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

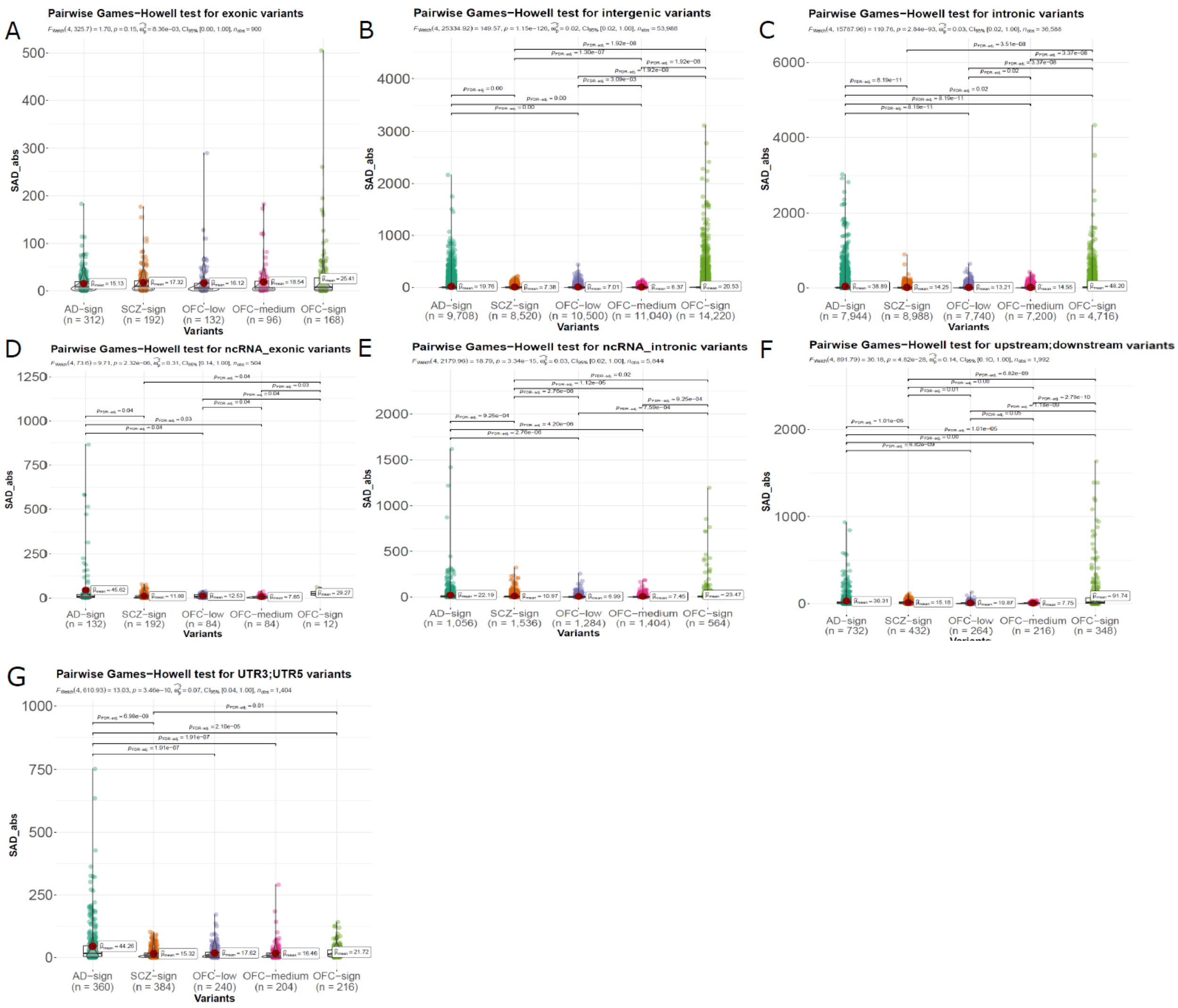

Figure S1. Pairwise Games-Howell test for seven variant categories across five different SNP sets

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.

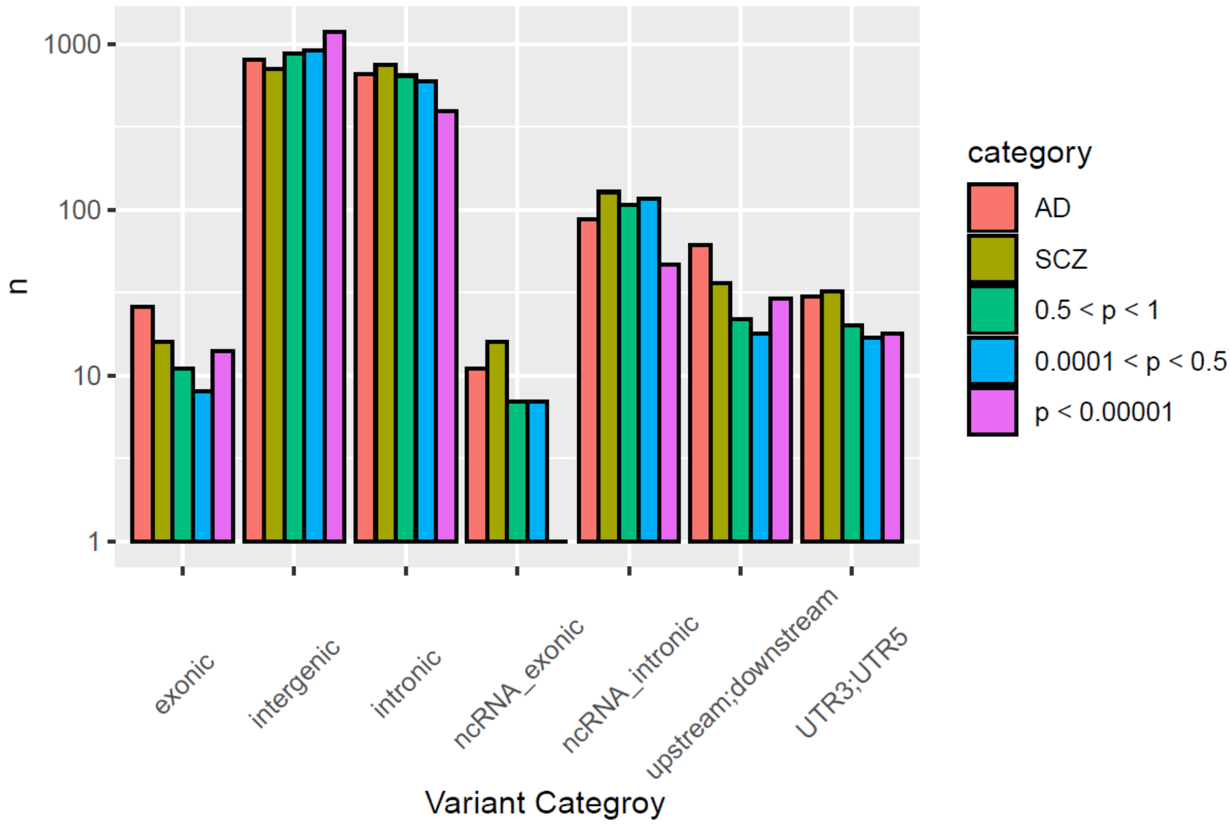

Figure S2. Number of SNP by variant category for five different SNP sets

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

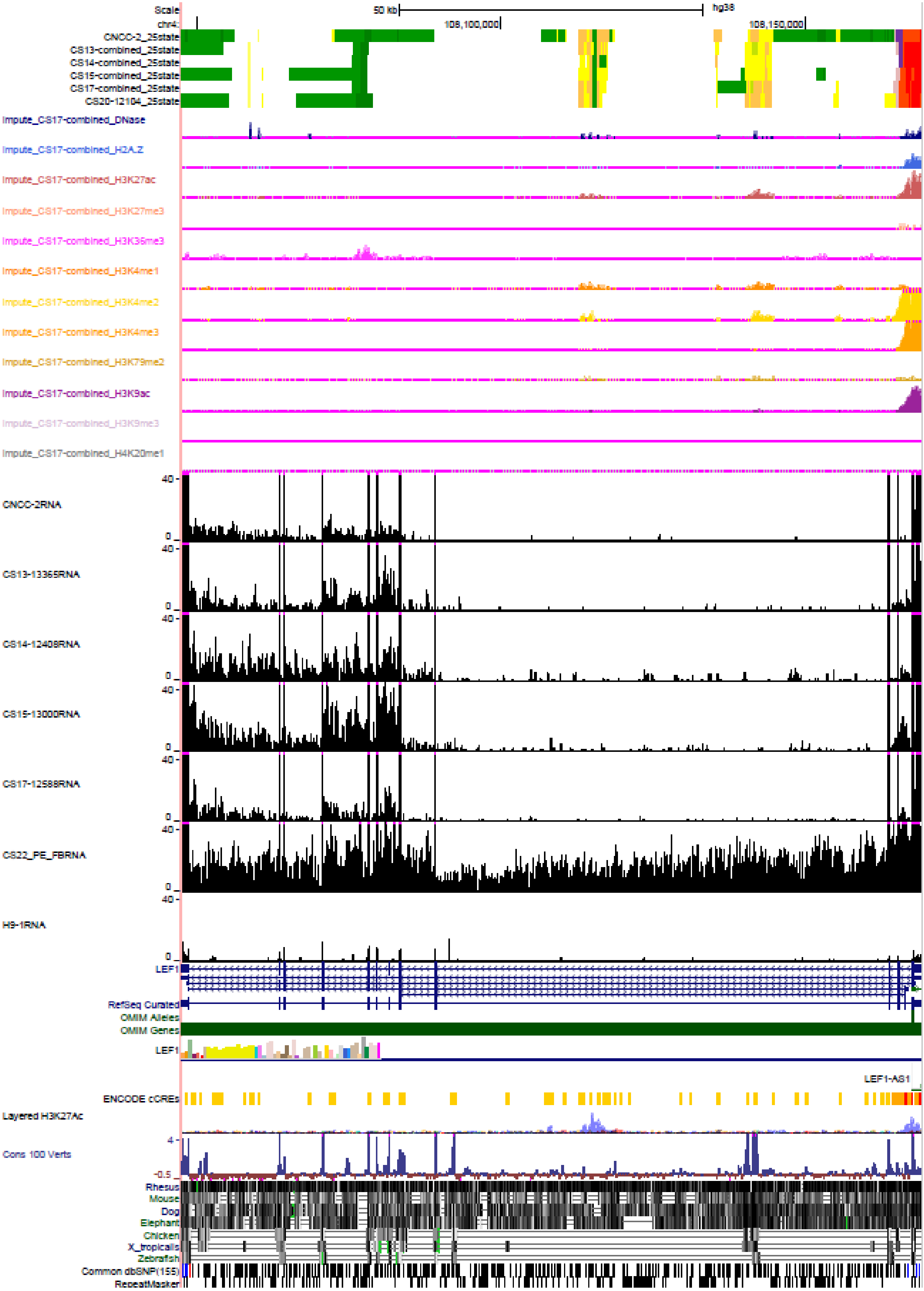
Scale
50 kb
hg38
chr4:
108,100,000
108,150,000
CNCC-2_25state
CS13-combined_25state
CS14-combined_25state
CS15-combined_25state
CS17-combined_25state
CS20-12104_25state
Impute_CS17-combined_DNase
Impute_CS17-combined_H2A.Z
Impute_CS17-combined_H3K27ac
Impute_CS17-combined_H3K27me3
Impute_CS17-combined_H3K36me3
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
40
0
CNCC-2RNA
40
0
CS13-13365RNA
40
0
CS14-12408RNA
40
0
CS15-13000RNA
40
0
CS17-12588RNA
40
0
CS22_PE_FBRNA
40
0
H9-1RNA
LEF1
RefSeq Curated
OMIM Alleles
OMIM Genes
LEF1
LEF1-AS1
ENCODE cCREs
Layered H3K27Ac
4
-0.5
Cons 100 Verts
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.

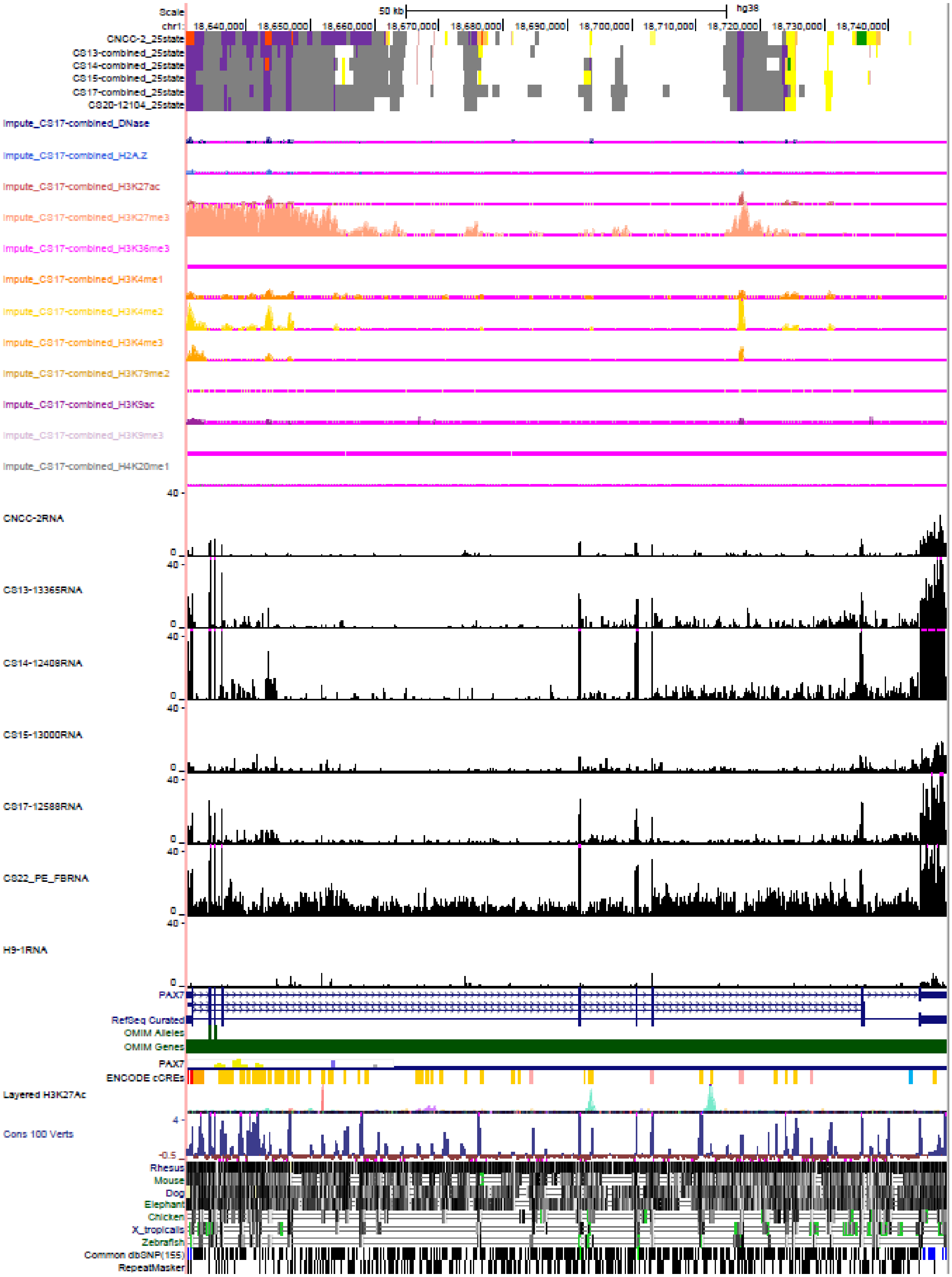
Scale
50 kb
hg38
chr1:
18,640,000
18,650,000
18,660,000
18,670,000
18,680,000
18,690,000
18,700,000
18,710,000
18,720,000
18,730,000
18,740,000
CNCC-2_25state
CS13-combined_25state
CS14-combined_25state
CS15-combined_25state
CS17-combined_25state
CS20-12104_25state
impute_CS17-combined_DNase
impute_CS17-combined_H2A.Z
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
40
CNCC-2RNA
0
40
CS13-13365RNA
0
40
CS14-12408RNA
0
40
CS15-13000RNA
0
40
CS17-12588RNA
0
40
CS22_PE_FBRNA
0
40
H9-1RNA
0
PAX7
RefSeq Curated
OMIM Alleles
OMIM Genes
PAX7
ENCODE cCREs
Layered H3K27Ac
4
Cons 100 Verts
-0.5
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.

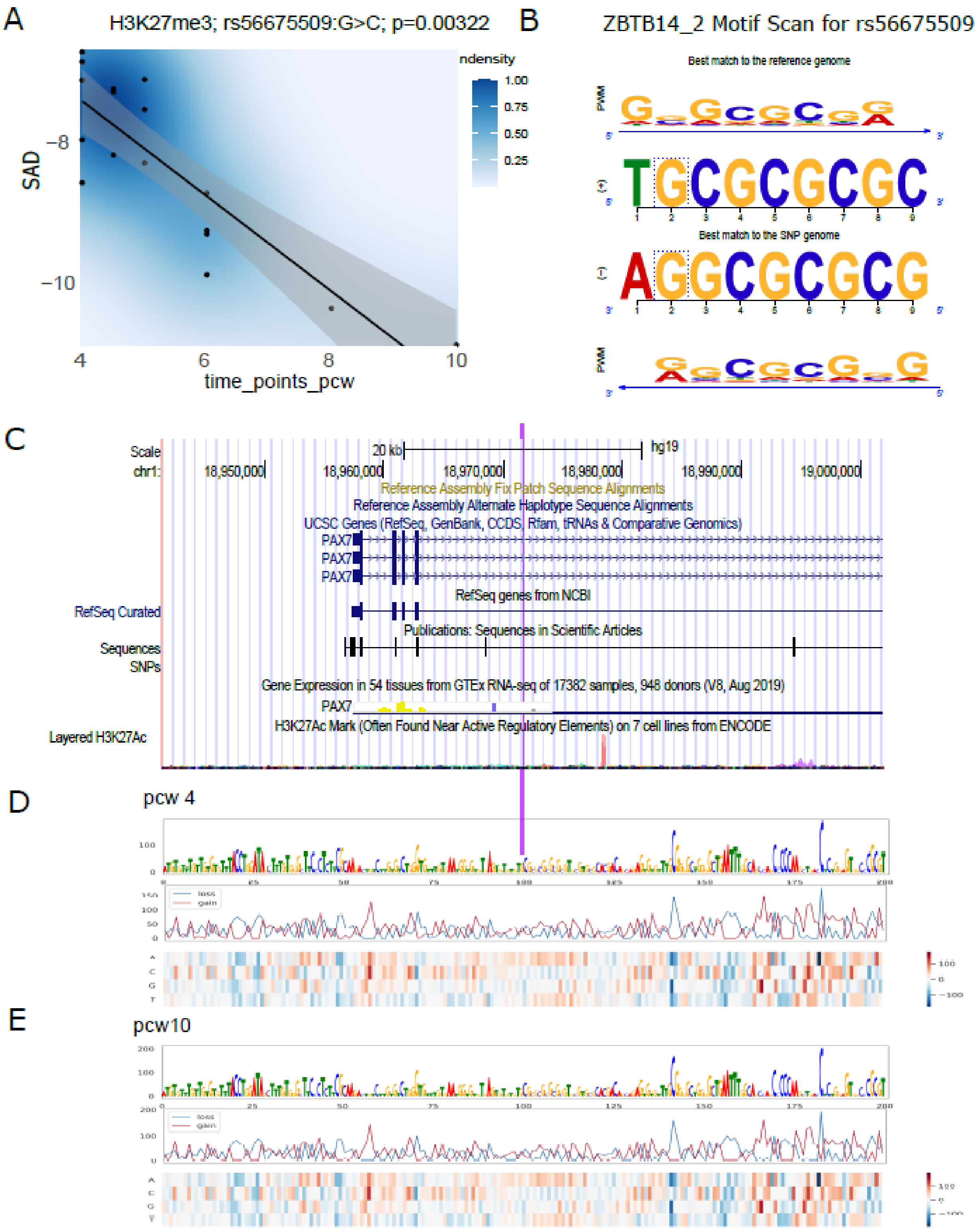

Figure S5. Motif analysis of DeepFace variant rs56675509.

(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.

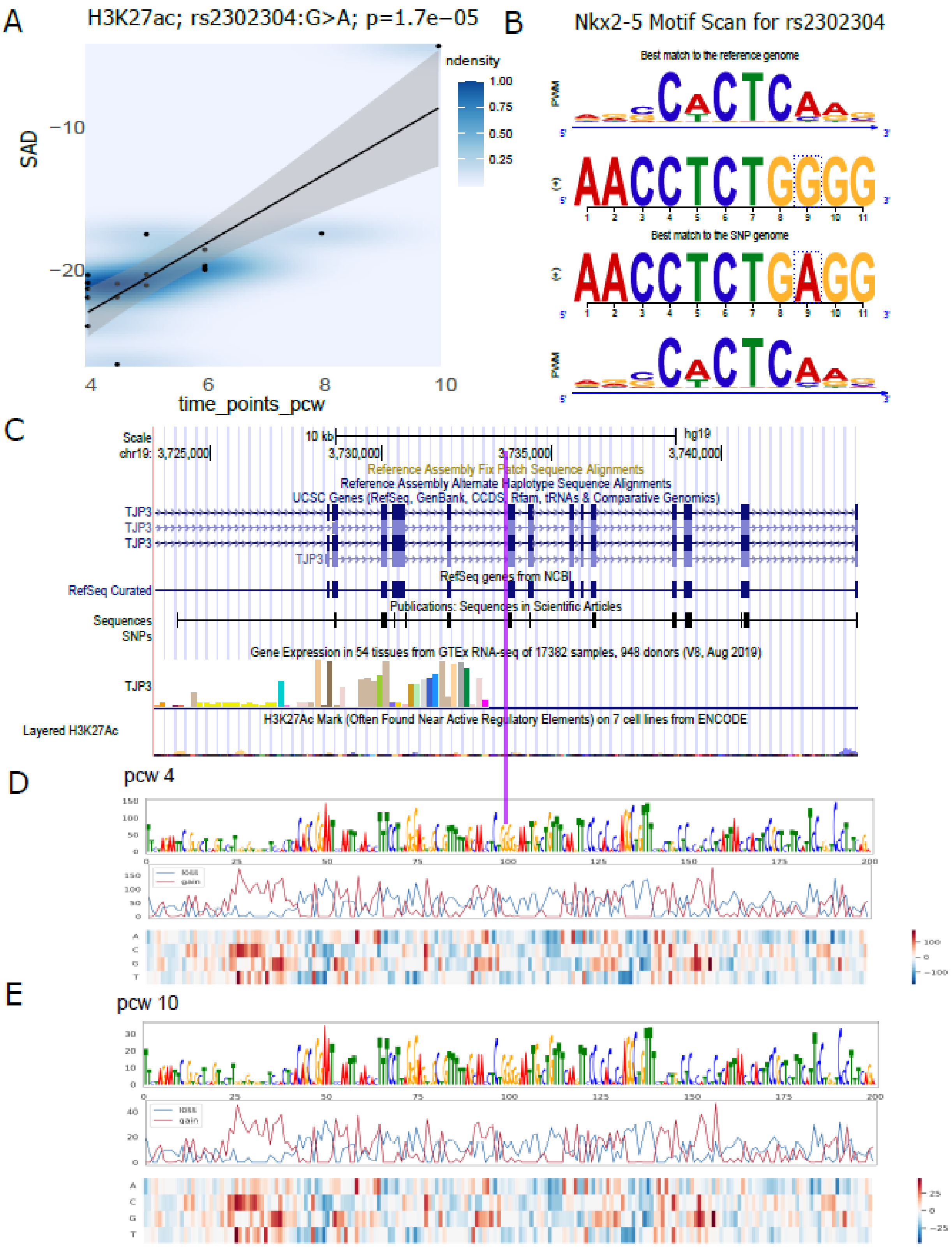

Figure S6. Motif analysis of DeepFace variant rs2302304.

(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.

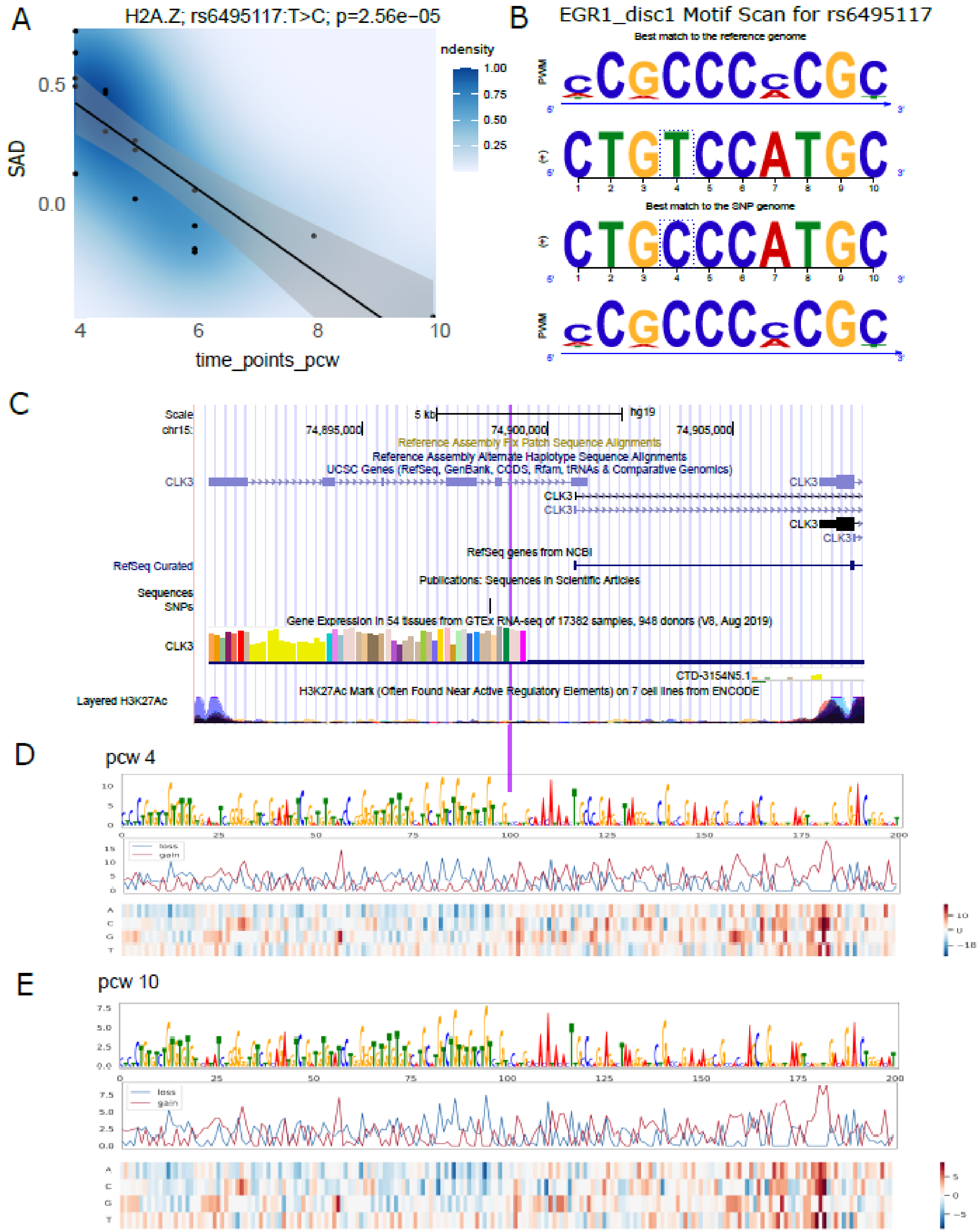

Figure S7. Motif analysis of DeepFace variant rs6495117.

(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.

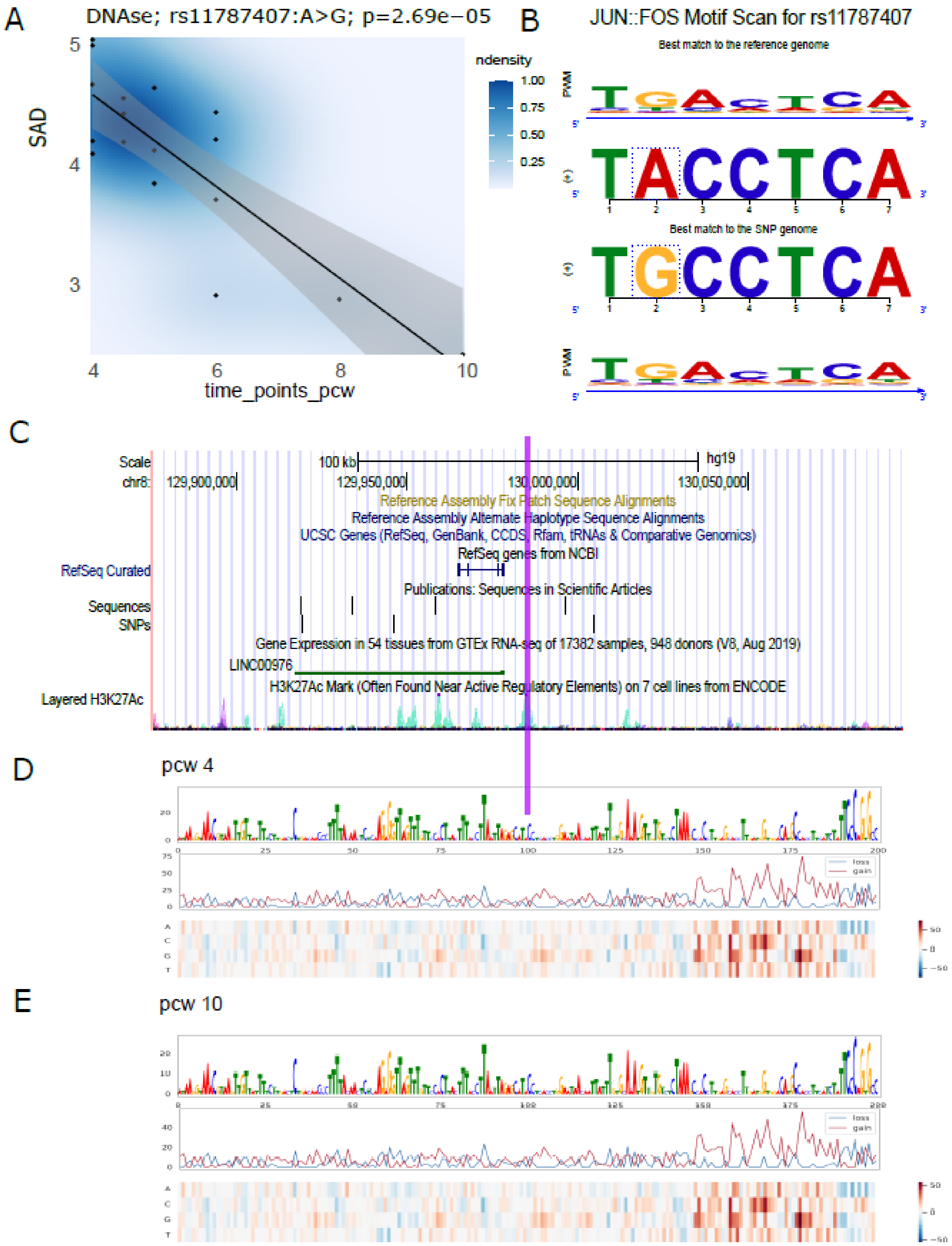

Figure S8. Motif analysis of DeepFace variant rs11787407.

(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.

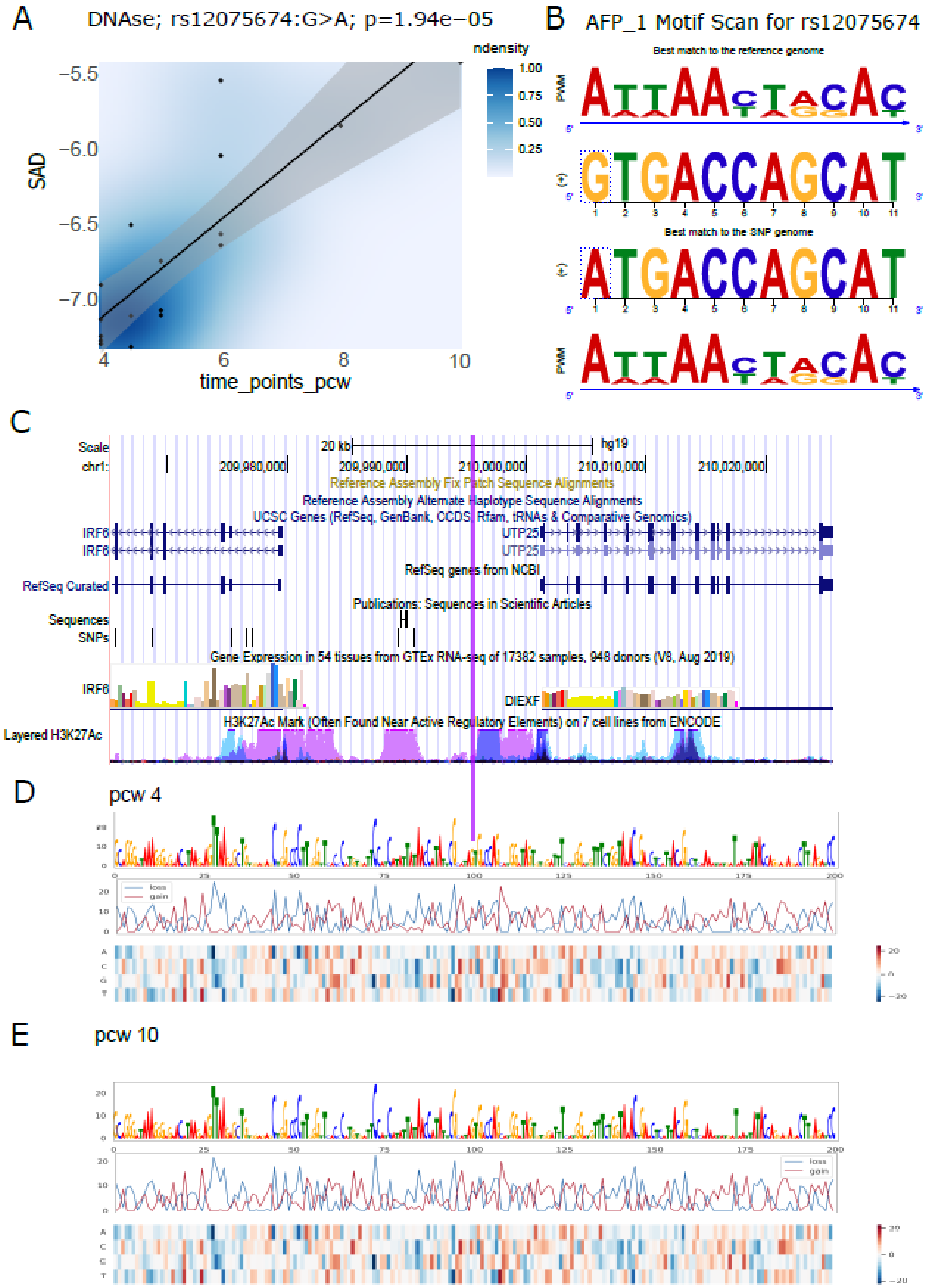
A
DNAse; rs12075674:G>A; p=1.94e−05
ndensity
1.00
0.75
0.50
0.25
SAD
−5.5
−6.0
−6.5
−7.0
4
6
8
10
time_points_pcw
B
AFP_1 Motif Scan for rs12075674
Best match to the reference genome
PWM
(+)
GTGACCAGCAT
Best match to the SNP genome
ATGACCAGCAT
5'
3'
1
2
3
4
5
6
7
8
9
10
11
C
Scale
20 kb
hg19
chr1:
209,980,000
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
0
25
50
75
100
125
150
175
200
loss
gain
A
C
G
T
20
0
−20
E
pcw 10
20
10
0
loss
gain

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.