Impact of low-frequency coding variants on human facial shape

Dongjing Liu[1], Nora Alhazmi[2,3], Harold Matthews[4,5], Myoung Keun Lee[6], Jiarui Li[7], Jacqueline T. Hecht[8], George L. Wehby[9], Lina M. Moreno[10], Carrie L. Heike[11], Jasmien Roosenboom[6], Eleanor Feingold[1,12], Mary L. Marazita[1,6], Peter Claes[4,7], Eric C. Liao[13], Seth M. Weinberg[1,6*], John R. Shaffer[1,6*]

[1]Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America

[2] Department of Oral Biology, Harvard School of Dental Medicine, Boston, Massachusetts, United States of America

[3] King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia

[4] Department of Human Genetics, KU Leuven, Leuven, Belgium

[5] Medical Imaging Research Center, UZ Gasthuisberg, Leuven, Belgium

[6] Center for Craniofacial and Dental Genetics, Department of Oral Biology, School of Dental Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America

[7] Department of Electrical Engineering, ESAT/PSI, KU Leuven, Leuven, Belgium

[8] Department of Pediatrics, University of Texas McGovern Medical Center, Houston, Texas, United States of America

[9] Department of Health Management and Policy, University of Iowa, Iowa City, Iowa, United States of America

[10] Department of Orthodontics, University of Iowa, Iowa City, Iowa, United States of America

[11] Department of Pediatrics, Seattle Children's Craniofacial Center, University of Washington, Seattle, Washington, United States of America

[12] Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania, United States of America

[13] Department of Surgery, Center for Regenerative Medicine, Massachusetts General Hospital, Shriners Hospital, Boston, Massachusetts, United States of America

[*] Corresponding author
E-mail: smwst46@pitt.edu (SMW)
E-mail: john.r.shaffer@pitt.edu (JRS)

Abstract

The contribution of low-frequency variants to the genomic architecture of normal-range facial traits is unknown. Therefore, we studied the influence of 31347 low-frequency coding variants (MAF < 1%) in 8091 genes on multi-dimensional facial shape phenotypes in a European cohort of 2329 healthy individuals. Using three-dimensional facial images, we partitioned the full face into 31 hierarchically arranged segments to model global-to-local features, and generated multi-dimensional phenotypes representing the shape variation within each segment. We used MultiSKAT, a multivariate kernel regression approach to scan the exome for face-associated low-frequency variants in a gene-based manner. After accounting for multiple tests, seven genes (*AR*, *CARS2*, *FTSJ1*, *HFE*, *LTB4R*, *TELO2*, *NECTIN1*) were significantly associated with morphology of the cheek, chin, nose and philtrum. These genes displayed a wide range of phenotypic effects, with some impacting the full face and others affecting localized regions. Notably, *NECTIN1* is an established craniofacial gene that underlies both syndromic and isolated forms of cleft lip and palate. The missense variant rs142863092 in *NECTIN1* had a significant individual effect on chin morphology, and it is predicted bioinformatically to be deleterious on the nectin-1 protein. We show that the zebrafish *nectin1a* mutation affects craniofacial development and leads to abnormal size and shape of the palate and Meckel's cartilage. These results expand our understanding of the genetic basis of normal-range facial shape by highlighting the role of low-frequency coding variants in novel genes.

Introduction

Significant progress has been made in elucidating the genetic basis of human facial traits [1-3]. Genome-wide association studies (GWAS) have now identified and replicated several common genetic variants with phenotypic effects on normal-range facial morphology [4-12]; however, these variants cumulatively explain only a small fraction of the heritable phenotypic variation. Based on large-scale genomic studies of other complex morphological traits such as height [13-15], we hypothesize that functional variants at hundreds or perhaps thousands of loci have yet to be discovered. While we expect that common variants, with a minor allele frequency (MAF) greater than 1%, account for much of the heritable variation in facial morphology, low frequency (<1% MAF) genetic variants may also play an important role. An exome-wide study of human height, for example, discovered 29 low-frequency coding variants with large effects of up to 2 centimeters per allele [13].

Our previous GWAS in a modestly sized cohort of healthy individuals identified 1932 common genetic variants associated with facial variation at 38 loci, 15 of which were independently replicated [5]. The success of this GWAS was attributed in part to an innovative data-driven phenotyping approach, in which 3D facial surfaces were partitioned into hierarchically organized regions, each defined by multiple axe of shape variation. This approach allowed for simultaneous testing of genetic variants on facial morphology at multiple levels of scale – from the entire face to highly localized facial region. Extending this global-to-local analysis of facial traits to the analysis of low-frequency variants requires an appropriate and scalable statistical framework capable of accommodating the multivariate nature of the facial shape variables. A recently developed approach, MultiSKAT [16] has been proposed for this purpose and showed desirable performance in its original development.

In this study, we evaluate the influence of low frequency coding variants, captured by the Illumina HumanExome BeadChip, on normal-range facial morphology in 2,329 individuals. We apply multivariate gene-based association testing methods to multi-dimensional facial shape phenotypes derived from 3D facial images. The results of our analyses point to novel genes, including at least one with a role in orofacial clefting and several others with no previously described role in craniofacial development or disease. Moreover, we provide experimental validation of our genetic association results through expression screening and knockout experiments in a zebrafish model. These results enhance our understanding of the genetic architecture of human facial variation.

Materials and Methods

Ethics statement

Institutional ethics (IRB) approval was obtained at each recruitment site and all subjects gave their written informed consent prior to participation (University of Pittsburgh Institutional Review Board #PRO09060553 and #RB0405013; UT Health Committee for the Protection of Human Subjects #HSC-DB-09-0508; Seattle Children's Institutional Review Board #12107; University of Iowa Human Subjects Office/Institutional Review Board #200912764 and #200710721).

Sample and Phenotyping

The study cohort comprised 2,329 unrelated, healthy individuals of European ancestry aged 3 to 40 years. Participants were eligible if they had not experienced facial trauma, major surgery, congenital facial anomalies that could potentially affect their natural facial structure. 3D images of each participant's resting face were captured via digital stereophotogrammetry using the 3dMD face camera system. The data-driven phenotyping approached has been described in detail previously [5]. Briefly, approximately 10,000 points—"quasi-landmarks"—were automatically placed across the facial surface, by a non-rigid registration of a standard facial template onto each surface. The result is that each quasi-landmark represents the same facial position across all participants [17]. The configurations were the co-aligned to their mean with generalized Procrustes analysis (GPA). The quasi-landmarks were then clustered into two groups of co-varying points in order to partition the full face into two segments. GPA was repeated within each segment such that the segments were further and further partitioned. This process was continued for a total of four iterations to generate a hierarchy of 31 facial segments comprising overlapping groups of quasi-landmarks. The hierarchical structure is illustrated in Fig 1, where the segments within each layer collectively constitute the whole face, and the successive layers represent the shift from more globally integrated to more locally focused morphology. In this way a total of 31 partially-overlapping segments, which we called modules, were generated. The shape variation characterizing a module is represented by the 3D coordinates of all quasi-landmarks contained therein. To reduce the dimensionality of the shape variation within each module, principal components analysis and parallel analysis were performed on the quasi-landmarks to form multi-dimensional phenotypes in which shape variation is represented by principal component scores (PCs). This procedure resulted in a total of 31 modules, each of which is made up of 8 to 50 PCs that jointly captured near complete shape variance. The effects of sex, age, height, weight, facial size

and genetic ancestry were corrected for at the phenotyping stage. These facial module phenotypes were successfully used in our previous GWAS of common variants [5] which demonstrated advantages of this data-driven multivariate modeling of multipartite traits in the context of gene mapping studies compared to *a priori* [8] and univariate [7] phenotypes in this sample.

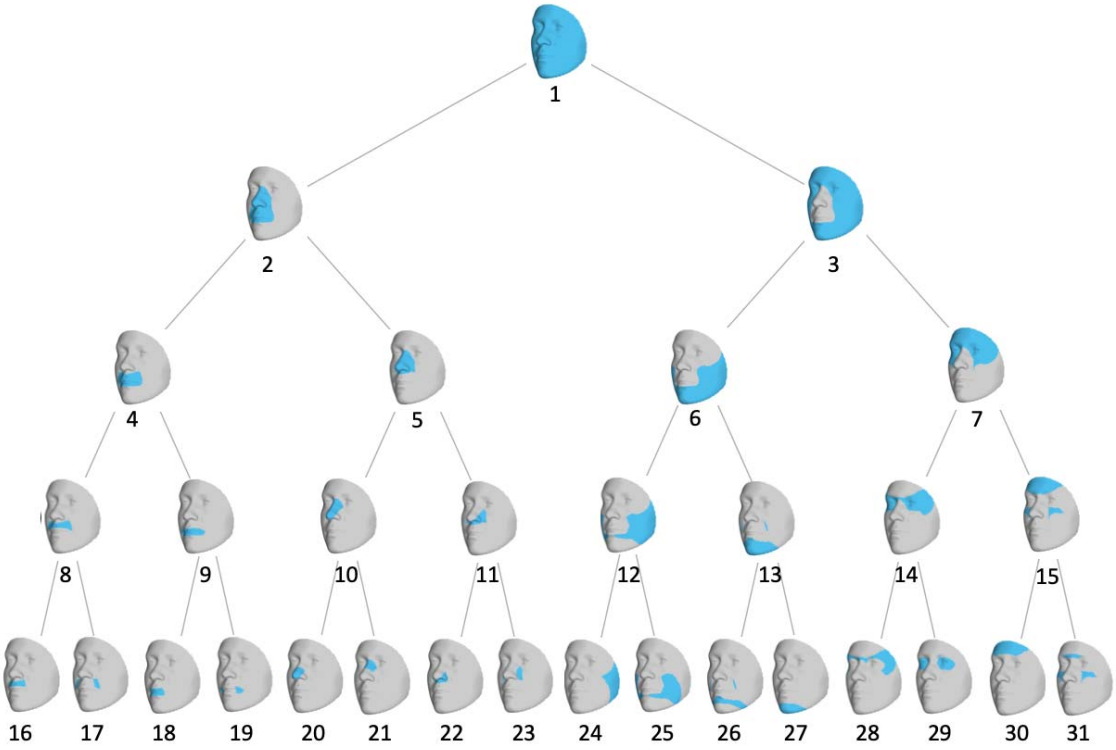

Fig 1. Hierarchical clustering of facial shape.
Global-to-local facial segmentation obtained using hierarchical spectral clustering. Segments are colored in blue. The highest-level segment representing the full face is split into two distinct sub-segments, and this bifurcation process is repeated until a five-level hierarchy comprising 31 segments has been reached.

In addition to the phenotype quality control process described in [5], we further examined the phenotypic distribution of each module for extreme outlier faces, as phenotypic outliers may adversely impact rare variant tests [18]. To accomplish this, we looked at both the joint and pairwise distribution of all PCs underlying each module. We visualized quantile-quantile (Q-Q) plots of chi-squared quantiles versus robust squared Mahalanobis distances to identify outliers that fell far apart from the rest of the sample. Mahalanobis distance is a metric measuring how far each observation is to the center of the joint distribution, which

can be thought of as the centroid in multivariate space. Facial images associated with outlier observations were revisited to confirm the data validity and sample eligibility. Finally, one outlier face was excluded for analysis involving module 27 representing variation of the chin.

Genotyping

The cohort was genotyped for the Illumina OmniExpress + Exome v1.2 array, which included approximately 245,000 coding variants in the exome panel. Standard data cleaning and imputation procedures were implemented. Imputed genotypes with a certainty above 0.9 were used to fill in any sporadic missingness among genotype calls of the directly genotyped variants. We did not include any wholly unobserved, imputed SNPs in this analysis. Ancestry PCs based on common LD-pruned SNPs were constructed and regressed out from the modular traits to adjust for population structure.

MultiSKAT

MultiSKAT [16] is a recently developed statistical approach for testing sets of variants, in this case coding variants within genes, for association with a multivariate trait. The strategy of testing low-frequency variants in aggregate improves power compared to individual tests of each variant. The tool is flexible in relating multiple variants collectively to multivariate phenotypes through the use of several choices of kernels, and includes an omnibus test to obtain optimal association p-values by integrating results across different kernels via Copula. This multivariate nature fits well in our facial module setting here, given that each module is composed of many independent components. MultiSKAT does not restrict the frequency of variants to be tested, but our analysis considers low-frequency variants exclusively.

MultiSKAT uses a phenotype kernel to model how one variant affects multiple traits and uses a genotype kernel to specify how multiple variants influence one trait. In reality, these effects are often not known as a priori, and the true relationship can be a mixture of effects. We used the heterogeneous and homogeneous phenotype kernels, which are appropriate when the set of traits analyzed are orthogonal PCs. We specified both the Sequence Kernel Association Test (SKAT) and burden test for the genotype kernel, and then let the tool aggregate results across these 2×2 kernel combinations to obtain the omnibus p-value.

Gene-level analysis

Genome-wide coding variants with MAF less than 1% were aggregated into genes. We filtered out any variants with three or fewer minor alleles in the sample, and excluded genes with less than two qualified variants. This led to 31347 variants in 8091 genes being

tested. While aggregating across variants within a gene, MultiSKAT assigns larger weights to rarer variants. Due to the burden of multiple comparisons, we applied a Bonferroni threshold to declare significance. To account for the apparent correlation among partially overlapping facial modules, we used the procedure based on eigenvalues proposed by Li and Ji [19] to determine that the effective number of independent modules was 19. The threshold for significance was therefore set to $p < 3.3 \times 10^{-7}$ (i.e., 0.05 divided by the product of 8091 and 19). The phenotypic effects of identified genes on face were visualized by creating and comparing the average facial morphs in people carrying variants and people who do not carry variants.

We interrogated genes showing significant effects using GREAT [20], FUMA [21] and ToppFun [22] for gene set enrichment, and we looked up their expression in GTEx [23]. Following our hypothesis that genes influencing typical facial presentation may also be involved in facial anomalies, we examined whether any genes nominated in this study were associated with non-syndromic cleft palate with or without cleft lip (NSCL/P) by retrieving summary statistics from a past study of our group where we performed a gene-based low-frequency variant association scan on NSCL/P [24].

Variant-level analysis

For genes highlighted by MultiSKAT, we scrutinized the quality of genotype calling by inspecting clustering in allele intensity plots, and further performed association tests of SNPs individually. We use the MultiPhen approach [25], which finds the linear combination of PCs from a facial segment most associated with the genotypes at each SNP, and has the advantage of being robust when variants with low frequency are tested against non-normal phenotypes, as is the case in our study. Variant level functional prediction was done using CADD [26]. CADD is a comprehensive metric that weights and integrates diverse sources of annotation, by contrasting variants that survived natural selection with simulated. The scaled CADD score expresses the deleteriousness rank in terms of order of magnitude. A score of 10, for instance, is interpreted as ranking in the top 10% in terms of the damaging degree amongst reference genome SNPs, and a score of 20 refers to 1%, 30 to 0.1%, and so on. Variant identifiers and chromosomal locations are indicated in the hg19 genome build. Single exonic variants were looked up in literature and PhenoScanner [27] existing associations.

We quantified the magnitude of phenotypic effect of individual low-frequency variants by the difference between averaged faces of variant carriers and non-carries, which was further compared with the effects of significant common variants identified in the prior GWAS of these multidimensional traits [5]. Specifically, the centroids of the

multidimensional space defined by PCs in a certain module were computed separately for people carrying the variant and people who do not carry the variant. Then the Euclidean distance between the two centroids was calculated as a measure of variant effect size.

Expression screen of candidate genes in zebrafish:

The whole-mount RNA in situ hybridization (WISH) for *ar, cars2, ftsj1, hfe, itb4r, telo2, nectin1a* and *nectin1b* was performed on wild type zebrafish embryos at 24 hpf and 48 hpf as described by Thisse et al. [28]. All wild type embryos were collected synchronously at the corresponding stages and fixed in 4% paraformaldehyde (PFA) overnight. T7 RNA polymerase promoter was added to the reverse primers and was synthesized with antisense DIG-labeled probe in order to generate antisense RNA probe. The probe primers for *ar* are: forward 5'- GTCCTACAAGAACGCCAACG-3' and reverse 5'-GGTCACAGACTTGGAAAGGG-3' at 59°C. The *cars2* probe primers are: forward 5'-ATCTGGGTCATGCGTGTTCA-3' and reverse 5'- GGATTCCTGTGGTGCTTGGT at 59°C. The *ftsj1* probe primers are: forward 5'- GGCGAGAAGTGCCTTCAAAC-3' and reverse 5'-AGTCGTGCTTGTGTCTGGTT-3' and *hfe* probe primers are: forward 5'-GGGGATGGATGCTTCTACGA-3' and reverse 5'- CGCGCACACAAAATCATCAC-3' at 59°C. The *itb4r* probe primers are: forward 5'-GACGGTGCATTACCTGTGC-3' and reverse 5'-AGTCTTGTCCGCCAAGGTC-3' at 58°C. The primers for *telo2* are: forward 5'-GCTCCACTGGTGAGAGTGAG-3' and reverse 5'-GTCAGCTGAGGAGAGTCTGCG-3'. The primers for *nectin1a* probe are: forward 5'-AACACCCAGGAGATCAGCAA-3' and reverse 5'-CCTCCACCTCAGATCCGTAC-3' at 57°C and the *nectin1b* probe primers are: forward 5'-TGCTAACCCAGCATTGGGAG-3' and reverse 5'-GGTTCTTGGGCATTGGAGGA-3' at 59°C. Embryos were mounted using glycerol and imaged using Nikon AZ100 multizoom microscope.

Phenotype of mutant zebrafish:

As described by Kimmel et al. [29], zebrafish adults and embryos were obtained and maintained. Zebrafish nectin1a mutants were generated by transgene insertion Tg(Nlacz-GTvirus) in Chr 21: 21731876 - 21731886 (Zv9), and obtained from Zebrafish International Resource Center, allele Ia021885Tg (ZIRC catalog ID: ZL6899.07). The retroviral-mediated insertional mutagenesis inserts a molecular tag in the DNA and isolates the allele of interest. Therefore, this will induce a frameshift and probably causing either nonsense-mediated mRNA decay or a truncated protein [30,31]. The PCR genotyping primers for *nectin1a* are: forward 5'-TTAGACCAGCCCACCTCA-3' and reverse 5'-AATATGAAATAGCGCCGTTGTG-3' at 62°C.

Alcian blue staining was performed as described by Walker et al. [32]. The craniofacial cartilages were dissected and flat-mounted and then imaged using Nikon AZ100 multizoom microscope. After imaging, each embryo tail was placed in a PCR tube for genotyping. The protocol was used as described by [33] with modification of using fresh embryos without fixation.

Results

Gene-based tests detected seven genes significantly associated with one or more facial modules (Fig 2; Table 1): *HFE, NECTIN1, CARS2, LTB4R, TELO2, AR,* and *FTSJ1*. Three of them showed associations with more than one module. Fig 3 and S1 Table show the results of these genes across multiple phenotypes. Fig 3 shows the association signals propagating along the branching paths from the more global segments to the more local segments. Four genes (*HFE, CARS2, LTB4R,* and *TELO2*) were associated with nose-related modules, and the others were associated with modules of the chin, mouth, and cheek. *FTSJ1* had broad signals in the whole face as well as local regions, while effects of other genes were more confined to local modules. We observed well-calibrated test statistics and little evidence of inflation as shown in the Q-Q plots (Fig S1).

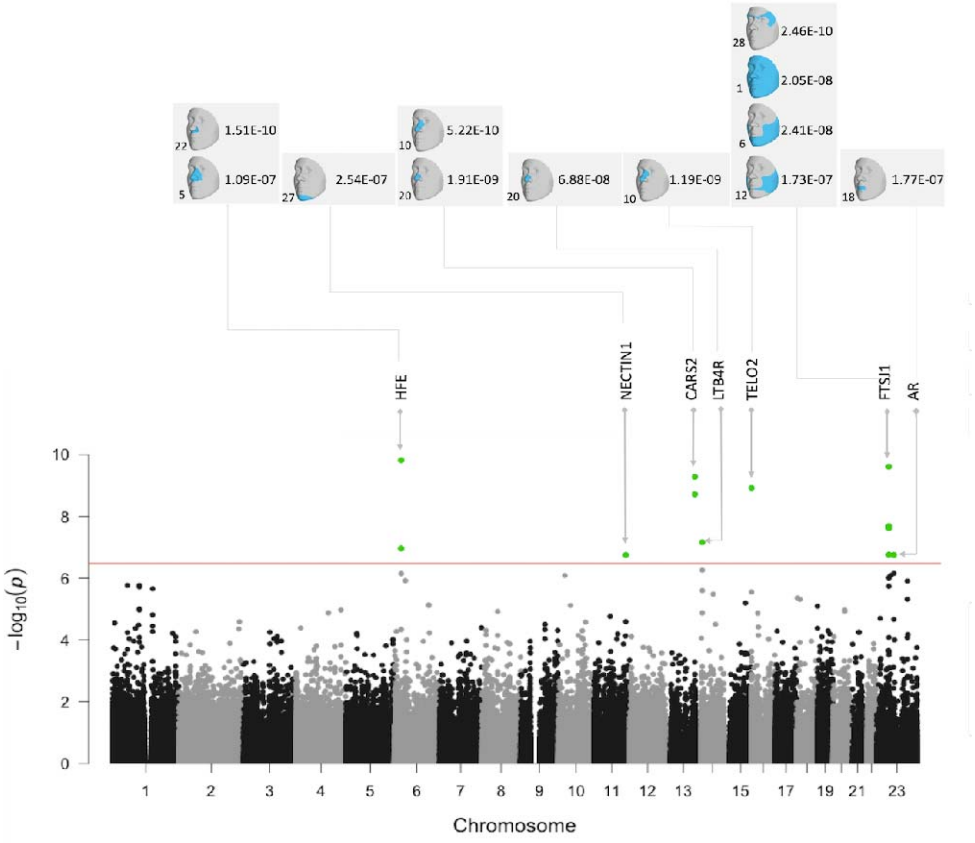

Fig 2. Composite Manhattan plot showing results across 31 facial modules.

Manhattan plot showing the position of genes on the x axis and MultiSKAT p-values on the y axis. A total of 31 points are plotted for each gene, representing p-values obtained by testing their association with 31 modules respectively. The red horizontal line indicates the study-wide significance threshold. The associated facial modules and the corresponding p-value for each gene that surpassed the threshold are shown above the Manhattan plot. The number to the left bottom of the facial image represents the module number from Fig 1.

Table 1. Single variant association and functional prediction for variants contributing to the gene-level significance

Chr	Gene	Gene Info	Number of Variants	SNP	Pos (hg19)	Ref/Alt[a]	Function[b]	CADD score[c]	MAF	Module[d]	MultiPhen P-value[d]
6	*HFE*	Homeostatic iron regulator, binds to transferrin receptor (TFR) and reduces its affinity for iron-loaded transferrin	2	rs149342416	26087686	G/C	Arg6Ser	15.3	0.087%	22	0.07
				rs143662783	26087718	C/G	Thr17Ile	13.4	0.086%	5	0.87
11	*NECTIN1*	Nectin 1, cell adhesion molecule	2	rs142863092	119548369	G/A	Arg210His	25.2	0.086%	27	1.08E-03
				rs137991779	119549425	G/A	Gly44Ser	29.2	0.108%	27	0.15
13	*CARS2*	Cysteinyl-tRNA synthetase 2, mitochondrial	2	rs151097801	111296817	C/T	Pro138Leu	22.4	0.086%	20	0.12
				rs117788141	111357899	G/A	Val69Ile	28.0	0.086%	10	0.01
14	*LTB4R*	Leukotriene B4 receptor 1, receptor for extracellular ATP > UTP and ADP	2	rs143666989	24780865	A/G	Gln332Arg	16.6	0.108%	20	0.11
				rs148153989	24780915	A/T	Met349Leu	12.5	0.086%	20	0.59
16	*TELO2*	Telomere length regulation protein homolog, regulate DNA damage response	3	rs140903666	1544313	G/A	Ala11Thr	6.3	0.215%	10	8.21E-04
				rs144863771	1544314	C/A	Ala11Asp	10.7	0.215%	10	8.21E-04
				rs147858841	1555541	C/T	Ala132Val	9.4	0.108%	10	0.43
23	*AR*	Androgen receptor, steroid hormone receptors	2	rs142280455	66905875	A/G	Ser598Gly	22.4	0.133%	18	0.81
				rs137852591	66941751	C/G	Gln267Glu	25.0	0.133%	18	3.91E-03
23	*FTSJ1*	Putative tRNA (cytidine(32)/guanosine(34)-2'-O)-methyltransferase	2	rs142932029	48341118	G/A	Ser161Asn	7.4	0.080%	28	1.59E-14
				rs201095751	48341414	C/T	Splice site	0.1	0.107%	12	0.10

[a] Alleles are listed as alternative/reference alleles on the forward strand of the reference genome.
[b] For missense variant, amino acid substitution is given
[c] Bioinformatic prediction of variant effect, higher score indicates greater damaging effect
[d] Variants were tested against the significant module(s) corresponding to the gene-based test, and for genes associated with multiple modules, the module yielding the smallest p-value is shown here.

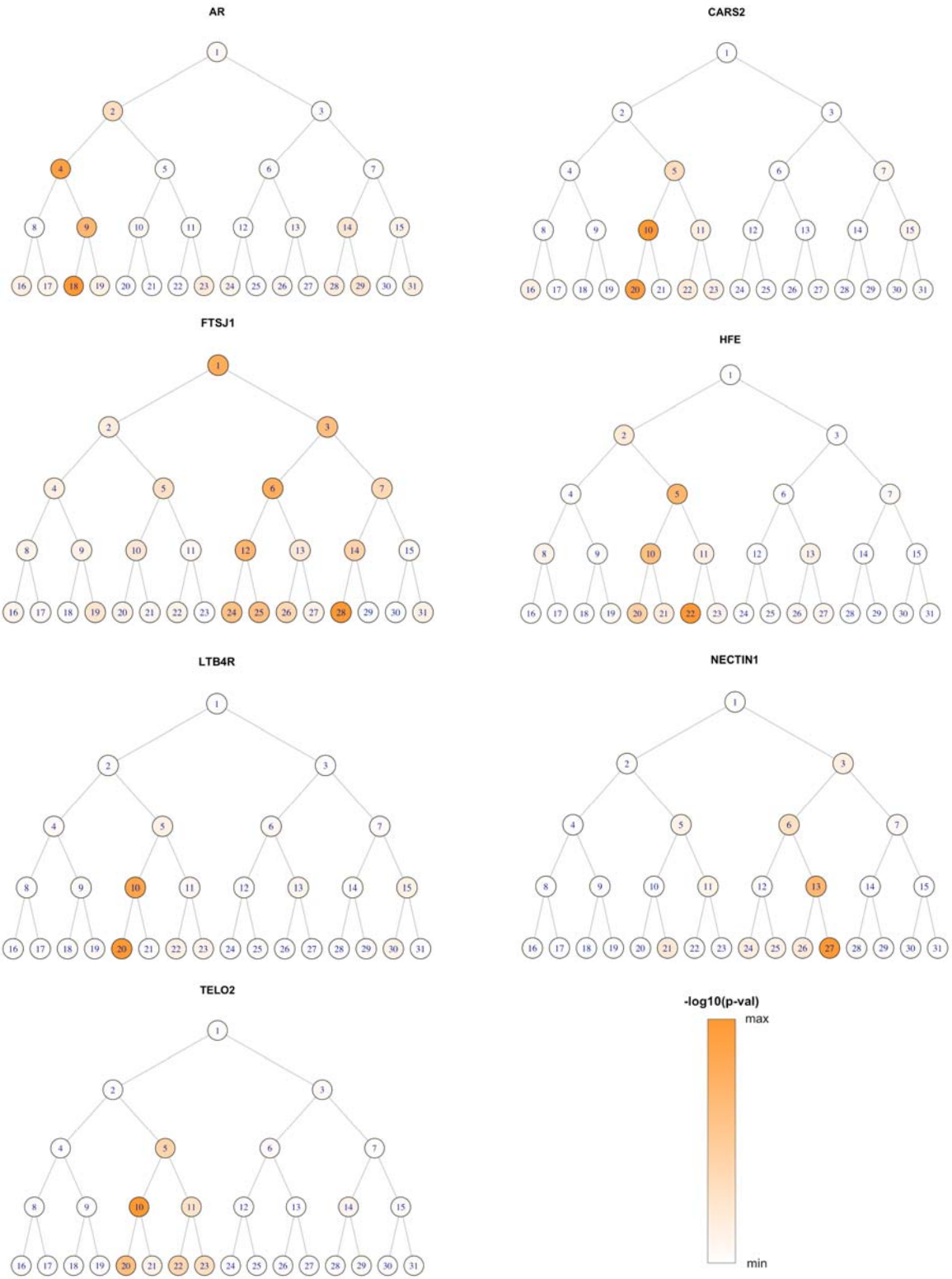

Fig 3. Module-wide association results for significant genes.

For each gene, the –log10 p-value is shown as color shades ranging from min to max, for 31 facial segments arranged the same way as Fig 1. The global-to-local phenotyping enabled the discovery of genetic effects at different scales.

To visualize gene effects on facial shape, we created the average module shape in non-carriers of the low-frequency variants for each gene, and the respective morph showing the change in shape from non-carriers to carriers (Fig 4). Blue and red indicate a local shape depression and protrusion, respectively, due to the low-frequency variants. For example, panel B in Fig 4 shows that *NECTIN1* variants shape the chin into a sharper and more protruding structure.

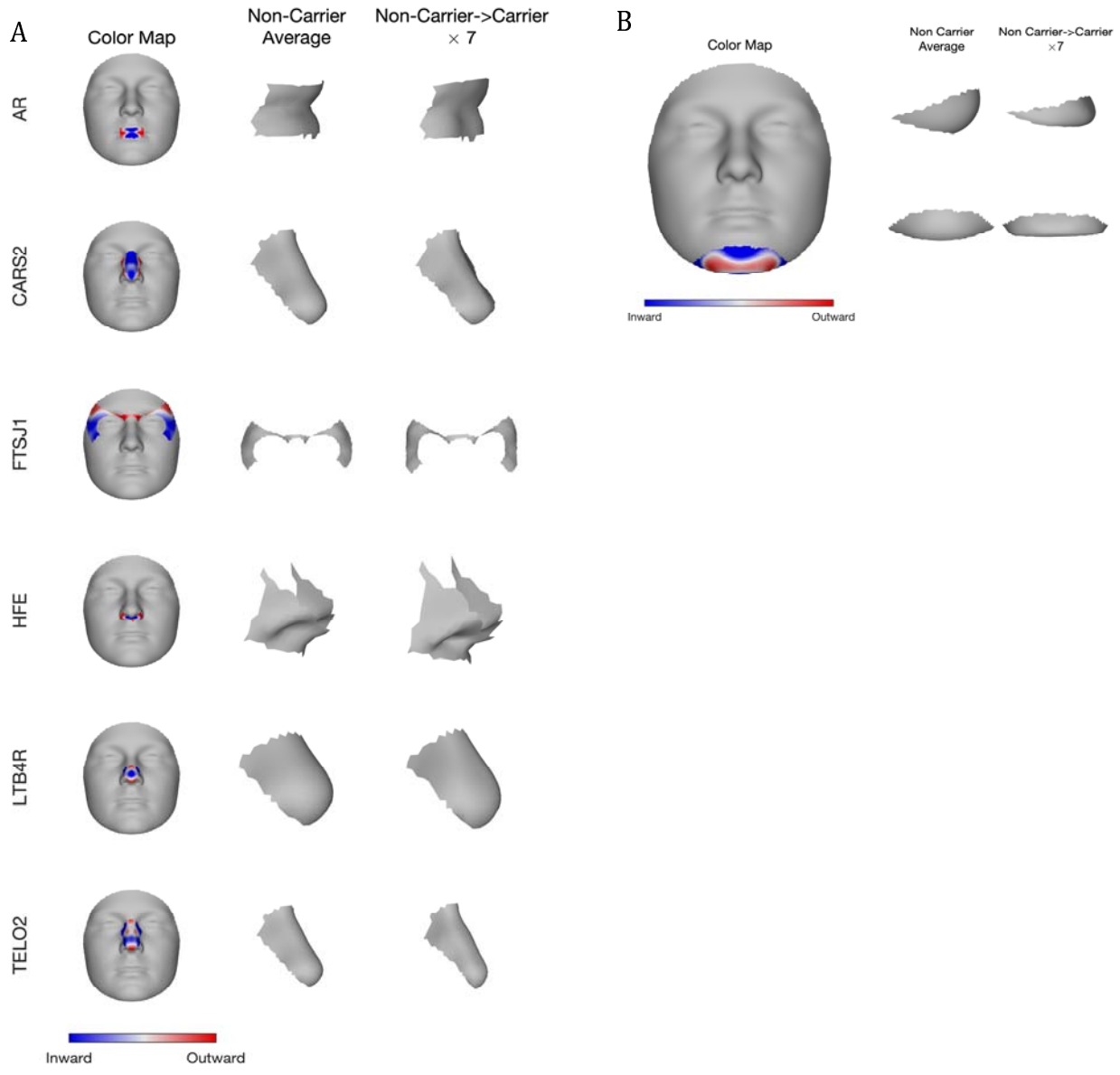

Fig 4. Phenotypic effect of the seven identified genes on their top associated module. Blue and red indicate a local shape depression and protrusion, respectively, due to the low-

frequency variants in the gene. A) First column shows gene effect on a representative module placing on the full face; middle column shows the lateral view of the average shape of the corresponding module among people who do not carry any variant in a gene; right column shows shape change of the same module, from non-carrier to carrier, multiplied by 7, to make the changes more clearly visible. B) For *NECTIN1* gene, we show both lateral (top) and frontal (bottom) view of its effect on chin shape. *NECTIN1* variants display a sharper, more protruding chin.

We employed various bioinformatics tools to explore the functions associated with the set of identified genes. Enrichment was detected for a variety of biological processes (Fig S2), especially ion-, metabolism-, transport- and regulation-related processes. Enriched gene ontology (GO) molecular functions tended to be housekeeping and general processes, e.g. signaling receptor and protein binding activity. Two genes with relatively well characterized functions, i.e. *HFE* and *AR* contributed a lot to these enrichment results. In GTEx database, these seven genes showed measurable expression level in adipose, skin and muscle-skeletal tissue (Fig S3), among which the strongest expression was seen for *NECTIN1* in skin.

To explore whether facial genes also affect the risk of orofacial clefts, results of gene-based associations of low-frequency (MAF<1%) variants with NSCL/P were retrieved from Leslie et al. 2017. Two out of the seven highlighted genes were not available from that study. S2 Table showed the SKAT and CMC test results for the other five genes, in the European, Asian, South American and total population. Two associations passed a Bonferroni corrected threshold for 40 tests (5 genes times 4 populations times 2 type of tests)—*TELO2* with a CMC p-value = 6.5×10^{-4}, and HFE with a CMC p-value = 1.1×10^{-3}, both in combined population of all ancestry groups.

Single variants were further tested individually with the corresponding facial module from the MultiPhen results (Table 1). Six SNPs showed nominal associations (p-value < 0.05) and the top association involved SNP rs142932029 in *FTSJ1* with module 28 (p-value = 1.59×10^{-14}). As shown in S4 Fig, these low-frequency variants had large effects compared to previously reported common variants [5].

Most of the individual variants appeared at frequencies much rarer than 1%, and all encode nonsynonymous substitutions except one splice site SNP in *FTSJ1*. Variants in *NECTIN1, CARS2* and *AR* are predicted to be deleterious based on CADD score (details in S3 Table). SNP rs137991779 in *NECTIN1* has a CADD score of 29.2, which ranks in the top 0.12% in deleteriousness among variants across the whole genome. PhenoScanner linked these

variants with a variety of human traits/disorders in previous studies (S4 Table, mostly from UK Biobank), including height, vascular diseases, osteoporosis, neoplasms etc., suggesting that coding variants influencing facial shape may be pleiotropic and play roles in other biological processes.

Zebrafish WISH was used to identify *ar, cars2, ftsj1, hfe, itb4r, telo2, nectin1a* and *nectin1b* expression pattern in craniofacial region across key developmental stages (Fig 5). At 24 hours post fertilization (hpf), *ftsj1* was expressed in the hindbrain, and *hfe* and *itb4r* were expressed in the forebrain. We detected nectin1a and nectin1b transcripts in the eyes, diencephalon, midbrain and hindbrain at 24 hpf. At 48 hpf, *ar* expression was detected in the epiphysis, and *cars2, nectin1a* and *nectin1b* were expressed in the palate (Fig 5 solid arrow). Moreover, *nectin1a* is also expressed in the lower jaw (Fig 5 hollow arrow).

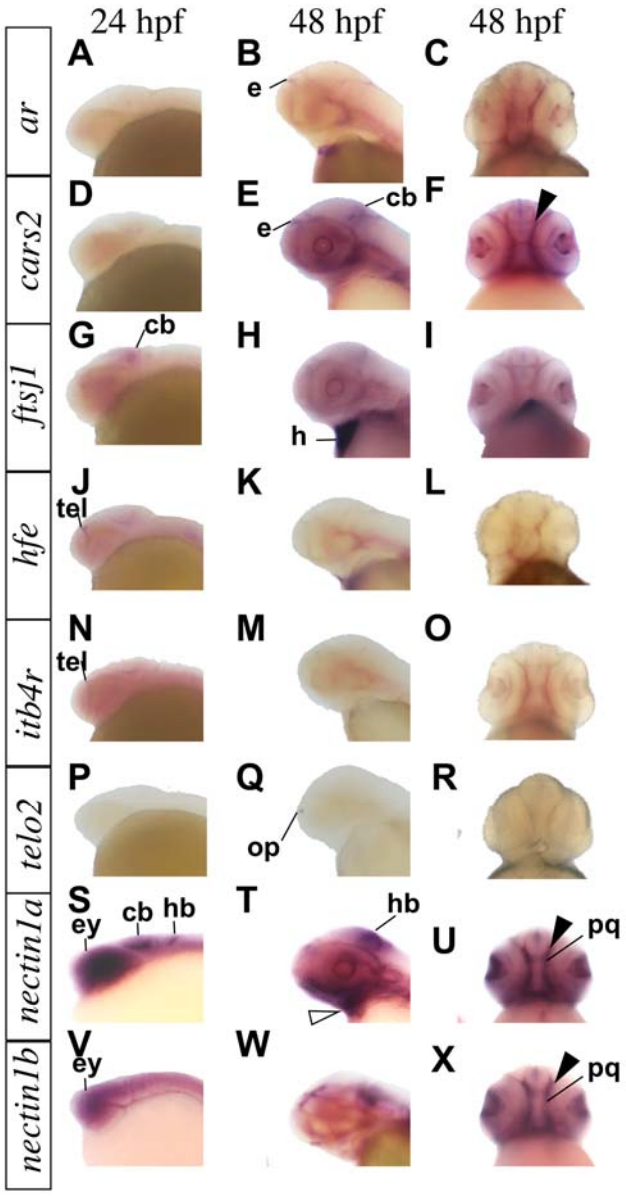

Fig 5. Whole-mount RNA in situ hybridization demonstrating genes expression in zebrafish.

Genes expression pattern in lateral and ventral views at the indicated embryonic stages as hours per fertilization (hpf). *cars2, nectin1a* and *nectin1b* are expressed in zebrafish palate (solid arrow). *nectin1a* is expressed in the lower jaw at 48 hpf (hollow arrow). cb: cerebellum, e: epiphysis, ey: eye, h: heart, hb: hindbrain, op: olfactory placode, pq: palate quadrate, tel: telencephalon.

To determine genetic requirement of *nectin1a* in craniofacial development, we analyzed the *nectin1a* mutant allele Ia021885Tg. Breeding of *nectin1a*+/- intercross generated embryos with Mendelian ratio (3 individuals with at least one wild type allele: 1 individual

homozygous for the mutant allele) demonstrating a mutant craniofacial phenotype, characterized by small head structures (Fig 6). Using Alcain blue staining at 120 hpf, *nectin1a* mutants displayed dysmorphic craniofacial development with smaller and distorted palate and abnormal Meckel's cartilage compared to age-matched wild type zebrafish embryos from the same intercross. These results show that *nectin1a* is expressed in the palate and is genetically required for normal palate and mandible morphogenesis.

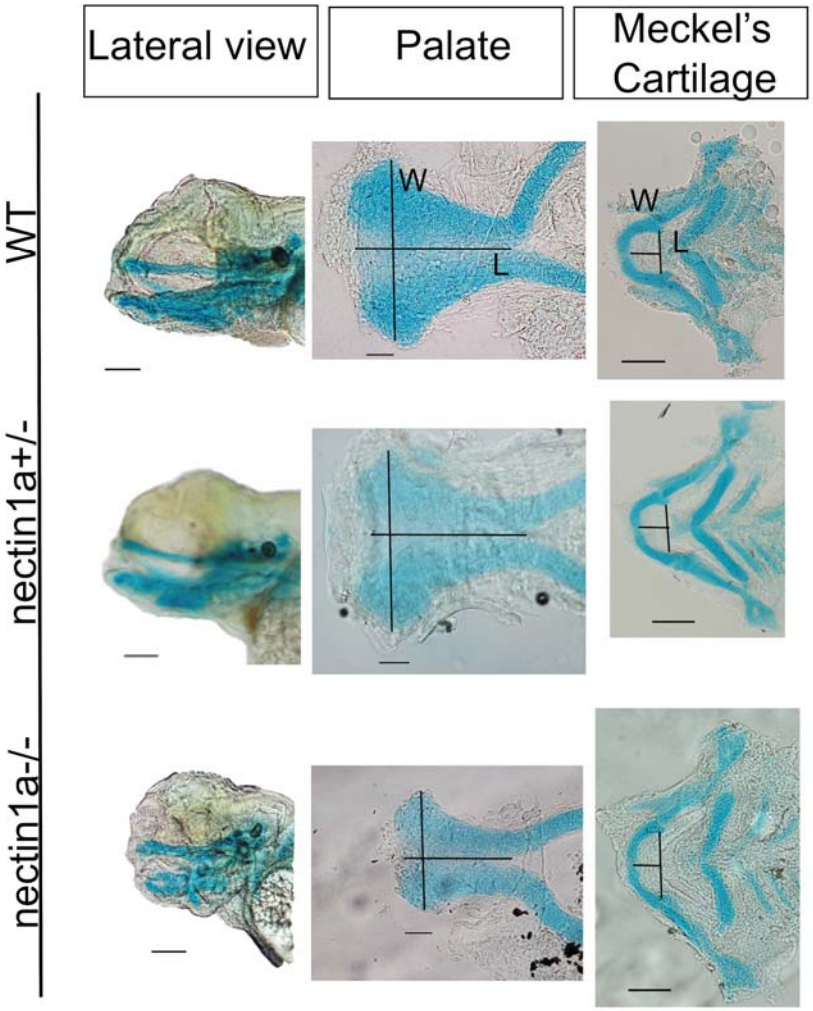

Fig 6. Alcian blue images for *nectin1a* zebrafish mutant comparted to wild type at day 5.

Wild type alcian blue lateral view, palate and Meckel's cartilage are the top images. Heterozygous *nectin1a* embryo alcian blue are the middle images. *nectin1a* mutant lateral images are below the heterozygous images. The length of the palate is measured from the anterior midpoint to the posterior midpoint of the palate. The width is measured as the maximum distance between the 2 lateral borders at the anterior area. However, the length of the Meckel's cartilage is measured from the midline of the Meckel's cartilage to the

midline of an imaginary line drawn joining the joints between the Meckel's cartilage and the palatoquadrate. The width is measured from the junction of the Meckel's cartilage and the palatoquadrate of one side to the other side. *nectin1a* mutant have smaller and shorter palate, whereas shorter and wider Meckel's cartilage compared to wild type. L: length, W: width. Scale bar: 10 μm

Discussion

This study presented a discovery effort to identify low-frequency coding variants associated with normal-range human facial morphology by undertaking gene-based association tests and subsequent analyses on a carefully phenotyped cohort genotyped on the Illumina Exome chip. Overall, we demonstrated that part of the morphological variation of facial shape is attributable to low-frequency coding variants, and pinpointed putative functional genes involved. *AR*, *CARS2*, *FTSJ1*, *HFE/LOC108783645*, *LTB4R*, *TELO2* and *NECTIN1* were implicated, with phenotypic effects in the area of cheek, chin, nose and philtrum. Notably, *NECTIN1* is known to cause orofacial clefts, a craniofacial malformation that can be associated with alterations in facial shape. Using a zebrafish model, we confirmed the expression of *nectin1a* and *nectin1b* in the developing head and the abnormal craniofacial phenotype in *nectin1a* mutant. Taken together, these findings support the contribution of low-frequency coding variants to the morphogenesis of facial structures and the genetic architecture of normal-range facial shape.

The seven genes identified by the multivariate approach are implicated in normal facial morphology for the first time to our knowledge. Their related-cellular processes/functions include metal ion transport (*HFE*), signaling (*AR*, *LTB4R*), tRNA metabolism (*CARS2, FTSJ1*), DNA repair (*TELO2*) and cell adhesion (*NECTIN1*). This diversity in functions led to a variety of functional pathways/categories showing up in our enrichment analysis, yet without strong signal in any particular one, probably due to the small number of genes as input and the nature of genetic architecture of morphological traits. With the exception *NECTIN1*, the role of these genes in patterning craniofacial tissue is largely unknown, and further investigation may yield more insights into normal and abnormal facial development.

Previous GWASs and studies of facial dysmorphology have demonstrated that there are common genetic factors underlying both normal-range facial variation and orofacial clefting [5,11,34]. Our findings suggest that low-frequency coding variants may also help explain this relationship. Although none of the other genes implicated here have been shown to have a direct involvement in craniofacial development, *NECTIN1* is an established player that has been linked both syndromic and isolated forms of orofacial clefting [35-37].

Individuals with cleft lip/palate-ectodermal dysplasia syndrome (OMIM:225060) have distinctive facial features including an underdeveloped lower jaw [38], which is consistent with the facial segment (chin) where *NECTIN1* association was observed. NECTIN1 protein belongs to the subfamily of immunoglobulin-like adhesion molecules that are key components of cell adhesion junctions, playing important roles in the fusion of palatal shelves during palatogenesis [39]. A handful of *NECTIN1* mutations that can potentially disrupt gene function have also been documented in non-syndromic cleft patients [40-42]. In our cohort, two coding variants *NECTIN1* were implicated, and both are predicted to be deleterious to protein function. We did lookups of the face-associated genes in a previous exome scan of NSCL/P cohort [24], and *NECTIN1* yielded a p-value of 0.004, although not passing the Bonferroni significance threshold. Two other genes, *TELO2* and *HFE*, did pass that threshold. These results are in line with previous evidence suggesting a role for same genes in normal and abnormal facial development.

The detected expression of all genes identified except *cars2* in the zebrafish embryonic head demonstrated their potential involvement in craniofacial development. Furthermore, our *nectin1a* knockout displayed altered shape and size of Meckel's cartilage. This affected structure is in accordance with the human anatomical region (mandible, lower jaw), which was associated with *NECTIN1* in our MultiSKAT test. We highlight the approach of interrogating human candidate genes in a biological context using the zebrafish model, where dynamic gene expression can be assayed in a high throughput fashion. Those candidate genes with spatiotemporal gene expression in the craniofacial domains then can be evaluated in functional studies, were mutants may already be available from large scale mutagenesis projects, or can be generated by CRISPR mediated gene editing.

With the hierarchical facial modules, we were able to pinpoint genetic effects at different scales. For example, the effect of *FTSJ1* was observed globally on the whole face, and also locally in specific modules on the side of the face. By contrast, the effect of *NECTIN1* was confined to localized facial parts only. These patterns may help with understanding the mechanisms by which genes take effect along the growth of facial structure. Our multivariate data-driven phenotyping approach eliminates the need of preselecting traits, captures more variation in the facial shape and thus displayed high efficiency for gene mapping.

The current study is an important extension and complement of our prior work [5]. Here we exclusively focus on coding variants with MAF below 1%, which have been omitted based on MAF filter from previous facial GWAS attempts. Importantly, our results generated distinct, non-overlapping knowledge about facial genetics. When we compared

our new results to those from a prior GWAS of this cohort [5], common variants in or near (within 500kb) the seven associated genes showed no evidence of association ($p > 0.001$ for all) at the same facial modules. Nonetheless, it is possible that there are trans-acting common GWAS SNPs that regulate the expression of the seven newly identified genes during facial morphogenesis. Low-frequency variants showed large magnitude of effects compared to common variants in [5]. It is necessary, however, to point out that this difference could partially or completely be a result of the drastically smaller group of variant carriers, and we therefore refrain from overinterpreting the comparison.

Our study demonstrated the power of applying gene-based tests of low-frequency variants that are usually untestable individually. While some significant genes harbor variants with a small p-value in our single-variant association test, others would have been missed if not tested in aggregate. With a moderate sample size of 2329, it is highly desirable to collapse low-frequency variants into putative functional units and perform burden-style tests. One explanation for the observed weak enrichment signals could be insufficient power, and we expect future well-powered studies to discover more biological pathways emerging from analyses of low-frequency coding variants.

For variants occurring at a low frequency in the population, attempting to replicate the signal is difficult. The prominent barrier is the limited sample size. The low numbers or even absence of the carriers in independent populations hindered the replication efforts of our findings. We found that six out of the seven genes were not testable in a separate cohort of 664 Whites participants due to low copy number of the variants. Despite all the benefits of targeting low-frequency variants, those in the lower extreme of the MAF spectrum were still missed from the current analysis. Given the limited sample size and the ExomeChip design, this study was not adequately powered to identify genes harboring very rare variants that may also contribute to facial traits. By our filtering criteria, variants with a MAF < 0.08% were out of the scope of the analyses. Although complex traits are not expected to have a large fraction of the heritability explained by rare and private variants, such variants may be influential, predictive, and actionable at the individual level. In this regard, whole exome or whole genome sequencing of large samples holds promise.

Like many other complex traits, research focused on uncovering the genetic architecture of facial morphology is confronted with the challenge of missing heritability [43,44]. Our study has extended the paradigm of genetic involvement in facial development from common to low frequency variants, and highlighted novel candidate genes that may lead to encouraging follow-ups. Given that rare and low-frequency genetic variation might be highly specific to certain populations, and facial shapes have distinctive ancestry features,

future studies may benefit from extending the discovery of influential low-frequency variants to other ethnic groups.

Acknowledgements

The authors thank all the dedicated staff, collaborators, and participants for their contribution to the study.

Funding

This study was supported by the National Institute for Dental and Craniofacial Research (NIDCR, (http://www.nidcr.nih. gov/) through the following grants: U01-DE020078 to SMW and MLM, R01-DE016148 to SMW and MLM, R01-DE027023 to SMW and JRS, and X01-HG007821 to MLM, SMW, and EF. Funding for initial genomic data cleaning by the University of Washington was provided by contract #HHSN268201200008I from the NIDCR awarded to the Center for Inherited Disease Research (CIDR, https://www.cidr.jhmi.edu/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability

All of the genotypic markers are available to the research community through the dbGaP controlled-access repository (https://dbgap.ncbi.nlm.nih.gov/) at accession phs000949.v1.p1. The raw source data for the phenotypes – the 3D facial surface models – are available for the 3D Facial Norms dataset through the FaceBase Consortium (www.facebase.org). Access to these 3D facial surface models requires proper institutional ethics approval and approval from the FaceBase data access committee. KU Leuven provides the spatially dense facial mapping software, free to use for academic purposes: MeshMonk (https://github.com/TheWebMonks/meshmonk).

Reference

1. Weinberg SM, Roosenboom J, Shaffer JR, Shriver MD, Wysocka J, Claes P. Hunting for genes that shape human faces: Initial successes and challenges for the future. Orthod Craniofac Res. 2019;22 Suppl 1: 207–212. doi:10.1111/ocr.12268

2. Weinberg SM, Cornell R, Leslie EJ. Craniofacial genetics: Where have we been and where are we going? PLoS Genet. 2018;14: e1007438. doi:10.1371/journal.pgen.1007438

3. Richmond S, Howe LJ, Lewis S, Stergiakouli E, Zhurov A. Facial Genetics: A Brief Overview. Front Genet. 2018;9: 462. doi:10.3389/fgene.2018.00462

4. Cha S, Lim JE, Park AY, Do J-H, Lee SW, Shin C, et al. Identification of five novel genetic loci related to facial morphology by genome-wide association studies. BMC GENOMICS. BMC Genomics; 2018;19: 481–17. doi:10.1186/s12864-018-4865-9

5. Claes P, Roosenboom J, White JD, Swigut T, Sero D, Li J, et al. Genome-wide mapping of global-to-local genetic effects on human facial shape. Nat Genet. Springer US; 2018;50: 1–16. doi:10.1038/s41588-018-0057-4

6. Crouch DJM, Winney B, Koppen WP, Christmas WJ, Hutnik K, Day T, et al. Genetics of the human face: Identification of large-effect single gene variants. Proc Natl Acad Sci USA. 2018;115: E676–E685. doi:10.1073/pnas.1708207114

7. Lee MK, Shaffer JR, Leslie EJ, Orlova E, Carlson JC, Feingold E, et al. Genome-wide association study of facial morphology reveals novel associations with FREM1 and PARK2. Li Y, editor. PLoS ONE. 2017;12: e0176566–13. doi:10.1371/journal.pone.0176566

8. Shaffer JR, Orlova E, Lee MK, Leslie EJ, Raffensperger ZD, Heike CL, et al. Genome-Wide Association Study Reveals Multiple Loci Influencing Normal Human Facial Morphology. Barsh GS, editor. PLoS Genet. Public Library of Science; 2016;12: e1006149–21. doi:10.1371/journal.pgen.1006149

9. Cole JB, Manyama M, Kimwaga E, Mathayo J, Larson JR, Liberton DK, et al. Genomewide Association Study of African Children Identifies Association of SCHIP1 and PDE8A with Facial Size and Shape. Barsh GS, editor. PLoS Genet. 2016;12: e1006174. doi:10.1371/journal.pgen.1006174

10. Adhikari K, Fuentes-Guajardo M, Quinto-Sánchez M, Mendoza-Revilla J, Camilo Chacón-Duque J, Acuña-Alonzo V, et al. A genome-wide association scan implicates DCHS2, RUNX2, GLI3, PAX1 and EDAR in human facial variation. Nature Communications. 2016;7: 11616. doi:10.1038/ncomms11616

11. Liu F, van der Lijn F, Schurmann C, Zhu G, Chakravarty MM, Hysi PG, et al. A Genome-Wide Association Study Identifies Five Loci Influencing Facial Morphology in Europeans. Gibson G, editor. PLoS Genet. Public Library of Science; 2012;8: e1002932–13. doi:10.1371/journal.pgen.1002932

12. Paternoster L, Zhurov AI, Toma AM, Kemp JP, Pourcain BS, Timpson NJ, et al. REPOR T Genome-wide Association Study of Three-Dimensional Facial Morphology Identifies a Variant in PAX3 Associated with Nasion Position. The American Journal of Human Genetics. The American Society of Human Genetics; 2012;90: 478–485. doi:10.1016/j.ajhg.2011.12.021

13. Marouli E, Graff M, Medina-Gomez C, Lo KS, Wood AR, Kjaer TR, et al. Rare and low-frequency coding variants alter human adult height. Nature. Nature Publishing Group; 2017;542: 186–190. doi:10.1038/nature21039

14. Lu X, Peloso GM, Liu DJ, Wu Y, Zhang H, Zhou W, et al. Exome chip meta-analysis identifies novel loci and East Asian–specific coding variants that contribute to lipid levels and coronary artery disease. Nat Genet. 2017;49: 1722–1730. doi:10.1038/ng.3978

15. Liu DJ, Peloso GM, Yu H, Butterworth AS, Wang X, Mahajan A, et al. Exome-wide association study of plasma lipids in >300,000 individuals. Nat Genet. Nature Publishing Group; 2017;49: 1758–1766. doi:10.1038/ng.3977

16. Dutta D, Scott L, Boehnke M, Lee S. Multi-SKAT: General framework to test for rare-variant association with multiple phenotypes. 2019;43: 4–23. doi:10.1002/gepi.22156

17. White JD, Ortega-Castrillón A, Matthews H, Zaidi AA, Ekrami O, Snyders J, et al. MeshMonk: Open-source large-scale intensive 3D phenotyping. scientific reports. Nature Publishing Group; 2019;9: 6085. doi:10.1038/s41598-019-42533-y

18. Auer PL, Reiner AP, Leal SM. The effect of phenotypic outliers and non-normality on rare-variant association testing. Eur J Hum Genet. Nature Publishing Group; 2016;24: 1188–1194. doi:10.1038/ejhg.2015.270

19. Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity. Nature Publishing Group; 2005;95: 221–227. doi:10.1038/sj.hdy.6800717

20. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. Nature Biotechnology. Nature Publishing Group; 2010;28: 495–501. doi:10.1038/nbt.1630

21. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. Nature Communications. Nature Publishing Group; 2017;8: 1826. doi:10.1038/s41467-017-01261-5

22. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. 2009;37: W305–W311. doi:10.1093/nar/gkp427

23. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet. Nature Publishing Group; 2013;45: 580–585. doi:10.1038/ng.2653

24. Leslie EJ, Carlson JC, Shaffer JR, Buxó CJ, Castilla EE, Christensen K, et al. Association studies of low-frequency coding variants in nonsyndromic cleft lip with or without cleft palate. Am J Med Genet. 2017;173: 1531–1538. doi:10.1002/ajmg.a.38210

25. O'Reilly PF, Hoggart CJ, Pomyen Y, Calboli FCF, Elliott P, Jarvelin M-R, et al. MultiPhen: Joint Model of Multiple Phenotypes Can Increase Discovery in GWAS.

Cherny S, editor. PLoS ONE. Public Library of Science; 2012;7: e34861–12. doi:10.1371/journal.pone.0034861

26. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47: D886–D894. doi:10.1093/nar/gky1016

27. Kamat MA, Blackshaw JA, Young R, Surendran P, Burgess S, Danesh J, et al. PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. Kelso J, editor. Bioinformatics. 2019;35: 4851–4853. doi:10.1093/bioinformatics/btz469

28. Thisse C, Thisse B. High-resolution in situ hybridization to whole-mount zebrafish embryos. Nat Protoc. Nature Publishing Group; 2008;3: 59–69. doi:10.1038/nprot.2007.514

29. Kimmel CB, Ballard WW, Kimmel SR, Ullmann B, Schilling TF. Stages of Embryonic-Development of the Zebrafish. Developmental Dynamics. 1995;203: 253–310. doi:10.1002/aja.1002030302

30. Sivasubbu S, Balciunas D, Amsterdam A, Ekker SC. Insertional mutagenesis strategies in zebrafish. Genome Biol. BioMed Central; 2007;8: S9. doi:10.1186/gb-2007-8-s1-s9

31. Amsterdam A, Hopkins N. Retroviral-Mediated Insertional Mutagenesis in Zebrafish. Methods in Cell Biology. Academic Press; 2003;77: 3–20. doi:10.1016/S0091-679X(04)77001-6

32. Walker MB, Kimmel CB. A two-color acid-free cartilage and bone stain for zebrafish larvae. Biotech Histochem. 2007;82: 23–28. doi:10.1080/10520290701333558

33. Westerfield M. A guide for the laboratory use of zebrafish Danio (Brachydanio) rerio. In: University of Oregon Press, Eugene; 1994.

34. Weinberg SM, Naidoo SD, Bardi KM, Brandon CA, Neiswanger K, Resick JM, et al. Face shape of unaffected parents with cleft affected offspring: combining three-dimensional surface imaging and geometric morphometrics. Orthod Craniofac Res. John Wiley & Sons, Ltd (10.1111); 2009;12: 271–281. doi:10.1111/j.1601-6343.2009.01462.x

35. Sozen MA, Suzuki K, Tolarova MM, Bustos T, Fernández Iglesias JE, Spritz RA. Mutation of PVRL1 is associated with sporadic, non-syndromic cleft lip/palate in northern Venezuela. Nat Genet. Nature Publishing Group; 2001;29: 141–142. doi:10.1038/ng740

36. Avila JR, Jezewski PA, Vieira AR, Orioli IM, Castilla EE, Christensen K, et al. PVRL1 variants contribute to non-syndromic cleft lip and palate in multiple populations.

American Journal of Medical Genetics Part A. 2006;140: 2562–2570. doi:10.1002/ajmg.a.31367

37. Suzuki K, Hu D, Bustos T, Zlotogora J, Richieri-Costa A, Helms JA, et al. Mutations of PVRL1, encoding a cell-cell adhesion molecule/herpesvirus receptor, in cleft lip/palate-ectodermal dysplasia. Nat Genet. 2000;25: 427–430. doi:10.1038/78119

38. Zlotogora J. Syndactyly, Ectodermal Dysplasia, and Cleft Lip/Palate. Journal of Medical Genetics. 1994;31: 957–959. doi:10.1136/jmg.31.12.957

39. Cobourne MT. The complex genetics of cleft lip and palate. Eur J Orthod. 2004;26: 7–16.

40. Oner DA, Tastan H. Identification of Novel Variants in the PVRL1 Gene in Patients With Nonsyndromic Cleft Lip With or Without Cleft Palate. Genetic Testing and Molecular Biomarkers. 2016;20: 269–272. doi:10.1089/gtmb.2015.0276

41. Tongkobpetch S, Suphapeetiporn K, Siriwan P, Shotelersuk V. Study of the poliovirus receptor related-1 gene in Thai patients with non-syndromic cleft lip with or without cleft palate. International Journal of Oral & Maxillofacial Surgery. 2008;37: 550–553. doi:10.1016/j.ijom.2008.01.024

42. Scapoli L, Palmieri A, Martinelli M, Vaccari C, Marchesini J, Pezzetti F, et al. Study of the PVRL1 gene in Italian nonsyndromic cleft lip patients with or without cleft palate. Annals of Human Genetics. Blackwell Publishing, Inc; 2006;70: 410–413. doi:10.1111/j.1529-8817.2005.00237.x

43. Tsagkrasoulis D, Hysi P, Spector T, Montana G. Heritability maps of human face morphology through large-scale automated three-dimensional phenotyping. scientific reports. Nature Publishing Group; 2017;7: 1–18. doi:10.1038/srep45885

44. Cole JB, Manyama M, Larson JR, Liberton DK, Ferrara TM, Riccardi SL, et al. Human Facial Shape and Size Heritability and Genetic Correlations. Genetics. Genetics; 2017;205: 967–978. doi:10.1534/genetics.116.193185

Supporting Information

S1 Fig. Q-Q plot of gene-based MultiSKAT tests by facial module

S2 Fig. FUMA enrichment results

S3 Fig. GTEx expression of MultiSKAT significant genes in tissues relevant to facial morphology. Dendrogram denotes similarity in expression level. TPM, transcripts per million

S4 Fig. Magnitude of variant effect on facial modules, quantified by the Euclidean distance between averaged faces of different genotype groups. The 95% confidence

interval was obtained by 5000 bootstraps. The farther away the blue (common) or red (low-freq) rectangular boxes fall from line x=0, the larger the group distances and the greater the magnitude of effects. Common variants that yielded significant GWAS association in the same cohort with the same modules are used as a comparison to low-frequency variants. Genotype groups column indicates the two groups of people of whom the faces were averaged and distance was computed. For example, 0 vs 1/2 means minor allele homozygotes vs the remaining. The following two columns indicate sizes of the two groups in comparison. Low-frequency variants had large effects compared to previously reported common variants, although this could be a result from the much smaller size of carrier group and may not reflect genuine greater effects of low-frequency variants.

S1 Table. Module-wide association results of genes identified by MultiSKAT. Show modules with a p-value < 10E-4.

S2 Table. SKAT and CMC test results of the association between the seven facial genes and NSCL/P, retreived from a previous exome-wide gene-based association study of NSCL/P

S3 Table. Functional prediction of individual variants in significant genes by CADD GRCh37-v1.4

S4 Table. PhenoScanner lookups for variants in seven significant genes. Show existing associations involving these variants with a p value < 10e-4.