


Copy Number Changes Identified Using Whole Exome Sequencing in Nonsyndromic Cleft Lip and Palate in a Honduran Population

Yi Cai ^{1,2}, Karynne E. Patterson³, Frederic Reinier³, Sarah E. Keesecker², Elizabeth Blue⁴, Michael Bamshad^{3,4}, and Joseph Haddad Jr^{*2}

Background: The majority of cleft lip with or without cleft palate cases appear as an isolated, nonsyndromic entity (NSCLP). With the advent of next generation sequencing, whole exome sequencing (WES) has been used to identify single nucleotide variants and insertion/deletions which cause or increase the risk of NSCLP. However, to our knowledge, there are no published studies using WES in NSCLP to investigate copy number changes (CNCs), which are a major component of human genetic variation. Our study aimed to identify CNCs associated with NSCLP in a Honduran population using WES. **Methods:** WES was performed on two to four members of 27 multiplex Honduran families. CNCs were identified using two algorithms, CoNIFER and XHMM. Priority was given to CNCs that were identified in more than one patient and had variant frequencies of less than 5% in reference data sets. **Results:** WES completion was defined as >90% of the WES target at >8 × coverage and >80% of the WES target at >20 × coverage. Twenty-

four CNCs that met our inclusion criteria were identified by both CoNIFER and XHMM. These CNCs were confirmed using quantitative PCR. Pedigree analysis produced three CNCs corresponding to *ADH7*, *AHR*, and *CRYZ* segregating with NSCLP. Two of the three CNCs implicate genes, *AHR* and *ADH7*, whose known biological functions could plausibly play a role in NSCLP. **Conclusion:** WES can be used to detect candidate CNCs that may be involved in the pathophysiology of NSCLP.

Birth Defects Research 109:1257–1267, 2017.

© 2017 Wiley Periodicals, Inc.

Key words: cleft lip; cleft palate; copy number changes; copy number variants; whole exome sequencing

Introduction

Cleft lip with or without palate (CL/P) is one of the most common birth defects worldwide, and is associated with major comorbidities, including feeding difficulties, speech delay, impaired hearing, dental problems, and psychiatric disease. CL/P occurs in approximately 1 in every 700 live births worldwide, but the prevalence may be as high as 1 in 500 in Asian and Amerindian populations (Dixon et al., 2011; Marazita, 2012). CL/P may occur as part of a syndrome, but the majority of cases (70%) appear as an isolated entity, nonsyndromic cleft lip with or without palate (NSCLP) (Calzolari et al., 2007).

It has long been thought that risk of developing NSCLP has a large genetic component (Dixon et al., 2011). However, identifying the specific genetic factors underlying risk

for NSCLP has proven challenging for several reasons. For example, inheritance of NSCLP often departs from traditional Mendelian modes of inheritance and, in some families, penetrance appears to be incomplete (Wyszynski, 2002; Beiraghi et al., 2007). Additionally, assigning affection status can be difficult because subclinical features, such as orbicularis oris defects, are part of the phenotypic spectrum of NSCLP (Neiswanger et al., 2007).

Prior targeted genetic studies have identified multiple candidate loci, including *IRF6*, *TGFA*, *RARA*, and *TGFB3* (Lidral and Moreno, 2005). Genome-wide linkage scans and genome-wide association studies (GWASs) have revealed additional candidate genes and susceptibility loci (Prescott et al., 2000; Wyszynski et al., 2003; Lidral and Moreno, 2005; Beaty et al., 2010; Mangold et al., 2010; Ludwig et al., 2012). In addition, environmental factors, such as smoking (Little et al., 2004; Zeiger et al., 2005; Honein et al., 2007), alcohol use (Shaw and Lammer, 1999; Chevrier et al., 2005; Romitti et al., 2007), nutrition, teratogens, and viral infections, also influence disease risk (Murray, 2002; Mossey et al., 2009; Dixon et al., 2011). These environmental factors may interact with genetic variants to modulate the risk of developing orofacial clefts. Such gene–environment interactions have been demonstrated with maternal smoking and *TGFA* (Zeiger et al., 2005), maternal folic acid consumption and *MTHFR* (van Rooij et al., 2003), and maternal multivitamin use and *NAT1* (Lammer et al., 2004).

The advent of next-generation sequencing has accelerated the discovery of new loci underlying Mendelian

Additional Supporting information may be found in the online version of this article.

¹Columbia University College of Physicians & Surgeons, New York, New York

²Department of Otolaryngology-Head and Neck Surgery, Columbia University Medical Center, New York, New York

³Department of Genome Sciences, University of Washington, Seattle, Washington

⁴Division of Medical Genetics, University of Washington, Seattle, Washington

Supported by the National Human Genome Research Institute and the National Heart, Lung and Blood Institute 2UM1HG006493.

*Correspondence to: Joseph Haddad, Suite 501N, 3959 Broadway, New York, NY 10032. E-mail: jh56@cumc.columbia.edu

Published online 27 July 2017 in Wiley Online Library (wileyonlinelibrary.com).
Doi: 10.1002/bdr2.1063

conditions and birth defects. Whole exome sequencing (WES) offers an efficient and powerful method for detecting pathogenic mutations, as exons comprise only 1% of the human genome but harbor 85% of disease-causing mutations (Choi et al., 2009). WES approaches have identified pathogenic single nucleotide variants (SNVs) and small insertions and deletions (indels) in NSCLP (Jezewski et al., 2003; Vieira et al., 2005; Bureau et al., 2014). However, fewer studies have investigated the role of copy number changes (CNCs) in NSCLP.

CNCs refer to portions of the genome present in variable number of copies between individuals or in comparison to a reference genome. CNCs have been defined at lengths of at least 50 base pairs to at least 1 kilobase (kb) (Feuk et al., 2006; Zarrei et al., 2015). In 2004, two GWASs revealed that CNCs are prevalent in human genomes and are an important source of genetic and phenotypic diversity (Iafrate et al., 2004; Sebat et al., 2004). Since then, CNCs have been studied and implicated in a wide breadth of human diseases, including autism, schizophrenia, obesity, type 1 diabetes, and various developmental disorders (Zarrei et al., 2015).

In the early years of CNC investigation, the predominant methods of CNC detection included fluorescent *in situ* hybridization, array comparative genomic hybridization, and single nucleotide polymorphism arrays. These methods could detect variants several kb to megabases in size (Stankiewicz and Lupski, 2010). With such methods, prior studies of CNCs in NSCLP have identified candidate regions (such as 7p14.1) or candidate genes (such as *SUMO1*, *BMP2*, and *CLPTM1L*), some of which have been validated in animal studies (Shi et al., 2009; Sahoo et al., 2011; Williams et al., 2012; Younkin et al., 2014; Klamt et al., 2016). In recent years, algorithms for calling CNCs from WES data have been developed, allowing for identification of CNCs as small as 50 base pairs (Fromer et al., 2012; Krumm et al., 2012; Tan et al., 2014). To our knowledge, no prior studies have used WES for this purpose in NSCLP.

Herein, we use WES to identify CNCs in a cohort of multiplex Honduran families with NSCLP. Studying multiplex families, those in which two or more persons have the same phenotype, increases the likelihood of finding alleles with larger effects that underlie NSCLP. Furthermore, it is advantageous to study the Honduran population because of their increased rate of clefting (given their Mesoamerican ancestry) and their relative genetic isolation due to limited influx of other ethnic populations (Dixon et al., 2011; Moreno-Estrada et al., 2013).

Materials and Methods

SAMPLES

Subjects were identified from patients at Hospital Escuela, a public hospital in Tegucigalpa, Honduras between 2001

and 2013. We recruited over 130 families with two or more members affected by NSCLP, although not all affected members were present to participate in this study. A medical history, family history, and physical exam were performed to characterize the type of cleft, exclude syndromic conditions, and construct pedigrees. Venous blood samples were obtained from both probands and available relatives. This study was approved by the Institutional Review Boards at Columbia University Medical Center and Hospital Escuela in Tegucigalpa, Honduras.

GENOTYPING AND WHOLE EXOME SEQUENCING

We selected 27 multiplex families with NSCLP for analysis, prioritizing families with DNA samples available for 2 or more affected individuals. The subjects included 52 affected individuals and 139 relatives. DNA was extracted from whole blood samples using the Qiagen Flexigene DNA kit (Qiagen, Valencia, CA). Two families were sequenced at Columbia University. The remaining samples were sent to the University of Washington Center for Mendelian Genomics (UW-CMG) in Seattle, Washington for sequencing. For sample quality control (QC), 191 subjects were genotyped using Illumina's Human Core Exome BeadChip. Variants missing >5% of genotypes were excluded.

PLINK v1.90 was used to confirm pedigree relationships using Mendelian error checking (<http://pngu.mgh.harvard.edu/purcell/plink/>) (Purcell et al., 2007). The BeadChip data were then used to estimate relationships using Kinship-based INference for GWASs (<http://people.virginia.edu/#wc9c/KING/>) (Manichaikul et al., 2010). QC included verification of sample/pedigree relationships. When discrepancies were observed, we collected new blood samples for confirmation testing. In cases where correction was not possible, QC led to the exclusion of two complete families as well as select individuals from three other families (one individual each from two families and two siblings from another). All other samples from affected individuals underwent WES. In four families, an unaffected member was included for variant phasing, bringing the total number of samples to 59 that underwent WES.

Exome capture was performed using Perkin-Elmer Janus II in 96-well plate format and Roche/Nimblegen SeqCap EZ at UW-CMG, as has been previously described (<http://uwcmg.org/#/instruction>) (Aylward et al., 2016). Library concentration was determined using quantitative PCR (qPCR). An Illumina HiSeq sequencer was used to massively parallel sequence samples and generate base calls. Unaligned BAM files were created using Picard Extract Illumina Barcodes and IlluminaBasecallsToSam and aligned to human reference GRCh37/hg19 using the Burrows-Wheeler Aligner v0.6.2 (Li and Durbin, 2009). QC measures were performed according to UW-CMG protocol (<http://uwcmg.org/#/instruction>) (Aylward et al., 2016).

Briefly, WES completion was defined as $>90\%$ of the WES target at $>8\times$ coverage and $>80\%$ of the WES target at $>20\times$ coverage.

CNC CALLING

CNC variant calling was performed using a large set of reference WES data from the UW-CMG ($N = 6085$) and the NHLBI GO Exome Sequencing Project (ESP; $N = 3635$) to reduce noise and improve calling quality (Fromer et al., 2012; Krumm et al., 2012). Because there is no gold standard for CNC calling using WES data, we restricted analysis to events detected by both XHMM v1.0 (Fromer et al., 2012) and CoNIFER v0.2.2 (Krumm et al., 2012) to decrease the number of false positives. The default software parameters were used for both tools. BED files from the two callsets were generated and Bedtools intersect was used to extract the intersecting calls (<http://bedtools.readthedocs.org/en/latest/content/tools/intersect.html>). Intersecting calls were defined as CNCs identified by both algorithms that shared the same genomic sections for 50% or more of their length.

CNC VALIDATION

qPCR was used to validate each CNC. Primers were designed using Primer3 (Rozen and Skaletsky, 2000) (Supplementary Table S1, which is available online), and qPCR was performed on an Applied Biosystems 7500 Real Time PCR System. Each sample was analyzed in triplicate, either in 25- μ l or 50- μ l reaction mixture, using SYBR Green I master mix (Roche Molecular Biochemicals), 200 nM of each primer, and 20 ng of genomic DNA. The default conditions supplied by the manufacturer (Applied Biosystems) were used for amplification. Data were normalized against the reference gene beta actin and relative gene expression, was determined using the Livak method (Livak and Schmittgen, 2001). Each qPCR experiment was performed in triplicate for each suspected CNC carrier, along with three control samples selected at random from a cohort of 100 healthy Honduran pediatric patients undergoing minor surgical procedures and not affected by clefting or other known genetic diseases. Relative gene expression values of ≥ 1.4 and ≤ 0.6 were considered evidence of duplication and deletion, respectively (Supplementary Table S1). For gene expression values between 0.6 and 1.4, we ruled out the corresponding CNCs as WES calling errors.

In addition, the population frequencies of the CNCs shown to segregate with NSCLP were calculated. Allele frequencies for CNCs were estimated by identifying events of the same type (duplication vs. deletion) with a minimum overlap of 50% of the observed CNC using Bedtools (v2.25.0) within reference data sets. Reference data included the 1000 genomes project integrated structural variant map (Sudmant et al., 2015) and the Exome

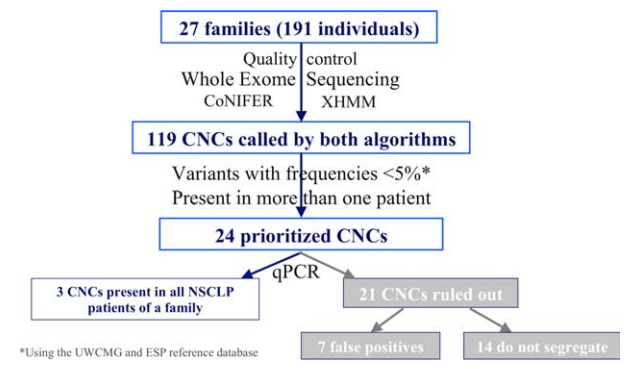


FIGURE 1. Summary of CNC calling, prioritization, and confirmation.

Aggregation Consortium (ExAC) release 0.3.1 data (Ruderman et al., 2016).

Results

CNC events observed in the WES data from multiplex Honduran families with a history of NSCLP were prioritized for validation as outlined in Figure 1.

XHMM identified 2545 CNCs while CoNIFER identified 1168 CNCs. After intersecting both call sets, 119 CNCs were identified by both XHMM and CoNIFER. To prioritize the CNCs most likely to be pathogenic, CNCs were filtered to focus on those identified in more than one patient and with variant frequencies $<5\%$ in the UW-CMG ($N = 6085$ individuals) and Exome Sequencing Project (ESP, $N = 3635$ individuals) reference data sets. This narrowed the list of CNCs of interest from 119 to 24 CNCs. These 24 CNCs ranged in size from approximately 4 kb to 226 kb, and the genes encompassed by the CNCs are listed in Table 1. After qPCR verification, seven CNCs were ruled out as CNC calling errors and 14 were excluded as they were not present in all patients affected by NSCLP within a pedigree (Fig. 1).

The three remaining CNCs were present in all members of a pedigree affected by NSCLP. Two of these CNCs were identified in a single family, observed in two affected brothers and their unaffected mother (Figs. 2 and 3). These CNCs were a duplication event of 7.7 kb on chromosome 4 and a deletion event of 10.3 kb on chromosome 1 corresponding to *ADH7* and *CRYZ*, respectively (Table 1). A CNC in *AHR* was identified in another family, observed in two affected brothers, their unaffected mother, and unaffected grandfather (Fig. 4). This was a deletion event of 13.3 kb and 23.6 kb in the affected brothers. The genomic contexts and population frequencies of these three CNCs are shown in Figure 5 and Table 2, respectively. These CNCs appear to fit inside single genes, and there are other events intersecting our variants in an online repository of CNCs with phenotypic information, DECIPHER (<https://decipher.sanger.ac.uk>) (Swaminathan et al., 2012). However, other than for the CNC associated with *ADH7*

TABLE 1. Twenty-Four CNVs Identified by both XHMM and CoNIFER and Present in More Than One NSCLP Patient

Gene	Event type	Length (kb)	Chrom-osome	Start/End coordinates ^a	Families	qPCR testing	Pedigree analysis ^b	CoNIFER Scores ^c	XHMM Scores ^c
ODFZL	Deletion	4.79	1	86847923 - 86852712	M28	TRUE	Does not segregate		
KIFAP3	Deletion	56.42, 60.35	1	169947224 - 170003639 169947224 - 170007574	M45, M94	FALSE			
HOOK1	Deletion	2.14, 8.89	1	60305970 - 60314857 60312716 - 60314857	M45	FALSE			
CRYZ	Deletion	10.27	1	75180237 - 75190507	M1	TRUE	Segregates	-1.71 - -1.64	-7.0 - -7.4
WDPCP	Deletion	50.1	2	63664553 - 63714656	M55, M62	FALSE			
APLF	Deletion	10.95	2	68729861 - 68740814	M28, M62	FALSE			
FSIP2 ^d	Deletion	8.82	2	186397377 - 186406199	M55, M62	TRUE	Does not segregate		
ARPP21	Deletion	9.26	3	35723242 - 35732499	M28, M62	TRUE	Does not segregate		
CACNA1D	Duplication	226.17	3	53699685 - 53925858	M71	TRUE	Does not segregate		
ADH7	Duplication	7.69	4	100334266 - 100341952	M1	TRUE	Segregates	1.72	4.56 - 5.93
ZNF608	Duplication	11.85	5	123973545 - 123985398	M55	TRUE	Does not segregate		
WDR36	Deletion	6.86, 9.38, 20.56	5	110436286 - 110443138 110436286 - 110456839 110439483 - 110448852	M28, M45, M46	TRUE	Does not segregate		
UFL1	Deletion	14.15, 14.56, 15.69, 19.27	6	96982117 - 97001381 96985248 - 96999808 96985248 - 96999401 96984118 - 96999808	M17, M28, M55, M62	FALSE			
TRDN	Deletion	31.02, 33.88	6	123539745 - 123573621 123542598 - 123573621	M28, M62	FALSE			
NKAIN2	Duplication	165.09	6	124979330 - 125144422	M46	TRUE	Does not segregate		
CACNA2D1	Deletion	4.74, 10.53	7	81593347 - 81603871 81596453 - 81601188	M45, M62	TRUE	Does not segregate		

TABLE 1. Continued

Gene	Event type	Length (kb)	Chrom-osome	Start/end coordinates ^a	Families	qPCR testing	Pedigree analysis ^b	CoNIFER Scores ^c	XHMM Scores ^c
AHR	Deletion	13.3, 23.6	7	17349559 - 17843203 17362123 - 17841282 17367382 - 17841282	M45	TRUE	Segregates	-0.735 - -0.517	-2.76
DDX18 ^d	Deletion	133.09	9	118289457 - 118422544	M28, M62	TRUE	Does not segregate		
DHTKD1	Duplication	36.35, 39.42, 51.86	10	12123469 - 12162889 12126537 - 12162889 12111031 - 12162889	M16, M67	TRUE	Does not segregate		
FAM76B	Deletion	8.4, 11.71	11	95504718 - 95513118 95504718 - 95516430	M28, M45	FALSE			
PCNX	Duplication	147.71	14	71428941 - 71576654	M94	TRUE	Does not segregate		
VPS13C	Deletion	32.2, 35.63	15	62304283 - 62336484 62300852 - 62336484	M28, M45	TRUE	Does not segregate		
CGNL1	Duplication	23.9	15	57730196 - 57754092	M53, M71	TRUE	Does not segregate		
C18orf54	Deletion	8.59, 11.53	18	57730196 - 57754092	M28, M45	TRUE	Does not segregate		

The CNC details and families in which they were called are listed.

^aStart and end coordinates from XHMM are listed.

^bSegregation of a CNC with NSCLP is defined as presence of the CNC in all family members affected by NSCLP within a family.

^cThe CoNIFER score represents the median signal strength of the calls (median_svdzpkm). The XHMM score shows the mean normalized read depth Z-scores over the interval.

^dFor CNC calls corresponding to intergenic regions, the gene for the closest exon is listed.

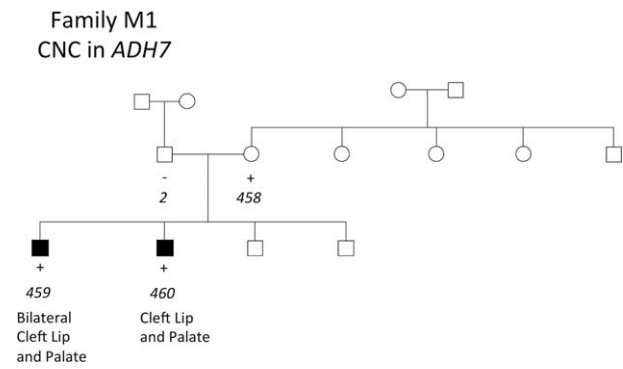


FIGURE 2. Presence of CNCs in *ADH7* in NSCLP patients in a Honduran family. +, positive for a CNC in *ADH7*. DNA was not available for subjects without + or - denoted. Sample identification numbers are italicized and correspond to Supplementary Table S1.

(described in the Discussion section), none of the overlapping CNCs correspond to a phenotype of cleft lip or palate.

Discussion

This study identifies candidate CNCs for NSCLP using WES technology. By using this technique on a cohort of multiplex families affected by NSCLP, we identified CNCs of potential significance corresponding to the genes *ADH7* (formerly referred to as *ADH3*), *AHR*, and *CRYZ* using our CNC filtering and prioritization criteria. Two of these genes, *ADH7* and *AHR*, share involvement in biological pathways linked to environmental factors known to influence NSCLP.

A CNC in *ADH7* was identified in two affected siblings as well as their unaffected mother. Of interest, prior analysis of WES data in this cohort did not identify causal SNVs, insertions, or deletions that segregated with NSCLP in this family (M1) (Aylward et al., 2016). *ADH7* encodes a member of the alcohol dehydrogenase family, class IV alcohol dehydrogenase, expressed ubiquitously during

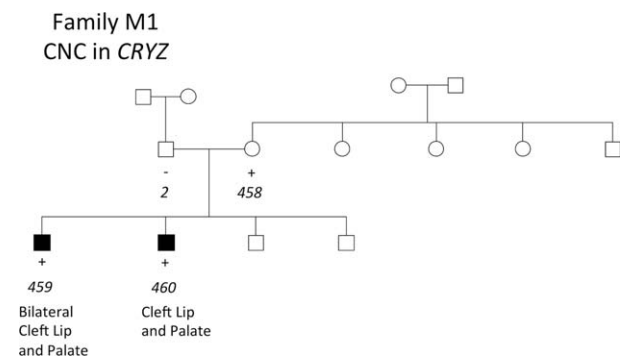


FIGURE 3. Presence of CNCs in *CRYZ* in NSCLP patients in family M1. +, positive for a CNC in *CRYZ*. Sample identification numbers are italicized and correspond to Supplementary Table S1.

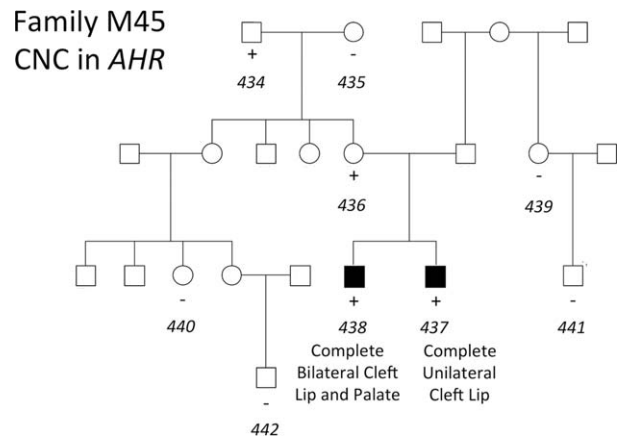


FIGURE 4. Presence of CNCs in *AHR* with NSCLP in family M45. +, positive for a CNC in *AHR*. Sample identification numbers are italicized and correspond to Supplementary Table S1.

embryogenesis (Molotkov et al., 2002). Compared with other members of its class, it is less efficient in ethanol oxidation and most active as a retinol dehydrogenase (Satre et al., 1994). Thus, *ADH7* may participate in the synthesis of retinoic acid, the active form of vitamin A. Retinoic acid plays an important role in cellular differentiation (Niederreither and Dolle, 2008) and is a well-established cause of cleft palate (Abbott and Birnbaum, 1990; Abbott et al., 1989b). Furthermore, prior studies of CL/P identified a significant association with a locus in the retinoic acid receptor (*RARA*) (Chenevix-Trench et al., 1992; Shaw et al., 1993).

There are no published studies linking *ADH7* to orofacial clefts, but there are additional cases with CNCs involving *ADH7* linked to CL/P in DECIPHER (Swaminathan et al., 2012). A duplication event of 8.08 Mb was associated with a nonmidline cleft lip in one patient (DECIPHER ID: 270855) and a deletion event of 5.81 Mb was associated with bilateral CL/P in the other patient (DECIPHER ID: 285906) (Fig. 5). Both patients had other related disorders, suggesting these presentations of CL/P were syndromic. This CNC has also been identified in a Non-Finnish European population in ExAC at a frequency of <0.001% (Table 2).

CNCs in *AHR* were identified in two affected siblings as well as their unaffected mother and maternal grandfather, consistent with an autosomal dominant mode of inheritance with incomplete penetrance. *AHR* encodes the aryl-hydrocarbon receptor (AHR), which is expressed in the developing mouse palate and upregulated early in palatogenesis (Abbott et al., 1999). This receptor mediates the toxicities of aromatic hydrocarbons that may result in teratogenesis, cancers, and birth defects (Whitlock, 1990; Nebert et al., 2004). Hydrocarbons bind to AHR and activate downstream signaling pathways resulting in various

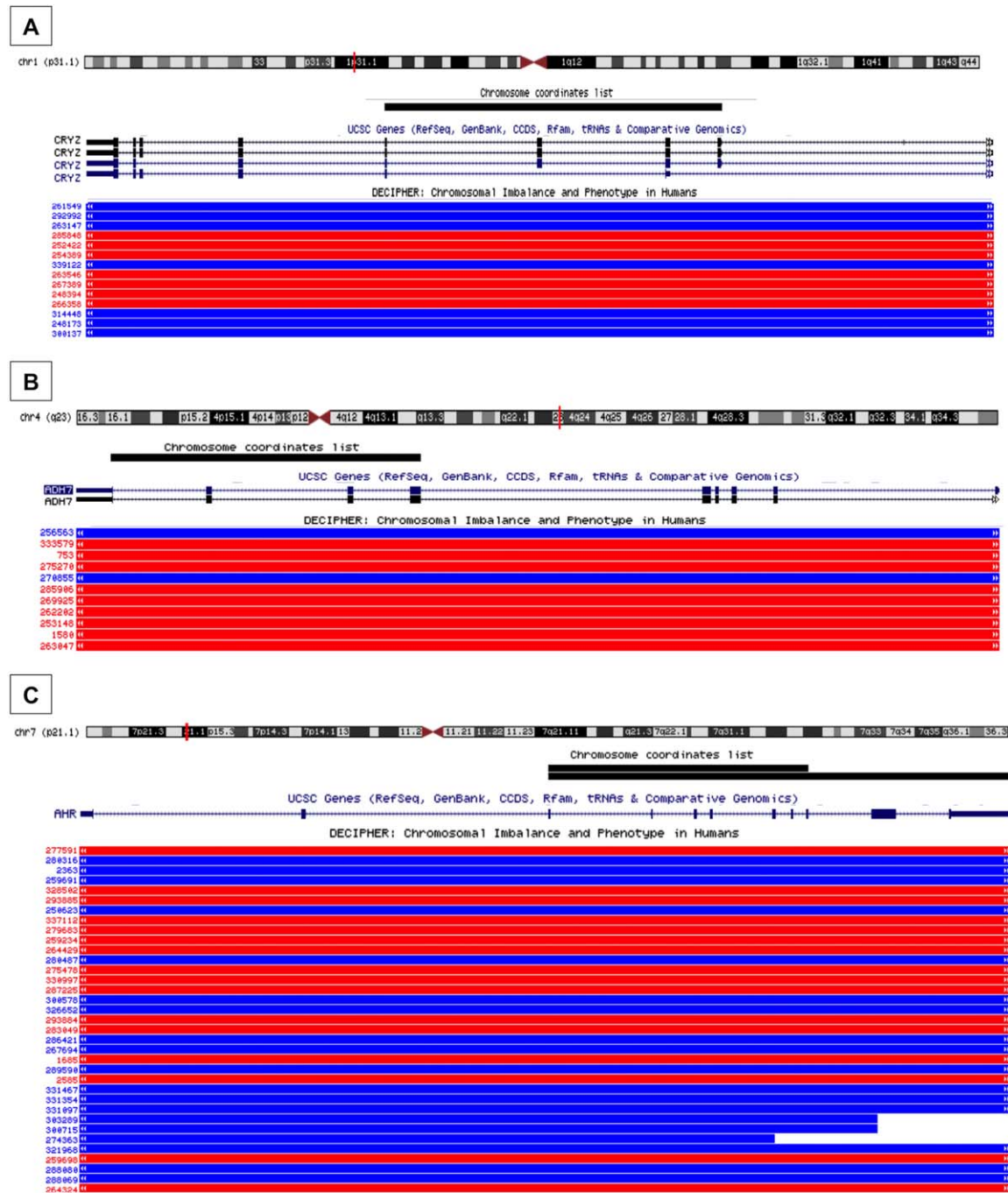


FIGURE 5. Genomic contexts of the CNCs identified in NSCLP patients. This figure was generated using University of California Santa Cruz Genome Browser. The genomic regions spanned by the CNCs corresponding to CRYZ (**A**), ADH7 (**B**), and AHR (**C**) are shown and represented by the black bar(s) labeled "Chromosome coordinates list." The boundaries of each CNC are indicated by the red box on the chromosome ideogram and the encompassed RefSeq genes are shown. In addition, other CNVs that have been previously reported in DECIPHER are shown in the red (deletion/loss) or blue (duplication/gain) bars.

toxicities (Poland and Knutson, 1982; Landers and Bunce, 1991). Of particular relevance to NSCLP, the aromatic hydrocarbon dioxin (2,3,7,8-tetrachlorodibenzo-p-dioxin) is a ligand of AHR and has been shown to induce cleft palate in

pregnant mice (Pratt et al., 1984; Abbott et al., 1989a; Takagi et al., 2000). Of interest, retinoic acid is involved in and necessary for the development of dioxin-induced cleft palate in mice by regulating AHR expression (Jacobs et al., 2011).

TABLE 2. Population Frequencies of CNCs in the Genes *CRYZ*, *ADH7*, and *AHR* That Segregated with NSCLP in a Cohort of Honduran Patients

Gene	Event	1000 Genomes frequencies					ExAC frequencies						
		AFR	AMR	EAS	EUR	SAS	AFR	AMR	EAS	FIN	NFE	OTH	SAS
<i>CRYZ</i>	deletion	0/1322	2/694	0/1008	0/1006	0/978	0/10406	1/11578	0/8654	0/6614	0/66740	0/908	0/16512
<i>ADH7</i>	duplication	0/1322	0/694	0/1008	0/1006	0/978	6/10406	0/11578	0/8654	0/6614	1/66740	0/908	0/16512
<i>AHR</i>	deletion	0/1322	0/694	0/1008	0/1006	0/978	0/10406	0/11578	0/8654	0/6614	2/66740	0/908	0/16512

Allele frequencies are presented as the number of observations over the maximum number of chromosomes within a data set, as exact numbers of chromosomes observed at a given position were not provided in the ExAC data release. These values should then be considered minimum frequencies in the reference data set. Subpopulations include African (AFR), American (AMR), East Asian (EAS), European (EUR), Finnish (FIN), Non-Finnish European (NFE), other (OTH), and South Asian (SAS) populations.

There is considerable support for a role for *AHR* in cleft palate development in mice, but it may have less significance to cleft palate development in humans. In one study, *AHR* mRNA in human embryos was found at levels 350-fold less than in mouse embryos (Abbott et al., 1999). In addition, humans may be less sensitive to the pathogenic effects of dioxin than are mice (Moriguchi et al., 2003). Induction of cleft palate using dioxin required a much higher concentration of dioxin in human embryos than in mouse embryos (Abbott and Birnbaum, 1991). Nonetheless, CNCs in *AHR* warrant further investigation given their segregation with NSCLP in one of our multiplex families and the gene's overall association with craniofacial development. Furthermore, single nucleotide polymorphisms in *AHR*'s cofactor, *AHR* nuclear translocator (*ARNT*), have been associated with nonsyndromic orofacial clefts in the Japanese population (Kayano et al., 2004). From our literature search, this CNC does not overlap with those that have been previously reported in association with NSCLP. This CNC has been identified in the African subpopulation in ExAC at a frequency of 0.06% and the Non-Finnish European subpopulation at frequency <0.01% (Table 2).

We also identified a CNC in *CRYZ* present in patients affected by NSCLP. The pedigree corresponding to *CRYZ* suggests an autosomal dominant inheritance pattern with incomplete penetrance in family M1 (Fig. 3). It is worth noting that members in this same family also had CNCs in *ADH7*. This finding of multiple candidate variants within the same patient and/or family has also been demonstrated by prior studies of SNVs or indels in our Honduran families (Aylward et al., 2016) and of CNCs in other cohorts (Simioni et al., 2015). These findings would be consistent with the polygenic nature of NSCLP and incomplete penetrance. However, there is less evidence supporting a major role for this gene in the pathogenesis of NSCLP. *CRYZ* encodes crystallin zeta, a quinone reductase expressed in the eye lens of vertebrates (Gonzalez et al., 1994). Its biologic function in humans is less well established, although a GWAS associated *CRYZ* with regulation

of resistin, a hormone implicated in diabetes in cardiovascular disease (Qi et al., 2012). Our literature search did not find reports of an association of *CRYZ* with NSCLP, although this CNC has been identified in the American subpopulation in 1000 Genomes project and ExAC at frequencies of 0.29% and 0.09%, respectively (Table 2).

Conclusion

We have identified candidate CNCs that rank highly based on prioritization criteria in three separate genes. Two of these genes, *ADH7* and *AHR*, play roles in craniofacial development. Interestingly, both *ADH7* and *AHR* interact with environmental factors, retinoic acid and dioxin, respectively, both of which have well-established roles in clefting. Specifically, *ADH7* participates in retinoic acid synthesis and retinoic acid regulates the expression of *AHR*. Our preliminary results suggest that these candidate genes identified by means of CNC analysis in exome sequencing data warrant further investigation as risk factors for NSCLP. Replication of these findings in a larger cohort of families with NSCLP or a different population is necessary to confirm or refute their role in the pathogenesis of NSCLP.

Acknowledgments

We thank the staff at Hospital Escuela, Tegucigalpa, Honduras, and the Honduran Medical Institute for their assistance in identifying patients and obtaining samples. Sequencing was provided by UW-CMG and was funded by the National Human Genome Research Institute and the National Heart, Lung and Blood Institute to D.N., M.B., and S.L. This study makes use of data generated by the DECIPHER Consortium. A full list of centers who contributed to the generation of the data is available from <https://decipher.sanger.ac.uk/> and by means of email from decipher@sanger.ac.uk. Funding for the project was provided by the Wellcome Trust. Those who carried out the original analysis and collection of the data bear no responsibility for the further analysis or interpretation of it by the recipient or its registered users. The DDD study

presents independent research commissioned by the Health Innovation Challenge Fund (grant number HICF-1009-003), a parallel funding partnership between the Wellcome Trust and the Department of Health, and the Wellcome Trust Sanger Institute (grant number WT098051). The views expressed in this publication are those of the author(s) and not necessarily those of the Wellcome Trust or the Department of Health. The study has UK Research Ethics Committee approval (10/H0305/83, granted by the Cambridge South REC, and GEN/284/12 granted by the Republic of Ireland REC). The research team acknowledges the support of the National Institute for Health Research, through the Comprehensive Clinical Research Network.

References

- Abbott BD, Birnbaum LS. 1990. Retinoic acid-induced alterations in the expression of growth factors in embryonic mouse palatal shelves. *Teratology* 42:597–610.
- Abbott BD, Birnbaum LS. 1991. TCDD exposure of human embryonic palatal shelves in organ culture alters the differentiation of medial epithelial cells. *Teratology* 43:119–132.
- Abbott BD, Diliberto JJ, Birnbaum LS. 1989a. 2,3,7,8-Tetrachlorodibenzo-p-dioxin alters embryonic palatal medial epithelial cell differentiation in vitro. *Toxicol Appl Pharmacol* 100:119–131.
- Abbott BD, Harris MW, Birnbaum LS. 1989b. Etiology of retinoic acid-induced cleft palate varies with the embryonic stage. *Teratology* 40:533–553.
- Abbott BD, Held GA, Wood CR, et al. 1999. AhR, ARNT, and CYP1A1 mRNA quantitation in cultured human embryonic palates exposed to TCDD and comparison with mouse palate in vivo and in culture. *Toxicol Sci* 47:62–75.
- Aylward A, Cai Y, Lee A, et al. 2016. Using whole exome sequencing to identify candidate genes with rare variants in nonsyndromic cleft lip and palate. *Genet Epidemiol* 40:432–441.
- Beatty TH, Murray JC, Marazita ML, et al. 2010. A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. *Nat Genet* 42:525–529.
- Beiraghi S, Nath SK, Gaines M, et al. 2007. Autosomal dominant nonsyndromic cleft lip and palate: significant evidence of linkage at 18q21.1. *Am J Hum Genet* 81:180–188.
- Bureau A, Parker MM, Ruczinski I, et al. 2014. Whole exome sequencing of distant relatives in multiplex families implicates rare variants in candidate genes for oral clefts. *Genetics* 197:1039–1044.
- Calzolari E, Pierini A, Astolfi G, et al. 2007. Associated anomalies in multi-malformed infants with cleft lip and palate: An epidemiologic study of nearly 6 million births in 23 EUROCAT registries. *American journal of medical genetics Part A* 143A:528–537.
- Chenevix-Trench G, Jones K, Green AC, et al. 1992. Cleft lip with or without cleft palate: associations with transforming growth factor alpha and retinoic acid receptor loci. *Am J Hum Genet* 51:1377–1385.
- Chevrier C, Perret C, Bahuau M, et al. 2005. Interaction between the ADH1C polymorphism and maternal alcohol intake in the risk of nonsyndromic oral clefts: an evaluation of the contribution of child and maternal genotypes. *Birth Defects Res A Clin Mol Teratol* 73:114–122.
- Choi M, Scholl UI, Ji W, et al. 2009. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A* 106:19096–19101.
- Dixon MJ, Marazita ML, Beatty TH, Murray JC. 2011. Cleft lip and palate: understanding genetic and environmental influences. *Nat Rev Genet* 12:167–178.
- Feuk L, Carson AR, Scherer SW. 2006. Structural variation in the human genome. *Nat Rev Genet* 7:85–97.
- Fromer M, Moran JL, Chambert K, et al. 2012. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet* 91:597–607.
- Gonzalez P, Rao PV, Zigler JS Jr. 1994. Organization of the human zeta-crystallin/quinone reductase gene (CRYZ). *Genomics* 21:317–324.
- Honein MA, Rasmussen SA, Reefhuis J, et al. 2007. Maternal smoking and environmental tobacco smoke exposure and the risk of orofacial clefts. *Epidemiology* 18:226–233.
- Iafrate AJ, Feuk L, Rivera MN, et al. 2004. Detection of large-scale variation in the human genome. *Nat Genet* 36:949–951.
- Jacobs H, Dennefeld C, Feret B, et al. 2011. Retinoic acid drives aryl hydrocarbon receptor expression and is instrumental to dioxin-induced toxicity during palate development. *Environ Health Perspect* 119:1590–1595.
- Jezewski PA, Vieira AR, Nishimura C, et al. 2003. Complete sequencing shows a role for MSX1 in non-syndromic cleft lip and palate. *J Med Genet* 40:399–407.
- Kayano S, Suzuki Y, Kanno K, et al. 2004. Significant association between nonsyndromic oral clefts and arylhydrocarbon receptor nuclear translocator (ARNT). *Am J Med Genet A* 130A:40–44.
- Klamt J, Hofmann A, Bohmer AC, et al. 2016. Further evidence for deletions in 7p14.1 contributing to nonsyndromic cleft lip with or without cleft palate. *Birth Defects Res A Clin Mol Teratol* 106:767–772.
- Krumm N, Sudmant PH, Ko A, et al. 2012. Copy number variation detection and genotyping from exome sequence data. *Genome Res* 22:1525–1532.
- Lammer EJ, Shaw GM, Iovannisci DM, Finnell RH. 2004. Periconceptional multivitamin intake during early pregnancy, genetic variation of acetyl-N-transferase 1 (NAT1), and risk for orofacial clefts. *Birth Defects Res A Clin Mol Teratol* 70:846–852.
- Landers JP, Bunce NJ. 1991. The Ah receptor and the mechanism of dioxin toxicity. *Biochem J* 276(Pt 2):273–287.

- Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760.
- Lidral AC, Moreno LM. 2005. Progress toward discerning the genetics of cleft lip. *Curr Opin Pediatr* 17:731–739.
- Little J, Cardy A, Munger RG. 2004. Tobacco smoking and oral clefts: a meta-analysis. *Bull World Health Organ* 82:213–218.
- Livak KJ, Schmittgen TD. 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2^{(-Delta Delta C(T))} Method. *Methods* 25:402–408.
- Ludwig KU, Mangold E, Herms S, et al. 2012. Genome-wide meta-analyses of nonsyndromic cleft lip with or without cleft palate identify six new risk loci. *Nat Genet* 44:968–971.
- Mangold E, Ludwig KU, Birnbaum S, et al. 2010. Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. *Nat Genet* 42:24–26.
- Manichaikul A, Mychaleckyj JC, Rich SS, et al. 2010. Robust relationship inference in genome-wide association studies. *Bioinformatics* 26:2867–2873.
- Marazita ML. 2012. The evolution of human genetic studies of cleft lip and cleft palate. *Annu Rev Genomics Hum Genet* 13:263–283.
- Molotov A, Fan X, Deltour L, et al. 2002. Stimulation of retinoic acid production and growth by ubiquitously expressed alcohol dehydrogenase Adh3. *Proc Natl Acad Sci U S A* 99:5337–5342.
- Moreno-Estrada A, Gravel S, Zakharia F, et al. 2013. Reconstructing the population genetic history of the Caribbean. *PLoS Genet* 9:e1003925.
- Moriguchi T, Motohashi H, Hosoya T, et al. 2003. Distinct response to dioxin in an arylhydrocarbon receptor (AHR)-humanized mouse. *Proc Natl Acad Sci U S A* 100:5652–5657.
- Mossey PA, Little J, Munger RG, et al. 2009. Cleft lip and palate. *Lancet* 374:1773–1785.
- Murray JC. 2002. Gene/environment causes of cleft lip and/or palate. *Clin Genet* 61:248–256.
- Nebert DW, Dalton TP, Okey AB, Gonzalez FJ. 2004. Role of aryl hydrocarbon receptor-mediated induction of the CYP1 enzymes in environmental toxicity and cancer. *J Biol Chem* 279:23847–23850.
- Neiswanger K, Weinberg SM, Rogers CR, et al. 2007. Orbicularis oris muscle defects as an expanded phenotypic feature in non-syndromic cleft lip with or without cleft palate. *Am J Med Genet A* 143A:1143–1149.
- Niederreither K, Dolle P. 2008. Retinoic acid in development: towards an integrated view. *Nat Rev Genet* 9:541–553.
- Poland A, Knutson JC. 1982. 2,3,7,8-tetrachlorodibenzo-p-dioxin and related halogenated aromatic hydrocarbons: examination of the mechanism of toxicity. *Annu Rev Pharmacol Toxicol* 22:517–554.
- Pratt RM, Dencker L, Diewert VM. 1984. 2,3,7,8-Tetrachlorodibenzo-p-dioxin-induced cleft palate in the mouse: evidence for alterations in palatal shelf fusion. *Teratog Carcinog Mutagen* 4:427–436.
- Prescott NJ, Lees MM, Winter RM, Malcolm S. 2000. Identification of susceptibility loci for nonsyndromic cleft lip with or without cleft palate in a two stage genome scan of affected sib-pairs. *Hum Genet* 106:345–350.
- Purcell S, Neale B, Todd-Brown K, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575.
- Qi Q, Menzaghi C, Smith S, et al. 2012. Genome-wide association analysis identifies TYW3/CRYZ and NDST4 loci associated with circulating resistin levels. *Hum Mol Genet* 21:4774–4780.
- Romitti PA, Sun L, Honein MA, et al. 2007. Maternal periconceptional alcohol consumption and risk of orofacial clefts. *Am J Epidemiol* 166:775–785.
- Rozen S, Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 132:365–386.
- Ruderfer DM, Hamamsy T, Lek M, et al. 2016. Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. *Nat Genet* 48:1107–1111.
- Sahoo T, Theisen A, Sanchez-Lara PA, et al. 2011. Microdeletion 20p12.3 involving BMP2 contributes to syndromic forms of cleft palate. *Am J Med Genet A* 155A:1646–1653.
- Satre MA, Zgombic-Knight M, Duester G. 1994. The complete structure of human class IV alcohol dehydrogenase (retinol dehydrogenase) determined from the ADH7 gene. *J Biol Chem* 269:15606–15612.
- Sebat J, Lakshmi B, Troge J, et al. 2004. Large-scale copy number polymorphism in the human genome. *Science* 305:525–528.
- Shaw D, Ray A, Marazita M, Field L. 1993. Further evidence of a relationship between the retinoic acid receptor alpha locus and nonsyndromic cleft lip with or without cleft palate (CL + /- P). *Am J Hum Genet* 53:1156–1157.
- Shaw GM, Lammer EJ. 1999. Maternal periconceptional alcohol consumption and risk for orofacial clefts. *J Pediatr* 134:298–303.
- Shi M, Mostowska A, Jugessur A, et al. 2009. Identification of microdeletions in candidate genes for cleft lip and/or palate. *Birth Defects Res A Clin Mol Teratol* 85:42–51.
- Simioni M, Araujo TK, Monlleo IL, et al. 2015. Investigation of genetic factors underlying typical orofacial clefts: mutational screening and copy number variation. *J Hum Genet* 60:17–25.
- Stankiewicz P, Lupski JR. 2010. Structural variation in the human genome and its role in disease. *Annu Rev Med* 61:437–455.

- Sudmant PH, Rausch T, Gardner EJ, et al. 2015. An integrated map of structural variation in 2,504 human genomes. *Nature* 526:75–81.
- Swaminathan GJ, Bragin E, Chatzimichali EA, et al. 2012. DECIPHER: web-based, community resource for clinical interpretation of rare variants in developmental disorders. *Hum Mol Genet* 21(R1):R37–R44.
- Takagi TN, Matsui KA, Yamashita K, et al. 2000. Pathogenesis of cleft palate in mouse embryos exposed to 2,3,7, 8-tetrachlorodibenzo-p-dioxin (TCDD). *Teratog Carcinog Mutagen* 20:73–86.
- Tan R, Wang Y, Kleinstein SE, et al. 2014. An evaluation of copy number variation detection tools from whole-exome sequencing data. *Hum Mutat* 35:899–907.
- van Rooij IA, Vermeij-Keers C, Kluijtmans LA, et al. 2003. Does the interaction between maternal folate intake and the methylenetetrahydrofolate reductase polymorphisms affect the risk of cleft lip with or without cleft palate? *Am J Epidemiol* 157:583–591.
- Vieira AR, Avila JR, Daack-Hirsch S, et al. 2005. Medical sequencing of candidate genes for nonsyndromic cleft lip and palate. *PLoS Genet* 1:e64.
- Whitlock JP Jr. 1990. Genetic and molecular aspects of 2,3,7,8-tetrachlorodibenzo-p-dioxin action. *Annu Rev Pharmacol Toxicol* 30:251–277.
- Williams ES, Uhas KA, Bunke BP, et al. 2012. Cleft palate in a multigenerational family with a microdeletion of 20p12.3 involving BMP2. *Am J Med Genet A* 158A:2616–2620.
- Wyszynski DF. 2002. Cleft lip and palate: from origin to treatment. New York: Oxford University Press.
- Wyszynski DF, Albacha-Hejazi H, Aldirani M, et al. 2003. A genome-wide scan for loci predisposing to non-syndromic cleft lip with or without cleft palate in two large Syrian families. *Am J Med Genet A* 123A:140–147.
- Younkin SG, Scharpf RB, Schwender H, et al. 2014. A genome-wide study of de novo deletions identifies a candidate locus for non-syndromic isolated cleft lip/palate risk. *BMC Genet* 15:24.
- Zarrei M, MacDonald JR, Merico D, Scherer SW. 2015. A copy number variation map of the human genome. *Nat Rev Genet* 16:172–183.
- Zeiger JS, Beaty TH, Liang KY. 2005. Oral clefts, maternal smoking, and TGFA: a meta-analysis of gene-environment interaction. *Cleft Palate Craniofac J* 42:58–63.