NOVEL CANDIDATE GENES IN ESOPHAGEAL ATRESIA/TRACHEOESOPHAGEAL FISTULA IDENTIFIED BY EXOME SEQUENCING

Jiayao Wang[1,2,*], Priyanka R. Ahimaz[1,*], Somaye Hashemifar[1,2,*], Julie Khlevner[4], Joseph A. Picoraro[4], William Middlesworth[5], Mahmoud M. Elfiky[6], Jianwen Que[3], Yufeng Shen[2,#], and Wendy K. Chung[1,5,#]

[1]Division of Molecular Genetics, Department of Pediatrics, Columbia University Medical Center, New York, NY, USA
[2]Department of Systems Biology and Biomedical Informatics, Columbia University Medical Center, New York, NY, USA
[3]Department of Medicine and Columbia Center for Human Development, Columbia University, New York, NY, USA
[4]Division of Pediatric Gastroenterology, Hepatology and Nutrition, Department of Pediatrics, Columbia University Medical Center, New York, NY, USA
[5]Division of Pediatric Surgery, Department of Surgery, Columbia University Medical Center, New York, NY, USA
[6]Pediatric Surgery, Faculty of Medicine, Cairo University, Cairo, Egypt

* Equal contribution
[#]Corresponding authors: WKC: wkc15@cumc.columbia.edu; YS: ys2411@cumc.columbia.edu

Abstract

The various malformations of the aerodigestive tract collectively known as esophageal atresia/tracheoesophageal fistula (EA/TEF) constitute a rare group of birth defects of largely unknown etiology. Previous studies have identified a small number of rare genetic variants causing syndromes associated with EA/TEF. We performed a pilot exome sequencing study of 45 unrelated simplex trios (probands and parents) with EA/TEF. Thirteen had isolated and thirty-two had non-isolated EA/TEF; none had a family history of EA/TEF. We identified *de novo* variants in protein-coding regions, including 19 missense variants predicted to be deleterious (D-mis) and 3 likely-gene-disrupting variants (LGD). Consistent with previous studies of structural birth defects, there is a trend of increased burden of *de novo* D-mis in cases (1.57 fold increase over the background mutation rate), and the burden is greater in constrained genes (2.55 fold, p=0.003). There is a frameshift *de novo* variant in *EFTUD2*, a known EA/TEF risk gene involved in mRNA splicing. Strikingly, 15 out of 19 *de novo* D-mis variants are located in genes that are putative target genes of *EFTUD2* or *SOX2* (another known EA/TEF gene), much greater than expected by chance (3.34 fold, p-value=7.20e-5). We estimated that 33% of patients can be attributed to *de novo* deleterious variants in known and novel genes. We identified *APC2*, *AMER3, PCDH1*, *GTF3C1*, *POLR2B, RAB3GAP2,* and *ITSN1* as plausible candidate genes in the etiology of EA/TEF. We conclude that further genomic analysis to identify *de novo* variants will likely identify previously undescribed genetic causes of EA/TEF.

Keywords: Esophageal atresia, tracheoesophageal fistula, single nucleotide variant, exome sequencing, deleterious missense

INTRODUCTION

Esophageal atresia/tracheoesophageal fistula (EA/TEF) is a rare, complex congenital aerodigestive anomaly with an estimated incidence of 1 in 2500 to 1 in 4000 live births [1,2]. Almost half of infants born with this congenital anomaly have associated congenital malformations of other organ systems, most commonly cardiovascular, digestive [1], urogenital, and musculoskeletal [3]. These defects have been observed together as the VACTERL (vertebral defects, anal atresia, cardiac defects, tracheoesophageal fistula, renal anomalies, and limb abnormalities) association [4]. While there have been rare reports of variants in *FOXF1* and *ZIC3* in VACTERL-association patients [5], the molecular etiology for the majority of VACTERL cases remains unknown. Chromosome anomalies including aneuploidies and microdeletions are observed in 6-10% of non-isolated EA/TEF [3,5] patients. These anomalies include trisomy 13, 18, and 21, monosomy X [6], and several copy number variants (CNVs). Several monogenic causes of syndromes that include EA/TEF have also been elucidated and include mutations in *MYCN, SOX2, CHD7* and *MID1*. Monogenetic causes account for only about 5% of EA/TEF cases, and are mostly *de novo* (with the exception of mutations in recessive Fanconi anemia-related genes) [5-7].

SOX2 has been reported as an important gene for esophagus and anterior stomach development [8]. *SOX2* is involved in Wnt signaling by binding β-catenin, a central mediator of the Wnt pathway [9]. Deletion of the Wnt signaling downstream mediator β-catenin leads to lung agenesis, and the foregut fails to separate [10]. EFTUD2 is associated with esophageal atresia and other developmental disorders such as mandibulofacial dysostosis with microcephaly with heterozygous loss of function variants **[11,12 13]**. EFTUD2 is required for pre-mRNA splicing as component of the spliceosome. [14,15]

There have been few studies investigating the genetic causes of non-isolated EA/TEF, and it is still widely considered to have a multifactorial etiology. Small scale twin studies, however, have shown a higher concordance rate between monozygotic twins (67%) compared to dizygotic twins (42%), suggesting a genetic contribution [16,17]. Animal studies have identified genes in several developmental pathways associated with tracheoesophageal anomalies, among them sonic hedgehog pathway genes. Murine models with homozygous deficiencies of *SHH* and *GLI2* exhibit foregut anomalies including EA, TEF, and tracheoesophageal stenosis and hypoplasia [18]. Other developmental genes involved with foregut development in animal studies include transcription factors *Foxf1*, vitamin A effectors (*Rarα, Rarβ*) homeobox-containing transcription factors and their regulators (*Nkx2.1* [19], *Hoxc4*, *Pcsk5*), and developmental transcriptional regulators (*Tbx4*, *Sox2*) [3,20].

EA/TEF is identified prenatally in about 50% of cases. When the diagnosis is suspected (usually by sonographic findings of polyhydramnios and a small stomach), prognostic clinical information about associated birth defects is commonly sought. Definitive prognostic information is usually limited unless a chromosomal anomaly is identified. In an effort to identify novel genetic variants associated with EA/TEF, we studied 45 individuals with EA/TEF and their biological parents, none of whom had a family history of EA/TEF. We sought to identify novel genetic causes of EA/TEF using exome sequencing (WES). Our goal is to understand the genomic architecture of EA/TEF, and to better characterize the syndromes and conditions associated with EA/TEF. We designed this pilot study to assess whether genomic characterization of EA/TEF would provide more accurate prognostic information and help tailor therapy based on predicted phenotype. We plan to combine these data with that of other congenital malformations to provide a more comprehensive understanding of human development.

METHODS

Subject recruitment

Patients with isolated and non-isolated EA/TEF were recruited from two medical centers-Columbia University Medical Center (CUMC) in New York, USA and Cairo University General Hospital in Cairo, Egypt. Subjects eligible for the study included individuals diagnosed with known forms of EA/TEF and no family history of EA/TEF, based upon medical record review. All participants provided informed consent. The study was approved by the Columbia University institutional review board. Blood and/or saliva samples were obtained from the probands and both biological parents. A three-generation family history was taken at the time of enrollment and clinical data were extracted from the medical records and by patient and parental interview.

Exome sequencing

Exome sequencing was performed at Novogene Genome Sequencing Company (Chula Vista, CA). A total of 1.0 μg genomic DNA was used as input material. Sequencing libraries were generated using Agilent SureSelect Human All ExonV6 kit (Agilent Technologies, CA, USA) following manufacturer's recommendations. Briefly, fragmentation was carried out by hydrodynamic shearing system (Covaris, Massachusetts, USA) to generate 180-280 bp fragments. Remaining overhangs were converted into blunt ends via exonuclease/polymerase activities, and enzymes were removed. After adenylation of 3' ends of DNA fragments, adapter oligonucleotides were ligated. DNA fragments with ligated adapter molecules on both ends were selectively enriched in a PCR reaction. Captured libraries were enriched in a PCR reaction to add index tags to prepare for hybridization. Products were purified using AMPure XP system (Beckman Coulter, Beverly, USA) and quantified using the Agilent high sensitivity DNA assay on the Agilent Bioanalyzer 2100 system. The qualified libraries were sequenced on an Illumina HiSeq sequencer after pooling according to effective concentration and expected data volume. Read length were paired-end 150 bp.

Bioinformatics analysis and calling of *de novo* variants

We used GATK-recommended best practices for calling single nucleotide variants (SNVs) and short insertions and deletions (indels) from exome sequencing data. Specifically, we used BWA-mem [21] to align reads to human reference genome (GRCh37), Picard Tools to mark PCR duplicates, and GATK [22] haplotypeCaller for calling variants jointly from all sequenced samples, and GATK variant quality score recalibration (VQSR) to recalibrate variant quality. We applied multiple heuristic filtering rules to remove potential technical artifacts as previously described [23,24]. Specifically, we only retained variants that met all the following criteria: GQ >=30, FS <= 25, QD >=2(SNV), QD >=1 (INDEL), ReadPosRankSum >= -3 (INDEL), read depth on alt allele >=5, alt allele depth to total depth >= 0.1, VQSRSNP <= 99.80, VQSRINDEL <= 99.70 and mappability (based on 200 insert length) = 1.

To call *de novo* variants, we applied a previously published procedure [23,24] and used IGV [25] to visualize candidate *de novo* variants and remove potential artifacts. All non-synonymous *de novo* variants were sanger confirmed. In addition, we used PLINK to infer population structure and kinship. We used xHMM[26] to infer large CNVs to ruled out patients who potentially get EA/TEF due to chromosomal anomalies.

Annotation and in silico prediction

We used ANNOVAR [27] to annotate variants and aggregate population frequency (Exome Aggregation Consortium (ExAC) and Genome Aggregation Database (gnomAD) [28], protein-coding consequence, and multiple in silico predictions on genetic variants, including CADD [29] and REVEL [30].

Putative targets of *EFTUD2* or *SOX2*

We obtained putative targets of *EFTUD2* based on RNA binding protein (RBP) binding sites profiled by eCLIP in a HepG2 cell line from ENCODE [31] and processed using a recently published pipeline [32]. We selected the genes for which the peak count is equal to or greater than 2. We obtained target genes of transcription factor *SOX2* based on ChIP data from glandular mouse stomach [33] curated by ChEA [34].

Statistical analysis

For *de novo* variants, we determined the overall burden of four variant types including synonymous, likely gene disrupting (LGD, i.e. stopgain, frameshift, and splice site), missense and deleterious missense (D-mis, defined by REVEL ≥ 0.5 or CADD Phred score ≥ 25) in all genes and constrained genes (defined by ExAC [28] pLI ≥ 0.5). We used a less stringent pLI threshold for defining constrained genes, because it captures more known haploinsufficient genes [35]. We obtained estimated background mutation rate in previous publications calibrated for exome sequencing data [36]. The expected number of variants in different gene sets were calculated by summing up the background mutation rate of the specific variant class in the gene-set multiplied by twice the number of cases. We then test the burden of *de novo* variants in a gene set by a Poisson test with the baseline expectation as the mean under the null model. To estimate the proportion of cases that can be attributed to *de novo* deleterious variants, the difference between the observed number and expected number of *de novo* deleterious variants is divided by the number of cases [37].

RESULTS

Exome Sequencing data

A total of 45 individuals with EA/TEF were enrolled into the study. Probands were between the ages of 1.5 years and 55.7 years with an average of 10.2 years old (Table 1). Thirteen probands had isolated EA/TEF and thirty-two probands had neurodevelopmental delay and/or at least one additional congenital defect and were classified as non-isolated. Fourteen of the probands had congenital heart defects, eight had neurodevelopmental delay, four had gastrointestinal defects, twelve had genitourinary defects (non-renal), eight had skeletal defects, two had craniofacial defects and two had other defects. The majority of probands were of European ancestry (60%), and the remaining were of African-American (15%), Egyptian (15%) and Asian (10%) ancestry. None of the 45 probands reported a family history of EA/TEF.

Overall burden of *de novo* variants

We identified 57 *de novo* variants in 45 probands (Supplemental Table 1). We compared overall burden of *de novo* variants in 45 cases to expectations from a background mutation model [36]. We classified protein-coding variants into four groups: synonymous, missense, deleterious missense (D-mis), and likely gene disruptive (LGD). Overall the frequency of synonymous variants in cases is close to expectation from background mutation rate (p-value=0.68, enrichment rate=1.1x). There is a trend of enrichment of missense variants (p = 0.12, enrichment rate =1.3x) and D-mis variants (p = 0.06, enrichment rate =1.6x) in cases compared to expectation (Table 2).

Consistent with previous studies of other types of birth defects [24,38,39], the enrichment of D-mis variants is more pronounced (p-value = 0.003, enrichment rate=2.6x) in constrained genes that are intolerant of loss of function variants (ExAC pLI≥0.5) (Table 2).

Most of genes with deleterious *de novo* variants are putative targets of *EFTUD2* or *SOX2*

One patient has a *de novo* frameshift deletion (c.2314delC, p.Q772fs) in *EFTUD2* (elongation factor Tu GTP binding domain containing 2). The phenotype of the patient includes EA/TEF, bilateral clubfoot, hydrocele, atrial septal defect, and pylectasis which overlaps with features of Guion-Almeida type of mandibulofacial dysostosis caused by heterozygous *EFTUD2* variants. **[13]** *De novo* variants in *EFTUD2* are known to be associated with EA [11,12]. *EFTUD2* encodes a component of the splicesome complex that regulates mRNA splicing, a master regulator that potentially regulates the expression of thousands of genes. We hypothesized that genes regulated by *EFTUD2* and other master regulators relevant to EA/TEF (such as *SOX2* [8]) are more likely to be EA/TEF risk genes and therefore enriched with *de novo* variants. To test this, we obtained putative targets of *EFTUD2* based on eCLIP data in a HepG2 cell line from ENCODE [31] and targets of *SOX2* based on ChIP-seq data in mouse stomach [33]. There are 1629 and 4463 targets of *SOX2* and *EFTUD2*, respectively; and the union of the targets is 5454. Among 19 genes with D-mis *de novo* variants, 15 are targets of *SOX2* or *EFTUD2*, much larger than expected by background (enrichment rate=3.34, p-value=6.6e-05). Overall, the burden indicates that 33% of EA/TEF patients are attributable to deleterious *de novo* variants in genes that are *SOX2* or *EFUD2* targets.

Table 3 summarizes the associated clinical features and variants in candidate genes prioritized by intolerance to loss of function variants and biological pathways implicated in developmental disorders. Seven genes, *ADD1*, *APC2*, *GLS*, *SMAD6*, *RAB3GAP2*, *PTPN14,* and *EFTUD2* are OMIM genes and are associated with Mendelian diseases (Table 3). *ITSN1* was recently discovered as a risk gene for autism spectrum disorder [40]. The *ITSN1* variant carrier was only 18 months at the time of enrollment which is too young to make the diagnosis of autism.

DISCUSSION

In this pilot study, we report exome sequencing results on 45 proband-parent trios with isolated or non-isolated EA/TEF with no family history of EA/TEF. We identified 22 LGD or D-mis *de novo* variants. Consistent with previous studies of structural birth defects or developmental disorders, genes that are constrained are enriched with deleterious variants, likely due to an historical reduction of reproductive fitness by such predicted deleterious variants. The majority of the genes with deleterious *de novo* variants are putative targets of *SOX2* or *EFTUD2*, two master regulators that are known to cause EA/TEF through haploinsufficiency and may provide a biological mechanism for the etiology of some EA/TEF. Figure 1 shows genes with LGD or D-mis *de novo* variants and their relationships with *EFTUD2* and *SOX2*. We did not identify any *de novo* variants in *SOX2* gene in our small cohort. Given the overall high enrichment rate of 3.34, we expect that more than half of target genes of *SOX2* or *EFTUD2* with *de novo* predicted pathogenic variants are candidate EA/TEF risk genes [37,41].

Three genes, *ADD1*, *GLS,* and *RAB3GAP2*, are putative targets of both *EFTUD2* and *SOX2*.[31,33]Notably, *ITSN1*, *AP1G2*, *TECPR1*, and *RAB3GAP2* are involved in membrane trafficking pathway or autophagy.[42-45] *KLHL17*, *ADD1*, *CELSR2*, *PCDH1*, and *ITSN1* are involved in cytoskeleton or cell adhesion[42,46,47]. *AMER3 and APC2* are both key regulators in Wnt signaling, a process known to be implicated in EA/TEF and other birth defects[48]. A few other genes, *SMAD6*, *PTPN14*, and *PIK3C2G*, are involved in signaling pathways that are critical during development.[46,49,50]

Our current analysis is limited by the source of ChIP-seq of *SOX2* from stomach [33] and eCLIP of *EFTUD2* from a liver cancer cell line [31]. The availability of data from relevant tissues, e.g. ChIP-seq of *SOX2* and eCLIP-seq of *EFTUD2* in developing foregut, will enable more precise analysis of *de novo* and rare variants. Additionally, gene expression data, especially single cell sequencing data, of developing esophagus and trachea, will also allow us to refine the analysis and improve the ability to identify the most relevant EA/TEF genes.

Finally, it will be important to increase the sample size of future genomic studies to more precisely estimate the contribution of *de novo* variants to EA/TEF, and to identify novel risk genes with high confidence and relate the genetic factors to clinical outcomes.

ACKNOWLEDGEMENTS

We would like to acknowledge the patients and their families who participated in the study and are grateful for their tremendous contribution. We thank Steve Wyles and Sue Paul from EA for Adults, as well as the TOFS UK, OARA, Bridging the Gap of EA/TEF and the Canadian EA network organizations for publicizing the study and assisting with recruitment. We are also appreciative for the technical assistance provided by Patricia Lanzano, Jiangyuan Hu, Liyong Deng, and Charles LeDuc from Columbia University and the study team at Cairo University General Hospital. We thank Na Zhu for help with the calculation of background mutation rate. Funding support provided by P01HD093363 (JW, YS, and WKC) and R01GM120609 (YS).

References:

1. Pinheiro PFM, e Silva ACS, Pereira RM: Current knowledge on esophageal atresia. *World journal of gastroenterology: WJG* 2012; **18:** 3662.

2. Krishnan U, Mousa H, Dall'Oglio L *et al*: ESPGHAN-NASPGHAN guidelines for the evaluation and treatment of gastrointestinal and nutritional complications in children with esophageal atresia-tracheoesophageal fistula. *Journal of pediatric gastroenterology and nutrition* 2016; **63:** 550-570.

3. Stoll C, Alembik Y, Dott B, Roth M-P: Associated malformations in patients with esophageal atresia. *European journal of medical genetics* 2009; **52:** 287-290.

4. Shaw-Smith C: Genetic factors in esophageal atresia, tracheo-esophageal fistula and the VACTERL association: roles for FOXF1 and the 16q24. 1 FOX transcription factor gene cluster, and review of the literature. *European journal of medical genetics* 2010; **53:** 6-13.

5. Geneviève D, de Pontual L, Amiel J, Lyonnet S: Genetic factors in isolated and syndromic esophageal atresia. *Journal of pediatric gastroenterology and nutrition* 2011; **52:** S6-S8.

6. Felix JF, Tibboel D, de Klein A: Chromosomal anomalies in the aetiology of oesophageal atresia and tracheo-oesophageal fistula. *European journal of medical genetics* 2007; **50:** 163-175.

7. Murphy AJ, Li Y, Pietsch JB, Chiang C, Lovvorn HN: Mutational analysis of NOG in esophageal atresia and tracheoesophageal fistula patients. *Pediatric surgery international* 2012; **28:** 335-340.

8. Que J, Okubo T, Goldenring JR *et al*: Multiple dose-dependent roles for Sox2 in the patterning and differentiation of anterior foregut endoderm. *Development* 2007; **134:** 2521-2531.

9. Kormish JD, Sinner D, Zorn AM: Interactions between SOX factors and Wnt/β-catenin signaling in development and disease. *Developmental Dynamics* 2010; **239:** 56-68.

10. Morrisey EE, Hogan BL: Preparing for the first breath: genetic and cellular mechanisms in lung development. *Developmental cell* 2010; **18:** 8-23.

11. Gordon CT, Petit F, Oufadem M *et al*: EFTUD2 haploinsufficiency leads to syndromic oesophageal atresia. *Journal of medical genetics* 2012; **49:** 737-746.

12. Voigt C, Mégarbané A, Neveling K *et al*: Oto-facial syndrome and esophageal atresia, intellectual disability and zygomatic anomalies-expanding the phenotypes associated with EFTUD2 mutations. *Orphanet journal of rare diseases* 2013; **8:** 110.

13. Lines MA, Huang L, Schwartzentruber J *et al*: Haploinsufficiency of a spliceosomal GTPase encoded by EFTUD2 causes mandibulofacial dysostosis with microcephaly. *Am J Hum Genet* 2012; **90:** 369-377.

14. Zhang X, Yan C, Hang J, Finci LI, Lei J, Shi Y: An Atomic Structure of the Human Spliceosome. *Cell* 2017; **169:** 918-929 e914.

15. Bertram K, Agafonov DE, Dybkov O *et al*: Cryo-EM Structure of a Pre-catalytic Human Spliceosome Primed for Activation. *Cell* 2017; **170:** 701-713 e711.

16. Schulz AC, Bartels E, Stressig R *et al*: Nine new twin pairs with esophageal atresia: a review of the literature and performance of a twin study of the disorder. *Birth Defects Research Part A: Clinical and Molecular Teratology* 2012; **94:** 182-186.

17. Maroszyńska I, Fortecka-Piestrzeniewicz K, Niedźwiecka M, Żarkowska-Szaniawska A: Isolated esophageal atresia in both premature twins. *Pediatria Polska* 2015; **90:** 91-93.

18. Shaw-Smith C: Oesophageal atresia, tracheo-oesophageal fistula, and the VACTERL association: review of genetics and epidemiology. *Journal of medical genetics* 2006; **43:** 545-554.

19. : Development and stem cells of the esophagus. *Proceedings of the Seminars in cell & developmental biology*. Elsevier, pp. 25-35.

20. Al-Salem AH, Kothari M, Oquaish M, Khogeer S, Desouky MS: Morbidity and mortality in esophageal atresia and tracheoesophageal fistula: a 20-year review. *Annals of Pediatric Surgery* 2013; **9:** 93-98.

21. Li H: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:13033997* 2013.

22. DePristo MA, Banks E, Poplin R *et al*: A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* 2011; **43:** 491.

23. Homsy J, Zaidi S, Shen Y *et al*: De novo mutations in congenital heart disease with neurodevelopmental and other congenital anomalies. *Science (New York, NY)* 2015; **350:** 1262-1266.

24. Qi H, Yu L, Zhou X *et al*: De novo variants in congenital diaphragmatic hernia identify MYRF as a new syndrome and reveal genetic overlaps with other developmental disorders. *PLoS Genet* 2018; **14:** e1007822.

25. Thorvaldsdóttir H, Robinson JT, Mesirov JP: Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics* 2013; **14:** 178-192.

26. Fromer M, Moran JL, Chambert K *et al*: Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet* 2012; **91:** 597-607.

27. Wang K, Li M, Hakonarson H: ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* 2010; **38:** e164-e164.

28. Lek M, Karczewski KJ, Minikel EV *et al*: Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016; **536:** 285.

29. Kircher M, Witten DM, Jain P, O'roak BJ, Cooper GM, Shendure J: A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* 2014; **46:** 310.

30. Ioannidis NM, Rothstein JH, Pejaver V *et al*: REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *The American Journal of Human Genetics* 2016; **99:** 877-885.

31. Van Nostrand EL, Freese P, Pratt GA *et al*: A Large-Scale Binding and Functional Map of Human RNA Binding Proteins. *bioRxiv* 2018.

32. Feng H, Bao S, Rahman MA *et al*: Modeling RNA-Binding Protein Specificity In Vivo by Precisely Registering Protein-RNA Crosslink Sites. *Molecular cell* 2019; **74:** 1189-1204 e1186.

33. Sarkar A, Huebner AJ, Sulahian R *et al*: Sox2 Suppresses Gastric Tumorigenesis in Mice. *Cell reports* 2016; **16:** 1929-1941.

34. Lachmann A, Xu H, Krishnan J, Berger SI, Mazloom AR, Ma'ayan A: ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* 2010; **26:** 2438-2444.

35. Han X, Chen S, Flynn E, Wu S, Wintner D, Shen Y: Distinct epigenomic patterns are associated with haploinsufficiency and predict risk genes of developmental disorders. *Nature communications* 2018; **9:** 2138.

36. Samocha KE, Robinson EB, Sanders SJ *et al*: A framework for the interpretation of de novo mutation in human disease. *Nature genetics* 2014; **46:** 944.

37. Walsh R, Mazzarotto F, Whiffin N *et al*: Quantitative approaches to variant classification increase the yield and precision of genetic testing in Mendelian diseases: the case of hypertrophic cardiomyopathy. *Genome medicine* 2019; **11:** 5.

38. Jin SC, Homsy J, Zaidi S *et al*: Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. *Nat Genet* 2017.

39. Deciphering Developmental Disorders S: Prevalence and architecture of de novo mutations in developmental disorders. *Nature* 2017; **542:** 433-438.

40. Feliciano P, Zhou X, Astrovskaya I *et al*: Exome sequencing of 457 autism families recruited online provides evidence for novel ASD genes. *bioRxiv* 2019**:** 516625.

41. He X, Sanders SJ, Liu L *et al*: Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes. *PLoS genetics* 2013; **9:** e1003671.

42. Hussain NK, Jenna S, Glogauer M *et al*: Endocytic protein intersectin-l regulates actin assembly via Cdc42 and N-WASP. *Nature cell biology* 2001; **3:** 927-932.

43. Takatsu H, Sakurai M, Shin HW, Murakami K, Nakayama K: Identification and characterization of novel clathrin adaptor-related proteins. *J Biol Chem* 1998; **273:** 24693-24700.

44. Ogawa M, Yoshikawa Y, Kobayashi T *et al*: A Tecpr1-dependent selective autophagy pathway targets bacterial pathogens. *Cell Host Microbe* 2011; **9:** 376-389.

45. Spang N, Feldmann A, Huesmann H *et al*: RAB3GAP1 and RAB3GAP2 modulate basal and rapamycin-induced autophagy. *Autophagy* 2014; **10:** 2297-2309.

46. Gaudet P, Livstone MS, Lewis SE, Thomas PD: Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Brief Bioinform* 2011; **12:** 449-462.

47. Mische SM, Mooseker MS, Morrow JS: Erythrocyte adducin: a calmodulin-regulated actin-bundling protein that stimulates spectrin-actin binding. *J Cell Biol* 1987; **105:** 2837-2845.

48. Brauburger K, Akyildiz S, Ruppert JG *et al*: Adenomatous polyposis coli (APC) membrane recruitment 3, a member of the APC membrane recruitment family of

APC-binding proteins, is a positive regulator of Wnt-beta-catenin signalling. *FEBS J* 2014; **281:** 787-801.

49. Zhang X, Zhang J, Bauer A *et al*: Fine-tuning BMP7 signalling in adipogenesis by UBE2O/E2-230K-mediated monoubiquitination of SMAD6. *The EMBO journal* 2013; **32:** 996-1007.

50. Au AC, Hernandez PA, Lieber E *et al*: Protein tyrosine phosphatase PTPN14 is a regulator of lymphatic function and choanal development in humans. *Am J Hum Genet* 2010; **87:** 436-444.

Table 1. Patient characteristics of 45 patients with esophageal atresia.

	N=45
Mean age (range)	10.2 yrs (1.5 yrs-55.7 yrs)
Sex	
Male	25 (56%)
Female	20 (44%)
Type of EA	
Type A	3 (7%)
Type C	11 (24%)
Type D	1 (2%)
Type H (TEF only)	3 (7%)
Unknown	27 (60%)
Failure to Thrive	8 (18%)
Associated Anomalies	13 (65%)
Non-isolated cases	32 (71%)
Developmental Delay	8 (18%)
Other congenital defects	28 (64%)

Table 2. Overall burden of *de novo* heterozygous variants. Exp_Rate and Obs_Rate are respectively the expected and observed fraction of genes with a specific type of *de novo* mutation. Exp_Num and Obs_Num are the expected and observed number of genes with a specific type of *de novo* mutation, respectively. Constrained genes are defined by ExAC_pLI > 0.5. LGD: likely gene disrupting, including frameshift, stop-gain and variants at canonical splice site. D-mis: predicted deleterious missense variants.

Gene Sets	**Variant Class**	**Obs_Num**	**Obs_Rate**	**Exp_Num**	**Exp_Rate**	**Enrichment**	**P-value**
All Genes	Synonymous	15	0.333	13.7	0.304	1.1	0.68
	Missense	39	0.867	30.2	0.671	1.29	0.12
	D-mis	19	0.422	12.1	0.269	**1.57**	**0.06**
	LGD	3	0.066	4.04	0.089	0.743	0.81
Constrained Genes	Synonymous	8	0.178	4.98	0.111	1.61	0.17
	Missense	16	0.356	11.06	0.246	1.45	0.13
	D-mis	12	0.267	4.71	0.105	**2.55**	**0.003**
SOX2 or *EFTUD2* targets	Synonymous	8	0.178	4.84	0.108	1.65	0.16
	Missense	19	0.422	10.76	0.24	1.77	**2.2e-16**
	D-mis	15	0.333	4.49	0.099	**3.34**	**6.6e-05**

Table 3. *De novo* heterozygous variants in candidate genes. Eleven of these genes (*CELSR2, PCDH1, APC2, GLS, GTF3C1, ITSN1, MAP4K3, ADD1, POLR2B, PTPN14, RAB3GAP2*) are constrained genes with a D-mis variant. Three genes *AP1G2, KLHL17, SMAD6,* and *TECPR1* are non-constrained genes with D-mis variants. *EFTUD2*, *PIK3C2G* and *AMER3* have LGD variants and *EFTUD2* is a known candidate gene for EA. LGD: likely gene disrupting, including frameshift, stop-gain and variants at canonical splice site. D-mis: predicted deleterious mis-sense variants. EA/TEF: Esophageal atresia/tracheoesophageal fistula. AR: Autosomal Recessive; AD: Autosomal Dominant.

Study ID	Chr	Pos	Gene (OMIM#)	pLI	Coding Sequence Change	Type	REVEL	CADD	EA/TEF	Additional anomalies	OMIM condition (Inheritance) (OMIM #)
230	4	2877643	***ADD1*** (102680)	0.61	NM_001119:exon2: c.A1G:p.M1V	D-mis	0.53	25.1	EA+TEF	Extra wedge-shaped vertebrae, extra ribs, horseshoe kidney, bilateral radial hypoplasia with associated thumb and wrist anomaly	
48	2	131521881	***AMER3*** *(NA)*	0.00	NM_001105193:exon2:c.C2236T:p.R746X	LGD	N/A	35	EA+TEF	Atrial septal defect, bilateral clubfoot, hydrocele, renal pyelectasis	
15	14	24035494	***AP1G2*** *(*603534*)*	0.00	NM_001282475:exon3:c.G77A:p.R26H-AP1G2	D-mis	0.39	25.9	EA+TEF	Atrial septal defect, patent ductus arteriosis, short stature, small kidneys, hiatal hernia	
101	19	1456100	***APC2*** *(*612034*)*	0.99	NM_005883:exon7: c.T665C:p.I222T	D-mis	0.65	26.4	EA+TEF	Pierre Robin sequence, solitary kidney, cleft palate	Sotos syndrome 3 (AR) (617169); Cortical dysplasia, complex, with other brain malformations 10 (AR)(618677)
4	1	109795559	***CELSR2*** (604265)	0.99	NM_001408:exon1: c.A2858G:p.N953S	D-mis	0.57	24.1	EA+TEF	Duodenal atresia, Wolf-Parkinson White syndrome	
48	17	42930931	***EFTUD2*** (603892)	0.99	NM_001142605:exon23: c.2314delC: p.Q772fs	LGD	N/A	N/A	EA+TEF	Atrial septal defect, bilateral clubfoot, hydrocele, renal pyelectasis	Mandibulofacial dysostosis, Guion-Almeida type (AD)(610536)
48	2	191827642	***GLS*** *(*138280*)*	0.99	NM_014905:exon18:c.C1940T:p.T647I	D-mis	0.169	27.1	EA+TEF	Atrial septal defect, bilateral clubfoot, hydrocele, renal pyelectasis	Epileptic encephalopathy, early infantile, 71 (AR)(618328); Infantile cataract, skin abnormalities, glutamate excess, and impaired intellectual development (AD) (618339); Global developmental delay, progressive ataxia, and elevated glutamine (AR) (618412)

17	16	27473692	***GTF3C1*** (603246)	0.99	NM_001520:exon36:c.C6040A:p.P2014T	D-mis	0.60	26.4	EA+TEF	Multiple congenital hemangiomas	
2-8	21	35254586	***ITSN1*** (602442)	0.99	NM_001331010:exon34:c.C4366T:p.R1456C	D-mis	0.22	33	EA	Tetralogy of Fallot	
267	1	899892	***KLHL17*** *(*NA*)*	0.00	NM_198317:exon11:c.C1682A:p.A561E	D-mis	0.81	31	EA+TEF	Heart defect, kyphosis, tracheomalacia, right leg hemihypertrophy	
95	5	141248684	***PCDH11*** (603626)	0.87	NM_001278613:exon2:c.A401G:p.E134G	D-mis	0.66	26.7	EA+TEF	None	
125	12	18439865	***PIK3C2G*** *(*609001*)*	0.00	.N/A	LGD	.N/A	22.5	EA+TEF	Coarctation of aorta, total anomalous pulmonary venous return, congenital stricture in distal esophagus, hypospadias	
275	4	57883376	***POLR2B*** (180661)	0.99	NM_001303268:exon14:c.C1898T:p.A633V	D-mis	0.92	34	EA+TEF	Left multicystic dysplastic kidney, aortic plexus	
248	1	214557279	***PTPN14*** *(*603155*)*	0.99	NM_005401:exon13:c.G1919A:p.R640H	D-mis	0.119	25	EA+TEF	Dilated cardiomyopathy (not congenital-diagnosed in 30s)	Choanal atresia and lymphedema (AR)(613611)
2-6	1	220364518	***RAB3GAP2*** *(609275)*	0.99	NM_012414:exon14:c.G1379A:p.R460Q	D-mis	0.32	35	EA+TEF	None	Martsolf syndrome (AR)(212720) , Warburg micro syndrome 2 (AR) (614225)
275	15	67073475	***SMAD6*** (602931*)*	0.00	NM_005585:exon4:c.G1093A:p.G365S	D-mis	0.85	32	EA+TEF	Left multicystic dyplastic kidney, aortic plexus	Aortic valve disease 2 (AD)(614823)
15	7	97854186	***TECPR1*** *(*614781*)*	0.00	NM_015395:exon19:c.G2617A:p.D873N	D-mis	0.74	35	EA+TEF	Atrial septal defect, patent ductus arteriosus, short stature, small kidneys, hiatal hernia	

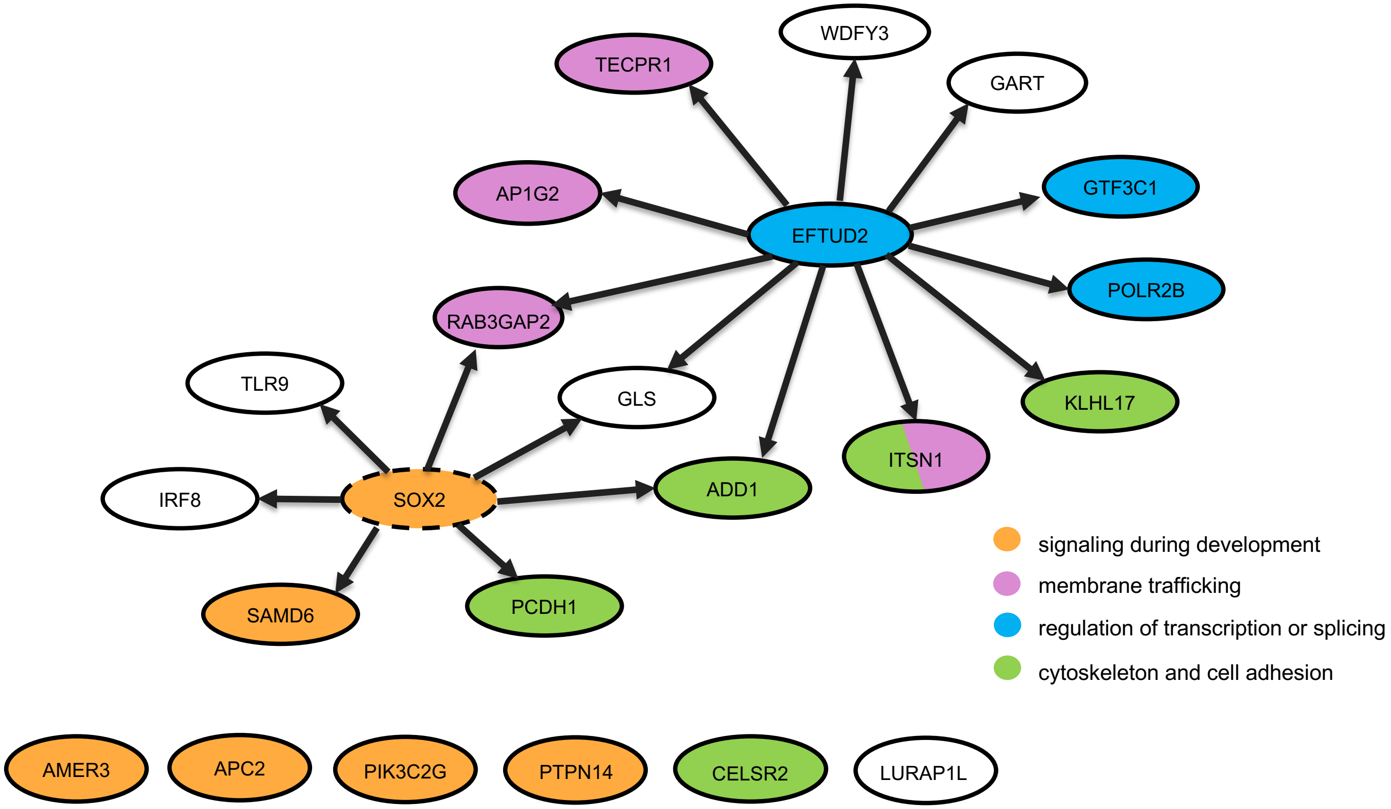

Figure 1. Genes with LGD or D-mis *de novo* variants and their relationship with *EFTUD2* and *SOX2*. Each gene is represented by a circle. Arrows indicate putative TF-target or RBP-target relationships. We did not observe *de novo* mutations in *SOX2* (dashed circle) in our cohort. Genes are colored by biological pathways. Only the pathways with at least three genes with LGD or D-mis variants are shown.