Luciano Abreu Brito

Variantes genéticas de risco às fissuras
orofaciais

Genetic risk variants for orofacial clefts

São Paulo

2016

Luciano Abreu Brito

Variantes genéticas de risco às fissuras
orofaciais

Genetic risk variants for orofacial clefts

Tese apresentada ao Instituto de
Biociências da Universidade de São
Paulo, para a obtenção de Título de
Doutor em Ciências, na Área de
Biologia/Genética.

Orientadora: Prof^ª. Dra. Maria Rita
dos Santos e Passos-Bueno

São Paulo

2016

Ficha Catalográfica

Brito, Luciano Abreu
Variantes genéticas de risco às
fissuras orofaciais
164 páginas

Tese (Doutorado) - Instituto de
Biociências da Universidade de São Paulo.
Departamento de Genética e Biologia
Evolutiva.

1. Fissuras labiopalatinas 2.
Sequenciamento de Exoma 3. CDH1
Universidade de São Paulo. Instituto de
Biociências. Departamento de Genética e
Biologia Evolutiva.

Comissão Julgadora:

Prof(a). Dr(a).

Prof(a). Dr(a).

Prof(a). Dr(a).

Prof(a). Dr(a).

Prof^ª. Dra. Maria Rita S. Passos-Bueno
orientadora

**A todos os pacientes com os quais
tive contato ao longo deste projeto.**

*Education is when you read the fine print;
experience is what you get when you don't.*

Pete Seeger

Agradecimentos

À minha família, em especial a meus pais e meu irmão, sem o apoio dos quais esta curta carreira já nem teria começado.

À Rita, pelo acolhimento, orientação, dedicação e disponibilidade durante todos esses anos.

Aos amigos do laboratório, que contribuíram para criar um ambiente de trabalho extremamente agradável: Gerson, Carol, Roberto, Lucas, Felipe, Van, Karina, May, Dani M, Bela, Bruno, Atique, Erika K, Joanna, Suzana, Tati, Dani B, Dani Y, Clarice, Ágatha, Camila M, Camila L, Lucas “Jr”, Gabi “Jra”, Cibele, Naila, Simone e Andressa.

A todos os organizadores e voluntários da Operação Sorriso, que tornaram este trabalho viável, e propiciaram momentos muito importantes de crescimento pessoal.

À equipe do Genoma, em especial ao pessoal do sequenciamento: Meire, Vanessa, Guilherme, Monize e, muito especialmente, Kátia, quem me ensinou o básico nos meus primeiros meses de laboratório.. Também aos demais colegas de departamento, em especial Vanessa S, Elaine, Michel, Natássia, Inês e Toninha.

To Eric, Christina, Antoine, Kushi, Renée, Mike, Yawei, Irving, François, Jullian, Vanessa, Maura, Chris, Leo, Sebastian, Takuya, Mary and Andy, who made my stay in Boston way easier and warmer.

Este trabalho contou com o apoio financeiro da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e do Ministério da Ciência, Tecnologia e Inovação do Brasil.

Notas

Esta tese de doutorado compreende um trabalho desenvolvido durante os anos de 2012 a 2015 no Laboratório de Genética do Desenvolvimento do Centro de Estudos do Genoma Humano e Células Tronco, Instituto de Biociências, Universidade de São Paulo.

A tese foi redigida no modelo de artigos e capítulos, no idioma inglês. Cinco artigos foram incluídos no corpo principal da tese. Publicações em co-autoria e não relacionadas ao tema principal da tese encontram-se sumarizadas nos Apêndices, ao final da tese.

O projeto que resultou na presente tese foi cadastrado na Plataforma Brasil e contou com o parecer consubstanciado do Comitê de Ética em Pesquisa do Instituto de Biociências da Universidade de São Paulo (Número 363.876/2013).

List of Abbreviations

<i>1kGP</i>	1000 Genomes Project	<i>HGDC</i>	Hereditary diffuse gastric cancer
<i>6500ESP</i>	Exome Variant Server database	<i>JPT</i>	Japanese in Tokyo, Japan
<i>AIM</i>	Ancestry informative marker	<i>LD</i>	Linkage disequilibrium
<i>CEGH60+</i>	Centro de Estudos do Genoma Humano database	<i>LoF</i>	Loss of function
<i>CDCV</i>	Common disease-common variant	<i>MAF</i>	Minor allele frequency
<i>CDRV</i>	Common disease-rare variant	<i>MNE</i>	Medionasal enhancer region
<i>CEU</i>	Central Europeans from Utah	<i>NCC</i>	Neural crest cells
<i>CHB</i>	Han Chinese in Beijing, China	<i>NGS</i>	Next-generation sequencing
<i>CI</i>	Confidence interval	<i>NHEJ</i>	Non-homologous end joining
<i>CLO</i>	Cleft lip only	<i>NSCL/P</i>	Nonsyndromic cleft lip with or without cleft palate
<i>CLP</i>	Cleft lip and palate	<i>NSCPO</i>	Nonsyndromic cleft palate only
<i>CL/P</i>	Cleft lip with or without cleft palate	<i>NSOFC</i>	Nonsyndromic orofacial clefts
<i>CPO</i>	Cleft palate only	<i>OFC</i>	Orofacial clefts
<i>dpf</i>	Days post fertilization	<i>OOM</i>	<i>Orbicularis oris</i> muscle
<i>DSB</i>	Double-strand break	<i>OOMMSC</i>	<i>Orbicularis oris</i> muscle mesenchymal stem cell
<i>EMT</i>	Epithelial-mesenchymal transition	<i>OR</i>	Odds ratio
<i>ExAC</i>	Exome Aggregation Consortium	<i>PCP</i>	Planar cell polarity
<i>eQTL</i>	Expression quantitative trait locus	<i>RR</i>	Relative risk
<i>FDR</i>	False discovery rate	<i>sgRNA</i>	Single-guide RNA
<i>GTEx</i>	Genotype-tissue expression project	<i>SNP</i>	Single nucleotide polymorphism
<i>GWAS</i>	Genome-wide association studies	<i>SNV</i>	Single nucleotide variant
<i>HDR</i>	Homology-dependent repair	<i>SRC</i>	Spearman's rank correlation
		<i>TS</i>	Target site
		<i>TSS</i>	Transcription start site
		<i>WT</i>	Wild type
		<i>YRI</i>	Yoruba in Ibadan, Nigeria

Table of contents

Chapter 1. General Introduction	13
Orofacial Clefts.....	13
Genetics of NSCL/P: Approaches and Risk Factors.....	18
Objectives.....	23
Chapter 2. <i>Exome analysis in multiplex families reveals novel candidate genes for nonsyndromic cleft lip / palate</i>.....	29
Main text.....	30
Supplementary information.....	48
Chapter 3. <i>Rare variants in the epithelial cadherin gene underlying the genetic etiology of nonsyndromic cleft lip with or without cleft palate</i>.....	61
Main text.....	62
Supplementary information.....	67
Chapter 4. <i>Establishment of cdh1-mutant zebrafish lines through CRISPR/Cas9-mediated genome editing</i>.....	83
Main text.....	84
Supplementary information.....	97
Chapter 5. <i>Association of GWAS loci with nonsyndromic cleft lip and/or palate in Brazilian population</i>.....	99
Main text.....	100
Supplementary information.....	117
Chapter 6. <i>eQTL mapping reveals MRPL53 (2p13) as a candidate gene for nonsyndromic cleft lip and/or palate</i>.....	125
Main text.....	126
Supplementary information.....	143
Chapter 7. General Discussion and Conclusions.....	153
Chapter 8. Abstract.....	157
Appendix: Additional publications.....	159

Chapter 1

General Introduction

1. Orofacial Clefts

1.1. Clinical and Epidemiological Aspects

Orofacial clefts (OFC) are congenital defects that arise from failure during the embryological process of closure of lip and palate, resulting in the cleft of these structures. Cleft lip may be unilateral or bilateral, and either can be restricted to the lip (cleft lip only, CLO) or reach the alveolus (gum) and the pre-incisive foramen palate (cleft lip and palate, CLP; Figure 1A-B). In the most severe cases, palate is affected anteriorly (pre-incisive foramen cleft) and posteriorly (post-incisive foramen cleft), being called complete cleft palate (Figure 1C-D). Cleft palate can also occur without cleft lip (cleft palate only, CPO), and is usually restricted to the posterior palate (Figure 1E; Schutte and Murray, 1999; Gorlin and Cohen Jr., 2001).

OFC constitute the most prevalent group of congenital craniofacial malformations, with a worldwide prevalence estimated as 1:700 liveborn babies (Mossey et al., 2009). Epidemiological findings support the division of OFC in two distinct disorders: cleft lip with or without cleft palate (CL/P) and CPO (Fogh-Andersen, 1942; Fraser, 1955; Gorlin and Cohen Jr., 2001). As it will be discussed in the following section, differences in embryonic development of lip and palate also support this division.

Figure 1 – Most common types of cleft affecting the palate. (A) Unilateral cleft lip with alveolar involvement; (B) Bilateral cleft lip with alveolar involvement; (C) Unilateral cleft lip with complete cleft palate; (D) Bilateral cleft lip with complete cleft palate; (E) Cleft palate only. Adapted from Brito et al. (2012b).

The prevalence of CL/P varies substantially across populations: it is lower in Africans (~0.3:1,000), intermediate in Europeans (~1:1,000) and higher in East Asians (1.4-2.1:1,000) and Amerindians (~3.6:1,000; Vanderas, 1987; Gorlin and Cohen Jr., 2001). In addition, low socioeconomic level is correlated with higher incidences of CL/P (Murray et al., 1997; Xu et al., 2012). On the other hand, the prevalence of CPO, frequently estimated as 1:2,000, does not show ethnic heterogeneity (Gorlin and Cohen Jr., 2001). In European populations, on which most of studies have been conducted, differences are also observed in sex ratio (with CL/P being more frequent in men – 60-80% of cases – and CPO prevailing in women), and in empirical recurrence risks among first-degree relatives (3-4% for CL/P and 2% for CPO; Gorlin and Cohen Jr., 2001).

Based on the presence of additional malformations or comorbidities, OFC can be classified as syndromic or nonsyndromic. Nonsyndromic cases account for 70% of CL/P (nonsyndromic CL/P, NSCL/P) and 50% of CPO (nonsyndromic CPO, NSCPO; Stanier and Moore, 2004; Jugessur et al., 2009). Generally, NSCL/P and NSCPO do not segregate

within a same family, reinforcing the etiological differences between these entities. Nevertheless, co-segregation of both forms may occur in some syndromic forms (Dixon et al., 2011). Up to date, there are more than 500 syndromes that include OFC as part of the phenotype, according to OMIM database (*Online Mendelian Inheritance in Man*).

Individuals affected by OFC often experience difficulties in feeding, which still implies in mortality, specially in developing countries (Carlson et al., 2013). Dental, speech and hearing problems may be also present. Since OFC are readily noticed facial defects, affected individuals habitually face serious adversities in social engagement, leading to a psychological burden (Marazita, 2012). Therefore, the complete rehabilitation of the patient with OFC demands reparative surgeries (multiple, starting from 3 months of age until adult life) coupled with a multidisciplinary treatment. Given the relatively high prevalence of OFC and its costly treatment (estimated as US\$100,000 for a single patient; Centers for Disease Control and Prevention, <http://www.cdc.gov>; Waitzman et al., 1994), these disorders represent an important problem to the health care system. Therefore, understanding the etiological factors and mechanisms that lead to OCF may, ultimately, help to treat and prevent these disorders.

1.2. *Embryology*

The normal closure of lip and palate comprehends a sequence of finely coordinated steps of cell growth, proliferation, migration, differentiation and apoptosis. Any punctual disturbance in the biological processes of this chain of events may perturb the subsequent events, eventually leading to OFC (Leslie and Marazita, 2013).

Lip morphogenesis starts in the 4th week of development, when the neural crest cells (NCC) delaminate from the neural folds and migrate to the developing craniofacial region through the mesenchymal tissue. NCC migration gives rise to the five facial primordia: one frontonasal, one pair of mandibular processes and one pair of maxillary processes (Figure 2A). In the following weeks, the frontonasal prominence originates, at its lower portion, the medial and lateral nasal processes (1 pair each; Figure 2B). Upper lip and primary palate are formed when the maxillary processes touch and fuse with the medial nasal processes, which occurs until the 7th week (Figure 2C; Jiang et al., 2006). Therefore, failure during growth or fusion of these prominences results in cleft of the lip, which may reach the alveolus and primary palate (Figure 1A-D).

The morphogenesis of secondary palate begins in the 6th week, when the maxillary processes originate a pair of palatal shelves, located laterally to the developing tongue (Figure 2D). Initially, the palatal shelves grow vertically and, during the 7th week, they advance horizontally above the tongue, until they contact each other (Figure 2E). The subsequent fusion of the palatal shelves depends on the degeneration of an epithelial seam at the midline (Figure 2F), which is achieved by cell death and epithelial-mesenchymal transition (Mossey et al., 2009; Twigg and Wilkie, 2015), allowing a homogeneous mesenchyme in the palatal tissue, at the 10th week (Kerrigan et al., 2000). At this time, oral and nasal cavities are completely separated, but failures at these events will lead to cleft palate (Figure 1E).

Figure 2 – Lip (A-C) and palate (D-F) embryogenesis. (A) Frontonasal prominence, maxillary processes and mandibular processes surrounding the oral cavity, at 4th week of development. (B) By the 5th week, medial nasal and lateral nasal processes are formed, as well as the nasal pits. (C) At the end of 6th week, medial nasal processes fuse with maxillary processes, giving rise to superior lip and primary palate; lateral nasal processes originate nasal alae, and mandibular processes originate the mandible. (D) By the 6th week, the palatal shelves originate from the maxillary processes, and grow vertically. (E) The elevated palatal shelves grow horizontally at the 7th week, positioned above the tongue, and reach each other. (F) By the 10th week, palatal shelves fuse with each other, following the degeneration of a midline epithelial layer. Adapted from Dixon et al. (2011).

1.3. *Etiology*

Syndromic OFC may arise from gene mutations, chromosomal abnormalities or environmental factors, such as exposures to teratogens during the first trimester of pregnancy. NSCL/P and NSCPO, on the other hand, are complex disorders, with most of cases presenting multifactorial inheritance, where genetic and environmental susceptibility factors may play a role (Dixon et al., 2011).

Several environmental factors have been associated with increased risk of NSCL/P and NSCPO but contradictory findings are often observed. Among these factors, are maternal exposure to tobacco, alcohol consumption, obesity, infection, poor nutrition (and lack of nutrients such as folate, zinc and vitamins in general) and teratogens (as valproic acid; Mossey et al., 2009; Dixon et al., 2011).

Evidence for a strong genetic role for NSCPO susceptibility has been obtained from studies on heritability and recurrence risk (Mitchell and Christensen, 1996; Nordstrom et al., 1996). However, probably due to the lower prevalence of NSCPO, most of studies have focused on NSCL/P, which is also our main interest.

The genetic contribution to NSCL/P has been evidenced by heritability studies in different populations. Twin studies indicate phenotypic concordance of 40-60% for monozygotic and 3-5% for dizygotic twins from Denmark (Christensen and Fogh-Andersen, 1993; Mitchell et al., 2002); high heritability has also been observed in other European countries (reaching 84% in Italy; Calzolari et al., 1988), China (78%; Hu et al., 1982) and Brazil (reaching 85%; Brito et al., 2011). Another evidence for this genetic role comes from recurrence risk, which is 20-30 times higher in 1st-degree relatives of affected individuals than the population risk (Sivertsen et al., 2008; Grosen et al., 2010). Extensive research has been conducted in order to uncover the genetic basis of NSCL/P, and several susceptibility loci have emerged in the recent years.

2. Genetics of NSCL/P: Approaches and Risk Factors

2.1. Linkage and Candidate Gene Association Studies

A variety of approaches has been used to explore the genetic etiology of NSCL/P. Gene mapping strategies such as linkage and association studies has historically been the most popular. Linkage analysis relies on the co-segregation between genetic markers and the disease in families (Altshuler et al., 2008). Although several loci had been suggested by genome-wide linkage analysis, significant LOD-scores were firstly reached only in a meta-analysis, for the chromosomal regions 1q32, 2p13, 3q27-28, 9q21, 14q21-24 and 16q24 (Marazita et al., 2004).

Association analysis, under case-control or family-based design, was initially applied to candidate genes. Therefore, this approach required previous knowledge about the genes, before including them in the studies (Altshuler et al., 2008). Although many susceptibility loci were suggested by candidate gene studies, the vast majority was nonreplicable across studies (Leslie and Marazita, 2013). A single remarkable exception was *IRF6* (1q32), firstly associated with NSCL/P by Zuccherro et al. (2004), and consistently replicated thenceforth (Jugessur et al., 2008; Rahimov et al., 2008). Moreover, heterozygous loss-of-function mutations in *IRF6* lead to van der Woude syndrome (VWS1, MIM#119300), the most common syndromic form of OFC (Kondo et al., 2002).

2.2. GWAS and the Common Susceptibility Variants

The scenario dramatically changed with the genome-wide association studies (GWAS), which allowed association studies to be performed in genomic level, without bias regarding the need of *a priori* knowledge of candidate genes (Kruglyak, 2008). GWAS relies on the common disease-common variant (CDCV) hypothesis for complex diseases, which predicts that the allelic spectrum of the disease (i.e., all disease-contributing variants) is predominantly composed of frequent variants (originated from a common ancestor and maintained in the population) of low individual effects (Reich and Lander, 2001; Schork et al., 2009). In this manner, these studies were made possible thanks to a deep characterization of the patterns of genetic variation in human

genome, provided by the Human Genome Project (Lander et al., 2001) and the HapMap Project (<http://hapmap.ncbi.nlm.nih.gov>).

Birnbaum et al. (2009) conducted the first GWAS on NSCL/P, and found association of a group of markers in a 640-kb interval at a gene desert in 8q24 region, which was confirmed shortly after by a second GWAS (Grant et al., 2009). The third GWAS came from an expansion of Birnbaum's sample, and implied two new loci (10q25 and 17q22), besides having replicated the associations of *IRF6* and 8q24 (Mangold et al., 2009). Differently from the three previous GWAS, which used case-control design and only populations of European origin, Beaty et al. (2010) carried out a family-based GWAS with a mixed sample of European and Asian individuals. This study reported, for the first time, significant associations of 1p22.1 and 20q12, and suggested that association of these and previously reported loci may vary across populations. A meta-analysis of Mangold's and Beaty's data uncovered new associations, expanding to 12 the number of variants implicated by GWAS (Ludwig et al., 2012). Moreover, it confirmed the 8q24 locus as the strongest association in NSCL/P (Box 1). Recently, Sun et al. (2015) conducted the fifth GWAS on NSCL/P, the first on a totally non-European sample (the Chinese population). A new susceptibility locus was revealed in this study (16p13), reinforcing the importance of testing populations other than Europeans. All loci associated by GWAS are summarized in Table 1.

Several studies have endeavored to replicate these associations in different populations. Not rarely, they failed in detecting association for some loci. As an example, 8q24 association was extensively replicated in European populations, but not in Asians or Africans (see Box 1). A drawback in many replication studies, however, is that they generally focus on testing the top-SNP at each GWAS-associated locus. In consequence, if this SNP lays in a different haplotypic block than in European populations, lack of association will probably be observed (Kruglyak, 2008). In addition, if the top-SNP is rare in a given population, the study's statistical power to detect association will dramatically decrease, as association studies are powered to detect common variant (Murray et al., 2012). Therefore, these possibilities should be considered before assuming non-association of a candidate locus in a new population.

In general, the major NSCL/P susceptibility loci uncovered by GWAS have been shown to increase only a small risk, which, collectively, do not explain a significant proportion of the populational risk to the disease (Leslie and Marazita, 2013). The arising question of where this non-explained genetic risk would be hiding was termed as

“missing heritability”, and it is a common debate for most of complex disorders, (Maher, 2008; Manolio et al., 2009). Nonetheless, if in one hand GWAS have failed to explain a vast component of NSCL/P heritability, on the other, they did provide insights on new pathways involved with the disease (Visscher et al., 2012; Yang et al., 2014).

Table 1 – Genomic loci significantly associated with NSCL/P by GWAS

Region	Top SNP	Main candidate gene(s)	P-value	Risk (95% CI)	Associations in GWAS
1p22.1	rs560426	<i>ARHGAP29</i>	3.1×10^{-12}	RR _{het} =1.42 (1.24–1.62); RR _{hom} =1.86 (1.56–2.23) ^a	Beaty et al., 2010; Ludwig et al., 2012
1p36	rs742071	<i>PAX7</i>	7.0×10^{-9}	RR _{het} =1.316 (1.13–1.54); RR _{hom} =1.878 (1.52–2.32) ^a	Ludwig et al., 2012
1q32.2	rs861020	<i>IRF6</i>	3.2×10^{-12}	RR _{het} =1.44 (1.27–1.64); RR _{hom} =2.04 (1.60–2.60) ^a	Birnbaum et al., 2009; Mangold et al., 2009; Beaty et al., 2010; Ludwig et al., 2012; Sun et al., 2015
2p21	rs7590268	<i>THADA</i>	1.3×10^{-8}	RR _{het} =1.42 (1.23–1.64); RR _{hom} =1.98 (1.47–2.66) ^a	Ludwig et al., 2012
3p11.1	rs7632427	<i>EPHA3</i>	3.9×10^{-8}	RR _{het} =0.73 (0.64–0.83); RR _{hom} =0.61 (0.49–0.76) ^a	Ludwig et al., 2012
8q21.3	rs12543318		1.9×10^{-8}	RR _{het} =1.27 (1.11–1.46); RR _{hom} =1.68 (1.40–2.01) ^a	Ludwig et al., 2012
8q24	rs987525	<i>MYC</i>	5.1×10^{-35}	RR _{het} =1.92 (1.66–2.22); RR _{hom} =4.38 (3.39–5.67) ^a	Birnbaum et al., 2009 Grant et al., 2009 Mangold et al., 2009; Beaty et al., 2010; Ludwig et al., 2012 Sun et al., 2015
10q25	rs7078160	<i>VAX1</i>	4.0×10^{-11}	RR _{het} =1.38 (1.21–1.58); RR _{hom} =1.94 (1.58–2.39) ^a	Mangold et al., 2010 Ludwig et al., 2012 Sun et al., 2015
13q31.1	rs8001641	<i>SPRY2</i>	2.6×10^{-10}	RR _{het} =1.31 (1.13–1.51); RR _{hom} =1.86 (1.54–2.26) ^a	Ludwig et al., 2012
15q22.2	rs1873147	<i>TPM1</i>	7.9×10^{-7}	RR _{het} =1.43 (1.23–1.67); RR _{hom} =1.65 (1.34–2.04) ^a	Ludwig et al., 2012
16p13	rs8049367	<i>CREBBP</i> <i>ADCY9</i>	9.0×10^{-12}	OR _{add} =0.74 (0.68–0.80) ^b	Sun et al., 2015
17p13*	rs4791774	<i>NTN1</i>	5.1×10^{-19}	OR _{add} =1.56 (0.71–0.83) ^b	Sun et al., 2015
17q22	rs227731	<i>NOG</i>	1.8×10^{-8}	RR _{het} =1.23 (1.08–1.40); RR _{hom} =1.67 (1.40–2.0) ^a	Mangold et al., 2010 Ludwig et al., 2012
20q12	rs13041247	<i>MAFB</i>	6.2×10^{-9}	RR _{het} =0.84 (0.74–0.94); RR _{hom} =0.55 (0.45–0.66) ^a	Beaty et al., 2010; Ludwig et al., 2012; Sun et al., 2015

RR: Relative risk; **hom:** homozygous; **het:** heterozygous; **ORadd:** Odds ratio using additive model.

^a Data retrieved from Ludwig et al. (2012).

^b Data retrieved from Sun et al. (2015)

* Also marginally associated in Beaty et al., 2010

BOX1: 8q24 locus

The association of a 640-kb interval at 8q24 represents the most prominent finding of GWAS on NSCL/P. The association of the top-SNP rs987525 has been consistently replicated in populations from Europe (Cura et al., 2015), Central America (Rojas-Martinez et al., 2010), Brazil (Brito et al., 2012c) and Middle-East (Aldhorae et al., 2014). Nonetheless, replication studies have failed in finding association in Asian and African populations (Beaty et al., 2010; Weatherley-White et al., 2011; Figueiredo et al., 2014). At least in Asian populations, this lack of association is thought to be consequence of low statistical power, due to low allele frequency, since larger studies find suggestive signals of association (Murray et al., 2012). In addition, Boehringer et al. (2011) and Liu et al. (2012) have reported association of 8q24 locus with normal variation of human facial traits. Because no known gene is present at this region, a regulatory role has been proposed since its identification (Birnbbaum et al., 2009). In fact, Uslu et al. (2014), studying the syntenic murine locus, found that a 280-kb region within the NSCL/P associated interval is enriched for long-range regulatory elements of the proximal gene *MYC*. In addition, deletions of these elements frequently led to facial dysmorphologies, including cleft lip / palate. At the cellular level, the authors verified that deletion of these regions were correlated with lower *Myc* expression and enriched expression of genes involved with ribosome assembly and transcriptional, suggesting that abnormal cell proliferation is a possible mechanism by which deletion of these elements causes

2.3. *Resequencing Studies and the Rare Variants*

One hypothesis that addresses to the missing heritability question relies on the role of rare variants, typically defined as <1% (Maher, 2008). Alternatively to the CDCV hypothesis, some researchers argue that the major genetic contributors to common diseases would be rare, moderate-to-high effect variants distributed in the population. According to this common disease-rare variant (CDRV) hypothesis, a combination of only few rare, high-effect variants would be necessary to cause the disease in an individual, and most of disease's phenotypic variation and expressivity observed in population would be attributed to different allele combinations, under additional influence of environmental factors (Bodmer and Bonilla, 2008; Schork et al., 2009; Gibson, 2012).

Sequencing strategies have been the most suitable approach to detect rare variants implicated with diseases (Manolio et al., 2009). Resequencing of genes associated with NSCL/P by GWAS has found possibly pathogenic rare variants in *ARGHAP29* (Leslie and Murray, 2012), *MAFB* and *PAX7* (Butali et al., 2014). In addition, the advent of next-generation sequencing (NGS) technologies stimulated a genome-wide hunt for rare variants, by means of exome and genome sequencing. With the progressive drop in NGS costs, coupled with increase in throughput, this approach has become accessible by many research groups studying common diseases (Do et al., 2012; O'Roak et al., 2012). Recently, exome sequencing in NSCL/P patients has enabled the identification of possibly pathogenic variants in new genes, such as *CDH1*, at 16q22.1, (Bureau et al., 2014) and *DLX4*, at 17q21.33 (Wu et al., 2014), among other putative candidates (Liu et al., 2015).

Nevertheless, attributing a pathogenic role for a given rare variant is not a trivial task. Firstly, classifying a given variant as rare requires the availability of large population databases; in this regard, databases such as the 1000 Genomes Project (1kGP; <http://www.1000genomes.org/>) and the Exome Variant Server / NHLBI Exome Sequencing Project (ESP6500; <http://evs.gs.washington.edu/EVS/>) are valuable starting points. However, many populations, including Brazilian, are poorly represented in these databases. Therefore, local databases are of great relevance for identifying local common variants. Secondly, incomplete penetrance and genetic heterogeneity are expected to occur within families segregating complex diseases under the CDRV model, which may represent a confounding factor in exome sequencing studies that seek for co-segregation of variants in families (Cooper et al., 2013). Even though, multiplex families still retain the best chances of finding a rare, pathogenic variant.

Objectives

Our main objective is to find the major susceptibility variants / loci underlying NSCL/P in the Brazilian population. We aimed to explore the broad spectrum of allele frequency (either rare or common variants) in NSCL/P, by means of different strategies. In this respect, our objective can be divided as follows:

a) Identify rare, moderate-to-high effect variants underlying NSCL/P in familial cases, under two main hypothesis: (i) affected relatives sharing a major causative locus, which may vary among families, and (ii) affected relatives presenting at least two moderate-effect risk variants, not necessarily the same (i.e., genetic heterogeneity of moderate-effect risk variants within a family).

b) Investigate the role of common, low-risk variants in NSCL/P etiology, by (i) characterizing the 8q24 susceptibility locus in the Brazilian population, attempting to narrow the 640-kb interval previously associated; (ii) replicating some of the GWAS hits and (iii) seek for new susceptibility factors, combining association analysis and expression quantitative trait loci mapping.

References

- Aldhore KA, Bohmer AC, Ludwig KU, Esmail AH, Al-Hebshi NN, Lippke B, Golz L, Nothen MM, Daratsianos N, Knapp M *et al.* 2014. Nonsyndromic cleft lip with or without cleft palate in arab populations: genetic analysis of 15 risk loci in a novel case-control sample recruited in Yemen. *Birth Defects Res A Clin Mol Teratol* 100:307-313.
- Altshuler D, Daly MJ, Lander ES. 2008. Genetic mapping in human disease. *Science* 322:881-888.
- Beaty TH, Murray JC, Marazita ML, Munger RG, Ruczinski I, Hetmanski JB, Liang KY, Wu T, Murray T, Fallin M *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.
- Birnbaum S, Ludwig KU, Reutter H, Herms S, Steffens M, Rubini M, Baluardo C, Ferrian M, Almeida de Assis N, Alblas M *et al.* 2009. Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24. *Nat Genet* 41:473-477.
- Bodmer W, Bonilla C. 2008. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* 40:695-701.
- Boehringer S, van der Lijn F, Liu F, Gunther M, Sinigerova S, Nowak S, Ludwig KU, Herberz R, Klein S, Hofman A *et al.* 2011. Genetic determination of human facial morphology: links between cleft-lips and normal variation. *Eur J Hum Genet* 19:1192-1197.
- Brito LA, Bassi CF, Masotti C, Malcher C, Rocha KM, Schlesinger D, Bueno DF, Cruz LA, Barbara LK, Bertola DR *et al.* 2012a. IRF6 is a risk factor for nonsyndromic cleft lip in the Brazilian population. *Am J Med Genet A* 158A:2170-2175.
- Brito LA, Cruz LA, Rocha KM, Barbara LK, Silva CB, Bueno DF, Aguená M, Bertola DR, Franco D, Costa AM *et al.* 2011. Genetic contribution for non-syndromic cleft lip with or without cleft palate (NS CL/P) in different regions of Brazil and implications for association studies. *Am J Med Genet A* 155A:1581-1587.
- Brito LA, Meira JG, Kobayashi GS, Passos-Bueno MR. 2012b. Genetics and management of the patient with orofacial cleft. *Plast Surg Int* 2012:782821.
- Brito LA, Paranaíba LM, Bassi CF, Masotti C, Malcher C, Schlesinger D, Rocha KM, Cruz LA, Barbara LK, Alonso N *et al.* 2012c. Region 8q24 is a susceptibility locus for nonsyndromic oral clefting in Brazil. *Birth Defects Res A Clin Mol Teratol* 94:464-468.
- Bureau A, Parker MM, Ruczinski I, Taub MA, Marazita ML, Murray JC, Mangold E, Nothen MM, Ludwig KU, Hetmanski JB *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.
- Butali A, Mossey P, Adeyemo W, Eshete M, Gaines L, Braimah R, Aregbesola B, Rigdon J, Emeka C, Olutayo J *et al.* 2014. Rare functional variants in genome-wide association identified candidate genes for nonsyndromic clefts in the African population. *Am J Med Genet A* 164A:2567-2571.
- Calzolari E, Milan M, Cavazzuti GB, Cocchi G, Gandini E, Magnani C, Moretti M, Garani GP, Salvioli GP, Volpato S. 1988. Epidemiological and genetic study of 200 cases of oral cleft in the Emilia Romagna region of northern Italy. *Teratology* 38:559-564.

- Carlson L, Hatcher KW, Vander Burg R. 2013. Elevated infant mortality rates among oral cleft and isolated oral cleft cases: a meta-analysis of studies from 1943 to 2010. *Cleft Palate Craniofac J* 50:2-12.
- Christensen K, Fogh-Andersen P. 1993. Cleft lip (+/- cleft palate) in Danish twins, 1970-1990. *Am J Med Genet* 47:910-916.
- Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H. 2013. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum Genet* 132:1077-1130.
- Cura F, Bohmer AC, Klamt J, Schunke H, Scapoli L, Martinelli M, Carinci F, Nothen MM, Knapp M, Ludwig KU *et al.* 2015. Replication analysis of 15 susceptibility loci for nonsyndromic cleft lip with or without cleft palate in an Italian population. *Birth Defects Res A Clin Mol Teratol*.
- Dixon MJ, Marazita ML, Beaty TH, Murray JC. 2011. Cleft lip and palate: understanding genetic and environmental influences. *Nat Rev Genet* 12:167-178.
- Do R, Kathiresan S, Abecasis GR. 2012. Exome sequencing and complex disease: practical aspects of rare variant association studies. *Hum Mol Genet* 21:R1-9.
- Figueiredo JC, Ly S, Raimondi H, Magee K, Baurley JW, Sanchez-Lara PA, Ihenacho U, Yao C, Edlund CK, van den Berg *Det al.* 2014. Genetic risk factors for orofacial clefts in Central Africans and Southeast Asians. *Am J Med Genet A* 164A:2572-2580.
- Fogh-Andersen P. 1942. Inheritance of Harelip and Cleft Palate. Copenhagen: Munksgaard.
- Fraser FC. 1955. Thoughts on the etiology of clefts of the palate and lip. *Acta Genet Stat Med* 5:358-369.
- Gibson G. 2012. Rare and common variants: twenty arguments. *Nat Rev Genet* 13:135-145.
- Gorlin RJ, Cohen Jr. MMH, R. C. M. 2001. Syndromes of the Head and Neck. New York, NY, US: Oxford University Press. p 905-907.
- Grant SF, Wang K, Zhang H, Glaberson W, Annaiah K, Kim CE, Bradfield JP, Glessner JT, Thomas KA, Garriss *Met al.* 2009. A genome-wide association study identifies a locus for nonsyndromic cleft lip with or without cleft palate on 8q24. *J Pediatr* 155:909-913.
- Grosen D, Chevrier C, Skytthe A, Bille C, Molsted K, Sivertsen A, Murray JC, Christensen K. 2010. A cohort study of recurrence patterns among more than 54,000 relatives of oral cleft cases in Denmark: support for the multifactorial threshold model of inheritance. *J Med Genet* 47:162-168.
- Hu DN, Li JH, Chen HY, Chang HS, Wu BX, Lu ZK, Wang DZ, Liu XG. 1982. Genetics of cleft lip and cleft palate in China. *Am J Hum Genet* 34:999-1002.
- Jiang R, Bush JO, Lidral AC. 2006. Development of the upper lip: morphogenetic and molecular mechanisms. *Dev Dyn* 235:1152-1166.
- Jugessur A, Rahimov F, Lie RT, Wilcox AJ, Gjessing HK, Nilsen RM, Nguyen TT, Murray JC. 2008. Genetic variants in IRF6 and the risk of facial clefts: single-marker and haplotype-based analyses in a population-based case-control study of facial clefts in Norway. *Genet Epidemiol* 32:413-424.
- Jugessur A, Shi M, Gjessing HK, Lie RT, Wilcox AJ, Weinberg CR, Christensen K, Boyles AL, Daack-Hirsch S, Trung T *Net al.* 2009. Genetic determinants of facial clefting: analysis of 357 candidate genes using two national cleft studies from Scandinavia. *PLoS One* 4:e5385.
- Kerrigan JJ, Mansell JP, Sengupta A, Brown N, Sandy JR. 2000. Palatogenesis and potential mechanisms for clefting. *J R Coll Surg Edinb* 45:351-358.
- Kondo S, Schutte BC, Richardson RJ, Bjork BC, Knight AS, Watanabe Y, Howard E, de Lima RL, Daack-Hirsch S, Sander *Aet al.* 2002. Mutations in IRF6 cause Van der Woude and popliteal pterygium syndromes. *Nat Genet* 32:285-289.

- Kruglyak L. 2008. The road to genome-wide association studies. *Nat Rev Genet* 9:314-318.
- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh *Wet al.* 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860-921.
- Leslie EJ, Marazita ML. 2013. Genetics of cleft lip and cleft palate. *Am J Med Genet C Semin Med Genet* 163C:246-258.
- Leslie EJ, Murray JC. 2012. Evaluating rare coding variants as contributing causes to non-syndromic cleft lip and palate. *Clin Genet* 84:496-500.
- Liu F, van der Lijn F, Schurmann C, Zhu G, Chakravarty MM, Hysi PG, Wollstein A, Lao O, de Bruijne M, Ikram MA *et al.* 2012. A genome-wide association study identifies five loci influencing facial morphology in Europeans. *PLoS Genet* 8:e1002932.
- Liu YP, Xu LF, Wang Q, Zhou XL, Zhou JL, Pan C, Zhang JP, Wu QR, Li YQ, Xia Y *et al.* 2015. Identification of susceptibility genes in non-syndromic cleft lip with or without cleft palate using whole-exome sequencing. *Med Oral Patol Oral Cir Bucal* 20:e763-770.
- Ludwig KU, Mangold E, Herms S, Nowak S, Reutter H, Paul A, Becker J, Herberz R, AlChawa T, Nasser *Eet 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.
- Maher B. 2008. Personal genomes: The case of the missing heritability. *Nature* 456:18-21.
- Mangold E, Ludwig KU, Birnbaum S, Baluardo C, Ferrian M, Herms S, Reutter H, de Assis NA, Chawa TA, Mattheisen *Met al.* 2009. Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. *Nat Genet* 42:24-26.
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti *Aet al.* 2009. Finding the missing heritability of complex diseases. *Nature* 461:747-753.
- Marazita ML. 2012. The evolution of human genetic studies of cleft lip and cleft palate. *Annu Rev Genomics Hum Genet* 13:263-283.
- Mitchell LE, Beaty TH, Lidral AC, Munger RG, Murray JC, Saal HM, Wyszynski DF. 2002. Guidelines for the design and analysis of studies on nonsyndromic cleft lip and cleft palate in humans: summary report from a Workshop of the International Consortium for Oral Clefts Genetics. *Cleft Palate Craniofac J* 39:93-100.
- Mitchell LE, Christensen K. 1996. Analysis of the recurrence patterns for nonsyndromic cleft lip with or without cleft palate in the families of 3,073 Danish probands. *Am J Med Genet* 61:371-376.
- Mossey PA, Little J, Munger RG, Dixon MJ, Shaw WC. 2009. Cleft lip and palate. *Lancet* 374:1773-1785.
- Murray JC, Daack-Hirsch S, Buetow KH, Munger R, Espina L, Paglinawan N, Villanueva E, Rary J, Magee K, Magee W. 1997. Clinical and epidemiologic studies of cleft lip and palate in the Philippines. *Cleft Palate Craniofac J* 34:7-10.
- Murray T, Taub MA, Ruczinski I, Scott AF, Hetmanski JB, Schwender H, Patel P, Zhang TX, Munger RG, Wilcox A *Jet al.* 2012. 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.
- Nordstrom RE, Laatikainen T, Juvonen TO, Ranta RE. 1996. Cleft-twin sets in Finland 1948-1987. *Cleft Palate Craniofac J* 33:340-347.
- O'Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, Girirajan S, Karakoc E, Mackenzie AP, Ng SB, Baker *Cet al.* 2012. Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat Genet* 43:585-589.

- Rahimov F, Marazita ML, Visel A, Cooper ME, Hitchler MJ, Rubini M, Domann FE, Govil M, Christensen K, Bille *Cet al.* 2008. Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip. *Nat Genet* 40:1341-1347.
- Reich DE, Lander ES. 2001. On the allelic spectrum of human disease. *Trends Genet* 17:502-510.
- Rojas-Martinez A, Reutter H, Chacon-Camacho O, Leon-Cachon RB, Munoz-Jimenez SG, Nowak S, Becker J, Herberz R, Ludwig KU, Paredes-Zenteno *Met al.* 2010. Genetic risk factors for nonsyndromic cleft lip with or without cleft palate in a Mesoamerican population: Evidence for IRF6 and variants at 8q24 and 10q25. *Birth Defects Res A Clin Mol Teratol* 88:535-537.
- Schork NJ, Murray SS, Frazer KA, Topol EJ. 2009. Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev* 19:212-219.
- Schutte BC, Murray JC. 1999. The many faces and factors of orofacial clefts. *Hum Mol Genet* 8:1853-1859.
- Sivertsen A, Wilcox AJ, Skjaerven R, Vindenes HA, Abyholm F, Harville E, Lie RT. 2008. Familial risk of oral clefts by morphological type and severity: population based cohort study of first degree relatives. *Bmj* 336:432-434.
- Stanier P, Moore GE. 2004. Genetics of cleft lip and palate: syndromic genes contribute to the incidence of non-syndromic clefts. *Hum Mol Genet* 13 Spec No 1:R73-81.
- Sun Y, Huang Y, Yin A, Pan Y, Wang Y, Wang C, Du Y, Wang M, Lan F, Hu *Zet al.* 2015. Genome-wide association study identifies a new susceptibility locus for cleft lip with or without a cleft palate. *Nat Commun* 6:6414.
- Twigg SR, Wilkie AO. 2015. New insights into craniofacial malformations. *Hum Mol Genet*.
- Uslu VV, Petretich M, Ruf S, Langenfeld K, Fonseca NA, Marioni JC, Spitz F. 2014. Long-range enhancers regulating Myc expression are required for normal facial morphogenesis. *Nat Genet* 46:753-758.
- Vanderas AP. 1987. Incidence of cleft lip, cleft palate, and cleft lip and palate among races: a review. *Cleft Palate J* 24:216-225.
- Visscher PM, Brown MA, McCarthy MI, Yang J. 2012. Five years of GWAS discovery. *Am J Hum Genet* 90:7-24.
- Waitzman NJ, Romano PS, Scheffler RM. 1994. Estimates of the economic costs of birth defects. *Inquiry* 31:188-205.
- Weatherley-White RC, Ben S, Jin Y, Riccardi S, Arnold TD, Spritz RA. 2011. Analysis of genomewide association signals for nonsyndromic cleft lip/palate in a Kenya African Cohort. *Am J Med Genet A* 155A:2422-2425.
- Wu D, Mandal S, Choi A, Anderson A, Prochazkova M, Perry H, Gil-Da-Silva-Lopes VL, Lao R, Wan E, Tang *PLet al.* 2014. DLX4 is associated with orofacial clefting and abnormal jaw development. *Hum Mol Genet* 24:4340-4352.
- Xu MY, Deng XL, Tata LJ, Han H, Chen XH, Liu TY, Chen QS, Yao XW, Tang SJ. 2012. Case-control and family-based association studies of novel susceptibility locus 8q24 in nonsyndromic cleft lip with or without cleft palate in a Southern Han Chinese population located in Guangdong Province. *DNA Cell Biol* 31:700-705.
- Yang T, Jia Z, Bryant-Pike W, Chandrasekhar A, Murray JC, Frittsch B, Bassuk AG. 2014. Analysis of PRICKLE1 in human cleft palate and mouse development demonstrates rare and common variants involved in human malformations. *Mol Genet Genomic Med* 2:138-151.
- Zuccherro TM, Cooper ME, Maher BS, Daack-Hirsch S, Nepomuceno B, Ribeiro L, Caprau D, Christensen K, Suzuki Y, Machida *Jet al.* 2004. Interferon regulatory factor 6 (IRF6) gene variants and the risk of isolated cleft lip or palate. *N Engl J Med* 351:769-780.

Chapter 2

Exome analysis in multiplex families reveals novel candidate genes for nonsyndromic cleft lip / palate

*Brito LA, Ezquina S, Savastano C, Hsia G, Malcher C, Yamamoto GL,
Passos-Bueno MR*

Centro de Estudos do Genoma Humano e Células-Tronco, Instituto de Biociências,
Universidade de São Paulo, SP, Brasil

Key words: E-cadherin, zebrafish, CRISPR-Cas9, embryonic lethality, orofacial clefts.

Abstract

Nonsyndromic cleft lip with or without cleft palate (NSCL/P) is a prevalent complex disorder. The role played by rare, high-effect variants in NSCL/P etiology has been focus of intense debate. Here, we used exome sequencing in familial cases of NSCL/P to explore the relevance of such variants to the disease. We sequenced, in total, 29 individuals from nine familial cases, segregating NSCL/P under autosomal dominant-like inheritance, in order to maximize chances of finding a true causative variant. After filtering strategies that prioritized rare variants in functionally relevant genes, we identified, for two families, a pathogenic variant in the epithelial-cadherin gene. In other family, no candidate gene was identified under a major-effect gene hypothesis, suggesting influence of genetic heterogeneity. For the six remaining families, we raised a list of 11 promising candidate variants in genes involved with planar cell polarity pathway, microtubules, cell adhesion, epithelial-mesenchymal transition or cell cycle control. In conclusion, our approach successfully identified the causative variant in two out of nine families, up to now; meanwhile, we shed light on new candidate genes and pathways, that should be investigated by future functional and population studies.

Resumo

Fissura labial com ou sem fissura de palato não sindrômica (FL/P NS) é uma doença complexa, para a qual a contribuição genética ainda é pouco conhecida. Após sucessivos estudos de associação de varredura genômica, pesquisadores têm direcionado atenção para o papel de variantes raras de grande efeito na etiologia das FL/P NS. Neste estudo, utilizamos sequenciamento de exoma para investigar a existência de variantes raras potencialmente patogênicas em famílias segregando FL/P NS. Foram sequenciados 29 indivíduos, pertencentes a nove famílias. A priorização de variantes levou em consideração a qualidade das variantes, frequência ($<0.5\%$), predições *in silico* de dano na proteína e conservação, além de informações funcionais depositadas em bancos de dados. Assumindo herança autossômica dominante com penetrância incompleta, nós identificamos uma variante patogênica no gene *CDH1* segregando em duas famílias. Em outra família, não foi possível priorizar nenhuma variante, possivelmente devido a efeito de heterogeneidade genética. Para as seis famílias restantes, nós geramos uma lista de 11 variantes candidatas, em genes relacionados à via de polaridade planar celular, microtúbulos, transição epitélio-mesênquima/adesão celular e controle de ciclo celular. Em conclusão, nossa abordagem foi capaz de identificar a variante causal em duas de nove famílias. Além disso este estudo sugere novos genes e vias candidatos que devem ser investigados em futuras análises funcionais e populacionais.

Introduction

The modest success of genome-wide association studies (GWAS) in implicating common variants in a substantial risk to nonsyndromic cleft lip with or without cleft palate (NSCL/P) has suggested that these variants may not be the main source of genetic variation for NSCL/P susceptibility (Leslie and Marazita, 2013). This observation confronts the validity of common disease – common variant hypothesis for NSCL/P, which assumes that numerous common, low-risk variants would be the main contributors to the allelic spectrum of common diseases (Schork et al., 2009). Prior to GWAS era, however, linkage analyses also failed in identifying causative *loci* for the majority of the families studied, under the hypothesis that a single causative variant (thus, presumably rare) would drive the phenotype (Marazita et al., 2004). As possible reasons behind this, are genetic heterogeneity within families and oligogenic inheritance, in which at least two moderately penetrant variants would drive the phenotype (in a combination that might vary among individuals from a same family).

With the advent of next generation sequencing (NGS), the role of rare, high-effect variants could be reassessed with better resolution, and it has been considered a powerful tool for the identification of such variants. Using exome sequencing in familial cases, rare and private variants have been implicated with NSCL/P in a limited, yet growing, number of families (Leslie and Murray, 2012; Wu et al., 2014). These studies have also indicated that a still unknown fraction of NSCL/P familial cases displays Mendelian-like inheritance, usually with incomplete penetrance, which may be caused by a variant in a single major gene, or by a group of few contributing variants in different genes. Under this assumption, a rare variant with largely deleterious effect might be sufficient to cause the phenotype, whereas moderately deleterious variants would require other contributing variants to do so. In addition, given the incomplete penetrance and phenotypic expressivity commonly observed in NSCL/P families, modifier variants and environmental components may also play an important role.

In this article, we report the exome sequencing of 9 families segregating NSCL/P. We raise new candidate variants underlying NSCL/P, and explore gene pathways that may be involved with the disease.

Material and Methods

Subjects

We ascertained 9 large Brazilian pedigrees with multiple affected members at Hospital Sobrapar (Campinas-SP; F1843), Hospital das Clínicas (São Paulo-SP; F8418), and during surgical missions of Operation Smile in Barbalha-CE (F2570 and F3196) and Fortaleza-CE (F617, F886, F2848, F3788, F7614). All affected individuals presented cleft lip with or without cleft palate, or lip scar, without any additional malformation or comorbidity.

Ethics Statements

This study was approved by the Ethics Committee of the Instituto de Biociências of Universidade de São Paulo, Brazil. All biological samples were obtained after informed consent from the patients or their legal guardians.

DNA Preparation

DNA was purified from peripheral blood (according to standard protocols) or saliva (collected with Oragene® saliva collection kits OG-500 and OG-575; DNA GenotekInc, Ottawa, Canada), following manufacturer's instructions.

Library Construction and Exome Sequencing

Library preparation and exome capture were performed with Illumina's TruSeq Sample Prep and Exome Enrichment Kits, for individuals from families F617, F886, F2570, F3196, F3788 and F7614. Nextera Rapid Capture Exome was used for individuals from families F1843, F2848 and F8418. Library quantification was performed with KAPA Library Quantification kit (KAPA Biosystems), through real-time quantitative PCR.

Paired-end sequencing was performed on a HiScanSQ (Illumina) for families F617, F886, F2570, F3195, F3788 and F7614, and on a HiSeq 2500 (Illumina) for families F1843, F2848 and F8418. Exome mean coverage for each family is as follows: F886: 65.1x; F1843: 43.8x; F2570: 66.7x; F2848: 48.1x; F3196: 62.3x; F7614: 64.9x; F8418: 53.8x. Additional sequencing details of F617 and F3788 are described elsewhere (Brito et al., 2015).

Exome Data Processing

Sequences were aligned to the hg19 reference genome with Burrows-Wheller Aligner (BWA; <http://bio-bwa.sourceforge.net>). Genome indexing, realignment of reads and duplicate removal were performed with Picard (<http://broadinstitute.github.io/picard/>). Variants were then called using Genome Analysis Toolkit package (GATK; <https://www.broadinstitute.org/gatk/>), and subsequently annotated with ANNOVAR (<http://www.openbioinformatics.org/annovar/>).

Variant Filtering

We applied a “frequency filter”, to exclude variants with minor allele frequency (MAF) > 0.5% in public databases (1000 Genomes Project [1kGP; Abecasis et al., 2012], NHLBI Exome Sequencing Project [ESP6500; <http://evs.gs.washington.edu/EVS/>], Exome Aggregation Consortium [ExAC; <http://exac.broadinstitute.org/>]). To account for local polymorphisms, we also used our in-house database (CEGH60+, a collection of exome sequencing data of 609 elderly Brazilians from the biobank of the Centro de Pesquisa sobre o Genoma Humano e Células Tronco, coordinated by M. Zatz), and additional exomes of patients affected by unrelated conditions.

To avoid false positive calls, we applied a “quality filter”, that removed variants with low quality (minimum GATK quality score threshold fixed as 30), low coverage (<10x), and displaying allelic imbalance greater than 25% : 75%. Synonymous variants, or variants located in hypervariable genes (Fuentes Fajardo et al., 2012), were also removed from further analysis.

After frequency and quality filters, the refined list of variants was then submitted to a variety of tools and strategies for gene prioritization. VarElect prioritization tool (<http://varelect.genecards.org>) was used to identify variants in genes that are either directly related with NSCL/P, or that interact with genes with known role in lip and palate morphogenesis. “Cleft lip” and “cleft palate” were used as VarElect queries. We also prioritized genes involved with epithelial-mesenchymal transition (EMT), according to dbEMT (<http://dmemt.bioinfo-minzhao.org>), based on the role that EMT plays during lip and palate morphogenesis (Griffith and Hay, 1992; Sun et al., 2000), and on the common association between NSCL/P and different types of carcinomas, that frequently undergo EMT (Seto-Salvia and Stanier, 2014). Gene expression and phenotypes in animal models were evaluated using Mouse Genome Informatics (<http://www.informatics.jax.org>) and ZFIN (The Zebrafish Model Organism Database, <http://zfin.org>); genes expressed in early craniofacial embryogenesis were prioritized, as well as those implicated in craniofacial abnormalities or embryonic lethality. The bioinformatics tool SysFACE (<http://bioinformatics.udel.edu/Research/SysFACE/>) was used to rank genes with enriched expression during mandible, maxilla, frontonasal region and palate development, based on data of gene expression profiles in mice, from E10 to E14.5, retrieved from the FaceBase consortium database (www.facebase.org). This tool prioritizes not only genes highly expressed in absolute levels in those tissues, but also considers relative expression, normalizing to whole embryo body tissue control. ExAC Browser’s constraint metrics (Lek et al., 2015) were used to prioritize genes significantly scored for intolerance to harbor missense or loss-of-function (LoF) variants. This analysis uses the collection of gene variants deposited in ExAC database to estimate the observed x expected number of missense and LoF mutations for each gene, classifying them in either tolerant or intolerant to harbor these mutations. Other bioinformatics tools were used to assess conservation and predictions of protein damage: SIFT (<http://sift.jcvi.org>), Polyphen-2 (<http://genetics.bwh.harvard.edu/pph2/>), Mutation Taster and PhastCons (<http://www.mutationtaster.org>). PubMed, OMIM and GeneCards databases were also examined. Collectively, these analyses provided the basis for raising a list of the best candidate variants.

Variant Validation

Variants classified as best candidates were visually inspected using Integrative Genomics Viewer software (Broad Institute of MIT and Harvard). Sanger sequencing was used for variant validation and, whenever appropriate, for mutation screening in additional relatives. PCR primers were designed with Primer Designing Tool web interface (NCBI). Primer sequences and PCR conditions are detailed in Supplementary Table 1. PCR products were sequenced with ABI 3730 DNA Analyzer (Applied Biosystems), and sequences were visualized using Sequencher® 5.2 analysis software (Gene Codes).

Results

Exome sequencing was performed in 29 individuals affected by NSCL/P with variable degree of severity, from 9 Brazilian families (F617:III-13, IV-1, IV-9, V-1; F886:II-8, II-17, II-18, IV-1; F1843: II-5, II-8, III-1; F2570: II-14, III-2, III-3; F2848: I-2, II-6; F3196: II-2, III-2, III-3; F3788: III-4, III-7, III-16, IV-2; F7614: I-3, I-4, III-6; F8418: II-3, II-5, III-7; Figure 1). On average, 26,127 (± 1330 , standard deviation) exonic variants were called for each sequenced individual. Based on the segregation pattern of the malformation within each family, autosomal dominant was considered the most probable NSCL/P mode of inheritance in all families, assuming a major-effect gene segregating with an incompletely penetrant phenotype. Therefore, for each family, we selected heterozygous variants shared by all affected relatives sequenced.

After applying the quality and frequency ($>0.5\%$) filters, excluding also hypervariable genes and synonymous variants, we obtained a total of 357 variants (mean of 45 per family, excluding family F886, for which no variant passed the filters). Of these variants, 68 were absent in all databases, including our in-house control database (Figure 2). All variants that persisted after these filters are listed in Supplementary Table 2. No obvious orofacial cleft gene was present in this list.

Figure 3 – Pedigrees of NSCL/P familial cases included in the study. Genotypes of candidate variants (denoted by gene's name) are displayed in the tables, where +/- refers to the presence of variant in heterozygosis. Individuals included in the exome analysis are identified in the tables (e).

	Raw exonic SNVs (mean per individual)	Heterozygous shared among affected individuals	Excluding low quality and hypervariable genes	MAF<0.005	Nonsynonymous, indels or LoF / number of non-described
F886	24,156	1,853	1,293	0	0
F1843	26,554	5,563	4,216	49	34 / 4
F2570	28,467	5,549	4,589	99	60 / 6
F2848	25,778	8,386	6,825	193	137/25
F3196	27,468	5,518	4,333	107	65 / 17
F7614	27,085	5,067	4,100	36	20 / 7
F8418	26,555	5,093	3,930	54	34 / 6
F617 (Brito et al., 2015)	25,096	1,933	1,298	3	3 / 2
F3788 (Brito et al., 2015)	25,346	1,912	1,179	6	4 / 1

Figure 2 – Number of remaining variants after each filtering step, for each family.

From this list, we elected, using bioinformatics prioritization tools and online databases, 11 best candidate variants (8 novel and 3 previously described in databases; Table 1). In this selection, we included genes implicated with craniofacial abnormalities in animal models, and excluded variants in genes with no expression in craniofacial development, according to SysFACE tool. In addition, genes with unknown expression profile in craniofacial development or unknown associated phenotypes, but harboring novel variants with *in silico* predictions of protein damage were also elected (e.g., *PAX8*). Variants with benign *in silico* predictions in all bioinformatics tools, on the other hand, were excluded. The elected variants were validated by Sanger sequencing (Supplementary Figure 1).

Among the novel variants, 2 were LoF (in *PPM1F* and *PRICKLE1*), while all remaining variants were missense. Functional data retrieved from SysFACE revealed that *PPM1F*, *KIF20B*, *ROR2*, *ZEB1*, *IGF2R*, *PRICKLE1*, *KIFAP3*, *CDK1* and *CDH1* are highly

expressed in murine craniofacial embryogenesis. In particular, some of these genes presented enriched expression in embryonic craniofacial structures (compared to whole embryonic body): *PPM1F* (in mandible, maxilla, frontonasal region and palate), *PRICKLE1* (in maxilla, frontonasal region and palate), *SETD5* (in maxilla and frontonasal region) and *IGF2R* (in maxilla and palate). Phenotypes of complete knockout mice were associated with craniofacial anomalies for 6 of these candidate genes (*KIF20B*, *ROR2*, *ZEB1*, *IGF2R*, *PRICKLE1*, and *KIFAP3*), and with variable degree of lethality for 10 candidate genes (*PPM1F*, *PAX8*, *KIF20B*, *PRICKLE1*, *IGF2R*, *KIFAP3*, *ROR2*, *ZEB1*, *CDK1* and *CDH1*).

Additional affected and unaffected relatives were screened for segregation analysis, whenever possible. Segregation of *PPM1F* variant in F1843 was not complete (Figure 1C), as the variant was absent in individual I-2, whose brother presents NSCL/P. For the other variants, segregation with NSCL/P was observed in all affected individuals available, including non-affected obligate carriers (Figure 1A,D-I). Some of the candidate variants also segregated in non-affected individuals (not considering obligate carriers), as observed for *ROR2* and *CDK1* variants, in families F2848 and F8418, respectively. On the other hand, variants in *PAX8* (F2570) and *ZEB1* (F2848) were absent in all non-affected relatives sequenced. Finally, families F617 and F3788 segregated the same candidate variant, c.760G>A, in *CDH1*.

Table 1 –Best candidate variants identified in our families

Family	Gene	Region	Position (hg19)	Variant	Type	Qual	<i>In silico predictions*</i>	Frequency	Relevant Informations
							SIFT / Polyphen-2 HD;HV / Mutation Taster / PhasCons / ExAC	ExAC / 1kGP / ESP6500 / CEGH60+	
F1843	PPM1F	22q11	22300402	c.19C>T:p.Q7X (NM_014634)	stopgain	1334	0.02 D / na;na / Disease causing (prob:1) / na / pLI=0 (T) (ENST263212)	0 / 0 / 0 / 0	Phosphatase, regulates kinesin-2 complex and cell adhesion. Enriched expression in murine craniofacial morphogenesis. Knockout mice may exhibit preweaning lethality.
	PAX8	2q24	113999636	c.550G>C:p.G184R (NM_003466)	missense	5167	0.36 T / 1 D ; 1 D / Disease causing (prob:1) / 1 C / z=0.5 (T) (ENST429538)	0 / 0 / 0 / 0	Gene of paired-box family, important for embryogenesis of several tissues. Knockout mice may exhibit postnatal lethality.
F2570	KIF20B	10q23	91522558	c.4835A>G:p.K1612R	missense	1326	0.19 T / 0.06 B; 0.03 B / Disease causing (prob:1) / 1 C / z=-3.5 (T)	0.0004 / 0.001 / 0.001 / 0	Kinesin-interacting phosphoprotein, regulates cell cycle progression. Expressed in murine craniofacial morphogenesis. Knockout mice present abnormal craniofacial development, including cleft palate, and may exhibit embryonic lethality.
			91532586	c.5263G>A:p.V1755M (NM_016195)	missense	1711	0.001 D / 1 D ; 1 D / Polymorphism (prob:1) / 0.02 U / z=-3.5 (T) (ENST260753)	0.001 / 0.001 / 0.002 / 0.0001	
F2848	ROR2	9q22	94487187	c.1589G>A:p.R530Q (NM_004560)	missense	865	0.05 D / 0.36 B / 0.01 B / Disease causing (prob:0.98) / 1 C / z=0.5 (T) (ENST375708)	0.002 / 0.001 / 0.002 / 0.001	Cell receptor, involved with noncanonical Wnt5a / PCP pathway during palatogenesis. Expressed in murine craniofacial morphogenesis. Knockout mice present abnormal craniofacial development, including cleft palate, and may exhibit neonatal lethality.
	ZEB1	10p11	31809536	c.1213A>G:p.I405V (NM_001174093)	missense	185	0.09 T / 0.76 P / 0.25 B / Disease causing (prob:1) / 1 C / z=0.7 (T) (ENST560721)	0 / 0 / 0 / 0	Zinc finger transcription factor, negatively regulates E-cadherin during EMT. Expressed in murine craniofacial morphogenesis. Knockout mice present abnormal craniofacial development, including cleft lip and palate, and may exhibit neonatal lethality.

F3196	<i>IGF2R</i>	6q25	160468235	c.2096C>T:p.S699L (NM_000876)	missense	1021	0.001 D / 1 D ; 1 D / Disease causing (prob:1) / 0.99 C / z=1.4 (T) (ENST356956)	0 / 0 / 0 / 0	Cell receptor, binds IGF2. Enriched expression in murine craniofacial morphogenesis. Knockout mice may exhibit abnormal craniofacial development and / or neonatal lethality.
	<i>PRICKLE1</i>	12q12	42853958	c.2149delA:p.S717fs (NM_001144881)	frameshift deletion	583	na / na; na / na / pLI=1 (I) (ENST455697)	0 / 0 / 0 / 0	Core PCP pathway protein. Enriched expression in murine craniofacial morphogenesis. Knockout mice present abnormal craniofacial development, including cleft palate, and may exhibit embryonic lethality.
F7614	<i>KIFAP3</i>	1q24	169961306	c.1142G>C:p.C381S (NM_001204516)	missense	1536	0 D / 0.12 B ; 0.25 B / Disease causing (prob:1) / 1 C / z=1.2 (T) (ENST538366)	0 / 0 / 0 / 0	Part of kinesin-2 complex, involved with intracellular trafficking of cell adhesion proteins and Hedgehog signaling. Expressed in murine craniofacial morphogenesis. Knockout mice may exhibit abnormal craniofacial development and/or embryonic lethality.
F8418	<i>CDK1</i>	10q21	62544513	c.88G>A:p.V30I (NM_001170406)	missense	2501	0.09 T / 0.97 D ; 0.77 D / Disease causing (prob:1) / 1 C / z=2.6 (T) (ENST395284)	0 / 0 / 0 / 0	Cyclin-dependent kinase 1. Role in cell cycle control and double-strand break repair. Expressed in murine craniofacial morphogenesis. Knockout mice may exhibit embryonic lethality.
F3788 and F617**	<i>CDH1</i>	16q22	68844172	c.760G>A:p.D254N (NM_004360)	missense	6991	0.02 D / 1 D ; 1 D / Disease causing (prob:1) / 1 C / z=0.8 (T) (ENST261769)	0 / 0 / 0 / 0	Expressed in murine craniofacial morphogenesis. Knockout mice exhibit embryonic lethality.

Qual: GATK base quality (minimum threshold fixed as 30); **HD:** Polyphen-2 HumDiv; **HV:** Polyphen-2 HumVar; **1kGP:** 1000 Genomes Project; **ExAC:** Exome Aggregation Consortium; **ESP6500:** Exome Sequencing Project database; **CEGH60+:** in-house database (Centro de Estudos do Genoma Humano e Células-Tronco); **na:** not available. **D:** damaging; **P:** possibly damaging; **B:** Benign; **T:** tolerated (for SIFT), tolerant (for ExAC); **C:** conserved; **U:** unconserved; **I:** intolerant. **ENST:** Ensembl transcript.

* Ranges of score variation and thresholds for categorical variant predictions obtained with bioinformatic tools are as follows: SIFT (0-1), damaging if <=0.05; Polyphen-2 (0-1), probably damaging if >=0.909 (HumVar) and >= 0.957 (HumDiv). PhastCons (0-1), conserved if >.9; ExAC classifies genes as either tolerant (T) or intolerant (I) to missense (z score, intolerant if z>= 3) or LoF (pLI score, ranges from 0 to 1, intolerant if pLI>=0.9) mutations.

** Also reported in Brito et al. (2015).

Discussion

After the GWAS era, the role of rare variants with moderate-to-high effect in NSCL/P genetic architecture has been in the spotlight, explored either by using NGS in genomic approaches, or by resequencing GWAS loci. Here, we used exome sequencing in multiplex families to investigate the contribution of rare variants to NSCL/P etiology.

Gene prioritization criteria included quality and frequency of variants, bioinformatics predictions of pathogenicity, and annotations in databases regarding gene's function, associated phenotypes (in humans or animal models), expression profile and amino acid conservation. The whole picture provided by these criteria allowed us to conclusively identify a major pathogenic variant in 2 families: the *CDH1* variant c.760G>A, in families F617 and F3788, reported in a separate article (Brito et al., 2015). On the other hand, the absence of good candidates for family F886 strongly suggests that genetic heterogeneity is underlying NSCL/P in this family, especially considering that both branches of the family segregate the disease. For the remaining families, we raised a list of best candidates, which were positively assessed for most of the following criteria: gene expression during craniofacial development, craniofacial phenotypes in animal models, *in silico* prediction of protein damage [in at least one bioinformatics tool], and interaction with genes involved with craniofacial development. In this gene list, we observed enrichment for genes involved with Planar Cell Polarity (PCP) pathway, microtubules, cell adhesion and cell cycle control, as will be briefly discussed.

PCP signaling pathway is responsible for orchestrating cell and tissue polarity, through asymmetrical distribution and dynamics of PCP components (Sebbagh and Borg, 2014). During early craniofacial embryogenesis, PCP pathway controls elongation and migration of neural crest cells (NCC; De Calisto et al., 2005), and it was also shown to regulate craniofacial cartilage formation in mice and zebrafish, leading to bone defects if the pathway is dysregulated (Topczewski et al., 2011; Le Pabic et al., 2014). We identified mutations in key PCP genes – *PRICKLE1* (F3196), and the Wnt receptor *ROR2* (F2848). PCP events are triggered by the noncanonical Wnt/PCP pathway, initiated by Wnt5a and its receptor Ror2. Downstream in this pathway, Prickle1 is an intracellular core PCP protein asymmetrically distributed in the cell (Sebbagh and Borg, 2014). In mice, expression of Wnt5a, Ror2 and Prickle1 have been detected in the early

craniofacial cartilage and developing palate, where they regulate cell proliferation and migration during palatogenesis (He et al., 2008). Murine phenotypes associated with defects in these genes may include limb and craniofacial defects, including short snout and cleft palate (Schwabe et al., 2004; He et al., 2008; Yang et al., 2013). In humans, two isolate NSCL/P individuals carrying rare missense variants in *PRICKLE1*, shared with their unaffected mothers, were previously reported (Yang et al., 2013). Here, we report, for the first time, a novel *PRICKLE1* variant segregating in NSCL/P individuals from a nuclear family. Variants in *ROR2*, on the other hand, have been implicated, when in homozygosis, with Robinow syndrome, that fairly recapitulates the murine phenotypes (Brunetti-Pierri et al., 2008). In addition, a genetic association between *ROR2* markers and nonsyndromic cleft palate only, but not NSCL/P, has been reported in an Asian population (Wang et al., 2012). Nevertheless, the variant here reported, c.1589G>A, is present in databases.

Three microtubule-related genes were prioritized among our best candidates: *KIF20B* (2 variants, in F2570), *KIFAP3* (F7614), and *PPM1F* (F1843). Although none of them has been directly associated with orofacial clefts, functional annotations provided indirect links towards a potential role in NSCL/P etiology. *KIF20B* codifies the vertebrate-specific kinesin-6, a cell cycle regulator required for cytokinesis of the polarized neuroepithelial cells during cerebral cortex growth (Janisch et al., 2013). In addition, craniofacial defects has been observed in the microcephalic *magoo* mouse, deficient for *KIF20B*, which often displays shortened snout with or without cleft palate, in association with highly penetrant eye and thalamocortical system abnormalities (Dwyer et al., 2011). Two rare variants in this gene were found segregating in family F2570 (probably in *cis*, given their frequency). Although individual *in silico* predictions have not consistently classified them as pathogenic, it is possible that, together, they confer deleterious effect.

KIFAP3 (F7614) codifies the kinesin-associated protein 3 (KAP3), a non-motor protein that binds cargo and interacts with motor kinesin subunits KIF3A and KIF3B, forming the motor complex KIF3 (Tanuma et al., 2009). It has been reported that KIF3 activity is required for the correct function of primary cilia in NCC. Conditional knockout of KIF3A leads to truncation of primary cilia, which in turn result in gain of Hedgehog function and abnormal NCC proliferation in the facial midline of avian embryos (Brugmann et al., 2010). In addition, the cancer-related proteins β -catenin and N-cadherin undergo intracellular trafficking via KIF3 complex (Jimbo et al., 2002; Tanuma

et al., 2009). Interestingly, KIF3-mediated transport of N-cadherin to cell periphery is impaired in fibroblasts overexpressing POPX2, a KAP3 interacting partner, codified by *PPM1F* (Phang et al., 2014), which is mutated in F1843. Accordingly, high levels of POPX2 result in defective cell adhesion and increased cell motility, promoting the invasive behavior of cancer cells (Susila et al., 2010). In addition, it has been shown that POPX2 regulates centrosome positioning, which is crucial for cell polarity establishment and migration (Hoon et al., 2014). The possible involvement of *KIFAP3* and *PPM1F* with NSCL/P is reinforced by the link between NSCL/P and cadherin-mediated cell adhesion, as recently suggested (Vogelaar et al., 2013; Brito et al., 2015). The variants in *KIFAP3* and *PPM1F* here reported are novel and probably deleterious, according to *in silico* analysis. Segregation analysis revealed that c.19C>T, in *PPM1F*, segregates from the unaffected branch of the family; it is possible, however, that individual I-1's phenotype is the result of genetic heterogeneity in NSCL/P.

A role in cell adhesion is also described for transcription factor Zeb1, which regulates EMT in development and cancer. Zeb1 expression is shown to repress the epithelial signature of cells, by inducing downregulation of E-cadherin expression. Concurrently, Zeb1 upregulates N-cadherin and matrix metalloproteinases (MMPs), contributing to the loss of adhesive behavior and acquisition of a mesenchymal and invasive phenotype (Peinado et al., 2007; Xu et al., 2009; Lamouille et al., 2014). In fact, mice homozygous for either a *ZEB1* LoF mutation or for a regulatory mutation leading to *ZEB1* super expression display several malformations, including cleft palate and skeletal abnormalities, and premature death (Takagi et al., 1998; Kurima et al., 2011). Although *ZEB1* variant c.1213A>G presented ambiguous *in silico* predictions of protein damage, its absence in databases and in all non-affected siblings from family F2848 reinforces its putative pathogenicity.

Cdk1 (cyclin-dependent kinase 1) is part of the cell cycle-related subfamily of cyclin-dependent kinases (CDKs), which associates with a regulatory subunit, cyclin, to promote cell cycle progression in eukaryotes (Malumbres, 2014). A role in activation of DNA damage checkpoint in response to double-strand break (DSB), and in DSB-induced homologous recombination, has also been reported for Cdk1 in yeast (Ira et al., 2004). In fact, dysregulation of gene networks involved in cell cycle control and DSB repair was recently suggested to contribute to NSCL/P etiology (Kobayashi et al., 2013). In addition, bone morphogenetic protein 4, a mesoderm and bone inductor with relevant roles during craniofacial morphogenesis, was shown to upregulate Cdk1 levels in

hepatocellular cell lines, accelerating their cell cycle progression. (Chiu et al., 2012). The *CDK1* missense variant c.88G>A (family F8418) is absent in databases and local controls, and also predicted to be deleterious, according to *in silico* predictions, underpinning the pathogenic potential of this variant.

Although no obvious relation with craniofacial development exists for *IGF2R* (F3196) and *PAX8* (F2570), variants in these genes were also prioritized. Both are novel, and predicted to be pathogenic, by *in silico* analysis. *IGF2R* codifies a multifunctional receptor, which binds IGF2 and other molecules (Bergman et al., 2013). An enriched expression during murine craniofacial development has been observed in SysFACE. *PAX8* is a member of Paired box (PAX) gene family, which encodes DNA-binding transcription factors that coordinates organogenesis and lineage determination during embryogenesis. Although *PAX8* has been implicated mainly in the development of thyroid, central nervous system and ear, the family-members *PAX3* and *PAX7* are involved with facial development (Blake and Ziman, 2014).

Based on gene function, segregation analysis, and presence in databases, we elected 4 probably pathogenic variants from our best candidate list: *KIFAP3*, *PRICKLE1*, *ZEB1* and *CDK1*. We favored these genes over the others based on absence in databases, *in silico* predictions and segregation with phenotype. Nevertheless, we still classify the others as good candidates, given their potential functional role, especially if we assume that rare variants in more than 1 major gene may be necessary to drive the phenotype.

In summary, exome analysis allowed us to find a major pathogenic mutation in 2 out of 9 families. In addition, we report probably pathogenic variants underlying NSCL/P in 4 families, while no candidate was raised in one family. In light of our findings, we suggest that mutations in pathways related to PCP, cadherin-mediated cell adhesion, microtubules and cell cycle control may contribute to NSCL/P. These findings suggest that an important confounder effect in NSCL/P is genetic heterogeneity and also provide a source of new NSCL/P candidate genes for further functional and population approaches.

References

- Bergman D, Halje M, Nordin M, Engstrom W. 2013. Insulin-like growth factor 2 in development and disease: a mini-review. *Gerontology* 59:240-249.
- Blake JA, Ziman MR. 2014. Pax genes: regulators of lineage specification and progenitor cell maintenance. *Development* 141:737-751.
- Brito LA, Yamamoto GL, Melo S, Malcher C, Ferreira SG, Figueiredo J, Alvizi L, Kobayashi GS, Naslavsky MS, Alonso N *et al.* 2015. Rare variants in the epithelial cadherin gene underlying the genetic etiology of nonsyndromic cleft lip with or without cleft palate. *Hum Mutat.*
- Brugmann SA, Allen NC, James AW, Mekonnen Z, Madan E, Helms JA. 2010. A primary cilia-dependent etiology for midline facial disorders. *Hum Mol Genet* 19:1577-1592.
- Brunetti-Pierri N, Del Gaudio D, Peters H, Justino H, Ott CE, Mundlos S, Bacino CA. 2008. Robinow syndrome: phenotypic variability in a family with a novel intragenic ROR2 mutation. *Am J Med Genet A* 146A:2804-2809.
- Chiu CY, Kuo KK, Kuo TL, Lee KT, Cheng KH. 2012. The activation of MEK/ERK signaling pathway by bone morphogenetic protein 4 to increase hepatocellular carcinoma cell proliferation and migration. *Mol Cancer Res* 10:415-427.
- De Calisto J, Araya C, Marchant L, Riaz CF, Mayor R. 2005. Essential role of non-canonical Wnt signalling in neural crest migration. *Development* 132:2587-2597.
- Dwyer ND, Manning DK, Moran JL, Mudbhary R, Fleming MS, Favero CB, Vock VM, O'Leary DD, Walsh CA, Beier DR. 2011. A forward genetic screen with a thalamocortical axon reporter mouse yields novel neurodevelopment mutants and a distinct *emx2* mutant phenotype. *Neural Dev* 6:3.
- Fuentes Fajardo KV, Adams D, Mason CE, Sincan M, Tifft C, Toro C, Boerkoel CF, Gahl W, Markello T. 2012. Detecting false-positive signals in exome sequencing. *Hum Mutat* 33:609-613.
- Griffith CM, Hay ED. 1992. Epithelial-mesenchymal transformation during palatal fusion: carboxyfluorescein traces cells at light and electron microscopic levels. *Development* 116:1087-1099.
- He F, Xiong W, Yu X, Espinoza-Lewis R, Liu C, Gu S, Nishita M, Suzuki K, Yamada G, Minami Y *et al.* 2008. Wnt5a regulates directional cell migration and cell proliferation via Ror2-mediated noncanonical pathway in mammalian palate development. *Development* 135:3871-3879.
- Hoon JL, Li HY, Koh CG. 2014. POPX2 phosphatase regulates cell polarity and centrosome placement. *Cell Cycle* 13:2459-2468.
- Ira G, Pelliccioli A, Balijja A, Wang X, Fiorani S, Carotenuto W, Liberi G, Bressan D, Wan L, Hollingsworth NM *et al.* 2004. DNA end resection, homologous recombination and DNA damage checkpoint activation require CDK1. *Nature* 431:1011-1017.
- Janisch KM, Vock VM, Fleming MS, Shrestha A, Grimsley-Myers CM, Rasoul BA, Neale SA, Cupp TD, Kinchen JM, Liem KF, Jr. *et al.* 2013. The vertebrate-specific Kinesin-6, Kif20b, is required for normal cytokinesis of polarized cortical stem cells and cerebral cortex size. *Development* 140:4672-4682.
- Jimbo T, Kawasaki Y, Koyama R, Sato R, Takada S, Haraguchi K, Akiyama T. 2002. Identification of a link between the tumour suppressor APC and the kinesin superfamily. *Nat Cell Biol* 4:323-327.
- Kobayashi GS, Alvizi L, Sunaga DY, Francis-West P, Kuta A, Almada BV, Ferreira SG, de Andrade-Lima LC, Bueno DF, Raposo-Amaral CE *et al.* 2013. Susceptibility to DNA damage as a molecular mechanism for non-syndromic cleft lip and palate. *PLoS One* 8:e65677.
- Kurima K, Hertzano R, Gavrilova O, Monahan K, Shpargel KB, Nadaraja G, Kawashima Y, Lee KY, Ito T, Higashi Y *et al.* 2011. A noncoding point mutation of Zeb1 causes multiple developmental malformations and obesity in Twirler mice. *PLoS Genet* 7:e1002307.
- Lamouille S, Xu J, Derynck R. 2014. Molecular mechanisms of epithelial-mesenchymal transition. *Nat Rev Mol Cell Biol* 15:178-196.
- Le Pabic P, Ng C, Schilling TF. 2014. Fat-Dachsous signaling coordinates cartilage differentiation and polarity during craniofacial development. *PLoS Genet* 10:e1004726.

- Lek M, Karczewski K, Minikel E, Samocha K, Banks E, Fennel T, O'Donnell-Luria A, Ware J, Hill A, Cummings B *et al.* 2015. Analysis of protein-coding genetic variation in 60,706 humans. *bioRxiv*.
- Leslie EJ, Marazita ML. 2013. Genetics of cleft lip and cleft palate. *Am J Med Genet C Semin Med Genet* 163C:246-258.
- Leslie EJ, Murray JC. 2012. Evaluating rare coding variants as contributing causes to non-syndromic cleft lip and palate. *Clin Genet* 84:496-500.
- Malumbres M. 2014. Cyclin-dependent kinases. *Genome Biol* 15:122.
- Marazita ML, Murray JC, Lidral AC, Arcos-Burgos M, Cooper ME, Goldstein T, Maher BS, Daack-Hirsch S, Schultz R, Mansilla MA *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.
- Peinado H, Olmeda D, Cano A. 2007. Snail, Zeb and bHLH factors in tumour progression: an alliance against the epithelial phenotype? *Nat Rev Cancer* 7:415-428.
- Phang HQ, Hoon JL, Lai SK, Zeng Y, Chiam KH, Li HY, Koh CG. 2014. POPX2 phosphatase regulates the KIF3 kinesin motor complex. *J Cell Sci* 127:727-739.
- Schork NJ, Murray SS, Frazer KA, Topol EJ. 2009. Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev* 19:212-219.
- Schwabe GC, Trepczik B, Suring K, Brieske N, Tucker AS, Sharpe PT, Minami Y, Mundlos S. 2004. Ror2 knockout mouse as a model for the developmental pathology of autosomal recessive Robinow syndrome. *Dev Dyn* 229:400-410.
- Sebbagh M, Borg JP. 2014. Insight into planar cell polarity. *Exp Cell Res* 328:284-295.
- Seto-Salvia N, Stanier P. 2014. Genetics of cleft lip and/or cleft palate: association with other common anomalies. *Eur J Med Genet* 57:381-393.
- Sun D, Baur S, Hay ED. 2000. Epithelial-mesenchymal transformation is the mechanism for fusion of the craniofacial primordia involved in morphogenesis of the chicken lip. *Dev Biol* 228:337-349.
- Susila A, Chan H, Loh AX, Phang HQ, Wong ET, Tergaonkar V, Koh CG. 2010. The POPX2 phosphatase regulates cancer cell motility and invasiveness. *Cell Cycle* 9:179-187.
- Takagi T, Moribe H, Kondoh H, Higashi Y. 1998. DeltaEF1, a zinc finger and homeodomain transcription factor, is required for skeleton patterning in multiple lineages. *Development* 125:21-31.
- Tanuma N, Nomura M, Ikeda M, Kasugai I, Tsubaki Y, Takagaki K, Kawamura T, Yamashita Y, Sato I, Sato M *et al.* 2009. Protein phosphatase Dusp26 associates with KIF3 motor and promotes N-cadherin-mediated cell-cell adhesion. *Oncogene* 28:752-761.
- Topczewski J, Dale RM, Sisson BE. 2011. Planar cell polarity signaling in craniofacial development. *Organogenesis* 7:255-259.
- Vogelaar IP, Figueiredo J, van Rooij IA, Simoes-Correia J, van der Post RS, Melo S, Seruca R, Carels CE, Ligtenberg MJ, Hoogerbrugge N. 2013. Identification of germline mutations in the cancer predisposing gene CDH1 in patients with orofacial clefts. *Hum Mol Genet* 22:919-926.
- Wang H, Hetmanski JB, Ruczinski I, Liang KY, Fallin MD, Redett RJ, Raymond GV, Chou YH, Chen PK, Yeow V *et al.* 2012. ROR2 gene is associated with risk of non-syndromic cleft palate in an Asian population. *Chin Med J (Engl)* 125:476-480.
- Wu D, Mandal S, Choi A, Anderson A, Prochazkova M, Perry H, Gil-Da-Silva-Lopes VL, Lao R, Wan E, Tang PL *et al.* 2014. DLX4 is associated with orofacial clefting and abnormal jaw development. *Hum Mol Genet* 24:4340-4352.
- Xu J, Lamouille S, Derynck R. 2009. TGF-beta-induced epithelial to mesenchymal transition. *Cell Res* 19:156-172.
- Yang T, Jia Z, Bryant-Pike W, Chandrasekhar A, Murray JC, Fritzsche B, Bassuk AG. 2013. Analysis of PRICKLE1 in human cleft palate and mouse development demonstrates rare and common variants involved in human malformations. *Mol Genet Genomic Med* 2:138-151.

Supplementary information

Supplementary Figure 1. Validation of NGS by Sanger Sequencing.

Supplementary Table 1. Primer sequences for variant validation.

Supplementary Table 2. Variant prioritized after quality and frequency filters.

Supplementary Table 1 – Primer sequences for variant validation

Gene*	Product Size	T _A (°C)	Sequences (5'-3')	
<i>CDH1</i>	312	57	F	CTTCCTCATCAGAGCTCAAGT
			R	CAAGAAGTTCTGTCCGTAGGA
<i>CDK1</i>	356	54	F	TCTTTAGTTTGTGGGGTGTGTC
			R	GTGCGGCATTCTCAACTACC
<i>IGF2R</i>	557	54	F	TTCTGGTGACTCCTCACGTC
			R	TTAGCCTCTGCTAAGTGGGG
<i>KIF20B exon29</i>	496	54	F	ACATACCATGTGTGAGCCTGAA
			R	TCTCTAACAGCAGTGCCTATGAA
<i>KIF20B exon32</i>	569	54	F	CAAGTCATTAAAGATGCCTGAGGTT
			R	AACACATGTGAGACTAAAAGCATGA
<i>KIFAP3</i>	597	54	F	TGGGAAATAGACATTCTTTCTTGGA
			R	AGAGAGAAGGTACCAAAGTGTT
<i>PAX8</i>	548	54	F	TGCACTCTCACTCCCTGAC
			R	AGGACTGACTCCAAGCTGAC
<i>PPM1F</i>	420	54	F	CTGACATTTCTCCCAGGGT
			R	ACATGGTGCTGCATAGGGAT
<i>PRICKLE1</i>	479	57	F	ATTTTAAATTCAGGGGGCCTG
			R	GAAGAATAAGGACTGAGGAGGGT
<i>ROR2</i>	480	54	F	CTGCGGTGACAGTGATGTTTC
			R	GGATAGGTAATCCATCCCCG
<i>ZEB1</i>	481	54	F	CCACACGACCACAGATACG
			R	CTAGGCTGCTCAAGACTGTAG

T_a: Annealing temperature; **F:** Forward; **R:** Reverse

*Genes harboring candidate variants

Supplementary Table 2 – Variant prioritized after quality and frequency filters.

NEWLY DESCRIBED VARIANTS							
Family	Gene	Variant	Type	Chr	Location	LJB23_SIFT	LJB23_PolyPhen2HumDiv
F617	<i>ECT2L</i>	NM_001195037:c.G2224C:p.D742H	nonsynonymous SNV	6	139206932	0.722063,NA	0.008,B
	<i>PHACTR2</i>	NM_001100164:c.G58A:p.D20N	nonsynonymous SNV	6	144033164	1,D	0.98,D
	<i>CDH1</i>	NM_004360:c.G760A:p.D254N	nonsynonymous SNV	16	68844172	NA	NA
F1843	<i>FHDC1</i>	NM_033393:c.G223T:p.A745S	nonsynonymous SNV	4	153896676	0.74,0.26,T	0.031,B
	<i>IFFO1</i>	NM_001039670:c.C421G:p.L141V	nonsynonymous SNV	12	6664775	0.38,0.62,T	0.062,B
	<i>KIZ</i>	na	nonsynonymous SNV	20	21143073	NA	NA
	<i>PPM1F</i>	NM_014634:c.C19T:p.Q7X	stopgain	22	22300402	0.02,0.98,D	NA
F2570	<i>PAX8</i>	NM_003466:c.G550C:p.G184R	nonsynonymous SNV	2	113999636	0.000000,D	1.0,D
	<i>ERBB4</i>	NM_001042599:c.G3379A:p.G1127R	nonsynonymous SNV	2	212251632	0.170000,T	0.928,P
	<i>PGAM5</i>	NM_001170543:c.T491G:p.L164R	nonsynonymous SNV	12	133294145	0.000000,D	1.0,D
	<i>PLCB2</i>	NM_004573:c.A546C:p.E182D	nonsynonymous SNV	15	40594194	0.020000,D	0.986,D
	<i>MGA</i>	NM_001080541:c.C335T:p.A1118V	nonsynonymous SNV	15	42005617	0.130000,T	0.039,B
	<i>DMXL2</i>	NM_001174116:c.A109C:p.I37L	nonsynonymous SNV	15	51868357	0.050000,D	0.314,B
	<i>CGNL1</i>	NM_032866:c.T3139C:p.Y1047H	nonsynonymous SNV	15	57820951	0.380000,T	0.002,B
F2848	<i>KIR2DL2</i>	NM_014219:c.G563A:p.G188D	nonsynonymous SNV	19	27360	NA	NA
	<i>FBXO2</i>	NM_012168:c.367_368insGCAACC:p.L123delinsRNL	nonframeshift insertion	1	11710546	NA	NA
	<i>LEPR</i>	NM_002303:c.G2902A:p.E968K	nonsynonymous SNV	1	66102102	0.48,0.52,T	0.131,B
	<i>C1orf229</i>	NM_207401:c.G146T:p.R49L	nonsynonymous SNV	1	247275381	0.1.00,D	0.876,P
	<i>CMKP2</i>	NM_001256478:c.G908A:p.R303K	nonsynonymous SNV	2	7001399	0.49,0.51,T	0.767,P
	<i>CAD</i>	NM_00434:c.G6568A:p.E2190K	nonsynonymous SNV	2	27466153	0.1.00,D	1.0,D
	<i>G6PC2</i>	NM_001081686:c.G184A:p.V62I	nonsynonymous SNV	2	169758025	0.2,0.80,T	1.0,D
	<i>C2orf72</i>	NM_001144994:c.C397G:p.R133G	nonsynonymous SNV	2	231902677	0.01,0.99,D	0.993,D
	<i>ITIH3</i>	NM_002217:c.T815C:p.F272S	nonsynonymous SNV	3	52833413	0.05,0.95,D	1.0,D
	<i>TYRP1</i>	NM_000550:c.C785G:p.T262R	nonsynonymous SNV	9	12698527	0.1.00,D	0.998,D
	<i>ZEB1</i>	NM_001174093:c.A1213G:p.I405Vp.I408V	nonsynonymous SNV	10	31809536	0.52,0.48,T	0.993,D
	<i>PLCE1</i>	NM_001165979:c.T2477C:p.V826A	nonsynonymous SNV	10	96014653	1,0.00,T	0.0,B
	<i>FRAT1</i>	NM_005479:c.G649C:p.V217L	nonsynonymous SNV	10	99079859	0.1.00,D	0.967,D
	<i>CNNM1</i>	NM_020348:c.G2536A:p.D846N	nonsynonymous SNV	10	101147920	1,0.00,T	0.006,B
	<i>ANKK1</i>	NM_178510:c.A2039C:p.E680A	nonsynonymous SNV	11	113270730	0.45,0.55,T	0.032,B
	<i>ARF3</i>	NM_001659:c.G290A:p.R97Q	nonsynonymous SNV	12	49333532	0.1.00,D	0.68,P
	<i>SCN8A</i>	NM_001177984:c.G76A:p.E26K	nonsynonymous SNV	12	52056677	0.02,0.98,D	0.982,D
	<i>TMEM19</i>	NM_018279:c.A514G:p.M172V	nonsynonymous SNV	12	72091191	0.04,0.96,D	0.101,B
	<i>FAM109A</i>	NM_001177997:c.G20A:p.S7N	nonsynonymous SNV	12	111801212	0.06,0.94,T	0.941,P
	<i>ARID4A</i>	NM_002892:c.A629C:p.Q210P	nonsynonymous SNV	14	58795001	0.02,0.98,D	0.998,D
	<i>PIGQ</i>	NM_148920:c.G1866T:p.W622C	nonsynonymous SNV	16	633217	NA	0.933,P
	<i>MYLK3</i>	NM_182493:c.G488A:p.R163Q	nonsynonymous SNV	16	46774049	0.56,0.44,T	0.026,B
	<i>C17orf107</i>	NM_001145536:c.C403G:p.Q135E	nonsynonymous SNV	17	4803658	0.1.00,D	0.007,B

	EMR2	NM_001271052:c.T914C:p.L305S	nonsynonymous SNV	19	14875415	0.02,0.98,D	0.529,P				
	KCNA7	NM_031886:c.214_216del:p.72_72del	nonframeshift deletion	19	49575627	NA	NA				
	BPIFA1	NM_001243193:c.T726G:p.I242M	nonsynonymous SNV	20	31829921	0.02,0.98,D	0.93,P				
F3196	TTLL3	NM_001025930:c.C352G:p.P118A	nonsynonymous SNV	3	9852042	0.690000,T	NA				
	PDE6B	NM_001145292:c.G1061A:p.G354E	nonsynonymous SNV	4	656954	0.070000,T	0.948,P				
	FAM26F	NM_001010919:c.G184C:p.A62P	nonsynonymous SNV	6	116783276	0.010000,D	0.997,D				
	IGF2R	NM_000876:c.C2096T:p.S699L	nonsynonymous SNV	6	160468235	0.020000,D	0.999,D				
	ABI1	NM_001012750:c.G445A:p.V149M	nonsynonymous SNV	10	27066011	0.020000,D	1.0,D				
	ZMIZ1	NM_020338:c.A2167G:p.T723A	nonsynonymous SNV	10	81063813	0.640000,T	0.0,B				
	C2CD3	NM_015531:c.A2428G:p.M810V	nonsynonymous SNV	11	73814328	0.060000,T	0.982,D				
	VWF	NM_000552:c.G7363T:p.D2455Y	nonsynonymous SNV	12	6085351	0.000000,D	1.0,D				
	PRICKLE1	NM_001144881:c.2149delA:p.S717fs	frameshift deletion	12	42853958	NA	NA				
	TEX29	NM_152324:c.G139T:p.E47X	stopgain SNV	13	111980610	1,T	NA				
	PCNX	NM_014982:c.T6233G:p.I2078S	nonsynonymous SNV	14	71572089	0.780000,T	0.0,B				
	CCDC78	NM_001031737:c.G50C:p.R17P	nonsynonymous SNV	16	776318	0.300000,T	0.003,B				
	STX10	NM_003765:c.G35delT:p.V212fs	frameshift deletion	19	13255429	NA	NA				
	ZNF98	NM_001098626:c.G1571A:p.G524D	nonsynonymous SNV	19	22574466	0.080000,T	0.901,P				
	SSC5D	NM_001144950:c.C3824T:p.P1275L	nonsynonymous SNV	19	56029467	0.020000,D	0.0,B				
	PTPN1	NM_002827:c.C1174T:p.P392S	nonsynonymous SNV	20	49197887	0.530000,T	0.001,B				
	PANX2	NM_052839:c.C1967A:p.A656D	nonsynonymous SNV	22	50617639	NA	0.037,B				
F3788	CDH1	NM_004360:c.G760A:p.D254N	nonsynonymous SNV	16	68844172	NA	NA				
F7614	GJA8	NM_005267:c.T18delG:p.V240X	stopgain SNV	1	147380800	NA	NA				
	KIFAP3	NM_001204516:c.G1142C:p.C381S	nonsynonymous SNV	1	169961306	NA	0.379,B				
	MTUS2	NM_001033602:c.A2623G:p.S875G	nonsynonymous SNV	13	29675056	NA	0.971,D				
	ALDOA	NM_001127617:c.G1087A:p.A363T	nonsynonymous SNV	16	30081525	0.330000,T	0.03,B				
	RSPH1	NM_080860:c.265_274del:p.89_92del	frameshift deletion	21	43912868	NA	NA				
	RSPH1	NM_080860:c.262delT:p.S88fs	frameshift deletion	21	43912880	NA	NA				
F8418	RSPH1	NM_080860:c.245_259del:p.82_87del	nonframeshift deletion	21	43912883	NA	NA				
	C8B	NM_000066:c.T112C:p.F38L	nonsynonymous SNV	1	57425830	0.64,0.36,T	0.0,B				
	DEPDC1	NM_017779:c.T1097A:p.F366Y	nonsynonymous SNV	1	68944990	0.01,0.99,D	1.0,D				
	TRANK1	NM_014831:c.C1409T:p.T470M	nonsynonymous SNV	3	36900340	0.31,0.69,T	NA				
	CDK1	NM_001170406:c.G88A:p.V30I	nonsynonymous SNV	10	62544513	0.17,0.83,T	0.999,D				
	SMTNL1	NM_001105565:c.G539T:p.G180V	nonsynonymous SNV	11	57310654	0.07,0.93,T	0.007,B				
VWF	NM_000552:c.A440G:p.Q147R	nonsynonymous SNV	12	6219632	0.27,0.73,T	0.976,D					
VARIANTS PREVIOUSLY REPORTED IN DATABASES											
Family	Gene / Variant	rs name*	Type	Chr	Location	LB23_SIFT	LB23 PolyPhen2 HD	ExAC	ESP6500	1kGP	CEGH60+
F617	VTA1 / NM_016485:c.C22A:p.P8T	rs137966407	nonsynonymous	6	142468446	0.01,0.99,D	0.78,NA	NA	0.000077	0	0.002463054
	PLA2G15 / NM_012320:c.C92T:p.A31V	rs142099674	nonsynonymous	16	68279421	0.06,0.94,T	0.012,B	NA	0.000154	0	0.001642036
	ZNF628 / NM_033113:c.C2558A:p.T853K	rs151142470	nonsynonymous	19	55995118	NA	NA	NA	0.000538	0	0.00410509

F1843	COL11A1 / NM_080630:exon37:c.C2573A:p.P858Q	rs78046647	nonsynonymous	1	103428312	0.14,0.86,T	1.0,D	1.85E-03	0.002307	0.001198	0.003284072
	MF12 / NM_005929:exon15:c.C2119T:p.Q707X	NA	stopgain	3	196730790	0.99,0.01,T	NA	8.13E-06	0	0	0.000821018
	GPR78 / NM_080819:exon1:c.A275G:p.N92S	rs376898264	nonsynonymous	4	8582984	0.15,0.85,T	0.961,D	8.25E-06	0.000077	0	0
	KCTD20 / NM_173562:exon3:c.C359T:p.T120M	rs146111204	nonsynonymous	6	36442764	0.1,0,D	1.0,D	2.14E-03	0.000846	0.001198	0.000821018
	PON2 / NM_000305:exon6:c.A692G:p.D231G	NA	nonsynonymous	7	95039216	1,0,0,T	0,0,B	2.44E-05	0	0	0
	TSPYL5 / NM_033512:exon1:c.A835G:p.S279G	rs151015596	nonsynonymous	8	98289238	0.04,0.96,D	0.051,B	3.54E-03	0.00469	0.001597	0.005747126
	COX6C / NM_004374:exon3:c.G169A:p.D57N	NA	nonsynonymous	8	100899792	0.04,0.96,D	0.944,P	3.25E-05	0	0	0
	FAM83A / NM_001288587:exon1:c.C263T:p.A88V	rs34007285	nonsynonymous	8	124195359	0.28,0.72,T	0.131,B	6.42E-04	0.001153	0.0001996	0
	RIC1 / NM_001135920:exon11:c.G1175T:p.R392M	rs138987836	nonsynonymous	9	5746010	0.1,0.90,T	0.216,B	7.32E-04	0.000769	0.0003993	0.003284072
	EHMT1 / NM_001145527:exon3:c.G154C:p.G52R	NA	nonsynonymous	9	140611146	0.16,0.84,T	0.991,D	1.63E-05	0	0	0
	TRPM5 / NM_014555:exon7:c.C933A:p.H311Q	rs138197888	nonsynonymous	11	2439033	0.66,0.34,T	0.99,D	1.46E-03	0.001618	0.0007987	0.002467105
	TRIM6 / NM_001198645:exon6:c.C925T:p.R309C	rs151311625	nonsynonymous	11	5632555	1,0,0,T	0.061,B	8.13E-04	0.00177	0	0.000821018
	ATG2A / NM_015104:exon17:c.C2399T:p.T800M	NA	nonsynonymous	11	64675328	0.01,0.99,D	0.606,P	2.44E-05	0	0	0
	NCAPD2 / NM_014865:exon4:c.G212A:p.R71Q	rs200920120	nonsynonymous	12	6619249	0.56,0.44,T	0.99,D	4.07E-05	0.000154	0	0
	SV2B / NM_001167580:exon8:c.A904G:p.K302E	rs146893681	nonsynonymous	15	91811819	1,0,0,T	0,0,B	7.97E-04	0.002232	0.002595	0.00410509
	KLHDC4 / NM_001184854:exon3:c.G265A:p.G89R	rs141582040	nonsynonymous	16	87782349	0.23,0.77,T	0.995,D	1.21E-03	0.003616	0.003793	0.003284072
	CDK10 / NM_001098533:exon9:c.G422A:p.G141E	NA	nonsynonymous	16	89760607	0.03,0.97,D	0.978,D	3.25E-05	0	0	0.001642036
	GAA / NM_000152:exon16:c.G2238C:p.W746C	rs1800312	nonsynonymous	17	78090815	0.1,0,D	1.0,D	2.93E-04	0.000308	0.0003993	0.002463054
	FSCN2 / NM_001077182:exon1:c.A340G:p.T114A	rs202158770	nonsynonymous	17	79495897	0.05,0.95,D	0.026,B	5.05E-04	0.000803	0.0001996	0.001642036
	ATP5A1 / NM_001257334:exon1:c.G25A:p.A9T	rs141639003	nonsynonymous	18	43678173	0.17,0.83,T	0,0,B	1.24E-03	0.000157	0.002595	0.00247117
	GPI / NM_001289790:exon9:c.A754G:p.I252V	NA	nonsynonymous	19	34884186	0.02,0.98,D	0.73,P	8.13E-06	0	0	0
	SPRED3 / NM_001042522:exon4:c.C553A:p.R185S	rs112201791	nonsynonymous	19	38885412	0.64,0.36,T	0.256,B	7.26E-04	0.001192	0.0003993	0.005794702
	ZNF337 / NM_001290261:exon4:c.1174_1176del:p.392_392del	NA	nonframeshift deletion	20	25656748	NA	NA	2.60E-04	0.00024	0	0.000821018
	DLGAP4 / NM_001042486:exon4:c.C378A:p.N126K	NA	nonsynonymous	20	35129001	0.23,0.77,T	0.822,P	8.32E-06	0	0	0
	KCNS1 / NM_002251:exon4:c.C206A:p.A69E	NA	nonsynonymous	20	43727207	1,0,0,T	0.008,B	2.19E-04	0	0.002595	0.000830565
	ZFP64 / NM_022088:exon5:c.G1420A:p.A474T	rs143029301	nonsynonymous	20	50769149	0.6,0.40,T	0.001,B	8.24E-05	0.000231	0.000199	0.001642036
	SLC2A11 / NM_001024938:exon9:c.G997A:p.E333K	rs79905160	nonsynonymous	22	24225971	0.1,0,D	0.999,D	3.86E-03	0.003537	0.002196	0.003284072
	SUSD2 / NM_019601:exon12:c.G1927A:p.D643N	rs114116915	nonsynonymous	22	24583574	0.1,0.90,T	1.0,D	1.30E-04	0	0.0005990	0
	TRABD / NM_025204:exon3:c.C74T:p.P25L	rs6010154	nonsynonymous	22	50632045	0.27,0.73,T	0,0,B	4.31E-04	0.000615	0.001996	0
	HDAC10 / NM_001159286:exon17:c.G1597A:p.G533S	rs138919111	nonsynonymous	22	50684515	0.26,0.74,T	0.911,P	6.92E-04	0.002077	0.002196	0.002463054
F2570	PEX10 / NM_002617:exon3:c.C427T:p.R143C	rs199667764	nonsynonymous	1	2340064	0.06,T	0.998,D	NA	0	0.0009	0.000821018
	SLC39A10 / NM_001127257:exon6:c.A1597C:p.K533Q	rs202179859	nonsynonymous	2	196578178	0.53,T	0.801,P	NA	0	0	0.000821018
	BOLL / NM_033030:exon10:c.C800T:p.A267V	NA	nonsynonymous	2	198607813	0.0,D	0.982,D	NA	0.000077	0	0
	GPD1L / NM_015141:exon4:c.A467G:p.N156S	rs139494055	nonsynonymous	3	32181820	0.56,T	0,0,B	NA	0.000154	0	0
	DZIP1L / NM_173543:exon16:c.C2233T:p.R745C	rs146029706	nonsynonymous	3	137781729	0.04,D	0.013,B	NA	0.002384	0.0009	0.000821018
	IGSF10 / NM_178822:exon4:c.T1991C:p.M664T	rs151226002	nonsynonymous	3	151165778	0.57,T	0,0,B	NA	0.004998	0.0037	0.002463054
	GPR149 / NM_001038705:exon1:c.C836A:p.A279E	NA	nonsynonymous	3	154146569	0.26,T	0.221,B	NA	0	0	0
	PLCH1 / NM_014996:exon22:c.G3607A:p.G1203S	rs148407866	nonsynonymous	3	155200118	0.31,T	0.002,B	NA	0.000692	0	0
	RNF212 / NM_001131034:exon6:c.T363G:p.S121R	rs143046808	nonsynonymous	4	1075407	0.76,T	0.324,B	NA	0.000077	0.0041	0.003284072
	REPS1 / NM_001128617:exon6:c.T803G:p.I268S	rs35796154	nonsynonymous	6	139265103	0.73,T	0.999,D	NA	0.000538	0	0
	MYOM2 / NM_003970:exon14:c.G1617C:p.K539N	rs137923713	nonsynonymous	8	2033495	0.06,T	0.171,B	NA	0.00223	0.0005	0.002463054
	PRSS55 / NM_198464:exon5:c.G834T:p.E278D	rs1133417	nonsynonymous	8	10396078	0.13,T	0.553,P	NA	0.003537	0.0018	0.002463054

	PAPPA / NM_002581:exon19:c.T4537C:p.S1513P	rs142592736	nonsynonymous	9	119129965	1,T	0.001,B	NA	0.000923	0	0.000821018
	FIBCD1 / NM_032843:exon2:c.C364T:p.H122Y	NA	nonsynonymous	9	133805142	0.09,T	0.005,B	NA	0.000161	0	0.000821018
	SETX / NM_015046:exon10:c.A3568G:p.K1190E	rs35473230	nonsynonymous	9	135203417	0.85,T	0.0,B	NA	0.003383	0.0046	0.001642036
	ADAMTS13 / NM_139025:exon25:c.C3445T:p.R1149W	rs141494468	nonsynonymous	9	136320602	NA	0.484,P	NA	0.002932	0.0009	0.003284072
	GLT6D1 / NM_182974:exon5:c.C566T:p.P189L	NA	nonsynonymous	9	138516208	0.07,T	0.587,P	NA	0.000249	0	0
	ADAMTS14 / NM_080722:exon22:c.G3520A:p.G1174R	rs61754838	nonsynonymous	10	72520457	0.74,T	0.001,B	NA	0.002537	0.0014	0.005747126
	SEC24C / NM_198597:exon5:c.C763G:p.P255A	rs150235476	nonsynonymous	10	75520057	0.48,T	0.835,P	NA	0.00469	0.0032	0.003284072
	ANKRD22 / NM_144590:exon3:c.C257T:p.T86I	rs139978877	nonsynonymous	10	90588380	0.2,T	0.986,D	NA	0.001076	0.0014	0
	KIF20B / NM_016195:exon29:c.A4835G:p.K1612R	rs143235231	nonsynonymous	10	91522558	0.27,T	0.058,B	NA	0.001	0.0009	0
	KIF20B / NM_016195:exon32:c.G5263A:p.V1755M	rs145242589	nonsynonymous	10	91532586	0.03,D	1.0,D	NA	0.002307	0.0009	0.000821018
	DCLRE1A / NM_014881:exon4:c.A2291G:p.H764R	rs35884667	nonsynonymous	10	115605531	1,T	0.0,B	NA	0.003076	0.0041	0.005747126
	BAG3 / NM_004281:exon2:c.C187G:p.P63A	rs144041999	nonsynonymous	10	121429369	0.77,T	0.0,B	NA	0.003767	0.0032	0.004926108
	BAG3 / NM_004281:exon4:c.C1138T:p.P380S	rs144692954	nonsynonymous	10	121436204	0.18,T	0.172,B	NA	0.003691	0.0032	0.004926108
	SEC23IP / NM_007190:exon13:c.C2153T:p.A718V	rs150192084	nonsynonymous	10	121685579	0.29,T	0.0,B	NA	0.003921	0.0037	0.003284072
	PIK3C2G / NM_004570:exon15:c.G2144C:p.R715T	rs138567313	nonsynonymous	12	18552733	0.38,T	0.053,B	NA	0.000925	0.0014	0.001642036
	GPR133 / NM_198827:exon7:c.C794A:p.S265Y	rs137909892	nonsynonymous	12	131475607	0.01,D	0.511,P	NA	0.004383	0.0027	0.003284072
	DUOX1 / NM_175940:exon19:c.C2383T:p.R795W	rs146747045	nonsynonymous	15	45439691	0.00,D	0.292,B	NA	0.000077	0.0014	0.002463054
	CGNL1 / NM_032866:exon13:c.T3139C:p.Y1047H	NA	nonsynonymous	15	57820951	0.38,T	0.002,B	NA	0	0	0
	VPS13C / NM_017684:exon61:c.C8582T:p.S2861L	rs115869241	nonsynonymous	15	62204043	0.24,T	0.868,P	NA	0.002768	0.0027	0.000821018
	C15orf39 / NM_015492:exon2:c.C58T:p.R20C	rs143806224	nonsynonymous	15	75498447	0.00,D	1.0,D	NA	0.002311	0.0023	0.003284072
	SEC11A / NM_014300:exon3:c.G258T:p.R86S	NA	nonsynonymous	15	85230909	0.10,T	0.135,B	NA	0	0	0
	FAM20A / NM_017565:exon1:c.G193A:p.G65S	rs143371801	nonsynonymous	17	66596615	0.87,T	0.006,B	NA	0.00054	0	0
	C17orf80 / NM_001100621:exon3:c.G889A:p.E297K	rs115090235	nonsynonymous	17	71232510	0.28,T	0.989,D	NA	0.002614	0.0014	0.000821018
	GRIN2C / NM_000835:exon2:c.T348G:p.H116Q	rs144762326	nonsynonymous	17	72850884	0.32,T	0.42,B	NA	0.000154	0.0005	0
	ARMC7 / NM_024585:exon3:c.G577A:p.V193M	rs139709035	nonsynonymous	17	73125113	0.12,T	0.733,P	NA	0.001697	0.0041	0.001642036
	EMILIN2 / NM_032048:exon8:c.C3157T:p.L1053F	rs116171748	nonsynonymous	18	2913397	0.03,D	0.976,D	NA	0.000769	0.0005	0
	PHLPP1 / NM_194449:exon9:c.G2750A:p.R917Q	NA	nonsynonymous	18	60582187	0.55,T	0.93,P	NA	0	0	0
	ARHGEF1 / NM_198977:exon20:c.A1820C:p.N607T	rs147783681	nonsynonymous	19	42408188	0.01,D	0.335,B	NA	0.000538	0.0005	0.001642036
	CEACAM1 / NM_001184813:exon7:c.C1165T:p.P389S	rs146856369	nonsynonymous	19	43015049	0.44,T	0.738,P	NA	0.000461	0	0.001642036
	XRCC1 / NM_006297:exon4:c.G361A:p.A121T	rs138284081	nonsynonymous	19	44058851	0.16,T	0.999,D	NA	0.000615	0.0009	0.000821018
	ZNF225 / NM_013362:exon5:c.C751T:p.R251C	rs62640902	nonsynonymous	19	44635518	0.09,T	0.007,B	NA	0.000308	0	0.001642036
	ZNF233 / NM_001207005:exon5:c.609dupA:p.V203fs	NA	frameshift insertion	19	44777421	0.71,T	NA	NA	0.001917	0	0.003284072
	ZFP112 / NM_013380:exon4:c.T455C:p.I152T	rs61733041	nonsynonymous	19	44833855	0.21,T	0.0,B	NA	0.002076	0.0014	0.003284072
	HIF3A / NM_022462:exon5:c.C437T:p.P146L	rs150500130	nonsynonymous	19	46812490	1,T	1.0,D	NA	0.000154	0.0009	0
	CCDC9 / NM_015603:exon5:c.C415T:p.H139Y	rs147315092	nonsynonymous	19	47764049	0.35,T	0.023,B	NA	0.001231	0	0
	GPR77 / NM_018485:exon2:c.G56T:p.R19L	rs150108068	nonsynonymous	19	47844112	0.72,T	0.0,B	NA	0.000231	0.0009	0
	HRC / NM_002152:exon1:c.G1357A:p.G453S	rs77728314	nonsynonymous	19	49657138	0.00,D	0.128,B	NA	0.004613	0.0027	0.002463054
	ALDH16A1 / NM_001145396:exon8:c.G977A:p.R326H	rs113108384	nonsynonymous	19	49967142	0.41,T	0.927,P	NA	0.002922	0.0037	0
	LILRB5 / NM_001081443:exon4:c.C565T:p.H189Y	rs685082	nonsynonymous	19	54759236	0.00,D	0.925,P	NA	0.001384	0.0005	0.000821018
	KIR2DL1 / NM_014218:exon8:c.A997G:p.T333A	rs201644346	nonsynonymous	19	55295215	0.00,D	0.017,B	NA	0	0	0
	DONSON / NM_017613:exon9:c.G1411A:p.E471K	rs140592434	nonsynonymous	21	34951808	NA	0.997,D	NA	0.000308	0.0009	0.000821018
F2848	KLHL17 / NM_198317:exon11:c.G1579A:p.V527M	rs199893924	nonsynonymous	1	899789	0.1,00,D	0.999,D	3.94E-04	0.000476	0.0002	0

PLEKHM2 / NM_015164:exon8:c.G891C:p.E297D	rs61738982	nonsynonymous	1	16051990	0.31,0.69,T	0.001,B	3.28E-03	0.004165	0.002396	0.00410509
SPEN / NM_015001:exon11:c.G3064A:p.V1022M	rs115566585	nonsynonymous	1	16255799	0.14,0.86,T	0.004,B	1.81E-03	0.002537	0.002396	0.001642036
PKN2 / NM_006256:exon2:c.G152A:p.R51Q	rs200490316	nonsynonymous	1	89206774	0.45,0.55,T	0.019,B	6.70E-04	0.000676	0	0
KIAA1107 / NM_015237:exon8:c.G1522T:p.V508F	rs200439049	nonsynonymous	1	92646076	0.17,0.83,T	0.834,P	1.17E-03	0	0	0.000821018
CYB561D1 / NM_001134404:exon4:c.A301G:p.S101G	rs202045101	nonsynonymous	1	110038311	0,1.00,D	0.0,B	1.03E-03	0.00219	0.0005990	0.002463054
EPS8L3 / NM_024526:exon4:c.T163A:p.F55I	rs75718950	nonsynonymous	1	110302392	0.51,0.49,T	0.016,B	2.21E-03	0.00246	0.001198	0.003284072
MTMR11 / NM_181873:exon14:c.G1346A:p.R449H	rs145659444	nonsynonymous	1	149902342	0,1.00,D	0.999,D	0	0.004921	0.001397	0.003284072
LYSMD1 / NM_001136543:exon2:c.G251A:p.R84H	rs77292984	nonsynonymous	1	151134362	0.18,0.82,T	0.0,B	1.29E-03	0.003767	0.002595	0.004926108
ASTN1 / NM_004319:exon23:c.C3850G:p.P1284A	rs146353059	nonsynonymous	1	176833455	1.0,0.0,T	0.012,B	1.79E-04	0.000538	0	0
DIEXF / NM_014388:exon11:c.G1903A:p.E635K	rs142546646	nonsynonymous	1	210016917	0.48,0.52,T	0.992,D	8.05E-04	0.001461	0.0007987	0.00410509
TLR5 / NM_003268:exon6:c.T2160A:p.S720R	rs142143294	nonsynonymous	1	223284214	0.13,0.87,T	0.979,D	1.28E-03	0.002614	0.0005990	0.000821018
CCDC88A / NM_018084:exon29:c.T4897C:p.S1633P	rs144009079	nonsynonymous	2	55523504	0.43,0.57,T	0.0,B	1.21E-03	0.000461	0.0003993	0
SPHKAP / NM_001142644:exon7:c.G1987A:p.V663I	rs115375305	nonsynonymous	2	228883583	0.4,0.60,T	0.002,B	1.67E-03	0.001615	0.0005990	0.005747126
VGLL3 / NM_016206:exon3:c.A679C:p.M227L	rs193078775	nonsynonymous	3	87017998	0.2,0.80,T	0.0,B	3.13E-03	0.002706	0.0001996	0
MORC1 / NM_014429:exon21:c.G2155C:p.D719H	rs35276036	nonsynonymous	3	108719436	0.58,0.42,T	0.79,P	3.42E-03	0.00346	0.001397	0.002463054
KALRN / NM_007064:exon2:c.A278G:p.Q93R	rs111238042	nonsynonymous	3	124351459	0.33,0.67,T	0.244,B	1.63E-04	0.000384	0	0
ATR / NM_001184:exon10:c.A2290G:p.K764E	rs77208665	nonsynonymous	3	142274770	0.58,0.42,T	0.919,P	3.25E-03	0.004536	0.001996	0.000821018
IGSF10 / NM_001178145:exon2:c.C69A:p.Y23X	NA	stopgain	3	151156361	0.26,0.74,T	NA	8.14E-06	0	0	0
ADIPOQ / NM_004797:exon2:c.22_24del:p.8_8del	NA	nonframeshift deletion	3	186570869	NA	NA	8.46E-04	0	0.0005990	0
FAM193A / NM_001256666:exon13:c.C1712T:p.P571L	rs146516940	nonsynonymous	4	2692479	0.36,0.64,T	0.167,B	2.05E-03	0.002153	0.0011980	0.00410509
WFS1 / NM_001145853:exon8:c.C2452T:p.R818C	rs35932623	nonsynonymous	4	6303974	0.03,0.97,D	1.0,D	4.74E-03	0.004767	0.0037939	0.00410509
MAN2B2 / NM_001292038:exon9:c.C1186G:p.R396G	rs149169745	nonsynonymous	4	6600015	0.34,0.66,T	0.577,P	1.14E-03	0.000846	0.0013977	0.000821018
QRFPR / NM_198179:exon1:c.C100T:p.L34F	rs139347755	nonsynonymous	4	122301703	0.59,0.41,T	1.0,D	3.91E-03	0.002845	0.0009984	0.000821018
GALNTL6 / NM_001034845:exon12:c.G1582A:p.V528I	rs143941607	nonsynonymous	4	173942720	0.29,0.71,T	0.009,B	1.63E-04	0.000231	0.0001996	0
WWC2 / NM_024949:exon11:c.G1349A:p.R450Q	rs141501417	nonsynonymous	4	184182125	0.32,0.68,T	0.681,P	3.54E-03	0.001845	0.0003993	0
PPIP5K2 / NM_015216:exon20:c.G2510A:p.R837H	NA	nonsynonymous	5	102509657	0,1.00,D	0.995,D	1.46E-04	0	0.0001996	0
FAM53C / NM_001135647:exon4:c.G471C:p.Q157H	rs140173423	nonsynonymous	5	137680848	NA	0.688,P	2.75E-03	0.000692	0.0041932	0.000821018
DIAPH1 / NM_001079812:exon15:c.G1958A:p.G653D	rs200735096	nonsynonymous	5	140953432	0.1,0.90,T	0.001,B	9.59E-04	0.000821	0.0001996	0
WWC1 / NM_001161661:exon21:c.G3140A:p.R1047H	rs145892564	nonsynonymous	5	167891939	0.01,0.99,D	0.999,D	2.44E-05	0.000231	0	0
PRPF4B / NM_003913:exon2:c.G622A:p.V208I	rs200965972	nonsynonymous	6	4032373	0.23,0.77,T	0.0,B	8.95E-05	0	0.0001996	0
PIM1 / NM_001243186:exon1:c.G253C:p.A85P	NA	nonsynonymous	6	37138331	NA	NA	4.89E-05	0	0	0
DNAH8 / NM_001206927:exon53:c.C7774T:p.R2592W	rs113332942	nonsynonymous	6	38843520	0.01,0.99,D	0.999,D	4.80E-04	0.000923	0.0001996	0
MEP1A / NM_005588:exon11:c.T1548A:p.D516E	rs142787710	nonsynonymous	6	46801214	0,1.00,D	0.997,D	3.25E-03	0.004306	0.001797	0.001642036
COL9A1 / NM_078485:exon28:c.G1430A:p.R477Q	rs192467838	nonsynonymous	6	70944597	0.47,0.53,T	1.0,D	4.39E-04	0.000461	0.0003993	0.000821018
MMS22L / NM_198468:exon15:c.2164_2168del:p.F722fs	NA	frameshift deletion	6	97634438	NA	NA	8.95E-05	0.003435	0	0
MAN1A1 / NM_005907:exon10:c.A1330G:p.I444V	rs375324542	nonsynonymous	6	119511045	0.49,0.51,T	0.068,B	8.13E-06	0.000077	0	0
LATS1 / NM_001270519:exon3:c.T452C:p.M151T	rs201939550	nonsynonymous	6	150016254	NA	0.276,B	4.88E-05	0.000231	0	0.000821018
SYTL3 / NM_001242395:exon14:c.A1232G:p.H411R	rs77838934	nonsynonymous	6	159183129	0.11,0.89,T	0.006,B	1.20E-03	0.001692	0.0001996	0
ZSCAN25 / NM_145115:exon4:c.C307T:p.R103W	rs145815306	nonsynonymous	7	99217536	0,1.00,D	0.009,B	1.48E-03	0.001845	0	0
MCPH1 / NM_024596:exon13:c.A2401G:p.S801G	rs45540031	nonsynonymous	8	6479161	0.22,0.78,T	0.003,B	2.92E-03	0.001907	0.001797	0
RGS20 / NM_001286673:exon1:c.G91A:p.A31T	rs144848624	nonsynonymous	8	54764550	0.23,0.77,T	0.001,B	2.88E-03	0.000692	0.003594	0

RPS20 / NM_001023:exon4:c.C356G:p.A119G	NA	nonsynonymous	8	56985653	0,1.00,D	0.014,B	1.64E-05	0	0	0
PLEC / NM_201378:exon32:c.G10390A:p.V3464M	rs375766567	nonsynonymous	8	144993557	0.04,0.96,D	0.999,D	1.63E-05	0.00008	0	0
SMARCA2 / NM_001289396:exon4:c.A695C:p.Q232P	rs143245740	nonsynonymous	9	2039805	0.7,0.30,T	0.0,B	0	0	0.0003993	0.002463054
PPAPDC2 / NM_203453:exon1:c.T530C:p.L177P	rs200674488	nonsynonymous	9	4662905	0,1.00,D	0.999,D	3.25E-05	0.000077	0	0
GNA14 / NM_004297:exon4:c.C512T:p.T171I	rs139345882	nonsynonymous	9	80046318	0,1.00,D	0.984,D	2.43E-03	0.001	0.0007987	0
ROR2 / NM_004560:exon9:c.G1589A:p.R530Q	rs35852786	nonsynonymous	9	94487187	0.02,0.98,D	0.398,B	2.15E-03	0.001615	0.0013977	0.000821018
LARP4B / NM_015155:exon17:c.C1889T:p.A630V	NA	nonsynonymous	10	860722	0.33,0.67,T	0.242,B	4.07E-05	0	0	0
ITIH5 / NM_032817:exon5:c.C599T:p.T200M	rs146565776	nonsynonymous	10	7621895	0.07,0.93,T	0.984,D	1.84E-03	0.002153	0.0011980	0.001642036
COMMD3-BMI1 / NM_001204062:exon1:c.C92A:p.A31E	rs140293380	nonsynonymous	10	22605438	0.01,0.99,D	0.999,D	2.63E-04	0.001182	0.0003993	0.000821018
KIAA1217 / NM_001282767:exon1:c.C47T:p.S16L	rs144874255	nonsynonymous	10	24498169	0.12,0.88,T	0.004,B	1.00E-03	0.000615	0.0003993	0
ZNF485 / NM_145312:exon5:c.A1211G:p.H404R	rs146214560	nonsynonymous	10	44112702	0,1.00,D	1.0,D	3.33E-04	0.000769	0.0003993	0.001642036
IDE / NM_004969:exon2:c.C248T:p.T83M	NA	nonsynonymous	10	94297158	0.02,0.98,D	1.0,D	1.63E-05	0	0	0
BAG3 / NM_004281:exon4:c.T983C:p.V328A	NA	nonsynonymous	10	121436049	1.0,0.0,T	0.0,B	1.63E-05	0	0	0
CUZD1 / NM_022034:exon5:c.G692A:p.R231H	rs144483251	nonsynonymous	10	124596472	0.2,0.80,T	0.988,D	1.64E-03	0.001615	0	0.002463054
AP2A2 / NM_001242837:exon2:c.A88G:p.I30V	rs200802126	nonsynonymous	11	959457	0.16,0.84,T	0.002,B	3.03E-04	0.000256	0	0.000821018
OR56B1 / NM_001005180:exon1:c.T86C:p.I29T	rs145028394	nonsynonymous	11	5757832	0,1.00,D	1.0,D	7.64E-04	0.001154	0.0001996	0.000821018
ALKBH3 / NM_139178:exon7:c.T421C:p.Y141H	NA	nonsynonymous	11	43913641	0,1.00,D	1.0,D	8.13E-05	0	0	0
SLC39A13 / NM_001128225:exon8:c.G833C:p.S278T	rs370516835	nonsynonymous	11	47436374	0.47,0.53,T	0.006,B	2.44E-05	0.000077	0	0
OSBP / NM_002556:exon2:c.A419G:p.N140S	rs376305185	nonsynonymous	11	59378006	0.41,0.59,T	0.014,B	4.07E-05	0.000077	0	0
TMEM216 / NM_001173990:exon4:c.C265T:p.L89F	NA	nonsynonymous	11	61165281	0.01,0.99,D	0.997,D	1.63E-05	0	0	0
RTN3 / NM_201428:exon2:c.G361A:p.V121I	rs144921220	nonsynonymous	11	63486392	0.7,0.30,T	0.001,B	8.95E-05	0.000462	0.0003993	0
OVOL1 / NM_004561:exon2:c.C304T:p.R102C	NA	nonsynonymous	11	65561705	0.18,0.82,T	0.013,B	4.07E-05	0	0	0
PC / NM_022172:exon5:c.G616T:p.V206L	rs147945506	nonsynonymous	11	66638540	0.03,0.97,D	0.603,P	3.33E-03	0.001463	0.0039936	0.003284072
C11orf24 / NM_022338:exon4:c.C725T:p.A242V	rs143548724	nonsynonymous	11	68029738	0.23,0.77,T	0.259,B	4.22E-03	0.003234	0.0023961	0.003284072
PPP6R3 / NM_001164160:exon8:c.G836A:p.R279Q	NA	nonsynonymous	11	68326138	0.11,0.89,T	1.0,D	1.63E-05	0	0	0
PAAF1 / NM_001267806:exon7:c.C379T:p.P127S	rs140276370	nonsynonymous	11	73620635	0.64,0.36,T	1.0,D	2.46E-03	0.002156	0.001198	0
VWF / NM_000552:exon16:c.T2020C:p.Y674H	NA	nonsynonymous	12	6161875	0.54,0.46,T	0.002,B	8.13E-06	0	0	0
RAPGEF3 / NM_001098531:exon2:c.C79T:p.R27W	rs199576694	nonsynonymous	12	48151789	0.03,0.97,D	0.74,P	1.29E-03	0.001159	0	0.001642036
HDAC7 / NM_001098416:exon12:c.G1382A:p.R461Q	rs149671930	nonsynonymous	12	48187337	0.61,0.39,T	1.0,D	5.74E-04	0.000656	0	0.001642036
ANKRD33 / NM_001130015:exon5:c.C416T:p.P139L	rs146634550	nonsynonymous	12	52284521	0,1.00,D	1.0,D	1.22E-04	0.000384	0	0
TBC1D15 / NM_001146213:exon2:c.G94A:p.G32S	NA	nonsynonymous	12	72265913	1.0,0.0,T	0.001,B	8.13E-06	0	0	0.000821018
MSI1 / NM_002442:exon10:c.G682A:p.A228T	rs143961492	nonsynonymous	12	120791153	0.29,0.71,T	0.095,B	2.90E-03	0.003844	0.000599	0.001642036
IPO5 / NM_002271:exon17:c.A1695C:p.Q565H	rs61750356	nonsynonymous	13	98658527	0.15,0.85,T	0.114,B	4.88E-05	0.000077	0.000199	0
DOCK9 / NM_001130049:exon3:c.G263A:p.R88Q	rs199901746	nonsynonymous	13	99582495	NA	1.0,D	1.08E-03	0.001619	0.000599	0
FAM179B / NM_015091:exon1:c.A842G:p.Q281R	rs146515749	nonsynonymous	14	45432466	0.25,0.75,T	0.981,D	3.25E-04	0.000384	0	0
NID2 / NM_007361:exon9:c.G2176A:p.V726M	rs35147930	nonsynonymous	14	52505546	0.05,0.95,D	1.0,D	4.30E-03	0.004229	0.000998	0.000821018
SPTB / NM_000347:exon22:c.A4670G:p.E1557G	rs140648376	nonsynonymous	14	65242015	0.21,0.79,T	0.903,P	1.12E-03	0.001076	0.000399	0.000821018
CGNL1 / NM_032866:exon13:c.A3049G:p.M1017V	rs143875419	nonsynonymous	15	57820861	0.17,0.83,T	0.0,B	4.12E-03	0.003085	0.00219	0.004926108
VPS13C / NM_017684:exon17:c.T1564A:p.S522T	rs141515062	nonsynonymous	15	62277084	0.4,0.60,T	0.015,B	2.60E-04	0.000615	0.0002	0
SNX1 / NM_001242933:exon1:c.G82A:p.A28T	rs200363020	nonsynonymous	15	64388294	0.53,0.47,T	0.784,P	3.13E-04	0.000156	0.000599	0
HAPLN3 / NM_178232:exon4:c.C674T:p.P225L	rs138635787	nonsynonymous	15	89422320	0,1.00,D	0.998,D	5.62E-04	0.000308	0.001198	0.000821018
ANPEP / NM_001150:exon13:c.G1858C:p.V620L	rs143245843	nonsynonymous	15	90342752	0.56,0.44,T	0.005,B	1.80E-03	0.002308	0.0008	0.00410509

	LRRK1 / NM_024652:exon32:c.G5104A:p.G1702S	rs78605716	nonsynonymous	15	101605746	0.09,0.91,T	0.999,D	1.12E-03	0.001193	0.0004	0.002463054
	E4F1 / NM_001288776:exon2:c.G251A:p.R84Q	NA	nonsynonymous	16	2278466	0.01,0.99,D	1.0,D	1.63E-05	0	0.0002	0.000821018
	ZNF597 / NM_152457:exon4:c.C413G:p.T138S	rs374686560	nonsynonymous	16	3487286	0.45,0.55,T	0.1,B	8.13E-06	0.000077	0	0
	BEAN1 / NM_001178020:exon3:c.C226A:p.R76S	rs200706119	nonsynonymous	16	66503705	0.02,0.98,D	NA	4.44E-03	0	0.00499	0.004926108
	HSF4 / NM_001040667:exon5:c.G259A:p.E87K	rs367654370	nonsynonymous	16	67199648	0.05,0.95,D	0.962,D	1.07E-04	0.000327	0	0
	FHOD1 / NM_013241:exon12:c.G1325A:p.R442Q	NA	nonsynonymous	16	67268370	0.39,0.61,T	1.0,D	8.16E-06	0	0	0
	ALOXE3 / NM_001165960:exon7:c.C1096T:p.R366X	rs121434233	stopgain	17	8015495	0.65,0.35,T	NA	1.14E-04	0.000077	0	0
	DNAH9 / NM_001372:exon3:c.C716G:p.P239R	NA	nonsynonymous	17	11513814	0.74,0.26,T	0.169,B	4.88E-05	0	0	0
	KRT16 / NM_005557:exon3:c.C644G:p.T215S	rs147423442	nonsynonymous	17	39767724	0.88,0.12,T	0.002,B	3.41E-03	0.004306	0.0012	0.004926108
	KPNA2 / NM_002266:exon5:c.C494G:p.P165R	rs11545989	nonsynonymous	17	66038392	0.1,0.0,D	1.0,D	3.60E-03	0.004078	0.0001	0
	SLC16A5 / NM_001271765:exon5:c.C1102T:p.L368F	rs149503044	nonsynonymous	17	73096860	0.27,0.73,T	0.719,P	3.09E-03	0.003383	0.0004	0
	TBCD / NM_005993:exon5:c.A553G:p.L185V	NA	nonsynonymous	17	80726413	NA	0.128,B	1.39E-04	0	0	0
	MYOM1 / NM_003803:exon15:c.A2062T:p.T688S	rs188677538	nonsynonymous	18	3135692	0.51,0.49,T	0.927,P	2.73E-03	0.001801	0.00179	0
	TRAPPC8 / NM_014939:exon21:c.G3257A:p.C1086Y	rs201669717	nonsynonymous	18	29435702	0.13,0.87,T	0.307,B	8.13E-06	0	0	0
	ZNF407 / NM_001146189:exon1:c.C499T:p.P167S	NA	nonsynonymous	18	72343474	0.5,0.50,T	0.15,B	1.64E-05	0	0	0
	ZNF516 / NM_014643:exon3:c.G388C:p.G130R	NA	nonsynonymous	18	74154623	0.44,0.56,T	1.0,D	2.83E-04	0	0	0
	DOT1L / NM_032482:exon24:c.G3157A:p.A1053T	rs144165419	nonsynonymous	19	2222325	0.1,0.0,D	0.006,B	1.82E-03	0.000708	0.00279	0.000821018
	LRRCE8 / NM_001268285:exon2:c.G1834A:p.V612M	NA	nonsynonymous	19	7965628	0.08,0.92,T	0.976,D	2.44E-05	0	0	0
	ZFP30 / NM_014898:exon4:c.G65T:p.C22F	rs112995701	nonsynonymous	19	38135582	0.7,0.30,T	0.002,B	1.97E-03	0.001692	0.0004	0
	WDR87 / NM_001291088:exon4:c.1192delG:p.D398fs	NA	frameshift deletion	19	38385151	NA	NA	3.08E-04	0.002286	0	0
	CEACAM6 / NM_002483:exon4:c.G919T:p.G307C	rs146516997	nonsynonymous	19	42266092	0.02,0.98,D	1.0,D	1.30E-03	0.000692	0.000399	0.002463054
	KLK5 / NM_001077492:exon4:c.A514G:p.L172V	rs2232534	nonsynonymous	19	51452193	0.66,0.34,T	0.09,B	3.59E-03	0.004921	0.000599	0.003284072
	TARM1 / NM_001135686:exon4:c.G386A:p.R129Q	rs376721355	nonsynonymous	19	54577444	0.18,0.82,T	0.0,B	2.95E-04	0.000438	0.000599	0
	CNOT3 / NM_014516:exon11:c.G929A:p.S310N	NA	nonsynonymous	19	54651917	0.24,0.76,T	0.001,B	5.71E-05	0	0	0
	CD93 / NM_012072:exon1:c.G1097A:p.R366H	rs142218043	nonsynonymous	20	23065733	0.63,0.37,T	0.006,B	4.08E-03	0.00223	0.00159	0.000821018
	KCNE2 / NM_172201:exon2:c.A22G:p.T8A	rs2234916	nonsynonymous	21	35742799	0.1,0.0,D	0.999,D	3.78E-03	0.004921	0.00139	0.002463054
	SLC35E4 / NM_001001479:exon1:c.T515G:p.L172R	rs150781327	nonsynonymous	22	31032952	0.19,0.81,T	1.0,D	3.89E-03	0.004846	0.000998	0.004926108
	TTL12 / NM_015140:exon14:c.C1899G:p.D633E	rs147704786	nonsynonymous	22	43564050	0.16,0.84,T	0.988,D	2.92E-03	0.003383	0.001996	0.001642036
F3196	KLHL17 / NM_198317:exon5:c.A781G:p.S261G	rs200158162	nonsynonymous	1	897804	0.13,T	0.173,B	NA	0	0	0
	CACHD1 / NM_020925:exon1:c.C4T:p.R2W	NA	nonsynonymous	1	64936584	0.27,T	0.938,P	NA	0	0	0.002463054
	KIAA1804 / NM_032435:exon7:c.C1790T:p.S597F	rs34984140	nonsynonymous	1	233511776	0.00,D	1.0,D	NA	0.001924	0.0037	0.00410509
	RYR2 / NM_001035:exon31:c.G3823A:p.G1275S	NA	nonsynonymous	1	237753955	0.01,D	0.005,B	NA	0	0	0
	FMN2 / NM_020066:exon12:c.C4619T:p.S1540L	rs150801382	nonsynonymous	1	240497221	0.03,D	1.0,D	NA	0.002307	0.0009	0.005747126
	EXO1 / NM_006027:exon14:c.C2480T:p.A827V	rs145975455	nonsynonymous	1	242052841	0.09,T	0.342,B	NA	0.000846	0	0.002463054
	SETD5 / NM_001080517:exon6:c.G365A:p.R122Q	NA	nonsynonymous	3	9476543	0.08,T	0.981,D	NA	0	0	0
	FLNB / NM_001164319:exon39:c.T6611C:p.I2204T	rs149629209	nonsynonymous	3	58140566	0.00,D	0.983,D	NA	0.001692	0.0009	0.000821018
	PXK / NM_017771:exon6:c.G501C:p.K167N	rs148642996	nonsynonymous	3	58376908	0.02,D	0.993,D	NA	0.001768	0.0027	0.001642036
	SLC2A2 / NM_000340:exon3:c.G158A:p.R53Q	rs145210664	nonsynonymous	3	170732471	0.41,T	0.282,B	NA	0.000308	0.0005	0.001642036
	TECRL / NM_001010874:exon5:c.G536A:p.R179H	rs147048390	nonsynonymous	4	65180381	0.53,T	0.99,D	NA	0.002691	0.0023	0.00410509
	HERC6 / NM_001165136:exon16:c.T2063C:p.M688T	rs187435909	nonsynonymous	4	89352378	0.00,D	0.054,B	NA	0.00116	0.0014	0.004926108
	SYNPO / NM_001166208:exon2:c.G319C:p.D107H	rs6868344	nonsynonymous	5	149998248	0.02,D	0.956,D	NA	0.001533	0	0.002463054
	EZR / NM_003379:exon7:c.A752G:p.N251S	rs139467617	nonsynonymous	6	159197483	0.46,T	0.191,B	NA	0.000461	0	0.00410509
	RRM2B / NM_001172477:exon1:c.C44T:p.P15L	rs201028777	nonsynonymous	8	103251007	NA	NA	NA	0.001461	0	0.000821018

	GOT1 / NM_002079:exon4:c.G479A:p.R160H	rs146049867	nonsynonymous	10	101165952	0.05,D	0.041,B	NA	0.000308	0.0005	0.000821018
	LDLRAD3 / NM_174902:exon2:c.G64A:p.G22R	rs144816501	nonsynonymous	11	36057670	0.09,T	1.0,D	NA	0.000308	0.0005	0
	ZP1 / NM_207341:exon11:c.C1718T:p.P573L	NA	nonsynonymous	11	60642665	0.09,T	0.979,D	NA	0.000077	0	0
	C11orf9 / NM_001127392:exon7:c.C1063T:p.P355S	rs200349251	nonsynonymous	11	61539372	0.10,T	0.876,P	NA	0	0	0.000821018
	IGHMBP2 / NM_002180:exon15:c.G2837A:p.R946Q	rs149824485	nonsynonymous	11	68707054	0.03,D	0.985,D	NA	0.000847	0	0.002463054
	GALNT8 / NM_017417:exon5:c.G988A:p.E330K	rs201387598	nonsynonymous	12	4854722	0.55,T	0.007,B	NA	0	0	0
	CCDC60 / NM_178499:exon12:c.G1262A:p.R421H	rs199921208	nonsynonymous	12	119966452	NA	0.0,B	NA	0	0.0005	0
	LRRC43 / NM_001098519:exon2:c.C278G:p.S93C	NA	nonsynonymous	12	122669193	0.00,D	1.0,D	NA	0	0	0.001642036
	AKAP11 / NM_016248:exon8:c.A2288G:p.E763G	rs199696994	nonsynonymous	13	42875170	0.02,D	0.999,D	NA	0.000154	0	0
	DLST / NM_001244883:exon7:c.G403A:p.E135K	NA	nonsynonymous	14	75357831	0.45,T	1.0,D	NA	0	0	0
	CKB / NM_001823:exon7:c.A806G:p.Y269C	rs146047573	nonsynonymous	14	103986620	0.03,D	0.984,D	NA	0.000077	0	0
	PLD4 / NM_138790:exon7:c.C756A:p.H252Q	rs201336317	nonsynonymous	14	105397117	0.66,T	0.001,B	NA	0.000573	0.0009	0.002463054
	CHTF18 / NM_022092:exon19:c.C2581T:p.R861W	rs201112333	nonsynonymous	16	846841	0.05,D	0.013,B	NA	0	0.0005	0
	CLCN7 / NM_001114331:exon7:c.G658A:p.V220M	NA	nonsynonymous	16	1507703	0.02,D	1.0,D	NA	0.000154	0	0.000821018
	SLC9A3R2 / NM_001252073:exon2:c.C178T:p.R60W	rs139491786	nonsynonymous	16	2086421	0.00,D	1.0,D	NA	0.004154	0.0023	0.002463054
	RNMTL1 / NM_018146:exon4:c.G859T:p.A287S	rs139632363	nonsynonymous	17	694905	0.02,D	0.999,D	NA	0.000923	0	0.001642036
	DNAH9 / NM_001372:exon17:c.A3050G:p.Y1017C	rs139596704	nonsynonymous	17	11572808	0.00,D	0.981,D	NA	0.000308	0.0046	0.003284072
	DHRS13 / NM_144683:exon5:c.1044delA:p.Q348fs	rs139871089	frameshift deletion	17	27225549	NA	NA	NA	0.002716	0	0.000821018
	DHRS13 / NM_144683:exon4:c.G626A:p.R209Q	rs150228941	nonsynonymous	17	27228064	0.06,T	1.0,D	NA	0.000384	0.0005	0.000821018
	NLE1 / NM_018096:exon3:c.C374T:p.T125M	rs115162358	nonsynonymous	17	33466874	0.00,D	0.999,D	NA	0.003383	0.0027	0.001642036
	KRT15 / NM_002275:exon3:c.G733A:p.E245K	rs140616866	nonsynonymous	17	39673065	0.00,D	1.0,D	NA	0.001307	0	0.003284072
	KRT17 / NM_000422:exon6:c.G964A:p.A322T	rs149778356	nonsynonymous	17	39777128	0.15,T	0.868,P	NA	0.000692	0	0.00410509
	HSD17B1 / NM_000413:exon1:c.G61T:p.V21L	rs143237971	nonsynonymous	17	40705012	0.28,T	0.047,B	NA	0.000308	0.0005	0.000821018
	ABCA6 / NM_080284:exon29:c.C3722T:p.A1241V	NA	nonsynonymous	17	67083591	0.62,T	0.001,B	NA	0	0	0.000821018
	SLC39A11 / NM_001159770:exon4:c.A256G:p.T86A	rs199758218	nonsynonymous	17	71027745	0.38,T	0.001,B	NA	0	0	0
	TSEN54 / NM_207346:exon2:c.C83T:p.S28L	rs201089582	nonsynonymous	17	73512853	0.07,T	0.918,P	NA	0.001452	0	0.000821018
	C17orf110 / NM_001162997:exon2:c.G100A:p.A34T	NA	nonsynonymous	17	73643556	NA	0.017,B	NA	0	0	0
	UNC13D / NM_199242:exon31:c.G2983C:p.A995P	rs138760432	nonsynonymous	17	73825036	0.28,T	0.001,B	NA	0.001231	0.0005	0.000821018
	UNC13D / NM_199242:exon26:c.A2542C:p.I848L	rs144968313	nonsynonymous	17	73827335	0.74,T	0.001,B	NA	0.001461	0.0005	0.000821018
	ESCO1 / NM_052911:exon4:c.G970A:p.E324K	rs148536942	nonsynonymous	18	19153835	0.03,D	0.996,D	NA	0.001153	0.0023	0.001642036
	ZNF491 / NM_152356:exon3:c.G1154A:p.C385Y	rs148440087	nonsynonymous	19	11917922	0.00,D	1.0,D	NA	0.000692	0	0.000821018
	CASP14 / NM_012114:exon3:c.C52T:p.R18C	NA	nonsynonymous	19	15164317	0.00,D	1.0,D	NA	0.000231	0	0
	FBL / NM_001436:exon2:c.G71A:p.R24H	rs145491904	nonsynonymous	19	40331367	0.1,T	0.808,P	NA	0.001473	0	0.001642036
	DHX34 / NM_014681:exon4:c.A1262C:p.D421A	rs199770744	nonsynonymous	19	47861367	0.26,T	0.019,B	NA	0	0.0005	0
	ZNF341 / NM_032819:exon15:c.G2248A:p.G750S	NA	nonsynonymous	20	32379027	0.70,T	0.001,B	NA	0	0	0
	SGSM1 / NM_001039948:exon4:c.C245T:p.P82L	NA	nonsynonymous	22	25243706	0.00,D	0.904,P	NA	0.000082	0	0
	LRP5L / NM_182492:exon1:c.C76T:p.Q26X	rs61740933	stopgain SNV	22	25755984	0.12,T	NA	NA	0.001154	0.0018	0.000821018
	PANX2 / NM_052839:exon3:c.G1966A:p.A656T	NA	nonsynonymous	22	50617638	NA	0.0,B	NA	0	0	0
	POP1 / NM_001145860:c.C194G:p.S65C	rs148625494	nonsynonymous	8	99139874	0.16,T	0.902,P	NA	0.000538	0	0.002463054
	RIMS2 / NM_014677:c.G2167C:p.D723H	rs200437104	nonsynonymous	8	104987598	0.01,D	1,D	NA	0.000334	0	0
	KCTD19 / NM_001100915:c.1511_1513del:p.504_505del	NA	nonframeshift deletion	16	67328560	NA	NA	NA	0.002628	0	0.000821018
F7614	LCE1D / NM_178352:exon2:c.G248C:p.R83P	NA	nonsynonymous	1	152770518	0.00,D	0.248,B	NA	0	0	0

	CEP350 / NM_014810:exon19:c.C4337T:p.T1446I	rs140855739	nonsynonymous	1	180010912	.	0.938,P	NA	0.000461	0	0
	PRG4 / NM_001127710:exon4:c.G904A:p.A302T	rs201278528	nonsynonymous	1	186276157	.	0.123,B	NA	0	0	0
	AAK1 / NM_014911:exon17:c.C2312T:p.P771L	rs34422616	nonsynonymous	2	69723170	0.00,D	1.0,D	NA	0.000251	0	0
	RMND5B / NM_022762:exon7:c.A566G:p.N189S	rs142121381	nonsynonymous	5	177570981	0.32,T	0.926,P	NA	0.000077	0	0
	COL23A1 / NM_173465:exon21:c.C1265T:p.P422L	NA	nonsynonymous	5	177674780	0.30,T	1.0,D	NA	0	0	0.000821018
	CBY3 / NM_001164444:exon2:c.T472C:p.W158R	NA	nonsynonymous	5	179105841	0.65,T	.	NA	0.003066	0	0.001642036
	USP54 / NM_152586:exon18:c.G3691A:p.D1231N	rs4619071	nonsynonymous	10	75276493	0.19,T	0.157,B	NA	0.003537	0.0032	0.003284072
	CCDC168 / NM_001146197:exon4:c.G16004A:p.R5335Q	rs189464016	nonsynonymous	13	103387043	0.46,T	0.0,B	NA	0.004599	0.0018	0.000943396
	ASPG / NM_001080464:exon6:c.C529T:p.Q177X	rs201162007	stopgain SNV	14	104565205	0.06,T	.	NA	0.000321	0	0.000821018
	C16orf88 / NM_001012991:exon5:c.T1276G:p.W426G	rs200976200	nonsynonymous	16	19718333	.	1.0,D	NA	0.001088	0.0014	0.002463054
	INO80E / NM_173618:exon5:c.C358T:p.L120F	rs148682647	nonsynonymous	16	30012323	0.03,D	0.997,D	NA	0.002078	0.0023	0.002463054
	C21orf33 / NM_198155:exon6:c.T631G:p.F211V	rs144882095	nonsynonymous	21	45564748	0.00,D	1.0,D	NA	0.001922	0.0009	0.000821018
	SCMH1 / NM_001172222:exon3:c.C463T:p.P155S	rs143365597	nonsynonymous	1	41540902	0.9,0.10,T	1.0,D	2.55E-03	0.002153	0.000798	0.001642036
	CEBPZ / NM_005760:exon10:c.G2515A:p.V839M	NA	nonsynonymous	2	37441037	0.1,00,D	1.0,D	2.44E-05	0	0	0.000821018
F8418	PREPL / NM_001042385:exon9:c.G1375A:p.A459T	rs373786632	nonsynonymous	2	44554036	0.18,0.82,T	0.675,P	1.63E-05	0.000077	0.000199	0
	ATP2B2 / NM_001683:exon11:c.A1784G:p.N595S	rs140327013	nonsynonymous	3	10400592	0.51,0.49,T	0.005,B	4.23E-04	0.000538	0.000399	0
	XIRP1 / NM_001198621:exon2:c.G1156A:p.E386K	rs145699338	nonsynonymous	3	39229781	0.1,00,D	1.0,D	4.42E-03	0.00469	0.000798	0
	CCDC170 / NM_025059:exon5:c.T767G:p.L256R	rs192947987	nonsynonymous	6	151869617	0.1,00,D	1.0,D	8.90E-04	0.00114	0.000798	0.003284072
	THAP5 / NM_001287601:exon2:c.A649G:p.N217D	rs184758838	nonsynonymous	7	108204688	NA	0.816,P	3.66E-04	0.000154	0.0009984	0.000821018
	MYOM2 / NM_003970:exon13:c.A1515T:p.E505D	rs199950789	nonsynonymous	8	2027693	0.16,0.84,T	0.997,D	2.44E-05	0.000077	0	0
	GFRA2 / NM_001165039:exon2:c.C346T:p.P116S	rs75502370	nonsynonymous	8	21608149	0.1,00,D	1.0,D	1.22E-04	0.000543	0	0
	SFTPC / NM_001172357:exon5:c.G482A:p.R161Q	rs34957318	nonsynonymous	8	22021460	0.35,0.65,T	0.424,B	1.36E-03	0.004894	0.0045926	0.00410509
	PEBP4 / NM_144962:exon6:c.C445A:p.R149S	rs201433464	nonsynonymous	8	22582428	0.1,00,D	1.0,D	2.78E-04	0.000813	0.000599	0.000821018
	ADAM28 / NM_014265:exon3:c.G202A:p.A68T	rs138423877	nonsynonymous	8	24167458	0.19,0.81,T	0.505,P	2.36E-04	0.000847	0.000399	0
	NUP214 / NM_005085:exon29:c.G5237C:p.S1746T	rs141655844	nonsynonymous	9	134074118	0.32,0.68,T	0.873,P	7.32E-05	0.000461	0	0
	PTPLA / NM_014241:exon7:c.A800G:p.Y267C	NA	nonsynonymous	10	17632430	0.1,00,D	1.0,D	1.63E-05	0	0	0
	KIAA1217 / NM_001282769:exon8:c.A1927G:p.N643D	NA	nonsynonymous	10	24813673	0.62,0.38,T	0.804,P	6.51E-05	0	0	0
	TMEM132A / NM_017870:exon11:c.C2850A:p.S950R	rs199829020	nonsynonymous	11	60704154	0.1,00,D	0.033,B	7.99E-04	0.000852	0	0
	RAB30 / NM_001286059:exon4:c.A334G:p.S112G	rs147083051	nonsynonymous	11	82698656	NA	0.01,B	2.52E-04	0.000538	0.00119	0.000821018
	FAT3 / NM_001008781:exon9:c.C8126G:p.T2709S	NA	nonsynonymous	11	92534305	1,0,00,T	0.001,B	4.08E-05	0	0	0
	DYNC2H1 / NM_001080463:exon38:c.A6047G:p.Y2016C	rs200190291	nonsynonymous	11	103048457	0.04,0.96,D	0.308,B	6.13E-04	0.001081	0.000399	0.00410509
	TRPC4 / NM_001135955:exon3:c.A775T:p.T259S	rs144103505	nonsynonymous	13	38320196	0.04,0.96,D	0.996,D	6.51E-05	0.000231	0	0
	ENOX1 / NM_001127615:exon5:c.A197G:p.Q66R	rs76824578	nonsynonymous	13	43986063	0.1,0.90,T	0.928,P	7.64E-04	0.001615	0.00339	0.000821018
	KIAA0226L / NM_001286762:exon4:c.A644T:p.D215V	rs139391192	nonsynonymous	13	46942359	0.1,00,D	1.0,D	2.36E-04	0.000923	0.00139	0
	ZNF469 / NM_001127464:exon2:c.C4922T:p.A1641V	rs200070902	nonsynonymous	16	88498884	0.1,00,D	0.062,B	3.37E-04	0	0.000399	0
	CBFA2T3 / NM_005187:exon2:c.C293T:p.T98M	rs143043820	nonsynonymous	16	88967923	0.1,00,D	1.0,D	5.82E-04	0.000696	0.000199	0.000821018
	RTN4RL1 / NM_178568:exon2:c.G973A:p.A325T	rs189434887	nonsynonymous	17	1840143	0.77,0.23,T	0.313,B	7.02E-04	0.002355	0.001198	0.001642036
	THEG / NM_016585:exon1:c.G28A:p.G10R	rs141910884	nonsynonymous	19	375943	0.39,0.61,T	0.759,P	5.77E-05	0.000154	0.000399	0
	LENG8 / NM_052925:exon9:c.C1288T:p.R430C	rs75472495	nonsynonymous	19	54967408	0.19,0.81,T	1.0,D	2.31E-03	0.002154	0.0008	0.000821018
	GGT5 / NM_001302465:exon6:c.C817T:p.R273W	NA	nonsynonymous	22	24622225	0.13,0.87,T	0.01,B	8.95E-05	0	0	0

*dbSNP137; D: Deleterious; P: Probably pathogenic; T: Tolerated

Chapter 3

Rare Variants in the Epithelial Cadherin Gene Underlying the Genetic Etiology of Nonsyndromic Cleft Lip with or without Cleft Palate

Brito LA¹, Yamamoto GL¹, Melo S^{2,3}, Malcher C¹, Ferreira SG¹, Figueiredo J^{2,3}, Alvizi L¹, Kobayashi GS¹, Naslavsky MS¹, Alonso N⁴, Felix TM⁵, Zatz M¹, Seruca R^{2,3,6}, Passos-Bueno MR¹.

¹ Centro de Estudos do Genoma Humano e Células-Tronco, Instituto de Biociências, Universidade de São Paulo, SP, Brasil.

² IPATIMUP, Institute of Molecular Pathology and Immunology, University of Porto, Porto, Portugal.

³ Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Portugal.

⁴ Hospital das Clínicas, Faculdade de Medicina, Universidade de São Paulo, São Paulo, Brasil

⁵ Laboratório de Medicina Genômica, Centro de Pesquisa Experimental, Hospital de Clínicas de Porto Alegre, Porto Alegre, Brasil.

⁶ Faculty of Medicine, University of Porto, Portugal.

Article published in Human Mutation journal, in June / 2015

Rare Variants in the Epithelial Cadherin Gene Underlying the Genetic Etiology of Nonsyndromic Cleft Lip with or without Cleft Palate

Luciano Abreu Brito,¹ Guilherme Lopes Yamamoto,¹ Soraia Melo,^{2,3} Carolina Malcher,¹ Simone Gomes Ferreira,¹ Joana Figueiredo,^{2,3} Lucas Alvizi,¹ Gerson Shigeru Kobayashi,¹ Michel Satya Naslavsky,¹ Nivaldo Alonso,⁴ Temis Maria Felix,⁵ Mayana Zatz,¹ Raquel Seruca,^{2,3,6} and Maria Rita Passos-Bueno^{1*}

¹Centro de Pesquisa sobre o Genoma Humano e Células-Tronco, Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade de São Paulo, São Paulo, Brasil; ²IPATIMUP, Institute of Molecular Pathology and Immunology, University of Porto, Porto, Portugal; ³Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal; ⁴Hospital das Clínicas, Faculdade de Medicina, Universidade de São Paulo, São Paulo, Brasil; ⁵Laboratório de Medicina Genômica, Centro de Pesquisa Experimental, Hospital de Clínicas de Porto Alegre, Porto Alegre, Brasil; ⁶Faculty of Medicine, University of Porto, Porto, Portugal

Communicated by Iain McIntosh

Received 2 February 2015; accepted revised manuscript 17 June 2015.

Published online 29 June 2015 in Wiley Online Library (www.wiley.com/humanmutation). DOI: 10.1002/humu.22827

ABSTRACT: Nonsyndromic orofacial cleft (NSOFC) is a complex disease of still unclear genetic etiology. To investigate the contribution of rare epithelial cadherin (*CDH1*) gene variants to NSOFC, we target sequenced 221 probands. Candidate variants were evaluated via in vitro, in silico, or segregation analyses. Three probably pathogenic variants (c.760G>A [p.Asp254Asn], c.1023T>G [p.Tyr341*], and c.2351G>A [p.Arg784His]) segregated according to autosomal dominant inheritance in four nonsyndromic cleft lip with or without cleft palate (NSCL/P) families (Lod score: 5.8 at $\theta = 0$; 47% penetrance). A fourth possibly pathogenic variant (c.387+5G>A) was also found, but further functional analyses are needed (overall prevalence of *CDH1* candidate variants: 2%; 15.4% among familial cases). *CDH1* mutational burden was higher among probands from familial cases when compared to that of controls ($P = 0.002$). We concluded that *CDH1* contributes to NSCL/P with mainly rare, moderately penetrant variants, and *CDH1* haploinsufficiency is the likely etiological mechanism.

Hum Mutat 36:1029–1033, 2015. © 2015 Wiley Periodicals, Inc.

KEY WORDS: *CDH1*; oral clefts; gastric cancer; two-hit model; rare variant

Nonsyndromic cleft lip with or without cleft palate (NSCL/P) and nonsyndromic cleft palate only (NSCPO) are two complex disorders within the nonsyndromic orofacial cleft (NSOFC)

spectrum [Gorlin et al., 2001]. While the genetic etiology of NSCPO is largely unclear, genetic *loci* have been systematically implicated in NSCL/P, such as common low-risk 8q24, 10q25, and *IRF6* variants [Rahimov et al., 2008; Birnbaum et al., 2009; Mangold et al., 2010; Brito et al., 2012aa, 2012b]. However, given the high heritability attributed to NSCL/P [Hu et al., 1982; Calzolari et al., 1988; Brito et al., 2011], searching for alternative genetic variants or mechanisms is necessary to bridge the missing heritability gap of these malformations.

Rare germline variants in the gene encoding the adhesion molecule epithelial cadherin, *CDH1* (MIM# 192090), have long been associated with diffuse gastric cancer and lobular breast cancer [van Roy and Berx, 2008]. Most recently, *CDH1* mutations have been reported in OFC patients in association with gastric cancer [Frebourg et al., 2006; Kluijdt, et al. 2012; Benusiglio et al., 2013] or not [Vogelaar et al., 2013; Bureau et al., 2014]. These findings raise the questions as to what the proportion of NSOFC cases underlain by *CDH1* variants and their attributed penetrance is, and which types of mutations or mechanisms lead to OFC, cancer, or both phenotypes.

Here, we performed a variant screening for *CDH1* (NM_004360.3) coding region in 221 NSOFC probands (affected by NSCL/P [$n = 189$] or NSCPO [$n = 32$], either from nonfamilial [$n = 138$] or familial cases [$n = 83$]; Supp. Table S1). Sequencing was performed by using next-generation sequencing (NGS, exome or targeted gene sequencing) and Sanger sequencing (SS), and applied for 65 and 156 probands, respectively. Additional NGS or SS was performed for extra members of familial cases, when available. When exome sequencing was performed in affected members of the same family, we filtered out variants with minor allele frequency greater than 1% in public databases (1000 Genomes Project, and NHLBI ESP exomes) and in our in-house database of 609 Brazilian control exomes (Supp. Methods). Among the 221 probands, we identified a total of 47 variants, of which 12 were absent in our controls (two missense, one nonsense, and nine noncoding or synonymous variants; Supp. Table S2). Variants were submitted to the LOVD database (at <http://www.lovd.nl/CDH1>).

The novel missense variant c.760G>A (p.Asp254Asn, exon 6) was the most likely causative variant among the main candidates detected by exome analysis (mean coverage of 60 $\times$; average of 25,140 variants called for each individual; Supp. Table S3) in families F3788

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

*Correspondence to: Maria Rita dos Santos e Passos-Bueno, Rua do Matão 277, sala 200, Cidade Universitária, São Paulo, SP, 05508-090, Brasil. Tel.: 5511 3091-7740, E-mail: passos@ib.usp.br

Contract grant sponsor(s): Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP; CEPID Project 2013/08028-1); Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq; 401952/2010-0); Fundos FEDER através do Programa Operacional Factores de Competitividade (COMPETE); Portuguese Foundation for Science and Technology (FCT; PTDC/SAU-ONC/110294/2009, PTDC/BIM-ONC/0171/2012, PEst-OE/EE/LA0009/2013, and JF post-doctoral grant SFRH/BPD/87705/2012).

and F617 (four affected individuals sequenced in each family; Supp. Fig. S1a and b). Both families segregate NSCL/P, and haplotype analysis of the exome data did not support a close relationship between these families (data not shown). SS of additional two affected and eight unaffected members from these two families supported segregation in accordance with an autosomal dominant model with incomplete penetrance estimated at 53%. Assuming this penetrance, a Lod score of 4.8 was obtained at recombination fraction (θ) 0, under an allele frequency of 0.0001. The other novel missense variant, c.2351G>A (p.Arg784His, exon 15), was found in the proband of family F1387 through SS. Segregation with NSCL/P was evidenced by its presence in three affected relatives (Supp. Fig. S1c), and penetrance was estimated at 62%. Further, the loss-of-function variant c.1023T>G (exon 8), predicted to create a stop codon at position 341 of *CDH1* (p.Tyr341*), was found in the proband of family F7618 (Supp. Fig. S1d) through SS. Segregation with NSCL/P was suggested by its presence in an affected first cousin once removed. Although this represents the first association between this variant and NSCL/P, an association with hereditary diffuse gastric cancer (HDGC) has been previously observed [Guilford et al., 2010]. Penetrance was estimated at 31% in this family. Considering the four pedigrees, an overall 47% penetrance of NSCL/P was estimated, with a maximum Lod score of 5.86 at $\theta = 0$ (individual Lod scores: F3788: 2.3; F617: 2.5; F1387: 0.9; F7618: 0.2).

Among the remaining nine noncoding or synonymous novel variants, only two variants were significantly scored by in silico tools for pathogenicity prediction (Supp. Table S4). Variant c.387+5G>A, which possibly decreases exon 3 splice donor site recognition, was found in a NSCL/P proband from a nonfamilial case (parental DNA unavailable for testing whether it is a de novo variant). We considered this variant as possibly pathogenic, although further functional studies are necessary. Variant c.2514C>T (exon 16), present in two unrelated probands (one isolated and one familial case), was discarded as pathogenic as it did not segregate with NSCL/P in the familial case.

E-cadherin consists of three major domains: a short cytoplasmic, a single transmembrane, and a large extracellular domain, with five repetitive subdomains [Paredes et al., 2012]. The p.Asp254Asn substitution is located at a calcium-binding site comprised by the amino acid sequence Asp-Gln-Asn-Asp, at position 254–257 of *CDH1* [Tepass et al., 2000]. This site, in turn, is located in the outmost extracellular subdomain, which plays a major role in the molecular adhesive properties between cadherin trans-dimers [Shapiro et al., 1995]. Calcium binding in the extracellular subdomains is necessary for the cis-dimerization of E-cadherin, and for conferring rigidity to the extracellular domain [Nagar et al., 1996; Pertz et al., 1999]. Importantly, an amino acid substitution in a nearby calcium-binding site has been reported to completely suppress the cellular adhesive properties of E-cadherin in vitro [Ozawa et al., 1990]. The p.Arg784His substitution is located in the cytoplasmic domain, which is important for the assembly of catenins and for promoting cellular signaling [Nelson and Nusse, 2004]. Analyses with in silico tools indicated that these two missense variants are located in highly conserved regions of E-cadherin and probably impair protein function (Supp. Table S4). To determine their pathogenic potential in vitro, we transiently transfected Chinese Hamster Ovary (ATCC number: CCL-61) cells, which are negative for E-cadherin expression, with vectors encoding the wild-type (WT) E-cadherin and variants p.Asp254Asn and p.Arg784His (Supp. Methods). As revealed by Western blot and immunocytochemistry analysis, Asp254Asn cells showed decreased total E-cadherin protein expression ($P = 0.00053$; Fig. 1A), as well as reduced amount of E-cadherin located in the plasma membrane (Fig. 1B), when compared with cells expressing the WT

protein. Furthermore, mutant cells were unable to form cellular aggregates and exhibited a scattered phenotype, contrary to the WT cells, thus clearly indicating impaired adhesive function (Fig. 1C). Even though no structural impact was predicted in the mutated *CDH1* protein (performed with FoldX, <http://foldx.crg.es/>; $\Delta\Delta G = -0.81$ kcal/mol), our in vitro assays suggest that p.Asp254Asn may lead to premature degradation, as shown for other cancer-related *CDH1* pathogenic variants [Simoes-Correia et al., 2008, 2012; Figueiredo et al., 2013]. The functional effect of this variant could also be related to disturbances in calcium ion binding, given its location. Arg784His cells, in turn, showed no observable difference from WT cells in total E-cadherin amount, its location in the plasma membrane, and its adhesive behavior (Fig. 1A–C). However, this result should not be sufficient to rule out the pathogenicity of this variant, since this in vitro assay may not be able to detect other types of functional effects, such as changes in interactions with other proteins, altering subsequent signaling pathways.

Families F3788 (p.Asp254Asn) and F1387 (p.Arg784His), the only two we were able to reascertain for cancer family history, include mutation carriers aged up to 70 years without cancer. This observation suggests that, under certain circumstances, *CDH1* variants might cause NSCL/P alone. The invasive potential of Asp254Asn and Arg784His cells, investigated by an in vitro Matrigel matrix invasion assay, was similar to that of WT cells (Fig. 1D). Thus, it is possible that some E-cadherin mutations increase the risk of NSCL/P alone, while others increase risk of gastric cancer (mutations associated with higher invasiveness). However, it is of note that the current landscape of *CDH1* mutations associated with gastric cancer and CL/P does not suggest any preferential distribution of mutations along the E-cadherin molecule (Fig. 2).

The overall prevalence of rare, possibly pathogenic *CDH1* variants here reported was 2% (5 out of 221 NSOFC probands). To date, the *CDH1* mutational repertoire in the literature associated with OFC includes 10 different mutations. Six of these mutations have been reported in families also segregating gastric cancer (four affecting mRNA splicing, one nonsense, and one frameshift deletion) [Frebouret et al., 2006; Kluijdt et al., 2012; Benusiglio et al., 2013] and four were found in individuals with uncertain history of gastric cancer (one nonsense [Bureau et al., 2014] and three missense in a European cohort [Vogelaar et al., 2013]). Revisiting the list of variants described in the European cohort, we observed that the missense variant c.88C>A, which was reported in two patients from that study, was also found in two of our Brazilian controls; after removing this variant, the prevalence of possibly pathogenic *CDH1* mutations in the European cohort becomes 2%, instead of the previously reported 5% and now similar to our estimate. Furthermore, considering that two of the possibly pathogenic mutations here reported (p.Asp254Asn and p.Tyr341*) and most of the 10 above-mentioned mutations are predicted to cause *CDH1* loss-of-function, haploinsufficiency in critical stages of embryonic development seems to be the most likely mechanism by which rare variants in *CDH1* lead to OFC.

To investigate whether the group of 221 NSOFC probands presents a higher burden of *CDH1* rare variants compared to that of 609 Brazilian controls, we performed a gene-based sequence kernel association test (SKAT) [Wu et al., 2011]. A complementary two-tailed Fisher's exact test was performed to compare the proportion of individuals carrying at least one rare *CDH1* variant between probands and controls. Only variants with minor allele frequency <1% and with nonneutral prediction in at least one in silico tool were selected for these tests. To avoid methodological bias in the tests, we only included variants from regions that were

Figure 1. In vitro functional characterization of *CDH1* missense variants p.Asp254Asn (D254N) and p.Arg784His (R784H). **A:** Protein expression levels of Chinese Hamster Ovary cells transfected with the WT E-cadherin, the variants D254N and R784H, and the empty vector (Mock). α -Tubulin was used as a loading control. Band intensity was quantified and normalized against WT E-cadherin-expressing cells. The graph shows the average $\pm$ SE of E-cadherin protein level in five independent experiments. **B:** Immunocytochemistry showing E-cadherin subcellular localization (green staining). Nuclei counterstained with DAPI (blue). Scale bar represents 100 μ m. **C:** Cell-cell adhesive properties assessed by slow aggregation assays. **D:** Invasive ability of cells expressing WT or the D254N and R784H variants. The graph shows the number of invasive cells $\pm$ SE of five independent experiments. Cells expressing a known invasive mutation associated with gastric cancer (Q16H), used as a positive control, showed higher invasive potential than the studied mutations.

Figure 2. Scheme of *CDH1* protein showing the distribution of *CDH1* germline mutations (missense $\times$ nonsense/splice site/frameshift) associated with gastric cancer [Corso et al., 2012], cleft lip/palate, and both phenotypes. Circles depict the proportion of mutations associated with each phenotype along the protein domains (signal, precursor, extracellular, transmembrane [TM], and cytoplasmic).

covered by both SS and NGS (with minimum coverage of 25 $\times$; Supp. Table S5). No significant differences in variant enrichment were detected by SKAT when comparing the 221 NSOFC probands with our 609 control exomes ($P = 0.25$). Similarly, no significant difference was detected in the number of individuals carrying these variants (two-tailed Fisher's exact test $P = 0.85$; patients: 9/221 [4%]; controls: 28/609 [5%]). Most of the probably pathogenic variants here

reported (p.Asp254Asn, p.Tyr341*, and p.Arg784His) were found in familial cases with at least two affected members aside from the probands (4 out of 26 families matching the same condition, or 15.4%). Considering only probands from families with at least two additional affected individuals, significant differences were detected by SKAT ($P = 0.002$) and by two-tailed Fisher's exact test ($P = 0.002$; patients: 6/26 [23%]; controls: 28/609 [5%]). These findings

suggest that the most noteworthy *CDH1* etiological contribution to NSOFC arises from the fraction of NSCL/P cases involving moderate penetrance, which is best represented by familial cases. Since the previously suggested association between common variants and NSCL/P [Letra et al., 2009; Hozyasz et al., 2014] has not been supported by a large meta-analysis with GWAS data [Ludwig et al., 2012], rare variants seem to be the major contribution of *CDH1* to NSCL/P etiology.

Given the lack of correlation between type/location of *CDH1* rare pathogenic variants and NSCL/P or HDGC (Fig. 2), we speculate that a common underlying molecular mechanism could explain both phenotypes. Penetrance of *CDH1* germline mutations implicated in HDGC depends on a second hit, which frequently occurs via promoter hypermethylation of the nonmutated allele, possibly triggered by environmental factors [Oliveira et al., 2009; Zeng et al., 2015]. In this regard, a lifetime exposure to such factors would be in agreement with the higher penetrance in HDGC (80%) [Pharoah et al., 2001], as compared to NSCL/P (47%). Germline, pathogenic variants in *CDH1* could determine the resultant phenotype (NSCL/P or gastric cancer) under the influence of the following factors: time (early development or later in life), tissue (craniofacial or gastric structures), and exposure to environmental factors. In addition, given the prevalence of *CDH1* pathogenic variants found in this study, we believe that the NSCL/P-associated *CDH1* mutations are currently underrepresented, and future research should focus on their identification.

In summary, our results indicate a consistent role of rare, loss-of-function, moderately penetrant *CDH1* variants in NSCL/P etiology. To better comprehend the mechanisms linking *CDH1* to NSCL/P, as well as the risk of gastric cancer among NSCL/P individuals with mutations in *CDH1*, further studies are needed. Finally, *CDH1* testing in NSCL/P familial cases should be discussed for genetic counseling purposes.

Acknowledgments

We thank the Operation Smile Brazil team and Dr. Daniela Bueno (Hospital Menino Jesus) for the productive collaboration, and Suzana Ezquina, Vitor Aguiar, Vanessa Simões, Dr. Paulo A. Otto, and Dr. Rui M. Ferreira for helpful discussions.

Disclosure statement

The authors declare no conflict of interest.

References

Benusiglio PR, Caron O, Consolino E, Duvillard P, Coulet F, Blayau M, Malka D. 2013. Cleft lip, cleft palate, hereditary diffuse gastric cancer and germline mutations in *CDH1*. *Int J Cancer* 132:2470.

Birnbaum S, Ludwig KU, Reutter H, Herms S, Steffens M, Rubini M, Baluardo C, Ferriani M, Almeida de Assis N, Alblas MA, Barth S, Freudenberg J, et al. 2009. Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24. *Nat Genet* 41:473–477.

Brito LA, Bassi CF, Masotti C, Malcher C, Rocha KM, Schlesinger D, Bueno DF, Cruz LA, Barbara LK, Bertola DR, Meyer D, Franco D, et al. 2012a. IRF6 is a risk factor for nonsyndromic cleft lip in the Brazilian population. *Am J Med Genet A* 158:2170–2175.

Brito LA, Cruz LA, Rocha KM, Barbara LK, Silva CB, Bueno DF, Aguiar M, Bertola DR, Franco D, Costa AM, Alonso N, Otto PA, et al. 2011. Genetic contribution for non-syndromic cleft lip with or without cleft palate (NS CL/P) in different regions of Brazil and implications for association studies. *Am J Med Genet A* 155A:1581–1587.

Brito LA, Paranaíba LM, Bassi CF, Masotti C, Malcher C, Schlesinger D, Rocha KM, Cruz LA, Bárbara LK, Alonso N, Franco D, Bagordakis E, et al. 2012b. Region 8q24

is a susceptibility locus for nonsyndromic oral clefting in Brazil. *Birth Defects Res A Clin Mol Teratol* 94:464–468.

Bureau A, Parker MM, Ruczinski I, Taub MA, Marazita ML, Murray JC, Mangold E, Noethen MM, Ludwig KU, Hetmanski JB, Bailey-Wilson JE, Cropp CD, 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, Milan M, Cavazzuti GB, Cocchi G, Gandini E, Magnani C, Moretti M, Garani GP, Salvioli GP, Volpato S. 1988. Epidemiological and genetic study of 200 cases of oral cleft in the Emilia Romagna region of northern Italy. *Teratology* 38:559–564.

Corso G, Marrelli D, Pascale V, Vindigni C, Roviello F. 2012. Frequency of *CDH1* germline mutations in gastric carcinoma coming from high- and low-risk areas: metanalysis and systematic review of the literature. *BMC Cancer* 12:8.

Figueiredo J, Soderberg O, Simoes-Correia J, Grannas K, Suriano G, Seruca R. 2013. The importance of E-cadherin binding partners to evaluate the pathogenicity of E-cadherin missense mutations associated to HDGC. *Eur J Hum Genet* 21:301–309.

Fitzgerald RC, Hardwick R, Huntsman D, Carneiro F, Guilford P, Blair V, Chung DC, Norton J, Ragnauth K, VanKrieken JH, Dwerryhouse S, Caldas C. 2010. Hereditary diffuse gastric cancer: updated consensus guidelines for clinical management and directions for future research. *J Med Genet* 47:436–444.

Frebourg T, Oliveira C, Hochain P, Karam R, Manouvrier S, Graziadio C, Vekemans M, Hartmann A, Baert-Desurmont S, Alexandre C, Lejeune Dumoulin S, Marroni C, et al. 2006. Cleft lip/palate and *CDH1*/E-cadherin mutations in families with hereditary diffuse gastric cancer. *J Med Genet* 43:138–142.

Gorlin RJ, Cohen MMJ, Hennekam RCM. 2001. Syndromes of the head and neck. New York: Oxford University Press. p 1344.

Guilford P, Humar B, Blair V. 2010. Hereditary diffuse gastric cancer: translation of *CDH1* germline mutations into clinical practice. *Gastric Cancer* 13:1–10.

Hozyasz KK, Mostowska A, Wojcicki P, Lasota A, Offert B, Balcerek A, Dunin-Wilczynska I, Jagodzinski PP. 2014. Nucleotide variants of the cancer predisposing gene *CDH1* and the risk of non-syndromic cleft lip with or without cleft palate. *Fam Cancer* 13:415–421.

Hu DN, Li JH, Chen HY, Chang HS, Wu BX, Lu ZK, Wang DZ, Liu XG. 1982. Genetics of cleft lip and cleft palate in China. *Am J Hum Genet* 34:999–1002.

Kluijft I, Siemerink EJ, Ausems MG, van Os TA, de Jong D, Simoes-Correia J, van Krieken JH, Ligtenberg MJ, Figueiredo J, van Riel E, Sijmons RH, Plukker JT, et al. 2012. *CDH1*-related hereditary diffuse gastric cancer syndrome: clinical variations and implications for counseling. *Int J Cancer* 131:367–376.

Letra A, Menezes R, Granjeiro JM, Vieira AR. 2009. *AXIN2* and *CDH1* polymorphisms, tooth agenesis, and oral clefts. *Birth Defects Res A Clin Mol Teratol* 85:169–173.

Ludwig KU, Mangold E, Herms S, Nowak S, Reutter H, Paul A, Becker J, Herberz R, AlChawa T, Nasser E, Böhmer AC, Mattheisen M, 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, Baluardo C, Ferriani M, Herms S, Reutter H, de Assis NA, Chawa TA, Mattheisen M, Steffens M, Barth 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.

Nagar B, Overduin M, Ikura M, Rini JM. 1996. Structural basis of calcium-induced E-cadherin rigidification and dimerization. *Nature* 380:360–364.

Nelson WJ, Nusse R. 2004. Convergence of Wnt, beta-catenin, and cadherin pathways. *Science* 303:1483–1487.

Oliveira C, Sousa S, Pinheiro H, Karam R, Bordeira-Carriço R, Senz J, Kaurah P, Carvalho J, Pereira R, Gusmão L, Wen X, Cipriano MA et al. 2009. Quantification of epigenetic and genetic 2nd hits in *CDH1* during hereditary diffuse gastric cancer syndrome progression. *Gastroenterology* 136:2137–2148.

Ozawa M, Engel J, Kemler R. 1990. Single amino acid substitutions in one Ca^{2+} binding site of uvomorulin abolish the adhesive function. *Cell* 63:1033–1038.

Paredes J, Figueiredo J, Albergaria A, Oliveira P, Carvalho J, Ribeiro AS, Caldeira J, Costa AM, Simoes-Correia J, Oliveira MJ, Pinheiro H, Pinho SS, et al. 2012. Epithelial E- and P-cadherins: role and clinical significance in cancer. *Biochim Biophys Acta* 1826:297–311.

Pertz O, Bozic D, Koch AW, Fauser C, Brancaccio A, Engel J. 1999. A new crystal structure, Ca^{2+} dependence and mutational analysis reveal molecular details of E-cadherin homoassociation. *EMBO J* 18:1738–1747.

Pharoah PD, Guilford P, Caldas C. 2001. Incidence of gastric cancer in *CDH1* (E-cadherin) mutation carriers from hereditary diffuse gastric cancer families. *Gastroenterology* 121:1348–1353.

Rahimov F, Marazita ML, Visel A, Cooper ME, Hitchler MJ, Rubini M, Domann FE, Govil M, Christensen K, Bille C, Melbye M, Jugessur A, et al. 2008. Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip. *Nat Genet* 40:1341–1347.

- Shapiro L, Fannon AM, Kwong PD, Thompson A, Lehmann MS, Grubel G, Legrand JF, Als-Nielsen J, Colman DR, Hendrickson WA. 1995. Structural basis of cell-cell adhesion by cadherins. *Nature* 374:327–337.
- Simoes-Correia J, Figueiredo J, Lopes R, Stricher F, Oliveira C, Serrano L, Seruca R. 2012. E-cadherin destabilization accounts for the pathogenicity of missense mutations in hereditary diffuse gastric cancer. *PLoS One* 7:e33783.
- Simoes-Correia J, Figueiredo J, Oliveira C, van Hengel J, Seruca R, van Roy F, Suriano G. 2008. Endoplasmic reticulum quality control: a new mechanism of E-cadherin regulation and its implication in cancer. *Hum Mol Genet* 17:3566–3576.
- Tepass U, Truong K, Godt D, Ikura M, Peifer M. 2000. Cadherins in embryonic and neural morphogenesis. *Nat Rev Mol Cell Biol* 1:91–100.
- van Roy F, Berx G. 2008. The cell-cell adhesion molecule E-cadherin. *Cell Mol Life Sci* 65:3756–3788.
- Vogelaar IP, Figueiredo J, van Rooij IA, Simoes-Correia J, van der Post RS, Melo S, Seruca R, Carels CE, Ligtenberg MJ, Hoogerbrugge N. 2013. Identification of germline mutations in the cancer predisposing gene CDH1 in patients with orofacial clefts. *Hum Mol Genet* 22:919–926.
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 89:82–93.
- Yamaoka Y, Kato M, Asaka M. 2008. Geographic differences in gastric cancer incidence can be explained by differences between *Helicobacter pylori* strains. *Intern Med* 47:1077–1083.
- Zeng W, Zhu J, Shan L, Han Z, Aexiding P, Quhai A, Zeng F, Wang Z, Li H. 2015. The clinicopathological significance of CDH1 in gastric cancer: a meta-analysis and systematic review. *Drug Des Devel Ther* 13:2149–2157.

SUPP. METHODS

Ethics Committee approval and patient consents

This study was approved by the Ethics Committee of the Instituto de Biociências (Universidade de São Paulo, Brazil) and all biological samples were obtained after informed consent from the patients or their legal guardians.

Patient and control details

Patients were ascertained at Hospital das Clínicas of Universidade de São Paulo (São Paulo, Brazil), Hospital de Clínicas de Porto Alegre (Rio Grande do Sul, Brazil) or during surgical programs of Operation Smile Brazil. Controls were obtained from the biorepository of the Centro de Pesquisa sobre o Genoma Humano e Células-Tronco, Universidade de São Paulo.

A database of exome sequencing data from 609 elderly Brazilians (60 years of age or more), stored at the Centro de Pesquisa sobre o Genoma Humano e Células-Tronco (Universidade de São Paulo), was used as control. This in-house bank is ethnically and socioeconomically matched with the general population from São Paulo, using the 2000 Census of the Brazilian Institute of Geography and Statistics as reference (unpublished data).

DNA extraction

All DNA samples were extracted from peripheral blood (following standard protocols) or saliva (collected with Oragene® DNA collection kit OG-500, and following prepIT-L2P manufacturer's instructions; DNA Genotek Inc, Ottawa, Canada).

CDH1 sequencing

The coding region and exon-intron boundaries of *CDH1* (NM_004360.3) were sequenced using NGS (exome [n=6 probands] and targeted gene sequencing [n=59 probands]) and SS

(n=156 probands). When DNA was available, additional family members were sequenced using NGS or SS.

Exome sequencing was performed in probands from 6 familial cases and their affected relatives on HiScanSQ (Illumina; exome mean coverage of 60x; *CDHI* mean coverage of 43x; 80% of bases presenting and minimal coverage of 20x). Library preparation and exome capture were performed with Illumina's TruSeq Sample Prep and Exome Enrichment kits. *CDHI* targeted sequencing was performed in Illumina MiSeq sequencer (*CDHI* mean coverage of 110x; 95% of bases presenting minimal coverage of 20x), using Illumina's Nextera kits for library preparation and custom capture of 143 genes involved in craniofacial and skeletal dysplasia disorders. KAPA Library Quantification kit (KAPA Biosystems) was used to quantify the libraries through real-time quantitative PCR. Sequence alignment, data processing, variant calling, and variant annotation were performed with Burrows-Wheller Aligner (BWA; <http://bio-bwa.sourceforge.net>), Picard (<http://broadinstitute.github.io/picard/>), Genome Analysis Toolkit package (GATK; <https://www.broadinstitute.org/gatk/>) and ANNOVAR (<http://www.openbioinformatics.org/annovar/>), respectively.

Exome sequencing from our in-house control database had mean coverage of 60x, and *CDHI* mean coverage of 49x (82% of bases presenting minimum of 20x coverage depth).

To identify the causative variant in each family, we filtered out variants with minor allele frequency greater than 1% (in either public databases – dbSNP, 1000 Genomes Project [Abecasis et al., 2012] and NHLBI Exome Sequencing Project [<http://evs.gs.washington.edu/EVS/>] - or in our in-house database of 609 control exomes). Only variants in heterozygosis in all sequenced affected individuals from the same family were considered. Among the residual variants, we prioritized nonsynonymous variants with *in silico* prediction of protein damage by SIFT (damaging if <0.05) and Polyphen-2 (damaging if >0.15), and those located in genes previously associated with craniofacial disorders. All *CDHI* rare variants detected by NGS were visually inspected in the Integrative Genomics Viewer software (Broad Institute of MIT and Harvard), and subsequently validated by SS in the same individuals.

PCR primers for SS covered all *CDHI* protein-coding regions and exon-intron boundaries (primer sequences available upon request). Capillary electrophoresis was performed

with ABI3730 DNA Analyzer (Applied Biosystems), and sequences were inspected in Sequencher® 5.2 sequence analysis software (Gene Codes).

Variant names throughout the article are based on cDNA sequence (where 1 is the first base of the translation initiation codon of NM_004360.3) or protein sequence, whenever appropriate.

***In silico* analyses**

Bioinformatics prediction analyses for the missense variants were performed using the Ensembl Transcript CDH1-001 (ENST00000261769; 882 amino acids), with the following tools: PROVEAN Protein Batch (which returns SIFT and PROVEAN scores; http://provean.jcvi.org/protein_batch_submit.php?species=human) [Kumar et al., 2009; Choi et al., 2012], Polyphen-2 (<http://genetics.bwh.harvard.edu/pph2/>) [Adzhubei et al., 2010], Mutation Taster (which returns the Mutation Taster prediction and the conservation scores of PhyloP and PhastCons; <http://www.mutationtaster.org>) [Siepel et al., 2005; Pollard et al., 2010; Schwarz et al., 2014], and Align GVGD (<http://agvgd.iarc.fr/index.php>) [Tavtigian et al., 2005].

Possible splice site alterations were predicted using NNSplice algorithm (<http://www.fruitfly.org>) [Reese et al., 1997] and Human Splicing Finder 3.0 (which returns MaxEntScan and Human Splicing Finder scores; <http://www.umd.be/HSF3/>) [Yeo et al., 2004; Desmet et al., 2009].

FoldX (<http://foldx.crg.es/>) algorithm was used to predict the impact of p.Asp254Asn on protein structure, based on structural models previously described [Simões-Correia et al., 2012]. The energetic difference ($\Delta\Delta G$) between the WT reference and mutant was then generated. A threshold of $\Delta\Delta G > 0.81$ was adopted to consider the mutation destabilizing.

Plasmid construction

We used vectors encoding the wild-type (WT) and the *CDH1* variants p.Asp254Asn and p.Arg784His, which was induced in the hCDH1 pIRES2-EGFP vector (Clontech, Takara Bio) by site-directed mutagenesis (primers sequences available upon request) [Wang and Wilkinson, 2000]. The corresponding empty vector was used as control.

Cell culture and transfection

CHO (Chinese Hamster Ovary) cells (ATCC number: CCL-61) were cultured in α -MEM medium (Gibco, Invitrogen) supplemented with 10% fetal bovine serum (HyClone, Perbio) and 1% penicillin/streptomycin (Gibco, Invitrogen), and maintained at 37°C under 5% CO₂ humidified air. Twenty-four hours after seeding, cells were transiently transfected with vectors encoding the WT protein, the variants p.Asp254Asn and p.Arg784His, or the empty vector (mock condition). Transfections were performed with Lipofectamine 2000 (Invitrogen), following the manufacturer's recommendations. To control transfection efficiency, GFP fluorescence was assessed in each experiment by flow cytometry.

Western blotting

Cell lysates were prepared by scraping the cells in cold lysis buffer composed of 1% Triton X-100 (Sigma) and 1% Nonidet P-40 (Sigma) in PBS (phosphate-buffered saline, pH7.4), supplemented with a cocktail of phosphatase and protease inhibitors (from Sigma and Roche, respectively). Protein concentration was determined using a modified Bradford assay (Bio-Rad Protein Assay; Bio-Rad). Fifteen micrograms of total protein were separated in 7.5% SDS-PAGE gels and electroblotted onto Hybond ECL membranes (Amersham Biosciences). Membrane blocking was performed in 5% non-fat milk and 0.5% Tween-20 in PBS, for 1h. Antibodies against E-cadherin (1:2000, Clone HECD1, Invitrogen) and α -Tubulin (1:10000, Sigma) diluted in blocking solution were used for protein detection. This was followed by incubation with a sheep anti-mouse HRP-conjugated secondary antibody (Amersham Biosciences) and detection with ECL reagents (Amersham Biosciences), following manufacturer's recommendations. Densitometric analysis of 5 independent experiments was carried out with Quantity One Software (Bio Rad). Relative protein expression was calculated by normalizing E-cadherin values against the loading control (α -Tubulin).

Immunofluorescence staining

Immunocytochemistry was applied to study E-cadherin expression and localization. Cells seeded on top of glass coverslips were washed in PBS and fixed in ice-cold methanol for 20min. Blocking was performed with 3% bovine serum albumin in PBS for 30min at room temperature. E-cadherin was detected with a mouse monoclonal antibody (BD Biosciences), diluted at 1:300 in blocking solution. Cells were subsequently dark-incubated with the Alexa Fluor 488 goat anti-mouse antibody (1:500, Invitrogen) for 1h. Coverslips were mounted on slides using Vectashield with DAPI (Vector Laboratories). Images were acquired on a Carl Zeiss Apo Tome Axiovert 200M Fluorescence Microscope with an AxioCam HRm camera, and processed with the Zeiss AxioVision 4.8 software.

Slow aggregation assay

To test cell-cell aggregation, wells of 96-well-plates were coated with 50µl of an agar solution, prepared by dissolving 100mg of Bacto-Agar in 15ml of sterile PBS. Cells were then trypsinized and resuspended in culture medium. A cellular suspension of 1×10^5 cells/ml was prepared for each condition, and 200µl (corresponding to 2×10^4 cells) were added to the agar-coated wells. Each condition was evaluated in triplicate. The plate was incubated at 37°C under 5% CO₂ humidified air. Aggregation phenotypes were evaluated 24h after seeding, under an inverted microscope (objective of 4x, ECLIPSE TS100-Nikon), and photographed with a Digital Sight Camera System from Nikon.

Matrigel invasion assay

Cell invasiveness was assessed with Matrigel invasion chambers suitable for 24-well-plates (BD Biocoat). Matrigel inserts were hydrated with α -MEM medium for 1h at 37°C. Thereafter, 500µl of cellular suspensions at 1×10^5 cells/ml (corresponding to 5×10^4 cells) were plated in each chamber. The plate was incubated at 37°C in a humidified atmosphere with 5% CO₂. Twenty-four hours after seeding, noninvasive cells and Matrigel from the upper side of the filters were removed with a pre-wet 'cotton swab'. Membranes were washed in PBS, fixed in ice-cold methanol for 15min, and mounted on slides with Vectashield with DAPI (Vector Laboratories). The total number of invasive nuclei was counted under a Leica DM2000

microscope. A previously studied germline missense mutation, associated with increased cell invasion, was included as a control (p.Gln16His)[Zhang et al., 2014].

Statistical analyses

A gene-based Sequence Kernel Association Test (SKAT) was used to test for overall burden of *CDHI* rare variants in the affected individuals, compared to the 609 controls. Two-tailed Fisher's exact test was performed to compare the proportion of individuals carrying at least one rare variant between probands and controls. Variant inclusion criteria for these tests were minor allele frequency <1% in probands and controls, and prediction to be non-neutral by at least one *in silico* tool. Given the different sequencing methods used, we restricted these test to variants simultaneously covered by SS and NGS (with a minimum coverage of 25x for both gene targeted and exome sequencing of patients and controls). Statistical significance was considered when $p \leq 0.05$.

Two-point Lod scores were calculated with Superlink Online SNP 1.1 (<http://cbl-hap.cs.technion.ac.il/superlink-snp/>), for familial cases in which DNA from additional family members were available for segregation analyses, using *CDHI* candidate variants as markers. Penetrance was estimated with PENCALC software (<http://www.usp.br/pencalcweb/web/>)[Horimoto et al., 2010].

In vitro invasion assays were analyzed with a two-tailed unpaired Student's *t*-test by comparing the number of invasive nuclei between mutants and WT cells, using $p \leq 0.05$ as threshold.

Supp. Figure S1. Pedigrees of familial NS CL/P segregating mutations in *CDH1*. Chromatograms are displayed for the probands of each family, and genotypes are informed for the individuals with DNA available. (A) Family F3788, segregating c.760G>A. Exome sequencing performed for individuals III-4, III-7, III-16 and IV-2. (B) Family F617, segregating c.760G>A. Exome sequencing performed for individuals III-13, IV-1, IV-9 and V-3. (C) Family F1387, segregating c.2351G>A. Note that individual IV-1, affected by submucous CPO, do not segregate this variant, possibly reflecting etiological differences between CPO and CL/P [Gorlin, et al., 2001]. (D) Family F7618, segregating c.1023T>G; additional DNA sample was obtained only from individual III-1 in this family. CL: cleft lip only; CLP: cleft lip and palate; LS: lip scar; SCP: submucous cleft palate; yo: years old at last ascertainment.

Supp. Table S1. Sample description of the 221 probands used in the study

	Total	Female / Male	Familial^a / Isolated
NS CL/P	189	71 / 118	74 / 115
NS CPO	32	19 / 13	9 / 23

^a at least one affected individual with minimum coefficient of relationship of 1/32 (e.g. second cousins).

Supp. Table S2. Total *CDH1* variants detected in the 221 probands, and minor allele frequencies among probands and 609 controls

Variant	rs name	Location	Chr	Genotype counting in probands			MAF	MAF
				CC/CC=	CC/AA=	AA/AA=	probands	controls
c.-124-5_6delinsAA	ND	upstream	392	195	1	0	AA = 0.0026	0
c.-84G>A	ND	5'UTR	400	GG=199	GA=1	AA=0	A = 0.0025	0.0026
c.-71C>G	rs34033771	5'UTR	400	CC = 197	CG = 3	GG = 0	G = 0.0075	0.013
c.48+22G>A	ND	intron 1	400	GG = 199	GA = 1	AA = 0	A = 0.0025	0
c.324A>G	rs116542018	exon 3 (syn)	438	AA = 218	AG = 1	GG = 0	G = 0.0023	0.0049
c.345G>A	rs1801023	exon 3 (syn)	438	GG = 217	GA = 2	AA = 0	A = 0.0046	0.0049
c.387+5G>A	ND	intron 3	438	GG = 218	GA = 1	AA = 0	A = 0.0023	0
c.387+27C>T	rs33932809	intron 3	438	CC = 217	CT = 2	TT = 0	T = 0.0046	0.0041
c.531+10G>C	rs33963999	intron 4	438	GG = 210	GC = 9	CC = 0	C = 0.0205	0.0287
c.532-27_-26insA	ND	intron 4	438	: : = 218	: A = 1	AA = 0	A = 0.0023	0
c.532-18C>T	rs200673941	intron 4	438	CC = 218	CT = 1	TT = 0	T = 0.0023	0.0033
c.687+144G>A	rs7188750	intron 5	378	GG = 124	GA = 57	AA = 8	A = 0.1931	0.162
c.760G>A	ND	exon 6 (mis)	434	GG = 215	GA = 2	GG = 0	A = 0.0046	0
c.833-16C>G	rs33984587	intron 6	436	CC = 217	CG = 1	GG = 0	G = 0.0023	0.0041
c.933C>G	rs35539711	exon 7 (syn)	436	CC = 217	CG = 1	GG = 0	G = 0.0023	0.0074
c.1008+47T>C	rs34374107	intron 7	436	TT = 216	TC = 2	CC = 0	C = 0.0046	0.0090
c.1009-130C>T	ND	intron 7	432	CC = 215	CT = 1	TT = 0	T = 0.0023	0
c.1009-75C>T	ND	intron 7	432	CC = 215	CT = 1	TT = 0	T = 0.0023	0
c.1023T>G	rs587776398	exon 8 (non)	432	TT = 215	TC = 1	CC = 0	C = 0.0023	0
c.1138-125G>T	ND	intron 8	436	GG = 217	GT = 1	TT = 0	T = 0.0023	0
c.1272C>T	rs61756284	exon 9 (syn)	436	CC = 216	CT = 2	TT = 0	T = 0.0046	0.0033
c.1320+35C>T	rs35423758	intron 9	436	CC = 217	CT = 1	TT = 0	T = 0.0023	0.0008
c.1320+45G>C	rs33950903	intron 9	436	GG = 207	GC = 11	CC = 0	C = 0.0252	0.0164
c.1321-88C>T	rs33934457	intron 9	410	CC = 204	CT = 1	TT = 0	T = 0.0024	0.0057
c.1711+47G>A	rs35667437	intron 11	440	GG = 209	GA = 11	AA = 0	A = 0.0250	0.0116
c.1712-52G>C	rs34582129	intron 11	442	GG = 216	GC = 5	CC = 0	C = 0.0113	0.0164
c.1712-37T>G	rs35831514	intron 11	442	TT = 219	TG = 2	GG = 0	G = 0.0045	0.0057
c.1774G>A ^a	rs35187787	exon 12 (mis)	442	GG = 220	GA = 1	AA = 0	A = 0.0023	0.0033
c.1849G>A ^a	rs33935154	exon 12 (mis)	442	GG = 216	GA = 5	AA = 0	A = 0.0113	0.0066
c.1896C>T	rs33969373	exon 12 (syn)	442	CC = 208	CT = 13	TT = 0	T = 0.0294	0.0386
c.1937-13T>C	rs2276330	intron 12	438	TT = 196	TC = 23	CC = 0	C = 0.0525	0.0747
c.2076T>C	rs1801552	exon 13 (syn)	428	TT = 32	TC = 79	CC = 109	T = 0.3250	0.3440
c.2253C>T	rs33964119	exon 14 (syn)	434	CC = 195	CT = 20	TT = 2	T = 0.0553	0.0378
c.2292C>T	rs61747636	exon 14 (syn)	434	CC = 213	CT = 4	TT = 0	T = 0.0092	0.0025
c.2295+53G>A	rs199498992	intron 14	434	GG = 216	GA = 1	AA = 0	A = 0.0023	0.0008

Variant	rs name	Location	Chr	Genotype counting in probands			MAF probands	MAF controls
c.2295+89delC	ND	intron 14	434	CC = 216	C : = 1	: : = 0	: = 0.0023	0.0016
c.2295+166C>T	rs143044978	intron 14	434	CC = 216	CT = 1	TT = 0	T = 0.0023	0.012
c.2295+199G>A	ND	intron 14	434	GG = 216	GA = 1	AA = 0	A = 0.0023	0.0025
c.2351G>A	ND	exon 15 (mis)	434	GG = 216	GA = 1	AA = 0	A = 0.0023	0
c.2439+22C>T	rs34751606	intron 15	434	CC = 216	CT = 1	TT = 0	T = 0.0023	0.0025
c.2440-143G>A	rs34238989	Intron 15	432	GG = 213	GA = 3	AA = 0	A = 0.0069	0.016
c.2440-126A>G	rs35316236	intron 15	432	AA = 213	AG = 3	GG = 0	G = 0.0069	0.0052
c.2514C>T	ND	exon 16 (syn)	434	CC = 215	CT = 2	TT = 0	T = 0.0046	0
c.2520C>T	rs140328601	exon 16 (syn)	434	CC = 213	CT = 4	TT = 0	T = 0.0092	0.0016
c.2634C>T	rs2229044	exon 16 (syn)	434	CC = 208	CT = 9	TT = 0	T = 0.0207	0.0300
c.*16C>T	ND	3'UTR	434	CC = 216	CT = 1	TT = 0	T = 0.0023	0
c.*54C>T	rs1801026	3'UTR	434	CC = 163	CT = 48	TT = 5	T = 0.1343	0.1650

Variant names based on cDNA sequence. **Chr**: number of chromosomes analyzed in the probands (Number of sequenced chromosomes varied per variant, as we included only those with good quality); **MAF**: Minor allele frequency; **ND**: non-described; **syn**: synonymous; **mis**: missense; **non**: nonsense.

^a Variants previously reported in patients in gastric cancer, and considered non-pathogenic by *in silico* or *in vitro* approaches [Suriano, et al., 2003; Keller, et al., 2004].

Supp. Table S3. Main candidate variants obtained through filtering strategies of the exome analysis from families F3788 and F617

	Gene (chromosome)	Variant	rs name	SIFT	Polyphen
Family	<i>YWHAZ</i> (8)	c.-12G>C (splicing)	rs138890711	-	-
	<i>RIMS2</i> (8)	c.946G>C (exonic)	rs200437104	0.01 D	0.92 D
	<i>CDHI</i> (16)^a	c.760G>A (exonic)	-	0.02 D	1 D
Family	<i>VTAI</i> (6)	c.22C>A (exonic)	rs137966407	0.01 D	0.78
	<i>PHACTR2</i> (6)	c.58G>A (exonic)	-	0.01 D	0.98 D
	<i>CDHI</i> (16)^a	c.760G>A (exonic)	-	0.02 D	1 D

Variant names based on cDNA sequence. **D**: Damaging

^a According to MIM data, *CDHI* is the only gene of the list that has been implicated in oral clefting. The shared variant c.760G>A was thus considered the most likely pathogenic.

Supp. Table S4. Predictions from *in silico* analyses^a, using *CDH1* Ensembl transcript 261769; only variants predicted to be non-neutral are displayed (variant names based on cDNA sequence)

Intronic / synonymous variants ^b							
	MaxEntScan (-20 – 20)	NNSplice (0-1)	HSF 3.0 (0-20)				
c.387+5G>A	9.80 → 7.88 <i>Decreased recognition of splice donor site</i>	0.99 → 0.64 <i>Decreased recognition of splice donor site</i>	100 → 99 <i>No effect</i>				
c.2514C>T;	-0.69 → 2.17	0.59 → 0.69	39 → 69				
p.Ser838=	<i>No effect</i>	<i>No effect</i>	<i>Activation of exonic cryptic donor site</i>				
Missense variants							
	Polyphen-2 (0 – 1)	SIFT (0 – 1)	PROVEAN (-13 – 4)	Align GV/GD (C0-C65)	Mutation Taster ^c	PhyloP (-14 – 6)	PhastCons (0 – 1)
c.760G>A;	1 (HumDiv),	0.022	-4.622	0 / 23.01	Disease	5.3	1
p.Asp254Asn	0.997 (HumVar) <i>Probably damaging</i>	<i>Damaging</i>	<i>Deleterious</i>	<i>Class</i> <i>C15</i>	causing (prob: 0.99)	<i>Conserved</i>	<i>Conserved</i>
c.2351G>A;	1 (HumDiv),	0.021	-4.364	0 / 28.82	Disease	5.8	1
p.Arg784His	0.999 (HumVar) <i>Probably damaging</i>	<i>Damaging</i>	<i>Deleterious</i>	<i>Class</i> <i>C25</i>	causing (prob: 0.99)	<i>Conserved</i>	<i>Conserved</i>

^a Ranges of score variation are informed below the program names. Categorical predictions based on the following cut-off values: Polyphen-2 HumDiv: Probably damaging if ≥ 0.957 ; Polyphen-2 HumVar: Probably damaging if ≥ 0.909 ; SIFT: Damaging if ≤ 0.05 ; PROVEAN: Deleterious if ≤ -2.5 ; Align GV/GD: Deleterious if Class $\geq C15$; PhyloP: Conserved if > 0 ; Phastcons: Conserved if > 0 ; MaxEntScan: splice site if ≥ 3 (alteration in splice site if variation between WT and mutant $> 10\%$); NNSplice: splice site if ≥ 0.95 (alteration in splice site if variation between WT and mutant $> 10\%$); HSF 3.0: splice site if ≥ 65 (alteration in splice site if variation between WT and mutant $> 30\%$).

^b Values displayed on the left and on the right side of the arrow refer, respectively, to WT and mutant alleles.

^c Mutation taster prediction is given as either “polymorphism” or “disease causing”. A “probability value” (prob) indicates the security of the prediction, being 1 the most secure.

Supp. Table S5. *CDHI* variants included in SKAT and two-tailed Fisher's exact test

Variant	AC in 609 controls	AC in 221 patients	AC in 26 probands from familial cases	<i>In silico</i> prediction tools	Scores (range) and predictions	Mean coverage in NGS		
						Exomes (patients)	TGS (patients)	Exomes (controls)
c.88C>A	2	0	0	Polyphen-2 PROVEAN	0.99 (0-1), probably damaging -3.07 (-13 - 4), deleterious	127x	70x	65x
c.324A>G	6	1	1	MaxEntScan	5.61→6.86 (creation of ESS)	32x	223x	98x
c.375C>A	1	0	0	MaxEntScan	8.87→7.22 (donor decreased)	25x	150x	55x
c.387+5G>A	0	1	0	MaxEntScan NNSplice	9.80→7.88 (donor decreased) 0.99→0.64 (donor decreased)	25x	130x	45x
c.670C>T	1	0	0	Polyphen-2 SIFT PROVEAN	0.65 (0-1), possibly damaging 0.01 (0-1), damaging -3.34 (-13 - 4), deleterious	26x	61x	25x
c.760G>A	0	2	2	Polyphen-2 SIFT PROVEAN	1 (0-1), probably damaging 0.022 (0-1), damaging -4.62 (-13 - 4), deleterious	55x	124x	80x
c.892G>A	1	0	0	Polyphen-2 SIFT	0.93 (0-1), possibly damaging 0.001 (0-1), damaging	67x	66x	166x
c.1008+8G>T	1	0	0	HSF 3.0	36→66 (activation of ICDS)	35x	99x	72x
c.1023T>G	0	1	1	No prediction (stopgain mutation)		40x	106x	75x
c.1138-3C>T	1	0	0	MaxEntScan	7.38→6.19 (acceptor decreased)	49x	196x	70x
c.1633C>G	1	0	0	Polyphen-2 PROVEAN	0.79 (0-1), possibly damaging -3.22 (-13 - 4), deleterious	50x	101x	34x
c.1774G>A	4	1	1	Polyphen-2 SIFT PROVEAN	0.49 (0-1), possibly damaging 0.002 (0-1), damaging -2.75 (-13 - 4), deleterious	72x	200x	165x
c.2201G>C	1	0	0	Polyphen-2 SIFT PROVEAN	0.98 (0-1), probably damaging 0.002 (0-1), damaging -4.76 (-13 - 4), deleterious	30x	105x	34x
c.2351G>A	0	1	1	Polyphen-2 SIFT PROVEAN	0.99 (0-1), probably damaging 0.02 (0-1), damaging -4.36 (-13 - 4), deleterious	33x	46x	131x
c.2369C>T	1	0	0	Polyphen-2 PROVEAN	0.99 (0-1), probably damaging -2.72 (-13 - 4), deleterious	36x	47x	142x
c.2440-6C>G	3	0	0	MaxEntScan	10.93→9.85 (acceptor decreased)	32x	33x	30x
c.2514C>T	0	2	1	HSF 3.0	39→69 (activation of ECDS)	68x	55x	50x
c.2635G>A	1	0	0	Polyphen-2 PROVEAN	1 (0-1), probably damaging -3.41 (-13 - 4), deleterious	74x	107x	62x

Variant names based on cDNA sequence. **AC:** Allele Count; **TGS:** targeted gene sequencing; **HSF:** Human Splicing Finder; **ESS:** exonic splice site; **ICDS:** intronic cryptic donor site; **ECDS:** exonic cryptic donor site.

SUPP. REFERENCES

- Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1092 human genomes. *Nature* 491:56-65.
- Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods* 7:248-249.
- Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. 2012. Predicting the functional effect of amino acid substitution and indels. *PLoS One* 7:e46688.
- Desmet FO, Hamroun D, Lalande M, Collod-Bérout G, Claustres M, Bérout C. 2009. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res* 37:e67.
- Gorlin RJ, Cohen MMJ, Hennekam RCM. 2001. Syndromes of the head and neck. New York: Oxford University Press. 1344 p.
- Horimoto AR, Onodera MT, Otto PA. 2010. PENCALC: A program for penetrance estimation in autosomal dominant diseases. *Genet Mol Biol.* 33:455-459.
- Keller G, Vogelsang H, Becker I, Plaschke S, Ott K, Suriano G, Mateus AR, Seruca R, Biedermann K, Huntsman D, Doring C, Holinski-Feder E, et al. 2004. Germline mutations of the E-cadherin(CDH1) and TP53 genes, rather than of RUNX3 and HPP1, contribute to genetic predisposition in German gastric cancer patients. *J Med Genet* 41:e89.

- Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 7:1073-1081.
- Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* 20:110-121.
- Reese MG, Eeckman FH, Kulp D, Haussler D. 1997. Improved splice site detection in Genie. *J Comput Biol* 4:311-323.
- Schwarz JM, Cooper DN, Schuelke M, Seelow D. 2004. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods* 11:361-362.
- Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15:1034-1050.
- Simões-Correia J, Figueiredo J, Lopes R, Stricher F, Oliveira C, Serrano L, Seruca R. 2012. E-cadherin destabilization accounts for the pathogenicity of missense mutations in hereditary diffuse gastric cancer. *PLoS One* 7:e33783.
- Suriano G, Oliveira MJ, Huntsman D, Mateus AR, Ferreira P, Casares F, Oliveira C, Carneiro F, Machado JC, Mareel M, Seruca R. 2003. E-cadherin germline missense mutations and cell phenotype: evidence for the independence of cell invasion on the motile capabilities of the cells. *Hum Mol Genet* 12:3007-3016.
- Tavtigian SV, Deffenbaugh AM, Yin L, Judkins T, Scholl T, Samollow PB, de Silva D, Zharkikh A, Thomas A. 2006. Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *J Med Genet* 43:295-305.
- Wang J, Wilkinson MF. 2000. Site-directed mutagenesis of large (13-kb) plasmids in a single-PCR procedure. *Biotechniques* 29:976-978.

Yeo G, Burge CB. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* 11:377-394.

Zhang L, Xiao A, Ruggeri J, Bacares R, Somar J, Melo S, Figueiredo J, Simões-Correia J, Seruca R, Shah MA. 2014. The germline CDH1 c.48 G>C substitution contributes to cancer predisposition through generation of a pro-invasive mutation. *Mutat Res* 770:106-111.

Chapter 4

Establishment of *cdh1*-mutant zebrafish lines through CRISPR/Cas9-mediated genome editing

Brito LA¹, Kong Y², Kague E¹, Grimaldi M², Ethier R, Liao EC², Passos-Bueno MR¹

- 1- Centro de Estudos do Genoma Humano e Células-Tronco, Instituto de Biociências, Universidade de São Paulo, SP, Brasil
- 2- Division of Plastic and Reconstructive Surgery, Center for Regenerative Medicine, Massachusetts General Hospital, Harvard Medical School.

Key words: E-cadherin, zebrafish, CRISPR-Cas9, embryonic lethality, orofacial clefts.

Abstract

Epithelial cadherin is a cell-cell adhesion protein. Mutations in its encoding gene, *CDH1*, has long been implicated with cancer, and recent studies have also suggested a role in nonsyndromic cleft lip / palate etiology (NSCL/P). The mechanisms linking *CDH1* mutations and NSCL/P, however, are unknown, and may involve somatic second hits, as observed in cancer. In order to investigate the mechanisms linking *CDH1* mutations and NSCL/P, we used CRISPR/Cas9-mediated genome editing approach to induce mutations in *CDH1* Zebrafish homolog. We were able to establish 4 zebrafish mutation lines, being one loss-of-function (LoF) 3 in-frame mutations. Homozygous fish for the LoF mutation (i.e., *cdh1* complete knockout) die within the first hours of development, corroborating the critical role of E-cadherin during gastrulation. No lethality or craniofacial phenotype (observed with alcian blue staining) was observed for any other fish (homozygous for the in-frame mutations or compound heterozygous between any two mutations). We believe that the standardization of CRISPR/Cas9 methodology, as well as the establishment of a zebrafish line harboring a LoF mutation in *cdh1*, will be of extreme relevance for further studies, in order to clarify the etiological role of *CDH1* in NSCL/P.

Resumo

A caderina epitelial, codificada pelo gene *CDH1*, é uma proteína que promove adesão celular entre células epiteliais vizinhas. *CDH1* tem papel amplamente conhecido como gene supressor de tumor, implicado em diversos tipos de câncer. Alguns estudos também relacionam mutações no *CDH1* à ocorrência de fissuras de lábio / palato não sindrômicas (FL/P NS), porém sem um mecanismo genético conhecido. Para investigar esse mecanismo, nós utilizamos o método de edição de genoma CRISPR/Cas9 para induzir mutações no gene homólogo de Zebrafish (*cdh1*). Nós estabelecemos linhagens estáveis de peixes com 4 mutações: uma deleção *frameshift*, levando ao truncamento precoce da proteína, e 3 inserções ou deleções *in-frame*. Nós verificamos que, em homozigose, a deleção *frameshift* é letal nas primeiras 24h do desenvolvimento. Por outro lado, os homozigotos para as outras mutações, bem como os heterozigotos compostos entre quaisquer das quatro mutações, são viáveis e não apresentam fenótipo relacionado ao desenvolvimento ósseo (avaliado por coloração de *alcian blue*). Em conclusão, a padronização da técnica de CRISPR/Cas9 e a geração de linhagens deficientes para *cdh1* constituem importantes avanços para a investigação dos mecanismos genéticos subjacentes à associação entre *CDH1* e NSCL/P.

Introduction

Epithelial cadherin, or E-cadherin, is a membrane protein, essential for establishing the adherens junctions of epithelial cells. The cadherin-mediated cell adhesion relies on a calcium-dependent interaction between two cadherin dimers from adjacent cells at their extracellular domains. Besides the adhesive function, E-cadherin and its intracellular interacting partner, β -catenin, are key regulators of epithelial-mesenchymal transition (EMT; Paredes et al., 2012). During craniofacial development, EMT is responsible for initiating neural crest cell migration, playing also decisive role in the closure of palate and upper lip (Kang and Svoboda, 2005; Thevenneau and Mayor, 2012; Ke et al., 2015; Twigg and Wilkie, 2015).

We previously reported the relevance of rare germline variants in the tumor suppressor gene *CDH1*, encoding E-cadherin, on nonsyndromic cleft lip with or without cleft palate (NSCL/P) genetic etiology (Brito et al., 2015). Among them, one nonsense (p.Tyr341*) and 2 missense variants (p.Asp254Asn and p.Arg784His) segregated with NSCL/P in 4 families. Based on the type of mutation, and on functional cellular studies, a loss-of-function (LoF) mechanism has been proposed for *CDH1* variants in NSCL/P etiology. In addition, we have also speculated the need of a second hit inactivating the wild-type allele at the embryonic tissue that will give rise to lip and palate, similarly to gastric cancer (Grady et al., 2000). In order to address these questions, as well as to functionally investigate the relevance of *CDH1* mutations in craniofacial development, we explored the Zebrafish model (*Danio rerio*) its homolog gene, *cdh1*. Since the elementary processes of craniofacial morphogenesis are similar among vertebrate species (Northcutt and Gans, 1983; Schilling et al., 1996), this animal model represents a good predictive for human craniofacial development, and provides some advantages over other models, such as rapid growing, easy maintenance, and optical transparency of embryos (Westerfield, 2000). Here, we used CRISPR/Cas9-mediated genome editing approach to create *cdh1* mutant zebrafish. This collaborative study was carried out in Eric C. Liao's lab (Craniofacial Developmental Biology Laboratory, Massachusetts General Hospital, Harvard Medical School), where zebrafish has been successfully adopted as a model for craniofacial development (Gfrerer et al., 2014; Kong et al., 2014).

Material and Methods

An Overview of CRISPR/Cas9 System

CRISPR/Cas9 is a genome-editing approach based on an adaptive immune system present in Bacteria and Archea, named CRISPR/Cas (clustered regularly interspaced short palindromic repeats / CRISPR associated; Sorek et al., 2008). With this system, exogenous DNA that enters the bacteria is incorporated in the genome at CRISPR locus, and its transcript will then be able to recognize, by homology, exogenous sequences in a second invasion. Following sequence recognition, the exogenous DNA is degraded in a complex that contains the endonuclease Cas and a trans-activating RNA (tracrRNA), which is necessary for hybridizing with crRNA and recruiting Cas. Among the different types of bacterial CRISPR/Cas systems, the type II needs only a single Cas enzyme (Cas9), and it has been applied in genome editing assays (Mali et al., 2013; Figure 1A).

Analogous to the bacterial system, the CRISPR/Cas9-mediated genome editing system requires the presence of a single-guide RNA (sgRNA), which contains a complementary region to the gene of interest, and a “scaffold” sequence, which recruits Cas9 (as a fusion of bacterial crRNA and tracrRNA; Figure 1B). When injected in a cell, this system induces a double-strand break (DSB) at the target site, which in turn will be repaired by one of two main paths: homology-dependent repair (HDR) or non-homologous end joining (NHEJ). While HDR uses the homologous DNA as a template, NHEJ does not depend on a template strand, and frequently introduces indels near the DSB. Thereby, repairing by NHEJ is preferable when the aim is to truncate a protein and inactivate a gene (knockout; Lieber, 2010). Alternatively, co-injecting a template oligo, with homology to both sides of the double break, may induce HDR, and it could be used to knock-in editing (Chang et al., 2013; Hwang et al., 2013).

Zebrafish Lines

Wild-type (WT) zebrafish of TU strain were raised and cared at the Center for Regenerative Medicine (Massachusetts General Hospital, Harvard Medical School), and at the Instituto de Biociências (Universidade de São Paulo), following standard protocols as per Subcommittee on Research Animal Care, Massachusetts General Hospital.

Single-guide RNA (sgRNA) and Cas9 mRNA Production

Three target sites (TS) to *cdh1* were designed using the program ZiFit Targeter Version 4.2 (<http://zifit.partners.org/ZiFiT/>). For each TS (namely, TS1, TS2 and TS3), we ordered two partially complementary oligos (Supplementary Table I), using assembly for T7 promoter.

The oligos were resuspended in nuclease-free water and each pair was annealed after 95°C incubation for 15min and gradual ramp to room temperature. Annealed oligos were then 5' phosphorylated using T4 Polynucleotide kinase enzyme (New England Biolabs, [NEB]) and cloned into a BsaI-linearized pDR274 plasmid (obtained from AddGene, plasmid #42250). This vector contains a T7 promoter upstream of the sgRNA sequence, and the overhangs of the annealed oligos are compatible with the digested vector backbone¹⁷. One Shot TOP10 Chemically competent *E. coli* (Life Technologies) was used for bacterial transformation according to manufacturer's recommendations, and the transformed bacterial stock was seeded onto LB-agar plates containing 50ug/ml kanamycin. For each TS, 4 isolated colonies were picked and grown overnight in LB broth containing 50ug/ml kanamycin. Plasmid DNA was purified (with QIAprep Spin Miniprep Kit, Qiagen) and sequenced for validation, with M13 primer.

The plasmids carrying the inserts were then digested with DraI (NEB R0129S), and run in a 1% agarose gel. sgRNAs were generated via *in vitro* transcription with MEGAshortscript T7 kit (Life Technologies).

The isolated Cas9 vector (Addgene 42251, MLM3613) was linearized with XbaI (NEB R0145S) and run in a 1% agarose diagnostic gel. Cas9 *in vitro* transcription was performed with the mMESSAGE mMACHINE T3 kit (Life Technologies). Both sgRNAs and the Cas9 mRNA were purified by phenol:chloroform extraction and isopropanol precipitation, and stored at -80°C.

Microinjections

Injection solutions were prepared with 200ng/ul Cas9 mRNA and variable concentrations of sgRNA (30, 45 and 60ng/ul), in order to achieve the best efficiency. Nuclease-free water was used for normalization.

WT zebrafish were set up in breeding tanks with dividers in the night prior to injections, in a ratio of one male to one female. In the next morning, the dividers were pulled to allow breeding, and the embryos were placed in a petri dish containing 1x zebrafish E3 embryo medium and a 1.5%-agarose scaffold, to prevent eggs from moving. Eggs were uniformly positioned as a single column, with the cells facing the microinjector. Needles were prepared by pulling a 10cm borosilicate glass (1mm outer diameter; 0.5mm inner diameter), using a P-97 micropipette puller (Sutter Instrument).

We injected, on average, 150 embryos with injection solution for each of the 9 sgRNA conditions (TS1-30ng/ul, TS1-45ng/ul, TS1-60ng/ul, TS2-30ng/ul, TS2-45ng/ul, TS2-60ng/ul, TS3-30ng/ul, TS3-45ng/ul and TS3-60ng/ul). Microinjections were performed into the blastomere during the one-cell stage embryos, using a volume of 1nl to 3nl of injection solution for each embryo.

After injection, the embryos were raised in E3 medium for 5 days at 28°C. Dead or deformed embryos were removed 1 day post fertilization (dpf). Five dpf, the embryos were placed into tanks with slow running system water at 28.5°C, according to standard protocols (Westerfield, 2000).

DNA Purification

We used alkaline-lysis method for DNA extraction (50mM NaOH followed by 95°C incubation and neutralization with 1M Tris, pH 8). For zebrafish embryos, DNA was purified from pools of 5 1-dpf embryos; DNA from adults was purified from fin clip.

Mutation Screening and Efficiency Evaluation

In order to detect the presence of indels, each TS was amplified by PCR with FAM-labeled primers (Supplementary Table II). Amplification was performed with Platinum PCR Supermix (Life Technologies), according to manufacturer's recommendations. PCR conditions were 94°C/4min (1x), 94°C/25s 59°C/25s 72°C/40s (25x), 72°C/5min (1x). Capillary electrophoresis was performed using ABI3730xl DNA Analyzer [Applied Biosystems], at the DNA Core Facility of Massachusetts General Hospital. GeneMapper v4.0 (Life Technologies) was used for allele calling.

To determine mutation sequences, when indels were detected in embryo pools, FAM-PCR products were cloned into pGEM-T Easy Vector (Promega), and in-house competent cells were used for transformation. Colonies were sequenced using T7 primer, with Sanger method. The proportion of indels detected among all individual clones was used to assess the method's efficiency. When indels were detected by FAM-PCR in adults, PCR products were directly sequenced by Sanger method

Alcian Blue Staining

Six dpf, larvae were fixed in 4% paraformaldehyde and stained with alcian blue, in order to observe cartilaginous structures (50% ethanol washing for 10min, followed by overnight alcian blue staining, ddH2O washing for 10min and bleaching until the pigment is gone)

Results

We firstly injected 9 combinations of CRISPR/Cas9 injection solution (with variable concentration of sgRNA for each TS), in ~150 embryos for each condition. DNA fragment analysis of 5 1dpf-embryo pools revealed mutant alleles in embryos injected with TS2 sgRNA in all concentrations, with higher number of mutants observed with 60ng/ul (Figure 2A). TS1 and TS3, on the other hand, presented poor efficiency. To characterize the mutations and evaluate the method's efficiency, we sequenced

individual clones of the PCR products of TS2-30ng/ul and TS2-60ng/ul conditions. For TS2-30ng/ul, 4 mutant alleles were detected out of 9 sequenced alleles (44%), while 4 mutants out of 5 alleles (80%) were found for TS2-60ng/ul (Figure 2B). After identifying TS2 as the most efficient target site, a larger set of ~500 embryos were re-injected with TS2-30ng/ul and TS -60ng/ul conditions, and raised.

Figure 2 – Injection efficiency of CRISPR/Cas9 assay for each TS and concentration. (A) DNA fragment analysis performed with FAM-PCR products. Each graph represents the genotyping of a pool of 5 1dpf embryos. The presence of peaks other than the expected WT allele size (391bp for TS1, 341bp for TS2 and 395bp for TS3, indicated by black arrows) is evidence of indels (red dashed squares). Image generated with GeneMapper 4.0. **(B)** Mutant allele sequences identified by Sanger sequencing of single clones, for the conditions TS2-30ng/ul and TS2-60ng/ul. In the reference sequence (WT), sgRNA target sequence is highlighted in blue, followed by PAM sequence in red. Lower case represents intronic sequence; capital letters represent exonic sequence, in which bold/non-bold triplets correspond to codons. For the mutant alleles, deletions are indicated by dashes, while insertions are highlighted in green

Even injecting during 1-cell embryo stage with optimal RNA concentrations, F0 fish are expected to be mosaic for *cdh1* mutations. In order to generate non-mosaic *cdh1* mutants, we firstly identified F0 injected fish harboring *cdh1* mutations in the germ line cells (founders), for further crossing. Founders were identified by genotyping 1dpf embryos, produced by crosses between F0 injected fish with WT, as the presence of indels in offspring indicates mutations in germ line cells. Mutant alleles were present in 3 out of 14 crosses, revealing 3 founders (namely A, F and I; Figure 3A). Among these 3 founders, founder A's progeny presented the highest number of mutants. Sequencing of individual clones of PCR products revealed 10 mutants out of 22 sequenced alleles (45%) from Founder A's progeny, and 2 mutant out of 9 alleles (22%) from Founder F's progeny (Figure 3B). Among the 4 mutation types observed in founder A's progeny, there was 1 frameshift deletion ("del5" – p.(Asp129Gly*4)), and 3 in-frame indels ("del3" – p.(Asp129del), "ins9" – p.(Asp129delinsTyrThrIleHis) and "del18" – p.(Val124_Asp129del)).

F1 generation produced by founder A was raised, and heterozygous fish for the 4 mentioned mutations were identified by Sanger sequencing using DNA extracted from fin clip. We did not detect morphological differences by alcian blue staining among heterozygous fish (del3/+, del5/+, ins9/+, del18/+) and WT (+/+). Heterozygous fish developed normally, and were intercrossed to generate homozygous and compound heterozygous (F2). An embryonic lethality of 35% was observed 1dpf in F2 generated by del5/+ intercross (expected genotypic proportions were 1 del5/del5 : 2 del5/+ : 1 +/+). Sequencing of 10 juveniles from this breed revealed 5 +/+ and 5 del5/+, but no del5/del5, suggesting a massive loss of this genotype. No substantial embryo lethality (<2.5%) was observed in the F2 progenies containing the homozygous genotypes ins9/ins9 and del18/del18, or the compound heterozygous del5/ins9 del5/del18 and ins9/del18. In addition, alcian blue staining in 6dpf embryos did not reveal abnormal development of craniofacial cartilage in any of them (data not shown). Del3 was not analyzed, as this deletion is contained in del18.

Figure 3 – Founder identification and F1 sequencing. (A) DNA Fragment analysis of pools of F1 embryos, generated by 3 different founders (A, F and I). All founders were injected with sgRNA for TS2. Black arrows indicate the expected WT-allele peak (341bp). Additional peaks correspond to indels (red dashed squares), indicating that the injected progenitor harbors *cdh1* mutations in germline cells. Image generated by GeneMapper 4.0. **(B)** Mutant alleles identified in F1 generation from founders A and F. In the reference sequence (WT), sgRNA target sequence is highlighted in blue, followed by PAM sequence in red. Lower case represents intronic sequence; capital letters represent exonic sequence, in which bold/non-bold triplets correspond to codons. For the mutant alleles, deletions are indicated by dashes, while insertions are highlighted in green.

Discussion

Genome editing with CRISPR/Cas9 method represents an efficient tool to investigate the phenotypic effects of gene mutations, given its simple design and higher

efficiency, compared to other genome editing methods (Mali et al., 2013). Here, we created stable lines of *cdh1* mutant zebrafish via CRISPR/Cas9-mediated genome editing. Among the mutations generated, “del5” is a frameshift deletion that creates a premature stop codon at position 133 of E-cadherin (p.(Asp129Gly*4)), resulting in LoF. We observed that the complete knockout fish (del5/del5) is lethal during early embryogenesis. This is compatible with E-cadherin’s key role in regulating cell movements during gastrulation (Babb and Marrs, 2004), and is also consistent with previously reported animal models (Kane et al., 2005; Shimizu et al., 2005; Schneider and Kolligs, 2015). Since the partial knockout fish were phenotypically normal, it’s possible that a somatic second hit in *cdh1* in craniofacial primordia, after the critical period of gastrulation, would lead to craniofacial deformities. However, this hypothesis needs to be tested.

No phenotypic alteration was observed for homozygous fish for the in-frame mutations (ins9/ins9 and del18/del18), or for any compound heterozygous (del5/ins9, del5/del18, and ins9/del18). This suggests that the amino acids codified by TS2 region can tolerate substitutions or in-frame variants. In fact, TS2 is located in a poorly conserved region of the precursor domain of *cdh1*, which is cleaved during maturation in endoplasmic reticulum, before transportation to plasma membrane (Ozawa and Kemler, 1990). In humans, disease-associated mutations in this domain are mostly implicated with gastric cancer, with 80% of LoF (nonsense, splice site or frameshift), and 20% of missense mutations (Corso et al., 2012). Therefore, the role in craniofacial morphogenesis played by in-frame or missense variants in *CDH1* precursor domain is still uncertain.

The *cdh1* mutant lines here reported represent a valuable tool for functional studies of *CDH1* or other NSCL/P candidate genes that lay in common gene pathways with *CDH1*. In addition, the standardization of CRISPR/Cas9 technique in zebrafish will help to define disease-implicated genes among candidates obtained from NGS studies.

References

- Babb SG, Marrs JA. 2004. E-cadherin regulates cell movements and tissue formation in early zebrafish embryos. *Dev Dyn* 230:263-277.
- Brito LA, Yamamoto GL, Melo S, Malcher C, Ferreira SG, Figueiredo J, Alvizi L, Kobayashi GS, Naslavsky MS, Alonso N *et al.* 2015. Rare variants in the epithelial cadherin gene

- underlying the genetic etiology of nonsyndromic cleft lip with or without cleft palate. *Hum Mutat*.
- Chang N, Sun C, Gao L, Zhu D, Xu X, Zhu X, Xiong JW, Xi JJ. 2013. Genome editing with RNA-guided Cas9 nuclease in zebrafish embryos. *Cell Res* 23:465-472.
- Corso G, Marrelli D, Pascale V, Vindigni C, Roviello F. 2012. Frequency of CDH1 germline mutations in gastric carcinoma coming from high- and low-risk areas: metanalysis and systematic review of the literature. *BMC Cancer* 12:8.
- Gfrerer L, Shubinets V, Hoyos T, Kong Y, Nguyen C, Pietschmann P, Morton CC, Maas RL, Liao EC. 2014. Functional analysis of SPECC1L in craniofacial development and oblique facial cleft pathogenesis. *Plast Reconstr Surg* 134:748-759.
- Grady WM, Willis J, Guilford PJ, Dunbier AK, Toro TT, Lynch H, Wiesner G, Ferguson K, Eng C, Park JG *et al*. 2000. Methylation of the CDH1 promoter as the second genetic hit in hereditary diffuse gastric cancer. *Nat Genet* 26:16-17.
- Hwang WY, Fu Y, Reyon D, Maeder ML, Tsai SQ, Sander JD, Peterson RT, Yeh JR, Joung JK. 2013. Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nat Biotechnol* 31:227-229.
- Kane DA, McFarland KN, Warga RM. 2005. Mutations in half baked/E-cadherin block cell behaviors that are necessary for teleost epiboly. *Development* 132:1105-1116.
- Kang P, Svoboda KK. 2005. Epithelial-mesenchymal transformation during craniofacial development. *J Dent Res* 84:678-690.
- Ke CY, Xiao WL, Chen CM, Lo LJ, Wong FH. 2015. IRF6 is the mediator of TGFbeta3 during regulation of the epithelial mesenchymal transition and palatal fusion. *Sci Rep* 5:12791.
- Kong Y, Grimaldi M, Curtin E, Dougherty M, Kaufman C, White RM, Zon LI, Liao EC. 2014. Neural crest development and craniofacial morphogenesis is coordinated by nitric oxide and histone acetylation. *Chem Biol* 21:488-501.
- Lieber MR. 2010. The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annu Rev Biochem* 79:181-211.
- Mali P, Esvelt KM, Church GM. 2013. Cas9 as a versatile tool for engineering biology. *Nat Methods* 10:957-963.
- Northcutt RG, Gans C. 1983. The genesis of neural crest and epidermal placodes: a reinterpretation of vertebrate origins. *Q Rev Biol* 58:1-28.
- Ozawa M, Kemler R. 1990. Correct proteolytic cleavage is required for the cell adhesive function of uvomorulin. *J Cell Biol* 111:1645-1650.
- Paredes J, Figueiredo J, Albergaria A, Oliveira P, Carvalho J, Ribeiro AS, Caldeira J, Costa AM, Simoes-Correia J, Oliveira MJ *et al*. 2012. Epithelial E- and P-cadherins: role and clinical significance in cancer. *Biochim Biophys Acta* 1826:297-311.
- Schilling TF, Piotrowski T, Grandel H, Brand M, Heisenberg CP, Jiang YJ, Beuchle D, Hammerschmidt M, Kane DA, Mullins MC *et al*. 1996. Jaw and branchial arch mutants in zebrafish I: branchial arches. *Development* 123:329-344.
- Schneider MR, Kolligs FT. 2015. E-cadherin's role in development, tissue homeostasis and disease: Insights from mouse models: Tissue-specific inactivation of the adhesion protein E-cadherin in mice reveals its functions in health and disease. *Bioessays* 37:294-304.
- Shimizu T, Yabe T, Muraoka O, Yonemura S, Aramaki S, Hatta K, Bae YK, Nojima H, Hibi M. 2005. E-cadherin is required for gastrulation cell movements in zebrafish. *Mech Dev* 122:747-763.
- Sorek R, Kunin V, Hugenholtz P. 2008. CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat Rev Microbiol* 6:181-186.
- Theveneau E, Mayor R. 2012. Neural crest delamination and migration: from epithelium-to-mesenchyme transition to collective cell migration. *Dev Biol* 366:34-54.
- Twigg SR, Wilkie AO. 2015. New insights into craniofacial malformations. *Hum Mol Genet*.
- Westerfield M. 2000. The zebrafish book: a guide for the laboratory use of zebrafish (*Danio rerio*). Eugene, OR: University of Oregon Press.

Supplementary Information

Supplementary Table 1- Sequences of *cdh1* target sites (TS1, TS2 and TS3).

	Exon	Target site (5' – 3')*	Oligonucleotide 1 (5' -3')	Oligonucleotide 2 (5' -3')
TS1	3	GGTGTTTCTGTGCATGCCT <u>GGG</u>	TAGGTGTTTCTGTGCATGCCT	AAACAGGCATGCACAGAAAACA
TS2	4	GGTGGAATCTTCCTCTGATG <u>TGG</u>	TAGGTGGAATCTTCCTCTGATG	AAACCATCAGAGGAAGATTCCA
TS3	7	GGATTTCTGTGCTGGAAAA <u>AGG</u>	TAGGGATTTCTGTGCTGGAAAA	AAACTTTCCAGCACAGAAATC

*PAM sequences adjacent to each TS are underlined

Supplementary Table 2- FAM-labeled PCR primers used for mutation screening after, and expected WT product size.

	Forward (5'-3')	Reverse (5'-3')	Expected size
TS1 (exon 3)	FAM-CTACCACTGCAATCGCTGAAC	GTACTTCACTGCATTGCACAAGA	391
TS2 (exon 4)	FAM-CCAGGCTACATACCTACTATTT	AATGCAGAGACCAATCAGAAA	341
TS3 (exon 7)	FAM-TGTTTTGTGTGCTGAAGTGA	GTCAGTGGCCGTAATAACCA	395

Chapter 5

Association of GWAS loci with nonsyndromic cleft lip and/or palate in Brazilian population

Brito LA¹, Savastano CP¹, Malcher C¹, Ferreira SG¹, Rocha, KM¹, Santos SE², Passos-Bueno MR¹

- 1- Centro de Estudos do Genoma Humano e Células-Tronco, Universidade de São Paulo, SP, Brasil
- 2- Departamento de Patologia, Universidade Federal do Pará, PA, Brasil.

Key words: orofacial clefts, structured association, 8q24, 20q12, medionasal enhancer region

Abstract

Nonsyndromic cleft lip with or without cleft palate (NSCL/P) is a prevalent disorder with complex etiology. Genome-wide association studies (GWAS) have provided insights into the etiological role of common susceptibility variants, unraveling the association of a dozen loci. In this study, we performed an association study to better characterize, in a Brazilian population, the most prominent associated locus, a gene desert at 8q24 region, and also validate additional GWAS associations. To avoid confounding effects due to population stratification, we used a structured association approach, with the information provided by ancestry informative markers. We genotyped, in 620 cases and 675 controls, a dense set of 81 SNP covering 8q24 locus, and also the GWAS hits reported for 1p22, 1q32 (*IRF6*), 10q25, 13q31, 17q22 and 20q12 associated regions. We found significant association for 8 markers in a 310-kb interval at 8q24 (top SNP: rs987525, $P=4.8 \times 10^{-8}$), and also for 20q12 (rs13041247, $P=2.7 \times 10^{-4}$). Stratifying our samples by percentage of European ancestry, we found stronger association for 8q24 in the group with high European ancestry. The 8q24 genomic interval associated in this study partially overlaps a recently characterized long-range regulatory region of *MYC*, that contains the putative enhancer hs1877. In addition, we found borderline statistical significance for markers in 1p22 and 10q25. In conclusion, our results narrow the 8q24 associated interval in Brazilian population and suggest that the candidate enhancer hs1877 may have a major etiological role in this region. We also provide the first evidence of association for 20q12 in our population.

Resumo

Fissura labial com ou sem fissura de palato não sindrômica (FL/P NS) é uma doença comum de etiologia multifatorial. Por meio de estudos de associação de varredura genômica (GWAS), múltiplos loci candidatos têm emergido. Neste artigo, nós conduzimos um estudo de associação para caracterizar melhor o locus mais fortemente associado às FL/P NS, um deserto gênico na região 8q24, e também para validar, na população Brasileira, outras associações reportadas em 1p22, 1q32 (*IRF6*), 10q25, 13q31, 17q22 e 20q12. Para contornar o problema da estratificação populacional decorrente da miscigenação, nós utilizamos a abordagem de *structured association*, por meio da caracterização ancestral da nossa amostra com SNPs informativos de ancestralidade. Nós encontramos associação significativa para 8 marcadores na região 8q24, compreendendo um intervalo de 310kb (associação mais significativa: rs987525, $P=4.8 \times 10^{-8}$), e também para o locus 20q12 (rs13041247, $P=2.7 \times 10^{-4}$). Estratificando nossa amostra de acordo com a porcentagem de ancestralidade europeia, nós verificamos que o grupo formado por indivíduos com mais de 50% de ancestralidade europeia têm associação mais significativa para SNPs na região 8q24. O intervalo associado na região 8q24 se sobrepõe parcialmente a uma região rica em enhancers para *MYC* recentemente caracterizada, incluindo o enhancer putativo hs1877. Nosso estudo também encontrou associação de significância estatística marginal para os loci 1p22 e 10q25. Em conclusão, nossos resultados restringiram, na população brasileira, o intervalo associado na região 8q24, e sugerem que o elemento hs1877 pode desempenhar um papel relevante na etiologia das FL/P NS. Ainda, nós também evidenciamos, pela primeira vez na população brasileira, associação para a região 20q12.

Introduction

Cleft lip with or without cleft palate (CL/P) is one of the most frequent human congenital defects. CL/P prevalence in European populations has been estimated as 1:700, but a pronounced variation has been observed among continental groups, ranging from 0.3:1,000 in Africans populations, to 3.6:1,000, in Amerindians (Gorlin and Cohen Jr., 2001). In 70% of cases, CL/P occurs without any other clinical alteration (nonsyndromic cleft lip with or without cleft palate; NSCL/P). NSCL/P is a complex disease, largely influenced by genetic contribution, as suggested by epidemiological studies (Jugessur and Murray, 2005; Brito et al., 2012b). Before genome-wide association studies (GWAS), several genes were suggested to be implicated with NSCL/P. However, as a general rule, conflicting or non-reproducible results were obtained, with exception for *IRF6* gene (Dixon et al., 2011). Recently, novel susceptibility variants have consistently emerged through GWAS, but they only account for a small fraction of the heritability attributed to the disease (Leslie and Marazita, 2013).

Birnbaum et al. (2009), in the first GWAS conducted on NSCL/P patients (Bonn-I study), reported a strong association signal arising from several markers in a 640-kb interval, within a gene desert in 8q24 region. Subsequent GWAS confirmed this association, placing 8q24 as the main NSCL/P-associated locus in European population, with rs987525 as the strongest associated marker (Grant et al., 2009 [Philadelphia study]; Mangold et al., 2009 [Bonn-II study]; Beaty et al., 2010 [Baltimore study]; Ludwig et al., 2012 [meta-analysis combining Bonn-II and Baltimore]). It has been recently shown that this region is enriched for long-range *MYC* enhancers, clustered in a region termed as “Medionasal Enhancer Region” (MNE), and active during craniofacial development (Uslu et al., 2014). On the other hand, association of this region has been controversial in Asian (Beaty et al., 2010; Sun et al., 2015) and African (Weatherley-White et al., 2011; Figueiredo et al., 2014) populations.

Additional associated loci has emerged in Bonn-II and Baltimore studies, and confirmed afterwards in a meta-analysis: 1p22.1, 1p36.13, 2p21, 3p11.1, 3q12.3, 8q21.3, 10q25.3, 13q31.1, 17p13, 17q22, and 20q12 (Ludwig et al., 2012). Collectively, these studies analyzed mostly European descent individuals, with a smaller contribution of Asian samples from Baltimore study. In addition, a recent GWAS (Beijing study),

identified association of a new locus, 16p13.3, in a Chinese population, reflecting the importance of testing non-European groups (Sun et al., 2015).

In previous reports from our group with a smaller sample, 8q24 was significantly associated in a 2-stage analysis (Brito et al., 2012c), while *IRF6* was associated only in a subset of the sample (Brito et al., 2012a). In this study, we aimed to provide further insights into 8q24 associated region and neighboring genes, in a large Brazilian case-control sample, and also validate the previously reported GWAS loci in 1p22, 1q32 (*IRF6*), 10q25, 13q31, 17q22 and 20q12 in the admixed Brazilian population.

Methods

Ethics

Informed consent was obtained from each individual or legal tutor. Biological samples were collected in accordance with the Research Ethics Committee of Instituto de Biociências (Universidade de São Paulo).

Subjects

A total of 631 patients and 689 controls were included in this study. Patients were recruited from the following Brazilian centers: Hospital das Clínicas (Universidade de São Paulo-SP), Hospital Menino Jesus (São Paulo-SP), Hospital SOBRAPAR (Campinas-SP), and during surgical missions of the non-profit organization Operation Smile (<http://www.operationsmile.org>) in Barbalha-CE, Fortaleza-CE, Maceió-AL, Santarém-PA, and Rio de Janeiro-RJ. All patients were clinically evaluated, and individuals affected by syndromic cleft lip / palate, or nonsyndromic cleft palate only, were removed from this study. Controls were obtained from CEGH-CEL biobank (University of São Paulo) and also from Hospital Regional do Baixo Amazonas (Santarém-PA).

SNP selection

We covered 3Mb in the 8q24 *locus* with 119 SNPs, of which 54 were distributed within the previously associated 640-kb interval; 5 SNPs were distributed along *MYC* gene, and 19 have been characterized as eQTLs in different tissues for the proximal genes *TMEM75*, *MYC*, *ASAP1*, *GSDMC*), according to Genevar database (Yang et al., 2010). Tag-SNP selection was made with Haploview, using Tagger algorithm (pairwise model, r^2 threshold=0.6), and linkage disequilibrium (LD) patterns from HapMap populations CEU (Utah residents with Northern and Western European ancestry), and JPT+CHB (Japanese in Tokyo, Japan, and Han Chinese in Beijing, China). We also targeted the previously suggested candidate loci *IRF6* (1q32.2; 2 SNPs), 10q25.5 (2 SNPs), 13q31.1 (1 SNP), 17q22 (1 SNP), 1p22.1 (2 SNPs) and 20q12 (1 SNP).

A panel of 122 ancestry-informative markers was selected from the 128-marker panel characterized by Kosoy et al. (2009); 6 markers from the original panel were excluded from this study, after evaluation by Illumina Technical Support.

DNA preparation and genotyping

DNA samples were purified from peripheral blood (according to standard protocols) or saliva (collected with Oragene saliva kits OG-500 and OG-575, and extracted following prepIT-L2P manufacturer's instructions; DNA Genotek Inc., Ottawa, Canada).

Genotyping was performed using a custom Illumina GoldenGate VeraCode assay, on the Illumina BeadXpress platform (Illumina, CA, USA), according to the manufacturer's instructions, with total DNA input of 250ng. Genotype calling was performed with GenomeStudio software (Genotyping module; Illumina).

Quality Control

After genotype calling, SNPs were removed from further analysis if call rate < 90%, minor allele frequency (MAF) < 0.05 or Hardy-Weinberg deviations were detected in controls ($P < 0.005$). Samples were removed if genotyping rate < 90% or if we

detected gender discrepancies (between our records and the GoldenGate assay's internal gender controls). After quality control steps, we proceeded to statistical analysis with 1,295 individuals (620 cases and 675 controls) and 231 SNPs: 120 AIMs (Supplementary Table 1), 81 in 8q24, 3 in *MYC*, 18 eQTLs for genes surrounding 8q24 gene desert, 2 in *IRF6*, 2 in 10q25, 1 in 13q31, 1 in 17q22, 2 in 1p22 and 1 in 20q12 (Supplementary Table 2). Male:female ratio was 1:0.8 in patients and 1:1.1 in controls. Cleft lip only:cleft lip and palate, and familial:isolated ratios were 1:4.1 and 1:1.6, respectively.

Statistical Analysis

Linkage disequilibrium (LD) was estimated with Haploview software, and plotted as Standard D'/LOD color scheme.

Population structure was assessed with Structure 2.3.4 software (Falush et al., 2007), using genotype data of 120 AIMs. The run was performed with 100,000 burn in steps and 100,000 Markov chain Monte Carlo repetitions, with the parameters *admixture model*, *allele frequencies correlated* and *USEPOPINFO=1*. A model of three parental populations (K=3) was assumed, based on the well-known tri-hybrid origin of Brazilian population (Salzano and Sans, 2014). In order to assist the ancestry estimates, we included genotypes from European, African and Amerindian samples, obtained from HapMap Project (57 CEU and 62 YRI; Altshuler et al., 2010) and from Kosoy's study (60 CEU, 128 European descent from New York [NYCPEA], 56 YRI, 19 Bini West African from Niger-Congo region, 23 Kanuri West African from Nigeria, 50 Mayan Amerindians from Guatemala, 26 Quechuan Amerindians from Peru and, 29 Nahuan Amerindians from Mexico; Kosoy et al., 2009). In order to improve the accuracy of the Amerindian ancestry estimates, we included, in our genotyping sample, 28 Northern Brazilian Indians, from different groups (Arara [n=7], Asurini [n=8], Gavião-Kyikateje [n=6], Parakanã [n=1], and other individuals from Xingu river area [n=6]). These samples were kindly provided by S.E.S.

Each SNP was tested for association with NSCL/P using STRAT software, which compares cases and controls with similar ancestry background (Pritchard et al., 2000). Statistical significance threshold was fixed at 4×10^{-4} (Bonferroni correction for 112 independent tests). Odds ratio (OR) with 95% confidence interval (CI) was calculated

for the homozygous and heterozygous genotypes (using homozygous for major allele as reference), and also under dominant and recessive models.

Results

Ancestry Discrimination

In order to provide accurate reference data for native Brazilians, which has never been done for Kosoy's AIM panel, we genotyped a sample of 28 Northern Brazilian Indians. Using $K=3$ populations, this group clustered with the other 3 Amerindian groups from Kosoy et al. (2009), and separated from the other continental groups (Supplementary Figure 1A). Compared exclusively to the other Amerindian groups, most of Brazilian Indians was discriminated from the other populations in a run assuming $K=2$ populations, while Quechuans, Mayans and Nahuans tended to form a single cluster (Supplementary Figure 1B). Even considering the limited resolution of this panel to separate intra-continental groups, this analysis suggests that the Brazilian Indians constituted the most genetically distinct group among all Amerindian groups included. In addition, our Indians seemed to be more similar to the Quechuan group, from Peru, than others (Supplementary Figure 1C), which is consistent with their spatial distribution across America. In particular, a subgroup of Brazilian Indians composed by individuals from Gavião-Kyikatejê population was genetically more similar to Quechuans, than to any Brazilian group, in all analysis performed (Supplementary Figure 1B).

Global ancestry estimates for our patients and controls revealed a highly admixed population, with proportions of European, African, and Amerindian components of, respectively, 0.579, 0.218 and 0.203 for patients and 0.687, 0.175 and 0.138 for controls (Figure 1). These estimates confirmed the expected ancestry profile, based on previous findings for Brazilian population (Pena et al., 2009; Santos et al., 2010). Subdividing our case-control sample by location of origin, the ancestry analysis revealed the European component as the most pronounced across all locations (Supplementary Figure 1A). The strongest Amerindian signal was found in cases and controls from Santarém-PA (0.384 and 0.316, respectively), while patients from Rio de

Janeiro showed the highest African component (0.300; Supplementary Table 2). These estimates were consistent with a previous study from our group, using a lower-resolution AIM panel (Brito et al., 2012c).

Figure 1 – Ancestry profile of case-control sample in Structure run assuming K=3 populations. “Ancestry controls” were incorporated to represent Brazilian parental populations, and assist ancestry inference. European, African and Amerindian components are represented in blue, red and green, respectively. Each single column represents an individual; individuals are ordered from the highest European ancestry to the lowest.

Association analysis

After Bonferroni correction for multiple comparisons, we found significant association for 8 markers in 8q24 ($P < 4 \times 10^{-4}$) and 4 additional markers with borderline statistical significance ($P < 6 \times 10^{-3}$; Table 1). This association block comprises a ~310kb region (~129,690,000-130,000,000), located within Birnbaum’s originally reported 640-kb association block (129,630,800-130,270,800; Figure 2A). The most significant association was found for rs987525 ($P = 4.8 \times 10^{-8}$; $OR_{het} = 2.10$ (1.65-2.68 95%CI), $OR_{hom} = 3.12$ (2.19-4.79 95% CI). This 310-kb association interval comprises multiple LD blocks (Figure 3), and some associated SNPs in this region are in low LD with the top-scored SNP rs987525, suggesting that they represent independent associations, such as rs7817486 ($D' = 0.55$, $r^2 = 0.16$), rs1157136 ($D' = 0.45$, $r^2 = 0.16$), rs16903635 ($D' = 0.38$, $r^2 = 0.14$), rs1367969 ($D' = 0.31$, $r^2 = 0.09$) and rs7388409 ($D' = 0.34$, $r^2 = 0.11$). No association was found for SNPs covering *MYC*, or for any eQTLs for neighboring genes (*TMEM75*, *MYC*, *GSDMC* and *ASAP1*).

Figure 2 – Significance distribution for SNPs in 8q24. (A) Distribution of all SNPs, from analysis based on all individuals. Our associated interval (black dashed square) is contained within the associated interval from Bonn-I study (red dashed square), and partially overlaps MNE region (green box) and the regulatory element *hs1877* (gray box). The proximal *MYC* gene is also represented (orange box). Stratified analyses were also performed, with individuals presenting **(B)** minimum European ancestry component of 50% or **(C)** maximum European ancestry component of 50% (zoom in Bonn-I associated interval)

Since 8q24 has been preferentially associated in European population, we stratified our case-control sample based on European contribution. Individuals classified by Structure as more than 50% European were grouped together (“EUR>50%”; 419 patients and 527 controls) while the remaining individuals were group in “EUR<50%” subsample (either with predominance of African or Amerindian contribution; 201 patients and 148 controls). This analysis revealed a stronger association in EUR>50% subsample than in EUR<50%, comprising the same 310-kb associated interval as in total sample, although individual *P* values were less significant than previously (top-EUR>50% SNPs: rs16903635, $P=1.04 \times 10^{-5}$; rs7388409, $P=1.42 \times 10^{-5}$; rs987525, $P=2.94 \times 10^{-5}$; Table 1 and Figure 2B). In EUR<50% group, fewer SNPs were associated, in a ~70-kb interval (top-EUR<50% SNPs: rs987525, $P=1.0 \times 10^{-4}$, rs11776303, $P=2.0 \times 10^{-4}$, rs748978, $P=4.0 \times 10^{-4}$; Table 1). Compared to the 310-kb interval associated in the European group, this 70-kb interval (~129,940,000-130,010,000) is shifted towards the distal SNP rs748978, which was exclusively associated in EUR<50% sample, and overlaps a wider portion of MNE region (~129,960,800-130,220,800; Uslu et al., 2014; Figure 2C).

Table 1 – Major associations of 8q24 SNPs

SNP	Position*	Allele	MAF_All	Genotypes	MAF_All	Genotypes	<i>P</i> _{ALL}	OR_het	OR_hom	<i>P</i> _{EUR>50%}	<i>P</i> _{EUR<50%}
			Patients	All Patients	Controls	All Controls		(95% CI)	(95% CI)		
rs7388409	129701778	A/G	0.329 (G)	248 / 278 / 93	0.289 (G)	342 / 276 / 56	1.6x10⁻⁷	1.39 (1.10-1.75)	2.29 (1.58-3.31)	1.4x10⁻⁵	3.3x10 ⁻²
rs1367969	129733232	A/G	0.355 (G)	264 / 272 / 84	0.290 (G)	337 / 284 / 54	2.1x10⁻⁵	1.22 (0.97-1.54)	1.99 (1.36-2.90)	1.5x10⁻⁴	7.7x10 ⁻²
rs16903635	129842198	T/G	0.367 (G)	247 / 290 / 82	0.285 (G)	348 / 269 / 58	1.9x10⁻⁶	1.52 (1.20-1.92)	1.99 (1.37-2.89)	1.0x10⁻⁵	1.3x10 ⁻¹
rs1519849	129896967	G/A	0.251 (A)	335 / 227 / 37	0.332 (A)	278 / 303 / 62	4.8x10 ⁻³	0.62 (0.49-0.79)	0.49 (0.32-0.77)	3.7x10 ⁻²	4.3x10 ⁻¹
rs1157136	129901910	T/C	0.423 (C)	200 / 311 / 106	0.349 (C)	283 / 312 / 79	1.7x10⁻⁵	1.41 (1.11-1.79)	1.90 (1.35-2.68)	1.1x10⁻⁴	1.6x10 ⁻¹
rs2395864	129903563	A/G	0.257 (G)	337 / 243 / 37	0.345 (G)	279 / 320 / 71	1.0x10 ⁻³	0.63 (0.50-0.79)	0.43 (0.28-0.66)	6.4x10 ⁻³	5.4x10 ⁻¹
rs7009139	129914248	A/G	0.492 (G)	158 / 311 / 148	0.542 (G)*	151 / 317 / 207	5.2x10 ⁻³	0.94 (0.71-1.23)	0.68 (0.50-0.93)	3.4x10 ⁻²	8.2x10 ⁻²
rs12542837	129926661	C/T	0.497 (T)	154 / 314 / 150	0.548 (T)*	147 / 315 / 212	3.2x10 ⁻³	0.95 (0.72-1.25)	0.67 (0.50-0.92)	4.1x10 ⁻²	6.0x10 ⁻²
rs987525	129946154	C/A	0.395 (A)	90 / 283 / 213	0.254 (A)	47 / 228 / 360	4.8x10⁻⁸	2.10 (1.65-2.68)	3.23 (2.19-4.79)	2.9x10⁻⁵	1.0x10⁻⁴
rs11776303	129965580	T/C	0.400 (C)	224 / 293 / 101	0.485 (C)	177 / 339 / 157	6.5x10⁻⁵	0.68 (0.53-0.88)	0.51 (0.37-0.70)	6.0x10 ⁻⁴	2.0x10⁻⁴
rs10956454	129965888	C/T	0.196 (T)	403 / 186 / 28	0.205 (T)	432 / 209 / 34	3.9x10⁻⁴	0.95 (0.75-1.21)	0.88 (0.53-1.48)	4.0x10⁻⁴	4.3x10 ⁻²
rs7817486	129987702	T/C	0.487 (C)	162 / 309 / 146	0.563 (C)*	127 / 332 / 212	2.0x10⁻⁴	0.73 (0.55-0.96)	0.54 (0.39-0.74)	6.6x10 ⁻³	2.2x10 ⁻³
rs748978	130003116	C/T	0.168 (T)	404 / 175 / 12	0.116 (T)	418 / 112 / 6	1x2x10 ⁻²	1.62 (1.23-2.13)	2.07 (0.77-5.57)	8.4x10 ⁻¹	4.0x10⁻⁴

MAF: Minor allele frequency; **OR:** Odds ratio; **CI:** Confidence interval; *P*_{ALL}: *P*-value for association test with the whole sample

*chromosome 8, hg19

Figure 3 – LD blocks, displayed by standard D' /LOD color scheme. LD is measured with D' ; darker red squares represents stronger LD. SNPs associated with NSCL/P are identified with *.

In addition, we tested SNPs in regions associated in previous GWAS: *IRF6*/1q32 (rs590223 and rs642961), 1p22 (rs4147811 and rs560426), 10q25 (rs4752028 and rs7078160), 13q31 (rs9574565), 17q22 (rs227730) and 20q12 (rs13041247; Table 3). Significant association, after Bonferroni correction, was only found for 20q12 SNP ($P=2.71 \times 10^{-4}$), $OR_{het} = 0.63$ (0.48-0.82 95% CI), $OR_{hom} = 0.41$ (0.27-0.60 95% CI). Borderline statistical significance was observed for rs4147811 (1p22, $P=3.2 \times 10^{-3}$), and rs4752028 (10q25, $P=5.2 \times 10^{-3}$), while no evidence for association was found for *IRF6*, 13q31 and 17q22 SNPs. Stratifying this analysis by ancestry uncovered a borderline association for rs560426 (1p22; $P=2.6 \times 10^{-3}$) in EUR<50% subsample (Table 2).

Discussion

It is assumed that the relevance of a given susceptibility locus for complex disease may vary among populations, due to factors such as differences in genetic background or interaction with environment (Vilhjammsson and Nordborg, 2013). Therefore, exploring GWAS loci in different populations is a fundamental step to understand the genetic architecture of common variants in NSCL/P, not only for validation purposes, but also to provide clues of the underlying biological mechanisms, and characterize population-specific risk variants.

In admixed populations, such as Brazilian, admixture may represent a strong confounder effect for association studies (Shriner et al., 2011). With the structured association approach, this issue can be minimized by the use of AIMs, as we previously have shown (Brito et al., 2012c). In order to further improve this analysis, we have genotyped an AIM panel containing 120 SNPs (Kosoy et al., 2009) in a group of 28 Northern Brazilian Indians. The characterization of this panel in Brazilian Indians will provide a more accurate ancestry inference for Brazilian individuals genotyped with the same panel. Likewise, it may also provide insights into the history of the Brazilian indigenous peoples.

Table 2 – Association of GWAS hits in our Brazilian case-control sample

SNP	Region	Allele	MAF Patients	Genotypes patients	MAF Controls	Genotypes Controls	P _{ALL}	OR _{het} (95%CI)	OR _{hom} (95%CI)	PEUR>50%	PEUR<50%
rs560426	1p22	C/T	0.479 (T)	162 / 321 / 136	0.458 (T)	125 / 328 / 178	5.8x10 ⁻²	0.78 (0.58-1.04)	0.61 (0.43-0.85)	6.0x10 ⁻²	2.6x10 ⁻³
rs4147811	1p22	G/A	0.318 (A)	285 / 277 / 59	0.353 (A)	280 / 306 / 83	3.2x10 ⁻³	0.89 (0.71-1.12)	0.70 (0.48-1.01)	3.4x10 ⁻¹	5.4x10 ⁻²
rs590223	1q32	A/G	0.362 (G)	254 / 283 / 83	0.344 (G)	295 / 291 / 86	7.29x10 ⁻¹	1.13 (0.89-1.43)	1.12 (0.79-1.58)	9.4x10 ⁻²	9.3x10 ⁻¹
rs642961	1q32	G/A	0.169 (A)	417 / 149 / 25	0.164 (A)	460 / 158 / 27	7.43x10 ⁻¹	1.04 (0.80-1.35)	1.02 (0.58-1.79)	2.9x10 ⁻¹	1.6x10 ⁻¹
rs7078160	10q25	G/A	0.229 (A)	369 / 214 / 34	0.172 (A)	465 / 181 / 24	1.72x10 ⁻²	1.49 (1.17-1.90)	1.79 (1.04-3.06)	2.9x10 ⁻²	2.8x10 ⁻¹
rs4752028	10q25	T/C	0.313 (C)	287 / 275 / 56	0.233 (C)	394 / 237 / 37	5.2x10 ⁻³	1.59 (1.26-2.01)	2.08 (1.34-3.23)	4.4x10 ⁻³	3.8x10 ⁻¹
rs9574565	13q31	C/T	0.265 (T)	321 / 245 / 37	0.281 (T)	187 / 148 / 28	3.24x10 ⁻¹	0.96 (0.73-1.27)	0.77 (0.46-1.30)	6.2x10 ⁻¹	5.2x10 ⁻¹
rs227730	17q22	A/G	0.278 (G)	290 / 246 / 36	0.270 (G)	191 / 141 / 26	5.8x10 ⁻¹	1.15 (0.87-1.51)	0.91 (0.53-1.56)	9.7x10 ⁻¹	6.3x10 ⁻¹
rs13041247	20q12	T/C	0.285 (C)	312 / 232 / 55	0.393 (C)	173 / 205 / 75	2.71x10⁻⁴	0.63 (0.48-0.82)	0.41 (0.27-0.60)	1.2x10 ⁻³	2.0x10 ⁻²

MAF: Minor allele frequency; **OR:** Odds ratio; **CI:** Confidence interval; **P_{ALL}:** *P*-value for association test with the whole sample

A regulatory role has been suggested for the gene desert at 8q24, since its first association with NSCL/P, in Bonn-I GWAS (Birnbaum et al., 2009). Within the 640-kb interval originally identified, a study using murine model reported a ~260-kb region distal to marker rs987525 containing several face-specific enhancers for *MYC* (Uslu et al., 2014). This *cis*-regulatory region (MNE region), was demonstrated to regulate craniofacial development. In addition, deletions in this region may cause orofacial clefts in mice. We show that, in an admixed Brazilian population, the NSCL/P-associated region is a ~310-kb interval, contained within Bonn-I 640-kb interval, and that partially overlaps MNE region. In addition, many of the significantly associated markers are in distinct LD blocks, suggesting that they may be targeting different regulatory elements. In accordance with previous reports, the association was stronger in the European subgroup, although it may have been due to limited statistical power of our small EUR<50% sample. Interestingly, the associated interval in this less-European sample is shifted towards the distal marker rs748978, and overlaps a broader region of MNE. The candidate enhancer element hs1877 showed overlapping position with our associated intervals, either stratifying by ancestry or not. Therefore, our data highlights the importance of this putative regulatory element in NSCL/P etiology in our population.

We also report a positive association for 20q12 SNP rs13041247. Association of this region has been consistently replicated, and seems to be more pronounced in populations of Asian origin than in Europeans (Beaty et al., 2010; Sun et al., 2015). Our study provides the first association of this region in a Brazilian population. In addition, this locus represented the second loci in strength of association in our Brazilian sample, among the SNPs here tested. Borderline associations were revealed for markers in 1p22 and 10q25 in our sample. Marginal significance has been observed for 1p22 markers in Brazilian population (Bagordakis et al., 2012), and association signal has been weaker in European than in Asian samples (Beaty et al., 2010). Similarly, our stratified analysis revealed stronger association signal of 1p22 in our sample with lower European ancestry. Nevertheless, it is possible that the effect sizes of 1p22 and 10q25 markers are small, and our study did not have statistical power to detect them.

IRF6 locus has been consistently associated with NSCL/P in all GWAS performed until present. The functional SNP rs642961 have previously shown to disrupt an AP-2a binding site in an *IRF6* regulatory element (Rahimov et al., 2008). This functional role boosted several association studies of this SNP, with many positive associations. We previously found suggestive association of this marker in a small sample set (Brito et al.,

2012). The larger sample size of the present study provides a more comprehensive view of the relevance of this locus on NSCL/P susceptibility. Here, we observed a lack of association of rs642961, which may indicate that this SNP is not the main susceptibility factor in *IRF6* locus, as suggested by others (Sun et al., 2015). Finally, we did not find association of markers in 13q31 and 17q22. Since a single SNP was tested for each region, it is possible this lack of association is due to different LD patterns in our population.

In summary, we found positive association of several unlinked markers in 8q24, which may be driven by different regulatory elements. We observed an overlap of our association interval with a recently characterized regulatory element of *MYC*, hs1877, which may play a major regulatory role in this region. In addition, we report significant association of a marker in 20q12 region, for the first time in a Brazilian population.

Acknowledgments

We are thankful to Operation Smile team, and its coordinators Luciana Garcia and Clovis Brito. We also thank Jocivan Pedroso, Domingos Neto and Hospital Regional do Baixo Amazonas board of directors, for helping with the ascertainment of control samples. This work was supported by grants from FAPESP and CNPq.

References

- Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Peltonen L *et al.* 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* 467:52-58.
- Bagordakis E, Paranaíba LM, Brito LA, de Aquino SN, Messetti AC, Martelli-Junior H, Swerts MS, Graner E, Passos-Bueno MR, Coletta RD. 2012. Polymorphisms at regions 1p22.1 (rs560426) and 8q24 (rs1530300) are risk markers for nonsyndromic cleft lip and/or palate in the Brazilian population. *Am J Med Genet A* 161A:1177-1180.
- Beatty TH, Murray JC, Marazita ML, Munger RG, Ruczinski I, Hetmanski JB, Liang KY, Wu T, Murray T, Fallin MD *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.
- Birnbaum S, Ludwig KU, Reutter H, Herms S, Steffens M, Rubini M, Baluardo C, Ferriani M, Almeida de Assis N, Alblas MA *et al.* 2009. Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24. *Nat Genet* 41:473-477.

- Brito LA, Bassi CF, Masotti C, Malcher C, Rocha KM, Schlesinger D, Bueno DF, Cruz LA, Barbara LK, Bertola DR *et al.* 2012a. IRF6 is a risk factor for nonsyndromic cleft lip in the Brazilian population. *Am J Med Genet A* 158A:2170-2175.
- Brito LA, Meira JG, Kobayashi GS, Passos-Bueno MR. 2012b. Genetics and management of the patient with orofacial cleft. *Plast Surg Int* 2012:782821.
- Brito LA, Paranaíba LM, Bassi CF, Masotti C, Malcher C, Schlesinger D, Rocha KM, Cruz LA, Barbara LK, Alonso N *et al.* 2012c. Region 8q24 is a susceptibility locus for nonsyndromic oral clefting in Brazil. *Birth Defects Res A Clin Mol Teratol* 94:464-468.
- Dixon MJ, Marazita ML, Beaty TH, Murray JC. 2011. Cleft lip and palate: understanding genetic and environmental influences. *Nat Rev Genet* 12:167-178.
- Falush D, Stephens M, Pritchard JK. 2007. Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol Ecol Notes* 7:574-578.
- Figueiredo JC, Ly S, Raimondi H, Magee K, Baurley JW, Sanchez-Lara PA, Ihenacho U, Yao C, Edlund CK, van den Berg D *et al.* 2014. Genetic risk factors for orofacial clefts in Central Africans and Southeast Asians. *Am J Med Genet A* 164A:2572-2580.
- Gorlin RJ, Cohen Jr. MMH, R. C. M. 2001. *Syndromes of the Head and Neck*. New York, NY, US: Oxford University Press. p 905-907.
- Grant SF, Wang K, Zhang H, Glaberson W, Annaiah K, Kim CE, Bradfield JP, Glessner JT, Thomas KA, Garrius M *et al.* 2009. A genome-wide association study identifies a locus for nonsyndromic cleft lip with or without cleft palate on 8q24. *J Pediatr* 155:909-913.
- Jugessur A, Murray JC. 2005. Orofacial clefting: recent insights into a complex trait. *Curr Opin Genet Dev* 15:270-278.
- Kosoy R, Nassir R, Tian C, White PA, Butler LM, Silva G, Kittles R, Alarcon-Riquelme ME, Gregersen PK, Belmont JW *et al.* 2009. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum Mutat* 30:69-78.
- Leslie EJ, Marazita ML. 2013. Genetics of cleft lip and cleft palate. *Am J Med Genet C Semin Med Genet* 163C:246-258.
- Ludwig KU, Mangold E, Herms S, Nowak S, Reutter H, Paul A, Becker J, Herberz R, AlChawa T, Nasser E *et al.* 2012. Genome-wide meta-analyses of nonsyndromic cleft lip with or without cleft palate identify six new risk loci. *Nat Genet* 44:968-971.
- Mangold E, Ludwig KU, Birnbaum S, Baluado C, Ferrian M, Herms S, Reutter H, de Assis NA, Chawa TA, Mattheisen M *et al.* 2009. Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. *Nat Genet* 42:24-26.
- Pena SD, Bastos-Rodrigues L, Pimenta JR, Bydlowski SP. 2009. DNA tests probe the genomic ancestry of Brazilians. *Braz J Med Biol Res* 42:870-876.
- Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. 2000. Association mapping in structured populations. *Am J Hum Genet* 67:170-181.
- Rahimov F, Marazita ML, Visel A, Cooper ME, Hitchler MJ, Rubini M, Domann FE, Govil M, Christensen K, Bille C *et al.* 2008. Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip. *Nat Genet* 40:1341-1347.
- Salzano FM, Sans M. 2014. Interethnic admixture and the evolution of Latin American populations. *Genet Mol Biol* 37:151-170.
- Santos NP, Ribeiro-Rodrigues EM, Ribeiro-Dos-Santos AK, Pereira R, Gusmao L, Amorim A, Guerreiro JF, Zago MA, Matte C, Hutz MH *et al.* 2010. Assessing individual interethnic admixture and population substructure using a 48-insertion-deletion (INSEL) ancestry-informative marker (AIM) panel. *Hum Mutat* 31:184-190.
- Shriner D, Adeyemo A, Ramos E, Chen G, Rotimi CN. 2011. Mapping of disease-associated variants in admixed populations. *Genome Biol* 12:223.
- Sun Y, Huang Y, Yin A, Pan Y, Wang Y, Wang C, Du Y, Wang M, Lan F, Hu Z *et al.* 2015. Genome-wide association study identifies a new susceptibility locus for cleft lip with or without a cleft palate. *Nat Commun* 6:6414.
- Uslu VV, Petretich M, Ruf S, Langenfeld K, Fonseca NA, Marioni JC, Spitz F. 2014. Long-range enhancers regulating Myc expression are required for normal facial morphogenesis. *Nat Genet* 46:753-758.
- Weatherley-White RC, Ben S, Jin Y, Riccardi S, Arnold TD, Spritz RA. 2011. Analysis of genomewide association signals for nonsyndromic cleft lip/palate in a Kenya African Cohort. *Am J Med Genet A* 155A:2422-2425.

Yang TP, Beazley C, Montgomery SB, Dimas AS, Gutierrez-Arcelus M, Stranger BE, Deloukas P, Dermitzakis ET. 2010. Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. *Bioinformatics* 26:2474-2476.

Supplementary Information

Supplementary Figure 1. Detailed ancestry profile by region of origin

Supplementary Table 1. Panel of 120 ancestry informative markers used in this study.

Supplementary Table 2. SNPs included in association analysis, after quality control.

Supplementary Table 3. Average proportions of ancestry contributions by local of ascertainment.

Supplementary Figure 1 – Detailed ancestry profile by region of origin.

(A) Ancestry discrimination for patients (upper panel) and controls (middle panel), in Structure run assuming $K=3$ populations. We incorporated ancestry controls in the run (lower panel), either from databases, or a genotyped sample of Brazilian Indians. European, African and Amerindian components are represented in blue, red and green, respectively. Each single column represents an individual. **(B)** Runs including only Amerindian samples: Brazilian Indians ($n=28$), Quechuans from Peru ($N=26$), Nahuans from Mexico ($N=29$) and Mayans from Guatemala ($N=50$). Even with the limited resolution of this AIM panel, it is possible to notice that Brazilian Indians tend to form a separate cluster in run assuming $K=2$ populations (upper panel); Quechuans are the second group to separate from Nahuans+Mayans ($K=3$, middle panel), and the four groups are roughly separated in $K=4$. It is also possible to observe that a subgroup of Brazilian Indians (Gavião-Kyikatejê, dotted square) tend to group together with Quechuans, in all runs, constituting a group genetically more distant than its Brazilian counterparts (Arara, Asurini, Parakanã and individuals from Xingu river area). **(C)** Composition of “population components” for each Amerindian group, inferred by Structure in runs assuming $K=2$ to $K=4$ populations.

Supplementary Table 1 – Panel of 120* ancestry informative markers used in this study.

rs name	Chr	Position	Alleles	rs name	Chr	Position	Alleles
rs2986742	1	6550376	[A/G]	rs7844723	8	122908503	[T/C]
rs647325	1	18170886	[T/C]	rs2001907	8	140241181	[A/G]
rs4908343	1	27931698	[A/G]	rs1408801	9	12672320	[T/C]
rs1325502	1	42360270	[T/C]	rs10511828	9	28628500	[T/C]
rs12130799	1	55663372	[T/C]	rs3793451	9	71659280	[T/C]
rs3118378	1	68849687	[T/C]	rs2306040	9	93641199	[A/G]
rs3737576	1	101709563	[T/C]	rs10513300	9	120130206	[A/G]
rs7554936	1	151122489	[T/C]	rs2073821	9	135933122	[A/G]
rs1040404	1	168159890	[A/G]	rs3793791	10	50841704	[T/C]
rs1407434	1	186149032	[A/G]	rs4746136	10	75300994	[A/G]
rs4951629	1	212786883	[A/G]	rs4918842	10	115316812	[A/G]
rs316873	1	242342504	[T/C]	rs4880436	10	134650103	[A/G]
rs798443	2	7968275	[A/G]	rs10839880	11	7850316	[T/C]
rs7421394	2	14756349	[A/G]	rs1837606	11	15838137	[A/G]
rs4666200	2	29538411	[A/G]	rs2946788	11	24010530	[A/C]
rs4670767	2	37941396	[A/C]	rs11227699	11	66898492	[A/G]
rs13400937	2	79864923	[T/G]	rs948028	11	120644447	[T/G]
rs260690	2	109579738	[A/C]	rs2416791	12	11701488	[A/G]
rs10496971	2	145769943	[T/G]	rs1513056	12	17407792	[A/G]
rs2627037	2	179606538	[A/G]	rs214678	12	47676950	[T/C]
rs1569175	2	201021954	[T/C]	rs772262	12	56163734	[T/C]
rs10510228	3	2208832	[T/C]	rs2070586	12	109277720	[A/G]
rs4955316	3	30415612	[T/G]	rs1503767	12	118889488	[A/C]
rs9809104	3	39146429	[A/G]	rs9319336	13	27624356	[A/G]
rs6548616	3	79399575	[A/G]	rs7997709	13	34847737	[A/G]
rs12629908	3	120522716	[A/G]	rs9530435	13	75993887	[A/G]
rs734873	3	147750355	[A/G]	rs9522149	13	111827167	[T/C]
rs2030763	3	179964727	[T/C]	rs1760921	14	20818131	[A/G]
rs1513181	3	188574996	[T/C]	rs2357442	14	52607967	[T/G]
rs9291090	4	5390637	[T/G]	rs1950993	14	58238687	[T/G]
rs10007810	4	41554364	[A/G]	rs8021730	14	67886781	[T/G]
rs1369093	4	73245191	[A/G]	rs946918	14	83472868	[T/G]
rs385194	4	85309078	[A/G]	rs3784230	14	105679055	[T/C]
rs7657799	4	105375423	[A/C]	rs12439433	15	36220035	[T/C]
rs2702414	4	179399523	[A/G]	rs2899826	15	74734500	[T/C]
rs316598	5	2364626	[A/G]	rs8035124	15	92105708	[T/G]
rs870347	5	6845035	[T/G]	rs4984913	16	740466	[T/C]
rs6451722	5	43711378	[A/G]	rs4781011	16	10975311	[A/C]
rs6556352	5	155471714	[A/G]	rs818386	16	65406708	[A/G]
rs1500127	5	165739982	[T/C]	rs2966849	16	85183682	[A/G]
rs6422347	5	177863083	[A/G]	rs1879488	17	1401613	[T/G]
rs1040045	6	4747159	[A/G]	rs2033111	17	53788280	[A/G]
rs2504853	6	12535111	[T/C]	rs11652805	17	62987151	[T/C]
rs7745461	6	21911616	[A/G]	rs10512572	17	69512099	[T/C]

rs192655	6	90518278	[A/G]	rs2125345	17	73782191	[A/G]
rs4463276	6	145055331	[A/G]	rs4798812	18	9420504	[A/G]
rs4458655	6	163221792	[T/C]	rs4800105	18	19651982	[T/C]
rs1871428	6	168665760	[T/C]	rs7238445	18	49781544	[A/G]
rs731257	7	12669251	[A/G]	rs881728	18	59333108	[A/C]
rs32314	7	32179124	[A/G]	rs4891825	18	67867663	[T/C]
rs2330442	7	42380071	[A/G]	rs874299	18	75056284	[T/C]
rs4717865	7	73454199	[A/G]	rs8113143	19	33652247	[T/G]
rs10954737	7	83533047	[A/G]	rs3745099	19	52901905	[A/G]
rs705308	7	97695363	[T/G]	rs2532060	19	55614923	[A/G]
rs7803075	7	130742066	[A/G]	rs6104567	20	10195433	[A/C]
rs10236187	7	139447377	[A/C]	rs3907047	20	54000914	[A/G]
rs10108270	8	4190793	[T/G]	rs2835370	21	37885625	[A/G]
rs3943253	8	13359500	[A/G]	rs1296819	22	18076546	[A/C]
rs1471939	8	28941305	[T/C]	rs4821004	22	32366359	[T/C]
rs12544346	8	86424616	[T/C]	rs5768007	22	48207872	[A/G]

Chr: chromosome; positions according hg19 reference

*120-marker version of the original panel (Kosoy et al., 2009)

Supplementary Table 2 – SNPs included in association analysis, after quality control

rs name	Inclusion criteria	Chr	Position*	Alleles	PALL
rs4147811	1p22	1	94575056	[A/G]	3.20E-03
rs560426	1p22	1	94553438	[T/C]	5.80E-02
rs590223	<i>IRF6</i>	1	209946707	[A/G]	7.29E-01
rs642961	<i>IRF6</i>	1	209989270	[A/G]	7.44E-01
rs6470563	8q24; eQTL (<i>TMEM75</i>)	8	128708570	[A/C]	7.90E-01
rs16902364	8q24; <i>MYC</i>	8	128745266	[A/C]	9.08E-02
rs4645943	8q24; <i>MYC</i>	8	128747471	[T/C]	5.61E-01
rs3891248	8q24; <i>MYC</i>	8	128750139	[A/T]	3.09E-01
rs2542419	8q24; eQTL (<i>MYC</i>)	8	129153280	[A/C]	4.33E-01
rs1827135	8q24; eQTL (<i>MYC</i>)	8	129158236	[T/C]	4.28E-01
rs4288340	8q24; eQTL (<i>MYC</i>)	8	129202654	[T/C]	3.98E-01
rs998018	8q24; eQTL (<i>MYC</i>)	8	129278791	[A/G]	9.37E-01
rs13248221	8q24	8	129452035	[A/C]	1.85E-01
rs12543535	8q24	8	129473012	[T/G]	2.79E-01
rs1400473	8q24	8	129474006	[A/G]	3.92E-01
rs6996647	8q24	8	129477436	[A/G]	6.51E-01
rs6989586	8q24	8	129479399	[T/C]	1.80E-01
rs16903006	8q24	8	129484159	[T/G]	4.96E-01
rs10095762	8q24	8	129498584	[A/G]	2.91E-01
rs10956429	8q24	8	129514987	[T/C]	3.73E-01
rs10505513	8q24	8	129525346	[T/C]	3.08E-02
rs17232510	8q24	8	129532907	[T/C]	2.46E-02
rs6995553	8q24	8	129540236	[A/G]	1.20E-02

rs1516971	8q24	8	129542100	[A/G]	2.60E-01
rs7839493	8q24	8	129546651	[T/C]	2.36E-01
rs10113762	8q24	8	129552202	[T/C]	5.00E-01
rs7824269	8q24	8	129552344	[A/G]	2.66E-02
rs1561927	8q24	8	129568078	[T/C]	5.86E-02
rs2165806	8q24	8	129569551	[C/G]	7.00E-01
rs16903109	8q24	8	129577629	[T/A]	9.32E-01
rs1443123	8q24	8	129592794	[A/G]	6.74E-02
rs11780763	8q24	8	129596286	[T/G]	7.58E-01
rs1372992	8q24	8	129597659	[A/G]	1.80E-03
rs7818548	8q24	8	129598701	[C/G]	2.04E-01
rs2099897	8q24; eQTL (MYC)	8	129608359	[A/G]	2.80E-01
rs6993958	8q24	8	129617654	[A/G]	5.90E-01
rs1476165	8q24; eQTL (MYC)	8	129620475	[A/G]	2.06E-01
rs16903308	8q24	8	129632852	[A/G]	2.83E-01
rs1432011	8q24	8	129662612	[A/C]	8.32E-01
rs2395848	8q24	8	129672688	[T/A]	1.89E-01
rs1073754	8q24	8	129688424	[T/C]	3.62E-01
rs7388409	8q24	8	129701778	[A/G]	1.35E-07
rs7845087	8q24	8	129717910	[A/G]	4.85E-01
rs1367969	8q24	8	129733232	[T/C]	2.08E-05
rs6986864	8q24	8	129746506	[G/C]	2.34E-01
rs6997983	8q24	8	129748718	[T/C]	1.48E-01
rs6994324	8q24	8	129832156	[T/C]	2.84E-01
rs16903635	8q24	8	129842198	[T/G]	1.87E-06
rs4733653	8q24	8	129844671	[T/C]	6.04E-02
rs17240568	8q24	8	129861680	[A/G]	2.42E-02
rs1519849	8q24	8	129896967	[T/C]	4.80E-03
rs1157136	8q24	8	129901910	[T/C]	1.35E-05
rs2395864	8q24	8	129903563	[T/C]	1.00E-03
rs7009139	8q24	8	129914248	[T/C]	5.20E-03
rs12542837	8q24	8	129926661	[T/C]	3.20E-03
rs987525	8q24	8	129946154	[T/G]	4.83E-08
rs11776303	8q24	8	129965580	[T/C]	6.50E-05
rs10956454	8q24	8	129965888	[A/G]	3.91E-04
rs7817486	8q24	8	129987702	[A/G]	2.00E-04
rs7829061	8q24	8	130000524	[A/C]	1.86E-01
rs12155905	8q24	8	130002469	[A/T]	4.28E-01
rs748978	8q24	8	130003116	[T/C]	1.16E-02
rs894263	8q24	8	130003263	[A/C]	1.90E-01
rs10090304	8q24	8	130003883	[A/C]	2.40E-01
rs7836429	8q24	8	130014200	[A/C]	8.73E-01
rs10956458	8q24	8	130021709	[T/C]	6.98E-01
rs10092510	8q24	8	130022619	[A/G]	5.69E-01
rs9969663	8q24	8	130045034	[C/G]	5.43E-01
rs7013044	8q24	8	130045495	[A/G]	8.06E-01
rs7821930	8q24	8	130089946	[T/C]	6.36E-01
rs7014071	8q24	8	130097929	[G/C]	8.28E-01
rs6989059	8q24	8	130118208	[A/G]	3.49E-01

rs4236738	8q24	8	130156329	[T/C]	3.12E-02
rs4418312	8q24	8	130157804	[G/C]	6.46E-02
rs6470699	8q24	8	130182692	[A/G]	5.45E-01
rs6992168	8q24	8	130189014	[T/G]	7.41E-01
rs7830899	8q24	8	130201569	[A/G]	9.86E-02
rs7819161	8q24	8	130218506	[G/C]	9.03E-01
rs7826130	8q24	8	130224742	[A/G]	7.86E-02
rs11777669	8q24	8	130225474	[A/G]	5.30E-01
rs4733683	8q24	8	130232692	[A/C]	3.05E-01
rs4236741	8q24	8	130241880	[A/C]	2.79E-01
rs10093780	8q24	8	130251745	[A/G]	5.89E-01
rs4403373	8q24	8	130302821	[T/G]	3.94E-01
rs1368137	8q24	8	130332107	[A/C]	4.91E-01
rs1835851	8q24	8	130353227	[T/C]	5.27E-01
rs16904030	8q24	8	130353671	[T/C]	9.90E-02
rs2719221	8q24	8	130367270	[A/G]	3.53E-01
rs2568436	8q24	8	130368583	[G/C]	6.81E-01
rs12544748	8q24	8	130369564	[T/C]	8.37E-01
rs16904056	8q24	8	130386734	[G/C]	3.71E-01
rs10102588	8q24	8	130386915	[T/C]	5.78E-01
rs7004735	8q24	8	130390634	[T/G]	8.12E-01
rs2630520	8q24	8	130402414	[T/G]	6.40E-01
rs2579871	8q24	8	130430411	[T/C]	1.48E-01
rs4076446	8q24; eQTL (<i>ASAP1</i>)	8	130706301	[T/G]	5.41E-01
rs4733729	8q24; eQTL (<i>GSDMC</i>)	8	130732295	[A/C]	9.75E-01
rs4425724	8q24; eQTL (<i>GSDMC</i>)	8	130733030	[A/C]	9.63E-01
rs5006712	8q24; eQTL (<i>GSDMC</i>)	8	130759038	[T/C]	9.87E-01
rs4144738	8q24; eQTL (<i>GSDMC</i>)	8	130760850	[T/C]	1.00E+00
rs4285452	8q24; eQTL (<i>GSDMC</i>)	8	130764791	[T/C]	3.98E-01
rs11775398	8q24; eQTL (<i>ASAP1</i>)	8	131439767	[A/G]	4.86E-01
rs7461215	8q24; eQTL (<i>ASAP1</i>)	8	131458713	[A/C]	2.22E-01
rs1467065	8q24; eQTL (<i>GSDMC</i>)	8	131709637	[T/C]	8.22E-02
rs2376823	8q24; eQTL (<i>ASAP1</i>)	8	131813922	[T/C]	8.69E-01
rs6993816	8q24; eQTL (<i>ASAP1</i>)	8	131814299	[T/C]	2.56E-01
rs4752028	10q25	10	118834991	[A/G]	5.20E-03
rs7078160	10q25	10	118827560	[T/C]	1.72E-02
rs9574565	13q31.1	13	80668874	[T/C]	3.24E-01
rs227730	17q22	17	54773951	[A/G]	7.44E-01
rs13041247	20q12	20	39269074	[T/C]	2.71E-04

Chr: chromosome; **PALL:** *P*-value of structured association test with the whole case-control sample.

*Positions according hg19 genome reference.

Supplementary Table 3 – Average proportions of ancestry contribution by local of ascertainment

		Ancestry components			
	Subpopulation	N	European	African	Amerindian
Patients	Fortaleza-CE	141	0.637	0.200	0.163
	São Paulo-SP	40	0.723	0.197	0.080
	Maceió-AL	12	0.659	0.232	0.109
	Rio de Janeiro-RJ	222	0.614	0.297	0.089
	Santarém-PA	205	0.468	0.148	0.384
Controls	São Paulo-SP	507	0.740	0.180	0.080
	Santarém-PA	168	0.526	0.159	0.316

N: number of individuals

Chapter 6

eQTL mapping reveals *MRPL53* (2p13) as a candidate gene for nonsyndromic cleft lip and/or palate

Masotti C^{1,2}, Brito LA¹, Nica AC³, Sunaga DY¹, Malcher C¹, Ferreira SG¹, Kobayashi GS¹, Savastano CP¹, Agüena M¹, Meyer D⁴, Bueno D², Alonso N⁵, Franco D⁶, Dermitzakis E³, Passos-Bueno MR¹

- 1- Centro de Estudos do Genoma Humano e Células-Tronco, Instituto de Biociências, Universidade de São Paulo, SP, Brasil
- 2- Hospital Sírio-Libanês, São Paulo, SP, Brasil
- 3- Department of Human Genetics and Development, University of Geneva, Switzerland
- 4- Instituto de Biociências, Universidade de São Paulo, SP, Brasil
- 5- Divisão de Cirurgia Plástica, Faculdade de Medicina, Universidade de São Paulo, SP, Brasil
- 6- Departamento de Cirurgia Plástica, Hospital Clementino Braga Filho, Faculdade de Medicina, Universidade Federal do Rio de Janeiro, RJ, Brasil

Key words: orofacial clefts, eQTLs, *MRPL53*, 2p13, orbicularis oris muscle mesenchymal stem cells

Author's note:

This work was conducted by Dra Masotti C, during her postdoc at University of São Paulo and at University of Geneva (Switzerland). She was the main responsible for the eQTL mapping analysis. The main contribution of the second author, Brito LA, lays in the association analysis and gene sequencing.

Abstract

As the vast majority of GWAS hits reported for NSCL/P fall in noncoding regions or gene deserts, it has been speculated that regulatory regions might help to explain the missing heritability of NSCL/P. In this work, we mapped expression quantitative trait loci (eQTLs) in mesenchymal stem cell cultures derived from *orbicularis oris* muscle (OOMMSC) obtained from NSCL/P individuals, and selected the best eQTLs to conduct an association study. *Orbicularis oris* muscle might be compromised in NSCL/P patients, and is an accessible source of mesenchymal stem cells, as it is discarded during reconstructive surgeries. By correlating genome-wide expression levels and genotypes of OOMMSC, we mapped 119 eQTLs related to 18 genes ($P < 0.0001$; FDR=14%). Through a case-control study of 624 patients and 668 controls, we detected one OOMMSC *cis*-eQTL significantly associated with the disease (rs1063588, $P = 0.0008$). We herein describe a novel susceptibility locus for NSCL/P, an eQTL for the *MRPL53* gene, which encodes a 39S protein subunit of mitochondrial ribosomes, located at a chromosomal region previously reported as OFC2 (orofacial cleft 2 locus, 2p13).

Resumo

Uma grande parcela das variantes associadas a fissura de lábio / palato não sindrômica (FL/P NS) por meio dos *genome-wide association studies* localiza-se em regiões não-gênicas ou não-codificantes. Por conta disso, frequentemente atribui-se um papel regulatório a estas variantes, e especula-se que uma parcela da herdabilidade perdida em FL/P NS pode ser explicada por elas. Com o intuito de explorar variantes regulatórias, nós conduzimos um estudo de associação de eQTLs (*expression quantitative trait loci*), mapeados em células tronco mesenquimais derivadas de amostras de músculo orbicular do lábio (OOMMSC), obtidas de 46 indivíduos com FL/P NS. Este tecido é possivelmente comprometido em indivíduos FL/P NS, e é frequentemente descartado durante a cirurgia de correção da fissura labial. Por meio da correlação entre dados genômicos genotípicos e de expressão gênica na amostra de OOMMSC, nós mapeamos 119 eQTLs relacionados a 18 genes ($P < 0.0001$; FDR=14%). Por meio de um estudo de associação com 624 pacientes e 668 controles, nós detectamos associação significativa entre um eQTL (rs1063588, $P = 0.0008$) e NSCL/P. Esta variante está localizada na região cromossômica 2p13, anteriormente denominada OFC2 (*orofacial cleft 2*), e é um eQTL para *MRPL53*, que codifica a subunidade ribossômica mitocondrial 39S.

Introduction

Nonsyndromic cleft lip with or without cleft palate (NSCL/P) is a multifactorial complex disorder, with several environmental and genetic susceptibility loci reported (Dixon et al., 2011; Ludwig et al., 2012). Not only for NSCL/P, but also for the majority of other complex disorders, genome-wide association studies (GWAS) have predominantly discovered non-coding genetic variants implicated in disease susceptibility (Hindorff et al., 2009); as a consequence, regulatory roles of GWAS variants have been focus of research (Stranger et al., 2007; Cookson et al., 2009; Nica et al., 2010; Albert and Kruglyak, 2015). As examples, the strongest association reported for NSCL/P lays on a gene desert at 8q24, recently implicated in *MYC* regulation (Uslu et al., 2014). In addition, the associated SNP near *IRF6* (1q32), was shown to disrupt an AP-2 α binding site of a murine craniofacial development enhancer (Rahimov et al., 2008).

In this context, a valuable approach to mine the functional effect genetic variants on a genome-wide basis, specially those with effect over gene expression, is looking for expression quantitative trait loci (eQTLs) in different cell or tissue types (Dimas et al., 2009; Grundberg et al., 2012; Mele et al., 2015). In particular, the use of a disease-related tissue type may be advantageous to characterize the genetic architecture of the disease (Dermitzakis, 2012).

NSCL/P is a developmental disorder: craniofacial structures originate from first and second embryonic pharyngeal arches, and human lip and palate development occur from 4th to 12th weeks of pregnancy (Wilkie and Morriss-Kay, 2001; Gritli-Linde, 2008). The *orbicularis oris* muscle (OOM) is one of the possible affected tissues in NSCL/P patients, and our group has previously demonstrated that OOM is an accessible source of mesenchymal stem cells, capable to undergo chondrogenic, adipogenic, osteogenic, and skeletal muscle differentiation (Bueno et al., 2009). In this work, we dissected the gene expression profile of OOM-derived mesenchymal stem cells (OOMMSC), which can be a suitable cell model, as tissue fragments are regularly discarded during cleft lip surgical repair. By correlating genome-wide expression levels and genotypes obtained from 46 OOMMSC samples, we mapped *cis*-eQTLs, and selected the best candidates to conduct a case-control association study in 631 patients and 689 controls.

Subjects and Methods

Ethics

This study was approved by the Research Ethics Committee of the Instituto de Biociências of Universidade de São Paulo (Brazil), and informed consent was obtained from patients, controls or legal tutors.

Orbicularis Oris Muscle Samples

We obtained samples of OOM from 47 individuals (43 affected and 4 non-affected), subjected to reconstructive plastic surgery in 3 different collaborating centers: Hospital das Clínicas (Divisão de Cirurgia Plástica, Universidade de São Paulo; São Paulo-SP); Hospital Sobrapar (Campinas-SP), and the non-governmental organization Operation Smile, during surgical missions in Rio de Janeiro-RJ). Subjects were clinically evaluated, in order to exclude syndromic and cleft palate only cases (Supplementary Table 1). One patient was excluded from eQTL analysis due to missing expression data.

Cell Culture Conditions

We established *orbicularis oris* muscle mesenchymal stem cell (OOMMSC) primary cultures according to previously published protocol (Bueno et al., 2009) and expanded to 70-90% confluence. RNA and DNA extractions of cell cultures in 3rd to 4th passages followed the manufacturer's protocols (Nucleo Spin Tissue and RNAII column kits; Macherey-Nagel, Germany). We confirmed the mesenchymal immunophenotype of a representative sample of cell cultures (N=10) through flow cytometry analysis (EasyCyte Flow cytometer, Guava Technologies, Hayward, CA) with fluorescent antibodies for markers CD29 (adhesion), CD31 (endothelial), CD45 (hematopoietic), CD73 or CD166, CD90, and CD105 (mesenchymal; Supplementary Table 2)

Gene Expression Quantification

To measure transcript levels, we used the Human Gene Chip 1.0 ST v1 microarrays (Affymetrix, USA), according to the manufacturer's protocol, using 300ng of total RNA. We processed all samples in batches, but with randomized phenotypic order (cleft lip only [CLO], cleft lip and palate [CLP], or controls). Gene expression values were obtained with the three-step Robust Multi-array Average method (RMA), implemented in the software Expression Console (Affymetrix, USA; Irizarry et al., 2003). The hybridization intensity values were background corrected, $\log_2$ transformed and then quantile normalized. We removed batch effect from $\log_2$ RMA data using an empirical Bayesian method (ComBat, parametric test; Johnson et al., 2007), and ComBat output data was used as input for eQTL analysis. Further, we excluded from expression data the Affymetrix controls and probe sets classified as '*rescue->FLmRNA->unmapped*' (poorly aligning to transcriptome, *Homo sapiens* NetAffx annotation release 32, 2011-06-21), removing also probe sets with cross hybridization potential.

Microarray Genotyping

We genotyped 500K SNPs with *GeneChip Human Mapping 250K Nsp/250K Sty Arrays*, according to manufactures' protocol (Affymetrix, USA), using 250ng of total DNA. Genotypes were obtained through BRLMM algorithm analysis, implemented in Genotyping Console software (Affymetrix, USA). For eQTL mapping, we rejected SNPs with <100% call rates, with minor allele frequencies (MAF) <5%, or deviating from Hardy-Weinberg equilibrium (HWE; $P < 1 \times 10^{-7}$).

eQTL mapping

For each transcript, obtained from adjusted expression data, (N=25,506 probe-sets) each SNP (N=237,523) we fit a Spearman's Rank Correlation (SRC) model as previously described (Nica et al., 2011; Gutierrez-Arcelus et al., 2013; Bryois et al., 2014). We limited our analysis to *cis*-eQTLs, i.e., we only tested SNPs located within a 1Mb window to either side of the transcription start site (TSS). The significance level to call an eQTL was determined through permutations: we permuted each transcript (Affymetrix probe-set) 10,000 times relative to genotypes and kept the best SRC P value

after each permutation. For each transcript, we ranked 10,000 permutation P values and kept the lowest, the 10th, the 100th and the 500th lowest SRC P values. The lowest permuted P value is the permutation P threshold of 0.0001 (PT1) for the transcript, as well as the 10th lowest P value is the permutation P threshold of 0.001 (PT10). The 100th and the 500th P values are the permutation P thresholds of 0.01 (PT100) and 0.05 (PT500), respectively. We then assessed whether a SNP-transcript association from non-permuted observations had a lower SRC P value than the transcript's permutation P threshold. In a per gene basis, we computed the false discovery rate (FDR) given by each permutation threshold (Stranger et al., 2007; Gutierrez-Arcelus et al., 2013; Bryois et al., 2014). We considered as “best *cis*-eQTLs” the most significant SNP-transcript associations for each transcript; for transcripts with multiple and tied best P values, we chose the closest SNP relative to TSS.

Cross-references databases

To identify OOMMSC eQTLs of NSCL/P candidate genes, we crossed eQTL data with a modified version of a 357-candidate genes list (Jugessur et al., 2009), in which we added 15 other loci based on recent literature (Dixon et al., 2011; Ludwig et al., 2012). We searched for NSCL/P eQTL genes and OOMMSC eQTLs located within ± 1 Mb relative to NSCL/P genes' TSS. Linkage disequilibrium (LD) estimations for eQTL SNPs was calculated in *PLINK* (Purcell et al., 2007) using eQTL-samples' genotypes.

We used GSEA (*Gene Set Enrichment Analysis*) to compute the overlap between significant OOMMSC eQTL genes and annotated sets of genes from Molecular Signatures Database v5.0 (MSigDB), such as Gene Ontology (GO) “Biological Process” and “Molecular Function”, and Curated Gene Sets from pathway databases (Subramanian et al., 2005). Overlaps are computed using HUGO gene symbols and P values estimations based on the hypergeometric distribution for $(k-1, K, N-K, n)$, where k is the number of eQTL genes intersecting the set from MSigDB, K is the number of genes in MSigDB set, N is the total known gene symbols, and n the total genes in the query set.

eQTL Replication Cohort

The Genotype-Tissue Expression (GTEx) Project identified eQTLs for several “nondiseased” human tissues and cell lines by using RNA-seq and imputed genotypes (Mele et al., 2015). GTEx Analysis V6 of Single Tissue eQTL Data (dbGAP accession phs000424.v6.p1) is publicly available, and we accessed seven tissue files with every SNP-gene association test to validate OOMMSC eQTLs. Evaluated tissues were adipose (subcutaneous), brain frontal cortex, skeletal muscle, esophagus mucosa, *esophagus muscularis*, in addition to cell lines of cultured skin fibroblasts and EBV-transformed lymphocytes (LCLs). For quantifying eQTL replication, we estimated q-values for FDR control, using the R package *QVALUE* 2.0.0 default recommended settings (Storey and Tibshirani, 2003). The program takes a list of P values and computes π_0 (the proportion of true null features) based on their distribution (the assumption is that P values of null features are uniformly distributed between $[0,1]$). The quantity $\pi_1=1-\pi_0$ estimates the proportion of true positives. Replication between two samples is reported as the π_1 , estimated from the P value distribution of independent eQTLs (best per transcript) discovered in sample 1 in the second sample (Nica et al., 2011; Grundberg et al., 2012; Bryois et al., 2014).

Case-control sample

A case-control sample, composed by 631 unrelated NSCL/P patients and 689 controls, were used to test association of the best-scored eQTLs. This sample included 38 from the 46 patients used in eQTL mapping. Patients were ascertained during Operation Smile programs in Rio de Janeiro-RJ, Maceió-AL, Barbalha-CE, Fortaleza-CE and Santarém-PA, according to previous reports (Brito et al., 2012), and also in Hospital das Clínicas (Universidade de São Paulo), Hospital Menino Jesus (São Paulo-SP) and Hospital Sobrapar (Campinas-SP). All patients were clinically evaluated, to ensure that no syndromic or cleft palate only patients were included. We used controls from CEGH-CEL biobank (coordinated by M. Zatz; Instituto de Biociências, Universidade de São Paulo), and from Hospital Regional do Baixo Amazonas (Santarém-PA). DNA was extracted from saliva (using Oragene saliva kits OG-500 and OG-575; DNA Genotek Inc., Canada) or peripheral blood, following standard protocols.

eQTL and Ancestry Informative Marker Genotyping

Our case-control sample was genotyped for 35 eQTLs SNVs and 122 ancestry-informative markers (AIMs), described elsewhere (Kosoy et al., 2009). We used the Illumina GoldenGate VeraCode assay, on Illumina BeadXpress platform, following manufacturer's instructions. Allele calling was performed with GenomeStudio software (Illumina). Quality control procedures removed variants with call rate <90%, MAF < 0.05, or departure from HWE in controls ($P < 0.01$). Individuals genotyped for <90% of SNPs, or that showed gender inconsistency were also removed. After quality control steps, 624 patients and 668 controls proceeded to association analysis, with 35 eQTLs and 119 AIMs.

Structured Association Analysis of OOMMSC eQTLs

We used a structured association approach, in order to account for Brazilian population stratification. Structure 2.3.4 software was run to infer ancestry components of each individual (Falush et al., 2007), with 100,000 burn-in steps and 100,000 Markov chain Monte Carlo repetitions. *Admixture model*, *allele frequencies correlated*, and *USEPOPINFO=1* were used as parameters. The run assumed $K=3$ parental populations, based on the tri-hybrid origin of Brazilian population (Salzano and Sans, 2014). As ancestry controls, we incorporated genotypes from European, African and Amerindian populations, obtained from HapMap (57 Utah residents with Northern and Western European ancestry [CEU] and 62 Yoruban from Ibadan-Nigeria [YRI]) and from Kosoy et al. (2009; 60 CEU, 128 New York City residents with European ancestry [NYCPEA], 56 YRI, 19 Bini West African from Niger-Congo, 23 Kanuri West African from Nigeria, 50 Mayan Amerindians from Guatemala, 26 Quechuan Amerindians from Peru, 29 Nahuan Amerindians from Mexico, and a collection of 28 Amerindians of different peoples from North Brazil). Following ancestry inference by Structure, STRAT software was used to test association for each SNP, conditioning on individual ancestry proportions, using 5,000 simulated tests per locus (Pritchard et al., 2000). Significant associations were considered if $P < 0.0014$ (Bonferroni adjustment for 35 independent test).

Sanger Sequencing of MRPL53

MRPL53 coding and intronic regions of 203 NSCL/P individuals, from the case-control sample, were amplified, with a single pair of PCR primers (forward: 5' GAAGCGCAGCCATTAC 3'; reverse: 5' AGGGTTAAGCAGTGAGTGAT 3'; annealing temperature: 54°C). PCR products were submitted to Sanger sequencing, and capillary electrophoresis was performed with ABI3730 DNA Analyzer (Applied Biosystems). Sequences were visualized in Sequencher5.2 analysis software (Gene Codes).

Results

OOMMSC cis-eQTLs

Using genotype and gene expression microarray data from OOMMSC of 46 individuals, we obtained 25,506 transcripts and 237,523 SNPs, for eQTL mapping. Cis-association tests (SRC) were performed between transcripts and SNPs, resulting in 3,955,574 nominal associations. A subset of 38 patients was genotyped for an AIM panel, revealing, on average, 65%, 28% and 7% of European, African and Amerindian contributions, respectively.

Due to our limited sample size to map eQTLs, we firstly explored the abundance of tests with low P values in the observed data. Moreover, we performed the same *cis*-association tests (SRC) between transcripts and genotypes, but using a single permuted dataset of expression values, and compared the P value distributions of permuted with observed nominal associations (Figure 1). We found an enrichment of low P values in the observed data, especially at the tail of the distribution, either by comparing observed over permuted (Figure 1A) or observed over expected (random uniform distribution of 3,955,574 values ranging from 0 to 1, Figure 1B).

We then assessed the statistical significance of nominal associations through permutations, estimating P value cutoffs from permutation thresholds for each transcript: 119 *cis*-eQTLs remained significant under the most stringent permutation P threshold (0.0001, $P < 1.6 \times 10^{-6}$, 14% FDR); 205, 614, and 2,201 *cis*-eQTLs were significant under permutation P thresholds 0.001 ($P < 1.1 \times 10^{-5}$), 0.01 ($P < 8 \times 10^{-4}$), and 0.05 ($P < 1 \times 10^{-3}$), respectively, but with high FDRs (>40% FDR, Table 1). Most of significant OOMMSC *cis*-eQTLs (permutation threshold 0.0001) are located close to TSS

(median distance<500KB; Supplementary Figure 1). There are multiple significant *cis*-eQTLs for the same transcript, and, in total, we observed 18 transcripts with at least one significant *cis*-eQTL in permutation threshold 0.0001 (Table 1). These multiple *cis*-eQTLs per transcript do not seem to be independent, as LD estimates between most significant eQTL-SNPs and most distant eQTL-SNPs (relative to TSS) evidenced average $r^2=0.893$ and average $D'=0.945$ (data not shown).

Table 1 – Significant OOMMSC *cis*-eQTLs under different permutation thresholds.

	Nominal Associations	Permutation <i>P</i> Thresholds			
		0.0001	0.001	0.01	0.05
Transcripts	25,506	18	60	294	1,212
SNPs	237,523	116	201	597	2,047
<i>cis</i> -eQTLs	3,955,574	119	205	614	2,201
Maximum <i>P</i> value cutoff	1	1.6×10^{-6}	1.1×10^{-5}	8×10^{-4}	1×10^{-3}
FDR based on PT	-	0.14	0.43	0.87	1.05
<i>cis</i> -eQTLs for NSCL/P genes	51,030	0	0	2	37

Figure 4 - QQ plots comparing $-\log_{10}(P)$ values distributions from observed over permuted nominal associations (A), as well as observed over expected (B). Permuted distribution corresponds to 3,955,574 *P* values from *cis*-associations using a single permuted dataset of expression values. Expected is the random uniform distribution of 3,955,574 values ranging from 0 to 1.

Using cross-reference databases, we did not observe any functional annotation term significantly enriched among genes for which we identified eQTLs under permutation thresholds 0.001, 0.001, and 0.01. However, 875 genes with identifiable HUGO gene symbols from the 1,212-transcript list (permutation threshold 0.05) evidenced enrichment for several terms from *GO Database* (under the ontologies “Biological Processes” and “Molecular Function”) and *MYC Target Gene Database* (Supplementary Tables 3, 4 and 5). A notable finding was that MYC transcription factor binds to the promoter of 74 OOMMSC eQTL-genes (“DANG BOUND BY MYC” from Chemical and Genetic Perturbations Gene Set, FDR q-value $<5.27 \times 10^{-17}$).

Replication and evidence of tissue-independency of OOMMSC cis-eQTLs

We performed an *in silico* replication study of OOMMSC *cis*-eQTLs in seven independent datasets of different human tissues provided by the Genotype-Tissue Expression (GTEx) project. We observed a high replication rate for the best *cis*-eQTLs per transcript identified at permutation threshold 0.0001 in all seven tissues, and, on average, 77% of the SNP-transcript associations equally tested in OOMMSC and GTEx tissues shared the direction of allelic effects ($P < 0.05$ cutoff; Supplementary Table 6). Although π_1 could not be estimated in all validation datasets (due to non-uniform distribution of $N=11$ or 12 P values), we observed that all best OOMMSC *cis*-eQTLs were shared by at least two tissues and ~60-85% of best *cis*-eQTLs were shared by all tissues (Supplementary Table 6). For best *cis*-eQTLs under permutation threshold 0.05, the replication rate was remarkable lower, with π_1 estimates ranging from 0.02 (*esophagus muscularis*) to 0.37 (*esophagus mucosa*) (Supplementary Table 6; Supplementary Figure 3). Taking the direction of allelic effect into account and $P < 0.05$ cutoff, we observed proportions of 13 to 19% of validated best *cis*-eQTLs. Tissue-independency was also observed at this permutation threshold (0.05), where ~82-90% of best *cis*-eQTLs were shared by at least two tissues.

Evidence of association between NSCL/P and MRPL53 cis-eQTLs

None of our eQTLs was located near GWAS-associated loci. Therefore, we selected 35 OOMMSC *cis*-eQTLs to test for association with NSCL/P based on two criteria: (1) best eQTLs significant at permutation threshold 0.01 near NSCL/P candidate *loci* (N=13) and (2) best eQTLs significant at permutation threshold 0.05, implicated with NSCL/P candidate genes (N=22; Supplementary Table 7). Cases and controls were highly admixed, as indicated by Structure analysis. On average, proportion of European, African and Amerindian components were, respectively, 58%, 21%, and 21% for patients, and 69%, 17% and 14% for controls (Supplementary Fig 2).

We observed a significant association for rs1063588 (Table 2; $P=0.0008$; odds ratio [OR]= 1.26 [1.00-1.59 95%CI] for heterozygous and 1.23 [0.86-1.77 95%CI] for homozygous genotypes), an eQTL for *MRPL53* (mitochondrial ribosomal protein L53, located at 2p13 region). Supporting this result, another eQTL for *MRPL53*, rs6546909, was marginally associated ($P=0.0036$), as they probably belong to the same haplotype (LD estimates between rs1063588 and rs6546909 are $r^2 = 0.795$, $D' = 0.952$). The genotype distributions for these SNPs in cases and controls are compatible with a dominant ($OR_{dom_rs1063588} = 1.25$ [1.01-1.56 95%CI]; $OR_{dom_rs6546909} = 1.31$ [1.05-1.63 95%CI]) but not recessive, additive or multiplicative mode of inheritance (Table 2). Rs1063588 is a missense variant (NM 001146158.1: c.397G>A; NP 001139630.1: p.Asp133Asn) located at exon 4 of *MOGS* gene, ~8.7KB upstream from *MRPL53* TSS, while rs6546909 is a synonymous variant (NM 133637.2: c.1842T>A) located at exon 11 of *DQX1* gene, ~47KB downstream from *MRPL53* TSS (Figure 2).

Table 2 – *cis*-eQTLs for *MRPL53* in the associated region 2p13.

eQTL	MAF	Genotypes	P value	OR het (95%CI)	OR hom (95%CI)	OR dom (95%CI)	OR rec (95% CI)
rs1063588 (A/G)	P: 0.351 (A) C: 0.318 (A)	P: 74/289/260 C: 73/279/316	0.0008	1.26 (1.00-1.59)	1.23 (0.86-1.77)	1.25 (1.01-1.56)	1.10 (0.78-1.55)
rs6546909 (A/T)	P: 0.354 (T) C: 0.311 (T)	P: 258/289/76 C: 319/277/68	0.0036	1.29 (1.02-1.63)	1.38 (0.96-1.99)	1.31 (1.05-1.63)	1.21 (0.86-1.72)

MAF: minor allele frequency; **P:** patients; **C:** controls; **OR:** odds ratio; **CI:** confidence interval. Odds ratios were calculated for heterozygous (het), homozygous (hom), dominant (dom) and recessive (rec) models.

Figure 2 - OOMMSC cis-eQTLs for *MRPL53* gene. SNPs rs1063588 **(A)** and rs6546909 **(B)** are eQTLs for *MRPL53* ($P < 0.001$, 14% FDR) and low expression alleles are associated with NSCL/P. **(C)** Genomic location (2p13.1, chr2:74,455,000-74,520,000 – GRCh37/hg19 Assembly) of these SNPs (bottom red arrows) relative to *MRPL53* gene (red circle), and to *WBP1* gene (blue circle).

Mutation screening of *MRPL53*

We selected 203 NSCL/P individuals for resequencing of *MRPL53* coding region and introns. We identified the common variant rs1047911 (MAF=0.38), and the rare variants rs141704877 (MAF=0.005, 2 patients), rs78834087 (MAF=0.012, 5 patients), rs148007344 (MAF=0.005, 2 patients) and rs139817903 (MAF=0.005, 2 patients). All variants are reported in dbSNP, and also present in our in-house exome database in similar frequencies (Table 3).

Table 3 –*MRPL53* variants identified in 201 NSCL/P individuals

Variant	Location*	Exon	Substitution; type	Amino acid change	Genotypes (NSCL/P) / MAF	Genotypes (controls) / MAF
rs1047911	74699778	1	c.10G>T; missense	p.(A4S)	74GG-99GT-27TT MAF = 0.38	298GG-242GT-64TT MAF = 0.31
rs141704877	74699715	1	c.73A>G; missense	p.(K25E)	201AA-2AG-0GG MAF = 0.005	593AA-11AG: 0GG MAF = 0.009
rs78834087	74699595	2	c.95C>A; missense	p.(T32N)	198CC-5CA-0AA MAF = 0.012	583CC-21CA-0 AAMAF = 0.017
rs148007344	74699584	2	c.106A>G; missense	p.(T36A)	201AA-2AG-0GG MAF = 0.005	597AA-7AG-0GG MAF = 0.006
rs139817903	74699510	2	c.180C>G; synonymous	p.(S60S)	201CC-2CG-0GG MAF = 0.005	599CC-5CG-0GG MAF = 0.004

*hg19, chromosome 2.

Discussion

A common challenge in GWAS has been establishing a functional link between non-coding variants and regulatory effects. Here, we explore this issue through an approach that connects genetic variants, gene expression levels and phenotype. We mapped 119 *cis*-eQTLs for 18 genes in OOMMSC, under the most stringent permutation *P* threshold (0.0001, $P < 1.6 \times 10^{-6}$, 14% FDR). Although we did not observe, for these genes, enrichment of any functional annotation term, when we adopted a less stringent permutation threshold (0.05), we were able to observe that our transcript list is enriched with MYC-interacting genes, as MYC transcription factor binds to the promoter of 74 OOMMSC eQTL-genes. *MYC* has been implicated with the strongest association signals in GWAS on NSCL/P (Birnbaum et al., 2009), strengthening our approach.

From the 119 *cis*-eQTL identified, we selected 35 for association analysis, and identified a *MRPL53* regulatory region as a susceptibility locus for NSCL/P. *MRPL53*, mapped at *2p13*, codifies a component of the large subunit of the mitochondrial ribosome. A proper functioning of mitochondrial translation machinery is required for mitochondrial oxidative phosphorylation system. It has been shown that mitochondrial activity is essential for crucial cellular functions during development, such as cell growth, differentiation and migration (Sylvester et al., 2004). However, functional

annotations for this gene are scarce; for other members of the large MRPL gene family, genetic variants are usually associated with sensory disorders (Sylvester et al., 2004).

The chromosomal region 2p13 comprehends the putative susceptibility locus OFC2, firstly reported by candidate gene association analysis (Ardinger et al., 1989). Although other studies, in the pre-GWAS era, implicated *TGFA* as the probable gene involved in this region, conflicting results were often produced (Passos-Bueno et al., 2004; Carinci et al., 2007). Recently, a meta-analysis of GWAS data, including mostly samples of European origin, and which represents the largest genetic study on NSCL/P to date, did not find any evidence of association at 2p13, even considering marginal associations ($P < 10^{-4}$; Ludwig et al., 2012). Furthermore, a dense SNP imputation in a 1Mb region surrounding *MRPL53* in the meta-analysis sample did not exposed any suggestive association (Ludwig, personal communication). Therefore, we ruled out differences in LD patterns as a possible reason behind this lack of association in the European population. Nevertheless, the locus here reported may represent a population-specific risk factor. In this regard, a recent GWAS in Chinese population revealed association of locus 16p13.3, which had never been reported by the previous GWAS with European populations (Sun et al., 2015).

In an attempt to find supporting evidence for the role of *MRPL53* in NSCL/P etiology, we sequenced this gene in 203 patients, but the variants found were also present in our controls with similar allele frequencies. Expression of *MRPL53* during craniofacial development has been quite unexplored. In a 1-Mb window around *MOGS* gene (which harbors the eQTL rs1063588 in humans), there is evidence of high expression of the proximal gene *WBP1* in frontonasal region and maxilla, during embryogenesis of humans and mice (according to SysFACE database). *WBP1* is located ~3kb downstream of rs1063588 (Figure 2), and encodes a WW domain-binding protein (McDonald et al., 2011). Although rs1063588 was identified here as an eQTL for *MRPL53* in *orbicularis oris* muscle, it is possible that this association reflects a common regulatory region for other proximal genes in different tissues. In this scenario, *WBP1* emerges as a second candidate.

In summary, we characterize 119 variants as eQTLs in OOMMSC, and report the association between a *MRLP53* eQTL, at chromosome 2p13, and NSCL/P. Nevertheless, further studies are necessary to explore *MRPL53* function during development and its role in NSCLP susceptibility.

References

- Albert FW, Kruglyak L. 2015. The role of regulatory variation in complex traits and disease. *Nat Rev Genet* 16:197-212.
- Ardinger HH, Buetow KH, Bell GI, Bardach J, VanDemark DR, Murray JC. 1989. Association of genetic variation of the transforming growth factor-alpha gene with cleft lip and palate. *Am J Hum Genet* 45:348-353.
- Birnbaum S, Ludwig KU, Reutter H, Herms S, Steffens M, Rubini M, Baluardo C, Ferrian M, Almeida de Assis N, Alblas MA *et al.* 2009. Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24. *Nat Genet* 41:473-477.
- Brito LA, Paranaíba LM, Bassi CF, Masotti C, Malcher C, Schlesinger D, Rocha KM, Cruz LA, Barbara LK, Alonso N *et al.* 2012. Region 8q24 is a susceptibility locus for nonsyndromic oral clefting in Brazil. *Birth Defects Res A Clin Mol Teratol* 94:464-468.
- Bryois J, Buil A, Evans DM, Kemp JP, Montgomery SB, Conrad DF, Ho KM, Ring S, Hurles M, Deloukas P *et al.* 2014. Cis and trans effects of human genomic variants on gene expression. *PLoS Genet* 10:e1004461.
- Bueno DF, Kerkis I, Costa AM, Martins MT, Kobayashi GS, Zucconi E, Fanganiello RD, Salles FT, Almeida AB, do Amaral CE *et al.* 2009. New source of muscle-derived stem cells with potential for alveolar bone reconstruction in cleft lip and/or palate patients. *Tissue Eng Part A* 15:427-435.
- Carinci F, Scapoli L, Palmieri A, Zollino I, Pezzetti F. 2007. Human genetic factors in nonsyndromic cleft lip and palate: an update. *Int J Pediatr Otorhinolaryngol* 71:1509-1519.
- Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. 2009. Mapping complex disease traits with global gene expression. *Nat Rev Genet* 10:184-194.
- Dermitzakis ET. 2012. Cellular genomics for complex traits. *Nat Rev Genet* 13:215-220.
- Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, Ingle C, Beazley C, Gutierrez Arcelus M, Sekowska M *et al.* 2009. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* 325:1246-1250.
- Dixon MJ, Marazita ML, Beaty TH, Murray JC. 2011. Cleft lip and palate: understanding genetic and environmental influences. *Nat Rev Genet* 12:167-178.
- Falush D, Stephens M, Pritchard JK. 2007. Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol Ecol Notes* 7:574-578.
- Gritli-Linde A. 2008. The etiopathogenesis of cleft lip and cleft palate: usefulness and caveats of mouse models. *Curr Top Dev Biol* 84:37-138.
- Grundberg E, Small KS, Hedman AK, Nica AC, Buil A, Keildson S, Bell JT, Yang TP, Meduri E, Barrett A *et al.* 2012. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat Genet* 44:1084-1089.
- Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, Buil A, Ongen H, Yurovsky A, Bryois J, Giger T, Romano L, Planchon A *et al.* 2013. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife* 2:e00523.
- Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* 106:9362-9367.
- Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4:249-264.
- Johnson WE, Li C, Rabinovic A. 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8:118-127.
- Jugessur A, Shi M, Gjessing HK, Lie RT, Wilcox AJ, Weinberg CR, Christensen K, Boyles AL, Daack-Hirsch S, Trung TN *et al.* 2009. Genetic determinants of facial clefting: analysis of 357 candidate genes using two national cleft studies from Scandinavia. *PLoS One* 4:e5385.

- Kosoy R, Nassir R, Tian C, White PA, Butler LM, Silva G, Kittles R, Alarcon-Riquelme ME, Gregersen PK, Belmont JW *et al.* 2009. Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum Mutat* 30:69-78.
- Ludwig KU, Mangold E, Herms S, Nowak S, Reutter H, Paul A, Becker J, Herberz R, AlChawa T, Nasser E *et al.* 2012. Genome-wide meta-analyses of nonsyndromic cleft lip with or without cleft palate identify six new risk loci. *Nat Genet* 44:968-971.
- McDonald CB, McIntosh SK, Mikles DC, Bhat V, Deegan BJ, Seldeen KL, Saeed AM, Buffa L, Sudol M, Nawaz Z *et al.* 2011. Biophysical analysis of binding of WW domains of the YAP2 transcriptional regulator to PPXY motifs within WBP1 and WBP2 adaptors. *Biochemistry* 50:9616-9627.
- Mele M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, Young TR, Goldmann JM, Pervouchine DD, Sullivan TJ *et al.* 2015. Human genomics. The human transcriptome across tissues and individuals. *Science* 348:660-665.
- Nica AC, Montgomery SB, Dimas AS, Stranger BE, Beazley C, Barroso I, Dermitzakis ET. 2010. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet* 6:e1000895.
- Nica AC, Parts L, Glass D, Nisbet J, Barrett A, Sekowska M, Travers M, Potter S, Grundberg E, Small K *et al.* 2011. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet* 7:e1002003.
- Passos-Bueno MR, Gaspar DA, Kamiya T, Tescarollo G, Rabanea D, Richieri-Costa A, Alonso N, Araujo B. 2004. Transforming growth factor- α and nonsyndromic cleft lip with or without palate in Brazilian patients: results of a large case-control study. *Cleft Palate Craniofac J* 41:387-391.
- Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. 2000. Association mapping in structured populations. *Am J Hum Genet* 67:170-181.
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ *et al.* 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559-575.
- Rahimov F, Marazita ML, Visel A, Cooper ME, Hitchler MJ, Rubini M, Domann FE, Govil M, Christensen K, Bille C *et al.* 2008. Disruption of an AP-2 α binding site in an IRF6 enhancer is associated with cleft lip. *Nat Genet* 40:1341-1347.
- Salzano FM, Sans M. 2014. Interethnic admixture and the evolution of Latin American populations. *Genet Mol Biol* 37:151-170.
- Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D *et al.* 2007. Population genomics of human gene expression. *Nat Genet* 39:1217-1224.
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al.* 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102:15545-15550.
- Sun Y, Huang Y, Yin A, Pan Y, Wang Y, Wang C, Du Y, Wang M, Lan F, Hu Z *et al.* 2015. Genome-wide association study identifies a new susceptibility locus for cleft lip with or without a cleft palate. *Nat Commun* 6:6414.
- Sylvester JE, Fischel-Ghodsian N, Mougey EB, O'Brien TW. 2004. Mitochondrial ribosomal proteins: candidate genes for mitochondrial disease. *Genet Med* 6:73-80.
- Uslu VV, Petretich M, Ruf S, Langenfeld K, Fonseca NA, Marioni JC, Spitz F. 2014. Long-range enhancers regulating Myc expression are required for normal facial morphogenesis. *Nat Genet* 46:753-758.
- Wilkie AO, Morriss-Kay GM. 2001. Genetics of craniofacial development and malformation. *Nat Rev Genet* 2:458-468.

Supplementary Information

Supplementary Figure 1: OOMMSC *cis*-eQTLs distances from TSS.

Supplementary Figure 2: Ancestry profile of case-control sample.

Supplementary Figure 3: Replication of best OOMMSC *cis*-eQTL (permutation threshold 0.05) in independent datasets of human tissues (GETx).

Supplementary Table 1: Clinical data of 43 NSCL/P patients and 4 controls used for eQTL mapping

Supplementary Table 2: OOMMSC immunophenotype for adhesion, endothelial, hematopoietic, and mesenchymal markers.

Supplementary Table 3: GSEA Molecular Signatures Database (MSigDB) - Gene Sets Overlap Analysis of PT500 eQTL genes 1 - GO Biological Process.

Supplementary Table 4: GSEA Molecular Signatures Database (MSigDB) - Gene Sets Overlap Analysis of PT500 eQTL genes 1 - GO Molecular Function.

Supplementary Table 5: GSEA Molecular Signatures Database (MSigDB) - Gene Sets Overlap Analysis of PT500 eQTL genes 1 – Curated.

Supplementary Table 6. Validation of OOMMSC *cis*-eQTLs (best per transcript) in independent samples (GTEx).

Supplementary Table 7. OOMMSC *cis*-eQTLs tested for association with NSCL/P.

Supplementary Figure 1. OOMMSC *cis*-eQTLs distances from TSS
(permutation threshold 0.0001, N=119). Best *cis*-eQTLs (per transcript) are highlighted as red triangles (N=18).

Supplementary Figure 2. Ancestry profile of case-control sample.

Proportions of European (blue) African (red) and Amerindian (green) contribution for each individual, represented by a single column. Among the patient sample (upper panel), 38 of 43 individuals with OOMMSC, from eQTL mapping, were also included in this analysis. Percentages of European-African-Amerindian ancestries were, respectively, 58-21-21 for all patients, 65-28-7 for OOMMSC patient subsample (upper panel), and 69-17-14 for controls (middle panel). Individuals from European, African and Amerindians populations were included as ancestry controls (lower panel).

Supplementary Figure 3: Replication of best OOMMSC *cis*-eQTL (permutation threshold 0.05) in independent datasets of human tissues (GETx). *P* values distributions of *cis* hits tested in seven tissues (in white, associations with discordant directions of allelic effect). Replication and sharing between two samples is reported as the proportion of true positives (π_1 , or “ π_1 ”))

Supplementary Table 1. Clinical data of 43 patients and 4 controls used for eQTL mapping.

Sample	Register	Age (years)	Gender	Disorder status	CLO / CLP	Familial
1	F1403	NA	Male	Patient	CLP	Isolated
2	F3292	0.33	Female	Patient	CLP	NA
3	F3324	NA	Female	Patient	CLP	NA
4	F3404	17	Female	Patient	CLP	Isolated
5	F3413	6	Female	Patient	CLP	Isolated
6	F3427	12	Male	Patient	CLO	Isolated
7	F3430	13	Male	Patient	CLP	Familial
8	F3434	9	Female	Patient	CLP	Isolated
9	F3436	5	Male	Patient	CLP	Isolated
10	F3439	1	Male	Patient	CLO	Familial
11	F3440	4	Male	Patient	CLP	Familial
12	F3458	4	Female	Patient	CLO	Isolated
13	F3459	3	Female	Patient	CLP	Isolated
14	F3462	0.75	Female	Patient	CLO	Isolated
15	F3463	8	Male	Patient	CLO	Isolated
16	F3468	1	Female	Patient	CLO	Familial
17	F3474	6	Male	Patient	CLO	NA
18	F3476	0.58	Male	Patient	CLO	NA
19	F3478	4	Male	Patient	CLO	NA
20	F3479	0.66	Male	Patient	CLP	NA
21	F3480	14	Female	Patient	CLO	NA
22	F3484	22	Male	Patient	CLO	NA
23	F3488	15	Male	Patient	CLO	NA
24	F3491	7	Female	Patient	CLP	NA
25	F3493	9	Male	Patient	CLO	NA
26	F3496	4	Male	Patient	CLP	NA
27	F3497	2	Female	Patient	NA	NA
28	F3507	17	Male	Patient	NA	NA
29	F4219	NA	Male	Patient	CLP	NA
30	F4257	24	Female	Patient	CLP	NA
31	F4317	0.58	Female	Patient	CLP	NA
32	F4437	0.25	Male	Patient	CLO	NA
33	F5610	14	Female	Patient	CLP	Familial
34	F5614	7	Female	Patient	CLO	Isolated
35	F5647	37	Male	Patient	CLP	Familial
36	F5662	0.33	Male	Patient	CLP	NA
37	F5686	0.75	Female	Patient	CLO	NA
38	F5715	NA	Male	Control	NA	NA
39	F5716	18	Male	Control	NA	NA
40	F5959	0.25	Male	Patient	CLP	NA
41	F5992	28	Female	Control	NA	NA
42	F6010	0.25	Female	Patient	CLO	Isolated
43	F6023	0.42	Male	Patient	CLO	NA
44	F6081	0.33	Male	Patient	CLP	Familial
45	F6130	0.5	Female	Patient	CLP	NA
46	F8000	NA	Female	Control	NA	NA
47	F3506*	30	Male	Patient	NA	NA

CLO: cleft lip only; **CLP:** cleft lip and palate; **NA:** information not available; Isolated: cases with no relatives bearing NSCL/P; Familial: cases with at least one relative (from 1st to 3rd degree) bearing NSCL/P.

*Excluded from eQTL due to missing expression data

Supplementary Table 2. OOMMSC immunophenotype for adhesion, endothelial, hematopoietic, and mesenchymal markers.

Sample	CD29	CD31	CD45	CD73 or CD166*	CD90	CD105
F3440	99.69%(4164)	1.1%(36)	0.77%(25)	93.14%(3121)	99.93%(2761)	98.10%(2945)
F3474	99.88%(4994)	0.42%(21)	0.68%(34)	91.86%(4593)	99.54%(4977)	99.66%(4983)
F3476	99.96%(4998)	0.32%(16)	2.62%(131)	99.02%(4951)	99.98%(4999)	99.98%(4999)
F3478	99.86%(4993)	0.48%(24)	0.20%(10)	94.52%(4726)	99.64%(4982)	97.88%(4894)
F3491	99.96%(4998)	0.44%(22)	0.32%(16)	95.43%(4771)	100%(5000)	94.82%(4741)
F3493	99.74%(3987)	0.12%(6)	0.42%(21)	96.6%(4262)	99.38%(3820)	NA
F3496	99.86%(4993)	0.64%(32)	0.88%(44)	98.84%(4942)*	99.98%(4999)	99.88%(4994)
F5614	99.94%(4997)	1.39%(31)	1.91%(41)	99.26%(3899)	99.25%(3708)	93.1%(513)
F5715	99.98%(4999)	0.6%(30)	1.58%(79)	97.48%(4874)	99.98%(4999)	99.98%(4999)
F5716	99.94%(4997)	0.04%(2)	0.94%(47)	91.30%(4565)*	91.84%(4592)	97.66%(4883)

Percentage of CD+ cells and absolute count between parentheses

Adhesion marker: CD29; endothelial marker: CD31; hematopoietic marker: CD31; mesenchymal markers: CD73, CD166, CD90 and CD105.

Supplementary Table 3. GSEA Molecular Signatures Database (MSigDB) - Gene Sets Overlap Analysis of PT500 eQTL genes 1 - GO Biological Process.

Gene Set Name (GO category)	# Genes in Gene Set (K)	Description	# Genes in Overlap (k)	k/K	p-value	FDR q-value
PROTEIN METABOLIC PROCESS	1231	Genes annotated by the GO term GO:0019538. The chemical reactions and pathways involving a specific protein, rather than of proteins in general. Includes protein modification.	69	0.0561	2.35E-15	1.94E-12
SIGNAL TRANSDUCTION	1634	Genes annotated by the GO term GO:0007165. The cascade of processes by which a signal interacts with a receptor, causing a change in the level or activity of a second messenger or other downstream target, and ultimately effecting a change in the functioning of the cell.	79	0.0483	5.90E-14	2.43E-11
CELLULAR MACROMOLECULE METABOLIC PROCESS	1131	Genes annotated by the GO term GO:0044260. The chemical reactions and pathways involving macromolecules, large molecules including proteins, nucleic acids and carbohydrates, as carried out by individual cells.	62	0.0548	1.73E-13	4.75E-11
CELLULAR PROTEIN METABOLIC PROCESS	1117	Genes annotated by the GO term GO:0044267. The chemical reactions and pathways involving a specific protein, rather than of proteins in general, occurring at the level of an individual cell. Includes protein modification.	61	0.0546	3.21E-13	6.63E-11
ESTABLISHMENT OF	870	Genes annotated by the GO term GO:0051234. The directed movement of a cell, substance or	49	0.0563	2.50E-11	4.13E-09

LOCALIZATION		cellular entity, such as a protein complex or organelle, to a specific location.					
TRANSPORT	795	Genes annotated by the GO term GO:0006810. The directed movement of substances (such as macromolecules, small molecules, ions) into, out of, within or between cells.	44	0.0553	4.52E-10	6.22E-08	
BIOPOLYMER METABOLIC PROCESS	1684	Genes annotated by the GO term GO:0043283. The chemical reactions and pathways involving biopolymers, long, repeating chains of monomers found in nature e.g. polysaccharides and proteins.	71	0.0422	5.29E-10	6.24E-08	
ORGANELLE ORGANIZATION AND BIOGENESIS	473	Genes annotated by the GO term GO:0006996. A process that is carried out at the cellular level which results in the formation, arrangement of constituent parts, or disassembly of any organelle within a cell.	32	0.0677	9.41E-10	9.70E-08	
NEGATIVE REGULATION OF BIOLOGICAL PROCESS	677	Genes annotated by the GO term GO:0048519. Any process that stops, prevents or reduces the frequency, rate or extent of a biological process. Biological processes are regulated by many means; examples include the control of gene expression, protein modification or interaction with a protein or substrate molecule.	36	0.0532	4.74E-08	4.34E-06	
CELL SURFACE RECEPTOR LINKED SIGNAL TRANSDUCTION GO 0007166	641	Genes annotated by the GO term GO:0007166. Any series of molecular signals initiated by the binding of an extracellular ligand to a receptor on the surface of the target cell.	34	0.053	1.18E-07	9.75E-06	
1 Subramanian, Tamayo, et al. (2005, PNAS 102, 15545-15550)							
2 Mootha, Lindgren, et al. (2003, Nat Genet 34, 267-273)							

Supplementary Table 4. GSEA Molecular Signatures Database (MSigDB) - Gene Sets Overlap Analysis of PT500 eQTL genes 1 - GO Molecular Function.

Gene Set Name (GO category)	# Genes in Gene Set (K)	Description	# Genes in Overlap (k)	k/K	p-value	FDR q-value
DNA BINDING	602	Genes annotated by the GO term GO:0003677. Interacting selectively with DNA (deoxyribonucleic acid).	30	0.0498	2.30E-06	8.35E-04
ENZYME REGULATOR ACTIVITY	323	Genes annotated by the GO term GO:0030234. Modulates the activity of an enzyme.	20	0.0619	4.94E-06	8.35E-04
TRANSFERASE ACTIVITY TRANSFERRING PHOSPHORUS CONTAINING GROUPS	424	Genes annotated by the GO term GO:0016772. Catalysis of the transfer of a phosphorus-containing group from one compound (donor) to another (acceptor).	23	0.0542	8.89E-06	8.35E-04
GROWTH FACTOR ACTIVITY	55	Genes annotated by the GO term GO:0008083. The function that stimulates a cell to grow or proliferate. Most growth factors have other actions besides the induction of cell growth or proliferation.	8	0.1455	9.24E-06	8.35E-04
KINASE ACTIVITY	369	Genes annotated by the GO term GO:0016301. Catalysis of the transfer of a phosphate group, usually from ATP, to a substrate molecule.	21	0.0569	1.05E-05	8.35E-04
RECEPTOR BINDING	377	Genes annotated by the GO term GO:0005102. Interacting selectively with one or more specific sites on a receptor molecule, a macromolecule that undergoes combination with a hormone, neurotransmitter, drug or intracellular messenger to initiate a change in cell function.	21	0.0557	1.45E-05	8.89E-04
NUCLEOSIDE TRIPHOSPHATASE ACTIVITY	212	Genes annotated by the GO term GO:0017111. Catalysis of the reaction: a nucleoside triphosphate + H ₂ O = nucleoside diphosphate + phosphate.	15	0.0708	1.57E-05	8.89E-04
PYROPHOSPHATASE ACTIVITY	226	Genes annotated by the GO term GO:0016462. Catalysis of the hydrolysis of a pyrophosphate bond between two phosphate groups, leaving one phosphate on each of the two fragments.	15	0.0664	3.31E-05	1.53E-03

HYDROLASE ACTIVITY ACTING ON ACID ANHYDRIDES	228	Genes annotated by the GO term GO:0016817. Catalysis of the hydrolysis of any acid anhydride.	15	0.0658	3.67E-05	1.53E-03
PROTEIN KINASE ACTIVITY	285	Genes annotated by the GO term GO:0004672. Catalysis of the phosphorylation of an amino acid residue in a protein, usually according to the reaction: a protein + ATP = a phosphoprotein + ADP.	17	0.0596	3.97E-05	1.53E-03

1 Subramanian, Tamayo, et al. (2005, PNAS 102, 15545-15550)
2 Mootha, Lindgren, et al. (2003, Nat Genet 34, 267-273)

Supplementary Table 5. GSEA Molecular Signatures Database (MSigDB) - Gene Sets Overlap Analysis of PT500 eQTL genes 1 - Curated

<u>Gene Set Name (C2 - Curated Gene Sets)</u>	<u># Genes in Gene Set (K)</u>	<u>Description</u>	<u># Genes in Overlap (k)</u>	<u>k/K</u>	<u>p-value</u>	<u>FDR q-value</u>
DANG BOUND BY MYC	1103	Genes whose promoters are bound by MYC [GeneID=4609], according to MYC Target Gene Database.	74	0.0671	1.11E-20	5.27E-17
PUJANA BRCA1 PCC NETWORK	1652	Genes constituting the BRCA1-PCC network of transcripts whose expression positively correlated (Pearson correlation coefficient, PCC >= 0.4) with that of BRCA1 [GeneID=672] across a compendium of normal tissues.	90	0.0545	5.87E-19	1.15E-15
GRAESSMANN APOPTOSIS BY DOXORUBICIN DN	1781	Genes down-regulated in ME-A cells (breast cancer) undergoing apoptosis in response to doxorubicin [PubChem=31703].	94	0.0528	7.29E-19	1.15E-15
DIAZ CHRONIC MEYLOGENOUS LEUKEMIA UP	1382	Genes up-regulated in CD34+ [GeneID=947] cells isolated from bone marrow of CML (chronic myelogenous leukemia) patients, compared to those from normal donors.	77	0.0557	7.12E-17	8.41E-14
BENPORATH MYC MAX TARGETS	775	Set 'Myc targets2': targets of c-Myc [GeneID=4609] and Max [GeneID=4149] identified by ChIP on chip in a Burkitt's lymphoma cell line; overlap set.	55	0.071	1.08E-16	1.02E-13
BLALOCK ALZHEIMERS DISEASE UP	1691	Genes up-regulated in brain from patients with Alzheimer's disease.	84	0.0497	1.96E-15	1.54E-12
PILON KLF1 TARGETS DN	1972	Genes down-regulated in erythroid progenitor cells from fetal livers of E13.5 embryos with KLF1 [GeneID=10661] knockout compared to those from the wild type embryos.	89	0.0451	6.88E-14	4.65E-11
MARTENS TRETINOIN RESPONSE DN	841	Genes down-regulated in NB4 cells (acute promyelocytic leukemia, APL) in response to tretinoin [PubChem=444795]; based on Chip-seq data.	51	0.0606	6.39E-13	3.77E-10
GRADE COLON CANCER UP	871	Up-regulated genes in colon carcinoma tumors compared to the matched normal mucosa samples.	51	0.0586	2.35E-12	1.08E-09
GOBERT OLIGODENDROCYTE DIFFERENTIATION DN	1080	Genes down-regulated during differentiation of Oli-Neu cells (oligodendroglial precursor) in response to PD174265 [PubChemID=4709].	58	0.0537	2.46E-12	1.08E-09

Curated Gene Sets (C2):

Chemical and Genetic Perturbations, Canonical Pathways, BioCarta gene sets, KEGG gene sets, Reactome gene sets.

1 Subramanian, Tamayo, et al. (2005, PNAS 102, 15545-15550) ; 2 Mootha, Lindgren, et al. (2003, Nat Genet 34, 267-273)

Supplementary Table 6. Validation of OOMMSC *cis*-eQTLs (best per transcript) in independent samples (GTEx).

GTEx Tissue Sample	N	OOMMSC Permutation <i>P</i> Threshold 0.0001					OOMMSC Permutation <i>P</i> Threshold 0.05				
		Tested /Total	<i>P</i> <0.05 Same direction	π_1^a	Shared all tissues ^b	Shared any tissue ^c	Tested /Total	<i>P</i> <0.05 Same direction	π_1	Shared all tissues ^b	Shared any tissue ^c
Fibroblasts cell line	272	11/18	10/11 (91%)	-	6/10	10/10	741/1,212	143/741 (19%)	0.15	44/143	124/143
LCLs	114	11/18	8/11 (73%)	-	6/8	8/8	699/1,212	93/699 (13%)	0.32	44/93	84/93
Adipose	298	12/18	10/12 (83%)	-	6/10	10/10	797/1,212	130/797 (16%)	0.17	44/130	108/130
Subcutaneous											
Brain Frontal Cortex	92	11/18	7/11 (64%)	-	6/7	7/7	754/1,212	102/754 (14%)	0.33	44/102	84/102
Muscle Skeletal	361	11/18	8/11 (73%)	0.54	6/8	8/8	744/1,212	115/744 (15%)	0.20	44/115	94/115
Esophagus Mucosa	241	12/18	10/12 (83%)	0.79	6/10	10/10	786/1,212	128/786 (16%)	0.37	44/128	108/128
Esophagus Muscularis	218	11/18	8/11 (73%)	0	6/8	8/8	775/1,212	118/775 (15%)	0.02	44/118	102/118

^a “-” corresponds to non-estimated π_1 : *P* values are not uniformly distributed, and $\pi_0 \leq 0$.

^b SNP-gene associations (best per transcript) with *P*<0.05 (same direction of allelic effect) in all tested GTEx tissues.

^c SNP-gene associations (best per transcript) with *P*<0.05 (same direction of allelic effect) in at least two tested GTEx tissues.

Supplementary Table 7. OOMMSC *cis*-eQTLs tested for association with NSCL/P.

SNV	Region	Position*	P value	Selection Criteria	Gene Name
rs10874274	1p31.1	75252134	0.669	eQTL PT1 near CLPgenes	CRYZ
rs2770191	1p22.1	92391478	0.961	CLP candidates	TGFBR3
rs1063588	2p13.1	74690378	0.0008	eQTL PT1 near CLPgenes	MRPL53
rs6546909	2p13.1	74746322	0.0036	eQTL PT1 near CLPgenes	MRPL53
rs6743068	2q33.1	202153920	0.747	eQTL PT1 near CLPgenes	AK055262
rs9310769	3p24.2	25092767	0.1	CLP candidates	THRB
rs4429624	3p21.2	50757396	0.0166	CLP candidates	HYAL2
rs6852065	4p16.3	1684409	0.88	CLP candidates	NELFA
rs9376026	6q23.2	134602454	0.242	CLP candidates	TCF21
rs4355658	7q11.23	73225692	0.505	eQTL PT1 near CLPgenes	WBSCR27
rs1127155	7q11.23	73246461	0.398	eQTL PT1 near CLPgenes	WBSCR27
rs6963245	7q36.2	154602676	0.193	CLP candidates	SHH
rs6977778	7q36.2	154724934	0.902	CLP candidates	SHH
rs10249191	7q36.2	154731120	0.689	CLP candidates	SHH
rs11135681	8p21.3	22703333	0.233	CLP candidates	TNFRSF10B
rs637217	10p11.21	35996972	0.669	CLP candidates	FZD8
rs10769945	11p15.5	1985127	0.397	CLP candidates	CDKN1C
rs11235120	11q14.2	87080713	0.22	CLP candidates	FZD4
rs10898681	11q14.2	87081370	0.514	CLP candidates	FZD4
rs8021741	14q24.3	76022400	0.402	CLP candidates	TGFB3
rs7166105	15q24.2	75343942	0.444	eQTL PT1 near CLPgenes	SCAMP5
rs12595945	16q12.1	51734008	0.04	CLP candidates	SALL1
rs2352950	16q24.1	86249038	0.52	CLP candidates	FOXC2
rs1682280	17p11.2	18289210	0.0888	CLP candidates	SHMT1
rs981684	17q21.2	38935812	0.307	CLP candidates	KRT24
rs4074462	17q21.31	43855228	0.61	eQTL PT1 near CLPgenes	CRHR1-IT1
rs4640231	17q21.31	43912786	0.794	eQTL PT1 near CLPgenes	CRHR1-IT1
rs17690703	17q21.31	43925297	0.0182	eQTL PT1 near CLPgenes	CRHR1-IT1
rs17574228	17q21.31	44104509	0.813	eQTL PT1 near CLPgenes	CRHR1-IT1
rs8070942	17q21.31	44208674	0.0078	eQTL PT1 near CLPgenes	CRHR1-IT1
rs2532276	17q21.31	44246624	0.699	eQTL PT1 near CLPgenes	CRHR1-IT1
rs606911	17q21.32	46908228	0.421	CLP candidates	HOXB2
rs17809734	22q11.21	18584588	0.331	CLP candidates	UFD1L
rs737857	22q11.21	19575953	0.991	CLP candidates	UFD1L
rs6629443	Xp22.11	21970068	0.595	CLP candidates	SMS

*dbSNP build 141 (hg19)

Chapter 7

General Discussion and Conclusions

Although several research groups have been engaged in the search for genetic factors underlying NSCL/P, much of its heritability is still barely understood. In this regard, the CDCV x CDRV debate in NSCL/P addresses part of this question, exploring the role of common and rare variants in NSCL/P etiology.

We successfully identified rare variants in the Epithelial-Cadherin gene leading to NSCL/P. This work consisted in one of the first publications to correlate *CDH1* variants with NSCL/P, and comprises, up to now, the largest collection of NSCL/P patients with *CDH1* mutations. A remarkable finding of this study was that a high proportion (15%) of our families with more than 2 affected individuals harbors a causal *CDH1* variant. Similar studies with different populations would be of extreme importance in order to corroborate the importance of this gene in familial cases of NSCL/P. Alternatively, a higher prevalence of *CDH1* mutations in Brazilian population might be related to founder effect of a few pathogenic alleles.

From a total of nine families with exomes sequenced, we were able to identify the causal variant in two (both harboring a mutation in *CDH1*). The success obtained for these two families was a direct consequence of the number of affected members and the availability of DNA samples. In addition, the big size of the families allowed us to sequence distantly related individuals, which dramatically reduced the variants that remained after filtering steps. That was not the situation for the majority of our families; for those, we needed to use literature data for prioritizing the best candidate variants. A potential drawback of this strategy is that we tend to prioritize only genes with known functional data, particularly those with some relation with craniofacial structures. Broadly speaking, if the analysis is based on our current knowledge, we diminish the chances of implicating a poorly studied gene, with unknown function or relation with craniofacial development. To overcome this limitation, we are putting effort in sequencing the exome of extra relatives from some of our families (e.g., F886 and F1843). Notwithstanding, we believe that the enrichment of candidate genes related to

PCP pathway, microtubules and cell adhesion are unbiased findings, since it was found after variant prioritization. Taken these results together, we suggest that genetic heterogeneity is underlying NSCL/P in at least one of our families (F886), and rare variants associated with high penetrance may explain the phenotype in six out of our nine families. In our evaluation, these results deeply encourage the application of exome sequencing in further familial cases.

In the meanwhile, we established *cdh1*-mutant zebrafish lines, in order to investigate how *cdh1* mutations lead to OFC. Among all homozygous mutants we generated, we only observed phenotype (embryo lethality) for the frameshift deletion (double knockout). We then generated compound heterozygous with the frameshift deletion and the in-frame mutations, to explore the possibility that one of our in-frame mutations could lead to a degree of protein impairment, but without causing embryo lethality. However, none of the compound heterozygous led to a phenotype, suggesting that these variants do not compromise the function of the protein. In face of this limitation, we are currently performing alternative strategies to phenotypically evaluate the absence of E-cadherin in zebrafish, to test the two-hit model for variants in *cdh1*. As an example, we plan to inject a small quantity of wild-type mRNA in double-knockout embryos. We believe that, overpassing the critical period of gastrulation, we will be able to assess *cdh1* knockout phenotypes in late embryo stages.

To explore the role of common variants in NSCL/P etiology, we choose to characterize the best-associated locus 8q24, and search for new loci through an eQTL mapping-based association analysis. The high admixture degree of Brazilian population introduces a powerful confounding factor that needs to be accounted for. For this reason, we have used a structured association approach, which takes advantage of information on individual ancestry components of cases and controls before performing the association tests. Differently from our previous reports (Brito et al., 2012a; Brito et al., 2012b) we used here an AIM panel composed of biallelic SNPs and a larger sample set.

Our association study narrowed the previously associated 8q24 locus to a 310-kb interval, composed of multiple linkage disequilibrium blocks. The most significant SNP reported here (rs987525; $p=4.8 \times 10^{-8}$; $OR_{het}=2.10$ [1.65-2.68 95%CI]; $OR_{hom}=3.23$ [2.19-4.79 95%CI]) coincided with the same found in GWAS. Our odds ratio estimates corroborate the moderate effect previously suggested (Birnbaum et al., 2009); nevertheless, we should bear in mind that they are based exclusively in allele

frequencies, without any correction for population structure, and thus subjected to stratification bias, which could lead to under or overestimated values. Our associated interval overlaps with a putative regulatory element, *hs1877*, in the previously defined *MYC* Medionasal Enhancer Region. These results may indicate that this regulatory element confers the most critical risk among the regulatory elements implicated with facial morphogenesis of this gene desert. On the other hand, we diminish the relevance of *IRF6* SNP rs642961 in susceptibility to NSCL/P in our population. Our results are consistent with recent findings that suggest that the association observed for *IRF6* may be driven by other variants (Sun et al., 2015), even though a functional role has been attributed to this variant (Rahimov et al., 2008). Finally, we also reported, for the first time in our population, the association of a marker in 20q12 region. It has been suggested that *MAFB* gene may be driving this association (Beaty et al., 2010). Accordingly, we have recently included *MAFB* gene in a targeted NGS panel and we will be able, in the near future, to analyze its coding region in ~200 NSCL/P patients.

Lastly, we applied an eQTL-based association study to seek for new NSCL/P susceptibility genes. We opted for this approach as an alternative to GWAS, since it directly investigates regulatory variants. We revealed, for the first time, the association of eQTLs for *MRPL53* gene (2p13.1). To endorse the validity of this association (and, in consequence, of this approach), we have tried to find a second evidence of an etiological role for this gene. Nevertheless, we have failed in finding association in a re-analysis of the meta-analysis study (Ludwig et al., 2012) with imputed SNPs at 2p13.1 locus (Ludwig, personal communication); furthermore, no pathogenic variant was found in the resequencing of *MRPL53* in Brazilian NSCL/P patients. Therefore, the emerging challenge is to establish a functional link between this gene and the disease.

In conclusion, the present work evidences the role of rare variants in NSCL/P etiology, suggesting *CDH1* as a major contributor for moderate-to-high effect variants. In addition, we also provide insights into the association of major susceptibility locus at 8q24, and report association of 2p13.1 locus, possibly implicated with *MRPL53* gene.

References

- Beaty TH, Murray JC, Marazita ML, Munger RG, Ruczinski I, Hetmanski JB, Liang KY, Wu T, Murray T, Fallin MD *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.
- Birnbaum S, Ludwig KU, Reutter H, Herms S, Steffens M, Rubini M, Baluardo C, Ferrian M, Almeida de Assis N, Alblas MA *et al.* 2009. Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24. *Nat Genet* 41:473-477.
- Brito LA, Bassi CF, Masotti C, Malcher C, Rocha KM, Schlesinger D, Bueno DF, Cruz LA, Barbara LK, Bertola DR *et al.* 2012a. IRF6 is a risk factor for nonsyndromic cleft lip in the Brazilian population. *Am J Med Genet A* 158A:2170-2175.
- Brito LA, Paranaíba LM, Bassi CF, Masotti C, Malcher C, Schlesinger D, Rocha KM, Cruz LA, Barbara LK, Alonso N *et al.* 2012b. Region 8q24 is a susceptibility locus for nonsyndromic oral clefting in Brazil. *Birth Defects Res A Clin Mol Teratol* 94:464-468.
- Ludwig KU, Mangold E, Herms S, Nowak S, Reutter H, Paul A, Becker J, Herberz R, AlChawa T, Nasser E *et al.* 2012. Genome-wide meta-analyses of nonsyndromic cleft lip with or without cleft palate identify six new risk loci. *Nat Genet* 44:968-971.
- Rahimov F, Marazita ML, Visel A, Cooper ME, Hitchler MJ, Rubini M, Domann FE, Govil M, Christensen K, Bille C *et al.* 2008. Disruption of an AP-2alpha binding site in an IRF6 enhancer is associated with cleft lip. *Nat Genet* 40:1341-1347.
- Sun Y, Huang Y, Yin A, Pan Y, Wang Y, Wang C, Du Y, Wang M, Lan F, Hu Z *et al.* 2015. Genome-wide association study identifies a new susceptibility locus for cleft lip with or without a cleft palate. *Nat Commun* 6:6414.

Chapter 8

Abstract

Orofacial clefts (or cleft lip / palate) are congenital malformations with high prevalence in population (~1:700 births). Among the orofacial cleft types, an etiologically distinct group is composed by cleft lip with or without cleft palate, which, in 70% of cases, is not accompanied by other malformations (nonsyndromic cleft lip with or without cleft palate, NSCL/P). NSCL/P presents complex etiology, often with multifactorial inheritance. Although important, the genetic contribution to NSCL/P is still poorly comprehended, and the susceptibility loci that have been associated with NSCL/P do not explain the totality of the disease's heritability. In light of this, our aim was to investigate risk variants for NSCL/P by means of different strategies. With exome sequencing for NSCL/P familial cases, we report that the epithelial cadherin-encoding gene contributes with rare, moderate-to-high risk variants to NSCL/P etiology. In addition, we suggest an etiological contribution of genes laying in planar cell polarity pathway, or involved with epithelial-mesenchymal transition, cell adhesion, cell cycle regulation, and interaction with microtubules. Using structured association approach, we narrowed the associated interval of 8q24 region in a Brazilian population, and also validated the association for 20q12. Finally, by combining association analysis with eQTL mapping, we found association of regulatory variants of *MRPL53*, in 2p13, with NSCL/P. In conclusion, this study contributes with a deeper comprehension of the etiological role of rare and common variants for NSCL/P.

Resumo

As fissuras orofaciais, ou fissuras labiopalatinas, são malformações prevalentes na população mundial, presente em cerca de um a cada 700 nascimentos. Dentro das fissuras orofaciais, um grupo etiológicamente distinto é composto pelas fissuras de lábio com ou sem fissura de palato, que, em 70% dos casos, não estão associadas a nenhuma comorbidade (fissuras de lábio com ou sem palato não sindrômicas, FL/P NS). A etiologia das FL/P NS é complexa, e em muitos casos apresenta herança multifatorial. A contribuição genética para as FL/P NS, embora sabidamente relevante, ainda é pouco conhecida. Ainda, os loci de suscetibilidade consistentemente associados às FL/P NS, não conferem um risco que explique a herdabilidade total da doença. O objetivo do presente trabalho foi investigar, por meio de diferentes estratégias, variantes de risco às FL/P NS. Utilizando sequenciamento de exoma em casos familiares, verificamos que o gene codificante da caderina epitelial, *CDH1*, contribui importantemente com variantes raras de efeito moderado a alto na etiologia das FL/Ps. Além disso, propusemos que também podem ter relevância etiológica genes envolvidos na via de polaridade planar celular, transição epitélio-mesênquima, adesão celular, regulação de ciclo celular ou de interação com microtúbulos. Por meio de um estudo de associação com correção para estratificação populacional, caracterizamos o intervalo de associação da região 8q24, o principal locus de suscetibilidade às FL/P, e identificamos associação significativa também para a região 20q12. Por fim, combinando o estudo de associação com mapeamento de eQTLs, encontramos pela primeira vez a associação entre marcadores na região 2p13, que regulam *MRPL53*, em FL/P NS. Em conclusão, este trabalho contribui para o melhor entendimento da relevância de variantes raras, de efeito moderado a alto, e comuns, de efeito pequeno, na etiologia das FL/P NS.

Appendix: Additional publications

I - Genetics and Management of the Patient with Orofacial Cleft. **Brito et al.** *Plast Surg Int.* 2012;2012:782821. doi: 10.1155/2012/782821. Epub 2012 Nov 1.

Hindawi Publishing Corporation
Plastic Surgery International
Volume 2012, Article ID 782821, 11 pages
doi:10.1155/2012/782821

Review Article

Genetics and Management of the Patient with Orofacial Cleft

**Luciano Abreu Brito, Joanna Goes Castro Meira,
Gerson Shigeru Kobayashi, and Maria Rita Passos-Bueno**

Human Genome Research Center, Institute of Biosciences, University of São Paulo, 05508-090 São Paulo, SP, Brazil

Correspondence should be addressed to Maria Rita Passos-Bueno, passos@ib.usp.br

Received 4 August 2012; Accepted 1 October 2012

Academic Editor: Renato Da Silva Freitas

Copyright © 2012 Luciano Abreu Brito et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Cleft lip or palate (CL/P) is a common facial defect present in 1 : 700 live births and results in substantial burden to patients. There are more than 500 CL/P syndromes described, the causes of which may be single-gene mutations, chromosomopathies, and exposure to teratogens. Part of the most prevalent syndromic CL/P has known etiology. Nonsyndromic CL/P, on the other hand, is a complex disorder, whose etiology is still poorly understood. Recent genome-wide association studies have contributed to the elucidation of the genetic causes, by raising reproducible susceptibility genetic variants; their etiopathogenic roles, however, are difficult to predict, as in the case of the chromosomal region 8q24, the most corroborated locus predisposing to nonsyndromic CL/P. Knowing the genetic causes of CL/P will directly impact the genetic counseling, by estimating precise recurrence risks, and the patient management, since the patient, followup may be partially influenced by their genetic background. This paper focuses on the genetic causes of important syndromic CL/P forms (van der Woude syndrome, 22q11 deletion syndrome, and Robin sequence-associated syndromes) and depicts the recent findings in nonsyndromic CL/P research, addressing issues in the conduct of the geneticist.

In this review paper, our group provided an overview on the current knowledge of orofacial cleft etiology. We explored nonsyndromic clefts and the major cleft syndromes: van der Woude, velocardiofacial (DiGeorge) and Robin sequence-associated syndromes, focusing on the implications on genetic counseling and patient care.

II – Polymorphisms at regions 1p22.1 (rs560426) and 8q24 (rs1530300) are risk markers for nonsyndromic cleft lip and/or palate in the Brazilian population. **Bagordakis et al.** Am J Med Genet A. 2013 May;161A(5):1177-80. doi: 10.1002/ajmg.a.35830. Epub 2013 Mar 26.

RESEARCH LETTER

AMERICAN JOURNAL OF
medical genetics

PART
A

Polymorphisms at Regions 1p22.1 (rs560426) and 8q24 (rs1530300) Are Risk Markers for Nonsyndromic Cleft Lip and/or Palate in the Brazilian Population

Elizabete Bagordakis,¹ Livia Máris Ribeiro Paranaíba,² Luciano Abreu Brito,³ Sibeles Nascimento de Aquino,¹ Ana Camila Messetti,¹ Hercílio Martelli-Junior,^{2,4} Mario Sergio Oliveira Swerts,⁴ Edgard Graner,¹ Maria Rita Passos-Bueno,³ and Ricardo D. Coletta^{1*}

¹Department of Oral Diagnosis, School of Dentistry, State University of Campinas, Piracicaba, São Paulo, Brazil
²Stomatology Clinic, Dental School, State University of Montes Claros, Montes Claros, Minas Gerais, Brazil
³Human Genome Research Center, Institute of Biosciences, University of São Paulo, São Paulo, Brazil
⁴Center for Rehabilitation of Craniofacial Anomalies, Dental School, University of José Rosário Vellano, Minas Gerais, Brazil

Manuscript Received: 12 April 2012; Manuscript Accepted: 2 December 2012

This letter describes suggestive associations for markers in 1p22.1 (also tested in the presente thesis) and in 8q24. Dr. Coletta's group is also interested in investigating risk factors for orofacial clefts, and our collaboration was important to help with genotyping and implementation of structured association approach.

III) Contribution of polymorphisms in genes associated with craniofacial development to the risk of nonsyndromic cleft lip and/or palate in the Brazilian population. **Paranaíba et al.** Med Oral Patol Oral Cir Bucal. 2013 May 1;18(3):e414-20.

Med Oral Patol Oral Cir Bucal-AHEAD OF PRINT
Genetic variants in nonsyndromic cleft lip and/or palate

Journal section: Oral Medicine and Pathology

Publication Types: Research

doi:10.4317/medoral.18357

<http://dx.doi.org/doi:10.4317/medoral.18357>

Contribution of polymorphisms in genes associated with craniofacial development to the risk of nonsyndromic cleft lip and/or palate in the Brazilian population

Lívia-Máris-Ribeiro Paranaíba¹, Sibele-Nascimento de Aquino², Andreia Bufalino², Hercílio Martelli-Júnior¹, Edgard Graner³, Luciano-Abreu Brito⁴, Maria-Rita dos Santos e Passos-Bueno⁵, Ricardo-D. Coletta³, Mário-Sérgio-Oliveira Swerts⁶

¹ Professor, Stomatology Clinic, Dental School, State University of Montes Claros, Montes Claros, Minas Gerais, Brazil

² Pos Graduate, Department of Oral Diagnosis, School of Dentistry, State University of Campinas, Piracicaba, São Paulo, Brazil

³ Professor, Department of Oral Diagnosis, School of Dentistry, State University of Campinas, Piracicaba, São Paulo, Brazil

⁴ Pos Graduate, Human Genome Research Center, Institute of Biosciences, University of São Paulo, São Paulo, Brazil

⁵ Professor, Human Genome Research Center, Institute of Biosciences, University of São Paulo, São Paulo, Brazil

⁶ Professor, Center for Rehabilitation of Craniofacial Anomalies, Dental School, University of Alfenas, Alfenas, Minas Gerais, Brazil

Please cite this article in press as: Paranaíba LMP, de Aquino SN, Bufalino A, Martelli-Júnior H, Graner E, Brito LA, Passos-Bueno MR, Coletta RD, Swerts MSO. Contribution of polymorphisms in genes associated with craniofacial development to the risk of nonsyndromic cleft lip and/or palate in the Brazilian population. Med Oral Patol Oral Cir Bucal. (2013), doi:10.4317/medoral.18357

Correspondence:

Av: Cula Mangabeira, 562,
Santo Expedito, Montes Claros,
Minas Gerais, Brazil,
liviaparanaib@gmail.com

Received: 20-03-2012
Accepted: 14-11-2012

Abstract

Background and Objective: Nonsyndromic cleft lip and/or palate (NSCL/P) is a complex disease associated with both genetic and environmental factors. One strategy for identifying of possible NSCL/P genetic causes is to evaluate polymorphic variants in genes involved in the craniofacial development.

Design: We carried out a case-control analysis of 13 single nucleotide polymorphisms in 9 genes related to craniofacial development, including *TBX1*, *PVRL1*, *MID1*, *RUNX2*, *TP63*, *TGFβ3*, *MSX1*, *MYH9* and *JAG2*, in 367 patients with NSCL/P and 413 unaffected controls from Brazil to determine their association with NSCL/P.

Results: Four out of 13 polymorphisms (rs28649236 and rs4819522 of *TBX1*, rs7940667 of *PVRL1* and rs1057744 of *JAG2*) were presented in our population. Comparisons of allele and genotype frequencies revealed that the G variant allele and the AG/GG genotypes of *TBX1* rs28649236 occurred in a frequency significantly higher in controls than in the NSCL/P group (OR: 0.41; 95% CI: 0.25-0.67; p=0.0002). The frequencies of rs4819522, rs7940667 and rs1057744 minor alleles and genotypes were similar between control and NSCL/P group, without significant differences. No significant associations among cleft types and polymorphisms were observed.

Conclusion: The study suggests for the first time evidences to an association of the G allele of *TBX1* rs28649236 polymorphism and NSCL/P.

This study provides evidence of association between *TBX1* and NSCL/P. Our collaboration in this work relies on the genotyping of AIMS, to perform the structured association analysis

IV) *MTHFR* rs2274976 polymorphism is a risk marker for nonsyndromic cleft lip with or without cleft palate in the Brazilian population. **de Aquino et al.** Birth Defects Res A Clin Mol Teratol. 2014 Jan;100(1):30-5. doi: 10.1002/bdra.23199. Epub 2013 Nov 19.

***MTHFR* rs2274976 Polymorphism Is a Risk Marker for Nonsyndromic Cleft Lip with or without Cleft Palate in the Brazilian Population**

Sibele Nascimento de Aquino¹, Ryuichi Hoshi², Elizabete Bagordakis¹, Maria Giulia Rezende Pucciarelli¹, Ana Camila Messetti¹, Helenara Moreira³, Andreia Bufalino¹, Andréa Borges², Ana Lucia Rangel⁴, Luciano Abreu Brito⁵, Mario Sergio Oliveira Swerts⁶, Hercilio Martelli-Junior⁷, Sergio R. Line⁸, Edgard Graner¹, Sílvia R.A. Reis², Maria Rita Passos-Bueno⁵, and Ricardo D. Coletta*¹

Background: Polymorphisms within the *MTHFR* (rs2274976) and *MTHFD1* (rs2236225) genes were previously associated with maternal susceptibility for having an offspring with nonsyndromic cleft lip with or without cleft palate (NSCL/P) in the Brazilian population. However, as the genotypes of the patients with NSCL/P were not evaluated, it is not clear whether the effects are associated with maternal or offspring genotypes. The aim of this study was to evaluate the association of rs2274976 and rs2236225 in the pathogenesis of NSCL/P. **Methods:** By using the TaqMan 5'-exonuclease allelic discrimination assay, the present study genotyped the rs2274976 and rs2236225 polymorphisms in 147 case-parent trios, 181 isolated samples of NSCL/P and 478 healthy controls of the Brazilian population. Transmission disequilibrium test and structured case-control analysis based on the individual ancestry proportions were performed. **Results:** The transmission

disequilibrium test showed a significant overtransmission of the rs2274976 A allele ($p = 0.004$), but no preferential parent-of-origin transmission was detected. The structured case-control analysis supported those findings, revealing that the minor A allele of rs2274976 was significantly more frequent in NSCL/P group compared with control group ($p = 0.001$), yielding an odds ratio of 3.46 (95% confidence interval, 2.05–5.85). No association of rs2236225 polymorphism with NSCL/P was observed in both transmission disequilibrium test and case-control analysis. **Conclusion:** The results of the study revealed that the presence of the rs2274976 A allele is a risk marker for the development of NSCL/P in the Brazilian population.

Birth Defects Research (Part A) 100:30–35, 2014.
© 2013 Wiley Periodicals, Inc.

This two-stage association analysis showed evidence for association of a marker in *MTHFR* with NSCL/P, using transmission disequilibrium test in trios and a case-control structured association analysis. Our contribution in this article lays on AIM genotyping, for structure association analysis.

V) Genetics and genomics in Brazil: a promising future. **Passos-Bueno et al.** Mol Genet Genomic Med. 2014 Jul;2(4):280-91. doi: 10.1002/mgg3.95.

Molecular Genetics & Genomic Medicine

Open Access

INVITED COMMENTARY

Genetics and genomics in Brazil: a promising future

Maria Rita Passos-Bueno¹, Debora Bertola^{1,2},
Dafne Dain Gandelman Horovitz³,
Victor Evangelista de Faria Ferraz⁴ & Luciano Abreu Brito¹

¹Centro de Pesquisa sobre o Genoma Humano, Departamento de Genética e Biologia Evolutiva, Instituto de Biociências Universidade de São Paulo, São Paulo, Brazil

²Instituto da Criança do Hospital das Clínicas da Faculdade de Medicina Universidade de São Paulo, São Paulo, Brazil

³Centro de Genética Médica, Instituto Nacional de Saúde da Mulher, da Criança e do Adolescente Fernandes Figueira, Fundação Oswaldo Cruz, Rio de Janeiro, Brazil

⁴Departamento de Genética, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, São Paulo, Brazil

doi: 10.1002/mgg3.95

In this invited article, the authors explore the challenges faced by Medical Genetics research in Brazil, from the perspective of different genetic centers. The authors also provide an overview, to foreign public, on Brazilian health care system and on the recently implemented National Policy for Rare Diseases.