Research Networking Programme

Final Report

Conte Federica

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

ESF/EUROCleftNet exchange visit (May 1st - October 31st, 2014)

INDEX:

1. *SCIENTIFIC BACKGROUND AND AIM OF THE VISIT*

Genetic studies of orofacial clefts (especially cleft lip and/or palate, CL/P, and cleft palate only, CPO) in combination with other malformations 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*, are focused mainly on the coding regions that occupy about 2% of the genome. Considering cleft lip and/or palate (CL/P), so far a limited number of CL/P genes have been identified by exome sequencing, mainly in syndromic forms. For the more common sporadic CL/P, *genome-wide association studies* (GWAS) are often applied. These 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) may be contained.

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 these studies are reports of single cases with large genomic variants in which the causative genes or REs are often not clear.

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*). A consensus binding motif that is recognized by p63 for its binding to DNA has been identified (*Kouwenhoven et al., 2010*). In general, the binding motif sequence is composed of 19 nt (**Fig.1**) and four of them are highly conserved in p63 binding sites: C (5th nt), G (8th nt), C (15th nt) and G (18th nt).

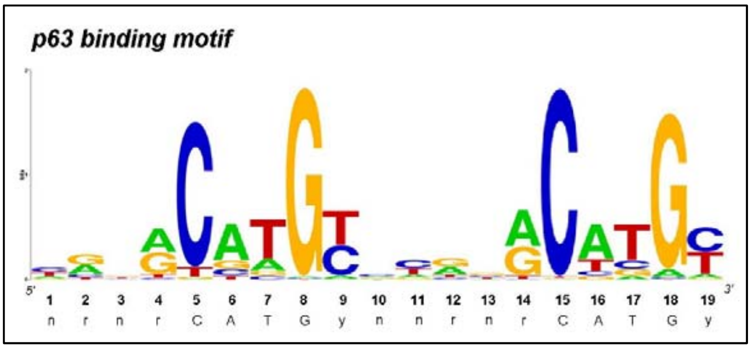

***Figure 1**: p63 consensus binding motif identified from 11,329 binding sites (Wolchinsky et al., 2014).*

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 bioinformatics analysis by comparing and intersecting CNV regions found in cleft patients recorded in two online databases, DECIPHER and ECARUCA (**Fig.2**) to identify causative genes and p63-bound REs that can play a rule in OFC etiology.

Considering this background, the main idea of our project is to set up an original bioinformatic approach to predict genes and REs that can play a rule in OFC etiology, by comparing and intersecting CNV regions found in cleft patients recorded in two online databases, DECIPHER and ECARUCA (**Fig.2**).

Concerning the pipeline, after patients' selection, the CNV data were collected, compared and overlapped by applying a new computational approach aimed to identify overlapping sequences, deleted or duplicated, shared among cleft patients. Subsequently, these regions were prioritized at different levels, considering multiple criteria to assess whether they were more likely associated with cleft phenotypes. Afterwards, the genes and p63 regulatory elements contained in the overlapping regions were separately investigated and again prioritized, evaluating diverse parameters for highlighting the most suitable candidates which could be causative for clefting. In this regard, the future part of this project would be focused on proving concretely the contribution of predicted genes and p63 regulatory elements in OFC development, through a functional validation based on different approaches both *in vitro* and *in vivo*.

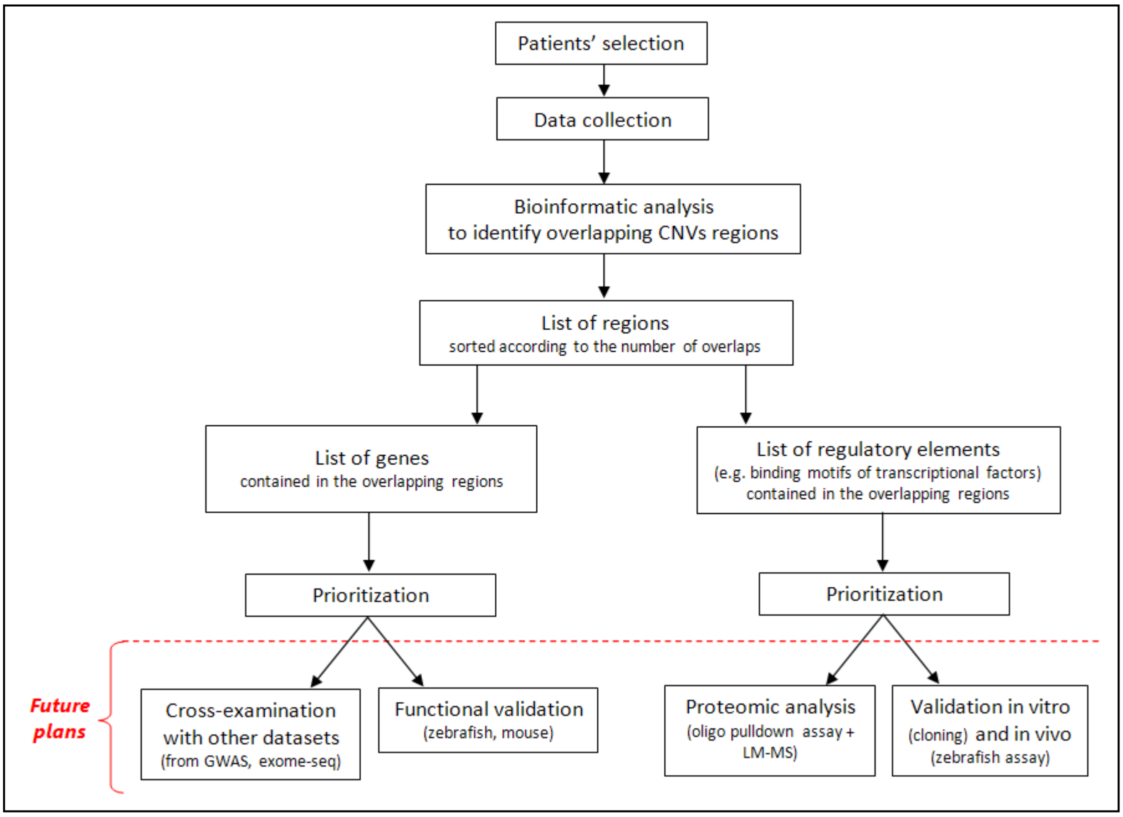
Patients' selection
Data collection
Bioinformatic analysis to identify overlapping CNVs regions
List of regions sorted according to the number of overlaps
List of genes contained in the overlapping regions
List of regulatory elements (e.g. binding motifs of transcriptional factors) contained in the overlapping regions
Prioritization
Prioritization
Future plans
Cross-examination with other datasets (from GWAS, exome-seq)
Functional validation (zebrafish, mouse)
Proteomic analysis (oligo pulldown assay + LM-MS)
Validation in vitro (cloning) and in vivo (zebrafish assay)

***Figure 2**: Project workflow.*

2. WORK DESCRIPTION AND RESULTS

2.1. Bioinformatic analysis of overlapping CNV regions

2.1.1. CNV databases used in our study

Patients included in our study were retrieved from two freely-available CNV databases: 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 genetic medicine and 1600 clinical geneticists and diagnostic laboratory scientists, belonging to 30 different Countries (*Bragin et al., 2014*). At the time of writing this report, in this database are recorded more than 10,000 clinic cases and over 25,000 patients: of these, about 300 patients present cleft phenotype, including CL, CLP, CPO, alveolar ridge cleft and other types.

Similarly, ECARUCA (*European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations*, *www.ecaruca.net*) is another CNV database that collects and provides detailed 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 we found only 17 OFC patients with CNV array data available.

2.1.2. Patient selection and data collection

The first step of our project was to select the patients according to their orofacial cleft phenotype and the availability of their CNV data. So far, our study includes 312 patients affected by OFCs (**Table 1**), 295 from DECIPHER and 17 from ECARUCA. Their genomic CNV regions were analyzed in order to identify the overlapping regions (common deleted or duplicated regions). Although the etiologic bases and developmental processes may vary depending on the cleft type, initially we decided to include patients with different OFCs (*mixed population*) in order to reach a larger cohort. To give an overview, the group of CPO patients was the greatest in number (196), followed by the group of CLP patients (42), and then the one of bifid uvula patients' (30) and CL patients' group (30), while the patients affected by other types of orofacial clefts have smaller numbers.

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 1*: 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.*

After selecting the relevant subjects, we gathered patients' details (*ID num., cleft type and other phenotypes, presence of syndromes*) and CNV information (*CNV type, genomic location, size, num. of involved genes*) from the databases. Subsequently, patients' CNVs were grouped in two lists based on the type, deletion (DEL) or duplication (DUP), in order to work with them separately.

2.1.3. Analysis of overlapping genomic regions

To indentify overlapping genomic regions in these patients' CNVs*,* initially *Galaxy* platform (http://usegalaxy.org, *Goecks et al., 2010*) has been used.

Two lists of overlapping genomic regions (one for overlapping DELs and one for overlapping DUPs) were generated, comprising of their genomic coordinates and number of overlaps.

According to our hypothesis, those overlapping regions which are affected by a chromosomal aberration (deletion or duplication) in a high number of cleft patients may contain genes or regulatory elements that more likely contribute to OFCs development. For this reason, a sorting of the overlapping regions based on the number of overlaps, was performed to highlight the most informative ones in our lists. Altogether, two genomic regions with duplications were found to be shared in eight cleft patients (highest for duplication regions) and one genomic region with deletion in other eight cleft patients (highest for deleted regions).

Subsequently, 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 (**Fig.1**).

(A) Gene selection

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

The genes of this list were prioritized based on (a) the number of patients with these CNVs and (b) the size of the genomic regions of these CNVs. Afterwards, genes in overlapping regions that contained less genes than our cut-off value (arbitrary) were separated to those in overlapping regions containing more genes, creating a *sub-lists* of about one hundred top genes (117 genes from DELs; 88 genes from DUPs).

In theory, the regions with a high number of overlaps and a low number of genes should be more informative that the regions overlapped by a few patients or containing many genes, because small regions shared in several patients may contain genes which were most likely crucial for cleft etiology during embryogenesis in those subjects.

The resulting sub-lists were prioritized using ENDEAVOUR (*http://www.esat.kuleuven.be/endeavour*), an online tool for prioritizing variants based on the prediction of candidates by analyzing inherent sequence characteristics, sequence similarity to known disease genes and functional annotation of candidate genes. To run the prioritization, this software required a *training-list* composed of reference genes. Then, the analysis was run on the basis of a three step algorithm: (1) the program collected all the data about pathways and diseases in which the genes are involved, accessing up to twenty sources (to mention a few: *GeneOntology, SwissProt, BioGrid interactions, ProspectR, BLAST score protein, etc.*); (2) for each source, the candidates were ranked according to their similarities, so a set of *n* rankings was generated (one per data source); (3) the *n* rankings were combined into one global prioritization list of candidate using the order statistics.

From the resulting prioritized lists of genes, the first 25 genes were further analyzed by examining whether these candidates were already known OFC genes and, if not, the presence of known OFC genes in the same overlapping CNVs. On the top of the gene lists, we found several known cleft-related genes. In the list from overlapping deletions, eight were known cleft genes (MEIS2, SABT2, FGF2, TSHZ1, FRZB, SPRY1, WHSC1, WHSC2). In the list from overlapping duplications, five were known cleft genes (TBX1, MAPK3, CYFIP1, DGCR6 and GNB1L). These data show that our methodology can identify OFC causative genes. Interestingly, we also identified some novel genes that are not yet known to be associated with OFCs. In the list from overlapping deletions, 17 were new candidates, and in the list from overlapping duplications, 20 were novel new candidates (*note: all gene details will be reported in our future paper in 2015; see chapter 3, "Future plans and publications"*).

To better clarify how these genes could be involved in OFC developmental process, others databases were used to collect details about their functions in human, such as DAVID (*Database for Annotation, Visualization and Integrated Discovery*, david.abcc.ncifcrf.gov/) and GeneCards (*www.genecards.org/*).

The list of new candidate genes in the deleted regions (17) includes one structural protein, two transcriptional factors, two transporters, three metabolic enzymes and four proteins related with different developmental processes which were as many as the receptors/co-receptors.

Among the new top genes (20) in the overlapping duplication regions, we found two transcriptional factors, four receptors, four proteins involved in signalling pathways, four the metabolic proteins and three transporters, as many as proteinases related to cellular degradative pathways.

Particularly, six candidates from deletions and five from duplications showed functions that could fit with OFC etiological model and embryogenesis processes (*note: all gene details will be reported in our future paper in 2015; see chapter 3, "Future plans and publications"*).

In conclusion, we prioritized candidate genes according to their known functions and expression in embryonic mouse palate. Nevertheless, they need to be further investigated for determining their contribution to OFC development (*see chapter 3, "Future plans and publications"*).

(B) Regulatory element selection

To understand whether the overlapping genomic CNV regions contain regulatory elements, especially those bound by p63, we performed analysis to search for p63-bound regulatory elements in CNV regions by using *UNIX BEDtools* (http://bedtools.readthedocs.org/) and p63 ChIP-seq datasets provided in the host lab (*E.Kouwenhoeven, K.Khandelwal*). In the host lab, p63 binding motifs that contain SNPs in any of four important positions (**Fig.1**) have been identified (*K.Khandelwal*). Intersecting her dataset with our list of overlapping regions, several shared p63 binding motifs came out. Subsequently we prioritized them according to the number of overlaps of these region in patients. Briefly, we found 43 p63 binding motifs present in overlapping CNVs with 3 overlaps or higher. Particularly, on the top of this prioritized list of binding motifs, six motifs appeared to be shared in seven patients: two of them were present in two duplications, other three motifs were contained in the same deleted region while the last motif was found in another deletion on a different chromosome (*note: regulatory elements details will be mentioned in our future paper in 2015; see chapter 3, "Future plans and publications"*).

3. FUTURE PLANS AND PUBLICATIONS

3.1. Future bioinformatic work

As statistics has not been applied to the current bioinformatics analysis, a statistical validation of our bioinformatic approach should be performed to demonstrate that the overlapping region distribution among the chromosomes is not due to the contingency. About that, we also think that could be interesting to modify the cohort of OFC patients, for example, to excluding syndromic cases and to examine how the distribution of overlaps changes and whether some overlapping regions could be specifically related with non syndromic forms.

Moreover, the lists of candidate genes can be applied to two downstream analyses, to cross-check genes identified with other techniques: (a) cross-examination between these prioritized gene lists and the candidates identified by exome-seq analysis carried out on 10 Dutch cleft families in the host lab (*C.Ockeloen, K.Khandelwal*); (b) cross-checking of overlapping regions and GWAS dataset (*K.Khandelwal*). This is to test whether our method can be used for prioritization for exome sequencing studies for OFC.

Another bioinformatic analysis that could be performed is to check whether these selected candidate genes are directly regulated by p63, using *p53FamTag* database and many ChIP-seq datasets generated in the host lab (*E.Kouwenhoeven*), because the presence of p63 targets can highlight a stronger relation between those genes and clefting.

Furthermore, we have identified miRNA genes in our high priority lists. This suggests a possible association of non-coding RNA with cleft-related processes, such as embryo morphogenesis, neural tube development, cell migration and proliferation, skeletal development. To test this hypothesis, we are planning to use other bioinformatic tools and miRNA databases for collecting data systematically and perform a prioritization process to investigate which miRNA may be cleft-related.

3.2. Functional validation of top genes

Concerning the genes identified through our computational analysis, we could evaluate their expression and tissue specificity in human by checking online databases. In addition, their contribution in clefting can also be tested in animal models such as zebrafish and mouse.

3.3. Validation of selected p63 motifs

In the context of regulatory elements, we could apply systematically the oligo pulldown assay followed by LM-MS to investigate other p63 binding motifs, such as those identified through the computational analysis in the overlapping regions. As p63-bound regulatory elements are relevant to etiology of OFCs (*Thomason et al., 2010; Fakhouri et al., 2014*), it is important to identify proteins that can cooperate with p63 during orofacial development. The *DNA pulldown followed by mass spectrometry assay* is a highly sensitive method to detect all the DNA-binding factors that recognize the same consensus sequences on the DNA. For our experiments, we would use oro-epithelial cells, to identify proteins that may be important for orofacial structure. For instance, we plan to select the p63 motifs which present SNPs in linkage disequilibrium with OFCs and, using this proteomic approach, to investigate the mutation effects on p63 bound and activity. For each selected p63 motif, we plan to design three oligo pairs containing: (a) wildtype p63 motif (WT oligo); (b) all four conserved Cs and Gs mutated (AM oligo); (c) the SNP in any of for important position (SNP oligo).

Subsequently, we would check one-by-one the lists of DNA-binding factors that bind the oligos containing p63 binding motif, to evaluate their similarity and to look for possible cofactors. In relation to this aspect, also another type of pulldown, the so called *antibody pulldown assay*, could be performed to identify the real cofactors of p63, which are still unknown.

Furthermore, the selected p63 binding motifs can be tested *in vitro*, through cloning approaches (*transient transfection assay*) and *in vivo*, using animal models. The prioritized p63 binding sites will be cloned into Gateway system vectors for transient transfection assay in human cell lines and in transgenesis in the zebrafish model (*in collaboration with J.L. Gomez-Skarmeta, CABD Sevilla*). These experiments aim at understanding whether these p63 binding sites drive gene expression during orofacial development.

3.4. Future publications

Once the statistic analysis of the distribution of overlapping regions and candidate genes is complete, we plan to publish our bioinformatics analysis of CNVs. So far, we have started applying the non-parametric statistic validation on our data, and writing the first draft of the paper. We plan to submit the manuscript early next year.

In addition, this work has provided a lot of information regarding genomic regions involved in OFC and established new techniques that can be used in subsequent experiments. It is conceivable that this work will contribute to other publications in the future.

3.5. Final words

Thanks to the second *European Science Foundation/EUROCleftNet* grant approved on October 13th awarded to me, I have the great opportunity to continue the work proposed above for the next six months. Now I am focusing on the functional validation of our selected genes and p63-bound regulatory elements in order to test their contribution in orofacial cleft development.

When preparing this report, details of some data have been have been left out for data protection and data confidentiality reasons. These data will be published and become publically available in 2015.

REFERENCES

Barber JC1, 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.

Dixon MJ, Marazita ML, Beaty TH, Murray JC. **Cleft lip and palate: understanding genetic and environmental influences.** Nat Rev Genet, 2011; 12(3):167-178.

Döcker D, Schubach M, Menzel M, Munz M, Spaich C, Biskup S, Bartholdi D. **Further delineation of the SATB2 phenotype.** Eur J Hum Genet. 2014 Aug;22(8):1034-9.

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.

Funato N, Nakamura M, Richardson JA, Srivastava D, Yanagisawa H. **Loss of Tbx1 induces bone phenotypes similar to cleidocranial dysplasia.** Hum Mol Genet. 2014 Sep;10.

Goecks J, Nekrutenko A, Taylor J; Galaxy Team. **Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.** Genome Biol. 2010;11(8):R86.

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.

Johansson S, Berland S, Gradek GA, Bongers E, de Leeuw N, Pfundt R, Fannemel M, Brendehaung A, Haukanes BI, Hovland R,Hellend G, Houge G. **Haploinsufficiency of MEIS2 is associated with orofacial clefting and learning disability.** Am J Med Genet 2014 Nov;164A:1662-26.

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.

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.

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.