


ORIGINAL ARTICLE

Analysis of sequence data to identify potential risk variants for oral clefts in multiplex families

Emily R. Holzinger^{1,2,*} , Qing Li¹, Margaret M. Parker^{3,4}, Jacqueline B. Hetmanski⁵, Mary L. Marazita⁶, Elisabeth Mangold⁷, Kerstin U. Ludwig^{7,8}, Margaret A. Taub⁹, Ferdouse Begum⁵, Jeffrey C. Murray¹⁰, Hasan Albacha-Hejazi¹¹, Khalid Alqosayer¹², Giath Al-Souki¹³, Abdullatiff Albasha Hejazi¹⁴, Alan F. Scott^{15,16}, Terri H. Beaty⁵ & Joan E. Bailey-Wilson^{1,*}

¹Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland

²National Institute of General Medical Sciences, National Institutes of Health, Bethesda, Maryland

³Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, Massachusetts

⁴Harvard Medical School, Cambridge, Massachusetts

⁵Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland

⁶Department of Oral Biology, Center for Craniofacial and Dental Genetics, School of Maryland Dental Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania

⁷Institute of Human Genetics, University of Bonn, Bonn, Germany

⁸Department of Genomics, Life & Brain Center, University of Bonn, Bonn, Germany

⁹Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland

¹⁰Department of Pediatrics, University of Iowa, Iowa City, Iowa

¹¹Hejazi Clinic, Damascus, Syrian Arab Republic

¹²Prime Health Clinic, Jeddah, Saudi Arabia

¹³Saudi Red Crescent, Jeddah, Saudi Arabia

¹⁴Al-Eqtisad Est., Jeddah, Saudi Arabia

¹⁵Center for Inherited Disease Research, Johns Hopkins School of Medicine, Baltimore, Maryland

¹⁶Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland

Keywords

DNA sequence data analysis, genetic risk variants, oral clefts, rare variants, statistical genetics

Correspondence

Emily R. Holzinger, Triad Technology Center, 333 Cassell Drive, Suite 1200, Baltimore, MD 21224. Tel: 443 740 2915; Fax: 443 740 2165; E-mail: emily.holzinger@nih.gov

and

Joan E. Bailey-Wilson, Triad Technology Center, 333 Cassell Drive, Suite 1200, Baltimore, MD 21224. Tel: 443 740 2921; Fax: 443 740 2165; E-mail: jebw@mail.nih.gov

Funding Information

This study was supported by the National Institutes of Health (grant no.: P50-DE-016215, R01-DE-009886, R01-DE-014581, R01-DE-016148, R37-DE-08559, U01-DE-018993, U01-DE-020073, X01 HG006177; contract no.: HHSN268200782096C).

Received: 18 February 2017; Revised: 12 June 2017; Accepted: 14 June 2017

Molecular Genetics & Genomic Medicine
2017; 5(5): 570–579

doi: 10.1002/mgg3.320

570

Abstract

Background

Nonsyndromic oral clefts are craniofacial malformations, which include cleft lip with or without cleft palate. The etiology for oral clefts is complex with both genetic and environmental factors contributing to risk. Previous genome-wide association (GWAS) studies have identified multiple loci with small effects; however, many causal variants remain elusive.

Methods

In this study, we address this by specifically looking for rare, potentially damaging variants in family-based data. We analyzed both whole exome sequence (WES) data and whole genome sequence (WGS) data in multiplex cleft families to identify variants shared by affected individuals.

Results

Here we present the results from these analyses. Our most interesting finding was from a single Syrian family, which showed enrichment of nonsynonymous and potentially damaging rare variants in two genes: *CASP9* and *FAT4*.

Conclusion

Neither of these candidate genes has previously been associated with oral clefts and, if confirmed as contributing to disease risk, may indicate novel biological pathways in the genetic etiology for oral clefts.

Introduction

Nonsyndromic oral clefts, including cleft lip with or without cleft palate (CL/P) and cleft palate (CP) alone, are the most common craniofacial malformations in humans. The etiology of oral clefts is complex and heterogeneous with different environmental and genetic factors contributing to risk. Previous linkage and genome-wide association studies (GWAS) have identified multiple genes and regions associated with risk for CL/P. However, it is estimated that these regions only account for 20–25% of the heritability. Improvements in sequencing technology allows us to expand our search for causal variants even further (Beaty *et al.* 2016). Recently, we have identified a novel, potentially damaging variant in *CDH1* in one multiplex CL/P family based on whole exome sequence (WES) data (Bureau *et al.* 2014). For this analysis, we used WES and whole genome sequence (WGS) data in families with distantly related affected individuals (second or third degree relationships) to identify genes containing shared rare variants.

The goal of this study was to identify novel rare variants shared at the population or the family level that may contribute to risk for oral cleft phenotypes. Our most notable finding came from a single Syrian family where we identified enrichment of nonsynonymous and potentially damaging rare variants in two genes: *CASP9* and *FAT4*. Furthermore, the *CASP9* variant was not present in any of the other affected individuals in this study, including the other Syrian families, nor is it reported in any of the major variant databases. Here we provide a summary of our results with a focus on the most interesting findings to assess the possible biological roles of these rare variants in influencing risk for oral clefts.

Materials and Methods

Ethical compliance

All studies were approved by the local Institutional Review/Ethics Boards and followed the tenets of the Declaration of Helsinki.

Data collection

The multiplex cleft families studied here were originally ascertained and recruited by different studies for linkage analysis. Families were enrolled because they had at least two biological relatives affected with an apparent nonsyndromic oral cleft. Some families have been previously genotyped and included in published linkage analyses (Wyszynski *et al.* 2003; Field *et al.* 2004; Marazita *et al.*

2004, 2009; Schultz *et al.* 2004; Riley *et al.* 2007; Mangold *et al.* 2009), but the specific marker panels varied and provided sparse coverage of the genome. Families were enrolled in studies in Germany, India, the Philippines, and the Syrian Arab Republic. Each study was conducted somewhat differently, but, in general, a patient with nonsyndromic oral cleft was identified, a preliminary family history investigation revealed at least one additional affected relative existed, and the family was evaluated for potential informativeness for linkage studies. Multiplex families identified as informative were enrolled, and both affected and unaffected relatives were consented and recruited. Study participants were examined to confirm their phenotypic status, a DNA sample was collected, and, for some individuals, limited information concerning potential environmental risk factors (such as mother's smoking history during pregnancy) was available. All three types of oral clefts were identified in these families: cleft lip and palate, cleft palate only, and cleft lip only. We also obtained information on the location of the cleft (right, left, bilateral, or midline), and whether it was a complete or incomplete cleft. The specific oral cleft phenotype varied within and between families. Families were selected for this study if DNA samples were available for at least two second or third degree affected relatives who had given informed consent adequate for DNA sequence analysis. The second degree relatives included half-sibs, avuncular, or grandparental pairs, while the third degree relatives included first cousins and great-avuncular pairs). Some more distant affected relatives such as second cousins and first cousins once removed were also included. Due to funding constraints, only affected family members were sequenced in almost all families and parents of affected individuals were not sequenced.

For the WES portion of the study, we sequenced 108 affected individuals from 52 families (four of the families each had three affected individuals sequenced, and the remaining 48 families each had two affected individuals sequenced, four duplicate subjects for quality control, and two unrelated controls from the CEU HAPMAP population (Utah residents with Northern and Western European ancestry from the Centre d'Etude de Polymorphisme Humain (CEPH) data collection)). These multiplex families were from Syrian, Filipino, Indian, and German populations (Table 1). All samples were sequenced at the Center for Inherited Disease Research (CIDR).

For the WGS data, there were 113 sequenced individuals (107 affected, six unaffected) from 32 families, all sequenced by Illumina (13 families with two affected individuals sequenced, two families with three affected individuals, seven families with four affected individuals, two families with three affected individuals and one

Table 1. Number of affected individuals with nonsyndromic oral clefts¹ and DNA sequence data.

Population	Individuals (families) with WES data	Individuals (families) with WGS data
Syrian ²	22 (10)	37 (14)
Filipino	22 (11)	76 (18) ³
Indian	26 (12)	0
German	38 (19)	0

¹Multiple affected individuals were sequenced from multiplex families.

²Three Syrian individuals from two families (total of six) have both WES and WGS data.

³Seventy affected and six unaffected individuals.

unaffected, five families with five affected individuals, two families with four affected individuals and two unaffected, and one family with eight affected individuals). All of these families were from either Syrian or Filipino populations, and all of the unaffected individuals were Filipino. Table 1 shows the country of origin of these families along with individual and family counts noting the country of origin.

Sequence data generation

Whole exome sequencing

Exome sequencing and genotyping was done at the CIDR. DNA sequencing was performed on an Illumina[®] HiSeq 2500 instrument using standard protocols for a 100-bp paired-end run. Six samples were run per flowcell, guaranteeing >90–95% completeness at a minimum of 20× coverage.

Illumina HiSeq reads were processed through Illumina's Real-Time Analysis (RTA) software generating base calls and corresponding quality scores. Resulting data were aligned to a reference genome with the Burrows-Wheeler Alignment (Li and Durbin 2010) (BWA) tool creating a SAM/BAM file. Postprocessing of the aligned data includes local realignment around indels, base call quality score recalibration performed by the Genome Analysis Tool Kit (GATK) (McKenna et al. 2010; DePristo et al. 2011; Van der Auwera et al. 2013), and flagging of molecular/optical duplicates using software from the Picard program suite. Multisample variant calling was performed using GATK 2.0's Unified Genotyper. Variant quality score recalibration (VQSR) was done in GATK 2.0. CIDR required a minimum mean of 8× coverage before calling any single-nucleotide variant (SNV), but the overall coverage averaged 84× across all exons. Further details of this process are provided in the methods section of Bureau et al. (2014).

Whole genome sequencing

WGS on genomic DNA samples was performed by Illumina, Inc. (San Diego, CA, USA) using TruSeq SBS v3 Reagents, HiSeq Control Software (HCS) and RTA on a HiSeq 2000 machine for real-time image analysis and base calling. Genome assembly, genotype calling, and QC filtering was performed using tools in the CASAVA package. Multisample VCF files were generated using VCFtools (Danecek et al. 2011) and were backfilled with custom scripts to include homozygous reference genotypes and depth of coverage. Full details of the sequencing, alignment, and variant calling process are provided in the supporting information of Mathias et al. (2016).

Sequence data filtering and annotation

WES data

To find potentially causal variants for oral cleft phenotypes in the WES data, we performed an initial filtering step to remove variants of low quality based on specific metrics (described below), and variants in genes with extremely high variation (Table S1) (Schmidt et al. 2013). Called variants were dropped if they failed the following quality metrics: mapping quality <30, depth <8 or depth >20,000, non-Y SNP call rate <98%, replicate errors occurring in >1 pairs (among six duplicate subjects), monomorphic, and failing *both* the GATK VQSR filter and an in-house machine learning metric that combines many of these QC measures to estimate the probability of being a low-quality variant. The variant was dropped if this probability of being low quality exceeded 0.70. We then performed variant-based counting steps at the population level and family level to identify potential risk variants for oral clefts. Here we define population as being one of the four ethnic groups (Syrian, German, Indian, and Filipino).

For both family and population level analyses, we further filtered out common variants (minor allele frequency >5% in 1000 Genomes Phase I data, or in the dbSNP common variant set) and variants with an alternate allele present in either of the two controls (Fig. 1).

For the population-specific analyses, we further screened the rare variants identifying those that were homozygous for the alternate allele in at least one case *and* at least 20% of all cases had the alternate allele in either the homozygous or heterozygous state. This criterion was chosen to allow for some within-population heterogeneity, while enriching for potential recessive candidate variants by requiring at least one individual be homozygous for the rare allele. Since many of the families

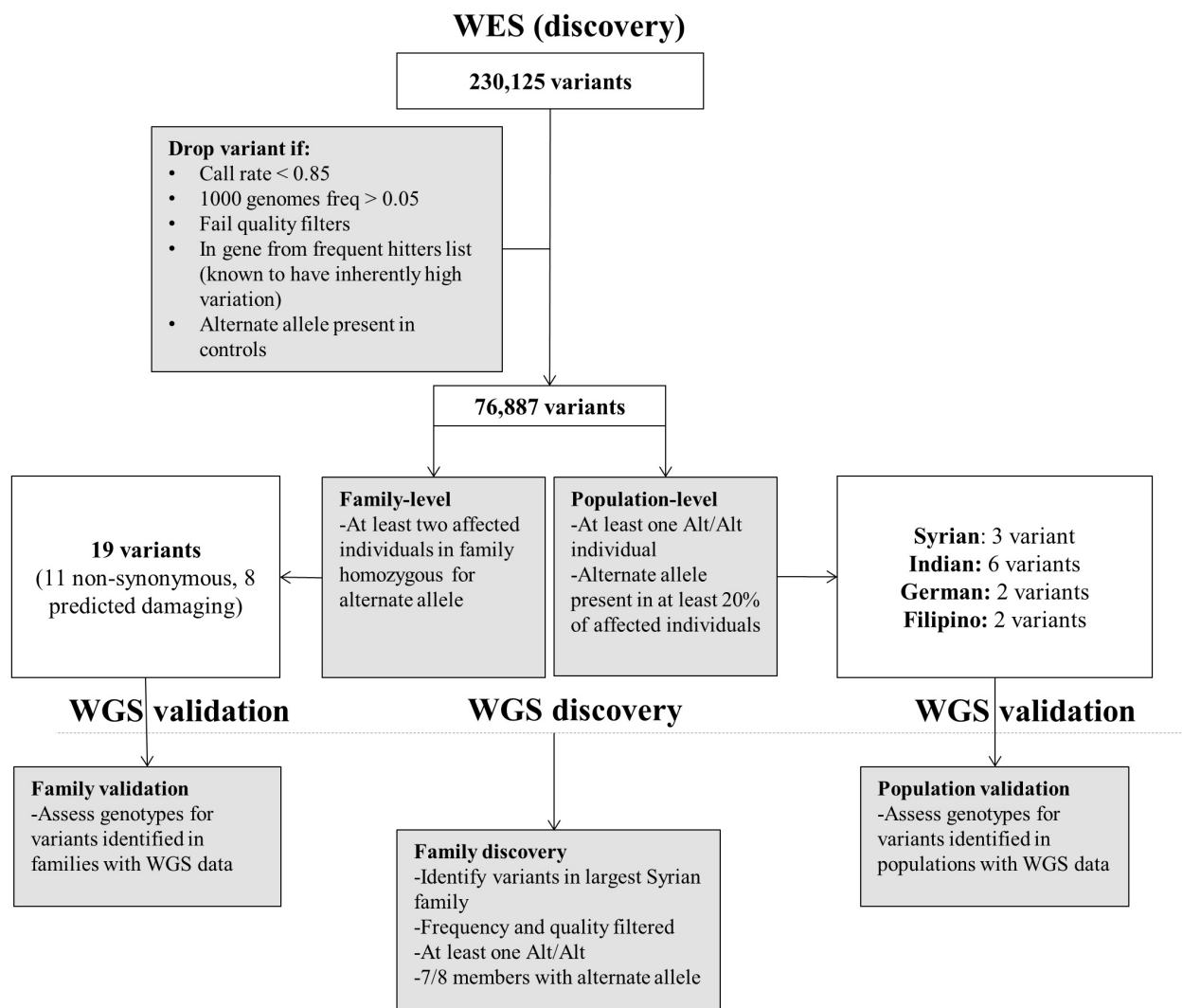


Figure 1. Flowchart showing single variant analysis steps for WES and WGS data.

included in this study exhibit at least some degree of consanguinity, this was deemed a reasonable criterion.

For the family-specific analysis, we identified variants that were homozygous for the alternate allele in at least two affected members of any given family. This is a relatively stringent criterion since these families had sequence data available on only two or three affected individuals. We purposely chose more stringent criteria for the family-specific analysis because it is more likely that the risk variant is the same within the consanguineous families that made up a large proportion of the studied families as opposed to within a population. We selected homozygous variants to enrich for potentially causal recessive candidate variants since even distantly related individuals within the same family share

a large number of variants by chance. Furthermore, there is a higher level of observed variant sharing than would be expected (most likely due to extensive consanguinity) in some of these families. Two families contained only one affected individual due to sequencing failure of one member and were dropped from the family-specific analysis.

We also incorporated annotation information from wAnnoVar (Wang *et al.* 2010; Chang and Wang 2012; Yang and Wang 2015) to assess our set of interesting genes based on potential pathogenicity, along with gene location (exonic or intronic), and predicted variant function (nonsynonymous or synonymous). Nine sources were used from wAnnoVar to assess the potential pathogenicity of each variant: SIFT, Polyphen2 HDIV,

Polyphen2 HVAR, LRT, Mutation Taster, Mutation Assessor, FATHMM, Radial SVM, and LR.

WGS data

In the WGS data, we performed both validation and discovery analyses (Fig. 1). There was very little overlap between individuals in the WES and WGS datasets (only six individuals from two families; Table 1). For our population-specific validation analyses, we assessed genotype counts in the WGS data for variants and genes identified in the WES analyses. Validation in the population-specific analyses was limited to the Syrian and Filipino families. Validation in the family-specific analyses was limited to the two Syrian families that had both WES and WGS data.

As a follow-up to the results from the single variant validation analysis for Syrian Family 1, we further analyzed the WES data to identify genes with potential compound heterozygous individuals (i.e., having two or more rare variants in the same gene). We specifically identified genes where all three affected individuals had more than one exonic, nonsynonymous variant in the WES data. We validated these results in the WGS data by assessing genotypes for all exonic, nonsynonymous variants in the genes identified in the WES analysis. We define validated genes as those with at least one WGS member who was a potential compound heterozygote.

Discovery in the WGS data was limited to the largest Syrian family (Family 1) with sequence data available on eight affected individuals. There are a total of 22 affected individuals in this large, highly consanguineous family. Almost all parents of affected individuals are consanguineous, often related through multiple paths. The eight individuals were chosen for sequencing such that they formed the most distantly related relative pairs to try to limit chance identity by descent allele sharing. After performing the same genotype quality and frequency filtering steps completed for the WES data, we identified variants with at least one family member who was homozygous for the alternate allele and at least seven of the eight family members carrying the alternate allele (this allowed for some within-family heterogeneity, as the results from the WES analysis suggested this might be present). To reduce the number of potential risk variants to those that were more likely to be functional, we only considered exonic, nonsynonymous variants. We also performed a follow-up analysis in the WGS data to assess rare variants in known enhancer and promoter regions of identified candidate genes.

We used the SNP and Variation Suite v8.3.4 (Golden Helix, Inc., Bozeman, MT, www.goldenhelix.com) to perform filtering, counting, annotation, and validation steps in both the WES and WGS datasets (Bozeman 2016).

Results

WES single variant discovery analysis

After performing the filtering steps shown in Figure 1, we used wAnnotator to assess the potential function of the most interesting variants. We also included allele frequency information from a new databases created by the Greater Middle East (GME) Variome project (Scott et al. 2016) and the Qatar Genome (Fakhro et al. 2016) for a better assessment of variants found in the Syrian population. Table 2 shows the variants identified in the family-based analyses, along with annotation information. Tables S2–S5 show the results for the population-specific analyses. We excluded all variants in the HLA region, as they have been shown to be highly population-specific (Sanchez-Mazas and Meyer 2014).

WGS single variant validation analysis

We performed several validation analyses for our most interesting findings from the WES analysis in the WGS data. For the population-specific analyses, we were able to perform validation in the Syrian and Filipino populations. All of the population-specific discovery and validation analysis results are shown in Tables S2–S5. In the Syrian population, all of the WGS individuals were homozygous for the reference allele for the two variants identified in the *CTSL3P* gene. The variant identified in the *SYT17* gene was not present in any other subjects (Table S2). For the individuals from the Filipino population, neither of the two variants identified in the WES analysis were present in the WGS data, most likely due to low quality (Table S4).

For the family-specific variants, we were able to perform validation in two of the Syrian families (1 and 3). As previously stated, there was little overlap between the WES and WGS individuals (Table 1), and we removed any overlapping individuals from this validation analysis.

For the eight affected individuals with WGS data in Syrian Family 1, we confirmed the homozygous genotypes observed in the three WES individuals for the nonsynonymous *CASP9* variant shown in Table 2. For the five additional affected individuals in this family, one was homozygous for the alternate allele, three were heterozygous, and one was homozygous for the reference allele (Table 3).

For Syrian Family 3, there were five individuals with WGS data, three of which were also in the WES analysis. We could not confirm the genotype for the novel variant

Table 2. Variants passing family-specific analysis filter in WES data for all cohorts.

Pop.	Fam. ID	Gene	Chr.	BP	A	R	AA	AR	RR	Location	Func.	1000G Freq.	Predicted damaging
Syrian	1	<i>CASP9</i>	1	15831171	C	T	3	0	0	Exonic	NS	–	2
	3	<i>HCN2</i>	19	603971	G	A	2	1	0	Intronic	–	0.006	–
	6	<i>CHRNA4</i>	2	233408449	T	A	2	0	0	Intronic	–	0.0012	–
	7	<i>SLC24A4</i>	14	92792313	G	A	2	0	0	Exonic	NS	0.0004	2
	7	<i>LGMIN</i>	14	93179134	T	C	2	0	0	Intronic	–	0.0002	–
	7	<i>SERPINA6</i>	14	94776036	A	G	2	0	0	Intronic	–	0.0002	–
	7	<i>HHIPL1</i>	14	100126748	A	G	2	0	0	Intronic	–	0.0006	–
	9	<i>PTGDR</i>	14	52734696	A	G	2	0	0	Exonic	NS	0.0006	0
	10	<i>FCHO1</i>	19	17889669	A	G	2	0	0	Exonic	NS	–	3
	10	<i>SUMO3</i>	21	46228597	T	C	2	0	0	Intronic	–	–	–
German	7	<i>FTCD</i>	21	47572892	G	A	2	0	0	Exonic	NS	–	1
	7	<i>MYO16</i>	13	109792825	T	C	2	0	0	Exonic	NS	0.0058	0
	7	<i>PRCD</i>	17	74534592	C	A	2	0	0	Upstream	–	0.0002	–
	10	<i>PALM</i>	19	740436	A	G	2	0	0	Exonic	NS	0.0018	0
	12	<i>CHAC1</i>	15	41245692	G	A	2	0	0	Exonic	NS	0.0008	1
Indian	20	<i>HMHA1</i>	19	1081558	A	G	2	0	0	Exonic	NS	–	7
	60	<i>DGKQ</i>	4	967071	A	G	2	0	0	Exonic	NS	0.027	5
Filipino	8	<i>TNK2</i>	3	195595358	T	A	2	0	0	Exonic	NS	0.0004	4
	10	<i>HLA-DPA2</i>	6	33059894	G	A	2	0	0	Intergenic	–	0.013	–

We show the family ID for each population (Fam. ID). For each variant we give the gene name, chromosome (Chr.), base pair (BP), alternate allele (A), reference allele (R), number of individuals homozygous for the alternate allele (AA), number of heterozygous individuals (AR), number of individuals homozygous for the reference allele (RR), the gene location (Location), the function of the variant if it is exonic (NS, nonsynonymous; S, synonymous), the frequency of the alternate allele for all populations in 1000 Genomes (1000G Freq.), the frequency from the Greater Middle East Variome Project (GME Freq.), and the number of sources that predict the base pair change to be damaging out of the nine present in wAnnovar. The dashes (–) represent the following: Func. column: variants in non-exonic regions with no defined function; 1000 Freq column: Not present in 1000 Genomes; Predicted damaging column: No pathogenicity predicted.

Table 3. Nonsynonymous and potentially damaging variants from Syrian Family 1 in WGS data.

Gene	Chr.	BP	A	R	AA	AR	RR	Loc.	Func.	1000G Freq.	GME Freq.	QG Freq.	Predicted damaging
<i>CASP9</i>	1	15831171	C	T	4	3	1	Exonic	NS	–	–	–	2
<i>FAT4</i>	4	126367606	T	G	2	5	1	Exonic	NS	0.003	0.006	0.002	3
<i>FAT4</i>	4	126336105	G	A	2	5	1	Exonic	NS	0.002	0.007	0.002	0
<i>FAT4</i>	4	126400922	T	C	2	5	1	Exonic	NS	0.004	0.006	–	0

For each variant we give the gene name, chromosome (Chr.), base pair (BP), alternate allele (A), reference allele (R), number of individuals homozygous for the alternate allele (AA), number of heterozygous individuals (AR), number of individuals homozygous for the reference allele (RR), the gene location (Location), the function of the variant (NS, nonsynonymous; S, synonymous), the frequency of the alternate allele for all populations in 1000 Genomes (1000G Freq.), the frequency from the Greater Middle East Variome Project (GME Freq.), the frequency from the Qatar Genome data (QG Freq.), and the number of sources that predict the base pair change to be damaging out of the nine present in wAnnovar. The dashes (–) represent variants that were not present in the specific frequency database.

in *HCN2* for two of the individuals from the WES analysis, because they were not included in the whole genome sequencing project. For the one WES individual who also had WGS data, the homozygous genotype for the alternate allele was confirmed. For the two individuals who were not in the WES analysis, one was heterozygous and one was homozygous for the reference allele. Thus, strong validation for this variant identified in Syrian Family 3 was not achieved due to the low number of individuals with WGS data.

Compound heterozygous analysis (WES discovery and WGS validation)

To determine if individuals in Syrian Family 1 had any other variants potentially contributing to risk for oral clefts, we searched for genes with more than one heterozygous variant in the WES data. Such an observation may indicate that the affected individuals have two distinct deleterious alleles at the same locus. Because we do not have phase information on these

individuals, we cannot rule out the possibility that the alternate alleles were inherited from the same parent as a rare haplotype. We still deem this as interesting as it could identify compound heterozygotes or the rare haplotype could itself confer increased risk for oral clefts. We defined one gene as possibly having compound heterozygotes when all three WES individuals had more than one exonic, nonsynonymous variant in that particular gene.

While there were no other exonic and/or predicted deleterious rare variants identified in *CASP9*, we did identify four genes meeting our definition of compound heterozygosity (two variants in *COL7A1*, two variants in *CELSR3*, two variants in *TKT*, and three variants in *NLRP14*) (Table 4). We then performed a validation analysis of these results using the WGS data by assessing the genotypes for all exonic, nonsynonymous variants present in these four genes. No other exonic, nonsynonymous variants were found in these genes in any of the eight family members with both WES and WGS data. We were able to confirm the heterozygous genotypes for individuals with WES data in the WGS data. Two of the five additional WGS individuals were compound heterozygous for two identified genes: *COL7A1* and *TKT* (Table 4).

WGS discovery and follow-up analyses

We performed a discovery analysis for the eight individuals with WGS data in Syrian Family 1 to search for potentially damaging variants that may have been missed in the WES analysis. First, we performed the same quality and allele frequency filtering steps as in the WES discovery analysis. We then performed genotype filtering requiring at least one family member to be homozygous for the alternate allele, and for the alternate allele to be present

in at least seven of the eight other family members. We did not require that all eight of the individuals carry the alternate allele given the complex and distant relatedness patterns in this family. We further filtered variants based on annotation from wAnnovar. Specifically, we selected only rare, exonic, nonsynonymous variants. Using these steps, we identified one gene, *FAT4*, with three exonic, nonsynonymous variants, one of which was predicted to be damaging by three different sources in wAnnovar (Table 3). Genotypes for the eight total variants passing our filtering steps in this WGS validation and discovery results (all from Syrian Family 1) are listed in Table 5.

We also performed a follow-up analyses to determine if there were any rare variants in eight known enhancer regions and two promoter regions of *CASP9* (Table S6). Enhancers were selected from the GeneHancer database based on their gene enhancer score (>5) (Fishilevich et al. 2017). We assessed promoters that were <200 kb from the transcription start site of *CASP9* (Stelzer et al. 2016). Table S7 shows the results from this analysis. We did identify several heterozygous variants in these potential *CASP9* regulatory regions. Interestingly, the individual with the most heterozygous variants in these regions was homozygous for the reference allele. Furthermore, no regulatory region variants were identified in the affected individuals that were homozygous for the alternate allele of the nonsynonymous, exonic *CASP9* variant. This may indicate within-family allelic heterogeneity.

Discussion

In this study, our goal was to identify rare potential risk variants shared by distantly related individuals with oral clefts. To do this, we performed genotype and annotation filtering steps to find genes with variants that are not

Table 4. Variants identified in the compound heterozygous analysis in Syrian Family 1 in the WES data.

Gene	Chr.	BP	A	R	AA	AR	RR	Location	Function	1000G Freq.	GME Freq.	QG Freq.	Predicted damaging
<i>COL7A1</i>	3	48602623	A	G	0	3	0	Exonic	NS	0.001	0.005	0.003	5
<i>COL7A1</i>	3	48620046	A	G	0	3	0	Exonic	NS	0.001	0.005	0.002	7
<i>CELSR3</i>	3	48677114	G	C	0	3	0	Exonic	NS	0.019	0.011	0.013	4
<i>CELSR3</i>	3	48691197	T	C	0	3	0	Exonic	NS	0.005	0.005	0.005	0
<i>TKT</i>	3	53267183	T	C	0	3	0	Exonic	NS	0.002	0.011	0.010	3
<i>TKT</i>	3	53269028	T	G	0	3	0	Exonic	NS	0.002	0.011	0.010	1
<i>NLRP14</i>	11	7060948	T	C	0	3	0	Exonic	NS	0.014	0.040	0.046	0
<i>NLRP14</i>	11	7083610	A	T	0	3	0	Exonic	NS	0.015	0.039	0.048	5
<i>NLRP14</i>	11	7083620	C	T	0	3	0	Exonic	NS	0.022	0.043	0.051	0

For each variant we give the gene name, chromosome (Chr.), base pair (BP), alternate allele (A), reference allele (R), number of individuals homozygous for the alternate allele (AA), number of heterozygous individuals (AR), number of individuals homozygous for the reference allele (RR), the gene location (Location), the function of the variant (NS, nonsynonymous; S, synonymous), the frequency of the alternate allele for all populations in 1000 Genomes (1000G Freq.), the frequency from the Greater Middle East Variome Project (GME Freq.), and the number of sources that predict the base pair change to be damaging out of the nine present in wAnnovar.

Table 5. Genotypes for the eight individuals with WGS data in Syrian Family 1 for the WES and WGS validation and discovery results.

Ind. ID	Phen.	Single variant analyses				Compound Het. analyses			
		1:15831171 (CASP9)	4:126367606 (FAT4)	4:126336105 (FAT4)	4:126400922 (FAT4)	3:48602623 (COL7A1)	3:48620046 (COL7A1)	3:53267183 (TKT)	3:53269028 (TKT)
1 (111)*	L.CL	AA	AR	AR	AR	AR	AR	AR	AR
2 (118)*	B.CL, M.CP	AA	RR	RR	RR	AR	AR	AR	AR
3 (125)*	R.CL	AA	AR	AR	AR	AR	AR	AR	AR
4 (38)	L.CL	AA	AR	AR	AR	AR	AR	AR	AR
5 (114)	B.CL	AR	AR	AR	AR	RR	RR	RR	RR
6 (129)	L.CP-I	RR	AA	AA	AA	AR	AR	AR	AR
7 (150)	R.CL	AR	AA	AA	AA	RR	RR	RR	RR
8 (157)	L.CL, M.CP	AR	AR	AR	AR	RR	RR	RR	RR

The first three individuals (111, 118 and 125) have WES and WGS data, as indicated by the asterisk. We identify each variant using *Chromosome: Base Pair* along with the gene name in parentheses. Homozygous for the alternate allele = AA, heterozygous = AR, homozygous for the reference allele = RR. We also show the specific cleft phenotypes for each individual (Phen. column), where L. = left, R. = right, B. = bilateral, M. = midline, I = incomplete, CL = cleft lip, and CP = cleft palate.

common in population databases such as 1000 Genomes, but for which affected individuals were enriched at both the population and family levels. We had the most power to do this in a single Syrian family, because it had the highest number of affected individuals with sequence data and the greatest overlap between WES and WGS data. Our results reflected this, as we were able to detect eight exonic, nonsynonymous variants (four passing the single variant analysis filter and four passing the compound heterozygous filter) that show extensive sharing among these affected relatives from this highly consanguineous family. Thus, one or more of these variants may contribute to risk for oral clefts in this family.

Our most intriguing finding was a nonsynonymous variant in *CASP9* that occurred in seven of the eight family members (four of whom were homozygous for the alternate allele). We identified this as our most interesting finding for several reasons. First, this allele is not present in any of the other Syrian families, 1000 Genomes, the Qatar Genome Database, or the GME Variome project database. Second, it is a nonsynonymous, exonic variant predicted to be pathogenic by two different sources in wAnnoVar. Finally, there is strong evidence that apoptotic genes play a role in the etiology of oral clefts. Among many other aspects of embryonic development, apoptosis plays a crucial role in craniofacial development. Failure of apoptosis during development may result in oral clefts (Smane *et al.* 2013). While no other studies to date have identified variants in *CASP9*, it is directly involved in an apoptotic signaling pathway shown to result in a facial cleft phenotype in mouse models (D'Amelio *et al.* 2010).

Three of the other genes identified in this study contain variants known to cause severe Mendelian syndromes

(*FAT4* in Van Maldergem syndrome [Alders *et al.* 2014, p. 4], *COL7A1* in recessive dystrophic epidermolysis bullosa [Hovnanian *et al.* 1997], and *TKT* in a syndrome which includes short stature, developmental delay, and congenital heart disease [Boyle *et al.* 2016]). While none of these syndromes include oral clefts as a key phenotype, variants in *FAT4* leading to Van Maldergem syndrome can present with craniofacial abnormalities (Alders *et al.* 2014, p. 4). Interestingly, the one individual from Syrian Family 1 who was homozygous for the reference allele for the variant in *CASP9* was homozygous for the alternate allele at all three of the *FAT4* variants and, further, had more than one variant predicted to be pathogenic in both *COL7A1* and *TKT*. This same individual also had the highest number of rare variants (four) in enhancer and promoter regions of *CASP9*. It is also important to note that this individual had a phenotype (incomplete cleft palate) that was distinct from the other seven family members (all others had cleft lip with or without cleft palate) (Table 5). Together, this may represent within-family heterogeneity for genetic factors contributing to risk to nonsyndromic oral clefts.

One notable limitation of our study is there were no unaffected family members with WES data nor were there any sequenced unaffected individuals in the Syrian WGS data. This was a major limitation for our population-based analyses, as many of these populations are not well represented in available allele frequency databases (e.g., 1000 Genomes and ExAC). Therefore, any of our interesting findings at the population-level may be population-specific and not truly phenotype specific. While this is also a limitation in our family-based analyses, we have partly addressed this by limiting our results to those with a higher likelihood of being functional based on multiple

annotation information (exonic, nonsynonymous, with some evidence of being pathogenic).

In this analysis we have identified rare and potentially damaging variants shared by affected family members in a single Syrian family. While these candidate genes and variants need to be assessed further in the unaffected and other affected members of this family, none have been previously identified as risk factors for nonsyndromic oral clefts and may indicate novel genetic underpinnings for this phenotype.

Acknowledgments

We thank the members of families who participated in this study, and the field and laboratory staff who made this analysis possible. This work was supported by the National Institutes of Health (R01-DE-014581, U01-DE-018993, and U01-DE020073 to T. H. B.; R01-DE-016148 and R01-DE-009886 to M. L. M.; P50-DE-016215 and R37-DE-08559 to J. C. M.). Recruitment of Syrian families was supported by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health, USA, and the Ibn Al-Nafees Hospital, Syrian Arab Republic. Additional support from X01 HG006177 to T. H. B., M. L. M., and J. C. M. for whole exome sequencing at the Center for Inherited Disease Research, which is funded through a federal contract from the NIH to Johns Hopkins University (contract no. HHSN268200782096C). This study was funded, in part, by the Intramural Research Program of the National Human Genome Research Institute of the National Institutes of Health. E. R. H. is funded by the PRAT Program of the National Institute for General Medical Sciences of the National Institutes of Health.

Conflict of Interest

There are no conflicts of interest to disclose.

References

- Alders, M., L. Al-Gazali, I. Cordeiro, B. Dallapiccola, L. Garavelli, B. Tuysuz, et al. 2014. Hennekam syndrome can be caused by FAT4 mutations and be allelic to Van Maldergem syndrome. *Hum. Genet.* 133:1161–1167.
- Beaty, T. H., M. L. Marazita, and E. J. Leslie. 2016. Genetic factors influencing risk to orofacial clefts: today's challenges and tomorrow's opportunities. *F1000Research* 5:2800.
- Boyle, L., M. M. C. Wamelink, G. S. Salomons, B. Roos, A. Pop, A. Dauber, et al. 2016. Mutations in TKT are the cause of a syndrome including short stature, developmental delay, and congenital heart defects. *Am. J. Hum. Genet.* 98:1235–1242.
- Bozeman, M. T. 2016. SNP & Variation Suite. Golden Helix, Inc.
- Bureau, A., M. M. Parker, I. Ruczinski, M. A. Taub, M. L. Marazita, J. C. Murray, 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.
- Chang, X., and K. Wang. 2012. wANNOVAR: annotating genetic variants for personal genomes via the web. *J. Med. Genet.* 49:433–436.
- D'Amelio, M., V. Cavallucci, and F. Cecconi. 2010. Neuronal caspase-3 signaling: not only cell death. *Cell Death Differ.* 17:1104–1114.
- Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, et al. 2011. Genomes Project Analysis Group. 2011. The variant call format and VCFtools. *Bioinformatics* 27:2156–2158.
- DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 43:491–498.
- Fakhro, K. A., M. R. Staudt, M. D. Ramstetter, A. Robay, J. A. Malek, R. Badii, et al. 2016. The Qatar genome: a population-specific tool for precision medicine in the Middle East. *Hum. Genome Var.* 3:16016.
- Field, L. L., A. K. Ray, M. E. Cooper, T. Goldstein, D. F. Shaw, and M. L. Marazita. 2004. Genome scan for loci involved in nonsyndromic cleft lip with or without cleft palate in families from West Bengal, India. *Am. J. Med. Genet. A* 130A:265–271.
- Fishilevich, S., R. Nudel, N. Rappaport, R. Hadar, I. Plaschkes, T. Iny Stein, et al. 2017. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database 2017.
- Hovnanian, A., A. Rochat, C. Bodemer, E. Petit, C. A. Rivers, C. Prost, et al. 1997. Characterization of 18 new mutations in COL7A1 in recessive dystrophic epidermolysis bullosa provides evidence for distinct molecular mechanisms underlying defective anchoring fibril formation. *Am. J. Hum. Genet.* 61:599–610.
- Li, H., and R. Durbin. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26:589–595.
- Mangold, E., H. Reutter, S. Birnbaum, M. Walier, M. Mattheisen, H. Henschke, et al. 2009. Genome-wide linkage scan of nonsyndromic orofacial clefting in 91 families of central European origin. *Am. J. Med. Genet. A* 149A:2680–2694.
- Marazita, M. L., J. C. Murray, A. C. Lidral, M. Arcos-Burgos, M. E. Cooper, T. Goldstein, et al. 2004. Meta-analysis of 13 genome scans reveals multiple cleft lip/palate genes with novel loci on 9q21 and 2q32-35. *Am. J. Hum. Genet.* 75:161–173.

- Marazita, M. L., A. C. Lidral, J. C. Murray, L. L. Field, B. S. Maher, T. Goldstein McHenry, et al. 2009. Genome scan, fine-mapping, and candidate gene analysis of non-syndromic cleft lip with or without cleft palate reveals phenotype-specific differences in linkage and association results. *Hum. Hered.* 68:151–170.
- Mathias, R. A., M. A. Taub, C. R. Gignoux, W. Fu, S. Musharoff, T. D. O'Connor, et al. 2016. A continuum of admixture in the Western Hemisphere revealed by the African Diaspora genome. *Nat. Commun.* 7:12522.
- McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernysky, et al. 2010. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20:1297–1303.
- Riley, B. M., R. E. Schultz, M. E. Cooper, T. Goldstein-McHenry, S. Daack-Hirsch, K. T. Lee, et al. 2007. A genome-wide linkage scan for cleft lip and cleft palate identifies a novel locus on 8p11-23. *Am. J. Med. Genet. A* 143A:846–852.
- Sanchez-Mazas, A., and D. Meyer. 2014. The relevance of *HLA* sequencing in population genetics studies. *J. Immunol. Res.* 2014:1–12.
- Schmidt, E. E., O. Pelz, S. Buhlmann, G. Kerr, T. Horn, and M. Boutros. 2013. GenomeRNAi: a database for cell-based and in vivo RNAi phenotypes, 2013 update. *Nucleic Acids Res.* 41:D1021–D1026.
- Schultz, R. E., M. E. Cooper, S. Daack-Hirsch, M. Shi, B. Nepomucena, K. A. Graf, et al. 2004. Targeted scan of fifteen regions for nonsyndromic cleft lip and palate in Filipino families. *Am. J. Med. Genet. A* 125A:17–22.
- Scott, E. M., A. Halees, Y. Itan, E. G. Spencer, Y. He, M. A. Azab, et al. 2016. Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. *Nat. Genet.* 48:1071–1076.
- Smame, L., M. Pilmane, and I. Akota. 2013. Apoptosis and MMP-2, TIMP-2 expression in cleft lip and palate. *Stomatologija* 15:129–134.
- Stelzer, G., N. Rosen, I. Plaschkes, S. Zimmerman, M. Twik, S. Fishilevich, et al. 2016. The GeneCards suite: from gene data mining to disease genome sequence analyses: the GeneCards suite. Pp 1.30.1–1.30.33 in A. Bateman, W. R. Pearson, L. D. Stein, G. D. Stormo and J. R. Yates, eds. *Current protocols in bioinformatics*. John Wiley & Sons, Inc., Hoboken, NJ, USA.
- Van der Auwera, G. A., M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel, A. Levy-Moonshine, et al. 2013. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline: the genome analysis toolkit best practices pipeline. Pp 11.10.1–11.10.33 in A. Bateman, W. R. Pearson, L. D. Stein, G. D. Stormo and J. R. Yates, eds. *Current protocols in bioinformatics*. John Wiley & Sons, Inc., Hoboken, NJ, USA.
- Wang, K., M. Li, and H. Hakonarson. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38:e164.
- Wyszynski, D. F., H. Albacha-Hejazi, M. Aldirani, M. Hammod, H. Shkair, A. Karam, 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.
- Yang, H., and K. Wang. 2015. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat. Protoc.* 10:1556–1566.

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

Table S1. Frequent hitter gene names.

Table S2. Genes with variants that passed population-specific analysis filter in WES data in the Syrian cohort with WGS validation.

Table S3. Genes with variants that passed population-specific analysis filter in WES data in the Indian cohort.

Table S4. Genes with variants that passed population-specific analysis filter in WES data in the Filipino cohort with WGS validation.

Table S5. Genes with variants that passed population-specific analysis filter in WES data in the German cohort.

Table S6. The IDs and genomic regions for the eight enhancer regions and two promoter regions for *CASP9*.

Table S7. Genotypes for the eight individuals with WGS data in Syrian Family 1 for the variants in enhancer and promoter regions of *CASP9*.