Interpretation of allele-specific chromatin accessibility using cell state-aware deep learning

Zeynep Kalender Atak[1,2,$,#], Ibrahim Ihsan Taskiran[1,2,#], Jonas Demeulemeester[1,2,3], Christopher Flerin[1,2], David Mauduit[1,2], Liesbeth Minnoye[1,2], Gert Hulselmans[1,2], Valerie Christiaens[1,2], Ghanem-Elias Ghanem[4], Jasper Wouters[1,2], and Stein Aerts[1,2,*].

1. VIB-KU Leuven Center for Brain & Disease Research, Leuven, Belgium.
2. KU Leuven, Department of Human Genetics KU Leuven, Leuven, Belgium.
3. Cancer Genomics Laboratory, The Francis Crick Institute, London, UK.
4. Institut Jules Bordet, Université Libre de Bruxelles, Brussels, Belgium.

# equal contribution
* corresponding author
$ current address: Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge UK.

Abstract

Genomic sequence variation within enhancers and promoters can have a significant impact on the cellular state and phenotype. However, sifting through the millions of candidate variants in a personal genome or a cancer genome, to identify those that impact cis-regulatory function, remains a major challenge. Interpretation of non-coding genome variation benefits from explainable artificial intelligence to predict and interpret the impact of a mutation on gene regulation. Here we generate phased whole genomes with matched chromatin accessibility, histone modifications, and gene expression for 10 melanoma cell lines. We find that training a specialized deep learning model, called DeepMEL2, on melanoma chromatin accessibility data can capture the various regulatory programs of the melanocytic and mesenchymal-like melanoma cell states. This model outperforms motif-based variant scoring, as well as more generic deep learning models. We detect hundreds to thousands of allele-specific chromatin accessibility variants (ASCAVs) in each melanoma genome, of which 15-20% can be explained by gains or losses of transcription factor binding sites. A considerable fraction of ASCAVs are caused by changes in AP-1 binding, as confirmed by matched ChIP-seq data to identify allele-specific binding of JUN and FOSL1. Finally, by augmenting the DeepMEL2 model with ChIP-seq data for GABPA, the TERT promoter mutation as well as additional ETS motif gains can be identified with high confidence. In conclusion, we present a new integrative genomics approach and a deep learning model to identify and interpret functional enhancer mutations with allelic imbalance of chromatin accessibility and gene expression.

Introduction

Understanding the functional consequences of non-coding variants is still a fundamental challenge in human genetics. Genome-wide association studies indicate that almost 90% of disease-related variants reside in non-coding regions (Hindorff et al. 2009; Maurano et al. 2012), and these regions are enriched for transcription factor binding sites (Khurana et al. 2013). A large body of work has been devoted to identifying non-coding variants that alter gene regulation by linking them to functional genomics data (Gaffney et al. 2012; GTEx Consortium 2015; Banovich et al. 2018; Chen et al. 2016). Broadly, the approaches taken to address this problem can be classified into two groups. The first one is quantitative trait loci (QTL) analysis in which a variant is correlated to a cellular trait (e.g. expression, binding, accessibility) across a large number of samples. This strategy is widely used with expression, chromatin immunoprecipitation, and chromatin accessibility data, for detecting, respectively, QTL associated with gene expression (eQTL), transcription factor binding (bQTL) (Kilpinen et al. 2013), histone modifications (hQTL) (McVicker et al. 2013), or chromatin accessibility (caQTL) (Maurano et al. 2015). This type of analysis is cost-efficient and can be conducted with array-based data, but requires large sample sizes, since effect sizes are usually low (Do et al. 2017). Moreover, the resolution is typically too low to pinpoint a single variant due to linkage disequilibrium (typically spanning 10 to 100kb) (Do et al. 2017). Additionally, structural variation and rare variants (minor allele frequency < 0.05) are often ignored in these studies (Chen et al. 2016; Audano et al. 2019). The alternative approach is to assess allelic imbalance at a heterozygous site directly. This allele counting approach has been extensively used with RNA-seq data to identify allele-specific expression (Castel et al. 2015), but is also applicable to other types of functional genomics data (Chen et al. 2016; Rozowsky et al. 2011). Here, the strategy relies on finding the allelic origin of the observed signal. Unlike QTL analysis, this approach does not depend on large sample sizes and can be used to find rare or even *de novo* regulatory variants; however it requires higher genomic coverage and more complex data processing. Technical issues inherent to alignment and variant calling procedures such as reference bias (i.e. reads originating from the reference allele map better than those containing the variant), ambiguous alignments, and copy number alterations need to be addressed in order to obtain accurate measures of allelic imbalance (Chen et al. 2016; Rozowsky et al. 2011; de Santiago et al. 2017). The use of personalized diploid genomes instead of a haploid reference has been suggested to prevent some of these technical biases (Chen et al. 2016; Castel et al. 2015; Rozowsky et al. 2011).

Both QTL analysis and inference of allelic imbalance can lead to the identification of candidate gene regulatory variants. However, they typically yield little information as to the precise regulatory mechanisms affected by these variants. More than 70% of non-

coding variants associated with common diseases overlap with transcription factor binding sites (Maurano et al. 2012), however studies so far showed that the majority of the variants associated with allele-specific enhancer activity cannot be explained by TF motif alterations (Kilpinen et al. 2013; Maurano et al. 2015; Deplancke, Alpern, and Gardeux 2016; Kumasaka, Knights, and Gaffney 2016; Tehranchi et al. 2019; Degner et al. 2012; Waszak et al. 2015). This might be due to the inadequacy of current TF motif models (such as position weight matrices) that do not take other enhancer features into account, such as flanking sequence context, DNA shape or combinatorial TF binding (Inukai, Kock, and Bulyk 2017). Leveraging the extensive chromatin and TF binding data available, machine learning approaches hold promise to predict TF-bound regions and chromatin changes with single-nucleotide resolution. However, these models require correct training and rigorous validation, and are typically trained either for a single TF (Alipanahi et al. 2015; Quang and Xie 2019; Lee 2016; Avsec et al. 2021) or hundreds of epigenomic features (Zhou and Troyanskaya 2015; Kelley, Snoek, and Rinn 2016). These models tend to be cell type-specific, resulting in reduced performance when applied to other cell types (Banovich et al. 2018). We have previously shown that specialized deep learning models can outperform generic ones at predicting regulatory features and the effect of sequence variation across species (Minnoye et al. 2020).

Here, we perform a comprehensive analysis of 10 melanoma whole genomes to identify and characterize functional non-coding variants. By integrating sample-matched phased whole-genome sequencing, Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq), chromatin immunoprecipitation against H3K27ac (ChIP-seq) and transcriptome sequencing (RNA-seq) data, we identify allele-specific regulatory changes. To interpret how sequence variation affects the gain or loss of transcription factor binding sites, we used a deep learning model, called DeepMEL2, that is trained on different melanoma cell states. We investigate the benefits and limitations of cell state-specific deep learning and motif analysis to unravel how enhancer mutations affect gene regulation.

Results

Identification of ASCAVs using linked-read genome sequencing and ATAC-seq

We obtained haplotype-resolved whole-genome sequencing (WGS) data of 10 patient-derived melanoma cultures (MM lines) using linked-reads technology from 10x Genomics (**Figure 1A, Figure S1, Table S1,2**). Samples were sequenced to an average depth of 38×, apart from MM087 and MM099, which were sequenced more deeply (68× and 133× coverage, respectively). We also profiled chromatin accessibility of the same melanoma lines using OmniATAC-seq (9 samples were reanalysed from (Wouters et al. 2020); while data for A375 was generated in this study).

We find a total of 16 million phased variants across the 10 genomes (**Figure 1B**) and pinpoint as likely somatic in origin between 206,724 and 304,754 of these per sample, based on their absence from the Genome Aggregation Database (gnomAD, v3.0). To dissect the contributions of distinct mutational processes, we estimated exposures to the COSMIC v3 single base substitution signatures. We used the Bayesian approach implemented in SigFit (Gori and Baez-Ortega 2020) and fitted those signatures that have previously been reported as active in melanoma (**Figure S2**). Overall, while a mean 10.9% of somatic mutations showed the unique footprint of UV-damage (COSMIC v3 SBS7a–d & 38; range 10.8–26.2%), 88.5% found its origin in the endogenous “clock-like” mutational processes SBS1, 5, and 40 (Alexandrov et al. 2020)(range 73.7–98.7%; (**Figure S2**). Note that in addition to true somatic SNVs having arisen during tumorigenesis, these “clock-like” variants will contain contributions from germline SNPs which were absent from gnomAD as well as SNVs having arisen during passaging in culture. Only an average 0.5% of somatic mutations were assigned to the remaining signatures (range 0.1–2.3%).

Next, we constructed a personal genome for each sample, to optimize the mapping accuracy at variant positions (see Methods). Combining these personal genomes with ATAC-seq, we found 231,370 variants (of the 16 million) that overlap with ATAC-seq peaks. We then tested each of these variants for allelic imbalance of the overlapping ATAC-seq signal using a modified *alleleseq* pipeline, yielding allele-specific chromatin accessibility variants (ASCAVs; **Figure 1C,D**; **Materials and Methods**) (Chen et al. 2016). Although binomial or beta-binomial tests are used for the detection of allele-specific events, the ubiquitous presence of copy number aberrations in cancer genomes violates the assumptions of these methods. Therefore, we plugged in the Bayesian framework of BaalChIP, which specifically addresses this problem (de Santiago et al. 2017). BaalChIP enabled us to correct the allelic ratios observed in ATAC-seq reads using the genomic allelic ratios from the WGS data to correct for the extensive copy number variation and frequent whole-genome doubling in these lines (**Figure S3,4**). This pipeline resulted in 19,983 significant ASCAVs (8.6% of the variants that overlap

with an ATAC-seq peak) across the 10 genomes (range 451-7,183 per sample) (**Table S3,4**). The majority of ASCAVs are unique to one MM line, and a small proportion is shared between multiple samples (1,073 out of 19,983; 5.4%) (**Figure S5**). Only two of the shared ASCAVs are called discordantly between the samples, all of which are known multiallelic polymorphic SNPs (rs138784536, rs9880846), illustrating the accuracy of the ASCAV pipeline. We also assembled a set of control heterozygous variants within ATAC-seq peaks that show no allelic bias. The genomic distribution of both sets, ASCAVs and control variants, is highly similar (**Figure S6, Table S5**).

Even though most ASCAVs are germline polymorphisms (88.9%), 2,201 ASCAVs are likely somatic (i.e. absent from gnomAD v3.0). Somatic variants hence appear more likely to constitute ASCAVs than do germline variants (chi-square test per sample, all $p \leq 3.28E\text{-}8$). This may be explained in part by increased local mutation rates at TF-bound motifs, negative selection in the germline and/or positive selection in the tumour. Furthermore, if a mutational process would be more (or less) prone to introduce ASCAVs, this may be detectable as a disproportionate contribution of the corresponding mutation signature to somatic ASCAVs compared to all somatic variants. We therefore re-estimated mutational signature exposures specifically from the somatic ASCAVs and contrasted these estimates with the signature exposures obtained from all somatic variants. No consistent differences in signature activities could be detected (Figure S7). Nevertheless, the 5 samples with the highest overall proportion of SBS7a mutations showed a smaller contribution of this type of mutations to ASCAVs, suggesting UV-induced lesions do not systematically affect chromatin accessibility.

To further evaluate whether our ASCAV detection pipeline is robust to copy number variation, we inferred allele-specific copy number from the WGS data (Van Loo, PNAS, 2010), evidencing extensive aneuploidy (Figure 1E, Figure S4). In addition, using sample ploidy and the fraction of the genome with loss of heterozygosity, we were able to classify 8 of our 10 lines as having undergone a whole-genome doubling (Figure S8) (Dentro et al. 2020). By considering the number of chromosome copies carrying an ASCAV, we can time the variants with respect to copy number gains: if a variant arose on a chromosome prior to its duplication, it will be duplicated as well (mutation copy number ≥ 2, an “early” variant). If it arose after, only one copy will be present (mutation copy number = 1, a “late” variant). In regions with loss of heterozygosity or gains on both alleles, these two scenarios can be readily distinguished (Gerstung et al. 2020). Apart from a reduction in regions with loss of heterozygosity (i.e. only somatic and no germline variants can be tested for allelic skewing), ASCAVs are called across all copy number states, and both early and late variants are detectable as ASCAV, confirming that our pipelines are robust in the face of copy number changes (**Figure 1E**).

A subset of ASCAVs overlap with allele-specific gene expression and allele-specific histone modifications

To investigate whether allele-specific chromatin accessibility is associated with allelically skewed gene expression or histone modifications, we analyzed matching RNA-seq and H3K27ac ChIP-seq data, for all 10 samples (Verfaillie et al. 2015). RNA-seq and ChIP-seq reads were processed with the same analysis pipeline (**Figure 1C**). For identifying allele-specific expression variants (ASEVs) in the presence of copy number alterations, we employed a beta-binomial test of the RNA-seq allele counts, where the shape of the beta distribution is informed by the corresponding WGS allele counts (**Methods**). We identified 11,578 distinct autosomal ASEVs, associated with 6,029 genes (**Table S6**). One gene, *MAP2K3* (also known as MEK3), shows ASE in all 10 samples and was previously reported to be allele-specifically expressed in various human and mouse tissues (Kukurba et al. 2014; Tuskan et al. 2008). Globally, ASCAVs are enriched near genes with ASE (p-value=0.005; Fisher's exact test) with 6% of ASCAVs located in promoters (< 2kb upstream of a TSS) or introns of ASE genes (**Table S7**).

We also tested for allele-specific H3K27ac ChIP-seq signal using the same pipeline coupled to BaalChIP. Across the 10 lines we identified 4,016 allele-specific histone variants (ASHVs) (**Table S8**), 343 are both ASCAV and ASHV and an additional 170 are within 1kb of an ASCAV. Similar to ASEVs, ASCAVs are enriched near ASHVs when compared to control variants (p-value <2.2E-16; Fisher's exact test) (**Table S9**). When combined, there are 1,589 ASCAVs that are either close to a gene with ASE or close to an ASHV, and 89 of them show significant changes on all three levels (odds ratio of 1.6 and 3.9, respectively, when compared to control variants, with p-value <2.2E-16). One such example is observed near the pigment associated factor TYR (**Figure 1F)** in which four loci exhibit allele-specific events on all three levels (ATAC-seq, H3K27ac ChIP-seq, and RNA-seq). Taken together, our finding that a significant fraction of ASCAVs are linked with ASHV and ASE, support their functional relevance.

Transcription factor motifs are enriched on ASCAVs, with AP-1 being dominant

Next, we next investigated whether ASCAVs affect transcription factor binding sites. We evaluated a variety of regulatory sequence analysis tools to assess which ASCAVs may have arisen through direct *cis*-regulatory changes, such as gains or a losses of transcription factor binding sites (Maurano et al. 2015; Fu et al. 2014; Deplancke, Alpern, and Gardeux 2016), and which ASCAVs are more likely to result from indirect events. We first evaluate simple models, namely position weight matrices (PWM) before

moving on to more advanced deep learning-based models and comparing their prediction accuracy.

Binding site predictions using PWMs are notorious for their high false positive rates (Wasserman and Sandelin 2004), and this problem is aggravated as our collection of PWMs is very large (more than 22,000 PWMs (Janky et al. 2014)). To overcome this problem, we asked whether, for any given transcription factor, multiple binding sites are gained or lost in a sample, or across the cohort. This provides a statistical cue, as we can exploit motif enrichment across all variants, testing which PWM yields a disproportionate number high "delta-PWM scores", compared to control variants. A similar motif enrichment technique has been applied before to identify pioneer factors from chromatin accessibility QTL data (Jacobs et al. 2018). Out of all 22,000 motifs tested, 719 are significantly altered by ASCAVs compared to control SNPs (Fisher's exact Test, FDR 0.05) (**Figure 2A**). As our collection of motifs is highly redundant (multiple PWMs are present per TF), we clustered the 719 significant PWMs into 47 distinct families. We then focused on 13 of these clusters for which the associated transcription factor is known, and that contained at least 6 motifs (**Figure 2B**). This analysis revealed the AP-1 family as the top hit, with a total of 191 enriched PWMs (FET-adjusted p-value threshold 0.05) in 4,011 allele-specific ATAC-seq peaks across the 10 samples (**Figure 2C,D Table S10, Figure S9**). We observed a significant correlation (Kendall's tau 0.68 with p-value =0.035) between expression of AP-1 factors (all JUN and FOS paralogs together) and the fraction of explainable ASCAVs per sample (**Figure 2E**). Indeed, MM lines of the mesenchymal subtype (MES; MM099, MM047, and MM029) have higher AP-1 activity and more AP-1 motif gains and losses at ASCAVs compared to MM lines of the melanocytic subtype (MEL; MM031 and MM001). MM011 represents an exception in this case, being of the MEL subtype, but with high AP-1 activity. The remaining lines (MM087, MM057, and MM074) are in an intermediate state (Wouters et al. 2020). Overall, these findings suggest that AP-1 binding sites are strongly correlated with changes in chromatin accessibility, and confirm the power of allele-specific chromatin accessibility profiling to identify both gain- and loss-of-function enhancer mutations. AP-1 has indeed been reported to act as a pioneer factor, resulting in nucleosome displacement at enhancers in murine mammary epithelial cells (Biddie et al. 2011).

Using the entire set of 719 enriched motifs, we calculated the delta-PWM score across all ASCAVs. This allows us to evaluate the sensitivity and specificity of the PWM approach for predicting which variants induce allele-specific chromatin accessibility (**Figure 2F**). At 95% specificity, 1,919 variants are predicted to be ASCAVs. Finally, we also tested whether some motif gains or losses can be negatively associated with accessibility, i.e. the delta-PWM and accessibility are negatively correlated. We only identified motifs linked to transcription factors of the ZEB/SNAI family, which are known repressors in the neural crest lineage, including in melanomas (Caramel et al. 2013;

Peinado, Olmeda, and Cano 2007; Postigo and Dean 1999; Postigo et al. 1999; Denecker et al. 2014) (**Figure S10**).
In conclusion, motifs of relevant TFs are enriched at ASCAVs, suggesting that around 9.6% of variants in ATAC-seq peaks create or break a binding site. In turn, such motif gains or losses likely underlie the observed allele-specific chromatin accessibility signals.

A cell state-aware deep learning model can interpret ASCAVs

We next tried to improve the accuracy obtained with the PWM approach using more advanced enhancer modeling. Machine-learning models can be trained on enhancers and take flanking sequence information into account. Examples of deep learning models are Basset (Kelley, Snoek, and Rinn 2016) and DeepSEA (Zhou and Troyanskaya 2015), which are available in Kipoi (Avsec et al. 2018), and can readily be applied to score *cis*-regulatory variants. These generic models have been trained on large collections of epigenomic data (DeepSEA was trained on 919 cell type-specific epigenomic features, Basset was trained on DNase-seq from 164 cell types), allowing their application to "any" cell type. Their prediction accuracy on our MM lines (to discriminate ASCAVs from control variants) is usually higher than the motif-based approach (**Figure S11**). Particularly on the MES lines Basset and DeepSEA achieve high accuracy, explaining 14–16% of ASCAVs by motif changes, at 95% specificity. This is likely due to the fact that the training data was rich in AP-1 bound enhancers, which are well represented in the ENCODE repositories on which DeepSEA and Basset were trained.

Next, we train our own deep learning model that takes the main melanoma cell states into account, namely the melanocytic state (MEL) expressing melanocyte specific transcription factors and pigmentation genes, and the mesenchymal-like state (MES) where cells are more invasive and therapy resistant (Verfaillie et al. 2015; Wouters et al. 2020; Bravo González-Blas et al. 2019). In our previous work (Minnoye et al. 2020), we trained a deep learning model, DeepMEL, on ATAC-seq from a cohort of 16 melanoma samples, including the 10 MM lines used in this work. Applying DeepMEL to discriminate ASCAVs from control variants outperforms the generic models (Basset and DeepSEA) for MEL, but not for MES lines. This is likely because the generic models were trained on a larger data set with a high amount of MES-like genomic enhancers. In contrast, melanocyte and MEL-melanoma states are likely under-represented in ENCODE resulting in models that are not fully "aware" of this regulatory program.
We then asked whether we could further improve DeepMEL including additional ATAC-seq data (and further below also ChIP-seq). We extended our DeepMEL cohort with 14

new samples, to a total of 30 melanoma lines. A cisTopic (Bravo González-Blas et al. 2019) analysis on this larger cohort identifies 47 *cis*-regulatory topics, where 2 of them are generally accessible across all cell lines (Topic-14 and Topic-31) and 9 state-specific MEL and MES topics (**Figure 3A**). We also enhanced the deep learning framework by including the 283 known clustered and partitioned PWMs from the JASPAR database (Fornes et al. 2020) in the convolutional filters that serve as priors (**Figure 3B**). After training DeepMEL2 on the 47 topics, we evaluated its classification performance on left-out data (**Figure S12**). Particularly promoter topics, MEL-topics, and MES-topics achieve high performance, while cell line specific topics are difficult to predict (**Figure 3C**). For the latter, we believe this is due to the fact that the cell line specific topics represent sample-specific copy number variation, rather than differentially accessible regions (**Figure S4**). Next, we applied *in silico* saturation mutagenesis on the MEL-type *IRF4* enhancer. This explainable AI technique, in which each possible mutation in the enhancer sequence is evaluated by re-classification using the model, highlights the outperformance of DeepMEL2 compared to DeepMEL (**Figure 3D**).

Next, we used DeepMEL2 to score all ASCAVs (**Materials and Methods**). On the six MEL lines, DeepMEL2 identifies more ASCAVs compared to all other methods at the same false positive rate. (**Methods**, **Figure 3G, S11**). When we score ASCAVs by using a MEL-specific topic (Topic-17), which represents MEL enhancers, explainable mutations occur more frequently in the samples of the melanocytic subtype, where MEL enhancers are operational (**Figure 3E**). A MES-specific topic (Topic-19), on the other hand, is mostly affected in samples of the mesenchymal subtype (**Figure 3F**). Note that, in agreement with the motif enrichment analysis, AP-1 motif changes (the main drivers of the MES scores, **Figure 2C,D**), are also found enriched for ASCAVs in the melanocytic lines, except for MM001 which has no AP-1 activity (**Figure S9**).

In conclusion, additional enhancement and training of DeepMEL2 further improves prediction of functional *cis*-regulatory changes, particularly on melanoma samples of the MEL subtype, and provides high-resolution insight into precise enhancer changes.

DeepMEL2 predictions on ASCAVs are confirmed by allele-specific TF binding

We predicted a large fraction of allele-specific AP-1 binding sites that are associated with an allele-specific ATAC-seq peak. To test whether AP-1 factors indeed bind preferentially to the predicted allele, we performed ChIP-seq against four AP-1 family members (JUN, JUNB, FOS, FOSL1) that are expressed in the MES-type MM099 line. The ChIP-seq peaks of all four data sets are enriched for the AP-1 motif (**Figure S13**). The FOSL1 and JUN ChIP-seq yield the highest quality peaks, suggesting that these play a role in MM099 and that these antibodies are of high quality (**Figure S13**). Using

the pipeline developed above to infer ASCAVs, we now identify 583 significant allele-specific binding (ASB) events for JUN and 241 for FOSL1 (JUNB and FOS yield only 138 ASBs in total) and some of them are identified as ASCAVs in other cell lines as well (**Figure 4A**). The MES-specific topics are able to predict allele-specific AP-1 binding events (**Figure 4B**). When we rank all MM099 ASCAVs by their maximum score from the different models (i.e. delta between the two alleles for the PWM approach, Basset, DeepSEA, DeepMEL, and DeepMEL2), we find that DeepMEL2 performs best at enriching for ASB events (**Figure 4C**). This means that a significant fraction of the ASCAVs with high DeepMEL2 delta scores are indeed ASB for JUN or FOSL1. Note that whereas Basset and DeepSEA are better at distinguishing ASCAVs from control variants in MEL099, this is not the case for predicting ASB. Since DeepMEL2 was trained to distinguish 47 different melanoma *cis*-regulatory topics, we can also score ASCAVs using specific topics. Leveraging the best performing MES topic (Topic-19) indeed further improves the prediction of allele-specific AP-1 binding (**Figure 4C**). Note that we did not search specifically for AP-1 sites, but rather exploit the regulatory topic of the matching cell state to score the genomic variants.

In a second validation experiment we evaluated the putative effect of ASCAVs on enhancer activity. Our phased genomes allow direct linking of ASCAVs to allele-specific expression of nearby genes. A total of 6.5% of all ASCAVs that can be explained by DeepMEL2 are located in the promoter or body of genes with ASE. Therefore, these enhancer mutations may underlie the expression imbalance of the target gene. To further examine this, we selected three enhancers in MM057 for which the predicted target gene shows ASE. The first two examples, *PEPD* and *MITF*, have a DeepMEL2-predicted AP-1 motif gain (**Figure 4D,E)**. Luciferase reporter assays in this cell line, using sequences of both haplotypes, confirm the potency of these variants to drive enhancer activity and gene expression, only when the AP-1 site is present (**Figure 4D,E**). The third example is an enhancer in the first intron of *EVA1C* with a predicted SOX10 motif gain. This variant (rs2833812) is identified as a phased heterozygous SNP in four lines (MM031, MM057. MM074, MM087) and results in allele specific accessibility in all cases (**Figure S14**). Again, when assessed in a luciferase reporter assay (in MM057), only the enhancer sequence that carries the allele generating the SOX10 motif is able to drive luciferase expression (**Figure 4F**). This confirms that enhancer mutations associated with changes in chromatin accessibility can have an effect on the expression of nearby genes.

Analysis of *TERT* promoter mutations by augmenting DeepMEL2 with ChIP-seq data

As a final analysis we asked whether DeepMEL2 can identify oncogenic mutations in the *TERT* promoter (Horn et al. 2013; Huang et al. 2013). Two *TERT* promoter hotspots are recurrently mutated (C228T and C250T) across a large fraction of cancers of the central nervous system (43%), bladder cancer (59%), melanoma (29 %), and other cancer types (Vinagre et al. 2013). These gain-of-function mutations create a binding site for the ETS-family transcription factors, notably GAPBA, leading to up-regulation of the *TERT* oncogene (Bell et al. 2015). The A375 cell line contains one of these mutations, which is predicted as an ASCAV (**Figure 5A**).

First, we evaluated the accuracy of DeepMEL2 and other methods to predict functional changes across the *TERT* promoter, by comparing the predictions to a previously published saturation mutagenesis screen, performed in a glioblastoma cell line (Kircher et al. 2019). In silico saturation mutagenesis of the *TERT* promoter, scored with DeepMEL2, correlates strongly (54%) with the experimental data, outperforming other methods (**Figure 5B**). Despite the ability of DeepMEL2 to interpret overall *cis*-regulatory variation in the *TERT* promoter, the model does not predict the oncogenic gains of GABPA sites themselves to cause an increase in enhancer activity.

In an attempt to improve interpretation of oncogenic *TERT* mutations, we retrained DeepMEL2 by adding a 48th topic representing GABPA binding. Particularly, we labeled all ATAC-seq peaks that overlap with GABPA ChIP-seq peaks (ENCODE accession ENCSR000BJK) as topic 48. With this fine-tuned model, the explainability of the entire *TERT* promoter increases to 68% (**Figure 5B**), and both *TERT* mutations are identified (**Figure 5C,D,E**).

This provides us with a model that can potentially identify other functional gains or losses of GABPA binding sites. In fact, 43 ASCAVs across the 10 lines demonstrate a higher Topic-48 delta score than the known *TERT* promoter mutations. Thirteen of these are observed in other cancers (listed in COSMIC), and five are located in the promoter of a gene with ASE (**Figure 5D**, **Figure S15, Table S11**). Thus, the augmented DeepMEL2 model can be used to interpret *cis*-regulatory variation in melanoma genomes, with awareness of the MEL and MES enhancer code, the proximal promoter code, and *cis*-regulatory elements used by ETS-family members (**Figure 5F,G,H**).

Discussion

Functional variants affecting crucial genes and pathways are underpinning the fitness advantages of cancer cells. Such mutations may be found by recurrence analysis across patients and even across cancer types, at least in the coding fraction of the genome (Bailey et al. 2018). In the non-coding genome, however, this approach typically breaks down (Melton et al. 2015; Zhang et al. 2018), and *TERT* promoter mutations are a notable exception (Horn et al. 2013; Huang et al. 2013). Overall, non-coding mutations tend not to affect the same nucleotide or the same enhancer across samples. A recent large-scale and comprehensive whole-genome pan-cancer study from the ICGC-TCGA PCAWG consortium identified only 30 regionally recurrent *cis*-regulatory changes (Zhu et al. 2020). One reason for this might be the complex nature of gene regulation: often multiple enhancers are brought into close proximity of a promoter to initiate transcription, and redundancy or cooperativity of these enhancers remains difficult to disentangle (Gasperini, Tome, and Shendure 2020).

Here we address the challenge of identifying functional non-coding variants by focusing on allelic imbalances in chromatin accessibility, and by linking those to changes in the enhancer sequence using an explainable AI model. Then, by exploiting phased genomes, we further link these potentially causal enhancer changes to allele-specific gene expression, TF binding and histone acetylation.

Earlier work has shown that sample-matched epigenomic and transcriptomic data is instrumental to obtain a functional readout for genomic alterations (Chen et al. 2016; Stevenson, Coolon, and Wittkopp 2013; Castel et al. 2015). An intuitive approach to further establish causality is to assess how these variants affect the binding of transcription factors. Indeed, this strategy has previously been applied in several studies using PWMs to explain the impact of allele-specific non-coding variants (Deplancke, Alpern, and Gardeux 2016). However, as PWM scoring requires stringent thresholds to limit the number of false positive predictions, these models typically have low sensitivity (e.g., 3.3–6.2% explainable variants (Maurano et al. 2015) (Tehranchi et al. 2019) (Degner et al. 2012) (Chen et al. 2016)). More sophisticated machine learning methods have been developed that can overcome these limitations. Models including Support Vector Machines (Ghandi et al. 2014; Svetlichnyy et al. 2015) and neural networks (Alipanahi et al. 2015; Zhou and Troyanskaya 2015; Kelley, Snoek, and Rinn 2016; Liu et al. 2018) can be trained on enhancer sequences and used to predict the impact of mutations. In the context of allele-specific variant interpretation in normal genomes, Banovich et al. developed OrbWeaver, a four-layered neural network with log-transformed PWMs of 1,320 TFs as the first layer (Banovich et al. 2018). OrbWeaver was used to predict features of accessible chromatin in induced pluripotent stem cells and successfully captures the effect of chromatin accessibility QTLs in a cell-type specific manner. In another study, Hoffman et al. developed DeepFIGV using DNase-

seq and histone modification data from 75 lymphoblastoid cell lines, and used it to predict allele-specific binding in an independent set of TF ChIP-seq data (Hoffman et al. 2019).

In addition to normal genomes, deep learning models have also been used to understand non-coding mutations in a disease context, such as autism spectrum disorders (Zhou et al. 2019) and pancreatic cancer (Feigin et al. 2017). These studies trained models on regulatory features from a diverse set of tissues and cell types profiled by the ENCODE and Roadmap Epigenomics projects. While broadly applicable, this approach might limit model performance as regulatory activity is context-dependent (The ENCODE Project Consortium 2012). Indeed, models trained on cell-type specific enhancers have been shown to yield better predictions (Banovich et al. 2018; Minnoye et al. 2020). The model we developed here is also context-dependent and captures regulatory information from different melanoma cell states. The mesenchymal-like melanoma state, with a dominant role for AP-1, is shared with other cancer types (Baron et al. 2020) and is therefore well-represented within ENCODE and other resources. As such, models trained on these large compendia (e.g. DeepSEA and Basset), can effectively identify AP-1 motif gains and losses. The melanocytic cell state, on the other hand, is less well represented. As a result, DeepMEL2 achieves a higher accuracy for all melanocytic samples than DeepSEA and Basset.

We combined three components that improve the efficiency to detect enhancer mutations. Firstly, we performed linked-read whole-genome sequencing on pure cancer cells (avoiding normal cell admixture); secondly, we incorporated matched ATAC-seq and RNA-seq data; thirdly, we developed a context-dependent deep learning model, DeepMEL2. The use of pure cancer samples allowed us to accurately correct allelic signals in ATAC-, ChIP- as well as RNA-seq data for genomic copy number alterations, leading to robust inference of functional allelic imbalances (Rozowsky et al. 2011; Chen et al. 2016; de Santiago et al. 2017).

Using DeepMEL2, 10–16% of ASCAVs per MM line can be explained by changes in the *cis*-regulatory code. A sizable fraction of these were attributed to gains or losses of AP-1 binding sites, particularly in mesenchymal enhancers, which, in turn, could be confirmed by assaying allele-specific binding of AP-1 family members JUN and FOSL1. In melanocytic enhancers we found that gains and losses of SOX10 binding sites were most commonly linked with allele-specific chromatin accessibility. Although a large fraction of these events influences binding of a TF, yet without any other observed consequences, a subset does impact enhancer activity and are associated with changes in gene expression.

Finally, we investigated how context-specific models such as DeepMEL2, trained on epigenomic data, can be fine-tuned to better understand under-represented enhancer logic. By examining the recurrently mutated *TERT* promoter, we found that the DeepMEL2 model did not utilize ETS motifs to classify melanoma enhancers. This may

be due to ETS binding sites not discriminating between MEL and MES enhancers, while other sequence features were more informative to predict classes. We resolved this limitation by providing the model with specific ChIP-seq data as an additional label. The augmented model achieved high prediction accuracies on the entire *TERT* promoter, including the known oncogenic mutations.

Our study shows that deep learning models provide a powerful means to pinpoint functional non-coding variation in cancer genomes, which may translate into clinical benefits for future patients. However, our results also suggest that each cancer type may require its own "matched" deep learning model, trained on epigenomic data from the relevant cancer cell states. Whereas the current work uses patient-derived cell cultures to infer ASCAVs, our framework is in principle also applicable to data obtained from bulk tumor biopsies. In a clinical setting, we envision that (1) deep learning models can be trained on ATAC-seq data from pure cancer cell clusters, when single-cell ATAC-seq is applied to a cohort of biopsies; and (2) genomics methods for single-nucleotide variant and copy number calling take normal cell admixture into account. It is worth noting that both points have been successfully demonstrated in the literature (Satpathy et al. 2019; Dentro et al. 2020).

In conclusion, we have shown that high-confidence *cis*-regulatory variants can be detected by directly comparing the alleles of a cancer genome and using specialized predictive and explainable deep learning models trained on corresponding epigenomics profiles. Our compilation of multi-ome melanoma data and a melanoma-specific deep learning model provides unique data sets and a novel framework for understanding the impact of non-coding variants. Our approach is applicable to pure samples of any cancer type and may contribute to the identification of *cis*-regulatory driver mutations.

Materials and Methods

Cell culture

The melanoma MM lines are derived from patient biopsies by the Laboratory of Oncology and Experimental Surgery (Prof. Dr. Ghanem Ghanem) at the Institut Jules Bordet, Brussels (Verfaillie et al. 2015; Wouters et al. 2020; Gembarska et al. 2012). See Supplemental Methods for further details on culture conditions.

Phased whole genome library preparation and sequencing

The extraction of high molecular weight (HMW) genomic DNA (gDNA) and subsequent preparation of phased whole genome libraries was performed using the Chromium instrument and the Linked-Reads Genome Kit v2 (10x Genomics), according to the manufacturer's protocol (Rev A). Experimental details are elaborated in Supplemental Methods. Genomic aberrations per samples was visualized using Circos (Krzywinski et al. 2009).

Copy number analysis

Allele-specific copy number calls were generated using ASCAT v2.5.2 (Van Loo et al. 2010). The analyses are further detailed in the Supplemental Methods.

We also assessed our ability to infer ASCAVs having arisen pre- and post-copy number gains (**Figure S4**). By leveraging the WGS variant allele-frequency (VAF), and the total tumour copy number and sample purity estimates from ASCAT (ntot and ρ = 100%, respectively), we can compute the number of chromosome copies carrying a variant (mutation copy number, mcn) as $mcn = |ntot \times VAF/\rho|$. In turn, this can be used to "time" variants with respect to copy number gains: if a variant arose on a chromosome prior to its duplication, it will be duplicated as well ($mcn \geq 2$, an "early" variant). If it arose after, only one copy will be present ($mcn = 1$, a "late" variant). In regions with loss of heterozygosity or gains on both alleles, these two scenarios can be readily distinguished. Note that, in our case, early variants will include germline SNPs.

Personalized genome construction

Indels and SNVs (as generated by GATK/longranger) were used to construct personalized genomes with Crossstich. This procedure resulted in the generation of personalized reference sequences per haplotype in FASTA format as well as chain files that link reference genome to personalized genomes. To obtain chain files to link personalized genomes to the reference genome, we performed whole genome alignment between personalized genomes and reference genome using BLAT. The analyses are further detailed in Supplemental Methods.

Mutational signatures

We use the Bayesian approach SigFit (Gori and Baez-Ortega 2020) to estimate mutation signature exposures in our MM lines for those COSMIC v3 signatures which have previously been reported as active in melanoma (SBS1, 2, 5, 7a–d, 9, 13, 14, 36, 38 & 40). SigFit was run on all likely somatic variants, i.e. called variants not present in the Genome Aggregation Database (gnomAD, v3.0), as well as on the likely somatic ASCAVs.

ATAC-seq library preparation and sequencing

ATAC-seq data was generated using the OmniATAC-seq technique as described previously (Corces et al. 2017). See Supplemental Methods for further details.

ATAC-seq alignment to the reference and personalized genomes

We aimed at obtaining minimum 15M usable reads per sample, and eventually achieved 65M reads on average across 10 sequenced samples (**Table S2**). Paired-end reads were mapped to the human genome (hg38) and sample specific personalized genomes using Bowtie 2 with --very-sensitive option (v2.2.6). 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 chromosome M reads and filtering for Q>2 using SAMtools. Usable reads is defined as the number of reads retained after these filtering steps. We observed that the number of reads mapping to personalized genomes were slightly higher than the number of reads mapping to the reference genome (hg38), which has been reported previously for ChIP-seq data (Rozowsky et al. 2011; Mayba et al. 2014) (**Table S2**).

ATAC-seq peak calling

Peaks from ATAC-seq data was called for reference mapped and personalize genome mapped data using MACS2 (v2.1.2) (Zhang et al. 2008) using the parameters -q 0.05, --nomodel, --call-summits, --shift -75 --keep-dup all and --extsize 150. Summits for personalized genome mapped samples were lifted over to reference genome using liftOver. Summits were extended by 250bp up- and downstream using slopBed (BEDtools; v2.28.0), providing human chromosome sizes and filtered for blacklisted regions of the reference genome (ENCSR636HFF). Per sample, we obtained reference mapped peaks, HAP1 mapped peaks and HAP2 mapped peaks. To obtain a consolidated peak set per sample, we followed the strategy described by Corces et al (Corces et al. 2018), as elaborated in Supplemental Methods.

Identification of allele-specific events in ATAC-seq data

We have built a new allele specific variant detection pipeline using the backbone of alleleSeq pipeline (Chen et al. 2016). We used Bowtie 2 (Langmead and Salzberg 2012, 2) to map ATAC-seq reads to personalized genomes as described above. Next,

we marked duplicate reads in each alignment using Picard. Then, we evaluated two alignment files (ie. haplotype 1 mapped and haplotype 2 mapped) to identify the most likely origin of each read. Each read was evaluated iteratively using mapping quality (MapQ), CIGAR string and XM tag (which reports the number of mismatches in the alignment) If the read had the same mapQ, same CIGAR string (or the same number of Ms) and same XM tag for both alignments, it was marked as commonly mapping. This step resulted in four BAM files: haplotype1.exclusive, haplotype2.exclusive, haplotype1.common and haplotype2.common.

After identifying the source of each read, we filtered out duplicate reads. We also filtered out ambiguously mapping reads by evaluating the reads that map equally well to both haplotypes (ie. reads in haplotype1.common and haplotype2.common alignment files) in the reference genome. We lifted these positions over to hg38 coordinates, and discarded reads if it mapped to different locations. Next, we overlapped phased heterozygous variants obtained from whole genome sequence data with consolidated peak set (as described above). The variant positions that overlapped with the peaks were lifted over to haplotype 1 and 2 coordinates, and allele counts were obtained using *samtools mpileup* command with all four alignment files (allelic counts coming from common alignment files were compared and no major differences were found). Then allelic counts over heterozygous sites were merged, and variants that had at least 6 reads were further processed for allele specific accessibility analysis with BaalChIP (de Santiago et al. 2017) package in R/Bioconductor (R Core Team 2019). Count tables containing number of reference and alternative supporting reads per variant, together with allelic ratio of the same variant from whole genome sequence data was provided to runBayes command of BaalChIP, and allele-specific chromatin accessibility variants were identified. The remaining variants were defined as control variants.

Genomic annotations of both sets of variants were done using ChIPseeker (Yu et al. 2015) with UCSC hg38 knownGene table (TxDb.Hsapiens.UCSC.hg38.knownGene package in R/Bioconductor).

Motif enrichment analysis with allele-specific binding events

ASCA and control variants per sample were overlapped with consolidated ATAC-seq peaks. Peaks that had multiple variants were filtered out if the allelic bias between variants was inconsistent. Allelic counts were used to determine preferred allele (i.e. allele that has the highest ATAC-seq signal). Peak sequences were extracted from the FASTA sequence of the preferred allele (using fastaFromBed command from BEDtools (Quinlan and Hall 2010)) Reference sequence for each variant was extracted from hg38 using the same command. For each peak, we calculated the CRM score for the preferred allele and the other allele (reference sequence) using a set of 22,000 position weight matrices (Janky et al. 2014), and calculated a "delt a CRM score" for each peak and for each motif (**Figure 1D**). We evaluated the enrichment of CRM delta's in ATAC-

seq peaks using one-sided Fishers' exact test with a control set of 152,999 peaks containing non-ASCA variants. We performed enrichment analysis individually (per MM-line) and globally (across all MM-lines) (**Figure 2C**, **Figure S9**). Haplotype resolved ATAC-seq alignment figures were created with fluff (Georgiou and Heeringen 2016).

Identification of allele-specific expression variants

RNA-seq reads were mapped to the personalized genomes using Bowtie 2 (Langmead and Salzberg 2012, 2) with *--very-sensitive* option. We implemented the same post-processing steps as in the ATAC-seq analysis; this included choosing the best alignment between two mappings based on mapping quality and number of mismatches, as well as removal of ambiguously mapping or duplicate reads. Next, we overlaid phased heterozygous variants obtained from the whole genome sequencing data with the coding genome (hg38 CDS regions). Variant positions falling inside genes were lifted over to haplotype 1 and 2 coordinates, and allele counts were obtained using samtools mpileup. Then allelic counts over heterozygous sites were merged, and variants that had at least 10 reads were further processed for allele specific expression variant analysis. To assess allele specific expression in the presence of copy number changes, we used a beta-binomial model of the RNA-seq allele counts, informed by the WGS data. Briefly, for every variant, we obtain the posterior estimate Beta(1+#A, 1+#B) of the WGS B-allele frequency using a uniform Beta(1, 1) prior and a binomial likelihood to describe the WGS allelic read counts (#A and #B). This posterior BAF estimate is then used to perform a two-tailed beta-binomial test of the observed RNA-seq allele counts. Multiple testing correction was implemented with Benjamini & Hochberg method, and variants with FDR < 0.05 were reported as allele specific expression variants.

Identification of allele-specific variants in H3K27ac ChIP-seq data

ChIP-seq reads were mapped personalized genomes using Bowtie 2 (Langmead and Salzberg 2012, 2) with *--very-sensitive* option. We implemented the same pipeline as in ATAC-seq analysis for allele-specific variant detection.

cisTopic analysis

In order to train DeepMEL2 on, we used cisTopic (Bravo González-Blas et al. 2019) to obtain sets of co-accessible regions as in previous work where we trained DeepMEL (Minnoye et al. 2020). To be able to use cisTopic, single cells were simulated from bulk OmniATAC-seq data on the 30 human melanoma cell lines. Bootstrapping was used for the single cell simulation and 50 single cells were simulated for each melanoma line. Each single cell contains 50.000 random reads from its original bulk OmniATAC-seq data. Then, cisTopic was run on these simulated single cells (parameters: α = 50/T, β =

0.1, burn-in iterations = 500, recording iterations = 1000). The best model (47 topics) was selected according to log-likelihood value.

The DeepMEL2 neural network

In previous work we developed DeepMEL on 16 human ATAC-seq samples (Minnoye et al. 2020). Here, we developed an updated model, DeepMEL2, on 30 ATAC-seq samples. DeepMEL2 is, similarly to DeepMEL, a hybrid CNN-RNN deep learning enhancer classification model composed of convolutional, max pooling, time-distributed dense, bidirectional LSTM, and dense layers between input and output. The 128 convolutional filters of DeepMEL were initialized with random numbers, while in DeepMEL2 the 283 of 300 filters are populated with JASPAR position weight matrices that are clustered for five taxonomic groups (Fornes et al. 2020). Number of filters is increased from 128 to 300 and filter size is increased from 20 to 30 in order to populate convolutional filters with JASPAR motif collection. The detailed model architecture is shown in **Table S12.** DeepMEL2 is trained on melanoma-specific co-accessible region classes. It takes a 500bp DNA sequence and predicts an output vector corresponding to binarised 47 topics. To evaluate its performance auROC and auPR on training (%80), validation (%10) and test (%10) sets were calculated for each topic.

DeepMEL2+GABPA model was trained on 48 classes. On top of 47 topics, we added a 48th class where regions in input data were labeled as 1 if it overlaps with GABPA ChIP-seq peaks (ENCSR000BJK). The same architecture that was used to train DeepMEL2 to was used for DeepMEL2+GABPA

Scoring enhancers and ASCAVs with DeepMEL2

To score ASCAVs, we perturbed the 500bp ATAC-seq peaks by doing a single nucleotide change according to variants coming from two alleles. We calculated delta prediction score for each of the ASCAVs and the control variants for each of the classes. Then, we evaluated the delta prediction scores for each class to identify the fraction of explainable ASCAVs. We used a one-sided Fishers' exact test with a control set of non-ASCA variants at 5% false positive rate.

In order to compare different models, the maximum delta score for each variant was calculated.

Calculating the contribution of each nucleotide to the final output

We initialised DeepExplainer (Lundberg and Lee 2017) with randomly selected sequences (500) and calculated the importance scores of the sequence of interest with respect to any of the 47 classes. We multiplied this importance score by the one-hot encoded matrix of the sequence. Finally, we visualised the sequence by adjusting the nucleotide heights based on their importance score, similar to earlier work (Shrikumar, Greenside, and Kundaje 2019).

In silico saturation mutagenesis

For a 500bp sequence, we generated mutated sequences by changing each single nucleotide into the three other possible nucleotides. We scored the initial sequence without mutations, as well as all 1500 generated sequences with DeepMEL2 and calculated the delta prediction score for each class and for each mutation by comparing the final prediction relative to prediction for the initial sequence.

Luciferase assays

501 bp regions, surrounded by 20 bp flanking adaptors, were synthesized (TWIST Bioscience) and then individually cloned in a pGL4.23 plasmid. Luciferase activity in MM057 was measured using the Dual-luciferase Reporter Assay System (Promega). Experimental details are elaborated in the Supplemental Methods.

AP1 ChIP-seq library preparation and sequencing

The melanoma MM lines were grown to approximately 85% confluence, and a total of 20 million cells per ChIP sample was collected. ChIP samples were prepared following the 'Myers Lab ChIP-seq Protocol v011014', using the following antibodies at a concentration of 5 µg per ChIP: FOS (c-Fos; sc-166940 X, Santa Cruz Biotechnology), FOSL1 (Fra-1; sc-376148, Santa Cruz Biotechnology), JUN (c-Jun; sc-74543 X, Santa Cruz Biotechnology), JUNB (Jun-B; sc-8051 X, Santa Cruz Biotechnology). See Supplemental Methods for experimental details.

Analysis of AP1 ChIP-seq data

Sequence reads were mapped to human reference genome (hg38) using Bowtie 2 with --very-sensitive option (v2.2.6). Mapped reads were sorted using SAMtools (v1.8) and duplicates were removed using Picard MarkDuplicates (v1.134). Reads were filtered for mapping quality of 30 (MAPQ>30) using SAMtools. We implemented the same pipeline as in ATAC-seq analysis for allele-specific variant detection.

Publicly available data used in this work

RNA-seq data and H3K27ac ChIP-seq (data for A375, MM001, MM011, MM029, MM031, MM047, MM057, MM074, MM087 and MM099) was downloaded from GSE60666 (Verfaillie et al. 2015).

Data Access

All raw and processed sequencing data, except the whole genome 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 numbers GSE134432, GSE142238, and GSE159965. Genome sequencing data have been submitted to the

European Genome-phenome Archive (EGA; http://www.ebi.ac.uk/ega/) under accession number EGAS00001004136. The code for the analysis of the WGS data and detection of ASCAV is deposited at https://github.com/aertslab/AS_variant_pipeline. The DeepMEL2 and DeepMEL2_GABPA models are available from http://kipoi.org/models/DeepMEL/. The Jupyter notebooks to train DeepMEL and DeepMEL2 are available at https://github.com/aertslab/DeepMEL and the notebooks to train DeepMEL2 are provided as supplementary files (Supplemental Code).

Acknowledgements
This work was supported by an ERC Consolidator Grant to S.A. (no. 724226_cis-CONTROL), by the KU Leuven (grant no. C14/18/092 to S.A.), by the Foundation Against Cancer (grant no, 2016-070 to S.A.), a PhD and a postdoctoral fellowship from the FWO (L.M., no. 1S03317N, J.D. no. 12J6916N, respectively) and a postdoctoral research fellowship from Kom op tegen Kanker (Stand up to Cancer), the Flemish Cancer Society, and from Stichting tegen Kanker (Foundation against Cancer), the Belgian Cancer Society (Z.K.A and J.W.). Computing was performed at the Vlaams Supercomputer Center and high-throughput sequencing 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.

Competing interests
The authors declare no competing interests.

Author contributions
Z.K.A., I.I.T., J.W., and S.A. conceived the study. J.W., D.M., and V.C. performed the experimental work for the whole genome sequencing and ATAC-seq. G.E.G. established and provided the cell lines. Z.K.A., J.D., C.F., L.M., and G.H. performed the bioinformatics analyses. I.I.T. established the neural network and performed all analyses regarding the deep learning. Z.K.A., I.I.T., J.D., J.W., and S.A. wrote the manuscript.

Figure captions

Figure 1. Detection of allele-specific chromatin accessibility

A. Circos plot for sample MM074. Circos plots for the remaining samples are shown in Figure S1. **B.** Sankey diagram of the number of variants that went through our ASCAV discovery pipeline. **B.** Analysis pipeline for identification of allele-specific events from matched phased whole-genome data and functional genomics data (ATAC-seq, RNA-seq or ChIP-seq). **D.** Phased whole genome sequencing is applied to 10 melanoma cell lines, and used together with the reference genome to create personalized diploid genomes. Matched ATAC-seq, RNA-seq and ChIP-seq data (against H3K27ac mark and transcription factors) is used to detect allelic imbalance in chromatin accessibility (ASCA), gene expression (ASE), histone acetylation (ASHV) or allele-specific binding (ASB). By combining a melanoma-specific deep learning model (DeepMEL2), and motif discovery, *cis*-regulatory variants are predicted. **E**. Genome-wide allele-specific copy number is shown for sample MM074. Superposed are the identified ASCAVs in this cell line, of which the mutation copy number is plotted. The colour of the ASCAVs indicates whether they can be classified as either early or late. If their copy number context does not allow timing, they are labelled “na”. Allele-specific copy numbers for the remaining samples are shown in Figure S4 F. **F.** Concordant allele-specific events are detected around TYR, a gene encoding an enzyme involved in pigmentation. Inset shows the reads from whole-genome and ATAC-seq data for one of the allele-specific SNPs (rs1799989). Whole genome data indicates a haplotype 1 specific heterozygous SNP (i.e. GT=1|0) with variant allele frequency of 0.33, while ATAC-seq data indicates the reads are coming from one allele (haplotype 1). There are a further 6 allele-specific variants in *TYR* that are either haplotype 1 (i.e. GT = 1|0) or haplotype 2 (i.e. GT = 0|1) specific in the WGS data, yet all the variants manifest a haplotype specific activity in matched functional genomics data. The inset plots for all these 7 variants show ATAC-seq, H3K27ac ChIP-seq or RNA-seq reads in these loci segregated into haplotypes. Reads mapping exclusively to haplotype 1 are shown at the top (red), while the ones mapping exclusively to haplotype 2 are shown in the middle (blue). We can detect exclusive mapping only at variant locations, hence the majority of the reads map equally well to both haplotypes, and are shown at the bottom (green). Additionally, reference allele fractions (RAF) are shown for all the variants (corrected RAFs are obtained via BaalChIP for ASCAVs and ASHV).

Figure 2. Transcription factor motif enrichment on ASCAVs.

A. Selection of ASCAVs and control variants used to assess association between sequence content and allele-specific accessibility. **B.** Heatmap showing the clustering of all 719 ASCAV-enriched motifs into 47 families (color-coded margins). The 13 major families are labeled with their cognate TF on the diagonal. **C.** Scatter plot of motifs that are associated with chromatin accessibility. Each dot indicates a motif and is colored

based on the motif cluster they belong to. The x- and y-axes represent the delta cluster-buster motif score and the negative log-scaled FDR corrected p-value, respectively. **D.** Bar plot showing the number of ASCAVs explained by each motif cluster. For each family, the consensus motif is shown. **E.** Scatter plot of the average expression of AP-1 family members (JUN, JUNB, JUND, FOS, FOSB, FOSL1, FOSL2) and the fraction of ASCAVs that affect an AP-1 binding site. Correlation coefficient (Kendall's tau) and p-value are shown. **f.** Fractions of ASCAVs explained at different false positive rates are shown as curves for each MM line. Dashed lines represent the control for each MM line, where labels of ASCAVs and control variants are shuffled.

Figure 3. Cell state-aware DeepMEL2 can interpret ASCAVs.
A. Normalised cisTopic cell-topic heatmap of 30 melanoma cell lines showing general, state-specific, and cell line-specific sets of co-accessible regions. **B.** Schematic overview of DeepMEL2 highlighting improvements compared to DeepMEL. **C.** Scatter plot of auROC and auPR values shows the performance of DeepMEL2 on each topic. Promoter, state-specific, and cell line-specific topics are represented by red, blue, and green colors, respectively. **D.** Performance of DeepMEL2 and other models at predicting variant effects on *IRF4* enhancer activity. **E,F.** Curves indicate fractions of ASCAVs explained by Topic-17 score (MEL; **e**) and Topic-19 score (MES; **f**) at different false positive rates for each MM line. Bar chart insets show the exact fraction of the explained ASCAVs at 5% false positive rate. **G.** Bar charts showing the fraction of ASCAVs explained at 5% false positive rate for each MM line using either DeepMEL2, DeepMEL, DeepSEA, Basset, and PWM. The black bar represents the fraction when ASCAVs and control variants are shuffled.

Figure 4. Model explanation and experimental validation of three *cis*-regulatory variants.
A. C>T intronic SNP (rs2322683) in *SUMF1* is an ASCAV and AP-1 ASB (JUN and FOSL1 ChIP-seq datasets). The left-hand side plot shows haplotype 1, 2, and unphased reads (color-coded) from this locus in MM099 JUN and FOSL1 ChIP-seq and ATAC-seq reads. The right-hand side panel shows the same locus in three additional MM lines (MM011, MM047 and MM087) in which rs2322683 is also inferred as an ASCAV. WGS genotypes (GT) and BaalChIP allele ratios are shown in parentheses. **B.** DeepExplainer plot of the rs2322683 locus (position indicated with dashed lines), where the height of the nucleotides indicates their importance for the final prediction. Scoring using Topic-19 on both haplotypes shows C>T substitution generates an AP-1 binding site. *In silico* saturation mutagenesis on the reference sequence reveals the effect of each possible variant as a delta Topic-19 prediction score **C.** The curves represent the number of FOSL1 or JUN ASB variants found among the top-*n* MM099 ASCAVs ranked by the maximum delta prediction score of the different models. **D,E,F.** Each row

showcases: (**I**) an ASCAV and its allele-specific accessibility peak; (**II**) DeepExplainer and *in silico* mutagenesis results of the two haplotypes; (**III**) the DeepMEL2 score for both haplotypes; (**IV**) and the luciferase enhancer-reporter activity for both haplotypes. **D.** C>T intronic variant in *PEPD* is identified as an ASCAV and predicted to generate an AP-1 binding site, with an increase in MES enhancer score. The *in silico* mutagenesis plot shows that only a single mutation to T at position 269 increases the MES enhancer prediction significantly, and this is exactly the location of the ASCAV. **E.** C>T intronic variant in *MITF* is identified as an ASCAV and predicted to generate an AP-1 binding site. **F.** G>A intronic variant in *EVA1C* is identified as an ASCAV and predicted to generate a SOX10 binding site.

Figure 5. Analysis of *TERT* promoter mutations.
A. *TERT* promoter hotspot mutation in A375 is detected as an ASCAV as evidenced by ATAC-seq reads segregated into haplotypes (color-coded). In A375, haplotype 2 harbors the mutant allele T (according to WGS data, see **Figure S16**), and ATAC-seq evidences exclusive accessibility for this allele. The corrected reference ATAC-seq allele ratio is indicated in parentheses. **B.** Bar chart of model variant effect prediction performance on *TERT* promoter activity assessed by experimental saturation mutagenesis. **C.** Scatter plot showing the effect of each variant in the in vitro (x-axis) and in silico (y-axis) mutagenesis of the *TERT* promoter. The two hotspot gain-of-function mutations are highlighted. **D.** Scatter plot of delta Topic-14 score (promoter topic) vs. delta Topic-48 score (GABPA topic) of all ASCAVs from 10 MM lines calculated by using the DeepMEL2+GABPA model. ASCAVs are colored by their maximum delta prediction score. The *TERT* mutation of A375, and two newly predicted GABPA gains in MM047 and MM001 that are discussed in the text are encircled. **E.** DeepMEL2 prediction score for each topic for both haplotype 1 (red) and haplotype 2 (blue) of the A375 *TERT* locus is shown on the left, and the delta prediction scores between two haplotypes are shown on the right. The delta prediction scores for both Topic-14 (promoter topic) and Topic-48 (GABPA topic) are above the 0.05 detection threshold. **F, H.** Haplotype-specific DeepExplainer plots of the A375 TERT promoter locus by using Topic-14 (**F**) and Topic-48 (**H**), annotated with the corresponding TFs. **H.** Comparison of *in silico* (top, DeepMEL2 delta Topic-14 prediction scores) and *in vitro* (bottom, fold change in promoter activity) saturation mutagenesis assay. Each variant is color-coded.

References

Alexandrov, Ludmil B., Jaegil Kim, Nicholas J. Haradhvala, Mi Ni Huang, Alvin Wei Tian Ng, Yang Wu, Arnoud Boot, et al. 2020. “The Repertoire of Mutational Signatures in Human Cancer.” Nature 578 (7793): 94–101. https://doi.org/10.1038/s41586-020-1943-3.

Alipanahi, Babak, Andrew Delong, Matthew T. Weirauch, and Brendan J. Frey. 2015. “Predicting the Sequence Specificities of DNA- and RNA-Binding Proteins by Deep Learning.” Nature Biotechnology 33 (8): 831–38. https://doi.org/10.1038/nbt.3300.

Audano, Peter A., Arvis Sulovari, Tina A. Graves-Lindsay, Stuart Cantsilieris, Melanie Sorensen, AnneMarie E. Welch, Max L. Dougherty, et al. 2019. “Characterizing the Major Structural Variant Alleles of the Human Genome.” Cell 176 (3): 663-675.e19. https://doi.org/10.1016/j.cell.2018.12.019.

Avsec, Žiga, Roman Kreuzhuber, Johnny Israeli, Nancy Xu, Jun Cheng, Avanti Shrikumar, Abhimanyu Banerjee, et al. 2018. “Kipoi: Accelerating the Community Exchange and Reuse of Predictive Models for Genomics.” Preprint. Bioinformatics. https://doi.org/10.1101/375345.

Avsec, Žiga, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, et al. 2021. “Base-Resolution Models of Transcription-Factor Binding Reveal Soft Motif Syntax.” Nature Genetics 53 (3): 354–66. https://doi.org/10.1038/s41588-021-00782-6.

Bailey, Matthew H., Collin Tokheim, Eduard Porta-Pardo, Sohini Sengupta, Denis Bertrand, Amila Weerasinghe, Antonio Colaprico, et al. 2018. “Comprehensive Characterization of Cancer Driver Genes and Mutations.” Cell 173 (2): 371-385.e18. https://doi.org/10.1016/j.cell.2018.02.060.

Banovich, Nicholas E., Yang I. Li, Anil Raj, Michelle C. Ward, Peyton Greenside, Diego Calderon, Po Yuan Tung, et al. 2018. “Impact of Regulatory Variation across Human IPSCs and Differentiated Cells.” Genome Research 28 (1): 122–31. https://doi.org/10.1101/gr.224436.117.

Baron, Maayan, Mohita Tagore, Miranda V. Hunter, Isabella S. Kim, Reuben Moncada, Yun Yan, Nathaniel R. Campbell, Richard M. White, and Itai Yanai. 2020. “The Stress-Like Cancer Cell State Is a Consistent Component of Tumorigenesis.” Cell Systems 11 (5): 536-546.e7. https://doi.org/10.1016/j.cels.2020.08.018.

Bell, Robert J.A., H. Tomas Rube, Alex Kreig, Andrew Mancini, Shaun D. Fouse, Raman P. Nagarajan, Serah Choi, et al. 2015. “The Transcription Factor GABP Selectively Binds and Activates the Mutant TERT Promoter in Cancer.” Science (New York, N.Y.) 348 (6238): 1036–39. https://doi.org/10.1126/science.aab0015.

Biddie, Simon C., Sam John, Pete J. Sabo, Robert E. Thurman, Thomas A. Johnson, R. Louis Schiltz, Tina B. Miranda, et al. 2011. “Transcription Factor AP1 Potentiates Chromatin Accessibility and Glucocorticoid Receptor Binding.” Molecular Cell 43 (1): 145–55. https://doi.org/10.1016/j.molcel.2011.06.016.

Bravo González-Blas, Carmen, Liesbeth Minnoye, Dafni Papasokrati, Sara Aibar, Gert Hulselmans, Valerie Christiaens, Kristofer Davie, Jasper Wouters, and Stein Aerts. 2019. “CisTopic: Cis-Regulatory Topic Modeling on Single-Cell ATAC-Seq Data.” Nature Methods 16 (5): 397–400. https://doi.org/10.1038/s41592-019-0367-1.

Caramel, Julie, Eftychios Papadogeorgakis, Louise Hill, Gareth J. Browne, Geoffrey Richard, Anne Wierinckx, Gerald Saldanha, et al. 2013. “A Switch in the Expression of Embryonic EMT-Inducers Drives the Development of Malignant Melanoma.” Cancer Cell 24 (4): 466–80. https://doi.org/10.1016/j.ccr.2013.08.018.

Castel, Stephane E., Ami Levy-Moonshine, Pejman Mohammadi, Eric Banks, and Tuuli Lappalainen. 2015. “Tools and Best Practices for Data Processing in Allelic Expression Analysis.” Genome Biology 16 (September): 195. https://doi.org/10.1186/s13059-015-0762-6.

Chen, Jieming, Joel Rozowsky, Timur R. Galeev, Arif Harmanci, Robert Kitchen, Jason Bedford, Alexej Abyzov, Yong Kong, Lynne Regan, and Mark Gerstein. 2016. “A Uniform Survey of Allele-

Specific Binding and Expression over 1000-Genomes-Project Individuals." Nature Communications 7 (1): 1–13. https://doi.org/10.1038/ncomms11101.

Consortium, The GTEx. 2015. "The Genotype-Tissue Expression (GTEx) Pilot Analysis: Multitissue Gene Regulation in Humans." Science 348 (6235): 648–60. https://doi.org/10.1126/science.1262110.

Corces, M. Ryan, Jeffrey M. Granja, Shadi Shams, Bryan H. Louie, Jose A. Seoane, Wanding Zhou, Tiago C. Silva, et al. 2018. "The Chromatin Accessibility Landscape of Primary Human Cancers." Science 362 (6413). https://doi.org/10.1126/science.aav1898.

Corces, M. Ryan, Alexandro E. Trevino, Emily G. Hamilton, Peyton G. Greenside, Nicholas A. Sinnott-Armstrong, Sam Vesuna, Ansuman T. Satpathy, et al. 2017. "An Improved ATAC-Seq Protocol Reduces Background and Enables Interrogation of Frozen Tissues." Nature Methods 14 (10): 959–62. https://doi.org/10.1038/nmeth.4396.

Degner, Jacob F., Athma A. Pai, Roger Pique-Regi, Jean-Baptiste Veyrieras, Daniel J. Gaffney, Joseph K. Pickrell, Sherryl De Leon, et al. 2012. "DNase I Sensitivity QTLs Are a Major Determinant of Human Expression Variation." Nature 482 (7385): 390–94. https://doi.org/10.1038/nature10808.

Denecker, G., N. Vandamme, O. Akay, D. Koludrovic, J. Taminau, K. Lemeire, A. Gheldof, et al. 2014. "Identification of a ZEB2-MITF-ZEB1 Transcriptional Network That Controls Melanogenesis and Melanoma Progression." Cell Death and Differentiation 21 (8): 1250–61. https://doi.org/10.1038/cdd.2014.44.

Dentro, Stefan C., Ignaty Leshchiner, Kerstin Haase, Maxime Tarabichi, Jeff Wintersinger, Amit G. Deshwar, Kaixian Yu, et al. 2020. "Characterizing Genetic Intra-Tumor Heterogeneity across 2,658 Human Cancer Genomes." BioRxiv, April, 312041. https://doi.org/10.1101/312041.

Deplancke, Bart, Daniel Alpern, and Vincent Gardeux. 2016. "The Genetics of Transcription Factor DNA Binding Variation." Cell 166 (3): 538–54. https://doi.org/10.1016/j.cell.2016.07.012.

Do, Catherine, Alyssa Shearer, Masako Suzuki, Mary Beth Terry, Joel Gelernter, John M. Greally, and Benjamin Tycko. 2017. "Genetic–Epigenetic Interactions in Cis: A Major Focus in the Post-GWAS Era." Genome Biology 18 (1): 120. https://doi.org/10.1186/s13059-017-1250-y.

ENCODE Project Consortium. 2012. "An Integrated Encyclopedia of DNA Elements in the Human Genome." Nature 489 (7414): 57–74. https://doi.org/10.1038/nature11247.

Feigin, Michael E., Tyler Garvin, Peter Bailey, Nicola Waddell, David K. Chang, David R. Kelley, Shimin Shuai, et al. 2017. "Recurrent Noncoding Regulatory Mutations in Pancreatic Ductal Adenocarcinoma." Nature Genetics 49 (6): 825–33. https://doi.org/10.1038/ng.3861.

Fornes, Oriol, Jaime A Castro-Mondragon, Aziz Khan, Robin van der Lee, Xi Zhang, Phillip A Richmond, Bhavi P Modi, et al. 2020. "JASPAR 2020: Update of the Open-Access Database of Transcription Factor Binding Profiles." Nucleic Acids Research 48 (D1): D87–92. https://doi.org/10.1093/nar/gkz1001.

Fu, Yao, Zhu Liu, Shaoke Lou, Jason Bedford, Xinmeng Mu, Kevin Y. Yip, Ekta Khurana, and Mark Gerstein. 2014. "FunSeq2: A Framework for Prioritizing Noncoding Regulatory Variants in Cancer." Genome Biology 15 (10): 480. https://doi.org/10.1186/s13059-014-0480-5.

Gaffney, Daniel J., Jean-Baptiste Veyrieras, Jacob F. Degner, Roger Pique-Regi, Athma A. Pai, Gregory E. Crawford, Matthew Stephens, Yoav Gilad, and Jonathan K. Pritchard. 2012. "Dissecting the Regulatory Architecture of Gene Expression QTLs." Genome Biology 13 (1): R7. https://doi.org/10.1186/gb-2012-13-1-r7.

Gasperini, Molly, Jacob M. Tome, and Jay Shendure. 2020. "Towards a Comprehensive Catalogue of Validated and Target-Linked Human Enhancers." Nature Reviews. Genetics 21 (5): 292–310. https://doi.org/10.1038/s41576-019-0209-0.

Gembarska, Agnieszka, Flavie Luciani, Clare Fedele, Elisabeth A. Russell, Michael Dewaele, Stéphanie Villar, Aleksandra Zwolinska, et al. 2012. "MDM4 Is a Key Therapeutic Target in Cutaneous Melanoma." Nature Medicine 18 (8): 1239–47. https://doi.org/10.1038/nm.2863.

Georgiou, Georgios, and Simon J. van Heeringen. 2016. "Fluff: Exploratory Analysis and Visualization of

High-Throughput Sequencing Data." PeerJ 4 (July): e2209. https://doi.org/10.7717/peerj.2209.
Gerstung, Moritz, Clemency Jolly, Ignaty Leshchiner, Stefan C. Dentro, Santiago Gonzalez, Daniel Rosebrock, Thomas J. Mitchell, et al. 2020. "The Evolutionary History of 2,658 Cancers." Nature 578 (7793): 122–28. https://doi.org/10.1038/s41586-019-1907-7.

Ghandi, Mahmoud, Dongwon Lee, Morteza Mohammad-Noori, and Michael A. Beer. 2014. "Enhanced Regulatory Sequence Prediction Using Gapped K-Mer Features." PLOS Computational Biology 10 (7): e1003711. https://doi.org/10.1371/journal.pcbi.1003711.
Hindorff, Lucia A., Praveen Sethupathy, Heather A. Junkins, Erin M. Ramos, Jayashri P. Mehta, Francis S. Collins, and Teri A. Manolio. 2009. "Potential Etiologic and Functional Implications of Genome-Wide Association Loci for Human Diseases and Traits." Proceedings of the National Academy of Sciences 106 (23): 9362–67. https://doi.org/10.1073/pnas.0903103106.
Hoffman, Gabriel E., Jaroslav Bendl, Kiran Girdhar, Eric E. Schadt, and Panos Roussos. 2019. "Functional Interpretation of Genetic Variants Using Deep Learning Predicts Impact on Chromatin Accessibility and Histone Modification." Nucleic Acids Research 47 (20): 10597–611. https://doi.org/10.1093/nar/gkz808.
Horn, Susanne, Adina Figl, P. Sivaramakrishna Rachakonda, Christine Fischer, Antje Sucker, Andreas Gast, Stephanie Kadel, et al. 2013. "TERT Promoter Mutations in Familial and Sporadic Melanoma." Science (New York, N.Y.) 339 (6122): 959–61. https://doi.org/10.1126/science.1230062.
Huang, Franklin W., Eran Hodis, Mary Jue Xu, Gregory V. Kryukov, Lynda Chin, and Levi A. Garraway. 2013. "Highly Recurrent TERT Promoter Mutations in Human Melanoma." Science (New York, N.Y.) 339 (6122): 957–59. https://doi.org/10.1126/science.1229259.
Inukai, Sachi, Kian Hong Kock, and Martha L. Bulyk. 2017. "Transcription Factor-DNA Binding: Beyond Binding Site Motifs." Current Opinion in Genetics & Development 43 (April): 110–19. https://doi.org/10.1016/j.gde.2017.02.007.
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.
Jacobs, Jelle, Mardelle Atkins, Kristofer Davie, Hana Imrichova, Lucia Romanelli, Valerie Christiaens, Gert Hulselmans, et al. 2018. "The Transcription Factor Grainy Head Primes Epithelial Enhancers for Spatiotemporal Activation by Displacing Nucleosomes." Nature Genetics 50 (7): 1011–20. https://doi.org/10.1038/s41588-018-0140-x.
Janky, Rekin's, Annelien Verfaillie, Hana Imrichová, Bram Van de Sande, Laura Standaert, Valerie Christiaens, Gert Hulselmans, et al. 2014. "IRegulon: From a Gene List to a Gene Regulatory Network Using Large Motif and Track Collections." PLoS Computational Biology 10 (7): e1003731. https://doi.org/10.1371/journal.pcbi.1003731.
Kelley, David R., Jasper Snoek, and John Rinn. 2016. "Basset: Learning the Regulatory Code of the Accessible Genome with Deep Convolutional Neural Networks." Genome Research, May, gr.200535.115. https://doi.org/10.1101/gr.200535.115.
Khurana, Ekta, Yao Fu, Vincenza Colonna, Xinmeng Jasmine Mu, Hyun Min Kang, Tuuli Lappalainen, Andrea Sboner, et al. 2013. "Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics." Science (New York, N.Y.) 342 (6154): 1235587. https://doi.org/10.1126/science.1235587.
Kilpinen, Helena, Sebastian M. Waszak, Andreas R. Gschwind, Sunil K. Raghav, Robert M. Witwicki, Andrea Orioli, Eugenia Migliavacca, et al. 2013. "Coordinated Effects of Sequence Variation on DNA Binding, Chromatin Structure, and Transcription." Science (New York, N.Y.) 342 (6159): 744–47. https://doi.org/10.1126/science.1242463.
Kircher, Martin, Chenling Xiong, Beth Martin, Max Schubach, Fumitaka Inoue, Robert J. A. Bell, Joseph

F. Costello, Jay Shendure, and Nadav Ahituv. 2019. “Saturation Mutagenesis of Twenty Disease-Associated Regulatory Elements at Single Base-Pair Resolution.” Nature Communications 10 (1): 1–15. https://doi.org/10.1038/s41467-019-11526-w.

Krzywinski M, Schein J, Birol İ, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: An information aesthetic for comparative genomics. Genome Res 19: 1639–1645.

Kukurba, Kimberly R., Rui Zhang, Xin Li, Kevin S. Smith, David A. Knowles, Meng How Tan, Robert Piskol, et al. 2014. “Allelic Expression of Deleterious Protein-Coding Variants across Human Tissues.” PLOS Genetics 10 (5): e1004304. https://doi.org/10.1371/journal.pgen.1004304.

Kumasaka, Natsuhiko, Andrew J. Knights, and Daniel J. Gaffney. 2016. “Fine-Mapping Cellular QTLs with RASQUAL and ATAC-Seq.” Nature Genetics 48 (2): 206–13. https://doi.org/10.1038/ng.3467.

Landi, Maria Teresa, D. Timothy Bishop, Stuart MacGregor, Mitchell J. Machiela, Alexander J. Stratigos, Paola Ghiorzo, Myriam Brossard, et al. 2020. “Genome-Wide Association Meta-Analyses Combining Multiple Risk Phenotypes Provide Insights into the Genetic Architecture of Cutaneous Melanoma Susceptibility.” Nature Genetics 52 (5): 494–504. https://doi.org/10.1038/s41588-020-0611-8.

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” Nature Methods 9 (4): 357–59. https://doi.org/10.1038/nmeth.1923.

Lee, Dongwon. 2016. “LS-GKM: A New Gkm-SVM for Large-Scale Datasets.” Bioinformatics 32 (14): 2196–98. https://doi.org/10.1093/bioinformatics/btw142.

Liu, Qiao, Fei Xia, Qijin Yin, and Rui Jiang. 2018. “Chromatin Accessibility Prediction via a Hybrid Deep Convolutional Neural Network.” Bioinformatics (Oxford, England) 34 (5): 732–38. https://doi.org/10.1093/bioinformatics/btx679.

Lundberg, Scott M, and Su-In Lee. 2017. “A Unified Approach to Interpreting Model Predictions.” In Advances in Neural Information Processing Systems 30, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 4765–74. Curran Associates, Inc. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Maurano, Matthew T., Eric Haugen, Richard Sandstrom, Jeff Vierstra, Anthony Shafer, Rajinder Kaul, and John A. Stamatoyannopoulos. 2015. “Large-Scale Identification of Sequence Variants Influencing Human Transcription Factor Occupancy in Vivo.” Nature Genetics 47 (12): 1393–1401. https://doi.org/10.1038/ng.3432.

Maurano, Matthew T., Richard Humbert, Eric Rynes, Robert E. Thurman, Eric Haugen, Hao Wang, Alex P. Reynolds, et al. 2012. “Systematic Localization of Common Disease-Associated Variation in Regulatory DNA.” Science (New York, N.Y.) 337 (6099): 1190–95. https://doi.org/10.1126/science.1222794.

Mayba, Oleg, Houston N. Gilbert, Jinfeng Liu, Peter M. Haverty, Suchit Jhunjhunwala, Zhaoshi Jiang, Colin Watanabe, and Zemin Zhang. 2014. “MBASED: Allele-Specific Expression Detection in Cancer Tissues and Cell Lines.” Genome Biology 15: 405. https://doi.org/10.1186/s13059-014-0405-3.

McVicker, Graham, Bryce van de Geijn, Jacob F. Degner, Carolyn E. Cain, Nicholas E. Banovich, Anil Raj, Noah Lewellen, Marsha Myrthil, Yoav Gilad, and Jonathan K. Pritchard. 2013. “Identification of Genetic Variants That Affect Histone Modifications in Human Cells.” Science 342 (6159): 747–49. https://doi.org/10.1126/science.1242429.

Melton, Collin, Jason A. Reuter, Damek V. Spacek, and Michael Snyder. 2015. “Recurrent Somatic Mutations in Regulatory Regions of Human Cancer Genomes.” Nature Genetics 47 (7): 710–16. https://doi.org/10.1038/ng.3332.

Minnoye, Liesbeth, Ibrahim Ihsan Taskiran, David Mauduit, Maurizio Fazio, Linde Van Aerschot, Gert Hulselmans, Valerie Christiaens, et al. 2020. “Cross-Species Analysis of Enhancer Logic Using Deep Learning.” Genome Research, July, gr.260844.120. https://doi.org/10.1101/gr.260844.120.

Peinado, Héctor, David Olmeda, and Amparo Cano. 2007. “Snail, Zeb and BHLH Factors in Tumour

Progression: An Alliance against the Epithelial Phenotype?" Nature Reviews. Cancer 7 (6): 415–28. https://doi.org/10.1038/nrc2131.

Postigo, A. A., and D. C. Dean. 1999. "ZEB Represses Transcription through Interaction with the Corepressor CtBP." Proceedings of the National Academy of Sciences of the United States of America 96 (12): 6683–88. https://doi.org/10.1073/pnas.96.12.6683.

Postigo, A. A., E. Ward, J. B. Skeath, and D. C. Dean. 1999. "Zfh-1, the Drosophila Homologue of ZEB, Is a Transcriptional Repressor That Regulates Somatic Myogenesis." Molecular and Cellular Biology 19 (10): 7255–63. https://doi.org/10.1128/mcb.19.10.7255.

Quang, Daniel, and Xiaohui Xie. 2019. "FactorNet: A Deep Learning Framework for Predicting Cell Type Specific Transcription Factor Binding from Nucleotide-Resolution Sequential Data." Methods, Deep Learning in Bioinformatics, 166 (August): 40–47. https://doi.org/10.1016/j.ymeth.2019.03.020.

Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." Bioinformatics (Oxford, England) 26 (6): 841–42. https://doi.org/10.1093/bioinformatics/btq033.

R Core Team. 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org/.

Rozowsky, Joel, Alexej Abyzov, Jing Wang, Pedro Alves, Debasish Raha, Arif Harmanci, Jing Leng, et al. 2011. "AlleleSeq: Analysis of Allele-Specific Expression and Binding in a Network Framework." Molecular Systems Biology 7 (1): 522. https://doi.org/10.1038/msb.2011.54.

Santiago, Ines de, Wei Liu, Ke Yuan, Martin O'Reilly, Chandra Sekhar Reddy Chilamakuri, Bruce A. J. Ponder, Kerstin B. Meyer, and Florian Markowetz. 2017. "BaalChIP: Bayesian Analysis of Allele-Specific Transcription Factor Binding in Cancer Genomes." Genome Biology 18 (1): 39. https://doi.org/10.1186/s13059-017-1165-7.

Satpathy, Ansuman T., Jeffrey M. Granja, Kathryn E. Yost, Yanyan Qi, Francesca Meschi, Geoffrey P. McDermott, Brett N. Olsen, et al. 2019. "Massively Parallel Single-Cell Chromatin Landscapes of Human Immune Cell Development and Intratumoral T Cell Exhaustion." Nature Biotechnology 37 (8): 925–36. https://doi.org/10.1038/s41587-019-0206-z.

Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje. 2019. "Learning Important Features Through Propagating Activation Differences." ArXiv:1704.02685 [Cs], October. http://arxiv.org/abs/1704.02685.

Stevenson, Kraig R., Joseph D. Coolon, and Patricia J. Wittkopp. 2013. "Sources of Bias in Measures of Allele-Specific Expression Derived from RNA-sequence Data Aligned to a Single Reference Genome." BMC Genomics 14 (August): 536. https://doi.org/10.1186/1471-2164-14-536.

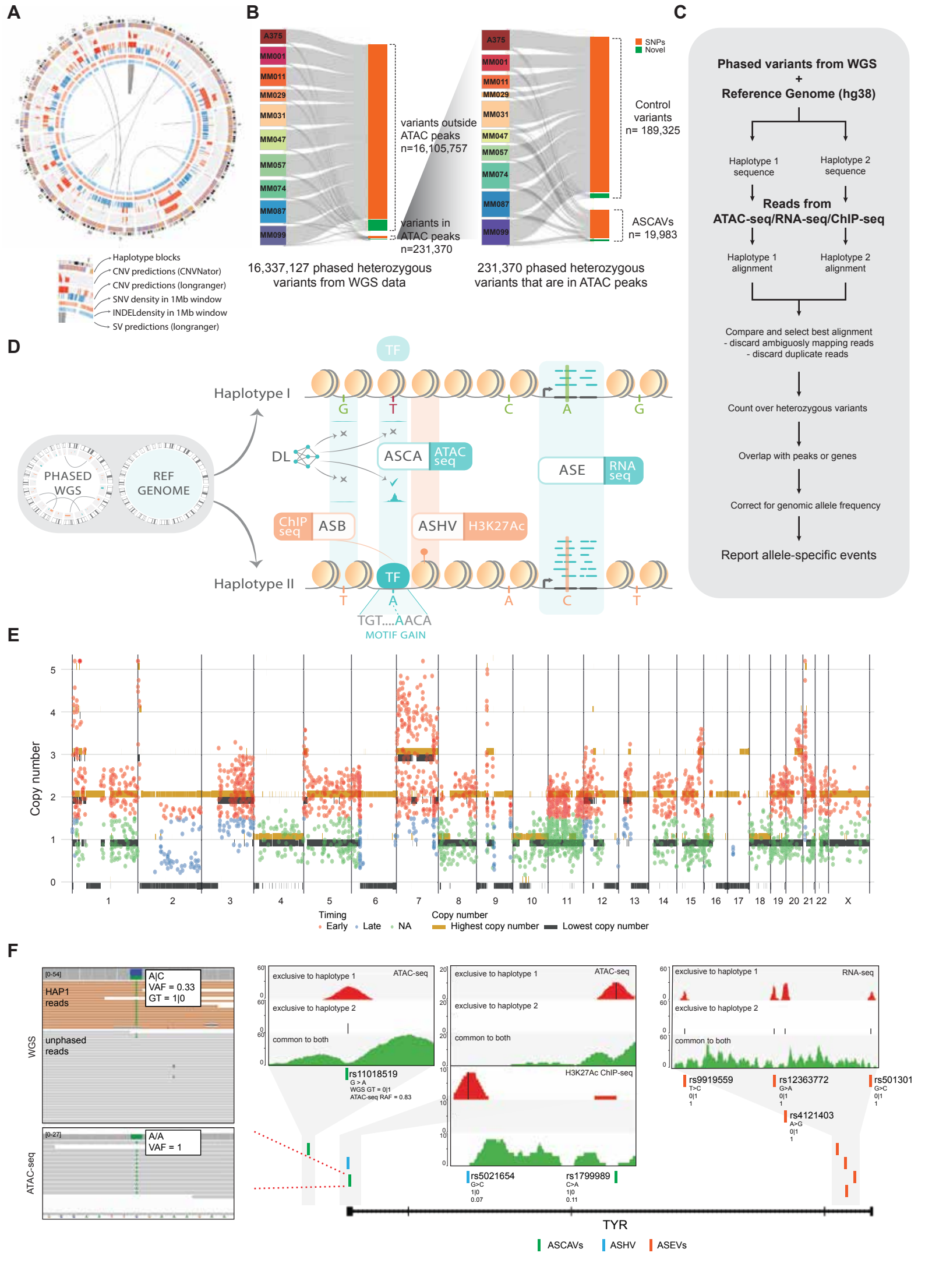
Svetlichnyy, Dmitry, Hana Imrichova, Mark Fiers, Zeynep Kalender Atak, and Stein Aerts. 2015. "Identification of High-Impact Cis -Regulatory Mutations Using Transcription Factor Specific Random Forest Models." PLOS Comput Biol 11 (11): e1004590. https://doi.org/10.1371/journal.pcbi.1004590.

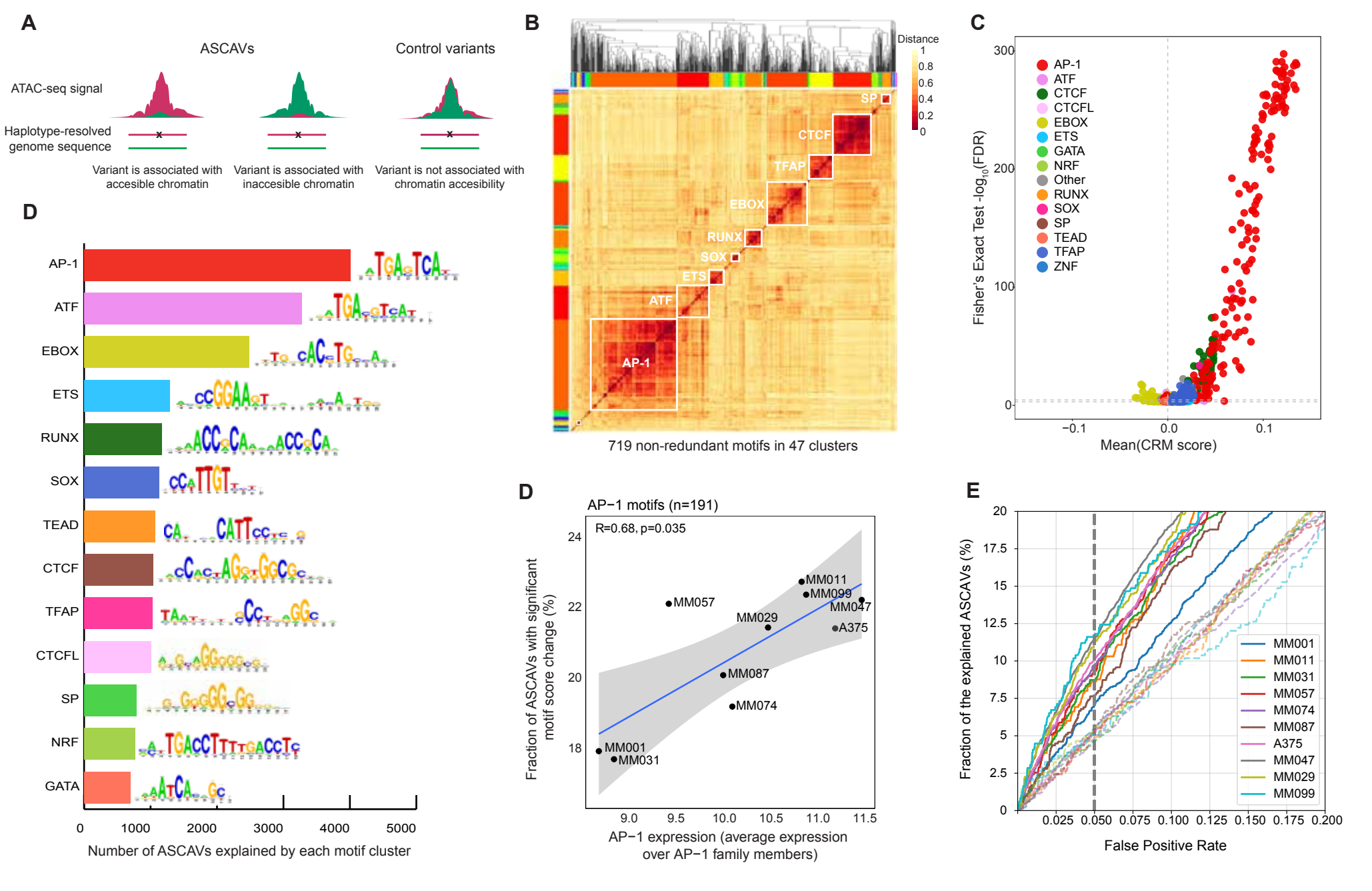
Tehranchi, Ashley, Brian Hie, Michael Dacre, Irene Kaplow, Kade Pettie, Peter Combs, and Hunter B Fraser. 2019. "Fine-Mapping Cis-Regulatory Variants in Diverse Human Populations." ELife 8 (January): e39595. https://doi.org/10.7554/eLife.39595.

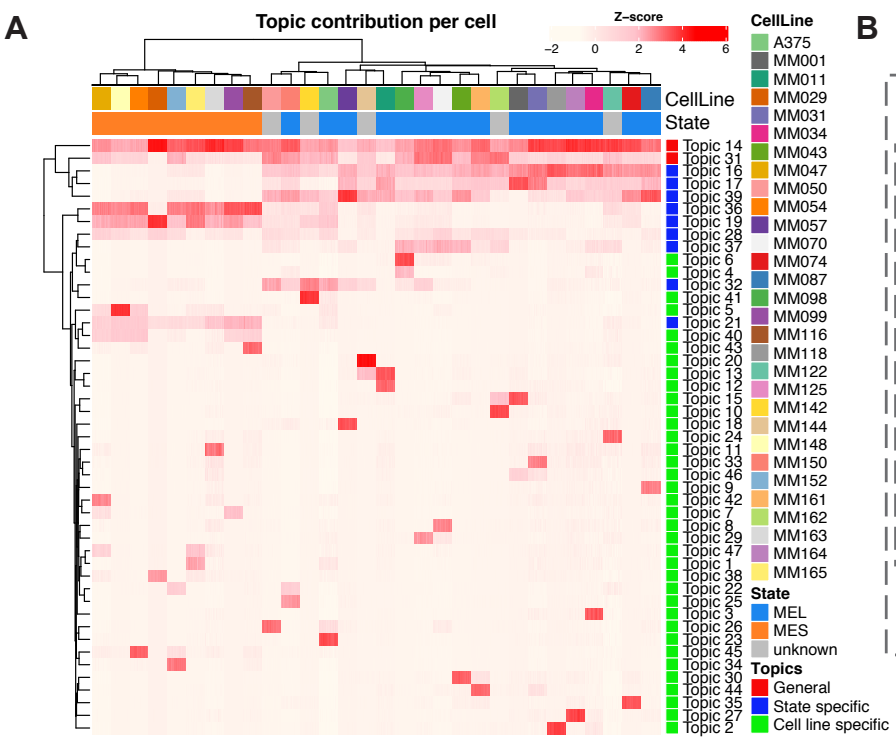
Tuskan, Robert G., Shirley Tsang, Zhonghe Sun, Jessica Baer, Ester Rozenblum, Xiaolin Wu, David J. Munroe, and Karlyne M. Reilly. 2008. "Real-Time PCR Analysis of Candidate Imprinted Genes on Mouse Chromosome 11 Shows Balanced Expression from the Maternal and Paternal Chromosomes and Strain-Specific Variation in Expression Levels." Epigenetics 3 (1): 43–50. https://doi.org/10.4161/epi.3.1.5469.

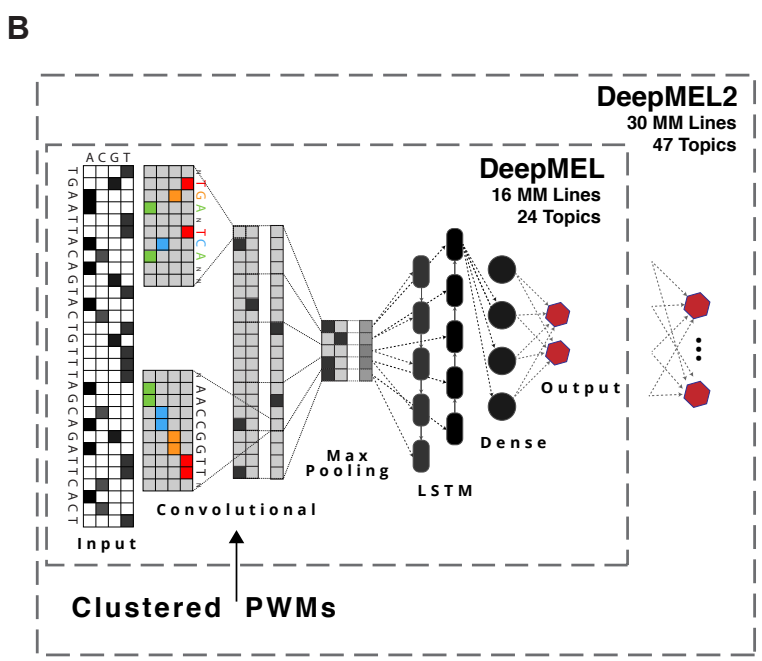
Van Loo, Peter, Silje H. Nordgard, Ole Christian Lingjærde, Hege G. Russnes, Inga H. Rye, Wei Sun, Victor J. Weigman, et al. 2010. "Allele-Specific Copy Number Analysis of Tumors." Proceedings of the National Academy of Sciences of the United States of America 107 (39): 16910–15.

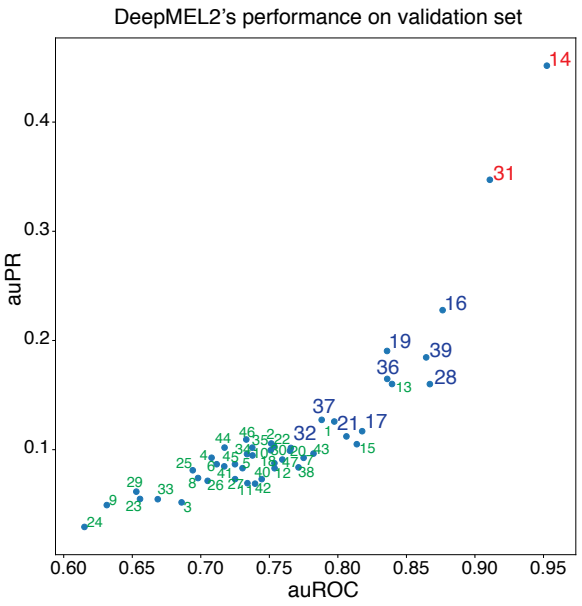
https://doi.org/10.1073/pnas.1009843107.
Verfaillie, Annelien, Hana Imrichova, Zeynep Kalender Atak, Michael Dewaele, Florian Rambow, Gert Hulselmans, Valerie Christiaens, et al. 2015. “Decoding the Regulatory Landscape of Melanoma Reveals TEADS as Regulators of the Invasive Cell State.” Nature Communications 6 (April). https://doi.org/10.1038/ncomms7683.
Vinagre, João, Ana Almeida, Helena Pópulo, Rui Batista, Joana Lyra, Vasco Pinto, Ricardo Coelho, et al. 2013. “Frequency of TERT Promoter Mutations in Human Cancers.” Nature Communications 4 (1): 1–6. https://doi.org/10.1038/ncomms3185.
Wasserman, Wyeth W., and Albin Sandelin. 2004. “Applied Bioinformatics for the Identification of Regulatory Elements.” Nature Reviews Genetics 5 (4): 276–87. https://doi.org/10.1038/nrg1315.
Waszak, Sebastian M., Olivier Delaneau, Andreas R. Gschwind, Helena Kilpinen, Sunil K. Raghav, Robert M. Witwicki, Andrea Orioli, et al. 2015. “Population Variation and Genetic Control of Modular Chromatin Architecture in Humans.” Cell 162 (5): 1039–50. https://doi.org/10.1016/j.cell.2015.08.001.
Wouters, Jasper, Zeynep Kalender-Atak, Liesbeth Minnoye, Katina I. Spanier, Maxime De Waegeneer, Carmen Bravo González-Blas, David Mauduit, et al. 2020. “Robust Gene Expression Programs Underlie Recurrent Cell States and Phenotype Switching in Melanoma.” Nature Cell Biology 22 (8): 986–98. https://doi.org/10.1038/s41556-020-0547-3.
Yu G, Wang L-G, He Q-Y. 2015. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics 31: 2382–2383.
Zhang, Wei, Ana Bojorquez-Gomez, Daniel Ortiz Velez, Guorong Xu, Kyle S. Sanchez, John Paul Shen, Kevin Chen, et al. 2018. “A Global Transcriptional Network Connecting Noncoding Mutations to Changes in Tumor Gene Expression.” Nature Genetics 50 (4): 613–20. https://doi.org/10.1038/s41588-018-0091-2.
Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based Analysis of ChIP-Seq (MACS). Genome Biology 9: R137.
Zhou, Jian, Christopher Y. Park, Chandra L. Theesfeld, Aaron K. Wong, Yuan Yuan, Claudia Scheckel, John J. Fak, et al. 2019. “Whole-Genome Deep-Learning Analysis Identifies Contribution of Noncoding Mutations to Autism Risk.” Nature Genetics 51 (6): 973–80. https://doi.org/10.1038/s41588-019-0420-0.
Zhou, Jian, and Olga G. Troyanskaya. 2015. “Predicting Effects of Noncoding Variants with Deep Learning-Based Sequence Model.” Nature Methods 12 (10): 931–34. https://doi.org/10.1038/nmeth.3547.
Zhu, Helen, Liis Uusküla-Reimand, Keren Isaev, Lina Wadi, Azad Alizada, Shimin Shuai, Vincent Huang, et al. 2020. “Candidate Cancer Driver Mutations in Distal Regulatory Elements and Long-Range Chromatin Interaction Networks.” Molecular Cell 0 (0). https://doi.org/10.1016/j.molcel.2019.12.027.

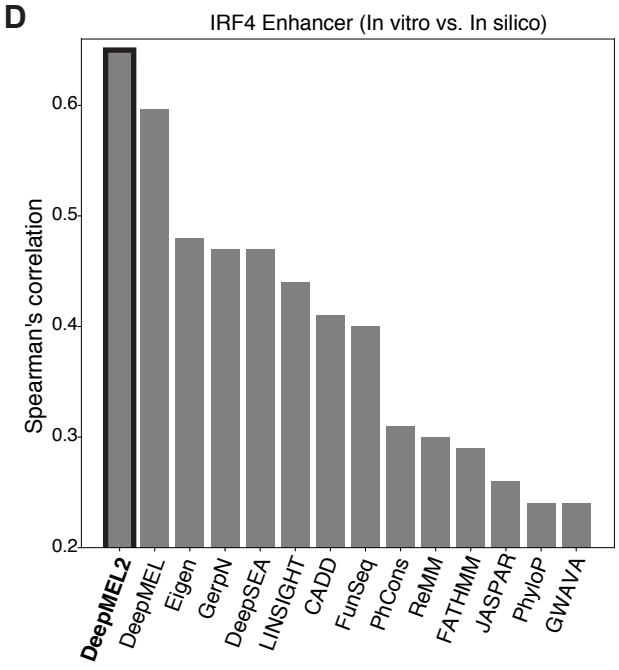

A
B
C
D
E
F
Haplotype blocks
CNV predictions (CNVNator)
CNV predictions (longranger)
SNV density in 1Mb window
INDELdensity in 1Mb window
SV predictions (longranger)
A375
MM001
MM011
MM029
MM031
MM047
MM057
MM074
MM087
MM099
variants outside ATAC peaks n=16,105,757
variants in ATAC peaks n=231,370
SNPs
Novel
Control variants n= 189,325
ASCAVs n= 19,983
16,337,127 phased heterozygous variants from WGS data
231,370 phased heterozygous variants that are in ATAC peaks
Phased variants from WGS
+
Reference Genome (hg38)
Haplotype 1 sequence
Haplotype 2 sequence
Reads from ATAC-seq/RNA-seq/ChIP-seq
Haplotype 1 alignment
Haplotype 2 alignment
Compare and select best alignment
- discard ambiguously mapping reads
- discard duplicate reads
Count over heterozygous variants
Overlap with peaks or genes
Correct for genomic allele frequency
Report allele-specific events
PHASED WGS
REF GENOME
Haplotype I
Haplotype II
TF
DL
ASCA
ATAC seq
ASE
RNA seq
ChIP seq
ASB
ASHV
H3K27Ac
TGT....AACA
MOTIF GAIN
Copy number
Timing
Early
Late
NA
Copy number
Highest copy number
Lowest copy number
WGS
ATAC-seq
HAP1 reads
unphased reads
A|C
VAF = 0.33
GT = 1|0
A/A
VAF = 1
exclusive to haplotype 1
exclusive to haplotype 2
common to both
ATAC-seq
RNA-seq
H3K27Ac ChIP-seq
rs11018519
G > A
WGS GT = 0|1
ATAC-seq RAF = 0.83
rs5021654
G>C
1|0
0.07
rs1799989
C>A
1|0
0.11
rs9919559
T>C
0|1
1
rs12363772
G>A
0|1
1
rs4121403
A>G
0|1
1
rs501301
G>C
0|1
1
TYR
ASCAVs
ASHV
ASEVs

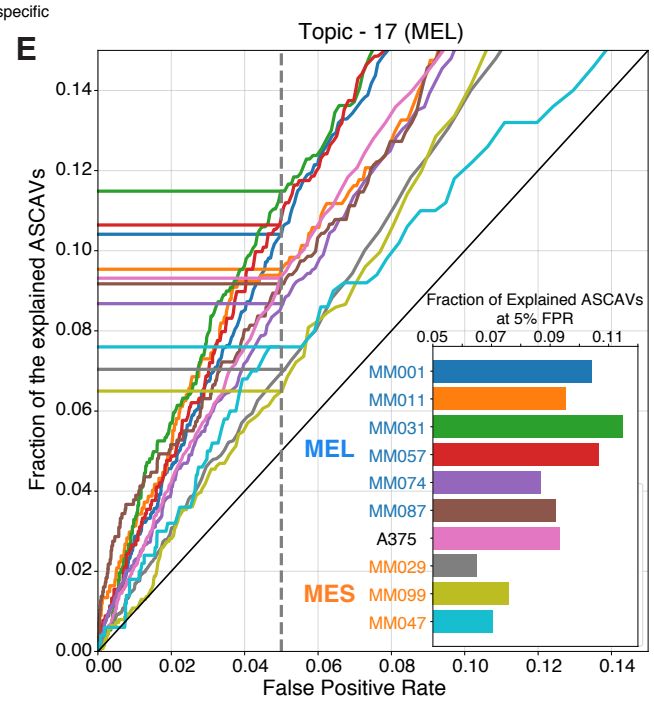

A
ASCAVs
Control variants
ATAC-seq signal
Haplotype-resolved genome sequence
Variant is associated with accesible chromatin
Variant is associated with inaccesible chromatin
Variant is not associated with chromatin accesibility
B
Distance
1
0.8
0.6
0.4
0.2
0
SP
CTCF
TFAP
EBOX
RUNX
SOX
ETS
ATF
AP-1
719 non-redundant motifs in 47 clusters
C
AP-1
ATF
CTCF
CTCFL
EBOX
ETS
GATA
NRF
Other
RUNX
SOX
SP
TEAD
TFAP
ZNF
Fisher's Exact Test -log10(FDR)
Mean(CRM score)
D
AP-1
ATF
EBOX
ETS
RUNX
SOX
TEAD
CTCF
TFAP
CTCFL
SP
NRF
GATA
Number of ASCAVs explained by each motif cluster
D
AP−1 motifs (n=191)
R=0.68, p=0.035
Fraction of ASCAVs with significant motif score change (%)
AP−1 expression (average expression over AP−1 family members)
MM011
MM099
MM047
MM057
MM029
A375
MM087
MM074
MM001
MM031
E
Fraction of the explained ASCAVs (%)
False Positive Rate
MM001
MM011
MM031
MM057
MM074
MM087
A375
MM047
MM029
MM099

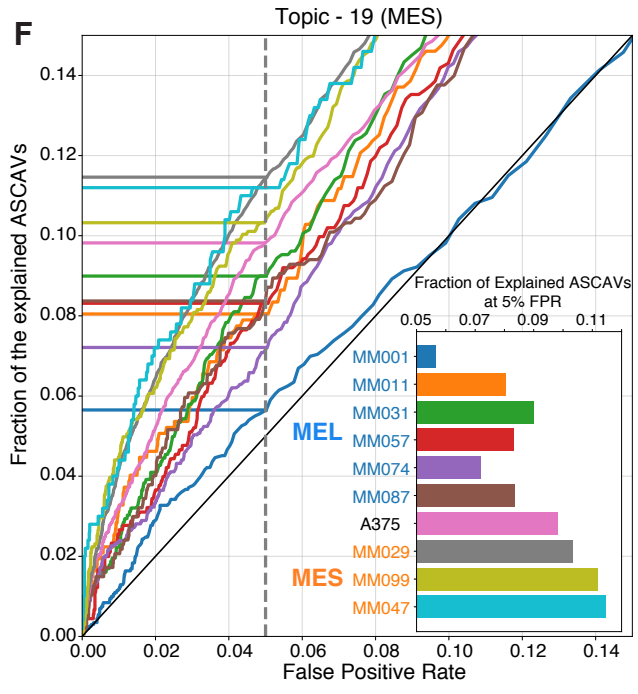

A
Topic contribution per cell
Z-score
−2 0 2 4 6
CellLine
State
Topic 14
Topic 31
Topic 16
Topic 17
Topic 39
Topic 36
Topic 19
Topic 28
Topic 37
Topic 6
Topic 4
Topic 32
Topic 41
Topic 5
Topic 21
Topic 40
Topic 43
Topic 20
Topic 13
Topic 12
Topic 15
Topic 10
Topic 18
Topic 24
Topic 11
Topic 33
Topic 46
Topic 9
Topic 42
Topic 7
Topic 8
Topic 29
Topic 47
Topic 1
Topic 38
Topic 22
Topic 25
Topic 3
Topic 26
Topic 23
Topic 45
Topic 34
Topic 30
Topic 44
Topic 35
Topic 27
Topic 2
CellLine
A375
MM001
MM011
MM029
MM031
MM034
MM043
MM047
MM050
MM054
MM057
MM070
MM074
MM087
MM098
MM099
MM116
MM118
MM122
MM125
MM142
MM144
MM148
MM150
MM152
MM161
MM162
MM163
MM164
MM165
State
MEL
MES
unknown
Topics
General
State specific
Cell line specific

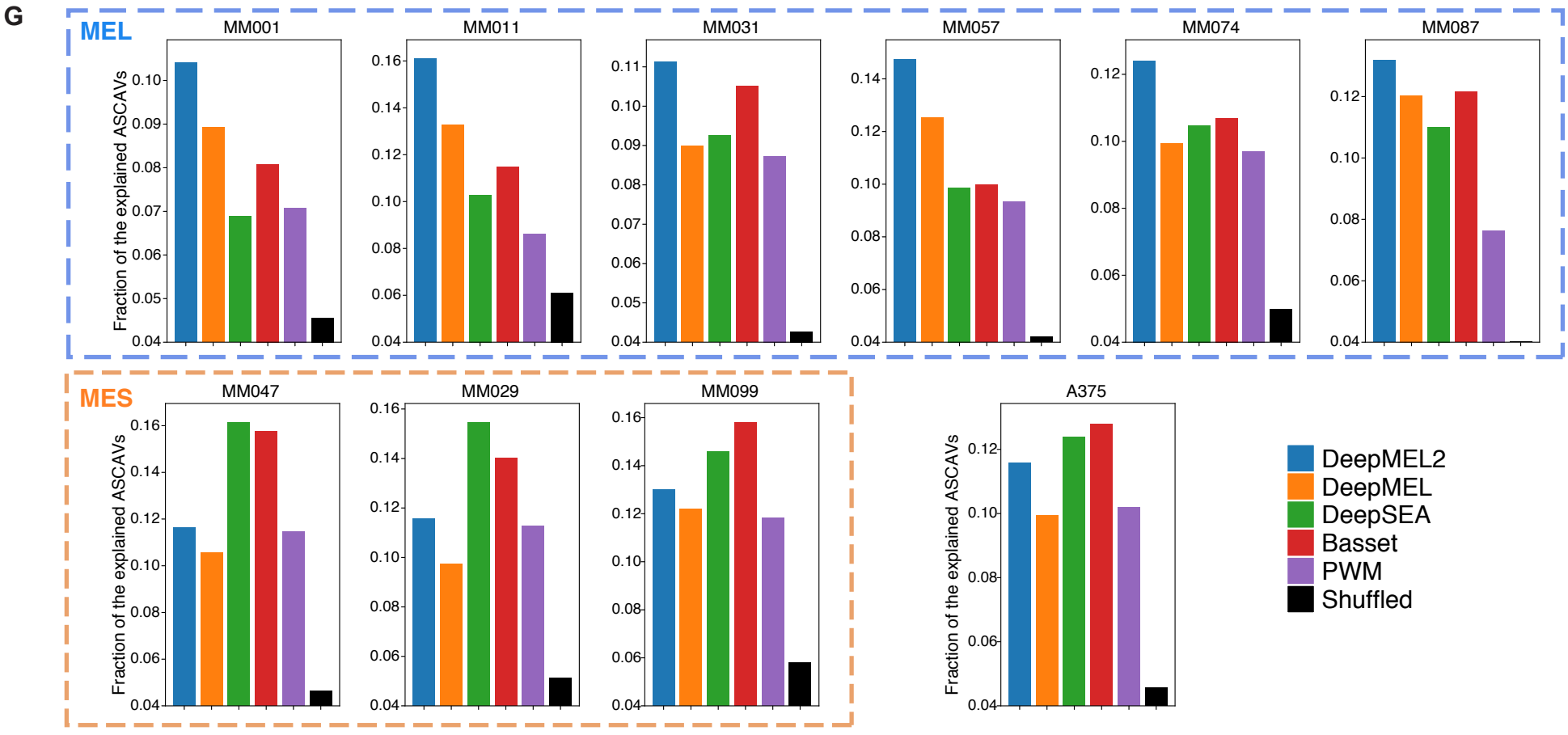

B
DeepMEL2
30 MM Lines
47 Topics
DeepMEL
16 MM Lines
24 Topics
Input
Convolutional
Max Pooling
LSTM
Dense
Output
Clustered PWMs

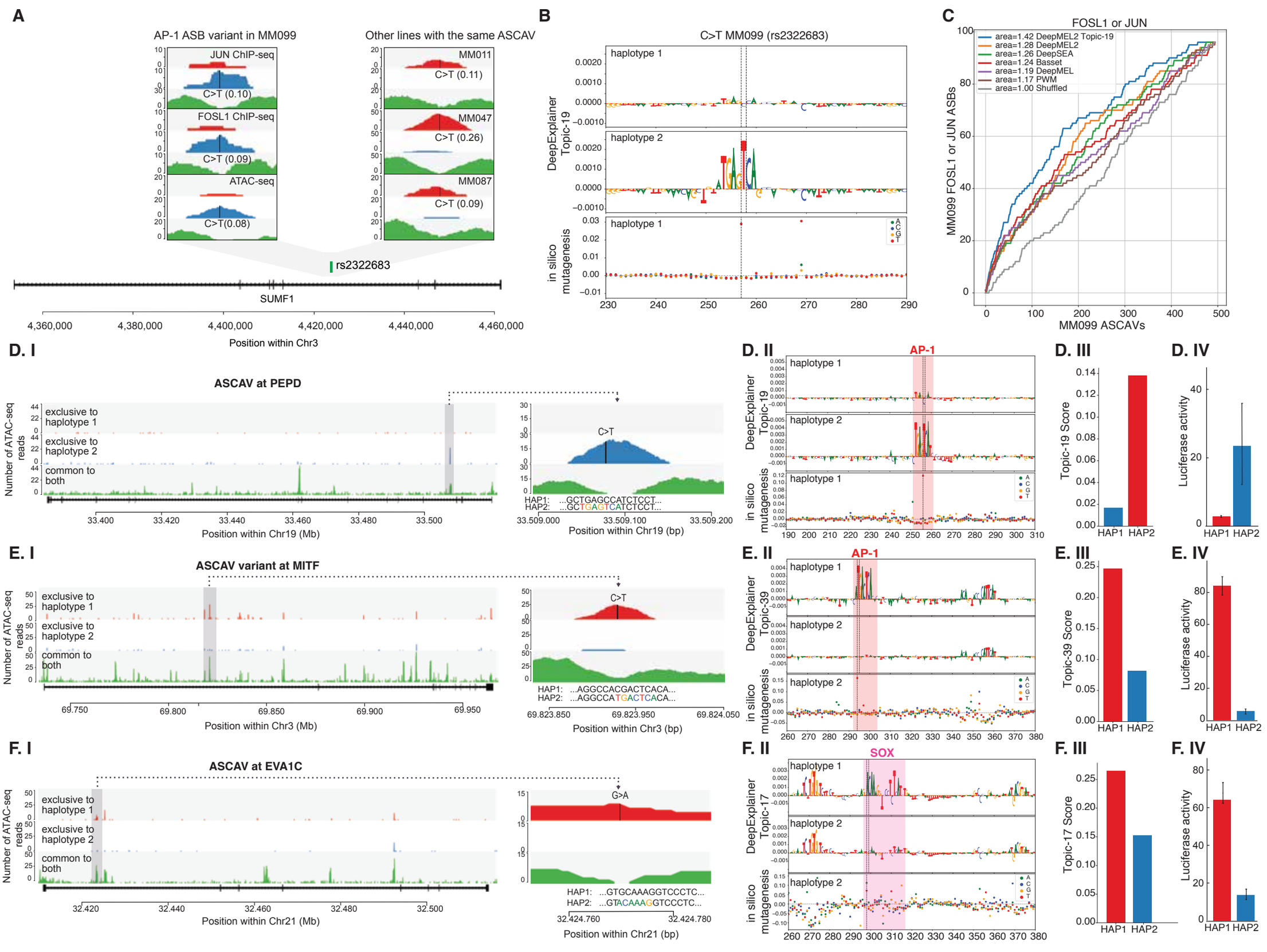

C
DeepMEL2's performance on validation set
auPR
auROC

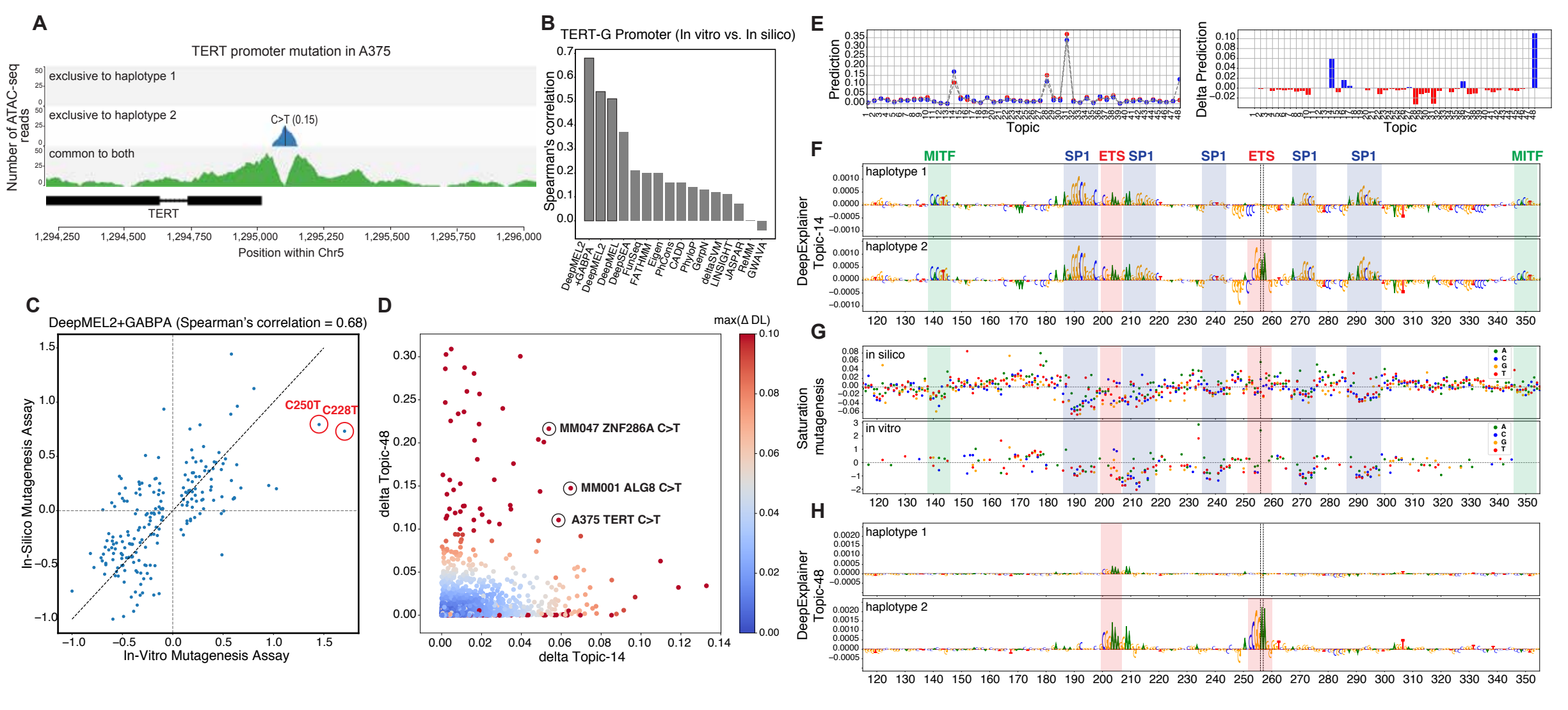

D
IRF4 Enhancer (In vitro vs. In silico)
Spearman's correlation
DeepMEL2
DeepMEL
Eigen
GerpN
DeepSEA
LINSIGHT
CADD
FunSeq
PhCons
ReMM
FATHMM
JASPAR
PhyloP
GWAVA


E
Topic - 17 (MEL)
Fraction of the explained ASCAVs
False Positive Rate
Fraction of Explained ASCAVs at 5% FPR
MEL
MES
MM001
MM011
MM031
MM057
MM074
MM087
A375
MM029
MM099
MM047


F
Topic - 19 (MES)
Fraction of the explained ASCAVs
False Positive Rate
Fraction of Explained ASCAVs at 5% FPR
MEL
MES
MM001
MM011
MM031
MM057
MM074
MM087
A375
MM029
MM099
MM047


G
MEL
MES
MM001
MM011
MM031
MM057
MM074
MM087
MM047
MM029
MM099
A375
Fraction of the explained ASCAVs
DeepMEL2
DeepMEL
DeepSEA
Basset
PWM
Shuffled


A
AP-1 ASB variant in MM099
Other lines with the same ASCAV
JUN ChIP-seq
C>T (0.10)
FOSL1 ChIP-seq
C>T (0.09)
ATAC-seq
C>T(0.08)
MM011
C>T (0.11)
MM047
C>T (0.26)
MM087
C>T (0.09)
rs2322683
SUMF1
4,360,000
4,380,000
4,400,000
4,420,000
4,440,000
4,460,000
Position within Chr3
B
C>T MM099 (rs2322683)
haplotype 1
haplotype 2
DeepExplainer Topic-19
in silico mutagenesis
C
FOSL1 or JUN
area=1.42 DeepMEL2 Topic-19
area=1.28 DeepMEL2
area=1.26 DeepSEA
area=1.24 Basset
area=1.19 DeepMEL
area=1.17 PWM
area=1.00 Shuffled
MM099 FOSL1 or JUN ASBs
MM099 ASCAVs
D. I
ASCAV at PEPD
exclusive to haplotype 1
exclusive to haplotype 2
common to both
Number of ATAC-seq reads
C>T
HAP1: ...GCTGAGCCATCTCCT...
HAP2: ...GCTGAGTCATCTCCT...
Position within Chr19 (Mb)
Position within Chr19 (bp)
D. II
AP-1
D. III
Topic-19 Score
HAP1 HAP2
D. IV
Luciferase activity
E. I
ASCAV variant at MITF
HAP1: ...AGGCCACGACTCACA...
HAP2: ...AGGCCATGACTCACA...
Position within Chr3 (Mb)
Position within Chr3 (bp)
E. II
AP-1
E. III
Topic-39 Score
E. IV
F. I
ASCAV at EVA1C
G>A
HAP1: ...GTGCAAAGGTCCCTC...
HAP2: ...GTACAAAGGTCCCTC...
Position within Chr21 (Mb)
Position within Chr21 (bp)
F. II
SOX
F. III
Topic-17 Score
F. IV


A
TERT promoter mutation in A375
exclusive to haplotype 1
exclusive to haplotype 2
common to both
C>T (0.15)
Number of ATAC-seq reads
TERT
1,294,250
1,294,500
1,294,750
1,295,000
1,295,250
1,295,500
1,295,750
1,296,000
Position within Chr5
B
TERT-G Promoter (In vitro vs. In silico)
Spearman's correlation
DeepMEL2 +GABPA
DeepMEL2
DeepMEL
DeepSEA
FunSeq
FATHMM
Eigen
PhCons
CADD
PhyloP
GerpN
deltaSVM
LINSIGHT
JASPAR
ReMM
GWAVA
C
DeepMEL2+GABPA (Spearman's correlation = 0.68)
In-Silico Mutagenesis Assay
In-Vitro Mutagenesis Assay
C250T
C228T
D
max(Δ DL)
delta Topic-48
delta Topic-14
MM047 ZNF286A C>T
MM001 ALG8 C>T
A375 TERT C>T
E
Prediction
Topic
Delta Prediction
Topic
F
MITF
SP1
ETS
SP1
SP1
ETS
SP1
SP1
MITF
DeepExplainer Topic-14
haplotype 1
haplotype 2
G
Saturation mutagenesis
in silico
in vitro
H
DeepExplainer Topic-48
haplotype 1
haplotype 2