HGG Advances

ARTICLE

DeepFace: Deep-learning-based framework to contextualize orofacial-cleft-related variants during human embryonic craniofacial development

Yulin Dai,[1] Toshiyuki Itai,[1] Guangsheng Pei,[1] Fangfang Yan,[1] Yan Chu,[2] Xiaoqian Jiang,[2] Seth M. Weinberg,[3,4] Nandita Mukhopadhyay,[3] Mary L. Marazita,[3,4,5] Lukas M. Simon,[6] Peilin Jia,[1] and Zhongming Zhao[1,7,8,*]

Summary

Orofacial clefts (OFCs) are among the most common human congenital birth defects. Previous multiethnic studies have identified dozens of associated loci for both cleft lip with or without cleft palate (CL/P) and cleft palate alone (CP). Although several nearby genes have been highlighted, the "casual" variants are largely unknown. Here, we developed DeepFace, a convolutional neural network model, to assess the functional impact of variants by SNP activity difference (SAD) scores. The DeepFace model is trained with 204 epigenomic assays from crucial human embryonic craniofacial developmental stages of post-conception week (pcw) 4 to pcw 10. The Pearson correlation coefficient between the predicted and actual values for 12 epigenetic features achieved a median range of 0.50–0.83. Specifically, our model revealed that SNPs significantly associated with OFCs tended to exhibit higher SAD scores across various variant categories compared to less related groups, indicating a context-specific impact of OFC-related SNPs. Notably, we identified six SNPs with a significant linear relationship to SAD scores throughout developmental progression, suggesting that these SNPs could play a temporal regulatory role. Furthermore, our cell-type specificity analysis pinpointed the trophoblast cell as having the highest enrichment of risk signals associated with OFCs. Overall, DeepFace can harness distal regulatory signals from extensive epigenomic assays, offering new perspectives for prioritizing OFC variants using contextualized functional genomic features. We expect DeepFace to be instrumental in accessing and predicting the regulatory roles of variants associated with OFCs, and the model can be extended to study other complex diseases or traits.

Introduction

Nonsyndromic orofacial clefts (OFCs) are among some of the most common human birth defects, occurring in 1 in 700 live births worldwide.[1] OFCs occur in various forms, including cleft lip alone (CL), cleft palate alone (CP), and a combination of both (CLP), with a spectrum of severity in each case.[1,2] Nonsyndromic OFCs arise without accompanying major cognitive or structural abnormalities and exhibit complex etiology. This complexity is due to the interplay of multiple genetic and environmental risk factors contributing to their development.

In recent years, multiple genome-wide association studies (GWASs) have successfully depicted the genetic architecture of OFCs in multiethnic populations.[3–13] Although dozens of loci have been identified through GWASs, most genetic discoveries are situated within noncoding and regions of linkage disequilibrium.[14] Consequently, delineating the regulatory roles of these associated variants necessitates comprehensive functional genomics data to accurately interpret their biological mechanisms.[15]

During recent years, large-scale experimental mapping of epigenomic modification assays have been conducted by several large consortia, including the Encyclopedia of DNA Elements (ENCODE)[16] and the Roadmap Epigenomics Project,[17] which provide insights for annotating the function of noncoding variants by considering their overlap with regulatory elements in related contexts (tissue, cell type, and developmental stage).[18,19] Furthermore, convolutional neural network (CNN) models have been recognized as a robust approach for investigating regulatory motifs within the genomic context. They are specifically designed to capture high-level information from long sequences, offering valuable insights into the complex patterns of genomic regulation.[20] Currently, many CNN-based frameworks have been implemented to access the function of noncoding variants, such as DeepSEA,[21] Basenji,[22] ExPecto,[23] and our previous work, DeepFun.[24,25] These CNN models provide a computational assessment of

[1]Center for Precision Health, McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA; [2]Center for Secure Artificial Intelligence for Healthcare, McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA; [3]Department of Oral and Craniofacial Sciences, School of Dental Medicine, Center for Craniofacial and Dental Genetics, University of Pittsburgh, Pittsburgh, PA 15213, USA; [4]Department of Human Genetics, School of Public Health, University of Pittsburgh, Pittsburgh, PA 15261, USA; [5]Clinical and Translational Science Institute, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15213, USA; [6]Therapeutic Innovation Center, Baylor College of Medicine, Houston, TX 77030, USA; [7]MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences, Houston, TX 77030, USA
[8]Lead contact
*Correspondence: zhongming.zhao@uth.tmc.edu
https://doi.org/10.1016/j.xhgg.2024.100312.

the regulatory effects resulting from genomic variations by detecting disruptions or creations of regulatory motifs identified through convolutional filters. Consequently, they enable the downstream prediction of chromatin accessibility and regulatory modifications.[21,22] However, these current methodologies predominantly focus on proximal sequences adjacent to risk variants, neglecting the potential for *cis*-regulatory elements to engage in looping interactions extending up to one million base pairs away.[26,27] Moreover, the epigenomic regulation of embryonic craniofacial development is highly context specific.[28] None of the current methods have trained a craniofacial development model. Therefore, predictions based on the noncontextual model will not reflect the dynamic epigenomic signals during craniofacial development.

To address these challenges, we obtained 204 human craniofacial epigenomic arrays, including datasets across six craniofacial developmental stages and 12 epigenetic indicators for enhancers, promoters, and gene bodies. These chromatin feature annotations could complement the modeling of the epigenomic map in craniofacial development. Moreover, we trained a deep learning model specifically for cleft development, DeepFace, to learn the epigenetic feature association with the long-range DNA sequence feature. Therefore, DeepFace predicts the impact of variants on DNA sequence, enabling us to understand how alterations in the DNA sequence influence epigenetic modifications. Next, we applied the DeepFace model to systematically assess curated CP and cleft lip with or without cleft palate (CL/P) risk variants. Then, we characterized variants with the largest accessibility alteration and the development-specific stage of variants. We anticipated that the CNN model on dense epigenomic maps would be a valuable approach for both gene-regulatory studies and disease studies seeking to elucidate the molecular basis of OFCs.

Material and methods

Primary chromatin feature collection and processing

The 204 chromatin immunoprecipitation of post-translational epigenetic modifications coupled with next-generation sequencing (ChIP-seq) data were collected from human embryonic craniofacial tissues[28] and downloaded from the Gene Expression Omnibus (GEO) (accessed on June 8, 2021, GEO: GSE97752).

Briefly, these 204 assays were extracted from 17 individual human embryos during a crucial developmental period. This period encompasses the formation of the human orofacial apparatus, spanning Carnegie stages (CSs) from post-conception weeks (pcw) 4 to pcw 10, including stages CS13, CS14, CS15, CS17, CS20, and F2.[29] For each sample, 11 post-translational histone modifications[28] were profiled, including the repressive marks (H3K27me3 and H3K9me3), promoter activation marks (H3K4me3 and H3K9ac), transcription regulation marks (H3K36me3, H4K20me1,[30,31] and H3K79me2[32]), active regulatory marks (H3K4me1, H3K4me2, H3K27ac, and H2A.Z), and open chromatin signal DNase. Then, we extracted nonoverlapping sequences across the chromosomes, each spanning approximately 131,072 bp (~131 kb) as the segment length of the input. Sequences with more than 35% unmappable content were discarded, collectively covering approximately 81% of the human genome. Each epigenomic data in bigWig format was further converted and split into these segments, resulting in 14,990 segments for training, 1,805 segments for validation, and 1,798 segments for testing, based on a distribution scheme across various samples in an 8:1:1 ratio.

Curation for OFC-significant variants

We collected a diverse set of orofacial variants from the GWAS Catalog[33] (accessed May 16, 2021) using the keyword "oral cleft," resulting in 306 variants with p value $<1 \times 10^{-5}$ from 33 studies (Table S1). These variants included a total number of 234 unique SNPs. We further obtained two multiethnic craniofacial GWAS datasets for CL/P[7,34] and CP[13] (available from dbGaP: phs000884.v2.p1). We selected the SNP with at least one SNP with nominal significance p value $<1 \times 10^{-5}$. Our craniofacial-sign (OFC-sign) dataset has 1,787 SNPs in total.

Curation for control variants

From previous studies,[35–38] we observed that trait-related variants tended to manifest their effect in the trait-related tissues. We investigated whether OFC-significant variants exhibit higher absolute SNP activity difference (SAD) scores than variants from unrelated traits using control variant collections equal in size to the OFC-significant collection from two resources: (1) nonsignificant craniofacial development variants and (2) irrelevant trait variant collections. The first variant collection was obtained from the aforementioned two GWAS datasets (CL/P[34] and CP only [CPO][13]). The OFC-low group was defined as randomly sampled variants with p value >0.5 in both GWAS datasets. The OFC-medium group was defined as randomly sampled variants with p values ranging from 1×10^{-5} to 0.5 in both GWAS datasets. The second irrelevant trait variant was chosen from two GWAS summary statistics datasets: neurodegenerative disease Alzheimer disease[39] (AD) and psychiatric disorder schizophrenia[40] (SCZ), each with p value $<1 \times 10^{-5}$ significance (AD-sign and SCZ-sign). We further employed random downsampling to ensure that all control datasets contained the same number of SNPs as the OFC-sign collection, thereby enhancing comparability. All the control variant groups were randomly sampled to match the size of the OFC-sign group.

Variant annotation

SNPs were annotated with the ANNOVAR function "table_annovar.pl" (v.4/16/2018) with hg19 reference genome and dbsnp150 version annotation.[41] The function of the SNPs was annotated and merged into the following categories: exon (variant overlaps a coding), intergenic (variant is in intergenic region), intronic (variant overlaps an intron), ncRNA_exonic (variant overlaps a transcript coding region without coding annotation in the gene definition), ncRNA_intronic (variant overlaps a transcript intron region without coding annotation in the gene definition), upstream/downstream (variant overlaps 1-kb region upstream or downstream of transcription start site), and untranslated region (UTR) 3′/UTR 5′ (variant overlaps a 3′ or 5′ UTR).

Training the CNN model and model performance evaluation

We utilized the CNN framework Basenji[22] and our in-house DeepFun[24] to train the 204 epigenomic assays. The CNN

architecture consists of seven dilated convolution layers with max pooling (in windows of two, four, four, and four) to obtain representations that describe 128-bp bin size, aligning with the 146-bp distance between nucleosome core particles.[22] This design allows information sharing across distal regulatory interactions ($128 \times 2^7 \times 2 = 32{,}000$). We applied seven layers of dilated convolutions to encompass these 128-bp bin representations, transforming every sequence feature (131,000) and the epigenetic signals into a length of 1,024-bp subsequence representations. Our previous work[24] has demonstrated that training on the complete features from the ENCODE dataset[42] outperforms the training on individual features. Therefore, we trained these sequence features across all 204 chromatin assays. We evaluated the performance on the validation and testing sets based on the Pearson's correlation coefficient (r) of predicted and real epigenetics features. Each assay's predicted epigenomic intensity was computed individually. Lastly, we fine-tuned the hyperparameters, learning rate, and batch size and stopped training when there was no reduction in r in the validation set loss over 15 consecutive epochs.

To evaluate the peak binary classification, we followed the Basenji[22] model to evaluate the peak binary classification comparison with one well-known method, Model-based Analysis of ChIP-Seq (MACS2).[43] We transformed the training and testing datasets to binary peak calls on shorter sequences. Each 131,000 (~1,024 binned subsequence × 128 bps/bin) sequence was segmented into subsequences of 1,024 bin features, with each subsequence encapsulating a 128-bp binned representative of the functional element. The aim of this deep learning model is to accurately predict the read coverage in 128-bp bins. We identified peaks within the central 256 bps of the subsequences for each dataset by applying a Poisson model to the smoothed, normalized counts. This model was parameterized by the higher value of a global and local null lambda, akin to the MACS2 methodology. We then established a 0.01 false discovery rate (FDR) cutoff to define the ground truth. The area under the precision-recall curve (AUPRC) was used to measure the model performance of prediction. More details about the model can be found in the Basenji model.[22]

SAD score

The DeepFace model is crafted to forecast the functional impacts of sequence alterations at a single-nucleotide resolution. For each variant, DeepFace considers contextual information within a 1,024-bp subsequence transformed from a 131-kb sequence, predicting the epigenomic activity probability for sequences containing the reference allele or alternative allele. In this context, activity denotes the binding affinity for DNase-seq or histone modifications, respectively. To assess the variant's impact, we employed the SAD, SAD = SA(alt_allele) − SA(ref_allele), where SA(alt_allele) and SA(ref_allele) are from the predicted matrices, to represent the predicted SNP activity for the alternative allele and the reference allele sequence, respectively. An elevated positive SAD score for genetic variants denotes that the alternative allele augments the epigenetic signal in comparison to the reference allele. Conversely, a negative SAD value denotes a diminution of the epigenetic signal. Notwithstanding the collective training of DeepFace models utilizing an extensive dataset, the functional score predicted for each variant is distinct and autonomous.

Motif mapping and visualization

We used the R package "atSNP" to search the potential transcription factor (TF) binding motif of variants in the JASPAR[44] and ENCODE[42] motif databases. We utilized the "ComputePValues" function within the atSNP toolkit to calculate the p values for all potential motifs. We identified significant motifs as those with Benjamini-Hochberg procedure-adjusted p values <0.05 in either the JASPAR or ENCODE database. Additionally, we employed the "plotMotifMatch" function from the atSNP package to visualize the motif pattern of the significant SNPs.

Cell-type specificity analysis of OFC-sign SNP set

Considering the epigenomic data of DeepFace are limited to the tissue of embryonic craniofacial development, we used two in-house methods, web-based cell-type-specific enrichment analysis of genes (WebCSEA)[37] and DeepFun[24,25] to contextualize the most relevant cell types of OFC-sign genes. WebCSEA (https://bioinfo.uth.edu/webcsea/) curated a total of 111 single-cell RNA-seq panels of human tissues and 1,355 tissue cell types from 61 different general tissues across 11 human organ systems and used the decoding tissue specificity algorithm[35] to measure the enrichment for each cell type.[37] We input the most nearby genes of the OFC-sign SNP set and visualized the most enriched cell type with a nominal significance of 1×10^{-3}.

The DeepFun web server (https://bioinfo.uth.edu/deepfun/) leverages a CNN architecture trained on approximately 8,000 chromatin feature assays from 225 distinct tissues or cell types from the ENCODE and Roadmap projects. We input all OFC-sign SNPs to the DeepFun web server to assess the SAD scores, which is the normalized version (range from −1 to 1) of SAD used in this study. For every SNP in each cell type, we calculated the mean absolute SAD and then identified the cell types with the highest absolute SAD values across the OFC-sign SNPs. The top count of cell types was defined as the cell type most related to the OFC-sign SNP set.

Results

Narrow peak epigenetic chromatin features had better prediction than broad peak features

Following the DeepFace design in Figure 1, the trained 204 chromatin feature assays were evaluated for the prediction performance on the r of predicted and real continuous epigenetics features (median ranging from 0.50 to 0.83) and the AUPRC of the binary predicted peak and real peak called by MACS (ranging from 0.54 to 0.81). As shown in Figures 2A and 2B, both continuous and binary epigenomic features shared the same trend over the chromatin features. Specifically, H3K4me3 and H3K79me2 are on the top in Pearson's r and AUPRC values, respectively. The two broad repressive marks, H3K27me3 and H3K9me3, have the lowest medium performance across the samples over development stages, suggesting that the narrow peak histone modification features tend to have a more accurate prediction than the broad histone modification features.[24]

OFC-sign variants show greater enrichment in embryonic craniofacial development than other sets of variants

We implemented our pretrained DeepFace model to predict the SAD scores of curated SNP sets (OFC-sign, OFC-medium, OFC-low, AD-sign, and SCZ-sign, see

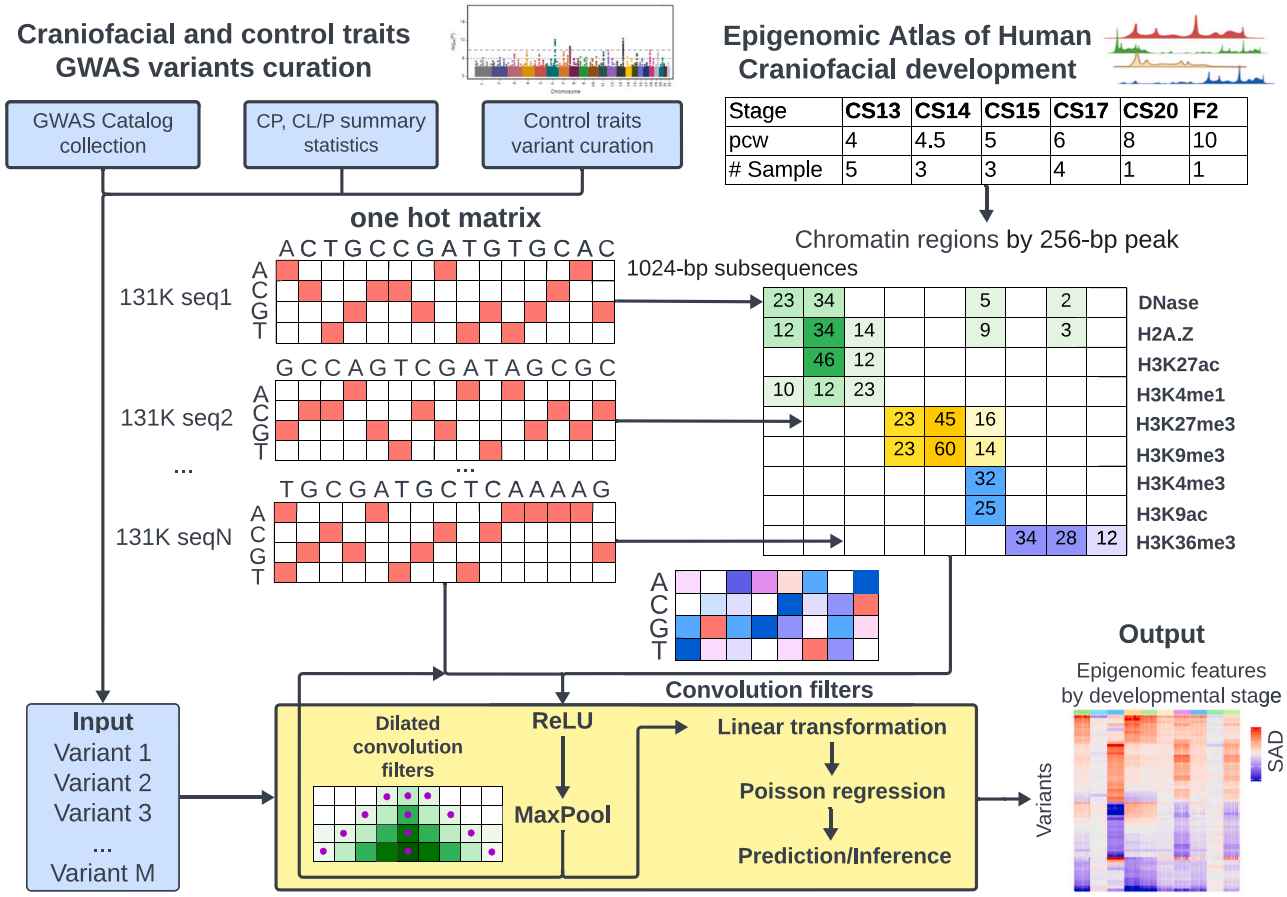

Figure 1. Overview of DeepFace workflow
DeepFace is a dilated convolutional neural network (CNN) framework to contextualize the function of common orofacial cleft (OFC) variants trained from 204 human embryonic craniofacial epigenomic arrays (six stages of craniofacial development and 12 histone modification markers for enhancers, promoters, and gene bodies, Wilderman et al.[28]).

material and methods). As shown in Figure 2C, while most SNP sets showed minimal variation in absolute SAD scores for histone modifications, notable exceptions were found in transcription regulation markers (H3K36me3, H3K79me2, and H4K20me1). Among them, the AD-sign and SCZ-sign SNP sets exhibited significantly higher median absolute SAD scores. Conversely, Figure 2D revealed that the OFC-sign group consistently presented higher median absolute SAD scores across various variant categories, particularly in intronic (OFC vs. AD $p_{FDR} = 0.02$; OFC vs. SCZ $p_{FDR} = 3.51 \times 10^{-8}$), ncRNA_exonic (OFC vs. AD $p_{FDR} > 0.05$; OFC vs. SCZ $p_{FDR} = 0.02$), and upstream/downstream regions (OFC vs. AD $p_{FDR} = 1.01 \times 10^{-5}$; OFC vs. SCZ $p_{FDR} = 6.82 \times 10^{-9}$) (Figure S1). Meanwhile, OFC-low or OFC-medium tends to be the lowest absolute SAD score in any category. This observation suggested a higher enrichment of functionally variant OFC-sign sets affecting SAD scores when compared to the other groups. Interestingly, the AD-sign set stood out with higher median absolute SAD scores in the UTR3/UTR5 regions. Moreover, the AD-sign set stood out with higher median absolute SAD scores in the upstream; downstream category as well. Figure S2 indicated a higher prevalence of SNPs in the exonic, ncRNA_exonic, UTR3/UTR5, and upstream/downstream categories for both AD and SCZ groups, as these variants may play a more significant role in altering gene function. Coherent with Figure 2C, these genomic regions are typically enriched with transcription regulation signals (H3K36me3, H3K79me2, and H4K20me1). Consequently, the elevated medium absolute SAD scores of transcription regulation markers in AD and SCZ could be attributed to the transcription regulation within the upstream, downstream, and UTR3, and UTR5 regions.

SAD scores offer the promise of interpretation function of known OFC-related variants

SAD scores could link the function to OFC-sign variants (Table S2). For example, rs117496742 (risk-allele A, lead SNP) is an intronic variant located within the *YAP1* on chromosome 11q22.1. This variant has garnered genome-wide significance in European populations, as documented in CPO.[14] Notably, within the CS20 stage, characterized by active regulatory mark H3K4me1, rs117496742 boasts the highest absolute SAD score (75.44), underscoring its potential regulatory impact. In contrast, during the CS14 stage, the same variant exhibits the lowest SAD score (70.9)

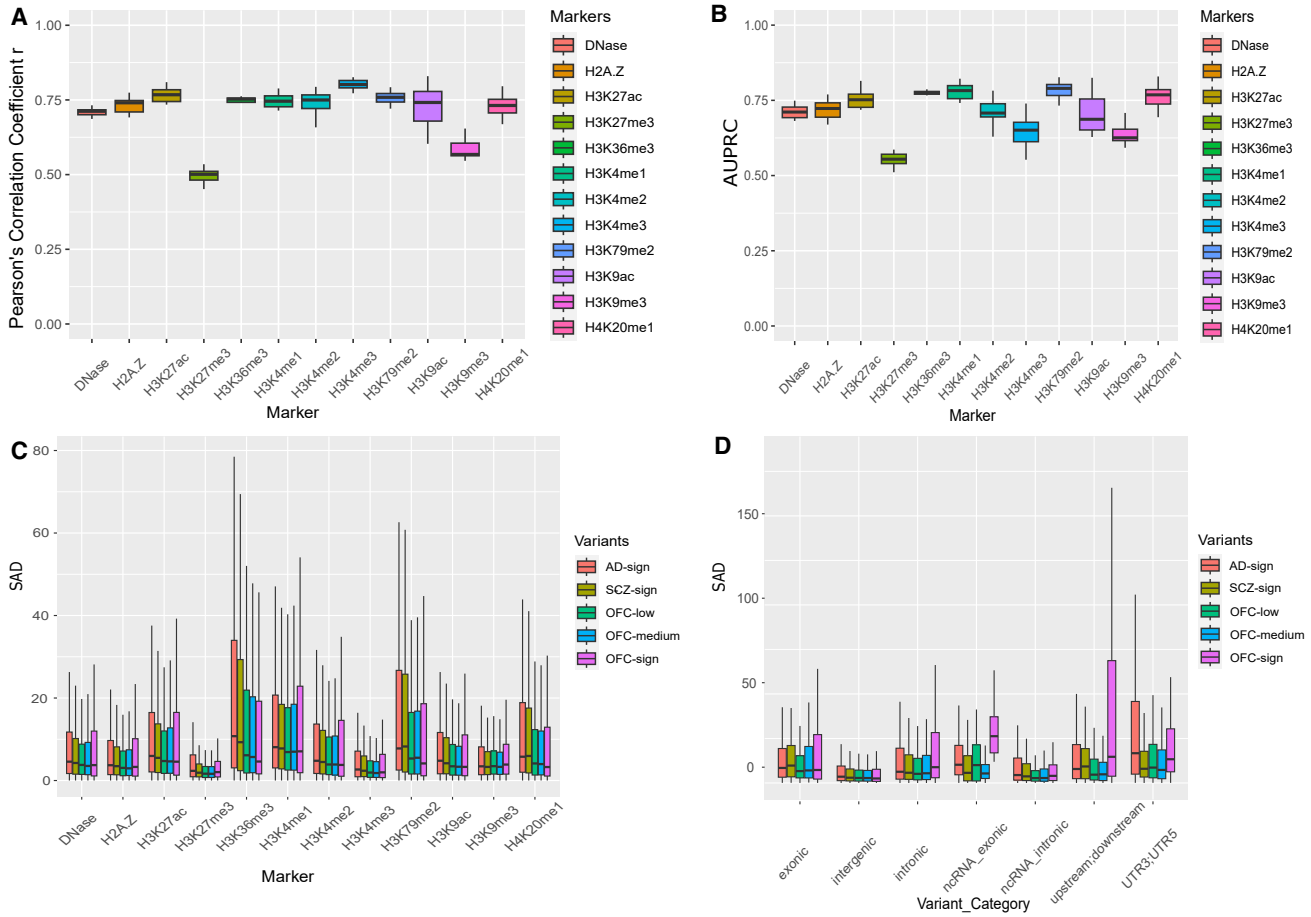

Figure 2. DeepFace performance

Boxplot indicates the minimum, 25th percentile, median, 75th percentile, and maximum values for each category.

(A) The Pearson's correlation coefficient (r) of predicted and real epigenetics features among samples across developmental stages by each chromatin feature.

(B) The area under the precision-recall curve (AUPRC) for predicted peak and real peaks called by MACS among the samples across developmental stages by each chromatin feature.

(C) SNP activity difference (SAD) scores of the variants in five different variant categories: AD-sign, SCZ-sign, OFC-low, OFC-medium, and OFC-sign stratified by each chromatin feature.

The function epigenetic features in (A)–(C) are summarized: repressive marks (H3K27me3 and H3K9me3), promoter activation marks (H3K4me3 and H3K9ac), transcription regulation marks (H3K36me3, H4K20me1, and H3K79me2), active regulatory marks (H3K4me1, H3K4me2, H3K27ac, and H2A.Z), and open chromatin signal DNase.

(D) SAD scores of variants from five genetic background categories: AD-sign, SCZ-sign, OFC-low, OFC-medium, and OFC-sign stratified by variant category.

within the H3K4me1 profile. Similarly, rs12543318 (risk-allele C, lead SNP) is an intergenic variant proximal to *DCAF4L2* and *MMP16* on chromosome 8q21. This variant has been identified as nominally significant in multiethnic populations, as reported in CL/P.[7] Noteworthy is its behavior within the CS13 stage, marked by active regulatory mark H3K4me2, where it registers the highest absolute SAD score (52.78), indicating its potential regulatory influence. Conversely, in the CS15 stage, this variant displays the lowest SAD score (36.0) within the H3K4me2 profile.

SAD scores of SNPs may reflect temporal regulation roles

The epigenomic assays from different developmental stages provided us with opportunities to explore the potential temporal epigenetic alteration by variant. We hypothesized whether the SAD scores are associated with the temporally regulatory roles of SNPs. Here, we mainly explored the potential linear regulatory roles of SNPs across the craniofacial development course. To this end, we employed a generalized linear model to assess whether the SAD scores for each SNP exhibited a significant linear relationship with various developmental stages. The linear model coefficient *p* values were further adjusted by Bonferroni correction of the 1,590 nonzero SAD SNPs from 1,787 OFC-sign SNPs. This procedure revealed six SNPs with significant linear association, suggesting their roles in influencing features throughout the course of craniofacial development. As illustrated in Figure 3, rs1339063 (A>T) is an intronic variant in gene *PAX7*. The predicted SAD

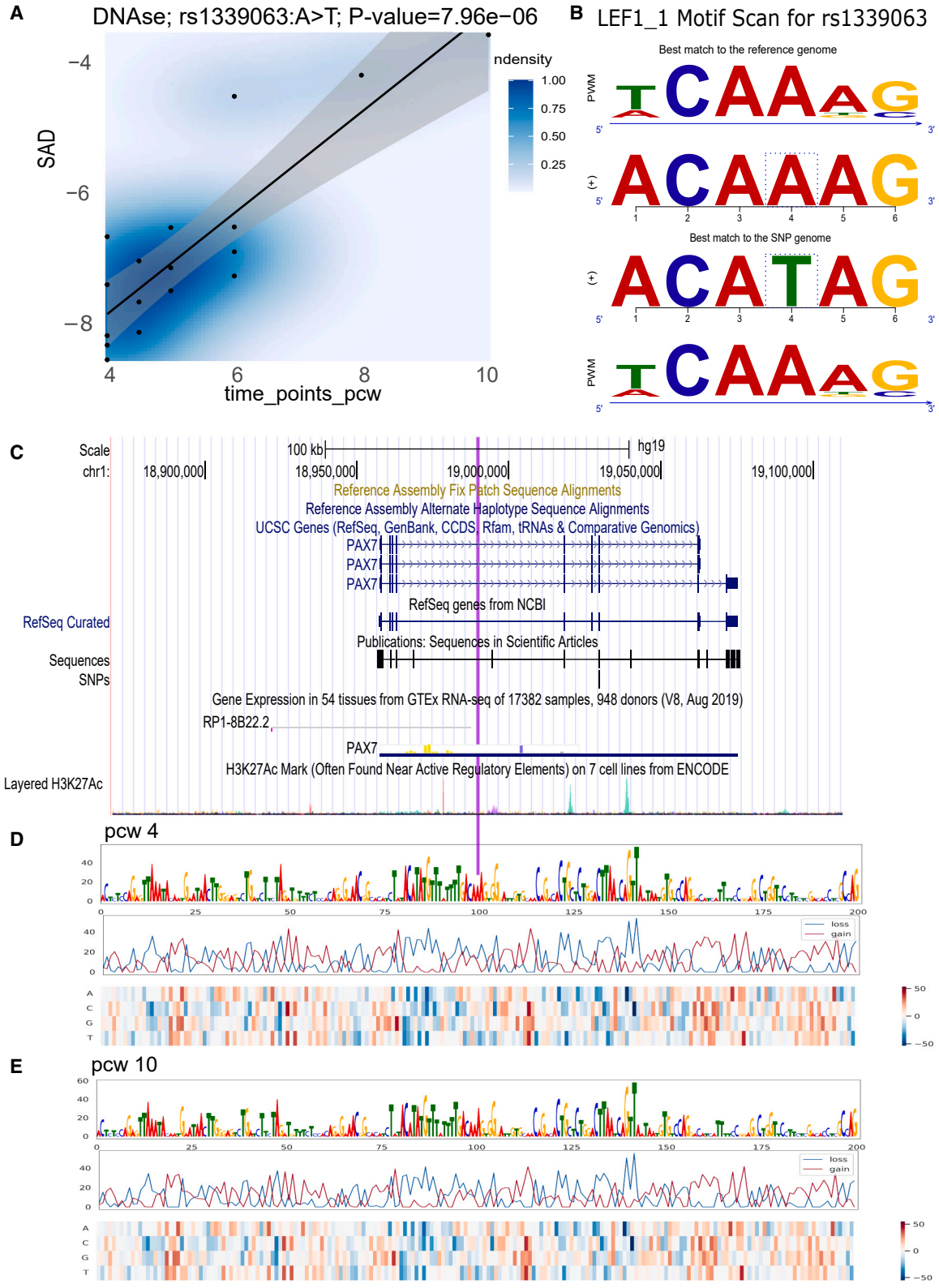

Figure 3. DeepFace variant rs1339063: SAD and motif analysis

(A) SAD for rs1339063 (A>T) associated with the post-conception week (pcw) trajectory using DNAse data. The blue area of intensity is positively correlated with the SAD score point density. The black line and gray confidence interval (95%) model the linear relationship between SAD scores and pcw trajectory.

(B) Sequence logo stacks from top to bottom: sequence logo of reference allele matching position weight matrix, reference subsequences, alternative allele subsequences, and sequence logo of SNP allele matching position weight matrix.

(legend continued on next page)

in H3K27me3 showed a significant decrease during the pcw trajectory. The mapped motif is LEF1, a member of the T cell factor/LEF1 family of high-mobility group TFs, which is a downstream mediator of the Wnt/β-catenin signaling pathway.[45] The alternative allele (T) decreased the binding affinity of LEF1 on this site. Figures 3D and 3E illustrate that pcw 10 exhibited a lower SAD score when compared to pcw 4, suggesting that the variant had a stronger effect during pcw 4. Both *LEF1* and *PAX7* were actively expressed in craniofacial tissue from CS13 to CS22, as shown in Cotney lab's craniofacial Genome Browser[46] (Figures S3 and S4).

Cell-type specificity analysis of OFC-sign SNP set

The pleiotropic nature of OFC genes underscores the necessity of understanding the specific tissue cell types and contexts where disease-related variants predominantly exert their effects. So far, no human craniofacial context epigenomic data have been available in the current deep learning framework. Therefore, we applied two alternative methods, WebCSEA[37] and DeepFun.[24,25] The top three enriched cell types identified by WebCSEA were endothelial cells, trophoblast cells, and stromal cells (Figure 4A). The top three ranked primary cell types by DeepFun were foreskin_melanocyte, trophoblast_cell, and T-helper_2_cell (Figure 4B). Both methods identified trophoblast cells among the top three ranked cell types, suggesting that OFC-risk genes manifest their function most during the embryonic developmental stage.[47] This cell type is associated with the embryonic stage of craniofacial development,[48] aligning with the finding of their similarity to stem cells revealed by Wilderman et al.[28] The melanocyte cell originates from the neural crest, which itself emerges from the neural tube. After formation, neural crest cells undergo a process known as the epithelial-to-mesenchymal transition, during which they detach from the uppermost part of the neural tube.[49] Recent studies[50,51] have revealed that endothelial cells and the vasculature play a pivotal role in guiding tissue morphogenesis and cell differentiation in various cranial structures. Additionally, genes from the vascular endothelial growth factor (VEGF) family have been observed in the mesenchyme surrounding Meckel's cartilage.[52] Furthermore, rare variants in the VEGFA gene have been associated with nonsyndromic CL/P,[53] underscoring the significant role of endothelial cells in craniofacial development. In addition, our enrichment analyses identified epithelial cells and stromal cells (mesenchymal cells), both well documented for their involvement in OFC disorders,[54] as top related cell types. Specifically, epithelial cells were ranked fifth and fourth in the WebCSEA and DeepFun analyses, respectively. This high ranking suggests that their signals are prominent, as many genes play pleiotropic roles across various cell types. Furthermore, the stromal cell type, a subset of mesenchymal cells crucial for structural support and craniofacial development, was ranked third in the WebCSEA analysis, underscoring their importance in the context of OFCs.

Discussion

So far, GWASs from both genotyping and genome sequencing have been extensively performed, leading to many thousands of variants with association signals of the disease or traits under investigation. However, great challenges remain because the roles of most of these variants are not clear, impeding the understanding of molecular mechanisms of disease and further development of disease prevention and therapeutic strategies. Therefore, prioritizing potential causal variants, particularly the thousands of noncoding variants with association signals, is crucial for fully understanding pathogenic mechanism of OFCs. In this work, we aimed to contextualize the function of a comprehensive collection of OFC-related variants during human craniofacial development. To achieve this, we built a deep-learning-based framework, namely DeepFace, by leveraging a spectrum of epigenetics features during the key human embryonic craniofacial development stages. Our DeepFace model pinpointed the high-risk OFC coding and noncoding variants that tended to have the largest predicted SAD scores in several variant categories, including intronic, ncRNA_exonic, and upstream/downstream. Our temporal association analysis further identified six high-risk craniofacial SNPs that exhibited a significant linear relationship between epigenetic impact and the craniofacial developmental process. Overall, DeepFace leveraged the *cis*-regulatory features to provide a high-resolution prediction on the functional changes caused by OFC-related variants during human craniofacial development. To our knowledge, this is the first deep learning model specifically for craniofacial development by leveraging 204 human functional genomics datasets.

As summarized in Table 1, two SNPs, rs1339063 and rs56675509, were in the intronic region of gene *PAX7*, which encodes paired box 7. *PAX7* belongs to the paired box gene family and plays a role in neural crest development, contributing to various tissues, including craniofacial

(C–E) The best match reference sequence and alternative allele sequence for the motif LEF1 were visualized in (C) UCSC genome browser for the rs1339063 and its surrounding genes. The purple vertical line indicates the exact genomic region of 200 bps for pcw 4 (D) and pcw 10 (E). (D) and (E) show the dynamic gain and loss of SAD scores for all possible substitutions in each of the 200-bp genomic positions around the rs1339063 in pcw 4 and pcw 10, respectively. The alteration between (D) and (E) is relatively small in the figure. These SAD score dynamics were visualized in three rows. Top row: sequence logo weight by the loss of SAD across 200-bp sequence; middle row: the blue and red lines indicating the minimum (loss) and maximum (gain) changes among the possible substitutions from reference allele; and bottom: the quantities in the heatmap, which reflects the change in SAD after substituting the reference allele with the alternative allele.

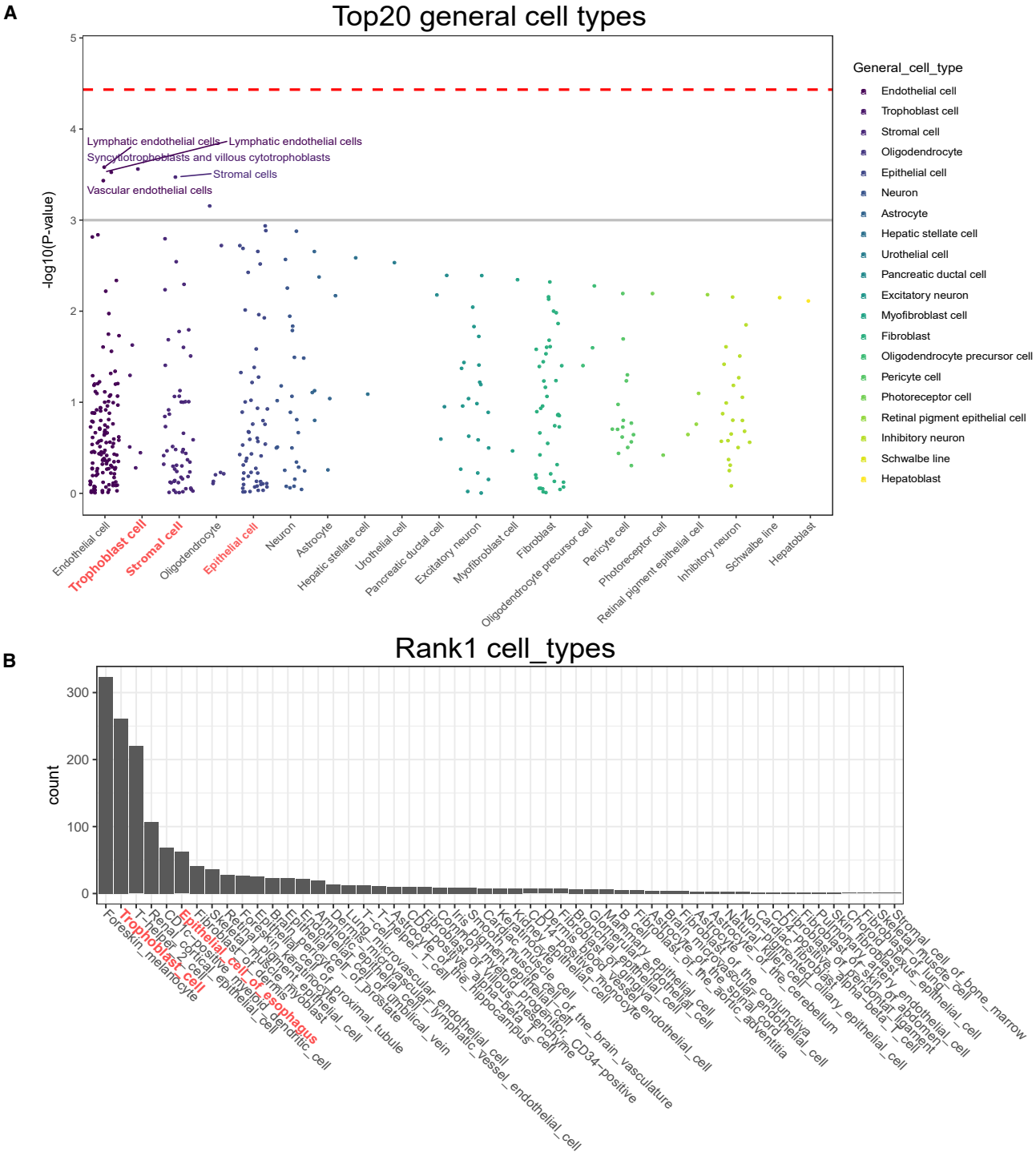

Figure 4. Cell-type specificity analysis for OFC variants

(A) WebCSEA result: the red dashed line represents the Bonferroni-corrected significance threshold at $-\log_{10} p$ value cutoff 3.69×10^{-5}. The gray solid line marks the nominal significance level at $-\log_{10} p$ value cutoff 1×10^{-3}. In every general cell-type category, each dot represents a specific tissue cell type within that category, differentiated by color according to the column it belongs to. We highlighted the top five tissue cell types.

(B) DeepFun results: for each of the SNPs from the OFC-sign SNP set, we calculated the mean absolute SAD and then identified the cell types with the highest absolute SAD values. The count of primary cell types with the highest absolute SAD values is visualized in the bar plot.

bones and cartilage.[57,58] Several SNPs have been reported to increase the risk of nonsyndromic CL/P.[59,60] Our finding thus supported the utilities of the DeepFace model; however, further experiment validation of the regulatory roles of these two SNPs will be warranted. We further discuss their roles below.

Table 1. Summary of six SNPs with significant linear association

SNP ID	Chr	Pos[a]	Minor allele	Gene	Motif	Trait	Reference (PubMed)
rs56675509	1	18971634	C	*PAX7*	ZBTB14	CL/P_all_pop	Mukhopadhyay et al.[34]
rs1339063	1	18989575	T	*PAX7*	LEF1	CL/P_all_pop	Mukhopadhyay et al. [34]
rs2302304	19	3733651	A	*TJP3*	Nkx2-5	cleft lip with or without cleft palate x maternal periconceptional vitamin use interaction (parent of origin effect)	Haaland et al.[55]
rs6495117	15	74899500	A	*CLK3*	EGR1	nonsyndromic cleft lip with cleft palate	Yu et al.[56]
rs11787407	8	129985440	G	*LINC00976/CCDC26*	JUN/FOS	csa_CL/P,eur_CL/P, CL/P_all_pop	Mukhopadhyay et al.[34]
rs12075674	1	209995470	A	*IRF6*	AFP	csa_CL/P, CL/P_all_pop	Mukhopadhyay et al.[34]

csa_CL/P, cleft lip with or without palate in Central/South Asian ancestry; eur_CL/P, cleft lip with or without palate in European ancestry; CL/P_all_pop, cleft lip with or without palate in all populations (European and Central/South Asian).
[a]hg19.

The two representative significant motifs on rs1339063 and rs56675509 are LEF1 and ZBTB14, respectively (Figures S4 and S5). Gene *LEF1* is expressed in the neural crest[61] and plays a role in patterning the mesoderm and ectoderm in *Xenopus*.[62] In mice, *Lef1* plays an important role in epithelial-mesenchymal transition during palatal fusion.[63] ZBTB14 belongs to the zinc-finger and BTB/POZ (broad-complex, tramtrack, and bric-a-brac/poxvirus and zinc-finger) domain-containing protein family, which regulates organ morphogenesis and development.[64,65] In *Xenopus*, *Zbtb14* plays a crucial role in the formation of dorsal-ventral and anterior-posterior axes by regulating BMP and Wnt signaling pathways, both of which are crucial to midfacial development.[66,67]

SNP rs2302304 (Figure S6) is an intronic variant in gene *TJP3* encoding tight junction protein 3, which is a member of the family of membrane-associated guanylate kinase-like proteins that are associated with intracellular junctions.[68] Silencing tjp3/zo-3 using morpholinos leads to edema, loss of blood circulation, and tail fin malformations in zebrafish embryos.[69] The TF binding motif of this variant is NK2 homeobox 5 (Nkx2-5), which has been reported to play an important role in craniofacial development in zebrafish through regulating the endothelin.[70] Funato et al.[71] also found that NKX2-5 is involved in molecular function and biological pathways of CPO, incomplete CP, and submucous CP.

SNP rs6495117 (Figure S7) is an intronic variant in gene *CLK3* encoding CDC-like kinase 3, which is a member of the cdc2-like kinases with four isoforms.[72] In *Xenopus*, *Clk3* knockdown leads to severe developmental defects such as reduced head and eye size and a shortened anterior-posterior axis.[73] The TF binding motif of this variant is early growth response 1 (EGR1), which is an EGR gene that regulates the skeleton's normal development.[74,75] In our previous work,[76] using a developmental-stage-specific network approach integrating TFs and microRNAs, our results showed that Egr1 was a crucial regulator in mice embryogenesis from embryonic day (E) 11.5 to 13.5.

SNP rs11787407 (Figure S8) is an intergenic variant nearby gene *LINC00976/CCDC26*, a long noncoding RNA that is related to cancers,[77] though its functions remain to be elucidated. It is suggested that rs987525, located near *CCDC26*, increases the risk of nonsyndromic CL/P.[78–80] The motif of the variant is FOS/JUN, which is a transcriptional regulator consisting of members of the Fos and Jun families.[81] Fos disruption causes craniofacial anomalies in zebrafish.[82] The recent single-cell RNA-seq and single-cell multiome studies in mice also showed that Fos and Jun were involved in secondary palate development[54] and all-*trans* retinoic-acid-induced CP.[83]

SNP rs12075674 (Figure S9) is an intronic variant in gene *IRF6* encoding interferon regulatory factor 6, which is one of nine TFs that share a highly conserved helix-turn-helix DNA-binding domain.[84] IRF6-related disorders, which are caused by both common and rare variants, have a wide variety of symptoms, including nonsyndromic CL/P and CPO and Van der Woude syndrome (MIM: 119300) at the mild end to the more severe popliteal pterygium syndrome (MIM: 119500).[85,86] The alpha-fetoprotein enhancer binding protein (AFP-1) motif currently lacks direct evidence linking it to craniofacial development.

In summary, our DeepFace framework provided a quantitative measurement of craniofacial-related SNPs during craniofacial development stages. We acknowledged that these six SNPs only represent a monotonic trend of regulatory role. Although specifically trained and applied to craniofacial development, DeepFace is limited by several factors, including a limited number of functional genomics datasets, low prediction performance on broad repressive marks (H3K27me3 and H3K9me3), and a lack of extensive comparison with many other tissues or development stages. Although significant SNPs with temporal effects were observed, their impacts were relatively small. It is expected that many more variants with significant correlations between SAD score trends and developmental stages can be identified. There are additional facets of SNP characteristics that warrant further exploration. This includes those SNPs with the strongest impact, those with specific influences at specific development stages or cell types, and those related to particular chromatin

features, all of which could be investigated in future studies. Therefore, we provide the OFC-sign SAD matrix (Table S2) to the research community, which is composed by SAD scores for 204 epigenomic features by 1,787 OFC--sign SNPs. Finally, as sequencing technologies are evolving quickly, we expect that many more genomics datasets will be generated, especially those by assay for transposase-accessible chromatin with sequencing (ATAC-seq) and single-cell multiome,[54] in craniofacial development. Such data will allow us to refine the DeepFace model for both accuracy and precision toward the development stages and tissue and cell types.

Conclusion

We trained a deep-learning-based model to *in silico* evaluate the SNP alleles on epigenomic alteration across human craniofacial development during embryonic stages from pcw 4 to pcw 10. Our deep learning model, DeepFace, identified that the OFC-related significant SNP set tended to have stronger SAD scores in several variant categories than other groups, suggesting that these high-risk variants manifest their functional impact during these development stages. We pinpointed six SNPs with a significant linear relationship with SAD scores across developmental progression. Those SNPs may have critical roles in OFCs, and further investigation is warranted. Our study demonstrates that DeepFace has great potential to harness the long-range regulatory element signals from comprehensive epigenomic assays and thus to prioritize, interpret, and decode the dynamic influence of variants related to OFCs and other traits.

Data and code availability

All datasets analyzed in this study are publicly available. The 204 ChIP of post-translational histone modifications from human embryonic craniofacial tissues were obtained from GEO: GSE97752. The OFC-related variants were obtained from the GWAS Catalog (https://www.ebi.ac.uk/gwas/). Multiethnic craniofacial raw data for CL/P and CP are available from dbGaP: phs000884.v2.p1. Other data can be accessed from public resources described in the material and methods. The source code for the pre-trained DeepFace model and SAD scores are available at the following GitHub repository: https://github.com/bsml320/DeepFace/.

Web resources

dbGaP, https://www.ncbi.nlm.nih.gov/gap/
DeepFace, https://github.com/bsml320/DeepFace/
DeepFun, https://bioinfo.uth.edu/deepfun/
GWAS Catalog, https://www.ebi.ac.uk/gwas/
Online Mendelian Inheritance in Man (OMIM), https://omim.org/
WebCSEA, https://bioinfo.uth.edu/webcsea/

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2024.100312.

Acknowledgments

We are thankful for the National Institutes of Health (NIH) grant R01DE030122 for supporting this project. Z.Z. was partially supported by R01DE029818, R01LM012806, U01AG079847, and Chair Professorship for Precision Health funds. T.I. and F.Y. are CPRIT Postdoctoral Fellow and Predoctoral Fellow in the Biomedical Informatics, Genomics, and Translational Cancer Research Training Program (BIG-TCR) funded by the Cancer Prevention and Research Institute of Texas (CPRIT RP210045), respectively. M.L.M. was partially supported by NIH grants R01DE016148, R37DE008559, R01DE032319, R01DE031261, R01DE031855, R01DE032122, and X01HG007845. We are thankful for the technical support from the Cancer Genomics Core, funded by the Cancer Prevention and Research Institute of Texas (CPRIT RP180734). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. We are thankful for the technical support from Luyao Chen from the Center for Secure Artificial Intelligence for Healthcare, McWilliams School of Biomedical Informatics, the University of Texas Health Science Center at Houston. We thank all members of the Bioinformatics and Systems Medicine Laboratory for discussions.

Author contributions

Z.Z., P.J., Y.D., and G.P. contributed to the conception and design of the study; Y.D., T.I., G.P., N.M., M.L.M., F.Y., and Y.C. collected the data and performed the analysis; Y.D., T.I., G.P., F.Y., Y.C., X.J., and L.M.S. interpreted the results; and Y.D., T.I., G.P., and Z.Z. wrote the manuscript. All authors read and approved the final manuscript.

Declaration of interests

The authors declare no competing interests.

Received: February 8, 2024
Accepted: May 23, 2024

References

1. Leslie, E.J., and Marazita, M.L. (2013). Genetics of cleft lip and cleft palate. Am. J. Med. Genet. C Semin. Med. Genet. *163C*, 246–258.
2. Dixon, M.J., Marazita, M.L., Beaty, T.H., and Murray, J.C. (2011). Cleft lip and palate: understanding genetic and environmental influences. Nat. Rev. Genet. *12*, 167–178.
3. Beaty, T.H., Murray, J.C., Marazita, M.L., Munger, R.G., Ruczinski, I., Hetmanski, J.B., Liang, K.Y., Wu, T., Murray, T., Fallin, M.D., et al. (2010). A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. Nat. Genet. *42*, 525–529.
4. Birnbaum, S., Ludwig, K.U., Reutter, H., Herms, S., Steffens, M., Rubini, M., Baluardo, C., Ferrian, M., Almeida de Assis, N., Alblas, M.A., et al. (2009). Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24. Nat. Genet. *41*, 473–477.

5. Camargo, M., Rivera, D., Moreno, L., Lidral, A.C., Harper, U., Jones, M., Solomon, B.D., Roessler, E., Vélez, J.I., Martinez, A.F., et al. (2012). GWAS reveals new recessive loci associated with non-syndromic facial clefting. Eur. J. Med. Genet. *55*, 510–514.
6. Grant, S.F.A., Wang, K., Zhang, H., Glaberson, W., Annaiah, K., Kim, C.E., Bradfield, J.P., Glessner, J.T., Thomas, K.A., Garris, M., et al. (2009). A genome-wide association study identifies a locus for nonsyndromic cleft lip with or without cleft palate on 8q24. J. Pediatr. *155*, 909–913.
7. Leslie, E.J., Carlson, J.C., Shaffer, J.R., Feingold, E., Wehby, G., Laurie, C.A., Jain, D., Laurie, C.C., Doheny, K.F., McHenry, T., et al. (2016). A multi-ethnic genome-wide association study identifies novel loci for non-syndromic cleft lip with or without cleft palate on 2p24.2, 17q23 and 19q13. Hum. Mol. Genet. *25*, 2862–2872.
8. Mangold, E., Ludwig, K.U., Birnbaum, S., Baluardo, C., Ferrian, M., Herms, S., Reutter, H., de Assis, N.A., Chawa, T.A., Mattheisen, M., et al. (2010). Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. Nat. Genet. *42*, 24–26.
9. Sun, Y., Huang, Y., Yin, A., Pan, Y., Wang, Y., Wang, C., Du, Y., Wang, M., Lan, F., Hu, Z., et al. (2015). Genome-wide association study identifies a new susceptibility locus for cleft lip with or without a cleft palate. Nat. Commun. *6*, 6414.
10. Wolf, Z.T., Brand, H.A., Shaffer, J.R., Leslie, E.J., Arzi, B., Willet, C.E., Cox, T.C., McHenry, T., Narayan, N., Feingold, E., et al. (2015). Genome-wide association studies in dogs and humans identify ADAMTS20 as a risk variant for cleft lip and palate. PLoS Genet. *11*, e1005059.
11. Ludwig, K.U., Mangold, E., Herms, S., Nowak, S., Reutter, H., Paul, A., Becker, J., Herberz, R., AlChawa, T., Nasser, E., et al. (2012). Genome-wide meta-analyses of nonsyndromic cleft lip with or without cleft palate identify six new risk loci. Nat. Genet. *44*, 968–971.
12. Beaty, T.H., Ruczinski, I., Murray, J.C., Marazita, M.L., Munger, R.G., Hetmanski, J.B., Murray, T., Redett, R.J., Fallin, M.D., Liang, K.Y., et al. (2011). Evidence for gene-environment interaction in a genome wide study of nonsyndromic cleft palate. Genet. Epidemiol. *35*, 469–478.
13. Leslie, E.J., Liu, H., Carlson, J.C., Shaffer, J.R., Feingold, E., Wehby, G., Laurie, C.A., Jain, D., Laurie, C.C., Doheny, K.F., et al. (2016). A genome-wide association study of nonsyndromic cleft palate identifies an etiologic missense variant in GRHL3. Am. J. Hum. Genet. *98*, 744–754.
14. Xu, H., Yan, F., Hu, R., Suzuki, A., Iwaya, C., Jia, P., Iwata, J., and Zhao, Z. (2021). CleftGeneDB: a resource for annotating genes associated with cleft lip and cleft palate. Sci. Bull. *66*, 2340–2342.
15. Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.
16. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.
17. Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., et al. (2010). The NIH Roadmap Epigenomics Mapping Consortium. Nat. Biotechnol. *28*, 1045–1048.
18. Huang, C., Thompson, P., Wang, Y., Yu, Y., Zhang, J., Kong, D., Colen, R.R., Knickmeyer, R.C., Zhu, H.; and Alzheimer's Disease Neuroimaging Initiative (2017). FGWAS: Functional genome wide association analysis. Neuroimage *159*, 107–121.
19. Hu, R., Pei, G., Jia, P., and Zhao, Z. (2021). Decoding regulatory structures and features from epigenomics profiles: A Roadmap-ENCODE Variational Auto-Encoder (RE-VAE) model. Methods *189*, 44–53.
20. Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol. *33*, 831–838.
21. Zhou, J., and Troyanskaya, O.G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Methods *12*, 931–934.
22. Kelley, D.R., Reshef, Y.A., Bileschi, M., Belanger, D., McLean, C.Y., and Snoek, J. (2018). Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Genome Res. *28*, 739–750.
23. Zhou, J., Theesfeld, C.L., Yao, K., Chen, K.M., Wong, A.K., and Troyanskaya, O.G. (2018). Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nat. Genet. *50*, 1171–1179.
24. Pei, G., Hu, R., Dai, Y., Manuel, A.M., Zhao, Z., and Jia, P. (2021). Predicting regulatory variants using a dense epigenomic mapped CNN model elucidated the molecular basis of trait-tissue associations. Nucleic Acids Res. *49*, 53–66.
25. Pei, G., Hu, R., Jia, P., and Zhao, Z. (2021). DeepFun: a deep learning sequence-based model to decipher non-coding variant effect in a tissue- and cell type-specific manner. Nucleic Acids Res. *49*, W131–W139.
26. Gasperini, M., Tome, J.M., and Shendure, J. (2020). Towards a comprehensive catalogue of validated and target-linked human enhancers. Nat. Rev. Genet. *21*, 292–310.
27. Xi, W., and Beer, M.A. (2021). Loop competition and extrusion model predicts CTCF interaction specificity. Nat. Commun. *12*, 1046.
28. Wilderman, A., VanOudenhove, J., Kron, J., Noonan, J.P., and Cotney, J. (2018). High-resolution epigenomic atlas of human embryonic craniofacial development. Cell Rep. *23*, 1581–1597.
29. Schoenwolf, G.C., Bleyl, S.B., Brauer, P.R., and Francis-West, P.H. (2021). Larsen's Human Embryology (Elsevier - Health Sciences Division).
30. Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Peng, W., Zhang, M.Q., et al. (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. Nat. Genet. *40*, 897–903.
31. Beck, D.B., Oda, H., Shen, S.S., and Reinberg, D. (2012). PR-Set7 and H4K20me1: at the crossroads of genome integrity, cell cycle, chromosome condensation, and transcription. Genes Dev. *26*, 325–337.
32. Ljungman, M., Parks, L., Hulbatte, R., and Bedi, K. (2019). The role of H3K79 methylation in transcription and the DNA damage response. Mutat. Res. Rev. Mutat. Res. *780*, 48–54.
33. Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., Groza, T., Güneş, O., Hall, P., Hayhurst, J., et al. (2023). The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. Nucleic Acids Res. *51*, D977–D985.
34. Mukhopadhyay, N., Feingold, E., Moreno-Uribe, L., Wehby, G., Valencia-Ramirez, L.C., Muñeton, C.P.R., Padilla, C., Deleyiannis, F., Christensen, K., Poletta, F.A., et al. (2021).

Genome-wide association study of non-syndromic orofacial clefts in a multiethnic sample of families and controls identifies novel regions. Front. Cell Dev. Biol. *9*, 621482.
35. Pei, G., Dai, Y., Zhao, Z., and Jia, P. (2019). deTS: tissue-specific enrichment analysis to decode tissue specificity. Bioinformatics *35*, 3842–3845.
36. Dai, Y., Hu, R., Manuel, A.M., Liu, A., Jia, P., and Zhao, Z. (2021). CSEA-DB: an omnibus for human complex trait and cell type associations. Nucleic Acids Res. *49*, D862–D870.
37. Dai, Y., Hu, R., Liu, A., Cho, K.S., Manuel, A.M., Li, X., Dong, X., Jia, P., and Zhao, Z. (2022). WebCSEA: web-based cell-type-specific enrichment analysis of genes. Nucleic Acids Res. *50*, W782–W790.
38. Jia, P., Dai, Y., Hu, R., Pei, G., Manuel, A.M., and Zhao, Z. (2019). TSEA-DB: a trait–tissue association map for human complex traits and diseases. Nucleic Acids Res. *48*, D1022–D1030.
39. Kunkle, B.W., Grenier-Boley, B., Sims, R., Bis, J.C., Damotte, V., Naj, A.C., Boland, A., Vronskaya, M., van der Lee, S.J., Amlie-Wolf, A., et al. (2019). Author Correction: Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. Nat. Genet. *51*, 1423–1424.
40. Trubetskoy, V., Pardiñas, A.F., Qi, T., Panagiotaropoulou, G., Awasthi, S., Bigdeli, T.B., Bryois, J., Chen, C.-Y., Dennison, C.A., Hall, L.S., et al. (2022). Mapping genomic loci implicates genes and synaptic biology in schizophrenia. Nature *604*, 502–508.
41. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.
42. Kheradpour, P., and Kellis, M. (2014). Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Res. *42*, 2976–2987.
43. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.
44. Castro-Mondragon, J.A., Riudavets-Puig, R., Rauluseviciute, I., Lemma, R.B., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Manosalva Pérez, N., et al. (2022). JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Res. *50*, D165–D173.
45. Santiago, L., Daniels, G., Wang, D., Deng, F.-M., and Lee, P. (2017). Wnt signaling pathway protein LEF1 in cancer, as a biomarker for prognosis and a target for treatment. Am. J. Cancer Res. *7*, 1389–1406.
46. Yankee, T.N., Oh, S., Winchester, E.W., Wilderman, A., Robinson, K., Gordon, T., Rosenfeld, J.A., VanOudenhove, J., Scott, D.A., Leslie, E.J., et al. (2023). Integrative analysis of transcriptome dynamics during human craniofacial development identifies candidate disease genes. Nat. Commun. *14*, 4623.
47. Roberts, R.M., Loh, K.M., Amita, M., Bernardo, A.S., Adachi, K., Alexenko, A.P., Schust, D.J., Schulz, L.C., Telugu, B.P.V.L., Ezashi, T., et al. (2014). Differentiation of trophoblast cells from human embryonic stem cells: to be or not to be? J. Reprod. Fertil. *147*, D1–D12.
48. Roth, D.M., Bayona, F., Baddam, P., and Graf, D. (2021). Craniofacial development: Neural crest in molecular embryology. Head Neck Pathol. *15*, 1–15.
49. Mort, R.L., Jackson, I.J., and Patton, E.E. (2015). The melanocyte lineage in development and disease. Development *142*, 620–632.
50. Asrar, H., and Tucker, A.S. (2022). Endothelial cells during craniofacial development: Populating and patterning the head. Front. Bioeng. Biotechnol. *10*, 962040.
51. Lewis, A.E., Hwa, J., Wang, R., Soriano, P., and Bush, J.O. (2015). Neural crest defects in ephrin-B2 mutant mice are non-autonomous and originate from defects in the vasculature. Dev. Biol. *406*, 186–195.
52. Wiszniak, S., Mackenzie, F.E., Anderson, P., Kabbara, S., Ruhrberg, C., and Schwarz, Q. (2015). Neural crest cell-derived VEGF promotes embryonic jaw extension. Proc. Natl. Acad. Sci. USA *112*, 6086–6091.
53. Sun, B., Liu, Y., Huang, W., Zhang, Q., Lin, J., Li, W., Zhang, J., and Chen, F. (2021). Functional identification of a rare vascular endothelial growth factor a (VEGFA) variant associating with the nonsyndromic cleft lip with/without cleft palate. Bioengineered *12*, 1471–1483.
54. Yan, F., Suzuki, A., Iwaya, C., Pei, G., Chen, X., Yoshioka, H., Yu, M., Simon, L.M., Iwata, J., and Zhao, Z. (2024). Single-cell multiomics decodes regulatory programs for mouse secondary palate development. Nat. Commun. *15*, 1–17.
55. Haaland, Ø.A., Romanowska, J., Gjerdevik, M., Lie, R.T., Gjessing, H.K., and Jugessur, A. (2019). A genome-wide scan of cleft lip triads identifies parent-of-origin interaction effects between ANK3 and maternal smoking, and between ARHGEF10 and alcohol consumption. F1000Res. *8*, 960. https://doi.org/10.12688/f1000research.19571.2.
56. Yu, Y., Zuo, X., He, M., Gao, J., Fu, Y., Qin, C., Meng, L., Wang, W., Song, Y., Cheng, Y., et al. (2017). Genome-wide analyses of non-syndromic cleft lip with palate identify 14 novel loci and genetic heterogeneity. Nat. Commun. *8*, 14364. https://doi.org/10.1038/ncomms14364.
57. Chi, N., and Epstein, J.A. (2002). Getting your Pax straight: Pax proteins in development and disease. Trends Genet. *18*, 41–47.
58. Murdoch, B., DelConte, C., and García-Castro, M.I. (2012). Pax7 lineage contributions to the mammalian neural crest. PLoS One *7*, e41089.
59. Gaczkowska, A., Biedziak, B., Budner, M., Zadurska, M., Lasota, A., Hozyasz, K.K., Dąbrowska, J., Wójcicki, P., Szponar-Żurowska, A., Żukowski, K., et al. (2019). PAX7 nucleotide variants and the risk of non-syndromic orofacial clefts in the Polish population. Oral Dis. *25*, 1608–1618.
60. Khan, M.I., Cs, P., and Srinath, N. (2022). Role of PAX7 gene rs766325 and rs4920520 polymorphisms in the etiology of non-syndromic cleft lip and palate: A genetic study. Glob. Med. Genet. *9*, 208–211.
61. van Genderen, C., Okamura, R.M., Fariñas, I., Quo, R.G., Parslow, T.G., Bruhn, L., and Grosschedl, R. (1994). Development of several organs that require inductive epithelial-mesenchymal interactions is impaired in LEF-1-deficient mice. Genes Dev. *8*, 2691–2703.
62. Roël, G., Gent, Y.Y.J., Peterson-Maduro, J., Verbeek, F.J., and Destrée, O. (2009). Lef1 plays a role in patterning the mesoderm and ectoderm in Xenopus tropicalis. Int. J. Dev. Biol. *53*, 81–89.
63. Shu, X., Shu, S., and Cheng, H. (2019). Genome-wide mRNA-seq profiling reveals that LEF1 and SMAD3 regulate epithelial-mesenchymal transition through the Hippo signaling pathway during palatal fusion. Genet. Test. Mol. Biomarkers *23*, 197–203.
64. Lee, S.-U., and Maeda, T. (2012). POK/ZBTB proteins: an emerging family of proteins that regulate lymphoid development and function. Immunol. Rev. *247*, 107–119.

65. Siggs, O.M., and Beutler, B. (2012). The BTB-ZF transcription factors. Cell Cycle *11*, 3358–3369.
66. Takebayashi-Suzuki, K., Konishi, H., Miyamoto, T., Nagata, T., Uchida, M., and Suzuki, A. (2018). Coordinated regulation of the dorsal-ventral and anterior-posterior patterning of Xenopus embryos by the BTB/POZ zinc finger protein Zbtb14. Dev. Growth Differ. *60*, 158–173.
67. Suzuki, A., Sangani, D.R., Ansari, A., and Iwata, J. (2016). Molecular mechanisms of midfacial developmental defects. Dev. Dynam. *245*, 276–293.
68. Itoh, M., Furuse, M., Morita, K., Kubota, K., Saitou, M., and Tsukita, S. (1999). Direct binding of three tight junction-associated MAGUKs, ZO-1, ZO-2, and ZO-3, with the COOH termini of claudins. J. Cell Biol. *147*, 1351–1363.
69. Kiener, T.K., Selptsova-Friedrich, I., and Hunziker, W. (2008). Tjp3/zo-3 is critical for epidermal barrier function in zebrafish embryos. Dev. Biol. *316*, 36–49.
70. Iklé, J.M., Tavares, A.L.P., King, M., Ding, H., Colombo, S., Firulli, B.A., Firulli, A.B., Targoff, K.L., Yelon, D., and Clouthier, D.E. (2017). Nkx2.5 regulates endothelin converting enzyme-1 during pharyngeal arch patterning. Genesis *55*, e23021.
71. Funato, N., and Nakamura, M. (2017). Identification of shared and unique gene families associated with oral clefts. Int. J. Oral Sci. *9*, 104–109.
72. Jain, P., Karthikeyan, C., Moorthy, N.S.H.N., Waiker, D.K., Jain, A.K., and Trivedi, P. (2014). Human CDC2-like kinase 1 (CLK1): a novel target for Alzheimer's disease. Curr. Drug Targets *15*, 539–550.
73. Virgirinia, R.P., Nakamura, M., Takebayashi-Suzuki, K., Fatchiyah, F., and Suzuki, A. (2021). The dual-specificity protein kinase Clk3 is essential for Xenopus neural development. Biochem. Biophys. Res. Commun. *567*, 99–105.
74. Sukhatme, V.P., Cao, X.M., Chang, L.C., Tsai-Morris, C.H., Stamenkovich, D., Ferreira, P.C., Cohen, D.R., Edwards, S.A., Shows, T.B., and Curran, T. (1988). A zinc finger-encoding gene coregulated with c-fos during growth and differentiation, and after cellular depolarization. Cell *53*, 37–43.
75. Mcmahon, A.P., Champion, J.E., Mcmahon, J.A., and Sukhatme, V.P. (1990). Developmental expression of the putative transcription factor *Egr-1* suggests that *Egr-1* and *c-fos* are coregulated in some tissues. Development *108*, 281–287.
76. Yan, F., Jia, P., Yoshioka, H., Suzuki, A., Iwata, J., and Zhao, Z. (2020). A developmental stage-specific network approach for studying dynamic co-regulation of transcription factors and microRNAs during craniofacial development. Development *147*, dev192948.
77. Hirano, T., Tsuruda, T., Tanaka, Y., Harada, H., Yamazaki, T., and Ishida, A. (2021). Long noncoding RNA CCDC26 as a modulator of transcriptional switching between fetal and embryonic globins. Biochim. Biophys. Acta Mol. Cell Res. *1868*, 118931.
78. Yildirim, M., Seymen, F., Deeley, K., Cooper, M.E., and Vieira, A.R. (2012). Defining predictors of cleft lip and palate risk. J. Dent. Res. *91*, 556–561.
79. Mostowska, A., Hozyasz, K.K., Wojcicki, P., Biedziak, B., Paradowska, P., and Jagodzinski, P.P. (2010). Association between genetic variants of reported candidate genes or regions and risk of cleft lip with or without cleft palate in the polish population. Birth Defects Res. A Clin. Mol. Teratol. *88*, 538–545.
80. Boehringer, S., van der Lijn, F., Liu, F., Günther, M., Sinigerova, S., Nowak, S., Ludwig, K.U., Herberz, R., Klein, S., Hofman, A., et al. (2011). Genetic determination of human facial morphology: links between cleft-lips and normal variation. Eur. J. Hum. Genet. *19*, 1192–1197.
81. Zenz, R., Eferl, R., Scheinecker, C., Redlich, K., Smolen, J., Schonthaler, H.B., Kenner, L., Tschachler, E., and Wagner, E.F. (2008). Activator protein 1 (Fos/Jun) functions in inflammatory bone and skin disease. Arthritis Res. Ther. *10*, 201.
82. Maili, L., Tandon, B., Yuan, Q., Menezes, S., Chiu, F., Hashmi, S.S., Letra, A., Eisenhoffer, G.T., and Hecht, J.T. (2023). Disruption of fos causes craniofacial anomalies in developing zebrafish. Front. Cell Dev. Biol. *11*, 1141893.
83. Wang, B., Xu, M., Zhao, J., Yin, N., Wang, Y., and Song, T. (2023). Single-cell transcriptomics reveals activation of macrophages in all-trans retinoic acid (atRA)-induced cleft palate. J. Craniofac. Surg. *35*, 177–184. https://doi.org/10.1097/SCS.0000000000009782.
84. Kondo, S., Schutte, B.C., Richardson, R.J., Bjork, B.C., Knight, A.S., Watanabe, Y., Howard, E., de Lima, R.L.L.F., Daack-Hirsch, S., Sander, A., et al. (2002). Mutations in IRF6 cause Van der Woude and popliteal pterygium syndromes. Nat. Genet. *32*, 285–289.
85. Schutte, B.C., Saal, H.M., Goudy, S., and Leslie, E.J. (2021). IRF6-Related Disorders (University of Washington, Seattle).
86. Zucchero, T.M., Cooper, M.E., Maher, B.S., Daack-Hirsch, S., Nepomuceno, B., Ribeiro, L., Caprau, D., Christensen, K., Suzuki, Y., Machida, J., et al. (2004). Interferon regulatory factor 6 (IRF6) gene variants and the risk of isolated cleft lip or palate. N. Engl. J. Med. *351*, 769–780.

HGGA, Volume *5*

Supplemental information

DeepFace: Deep-learning-based framework to contextualize orofacial-cleft-related variants during human embryonic craniofacial development

Yulin Dai, Toshiyuki Itai, Guangsheng Pei, Fangfang Yan, Yan Chu, Xiaoqian Jiang, Seth M. Weinberg, Nandita Mukhopadhyay, Mary L. Marazita, Lukas M. Simon, Peilin Jia, and Zhongming Zhao

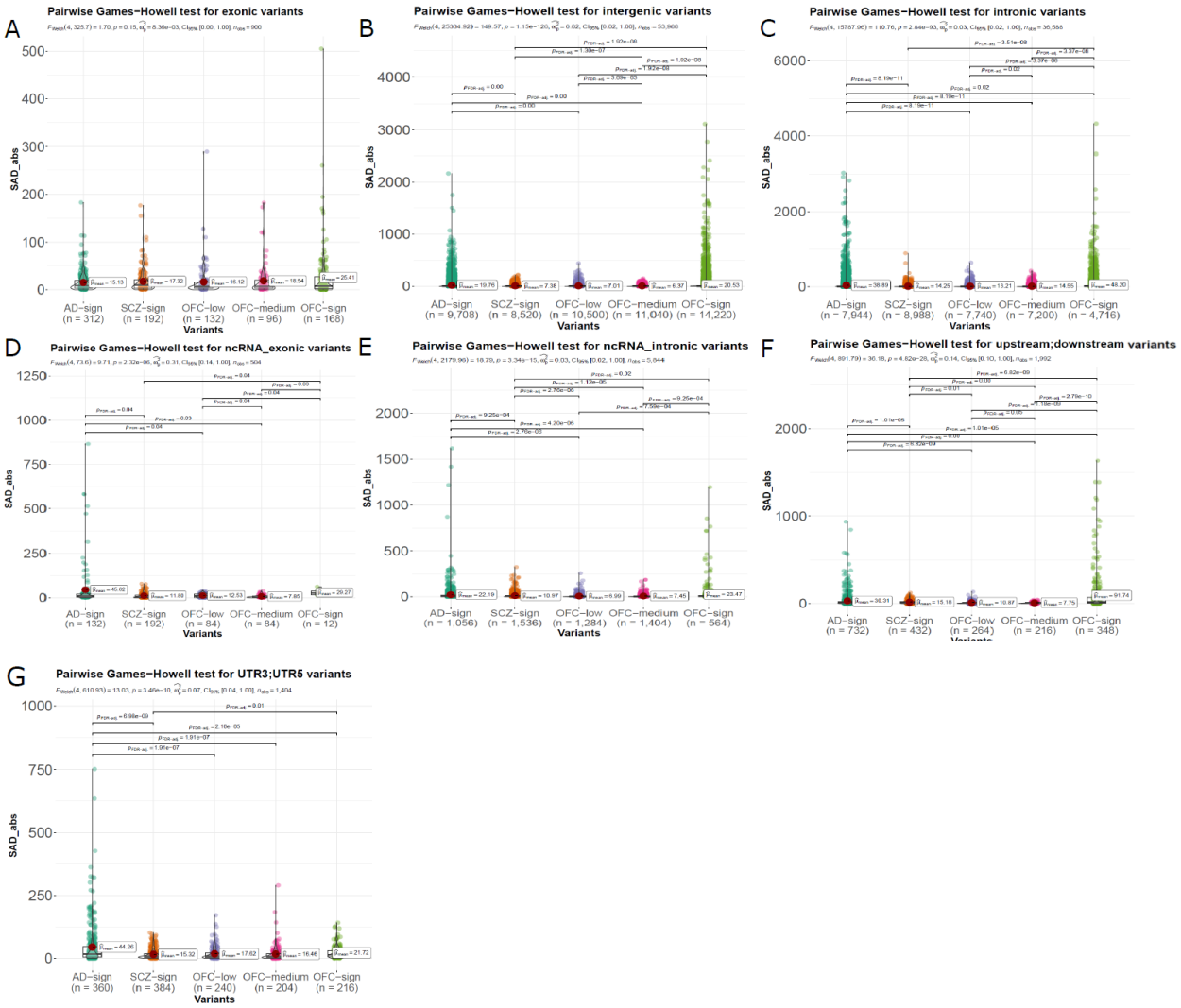

Figure S1. Pairwise Games-Howell test for seven variant categories across five different SNP sets

Panels (A) to (G) depict the results of Welch's F-test from ANOVA, conducted to determine whether there is a mean difference across comparison groups. Subsequently, the non-parametric Games-Howell test was used to evaluate whether pairwise mean rank differences exist in the absolute SAD scores between the five variant categories: AD-sign, SCZ-sign, OFC-low, OFC-medium, and OFC-sign. In each sub-panel, only the comparison groups with significant P_{FDR}-adjusted p-values are highlighted.

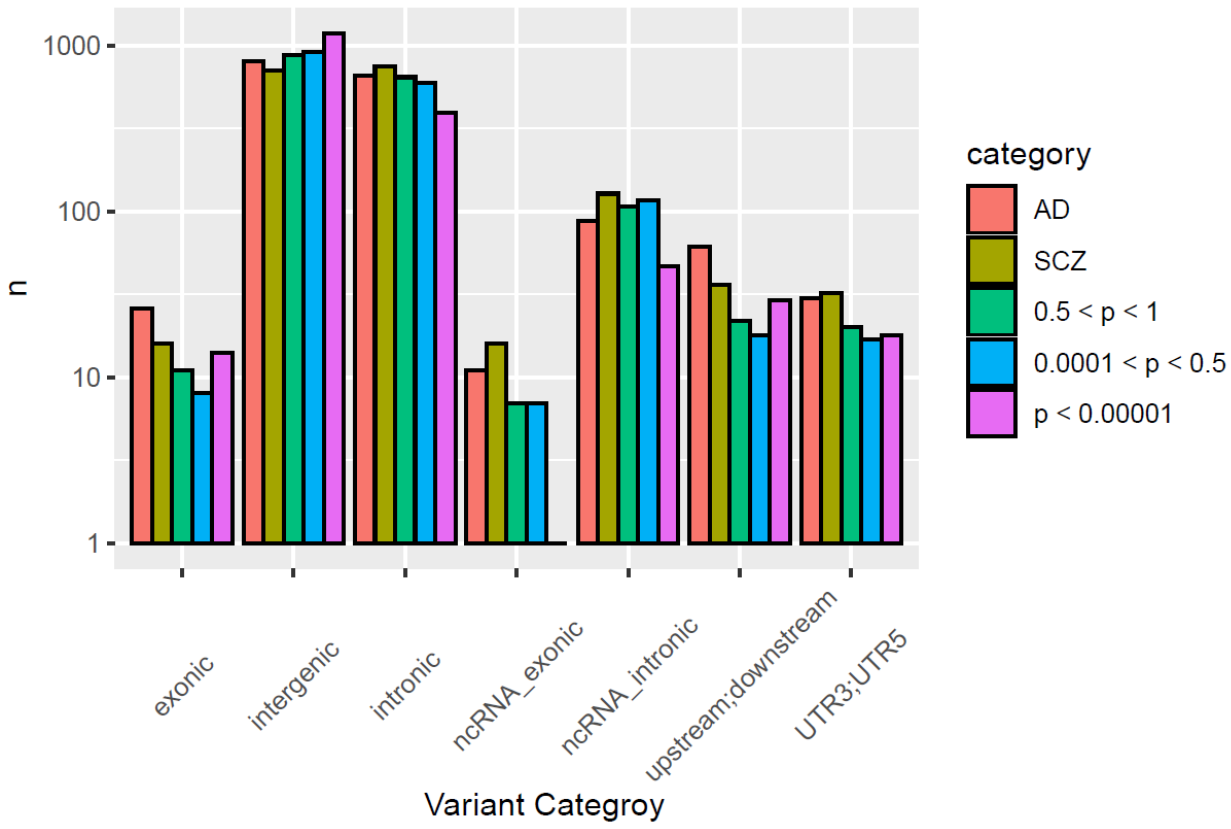

Figure S2. Number of SNP by variant category for five different SNP sets

Number of variants from five different categories AD-sign, SCZ-sign, OFC-low, OFC-medium, and OFC-sign stratified by variant category.

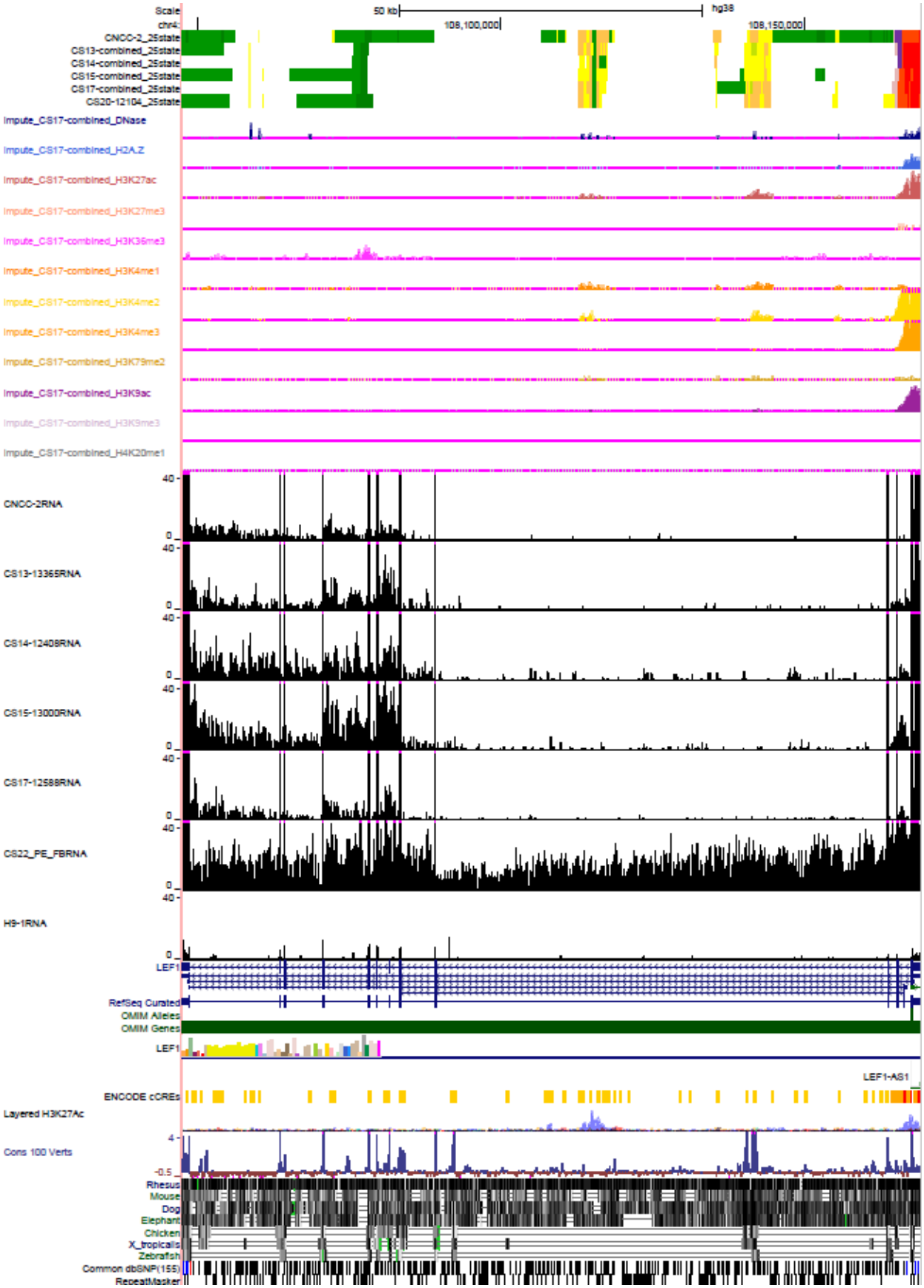
Scale
50 kb
hg38
chr4:
108,100,000
108,150,000
CNCC-2_25state
CS13-combined_25state
CS14-combined_25state
CS15-combined_25state
CS17-combined_25state
CS20-12104_25state
Impute_CS17-combined_DNase
Impute_CS17-combined_H2A.Z
Impute_CS17-combined_H3K27ac
Impute_CS17-combined_H3K27me3
Impute_CS17-combined_H3K36me3
Impute_CS17-combined_H3K4me1
Impute_CS17-combined_H3K4me2
Impute_CS17-combined_H3K4me3
Impute_CS17-combined_H3K79me2
Impute_CS17-combined_H3K9ac
Impute_CS17-combined_H3K9me3
Impute_CS17-combined_H4K20me1
CNCC-2RNA
CS13-13365RNA
CS14-12408RNA
CS15-13000RNA
CS17-12588RNA
CS22_PE_FBRNA
H9-1RNA
40
0
LEF1
RefSeq Curated
OMIM Alleles
OMIM Genes
LEF1
LEF1-AS1
ENCODE cCREs
Layered H3K27Ac
Cons 100 Verts
4
-0.5
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Common dbSNP(155)
RepeatMasker

Figure S3. LEF1 in UCSC from Cotney lab craniofacial Genome Browser for transcriptome and epigenomic features for craniofacial tissue from Carnegie stages CS 13 to CS 22.

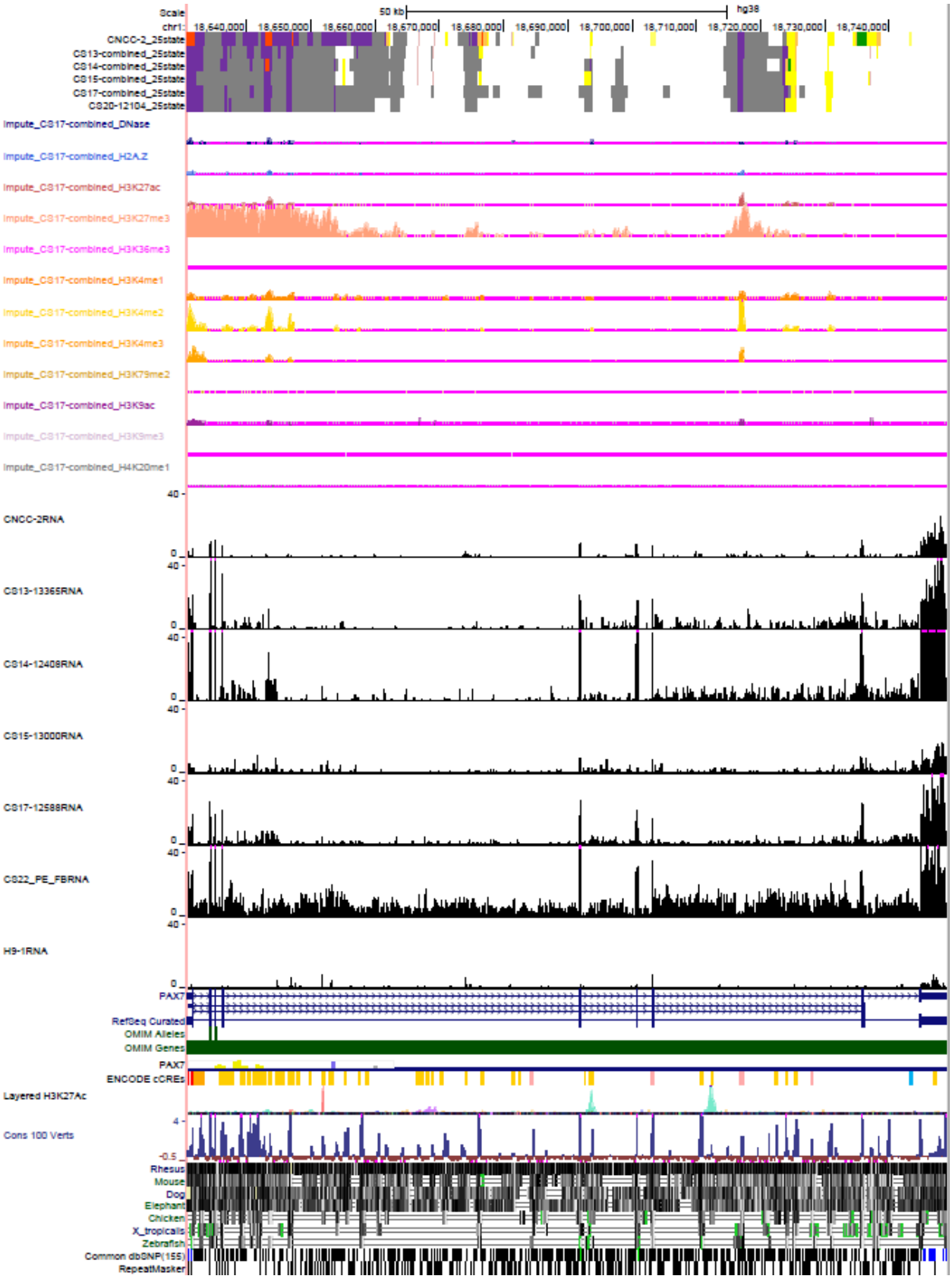
Scale
50 kb
hg38
chr1: 18,640,000 18,650,000 18,660,000 18,670,000 18,680,000 18,690,000 18,700,000 18,710,000 18,720,000 18,730,000 18,740,000
CNCC-2_25state
CS13-combined_25state
CS14-combined_25state
CS15-combined_25state
CS17-combined_25state
CS20-12104_25state
Impute_CS17-combined_DNase
Impute_CS17-combined_H2A.Z
Impute_CS17-combined_H3K27ac
Impute_CS17-combined_H3K27me3
Impute_CS17-combined_H3K36me3
Impute_CS17-combined_H3K4me1
Impute_CS17-combined_H3K4me2
Impute_CS17-combined_H3K4me3
Impute_CS17-combined_H3K79me2
Impute_CS17-combined_H3K9ac
Impute_CS17-combined_H3K9me3
Impute_CS17-combined_H4K20me1
CNCC-2RNA
CS13-13365RNA
CS14-12408RNA
CS15-13000RNA
CS17-12588RNA
CS22_PE_FBRNA
H9-1RNA
PAX7
RefSeq Curated
OMIM Alleles
OMIM Genes
PAX7
ENCODE cCREs
Layered H3K27Ac
Cons 100 Verts
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Common dbSNP(155)
RepeatMasker

Figure S4. PAX7 in UCSC from Cotney lab craniofacial Genome Browser for transcriptome and epigenomic features for craniofacial tissue from Carnegie stages CS 13 to CS 22.

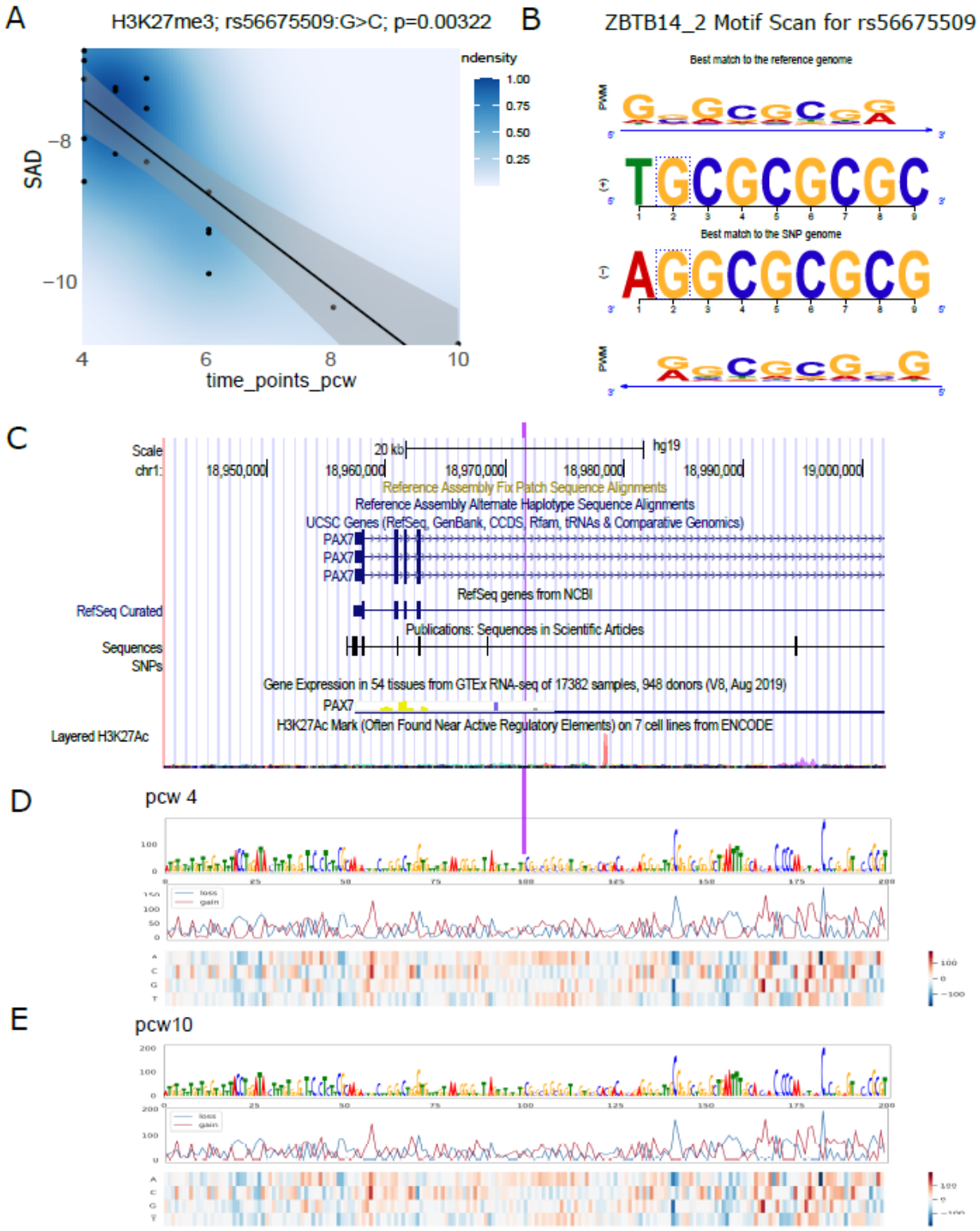

Figure S5. Motif analysis of DeepFace variant rs56675509.

(A) SNP activity difference (SAD) for rs56675509 (G>C) associated with post-conception weeks (pcw) increase for H3K27me3. The blue area of intensity is positively correlated with the SAD score point density. The black line and grey confidence interval (95%) model the linear relationship between SAD scores and pcw. P-value indicates the possibility of null hypothesis that the coefficient is equal to zero is true. (B) Sequence logo stacks from top to bottom: sequence logo of reference allele matching position weight matrix; Reference subsequences, alternative allele subsequences, and sequence logo of SNP allele matching position weight matrix. The best match reference sequence and alternative allele sequence for the motif ZBTB14 were visualized. (C) UCSC genome browser for the rs56675509 and its surrounding gene. The purple vertical line indicates the exact genomic region of 200 bps for (D) and (E). (D) & (E) show the dynamic gain and loss of SAD score for all possible substitutions in each of the 200-bp genomic positions around the rs56675509 in pcw 4 and pcw 10, respectively. The alteration between D & E is relatively small in figure. These SAD score dynamics were visualized in three ways: 1) sequence logo weight by the loss of SAD across 200-bp sequence; 2) The blue and red lines indicate the minimum (loss) and maximum (gain) change among the possible substitutions from reference allele; 3) The quantities in the heatmap display the change in SAD after substituting nucleotide from reference allele.

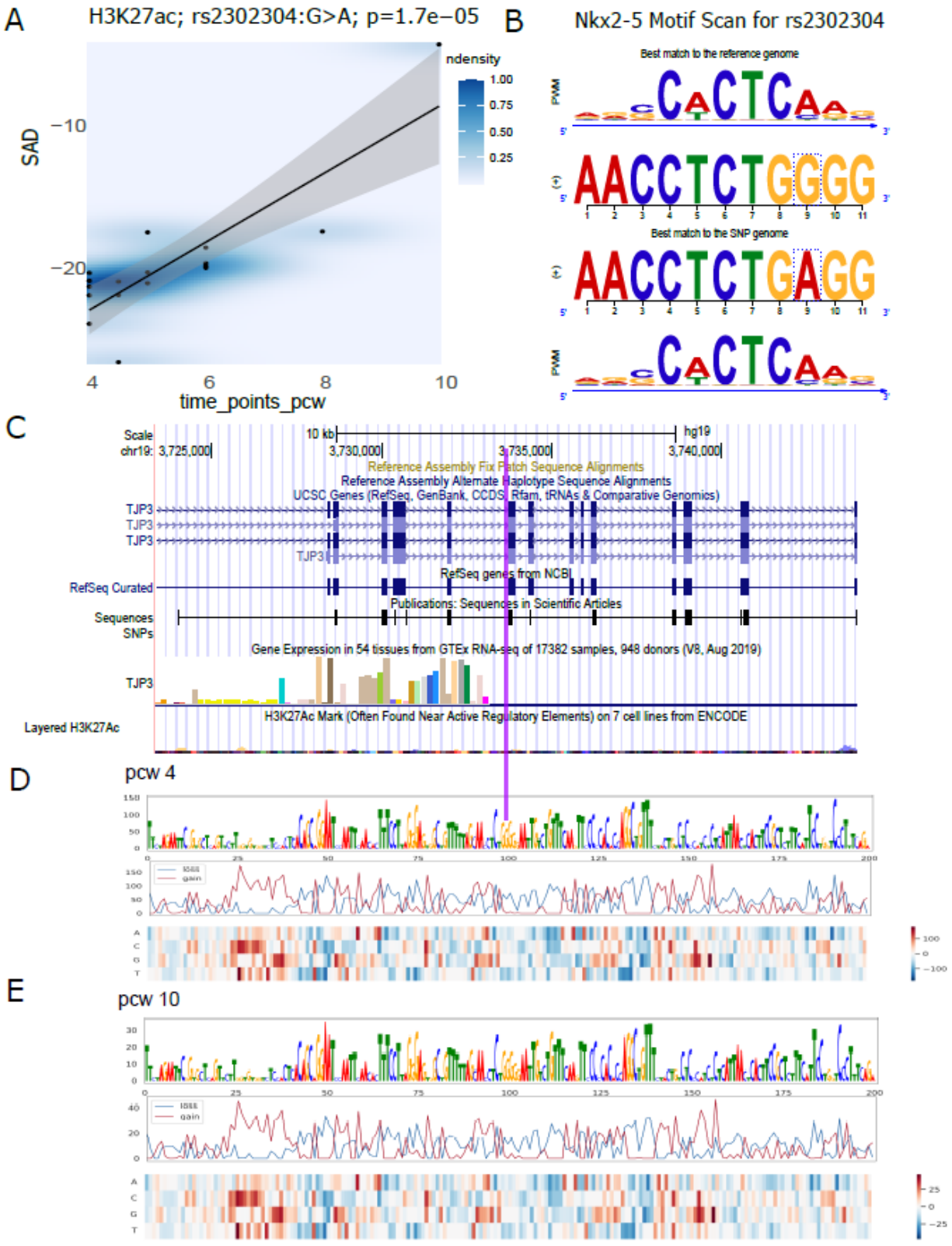

Figure S6. Motif analysis of DeepFace variant rs2302304.

(A) SNP activity difference (SAD) for rs2302304 (G>A) associated with post-conception weeks (pcw) increase for H3K27ac. The blue area of intensity is positively correlated with the SAD score point density. The black line and grey confidence interval (95%) model the linear relationship between SAD scores and pcw. P-value indicates the possibility of null hypothesis that the coefficient is equal to zero is true. (B) Sequence logo stacks from top to bottom: sequence logo of reference allele matching position weight matrix; Reference subsequences, alternative allele subsequences, and sequence logo of SNP allele matching position weight matrix. The best match reference sequence and alternative allele sequence for the motif Nkx2-5 were visualized (C) UCSC genome browser for the rs2302304 and its surrounding gene. The purple vertical line indicates the exact genomic region of 200 bps for (D) and (E). (D) & (E) show the dynamic gain and loss of SAD score for all possible substitutions in each of the 200-bp genomic positions around the rs2302304 in pcw 4 and pcw 10, respectively. The alteration between D & E is relatively small in figure.These SAD score dynamics were visualized in three ways: 1) sequence logo weight by the loss of SAD across 200-bp sequence; 2) The blue and red lines indicate the minimum (loss) and maximum (gain) change among the possible substitutions from reference allele; 3) The quantities in the heatmap display the change in SAD after substituting nucleotide from reference allele.

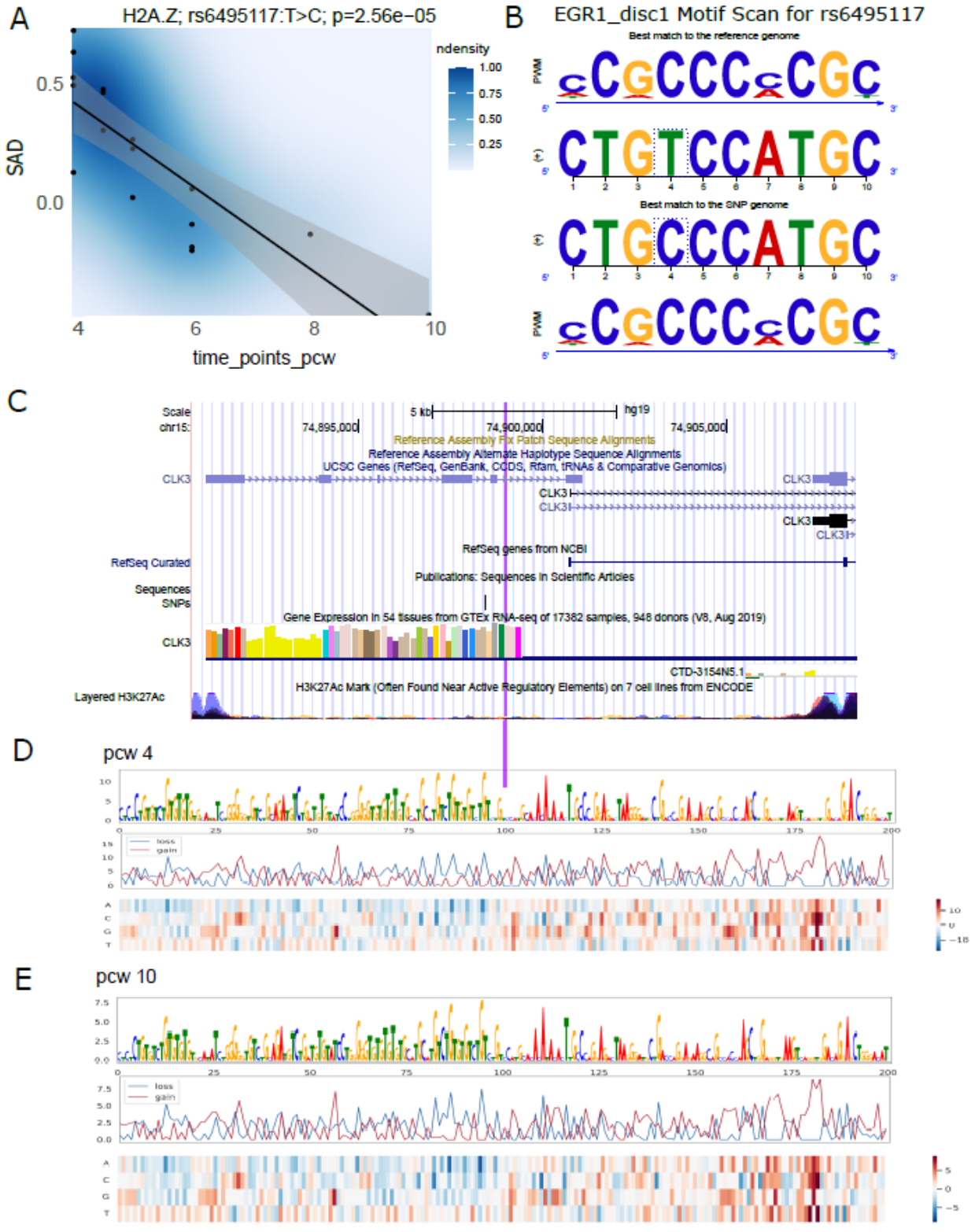

Figure S7. Motif analysis of DeepFace variant rs6495117.

(A) SNP activity difference (SAD) for rs6495117 (T>C) associated with post-conception weeks (pcw) increase for H2A.Z. The blue area of intensity is positively correlated with the SAD score point density. The black line and grey confidence interval (95%) model the linear relationship between SAD scores and pcw. P-value indicates the possibility of null hypothesis that the coefficient is equal to zero is true. (B) Sequence logo stacks from top to bottom: sequence logo of reference allele matching position weight matrix; Reference subsequences, alternative allele subsequences, and sequence logo of SNP allele matching position weight matrix. The best match reference sequence and alternative allele sequence for the motif EGR1 were visualized (C) UCSC genome browser for the rs6495117 and its surrounding gene. The purple vertical line indicates the exact genomic region of 200 bps for (D) and (E). (D) & (E) show the dynamic gain and loss of SAD score for all possible substitutions in each of the 200-bp genomic positions around the rs6495117 in pcw 4 and pcw 10, respectively. The alteration between D & E is relatively small in figure. These SAD score dynamics were visualized in three ways: 1) sequence logo weight by the loss of SAD across 200-bp sequence; 2) The blue and red lines indicate the minimum (loss) and maximum (gain) change among the possible substitutions from reference allele; 3) The quantities in the heatmap display the change in SAD after substituting nucleotide from reference allele.

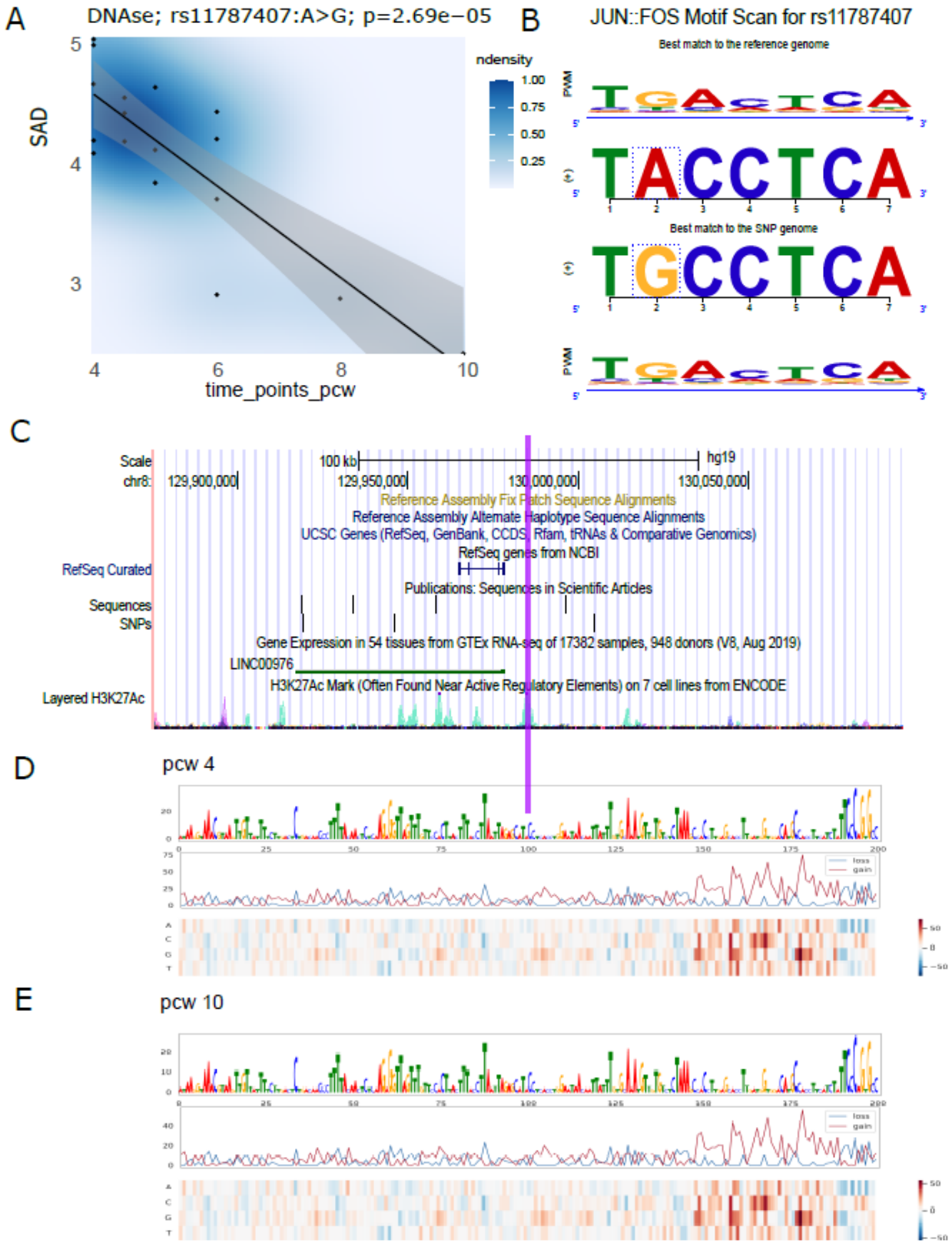

Figure S8. Motif analysis of DeepFace variant rs11787407.

(A) SNP activity difference (SAD) for rs11787407 (A>G) associated with post-conception weeks (pcw) increase for DNase. The blue area of intensity is positively correlated with the SAD score point density. The black line and grey confidence interval (95%) model the linear relationship between SAD scores and pcw. P-value indicates the possibility of null hypothesis that the coefficient is equal to zero is true. (B) Sequence logo stacks from top to bottom: sequence logo of reference allele matching position weight matrix; Reference subsequences, alternative allele subsequences, and sequence logo of SNP allele matching position weight matrix. The best match reference sequence and alternative allele sequence for the motif JUN/FOS were visualized (C) UCSC genome browser for the rs11787407 and its surrounding gene. The purple vertical line indicates the exact genomic region of 200 bps for (D) and (E). (D) & (E) show the dynamic gain and loss of SAD score for all possible substitutions in each of the 200-bp genomic positions around the rs11787407 in pcw 4 and pcw 10, respectively. These SAD score dynamics were visualized in three ways: 1) sequence logo weight by the loss of SAD across 200-bp sequence; 2) The blue and red lines indicate the minimum (loss) and maximum (gain) change among the possible substitutions from reference allele; 3) The quantities in the heatmap display the change in SAD after substituting nucleotide from reference allele.

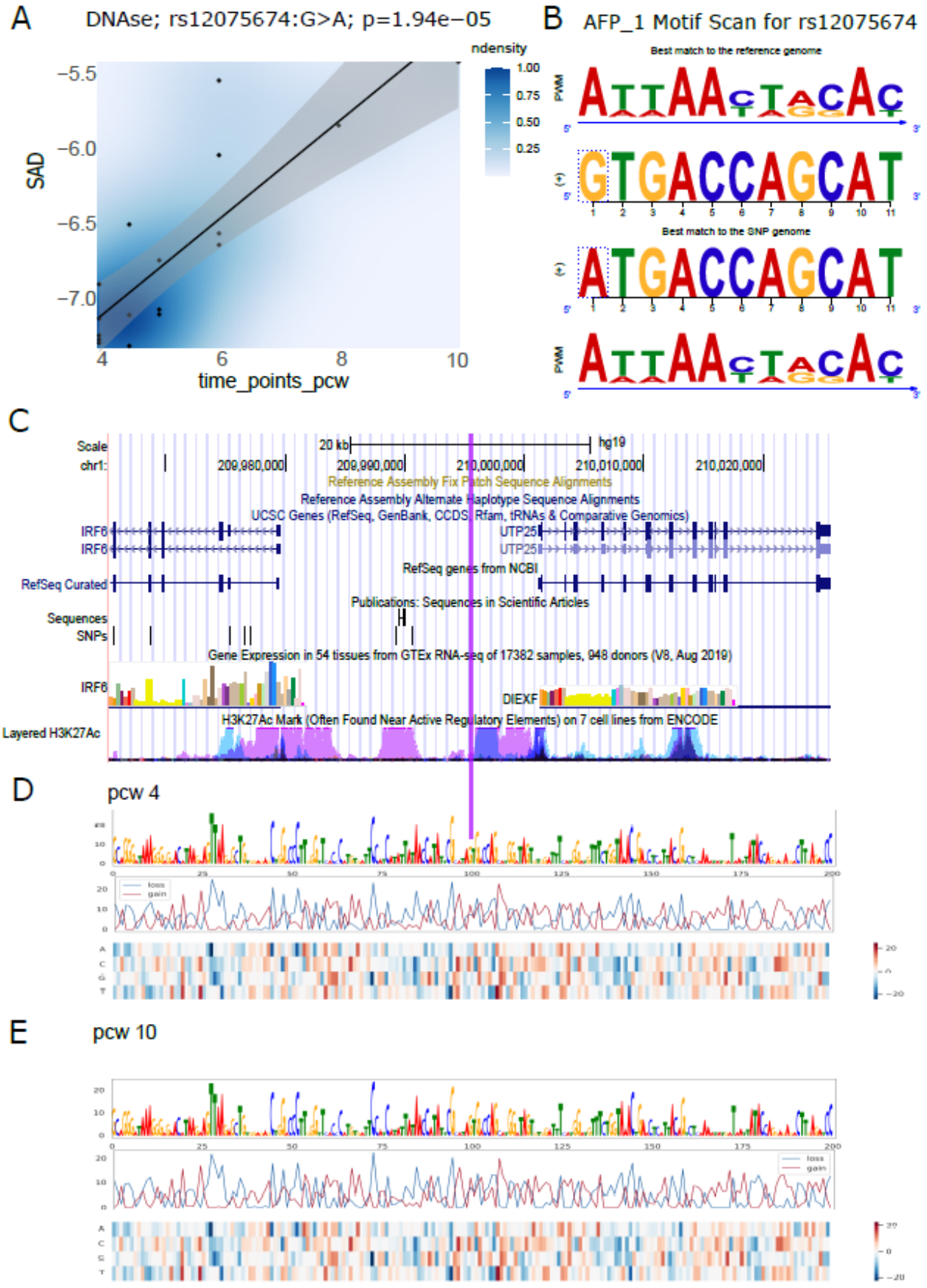
A
DNAse; rs12075674:G>A; p=1.94e−05
ndensity
1.00
0.75
0.50
0.25
SAD
−5.5
−6.0
−6.5
−7.0
4
6
8
10
time_points_pcw
B
AFP_1 Motif Scan for rs12075674
Best match to the reference genome
PWM
(+)
Best match to the SNP genome
C
Scale
chr1:
20 kb
hg19
209,980,000
209,990,000
210,000,000
210,010,000
210,020,000
Reference Assembly Fix Patch Sequence Alignments
Reference Assembly Alternate Haplotype Sequence Alignments
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
IRF6
IRF6
UTP25
UTP25
RefSeq genes from NCBI
RefSeq Curated
Publications: Sequences in Scientific Articles
Sequences
SNPs
Gene Expression in 54 tissues from GTEx RNA-seq of 17382 samples, 948 donors (V8, Aug 2019)
IRF6
DIEXF
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
D
pcw 4
loss
gain
E
pcw 10

Figure S9. Motif analysis of DeepFace variant rs12075674.

(A) SNP activity difference (SAD) for rs12075674 (G>A) associated with post-conception weeks (pcw) increase for DNase. The blue area of intensity is positively correlated with the SAD score point density. The black line and grey confidence interval (95%) model the linear relationship between SAD scores and pcw. P-value indicates the possibility of null hypothesis that the coefficient is equal to zero is true. (B) Sequence logo stacks from top to bottom: sequence logo of reference allele matching position weight matrix; Reference subsequences, alternative allele subsequences, and sequence logo of SNP allele matching position weight matrix. The best match reference sequence and alternative allele sequence for the motif AFP were visualized (C) UCSC genome browser for the rs12075674 and its surrounding gene. The purple vertical line indicates the exact genomic region of 200 bps for (D) and (E). (D) & (E) show the dynamic gain and loss of SAD score for all possible substitutions in each of the 200-bp genomic positions around the rs12075674 in pcw 4 and pcw 10, respectively. These SAD score dynamics were visualized in three ways: 1) sequence logo weight by the loss of SAD across 200-bp sequence; 2) The blue and red lines indicate the minimum (loss) and maximum (gain) change among the possible substitutions from reference allele; 3) The quantities in the heatmap display the change in SAD after substituting nucleotide from reference allele.