Genome-wide analyses of non-syndromic cleft lip with palate identify 14 novel loci and genetic heterogeneity

Yanqin Yu1,*, Xianbo Zuo2,3,4,5,6,*, Miao He1,7,*, Jinping Gao2,3,4,5,6,*, Yuchuan Fu8, Chuanqi Qin1,8, Liuyan Meng1, Wenjun Wang2,3,4,5,6, Yaling Song1, Yong Cheng1, Fusheng Zhou2,3,4,5,6, Gang Chen2,3,4,5,6, Xiaodong Zheng2,3,4,5,6, Xinhuan Wang1, Bo Liang2,3,4,5,6, Zhengwei Zhu2,3,4,5,6, Xiazhou Fu9, Yujun Sheng2,3,4,5,6, Jiebing Hao10, Zhongying Liu11, Hansong Yan12, Elisabeth Mangold13, Ingo Ruczinski14, Jianjun Liu2,3,4,5,6, Mary L. Marazita15,16,17, Kerstin U. Ludwig13,18, Terri H. Beaty19, Xuejun Zhang2,3,4,5,6,20,21, Liangdan Sun2,3,4,5,6,22,** & Zhuan Bian1,**

Non-syndromic cleft lip with palate (NSCLP) is the most serious sub-phenotype of non-syndromic orofacial clefts (NSOFC), which are the most common craniofacial birth defects in humans. Here we conduct a GWAS of NSCLP with multiple independent replications, totalling 7,404 NSOFC cases and 16,059 controls from several ethnicities, to identify new NSCLP risk loci, and explore the genetic heterogeneity between sub-phenotypes of NSOFC. We identify 41 SNPs within 26 loci that achieve genome-wide significance, 14 of which are novel (RAD54B, TMEM19, KRT18, WNT9B, GSC/DICER1, PTCH1, RPS26, OFCC1/TFAP2A, TAF1B, FGF10, MSX1, LINC00640, FGFR1 and SPRY1). These 26 loci collectively account for 10.94% of the heritability for NSCLP in Chinese population. We find evidence of genetic heterogeneity between the sub-phenotypes of NSOFC and among different populations. This study substantially increases the number of genetic susceptibility loci for NSCLP and provides important insights into the genetic aetiology of this common craniofacial malformation.

1 The State Key Laboratory Breeding Base of Basic Science of Stomatology (Hubei-MOST) and Key Laboratory of Oral Biomedicine Ministry of Education, School and Hospital of Stomatology, Wuhan University, Wuhan, Hubei 430079, China. 2 Institute of Dermatology and Department of Dermatology at No. 1 Hospital, Anhui Medical University, Hefei, Anhui 230032, China. 3 State Key Lab Incubation of Dermatology, Ministry of Science and Technology, Hefei, China. 4 Key Lab of Dermatology, Ministry of Education, Hefei, China. 5 Key Lab of Gene Resources Utilization for Severe Inherited Disorders, Anhui 230032, China. 6 Collaborative Innovation Center of Complex and Severe skin Disease, Anhui Medical University, Hefei, Anhui 230032, China. 7 Department of Pediatric Dentistry, School and Hospital of Stomatology, Wuhan University, Wuhan, Hubei 430079, China. 8 Department of Oral and Maxillofacial Surgery, School and Hospital of Stomatology, Wuhan University, Wuhan, Hubei 430079, China. 9 Department of Genetics and Centre for Developmental Biology, College of Life Science, Wuhan University, Hubei, 430072, China. 10 Stomatological Hospital of Xiangyang, Xiangyang, Hubei 441011, China. 11 Institute of Human Genetics, Life and Brain Center, University of Bonn, 53127 Bonn, Germany. 12 Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland 21205, USA. 13 Department of Oral Biology and Center for Craniofacial and Dental Genetics, School of Dental Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania 15219, USA. 14 Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania 15261, USA. 15 Department of Epidemiology, School of Public Health, Johns Hopkins University, Baltimore, Maryland 21205, USA. 16 Department of Dentistry at No. 2 Hospital, Anhui Medical University, Hefei, Anhui 230022, China. 17 Institute of Dermatology and Department of Dermatology, Huashan Hospital of Fudan University, Shanghai 200040, China. 18 Key Laboratory of Major Autoimmune Diseases, Anhui Province, Anhui 230032, China. * These authors contributed equally to this work. ** These authors jointly supervised this work. Correspondence and requests for materials should be addressed to L.S. (email: ahmusld@163.com) or to Z.B. (email: bianzhuan@whu.edu.cn).
Orofacial clefts (OFCs) are the most common craniofacial malformations in humans and present a major public health burden, imposing substantial health care and financial burdens on the individual, their family and society. In general, the highest birth prevalence rates of OFCs are reported in Asia (especially in China and Japan), often as high as 1 in 500 and affecting more than 2.6 million people in China. According to whether the patients have other malformations or anomalies, OFCs can be divided into syndromic and non-syndromic forms. Approximately 70% of cleft lip (CL) with or without cleft palate (CP) cases and 50% of CP only (CPO) cases occur, as isolated entities with no other abnormal phenotypes are considered to be non-syndromic (referred to as NSOF/C)1–3. NSOF/C is further classified into non-syndromic cleft lip with palate (NSCL/P), non-syndromic CL only (NSCLO) and non-syndromic CPO (NSCPO), based on the anatomical morphology. As they share common epidemiological patterns and occur during the same embryological period, NSCL/P and NSCLO are often grouped together as non-syndromic CL with or without CP (NSCL/P), differing only in severity. However, there is some evidence showing that NSCL/P and NSCLO might harbour different genetic aetiologies4–9.

Multiple genome-wide association study (GWAS) and relative extension studies of NSCLO/P have been performed, and 22 susceptibility loci were identified15–16, including the 1q32 (IRF6) locus, which was observed in previous candidate gene studies and subsequently confirmed in several GWASs7,10,12,14–16. However, only one GWAS of NSCL/P was conducted in a Chinese population15 and thus the heritability in the risk of NSOF/C remains unexplained in China, especially for the three distinct sub-phenotypes of NSOFC, here we perform a NSCL/P GWAS using two independent case–control samples from China and replicate interesting markers in a total of 23,463 samples from sub-phenotypes of NSOF/C and multiple ethnic groups. We identify 14 new loci and confirm 12 previously reported ones for NSCL/P. These susceptibility variants identified in the current study collectively account for 10.94% of the heritability for NSCL/P in Chinese population. In addition, evidence of genetic heterogeneity is observed between the three sub-phenotypes of NSOF/C and among different populations.

**Results**

**Identification of 26 NSCLP-associated loci.** In the discovery stage, we genotyped 900,015 single-nucleotide polymorphisms (SNPs) using the Illumina HumanOmniZhongHua-8 BeadChip in 2,096 cases and 4,051 controls of Chinese ancestry (cohort 1). After quality control, 803,202 SNPs (call rate >95% and minor allele frequency (MAF) >1%) in 2,033 NSCLP cases and 4,051 controls of Chinese ancestry were used in the GWAS discovery analysis (Fig. 1 and Supplementary Table 1). The Manhattan plot of P-values using Cochran–Armitage trend test with adjustment for gender is shown in Supplementary Fig. 1. All cases and controls were assessed by principal components analysis for population stratification and were confirmed to be of Chinese ancestry (Supplementary Fig. 2). Quantile–quantile plots were constructed and genomic control values were calculated ($\lambda_{GC} = 1.04$) (Supplementary Fig. 3). Both of these results indicate negligible inflation of the genome-wide association signals caused by population stratification, further suggesting that the deviated tail of the P-values’ distribution reflects some true genetic associations with NSCLP. We then conducted logistic regression analysis to assess the genotype–phenotype association.

To perform a fast-track replication study, we selected and genotyped 152 SNPs ($P < 1 \times 10^{-4}$) within 79 loci for a follow-up analysis in an additional 1,346 NSCLP cases and 4,542 controls of Chinese ancestry (cohort 2). Of the 146 successfully genotyped SNPs, 64 showed nominal association ($P < 0.05$ using logistic regression) in the validation stage and 61 of them showed a consistent direction in their estimated effects on risk between the discovery (cohort 1) and validation (cohort 2) stages (Supplementary Table 2). A fixed-effects meta-analysis of the combined cohorts 1 and 2, totalling 3,379 NSCLP cases and 8,593 controls, identified 14 new loci (20 SNPs) ($P < 5.00 \times 10^{-5}$ using Cochran–Mantel–Haenszel test), namely 2p25.1, 4p16.2, 4q28.1, 5p12, 6p24.3, 8p11.23, 8q22.1, 9q22.32, 12q13.13, 12q21.1, 14q22.1, 14q32.13 and 17q21.32, and three suggestive loci 2q35, 8q22.2 and 20q13.2 (Table 1, Fig. 2 and Supplementary Table 3). We also confirmed 12 reported loci (21 SNPs): 1p22.1, 1q32.2, 2p24.2, 8q21.3, 8q24.21, 9q22.2, 10q25.3, 13q31.1, 16p13.3, 17p13.1, 17q22 and 20q12 ($P < 5.00 \times 10^{-5}$) (Fig. 2 and Supplementary Table 4). All these 26 susceptibility loci collectively account for 10.94% of the NSCLP heritability. In addition, conditional analyses were performed for all 26 loci and we identified a secondary signal in one previously reported locus at 1q32.2 (Supplementary Table 5). After reviewing the published GWASs of NSCL/P and the present study, we summarize the susceptibility loci identified in different populations in Supplementary Fig. 4.

**Replications of the 26 loci in sub-phenotype groups of NSOFC.** We successfully genotyped 40 of the 41 SNPs (1 SNP, rs481931 at 1p22.1, was unsuccessfully genotyped) from the 26 loci in cohort 3 (NSCLO) and cohort 4 (NSCPO). Two novel (1q32.13 and 17q21.32) and eight reported loci (1p22.1, 1q32.2, 2p24.2, 8q21.3, 9q22.2, 10q25.3, 17p13.1 and 20q12) showed significant associations ($P_{\text{Bonferroni}} < 1.25 \times 10^{-3}$ using logistic regression test and Bonferroni correction; 0.05 out of 40) with NSCLO (Supplementary Table 6). All the associated SNPs from the above ten loci have concordant associations in the effect sizes and direction in both NSCL/P and NSCLO (Supplementary Tables 3, 4 and 6). Two loci (13q31.1 and 15q13.3) were reported to be more strongly associated with NSCLO than NSCL/P7–9. We also found rs9545308 at 13q31.1 to be significantly associated with NSCLO ($P_{\text{NSCLO}} = 4.95 \times 10^{-3}$, OR = 1.23) in our Chinese samples. The marker at 15q13.3 was not successfully replicated in NSCL/P and thus was not chosen to be replicated in NSCLO in our study.

One novel (9q22.32) and two reported loci (1q32.2 and 8q24.21) showed significant associations with NSCPO (Supplementary Table 7). The marker in 1q32.2 showed opposite directions of association between the NSCLO and NSCPO groups (rs9430019; OR$_{\text{NSCLO}} = 1.25$ and OR$_{\text{NSCPO}} = 0.66$), whereas the markers in the 8q24.21 and 9q22.32 loci were concordant in the estimated direction of association with NSCLP (Supplementary Tables 3, 4 and 7). It is worth mentioning that the recent GWAS57 and sequencing study58 revealed an aetiological missense variant in GRHL3 (1p36.11) for NSCPO. The additional locus 9q22.33 (FOXE1) was identified potentially accounting for linkage to both NSCL/P and NSCPO59. The markers at 1p36.11 and 9q22.33 were not significant at the GWAS stage in our study and thus were not replicated in NSCL/P and NSCPO.

Tests for heterogeneity showed that the SNPs at 1, 8 and 5 loci yielded significant evidence of heterogeneity ($P < 1 \times 10^{-3}$ using logistic regression test and Bonferroni correction; 0.05 out of 40) between NSCL/P and NSCLO, NSCPO and NSCLO, and NSCPO and NSCLO, respectively (Table 2 and Table 3).
Interestingly, gender stratification analysis revealed that one previously identified locus (1q32.2) showed strong evidence of heterogeneity ($P = 1.38 \times 10^{-4}$) in the evidence of association with NSCLP from male and female cases. The marker on 8q21.3 was observed to exhibit significant evidence of heterogeneity in its estimated effect between older mothers (>35 years) and the reference age of mothers (25–35 years) (Supplementary Table 8).

Replications of 16 NSCLP loci in multi-ethnic groups. We further checked for associations of the 14 novel loci and two recently reported NSCLP loci (16p13.3 in China\textsuperscript{15} and 2p24.2 in a multi-ethnic study\textsuperscript{16}) using cohorts 5–7. Different loci showed evidence of association in different cohorts ($P < 0.05$ using logistic regression test), specifically three loci in Central Europeans, seven loci in Asians and five loci in European ancestry (Table 1 and Supplementary Table 9). For the majority of the 16 loci mentioned, the direction and magnitude of the effect of ORs were consistent across Chinese and non-Chinese samples. However, we observed an apparent difference in risk allele frequencies (AFs) for most of these 16 risk loci. For example, AFs in the cases of the markers at 4p16.2 (rs1907989, $AF_{\text{Chinese}} = 0.46$, $AF_{\text{European}} = 0.57$) and 17q21.32 (rs1838105, $AF_{\text{Chinese}} = 0.45$, $AF_{\text{European}} = 0.39$) showed a certain degree of difference between the Chinese and Central European populations (Table 1), whereas the AFs of the markers at 8q24.21 (rs987525, $AF_{\text{Chinese}} = 0.07$, $AF_{\text{European}} = 0.38$ and rs7017252, $AF_{\text{Chinese}} = 0.08$, $AF_{\text{European}} = 0.55$) were highly different between the Chinese and European populations (Supplementary Table 10).

Biological implications analyses for the 26 NSCLP loci. Of the 135 SNPs associated with the risk of NSCLP at these 26 loci ($r^2 \geq 0.7$ with the index SNPs), 33, 99 and 113, respectively, were found in known or predicted regulatory elements such as promoters, enhancers or motifs biochemically characterized to
regulate transcription for 50 reference genes (Supplementary Data 1). Expression profiles of the 49 genes within 500 kb and showing strong linkage disequilibrium (LD) with index SNPs ($R^2 ≥ 0.7$) in the 26 loci based on craniofacial gene expression patterns in mouse (EMAGE database) showed that a total of these 20 genes were related to embryonic development, which were also expressed in the related tissue for OFC (Supplementary Table 11). Twelve genes were found to produce CL/CP malformations in branchio-ocular-facial syndrome (Table 1). Seven genes were related to be associated with non-recognition malformation syndromes including OFC as a clinical phenotype (Supplementary Table 13). By manually reviewing the literature related to these 49 genes and their functions, annotated systematically using several databases, 28 of these genes were classified into different biological categories according to morphogenesis, development, molecular and cellular function, and among these categories contained 3 or more genes. The most relevant category was the morphogenesis- and development-related traits, which included the bone, limb, brain, eye and other organs. Several important signalling pathways were also reviewed, among which epidermis/epithelium development and morphogenesis, and the fibroblast growth factor receptor (FGFR) signalling pathway involved the most genes, whereas the bone morphogenic proteins (BMP), WNT and Notch signalling pathways have also been reported to be associated with lip and palate development (Supplementary Table 14).

### Discussion

In this study, we identified 14 novel risk loci associated with NSCLP and confirmed 12 previously reported loci in the Chinese population. Several lines of evidence support the hypothesis that these newly identified loci contribute to orofacial clefting. At 4p16.2, two markers in moderate LD with one another ($r^2 = 0.65$) are located 50 kb upstream of MSX1 (Supplementary Fig. 5b). Mutations in MSX1 contribute to CL/P and tooth agenesis in humans. MSX1 homozygotic mice exhibit CP, deficiency of the alveolar mandible and maxilla, and failure of tooth development.

The orthologous Spry1 of the optimized gene SPRY1 at 4q28.1 has been shown to exhibit facial clefting and CP in transgenic mice. The marker at 5p12 is adjacent to FGF10, whose encoded protein has been reported to play roles in the epithelial–mesenchymal transition, and Fgf10 knockout mice show the symptoms of CP. The 6p24.3 region harbours two candidate genes for the OFC: OFCC1, which contribute to CL/P and tooth agenesis, and MSX1, which contribute to orofacial clefting. At 4q14.2–4q14.3, the marker at 4q22.3 is adjacent to FGF10, whose encoded protein has been shown to play roles in the epithelial–mesenchymal transition, and Fgf10 knockout mice show the symptoms of CP. The 6p24.3 region harbours two candidate genes for the OFC: OFCC1, which contribute to CL/P and tooth agenesis, and MSX1, which contribute to orofacial clefting.

The SNP at 8p11.23 is located in an enhancer within FGFR1. In humans, mutations of FGFR1 have been found in syndromic CL/P, whereas in mice a homozygous hypomorphic allele at Fgfr1 caused CP. A marker at 9q22.32 is located at the intronic region of PTK1, mutations of which have been identified in patients of Gorlin syndrome with CL/P as one of its clinical manifestations. Ptk1D1 mouse models showed craniofacial defects, including underdeveloped palatal shelves and clefting of the secondary palate.

The marker rs705704 is located at the locus 12q13.2, which includes the gene RPS26, which may be relevant to orofacial clefting, as mutations of RPS26 have been reported in Diamond–Blackfan anaemia (DBA) with CPO as one
of its clinical phenotypes. Two highly correlated markers \( r^2 = 0.99 \) at 14q32.13 are located between GSC and DICER1. GSC modulates the epithelial–mesenchymal transition and mutations in GSC lead to a syndrome defined by short stature, auditory canal atresia, mandibular hypoplasia and skeletal abnormalities, whereas DICER1 mutations have been reported to cause pleuropulmonary blastoma and multinodular goiter-1, with or without Sertoli–Leydig cell tumours (MNG1), and Dicer1 conditional knockout mice exhibit secondary palate clefting and other severe craniofacial dysmorphisms. The excellent candidate gene WNT9B at 17q21.32 has already been functionally implicated in craniofacial development, as mice with Wnt9b targeted mutation were described as presenting CL/P phenotypes.

Among the other new signals, two markers at 2p25.1 are in perfect LD \( r^2 = 1 \) with one another and are located 12 kb upstream of TAF1B, encoded protein of which is important for polymerase (Pol) I transcription. At 8q22.1, a synonymous codon SNP rs957448 (KIAA1429) is correlated \( r^2 = 0.65 \) with rs12681366 (an intronic SNP of RAD54B). Human RAD54B was first identified as a homologue of RAD54, which plays an important role in DNA repair. The strongest associated marker at 12q13.13 is located 500 bp downstream of KRT18, which encodes a protein in the large family of cytoskeletal proteins with specific expression in epithelial cells. At 12q21.1, the signals are near the TMEM19 gene, involving the SNPs rs2304269 and rs7967428, which are in strong LD with each other \( r^2 = 0.98 \). Rs2304269 and rs7967428 are respectively located at one active promoter and five strong enhancers in epidermal keratinocytes according to ENCODE data. The 14q22.1 signal is close to LINC00640, a gene of unknown function. In addition, to gain further insight into the possible involvement of genes at some
identified loci in the development of NSCLP, immuno-histochromy (IHC) analysis performed in mice at different embryonic stages found positive IHC staining of three genes of interest (Rad54b, Rps26 and Fam49a) in the palatal mesenchymal cells and epithelium cells (Supplementary Fig. 6).

Notably, in our study, two members of the FGF signalling pathway, including FGFR1 and FAM49A, were connected (Fig. 3). The FGF signalling pathway was proposed to contribute to craniofacial development. Intriguingly, we also found a potential link between the FGF signalling pathway and NSCLP, which showed that several FGF signalling association study strengthen the hypothesis that the FGF signalling pathway might play important roles in craniofacial development. 

GWAS stage (Supplementary Table 4).

Table 2 | SNPs showing significance in stratified analysis among the three anatomical types of orofacial clefts: NSCLP, NSCLO and NSCPO in Chinese population.

| Phenotype          | Loci       | SNP       | BP (hg19)   | Allele | F_A | F_U | P*  | OR |
|--------------------|------------|-----------|-------------|--------|-----|-----|-----|----|
| NSCLP versus NSCLO | 1q32.2     | rs861020  | 209977111   | A/G    | 0.23| 0.30| 2.05E-09 | 0.72 (0.65-0.80) |
|                    | 1q32.2     | rs642961  | 209989270   | A/G    | 0.23| 0.30| 1.16E-09 | 0.72 (0.64-0.80) |
|                    | 1q32.2     | rs861020  | 209977111   | A/G    | 0.23| 0.17| 8.69E-11 | 1.51 (1.34-1.72) |
|                    | 1q32.2     | rs642961  | 209989270   | A/G    | 0.23| 0.17| 9.08E-11 | 1.51 (1.33-1.72) |
|                    | 1q32.2     | rs2064163 | 210048819   | A/C    | 0.38| 0.44| 6.41E-08 | 0.76 (0.69-0.84) |
|                    | 1q32.2     | rs9430019 | 210050794   | A/G    | 0.31| 0.19| 1.29E-05 | 1.87 (1.66-2.11) |
|                    | 2p25.1     | rs287980  | 9973166     | G/A    | 0.23| 0.27| 7.67E-04 | 0.83 (0.74-0.92) |
|                    | 2p25.1     | rs287980  | 9972442     | G/A    | 0.23| 0.27| 8.29E-04 | 0.83 (0.74-0.93) |
|                    | 2p24.2     | rs10172734| 16733054    | G/A    | 0.26| 0.33| 6.31E-09 | 0.73 (0.66-0.81) |
|                    | 2p24.2     | rs7552    | 16733928    | A/G    | 0.24| 0.28| 6.33E-04 | 0.83 (0.74-0.92) |
|                    | 8q21.3     | rs1034832 | 88918331    | C/A    | 0.31| 0.35| 4.11E-04 | 0.83 (0.75-0.92) |
|                    | 12q21.1    | rs2304269 | 72080272    | G/A    | 0.39| 0.45| 8.58E-07 | 0.78 (0.71-0.86) |
|                    | 12q21.1    | rs7967428 | 72089040    | G/A    | 0.40| 0.45| 1.98E-06 | 0.79 (0.72-0.87) |
|                    | 16p13.3    | rs2283487 | 3969886     | G/A    | 0.42| 0.47| 3.21E-04 | 0.84 (0.76-0.92) |
|                    | 16p13.3    | rs17136624| 3996262     | G/A    | 0.26| 0.22| 8.59E-04 | 1.12 (1.08-1.16) |
|                    | 17p13.1    | rs2872615 | 814693      | G/A    | 0.43| 0.49| 2.59E-07 | 0.78 (0.70-0.85) |
|                    | 17p13.1    | rs1880646 | 8929845     | G/A    | 0.48| 0.53| 2.45E-04 | 0.83 (0.76-0.92) |
|                    | 17q21.32   | rs1838105 | 45008935    | A/G    | 0.44| 0.38| 2.28E-05 | 1.24 (1.12-1.37) |
| NSCLO versus NSCPO | 1q32.2     | rs861020  | 209977111   | A/G    | 0.30| 0.17| 3.98E-04 | 2.10 (1.81-2.42) |
|                    | 1q32.2     | rs642961  | 209989270   | A/G    | 0.30| 0.17| 2.09E-04 | 2.11 (1.82-2.43) |
|                    | 1q32.2     | rs2064163 | 210048819   | A/C    | 0.39| 0.44| 1.21E-04 | 0.79 (0.70-0.89) |
|                    | 1q32.2     | rs9430019 | 210050794   | A/G    | 0.30| 0.19| 8.70E-18 | 1.84 (1.60-2.12) |
|                    | 2p24.2     | rs10172734| 16733054    | G/A    | 0.25| 0.33| 6.20E-09 | 0.68 (0.59-0.77) |
|                    | 8q21.3     | rs1034832 | 88918331    | C/A    | 0.28| 0.35| 1.13E-06 | 0.73 (0.64-0.83) |
|                    | 17p13.1    | rs2872615 | 8914693     | G/A    | 0.42| 0.49| 1.34E-05 | 0.77 (0.68-0.86) |

*F_A, minor allele frequency in cases; F_U, minor allele frequency in controls; NSCLO, non-syndromic cleft lip only; NSCLP, non-syndromic cleft lip with palate; NSCPO, non-syndromic cleft palate only; OR is calculated based on minor allele; alleles are shown as minor allele/major allele. *P-value below 1.25 x 10^-3 (0.05 out of 40, the P-value using logistic regression test and Bonferroni correction) was considered to be statistically significant.
our study, 1q32.2, 8q24.21 and 9q22.32 were first demonstrated to have an effect on NSCPO in the Chinese population. Importantly, our study provided evidence that 1q32.2 exhibits an overlapping effect on all three sub-phenotypes of NSCLP, NSCLO and NSCPO. The evidence of association at 1q32.2 was stronger among males than females, which may reflect the higher prevalence of NSCLO and NSCPO. The evidence of association at 1q32.2 was an overlapping effect on all three sub-phenotypes of NSCLP, as suggested by some previous studies of NSCLP cases and 8,593 controls of Chinese Han and prioritized genes in each significant SNP.

Table 3 | Markers achieving genome-wide significance in GWAS of 3,379 NSCLP cases and 8,593 controls of Chinese Han and prioritized genes in each significant SNP.

| Loci | SNP | BP | Allele | \(P_{\text{meta}}\) | OR | \(P_{\text{het}}\) | Notable gene(s) |
|------|-----|----|--------|-----------------|----|----------------|----------------|
| Novel loci: | | | | | | | |
| 2p25.1 | rs287980 | 9971366 | G/A | 1.94E–08 | 0.83 | 0.8120 | TAF1B |
| 2p25.1 | rs287982 | 9972442 | G/A | 6.15E–09 | 0.82 | 0.8981 | TAF1B |
| 4p16.2 | rs34246903 | 4794195 | C/A | 4.45E–08 | 0.85 | 0.2344 | MSX1 |
| 4p16.2 | rs1907989 | 4818925 | A/G | 1.58E–08 | 0.85 | 0.1130 | MSX1 |
| 4q28.1 | rs908822 | 124906257 | A/G | 4.33E–08 | 1.31 | 0.0545 | SPRY1 |
| 5p12 | rs10462065 | 44068846 | A/C | 1.12E–08 | 1.22 | 0.4835 | FGF10 |
| 6p24.3 | rs9381107 | 9469238 | A/G | 2.72E–09 | 0.83 | 0.9090 | OFCC1/TFAP2A |
| 8p11.23 | rs13317 | 38269514 | G/A | 3.96E–08 | 0.85 | 0.4691 | FGF1 |
| 8q22.1 | rs12681366 | 95401265 | C/A | 5.10E–10 | 0.82 | 0.2026 | PTCH1 |
| 12q13.13 | rs3741442 | 53346750 | A/G | 3.72E–12 | 1.22 | 0.9598 | KRT1B |
| 12q13.2 | rs705704 | 56435412 | A/G | 1.29E–09 | 1.22 | 0.9839 | RPS26 |
| 12q21.1 | rs2304269 | 72080272 | G/A | 1.32E–12 | 0.81 | 0.3903 | TMEM19 |
| 12q21.1 | rs7967428 | 72089040 | A/G | 3.08E–12 | 0.81 | 0.3871 | TMEM19 |
| 14q22.1 | rs7148069 | 5139645 | A/G | 1.69E–08 | 1.22 | 0.2538 | LINCO0640 |
| 14q22.13 | rs1243572 | 93739499 | A/G | 3.52E–10 | 1.20 | 0.1138 | GSC/DICER1 |
| 14q22.13 | rs1243573 | 93739583 | C/A | 8.61E–10 | 1.20 | 0.1178 | GSC/DICER1 |
| 17q21.32 | rs4968247 | 44988703 | A/G | 8.70E–10 | 0.83 | 0.8605 | WNT9B |
| 17q21.32 | rs1838105 | 45008935 | A/G | 1.31E–11 | 1.22 | 0.3543 | WNT9B |

Reported loci:

| Loci | SNP | BP | Allele | \(P_{\text{meta}}\) | OR | \(P_{\text{het}}\) | Notable gene(s) |
|------|-----|----|--------|-----------------|----|----------------|----------------|
| 1p22.1 | rs481931 | 94570016 | A/C | 1.06E–12 | 0.80 | 0.3687 | ARHGAP29 |
| 1p22.1 | rs4174703 | 94582293 | G/C | 7.97E–12 | 0.81 | 0.8369 | ARHGAP29 |
| 1q22.2 | rs861020 | 20977111 | A/G | 1.30E–14 | 1.31 | 0.5428 | IRF6 |
| 1q22.2 | rs642961 | 20959270 | A/G | 2.76E–15 | 1.32 | 0.6061 | IRF6 |
| 1q22.2 | rs2064183 | 21004819 | A/G | 8.60E–19 | 0.77 | 0.9625 | IRF6 |
| 1q22.2 | rs943009 | 21005794 | A/G | 1.68E–12 | 1.25 | 0.6420 | IRF6 |
| 2p24.2 | rs10172734 | 16733054 | A/G | 2.89E–20 | 0.74 | 0.4992 | FAM49A |
| 2p24.2 | rs7552 | 16733928 | A/G | 5.83E–22 | 0.73 | 0.5814 | FAM49A |
| 8q21.3 | rs12543318 | 88686340 | A/C | 8.80E–12 | 0.81 | 0.2050 | MMP16 |
| 8q21.3 | rs1034832 | 8891331 | C/A | 1.35E–10 | 0.82 | 0.2243 | MMP16 |
| 8q24.21 | rs7854651 | 12988794 | A/G | 1.03E–10 | 1.27 | 0.8110 | MYC |
| 8q24.21 | rs7017252 | 12995084 | A/G | 8.47E–16 | 1.60 | 0.8960 | MYC |
| 9q22.2 | rs7871995 | 92209587 | A/G | 6.06E–09 | 1.21 | 0.5782 | GADD45G |
| 10q25.3 | rs6585429 | 11889232 | A/G | 7.14E–13 | 0.81 | 0.9967 | VAX1 |
| 13q31.1 | rs954308 | 80639405 | A/C | 2.00E–09 | 1.29 | 0.8103 | SPRY2 |
| 16p13.3 | rs2283487 | 3969886 | G/A | 1.27E–10 | 0.83 | 0.9121 | CREBBP |
| 16p13.3 | rs17136624 | 3996282 | A/G | 3.82E–10 | 1.24 | 0.5269 | CREBBP |
| 17p13.1 | rs2878265 | 8914693 | A/G | 8.81E–12 | 0.82 | 0.5224 | NNT1 |
| 17p13.1 | rs1800646 | 8928945 | A/G | 1.69E–11 | 1.22 | 0.4104 | NNT1 |
| 17q22 | rs227731 | 54773238 | C/A | 8.83E–09 | 1.19 | 0.5623 | NOG |
| 20q12 | rs12962563 | 39275603 | A/G | 8.57E–12 | 1.23 | 0.5970 | MAF8 |

GWAS, genome-wide association study; NSCLP, non-syndromic cleft lip with palate; OR, odds ratio; SNP, single-nucleotide polymorphism.

Genome-wide significance is defined as \(P < 5 \times 10^{-8}\); SNP positions are reported according to Build 37 and their alleles are coded based on the positive strand; alleles (minor/major); meta-analysis is of NSCLP GWAS and NSCLP replication using fixed model; the \(P\)-value using Cochran-Mantel-Haenszel test; OR is calculated based on minor allele.

* \(P_{\text{het}}\): P-value for heterozygosity test using logistic regression test and Bonferroni correction and \(P_{\text{het}} > 0.05\) was considered to signify no heterogeneity.
diabetes autoantibodies and alopecia at 12q13.2, and height at 9q22.32. The SNPs reported to be associated with NSCLP and other diseases/traits were in strong LD ($r^2 \geq 0.7$), which showed significant and non-independent association of the risk of NSCLP and other diseases/traits. Interestingly, NSCLP and adult height shared the same index SNP (rs10512248 at PTCH1 in 9q22.32), suggesting that some shared genetic factors might underlie these two very distinct phenotypes. Furthermore, some reported GWAS loci had susceptibility genes shared between NSCLP and other diseases/traits, such as MAFB at 20q12 for Dupuytren’s disease, low-density lipoprotein cholesterol and total cholesterol, which showed a clearly independent association with NSCLP index SNPs and suggested pleiotropic effects of these genes on other diseases/traits.

Overall, our current study has advanced the understanding of the genetic architecture controlling the risk of NSOFC by substantially increasing the number of genetic risk factors and has highlighted potential candidate genes through subsequent genetic and biological analyses. This study has also provided further insight into the possible pleiotropic effects of genetic risk factors on different sub-phenotypes, in different populations and among different diseases/traits. Through a comprehensive analysis of cases and controls from a Chinese population, we have identified 14 new genetic risk factors and validated associations in a large majority of previously reported loci. Further sequencing and functional investigations will probably identify causal mutational events and true susceptibility genes in or near these tagging SNPs and further elucidate the disease pathogenesis of these common congenital birth defects.

Methods

Samples. In the current study, we carried out a two-stage GWAS and further replications of NSOFC. The discovery stage included 2,096 NSCLP cases and 4,051 controls (cohort 1). Replication studies were performed in an additional 1,346 cases.
unrelated NSCLP cases and 4,542 controls (cohort 2). Further replications consisted of cohort 3 (1,104 NSCLO cases versus 3,312 controls), cohort 4 (1,104 NSCP/O cases versus 3,312 controls shared with cohort 3), cohort 5 (3,99 NSCL/P cases versus 3,138 controls), cohort 6 (861 NSCL/P case–parent trios) and cohort 7 (557 NSCL/P case–parent trios). Samples of cohorts 1–4 were recruited from the Chinese population through collaboration with multiple hospitals in Hubei, Henan and Anhui province. All cases were interviewed and clinically assessed by at least two experienced clinicians, and a detailed questionnaire was completed to identify any further anomalies, such as congenital heart disease, hypospadia, accessory auricle, lip pits and so on, which would suggest an underlying syndrome. We collected clinical information from the subjects through a full clinical checkup and additional demographic information from the cases was obtained through a structured questionnaire that mainly included four parts: basic information, clinical symptom, additional demographic information, and personal history. Peripheral blood samples were collected after the written informed consents were obtained from all the participants or their guardians. The study was approved by the institutional ethics committee of each hospital (Hospital of Stomatology Wuhan University, The Second Charity Hospital of Henan Province, Stomatological Hospital of Nanyang, Stomatological Hospital of Xiangyang and The First Affiliated Hospital of Anhui Medical University) and was conducted according to the Declaration of Helsinki principles. The replication data in cohort 5 from the GWAS in Central Europe was provided by Mangold et al.\textsuperscript{13}, whereas the replication data in cohorts 6 and 7 were from the GWAS of case–parent trios of Asians and European ancestry provided by Beatty et al.\textsuperscript{12}. All the controls and cases for each replication cohort were sampled from the same locality and the same population in each study, to assure minimal population stratification effects for each replication.

DNA extraction. Approximately 4 ml EDTA anticoagulated venous blood sample was collected from each participant. Genomic DNAs of the cases were extracted from peripheral blood lymphocytes using the standard SDS–proteinase K-phenol/chloroform method. For the controls, DNAs were isolated by standard procedures using Flexi Gene DNA kits (Qiagen) according to the manufacturer’s protocol. For quality control, DNAs were diluted to working concentrations of 45–55 ng\textsuperscript{1} for genome-wide genotyping and 20–30 ng\textsuperscript{1} for the validation studies, respectively.

Genotyping and quality controls in GWAS. The discovery-stage genotyping was conducted according to the Infinium HD protocol using the Illumina HumanOmniZhongHua-8 v1.1 BeadChip (Illumina, San Diego, CA, USA) at the Key Laboratory of Dermatology at Anhui Medical University (Ministry of Education), Hefei, Anhui, China. Genotyping was performed as described in the Infinium HD protocol from Illumina\textsuperscript{2,3}. In the GWAS stage, a total of 900,015 SNPs were genotyped in 2,096 cases and 4,051 controls. A standard quality-control criterion was applied to select SNPs and samples for further analysis. SNPs were excluded if they had (i) a call rate <95% in cases or controls; (ii) MAF of <0.01 in cases or controls; (iii) significant linkage disequilibrium (LD) or Hardy–Weinberg equilibrium (HWE) in the controls (P\textsubscript{e} < 10\textsuperscript{−4}). In addition, all the SNPs on the Y, X and mitochondrial chromosomes, as well as the copy number variation-related SNPs and probes, were excluded from statistical analysis. Meanwhile, samples were removed if they (i) had an overall genotyping rate of <95% across all SNPs; (ii) had MAF of <0.01 in cases or controls; (iii) had a call rate <95% from Hardy–Weinberg equilibrium (HWE) in the controls (P\textsubscript{e} < 10\textsuperscript{−4}); (iv) showed inconsistent genetic gender with epidemiological or clinical data; and (v) were ancestry outliers or heterozygosity outliers. Samples were assessed for population stratification using the software package EIGENSTRAT\textsuperscript{4,5}. The original script from EIGENSTRAT was modified to extract the principal components for plotting. In total, 63 samples were removed from analysis. After quality control, the genotype data of 803,202 autosomal SNPs in 2,033 cases and 4,051 controls remained for further analysis.

SNP selection for replication studies. SNPs were selected for NSCLP replication according to the following steps: (i) we first picked out all the top SNPs with P < 1.0 \times 10\textsuperscript{−4} in the initial stage and excluded the EPIS devised with ambiguous genotype scatter plots; (ii) then we selected at least one SNP with the lowest P-values in each of the novel loci, which defined by using the PLINK option ‘--indep-pairwise 50 5 0.2’; (iii) in addition, one to four top SNPs were chosen from the previously reported loci; (iv) we also selected SNPs that were located within or close to the susceptibility genes with gene expression profiling evidence for EOC or for syndromes with OFC symptoms. In total, 152 SNPs were selected for the NSCLP replication stage. Furthermore, all the promising SNPs were selected for the NSCLO and NSCPO replications. These SNPs had the lowest P-values in NSCLP-meta stage and showed P < 0.05, as well as with call rate >98% and MAF >0.01 and P\textsubscript{het} > 10\textsuperscript{−4} in NSCLP replication stage; thus, 41 SNPs reached genome-wide significance. The above 41 SNPs were distributed in 12 previously reported NSCL/P associated loci and 14 novel loci. In cohort 5–7 replications, 24 SNPs were selected for replication in 3 GWAS data sets from Central Europeans, Asians and European ancestry groups. Of the 24 SNPs, 20 were picked out from the 14 novel loci and 4 were from 2 newly reported NSCL/P loci (16p13.3 (ref. 15) and 2p24.2 (ref. 16)) and all of them were from the 41 significant SNPs.

Genotyping and quality control in replication studies. Genotyping analyses of replications in cohorts 2–4 were conducted by using the Sequenom MassARRAY system, at the Key Laboratory of Dermatology at Anhui Medical University (Ministry of Education), Hefei, Anhui, China. Locus-specific PCR primers (Supplementary Table 15) were designed using MassARRAY ASSAY Design 3.0 software following the manufacturer’s instruction. Sequenom MassARRAY Massarray\textsuperscript{8} was performed in each data set separately using PLINK 1.07. In each case–control replications (cohorts 2–4), we excluded SNPs with a call rate <90% in cases or controls, or deviation from HWE proportions (P\textsubscript{e} < 1.0 \times 10\textsuperscript{−4}) in the controls.

To evaluate the quality of the genotype data for the validation analyses, 100 samples were randomly selected from the GWAS stage and were replications using the Sequenom system. The concordance rate between the genotypes from the Illumina HumanOmniZhongHua-8 v1.1 BeadChip and the Sequenom MassARRAY assay analyses was >99%. The cluster plots from the Illumina and Sequenom analyses were checked to confirm their good quality. After quality control, 146 SNPs were left for further replications in cohort 3 and 4 analyses, respectively.

Statistical analyses. In the GWAS stage, we examined potential genetic relatedness based on pairwise identity by state for all of the successfully genotyped samples using the PLINK 1.07 software. For the duplicated samples and all pairs of first-, second- and third-degree relatives, the detected pair from the subject with each other lower call rate was removed from further analysis. All cases and controls were assumed to be principal component analyses for population stratification. SNPs were confirmed to be of Chinese ancestry. Quantile–quantile plots were constructed and calculations of genomic control values (\lambda\textsubscript{GC} = 1.04 indicated a negligible inflation of the genome-wide statistical results due to population stratification) were performed by using the software R (http://www.r-project.org/) to evaluate the overall significance of genome-wide association results and the potential impact of population stratification, respectively, in the GWAS stage.

Association of GWAS and replication analysis were performed using the Cochran–Armitage trend test. Single-marker association analyses were performed to test for disease–SNP associations using logistic regression in each stage. Fixed effects analyses of cohort 2 (NSCLP combined) and 3 were performed using the Cochran–Mantel–Haenszel test, where P-values and heterogeneity index Q-values from Cochran’s Q statistics were also obtained. Assessment of heterogeneity across studies was carried out by evaluating the P\textsubscript{het} values from Cochran’s Q statistics (Bonferroni-corrected heterogeneity Q-values P\textsubscript{het} < 0.05 were considered significant)\textsuperscript{32,36}. OR values were measured as OR per allele and presented for the minor allele of a SNP, unless otherwise stated. A threshold of P < 5 \times 10\textsuperscript{−8} was adopted to define novel loci with genome-wide significance. The regional association plots for each susceptibility locus were generated in R using information from the HapMap project (CHB and JPT samples). After applying quality control and removing SNPs with population stratification, the 26 NSCLP risk loci. The proportion of variance in NSCLP risk was examined via the residual maximum likelihood method in the program genome-wide complex trait analysis and estimated assuming a disease prevalence of 0.001 (1 out of 1,000) and log additive risk\textsuperscript{35,37}. All power calculations were performed using the genetic power calculator assuming a disease prevalence of 0.001 and log-additive risk. We carried out conditional analyses to identify additional association signals after accounting for the effects of known and newly discovered susceptibility loci.

To investigate more than two association signals per locus, we used a stepwise procedure in which additional SNPs were added to the model according to their conditional values, as programatically described in the program EMMAX. When SNPs with MAF < 0.05 were included in the models, the significance of the 26 NSCLP risk loci. The proportion of variance in NSCLP risk was examined via the residual maximum likelihood method in the program genome-wide complex trait analysis and estimated assuming a disease prevalence of 0.001 (1 out of 1,000) and log additive risk\textsuperscript{35,37}. All power calculations were performed using the genetic power calculator assuming a disease prevalence of 0.001 and log-additive risk. We carried out conditional analyses to identify additional association signals after accounting for the effects of known and newly discovered susceptibility loci.

To investigate more than two association signals per locus, we used a stepwise procedure in which additional SNPs were added to the model according to their conditional values, as programatically described in the program EMMAX. When SNPs with MAF < 0.05 were included in the models, the significance of the 26 NSCLP risk loci. The proportion of variance in NSCLP risk was examined via the residual maximum likelihood method in the program genome-wide complex trait analysis and estimated assuming a disease prevalence of 0.001 (1 out of 1,000) and log additive risk\textsuperscript{35,37}. All power calculations were performed using the genetic power calculator assuming a disease prevalence of 0.001 and log-additive risk. We carried out conditional analyses to identify additional association signals after accounting for the effects of known and newly discovered susceptibility loci.
replication stages. Then, we performed stratification analyses on gender and maternal gestational age in NSCLP. P-value below 1.22 × 10−3 was used to achieve significance (0.05 out of 41, Bonferroni correction) was considered to be statistically significant.

Heterogeneity analyses among NSCLP, NSCLO and NSCPO were performed by using PLINK 1.07 software based on the 40 significant associated markers in NSCLP meta-stage. Genotype data were extracted from discovery and replication stages of NSCLP, NSCLO and NSCPO (cohorts 1–4). We first divided the cases into three sub-phenotypes NSCLP, NSCLO and NSCPO, then extracted genotype of each case from the above four cohorts and calculated the association between each combination of two sub-phenotypes. P-value below 1.25 × 10−3 using logistic regression (0.05 out of 40, Bonferroni correction) was considered to be statistically significant.

**Locus annotation and candidate gene prioritization.** To prioritize candidate genes, besides the nearest genes to the index SNPs, the following methods were used to help prioritize potential causal genes in each associated region. All genes located in the same LD block as the index SNPs (r2 ≥ 0.7) were selected and annotated for function in molecular, cellular, animal model and tissue/organ levels using several databases, including PubMed (http://www.ncbi.nlm.nih.gov/pubmed/), EMAGE (http://www.emouseatlas.org/emage/home.php), MGI (http://www.informatics.jax.org/), OMIM (http://www.omim.org/), Gene (http://www.ncbi.nlm.nih.gov/gene/), UCSC (http://genome.ucsc.edu/) and Ensembl (http://www.ensembl.org/index.html). The nearest genes on both sides of the index SNP were annotated when no gene was located within the LD block. A total of 135 SNPs at these 26 NSCLP risk loci (all with r2 ≥ 0.7 with the SNPs found to be genome-wide significant here) were MAF > 0.05 and P-value > 1.0 × 10−5 were annotated by using the following methods: regulatory features from ENCODE Consortium/ENCODE/Roadmap Epigenomics Project (http://www.roadmapepigenomics.org/).

**GWAS catalogue reviews.** We evaluated all the SNPs within ± 500 kb of the index SNPs (from the 26 loci) and with F < 5 × 10−5 recorded in National Human Genome Research Institute GWAS catalogue (http://www.genome.gov/gwastudies) updated on 20 February 2015. The LD patterns of the index SNPs and the recorded SNPs in GWAS catalogue were queried using SNAP version 2.2 in Asian (CHB + JPT) and European (CEU) populations using data from the 1,000 Genomes Project Pilot 1.

**Expression studies in the mouse.** Eight to 14-week-old wild-type Kunming mice were housed in approved specific pathogen-free conditions and mated for 12 h, the resulting embryos (E0.5) were harvested and fixed in 4% paraformaldehyde overnight at 4°C. Embryos with death or other malformations were ruled out. Normal fetuses were weighed and enumerated. The sections were then counterstained with haematoxylin. The results were analyzed using several databases, including PubMed (http://www.ncbi.nlm.nih.gov/pubmed/), EMAGE (http://www.emouseatlas.org/emage/home.php), MGI (http://www.informatics.jax.org/), OMIM (http://www.omim.org/), Gene (http://www.ncbi.nlm.nih.gov/gene/), UCSC (http://genome.ucsc.edu/) and Ensembl (http://www.ensembl.org/index.html). The nearest genes on both sides of the index SNP were annotated when no gene was located within the LD block. A total of 135 SNPs at these 26 NSCLP risk loci (all with r2 ≥ 0.7 with the SNPs found to be genome-wide significant here) were MAF > 0.05 and P-value > 1.0 × 10−5 were annotated by using the following methods: regulatory features from ENCODE Consortium/ENCODE/Roadmap Epigenomics Project (http://www.roadmapepigenomics.org/).

**Network analysis.** We expanded the global network by including the Human Net protein interaction database61 and literature-curated interactions from STRING62,63 to derive an expanded global network based on known protein–protein interactions using the notables of the 26 NSCLP associated loci from the present study (Fig. 3).

**GWAS catalogue reviews.** We evaluated all the SNPs within ± 500 kb of the index SNPs (from the 26 loci) and with F < 5 × 10−5 recorded in National Human Genome Research Institute GWAS catalogue (http://www.genome.gov/gwastudies) updated on 20 February 2015. The LD patterns of the index SNPs and the recorded SNPs in GWAS catalogue were queried using SNAP version 2.2 in Asian (CHB + JPT) and European (CEU) populations using data from the 1,000 Genomes Project Pilot 1.

**Expression studies in the mouse.** Eight to 14-week-old wild-type Kunming mice were housed in approved specific pathogen-free conditions and mated for 12 h, the presence of a vaginal plug was designated as E0.5. Pregnant mice were randomly housed in approved specific pathogen-free conditions and mated for 12 h, the resulting embryos (E0.5) were harvested and fixed in 4% paraformaldehyde overnight at 4°C. Embryos with death or other malformations were ruled out. Normal fetuses were weighed and enumerated. The sections were then counterstained with haematoxylin. The results were analyzed using several databases, including PubMed (http://www.ncbi.nlm.nih.gov/pubmed/), EMAGE (http://www.emouseatlas.org/emage/home.php), MGI (http://www.informatics.jax.org/), OMIM (http://www.omim.org/), Gene (http://www.ncbi.nlm.nih.gov/gene/), UCSC (http://genome.ucsc.edu/) and Ensembl (http://www.ensembl.org/index.html). The nearest genes on both sides of the index SNP were annotated when no gene was located within the LD block. A total of 135 SNPs at these 26 NSCLP risk loci (all with r2 ≥ 0.7 with the SNPs found to be genome-wide significant here) were MAF > 0.05 and P-value > 1.0 × 10−5 were annotated by using the following methods: regulatory features from ENCODE Consortium/ENCODE/Roadmap Epigenomics Project (http://www.roadmapepigenomics.org/).

**Network analysis.** We expanded the global network by including the Human Net protein interaction database61 and literature-curated interactions from STRING62,63 to derive an expanded global network based on known protein–protein interactions using the notables of the 26 NSCLP associated loci from the present study (Fig. 3).

**GWAS catalogue reviews.** We evaluated all the SNPs within ± 500 kb of the index SNPs (from the 26 loci) and with F < 5 × 10−5 recorded in National Human Genome Research Institute GWAS catalogue (http://www.genome.gov/gwastudies) updated on 20 February 2015. The LD patterns of the index SNPs and the recorded SNPs in GWAS catalogue were queried using SNAP version 2.2 in Asian (CHB + JPT) and European (CEU) populations using data from the 1,000 Genomes Project Pilot 1.

**Expression studies in the mouse.** Eight to 14-week-old wild-type Kunming mice were housed in approved specific pathogen-free conditions and mated for 12 h, the presence of a vaginal plug was designated as E0.5. Pregnant mice were randomly housed in approved specific pathogen-free conditions and mated for 12 h, the resulting embryos (E0.5) were harvested and fixed in 4% paraformaldehyde overnight at 4°C. Embryos with death or other malformations were ruled out. Normal fetuses were weighed and enumerated. The sections were then counterstained with haematoxylin. The results were analyzed using several databases, including PubMed (http://www.ncbi.nlm.nih.gov/pubmed/), EMAGE (http://www.emouseatlas.org/emage/home.php), MGI (http://www.informatics.jax.org/), OMIM (http://www.omim.org/), Gene (http://www.ncbi.nlm.nih.gov/gene/), UCSC (http://genome.ucsc.edu/) and Ensembl (http://www.ensembl.org/index.html). The nearest genes on both sides of the index SNP were annotated when no gene was located within the LD block. A total of 135 SNPs at these 26 NSCLP risk loci (all with r2 ≥ 0.7 with the SNPs found to be genome-wide significant here) were MAF > 0.05 and P-value > 1.0 × 10−5 were annotated by using the following methods: regulatory features from ENCODE Consortium/ENCODE/Roadmap Epigenomics Project (http://www.roadmapepigenomics.org/).

**Network analysis.** We expanded the global network by including the Human Net protein interaction database61 and literature-curated interactions from STRING62,63 to derive an expanded global network based on known protein–protein interactions using the notables of the 26 NSCLP associated loci from the present study (Fig. 3).

**GWAS catalogue reviews.** We evaluated all the SNPs within ± 500 kb of the index SNPs (from the 26 loci) and with F < 5 × 10−5 recorded in National Human Genome Research Institute GWAS catalogue (http://www.genome.gov/gwastudies) updated on 20 February 2015. The LD patterns of the index SNPs and the recorded SNPs in GWAS catalogue were queried using SNAP version 2.2 in Asian (CHB + JPT) and European (CEU) populations using data from the 1,000 Genomes Project Pilot 1.

**Expression studies in the mouse.** Eight to 14-week-old wild-type Kunming mice were housed in approved specific pathogen-free conditions and mated for 12 h, the presence of a vaginal plug was designated as E0.5. Pregnant mice were randomly housed in approved specific pathogen-free conditions and mated for 12 h, the resulting embryos (E0.5) were harvested and fixed in 4% paraformaldehyde overnight at 4°C. Embryos with death or other malformations were ruled out. Normal fetuses were weighed and enumerated. The sections were then counterstained with haematoxylin. The results were analyzed using several databases, including PubMed (http://www.ncbi.nlm.nih.gov/pubmed/), EMAGE (http://www.emouseatlas.org/emage/home.php), MGI (http://www.informatics.jax.org/), OMIM (http://www.omim.org/), Gene (http://www.ncbi.nlm.nih.gov/gene/), UCSC (http://genome.ucsc.edu/) and Ensembl (http://www.ensembl.org/index.html). The nearest genes on both sides of the index SNP were annotated when no gene was located within the LD block. A total of 135 SNPs at these 26 NSCLP risk loci (all with r2 ≥ 0.7 with the SNPs found to be genome-wide significant here) were MAF > 0.05 and P-value > 1.0 × 10−5 were annotated by using the following methods: regulatory features from ENCODE Consortium/ENCODE/Roadmap Epigenomics Project (http://www.roadmapepigenomics.org/).

**Network analysis.** We expanded the global network by including the Human Net protein interaction database61 and literature-curated interactions from STRING62,63 to derive an expanded global network based on known protein–protein interactions using the notables of the 26 NSCLP associated loci from the present study (Fig. 3).

**GWAS catalogue reviews.** We evaluated all the SNPs within ± 500 kb of the index SNPs (from the 26 loci) and with F < 5 × 10−5 recorded in National Human Genome Research Institute GWAS catalogue (http://www.genome.gov/gwastudies) updated on 20 February 2015. The LD patterns of the index SNPs and the recorded SNPs in GWAS catalogue were queried using SNAP version 2.2 in Asian (CHB + JPT) and European (CEU) populations using data from the 1,000 Genomes Project Pilot 1.
33. Barritt, L. C. et al. Conditional deletion of the human ortholog gene Dicer1 in Pax2-Cre expression domain impairs orofacial development. Indian J. Hum. Genet. 18, 310–319 (2012).

34. Juriloff, D. M., Harris, M. J., McMahon, A. P., Carroll, T. J. & Lidral, A. C. Wnt9b is the mutated gene involved in multifactorial nonsyndromic cleft lip with or without cleft palate in A/WySn mice, as confirmed by a genetic complementation test. Birth Defects Res. A. Clin. Mol. Teratol. 76, 574–579 (2006).

35. Naidu, S., Friedrich, J. K., Russell, J. & Zomerdiijk, J. C. TAF1B is a TFIB-like component of the basal transcription machinery for RNA polymerase I. Science 333, 1640–1642 (2011).

36. Sarai, N. et al. Biochemical analysis of the N-terminal domain of human RADS4B. Nucleic Acids Res. 36, 5441–5450 (2008).

37. Fortier, A. M., Asselin, E. & Cadrin, M. Keratin 8 and 18 loss in epithelial claudin1 up-regulation. Nucleic Acids Res. 42, 2515–2525 (1999).

38. Kramer, S., Okabe, M., Hacohen, N., Krasnow, M. A. & Hiromi, Y. Sprouty: a common antagonist of FGF and EGF signaling pathways in Drosophila. Development 126, 2515–2525 (1999).

39. Metzis, V. et al. Patched1 is required in neural crest cells for the prevention of orofacial clefts. Hum. Mol. Genet. 22, 5026–5035 (2013).

40. Jin, Y. R., Han, X. H., Taketo, M. M. & Yoon, J. K. Wnt9b-dependent FGF signaling is crucial for outgrowth of the nasal and maxillary processes during upper jaw and lip development. Development 139, 1821–1830 (2012).

41. Riley, B. M. et al. Impaired FGF signaling contributes to cleft lip and palate. Proc. Natl Acad. Sci. USA 104, 4517–4521 (2007).

42. Pauws, E. & Stanier, P. FGF signalling and SUMO modification: new players in postoperative distant recurrence in colorectal cancer patients. Proc. Natl Acad. Sci. USA 104, 5441–5450 (2007).

43. Narla, A. & Ebert, B. L. Ribosomopathies: human disorders of ribosome dysfunction. Blood 115, 3196–3205 (2010).

44. Cui, D. et al. The ribosomal protein S26 regulates p35 activity in response to DNA damage. Oncogene 33, 2225–2235 (2014).

45. Nagai, Y. et al. High RADS4B expression: an independent predictor of postoperative distant recurrence in colorectal cancer patients. Oncotarget 6, 21064–21073 (2015).

46. Murray, T. et al. Examining markers in 8q24 to explain differences in evidence for association with cleft lip with/without cleft palate between Asians and Europeans. Genet. Epidemiol. 36, 392–399 (2012).

47. Udou, V. V. et al. Long-range enhancers regulating Myc expression are required for normal facial morphogenesis. Nat. Genet. 46, 753–758 (2014).

48. Watkins, S. E., Meyer, R. E., Strauss, R. P. & Aylsworth, A. S. Classification, epidemiology, and genetics of orofacial clefts. Birth Defects Res. A Clin. Mol. Teratol. 61, 197–206 (2015).

49. Meng, L., Bian, Z., Torensma, R. & Von den Hoff, J. W. Biological mechanisms of cleft lip with palate identify 14 novel loci and genetic heterogeneity. Nat. Rev. Genet. 15, 56–68 (2011).

50. Bernstein, B. E. et al. The NIH Roadmap Epigenomics Mapping Consortium. Nat. Biotechnol. 28, 1045–1048 (2010).

51. Caro, C. et al. The Gene Environment Association Studies consortium (GENEVA): maximizing the knowledge obtained from GWAS by collaboration across studies of multiple conditions. Genet. Epidemiol. 34, 364–372 (2010).

52. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).