BRIEF COMMUNICATION   OPEN

Check for updates

Joint analysis of functionally related genes yields further candidates associated with Tetralogy of Fallot

Alexandru Chelu[1], Simon G. Williams[1], Bernard D. Keavney (ID)[1] and David Talavera (ID)[1✉]

Although several genes involved in the development of Tetralogy of Fallot have been identified, no genetic diagnosis is available for the majority of patients. Low statistical power may have prevented the identification of further causative genes in gene-by-gene survey analyses. Thus, bigger samples and/or novel analytic approaches may be necessary. We studied if a joint analysis of groups of functionally related genes might be a useful alternative approach. Our reanalysis of whole-exome sequencing data identified 12 groups of genes that exceedingly contribute to the burden of Tetralogy of Fallot. Further analysis of those groups showed that genes with high-impact variants tend to interact with each other. Thus, our results strongly suggest that additional candidate genes may be found by studying the protein interaction network of known causative genes. Moreover, our results show that the joint analysis of functionally related genes can be a useful complementary approach to classical single-gene analyses.

Tetralogy of Fallot (TOF) is the most common cyanotic congenital heart defect (CHD) [1]. A full understanding of the aetiology of TOF has remained elusive, especially the major genetic mechanisms that contribute to the development of non-syndromic cases. Some genes have recently been identified as the main contributors to the development of TOF [2–6]. Nonetheless, those genes only explain a minority of cases. Many more genes are likely to be involved in the development of TOF, but novel approaches may be necessary to identify them. Causative genes are usually identified by comparing their allele frequencies in cases and controls, i.e., as a principle, causative variants should be found in cases but not in controls. These gene-by-gene survey studies encounter two main difficulties when applied to the study of oligogenic/polygenic diseases: (1) since many genes are involved in the disease, their individual effect size is pretty small for most of them, and (2) the total number of tests to perform creates a big multiple comparison burden. As a result, most single-gene tests lack statistical power [7, 8].

The joint analysis strategy attempts to overcome these difficulties: (1) the total number of tests is smaller, and (2) the effect size may be greater if the effects are in the same direction. Therefore, one essential consideration of a joint analysis is which genes need to be analysed together, and which ones should not be merged. Since most proteins interact with other proteins in order to carry out their function, we hypothesised that variants in either member of an interacting pair working together in a biological process might have similar effects. This hypothesis is supported by previous research that found that clusters of functionally related proteins were associated with particular diseases [9–11]. Indeed, Reuter et al. found that most genes known/suspected to be involved in TOF participate in a tightly packed protein interaction network [5]. In this study, we used the

joint analysis approach in order to reanalyse whole-exome data from 829 patients of isolated, non-syndromic TOF. The cohort and the sequencing data have been described previously [4]. Instead of a gene-by-gene survey, we jointly analysed groups of genes that had been clustered based on current biological knowledge. This is an alternative approach to the network propagation method that has been used to identify new candidate genes interacting with genes known to be involved in a particular disease [12–14]. We grouped human proteins based on two conditions: (1) proteins had to participate in the same biological process as defined by the Gene Ontology [15, 16] (disregarding author/curator statements and electronic annotations), and (2) proteins had to physically interact with at least another protein within that group as reported by the BioGRID database [17, 18] (all reported interactions were included in the analysis). We focused our analysis on high-impact SNVs, i.e., single-nucleotide variants affecting splice sites, removing existent start or stop codons, or introducing novel stop codons. These variants are the most likely to affect the protein interaction network and hence the biological processes they participate in. Although moderate-effect variants (e.g., missense variants) may also be detrimental their effect on the biological processes is more difficult to assess. We used a permutation test in order to test if the number of patients with high-impact variants in particular groupings of functionally related genes was greater than those expected by chance in subsets of identical size (see Fig. 1 for a graphical representation of the analysis workflow). Similarly, we used a permutation test for assessing if there were more protein–protein interactions than expected by chance within the identified groupings of genes.

Our results show that 12 functional groupings exceed the number of expected patients with high-impact variants (Bonferroni-adjusted *p* value < 0.01; Fig. 2A, B). Those 12 groupings

---

[1]Division of Cardiovascular Sciences, School of Medical Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, UK. ✉email: David.Talavera@manchester.ac.uk
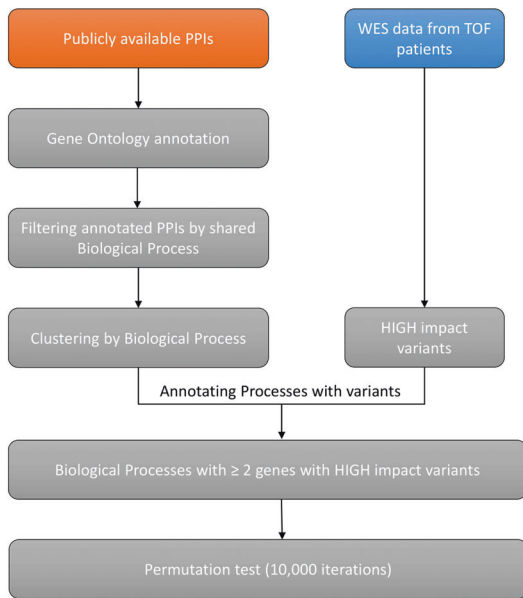
Fig. 1 Diagram representing the analysis workflow. PPIs were extracted from the public repository. Each protein was annotated with the GO Biological Process terms assigned to them. Only PPIs where both proteins participate in the same biological process were kept. All biological process-based subnetworks of PPIs were identified. WES data from TOF patients were obtained. Only high-impact variants were kept for further analysis. PPIs subnetworks were annotated with the genetic variation information. Subnetworks containing at least two genes with high-impact variants were kept. A permutation test was used in order to assess if a particular subnetwork (biological process) was associated with more patients than expected by chance. The Python scripts used in this analysis are available from https://github.com/AlexUOM/JHG-Manuscript-code. PPIs protein–protein interactions, GO Gene Ontology, WES whole-exome sequencing, TOF Tetralogy of Fallot

contain 222 genes that were identified as having at least one high-impact variant in at least one patient. Although high-impact variants are likely to disrupt the protein function, we cannot be sure if the cell/organism can tolerate that effect. One way of assessing this is by studying if genes are under strong selective pressure, i.e., the number of variants observed in the population is smaller than expected [19]. Of those 222 genes, 69 have a pLI ≥0.9 (Supplementary Table I), showing enrichment for genes intolerant to loss-of-function (31.2% in the set of candidate genes vs 15.8% in the rest of the genome) (*p* value < 0.01; proportion test). A total of 165 patients (19.9 %) contain a high-impact variant in a single gene in those groupings, while 24 additional patients (2.9 %) contain variants in more than one gene. Thus, high-impact variants affecting those 12 functional groupings were found in 22.8% of the patients. In addition to biological processes assumed to be involved in TOF such as signalling pathways (19 patients) or regulation of transcription (130 patients), there were groupings involved in post-translational protein modification (41 patients), intracellular protein transport (10 patients), and cilium assembly (23 patients). This latter result in particular is in accord with the growing evidence showing that ciliopathies are linked to many cases of congenital heart disease, including TOF [3, 20]. Groupings are extremely sensitive to the functional annotations used, i.e., other annotation systems would likely lead to slightly different results (Supplementary Tables II and III and Supplementary Figs. S1–S3). Our results also confirmed our expectation that TOF-associated variants should be in interacting partners within the biological process (Fig. 2C, D and Supplementary Tables IV and V): interactions between proteins with high-impact variants exceed their expected-by-chance number in 8 out of the 12 groupings (*p* value < 0.05; Bonferroni-corrected permutation test), with some proteins bridging distinct biological processes.

Reassuringly, our approach recapitulated previous findings obtained in two gene-by-gene analyses of the same data [4, 5], i.e., we identified *FLT4*, *NOTCH1*, *KDR*, *JAG1* and *GATA6* as members
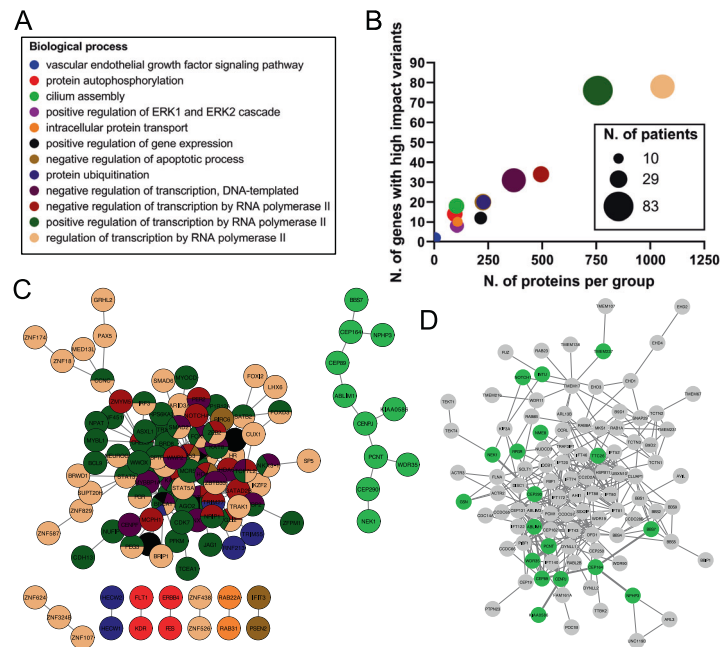
Fig. 2 Functional groupings of genes associated with TOF. **A** List of functional groupings whose high-impact variants are associated with more patients than expected by chance. Colours are used to represent the functional groupings in the subsequent panels. **B** Description of the 12 statistically significant functional groupings. The scatter plot shows the number of proteins within each grouping, the number of genes with high-impact variants within the groupings and the number of patients with at least a variant in those groupings. **C** Network of protein interactions between proteins with high-impact variants. Out of the 222 proteins, 111 are present in the network. Nodes are coloured according to the within-biological process interactions. **D** Interaction network of the cilium assembly grouping. Green nodes represent proteins with high-impact variants, and grey nodes represent other proteins in the functional grouping. Interactions between proteins with high-impact variants are highlighted

of those functional groupings with high-impact variants. Indeed, 9 out of the 26 genes recently highlighted by Reuter et al. are members of these groupings. Importantly, we were also able to identify other genes involved in those functional groupings that had not previously been linked to TOF. Some of those genes had been associated with other types of CHD (e.g., *CEP290* [21], *KIAA0586* [21], *TCF12* [22]), or other cardiac phenotypes unrelated to TOF (e.g., *PSEN2* [23] and *CD36* [24]). Nonetheless, there was no known association between cardiovascular diseases and the vast majority of genes susceptible to altering the highlighted biological processes (e.g., *ARID3A*, *BRWD1*, *CBLC*, *CENPF*, *GSN*, *LHX6*, *MAP3K3*, *NRIP1*, *RNF213*, *TNIK*, *ZNF274*, *ZNF407*, and *ZNF808*).

Our analysis has not only been able to recapitulate the most prominent general biological processes known to be involved in the development of TOF such as signalling pathways and regulation of transcription but has also identified more specific processes known to be involved in CHD such as cilium assembly [3, 20]. Importantly, this approach can be used for highlighting possible candidates that would be overlooked in classical gene-by-gene analyses due to a lack of statistical power. Finally, our results suggest that the interaction partners of some known or emerging TOF candidates should be prioritised in future functional analyses. Our findings suggest that the joint analysis of groups of functionally related genes may be a powerful tool for identifying novel putative candidates involved in the development of congenital diseases.

REFERENCES

1. Bailliard F, Anderson RH. Tetralogy of Fallot. Orphanet J Rare Dis. 2009;4:2.
2. Jin SC, Homsy J, Zaidi S, Lu Q, Morton S, DePalma SR, et al. Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands. Nat Genet. 2017;49:1593–601.
3. Pierpont ME, Brueckner M, Chung WK, Garg V, Lacro RV, McGuire AL, et al. Genetic basis for congenital heart disease: revisited: a scientific statement from the American Heart Association. Circulation. 2018;138:e653–e711.
4. Page DJ, Miossec MJ, Williams SG, Monaghan RM, Fotiou E, Cordell HJ, et al. Whole exome sequencing reveals the major genetic contributors to non-syndromic Tetralogy of Fallot. Circ Res. 2019;124:553–63.
5. Reuter MS, Chaturvedi RR, Jobling RK, Pellecchia G, Hamdan O, Sung WWL, et al. Clinical genetic risk variants inform a functional protein interaction network for Tetralogy of Fallot. Circ Genom Precis Med. 2021;14:e003410.
6. Skoric-Milosavljevic D, Lahrouchi N, Bosada FM, Dombrowsky G, Williams SG, Lesurf R, et al. Rare variants in KDR, encoding VEGF Receptor 2, are associated with Tetralogy of Fallot. Genet Med. 2021;23:1952–60.
7. Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. Nat Rev Genet. 2014;15:335–46.
8. Tong DMH, Hernandez RD. Population genetic simulation study of power in association testing across genetic architectures and study designs. Genet Epidemiol. 2020;44:90–103.
9. Aibar S, Fontanillo C, Droste C, De Las Rivas J. Functional Gene Networks: R/Bioc package to generate and analyse gene networks derived from functional enrichment and clustering. Bioinformatics. 2015;31:1686–8.
10. Ghiassian SD, Menche J, Barabasi AL. A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. PLoS Comput Biol. 2015;11:e1004120.
11. Sun PG, Gao L, Han S. Prediction of human disease-related gene clusters by clustering analysis. Int J Biol Sci. 2011;7:61–73.
12. Siitonen A, Kytovuori L, Nalls MA, Gibbs R, Hernandez DG, Ylikotila P, et al. Finnish Parkinson's disease study integrating protein-protein interaction network data with exome sequencing analysis. Sci Rep. 2019;9:18865.
13. Smedley D, Kohler S, Czeschik JC, Amberger J, Bocchini C, Hamosh A, et al. Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases. Bioinformatics. 2014;30:3215–22.
14. Yepes S, Tucker MA, Koka H, Xiao Y, Jones K, Vogt A, et al. Using whole-exome sequencing and protein interaction networks to prioritize candidate genes for germline cutaneous melanoma susceptibility. Sci Rep. 2020;10:17198.
15. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25:25–9.
16. Gene Ontology C. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. 2021;49:D325–D34.
17. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 2006;34:D535–9.
18. Oughtred R, Rust J, Chang C, Breitkreutz BJ, Stark C, Willems A, et al. The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci. 2021;30:187–200.
19. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536:285–91.
20. Li Y, Klena NT, Gabriel GC, Liu X, Kim AJ, Lemke K, et al. Global genetic analysis in mice unveils central role for cilia in congenital heart disease. Nature. 2015;521:520–4.
21. Alby C, Piquand K, Huber C, Megarbane A, Ichkou A, Legendre M, et al. Mutations in KIAA0586 cause lethal ciliopathies ranging from a hydrolethalus phenotype to short-rib polydactyly syndrome. Am J Hum Genet. 2015;97:311–8.
22. Morton SU, Shimamura A, Newburger PE, Opotowsky AR, Quiat D, Pereira AC, et al. Association of damaging variants in genes with increased cancer risk among patients with congenital heart disease. JAMA Cardiol. 2021;6:457–62.
23. Li D, Parks SB, Kushner JD, Nauman D, Burgess D, Ludwigsen S, et al. Mutations of presenilin genes in dilated cardiomyopathy and heart failure. Am J Hum Genet. 2006;79:1030–9.
24. Ma X, Bacci S, Mlynarski W, Gottardo L, Soccio T, Menzaghi C, et al. A common haplotype at the CD36 locus is associated with high free fatty acid levels and increased cardiovascular risk in Caucasians. Hum Mol Genet. 2004;13:2197–205.

AUTHOR CONTRIBUTIONS
DT conceived and designed the study. AC conducted the experiments. SGJ and BDK contributed data. All authors analysed and interpreted the results. DT wrote the manuscript. AC, SGJ and BDK critically reviewed the manuscript.

COMPETING INTERESTS
The authors declare no competing interests.

ADDITIONAL INFORMATION
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s10038-022-01051-y.

Correspondence and requests for materials should be addressed to David Talavera.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.