Supplementary Information

GATA6 haploinsufficiency causes pancreatic agenesis in humans

Hana Lango Allen[1,6], Sarah E Flanagan[1,6], Charles Shaw-Smith[1,6], Elisa De Franco[1,6], Ildem Akerman[2,3,4], Richard Caswell[1], the International Pancreatic Agenesis Consortium[5], Jorge Ferrer[2,3,4], Andrew T Hattersley[1] and Sian Ellard[1]

[1] Institute of Biomedical and Clinical Science, Peninsula College of Medicine and Dentistry, University of Exeter, Exeter, UK.

[2] Genomic Programming of Beta-cells Laboratory, Institut d'Investigacions August Pi i Sunyer (IDIBAPS), Barcelona, Spain.

[3] CIBER de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM), Barcelona, Spain.

[4] Department of Endocrinology and Nutrition, Hospital Clínic de Barcelona, Barcelona, Spain.

[5] A full list of members is in the Supplementary Note.

[6] These authors contributed equally to this work.

Correspondence should be addressed to A.T.H. (Andrew.Hattersley@pms.ac.uk).

Table of Contents

Supplementary Note

Members of the International Pancreatic Agenesis Consortium:

Henk-Jan Aanstoot[1], Erika Aberg[2], Peter Adolfsson[3], Mark Anthony[4], Marta Batlle[5], Jan Bruining[6], Patrick Bühr[7], Elizabeth Cummings[8], Julie Edge[9], Javier Garcia-Hurtado[5], Eba Hathout[10], Josephine Ho[11], Craig Jefferies[12], I Karen Temple[13], Ellen Kaufman[14], Vjosa Mulliqi Kotori[15], Ronald de Krijger[16], Mark Kummer[17], Deborah Mackay[13], John Mace[10], Cees Noordam[18], Frances O'Brien[4], Oscar Rubio-Cabezas[19], Julian Shield[20], David Skidmore[8], Sue White[21], Ulrike Zanier[22]

[1]Center for Pediatric and Adolescent Diabetes, Rotterdam, the Netherlands; [2]Maritime Medical Genetics, IWK Health Centre Halifax, Nova Scotia, Canada; [3]The Queen Silvia Children's Hospital, Institute of Clinical Sciences, Sahlgrenska Academy at University of Gothenburg, Göteborg, Sweden; [4]Department of Neonatology, John Radcliffe Hospital, Oxford, UK; [5]Genomic Programming of Beta-cells Laboratory, Department of Endocrinology and Nutrition, Hospital Clínic de Barcelona, Barcelona, Spain; [6]Department of Paediatrics, Maxima Medical Centre, Veldhoven, University Medical Centre, Groningen, the Netherlands; [7]Department of Pediatrics, Landeskrankenhaus Feldkirch, Feldkirch, Austria; [8]Department of Pediatrics, Dalhousie University, Halifax, Nova Scotia, Canada; [9]Department of Paediatric Endocrinology, Oxford Children's Hospital, Oxford, UK; [10]Division of Paediatric Endocrinology and Diabetes, Loma Linda University School of Medicine, California, USA; [11]Alberta Children's Hospital, Department of Pediatric Endocrinology, Calgary, Alberta, Canada; [12]Department of Pediatric Endocrinology, Starship Children's Hospital, Auckland, New Zealand; [13]Faculty of Medicine, University of Southampton, UK; [14]Presbyterian Medical Group, Albuquerque, New Mexico, USA; [15]University Clinical Center, Pediatric Clinic department of Endocrinology, Pristina, Kosovo; [16]Erasmus MC University Medical Center, Rotterdam, the Netherlands; [17]Nemours Children's Clinic, Pensacola, Florida, USA; [18]Department of Pediatrics and Children's Diabetes Center, Radboud University Nijmegen Medical Center, Nijmegen, the Netherlands; [19]Department of Endocrinology, Hospital Infantil Universitario Niño Jesús, Madrid, Spain and Institute of Biomedical and Clinical Science, PMS, Exeter, UK; [20]Department of Pediatric Endocrinology, Bristol Royal Hospital for Children, Bristol, UK; [21]Genetic Health Services, Royal Children's Hospital, Parkville, Victoria, Australia; [22]Krankenhaus Dornbirn, Dornbirn, Austria

The consortium members were responsible for analysing the clinical data for the subjects reported in this manuscript.

Supplementary Methods

Subject ascertainment and sample preparation

We studied a cohort of 27 subjects, born to non-diabetic parents, who had pancreatic agenesis defined as neonatal diabetes requiring insulin treatment and exocrine pancreatic insufficiency requiring enzyme replacement therapy. Subjects with pancreatic agenesis were recruited by their clinicians for molecular genetic analysis in the Exeter Molecular Genetics Laboratory. All subjects or their parents gave informed consent for genetic testing.

Genomic DNA was extracted from peripheral leukocytes using standard procedures and the coding exons and intron/exon boundaries of the *PTF1A* and *PDX1* genes were amplified by PCR (primers and conditions available on request). PCR products were sequenced using standard methods on an ABI 3730 (Applied Biosystems, Warrington, UK) and sequences were compared to the published sequences (*PDX1:* NM_000209.3 and *PTF1A*: NM_178161.2) using Mutation Surveyor v3.95 (Soft Genetics, PA, USA).

Exome sequencing and variant calling

Genomic regions corresponding to NCBI Consensus Coding Sequence (CCDS) database were captured and amplified using Agilent's SureSelect Human All Exon Kit (v1). Paired-end sequencing was performed on an Illumina GAII, one lane per sample, with 101 or 76 bp read length. The resulting reads were aligned to the hg19 reference genome with BWA[1], providing us with an average of 7.0Gb and 4.2Gb aligned sequence, and mean target coverage of 105x and 63x, respectively. In all cases at least 95% of the targeted bases were covered by at least 2 reads, and 77-83% by at least 20 reads (**Supplementary table 1**).

We applied Picard (http://picard.sourceforge.net) duplicates removal and GATK[2] local realignment around indels, and performed SNP and InDel discovery across all 6 samples using recommended hard filtering parameters[3] and minimum depth of 10 reads per base. We used the combination of ANNOVAR[4] and SeattleSEQ SNP Annotation server (http://snp.gs.washington.edu/SeattleSeqAnnotation131/) to functionally annotate variants, and in-house mysql queries to identify non-synonymous *de novo* variants. A single candidate mutation was identified in each subject, both in the *GATA6* gene (**Supplementary table 2**).

Mutation confirmation and further genetic analysis

We confirmed both *de novo* variants by Sanger sequencing of the subjects and their parents. We then sequenced exons 2-7 and intron/exon boundaries of *GATA6* in the remaining cases with pancreatic agenesis but no causative mutation (**Supplementary table 1**). Primers for *GATA6* exons 2-7 are provided in **Supplementary table 3.**

We identified *GATA6* mutations in further 13 subjects, bringing the total number of *GATA6* positive subjects to 15/27 (56% of the total cohort). Parental DNA samples were available for 10 of these additional cases and testing confirmed all of these mutations as *de novo*. Biological relationships were confirmed by microsatellite analysis using the PowerPlex kit (Powerplex 16 System, Promega, Southampton, UK).

In total 6 different *GATA6* missense mutations were identified in 7 probands. All mutations occurred within the DNA binding domain at residues that are highly conserved across species (conserved to Drosophila). In three subjects mutations within the canonical splice sites were identified. Two mutations occurred within the 5' splice site of intron 5 and are predicted to abolish (c.1516+1G>A) or reduce the strength (c.1516+4A>G) of the splice donor site. The third splicing mutation (c.1303-10C>G) is predicted to introduce a cryptic splice acceptor site within intron 3. The subsequent incorporation of 9 nucleotides (TGTTTCTAG) from intron 3 into the mRNA is predicted to result in a premature termination of translation at the third codon within exon 4 of the novel transcript (splicing prediction software was accessed through Alamut Interactive Biosoftware, version 1.5, Rouen, France). Five different frameshift mutations were identified in 5 probands, and in all cases the mutations were predicted to introduce a

premature termination codon. None of the mutations were present in the dbSNP132 or 1000 genomes databases (Jan 2011 release, based on 1094 individuals).

The non-coding exon 1 of *GATA6* was amplified and sequenced in the 11 remaining subjects but no mutations were identified. The primer sequences are provided in **Supplementary table 3**.

Structural modelling

The model in **Supplementary Figure 1** shows the predicted structure of GATA6 residues 438-493 complexed to DNA, and was generated using the Swiss-Model web server (http://swissmodel.expasy.org/) based on the structure of mouse GATA3 residues 311-366 bound to DNA.

Functional Studies of GATA6 Zinc-finger mutants.

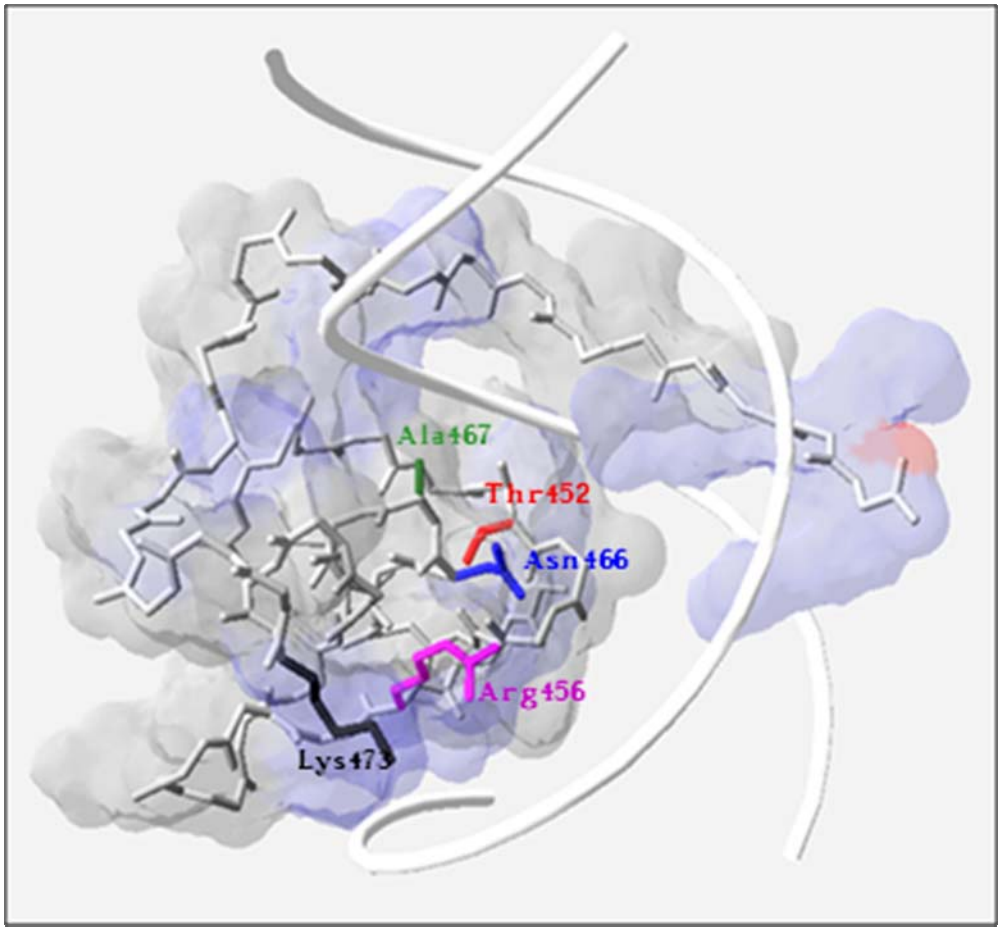
Mammalian expression vector pcDNA3.1(+)- GATA6 (Hiroyuki Yamagishi, Keio University, Japan) [5] was used to perform Dpn I-mediated site-directed mutagenesis to generate *GATA6* Arg456Cys (c.1366C>T), Asn466Asp (c.1396A>G), Ala467Thr (c.1399G>A) and Lys473Gln (c.1417A>C) mutations. Two independent clones were created for each base substitution, sequenced, and assayed in transient transfection assays in duplicates on three independent experiments, as described [6]. Briefly, cells were transfected with 0.15 µg of pGL3-*WNT2* promoter-Luciferase construct, 0.6 ng pRL Renilla Luciferase reporter vector, in conjunction with 0.35 µg of empty vector (pcDNA3.1), pcDNA3.1-GATA6 or pcDNA3-GATA6 mutants using Lipofectamine 2000. Firefly and Renilla luciferase activity was assayed using Dual-Luciferase Reporter Assay System (Promega). Binding of nuclear lysates that contain GATA6 and GATA6 mutant proteins Arg456Cys, Asn466Asp, Ala467Thr and Lys473Gln to P^{32}-labeled oligonucleotides that contain consensus binding sites for GATA6 was performed as described [6]. Sequence of oligonucleotides used in this assay include a predicted GATA6 binding site in the promoter of the pancreatic (P2) *HNF4A* proximal promoter (*HNF4A*-P2 5´-gatcATCAATAAGATAACCGCGCG - 3´ and *HNF4A*-P2 Mutant 5´-gatcATCAATAAGCCAACCGCGCG - 3´) or the TFF2 consensus sequence (TFF2 5´-gatcGCCAGCAGATAGCATGGAAAAG - 3´).

Specificity of retardation complex was assessed by preincubating nuclear extracts with 50-fold excess wild type or mutant unlabeled oligonucleotides, or GATA6 antiserum (SC-9055, Santa Cruz Biotechnology).

References

1. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).
2. McKenna, A. et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297-1303 (2010).
3. DePristo, M.A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491-498 (2011).
4. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
5. Kodo, K. et al. GATA6 mutations cause human cardiac outflow tract defects by disrupting semaphorin-plexin signaling. *Proc. Natl. Acad. Sci.* **106**, 13933-13938 (2009).
6. Boj, S.F., Parrizas, M., Maestro, M.A. & Ferrer, J. A transcription factor regulatory circuit in differentiated pancreatic cells. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 14481-6 (2001).

Supplementary Figure 1. Location of the zinc finger domain mutations. The protein backbone is shown as a solid grey line; the side chains of residues affected by missense substitutions are shown in various colours and labelled by position. The computed molecular surface is shown as a transparent layer (light grey shading) with regions of positive electrostatic charge (light blue shading); the backbones of DNA chains are shown as white ribbons.



Supplementary Figure 2. a) Evolutionary amino acid conservation in GATA6 zinc finger 2 region. Alignment to amino acids 441-490 of human GATA6 (UniprotKB entry Q92908) is shown for vertebrate orthologues with complete entries in the UniProtKB database; amino acid numbering is taken from database entries as follows: rhesus macaque, F7GIW0; mouse, Q61169; rat, P46153; opossum, F7G2T6; chicken, P43693; zebrafish, Q6NW63; *X. laevis*, Q91678; *X. tropicalis*, Q7T1R5. Positions of missense mutations reported in this study are shown in the human sequence in underlined, bold font; for orthologous sequences, positions of identity are shown in normal font, conserved amino acids by grey shading and non-conserved amino acids by inverted type. **b) The zinc finger 2 region of human GATA family proteins 1-6 aligned**. Amino acid numbering is taken from the UniProtKB entry for each protein (GATA1, P15976; GATA2, P23769; GATA3, P23771; GATA4, P43694; GATA5, Q9BWX5). Identity in the alignment is indicated by an asterisk underneath the sequence; the residues affected by missense mutations reported in this study are shown in bold font, and are conserved across all human GATA family members.

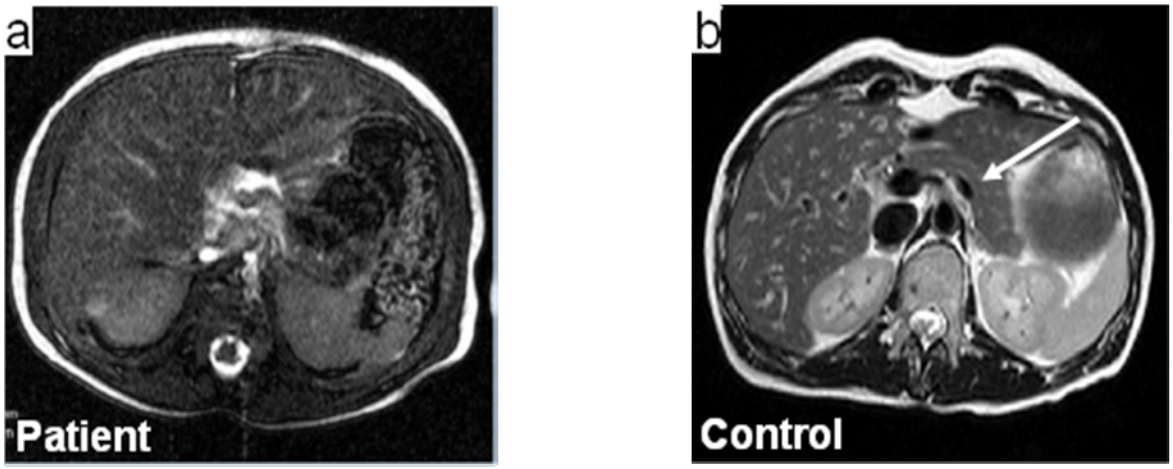
a

```
Human       441 GLSCANCHTTTTTLWRRNAEGEPVCNACGLYMKLHGVPRPLAMKKEGIQT 490
Rhesus      439 GLSCANCHTTTTTLWRRNAEGEPVCNACGLYMKLHGVPRPLAMKKEGIQT 488
Mouse       435 GLSCANCHTTTTTLWRRNAEGEPVCNACGLYMKLHGVPRPLAMKKEGIQT 484
Rat         435 GLSCANCHTTTTTLWRRNAEGEPVCNACGLYMKLHGVPRPLAMKKEGIQT 484
Opossum     436 GLSCANCHTTTTTLWRRNAEGEPVCNACGLYMKLHGVPRPLAMKKEGIQT 485
Chicken     232 GLSCANCHTTTTTLWRRNAEGEPVCNACGLYMKLHGVPRPLAMKKEGIQT 281
Zebrafish   348 GLSCANCQTSTTTLWRRNAEGEPVCNACGLYTKLHGVPRPLAMKKEGIQT 397
X.laevis    233 GLACANCHTSTTTLWRRNTEGEPVCNACGLYMKLHGVPRPLAMKKEGIQT 282
X.trop.     367 GLACANCHTTTTTLWRRNTEGEPVCNACGLYMKLHGVPRPLAMKKEGIQT 416
```

b

```
GATA1       245 GTQCTNCQTTTTTLWRRNASGDPVCNACGLYYKLHQVNRPLTMRKDGIQT 304
GATA2       346 GTCCANCQTTTTTLWRRNANGDPVCNACGLYYKLHNVNRPLTMKKEGIQT 395
GATA3       314 GTSCANCQTTTTTLWRRNANGDPVCNACGLYYKLHNINRPLTMKKEGIQT 363
GATA4       268 GLSCANCQTTTTTLWRRNAEGEPVCNACGLYMKLHGVPRPLAMRKEGIQT 317
GATA5       240 GLCCTNCHTTNTTLWRRNSEGEPVCNACGLYMKLHGVPRPLAMKKESIQT 289
GATA6       441 GLSCANCHTTTTTLWRRNAEGEPVCNACGLYMKLHGVPRPLAMKKEGIQT 490
                    *  *:**:**.*******:.*:********* *** : ***:*:*:.***
```

Supplementary Figure 3. Mutations in *GATA6* are the most common cause of pancreatic agenesis. (a) MRI image (proband 5) showing total pancreatic agenesis; (b) control subject with pancreas indicated by white arrow.



Supplementary table 1. Exome sequence alignment and coverage data.

	Family 1			Family 2		
	Proband	**Mother**	**Father**	**Proband**	**Mother**	**Father**
Read length (bp)	101	101	101	76	76	76
Total reads	78,961,776	78,580,458	76,521,764	62,920,527	65,708,268	65,803,360
Uniquely aligned reads	72,593,975	72,223,629	70,498,304	57,321,731	59,824,413	60,002,718
Uniquely aligned bases	7,029,126,684	7,017,021,984	6,858,849,488	4,183,749,056	4,330,030,091	4,366,342,386
Total CCDS bases	2,909,282,613	2,944,353,630	2,876,886,228	1,707,269,060	1,774,334,679	1,764,345,195
Mean CCDS coverage	104.5	105.9	103.4	61.5	63.9	63.6
% CCDS bases covered >=2	96.1	96.0	96.1	95.5	95.6	95.4
% CCDS bases covered >=10	90.0	90.0	90.1	87.1	88.0	87.4
% CCDS bases covered >=20	83.3	83.4	83.4	76.9	78.2	78.0
% CCDS bases covered >=30	77.0	77.2	77.2	67.2	68.4	68.7

Supplementary Table 2. Breakdown of variants identified by exome sequencing of the two subjects. Various filtering steps narrowed down the list of possible pathogenic mutations to a single variant in each subject, and both of these were in the same gene, *GATA6*. *Some putative *de novo* variants were clearly present in at least one of the parents, but with total coverage of <10 reads so they were not initially called, but were identified upon manual inspection of the reads.

	Proband 1		Proband 2	
	substitutions	**indels**	**substitutions**	**indels**
Total passing quality filters	23,028	1493	22,439	1186
After dbSNP131 filtering	675	674	721	484
After 1000Genomes filtering	377	674	412	484
After excluding non-coding	274	94	271	56
After excluding those in parents (*de novo*)	8	6	6	4
After excluding synonymous/non-frameshift	7	2	0	2
After manual inspection of reads*	1	0	0	1
Gene	*GATA6*	n/a	n/a	*GATA6*

Supplementary Table 3. *GATA6* primer sequences.

Exon	Forward (M13 tailed) (All primers start 5' TGTAAAACGACGGCCAGT)	Reverse (M13 tailed) (All primers start 5' CAGGAAACAGCTATGACC)
Ex 1	TCCCCACCCTCTTTTCTCTC	AGGAGGAAGCAGCCGAAC
Ex2-a	AAGGAGTGGAGGCGAGGTAG	GGTCGAGGTCAGTGAACAGC
Ex2-b	CTCAGCTCGACACGGAGG	GTATGGAGGGCTGTCGGC
Ex2-c	CTCTCTCCAGCCAGGGTCC	GACAGCGAGCTGTACTGGG
Ex2-d	GCGCTTCCCCTACTCTCC	CTGCAAATCCTTCCTGGGAC
Ex3	AAAAGCTCAGCCGGGAAG	CTAGGGTGGGACCGCAG
Ex4	TCTTGGCCCAGAAAAGTCAG	AAAAGCACCTTCAATTCAATAA
Ex5_6	CATGCGCCAACAAGTCTG	CCAAAACTTCTTGCTTTTACTTGG
Ex7	GAGCAGCTCTGGCCCTG	CAAAATAAAGGCACGAGAATCAC

Supplementary Table 4. Clinical and molecular genetics characteristics of the subjects with *GATA6* mutations.

Proband	Mutation	Protein	*De novo*	Cardiac malformations	Additional endocrine abnormalities	Hepatobiliary malformations	Neurological abnormalities	Gut abnormalities
1	c.1354A>G	p.Thr452Ala	Yes	Atrial septal defect			Developmental delay	Colonic perforation
2	c.1448_1455del	p.Met483ArgfsX11	Yes	Multiple ventricular septal defects, atrial septal defect, mild hypoplasia of right ventricle and tricuspid valve, pulmonary stenosis, patent ductus arteriosus		Gall bladder agenesis		
3	c.1399G>A	p.Ala467Thr	No parental samples available (subject adopted)	Atrial septal defect, pulmonary stenosis	Pituitary agenesis		Moderate learning difficulties and seizures	
4	c.1303-10C>G	p.?	Yes	Interrupted aortic arch		Gall bladder agenesis		
5	c.1366C>T	p.Arg456Cys	Yes	Truncus arteriosus, perimembranous ventricular septal defect			Developmental delay and seizures	
6	c.1516+1G>C	p.?	Yes	Atrial septal defect, persistent foramen ovale	Transient hypothyroidism			
7	c.1366C>T	p.Arg456Cys	Yes	Tetralogy of Fallot			Developmental delay	Umbilical hernia
8	c.1498_1501del	p.Lys500GlnfsX14	Yes	Double outlet left ventricle, ventricular septal defect, hypoplastic pulmonary artery, valvular pulmonary stenosis, patent foramen ovale, persistent ductus arteriosus				
9	c.1396A>G	p.Asn466Asp	Paternal sample not available	Patent ductus arteriosus	Transient hypothyroidism	Gall bladder agenesis	Developmental delay and epilepsy	Intestinal malrotation and microcolon
10	c.1417A>C	p.Lys473Gln	Yes	Atrial septal defect		Gall bladder agenesis		
11	c.1516+4A>G	p.?	Yes	Normal echocardiogram				Left diaphragmatic hernia
12	c.877_880delinsTAC	p.Val293TyrfsX27	No parental samples available (subject adopted)	Tetralogy of Fallot		Biliary atresia	Microcephaly and learning difficulties	Inguinal hernia
13	c.701delC	p.Pro234HisfsX60	Yes	Atrial and ventricular septal defects				
14	c.1108_1121dup	p.Gly375SerfsX22	Yes	Tetralogy of Fallot				
15	c.1367G>A	p.Arg456His	Yes	Patent ductus arteriosus, ventricular septal defect, hypoplastic left pulmonary artery			Severe developmental delay	