HGGA, Volume *2*

Supplemental information

Integrative approaches generate insights into the architecture of non-syndromic cleft lip with or without cleft palate

Julia Welzenbach, Nigel L. Hammond, Miloš Nikolić, Frederic Thieme, Nina Ishorst, Elizabeth J. Leslie, Seth M. Weinberg, Terri H. Beaty, Mary L. Marazita, Elisabeth Mangold, Michael Knapp, Justin Cotney, Alvaro Rada-Iglesias, Michael J. Dixon, and Kerstin U. Ludwig

Supplemental Material and Methods

Cohort Description

The meta-analysis included data from three previously published individual GWAS on nsCL/P (Bonn case-control GWAS cohort[1], GENEVA trio cohort[2], POFC GWAS cohort[3], Table S1). We included all nsCL/P summary statistics that were publicly accessible until June 2018. In the present study we combined the three GWAS cohorts to generate the largest nsCL/P meta-analysis to date. In accordance with previous studies[4-6], two meta-analyses were performed: (1) using all individuals with diverse population backgrounds (MAiC, for **M**eta-**A**nalysis **i**n **C**lefting) and (2) using the European data sets only ($MAiC_{Euro}$). Data quality control (QC) included the detection and removal of overlapping individuals, confirmation of ethnicity, and data re-analysis as described in the following:

Data QC on the individual Bonn and GENEVA cohorts was done as previously described[4,5]. For the POFC GWAS cohort, imputed genotypes were retrieved from dbGaP. Briefly, this data comprised genotypes for 557,677 single nucleotide polymorphisms (SNPs) in 11,855 individuals of diverse ethnicities (Ethiopia, Nigeria, China, India, Philippines, Denmark, Hungary, Spain, Turkey, Argentina, Colombia, Guatemala, Puerto Rico, United States), representing 3,981 families. Based on these genotype data, a dataset was constructed that had maximal overlap to the original published POFC GWAS[3] while excluding individuals that have already been analysed previously as part of the GENEVA cohort. First we generated a combined POFC-GENEVA dataset and used KING[7] on a set of 115,380 genotyped variants, to estimate relatedness between individuals. Relationship was defined based on a KING kinship coefficient ≥0.0884 (representing 2nd degree relationship), and affected individuals or families were removed from the POFC cohort. In the remaining POFC individuals, we then aimed at maximizing the number of case-parent trios, resulting in 1,328 complete trios (1,319 had been included in Leslie et al. 2016[3]; nine additional nsCL/P trios were identified based on inference of family structure). This data set formed the final $POFC_{trio}$ cohort. From the remaining families (where no case-parent trio had been drawn from), independent individuals were selected to construct the $POFC_{case\text{-}control}$ cohort.

Individuals were designated “cases” if affected with nsCL/P, based on the phenotypic data provided. In situations when multiple individuals within one family were affected, the individual with the lowest number of missing genotypes was included. This resulted in 848 cases. For the control set, data from 1,568 families without any affected individual were available. From these families, the individual with lowest number of missing genotypes was selected as control and included in the $POFC_{case\text{-}control}$ cohort (n=1,568 controls).

Ethnicity was identified based on the information provided in column “country of origin”, and genotype data. Individuals were classified into “European” and “Non-European” based on the smallest distance in mean and standard deviation of the first two principal component analysis (PCA)-Eigenvectors, to the defined Central European country category ‘Denmark/Hungary/Spain/Turkey’. To exclude individuals that were identical within studies (or were part of a “superfamily” with other individuals in the cohorts), the relationship between individuals was calculated with KING based on the shared variants in the Bonn, GENEVA and POFC cohorts. Again, individuals showing a kinship-coefficient of ≥ 0.0884 were excluded, this resulted in removal of 90 case-parent trios, one case, and seven controls from the POFC cohort. Thus, the final MAiC dataset resulted in 848 cases, 1,561 controls and 1,238 case-parent trios after sample QC (Table S1).

Statistical analyses

Statistical analyses were performed separately for case-control cohorts and case-parent trios, respectively. Notably, the Ludwig et al. (2017) meta-analysis had applied the FBAT-dosage method to generate genotypes for the trio cohorts, lacking individual relative risk (RR) information. In the present study, best-guess genotypes were assigned based on *a-posteriori* genotype probabilities of ≥0.6. In the case-control cohorts, GWAS was performed using SNPTEST and -method expected, by incorporating 5-18 dimensions of the multi-dimensional-scaling coordinates[8], respectively. For the case-parent trios a transmission disequilibrium test (TDT) was performed on the best-guess genotypes[9]. Given the present sample sizes, we accounted for the limited power of imputation approaches to correctly predict rare and low-frequency variants by retaining robust SNPs only (info-score ≥0.4, minor allele frequency

(MAF) >1 % in controls and non-transmitting parents). Moreover, for the case-parent trios, SNPs had to be present in ≥75 % of the families. After data cleaning procedures, we meta-analyzed the GWAS data of all four sub-cohorts (Bonn case-control, GENEVA case-parent trios, POFC case-control and POFC case-parent trios) using METAL[10]. METAL combines p-values across studies while considering directions of effects and effective sub-cohort size, as indicated by N_{eff}. Here, N_{eff} is defined as

$$case/control\ N_{eff} = \frac{4}{\frac{1}{n\ cases} + \frac{1}{n\ controls}} \qquad \text{and} \qquad trio\ N_{eff} = \frac{4}{\frac{1}{n\ trio} + \frac{1}{n\ trio}}$$

Post-analysis SNPs that were absent from more than one sub-cohort (Bonn case-control, GENEVA case-parent trios, POFC case-control and POFC case-parent trios) were removed. Thus, the final MAiC dataset contained 7,744,527 SNPs in MAiC, and 7,690,843 SNPs in $MAiC_{Euro}$. To estimate the SNP-based heritability (h^2) for nsCL/P on the liability scale, we generated a European case-control-only dataset that excluded the case-parent trios (Table S1). Using this data set we performed linkage disequilibrium (LD) score regression as implemented in ldsr[11], for individuals of European ethnicity. Because LD score regression requires a homologues data structure, it was not possible to apply ldsr on the whole nsCL/P dataset with mixed ethnicities and complex cohort structure (case/control and case-parent trios).

Identification of novel nsCL/P risk loci

We defined 'previously identified risk loci for nsCL/P' as those having reached genome-wide significance ($P<5\times10^{-08}$) in either GWAS, meta-analysis or large-scale systematic study before (n=40, Table S2). For each of the lead variants (as defined in the respective original study) a window of strong LD ($r^2 \geq 0.6$) was defined. The most distant SNPs at this threshold determined the boundaries for known genetic risk loci. For those of the previously identified risk loci that reached genome-wide significance in the MAiC dataset (n=26), the 1000 Genome Phase3 reference panel, Central European backbone, was used to compute r^2 in the FUMA (Functional Mapping and Annotation of Genome-Wide Association Studies) platform, v1.3.2[12]. For the remaining 14 loci, $r^2 = 0.6$ boundaries were estimated using LDproxy provided in LDlink 3.2.0 suite[13], in a mixed (European/East Asian) population. 'Novel risk loci' were defined as those with $P<5\times10^{-08}$ in the MAiC analyses if located outside of the previously

identified loci. For each of the novel risk loci identified in MAiC, we defined the associated region using the same parameters (Table S2). For follow-up analyses, recent data from GTEx (v8)[14] were assessed using the GTEx browser.

Gene-based and pathway analyses

Gene-based analyses in MAiC and $MAiC_{Euro}$ were performed using MAGMA[15] (v1.06), implemented in FUMA. MAGMA's gene analysis uses a multiple regression approach to properly incorporate LD between markers and to detect multi-marker effects. We run the gene-based analysis with default parameters (SNP-wide mean model), using 1000 Genome Phase3 as reference panel and the full distribution of imputed input SNPs of MAiC (N=7,744,527; info-score≥0.4; MAF>1%). These were mapped to 17,911 protein-coding genes based to a distance of 0kb upstream/downstream of the genes, resulting in threshold of test-wide significance of $P = 2.79\times10^{-6}$ (i.e., 0.05/17,911).

Epigenetic datasets for mid-facial development

Identification of datasets relevant to mid-facial development

Human cell-type and developmental-stage specific data for mid-facial development is underrepresented (or not represented at all) in large consortia data such as ENCODE[16]. However, available data in the Gene Expression Omnibus (GEO) covered mid-facial development from (i) early stages (hNCC[17], accessed through GSE28874), (ii) differentiated human cNCC[18] (accessed through GSE70751), and (iii) embryonic craniofacial human tissue of different CS (accessed through GSE97752)[19]. Each of the datasets is briefly described in the following section:

(i) *Human neural crest cells.* For the establishment of hNCCs, an *in vitro* differentiation model had been used in which hESC (H9 cell line) were first induced to form neuroectodermal spheres (hNEC) that subsequently gave rise to migratory cells expressing early NC markers and recapitulating neuronal, mesenchymal and melanocytic differentiation potential of the neural crest[17,20]. Sequential ChIP assays

in hNCCs had been performed from approximately 10^7 hNCC cells per experiment[21], as described before[22] and included chromatin modifications H3K27ac, H3K4me1, H3K4me3 and H3K27me3.

(ii) *Human cranial neural crest cells.* Human cNCCs had been differentiated *in vitro* from iPSCs (H9 cell line), first forming hNECs which then later attached and gave rise to migratory cNCCs which could be maintained up to 18 passages[18]. In cNCC, ChIPs had been performed using approximately 0.5×10^{-7} to 1.0×10^{-7} cells per experiment, with the same protocol as described for hNCC. ChIP-Seq had been performed for chromatin modifications H3K27ac, H3K4me1, H3K4me3 and H3K27me3, as for hNCC.

(iii) *Craniofacial tissue from different stages.* Human embryonic CT was collected, staged, and provided by the Joint MRC/Wellcome Trust Human Developmental Biology Resource (HDBR, www.hdbr.org). Information describing the developmental stage, termination method, collection site, and karyotype of each embryo and the ChIP protocol is provided in the original study[19]. Briefly, samples encompassed CS 13 (4.5 weeks *post-conception*, wpc, 5 embryos), 14 (5 wpc, 3 embryos), 15 (5.5 wpc, 3 embryos), 17 (6 wpc, 4 embryos) and 20 (8 wpc, one embryo), and one sample of 10 wpc. For samples of CS13-CS17, ChIP-Seq was performed for chromatin modifications H3K27ac, H3K4me1, H3K4me3, H3K27me3 and H3K36me3 and resulted in a mean of 37.3 million uniquely aligned reads per sample and chromatin mark (Table S4). Data for H3K9me3 marks were imputed. For CS20 and 10 wpc, H3K27ac3 ChIP-Seq data was experimentally derived, all other marks were imputed (Table S3).

ChIP-Seq Data processing

For hNCC and cNCC, raw data were available in fastq format. ChIP-seq data from craniofacial data Wilderman et al. (2018)[19] comprise processed formats, including imputed signals, peaks and segmentation data. In order to ensure comparability between the three data sources, computational processing of ChIP-seq data as published in Wilderman et al. 2018 (QC, alignment, peak calling, epigenetic imputation, chromatin segmentation) was adopted to the hNCC/cNCC bioinformatics pipeline. Scripts have been used as deposited on https://github.com/cotneylab/ChIP-Seq.

Briefly, FastQC (v0.11.7) was used to combine multiple fastq files of one ChIP experiment and perform QC. Alignment of the single-end reads with length of 36 bp (hNCC) and 36-50 bp (cNCC) to the human genome (hg19) was performed with Bowtie2 (v2.3.2)[23]. Fragment sizes of each library were estimated using PhantomPeakQualTools (v1.14)[24]. For hNCC and cNCC, the number of uniquely aligned reads per sample and chromatin mark is given in Table S5. In the following, analysis of treatment against input sample was performed to generate p-value based signal tracks and peak files based on estimated library fragment size using MACS2 (v2.1.1.20160309)[25].

Chromatin imputation and segmentation.

To obtain uniform data sets, chromatin imputation followed by chromatin state segmentation was performed. First, H3K9me3 and H3K36me3 marks in both hNCC and cNCC were imputed using ChromImpute (v1.0.1)[26], based on 127 cell types from the Roadmap Epigenome Project[27]. Briefly we used *P*-value-based signal files for marks H3K27ac, H3K4me1, H3K4me3, H3K27me3, H3K9me3 and H3K36me3 for 127 tissues and cell types. Conversion from bigWig to bedGraph format was done using the ENCODE function 'BigWigToBedGraph'. Both the hNCC and cNCC p-value signals in bedGraph format as well as the Roadmap reference samples were converted to 25 bp resolution and processed for model training and generation of imputed signals.

Imputed hNCC and cNCC signal files for each individual chromosome and each chromatin mark were binarized, and segmentation was performed using the core+K27ac 18-state chromatin model provided by Roadmap with ChromHMM[28] which uses a multivariate Hidden Markov Model that explicitly models the combinatorial presence or absence of each mark to predict 18 chromatin states. This procedure can identify tissue-specific regulatory information from initial tissue/cells for which no gene expression data is available. Because of the low number of chromatin marks measured in the NCC samples, epigenetic imputation issues and the higher risk of batch effect between hNCC, cNCC and CT, we adopted a robust strategy and condensed the 18 generated states into eight states.

Translation of genetic associations into tissue- and time point-specific regulatory effects at a systematic level

Enrichment analyses using GREGOR

Based on chromatin segments obtained from hNCC, cNCC and CT (see section "Chromatin imputation and segmentation"), we used the GREGOR software (Genomic Regulatory Elements and Gwas Overlap algorithm)[29] to evaluate the enrichment of significant SNPs from the MAiC data in the available regulatory features (i.e., eight predicted chromatin states). A set of samples from the Roadmap Epigenomics project (comprising both fetal and adult tissue samples) was selected as independent dataset for comparison. These included ESC H1 cell line (ESC.H1), iPS cell line (IPSC.DF.19.11), bone marrow derived cultured mesenchymal stem cells (STRM.MRW.MSC), primary B cells from peripheral blood (BLD.CD19.PPC), foreskin fibroblast primary cells skin01 (SKIN.PEN.FRSK.FIB.01), fetal muscle trunk (MUS.TRNK.FET), fetal muscle leg (MUS.LEG.FET), fetal stomach (GI.STMC.FET), psoas muscle (MUS.PSOAS), rectal mucosa donor 29 (GI.RECT.MUC.29), and HeLa-S3 cervical carcinoma cell line (CRVX.HELAS3.CNCR). As input we used MAiC nsCL/P variants with P≤0.001 without additional variants in LD (N=22,999); this threshold was selected to balance between adequate statistical power and true positive association signals. To test if the observed enrichment is specific for nsCL/P, we configured a control SNP set comprising an equal number of SNPs with *P*>0.1 from MAiC data, which were selected to represent the same MAF distribution as the test input.

Characterization of nsCL/P risk variants and candidate gene priorization in context of epigenetic mid-facial time line.

For comprehensive insights in regulatory mechanisms at nsCL/P risk loci, we finally integrated all available genetic and functional data (MAiC associations, $GWAS_{TAD}$- and r^2≥0.6-region boundaries, NCC- and CT-specific active chromatin sites, chromatin segmentation tracks and pCHi-C *cis* interactions). Based on this approach we attempt to prioritize genetic variants with regulatory effect and potential

downstream target genes and to detect relevant regulatory elements specific for the early (hNCC/cNCC) or later mid-facial development (CT). Chromatin signals that were either only predicted in individual or few cells/tissues were discarded as artefacts. Risk loci without interpretable chromatin signals at the associated region (either no chromatin signals or artefacts) and too complex risk loci (either with many overlapping genes or very broad associated regions that both makes it impossible to derive clear conclusions on regulatory effects on particular genes) were excluded from this part of functional interpretation.

Supplemental Figures

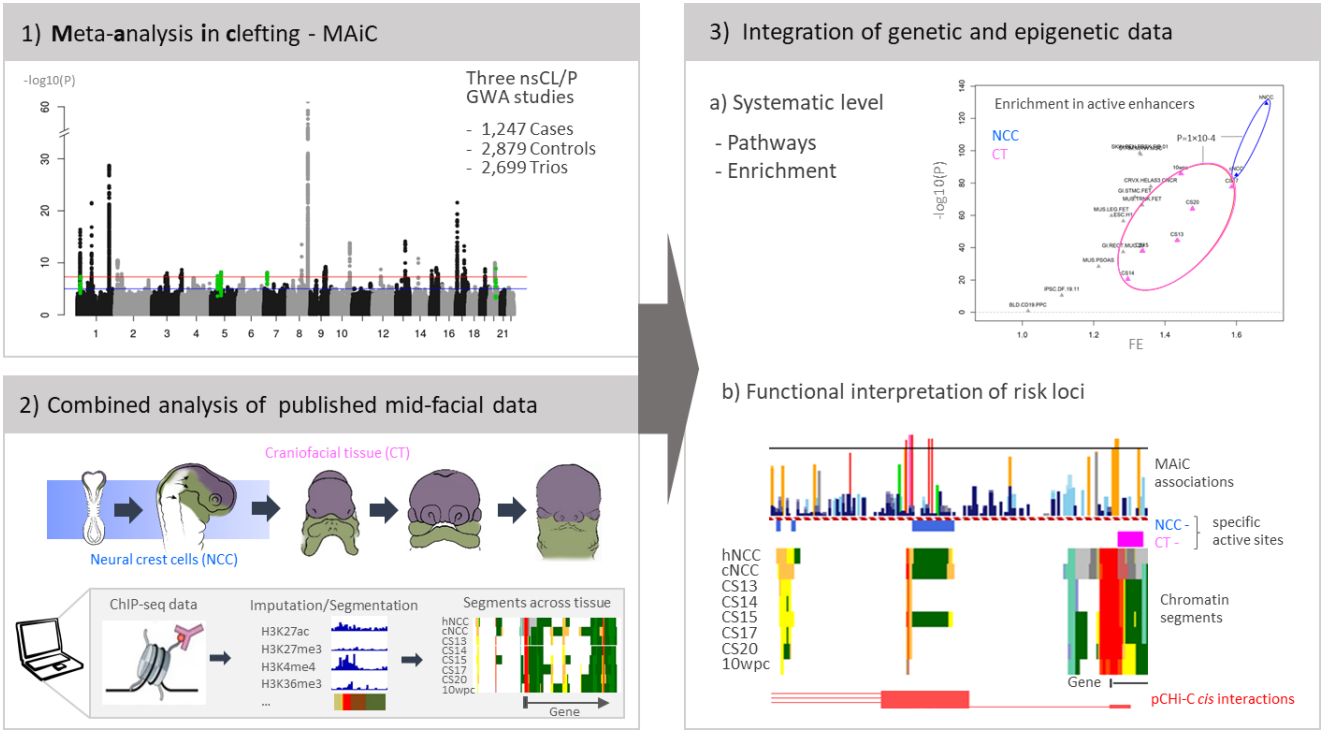

Figure S1 Graphical workflow of the study. **1)** Manhattan plot of summary statistics for MAiC in nonsyndromic cleft lip with/without cleft palate (nsCL/P). Blue and red lines indicate suggestive (at $-\log_{10}(1\times10^{-5})$) and genome-wide (at $-\log_{10}(5\times10^{-8})$) significance. **2)** Available epigenetic datasets for cells and tissue relevant for craniofacial development were processed in a joint bioinformatic pipeline to generate a comparable map of epigenetic data for the functional analysis of nsCL/P across mid-facial development. **3)** Systematic integration of MAiC association and epigenetic data to analyse the enrichment of genetic variants, to reveal relevant biological pathways and to study regulatory mechanisms at nsCL/P risk loci. GWA - Genome-wide association; ChIP-seq - Chromatin immunoprecipitation followed by sequencing; hNCC - human neural crest cell; cNCC - cranial NCC; CS - Carnegie stage; wpc - weeks *post-conceptum*; pCHi-C – Promoter capture chromosome capture.

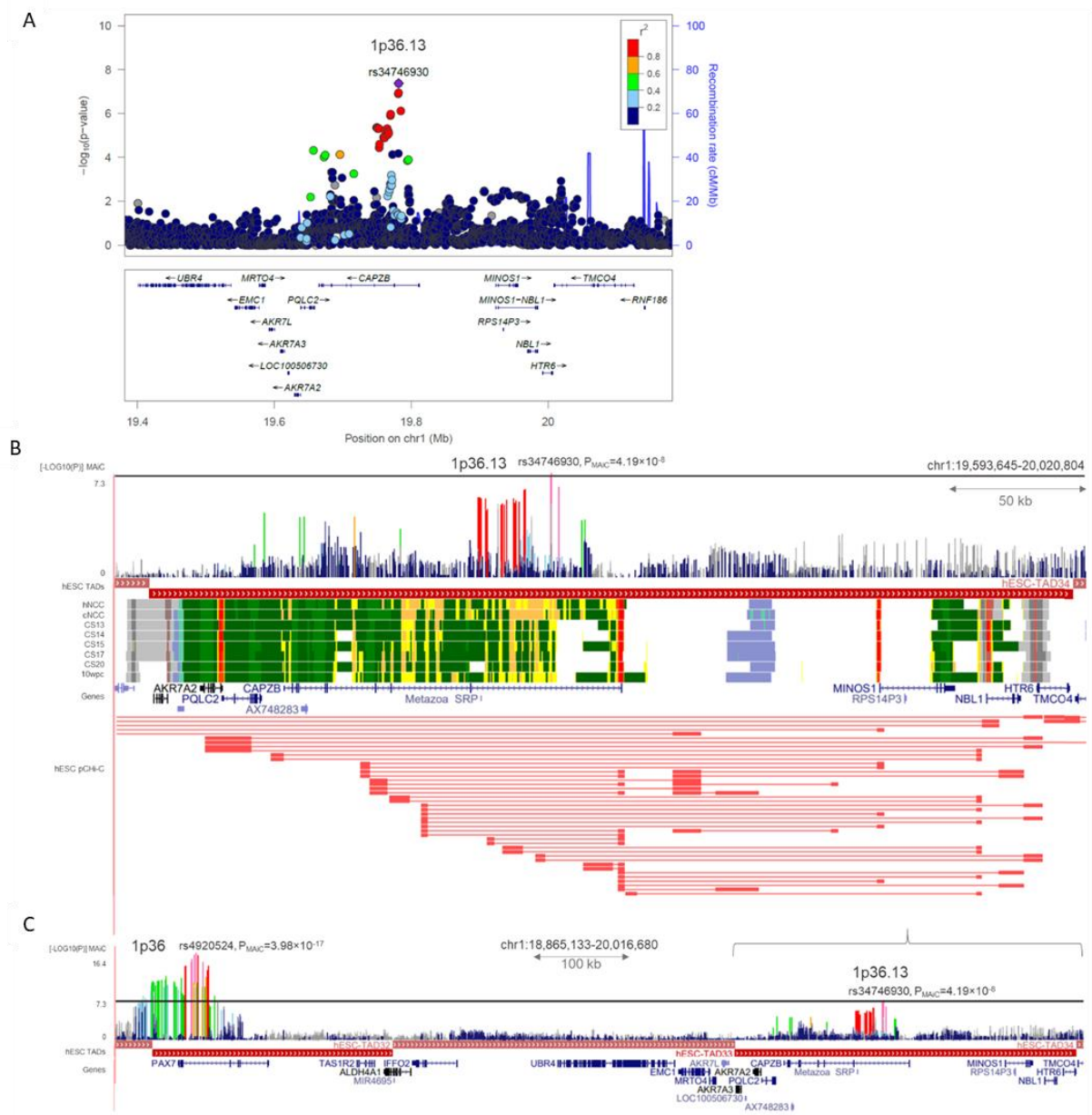

Figure S2 Regulatory architecture at novel risk locus for nonsyndromic cleft lip with/without cleft palate (nsCL/P) at 1p36.13. A) Regional association plot of chromosomal region 1p36.1 in vicinity of *CAPZB*, with lead variant rs34746930. Data from MAiC, plot generated with LocusZoom. **B)** Zoom into regulatory architecture at 1p36.1. Based on the extent of the topologically associated domain around rs34746930 and variants in linkage disequilibrium (LD) $r^2 \geq 0.6$, different layers of data were aggregated. Tracks include (top-down): MAiC association P-values with colour code based on LD to rs34746930; chromatin segmentation data from early human neural crest cell (hNCC), cranial NCC (cNCC), craniofacial tissue (CT) of different Carnegie stages (CS) and 10 weeks *post-conceptum* (wpc, colour code as in Figure 3 of main text); RefSeq gene positions; and promotor capture Hi-C *cis*-interactions collected in hESC (GSE8682133). **C)** Association structure at 1p36.13, indicating its independence from the known risk locus 1p36 around *PAX7*. Further information on the region is provided as Supplemental Text.

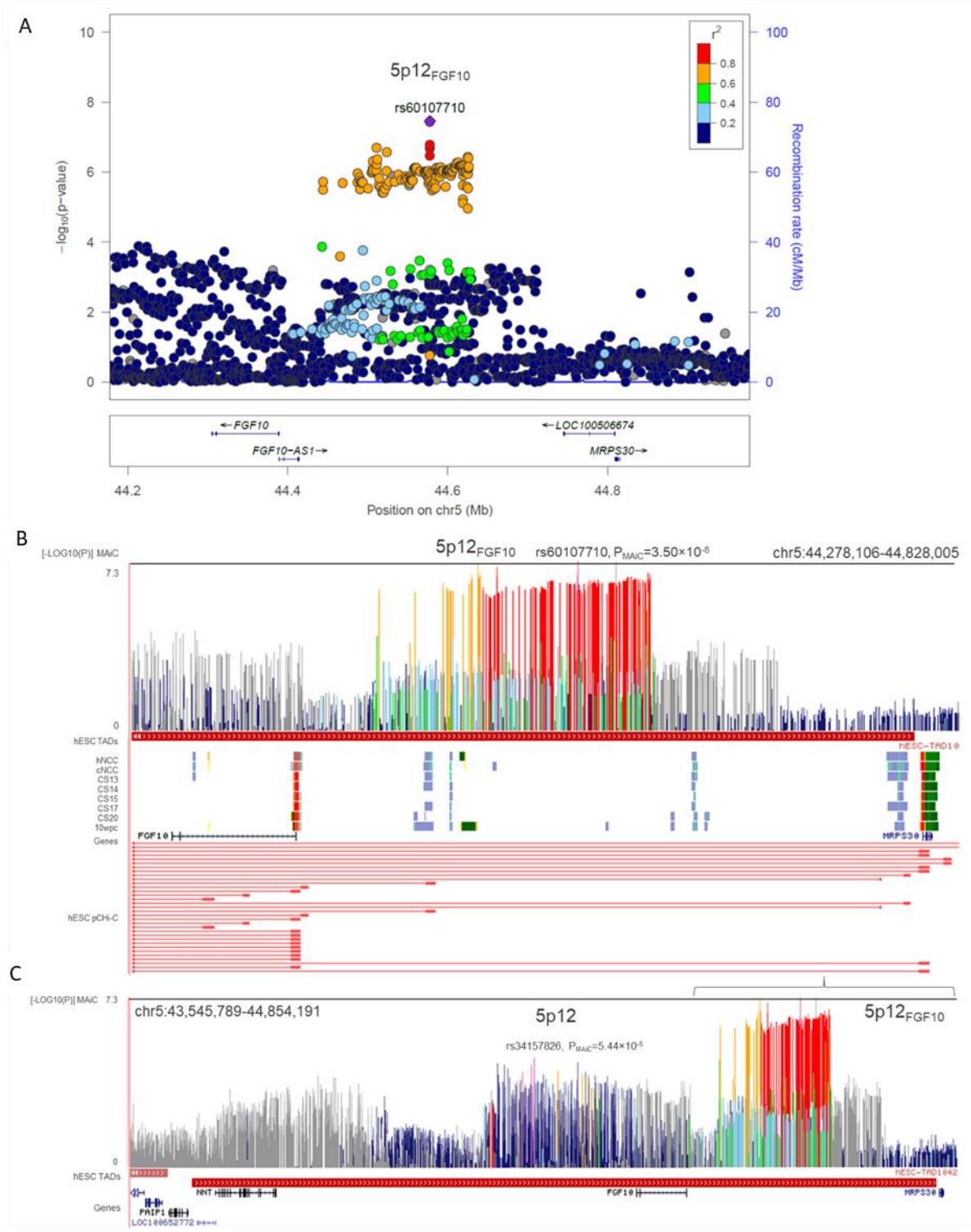

Figure S3 Regulatory architecture at novel risk locus for nonsyndromic cleft lip with/without cleft palate (nsCL/P) risk locus at 5p12$_{FGF10}$. A) Regional association plot of chromosomal region 5p12$_{FGF10}$ with lead variant rs60107710. Data from MAiC, plot generated with LocusZoom. **B)** Zoom into regulatory architecture at 5p12$_{FGF10}$. Based on the extent of the topologically associated domain around rs60107710 and variants in linkage disequilibrium (LD) $r^2 \geq 0.6$, different layers of data were aggregated. Tracks include (top-down): MAiC association P-values with colour code based on LD to rs60107710; chromatin segmentation data from early human neural crest cell (hNCC), cranial NCC (cNCC), craniofacial tissue (CT) of different Carnegie stages (CS) and 10 weeks *post-conceptum* (wpc, colour code as in Figure 3 of main text); RefSeq gene positions; and promotor capture Hi-C *cis*-interactions collected in hESC (GSE8682133). C**)** Association structure at 5p12$_{FGF10}$, indicating its independence from the known risk locus 5p12 previously reported in Chinese individuals. Further information on the region is provided as Supplemental Text.

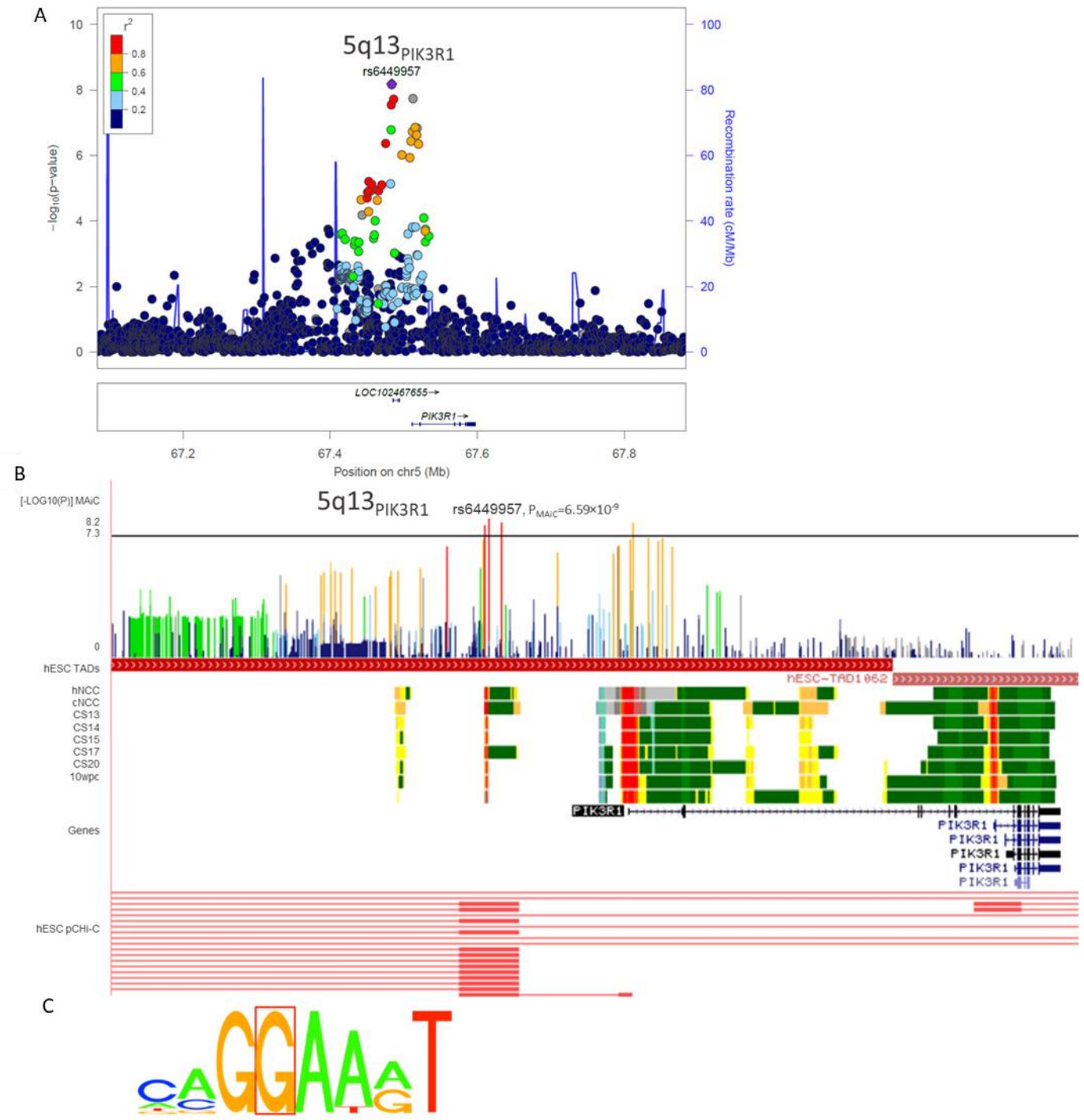

Figure S4 Regulatory architecture at novel risk locus for nonsyndromic cleft lip with/without cleft palate (nsCL/P) risk locus at 5q13$_{PIK3R1}$. A) Regional association plot of chromosomal region 5q13$_{PIK3R1}$ with lead variant rs6449957. Data from MAiC, plot generated with LocusZoom. **B)** Zoom into regulatory architecture at 5q13$_{PIK3R1}$. Based on the extent of the topologically associated domain around rs6449957 and variants in linkage disequilibrium (LD) $r^2 \geq 0.6$, different layers of data were aggregated. Tracks include (top-down): MAiC association P-values with colour code based on LD to rs6449957; chromatin segmentation data from early human neural crest cell (hNCC), cranial NCC (cNCC), craniofacial tissue (CT) of different Carnegie stages (CS) and 10 weeks *post-conceptum* (wpc, colour code as in Figure 3 of main text); RefSeq gene positions; and promotor capture Hi-C *cis*-interactions collected in hESC (GSE8682133). **C)** According to JASPAR 2018, rs6449956 (G>T), in high LD with rs6449957, is predicted to disrupt a binding motif for FEV, a member of the Ets-family of transcription factors. Further information on the region is provided as Supplemental Text.

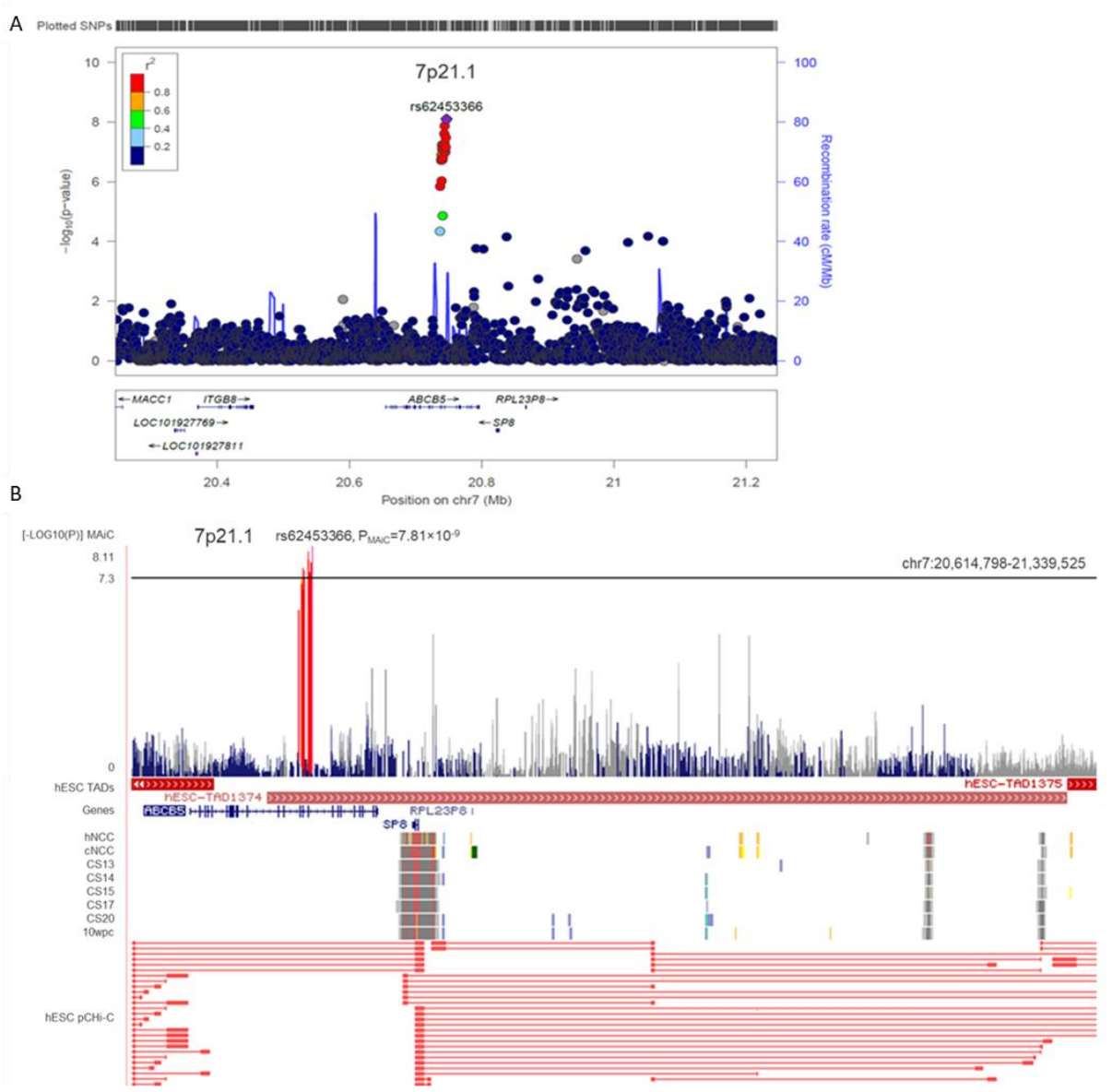

Figure S5 Regulatory architecture at novel risk locus for nonsyndromic cleft lip with/without cleft palate (nsCL/P) risk locus at 7p21.1. **A)** Regional association plot of chromosomal region 7p21.1 with lead variant rs62453366. Data from MAiC, plot generated with LocusZoom. **B)** Zoom into regulatory architecture at 7p21.1. Based on the extent of the topologically associated domain around rs62453366 and variants in linkage disequilibrium (LD) $r^2 \geq 0.6$, different layers of data were aggregated. Tracks include (top-down): MAiC association P-values with colour code based on LD to rs62453366; chromatin segmentation data from early human neural crest cell (hNCC), cranial NCC (cNCC), craniofacial tissue (CT) of different Carnegie stages (CS) and 10 weeks *post-conceptum* (wpc, colour code as in Figure 3 of main text); RefSeq gene positions; and promotor capture Hi-C *cis*-interactions collected in hESC (GSE8682133). Further information on the region is provided as Supplemental Text.

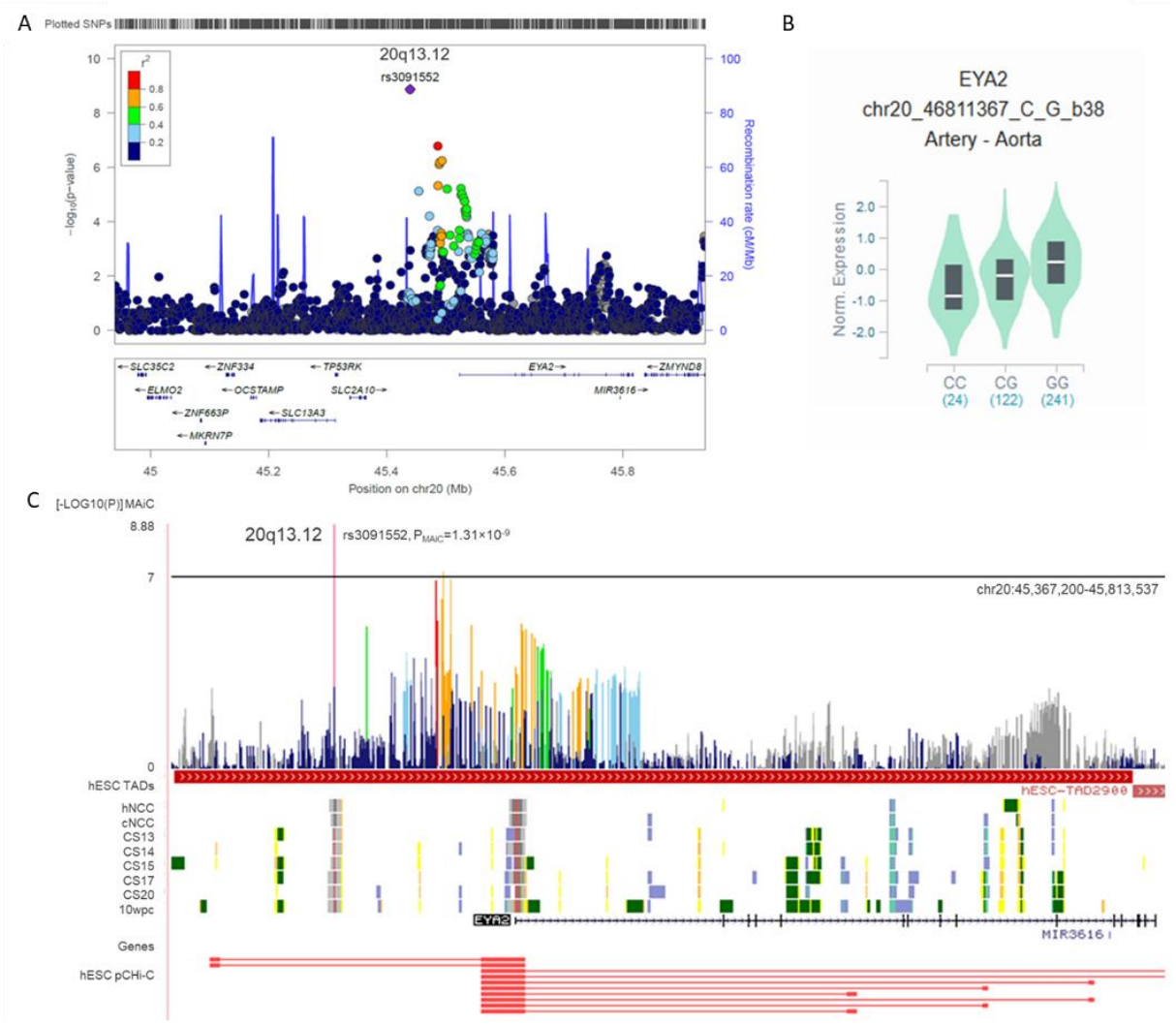

Figure S6 Regulatory architecture at novel risk locus for nonsyndromic cleft lip with/without cleft palate (nsCL/P) risk locus at 20q13.12. **A)** Regional association plot at chromosomal region 20q13.12, with lead variant rs3091552. Data from MAiC, plot generated with LocusZoom. **B)** According to GTEx(v8) data, rs3091552 is an eQTL for *EYA2* in artery/aorta tissue, the risk allele being. **C)** Zoom into regulatory architecture at 20q13.12. Based on the extent of the topologically associated domain around rs3091552 and variants in linkage disequilibrium (LD) $r^2 \geq 0.6$, different layers of data were aggregated. Tracks include (top-down): MAiC association P-values with colour code based on LD to rs3091552; chromatin segmentation data from early human neural crest cell (hNCC), cranial NCC (cNCC), craniofacial tissue (CT) of different Carnegie stages (CS) and 10 weeks *post-conceptum* (wpc, colour code as in Figure 3 of main text); RefSeq gene positions; and promotor capture Hi-C *cis*-interactions collected in hESC (GSE8682133). Further information on the region is provided as Supplemental Text.

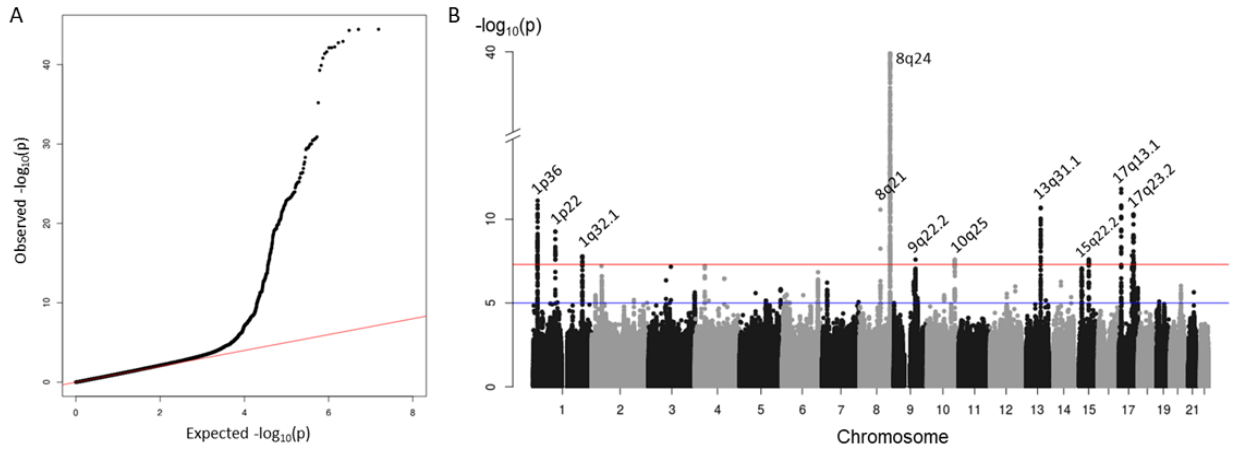

Figure S7 M**eta-analysis in clefting (MAiC) performed in the European cohorts. A)** In the quantile-quantile-plot, the number and magnitude of observed associations between single variants and nonsyndromic cleft lip with or without cleft palate (nsCL/P) is compared to the association statistics expected under the null hypothesis of no association. The lambda value is 1.04. **B)** In the Manhattan plot, associations of genetic variants across all autosomes and nsCL/P are plotted against the -log10 transformed P-values of MAiC. Level of suggestive significance is highlighted in blue at -log10(1×10^{-5}), and genome-wide significance in red at -log10(5×10^{-8}). N=716 SNPs at ten nsCL/P risk loci were detected as genome-wide significant, the most significant locus being the established 8q24 nsCL/P risk locus (Birnbaum et al. 2009). The SNP with the lowest P-value in $MAiC_{euro}$ is rs72728734 (chr8:129,933,720, P=3.58×10^{-45}).

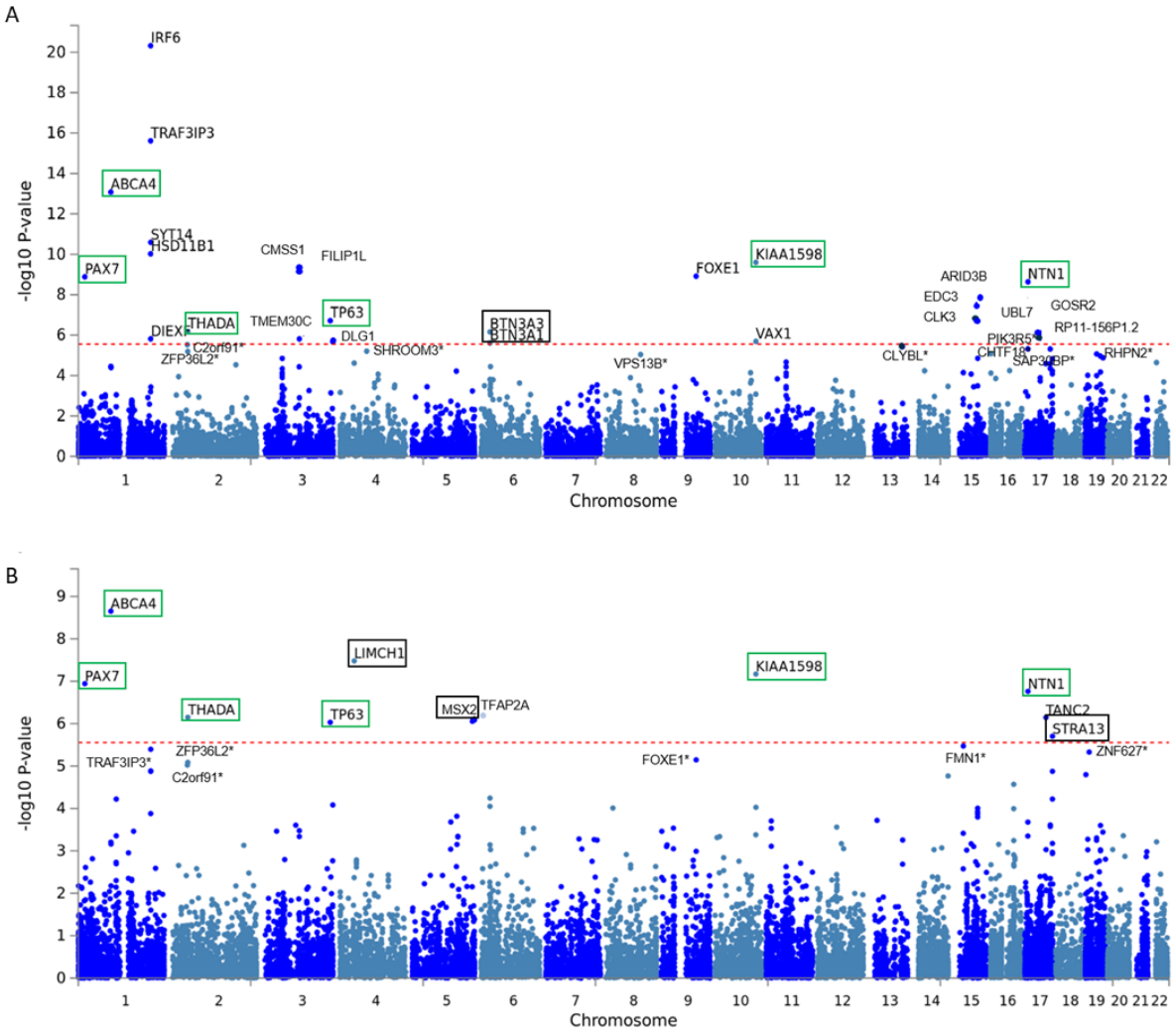

Figure S8 Gene-based Manhattan plot for the multi-ethnic analysis (A) and the European cohort (B). MAiC summary statistics for n=7,744,527 (MAiC) / n=7,690,843 ($MAiC_{Euro}$) SNPs were mapped to 17,921 protein coding genes based to a distance of 0kb upstream/downstream of the genes. Of those, nominally significant gene-based associations (P<0.05) were obtained for 1,358 (MAiC) and 1,222 ($MAiC_{Euro}$) genes, respectively. Test-wide significance (indicated by red dashed line) was defined at $P=2.79\times10^{-6}$ (Online Methods). Twenty-five (MAiC) and eleven ($MAiC_{Euro}$) genes, respectively, reached test-wide significance. Additionally, nine (MAiC) and six ($MAiC_{Euro}$) genes were significant at suggestive level ($2.79\times10^{-6}<P<10^{-5}$, highlighted by '*'). Six genes (*ABCA4*, *KIAA1598*, *PAX7*, *NTN1*, *TP63* and *THADA*; highlighted in green boxes) were identified with test-wide significance in both analyses. Across both analyses, five genes were identified at test-wide significance (*BTN3A3*, *BTN3A1; LIMCH1*, *MSX2* and *STRA13*; highlighted in black boxes) which do not map to any of the known nsCL/P risk loci. Data generated in FUMA.

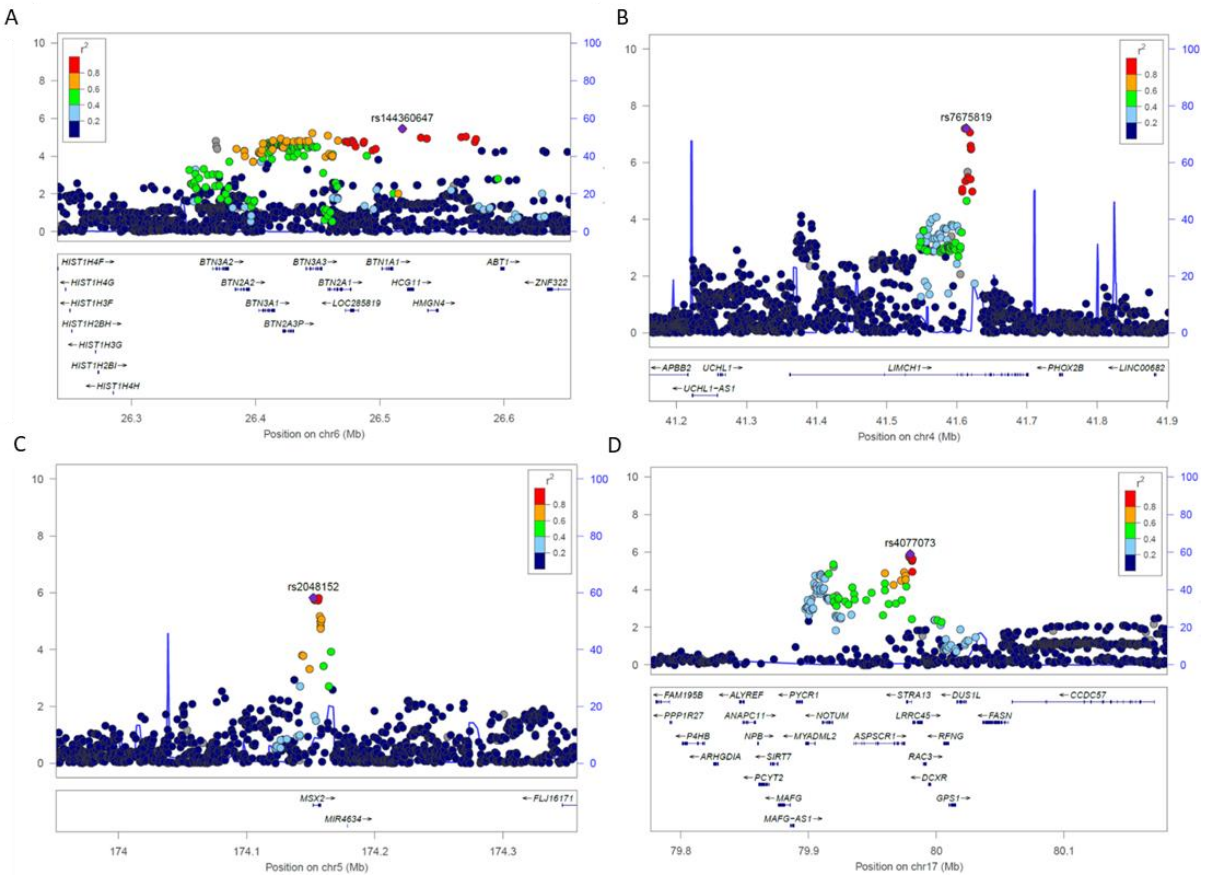

Figure S9 Regional association plots of genes from gene-based analysis which had not been previously reported. Genes were detected in gene-based analysis as implemented in FUMA, applied to MAiC and MAiC$_{Euro}$ summary statistics. Input SNPs were mapped to 17,911 protein coding genes based to a distance of 0kb upstream/downstream. **(A)** Gene-based analysis in MAiC revealed two test-wide significant genes *BTN3A3* ($P=6.96\times10^{-7}$) and *BTN3A1* ($P=2.44\times10^{-6}$). This region does not map to any of the known nsCL/P risk loci. **(B-D)** In MAiC$_{Euro}$ three test-wide significant genes - *LIMCH1* ($P=3.31\times10^{-8}$), *MSX2* ($P=8.80\times10^{-7}$) and *STRA13* ($P=1.99\times10^{-6}$) were identified outside of the established risk loci for nsCL/P. Plots were generated in LocusZoom.

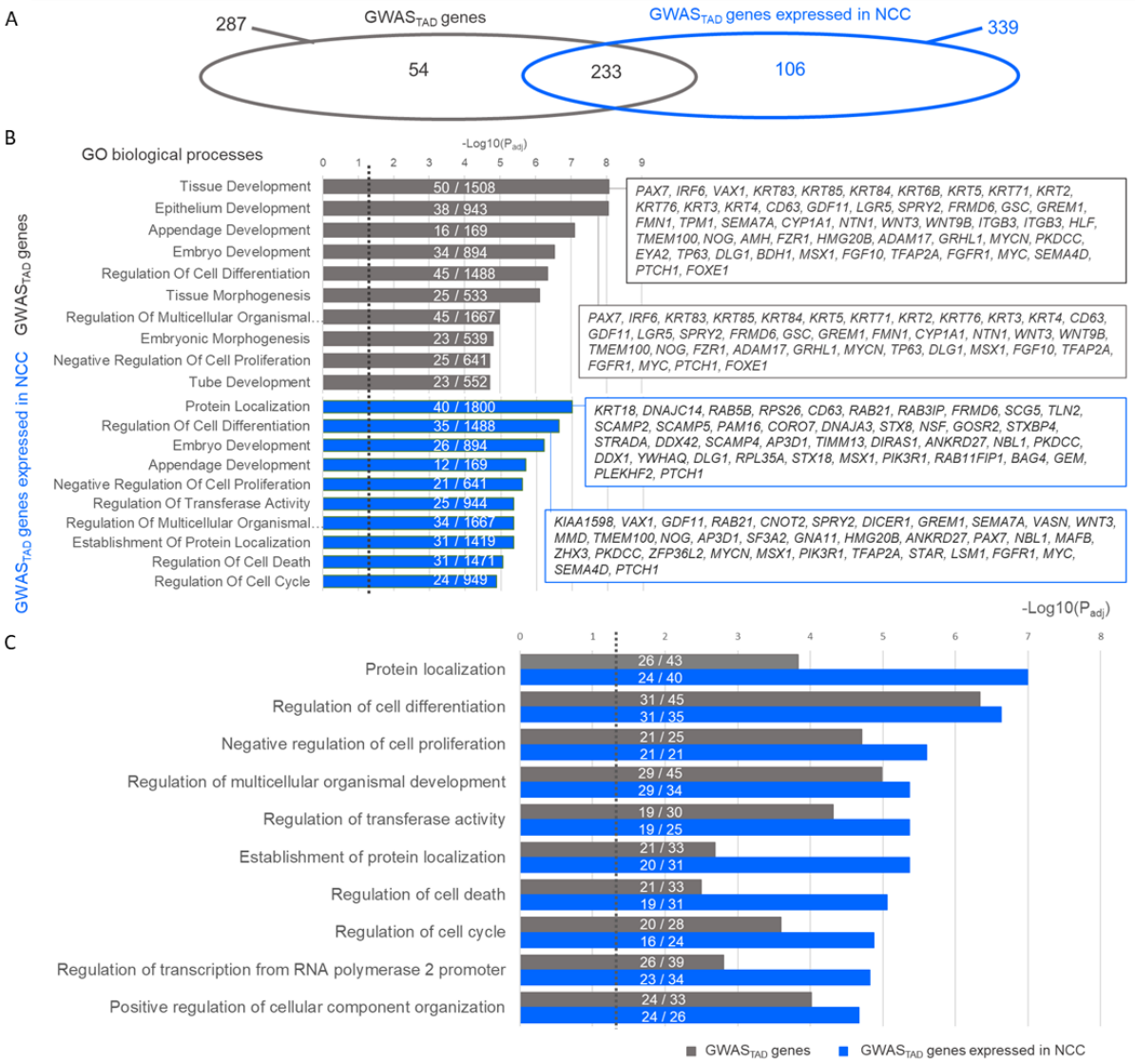

Figure S10 Gene ontology (GO) analysis of 'biological processes' in MAiC data. For each of the 45 nsCL/P risk loci, genes located within the corresponding topological associating domain (TAD) regions were extracted (GWAS$_{TAD}$-genes). This set of 407 genes was cross-referenced with expression data from neural crest cells (NCC; Laugsch *et al.* (2018) (GSE108522)), revealing expression of 240 GWAS$_{TAD}$ genes (GWAS$_{TAD}$-genes expressed in NCC). Using the 'GENE2FUNC' application of FUMA (v1.3.4b), enrichment analysis of both gene sets was performed. **A)** Number of significant GO-terms ($P_{adj} \leq 0.05$) in both analyses (n=287 and 339, respectively), with their overlap (n=233) indicated. **B)** Top10 GO biological processes of the individual analyses for 'GWAS$_{TAD}$' (gray) and 'GWAS$_{TAD}$ genes expressed in NCC' (blue). Dashed line indicates the significance threshold of $P_{adj}=0.05$. Within the bars, the numbers of nsCL/P risk loci / genes represented in this pathway are provided. **C)** Across both analyses, 233 pathways were shared. Of those, 157 had lower P-values in the subset of 'GWAS$_{TAD}$ genes expressed in NCC'. Here, the top10 GO biological processes are shown.

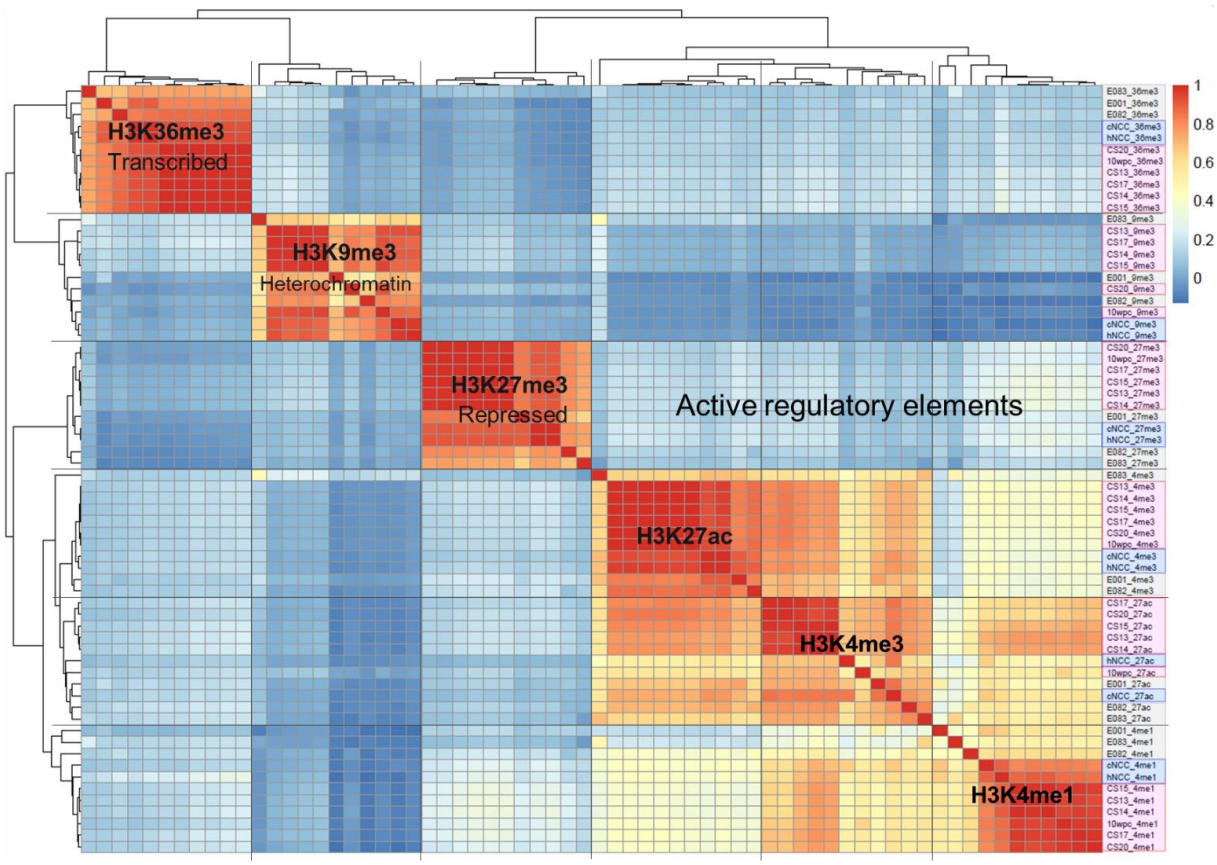

Figure S11 Chromatin modifications in mid-facial development. Hierarchical clustering of pairwise Pearson correlations of epigenetic data. ChIP-seq signals of six histone modifications obtained in human neural crest cells (hNCC); cranial NCC (cNCC, both highlighted in blue); six craniofacial tissue samples (Carnegie stage (CS) 13, CS14, CS15, CS17, CS20, 10 weeks *post-conceptum*, all highlighted in red); and three Roadmap samples (embryonal stem cells (ESC) I3, fetal brain, fetal heart, highlighted in gray).

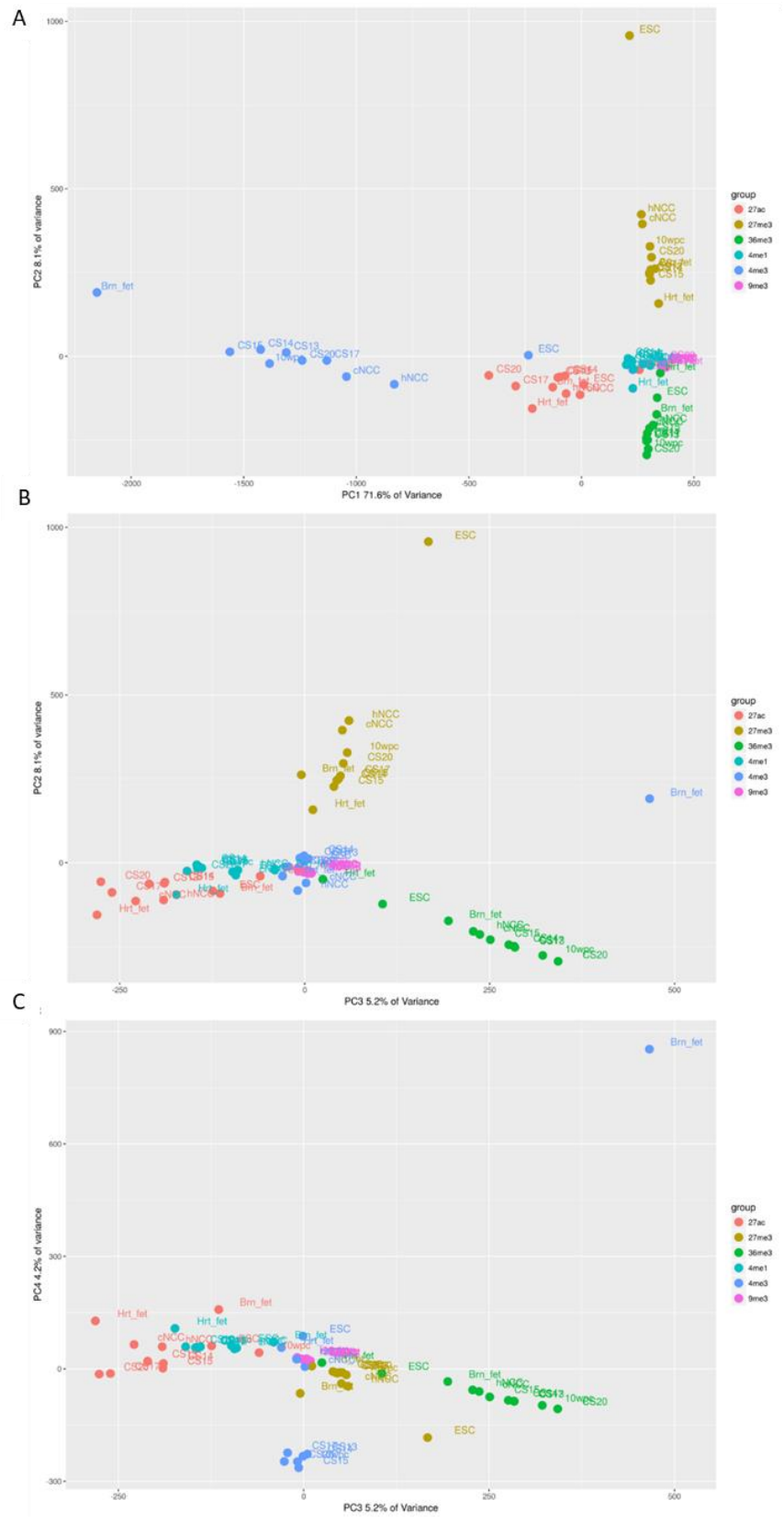

Figure S12 Principal component analysis plot of all imputed chromatin mark signals in neural crest cells (NCC), craniofacial tissue of different Carnegie stages (CS) and selected Roadmap samples. Projection of first vs. second **(A)**, second vs. third **(B)** and third vs fourth **(C)** principal component (PC) as analyses based on genome-wide signal profiles of H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3 and H3K9me3 in early human neural crest cells (hNCC), cranial NCC (cNCC), craniofacial tissue and Roadmap samples of ESC (E001), fetal heart (E083) and fetal brain (E082). Samples are color-coded by chromatin mark. Percentages of variance explained by each PC are indicated along each axis.

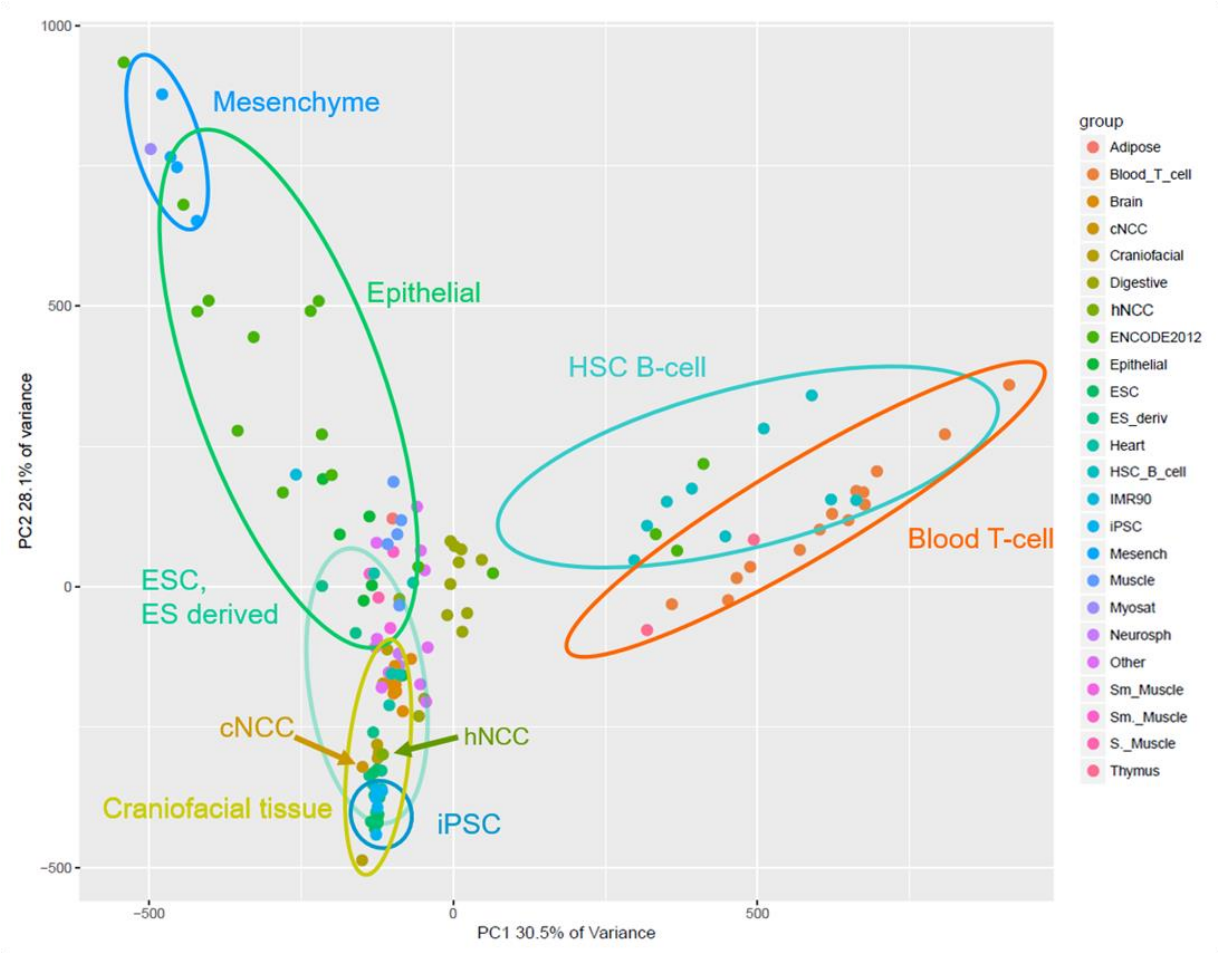

Figure S13 Principal component analysis (PCA) plot based on genome-wide H3K27ac signals in early human neural crest cells (hNCC), cranial NCC (cNCC), craniofacial tissue of different Carnegie stages (CS) all Roadmap samples based on chromatin mark H3K27ac. PCA projection shows the the first and second component dimensions for genome-wide signal profiles of H3K27ac in hNCC, cNCC, craniofacial tissue and all 127 Roadmap/ENCODE samples. Samples are grouped and colour coded as indicated in the legend. Percentages of variance across samples explained by each component are indicated along each axis.

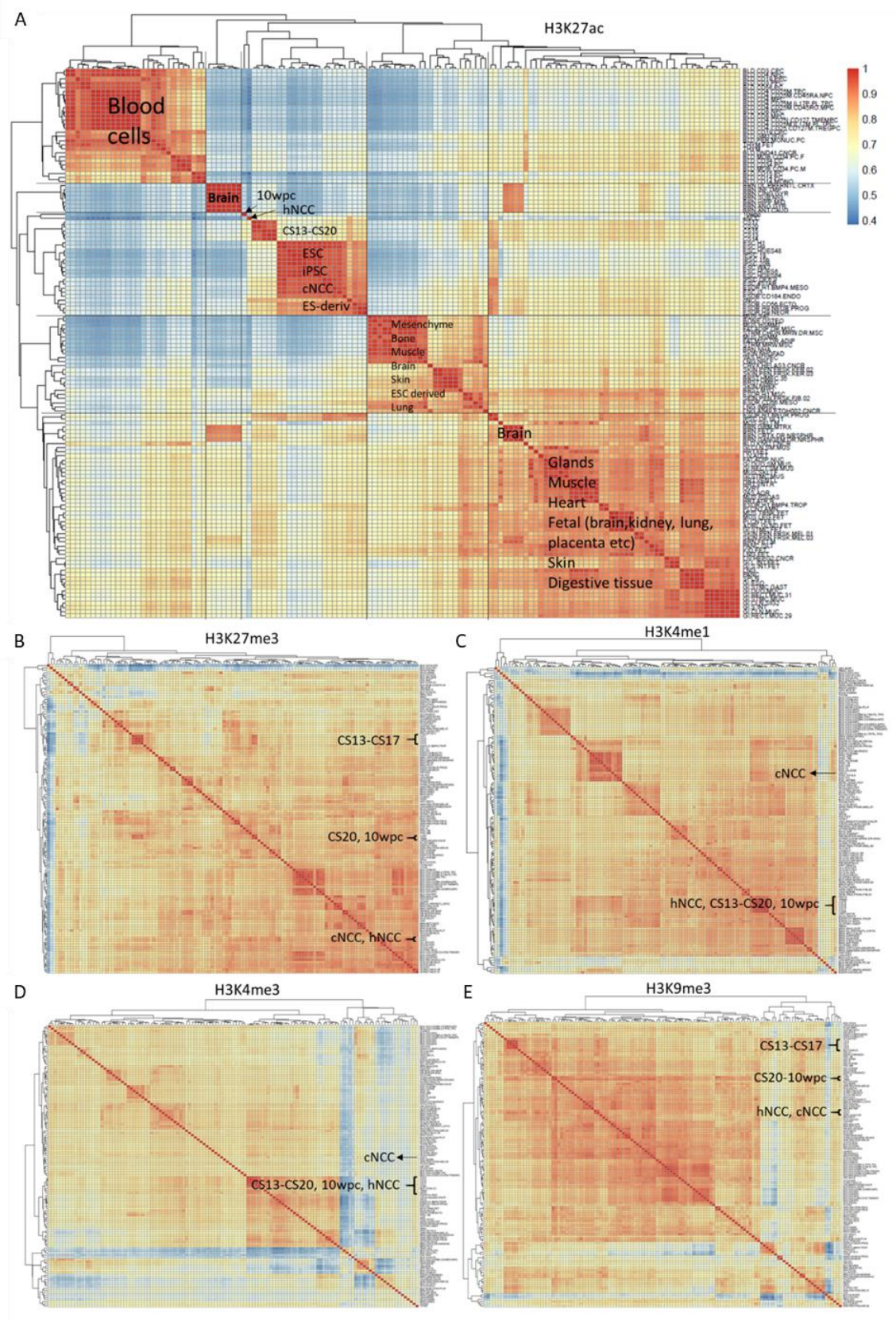

Figure S14 Heatmap and hierarchical clustering of pairwise Pearson correlations. (A) H3K27ac, **(B)** H3K27me3, **(C)** H3K4me1, **(D)** H3K4me3 and **(E)** H3K9me3 signals. For each chromatin modification, the heatmap was generated based on the respective genome-wide ChIP-Seq signals measured in early human neural crest cells (NCC), cranial NCC, six craniofacial tissue samples (Carnegie stage (CS) 13, CS14, CS15, CS17, CS20, 10 weeks *post-conceptum*) and all 127 Roadmap/ENCODE samples. Relatedness of epigenomic profiles by sample is indicated by dendrogram along the axes of the heatmap. Red indicates positive correlation between datasets. The underlying signal comparisons were calculated with deepTools 3.1.3 multiBigwigSummary in bins mode.

Figure S15 Cumulative percentage of chromatin segments on autosomes in neural crest cells (NCC), craniofacial tissue and selected Roadmap samples. Based on the 18-state model **(A)**, data was aggregated into eight states to increase robustness of the analyses **(B)**. For each chromatin state, fractions were calculated in human NCC (hNCC), cranial NCC (cNCC), six craniofacial tissue samples (Carnegie stage (CS) 13, CS14, CS15, CS17, CS20, 10 weeks *post-conceptum*) and a selection of segmentations generated by Roadmap Epigenome (Roadmap Epigenomics Consortium et al., 2015). Color code as presented in the legends, abbreviation of chromatin states as listed in Main Text.

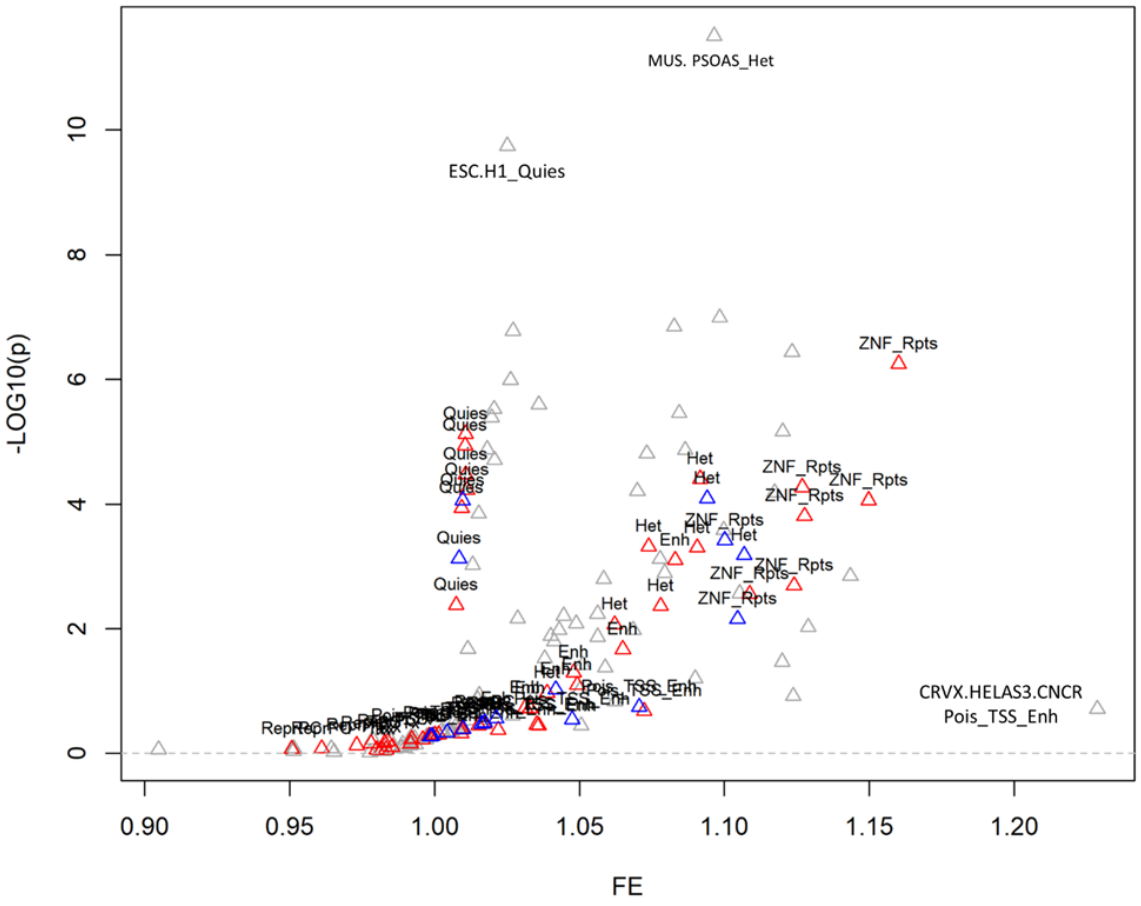

Figure S16 Enrichment analysis of a control SNP set in different chromatin states. Based on eight chromatin states retrieved in two neural crest cell (NCC), six craniofacial tissue (CT) and eleven Roadmap samples, the enrichment of n=22,999 SNPs (P_{MAiC}>0.1, matched for allele frequency distribution) was calculated using GREGOR (Schmidt et al. 2015). Roadmap samples (in gray) included three fetal (fetal muscle trunk (MUS.TRNK.FET), fetal muscle leg (MUS.LEG.FET), fetal stomach (GI.STMC.FET) and eight non-fetal samples (ESC H1 cell line (ESC.H1), iPS cell line (IPSC.DF.19.11), bone marrow derived cultured mesenchymal stem cells (STRM.MRW.MSC), primary B cells from peripheral blood (BLD.CD19.PPC), foreskin fibroblast primary cells skin01 (SKIN.PEN.FRSK.FIB.01), psoas muscle (MUS.PSOAS), rectal mucosa donor 29 (GI.RECT.MUC.29), and heLa-S3 cervical carcinoma cell line (CRVX.HELAS3.CNCR)). Enrichment of NCC (blue) and CT (red) samples is indicated by corresponding chromatin states. Roadmap samples with highest -log10(p) and log2(FE) are annotated by tissue type and chromatin state. The chromatin state transcription starting site (TSS) includes both active TSS and upstream/downstream flanking TSS, and the enhancers (Enh) include active and genic enhancers; FE - fold enrichment; ZNF_Rpts - Zink-finger genes and repeats; Het - heterochromatin; Pois_TSS_Enh - poised enhancers and bivalent TSS. This Figure complements Figure 3b of the Main Text.

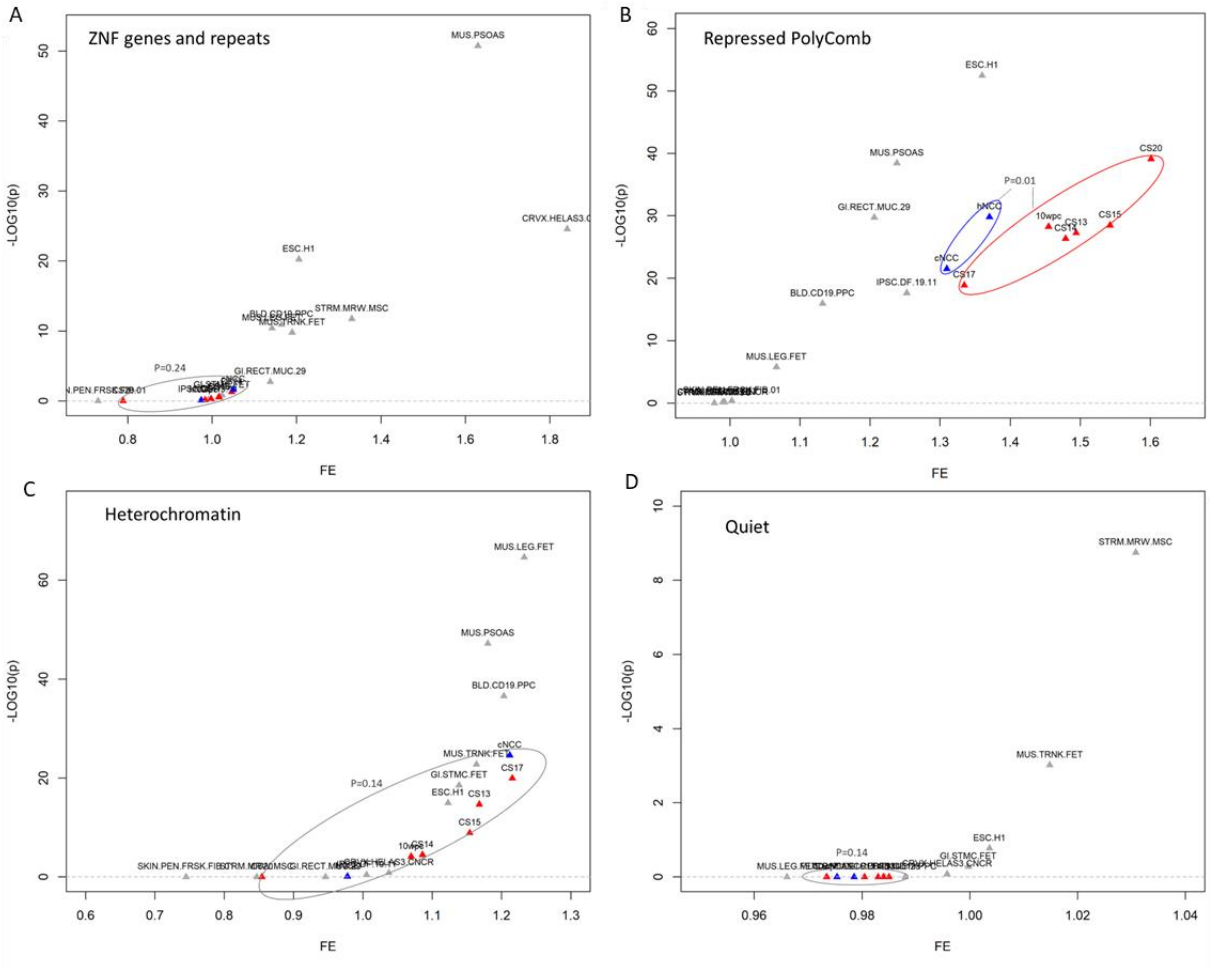

Figure S17 Association of meta-analysis in clefting (MAiC) data for four chromatin states. Based on eight chromatin states retrieved in two neural crest cell (NCC, blue), six craniofacial tissue (CT, pink) and eleven Roadmap samples (gray), enrichment analyses were performed for MAiC risk variants, at $P_{MAiC} \leq 0.001$ (n=22,999). Individual enrichment results for MAiC risk variants in four chromatin states. P-values represent difference in enrichment between NCC and CT. **A)** ZNF genes and repeats, **B)** repressed polycomb, **C)** heterochromatin, and **D)** quiescent regions. Abbreviations of tissues as provided by Roadmap. This Figure complements Figures 3c-3f of the Main Text. FE - fold enrichment; ZNF - zinc finger

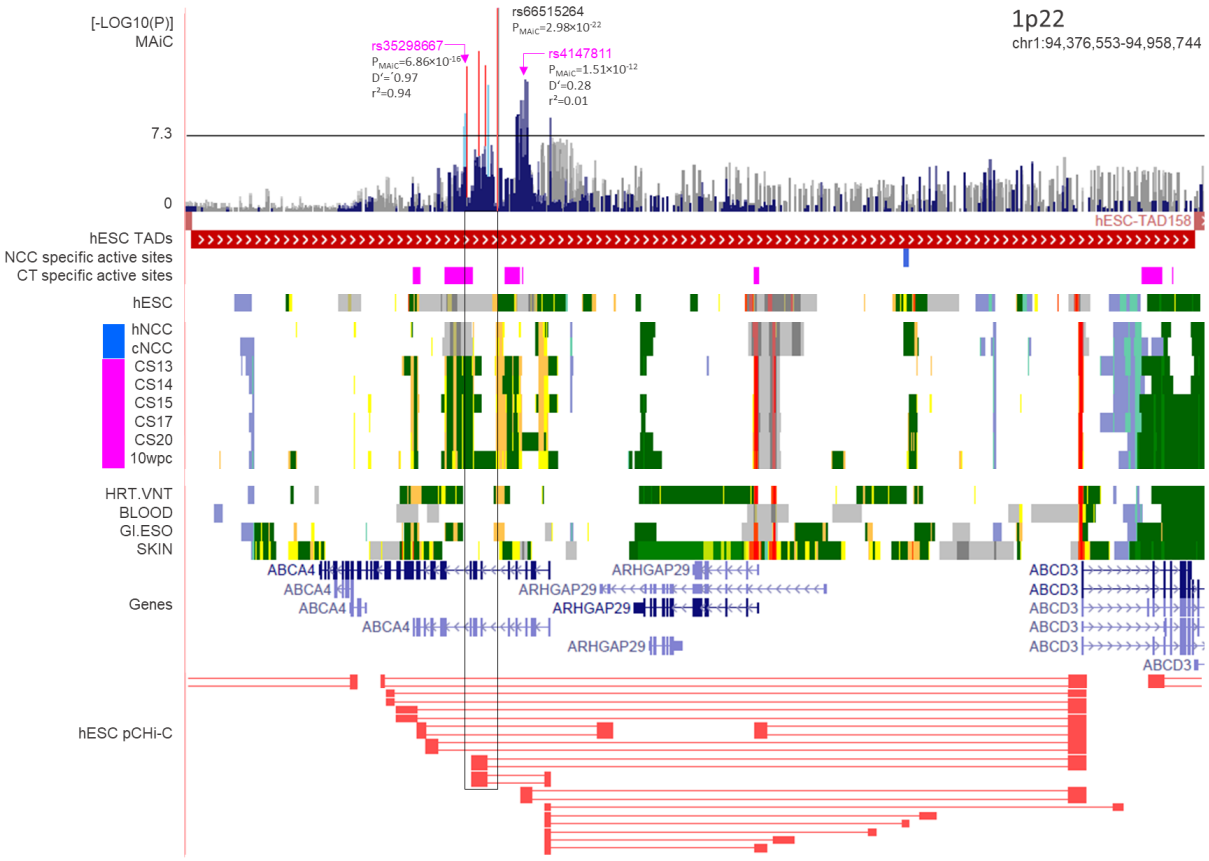

Figure S18 Regulatory architecture at nsCL/P risk locus 1p22. Based on the extent of the topologically associated domain around rs66515264 and variants in linkage disequilibrium (LD) $r^2 \geq 0.6$, different layers of data were aggregated. Tracks include (top-down): MAiC association P-values with colour code based on LD to rs66515264; chromatin segmentation data from early human neural crest cell (hNCC), cranial NCC (cNCC), craniofacial tissue (CT) of different Carnegie stages (CS) and 10 weeks *post-conceptum* (wpc, colour code as in Figure 3 of main text); RefSeq gene positions; and promotor capture (pC) Hi-C cis-interactions collected in hESC (GSE8682133). Black box highlights region of strongest association that contains craniofacial-tissue specific active sites and 3D connections to the promoter of *ABCD3*, which is indicated as active.

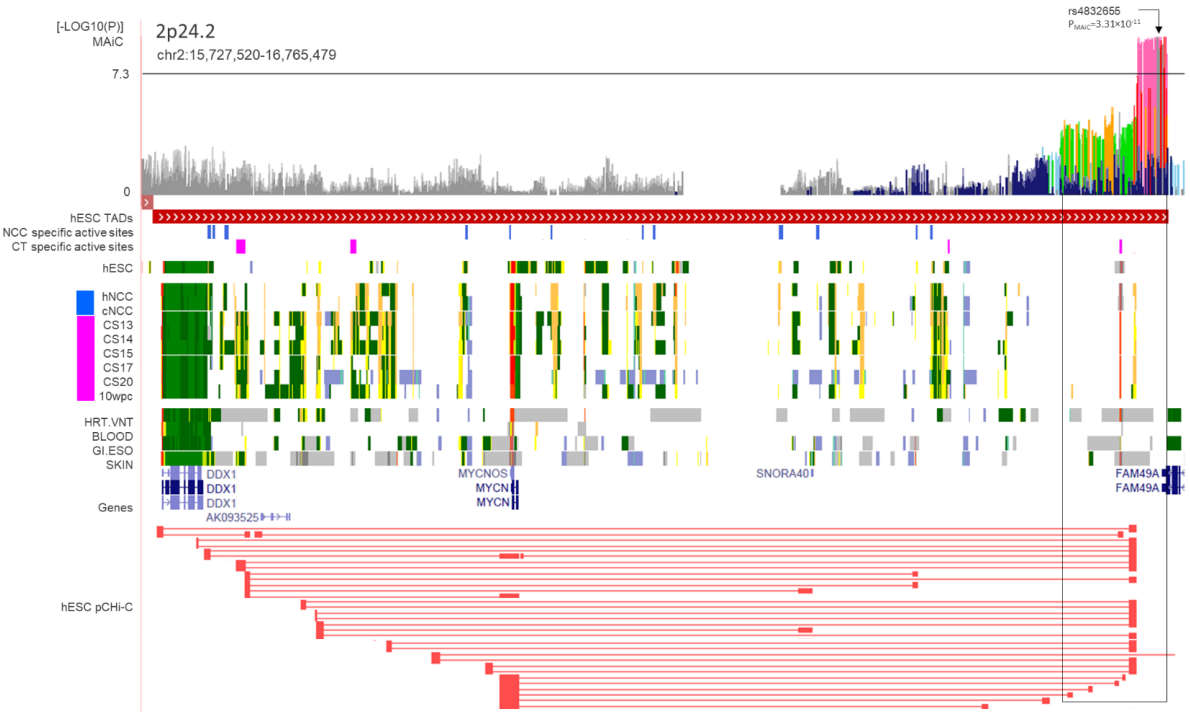

Figure S19 Regulatory architecture at 2p24.2. Based on the extent of the topologically associated domain (TAD) around rs4832655 and variants in linkage disequilibrium (LD) $r^2 \geq 0.6$, different layers of data were aggregated. Tracks include (top-down): MAiC association P-values with colour code based on LD to rs4832655; chromatin segmentation data from early human neural crest cell (hNCC), cranial NCC (cNCC), craniofacial tissue (CT) of different Carnegie stages (CS) and 10 weeks *post-conceptum* (wpc, colour code as in Figure 3 of main text); RefSeq gene positions; and promotor capture Hi-C *cis*-interactions collected in hESC (GSE8682133). Black box highlights region of strongest association that contains SNPs in strong LD to rs4832655 which map to an active region across midfacial development (orange, right) and an active site predominantly in craniofacial tissue (red, left). The presumed enhancer region interacts with diverse genes within the TAD, including *MYCN* and *DDX1*.

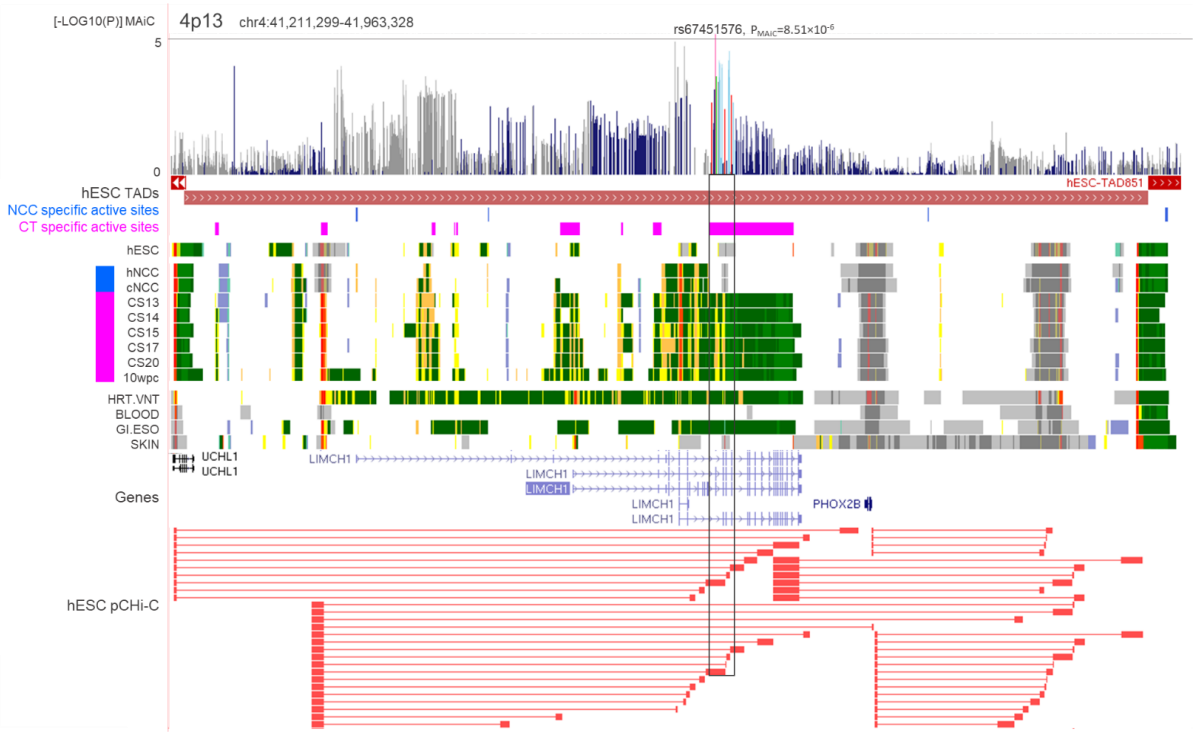

Figure S20 Regulatory architecture at 4p13. Based on the extent of the topologically associated domain (TAD) around rs67451576 and variants in linkage disequilibrium (LD) $r^2 \geq 0.6$, different layers of data were aggregated. Tracks include (top-down): MAiC association P-values with colour code based on LD to rs67451576; chromatin segmentation data from early human neural crest cell (hNCC), cranial NCC (cNCC), craniofacial tissue (CT) of different Carnegie stages (CS) and 10 weeks *post-conceptum* (wpc, colour code as in Figure 3 of main text); RefSeq gene positions; and promotor capture (pC) Hi-C cis-interactions collected in hESC (GSE8682133). Black box highlights region of strongest association, which maps to a region of strong craniofacial-tissue specific activity. Chromosomal interactions are indicated to a yet un-characterised genetic region upstream of the *LIMCH* promoter. Of note is also the lack of expression for the 3'-part of the *LIMCH1* gene in NCC, suggesting the presence of specific isoforms, whose role will have to be further investigated.

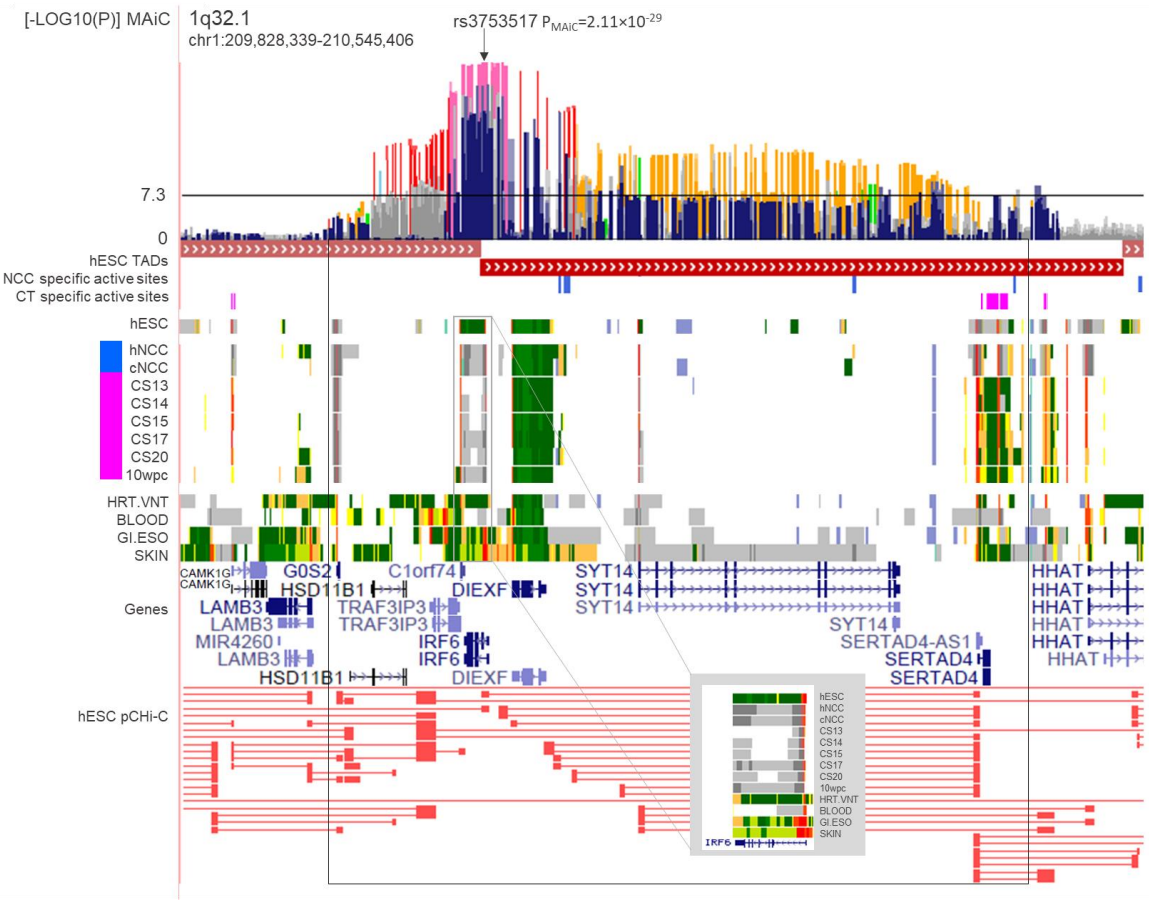

Figure S21 Complex regulatory architecture at 1q32.1. Based on the extent of the topologically associated domain (TAD) around rs3753517 and variants in linkage disequilibrium (LD) $r^2 \geq 0.6$, different layers of data were aggregated. Tracks include (top-down): MAiC association P-values with colour code based on LD to rs3753517; chromatin segmentation data from early human neural crest cell (hNCC), cranial NCC (cNCC), craniofacial tissue (CT) of different Carnegie stages (CS) and 10 weeks *post-conceptum* (wpc, colour code as in Figure 3 of main text); RefSeq gene positions; and promotor capture Hi-C *cis*-interactions collected in hESC (GSE8682133). The region of strongest association maps upstream of the *IRF6* promoter, encompassing a previously identified causal element (Rahimov et al. 2008). Notably, this putative enhancer region is poised in both NCC and CT, which matches the signals of the *IRF6* coding region (grey box). This is likely due to the established function of IRF6 in periderm / epithelial lineages, which are underrepresented in NCC and CT.

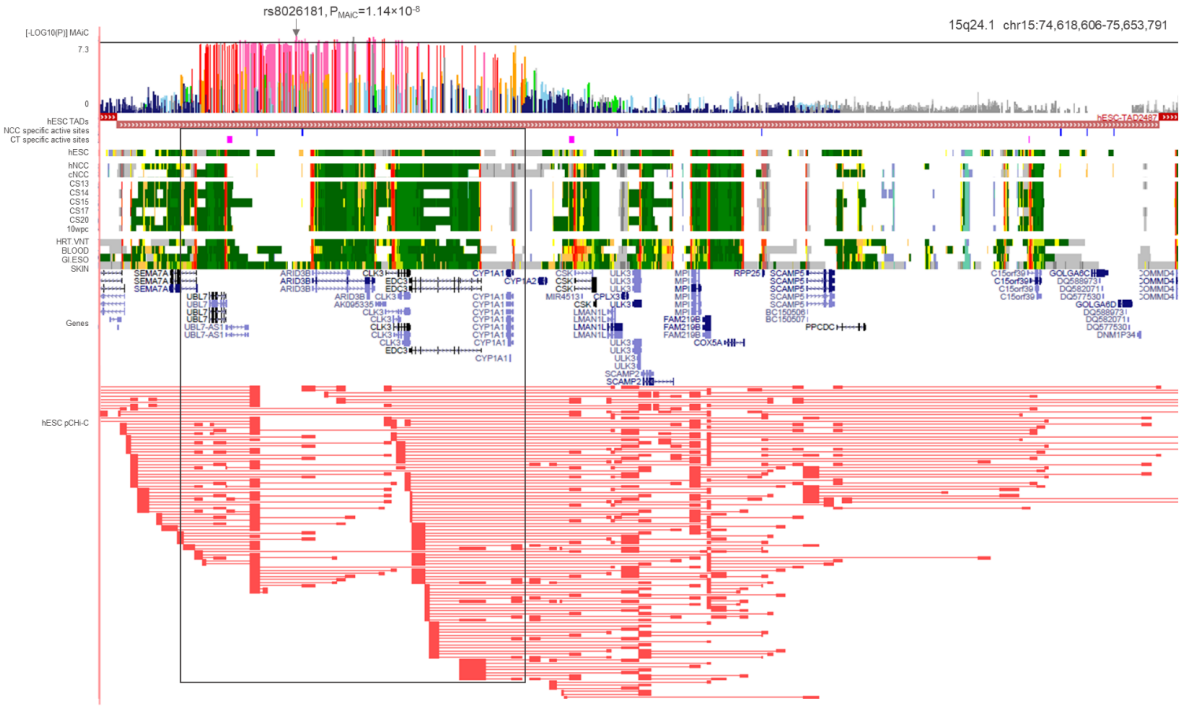

Figure S22 Complex regulatory architecture at 15q24.1. Based on the extent of the topologically associated domain (TAD) around rs8026181 and variants in linkage disequilibrium (LD) $r^2 \geq 0.6$, different layers of data were aggregated. Tracks include (top-down): MAiC association P-values with colour code based on LD to rs8026181; chromatin segmentation data from early human neural crest cell (hNCC), cranial NCC (cNCC), craniofacial tissue (CT) of different Carnegie stages (CS) and 10 weeks *post-conceptum* (wpc, colour code as in Figure 3 of main text); RefSeq gene positions; and promotor capture Hi-C *cis*-interactions collected in hESC (GSE8682133).

Supplemental Tables:

This document contains Tables S1, S3, S5 and S15. All other Supplemental Tables can be found in the Spreadsheet.

Table S1 Overview of individual cohorts used in our meta-analysis in clefting (MAiC).

GWAS (PubMed ID)	Cohort structure	Ethnicity	N individuals (case/control, in trios)	N_{eff} for MAiC ($MAiC_{Euro}$)
Bonn (20023658)	case/control	European	399/1318	1225
GENEVA (20436469)	case-parent trios	European	666	2992 (1332)
		Asian	795	
POFC (27033726)	case/control	European	163/733	2198 (533)
		Asian/Latin American	685/828	
	case-parent trios	European Asian/Latin American	289 949	2476 (578)

Three GWAS cohorts comprise four independent sub-cohorts. Abbreviations: N = number after removal of individuals based on inter-individual relationship, N_{eff} = effective number of individuals, calculated as described in "Statistical analysis".

Table S3 Antibodies used in ChIP-Seq experiments and information about imputation.

<table>
<tr><th>Cell/tissue[a]</th><th>H3K27ac</th><th>H3K4me1</th><th>H3K4me3</th><th>H3K27me3</th><th>H3K9me3</th><th>H3K36me3</th><th>Reference</th></tr>
<tr><td>Prediction</td><td>Enhancer</td><td>Enhancer</td><td>Promotor activating</td><td>Repressed</td><td>Hetero-chromatin</td><td>Active transcription</td><td></td></tr>
<tr><td>hNCC</td><td>ab4729, Abcam</td><td rowspan="4">ab8895, Abcam</td><td rowspan="3">39159, Active Motif</td><td rowspan="3">39536, Active Motif</td><td rowspan="3">Imputed[1]</td><td rowspan="3">Imputed[1]</td><td>Rada-Iglesias et al. 2012 (GSE28874)</td></tr>
<tr><td>cNCC</td><td>39133, Active Motif</td><td>Prescott et al. 2015 (GSE70751)</td></tr>
<tr><td>CT_CS13</td><td rowspan="5">ab4729, Abcam</td><td rowspan="3">ab8580, Abcam</td><td rowspan="3">07-449, EMD Millipore</td><td rowspan="4">Imputed²</td><td rowspan="3">ab9050, Abcam</td><td rowspan="4">Wilderman et al. 2018 (GSE97752)</td></tr>
<tr><td>CT_CS14</td></tr>
<tr><td>CT_CS15</td></tr>
<tr><td>CT_CS17</td><td rowspan="2">Imputed²</td><td rowspan="2">Imputed²</td><td rowspan="2">Imputed²</td><td rowspan="2">Imputed²</td></tr>
<tr><td>CT_CS20</td></tr>
<tr><td>10wpc</td><td>Imputed²</td><td>Imputed²</td><td>Imputed²</td><td>Imputed²</td></tr>
</table>

[a] - Cell/tissue as origin of chromatin immunoprecipitation followed by sequencing (ChIP-seq); hNCC - early human neural crest cell; cNCC - cranial NCC; CT - craniofacial tissue; CS - Carnegie stage; wpc - weeks *post-conceptum.* [1] - Imputation performed using ChromImpute v1.0.1 (Ernst and Kellis, 2015) based on 127 cell types from Roadmap Epigenome Project (Roadmap Epigenomics Consortium et al. 2015) and the available chromatin marks in hNCC and cNCC in present study. [2] - Imputation performed using ChromImpute v1.0.1 (Ernst and Kellis, 2015) based on 127 cell types from Roadmap Epigenome Project (Roadmap Epigenomics Consortium et al. 2015) and available chromatin marks in CT by Wilderman et al. 2018.

Table S5 Uniquely aligned reads per sample and chromatin mark in neural crest cell samples

Chromatin mark	hNCC[a]	cNCC[b]
H3K27C	19,673,201	29,813,573
H3K27me3	15,902,493	19,534,085
H3K4me1	17,415,485	24,603,013
H3K4me3	18,482,513	22,506,230
Input	19,733,798	25,999,759
Sum	**91,207,490**	**122,456,660**
Average	**18,241,498**	**24,491,332**

[a] - Number of read in raw data (fastq file) per early human neural crest cell sample (hNCC) and chromatin mark downloaded from GEO (GSE28874); Rada-Iglesias et al. 2012. [b] - Number of read in raw data (fastq file) per cranial NCC (cNCC) sample and chromatin mark downloaded from GEO (GSE70751); Prescott et al. 2015

Table S15 Transfer of 18-state model of ChromHMM to 8-state model.

Original 18 state model				Condensed 8 state model	
States[a]	**Description**	**Color name**	**RGB code**	**RGB code**	**State[a]**
TssA	Active TSS	Red	255,0,0	255,0,0	TSS
TssFlnk	Flanking TSS	Orange Red	255,69,0		
TssFlnk	Flanking TSS Upstream	Orange Red	255,69,0		
TssFlnk	Flanking TSS Downstream	Orange Red	255,69,0		
Tx	Strong transcription	Green	0,128,0	0,128,0	Tx
TxWk	Weak transcription	DarkGreen	0,100,0		
EnhG1	Genic enhancer1	GreenYellow	194,225,5	255,255,0	Enh
EnhG2	Genic enhancer2	GreenYellow	194,225,5		
EnhA1	Active Enhancer 1	Orange	255,195,77		
EnhA2	Active Enhancer 2	Orange	255,195,77		
EnhWk	Weak Enhancer	Yellow	255,255,0		
ZNF/Rp	ZNF genes & repeats	Medium Aquamarine	102,205,170	102,205,170	ZNF_Rpts
Het	Heterochromatin	PaleTurquoise	138,145,208	138,145,208	Het
TssBiv	Bivalent/Poised TSS	IndianRed	205,92,92	233,150,122	TssBiv_Enh
EnhBiv	Bivalent Enhancer	DarkKhaki	189,183,107		
ReprPC	Repressed PolyComb	Silver	128,128,128	128,128,128	ReprPC
ReprPC	Weak Repressed PolyComb	Gainsboro	192,192,192		
Quies	Quiescent/Low	White	255,255,255	255,255,255	Quies

a - TSS - transcription starting site; Enh - enhancer; ReprPC - repressed PolyComb; Tx - transcribed sites; Het - Heterochromatin; Pois_TSS_Enh - poised TSS and bivalent enhancers; ZNF_Rpts - Zinc finger genes and repeats.

Supplemental Text - Description of novel risk loci

1) Risk locus $1p36_{CAPZB}$

The 1p36 locus was previously suggested as nsCL/P risk locus without reaching formal statistical thresholds, in an association study containing a subsample of the present study[30]. We here confirm this association at genome-wide significance. The top associated variant in MAiC is rs34746930, located within the genic region of the capping protein (actin filament) muscle Z-line, beta gene (*CAPZB*). Notably, this locus is independent from another risk locus previously

reported at 1p36, around the *PAX7* gene ($1p36_{PAX7}$, leadSNP rs742071)[31], as indicated by the location in two different topologically-associated domains (TAD), and the lack of linkage disequilibrium (LD) between the two lead variants (Figure S2); CEU: r^2=0.0017 / D'=0.12; East Asians: r^2=0.0058 / D'=0.18; South Asians: r^2=0.0002 / D'=0.16, assessed using LDpair in LDlink, 1000 genomes phase 3). The TAD around rs34746930 contains several genes, three of which can be considered strong candidates for an involvement in nsCL/P:

CAPZB has been shown to be highly expressed in the first pharyngeal arch during human development, an embryonic structure that hosts cells required for the formation of facial structures[32]. A *de novo* balanced translocation, disrupting *CAPZB*, was previously reported in a female individual presenting with craniofacial defects (e.g., cleft palate, micrognathia, low-set and rotated ears), hypotonia and developmental delay[33]. In addition, several deletions of different sizes encompassing *CAPZB* have been reported in individuals of the DECIPHER database, all of which presented with some degree of craniofacial malformation[34]. In zebrafish, loss of *capzb* leads to craniofacial phenotypes, and molecular data showed cell migration defects in zebrafish larvae, and differential expression of *pax3a* and *pax7a* in zebrafish neural crest cells[33]. Notably, *PAX7* is a candidate gene at the neighbouring $1p36_{PAX7}$-locus, suggesting further follow-up-studies including interaction analyses between *CAPZB* and *PAX7*.

NBL1 (neuroblastoma, suppression of tumorigenicity 1) encodes a bone morphogenic protein (BMP) antagonist of the DAN family, the latter is strongly evolutionary conserved. These secreted proteins are involved as antagonists in the BMP pathway: they bind BMP and, thereby, prevent its interaction with other receptors. This suggests an important role during growth and development. Recently, a role for the encoded NBL1-protein in neural crest cell migration was observed: Through *in vivo* analyses in the chicken and *in silico* simulations, it was shown that Nbl1 restrains cell migration through the regulation of cell speed. Thus, NBL1 is suggested to inhibit uncontrolled neural crest invasion and promotes collective migration[35].

Finally, the gene ***HTR6***, encoding the Serotonin receptor 6, is located ~220kb away from the sentinel SNP. No direct evidence for a role of this gene in craniofacial development has yet been reported. However, it has been suggested that a five-SNP haplotype in *HTR6* is associated with the risk of becoming a smoker[36]. Given increasing evidence for smoking being an environmental risk factor for orofacial clefing[37], this gene might be considered in further analyses of gene-environment analyses (taking into consideration maternal-fetal interactions). However, epigenetic data across mid-facial development do not indicate expression of *HTR6*.

The core associated region (defined as the region containing variants with r^2>0.8) extends over 35 kb and contains 25 common variants. One of these variants, rs6682099 (CEU: r^2=0.92 / D'=1.0 to rs34746930) is highly conserved, has a CADD score >20 and a Regulome-db-Score of 2b. The most prominent position weight matrix (PWM) altered by the C/T exchange of rs6682099 is a 7bp core motif for *PITX2*. Mutations in *PITX2* cause Axenfeld-Rieger Syndrome type 1, an autosomal-dominant disorder affecting primarily facial structures [OMIM 180500]. Patients with Axenfeld-Rieger Syndrome type 1 present with maxillary hypoplasia, short philtrum and thin upper lip, hypodontia (in particular

maxillary incisors) as well as complex eye phenotypes. In GTEX-data (v8), rs6682099 is an eQTL for *NBL1* (in sun-exposed skin tissue, and stomach), *CAPZB* (skin and adrenal gland), and *PQLC2* (in lung tissue), together with a set of other SNPs in high LD.

2) Risk locus $5p12_{FGF10}$

The MAiC lead SNP at 5p12, rs60107710, is located about 510 kb away from another previously reported variant at 5p12, i.e. rs10462065[38]. Although both lead variants are located within the same TAD, there is evidence from haplotype data to be independent from one another, as the lead variants do not share any LD in the three main investigated populations (i.e., CEU: r^2=0.0044 / D'=0.09; East Asians: r^2=0.0017 / D'=0.08; South Asians: r^2=0.0024 / D'=0.14, assessed using LDpair in LDlink, 1000 genomes phase 3 data; Figure S3). The 5p12-associated region around rs10462065 ($5p12_{rs10462065}$) is located about 320 kb downstream of the *FGF10* transcription start site (TSS), while the newly identified 5p12-association region around rs60107710 ($5p12_{rs60107710}$) maps about 190 kb upstream of the TSS. The core associated region of the $5p12_{rs60107710}$-region contains ~200 SNPs and extends over 180 kb. Within that region, the common variant with the lowest Regulome db-Score is rs1482664 (2b), however, no further annotation is provided in support for this variant being causal (including the absence of a significant eQTL effect in GTex v8).

Within the TAD around rs60107710, two protein-coding genes are located. ***FGF10*** (fibroblast growth factor 10) is a signalling growth factor that predominantly acts in mesenchymal and epithelial tissue, and is required for the development of multiple organs including the craniofacial complex[39]. *FGF10* is well studied for its role in palatal growth, in particular within the *PAX9* palatogenesis pathway[40]. When *Fgf10* is conditionally knocked-out in murine neural crest cells, many of the phenotypes observed in constitutive $Fgf10^{-/-}$ mice are recapitulated, including the frequent occurrence of cleft palate[41]. However, in our epigenetic data *FGF10* shows only limited evidence for being actively described in NCC.

The second gene within this TAD, ***NNT*** (nicotinamide nucleotide transhydrogenase), encodes for an integral protein of the inner mitochondrial membrane. It is ubiquitously expressed. Knocking down *NNT* in Hep1-cells results in impaired homeostasis of cells and reduced cell proliferation[42]. So far, no specific role in craniofacial development or disease has been reported.

3) Risk locus 5q13.1

The association at 5q13.1 is characterized by the lead variant rs6449957, which is located ~28 kb upstream of the TSS of ***PIK3R1*** (phosphoinositide-3-Kinase regulatory subunit 1, alias: *GRB1*). However, the associated region extends into the first coding exons of *PIK3R1* (Figure 4c, Figure S4), which is also the only protein-coding gene located within the TAD.

PIK3R1 has been shown to play an important role in the metabolism of insulin[43,44]. Moreover, heterozygous mutations in the *PIK3R1* gene have been described as causal for SHORT syndrome [OMIM #269880], clinical symptoms of which include teething delay, short stature, hernia and ocular depression[45]. In a recent systems genetics study, *PIK3R1* was identified as candidate gene for nsCL/P based on a re-analysis of a previously published expression dataset that compared dental pulp stem cells from nsCL/P patients with non-affected control children[46]. Moreover, another study employing systems genetics suggested PIK3R1 as mediator for viral cancerogenesis and interaction with cancer genes[47], which is noteworthy given some suggestive evidence of orofacial clefting being associated with an increased risk of different cancer types[48].

The association structure at 5q13.1 is described in the Main Text. Briefly, the core associated region comprises ~30 variants, none of which is an eQTL in GTex v8. However, we observed rs6449957 to be reported as splice QTL for both *PIK3R1* and a long non-coding RNA (*LINC02219*), respectively, in testis. The lowest Regulome-db score at 5p13.1 is observed for rs6449956 (score 2b), which is predicted to disrupt the binding site for transcription factor FEV, a member of the Ets-family of transcription factors (according to JASPAR2018, Figure S4), however, no role of FEV in craniofacial development has yet been reported. Interestingly, our integrative data suggests some evidence for *MAST4*, located in the adjacent TAD, as second candidate gene at this locus (see Main Text).

4) Risk locus 7p21.1

The 7p21.1 risk region is characterized by its lead SNP rs62453366, which is located intronically within the *ABCB5* gene (Figure S5). The core associated region is very narrow, encompassing 10 kb only. None of the 20 variants located within the $GWAS_{SNP}$-region is an eQTL in GTEx v8 data, and none of the variants has CADD>10 or Regulome-score better than 4. Within the TAD, three genes are located – *ABCB5*, *SP8* and *RPLS23P8*, the latter of which is a processed ribosomal protein pseudogene for which no functional information is available. We therefore here describe the two other genes, which both can be considered interesting candidate genes for nsCL/P.

ABCB5 (ATP binding cassette subfamily B member 5) represents a member of the ABC transporter superfamily of integral membrane proteins. ABCB5 is a marker for progenitor cells in both skin and human melanoma, and plays an important role as regulator of cellular differentiation[49]. *ABCB5* is also expressed in specific fractions of limbal stem (LS) cells, lack of which represent a major cause of blindness[50]. LS-cells positive for *ABCB5* expression co-express the *deltaNp63alpha isoform* of p63, but not the differentiation marker *KRT12*[51]. The striking co-expression of *ABCB5* and p63 in limbal stem cells is noteworthy, for the following reasons: (i) p63 is the causal gene for different types of ectodermal dysplasias (e.g. AEC/EEC-syndrome), (ii) the *p63* gene maps to an nsCL/P risk locus itself (chromosome 3q28), and (ii) p63 has recently been shown to be critically relevant for establishing enhancer marks at genes relevant in craniofacial development and disease[53]. Notably, a recent study identified that p63 is superior to ABCB5 as marker for stem cells, and also associates with LS-cells with increased pigmentation[53].

SP8 encodes the zinc finger transcription factor SP8. *Sp8*$^{-/-}$-mice show severe defects in limb development and craniofacial malformations. Specifically, at E14.5, facial prominences of *Sp8*$^{-/-}$ mice are underdeveloped in both size and structure, which results in severe craniofacial hypoplasia. Later in development, this is still recapitulated through severe midline defects, exencephaly, cleft palate, and a loss of neural crest cell. In that study, Sp8 was identified as craniofacial signalling center that regulates proliferation and apoptosis of NCC, with molecular downstream effects on *Fgf8* and *Fgf17* expression. Partial rescue of the phenotype in the *Sp8*$^{-/-}$ mice was obtained through reduction of Sonic hedgehog signalling, indicative of role for *Sp8* in the Shh-Fgf signalling pathway[54]. Recently, a truncating mutation within *SP8* was identified in a patient with nsCL/P (p.S261X) in a resequencing study[55], which is highly interesting as no *SP8* loss-of-function variant is currently reported in gnomAD database.

5) Risk locus 20q13.12

The 20q13.12 risk locus is characterized by its lead SNP rs3091552, which is located upstream of the gene eyes absent homologue 2 (***EYA2***). EYA2 is also the only protein-coding gene located in this single-gene TAD (Figure S6). The core-associated region encompasses rs3091552 and two additional variants

in strong LD (rs12481092: D' = 1.0 / r^2 = 0.88; rs6066089: D' = 1.0 / r^2 = 0.83). In GTEx (v8), all three SNPs represent eQTLs for *EYA2* in artery/aorta and additional tissues, with the risk alleles being associated with decreased *EYA2* expression. The strongest eQTL effect at this locus is demonstrated for rs8125695 (D' = 1.0 / r^2 = 0.79 to rs3091552, with P=8.6×10^{-14}, effect size 0.48 in GTExv8).

The gene *EYA2* encodes for a transcription factor with profound role in a variety of cellular and developmental processes, including cardiac[56] and muscle[57] development. Members of the EYA-family (including EYA2) are centrally involved in embryonic organogenesis through the promotion of proliferation and/or survival of progenitor-cell populations[58], and loss of function mutations have been shown to cause branchio-oto-renal (BOR) syndrome[59] which, among others, is characterized by malformations of anatomical structures derived from the human branchial arches (e.g., ears, OMIM: #113650). In mice it was shown that during eye morphogenesis, retinoic acid targets the neural crest-cell-derived mesenchyme in which Eya2-related apoptosis has been observed[60]. EYA-proteins largely functions through formation of a protein-complex together with SIX1 and DACH[58], and the important role for this transcription complex has been shown through both functional assays[58] and structural modelling[61], respectively. Notably, disruption of this process contributes to epithelial-mesenchymal-transition, and metastasis[61]. Recently, it was also shown in mice that disrupting Eya2 phosphatase activity through chemical inhibits Eya2-mediated cell migration[62].

References

1. Mangold, E., Ludwig, K.U., Birnbaum, S., Baluardo, C., Ferrian, M., Herms, S., Reutter, H., De Assis, N.A., Chawa, T. Al, Mattheisen, M., et al. (2010). Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. Nat. Genet.

2. Beaty, T.H., Murray, J.C., Marazita, M.L., Munger, R.G., Ruczinski, I., Hetmanski, J.B., Liang, K.Y., Wu, T., Murray, T., Fallin, M.D., et al. (2010). A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. Nat. Genet.

3. Leslie, E.J., Carlson, J.C., Shaffer, J.R., Feingold, E., Wehby, G., Laurie, C.A., Jain, D., Laurie, C.C., Doheny, K.F., McHenry, T., et al. (2016). A multi-ethnic genome-wide association study identifies novel loci for non-syndromic cleft lip with or without cleft palate on 2p 24.2, 17q23 and 19q13. Hum. Mol. Genet.

4. Ludwig, K.U., Mangold, E., Herms, S., Nowak, S., Reutter, H., Paul, A., Becker, J., Herberz, R., Alchawa, T., Nasser, E., et al. (2012). Genome-wide meta-analyses of nonsyndromic cleft lip with or without cleft palate identify six new risk loci. Nat. Genet.

5. Ludwig, K.U., Böhmer, A.C., Bowes, J., Nikolić, M., Ishorst, N., Wyatt, N., Hammond, N.L., Gölz, L., Thieme, F., Barth, S., et al. (2017). Imputation of orofacial clefting data identifies novel risk loci and sheds light on the genetic background of cleft lip ± cleft palate and cleft palate only. Hum. Mol. Genet.

6. Leslie, E.J., Carlson, J.C., Shaffer, J.R., Butali, A., Buxó, C.J., Castilla, E.E., Christensen, K., Deleyiannis, F.W.B., Leigh Field, L., Hecht, J.T., et al. (2017). Genome-wide meta-analyses of nonsyndromic orofacial clefts identify novel associations between FOXE1 and all orofacial clefts, and TP63 and cleft lip with or without cleft palate. Hum. Genet.

7. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics.

8. Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet.

9. Spielman, R.S., McGinnis, R.E., and Ewens, W.J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am. J. Hum. Genet.

10. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: Fast and efficient meta-analysis of genomewide association scans. Bioinformatics.

11. Bulik-Sullivan, B., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., Neale, B.M., Corvin, A., et al. (2015). LD score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet.

12. Watanabe, K., Taskesen, E., Van Bochoven, A., and Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. Nat. Commun.

13. Machiela, M.J., and Chanock, S.J. (2015). LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics.

14. Aguet, F., Brown, A.A., Castel, S.E., Davis, J.R., He, Y., Jo, B., Mohammadi, P., Park, Y.S., Parsana, P., Segrè, A. V., et al. (2017). Genetic effects on gene expression across human tissues. Nature.

15. de Leeuw, C.A., Mooij, J.M., Heskes, T., and Posthuma, D. (2015). MAGMA: Generalized Gene-Set Analysis of GWAS Data. PLoS Comput. Biol.

16. Rada-Iglesias, A., Bajpai, R., Prescott, S., Brugmann, S.A., Swigut, T., and Wysocka, J. (2012). Epigenomic annotation of enhancers predicts transcriptional regulators of human neural crest. Cell Stem Cell.

17. Prescott, S.L., Srinivasan, R., Marchetto, M.C., Grishina, I., Narvaiza, I., Selleri, L., Gage, F.H., Swigut, T., and Wysocka, J. (2015). Enhancer Divergence and cis-Regulatory Evolution in the Human and Chimp Neural Crest. Cell.

18. Wilderman, A., VanOudenhove, J., Kron, J., Noonan, J.P., and Cotney, J. (2018). High-Resolution Epigenomic Atlas of Human Embryonic Craniofacial Development. Cell Rep.

19. Bajpai, R., Chen, D.A., Rada-Iglesias, A., Zhang, J., Xiong, Y., Helms, J., Chang, C.P., Zhao, Y., Swigut, T., and Wysocka, J. (2010). CHD7 cooperates with PBAF to control multipotent neural crest formation. Nature.

20. Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature.

21. Furlan-Magaril, M., Rincón-Arano, H., and Recillas-Targa, F. (2009). Sequential chromatin immunoprecipitation protocol: ChIP-reChIP. Methods Mol. Biol.

22. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods.

23. Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P., et al. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res.

24. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol.

25. Ernst, J., and Kellis, M. (2015). Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nat. Biotechnol.

26. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature.

27. Ernst, J., and Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. Nat. Protoc.

28. Laugsch, M., Bartusel, M., Alirzayeva, H., Karaolidou, A., Rehimi, R., Crispatzu, G., Nikolic, M., Bleckwehl, T., Kolovos, P., van Ijcken, W.F.J., et al. (2018). Disruption of the TFAP2A Regulatory Domain Causes Banchio-Oculo-Facial Syndrome (BOFS) and Illuminates Pathomechanisms for Other Human Neurocristopathies.

29. Schmidt, E.M., Zhang, J., Zhou, W., Chen, J., Mohlke, K.L., Chen, Y.E., and Willer, C.J. (2015). GREGOR: Evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. Bioinformatics.

30. Carlson, J. C. *et al.* (2019). A systematic genetic analysis and visualization of phenotypic heterogeneity among orofacial cleft GWAS signals. *Genet Epidemiol* 43, 704-716.

31. Ludwig, K. U. *et al.* (2012). Genome-wide meta-analyses of nonsyndromic cleft lip with or without cleft palate identify six new risk loci. *Nat Genet* 44, 968-971.

32. Cai, J. *et al.* (2005). Gene expression in pharyngeal arch 1 during human embryonic development. *Hum Mol Genet* 14, 903-912.

33. Mukherjee, K. *et al.* (2016). Actin capping protein CAPZB regulates cell morphology, differentiation, and neural crest migration in craniofacial morphogenesisdagger. *Hum Mol Genet* 25, 1255-1270.

34. Kang, S. H. *et al.* (2007). Identification of proximal 1p36 deletions using array-CGH: a possible new syndrome. *Clin Genet* 72, 329-338.

35. McLennan, R. *et al.* (2017). DAN (NBL1) promotes collective neural crest migration by restraining uncontrolled invasion. *J Cell Biol* 216, 3339-3354.

36. Lerer, E., Kanyas, K., Karni, O., Ebstein, R. P. & Lerer, B. (2006). Why do young women smoke? II. Role of traumatic life experience, psychological characteristics and serotonergic genes. *Mol Psychiatry* 11, 771-781.

37. Sabbagh, H. J. *et al.* (2015). Passive smoking in the etiology of non-syndromic orofacial clefts: a systematic review and meta-analysis. *PLoS One* 10, e0116963.

38. Yu, Y. *et al.* (2017). Genome-wide analyses of non-syndromic cleft lip with palate identify 14 novel loci and genetic heterogeneity. *Nat Commun* 8, 14364.

39. Prochazkova, M., Prochazka, J., Marangoni, P. & Klein, O. D. (2018). Bones, Glands, Ears and More: The Multiple Roles of FGF10 in Craniofacial Development. *Front Genet* 9, 542.

40. Li, R., Chen, Z., Yu, Q., Weng, M. & Chen, Z. (2019). The Function and Regulatory Network of Pax9 Gene in Palate Development. *J Dent Res* 98, 277-287.

41. Teshima, T. H., Lourenco, S. V. & Tucker, A. S. (2016). Multiple Cranial Organ Defects after Conditionally Knocking Out Fgf10 in the Neural Crest. *Front Physiol* 7, 488.

42. Ho, H. Y., Lin, Y. T., Lin, G., Wu, P. R. & Cheng, M. L. (2017). Nicotinamide nucleotide transhydrogenase (NNT) deficiency dysregulates mitochondrial retrograde signaling and impedes proliferation. *Redox Biol* 12, 916-928.

43. Thauvin-Robinet, C. *et al.* (2013). PIK3R1 mutations cause syndromic insulin resistance with lipoatrophy. *Am J Hum Genet* 93, 141-149.

44. Kuo, T. *et al.* (2017). Pik3r1 Is Required for Glucocorticoid-Induced Perilipin 1 Phosphorylation in Lipid Droplet for Adipocyte Lipolysis. *Diabetes* 66, 1601-1610.

45. Avila, M. *et al.* (2016). Clinical reappraisal of SHORT syndrome with PIK3R1 mutations: toward recommendation for molecular testing and management. *Clin Genet* 89, 501-506.

46. Kobayashi, G. S. *et al.* (2013). Susceptibility to DNA damage as a molecular mechanism for non-syndromic cleft lip and palate. *PLoS One* 8, e65677.

47. Wang, H. *et al.* (2016). Gene expression profiling analysis contributes to understanding the association between non-syndromic cleft lip and palate, and cancer. *Mol Med Rep* 13, 2110-2116.

48. Bille, C. *et al.* (2005). Cancer risk in persons with oral cleft--a population-based study of 8,093 cases. *Am J Epidemiol* 161, 1047-1055.

49. Frank, N. Y. *et al.* (2003). Regulation of progenitor cell fusion by ABCB5 P-glycoprotein, a novel human ATP-binding cassette transporter. *J Biol Chem* 278, 47156-47165.

50. Dua, H. S., Joseph, A., Shanmuganathan, V. A. & Jones, R. E. (2003). Stem cell differentiation and the effects of deficiency. *Eye (Lond)* 17, 877-885.

51. Ksander, B. R. *et al.* (2014). ABCB5 is a limbal stem cell gene required for corneal development and repair. *Nature* 511, 353-357.

52. Lin-Shiao, E. *et al.* (2019). p63 establishes epithelial enhancers at critical craniofacial development genes. *Sci Adv* 5, eaaw0946.

53. Liu, L. *et al.* (2018). Pigmentation Is Associated with Stemness Hierarchy of Progenitor Cells Within Cultured Limbal Epithelial Cells. *Stem Cells* 36, 1411-1420.

54. Kasberg, A. D., Brunskill, E. W. & Steven Potter, S. (2013). SP8 regulates signaling centers during craniofacial development. *Dev Biol* 381, 312-323.

55. Marini, N. J., Asrani, K., Yang, W., Rine, J. & Shaw, G. M. (2019). Accumulation of rare coding variants in genes implicated in risk of human cleft lip with or without cleft palate. *Am J Med Genet A* 179, 1260-1269.

56. Lee, S. H. *et al.* (2009). The transcription factor Eya2 prevents pressure overload-induced adverse cardiac remodeling. *J Mol Cell Cardiol* 46, 596-605.

57. Grifone, R. *et al.* (2007). Eya1 and Eya2 proteins are required for hypaxial somitic myogenesis in the mouse embryo. *Dev Biol* 302, 602-616.

58. Li, X. *et al.* (2003). Eya protein phosphatase activity regulates Six1-Dach-Eya transcriptional effects in mammalian organogenesis. *Nature* 426, 247-254.

59. Abdelhak, S. *et al.* (1997). A human homologue of the Drosophila eyes absent gene underlies branchio-oto-renal (BOR) syndrome and identifies a novel gene family. *Nat Genet* 15, 157-164.

60. Matt, N. *et al.* (2005). Retinoic acid-dependent eye morphogenesis is orchestrated by neural crest cells. *Development* 132, 4789-4800.

61. Patrick, A. N. *et al.* (2013). Structure-function analyses of the human SIX1-EYA2 complex reveal insights into metastasis and BOR syndrome. *Nat Struct Mol Biol* 20, 447-453.

62. Krueger, A. B. *et al.* (2014). Allosteric inhibitors of the Eya2 phosphatase are selective and inhibit Eya2-mediated cell migration. *J Biol Chem* 289, 16349-16361.