UNIVERSITA' DEGLI STUDI di FERRARA - *Dep. of*

St. RADBOUD UNIVERSITY *(NIJMEGEN)* - *Dep. of Orthodontics and Craniofacial Biology (Prof. C. Carels)*
- *Dep. of Human Genetics (Dr. J.H. Zhou)*

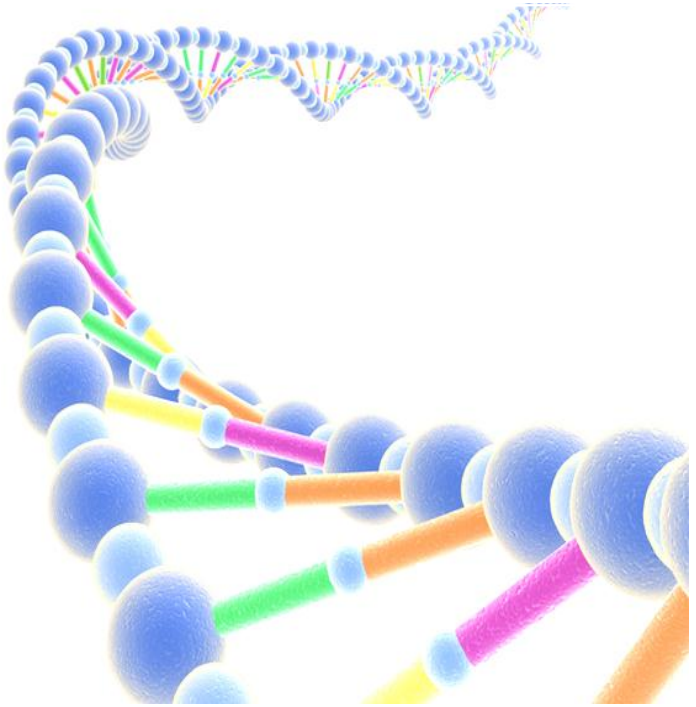
Exome sequence analysis of an oral cleft family: from variant validation to top candidate gene selection.

Laureanda: ***Conte Federica***

Relatore: *Prof.* ***Michele Rubini***

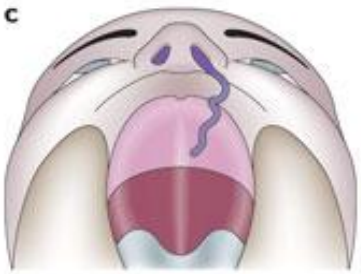
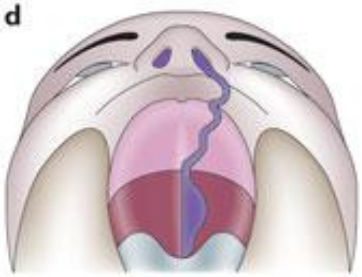
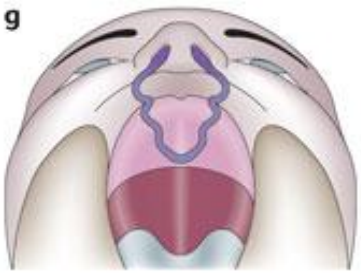
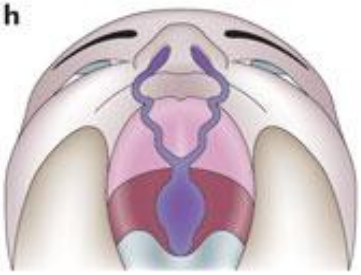
Correlatore: *Dr.* ***Elia Bonomo Roversi***

Dutch collaborators: *Prof.* ***Carine Carels,***
Dr. ***Jo H. Zhou,***
Dr. ***Kriti D. Khandelwal***.



NONSYNDROMIC CLEFT LIP AND PALATE (*NSCLP*)

- A *congenital disorder*, or *birth defect*, is a condition existing at birth and developed before parturition, during intrauterine life. Particularly, neural tube defects (such as *cleft* and *spina bifida*) arise in the first three months of embryonic development.
- *Cleft lip and palate* (CLP) is a clefting of the upper lip which continues in a clefting of the alveolar ridge and palate. The term *"nonsyndromic"* means that only this defect is present in a patient, without other malformations in different anatomical regions. The nonsyndromic forms are complex traits involving genetic heterogeneity, low penetrance and the influence of genetic and environmental factors.

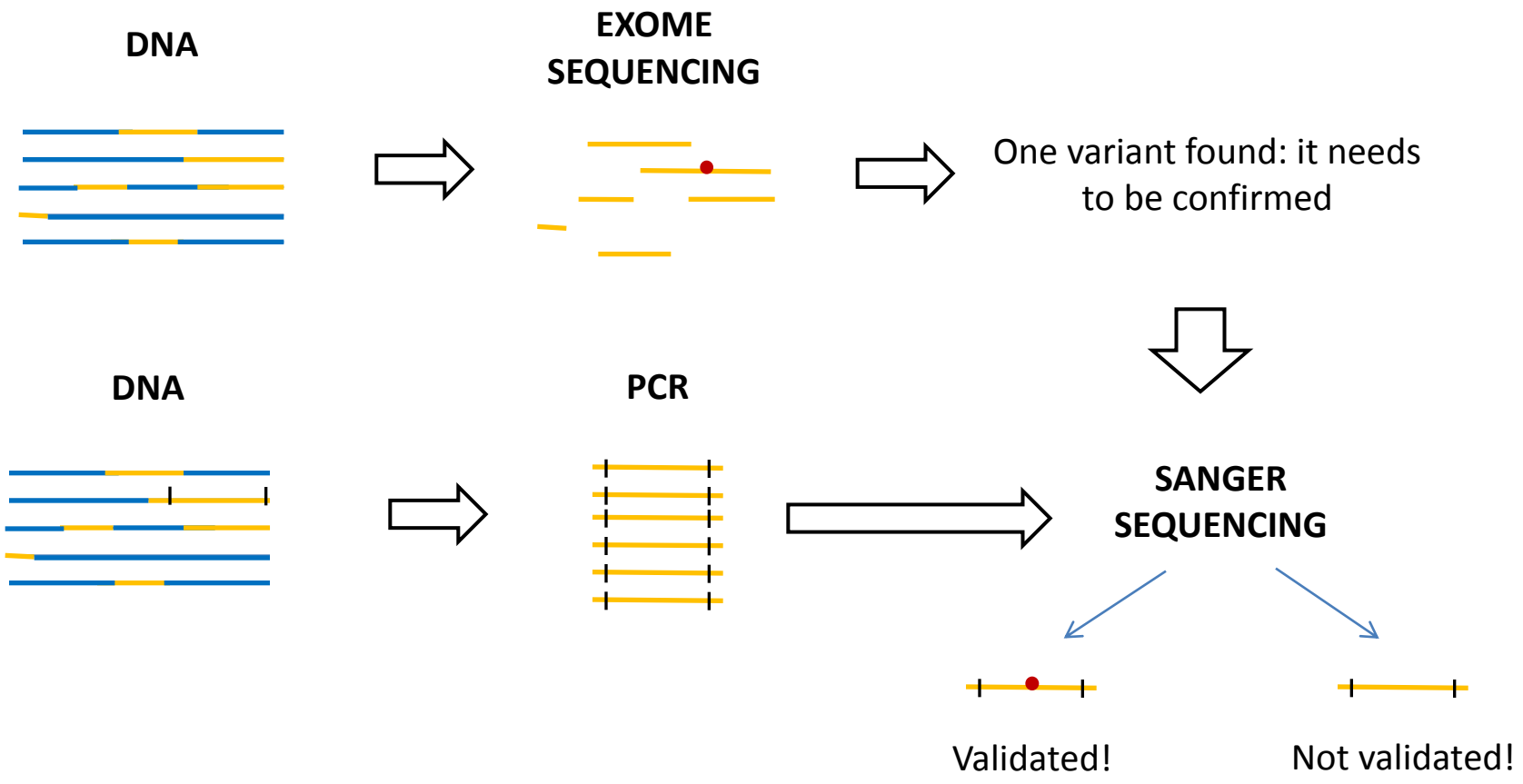
	Incomplete	*Complete*
Unilateral		
Bilateral		

Source: Dixon M.J. et al., 2011

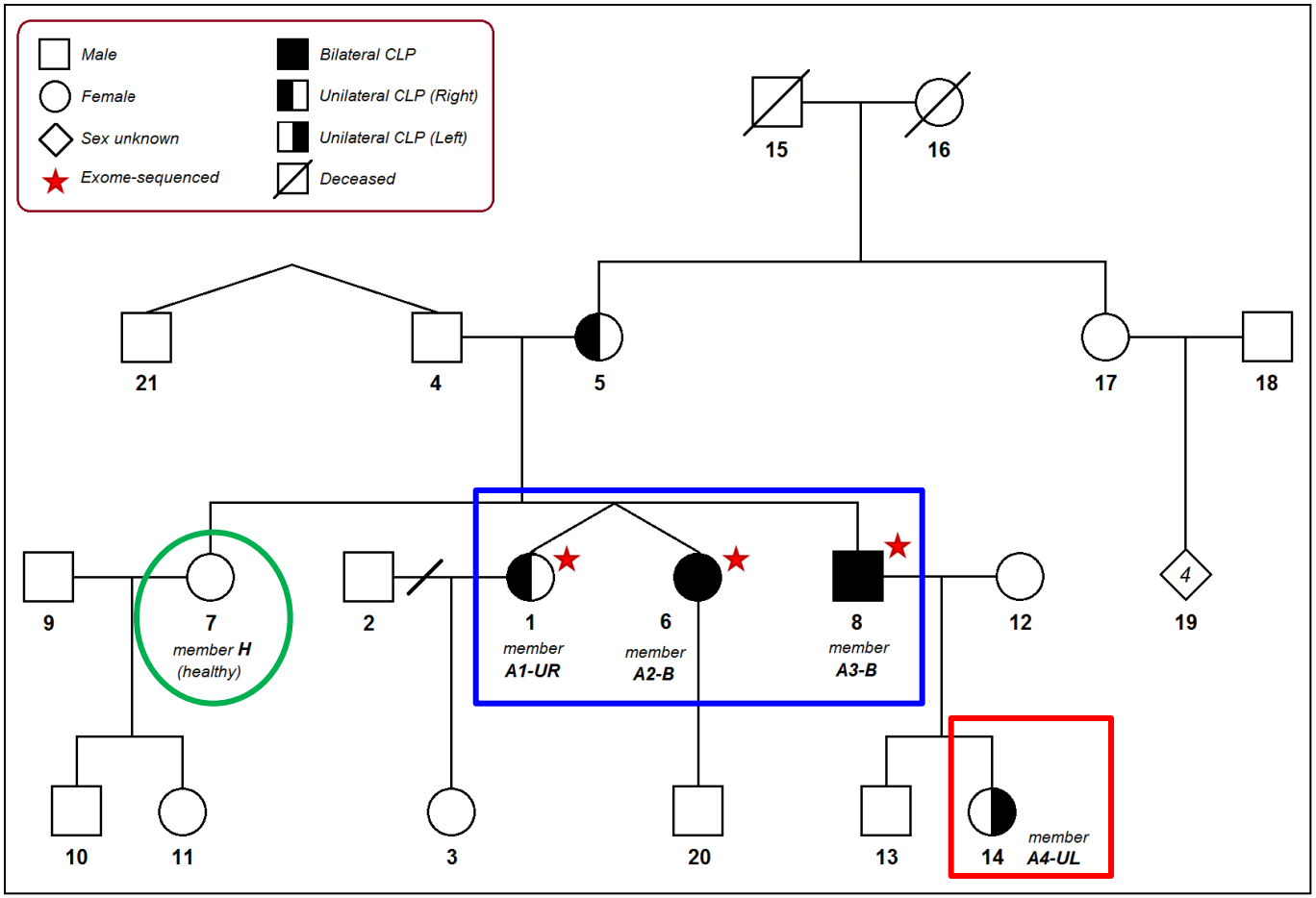
In the past 20 years, genetic studies of human diseases have evolved from low-resolution cytogenetic techniques and candidate gene-based molecular analyses to a single nucleotide resolution techniques, such as genome-wide arrays and *next generation sequencing* (whole-genome and **whole exome sequencing**) *(Khandelwal K.D. et al., 2013).*

FROM EXOME SEQUENCING TO SANGER SEQUENCING

Human complex traits can be caused by gene mutations: most of these mutations are found in the protein-coding area of genes, also known as the **exons**.
The specific method used for analyzing the exons is called **exome sequencing** (*NGS)*.
However its results need to be validated with other methods, such as **Sanger sequencing**.



I./R. FAMILY PEDIGREE



Member A1-UR (n. 1):
affected by ***unilateral CLP***
Exome-sequenced

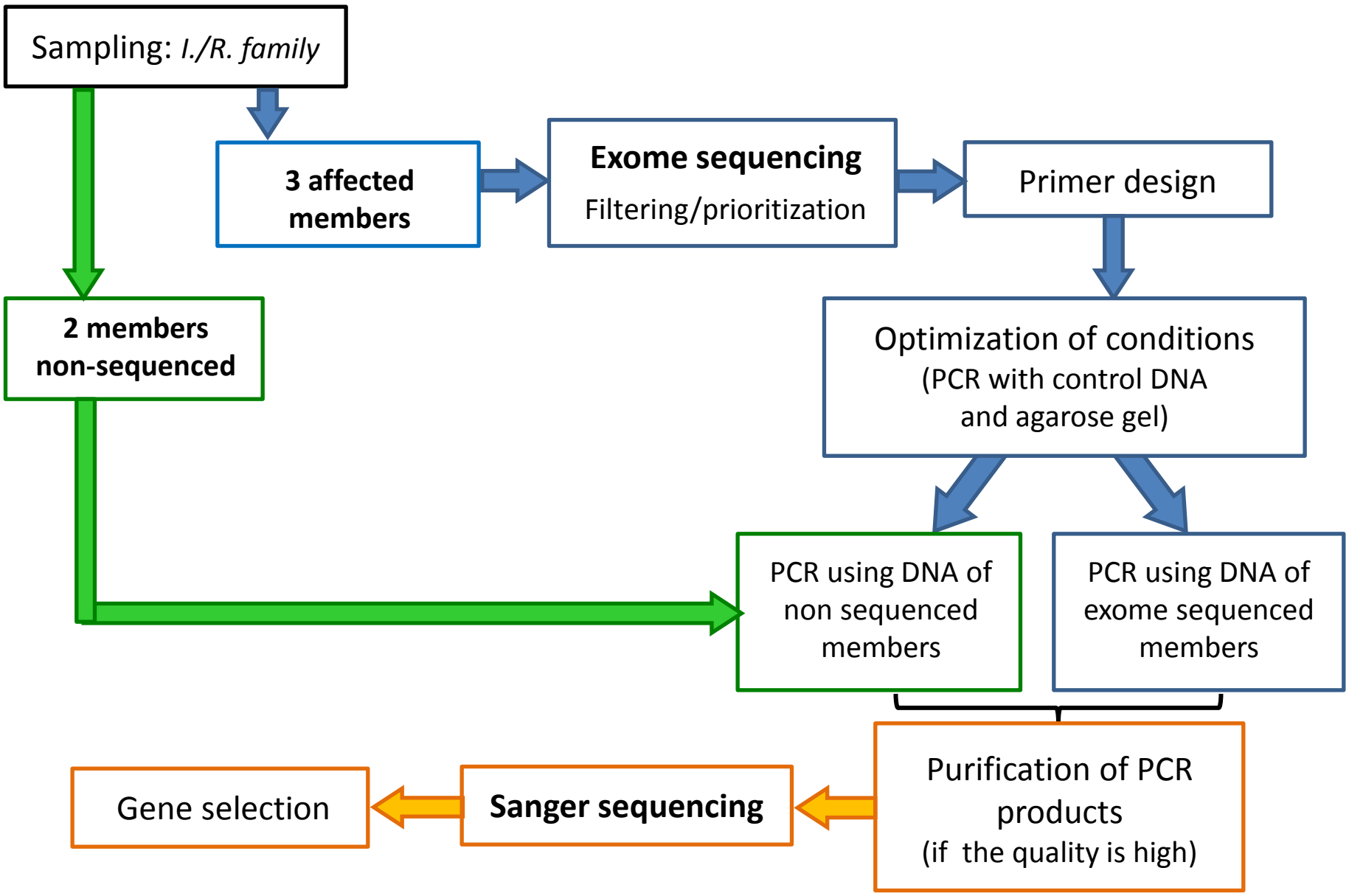
Member A2-B (n. 6):
affected by ***bilateral CLP***
Exome-sequenced

Member A3-B (n. 8):
affected by ***bilateral CLP***
Exome-sequenced

Member A4-UL (n. 14):
affected by ***unilateral CLP***
Not exome-sequenced

Member H (n. 7):
not affected (***healthy***)
Not exome-sequenced

PROJECT OVERVIEW



RAW RESULTS: from 25,000 to 31 variants

~**25,000** variants were identified using ***exome sequencing*** *(Illumina/Solexa platform)*

~**200** variants were filtered using ***Microsoft Excel***

- number of reading
- delete all intronic varinats
- delete all non-overlapping variants
- delete homozygous variants

31 variants were prioritized by ***Endeavour®*** (based on more than twenty sources)

…and these 31 variants were further checked with ***Sanger Sequencing***.

SANGER SEQUENCING VALIDATION

	PCR e primers			Sanger seq. results					
Locus (according to GeneCards - EntrazGene)	Band size	Annealing temp. used	Exome Seq. result	member 61808 AFFECTED	member 61810 AFFECTED	member 61812 AFFECTED	member B.I. (no exome seq.) AFFECTED	member MONICA (no exome seq.) NOT-AFFECTED	Type of variant
10q24.2	428 bp	60.0 °C	G > A	G > A	G > A	G > A	G	G	missense G>D
5q11.2	403 bp	60.0 °C	G > C	G > C	G > C	G > C	G	G	missense C>S
16p12.2	498 bp	60.0 °C	T > G	T > G	T > G	T > G	T	T	missense F>C
20q13.2	38								
2q31.1	48								missense V>I
12q13.13	44								missense L>V
2q31.2	45								missense Y>S
18q12.2	49							T	missense V>I
14q11.2	460 bp								missense A>G
2q35	413 bp								
16q12.2	477 bp								
1p36.13	444 bp								
1q42.3	575 bp								
3q26.31	372 bp								
17q24.3	364 bp								
9p24.3	463 bp								
1p32.3	530 bp	60.0 °C	C > T	C > C/T	C > C/T	C > C/T	C > C/T	C	missense R>W
18q12.1	475 bp	60.0 °C	G > A	G > G/A	G > G/A	G > G/A	G > G/A	G	missense V>I
11q21-22	476 bp	60.0 °C	C > A	C > C/A	C > C/A	C > C/A	C > C/A	C	missense L>I
4q22.1	452 bp	60.0 °C	C > T	C > C/T	C > C/T	C > C/T	C > C/T	C > C/T	missense T>M
21q22.3	520 bp	60.0 °C	G > T	G > G/T	G > G/T	G > G/T	G	G	missense G>V
16p12.2	404 bp	60.0 °C	C > A	C > C/A	C > C/A	C > C/A	C	C	missense P>T
6q22.2-3	452 bp	60.0 °C	G > A	G > G/A	G > G/A	G > G/A	G > G/A	G > G/A	missense H>Y
19p13.3	487 bp	60.0 °C	T > C	T > C/T	T > C/T	T > C/T	T > C/T	T > C/T	missense F>L
15q14	445 bp	60.0 °C	T > A	T > T/A	T > T/A	T > T/A	T	T > T/A	missense M>L
6q25.1	495 bp	60.0 °C	G > A	G > G/A	G > G/A	G > G/A	G	G	missense R>W

…all the 31 variants were validated (confirmed)

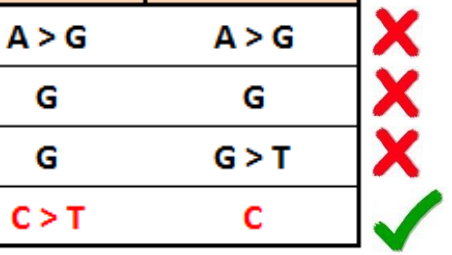
by using Sanger sequencing.

First selection: SEGRAGATION PATTERN

The first selection was based on the **segregation pattern**.

Four possible patterns:

	---------- exome-sequenced ----------			*--- not exome-sequenced ---*	
Gene name	**member A1-UR affected**	**member A2-B affected**	**member A3-B affected**	**member A4-UL affected**	**member H healthy**
Gene A	**A > G**	**A > G**	**A > G**	**A > G**	**A > G**
Gene B	**G > T**	**G > T**	**G > T**	**G**	**G**
Gene C	**G > T**	**G > T**	**G > T**	**G**	**G > T**
Gene D	**C > T**	**C > T**	**C > T**	**C > T**	**C**



We chose the variants which were present ONLY in the four affected members and absent in the non-affected one.

First selection: THE SIX TOP CANDIDATE GENES

After the first selection (based on the segregation pattern), **6 genes** were selected:

	Exome-sequenced			Non-sequenced	
Gene locus	*Member A1-UR*	*Member A2-B*	*Member A3-B*	*Member A4-UL*	*Member H*
1p32.3	C > T	C > T	C > T	C > T	C
1p36.13	G > A	G > A	G > A	G > A	G
3q26.31	A > G	A > G	A > G	A > G	A
11q21-22	C > A	C > A	C > A	C > A	C
12q24.31	C > G	C > G	C > G	C > G	C
18q12.1	G > A	G > A	G > A	G > A	G

First selection: ONE-POINT LOD SCORE

Additionally, for each variant we also calculated the **one-point LOD score** to assess the likelihood of linkage between the variant and cleft phenotype.

The formula of one-point LOD score (considering the *phase uncertainty*) is edited as follows:

$$\log_{10}\left[\left(\frac{1}{2}\cdot\frac{(1-d)^{P}\cdot d^{R}}{0.5^{(P+R)}}\right)+\left(\frac{1}{2}\cdot\frac{(1-d)^{R}\cdot d^{P}}{0.5^{(P-R)}}\right)\right]$$

d : *Recombinant distance (d = 0)*
P : *Parental frequency*
R : *Recombinant frequency*

First selection: ONE-POINT LOD SCORE

Four types of variant segregation pattern were present in *I./R. family*, so the genes were divided into four groups. The variants in the same group showed the same LOD score.

Loci	*Description*	*Parental*	*Recombinant*	*LOD*
1p32.3 1p36.13 3q26.31 11q21-22 12q24.31 18q12.1	*Present in 4 affected members. Absent in healthy member.*	5	0	**1,204098**
1q42.3 2q35 4q22.1 6q22.2-3 6q25.1 12q13.13 19p13.3	*Present in 4 affected members. Present in healthy member.*	4	1 (member H)	**-3,7959**
3p22.2 3q21.3 5q11.2 10q24.2 21q22.3 17q24.3 18q12.2 16p12.2 (I) 16p12.2 (II)	*Present in 3 affected members. Absent in healthy member. Absent in 1 affected member.*	4	1 (member A4-UL)	**-3,7959**
2q31.1 2q31.2 7q32.1 9p24.3 13q12.11 14q11.2 15q14 16q12.2 20q13.2	*Present in 3 affected members. Present in healthy member. Absent in 1 affected member.*	3	2 (members H and A4-UL)	**-8,79589**

Second selection: DIFFERENT PARAMETERS

The further selection was based on *different parameters*:

1) severity of amino acid substitution;
2) gene and protein characteristics;
3) presence of the selected variants in EVS database;
4) presence of recorded patients with chromosomal alterations involving these genes, who show pathological phenotypes;
5) gene expression levels in mouse palate tissue;
6) availability of KO mouse models.

Second selection: DIFFERENT PARAMETERS

The further selection was based on *different parameters*:

1) **severity of amino acid substitution**;
2) gene and protein characteristics;
3) presence of the selected variants in EVS database;
4) presence of recorded patients with chromosomal alterations involving these genes, who show pathological phenotypes;
5) gene expression levels in mouse palate tissue;
6) availability of KO mouse models.

Second selection: SEVERITY OF AMINO ACID SUBSTITUTION

Using **PolyPhen-2®**, we checked the effect of amino acid substitutions due to these variants in the corresponding proteins.

Gene locus	*PolyPhen-2® (HumDiv)*			
	Severity prediction	*Score*	*Sensitivity*	*Specificity*
1p32.3	Probably damaging	1.000	0.00	1.00
1p36.13	Probably damaging	1.000	0.00	1.00
3q26.31	Probably damaging	0.997	0.41	0.98
11q21-22	Probably damaging	0.995	0.68	0.97
12q24.31	Benign	0.004	0.97	0.59
18q12.1	Possibly damaging	0.799	0.84	0.93

Second selection: DIFFERENT PARAMETERS

The further selection was based on *different parameters*:

1) severity of amino acid substitution;
2) **gene and protein characteristics**;
3) presence of the selected variants in EVS database;
4) presence of recorded patients with chromosomal alterations involving these genes, who show pathological phenotypes;
5) gene expression levels in mouse palate tissue;
6) availability of KO mouse models.

Second selection: GENE/PROTEIN CHARACTERISTICS

We searched for details regarding each gene and its corresponding protein, by using ***GeneCards®*** and ***UniProt®*** databases.

Gene locus	***Protein size***	***Domain containing aa substitution***	***Protein name and function***	***Disease related to gene defects (link from OMIM®)***
1p32.3	74 kDa	***CoA-binding domain***	Inner mitochondrial protein; involved in the palmitoyl-CoA shuttle system (LCFA) and in the metabolism of lipid and lipoproteins (fatty acids oxidation and PPAR-α pathway).	CPT deficiency late-onset / lethal neonatal; hepato-cardio muscular manifestations; encephalopathy acute infection-induced type 4
1p36.13	18 kDa	*unknown function*	Uncharacterized protein in human	No clinical evidences
3q26.31	133 kDa	*unknown function*	Single-pass membrane protein (cytosol); matrix-cell adhesion; cell differentiation in embryonic development; involved in wound healing; regulator of adipogenesis.	Cancer; fibrosis; altered embryonic development
11q21-22	493 kDa	***Stem domain***	Motor for the intraflagellar retrograde transport; component of cilium; intracellular transport RE-Golgi.	Asphyxiating thoracic dystrophy type 3; short rib-polydactyly syndrome type 3/2B
12q24.31	28 kDa	*unknown function*	Interaction with p63 in the cellular cycle regulation (related to tumorigenesis).	Tumorigenesis (cervical/breast cancer); leukemia; lymphoma; cervicitis; leptospirosis
18q12.1	122 kDa	***5th structural repeat***	Single-pass membrane protein; component of intercellular desmosomes; role in the apoptosis and in the degradation of cell adhesion proteins.	Familial arrhythmogenic right ventricular dysplasia type 10; susceptibility to cardiomyopathy dilated type 1BB; signet ring cell adenocarcinoma; renal clear cell carcinoma; keratosis; keratoacanthoma: squamous cell carcinoma; arachnoiditis

Second selection: DIFFERENT PARAMETERS

The further selection was based on *different parameters*:

1) severity of amino acid substitution;
2) gene and protein characteristics;
3) **presence of the selected variants in EVS database**;
4) presence of recorded patients with chromosomal alterations involving these genes, who show pathological phenotypes;
5) gene expression levels in mouse palate tissue;
6) availability of KO mouse models.

Second selection: PRESENCE IN EVS

Afterwards, we used ***EVS®*** (*Exome Variant Server*) database in order to verify if the variants, identified by exome sequencing, had ever been seen before.

Gene locus	*Presence in EVS®*	*Clinical link*
1p32.3	no	---
1p36.13	**yes**	Unknown
3q26.31	no	---
11q21-22	**yes**	Unknown
12q24.31	no	---
18q12.1	**yes**	Unknown

Second selection: DIFFERENT PARAMETERS

The further selection was based on *different parameters*:

1) severity of amino acid substitution;
2) gene and protein characteristics;
3) presence of the selected variants in EVS database;
4) **presence of recorded patients** with **chromosomal alterations** involving these genes, who show **pathological phenotypes**;
5) gene expression levels in mouse palate tissue;
6) availability of KO mouse models.

Second selection: DIFFERENT PARAMETERS

Furthermore, we checked the presence of patients (recorded in ***Decipher***®) with chromosomal alterations involving these genes, who show pathological phenotypes, such as: *cleft lip*, *cleft palate*, *cleft lip and palate*, other craniofacial malformations.

Gene locus	*Patient ID num.*	*Type of alteration*	*Dimension*	*Location*	*Pathologic phenotypes (cranio-facial malformations)*
3q26.31	2757	Deletion (7 genes)	1.26 Mb	chr3:171214816-172472424	Ptosis, Abnormality of the outerear
1p32.3	253629	Duplication (105 genes)	16.02 Mb	chr1:46713932-62729246	**Thick lower lip vermilion**, **Thick upper lip vermilion**, Prominent ears
	272313	Deletion (57 genes)	7.30 Mb	chr1:47090879-54388982	**Hypoplasia of the maxilla**, **Exaggerated cupid's bow**, Microcephaly, Prominent glabella
1p36.13	**2483**	Deletion (149 genes)	12.57 Mb	chr1:4795388-17364849	**Submucous cleft hard palate**, **Thin upper lip vermilion**, **Thick lower lip vermilion**, Downturned corners of mouth, Midface retrusion, Prominent ears
	254939	Deletion (83 genes)	5.82 Mb	chr1:15325975-21141171	**High palate**, epicanthus, hypoplasti nasal alae, Microcephaly
	259769	Duplication (318 genes)	24.41 Mb	chr1:712577-25120526	Depressed nasal bridge
11q21-22	767	Duplication (51 genes) - related to the first mutation	**11.48 Mb** / 5.47 Mb / 20.42 Mb	**chr11:9410383131-105588056** chr11:107792125-113258885 chr11:114235228-134651277	**Bifid uvula**, **Abnormality of the labia**, Preauricular pit, Hydrocephalus
	248786	Duplication (246 genes)	27.11 Mb	chr11:100747981-127862005	**High palate**, Encephalocele, Frontal bossing, Trigonocephaly, Brechycephaly, Prominent nose
	249758	Duplication (68 genes)	14.49 Mb	chr11:92254068-106746797	**Thick upper lip vermilion**, **Thick lower lip vermilion**, **Short philtrum**, Wide nasal bridge, Bifid nasal bridge, Brechycephaly, Microcephaly, Synophrys, Blepharophimosis
	251725	Deletion (84 genes)	18.74 Mb	chr11:84405018-103141743	Midface retrusion
	262828	Deletion (66 genes)	13.90 Mb	chr11:92765018-106662479	Abnormality of the face (*not specifie* d)
18q12.1	**1581**	Duplication (177 genes) - related to the first mutation	**55.90 Mb** / 14.87 Mb	**chr18:21936109-77839271** chr18:140336-15008636	**Cleft palate**, **Non-midline cleft lip**, Long face, Blepharophimosis
	260121	Deletion (46 genes)	13.40 Mb	chr18:22032122-35430900	Abnormality of the face (*not specified*)
	266270	Duplication (263 genes)	77.92 Mb	chr18:83701-78001525	**Micrognathia**, Depressed nasal bridge
12q24.31	263700	Deletion (43 genes)	9.03 Mb	chr12:124743122-133773534	Abnormality of the face (*not specified*), Hydrocephalus

Second selection: DIFFERENT PARAMETERS

The further selection was based on *different parameters*:

1) severity of amino acid substitution;
2) gene and protein characteristics;
3) presence of the selected variants in EVS database;
4) presence of recorded patients with chromosomal alterations involving these genes, who show pathological phenotypes;
5) **gene expression levels in mouse palate tissue**;
6) availability of KO mouse models.

Second selection: GENE EXPRESSION IN MOUSE PALATE

Moreover, we evaluated also the expression levels of these genes in the embryonic mouse palate (mesenchymal tissue).
The genic expression of each gene in this tissue was evaluated using the ***RNA-seq data*** generated by an external collaborator, ***Dr. M.J. Dixon***, and his group (*Academic Health Sciences Centre, University of Manchester*).

Gene locus	*Expression level in mouse palate*
1p32.3	Low
1p36.13	(not expressed)
3q26.31	Very high
11q21-22	High
12q24.31	Very high
18q12.1	Medium

Second selection: DIFFERENT PARAMETERS

The further selection was based on *different parameters*:

1) severity of amino acid substitution;
2) gene and protein characteristics;
3) presence of the selected variants in EVS database;
4) presence of recorded patients with chromosomal alterations involving these genes, who show pathological phenotypes (cleft lip, palate, lip and palate or other cranio-facial malformations);
5) gene expression levels in mouse palate tissue;
6) **availability of KO mouse models**.

Second selection: KO MOUSE STRAINS

Finally, we checked the gene inactivation effect on the phenotype, analyzing the data of knockout mouse models available online (on ***MGI***® and ***IMSR***®), which would be also useful to plan further functional analyses *in vivo*.

Gene locus	*Strains available*	*Description*	*Phenotypic summary*
1p32.3	no	---	---
1p36.13	no	---	---
3q26.31	2 strains available	1st strain: targeted (KO); intragenic deletion	Lethal for homozygous mutants (-/-); increased level of IgG2a in serum
		2nd strain: targeted (KO); intragenic deletion	Lethal for homozygous mutants (-/-); abnormal cell differentiation/adhesion; decreased fibroblast migration and proliferation; abnormal bone ossification
11q21-22	2 strains available	1st strain: targeted (KO); insertion	Lethal for homozygous mutants (-/-); abnormal cell proliferation; increased apoptosis; loss of Shh-dependent signaling in the neural tube (no other information available)
		2st strain: targeted (KO); chemically induced mutation	Pulmonary atresia with ventricular septal defect; atrioventricular septal defect; major aortopulmonary collateral arteries; micrognathia; hypotelorism; duplex kidney and agenesis; polydactyly; syndactyly; oligodactyly; tracheoesphageal fistula; eye malformation; mouth malformation
12q24.31	3 strains recorder (the details are reserved)		
18q12.1	1 strains available	Targeted (KO); disruption by insertion of vector	Homozygous (-/-) embryos die around implantation; abnormal cell death; decreased fibroblast proliferation

THE THREE BEST CANDIDATE GENES

Comparing all these data, we have chosen three top candidate genes which should be analyzed further: these genes are located at **1p32.3**, **3q26.31** and **12q24.31**.

Gene locus	*Type of amino acids change*	*Protein dimension*	*Position of aa mutation in the protein structure*	*Protein function*	*Expression in mouse palate*	*KO mouse models available*	*Patients with alterations involving the gene (Decipher®)*	*Presence of the variant in EVS® database*	*Selected*
1p32.3	R > W *(Probably damaging)*	74 kDa	***CoA-binding domain***	Inner mitochondrial protein; involved in the palmitoyl-CoA shuttle system and in the metabolism of lipid and lipoproteins.	Low	*NO*	2	No	Yes
1p36.13	D > N *(Probably damaging)*	169 KDa	*Portion without particular function*	Uncharacterized protein in human	Not expressed	*NO*	3 *(one presents cleft of palate)*	Yes	No
3q26.31	Y > C *(Probably damaging)*	133 kDa	*Portion without particular function*	Single-pass membrane protein (cytosol); regulator of adipogenesis.	Very high	2 KO strains	1	No	Yes
11q21-22	L > I *(Probably damaging)*	493 kDa)	***Stem domain***	Motor for the intraflagellar retrograde transport, component of cilium, intracellular transport RE-Golgi.	High	2 KO strains	5 *(one presents bifid uvula)*	Yes	No
12q24.31	L > V *(Benign)*	28 kDa)	*Portion without particular function*	Interaction with p63 in the cellular cycle regulation (related to tumorigenesis).	Very High	*3 KO strains (no data available)*	1	No	Yes
18q12.1	V > I *(Possibly damaging)*	122 kDa)	***5th structural repeat***	Single-pass membrane protein; component of intercellular desmosomes; role in the apoptosis and in the degradation of cell adhesion proteins.	Medium	1 KO strain	3 *(one presents cleft of lip and palate)*	Yes	No

CONCLUSION

In conclusion, comparing all the data found during the second selection phase, for *I./R. family* three genes have been selected.

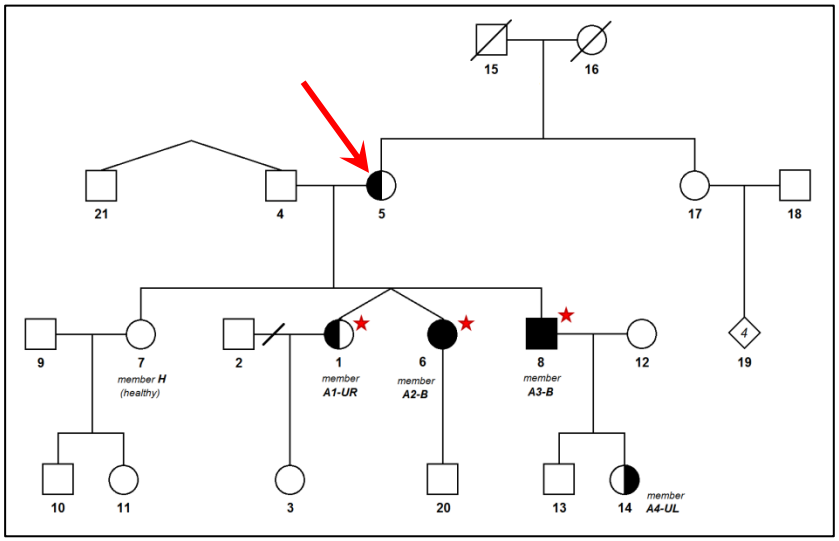
These genes are located at:
- *1p32.3*
- *3q26.31*
- *12q24.31*

In all probability, the most interesting gene could be the one whose locus is ***3q26.31***. It encodes for a structural protein involved in the cell-matrix adhesion and in cell differentiation, particularly during embryonic development. Interestingly, defects in this gene are associated with anomalies in embryogenesis and morphogenesis, both in human and in mouse.

Anyway, we are not yet able to conclude if one of these genes is a causative gene for NSCLP...

FUTURE PERSPECTIVES

The next step will be to improve the statistical power of this family by analyzing other members (such as **member n. 5**) and/or checking the presence of these candidate variants also in other Australian families, which show high recurrence of NSCLP.



So, if the variants will be identify also in other families, the different LOD scores (regarding the same variant) will be added together: in the end, if the resulting value reaches the significance cut-off value (3), we will be able to confirm the association between the variant and the phenotype.

In this case, further analyses will be performed (e.g. functional analyses), both *in vitro* and *in vivo* models, in order to investigate the role of this potential causative variant in the nonsyndromic cleft lip and palate development.