Whole genome sequencing of tracheoesophageal trios identifies novel candidate risk genes in endocytosis and membrane vesicle trafficking

Draft of EA/TEF

**Abstract**

Esophageal atresia/tracheoesophageal fistula (EA/TEF) is a rare birth defect caused by incomplete or dysfunctional development of the foregut. Previous studies indicate that rare or de novo variants have significant contribution to EA/TEF genetic risk, but the vast majority of individuals with EA/TEF do not have known genetic causes. To identify novel genetic etiologies of EA/TEF, we performed a whole genome sequencing study of 185 unrelated simplex trios (probands and parents) with EA/TEF, where 59 had isolated and 126 had complex EA/TEF with additional congenital anomalies or neurodevelopmental disorders. There is an enriched number of damage variants in 126 non-isolated cases and the burden is higher in constrained genes. We identified several autophagy related pathways as potential etiology of EA/TEF, indicating their importance in epithelia cell shape regulation during foregut development. This research can have general implications to the mechanisms of rare birth defects.

**Introduction**

Esophageal atresia (EA) is a congenital abnormality of the esophagus, with which a tracheoesophageal fistula (TEF) co-occurs in 70-90% cases1. The overall worldwide prevalence of EA/TEF is 2.4 per 100,000 births2. Approximately 55% of people born with EA/TEF have associated birth defects or other anomalies2, including cardiovascular, musculoskeletal, urinary, gastrointestinal, and central nervous system anomalies3. The genetic causes of EA/TEF include chromosome anomalies or mutations of genes involved in critical developmental processes that prove to be dosage sensitive4. Several disease risk genes include SOX2, MYCN, CHD7, FANCB and members of FOX transcription factor gene has been reported to be causes of syndromes that include EA/TEF4.

Several mouse models have been studied to assess the involvement of certain genes in the development of EA. It has been reported that precise regulations of Nkx2.1, SOX2, the bone morphogenetic protein 4 (BMP4) as well as WNT signaling pathways are required for separation of the esophagus and trachea2,5–7, without which failure of tracheoesophageal separation occurs. Moreover, EFTUD2 haploinsufficiency has been reported to lead to syndromic EA, emphasizing the necessity of mRNA maturation through the spliceosome complex for global growth and within specific regions of the embryo during development8.

Despite the existing studies of genetics in several syndromes that include EA/TEF and mouse models, the research of multifactorial etiology of EA/TEF is still limited. To identify novel genetic etiologies of EA/TEF, we performed whole genome sequencing (WGS) of 185 individuals with EA/TEF and their biological parents, none of whom had a family history of EA/TEF. We confirmed an overall enrichment of damaging de novo variants in all cases, and identified several autophagy related biological pathways as well as several new candidate EA/TEF genes.

**Methods**

**Patients Recruitment**

We recruited and performed whole genome sequencing on 185 probands diagnosed with EA/TEF and their parents. Patients with only EA/TEF were classified as isolated cases (59 in total) and patients with other type of congenital abnormalities were classified as complex cases (126 in total, Table S1).

**WGS analysis**

We identified the de novo coding variants using previously published procedures with heuristic filters9–11 {Qi 2018; Richter 2020; Wang 2021} augmented with in silico confirmation by DeepVariant12 {DeepVariant} (Table S2). We used ANNOVAR and VEP to annotate variants with population allele frequency13,14 (gnomAD and ExAC), protein-coding consequences, and predicted damaging scores of missense variants. Variants were classified as LGD (Likely Gene Disruption, including frameshift, stop gained/lost, start lost, splice acceptor/donor and protein altering variants), missense or synonymous. In frame deletions/insertions and splice region variants were excluded in following analysis. Variants in olfactory receptor genes, HLA genes or MUC gene family were filtered out of following analysis.

**EFTUD2 putative targets**

We calculated the putative binding score of EFTUD2 to all genes in all cell types of mouse fetal foregut based on eCLIP data and scRNA expression data15 {Han 2020, POLARIS}. The cell type expression value was defined as average of normalized counts in all corresponding single cells. For each gene, we calculated the binding scores of all possible binding windows (1kb) in both exon and intron region and average it across cell types. The final gene level score was defined as the maximum score of all binding windows in that gene. Genes with score≥0.7 were considered as EFTUD2 putative targets (1596 in total).

**Hedgehog/Gli signaling pathways**

Obtained from Aaron, method?

**Burden test**

We divided the patient into 2 cohorts based on their phenotypes (isolated and complex) and did burden test on both cohorts and the aggregated cohort. For each cohort, we divided de novo coding variants into four type: synonymous, LGD, missense and Damage (defined as combination of LGD and missense variants). For each type, we calculated the expected number of variants based on a background mutation rate model16,17 {Samocha 2014; Ware 2015}. We used single-sided Poisson test to test whether the number of observed variants is significantly higher than expected. We did this test on de novo coding variants of all genes, genes intolerant of loss of function variants (“constrained genes” based on gnomAD pLI≥0.5), non-constrained genes and EFTUD2 putative target genes (EFTUD2 binding score≥0.7) respectively.

**Pathway enrichment analysis**

To identify the pathways that damage de novo variants are involved in, we applied a pathway enrichment analysis on the GO and HPO pathways from GSEA18,19 {gsea} database. We removed the pathways with less than 2 expected damage variants (defined by combination of LGD and missense variants) in the complex cases cohort based on background mutation rate model16,17 {Samocha 2014; Ware 2015}. There is a total of 907 pathways used for analysis. The mutation rates were rescaled genome widely according to the fold ratio of synonymous variants in all cases as an estimated correction of sequencing depth. We did single-sided Poisson test of observed damage variants versus expectation on all pathways. To control the Family Wise Error Rate (FWER), we simulated 20000 times under null hypothesis based on background mutation rate. The FWER of each pathway were calculated as , where null is estimated from simulation. We used both Jaccard Index and correlation to show the overlapping of two pathways. For each pair of pathways, the Jaccard Index were defined as the aggregated mutation rate of overlapping genes divided by aggregated mutation rate of all genes, and correlation were calculated as the Pearson correlation during simulation. Network layout is generated by “Prefuse Forced Directed OpenCL Layout” algorithm in Cytoscape.

**Protein-protein interaction analysis**

We tested the protein level interaction of damage de novo variants using StringDB20 {stringDB}. Interactions were filtered by stringDB score≥0.4 and visualized by Cytoscape21 {Cytoscape}. Proteins that are not connected to any other genes after interaction filtration were removed from the network. Network layout is generated by “Prefuse Forced Directed OpenCL Layout” algorithm in Cytoscape. For each gene, Degree was calculated as the sum of all stringDB score of it.

**Single cell analysis**

Single cell data was downloaded from a recent publication of developmental mouse foregut15 {Han 2020} processed by the same procedures as in their paper.

**Results**

**Burden of de novo coding variants**

We identified 250 de novo coding variants on 185 probands diagnosed with EA/TEF (Table S3). The average number of de novo coding variants per proband is 1.35. Patients were classified as isolated cases and complex cases based on their phenotype {method part}. We classified likely gene disruption (LGD) and missense variants as damaging variants. We identified 192 damaging variants in all probands, including 48 (25%) in isolated cases and 144 (75%) in complex cases.

We performed a burden test for enrichment of the de novo coding variants in all cases, isolated cases and complex cases respectively. The number of synonymous variants is close to expectation (fold=0.94, p=0.695). Both categories of variants are more enriched in constrained genes than non-constrained genes(fold=1.92 vs 0.89 for LGD; fold=1.36 vs 1.18 for missense; Table 1). The enrichment of LGD variants and missense variants (fold=1.33, p=1.5E-03) are higher in complex cases compared to isolated cases (fold=1.53 vs 0.49 for LGD, fold=1.33 vs 1.03 for missense). We further tested the enrichment of variants in complex cases in gene sets divided by rank score of gnomAD MisOEUF (Table S4), both LGD and missense variants were significantly enriched in top 30% MisOEUF ranked genes (fold=2.51, p-value=5.50E-03 for LGD; fold=1.52, p-value=4.36E-03 for missense). Three genes, ERCC1, GLS, TLR9 have recurrent missense variants in our cohort.

Furthermore, we observed 21 damaging variants (10.9%) in the genes that are putative targets of EFTUD2{POLARIS, method part}, including an LGD variant in EFTUD2 gene and 3 genes with LGD variants: MANBAL, GTF2I and IVD. The enrichment of LGD and missense variants are higher in those genes compared to genome baseline (fold=2.51 vs 1.20 for LGD; fold=1.46 vs 1.23 for missense). Overall, the burden indicates that LGD variants as well as missense variants in complex cases are likely to be pathogenic, and EFTUD2 targeted genes plays an important role in disease.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | All cases (n=185) | | | | Isolated cases (n=59) | | | | Complex cases (n=126) | | | |
| obs | exp | fold | p | obs | exp | fold | p | obs | exp | fold | p |
| ALL 19384 genes | Synonymous | 58 | 61.6 | 0.94 | 6.95E-01 | 23 | 19.7 | 1.17 | 2.53E-01 | 35 | 42.0 | 0.83 | 8.78E-01 |
| LGD | 23 | 19.2 | 1.20 | 2.21E-01 | 3 | 6.1 | 0.49 | 9.43E-01 | 20 | 13.1 | 1.53 | 4.48E-02 |
| Missense | 169 | 137.3 | 1.23 | 4.93E-03 | 45 | 43.8 | 1.03 | 4.48E-01 | 124 | 93.5 | 1.33 | 1.50E-03 |
| Damage (LGD+missense) | 192 | 156.5 | 1.23 | 3.34E-03 | 48 | 49.9 | 0.96 | 6.26E-01 | 144 | 106.6 | 1.35 | 3.29E-04 |
| pLI 4365 genes | Synonymous | 22 | 18.7 | 1.18 | 2.50E-01 | 9 | 6.0 | 1.51 | 1.49E-01 | 13 | 12.7 | 1.02 | 5.06E-01 |
| LGD | 11 | 5.7 | 1.92 | 3.22E-02 | 0 | 1.8 | 0.00 | 1.00E+00 | 11 | 3.9 | 2.82 | 2.34E-03 |
| Missense | 56 | 41.2 | 1.36 | 1.64E-02 | 12 | 13.1 | 0.91 | 6.62E-01 | 44 | 28.1 | 1.57 | 3.27E-03 |
| Damage (LGD+missense) | 67 | 47.0 | 1.43 | 3.41E-03 | 12 | 15.0 | 0.80 | 8.14E-01 | 55 | 32.0 | 1.72 | 1.35E-04 |
| other 15019 genes | Synonymous | 36 | 42.9 | 0.84 | 8.74E-01 | 14 | 13.7 | 1.02 | 5.03E-01 | 22 | 29.2 | 0.75 | 9.29E-01 |
| LGD | 12 | 13.5 | 0.89 | 6.93E-01 | 3 | 4.3 | 0.70 | 8.02E-01 | 9 | 9.2 | 0.98 | 5.68E-01 |
| Missense | 113 | 96.1 | 1.18 | 5.00E-02 | 33 | 30.6 | 1.08 | 3.59E-01 | 80 | 65.5 | 1.22 | 4.47E-02 |
| Damage (LGD+missense) | 125 | 109.6 | 1.14 | 7.93E-02 | 36 | 34.9 | 1.03 | 4.52E-01 | 89 | 74.6 | 1.19 | 5.73E-02 |
| EFTUD2 targets (POLARIS >= 0.7) 1596 genes | Synonymous | 4 | 5.2 | 0.77 | 7.63E-01 | 1 | 1.7 | 0.60 | 8.10E-01 | 3 | 3.5 | 0.85 | 6.88E-01 |
| LGD | 4 | 1.6 | 2.51 | 7.80E-02 | 1 | 0.5 | 1.97 | 3.99E-01 | 3 | 1.1 | 2.76 | 9.67E-02 |
| Missense | 17 | 11.6 | 1.46 | 8.21E-02 | 7 | 3.7 | 1.89 | 8.24E-02 | 10 | 7.9 | 1.26 | 2.73E-01 |
| Damage (LGD+missense) | 21 | 13.2 | 1.59 | 2.91E-02 | 8 | 4.2 | 1.90 | 6.50E-02 | 13 | 9.0 | 1.44 | 1.24E-01 |

Table 2. Burden of de novo variants. Burden were calculated for all cases, isolated cases and complex cases. Damage variants were defined as LGD and missense variants.

**Damage variants in complex cases are involved in critical pathways**

We performed pathway-wide enrichment analysis on damage variants of complex cases (n=126) to identify the underlying pathogenesis. We focused on GO (Gene Ontology) and HPO (Human Phenotype Ontology) pathways that are expected to have at least 2 LGD or missense variants by chance in the complex cases cohort (n=126). We compared the observed variants of each pathway to the expected variant from background mutation rate, and tested the enrichment using Poisson test. We corrected the multi-testing p-values to family wise error rate (FWER) based on 20000 times simulations {method part}. The damage variants were significantly enriched in 13 pathways with FWER≤0.05 (Fig. 1, Table S5). Remarkably, we identified three pathways that are involved in autophagy related process and three pathways in protein transport and localization, which are mostly driven by missense variants. Those pathways indicate the important role of transcytosis in early fetal foregut development, during which the trachea and esophagus arise from the separation of a common foregut. Transcytosis has been shown to be a regulator of epithelial cell shape and tubulogenesis during development in several model organisms22 {Serra 2021}, where it can synchronize apical and basal membrane growth during early Drosophila tracheal terminal cell development22–24 {Serra 2021; Best, 2019; Mathew 2020}. Both endocytosis and transcytosis can contribute to reshape of epithelial cells through the redistribution of membrane. It has been further shown in mouse models that this endosome-mediated epithelial remodeling is been regulated by Hedgehog-Gli signaling pathways, mutations in which can disrupt tracheoesophageal morphogenesis and cause life-threatening birth defects25 {Nasr, Mancini 2019}. There are 7 Hedgehog-Gli target genes with LGD variants in the patient cohort, with a high enrichment compared to background mutation rate (fold=3.66, p-value=6.74E-03, Supplementary Table 1). Among those genes, MYCN is a known gene to cause Feingold syndrome [OMIM 164280] that includes Esophageal & duodenal atresias and cardiac effects. CAM2B and KIF17 variants were found in GeneDx EA/TEF patients as well.

Moreover, we identified several core developmental pathways implicated in other developmental disorders, which are driven by both LGD and missense variants. Including KMT2D, a well-known congenital heart disease (CHD) risk gene.

Chart, radar chart

Description automatically generated

Fig 1. Pathway enrichment analysis. Dot size represents the number of observed de novo variants in the pathway; Dot color represents the FWER; Edge width represents the Jaccard Index between two pathways, and Edge color represents the correlation of the two pathways based on simulation.

To further reveal the protein level associations of those autophagy related genes. We used stringDB to show the genes that harbor LGD and missense genes in complex cases (150 genes in total). We observed a significant enriched number of edges in those genes (PPI enrichment p-value=0.00241, Fig.2). A total of 85 genes are involved in at least one significant pathway. 64 genes were involved in endocytosis and transcytosis pathways, among which 4 genes (AP3D1, EEF1A2, PHB2, TOMM40) were involved in more than 4 pathways. 44 genes were involved in pathways related to other developmental disorders, including EFTUD2 and KMT2D. Furthermore, we identified several core genes that are prioritized in the connections with other genes, including POLR2B (degree=7.233), UBR4(degree=6.469), KMT2D (degree=5.639). Overall, those genes are likely to act as core elements in the pathogenesis.



Fig.2 String-DB of LGD and missense genes in complex cases. Dots were colored to indicate whether it is involved in one of the significant pathways. Edge width represents the stringDB score. Genes without connections were not shown.

**Discussion**

In this study, we identified 250 de novo variants in 185 EA/TEF patients, including 23 LGD variants and 169 missense variants. Damage variants are enriched in complex cases. Consistant with previous studies of birth defects, those damage variants showed higher enrichment in constrained genes. Pathway analysis showed that damage variants are likely to cause abnormalities in autophagy related processes, which are important regulators of epithelial cell shape remodeling during tracheoesophageal separation. Furthermore, LGD variants are highly enriched in putative targets of EFTUD2 and Hedgehog/Gli signaling pathways, which is consistent with findings in the mouse model. Both LGD and missense variants in complex cases are involved in several pathways implicated in other developmental disorders. Overall, our study revealed the potential etiologies of EA/TEF and could further benefit from larger sample size.

**Genes with LGD variants are highly expressed during mouse fetal foregut development**

Single cell (Ongoing)

1. Cluster/time specific expression of LGD+missense genes

No significant cluster/time specific expression. Apob is in extra\_embryonic, SLC4A1 is in blood.

1. RNA velocity of LGD+missense genes

Most genes have 0 velocity because the expression level is too low.

Reference:

1. Nassar, N. *et al.* Prevalence of esophageal atresia among 18 international birth defects surveillance programs. *Birth Defects Research Part A: Clinical and Molecular Teratology* **94**, 893–899 (2012).

2. van Lennep, M. *et al.* Oesophageal atresia. *Nature Reviews Disease Primers* **5**, 1–21 (2019).

3. Stoll, C., Alembik, Y., Dott, B. & Roth, M.-P. Associated anomalies in cases with esophageal atresia. *American Journal of Medical Genetics Part A* **173**, 2139–2157 (2017).

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* **53**, 6–13 (2010).

5. Li, Y., Litingtung, Y., Dijke, P. T. & Chiang, C. Aberrant Bmp signaling and notochord delamination in the pathogenesis of esophageal atresia. *Developmental Dynamics* **236**, 746–754 (2007).

6. Que, J., Choi, M., Ziel, J. W., Klingensmith, J. & Hogan, B. L. M. Morphogenesis of the trachea and esophagus: current players and new roles for noggin and Bmps. *Differentiation* **74**, 422–437 (2006).

7. Que, J. *et al.* Multiple dose-dependent roles for Sox2 in the patterning and differentiation of anterior foregut endoderm. *Development* **134**, 2521–2531 (2007).

8. Gordon, C. T. *et al.* EFTUD2 haploinsufficiency leads to syndromic oesophageal atresia. *Journal of Medical Genetics* **49**, 737–746 (2012).

9. Qi, H. *et al.* De novo variants in congenital diaphragmatic hernia identify MYRF as a new syndrome and reveal genetic overlaps with other developmental disorders. *PLOS Genetics* **14**, e1007822 (2018).

10. Richter, F. *et al.* Genomic analyses implicate noncoding de novo variants in congenital heart disease. *Nature Genetics* **52**, 769–777 (2020).

11. Wang, J. *et al.* Novel candidate genes in esophageal atresia/tracheoesophageal fistula identified by exome sequencing. *European Journal of Human Genetics* **29**, 122–130 (2021).

12. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology* **36**, 983–987 (2018).

13. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

14. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).

15. Han, L. *et al.* Single cell transcriptomics identifies a signaling network coordinating endoderm and mesoderm diversification during foregut organogenesis. *Nature Communications* **11**, 4158 (2020).

16. Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nature Genetics* **46**, 944–950 (2014).

17. Ware, J. S., Samocha, K. E., Homsy, J. & Daly, M. J. Interpreting de novo Variation in Human Disease Using denovolyzeR. *Current Protocols in Human Genetics* **87**, 7.25.1-7.25.15 (2015).

18. Mootha, V. K. *et al.* PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics* **34**, 267–273 (2003).

19. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS* **102**, 15545–15550 (2005).

20. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research* **47**, D607–D613 (2019).

21. Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).

22. Serra, N. D. & Sundaram, M. V. Transcytosis in the development and morphogenesis of epithelial tissues. *The EMBO Journal* **n/a**, e106163 (2021).

23. Best, B. T. Single-cell branching morphogenesis in the Drosophila trachea. *Developmental Biology* **451**, 5–15 (2019).

24. Mathew, R., Rios-Barrera, L. D., Machado, P., Schwab, Y. & Leptin, M. Transcytosis via the late endocytic pathway as a cell morphogenetic mechanism. *The EMBO Journal* **39**, e105332 (2020).

25. Nasr, T. *et al.* Endosome-Mediated Epithelial Remodeling Downstream of Hedgehog-Gli Is Required for Tracheoesophageal Separation. *Developmental Cell* **51**, 665-674.e6 (2019).