Two non-coding variants associated with isolated orofacial cleft promote binding of transcriptional repressors FOXE1 or ETS2 and reduce expression of *IRF6*

Priyanka Kumari[1], Ryan Z. Friedman[2,3], Lira Pi[4], Sarah Curtis[5], Kitt Paraiso[6], Axel Visel[6,7,8], Lindsey Rhea[9], Martine Dunnwald[9], Anjali P. Patni[1,10,11,12], Daniel Mar[11,13], Karol Bomsztyk[11,13,14], Julie Mathieu[11,15], Hannele Ruohola-Baker [1,10,11,16,17], Elizabeth Leslie[5], Michael A. White[2,3], Barak A. Cohen[2,3], Robert A. Cornell[1]

[1]Department of Oral Health Sciences, University of Washington, Seattle, Washington, USA

[2]Department of Genetics, Washington University School of Medicine in St. Louis, St. Louis, Missouri, USA

[3]The Edison Family Center for Genome Sciences & Systems Biology, Washington University School of Medicine in St. Louis, St. Louis, Missouri, USA

[4]PharmaLex, 1700 District Ave, Burlington, MA, USA

[5]Department of Human Genetics, Emory University School of Medicine, Atlanta, Georgia, USA

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

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

[8]University of California, Merced, California, USA

[9]Department of Anatomy and Cell Biology, Carver College of Medicine, University of Iowa, Iowa City, Iowa, USA

[10]Department of Biochemistry, University of Washington School of Medicine, Seattle, WA 98195, USA

[11]Institute for Stem Cell and Regenerative Medicine, University of Washington School of Medicine, Seattle, WA 98109, USA

[12]Cancer Biology and Stem Cell Biology Laboratory, Department of Genetic Engineering, School of Bioengineering, College of Engineering and Technology, SRM Institute of Science and Technology, Chennai 603203, India

[13]UW Medicine South Lake Union, University of Washington, Seattle, WA 98109, USA

[14]Matchstick Technologies, Inc, Kirkland, WA 98033, USA

[15]Department of Comparative Medicine, University of Washington School of Medicine, Seattle, WA 98195, USA

[16]Brotman Baty Institute for Precision Medicine, Seattle, WA 98195, USA

[17]Department of Bioengineering, University of Washington, Seattle, WA 98195, USA

Correspondence: cornellr@uw.edu

Running title: *MPRA on SNPs associated with OFC*

Key words: MPRA, orofacial cleft, IRF6, FOXE1, ETS2

Abstract

Genome wide association studies of isolated orofacial cleft (OFC) have identified common single nucleotide polymorphisms (SNPs) at the 1q32/*IRF6* locus and seven other loci where, like *IRF6*, the presumed OFC-relevant gene is expressed in oral epithelium. To identify the functional subset of SNPs at these loci we conducted a massively parallel reporter assay (MPRA) in a cell line derived from fetal oral mucosa. This assay identified SNPs with allele-specific effects on reporter activity levels at all eight loci; at four of the loci, these SNPs were within chromatin marked as an active enhancer in human embryonic faces. Results from traditional luciferase reporter assays concorded with those from the MPRA at nine of eleven SNPs tested. We separately engineered two promising SNPs from the 1q32/*IRF6* locus to homozygosity in the genome of induced pluripotent stem cells and then differentiated the cells into embryonic oral epithelium. We found that cells homozygous for the risk-associated alleles had lower levels of *IRF6* expression, and that these alleles favored the binding of transcriptional repressors, FOXE1 and ETS2, respectively. Mutations in *FOXE1* also cause a syndromic form of OFC, and ETS2 suppresses differentiation of epithelia. Conditional analyses of a meta-analysis of multi-ethnic genome wide association studies suggest that the two SNPs account for the majority of risk for OFC associated with variation at 1q32/*IRF6*. This study connects genetic variation associated with orofacial cleft to mechanisms of pathogenesis.

Introduction

Oral facial cleft (OFC) is a multifactorial disorder with genetic predisposition. More than one hundred syndromes include OFC as a phenotype and overall about 30% of cases are syndromic [1]. Such syndromes are generally caused by mutations in single genes and follow a Mendelian inheritance pattern. The remaining cases are non-syndromic (isolated) meaning there is cleft in the lip and/or in the palate in the absence of other structural abnormalities. The etiology of non-syndromic OFC is incompletely understood but includes a role for genetics as the concordance of non-syndromic OFC is 50% in identical twins as opposed to 3-5% in other first-degree relatives [2-4]. Multiple independent genome wide association studies (GWAS), and meta-analyses thereof, have identified more than 40 loci where a specific allele (i.e., the risk allele) of common single nucleotide polymorphisms (SNPs) are over-represented in cases versus in controls with the same ancestry (reviewed in [5]). Importantly, however, the majority of the heritable risk for isolated OFC has not been assigned to any gene or locus.

A portion of the missing heritability for OFC is likely to be found in genes belonging to the same transcriptional regulatory networks known as OFC-risk genes. GWAS loci may reveal the membership of those networks, because most SNPs identified in GWAS are in non-coding DNA, and the subset of these SNPs that are functional presumably alters the binding of transcription factors. However, pinpointing the functional variants is challenging for two reasons. First, at most GWAS-identified loci there are multiple SNPs in strong linkage disequilibrium with the lead SNP and any one of them may be functional. Second, functional SNPs disrupt *cis*-regulatory sequences (i.e., enhancers and promoters) whose sequence constraints remain poorly understood. Investigators have mined sets of tissue-specific enhancers for sequence patterns [6], and there are *in silico* tools for inferring variants that affect enhancer function [7], but currently there is no tool that can robustly identify non-coding variants that alter enhancer activity. An alternative approach is the massively parallel reporter assay (MPRA). MPRA have been widely used to detect elements with enhancer activity, and in some cases to detect the effect of variants on enhancer activity [8-17]. A challenge in deploying MPRA to identify functional SNPs is that the results are highly dependent on the cellular context [12] because many enhancers exhibit tissue-specific activity.

The 1q32/*IRF6* locus is strongly associated with OFC in multiple ethnic groups [18-29]. Moreover, it is estimated that variation at this locus accounts for about 12% of genetic contribution to isolated OFC [30], the most for any single locus. A strong candidate for the risk-relevant gene in

this locus is *IRF6,* encoding the transcription factor Interferon Regulatory Factor 6, because mutations in *IRF6* are found in about 70% of patients with Van der Woude syndrome, the most common syndromic form of OFC (VWS1, OMIM # 119300). Finding the functional SNPs at this locus, and the transcription factors whose binding is affected by these SNPs, could be particularly helpful in revealing loci that contribute to the missing heritability for OFC. However, a meta-analysis of several GWAS identified more than 600 SNPs with p-values indicating at least a suggestive association with OFC [20]. The large number of OFC-associated SNPs at this locus makes it difficult to determine which are functional.

Here we deployed an MPRA to nominate which OFC-associated SNPs are functional at *IRF6*/1q32 and several other loci where the OFC-risk-relevant gene is expressed in oral epithelium. We tested 889 SNPs in an MPRA in an oral epithelial cell line and validated a subset of the MPRA results using traditional luciferase assays in the cell line and in primary keratinocytes. For two promising SNPs near *IRF6*, we carried out extensive follow up experiments in oral epithelial cells derived from induced pluripotent stem cells whose genotype had been engineered. These studies support the two SNPs as being functional, explain how they are functional, and show that they account for most of the heritable risk for OFC phenotypes attributed to this locus.

Materials and Methods

SNPs Selection of GWAS-identified loci/variants

We selected the OFC-associated SNPs that lie near a gene known to be expressed in oral epithelium or epidermis. For this study we first re-performed study on meta-analysis data of 4 NSCL/P GWASs from European and Asian ancestries [20]. With this criterion we selected SNPs from 6 loci: 1q32/*IRF6*, 2p21/*THADA*, 3q28/*TP63*, 9q22.2/*GADD45G*, 12q13.13/*KRT18*, 20q12/*MAFB* with *IRF6* locus alone harboring 608 SNPs (**Table 1, Table S1**). We also included 9 SNPs from the 9q22.33/*FOXE1* locus that we previously found to be present in predicted regulatory elements, and 112 SNPs from the 6p24.3/*TFAP2A* locus reported in the one GWAS to detect a signal at this locus [23].

Reporter Library Construction

We designed an oligonucleotide library in which each oligo contained the sequence of the SNP of interest, a unique barcode, and priming and restriction sites necessary for amplification and

cloning into a pGL-hsp68 reporter plasmid [31]. The sequence of human genomic DNA corresponding to 161 bp windows, centered on each of 889 GWAS-identified SNPs, was downloaded from the UCSC Table browser tool (genome build: hg19). We generated two versions of each cis regulatory element (CRE) that were identical except for harboring the major or minor allele of the SNP, respectively (1778 elements) (**Fig.1, Table S2**). We ordered a pool of 7,500 unique 230-mer oligonucleotides (elements) through a limited licensing agreement with Agilent Technologies (Santa Clara, CA). To establish an empirical null distribution of activity in the assay, we included 90 sequences that were scrambled versions of the sequences of other elements. To account for noise in the assay, each sequence was assigned four unique barcodes. Finally, to create constructs with only the basal promoter and no upstream regulatory sequence, we included an additional random 28 filler elements, which were excised in the second step of cloning, yielding 28 basal promoter only controls with one barcode each. In total we synthesized 7500 elements (7112+360+28) (**Table S3, Fig. 1**).

Oligos were designed with the following structure (**Fig. S1**): a forward priming sequence (PS1) with NheI restriction site, (primer for this site is KC_1F), 161 base cis-regulatory element (**Table S2**), AgeI restriction site, NruI restriction site, MluI restriction site, 9 base barcode, and a reverse priming sequence (PS2) with XhoI site (primer for this site is KC_1R) (**Fig. 1A, Fig. S1, Table S3**).

We amplified the oligo pool using five cycles of PCR with Phusion High-Fidelity polymerase (New England Biolabs) and primer set KC_1 (**Table S4**). We cloned the pool of amplicons into the pGL-hsp68 backbone using NheI and XhoI which excludes the hsp68 promoter fused to DsRed. We prepared DNA from 41,000 colonies to generate library KC_L1. We then amplified a DNA fragment containing the hsp68 promoter and the DsRed cDNA using primers KC_2F and KC_2R (**Table S4**) containing AgeI and MluI restriction sites, respectively, and cloned this fragment into the KC_L1 library such that the CREs lie upstream to the hsp68 promoter driving DsRed and the barcodes; this procedure created library KC_L2.

To assess the completeness of the cloned library, we sequenced the barcodes in the final plasmid pool in library KC_L2. It was achieved by amplifying the barcode fragments using primer set KC_BC containing with NheI and MfeI restriction sites (**Table S4**), followed by specific Illumina adapters ligation (P1 and PE2 adapters; **Table S4**), and Illumina enrichment PCR using KC_3 set of primers (**Table S4**), and sequencing.

Cell lines

The human fetal oral epithelial cell line (GMSM-K)[32] (a kind gift from Dr. Daniel Grenier, Université Laval) was maintained in keratinocyte serum-free medium (KSFM) supplemented with human recombinant epidermal growth factor (EGF 30µg/ml) and bovine pituitary extract (BPE 0.2 ng/ml) (Thermo Fisher Scientific). Primary Epidermal Keratinocytes: Normal, Human, Neonatal Foreskin (HEKn) purchased from ATCC (ATCC Catalog no. PCS-200-010) were maintained and cultured as per manufacturer's instruction in dermal cell basal medium (ATCC Catalog no. PCS-200-030) supplemented with keratinocyte growth kit (ATCC Catalog no. PCS-200-040). Normal Human Skin Keratinocytes (NHSK) [33] were grown in KSFM or DMEHAM according to Rheinwald and Green [34].The human induced pluripotent stem cells (iPSCs), WTC-11 (Coriell Institute for Medical Research, Catalog no. GM25256), were maintained in mTeSR Plus media (StemCell Technologies) on Matrigel (Corning) coated plates and passaged with Accutase. Cells were incubated at 37°C in 5% CO_2.

Transfection of CRE-library (plasmid pool)

A pool of plasmids (from library KC_L2; as described in method section - Reporter library construction) in four replicates, 2.5 µg each, was transfected into 1 million GMSM-K with Amaxa Cell Line Nucleofector Kit V (Lonza, Cologne, Germany) using Nucleofector II (Lonza) (program: T-020). Forty eight hours post transfection RNA and genomic DNA was extracted from each set using Quick-DNA/RNA Miniprep kit (Zymo Research Catalog no. D7001) as per manufacturer's instructions.

Preparing Samples for RNA-Seq

Isolated RNA was treated with Turbo DNaseI (ThermoFisher Catalog no. AM1907) as per manufacturer's instructions. First strand of cDNA was synthesized with SuperScript III First-Strand Synthesis System (Thermo Fisher Scientific Catalog no. 18080051). The barcodes including the flanking sequence were amplified from both cDNA, genomic DNA (transfected plasmid DNA) and plasmid DNA samples (that was used for transfection) with Phusion High-Fidelity PCR Master Mix (NEB Catalog no. M0531S) using primer set KC_BC (**Table S4**) and that yielded barcode sequences that were flanked with NheI and MfeI restriction enzyme sites. The products were digested with NheI and MfeI and then ligated with Illumina adapter sequences P1 and PE2 (**Table S4**). This product was further amplified using primer set KC_3 (**Table S4**) followed by sequencing.

To measure expression of the barcoded library, electroporation replicates were multiplexed and run on two lanes of an Illumina HiSeq machine, which generated 48.2 million sequence reads corresponding to cDNA and 48.8 million reads corresponding to DNA (from plasmids). Sequencing reads that matched the first 20 nucleotides of designed sequence were counted, regardless of quality score.

Statistical analysis to determine regulatory effects of SNPs by MPRA

We filtered out barcodes that were low abundance in the plasmid pool (<50 reads), then set any cDNA barcodes that were detection-limited in at least one sample (<50 reads for replicates run on the first lane, <100 reads for the replicates run on the second lane) to zero across all cDNA samples. We normalized each sample to the total counts per million (CPM). RNA barcode levels were highly correlated among 4 replicates (**Fig. S2**). There was a strong correlation in the barcode counts in the plasmid library before transfection and after harvesting plasmid from transfected cells (**Fig. S3**). Therefore, we used plasmid barcode counts (pre-transfection) for further analyses. We further normalized RNA barcode counts by the plasmid barcode counts (RNA/DNA). We checked the reproducibility based on log2RNA/DNA counts (**Fig. S4**). To calculate the reporter activity for each sequence we averaged the log2RNA/DNA counts within a replicate, normalized to the within-replicate mean basal barcode activity and then averaged across all the replicates (**Fig. S5A**). The standard error of mean (SEM) was calculated using the basal-normalized activities across replicates. Variant significance was determined by performing Welch's t-test between major and minor alleles followed by FDR correction (**Fig.1B**). We used our previously published [35] Python code for processing and analyzing MPRA reads, which is available on Github at [https://github.com/barakcohenlab/CRX-Information-Content].

Activity-by-Contact to assign SNP-containing-enhancers to promoters

The activity-by-contact method (ABC) utilizes cell-type specific chromatin accessibility (ATAC-seq or DNase hypersensitivity) and chromatin activity (H3K27ac) and HiC data that is either cell-type specific or averaged over 10 cell types –the performance of this tool is similar with either type of HiC data [36]. We used Normal Human Epidermal Keratinocytes (NHEK) specific publicly available datasets to calculate ABC score: GSM733674 for H3K27ac, GSM736545 for DNase hypersensitivity, GSM958736 for RNA-seq and averaged HiC data from 10 cell types as reported in Fulco et al [36]. Elements with ABC score >0.01 was included in the analysis.

Electroporation and dual luciferase reporter assay

For dual luciferase assays, each reporter construct (**Table S5**) was cloned into the pGL3 promoter vector and constructs were co-transfected with Renilla luciferase plasmid. Briefly, GMSM-K or HEKn were electroporated with plasmid using the Amaxa Cell Line Nucleofector Kit (Lonza) and the Nucleofector II instrument (Lonza) program T-020. We used a dual-luciferase reporter assay system (Promega) and FB12 Luminometer (Berthold Detection Systems) to evaluate the luciferase activity following manufacturer's instructions. Relative luciferase activity was the ratio of values for the Firefly and Renilla enzymes. At least three independent measurements were performed for each transfection group. All results are presented as mean ± standard deviation. Statistical significance was determined using two-tailed Student's t-test.

Immunofluorescence detection and imaging

GMSM-K and NHSK were grown on glass coverslips, fixed in 4% paraformaldehyde for 10 min, permeabilized in 0.2% Triton-X100 for 2 min, then rinsed in PBS. Cells were incubated in 3% goat serum (Vector Laboratories, Catalog no. S-1000-20) in 2% PBS-Bovine Serum Albumin for 30 min to block non-specific binding, followed by incubation in primary then secondary fluorescent antibodies 1 h at room temperature with PBS washes in between the two incubations. Primary antibodies that were used are: mouse monocolonal anti-E-cadherin (BD Bioscience, Catalog no. 610181, 1/100), mouse monoclonal anti P63 (clone 4A4, Santa Cruz, Catalog no. SC-8431, 1/250), mouse monoclonal anti human Keratin 14 (clone LL002, Novacastra, Catalog no. NCL-LL002, 1/200). Goat anti-mouse Alexafluor 488 (In Vitrogen, Catalog no. A-110001) was used as secondary antibody. Following PBS washes, coverslips were dried and mounted on a microscope slide using Prolong antifade mounting medium containing DAPI (ThermoFisher Scientific, Catalog no. P36931) and let cured for 24 h at room temperature in the dark. Images were collected with Zeiss 880 or 980 confocal microscopes using the ZEN software (Zeiss). Confocal images were processed using Fiji software.

CRISPR-Cas9 mediated genome editing

Genome engineering for both SNPs (rs11119348 and rs661849) in iPSCs or GMSM-K was achieved using a previously described protocol [37]. Sequences of CRISPR RNAs (crRNA) and repair template specific to each SNPs are provided in (**Fig. S6A-B**). Briefly, equimolar concentrations of specific crRNA and trans-activating crRNA (tracrRNA, IDT Catalog no. 1072532) were annealed at 95°C for 5 min followed by cooling at room temperature (RT) to

form functional gRNA duplexes. Further, the ribonucleoprotein (RNP) complex was prepared 15 min before transfection by mixing gRNAs (1.5 µM) with Cas9 protein (0.45 µM) (Sigma, Catalog no. CAS9PROT for iPSCs and IDT Catalog no. 1081058 for GMSM-K) at RT. RNP complexes together with the repair template were transfected into 1 million iPSCs with Amaxa nucleofector (Lonza, Human Stem Cell kit 2, program: A-023 for iPSCs) in presence of ROCK inhibitor (Y-27632, Catalog no. 04-0012-10, Stemgent; 10 mM stock). Lonza Kit V, program: T-020 was used for transfecting GMSM-K. Individual colonies were hand-picked and plated into 96 well plates. DNA was extracted using Quick Extract DNA extraction solution (Epicentre Catalog no. QE09050). Colonies were screened by PCR and sequenced using primers flanking each SNP (Primer sequences are provided in **Table S6A**). From iPSCs transfection experiment, two independent clones each of homozygous risk (AA for rs11119348; CC for rs661849), and homozygous non-risk (CC for rs11119348; TT for rs661849) genotype were isolated. From GMSM-K transfection experiment, multiple independent clones, six clones of homozygous risk (AA for rs11119348), seven clones of homozygous non-risk (CC for rs11119348) and four independent clones each of the homozygous risk (CC for rs661849), and homozygous non-risk (TT for rs661849) were isolated. Deletion of *FOXE1* gene in GMSM-K was achieved using a pair of gRNA (**Fig. S7**). A set of three primers was used for colony screening (**Table S6B**). Sequencing of three independent clones revealed that all but 30 bp at the 3' end of the single coding exon gene was deleted (**Fig. S7**).

In vitro differentiation of iPSCs into embryonic oral epithelial cells

Two independent samples of each genotype (2 clones of the homozygous risk genotype, and 2 clones of the homozygous non-risk genotype) for both SNPs were seeded in triplicate in 12-well plate (50,000 cells per well). Further, the cells were induced to differentiate into embryonic oral epithelial (iOE) cells using a 10-day differentiation protocol described previously [46].

Quantitative RT-PCR on induced embryonic oral epithelial (iOE) cells or GMSM-K

RNA was extracted from multiple samples of each genotype (homozygous risk and homozygous non-risk) for rs11119348 and rs661849 (iOE cells – 6 each genotype for both SNPs; GMSM-K - 6 non-risk and 7 risk genotype for rs11119348 and 4 each genotype for rs661849) using Quick-DNA/RNA Miniprep kit (Zymo Research #D7001) followed by DNase treatment using manufacturer's instructions (Thermo Fisher Scientific #AM1907). Reverse transcription was performed using High-Capacity cDNA Reverse Transcription Kit (Thermo Fisher Scientific #

4368814). Real-time PCR was performed using SYBR Green qPCR mix (Bio-Rad) in Bio-Rad CFX Connect Real-Time System. Quantitative RT-PCR reaction for each sample was performed in triplicate. Data analysis was performed using previously described methods [38, 39]. Briefly, expression levels of *IRF6* were normalized against *ACTB* [38] for GMSM-K or HEKn and expression levels of *IRF6* are normalized against *ACTB, GAPDH, HPRT, UBC* and *CDH1* [39] for iOE cells (mentioned in the respective figure legends). Data were presented as mean ± s. d. Statistical significance was determined using the Student's *t*-test (two-tailed). Primer sequences used in the study for qRT-PCR are provided in **Table S6C**.

Transcription factor binding site analysis using JASPAR 2022

Ten nucleotides flaking each side of the *IRF6* -10 kb and -22 kb SNPs (21 nucleotides long), individually with non-risk and risk allele was used for transcription factor binding site (TFBS) analysis using JASPAR 2022 [40]. An 80% threshold was used to detect transcription factor binding sites, and all those with a score below 4.0 were discarded, to focus on binding sites of at least medium affinity. For TFBS that, at 80% threshold, were specific for the risk or non-risk allele, we dropped the threshold to 50% to learn the affinity of such TFBSs with the other allele.

Chromatin immunoprecipitation qPCR

Induced oral epithelial cells (iOE cells) derived from iPSC that are heterozygous for risk and non-risk alleles of rs11119348 and rs661849 or GMSM-K homozygous for risk or non-risk alleles for rs11119348 and rs661849 were harvested (~3.5 million cells) and fixed in 1% formaldehyde for 15 min at RT. Fixation was stopped with 125 mM glycine for 5 min at RT. Chromatin immunoprecipitation (ChIP) on iOE cells was performed using a semi-automated protocol described previously [41]. Briefly, cells were sheared using PIXUL or VirSonic 600 (5 min with 30/40 second on/off cycles for GMSM-K for *IRF6* -22 kb SNP) to obtain genomic DNA fragments averaging 300 to 600 bp. The sonicated cell lysates were subjected to chromatin immunoprecipitation with specific antibody (1 µg of anti-H3K27Ac from Millipore-Sigma Catalog no. 07-360 and 1 µg of anti-ETS2 from GeneTex Catalog no. GTX104527 anti-FOXE1 (a kind gift from Pilar Santisteban, Universidad Autónoma de Madrid, as previously described) [42]. Normal rabbit IgG (1 µg) was used as negative antibody control. Quantitative PCR was performed with equal volume of DNA in triplicate from three ChIP replicates of each antibody and percent input DNA was calculated using the method described previously [43]. Primer sequences are shown in **Table S6D**. Negative control PCR primers were designed to target a

sequence 103.7 kb upstream of the *IRF6* transcription start site that did not harbor active regulatory elements identified from ATAC-Seq and H3K27Ac ChIP-Seq in Human immortalized oral epithelial cells (HIOEC) or NHEK. Data were presented as mean ± s. d. Statistical significance was determined using the Student's *t*-test (two-tailed).

***ETS2* and *FOXE1* knockdown**

siRNAs targeting *FOXE1* (Santa Cruz Catalog no. sc-44175) or *ETS2* (Santa Cruz Catalog no. sc-37855) and siRNA controls (Santa Cruz Catalog no.sc-37007) were transfected into one million HEKn (heterozygous for *IRF6* -22 kb SNP) or GMSM-K (homozygous risk for *IRF6* -10 kb or -22 kb SNPs). RNA extraction and qRT-PCR were performed on three replicates. Data were presented as mean ± s. d. Statistical significance was determined using the Student's *t*-test (two-tailed). Primer sequences used in the study for qRT-PCR are provided in **Table S6C**. RT-PCR primer pair B from Santa Cruz (Catalog no. sc-44175-PR) for *FOXE1* was used.

Further, siRNA targeting *FOXE1* or *ETS2* and control siRNA were co-transfected with respective risk harboring luciferase constructs along with Renilla luciferase plasmid. Three independent transfections were performed. Cells were harvested 48 hours post transfection and dual luciferase reporter assays were performed as described earlier. The RNA expression and luciferase reporter assays were carried out on *FOXE1* KO GMSM-K as well. GraphPad Prism 10 was used for the statistical data analysis for luciferase reporter assays, RNA expression and ChIP-qPCR experiments.

Conditional Analysis Methods

Conditional analyses were conducted using genome-wide genotyping data from the Pittsburgh Orofacial Cleft (POFC) Study. The details of the sample collection and genotype quality control (QC) have been described previously [21]. Briefly, the original study recruited individuals with OFCs, their unaffected relatives, and unrelated controls (individuals with no known family history of OFCs or other craniofacial anomalies; N = 1626). Samples were genotyped for approximately 580,000 single nucleotide polymorphic (SNP) markers from the Illumina HumanCore+Exome array, of which approximately 539,000 SNPs passed quality control filters recommended by the Center for Inherited Disease Research (CIDR) and the Genetics Coordinating Center (GCC) at the University of Washington [21]. This data was then phased with SHAPEIT2 [44] and imputed with IMPUTE2 [45] to the 1000 Genomes Project Phase 3 release (September 2014) reference panel. The most-likely imputed genotypes were selected for statistical analysis if the highest probability

(r2) > 0.9. SNP markers showing deviation from Hardy-Weinberg equilibrium in European controls, a minor allele frequency or MAF < 5%, or imputation INFO scores < 0.5 were filtered out of all subsequent analyses. The information for the genotyped markers was retained after imputation and the imputed values for these variants were only used to assess concordance. Common variants (MAF > 0.05%) near the *IRF6* locus were included in this analysis. All analyses were done by comparing cases with either cleft lip only (N = 394) or cleft lip and palate (N = 1995) to the group of unrelated controls, controlling for sex and 10 PCs of ancestry. Conditional analyses also controlled for either the genotype of rs661849, rs11119348, or the genotypes of both rs661849 and rs11119348. The association between every genetic variant and OFCs were calculated using a logistic regression in PLINK (v1.9) [46]. Since the data came from a genome-wide analysis, SNPs with association p-values less than 5 × 10-8 were considered genome-wide significant. Regional association plots were made with LocusZoom, where the linkage disequilibrium (LD) blocks and recombination rates were estimated from European populations [47]. LD between pairs of SNPs was calculated in the data from the 1000 Genomes Project using LDLink [48, 49].

Results

A massively parallel reporter assay reveals candidate functional SNPs at 8 OFC-associated loci

We conducted an MPRA on OFC-associated SNPs at eight loci where the presumed OFC risk-relevant gene is expressed in the oral epithelium. For seven loci we picked all SNPs whose association with OFC was suggestive of significance ($p < 10^{-4}$) in a newly-performed meta-analysis of GWA studies (**Table 1**, **Table S1**) [20]. These loci were 1q32/*IRF6* [50], 2p21/*THADA* [51], 3q28/*TP63* [52], 6p24.3/*TFAP2A* [53], 9q22.2/*GADD45G* [54], 12q13.13/*KRT18* [55], and 20q12/*MAFB* [27] . For an eighth locus 9q22.33/*FOXE1* [20, 56, 57], we instead picked 9 SNPs that were in linkage disequilibrium with the lead SNP (rs1234719) from a European population and that were annotated by Haploreg as being in regulatory elements [20] (**Table 1, Table S1**). As outlined in **Fig. 1A**, we synthesized a library of reporter plasmids containing 161 base-pair genomic elements centered on an OFC-associated SNP; OFC risk and non-risk alleles of each SNP were represented in four replicate constructs with distinct barcodes (**Table S3**). We performed the MPRA in GMSM-K, a cell line derived from human fetal oral mucosa [32]. The enhancer activity of specific test elements was rarely higher than the basal promoter alone, which reflects that most SNPs do not lie in active enhancers (**Fig. 1B**). The enhancer activity of test elements

was modestly correlated with levels of H3 lysine 27 acetylation (H3K27Ac), a mark of active enhancers, in primary normal human epidermal keratinocyte (NHEK) cells (Pearson correlation coefficient r = 0.02) (**Fig. S5B**). Of the 889 SNPs tested, 65 had allele-specific effects on the level of enhancer activity (**Fig. 1B**)**,** nominating these as candidate functional SNPs. There was at least one such candidate at each of the loci included in the study; at the 1q32/*IRF6* locus there were 46 such candidates (**Fig. 1B** and **Table 1, Table S7**).

Filtering candidate SNPs for those present in annotated enhancers

We next filtered candidate functional SNPs for those lying in predicted enhancers. First, we used the activity-by-contact (ABC) method [36] to predict enhancers for the eight risk-relevant genes in NHEK. This method predicts enhancers for a given gene in a given cell type based on (1) RNA-seq, DNAse hypersensitivity and H3K27Ac ChIP-seq data from the cell type (here, from NHEK), and (2) averaged Hi-C chromatin-contact data from 10 ENCODE cell types [36, 58]. Second, we obtained enhancers predicted by chromatin marks in explants of human embryonic faces [59]. At two of the eight loci there were candidate SNPs lying in enhancers predicted by both methods (i.e., two SNPs near *IRF6*, one SNP near *FOXE1*) (**Figs. 2A,B** and **S8A**), and at two solely by the latter method (one SNP each near *MAFB* and *TFAP2A*) (**Fig. S8B,C**). The first candidate SNP near *IRF6*, rs11119348, lies within an element named "multi species conserved sequence 9.7 kb from *IRF6*" (MCS9.7) [60] that has enhancer activity in murine embryonic oral epithelium [61] and in zebrafish periderm [6]. This element is the site of a single-base-pair duplication that appears to cause Van der Woude syndrome [61], and contains three other SNPs, in addition to rs11119348, that are associated with non-syndromic OFC, including rs642961, the focus of an earlier study[60]. All four SNPs were tested separately in the MPRA but only rs11119348 had allele-specific effects in it. We will refer to this as the *"IRF6* -10 kb SNP" (**Fig. 2B**). The second SNP in this locus, rs661849, resides within an evolutionarily conserved sequence 21.7 kb upstream of the *IRF6* promoter and near the transcription start site of *UTP25.* This SNP is an eQTL for *IRF6* in the brain (GTEx). There are no other OFC-associated SNPs within the same apparent regulatory element. We will refer to this SNP as the "*IRF6* -22 kb SNP" (**Fig. 2B**). Of note, rs642961, mentioned above as the subject of a previous functional study [60], is in LD with *IRF6* -22 kb SNP (R^2 = 0.76) but not with *IRF6* -10 kb SNP (R^2 = 0.03) (based on data from the 1000 Genomes project) (**Fig. S9** and **Table S8**). In summary, intersecting candidate functional SNPs with apparent enhancers strengthened the candidacy of five SNPs as being functional, including two SNPs in the 1q32/*IRF6* locus.

MPRA results tested with luciferase reporter assays

To confirm that the MPRA correctly identifies SNPs with allele-specific effects, we performed traditional luciferase reporter assays on 12 SNPs using 701 bp elements centered on the SNP. For these tests we picked six SNPs with allele-specific effects in the MPRA, including the two near *IRF6* and one near *FOXE1* discussed above, and six that did not. Every SNP that had allele-specific effects in the MPRA also had them in the luciferase reporter assays (**Fig. 2C**), although in two cases, rs201265 near *TFAP2A* and rs10984103 near *FOXE1,* the direction of the effect was reversed (**Fig. 2C**). We infer that the shorter elements used in MPRA bind an incomplete set of transcription factors but nonetheless reflect the importance of the SNP. Among the six SNPs that did not have allele-specific effects in the MPRA, four also did not have them in the luciferase assays (**Fig. 2C**). In summary, the luciferase assays confirmed whether enhancer activity was affected by a SNP allele in 10 of 12 cases; this rate of concordance between MPRA and luciferase results is significant ($p = 0.03$, one-sided Fisher's exact test) and matches or exceeds that of other studies using MPRAs to detect functional non-coding SNPs [9, 10].

We conducted RNA-seq on our stock of GMSM-K and discovered that, relative to primary neonatal human epidermal keratinocytes (HEKn), GMSM-K express higher levels of *PITX1* and *FOXE1* transcripts*,* consistent with their origin in the oral cavity, but lower levels of epithelial markers *IRF6*, *CDH1, KRT8*, *KRT18* and *TP63* (**Table S9, Fig. S10A, B**). Therefore, we repeated the luciferase assays in HEKn. At 9 of 11 SNPs tested, the luciferase assays confirmed whether enhancer activity was affected by a SNP allele (**Figs. S11A, 2C, 2D**); one of the SNPs that did not validate was shared with luciferase assays conducted in GMSM-K. Thus, while our aliquot of GMSM-K lacks the robust epithelial features it was reported to have initially [32], the results from an MPRA conducted in GMSM-K largely concorded with those from luciferase reporter studies in primary neonatal keratinocytes.

Because loss-of-function mutations in *IRF6* cause Van der Woude syndrome [62], it was unexpected that elements harboring the risk alleles of the *IRF6* -10 kb or the *IRF6* -22 kb SNPs yielded higher enhancer activity than those harboring the non-risk alleles of these SNPs, both in the MPRA and luciferase assays (**Fig. 2C**). However, this trend was reversed when we conducted the luciferase assays with longer reporter elements, both in HEKn (**Fig. 2D, E**) and in GMSM-K (**Fig. S11B, C**). In summary, luciferase assays support the notion that risk associated alleles of two SNPs near *IRF6* reduce activity of the enhancers in which they reside.

Genome engineering reveals rs11119348 and rs661849 affect *IRF6* expression

To test whether the *IRF6* -10 kb and *IRF6* -22 kb SNPs affect enhancer activity in their native genomic context we engineered the genotype of these SNPs in induced pluripotent stem cells (iPSCs). For each SNP we raised two clones each that were homozygous for the risk allele, homozygous for the non-risk allele, or heterozygous for the risk and non-risk alleles (i.e., parental genotype) (**Fig. S12A, B**). Next, we subjected three replicates of each clone to a 10-day protocol that induces iPSCs to differentiate into embryonic oral epithelium (induced oral epithelium, iOE)[63]. Quantitative RT-PCR revealed that the average level of *IRF6* expression in cells homozygous for the risk allele of either SNP (and heterozygous at the other) was lower than that in clones homozygous for the non-risk allele of that SNP (**Figs. 3A, 4A**), consistent with the reporter assays. We engineered the genotype of these SNPs in GMSM-K as well. Despite the much lower expression of *IRF6* in GMSK-K versus iOE cells, there was the same trend of lower *IRF6* expression in cells homozygous for risk alleles compared to in cell homozygous for non-risk alleles (**Fig. S13 A-F**). These results indicate that the *IRF6* -10 kb SNP and the *IRF6* -22 kb SNP both directly affect *IRF6* expression.

The *IRF6* -10 kb SNP risk allele promotes binding the OFC-associated transcriptional repressor FOXE1

Using online tools [40] we discovered that the risk allele of the *IRF6* -10kb SNP diminished the predicted affinity for two transcription factors (i.e., AR and SOX10) and elevated it for POU5F1B plus sixteen members of the FOX family, including FOXE1 (JASPAR score for FOXE1 site MA1487.1: risk, 14.1; non-risk, 8.6) (**Fig. 3B, Table S10 A, B**). Members of this family generally function as transcriptional repressors[64]. In a published report, chromatin immuno-precipitation with sequencing (ChIP-seq) revealed binding of FOXA1 and FOXA2 in bronchial epithelial cells at the *IRF6* -10 kb enhancer; and knocking down expression of either transcription factor led to upregulation of *IRF6* [65]. However, we found no published reports connecting FOXA1 or FOXA2 to craniofacial defects. By contrast, mutations in *FOXE1* cause Bamforth-Lazarus syndrome [66] and SNPs near it are associated with risk for non-syndromic OFC [20, 57, 67]. Using an antibody against FOXE1 we conducted ChIP in iOE cells. qPCR revealed more chromatin precipitated at the *IRF6* -10 kb SNP locus by anti-FOXE1, and by anti-H3K27Ac, than at an intergenic region, consistent with the SNP being in an enhancer bound by FOXE1 (**Fig. 3C, D**). While the iOE cells were heterozygous at this SNP (**Fig. 3E**), PCR and sequencing revealed that chromatin precipitated by anti-FOXE1 was enriched for the risk allele (**Fig. 3E**, two additional replicates in

Fig. S14A). By contrast, anti-H3K27Ac pulled down more of the non-risk allele (Fig. 3E, S14A), consistent with higher activity of the enhancer and higher expression of *IRF6* in cells with the non-risk genotype. Similarly, anti-FOXE1 ChIP in GMSM-K homozygous for the risk or non-risk allele of *IRF6* -10 kb SNP revealed stronger binding of FOXE1 to the risk allele (Fig. S15A). These results indicate the *IRF6* -10 kb SNP risk allele favors the binding of repressive FOX family members, including FOXE1.

To test the notion that FOXE1 inhibits activity of the *IRF6* -10 kb enhancer we returned to GMSM-K because although they exhibit low *IRF6* levels relative to NHEK, the activity of the *IRF6* -10 kb reporter construct is comparable in both cell types (compare **Figs. 2D** and **S11B,** which use the same scale), and the expression of *FOXE1* is higher in GMSM-K compared to in NHEK (**Fig. S10A** and **Table S9**). We used CRISPR/Cas9 reagents to delete most of the *FOXE1* gene from GMSM-K (hereafter, *FOXE1* knockout); *FOXE1* expression was not detected in these cells (**Fig. S15B**). Interestingly, the activity of the *IRF6* -10 kb reporter was higher in *FOXE1*-knockout than in wild-type GMSM-K (**Fig. 3F**); it was also higher in GMSM-K depleted of *FOXE1* with siRNA than in those transfected with a control siRNA (**Fig S15C, D**). Moreover, levels of *IRF6* expression were higher in *FOXE1*-knockout than in wild type GMSM-K (**Fig. 3G**); it was also higher in GMSM-K depleted of *FOXE1* with siRNA than in those transfected with a control siRNA (**Fig S15E**)**.** Together, these results support the model that the risk allele of the *IRF6* -10 kb SNP results in lower expression of *IRF6* by promoting the binding of transcriptional repressors in the FOX family members, including FOXE1, in cells of the oral cavity (**Fig. 3H**).

The *IRF6* -22 kb SNP risk allele promotes binding of the transcriptional repressor ETS2

Similarly, we discovered that the risk allele of the *IRF6* -22kb SNP diminished the predicted affinity for 2 transcription factors (GATA2 and HSF4) and elevated it for NFKB2, ZNF454, and ten members of the ETS family, including ETS2 (JASPAR score for ETS2 site MA1484.1, risk: 4.9; non-risk, -4.1) **(Fig 4B, Table S11 A, B**). ETS2 can function as a transcriptional repressor [68, 69] and is regulates differentiation of epithelia [70]. Anti-ETS2 and Anti-H3K27Ac antibodies yielded higher ChIP-qPCR signals at the *IRF6* -22 kb SNP than at an intergenic region (**Fig 4C, D**). The iOE cells are heterozygous for risk and non-risk alleles of this SNP, and PCR and sequencing of chromatin precipitated by anti-ETS2 revealed an enrichment of ETS2 binding to the risk allele (**Fig. 4E,** two additional replicates in **Fig. S14B**), while anti-H3K27Ac pulled down more of the non-risk allele (**Fig. 4E, S14B**), consistent with higher activity of the enhancer and higher expression of *IRF6* in cells with the non-risk genotype. Similarly, anti-ETS2 ChIP in

GMSM-K homozygous for the risk or non-risk allele of *IRF6* -22 kb SNP revealed stronger binding of ETS2 to the risk allele (**Fig. S15F**).

We also found that reporter activity of the *IRF6* -22 kb enhancer harboring the risk allele of the *IRF6* -22 kb SNP was higher in ETS2-depleted HEKn cells relative to those transfected with a control siRNA (**Figs. 4F, S15G**); we observed analogous results in GMSM-K (**Fig S15H-J**). *IRF6* expression was also higher in in ETS2-depleted HEKn than in control transfected cells (**Fig. 4G**). Together these results support the model that the risk allele of the *IRF6* -22 kb SNP results in lower expression of *IRF6* by promoting the binding of ETS2, and potentially other transcriptional repressors predicted to bind here (**Fig. 4H) (Table S11)**.

Conditional analysis provides additional evidence for both SNPs at *IRF6* as functional variants

Finally, to assess whether the *IRF6* -10 kb and *IRF6* -22 kb SNPs can explain the heritable risk for OFC associated with 1p22 we conducted conditional analyses by comparing cleft lip only (CL) and cleft lip and palate (CLP) cases separately to unrelated controls (we lacked sufficient cleft-palate-only cases to do a conditional analysis) (**Fig. 5**). As expected, the *IRF6* locus was associated with CL (lead SNP *P*-value = 3.29×10-9) (**Fig. 5A**) and CLP (lead SNP *P*-value = 5.24×10-10) (**Fig. 5E**). When compared to the unconditioned analysis, both rs661849 and rs11119348 explained some of the signal at the *IRF6* locus. However, rs661849 explained more signal in CL cases (lead SNP *P*-value in CL = 7.48×10-6; lead SNP *P*-value in CLP = 3.34×10-8) (**Fig. 5B, F**) and rs11119348 explained more signal in CLP cases (lead SNP *P* -value in CL = 5.05×10-6; lead SNP *P*-value in CLP = 0.0001) (**Fig. 5C, G**). Analyses conditioned on both SNPs explained most of the signal at *IRF6*, especially in CLP cases (lead SNP *P*-value in CL = 0.0003; lead SNP *P*-value in CLP = 0.0001) (**Fig. 5D, H**). In summary, *IRF6* -10 kb (rs11119348) and *IRF6* -22 kb (rs661849) SNPs appear to explain most of the attributable risk for OFC due to variation at this locus within ethnic groups included in the meta-analysis.

Discussion

We sought to identify the functional subset of SNPs associated with OFC to understand how common variants predispose individuals to orofacial cleft (OFC). We first re-performed a meta-analysis on several GWAS of OFC and identified all SNPs with p-values suggestive of association with OFC ($p < 10^{-4}$). We tested 889 such SNPs in a massively parallel reporter assay (MPRA) as has been done in other studies seeking to identify SNPs that affect enhancer

activity [10, 71, 72]. Because enhancers are often cell-type specific, we limited the study to SNPs from loci where the presumed OFC risk-relevant gene is expressed in oral epithelium (as opposed to oral mesenchyme or brain) and used a cell line model of this tissue for the assay. We filtered SNPs with allele-specific effects in the MPRA for histone-mark evidence of being enhancers in primary epidermal keratinocytes or in human embryonic faces. At four of the eight loci, we found SNPs that passed both of these tests of being functional. For the final tests of two candidate SNPs we switched to embryonic oral epithelium differentiated from iPSC, which is currently the best available model of this cell type that can be experimentally manipulated. We found that both SNPs had allele-specific effects on *IRF6* expression levels, supporting the notion that these are bona fide functional SNPs. We identified transcription factors whose binding is affected by the SNP allele, and showed that they regulate the expression of *IRF6*, and of enhancers harboring the SNPs, in the predicted direction. Finally, we returned to the GWAS data and found that the two SNPs explain most of the OFC association signal at this locus in the populations included in the GWAS.

Identifying functional SNPs can illuminate transcriptional regulatory networks (TRNs) that are connected to OFC risk. For instance, analysis of transcription factor binding sites affected by the SNPs pointed to possible transcription factor families affected by each, and the *IRF6* -10 kb SNP affects a FOX binding site. We found that binding of FOXE1 is affected by the allele of this SNP; that of FOXA1 and FOXA2 may also be affected. We showed that deletion of *FOXE1* elevates *IRF6* expression, which is consistent with FOXE1 binding to the *IRF6* -10 kb element and thereby dampening *IRF6* expression. *FOXE1* is implicated in risk for OFC by genome wide linkage analysis [20, 56, 57] and GWAS [23], and mutations in it cause Bamforth-Lazarus syndrome that includes OFC [73]. This example illustrates that identifying a functional SNP for OFC can indeed lead one to another OFC risk-relevant gene. Of note, these results support the possibility of a greater-than-additive effect on OFC risk if risk-associated variants near both genes are present in the same individual, but current GWAS data are underpowered to reveal such a gene-by-gene interaction. Similarly, the risk allele of the *IRF6* -22 kb SNP creates a low affinity ETS family binding site, and ChIP-qPCR indicated that the risk allele promotes binding of ETS2. Of note, low affinity ETS binding sites are important for the regulation of SHH in the limb [74]. In mice it is unclear if ETS2 governs craniofacial development because a strong loss-of-function mutation in *ETS2* causes embryonic lethality by E8.5 [75]. *ETS2* is expressed in differentiated keratinocytes and glandular epithelium [70, 76]; it suppresses epithelial differentiation, and may contribute to carcinogenesis [77]. Our findings suggest that *ETS2* and the genes encoding other

transcription factors whose binding is predicted to be affected by the *IRF6* -22 kb SNP are worth exploring as OFC candidate genes.

Interestingly, *IRF6* -10kb SNP is the fourth example of a potentially functional SNP associated with orofacial cleft within the same evolutionarily conserved enhancer of IRF6, MCS9.7 [6, 37, 53, 60, 61, 78]. The first of these to be identified, rs642961, is associated with CL and the risk-associated allele disrupted binding of the transcription factor AP2-α (TFAP2A) in an electrophoretic mobility shift assay [60]. However, luciferase reporter assays using a 540bp test element did not reveal an allele-specific effect of rs642961 on enhancer activity (tested in human foreskin keratinocytes) [60], making it unclear if this SNP is functional. It is possible that the *IRF6* -22kb SNP, which is in linkage disequilibrium with rs642961, is the functional SNP explaining the OFC signal leading investigators to rs642961. Subsequently, a rare, single nucleotide insertion in MCS9.7 was found to segregate with cleft palate and other characteristic symptoms in a Brazilian family with Van der Woude syndrome [61]. In stable transgenic mice, MCS9.7 (from the human genome) drove reporter expression in oral epithelium [78]; in stable transgenic zebrafish (*Tg*(*MCS9.7:gfp*), it drove reporter expression in periderm in the entire embryo including in the oral cavity [6]. We recently found that a SNP in the same enhancer, rs570516915, is associated with CP only in populations from Finland and Estonia and that the risk allele reduces *IRF6* expression level by disrupting the *IRF6* binding site [37]. The current study shows that rs11119348 in the same enhancer affects *IRF6* expression by altering the binding of FOXE1. Taken together, these studies emphasize the extraordinary importance of a single enhancer in genetic predisposition for orofacial cleft, and raise the question of how different mutations in this enhancer elevate risk for cleft lip with or without cleft palate, or cleft palate only.

There remains a need for high-throughput methods to screen functional SNPs in OFC and other common disorders. The MPRA we performed required synthesis of a library of oligos. This step was avoided in an MPRA that used reporter constructs from sheared chromatin from individual of distinct genotypes; this study managed to score more than 50 million SNPs [12]. However, the impact of this large scale MPRA was limited by only being performed in two cell types. The success of MPRAs at identifying functional SNPs relies on the cell line deployed being at least a reasonable model of the embryonic tissue where the SNP affects disease risk in vivo. While results from MPRA in GMSM-K helped us identify SNPs that affected *IRF6* expression in induced oral epithelium cells, candidate SNPs from the other loci will require additional follow up experiments to confirm whether they are functional. The step of engineering the genome in iPSCs and creating iOE cells *in vitro* remains laborious and expensive. One promising approach

to prioritize non-coding SNPs is chromatin-accessibility QTL analyses [79] but these too require appropriate cell line models. Despite these challenges, identifying functional SNPs remains an essential step in translating the findings of GWAS into an understanding of how common genetic variants contribute to pathogenesis of OFC and other congenital anomalies.

Supplemental information: Eleven supplementary tables provided as Excel sheets and supplementary Figs. (1-15) provided in one PDF document.

Acknowledgments:

The authors thank Pilar Santisteban, Universidad Autónoma de Madrid, for the FOXE1 antibody, members of the Craniofacial Interest Group at the University of Iowa for helpful comments, and Greg Bonde (University of Iowa) and Josh Rosswork (University of Washington) for technical assistance throughout the project. We thank Ellison Stem Cell Core at ISCRM and Jennifer Hesson for technical assistance in screening and genotyping of iPSCs. We are grateful for many discussions with members of the Craniofacial Interest Group weekly meeting hosted at the University of Iowa.

This project was funded by grants from the U.S. National Institutes of Health, DE027362 (R.A.C.), R01DE028599 (A. V.), R01DE033016 (R.A.C., H.R.B., J.M.) HG010855 (K. B.) and CA246503 (K. B.). Research conducted at the E.O. Lawrence Berkeley National Laboratory was performed under Department of Energy Contract DE-AC02-05CH11231, University of California.

Competing interests: KB is co-founder of Matchstick Technologies, Inc. KB is a co-inventor of PIXUL (US Patents 10809166, 11592366).

References:

1. Stanier P, Moore GE. Genetics of cleft lip and palate: syndromic genes contribute to the incidence of non-syndromic clefts. Human molecular genetics. 2004;13(suppl_1):R73-R81.

2. Klotz CM, Wang X, DeSensi RS, Grubs RE, Costello BJ, Marazita ML. Revisiting the recurrence risk of nonsyndromic cleft lip with or without cleft palate. American Journal of Medical Genetics Part A. 2010;152(11):2697-702.

3. Sivertsen Å, Wilcox AJ, Skjærven R, Vindenes HA, Åbyholm F, Harville E, Lie RT. Familial risk of oral clefts by morphological type and severity: population based cohort study of first degree relatives. Bmj. 2008;336(7641):432-4.

4. Grosen D, Chevrier C, Skytthe A, Bille C, Mølsted K, Sivertsen Å, Murray JC, Christensen K. A cohort study of recurrence patterns among more than 54 000 relatives of oral cleft cases in Denmark: support for the multifactorial threshold model of inheritance. Journal of medical genetics. 2010;47(3):162-8.

5. Nasreddine G, El Hajj J, Ghassibe-Sabbagh M. Orofacial clefts embryology, classification, epidemiology, and genetics. Mutation Research/Reviews in Mutation Research. 2021:108373.

6. Liu H, Duncan K, Helverson A, Kumari P, Mumm C, Xiao Y, Carlson JC, Darbellay F, Visel A, Leslie E. Analysis of zebrafish periderm enhancers facilitates identification of a regulatory variant near human KRT8/18. Elife. 2020;9:e51325.

7. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S. Annotation of functional variation in personal genomes using RegulomeDB. Genome research. 2012;22(9):1790-7.

8. Inoue F, Kircher M, Martin B, Cooper GM, Witten DM, McManus MT, Ahituv N, Shendure J. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. Genome research. 2017;27(1):38-52.

9. Klein JC, Keith A, Rice SJ, Shepherd C, Agarwal V, Loughlin J, Shendure J. Functional testing of thousands of osteoarthritis-associated variants for regulatory activity. Nature communications. 2019;10(1):2434.

10. Abell NS, DeGorter MK, Gloudemans MJ, Greenwald E, Smith KS, He Z, Montgomery SB. Multiple causal variants underlie genetic associations in humans. Science. 2022;375(6586):1247-54.

11. Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, Zhang X, Alston J, Mikkelsen TS, Kellis M. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. Genome research. 2013;23(5):800-11.

12. van Arensbergen J, Pagie L, FitzPatrick VD, de Haas M, Baltissen MP, Comoglio F, van der Weide RH, Teunissen H, Võsa U, Franke L. High-throughput identification of human SNPs affecting regulatory element activity. Nature genetics. 2019;51(7):1160-9.

13. Vockley CM, Guo C, Majoros WH, Nodzenski M, Scholtens DM, Hayes MG, Lowe WL, Reddy TE. Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. Genome research. 2015;25(8):1206-14.

14. Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK, Andersen KG, Mikkelsen TS, Lander ES, Schaffner SF. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. Cell. 2016;165(6):1519-29.

15. Ulirsch JC, Nandakumar SK, Wang L, Giani FC, Zhang X, Rogov P, Melnikov A, McDonel P, Do R, Mikkelsen TS. Systematic functional dissection of common genetic variation affecting red blood cell traits. Cell. 2016;165(6):1530-45.

16. Liu S, Liu Y, Zhang Q, Wu J, Liang J, Yu S, Wei G-H, White KP, Wang X. Systematic identification of regulatory variants associated with cancer risk. Genome biology. 2017;18:1-14.

17. Zhang P, Xia J-H, Zhu J, Gao P, Tian Y-J, Du M, Guo Y-C, Suleman S, Zhang Q, Kohli M. High-throughput screening of prostate cancer risk loci by single nucleotide polymorphisms sequencing. Nature communications. 2018;9(1):2022.

18. Butali A, Suzuki S, Cooper ME, Mansilla AM, Cuenco K, Leslie EJ, Suzuki Y, Niimi T, Yamamoto M, Ayanga G. Replication of Genome Wide Association Identified Candidate Genes Confirm the Role of Common and Rare Variants in PAX 7 and VAX 1 in the Etiology of Nonsyndromic CL (P). American Journal of Medical Genetics Part A. 2013;161(5):965-72.

19. Camargo M, Rivera D, Moreno L, Lidral AC, Harper U, Jones M, Solomon BD, Roessler E, Vélez JI, Martinez AF. GWAS reveals new recessive loci associated with non-syndromic facial clefting. European journal of medical genetics. 2012;55(10):510-4.

20. Leslie EJ, Carlson JC, Shaffer JR, Butali A, Buxó CJ, Castilla EE, Christensen K, Deleyiannis FW, Field LL, Hecht JT. 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. Human genetics. 2017;136(3):275-86.

21. Leslie EJ, Carlson JC, Shaffer JR, Feingold E, Wehby G, Laurie CA, Jain D, Laurie CC, Doheny KF, McHenry T. A multi-ethnic genome-wide association study identifies novel loci for non-syndromic cleft lip with or without cleft palate on 2p24. 2, 17q23 and 19q13. Human molecular genetics. 2016;25(13):2862-72.

22. Sun Y, Huang Y, Yin A, Pan Y, Wang Y, Wang C, Du Y, Wang M, Lan F, Hu Z. Genome-wide association study identifies a new susceptibility locus for cleft lip with or without a cleft palate. Nature communications. 2015;6(1):1-7.

23. Yu Y, Zuo X, He M, Gao J, Fu Y, Qin C, Meng L, Wang W, Song Y, Cheng Y. Genome-wide analyses of non-syndromic cleft lip with palate identify 14 novel loci and genetic heterogeneity. Nature communications. 2017;8(1):1-11.

24. Birnbaum S, Ludwig KU, Reutter H, Herms S, Steffens M, Rubini M, Baluardo C, Ferrian M, de Assis NA, Alblas MA. Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24. Nature genetics. 2009;41(4):473-7.

25. Rojas-Martinez A, Reutter H, Chacon-Camacho O, Leon-Cachon RB, Munoz-Jimenez SG, Nowak S, Becker J, Herberz R, Ludwig KU, Paredes-Zenteno M. Genetic risk factors for nonsyndromic cleft lip with or without cleft palate in a Mesoamerican population: Evidence for IRF6 and variants at 8q24 and 10q25. Birth Defects Research Part A: Clinical and Molecular Teratology. 2010;88(7):535-7.

26. Ludwig KU, Mangold E, Herms S, Nowak S, Reutter H, Paul A, Becker J, Herberz R, AlChawa T, Nasser E. Genome-wide meta-analyses of nonsyndromic cleft lip with or without cleft palate identify six new risk loci. Nature genetics. 2012;44(9):968-71.

27. Beaty TH, Murray JC, Marazita ML, Munger RG, Ruczinski I, Hetmanski JB, Liang KY, Wu T, Murray T, Fallin MD. A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. Nature genetics. 2010;42(6):525-9.

28. Lee T-H, Liu TT, Sung CW, Ou JC, Yeh J-J, Kuo Y-T, Wang Y-H. Association between IRF6 rs642961 polymorphism and non-syndromic cleft lip with or without a cleft palate: a systematic review and meta-analysis. International Journal of Clinical and Experimental Medicine. 2018:3033-42.

29. Butali A, Mossey P, Adeyemo W, Eshete M, Gaines L, Braimah R, Aregbesola B, Rigdon J, Emeka C, Olutayo J. Rare functional variants in genome–wide association identified candidate genes for nonsyndromic clefts in the African population. American Journal of Medical Genetics Part A. 2014;164(10):2567-71.

30. Zucchero TM, Cooper ME, Maher BS, Daack-Hirsch S, Nepomuceno B, Ribeiro L, Caprau D, Christensen K, Suzuki Y, Machida J. Interferon regulatory factor 6 (IRF6) gene variants and the risk of isolated cleft lip or palate. New England Journal of Medicine. 2004;351(8):769-80.

31. Kwasnieski JC, Fiore C, Chaudhari HG, Cohen BA. High-throughput functional testing of ENCODE segmentation predictions. Genome research. 2014;24(10):1595-602.

32. Gilchrist EP, Moyer MP, Shillitoe EJ, Clare N, Murrah VA. Establishment of a human polyclonal oral epithelial cell line. Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology. 2000;90(3):340-7.

33. Darbro B, Lee K, Nguyen N, Domann F, Klingelhutz A. Methylation of the p16INK4a promoter region in telomerase immortalized human keratinocytes co-cultured with feeder cells. Oncogene. 2006;25(56):7421-33.

34. Rheinwald JG, Green H. Epidermal growth factor and the multiplication of cultured human epidermal keratinocytes. Nature. 1977;265(5593):421-4.

35. Friedman RZ, Granas DM, Myers CA, Corbo JC, Cohen BA, White MA. Information content differentiates enhancers from silencers in mouse photoreceptors. Elife. 2021;10:e67403.

36. Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, Grossman SR, Anyoha R, Doughty BR, Patwardhan TA. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. Nature genetics. 2019;51(12):1664-9.

37. Rahimov F, Nieminen P, Kumari P, Juuri E, Nikopensius T, Karjalainen J, Kurki M, Palotie A, FinnGen N, Heliövaara A. High incidence and regional distribution of cleft palate in Finns are associated with a functional variant in an IRF6 enhancer2021.

38. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2− ΔΔCT method. methods. 2001;25(4):402-8.

39. Taylor SC, Nadeau K, Abbasi M, Lachance C, Nguyen M, Fenrich J. The ultimate qPCR experiment: producing publication quality, reproducible data the first time. Trends in biotechnology. 2019;37(7):761-74.

40. Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Berhanu Lemma R, Turchi L, Blanc-Mathieu R, Lucas J, Boddie P, Khan A, Manosalva Pérez N. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic acids research. 2022;50(D1):D165-D73.

41. Bomsztyk K, Mar D, Wang Y, Denisenko O, Ware C, Frazar CD, Blattler A, Maxwell AD, MacConaghy BE, Matula TJ. PIXUL-ChIP: integrated high-throughput sample preparation and analytical platform for epigenetic studies. Nucleic Acids Research. 2019;47(12):e69-e.

42. Morillo-Bernal J, Fernández LP, Santisteban P. FOXE1 regulates migration and invasion in thyroid cancer cells and targets ZEB1. Endocrine-related cancer. 2020;27(3):137-51.

43. Solomon ER, Caldwell KK, Allan AM. A novel method for the normalization of ChIP-qPCR data. MethodsX. 2021;8:101504.

44. Delaneau O, Zagury J-F, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. Nature methods. 2013;10(1):5-6.

45. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS genetics. 2009;5(6):e1000529.

46. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4(1):s13742-015-0047-8.

47. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR, Willer CJ. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010;26(18):2336-7.

48. Alexander TA, Machiela MJ. LDpop: an interactive online tool to calculate and visualize geographic LD patterns. BMC bioinformatics. 2020;21(1):1-4.

49. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics. 2015;31(21):3555-7.

50. Knight AS, Schutte BC, Jiang R, Dixon MJ. Developmental expression analysis of the mouse and chick orthologues of IRF6: the gene mutated in Van der Woude syndrome. Developmental dynamics: an official publication of the American Association of Anatomists. 2006;235(5):1441-7.

51. Li M-X, Li Z, Zhang R, Yu Y, Wang L-S, Wang Q, Ding Z, Zhang J-P, Zhang M-R, Xu L-C. Effects of small interfering RNA-mediated silencing of susceptibility genes of non-syndromic cleft lip with or without cleft palate on cell proliferation and migration. International Journal of Pediatric Otorhinolaryngology. 2020;138:110382.

52. Laurikkala J, Mikkola ML, James M, Tummers M, Mills AA, Thesleff I. p63 regulates multiple signalling pathways required for ectodermal organogenesis and differentiation2006.

53. Kousa YA, Fuller E, Schutte BC. IRF6 and AP2A interaction regulates epidermal development. Journal of Investigative Dermatology. 2018;138(12):2578-88.

54. Lu Y, Liang M, Zhang Q, Liu Z, Song Y, Lai L, Li Z. Mutations of GADD45G in rabbits cause cleft lip by the disorder of proliferation, apoptosis and epithelial-mesenchymal transition (EMT). Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease. 2019;1865(9):2356-67.

55. Dixon M, Robinson V, White A, Ferguson M. Monoclonal antibodies recognising stage and region specific epitopes in embryonic mouse palatal epithelial cells. Journal of anatomy. 1993;183(Pt 2):423.

56. Marazita ML, Lidral AC, Murray JC, Field LL, Maher BS, Goldstein McHenry T, Cooper ME, Govil M, Daack-Hirsch S, Riley B. Genome scan, fine-mapping, and candidate gene analysis of non-syndromic cleft lip with or without cleft palate reveals phenotype-specific differences in linkage and association results. Human heredity. 2009;68(3):151-70.

57. Moreno LM, Mansilla MA, Bullard SA, Cooper ME, Busch TD, Machida J, Johnson MK, Brauer D, Krahn K, Daack-Hirsch S. FOXE1 association with both isolated cleft lip with or without cleft palate, and isolated cleft palate. Human molecular genetics. 2009;18(24):4879-96.

58. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57.

59. Wilderman A, VanOudenhove J, Kron J, Noonan JP, Cotney J. High-resolution epigenomic atlas of human embryonic craniofacial development. Cell reports. 2018;23(5):1581-97.

60. Rahimov F, Marazita ML, Visel A, Cooper ME, Hitchler MJ, Rubini M, Domann FE, Govil M, Christensen K, Bille C. Disruption of an AP-2α binding site in an IRF6 enhancer is associated with cleft lip. Nature genetics. 2008;40(11):1341-7.

61. Fakhouri WD, Rahimov F, Attanasio C, Kouwenhoven EN, Ferreira De Lima RL, Felix TM, Nitschke L, Huver D, Barrons J, Kousa YA. An etiologic regulatory mutation in IRF6 with loss-and gain-of-function effects. Human molecular genetics. 2014;23(10):2711-20.

62. Kondo S, Schutte BC, Richardson RJ, Bjork BC, Knight AS, Watanabe Y, Howard E, Ferreira de Lima RL, Daack-Hirsch S, Sander A. Mutations in IRF6 cause Van der Woude and popliteal pterygium syndromes. Nature genetics. 2002;32(2):285-9.

63. Alghadeer A, Hanson-Drury S, Patni AP, Ehnes DD, Zhao YT, Li Z, Phal A, Vincent T, Lim YC, O'Day D. Single-cell census of human tooth development enables generation of human enamel. Developmental Cell. 2023;58(20):2163-80. e9.

64. Katoh M, Igarashi M, Fukuda H, Nakagama H, Katoh M. Cancer genetics and genomics of human FOX family genes. Cancer letters. 2013;328(2):198-206.

65. Paranjapye A, Mutolo MJ, Ebron JS, Leir S-H, Harris A. The FOXA1 transcriptional network coordinates key functions of primary human airway epithelial cells. American Journal of Physiology-Lung Cellular and Molecular Physiology. 2020;319(1):L126-L36.

66. Bamforth J, Hughes I, Lazarus J, Weaver C, Harper P. Congenital hypothyroidism, spiky hair, and cleft palate. Journal of medical genetics. 1989;26(1):49.

67. Ludwig K, Böhmer A, Rubini M, Mossey P, Herms S, Nowak S, Reutter H, Alblas M, Lippke B, Barth S. Strong association of variants around FOXE1 and orofacial clefting. Journal of Dental Research. 2014;93(4):376-81.

68. Panagoulias I, Karagiannis F, Aggeletopoulou I, Georgakopoulos T, Argyropoulos CP, Akinosoglou K, Gogos C, Skoutelis A, Mouzaki A. ets-2 acts as a Transcriptional repressor of the human immunodeficiency Virus Type 1 through Binding to a repressor–activator Target sequence of 5′-lTr. Frontiers in Immunology. 2018;8:1924.

69. Wei G, Schaffner AE, Baker KM, Mansky KC, Ostrowski MC. Ets-2 interacts with co-repressor BS69 to repress target gene expression. Anticancer research. 2003;23(3):2173-8.

70. Oettgen P, Kas K, Dube A, Gu X, Grall F, Thamrongsak U, Akbarali Y, Finger E, Boltax J, Endress G. Characterization of ESE-2, a novel ESE-1-related Ets transcription factor that is restricted to glandular epithelium and differentiated keratinocytes. Journal of Biological Chemistry. 1999;274(41):29439-52.

71. Klein JC, Keith A, Rice SJ, Shepherd C, Agarwal V, Loughlin J, Shendure J. Functional testing of thousands of osteoarthritis-associated variants for regulatory activity. Nature communications. 2019;10(1):1-9.

72. Ajore R, Niroula A, Pertesi M, Cafaro C, Thodberg M, Went M, Bao EL, Duran-Lozano L, Lopez de Lapuente Portilla A, Olafsdottir T. Functional dissection of inherited non-coding variation influencing multiple myeloma risk. Nature communications. 2022;13(1):1-15.

73. Clifton-Bligh RJ, Wentworth JM, Heinz P, Crisp MS, John R, Lazarus JH, Ludgate M, Chatterjee VK. Mutation of the gene encoding human TTF-2 associated with thyroid agenesis, cleft palate and choanal atresia. Nature genetics. 1998;19(4):399-401.

74. Lim F, Solvason JJ, Ryan GE, Le SH, Jindal GA, Steffen P, Jandu SK, Farley EK. Affinity-optimizing enhancer variants disrupt development. Nature. 2024;626(7997):151-9.

75. Yamamoto H, Flannery ML, Kupriyanov S, Pearce J, McKercher SR, Henkel GW, Maki RA, Werb Z, Oshima RG. Defective trophoblast function in mice with a targeted mutation of Ets2. Genes & development. 1998;12(9):1315-26.

76. Andreoli JM, Jang S-I, Chung E, Coticchia CM, Steinert PM, Markova NG. The expression of a novel, epithelium-specific ets transcription factor is restricted to the most differentiated layers in the epidermis. Nucleic acids research. 1997;25(21):4287-95.

77. Feldman RJ, Sementchenko VI, Watson DK. The epithelial-specific Ets factors occupy a unique position in defining epithelial proliferation, differentiation and carcinogenesis. Anticancer research. 2003;23(3):2125-32.

78. Fakhouri WD, Rhea L, Du T, Sweezer E, Morrison H, Fitzpatrick D, Yang B, Dunnwald M, Schutte BC. MCS9. 7 enhancer activity is highly, but not completely, associated with expression of Irf6 and p63. Developmental dynamics. 2012;241(2):340-9.

79. Tehranchi A, Hie B, Dacre M, Kaplow I, Pettie K, Combs P, Fraser HB. Fine-mapping cis-regulatory variants in diverse human populations. Elife. 2019;8:e39595.

Table 1: The number of SNPs tested, and the number with allele-specific effects, in the MPRA at each locus

Chr	**Locus**	**No. of SNPs** (889 total)	**MPRA-nominated SNPs** (65 total)
1	*1q32/IRF6*	608	46
2	*2p21/THADA*	40	1
3	*3q28/TP63*	14	1
6	*6p24.3/TFAP2A*	112	8
9	9q22.2/*GADD45G*	36	2
9	9q22.33/*FOXE1*	11	1
12	12q13.13/*KRT18*	14	2
20	20q12/*MAFB*	54	4

Figures and figure legends

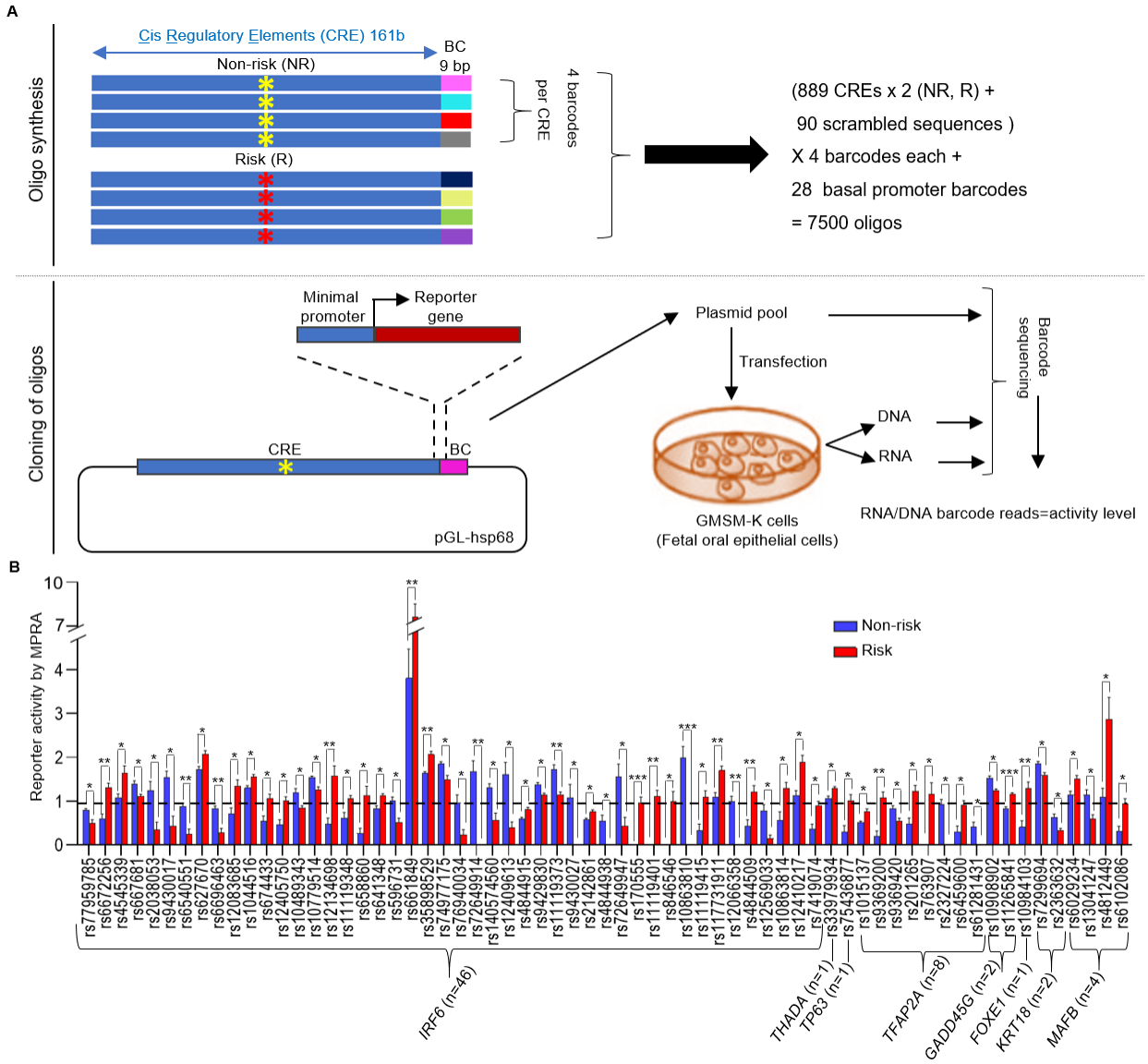

Fig. 1: MPRA identified SNPs with allele-specific significant effects on reporter activity.

A) MPRA schematic, design, cloning and transfection of oligos library. BC, barcode. Yellow and red asterisks, non-risk and risk allele respectively. **B)** Bar chart showing SNPs with significant allele-specific effects on reporter activity MPRA identified. The number of SNPs at each locus is shown in brackets. Dotted line represents the mean of reporter activity of 90 scrambled elements. Data are represented as mean values +/- SEM. Statistical significance is determined by Welch's t-test (*P<0.05, ** P<0.01, *** P<0.001)**;** NS, not significant.

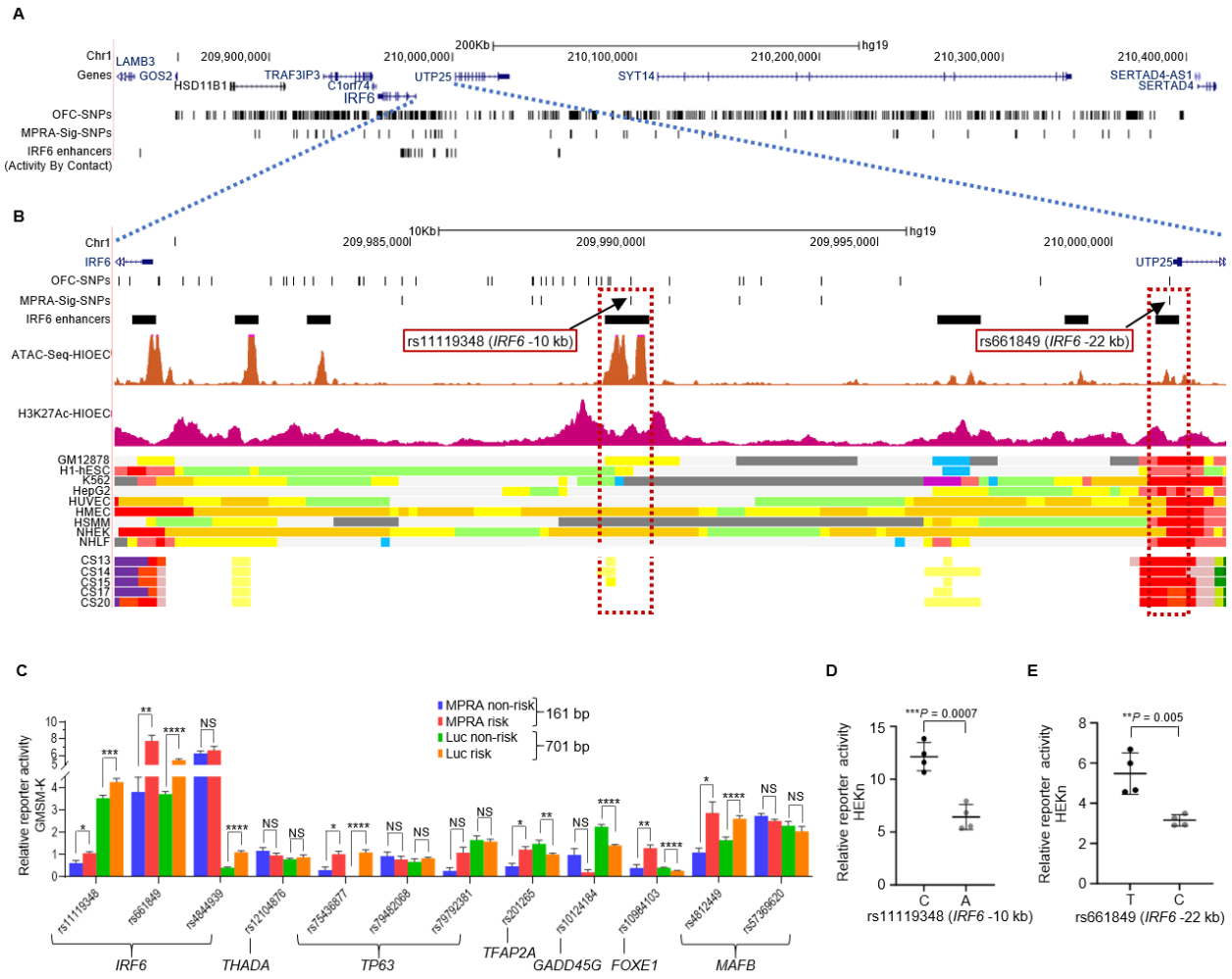

Fig. 2: Two SNPs (rs11119348 and rs661849) at *IRF6* as regulatory variants

A) Browser view of the human genome, GRCh37/hg19, focused on the GWAS-identified SNPs at *IRF6* associated with orofacial cleft. First two tracks represent the total SNPs (608) identified by GWAS at *IRF6* and MPRA-identified SNPs (46) at *IRF6* respectively. Next track represents IRF6 enhancers revealed by the ABC (Activity by Contact) method. **B)** Browser view (zoomed in from Figure 2A) of the human genome, GRCh37/hg19, focused on the SNPs at *IRF6* (rs11119348 and rs661849) that lie in active elements. First three tracks represent the same as in Fig. 2A. Next two tracks represent open chromatin elements detected by ATAC-Seq and active chromatin revealed by H3K27Ac in HIOEC, respectively, illustrating rs11119348 and rs661849 lie within *IRF6* enhancers. Following color coded bars represent chromatin status revealed by ChIP-Seq to various chromatin marks from the ENCODE Project cell lines and facial explants from human embryos at Carnegie stage (CS) 13-20, encompassing the time when palatal shelves fuse, where orange and yellow bars represent the active and weak

enhancer element, respectively; bright red, active promoter; light red, weak promoter; purple, inactive/poised promoter; blue, insulator; light green, weak transcribed; grey, Polycomb repressed; light grey, heterochromatin/repetitive; green and dark purple in facial explants dataset, transcription and bivalent promoter respectively; GM12878, B-cell derived cell line; ESC, embryonic stem cells; K562, myelogenous leukemia cell line; HepG2, liver cancer cell line; 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, CS13-CS20 are facial explants from human embryos. **C)** Bar chart of MPRA and luciferase reporter activities for the indicated SNPs at various loci in GMSM-K. For luciferase reporter activities, data are represented as mean values +/- s. d. from four independent experiments. Statistical significance is determined by Student's t-test, two-tailed *P* value (**P*<0.05, ***P*<0.01, ****P*<0.001, *****P*<0.0001) and NS, non-significant; Luc, luciferase. **D** and **E)** Scattered dot plot of relative luciferase activity for non-risk and risk alleles of rs11119348 and rs661849 respectively in HEKn. Data are represented as mean values +/- s. d. from four independent experiments. Statistical significance is determined by Student's *t*-test. *P* value (two-tailed) is indicated on the plot.

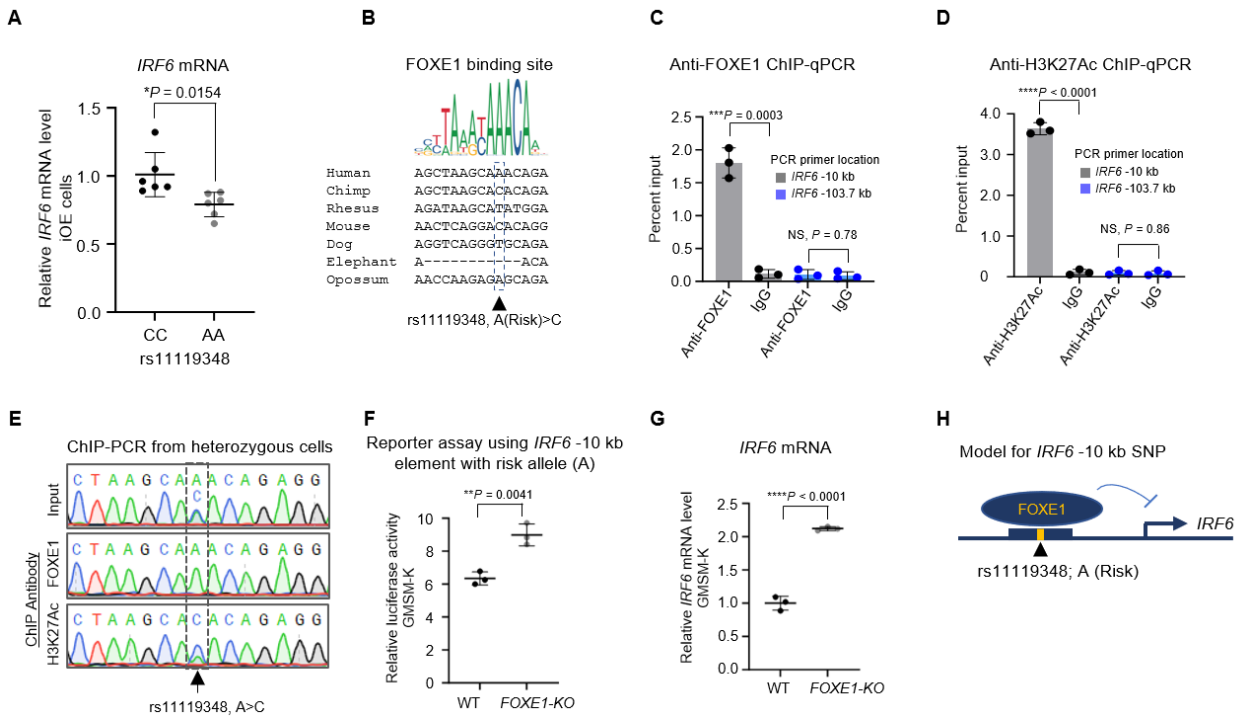

Fig. 3: Risk allele of *IRF6* -10 kb SNP (rs11119348) reduces the *IRF6* expression by binding to the FOXE1 repressor. A) Scattered dot plot of relative levels of *IRF6* mRNA in edited iOE cells harboring non-risk (CC) and risk (AA) genotype of rs11119348 assessed by qRT-PCR. Expression levels of *IRF6* are normalized against *ACTB, GAPDH, HPRT, UBC* and *CDH1*. Data are represented as mean values +/- s. d. from six replicates of cells harboring each genotype, as indicated in the plot. Statistical significance is determined by Student's *t*-test (two-tailed). **B)** Consensus FOXE1 binding motif from the JASPAR database of transcription factor DNA-binding preferences (Matrix ID: MA1847.1) and alignment of the variant site in different species. At this SNP, the risk allele A has a higher frequency than the non-risk allele, C, in most populations. **C** and **D)** Percent input identified by ChIP-qPCR for anti-FOXE1 and anti-H3K27Ac respectively in iOE cells heterozygous for rs11119348 using primers specific to the *IRF6* -10 kb enhancer site or, as a negative control, to a region 103.7 kb upstream of the *IRF6* transcription start site that did not harbor active elements identified from ATAC-Seq and H3K27Ac ChIP-Seq in HIOEC or NHEK. Error bars refer to three ChIP replicates and expressed as mean values +/- s. d. Statistical significance is determined by Student's *t*-test. *P* value (two-tailed) is indicated on the plot and NS, non-significant. **E)** Sequencing of anti-FOXE1 and anti-H3K27Ac ChIP-PCR product of cells heterozygous for rs11119348. **F)** Scattered dot plot of relative luciferase activity

using risk allele harboring element in wildtype (WT) and *FOXE1*-KO GMSM-K. Data are represented as mean values +/- s. d. from three independent experiments. Statistical significance is determined by Student's *t*-test. *P* value (two-tailed) is indicated on the plot. **G)** Scattered dot plot of relative levels of *IRF6* mRNA in WT and *FOXE1*-KO GMSM-K assessed by qRT-PCR. Expression levels of *IRF6* are normalized against *ACTB*. Data are represented as mean values +/- s. d. from three replicates. Statistical significance is determined by Student's *t*-test. *P* value (two-tailed) is indicated on the plot. **H)** Model for the risk allele of rs11119348 meditated binding of FOXE1 repressor to the *IRF6* -10 kb enhancer leading to the reduced *IRF6* expression.

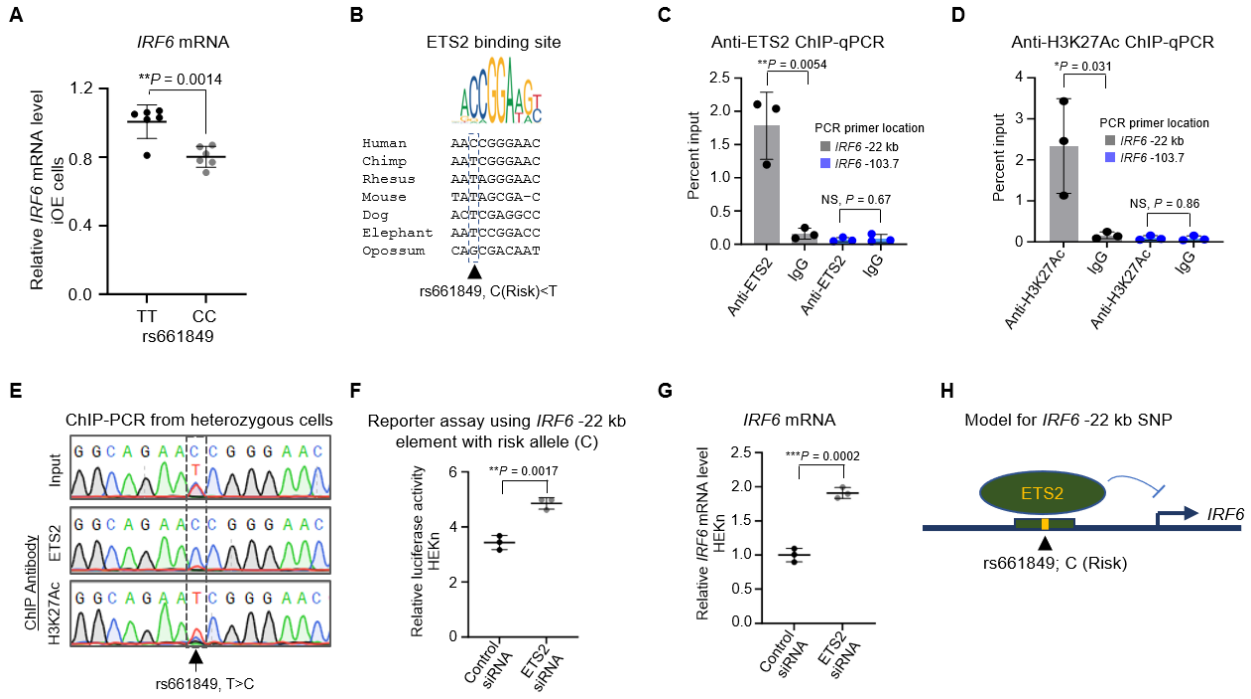

Fig. 4: Risk allele of *IRF6* -22 kb SNP (rs661849) reduces the *IRF6* expression by binding to the ETS2 repressor. A) Scattered dot plot of relative levels of *IRF6* mRNA in edited iOE cells harboring non-risk (CC) and risk (AA) genotype of rs661849 assessed by qRT-PCR. Expression levels of *IRF6* are normalized against *ACTB, GAPDH, HPRT, UBC* and *CDH1*. Data are represented as mean values +/- s. d. from six replicates of cells harboring each genotype, as indicated in the plot. Statistical significance is determined by Student's *t*-test (two-tailed). **B)** Consensus ETS2 binding motif from the JASPAR database of transcription factor DNA-binding preferences (Matrix ID: MA1484.1) and alignment of the variant site in different species. At this SNP, the risk allele C has a lower frequency than the non-risk allele, T, in most populations. Even the risk allele is an imperfect match for the ETS2 binding site, which accounts for its low score. **C** and **D)** Percent input identified by ChIP-qPCR for anti-ETS2 and anti-H3K27Ac respectively in iOE cells heterozygous for rs661849 using primers specific to the *IRF6* -22 kb enhancer site or, as a negative control, to a region 103.7 kb upstream *IRF6* transcription start site that did not harbor active elements identified from ATAC-Seq and H3K27Ac ChIP-Seq in HIOEC or NHEK. Error bars refer to three ChIP replicates and expressed as mean values +/- s. d. Statistical significance is determined by Student's *t*-test. *P* value (two-tailed) is indicated on the plot and NS, non-significant. **E)** Sequencing of anti-ETS2 and anti-H3K27Ac ChIP-PCR product of cells heterozygous for rs661849. **F)** Scattered dot plot of relative luciferase activity using risk allele harboring element in control vs ETS2 depleted HEKn. Data are represented as

mean values +/- s. d. from three independent experiments. Statistical significance is determined by Student's *t*-test. *P* value (two-tailed) is indicated on the plot. **G)** Scattered dot plot of relative levels of *IRF6* mRNA in control vs ETS2-depleted HEKn assessed by qRT-PCR. Expression levels of *IRF6* are normalized against *ACTB*. Data are represented as mean values +/- s. d. from three replicates. Statistical significance is determined by Student's *t*-test. *P* value (two-tailed) is indicated on the plot. **H)** Model for the risk allele of rs661849 meditated binding of ETS2 repressor to the *IRF6* -22 kb enhancer leading to the reduced *IRF6* expression.

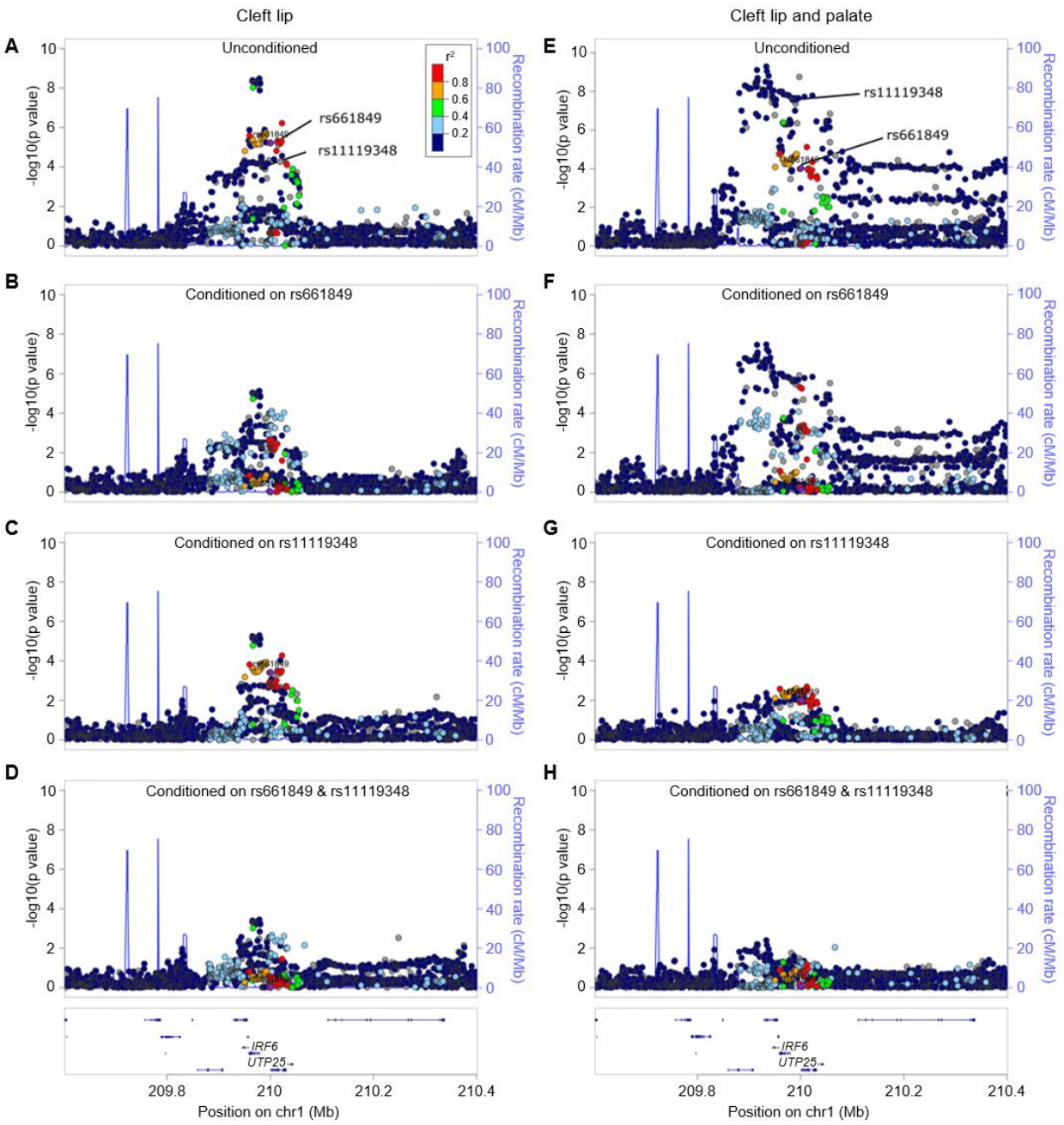

Fig. 5: Conditional analysis supports rs11119348 and rs661849 as functional variants.

Locus Zoom plot of cleft lip and cleft lip and palate respectively; unconditioned (**A, E**), conditioned on rs661849 (**B, F**), conditioned on rs1119348 (**C, G**) and conditioned on both SNPs together (**D, H**). Points are color-coded based on linkage disequilibrium (r^2) in Europeans.

Supplementary Figures

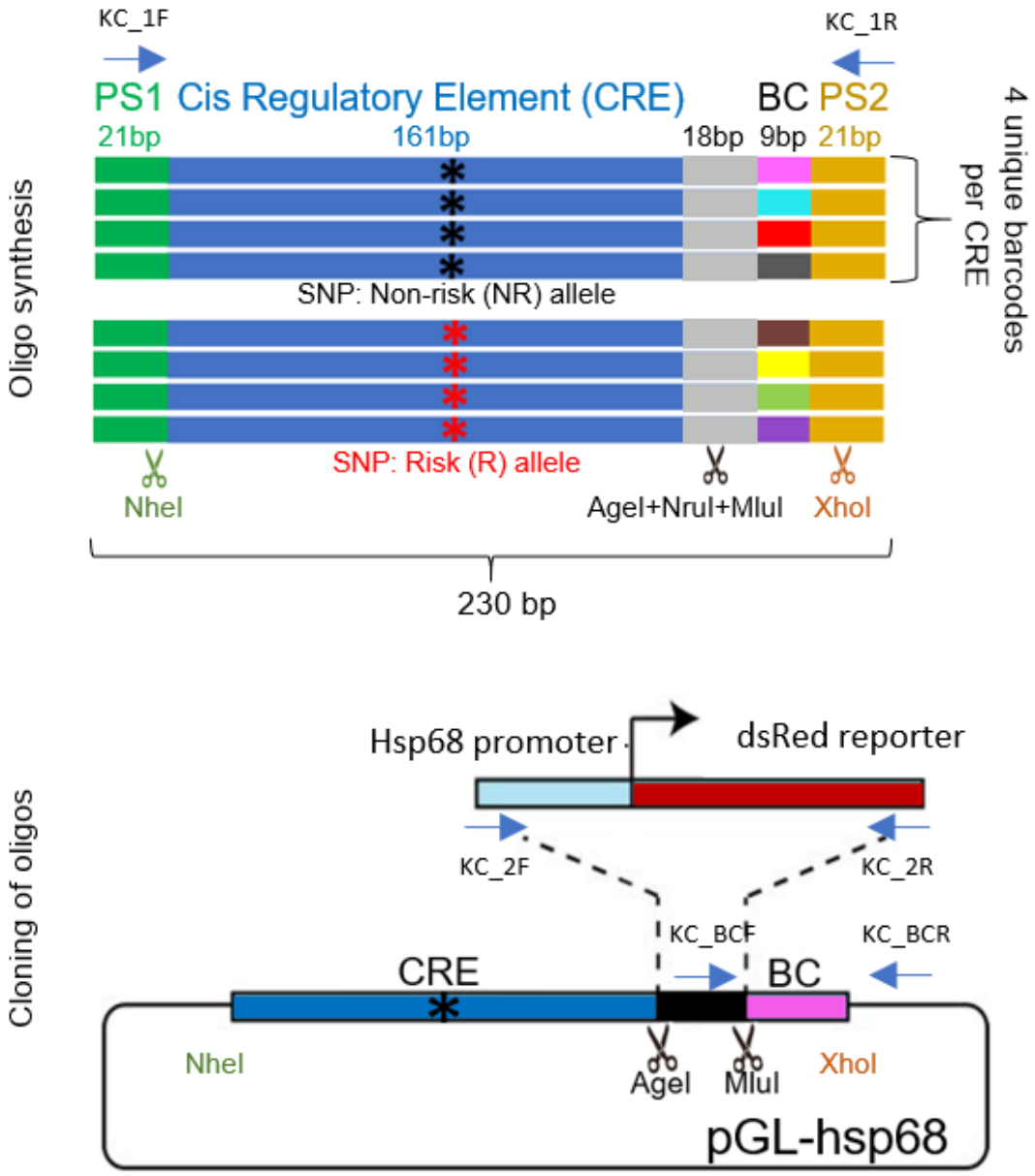

Fig. S1: Details of MPRA library construction. PS represents priming sequences; BC represents barcode. KC represents different set of primers used in the MPRA library process (see method section and table S4 for details).

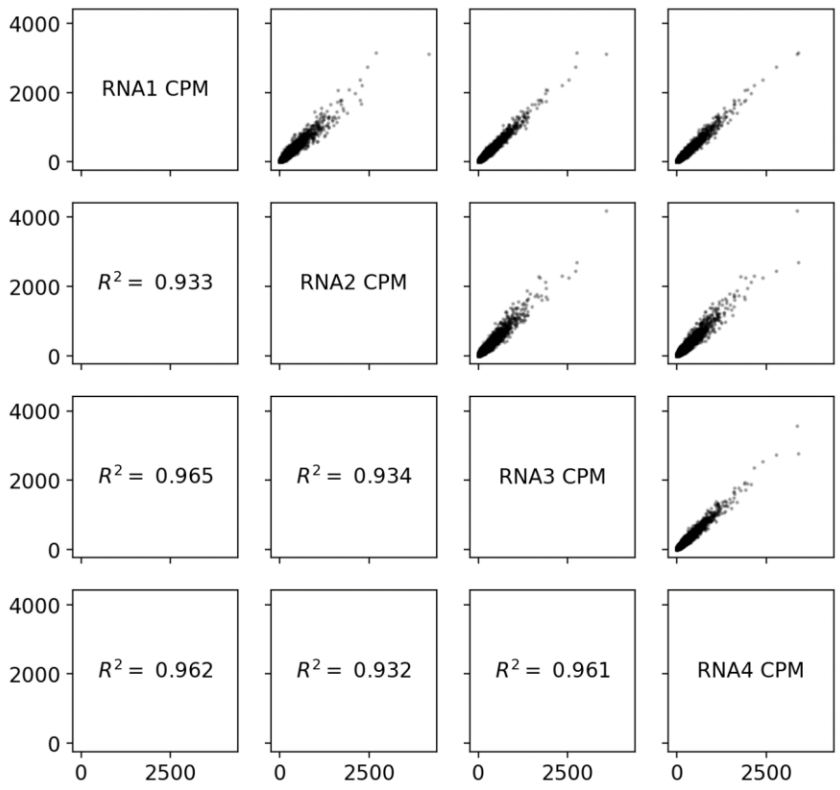

Fig. S2: Correlation plot showing distribution of RNA barcode counts across four RNA replicates.

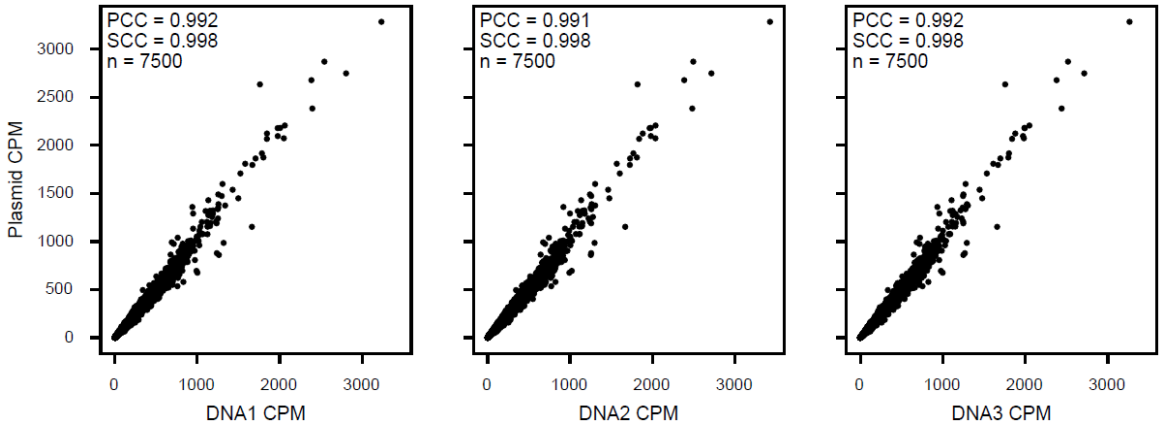

Fig. S3: Correlation plot showing DNA barcode counts (post-transfection) vs plasmid DNA barcode counts (pre-transfection).

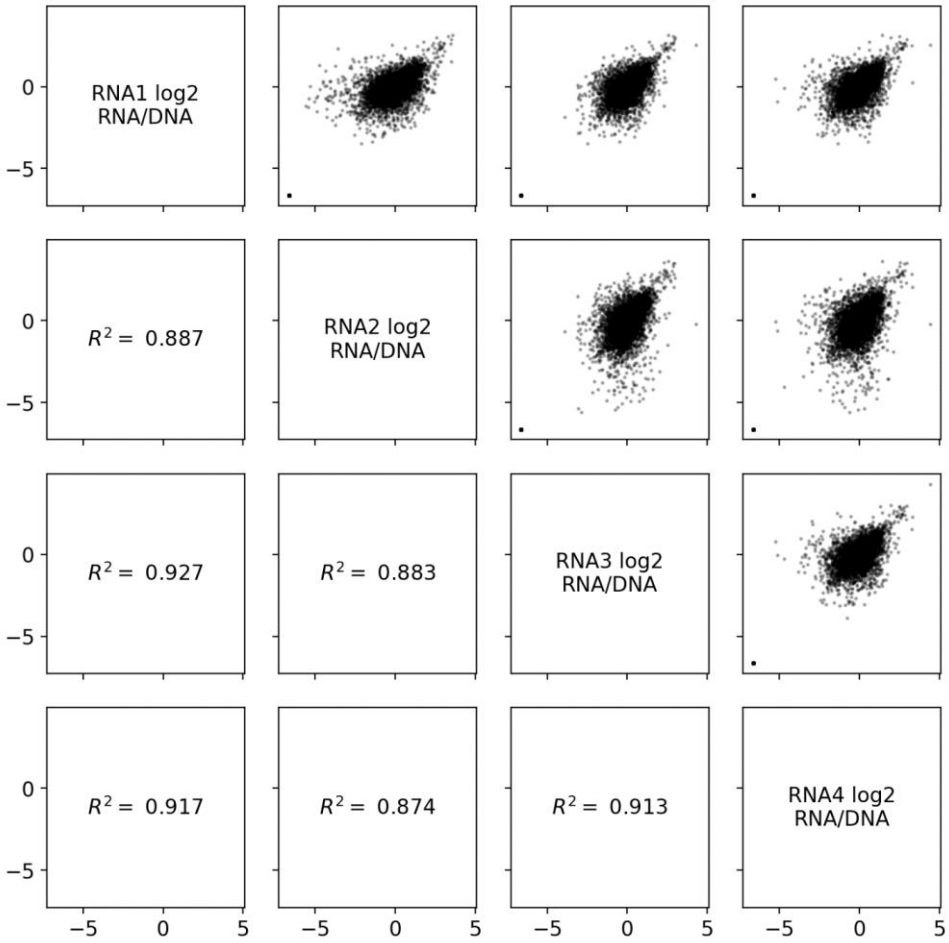

Fig. S4: Correlation plot showing RNA barcode counts normalized to plasmid DNA barcode counts (pre-transfection) across four replicates.

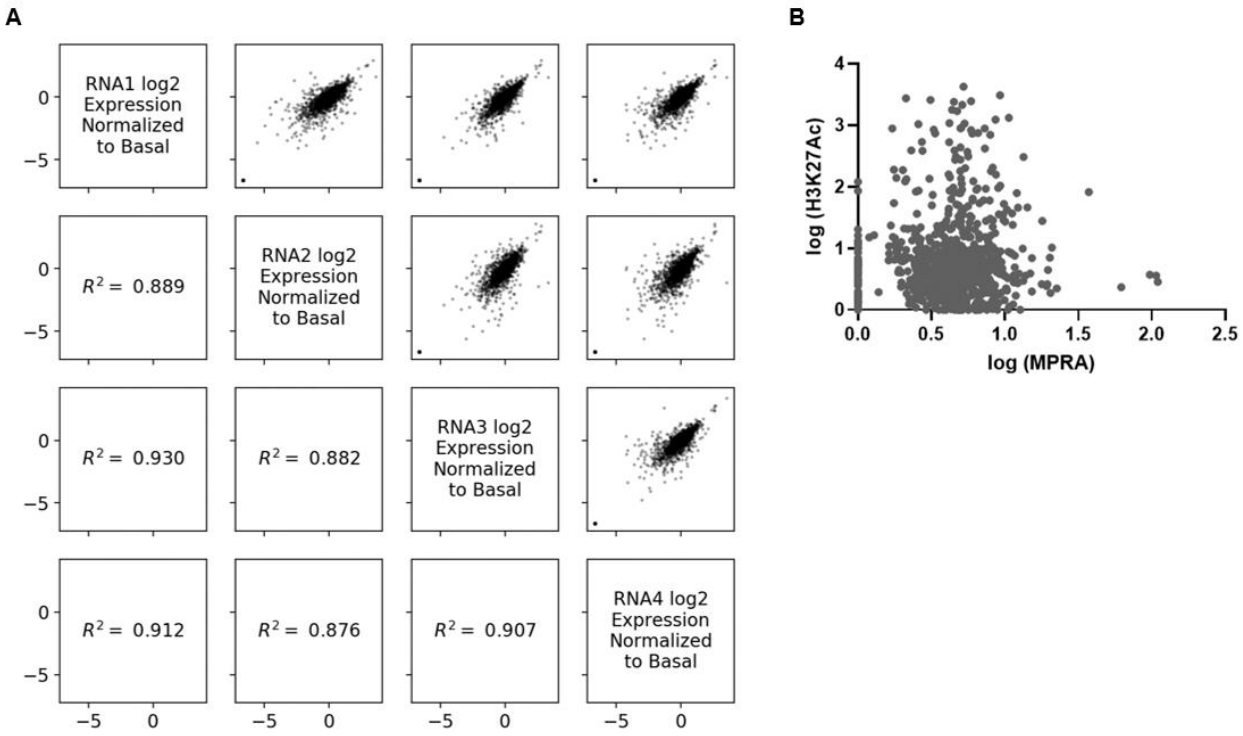

Fig. S5: A) Correlation plot showing reporter activity, normalized to basal expression, across four replicates. **B)** Correlation plot between MPRA data from GMSM-K and publicly available H3K27Ac data (GSM733674) from NHEK.

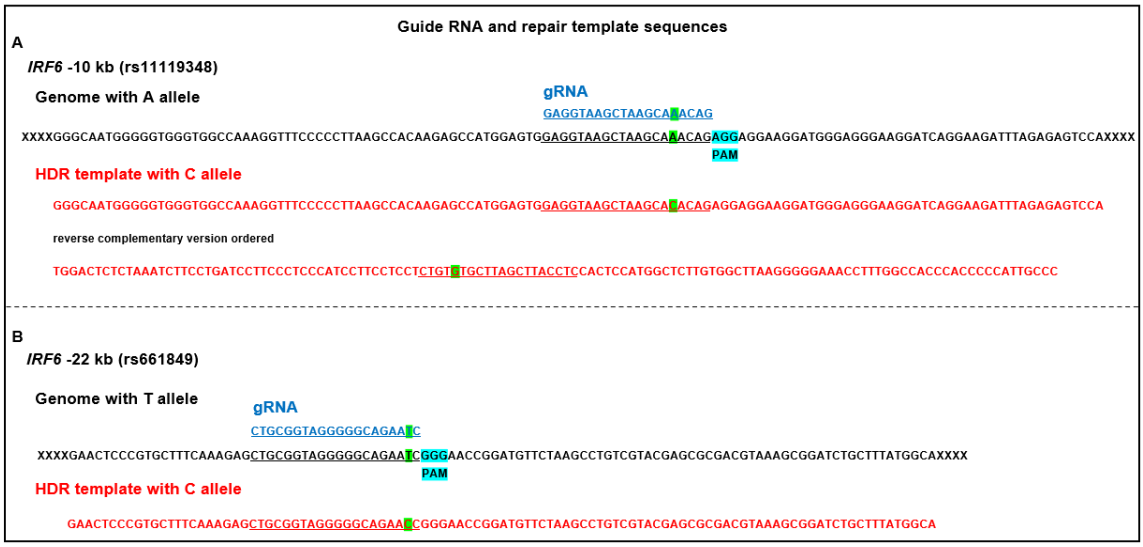

Fig. S6: Sequences of the guide RNA (gRNA) and homology-directed repair (HDR) template DNA used for CRISPR-Cas9 editing of the **A)** rs11119348 and **B)** rs661849 variants. PAM represents the protospacer adjacent motif for the gRNAs used in the study.

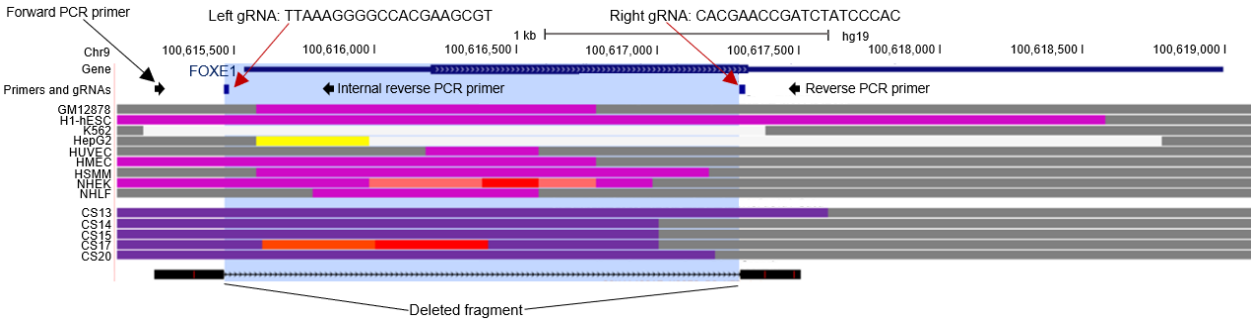

Fig. S7: *FOXE1* KO GMSM-K: Browser view of the human genome, GRCh37/hg19, focused on *FOXE1* gene. First track represents a pair of guide RNAs (marked with a block) and primers (marked with arrows) used to generate *FOXE1* KO cells. Primer sequences are provided in Table S5. Following color coded bars represent chromatin status revealed by ChIP-Seq to various chromatin marks from the ENCODE Project cell lines and facial explants from human embryos at Carnegie stage (CS) 13-20, encompassing the time when palatal shelves fuse, where yellow bars represent the enhancer element, bright red, active promoter; light red, weak promoter; purple, inactive/poised promoter; grey, Polycomb repressed; dark purple in facial explants dataset, bivalent promoter; 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, CS13-CS20 are facial explants from human embryos. The blue region highlights the deleted fragment after CRISPR/Cas9 targeting.

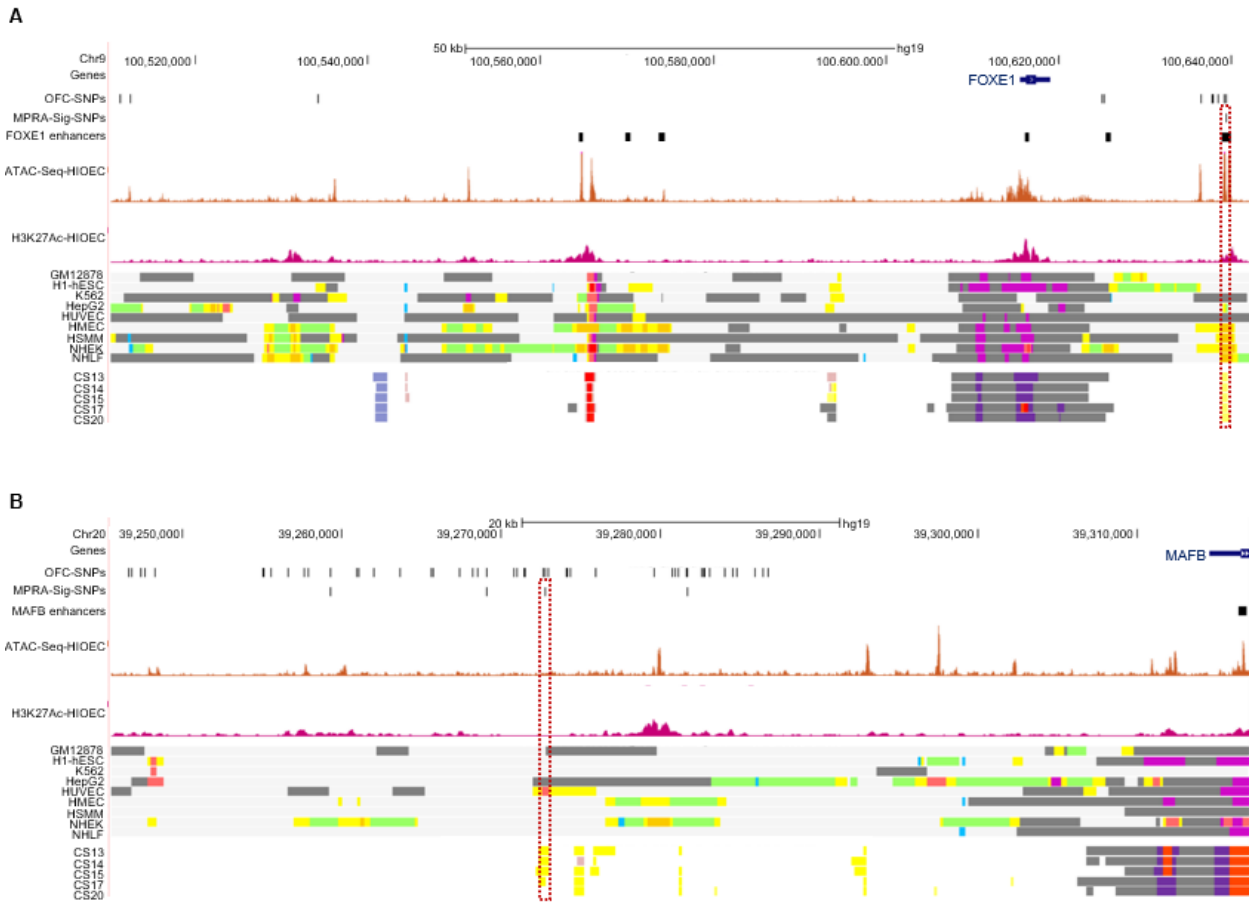

Fig. S8A, B: SNPs with significant effect in MPRA near A) *FOXE1* and B) *MAFB*: A) Browser view of the human genome, GRCh37/hg19, focused on the GWAS-identified SNPs at *FOXE1.* First two tracks represent the OFC associated SNPs at *FOXE1* and MPRA-identified SNPs at *FOXE1* respectively. Next track represents the ABC (Activity by Contact) elements scored using HiC data from NHEK. Next two tracks represent open chromatin peaks detected by ATAC-Seq and H3K27Ac marks in HIOEC respectively. Following color coded bars represent chromatin status revealed by ChIP-Seq to various chromatin marks from the ENCODE Project cell lines and facial explants from human embryos at Carnegie stage (CS) 13-20, encompassing the time when palatal shelves fuse, where orange and yellow bars represent the active and weak enhancer element, respectively; bright red, active promoter; light red, weak promoter; purple, inactive/poised promoter; blue, insulator; light green, weak transcribed; grey, Polycomb repressed; light grey, heterochromatin/repetitive; pale turquoise, light purple and dark purple in facial explants dataset, heterochromatin poised promoter and

bivalent promoter respectively; 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, CS13-CS20 are facial explants from human embryos.
Box in **A**, a SNP (rs10984103) with significant effect in MPRA lying within an enhancer for *FOXE1* in HEKn by the ABC method*,* and within chromatin marked as an active enhancer in human embryonic faces. **B)** Browser view of the human genome, GRCh37/hg19, focused on the GWAS-identified SNPs at *MAFB* associated with orofacial cleft. Tracks and color code are similar as in Fig. 2B. Box in **B**, a SNP (rs4812449) with significant effect in MPRA lying within chromatin marked as an active enhancer in human embryonic faces.

Fig. S8C: SNPs with significant effect in MPRA near *TFAP2A*:

Browser view of the human genome, GRCh37/hg19, focused on the GWAS-identified SNPs at *TFAP2A* associated with orofacial cleft. Tracks and color code are similar as in Fig. S8A, B. Blue and green dashed lines indicate the zoomed-in browser view from the left-hand side and right-hand side OFC-associated SNPs at *TFAP2A* respectively. Box, SNPs with significant effect in MPRA lying within chromatin marked as an active enhancer in NHEK (top-box: rs9369420) human embryonic faces (bottom-box: rs201265).

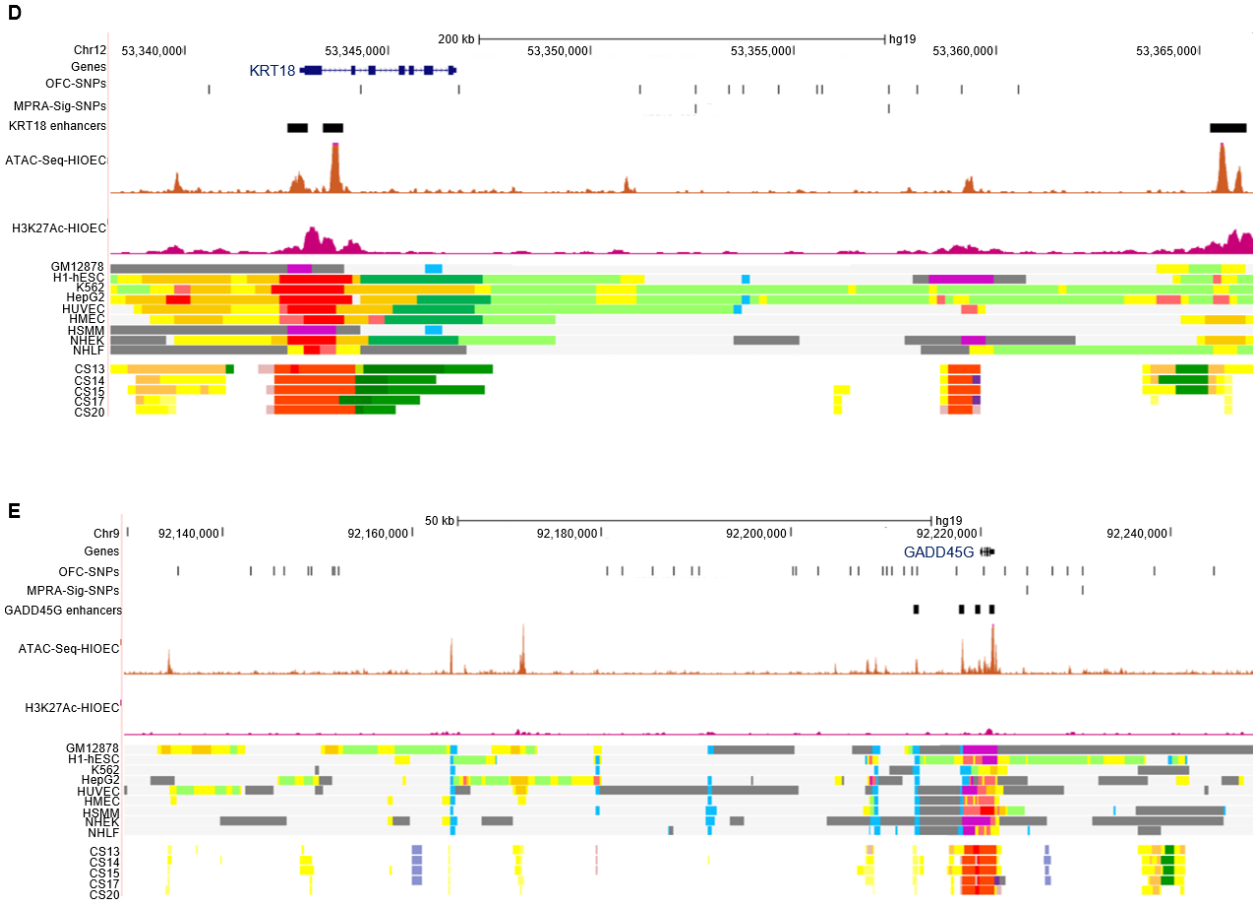

Fig. S8D-E: SNPs with significant effect in MPRA near D) *KRT18* and E) *GADD45G*: Browser view of the human genome, GRCh37/hg19, focused on the GWAS-identified SNPs at **D)** *KRT18* and **E)** *GADD45G* associated with orofacial cleft. Tracks and color code are similar as in Fig. S8A, B.

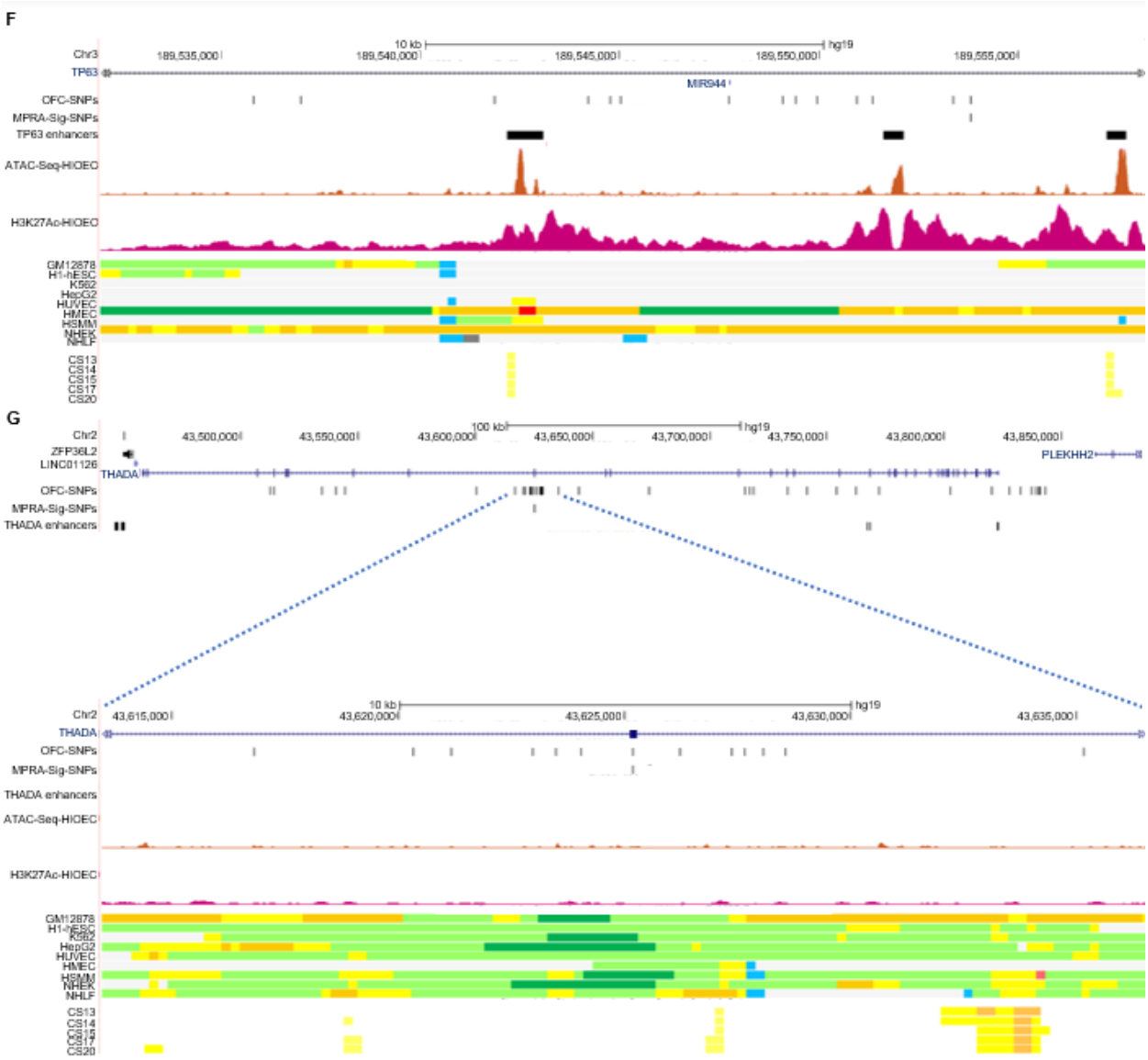

Fig. S8F-G: SNPs with significant effect in MPRA near F) *TP63* and G) *THADA*: Browser view of the human genome, GRCh37/hg19, focused on the GWAS-identified SNPs at **F)** *TP63* and **G)** *THADA* associated with orofacial cleft. Tracks and color code are similar as in Fig. S8A, B.

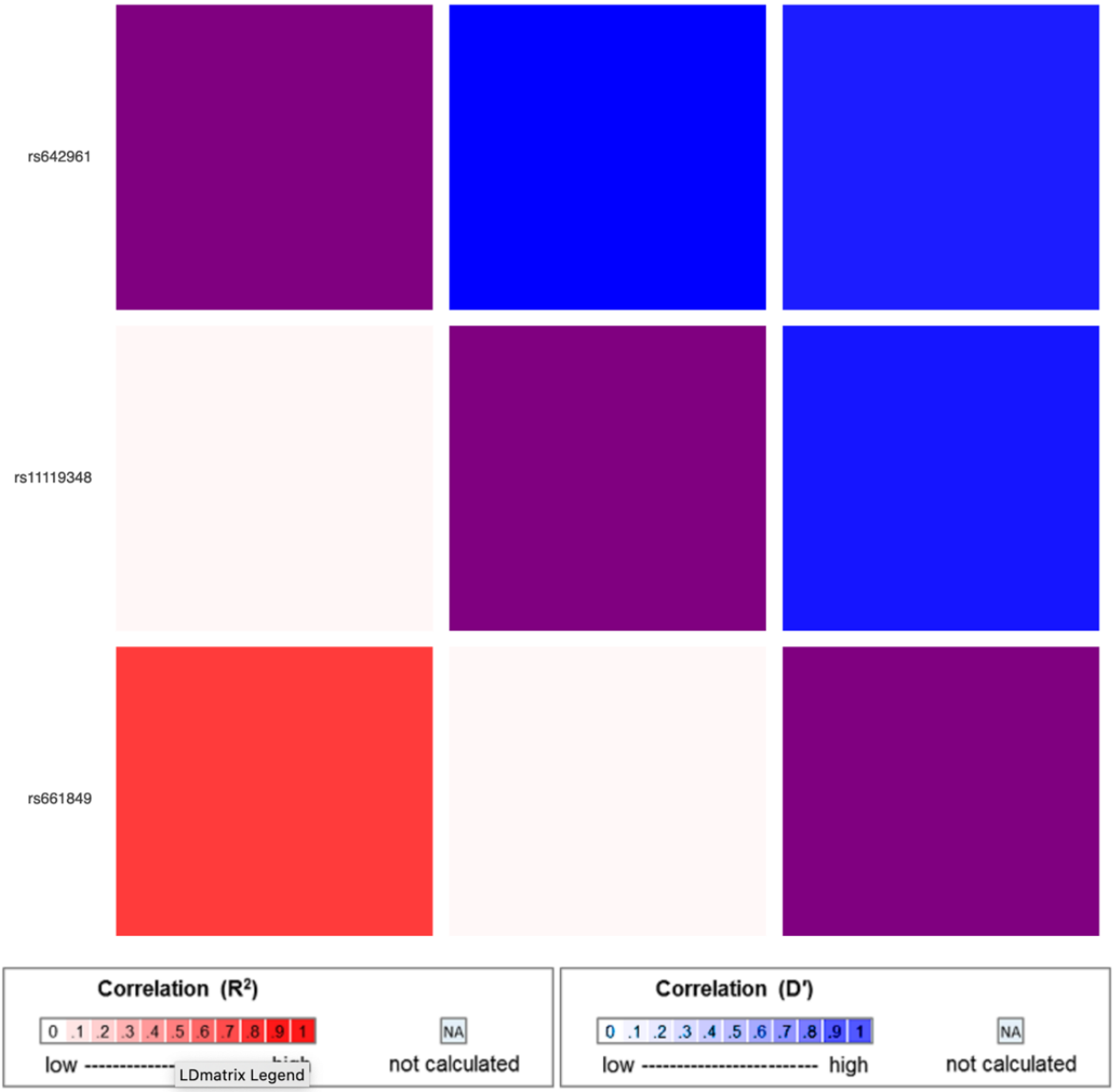

Fig. S9: Linkage disequilibrium plot between rs642961 and *IRF6* -10 kb (rs11119348) and -22 kb (rs661849) SNPs.

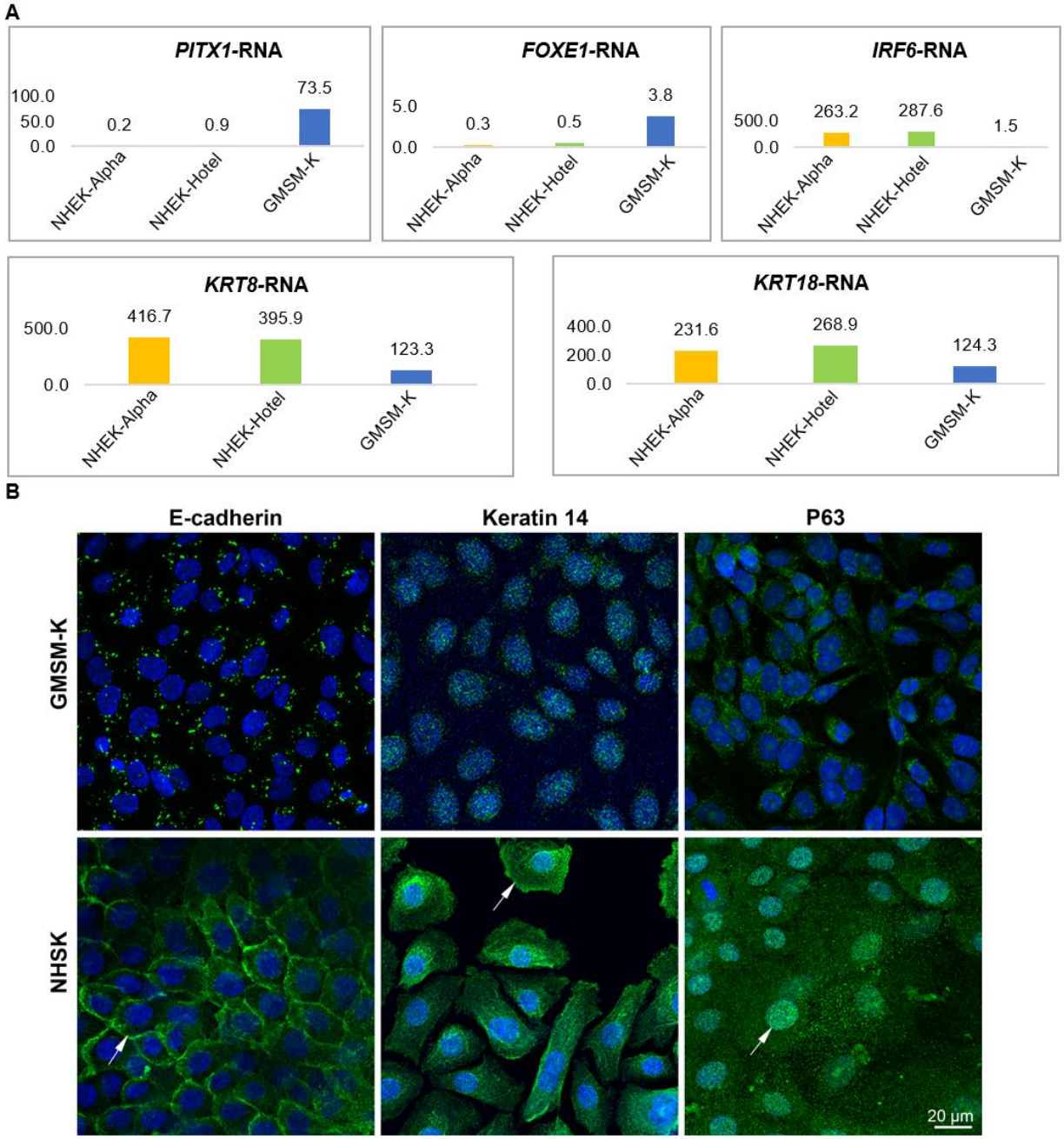

Fig. S10: A) RNA expression levels of different epithelial markers and *IRF6* gene plotted from RNA-Seq data. For NHEK two different publicly available RNA-seq profiles were used (GSM6050574: NHEK-Alpha and GSM6050576: NHEK-Hotel). **B)** Immunofluorescence staining for E-cadherin, Keratin 14, and P63 (green) of human GMSM-K and normal human skin keratinocytes (NHSK). White arrows indicate positive staining in NHSK absent in GMSM-K. Nuclear DNA is labeled with DAPI (blue). Scale bar = 20 µm.

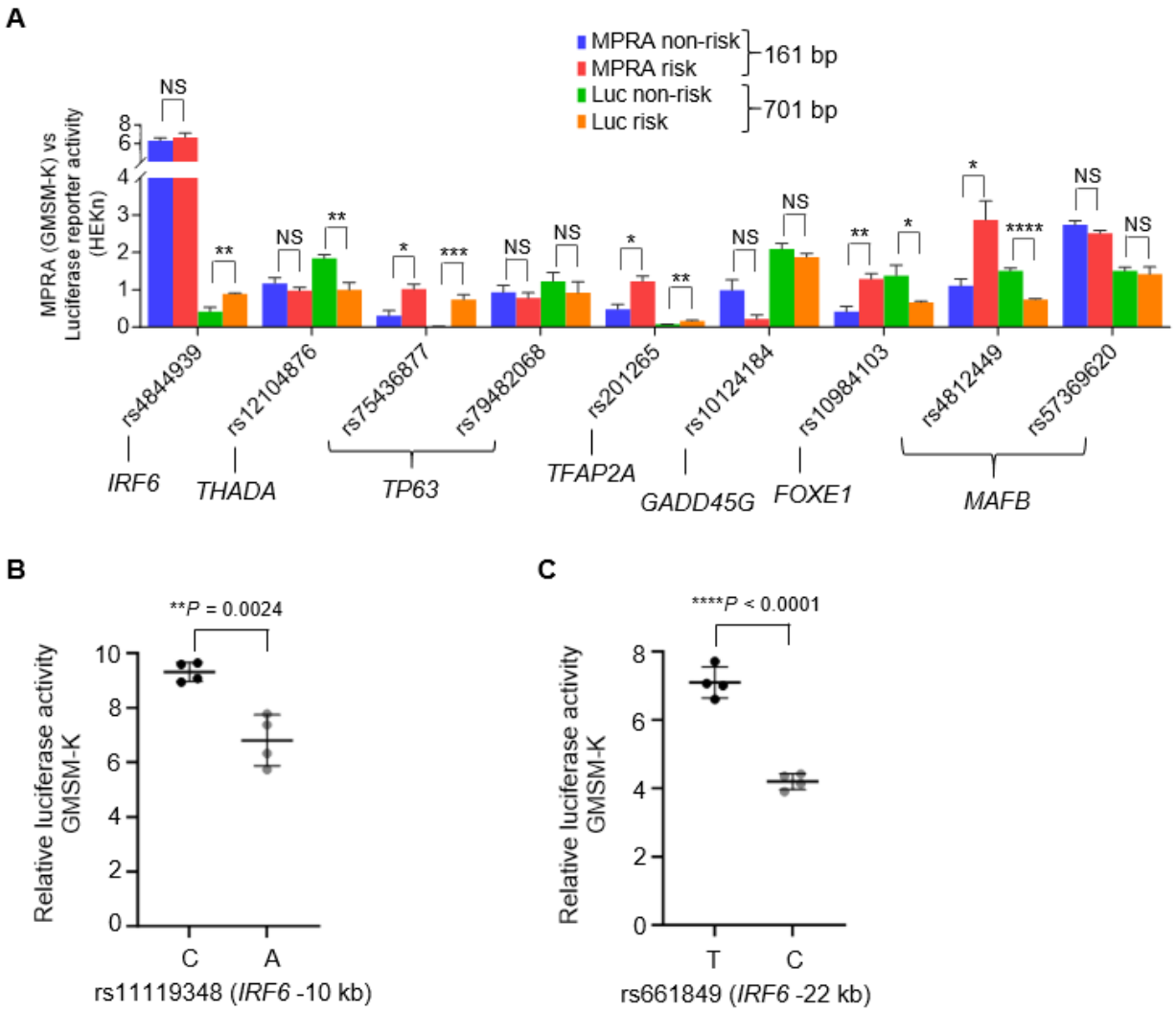

Fig. S11: Luciferase reporter assay in HEKn and GMSM-K: **A)** Bar chart of MPRA and luciferase reporter activities for the indicated SNPs at various loci in HEKn. For luciferase reporter activities, data are represented as mean values +/- s. d. from three independent experiments. Statistical significance is determined by Student's *t*-test, two-tailed *P* value (**P*<0.05, ***P*<0.01, ****P*<0.001, *****P*<0.0001) and NS represents non-significant. Luc represents luciferase. **B** and **C)** Scattered dot plot of relative luciferase activity for longer constructs of non-risk and risk alleles of rs11119348 and rs661849 respectively in GMSM-K. Data are represented as mean values +/- s. d. from four independent experiments. Statistical significance is determined by Student's *t*-test. *P* value (two-tailed) is indicated on the plot.

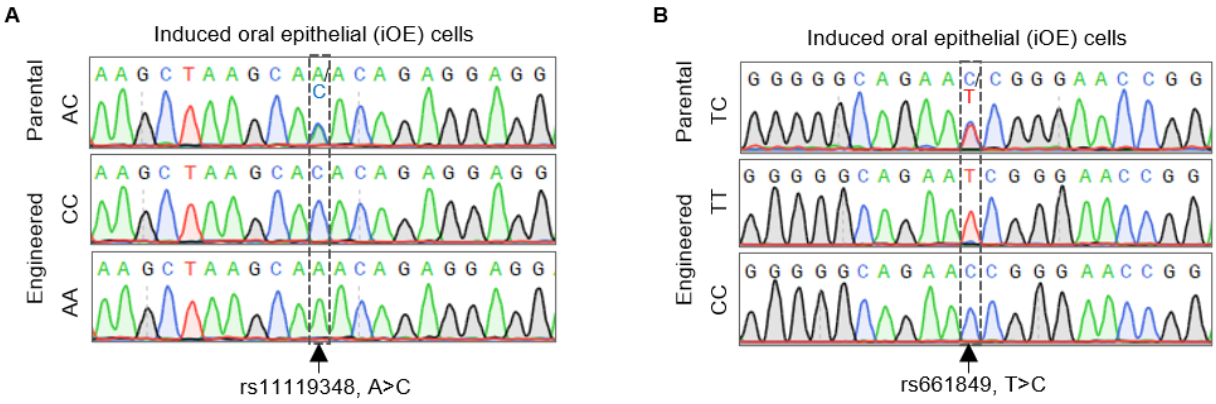

Fig. S12: Genome of iPSCs engineered to homozygosity for risk and non-risk alleles of the *IRF6* -10 kb SNP and, separately, the *IRF6* -22 kb SNP.

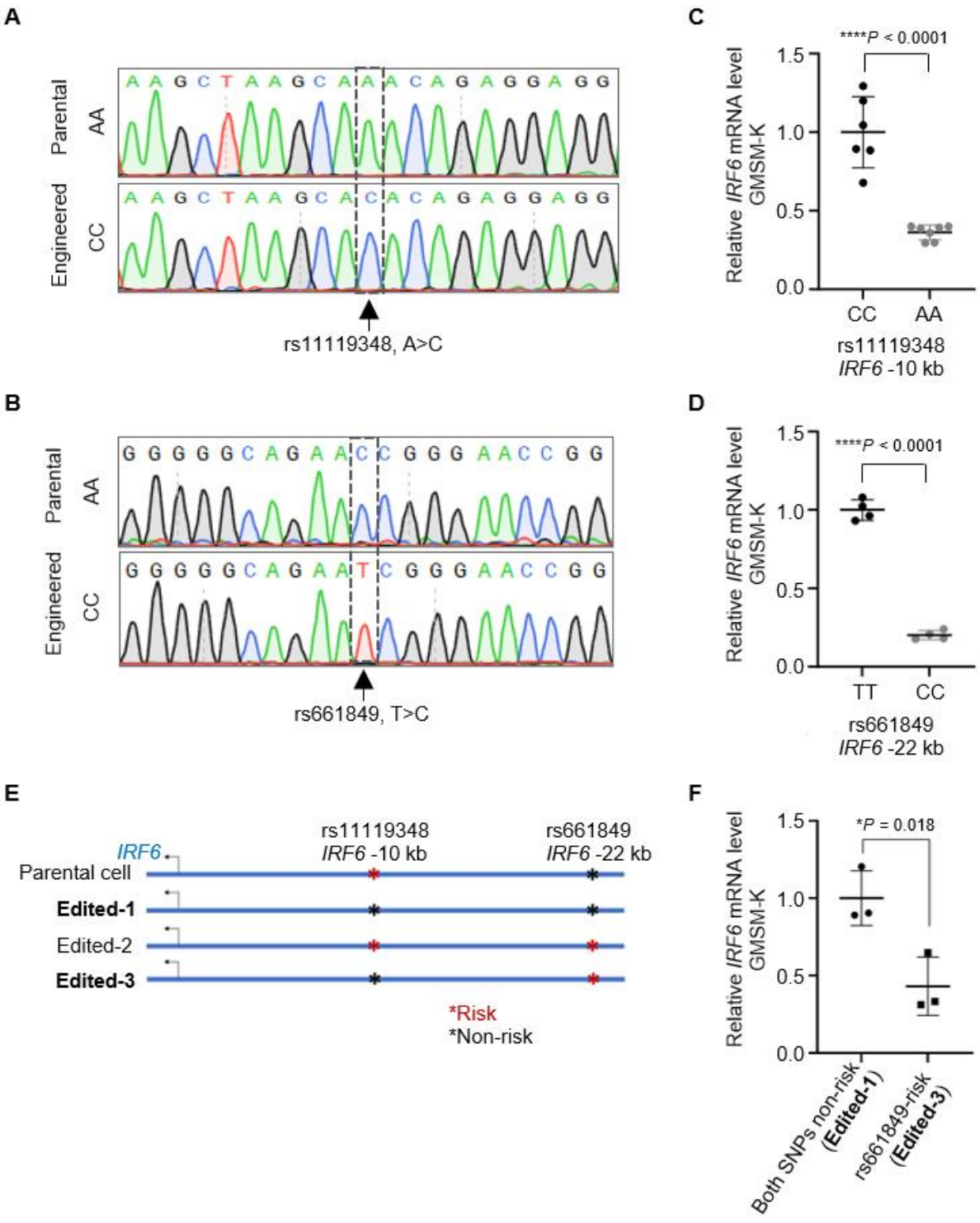

Fig. S13: *IRF6* RNA expression in engineered GMSM-K: **A** and **B)** GMSM-K engineered to homozygosity for risk and non-risk alleles of the *IRF6* -10 kb SNP and, separately, the *IRF6* -22 kb SNP respectively: **C** and **D)** Scattered dot plot of relative

levels of *IRF6* mRNA in GMSM-K harboring non-risk (CC) and risk (AA) genotype of rs11119348 **(C)** and in GMSM-K harboring non-risk (TT) and risk (CC) genotype of rs661849 **(D)** assessed by qRT-PCR. Expression levels of *IRF6* are normalized against *ACTB.* Data are represented as mean values +/- s. d from multiple replicates as indicated in the plot. Statistical significance is determined by Student's *t*-test (two-tailed). Please note that GMSM-K harboring non-risk (TT) or risk (CC) genotypes of rs661849 in **B** and **D** harbors the parental risk (AA) genotype of the other SNP (rs11119348) as indicated in E. **E)** Engineered genome of GMSM-K harboring non-risk (TT) or risk (CC) genotypes of rs661849 in the background of non-risk (CC) genotype of rs11119348. **F)** Scattered dot plot of relative levels of *IRF6* mRNA in GMSM-K harboring non-risk (TT) and risk (CC) genotype of rs661849 in the background of non-risk (CC) genotype of rs11119348 assessed by qRT-PCR. Expression levels of *IRF6* are normalized against *ACTB.* Data are represented as mean values +/- s. d from multiple replicates as indicated in the plot. Statistical significance is determined by Student's *t*-test (two-tailed).

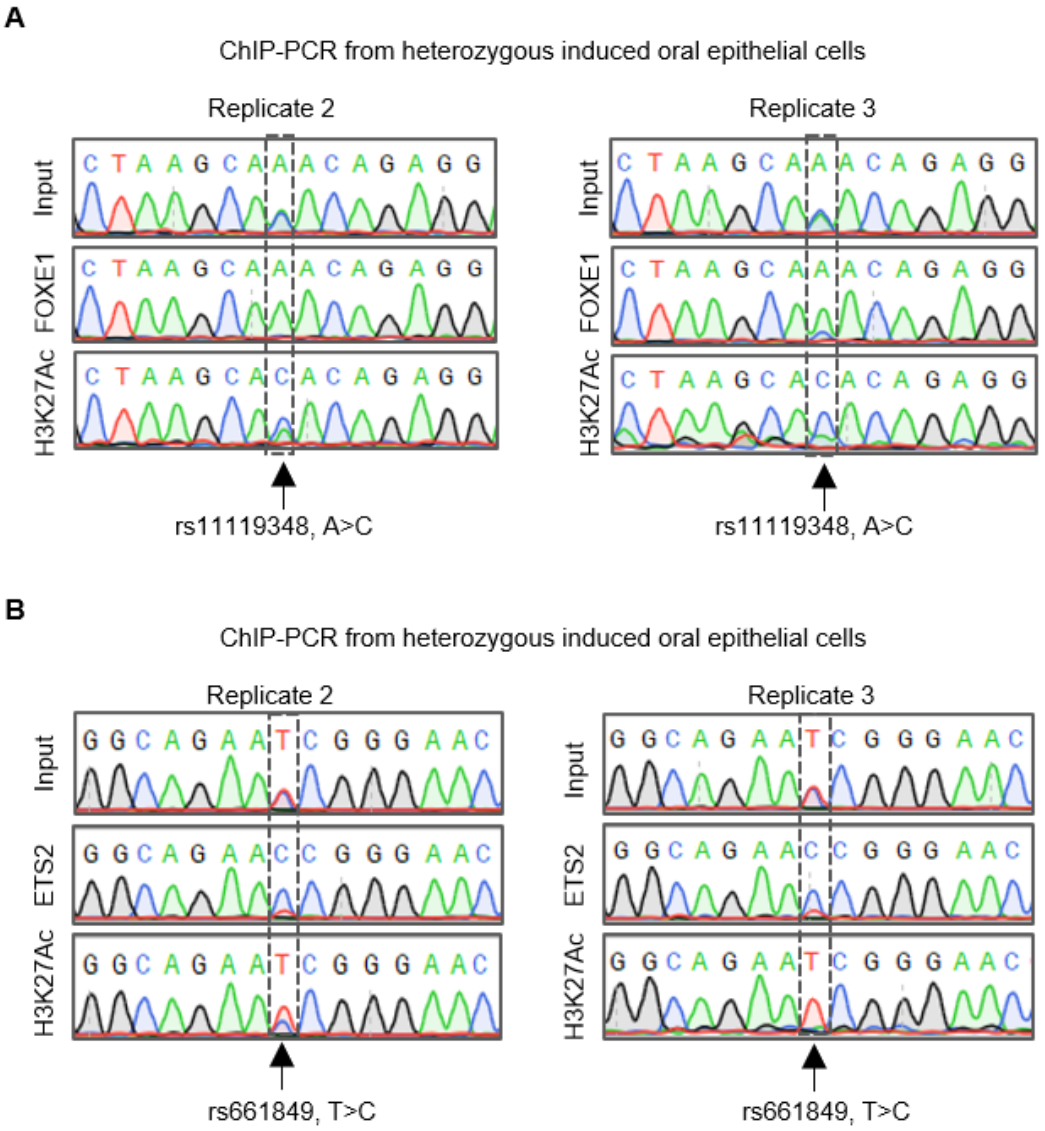

Fig. S14: Sequencing of **A)** anti-FOXE1 and anti-H3K27Ac ChIP-PCR product of cells heterozygous for rs11119348 from two ChIP replicates and **B)** anti-ETS2 and anti-H3K27Ac ChIP-PCR product of cells heterozygous for rs661849 from two ChIP replicates.

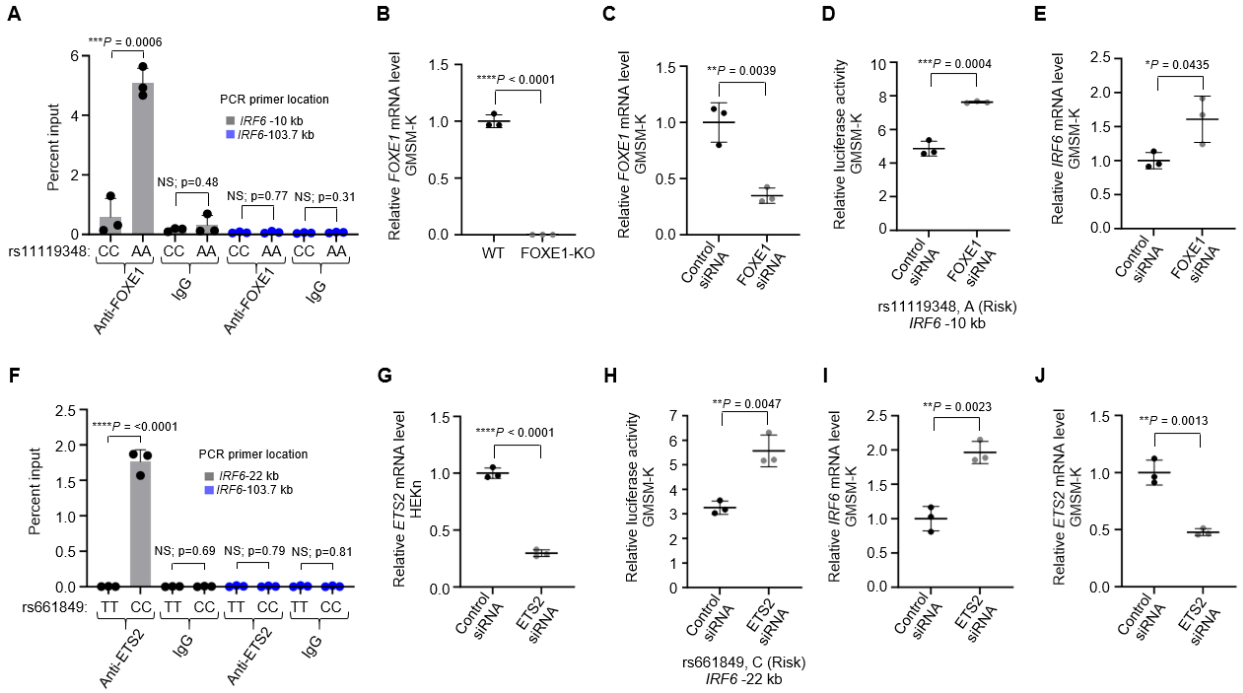

Fig. S15: A) Percent input identified by ChIP-qPCR for anti-FOXE1 in GMSM-K harboring non-risk (CC) and risk (AA) genotype of rs11119348 using primers specific to the *IRF6* -10 kb enhancer site or, as a negative control, to a region 103.7 kb upstream to the *IRF6* transcription start site that did not harbor active elements identified from ATAC-Seq and H3K27Ac ChIP-Seq in HIOEC or NHEK. Error bars refer to 3 ChIP replicates and expressed as mean values +/- s. d. Statistical significance is determined by Student's *t*-test. *P* value (two-tailed) is indicated on the plot. NS, non-significant. **B and C)** Scattered dot plot of relative levels of *FOXE1* mRNA in WT and *FOXE1* KO GMSM-K **(B)** or control siRNA and FOXE1 targeting siRNA transfected GMSM-K **(C)** assessed by qRT-PCR. Expression levels of *FOXE1* are normalized against *ACTB.* Data are represented as mean values +/- s. d. from three replicates. Statistical significance is determined by Student's *t*-test (two-tailed). **D)** Scattered dot plot of relative luciferase activity for longer construct with risk allele (A) of rs11119348 in GMSM-K transfected with control siRNA or FOXE1 targeting siRNA. Data are represented as mean values +/- s. d. from three independent experiments. Statistical significance is determined by Student's *t*-test. *P* value (two-tailed) is indicated on the plot. **E)** Scattered dot plot of relative levels of *IRF6* mRNA in control siRNA and FOXE1

targeting siRNA transfected GMSM-K assessed by qRT-PCR. Expression levels of *IRF6* are normalized against *ACTB.* Data are represented as mean values +/- s. d. from three replicates. Statistical significance is determined by Student's *t*-test (two-tailed). **F)** Percent input identified by ChIP-qPCR for anti-ETS2 in GMSM-K harboring non-risk (TT) and risk (CC) genotype of rs661849 using primers specific to the *IRF6* -22 kb enhancer site or, as a negative control, to a region 103.7 kb upstream *IRF6* transcription start site that did not harbor active elements identified from ATAC-Seq and H3K27Ac ChIP-Seq in HIOEC or NHEK. Error bars refer to 3 ChIP replicates and expressed as mean values +/- s. d. Statistical significance is determined by Student's *t*-test. *P* value (two-tailed) is indicated on the plot and NS represents non-significant. **G)** Scattered dot plot of relative levels of *ETS2* mRNA in control siRNA and ETS2 targeting siRNA transfected HEKn assessed by qRT-PCR. Expression levels of *ETS2* are normalized against *ACTB.* Data are represented as mean values +/- s. d. from three replicates. Statistical significance is determined by Student's *t*-test (two-tailed). **H)** Scattered dot plot of relative luciferase activity for longer construct with risk allele (C) of rs661849 in GMSM-K transfected with control siRNA or ETS2 targeting siRNA. Data are represented as mean values +/- s. d. from three independent experiments. Statistical significance is determined by Student's *t*-test. *P* value (two-tailed) is indicated on the plot. **I** and **J)** Scattered dot plot of relative levels of *IRF6* mRNA **(I)** or *ETS2* mRNA **(J)** in control siRNA and ETS2 targeting siRNA transfected GMSM-K assessed by qRT-PCR. Expression levels are normalized against *ACTB.* Data are represented as mean values +/- s. d. from three replicates. Statistical significance is determined by Student's *t*-test (two-tailed).