Cross-species analysis of enhancer logic using deep learning

Liesbeth Minnoye, Ibrahim Ihsan Taskiran, David Mauduit, Maurizio Fazio, Linde van Aerschot, Gert Hulselmans, Valerie Christiaens, Samira Makhzami, Monika Seltenhammer, Panagiotis Karras, et al.

Method

Cross-species analysis of enhancer logic using deep learning

Liesbeth Minnoye,[1,2,13] Ibrahim Ihsan Taskiran,[1,2,13] David Mauduit,[1,2] Maurizio Fazio,[3,4] Linde Van Aerschot,[1,2,5] Gert Hulselmans,[1,2] Valerie Christiaens,[1,2] Samira Makhzami,[1,2] Monika Seltenhammer,[6,7] Panagiotis Karras,[8,9] Aline Primot,[10] Edouard Cadieu,[10] Ellen van Rooijen,[3,4] Jean-Christophe Marine,[8,9] Giorgia Egidy,[11] Ghanem-Elias Ghanem,[12] Leonard Zon,[3,4] Jasper Wouters,[1,2] and Stein Aerts[1,2]

[1]VIB-KU Leuven Center for Brain and Disease Research, 3000 Leuven, Belgium; [2]KU Leuven, Department of Human Genetics KU Leuven, 3000 Leuven, Belgium; [3]Howard Hughes Medical Institute, Stem Cell Program and the Division of Pediatric Hematology/Oncology, Boston Children's Hospital and Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts 02115, USA; [4]Department of Stem Cell and Regenerative Biology, Harvard Stem Cell Institute, Cambridge, Massachusetts 02138, USA; [5]Laboratory for Disease Mechanisms in Cancer, KU Leuven, 3000 Leuven, Belgium; [6]Center for Forensic Medicine, Medical University of Vienna, 1090 Vienna, Austria; [7]Division of Livestock Sciences (NUWI) - BOKU University of Natural Resources and Life Sciences, 1180 Vienna, Austria; [8]VIB-KU Leuven Center for Cancer Biology, 3000 Leuven, Belgium; [9]KU Leuven, Department of Oncology KU Leuven, 3000 Leuven, Belgium; [10]CNRS-University of Rennes 1, UMR6290, Institute of Genetics and Development of Rennes, Faculty of Medicine, 35000 Rennes, France; [11]Université Paris-Saclay, INRA, AgroParisTech, GABI, 78350 Jouy-en-Josas, France; [12]Institut Jules Bordet, Université Libre de Bruxelles, 1000 Brussels, Belgium

Deciphering the genomic regulatory code of enhancers is a key challenge in biology because this code underlies cellular identity. A better understanding of how enhancers work will improve the interpretation of noncoding genome variation and empower the generation of cell type–specific drivers for gene therapy. Here, we explore the combination of deep learning and cross-species chromatin accessibility profiling to build explainable enhancer models. We apply this strategy to decipher the enhancer code in melanoma, a relevant case study owing to the presence of distinct melanoma cell states. We trained and validated a deep learning model, called DeepMEL, using chromatin accessibility data of 26 melanoma samples across six different species. We show the accuracy of DeepMEL predictions on the CAGI5 challenge, where it significantly outperforms existing models on the melanoma enhancer of *IRF4*. Next, we exploit DeepMEL to analyze enhancer architectures and identify accurate transcription factor binding sites for the core regulatory complexes in the two different melanoma states, with distinct roles for each transcription factor, in terms of nucleosome displacement or enhancer activation. Finally, DeepMEL identifies orthologous enhancers across distantly related species, where sequence alignment fails, and the model highlights specific nucleotide substitutions that underlie enhancer turnover. DeepMEL can be used from the Kipoi database to predict and optimize candidate enhancers and to prioritize enhancer mutations. In addition, our computational strategy can be applied to other cancer or normal cell types.

[Supplemental material is available for this article.]

A cell's phenotype arises from the expression of a unique set of genes, which is regulated through the binding of transcription factors (TFs) to *cis*-regulatory regions, such as promoters and enhancers. Deciphering gene regulatory programs entails mapping the network of TFs and *cis*-regulatory regions that govern the identity of a given cell type, as well as understanding how the specificity of such a network is encoded in the DNA sequence of genomic enhancers. Profiling accessible chromatin via DNase I hypersensitive sequencing (DNase-seq) or via the assay for transposase-accessible chromatin using sequencing (ATAC-seq) represents a useful approach for identifying putative enhancers (Song and Crawford 2010; Buenrostro et al. 2013; Klemm et al. 2019). Indeed, active enhancers are typically depleted of one or more nucleosomes owing to the binding of TFs. Initial changes in DNA accessibility can be facilitated through a special class of TFs that bind with high affinity to their recognition sites and that have a long residence time at the enhancer, sometimes referred to as pioneer TFs (Zaret and Carroll 2011; Klemm et al. 2019). By displacing nucleosomes or thermodynamically outcompeting nucleosome binding, they allow other TFs to cobind, thereby further stabilizing the nucleosome-depleted region and/or actively enhancing transcription of target genes (Grossman et al. 2018; Jacobs et al. 2018; Dodonova et al. 2020).

Because the presence and architecture of TF binding sites within enhancers determines which TFs can bind with high affinity, understanding this "enhancer logic" can help interpret the functional role of enhancers within a gene regulatory network.

[13]These authors contributed equally to this work.
Corresponding author: stein.aerts@kuleuven.vib.be
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.260844.120. Freely available online through the *Genome Research* Open Access option.

Several techniques exist to study the enhancer code, including (1) motif discovery tools (Bailey et al. 2009; Heinz et al. 2010; Thomas-Chollier et al. 2011, 2012; Janky et al. 2014; Imrichová et al. 2015); (2) comparative genomics (Ballester et al. 2014; Prescott et al. 2015; Villar et al. 2015); (3) genetic screens (Gasperini et al. 2019; Kircher et al. 2019); and (4) machine learning techniques (Park and Kellis 2015). In particular, the latter has seen a strong boost in recent years with the advent of large training sets derived from genome-wide profiling. Three pivotal methods based on deep learning include DeepBind (Alipanahi et al. 2015), DeepSEA (Zhou and Troyanskaya 2015), and Basset (Kelley et al. 2016), the first convolutional neural networks (CNNs) applied to genomics data (Eraslan et al. 2019). Since their emergence in the genomics field, machine learning techniques, and especially CNNs, have been applied to model a range of regulatory aspects, including cross-species enhancer predictions (Min et al. 2016; Quang and Xie 2016; Chen et al. 2018), TF binding sites (Wang et al. 2018; Avsec et al. 2020), DNA methylation (Angermueller et al. 2017), and 3D chromatin architecture (Schreiber et al. 2017).

Deciphering gene regulation and the underlying enhancer code is not only important during dynamic processes such as development, but also in disease contexts such as cancer, where gene regulatory networks are typically misregulated owing to mutations. Particularly in melanoma, a type of skin cancer that develops from melanocytes, gene expression is misregulated and highly plastic (Shain and Bastian 2016; Rambow et al. 2019). This gives rise to two main melanoma cell states: the melanocytic (MEL) state, which still resembles the cell of origin, expressing high levels of the melanocyte-lineage specific transcription factors MITF, SOX10, and TFAP2A, as well as typical pigmentation genes such as *DCT*, *TYR*, *PMEL*, and *MLANA*; and the mesenchymal-like (MES) state, in which the cells are more invasive and therapy resistant, expressing high levels of genes involved in TGFB signaling and epithelial-to-mesenchymal transition (EMT)-related genes (Hoek et al. 2006, 2008; Verfaillie et al. 2015; Rambow et al. 2019; Wouters et al. 2020). These transcriptomic differences have also been studied at the epigenomics level, with AP-1 and TEAD factors as master regulators of the MES state and binding sites for SOX10 and MITF significantly enriched in MEL-specific regulatory regions (Verfaillie et al. 2015; Bravo González-Blas et al. 2019; Wouters et al. 2020). However, it remains unclear how these regulatory states are encoded in particular enhancer architectures and whether such architectures are evolutionary conserved. Besides human cell lines and human patient–derived cultures, several animal models have been established in melanoma research, including mouse, pig, horse, dog, and zebrafish (Egidy et al. 2008; Seltenhammer et al. 2014; van der Weyden et al. 2016; van Rooijen et al. 2017; Segaoula et al. 2018; Prouteau and André 2019). Although these models are widely used, it is unknown whether their enhancer landscapes and regulatory programs are conserved with human. Here, we take advantage of these animal model systems and combine cross-species chromatin accessibility profiling with deep learning, to investigate enhancer logic in melanoma.

Results

Melanoma chromatin accessibility landscapes are conserved across species

We profiled chromatin accessibility using ATAC-seq on a collection of melanoma cell lines across six species, for a total of 26 samples (Fig. 1A). These include 16 human patient–derived cultures (MM lines) (Gembarska et al. 2012; Verfaillie et al. 2015), one mouse cell line (Dankort et al. 2009), primary melanoma cells from the pig melanoma model MeLiM (MeLiM) (Egidy et al. 2008), two horse melanoma lines derived from a Grey Lipizzaner horse (HoMel-L1) and from an Arabian horse (HoMel-A1) (Seltenhammer et al. 2014), two dog melanoma cell lines from oral and uveal sites (Dog-OralMel-18249 and Dog-IrisMel-14205, respectively; Cani-DNA BRC: https://dog-genetics.genouest.org), and four melanoma lines established from zebrafish (ZMEL1, EGFP-121-1, EGFP-121-5, and EGFP-121-3) (White et al. 2008, 2011). Per sample, between 65,475 and 176,695 ATAC-seq peaks were called, with distinct levels of conservation of accessibility across the species (Fig. 1A; Supplemental Fig. S1A). The difference in the number of peaks across the samples is attributable, on the one hand, to genome size (Supplemental Fig. S1B), and on the other hand, to data quality (measured as the fraction of reads in peaks [FRiP]) (Supplemental Fig. S1C).

Unsupervised clustering of the 16 human lines revealed two distinct groups (Supplemental Fig. S1D), which correspond to the two main cell states in human melanoma, that is, the melanocytic state (MEL) and mesenchymal-like state (MES), as was further confirmed for most of the cell lines by previously generated RNA-seq data (Supplemental Fig. S1E; Verfaillie et al. 2015) and corroborated by previous studies using epigenomics data (Verfaillie et al. 2015; Wouters et al. 2020). Indeed, regulatory regions near MEL-specific genes such as *SOX10* are accessible in human lines in the MEL state (MM001, MM011, MM031, MM034, MM052, MM057, MM074, MM087, MM118, MM122, and MM164), whereas they are closed in MES melanoma lines (MM029, MM099, MM116, MM163, and MM165) (Fig. 1B). As in Wouters et al. (2020), we observed heterogeneity between samples of the MEL state (Supplemental Fig. S1D).

To enable the comparison of chromatin accessibility between human and other species, we first identified regulatory regions that are alignable (i.e., have a high sequence similarity) between species using the liftOver tool (at least 10% of bases must remap) (Meyer et al. 2012). When such an alignable region contains an ATAC-seq peak in the compared species, it is referred to as a "conserved accessible" region. Between 1.1% and 40.9% of the ATAC-seq regions in non-human lines were conserved accessible in human (Fig. 1C), and between 0.9% and 18.4% of the human peaks were conserved accessible in the other species (Supplemental Fig. S1F). Accordingly, we identified 303,392 alignable and 10,592 conserved accessible regions across all mammalian species. This number decreases when including zebrafish, to 29,619 alignable regions and, only 116 conserved accessible regions. Nearly half of the 10,592 conserved accessible mammalian regions were promoters within 1 kb of a transcription start site (Supplemental Fig. S1G). Indeed, high conservation of proximal promoters has previously been reported (Villar et al. 2015). In each of the mammalian species, the 10,592 conserved accessible regions were more accessible compared to all ATAC-seq regions; in addition, they show a higher ChIP-seq signal for acetylation of histone H3 at lysine 27 (H3K27ac) in human, a mark for active regulatory regions (Supplemental Fig. S1H,I; Creyghton et al. 2010), and higher sequence conservation compared to alignable regions as measured by phastCons and phyloP (Supplemental Fig. S1J; Siepel 2005; Pollard et al. 2010). Nevertheless, although ATAC-seq regions are nucleosome depleted and often bound by several TFs, they are not necessarily active enhancers, because accessibility does not directly translate to enhancer activity (Shlyueva et al. 2014).

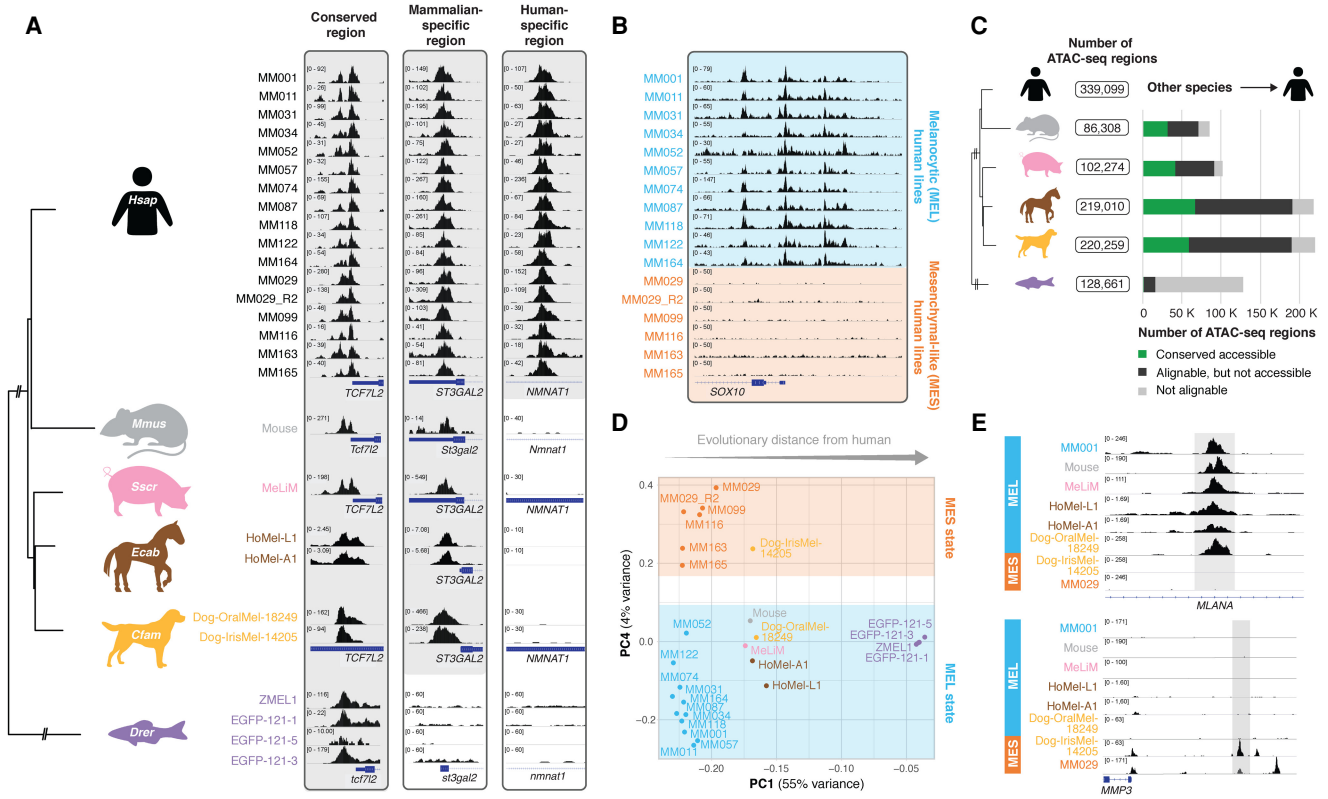

Figure 1. Comparative epigenomics reveals conservation of two main melanoma states. (*A*) Evolutionary relationship between the six studied species, represented by a phylogenetic tree (NCBI taxonomy tree). ATAC-seq profiles of the 26 melanoma cell lines are shown for three regulatory regions. (*B*) ATAC-seq profiles of the human melanoma lines for the *SOX10* locus. Lines are colored by the melanocytic (MEL, in blue) or mesenchymal-like (MES, in orange) melanoma state. (*C*) Total number of ATAC-seq regions observed across all samples of a species are colored based on whether they are not alignable, alignable, or conserved accessible in human. (*D*) PCA clustering based on the accessibility of the 29,619 alignable regions across all six species. (*E*) ATAC-seq profiles of MEL and MES lines of different species for an intronic *MLANA* enhancer and the upstream region of *MMP3*.

Next, we examined whether the MEL and MES melanoma states are conserved in the other species of our cohort. Clustering all mammalian samples based on the accessibility of the 303,392 alignable regions (Supplemental Fig. S1K), or of all samples (including zebrafish) using the 29,619 alignable regions (Fig. 1D), revealed two axes of variation between the samples, namely (1) the evolutionary variation between species and (2) the distinction between the melanoma states. All human MEL samples are clustered together with nine of the 10 non-human lines, indicating that most of the non-human cell lines are epigenomically similar to the human MEL lines. Conversely, the dog cell line Dog-IrisMel-14205 clustered together with the human MES samples, which indicates that Dog-IrisMel-14205 belongs to the MES state. This classification of melanoma samples was reflected in their accessibility at known MEL and MES regulatory regions such as the intronic enhancer of *MLANA*, a MEL-specific gene involved in melanosome biogenesis (De Mazière et al. 2002), and an enhancer upstream of *MMP3*, a gene that increases metastatic potential in melanoma cell lines (Fig. 1E; Shoshan et al. 2016). Classifying the cross-species samples based on a principal component analysis (PCA) of only the conserved accessible regions (i.e., without species-specific or clade-specific peaks) clearly revealed the MEL-MES distinction, whereas the species variation was less outspoken (Supplemental Fig. S1L,M).

In conclusion, by using ATAC-seq on a panel of 26 melanoma lines across six species, conserved accessible regulatory regions could be identified. These regions allowed clustering of the melanoma samples into two groups that correspond to the two main melanoma cell states, indicating conservation of the MES melanoma state in dog and the MEL melanoma state in pig, mouse, horse, dog, and even zebrafish melanoma samples.

Conservation of transcription factor motifs in state-specific enhancers

Next, we investigated whether TF binding motifs that are specific to the MEL and MES states are conserved across species. To this end, we performed differential motif enrichment between MEL and MES accessible regions for human and dog, because these were the two species in our cohort for which cell lines of both states were identified above. Differential peak calling ($\log_2$FC > 2.5 and $P_{Adj} < 0.0005$), followed by motif enrichment using HOMER (Heinz et al. 2010), revealed a highly similar enrichment of SOX, TFAP2 family, E-box, RUNX, and ETS TF binding motifs in both the human and dog MEL-specific peaks (Fig. 2A,B; for complete HOMER output, see Supplemental Table S1). The enriched motifs of the TFAP2 family can most likely be linked to TFAP2A because this is a master regulator in human melanocytes and melanoma (Seberg et al. 2017). Similarly, the observed E-box and SOX motifs most likely represent MITF and SOX10, respectively, because they are among the previously reported master regulators in human MEL lines (Hoek et al. 2006; Verfaillie et al. 2015; Bravo González-Blas et al. 2019; Wouters et al. 2020). Likewise, motif

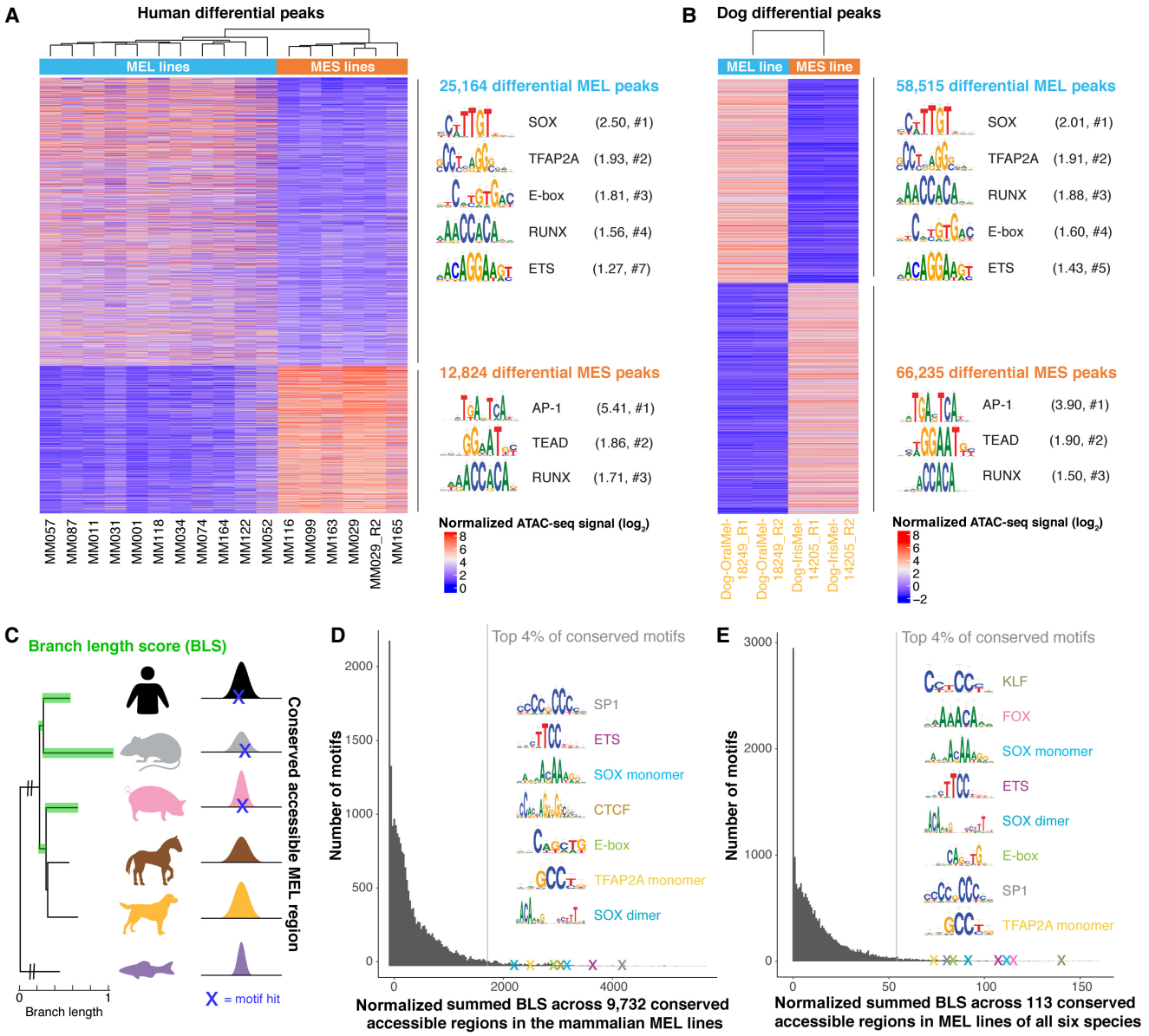

Figure 2. Conservation of binding motifs of master regulators of MEL and MES melanoma states. (*A*,*B*) Heatmap of differential ATAC-seq regions when comparing human MEL versus human MES lines (*A*) and the MEL dog line "Dog-OralMel-18249" versus the MES dog line "Dog-IrisMel-14205" (two biological replicates each) (*B*), colored by normalized ATAC-seq signal. Enriched TF binding motifs in the differential peaks were identified via HOMER (Heinz et al. 2010), and the first logo of enriched TF families is shown. The ratio of the percentage of target and background sequences with the motif is indicated between brackets, as well as the rank of the TF class within the HOMER output (#). (*C*) Schematic overview of cross-species motif analysis using the branch length score (BLS) as a measure for the evolutionary conservation of a motif hit across conserved accessible regions. The BLS was summed across a set of conserved accessible regions. (*D*,*E*) Histogram of the normalized summed BLS score for 20,003 motifs on 9732 conserved accessible regions across the mammalian MEL lines (*D*) and on 113 conserved accessible regions across MEL lines of all six species (*E*). The first hit of the top recurrent TF binding motifs within the top 4% conserved motifs is indicated as a cross and is accompanied by the logo of the motif.

enrichment in the MES regions is very similar between human and dog, revealing AP-1 and TEAD motifs as most highly enriched (Fig. 2A,B), corroborating earlier findings (Verfaillie et al. 2015). Together, these observations indicate that the MEL and MES melanoma cell states are conserved in dog and that they are likely governed by the same master regulators, based on the concordance of motif enrichment.

To further verify the importance of the MEL-specific master regulators in MEL cell lines of the remaining four species, we applied a different strategy because we could not contrast MEL and MES lines for horse, pig, mouse, and zebrafish. We analyzed 9732 accessible regions that are conserved accessible across all mammalian MEL lines to identify conserved TF binding sites. We scanned these regions using the cisTarget motif collection (v8) (Herrmann et al. 2012; Janky et al. 2014; Imrichová et al. 2015) containing 20,003 TF position-weight matrices (PWMs) and used a branch length score (BLS) to calculate the level of evolutionary conservation of each TF binding motif (Fig. 2C), a

strategy applied before in other systems (Stark et al. 2007; Jacobs et al. 2018). Among the 4% most conserved motifs were SP1, ETS, SOX, CTCF, MITF, and TFAP2A motifs (Fig. 2D). The top conserved motifs were members of the SP/KLF TF family, which bind to GC-rich motifs in promoters (Dynan and Tjian 1983). Indeed, 47% of the 9732 conserved accessible regions in mammalian MEL lines are proximal promoters (≤1 kbp from TSS). BLS scoring on the remaining 5196 more distal conserved accessible regions revealed similar highly conserved motifs, except for SP/KLF TF family motifs, indicating that distal regions, such as enhancers, mostly contain the state-specific TF binding motifs (Supplemental Fig. S1N). In the 113 conserved accessible regions across the MEL cell lines across all six species, BLS scoring again revealed SOX, ETS, MITF, and TFAP2A motifs among the most conserved motifs (Fig. 2E).

In conclusion, two independent strategies of motif analysis suggest conservation of TF binding sites for known melanoma master regulators, with conserved SOX10, MITF, TFAP2A, and ETS TF family motif enrichment in MEL enhancers across all six studied species.

Deep neural network DeepMEL reveals nucleotide-resolution enhancer logic

Although motif enrichment can predict candidate regulators, we sought to build a more comprehensive model of the MEL enhancers, which would allow cross-species predictions and in-depth analysis of enhancer architecture. To this end, we trained a deep learning (DL) model on the human ATAC-seq data. First, to construct an unsupervised training set, we clustered all 339,099 human ATAC-seq peaks using cisTopic—a probabilistic framework to analyze scATAC-seq data that can also be applied to bootstrapped bulk ATAC-seq data (Bravo González-Blas et al. 2019; Methods)—into 24 "topics" or sets of coaccessible regions (Fig. 3A; Supplemental Fig. S2A,B). This provided a nuanced classification, with topic 4 and topic 7 representing the MEL- and MES-specific enhancers, respectively, being accessible across all MEL or MES samples (Fig. 3A; Supplemental Fig. S2C). In addition, we found two topics with regions that are generally accessible across all cell lines (topic 1 and topic 19) (Fig. 3A; Supplemental Fig. S2C). These ubiquitously accessible regions are highly enriched for proximal promoters (Supplemental Fig. S2D) and for known promoter-specific TF binding motifs linked to SP and NFY TF families (Supplemental Fig. S2C; Dynan and Tjian 1983; Maity and de Crombrugghe 1998). Other topics were more specific to one or a small group of cell lines (Fig. 3A). We verified the biological relevance of these topics by Gene Ontology (GO) enrichment of flanking genes using GREAT (McLean et al. 2010). Genes near topic 4 regions are significantly enriched for GO terms such as pigmentation (FDR = 1.95×10^{-8}) and neural crest cell differentiation (FDR = 4.26×10^{-7}), whereas genes near topic 7 regions were enriched for GO terms involved in cell–cell adhesion (1.56×10^{-13}). Motif discovery on the top regions assigned to each topic confirmed enrichment of SOX, ETS, TFAP2A, and MITF motifs in the MEL topic regions (topic 4) and AP-1 in the MES topic (topic 7) (Supplemental Fig. S2C). An example topic 4 region in the promoter of the SOX10 target gene *MIA* (Graf et al. 2014) is shown in Figure 3B, as well as two topic 7 regions upstream of *SERPINE1,* a gene expressed in metastatic melanoma (Klein et al. 2012).

Using the 24 topics as classes, we trained a multiclass, multilabel classifier using a neural network, called "DeepMEL" (Fig. 3C). As input, we used the forward and reverse complement of 500-bp sequences centered on the ATAC-seq summit. As topology, we used the DanQ CNN-RNN hybrid architecture (Quang and Xie 2016) consisting of four main layers: a convolution layer to discover local patterns in sequential data, followed by a max-pooling layer to reduce the dimensionality of the data and generalize the model effectively, a bidirectional recurrent layer (LSTM) to detect long-range dependencies of the local patterns discovered in the first layer, and finally a fully connected (dense) layer just before the output layer to help the classification after the feature extraction layers (Fig. 3C). Note that several hyperparameters, including the number and size of the convolutional filters and the length of the input DNA sequence, were optimized to yield the final model (Supplemental Fig. S3; Supplemental Note). After successful training of DeepMEL—area under the receiver operating characteristic curve (auROC) = 0.863 and area under the precision recall curve (auPR) = 0.374 on test data for topic 4 regions (Fig. 3D,E; Supplemental Fig. S4A)—we used the weights of the neurons from the convolutional filters to extract local patterns learned by the model. We transformed these convolution filters into PWMs and found the importance of each filter for each topic (Methods). Filters that represent SOX, MITF, TFAP2A, and RUNX motifs were most relevant for the MEL-specific topic 4; filters that represent AP-1, TEAD, and RUNX binding sites were assigned to the MES-specific topic 7 (Fig. 3F). Thus, DeepMEL learned the relevant features de novo from the sequence. The 3885 regions classified as MEL-specific in MM001 (topic 4 scores above threshold of 0.16) (Methods) were not only highly accessible in MEL lines and closed in MES lines (Supplemental Fig. S4B), but were also accessible in human melanocytes (Supplemental Fig. S4C), indicating that MEL-specific melanoma regions are not cancer-specific but already accessible in their cell of origin, that is, the melanocytes. As a consequence, we can potentially extrapolate the observations on this topic to normal melanocyte enhancers. Although in the remainder of this work we will score accessible regions to identify functional enhancers, it is also possible to score the entire genome, without filtering for ATAC-seq peaks (Supplemental Fig. S4D).

To examine the TF binding site architecture within enhancers, we used a model interpretation tool, DeepExplainer (Lundberg and Lee 2017; Avsec et al. 2020; Lundberg et al. 2020). For a MEL enhancer located on the fourth intron of *IRF4,* nucleotides important for classifying this enhancer as topic 4 emerge as motifs for SOX10, MITF, TFAP2A, and RUNX factors (Fig. 3G, top two rows; for another example, see Supplemental Fig. S4E,F).

It is known that enhancer accessibility does not directly translate to enhancer activity (Shlyueva et al. 2014). To test whether the same TF binding motifs contribute to the activity of MEL enhancers, we used the *IRF4* enhancer as case study. For this enhancer, Kircher et al. (2019) performed saturation mutagenesis followed by an in vitro massively parallel reporter assay (MPRA), testing the effect of every possible single-nucleotide mutation on enhancer activity (Fig. 3G, third row). The most deleterious mutations coincided with the DeepMEL-predicted SOX, E-box, and RUNX-like motifs, overlapping with nucleotides that also have the strongest in silico effect (Fig. 3G, last row), indicating that the predicted motifs are actually contributing to enhancer activity. In addition, the magnitude of the in silico predicted effect highly correlates with the effect of the in vitro mutations (Spearman's correlation of 0.60) (Fig. 3G,H). These observations indicate that, although DeepMEL was trained to predict binary enhancer accessibility, it is also a good predictor of enhancer activity of this specific enhancer. DeepMEL predictions outperform other classifiers and deep

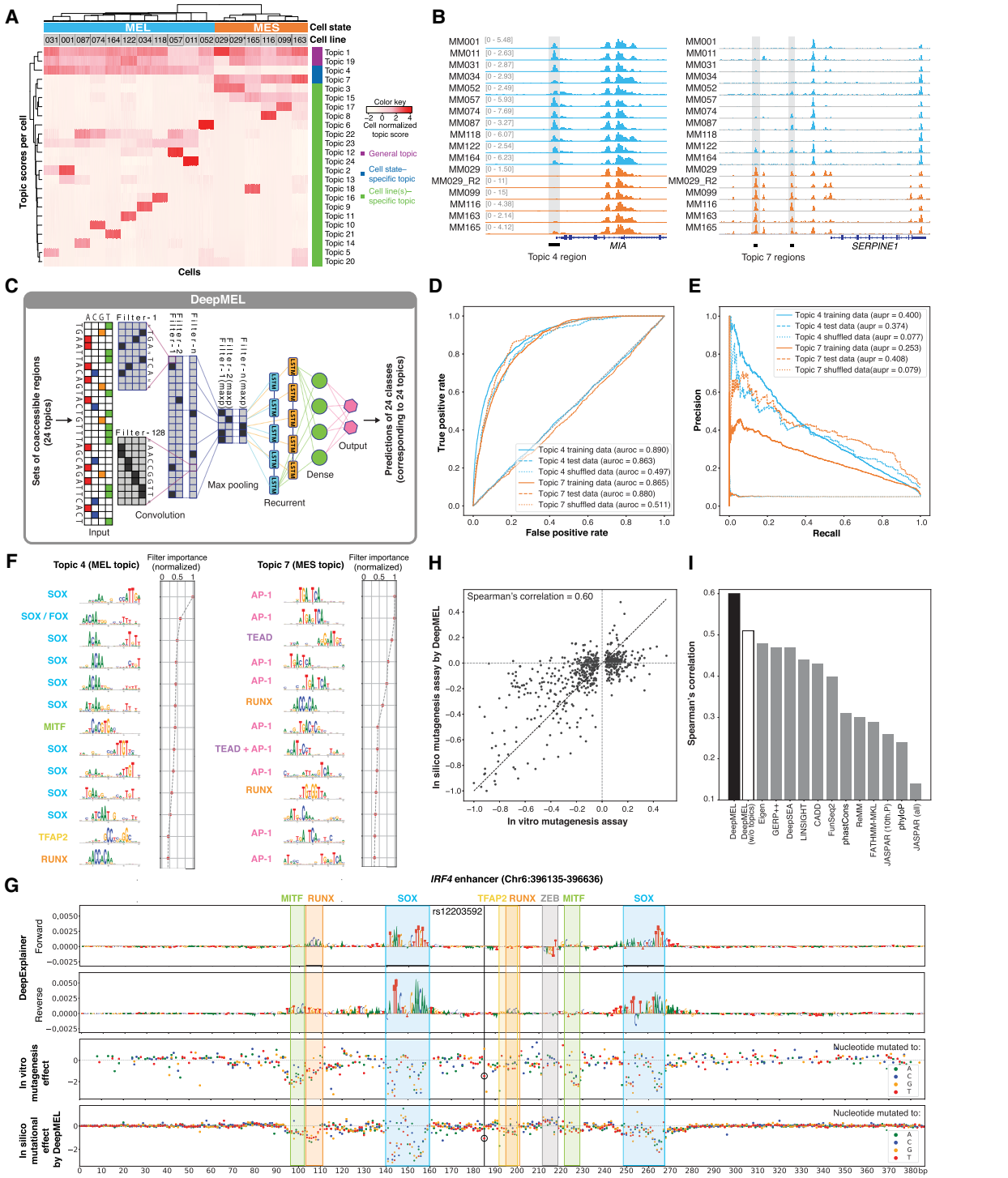

Figure 3. DeepMEL classifies melanoma enhancers and predicts important TF binding motifs. (*A*) Cell-topic heatmap of cisTopic applied to 339,099 ATAC-seq regions across the 16 human melanoma lines, colored by normalized topic scores. (029*) MM029_R2. (*B*) Example regions of a MEL-specific (topic 4) region near *MIA* and MES-specific (topic 7) regions upstream of *SERPINE1*. (*C*) Schematic overview of DeepMEL. Twenty-four topics or sets of coaccessible regions were used as input for training of a multiclass multilabel neural network. (*D*,*E*) Receiver operating characteristic curve (*D*) and precision recall curve (*E*) for DeepMEL on training, test, and shuffled data of topic 4 and topic 7 regions. (*F*) Top enriched filters learned by DeepMEL to classify regions as MEL (topic 4) or MES (topic 7). Normalized filter importance is shown per filter. (*G*) Example of a MEL-predicted enhancer near *IRF4*. (First and second rows) DeepExplainer view of the forward and reverse strand, with the height of the nucleotides indicating the importance for prediction of the MEL enhancer. (Third row) In vitro effect of point mutations on enhancer activity as measured by MPRA (Kircher et al. 2019). Colors represent the nucleotide to which the wild-type nucleotide is mutated. (Fourth row) In silico effect of point mutations as predicted by DeepMEL. (*H*) Correlation between the in vitro mutational effects on the *IRF4* enhancer and the in silico mutagenesis predictions. (*I*) Performance of variant effect prediction of DeepMEL using topics (black bar, model used in this paper) or using ATAC-seq signal (white bar), and several previously tested models on the *IRF4* enhancer case (Kircher et al. 2019).

learning models that were benchmarked in Kircher et al. (2019) (Fig. 3I). One possible explanation for this improvement is that DeepMEL uses more nuanced topics (Fig. 3I, black bar) rather than the ATAC-seq signal of the different MM lines as labels (Fig. 3I, white bar). Enhancer accessibility and activity cannot only be influenced by mutations that break a motif for an activating TF, but also by the creation of a repressor binding motif, as was, for instance, the case for the SNP rs12203592 (Fig. 3G; Supplemental Fig. S4G).

In conclusion, DeepMEL, trained on topics of human coaccessible regions, is performant in classifying melanoma regulatory regions into different classes based on purely the DNA sequence. Features learned by DeepMEL correspond to TF binding motifs of master regulators of specific classes. These motifs can also be located and visualized within regions using a model interpretation tool, allowing examination of the motif architecture within specific enhancers, and predicting the effect of mutations on enhancer accessibility.

Cross-species scoring identifies orthologous melanoma enhancers

Next, we asked whether the human-trained model DeepMEL can be used to predict MEL and MES enhancers in other species. We started with the dog genome as a test case, because the differential ATAC-seq peaks between the MEL (Dog-OralMel-18249) and MES (Dog-IrisMel-14205) dog cell lines can serve as true positives (Fig. 4A). DeepMEL reached similar performance in human and dog for predicting MEL and MES regions, and this accuracy is significantly higher compared to using *cis*-regulatory module (CRM) scoring with PWMs (Fig. 4A). Having confirmed that the human model can identify enhancers in the dog genome, we predicted MEL and MES enhancers across all six species. This furthermore allowed us to order all samples according to the MEL-MES axis (Fig. 4B). Between 2093 and 5400 MEL enhancers were predicted, and between 7459 and 10,743 MES enhancers, in samples of the MEL and MES state, respectively (Fig. 4B). The majority of these enhancers could not have been detected using whole-genome alignments (liftOver) (Supplemental Fig. S5A–E). Of note, predicted MEL enhancers in the pig melanoma cells (MeLiM) were similarly accessible in pig melanocytes (Supplemental Fig. S5F), again indicating that MEL melanoma enhancers can be used as a model for melanocyte enhancers.

Next, we compared the occurrence of MEL enhancers between species in relation to putative target genes. Particularly, we looked at enhancers located near a set of 379 human genes that are specifically expressed in the MEL state (Methods). Of these 379 genes, 217 (67%) had at least one MEL-predicted enhancer within 200 kb upstream of and downstream from the gene. Between 70% and 85% of the orthologous MEL genes in other species had at least one MEL enhancer 200 kb upstream of and downstream from the gene (Supplemental Fig. S5G). Only a small subset of these enhancers could have been found using liftOver (2%–43%, depending on the species). Of these genes, 32 form a core set of conserved MEL-specific genes throughout all species including zebrafish, each having a MEL enhancer nearby. Examples of genes in the core set are *MITF*, *PMEL*, and *TYRP1*, genes known to be involved in melanocyte development, melanosome formation, and melanin production (D'Mello et al. 2016).

A long-standing question in enhancer studies is how to compare enhancers with each other, if their sequences do not align (Cliften et al. 2001; Arunachalam et al. 2010). Here, we tackle this question by using the dense layer of DeepMEL as a reduced dimensional space to calculate the correlation between enhancers. Using this measure we found that MEL-predicted enhancers in proximity of orthologous MEL genes are significantly more similar to each other compared to both MEL-predicted enhancers in proximity of different MEL genes within the same species (Fig. 4C), and redundant (or shadow) (Hong et al. 2008) enhancers linked to the same MEL gene in a species, as well as random non-MEL ATAC-seq peaks near homologous MEL genes (Supplemental Fig. S5H). This altogether supports the idea that MEL enhancers near orthologous genes are indeed orthologous enhancers.

Last, we studied an example of a MEL enhancer in more detail, namely the enhancer near *ERBB3*. DeepMEL predicts a MEL enhancer upstream or intronic of *ERBB3* in each of the mammalian species, which were also found by liftOver of the human *ERBB3* enhancer (Fig. 4D, II). However, in the zebrafish genome, liftOver was unable to identify the homologous region, whereas DeepMEL predicted two MEL enhancers, one upstream of the TSS of *erbb3b* and another in the first intron. Both zebrafish enhancers were highly correlated with the human *ERBB3* enhancer (deep layer Pearson's correlation of 0.812 and 0.797 for the upstream and intronic zebrafish enhancer, respectively), suggesting that both enhancers are orthologous to the human *ERBB3* enhancer. Applying DeepExplainer to the multiple-aligned sequences revealed a conserved motif architecture in the orthologous mammalian *ERBB3* enhancers containing each three SOX motifs and one TFAP2A motif (Fig. 4D, III). In mouse, one SOX binding site was lost, and mouse is also the mammalian species that is most distant from human, among the included mammals in this study (Fig. 4D, I). The two zebrafish enhancers have a highly similar motif architecture, suggesting that they arose by duplication from a common ancestor enhancer.

In conclusion, we showed that DeepMEL is able to identify MEL- and MES-specific enhancers in different species, which allows studying evolutionary events and enhancer logic within orthologous enhancers, even in distant species such as zebrafish.

Motif architecture of the MEL enhancer

To study the architecture of MEL enhancers in more detail, including motif composition, motif order and distance, and relationships to the position of nucleosomes, we set out to obtain high-confidence motif annotations in each of the 3885 MEL enhancers in human (MM001, the most MEL-like human cell line), for each of the predicted core regulatory factors (SOX10, MITF, TFAP2A, RUNX). To achieve this, we devised an optimized motif scoring method that obtains precise positions of TF binding motifs by multiplying DeepMEL activation scores of convolutional filters (i.e., motifs) with the DeepExplainer profile of each enhancer (Fig. 5A; Methods; Shrikumar et al. 2019).

The first observation was that each MEL enhancer contains at least one SOX10 motif hit, and often two or more (Fig. 5B). This suggests that SOX10 plays a central role in MEL enhancer accessibility. Indeed, knockdown (KD) of SOX10 in MM001 significantly decreases the accessibility of MEL enhancers (Supplemental Fig. S6A), and the regions that close after SOX10-KD are highly enriched for SOX motifs (NES = 28.5), possibly revealing a pioneering-role of SOX10 in MEL enhancers. Next to SOX motifs, a combination of one or multiple TFAP2A, MITF, or RUNX-like motif hits were present in 84% of the MEL-predicted enhancers (Fig. 5B). Next, to facilitate a systematic study of the MEL enhancer logic, we binarized the motif-region matrix to simplify the region clustering (Fig. 5C). We obtained eight different enhancer classes,

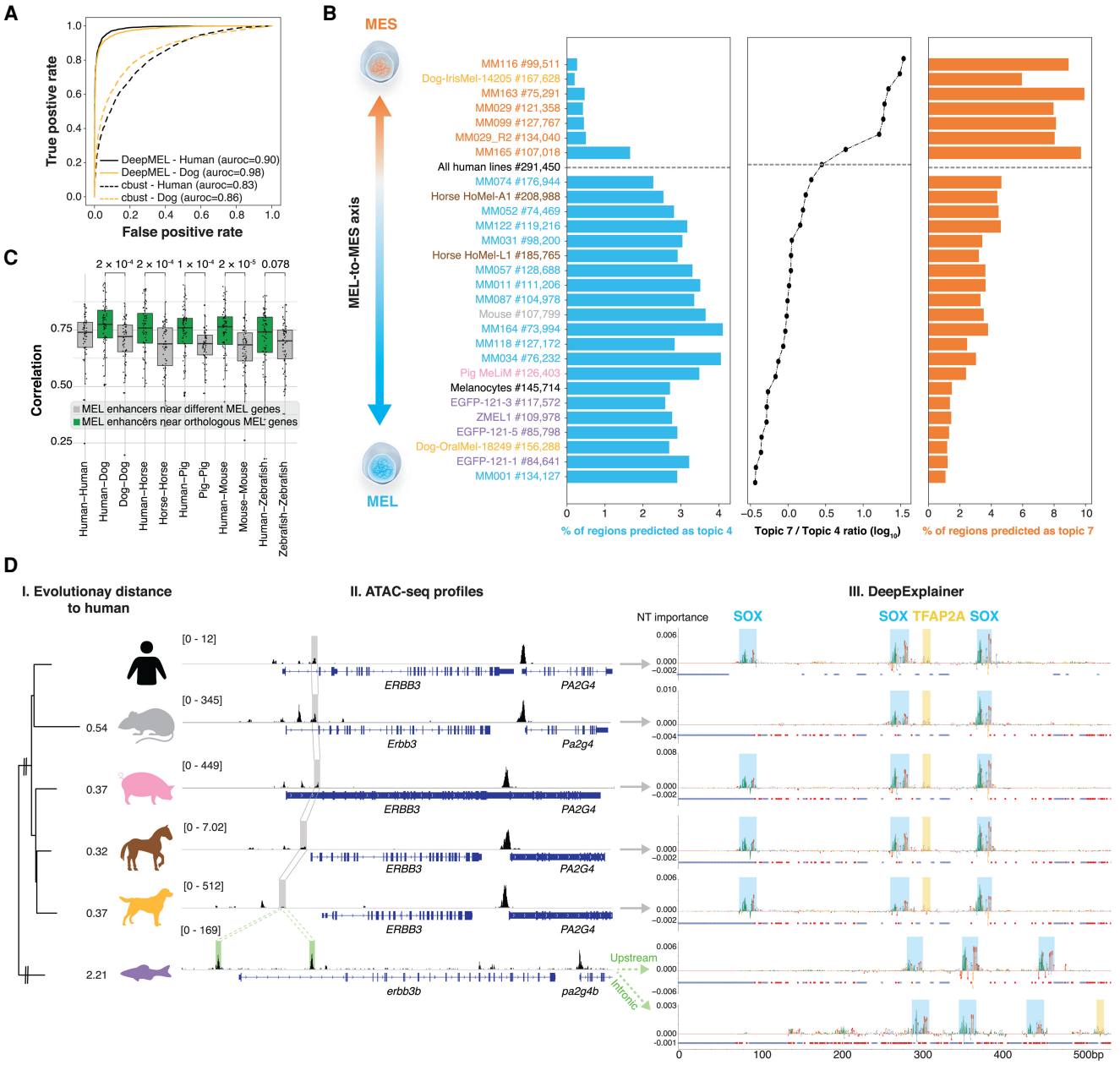

Figure 4. Human-trained deep learning model applied to cross-species ATAC-seq data. (*A*) Performance of DeepMEL and Cluster-Buster (cbust) in classifying MEL and MES differential peaks in human and dog. (*B*) Percentage of MEL- and MES-predicted ATAC-seq regions across all samples in our cohort and in human melanocytes. Samples are ordered according to the ratio of the number of MES/MEL-predicted regions. (*C*) Pearson's correlation of deep layer scores between MEL-predicted regions near orthologous MEL genes between human and another species (Human-Species) or between MEL-predicted regions near different MEL genes within one species (Species-Species). *P*-values of unpaired two-sample Wilcoxon tests are reported. (*D*) (I) Evolutionary distance between human and other species in branch length units. (II) ATAC-seq profiles of the *ERBB3* locus in the six species. MEL-specific enhancers that were predicted by DeepMEL and that were also found (gray) or not found (green) via liftOver of the human MEL enhancer are highlighted. (III) DeepExplainer plots for the multiple-aligned MEL-predicted *ERBB3* enhancers. Red and blue dots represent point and indel mutations, respectively.

each with a different motif composition (Fig. 5C). As validation of the clusters and the predicted TF binding sites, we used human ChIP-seq data of SOX10, MITF, and TFAP2A in melanoma or melanocytes (Fig. 5D; Laurette et al. 2015; Seberg et al. 2017). All clusters were indeed highly bound by SOX10, validating the prevalence of the SOX10 motif in MEL enhancers. In contrast, MITF and TFAP2A ChIP-seq data revealed that MITF and TFAP2A bind, respectively, more to enhancers with MITF and TFAP2A sites compared to regions without a predicted MITF or TFAP2A site. These observations indicate that the MEL enhancer architecture does not entail indirect DNA binding of the core regulatory factors because MITF and TFAP2A are only bound when their motifs are present within the enhancer. We further observed that regions containing a TFAP2A site, next to the SOX10 site(s) and possible others, showed a modest increase in accessibility (Supplemental Fig. S6B), which could be in line with the previously described

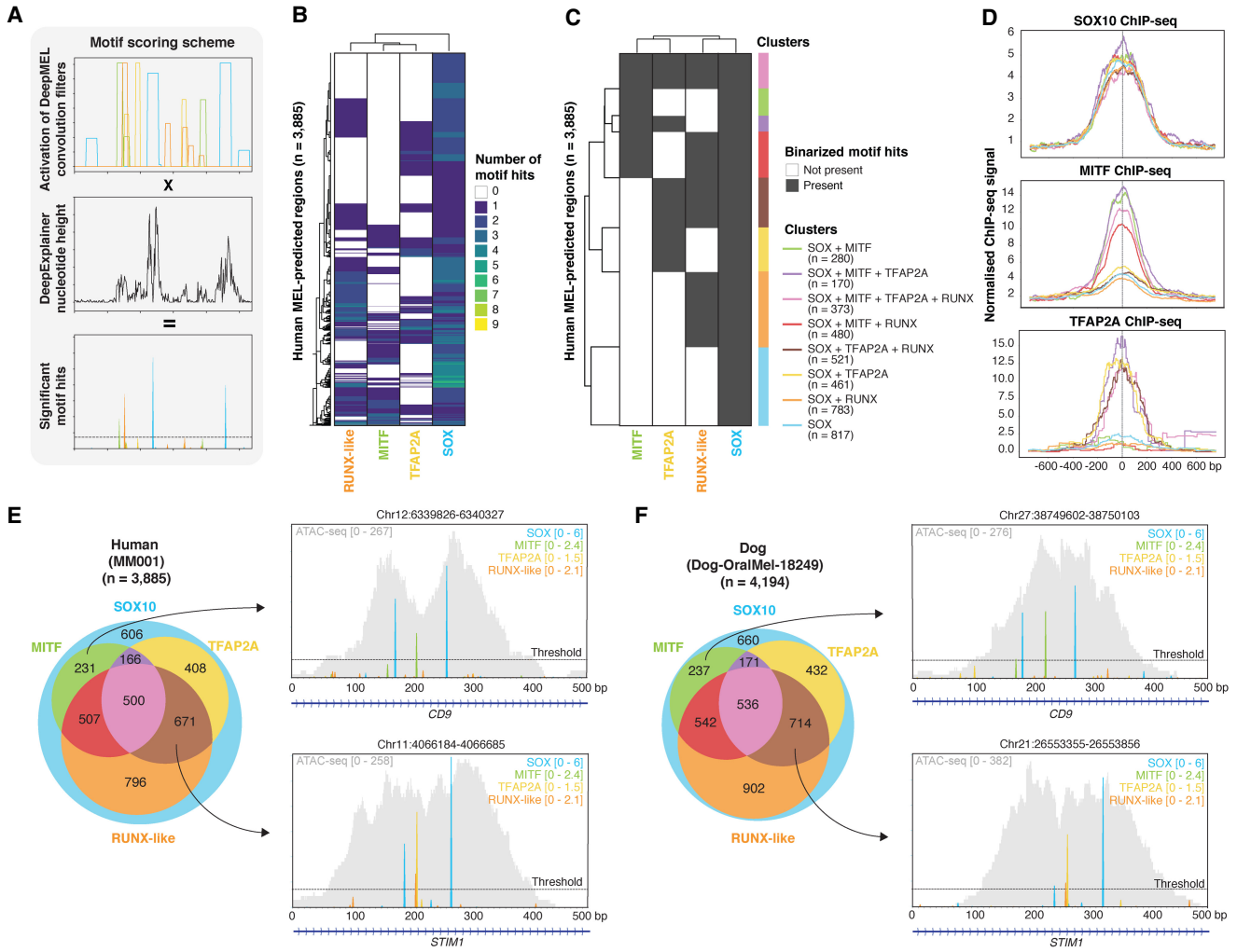

Figure 5. Core Regulatory Complex of MEL melanoma enhancers. (*A*) Schematic overview of motif scoring method in which extended convolutional filter hits from DeepMEL are multiplied by DeepExplainer profiles to yield significant motif hits. (*B*,*C*) Heatmap (*B*) and binarized heatmap (*C*) of the number of significant SOX, TFAP2A, MITF, and RUNX-like motif hits on the 3885 MEL-predicted regions in the human cell line MM001. (*D*) Aggregation plot of normalized ChIP-seq signal of SOX10, MITF, and TFAP2A on the human enhancer clusters. (*E*,*F*) Venn diagram of regions clusters on the 3885 MEL-predicted regions in human (in MM001) (*E*) and the 4194 MEL-predicted regions in dog (in Dog-OralMel-18249) (*F*). Example MEL-predicted enhancers in human and dog are shown for two of the region clusters. The ATAC-seq signal of the regions is shown in gray.

role of TFAP2A as a stabilizer of nucleosome-depleted regions (Grossman et al. 2018). The opposite was true for regions containing RUNX-like binding sites (Supplemental Fig. S6B), suggesting a repressive role of RUNX factors. The presence of a MITF site did not seem to alter the accessibility of enhancers compared to SOX-only enhancers but did increase H3K27ac signal (Supplemental Fig. S6C), possibly indicating that MEL enhancers bound by MITF are more active.

To validate these MEL enhancer classes in other species, we applied the same motif scoring and binarization to DeepMEL-predicted MEL regions in the other species in our cohort. MEL enhancers in other species also clustered into the same eight clusters, with a similar distribution of regions per cluster (Fig. 5E, F; Supplemental Fig. S6D). In addition, liftOver of the clusters showed that the regions of a human cluster correspond more to the same cluster in the other species (Supplemental Fig. S6E), indicating conservation of the MEL enhancer clusters across species. For instance, the dog orthologs of two human MEL enhancers belonging to either the [SOX10 + MITF] cluster (intronic enhancer of *CD9*) or to the cluster containing [SOX10 + TFAP2A + RUNX] (intronic enhancer of *STIM1*) (Fig. 5E) were part of the corresponding clusters in dog (Fig. 5F).

Altogether, these data suggest a Core Regulatory Complex (CoRC) (Arendt et al. 2016) of SOX10, TFAP2A, MITF, and RUNX factors in regulating melanoma MEL enhancers, encoded by a mixed enhancer model (Long et al. 2016), with high flexibility in the combination of binding sites for these four TFs, but with some rigidity (or hierarchy) in the code as at least one SOX10 dimer site is required.

Putative roles of SOX10 as a pioneer and TFAP2A as a stabilizer in melanoma MEL enhancers

Because previous results suggested a pioneering and stabilizer function for SOX10 and TFAP2A, respectively, we wanted to further investigate these putative roles and how they are

mechanistically affecting chromatin accessibility. First, we analyzed the location of binding sites relative to the position of the nucleosome, focusing on a human and dog MEL enhancer that contain a combination of one SOX10 and one TFAP2A site (Fig. 6A,B). We predicted the nucleosome start and middle point using a previously published model (Kaplan et al. 2009) and observed that SOX10 binding sites are situated within the borders of the nucleosome, near the nucleosome start point, whereas TFAP2A binding occurs preferentially near the center of the nucleosome (Fig. 6A,B). KD of TFAP2A halved the accessibility of this specific human region, whereas SOX10-KD completely abolished the ATAC-seq peak (Fig. 6A), indicating that SOX10 is necessary for accessibility, and that TFAP2A further increases the accessibility, which is in line with our previous observations (Supplemental Fig. S6A,B).

These example enhancers raised an interesting positional preference of SOX10 and TFAP2A. To assess whether this occurs globally, we centered human MEL enhancers on the SOX10 and TFAP2A motif hits and calculated the aggregated location of the nucleosome start and middle point (Fig. 6C–E). SOX10 shows a consistent preference for binding within the nucleosome borders, ~40 bp away from the nucleosome start point (Fig. 6D). Other pioneering factors have also been shown to bind near the borders of the nucleosome, for instance, FOX factors which bind ~60 bp from the center of the nucleosome, displacing linker histones and destabilizing the central nucleosome (Iwafuchi-Doi et al. 2016; Grossman et al. 2018). In contrast, when centering the MEL regions based on the TFAP2A motif, we did not observe a strong preference in the location of the nucleosome start point relative to the TFAP2A binding site (Fig. 6D), but in fact TFAP2A consistently binds in a wide range on and around the nucleosome middle point (Fig. 6E). Stabilizers, such as NFIB, have been reported to directly compete with the central nucleosomes to stabilize the accessible chromatin configuration (Denny et al. 2016; Grossman et al. 2018). Centering based on the SOX10 or TFAP2A motif hit revealed protection of Tn5 cutting on important nucleotides of the dimer motif (Supplemental Fig. S7A,B). We did not observe strong positional preferences of MITF and RUNX motifs relative to the nucleosome start or middle point (Supplemental Fig. S7C,D).

Altogether these data suggest that SOX10 functions as a pioneer in the CoRC of MEL enhancers, leading to their accessibility by binding to the central nucleosome, near the nucleosome start point. Conversely, TFAP2A appears to act as stabilizer of SOX-

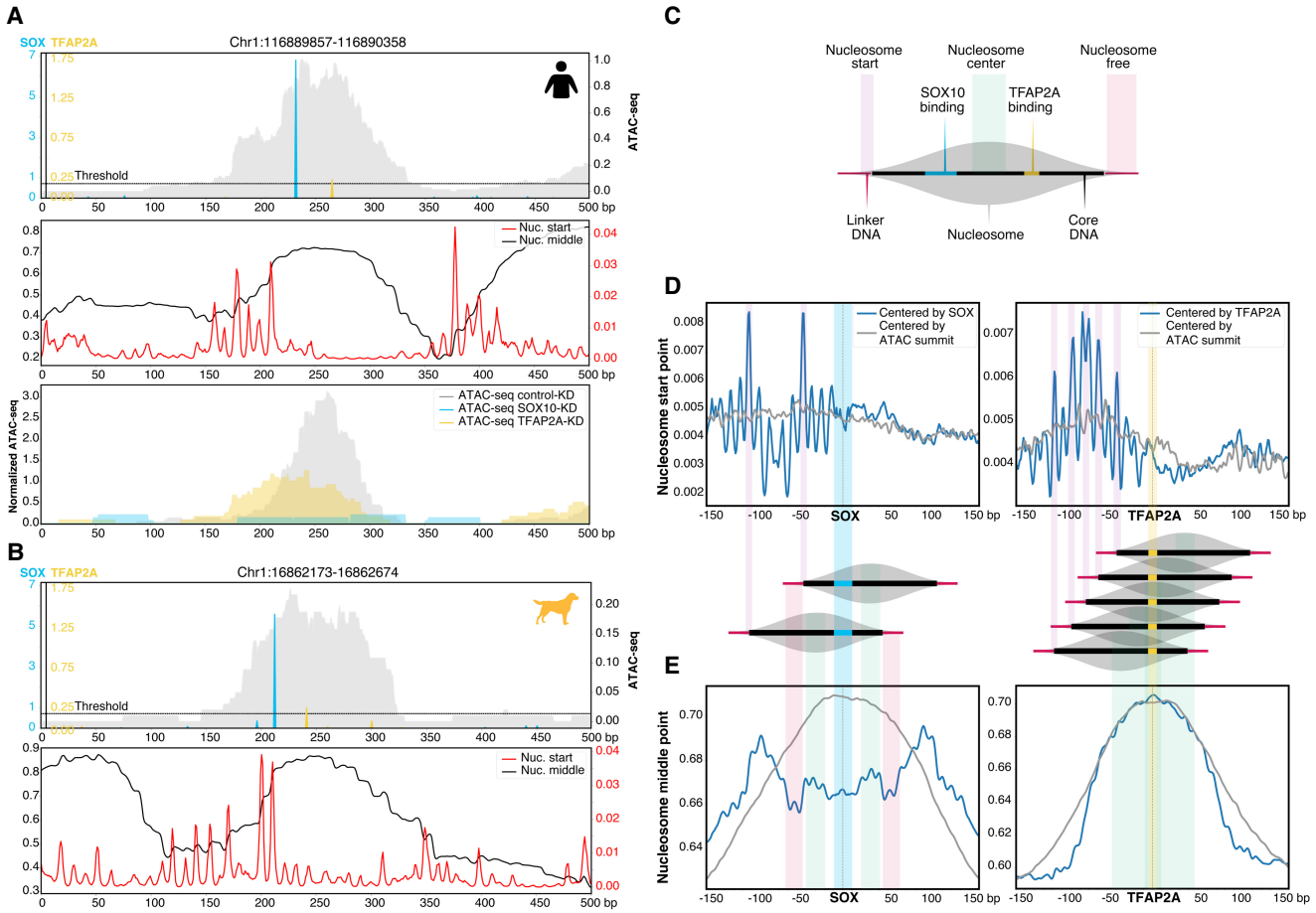

Figure 6. Positional specificity of SOX10 and TFAP2A in MEL melanoma enhancers. (*A,B, top*) Example human (*A*) and dog (*B*) MEL-predicted enhancer containing significant SOX10 and TFAP2A motifs. The ATAC-seq signal is shown in gray. (*A, middle*; *B, bottom*) Imputed nucleosome start and middle point profiles. (*A, bottom*) For the human example region, ATAC-seq profiles of MM001 in control condition, after 72 h of SOX10 knockdown or TFAP2A knockdown are shown. (*C*) Schematic overview of the nucleosome structure explaining the colors used in *D* and *E*. (*D,E*) Nucleosome start point (*D*) and nucleosome middle point predictions (*E*) on MEL-predicted regions containing one SOX10 (*left*) or one TFAP2A motif (*right*) next to possible other motifs, where the regions are either centered on the ATAC-seq summit (gray) or on the SOX10 or TFAP2A motif (blue).

dependent nucleosome-depleted regions by binding around the nucleosome middle point, possibly going in competition with the central nucleosome.

DeepMEL predicts evolutionary changes in MEL enhancer accessibility and activity

To further validate our findings on the MEL enhancer logic, we compared motif architectures between species and investigated how turnover of TF binding sites affects enhancer accessibility and function. To this end, we compared pairs of highly probable orthologous MEL enhancers that are only accessible in one of the species (Methods; Supplemental Fig. S8A). For example, an enhancer upstream of *APPL2* is predicted as a MEL enhancer in the dog line Dog-OralMel-18249 (topic 4 DL score of 0.35), whereas the orthologous enhancer in human is not accessible (Fig. 7A). Not only the accessibility of the human homolog was lost, but also its activity, as we confirmed by a luciferase assay (Fig. 7B). The topic 4 DeepMEL score for this enhancer was six times lower in human compared to dog (0.06 in human versus 0.35 in dog) (Fig. 7C), falling below the topic 4 significance threshold of 0.16, indicating that the model detected critical changes in the human enhancer sequence that could explain the loss of accessibility and activity of this MEL enhancer. The functional dog enhancer contains a SOX10, MITF, and TFAP2A binding site, which are all affected by substitutions in the nonfunctional human homologous sequence and might therefore be causal for the loss in accessibility (and activity) (Fig. 7D,E). The SOX10 motif mutation had the strongest effect, as it caused a 45% drop in the MEL-prediction score (Fig. 7D).

Next, we performed this analysis on a larger scale. First, per species pair, we observed that differences in DeepMEL predictions between species (delta-DeepMEL score) are highly predictive for differences in accessibility (Spearman's correlation of 0.43) (Supplemental Fig. S8B,C). Among the four studied regulators, mostly the disruption or gain of one or more SOX10 binding sites

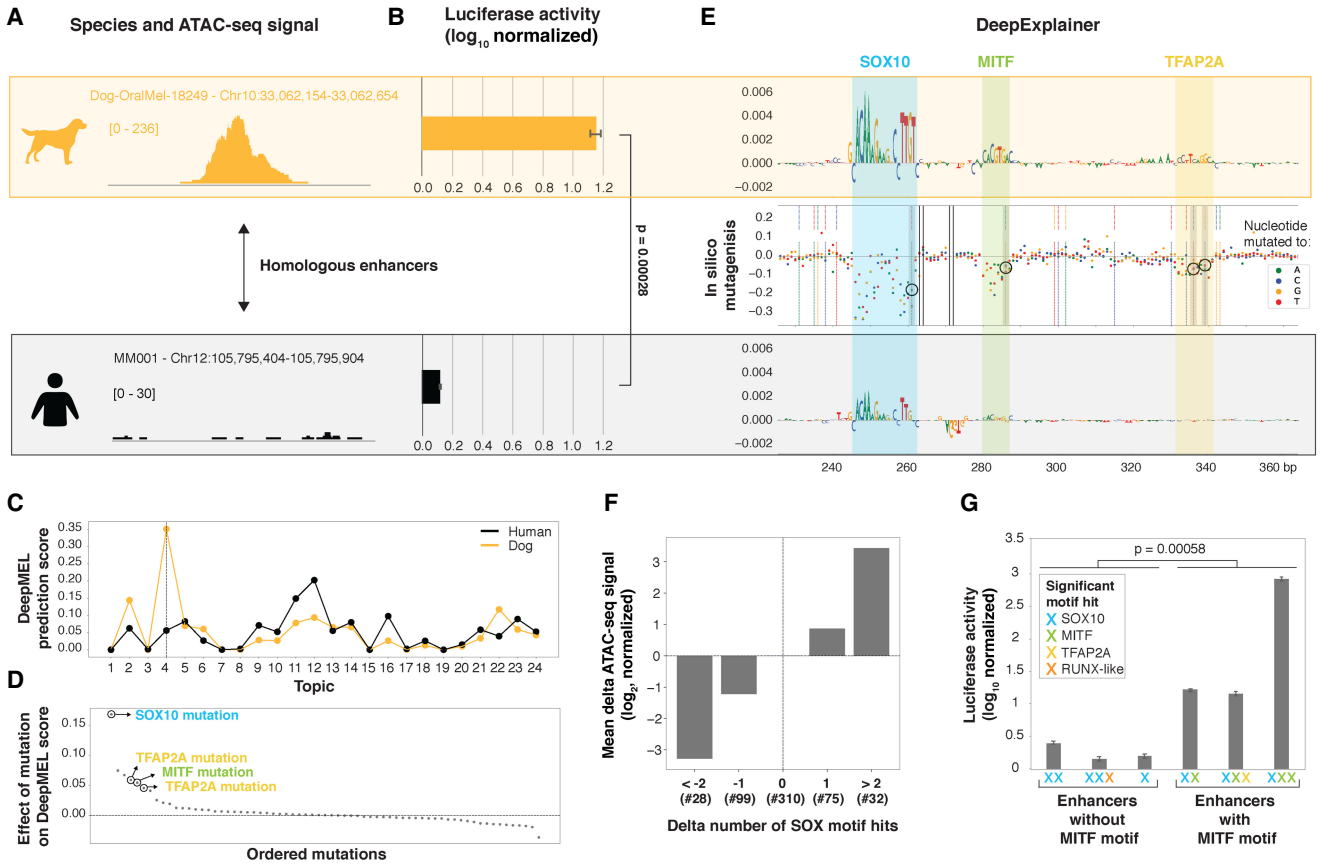

Figure 7. Predicting causal mutations of evolutionary changes in MEL enhancers. (*A*,*B*) Example region upstream of *APPL2* that is accessible (*A*) and active (*B*) in the MEL dog line Dog-OralMel-18249 but not in human MEL lines. (*C*) DeepMEL prediction score of each of the 24 topics for the dog and human *APPL2* enhancer. (*D*) Effect on topic 4 DeepMEL score on the dog sequence when in silico simulating each of the single detected point mutations between the dog and human *APPL2* enhancer. (*E*) DeepExplainer plots of the middle 120 bp of the dog and human *APPL2* enhancer. In the *middle*, the effect of each possible point mutation between the dog and human sequence on the MEL DeepMEL score was in silico calculated and is represented by colored dots depending on the nucleotide to which the original dog nucleotide was in silico mutated. Truly existing point mutations between the dog and human sequence are highlighted by color-coded vertical dashed lines. Four mutations that decrease the motif score of the SOX10, MITF, and TFAP2A motifs are highlighted by a gray box and are encircled. (*F*) Bar plot showing the mean effect on the $\log_2$ delta ATAC-seq signal of a non-human region compared to the human homolog depending on the number of SOX10 motif hits lost or gained. Only regions having no change in the number of significant TFAP2A, MITF, and RUNX motifs hits were used. The *y*-axis is normalized to the category with no changes in the number of significant SOX10 motif hits. The number of regions in each of the categories is mentioned (#). (*G*) Luciferase assay on six human or dog enhancers. Significant motif hits per enhancer are shown with colored crosses. For the luciferase assays: luciferase activity in MM001 is shown relative to *Renilla* signal and is $\log_{10}$ transformed. *P*-values were determined using Student's *t*-test, and the error bars represent the standard deviation over three biological replicates.

between orthologous enhancers quantitatively altered the ATAC-seq signal in a concordant way (Fig. 7F; Supplemental Fig. S8D), indicating that SOX10 mutations are most causal for changes in MEL enhancer accessibility, and possibly also in enhancer activity, as was the case in the *APPL2* enhancer above. However, concordance between accessibility and activity was not always observed (Supplemental Fig. S9). Furthermore, luciferase assays of six human or dog MEL-predicted enhancers suggested that enhancers with at least one MITF motif ($n = 3$) are significantly more active compared to enhancers without any MITF motif ($n = 3$) (Fig. 7G). Although the number of tested enhancers is small, this trend, together with the fact that MEL enhancers containing a MITF binding site showed increased H3K27ac signal (Supplemental Fig. S6C), indicates that MITF could function as an activator in MEL enhancers. Indeed, MITF has been shown to activate genes involved in pigmentation by recruitment of cofactors and chromatin remodeling complexes (Kawakami and Fisher 2017) and was previously classified as a TF involved in cofactor recruitment and activation (Grossman et al. 2018). SOX10 binding is insufficient but appears necessary for enhancer activity, because mutations in SOX10 binding sites disrupt enhancer activity in the *IRF4* case study (Fig. 3G).

In conclusion, DeepMEL provides a suitable platform to study the effect of evolutionary mutations on MEL enhancer accessibility and, in some cases, activity across species. Together, these results validate that SOX10 is crucial for enhancer accessibility in MEL enhancers, and necessary but insufficient for MEL enhancer activity, because activity appears to be mainly dependent on MITF binding.

Discussion

Here, we present an in-depth study of melanoma enhancer logic, especially in enhancers specific to the melanocytic (MEL) state, by exploiting both cross-species data and machine learning. Although the MEL and MES melanoma cell states have been studied extensively on a transcriptomic and epigenomic level, the combinatorial code of binding sites of their regulatory factors in state-specific enhancers had not yet been explored. Understanding the enhancer logic and the mechanism by which TFs bind and direct active enhancers will become increasingly important, because it will be essential for the development of new therapies that influence cell state-specific enhancer functions in a targeted way (e.g., for enhancer therapy) (Johnson et al. 2008; Hamdan and Johnsen 2019), or to prioritize noncoding variants in whole-genome sequencing studies of personal or cancer genomes (Atak et al. 2019).

Predicting enhancers and determining their functional role within gene regulatory networks has been an active field for years. Despite the well-established power of cross-species approaches in this field, to our knowledge, a large comparative epigenomics study in melanoma has not yet been conducted, although several non-human models are commonly used in melanoma research (van der Weyden et al. 2016) and have been studied on an intraspecies level (Rambow et al. 2008; Rosengren Pielberg et al. 2008; Sundström et al. 2012; Jiang et al. 2014; Seltenhammer et al. 2014; Kaufman et al. 2016; Hitte et al. 2019) or in relation to human melanoma (Egidy et al. 2008; Segaoula et al. 2018; Rahman et al. 2019). Here, we show that the MEL and MES states are conserved across species, as well as the key regulators of these states.

Despite their proven advantages, sequence-based comparative approaches have limited power to identify orthologous regulatory regions in distant species, in part because of the rapid evolution of distal enhancers (Dermitzakis and Clark 2002; Lindblad-Toh et al. 2011). Methods, such as enhancer element locator (EEL), try to tackle this question by aligning TF binding sites to identify conserved enhancer elements (Hallikas et al. 2006) or by calculating the co-occurrence of sequence patterns (Arunachalam et al. 2010). However, these methods are either supervised because they require user-provided PWMs (Hallikas et al. 2006), or it is difficult to extract the important biologically relevant features from these methods (Arunachalam et al. 2010). In addition, the identification and exact localization of important (de novo) TF binding sites within enhancers is complex because motif discovery tools are often dependent on user-provided databases and motif-specific thresholds. Recently, deep learning approaches, which are commonly used in disciplines such as speech recognition and image analysis, found their way into the regulatory genomics field to overcome these concerns (Park and Kellis 2015). Deep learning models, such as DeepBind, are particularly powerful in learning complex patterns by leveraging large epigenomics data sets; therefore, they are well suited to function as de novo motif detectors, as well as to uncover more complex sequence features (Alipanahi et al. 2015; Park and Kellis 2015). By designing DeepMEL, a multiclass, multilabel neural network trained on melanoma human regulatory topics of coaccessible regions, and by using the model interpretation tool DeepExplainer and our newly developed motif scoring scheme (Lundberg and Lee 2017; Lundberg et al. 2020), we were able to perform a thorough and unsupervised analysis of important TF binding sites in melanoma enhancers. Specifically, in MEL enhancers, our data suggest conserved cobinding of a CoRC of three main TFs, consisting of SOX10, TFAP2A, and MITF. DeepMEL also finds motifs for RUNX factors, but their role in the melanocyte or melanoma is less clear. Evidence for cobinding of SOX10, MITF, and TFAP2A was previously observed by enrichment of both MITF and TFAP2A motifs in SOX10 ChIP-seq data in melanoma cells (Laurette et al. 2015). We observed high flexibility in the organization of TF binding sites of the CoRC because eight different modalities were found, formed by all permutations of the CoRC factors, with the exception that all MEL enhancers contained at least one SOX10 binding site. MEL enhancers thereby adhere to a "mixed modes enhancer" model, a billboard-like model with mostly high flexibility in the TF motif organization, except for the ever-present SOX10 binding sites (Long et al. 2016). In addition, ChIP-seq data of MITF and TFAP2A indicated no indirect DNA binding of these CoRC factors within MEL enhancers, but that the bound TFs are largely determined by their individual motif presence. Although DeepMEL was trained on melanoma ATAC-seq data, the human- and pig-predicted MEL enhancers were also accessible in human and pig melanocytes, respectively, indicating that we could extend these observations on the MEL enhancer logic to enhancers in melanocytes, and that our methodology could be applied to nondisease states.

It is well established that distinct functional classes of TFs exist, with respect to enhancer binding. Pioneer TFs, such as POU5F1, SOX2, Grh-like TFs, and FOXA1, are able to bind nucleosomal DNA, leading to displacement of the nucleosome and facilitating the binding of other TFs to the accessible enhancer (Zaret and Carroll 2011; Long et al. 2016; Jacobs et al. 2018). SOX2 and other SOX factors have a HMG domain that interacts with the minor groove of the DNA, causing the DNA to bend in a 60°–70° angle, a property that has been suggested to contribute to the pioneering activity of SOX2, and possibly of other SOXs (Hou

et al. 2017). Dodonova et al. (2020) indicate that SOX2 and SOX11 can bind to their binding motif on nucleosomal DNA and that they use their binding energy to initiate chromatin opening. However, there is still some dispute on the pioneering properties of SOX TFs, as another study classified SOXs as "migrant TFs," that is, nonpioneering TFs that only bind sporadically to (non)-chromatinized DNA (Sherwood et al. 2014). Nonetheless, we find strong evidence for a pioneering function of SOX10 in MEL melanoma cells. Our current and previous study (Bravo González-Blas et al. 2019) have shown that knockdown of SOX10 induces closure of SOX10-bound ATAC-seq peaks containing a SOX10 motif. In fact, DeepMEL predicts SOX10 binding sites as essential for MEL enhancer accessibility. Next to pioneer factors, other functional classes of TFs exist, including factors that stabilize the accessibility of the nucleosome-depleted regions. TFAP2A was previously classified as such a chromatin stabilizer (Grossman et al. 2018), and it has been shown that evolutionary divergence from the TFAP2A consensus motif correlates with loss of chromatin accessibility and H3K27ac ChIP-seq signal (Prescott et al. 2015). These reports support our observations of TFAP2A as a stabilizer of SOX10-dependent accessible MEL enhancers, likely caused by direct competition of TFAP2A with the nucleosome, because TFAP2A binding sites were highly enriched at the predicted center of the central nucleosome. The dependence of SOX10 for opening MEL enhancers before TFAP2A binding is in line with the reported classification of TFAP2A as a "settler," a TF whose binding depends predominantly on the accessibility of the chromatin at their binding sites (Sherwood et al. 2014).

Besides classifying accessible (orthologous) regions and predicting important TF motifs within them, DeepMEL is an accurate predictor of the effect of mutations on enhancer accessibility and, for some enhancers, also the activity. This was for instance the case for the *IRF4* MEL enhancer, where DeepMEL outperformed existing methods tested in Kircher et al. (2019). However, the other models in the benchmark were trained to predict the activity of a total of 20 regulatory regions ranging across different cell types, whereas our DL model is specialized for melanoma regulatory regions. This shows the value of using case-specific training data, such as the data set generated in this study for melanoma. Not all predicted MEL enhancers were in fact active, as MITF binding seems to be required to activate SOX10-dependent melanoma enhancers. Fufa et al. (2015) support this hypothesis, because activating SOX10-regions in mouse melanocytes showed significant enrichment of E-box motifs (bound by the bHLH protein family, which includes MITF), indicating that MITF cooperates with SOX10 to execute melanocyte-specific gene activation. In addition, MITF was previously classified as a TF involved in cofactor recruitment and activation (Kawakami and Fisher 2017; Grossman et al. 2018). Although SOX10 binding is not sufficient for enhancer activity, it appears to be necessary, because disruption of the SOX10 binding site in the *IRF4* enhancer had a strong effect on activity, probably owing to the reappearance of the central nucleosome.

In conclusion, the combination of comparative epigenomics with deep learning allowed us to perform an in-depth analysis of the melanoma enhancer logic. This work presents an overall framework that can be applied to decipher the enhancer logic in a cell type or cell state of interest, starting from the generation of an extensive cell type–specific (cross-species) epigenomics data set, all the way through the training and exploitation of a deep neural network to decode enhancer features across species, and to utilize it to assess the impact of *cis*-regulatory variation.

Methods

Cell culture

Human melanoma cell lines

Human melanoma cultures (MM lines) are short-term cultures derived from patient biopsies (Gembarska et al. 2012; Verfaillie et al. 2015). Cells were cultured at 37°C with 5% CO_2 and were maintained in Ham's F10 nutrient mix (Thermo Fisher Scientific) supplemented with 10% fetal bovine serum (FBS; Thermo Fisher Scientific) and 100 µg mL^{-1} penicillin/streptomycin (Thermo Fisher Scientific).

Zebrafish melanoma cell lines

Experiments were performed as previously outlined (Ceol et al. 2011). Briefly, 25 pg of MCR:EGFP were microinjected together with 25 pg of Tol2 transposase mRNA into one-cell Tg (BRAFV600E); $tp53^{-/-}$; $mitf^{-/-}$ zebrafish embryos. Embryos were scored for melanocyte rescue at 48–72 h post-fertilization, and equal numbers were raised to adulthood (15–20 zebrafish per tank) and scored weekly (from 8 to 12 wk post-fertilization) or biweekly (>12 wk post-fertilization) for the emergence of raised melanoma lesions (van Rooijen et al. 2017). For in vitro culture, large tumors were isolated from MCR/MCR:EGFP (14–28 wk post-fertilization). Zebrafish were maintained under IACUC-approved conditions. Zebrafish primary melanoma ZMEL1 cell line was previously described (White et al. 2008, 2011), and EGFP 121-1, EGFP 121-2, EGFP 121-3, and EGFP 121-5, were generated as described (Heilmann et al. 2015; Wojciechowska et al. 2016). All cell lines were cultured in DMEM medium (Thermo Fisher Scientific) supplemented with 10% heat-inactivated FBS (Atlanta Biologicals), 1× GlutaMAX (Thermo Fisher Scientific), and 1% penicillin/streptomycin (Thermo Fisher Scientific), at 28°C, 5% CO_2. Zebrafish melanoma lines were authenticated by qPCR and Western for EGFP transgene expression, and periodically checked for mycoplasma using the Universal Mycoplasma Detection Kit (ATCC).

Horse melanoma cell lines

The horse cell lines HoMel-L1 and HoMel-A1 are melanoma cell lines derived from a Lipizzaner stallion and Shagya-Arabian mare, respectively, and were established in Seltenhammer et al. (2014). Cells were cultured at 37°C with 5% CO_2 in Roswell Park Memorial Institute (RPMI) medium (Thermo Fisher Scientific) supplemented with 10% FBS (Thermo Fisher Scientific) and 1% penicillin/streptomycin (Thermo Fisher Scientific).

Pig melanoma and melanocyte cell line

The immortal line of pigmented melanocytes (PigMel) was previously derived (Julé et al. 2003), and the 30-d-old piglet primary melanoma cells (MeLiM) were isolated as described (Egidy et al. 2008). PigMel cells were cultured at 37°C with 10% CO_2 in MEM medium supplemented with 1× MEM nonessential amino acids (Thermo Fisher Scientific), 1 mM Na pyruvate, 2 mM glutamine, 100 units/mL penicilin/streptomycin (Thermo Fisher Scientific), 10% FCS and 3.7 g/mL Na bicarbonate. MeLiM cells were cultured in DMEM high glucose (Thermo Fisher Scientific), 10% FCS, Pen/Strep, and 5% CO_2.

Dog melanoma cell lines

The dog cell lines Dog-IrisMel-14205 and Dog-OralMel-18249 were established by Aline Primot, and were derived from an uveal

melanoma from a beagle crossed dog and an oral melanoma from the palate of a Shih Tzu, respectively. Cells were cultured at 37°C with 5% CO_2 in Ham's F-12 Nutrient Mixture medium (Thermo Fisher Scientific) supplemented with 10% FBS (Thermo Fisher Scientific) and 1% penicillin/streptomycin (Thermo Fisher Scientific).

Mouse melanoma cell lines

The mouse melanoma cell line was generated as described (Dankort et al. 2009). Cells were cultured at 37°C with 5% CO_2 in Dulbecco's Modified Eagle Medium (DMEM) (Thermo Fisher Scientific) supplemented with 10% FBS (Thermo Fisher Scientific) and 1% penicillin/streptomycin (Thermo Fisher Scientific).

Knockdown experiments

SOX10, TFAP2A, and the control knockdown (KD) were performed in MM001 using a SMARTpool of four siRNAs against, respectively, *SOX10* (SMARTpool: ON-TARGETplus SOX10 siRNA, number L017192-00-0005, Dharmacon), *TFAP2A* (SMARTpool: ON-TARGETplus TFAP2A siRNA, number L-006348-02-0005, Dharmacon), and a negative control pool (ON-TARGETplus non-targeting pool, number D-001810-10-05, Dharmacon) at a concentration of 20 nM for SOX10-KD, and 40 nM for TFAP2A-KD and the control using as medium Opti-MEM (Thermo Fisher Scientific) and omitting antibiotics. The cells were incubated for 72 h before processing.

OmniATAC-seq data generation, data processing, and follow-up analyses

OmniATAC-seq on mammalian lines

Omni-assay for transposase-accessible chromatin using sequencing (OmniATAC-seq) was performed as described previously (Corces et al. 2017). After the final amplification was done with the additional number of cycles, samples were cleaned-up by MinElute and libraries were prepped using the KAPA Library Quantification Kit as previously described (Corces et al. 2017). Samples were sequenced on a HiSeq 4000 or NextSeq 500 High Output chip.

ATAC-seq on zebrafish lines

Fifty thousand cells per line were lysed and subjected to a tagmentation reaction and library construction as described in Buenrostro et al. (2013). Libraries were run on an Illumina HiSeq 2000.

Data processing of ATAC-seq and OmniATAC-seq samples

Paired-end or single-end reads were mapped to the human genome (hg19-GENCODE v18) using Bowtie 2 (v2.2.6) (Langmead and Salzberg 2012) or STAR (v2.5.1b) (Dobin et al. 2013) to species-specific genomes, which were downloaded from UCSC (https://hgdownload.soe.ucsc.edu/downloads.html) (for human: hg19-GENCODE v18; for dog: canFam3; for horse: equCab2; for pig: susScr11; for mouse: mm10; for zebrafish: danRer10) and by applying the parameters --alignIntronMax 1 and --alignIntronMin 2. For the human data, we used hg19 as genome assembly instead of the more recent GRCh38 assembly because i-cisTarget (Herrmann et al. 2012; Janky et al. 2014; Imrichová et al. 2015) and GREAT (McLean et al. 2010) are or were not (yet) available for GRCh38 at the time of the analyses. However, the use of GRCh38 instead of hg19 would not significantly affect conclusions. We, for instance, validated this by rescoring MEL-predicted regions by DeepMEL in MM057 after liftOver (Kuhn et al. 2013) from hg19 to GRCh38, in which we observed that changing genome assembly yields the same DeepMEL score for all 4244 regions except for eight of them. Also note that for MM029, two biological replicates were used. Mapped reads were sorted using SAMtools (v1.8) (Li et al. 2009), and duplicates were removed using Picard MarkDuplicates (v1.134). Reads were filtered by removing mitochondrial reads and filtering for $Q > 30$ using SAMtools. BAM files of technical replicates of the same cell line were merged at this point using SAMtools merge. Peaks were called using MACS2 (v2.1.2) (Gaspar 2018) callpeak using the parameters -q 0.05, --nomodel, --call-summits, --shift -75 --keep-dup all and --extsize 150 per sample. Blacklisted regions (ENCODE) and peaks overlapping with alternative chromosomes and ChrM were removed. Summits were extended by 250 bp up- and downstream using slopBed (BEDTools; v2.28.0) (Quinlan and Hall 2010), providing human chromosome sizes. Peaks were normalized for the library size using a custom script, and overlapping peaks were filtered using the peak score by keeping the peak with the highest score. Normalized bigWigs were either made from normalized bedGraphs using as scaling parameter (-scale) 1×10^6/(number of nonmitochondrial mapping reads); or made by bamCoverage (deepTools, v3.3.1) (Ramírez et al. 2016), using as parameters --normalizeUsing None, -bl EncodeBlackListedRegions --effectiveGenomeSize 2913022398 and as scaling parameter (-scaleFactor) $1/(\text{RIP}/1 \times 10^6)$, in which RIP stands for the number of reads in peaks.

HOMER on human and dog differential accessible peaks

Count matrices were produced by featureCounts (v1.6.5) (Liao et al. 2014) for five melanocytic (MEL) and five mesenchymal-like (MES) lines for human, and for Dog-OralMel-18249 and Dog-IrisMel-14205 for dog. Differential peaks were identified using DESeq2 (v1.22.2, R v3.5.2) (R Core Team 2018; Love et al. 2014) with a $\log_2$FC higher than 2.5 and a P_{Adj} lower than 0.0005. HOMER (Heinz et al. 2010) was performed on the differentially accessible regions using findMotifsGenome.pl, providing the differential regions as a BED file and a FASTA file of the human or dog genome, with parameters -mask, -size given, and -len 6,8,10,11,12,17,18.

Defining sets of alignable and conserved accessible ATAC-seq regions

ATAC-seq regions of non-human species were defined as alignable regions when they could be converted to hg19 coordinates using liftOver (Kent-tools, -minMatch = 0.1) (Kuhn et al. 2013) by providing the appropriate liftOver chain (UCSC). Alignable regions were intersected with accessible peaks in human using intersectBed (BEDTools, v2.28.0) (Quinlan and Hall 2010) with -f 0.6 to define sets of conserved accessible regions across species.

Clustering of species based on globally alignable ATAC-seq regions

Per species, a count matrix was made on the alignable union ATAC-seq regions by featureCounts (v1.6.5) (Liao et al. 2014). The count matrices of different species were merged and the final count matrix was CPM normalized (edgeR v3.22.5, R v3.5.2) (Robinson et al. 2010; R Core Team 2018), followed by quantile normalization. A principal component analysis (PCA) on the normalized count matrix was performed using irlba (v2.3.3, R v3.5.2) (Baglama and Reichel 2005).

Branch length scoring across species

Conserved accessible ATAC-seq regions were identified as described above, and for each of the species, the set of conserved

accessible regions was converted to the coordinate system per species and FASTA sequences were retrieved. All sequences were scored with the cisTarget motif collection (v8) (http://iregulon.aertslab.org/collections.html) (Herrmann et al. 2012; Janky et al. 2014; Imrichová et al. 2015) containing 20,003 TF position-weight matrices (PWMs) using Cluster-Buster (Frith et al. 2003) with parameters -m 0, -c 0, and -r 10000. For each motif, the highest *cis*-regulatory module (CRM) score per conserved accessible sequence was used to calculate the branch length score (BLS) across species according to Stark et al. (2007) and Jacobs et al. (2018). The branch length was taken from the phylogenetic data from http://hgdownload.cse.ucsc.edu/goldenpath/hg19/phyloP100way/ (UCSC). The sum of the BLSs for all the conserved accessible sequences across the mammalian or all six species was used as a total score for each motif. We normalized these scores by performing BLS on a shuffled variant of all sequences by shuffleseq (EMBOSS, v6.6.0.0), keeping the same base-pair compositions and sequence lengths, and subtracting the shuffled BLS from the true BLS per motif.

CisTopic analysis to obtain sets of coaccessible regions in human OmniATAC-seq data

To apply cisTopic (Bravo González-Blas et al. 2019), a tool designed for single-cell ATAC-seq analysis, we first simulated single cells from the bulk OmniATAC-seq data of the 16 human melanoma lines via bootstrapping. Per cell line, 50 simulated single-cell BAM files were generated containing each 50,000 random reads that were bootstrapped from the bulk BAM files. These simulated single-cell BAM files were provided as input for cisTopic (v0.2.0, R v3.4.1) (R Core Team 2017), together with the merged BED file of ATAC-seq regions across all 16 samples, after removing blacklisted regions (ENCODE). We ran cisTopic (parameters: $\alpha = 50/T$, $\beta = 0.1$, burn-in iterations = 500, recording iterations = 1000) for models with a number of topics (sets of coaccessible regions) between 2 and 30 (2 by 2). The best model, containing 24 topics, was selected on the basis of the highest log-likelihood. Topics were binarized using a probability threshold of 0.995 (resulting in a total of 35,940 binarized topic regions across the 24 topics), and we performed motif enrichment analysis with cisTarget (Imrichová et al. 2015).

Deep learning

Data preparation

The deep learning (DL) model, DeepMEL, was trained on the binarized regions of the 24 topics obtained from the cisTopic analysis explained above. To increase the amount of training data, the 500-bp regions in the merged BED file of all 339,099 ATAC-seq regions across the 16 human cell lines (see "Data processing of ATAC-seq and OmniATAC-seq samples"), were augmented by extending them to 700 bp around the summit and sliding a 500-bp window over these elongated regions with a 10-bp stride. This augmented master region BED file was intersected with each topic BED file separately (using BEDTools) (Quinlan and Hall 2010), and a region was labeled with a topic number if there was at least 60% overlap. If regions overlapped with multiple topics, they were assigned with multiple topic labels, allowing for a multilabel and multiclass DL model. This augmentation and intersection resulted in 696,654 training regions in total, excluding the 58,086 regions on Chr 2 that were used for testing.

DeepMEL model architecture and training parameters

The DeepMEL architecture was built with four layers between input and output layer: a Conv1D layer (containing 128 filters and setting the parameters kernel_size as 20, the strides as 1 and the activation as relu), MaxPooling1D layer (with the pool_size 10 and strides 10), TimeDistributed Dense layer together with Bidirectional LSTM layer (with 128 unit and setting the dropout as 0.1 and the recurrent_dropout as 0.1), and Dense layer (with 256 units and setting the activation as relu). After MaxPooling1D, Bidirectional LSTM, and Dense layer, a Dropout layer was used each time with the fraction of dropout set as 0.2, 0.2, and 0.4, respectively. For each region in the training data, DeepMEL takes the one-hot encoded (500 bp × 4 nt) forward and reverse strand and passes them separately through the model. To make the final prediction, DeepMEL takes the average activation (*average* function) of the neurons in the final Dense layer (which contains 24 units corresponding to the 24 topics; with a sigmoid activation function). The model was compiled using the Adam optimizer with the default learning rate, which is 0.001. To calculate the loss, the binary cross entropy (binary_crossentropy) was used. The model was trained for two epochs with a batch size of 128, which took 67 min. Keras 2.2.4 (https://keras.io) with tensorflow 1.14.0 (Abadi et al. 2016) was used. A Tesla P100-SXM2-16GB GPU was used for training on VSC servers (Flemish Supercomputer Center).

Performance evaluation

The performance of the model was evaluated for each topic separately because it was a multilabel classifier. The auROC and auPR were calculated for the combined training and validation data (regions on all chromosomes except Chr 2), test (regions on Chr 2), and label-shuffled regions.

Converting convolution filters to PWMs, filter-topic assignment, and filter annotation

Filters of the convolution layer were converted to position-weight matrices (PWMs) by the following strategy: (1) 4,000,000 unique 20-bp-long (size of the filters) sequences were randomly generated; (2) the activation score of each filter for each sequence was calculated and the top 100 sequences were selected; (3) a count matrix was generated from these 100 sequences obtained for each filter; and (4) finally, the count matrices were converted into PWMs. To assign the filters to topics, a similar strategy that is mentioned in Basset (Kelley et al. 2016) was used. After setting the activation score of a filter to its mean activation score over all the sequences, the loss/accuracy score on the prediction was calculated for each topic. Filters were ordered based on their effect on a certain topic. To annotate the filters to known transcription factor binding motifs, the Tomtom motif annotation tool (Gupta et al. 2007) was used together with our curated cisTarget motif collection (v9) (http://iregulon.aertslab.org/collections.html) (Herrmann et al. 2012; Janky et al. 2014; Imrichová et al. 2015) of 24,453 PWMs (cutoff for the *Q*-value was set to 0.3).

DeepExplainer

From the 35,940 topic regions that were obtained after binarization of the 24 topics within the selected cisTopic model (see methods on cisTopic analysis above), 500 regions were randomly selected to initialize the DeepExplainer pipeline (Lundberg and Lee 2017). A hypothetical importance score for each position of the sequence of interest was calculated for any of the 24 topics. For each sequence, these DeepExplainer-obtained importance scores were multiplied by the one-hot encoded matrix of the sequences. Finally, the 500-bp sequences were visualized by adjusting the nucleotide heights based on their importance score by using the modified viz_sequence function from the DeepLift repository (Shrikumar et al. 2017).

In silico saturation mutagenesis

In silico saturation mutagenesis of a region was performed by separately changing each nucleotide on the 500-bp sequence into the three other nucleotides and scoring these mutated sequences with DeepMEL. The delta prediction score for each mutation was calculated for each of the 24 topics by comparing the prediction score of the mutated sequence relative to the prediction score for the initial sequence. For the *IRF4* enhancer case, the actual *IRF4* enhancer sequence used in the in vitro saturation mutagenesis assay (Chr 6: 396,143–396,593) overlapped with a predicted MEL enhancer in human MEL cell lines in our cohort (Chr 6: 396,135–396,636). The delta prediction score of topic 4 (MEL topic) was calculated following an in silico saturation mutagenesis on this region, and a Pearson's correlation was calculated on the overlapping nucleotides between the in silico and in vitro assays (451 bp).

Motif scoring method

We designed an optimized motif scoring method, in which activation scores of the filters on each sequence are multiplied by the DeepExplainer importance scores of the sequence. Then, after the output of this multiplication was normalized, a threshold was calculated for each motif by comparing MEL and MES enhancers. This approach yielded significant motif hits with their precise location.

Nucleosome positioning

Nucleosome start and middle point predictions were calculated by using the executable nucleosome prediction tool Kaplan_v3 (Kaplan et al. 2009) that takes just the DNA sequence and calculates the nucleosome positioning for each nucleotide. To get more precise results, as the authors of Kaplan_v3 suggest, enhancers were extended 3 kb from both ends. After obtaining the predictions, the middle 500-bp part of the 6.5-kb nucleosome prediction score was used.

Tn5 footprinting

Footprints of the Tn5 were determined by inferring Tn5 cut sites from the start point of each ATAC-seq read in a BAM file using a custom script.

AUROC on human and dog of DeepMEL and Cluster-Buster

The performance of DeepMEL to discriminate between MEL and MES regions in human and dog was calculated by scoring the top 5000 differential MEL and MES regions in human and dog (described above) with DeepMEL and calculating the precision of correct assignment (i.e., topic 4 score for the MEL regions and topic 7 scores for the MES regions). The performance of DeepMEL was compared with the motif scoring tool Cluster-Buster (Frith et al. 2003) by scoring the same sets of regions with Cluster-Buster using a merged motif file of (some of) the top filters identified by the model in either topic 4 or topic 7. The obtained CRM scores were used to estimate the performance of Cluster-Buster.

Identification of homologous MEL genes and MEL enhancers

To identify genes differentially expressed in human MEL cell lines, we performed DESeq2 (v1.22.2, R v3.5.2) (R Core Team 2018; Love et al. 2014) on RNA-seq data of seven MEL (MM031, MM034, MM057, MM074, MM087, MM118, MM164) and five MES (MM029, MM099, MM116, MM163, MM165) human lines. Three hundred seventy-nine genes were found differentially expressed in MEL lines ($\log_2$FC > 2.5 and $P_{Adj} < 0.005$). We converted the gene symbols to Ensembl gene IDs using biomaRt (v2.38.0, R v3.5.2) (Durinck et al. 2005) and found back the genomic locations of the genes using GenomicFeatures (v1.34.8, R v3.5.2) (Lawrence et al. 2013). For the human differential MEL genes with at least one MEL-predicted peak in their extended gene locus (200 kbp upstream and downstream), the homologous genes in the other six species were identified using biomaRt to convert the human Ensembl gene IDs to Ensembl gene IDs of the other species. We identified the MEL enhancers that overlapped with the extended gene loci of each of the homologous genes using BEDTools intersect (Quinlan and Hall 2010). liftOver (-minMatch = 0.1) (Kuhn et al. 2013) was used to calculate the number of these regions that could be identified by performing coordinate conversion.

Correlation of MEL enhancers using deep layers of DeepMEL

Conserved accessible MEL enhancers in the extended loci of conserved MEL-specific genes across the six species (see above) were scored by DeepMEL. A matrix was generated consisting of a score for each of the 256 nodes in the Dense layer for each of the regions. A Pearson's correlation matrix was generated to calculate the pairwise similarity between each of the regions.

Genome-wide prediction of MEL enhancers

The first chromosome of the human genome (hg19) was tiled with a sliding window of 500 bp and a 100-bp shift using BEDTools makewindows (v2.28.0) (Quinlan and Hall 2010). Tiles containing "N" were deleted and the remaining tiles were scored by DeepMEL, and the number of MEL-predicted tiles (topic 4 score > 0.16) was calculated.

Mutations in orthologous enhancers across species

We defined highly probable orthologous MEL enhancers between human and another species as regions that were predicted as MEL in one species and for which there was a stringent liftOver (-minMatch = 0.995) (Kuhn et al. 2013) and high sequence identity, that is, >80% after pairwise alignment via needle (EMBOSS, v6.6.0.0) (Madeira et al. 2019), using parameters -gapopen 10.0 -gapextend 0.5, in the other species. featureCounts (v1.6.5) (Liao et al. 2014) was used to generate count matrices per species on these regions, which was followed by library size normalization. Delta ATAC-seq scores were calculated for the pairs of orthologous regions by dividing the normalized counts of the two species (human counts/non-human counts) after adding a pseudocount. Mutations were identified by alignment via needle, using the parameters -gapopen 10.0 and -gapextend 0.5.

Luciferase assay

Six MEL-predicted enhancers (three in the dog line Dog-OralMel-18249 and three in the human line MM001) were synthetically generated and cloned into a pTwist ENTR plasmid (Twist Bioscience) via Twist Bioscience. Regions were transferred from the Gateway entry clone into the destination vector (pGL4.23-GW, Addgene) via a LR reaction by mixing 2 µL of the entry clone (100 ng/µL) with 1 µL of the destination plasmid (150 ng/µL), 1 µL TE buffer, and 1 µL LR enzyme (LR Clonase II Plus enzyme mix, Thermo Fisher Scientific), and incubating this mixture at 25°C. Afterwards, 1 µL of Proteinase K (Thermo Fisher Scientific) was added and reactions were incubated for 1 h at 37°C for 10 min. Then, 3 µL of each LR reaction was transformed into 50 µL of Stellar competent cells (Takara Bio) via heat shock. Next, 200 µL of SOC medium was added and the cells were incubated for 1 h in a shake incubator at 37°C, before plating the transformed cells

on LB agar plates with 1/1000 carbenicillin and incubation overnight at 37°C. The next day, one colony per construct was picked and grown overnight in 5 mL of LB medium with 1/1000 carbenicillin in a shake incubator at 37°C before plasmid extraction using the NucleoSpin Plasmid Transfection-grade kit (Macherey-Nagel). For each construct, three biological replicates were performed by transfecting the plasmids into 80% confluent cells of MM001 in a 24-well plate. Per transfection, 400 ng of the construct was transfected together with 40 ng of *Renilla* plasmid (Promega) using lipofectamine 2000 (Thermo Fisher Scientific). Luciferase activity of each construct was measured using the Dual-Luciferase Reporter Assay (Promega) according to the manufacturer's instructions. Enhancer luciferase activity was normalized against the *Renilla* luciferase activity.

Publicly available data used in this work

SOX10 ChIP-seq and MITF ChIP-seq data on the 501Mel melanoma cell lines were downloaded as raw FASTQ files from the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) through accession number GSE61965 (Laurette et al. 2015) and were mapped to the human genome using Bowtie 2 (v2.1.0) (Langmead and Salzberg 2012) and peaks were called by MACS2 (v2.1.1) (Gaspar 2018). TFAP2A ChIP-seq data on human primary melanocytes from neonatal foreskin were retrieved from Seberg et al. (2017) (GSE67555) as a BED file, which was converted to a bedGraph and bigWig using the peak height from the BED file. Histone H3 at lysine 27 (H3K27ac) and H3 monomethylation at K3 (H3K4me1) ChIP-seq data for MM001 (GSE60666); and RNA-seq data (for MM031, MM034, MM057, MM074, MM087, MM099, and MM118 downloaded from GSE60666; for MM029, MM116, MM0163, MM164, and MM165 from GSE134432) were processed as explained in Verfaillie et al. (2015). OmniATAC-seq data for the human lines MM001, MM011, MM029, MM031, MM074, MM057, MM087, and MM099 were obtained through GSE134432 (Wouters et al. 2020) and were processed as described above in "Data processing of ATAC-seq and OmniATAC-seq samples"; which was also the case for ATAC-seq data from normal human melanocytes on foreskin (NHM1), which were downloaded as raw FASTQ files from GSE94488 (GSM2476338) (Fontanals-Cirera et al. 2017). The massively parallel reporter assay (MPRA) data on the *IRF4* enhancer was downloaded from https://mpra.gs.washington.edu/satMutMPRA/ and was processed as described above.

Data access

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE142238. This includes OmniATAC-seq data of human melanoma cell lines (MM029, MM034, MM052, MM116, MM118, MM122, MM163, MM164, MM165; data for the other lines used in this study were published before [see "Publicly available data used in this work"]), two dog melanoma cell lines, two horse melanoma cell lines, one pig melanoma sample, one pig melanocyte cell line, and one mouse melanoma cell line; ATAC-seq data of four zebrafish cell lines; and OmniATAC-seq data of SOX10 and TFAP2A knockdown in the human melanoma cell line MM001. The DeepMEL model was deposited in Kipoi (Avsec et al. 2019) (http://kipoi.org/models/DeepMEL/). Code and custom scripts for training DeepMEL, DeepMEL predictions, DeepExplainer usage, and BLS scoring are provided in GitHub (https://github.com/aertslab/DeepMEL) and as Supplemental Code.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This work was supported by a European Research Council Consolidator Grant to S.A. (No. 724226_cis-CONTROL), the Catholic University of Leuven (Grant No. C14/18/092 to S.A.), the Foundation Against Cancer (Grant No. 2016-070 to S.A.), a PhD fellowship from the Fonds Wetenschappelijk Onderzoek (to L.M., No. 1S03317N), and a postdoctoral research fellowship from Kom op tegen Kanker (Stand up to Cancer; the Flemish Cancer Society) and Stichting tegen Kanker (Foundation against Cancer; the Belgian Cancer Society) (to J.W.). We thank Odessa Van Goethem and Véronique Benne for their contribution in establishing and providing the mouse melanoma cell line and Leif Andersson for sharing the horse melanoma cell lines. We thank Catherine André (National Centre for Scientific Research–University of Rennes 1, UMR6290, Institute of Genetics and Development of Rennes, Faculty of Medicine, Rennes, France) and Cani-DNA Biological Resource Centre (BRC) (Biosit, Rennes, France) for sharing the in-house canine oral and uveal melanoma cell lines. The Cani-DNA BRC (https://dog-genetics.genouest.org) is funded through the BRC-Anim PIA1 funding (2012–2022) ANR-11-INBS-0003. In addition, we thank Austin George for his help with the hyperparameter optimization. Computing was performed at the Vlaams Supercomputer Center, and high-throughput sequencing was done via the Genomics Core Leuven. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions: L.M., I.I.T., and S.A. conceived the study. L.M. performed the experimental work for the mammalian OmniATAC-seq data set, with the help of L.V.A., S.M., V.C., and J.W. M.F., E.v.R., and L.Z. established and maintained the zebrafish cell lines and performed ATAC-seq on these. G.E. maintained and provided the pig cell lines. A.P. and E.C. established and provided the dog cell lines. P.K. and J.-C.M. established and provided the mouse melanoma cell line. M.S. established and provided the horse cell lines. G.-E.G. established and provided the human cell lines. L.M. performed the experimental work and analysis of the luciferase assays together with D.M. L.M. performed the bioinformatic analyses of the OmniATAC-seq data set. G.H. wrote the scripts to perform the branch length scoring analysis. I.I.T. established the neural network and performed all bioinformatic analyses regarding the model. L.M., I.I.T., J.W., and S.A. wrote the manuscript.

References

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, et al. 2016. Tensorflow: large-scale machine learning on heterogeneous distributed systems. arXiv:1603.04467 [cs.DC].

Alipanahi B, Delong A, Weirauch MT, Frey BJ. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* **33:** 831–838. doi:10.1038/nbt.3300

Angermueller C, Lee HJ, Reik W, Stegle O. 2017. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol* **18:** 67. doi:10.1186/s13059-017-1189-z

Arendt D, Musser JM, Baker CVH, Bergman A, Cepko C, Erwin DH, Pavlicev M, Schlosser G, Widder S, Laubichler MD, et al. 2016. The origin and evolution of cell types. *Nat Rev Genet* **17:** 744–757. doi:10.1038/nrg.2016.127

Arunachalam M, Jayasurya K, Tomancak P, Ohler U. 2010. An alignment-free method to identify candidate orthologous enhancers in multiple *Drosophila* genomes. *Bioinformatics* **26:** 2109–2115. doi:10.1093/bioinformatics/btq358

Atak ZK, Taskiran II, Flerin C, Mauduit D, Minnoye L, Hulsemans G, Christiaens V, Ghanem GE, Wouters J, Aerts S. 2019. Prioritization of enhancer mutations by combining allele-specific chromatin accessibility with deep learning. bioRxiv doi:10.1101/2019.12.21.885806

Avsec Ž, Kreuzhuber R, Israeli J, Xu N, Cheng J, Shrikumar A, Banerjee A, Kim DS, Beier T, Urban L, et al. 2019. The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. *Nat Biotechnol* **37:** 592–600. doi:10.1038/s41587-019-0140-0

Avsec Ž, Weilert M, Shrikumar A, Krueger S, Alexandari A, Dalal K, Fropf R, McAnany C, Gagneur J, Kundaje A, et al. 2020. Base-resolution models of transcription factor binding reveal soft motif syntax. bioRxiv doi:10.1101/737981

Baglama J, Reichel L. 2005. Augmented implicitly restarted Lanczos bidiagonalization methods. *SIAM J Sci Comput* **27:** 19–42. doi:10.1137/04060593X

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37:** W202–W208. doi:10.1093/nar/gkp335

Ballester B, Medina-Rivera A, Schmidt D, Gonzàlez-Porta M, Carlucci M, Chen X, Chessman K, Faure AJ, Funnell APW, Goncalves A, et al. 2014. Multi-species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways. *eLife* **3:** e02626. doi:10.7554/eLife.02626

Bravo González-Blas C, Minnoye L, Papasokrati D, Aibar S, Hulselmans G, Christiaens V, Davie K, Wouters J, Aerts S. 2019. CisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat Methods* **16:** 397–400. doi:10.1038/s41592-019-0367-1

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10:** 1213–1218. doi:10.1038/nmeth.2688

Ceol CJ, Houvras Y, Jane-Valbuena J, Bilodeau S, Orlando DA, Battisti V, Fritsch L, Lin WM, Hollmann TJ, Ferré F, et al. 2011. The histone methyltransferase SETDB1 is recurrently amplified in melanoma and accelerates its onset. *Nature* **471:** 513–517. doi:10.1038/nature09806

Chen L, Fish AE, Capra JA. 2018. Prediction of gene regulatory enhancers across species reveals evolutionarily conserved sequence properties. *PLoS Comput Biol* **14:** e1006484. doi:10.1371/journal.pcbi.1006484

Cliften PF, Hillier LW, Fulton L, Graves T, Miner T, Gish WR, Waterston RH, Johnston M. 2001. Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res* **11:** 1175–1186. doi:10.1101/gr.182901

Corces MR, Trevino AE, Hamilton EG, Greenside PG, Sinnott-Armstrong NA, Vesuna S, Satpathy AT, Rubin AJ, Montine KS, Wu B, et al. 2017. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat Methods* **14:** 959–962. doi:10.1038/nmeth.4396

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci* **107:** 21931–21936. doi:10.1073/pnas.1016071107

Dankort D, Curley DP, Cartlidge RA, Nelson B, Karnezis AN, Damsky WE Jr, You MJ, DePinho RA, McMahon M, Bosenberg M. 2009. *Braf*^{V600E} cooperates with Pten loss to induce metastatic melanoma. *Nat Genet* **41:** 544–552. doi:10.1038/ng.356

De Mazière AM, Muehlethaler K, van Donselaar E, Salvi S, Davoust J, Cerottini JC, Lévy F, Slot JW, Rimoldi D. 2002. The melanocytic protein Melan-A/MART-1 has a subcellular localization distinct from typical melanosomal proteins. *Traffic* **3:** 678–693. doi:10.1034/j.1600-0854.2002.30909.x

Denny SK, Yang D, Chuang CH, Brady JJ, Lim JS, Grüner BM, Chiou SH, Schep AN, Baral J, Hamard C, et al. 2016. Nfib promotes metastasis through a widespread increase in chromatin accessibility. *Cell* **166:** 328–342. doi:10.1016/j.cell.2016.05.052

Dermitzakis ET, Clark AG. 2002. Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol Biol Evol* **19:** 1114–1121. doi:10.1093/oxfordjournals.molbev.a004169

D'Mello SAN, Finlay GJ, Baguley BC, Askarian-Amiri ME. 2016. Signaling pathways in melanogenesis. *Int J Mol Sci* **17:** 1144. doi:10.3390/ijms17071144

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29:** 15–21. doi:10.1093/bioinformatics/bts635

Dodonova SO, Zhu F, Dienemann C, Taipale J, Cramer P. 2020. Nucleosome-bound SOX2 and SOX11 structures elucidate pioneer factor function. *Nature* **580:** 669–672. doi:10.1038/s41586-020-2195-y

Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W. 2005. Biomart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21:** 3439–3440. doi:10.1093/bioinformatics/bti525

Dynan WS, Tjian R. 1983. The promoter-specific transcription factor Sp1 binds to upstream sequences in the SV40 early promoter. *Cell* **35:** 79–87. doi:10.1016/0092-8674(83)90210-6

Egidy G, Julé S, Bossé P, Bernex F, Geffrotin C, Vincent-Naulleau S, Horak V, Sastre-Garau X, Panthier JJ. 2008. Transcription analysis in the MeLiM swine model identifies RACK1 as a potential marker of malignancy for human melanocytic proliferation. *Mol Cancer* **7:** 34. doi:10.1186/1476-4598-7-34

Eraslan G, Avsec Ž, Gagneur J, Theis FJ. 2019. Deep learning: new computational modelling techniques for genomics. *Nat Rev Genet* **20:** 389–403. doi:10.1038/s41576-019-0122-6

Fontanals-Cirera B, Hasson D, Vardabasso C, Di Micco R, Agrawal P, Chowdhury A, Gantz M, de Pablos-Aragoneses A, Morgenstern A, Wu P, et al. 2017. Harnessing BET inhibitor sensitivity reveals AMIGO2 as a melanoma survival gene. *Mol Cell* **68:** 731–744.e9. doi:10.1016/j.molcel.2017.11.004

Frith MC, Li MC, Weng Z. 2003. Cluster-Buster: finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res* **31:** 3666–3668. doi:10.1093/nar/gkg540

Fufa TD, Harris ML, Watkins-Chow DE, Levy D, Gorkin DU, Gildea DE, Song L, Sa A, Crawford GE, Sviderskaya EV, et al. 2015. Genomic analysis reveals distinct mechanisms and functional classes of SOX10-regulated genes in melanocytes. *Hum Mol Genet* **24:** 5433–5450. doi:10.1093/hmg/ddv267

Gaspar JM. 2018. Improved peak-calling with MACS2. bioRxiv doi:10.1101/496521

Gasperini M, Hill AJ, McFaline-Figueroa JL, Martin B, Kim S, Zhang MD, Jackson D, Leith A, Schreiber J, Noble WS, et al. 2019. A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* **176:** 377–390.e19. doi:10.1016/j.cell.2018.11.029

Gembarska A, Luciani F, Fedele C, Russell EA, Dewaele M, Villar S, Zwolinska A, Haupt S, de Lange J, Yip D, et al. 2012. MDM4 is a key therapeutic target in cutaneous melanoma. *Nat Med* **18:** 1239–1247. doi:10.1038/nm.2863

Graf SA, Busch C, Bosserhoff AK, Besch R, Berking C. 2014. SOX10 promotes melanoma cell invasion by regulating melanoma inhibitory activity. *J Invest Dermatol* **134:** 2212–2220. doi:10.1038/jid.2014.128

Grossman SR, Engreitz J, Ray JP, Nguyen TH, Hacohen N, Lander ES. 2018. Positional specificity of different transcription factor classes within enhancers. *Proc Natl Acad Sci* **115:** E7222–E7230. doi:10.1073/pnas.1804663115

Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. 2007. Quantifying similarity between motifs. *Genome Biol* **8:** R24. doi:10.1186/gb-2007-8-2-r24

Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, Ukkonen E, Taipale J. 2006. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* **124:** 47–59. doi:10.1016/j.cell.2005.10.042

Hamdan FH, Johnsen SA. 2019. Perturbing enhancer activity in cancer therapy. *Cancers (Basel)* **11:** 634. doi:10.3390/cancers11050634

Heilmann S, Ratnakumar K, Langdon E, Kansler E, Kim I, Campbell NR, Perry E, McMahon A, Kaufman C, van Rooijen E, et al. 2015. A quantitative system for studying metastasis using transparent zebrafish. *Cancer Res* **75:** 4272–4282. doi:10.1158/0008-5472.CAN-14-3319

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38:** 576–589. doi:10.1016/j.molcel.2010.05.004

Herrmann C, Van de Sande B, Potier D, Aerts S. 2012. i-cisTarget: an integrative genomics method for the prediction of regulatory features and *cis*-regulatory modules. *Nucleic Acids Res* **40:** e114. doi:10.1093/nar/gks543

Hitte C, Le Béguec C, Cadieu E, Wucher V, Primot A, Prouteau A, Botherel N, Hédan B, Lindblad-Toh K, André C, et al. 2019. Genome-wide analysis of long non-coding RNA profiles in canine oral melanomas. *Genes (Basel)* **10:** 477. doi:10.3390/genes10060477

Hoek KS, Schlegel NC, Brafford P, Sucker A, Ugurel S, Kumar R, Weber BL, Nathanson KL, Phillips DJ, Herlyn M, et al. 2006. Metastatic potential of melanomas defined by specific gene expression profiles with no BRAF signature. *Pigment Cell Res* **19:** 290–302. doi:10.1111/j.1600-0749.2006.00322.x

Hoek KS, Eichhoff OM, Schlegel NC, Döbbeling U, Kobert N, Schaerer L, Hemmi S, Dummer R. 2008. In vivo switching of human melanoma cells between proliferative and invasive states. *Cancer Res* **68:** 650–656. doi:10.1158/0008-5472.CAN-07-2491

Hong JW, Hendrix DA, Levine MS. 2008. Shadow enhancers as a source of evolutionary novelty. *Science* **321:** 1314. doi:10.1126/science.1160631

Hou L, Srivastava Y, Jauch R. 2017. Molecular basis for the genome engagement by Sox proteins. *Semin Cell Dev Biol* **63:** 2–12. doi:10.1016/j.semcdb.2016.08.005

Imrichová H, Hulselmans G, Kalender Atak Z, Potier D, Aerts S. 2015. i-cisTarget 2015 update: generalized cis-regulatory enrichment analysis in human, mouse and fly. *Nucleic Acids Res* **43:** W57–W64. doi:10.1093/nar/gkv395

Iwafuchi-Doi M, Donahue G, Kakumanu A, Watts JA, Mahony S, Pugh BF, Lee D, Kaestner KH, Zaret KS. 2016. The pioneer transcription factor foxA maintains an accessible nucleosome configuration at enhancers for tissue-specific gene activation. *Mol Cell* **62:** 79–91. doi:10.1016/j.molcel.2016.03.001

Jacobs J, Atkins M, Davie K, Imrichova H, Romanelli L, Christiaens V, Hulselmans G, Potier D, Wouters J, Taskiran II, et al. 2018. The transcription factor grainy head primes epithelial enhancers for spatiotemporal activation by displacing nucleosomes. *Nat Genet* **50:** 1011–1020. doi:10.1038/s41588-018-0140-x

Janky R, Verfaillie A, Imrichová H, van de Sande B, Standaert L, Christiaens V, Hulselmans G, Herten K, Naval Sanchez M, Potier D, et al. 2014. iRegulon: from a gene list to a gene regulatory network using large motif and track collections. *PLoS Comput Biol* **10:** e1003731. doi:10.1371/journal.pcbi.1003731

Jiang L, Campagne C, Sundström E, Sousa P, Imran S, Seltenhammer M, Pielberg G, Olsson MJ, Egidy G, Andersson L, et al. 2014. Constitutive activation of the ERK pathway in melanoma and skin melanocytes in Grey horses. *BMC Cancer* **14:** 857. doi:10.1186/1471-2407-14-857

Johnson LA, Zhao Y, Golden K, Barolo S. 2008. Reverse-engineering a transcriptional enhancer: a case study in *Drosophila*. *Tissue Eng Part A* **14:** 1549–1559. doi:10.1089/ten.tea.2008.0074

Julé S, Bossé P, Egidy G, Panthier JJ. 2003. Establishment and characterization of a normal melanocyte cell line derived from pig skin. *Pigment Cell Res* **16:** 407–410. doi:10.1034/j.1600-0749.2003.00071.x

Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J, et al. 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458:** 362–366. doi:10.1038/nature07667

Kaufman CK, Mosimann C, Fan ZP, Yang S, Thomas AJ, Ablain J, Tan JL, Fogley RD, van Rooijen E, Hagedorn EJ, et al. 2016. A zebrafish melanoma model reveals emergence of neural crest identity during melanoma initiation. *Science* **351:** aad2197. doi:10.1126/science.aad2197

Kawakami A, Fisher DE. 2017. The master role of microphthalmia-associated transcription factor in melanocyte and melanoma biology. *Lab Invest* **97:** 649–656. doi:10.1038/labinvest.2017.9

Kelley DR, Snoek J, Rinn JL. 2016. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res* **26:** 990–999. doi:10.1101/gr.200535.115

Kircher M, Xiong C, Martin B, Schubach M, Inoue F, Bell RJA, Costello JF, Shendure J, Ahituv N. 2019. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat Commun* **10:** 3583. doi:10.1038/s41467-019-11526-w

Klein RM, Bernstein D, Higgins SP, Higgins CE, Higgins PJ. 2012. SERPINE1 expression discriminates site-specific metastasis in human melanoma. *Exp Dermatol* **21:** 551–554. doi:10.1111/j.1600-0625.2012.01523.x

Klemm SL, Shipony Z, Greenleaf WJ. 2019. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet* **20:** 207–220. doi:10.1038/s41576-018-0089-8

Kuhn RM, Haussler D, Kent WJ. 2013. The UCSC Genome Browser and associated tools. *Brief Bioinform* **14:** 144–161. doi:10.1093/bib/bbs038

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9:** 357–359. doi:10.1038/nmeth.1923

Laurette P, Strub T, Koludrovic D, Keime C, Le Gras S, Seberg H, Van Otterloo E, Imrichova H, Siddaway R, Aerts S, et al. 2015. Transcription factor MITF and remodeller BRG1 define chromatin organisation at regulatory elements in melanoma cells. *eLife* **4:** e06857. doi:10.7554/eLife.06857

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ. 2013. Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9:** e1003118. doi:10.1371/journal.pcbi.1003118

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079. doi:10.1093/bioinformatics/btp352

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30:** 923–930. doi:10.1093/bioinformatics/btt656

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, et al. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478:** 476–482. doi:10.1038/nature10530

Long HK, Prescott SL, Wysocka J. 2016. Ever-changing landscapes: transcriptional enhancers in development and evolution. *Cell* **167:** 1170–1187. doi:10.1016/j.cell.2016.09.018

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15:** 550. doi:10.1186/s13059-014-0550-8

Lundberg S, Lee S-I. 2017. A unified approach to interpreting model predictions. arXiv:1705.07874 [cs.AI].

Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee S-I. 2020. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell* **2:** 56–67. doi:10.1038/s42256-019-0138-9

Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, et al. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res* **47:** W636–W641. doi:10.1093/nar/gkz268

Maity SN, de Crombrugghe B. 1998. Role of the CCAAT-binding protein CBF/NF-Y in transcription. *Trends Biochem Sci* **23:** 174–178. doi:10.1016/S0968-0004(98)01201-8

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of *cis*-regulatory regions. *Nat Biotechnol* **28:** 495–501. doi:10.1038/nbt.1630

Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, et al. 2012. The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res* **41:** D64–D69. doi:10.1093/nar/gks1048

Min X, Chen N, Chen T, Jiang R. 2016. DeepEnhancer: predicting enhancers by convolutional neural networks. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 637–644. IEEE, Shenzhen, China. http://ieeexplore.ieee.org/document/7822593/. doi:10.1109/BIBM.2016.7822593

Park Y, Kellis M. 2015. Deep learning for regulatory genomics. *Nat Biotechnol* **33:** 825–826. doi:10.1038/nbt.3313

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20:** 110–121. doi:10.1101/gr.097857.109

Prescott SL, Srinivasan R, Marchetto MC, Grishina I, Narvaiza I, Selleri L, Gage FH, Swigut T, Wysocka J. 2015. Enhancer divergence and *cis*-regulatory evolution in the human and chimp neural crest. *Cell* **163:** 68–83. doi:10.1016/j.cell.2015.08.036

Prouteau A, André C. 2019. Canine melanomas as models for human melanomas: clinical, histological, and genetic comparison. *Genes (Basel)* **10:** 501. doi:10.3390/genes10070501

Quang D, Xie X. 2016. Danq: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res* **44:** e107. doi:10.1093/nar/gkw226

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842. doi:10.1093/bioinformatics/btq033

Rahman MdM, Lai Y, Husna A, Chen H, Tanaka Y, Kawaguchi H, Hatai H, Miyoshi N, Nakagawa T, Fukushima R, et al. 2019. Transcriptome analysis of dog oral melanoma and its oncogenic analogy with human melanoma. *Oncol Rep* **43:** 16–30. doi:10.3892/or.2019.7391

Rambow F, Malek O, Geffrotin C, Leplat JJ, Bouet S, Piton G, Hugot K, Bevilacqua C, Horak V, Vincent-Naulleau S. 2008. Identification of differentially expressed genes in spontaneously regressing melanoma using the meLiM swine model: differential gene expression in swine melanoma. *Pigment Cell Melanoma Res* **21:** 147–161. doi:10.1111/j.1755-148X.2008.00442.x

Rambow F, Marine JC, Goding CR. 2019. Melanoma plasticity and phenotypic diversity: therapeutic barriers and opportunities. *Genes Dev* **33:** 1295–1318. doi:10.1101/gad.329771.119

Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, Manke T. 2016. Deeptools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44:** W160–W165. doi:10.1093/nar/gkw257

R Core Team. 2017. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

R Core Team. 2018. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Robinson MD, McCarthy DJ, Smyth GK. 2010. Edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26:** 139–140. doi:10.1093/bioinformatics/btp616

Rosengren Pielberg G, Golovko A, Sundström E, Curik I, Lennartsson J, Seltenhammer MH, Druml T, Binns M, Fitzsimmons C, Lindgren G, et al. 2008. A *cis*-acting regulatory mutation causes premature hair graying and susceptibility to melanoma in the horse. *Nat Genet* **40:** 1004–1009. doi:10.1038/ng.185

Schreiber J, Libbrecht M, Bilmes J, Noble WS. 2017. Nucleotide sequence and DNaseI sensitivity are predictive of 3D chromatin architecture. bioRxiv doi:10.1101/103614

Seberg HE, Van Otterloo E, Loftus SK, Liu H, Bonde G, Sompallae R, Gildea DE, Santana JF, Manak JR, Pavan WJ, et al. 2017. TFAP2 paralogs regulate melanocyte differentiation in parallel with MITF. *PLoS Genet* **13:** e1006636. doi:10.1371/journal.pgen.1006636

Segaoula Z, Primot A, Lepretre F, Hedan B, Bouchaert E, Minier K, Marescaux L, Serres F, Galiègue-Zouitina S, André C, et al. 2018. Isolation and characterization of two canine melanoma cell lines: new models for comparative oncology. *BMC Cancer* **18:** 1219. doi:10.1186/s12885-018-5114-y

Seltenhammer MH, Sundström E, Meisslitzer-Ruppitsch C, Cejka P, Kosiuk J, Neumüller J, Almeder M, Majdic O, Steinberger P, Losert UM, et al. 2014. Establishment and characterization of a primary and a metastatic melanoma cell line from Grey horses. *Vitro Cell Dev Biol - Anim* **50:** 56–65. doi:10.1007/s11626-013-9678-1

Shain AH, Bastian BC. 2016. From melanocytes to melanomas. *Nat Rev Cancer* **16:** 345–358. doi:10.1038/nrc.2016.37

Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP, Karun V, Jaakkola T, Gifford DK. 2014. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat Biotechnol* **32:** 171–178. doi:10.1038/nbt.2798

Shlyueva D, Stampfel G, Stark A. 2014. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet* **15:** 272–286. doi:10.1038/nrg3682

Shoshan E, Braeuer RR, Kamiya T, Mobley AK, Huang L, Vasquez ME, Velazquez-Torres G, Chakravarti N, Ivan C, Prieto V, et al. 2016. NFAT1 directly regulates IL8 and MMP3 to promote melanoma tumor growth and metastasis. *Cancer Res* **76:** 3145–3155. doi:10.1158/0008-5472.CAN-15-2511

Shrikumar A, Greenside P, Kundaje A. 2017. Learning important features through propagating activation differences. arXiv:1704.0268 [cs.CV].

Shrikumar A, Tian K, Avsec Ž, Shcherbina A, Banerjee A, Sharmin M, Nair S, Kundaje A. 2019. Technical note on transcription factor motif discovery from importance scores (TF-MoDISco) version 0.5.6.5. arXiv:1811.00416 [cs.LG].

Siepel A. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15:** 1034–1050. doi:10.1101/gr.3715005

Song L, Crawford GE. 2010. DNase-seq: A high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc* **2010:** pdb.prot5384. doi:10.1101/pdb.prot5384

Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA, Rasmussen MD, Roy S, Deoras AN, et al. 2007. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* **450:** 219–232. doi:10.1038/nature06340

Sundström E, Komisarczuk AZ, Jiang L, Golovko A, Navratilova P, Rinkwitz S, Becker TS, Andersson L. 2012. Identification of a melanocyte-specific, microphthalmia-associated transcription factor-dependent regulatory element in the intronic duplication causing hair greying and melanoma in horses: a melanocyte-specific regulatory element in the duplicated sequence causing greying and melanoma in horses. *Pigment Cell Melanoma Res* **25:** 28–36. doi:10.1111/j.1755-148X.2011.00902.x

Thomas-Chollier M, Hufton A, Heinig M, O'Keeffe S, Masri NE, Roider HG, Manke T, Vingron M. 2011. Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. *Nat Protoc* **6:** 1860–1869. doi:10.1038/nprot.2011.409

Thomas-Chollier M, Herrmann C, Defrance M, Sand O, Thieffry D, van Helden J. 2012. RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res* **40:** e31–e31. doi:10.1093/nar/gkr1104

van der Weyden L, Patton EE, Wood GA, Foote AK, Brenn T, Arends MJ, Adams DJ. 2016. Cross-species models of human melanoma. *J Pathol* **238:** 152–165. doi:10.1002/path.4632

van Rooijen E, Fazio M, Zon LI. 2017. From fish bowl to bedside: The power of zebrafish to unravel melanoma pathogenesis and discover new therapeutics. *Pigment Cell Melanoma Res* **30:** 402–412. doi:10.1111/pcmr.12592

Verfaillie A, Imrichova H, Atak ZK, Dewaele M, Rambow F, Hulselmans G, Christiaens V, Svetlichnyy D, Luciani F, Van den Mooter L, et al. 2015. Decoding the regulatory landscape of melanoma reveals TEADS as regulators of the invasive cell state. *Nat Commun* **6:** 6683–6683. doi:10.1038/ncomms7683

Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, Park TJ, Deaville R, Erichsen JT, Jasinska AJ, et al. 2015. Enhancer evolution across 20 mammalian species. *Cell* **160:** 554–566. doi:10.1016/j.cell.2015.01.006

Wang M, Tai C, E W, Wei L. 2018. Define: deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding variants. *Nucleic Acids Res* **46:** e69. doi:10.1093/nar/gky215

White RM, Sessa A, Burke C, Bowman T, LeBlanc J, Ceol C, Bourque C, Dovey M, Goessling W, Burns CE, et al. 2008. Transparent adult zebrafish as a tool for in vivo transplantation analysis. *Cell Stem Cell* **2:** 183–189. doi:10.1016/j.stem.2007.11.002

White RM, Cech J, Ratanasirintrawoot S, Lin CY, Rahl PB, Burke CJ, Langdon E, Tomlinson ML, Mosher J, Kaufman C, et al. 2011. DHODH modulates transcriptional elongation in the neural crest and melanoma. *Nature* **471:** 518–522. doi:10.1038/nature09882

Wojciechowska S, van Rooijen E, Ceol C, Patton EE, White RM. 2016. Generation and analysis of zebrafish melanoma models. *Methods Cell Biol* **134:** 531–549. doi:10.1016/bs.mcb.2016.03.008

Wouters J, Kalender-Atak Z, Minnoye L, Spanier KI, De Waegeneer M, Bravo González-Blas C, Mauduit D, Davie K, Hulselmans G, Najem A, et al. 2020. Robust gene expression programs underlie recurrent cell states and phenotype switching in melanoma. *Nat Cell Biol* **22:** 986–998. doi:10.1038/s41556-020-0547-3

Zaret KS, Carroll JS. 2011. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev* **25:** 2227–2241. doi:10.1101/gad.176826.111

Zhou J, Troyanskaya OG. 2015. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* **12:** 931–934. doi:10.1038/nmeth.3547

Received January 30, 2020; accepted in revised form June 15, 2020.

Cross-species analysis of enhancer logic using deep learning

Liesbeth Minnoye, Ibrahim Ihsan Taskiran, David Mauduit, et al.

Genome Res. 2020 30: 1815-1834 originally published online July 30, 2020
Access the most recent version at doi:10.1101/gr.260844.120