Analysis of zebrafish periderm enhancers facilitates identification of a regulatory variant near human *KRT8/18*.

Huan Liu[1,2,3,+], Kaylia Duncan[4], Annika Helverson[2], Priyanka Kumari[2], Camille Mumm[2], Yao Xiao[1], Jenna Carlson[5], Fabrice Darbellay[6], Axel Visel[6,10,11], Elizabeth Leslie[7], Patrick Breheny[8], Albert Erives[9], Robert A. Cornell[2,4,+]

[1] State Key Laboratory Breeding Base of Basic Science of Stomatology (Hubei-MOST) and Key Laboratory for Oral Biomedicine of Ministry of Education (KLOBM), School and Hospital of Stomatology, Wuhan University, Wuhan 430079, China.

[2] Department of Anatomy and Cell Biology, University of Iowa, Iowa City, Iowa

[3] Department of Periodontology, School of Stomatology, Wuhan University, Wuhan, China

[4] Interdisciplinary Program in Molecular Medicine, University of Iowa, Iowa City, Iowa

[5] Department of Biostatistics, University of Pittsburgh

[6] Environmental Genomics and Systems Biology Division, Lawrence Berkeley Laboratories, Berkeley, California

[7] Department of Human Genetics, Emory University School of Medicine

[8] Department of Biostatistics, University of Iowa, Iowa City, Iowa

[9] Department of Biology, University of Iowa, Iowa City, Iowa

[10] U.S. Department of Energy Joint Genome Institute, Lawrence Berkeley Laboratories, Berkeley, California

[11] University of California at Merced, Merced, California

+, Authors for correspondence: robert-cornell@uiowa.edu, liu.huan@whu.edu.cn

Acknowledgements: This work supported by grants from the NIH including DE023575 (RAC), DE027362 (RAC), DE025060 (EJL), DE024427 (AV), DE028599 (AV), and NCI P30CA086862 (University of Iowa Flow Cytometry Facility), from the National Natural Science Foundation of China (NO. 81771057; 81400477) (HL), the Natural Science Foundation of Hubei Province (NO. 2017CFB515) (HL) and the Young Elite Scientist Sponsorship Program by CAST (NO. 2017QNRC001) (HL). Research at Lawrence Berkeley National Laboratory was performed under Department of Energy Contract DE-AC02-05CH11231, University of California.

Abstract

Genome wide association studies for non-syndromic orofacial cleft (OFC) have identified single nucleotide polymorphisms (SNPs) at loci where the presumed risk-relevant gene is expressed in oral periderm. The functional subsets of such SNPs are difficult to predict because the sequence underpinnings of periderm enhancers are unknown. We applied ATAC-seq to models of human palate periderm, including zebrafish periderm, mouse embryonic palate epithelia, and a human oral epithelium cell line, and to complementary mesenchymal cell types. We identified sets of enhancers specific to the epithelial cells and trained gapped-kmer support-vector-machine classifiers on these sets. We used the classifiers to predict the effect of 14 OFC-associated SNPs at 12q13 near *KRT18*. All the classifiers picked the same SNP as having the strongest effect, but the significance was highest with the classifier trained on zebrafish periderm. Reporter and deletion analyses support this SNP as lying within a periderm enhancer regulating *KRT18*/*KRT8* expression.

Introduction

Orofacial clefting (OFC), which can include cleft lip, cleft palate, or both, is among the most common of structural birth defects. Concordance for non-syndromic cleft lip with or without cleft palate (NSCLP) is about 50% in monozygotic twins, suggesting a strong genetic contribution to its etiology [1; 2]. Genome-wide association studies (GWAS) have advanced our understanding of this contribution as multiple independent GWAS, and meta-analyses of them, have identified more than 40 associated loci [3]. However, GWAS methods cannot distinguish SNPs that directly influence risk (i.e., functional SNPs) from those merely in linkage disequilibrium with such SNPs (i.e., rider SNPs). Functional SNPs that lie in non-coding DNA may alter the activity of tissue-specific enhancers. Such functional SNPs can be identified through systematic reporter assays: functional SNPs will have allele-specific effects on the enhancer activity of the encompassing element, while rider SNPs may not have this quality [4]. A bioinformatics-based approach to predicting functional SNPs requires training a machine-learning classifier on a set of enhancers of the type that the SNPs are expected to impact. The trained classifier can subsequently be used to score any element for its similarity to the training set, and also to evaluate the effect of a SNP on the score of the element containing it. The gapped-kmer support vector machine (gkmSVM) is a version of supervised machine learning based on the enrichment weights of all 10-mers in a training set[5; 6]. Using gkmSVM, a SNP's impact on the score of a 19-base-pair (bp) element encompassing it is called its deltaSVM: a SNP that alters a 10-mer with a significant weight will have a larger deltaSVM than one that does not do so [7]. Importantly, deltaSVM values were found to correlate reasonably well with the effect of SNPs on enhancer activity in reporter assays[7]. Therefore, the deltaSVM value prioritizes disease-associated SNPs for their likelihood of being functional. A challenge for the investigator can be acquiring an appropriate set of enhancers upon which to train the classifier.

A meta-analysis of OFC GWASs of several populations, and a single OFC GWAS of Han Chinese identified lead SNPs adjacent to the *KRT8* and *KRT18* genes [8; 9]. These genes are highly expressed in periderm, a simple squamous epithelium that comprises the most superficial layer of embryonic skin and oral epithelium, among other tissues [10 11; 12]. We reasoned that functional SNP (or SNPs) at this locus may disrupt a periderm enhancer. Here we revisit the meta-analysis results and find an additional 13 SNPs in this locus that have at least a suggestive statistical association to OFC. To prioritize these SNPs for functional tests using the deltaSVM method it would be optimal to train a classifier on a set of enhancers that are active in human palate-shelf periderm. To our knowledge, no periderm cell line, from any species, exists. However, primary periderm cells are readily isolated from zebrafish embryos. Although the DNA sequences of tissue-specific enhancers are rarely overtly conserved between mammals and fish, enrichment for specific transcription factor binding sites can be a conserved feature [13-15]. If the binding site features of zebrafish periderm enhancers are conserved with those of human periderm enhancers then a classifier trained on the former could be used to conduct a successful deltaSVM-based screen for SNPs that disrupt the latter.

This approach is promising in that the genetic pathways that underlie periderm development in mice and zebrafish are shared, in spite of the fact that the embryonic origins of periderm in the two species are distinct. In mouse embryos, the periderm develops from an epithelium that expresses the basal-keratinocyte marker p63 at embryonic day 9 (E9), well after gastrulation is complete [16]. In zebrafish embryos the periderm (initially called the enveloping layer [EVL]) develops from superficial blastomeres and becomes a distinct lineage at about 4 hours post fertilization (hpf), shortly before gastrulation [17]. EVL cells proliferate and cover the entire animal until at least 7 days post fertilization (dpf); by 30 dpf, the periderm is replaced by a periderm-like epithelium derived from basal keratinocytes [18]. The gene regulatory networks that govern differentiation of the murine periderm and zebrafish EVL share a dependence on IRF6 (Irf6 in zebrafish) and CHUK (Ikk1 in zebrafish)[16; 19; 20]. Differentiation of the zebrafish EVL also depends on Grhl paralogs, Klf17, and simple-epithelium keratins (e.g., Cyt1, Krt4, Krt8, Krt18) [21-24]. We predict that differentiation of murine and human periderm is similarly controlled as genes encoding orthologs of these proteins are implicated in risk for orofacial clefting (e.g., *GRHL3*, *KLF4*, and *KRT18*) [8; 22; 25; 26]. These findings, and the relative ease of isolating zebrafish periderm through cell sorting [21], motivated us to study zebrafish periderm enhancers.

Here we report on our identification of a set of zebrafish periderm enhancer candidates, an evaluation of enriched sequence transcription factor binding sites, and reporter assays with binding-site deletion analyses. Interestingly, the enriched binding sites indicate novel members of the periderm gene regulatory network. For comparison, we used the same method to identify sets of enhancer candidates in mouse palate epithelium and in a human oral epithelium cell line. We trained gkmSVM classifiers on each, and used them in deltaSVM to prioritize OFC-associated SNPs the *KRT18* gene. Finally, we tested the top candidate in reporter assays.

Results:

Identification of periderm-specific enhancers throughout the zebrafish genome

In *Tg(krt4:gfp)*^{gz7TG} transgenic embryos GFP was reported to be present exclusively in the most superficial layer of the embryo, called the enveloping layer (EVL) or periderm, after 8 hours post fertilization (hpf) [27]. Because a separate transgenic line built from the same element shows reporter expression in both basal and superficial epidermal layers at 54 hpf [28], we sectioned *Tg(krt4:gfp)* ^{gz7TG} embryos and confirmed that GFP was only present at high levels in the periderm (**Figure 1A**). We dissociated such embryos at 11 hpf (4-somite stage), isolated GFP-positive and GFP-negative cells, and performed ATAC-seq on both populations (**Figure 1A**). We then mapped the small (<100bp) ATAC-seq fragments, which are indicative of nucleosome free regions (NFRs), within the zebrafish genome [29]. The concordance of peaks called between two replicates of this experiment was strong (**Figure 1—figure supplement 1**). At the majority of NFRs, the density of mapped reads was comparable in GFP-positive and GFP-negative cells, but at about 5% of elements (i.e., 12865 and 13947 peaks, respectively; **Supplementary File 1a**), the normalized density of mapped reads was more enriched in one or the other cell type ($\log_{2(}$fold change) > 0.5 or <-0.5, FDR <0.01); we refer to these elements as GFP-positive NFRs (**Supplementary File 1b**) and GFP-negative NFRs (**Supplementary File 1c**), respectively (**Figure 1B**). Consistent with previous reports, overall ATAC-seq signal was high at transcription start sites [29; 30] (**Figure 1—figure supplement 2A**), but the majority of cell-type-specific NFRs were located in intergenic regions (**Figure 1—figure supplement 2B**). In both GFP-positive and GFP-negative NFRs, the average evolutionary conservation was higher within the NFR than in flanking DNA (**Figure 1—figure supplement 3**).

ATAC-seq identifies nucleosome free regions (NFRs), which include active enhancers, active and inactive promoters, and CTCF-bound regions, some of which are insulators [31; 32]. We reasoned that the subset of NFRs that are active enhancers and promoters would be flanked by nucleosomes with histone H3 acetylated on lysine 27 (H3K27Ac), a mark of active chromatin[33]. We used published data sets from whole embryos at 8 hpf or 24 hpf [34]. Although periderm comprises a small fraction of the embryo, we found examples of elements with ATAC-seq signal virtually specific to GFP-positive cells that nonetheless overlapped or were flanked by peaks of H3K27Ac signal detected in whole embryos (**Figure 1C**).

We split the GFP-positive NFRs into clusters of high or low H3K27Ac density detected in whole-embryo lysates at 8 hpf and/or at 24 hpf [35] (i.e., clusters 1 and 2) (**Figure 1B**). In the H3K27Ac^{High} cluster, the average H3K27Ac ChIP-seq signal dipped in the center of the NFR, consistent with NFRs being flanked by regions populated with nucleosomes bearing the H3K27Ac modification (**Figure 1D**). The average density of ATAC-seq reads did not differ between the H3K27Ac^{High} (**Supplementary File 1d**) and H3K27Ac^{Low} (**Supplementary File 1e**) clusters (p>0.05, Kolmogorov–Smirnov test) (**Figure 1D**), indicating that nucleosome-depletion alone does not signify an active regulatory element. We employed the Genomic Regions Enrichment of Annotations Tool (GREAT) (v 3.0) [36] (assignment rule: two nearest genes within 100 kb) to identify the sets of genes associated with each cluster. This set of genes associated with GFP-positive NFRs was strongly enriched for the Gene Ontology (GO) term in the zebrafish anatomy categories including “EVL” and “periderm” (e.g., at 10-10.3hpf, **Figure 1E, Figure 1—figure supplement 5A**). The significance of the association was much stronger for GFP-positive NFRs in the H3K27Ac^{High} cluster than in the H3K27Ac^{Low} cluster (**Figure 1E**). H3K27Ac is a dynamic mark of enhancers, however filtering on GFP-positive NFRs that are H3K27Ac-positive at 8hpf, H3K27Ac-positive at 24 hpf, H3K27Ac-positive at 8hpf and/or at 24 hpf, or H3K4me1-positive (a more stable mark of enhancers) at 8 hpf, all yielded sets of elements whose associated genes were enriched for the same GO terms at very similar significance levels **(Figure 1E)**. As mentioned above, GFP-positive NFRs have >2 fold more ATAC-seq reads in GFP-positive cells than in GFP-negative cells; perhaps surprisingly, the subset of GFP-positive NFRs with >4 fold more was less strongly associated with genes expressed in periderm than were GFP-positive NFRs overall (**Figure 1E**). In subsequent analyses, we used the set of GFP-positive NFRs positive for H3K27Ac signal detected in whole embryos at 8 hpf and/or at 24 hpf.

Although GFP-negative cells comprise a variety of non-periderm cell types, the set of genes associated with GFP-negative cells was also enriched for certain GO terms, including “brain development” (**Figure 1—figure supplement 5B**).

We tested the enhancer activity of ten elements of the H3K27Ac^{High} cluster, each adjacent to a gene expressed in periderm, using reporter assays. For example, *cldne* is expressed in zebrafish periderm from 6 hpf to 15 hpf (www.zfin.org). We amplified five ~400 bp elements , located approximately +6 kb, +3 kb, -8 kb, -11 kb, and 0 kb from the transcription start site of *cldne* and engineered them into a reporter vector upstream of a minimal promoter and cDNA encoding GFP [37] (**Figure 1C**). Embryos injected with these constructs exhibited mosaic GFP expression in the periderm at 11 hpf (e.g., **Figure 1F and G, Figure 1—figure supplement 4**). Four additional examples near other genes expressed in periderm (i.e., *gadd45ba*, *cavin2b*, *klf17*, and *ppl*) also had strong periderm enhancer activity (discussed later in the manuscript). Finally, we filtered out all elements that overlapped transcription start sites, in order to focus subsequent analyses on enhancers; the residual set of 3947 elements (**Supplementary File 1f**) we refer to as zebrafish GFP-positive active enhancers (zGPAEs). The analogous set of GFP-negative elements are GFP-negative active elements (zGNAEs).

zGPAEs are associated with genes expressed at high levels in the periderm

To gain a genome-wide view of the association between zGPAEs and genes whose expression is enriched in periderm, we again sorted GFP-positive and GFP-negative cells from *Tg(krt4:gfp)* embryos at 11 hpf and generated expression profiles for both populations using RNA-seq. We identified 4331 genes with higher expression in GFP-positive cells and 4216 genes with higher expression in GFP-negative cells (q value < 0.05, beta < 0, average TPM in GFP-positive cells > 1) (**Figure 1H, Figure 1—figure supplement 6A**). As expected, genes enriched in GFP-positive cells correlated positively with genes annotated at ZFIN, an online gene expression atlas, as being expressed in the EVL (**Figure 1—figure supplement 6B)** (www.zfin.org). Differentially accessible elements associated with genes whose expression is enriched in GFP-positive cells had significantly higher average accessibility in GFP-positive cells than in GFP-negative cells, and vice versa (**Figure 1I**). For instance, there is a zGPAE near *cyt1*, which has higher expression in GFP-positive versus GFP-negative cells (**Figure 1—figure supplement 7A**) and there is a GNAE near *her4.3* with the opposite enrichment (**Figure 1—figure supplement 7B**). Some exceptions to the general pattern were observed (e.g., *hspb9* and *npm1b,* **Figure 1H**); this might reflect the fact that enhancers do not always regulate an adjacent gene. We conclude that most or perhaps all zGPAEs are enhancers active in periderm in embryos 8hpf to 24 hpf.

Transcription factor binding sites overrepresented within zGPAEs

Using HOMER, we identified 12 short sequence motifs, corresponding to the preferred binding sites of specific transcription factors, that are enriched in zGPAEs and present in at least 5% them (**Figure 2A**); this list of zGPAE signature-motifs prompted testable hypothesis regarding the membership and structure of the periderm GRN. For instance, analysis of our RNA-seq profile of GFP-positive cells, and of available single-cell sequencing data (sc-seq) from zebrafish embryos [38] [39], revealed transcription factors expressed in the EVL at 10-14 hpf; the subset of these that bind zGPAE-enriched motifs are candidates to participate in the periderm GRN (**Figure 2A,** “best match”). In addition, clustering zGPAEs by virtue of the stage when the H3K27Ac signal is strongest (**Figure 2—figure supplement 1A**) [35] and then reassessing motif enrichment in each cluster revealed that the IRF6 site is more strongly enriched in early-acting zGPAEs (i.e., with H3K27Ac signal stronger earlier) (**Figure 2—figure supplement 1B, cluster 1**), and the TFAP2 and GATA sites in more strongly enriched in late-acting zGPAEs (**Figure 2—figure supplement 1B, cluster 2 and 3**), than in zGPAEs overall (**Figure 2A**). This implies that IRF6 acts earlier in the periderm GRN than do TFAP2 and GATA paralogs.

We predicted that zGPAE signature motifs would be essential for the periderm enhancer activity of zGPAEs. A zGPAE 3kb downstream of the *cldne* transcriptional start site possesses a GRHL motif but lacks other signature motifs (**Figure 2B**). ATAC-seq data from GFP-positive cells included fewer TN5-mediated cleavage events within this motif than in flanking DNA, indicating that a transcription factor bound at the motif (**Figure 2C**), i.e., footprint analysis [40]. We amplified the zGPAE, deleted the motif by site-directed mutagenesis, engineered both the intact and motif-deleted versions into a GFP reporter vector (separately), and injected each into wild-type zebrafish embryos at the 1-cell stage. The periderm enhancer activity of the intact zGPAE was strong and specific at 11 hpf (4-somite stage) (**Figure 2D**), but that of the GRHL motif-deleted form was weaker (i.e., fewer injected embryos exhibited GFP in the periderm) (**Figure 2E, F**). Similarly, we identified zGPAEs in which motifs matching the KLF (**Figure 2—figure supplement 2A**), TFAP2 (**Figure 2—figure supplement 2B**) and C/EBP (**Figure 2—figure supplement 2C**) motifs were the only ones detected. In each case this sequence was protected from transposase access and its deletion reduced periderm enhancer reporter activity. Collectively, these assays support the assumption that transcription factor binding motifs are essential for the function of periderm enhancers.

We then built a network using the best-match candidate transcription factors as nodes and linked the nodes using their putative target motifs as directional edges (from the regulating factor to its target motif, **Figure 2—figure supplement 3**). This network makes a number of testable hypotheses, including that GRHL transcription factors regulate expression of the other transcription factors enriched in periderm. Finally, analyzing the genes associated with zGPAEs that contain specific motifs revealed that although only about 27% of zGPAEs contain a GRHL binding site, more than 70% of genes whose expression is enriched in GFP-positive cells are associated with a zGPAE containing a GRHL site (**Figure 2F**). This implies that GRHL paralogs contribute to the regulation of most genes expressed in periderm.

To determine if particular combinations of signature motifs are over represented in GPAEs we counted the numbers of GPAEs with various two-motif or three-motif combinations (**Figure 2—figure supplement 4B** and **C)**. While some combinations were present more frequently than others, GPAEs with the most frequent three-motif-combination and with the least frequent three-motif-combination both were highly associated with the GO term “EVL (6-8h)”. Interestingly, however, the target genes of these two types of GPAEs rarely overlapped (**Figure 2—figure supplement 4 D**), implying there is underlying logic to when certain combinations are deployed. We also carried out unsupervised hierarchical clustering based on the frequency and combinations of motifs they contained (**Figure 2—figure supplement 4A**), but did not detect any striking patterns. Next, to detect patterns we switched to a supervised learning approach.

A gapped k-mer support vector machine classifier trained on zebrafish periderm enhancer candidates

To convert the observation of enriched binding motifs into a scoring function we trained a supervised machine-learning classifier called gapped k-mer support vector machine (gkmSVM) on zGPAEs [5] (**Figure 3A**). The trained classifier consists of a set of weights quantifying the contribution of each possible 10-mer to an element's membership in a training set. The resulting scoring function quantifies the degree to which a given test sequence resembles the training set [5]. Five-fold cross validation on subsets of zGPAEs reserved from the training set revealed that the gkmSVM trained on zGPAEs had an area under the receiver operating characteristic curve (auROC) of 0.88, and area under the precision-recall curve (auPRC) of 0.87 (**Figure 3B**). The latter shows we can identify over 50% of all zebrafish enhancers (recall) at a false positive rate of under 10% (precision = 1 - false positive rate). These performance measures compare favorably to those of sequence-based classifiers trained on tissue-specific enhancers in other studies [14; 15] and support the validity of the parameters we chose to use for identifying zGPAEs.

The performance measures indicated that the classifier should be able to distinguish elements with periderm enhancer activity based on their sequence. To test this prediction, we partitioned the genome into 400 bp tiles, each overlapping the preceding one by 300 bp, and scored each tile with the classifier. As expected, the average score of the tiles that overlap zGPAEs (i.e., the training set) was higher than that for tiles that do not overlap zGPAEs (**Figure 3C**). Moreover, the average H3K27Ac signal (in 24 hpf embryos) at the top scoring 30,000 tiles is higher than in the lowest-scoring 30,000 tiles (**Figure 3D**). Most importantly, genes associated with the top-scoring 10,000 tiles are enriched for the GO terms ectoderm, EVL, and periderm (**Figure 3E**), fulfilling our prediction. In addition, average expression of such genes was higher in GFP-positive versus GFP-negative cells in our RNA-seq profiles of these two cell types ($p<1.46e-05$, Mann-Whitney-Wilcoxon Test). Interestingly, the top-scoring 10,000 tiles that do not overlap zGPAEs are not enriched for GO terms related to EVL (**Figure 3—figure supplement 2**). Thus, even though the classifier overall has a low false discovery rate, given the large size of the genome there are many high-scoring elements are not periderm enhancers (false positives).

Zebrafish periderm enhancers share a binding site code with mouse and human periderm enhancers

While tissue-specific enhancers are rarely conserved between mammals and zebrafish (with some exceptions [41]), inter-species reporter tests have shown that they nonetheless can be composed of the same binding site code [13]. Therefore, we predicted that elements of the human genome that receive a high score from the classifier trained on zGPAEs will be enriched for human periderm enhancers. To test this notion, we divided the human genome into 400-bp tiles and scored each tile using the classifier trained on zGPAEs. We identified the top-scoring 0.1% bin of tiles (28,595 tiles) and examined their overlap with active enhancers, defined by ChIP-seq with antibodies to various chromatin-marks, in 125 cell/tissue types evaluated by the Roadmap Epigenomics project [42]. Although periderm was not among the tissues evaluated by this project, enhancers for several epithelial cell types were enriched beyond the overlap expected by chance (**Figure 4A**), suggesting such enhancers share a binding site code with zebrafish periderm enhancers. For instance, the average H3K27Ac signal in the top 0.1% tiles was much higher in normal human epidermal keratinocytes (NHEK) compared to in a transformed lymphocyte cell line (GM12878) (**Figure 4B**).

If top-scoring tiles in the human genome include human periderm enhancers, then the set of genes associated with such tiles should be enriched for those expressed in periderm. Unfortunately, there is no available expression profile of human periderm. However, a recent single cell-seq analysis of murine embryonic faces reported a cluster of 248 genes co-expressed with the canonical periderm marker *Krt17* [43]. Genes associated with the top-scoring 0.1% bin of tiles are enriched for those that are expressed in mouse periderm (hypergeometric p-value = 0.044) [43]. Similarly, the zebrafish orthologs of such genes are expressed, on average, at higher levels in GFP-positive versus GFP-negative cells described above (Wilcoxon rank sum test, p-value = 8.371e-06). Finally, we analyzed tiles of the mouse genome using the classifier trained on zGPAEs and found that tiles in the top-scoring 0.1% bin were strongly associated with genes expressed in mouse oral periderm (Fisher's Exact test, p = 2.2e-16) [43]. These findings suggest periderm enhancers in zebrafish, human and mouse genomes are enriched for the same transcription factor binding sites.

An enhancer 9.7 kb upstream of the human *IRF6* transcription start site (i.e., *IRF6-9.7*) has been shown to drive reporter expression in oral periderm of transgenic mouse embryos (**Figure 4C**) [44]. This element is of clinical interest as it harbors a mutation (MCS9.7-*350dupA*) that causes Van der Woude syndrome and diminishes the periderm-enhancer activity of the element containing it [45]; it also harbors a single nucleotide polymorphism associated with risk for non-syndromic orofacial clefting [46]. We engineered a 606 bp element within *IRF6-9.7* into the GFP reporter vector and injected it into wild-type embryos. In stable transgenic lines, GFP expression was detectable in the enveloping layer by shield stage (**Figure 4—figure supplement 1B-E)** and until at least 5 dpf (**Figure 4E**); it was also evident in pharyngeal epithelium at this stage **(Figure 4—figure supplement 1F and G**). We created a double-transgenic embryo, harboring a tdTomato reporter whose expression is driven by the *krt4* promoter, and confirmed that tdTomato and GFP expression overlap **(Figure 4F)**. Although IRF6-9.7 is not overtly conserved to the zebrafish genome (**Figure 4C**), the part of it tested in mouse and zebrafish reporter contains a tile whose score matches the median score of tiles overlapping the training set (**Figure 3C**), and is in the top-scoring 1.0-1.5% bin of tiles in the human genome (**Figure 4C**). Interestingly, the highest-scoring tile within the larger enhancer, marked by H3K27Ac in normal human epidermal keratinocytes (NHEK), is in the top-scoring 0.2% bin of tiles (**Figure 4C**).

We discovered a second periderm enhancer by searching the genome for elements sharing several ChIP-seq features of *IRF6*-9.7; specifically we filtered on strong H3K27Ac signal (ENCODE data) and peaks of IRF6 [47], KLF4 [48], and TP63 [49] binding, all assessed in normal human epidermal keratinocytes (NHEK). There are only 5 elements in the genome where all of these features converge, and each score in the top 2% bin or higher (**Figure 4—figure supplement 2**). We focused on one that lies 36.7 kb upstream of *ZNF750* (*ZNF750-36.7*) (**Figure 4D**). Chromatin configuration data indicate it binds to the *ZNF750* promoter in keratinocytes [50]. The mouse ortholog of *ZNF750* (i.e., *ZFP750*) is expressed in murine oral epithelium [51], and oral periderm [43], and the zebrafish ortholog *znf750* is expressed in EVL [38]. We amplified a 2.8 kb element overlapping the H3K27Ac signal from NHEK cells and made stable transgenic reporter fish. In them GFP expression is detectable in the enveloping layer starting at 5.25 hpf (50% epiboly) (**Figure 4—figure supplement 1I'**) and still visible at 5 dpf, although with lower intensity than in *Tg(IRF6-9.7:gfp)* transgenic animals at this stage (**Figure 4F**), consistent with lower expression levels of *znf750* in comparison to *irf6*. The highest scoring tiles in this apparent human periderm enhancer lies in the 0.5-1.0% bin (**Figure 4D**), supporting our prediction.

We found a third periderm enhancer 8.3 kb upstream of the transcriptional start site of *PPL* (encoding Periplakin*)* (*PPL*-8.3), whose mouse ortholog is highly expressed in periderm [43] and contributes to epidermal barrier formation [52]. Referring to ChIP seq experiments in NHEK cells, this element is bound by KLF4, ZNF750, and GRHL3, all transcription factors implicated in differentiation of keratinocytes (reviewed in [53]); it is on the flank of an island of H3K27Ac signal in NHEKs (ENCODE). In transient transgenic reporter assays in zebrafish it is a potent periderm enhancer (**Figure 4—figure supplement 3A and B**). We also amplified a zGPAEs 10 kb upstream of zebrafish *ppl* (*ppl*-10) **(Figure 4—figure supplement 3D and E)**. Interestingly, deletion of KLF4 binding sites from the human element *PPL-8.3* or from the zebrafish element *ppl-10* strongly diminished the periderm enhancer activity in both cases **(Figure 4—figure supplement 3C and F),** suggesting that the two enhancers might be at least partially functionally conserved. As predicted, the highest scoring tile within *PPL*-8.3 is in the top-scoring 1.5-2% bin.

To determine whether the shared binding sites reflect sequence homology between *ppl*-10 and *PPL*-8.3, we performed sequence alignments. We found that a 467 bp core sequence from the zebrafish enhancer (plus-strand) is marginally more identical to a 400 bp core sequence from the human enhancer (plus-strand) relative to several control sequences including: the zebrafish minus-strand (reverse-complement), the non-biological reverse sequence, and non-biological sequences of similar lengths produced by Fisher-Yates shuffling of the plus-strand sequence (see Material & Methods, **Supplementary File 2a** and **Supplementary File 2b**). Furthermore, three-way alignments of the human and mouse plus-strand sequences with each of the zebrafish test and control sequences indicates that the zebrafish plus-strand engenders a need for a number of null characters (dashes) in the three-way alignments that is almost one standard deviation smaller than the controls (158 insertions versus an average number of insertions of 200.2 +/- 44.1 s.d. amongst minus-strand, reverse, and three Fisher-Yates shuffled sequences of the plus-strand). Last, the Hu_400+ and Zf_467+ pairwise-alignment has more 5 bp-long blocks of perfect identity (5 such blocks) relative to all five of the zebrafish controls (average two 5 bp-long blocs). Much of this potentially faint conservation overlies the elements conserved in the mammalian enhancer sequences (**Supplementary File 2b**). In summary, there is modest but detectable sequence homology between human *PPL*-8.3 and zebrafish *ppl*-10.

Defining sets of enhancers in mouse palate epithelium and a human oral epithelium cell line with ATAC-seq

Given the preceding findings, we reasoned that the gkmSVM classifier trained on the zebrafish periderm enhancers might be able to identify SNPs that disrupt periderm enhancers, and might perform as well as, or better than classifiers trained on enhancers from other relevant tissues, i.e., mouse primary palate epithelium or a human oral epithelium cell line. To test this prediction, we dissected palate shelves from mouse embryos at embryonic day (E)14.5 and manually removed epithelium after brief incubation in trypsin (**Figure 5A**). Subsequently we subjected both the palate epithelium and the residual palate mesenchyme to ATAC-seq; there was a strong correlation among peaks in the three replicates (**Figure 5—figure supplement 1A**). In palate epithelium, we identified elements with more ATAC-seq reads than in palate mesenchyme (i.e., palate epithelium-specific NFRs, listed in **Supplementary File 1g** and **h**); these were binned as high- or low-density H3K27Ac signal (6079 and 8177 elements, respectively; listed in **Supplementary File 1i**) based on H3K27Ac ChIP-seq data from embryonic facial prominences (E14.5, ENCODE database, GEO:GSE82727) [54] (**Figure 5B**). In the H3K27Ac^{High} cluster, the average H3K27Ac ChIP-seq signal dipped in the center of the NFR (**Figure 5C**) as in zGPAEs. The average density of ATAC-seq reads was slightly higher in H3K27Ac^{High} vs. H3K27Ac^{Low} cluster (**Figure 5—figure supplement 1B**). Genes assigned to H3K27Ac^{High} elements were strongly enriched for the GO term “oral epithelium” ($\log_{10}$(binomial-

FDR)<75), as were those assigned to H3K27Ac^{Low} elements ($\log_{10}$(binomial-FDR)<49) (**Figure 5D**, **Figure 5—figure supplement 1C**). Significantly, we found H3K27Ac^{High} cluster elements both near genes expressed at high levels in superficial palate epithelium (palate periderm) (e.g., *Krt17*, **Figure 5E**) and near those expressed in basal palate epithelium (e.g., *Krt14,* **Figure 5—figure supplement 1D**), showing that the isolated epithelium contained both of these layers. We also observed mesenchyme-specific NFRs in the H3K27Ac^{High} cluster near genes whose expression is high in mesenchyme (e.g., *Runx2*, **Figure 5F**), and shared NFRs in the H3K27Ac^{High} cluster near genes expressed in both (e.g., *Klf4*, **Figure 5—figure supplement 1E**). Elements in the H3K27Ac^{High} cluster that do not overlap transcription start sites were named mouse palate epithelium active enhancers (mPEAEs) (**Supplementary File 1i**). HOMER revealed 18 short sequences for which mPEAEs are enriched and that are present in at least 5% of them (**Figure 5G**). Of note, 6 were predicted to be bound by transcription factors also predicted to bind one of the 11 zebrafish periderm signature motifs (**Figure 5G,** bold text), suggesting epithelial enhancers share a set of core transcription factors.

Similarly, to identify enhancers active in an oral epithelium cell line, we carried out ATAC-seq, and H3K27Ac ChIP-seq on human immortalized oral epithelial cells (HIOEC) induced to differentiate by incubation in calcium. For comparison, we also carried out ATAC-seq in human embryonic palate mesenchyme cells (HEPM). Focusing on ATAC-seq peaks that concorded among three replicates in each cell type, 31,296 NFRs present in HIOEC cells were absent in HEPM cells (**Supplementary File 1j**). Among such HIOEC-specific peaks, 15,972 overlapped (or were flanked by) H3K27Ac peaks called in two or more of the three replicates (**Figure 5—figure supplement 2A**; listed in **Supplementary File 1k**) while 15,324 peaks neither overlapped nor were flanked (within 1500bp) by H3K27Ac peaks. GO enrichment assay for these two clusters NFRs showed the nearest genes of cluster 1 HIOEC-specific peaks were highly enriched with epithelial structure (**Figure 5—figure supplement 2B**), while cluster 2 did not exhibited such enrichment (**Figure 5—figure supplement 2C**). For instance, the *KRT17* gene is expressed in human epithelium, and chromatin regions within this locus were specifically open in HIOEC cells and overlapped with human embryonic craniofacial super enhancer (**Figure 5—figure supplement 2D**). However, chromatin regions within *RUNX2* locus were specifically open in HEPM cells (**Figure 5—figure supplement 2E**). After we filtered out elements containing transcription start sites, the subset that remained was the human oral-epithelium active enhancers (hOEAEs). They were enriched for a set of binding sites, such as TEAD, JUN, C/EBP, GRHL and TFAP2; among these several were shared with zGPAEs and mouse PEAEs (**Figure 5—figure supplement 3**). We trained a gkmSVM classier on hOEAEs and used it below.

Ranking OFC-associated SNPs using classifiers trained on zGPAEs, mPEAEs, and hOEAEs

Next, we used the classifiers trained on zGPAEs, mPEAEs, and hOEAEs to predict which single nucleotide polymorphisms (SNPs) associated with risk for orofacial cleft near the *KRT18* gene are most likely to disrupt an enhancer of the type upon which the classifier was trained. Revisiting our previously published GWAS data [8], including imputed SNPs, we found 14 SNPs with at least suggestive p-values for association to risk for orofacial clefting (p<1e-5) and in strong linkage disequilibrium with the lead SNP at this locus (i.e., SNPs 1-14) (SNP labels and p-values, and **Supplementary File 1j**) (**Figure 6A**). Functional SNPs are predicted to a) lie in enhancers active in a relevant tissue and b) have allele-specific effects on enhancer activity. To determine which SNPs lie in enhancers we evaluated published chromatin-state data from human embryonic faces [55] and 111 cell types characterized by the Roadmap Epigenomics project [42]. Interestingly just three of the SNPs, i.e., SNP1, SNP2, and SNP13, lie in chromatin predicted to be active in one or more of these tissues while the others lay in relatively inert chromatin (**Figure 6B,** 9 representative Roadmap cell lines are shown). Using the classifier trained on zGPAEs, we calculated deltaSVM scores of the 14 SNPs, and for comparison, of 1000 additional SNPs within 100 kb (**Figure 6 C and D**). The deltaSVM scores for most of the OFC-risk-associated SNPs were within one standard deviation of the median score of 1000 SNPs. By contrast, the deltaSVM score of SNP2 lower than that of 998 of 1000 randomly selected SNPs, and thus was an outlier (Bonferroni corrected p value = 0.028) (**Figure 6 C and D**). Interestingly, SNP2 also had the strongest negative deltaSVM of all the risk-associated SNPs when the classifier was trained on mPEAEs or on hOEAEs, although in neither case were the deltaSVM values significant outliers in comparison to those of the 1000 randomly-selected SNPs (SNP2, Bonferroni corrected p values of 0.126 and 0.238, respectively) (**Supplementary File 3b-e**). Because *KRT18* is not expressed in palate mesenchyme, we used the classifier was trained on mouse palate mesenchyme active elements (mPEAEs) as a negative control. As expected the deltaSVM for SNP2 was unremarkable (within the middle 50% of scores of 1000 SNPs) **(Figure 6D** and **Supplementary File 3b-e**).

Reporter assays in human oral epithelium cells support SNP2 being functional variant

Previously, upon training a gkmSVM classifier on melanocyte enhancers, deltaSVM scores of SNPs within known melanocyte enhancers were found to correlate with the observed differences in reporter activity between mutant and wild-type enhancer constructs tested in a melanocyte cell line [7]. We asked whether, similarly, upon training a classifier on zGPAEs, mPEAEs, or hOEAEs, would deltaSVM scores of SNPs within known epithelium enhancers similarly correlate with the differences in reporter activity between risk and non-risk constructs tested in a human basal oral keratinocyte cell line. Of the 14 OFC-risk associated SNPs at this locus, only SNP1 and SNP2 lie in chromatin marked as an active enhancer in normal human epidermal keratinocytes (NHEK) cells, which we predict are similar to oral keratinocytes (**Figure 6B**). SNP1 has a neutral deltaSVM when the classifier was trained on zGPAEs or mPEAEs (**Figure 6C and D** and **Figure 6—figure supplement 1**) and a deltaSVM of -3, in the lowest quartile of 1000 SNPs, when the trained on hOEAEs (**Figure 6D** and **Figure 6—figure supplement 1**). As mentioned above, SNP2 had a strongly negative deltaSVM with the classifier trained on zGPAEs, mPEAEs, and hOEAEs (**Figure 6 C and D**). We amplified 700-base-pair elements centered on each SNP, engineered them to harbor either the risk-associated or non-risk associated allele of the SNP, introduced them (separately) into a luciferase-based reporter vector (with a basal SV40 promoter), and transfected these constructs into GMSMK basal oral epithelium cells. Both elements drove luciferase levels above background, suggesting that both SNPs lie in enhancers active in oral keratinocytes. SNP2 but not SNP1 had significant allele-specific effects on this enhancer activity, with the risk-associated variant driving lower reporter activity than the non-risk variant (**Figure 6E**). Thus, for these two SNPs, the deltaSVM scores and reporter effects were correlated.

We next tested the prediction that the enhancer in which SNP2 lies regulates expression of *KRT8* and/or *KRT18*. We transfected GMSMK cells with Cas9 ribonucleotide protein (RNP) and, in experimental cells, with two gRNAs targeting sites separated by 109 bp and on either side of SNP2, or, in control cells, with a non-targeting gRNA. Two days post-transfection, quantitative RT-PCR on RNA harvested from the pools of cells revealed that *KRT18* mRNA, and at lower levels, *KRT8*, could be detected in control cells, and that levels of both transcripts trended lower in experimental cells (**Figure 6—figure supplement 2**). We isolated single-cells and expanded clones from both the control and experimental cells. Among the latter, we used PCR and sequencing to identify three independent clones that were homozygous for a 109 bp deletion between the two gRNA target sites. The average expression level of *KRT18* and *KRT8* in these three colonies was lower than in a single colony we isolated from the control cells (**Figure 6G, H**). These results support the notion that the region containing SNP2 is an enhancer driving expression of *KRT8* and *KRT18* in human oral epithelial cells.

KRT18 is expressed in many epithelia other than periderm, including trophectoderm [56], embryonic surface ectoderm [57; 58], oral epithelia [59], embryonic cornea [59], gonad [60], bladder [61], choroid plexus [62] and others. To test the prediction that SNP2 lies in a periderm enhancer, we engineered a 701 bp element centered on SNP2, and harboring either the risk or non-risk allele of it, into the GFP reporter vector. We similarly engineered two GFP reporter constructs from 701bp elements centered on SNP1 with risk or non-risk alleles of it. Unexpectedly, in zebrafish embryos injected with these constructs and monitored up until 4 days post fertilization, we detected very little GFP expression and no consistent pattern of it (N > 100 embryos each construct). We considered the possibility that, against our prediction, the elements are enhancers in mammals but not zebrafish. We engineered all four elements (separately) into reporter vectors with a minimal *Shh* promoter and the *LacZ* gene and carried out transgenic reporter assays in F0 mouse embryos using site-directed transgene integration [63]. Across 18 transgenic embryos injected with SNP1 element, 11 with non-risk allele, 7 with risk allele, we did not observe reproducible reporter expression, defined as expression in the same anatomical structure in at least two embryos injected with the same construct, and the majority of transgenic embryos did not show any reporter staining (**Figure 6—figure supplement 3A**). Similarly, in 14 transgenic embryos injected with SNP2 element, 7 with each allele, we did not observe reproducible reporter expression, and most showed none at all (**Figure 6—figure supplement 3B**). These results indicate that the elements centered on SNP1 and on SNP2, at least when paired with the basal *Shh* promoter and integrated at this locus, do not reproducibly drive high level reporter activity. Interestingly, however, a single embryo transgenic for the risk allele of the SNP2 element exhibited mild reporter expression in the periderm of the face and limbs (**Figure 6I, I', and Figure 6 supplement 3B**). We hypothesized that this embryo had a higher copy number of the reporter vector than the other embryos. PCR analysis confirmed that this embryo carried 8 copies of the reporter constructs, whereas all other embryos transgenic for SNP2 carried only 2 copies. While not conclusively demonstrating reproducible reporter activity, this result is consistent with the possibility that SNP2 lies in a sequence that has quantitatively mild periderm enhancer activity and may causatively contribute to the phenotypes observed in patients.

Allele-specific effects of SNP1 and SNP2 transcription factor binding sites

We used JASPAR to assess the transcription factor consensus binding sites in 19 bp windows centered on SNP1 (rs11170342) or SNP2 (rs2070875**)**. The risk allele of SNP1 strongly reduced the score of Plagl1 and Spz1 sites, and created high-scoring sites for SP4/8/9 and KLF14/15. The risk allele of SNP2 strongly reduced the score of SNAI1/2, NFATC1/2/4, and SIX1/2 sites, and created high-scoring sites for HNF4A/G and NR2F1 (**Supplementary File 3f**). Interestingly, Snai2 is expressed in mouse palate epithelium, and Snai1/Snai2 double mutants exhibit abnormal migration of periderm at medial edge palate epithelium [64].

Discussion

Here we undertook an analysis of zebrafish periderm enhancers with two objectives relevant to the genetic underpinnings of orofacial clefting. The first was to learn more about the gene regulatory network (GRN) governing periderm differentiation. Already, the human orthologs of four known elements of this GRN, Irf6, Grhl3, Klf17, and simple epithelium keratins, are implicated in risk for non-syndromic orofacial clefting. Therefore, the orthologs of additional members of this GRN, in particular that hubs, are candidates to harbor the mutations that constitute the missing heritability for orofacial clefting. The second objective was to use zebrafish periderm enhancers to train a classifier with which to prioritize SNPs in non-coding DNA for their likelihood of disrupting periderm enhancers. Until recently, identifying enhancers of a given specificity has required testing enhancers individually in reporter assays. Chromatin mark ChIP-seq, or ATAC-seq which requires fewer cells, has permitted the identification of candidate enhancers in a specific tissue or cell type in large numbers [30; 65]. However, there is currently no cell line model of human palate periderm. Periderm cells could be isolated from murine embryonic palate shelves using fluorescence-activated cell sorting (FACS) and the *Krt17*-GFP transgenic mouse line [57]. However, the zebrafish periderm is far more accessible. While tissue-specific enhancers are rarely strongly conserved between fish and mammals, the faithful performance of tissue-specific enhancers near the human *RET* gene in zebrafish transgenic reporter assays implied that, at least in some cases, such enhancers composed of the same transcription factor binding sites in the two clades [13]. Here we report evidence that this is the case for periderm enhancers – because tiles of the murine genome that are high scoring with the classifier trained on zebrafish periderm enhancers are enriched near genes expressed in periderm. Both of these objectives were met.

To identify a set of enhancers active in zebrafish periderm we sorted GFP-positive and GFP-negative cells from *Tg(krt4:gfp)* transgenic embryos at 11 hpf and performed ATAC-seq on both populations. About 5% of all ATAC-seq positive elements had at least 2-fold more ATAC-seq reads in GFP-positive cells. For comparison, ATAC-seq on GFP-positive and GFP-negative cells sorted from *Tg(fli1:gfp)* zebrafish embryos at 24 hpf [30], or pro-sensory cells sorted from cochlear ducts in *sox2-EGFP* transgenic mice[65], revealed about 9% and 15%, respectively, of all ATAC-seq elements to have greater read depth in GFP -positive cells. The difference in the fraction of tissue-specific elements may simply reflect differences in the fold-change-in-ATAC-seq-reads filter applied to such elements. Less than half of the NFRs enriched in GFP-positive cells were marked by H3K27Ac in whole embryo lysates from near the same stage [34]. This observation is consistent with other studies showing the correlation between nucleosome-free status and H3K27Ac signal it is not absolute; for instance, inactive enhancers can be nucleosome free [32]. Genes flanking elements meeting both criteria are strongly enriched for those expressed in periderm. Genes flanking elements meeting only the first criterion (more ATAC-seq reads in GFP-positive cells than GFP-negative cells) were similarly enriched, but to a lesser degree. Periderm-specific NFRs lacking H3K27Ac at 8hpf or 24 hpf but associated with genes expressed in periderm may be enhancers that are active at another stage; consistent with the existence of such elements, a recent study shows the expression profile in zebrafish periderm changes over time [66]. Further, 10 of 10 tested elements meeting both criteria and proximal to genes known to be highly and specifically expressed in periderm functioned as periderm enhancers in zebrafish reporter assays. Together these findings support our conclusion that elements with more ATAC-seq reads in GFP-positive versus GFP-negative cells sorted from Tg(krt4:gfp) embryos at 11 hpf, and marked with H3K27Ac at 8 hpf and or at 24 hpf, have periderm enhancer (or promoter) activity in embryos at these stages.

Assessment of transcription factor binding sites enriched in ATAC-seq peaks can yield inferences into transcriptional regulatory networks [67; 68]. Transcription factors known to bind motifs enriched in the zebrafish periderm enhancer candidates include those previously implicated in periderm development, like Irf6, Grhl3 and Klf17, and novel ones, including C/ebp, Fosl, Gata3 and Tead. Interestingly, orthologs (or, in the case of Klf17, the paralog, Klf4) of each these transcription factors are necessary for mammalian skin development, showing the similarity of the relevant GRNs (e.g., GRHL [69; 70],TEAD-YAP [71], FOS-JUN /AP-1 [72], KLF4 [73], TFAP2 [74], GATA6 [75], C/EBP [76], ETS1 [77; 78], IRF6 [79]). A role for Tead family members in periderm development is supported by the observation that a loss of function mutation in *yap*, encoding a cofactor of Tead, disrupts periderm development in medaka [80]. Knowledge of the key elements of the periderm GRN may help in prioritizing variants that are discovered in patients with orofacial clefting through whole exome or whole genome analyses. Second, whereas Irf6 sites are found in just 5% of enhancer candidates, Grhl sites are present in most. Our findings imply that, although both of these transcription factor families are essential for periderm development, Grhl proteins function much more broadly than their Irf6 counterparts. Third, the Irf6 binding site was enriched to a greater extent in the enhancer candidates that were associated with more anti-H3K27Ac ChIP-seq reads at 8 hours post fertilization (hpf) than 11 hpf. This implies that Irf6 acts early in the GRN; the notion that Irf6 activates a periderm program and is then no longer necessary is consistent with the fact that zygotic *irf6* mutants are viable [81]. Finally, the GPAEs associated with genes encoding candidate members of the GRN contain the binding sites for other such members, implying that mutual cross-regulation of these transcription factor is extensive. While direct connections inferred in this way await confirmation by ChIP-seq or CUT&RUN, cross-regulation among members of a given GRN layer is a common feature in development [82]. ATAC-seq data combined with RNA-seq data can be combined to yield transcriptional regulatory networks that approach the more difficult to achieve networks achieved with ChIP-seq and transcription-factor knockout studies [68].

Next, we applied machine-learning classifier to zebrafish periderm enhancer candidates and used it to test the prediction that zebrafish periderm enhancers are enriched for the same sequence motifs as their mammalian counterparts. While gapped kmer support vector machine classifier was the top-performing algorithm at the time we began the study, others are available now that may function better in some circumstances [83]. The trained classifier had a low false positive rate and auROC and auPR curves comparable to those in similar published studies [14; 15]. We applied gkmSVM to tiles of the zebrafish genome. Plotting the average score of tiles overlapping the training set and that of those that do not was useful in conveying the potential of this tool to distinguish periderm enhancers based on sequence alone. On average periderm enhancers have higher scores than non-periderm enhancers, but plots of these two distributions overlap, indicating that a high score is not a guarantee that an element is a periderm enhancer. We also applied the classifier to the human genome and discovered that enhancers in various epithelial cell types relative to other classes of enhancers are enriched for higher scoring tiles. This suggests that epithelial enhancers are similarly constructed in all organisms. In some cases, high-scoring tiles in the human genome may be orthologs of zebrafish periderm enhancers. An example of such an instance may be a high-scoring tile 8.3 kb upstream of the human *PPL* gene which we found had detectable sequence conservation to an GFP-positive NFR 10 kb upstream of the zebrafish *ppl* gene. Applying the classifier to the murine genome revealed that high-scoring tiles were enriched near genes expressed in periderm, revealing that periderm enhancers in zebrafish and mouse are enriched for the same sequence motifs. This implies human periderm enhancers share this feature and supports the use of the classifier in deltaSVM analysis of disease-associated SNPs near genes expressed in periderm.

Finally, we used the classifier trained on zGPAEs to prioritize orofacial-cleft associated SNPs near *KRT18,* a gene expressed in periderm*,* for those that are likely to affect a periderm enhancer. Interestingly, classifiers trained on enhancers apparently selective for a) zebrafish periderm versus non-periderm, b) mouse palate epithelium versus palate mesenchyme, or c) a human oral epithelium cell line versus a human palate mesenchyme cell line all picked SNP2 (i.e., rs2070875) as having the strongest Delta SVM among the 14 OFC-associated SNPs at this locus. This again supports the notion that epithelial enhancers are similarly constructed in all vertebrates. Although the sets of enriched transcription factor binding sites enriched in the three sets of enhancer were similar, only the classifier trained on zebrafish periderm enhancers yielded a deltaSVM for SNP2 that met formal significance. Assuming that SNP2 is indeed functional, successful use of deltaSVM in identifying functional SNPs depends more on the enhancers being derived from the correct tissue (i.e., periderm) than the correct species (i.e., human).

Is SNP2 the orofacial-cleft (OFC)-associated SNP at this locus that directly affects risk for the disorder? Luciferase assays in human oral epithelium cells support the notion that SNP1 and SNP2 lie in enhancers active in oral epithelium, and support the predictions of the deltaSVM that SNP2 but not SNP1 affects the activity of their surrounding enhancers. The fact that homozygous deletion of the 109 bp region flanking SNP2 reduced expression of *KRT18* and *KRT8,* also supports SNP2 affecting OFC risk, because periderm integrity is compromised in zebrafish embryos depleted of several keratin genes [23]. Unexpectedly these elements were not consistently active in reporter assays carried out in zebrafish and mouse embryos. However, in a single mouse embryo with 8 copies of the SNP2 transgene construct, reporter expression was clearly detected in external periderm. The construct had the risk allele of SNP2, but as reporter copy number varied among embryos, we could not assess whether the allele affected reporter level in this assay. Generating embryos with multiple integrated copies of the constructs, or perhaps targeting the constructs to a different locus, will be necessary to determine if the enhancer harboring SNP2 is active in oral periderm as would be predicted it is relevant to risk for orofacial clefting. Robust chromatin mark evidence from human embryonic faces and human epidermal keratinocytes suggests the lack (or very low) activity of these elements in the zebrafish and mouse embryo assays reflects a technical artifact; for instance it is possible these enhancers are not compatible with the basal promoter in the *GFP* vector (from the FOS gene [37]) or in the *LacZ* vector (from the *Shh* gene). In zebrafish embryo reporter assays, enhancers perform more robustly when paired with their cognate promoters than with an exogenous one [30]. In summary, the data gathered support SNP2 as affecting expression of an enhancer active in periderm and regulating *KRT18* and possibly *KRT8* expression.

Finally, the classifiers presented here trained on the various epithelia-specific enhancers may be useful in nominating functional SNPs at the additional OFC-associated loci identified in genome wide association studies. At loci where the candidate risk gene is expressed in oral epithelium, e.g., *IRF6, MAFB, FOXE1, TP63*, the classifiers trained on zGPAEs, mPEAEs, and hOEAEs should all be applied; if the risk gene is expressed in basal epithelium, like TP63, the classifier trained on mPEAEs may work better than the one trained on zGPAEs. Where the candidate risk gene is expressed in mesenchyme (e.g., *PAX7*), the classifier trained on mPMAEs is expected to be the most accurate. It is important to note that in a study of over 100 SNPs, the deltaSVM score and the effect of the SNP on reporter level (in an appropriate cell type) was significant but modest[7]. Therefore, machine learning analyses can prioritize SNPs but functional tests remain important essential. These include quantification of a SNP's effects on enhancer activity, either by reporter assays in vitro, or more powerfully, through genome engineering of an appropriate cell line to render the SNP homozygous for risk or non-risk allele and then quantification of RNA levels. The efficiency of homology directed repair for this purpose remains highly locus-dependent, and we did not succeed in applying it here, although we did so at another locus [4]. Fortunately, single-nucleotide editing tools are improving rapidly [84]. Finally, in vivo reporter assays will remain essential for testing the tissue specificity of enhancers harboring the candidate functional SNPs. In such assays, safe harbor chromatin integration is clearly desired, although such safe harbors may be more permissive for expression in some tissues than others, and enhancer-promoter compatibility may additionally affect efficiency of in vivo reporter assays.[30]

Materials and Methods

Key resources table

Key Resources Table

Reagent type (species) or resource	**Designation**	**Source or reference**	**Identifiers**	**Additional information**
strain, strain background (*Escherichia coli*)	One Shot® TOP10	Life technologies	Cat.no. C4040-10	Chemically competent cells
cell line (*Homo-sapiens*)	GMSM-K (human embryonic oral epithelial cell line)	[85]	RRID:CVCL_6A82	a kind gift from Dr. Daniel Grenier
cell line (*Homo-sapiens*)	HIOEC (human immortalized oral epithelial cells)	[86]	RRID:CVCL_6E43	
cell line (*Homo-sapiens*)	HEPM (human embryonic palatal mesenchyme cells)	ATCC	ATCC Cat# CRL-1486, RRID:CVCL_2486	
antibody	anti-Histone H3, Acetylated Lysine 27 (Rabbit polyclonal)	Abcam	Abcam Cat# ab4729, RRID:AB_2118291; lot NO. GR3211959-1;	ChIP (4ug per 500,000 HIOEC cells)
recombinant DNA reagent	pXX330 (plasmid)	Addgene;{Cong, 2013 #2;Ran, 2013 #1}	RRID :Addgene_42230	

recombinant DNA reagent	cFos-GFP	[37]		a gift from Shannon Fisher
recombinant DNA reagent	cFos-tdTomato	This paper		Modified from cFos-GFP
recombinant DNA reagent	pENTR/D-TOPO	Life technologies	Invitrogen™ Cat# K240020	
recombinant DNA reagent	cFos-FFLuc	[4]		
sequence-based reagent	cFos-RLuc	[4]		
sequence-based reagent	Klf17_+1.8_F	This paper	PCR primers	ATGCTGACTCCACCATCCTC
sequence-based reagent	Klf17_+1.8_R	This paper	PCR primers	CACCTACCCCTTGGCTAATCGTTG
sequence-based reagent	Cavin2b_+18_F	This paper	PCR primers	TTCTGTTTTTGCCATCAGCA
sequence-based reagent	Cavin2b_+18_R	This paper	PCR primers	CACCTTTTAATCACCGCCTTTCCA
sequence-based reagent	Gadd45ba_-0.7_F	This paper	PCR primers	TGGTTGGGTTCAGAGGTAGG
sequence-based reagent	Gadd45ba_-0.7_R	This paper	PCR primers	CACCATGACTCGACGAAAGCAAA

sequence-based reagent	SNP2_gRNA_left	This paper	gRNA target	CTAAGAAGGATCTGCTCCCC
commercial assay or kit	SNP2_gRNA_right	This paper	gRNA target	GAGGACAGTATTCTTAAACG
commercial assay or kit	RNAqueous® Total RNA Isolation Kit	Ambion	Cat. NO. AM1912	
commercial assay or kit	RNA Clean &Concentrator™-5 Kit	Zymo Research	Cat. NO. R1013	
commercial assay or kit	SMART-Seq® v4 Ultra® Low Input RNA Kit	TAKARA	Cat. NO. 634888	
commercial assay or kit	Agilent RNA 6000 Pico	Agilent Technologies	Cat. NO. 5067-1513	
commercial assay or kit	Nextera XT DNA Sample Preparation Kit	Illumina	Cat. NO. FC-131-1002	
commercial assay or kit	Nextera DNA Sample Preparation Kit	Illumina	Cat. NO. FC-121-1030	
commercial assay or kit	VAHTS Universal DNA Library Prep Kit for Illumina	Vanzyme	Cat. NO. ND606-01	

commercial assay or kit	KAPA Library Quantification Kit	Roche	Cat. NO. KK4824	
commercial assay or kit	NEBNext® High-Fidelity 2x PCR Master Mix	New England Biolabs	Cat. NO. M0541S	
chemical compound, drug	Ampure XP beads	Beckman Coutler	Cat. NO. A63881	
chemical compound, drug	0.25% trypsin-EDTA	Life Technologies	Cat. NO. 25200056	
chemical compound, drug	Defined trypsin inhibitor	Life Technologies	Cat. NO. R007100	
software, algorithm	Turbo™ DNase I	Ambion	Cat. NO. AM2238	
software, algorithm	R	R	RRID:SCR_001905	v 3.5.1 v 3.3.2
software, algorithm	Bowtie2	[87]	RRID:SCR_005476	v 2.3.4.1
software, algorithm	Trimmomatic	[88]	RRID:SCR_011848	v.0.38
software, algorithm	DiffBind	[89]	RRID:SCR_012918	

software, algorithm	seqMINER	[90]	RRID:SCR_013020	v 1.2.1
software, algorithm	HOMER	[91]	RRID:SCR_010881	v 3.0
software, algorithm	Gapped k-mer support vector machine	[6]	https://rdrr.io/cran/gkmSVM/	v 0.79.0
software, algorithm	BEDTools	[92]	RRID:SCR_006646	v 2.24.0
software, algorithm	Picard Tools	http://broadinstitute.github.io/picard/	RRID:SCR_006525	v 0.35
software, algorithm	SAMtools	[93]	RRID:SCR_002105	v 1.7
software, algorithm	MACS2	[94]	RRID:SCR_013291	v 2.1.1
software, algorithm	DeepTools	[95]	RRID:SCR_016366	v 2.0

Identification of SNPs associated with human orofacial clefting

We re-analyzed our published meta-analysis of two GWASs for orofacial clefting [8].The details of each contributing GWAS have been extensively described. Data are from a total of 823 cases, 1,700 controls, and 2,811 case-parent trios and were obtained by genotyping using the Illumina HumanCore+Exome array or the Illumina Human610-Quad array. In our re-analysis, genotype probabilities for imputed SNPs were converted to most-likely genotype calls using GTOOL (http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html). Genotype calls were retained for analysis only if the genotype with the highest probability was greater than 0.9. SNPs were excluded if the minor allele frequency was low (<1%), the imputation quality scores was low (INFO<0.5) or deviating from Hardy-Weinberg equilibrium in genetically defined Europeans (p<0.0001). Statistical analyses were performed as previously described [8], using an inverse variance-weighted fixed-effects meta-analysis. At the 12q13.13 locus, we prioritized 14 SNPs whose p-values indicated strong linkage disequilibrium (<1E-5) (listed in **Supplementary File 1j**). These included two SNPs with p-values reaching formal genome-wide significance (5E-8): rs11170344, the lead SNP in our re-analysis; and rs3741442, which was identified in a recent GWAS in a Chinese orofacial cleft population. [9]

Zebrafish lines and maintenance

D. rerio were maintained in the University of Iowa Animal Care Facility according to a standard protocol (protocol no. 6011616). [96] All zebrafish experiments were performed in compliance with the ethical regulations of the Institutional Animal Care and Use Committee at the University of Iowa and in compliance with NIH guidelines. Zebrafish embryos were maintained at 28.5°C, and staged by hours or days post-fertilization (hpf or dpf).

Mouse maintenance

All C57BL/6 mouse experiments used for ATAC-seq library prepara were performed in accordance with approval of the Institutional Animal Care and Use Committees at the School and Hospital of Stomatology of Wuhan University (protocol no. 00271454). Mouse experiments for *LacZ* reporter transgenic animal work performed at the Lawrence Berkeley National Laboratory (LBNL) were reviewed and approved by the LBNL Animal Welfare and Research Committee. Transgenic mouse assays were performed in *Mus musculus* FVB strain mice.

Cell culture

GMSM-K human embryonic oral epithelial cell line (a kind gift from Dr. Daniel Grenier) [85] and the human immortalized oral epithelial cell (HIOEC) line [86] were maintained in keratinocyte serum-free medium (Life Technologies, Carlsbad, CA) supplemented with EGF and bovine pituitary extract (Life Technologies). HEPM human embryonic palatal mesenchyme cells (ATCC CRL-1486) were maintained in DMEM (Hyclone, Pittsburgh, PA) supplemented with 10% fatal bovine serum (Hyclone). All the cell lines used in this study were tested for mycoplasma contamination and authenticated by genetic profiling using polymorphic short tandem repeats.

Electroporation and dual luciferase assay

For dual luciferase assays, each reporter construct was co-transfected with Renilla luciferase plasmid and three biological replicates were used. Briefly, GMSM-K cells were electroporated with plasmid using the Amaxa Cell Line Nucleofector Kit (Lonza, Cologne, Germany) and the Nucleofector II instrument (Lonza). We used a dual-luciferase reporter assay system (Promega, Madison, WI) and 20/20n Luminometer (Turner Biosystems, Sunnyvale, CA) to evaluate the luciferase activity following manufacturer's instructions. Relative luciferase activity was calculated as the ratio between the value for the firefly and Renilla enzymes. Three independent measurements were performed for each transfection group. All results are presented as mean ± s.d. Statistical significance was determined using the Student's *t*-test.

Plasmid constructs and transient reporter analysis of potential periderm enhancer in zebrafish and mouse

All candidate enhancer elements described were cloned using zebrafish or human genomic DNA, and were harvested from either zebrafish embryos or a human immortalized oral keratinocyte cell line (GMSM-K) [85]. Products were cloned into the pENTR/D-TOPO plasmid (Life Technologies, Carlsbad, CA) and validated by Sanger sequencing. Site-directed mutagenesis was used to generate elements lacking the corresponding motifs or containing a risk variant. All elements were subcloned into the *cFos-GFP* plasmid (a gift from Shannon Fisher) [37] or *cFos-tdTomato*, a derivative of *cFos-GFP*, or *cFos-GFP; Cry-GFP*, a derivative that includes a lens-specific promoter (cloning details available upon request). For each reporter construct, at least 100 embryos at the 1-cell to 2-cell stage were injected (20 pg reporter construct plus 20 pg tol2 mRNA); three replicates were performed, each on a different day [37]. Embryos injected were examined by epifluorescence microscopy first at approximately 11 hpf, then each day subsequently until approximately 4 dpf. SNP1 and SNP2 701 bp elements, used in GFP and *LacZ* reporter constructs were (SNP1) chr12:53,340,250-53,340,950 and (SNP2) chr12:53,343,968-53,344,668 (hg19). Specifically, for mouse reporter assay, candidate enhancers were PCR-amplified and cloned upstream of a *Shh*-promoter-*LacZ*-reporter cassette. We used a mouse enhancer-reporter assay that relies on site-specific integration of a transgene into the mouse genome [63]. In this assay, the reporter cassette is flanked by homology arms targeting the H11 safe harbor locus [97]. Cas9 protein and a sgRNA targeting H11 were co-injected into the pronucleus of FVB single cell stage mouse embryos (E0.5) together with the reporter vector [63]. Embryos were sampled and stained at E13.5. Embryos were only excluded from further analysis if they did not carry the reporter transgene. Transgene copy number was estimated by qPCR using a TaqMan probe targeting *Shh* promoter.

Dissociation of zebrafish embryos and FACS

About 500 *Tg(krt4:GFP)* [27] embryos were collected at the 4-somite stage and rinsed with PBS without Ca^{2+} or Mg^{2+} (Life Technologies). Embryos were dechorionated using pronase (Sigma, St. Louis, MO, 1 mg/mL in fish water 1 mg/mL in fish water) at room temperature for 10 minutes, rinsed in PBS, and then dissociated cells using a pestle and incubated in trypsin (0.25%)-EDTA (Life technology) at 33°C for 30 minutes. Reactions were stopped by adding PBS supplemented with 5% fetal bovine serum (Life technologies). Dissociated cells were re-suspended into single-cell solution and analyzed at the University of Iowa Flow Cytometry Facility, using an Aria Fusion instrument (Becton Dickinson, Franklin Lakes, NJ)

Dissociation of mouse palatal epithelium

Mouse embryos at were collected at E14.5 and palate shelves were dissected. The multi-layered palate shelf epithelium was manually isolated as described previously [98]. Briefly, palatal shelves isolated from the frontal facial prominence were incubated in 0.25% trypsin-EDTA (Life Technologies) at 4°C for 10 minutes, after which the reaction was stopped using Trypsin Inhibitor (Life Technologies). Under a dissecting microscope, the epithelium was isolated for ATAC-seq by gently peeling using microforceps. The remaining tissue, comprised largely of mesenchymal cells, was also collected for control samples. This isolation protocol was previously described as a method for harvesting oral periderm [98], but which sample(s) contained basal oral epithelial cells remained unclear. Approximately 20,000 epithelial cells and the same number of mesenchymal cells were harvested from 7 embryos at E14.5. And 20,000 cells were used to prepare one ATAC-seq library.

Preparation of RNA-seq libraries and high-throughput sequencing

Three independent biological replicates were subjected to RNA-seq profiling. In each replicate, we isolated 20,000 peridermal and non-peridermal cells from *Tg(krt4:GFP)* embryos at the 4-somite stage. Total RNA was extracted from the sorted cells using the RNAqueous® Total RNA Isolation Kit (Ambion, Foster City, CA) and treated with Turbo DNase I (Ambion, Austin, TX) to remove residual genomic DNA. The treated RNA was then purified and concentrated using the RNA Clean & Concentrator™-5 Kit (Zymo Research, Irvine, CA). After quantification with Qubit 3.0 (Life Technologies) and quality control with the Agilent RNA 6000 Pico Kit on Agilent 2100 (Agilent Technologies, Santa Clara, CA), 1 ug of RNA was subjected to first-strand cDNA synthesis and cDNA amplification using the SMART-Seq® v4 Ultra® Low Input RNA Kit (Takara Bio, Kusatsu, Shiga, Japan). Purified cDNA was quantified using Qubit (Life Technologies), and for each library 150 pg cDNA was used as input with the Nextera XT DNA Sample Preparation Kit (Illumina, San Diego, CA), following the manufacturer's instructions. Each DNA library was quantified using a KAPA Library Quantification Kit (Roche, Mannheim, Germany) and pooled for HiSeq4000 (Illumina) high-throughput sequencing at same molarity.

RNA-seq data analysis

RNA-seq raw reads data was trimmed using the Trimmomatic (Usadel Lab, Aachen, Germany. v0.36) [88] (parameter: ILLUMINACLIP:TrueSeq3PE-PE.fa:2:30:10:8:TRUE LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:30), and aligned to zv10 cDNA reference sequence using the kallisto (Pachter Lab, California Institute of Technology, Pasadena, CA) aligner [99] with the default parameters. The output from kallisto was quantified using the sleuth R package (Pachter Lab) [100] according to a standard protocol. We used q-value <0.01 and p-value<0.01 as the threshold for significant difference. Gene set enrichment analysis was performed using GSEA (v 3.0) [101]. Gene ontology (GO) enrichment analysis was performed using the Metascape online tool (http://metascape.org/gp/index.html) [102], and the top GO categories were selected according to the binomial *P* values. Raw and processed sequencing data for RNA-seq were deposited in GEO repository (GSE140241).

Chromatin immunoprecipitation of H3K27 acetylation (H3K27Ac) combined with high throughput sequencing (ChIP-seq)

HIOEC cells were seeded at 1x10^5 cells per 100mm plate (1 plate per biological replicate), grown to 90-100% confluency (refreshed medium every other day) and subjected to 1.2 mM Ca^{2+} in culture medium for 3 days. Cell were washed with ice-cold PBS (Hyclone) and fixed with 1% paraformaldehyde (PFA) for 10min at room temperature. PFA was then quenched in 134M Glycine (Sigma) for 5 min at room temperature, and cells were collected with a cell scraper in ice-cold PBS. After centrifuge for 10 mins at 500g, the cell pellets were resuspended with 5mM PIPES pH8.5, 85mM KCl, 1% (v/v) IGEPAL CA-630, 50mM NaF, 1mM PMSF, 1mM phenylarsine oxide, 5mM Sodium Orthovanadate and protease inhibitor cocktail (Roche, Germany). After sonication, chromatin immunoprecipitation was performed using 4ug of anti-Histone H3, Acetylated Lysine 27 (H3K27Ac) (Abcam, Cambridge, UK, ab4729, lot NO. GR3211959-1) per 500,000 cells. ChIP-seq library were indexed with kit (ND606-01, Vanzyme, China). 150-bp-paired-end sequencing was performed using the HiSeq X Ten sequencer (Illumina, provided by Annoroad Genomics, China). Output sequences were trimmed using Trimmomatic (v0.38) [88] (parameter: ILLUMINACLIP:TrueSeq3PE-PE.fa:2:30:10:8:TRUE LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:30). All trimmed, paired reads were aligned to human genome assembly 19 (hg19) using Bowtie 2 (Johns Hopkins

University, Baltimore, MD, default parameters).[87] Peaks were called using MACS2 (v2.1.1) [94]. DeepTools (Max Planck Institute for Immunobiology and Epigenetics, Freiburg, Germany, v 2.0) was used to confirm reproducibility of the biological replicates and generate bigWig coverage files for visualization [95]. Raw and processed sequencing data for this H3K27Ac ChIP-seq were deposited in GEO repository (GSE139809).

Preparation of ATAC-seq library and high-throughput sequencing

We prepared the ATAC-seq library according to a previously published protocol [31]. Briefly, sorted cells were lysed with 50 µL cold lysis buffer (10 mM Tris--HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% NP-40; all components purchased from Sigma) in a centrifuge at 500 x g for 15 minutes at 4°C. Pelleted nuclei were resuspended in 50 µL tagmentation reaction mix (25 µL Nextera TD Buffer, 2.5 µL Nextera TD Enzyme, and 22.5 µL H_2O, all from the Nextera DNA Sample Preparation Kit [Illumina]). Tagmentation was performed at 37°C for 30 min in a thermocycler and, immediately after the reaction was completed, the DNA was purified using a Qiagen PCR Purification MinElute Kit (QIAGEN, Germantown, MD) and eluted with 10 µL elution buffer. Eluted DNA was subjected to PCR amplification and library indexing, using the NEBNext® High-Fidelity 2x PCR Master Mix (New England Biolabs, Ipswich, MA) with a customized Nextera PCR primer pair, according to the following program: 72°C for 5 minutes; 98°C for 30 seconds; 11 cycles of 98°C for 10 seconds, 63°C for 30 seconds, and 72°C for 1 minute; and hold at 4°C. The PCR product was purified with 1.8 x volume (90 µL for each sample) of Ampure XP beads (Beckman Coulter, Brea, CA) to produce 18 µL of final library. Library quality was assessed using 1 µL of the final purified DNA on a BioAnalyzer 2100 High Sensitivity DNA Chip (Agilent Technologies). All DNA libraries that exhibited a nucleosome pattern in the BioAnalyzer 2100 Assay passed the pre-sequencing QC process and were pooled for high-throughput sequencing in HiSeq 2500, HiSeq4000, or HiSeq X Ten (Illumina, provided by Annoroad Genomics Company (China)).

Mapping of ATAC-seq reads and calling of peaks and differential peaks

Raw ATAC-seq reads were trimmed using Trimmomatic (v 0.38) [88] (parameter : ILLUMINACLIP:NexteraPE-PE.fa:2:30:10:8:TRUE LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:5) and mapped to either the danRer7, hg19 or mm10 reference genome using Bowtie 2 [87] (default parameters). Sorting, removal of PCR duplicates and conversion from SAM to BAM files were performed using SAMtools [93]. A customized Python script was used to identify fragments shorter than 100 bp as the nucleosome-free-regions (NFRs), as previously described (custom scripts and piplines we deployed are available at https://github.com/Badgerliu/periderm_ATACSeq) [31]. The Picard toolset (http://broadinstitute.github.io/picard/) was used to check fragment size distribution. We employed DeepTools (v 2.0) to check the reproducibility of the biological replicates and generate bigWig coverage files for visualization [95]. Peaks were called using MACS2 (v2.1.1) [94] (parameter: --nomodel --nolambda --gsize 1.4e9 --keep-dup all --slocal 10000 --extsize 54 for zebrafish periderm ATAC-seq, and --nomodel --nolambda --gsize 2.7e9 --keep-dup all --slocal 10000 for mouse periderm or HOIEC ATAC-seq). Differentially accessible NFRs were identified using an R package DiffBind [89] with a fold-change threshold of 0.5, and FDR < 0.01.

Raw and processed sequencing data for zebrafish, mouse and human ATAC-seq were deposited in GEO repository (GSE140241, GSE139945 and GSE139809).

Integration of ATAC-seq and H3K27Ac ChIP-seq data

To compare our zebrafish periderm ATAC-seq results to those in previously published whole-embryo H3K27Ac ChIP-seq studies[34], we retrieved the single-end, raw read data from GEO Series GSE32483 H3K27Ac ChIP-seq from whole zebrafish embryos at several stages. Raw data were aligned to the danRer7 reference genome using Bowtie 2, and peaks were called using MACS2. Following an earlier study[14], we defined “H3K27Ac-flanked” regions as those between adjacent H3K27Ac peaks separated by up to 1500 bp. NFRs were identified in cells isolated at 11 hpf; GFP-positive active elements (GPAEs) were the GFP-positive NFRs that overlap H3K27Ac peaks, or H3K27Ac flanked regions, at 8.3 hpf (80% epiboly) and/or at 24 hpf embryos; GNAEs were the GFP-negative NFRs that did so.

The approach used to identify enhancers that are active in the mouse palate epithelium was similar. It involved integrating mouse palate-epithelium-specific NFRs found in this study with previously published H3K27Ac ChIP-seq data for E14.5 mouse embryonic facial prominences (GSE82727) [54]. We also identified enhancers that are active in HIOECs by integrating HIOEC-enriched NFRs with the H3K27Ac ChIP-seq results from this study.

We used seqMINER (v 1.2.1) [90] to calculate the normalized reads matrix for each NFR of interest, generating a matrix file for the downstream heatmap and density plot in R.

Assignment of ATAC-seq peaks to genes and gene ontology analysis

The Genomic Regions Enrichment of Annotations Tool (GREAT, http://great.stanford.edu/public/html/) [36] was employed to assign genes proximal to genomic regions of interest (i.e., ATAC-seq and high-scoring elements), using the following rule set: two nearest genes within 100kb. We also used GREAT to identify GO terms for which the set of “closest genes” was enriched.

Comparisons of peak accessibility and gene expression

To compare ATAC-seq accessibility and relative gene expression between tissues, we first identified genes for which zebrafish peridermal and non-peridermal tissue was enriched using Sleuth as described above. We then identified the normalized ATAC-seq accessibility using the EdgeR package embedded in the DiffBind analysis suite (cutoff: fold change >0.5 or <-0.5, p-value<0.01). We used GREAT to associate the periderm- and non-periderm-enriched genes with their tissue-specific ATAC-seq peaks. To determine whether the accessibility of periderm-specific ATAC-seq correlates with gene expression in the periderm, we compared the accessibility of tissue-specific ATAC-seq peaks (values normalized) within genes for which either peridermal or non-peridermal tissue is enriched, as well as the levels of expression of genes associated with periderm or non-periderm ATAC-seq peaks.

Motif enrichment analysis and footprinting for periderm-enriched motifs

We identified the *de novo* motifs for which the genomic regions of interest are enriched using the findMotifsGenome.pl function of HOMER [91] (parameter: -len 8,10,12), and assigned the most enriched motifs to the transcription factors with highest expression in related tissues. For the Tn5 footprint analysis, we shifted all reads aligned to the plus strand by +4 bp, and all reads aligned to the minus strand by -5 bp. To predict the binding of members of the GRHL, KLF, TFAP2, and C/EBP transcription factor families, we downloaded the related motifs in all transcription factors of interest (http://cisbp.ccbr.utoronto.ca/) [103], and calculated the Tn5 cleavage frequency in the +/-100 bp sequence flanking the motifs of interest, using CENTEPEDE [40].

For analysis of the potential clustering pattern of periderm-enriched motif combination within zebrafish GPAEs, we firstly annotated all the GPAEs using HOMER annotatePeaks function with the motif files for GRHL, TEAD, KLF, FOS, TFAP2, GATA and CEBP, and counted the occurrence of each motif in each peak. Hierarchy clustering was then performed on the occurrence of different motif in each peak using the “hierarchy_cluster_motif_combination_pattern_in_GPAE.R” script deposited in periderm_ATACSeq github repository. We also counted the sum of every two-motif-combination and three-motif-combination in each GPAE using “motif_combination_count.R” script deposited in github.

Comparison of TFBS enrichment

To determine the number of binding sites that would have been shared by chance, we generated 10 sets of 4,000 randomly-selected 400-bp sequences from two species and assessed the average number of transcription factors predicted to bind sequences enriched in both species. Transcription factors receiving a score of 0.8 in HOMER output [91] were considered to match the binding site.

Construction of network depicting the regulatory relationships among periderm-enriched transcription factors

To assign periderm signature motifs to the periderm-enriched genes, we first annotated all periderm-specific NFRs with signature motifs using HOMER, then calculated the total number of times each motif was present in all of periderm-specific NFRs near the periderm-enriched transcription factor. Expression levels of each transcription factor in the periderm were also taken into account.

Training of a gapped k-mer support vector machine on zebrafish periderm enhancers

All GFP-positive active enhancers (GPAEs) were resized into 400-bp regions that maximize the overall ATAC-seq signal within each NFR; all GPAEs >70% repeats were removed. Repeat fractions were calculated using repeat masked sequence data (danRer7) from the UCSC Genome Browser (http://genome.ucsc.edu/). A supervised-machine-learning classifier, gapped k-mer support vector machine (gkmSVM) was used to generate a 10-fold larger set of random genomic 400-bp sequences in the danRer7 reference genome, based on matching of GC and repeat fraction of the positive training set. gkmSVM were performed to generate a scoring vector (parameter: K=6, L=10). Related ROC and PRC were generated using gkmSVM output.[6] For genome-wide enhancer predictions, mouse (mm10) and human (hg19) genomes were segmented into 400-bp regions with 300-bp overlap, and all regions were scored using a gkmSVM script.

Tests of enhancer homology

The following DNA fragments were used to test homology of the human and zebrafish enhancers. The 489 bp sequence corresponding to the human *ppl* periderm enhancer (plus-strand) was trimmed to a conserved 400 bp core block (“Hu_400+”), which lacks an overlapping non-conserved AluJr4 SINE on the *ppl*-proximal side and a non-conserved MIR SINE on the *ppl*-distal side. The homologous mouse sequence to the human 400 bp core enhancer was determined to be a 409 bp fragment (“Mm_409+”). A zebrafish 467 bp core fragment (“Zf_467+”) was identified to correspond to be the block most similar to the mammalian *ppl* enhancer core. All three core fragments lie 8.5 kb (human), 4.0 kb (mouse), and 10.4 kb (zebrafish) upstream of *ppl*, which is transcribed to the left in each case. To evaluate homology between Hu_400+ and Zf_467+ enhancer fragments we performed pairwise alignments between various sequences using CLUSTALW and default parameters as available via Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/). The Hu_400+ | Mm_409+ pairwise alignment constituted the homologous control to compare to the Hu_400+ | Zf_467+ test alignment. For different types of negative controls, we also aligned the Hu_400+ sequence to the mouse and fish sequences corresponding to the minus-strand (i.e., reverse-complements "Mm_409-" and "Zf_467-"), the non-biological reverse sequence ("Mm_409R" and "Zf_467R"), and three different Fisher-Yates shuffled versions of the mouse and zebrafish plus-strand sequences ("Mm_409+S1/S2/3" and "Zf_467+S1/S2/S3"). Last, we also compared different 3-way alignments involving Hu_400+, Mm_409+, and the various zebrafish test and controls (**Supplementary File 2a** and **Supplementary File 2b**).

Annotation of potential zebrafish periderm enhancer candidates using ENCODE/Roadmap data

All mapped data from the Roadmap Epigenomics Project [42] (http://www.roadmapepigenomics.org/) were downloaded as BAM files; imputed enhancer regions in each cell/tissue type were also downloaded. Using the BEDTools (v2.24.0) [92] intersect function, we evaluated the fraction of high-scored elements that overlapped with enhancer regions in each cell/tissue type.

Analysis of transcription factor binding sites affected by alleles of SNP1 and SNP2.

We used 19bp sequences centered on the SNP1 and SNP2, with risk or non-risk alleles of these SNPs, as input to JASPAR (http://jaspar.genereg.net)[104]. We queried all 1011 transcription factor binding site profiles using a relative profile score threshold of 80%. "Sites lost" were those at a particular start position with a score of 5.0 or higher in the sequence with the non-risk allele and with a score less than 2.0, or not detected, in the sequence with the risk allele. "Sites gained" had the opposite pattern.

References:

1. Little, J., and Bryan, E. (1986). Congenital anomalies in twins. Semin Perinatol 10, 50-64.
2. Takahashi, M., Hosomichi, K., Yamaguchi, T., Nagahama, R., Yoshida, H., Maki, K., Marazita, M.L., Weinberg, S.M., and Tajima, A. (2018). Whole-genome sequencing in a pair of monozygotic twins with discordant cleft lip and palate subtypes. Oral Dis 24, 1303-1309.
3. Saleem, K., Zaib, T., Sun, W., and Fu, S. (2019). Assessment of candidate genes and genetic heterogeneity in human non syndromic orofacial clefts specifically non syndromic cleft lip with or without palate. Heliyon 5, e03019.
4. Liu, H., Leslie, E.J., Carlson, J.C., Beaty, T.H., Marazita, M.L., Lidral, A.C., and Cornell, R.A. (2017). Identification of common non-coding variants at 1p22 that are functional for non-syndromic orofacial clefting. Nature communications 8, 14759.
5. Ghandi, M., Lee, D., Mohammad-Noori, M., and Beer, M.A. (2014). Enhanced regulatory sequence prediction using gapped k-mer features. PLoS Comput Biol 10, e1003711.
6. Ghandi, M., Mohammad-Noori, M., Ghareghani, N., Lee, D., Garraway, L., and Beer, M.A. (2016). gkmSVM: an R package for gapped-kmer SVM. Bioinformatics 32, 2205-2207.
7. Lee, D., Gorkin, D.U., Baker, M., Strober, B.J., Asoni, A.L., McCallion, A.S., and Beer, M.A. (2015). A method to predict the impact of regulatory variants from DNA sequence. Nat Genet 47, 955-961.
8. Leslie, E.J., Carlson, J.C., Shaffer, J.R., Butali, A., Buxo, C.J., Castilla, E.E., Christensen, K., Deleyiannis, F.W., Leigh Field, L., Hecht, J.T., et al. (2017). Genome-wide meta-analyses of nonsyndromic orofacial clefts identify novel associations between FOXE1 and all orofacial clefts, and TP63 and cleft lip with or without cleft palate. Hum Genet.
9. 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. Nature communications 8, 14364.
10. Vaziri Sani, F., Kaartinen, V., El Shahawy, M., Linde, A., and Gritli-Linde, A. (2010). Developmental changes in cellular and extracellular structural macromolecules in the secondary palate and in the nasal cavity of the mouse. European journal of oral sciences 118, 221-236.
11. Dale, B.A., Holbrook, K.A., Kimball, J.R., Hoff, M., and Sun, T.T. (1985). Expression of epidermal keratins and filaggrin during human fetal skin development. J Cell Biol 101, 1257-1269.
12. Moll, R., Moll, I., and Wiest, W. (1982). Changes in the pattern of cytokeratin polypeptides in epidermis and hair follicles during skin development in human fetuses. Differentiation; research in biological diversity 23, 170-178.
13. Fisher, S., Grice, E.A., Vinton, R.M., Bessling, S.L., and McCallion, A.S. (2006). Conservation of RET regulatory function from human to zebrafish without sequence similarity. Science 312, 276-279.
14. Gorkin, D.U., Lee, D., Reed, X., Fletez-Brant, C., Bessling, S.L., Loftus, S.K., Beer, M.A., Pavan, W.J., and McCallion, A.S. (2012). Integration of ChIP-seq and machine learning reveals enhancers and a predictive regulatory sequence vocabulary in melanocytes. Genome Res 22, 2290-2301.
15. Chen, L., Fish, A.E., and Capra, J.A. (2018). Prediction of gene regulatory enhancers across species reveals evolutionarily conserved sequence properties. PLoS Comput Biol 14, e1006484.
16. Richardson, R.J., Hammond, N.L., Coulombe, P.A., Saloranta, C., Nousiainen, H.O., Salonen, R., Berry, A., Hanley, N., Headon, D., Karikoski, R., et al. (2014). Periderm prevents pathological epithelial adhesions during embryogenesis. J Clin Invest 124, 3891-3900.
17. Warga, R.M., and Kimmel, C.B. (1990). Cell movements during epiboly and gastrulation in zebrafish. Development 108, 569-580.
18. Lee, R.T., Asharani, P.V., and Carney, T.J. (2014). Basal keratinocytes contribute to all strata of the adult zebrafish epidermis. PloS one 9, e84858.
19. Fukazawa, C., Santiago, C., Park, K.M., Deery, W.J., Gomez de la Torre Canny, S., Holterhoff, C.K., and Wagner, D.S. (2010). poky/chuk/ikk1 is required for differentiation of the zebrafish embryonic epidermis. Dev Biol.

20. Sabel, J.L., d'Alencon, C., O'Brien, E.K., Otterloo, E.V., Lutz, K., Cuykendall, T.N., Schutte, B.C., Houston, D.W., and Cornell, R.A. (2009). Maternal Interferon Regulatory Factor 6 is required for the differentiation of primary superficial epithelia in Danio and Xenopus embryos. Dev Biol 325, 249-262.
21. de la Garza, G., Schleiffarth, J.R., Dunnwald, M., Mankad, A., Weirather, J.L., Bonde, G., Butcher, S., Mansour, T.A., Kousa, Y.A., Fukazawa, C.F., et al. (2013). Interferon regulatory factor 6 promotes differentiation of the periderm by activating expression of Grainyhead-like 3. The Journal of investigative dermatology 133, 68-77.
22. Liu, H., Leslie, E.J., Jia, Z., Smith, T., Eshete, M., Butali, A., Dunnwald, M., Murray, J., and Cornell, R.A. (2016). Irf6 directly regulates Klf17 in zebrafish periderm and Klf4 in murine oral epithelium, and dominant-negative KLF4 variants are present in patients with cleft lip and palate. Hum Mol Genet 25, 766-776.
23. Pei, W., Noushmehr, H., Costa, J., Ouspenskaia, M.V., Elkahloun, A.G., and Feldman, B. (2007). An early requirement for maternal FoxH1 during zebrafish gastrulation. Dev Biol.
24. Miles, L.B., Darido, C., Kaslin, J., Heath, J.K., Jane, S.M., and Dworkin, S. (2017). Mis-expression of grainyhead-like transcription factors in zebrafish leads to defects in enveloping layer (EVL) integrity, cellular morphogenesis and axial extension. Sci Rep 7, 17607.
25. 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.
26. Peyrard-Janvid, M., Leslie, E.J., Kousa, Y.A., Smith, T.L., Dunnwald, M., Magnusson, M., Lentz, B.A., Unneberg, P., Fransson, I., Koillinen, H.K., et al. (2014). Dominant mutations in GRHL3 cause Van der Woude Syndrome and disrupt oral periderm development. Am J Hum Genet 94, 23-32.
27. Gong, Z., Ju, B., Wang, X., He, J., Wan, H., Sudha, P.M., and Yan, T. (2002). Green fluorescent protein expression in germ-line transmitted transgenic zebrafish under a stratified epithelial promoter from keratin8. Dev Dyn 223, 204-215.
28. O'Brien, G.S., Rieger, S., Wang, F., Smolen, G.A., Gonzalez, R.E., Buchanan, J., and Sagasti, A. (2012). Coordinate development of skin cells and cutaneous sensory axons in zebrafish. J Comp Neurol 520, 816-831.
29. Buenrostro, J.D., Wu, B., Chang, H.Y., and Greenleaf, W.J. (2015). ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr Protoc Mol Biol 109, 21 29 21-29.
30. Quillien, A., Abdalla, M., Yu, J., Ou, J., Zhu, L.J., and Lawson, N.D. (2017). Robust Identification of Developmentally Active Endothelial Enhancers in Zebrafish Using FANS-Assisted ATAC-Seq. Cell Rep 20, 709-720.
31. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature methods 10, 1213-1218.
32. Iwafuchi-Doi, M., Donahue, G., Kakumanu, A., Watts, J.A., Mahony, S., Pugh, B.F., Lee, D., Kaestner, K.H., and Zaret, K.S. (2016). The Pioneer Transcription Factor FoxA Maintains an Accessible Nucleosome Configuration at Enhancers for Tissue-Specific Gene Activation. Mol Cell 62, 79-91.
33. Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A 107, 21931-21936.
34. Bogdanovic, O., Fernandez-Minan, A., Tena, J.J., de la Calle-Mustienes, E., Hidalgo, C., van Kruysbergen, I., van Heeringen, S.J., Veenstra, G.J., and Gomez-Skarmeta, J.L. (2012). Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis. Genome Res.
35. Bogdanovic, O., van Heeringen, S.J., and Veenstra, G.J. (2012). The epigenome in early vertebrate development. Genesis 50, 192-206.
36. McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nature biotechnology 28, 495-501.

37. Fisher, S., Grice, E.A., Vinton, R.M., Bessling, S.L., Urasaki, A., Kawakami, K., and McCallion, A.S. (2006). Evaluating the biological relevance of putative enhancers using Tol2 transposon-mediated transgenesis in zebrafish. Nature protocols 1, 1297-1305.
38. Farrell, J.A., Wang, Y., Riesenfeld, S.J., Shekhar, K., Regev, A., and Schier, A.F. (2018). Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. Science 360.
39. Wagner, D.E., Weinreb, C., Collins, Z.M., Briggs, J.A., Megason, S.G., and Klein, A.M. (2018). Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. Science 360, 981-987.
40. Pique-Regi, R., Degner, J.F., Pai, A.A., Gaffney, D.J., Gilad, Y., and Pritchard, J.K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res 21, 447-455.
41. Visel, A., Prabhakar, S., Akiyama, J.A., Shoukry, M., Lewis, K.D., Holt, A., Plajzer-Frick, I., Afzal, V., Rubin, E.M., and Pennacchio, L.A. (2008). Ultraconservation identifies a small subset of extremely constrained developmental enhancers. Nat Genet 40, 158-160.
42. Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature 518, 317-330.
43. Li, H., Jones, K.L., Hooper, J.E., and Williams, T. (2019). The molecular anatomy of mammalian upper lip and primary palate fusion at single cell resolution. Development.
44. Fakhouri, W.D., Rhea, L., Du, T., Sweezer, E., Morrison, H., Fitzpatrick, D., Yang, B., Dunnwald, M., and Schutte, B.C. (2012). MCS9.7 enhancer activity is highly, but not completely, associated with expression of Irf6 and p63. Dev Dyn 241, 340-349.
45. Fakhouri, W.D., Rahimov, F., Attanasio, C., Kouwenhoven, E.N., Ferreira De Lima, R.L., Felix, T.M., Nitschke, L., Huver, D., Barrons, J., Kousa, Y.A., et al. (2014). An etiologic regulatory mutation in IRF6 with loss- and gain-of-function effects. Hum Mol Genet 23, 2711-2720.
46. Rahimov, F., Marazita, M.L., Visel, A., Cooper, M.E., Hitchler, M.J., Rubini, M., Domann, F.E., Govil, M., Christensen, K., Bille, C., et al. (2008). Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip. Nat Genet 40, 1341-1347.
47. Botti, E., Spallone, G., Moretti, F., Marinari, B., Pinetti, V., Galanti, S., De Meo, P.D., De Nicola, F., Ganci, F., Castrignano, T., et al. (2011). Developmental factor IRF6 exhibits tumor suppressor activity in squamous cell carcinomas. Proceedings of the National Academy of Sciences of the United States of America.
48. Boxer, L.D., Barajas, B., Tao, S., Zhang, J., and Khavari, P.A. (2014). ZNF750 interacts with KLF4 and RCOR1, KDM1A, and CTBP1/2 chromatin regulators to repress epidermal progenitor genes and induce differentiation genes. Genes Dev 28, 2013-2026.
49. Kouwenhoven, E.N., van Heeringen, S.J., Tena, J.J., Oti, M., Dutilh, B.E., Alonso, M.E., de la Calle-Mustienes, E., Smeenk, L., Rinne, T., Parsaulian, L., et al. (2010). Genome-wide profiling of p63 DNA-binding sites identifies an element that regulates gene expression during limb development in the 7q21 SHFM1 locus. PLoS genetics 6, e1001065.
50. Rubin, A.J., Barajas, B.C., Furlan-Magaril, M., Lopez-Pajares, V., Mumbach, M.R., Howard, I., Kim, D.S., Boxer, L.D., Cairns, J., Spivakov, M., et al. (2017). Lineage-specific dynamic and pre-established enhancer-promoter contacts cooperate in terminal differentiation. Nat Genet.
51. Richardson, R., Mitchell, K., Hammond, N.L., Mollo, M.R., Kouwenhoven, E.N., Wyatt, N.D., Donaldson, I.J., Zeef, L., Burgis, T., Blance, R., et al. (2017). p63 exerts spatio-temporal control of palatal epithelial cell fate to prevent cleft palate. PLoS genetics 13, e1006828.
52. Sevilla, L.M., Nachat, R., Groot, K.R., Klement, J.F., Uitto, J., Djian, P., Maatta, A., and Watt, F.M. (2007). Mice deficient in involucrin, envoplakin, and periplakin have a defective epidermal barrier. J Cell Biol 179, 1599-1612.
53. Klein, R.H., and Andersen, B. (2015). Dynamic networking for epidermal differentiation. Developmental cell 32, 661-662.
54. Consortium, E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57-74.

55. 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.
56. Foshay, K.M., and Gallicano, G.I. (2009). miR-17 family miRNAs are expressed during early mammalian development and regulate stem cell differentiation. Dev Biol 326, 431-443.
57. McGowan, K.M., and Coulombe, P.A. (1998). Onset of keratin 17 expression coincides with the definition of major epithelial lineages during skin development. J Cell Biol 143, 469-486.
58. Tadeu, A.M., and Horsley, V. (2013). Notch signaling represses p63 expression in the developing surface ectoderm. Development 140, 3777-3786.
59. Gong, S.G., Gong, T.W., and Shum, L. (2005). Identification of markers of the midface. J Dent Res 84, 69-72.
60. Appert, A., Fridmacher, V., Locquet, O., and Magre, S. (1998). Patterns of keratins 8, 18 and 19 during gonadal differentiation in the mouse: sex- and time-dependent expression of keratin 19. Differentiation; research in biological diversity 63, 273-284.
61. Erman, A., Veranic, P., Psenicnik, M., and Jezernik, K. (2006). Superficial cell differentiation during embryonic and postnatal development of mouse urothelium. Tissue Cell 38, 293-301.
62. Diez-Roux, G., Banfi, S., Sultan, M., Geffers, L., Anand, S., Rozado, D., Magen, A., Canidio, E., Pagani, M., Peluso, I., et al. (2011). A high-resolution anatomical atlas of the transcriptome in the mouse embryo. PLoS Biol 9, e1000582.
63. Kvon, E.Z., Zhu, Y., Kelman, G., Novak, C.S., Plajzer-Frick, I., Kato, M., Garvin, T.H., Pham, Q., Harrington, A.N., Hunter, R.D., et al. (2020). Comprehensive In Vivo Interrogation Reveals Phenotypic Impact of Human Enhancer Variants. Cell submitted.
64. Murray, S.A., Oram, K.F., and Gridley, T. (2007). Multiple functions of Snail family genes during palate development in mice. Development 134, 1789-1797.
65. Wilkerson, B.A., Chitsazan, A.D., VandenBosch, L.S., Wilken, M.S., Reh, T.A., and Bermingham-McDonogh, O. (2019). Open chromatin dynamics in prosensory cells of the embryonic mouse cochlea. Sci Rep 9, 9060.
66. Cokus, S.J., De La Torre, M., Medina, E.F., Rasmussen, J.P., Ramirez-Gutierrez, J., Sagasti, A., and Wang, F. (2019). Tissue-Specific Transcriptomes Reveal Gene Expression Trajectories in Two Maturing Skin Epithelial Layers in Zebrafish Embryos. G3 (Bethesda) 9, 3439-3452.
67. Lowe, E.K., Cuomo, C., Voronov, D., and Arnone, M.I. (2019). Using ATAC-seq and RNA-seq to increase resolution in GRN connectivity. Methods Cell Biol 151, 115-126.
68. Miraldi, E.R., Pokrovskii, M., Watters, A., Castro, D.M., De Veaux, N., Hall, J.A., Lee, J.Y., Ciofani, M., Madar, A., Carriero, N., et al. (2019). Leveraging chromatin accessibility for transcriptional regulatory network inference in T Helper 17 Cells. Genome Res 29, 449-463.
69. Gordon, W.M., Zeller, M.D., Klein, R.H., Swindell, W.R., Ho, H., Espetia, F., Gudjonsson, J.E., Baldi, P.F., and Andersen, B. (2014). A GRHL3-regulated repair pathway suppresses immune-mediated epidermal hyperplasia. J Clin Invest 124, 5205-5218.
70. Nishino, H., Takano, S., Yoshitomi, H., Suzuki, K., Kagawa, S., Shimazaki, R., Shimizu, H., Furukawa, K., Miyazaki, M., and Ohtsuka, M. (2017). Grainyhead-like 2 (GRHL2) regulates epithelial plasticity in pancreatic cancer progression. Cancer Med 6, 2686-2696.
71. Elbediwy, A., Vincent-Mistiaen, Z.I., Spencer-Dene, B., Stone, R.K., Boeing, S., Wculek, S.K., Cordero, J., Tan, E.H., Ridgway, R., Brunton, V.G., et al. (2016). Integrin signalling regulates YAP and TAZ to control skin homeostasis. Development 143, 1674-1687.
72. Uluckan, O., Guinea-Viniegra, J., Jimenez, M., and Wagner, E.F. (2015). Signalling in inflammatory skin disease by AP-1 (Fos/Jun). Clin Exp Rheumatol 33, S44-49.
73. Segre, J.A., Bauer, C., and Fuchs, E. (1999). Klf4 is a transcription factor required for establishing the barrier function of the skin. Nature genetics 22, 356-360.
74. Leask, A., Byrne, C., and Fuchs, E. (1991). Transcription factor AP2 and its role in epidermal-specific gene expression. Proc Natl Acad Sci U S A 88, 7948-7952.
75. Yang, H., Lu, M.M., Zhang, L., Whitsett, J.A., and Morrisey, E.E. (2002). GATA6 regulates differentiation of distal lung epithelium. Development 129, 2233-2246.

76. Sato, A., Xu, Y., Whitsett, J.A., and Ikegami, M. (2012). CCAAT/enhancer binding protein-alpha regulates the protease/antiprotease balance required for bronchiolar epithelium regeneration. Am J Respir Cell Mol Biol 47, 454-463.
77. Chin, S.S., Romano, R.A., Nagarajan, P., Sinha, S., and Garrett-Sinha, L.A. (2013). Aberrant epidermal differentiation and disrupted DeltaNp63/Notch regulatory axis in Ets1 transgenic mice. Biol Open 2, 1336-1345.
78. Nagarajan, P., Chin, S.S., Wang, D., Liu, S., Sinha, S., and Garrett-Sinha, L.A. (2010). Ets1 blocks terminal differentiation of keratinocytes and induces expression of matrix metalloproteases and innate immune mediators. Journal of cell science 123, 3566-3575.
79. Ingraham, C.R., Kinoshita, A., Kondo, S., Yang, B., Sajan, S., Trout, K.J., Malik, M.I., Dunnwald, M., Goudy, S.L., Lovett, M., et al. (2006). Abnormal skin, limb and craniofacial morphogenesis in mice deficient for interferon regulatory factor 6 (Irf6). Nat Genet 38, 1335-1340.
80. Porazinski, S., Wang, H., Asaoka, Y., Behrndt, M., Miyamoto, T., Morita, H., Hata, S., Sasaki, T., Krens, S.F.G., Osada, Y., et al. (2015). YAP is essential for tissue tension to ensure vertebrate 3D body shape. Nature 521, 217-221.
81. Li, E.B., Truong, D., Hallett, S.A., Mukherjee, K., Schutte, B.C., and Liao, E.C. (2017). Rapid functional analysis of computationally complex rare human IRF6 gene variants using a novel zebrafish model. PLoS genetics 13, e1007009.
82. Davidson, E.H. (2009). Network design principles from the sea urchin embryo. Curr Opin Genet Dev 19, 535-540.
83. Liu, Q., Gan, M., and Jiang, R. (2017). A sequence-based method to predict the impact of regulatory variants using random forest. BMC Syst Biol 11, 7.
84. Zafra, M.P., Schatoff, E.M., Katti, A., Foronda, M., Breinig, M., Schweitzer, A.Y., Simon, A., Han, T., Goswami, S., Montgomery, E., et al. (2018). Optimized base editors enable efficient editing in cells, organoids and mice. Nature biotechnology 36, 888-893.
85. Gilchrist, E.P., Moyer, M.P., Shillitoe, E.J., Clare, N., and Murrah, V.A. (2000). Establishment of a human polyclonal oral epithelial cell line. Oral Surg Oral Med Oral Pathol Oral Radiol Endod 90, 340-347.
86. Sdek, P., Zhang, Z.Y., Cao, J., Pan, H.Y., Chen, W.T., and Zheng, J.W. (2006). Alteration of cell-cycle regulatory proteins in human oral epithelial cells immortalized by HPV16 E6 and E7. Int J Oral Maxillofac Surg 35, 653-657.
87. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods 9, 357-359.
88. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114-2120.
89. Ross-Innes, C.S., Stark, R., Teschendorff, A.E., Holmes, K.A., Ali, H.R., Dunning, M.J., Brown, G.D., Gojis, O., Ellis, I.O., Green, A.R., et al. (2012). Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature 481, 389-393.
90. Ye, T., Krebs, A.R., Choukrallah, M.A., Keime, C., Plewniak, F., Davidson, I., and Tora, L. (2011). seqMINER: an integrated ChIP-seq data interpretation platform. Nucleic Acids Res 39, e35.
91. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38, 576-589.
92. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841-842.
93. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.
94. 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 biology 9, R137.
95. Ramirez, F., Ryan, D.P., Gruning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dundar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res 44, W160-165.

96. Westerfield, M. (1993). The Zebrafish Book.(Eugene, OR: University of Oregon Press).
97. Tasic, B., Hippenmeyer, S., Wang, C., Gamboa, M., Zong, H., Chen-Tsai, Y., and Luo, L. (2011). Site-specific integrase-mediated transgenesis in mice via pronuclear injection. Proc Natl Acad Sci U S A 108, 7902-7907.
98. Zhang, Y.D., Dong, S.Y., and Huang, H.Z. (2017). Inhibition of periderm removal in all-trans retinoic acid-induced cleft palate in mice. Exp Ther Med 14, 3393-3398.
99. Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nature biotechnology 34, 525-527.
100. Pimentel, H., Bray, N.L., Puente, S., Melsted, P., and Pachter, L. (2017). Differential analysis of RNA-seq incorporating quantification uncertainty. Nature methods.
101. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102, 15545-15550.
102. Tripathi, S., Pohl, M.O., Zhou, Y., Rodriguez-Frandsen, A., Wang, G., Stein, D.A., Moulton, H.M., DeJesus, P., Che, J., Mulder, L.C., et al. (2015). Meta- and Orthogonal Integration of Influenza "OMICs" Data Defines a Role for UBR4 in Virus Budding. Cell Host Microbe 18, 723-735.
103. Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., et al. (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell 158, 1431-1443.
104. Fornes, O., Castro-Mondragon, J.A., Khan, A., van der Lee, R., Zhang, X., Richmond, P.A., Modi, B.P., Correard, S., Gheorghe, M., Baranasic, D., et al. (2019). JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res.

Figure and Figure legends

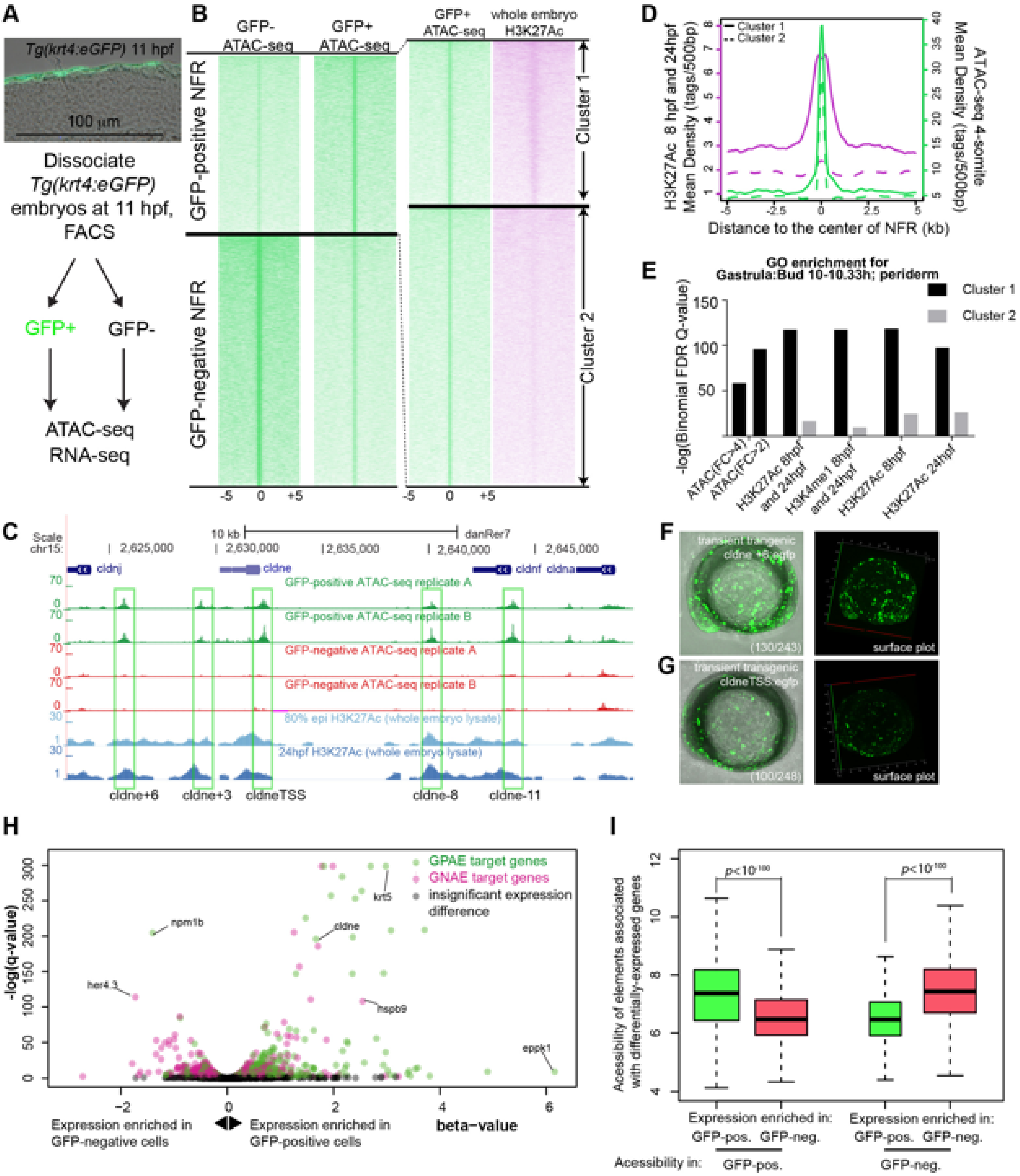


Figure 1 Identification of zebrafish periderm-specific active enhancers by integrating ATAC-seq and H3K27Ac ChIP-seq. A Transverse section of an 11 hpf (4-somite stage) *Tg(krt4:gfp)* embryo, showing GFP is confined to the superficial layer of cells, and workflow of ATAC-seq in periderm and non-periderm cells . **B** Density plots of ATAC-seq results. Each line is centered on a nucleosome free region (NFR) with significantly more ATAC-seq reads in GFP-positive or GFP-negative cells; the majority of ATAC-seq peaks were not enriched in either cell type. Density plots also show H3K27Ac ChIP-seq signal in whole embryos at 8 hpf and/or at 24 hpf (data from [34]) at each

of the GFP-positive NFRs; the latter are sorted in to those that overlap (or are flanked by within 100-1500bp) peaks of H3K27Ac signal (cluster 1, 4301 elements) and or not (cluster 2, 7952 elements). **C** UCSC Genome browser tracks showing the ATAC-seq peaks in GFP-positive and GFP-negative cells, and H3K27Ac signal from whole embryos at 8 hpf and at 24 hpf (data from [34]) at the *cldne* locus. Boxes, examples of cluster 1 elements, also known as zebrafish GFP-positive active enhancers (zGPAEs). Elements are *cldne*+6 kb (zv9：chr15:2625460-2625890), *cldne* +3 kb (chr15:2629012-2629544), *cldne* -8 kb (chr15:2639873-2640379), *cldne* -11 kb (chr15:2643578-2644160), and *cldne* TSS (chr15:2631981-2632513). **D** Plot of average density of H3K27Ac ChIP-seq signal (purple) and ATAC-seq signal (green). **E** GO enrichment for term “Gastrula:Bud 10-10.33h; periderm” among NFRs enriched in GFP-positive cells with normalized fold change greater than 2 (ATAC(FC>2)) and 4 (ATAC(FC>4)), NFRs enriched in GFP-positive cells flanked or overlapped by 24hpf and 80% epiboly H3K27Ac ChIP-seq peaks (cluster 1) and depleted with H3K27Ac (cluster 2), NFRs enriched in GFP-positve cells flanked or overlapped by 24hpf and 80% epiboly H3K4me1 ChIP-seq peaks (cluster 1) and depleted with H3K4me1(cluster 2), NFRs enriched in GFP-positive cells flanked or overlapped by 24hpf H3K27Ac ChIP-seq peaks (cluster 1) and depleted with H3K27Ac (cluster 2), and NFRs enriched in GFP-positive cells flanked or overlapped by 80% epiboly H3K27Ac ChIP-seq peaks (cluster 1) and depleted with H3K27Ac (cluster 2). **F**, **G** Lateral views of wild-type embryos at 11 hpf injected at the 1-cell stage with GFP reporter constructs built from **F** *cldne* +6 and **G** *cldne* transcription start site (TSS) elements. Left panels are stack views of the embryo, and right panels are surface plot for the embryos indicating most GFP signals are from the surface (periderm) of the embryos. Number in parentheses is the ratio of embryos with at least 10 GFP-positive periderm cells over injected embryos surviving at 11 hpf. **H** Volcano plot of RNA seq data, showing the expression of genes associated (by GREAT) with zGPAEs (green dots) or with zGNAEs (pink dots) in GFP-positive cells (beta-value>0) or in GFP-negative cells (beta-value <0). **I** Plot of accessibility scores of elements with differential accessibility (i.e., both zGPAEs and zGNAEs) associated with genes that are differentially expressed in GFP-positive and GFP-negative cells, showing that elements with increased accessibility in GFP-positive cells tend to be associated with genes whose expression is enriched in GFP-positive cells, and vice versa.

Figure 1-source data 1 Density plot for ATAC-seq and H3K27Ac ChIP-seq, as plotted in Figure 1D

Figure 1-source data 2 Barchart for GO enrichment, as plotted in Figure 1E

Figure 1-source data 3 Scatter plot for the genes near GPAEs and GNAEs, as plotted in Figure 1H

Figure 1-source data 4 Box plot for the normalized chromatin accessibility of periderm- and non-periderm enriched genes in GFP positive or negative cells, as plotted in Figure 1I

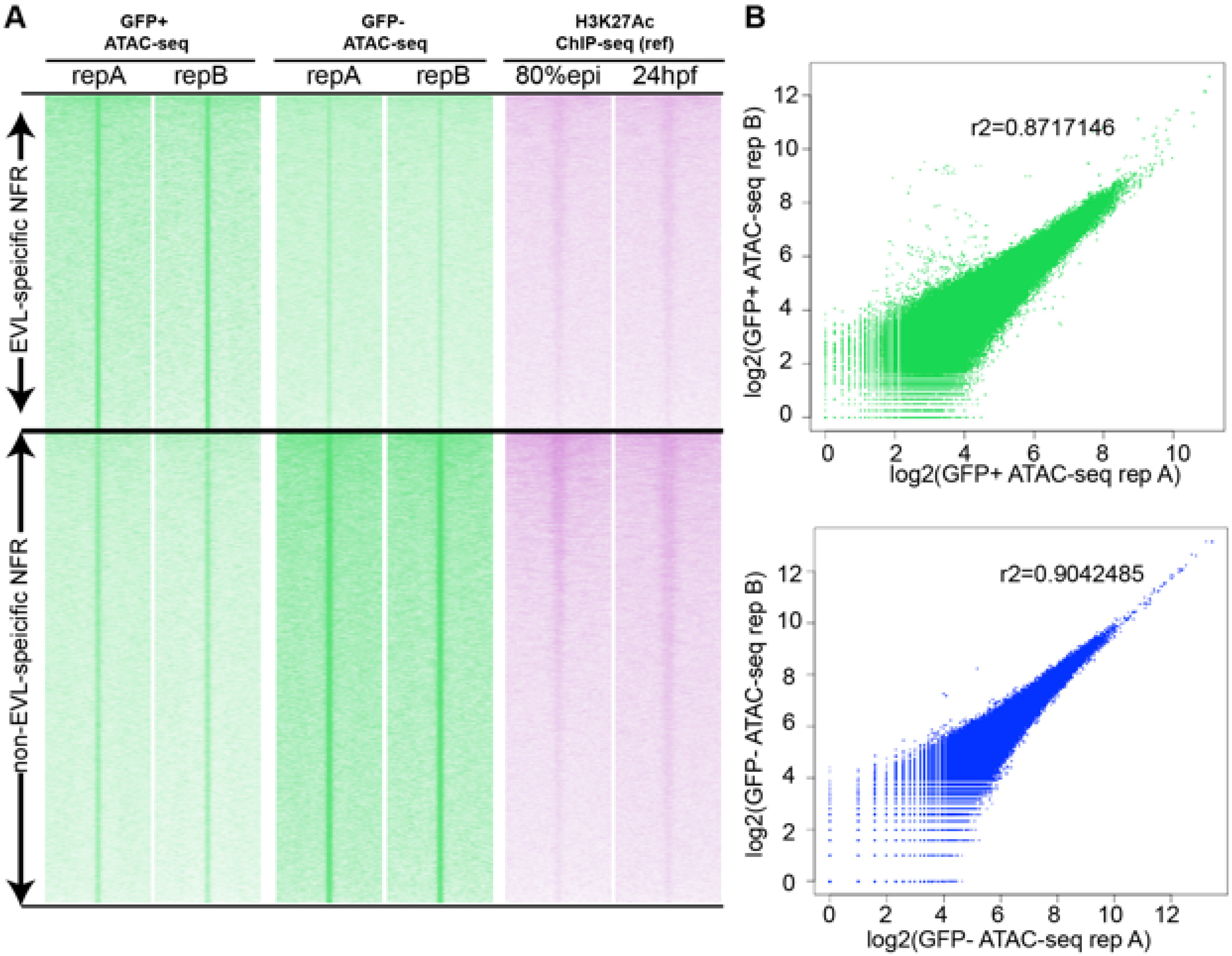


Figure 1—figure supplement 1 Correlation of zebrafish periderm ATAC-seq two biological replicates. **A** ATAC-seq summit centered heatmap of ATAC-seq signals from two biological replicates. **B** Scatter plots showing the ATAC-seq signal correlation between two biological replicates.

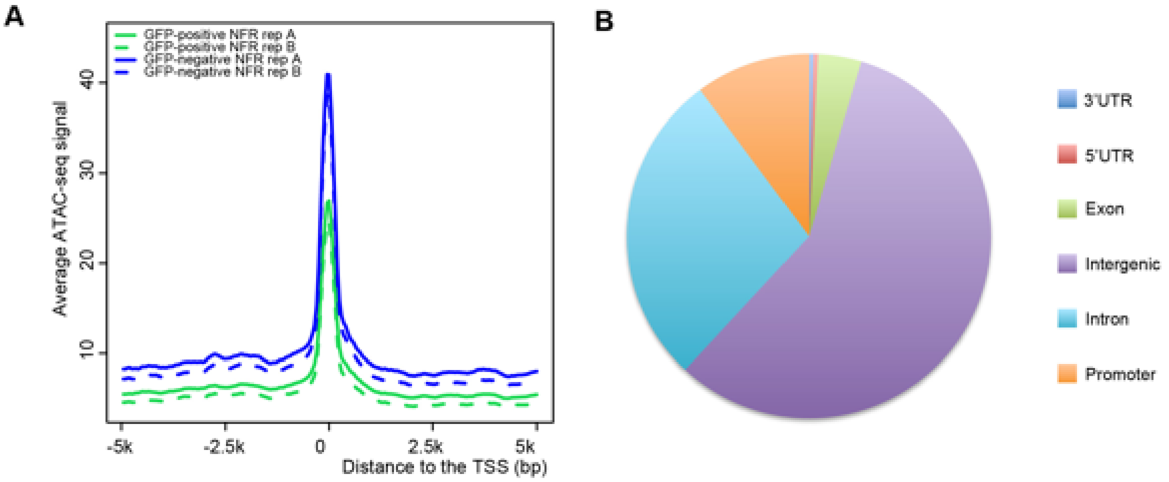


Figure 1—figure supplement 2 Annotation of ATAC-seq peaks relative to transcription start sites. **A** Histogram of read density of ATAC-seq in 10 kb flanking transcription start sites (TSS). **B** Pie chart showing the genomic location of GFP-positive NFRs (from ATAC-seq biological replicate 1).

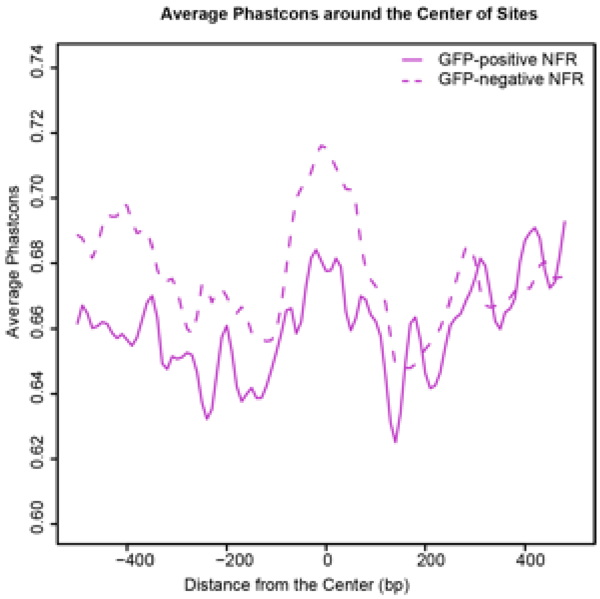


Figure 1—figure supplement 3 Average Vertebrate PhastCons Score (danRer7 genome) at different distances from the center of nucleosome free regions (NFRs) in GFP-positive and GFP-negative (flow through) cells sorted from Tg(krt4:gfp) embryos at 11 hpf.

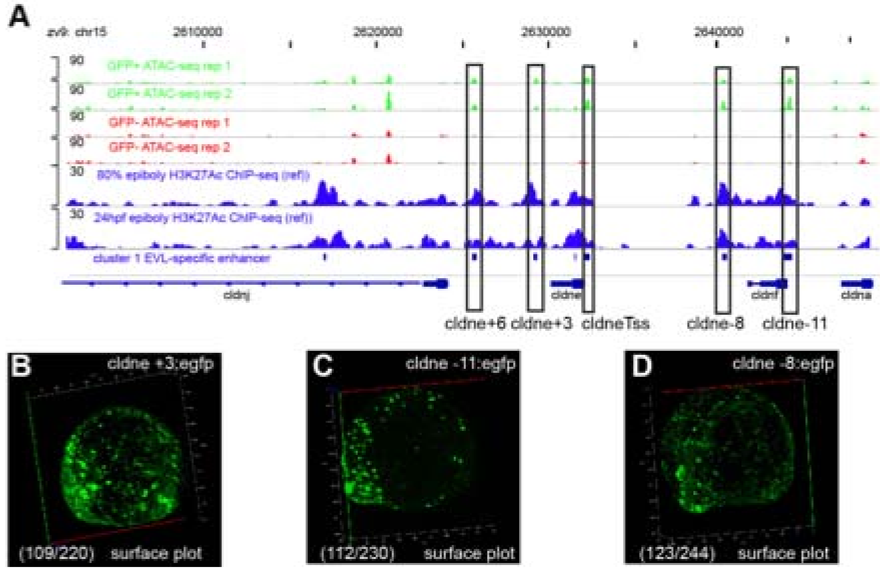


Figure 1—figure supplement 4 Transient reporter assay validation for *cldne* +3, *cldne* -11, and *cldne* -8 elements. Number in parentheses is the ratio of embryos with at least 10 GFP-positive periderm cells over injected embryos surviving at 11 hpf.

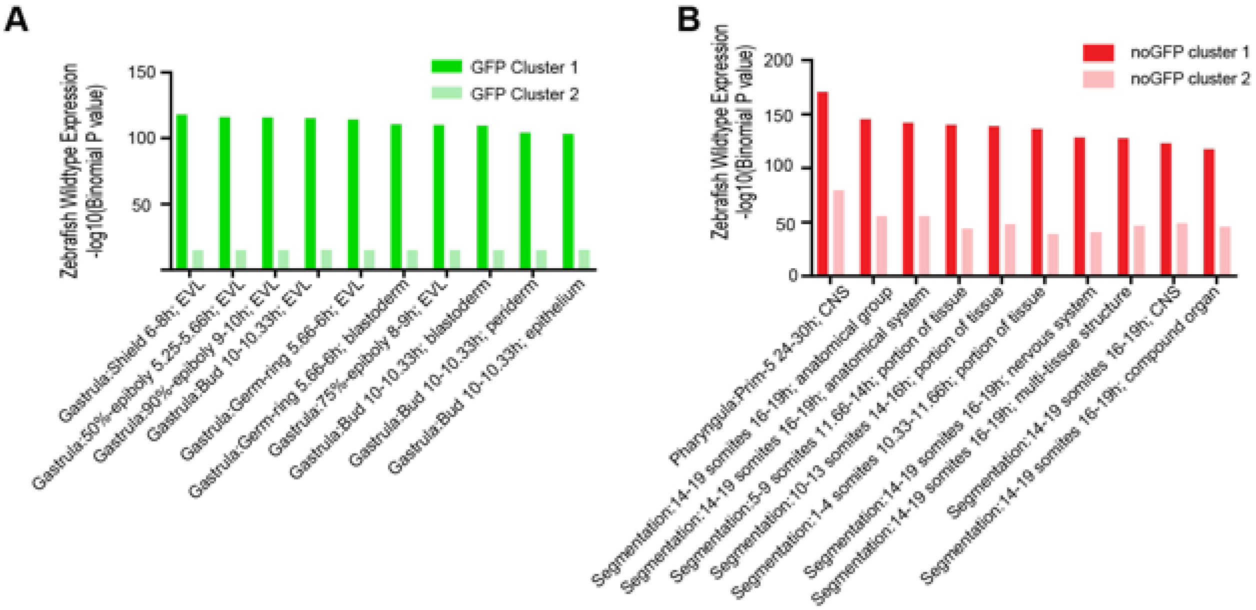


Figure 1—figure supplement 5 GO enrichment analysis for different clusters of GFP-positive or GFP-negative specific NFRs.

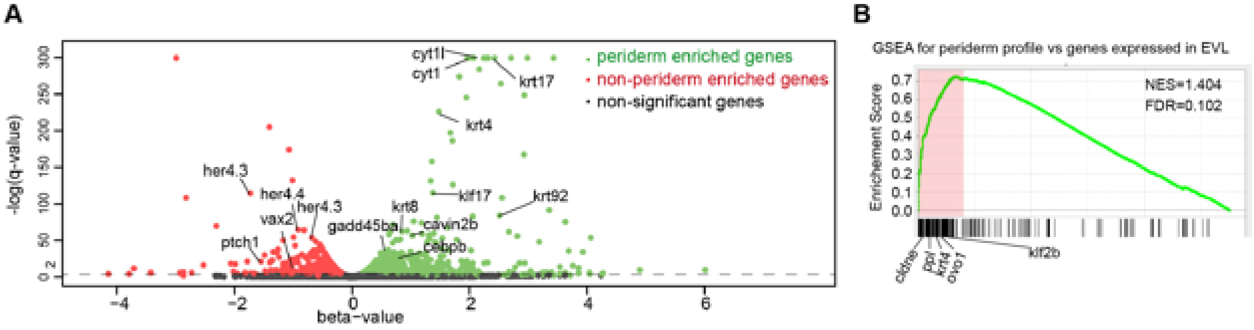


Figure 1—figure supplement 6 Summary for RNA-seq for krt4:GFP-positive and krt4:GFP-negative cells at 4-somite stage. **A)** Volcano plot for genes expressed in GFP-positive (in green) and –negative (in red) cells. **B)** GSEA for genes expressed in GFP-positive cells using EVL gene set (www.zfin.org).

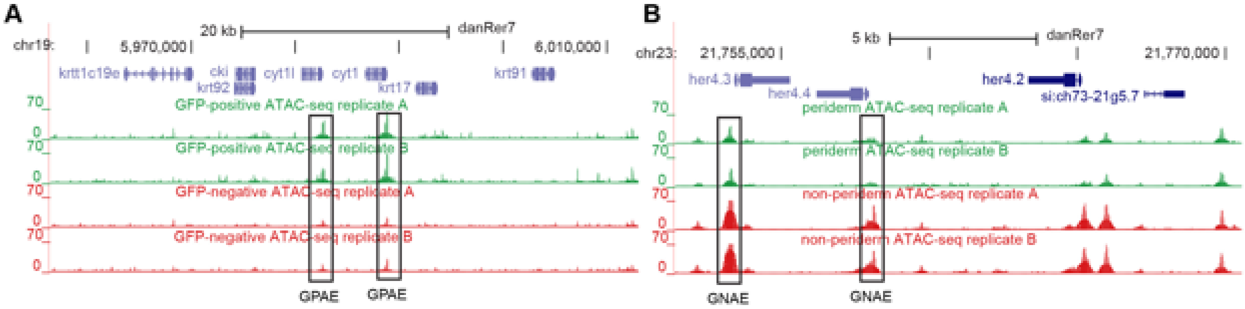
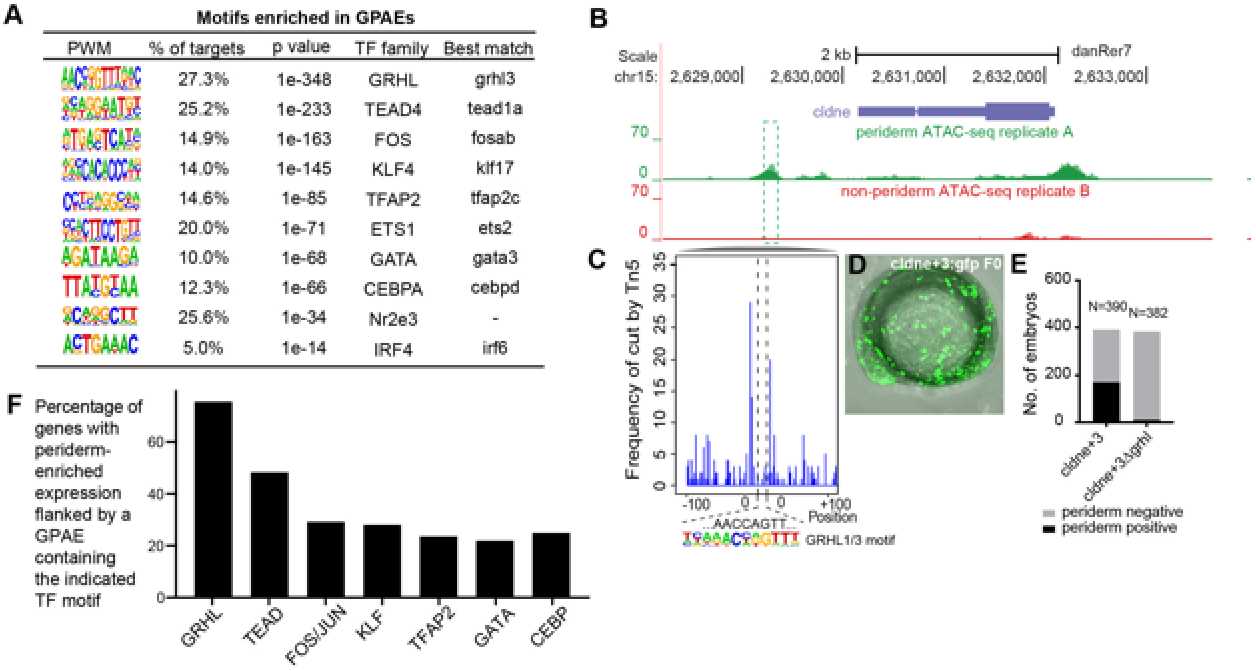


Figure 1—figure supplement 7 ATAC-seq near **A** keratin and **B** *her4* cluster genes

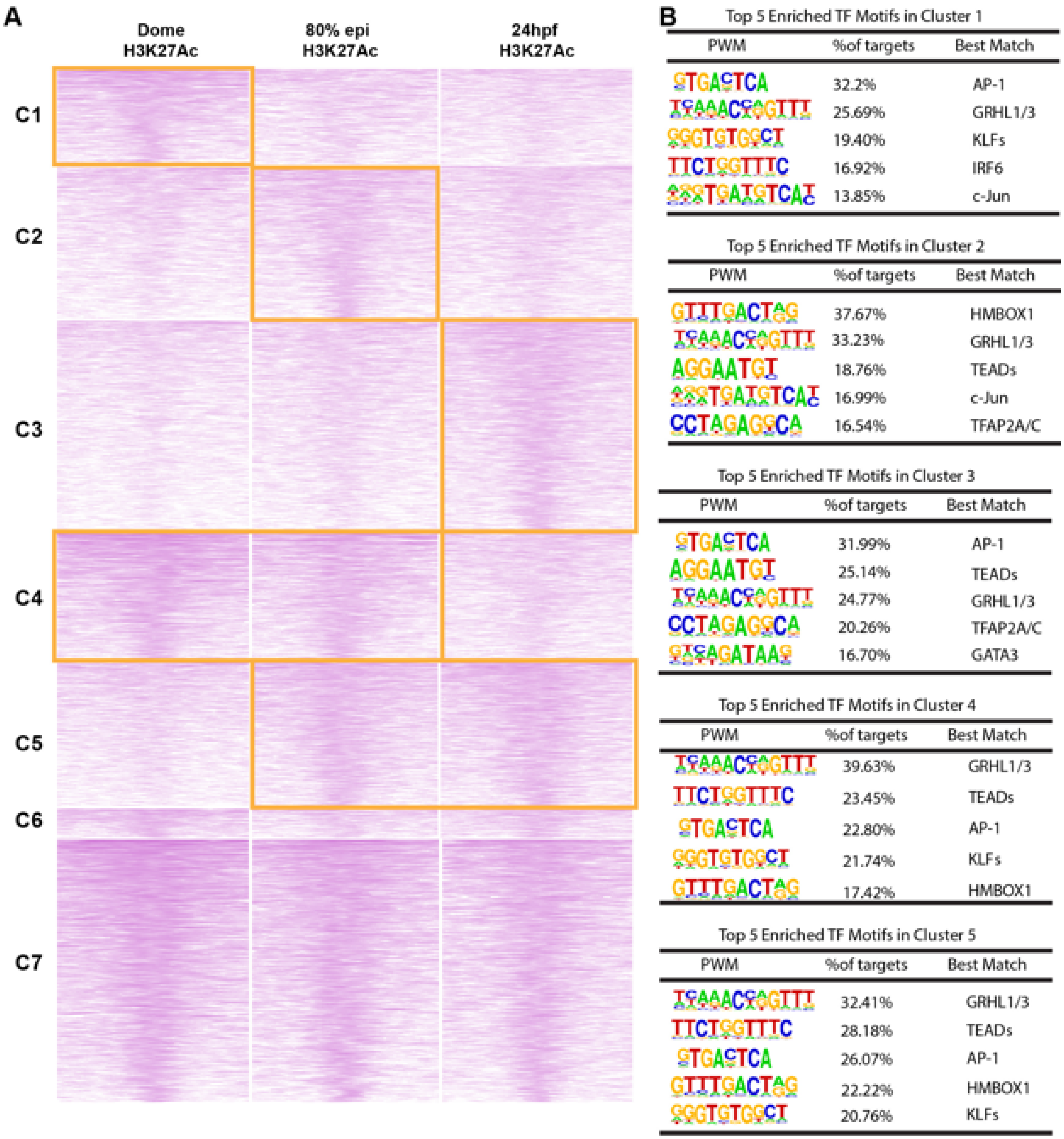


Motifs enriched in GPAEs

PWM	% of targets	p value	TF family	Best match
	27.3%	1e-348	GRHL	grhl3
	25.2%	1e-233	TEAD4	tead1a
	14.9%	1e-163	FOS	fosab
	14.0%	1e-145	KLF4	klf17
	14.6%	1e-85	TFAP2	tfap2c
	20.0%	1e-71	ETS1	ets2
	10.0%	1e-68	GATA	gata3
	12.3%	1e-66	CEBPA	cebpd
	25.6%	1e-34	Nr2e3	-
	5.0%	1e-14	IRF4	irf6

Figure 2.

Features of zGPAEs. A Enriched motifs in zGPAEs. PWM, position weighted matrix. TF, transcription factors. Best match, transcription factor in the indicated family with highest expression in GFP-positive cells, i.e., whether or not the family member is enriched in GFP-positive cells. **B** Genome browser view showing a GFP-positive nucleosome free region (NFR) about 3 kb downstream of the transcription start site of *cldne* gene. **C** Schematic of frequency of Tn5 cleavage sites at within this NFR, indicating reduced frequency of cleavage at a motif matching the GRHL binding site relative to in flanking DNA. **D** Confocal image of a wild-type embryo at 10 hpf (2-somite stage) injected at the one-cell stage with a reporter construct containing this NFR. **E** Bar chart showing number of embryos positive for GFP signal in the periderm after being injected with the intact reporter or one in which the GRHL motif was deleted. **F** Bar chart showing the percentage of genes whose expression is higher in GFP-positive cells than in GFP-negative cells that are flanked by a zGPAE possessing the indicated binding site.



Top 5 Enriched TF Motifs in Cluster 1

PWM	%of targets	Best Match
GTGACTCA	32.2%	AP-1
TCAAACCAGTTT	25.69%	GRHL1/3
GGGTGTGGCT	19.40%	KLFs
TTCTGGTTTC	16.92%	IRF6
AGGTGATGTCAT	13.85%	c-Jun

Top 5 Enriched TF Motifs in Cluster 2

PWM	%of targets	Best Match
GTTTGACTAG	37.67%	HMBOX1
TCAAACCAGTTT	33.23%	GRHL1/3
AGGAATGT	18.76%	TEADs
AGGTGATGTCAT	16.99%	c-Jun
CCTAGAGGCA	16.54%	TFAP2A/C

Top 5 Enriched TF Motifs in Cluster 3

PWM	%of targets	Best Match
GTGACTCA	31.99%	AP-1
AGGAATGT	25.14%	TEADs
TCAAACCAGTTT	24.77%	GRHL1/3
CCTAGAGGCA	20.26%	TFAP2A/C
GTCAGATAAG	16.70%	GATA3

Top 5 Enriched TF Motifs in Cluster 4

PWM	%of targets	Best Match
TCAAACCAGTTT	39.63%	GRHL1/3
TTCTGGTTTC	23.45%	TEADs
GTGACTCA	22.80%	AP-1
GGGTGTGGCT	21.74%	KLFs
GTTTGACTAG	17.42%	HMBOX1

Top 5 Enriched TF Motifs in Cluster 5

PWM	%of targets	Best Match
TCAAACCAGTTT	32.41%	GRHL1/3
TTCTGGTTTC	28.18%	TEADs
GTGACTCA	26.07%	AP-1
GTTTGACTAG	22.22%	HMBOX1
GGGTGTGGCT	20.76%	KLFs

Figure 2—figure supplement 1 Different clusters of H3K27Ac ChIP-seq at different developmental stages in zGPAEs. **A** ATAC-seq summit centered heatmap of H3K27Ac ChIP-seq at 4.5 hpf, 8hpf and 24hpf data from [34], cluster performed by *k*-mean. **B** Motif enriched in zGPAEs with high H3K27Ac at 4.5hpf. **C** Motifs enriched in zGPAEs with high H3K27Ac at 24 hpf.

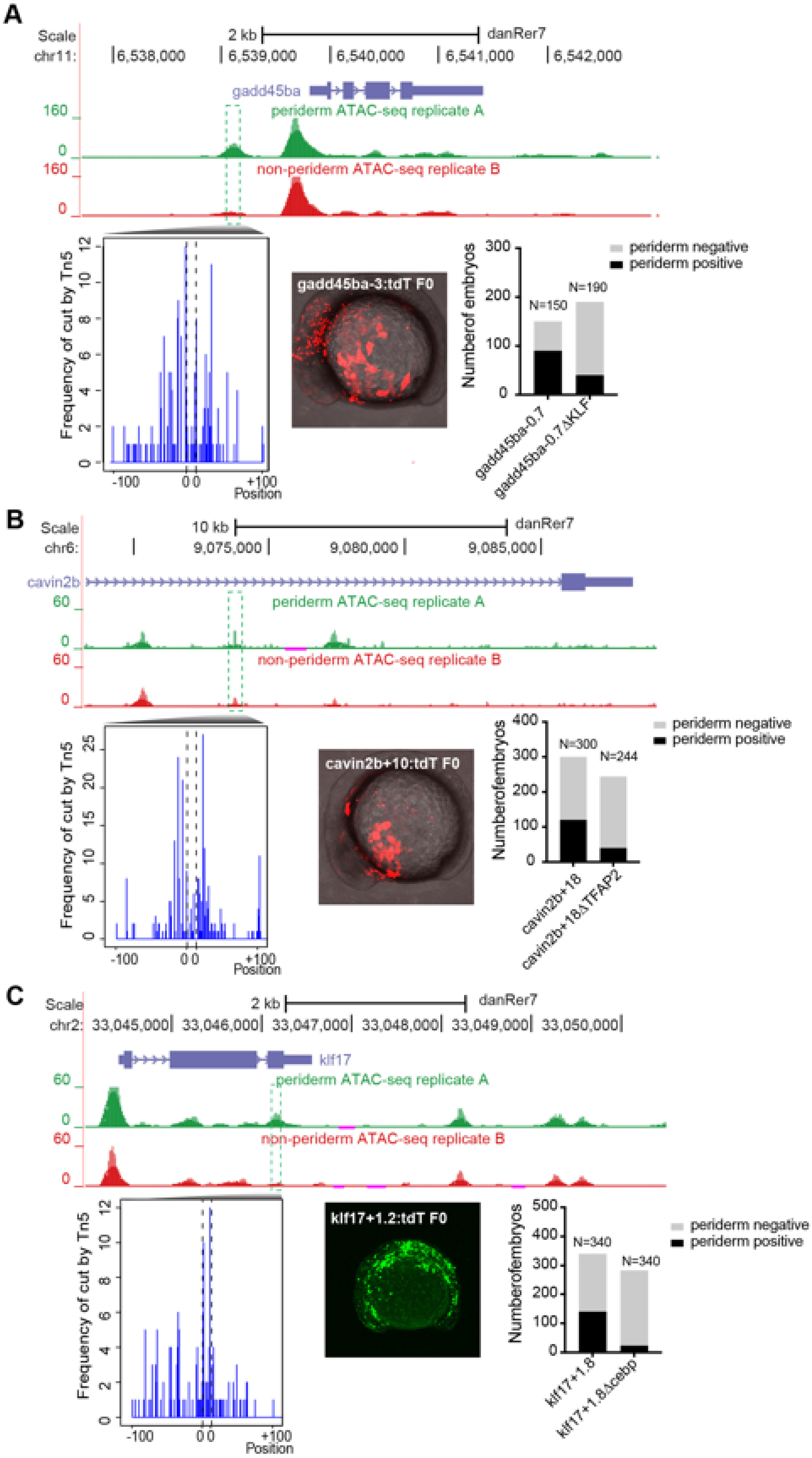


Figure 2—figure supplement 2 Transient reporter assay of **A** gadd45ba-3 with or without KLF motif, **B** cavin2b-+18 with or without TFAP2 motif and **C** klf17-+1.2 with or without C/EBP motif.

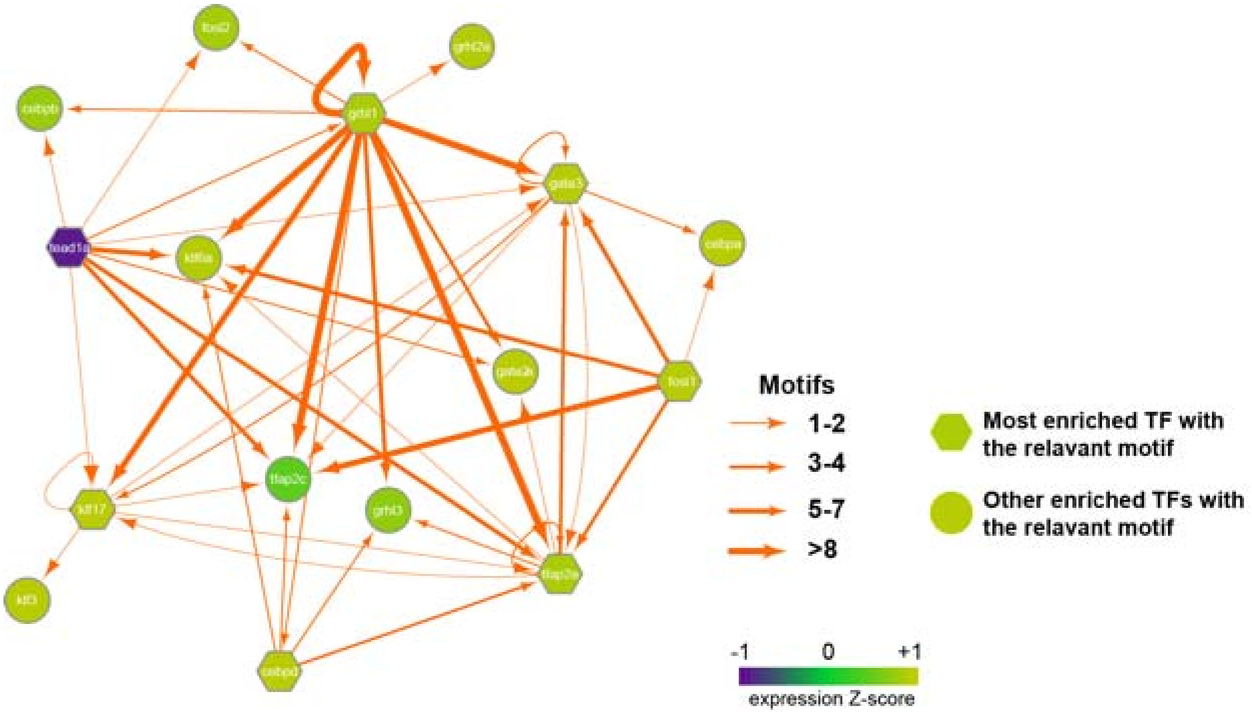


Figure 2—figure supplement 3 Putative regulatory interactions of major periderm-enriched transcription factors governing transcriptomic state in periderm cells at 4-somite stage. Depending on the expression level in periderm cells (GFP-positive cells) most enriched transcription factors with the relevant motifs are in hexagon while other enriched transcription factors with the relevant motifs are in round. Each TF node is colored according to the normalized expression z-score (related to periderm genes). The thickness of each edge represents the number of motifs located in the all nearby enhancers to each TF (within 100kbp to the transcription start site).

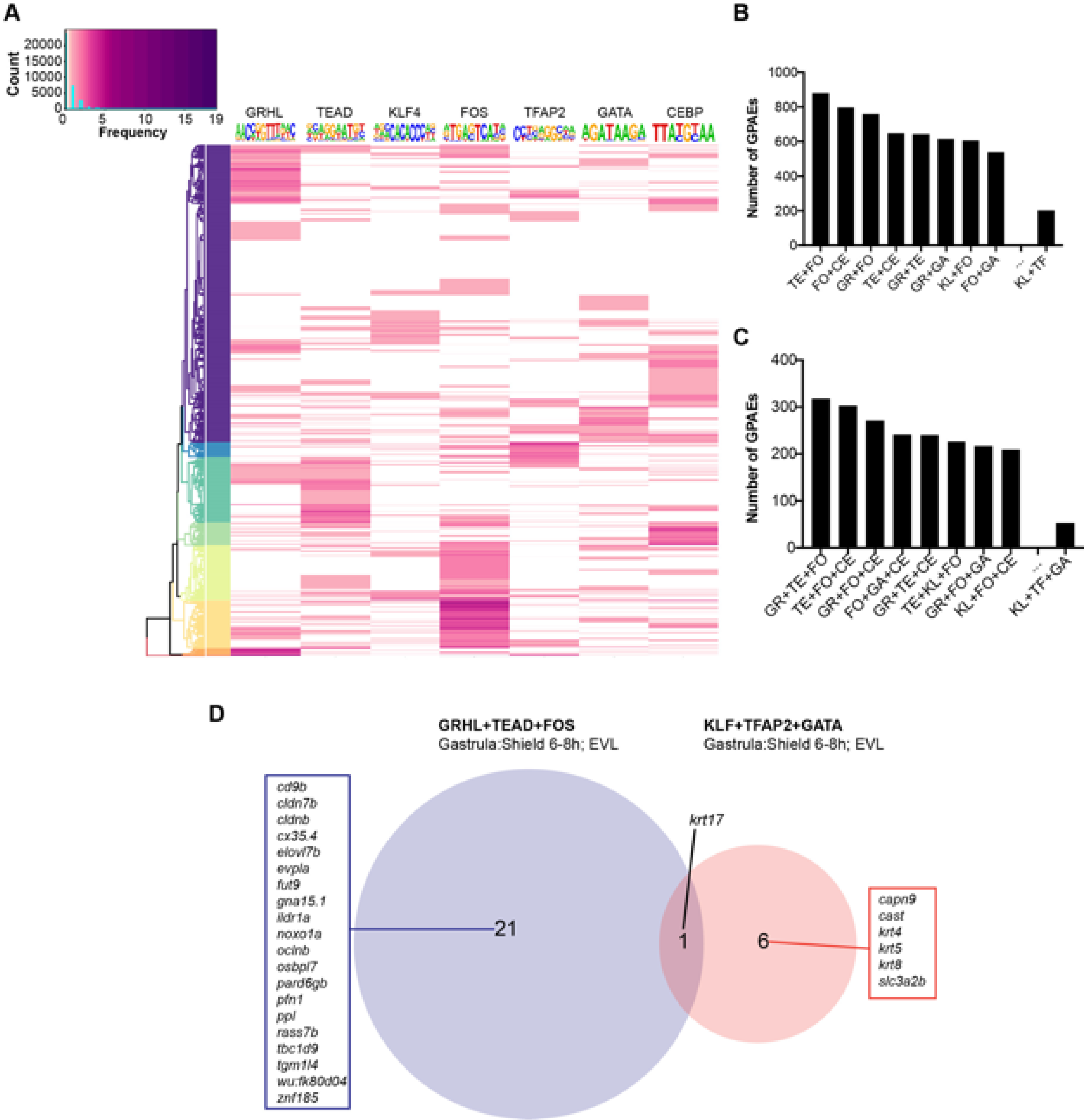


Figure 2—figure supplement 4 Motif combination in GPAEs **A)** Hierarchy clustering for the number of enriched motifs in all GPAEs. “Count” in the color key indicates the sum for different number of each motif “frequency”. **B)** Bar chart for the number of GPAEs with different two-motif combination. **C)** Bar chart for the number of GPAEs with different three-motif combination. **D)** Nearest EVL genes (within 100.0 kbp) of the GPAEs with “GRHL+TEAD+FOS” and “KLF+TFAP2+GATA” combination. GR: GRHL, TE: TEAD, FO: FOS, TF: TFAP2, GA: GATA, CE: CEBP, KL: KLF

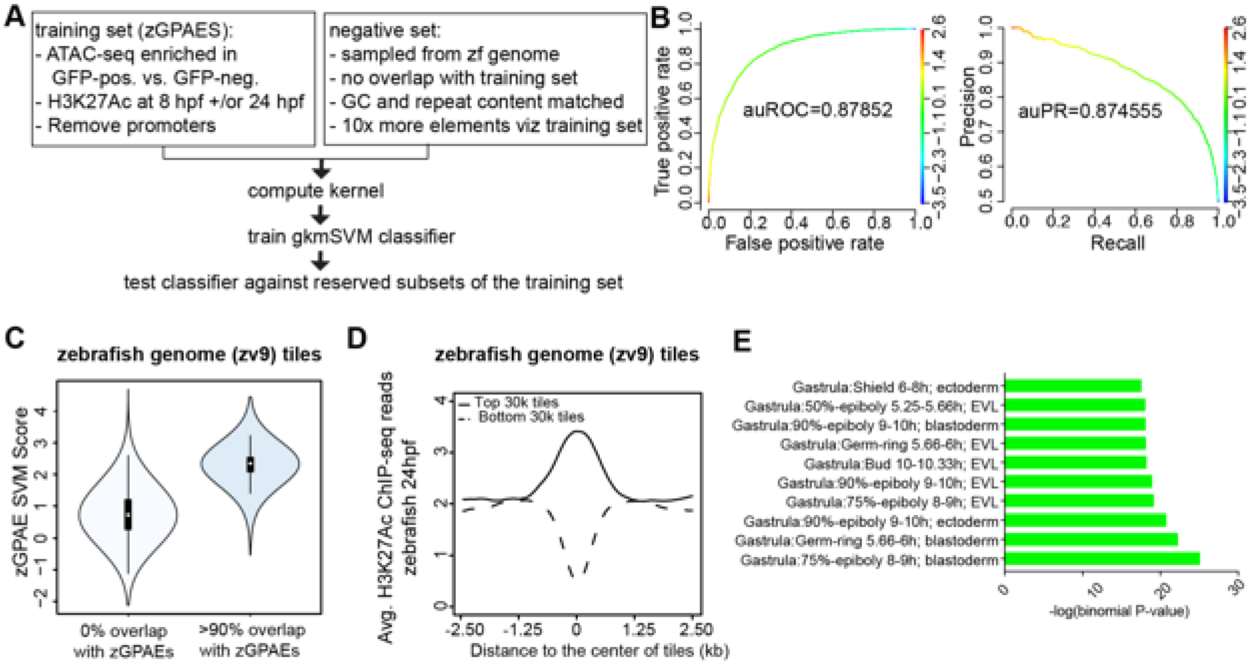


Figure 3 Training a gapped kmer support vector machine (gkmSVM) classifier trained on zGPAEs **A** Pipeline for training and cross-validation of gkmSVM classifier on zebrafish periderm enhancer candidates. **B** Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves using the gkmSVM trained on zGPAEs. au, area under. Color of curves corresponds to SVM scores. **C** Violin plots showing SVM scores of zebrafish genome tiles with 0% or at least 90% overlapped with the training set (GPAEs). **D** Average H3K27Ac ChIP-seq reads at the 30,000 elements with the highest or lowest scores from the gkmSVM trained on zGPAEs. **E** GO enrichment assay for genes associated with the top-scoring tiles 10,000 tiles, i.e., including those that overlap by the training set.

Figure 3-source data 1 Density plot for H3K27Ac ChIP-seq reads, as plotted in Figure 3D

Figure 3-source data 2 Barchart for GO enrichment assay, as plotted in Figure 3E

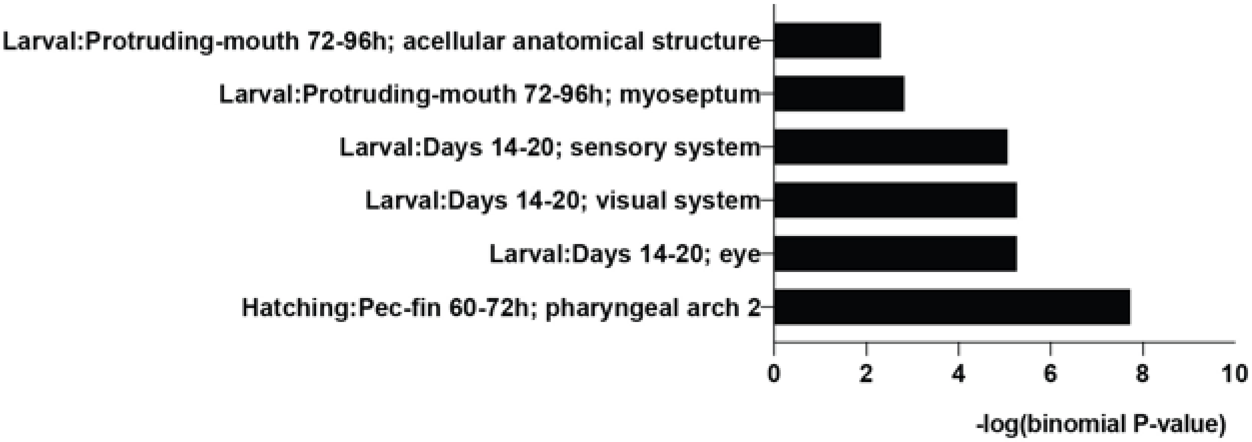


Figure 3—figure supplement 1 GO enrichment assay of gene expression for the top-scoring 10 K tiles that do not overlap zGPAEs

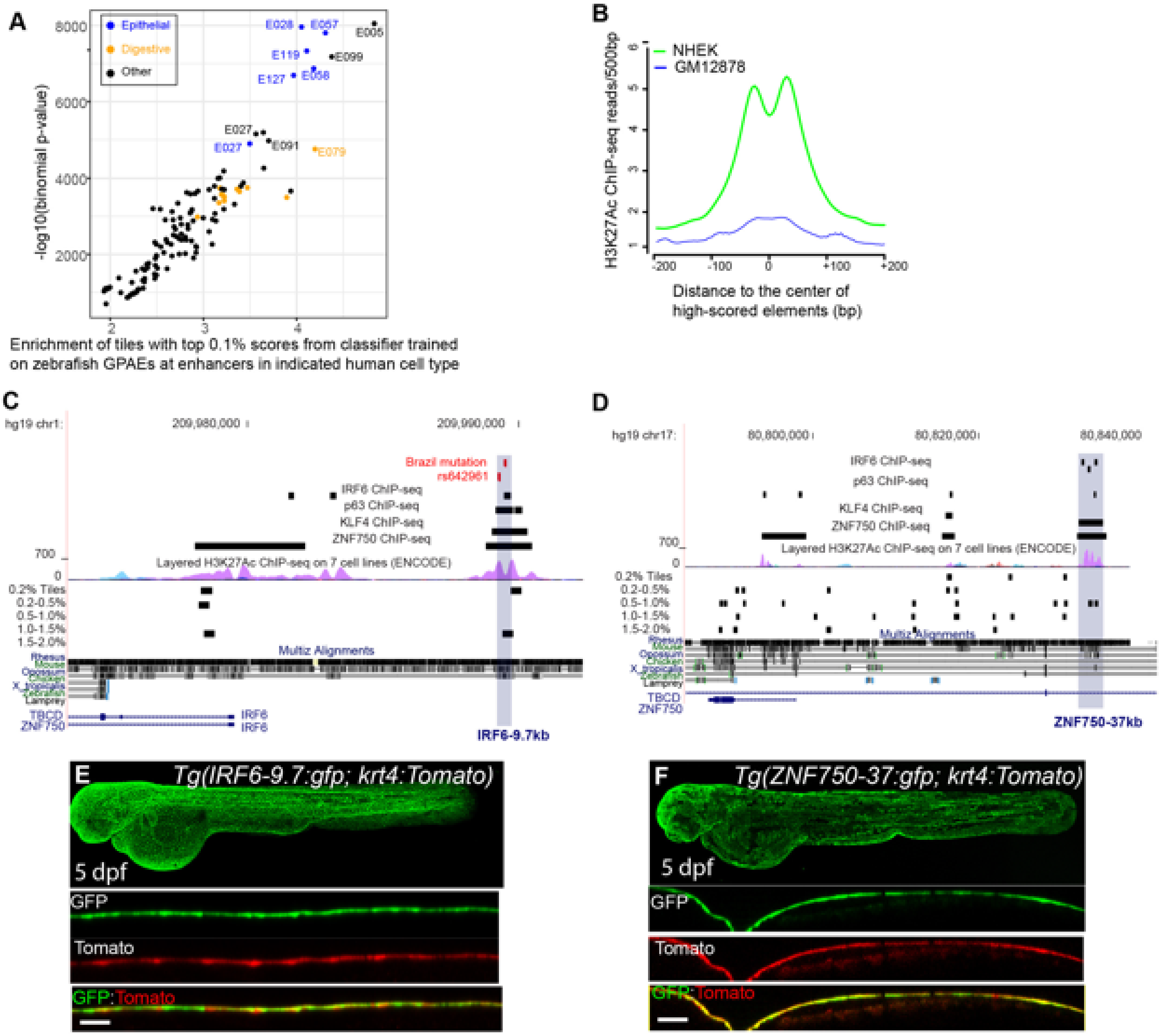


Figure 4 A classifier trained on zGPAEs applied to the human genome. **A** Enrichment of top-scoring 0.1% bin of human genome (hg19) tiles by fish classifier trained on zGPAEs at active enhancers in the indicated cell type, defined by ChIP-seq to chromatin marks by the Roadmap project [42]. Blue dots, epithelial tissues; orange dots, digestive tissues. [E05: H1 BMP4 Derived Trophoblast Cultured Cells; E027: Breast Myoepithelial Primary Cells; E028: Breast variant Human Mammary Epithelial Cells; E057, E058: Foreskin Keratinocyte Primary Cells; E079: Esophagus; E091: Placenta; E099: Placenta amnion; E119, Mammary Epithelial Primary Cells (HMEC); E127:NHEK-Epidermal Keratinocyte Primary Cells]. **B** Average density of H3K27Ac ChIP-seq signal in NHEK and GM12878 cells [42] at top 0.1% tiles predicted using zGPAEs. **C** Genome browser view focused on IRF6-9.7, also known as multispecies conserved sequence MCS9.7 (hg19 chr1:209989050-209989824). A SNP within it rs642961 (chr1: 209989270) is associated with risk for non-syndromic orofacial cleft. Brazil mutation refer to rare mutation (350dupA) reported previously [45]. Multiz Alignments of 100 vertebrate species revealed this high-score element within 1.0-1.5% tiles are not conserved in zebrafish but overlapped with IRF6 ChIP-seq, TP63 ChIP-seq and KLF4 ChIP-seq in normal human keratinocytes (reviewed in [53]). **D** Genome browser view focused on ZNF750-37 (hg19 chr17:80832267-80835105). This element, though not conserved in zebrafish, overlapped with IRF6 ChIP-seq, TP63 ChIP-seq and KLF4 ChIP-seq in normal human keratinocytes. **E** GFP expression pattern of *Tg(IRF6-9.7:gfp; krt4:Tomato)* at 5 dpf. **F** GFP expression pattern of *Tg(ZNF750-37:gfp; krt4:Tomato)* at 5 dpf.

Figure 4-source data 1Scatter plot for the enrichment of top scoring human genome tiles, as plotted in Figure 4A

Figure 4-source data 2 Density plot for H3K27Ac ChIP-seq in NHEK and GM12878 cells within the top scoring human genome ties, as plotted in Figure 4B

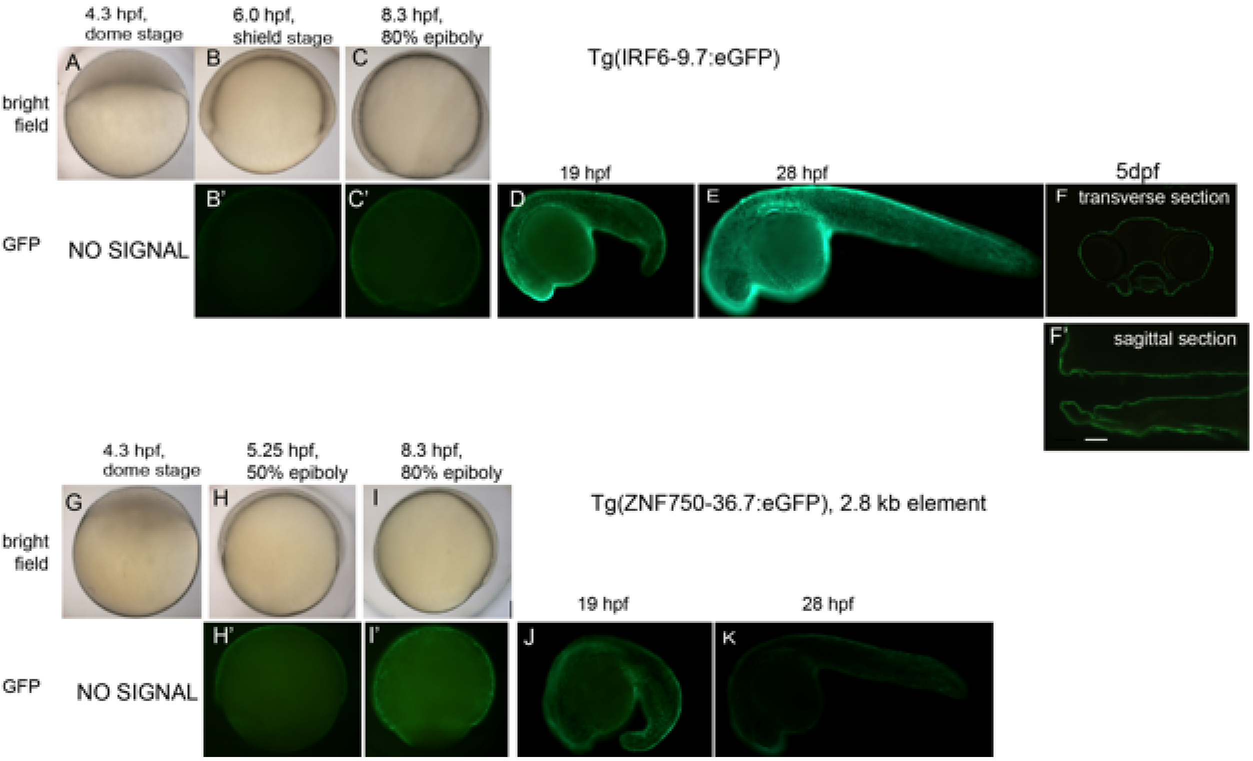
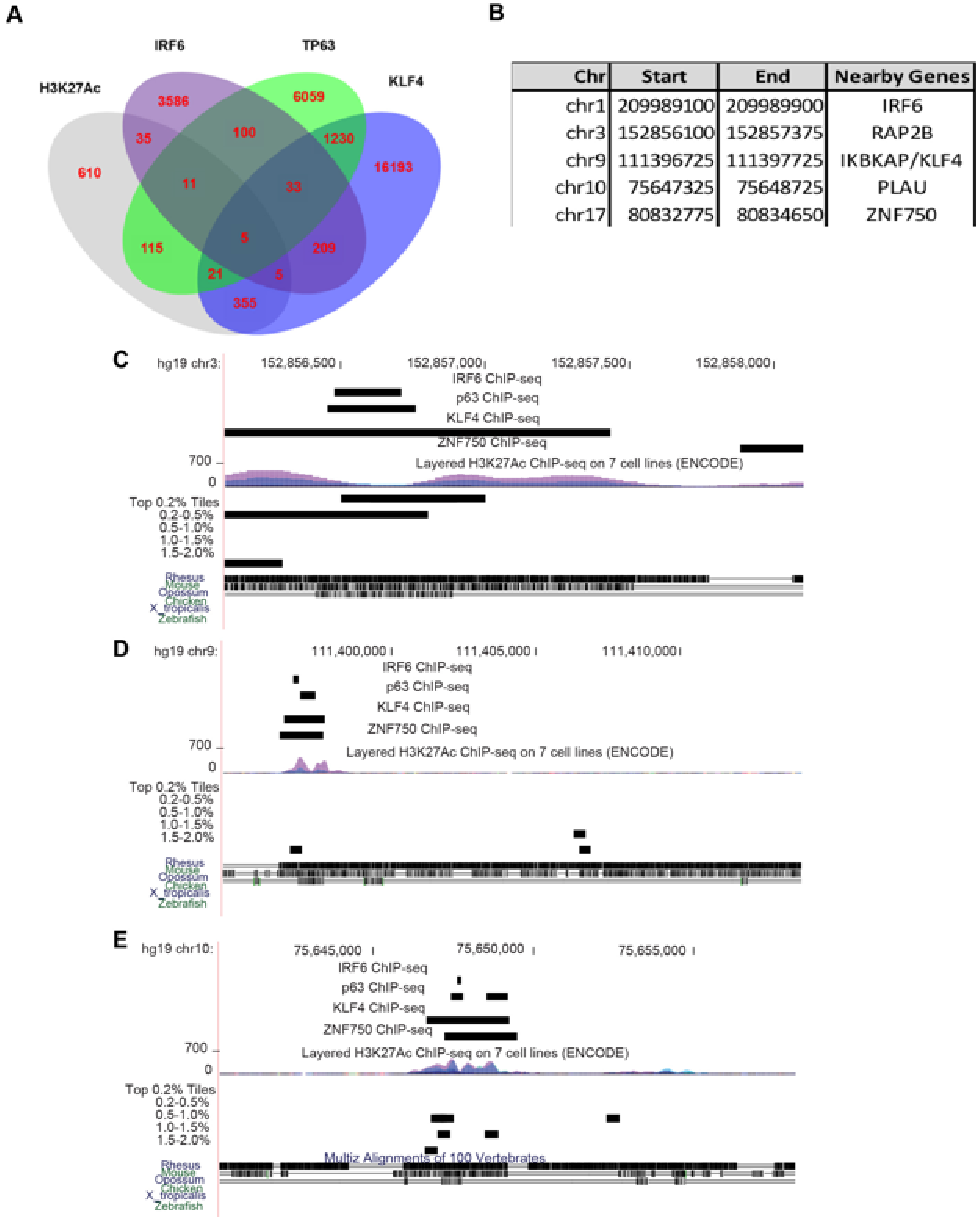


Figure 4—figure supplement 1 Detailed description of enhancer activity pattern of *Tg(IRF6-9.7:gfp)* and *Tg(ZNF750-36.7:gfp)*.



Chr	Start	End	Nearby Genes
chr1	209989100	209989900	IRF6
chr3	152856100	152857375	RAP2B
chr9	111396725	111397725	IKBKAP/KLF4
chr10	75647325	75648725	PLAU
chr17	80832775	80834650	ZNF750

Figure 4—figure supplement 2 Browser views of all loci with mcs9.7 ChIP-seq features. **A** Intersection of TP63, IRF6, KLF4 and H3K27Ac ChIP-seq peaks in human NHEK cells. **B** Coordinates for five genomic regions with overlapped TP63, IRF6, KLF4 and H3K27Ac ChIP-seq peaks. **C-E** Genome browser view for regions sharing this feature near *RAP2B*, *KLF4*, and *PLAU*.

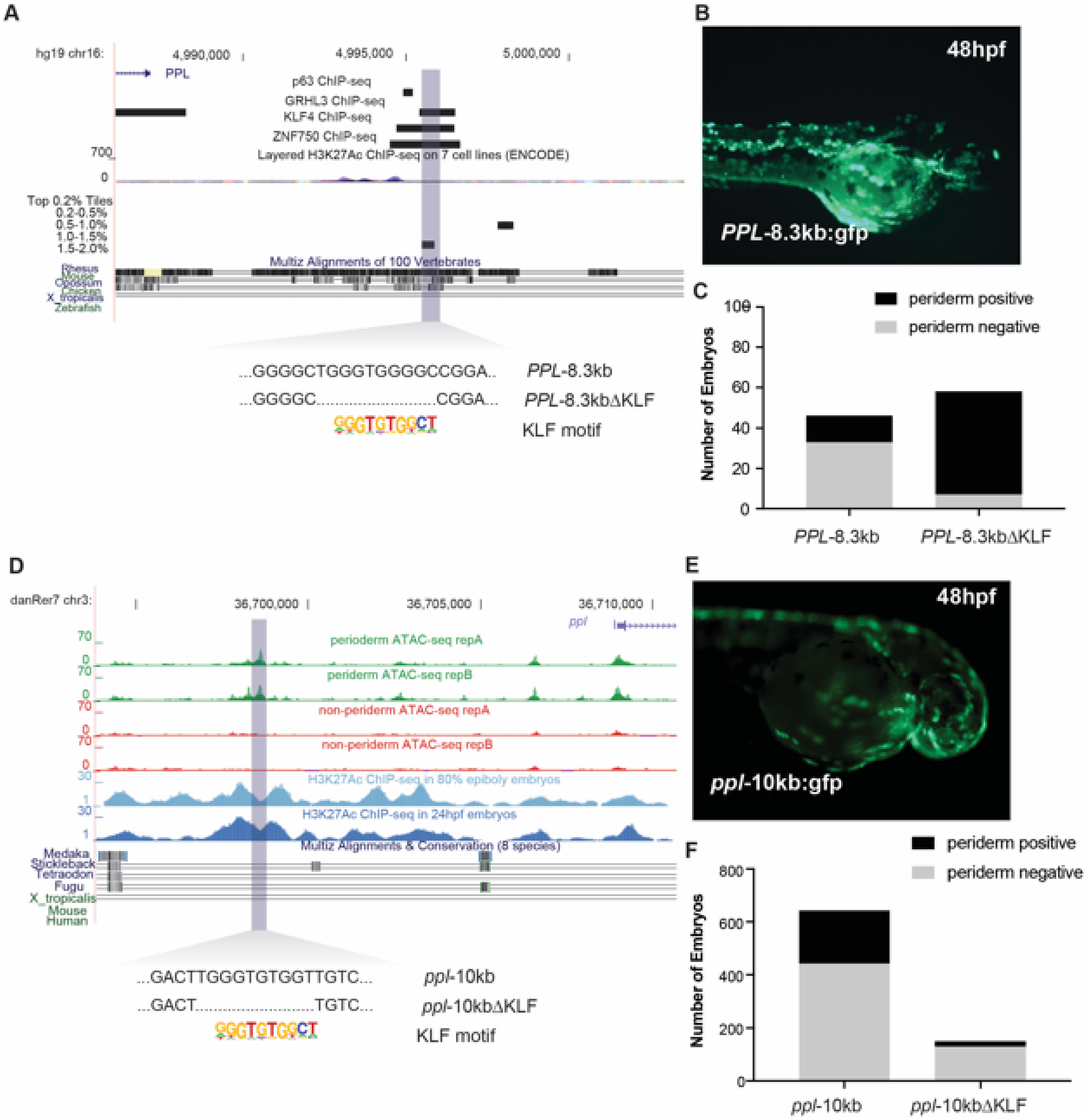
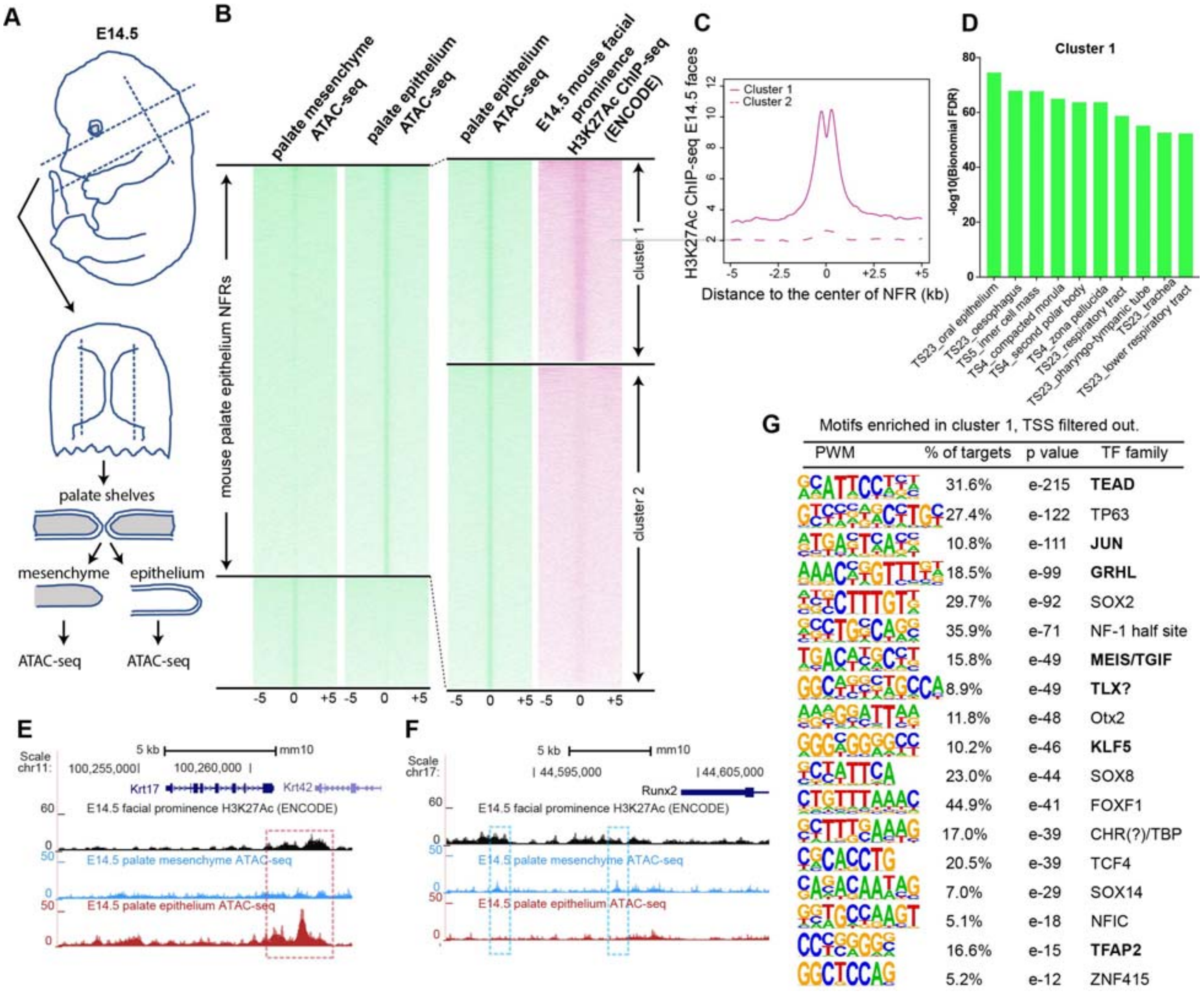


Figure 4—figure supplement 3 Reporter assay for human and zebrafish *PPL* elements predicted by zebrafish classifier. **A** Genome browser view for *ppl*-10kb **B** Transient reporter for *ppl*-10kb:gfp **C** Genome browser view for PPL-8.3kb. **D** Transient reporter for *PPL*-8.3kb:gfp.



Motifs enriched in cluster 1, TSS filtered out.

PWM	% of targets	p value	TF family
	31.6%	e-215	**TEAD**
	27.4%	e-122	TP63
	10.8%	e-111	**JUN**
	18.5%	e-99	**GRHL**
	29.7%	e-92	SOX2
	35.9%	e-71	NF-1 half site
	15.8%	e-49	**MEIS/TGIF**
	8.9%	e-49	**TLX?**
	11.8%	e-48	Otx2
	10.2%	e-46	**KLF5**
	23.0%	e-44	SOX8
	44.9%	e-41	FOXF1
	17.0%	e-39	CHR(?)/TBP
	20.5%	e-39	TCF4
	7.0%	e-29	SOX14
	5.1%	e-18	NFIC
	16.6%	e-15	**TFAP2**
	5.2%	e-12	ZNF415

Figure 5 Identification of mouse embryonic palatal epithelium-specific active enhancers. A Workflow of ATAC-seq in palatal epithelium and non-epithelium cells isolated at E14.5 embryos palate. **B** Heatmap plots of ATAC-seq and E14.5 mouse facial prominence H3K27Ac ChIP-seq [54] in tissue-specific NFRs. **C** Plot of average density of H3K27Ac ChIP-seq signal, showing higher signal at cluster 1 elements than cluster 2 elements. **D** GO enrichment (MGI mouse gene expression pattern) of nearest genes of cluster1 of palate epithelium NFR. **E** and **F** UCSC Genome browser views of the mouse genome (mm10 build) showing the ATAC-seq and H3K27Ac ChIP-seq signals near the *Krt17* and *Runx2* genes. Red box, an example of a palate-epithelium active enhancer (PEAE). Blue boxes, examples of palate mesenchyme active enhancers (PMAEs). **G** Motifs enriched in cluster 1 of E14.5 palate epithelium specific NFRs (i.e., PEAEs). Motifs shared with zGPAEs are in bold.

Figure 5-source data 1 Density plot for H3K27Ac ChIP-seq in two clusters, as plotted in Figure 5C

Figure 5-source data 2 Barchart for GO enrichment, as plotted in Figure 5D

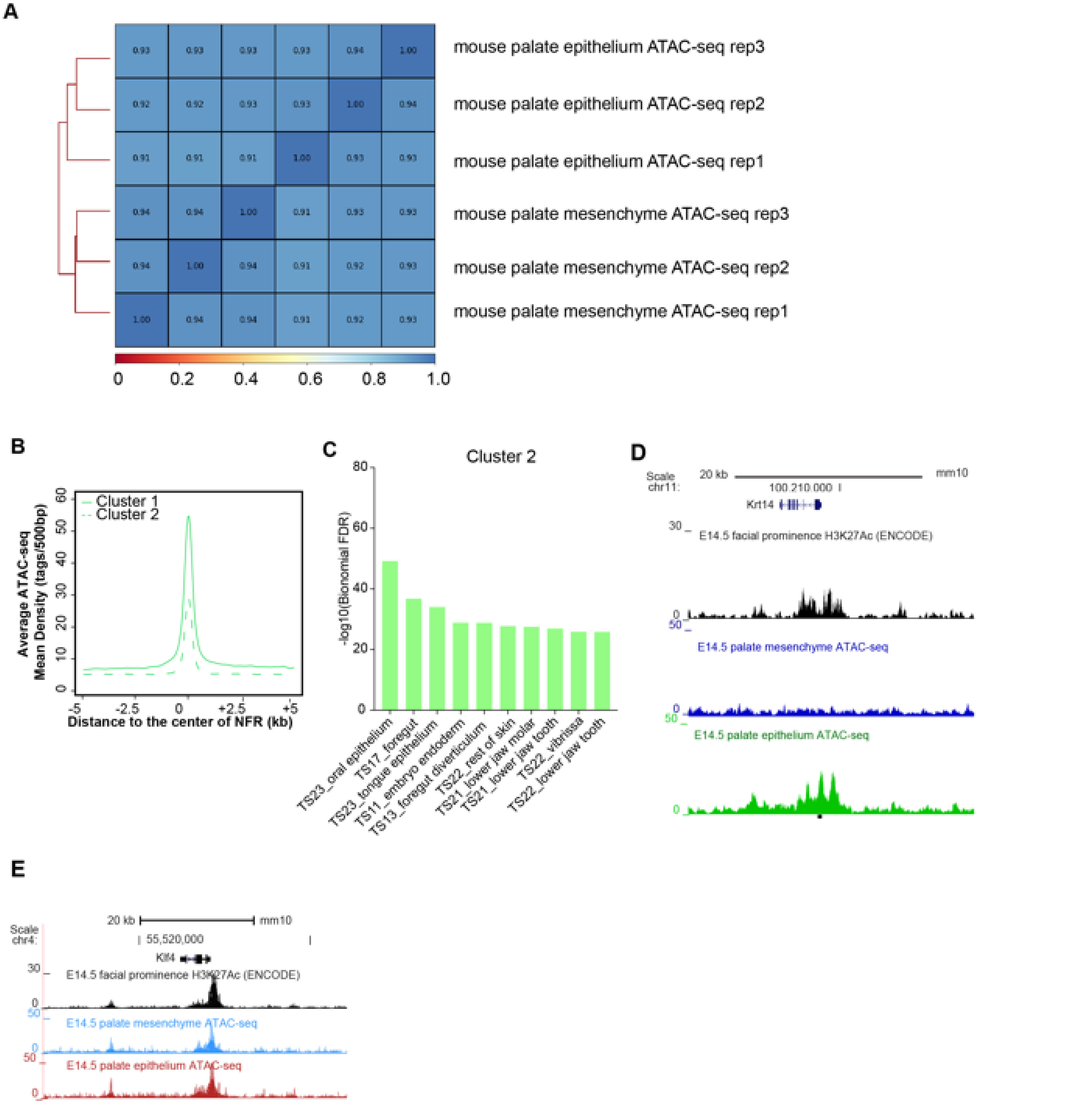


Figure 5—figure supplement 1 Concordance of replicates of mouse embryonic palatal epithelium ATAC-seq. **A** Correlation of three biological replicates of E14.5 mouse palate epithelium and mesenchyme ATAC-seq results. **B** ATAC-seq density plot of different clusters of E14.5 mouse palate epithelium specific NFRs. **C** GO enrichment (MGI mouse gene expression pattern) of nearest genes of cluster2 of palate epithelium NFR. **D** and **E** UCSC Genome browser view showing the ATAC-seq and H3K27Ac ChIP-seq signals in *Krt14* and *Klf4* locus.

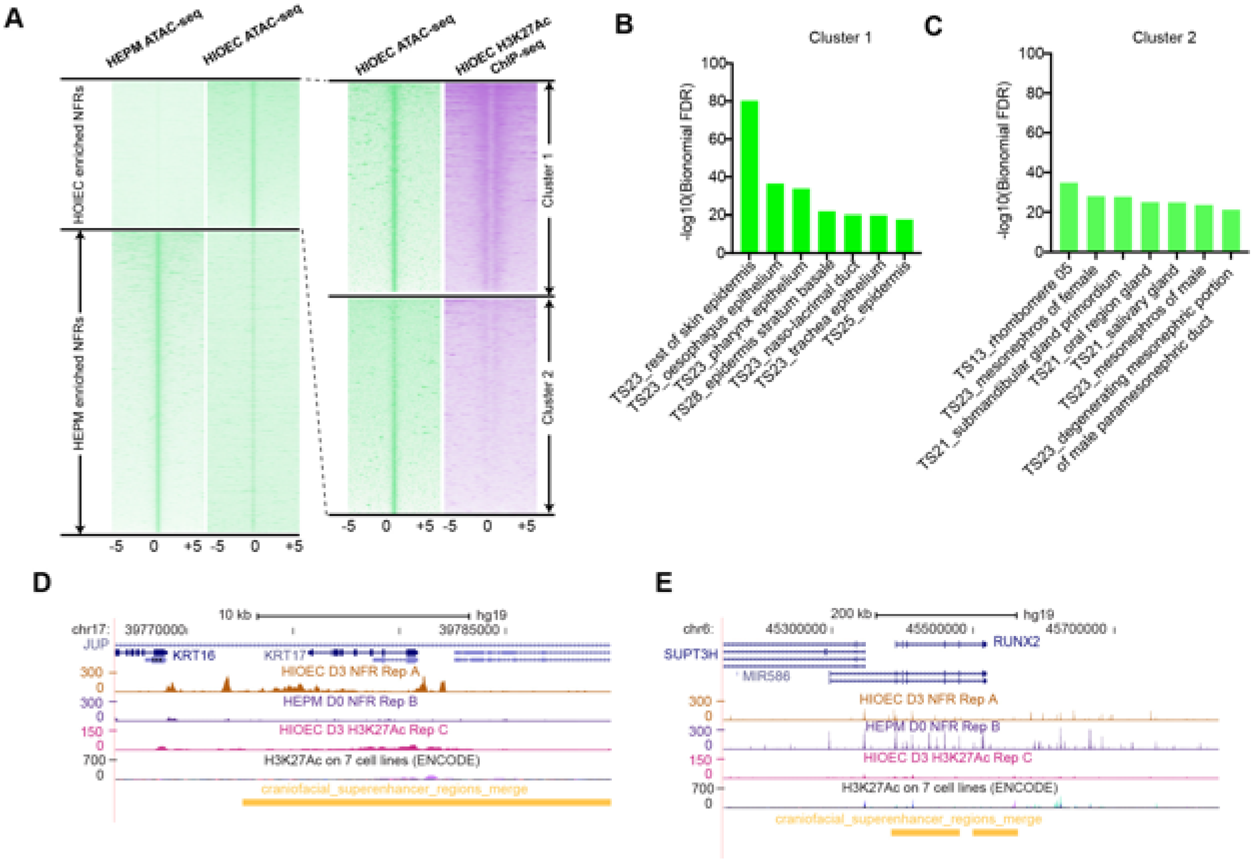
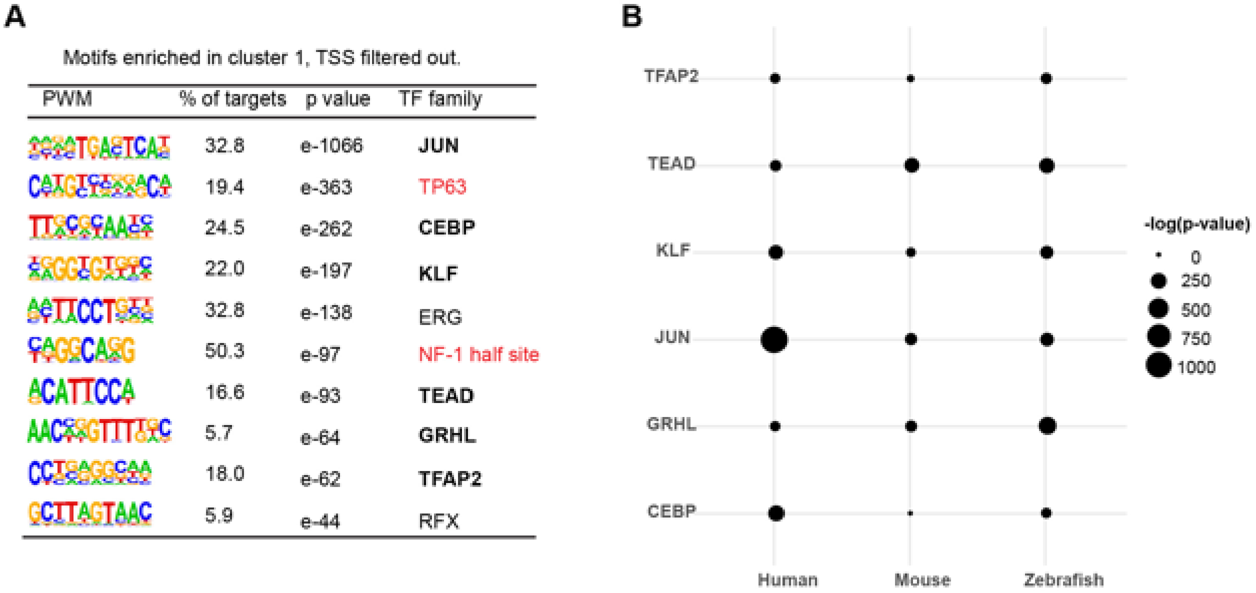


Figure 5—figure supplement 2 Summary of ATAC-seq in HIOEC and HEPM cells. **A** Heatmap plots of ATAC-seq of HIOEC- and HEPM-specific NFRs. **B**. GO enrichment for the genes near cluster 1 of HIOEC-specific NFRs. **C**. GO enrichment for the genes near cluster 2 of HIOEC-specific NFRs. **D and E** UCSC Genome browser tracks showing the HIOEC and HEPM ATAC-seq and NHEK H3K27Ac ChIP-seq signals in *IRF6* **(D)** and *RUNX2* **(E)** locus.



Motifs enriched in cluster 1, TSS filtered out.

PWM	% of targets	p value	TF family
	32.8	e-1066	**JUN**
	19.4	e-363	TP63
	24.5	e-262	**CEBP**
	22.0	e-197	**KLF**
	32.8	e-138	ERG
	50.3	e-97	NF-1 half site
	16.6	e-93	**TEAD**
	5.7	e-64	**GRHL**
	18.0	e-62	**TFAP2**
	5.9	e-44	RFX

Figure 5—figure supplement 3 Motifs enriched in hOEAEs and shared among zGPAEs, mPEAEs and hOEAEs. Bold, shared motifs enriched in all three epithelial tissues. **B** The significance of enrichment of each of the shared motifs among hOEAEs, mPEAEs and zPEAEs.

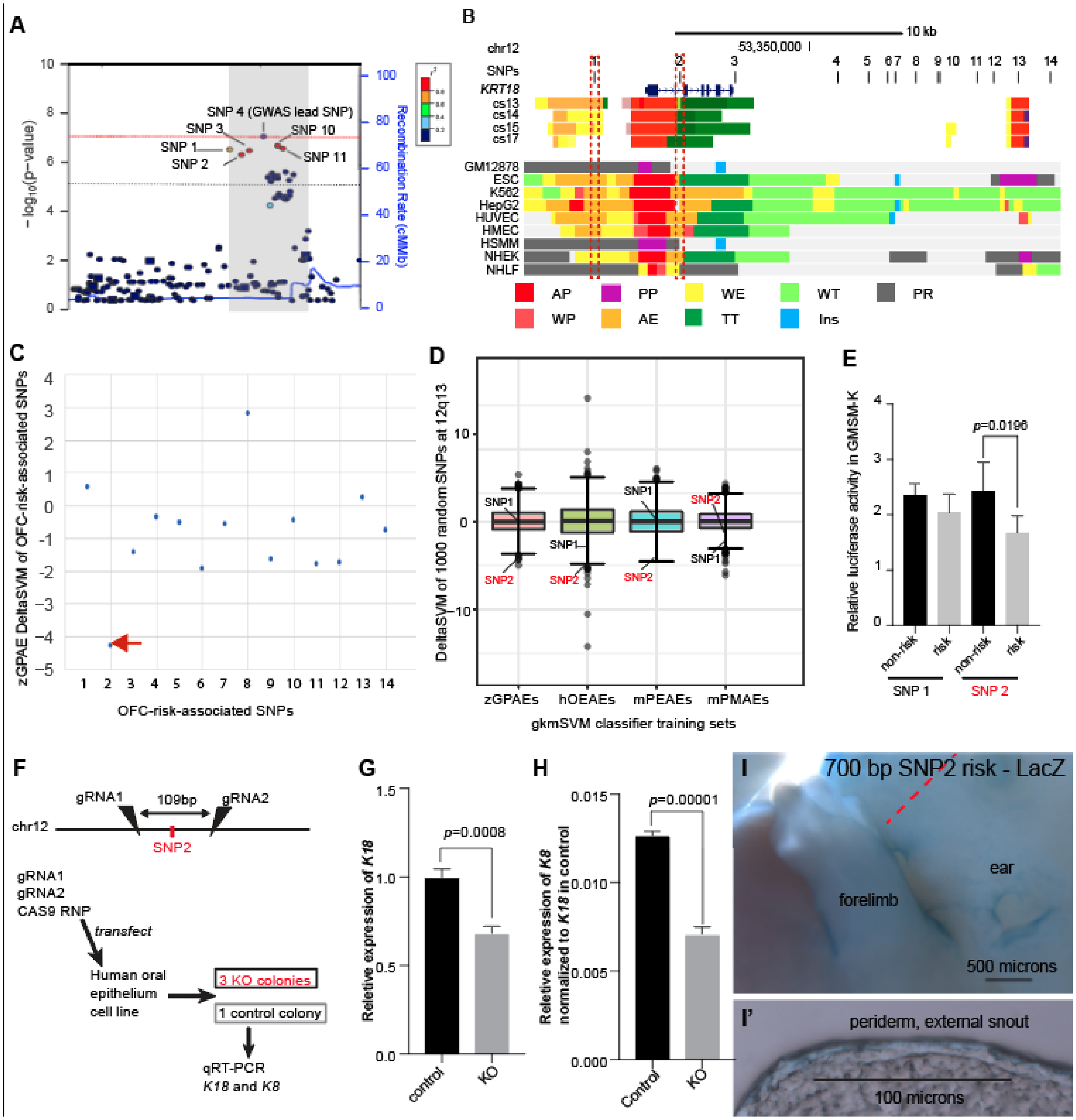


Figure 6 Classifier trained by zGPAEs helps prioritize orofacial clefting (OFC) functional variants in *KRT18* locus. **A** Regional plot showing OFC-risk-associated single nucleotide polymorphism (SNPs) near *KRT18* from this study. SNP4 is the lead SNP from our meta-analysis of OFC GWAS [8]. **B** Browser view of the human genome, hg19, focused on this locus. Tracks: **SNPs:** OFC-risk-associated SNPs. SNP1: rs11170342, SNP2: rs2070875, SNP3: rs3741442, SNP4: rs11170344, SNP5: rs7299694, SNP6: rs6580920, SNP7: rs4503623, SNP8: rs2363635, SNP9: rs2682339, SNP10: rs111680692, SNP11: rs2363632, SNP12: rs4919749, SNP13: rs2638522, SNP14: rs9634243. **Color coded bars**: Chromatin status (color code explained in key), revealed by

ChIP-seq to various chromatin marks. Cs13-cs17, facial explants from human embryos at Carnegie stage (cs) 13-17, encompassing the time when palate shelves fuse [55]. Roadmap Epigenomics Project cell lines [42]: GM12878, B-cell derived cell line; ESC, Embryonic stem cells; K562, myelogenous leukemia; HepG2, liver cancer; HUVEC, Human *umbilical vein endothelial cells*; HMEC, human mammary epithelial cells; HSMM, human skeletal muscle myoblasts; NHEK, normal human epidermal keratinocytes; NHLF, normal human lung fibroblasts. AP, active promoter; WP, weak promoter; PP, poised promoter; AE, active enhancer; WE, weak enhancer; TT, transcriptional transition; WT, weakly transcribed; Ins, insulator; PR, polycomb-repressed. **C** deltaSVM scores predicted by zGPAEs-derived classifier for the 14 OFC associated SNPs near *KRT18*. **D** Box and whisker plots of deltaSVM scores of 1000 randomly-selected SNPs near KRT18, scored by classifiers trained by zGPAEs (zebrafish periderm active enhancers), hOEAEs (human oral epithelium active enhancers), mPEAEs (mouse palatal epithelium active enhancers) and mPMAEs (mouse palatal mesenchyme active enhancers). The line is the median scoring SNP, the box contains the middle-scoring two quartiles, and the whisker represent the top and lower quartiles. Dots are outliers. deltaSVM scores for SNP1 and SNP2 are indicated. Number out of 1000 randomly selected SNPs with a lower deltaSVM than SNP2 with classifier trained on zGPAEs, 2; on mPEAEs, 9; on hOEAEs, 17; on mPMAEs, 186. **E** Dual luciferase assay for non-risk and risk alleles of rs11170342 (SNP1) and rs2070875 (SNP2) in GMSM-K cells. **F** Schematic diagram showing the workflow of generating GMSM-K cell colonies with 109bp flanking SNP2 deleted by CRISPR-Cas9. **G,H** qRT-PCR showing relative RNA expression of *KRT18* **(G)** and *KRT8* **(H)** in three homozygous knockout colonies (KO) and one isolated wildtype colony (Control) of GMSM-K cell lines. **I** Lateral view of transgenic mice LacZ reporter assay for the 701bp DNA fragment overlapping SNP2. **I'** Section of the facial prominence from I (red circled region)

Figure 6-source data 1 Barchart for relative dual luciferase activity in GMSM-K cells, as plotted in Figure 6E

Figure 6-source data 2 Barchart for relative gene expression of *K18* and *K8* in GMSM-K cells, as plotted in Figure 6G and H

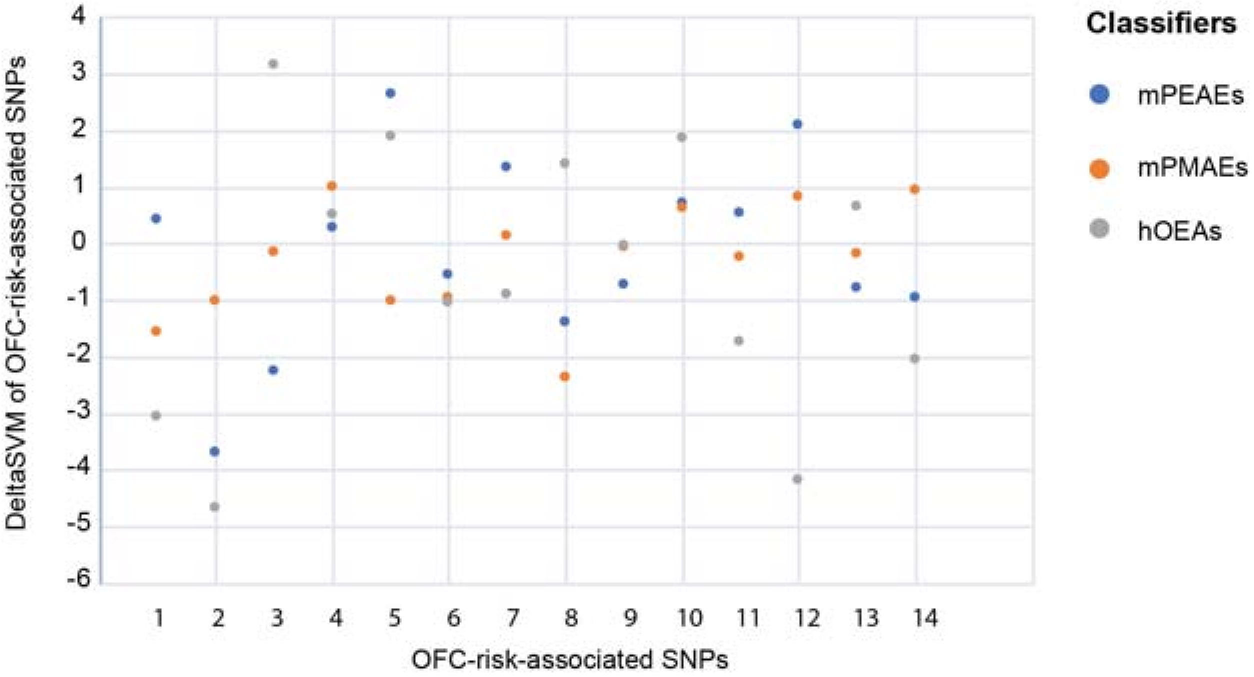


Figure 6—figure supplement 1 Dotplot of deltaSVM scores for each SNP calculated with classifiers trained on the indicated set of enhancer candidates.

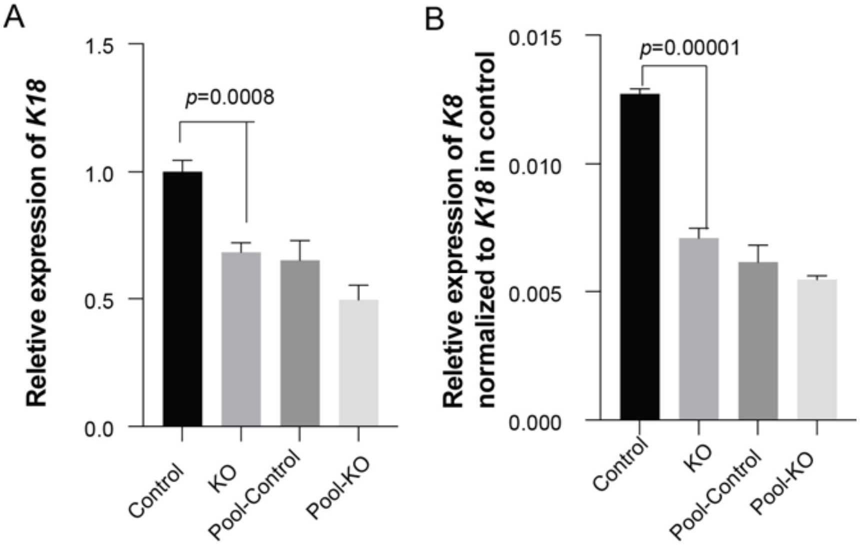


Figure 6—figure supplement 2 Bargraphs showing relative RNA expression of K18 (**A**) and K8 (**B**) in GMSM-K cells. KO: three homozygous knockout colonies; Control: one isolated wildtype colony; Pool-control: pool of GMSM-K cells transfected with two gRNAs only; Pool-KO: Pool of GMSM-K cells transfected with two gRNA along with Cas9 RNP.

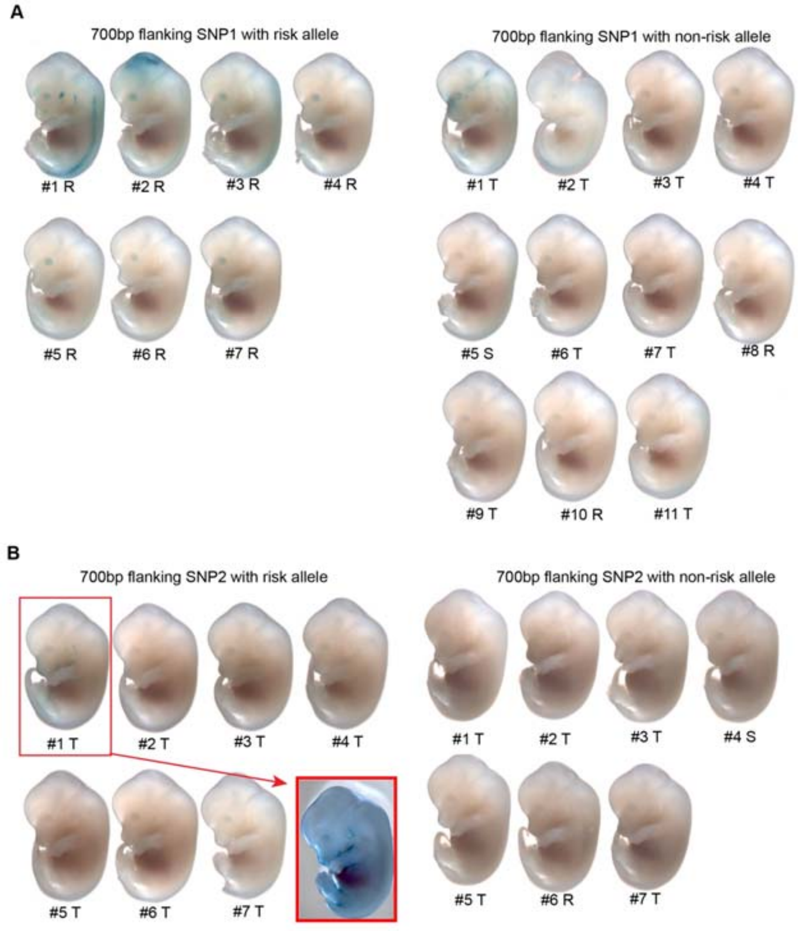

A
700bp flanking SNP1 with risk allele
#1 R
#2 R
#3 R
#4 R
#5 R
#6 R
#7 R
700bp flanking SNP1 with non-risk allele
#1 T
#2 T
#3 T
#4 T
#5 S
#6 T
#7 T
#8 R
#9 T
#10 R
#11 T
B
700bp flanking SNP2 with risk allele
#1 T
#2 T
#3 T
#4 T
#5 T
#6 T
#7 T
700bp flanking SNP2 with non-risk allele
#1 T
#2 T
#3 T
#4 S
#5 T
#6 R
#7 T

Figure 6—figure supplement 3

Lateral views of all wild-type mouse embryos for *LacZ* reporter assay**A** Embryos injected with a reporter construct built from a 700 bp element centered on SNP1, harboring the risk or non-risk allele as indicated. The large majority of embryos with SNP1 constructs, of either allele, were not blue, and no two blue embryos showed the same pattern. No further copy number analysis was not carried out. **B** Embryos injected with a reporter construct built from a 700 bp element centered on SNP2, harboring the risk or non-risk allele as indicated. Using the genomic DNA isolated from each embryo, PCR was carried out to determine if the reporter construct was present at all, and whether it was (**S** - single) present at the safe harbor locus in a single copy, (**T** - tandem), present at the safe harbor locus in more than one copy, or (**R**-random) was detectable but absent from the safe harbor locus, suggesting it integrated randomly into the genome. One embryo (number 1, boxed) injected with a SNP2 construct (risk-allele) showed reporter activity in the periderm, as predicted. Quantitative PCR indicated this embryo had 8-10 copies of the reporter construct while the other T embryos had 2.

Supplementary File 1 Coordinates of ATAC-seq and ChIP-seq peaks identified in this study

Supplementary File 1a Summary of peak numbers for all ATAC-seq and H3K27Ac ChIP-seq generated in this study

Supplementary File 1b Coordinates of GFP-positive NFRs flanked by H3K27Ac^{High} (zGPAEs)

Supplementary File 1c Coordinates of GFP-positive NFRs flanked low in H3K27Ac signals

Supplementary File 1d Coordinates of GFP-negative NFRs flanked by H3K27Ac^{High} (GNAEs)

Supplementary File 1e Coordinates of GFP-negative NFRs flanked low in H3K27Ac signals

Supplementary File 1f Coordinates of fish zGPAEs training set (zv9)

Supplementary File 1g Coordinates of mouse palate mesenchyme enriched NFR

Supplementary File 1h Coordinates of mouse palate epithelium enriched NFR

Supplementary File 1i Coordinates of mouse palate epithelium specific active enhancers

Supplementary File 1j Coordinates of HIOEC-specific NFRs

Supplementary File 1k Coordinates of HIOEC-specific active NFRs (flanked or overlapped with H3K27Ac ChIP-seq in HIOEC)

Supplementary File 2 Zebrafish ppl and human PPL enhancer alignments using ClustalO

Supplementary File 2a Alignments summary for enhancer homology test between *ppl-10* and *PPL-8.3*.

Supplementary File 2b Alignments details for enhancer homology test between *ppl-10* and *PPL-8.3*. All alignments were conducted using the CLUSTALW algorithm with default parameters via the Clustal Omega server (https://www.ebi.ac.uk/Tools/msa/clustalo/). Alignments were then annotated to highlight identical blocks of length 5 to 6 bp long (cyan) or longer (yellow). See Materials & Methods for further details on the choice of enhancer fragments used in these alignments.

Supplementary File 3 deltaSVM score and JASPAR predicted TF binding changes in the KRT18 locus

Supplementary File 3a List of OFC-associated SNPs near KRT18 locus

Supplementary File 3b deltaSVM scores for 14 OFC-associated SNPs near KRT18 locus and 1000 random SNPs using classifiers trained by zGPAEs

Supplementary File 3c deltaSVM scores for 14 OFC-associated SNPs near KRT18 locus and 1000 random SNPs using classifiers trained by mPEAEs

Supplementary File 3d deltaSVM scores for 14 OFC-associated SNPs near KRT18 locus and 1000 random SNPs using classifiers trained by hOEAEs

Supplementary File 3e deltaSVM scores for 14 OFC-associated SNPs near KRT18 locus and 1000 random SNPs using classifiers trained by mPMAEs

Supplementary File 3f Effects of different alleles of SNP1 and SNP2 on transcription factor binding sites, predicted by JASPAR