Pioneer activity of an oncogenic fusion transcription factor at inaccessible chromatin

Authors

Benjamin Sunkel[1,*], Meng Wang[1,*], Stephanie LaHaye[2,*], Benjamin J. Kelly[2], James R. Fitch[2], Frederic G. Barr[3], Peter White[2,4], Benjamin Z. Stanton[1,4,5,†]

Author Affiliations

[1] Nationwide Children's Hospital, Center for Childhood Cancer and Blood Diseases.

[2] The Steve and Cindy Rasmussen Institute for Genomic Medicine, Nationwide Children's Hospital.

[3] Laboratory of Pathology, National Cancer Institute.

[4] Department of Pediatrics, The Ohio State University College of Medicine.

[5] Department of Biological Chemistry and Pharmacology, The Ohio State University College of Medicine.

* Contributed equally to this work as joint co-first authors

[†] Correspondence should be addressed to B.Z.S. (benjamin.stanton@nationwidechildrens.org)

Abstract

Recent advances in the characterization of pioneer factors have led to new questions about domain structures and chromatin recognition. Seemingly separate fields have integrated common scientific problems with the identification of fusion pioneer factor (PF) pairs in childhood cancers. Excitingly, this represents a convergence of fields within science, of pediatric oncology, genomics, epigenetics, and biophysics. With PF-pairs emerging as driver oncogenes in childhood rhabdomyosarcoma (RMS), our studies investigate these fusion genes as a *gestalt phenomenon*, where the whole might equal more than the sum of the parts. We developed a new method for ChIP-seq with per-cell normalization (pc-ChIP-seq) to simultaneously define genome size and PAX3-FOXO1 localization in RMS. We report novel pioneer function for the major driver oncogene in RMS, with nucleosome-motif targeting and kinetic displacement of nucleosomes in human cells.

Introduction

The pioneer factor subset of the broad transcription factor (TF) class of proteins, possesses conserved motifs common to TFs, including a structured DNA-binding domain (DBD) responsible for motif recognition and a flexible transactivation domain mediating regulated recruitment (*1*). Distinct structural elements of pioneer factors (PFs) provide a unique capacity for high-affinity DNA binding despite steric deterrents, including nucleosomes. However, the combinatorial rules or "logic" of domain organization within PFs, requisite to achieve nucleosomal-motif recognition, remain obscure (*2*). Moreover, it is presently unknown if two pioneer factors fused together will possess pioneer activity in the resulting chimera.

In alveolar rhabdomyosarcoma, two pioneer factors, PAX3 and FOXO1, are fused in the recurrent driving translocation between chromosome arms 2p and 13q (*3*). The resulting PAX3-FOXO1 fusion oncogene has been described

as binding active regulatory elements alongside myogenic TFs (*4*), while its heterochromatin targeting function has remained unknown. Neither retention of canonical pioneer activity nor the emergence of novel function has been rigorously defined for PAX3-FOXO1 in fusion-positive rhabdomyosarcoma (FP-RMS). Given the relatively low mutational frequencies in FP-RMS, which can be approximated at 0.1 protein-coding mutations per Mb (*5*), we hypothesized that the chromatin-targeting function of PAX3-FOXO1 might drive this tumor. However, the mechanisms through which PAX3-FOXO1 engages distinct classes of chromatin have remained poorly understood.

The cascade of initiation events for a tumor are unlikely to involve mere stabilization of pre-existing transcriptional networks or DNA-accessibility, but rather, a restructuring of regulatory elements into a new state from a cell of origin (*6, 7*). Presently, the cell of origin for FP-RMS remains unknown, while the components of the major translocation, and other highly-expressed transcription factors (TFs), are reminiscent of neuronal tissue and developing muscle (*3, 8, 9*). Often, transcriptional reprogramming represents the functional output of tissue-specific pioneer factors (*10, 11*). It is noteworthy that tumors prevalent in children are frequently defined by a profound failure of cellular differentiation (*12*). In these relatively low-mutation-burden tumors, increasing evidence suggests that disruption of transcription factors drives reprogramming into altered epigenetic states. The role of PFs like SOX2 in developmental reprogramming may be analogous to pioneer activity in establishing cell-fate decisions in pediatric cancers, including synovial sarcoma and MPNST (*13, 14*), where mis-regulation of SOX-family pioneer factors occurs.

In the Forkhead family of PFs, a winged-helix domain mimics the structural features of the linker histone H1 (*15*), allowing displacement of H1 in compacted chromatin (*16*). Variations within the common helix-turn-helix DNA recognition motif are widely represented among PFs, including the PAX family, allowing DBD interactions with the major groove without sterically clashing with the associated nucleosome. The ability to recognize nucleosomal DNA sequences, inaccessible to traditional TFs, is fundamental to the pioneer factor's role in activating regulatory elements that establish new gene expression programs and differential phenotypes.

Results and Discussion

We asked whether the fused PFs in rhabdomyosarcoma exhibited fundamental properties of known pioneers. Homology analysis of PAX3 and FOXO1 amino acid sequences revealed that evolutionarily conserved residues within the PAX3 paired domain and homeobox domain are completely retained in the fusion TF (**Fig. 1A**). While the conserved nuclear-export and -localization sequences (NES and NLS), as well as the transactivation domain (TAD) of FOXO1 remain intact in the fusion, the forkhead or winged-helix domain is N-terminally truncated. The PAX3-FOXO1 fusion product retains the essential amino acid residues within alpha-helix 3 for recognition of FOXO1 DNA sequence motifs (5'-TGTTTAC-3') and the flexible wings W1 and W2, which have been shown to stabilize FOXO1:DNA interactions through direct phosphate-backbone contacts (**Fig. 1A**) (*17*). Despite the truncation, it was predicted by analysis of the primary amino acid sequence that the ordered DNA-binding domains and the disordered transactivation domains, originating from both PAX3 and FOXO1, are retained in the resulting fusion product (**fig. S1A**). As the PAX3-FOXO1 chimera retains many of the conserved features of these pioneer protein families, we went on to systematically assess the ability of this fusion TF to initiate reprogramming at the chromatin level to better understand epigenetic initiation in FP-RMS.

We were first motivated to address the enigmatic question of how a driver oncogene like PAX3-FOXO1 can initiate chromatin reprogramming unless it has capacity to recognize motifs outside of accessible DNA. Thus, we tested whether PAX3-FOXO1 genomic localization is distinct from non-pioneer TFs and whether it has the capacity to bind regions with lower levels of accessibility. We conducted a genome-wide correlation between various ChIP-seq datasets from the FP-RMS cell line RH4, available publicly or generated for this study (*18, 19*), and we compared these chromatin factors with PAX3-FOXO1 localization (*19*). Our results revealed remarkable dissimilarity between PAX3-FOXO1 binding and localization of FP-RMS core regulatory TFs (CRTFs; MYCN, MYOG, MYOD1), chromatin structural components (CTCF, RAD21), enhancer binding/regulatory factors (MED1, BRD4, p300), euchromatic histone modifications (H3K27ac, H3K27me3, H3K4me1), accessibility (ATAC-seq), and SWI/SNF chromatin remodeling complex subunits (BRD9, DPF2) (**fig. S1B**). Strikingly, we observed a unique genomic binding preference for PAX3-FOXO1 compared to other CRTFs in FP-RMS cells and suggested PAX3-FOXO1 may significantly reside in inactive regions of the genome.

To determine whether the pattern of PAX3-FOXO1 occupancy we observed could be attributed to its localization outside the active, euchromatic nuclear compartment, we employed a stringent, sequential fractionation protocol. We aimed at distinguishing cellular components associated with increasingly insoluble compartments of the nucleus in a panel of RMS cells, including PAX3-FOXO1 fusion-negative cell lines (RD, SMS-CTR) and fusion-positive cell lines (RH4, RH30). We found that the euchromatic SWI/SNF subunit BAF155 and the CRTF, MYCN, were readily extracted from the chromatin fiber in all cells when exposed to 500 mM NaCl extraction buffer (soluble nucleus; **Fig. 1B**). Residual chromatin-bound BAF155 was almost exclusively observed in the sonication-sensitive chromatin fraction (soluble chromatin; **Fig. 1B**). Consistent with previous characterization of this euchromatic fraction (*20*), we found that it was largely depleted of constitutive heterochromatin components such as HP1α as well as repression- and compaction-related histone modifications such as H3K9me3 and H4K20me1 (*20, 21*). TATA-binding protein (TBP), with its dual role in active gene transcription and mitotic bookmarking was readily observed in all nuclear and chromatin fractions including the residual insoluble chromatin pellet (**Fig. 1B**) (*22*). Blotting with an antibody recognizing the C-terminus of FOXO1 (epitope retained in the PAX3-FOXO1 fusion), we observed nuclear staining for wild type FOXO1 in RD and SMS-CTR cells that was largely excluded from the insoluble chromatin fraction. Unexpectedly, nuclear PAX3-FOXO1 in RH4 and RH30 cells, predominantly localized to the sonication-resistant, insoluble portion of the chromatin (**Fig. 1B and fig. S1C**). Thus, the PAX3-FOXO1 fusion protein appears to be uniquely capable of invading compact chromatin with higher affinity than its wild type constituents. This behavior, necessary for PF-mediated cell fate transition, is an unexplored feature of this oncogenic factor, previously unknown in this tumor.

To further delineate the unique chromatin binding features of PAX3-FOXO1 in FP-RMS, we next sought to define its genome-wide occupancy profile by optimizing high-specificity, spike-in normalized ChIP-seq conditions in RH4 cells using a C-terminal FOXO1 antibody (**fig. S2A, Supplemental Methods**). Our method, which we have called per-cell ChIP-seq (pc-ChIP-seq), addresses global changes in ChIP signal resulting from differential chromatin content or output between cell lines and treatment conditions, by introducing known ratios of mouse spike-in cells prior to sonication (*23*). We justified this methodology by analyzing input sequencing libraries to reveal vastly different relative genome sizes across RMS cell lines (Range: 5.25-10.02 Gb, see **Supplemental Methods**). Upon sequencing, we benchmarked our data against previously published PAX3-FOXO1 ChIP-seq generated with a non-commercial antibody, pFM2 (epitope spanning the fusion

breakpoint) (*19*). We found that our ChIP-eq conditions identified 69% of reported binding events (**fig. S2B**), and with improved signal, we revealed 7,341 novel and high-strength PAX3-FOXO1 sites (**fig. S2, B and C**). Employing identical ChIP conditions in additional RMS cell lines (**fig. S2D**), integrated analysis of our spike-in normalized anti-FOXO1 ChIP-seq revealed reproducible binding profiles that were concordant between FP-RMS cell lines (**Fig. 1C**). Additionally, while fusion-negative RMS cells (FN-RMS), particularly SMS-CTR, express nuclear FOXO1 (**Fig. 1B**), anti-FOXO1 ChIP-seq in these cells did not result in high confidence peak assignments (9 and 0 peaks called in RD and SMS-CTR, respectively), demonstrating that our approach is highly specific for generating PAX3-FOXO1 ChIP-seq even in the presence of chromatin-bound FOXO1.

We further investigated the quality and specificity of our PAX3-FOXO1 ChIP-seq by performing motif analysis. Among known motifs curated by HOMER, a PAX3:FKHR motif derived from previously published PAX3-FOXO1 ChIP-seq was the top enriched sequence (p-value = 10^{-1625}) (**Fig. 1D**) (*18, 19*). *De novo* motif identification also returned a sequence with highest similarity to the known PAX3:FKHR motif as the most enriched (p-value = 10^{-1753}). This motif is comprised of a paired-domain and a homeobox recognition sequence arranged in a divergent head-to-head orientation separated by a single cytosine. The remaining motifs enriched within PAX3-FOXO1 peaks (RH4 and RH30) were mainly E-Box motifs recognized by the non-pioneer (*2*), extended alpha-helical DBDs of basic helix-loop-helix (bHLH) family TFs such as MYF5, MYOG, MYOD1, and TCF, many of which are expressed in FP-RMS cells (**Fig. 1D**) (*24*). Notably, forkhead motifs were only modestly enriched in PAX3-FOXO1 binding sites relative to background (FOXO1 motif p-value = 10^{-31}), suggesting sequence-specific binding of the fusion is mediated predominantly through the intact PAX3 DBD. This was further supported by the existence of a transposase-protected footprint centered on high-confidence PAX3-FKHR motifs in RH4 binding sites, while no clear footprint is observed over forkhead motif positions by ATAC-seq (**Fig. 1E**). We proceeded with high confidence in our PAX3-FOXO1 (P3F) dataset as a foundation for discovering novel genomic binding patterns and potential pioneer function of this oncogenic fusion protein.

Having identified thousands of novel P3F binding sites in the FP-RMS genome, we were motivated to understand the genomic context and characteristic epigenetic state at these regions. As previously described, we found that P3F mainly occupies non-promoter, predominantly distal intergenic regions (**fig. S3A**) (*19*). Linking each P3F peak to nearby transcription start sites (TSS) with GREAT (*25*), we found that the limited number of promoter and promoter-proximal P3F binding sites were associated with general cellular and metabolic pathways (**fig. S3B**). Interestingly, more distal P3F binding sites up to 500 kilobases from a TSS were strongly linked to neurogenic differentiation processes, though FP-RMS cell lines do not exhibit differential expression of these genes compared to FN-RMS cell lines (**fig. S3C**). Thus, our discovery of P3F binding sites reveals associations of this fusion oncogene with gene expression programs unrelated to the myogenic transcriptional circuitry, which might provide clues regarding a cell-of-origin.

As in previous studies (*4, 18, 19*), we confirmed that P3F occupies gene-regulatory enhancers, including 1,520 individual binding sites within 932 TSS-distal, high intensity H3K27ac clusters stitched together by ROSE (**Fig. 1F** and **fig. S3D**). However, we observed no correlation between H3K27ac strength and P3F occupancy, as stitched enhancer regions lacking the fusion exhibited nearly equal levels of H3K27ac compared to enhancers bound by P3F (**Fig. 1G**). This may simply be due to the fact that P3F occupies a narrow footprint within discrete enhancer elements, and thus does not broadly influence the entire landscape of these broad regions (**Fig. 1G**). However, with a general lack of correlation between P3F

and H3K27ac enrichment across all enhancers, our results are consistent with a limited causal role for P3F in enhancer activation (**Fig. 1H**). Importantly, we found that 2,565 P3F sites occur outside of enhancers and confirmed that these regions are in fact depleted of both H3K27ac and H3K4me1 while maintaining DNA accessibility (**Fig. 1F**). Despite evolutionary conservation of the DNA sequences within these P3F sites (**fig. S3E**), we could not readily discern their regulatory role within the FP-RMS epigenome using publicly available datasets (**fig. S3F**). This suggests that these non-enhancer P3F peaks exhibit lineage-restricted epigenetic signatures, a feature common to promoter-distal gene-regulatory elements. When we evaluated the expression of genes linked to each type of P3F site, including novel non-enhancer regions, we observed little change in gene expression resulting from P3F knockdown, P3F add-back, or P3F expression in FP-RMS vs. FN-RMS cell lines regardless of P3F-site classification (**Fig. 1I**). These results are consistent with a pioneer role for PAX3-FOXO1 whose binding is associated with chromatin accessibility but which may be insufficient for chromatin activation and highly context-dependent for gene regulation.

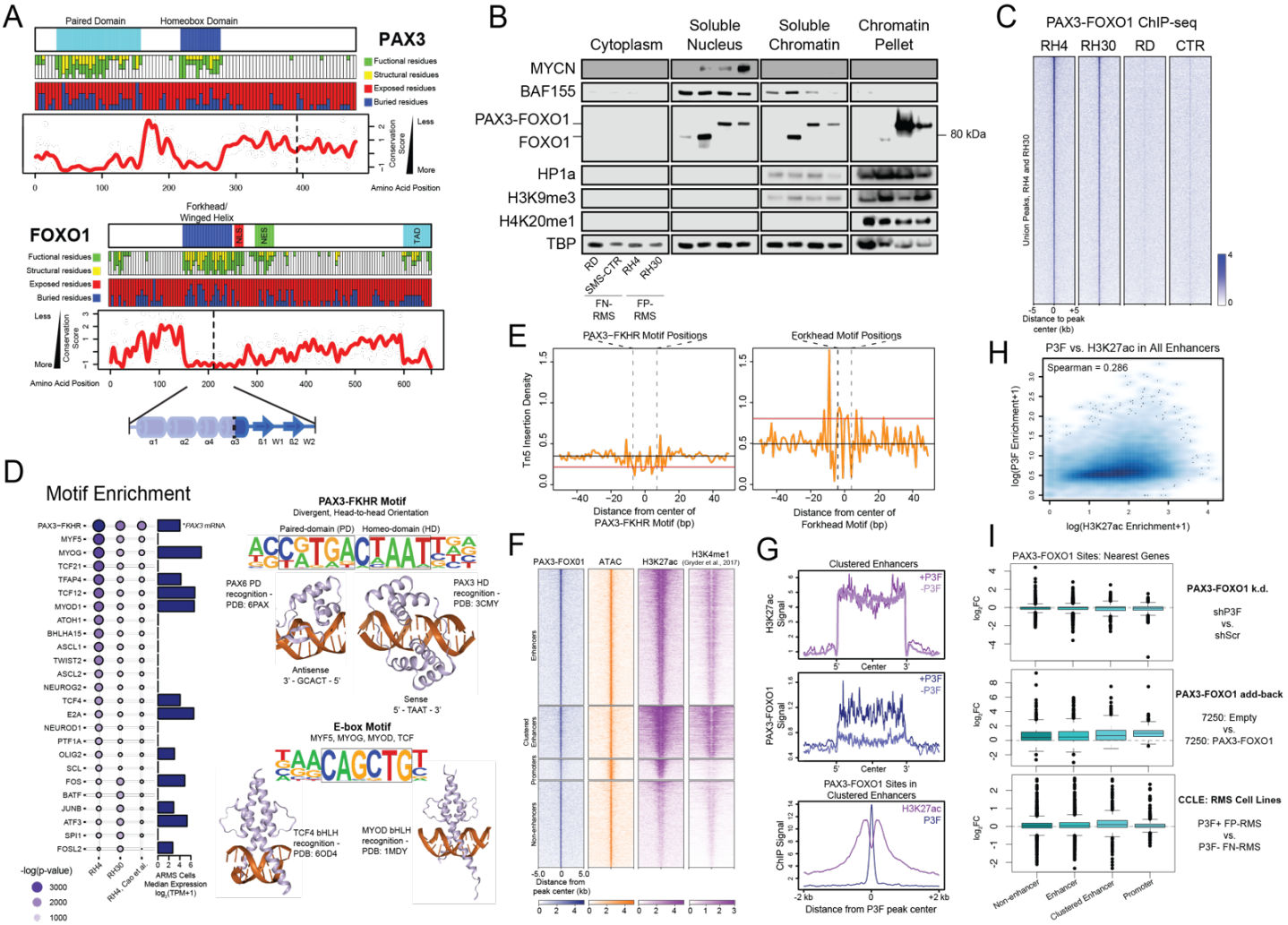


Fig. 1. Cellular and genomic localization of the PAX3-FOXO1 fusion transcription factor to inactive chromatin. (A) Evolutionary conservation of PAX3 (top) and FOXO1 (bottom) amino acid sequences. Vertical dashed lines indicate the breakpoint position resulting in the formation of the PAX3-FOXO1 fusion. FOXO1 inset details N-terminal truncation of the winged helix domain. (**B**) FN-RMS (RD, SMS-CTR) and FP-RMS (RH4, RH30) cells were fractionated to interrogate protein localization within the cytoplasm, soluble nucleus, soluble chromatin (sonication sensitive), and chromatin pellet (sonication insensitive) compartments of the cell. (**C**) A C-terminal FOXO1 antibody was utilized to perform per-cell ChIP-seq (pc-ChIP-seq) of the PAX3-FOXO1 fusion oncogene in a panel of RMS cells. Spike-in normalized signal intensity from each cell line was displayed across the union of PAX3-FOXO1 sites identified in RH4 and RH30 cells. (**D**) HOMER motif analysis was performed on PAX3-FOXO1 binding sites from anti-FOXO1 pc-ChIP-seq (RH4 and RH30), and pFM2 ChIP-seq (RH4, Cao *et al*., *Cancer Res* 2010). RNA-seq expression for the transcription factor corresponding to each motif is displayed as the median expression for PAX3-FOXO1+ alveolar RMS (ARMS) cells in the Cancer Cell Line Encyclopedia (CCLE). The PAX3-FOXO1 motif consists of a Paired Domain (PD) and a Homeobox Domain (HD) motif arranged in a divergent, head-to-head orientation. Representative PDB structures demonstrate the recognition of PD and HD motifs by short alpha-helices, where E-box motifs are recognized by the extended alpha-helices of bHLH family transcription factors. (**E**) ATAC-seq was performed in RH4 cells, and Tn5 insertion positions were calculated by the NucleoATAC pipeline. Insertion rates are displayed with respect to high confidence PAX3-FOXO1 (left) and FOXO1 (right) motif positions (vertical dashed lines) within PAX3-FOXO1 binding sites identified using FIMO. The median insertion rate outside and within the motif positions is displayed as a black and red horizontal line, respectively. (**F**) Heatmap of PAX3-FOXO1, histone marks and ATAC-seq in P3F binding sites categorized as promoters, enhancers, clustered enhancers, non-enhancers in RH4 cells (**G**) Analysis of H3K27ac and P3F in enhancer clusters. (Top, Middle) H3K27ac (Top) and P3F (Middle) signal over clustered enhancer regions with or without an overlapping P3F binding site. (Bottom) P3F and H3K27ac signal centered over P3F binding sites occurring within enhancer clusters. (**H**) Comparison of P3F vs. H3K27ac signal enrichment in all enhancer regions. (**I**) Analysis of public RNA-seq data. Genes linked to P3F binding site categories show little correlation with P3F expression.

While binding to inactive chromatin is a feature differentiating PFs from traditional TFs, a more rigorous definition of nucleosomal motif binding may be applied to understand how a TF with transforming potential is capable of invading repressed sites (*2, 26*). However, this function cannot be fully addressed at equilibrium, in which pioneer binding may result in rapid nucleosome eviction (*27*). Observing this phenomenon requires kinetic regulation of PAX3-FOXO1 to monitor chromatin state changes over short time scales. To address the limitations of studying FP-RMS cells at steady state, we employed a model system of immortalized human myoblasts (Dbt), engineered with a doxycycline-inducible PAX3-FOXO1 construct (Dbt/iP3F; **fig. S4A**) (*28*). With kinetic control of P3F expression, we set out to establish the immediate-early targets of P3F binding and observe its effects on nucleosome occupancy. We first performed spike-in normalized P3F ChIP-seq in Dbt/iP3F cells with 0, 8, and 24 hours of doxycycline treatment. At 8 hours, we identified 28,740 high-confidence P3F binding sites enriched primarily with bZIP and bHLH motifs, while the PAX3:FKHR motif ranked 14th among overrepresented sequences (**Fig. 2A**). By 24 hours, the number of P3F sites was reduced to 9,552, with the PAX3:FKHR motif ranked 1st. In addition to P3F binding sites being less numerous at the 24- compared to the 8-hour timepoint, sites at 24-hours also had a lower average intensity (**Fig. 2B** and **fig. S4B**). These results are consistent with the slow mobility of pioneer factors throughout chromatin, as they sample non-specific nucleosomal sites before equilibration in a DNA sequence-driven manner (*26*).

ATAC-seq revealed that P3F sites in Dbt/iP3F cells nearly universally exhibited increases in accessibility over the doxycycline treatment period, though we noted that many induced P3F sites begin with relatively high accessibility prior to induction (t_0, **fig. S4C**). To understand if accessibility changes were limited to sites with low vs. high initial accessibility, we distinguished P3F binding sites in Dbt/iP3F cells according to their baseline accessibility. We defined P3F binding site *Groups 1*, *2*, and *3* with high, medium, and low initial accessibility at t_0, respectively (**Fig. 2C**). In each *Group*, accessibility increased with time of P3F induction (**Fig. 2C**). We refined *Group 1*, *2*, and *3* regions to include only P3F sites that are observed at equilibrium in RH4 cells (referred to as *Clusters 1*, *2*, and *3*). ChIP-seq profiles in these regions indicate that the high P3F occupancy observed at 8 hours of expression in Dbt/iP3F cells is a transient state prior to the cells approaching equilibrium at the 24-hour timepoint, more closely resembling the binding profiles observed in FP-RMS cells (**Fig. 2D**). Additionally, we observed that the accessibility differences between *Clusters 1*, *2*, and *3* in Dbt/iP3F cells were clearly visible in RH4 ATAC-seq data, revealing that the Dbt/iP3F model faithfully recapitulates the targeting and retention of P3F to relatively inaccessible regions (**Fig. 2D** and **fig. S4D**). Overall these patterns may reflect a scenario of accessibility induction resulting from initial PAX3-FOXO1 binding, followed by an equilibrium period of P3F chromatin binding and release as additional factors are recruited to these regions.

Finally, we applied the NucleoATAC pipeline to determine if nucleosome positioning is affected over the time course of P3F induction and genomic occupancy (*29*). At t_0 we observed strong evidence for nucleosome occupancy in *Cluster 1*, *2*, and *3* sites (**fig. S2E**). Aligning nucleosome positions with respect to P3F peaks centers, we found that *Cluster 1* and *2* regions, exhibiting clear accessibility, had a pre-established nucleosome-depleted region (NDR) flanked by up- and downstream nucleosomes prior to P3F induction (**Fig. 2E**). *Cluster 3* sites with low baseline accessibility, exhibited more uniform nucleosome occupancy prior to P3F induction, with no apparent NDR. By 24 hours of P3F expression, the central NDR of *Clusters 1* and *2* was further remodeled, as evidenced by a widening and deepening of the nucleosome occupancy

curve. Remarkably, *Cluster 3* sites showed clear, *de novo* establishment of an NDR directly overlapping induced P3F binding events while evidence for well-positioned nucleosomes neighboring the P3F peak remained. This provides evidence of P3F targeting nucleosome-occupied sites followed by rapid remodeling events in inaccessible regions of the genome. Nearly identical patterns of equilibrium nucleosome occupancy were observed in RH4 cells, further reinforcing that nucleosomes are generally not retained following P3F binding (**Fig. 2F**). We noted that NucleoATAC applied to our 8-hour ATAC-seq dataset was inefficient for inferring nucleosome positions in all *Clusters* (**Fig. 2E**). Considering the NucleoATAC model inputs, we hypothesized that this was the result of differences in nucleosome-length fragments mapping to P3F sites, where a relative depletion of longer nucleosomal vs. shorter sub-nucleosomal fragments would result in low confidence nucleosome occupancy scores. Using plot2DO (*30*), we in fact observed differences in fragment length distributions consistent with our NucleoATAC result (**Fig. 2G**). We interpret this anomaly in nucleosome-length fragment distribution as evidence of local nucleosome redistribution taking place during the transition between baseline (t_0) and P3F-driven epigenetic equilibrium (24-hour). Supporting the involvement of direct P3F binding in this nucleosome disruption event, we observed that Tn5 insertion rates generally increased over time in all P3F site *Clusters* (**fig. S4F**), and particularly in *Cluster 3* we observed the gradual formation of a Tn5-protected footprint centered on high confidence P3F motifs over the 8- and 24-hour timepoints (**fig. S4G**).

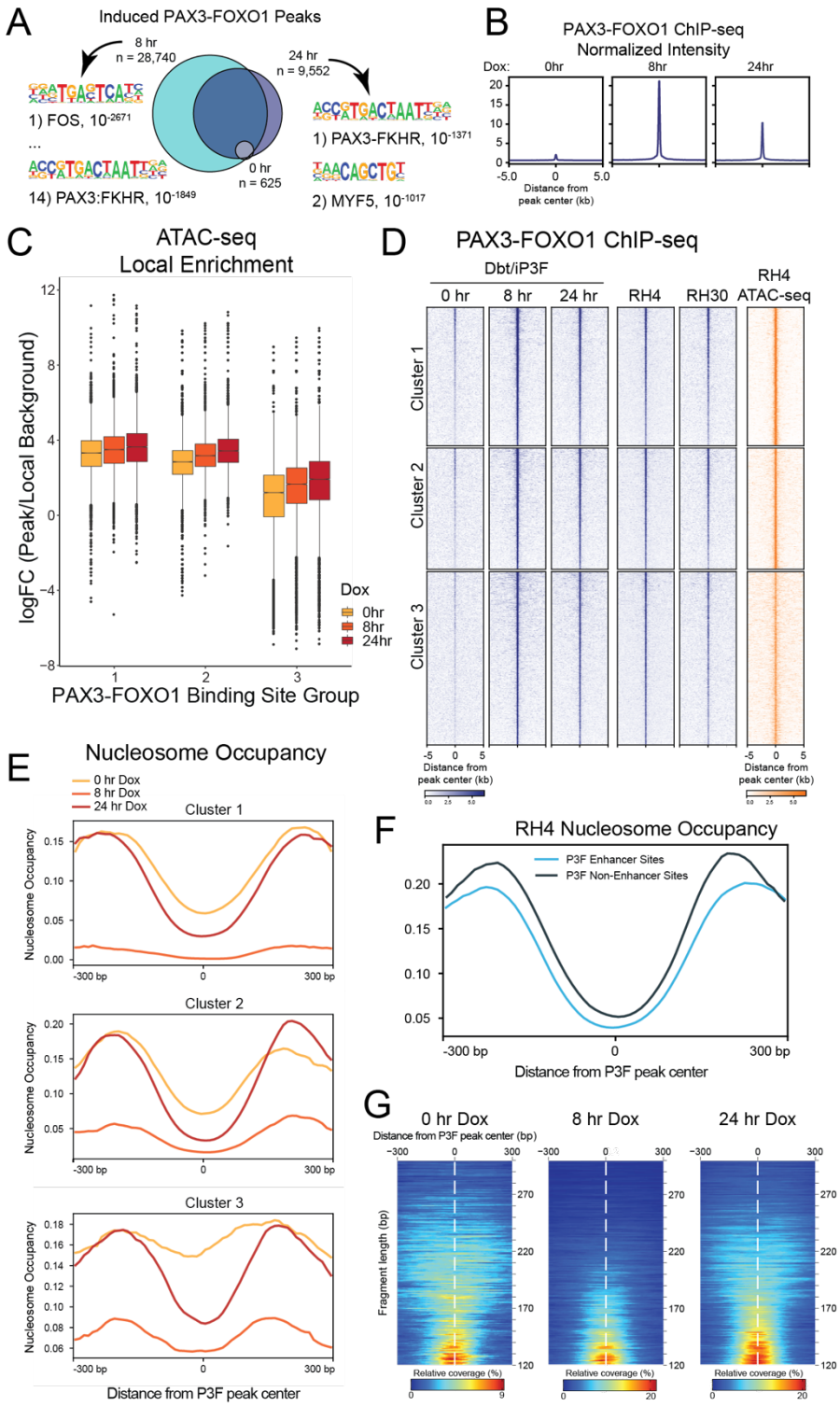


Fig. 2. Initial genomic targeting of PAX3-FOXO1 disrupts nucleosome occupancy. (**A**) PAX3-FOXO1 peaks called in Dbt/iP3F cells following treatment with 500 ng/mL doxycycline for 0, 8, or 24 hours. Motif analysis results are displayed for the 8 and 24 hour timepoints. (**B**) Spike-in normalized signal plot for PAX3-FOXO1 over the union of peaks identified at 0, 8, and 24 hours. (**C**) ATAC-seq data from Dbt/iP3F cells treated with doxycycline for 0, 8, and 24 hours was utilized for k-means clustering (k = 3) of induced PAX3-FOXO1 binding sites. Local enrichment of ATAC-seq signals (calculated as the fold change of the ATAC signal within binding sites vs. 10X-sized background regions immediately flanking each binding site) is displayed for PAX3-FOXO1 binding site *Groups 1*, *2*, and *3*. (**D**) Normalized ChIP-seq signal intensity and accessibility plotted across equilibrium PAX3-FOXO1 binding site *Clusters 1*, *2*, and *3*. (**E**) NucleoATAC was performed using time course ATAC-seq data in Dbt/iP3F cells across equilibrium PAX3-FOXO1 binding site *Clusters 1*, *2*, and *3*. Inferred nucleosome occupancy is displayed with respect to PAX3-FOXO1 peaks. (**F**) NucleoATAC of Enhancer and non-enhancer P3F binding sites in RH4 cells (**G**) plot2DO analysis of ATAC-seq fragment lengths present within induced PAX3-FOXO1 binding sites in Dbt/iP3F cells following 0, 8, and 24 hour doxycycline treatment.

Conclusion

We have critically evaluated the PAX3-FOXO1 fusion oncogene with respect to categorical properties intrinsic to the pioneer class of transcription factors. Advances in recent years have encouraged integrations across fields including pediatric oncology, genomics, and the biophysics of pioneer transcription factors (*2, 6, 12*). Our studies have revealed that for two pioneer factors fused as a chimera in a rare childhood tumor, chromatin recognition is consistent with PF function across the genome. In developing quantitative per-cell normalization for our genome-wide binding studies (pc-ChIP-seq; **Supplemental Methods**), we are able to infer nucleosome-targeting for PAX3-FOXO1, and also a non-diploid genome structure in rhabdomyosarcoma (*31*). We have discovered pioneer activity of the commonest driver alteration in rhabdomyosarcoma (*3*), which had previously uncharacterized function outside of active chromatin (*4, 18, 19*). Future efforts will be important to understand the kinetic rate constants of PAX3-FOXO1 dissociation from nucleosomes containing its motif (*32*), as well as focused chromatin sequencing of sonication-resistant binding sites (*20*). In the coming years we anticipate continued convergence of fields, with emerging evidence to understand and predict the logic of pioneer activity in development and disease.

Acknowledgements:

We thank M.G. Poirier, M. Parthun, C.E. Cottrell, B. Donovan, S.L. Lessnick, T.P. Cripe, and members of Stanton Lab for helpful conversations. B.Z.S. is grateful to the St. Baldrick's Foundation (Berry Strong fund), and Andrew McDonough B+ Foundation (Childhood Cancer Research Grant).

References:

1. M. Ptashne, A. Gann, Transcriptional activation by recruitment. *Nature* **386**, 569-577 (1997).
2. M. Fernandez Garcia *et al.*, Structural Features of Transcription Factors Associating with Nucleosome Binding. *Mol Cell* **75**, 921-932 e926 (2019).
3. N. Galili *et al.*, Fusion of a fork head domain gene to PAX3 in the solid tumour alveolar rhabdomyosarcoma. *Nat Genet* **5**, 230-235 (1993).
4. B. E. Gryder *et al.*, Histone hyperacetylation disrupts core gene regulatory architecture in rhabdomyosarcoma. *Nat Genet* **51**, 1714-1722 (2019).
5. J. F. Shern *et al.*, 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 (2014).
6. M. L. Suva, N. Riggi, B. E. Bernstein, Epigenetic reprogramming in cancer. *Science* **339**, 1567-1570 (2013).
7. B. Z. Stanton *et al.*, Smarca4 ATPase mutations disrupt direct eviction of PRC1 from chromatin. *Nat Genet* **49**, 282-288 (2017).
8. Y. Yu *et al.*, Expression profiling identifies the cytoskeletal organizer ezrin and the developmental homeoprotein Six-1 as key metastatic regulators. *Nat Med* **10**, 175-181 (2004).
9. S. S. Raghavan, K. L. Mooney, A. L. Folpe, G. W. Charville, OLIG2 is a marker of the fusion protein-driven neurodevelopmental transcriptional signature in alveolar rhabdomyosarcoma. *Hum Pathol* **91**, 77-85 (2019).

10. K. Takahashi, S. Yamanaka, Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663-676 (2006).
11. T. Vierbuchen *et al.*, Direct conversion of fibroblasts to functional neurons by defined factors. *Nature* **463**, 1035-1041 (2010).
12. B. A. Nacev *et al.*, The epigenomics of sarcoma. *Nat Rev Cancer* **20**, 608-623 (2020).
13. C. Kadoch, G. R. Crabtree, Reversible disruption of mSWI/SNF (BAF) complexes by the SS18-SSX oncogenic fusion in synovial sarcoma. *Cell* **153**, 71-85 (2013).
14. S. J. Miller *et al.*, Integrative genomic analyses of neurofibromatosis tumours identify SOX9 as a biomarker and survival gene. *EMBO Mol Med* **1**, 236-248 (2009).
15. L. A. Cirillo *et al.*, Binding of the winged-helix transcription factor HNF3 to a linker histone site on the nucleosome. *EMBO J* **17**, 244-254 (1998).
16. B. R. Zhou *et al.*, Distinct Structures and Dynamics of Chromatosomes with Different Human Linker Histone Isoforms. *Mol Cell*, (2020).
17. M. M. Brent, R. Anand, R. Marmorstein, Structural basis for DNA recognition by FoxO1 and its regulation by posttranslational modification. *Structure* **16**, 1407-1416 (2008).
18. B. E. Gryder *et al.*, PAX3-FOXO1 Establishes Myogenic Super Enhancers and Confers BET Bromodomain Vulnerability. *Cancer Discov* **7**, 884-899 (2017).
19. L. Cao *et al.*, Genome-wide identification of PAX3-FKHR binding sites in rhabdomyosarcoma reveals candidate target genes important for development and cancer. *Cancer Res* **70**, 6497-6508 (2010).
20. J. S. Becker *et al.*, Genomic and Proteomic Resolution of Heterochromatin and Its Restriction of Alternate Fate Genes. *Mol Cell* **68**, 1023-1037 e1015 (2017).
21. T. Hori *et al.*, Histone H4 Lys 20 monomethylation of the CENP-A nucleosome is essential for kinetochore assembly. *Dev Cell* **29**, 740-749 (2014).
22. S. S. Teves *et al.*, A stable mode of bookmarking by TBP recruits RNA polymerase II to mitotic chromosomes. *Elife* **7**, (2018).
23. B. E. Gryder, J. Khan, B. Z. Stanton, Measurement of differential chromatin interactions with absolute quantification of architecture (AQuA-HiChIP). *Nat Protoc* **15**, 1209-1236 (2020).
24. M. Ghandi *et al.*, Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503-508 (2019).
25. C. Y. McLean *et al.*, GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**, 495-501 (2010).
26. J. Lerner *et al.*, Two-Parameter Mobility Assessments Discriminate Diverse Regulatory Factor Behaviors in Chromatin. *Mol Cell* **79**, 677-688 e676 (2020).
27. C. Yan, H. Chen, L. Bai, Systematic Study of Nucleosome-Displacing Factors in Budding Yeast. *Mol Cell* **71**, 294-305 e294 (2018).
28. P. R. Pandey *et al.*, PAX3-FOXO1 is essential for tumour initiation and maintenance but not recurrence in a human myoblast model of rhabdomyosarcoma. *J Pathol* **241**, 626-637 (2017).

29. A. N. Schep *et al.*, Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Res* **25**, 1757-1770 (2015).
30. P. Beati, R. V. Chereji, Creating 2D Occupancy Plots Using plot2DO. *Methods Mol Biol* **2117**, 93-108 (2020).
31. L. Chen *et al.*, Clonality and evolutionary history of rhabdomyosarcoma. *PLoS Genet* **11**, e1005075 (2015).
32. B. T. Donovan, H. Chen, C. Jipa, L. Bai, M. G. Poirier, Dissociation rate compensation mechanism for budding yeast pioneer transcription factors. *Elife* **8**, (2019).

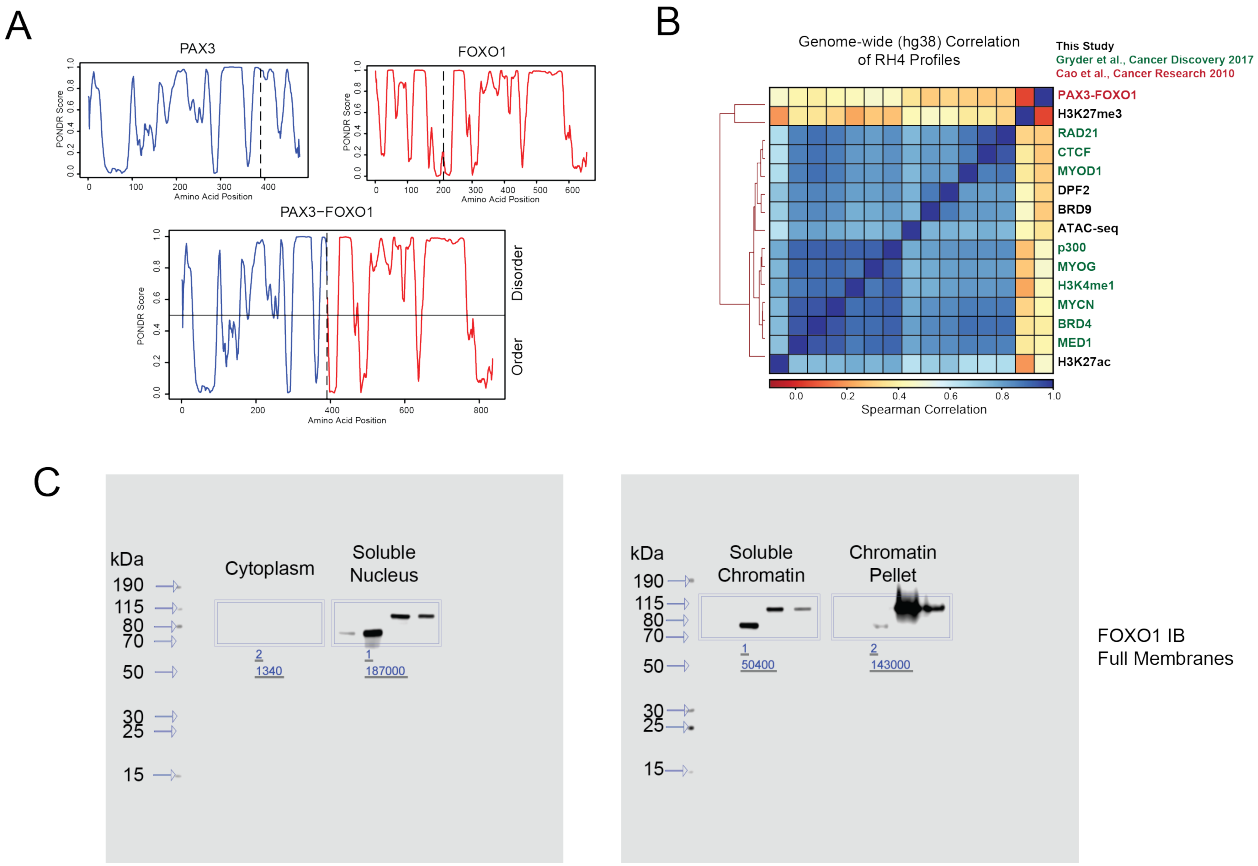
Supplementary Materials

Fig S1-S4

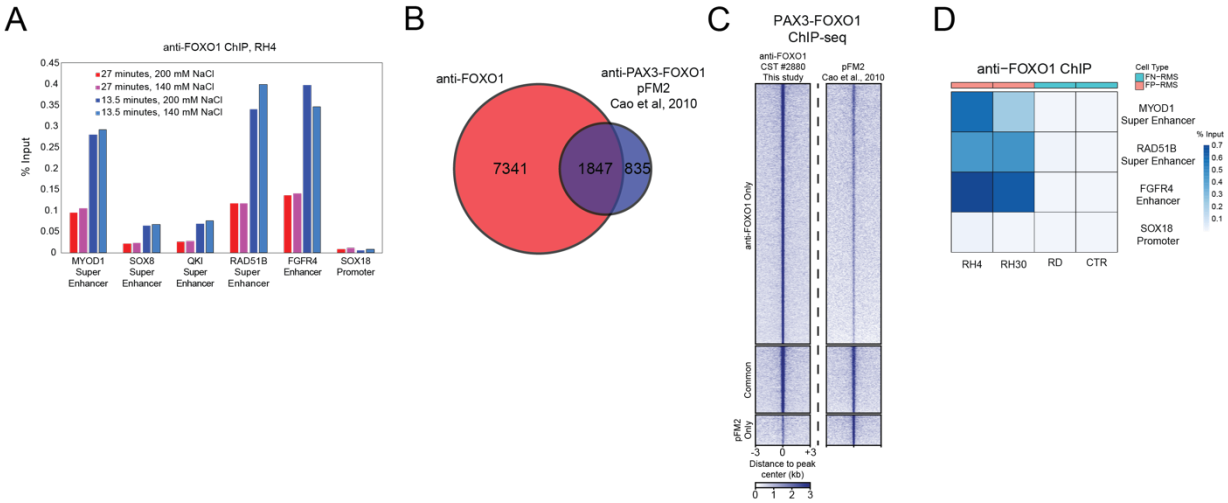
Materials and Methods

Additional References (1-17)

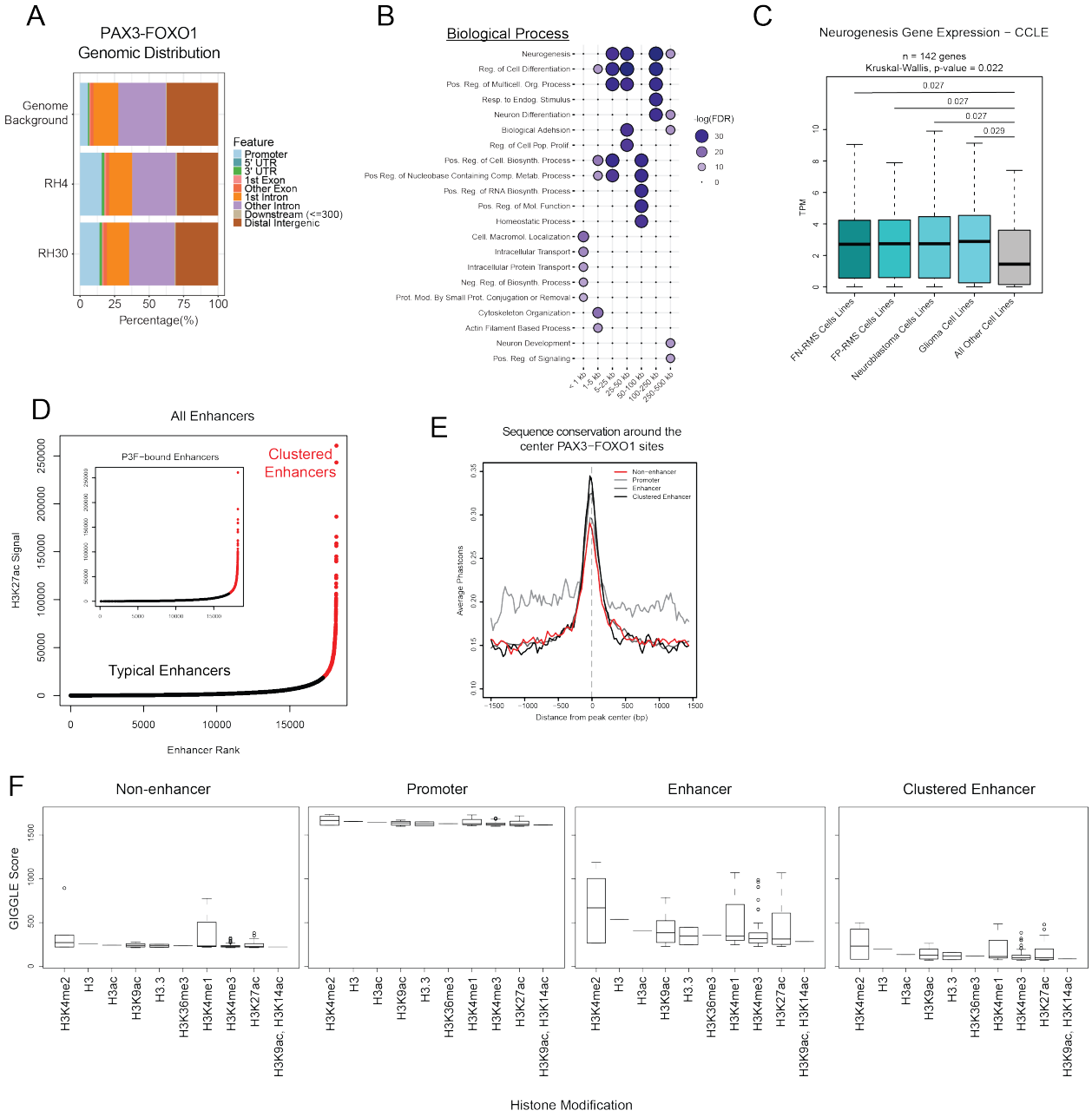
Supplementary Materials



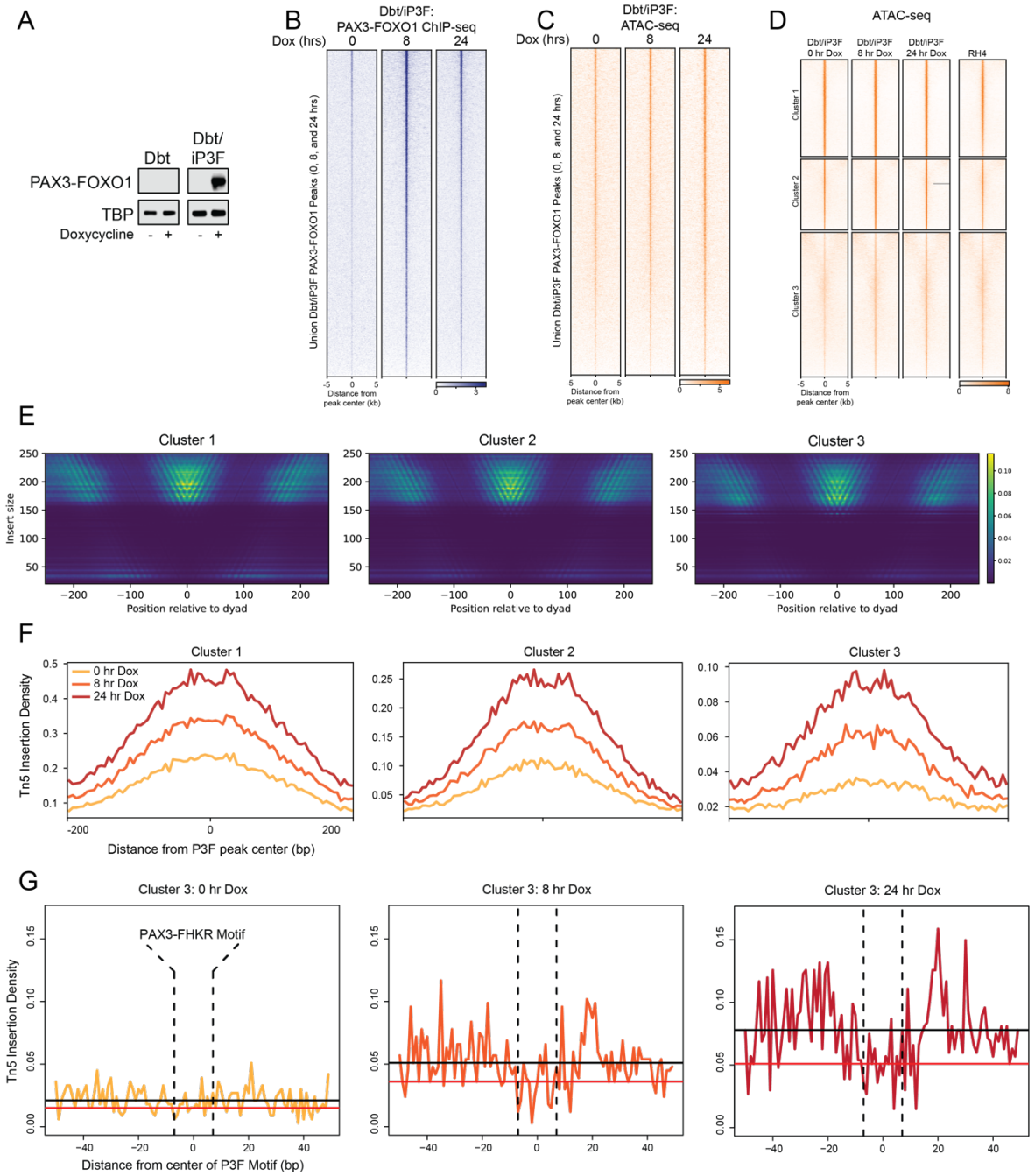
Supplemental fig. S1. (**A**) PAX3, FOXO1, and PAX3-FOXO1 amino acid sequences were analyzed by PONDR to calculate regions of ordered and disordered protein structure. (**B**) Epigenomic profiles prepared in this study (H3K27me3, H3K27ac, DPF2, BRD9, and ATAC-seq) or available publicly (RAD21, CTCF, MYOD1, p300, MYOG, H3K4me1, MYCN, BRD4, MED1 – Gryder *et al*., *Cancer Discovery* 2017 , and PAX3-FOXO1 – Cao *et al*., *Cancer Research* 2010) were used to compute genome-wide correlation values between each dataset. Clustered heatmap shows divergent binding profile for PAX3-FOXO1 with respect to all epigenomic features profiled. (**C**) Full membranes of anti-FOXO1 western blots probing cellular fractions prepared from RMS cell lines displayed in Fig. 1B.



Supplemental fig. S2. (**A**) ChIP-qPCR results following optimization of spike-in normalized PAX3-FOXO1 ChIP conditions in RH4 cells using a C-terminal FOXO1 antibody. Cells were sonicated for either 27 or 13.5 minutes and chromatin was immunoprecipitated in buffer containing either 200 or 140 mM NaCl. Optimal conditions were identified by comparing enrichment in known PAX3-FOXO1 binding sites (based on data from Cao *et al*., *Cancer Research* 2010) in the *MYOD1*, *SOX8*, *QKI*, *RAD51B*, and *FGFR4* loci compared to a negative control region in the *SOX18* promoter. (**B**) Overlap of PAX3-FOXO1 binding sites identified in this study with a FOXO1 antibody and sites identified by Cao *et al*. using an antibody (pFM2) spanning the PAX3-FOXO1 fusion breakpoint. (**C**) Heatmap of PAX3-FOXO1 enrichment in binding sites identified in this study, sites identified by Cao *et al*., or both. (**D**) Ant-FOXO1 ChIP-qPCR results in a panel of FP-RMS and FN-RMS cell lines.



Supplemental fig. S3. (**A**) Annotation of PAX3-FOXO1 binding sites identified in RH4 and RH30 cells. (**B**) PAX3-FOXO1 binding sites in RH4 cells were analyzed by GREAT. Genes linked to PAX3-FOXO1 binding sites were categorized according to distance from their TSS to the nearest binding site, and genes from from each category were utilized for gene ontology analysis. The top 5 ontologies for each category are displayed. (**C**) 142 genes belonging to the term "Neurogenesis" were queried against RNA-seq data from the Cancer Cell Line Encyclopedia. Expression level of these genes does not differ between FN-RMS, FP-RMS, Neuroblastoma, or Glioma cell lines, though these cell lines uniformly exhibit higher expression of neurogenic genes than all other cell lines (Kruskal-Wallis test, Benjamini-Hochberg p-value adjustment). (**D**) H3K27ac ChIP-seq data was analyzed by ROSE to identify weaker, typical enhancers (black points) vs. strong, clustered enhancers (red points). Inset: P3F occupies enhancers across the entire spectrum of H3K27ac strength. (**E**) Nucleotide sequence conservation of P3F binding site categories was calculated with phastCons. (**F**) The Cistrome DB Toolkit was used to assess similarity of publicly available histone modification/variant ChIP-seq profiles from various cell lines with P3F binding site categories in RH4 cells. High GIGGLE scores indicated stronger overlap



Supplemental fig. S4. (**A**) Western blot analysis of PAX3-FOXO1 protein expression in parental Dbt and inducible Dbt/iP3F cells following 24-hour treatment with 500 ng/mL doxycycline. (**B**) Heatmap of spike-in normalized PAX3-FOXO1 pc-ChIP-seq in Dbt/iP3F cells following 0, 8, or 24 hours of doxycycline treatment. (**C**) Heatmap of ATAC-seq in Dbt/iP3F cells following 0, 8, or 24 hours of doxycycline treatment. Vertical arrangement of peaks in **B** and **C** are independent. (**D**) Dbt/iP3F (0, 8, and 24 hour Dox) and RH4 ATAC-seq heatmaps over equilibrium PAX3-FOXO1 binding site *Clusters 1*, *2*, and *3*. (**E**) NucleoATAC analysis of baseline (0 hr) nucleosome positioning across equilibrium PAX3-FOXO1 binding site *Clusters* in Dbt/iP3F cells. Nucleosome- and sub-nucleosome-length ATAC-seq fragments are plotted with respect to high confidence nucleosome positions. (**F**) Tn5 insertion patterns from NucleoATAC were plotted over equilibrium PAX3-FOXO1 binding site *Clusters*. Each *Cluster* shows increasing Tn5 insertions with increasing time of doxycycline treatment. (**G**) Tn5 insertions were plotted with respect to high confidence PAX3-FKHR motif positions (indicated by vertical dashed lines) within *Cluster 3* binding sites. Tn5 insertion patterns reveal formation of a protected PAX3-FOXO1 footprint over time from 0 hours (left) to 24 hours (right) of doxycycline treatment. The median insertion rate outside and within the motif positions is displayed as a black and red horizontal line, respectively.

Materials and Methods

Cell Culture

RH4 (FP-RMS), RH30 (FP-RMS), RD (FN-RMS), and SMS-CTR (FN-RMS) cells, a gift from Dr. Peter Houghton (UTHSCSA), were cultured in high-glucose DMEM supplemented with 10% FBS, Glutamax, and penicillin/streptomycin. Immortalized Dbt myoblasts engineered with doxycycline-inducible PAX3-FOXO1 (Dbt/iP3F), engineered in the lab of Dr. Frederic Barr (NIH/NCI) (*1*), were cultured in Ham's/F-10 supplemented with 15% FBS, glutamine, sodium pyruvate, creatine monohydrate, uridine, and penicillin/streptomycin. Mouse C2C12 myoblasts were purchased from ATCC and cultured in high-glucose DMEM supplemented with 10% FBS, Glutamax, and penicillin/streptomycin. For PAX3-FOXO1 induction studies, Dbt/iP3F cells were seeded in normal culture media and grown to approximately 70% confluence before exchange with media containing 500 ng/mL doxycycline hyclate for the desired time period (8 or 24 hrs).

Cell Fractionation and Immunoblotting

Cells were collected from 15 cm plates and washed with ice-cold PBS containing protease inhibitor cocktail. Cell pellets were resuspended in Fractionation Buffer 1 (20 mM HEPES, 10 mM KCl, 0.2 mM EDTA) with protease inhibitor cocktail and incubated for 10 minutes on ice. NP-40 was added to a final concentration of 0.5% and the samples was vortexed on high for 15 seconds. Samples were incubated on ice for 1 minute, vortexed at high speed for 15 seconds, and pelleted at 14,000 rpm for 1 minutes at 4°C. The supernatant (cytoplasmic fraction) was transferred to a clean Eppendorf tube, and the nuclei pellet was resuspended in Fractionation Buffer 2 (10 mM Tris-HCl pH 8.0, 1 mM EDTA, 0.1% NP-40, 500 mM NaCl) with protease inhibitor cocktail and incubated at 4°C for 45-60 minutes with overhead rotation. The sample was centrifuged at 14,000 rpm for 10 minutes and the supernatant (soluble nuclear fraction) was transferred to a clean Eppendorf tube. The remaining chromatin pellet was resuspended in IP Buffer (50 mM Tris-HCl pH 8.0, 300 mM NaCl, 1 mM EDTA, 1% Triton X-100) with protease inhibitor cocktail and sonicated with an Active Motif EpiShear probe sonicator equipped with a cooled sonication platform for 5 minutes at 30% amplitude cycling from 30 seconds ON to 30 seconds OFF. The sample was centrifuged at 14,000 rpm for 10 minutes at 4°C and the supernatant (soluble chromatin fraction) was transferred to a clean Eppendorf tube. The remaining chromatin pellet was resuspended in 2X SDS protein sample buffer with 10% v/v 2-mercaptoethanol and heated at 95°C for 10 minutes. The other cell fractions were quantified with the Pierce Rapid Gold BCA Protein Assay Kit. Samples were resolved by SDS-PAGE using 2.5 µg of the cytoplasmic, soluble nuclear, and soluble chromatin fractions or 10 µL of the chromatin pellet fractions on a NuPAGE 4-12% Bis-Tris gel.

Proteins were transferred overnight at 4°C/30V to nitrocellulose membranes. Membranes were blocked at room temperature in 5% w/v milk solution in TBST (0.1% v/v Tween-20) for 1 hour before incubation for 2 hours at room temperature with primary antibodies detecting: MYCN (Cell Signlaing, 51705S), BAF155 (Cell Signaling, 11956S), FOXO1 (Cell Signaling, 2880S), HP1α (Cell Signaling, 2616S), H3K9me3 (Active Motif, 39062), H4K20me1 (Active Motif, 39727), or TBP (Cell Signaling, 44059S). Following 1 hour incubation with HRP-linked secondary antibodies, immunoblots were incubated with SuperSignal West Pico PLUS Chemiluminescent Substrate, images were acquired with a LI-COR C-DiGit Blot Scanner running Image Studio v5.2.

Amino Acid Primary Sequence Analysis

Human PAX3 (Uniprot ID: P23760) and FOXO1 (Uniprot ID: Q12778) amino acid sequence conservation was analyzed on the ConSurf Server using default parameters (*2*). ConSurf output files were used to project evolutionary conservation estimates, buried/exposed residue classifiers, and functional/structural residue classifiers onto the primary protein structure in **Fig. 1A**. Protein disorder analysis was conducted using the VL-XT PONDR algorithm (*3*) on the wild type PAX3 and FOXO1 protein sequences as well as the PAX3-FOXO1 fusion protein sequence.

PAX3-FOXO1 Spike-in ChIP Optimization

To facilitate quantitative normalization across ChIP-seq samples not only from different treatment conditions but also from distinct cell lines, we employed a spike-in ChIP strategy using known numbers of cells of human origin mixed in defined ratios with cells of mouse origin. This approach is similar in theory and in practice to the recently reported quantitative HiChIP method, known as AQuA-HiChIP (*4*). Our spike-in strategy addresses technical confounders introduced at two key points in the ChIP-seq protocol. Firstly, we introduce formaldehyde fixed mouse C2C12 cells to fixed human RMS or myoblast cells prior to sonication. In this study, the ratio of human:mouse cells is fixed at 3:1 for all experiments. This strategy ensures subsequent normalization steps retain read depth information on the basis of starting cell number. Failure to do so may obscure differences in chromatin output per cell across different cell types and treatment conditions. The assumption of equal chromatin produced per cell by sonication of any cell type implicit in commercial spike-in normalization reagents may be frequently violated when conducting experiments comparing aneuploid cancer cell lines, or in our case, RMS cells in which genome duplication may be a frequent event. Secondly, as in previous ChIP-Rx and commercial spike-in strategies (*5, 6*), we assume an equal number of mouse DNA fragments comprise the final sequencing libraries for input samples and for ChIP samples that will be directly compared. This permits us to properly correct for differences in sequencing depth across samples based on an internal and constant reference.

Our strategy utilizes one antibody, which we ideally expect to react with specific, conserved epitopes on human and mouse chromatin (e.g. histone modifications). In this ideal case, reliable and stable ChIP efficiency against epitopes on mouse chromatin ensures adequate and reproducible read numbers mapping to the spike-in mouse genome across all samples, while distinct human cell lines (e.g. RMS cells vs. myoblasts) under various treatment conditions (e.g. doxycycline induction) may produce a variable number of reads mapping to the human genome. In the case that an antibody has exquisite species-specific reactivity, or the desired epitope is not expressed in C2C12 cells (e.g. lineage restricted transcription factors), we leverage the inherently low signal-to-noise ratio of ChIP assays to our advantage. Here the number of non-specific reads mapping to the mouse genome is anticipated to remain constant across all ChIP samples, as even in highly efficient ChIP assays, background reads are present in relatively high proportions. Therefore, a single antibody approach is sufficient to produce ChIP samples with a constant number of spike-in reads mapping to the mouse genome for a given antibody, regardless of the reactivity of that antibody with mouse epitopes.

In addition to developing a novel, per-cell ChIP (pc-ChIP) approach, we tested two variables to optimize conditions for PAX3-FOXO1 ChIP using a FOXO1 antibody: 1) sonication time and 2) salt concentration in the ChIP buffer. The detailed pc-ChIP optimization strategy follows:

Cell fixation

FP-RMS, FN-RMS, Dbt/iP3F, or C2C12 cells were cultured in 15 cm plates, dissociated with trypsin, pelleted, and washed with PBS. Cell pellets were resuspended in Fixing Buffer (50 mM HEPES pH 7.3, 1 mM EDTA, 0.5 mM EDTA, 100 mM NaCl) and fresh, methanol-free formaldehyde was added to a final concentration of 1%. After 10 minutes of incubation at room temperature, the fixation was quenched by the addition of glycine to a final concentration of 125 mM and the cell suspension was placed on ice for 5 minutes. Fixed cells were pelleted at 1,200 x g for 5 minutes at 4°C and resuspended in ice-cold PBS containing a protease inhibitor cocktail. Fixed FP-RMS, FN-RMS, and Dbt/iP3F cells were aliquoted at $6x10^6$ cells/tube, and fixed C2C12 cells were aliquoted at $2x10^6$ cells/tube. Cells were pelleted at 3,000 rpm for 5 minutes at 4°C, the supernatant was removed, and pellets were snap frozen and stored at -80°C until further use.

Sonication

For initial optimization in RH4 cells, thawed aliquots of $6x10^6$ RH4 cells were combined with thawed aliquots of $2x10^6$ C2C12 cells in 800 uL of TE buffer pH 8.0 containing protease inhibitor cocktail and transferred to polystyrene sonication tubes. Samples were sonicated with an Active Motif EpiShear probe sonicator equipped with a cooled sonication platform for either 27 minutes or 13.5 minutes at 30% amplitude cycling from 30 seconds ON to 30 seconds OFF. After sonication, a 5 uL volume was aliquoted from each sample (input) and combined with 20 uL TE, 1 uL 10% SDS, and 1 uL 20 mg/mL Proteinase K for overnight decrosslinking at 65°C. The remaining sonicated chromatin was stored at 4°C. Input samples were purified using Qiagen MinElute PCR Purification columns and chromatin fragmentation was assessed by gel electrophoresis on an E-Gel 2% EX agarose gel.

ChIP

After evaluating fragmentation, sonicated chromatin in TE buffer was adjusted to ChIP Buffer by the addition of Triton X-100 (to 1% final concentration), SDS (to 0.1% final concentration), and sodium deoxycholate (to 0.1% final concentration). For initial optimization, we added sodium chloride to a final concentration of either 140 mM or 200 mM. Chromatin in ChIP Buffer was incubated on ice for 5 minutes and insoluble material was removed by centrifuging the sample at 13,000 rpm for 10 minutes at 4°C and transferring the supernatant to a new 1.5 mL tube. 5 uL of FOXO1 antibody (Cell Signaling Technology, #2880) was added, and samples were incubated at 4°C for 1 hour with overhead rotation. For each sample, 40 uL of Protein A Dynabeads were buffer exchanged with ChIP buffer containing either 140 mM or 200 mM NaCl before addition to the antibody:chromatin mixture, and the samples were incubated overnight at 4°C with overhead rotation. The beads were then washed twice with Low Salt Wash Buffer (0.1% SDS, 0.1% sodium deoxycholate, and 1% Triton X-100 in TE buffer pH 8.0), twice with ChIP Buffer (containing either 140 mM or 200 mM NaCl), twice with LiCl Wash Buffer (250 mM LiCl, 0.5% NP-40, 0.5% sodium deoxycholate in TE buffer pH 8.0), and twice with TE buffer pH 8.0. The beads were then resuspended in 100 uL TE buffer pH 8.0 with 2.5 uL 10% SDS and 5 uL 20 mg/mL Proteinase K, and ChIP samples were decrosslinked at 65°C overnight. ChIP DNA was purified with Qiagen MinElute PCR Purification columns.

Real-time PCR validation

Efficiency and specificity of our PAX3-FOXO1 ChIP conditions were first assessed by real-time PCR with primers designed against known PAX3-FOXO1 binding sites within the *MYOD1*, *SOX8*, *QKI*, *RAD51B*, and *FGFR4* loci (primer sequences listed below) (*7*). A negative control region in the *SOX18* promoter was also tested (*8*). This analysis revealed specific and robust ChIP enrichment within known PAX3-FOXO1 binding sites compared to the negative control region.

	Forward (5'-3')	Reverse (5'-3')	hg38 Region
MYOD1	CAGAACCATCCCATTCTCCG	GCCTGACCTTGAACGTGAAT	chr11:17650441-17650563
SOX8	GGATCGTGTAACCTGAGGGC	CAGTGAGTGTCTGCCTGCAA	chr16:1042251-1042350
QKI	TGCATGCTGGTGACAGATCA	ACAGCGTCCTCTTTCAGCTT	chr6:163777913-163778066
RAD51B	TTCCCATGAAAGGAGAAGCAGA	GGAAATGCCTCCACAGAAACG	chr14:68552591-68552741
FGFR4	AAATTTGACCTTCGTCGGCAC	CAGCTGTTGGCGATTTCACG	chr5:177105827-177105916
SOX18	GCTCTTGGTTCTCTGTCCCT	AGACAGACTGTGATGTGGGG	chr20:64050790-64051013

This optimization strategy revealed that 13.5 minutes of sonication followed by ChIP in buffer containing 200 mM NaCl was suitable for robust PAX3-FOXO1 ChIP enrichment using a FOXO1 antibody with limited background signal. All subsequent PAX3-FOXO1 ChIP assays in RH4, RH30, RD, SMS-CTR, and Dbt/iP3F cells were performed in an identical fashion.

Additional ChIP assays performed for this study were conducted in RH4 cells (with C2C12 spike-in) sonicated for 27 minutes, with immunoprecipitation performed in ChIP buffer with 200 mM NaCl using antibodies against DPF2 (Abcam, ab134942), BRD9 (Abcam, ab137245), H3K27ac (Active Motif , 39133), and H3K27me3 (Active Motif, 39155).

pc-ChIP-seq Considerations

We perform our pc-ChIP-seq spike-in and normalization on a "per cell" basis to retain information regarding the chromatin output across cell lines tested under various treatment conditions. Conversely, commercially available spike-in reagents recommend an equal starting amount of chromatin for each sample, regardless of starting cell number. For direct, quantitative comparison to be performed between distinct cell lines and treatments, the condition of equal chromatin produced per cell in all groups must be satisfied when normalizing on the basis of starting chromatin amount. In RMS cell lines and other aneuploid cancer cell lines, this condition is grossly violated. These cell lines generally contain variably sized genomes, and therefore they release different amounts of chromatin per cell that can globally influence the signal output from a standard ChIP assay. Using human/mouse read ratios in our input sequencing libraries, we can demonstrate the need to account for this by inferring the relative genome size for each cell line in this study. We used tetraploid C2C12 mouse cells for spike-in (estimated genome size, 5.4×10^9 bp) at a 3 to 1 ratio of human to mouse cells. We can therefore calculate the inferred genome sizes as:

$$Inferred\ Genome\ Size = (Observed\ Ratio \times 5.4 \times 10^9)/3$$

Observed Read Ratio (human/mouse)	Inferred Genome Size ($\times 10^9$ bp)
RH4 Input = 4.3	7.74
RH30 Input = 2.92	5.26
RD Input = 5.57	10.02
SMS-CTR Input = 3.1	5.58
Dbt/iP3F Input = 3.18	5.72

These figures reveal vastly different relative genome sizes across cell lines tested and showed consistent results across RH4 replicates (7.72 and 7.92 $\times 10^9$ bp, respectively). This is consistent with the scenario where chromatin output differs globally, and therefore relative read depth across both input and ChIP samples differ. Under these conditions, a spike-in and normalization procedure based on the starting amount of *chromatin* instead of the starting number of *cells* would obscure the differences between different biological samples. This prevents qualitative comparison of peak calls as wells as quantitative comparison of signal strength observed. Therefore, we needed a new strategy where the difference of amount of chromatin released by each cell is taken into consideration, and we developed pc-ChIP-seq to correct this bias. While the focus here is on chromosome ploidy resulting in differential chromatin output per cell, epigenetic repression and de-repression as well as cell cycle stage are examples of other conditions that may globally influence chromatin output per cell owing to differences in sonication sensitivity.

To justify the assumption that a one-antibody strategy is sufficient for a spike-in ChIP-seq normalization approach, even in the event that the chosen antibody does not recognize epitopes on mouse chromatin to produce high quality ChIP-seq data in the spike-in genome, we first analyzed our anti-FOXO1 ChIP-seq data from RH4/C2C12 cells with the ENCODE pipeline, aligning only to mm10. This analysis, prior to down-sampling, revealed just 135 FOXO1 peaks in the mouse genome. More importantly, just 0.18% of all reads aligning to the mouse genome mapped to these peaks. Second, in replicate PAX3-FOXO1 ChIP-seq datasets produced in RH4 cells, the ratio of human to mouse reads in the ChIP samples were 4.57 and 4.52. Together, these findings suggest even when non-specific mouse reads are not informative for peak calling in the mouse genome, they comprise a substantial and reproducible portion of a ChIP sample that can be leveraged for normalization. This reflects the inherently low signal-to-noise ratio of ChIP assays, where a typical transcription factor ChIP will often have a relatively low FRiP value.

pc-ChIP Library Construction and Sequencing

ChIP and input DNA from each sample were prepared for sequencing as before (*9*) by blunt end repair using the Lucigen End-It DNA End-Repair Kit, 3' A-tailing by Klenow fragment (3'-5' exo-), adaptor

ligation by T4 DNA ligase, and size selection on an E-Gel 2% EX agarose gel. Libraries were amplified with barcoded primers for 14 cycles and isolated from unreacted primers by gel purification. Pooled libraries were sequenced at the Nationwide Children's Hospital Institute for Genomic Medicine, Genomic Services Laboratory on a HiSeq4000 running in paired-end, 150 bp mode.

pc-ChIP-seq Normalization

As the pc-ChIP protocol rigorously controls for the starting number of cells and the ratio of human to mouse cells across all samples in each experimental group, read depth bias correction is achieved through normalization using mouse spike-in reads across all input and ChIP samples for a given comparison. We carried out normalization prior to peak calling and generation of signal files in order to perform quantitative comparisons across cell lines or test samples. In detail, our approach for random down-sampling of sequencing reads across samples is as follows:

BCL converted paired end fastq files were aligned to hg38 and mm10 reference genomes, separately, utilizing bowtie2 (*10*) and following ENCODE best practices. The resulting mouse BAMs were analyzed by samtools to calculate the number of reads aligned to each genome. The minimum number of mouse aligned reads (m) was identified across all samples, and was divided by the number of aligned mouse reads for each sample (s) to calculate the scaling factor (f), $f=m/s$. This scaling factor was subsequently used to normalize the hg38 aligned BAMs through subsampling with samtools view -s, which retains read pair information. Picard SamToFastq converted the resulting SAM files to paired end fastq files.

ChIP-seq Data Analysis

Normalized pc-ChIP-seq fastq files from this study as well as publicly available PAX3-FOXO1 ChIP-seq fastq files for RH4 cells (GSE19063, (*7*)) were processed with the ENCODE ChIP-seq pipeline with chip.xcor_exclusion_range_max set at 30. Normalized, paired-end fastqs were aligned with bowtie2 (version 2.3.4.3) to hg38, with parameters bowtie2 -X2000 --mm. Next, blacklisted region, unmapped, mate unmapped, not primary alignment, multi-mapped, low mapping quality (MAPQ<30), duplicate reads and PCR duplicates were removed. Peaks were called with MACS2 (version 2.2.4), with parameters -p 1e-2 --nomodel --shift 0 --extsize $[FRAGLEN] --keep-dup all -B –SPMR, where FRAGLEN is the estimated fragment length. IDR analyses were performed on peaks from replicate samples or pseudo-replicates, with threshold 0.05. Motif analysis with HOMER (version 4.11.1) (*11*) was then carried out on conservative IDR peaks. For visualization, bedGraph files were generated with MACS2 bdgcmp from the pile-up, and then converted to bigwig format with bedGraphToBigWig. Heatmaps were generated with deeptools (version 3.3.1). k-means classification from deeptools was used to generate peak clusters for **Fig. 2**. ROSE (v0.1) (*12*) was used with our H3K27ac IDR peaks to classify enhancers and to stitch proximate H3K27ac regions into clustered enhancers.

ATAC-seq and ATAC-seq Data Analysis

ATAC-seq was performed as previously described (*13*) with only minor modifications. $5x10^4$ cells per experiment were first washed with RSB buffer (10 mM Tris-HCl pH 8, 10 mM NaCl, 3 mM $MgCl_2$) and gently permeabilized with RSB lysis buffer (10 mM Tris-HCl pH 8, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% NP-40) on ice. Cells were suspended in 50 uL of tagmentation master mix prepared from Illumina Tagment DNA TDE1 Enzyme and Buffer Kit components (#20034197), and transposition was performed for 30 minutes at 37°C. Tagmented DNA fragments were isolated using Qiagen MinElute PCR Purification columns prior to library amplification. ATAC-seq libraries were amplified with barcoded Nextera primers for 14 cycles, and excess primers were removed by size selection with AMPure XP beads. Libraries were sequenced on the HiSeq4000 platform running in PEx150bp mode.

The ENCODE ATAC-seq pipeline (https://github.com/ENCODE-DCC/atac-seq-pipeline) with default parameters was used to process ATAC-seq data. First, reads are scanned for adaptor sequences and trimmed

with cutadapt (version 2.3). Reads are then mapped to hg38 with bowtie2 (version 2.3.4.3). Properly aligned, non-mitochondrial read pairs were retained for peak calling with MACS2 (version 2.2.4). After peaks are called, heatmaps are generated with deeptools (version 3.3.1) (*14*). Fragment length distributions were generated with plot2DO v1.0 (https://github.com/rchereji/plot2DO) (*15*). Local signal vs background enrichment is calculated with localEnrichmentBed (https://github.com/dariober/bioinformatics-cafe/tree/master/localEnrichmentBed). Nucleosome position, nucleosome occupancy and tn5 insertion density are estimated using NucleoATAC (version 0.3.4) (*16*). Transcription factor foot printing/motif protection was assessed by identifying PAX3-FKHR or FOXO1 motif positions within PAX3-FOXO1 binding sites using FIMO (*17*). Insertion rates were subsequently plotted over aligned motif positions using deeptools (version 3.3.1).

Data Accessibility

Raw sequence files for each dataset prepared for this manuscript have been uploaded to the Gene Expression Omnibus (GEO) and are freely available. Also available are processed files (hg38) generated for immediate use. These include all genome-wide signal files (bigwig format) and PAX3-FOXO1 peak calls (bed format). PAX3-FOXO1 signals and peaks are directly comparable across cell lines and conditions, as they have been collectively normalized for this study.

Code Availability

Scripts will be made available upon request.

Supplemental References

1. P. R. Pandey *et al.*, PAX3-FOXO1 is essential for tumour initiation and maintenance but not recurrence in a human myoblast model of rhabdomyosarcoma. *J Pathol* **241**, 626-637 (2017).
2. C. Berezin *et al.*, ConSeq: the identification of functionally and structurally important residues in protein sequences. *Bioinformatics* **20**, 1322-1324 (2004).
3. P. Romero *et al.*, Sequence complexity of disordered protein. *Proteins* **42**, 38-48 (2001).
4. B. E. Gryder, J. Khan, B. Z. Stanton, Measurement of differential chromatin interactions with absolute quantification of architecture (AQuA-HiChIP). *Nat Protoc* **15**, 1209-1236 (2020).
5. D. A. Orlando *et al.*, Quantitative ChIP-Seq normalization reveals global modulation of the epigenome. *Cell Rep* **9**, 1163-1170 (2014).
6. B. Egan *et al.*, An Alternative Approach to ChIP-Seq Normalization Enables Detection of Genome-Wide Changes in Histone H3 Lysine 27 Trimethylation upon EZH2 Inhibition. *PLoS One* **11**, e0166438 (2016).
7. L. Cao *et al.*, Genome-wide identification of PAX3-FKHR binding sites in rhabdomyosarcoma reveals candidate target genes important for development and cancer. *Cancer Res* **70**, 6497-6508 (2010).
8. M. E. Yohe *et al.*, MEK inhibition induces MYOG and remodels super-enhancers in RAS-driven rhabdomyosarcoma. *Sci Transl Med* **10**, (2018).
9. B. L. Kidder, K. Zhao, Efficient library preparation for next-generation sequencing analysis of genome-wide epigenetic and transcriptional landscapes in embryonic stem cells. *Methods Mol Biol* **1150**, 3-20 (2014).
10. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359 (2012).
11. S. Heinz *et al.*, Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589 (2010).
12. W. A. Whyte *et al.*, Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307-319 (2013).

13. J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, W. J. Greenleaf, Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213-1218 (2013).
14. F. Ramirez *et al.*, deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160-165 (2016).
15. P. Beati, R. V. Chereji, Creating 2D Occupancy Plots Using plot2DO. *Methods Mol Biol* **2117**, 93-108 (2020).
16. A. N. Schep *et al.*, Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Res* **25**, 1757-1770 (2015).
17. C. E. Grant, T. L. Bailey, W. S. Noble, FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017-1018 (2011).