Supplementary methods of article entitled:

Copy Number Variations in 375 patients with oesophageal atresia and/or tracheoesophageal fistula.

Authors and affiliations

Erwin Brosens[1,2]*,& Florian Marsch[3]*, Elisabeth M. de Jong[1,2], Hitisha P. Zaveri[4], Alina C. Hilger[3], Vera Gisela Choinitzki[3], Alice Hölscher[5], Per Hoffmann[3, 6], Stefan Herms[3, 6], Thomas M. Boemer[5], Benno M. Ure[7], Martin Lacher[7], Michael Ludwig[8], Bert H. Eussen[1], Robert M. van der Helm[1], Hannie Douben[1], Diane van Opstal[1], Rene M.H. Wijnen[2], H. Berna Beverloo[1], Yolande van Bever[1], Alice S. Brooks[1], Hanneke IJsselstijn[2], Daryl A. Scott[4], Johannes Schumacher[3], Dick Tibboel[2], Heiko Reutter[3,*], Annelies de Klein[1*]

*[1]Department of Clinical Genetics, Erasmus Medical Centre – Sophia Children's Hospital, Rotterdam, the Netherlands. [2]Department of Pediatric Surgery, Erasmus Medical Centre – Sophia Children's Hospital, Rotterdam, the Netherlands. [3]Institute of Human Genetics, University of Bonn, Bonn, Germany. [4]Department of Molecular and Human Genetics and [5]Departments of Molecular Physiology and Biophysics, Baylor College of Medicine, Houston, Texas, USA. [5]Department of Pediatric Surgery and Urology, Children's Hospital of Cologne, Cologne, Germany [6]Human Genomics Research Group, Department of Biomedicine, University Hospital and University of Basel, Basel, Switzerland. [7]Center of Pediatric Surgery Hannover, Hannover Medical School and Bult Children's Hospital, Hannover, Germany. [8]Department of Clinical Chemistry and Clinical Pharmacology, University of Bonn, Bonn, Germany. [9]Department of Neonatology and Pediatric Intensive Care, University of Bonn, Bonn, Germany. *authors contributed equally to the manuscript*

***authors contributed equally to the manuscript**

Corresponding Author:

#E. Brosens PhD

Erasmus Medical Centre – Sophia Children's Hospital

Room Ee971

P.O. Box 2040

3000 CA Rotterdam

Tel +31 10 70 37643

Fax: +31 10 70 44736

Study design

We assessed the CNVs according to the consensus statement for chromosomal micro-array analysis described by Miller and co-workers.[17] Our study design was based on the assumptions that CNVs are most likely to be causal for the abnormal phenotype in congenital anomalies if (I) a CNV is absent in large cohorts of unaffected individuals, (II) is absent in the unaffected parents of the affected individual and (III) and if it targets relevant genes or non-coding RNAs. Recurrence of loci affected by *de novo* CNVs in single cases could indicate loci harbouring genes mutated or otherwise affected in larger disease cohorts.

As OA/TOF disease frequency ranges between 2-3 in 10,000 live births (0.0002 to 0.0003), we classified CNVs to be rare if they were absent or present once in our in-house cohort of unaffected individuals. This cohort contains high-quality data of CNVs in autosomes (n=3,235 individuals) and sex chromosomes (n= 1,859 individuals). Overlap with one CNV in our in-house cohort reflects a population frequency of 0.0003 for autosomes and of 0.00054 for sex chromosomes. Inherited CNVs affect exactly the same locus as in their unaffected parent. As a single entity, these CNVs are therefore assumed not to be sufficient to give a severe phenotype. However, these CNVs might be of interest if present in more patients that share phenotypical characteristics, as they can act as a modifier in a multiple hit model, or as the second hit in a recessive condition. Special interest is given to shared anatomical malformations in addition to OA/TOF as these could hint at a specific sub-population in the cohort. Therefore, we also evaluated rare CNVs if they had overlapping loci in multiple patients with shared phenotypical characteristics.

Patient cohort

This study was approved by the institutional ethics committee of each participating centre, and was conducted in accordance with the principles of the Declaration of Helsinki. Patients with OA/TOF (isolated or non-isolated) were identified from the medical records. All patient records were reviewed by the treating physicians or geneticists of each participating centre. After retrieval of parental informed consent, blood was drawn from a total of 375 patients and their parents, comprising 239 patients from the Erasmus MC- Sophia, 28 from the Baylor College of Medicine, and 108 from a German multi-centre study regarding the genetic and environmental cause of OA/TOF ("The genetic risk for oesophageal atresia consortium [GREAT-consortium]"). Contact with patients and their families was established through their treating physicians or the German patient-support organization for patients with disorders of the oesophagus (KEKS e.V.; http://www.keks.org/). Previously, 68 patients of the Erasmus MC-Sophia and Baylor College of Medicine cohorts were evaluated in a study describing CNVs in VACTERL association.[16]

GTG-banded karyotyping and multiplex ligation-dependent probe amplification (MRC Holland, Amsterdam, the Netherlands) have ruled out the presence of large structural rearrangements

or aneuploidies in the majority of enrolled patients from the Erasmus MC- Sophia and the Baylor College of Medicine.[18] The German cohort was exclusively screened with micro-array. From 2009 onwards micro-array was the first tier-diagnostic tool used in the clinic, replacing GTG-banded karyotyping as a screening tool for large CNVs and aneuploidies. Patients born before 2009 were almost exclusively screened in a research setting and were enriched for the more complex non-isolated OA/TOF phenotypes. This causes over-representation for non-isolated OA/TOF patients in our cohort. Hence, over-representation of complex cases in this study is the result of selection bias and is not indicative for an actual deviance from the general isolated OA prevalence.

Literature search

We reviewed the literature (PubMed) in order to identify previously described disease-causing CNVs described in individuals with OA and / or TOF. Search terms included CNV, CNA (Copy Number Alteration), micro-array analysis, oesophageal atresia, tracheoesophageal fistula and VACTERL association.

DNA isolation

For the Dutch and American cohorts DNA for genomic analysis was extracted from peripheral blood, saliva and fibroblast cells with the Puregene DNA purification kit (Gentra Systems, USA) or the QIAamp DNA Blood Midi Kit (Qiagen, Inc., Hilden, Germany) or the Oragene DNA Kit. (DNA Genotek Inc., Ontario, Canada) DNA quality was evaluated with the Thermo Scientific NanoDrop 2000 (Thermo Fisher Scientific Inc., Waltham, USA), and dsDNA quantity with the Quant-iT™ PicoGreen ® dsDNA kit. (Invitrogen, Carlsbad, CA, USA). In the German Cohort DNA for genomic analysis was either extracted from peripheral blood or saliva, using the Chemagic Magnetic Separation Module I (ParkinElmer chemagen, Baesweiler, Germany) or the Oragene DNA Kit respectively.

Array-based molecular karyotyping

In the Dutch and American cohorts karyotyping was performed according to standard analytical methods. All DNA samples were tested for subtelomeric aberrations with Multiplex ligation-dependent probe amplification analysis, using the P036E1 and P070A2 Salsa telomere kits (MRC Holland, Amsterdam, and the Netherlands) as published previously. [1] High-resolution analyses for genomic imbalances present in the DNA of these patients were performed using single-nucleotide polymorphism (SNP) microarrays (Illumina Inc., San Diego, CA, USA and Affymetrix Inc. Santa Clara, CA, USA) and CGH oligonucleotide-based arrays (Agilent Inc., San Diego, CA, USA) using standard protocols. Micro-array analysis in the American and Dutch cohort was initially performed using three types of array chips (GeneChip Human Mapping 250K Nsp, 12-HumanCytoSNP DNA Analysis BeadChips v1-v2.1, and Agilent Human Genome CGH 105K and 244K), later these were replaced by chips with a higher resolution (Illumina Human 610-Quad Beadchip and Illumina HumanOmniExpress BeadChip, Illumina Infinium CytoSNP-850k BeadChip

and Agilent SurePrint G3 Human CGH 1M Oligo Microarray Kits G4411B/G4447). The majority of the Dutch cohort cases were processed on more than one array chip type. Material from all German patients enrolled in this study was screened using the HumanOmniExpress 12 v1.1 chip (Illumina Inc., San Diego, CA, USA).

Normalized output was generated using Feature Extraction software (v.9.1), alongside CGH Analytics software (v.3.3.28), with Affymetrix GTYPE (v.4.1, Affymetrix, Santa Clara, CA) or Illumina Genomestudio (v.2011.1, Illumina, San Diego, CA, USA), depending on chip type. CNVs in patient samples were visualized as log2-Ratios (Log2R) detected through comparison of patient probe intensity data with those of a virtual reference set of 270 HapMap samples of various ancestries (Illumina manifest file) or by comparison with sex matched controls of unaffected, unrelated individuals or reference DNA (Promega Corporation, Madison, WI). Since the German cohort was coherently processed on one chip type only, normalized output could be generated using the Genomestudio (v. 2011.1) genotyping module (v.1.9.4, Illumina, San Diego, CA, USA).

CNV filtering

For SNP-arrays, potential Copy Number Variants (CNVs) were estimated using a combination of allelic ratio (B-Allele Frequency, BAF) and SNP array copy number state (Log2R), utilizing Hidden-Markov-Model based algorithms. For this purpose Biodiscovery Nexus CN7.5 (Biodiscovery Inc, Hawthorne, CA, USA) was used for both cohorts. In the Nexus software, segmentation significance threshold was set at 5.0E-7 with a minimum of 5 probes per segment and a maximum probe spacing of 1,000 kb. The log2R-ratio thresholds were set at +0.18 (single copy gain), -0.18 (single copy loss), 0.4 (gain of two or more copies) and -1.1 (homozygous loss). The Homozygous Frequency/ Homozygous Value/ Heterozygous Imbalance Threshold were set at 0.95/0.8/0.4. The minimum LOH length was set at 100 kb and minimum SNP probe density, at 10 probes/Mb. Gender correction was used with a 3:1 sex chromosome gain threshold of 1.2 and a 4:1 sex chromosome gain threshold of 1.7. Log2R ratios of CGH-array results were determined with the ADM2 algorithm with filtering options of a minimum of 3 probes. (log2Ratio) >0.3. Each segment deviating from the normal situation was reviewed by visual inspection in Nexus CN 7.5. During the course of this study, the genome build switched from build hg18 to hg19. Not all arrays-chips could be re-processed or re-analyzed in the new genome build. For these array chips, the detected CNV regions were transformed using the UCSC lift-over tool and re-evaluated in the hg19 build for overlap with known polymorphisms (see below) and overlap with our in-house cohort (see below). Re-analysis of SNP-array in the new genome build, together with improved segmentation and waving correction algorithms resulted in the loss of several low confidence CNV calls.

Each observed CNV was compared to the frequency of that CNV at the particular locus in a modified version (i.e. excluding BAC arrays and small InDels) of the Database of Genomic Variants

(DGV, http://dgvbeta.tcag.ca/dgv/app/home?ref=NCBI36/hg19) incorporated in the Biodiscovery Nexus CN7.5 analysis package. Only those CNVs with a physical overlap with a DGV locus less than 70% and which occurred ≤5 times in the DGV were compared to an in-house control database of unrelated and unaffected parental samples and healthy control individuals. This cohort contains high-quality data of CNVs in autosomes (n=3235 individuals) and sex chromosomes (n= 1859 individuals). Overlap with one CNV in our in-house cohort reflects a population frequency of 0.0003 for autosomes and of 0.00054 for sex-chromosomes. We only considered CNVs which were either absent or present once for further analysis. All CNVs passing the filter criteria were evaluated manually by comparing the CNV to its presence in the most current versions of the database of genomic variation (DGV, http://dgv.tcag.ca/dgv/app/home), ISCA (http://dbsearch.clinicalgenome.org/search/), ClinGen (https://www.clinicalgenome.org/data-sharing/clinvar/) and DECIPHER. (http://decipher.sanger.ac.uk)

We searched for overlap in large CNV cohorts of control individuals published by Cooper *et al*.[2], Coe *et al*.[3] and Kaminsky et al.[4] We also evaluated the CNVs significantly different in these studies between patients and controls. As we used different array chips in this experiment, each with different marker spacing, distribution and content, we set a size CNV restriction threshold of minimum 30kb for single events, except when evaluating parental and patient SNP-arrays (trio analysis) . In trio analysis, all CNVs containing more than five probes were visually inspected using Nexus CN 7.5. We confirmed Copy Number (CN) state and inheritance pattern of private and rare CNVs if their size exceeded the restrictions thresholds and if the CNV contained genes or micro RNA's. All patient CNVs were visually re-evaluated using the Nexus software package e.g. we evaluated if CN state matched the allelic state of a CNV type.

Validation of microarray results

In the Dutch and American cohort, putative *de novo* events were confirmed via patient and parental copy number quantification using either additional SNP array, Fluorescence In Situ Hybridization (FISH) and/or Multiplex Amplicon Quantification (MAQ)(MultiplicoM N.V., Niel, Belgium) as well as real time quantitative PCR, the procedure also used for confirmation in the German cohort. At least two primer pairs for qPCR were designed within the putative CNVs using Primer Express Software v2.0 (Applied Biosystems, Foster City, CA, USA) for the Dutch cohort and Primer3web (v4.0.0.; http://primer3.ut.ee/)[5-7] for the German cohort. For the Dutch cohort the absence of SNPs in primer sequences was confirmed in dbSNP (build 135) and specificity of amplified region determined with the University of Santa Cruz (UCSC) *in silico* PCR and melting-curve analysis. Primer pairs with repeats in their resulting amplicon, according to CENSOR repeat masker[8] were excluded. Sequences for the German cohort were checked for SNPs and specificity using the UCSC human genome browser build 19 (http://genome.ucsc.edu).[9] QPCR experiments for the Dutch cohort were performed using a LightCycler 1.5 instrument and LightCycler FastStart DNA

Master SYBER Green I kit with C14ORF145 as a control locus. (Roche Molecular Diagnostics, Indianapolis, IN, USA) Locus exon 4 of the KIAA1279-gene was used as a control.[10] Cut-off values of <0.7 were used for deletions and values of >1.3 for duplications. For the German cohort QPCR experiments were performed on a LightCycler 480 II (Roche Molecular Diagnostics Indianapolis, IN, USA) in combination with Power SYBR Green Master Mix (Applied Biosystems, Foster City, USA). In order to normalize threshold cycle (Ct) values, each samples Ct value was compared to those of three reference genes (BCN1, CFTR, RNAseP subunit p38). Analysis was done as described by Scott *et al*.[11] In both cohorts each sample, including the no-template control (NTC) and control DNA, was run in triplicate.

For FISH confirmation, BAC-clones were selected from the UCSC genome browser and ordered from BACPAC Resources. After isolation of the BAC-DNA, the probes were amplified, labeled and used for FISH, according to standard protocols as described earlier.[12] The MAQ assay is a PCR-based-amplification method which uses 6 primer pairs on different loci for sample-internal copy number normalization, maximal 5 CNV specific primer pairs developed with the manufacturers' software package, and one type of FAM labeled primer specific to the sequence-tagged forward primers to amplify 20ng of dsDNA input. The DNA of 4 unaffected, unrelated individuals was used as sample-external copy number normalization. We amplified the DNA according to the manufacturer's protocol in a thermocycler with a heated lid, and analyzed the resulting FAM-labelled amplicons of patient, parental and controls using an automated sequencer (ABI 3730XL, Applied Biosystems, Foster City, CA, USA). This capillary electrophoresis step separates fragments on the basis of their length, these differences in amplicon length make multiplexing possible. The differences in fluorescence intensity reflect copy number state and were visualized in the MAQ-S analysis tool (Multiplicom Inc., Niel, Belgium) which compares amplicon size to the Genescan LIZ 500 size standard and normalizes copy number state to the internal amplicons and 4 external controls.

References

1. Van Opstal D, Boter M, de Jong D, et al. Rapid aneuploidy detection with multiplex ligation-dependent probe amplification: a prospective study of 4000 amniotic fluid samples. European journal of human genetics : EJHG 2009;**17**(1):112-21 doi: ejhg2008161 [pii]

10.1038/ejhg.2008.161 [doi][published Online First: Epub Date]|.
2. Cooper GM, Coe BP, Girirajan S, et al. A copy number variation morbidity map of developmental delay. Nat Genet 2011;**43**(9):838-46 doi: ng.909 [pii]

10.1038/ng.909 [doi][published Online First: Epub Date]|.
3. Coe BP, Witherspoon K, Rosenfeld JA-O, et al. Refining analyses of copy number variation identifies specific genes associated with developmental delay. (1546-1718 (Electronic))
4. Kaminsky EB, Kaul V, Paschall J, et al. An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities. Genetics in medicine : official journal of the American College of Medical Genetics 2011;**13**(9):777-84 doi: 10.1097/GIM.0b013e31822c79f9[published Online First: Epub Date]|.
5. Kent WJ. BLAT--the BLAST-like alignment tool. Genome research 2002;**12**(4):656-64 doi: 10.1101/gr.229202. Article published online before March 2002[published Online First: Epub Date]|.
6. Untergasser A, Cutcutache I, Koressaar T, et al. Primer3--new capabilities and interfaces. Nucleic acids research 2012;**40**(15):e115 doi: 10.1093/nar/gks596[published Online First: Epub Date]|.
7. Koressaar T, Remm M. Enhancements and modifications of primer design program Primer3. Bioinformatics (Oxford, England) 2007;**23**(10):1289-91 doi: 10.1093/bioinformatics/btm091[published Online First: Epub Date]|.
8. Kohany O, Gentles AJ, Hankus L, et al. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC bioinformatics 2006;**7**:474 doi: 10.1186/1471-2105-7-474[published Online First: Epub Date]|.
9. Speir ML, Zweig AS, Rosenbloom KR, et al. The UCSC Genome Browser database: 2016 update. Nucleic acids research 2015 doi: 10.1093/nar/gkv1275[published Online First: Epub Date]|.
10. Veenma D, Brosens E, de Jong E, et al. Copy number detection in discordant monozygotic twins of Congenital Diaphragmatic Hernia (CDH) and Esophageal Atresia (EA) cohorts. European journal of human genetics : EJHG 2012;**20**(3):298-304 doi: 10.1038/ejhg.2011.194[published Online First: Epub Date]|.
11. Scott DA, Klaassens M, Holder AM, et al. Genome-wide oligonucleotide-based array comparative genome hybridization analysis of non-isolated congenital diaphragmatic hernia. Human molecular genetics 2007;**16**(4):424-30 doi: 10.1093/hmg/ddl475[published Online First: Epub Date]|.
12. Eussen BH, van de Laar I, Douben H, et al. A familial inverted duplication 2q33-q34 identified and delineated by multiple cytogenetic techniques. European journal of medical genetics 2007;**50**(2):112-9 doi: S1769-7212(06)00106-6 [pii]

10.1016/j.ejmg.2006.10.006 [doi][published Online First: Epub Date]|.