valleys in an energetic landscape (Fig. 5a). The valleys correspond to discrete cellular states and their topological layout is defined by underlying gene regulatory networks (GRN) that actively shape and maintain cellular identity (Zaret & Mango, 2016). Walls between valleys restrict cell lineage capacity, by preventing cells to randomly "switch states" (*i.e.* hopping over to another valley), and epigenetic mechanisms effectively modulate the height of walls. Compacted and repressing chromatin, for example, prevents spurious activation of non-lineage gene regulatory factors, restricting changes in gene activity and increasing the height of energy walls between cell states, which blocks changes in cell state and cell type identity (Flavahan et al., 2017).

Initiation and progression of cancer can result from various mechanisms disrupting normal epigenetic regulation. Overly restrictive chromatin (*i.e.* high energy walls between valleys of Waddingtons landscape, Fig. 5b) can be achieved by gain-of-function

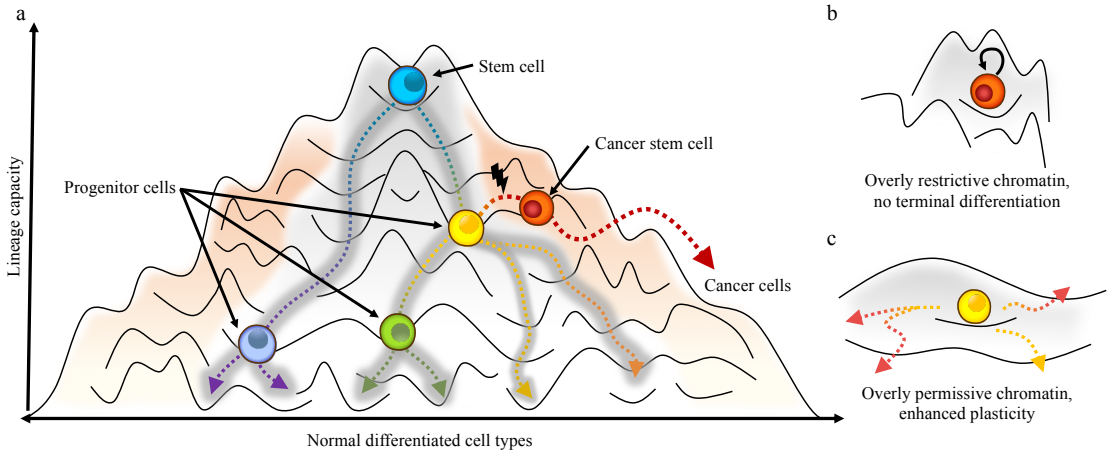


Figure 5: The epigenetic landscape in lineage development and cancer initiation. (a) Depiction of normal lineage development in the conceptualized epigenetic landscape from C. Waddington. Stem cells (blue) reside at the apex of a lineage and exhibit high lineage capacity. Progenitor cells proceed downhill along branching valleys into differentiated cell types. Oncogenetic events (black lightning) in lineage progeny with high lineage capacity can create cancer stem cells by switching cell states and might lead to development of cancer (red arrow). (b) Restrictive chromatin might arrest progenitor cells in their proliferative state, prevent their terminal differentiation and lead to cancer initiation. (c) Deteriorated and permissive chromatin confers enhanced cellular plasticity and might lead to spurious gene activation or cellular reprogramming and predisposes for cancer initation.

mutations of Ezh2, the catalytic subunit of Polycomb repressive complex 2 (Prc2). A hyperactive methyltransferase activity of Ezh2 leads to expansive, genome-wide H3K27 methylation (Sneeringer et al., 2010) and the loss of active chromatin marks. Overly restrictive chromatin arrests developing B-cells in a proliferative state and prevents terminal differentiation, leading to B-cell lymphoma (Béguelin et al., 2013). Epigenetic restriction can also arise from aberrations in DNA methylation: The CpG island methylator phenotype (CIMP) results from DNA hypermethylation and is characterized by silencing of tumour suppressor genes and DNA mismatch repair genes (Hitchins et al., 2007). Deterioration of overall chromatin topology (the layout of Waddingtons landscape, Fig. 5c) can be achieved by disruption of CTCF binding (Flavahan et al., 2016), a methylation-sensitive DNA binding protein that accomplishes partitioning of chromosomal loops into functional units by insulating TADs. Regulatory TAD boundaries protect against gene activation from overly promiscuous enhancers from neighboring TADs and loss thereof can lead to the activation of oncogenes (Hnisz et al., 2016). Upregulation of members of the Histone Lysine Demethylase (Kdm) protein family has been implicated in formation of overly permissive chromatin (*i.e.* low energy walls between valleys). Enhanced epigenetic plasticity allows for rapid cell reprogramming or adaptation and

drives diverse cancer types (Black et al., 2015; Liau et al., 2017; Roesch et al., 2013). Finally, loss of imprinting through DNA hypomethylation may permit reactivation of oncogenes, *e.g.* the insulin growth factor signalling pathway in some sarcomas (Rikhof, de Jong, Suurmeijer, Meijer, & van der Graaf, 2009). In such cases, both, the maternal and paternal copy of the Igf2 gene are transcribed, leading to elevated mRNA levels and predispose for cancer through autocrine signalling.

1.2.7 Cancer subtype detection using integrated molecular analysis

Identification and assignment of a tumour's molecular subtype an important step towards *precision medicine* and a prerequisite for tailoring therapy for individual patients (Senft et al., 2017; Vincent, 2017). Traditional clinical practice from the last decades is largely based on identification and assignment of tumour subtype on histopathology, cytology and expression or mutational status of known tumour markers. The clinical outcome is often determined by the individual expertise of clinicians and the classification scheme used (Ellis, 2006). With the advent of NGS technology and the establishment of novel clinical routines for sample processing and data analysis, molecular characterization of tumours became feasible (Noushmehr et al., 2010; Prat & Perou, 2011), promising to fulfil two goals of *precision medicine*: First, the discovery of molecular biomarkers that are predictive of disease outcome or effective drug treatment and second, a better mechanistic understanding of the molecular basis of tumour initiation and progression (Senft et al., 2017).

Increasing scale of NGS-based assays has so far been very useful in dissection of tumour heterogeneity. Although genome-wide screening of mutational status (Kuijjer, Paulson, Salzman, Ding, & Quackenbush, 2018), miRNA expression (Blenkiron et al., 2007), DNA methylation (Zhang et al., 2018) and RNA expression have led to the characterization of many cancer subtypes, the characterization of cancer using isolated assays suffers from certain limitations. For example, genomic profiling alone detects the presence or absence of genetic drivers, but fails to predict the activity of corresponding proteins and pathways. Simultaneous characterization using two or more assays might overcome such limitations and enhance clinical decision making towards targeted therapy, but requires effective integration strategies.

Data integration combines data from different sources, thereby enhancing accessibility and possibly enriching results from queries. Data are typically integrated across two axes: vertically, *i.e.* between different data types (*e.g.* genomic data, expression data or clinical data) and/or horizontally, *i.e.* within the same datatype, data from different providers or batches. The Cancer Genome Atlas Research Network provides a large collection of tumour samples that have been characterized using different assays (Hoadley et al., 2018) from numerous data generation centers. An early approach for vertical data integration used results from separate clustering of data types and performs clustering of cluster assignments (CoCA, Hoadley et al. (2014)). However, such an approach does not benefit from synergistic effects of combining evidence levels. It was succeeded by methods using simultaneous interrogation of subtype clustering from different data types, like iCluster (Shen, Olshen, & Ladanyi, 2009), which jointly models cancer sub-

types as latent variables from different data types, or tumorMap (Newton et al., 2017), which uses similarity of molecular tumour profiles to embed samples into a standardized similarity space. A natural approach to encode sample similarity is graph-based: Wang et al. (2014) constructed sample similarity networks for each data type individually and subsequently fused those networks into a common similarity space, thereby performing joint vertical and horizontal data integration. In their setup, cancer subtypes emerge as connected components in the graph and edges between samples provide information about the data type from which the evidence comes from. Interpretability is also a feature of Multi-Omics Factor Analysis (MOFA, Argelaguet et al. (2018)), a versatile statistical framework that infers a low-dimensional representation and captures major sources of variation across data types. Latent factors underlying the representation can be linked to most relevant features revealing shared variation between different omics layers. Recent technological advancement in the field of deep neural networks led to the development of methods for cancer subtype classification (Gao et al., 2019; Tabibu, Vinod, & Jawahar, 2019), using molecular or histological features for classifier training. However and in contrast to other applications, classification by neural networks is a supervised task and can only be used to classify new samples, but does not detect novel subtypes once more data are available.

1.3 Lack of knowledge and objectives

The question of how and why cancer develops already has led to numerous studies to characterise molecular tumour profiles and the cellular origin of different cancer types. Methods from molecular biology and assays based on high-density micro arrays or NGS are often used in conjunction to draw conclusions and validate results. A plethora of computational methods have been implemented for analysis of data from those assays and form the basis for approaches integrating across different data types. The upcoming sections will line out major knowledge gaps in computational approaches to handle data from such assays and in the search of cancer origin and cancer subtype detection, exemplified using ERMS. Further, objectives of this thesis were developed from identified knowledge gaps.

1.3.1 Knowledge gaps

In an extensive genomic analysis of tumours from 9 ERMS patients, Chen et al. (2013) set out to define biological signatures to predict patient outcome and assign targeted therapy for a high-risk subtype. They identified a subtype with defects in oxidative metabolism, but otherwise report a multiplicity of mutations in known cancer consensus pathways such as in Ras family genes, SHH/Wnt signalling or cell-cycle checkpoints. Analysis of therapy-resistant tumour subclones highlighted complex genetic changes and clonal evolution. The observations from Chen et. al., can answer two questions: (i) Which genomic mutations occur in ERMS and (ii) are mutations reoccurring? The ERMS subtype accumulated both, single nucleotide polymorphisms and larger copy number variations. Genomic alterations were not reoccurring among cancer samples

and indicate a large tumour heterogeneity - a result that complicates further subtype detection and follows observations from large genome sequencing projects. Extending the somatic mutation theory (see chapter 1.2.1), the genome theory explains cancer heterogeneity by introducing karyotype heterogeneity. Further, the genome theory assumes that initial genome context instability ignites a series of genome alterations from which the fittest (in terms of cell proliferation and expansion) is selected by an evolutionary process. Open question thus are

- (i) Which mechanisms induce genomic instability in ERMS?
- (ii) How do accumulated genome alterations confer to cellular fitness?
- (iii) Which hallmarks do accumulated genome alterations operate?
- (iv) Do accumulated genome alterations further specify ERMS subtypes?

Following the stem cell theory of cancer (see chapter 1.2.3), the cancer cell of origin of ERMS could possibly be either the muscle stem cell itself, or any of its potent progenitor cells, *e.g.* activated stem cells or pro-myogenic precursor cells. Elevated expression of Pax7, which occurs exclusively in fusion-negative RMS, led Tiffin, Williams, Shipley, & Pritchard-Jones (2003) to propose the origin of RMS within the myogenic lineage. Transformation of precursor cell populations from the myogenic lineage with either expression of oncogenic Kras (Blum et al., 2013; Hettmer et al., 2011), or lineage-specific deletion of tumour suppressor genes (Rubin et al., 2011), led to tumours that phenotypically resembled their *presumable* myogenic origin. However, it was not possible to discriminate the cellular origin of Kras-expressing tumours by transcriptional analysis (Hettmer et al., 2011). The studies presented above lack direct experimental evidence, leaving the possibility that other cell types in the muscle compartment might act as tumour initiator by *e.g.* cell migration mechanisms. Therefore, the cellular origin of RMS ultimately remained to be disclosed.

MiRNAs play a role in cancer initiation, progression and maintenance (see chapter 1.2.5) and discovery of the entire regulatory repertoire of these small molecules is crucial for understanding their function in a given biological system (Gomes et al., 2013). With the application of next generation sequencing in miRNA research (Tam, de Borja, Tsao, & McPherson, 2014), the numbers of identified miRNAs increased rapidly, as well as computational approaches to predict or detect them (Gomes et al., 2013). However, until recently, *isomiR* variation due to imperfect digestion of pre-miRs by Drosha and Dicer was dismissed as sequencing artefacts and led to underestimation of the *miRNome* complexity (Neilsen, Goodall, & Bracken, 2012), which constitute a challenge for proper computational detection and quantification of miRNAs. Further, miRNA detection often relies on the presence of a preferably complete genomic reference to align miRNA reads to a genomic locus or reference database, or extensive homology searches to exploit evolutionary conservation of a nearby species. Thus, two open questions in analyses of data from miRNA-seq are: (i) How can *isomiR* variation be detected and properly quantified and (ii) does a reference-free approach to miRNA detection exist and is it as

sensitive as the conventional approach described above? Can it additionally be used to incorporate knowledge from other organisms?

To functionally characterise miRNAs, the delineation of miRNA target genes, *i.e.* those genes that could be silence by complimentary miRNA binding, is required and often accomplished using *in silico* target prediction tools (reviewed in Oulas et al. (2015)). Although prior knowledge from databases harbouring validated miRNA target interactions (MTIs) exists, target prediction and integration of other experimental data, *e.g.* from expression studies, remains challenging (Bayer, Kuenne, Preussner, & Looso, 2016).

Epigenetic regulation plays an important role in cell lineage development and disruption of such regulation might serve tumour initiation as well as tumour maintenance (see chapter 1.2.6). However, mechanistic insight into how epigenetic lesions take effect (also in cooperation with or followed by ordinary genetic stimuli) is missing. Open questions with diagnostic and therapeutic implications include whether or not an initiating genetic hit (*e.g.* gain or loss of function) becomes secondary, once a downstream epigenetic lesion has occurred and altered the cellular state towards permanent tumourigenicity. Advances in microarray and next generation sequencing technology enable assaying different mechanisms of epigenetic regulation at high resolution and in large numbers of samples (Lister & Ecker, 2009; The ENCODE Project Consortium et al., 2007) but require specialised computational analysis (reviewed in Bock & Lengauer (2008)). Finally, DNA methylation as an important mechanism of epigenetic regulation is included, due to it's important contribution to cancer development and diagnosis (Kulis & Esteller, 2010; Seki et al., 2015; Sun et al., 2019). Profiling DNA methylation using microarray technology (Bibikova et al., 2011) allows researchers to assay large number of samples across the whole genome. Computational analysis of data from such technologies aims to identify differentially methylated regions between two or more groups and search for functional enrichment in those regions (Bock, 2012; Laird, 2010). However, such approaches are complicated by the spatial interdependencies of individual CpG sites. Open questions include how data from single CpG sites can be aggregated into regions in order to capture higher-order methylation patterns across broader genomic regions. It is unclear whether such data can be analysed without the need of prior knowledge, *e.g.* the definition of genomic regions, location of CpG islands, promoters or other gene regulatory regions. Annotation of differentially methylated regions to nearby genes or known genomic features for further downstream interpretation of results is not straightforward and requires flexible and fast software solutions (Kondili et al., 2017).

1.3.2 Objectives

Based on the previously described shortcomings of currently available methods for analysis of data obtained from high-density arrays or NGS experiments, two of the three main objectives of this thesis focus on methodological improvements in miRNA quantification and the analysis of DNA methylation pattern in arbitrary genomic regions. Another objective aims at the disclosure of cellular and molecular origins of ERMS, using -omics datasets, applied bioinformatics and advanced methods from cell biology.

Ultimately, methodological advancements will help to interpret and understand array-based data or data from NGS experiments in Rhabdomyosarcoma and potentially lead to the disclosure of molecular pathways acting during tumour initiation and progression.

Reference-free and fine-grained analysis of data from miRNA-sequencing

The large number of miRNAs detected from previous NGS experiments across many species and experimental conditions represents a great data resource to guide detection and quantification of newly-created datasets. Sensitive homology searches against existing miRNA sequences might be a promising starting point for genome reference-free miRNA detection. A graph-based data structure with known miRNA sequences as nodes, and edges representing sequencing read(s) matching known miRNA sequences, might enable further fine-grained analyses: We postulate that closely connected components of the graph represent miRNA families across different *isomiR*s and species, and this allowing improved quantification on either family or *isomiR* levels by modulation of sensitivity of the homology search. Further, the sum of edge weights might be fed into differential expression analysis, readliy enabling downstream analysis.

Analysis of DNA methylation patterns in arbitrary genomic regions

Although a plethora of methods exist for data normalization and single-site CpG methylation analysis, a robust method that takes the spatial interdependencies of CpG sites into account and allows for unbiased genome-wide analysis is critically missing. Combination of differences of CpG site methylation estimated via one-sided two-sample Wilcoxon rank tests with a recently published method for grouping and correction of spatially correlated p-values (Pedersen, Schwartz, Yang, & Kechris, 2012) might represent a novel and powerful statistical approach to detect differentially methylated regions. Importantly, this approach would eliminate the need for *a priori* knowledge of regulatory regions and allow for evaluation of arbitrary genomic regions. Integration of numerous existing normalization techniques and automated downstream gene set enrichment methods might additionally enhance usability and open the method for a wide variety of applications.

The cellular and molecular origin of embryonic rhabdomyosarcoma

Existing studies delineating the cellular and molecular origin of Rhabdomyosarcoma did not provide direct evidence for involvement of the myogenic lineage in tumour formation, but rather based their (valuable) findings on correlation with myogenic properties. The Cre/lox site-specific recombination system has been developed to create time- and tissue-specific mutations, *e.g.* to study effects of inducible gene knockouts (Feil, Valtcheva, & Feil, 2009). We reasoned that employment of an inducible recombination system allowing permanent fluorescence labelling of muscle stem cell as well as their progeny might provide direct evidence of muscle stem cell tumourigenicity. Further, it would enable subsequent transcriptomic, genomic and epigenomic analysis of purified

tumour-propagating cells, as well as cancer subtype detection and investigation of clonal evolution in secondary recipients.

1.4 Thesis contributions

This thesis comprises three peer-reviewed publications, which are presented in chronological order.

Publication 1

Carsten Kuenne, Jens Preussner, Mario Herzog, Thomas Braun and Mario Looso. **MIRPIPE: quantification of microRNAs in niche model organisms**. *Bioinformatics*, Vol. 30 (2014)

The publication introduces MIRPIPE, an algorithm for rapid miRNA detection and quantification from NGS data. The algorithm takes raw data from *total RNA sequencing* as input and initially performs quality control by filtering out reads with low quality base calls. Reads that are too long or too short and do not match the length assumption of mature miRNAs (18 - 28nt) are eliminated, and sequencing adapter contamination from the *3'*-end is eradicated. MIRPIPE optionally removes unique, or low abundant reads frequently denoting remaining sequencing errors or miRNA variations below the detection limit and clusters reads sharing the same *5'*-end to properly handle miRNAs originating from the same gene but imperfectly digested by Drosha/Dicer (isoMiRs). MIRPIPE builds a graph from homology searches against a reference database with reference miRNAs as nodes and edges when read(s) support two reference miRNAs equally well. This unique approach permits inclusion of reads that otherwise cannot be matched uniquely and detects miRNA families as connected components in the graph. Quantification results, *i.e.* counts per miRNA family and miRNA cluster (isoMiRs) can readily be used for downstream differential expression analysis.

Publication 2

Jens Preussner, Julia Bayer, Carsten Kuenne and Mario Looso. **ADMIRE: analysis and visualization of differential methylation in genomic regions using the Infinium HumanMethylation450 Assay**. *Epigenetics and Chromatin*, Vol. 8 (2015)

The publication introduces ADMIRE, a pipeline for differential methylation analysis within genomic regions. The algorithm takes raw data from *Infinium HumanMethylation450 BeadChips* and initially filters single CpG probes based on low signal-to-noise ratios in a variable proportion of analysed samples. ADMIRE offers several techniques to perform between-sample normalization, *e.g.* by regressing out variability observed between control probes, before calculating two one-sided two-sample rank tests per CpG probe and between any two sample groups. Intentionally, two p-values are obtained per CpG probe, indicating lower or higher methylation in either group. ADMIRE allows subsequent combination of spatially correlated p-values into arbitrary genomic regions

using weighted Z-scores. It controls the familywise error rate as well to accomplish multiple testing correction. Differential methylation results are used to perform gene set enrichment analysis, when genomic regions can be meaningfully related to genes, thereby facilitating target selection in clinical settings. Further, most significantly altered regions are readily visualised and exported for use in genome browsers or in spreadsheet viewers.

Publication 3

Jens Preussner, Jiasheng Zhong, Krishnamoorthy Sreenivasan, Stefan Günther, Thomas Engleitner, Carsten Künne, Markus Glatzel, Roland Rad, Mario Looso, Thomas Braun and Johnny Kim. **A molecular subtype of cancer originating from adult stem cells during regeneration is driven by Dux transcription factors**. *Cell Stem Cell*, Vol. 23 (2018)

The publication investigates the cellular and molecular origin of rhabdomyosarcoma leveraging lineage tracing, genomic and transcriptomic analysis. Using the mdx mouse model, we show that genetic inactivation of the Tp53 tumour suppressor in Pax7-expressing muscle stem cells located in continuously regenerating skeletal muscles will give rise to fusion-negative RMS. The approach allowed tracing of the cellular origin of tumours back to transformed muscle stem cells. Genomic analysis of purified tumour-propagating cells reveals large genomic instability and identifies diverse genomic lesions, among them known oncogenes and cancer-promoting pathways. An amplicon was identified that contained a member of the Dux family of transcription factors, pointing to a novel candidate oncogene in RMS. Expression analysis of several published RMS studies and extended analysis of data from the Cancer Genome Atlas suggests that Dux-expressing tumours represent a distinct subtype of fusion-negative RMS and a novel pan-cancer subtype. Analysis of data from forced expression of Duxbl in muscle stem cells revealed that Duxbl can elicit epithelialization/colonization via a MET-like program to initiate tumour formation. Lastly, we investigate therapeutic intervention of Duxbl tumours via short hairpin mediated knockdown.

1.5 Results and discussion

This thesis provides methodological development in three areas of cancer bioinformatics, targeting detection and quantification of miRNAs, analysis of CpG-site methylation affecting epigenetic mechanism and characterization of the cellular and molecular basis of Rhabdomyosarcoma including detection of cancer subtypes.

1.5.1 Sensitive computational quantification of miRNA sequences from NGS sequencing

MiRNA sequences from MIRPIPE were validated with two complimentary approaches based on genomic mapping and found to be as sensitive as existing methods, recovering

84% and 96% of reference miRNAs respectively. Quantification results of MIRPIPE effectively recapitulated quantification of two gold-standard datasets with Spearman rank correlation values of 0.68 and 0.69, respectively. Specificity of MIRPIPE was higher compared to the approach based on genomic mapping, based on MIRPIPEs strategy to filter out lowly abundant reads prior to graph-based analysis. Characterization of detected miRNA sequences by delineation of putative miRNA target genes was pursued in a follow-up project termed LimiTT (Bayer et al., 2016) with contributions from the author of this thesis. Briefly, LimiTT integrates several databases of experimentally validated miRNA-target interactions (MTIs) and additionally allows utilisation of data from RNA expression experiments to weight important MTIs via built-in MTI set enrichment analysis.

The employment of a graph-based data structure for results from homology searches against a database of known miRNAs is a novel and unique approach allowing handling of miRNA sequences miRNA family and *isoMiR* levels, a feature that was previously missing. The approach is similar to current methods of transcript-level quantification in analysis of RNA-sequencing data termed *pseudoalignment* (Bray, Pimentel, Melsted, & Pachter, 2016; Patro, Duggal, Love, Irizarry, & Kingsford, 2017). Pseudoalignment does not require mapping to a genomic reference, but performs probabilistic assignment of sequencing reads to known transcripts, producing a list of compatible transcripts per sequencing read using matching of *k-mer* contents. Aggregation of so-called transcript-compatibility counts results in gene-level quantification, similar to MIRPIPEs summation of *isoMiR* counts to produce miRNA family level counts. Since miRNA reference databases might grow in the near future, adoption of *pseudoalignment* for miRNA quantification seems to be a good replacement for time-consuming homology searches and promises to speed-up runtime by several orders of magnitude.

MIRPIPE has also been used for detection and quantification of microRNAs in skeletal muscle development, differentiation and regeneration (Boettger et al., 2014), circadian regulation of gene expression (Dagenais-Bellefeuille, Beauchemin, & Morse, 2017), transmission of LNA antimiRs in newborn mouse pubs (Hönig et al., 2018) and in novel, plant-derived exosome-like ultrastructures (Xiao et al., 2018).

1.5.2 Computational analysis of DNA methylation in arbitrary genomic regions

Most of the existing methods for analysis of CpG methylation data only feature detection of differential methylation at individual CpG sites. Thus, such approaces are limited to pre-defined genomic regions, such as CpG islands or gene regulatory promotors. In contrast, the unique statistical approach implemented in ADMIRE permits combination of methylation data from CpG sites with arbitrary genomic regions, while considering their spatial correlation. The approach has been shown to gain sensitivity when dealing with small sample numbers or when DNA methylation is changed globally, *e.g.* as discussed for the CpG island methylator phenotype (see chapter 1.2.6). Two datasets were used to assess sensitivity and significance of results obtained from ADMIRE: Investigation of DNA methylation changes in a study of permanent atrial fibrillation (AF) showed high

sensitivity of ADMIRE, which identified 20 regions differentially methylated, although only 11 samples were used as input. Its direct competitor, RnBeads, reported one region with higher methylation in AF, which was not reported by ADMIRE. Furthermore, ADMIRE detected 14 additional regions up to 10 kB and subsequent gene set enrichment analysis confirmed results of previously conducted GWAS studies. A second dataset was used to analyse ADMIRE's performance in large sample cohorts. 689 samples from a study analysing DNA methylation as an intermediary of genetic risk in rheumatic arthritis (RA) were analysed. In addition ADMIRE detected differential methylation in the T-cell activation and T-cell receptor signalling pathway to RA, thereby confirming implication of the MHC region from the original study and proving its scalability and applicability in large clinical studies.

ADMIRE has additionally been used to detect epigenetic inactivation of Laminin A/C in a subset of neuroblastomas (Rauschert et al., 2017) and to identify relevant differentially methylated regions in pulmonary arterial hypertension (Hautefort et al., 2017).

1.5.3 Lineage-tracing reveals cellular origin of ERMS and enables in-depth analysis of cancer stem cells

Mice expressing the Cre recombinase ($Pax7^{CreERT2}$) in muscle stem cells were crossed to a strain carrying two lox-p sites in the Trp53 gene and the Rosa26::lsl Tomato allele, thereby enabling muscle stem cell specific inactivation of the tumour suppressor p53 (SC^{p53}) and permanent fluorescent lineage tracing of p53-deficient muscle stem cells by Tomato expression upon treatment with Tamoxifen (TAM). Mdx mice harbouring the inducible system exhibited tumour formation at sites of musculature extremities or the trunk after TAM administration. Lineage-traced tumours were histopathologically classified as embryonic Rhabdomyosarcoma and were stained positive for myogenic factors, clearly indicating their origin from the muscular lineage. TAM-treated wildtype or mdx mice never developed tumours and TAM-treated SC^{p53} mice only developed tumours upon consecutive bouts of Cardiotoxin-induced injury of the Tibalis anterior muscle, demonstrating that muscle stem cell-specific loss of p53 in a regenerative environment is sufficient to generate RMS. Lineage-tracing enabled separation and purification of RMS cells into non-lineage-traced and lineage-traced tumour propagating cells (TPCs) using fluorescence-activated cell sorting (FACS). Importantly, transplantation of lineage-traced p53 deficient muscle stem cells into immunocompromised mdx-nude mice generated tumours already two weeks after injection. These data confirmed the hypothesis put forward by the stem cell theory introduced earlier (see chapter 1.2.3) and disclosed the cellular origin of embryonal RMS in the $p53^{-/-}$/mdx model.

1.5.4 Computational analysis of copy number variation reveals molecular origin of ERMS

Whole-exome DNA sequencing of purified TPCs and matched normal samples was followed by subsequent genome analysis to identify tumour-associated mutations. In 20

out of 21 specimen, discrete and dramatic copy number amplifications were identified as the prevailing mutations. Positional mapping revealed defined chromosomal regions harbouring known mutational targets in ERMS, including Yap1 (Tremblay et al., 2014), C-met (Taulli et al., 2006), Jun (Durbin et al., 2009), and Cdk4/Gli1/Os9 (Liu et al., 2014). Interestingly, TPCs did not accumulate somatic single-nucleotide variations, indicating that EMRS does not follow the classic mechanism proposed by the somatic mutation theory (see chapter 1.2.1) but are predominantly characterised by copy number changes.

It is widely accepted that overexpression of oncogenes or loss of tumour suppressor genes is a crucial molecular event resulting in tumour initiation, but it is unclear whether maintenance of tumourigenicity depends on the transforming molecular event as well. The phenomenon of *oncogene addiction* (*i.e.* the physiological dependence of cancer cells on oncogenes, (Weinstein (2002))) has been described for several cancer types and offers opportunities for therapeutic intervention by targeting oncogene expression with specific drugs. In fact, knockdown of Yap1 in Yap1-expressing TPCs using short hairpin RNA (shRNA) resulted in cell death, indicating the dependence of TPCs on distinct regulatory networks facilitated by Yap1 expression. However, such intervention requires personalized therapeutic approaches often not yet implemented in clinical settings. Additionally, the cancer phenotype might not be reversed by blocking expression of an oncogene, if oncogene-mediated genome instability induced subsequent mutations enable cells to escape oncogene dependence.

1.5.5 Integrative analysis of zygotic Dux factors defines a new cancer subtype

Several mice displayed amplification of a poorly described locus without any known oncogene on chromosome 14qA3. Analysis of genomic synteny (*i.e.* the physical co-localization of genetic loci) between different species revealed that Duxbl is located in synteny with human DuxB, a member of the Dux family of homeobox-containing transcription factors. Interestingly, DuxB and its paralog, DuxA, were recently shown to be expressed exclusively at the totipotent 8-cell stage in early zygotes (Madissoon et al., 2016). Furthermore, the founding member of the Dux transcription factor family, Dux4, and its murine homolog Dux, are responsible for driving cleavage-stage gene expression known as zygotic gene activation (ZGA) in totipotent embryonic stem cells (Hendrickson et al., 2017; Leidenroth & Hewitt, 2010; Whiddon, Langford, Wong, Zhong, & Tapscott, 2017). Those findings led to the speculation that Dux transcription factors might act at a putative interface of stem cell potency and tumour formation. So far, a more detailed analysis was difficult until recent technological advancements allowed experimental assessment of very little input material, as in the case of early zygotes.

To test whether Dux-driven activation of zygotic genes plays a role in human ERMS, an integrative analysis was designed to translate findings from whole-exome sequencing in mouse to transcriptome sequencing in human cancer patients. Intriguingly, 54 tumours (~10%), which expressed Dux4, DuxA or DuxB, showed a cleavage-stage-specific expression signature in a previously published discovery cohorts of human ERMS pa-

tients (Chen et al., 2013; Davicioni et al., 2009; Williamson et al., 2010). To test whether increased expression of Dux genes is restricted to ERMS or is associated with other malignancies, expression data from ~10,000 cancer patients from The Cancer Genome Atlas (Hoadley et al., 2014) was used for further molecular analysis. Interestingly, 349 patients displayed distinct expression of Dux family members either in combination or alone. The onset of cancer and the type of cancer was highly variable in these patients, suggesting that Dux transcription factors define a molecular subtype of a broad range of human cancers, including ERMS.

1.5.6 Gene expression analysis reveals epigenetic plasticity conferred by tumourigenic Duxbl

The selection of suitable targets for individual tumour therapeutics critically depends on molecular insight into mechanisms of tumour initiation and progression. To gain understanding on the action of Dux transcription factors, Duxbl was overexpressed in wild-type muscle stem cells *in vitro*, which resulted in the emergence of immortalised and morphologically rounded clones prone to spontaneously form epithelial-like spherical aggregates. Subcutaneous transplantation of clones formed neoplasia at the site of engraftment and clearly demonstrate that overexpression of Duxbl can transform muscle stem cells and elicit excessive growth *in vivo*. Interestingly, transformed cells contributed to myofiber formation when injected directly into the strong pro-differentiation environment of the *tibialis anterior* muscle, further supporting the suppressive role of functional tissue in cancer progression (Bissell & Radisky (2001), chapter 1.2.2). Expression analysis of isolated clones revealed upregulation of the histone lysine demethylase Kdm4d but no expression of myogenic determinants like Myf5, MyoD and MyoG, suggesting a lineage independent mechanism of cell transformation. Instead, dramatic induction of genes involved in epithelial cell proliferation and coding for integrins, collagens, cadherins and proto-cadherins was observed, along with expression of pluripotency factors Sox2 and Klf4. These genes are instrumental to facilitate mesenchymal-to-epithelial transition during reprogramming of somatic cells to induce pluripotent stem cells (Li et al., 2010). Taken together, the overexpression of Duxbl confers cellular plasticity through Kdm4-mediated permissive chromatin (Labbé, Holowatyj, & Yang, 2013), which allows induction of a MET-like transition that initiates growth of tumourigenic colonies. Most likely, the establishment of truly metastatic niches for tumour outgrowth requires a secondary oncogenic event but might not depend on sustained expression of Duxbl. In such a scenario, therapeutic supression of tumourigenicity conferred by DuxB/Duxbl via an epigenetic hit-and-run (Saunderson et al., 2017) event would be unable to take advantage of classical oncogene addiction, but would require novel therapeutic ideas.

1.6 Conclusion

The discovery of dynamics and complexity of cancer-causing mechanisms, including dysregulated expression of regulatory factors (miRNAs, transcription factors), structural

genomic and epigenomic alterations or disruption of normal tissue context, has benefited largely from technological and computational advancements in the last decade. Nowadays, cancer research faces unique challenges to manage, analyse and integrate complex and numerous datasets to identify relevant therapeutic targets and support precise treatment of cancer patients. On one hand, work presented in thesis seeks to provide missing methodological development within the field of computational analysis of cancer -omics data. On the other hand, work presented here investigates the molecular and cellular origin of a childhood cancer, embryonal rhabdomyosarcoma, and integrates data from genomics and transcriptomics with previously published findings to detect a novel cancer subtype.

Two algorithms have been contributed to enable characterisation and quantification of the entire regulatory repertoire of miRNAs from NGS experiments, and to facilitate analysis of DNA methylation. Importantly, the unique statistical approach applied during DNA methylation analysis overcomes previous limitations and enables differential analysis within arbitrary genomic regions in a case-control setting. Both algorithms are designed to work with all types of input from RNA sequencing or high-density DNA methylation arrays, respectively, and might be beneficial for future studies investigating the molecular origin of different types of cancer. By employment of genetic labelling and direct oncogenic transformation of muscle stem cells, work in this thesis unambiguously demonstrated that the cellular origin of ERMS lies within muscle stem cells. Further, analysis of data from genomics and transcriptomics in mouse was integrated with gene expression studies from ~10,000 human cancer patients to disclose the molecular origin of ERMS and discover a novel cancer subtype across a broad range of human cancer driven by oncogenic activation of zygotic Dux factors. To better understand how zygotic Dux factors confer tumour initiation, the thesis additionally provides evidence for Duxbl-conferred epigenetic plasticity and cellular transformation. Although an additional oncogenic hit is likely required for tumour metastasis, insights into tumour initiation is a useful starting point for selection of relevant therapeutic targets.

Taken together, findings and conclusions from this thesis allow future research in the areas of molecular biology, computational biology and precision medicine. For example, researching the role of Dux transcription factors for zygotic gene activation and cancer initiation has only started, leaving the definition of their gene regulatory network and their role in epigenetic restructuring for further investigation. Effective precision medicine for cancer treatment currently lacks suitable (computational) methods for data integration, which include different mechanisms of cancer initiation and the dynamics of cancer progression. Future research is needed to properly evaluate and select tailored treatment strategies based on identified cancer subtype properties, drug susceptibility, presence of neo-antigens, synthetic dosage lethality or unique tumour microenvironments. Technological advancement has just begun to enable integrative studies that are able to produce results for functional testing. So far, holistic analyses are not yet feasible in clinical routines, but might become available in a few years. Towards this end, the current thesis has contributed a certain share.

2 Publications

2.1 MIRPIPE: quantification of microRNAs in niche model organisms

Carsten Kuenne[1,‡], Jens Preussner[1,‡], Mario Herzog[1], Thomas Braun[2] and Mario Looso[1]

Affiliations	[1] Group of Bioinformatics and [2] Department of Cardiac Development and Remodeling Max Planck Institute for Heart and Lung Research ‡ These authors contributed equally
Journal	Bioinformatics, *Oxford Journals*
Date, Issue	2014, Vol. 30 no. 23
Pages	3412 - 3413
DOI	10.1093/bioinformatics/btu573
Supplementary data	available online

Contributions

The following contributions are attributed to the thesis autor:

Conceptualization	Contributed to definition of project goals
Methodology	Formulation of the graph-based isoMiR handling; Formulation of the benchmark tests
Investigation	Contributed to interpretation and discussion of results
Validation	Benchmarking reproducibility and predictive efficiency
Software	Contributed to overall workflow implementation; Implemented the graph-based isoMiR handling
Visualization	Created Figure 1 and Supplementary File S2
Resources	Setup and administration of MIRPIPEs web service; Contributed to Supplementary File S1 (software manual)
Writing	Contributed to manuscript draft; Review and editing of the manuscript

BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 23 2014, pages 3412–3413
doi:10.1093/bioinformatics/btu573

Sequence analysis

Advance Access publication August 26, 2014

MIRPIPE: quantification of microRNAs in niche model organisms

Carsten Kuenne[1,†], Jens Preussner[1,†], Mario Herzog[1], Thomas Braun[2] and Mario Looso[1,*]

[1]Group of Bioinformatics and [2]Cardiac Development and Remodelling, Max Planck Institute for Heart and Lung Research, Ludwigstrasse 43, D-61231 Bad Nauheim, Germany

Associate Editor: John Hancock

ABSTRACT

Summary: MicroRNAs (miRNAs) represent an important class of small non-coding RNAs regulating gene expression in eukaryotes. Present algorithms typically rely on genomic data to identify miRNAs and require extensive installation procedures. Niche model organisms lacking genomic sequences cannot be analyzed by such tools. Here we introduce the MIRPIPE application enabling rapid and simple browser-based miRNA homology detection and quantification. MIRPIPE features automatic trimming of raw RNA-Seq reads originating from various sequencing instruments, processing of isomiRs and quantification of detected miRNAs versus public- or user-uploaded reference databases.

Availability and implementation: The Web service is freely available at http://bioinformatics.mpi-bn.mpg.de. MIRPIPE was implemented in Perl and integrated into Galaxy. An offline version for local execution is also available from our Web site.

Contact: Mario.Looso@mpi-bn.mpg.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 24, 2014; revised on July 1, 2014; accepted on July 3, 2014

1 INTRODUCTION

MicroRNAs (miRNAs) are ~22 nucleotides long and belong to the class of snRNAs. miRNAs serve numerous roles in downregulation (transcript degradation and sequestering, translational suppression) of gene expression. In general, miRNAs are assumed to regulate multiple targets although effects on most targets are relatively mild (Ameres and Zamore, 2013). Isoforms of miRNAs resulting from imperfect digestion by Drosha and Dicer or RNA editing by specialized enzymes represent a challenge during the determination of correct read counts following RNASeq. miRNA variants might be 'silent' (3′ modification = isomiR) or target different mRNAs when changes occur in the 5′ regions responsible for complementary binding. Sequence differences between taxa hamper quantification, especially if no genomic or miRNA data for the studied organism are available as in the case of niche model organisms. Sequencing errors can further complicate the identification of miRNAs. These effects should ideally be addressed on multiple levels, including (i) isomiR handling, (ii) enforcement of a minimum read copy number, (iii) clustering of similar miRNAs, (iv) removal of relatively low abundance reads and (v) optional fallback to the miRNA family level. A set of applications in the field attempts to cover these features, but a Web-based tool able to unify all functionalities that can be applied to any organism is critically missing (An *et al.*, 2013; Giurato *et al.*, 2013; Wen *et al.*, 2012).

2 WORKFLOW AND FEATURES

MIRPIPE uses open-source binary tools including the FASTX-Toolkit (Pearson *et al.*, 1997), Cutadapt (Martin, 2011) and BLASTN (Boratyn *et al.*, 2013) for data processing. The pipeline was integrated into a Galaxy-based Web platform (Goecks *et al.*, 2010) but is also available for download and local execution. A detailed explanation of the algorithm can be found in Supplementary File S1.

The workflow starts with the upload of a compressed FASTQ/FASTA read file using the Web interface or the MIRPIPE FTP server. MIRPIPE can fully process raw reads originating from Illumina, 454, IonTorrent or Sanger sequencing instruments including adapter trimming. A reference FASTA database bearing mature target miRNAs can either be selected from current miRBase release (Griffiths-Jones *et al.*, 2006) or can be uploaded by the user.

The raw reads are processed to optionally remove an adapter sequence and trim for a minimum quality (default Q20). Only reads of the desired size range are selected to limit the pool to mature miRNAs. Duplicate reads are collapsed to decrease the number of necessary homology searches, and only those sequences represented by a minimum count are kept for further analyses. This measure is intended to remove unique reads, which frequently denote sequencing errors or miRNA variations that are expressed near to the detection limit, preventing reliable quantification.

Read counts from isomiRs of the same miRNA are combined. These isomiR read sequences may only differ by the 3′ end and are thus putatively encoded by the same gene. Only one nucleotide may differ between two sequences to be counted as isoforms of the same miRNA, and only the longest sequence is used in the next step to further reduce the amount of homology searches.

The remaining read sequences are used for a sequence similarity search versus the chosen reference database of miRNAs.

Mature reference miRNAs and their precursors are optionally collated by name on the family level to remove redundancy

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

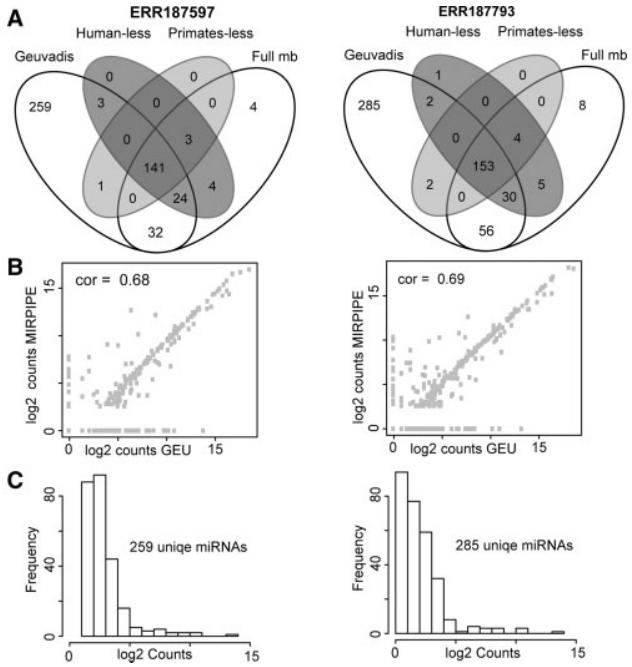


Fig. 1. **A**) Comparison of MIRPIPE prediction on two gold standard (GS) datasets using full miRBase and reduced miRBase as reference set. (**B**) Spearman correlation of absolute counts of GS and MIRPIPE. (**C**) The large number of GS-specific miRNA identifications is caused by low counts, filtered out by MIRPIPE default parameters

introduced by organism prefixes and precursor suffixes (e.g. bta-miR-200a, oan-miR-200a-3p > miR-200a).

For each read, the detected reference miRNA families are scored based on the minimum number of mismatches. If a read matched equally well versus multiple miRNA families, the respective families are joined by single linkage clustering. This permits the inclusion of reads that cannot be matched uniquely, as well as the exact measurement of the fraction of ambiguously matching reads and thereby the reliability of the match. By default, only those read sequences that are at least 5% as abundant as the most abundant sequence per miRNA family cluster are denoted to reduce the impact of sequencing errors and increase robustness.

Counts per miRNA family and cluster are presented for download. Currently, MIRPIPE can complete a job within 0.5–2 h, depending on the file size and the selected reference database. MIRPIPE quantification results can be directly used for differential expression analysis using other tools on our Web site (Supplementary File S1).

3 BENCHMARK

To demonstrate congruent results for MIRPIPE, we compared the results with an miRNA analysis based on a genomic mapping of Illumina HiSeq reads (Lawless *et al.*, 2013). We identified 96% of the published miRNAs (Supplementary File S2). Furthermore, we compared our tool with a similar approach without the need for a genome sequence by analyzing a public dataset (Zhang *et al.*, 2013) with the CLC Genomics Workbench. In this case, 84% of the miRNAs were identical (Supplementary File S2).

Finally, we checked the predictive efficiency of our tool for niche models based on a human RNA-Seq dataset (Lappalainen *et al.*, 2013). Here, we performed MIRPIPE versus a reference database bearing (i) the complete miRBase, (ii) miRBase excluding human miRNAs and (iii) miRBase excluding miRNAs of all primates. The absence of closely related reference sequences resulted in only a marginal loss of sensitivity for MIRPIPE, indicating its aptitude for the analysis of niche model organisms (Fig. 1, Supplementary File S2).

Funding: Excellence Cluster Cardio-Pulmonary System (ECCPS); MPI for Heart and Lung Research.

Conflict of interest: none declared.

REFERENCES

Ameres,S.L. and Zamore,P.D. (2013) Diversifying microRNA sequence and function. *Nat. Rev. Mol. Cell Biol.*, **14**, 475–488.

An,J. *et al.* (2013) miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. *Nucleic Acids Res.*, **41**, 727–737.

Boratyn,G.M. *et al.* (2013) BLAST: a more efficient report with usability improvements. *Nucleic Acids Res.*, **41**, W29–W33.

Giurato,G. *et al.* (2013) iMir: an integrated pipeline for high-throughput analysis of small non-coding RNA data obtained by smallRNA-Seq. *BMC Bioinformatics*, **14**, 362.

Goecks,J. *et al.* (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible: transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.

Griffiths-Jones,S. *et al.* (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.

Lappalainen,T. *et al.* (2013) Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, **501**, 506–511.

Lawless,N. *et al.* (2013) Next generation sequencing reveals the expression of a unique miRNA profile in response to a gram-positive bacterial infection. *PLoS One*, **8**, e57543.

Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.*, **17**, 10–12.

Pearson,W.R. *et al.* (1997) Comparison of DNA sequences with protein sequences. *Genomics*, **46**, 24–36.

Wen,M. *et al.* (2012) miREvo: an integrative microRNA evolutionary analysis platform for next-generation sequencing experiments. *BMC Bioinformatics*, **13**, 140.

Zhang,Z. *et al.* (2013) High-efficiency RNA cloning enables accurate quantification of miRNA expression by deep sequencing. *Genome Biol.*, **14**, R109.

2.2 ADMIRE: analysis and visualization of differential methylation in genomic regions using the Infinium HumanMethylation450 Assay

Jens Preussner[1], Julia Bayer[1], Carsten Kuenne[1] and Mario Looso[1]

Affiliations	[1] Group of Bioinformatics Max Planck Institute for Heart and Lung Research
Journal	Epigenetics and Chromatin, *BioMed Central*
Date, Issue	2015, Vol. 8 no. 1
Pages	1 - 10
DOI	10.1186/s13072-015-0045-1
Supplementary data	available online

Contributions

The following contributions are attributed to the thesis autor:

Conceptualization	Overall definition of project goals; Definition of visualizations
Methodology	Design of the software; Formulation of ADMIREs differential testing procedure; Formulation of the benchmark tests
Investigation	Conducted analysis of the RA and AF datasets; Contributed to interpretation and discussion of results
Validation	Benchmarking ADMIRE against RnBeads; Evaluation of ADMIREs performance
Software	Implementation of the ADMIRE workflow
Visualization	Creation of Fig. 1 to Fig. 4, Additional files 4 to 8
Resources	Setup and administration of ADMIREs web service; Wrote the manual
Writing	Contributed to the manuscript draft; Review and editing of the manuscript

Preussner *et al. Epigenetics & Chromatin (2015) 8:51*
DOI 10.1186/s13072-015-0045-1

Epigenetics & Chromatin

METHODOLOGY Open Access

ADMIRE: analysis and visualization of differential methylation in genomic regions using the Infinium HumanMethylation450 Assay

Jens Preussner, Julia Bayer, Carsten Kuenne and Mario Looso*

Abstract

Background: DNA methylation at cytosine nucleotides constitutes epigenetic gene regulation impacting cellular development and a wide range of diseases. Cytosine bases of the DNA are converted to 5-methylcytosine by the methyltransferase enzyme, acting as a reversible regulator of gene expression. Due to its outstanding importance in the epigenetic field, a number of lab techniques were developed to interrogate DNA methylation on a global range. Besides whole-genome bisulfite sequencing, the Infinium HumanMethylation450 Assay represents a versatile and cost-effective tool to investigate genome-wide changes of methylation patterns.

Results: Analysis of DNA Methylation In genomic REgions (ADMIRE) is an open source, semi-automatic analysis pipeline and visualization tool for Infinium HumanMethylation450 Assays with a special focus on ease of use. It features flexible experimental settings, quality control, automatic filtering, normalization, multiple testing, and differential analyses on arbitrary genomic regions. Publication-ready graphics, genome browser tracks, and table outputs include summary data and statistics, permitting instant comparison of methylation profiles between sample groups and the exploration of methylation patterns along the whole genome. ADMIREs statistical approach permits simultaneous large-scale analyses of hundreds of assays with little impact on algorithm runtimes.

Conclusions: The web-based version of ADMIRE provides a simple interface to researchers with limited programming skills, whereas the offline version is suitable for integration into custom pipelines. ADMIRE may be used via our freely available web service at https://bioinformatics.mpi-bn.mpg.de without any limitations concerning the size of a project. An offline version for local execution is available from our website or GitHub (https://github.molgen.mpg.de/loosolab/admire).

Background

Several epigenetic mechanisms control gene expression in cells [1]. One of these conserved mechanisms is DNA methylation, a process where cytosine bases of DNA are converted to 5-methylcytosine by the DNA methyltransferase (DNMT) enzymes. DNA methylation by these enzymes is a reversible regulator of gene expression. Methylated cytosine recruits proteins which are involved in gene repression and inhibit the binding of transcription factors. The pattern of DNA methylation in the genome undergoes changes during development and plays a role in a range of diseases, utilizing processes of de novo methylation and demethylation. In case of development and differentiation, differentiated cells display a stable, cell-type-specific methylation pattern, permanently switching off the expression of genes that are not essential for the respective cell type.

A number of lab techniques were developed to interrogate DNA methylation including whole-genome bisulfite sequencing (WGBS) and Infinium HumanMethylation450 Assays [2]. Although WGBS provides a comprehensive genome-wide coverage (around 28 million CpGs in humans), it is associated with relatively high costs for re-sequencing the whole genome. A similar method known as reduced representation bisulfite sequencing (RRBS) is intended to overcome this problem

*Correspondence: mario.looso@mpi-bn.mpg.de
Bioinformatics Group, Max Planck Institute for Heart and Lung Research, Ludwigstrasse 43, 61231 Bad Nauheim, Germany

by sequencing just DNA fragments enclosing at least one CpG site. While Infinium HumanMethylation450 Assays reveal a less comprehensive picture compared to sequencing-based methods (approximately 0.5 million CpGs are addressed), economical factors render them highly attractive for epigenome-wide association studies (EWAS) involving up to thousands of individual samples [3] and represent an effective tool to identify biomarkers of disease states and progression [4].

Although Infinium HumanMethylation450 Assays are widely used, just very recently a cohort of noncommercial analysis pipelines was introduced. However, most of these tools are designed as command line tools. This is frequently accompanied with complex usage requirements which pose a significant challenge to researchers with limited programming skills. Furthermore, the genome-wide visualization of methylation sites, the visualization of significantly differentially methylated sites and downstream analyses have not been addressed optimally, yet. Here we introduce ADMIRE, an easy to use web-based tool intended to simplify usage inside a comprehensive application accessible by web interface as well as programmatically. ADMIRE generates publication-ready graphical overviews of differentially methylated loci and genome-wide overview tracks (Additional file 1) including advanced statistical methods to increase sensitivity. An included gene set enrichment analysis provides an overview on the entities that might link the significant sites.

Results

Comparison to existing software

Very recently, a cohort of noncommercial analysis pipelines was introduced and a current selection of widely used packages is reviewed in [5]. While the total number of tools intended to perform at least individual steps of HumanMethylation450 assay analysis is estimated to be around 20, only a minority is accessible via a graphical user interface and often limited to specific operating systems. A detailed comparison of tool features is listed in Additional file 2. An easy to use web-based application is only provided by RnBeads [6], although this might be the best way for biologists with limited programming skills to access an analysis pipeline. In contrast to RnBeads (restricted to 24 arrays), the web-based version of ADMIRE does not restrict the number of input arrays and was tested with a sample set of 689 arrays from a GEO dataset described below. Additionally, since calculation of per-probe test statistics is the main computational task (see algorithm description below), the runtime of ADMIRE is virtually independent of the number of input arrays. While most of the available tools provide functions for probe filtering and normalization, only a small number include functionality to create scalable visualizations or to detect differentially methylated positions and regions simultaneously. Furthermore, regions of interest are often pre-calculated and only a small number of tools allow statistics on individual regions of interest that can be provided by the user. Finally, none of the available tools provides a downstream analysis that is able to discover the linkage of differentially methylated genes. In order to generate a tool that combines all these critical features, we developed ADMIRE, a web-based tool for users without any computational background.

ADMIREs calculation of test statistics

ADMIRE features five different normalization methods (see [7]) but can also work on raw methylation values. The pipeline performs two one-sided two-sample rank tests (Mann–Whitney *U* tests) based on the sample_group information provided. In contrast to the *t* test, the Mann–Whitney *U* test does not require normally distributed data. The one-sided two-sample tests are performed per Illumina probe on the array and between pairs of sample groups. Intentionally, two *p* values are obtained for each probe, indicating a higher probe methylation in a distinct group and allowing the subsequent combination of multiple single *p* values from within a genomic region of interest (tiles, promotors and the like) as suggested in [8]. The spatially correlated *p* values are combined with genomic regions by mapping probe specific *p* values onto pre-calculated or user-defined genomic regions, indicating no change or a higher methylation in either sample group. To create a *p* value for an entire region, the Stouffer–Liptak correction implemented in [9] is used. A 1-step Sidak correction for multiple testing is applied to obtain q-values (see [9]). In order to filter significantly differentially methylated regions, a user-defined q-value threshold is used.

The web-based analysis platform

The ADMIRE analysis platform is implemented as a web-based application (Fig. 1) and enables users with limited bioinformatics background to apply sophisticated methylation analysis. The web-based platform allows user accounts with the possibility to keep raw files and analyzed data in a workspace of unlimited size. The default output of a scanner system compatible to Illumina HumanMethylation450 Assay consists of a SampleSheet.csv file and a file directory named after the assays Sentrix-ID containing two compressed *.idat-files per sample. These raw files are supported by ADMIRE. Besides the original SampleSheet.csv, ADMIRE is also able to process a tab-separated sample definition file (see user manual, Additional file 3).

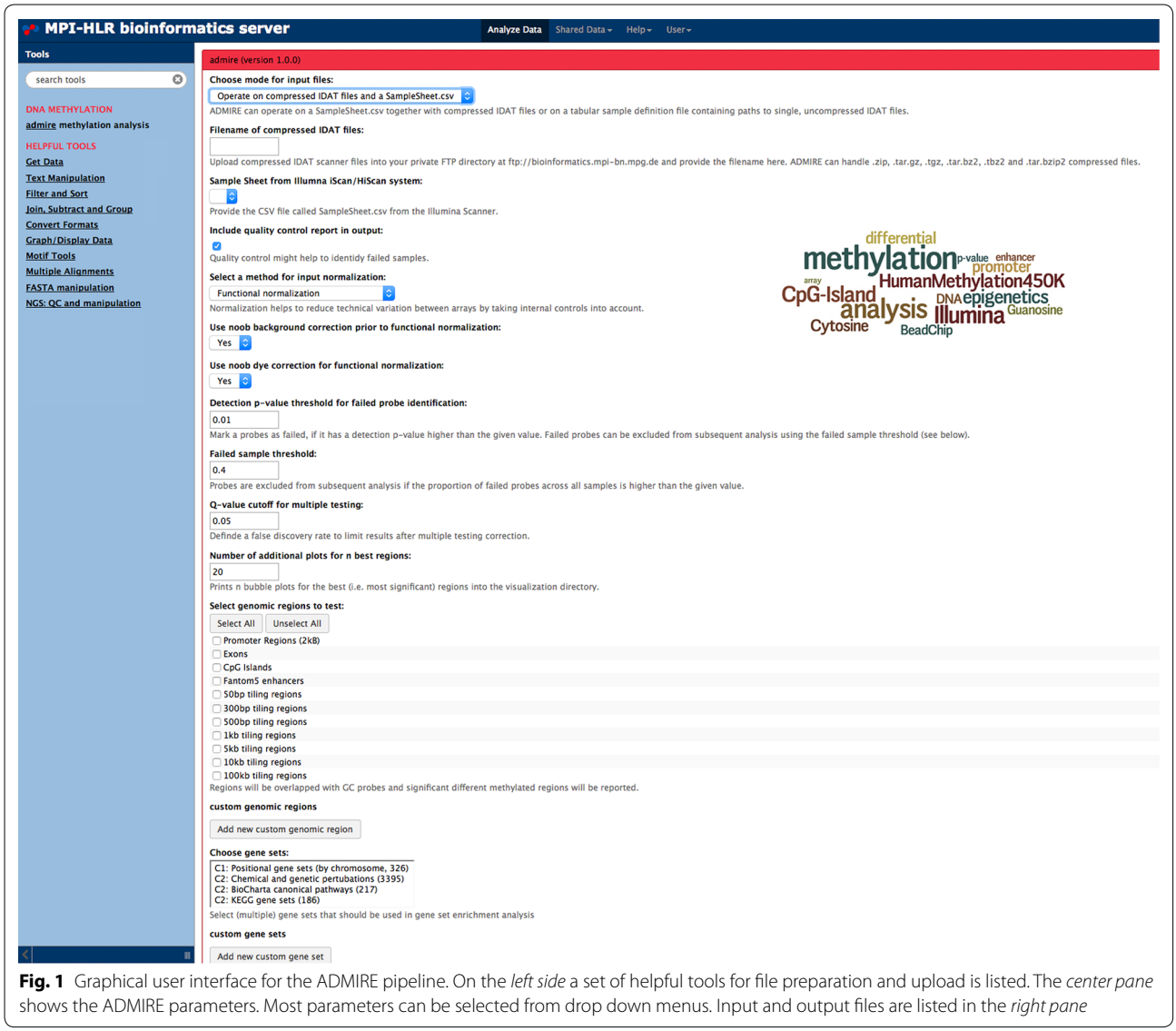


Fig. 1 Graphical user interface for the ADMIRE pipeline. On the *left side* a set of helpful tools for file preparation and upload is listed. The *center pane* shows the ADMIRE parameters. Most parameters can be selected from drop down menus. Input and output files are listed in the *right pane*

The settings file defines the groups that should be used for statistical testing. An all-vs-all comparison is performed with no limitation on the number of sample groups. Next, a wide range of analysis parameters can be adjusted, such as normalization method (SWAN, Functional, Quantile, Noob or Illumina), quality control filtering based on detection *p* values, failed sample threshold, Q-value cutoff for multiple testing as well as genomic regions for testing. A set of pre-calculated genomic regions are provided such as genome-wide tilings, annotations based on Gencode [10], as well as CpG islands and Fantom5 enhancers [11]. Furthermore, custom regions of interest can be uploaded to combine probes. To generate high-resolution graphics of differentially methylated regions, a numeric parameter is available to choose the number of graphics that will be generated from the most significantly altered regions. If the user is interested in a downstream analysis of differentially regulated regions, a gene set enrichment analysis can be performed on a selection of pre-defined gene sets [12] including chromosomal locations, pathways, diseases, and GO-terms. In addition to pre-defined sets, custom gene sets can be provided.

Workflow

Once the analysis is started, ADMIRE evaluates the sample definition file and prints out an error message in case files are missing or cannot be read. The raw files are preprocessed and filtered by the functions from the R package *minfi* [7], according to the parameters set. Aggregated

data is used to generate a quality control report in PDF format and normalized beta and m values are provided as tabular data (Fig. 2, step 1). In accordance to the groups defined earlier, all-vs-all pairwise comparisons of per-probe methylation are performed automatically. To call the significant differences in terms of methylation, ADMIRE performs statistical tests as described in the section above (Fig. 2, step 2).

Next, spatially correlated *p* values are combined with respect to the genomic regions defined by the user [9]. The generated result list includes all genomic regions, sorted by significance of methylation changes between the groups specified and the min/max/median change of methylation rate is calculated for further filtering (Fig. 2 step 3). For the most significant differentially methylated regions, a high-resolution image is generated (see Additional file 1). Finally, all results are transformed into BED format data tracks to allow visualization of differentially methylated regions in commonly used genome viewers such as IGV [13] or UCSC [14] (Fig. 2, step 4). Additionally, the output includes comma-separated tables that can be used to filter for specific genes, genomic locations, coverage, min/max/median change, *p* values, and/or *q* values. Details on the output files can be found in the methods section and in Additional file 3. Given that regions with a direct link to genes (indicated by a *gene_name* property) were chosen as regions of interest, a gene set enrichment analysis can be performed [12]. The enrichment analysis calculates an enrichment score (ES) for each gene set, depending on the ranks and differences in methylation of genes that are members of the gene set. In combination with graphs for enrichment score calculations, it can be inferred whether higher methylation in controls or cases contributed most to the enrichment of the gene set. Additionally, a heat map graphically represents a leading edge analysis that allows the detection of gene sets with a high overlap of core genes that mainly affect the ES (Fig. 2, step 5). All results listed above are generated in the workspace and can be downloaded as individual files or as a compressed archive from the web-based platform.

Performance evaluation and comparison to the existing gold standard

To demonstrate the ease of use, the robustness and applicability of ADMIRE, we downloaded 689 HumanMethylation450 Assay samples from a study analyzing DNA methylation as an intermediary of genetic risk in rheumatoid arthritis (GEO GSE42861) [15]. ADMIRE was invoked from the web interface using a custom sample-definition file (see "Methods") with default parameters. We selected all 2-kB promoter regions and chose positional gene sets as input for the enrichment analysis. Since the runtime of ADMIRE is virtually independent of input size, the results were obtained after 24 h with a maximum memory usage of 65 GB RAM. As the analysis in [15] was performed on single methylation sites and we did not intent to replicate the analysis, validation was done via an unbiased gene set enrichment analysis using positional gene sets as input. We identified the constant (*TRAC*) and variable (*TRAV*/*TRAJ*) segments of the T-cell receptor alpha chain on chr14q11 locus as higher methylated in arthritis patients. Additionally, four known members of the T-cell receptor signaling pathway, *CD28*, *CD3G*, *CD3D* as well as *PDCD1*, were found to be higher methylated in patients (Fig. 3).

In order to compare ADMIRE to RnBeads, the current gold standard for HumanMethylation450 Assay analysis, we used an additional dataset of smaller size since the RnBeads [16] web interface is restricted to 24 samples. Our test dataset contains 11 samples from a study analyzing permanent atrial fibrillation (GEO GSE62727). This dataset was analyzed by RnBeads using default parameters (5-kB pre-calculated tiling regions) as well as the ADMIRE pipeline. To match the output from RnBeads and enable a direct comparison, we selected all 5-kB tiling regions as input for ADMIRE (see "Methods"). Our tool found twenty 5-kB regions corresponding to protein coding genes to be higher methylated in fibrillating atria (see Additional file 4) with a median methylation change of up to 12 %. Next, we carried out a second run with ADMIRE using 10-kB tiling regions as input to test for reproducibility of statistically significantly changed regions. Besides nine genes present in both result files, another 14 genes were identified from 10-kB regions only, with a median methylation change up to 45 % (see Additional file 5). RnBeads identified only one region to be higher methylated in fibrillating atria. This genomic location was not reported by ADMIRE. Some representative significant regions found by ADMIRE and the single region found by RnBeads are shown in Fig. 4a–f. We chose an indirect way to evaluate specificity and significance of regions reported by ADMIRE but not by RnBeads. To evaluate the latter, we visualized the homogeneity of the methylation change over all 5-kB tiling regions detected by ADMIRE in Fig. 4g. The boxplots represent all single methylation sites, combined in accordance to the tiling region. Their level and spread present a global overview in order to investigate the magnitude of the methylation changes. The user can interpret this information to select an appropriate threshold. To evaluate the specificity of our findings, we performed a functional analysis. This showed an enrichment of transcriptional regulation, driven by transcription factors such as HOX A, TBX5, and PITX2 (Additional file 6). This is remarkable, as initial GWAS studies identified a major risk region where

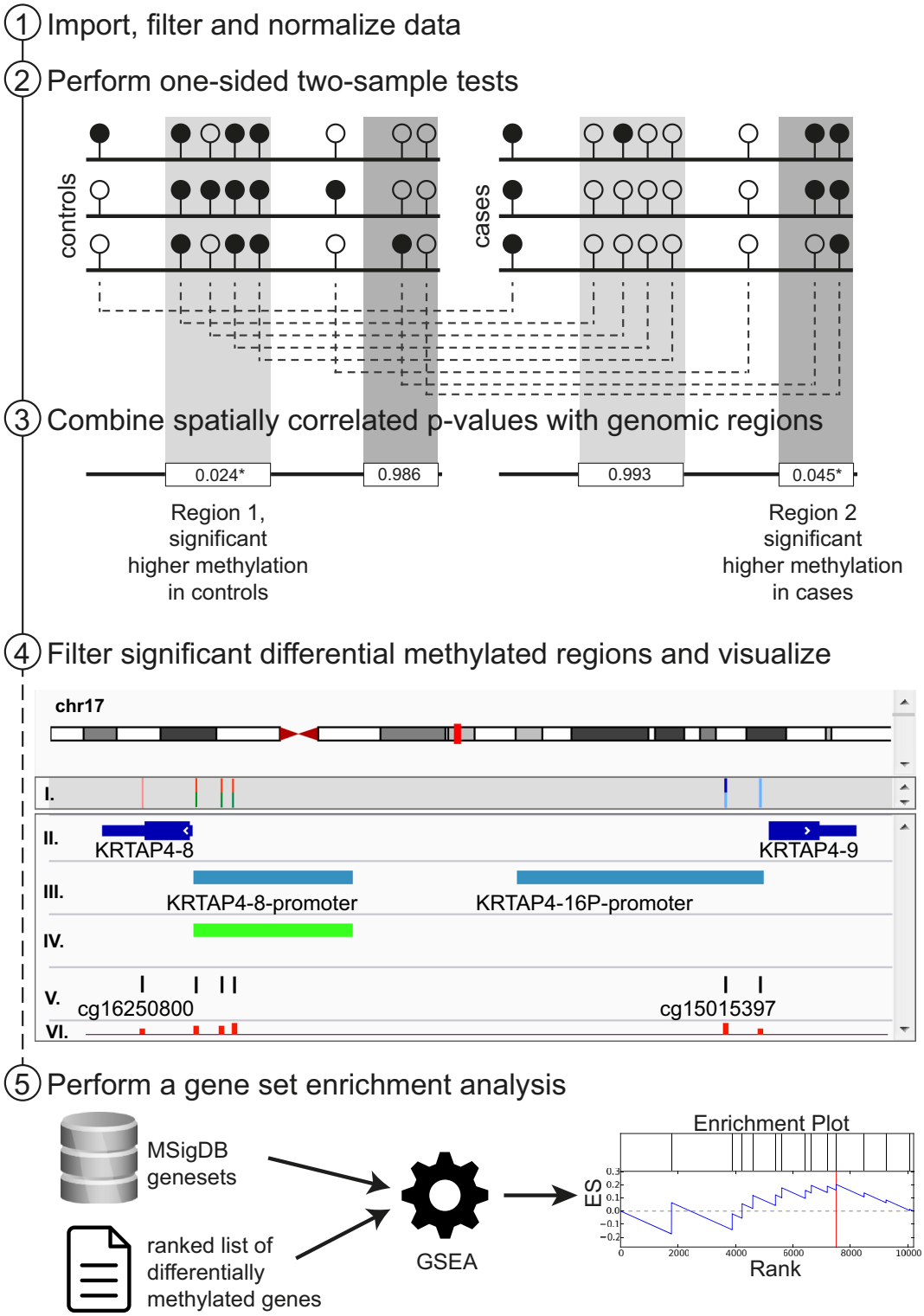


Fig. 2 Workflow is illustrated on the *left side* as five steps. *Step 2* Controls and cases are illustrated as replicates with methylated (*black*) and unmethylated (*white*) CpG sites. Single sites are compared between controls and cases (*dashed lines*). *Step 3* Site-specific *p* values are combined into genomic regions and a representative *q* value is calculated for each region (*light gray*: higher methylation in control; *dark gray*: higher methylation in cases). *Step 4* IGV screenshot of array visualization; tracks represent: (*I.*) single CpG site *q* values for two conditions with a *color code*, (*II.*) positions of known genes, (*III.*), selected regions of interest, (*IV.*) significant regions found by the pipeline, (*V.*) all probes represented on the array, and (*VI.*) *bar plot* track denoting absolute methylation change (up/down). *Step 5* An optional gene set enrichment analysis (GSEA) can be performed using pre-defined or custom gene sets and ranked lists of differentially methylated genes

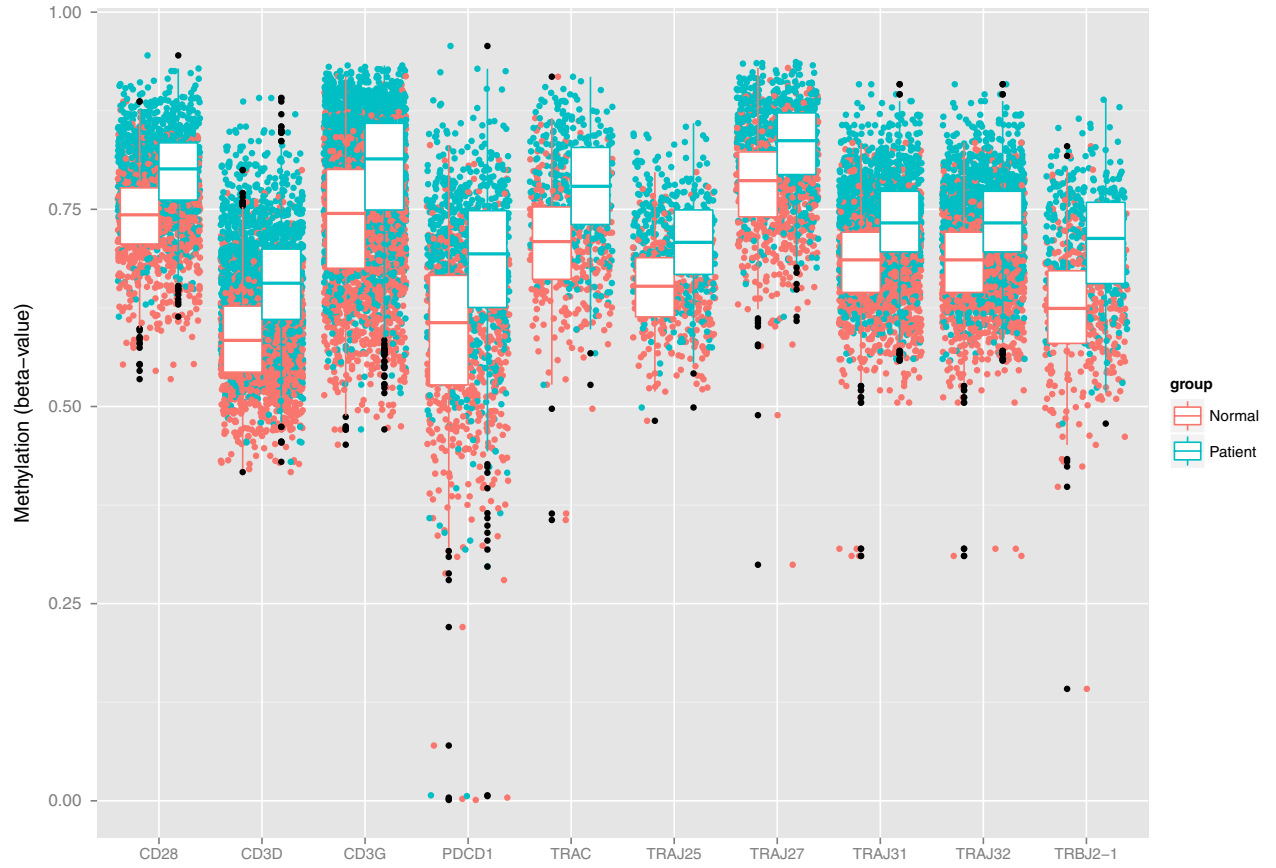


Fig. 3 Methylation of members of the T-cell receptor signaling pathway in rheumatic arthritis (*Patients*) vs. healthy individuals (*Normal*). Shown are the beta values from all methylation sites of 2-kb upstream regulatory regions across all 689 replicates. *Black dots* represent *boxplot outliers*

the presence of a variant increased the risk of AF up to 65 %. Located proximally to the variant, PITX2 is a transcription factor import for cardiogenesis, especially for left–right signaling and L/R atrial identity. Knockout of PITX2 lead to a shortened atrial action potential in haploinsufficient mice and increased the susceptibility to AF [17]. Expression analysis identified the Sinoatrial node (SAN) specific genes Shox2, Tbx3, and Hcn4 as upregulated in PITX2 null-mutant embryos [18]. A recent study additionally identified two microRNAs miR-17-92 and miR-106b-25 as direct targets of PITX2 that can repress Shox2 and Tbx3 upon transcription [19] and promote the expression of Cx43, a connexin protein forming gap junctions that allow the interchange of charged ions between adjacent cells [20]. Another GWAS study linked TBX5 to AF [21]. The homeobox transcription factor may play a role in heart development and specification of limb identity [22]. Interestingly, TBX5 was identified as interactor of Tbx3, a regulator of the SAN gene program [23]. Hoxa3 is another important gene in heart chamber morphogenesis, since Hoxa3-expressing progenitor cells in the second heart field give rise to the atria and parts of the outflow tract [24].

Summarizing these findings, we conclude that using genome-wide tiling regions as well as the positional gene sets in the implemented gene set enrichment provide a powerful and yet unbiased downstream analysis option to the user. As shown by the comparison to RnBeads, we assume ADMIRE to have a higher sensitivity to detect small changes in methylation rate, as the user can decide upon appropriate thresholds for absolute difference in methylation. Both datasets used for performance evaluation are available as shared data libraries on the ADMIRE web server (see Additional file 3 for loading shared data libraries).

Discussion

Integration and differential analysis of DNA methylation represents a major topic in clinical bioinformatics, most often addressed by whole-genome bisulfite sequencing or Infinium HumanMethylation450 Assays. Given the nature of methylation assay data, most of

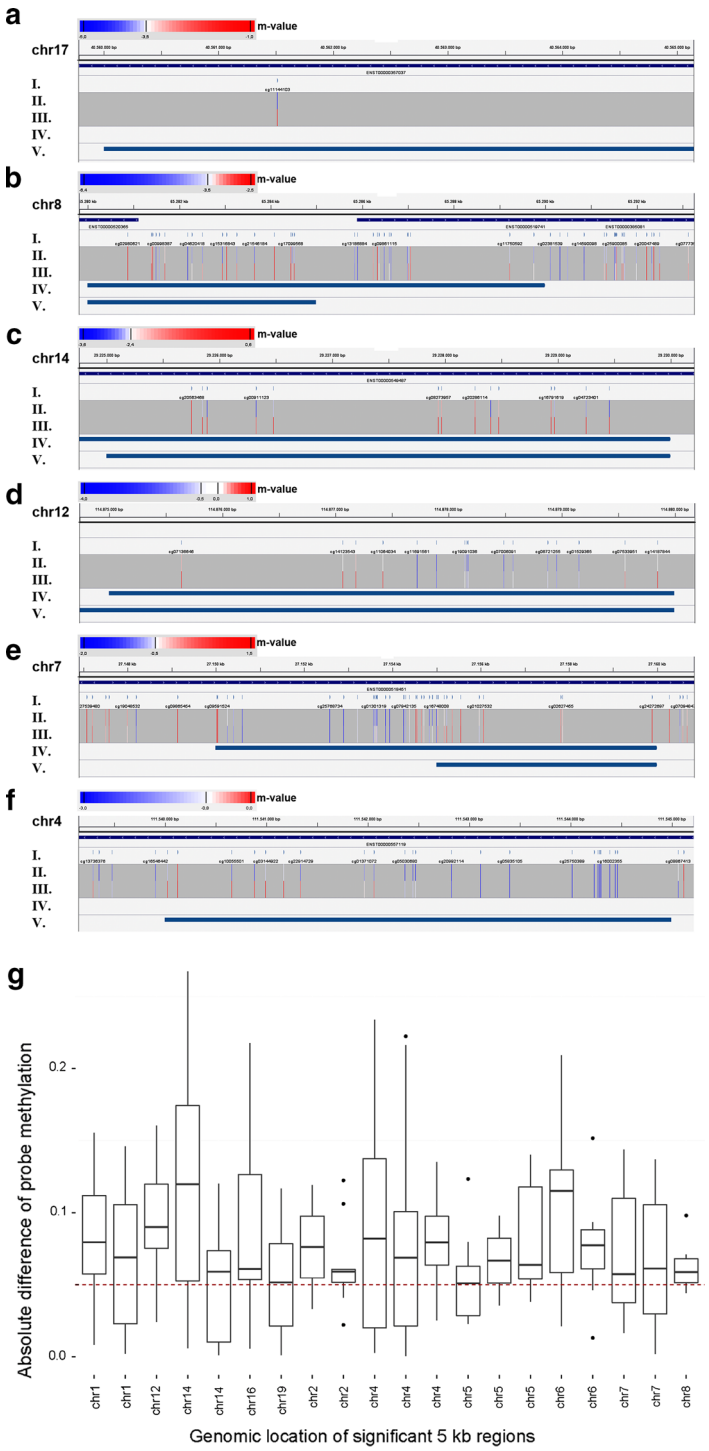


Fig. 4 IGV screenshots showing methylation across several genomic locations and *boxplots* for all significant sites. **a**–**f** Tracks shown are as follows: *I.* Methylation sites present on the HumanMethylation450 K Chip, *II. Color-coded* methylation values from control samples, *III.* Color-coded methylation values from AF samples, *IV.* Differentially methylated 10-kbp tiling regions called by ADMIRE, *V.* Differentially methylated 5-kbp tiling regions called by ADMIRE. The *color bar* encodes the *m* value, with blue indicating low methylation values and *red* indicating high methylation values. The absolute scale is created indvidually for each bar. Track *IV* and *V* are only used if the search with the corresponding input (5- or 10-kB tiling size) resulted in a significant region. **a** A 5-kbp region from chr17 called to be differentially methylated by RnBeads with an adjusted *p* value of 0.00008. **b**–**f** Top 5 differentially methylated regions from Admire with *q* values between 0.0004 and 0.003. **g** *Boxplots* for 20 significantly changed protein coding genes (higher in AF sample) identified by ADMIRE. *Each box* illustrates the distribution of absolute differences of the methylation values in the respective significantly changed region (see also Additional file 4). The cutoff at median methylation value of 5 % is shown as *red dashed line*

the analysis tools developed in the past are primarily focused on command line-based programming libraries, such as the R-based ChAMP [25] or minfi [7] packages, limiting the use of these tools to users with at least some programming skills. A second group of tools are intended to provide a comprehensive graphical interface to the user, including MethLAB [26], COHCAP [27], EpiDiff [28], and the Genome Studio (Illumina, proprietary license). Within this group, only two tools are available (RnBeads and ADMIRE) that are capable to provide their service not only on the command line but also as a web-based graphical user interface. While all of these programs are arguably valuable contributions to facilitate the analysis of Illumina HumanMethylation450 Assays, many may be too demanding to wet lab researchers and clinicians with limited computational skills. To face these needs, a web frontend might impose the least number of restrictions to the user. The intuitive, interactive, and relatively simple interface of ADMIRE facilitates the upload, analysis, and visualization of a complex technology. The input is limited to the raw files, a sample sheet describing the groups of interest and the selection of a few parameters. Common experimental setups in molecular studies that define more than two groups are addressed by automated all-vs-all comparisons. Genomic regions and gene sets are available as precomputed files, but the possibility to upload custom files offers a variety of downstream analysis options. Unfortunately, public web services frequently perform very limited in terms of throughput, since the workload has to be managed by the website provider. In case of HumanMethylation450Assays, the web-based analysis from RnBeads is limited to 24 arrays. In contrast, the algorithm of ADMIRE is designed to transfer the computational effort to the number of probes that are tested and is influenced only in a minor grade by the number of arrays under investigation. This focus permits the provision of the web service not only for small projects with a limited number of arrays, but also for large projects encompassing hundreds of input samples (performance evaluation with 689 input samples). Results from the original publication [15] handling these arrays, identify the MHC region as a major genetic risk loci in rheumatic arthritis. MHC peptides are bound by T-cell receptors together with their co-receptors *CD28* and *CD3*. ADMIRE highly supports this result, by linking differential methylation in the T-cell receptor signaling pathway as an alternative mechanism to rheumatic arthritis. Furthermore, the differential methylation of *PDCD1* (*PD-1*), a co-inhibitor of the T-cell receptor signaling pathway involved in T-cell activation [29] could represent another mechanism by disturbing the control of autoimmunity.

Conclusion

ADMIRE offers an intuitive interface to analyze DNA methylation patterns based on Infinium HumanMethylation450 Assays. Whereas most existing analysis tools are designed to be used on the command line, ADMIRE provides an easy to use web-based service as well as a version for local execution. A wide range of experimental and statistical settings can be adjusted, including normalization methods and detection of differentially methylated positions and regions. Whereas these regions are often precalculated in other tools, ADMIRE can calculate statistics on individual regions of interest provided by the user. As an optional step towards downstream analysis, ADMIRE additionally implements a gene set enrichment procedure. ADMIRE is freely accessible without a limit on experimental size at https://bioinformatics.mpi-bn.mpg.de.

Methods

Implementation

ADMIRE was implemented in Bash, R, and Python while making use of the open-source Bioconductor package minfi [7] and the comb-p [9] tool for data processing. Additionally, a variant of GSEA [12] is fully implemented in ADMIRE for gene set enrichment analysis. The pipeline was integrated into a Galaxy-based [30] platform similar to MIRPIPE [31] to provide online access but is also available for download and local execution. Input data can either be used immediately from Infinium HumanMethylation450 Assay compatible scanner systems (*SampleSheet.csv* and **.idat*-files) or the sample file can be prepared as a tab-separated text file. A detailed explanation of all input and output files is available in Additional file 3.

Generation of genetic regions and gene sets

Gene information from the GENCODE V19 [10] annotation was used to extract genomic regions for all exons (GTF feature type *exon*) and all 2-kB promoter regions downstream of the TSS. CpG islands were extracted from the Bioconductor annotation package *IlluminaHumanMethylation450kanno.ilmn12.hg19*. Enhancer information was downloaded from the Fantom5 project web site [11]. Bedtools *makewindows* function was used to generate genome-wide tiling regions of different sizes ranging from 50 bp up to 100 kB. All genomic regions were saved as bed files, keeping the *gene_name* property, if applicable. Gene sets for gene set enrichment analysis were downloaded from MSigDB [12] and are contained in the distribution of ADMIRE.

Benchmark and analysis of publicly available datasets

All raw **.idat*-files were downloaded from the respective GEO project site (GSE42861 and GSE62727). Tabular sample definition files were generated (see user manual). Admire was invoked using default parameters and the following genomic regions and gene sets: 2-kB promoter regions and positional gene sets for the rheumatic arthritis (RA) data and 5- and 10-kB genomic tiling regions for the atrial fibrillation (AF) data. Results from the RA data were limited to contain only protein coding genes and TR_C/TR_J genes with a Q-value below 0.01 and an absolute median difference in methylation between normal and patient samples of 5 % (Additional file 7). Remaining genes with higher methylation in patients were subjected to a GO analysis with two unranked lists of genes using GORILLA [32] (Additional file 8) and methylation values for significantly altered genes that map to the T-cell receptor signaling pathway were plotted in Fig. 3. Results from the AF data (Additional file 4) were annotated with their nearest gene using bedtools closest function and were limited to contain only protein coding genes with a median absolute difference of 5 %. Gene names were subjected to a GO analysis as described above. To analyze the sensitivity of ADMIRE, per-probe absolute differences were extracted using bedtools map function and plotted per chromosomal location in Fig. 4g.

Additional files

Additional file 1. Examples of publication ready graphical overviews.

Additional file 2. Comparison of available tools and packages for analysis of Illumina HumanMethylation450 Assays.

Additional file 3. ADMIRE documentation. The documentation provides description of all available parameters, input and output files as well as an example analysis of the atrial fibrillation data used in this publication.

Additional file 4. Significantly differentially methylated tiling regions in AF. Tiling regions were annotated with their nearest gene (see Methods).

Additional file 5. Absolute difference of methylation in 5 and 10 kB tiling regions in atrial fibrillation reported by ADMIRE. Boxplots give information about the magnitude of methylation change.

Additional file 6. Beta values of protein coding genes with significantly differential methylation between patients with atrial fibrillation and healthy individuals.

Additional file 7. Significantly differentially methylated promoter regions in RA.

Additional file 8. Enriched functional GO-Terms of genes with higher methylation values in RA. Background color codes for *p* values in the following way: >10–3 (white), 10–3 to 10–5 (light yellow) and 10–7 to 10–9 (orange).

Abbreviations

CpG: cytosine-phosphate-Guanine; DNA: deoxyribonucleic acid; GEO: gene expression omnibus database; GO: gene ontology; GTF: general transfer format; GWAS: genome-wide association study; IGV: integrative genomics viewer; kB: kilo basepairs; MHC: major histocompatibility complex; RAM: random access memory; SWAN: subset-quantile within array normalization; UCSC: University of California, Santa Cruz.

Authors' contributions

JP, CK, and ML conceived the algorithm; JP and JB implemented the algorithm; JP and ML analyzed the data and wrote manuscript with input from CK. All authors read and approved the final manuscript.

Acknowledgements

Funding Excellence Cluster Cardio-Pulmonary System (ECCPS); Max Planck Institute for Heart and Lung Research (MPI).

Competing interests

The authors declare that they have no competing interests.

Received: 24 August 2015 Accepted: 17 November 2015
Published online: 01 December 2015

References

1. Boland MJ, Nazor KL, Loring JF. Epigenetic regulation of pluripotency and differentiation. Circ Res. 2014;115(2):311–24. doi:10.1161/circresaha.115.301517.
2. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. Genomics. 2011;98(4):288–95. doi:10.1016/j.ygeno.2011.07.007.
3. Michels KB, Binder AM, Dedeurwaerder S, Epstein CB, Greally JM, Gut I, et al. Recommendations for the design and analysis of epigenome-wide association studies. Nat Methods. 2013;10(10):949–55. doi:10.1038/nmeth.2632.
4. Levenson VV. DNA methylation as a universal biomarker. Expert review of molecular diagnostics. 2010;10(4):481–8. doi:10.1586/erm.10.17.
5. Morris TJ, Beck S. Analysis pipelines and packages for Infinium HumanMethylation450 BeadChip (450 k) data. Methods. 2014;. doi:10.1016/j.ymeth.2014.08.011.
6. Assenov Y, Muller F, Lutsik P, Walter J, Lengauer T, Bock C. Comprehensive analysis of DNA methylation data with RnBeads. Nat Meth. 2014;11(11):1138-40. doi:10.1038/nmeth.3115. http://www.nature.com/nmeth/journal/v11/n11/abs/nmeth.3115.html#supplementary-information.
7. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30(10):1363–9. doi:10.1093/bioinformatics/btu049.
8. Bock C. Analysing and interpreting DNA methylation data. Nat Rev Genet. 2012;13(10):705–19. doi:10.1038/nrg3273.
9. Pedersen BS, Schwartz DA, Yang IV, Kechris KJ. Comb-p: software for combining, analyzing, grouping and correcting spatially correlated *P* values. Bioinformatics. 2012;28(22):2986–8. doi:10.1093/bioinformatics/bts545.
10. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22(9):1760–74. doi:10.1101/gr.135350.111.
11. Bertero T, Lu Y, Annis S, Hale A, Bhat B, Saggar R, et al. Systems-level regulation of microRNA networks by miR-130/301 promotes pulmonary hypertension. J Clin Investig. 2014;124(8):3514–28. doi:10.1172/JCI74773.
12. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci. 2005;102(43):15545–50. doi:10.1073/pnas.0506580102.
13. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14(2):178–92. doi:10.1093/bib/bbs017.
14. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002;12(6):996–1006. doi:10.1101/gr.229102 **(Article published online before print in May 2002)**.
15. Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat Biotechnol. 2013;31(2):142–7. doi:10.1038/nbt.2487.

16. Assenov Y, Muller F, Lutsik P, Walter J, Lengauer T, Bock C. Comprehensive analysis of DNA methylation data with RnBeads. Nat Methods. 2014;11(11):1138–40. doi:10.1038/nmeth.3115.
17. Zhou M, Liao Y, Tu X. The role of transcription factors in atrial fibrillation. J Thorac Dis. 2015;7(2):152–8. doi:10.3978/j.issn.2072-1439.2015.01.21.
18. Wang J, Klysik E, Sood S, Johnson RL, Wehrens XH, Martin JF. Pitx2 prevents susceptibility to atrial arrhythmias by inhibiting left-sided pacemaker specification. Proc Natl Acad Sci USA. 2010;107(21):9753–8. doi:10.1073/pnas.0912585107.
19. Wang J, Bai Y, Li N, Ye W, Zhang M, Greene SB, et al. Pitx2-microRNA pathway that delimits sinoatrial node development and inhibits predisposition to atrial fibrillation. Proc Natl Acad Sci USA. 2014;111(25):9181–6. doi:10.1073/pnas.1405411111.
20. Herve JC, Bourmeyster N, Sarrouilhe D, Duffy HS. Gap junctional complexes: from partners to functions. Prog Biophys Mol Biol. 2007;94(1–2):29–65. doi:10.1016/j.pbiomolbio.2007.03.010.
21. Zang X, Zhang S, Xia Y, Li S, Fu F, Li X, et al. SNP rs3825214 in TBX5 is associated with lone atrial fibrillation in Chinese Han population. PLoS One. 2013;8(5):e64966. doi:10.1371/journal.pone.0064966.
22. Tucker NR, Ellinor PT. Emerging directions in the genetics of atrial fibrillation. Circ Res. 2014;114(9):1469–82. doi:10.1161/CIRCRESAHA.114.302225.
23. Hoogaars WM, Engel A, Brons JF, Verkerk AO, de Lange FJ, Wong LY, et al. Tbx3 controls the sinoatrial node gene program and imposes pacemaker function on the atria. Genes Dev. 2007;21(9):1098–112. doi:10.1101/gad.416007.
24. Bertrand N, Roux M, Ryckebusch L, Niederreither K, Dolle P, Moon A, et al. Hox genes define distinct progenitor sub-domains within the second heart field. Dev Biol. 2011;353(2):266–74. doi:10.1016/j.ydbio.2011.02.029.
25. Morris TJ, Butcher LM, Feber A, Teschendorff AE, Chakravarthy AR, Wojdacz TK, et al. ChAMP: 450 k chip analysis methylation pipeline. Bioinformatics. 2014;30(3):428–30. doi:10.1093/bioinformatics/btt684.
26. Kilaru V, Barfield RT, Schroeder JW, Smith AK, Conneely KN. MethLAB: a graphical user interface package for the analysis of array-based DNA methylation data. Epigenet Off J of the DNA Methyl Soc. 2012;7(3):225–9. doi:10.4161/epi.7.3.19284.
27. Warden CD, Lee H, Tompkins JD, Li X, Wang C, Riggs AD, et al. COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis. Nucleic Acids Res. 2013;41(11):e117. doi:10.1093/nar/gkt242.
28. Zhang Y, Su J, Yu D, Wu Q, Yan H. EpiDiff: entropy-based quantitative identification of differential epigenetic modification regions from epigenomes. Conf Proc IEEE Eng Med Biol Soc. 2013;2013:655–8. doi:10.1109/EMBC.2013.6609585.
29. Sharpe AH, Wherry EJ, Ahmed R, Freeman GJ. The function of programmed cell death 1 and its ligands in regulating autoimmunity and infection. Nat Immunol. 2007;8(3):239–45. doi:10.1038/ni1443.
30. Goecks J, Nekrutenko A, Taylor J, Galaxy T. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010;11(8):R86. doi:10.1186/gb-2010-11-8-r86.
31. Kuenne C, Preussner J, Herzog M, Braun T, Looso M. MIRPIPE: quantification of microRNAs in niche model organisms. Bioinformatics. 2014;30(23):3412–3. doi:10.1093/bioinformatics/btu573.
32. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinform. 2009;10:48. doi:10.1186/1471-2105-10-48.

2.3 A molecular subtype of cancer originating from adult stem cells during regeneration is driven by Dux transcription factors

Jens Preussner[1,2,‡], Jiasheng Zhong[1,‡], Krishnamoorthy Sreenivasan[1], Stefan Günther[1,3], Thomas Engleitner[4], Carsten Künne[1,2], Markus Glatzel[5], Roland Rad[4], Mario Looso[2], Thomas Braun[1,6,7] and Johnny Kim[1,6]

Affiliations	[1] Department of Cardiac Development and Remodeling [2] Bioinformatics Core Unit (BCU) [3] ECCPS Deep Sequencing Platform Max Planck Institute for Heart and Lung Research [4] Institute of Molecular Oncology and Functional Genomics Translatum Cancer Center and Department of Medicine II Technical University of Munich [5] Institute of Neuropathology University Medical Center Hamburg-Eppendorf [6] German Center for Cardiovascular Research (DZHK) [7] German Center for Lung Research (DZL) [‡] These authors contributed equally
Journal	Cell Stem Cell, *Cell Press*
Date, Issue	2018, Vol. 23 no. 6
Pages	794 - 805
DOI	10.1016/j.stem.2018.10.011
Supplementary data	available online

Contributions

The following contributions are attributed to the thesis autor:

Conceptualization	Contributed to definition of project goals; Outlined data analysis strategy
Methodology	Hypothesis formulation of the role of Duxbl in epithelialization
Investigation	Analysis of recurring SNVs and CNVs, previously published discovery cohorts, data from TCGA; Investigation of the role of Dux family genes in ZGA; Analysis of previously published data from RNA-seq to investigate the role of Duxbl in epithelialization
Validation	Examined Dux-dependent zygotic gene activation in Chen et al. (2013), Davicioni et al. (2009) and Williamson et al. (2010)

Software	Implementation of software to analyse data from whole-exome sequencing
Visualization	Figures 3C, 3F, 3G, Figure 4, Figure 6 and graphical abstract; Supplementary Figures S3 to S7
Resources	Tables S1 and S2 supporting CN and SNV analysis; Tables S3, S4 supporting findings from TCGA analysis
Writing	Contributed to the manuscript draft; Review and editing of the manuscript

CellPress

Cell Stem Cell
Article

Oncogenic Amplification of Zygotic Dux Factors in Regenerating p53-Deficient Muscle Stem Cells Defines a Molecular Cancer Subtype

Jens Preussner,[1,2,8] Jiasheng Zhong,[1,8] Krishnamoorthy Sreenivasan,[1] Stefan Günther,[1,3] Thomas Engleitner,[4] Carsten Künne,[1,2] Markus Glatzel,[5] Roland Rad,[4] Mario Looso,[2] Thomas Braun,[1,6,7] and Johnny Kim[1,6,9,*]

[1]Department of Cardiac Development and Remodeling, Max-Planck-Institute for Heart and Lung Research, Bad Nauheim, Germany
[2]Bioinformatics Core Unit (BCU), Max Planck Institute for Heart and Lung Research, Bad Nauheim, Germany
[3]ECCPS Deep Sequencing Platform, Max Planck Institute for Heart and Lung Research, Bad Nauheim, Germany
[4]Institute of Molecular Oncology and Functional Genomics, Translatum Cancer Center and Department of Medicine II, Technical University of Munich, Munich, Germany
[5]Institute of Neuropathology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany
[6]German Center for Cardiovascular Research (DZHK), Rhine Main, Germany
[7]German Center for Lung Research (DZL), Giessen, Germany
[8]These authors contributed equally
[9]Lead Contact
*Correspondence: johnny.kim@mpi-bn.mpg.de
https://doi.org/10.1016/j.stem.2018.10.011

SUMMARY

The identity of tumor-initiating cells in many cancer types is unknown. Tumors often express genes associated with embryonic development, although the contributions of zygotic programs to tumor initiation and formation are poorly understood. Here, we show that regeneration-induced loss of quiescence in p53-deficient muscle stem cells (MuSCs) results in rhabdomyosarcoma formation with 100% penetrance. Genomic analyses of purified tumor cells revealed spontaneous and discrete oncogenic amplifications in MuSCs that drive tumorigenesis, including, but not limited to, the amplification of the cleavage-stage Dux transcription factor (TF) Duxbl. We further found that Dux factors drive an early embryonic gene signature that defines a molecular subtype across a broad range of human cancers. Duxbl initiates tumorigenesis by enforcing a mesenchymal-to-epithelial transition, and targeted inactivation of Duxbl specifically in Duxbl-expressing tumor cells abolishes their expansion. These findings reveal how regeneration and genomic instability can interact to activate zygotic genes that drive tumor initiation and growth.

INTRODUCTION

The cell of origin for many cancer types remains unknown, although the hypothesis has been put forward that cancerous stem cells (CSCs) typically arise out of healthy stem cells. In support of this hypothesis, prevalent types of cancer most often occur in tissues containing cells with increased proliferative potential inferred by tissue resident stem cells (SCs), which normally enable regeneration of the respective tissue (Morrison and Spradling, 2008; Tomasetti and Vogelstein, 2015). An excellent example and experimentally tractable model to study stem-cell-dependent regeneration is that of skeletal muscle, which is mediated by and dependent on rare Pax7-expressing muscle SCs that reside between the basal lamina and plasma membrane of mature skeletal muscle fibers (Almada and Wagers, 2016; Günther et al., 2013). Under resting conditions, muscle SCs are predominantly quiescent but become activated upon regenerative cues, such as an inflicting injury or during chronic regeneration of certain diseases. For example, muscle SCs from mdx mice that mimic certain features of Duchenne muscular dystrophy undergo continuous activation due to persistent muscle fiber degeneration and the consequent requirement for *de novo* fiber formation under steady-state conditions (Boldrin et al., 2015).

Recently, it was shown that germline inactivation of the tumor suppressor p53 in chronically regenerating mdx mice develop rhabdomyosarcoma (RMS) (Camboni et al., 2012; Chamberlain et al., 2007), a rare and aggressive childhood cancer and the most common soft-tissue sarcoma in children and adolescents (El Demellawy et al., 2017). The cancer cell of origin of RMS has yet remained unclear, in particular under these settings, although forced expression of common potent oncogenic drivers in muscle SCs, including, but not limited to, *kras* or *yap1*, can result in RMS formation (Blum et al., 2013; Chen et al., 2013; Hettmer et al., 2011; Shern et al., 2014; Tremblay et al., 2014). Other reports have indicated RMS to originate in mesenchymal cells (Wang et al., 2014), and it was recently demonstrated that RMS can arise through malignant myogenic trans-differentiation of endothelial progenitors via activation of the hedgehog pathway (Drummond et al., 2018). Indeed, RMS tumors are generally thought of as skeletal muscle tumors because they display features of myogenic differentiation reflected by the expression of myogenic determinants, such as MyoD, MyoG, and Desmin, all of which are sequentially expressed in activated muscle SCs during the progression of adult muscle regeneration and in muscle progenitors during

embryonic muscle development (Almada and Wagers, 2016; Braun and Gautel, 2011). Incidentally, the notion has been raised that re-expression of genes normally expressed in tissue-specific progenitors during embryonic development might be responsible for the stem-cell-like phenotypes of various poorly differentiated human tumors, including RMS, germ cell tumors, breast cancer tumors, glioblastoma, and bladder cell carcinoma (Ben-Porath et al., 2008). However, in which cell types, how, and by what means embryonic gene expression programs would be elicited to induce tumor formation and whether or not any of these associated genes are causally transformative has remained largely unknown.

Genomic sequencing of tumors from patients suffering from RMS and many other types of cancer has unveiled oncogenic mutations and copy number (CN) gains of certain genes associated with tumor formation (Chen et al., 2013; Editorial, 2015 [in *Nature Medicine*]; Shern et al., 2014). However, for most patients, both the specific cellular and genetic etiologies of tumor formation remain unknown, illustrating that mechanisms of tumorigenesis remain to be identified and by more refined methods. The constant proliferative and regenerative pressure on the muscle SC compartment in continuously regenerating mdx mice led us to reason that muscle SCs could be a cellular origin of RMS tumors. Here, we devised an inducible strategy (1) to clarify the cellular origin of RMS tumors, (2) to delineate the causal role of stem-cell-dependent regeneration in cancer progression, and (3) that would enable identification of causative mechanisms, leading to tumorigenic transformation of healthy stem cells *in vivo*.

RESULTS

Lineage Tracing Identifies Muscle SCs as a Cellular Origin of Embryonic RMS

To investigate the cell-autonomous role of muscle SCs in RMS formation, we generated mice that enable muscle-SC-specific deletion of p53 in both wild-type and constitutively regenerating mdx mice by intraperitoneal administration of tamoxifen (TAM), designated hereafter as SC^{p53} and $SC^{p53/MDX}$, respectively. We additionally introduced a Rosa26::lsl Tomato allele enabling permanent fluorescent lineage tracing of SCs after TAM treatment (Figure 1A). Strikingly, 20 weeks after TAM treatment, all $SC^{p53/MDX}$ mice developed tumors in, or immediate proximity to, the musculature of extremities or the trunk and were all histopathologically classified as embryonic RMS immunopositive for Desmin, MyoD, and MyoG (Figures 1A–1C and S1A). Remarkably, some mice developed several tumors; however, TAM-treated wild-type, mdx, or SC^{p53} mice never developed tumors up to an age of more than 52 weeks (Figure 1B). Consecutive bouts of cardiotoxin (CTX)-induced injury to the tibialis anterior (TA) muscle of SC^{p53} mice uniformly resulted in RMS formation at the site of injury, whereas control animals never developed tumors (Figure S1B). These data show that muscle-SC-specific loss of p53 in a regenerative environment is sufficient to generate RMS, or conversely, that a regenerative environment enables RMS formation upon muscle-SC-specific loss of p53. Moreover, these data support the notion that maintenance of SC quiescence provides a cellular mechanism to suppress tumorigenesis and are consistent with previous reports demonstrating that muscle injury is required to elicit RMS formation upon forced overexpression of oncogenic drivers, such as yap1 (Tremblay et al., 2014). All of the tumors were lineage traced by activation of the Rosa26-Tomato locus, clearly indicating muscle SCs as the cellular origin of the RMS tumors in these animals (Figure 1C). Consistently, strong Tomato fluorescence of the skeletal muscle, but not the liver, revealed prominent and specific contribution of muscle SCs toward *de novo* myofiber formation, as expected in chronically regenerating mdx muscles and over the period before tumor onset (Figure 1C). We next obtained single cell isolates from surgically excised skeletal muscles and tumors and subjected them to fluorescence-activated cell sorting (FACS), enabling separation and purification of muscle SCs and both lineage-traced and non-lineage-traced tumor-propagating cells (TPCs), respectively (Figures 1D, S1C, and S1D). PCR-based genotyping, immunofluorescent staining, and qRT-PCR of isolated tissues and FACS-purified SCs and TPCs confirmed that recombination of the p53 locus was highly efficient and specific to muscle SCs and disclosed that purified Tomato-positive cells were strictly deficient for p53 and purified Tomato negative cells only contained non-recombined p53 DNA (Figures 1E and S1C–S1F). In addition, immunofluorescent staining and qRT-PCR of the freshly isolated SCs revealed that chronic regeneration elicits expression of p53 in a subset of SCs, albeit at low levels (Figures S1E–S1G). Notably, regenerated skeletal muscles still contained DNA with intact p53 alleles, most likely reflecting the presence of myofibers derived from muscle SCs before and after the onset of p53 deletion and/or other muscle-resident cells. Likewise, DNA isolated from bulk tumors contained DNA with both intact and recombined p53 alleles (Figure 1E), in agreement with the observation that the tumors consisted of lineage-traced and non-lineage-traced TPCs (Figures 1C and 1D). These data demonstrate (1) the specific contribution of Pax7-expressing muscle SCs toward RMS formation and (2) the complex cellular composition of the developing tumors containing cells that are and are not derived from Pax7-expressing SCs. Essentially, FACS-separated TPCs were either strictly $p53^{+/+}Tom^{neg}$ or $p53^{-/-}Tom^{pos}$ (Figures 1D and 1E) that did or did not express myf5, myod, or myog, respectively, further confirming efficient labeling of $p53^{-/-}$ muscle SCs (Figure 1F). In contrast to $p53^{-/-}Tom^{pos}$ cells, $p53^{+/+}Tom^{neg}$ cells were highly enriched for cdkn1a mRNA transcripts (also known as p21), a primary target of p53 (Figure 1F), emphasizing that the loss of p53 functions in RMS formation specifically in muscle-SC-derived cells expressing myogenic markers. Importantly, only transplantation of $p53^{-/-}Tom^{pos}$, but not $p53^{+/+}Tom^{neg}$, cells into immunocompromised mdx-nude mice generated tumors at the site of injection already two weeks after transplantation (Figures 1G and 1H). Taken together, these data demonstrate that lineage tracing enables prospective purification of genuine tumor-propagating cells clearly originating from muscle SCs that are responsible for tumor formation.

Genomic Analyses of Purified TPCs

To gain insight on the molecular mechanism leading to transformation of SCs, we next analyzed proliferation and differentiation of wild-type, mdx, p53, and p53/mdx muscle SCs *in vitro*. Inactivation of p53 and/or dystrophin in muscle SCs did not impair terminal myogenic differentiation upon serum withdrawal (Figure 2A). However, under growth conditions, p53-deficient

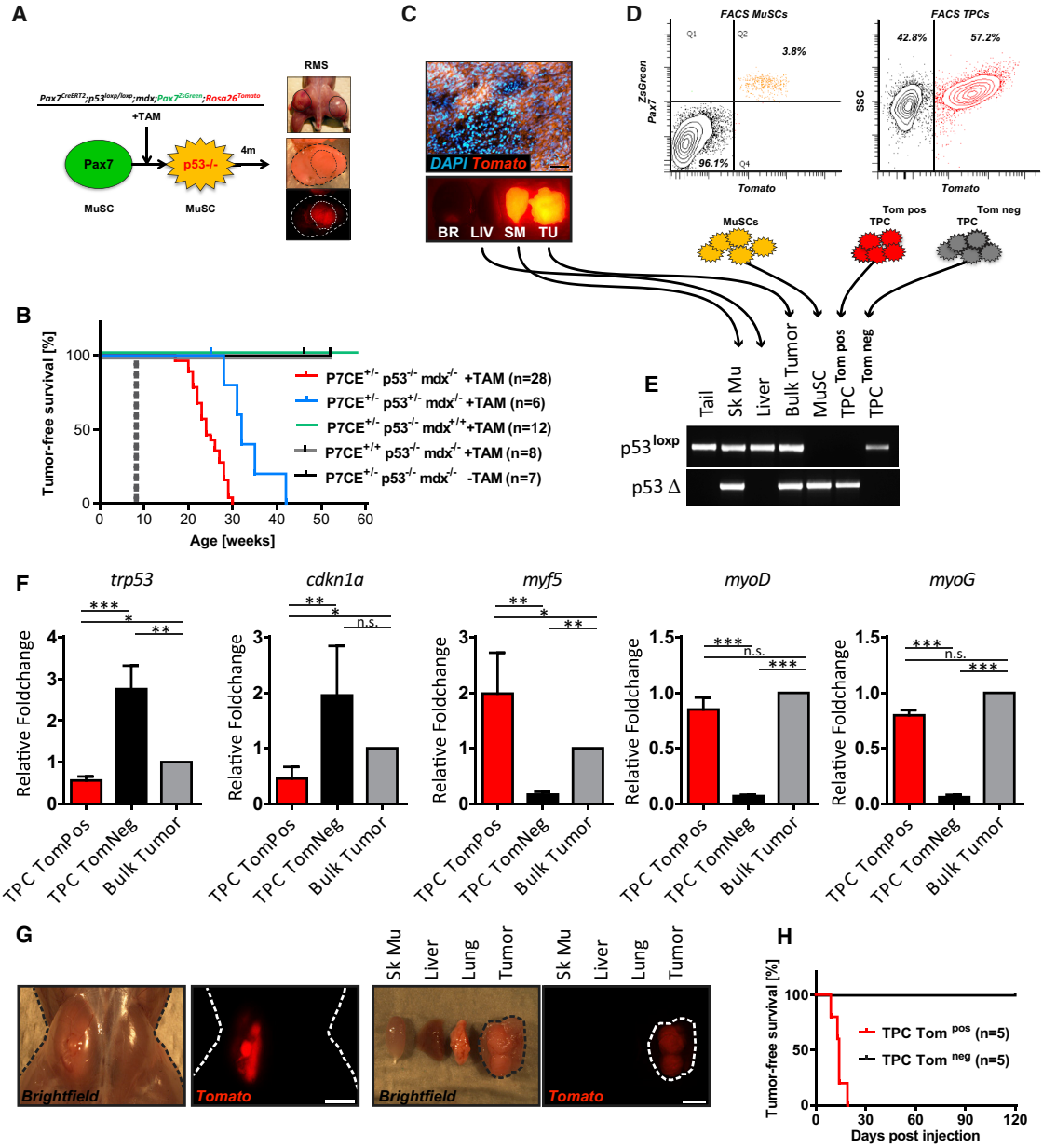


Figure 1. Loss of p53 in Muscle SCs Induces ERMS Tumors in mdx Mice

(A) Genetic components of the model system. p53 is deleted in muscle SCs expressing ZsGreen and simultaneously lineage traced *in vivo* via recombination of a Rosa26lsl::CAG^{Tomato} allele upon tamoxifen injection. Fluorescence enables FACS-based purification of muscle SCs and separation of lineage-traced and non-lineage-traced TPCs.

(B) Kaplan-Meier tumor-free survival curves are shown for indicated genotypes. Dashed line indicates timing of tamoxifen administration.

(C) Representative immunofluorescent images of isolated tissues (bottom panel) and cross-sectioned tumor (top panel). Note that not all cells within the tumor are lineage traced.

(D) FACS plots of purified SCs and TPCs.

(E) Genotyping of the p53 gene locus in indicated tissues, purified muscle SCs, and TPCs.

(F) mRNA expression analysis of p53, cdkn1a, and myogenic factors in purified TPCs and corresponding bulk tumors. Error bars indicate SD of the mean (t test: *p < 0.05; **p < 0.01; ***p < 0.001; n = 4).

(G) Macroscopic image of Tomato expression in tumors after injection of purified TPC^{TOMpos} in mdx-nude mice. Scale bar: 5 mm.

(H) Kaplan-Meier tumor-free survival curves for mdx-nude mice injected with either TPCs^{TOMpos} or TPCs^{TOMneg}.

muscle SCs displayed significantly enhanced proliferation and concomitant formation of DNA double-strand breaks (DSBs) reflected by a dramatic increase of EdU-incorporating, γH2AX, ATM, and 53bp1-positive SCs (Figures 2B–2E). These data demonstrate that loss of p53 does not impair differentiation but permits accumulation of mutations in actively proliferating

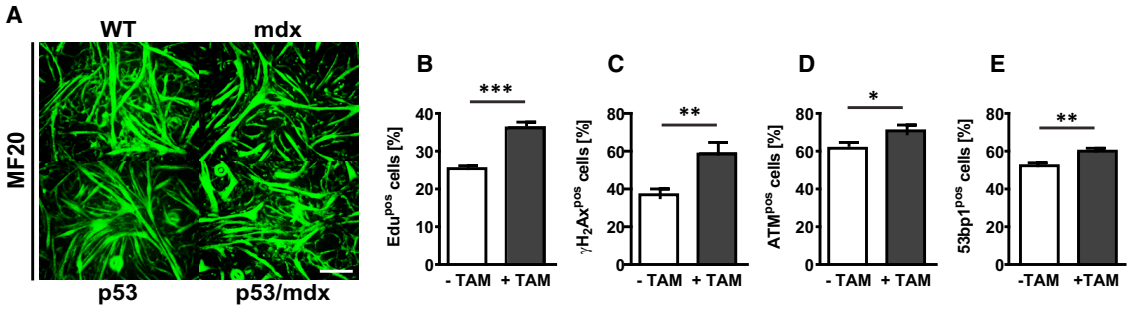


Figure 2. Loss of p53 Does Not Inhibit Myogenic Differentiation but Elicits Genomic Instability in Activated Muscle SCs

(A) Immunofluorescent staining of differentiated myotubes for indicated genotypes with MF20 antibody detecting sarcomeric actin. Scale bar: 50 μm.

(B–E) Quantification of percentage of (B) EdU, (C) γH2AX, (D) ATM, and (E) 53bp1-positive muscle SCs from $SC^{p53/mdx}$ mice in culture under growth conditions. Error bars represent SDs of the mean (t test: *p < 0.05; **p < 0.01; ***p < 0.001).

SCs, of which some might be responsible for tumorigenic transformation.

To identify mutations promoting tumor formation, we performed whole-exome sequencing on paired tumor-normal samples. Strikingly, CN analysis revealed discrete and dramatic genomic CN amplifications in almost every specimen (21 out of 22) when paired purified TPC^{TOMpos}-normal samples were used (Figures 3B–3G, S2A–S2S, S3A, and S3B). In contrast, we did not always detect CN variations when genomic DNA from bulk tumors was analyzed, demonstrating that non-tumorigenic stromal cells interfere with the analysis of genomic sequencing data (Figure 3A). Positional mapping of CN amplifications within each and across samples revealed genomic amplification of defined chromosomal regions harboring known mutational targets in ERMS, including yap1 (Tremblay et al., 2014; 8/22; 36%), c-met (Fleischmann et al., 2003; Taulli et al., 2006; 5/22; 23%), jun (Durbin et al., 2009; 1/22; 4.5%), and cdk4/gli1/os9 (Liu et al., 2014; 1/22; 4.5%). We also identified mutational targets that had not been associated with embryonic RMS (ERMS) so far but with other types of cancer, including rras (Flex et al., 2014), kdm4d (Soini et al., 2015), bap1 (Robertson et al., 2017), mcm4 (Polotskaia et al., 2015; Shima et al., 2007), and eloc (Sato et al., 2013; Figures 3F, 3G, S2A–S2S, S3A, and S3B). Several mice displayed amplification on chromosome 14qA3 (5/22; 23%), harboring a poorly described triplicated genomic locus in the mouse genome encoding for the genes plac9, tmem254, cphx, and duxbl (Figures 3D, 3F, and 3G). Notably, genomic analysis of two highly aggressive allografts, which harbored yap1/c-met and yap1/kdm4d amplifications, respectively, did not reveal significant accumulation of *de novo* mutations or CNAs two weeks after the tumor cell transplant (Figure 3F). In addition, a linkage between the amplified genes within and across samples was not detected, suggesting that, for each individual animal, a single discrete amplification was likely sufficient for SC transformation. Interestingly, we also noticed a dramatic reduction of mtDNA in many of the TPCs, suggesting those to be glycolytic and had likely undergone a Warburg effect (Figures 3B, 3D, 3E, and S2A–S2S). In contrast to oncogenic amplifications, somatic single-nucleotide variations (SNVs) were astonishingly low (Figures S4A and S4B), concordant with recent observations that soft-tissue sarcomas are predominantly characterized by copy number changes, with low mutational loads of only a few genes, including p53, atrx, and rb1 (The Cancer Genome Atlas Research Network, 2017; Chen et al., 2013; El Demellawy et al., 2017; Shern et al., 2014). Tables supporting CN and SNV analyses are provided in Tables S1 and S2.

Dux TFs Define a Molecular Subtype of Cancer

The subset of tumors containing amplification of 14qA3 was of particular interest, because this genomic region does not contain any known oncogene. We noticed that 14qA3 harbors the *duxbl* gene, which belongs to the Dux family of homeobox-containing transcription factors with human Dux4 as the founding member (Leidenroth and Hewitt, 2010). Recently it was shown that Dux4, or the murine homolog Dux, is responsible for driving cleavage-stage gene expression signatures known as zygotic gene activation (ZGA) in totipotent embryonic stem cells (ESCs) (De Iaco et al., 2017; Hendrickson et al., 2017; Whiddon et al., 2017). The human genome encodes two additional paralogs of Dux4, named DuxA and DuxB, which are expressed exclusively at the totipotent 8-cell stage in early zygotes (Madissoon et al., 2016). These observations led us to speculate that Dux transcription factors might act at a putative interface of stem cell potency and tumor formation.

To test this hypothesis, we analyzed previously published discovery cohorts of human ERMS patients (Chen et al., 2013; Davicioni et al., 2009; Williamson et al., 2010) for expression of Dux4, DuxA and DuxB, and/or ZGA. We assumed that a putative role of Dux genes in causing human ERMS might have been missed in the past due to poor annotation and in particular of the Dux family and ZGA-associated genes. In this context, it is noteworthy to mention that the human *duxA* and *duxB* genes were only recently identified and annotated, with human *duxB* being in genomic synteny with mouse *duxbl* and human *dux4* is in synteny with murine *dux* (Leidenroth et al., 2012; Leidenroth and Hewitt, 2010). Intriguingly, and consistent with the proposed role of Dux factors in driving cleavage-stage-specific gene expression signatures (Hendrickson et al., 2017), we identified 54 RMS tumors (~10%) that were strikingly positive for ZGA and/or Dux factor expression (Figures 4A and S5A–S5D; Table S3). Reanalysis of raw sequencing data of mRNA transcripts available from the dataset of Chen et al. (2013) disclosed four ZGA-positive patients of which two (X03D and X20A) expressed

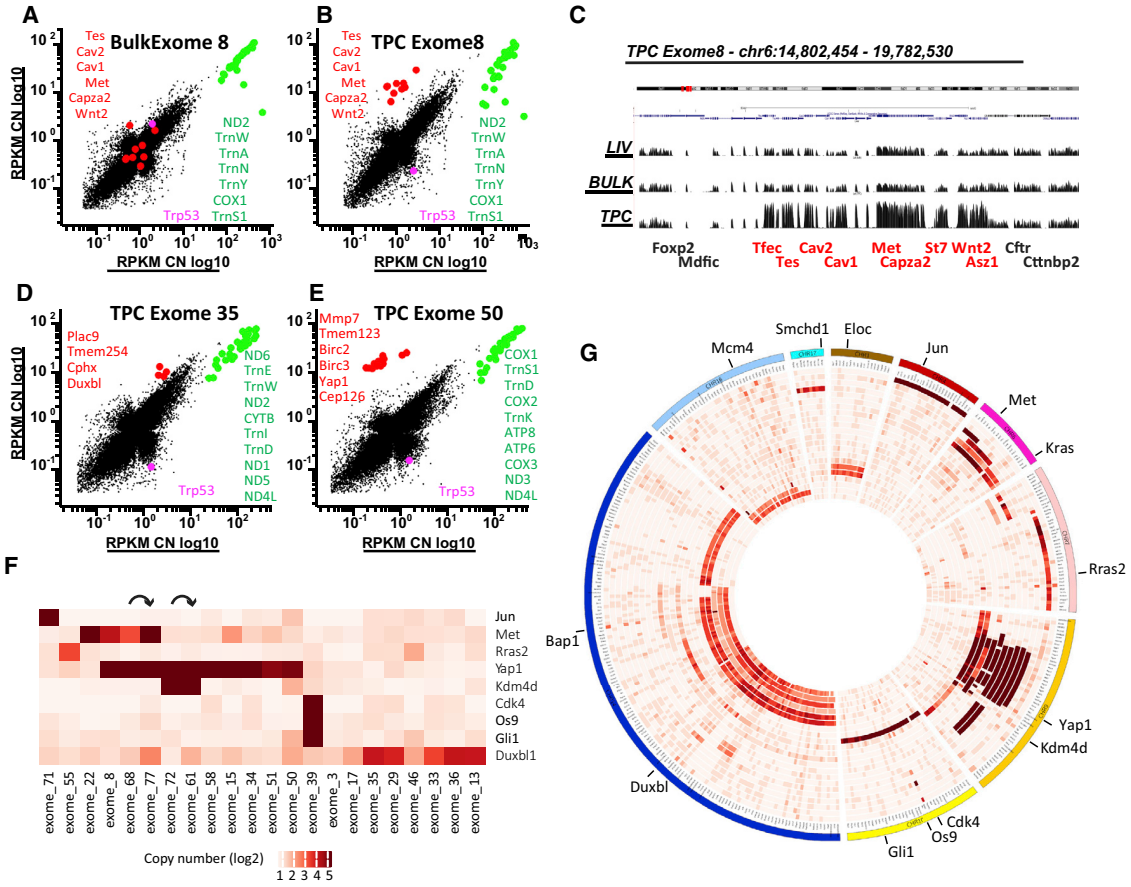


Figure 3. Identification of Distinct Copy Number Amplifications in RMS Tumors

(A, B, D, and E) Scatterplots depicting log-scaled RPKM values of genomic DNA in purified tumor cells (y axis) (B, D, and E) or in non-purified bulk tumor (A) versus liver control (x axis). Note that the same genes are highlighted in red in (A), (B), and (C). Green circles represent mitochondria-encoded genes. Red circles represent amplified genes. Magenta circle represents p53.

(C) Amplified genes highlighted in red in (A) and (B) displayed in physical genomic order.

(F) Heatmap summarizing amplified genes in the 22 analyzed tumors. Curved arrows indicate allografts of donor TPCs and TPCs after transplantation into recipient.

(G) Cumulative CIRCOS plot of amplified genes from all analyzed tumors. Note that genomic regions are displayed in physical order.

dramatically high levels of Dux4, DuxA, and DuxB (Figures 4A and S5A). These data indicate that Dux transcription-factor-driven zygotic gene activation defines a molecular signature of a new ERMS subtype. We next sought to investigate whether increased expression of Dux genes is restricted to ERMS or also associated with other malignancies. To this end, we re-screened <10,000 cancer patients of The Cancer Genome Atlas–Pan-Cancer (TCGA-PANCAN) dataset (Hoadley et al., 2014) for Dux4, DuxA, and DuxB exon expression. Intriguingly, we identified 349 patients that displayed distinct expression of Dux4, DuxA, and DuxB either in combination or alone. Cancer onset and type in these patients was highly variable, comprising of 32 different types of somatic cancer according to ICD-1O (International Classification of Diseases for Oncology) (Figure 4B; Table S3). These data show that Dux transcription factors driving early zygotic gene signatures define a molecular subtype of a broad range of human cancers.

We noticed that two tumors from the sequenced RMS cohort (X013D and X45D) displayed ZGA to a lower degree than X03D and X20A, which corresponded to clearly detectable but lower levels of Dux4, DuxA, or DuxB (Figures S5A and S5B). Similar expression patterns were visible in the larger cohort of patients from the PANCAN dataset, supporting the idea that Dux factors act upstream of ZGA to initiate tumorigenesis but that sustained Dux-mediated ZGA may not be required for maintenance of established tumors. Notably, tumors showing the most striking expression of Dux4, DuxA, DuxB, and ZGA were predominantly classified as testicular germ cell carcinomas (TGCs), more than half of which developed in subjects younger than 20 years old, thus indicating a particular vulnerability of Dux-ZGA–associated tumorigenesis in germ cell tissues (Figure 4C; Tables S3 and S4). Conspicuously, the most prevalent cancer types across the 349 patients (testis, breast, kidney, stomach, and lung) are thought to be of epithelial origin, and virtually all of the ZGA-positive tumors displayed prominent levels of Dux4, DuxA, and/or DuxB, but not vice versa (Figures 4B–4D, S6A, and S6B). Particularly apparent was that almost all of the breast cancer patients expressed dramatic levels of DuxB-Duxbl, but only few were additionally positive for ZGA,

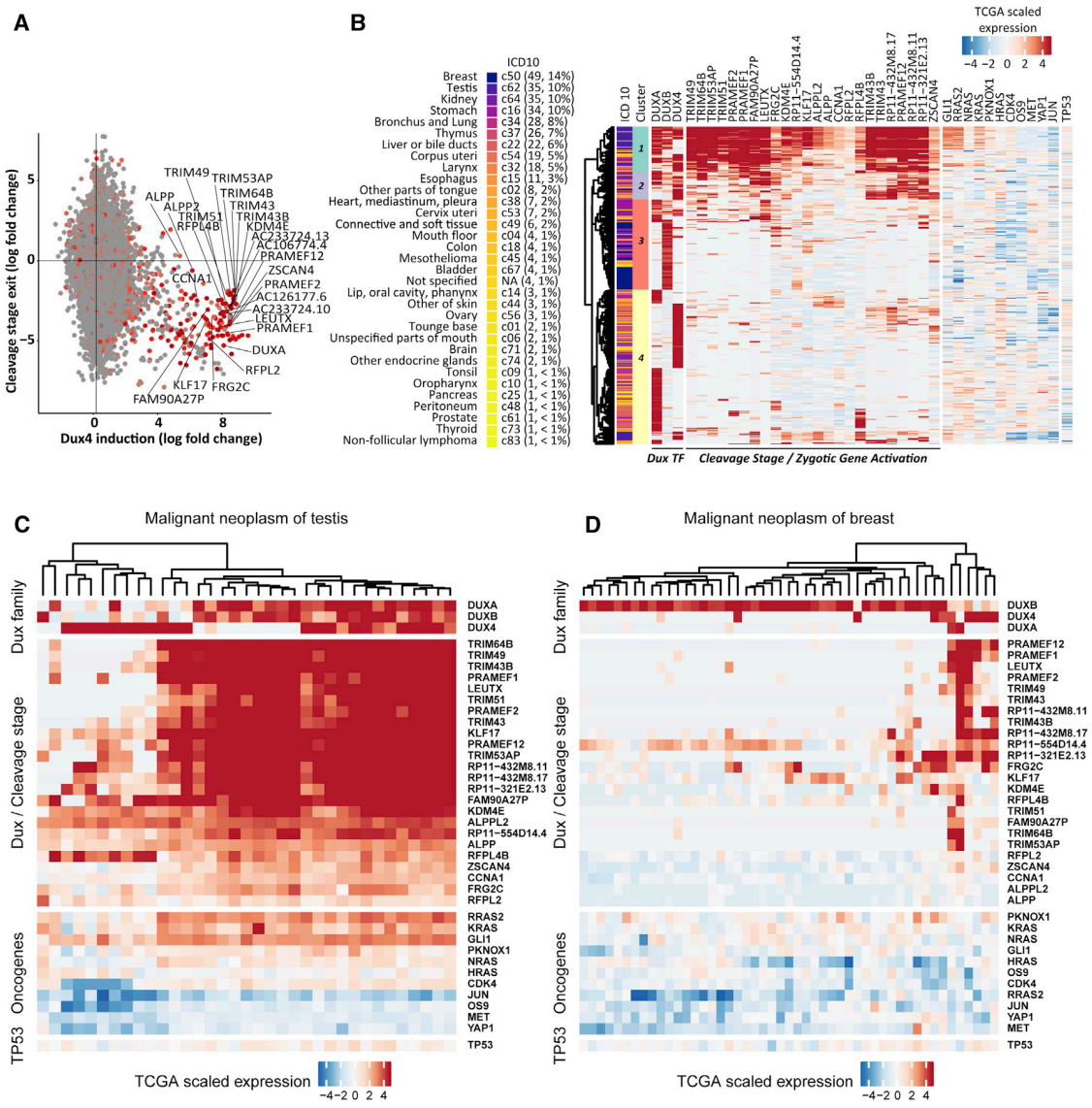


Figure 4. Dux Transcription Factors Define a Molecular Subtype of Cancer

(A) Integration analysis of Dux-dependent zygotic gene expression (x axis) and cleavage-stage-specific gene expression (y axis) in tumor from ERMS patient X20A.

(B) Unsupervised cluster analysis of tumors from the full TCGA-PANCAN dataset revealing four subgroups driven by mRNA expression of Dux factors and/or Dux-dependent zygotic gene activation. Percentages indicate prevalence of color-coded tumor type across 349 Dux-factor-ZGA–positive tumors. Clinical metadata of tumors stratified by age, gender, and cluster are provided in Tables S3 and S4.

(C and D) Unsupervised cluster analysis of breast cancer (C) and testicular germ cell carcinoma (D) positive for expression of Dux factors and/or Dux-dependent zygotic gene activation.

DuxA, and/or Dux4, suggesting Dux4 to act genetically upstream of DuxB (Figure 4D). Consistent with this hypothesis, acute overexpression of Dux4 elicited a profound and significant upregulation of ZGA-associated genes, as previously reported, and additionally DuxA and DuxB in human embryonic kidney cells (Figure S7A; Table S5; Hendrickson et al., 2017). Taken together, these data raise the idea that (1) Dux factors initiate tumorigenesis via activation of ZGA and/or (2) that Dux factor expression, and in particular DuxB-Duxbl, might facilitate tumorigenesis by an additional ZGA-independent mechanism.

Duxbl Initiates Tumor Formation via Eliciting a Mesenchymal-to-Epithelial-Transition-like Program

To gain a mechanistic understanding on the action of DuxB-Duxbl, we overexpressed Duxbl in wild-type muscle SCs via lentiviral transduction (Figure 5A). In wild-type control muscle SCs (designated hereafter as SC^{WT}), serial passaging resulted in exhaustion of proliferation concomitant with progressive formation of flattened myoblasts that robustly fused to form terminally differentiated myotubes (Figures 5B–5D). In striking contrast, overexpression of Duxbl resulted in the emergence of

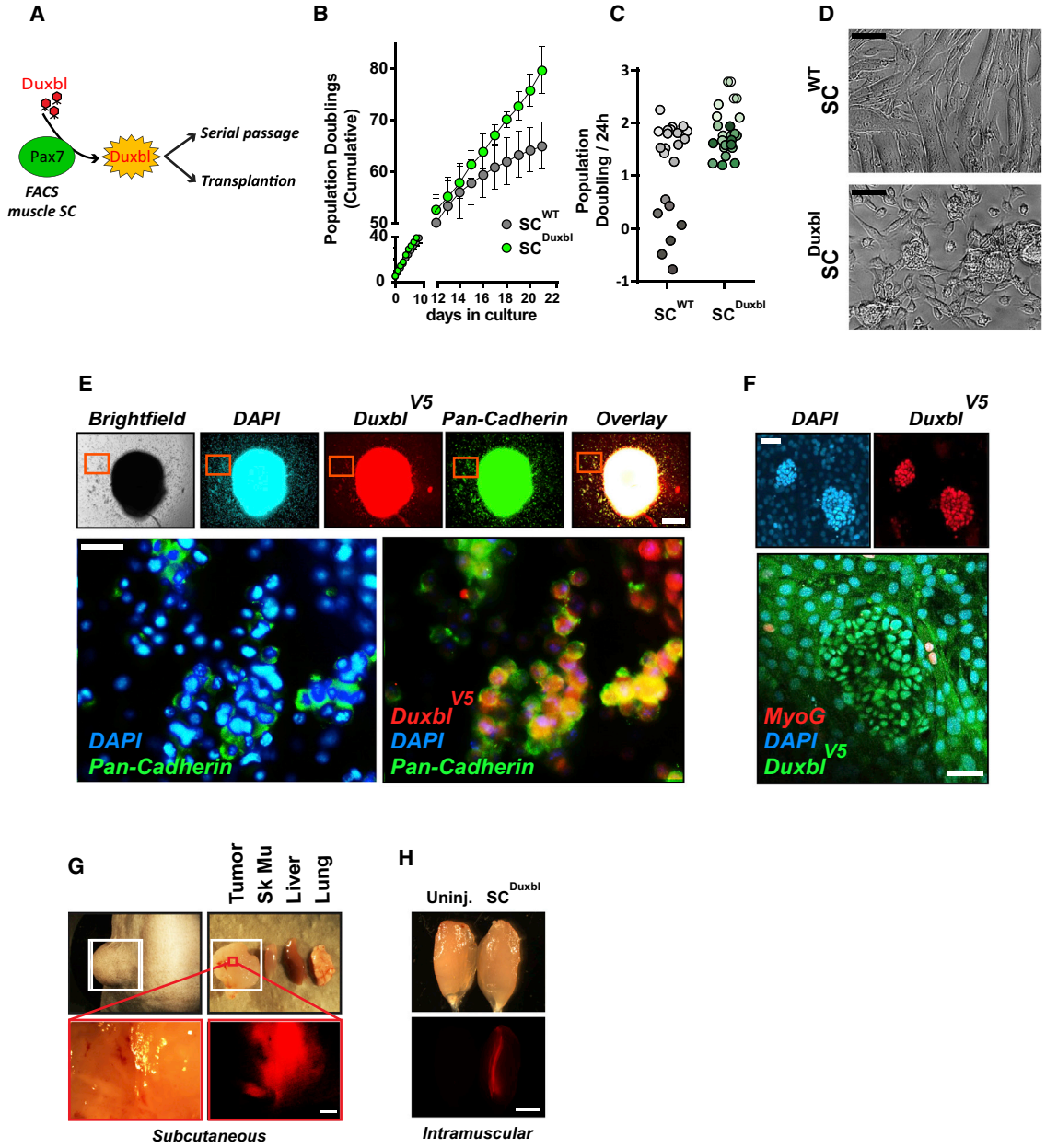


Figure 5. Overexpression of Duxbl Confers Plasticity and Promotes Tumor Formation

(A) Schematic of lentiviral-mediated transduction of V5-tagged Duxbl into wild-type (WT) muscle SCs.

(B) Cumulative population doublings over three weeks of culture. Error bars indicate SD of the mean.

(C) Population doubling rate calculated as PD = log(N1/N0)log2, where N1 is the final cell number and N0 is the initial number of seeded cells. Darker shading indicates increasing days in culture.

(D) Representative images of SCs^{WT} and SCs^{Duxbl} after 3 weeks in culture. Scale bars: 50 μm.

(E and F) Pseudo-colored immunofluorescent staining of SC^{Duxbl} colonies as indicated. Scale bars: 100 μm in top panels and 20 μm in lower panels of (E), and 20 μm in top panels and 50 μm in lower panels of (F).

(G and H) Macroscopic image of mCherry expression after injection of SC^{Duxbl} carrying an additional mCherry reporter (G) subcutaneously or (H) in TA muscle. Scale bar: 100 μM in (G) and 1 mm in (H).

immortalized and morphologically rounded clones prone to spontaneously form cadherin-positive, epithelial-like spherical aggregates that were devoid of myogenic differentiation and could be passaged indefinitely *in vitro* (designated hereafter as SC^{Duxbl}; Figures 5B–5F). Importantly, subcutaneous transplantation of SC^{Duxbl} resulted in tumor formation at the site of engraftment, clearly demonstrating that overexpression of Duxbl is sufficient for neoplastic transformation *in vivo* (Figure 5G). In

contrast, SC^{WT} did not give rise to any detectable neoplasias 3 months after subcutaneous transplantation as previously reported (data not shown; Irintchev et al., 1998). Interestingly, SC^{Duxbl} contributed to myofiber formation when injected directly into the strong pro-differentiation environment of TA muscle (Figure 5H). Collectively, these data provided an additional important mechanistic insight: it appeared that forced expression of Duxbl renders SCs to obtain increased plasticity and that the ability of SC^{Duxbl} to form tumors requires sustained reduction of pro-differentiation cues *in vivo*.

To gain a deeper molecular insight on how Duxbl might confer such plasticity, we performed RNA sequencing (RNA-seq) and compared gene expression profiles of isolated SC^{Duxbl} clones to various stages of myogenic differentiation of SC^{WT}. Linear principal-component analysis (PCA) revealed that separation of conditions (PC1; explaining 33.9% of variance) was primarily driven by differential expression of genes related to stem cell maintenance and differentiation (e.g., quiescent SC^{WT} expressed Pax7, but not MyoG, in comparison to differentiated SC^{WT} and vice versa; Figures 6A–6C). Notably, SC^{Duxbl} did not express any myogenic determinants, including Myf5, Myod, or MyoG, further demonstrating that forced expression of Duxbl profoundly impairs muscle SC differentiation. Intriguingly, separation of conditions across the second dimension (PC2; explaining 26.4% of variance) was enforced by dramatic expression of genes driving epithelialization (Figures 6A–6C). Indeed, pairwise differential expression analysis of SC^{Duxbl} with all stages of muscle SC differentiation (false discovery rate [FDR] < 0.01) followed by gene set enrichment analysis (GSEA) revealed a dramatic induction of genes involved in focal adhesion and proliferation of epithelial cells (Figures 6D–6F and S7B–S7D; Table S6). Consistently, SC^{Duxbl} revealed a dramatic upregulation or induction of numerous genes encoding for integrins, collagens, and most prominently cadherins and proto-cadherins (Figures 6C and S6B–S6D). Most intriguingly, SC^{Duxbl} expressed high levels of the neural and pluripotency factor *sox2*, which was expectedly absent in all stages of SC^{WT}. Interestingly, and similar to neural stem cells, undifferentiated quiescent SC^{WT} expressed high levels of the pluripotency factor *klf4* (Kim et al., 2009), the expression of which was downregulated during differentiation of SC^{WT} but sustained in SC^{Duxbl} (Figures 6A and 6C). Notably, both *sox2* and *klf4* are instrumental to facilitate an essential mesenchymal-to-epithelial transition (MET) event during reprogramming of somatic cells to induced pluripotent stem cells (iPSCs) (Li et al., 2010). In aggregate, these data indicate that forced expression of Duxbl most likely initiates tumorigenic transformation and colonization via a mechanism similar to MET (Figure 6G).

Targeting Oncogenic Duxbl

Finally, we sought out to test whether inactivation of DuxB-Duxbl could serve as a potential target in the primary tumor cells purified from the $SC^{p53/MDX}$ mice. To this end, we carried out short hairpin RNA (shRNA)-mediated knockdown in early passaged TPCs purified from primary tumors harboring either Duxbl (TPC^{Duxbl}) or Yap1 (TPC^{Yap1}) CN amplifications. Notably, expression of Duxbl mRNA was solely found in TPC^{Duxbl}, but not in TPC^{Yap1}. Likewise, expression of yap1 was restricted to TPC^{Yap1} (Figure 7A). Strikingly, shRNA-mediated knockdown of Duxbl in TPC^{Duxbl}, but not in TPC^{Yap1}, and vice versa, resulted in cell death, demonstrating that Duxbl expression is required specifically for maintenance of TPC^{Duxbl} *in vitro* (Figures 7B and 7C; related videos on Mendeley Data at https://doi.org/10.17632/7g2pbbrn4m.1). Collectively, these data demonstrate that DuxB-Duxbl can be targeted in tumor cells but disease-causing mutations in each individual tumor need to be identified before specific therapeutic intervention.

DISCUSSION

Genetic factors and certain environmental factors, such as exposure to viruses, radiation, or other carcinogens, are known to increase the risk of cancers, but for most cases, the cellular origin and cause of cancer remains unknown. Variation of cancer risk across tissues might be explained by the number of divisions of tissue-resident stem cells (Tomasetti and Vogelstein, 2015), which differs depending on developmental stages. In support of this claim, loss of p53 in muscle SCs only resulted in ERMS formation in mice undergoing continuous skeletal muscle regeneration and with astonishingly low variation of latency.

Interestingly, p53 single-knockout mice rarely develop rhabdomyosarcomas (Donehower et al., 1992), and although chronically regenerating mdx mice display elevated levels of DNA damage under steady-state conditions, their predisposition to spontaneously develop rhabdomyosarcomas predominantly occurs after more than one year of age (Camboni et al., 2012; Chamberlain et al., 2007). Previous observations revealing that compound p53/mdx germline knockouts almost only develop rhabdomyosarcomas raised several intriguing questions. Why and how does chronic muscle regeneration in p53 germline knockouts lead to predominant formation of RMS tumors and what is in fact the cellular origin of RMS tumors? Our data clearly show that sustained activation and division of muscle SCs promotes tumorigenesis in genomically instable SCs and additionally indicates maintenance of SC quiescence as a cellular mechanism to suppress tumorigenesis. Importantly, lineage tracing enabled us (1) to determine the cancer cell of origin, (2) to prospectively purify ''genuine'' tumor propagating cells, and (3) to identify discrete oncogenic amplifications associated with tumor formation via genomic sequencing. Separation of tumorigenic from stromal cells, which constitute a large part of solid tumors, was especially important in this respect, because the inherent complex cellular composition of tumors and non-tumorigenic cell therein interfered with analysis of genomic sequencing data obtained from bulk samples.

The Cancer Genome Atlas Research Network continues to report genome-wide signatures of pathologically defined tumor types, but past studies have used next generation sequencing (NGS) algorithms to detect mutations of well-annotated genomic loci, exempting recently discovered genes (Roychowdhury and Chinnaiyan, 2016). Accordingly, a more recent study revealed that oncogenic BAP1 alterations in uveal melanoma are missed by NGS mutation detection algorithms used in the past and can only be detected by more recently developed sequence-assembly-based methods (Robertson et al., 2017). Moreover, precise annotation of the human genome still remains incomplete (Chen et al., 2013; Steward et al., 2017). Therefore, tumor analyses need to be constantly repeated using updated annotated genomes and refined methods, including tumor cell purification

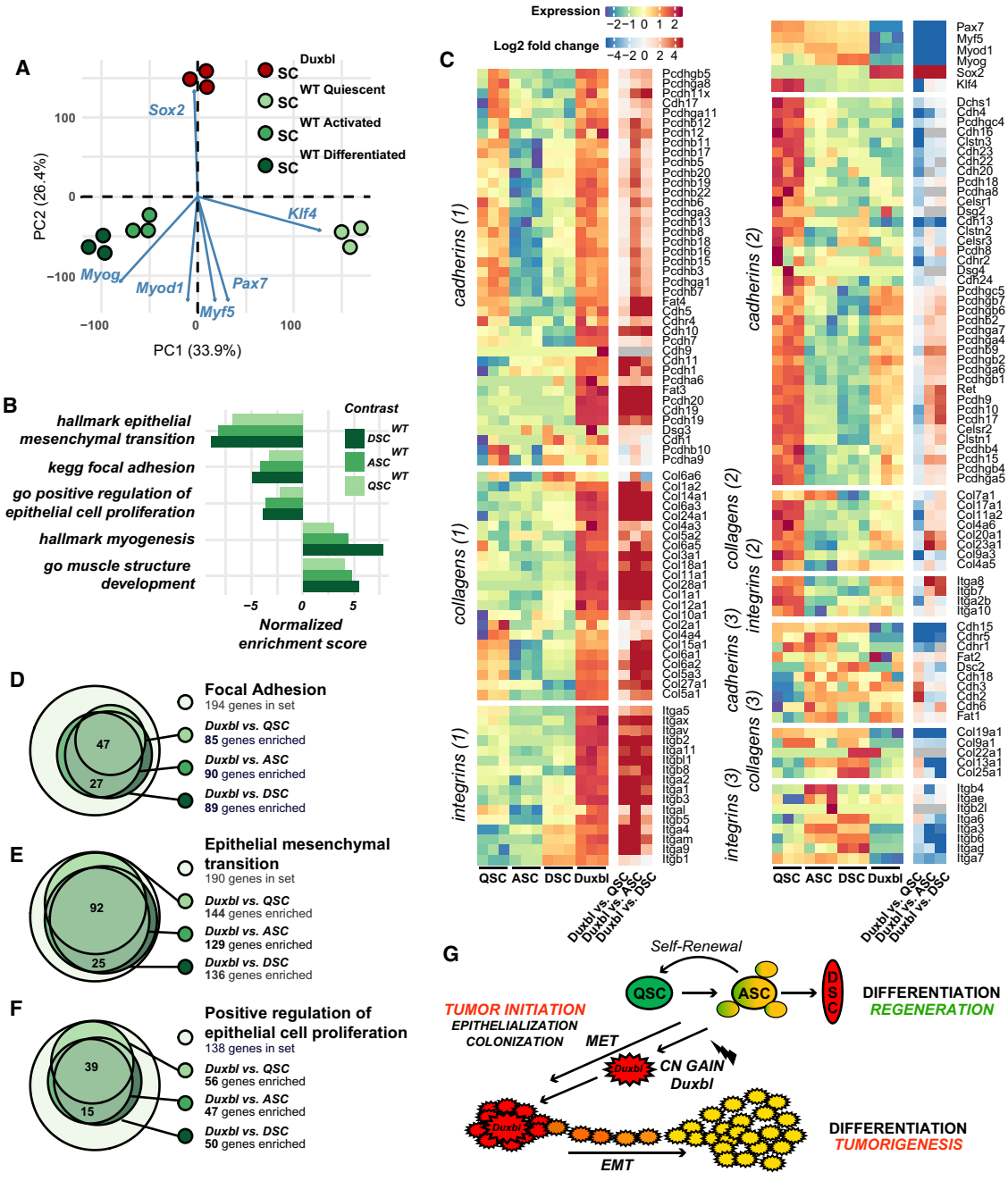


Figure 6. Overexpression of Duxbl in Muscle SCs Elicits Epithelialization

(A) Principal-component analyses of RNA-seq gene-expression data of SC^{Duxbl} and different stages of SC^{WT}.

(B and D–F) GSEA of SC^{Duxbl} and different stages of SC^{WT} with normalized enrichment score (B) and associated genes of focal adhesion (D), epithelial to mesenchymal transition (E), and regulation of epithelial cell proliferation (F).

(C) Heatmaps showing differential expression of genes driving focal adhesion, including cadherins, protocadherins, collagens, and integrins.

(G) Model depicting proposed mechanism of DuxB-Duxbl-mediated tumorigenesis. Healthy SCs contribute to muscle regeneration by differentiation of activated SCs upon injury. Copy number gain or expression of DuxB-Duxbl in activated SCs suppresses differentiation and promotes gain of SC plasticity accompanied by epithelialization and initiation of tumorigenic colonies. A secondary event likely involving EMT enables outgrowth of tumor cells from the tumor colony. ASC, activated stem cell; CN, copy number; DSC, differentiated stem cell; EMT, epithelial-to-mesenchymal transition; MET, mesenchymal-to-epithelial transition; QSC, quiescent stem cell.

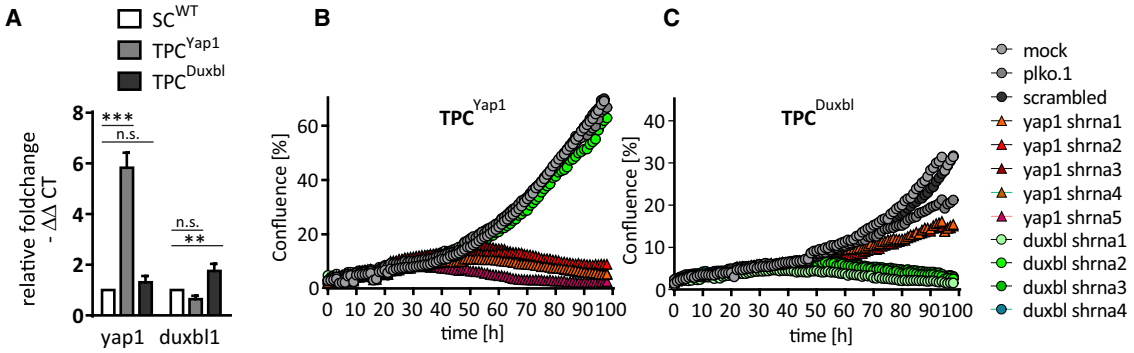


Figure 7. Targeting Duxbl

(A) qRT-PCR analysis of yap1 and duxbl in purified TPC^{Yap1} or TPC^{Duxbl}. Muscle stem cells purified from wild-type mice (SC^{WT}) served as controls. Expression levels were normalized to m36b4 mRNA. Error bars indicate SD of the mean (t test: $^{**}p < 0.01$; $^{***}p < 0.001$; n = 3).

(B and C) Growth curves of purified TPCs transduced with different shRNA lentiviruses targeting either yap1 or duxbl in TPC^{Yap1} (B) and TPC^{Duxbl} (C).

See also related videos deposited at Mendeley Data at https://doi.org/10.17632/7g2pbbrn4m.1.

or single-cell sequencing to identify new tumorigenic mutations. Here, we were able to unmask sample complexity through genomic profiling of purified TPCs, which enabled identification of distinct oncogenic amplifications in almost every analyzed animal.

Most importantly, we identified a novel oncogenic CNA of Duxbl, the murine paralog of human DuxB. Our data clearly indicate that redeployment of Dux transcription factors that define gene expression signatures of totipotent cleavage-stage ESCs (De Iaco et al., 2017; Hendrickson et al., 2017; Whiddon et al., 2017) confer stem cell expression profiles facilitating tumorigenesis in a broad range of human cancers and particularly in those of germ cell and epithelial origin. It did not miss our attention that three tumors harbored distinct copy number gains of the epigenetic regulators *smchd1* and *kdm4d* (Figure 3G), direct upstream and downstream effectors of Dux4, respectively, which indicates the Dux-ZGA axis to play a more prominent role in tumorigenesis than previously appreciated. Interestingly, derepression of silenced Dux4 in post-mitotic skeletal muscle fibers activates genes normally expressed in embryonic development and causes facioscapulohumeral muscular dystrophy (FSHD) (Lemmers et al., 2010). Ectopic expression of Dux4 in somatic cells causes cell death by yet unclarified mechanisms but seems to require a C-terminal domain specific to Dux4 that is not contained in DuxB-Duxbl (Bosnakovski et al., 2008a, 2017; Hewitt, 2015; Rickard et al., 2015). This indicates that different members of the Dux family appear to own different properties and/or exert different functions when expressed alone or with other Dux genes. In our mouse model, we observed oncogenic amplification of Duxbl, but not Dux, the human homolog of Dux4, and expression of Dux4 in human samples was almost always accompanied by co-expression of DuxB, which supports this conclusion. It is additionally tempting to speculate that higher resistance to cell death (e.g., by mutation of p53) might further render stem cells especially vulnerable to tumorigenic transformation by Dux factors.

The finding that Duxbl confers cellular plasticity and induces an MET-like process is particularly fascinating. The cancer stem cell (CSC) theory puts forward that most tumor cells lack tumor-initiating ability and that only a rare subpopulation of "stem-like" cells can lead to metastatic disease. Indeed, similar to pluripotent ESCs and iPSCs, CSCs show a plasticity that allows them to transition between epithelial- and mesenchymal-like states (Polyak and Weinberg, 2009). In accordance to the "seed and soil" hypothesis, we propose that forced expression of DuxB-Duxbl results in the initiation of a cancer stem cell that is a seed seeking the most fertile soil (a niche with constant low differentiation pressure) for it to grow in (Peinado et al., 2017). In such a scenario, it is reasonable that a chronic regeneration environment can generate focal niches that foster Duxbl-triggered MET, enabling initiation of plastic tumorigenic colonies, but a secondary event for the establishment of truly metastatic niches is likely required to enable tumor cell outgrowth (Figure 6G).

Finally, our results also revealed that inactivation of Duxbl in TPC^{Duxbl}, but not in TPC^{Yap1}, abolished tumor cell propagation and vice versa, indicating the dependence of individual tumors on distinct regulatory networks. The development of specific Dux inhibitors might allow personalized therapeutic interventions for patients suffering from Dux-factor-linked cancers, which can nowadays be easily diagnosed by sequencing approaches.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Mice
  - Cell culture
- METHOD DETAILS
  - mRNA expression analysis
  - Immunofluorescence and immunohistochemistry
  - Knockdown by shRNA transduction
  - Exome sequencing, bioinformatics
  - Analysis of public data
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and seven tables and can be found with this article online at https://doi.org/10.1016/j.stem.2018.10.011.

ACKNOWLEDGMENTS

This work was supported by the Max Planck Society, the DFG (Excellence Cluster Cardio-Pulmonary System [ECCPS]), the DFG Collaborative Research Centers SFB1213 (TP A02 and B02) and SFB TR81 (TP02), the LOEWE Center for Cell and Gene Therapy, the Foundation Leducq (3CVD01), and the German Center for Cardiovascular Research and the European Research Area Network on Cardiovascular Diseases project CLARIFY. M.G. is supported by a grant from the Herz Foundation "Infectophysics." We would like to thank J. Beetz and A. Romao for preparation of samples and help with animal work.

AUTHOR CONTRIBUTIONS

J.K. designed the study; J.P., J.Z., K.S., M.G., M.L., R.R., T.B., and J.K. interpreted and visualized data; J.P., T.E., C.K., M.L., and J.K. conducted bioinformatic analyses; J.K., J.Z., and K.S. performed experiments; S.G. performed sequencing; M.G. carried out pathological assessment; J.K. supervised the study; J.K. wrote the manuscript; and J.P., M.L., M.G., R.R., T.B., and J.K. edited the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 24, 2018
Revised: August 27, 2018
Accepted: October 8, 2018
Published: November 15, 2018

REFERENCES

Almada, A.E., and Wagers, A.J. (2016). Molecular circuitry of stem cell fate in skeletal muscle regeneration, ageing and disease. Nat. Rev. Mol. Cell Biol. *17*, 267–279.

Ben-Porath, I., Thomson, M.W., Carey, V.J., Ge, R., Bell, G.W., Regev, A., and Weinberg, R.A. (2008). An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. Nat Genet. *40*, 499–507.

Blum, J.M., Añó, L., Li, Z., Van Mater, D., Bennett, B.D., Sachdeva, M., Lagutina, I., Zhang, M., Mito, J.K., Dodd, L.G., et al. (2013). Distinct and overlapping sarcoma subtypes initiated from muscle stem and progenitor cells. Cell Rep. *5*, 933–940.

Boldrin, L., Zammit, P.S., and Morgan, J.E. (2015). Satellite cells from dystrophic muscle retain regenerative capacity. Stem Cell Res. (Amst.) *14*, 20–29.

Bosnakovski, D., Lamb, S., Simsek, T., Xu, Z., Belayew, A., Perlingeiro, R., and Kyba, M. (2008a). DUX4c, an FSHD candidate gene, interferes with myogenic regulators and abolishes myoblast differentiation. Exp. Neurol. *214*, 87–96.

Bosnakovski, D., Xu, Z., Li, W., Thet, S., Cleaver, O., Perlingeiro, R.C., and Kyba, M. (2008b). Prospective isolation of skeletal muscle stem cells with a Pax7 reporter. Stem Cells *26*, 3194–3204.

Bosnakovski, D., Toso, E.A., Hartweck, L.M., Magli, A., Lee, H.A., Thompson, E.R., Dandapat, A., Perlingeiro, R.C.R., and Kyba, M. (2017). The DUX4 homeodomains mediate inhibition of myogenesis and are functionally exchangeable with the Pax7 homeodomain. J. Cell Sci. *130*, 3685–3697.

Braun, T., and Gautel, M. (2011). Transcriptional mechanisms regulating skeletal muscle differentiation, growth and homeostasis. Nat. Rev. Mol. Cell Biol. *12*, 349–361.

Camboni, M., Hammond, S., Martin, L.T., and Martin, P.T. (2012). Induction of a regenerative microenvironment in skeletal muscle is sufficient to induce embryonal rhabdomyosarcoma in p53-deficient mice. J. Pathol. *226*, 40–49.

The Cancer Genome Atlas Research Network (2017). Comprehensive and integrated genomic characterization of adult soft tissue sarcomas. Cell *171*, 950–965.e28.

Chamberlain, J.S., Metzger, J., Reyes, M., Townsend, D., and Faulkner, J.A. (2007). Dystrophin-deficient mdx mice display a reduced life span and are susceptible to spontaneous rhabdomyosarcoma. FASEB J. *21*, 2195–2204.

Chen, X., Stewart, E., Shelat, A.A., Qu, C., Bahrami, A., Hatley, M., Wu, G., Bradley, C., McEvoy, J., Pappo, A., et al.; St. Jude Children's Research Hospital–Washington University Pediatric Cancer Genome Project (2013). Targeting oxidative stress in embryonal rhabdomyosarcoma. Cancer Cell *24*, 710–724.

Davicioni, E., Anderson, M.J., Finckenstein, F.G., Lynch, J.C., Qualman, S.J., Shimada, H., Schofield, D.E., Buckley, J.D., Meyer, W.H., Sorensen, P.H., and Triche, T.J. (2009). Molecular classification of rhabdomyosarcoma–genotypic and phenotypic determinants of diagnosis: a report from the Children's Oncology Group. Am. J. Pathol. *174*, 550–564.

De Iaco, A., Planet, E., Coluccio, A., Verp, S., Duc, J., and Trono, D. (2017). DUX-family transcription factors regulate zygotic genome activation in placental mammals. Nat. Genet. *49*, 941–945.

Editorial (2015). The future of cancer genomics. Nat. Med. *21*, 99.

Donehower, L.A., Harvey, M., Slagle, B.L., McArthur, M.J., Montgomery, C.A., Jr., Butel, J.S., and Bradley, A. (1992). Mice deficient for p53 are developmentally normal but susceptible to spontaneous tumours. Nature *356*, 215–221.

Drummond, C.J., Hanna, J.A., Garcia, M.R., Devine, D.J., Heyrana, A.J., Finkelstein, D., Rehg, J.E., and Hatley, M.E. (2018). Hedgehog pathway drives fusion-negative rhabdomyosarcoma initiated from non-myogenic endothelial progenitors. Cancer Cell *33*, 108–124.e5.

Durbin, A.D., Somers, G.R., Forrester, M., Pienkowska, M., Hannigan, G.E., and Malkin, D. (2009). JNK1 determines the oncogenic or tumor-suppressive activity of the integrin-linked kinase in human rhabdomyosarcoma. J. Clin. Invest. *119*, 1558–1570.

El Demellawy, D., McGowan-Jordan, J., de Nanassy, J., Chernetsova, E., and Nasr, A. (2017). Update on molecular findings in rhabdomyosarcoma. Pathology *49*, 238–246.

Fleischmann, A., Jochum, W., Eferl, R., Witowsky, J., and Wagner, E.F. (2003). Rhabdomyosarcoma development in mice lacking Trp53 and Fos: tumor suppression by the Fos protooncogene. Cancer Cell *4*, 477–482.

Flex, E., Jaiswal, M., Pantaleoni, F., Martinelli, S., Strullu, M., Fansa, E.K., Caye, A., De Luca, A., Lepri, F., Dvorsky, R., et al. (2014). Activating mutations in RRAS underlie a phenotype within the RASopathy spectrum and contribute to leukaemogenesis. Hum. Mol. Genet. *23*, 4315–4327.

Gross, J.G., and Morgan, J.E. (1999). Muscle precursor cells injected into irradiated mdx mouse muscle persist after serial injury. Muscle Nerve *22*, 174–185.

Günther, S., Kim, J., Kostin, S., Lepper, C., Fan, C.M., and Braun, T. (2013). Myf5-positive satellite cells contribute to Pax7-dependent long-term maintenance of adult muscle stem cells. Cell Stem Cell *13*, 590–601.

Hendrickson, P.G., Doráis, J.A., Grow, E.J., Whiddon, J.L., Lim, J.W., Wike, C.L., Weaver, B.D., Pflueger, C., Emery, B.R., Wilcox, A.L., et al. (2017). Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. Nat. Genet. *49*, 925–934.

Hettmer, S., Liu, J., Miller, C.M., Lindsay, M.C., Sparks, C.A., Guertin, D.A., Bronson, R.T., Langenau, D.M., and Wagers, A.J. (2011). Sarcomas induced in discrete subsets of prospectively isolated skeletal muscle cells. Proc. Natl. Acad. Sci. USA *108*, 20002–20007.

Hewitt, J.E. (2015). Loss of epigenetic silencing of the DUX4 transcription factor gene in facioscapulohumeral muscular dystrophy. Hum. Mol. Genet. *24* (R1), R17–R23.

Hoadley, K.A., Yau, C., Wolf, D.M., Cherniack, A.D., Tamborero, D., Ng, S., Leiserson, M.D.M., Niu, B., McLellan, M.D., Uzunangelov, V., et al.; Cancer Genome Atlas Research Network (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell *158*, 929–944.

Irintchev, A., Rosenblatt, J.D., Cullen, M.J., Zweyer, M., and Wernig, A. (1998). Ectopic skeletal muscles derived from myoblasts implanted under the skin. J. Cell Sci. *111*, 3287–3297.

Jonkers, J., Meuwissen, R., van der Gulden, H., Peterse, H., van der Valk, M., and Berns, A. (2001). Synergistic tumor suppressor activity of BRCA2 and p53 in a conditional mouse model for breast cancer. Nat. Genet. *29*, 418–425.

Kim, J., and Braun, T. (2014). Skeletal muscle stem cells for muscle regeneration. Methods Mol. Biol. *1213*, 245–253.

Kim, J.B., Sebastiano, V., Wu, G., Araúzo-Bravo, M.J., Sasse, P., Gentile, L., Ko, K., Ruau, D., Ehrich, M., van den Boom, D., et al. (2009). Oct4-induced pluripotency in adult neural stem cells. Cell *136*, 411–419.

Leidenroth, A., and Hewitt, J.E. (2010). A family history of DUX4: phylogenetic analysis of DUXA, B, C and Duxbl reveals the ancestral DUX gene. BMC Evol. Biol. *10*, 364.

Leidenroth, A., Clapp, J., Mitchell, L.M., Coneyworth, D., Dearden, F.L., Iannuzzi, L., and Hewitt, J.E. (2012). Evolution of DUX gene macrosatellites in placental mammals. Chromosoma *121*, 489–497.

Lemmers, R.J., van der Vliet, P.J., Klooster, R., Sacconi, S., Camaño, P., Dauwerse, J.G., Snider, L., Straasheijm, K.R., van Ommen, G.J., Padberg, G.W., et al. (2010). A unifying genetic model for facioscapulohumeral muscular dystrophy. Science *329*, 1650–1653.

Lepper, C., Conway, S.J., and Fan, C.M. (2009). Adult satellite cells and embryonic muscle progenitors have distinct genetic requirements. Nature *460*, 627–631.

Li, R., Liang, J., Ni, S., Zhou, T., Qing, X., Li, H., He, W., Chen, J., Li, F., Zhuang, Q., et al. (2010). A mesenchymal-to-epithelial transition initiates and is required for the nuclear reprogramming of mouse fibroblasts. Cell Stem Cell *7*, 51–63.

Liu, C., Li, D., Hu, J., Jiang, J., Zhang, W., Chen, Y., Cui, X., Qi, Y., Zou, H., Zhang, W., and Li, F. (2014). Chromosomal and genetic imbalances in Chinese patients with rhabdomyosarcoma detected by high-resolution array comparative genomic hybridization. Int. J. Clin. Exp. Pathol. *7*, 690–698.

Madissoon, E., Jouhilahti, E.M., Vesterlund, L., Töhönen, V., Krjutškov, K., Petropoulos, S., Einarsdottir, E., Linnarsson, S., Lanner, F., Månsson, R., et al. (2016). Characterization and target genes of nine human PRD-like homeobox domain genes expressed exclusively in early embryos. Sci. Rep. *6*, 28995.

Morrison, S.J., and Spradling, A.C. (2008). Stem cells and niches: mechanisms that promote stem cell maintenance throughout life. Cell *132*, 598–611.

Peinado, H., Zhang, H., Matei, I.R., Costa-Silva, B., Hoshino, A., Rodrigues, G., Psaila, B., Kaplan, R.N., Bromberg, J.F., Kang, Y., et al. (2017). Pre-metastatic niches: organ-specific homes for metastases. Nat. Rev. Cancer *17*, 302–317.

Polotskaia, A., Xiao, G., Reynoso, K., Martin, C., Qiu, W.G., Hendrickson, R.C., and Bargonetti, J. (2015). Proteome-wide analysis of mutant p53 targets in breast cancer identifies new levels of gain-of-function that influence PARP, PCNA, and MCM4. Proc. Natl. Acad. Sci. USA *112*, E1220–E1229.

Polyak, K., and Weinberg, R.A. (2009). Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits. Nat. Rev. Cancer *9*, 265–273.

Rickard, A.M., Petek, L.M., and Miller, D.G. (2015). Endogenous DUX4 expression in FSHD myotubes is sufficient to cause cell death and disrupts RNA splicing and cell migration pathways. Hum. Mol. Genet. *24*, 5901–5914.

Robertson, A.G., Shih, J., Yau, C., Gibb, E.A., Oba, J., Mungall, K.L., Hess, J.M., Uzunangelov, V., Walter, V., Danilova, L., et al. (2017). Integrative analysis identifies four molecular and clinical subsets in uveal melanoma. Cancer Cell *32*, 204–220.e15.

Roychowdhury, S., and Chinnaiyan, A.M. (2016). Translating cancer genomes and transcriptomes for precision oncology. CA Cancer J. Clin. *66*, 75–88.

Sato, Y., Yoshizato, T., Shiraishi, Y., Maekawa, S., Okuno, Y., Kamura, T., Shimamura, T., Sato-Otsubo, A., Nagae, G., Suzuki, H., et al. (2013). Integrated molecular analysis of clear-cell renal cell carcinoma. Nat. Genet. *45*, 860–867.

Shern, J.F., Chen, L., Chmielecki, J., Wei, J.S., Patidar, R., Rosenberg, M., Ambrogio, L., Auclair, D., Wang, J., Song, Y.K., et al. (2014). Comprehensive genomic analysis of rhabdomyosarcoma reveals a landscape of alterations affecting a common genetic axis in fusion-positive and fusion-negative tumors. Cancer Discov. *4*, 216–231.

Shima, N., Alcaraz, A., Liachko, I., Buske, T.R., Andrews, C.A., Munroe, R.J., Hartford, S.A., Tye, B.K., and Schimenti, J.C. (2007). A viable allele of Mcm4 causes chromosome instability and mammary adenocarcinomas in mice. Nat. Genet. *39*, 93–98.

Soini, Y., Kosma, V.M., and Pirinen, R. (2015). KDM4A, KDM4B and KDM4C in non-small cell lung cancer. Int. J. Clin. Exp. Pathol. *8*, 12922–12928.

Steward, C.A., Parker, A.P.J., Minassian, B.A., Sisodiya, S.M., Frankish, A., and Harrow, J. (2017). Genome annotation for clinical genomic diagnostics: strengths and weaknesses. Genome Med. *9*, 49.

Taulli, R., Scuoppo, C., Bersani, F., Accornero, P., Forni, P.E., Miretti, S., Grinza, A., Allegra, P., Schmitt-Ney, M., Crepaldi, T., and Ponzetto, C. (2006). Validation of met as a therapeutic target in alveolar and embryonal rhabdomyosarcoma. Cancer Res. *66*, 4742–4749.

Tomasetti, C., and Vogelstein, B. (2015). Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Science *347*, 78–81.

Tremblay, A.M., Missiaglia, E., Galli, G.G., Hettmer, S., Urcia, R., Carrara, M., Judson, R.N., Thway, K., Nadal, G., Selfe, J.L., et al. (2014). The Hippo transducer YAP1 transforms activated satellite cells and is a potent effector of embryonal rhabdomyosarcoma formation. Cancer Cell *26*, 273–287.

Wang, Y., Marino-Enriquez, A., Bennett, R.R., Zhu, M., Shen, Y., Eilers, G., Lee, J.C., Henze, J., Fletcher, B.S., Gu, Z., et al. (2014). Dystrophin is a tumor suppressor in human cancers with myogenic programs. Nat. Genet. *46*, 601–606.

Whiddon, J.L., Langford, A.T., Wong, C.J., Zhong, J.W., and Tapscott, S.J. (2017). Conservation and innovation in the DUX4-family gene network. Nat. Genet. *49*, 935–940.

Williamson, D., Missiaglia, E., de Reyniès, A., Pierron, G., Thuille, B., Palenzuela, G., Thway, K., Orbach, D., Laé, M., Fréneaux, P., et al. (2010). Fusion gene-negative alveolar rhabdomyosarcoma is clinically and molecularly indistinguishable from embryonal rhabdomyosarcoma. J. Clin. Oncol. *28*, 2151–2158.

Wüst, S., Dröse, S., Heidler, J., Wittig, I., Klockner, I., Franko, A., Bonke, E., Günther, S., Gärtner, U., Boettger, T., et al. (2018). Metabolic Maturation during Muscle Stem Cell Differentiation Is Achieved by miR-1/133a-Mediated Inhibition of the Dlk1-Dio3 Mega Gene Cluster. Cell Metab. *27*, 1026–1039.e6.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Anti-MyoD1 rabbit polyclonal antibody	Abcam	Cat# ab64159,RRID:AB_2266875
Anti-MyoD1 mouse monoclonal antibody	LSBio	Cat# LS-C9179-500, RRID:AB_835256
Anti-Myogenin rabbit polyclonal antibody	Sigma Aldrich	Cat# HPA038093,RRID:AB_10674546
Anti-Myogenin mouse monoclonal antibody	BD Bioscience	Cat# 556358, RRID:AB_396383
Anti-MF20 mouse monoclonal antibody	DSHB	Cat# MF 20, RRID:AB_2147781
Anti-Desmin rabbit polyclonal antibody	Sigma Aldrich	Cat# D8281, RRID:AB_476910
Anti-γH2AX rabbit polyclonal antibody	Cell Signaling Tech.	Cat# 2595, RRID:AB_10694556
Anti-ATM rabbit polyclonal antibody	Cell Signaling Tech.	Cat# 2851S, RRID:AB_330318
Anti-53bp1 rabbit polyclonal antibody	Abcam	Cat# ab36823, RRID:AB_722497
Anti-P/E Cadherin mouse monoclonal antibody	BD	Cat# 610182, RRID:AB_397581
Anti-V5 rabbit polyclonal antibody	Abcam	Cat# ab9116, RRID:AB_307024
Anti-V5 mouse monoclonal antibody	Invitrogen	Cat# 37-7500, RRID:AB_2533339
Anti-p53 rabbit polyclonal antibody	Leica	Cat# NCL-p53-CM5p, RRID:AB_563933
Bacterial and Virus Strains		
pMD2.G plasmid	Didier Trono Lab	Addgene # 12259
psPAX2 plasmid	Didier Trono Lab	Addgene # 12260
duxblV5 plasmid	Vectorbuilder	N/A
Yap1 shrna#1: GCAGACAGATTCCTTTGTTAA	Sigma Aldrich	N/A
Yap1 shrna#2: CCACCAAGCTAGATAAAGAAA	Sigma Aldrich	N/A
Yap1 shrna#3: CGGTTGAAACAACAGGAATTA	Sigma Aldrich	N/A
Yap1 shrna#4: GCGGTTGAAACAACAGGAATT	Sigma Aldrich	N/A
Yap1 shrna#5: CTGGTCAAAGATACTTCTTAA	Sigma Aldrich	N/A
duxbl shrna#1: GCAGGATAAACCTAGAGTTAA	Sigma Aldrich	N/A
duxbl shrna#2: GCTGAATGGATGCCTGACAAA	Sigma Aldrich	N/A
duxbl shrna#3: GCTTCAGTTATACTGCCTCTT	Sigma Aldrich	N/A
duxbl shrna#4: CCGCGCTTAGAAGATTGTACT	Sigma Aldrich	N/A
Scrambled shrna: CCTAAGGTTAAGTCGCCCT CGCTCGAGCGAGGGCGACTTAACCTTAGG	Sigma Aldrich	N/A
Stbl3 Chemically Competent *E. coli*	Invitrogen	Cat# C737303
Chemicals, Peptides, and Recombinant Proteins		
Tamoxifen	Sigma Aldrich	Cat# T5648
Dispase	BD	Cat# 354235
Collagenase, Type 2	Worthington Biochemicals	Cat# CLS-2
Percoll	Sigma Aldrich	Cat# P1644
Matrigel Matrix	BD	Cat# 356234
Trizol reagent	Invitrogen	Cat# 15596026
Critical Commercial Assays		
Click-iT EdU Kit	Invitrogen	Cat# C10337
SuperScript II Reverse Transcriptase Kit	Invitrogen	Cat# 18091050
DNeasy Blood & Tissue Kit	QIAGEN	Cat# 69504
SureSelect Mouse All Exon Kit	Agilent	Cat# G7550
Deposited Data		
Exome-Seq data	This paper	ENA: PRJEB23461
RNA-Seq data QSC, ASC, DSC	Wüst et al. (2018)	GEO: GSM2888361

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
RNA-Seq data Duxbl overexpression	This paper	ENA: PRJEB23461
RNA-Seq data Chen et al. (2013)	Chen et al. (2013)	ENA: EGAD00001000878
Videos of tumor cells subjected to shrna mediated inactivation of amplified oncogenes	This paper	https://doi.org/10.17632/7g2pbbrn4m.1
Experimental Models: Cell Lines		
HEK293FT	ATCC	Cat# PTA5077
Experimental Models: Organisms/Strains		
$p53^{loxP/loxP}$ mice	Jackson Laboratory	Stock No: 008462
$Rosa26^{Tomato}$ mice	Jackson Laboratory	Stock No: 007914
mdx mice	Jackson Laboratory	Stock No: 001801
mdx-nude mice	Dr. Jennifer Morgan	N/A
Pax7::ZsGreen mice	Dr. Michael Kyba	N/A
$Pax7^{CE}$ mice	Dr. Chenming Fan	N/A
Software and Algorithms		
R language (v3.4.1)	NA	http://www.r-project.org
GraphPad Prism 7	GraphPad Software	N/A
STAR(v2.5.2b)	N/A	https://bioconda.github.io/
Picard (v1.119)	N/A	https://bioconda.github.io/
FreeC (version 10.5)	N/A	https://bioconda.github.io/
Heatmaps and Circos (v0.69-3)	N/A	http://circos.ca/
Arraystar v.14	N/A	N/A
Gencode (version vM11)	N/A	https://www.gencodegenes.org/
DBsnp (version 142)	N/A	https://www.ncbi.nlm.nih.gov/projects/SNP/
VEP (version 90.6)	N/A	https://bioconda.github.io/
RSEM (v1.2.30)	N/A	https://bioconda.github.io/
GATK (v3.2.2)	N/A	https://software.broadinstitute.org/gatk
Oligonucleotides		
Oligonucleotides used for genotyping and mRNA expression analysis are provided in Table S7	N/A	N/A

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and reasonable requests for reagents may be directed to and will be fulfilled by the Lead Contact, Johnny Kim (johnny.kim@mpi-bn.mpg.de)

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Mice

All mice used in this study were bred on a C57BL/6 background, drug and test naive, healthy prior to the studies, not used in previous procedures and maintained in a barrier facility. Except for mdx-nude mice used for transplantation studies all mice were immune competent. Female and male animals to equal proportions were analyzed in this study. None of the determined parameters in this study correlated with animal sex. Genotypes of all animals and as indicated in this study were determined using verified protocols with DNA isolated from tail snip tissue biopsies collected upon weaning. All mice were mice were separated by sex and maintained with 4-5 mice per cage in low-noise, filtered ventilated cage racks. The $p53^{loxP/loxP}$ mouse strain was obtained from The Jackson Laboratory (6.129P2-Trp53tm1Brn/J) and described previously(Jonkers et al., 2001). B6.Cg-Gt(ROSA)26Sor tm14CAG-tdTomato) Hze/J ($Rosa26^{Tomato}$), C57BL/10ScSn-Dmdmdx/J (mdx) mouse strain were obtained from The Jackson Laboratory (Bar Harbor, ME). The mdx-nude strain was a kind gift from Jennifer Morgan and was described previously (Gross and Morgan, 1999). The generation of $Pax7^{CE}$ and Pax7::ZsGreen mice has been described previously (Bosnakovski et al., 2008b; Lepper et al., 2009). Primers used for genotyping are listed in the Supplemental Information. Tamoxifen (Sigma) was administered to two-month old mice intraperitoneally at 2mg per 40 g body weight per injection. Cardiotoxin (0.06 mg/ml, Sigma) was injected into tibialis anterior muscles in a volume of 50 μl. FACS purified TPCs ($1x10^5$) were injected intramuscularly into anesthesized mdx-nude mice as described in (Gross and Morgan, 1999). All animal experiments were done in accordance with the Guide for the Care and Use of Laboratory Animals

published by the US National Institutes of Health (NIH Publication No. 85-23, revised 1996) and according to the regulations issued by the Committee for Animal Rights Protection of the State of Hessen (Regierungspraesidium Darmstadt).

Cell culture

Satellite cell and tumor cell purification were performed according to established methods (Kim and Braun, 2014). Briefly limb and trunk muscles or surgically excised tumors were minced, digested with 100 CU Dispase (BD) and 0.2% type II collagenase (Worthington Biochemicals), and consecutively filtered through 100 μm, 70 μm, and 40 μm cell strainers (BD). Cells were applied to a discontinuous Percoll gradient consisting of 70% Percoll overlayed with 30% v/v Percoll. Mononuclear cells were collected at the 70/30 interphase and subjected to FACS (BD FACSAriaIII) using GFP fluorescence (for Pax7::ZsGreen) and/or RFP fluorescence for lineage traced Rosa26^{TOM} cells. FACS purified SCs and TPCs were cultured on Matrigel-coated μClear plates (BD Biosciences, Greiner) in DMEM medium with 20% FCS and bFGF (5ng/ml). EdU incorporation assay was performed by adding EdU with a final concentration of 10 μM 3 hours before fixation and then analyzed using the Click-iT EdU kit (Invitrogen) according to the manufacturer's protocol. Antibodies for immunohistochemical staining are listed in the Key Resources Table. Time lapse imaging and analysis was performed using an Incucyte Live-Cell Imaging System and software (Essen Instruments). In population doubling assays SCs were seeded at equal density (10^2/well), images were taken every hour and growth rates were calculated as a percentage of cell confluence per image and over time. Every two days SCs were trypsinized, undifferentiated SCs recollected, replated at equal densities and continuously reanalyzed as above.

METHOD DETAILS

mRNA expression analysis

Total RNA was isolated using RNeasy Micro Kit (QIAGEN) or Trizol reagent (Invitrogen) according to the manufacturers' protocols. Purified RNA was subjected to reverse transcriptase reaction in the presence of 25 ng/ml random primers and 2.5 mM dA/C/G/TTP with 10 U/ml SuperScript II Reverse Transcriptase (Invitrogen). Results were normalized to gapdh or m36b4 expression. Primers used for RT-qPCR are listed in the Key Resources Table.

Immunofluorescence and immunohistochemistry

Cells and tissue sections were fixed in 4% paraformaldehyde and blocked in PBS containing 5% BSA (Millipore) and 0.1%. Triton X-100 for 1 hour at room temperature, incubated with primary antibodies overnight at 4°C. Samples were then washed and incubated with secondary antibodies for at room temperature, washed, coverslipped and imaged on confocal (Leica), MetaXpress (Molecular Devices), or Axioimager (Zeiss) microscopes.

The list of antibodies used in this study is shown in Key Resources Table. Representative images for were combined into a single panel by adjusting the scale (Figures 1C, 4E, and 4F).

Knockdown by shRNA transduction

Production of lentivirus encoding shrnas was performed by $Ca_3(PO4)_2$ transfection of HEK293T cells with helper plasmids pMD2.G and psPAX2. Isolated satellite cells or tumor cells were seeded into tissue culture plates freshly coated with Matrigel (Greiner). 24 hours later, cells were transduced with lentiviral supernatants supplemented with 8μg/ml polybrene for 12 hours. After media exchange, cells were further incubated in growth media and imaged using am Incucyte live-imager (Essen Bioscience) over time or fixed in 4% PFA and whole well images were acquired and analyzed using an ImageXpress Micro automated high-throughput fluorescence microscope and MetaXpress software (Molecular Devices).

Exome sequencing, bioinformatics

Total genomic DNA from purified cells and isolated tissues was purified with DNeasy Blood & Tissue Kit (QIAGEN) and quantified by Qubit and NanoDrop measurement. Volume of samples with low concentrations (< 25ng/μl) was reduced by SpeedVac. 50ng of genomic DNA was used as input for SureSelect QXT library (Agilent) preparation using the standard protocol. Successful pre-hybridization library preparation was followed by Qubit and Labchip GX touch measurement for quantity and insert size. 200-1.500ng of library was used for hybridization with SureSelect Mouse All Exon Kit (Agilent) to enrich for Exome-containing library elements. Sequencing was performed on the NextSeq500 instrument (Illumina) using v2 chemistry, resulting in average of 30M reads per library with 2x300bp paired end setup. Raw reads were mapped against the mouse genome (mm10) using STAR (v2.5.2b), and alignments were deduplicated using Picard (v1.119), effectively removing PCR artifacts that lead to multiple copies of the same original fragment. Deduplicated input alignments of matched normal and tumor samples were used for analysis of copy number variations using FreeC (version 10.5) with capture regions from the SureSelect Mouse All Exon Kit (Agilent) as target windows. We obtained copy number ratios for targeted exons, as well as the median ratio per segmented amplification (spanning multiple target exons) from the FreeC output, which were calculated and normalized to log2 copy number values (log2CN). A segment was selected for further investigation after fulfillment of the following criteria: median log2CN was above 3, the segment contained exons from at least 5 consecutive genes, and the total spanning exonic length was greater than 5kB. Maximum log2CN values for a gene within a selected segment were used during visualization with Heatmaps and Circos (v0.69-3). Additionally, deduplicated input alignments were quantified using Arraystar v.14 and Qseq. Weighted RPKM-CN values wherein control and tumor samples from always the same specimen was used to

compare normalize read counts subsequently visualized using Perseus v1.6.0.7. Genes with read counts below the detection threshold were excluded from downstream analyses to eliminate low coverage exons. Analysis of single nucleotide polymorphisms (SNPs) closely followed the GATK best practices. Briefly, deduplicated input alignments were realigned to all exonic sequences from Gencode (version vM11) taking known variants from DBsnp into account. After base recalibration, SNP calling was performed within exons, allowing a padding of 100bp into flanking introns. Variant calls were annotated to known SNPs from DBsnp (version 142) and functional relevance of variant calls was predicted using VEP (version 90.6) using the Gencode annotation (version vM11) as source.

Analysis of public data

Raw data from Chen et al. (2013) was downloaded and mapped to the human genome (hg38) and transcriptome using STAR (v2.5.2b) with transcript annotations from Gencode (version 26). Gene expression was quantified as FPKMs using RSEM (v1.2.30). For downstream analysis and visualization, log2 FPKM values were centered to the mean and scaled to union standard deviation to obtain relative expression estimates across the cohort. For analysis of the TCGA data, metadata (including ICD10 diagnosis terms) from 11574 TCGA datasets across all primary sites with available raw counts were downloaded from the GDC Data Portal. In addition, slices from TCGA raw alignment data (dbgap project 11430, Validation of genomic mutations in human that result in tumor formation in mouse cancers) querying the genomic locations of DuxB (chr16:75690000-75710000) and Dux4 (chr4:190173774-190185911) were re-counted with HTSeq-Count and a custom annotation for DuxB and Dux4. Raw counts from TCGA sequencing data including DuxB and Dux4 were merged and normalized using the estimateSizeFactorsForMatrix method from DESeq2. Similar to data from Chen et al. (2013), complete expression data was centered to the mean and scaled to uniform standard variation.

QUANTIFICATION AND STATISTICAL ANALYSIS

Animal studies were performed without blinding and no animals were excluded from the analysis. For animal studies, a power test was used to estimate the sample size needed to observe a minimum of 2-fold difference in mean between groups with 0.8 power. All assays were repeated at least three times. Sample size for *in vitro* studies was chosen based on observed effect sizes and standard errors. Two-tailed unpaired t tests were performed using the GraphPad Prism 5.0a (GraphPad Software) program to determine statistical significance between groups. Error bars indicate SEM; p values of $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, and $^{****}p < 0.0001$ were considered to be statistically significant.

DATA AND SOFTWARE AVAILABILITY

All datasets generated and/or analyzed during the current study are presented in this published article or the accompanying Source Data or Supplemental Information files or are available from the corresponding authors upon reasonable request.

The accession numbers for the genomic datasets reported in this paper are ENA: PRJEB23461, ENA: EGAD00001000878, and GEO: GSM2888361.

Videos are deposited at https://doi.org/10.17632/7g2pbbrn4m.1

Cell Stem Cell, Volume *23*

Supplemental Information

Oncogenic Amplification of Zygotic Dux Factors in Regenerating p53-Deficient Muscle Stem Cells Defines a Molecular Cancer Subtype

Jens Preussner, Jiasheng Zhong, Krishnamoorthy Sreenivasan, Stefan Günther, Thomas Engleitner, Carsten Künne, Markus Glatzel, Roland Rad, Mario Looso, Thomas Braun, and Johnny Kim

Supplemental Table S4. ICD-10 classification of DUX/ZGA positive cancer patients. (Related to Figure 3)

ICD10		Age						Gender		Cluster			
diagnosis	neoplasm	<20	20–30	30–50	50–70	70–90	<NA>	female	male	1	2	3	4
c50	breast	0	1	8	32	8	0	48	1	1	2	4	42
c62	testis	5	15	14	1	0	0	0	35	25	9	0	1
c64	kidney	0	0	6	18	11	0	10	25	0	19	3	13
c16	stomach	0	0	6	17	11	0	9	25	4	5	5	20
c34	bronchus and lung	0	0	2	19	5	2	10	18	2	21	0	5
c37	thymus	0	0	6	14	5	1	10	16	3	18	5	0
c22	liver	0	0	5	14	3	0	9	13	0	22	0	0
c54	corpus uteri	0	0	2	11	5	1	19	0	3	9	5	2
c32	larynx	0	0	4	14	0	0	5	13	3	12	1	2
c15	esophagus	0	0	2	6	3	0	2	9	1	7	1	2

Supplemental Table S7. Oligonucleotides used in this study (Related to STAR Methods)

Genotyping Oligonucleotides		
Oligonucleotide sequence	**Purpose**	**Source**
ACTAGGCTCCACTCTGTCCTTC	Pax7 ^{CreERT2} Forward	Lepper et al., Nature 2009
GCAGATGTAGGGACATTCCAGTG	Pax7 ^{CreERT2} Reverse	Lepper et al., Nature 2009
CTGCATGTACCACGAGTCCA	ZsGreen Forward	this paper
GTCAGCTGCCACTTCTGGTT	ZsGreen Reverse	this paper
AAAGTCGCTCTGAGTTGTTAT	Rosa26 RosaFA	Jackson Laboratories
GGAGCGGGAGAAATGGATATG	Rosa26 RosaRF	Jackson Laboratories
CATCAAGGAAACCCTGGACTACTG	Rosa26 Rosa-SpliAC	Jackson Laboratories
GCGCGAAACTCATCAAATATGCGTG TTAGTGT	mdx Forward	Zhang et al., Nature Comm. 2015
GATACGCTGCTTTAATGCCTTTAGTC ACTCAGATAGTTGAAGCCATTTTG	mdx WT Reverse	Zhang et al., Nature Comm. 2015
CGGCCTGTCACTCAGATAGTTGAAG CCATTTTA	mdx MT Reverse	Zhang et al., Nature Comm. 2015
CTGTTCCTGTACGGCATGG	Tomato Forward	Jackson Laboratories
GGCATTAAAGCAGCGTATCC	Tomato Reverse	Jackson Laboratories
CACAAAAACAGGTTAAACCCAG	$p53^{loxp}/p53^{\Delta}$	Jonkers et al., Nature Genetics 2001
AGCACATAGGAGGCAGAGAC	$p53^{loxp}/p53^{\Delta}$	Jonkers et al., Nature Genetics 2001
GAAGACAGAAAAGGGGAGGG	$p53^{loxp}/p53^{\Delta}$	Jonkers et al., Nature Genetics 2001
RT-qPCR Oligonucleotides		
Oligonucleotide sequence	**Purpose**	**Source**
CTCTCCCCCGCAAAAGAAAAA	trp53 Forward	Zhang et al., Nature Comm. 2015
CGGAACATCTCGAAGCGTTTA	trp53 Reverse	Zhang et al., Nature Comm. 2015
TGCTGTGCAATTAAAGGCTGT	trp53 Forward	Zhang et al., Nature Comm. 2015
CGTGTTCTCCGAGATACTTGGT	trp53 Reverse	Zhang et al., Nature Comm. 2015
CGGTGTCAGAGTCTAGGGGA	cdkn1a (p21) Forward	Zhang et al., Nature Comm. 2015
ATTGGAGTCAGGCGCAGATC	cdkn1a (p21) Reverse	Zhang et al., Nature Comm. 2015
CCACCTCCAACTGCTCTGAC	myf5 Forward	Zhang et al., Nature Comm. 2015
GCTTCAGGGCTTCTTTTCCT	myf5 Reverse	Zhang et al., Nature Comm. 2015
GAATGGCTACGACACCGCCTACTAC	myoD Forward	Zhang et al., Nature Comm. 2015
CCTACGGTGGTGCGCCCTCTGC	myoD Reverse	Zhang et al., Nature Comm. 2015
TTGCTCAGCTCCCTCAACCA	myogenin Forward	Zhang et al., Nature Comm. 2015
TGGGCTGGGTGTTAGTCTTA	myogenin Reverse	Zhang et al., Nature Comm. 2015
AGATTCGGGATATGCTGTTGGC	m36b4 Forward	Zhang et al., Nature Comm. 2015
TCGGGTCCTAGACCAGTGTTC	m36b4 Reverse	Zhang et al., Nature Comm. 2015
GTGAAGGTCGGTGTGAACG	gapdh Forward	Judson et al., JCS 2012
ATTTGATGTTAGTGGGGTCTCG	gapdh Reverse	Judson et al., JCS 2012
ACCCTCGTTTTGCCATGAAC	yap1 Forward	Sakabe et al.,PNAS 2017
TGTGCTGGGATTGATATTCCGTA	yap1 Reverse	Sakabe et al.,PNAS 2017
GCATCTCTGAGTCTCAAATTATGACT TG	duxbl Forward	this paper
GCGTTCTGCTCCTTCTAGCTTCT	duxbl Reverse	this paper

Supplemental Figures and Legends

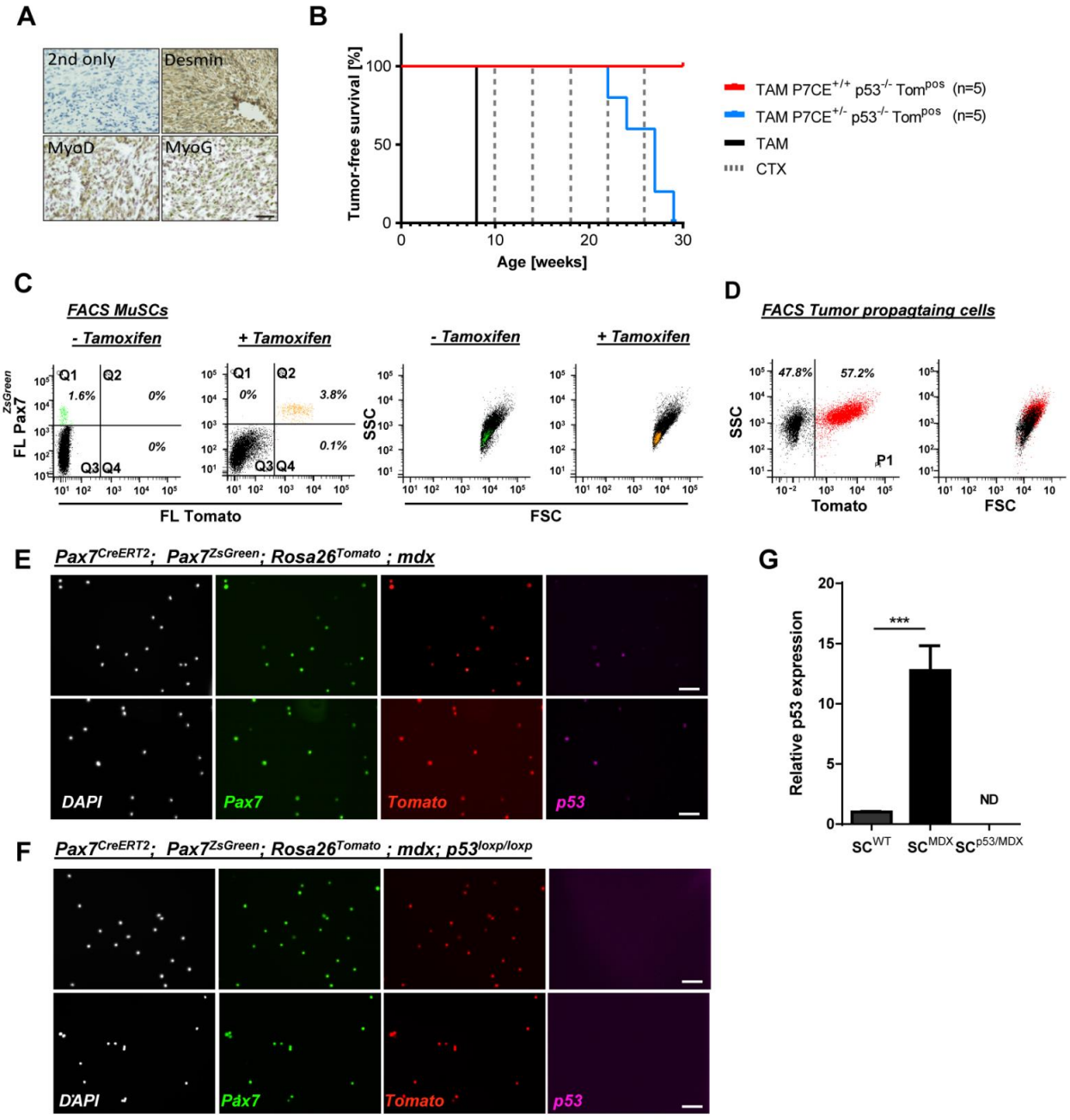


Figure S1, related to Figure 1. A regenerative environment is necessary to induce RMS formation upon muscle SC-specific loss of p53.

A) Representative immunohistochemical staining of cross-sectioned tumor with indicated antibodies. **B)** Kaplan-Meier tumor-free survival curves are shown for indicated genotypes. Solid black line indicates timing of tamoxifen administration. Dashed grey lines indicate timing of CTX induced muscle injury. **C-D)** FACS plots depicting efficient separation of muscle SCs (C) and TPCs (D) via fluorescence isolated from $SC^{p53/MDX}$ mice. Note that in (C) virtually all SCs are lineage-traced through activation of the Tomato reporter upon Tamoxifen treatment. **E-F)** Immunohistochemical staining of freshly FACS sorted SCs from indicated genotypes. Note that all $Pax7^{Zsgreen}$ SCs are lineage traced but only a subset expresses p53 in SC^{MDX} mice (E) but not in $SC^{p53/MDX}$ mice (F). **G)** RT-qPCR analysis of p53 in freshly sorted SCs from SC^{WT}, SC^{MDX} and $SC^{p53/MDX}$ mice. Expression levels were normalized to m36b4 mRNA. Error bars indicate standard deviation of the mean (t-test: P***<0.001, n=3, ND = not determinable).

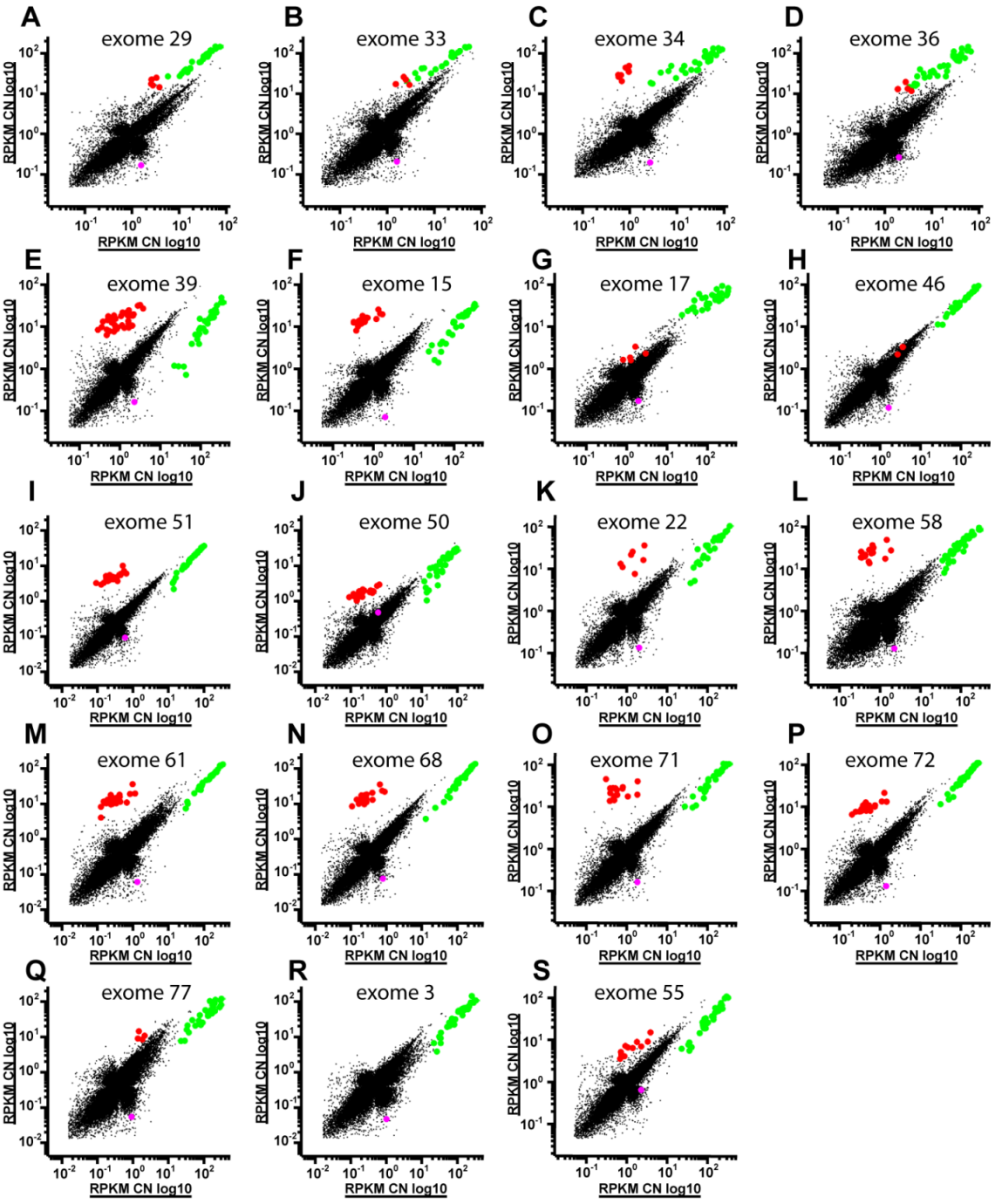


Figure S2, related to Figure 2. Identification of distinct copy number amplifications in purified tumor cells.

(A-S) Scatter plots depicting log-scaled RPKM values of genomic DNA in purified tumor cells (y-axis) vs liver control (x-axis). Green circles represent mitochondria encoded genes. Red circles represent amplified genes. Magenta circle represents p53. All CN values for each individual tumor are provided in Supplementary Table 1.

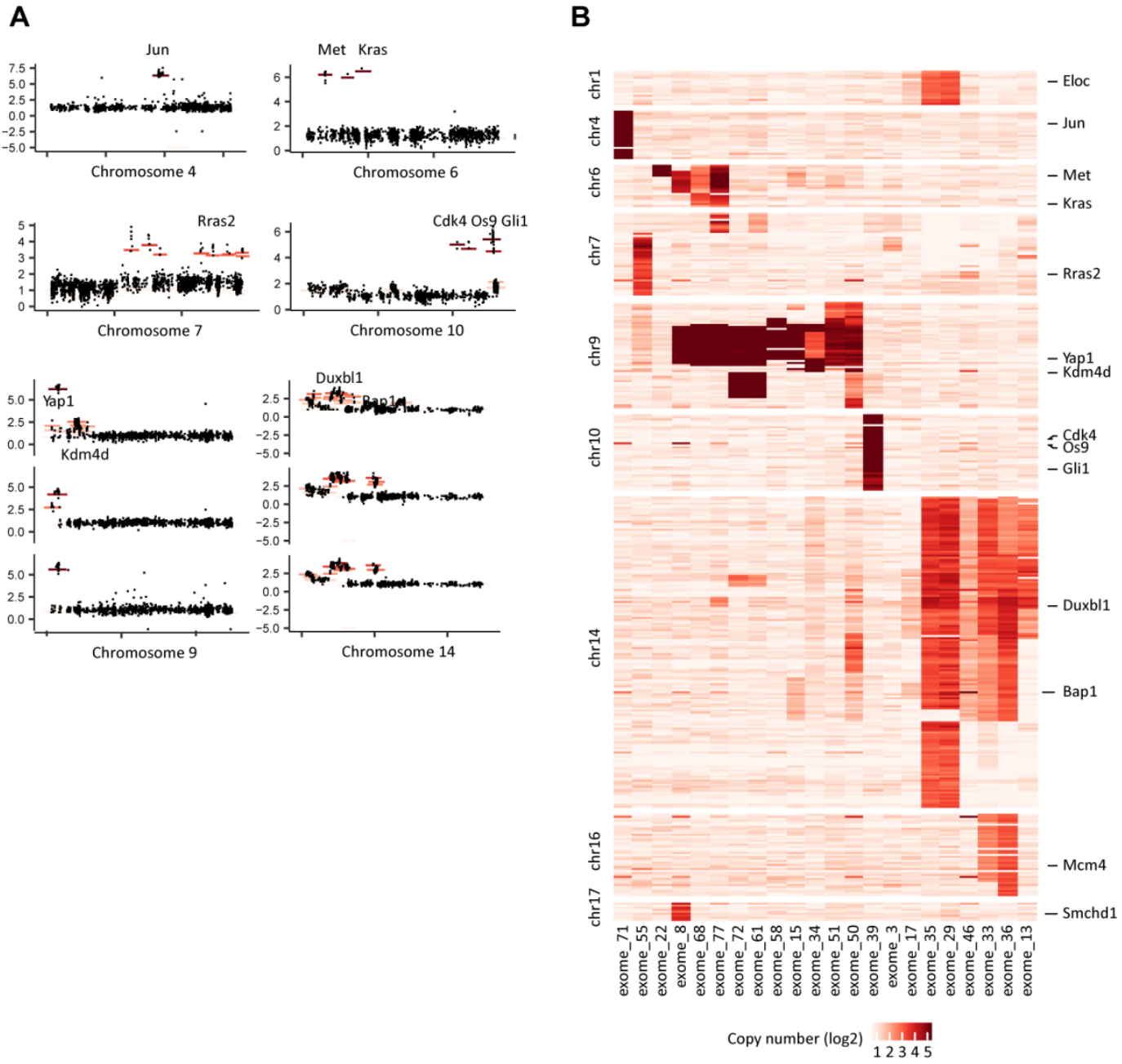


Figure S3, related to Figure 2. Positional analysis of copy number amplifications.

(A) Distribution of log2 scaled copy number values for amplified regions. **(B)** Positional heat map of amplified genes in all analyzed tumors. Note that genomic regions are displayed in physical order.

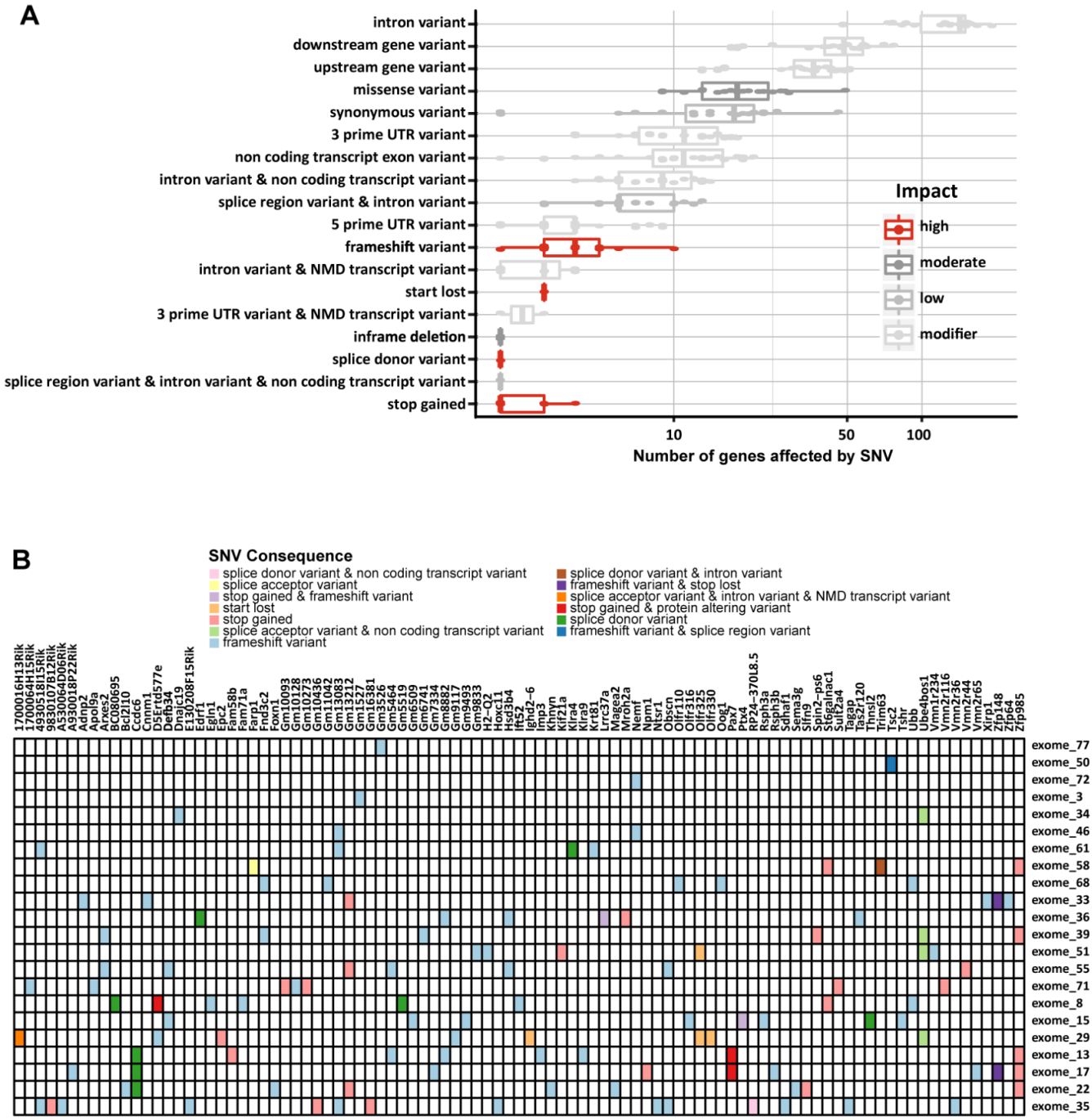
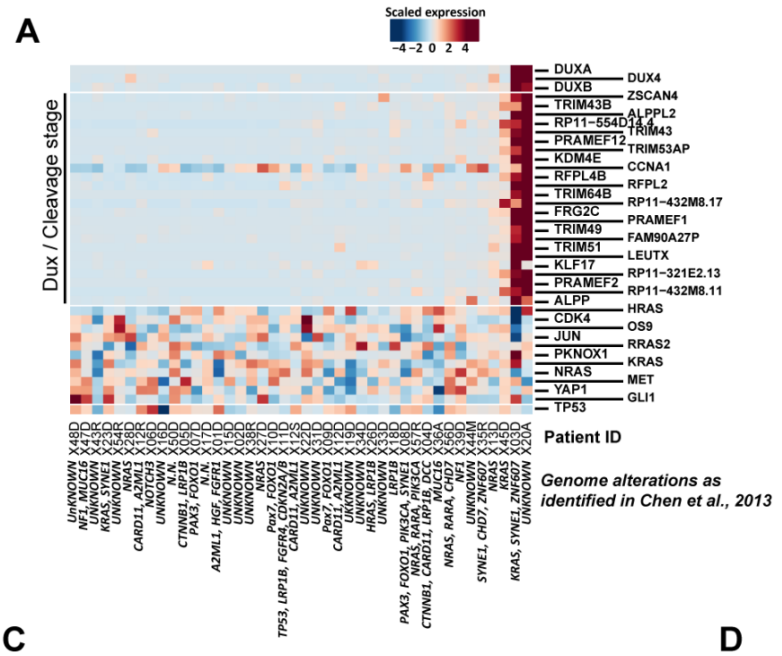


Figure S4, related to Figure 2. Mutational analysis.

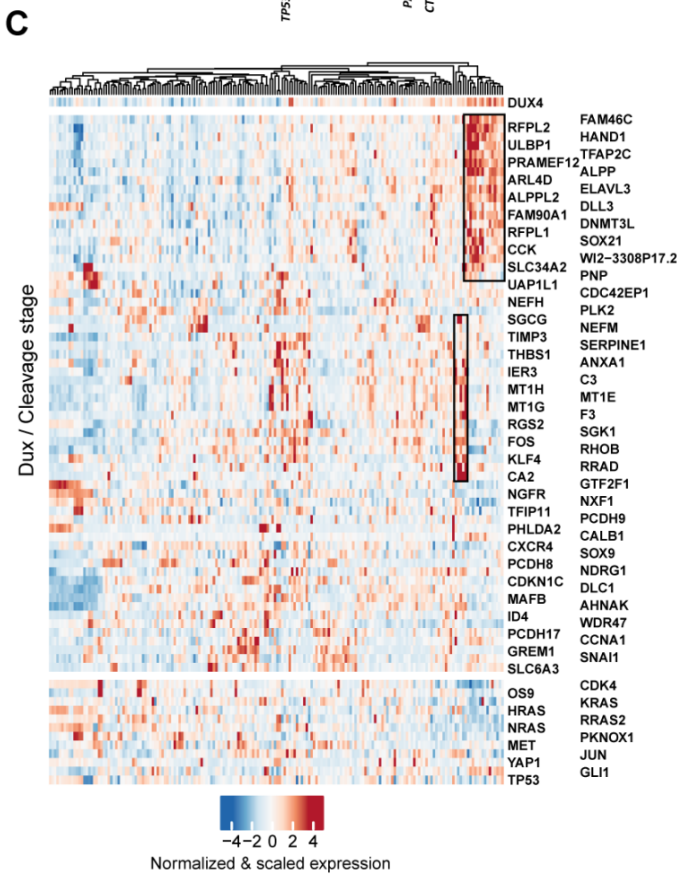
(A) Mutational load of purified TPCs. SNVs with high functional impact are highlighted in red.

(B) Summary of SNVs in the discovery cohort. All detected SNVs for each individual tumor are provided in Supplementary Table 2.



B

Patient ID	Common Dux4/2C	Stage	Risk
X03D	63.1	1	Low
X20A	57.7	4	High
X45D	16.1	3	Intermediate
X13D	14.8	3	Intermediate



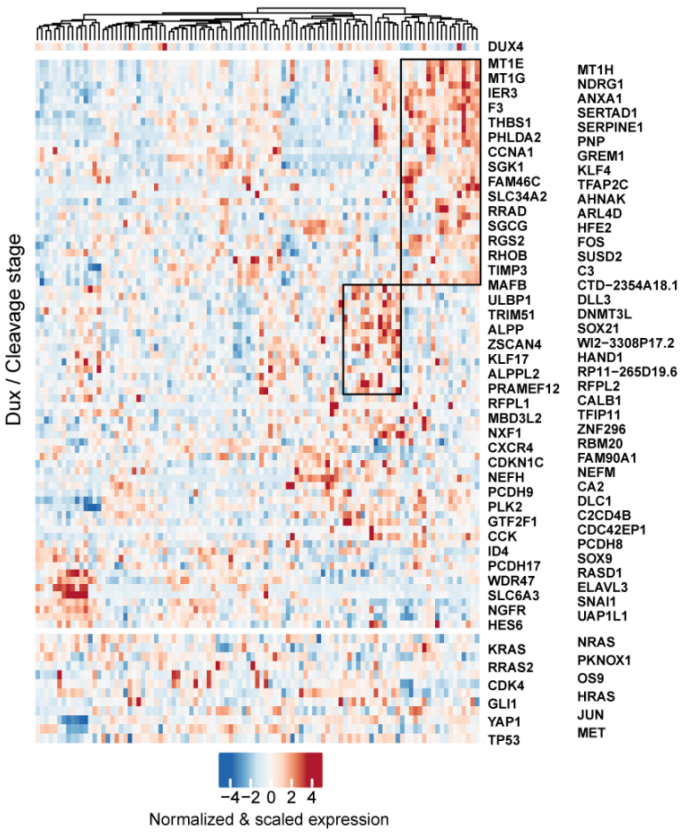


Figure S5, related to Figure 4. Dux transcription factors define a molecular subtype of ERMS.

(A) mRNA signatures of Dux dependent zygotic gene activation and common oncogenic drivers across 42 ERMS tumors from Chen et al., (Cancer Cell, 2013) **(B)** Pathological characterization of Dux factor and/or ZGA positive ERMS tumors and degree of ZGA. **(C)** mRNA signatures of Dux-dependent zygotic gene activation and common oncogenic drivers across 124 human RMS tumor samples from Davicioni et al. (J Clin Oncol., 2010). **(D)** mRNA signatures of Dux-dependent zygotic gene activation and common oncogenic drivers across 101 human tumor samples from Williamson et al. (J Clin Oncol., 2010).

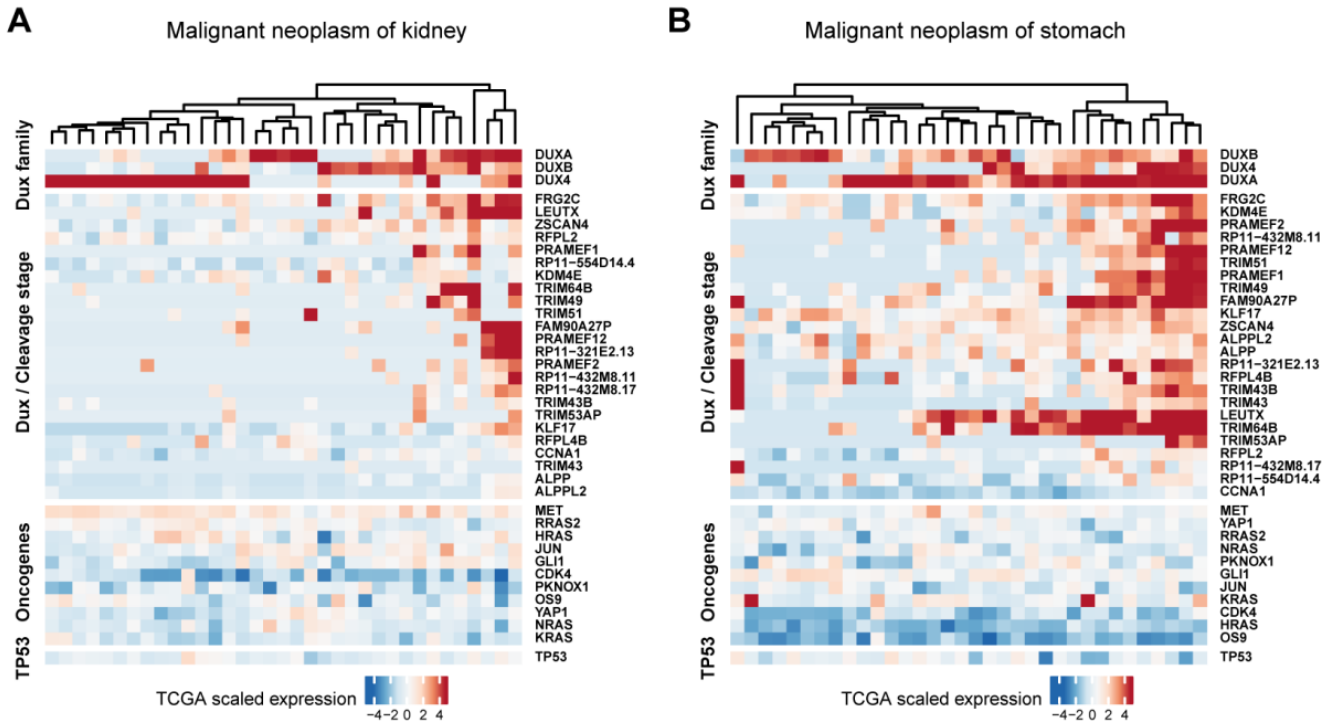


Figure S6, related to Figure 4. Dux transcription factors define a molecular subtype of a broad range of human cancers.

(A-B) Unsupervised cluster analysis of kidney (A) and stomach (B) cancers positive for Dux gene expression and/or zygotic gene activation from PANCAN-TCGA. All available clinical data for all tumor data sets are provided in Supplementary Tables 3 and 4.

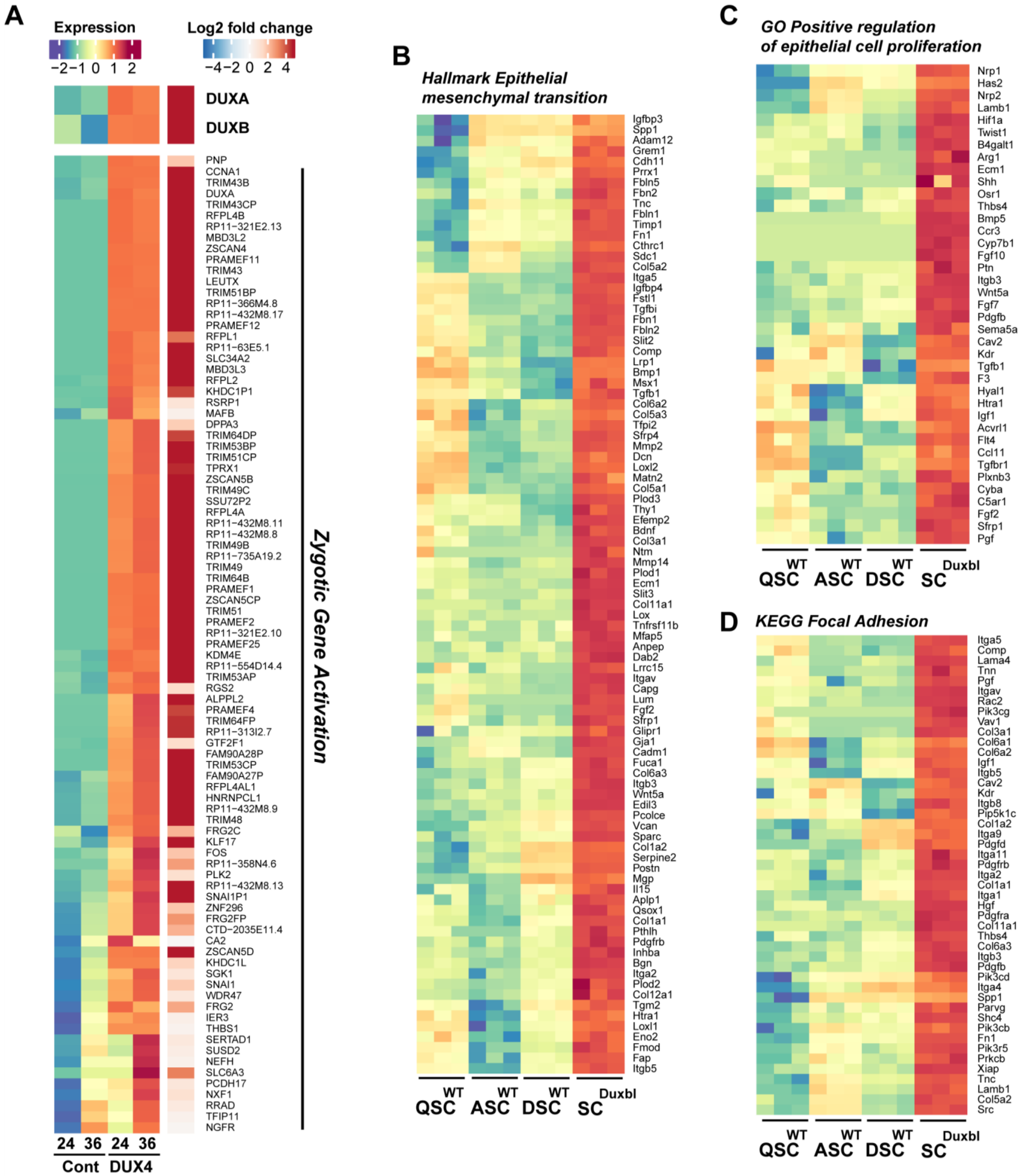

A
Expression
−2 −1 0 1 2
Log2 fold change
−4 −2 0 2 4
DUXA
DUXB
PNP
CCNA1
TRIM43B
DUXA
TRIM43CP
RFPL4B
RP11−321E2.13
MBD3L2
ZSCAN4
PRAMEF11
TRIM43
LEUTX
TRIM51BP
RP11−366M4.8
RP11−432M8.17
PRAMEF12
RFPL1
RP11−63E5.1
SLC34A2
MBD3L3
RFPL2
KHDC1P1
RSRP1
MAFB
DPPA3
TRIM64DP
TRIM53BP
TRIM51CP
TPRX1
ZSCAN5B
TRIM49C
SSU72P2
RFPL4A
RP11−432M8.11
RP11−432M8.8
TRIM49B
RP11−735A19.2
TRIM49
TRIM64B
PRAMEF1
ZSCAN5CP
TRIM51
PRAMEF2
RP11−321E2.10
PRAMEF25
KDM4E
RP11−554D14.4
TRIM53AP
RGS2
ALPPL2
PRAMEF4
TRIM64FP
RP11−313I2.7
GTF2F1
FAM90A28P
TRIM53CP
FAM90A27P
RFPL4AL1
HNRNPCL1
RP11−432M8.9
TRIM48
FRG2C
KLF17
FOS
RP11−358N4.6
PLK2
RP11−432M8.13
SNAI1P1
ZNF296
FRG2FP
CTD−2035E11.4
CA2
ZSCAN5D
KHDC1L
SGK1
SNAI1
WDR47
FRG2
IER3
THBS1
SERTAD1
SUSD2
NEFH
SLC6A3
PCDH17
NXF1
RRAD
TFIP11
NGFR
Zygotic Gene Activation
24 36 24 36
Cont DUX4
B
Hallmark Epithelial mesenchymal transition
Igfbp3
Spp1
Adam12
Grem1
Cdh11
Prrx1
Fbln5
Fbn2
Tnc
Fbln1
Timp1
Fn1
Cthrc1
Sdc1
Col5a2
Itga5
Igfbp4
Fstl1
Tgfbi
Fbn1
Fbln2
Slit2
Comp
Lrp1
Bmp1
Msx1
Tgfb1
Col6a2
Col5a3
Tfpi2
Sfrp4
Mmp2
Dcn
Loxl2
Matn2
Col5a1
Plod3
Thy1
Efemp2
Bdnf
Col3a1
Ntm
Mmp14
Plod1
Ecm1
Slit3
Col11a1
Lox
Tnfrsf11b
Mfap5
Anpep
Dab2
Lrrc15
Itgav
Capg
Lum
Fgf2
Sfrp1
Glipr1
Gja1
Cadm1
Fuca1
Col6a3
Itgb3
Wnt5a
Edil3
Pcolce
Vcan
Sparc
Col1a2
Serpine2
Postn
Mgp
Il15
Aplp1
Qsox1
Col1a1
Pthlh
Pdgfrb
Inhba
Bgn
Itga2
Plod2
Col12a1
Tgm2
Htra1
Loxl1
Eno2
Fmod
Fap
Itgb5
WT WT WT Duxbl
QSC ASC DSC SC
C
GO Positive regulation of epithelial cell proliferation
Nrp1
Has2
Nrp2
Lamb1
Hif1a
Twist1
B4galt1
Arg1
Ecm1
Shh
Osr1
Thbs4
Bmp5
Ccr3
Cyp7b1
Fgf10
Ptn
Itgb3
Wnt5a
Fgf7
Pdgfb
Sema5a
Cav2
Kdr
Tgfb1
F3
Hyal1
Htra1
Igf1
Acvrl1
Flt4
Ccl11
Tgfbr1
Plxnb3
Cyba
C5ar1
Fgf2
Sfrp1
Pgf
WT WT WT Duxbl
QSC ASC DSC SC
D
KEGG Focal Adhesion
Itga5
Comp
Lama4
Tnn
Pgf
Itgav
Rac2
Pik3cg
Vav1
Col3a1
Col6a1
Col6a2
Igf1
Itgb5
Cav2
Kdr
Itgb8
Pip5k1c
Col1a2
Itga9
Pdgfd
Itga11
Pdgfrb
Itga2
Col1a1
Itga1
Hgf
Pdgfra
Col11a1
Thbs4
Col6a3
Itgb3
Pdgfb
Pik3cd
Itga4
Spp1
Parvg
Shc4
Pik3cb
Fn1
Pik3r5
Prkcb
Xiap
Tnc
Lamb1
Col5a2
Src
WT WT WT Duxbl
QSC ASC DSC SC

Figure S7, related to Figure 5 and Figure 6. Overexpression of Duxbl in muscle SCs elicits epithelialization.

(A) Heatmap showing induction of ZGA upon overexpression of Dux4 at indicated time-points after transfection. Note that forced expression of Dux4 induces DuxA and DuxB expression, indicating Dux4 to act genetically upstream of DuxA and DuxB. **(B-D)** Heatmaps showing differential expression of genes involved in hallmarks of EMT/MET (B), positive regulation of epithelial cell proliferation (C), and focal adhesion (D).

Gratitude

I am deeply grateful for the support I received during the last seven years and in particular for the support during the preparation of this thesis and its defense.

I especially would like to thank Mario Looso, Thomas Braun, Alexander Goesmann and Johnny Kim, who have had a major influence on my thesis, my scientific success and my personal development. Thank you for your continous support, many discussions, suggestions, ideas and providing a healthy and productive working environment at the institute.

I further would like to thank Reinhard Dammann and Stefan Jannsen, who willingly agreed to be part of my thesis reviewers, for their time and effort before and during the defense.

Many thanks belong to all members of the Loosolab, i.e. Aditya, Annika, Arsenij, Basti, Carsten, Daniel, Fatemeh, Frank, Franz, Hendrik, Hosro, Julia, Kathrin, Marina, Mette, Mike, Nina, Peter, Philipp, Rene, Stefan, Sweta, Thomas and Vincent. It has been a pleasure working with you and being part of an inspiring group! Thank you, Johannes, Alica, Sylvia, Anja and Chris, for day-to-day inspiring conversations at lunch, in the lab and on the hallways.

Much love and hugs to my brother, Dirk, and my best friends Fabian, Julian, Till, Sebbl and Lukas for our long-lasting friendship and the wonderful time we have had together so far. Finally, I am deeply thankful for my parents, their presence, love and the unconditional support I receive day-to-day.

3 References

Abrahante, J. E., Daul, A. L., Li, M., Volk, M. L., Tennessen, J. M., Miller, E. A., & Rougvie, A. E. (2003). The Caenorhabditis elegans hunchback-like Gene lin-57/hbl-1 Controls Developmental Time and Is Regulated by MicroRNAs. *Developmental Cell*, *4*, 625–637.

Almada, A. E., & Wagers, A. J. (2016). Molecular circuitry of stem cell fate in skeletal muscle regeneration, ageing and disease. *Nature Reviews Molecular Cell Biology*, *17*, 267–279.

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., … Stegle, O. (2018). Multi-Omics Factor Analysisa framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, *14*. `https://doi.org/10.15252/msb.20178124`

Bayer, J., Kuenne, C., Preussner, J., & Looso, M. (2016). LimiTT: Link miRNAs to targets. *BMC Bioinformatics*, *17*. `https://doi.org/10.1186/s12859-016-1070-1`

Bell, A. C., & Felsenfeld, G. (2000). Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature*, *405*, 482–485.

Bernstein, E., Kim, S. Y., Carmell, M. A., Murchison, E. P., Alcorn, H., Li, M. Z., … Hannon, G. J. (2003). Dicer is essential for mouse development. *Nature Genetics*, *35*, 215–217.

Béguelin, W., Popovic, R., Teater, M., Jiang, Y., Bunting, K. L., Rosen, M., … Melnick, A. M. (2013). EZH2 Is Required for Germinal Center Formation and Somatic EZH2 Mutations Promote Lymphoid Transformation. *Cancer Cell*, *23*, 677–692.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., … Shen, R. (2011). High density DNA methylation array with single CpG site resolution. *Genomics*, *98*, 288–295.

Bissell, M. J., & Radisky, D. (2001). Putting tumours in context. *Nature Reviews Cancer*, *1*, 46–54.

Black, J. C., Atabakhsh, E., Kim, J., Biette, K. M., Van Rechem, C., Ladd, B., … Whetstine, J. R. (2015). Hypoxia drives transient site-specific copy gain and drug-resistant gene expression. *Genes & Development*, *29*, 1018–1031.

Blenkiron, C., Goldstein, L. D., Thorne, N. P., Spiteri, I., Chin, S.-F., Dunning, M. J., … Miska, E. A. (2007). MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype. *Genome Biology*, *8*, R214.

Blum, J. M., Añó, L., Li, Z., Van Mater, D., Bennett, B. D., Sachdeva, M., … Kirsch, D. G. (2013). Distinct and Overlapping Sarcoma Subtypes Initiated from Muscle Stem and Progenitor Cells. *Cell Reports*, *5*, 933–940.

Bock, C. (2012). Analysing and interpreting DNA methylation data. *Nature Reviews Genetics*, *13*, 705–719.

Bock, C., & Lengauer, T. (2008). Computational epigenetics. *Bioinformatics*, *24*, 1–10.

Boettger, T., Wüst, S., Nolte, H., & Braun, T. (2014). The miR-206/133b cluster is dispensable for development, survival and regeneration of skeletal muscle. *Skeletal Muscle*, *4*. https://doi.org/10.1186/s13395-014-0023-5

Boldrin, L., Zammit, P. S., & Morgan, J. E. (2015). Satellite cells from dystrophic muscle retain regenerative capacity. *Stem Cell Research*, *14*, 20–29.

Braun, T., & Gautel, M. (2011). Transcriptional mechanisms regulating skeletal muscle differentiation, growth and homeostasis. *Nature Reviews Molecular Cell Biology*, *12*, 349–361.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *68*, 394–424.

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*, 525–527.

Brennecke, J., Hipfner, D. R., Stark, A., Russell, R. B., & Cohen, S. M. (2003). Bantam Encodes a Developmentally Regulated microRNA that Controls Cell Proliferation and Regulates the Proapoptotic Gene hid in Drosophila. *Cell*, *113*, 25–36.

Bridge, J. A., Liu, J., Qualman, S. J., Suijkerbuijk, R., Wenger, G., Zhang, J., ... Barr, F. G. (2002). Genomic gains and losses are similar in genetic and histologic subsets of rhabdomyosarcoma, whereas amplification predominates in embryonal with anaplasia and alveolar subtypes. *Genes, Chromosomes and Cancer*, *33*, 310–321.

Camboni, M., Hammond, S., Martin, L. T., & Martin, P. T. (2012). Induction of a regenerative microenvironment in skeletal muscle is sufficient to induce embryonal rhabdomyosarcoma in p53-deficient mice. *The Journal of Pathology*, *226*, 40–49.

Chamberlain, J. S., Metzger, J., Reyes, M., Townsend, D., & Faulkner, J. A. (2007). Dystrophin-deficient mdx mice display a reduced life span and are susceptible to spontaneous rhabdomyosarcoma. *The FASEB Journal*, *21*, 2195–2204.

Chen, C.-Z. (2004). MicroRNAs Modulate Hematopoietic Lineage Differentiation. *Science*, *303*, 83–86.

Chen, E. Y., DeRan, M. T., Ignatius, M. S., Grandinetti, K. B., Clagg, R., McCarthy, K. M., ... Langenau, D. M. (2014). Glycogen synthase kinase 3 inhibitors induce the canonical WNT/ -catenin pathway to suppress growth and self-renewal in embryonal rhabdomyosarcoma. *Proceedings of the National Academy of Sciences*, *111*, 5349–5354.

Chen, J.-F., Tao, Y., Li, J., Deng, Z., Yan, Z., Xiao, X., & Wang, D.-Z. (2010). microRNA-1 and microRNA-206 regulate skeletal muscle satellite cell proliferation and differentiation by repressing Pax7. *The Journal of Cell Biology*, *190*, 867–879.

Chen, X., Stewart, E., Shelat, A. A., Qu, C., Bahrami, A., Hatley, M., ... Dyer, M. A. (2013). Targeting Oxidative Stress in Embryonal Rhabdomyosarcoma. *Cancer Cell*, *24*, 710–724.

Chendrimada, T. P., Gregory, R. I., Kumaraswamy, E., Norman, J., Cooch, N., Nishikura, K., & Shiekhattar, R. (2005). TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. *Nature*, *436*, 740–744.

Chiappalupi, S., Riuzzi, F., Fulle, S., Donato, R., & Sorci, G. (2014). Defective RAGE activity in embryonal rhabdomyosarcoma cells results in high PAX7 levels that sustain migration and invasiveness. *Carcinogenesis*, *35*, 2382–2392.

Chisholm, A. (2001). Cell Lineage. In *Encyclopedia of Genetics* (pp. 302–310). Elsevier.

Christofori, G., & Semb, H. (1999). The role of the cell-adhesion molecule E-cadherin as a tumour-suppressor gene. *Trends in Biochemical Sciences*, *24*, 73–76.

Ciarapica, R., Russo, G., Verginelli, F., Raimondi, L., Donfrancesco, A., Rota, R., & Giordano, A. (2009). Deregulated expression of miR-26a and Ezh2 in Rhabdomyosarcoma. *Cell Cycle*, *8*, 172–175.

Creton, S., Aardema, M. J., Carmichael, P. L., Harvey, J. S., Martin, F. L., Newbold, R. F., ... Yasaei, H. (2012). Cell transformation assays for prediction of carcinogenic potential: State of the science and future research needs. *Mutagenesis*, *27*, 93–101.

Dagenais-Bellefeuille, S., Beauchemin, M., & Morse, D. (2017). miRNAs Do Not Regulate Circadian Protein Synthesis in the Dinoflagellate Lingulodinium polyedrum. *PLOS ONE*, *12*, e0168817.

Davicioni, E., Anderson, M. J., Finckenstein, F. G., Lynch, J. C., Qualman, S. J., Shimada, H., ... Triche, T. J. (2009). Molecular Classification of Rhabdomyosarcoma Genotypic and Phenotypic Determinants of Diagnosis. *The American Journal of Pathology*, *174*, 550–564.

Drummond, C. J., Hanna, J. A., Garcia, M. R., Devine, D. J., Heyrana, A. J., Finkelstein, D., ... Hatley, M. E. (2018). Hedgehog Pathway Drives Fusion-Negative Rhabdomyosarcoma Initiated From Non-myogenic Endothelial Progenitors. *Cancer Cell*, *33*, 108–124.e5.

Durbin, A. D., Somers, G. R., Forrester, M., Pienkowska, M., Hannigan, G. E., & Malkin, D. (2009). JNK1 determines the oncogenic or tumor-suppressive activity of the integrin-linked kinase in human rhabdomyosarcoma. *Journal of Clinical Investigation*. https://doi.org/10.1172/JCI37958

El Demellawy, D., McGowan-Jordan, J., de Nanassy, J., Chernetsova, E., & Nasr, A. (2017). Update on molecular findings in rhabdomyosarcoma. *Pathology*, *49*, 238–246.

Ellis, I. O. (2006). Impact of a national external quality assessment scheme for breast pathology in the UK. *Journal of Clinical Pathology*, *59*, 138–145.

Feil, S., Valtcheva, N., & Feil, R. (2009). Inducible Cre Mice. In W. Wurst & R. Kühn (Eds.), *Gene Knockout Protocols* (Vol. 530, pp. 343–363). Totowa, NJ: Humana Press.

Flavahan, W. A., Drier, Y., Liau, B. B., Gillespie, S. M., Venteicher, A. S., Stemmer-Rachamimov, A. O., ... Bernstein, B. E. (2016). Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature*, *529*, 110–114.

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

Gao, F., Wang, W., Tan, M., Zhu, L., Zhang, Y., Fessler, E., ... Wang, X. (2019). DeepCC: A novel deep learning-based framework for cancer molecular subtype classification. *Oncogenesis*, *8*. https://doi.org/10.1038/s41389-019-0157-8

Garraway, L. A., & Sellers, W. R. (2006). Lineage dependency and lineage-survival oncogenes in human cancer. *Nature Reviews Cancer*, *6*, 593–602.

Gomes, C. P. C., Cho, J.-H., Hood, L., Franco, O. L., Pereira, R. W., & Wang, K. (2013). A Review of Computational Tools in microRNA Discovery. *Frontiers in Genetics*, *4*. https://doi.org/10.3389/fgene.2013.00081

Grizzi, F., Di Ieva, A., Russo, C., Frezza, E. E., Cobos, E., Muzzio, P. C., & Chiriva-Internati, M. (2006). Cancer initiation and progression: An unsimplifiable complexity. *Theoretical Biology and Medical Modelling*, *3*. https://doi.org/10.1186/1742-4682-3-37

Gupta, P. B., Kuperwasser, C., Brunet, J.-P., Ramaswamy, S., Kuo, W.-L., Gray, J. W., ... Weinberg, R. A. (2005). The melanocyte differentiation program predisposes to metastasis after neoplastic transformation. *Nature Genetics*, *37*, 1047–1054.

Günther, S., Kim, J., Kostin, S., Lepper, C., Fan, C.-M., & Braun, T. (2013). Myf5-Positive Satellite Cells Contribute to Pax7-Dependent Long-Term Maintenance of Adult Muscle Stem Cells. *Cell Stem Cell*, *13*, 590–601.

Han, J. (2004). The Drosha-DGCR8 complex in primary microRNA processing. *Genes & Development*, *18*, 3016–3027.

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of Cancer: The Next Generation. *Cell*, *144*, 646–674.

Hautefort, A., Chesné, J., Preussner, J., Pullamsetti, S. S., Tost, J., Looso, M., ... Perros, F. (2017). Pulmonary endothelial cell DNA methylation signature in pulmonary arterial hypertension. *Oncotarget*, *8*. https://doi.org/10.18632/oncotarget.18031

Hayashi, S., & McMahon, A. P. (2002). Efficient Recombination in Diverse Tissues by a Tamoxifen-Inducible Form of Cre: A Tool for Temporally Regulated Gene Activation/Inactivation in the Mouse. *Developmental Biology*, *244*, 305–318.

Hendrickson, P. G., Doráis, J. A., Grow, E. J., Whiddon, J. L., Lim, J.-W., Wike, C. L., ... Cairns, B. R. (2017). Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVL retrotransposons. *Nature Genetics*, *49*, 925–934.

Heng, H. H., Stevens, J. B., Bremer, S. W., Ye, K. J., Liu, G., & Ye, C. J. (2010). The evolutionary mechanism of cancer. *Journal of Cellular Biochemistry*, 1072–1084.

Hernández-Hernández, V., Rueda, D., Caballero, L., Alvarez-Buylla, E. R., & Benítez, M. (2014). Mechanical forces as information: An integrated approach to plant and animal development. *Frontiers in Plant Science*, *5*. https://doi.org/10.3389/fpls.2014.00265

Hettmer, S., Liu, J., Miller, C. M., Lindsay, M. C., Sparks, C. A., Guertin, D. A., ... Wagers, A. J. (2011). Sarcomas induced in discrete subsets of prospectively isolated skeletal muscle cells. *Proceedings of the National Academy of Sciences*, *108*, 20002–20007.

Hettmer, S., & Wagers, A. J. (2010). Muscling in: Uncovering the origins of rhabdomyosarcoma. *Nature Medicine*, *16*, 171–173.

Hirata, Y., & Hirata, S. (2002). Physio-mitotic theory and a new concept of cancer development. *Medical Hypotheses*, *58*, 361–364.

Hitchins, M. P., Wong, J. J., Suthers, G., Suter, C. M., Martin, D. I., Hawkins, N. J., & Ward, R. L. (2007). Inheritance of a Cancer-Associated *MLH1* Germ-Line Epimutation. *New England Journal of Medicine*, *356*, 697–705.

Hnisz, D., Weintraub, A. S., Day, D. S., Valton, A.-L., Bak, R. O., Li, C. H., ... Young, R. A. (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, *351*, 1454–1458.

Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., ... Mariamidze, A. (2018). Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell*, *173*, 291–304.e6.

Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., ... Stuart, J. M. (2014). Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. *Cell*, *158*, 929–944.

Hönig, J., Miíková, I., Nardiello, C., Surate Solaligue, D. E., Daume, M. J., Vadász, I., ... Morty, R. E. (2018). Transmission of microRNA antimiRs to mouse offspring via the maternalPlacentalFetal unit. *RNA*, *24*, 865–879.

Illmensee, K., & Mintz, B. (1976). Totipotency and normal differentiation of single teratocarcinoma cells cloned by injection into blastocysts. *Proceedings of the National Academy of Sciences of the United States of America*, *73*, 549–553.

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*, 860–921.

Jin, H. Y., Oda, H., Lai, M., Skalsky, R. L., Bethel, K., Shepherd, J., ... Xiao, C. (2013). MicroRNA-17∼92 plays a causative role in lymphomagenesis by coordinating multiple oncogenic pathways. *The EMBO Journal*, *32*, 2377–2391.

John, R. M., & Rougeulle, C. (2018). Developmental Epigenetics: Phenotype and the Flexible Epigenome. *Frontiers in Cell and Developmental Biology*, *6*. `https://doi.org/10.3389/fcell.2018.00130`

Jones, P. A. (2012). Functions of DNA methylation: Islands, start sites, gene bodies and beyond. *Nature Reviews Genetics*, *13*, 484–492.

Jones-Rhoades, M. W., Bartel, D. P., & Bartel, B. (2006). MicroRNAs AND THEIR REGULATORY ROLES IN PLANTS. *Annual Review of Plant Biology*, *57*, 19–53.

Judson, R. N., Tremblay, A. M., Knopp, P., White, R. B., Urcia, R., De Bari, C., ... Wackerhage, H. (2012). The Hippo pathway member Yap plays a key role in influencing fate decisions in muscle satellite cells. *Journal of Cell Science*, *125*, 6009–6019.

Kaelin, W. G. (2005). The Concept of Synthetic Lethality in the Context of Anticancer Therapy. *Nature Reviews Cancer*, *5*, 689–698.

Kim, H. K., Lee, Y. S., Sivaprasad, U., Malhotra, A., & Dutta, A. (2006). Muscle-specific microRNA miR-206 promotes muscle differentiation. *The Journal of Cell Biology*, *174*, 677–687.

Kondili, M., Fust, A., Preussner, J., Kuenne, C., Braun, T., & Looso, M. (2017). UROPA: A tool for Universal RObust Peak Annotation. *Scientific Reports*, *7*. `https://doi.org/10.1038/s41598-017-02464-y`

Koutsogiannouli, E., Papavassiliou, A. G., & Papanikolaou, N. A. (2013). Complexity in cancer biology: Is systems biology the answer? *Cancer Medicine*, *2*, 164–177.

Kuijjer, M. L., Paulson, J. N., Salzman, P., Ding, W., & Quackenbush, J. (2018). Cancer subtype identification using somatic mutation data. *British Journal of Cancer*, *118*, 1492–1501.

Kulis, M., & Esteller, M. (2010). DNA Methylation and Cancer. In *Advances in Genetics* (Vol. 70, pp. 27–56). Elsevier.

Labbé, R. M., Holowatyj, A., & Yang, Z.-Q. (2013). Histone lysine demethylase (KDM) subfamily 4: Structures, functions and therapeutic potential. *American Journal of Translational Research*, *6*, 1–15.

Laird, P. W. (2010). Principles and challenges of genome-wide DNA methylation analysis. *Nature Reviews Genetics*, *11*, 191–203.

Leidenroth, A., & Hewitt, J. E. (2010). A family history of DUX4: Phylogenetic analysis of DUXA, B, C and Duxbl reveals the ancestral DUX gene. *BMC Evolutionary Biology*, *10*, 364.

Li, R., Liang, J., Ni, S., Zhou, T., Qing, X., Li, H., ... Pei, D. (2010). A Mesenchymal-to-Epithelial Transition Initiates and Is Required for the Nuclear Reprogramming of Mouse Fibroblasts. *Cell Stem Cell*, *7*, 51–63.

Liau, B. B., Sievers, C., Donohue, L. K., Gillespie, S. M., Flavahan, W. A., Miller, T. E., ... Bernstein, B. E. (2017). Adaptive Chromatin Remodeling Drives Glioblastoma Stem Cell Plasticity and Drug Tolerance. *Cell Stem Cell*, *20*, 233–246.e7.

Lister, R., & Ecker, J. R. (2009). Finding the fifth base: Genome-wide sequencing of cytosine methylation. *Genome Research*, *19*, 959–966.

Liu, C., Li, D., Hu, J., Jiang, J., Zhang, W., Chen, Y., ... Li, F. (2014). Chromosomal and genetic imbalances in Chinese patients with rhabdomyosarcoma detected by high-resolution array comparative genomic hybridization. *International Journal of Clinical and Experimental Pathology*, *7*, 690–698.

Liu, W., & Wang, X. (2019). Prediction of functional microRNA targets by integrative modeling of microRNA binding and target expression data. *Genome Biology*, *20*. https://doi.org/10.1186/s13059-019-1629-z

MacConaill, L. E. (2013). Existing and Emerging Technologies for Tumor Genomic Profiling. *Journal of Clinical Oncology*, *31*, 1815–1824.

Madissoon, E., Jouhilahti, E.-M., Vesterlund, L., Töhönen, V., Krjutkov, K., Petropoulos, S., ... Kere, J. (2016). Characterization and target genes of nine human PRD-like homeobox domain genes expressed exclusively in early embryos. *Scientific Reports*, *6*. https://doi.org/10.1038/srep28995

Mardis, E. R. (2011). A decade's perspective on DNA sequencing technology. *Nature*, *470*, 198–203.

Margueron, R., & Reinberg, D. (2010). Chromatin structure and the inheritance of epigenetic information. *Nature Reviews Genetics*, *11*, 285–296.

Mohamad, T., Kazim, N., Adhikari, A., & Davie, J. K. (2018). EGR1 interacts with TBX2 and functions as a tumor suppressor in rhabdomyosarcoma. *Oncotarget*, *9*. https://doi.org/10.18632/oncotarget.24726

Neilsen, C. T., Goodall, G. J., & Bracken, C. P. (2012). IsomiRs the overlooked repertoire in the dynamic microRNAome. *Trends in Genetics*, *28*, 544–549.

Newton, Y., Novak, A. M., Swatloski, T., McColl, D. C., Chopra, S., Graim, K., ... Stuart, J. M. (2017). TumorMap: Exploring the Molecular Similarities of Cancer Samples in an Interactive Portal. *Cancer Research*, *77*, e111–e114.

Noushmehr, H., Weisenberger, D. J., Diefes, K., Phillips, H. S., Pujara, K., Berman, B. P., ... Aldape, K. (2010). Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma. *Cancer Cell*, *17*, 510–522.

Nurse, P. (1997). The ends of understanding. *Nature*, *387*, 657–657.

Ooi, S. K. T., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z., ... Bestor, T. H. (2007). DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature*, *448*, 714–717.

Oulas, A., Karathanasis, N., Louloupi, A., Pavlopoulos, G. A., Poirazi, P., Kalantidis, K., & Iliopoulos, I. (2015). Prediction of miRNA Targets. In E. Picardi (Ed.), *RNA Bioinformatics* (Vol. 1269, pp. 207–229). New York, NY: Springer New York.

Paduch, R. (2015). Theories of cancer origin: *European Journal of Cancer Prevention*, *24*, 57–67.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*, 417–419.

Pedersen, B. S., Schwartz, D. A., Yang, I. V., & Kechris, K. J. (2012). Comb-p: Software for combining, analyzing, grouping and correcting spatially correlated P-values. *Bioinformatics*, *28*, 2986–2988.

Perez-Losada, J., & Balmain, A. (2003). Stem-cell hierarchy in skin cancer. *Nature Reviews Cancer*, *3*, 434–443.

Prat, A., & Perou, C. M. (2011). Deconstructing the molecular portraits of breast cancer. *Molecular Oncology*, *5*, 5–23.

Pugh, T. J., Weeraratne, S. D., Archer, T. C., Pomeranz Krummel, D. A., Auclair, D., Bochicchio, J., ... Cho, Y.-J. (2012). Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. *Nature*, *488*, 106–110.

Rauschert, I., Aldunate, F., Preussner, J., Arocena-Sutz, M., Peraza, V., Looso, M., ... Agrelo, R. (2017). Promoter hypermethylation as a mechanism for Lamin A/C silencing in a subset of neuroblastoma cells. *PLOS ONE*, *12*, e0175953.

Reichek, J. L., Duan, F., Smith, L. M., Gustafson, D. M., O'Connor, R. S., Zhang, C., ... Barr, F. G. (2011). Genomic and Clinical Analysis of Amplification of the 13q31 Chromosomal Region in Alveolar Rhabdomyosarcoma: A Report from the Children's Oncology Group. *Clinical Cancer Research*, *17*, 1463–1473.

Ren, Y.-X., Finckenstein, F. G., Abdueva, D. A., Shahbazian, V., Chung, B., Weinberg, K. I., ... Anderson, M. J. (2008). Mouse Mesenchymal Stem Cells Expressing PAX-FKHR Form Alveolar Rhabdomyosarcomas by Cooperating with Secondary Mutations. *Cancer Research*, *68*, 6587–6597.

Rikhof, B., de Jong, S., Suurmeijer, A. J., Meijer, C., & van der Graaf, W. T. (2009). The insulin-like growth factor system and sarcomas. *The Journal of Pathology*, *217*, 469–482.

Roesch, A., Vultur, A., Bogeski, I., Wang, H., Zimmermann, K. M., Speicher, D., ... Herlyn, M. (2013). Overcoming Intrinsic Multidrug Resistance in Melanoma by

Blocking the Mitochondrial Respiratory Chain of Slow-Cycling JARID1Bhigh Cells. *Cancer Cell*, *23*, 811–825.

Rubin, B. P., Nishijo, K., Chen, H.-I. H., Yi, X., Schuetze, D. P., Pal, R., ... Keller, C. (2011). Evidence for an Unanticipated Relationship between Undifferentiated Pleomorphic Sarcoma and Embryonal Rhabdomyosarcoma. *Cancer Cell*, *19*, 177–191.

Sandhu, S. K., Fassan, M., Volinia, S., Lovat, F., Balatti, V., Pekarsky, Y., & Croce, C. M. (2013). B-cell malignancies in microRNA E -miR-17 92 transgenic mice. *Proceedings of the National Academy of Sciences*, *110*, 18208–18213.

Sarver, A. L., Li, L., & Subramanian, S. (2010). MicroRNA miR-183 Functions as an Oncogene by Targeting the Transcription Factor EGR1 and Promoting Tumor Cell Migration. *Cancer Research*, *70*, 9570–9580.

Saunderson, E. A., Stepper, P., Gomm, J. J., Hoa, L., Morgan, A., Allen, M. D., ... Ficz, G. (2017). Hit-and-run epigenetic editing prevents senescence entry in primary breast cells from healthy donors. *Nature Communications*, *8*. https://doi.org/10.1038/s41467-017-01078-2

Schaaf, G. J., Ruijter, J. M., van Ruissen, F., Zwijnenburg, D. A., Waaijer, R., Valentijn, L. J., ... Kool, M. (2005). Full transcriptome analysis of rhabdomyosarcoma, normal, and fetal skeletal muscle: Statistical comparison of multiple SAGE libraries. *The FASEB Journal*, *19*, 404–406.

Schwarz, D. S., Hutvágner, G., Du, T., Xu, Z., Aronin, N., & Zamore, P. D. (2003). Asymmetry in the Assembly of the RNAi Enzyme Complex. *Cell*, *115*, 199–208.

Seki, M., Nishimura, R., Yoshida, K., Shimamura, T., Shiraishi, Y., Sato, Y., ... Takita, J. (2015). Integrated genetic and epigenetic analysis defines novel molecular subgroups in rhabdomyosarcoma. *Nature Communications*, *6*. https://doi.org/10.1038/ncomms8557

Senft, D., Leiserson, M. D., Ruppin, E., & Ronai, Z. A. (2017). Precision Oncology: The Road Ahead. *Trends in Molecular Medicine*, *23*, 874–898.

Shen, R., Olshen, A. B., & Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, *25*, 2906–2912.

Sigston, E. A. W., & Williams, B. R. G. (2017). An Emergence Framework of Carcinogenesis. *Frontiers in Oncology*, *7*. https://doi.org/10.3389/fonc.2017.00198

Simon, J. A., & Kingston, R. E. (2013). Occupying Chromatin: Polycomb Mechanisms for Getting to Genomic Targets, Stopping Transcriptional Traffic, and Staying Put. *Molecular Cell*, *49*, 808–824.

Singer, J., Irmisch, A., Ruscheweyh, H.-J., Singer, F., Toussaint, N. C., Levesque, M. P., ... Beerenwinkel, N. (2017). Bioinformatics for precision oncology. *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bbx143

Sneeringer, C. J., Scott, M. P., Kuntz, K. W., Knutson, S. K., Pollock, R. M., Richon, V. M., & Copeland, R. A. (2010). Coordinated activities of wild-type plus mutant EZH2 drive tumor-associated hypertrimethylation of lysine 27 on histone H3 (H3K27) in human B-cell lymphomas. *Proceedings of the National Academy of Sciences*, *107*, 20980–20985.

Stephens, P. J., McBride, D. J., Lin, M.-L., Varela, I., Pleasance, E. D., Simpson, J. T., ... Stratton, M. R. (2009). Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, *462*, 1005–1010.

Sun, W., Chatterjee, B., Shern, J. F., Patidar, R., Song, Y., Wang, Y., ... Barr, F. G. (2019). Relationship of DNA methylation to mutational changes and transcriptional organization in fusion-positive and fusion-negative rhabdomyosarcoma. *International Journal of Cancer*, *144*, 2707–2717.

Tabibu, S., Vinod, P. K., & Jawahar, C. V. (2019). Pan-Renal Cell Carcinoma classification and survival prediction from histopathology images using deep learning. *Scientific Reports*, *9*. https://doi.org/10.1038/s41598-019-46718-3

Tam, S., de Borja, R., Tsao, M.-S., & McPherson, J. D. (2014). Robust global microRNA expression profiling using next-generation sequencing technologies. *Laboratory Investigation*, *94*, 350–358.

Taulli, R., Bersani, F., Foglizzo, V., Linari, A., Vigna, E., Ladanyi, M., ... Ponzetto, C. (2009). The muscle-specific microRNA miR-206 blocks human rhabdomyosarcoma growth in xenotransplanted mice by promoting myogenic differentiation. *Journal of Clinical Investigation*, JCI38075.

Taulli, R., Scuoppo, C., Bersani, F., Accornero, P., Forni, P. E., Miretti, S., ... Ponzetto, C. (2006). Validation of Met as a Therapeutic Target in Alveolar and Embryonal Rhabdomyosarcoma. *Cancer Research*, *66*, 4742–4749.

The ENCODE Project Consortium, Analysis Coordination, Chromatin and Replication, Genes and Transcripts, Integrated Analysis and Manuscript Preparation, Management Group, ... de Jong, P. J. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, *447*, 799–816.

Tiffin, N., Williams, R. D., Shipley, J., & Pritchard-Jones, K. (2003). PAX7 expression in embryonal rhabdomyosarcoma suggests an origin in muscle satellite cells. *British Journal of Cancer*, *89*, 327–332.

Tremblay, A. M., Missiaglia, E., Galli, G. G., Hettmer, S., Urcia, R., Carrara, M., ... Camargo, F. D. (2014). The Hippo Transducer YAP1 Transforms Activated Satellite Cells and Is a Potent Effector of Embryonal Rhabdomyosarcoma Formation. *Cancer Cell*, *26*, 273–287.

Ulitsky, I., Laurent, L. C., & Shamir, R. (2010). Towards computational prediction of microRNA function and activity. *Nucleic Acids Research*, *38*, e160–e160.

Venolia, L., & Gartler, S. M. (1983). Comparison of transformation efficiency of human active and inactive X-chromosomal DNA. *Nature*, *302*, 82–83.

Vincent, J.-L. (2017). The coming era of precision medicine for intensive care. *Critical Care*, *21*. https://doi.org/10.1186/s13054-017-1910-z

Visvader, J. E. (2011). Cells of origin in cancer. *Nature*, *469*, 314–322.

Wachtel, M., Rakic, J., Okoniewski, M., Bode, P., Niggli, F., & Schäfer, B. W. (2014). FGFR4 signaling couples to Bim and not Bmf to discriminate subsets of alveolar rhabdomyosarcoma cells: FGFR4 mediates survival in aRMS subgroups. *International Journal of Cancer*, *135*, 1543–1552.

Waddington, C. (2014). *The Strategy of the Genes* (1st ed.). Routledge.

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., ... Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, *11*, 333–337.

Wang, S., Guo, L., Dong, L., Guo, L., Li, S., Zhang, J., & Sun, M. (2010). TGF-$B1$ signal pathway may contribute to rhabdomyosarcoma development by inhibiting differentiation. *Cancer Science*, *101*, 1108–1116.

Weaver, V., Petersen, O., Wang, F., Larabell, C., Briand, P., Damsky, C., & Bissell, M. (1997). Reversion of the Malignant Phenotype of Human Breast Cells in Three-Dimensional Culture and In Vivo by Integrin Blocking Antibodies. *The Journal of Cell Biology*, *137*, 231–245.

Weinstein, I. B. (2002). CANCER: Enhanced: Addiction to Oncogenes–the Achilles Heal of Cancer. *Science*, *297*, 63–64.

Whiddon, J. L., Langford, A. T., Wong, C.-J., Zhong, J. W., & Tapscott, S. J. (2017). Conservation and innovation in the DUX4-family gene network. *Nature Genetics*, *49*, 935–940.

Williamson, D., Missiaglia, E., de Reyniès, A., Pierron, G., Thuille, B., Palenzuela, G., ... Delattre, O. (2010). Fusion GeneNegative Alveolar Rhabdomyosarcoma Is Clinically and Molecularly Indistinguishable From Embryonal Rhabdomyosarcoma. *Journal of Clinical Oncology*, *28*, 2151–2158.

Wu, D., Rice, C. M., & Wang, X. (2012). Cancer bioinformatics: A new approach to systems clinical medicine. *BMC Bioinformatics*, *13*, 71.

Xiao, J., Feng, S., Wang, X., Long, K., Luo, Y., Wang, Y., ... Li, M. (2018). Identification of exosome-like nanoparticle-derived microRNAs from 11 edible fruits and vegetables. *PeerJ*, *6*, e5186.

Yi, R. (2003). Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes & Development*, *17*, 3011–3016.

Yoder, J. A., Walsh, C. P., & Bestor, T. H. (1997). Cytosine methylation and the ecology of intragenomic parasites. *Trends in Genetics*, *13*, 335–340.

Zaret, K. S., & Mango, S. E. (2016). Pioneer transcription factors, chromatin dynamics, and cell fate control. *Current Opinion in Genetics & Development*, *37*, 76–81.

Zhang, S., Wang, Y., Gu, Y., Zhu, J., Ci, C., Guo, Z., ... Zhang, Y. (2018). Specific breast cancer prognosis-subtype distinctions based on DNA methylation patterns. *Molecular Oncology*, *12*, 1047–1060.