Research Networking Programme

Final Report

Conte Federica

Supervisors: ***PhD J.H.Zhou*** *(Radboud University Nijmegen),* ***Prof. M.Rubini*** *(University of Ferrara)*

ESF/EUROCleftNet exchange visit *(ref. number 4918)*

INDEX:

1. SCIENTIFIC BACKGROUND AND AIM OF THE VISIT

Orofacial clefts (OFCs) represent the most common craniofacial malformations, comprising a heterogeneous group of structural birth defects characterized by a consistently variable level of dysmorphological severity. Although OFCs are surgically repairable with only rare exceptions, these defects lead to a wide spectrum of lifelong complications which affect the quality of life of the patients and their families and cause a relevant social and economical burden.

In general, OFCs are defined as complex multifactorial polygenic traits arising from many different etiologies.

Genetic studies of orofacial clefts including cleft lip (CL), cleft lip and/or palate (CL/P), and cleft palate only (CPO) have identified a number of causative genes. These genes, however, only explain a low percentage of OFC cases, suggesting that novel disease mechanisms are still waiting to be discovered. Current genetic studies, such as exome sequencing, have identified a limited number of CL/P genes, mainly in syndromic forms that include other malformations in addition to OFCs. For the more common sporadic CL/P, genome-wide association studies (GWAS) are often applied. GWA studies implicated variants in many genomic loci contributing to the risk of CL/P by statistical analysis, and the majority of these variants are located in the non-coding regions of the genome, where regulatory elements (REs) are present.

In some cases, the patients show a deletion or duplication of a wide genomic sequence which contains several different genes and REs. These large alterations, called *genomic copy number variations* (CNVs), have been reported to associate with syndromic and non-syndromic forms of OFCs using various genetics analyses such classical FISH, CGH arrays or more recently SNP arrays [*FitzPatrick et al., 2003*; *Mulatinho et al., 2008*; *Barber et al., 2013*; *Izzo et al., 2013*]. Most of the CNV studies are reports of single cases with large genomic variants in which the causative genes or REs are often not clear.

With the current development of genomics and diagnostic tools, CNV data of patients are accumulating and many are deposited in online databases. Therefore, a systematic analysis of all reported CNVs associated with OFCs may identify common genomic regions including genes and regulatory elements and shed lights on causative elements and common molecular pathways involved in OFCs.

The objective of our project is to perform a systematic analysis by identifying CNV regions that are common in cleft patients recorded in two online databases, DECIPHER and ECARUCA, to identify genes and p63-bound REs that potentially play a role in OFC etiology. The analysis pipeline includes the selection of OFC patients, the collection of their CNV data and the identification of the common genomic deleted or duplicated regions, shared among OFC patients. Subsequently, the common regions were prioritized based on the number of overlaps and the number of

genes in the region. The genes and p63 regulatory elements encompassed in the overlapping regions were separately investigated. For candidate genes, the evaluation of genic expression levels in developing mouse embryo was used to prioritize them further, in order to highlight the gene most involved in palate morphogenesis. Afterwards, a thorough search was carried out using different databases to collect details about the top candidates, focusing particularly on their functions and possible common pathways. Concerning p63 REs, after the sorting and selection, we mainly focused on the set up of DNA pulldown assay protocol to be able to use this proteomic approach for the p63 REs validation. In addition, an extensive statistical analysis was performed on the common CNVs regions (randomization analysis, variability of genomic sequence), and on the top candidates (OFC gene enrichment, variability of genomic locations).

2. WORK DESCRIPTION AND RESULTS

2.1. Analysis of CNV regions

2.1.1. CNV databases and patient collection

OFC patients included in our study were retrieved from two publically available databases of chromosomal aberrations and phenotypes in human: DECIPHER and ECARUCA. DECIPHER (*https://decipher.sanger.ac.uk/*) is a specific database of chromosomal imbalance and phenotype in human, based on *Ensembl Resources*. Contributing to DECIPHER is an International Consortium of more than 200 academic clinical centers of medical genetics and 1600 clinical geneticists and diagnostic laboratory scientists from thirty different Countries [*Bragin et al., 2014*]. At the time of our analysis, in this database more than 10,000 clinic cases and over 25,000 patients were recorded: of these, about 300 patients presented cleft phenotype, including CL, CLP, CPO, alveolar ridge cleft and other minor types. Similarly, ECARUCA (*European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations*, *www.ecaruca.net*) is another CNV database that collects and provides clinical and molecular information on rare unbalanced chromosome aberrations. So far, ECARUCA contains over 4800 cases with a total of more than 6600 genetic aberrations and has over 3000 account holders worldwide [*Vulto-van Silfhout et al., 2013*]. In these second database, only 17 OFC patients with CNV coordinates available were found.

Specifically, our study included 312 OFC patients (Tab. I), 295 provided by DECIPHER and 17 by ECARUCA in July 2014.

Phenotypes	Num. of patients	
Cleft lip (CL)	25	*CL patients in total 30*
CL + Alveolar ridge cleft	3	
CL + Cleft mandible	2	
Cleft lower lip	1	
Cleft palate (CPO)	186	*CPO patients in total 196*
CPO + Bifid uvula	8	
CPO + Facial cleft (unspecified)	1	
CPO + Alveolar ridge cleft	1	
Cleft lip and palate (CLP)	40	*CLP patients in total 42*
CLP + Bifid uvula	1	
CLP + Cleft mandible	1	
Bifid uvula	30	
Oral cleft (unspecified)	10	
Facial cleft (unspecified)	1	
Alveolar ridge cleft	2	

Table I. *Cohort of OFC patients (312) recorded in DECIPHER and ECARUCA databases. The classification is based on their cleft phenotypes, which refer to the data inserted by the clinicians in those databases. For enlarging the cohort, we decided to consider a mixed population including patients with different types of OFCs.*

2.1.2. Identification of overlapping genomic regions

After retrieving the CNV coordinates of those patients, deletions and duplications were divided in two separated lists. Using a combination of different BEDtools [*Quinlan and Hall, 2010*], the deletions and the duplications of OFC patients were compared to identify the overlapped genomic regions. Thus a list of overlapping genomic deletions and another of overlapping duplications were generated and sorted based on the number of overlaps per region. According to our hypothesis, the overlapping regions which are affected by a chromosomal aberration (deletion or duplication) in a high number of cleft patients may more likely contain genes or REs that contribute to OFCs development.

The overlapping regions (overlaps ≥ 2) consisted of 146 deletions and 109 duplications, showing a maximum of 8 overlaps (**Tab. II**). Specifically, one deleted region of 0.48Mb on chr 2 and two duplicated regions of 1.8 Kb and 0.5 Kb both located on chr22 were shown to be shared among eight patients' CNVs.

Num. of overlaps	Deletions		Duplications	
	Num. of regions	Length average (bp)	Num. of regions	Length average (bp)
1	198	2242419.58	198	2078953.58
2	73	1808210.63	71	1747913.48
3	37	974799.30	23	471465.30
4	20	1076381.95	6	663154.00
5	8	1186606.63	2	330141.50
6	4	514565.50	2	515747.00
7	3	2487712.67	3	283718.67
8	1	484236.00	2	1177.50

Table II. Overview of identified overlapping genomic deletions and duplications.

To evaluate if the distributions of overlaps resulting from the overlapping process of deletions and duplications were occurred or not by chance, a statistical approach based on randomization was applied.

Using specific BEDtools, 1000 randomizations were created starting from the list of patients' deletions and other 1000 randomizations from the list of patients' duplications. Afterwards, the overlapping process was repeated for each randomization, defining the overlapping regions. For each of the resulting 1000 lists of overlapping regions, the mean of the overlaps was computed. Subsequently, the overall mean and the overall standard deviation were calculated and then used to computed the z score per randomization. Next, the 1000 z score values from the randomizations were plotted to evaluate the distribution: in both the cases, the z score distribution was confirmed not to be a normal distribution. Due to this, we decided to calculate a conservative empirical p-value by comparing all the randomization scores with the z score of the real overlapping region list. The z score was calculated also for the lists of genomic deletions and duplications, resulting in two values significantly higher than all the scores deriving from randomizations. Thus, in both cases the empirical p-value was lower than 0.001 ($p<0.001$), although the real p-value would be even lower, suggesting that the overlap distributions of the overlapping deletions and duplications are unlikely to be occurred by chance.

(*Note: data details and plots will be reported in our future article; see chapter 3, "Future plans and publications"*) .

As the overlapping regions are in general large and contain both coding and non-coding sequences, we analyzed both coding and non-coding regions separately in order to search for genes and REs, respectively.

2.2. Analysis of p63 regulatory elements

Notoriously, the transcription factor p63 is important in ectodermal and epithelial development, and mutations in TP63 gene give rise to developmental defects including OFCs. Since it has been shown that p63-bound regulatory elements (p63 binding sites) can drive expression of genes relevant to orofacial development and are important to

etiology of OFCs [*Thomason et al., 2010*; *Fakhouri et al., 2013*]. The consensus sequence of p63 binding motifs is generally composed of 19 nucleotides, four of them highly conserved in p63 binding sites: C (5^{th} nt), G (8^{th} nt), C (15^{th} nt) and G (18^{th} nt) [*Kouwenhoven et al., 2010*; *Wolchinsky et al., 2014*].

To understand whether the overlapping genomic CNV regions contain p63 REs, p63 ChIP-seq datasets provided in the host lab [*Kouwenhoven et al., 2010*] were intersected with the lists of overlapping CNVs. In the host lab, p63 binding motifs that contain SNPs in any of four highly conserved positions have been identified [*K.Khandelwal, unpublished data*]. By comparing this dataset with our list of overlapping CNV regions, several shared p63 binding motifs containing SNPs in any of highly conserved positions were identified and then prioritized according to their number of overlaps. Briefly, we found 43 p63 binding motifs present in overlapping CNVs, ranging from 3 to 7 overlaps. Particularly, six motifs were shown to be shared in seven patients: two of them were present in two duplications on chr22, other three motifs were contained in the same deleted region on chr18 while the last motif was found in another deletion on chr2.

Our initial idea was to validate p63 binding and other co-regulators on these binding motifs by using DNA pulldown assay followed by mass spectrometry, a highly sensitive method to detect all the DNA-binding factors . For our experiments, oro-epithelial cells (gingival keratinocytes) were cultured and their nuclear extract was used to perform the pulldown in order to identify proteins that may be important for orofacial structure. However, before to perform the pulldown assay on the p63 binding motifs identified by CNV analysis, the optimal conditions of the protocol needed to be optimized. To set up the protocol, one known p63 motif located on chr2 was selected, as it overlaps with a SNP in linkage disequilibrium with an OFC index SNP identified in GWAS. To investigate the effect of the SNP on p63 binding, we designed three oligo pairs containing: (a) wildtype p63 motif (*WT oligo*); (b) all four conserved Cs and Gs mutated (*AM oligo*); (c) motif containing the SNP (*SNP oligo*). A specific DNA pulldown protocol was applied, in which differentially labelled samples were combined two-by-two and read simultaneously with a single mass spec run, resulting in the ratio of proteins bound to the wildtype motif (control) against those mainly bound to the mutated motif sequence. This experiment was repeated four times for setting up the conditions, but the results were not consistent. Because the optimal conditions were not yet determined, the DNA pulldown assay has not been applied to the most interesting p63 binding motifs found in the overlapping CNVs so far.

2.3. Analysis of genes

2.3.1. Filtering and prioritization of candidates

The genes contained in overlapping regions were determined by combining UCSC Genome Browser (*http://genome.ucsc.edu/*) and BEDtools. A list of RefSeq genes that are either deleted (5812) or duplicated (5941)

in CNVs was generated.

The genes of these lists were prioritized based on two criteria: (a) the number of patients sharing the same CNV and (b) the total number of genes encompassed in their genomic regions. In details, the genes deleted or duplicated in only one patient were excluded. Furthermore, genes encompassed by overlapping CNV regions containing less than five genes in total (cut-off) were prioritized over those present in regions containing six or more genes. Thus, two smaller lists of 117 deleted genes and 88 duplicated genes were generated.

We further prioritized these genes by assessing their expression levels using RNA-seq data from developing mouse palate [provided by M. Dixon, Manchester Academic Health Sciences Centre, University of Manchester; *unpublished data*]. Altogether, 43 deleted and 24 duplicated candidates were shown to be highly expressed in embryonic mouse palate. After that, these top candidates were then thoroughly investigated using several online databases, including GeneCards (*www.genecards.org/*) and Ensembl (*www.ensembl.org/index.html*) for collecting general information, MGI (*www.informatics.jax.org/*) for checking the availability of mouse models, EntrezGene (*www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene*) and UniProtKB (*www.uniprot.org/help/uniprotkb*) for examining their functions, KEGG PATHWAY (*www.genome.jp/kegg/pathway.html*), STRING (string-db.org/), APID (bioinfow.dep.usal.es/apid/index.htm), Wiki-Pi (severus.dbmi.pitt.edu/wiki-pi/) and BioGRID v3.3 (http://thebiogrid.org/) for investigating shared pathways and interactions.

To give an overview about the functional aspect, the encoded proteins covered a wide spectrum of functions. In details, the deleted candidates encode seven structural proteins, four transporters, three enzymes and proteins involved in ubiquitin-dependent degradation, eleven effectors belonging to different signaling pathways with a number of regulatory functions and ten proteins implicated in transcription, translation or DNA repair mechanisms.

Among the duplicated candidates, seven genes codifies for regulatory proteins and signaling effectors, three for metabolic enzymes and three for structural proteins, only one gene encodes for transporters, two for proteins involved in transcription or translation and other three for proteins involved in degradation processes. The specific functions of five duplicated and five deleted candidates are still uncharacterized.

(*Note: gene details will be included in our future article; see chapter 3, "Future plans and publications"*) .

2.3.2. Enrichment of cleft genes in the top lists

Because the low number of genes in our top lists, the enrichment analysis performed by using DAVID (*Database for Annotation, Visualization and Integrated Discovery*, david.abcc.ncifcrf.gov/) and HPO (*Human Phenotype Ontology*, www.human-phenotype-ontology.org/) were not successful.

Due to this, we decided to evaluate the enrichment of OFC genes with a manual approach. Firstly, an extensive literature research was done to collect a list of traditionally known cleft genes and potential cleft-related genes. In this way, a table composed of 126 known/potential cleft genes from literature was generated.

Subsequently, the cleft gene table was compared with the initial lists composing of all the genes encompassed by CNVs (5812 deleted genes; 5941 duplicated genes), and with the top lists including the prioritized candidates (43 deleted genes; 24 duplicated genes), in order to verify how many known/potential OFC genes were contained.

Concerning the deletions, 47 known/potential cleft genes were identified in the initial list of 5812 genes. Instead, 10 known/potential cleft genes were found in the top list of 43 candidates. On the other hand, in the initial list from duplication analysis 30 known/potential genes out of 5941 were identified, while in the top list 3 genes out of 24 were recognized as known/potential cleft genes.

Next, the fold enrichment was computed and the p-value was calculated using the hypergeometric test run in R statistical program. The fold enrichment was higher than 20-fold and the p-value was significant ($\alpha<0.01$) in both the cases.

(*Note: the table containing known/potential cleft genes from literature and plots will be provided in our future article; see chapter 3, "Future plans and publications"*) .

2.3.3. Variability of candidate genomic locations

In principle, if a CNV is located in a hypervariable region (e.g. MHC) normally affected by a number of different structural anomalies in healthy subjects, the CNV is unlikely to be pathogenic. To evaluate the variability of CNV genomic regions that were obtained in our analyses, we used a list of structural variants found in healthy individuals retrieved from DGV (*Database of Genomic Variants,* http://dgv.tcag.ca/). DGV is a comprehensive curated catalogue of structural variation, defined as genomic alterations that involve segments of DNA larger than 1000 bp found in the genomes of healthy individuals from worldwide populations [*MacDonald et al., 2013*].

At the time of our analysis, 202,403 structural variants from healthy population were recorded in total.

The whole human genome was divided in windows of fixed-size using a combination of different BEDtools. In each window, the structural variants found in the healthy population reported in DGV were counted and the resulting value was used to calculate the z score. In this case, the z score was used as a measure of the genomic variability: the higher was the z score, the more variable was considered the genomic sequence in the healthy population. The z score values from all windows resulted not normal.

To assess how variable the identified overlapping CNV regions are, we firstly intersected with the lists of overlapping deletions and duplications with the DGV regions, hence identifying the z scores of the windows encompassing the CNVs. Subsequently, these z scores were plotted to compared their distribution with the distribution deriving from all windows. The z scores of the windows covering the overlapping deletions varied from -2.9 to +3.0 with a broad peak located between +0.3 and +1. On the other hand, the z scores of overlapping duplications ranged from -2.3 to +3.1 with an irregular shape showing a peak located between +0.7 and +1.1.

Furthermore, we evaluated the variability of the genomic locations of the selected top genes (43 from deletion

analysis; 24 from duplication analysis). In this case, the list of windows was intersected with the genomic locations of the top candidates retrieved by Ensembl. The z scores of the windows encompassing the deleted candidates are all contained in the interval (-2.1;+2.2), while the values of the windows containing the duplicated candidates are contained in the interval (-0.63;+1.7). Two deleted and three duplicated candidates were excluded from this analysis because encompassed by windows containing telomeric regions, which were previously removed to avoid bias in score calculation.

In conclusion, we showed that the overlapping deleted or duplicated regions are mainly located in not highly variable genomic sequences. In addition, we confirmed that no candidates genes were located in the hypervariable regions of the genome (e.g. MHC-II locus, z = 3.6), excluding the possibility that they were artifacts.

(*Note: data details and plots will be reported in our future article; see chapter 3, "Future plans and publications"*) .

3. FUTURE PLANS AND PUBLICATIONS

3.1. Alternative methods for functional validation of p63 regulatory elements

To further validate the p63 REs, a different type of oligo pulldown protocol can be used, which differs in the sample reading phase. In this alternative protocol, the pulldown reaction is performed separately for each sample, without mixing differentially labelled samples. This protocol allows to evaluate separately the specific p63 bound to each type of motif sequence (wildtype or mutated) without ratios, increasing the reliability, although it requires higher amounts of materials and longer time for sample preparation and reading.

Subsequently, the lists of DNA-binding factors that bind the oligos containing p63 binding motif can be checked to evaluate their similarity and to look for possible cofactors. In addition, an antibody pulldown assay that allows to pull down p63 and its co-regulators, can be also performed.

Furthermore, the selected p63 binding motifs can be tested in vitro, through cloning approaches (transient transfection assay) and in vivo, using animal models. For example, the prioritized p63 binding sites may be cloned into Gateway system vectors for transient transfection assay in human cell lines and in transgenesis in zebrafish model.

Altogether, these experiments aim at understanding whether these p63 binding sites drive gene expression during orofacial development.

3.2. Functional validation of top candidates

Concerning the genes identified in our analysis that are not yet associated to OFCs, further studies need to be performed both in vitro and in vivo to investigate their role in orofacial development and cleft etiology.

3.3. Future publications and final words

This work has provided a lot of information regarding CNVs involved in OFC. A manuscript that is based on the data obtained in this project and presented in this report, is in preparation and can be submitted soon.

Thanks to the two *European Science Foundation/EUROCleftNet* grants awarded to me (May 1st 2014 to May 1st 2015), I had the great opportunity to complete the project proposed and to publish our results. When preparing this report, the details of some analysis have been left out for data protection and data confidentiality reasons. These details will be published and become publically available this year.

REFERENCES

Barber JC, Rosenfeld JA, Foulds N, Laird S, Bateman MS, Thomas NS, Baker S, Maloney VK, Anilkumar A, Smith WE, Banks V, Ellingwood S, Kharbutli Y, Mehta L, Eddleman KA, Marble M, Zambrano R, Crolla JA, Lamb AN**. 8p23.1 duplication syndrome; common, confirmed, and novel features in six further patients.** Am J Med Genet A. 2013 Mar;161A(3):487-500.

Bragin E, Chatzimichali EA, Wright CF, Hurles ME, Firth HV, Bevan AP, Swaminathan GJ. **DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation.** Nucleic Acids Res. 2014 Jan;42(Database issue):D993-D1000.

Fakhouri WD, Rahimov F, Attanasio C, Kouwenhoven EN, Ferreira De Lima RL, Felix TM, Nitschke L, Huver D, Barrons J, Kousa YA, Leslie E, Pennacchio LA, Van Bokhoven H, Visel A, Zhou H, Murray JC, Schutte BC. **An etiologic regulatory mutation in IRF6 with loss- and gain-of-function effects.** Hum Mol Genet. 2014 May 15;23(10):2711-20.

FitzPatrick DR, Carr IM, McLaren L, Leek JP, Wightman P, Williamson K, Gautier P, McGill N, Hayward C, Firth H, Markham AF, Fantes JA, Bonthron DT. **Identification of SATB2 as the cleft palate gene on 2q32-q33.** Hum Mol Genet. 2003 Oct 1;12(19):2491-501.

Izzo G, Freitas ÉL, Krepischi AC, Pearson PL, Vasques LR, Passos-Bueno MR, Bertola DR, Rosenberg C. **A microduplication of 5p15.33 reveals CLPTM1L as a candidate gene for cleft lip and palate.** Eur J Med Genet. 2013 Apr;56(4):222-5.

Kouwenhoven EN, van Heeringen SJ, Tena JJ, Oti M, Dutilh BE, Alonso ME, de la Calle-Mustienes E, Smeenk L, Rinne T, Parsaulian L, Bolat E, Jurgelenaite R, Huynen MA, Hoischen A, Veltman JA, Brunner HG, Roscioli T, Oates E, Wilson M, Manzanares M, Gómez-Skarmeta JL, Stunnenberg HG, Lohrum M, van Bokhoven H, Zhou H. **Genome-wide profiling of p63 DNA-binding sites identifies an element that regulates gene expression during limb development in the 7q21 SHFM1 locus.** PLoS Genet. 2010 Aug 19;6(8):e1001065.

MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. **The Database of Genomic Variants: a curated collection of structural variation in the human genome.** Nucleic Acids Res. 2014 Jan;42(Database issue):D986-92.

Mulatinho M, Llerena J, Leren TP, Rao PN, Quintero-Rivera F. **Deletion (1)(p32.2-p32.3) detected by array-CGH in a patient with developmental delay/mental retardation, dysmorphic features and low cholesterol: A new microdeletion syndrome?** Am J Med Genet A. 2008 Sep 1;146A(17):2284-90.

Quinlan AR, Hall IM. **BEDTools: a flexible suite of utilities for comparing genomic features.** Bioinformatics. 2010 Mar 15;26(6):841-2.

Thomason HA, Zhou H, Kouwenhoven EN, Dotto GP, Restivo G, Nguyen BC, Little H, Dixon MJ, van Bokhoven H, Dixon J. **Cooperation between the transcription factors p63 and IRF6 is essential to prevent cleft palate in mice.** J Clin Invest. 2010 May;120(5):1561-9.

Vulto-van Silfhout AT, van Ravenswaaij CM, Hehir-Kwa JY, Verwiel ET, Dirks R, van Vooren S, Schinzel A, de Vries BB, de Leeuw N. **An update on ECARUCA, the European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations**. *Eur J Med Genet. 2013 Sep;56(9):471-4.*

Wolchinsky Z, Shivtiel S, Kouwenhoven EN, Putin D, Sprecher E, Zhou H, Rouleau M, Aberdam D. **Angiomodulin is required for cardiogenesis of embryonic stem cells and is maintained by a feedback loop network of p63 and Activin-A.** Stem Cell Res. 2014 Jan;12(1):49-59.