Genes and pathways implicated in tetralogy of Fallot revealed by ultra-rare variant burden analysis in 231 genome sequences

Roozbeh Manshaei PhD[1,§], Daniele Merico PhD[2,3,§], Miriam S. Reuter MD[1,3], Worrawat Engchuan PhD[3], Bahareh A. Mojarad PhD[4], Rajiv Chaturvedi MD, PhD[1,5], Tracy Heung BSc[7], Giovanna Pellecchia PhD[3], Mehdi Zarrei PhD[3,4], Thomas Nalpathamkalam BSc[3], Reem Khan MA[1], John B. A. Okello PhD[1], Eriskay Liston MSc[1], Meredith Curtis BSc[1], Ryan K.C. Yuen PhD[3,4], Christian R. Marshall PhD[3,6,7,8], Rebekah K. Jobling MD[1,7], Stephen W. Scherer PhD[3,4,6,9], Raymond H. Kim MD, PhD[1,10,11,ϕ], Anne S. Bassett MD[12,13,14,15,ϕ,*]

[1] Ted Rogers Centre for Heart Research, Cardiac Genome Clinic, The Hospital for Sick Children, Toronto, Ontario, Canada
[2] Deep Genomics Inc., Toronto ON Canada
[3] The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, Ontario, Canada
[4] Program in Genetics and Genome Biology, The Hospital for Sick Children, Toronto, Ontario, Canada
[5] Labatt Heart Centre, Division of Cardiology, The Hospital for Sick Children, Toronto, Ontario, Canada
[6] Centre for Genetic Medicine, The Hospital for Sick Children, Toronto, Ontario, Canada
[7] Genome Diagnostics, Department of Paediatric Laboratory Medicine, The Hospital for Sick Children, Toronto, Ontario, Canada
[8] Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada
[9] Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada
[10] Division of Clinical and Metabolic Genetics, The Hospital for Sick Children, Toronto, Ontario, Canada
[11] Fred A. Litwin Family Centre in Genetic Medicine, University Health Network, Department of Medicine, University of Toronto, Toronto, Ontario, Canada
[12] Clinical Genetics Research Program, Centre for Addiction and Mental Health, Toronto, Ontario, Canada
[13] Division of Cardiology, Toronto Congenital Cardiac Centre for Adults at the Peter Munk Cardiac Centre, Department of Medicine, University Health Network, Toronto, Ontario, Canada
[14] The Dalglish Family 22q Clinic for Adults with 22q11.2 Deletion Syndrome, Department of Psychiatry, and Toronto General Research Institute, University Health Network, Toronto, Ontario, Canada
[15] Department of Psychiatry, University of Toronto, Toronto, Ontario, Canada

§ These authors contributed equally to the work.
ϕ These authors shared the senior authorship.
* Correspondence: anne.bassett@utoronto.ca

Word count: Abstract = 279 words;   Text = 7870 words
3 Tables; 3 Figures; 14 Supplemental tables; 5 Supplemental figures

45 **<u>Abstract</u>**

46 Recent genome-wide studies of rare genetic variants have begun to implicate novel mechanisms

47 for tetralogy of Fallot (TOF), a severe congenital heart defect (CHD).

48 To provide statistical support for case-only data without parental genomes, we re-analyzed genome

49 sequences of 231 individuals with TOF or related CHD. We adapted a burden test originally

50 developed for *de novo* variants to assess singleton variant burden in individual genes, and in gene-

51 sets corresponding to functional pathways and mouse phenotypes, accounting for highly correlated

52 gene-sets, and for multiple testing.

53 The gene burden test identified a significant burden of deleterious missense variants in *NOTCH1*

54 (Bonferroni-corrected p-value <0.01). These *NOTCH1* variants showed significant enrichment for

55 those affecting the extracellular domain, and especially for disruption of cysteine residues forming

56 disulfide bonds (OR 39.8 vs gnomAD). Individuals with *NOTCH1* variants, all with TOF, were

57 enriched for positive family history of CHD. Other genes not previously implicated in TOF had

58 more modest statistical support and singleton missense variant results were non-significant for

59 gene-set burden. For singleton truncating variants, the gene burden test confirmed significant

60 burden in *FLT4.* Gene-set burden tests identified a cluster of pathways corresponding to VEGF

61 signaling (*FDR*=0%), and of mouse phenotypes corresponding to abnormal vasculature

62 (*FDR*=0.8%), that suggested additional candidate genes not previously identified (e.g., *WNT5A* and

63 *ZFAND5*). Analyses using unrelated sequencing datasets supported specificity of the findings for

64 CHD.

65 The findings support the importance of ultra-rare variants disrupting genes involved in VEGF and

66 NOTCH signaling in the genetic architecture of TOF. These proof-of-principle data indicate that

67 this statistical methodology could assist in analyzing case-only sequencing data in which ultra-rare

68   variants, whether *de novo* or inherited, contribute to the genetic etiopathogenesis of a complex

69   disorder.

70   **<u>Author summary</u>**

71   We analyzed the ultra-rare nonsynonymous variant burden for genome sequencing data from 231

72   individuals with congenital heart defects, most with tetralogy of Fallot. We adapted a burden test

73   originally developed for *de novo* variants. In line with other studies, we identified a significant

74   truncating variant burden for *FLT4* and deleterious missense burden for *NOTCH1*, both passing a

75   stringent Bonferroni multiple-test correction. For *NOTCH1*, we observed frequent disruption of

76   cysteine residues establishing disulfide bonds in the extracellular domain. We also identified genes

77   with BH-FDR <10% that were not previously implicated. To overcome limited power for

78   individual genes, we tested gene-sets corresponding to functional pathways and mouse phenotypes.

79   Gene-set burden of truncating variants was significant for vascular endothelial growth factor

80   signaling and abnormal vasculature phenotypes. These results confirmed previous findings and

81   suggested additional candidate genes for experimental validation in future studies. This

82   methodology can be extended to other case-only sequencing data in which ultra-rare variants make

83   a substantial contribution to genetic etiology.

84   **<u>Introduction</u>**

85   Congenital heart defects (CHD) occur in 8/1000 live births and are a leading cause of mortality

86   from birth defects (1), with a wide spectrum of severity (2). Among CHD, tetralogy of Fallot (TOF)

87   is the most common of the more severe (cyanotic) conditions. Individuals with TOF present with

88   a combination of abnormalities (pulmonary valve stenosis, right ventricular hypertrophy,

89   ventricular septal defect and overriding aorta) that together lead to insufficient tissue oxygenation.

90   Genetic factors are major contributors to the etiology of TOF. These include 20% of patients with

91   pathogenic copy number variants (CNV) or larger chromosomal anomalies (3,4). Recent studies

92   have also begun to elucidate the genome-wide role of rare variants at the sequence level, including

93   substitutions and small insertions/deletions.

94   In a multi-centre exome sequencing study of various CHD that focused on loss-of-function variants

95   and included parental sequencing data enabling *de novo* variant identification, the TOF sub-group

96   drove a significant genome-wide burden finding (p-value $\leq 1.3 \times 10^{-6}$) of *de novo* and rare inherited

97   heterozygous truncating variants for a novel gene, *FLT4* (5). Of 9 probands with *FLT4* truncating

98   variants, corresponding to 2.3% of the TOF group, 7 were inherited with evidence of incomplete

99   penetrance (5).

100   In an independent case-only study using whole genome sequencing (WGS), we investigated 175

101   adults with TOF for rare loss-of-function variants (including structural variants) disrupting *FLT4*

102   and other vascular endothelial growth factor (VEGF) pathway genes predicted to be

103   haploinsufficient based on the ExAC pLI index (6,7). We identified seven truncating variants in

104   *FLT4*, two in *KDR*, and one each in *BCAR1*, *FGD5*, *FOXO1*, *IQGAP1* and *PRDM1*, corresponding

105   in aggregate to 8% of participants; all variants were absent from public databases[6]. The results

106   suggested the importance of VEGF signaling; however, the statistical burden was not

107   systematically investigated. Another recent multi-centre exome sequencing study of 829 patients

108   with TOF reported genome-wide significant (p-value $\leq 5 \times 10^{-8}$) excess of ultra-rare (absent from a

109   public exome database and other reference data) deleterious variants for *FLT4* and *NOTCH1* (8).

110   Loss-of-function variants predominated for *FLT4*, and missense variants for *NOTCH1* (8).

111   In this study, we undertook a comprehensive statistical re-analysis of the cohort with WGS data

112   previously investigated for ultra-rare variants in the VEGF pathway (6). In an attempt to boost

113 power, we included the sequencing data available for 56 CHD cases as well as for the original 175

114 TOF cases (n=231 total). We focused on ultra-rare truncating (stop-gain, frameshift and splice site

115 altering) and missense variants that were not reported in the gnomAD database and were identified

116 in only one proband, i.e., that were singletons. We tested burden by adapting a test originally

117 developed for *de novo* variants by rescaling the mutation probability for singletons. Since

118 singletons are enriched in *de novo* variants and are likely to have arisen recently, this is an

119 appropriate extension of the test. To boost power, we additionally tested gene-sets corresponding

120 to (a) functional pathways, derived from Gene Ontology (GO) (9), BioCarta (http://cgap.nci.nih-

121 .gov/Pathways/BioCarta_Pathways/), Kyoto Encyclopedia of Genes and Genomes (KEGG) (10)

122 (http://www.genome.jp/kegg/), REACTOME (11), NCI-Nature Pathway Interaction Database

123 (PID) (http://pid.nci.nih.gov); and (b) phenotypes in mouse orthologues, derived from Mouse

124 Genome Informatics (MGI) and based on the Mouse Phenotype Ontology (MPO) classification

125 (12). To control for correlations between highly overlapping gene-sets that could lead to incorrect

126 multiple p-value corrections, we adopted a greedy step-down approach to cluster gene-sets with

127 highly overlapping genes. A sampling-based false discovery rate (FDR) was then estimated. We

128 did not analyze structural variants because no broadly-accepted probabilistic framework has yet

129 been developed to determine the statistical significance of their burden.

130 **<u>Results</u>**

131 **Identification of singleton variants**

132 Variant calls from the CHD WGS data-set were filtered to retain only high-quality singletons, that

133 were then categorized as truncating or missense based on their effect on the principal transcript

134 (see Materials and Methods for details). With respect to the 2,003 truncating singleton variants

135   initially identified, 868 variants remained after applying the low quality and frameshift indel filter,

136   764 after applying the principal transcript effect filter, 752 after applying the splice site alteration

137   filter, and finally 642 after considering a maximum of one singleton variant per gene per subject.

138   For the 4,324 missense variants initially identified, 3,521 remained after applying the low-quality

139   filter, 3,359 after applying the principal transcript filter, and finally, 3,293 singleton missense

140   variants after considering a maximum of one singleton variant per gene per subject. We then tested

141   these ultra-rare truncating and missense singleton variants for gene and gene-set burden (see Figure

142   1 for an overview of the analysis workflow; all ultra-rare singleton variants identified are listed in

143   Supplementary Table S1). For all analyses we tested truncating and missense variants separately

144   because of the likely differences in the genetic architecture of these variant types.
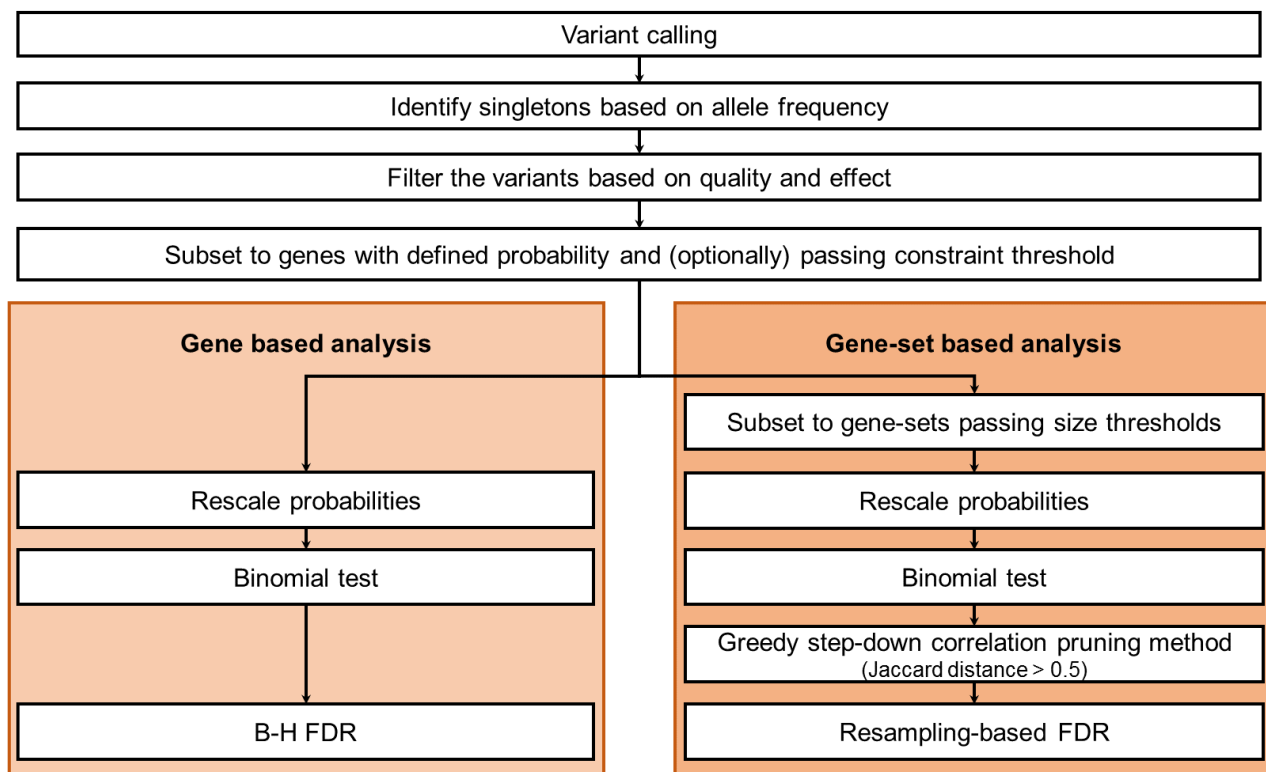
145

146   **Figure 1) General gene and gene-set burden analyses overview.**

147   **Gene burden results**

148   Genes were tested separately for the burden of singleton truncating and missense variants, using a

149   binomial test based on rescaled *de novo* mutation probabilities (as described in the Materials and

150   Methods). We performed multiple test correction on all genes with a defined probability, and also

151   on a more constrained subset: for truncating variants, gnomAD LOF o/e < 0.35; for missense

152   variants, gnomAD missense o/e < 0.75 (where *o/e* indicates observed/expected; see Supplementary

153   Figure S1 for the relation to the *pLI* and missense *z-score* constraint indexes). Constrained genes

154   are presumed to be more likely to contribute to disease, since they are under negative selection;

155   these thresholds were specifically set to include moderately constrained genes, considering the

156   incomplete penetrance observed for TOF (5,6,8). There were 603 genes with at least one truncating

157   singleton variant, of which 163 passed the constraint threshold; there were 2801 genes with at least

158   one singleton missense variant, of which 739 passed the constraint threshold (see Supplementary

159   Table S2 and Supplementary Table S3 for details). To assess the validity of the gene burden results,

160   we performed several additional analyses: (a) we checked the distribution of observed p-values

161   compared to expected p-values, to monitor for systematic p-value inflation; (b) we compared the

162   p-values obtained for CHD to those obtained for WGS data available for 263 individuals with

163   schizophrenia, processed in exactly the same way; (c) we reassessed the burden by comparing to

164   gnomAD singletons.

165   For truncating variants, there was only one constrained gene (of the 163 with at least one singleton

166   truncating variant) with significant burden: *FLT4* (uncorrected p-value = 9.56×10^{-12}, BH-FDR =

167   6.99×10^{-8}, Bonferroni-corrected p-value = 6.99×10^{-8}). When testing all genes (including 603 with

168   at least one singleton truncating variant), in addition to *FLT4*, we identified *CLDN9* as significant

169   at an FDR threshold of 10% (uncorrected p-value = 7.80×10^{-6}, BH-FDR = 0.073, Bonferroni-

170   corrected p-value = 0.145) (see Table 1 and Supplementary Table S2 for all details). There was no

171   evidence of genome-wide inflation in either analysis (see Figure 2 and Supplementary Figure S4).

172   Considering the top-associated CHD genes without using a constraint threshold, none had a similar

173   p-value in the schizophrenia sequencing data. When applying the constraint threshold, a single top-

174   associated gene that failed the 10% BH-FDR threshold (*ATXN3*) appeared to have a somewhat

175   similar p-value for schizophrenia. However, visualization of the bam files for the schizophrenia

176   data re-classified those variants to be in-frame polymorphisms (see Supplementary Table S4). For

177   *FLT4* and *CLDN9,* where BH-FDR was under the 10% threshold, we evaluated the truncating

178   **Table 1) Top six nominally significant genes with ultra-rare (singleton) variants identified in 231**
179   **individuals with CHD, as inferred from gene-based burden analyses for truncating and missense**
180   **singletons, respectively, with and without using a gene constraint cut-off.**

Gene Name	Number of observed variants[1]	All genes, no constraint		Genes with constraint[2]	
		p-value	BH-FDR[3]	p-value	BH-FDR[3]
Truncating variants					
FLT4	7	3.84×10^{-10}	7.15×10^{-6}	9.56×10^{-12}	6.99×10^{-8}
CLDN9	2	7.80×10^{-6}	0.0726	NA	NA
CCDC168	2	2.20×10^{-5}	0.1365	NA	NA
CSN2	2	0.0001	0.4823	NA	NA
L1TD1	2	0.0001	0.4823	NA	NA
TACC3	2	0.0002	0.6402	NA	NA
Missense variants					
NOTCH1	8	2.33×10^{-6}	0.0351	9.32×10^{-7}	0.0048
BCKDK	4	3.70×10^{-6}	0.0351	2.26×10^{-6}	0.0058
KL	4	1.54×10^{-5}	0.0973	NA	NA
DHH	3	3.01×10^{-5}	0.1369	2.08×10^{-5}	0.0361
PRRT4	3	3.78×10^{-5}	0.1369	NA	NA
VMAC	2	4.33×10^{-5}	0.1369	NA	NA

181   [1] All observed variants were in individuals with TOF, except 1 each in genes *CCDC36, CSN2, L1TD1, TACC3, KL,*
182   *DHH*, and *PRRT4*.
183   [2] Only those variants in genes with o/e score in gnomAD <0.35 for truncating variants and <0.75 for missense variants
184   are considered.
185   [3] The *Benjamini Hochberg* False Discovery Rate.
186   NA = not available, indicating that the respective gene was not in the gene list, thus data were not available.

187   singleton burden in CHD compared to that in gnomAD: *FLT4* had an even more significant

188   association (uncorrected p-value = 2.43×10^{-15}, BH-FDR = 4.01×10^{-11}), whereas *CLDN9* was less

189   significant (uncorrected p-value = 7.8×10^{-4}, BH-FDR =1), leading us to question the validity of its

190   association to CHD (see Supplementary Table S5 and Supplementary Figure S2). Restricting to

191   constrained genes may have some utility in prioritizing genes, but these results are too limited to

192   draw robust general conclusions.

193   Of the 739 genes with singleton missense variants that passed the constraint threshold, there were

194   three genes that passed the 10% FDR threshold: *NOTCH1* (uncorrected p-value = 9.32×10^{-7}, BH-

195   FDR=0.0048, Bonferroni-corrected p-value = 4.85×10^{-3}), *BCKDK* (uncorrected p-value = $2.26 \times 10^{-}$

196   ^{6}, BH-FDR=0.0058, Bonferroni-corrected p-value = 0.018), and *DHH* (uncorrected p-value =

197   2.08×10^{-5}, BH-FDR=0.0361, Bonferroni-corrected p-value = 0.108); see Table 1 and

198   Supplementary Table S3 for further details. When considering all 2801 genes with singleton

199   missense variants, regardless of constraint, the BH-FDR for gene *DHH* (0.1369) was less

200   significant, but another gene, *KL*, passed the 10% BH-FDR cut-off (uncorrected p-value = $1.54 \times 10^{-}$

201   ^{5}, BH-FDR=0.0973, Bonferroni-corrected p-value = 0.292) (see Table 1 and Supplementary Table

202   S3). There was no evidence of genome-wide inflation in either analysis (see Figure 2 and

203   Supplementary Table S4). When applying the constraint threshold, there was one top-associated

204   gene that did not meet the 10% BH-FDR threshold and that had somewhat similar results in the

205   schizophrenia cohort (*OLIG2*: CHD uncorrected p-value = 1.39×10^{-4}, BH-FDR=0.18;

206   schizophrenia uncorrected p-value = 0.017) (see Supplementary Table S6), indicating questionable

207   validity for CHD. For the genes identified without using the constraint threshold, none had a similar

208   p-value for schizophrenia. Comparing the missense singleton burden in CHD and in gnomAD,

209   genes *NOTCH1*, *BCKDK*, *DHH*, and *KL* displayed similar p-values, but only *NOTCH1* and

210   *BCKDK* passed the BH-FDR 10% threshold (see Supplementary Table S5 and Supplementary

211   Figure S3). These results suggest that for singleton missense gene burden in this study, there was

212   no apparent benefit in restricting to missense-constrained genes.
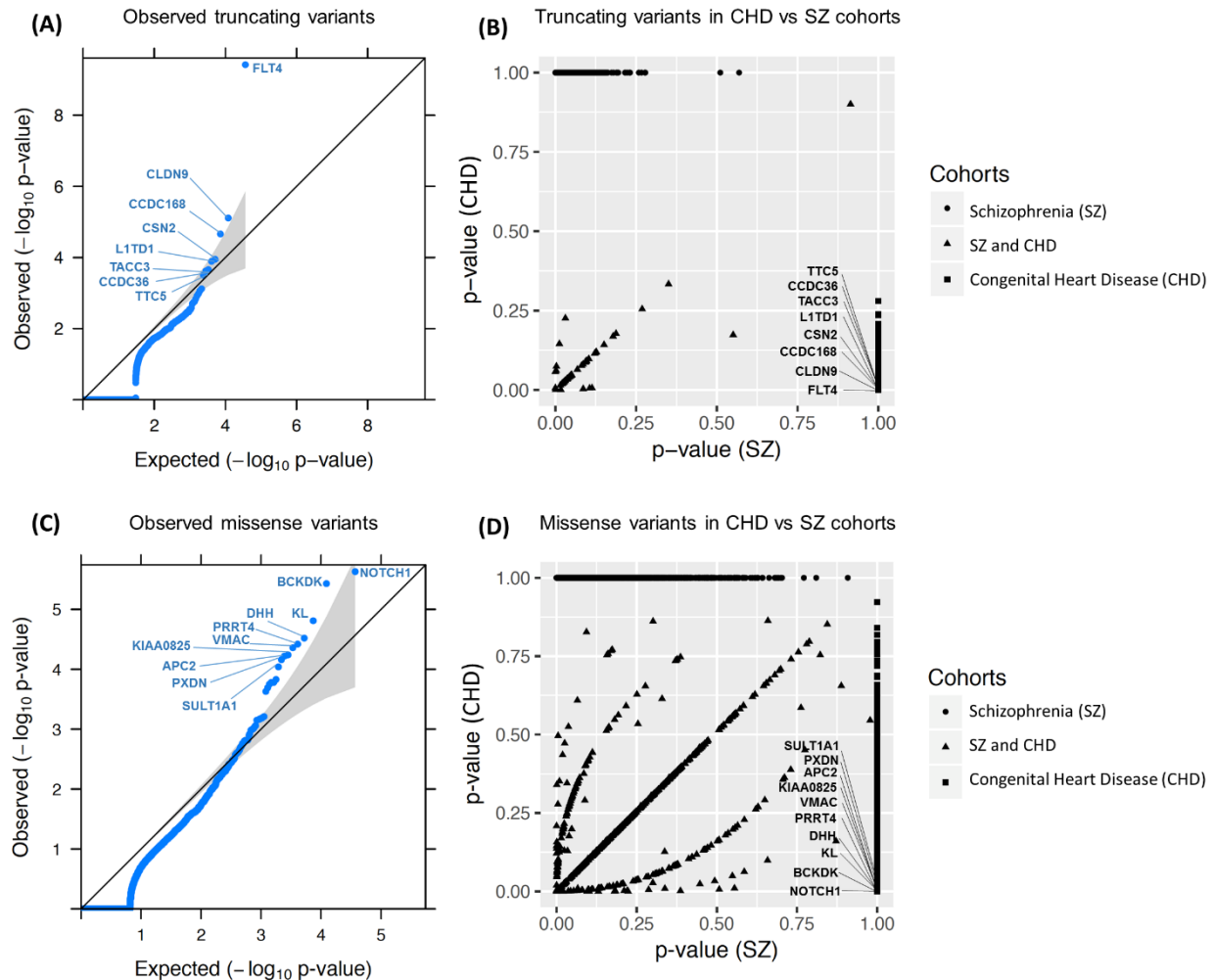
213
Figure 2) Gene burden analysis results for all genes. (A) and (C) show the quantile-quantile (QQ) plots
obtained for all ultra-rare truncating and missense variants in CHD, respectively (i.e., not setting any gene
constraint cutoff). The QQ plots represent the scatter plots of the –log10(p-value) expected under the null
hypothesis of no genetic association versus the observed –log10(p-value) for all 231 CHD samples. Grey
shading indicates the 95% confidence interval. (B) and (D) represent scatter plots of gene burden p-values
for truncating and missense variants respectively, comparing the CHD and schizophrenia WGS data. Names
of the top 8 and 10 genes identified for truncating (A) and missense (C) variants, respectively, are shown
(results for top 6 of each are presented in Table 1); *FLT4* (A) and *NOTCH1* (C) were the most significant
genes identified, neither with any observation in the comparison schizophrenia cohort (B, D). These plots
were generated based on the genes without constraint on o/e score. Supplementary Figure S4 shows results
for genes with constraint.

225   **Gene-set burden results**

226   Restricting to genes constrained for truncating variants, the gene-set burden analysis (as described

227   in the Materials and Methods) identified one cluster for GO and pathways, and one for MPO, both

228   of which were significant at the sampling FDR < 10%. The FDR approached 1.0 (non-significant)

229   for other clusters (see Table 2). Gene-set sub-clusters were manually identified with the aid of the

230   Cytoscape app EnrichmentMap (13) (see Table 3 and Supplementary Figure S5). The GO and

231   **Table 2) Top six gene-set clusters for truncating singleton variant burden analyses using Gene**
232   **Ontology (GO) / pathways and Mouse Phenotype Ontology (MPO), and restricting to constrained**
233   **genes.**
234

Gene-set clusters	Observed truncating variants in constrained genes (o/e score in gnomAD <0.35)	p-value	Permutation based FDR
GO and pathways			
VEGF signaling and blood vessel development	8	5.39×10^{-13}	0
Ion antiporter activity	5	0.0005	0.9564
Planar cell polarity pathway involved in neural tube closure	3	0.0013	0.9564
Positive regulation of vascular associated smooth muscle cell migration	4	0.0017	0.9564
Peptidyl-tyrosine autophorphorylation	5	0.0027	0.9564
Protein quality control for misfolded or incompletely synthesized proteins	3	0.0029	0.9564
MPO			
Abnormal lymphangiogenesis	7	9.64×10^{-11}	0.008
Abnormal cranial neural crest cell morphology	3	0.0010	0.9605
Neuronal cytoplasmic inclusions	2	0.0022	0.9605
Absent pharyngeal arches	4	0.0031	0.9605
Abnormal CD5-positive T cell number	2	0.0036	0.9605
Cochlear ganglion degeneration	4	0.0037	0.9605

235   pathway cluster (uncorrected p-value = 5.39×10^{-13}, sampling-based FDR = 0) comprised 30 gene-

236   sets, 20 of which were clearly related to VEGF signaling and/or blood vessel development

237   (angiogenesis); *FLT4* was by far the most significant gene (LOF variants N = 7, uncorrected p-

238   value = 9.56×10^{-12}), with other genes such as *KDR* (LOF variants N = 2, uncorrected p-value =

239   0.001), *ZFAND5* (LOF variant N = 1, uncorrected p-value = 0.008) and *WNT5A* (LOF variant N =

240   1, uncorrected p-value = 0.010) having more modest contributions (see Table 3 and Supplementary

241   Tables S7 and S8). The MPO cluster (uncorrected p-value = 9.64×10^{-11}, sampling-based FDR =

242   0.008) comprised 19 gene-sets, 15 of which corresponded to abnormalities of the cardiovascular

243   system such as abnormal vessel morphology and cardiac-related bleeding in mice (see Table 3 and

244   Supplementary Table S9 and S10). Again, *FLT4* encoding VEGFR3 was the largest contributor,

245   with other genes in the VEGF pathway including *KDR* encoding VEGFR2 and *FOXO1* (LOF N =

246   1, uncorrected p-value = 0.008) that were previously identified using manual curation (6). The GO

247   pathway and MPO cluster results however also identified other potential candidate genes for TOF

248   associated with functions of *FLT4* that were not identified in the previous study, including

249   *AKAP12*, *PKD1*, *ATF2*, and *EPN1* (Table 3). While other clusters were not significant after

250   multiple test correction, some top-scoring ones had a clear functional or phenotypic relation to

251   CHD (for instance, planar cell polarity in neural tube closure, ranking third for GO and pathways;

252   positive regulation of vascular smooth cell migration, ranking fourth for mouse phenotypes) and

253   included additional promising candidate genes (e.g., *DVL3*, *KIF3A*).

254   Since *FLT4* had such a prominent role in driving the gene-set signal for truncating variants, we

255   repeated the analysis without *FLT4*. No significant gene-set cluster was identified. Similar results

256   were obtained when considering all genes (i.e. without restricting to constrained genes), but the

257   MPO cluster had FDR slightly higher than the 10% threshold (see Supplementary Table S9). For

258 Table 3) Gene-set sub-clusters derived from the two gene-set clusters with significant truncating
259 singleton burden from Table 2.

Most significant composite gene-set sub-clusters	p-value	Genes [1] (contributing number of variants, p-value)
GO and pathways		
Positive regulation of protein kinase C signaling	5.39×10^{-13}	*FLT4*[2], *WNT5A* (1, 0.010)
Regulation of protein kinase C signaling	1.17×10^{-12}	*FLT4*[2], *WNT5A* (1, 0.010), *AKAP12* (1, 0.024)
VEGF and related pathways, and transmembrane receptor protein kinase activity	1.68×10^{-12}	*FLT4*[2], *KDR*[3] (2, 0.001)
Regulation of blood vessel remodeling, VEGFR3 signaling in lymphatic endothelium, and lung alveolus development	4.70×10^{-10}	*FLT4*[2]
Lymph vessel morphogenesis and development	6.73×10^{-9}	*FLT4*[2], *PKD1* (1, 0.046)
Respiratory system process and gaseous exchange	4.73×10^{-8}	*FLT4*[2], *ZFAND5* (1, 0.008)
Endothelial cell proliferation and migration	6.45×10^{-7}	*FLT4*[2], *KDR*[3] (2, 0.001), *WNT5A* (1, 0.010)
MPO		
Anterior cardinal vein development, abnormal lymph circulation, abnormal lymphatic system physiology, and ascites	9.64×10^{-11}	*FLT4*[2]
Abnormal lymphangiogenesis and abnormal lymphatic vessel morphology	1.55×10^{-10}	*FLT4*[2], *KDR*[3] (2, 0.001)
Heart hemorrhage	4.20×10^{-7}	*FLT4*[2], *KDR*[3] (2, 0.001), *ATF2* (1, 0.036), *PKD1* (1, 0.046)
Hemopericardium	2.33×10^{-6}	*FLT4*[2], *ATF2* (1, 0.036), *PKD1* (1, 0.046)
Skin edema and hydrops fetalis	5.84×10^{-5}	*FLT4*[2], *PKD1* (1, 0.046)
Abnormal vitelline vascular remodeling	2.41×10^{-4}	*FLT4*[2], *KDR*[3] (2, 0.001), *FOXO1*[3] (1, 0.008), *EPN1* (1, 0.015), *TTN* (1, 0.740)

260 [1] Genes listed are all those meeting the constraint threshold of o/e less than 0.35, e.g., including genes not reaching
261 significance (e.g., *TTN*).
262 [2] For each significant gene-set subcluster, for gene *FLT4* the number of variants contributing is 7, and the p-value is
263 9.56×10^{-12}.
264 [3] Candidate genes previously identified through manual curation methods, relevant to the VEGF pathway, in addition
265 to *FLT4* (6).

266 the missense variant analysis, we observed no significant gene-sets, with or without applying the

267 constraint cut-off (see Supplementary Table S11 and S12).

268   **Detailed *in-silico* analysis of missense variants in *NOTCH1* and other genes**

269   Given that our previous report had focused on truncating variants[6], we reviewed in detail the

270   singleton missense variants identified, considering amino acid conservation in orthologous

271   vertebrate sequences and *in-silico* predictors (SIFT, PolyPhen2, and Mutation Assessor) (14–16).

272   For *NOTCH1*, this manual review deemed 7 of the 8 ultra-rare missense variants to be either likely

273   deleterious (n=6) or potentially deleterious (n=1). For *BCKDK*, 1 of 4 was likely deleterious, and

274   1 of 4 potentially deleterious; for *KL*, 3 of 4 were potentially deleterious; for *DHH*, 1 of 3 was

275   likely deleterious and 1 of 3 potentially deleterious; see Supplementary Table S13 for details).

276   All 8 *NOTCH1* variants identified reside in the extracellular domain of the encoded protein (amino

277   acids 19-1735, see Figure 3), compared to 958 of 1,413 gnomAD v2.1 ultra-rare missense variants

278   (one-sided Fisher's Exact Test p-value = 0.045, odds ratio = +Inf). Similar to previously reported

279   exome sequencing findings (8), four of these 8 variants alter evolutionarily conserved cysteine

280   residues that establish disulfide bonds, located within the EGF-like repeats or the LNR (Lin12-

281   Notch) domain (17,18). This represents highly significant enrichment compared to such variants

282   from gnomAD v2.1 (23 of 958 variants; one-sided Fisher's Exact Test p-value = 3.15×10^{-5}, odds

283   ratio = 39.8).

284   Notably, all 8 of the ultra-rare missense variants in *NOTCH1* were identified within the 175

285   individuals with TOF, representing 4.6% of those studied. There was significant enrichment for

286   positive family history of CHD compared to the rest of the TOF sample (4 of 8 probands; two-

287   sided Fisher's Exact Test p-value = 0.003431, odds ratio = 11.49). Details of phenotype and family

288   history are provided for individuals with these 8 *NOTCH1* and 12 other variants in Supplementary

289   Table S14.

290
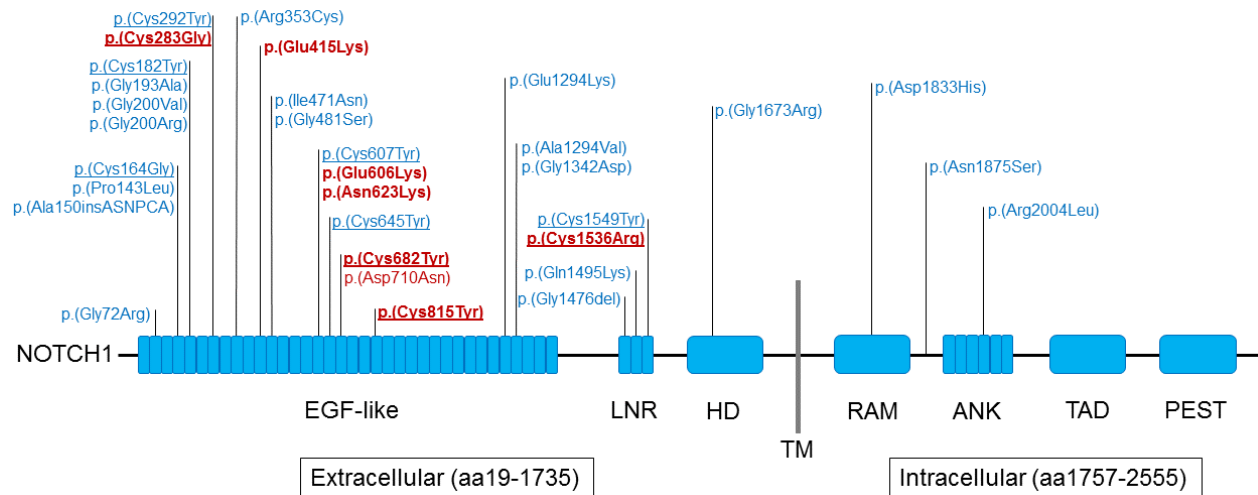
Figure 3) Schematic representation of *NOTCH1* domains (https://www.uniprot.org/uniprot/P46531)
and rare variants identified in individuals with tetralogy of Fallot. Findings from the current study
involving 8 of 175 probands with TOF are indicated in red font; 24 ultra-rare missense variants from the
Page et al. study (8) are indicated in blue font. The seven ultra-rare missense *NOTCH1* variants deemed to
be either likely deleterious (n=6) or potentially deleterious (p.(Asn623Lys)) are indicated in bold red font
(details in supplementary Table S13). Underline indicates those variants that alter evolutionarily conserved
cysteine residues; eight located within the EGF-like repeats domain and two in the LNR (Lin12-Notch)
domain. Abbreviations: aa, amino acid; ANK, ankyrin; EGF, epidermal growth factor; HD,
heterodimerization domain; LNR, Lin/NOTCH repeats; PEST, sequence rich in proline, glutamic acid,
serine, and threonine; RAM, RBP-JK–associated molecule region; TAD, transactivation domain; TM,
transmembrane domain (aa1736-1756).

Discussion

In this study, we re-analyzed WGS data available for 231 individuals with CHD, including 175

with TOF, to extend previously published results[6] using a statistical method modified to suit such

case-only data. By rescaling *de novo* mutation probabilities for singleton variants, we adapted a

burden test originally developed for *de novo* variants, and tested truncating and missense singleton

variants separately for increased burden in genes and in functionally relevant gene-sets.

Previous results suggested that ultra-rare nonsynonymous variants make an important contribution

to the genetic etiology of CHD, especially to TOF (5,6,8). Since constrained genes may be more

likely to contribute to disease, in order to maximize power, we performed multiple test correction

for all genes, and only for genes passing a constraint threshold. We assessed the validity of our

312   results by ensuring the absence of inflation when considering the burden test p-value distribution.

313   In addition, we compared burden results for CHD to a schizophrenia WGS data-set processed in

314   the same way (including variant calling and QC), to help identify potential artifacts. Finally, we

315   also retested burden by comparing results to gnomAD singletons that had been processed in the

316   same way with respect to variant effect and singleton definition.

317   **Gene burden**

318   Two genes passed a very stringent significance threshold of 0.01 after Bonferroni correction: *FLT4*

319   for truncating variants and *NOTCH1* for missense variants. Burden significance for these genes

320   was highly specific to CHD, compared to an unrelated schizophrenia sample, was confirmed by

321   the gnomAD singleton comparison analysis, and involved only individuals with TOF. The results

322   are consistent with exome sequencing results from an independent study of 829 individuals with

323   TOF, analyzed using a different approach, where excess of ultra-rare deleterious variants was

324   reported to be genome-wide significant (p-value $\leq 5\times10^{-8}$) for these two genes (8) and for *FLT4* in

325   another study restricted to LOF variants (5). These exome sequencing results serve to both help

326   validate our burden test methodology and provide independent replication, further cementing these

327   genetic findings for TOF. Collectively, the findings support study designs that focus on TOF.

328   For truncating variants, restricting to constrained genes did not result in identifying any other

329   significant genes, even when considering a more inclusive significance threshold of BH-FDR <

330   10%. Including all genes resulted in one other gene that passed BH-FDR < 10%, *CLDN9* (Claudin

331   9). *CLDN9* burden was not however confirmed by the comparison to gnomAD singleton variants

332   and the gene lacks evidence for involvement in cardiovascular development, thus at present we

333   consider this result to be likely artifactual. The results suggest that, in order to limit such artifacts,

334   considering only LOF-constrained genes may be especially important when a well matched data-

335   set (here, schizophrenia WGS) is not available. For example, artifacts can arise if *de novo* mutation

336   probabilities are derived from WGS data that were processed differently than the data available for

337   the case-only cohort (e.g., different variant calling pipeline, QC filters, and principal transcripts).

338   Also, *denovolyzer* probabilities were generated for exome analysis and adjusted for sequencing

339   depth, thus artifacts may arise in WGS studies where sequencing depth is greater. In contrast, for

340   missense variants, testing only genes passing a missense constraint threshold did not appear to be

341   beneficial. This is perhaps because missense constraint tends to be a characteristic of specific

342   protein regions rather than the full gene product and this is not adequately modelled by gnomAD

343   constraint indexes.

344   For ultra-rare missense variants, we identified slightly different sets of significant genes (BH-FDR

345   < 10%) when considering only constrained genes or all genes. *BCKDK* (Branched-chain keto acid

346   dehydrogenase kinase) was identified in both analyses; *DHH* (Desert hedgehog signaling

347   molecule) was identified only in the constrained analysis, and *KL* (Klotho) only in the all-gene

348   analysis. All displayed a similar p-value in the gnomAD singleton comparison, and only *BCKDK*

349   passed BH-FDR < 10%. *BCKDK* is a negative regulator of the branched-chain amino acids

350   catabolic pathways. In mice, complete *BCKDK* loss of function causes reduced size and

351   neurological abnormalities (19). In humans, recessive *BCKDK* loss of function causes recessive

352   'Branched-chain ketoacid dehydrogenase kinase deficiency' (OMIM: 614923), characterized by

353   autism, epilepsy, intellectual disability, and risk of developing schizophrenia (20). Alterations of

354   branched-chain amino acid metabolism have been described in relation to heart failure (21),

355   however, there is no evident link between *BCKDK* and CHD. *DHH* is required for Sertoli cells

356   development and peripheral nerve development: male Dhh-null mice are sterile and fail to produce

357   mature spermatozoa; in addition, peripheral nerves are highly abnormal (22,23). These features are

358   mirrored by the human recessive disorder '46XY partial gonadal dysgenesis, with minifascicular

359   neuropathy' (OMIM: 607080), whereas the recessive disorder '46XY sex reversal 7' (OMIM:

360   233420) does not present peripheral nerve abnormalities (24,25). No cardiac abnormalities were

361   reported for these disorders. However, *DHH* was also proposed to contribute to promoting

362   ischemia-induced angiogenesis by ensuring peripheral nerve survival (26). In humans, *KL* was

363   previously proposed as a candidate gene for TOF because of one patient with a broader 13q13

364   deletion and another patient with a narrower deletion at the same locus disrupting *KL* and *STARD13*

365   (27,28). Deficiency of *Kl* in mice has profound systemic effects, resulting in a phenotype

366   reminiscent of human ageing and characterized by reduced lifespan, stunted growth, skeletal

367   abnormalities, vascular calcification and atherosclerosis, cognitive impairment and other organ

368   alterations (29). Conversely, over-expression of *Kl* in mice protects against cardiovascular disease.

369   In addition, *Kl* is involved in the regulation of several pathways, including VEGF and Wnt (30).

370   However, considering the evidence reviewed above and that these three genes present a lower

371   fraction of singleton predicted deleterious missense variants compared to *NOTCH1*, caution is

372   needed when considering these genes as candidates for CHD/TOF. Replication in larger cohorts

373   and/or experimental follow-up is required.

374   **Functional gene-sets and candidate genes**

375   Since constrained genes would be expected to have a low rate of ultra-rare variation and thus

376   present power challenges in a cohort of this size, in addition to assessing variants by gene, we also

377   pooled variants by functional gene-sets and mouse ortholog phenotypes. To correct for strong

378   correlations introduced by highly overlapping gene-sets, which may result in inflated significance

379   after multiple test correction, we used a greedy clustering procedure to group highly overlapping

380   gene-sets; we then performed multiple test correction using a sampling-based FDR. Reassuringly,

381   given our previously published results (6), the gene-set burden analysis for truncating singleton

382   variants yielded a cluster corresponding to the VEGF pathway and blood vessel development (FDR

383   = 0), and also a cluster corresponding to abnormal vasculature (FDR = 0.008). As expected (6),

384   *FLT4* was the main gene driving these results. We additionally identified other genes that were

385   only nominally significant, but had suggestive functional or phenotypic evidence and could achieve

386   genome-wide significance in a larger cohort.

387   Although we had previously identified some of these genes (*KDR* and *FOXO1*) (6), *WNT5A* (Wnt

388   family member 5A) and *ZFAND5* (zinc finger AN1-type containing 5), were identified only in this

389   re-analysis and appear particularly promising candidates for TOF/CHD. *ZFAND5* is

390   transcriptionally activated by the platelet-derived growth factor (PDGF) pathway (31), and is

391   reported to be a member of the FoxO family signaling pathway by the NCI-Nature PID pathway

392   database. Mice homozygous for a *Zfand5* null mutation die postnatally due to widespread bleeding,

393   caused by loss of vascular smooth muscle cells (31). Mice homozygous for a *Wnt5a* null allele die

394   perinatally, with reduced growth and multiple organ system developmental abnormalities; notably,

395   the heart presents outflow tract defects and *Wnt5a* loss disrupts second heart field cell deployment;

396   heterozygous mice are apparently normal (32–34). Functional experiments in mice showed that

397   *Wnt5a* contributes to the vascular specification of cardiac progenitor cells and a role in pressure

398   overload-induced cardiac dysfuncion (35,36). In humans, heterozygous missense or homozygous

399   loss of function variants in *WNT5A* are associated with 'Robinow syndrome' (OMIM: 180700)

400   (37), which is characterized by short stature, macrocephaly, delayed bone age and limb shortening,

401   reproductive system and kidney abnormalities (38). CHD and specifically right ventricular outlet

402   obstruction are present in a fraction of the cases (39). The results also suggested other constrained

403   genes not previously identified, with evidence that supports a role in the VEGF pathway or other

404   complementary mechanisms for TOF, from human (e.g., *AKAP12*), mouse (e.g., *EPN1*, *ATF2*) or

405   both (*PKD1*) (Table 3) derived gene-sets (40–44).

406   We note that, collectively, genes *NOTCH1*, *FLT4*, *ZFAND5* and *WNT5A* present singleton variants

407   in probands with TOF but no other-CHD, representing significant enrichment (Fisher test two-

408   sided p-value = 0.01494) compared to background total sample. These account in total for about

409   11% of the individuals with TOF studied (see Supplementary Table 14).

410   One may wonder why certain VEGF pathway genes that were previously implicated in TOF using

411   manual curation of this data-set (6) were not found in the gene-set analysis in the current study.

412   There are several possible reasons. *BCAR1* was implicated by structural variation (thus not

413   analyzed in the current study), *VEGFA* does not have a defined *de novo* mutation probability in

414   *denovolyzer*, *FGD5* and *PRDM1* are not associated to any VEGF-related gene-sets among the GO

415   and pathway gene-sets used for this analysis, and *IQGAP1* was present only in a VEGF-related

416   gene-set that did not contain *FLT4* and thus did not achieve significance. If these genes were also

417   included, this would account for around 14% TOF individuals.

418   **Advantages and limitations**

419   Results from several published studies suggest that analyzing ultra-rare variant burden is a suitable

420   strategy for CHD, and especially for TOF (5,6,8), given a genetic architecture characterized by rare

421   variants of large effect, reduced penetrance and oligogenic contributions. The method we adopted

422   enables testing of ultra-rare genetic variant burden in a case-only cohort, without having access

423   either to parents to determine variant *de novo* status or to matched controls for case-control

424   analysis. This would be a relatively common circumstance for many studies, especially of rare and

425   under-funded conditions.

426   In our study design, we attempted to address issues that can produce artifacts, such as mismatch of

427   the variant calling, and/or processing pipelines, between those used for the disease data-set and for

428   the data-set supporting the calculation of *de novo* probabilities. We had the advantage of access to

429   a similarly sized sequencing data-set for an unrelated disease, processed in the same way, to aid in

430   identifying potential artifacts that may not be available for future applications of this statistical

431   burden method. Although we observed that restricting the burden analysis to genes constrained for

432   truncating variants may help minimize such artifacts, there was no apparent advantage using

433   constraint for missense variants, and we note that the findings may be disease or study specific.

434   Our primary gene burden analysis was based on *de novo* mutation probabilities rescaled to match

435   incidence of singleton variants, with probabilities defined separately for truncating and missense

436   variants. As a further confirmatory analysis to the unrelated cohort with similar WGS data, we

437   compared the singleton burden in CHD to that in gnomAD. Additional analyses using a benchmark

438   are required to establish whether one of these two methods is superior, in terms of power and

439   minimizing artifacts. The advantage of using gnomAD singletons is that, while variant calling

440   pipelines cannot be matched, other downstream processes like annotation can be matched to the

441   disease data-set of interest.

442   All results were limited by the size of the cohort available with WGS data. There were also

443   limitations to the design that are applicable to all analyses using gene-sets, including the lag in

444   updating bioinformatics databases (45) such as GO and MPO. These limitations could have had an

445   impact on that fact that there was no significant gene-set identified for missense variants. Also,

446   although the method identified highly relevant gene-set clusters for singleton truncating variants,

447   *FLT4* played a disproportionately large role in the analysis, likely influencing the fact that the

448   relatively few novel candidate genes identified largely converged on the VEGF pathway. For other

449   disorders that are even more genetically heterogeneous, the results suggest that optimizing the

450   analysis method at the gene-set level may be essential in order to identify significant results (45,46).

451   As for all studies using statistical methods to identify potential disease candidate genes, additional

452   experimental work would be required to conclusively implicate genes.

453   Sample size limitations, and the genetic architecture of TOF, also likely influenced the gene-based

454   analysis findings. Nonetheless, two genes passed a stringent Bonferroni correction, *FLT4*

455   (implicated by truncating variants) and *NOTCH1* (implicated by missense variants), consistent with

456   previous findings reported in two (5,8) and one (8) independent exome sequencing studies,

457   respectively. Notably, results of the current study indicated higher yields of ultra-rare variants in

458   these genes, perhaps related to differences in design and methods, including sequencing (WGS)

459   expected to have more uniform and complete coverage of coding regions, and perhaps the use of

460   an adult sample that would enrich for variants associated with survival. Future meta-analyses using

461   this and other data-sets, or studies using genetically relevant subsets of patients to reduce

462   heterogeneity, could reveal additional candidate genes with rare variants for TOF. In particular,

463   several gene-set clusters did not pass the multiple test correction yet appeared highly promising,

464   and could achieve significance in an expanded cohort.

465   Non-coding variants and structural variants, which can be detected using whole genome data (47),

466   were not studied. For rare non-coding variants, access to large samples of whole genome data may

467   offer interesting opportunities, especially if analyzing better understood functional elements like

468   promoters. The lack of a published mutation probability model for promoters could be

469   circumvented by performing only the gnomAD comparison analysis. For structural variants,

470   although reliably detectable using WGS (in contrast to exome sequencing), lack of a *de novo*

471   mutation model, and greater variability in variant calling pipelines, would prevent a direct gnomAD

472   comparison and represent major barriers at present.

473   **Conclusions**

474   The gene burden analysis method used, including a stringent Bonferroni correction, confirmed that

475   genes *FLT4* with ultra-rare truncating variants, and *NOTCH1* with ultra-rare deleterious missense

476   variants, are implicated in the etiology of TOF. The significant enrichment of *NOTCH1* missense

477   variants in the extracellular domain, and specifically altering cysteine residues forming disulfide

478   bonds, was also confirmed. Despite the small sample size, gene-set analysis identified ultra-rare

479   truncating variants in novel candidate genes, including *ZFAND5* and *WNT5A*, as potentially

480   implicated in the etiology of TOF. Other novel genes identified provide further confidence in the

481   importance of the VEGF pathway to TOF. While several of these candidate genes are compelling,

482   with supportive data from known functions and animal model phenotype, additional experimental

483   work and/or replication in other data-sets are required to appreciate their potential role in the

484   etiology and pathogenesis of TOF.

485   **Materials and Methods**

486   **Study participants and genome sequencing**

487   This study was authorized by the Research Ethics Boards at the University Health Network (REB

488   98-E156) (http://www.uhn.ca), and Centre for Addiction and Mental Health (REB 154/2002)

489   (http://www.camh.ca). Written consent was obtained from all participants or their legal guardians.

490   We performed genome sequencing on 231 probands (175 TOF, 49 transpositions of the great

491   arteries, 7 other CHD). DNA was sequenced on the Illumina HiSeq X system (https://www.illumin-

492   a.com/systems/sequencing-platforms/hiseq-x.html) at The Centre for Applied Genomics (TCAG)

493   (http://www.tcag.ca). Libraries were amplified by PCR prior to sequencing. Libraries were

494   assessed using Bioanalyzer DNA High Sensitivity chips and quantified by quantitative PCR using

495   Kapa Library Quantification Illumina/ABI Prism Kit protocol (KAPA Biosystems). Validated

496   libraries were pooled in equimolar quantities and paired-end sequenced on an Illumina HiSeq X

497   platform following Illumina's recommended protocol to generate paired-end reads of 150 bases in

498   length.

499   **Variant Calling, Annotation, and Truncating and Missense Variant extraction**

500   *Variant Calling*

501   The paired FASTQ reads were mapped to the GRCh37 reference sequence using the BWA-

502   backtrack algorithm (v0.7.12), and SNV and small indel variants were called using GATK (v3.7)

503   according to GATK Best Practices recommendations (48,49).

504   *Variant Annotation*

505   Variant calls were annotated using a custom pipeline based on ANNOVAR (July 2017 version)

506   (50). Allele frequencies were derived from 1000 genomes (Aug. 2015 version) (51), ExAC (Nov.

507   2015 version) (7), and gnomAD (Mar. 2017 version) (52).

508   *Classification of variants by truncating and missense effect*

509   Truncating   variants   (labelled   as   *LOF*   for   *loss   of   function*)   comprised   frameshift

510   insertions/deletions, alterations of the highly conserved intronic dinucleotide at splice sites and

511   substitutions creating a premature stop codon (stop gain). Missense variants are substitutions of

512   amino acids.

513 **Variant filters based on quality, allele frequency and effect**

514 *Allele frequency and singleton filter*

515 The burden test adopted in this study was originally developed for *de novo* variants, but we argue

516 that ultra-rare variants are not present in the general population and are likely to have arisen

517 recently from *de novo* mutations transmitted to the progeny. We defined ultra-rare singleton

518 variants as appearing only once in the CHD WGS data-set and never in population reference data-

519 sets (1000 genomes, ExAC, and gnomAD).

520 *Low quality filter*

521 We removed variants deemed to be low quality, which met at least one of these criteria: (i) low

522 sequencing depth (DP $\leq$ 10); (ii) low alternate allele read fraction or low genotype quality (for

523 heterozygous variants, alt_fraction < 0.3 or GQ $\leq$ 99, for homozygous variants, alt_fraction < 0.8

524 or GQ $\leq$ 25).

525 *Frameshift indel filter*

526 For each subject, whenever we found multiple indels on the same gene, we removed them from the

527 variants list if their cumulative size was a multiple of 3. Otherwise, we kept one of the indels as a

528 representative and removed the rest.

529 *Splice site alteration filter*

530 For insertions overlapping splice sites, we considered them as truncating variants only if the

531 alternate allele sequence did not encode a canonical AG/GT intronic dinucleotide.

532 *Principal transcript effect filter*

533   We used the APPRIS database (assembly version: GRCh37, gene dataset: RefSeq105, Oct. 2018)

534   to identify principal transcript isoforms (53) and retained only variants with an effect on a principal

535   transcript. APPRIS principal transcript identification is based on conservation, presence of protein

536   domains and other coding sequence characteristics.

537   *Final singleton counts*

538   We considered maximum only one singleton missense or truncating variant per gene per subject,

539   such that, for each variant type, the count of singleton variants in a given gene equals the count of

540   subjects with at least one singleton variant in that given gene.

541   **Gene burden analysis**

542   *De novo mutation probabilities*

543   We obtained *de novo* mutation probabilities for each gene from denovolyzeR (http://denovo-

544   lyzer.org/) (54). 1000 Genomes intergenic regions that are orthologous between humans and

545   chimps were used to derive mutation probabilities. The probabilities were based on substitution

546   type, trinucleotide context and other genome structure characteristics; in addition, they were

547   adjusted for exome sequencing depth (55).

548   *Rescale de novo mutation probability for singleton variants*

549   Since the original mutation probabilities were estimated for *de novo* variants, we applied a

550   multiplicative global scaling factor (SF), defined in equation [1], to obtain new rescaled

551   probabilities $P_{exp,(LOF\ or\ Missense),g}$; the scaling factor SF is computed so that the number of

552   predicted and observed singleton variants match.

553    $$SF = \frac{N_{Obs,(LOF\ or\ Missense)}}{\sum_{g=1,\ldots,G}\left(P_{exp,(LOF\ or\ Missense),g}\right)*N_S}$$ [1]

554   where $N_{Obs,(LOF\ or\ Missense),g}$ is the number of all observed truncating or missense singleton

555   variants; the denominator corresponds to the number of expected singletons using the original

556   unscaled probabilities: G is the total number of genes for which there is a defined mutation

557   probability (and optionally, that pass gnomAD constraint cut-offs); $P_{exp,(LOF\ or\ Missense),g}$ is the

558   expected *de novo* mutation probability for gene g with respect to truncating or missense variants;

559   and N_S is the number of subjects in the study.

560   *Binomial test*

561   Singleton truncating and missense burden was tested using a one-sided binomial test comparing

562   observed to expected singleton rates, where expected singleton rates correspond to the rescaled

563   mutation probabilities. The alternative hypothesis is defined as $P_{success} > N_{success}/N_{trials}$, i.e. that the

564   observed singleton rate for a given gene exceeds the expected rate based on rescaled mutation

565   probabilities.

566

567    $$\begin{cases} P_{Success} = P_{exp,(LOF\ or\ Missense),g} * SF \\ N_{trials} = N_S \\ N_{Success} = N_{Obs,(LOF\ or\ Missense),g} \end{cases} \qquad [2]$$

568

569   where $N_{Obs,(LOF\ or\ Missense),g}$ denotes the number of observed truncating or missense singleton

570   variants for gene g. Note that, for simplicity, we used singleton variant counts in equation [2], but

571   since we considered maximum only one singleton truncating or missense variant per subject per

572   gene, the singleton truncating or missense variant count per gene is equivalent to the count of

573   subjects with at least one truncating or missense singleton variant in that gene.

574   *gnomAD comparison analysis*

575   SNVs and indels data were obtained from the gnomAD v2.1.1 database, comprising WES (125,748

576   subjects) and WGS (15,708 subjects), after restricting to the interval list (hg19-v0-

577   wgs_evaluation_regions.v1.interval_list) used to generate the Exome Calling Intervals VCF file

578   (gnomad.genomes.r2.1.1.exome_calling_intervals.sites.vcf.bgz).   Singleton   variants   were

579   identified by using the allele counts provided in the gnomAD VCF file. Singletons were annotated

580   using the same ANNOVAR-based pipeline, followed by the same effect filters as in the main

581   analysis (including the selection of the same principal transcript) and finally categorized as

582   truncating or missense. Genes were tested for singleton burden by comparing CHD WGS

583   singletons to gnomAD singletons using a two-sided Fisher's Exact Test, and specifically by

584   constructing the 2×2 contingency matrix with counts: (a) CHD singletons in the gene of interest,

585   (b) CHD singletons in other genes, (c) gnomAD singletons in the gene of interest, (d) gnomAD

586   singletons in other genes; truncating and missense singletons were tested separately. For CHD,

587   only maximum one singleton per subject was considered (as in the main analysis).

588   *Multiple test correction*

589   For gene burden analyses, multiple test correction was performed using the Benjamini-Hochberg

590   False Discovery Rate (*BH-FDR*), as implemented in the R function *p.adjust*, and Bonferroni

591   correction, by multiplying the p-value by the number of genes tested. For both corrections, we

592   considered all genes with a defined probability, or all genes with a defined probability and passing

593   constraint cut-offs (o/e gnomAD score < 0.35 for truncating variants and o/e gnomAD score < 0.75

594   for missense variants).

595   **Gene-set burden analysis**

596   *Gene-set resources*

597   Gene-sets were derived from Gene Ontology (GO) annotations as provided by the Bioconductor

598   package org.Hs.eg.db v3.5 (9), BioCarta pathways (http://cgap.nci.nih.gov/Pathways/BioCarta-

599   Pathways/), KEGG pathways (http://www.genome.jp/kegg/) retrieved using the KEGG API (10),

600   REACTOME pathways (11), and National Cancer Institute (NCI) pathways (https://cactus.nci.nih-

601   .gov/download/nci/). Gene-sets corresponding to phenotypes of mouse orthologues were derived

602   from MPO gene annotations as provided by MGI (12).

603   *Gene-set filters*

604   We retained only the gene-sets with more than 5 genes and less than 100. Smaller gene-sets are

605   detrimental for power. Larger gene-sets are usually removed because they are overly general.

606   Considering the specific gene-level burden signal distribution observed for this data-set,

607   characterized by the presence of two "highly concentrated" burden genes (*FLT4* and *NOTCH1*),

608   some larger gene-sets could exceed the expected singleton rate just because of the presence of one

609   of these two genes. In addition, larger gene-sets are less suitable for the binomial test strategy, since

610   they are more likely to present with more than one singleton variant per subject and to contain

611   genes with heterogeneous mutation probabilities, which is detrimental when pooling counts (13).

612   For the analyses using a given gene constraint cut-off, we removed gene-sets with less than two

613   genes passing the constraint cut-offs.

614   *Binomial Test*

615   For the gene-set analysis, we used a binomial test (equation [3]) to compare the number of observed

616   and expected singleton variant in the gene-set, similar to the gene burden analysis. We additionally

617   ensured not to count more than one truncating or missense singleton per gene-set per subject.

618
$$\begin{cases} P_{Success} = \sum_{g \in GeneSet} P_{exp,(LOF \text{ or } Missense),g} * SF \\ N_{trials} = N_S \\ N_{Success} = \sum_{s=1,\dots,S} min\left(\sum_{g \in GeneSet} N_{Obs,(LOF \text{ or } Missense),g,s}, 1\right) \end{cases}$$ [3]

619 where $GeneSet$ represents the set of all genes in a particular gene-set; S are the study subjects; and

620 $N_{Obs,(LOF \text{ or } Missense),g,s}$ is the number of observed missense or truncating singleton variants in a

621 particular gene for subject S.

622 *Greedy step-down aggregation method to correct for gene-set correlations*

623 We addressed the problem of gene-set correlations, which are introduced by large gene overlaps

624 between related gene-sets, by using a greedy step-down clustering approach, similar to what was

625 adopted for highly correlated CNV locus gene testing in the *Marshall et al.* study (46). The

626 algorithm follows these steps, starting from an input list of gene-sets sorted by the singleton burden

627 binomial p-value:

628   1) Select the gene-set with the most significant p-value (i.e. the smallest p-value);

629   2) Identify other gene-sets that are highly correlated to the selected gene-set, using the *Jaccard*

630   similarity:

631       $\dfrac{|gs_i \cap gs_j|}{|gs_i \cup gs_j|}$ [4]

632   where gs_i and gs_j are the sets of singleton variants for gene-sets i and j, respectively. $|\;\;|$

633   is the number of singleton variants in the corresponding set.

634   3) Cluster gene-sets that have *Jaccard* similarity > 0.5 with the selected gene-set; these gene-

635   sets will not be considered for the multiple test correction calculation, only the selected

636   gene-set will be used (i.e. the p-value from the selected gene-set will be used as the p-value

637   for the gene-set cluster). Finally, remove the selected gene-set and its clustered gene-sets

638   from the sorted list.

639   Steps 1-3 are executed until no gene-set is left in the list.

640   **Resampling-based FDR**

641   Observed missense or truncating singleton variants are resampled based on each gene's rescaled

642   mutation probability (equation [1]), while maintaining the same total number of observed missense

643   or truncating singleton variants. After this step, gene-sets are tested as described in the previous

644   section. Finally, for each given p-value threshold p, the FDR is calculated as follows, considering

645   only gene-sets selected by the greedy step-down aggregation procedure:

646    $$FDR_p = \frac{mean_{i=1,...,1000}\left(N_{gs}^{permutation_i}\right)}{N_{gs}^{real}} \qquad [5]$$

647   where FDR_p is the FDR q-value for a given p-value threshold p, N_{gs}^{real} is the number of gene-

648   sets with binomial p-value ≤ p, and $N_{gs}^{permutation_i}$ corresponds to the number of gene-sets with

649   binomial p-value ≤ p at iteration i. As stated in the formula, we used 1,000 sampling iterations.

660   **Legends to Supplementary Figures and Tables**
661   **Supplementary Figures**

662   **Supplementary Figure 1**
663   **Relation between gnomAD genetic constraint indexes.**
664   (A) Relationship between pLI (x axis, discretized in three bins) and the ratio of observed/expected (o/e)
665   truncating variants (y axis). pLI > 0.9 has often been used as haploinsufficiency cutoff for clinical variant
666   interpretation, and gnomAD suggests using the upper bound of the o/e confidence interval < 0.35 for a
667   similar use. We preferred using a point estimate <= 0.35 to be more inclusive, i.e. including genes with
668   more moderate haploinsufficiency. For our analysis, we have considered genes with o/e score < 0.35. (B)
669   Relationship between the missense constraint z-score (x axis, discretized in two bins) and the ratio of
670   observed/expected missense variants (y axis). For our analysis, we have considered genes with o/e score <
671   0.75, which roughly corresponds to a z-score > 2, which in turn corresponds to a constraint p-value of
672   0.02275.
673
674   **Supplementary Figure 2**
675   **Relation between the number of singleton truncating variants per gene in the CHD data-set and in**
676   **gnomAD.**
677   The distribution (across genes) of the number of singleton truncating variants per gene is shown as an
678   overlaid boxplot and violin plot for singletons in gnomAD (x axis), stratified by the the number of
679   singleton variant in the CHD data-set (y axis); each dot represents a gene. The dashed line represents the
680   linear regression predictions, which are appear unreliable because of outliers and the small number of
681   unique CHD singleton counts. Only *FLT4* has 7 truncating singletons, but the trend for other strata
682   suggests that this is in large excess of singletons observed in gnomAD. Note that CHD singletons are not
683   observed in gnomAD, whereas gnomAD singletons are observed only once in gnomAD.
684
685   **Supplementary Figure 3**
686   **Relation between the number of singleton missense variants per gene in the CHD data-set and in**
687   **gnomAD.**
688   The distribution (across genes) of the number of singleton missense variants per gene is shown as an
689   overlaid boxplot and violin plot for singletons in gnomAD (x axis), stratified by the the number of
690   singleton variant in the CHD data-set (y axis); each dot represents a gene. The dashed line represents the
691   linear regression predictions, which appear robust. *KL* and *DHH* overlap with the lowest percentiles of the
692   distribution, whereas *BCKDK* is lower than any observed value; only NOTCH1 has 8 missense singletons,
693   but the trend for other strata suggests that this is in excess of singletons observed in gnomAD. Note that
694   CHD singletons are not observed in gnomAD, whereas gnomAD singletons are observed only once in
695   gnomAD.
696
697   **Supplementary Figure 4**
698   **QQ-plots and p-value CHD/SZ scatterplot for the gene burden analysis restricted to constrained**
699   **genes.**
700   (A) and (C) show the quantile-quantile (QQ) plots for gene burden p-values obtained for truncating
701   singletons restricted to constrained genes (gnomAD o/e < 0.35, A) or missense singletons restricted to
702   constrained genes (gnomAD o/e < 0.75, B). Only a few genes present p-values deviating from the null
703   distribution, suggesting absence of systematic p-value inflation. (B) and (D) show scatterplots of the
704   nominal p-values obtained for the gene burden analysis of truncating or missense singleton variants in
705   constrained genes, comparing CHD (y axis) versus schizophrenia (x axis). The most significant genes for
706   CHD are typically not significant for SZ, suggesting the absence of systematic confounders.
707

708   **Supplementary Figure 5**
709   **Cytoscape enrichment map for the gene-sets with significant burden of singleton truncating variants**
710   **in constrained genes.**
711   An enrichment map visualizes gene-sets as a network based on their overlaps. Nodes correspond to gene-
712   sets from the gene-set cluster with significant (FDR < 10%) burden for truncating singleton variants in
713   constrained genes, and edges correspond to the degree of overlap between gene-sets. Nodes are colored
714   based on the burden nominal p-value, with darker red corresponding to more significant gene-sets. Edge
715   thickness is proportional to the jaccard index obtained by considering singleton truncating variants as set
716   elements; only edges corresponding to jaccard index > 0.5 are displayed. Gene-set sub-clusters are
717   suggested by automated network layout. Gene Ontology and pathways (A) are shown separately from
718   mouse phenotypes (B).
719
720   **Supplementary Tables**

721   **Supplementary Table 1**
722   **Ultra-rare singleton variants observed for 231 CHD samples including gnomAD o/e**
723   **constraint scores**
724
725   **Supplementary Table 2**
726   **Gene burden statistics for CHD singleton truncating variants**
727
728   **Supplementary Table 3**
729   **Gene burden statistics for CHD singleton missense variants**
730
731   **Supplementary Table 4**
732   **Truncating variants observed in both SZ and CHD cohorts**
733
734   **Supplementary Table 5**
735   **Gene burden statistics obtained by comparing CHD singletons to gnomAD singletons and**
736   **tested using Fisher's Exact Test**
737
738   **Supplementary Table 6**
739   **Missense Variants observed in both SZ and CHD cohorts**
740
741   **Supplementary Table 7**
742   **Burden statistics for GO and pathway gene-set clusters and truncating variants in**
743   **constrained genes**
744
745   **Supplementary Table 8**
746   **Burden statistics for GO and pathway gene-sets and singleton truncating variants in**
747   **constrained genes**
748
749   **Supplementary Table 9**
750   **Burden statistics for mouse phenotype gene-set clusters and singleton truncating variants in**
751   **constrained genes**
752
753

754 **Supplementary Table 10**
755 **Burden statistics for mouse phenotype gene-sets and singleton truncating variants in**
756 **constrained genes**
757
758 **Supplementary Table 11**
759 **Burden statistics for GO and pathway gene-set clusters and singleton missense variants in**
760 **constrained genes**
761
762 **Supplementary Table 12**
763 **Burden statistics for GO and pathway gene-sets and singleton missense variants in**
764 **constrained genes**
765
766 **Supplementary Table 13**
767 **NOTCH1, BCKDK, DHH and KL missense variants details**
768
769 **Supplementary Table 14**
770 **Patient phenotype and family history for selected deleterious missense and truncating**
771 **variants**
772

773 **References**

774 1.   Glidewell J, Grosse SD, Riehle-Colarusso T, Pinto N, Hudson J, Daskalov R, et al. Actions
775   in Support of Newborn Screening for Critical Congenital Heart Disease — United States,
776   2011–2018. MMWR Morb Mortal Wkly Rep. 2019;68(5):107–11.

777 2.   Zaidi S, Brueckner M. Genetics and Genomics of Congenital Heart Disease. Circ Res.
778   2017;120(6):923–40.

779 3.   Morgenthau A, Frishman WH. Genetic Origins of Tetralogy of Fallot. Cardiol Rev.
780   2018;26(2):86–92.

781 4.   Mercer-Rosa L, Paridon SM, Fogel MA, Rychik J, Tanel RE, Zhao H, et al. 22q11.2
782   deletion status and disease burden in children and adolescents with tetralogy of Fallot. Circ
783   Cardiovasc Genet. 2015;8(1):74–81.

784 5.   Jin SC, Homsy J, Zaidi S, Lu Q, Morton S, DePalma SR, et al. Contribution of rare
785   inherited and de novo variants in 2,871 congenital heart disease probands. Nat Genet
786   [Internet]. 2017 Nov 9;49(11):1593–601. Available from:
787   http://dx.doi.org/10.1038/ng.3970

788 6.   Reuter MS, Jobling R, Chaturvedi RR, Manshaei R, Costain G, Heung T, et al.
789   Haploinsufficiency of vascular endothelial growth factor related signaling genes is
790   associated with tetralogy of Fallot. Genet Med [Internet]. 2019;21(4):1001–7. Available
791   from: http://dx.doi.org/10.1038/s41436-018-0260-9

792 7.   Lek M, Karczewski KJ, Minikel E V., Samocha KE, Banks E, Fennell T, et al. Analysis of
793   protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285–91.

794 8.   Page DJ, Miossec MJ, Williams SG, Monaghan RM, Fotiou E, Cordell HJ, et al. Whole
795   Exome Sequencing Reveals the Major Genetic Contributors to Nonsyndromic Tetralogy of

796   Fallot. Circ Res. 2019;124(4):553–63.

797   9.   Carlson M. (2019). org.Hs.eg.db: Genome wide annotation for Human. R package version
798   3.8.2. R package version 3.8.2. 2019.

799   10.   Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: New perspectives on
800   genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45(D1):D353–61.

801   11.   Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The
802   Reactome Pathway Knowledgebase. Nucleic Acids Res [Internet]. 2018;46(D1):D649–55.
803   Available from:
804   http://www.ncbi.nlm.nih.gov/pubmed/29145629%0Ahttp://www.pubmedcentral.nih.gov/ar
805   ticlerender.fcgi?artid=PMC5753187

806   12.   Bult CJ, Blake JA, Smith CL, Kadin JA, Richardson JE, Anagnostopoulos A, et al. Mouse
807   Genome Database (MGD) 2019. Nucleic Acids Res. 2019;47(D1):D801–6.

808   13.   Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, et al.
809   Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA,
810   Cytoscape and EnrichmentMap. Nat Protoc. 2019;14(2):482–517.

811   14.   Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for
812   genomes. Nat Protoc [Internet]. 2016 Jan 3;11(1):1–9. Available from:
813   http://www.nature.com/articles/nprot.2015.123

814   15.   Ivan A. Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E. Ramensky, Anna
815   Gerasimova, Peer Bork, Alexey S. Kondrashov and SRS. PolyPhen2 original paper.
816   October. 2010;7(4):248–9.

817   16.   Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations:
818   Application to cancer genomics. Nucleic Acids Res. 2011;39(17):37–43.

819   17.   Wouters MA, Rigoutsos I, Chu CK, Feng LL, Sparrow DB, Dunwoodie SL. Evolution of
820   distinct EGF domains with specific functions. Protein Sci. 2005;14(4):1091–103.

821   18.   Gordon WR, Arnett KL, Blacklow SC. The molecular logic of Notch: Biochemical
822   Perspective. Cell. 2009;121(Pt 19):3109–19.

823   19.   Joshi MA, Jeoung NH, Obayashi M, Hattab EM, Brocken EG, Liechty EA, et al. Impaired
824   growth and neurological abnormalities in branched-chain α-keto acid dehydrogenase
825   kinase-deficient mice. Biochem J [Internet]. 2006 Nov 15;400(1):153–62. Available from:
826   http://www.biochemj.org/cgi/doi/10.1042/BJ20060869

827   20.   Novarino G, El-Fishawy P, Kayserili H, Meguid N a, Scott EM, Schroth J, et al. Mutations
828   in BCKD-kinase Lead to a Potentially Treatable Form of Autism with Epilepsy. Science
829   (80-) [Internet]. 2012 Oct 19;338(6105):394–7. Available from:
830   http://www.sciencemag.org/cgi/doi/10.1126/science.1224631

831   21.   Sun H, Olson KC, Gao C, Prosdocimo DA, Zhou M, Wang Z, et al. Catabolic Defect of
832   Branched-Chain Amino Acids Promotes Heart Failure. Circulation [Internet]. 2016 May
833   24;133(21):2038–49. Available from:
834   https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.115.020226

835   22.   Bitgood MJ, Shen L, McMahon AP. Sertoli cell signaling by Desert hedgehog regulates

836   the male germline. Curr Biol. 1996;6(3):298–304.

837 23. Parmantier E, Turmaine M, Namini SS, Chakrabarti L, Jessen KR, Mirsky R, et al.
838   Schwann cell-derived desert hedgehog controls the development of peripheral nerve
839   sheaths. Neuron. 1999;23(4):713–24.

840 24. Umehara F, Tate G, Itoh K, Yamaguchi N, Douchi T, Mitsuya T, et al. A novel mutation
841   of desert hedgehog in a patient with 46, XY partial gonadal dysgenesis accompanied by
842   minifascicular neuropathy. Am J Hum Genet. 2000;67(5):1302–5.

843 25. Canto P, Söderlund D, Reyes E, Méndez JP. Mutations in the Desert hedgehog (DHH)
844   gene in patients with 46,XY complete pure gonadal dysgenesis. J Clin Endocrinol Metab.
845   2004;89(9):4480–3.

846 26. Renault MA, Chapouly C, Yao Q, Larrieu-Lahargue F, Vandierdonck S, Reynaud A, et al.
847   Desert hedgehog promotes ischemia-induced angiogenesis by ensuring peripheral nerve
848   survival. Circ Res. 2013;112(5):762–70.

849 27. Costain G, Silversides CK, Marshall CR, Shago M, Costain N, Bassett AS. 13q13.1-q13.2
850   deletion in tetralogy of Fallot: Clinical report and a literature review. Int J Cardiol
851   [Internet]. 2011;146(2):134–9. Available from:
852   http://dx.doi.org/10.1016/j.ijcard.2010.05.070

853 28. Costain G, Lionel AC, Ogura L, Marshall CR, Scherer SW, Silversides CK, et al. Genome-
854   wide rare copy number variations contribute to genetic risk for transposition of the great
855   arteries. Int J Cardiol [Internet]. 2016;204:115–21. Available from:
856   http://dx.doi.org/10.1016/j.ijcard.2015.11.127

857 29. Kuro-o M, Matsumura Y, Aizawa H, Kawaguchi H, Suga T, Utsugi T, et al. Mutation of
858   the mouse klotho gene leads to a syndrome resembling ageing. Nature [Internet]. 1997
859   Nov;390(6655):45–51. Available from: http://www.nature.com/articles/36285

860 30. Mencke R, Hillebrands JL. The role of the anti-ageing protein Klotho in vascular
861   physiology and pathophysiology. Ageing Res Rev [Internet]. 2017;35:124–46. Available
862   from: http://dx.doi.org/10.1016/j.arr.2016.09.001

863 31. Schmahl J, Raymond CS, Soriano P. PDGF signaling specificity is mediated through
864   multiple immediate early genes. Nat Genet. 2007;39(1):52–60.

865 32. Yamaguchi TP, Bradley A, McMahon AP, Jones S. A Wnt5a pathway underlies outgrowth
866   of multiple structures in the vertebrate embryo. Development [Internet].
867   1999;126(6):1211–23. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10021340

868 33. Schleiffarth JR, Person AD, Martinsen BJ, Sukovich DJ, Neumann A, Baker CVH, et al.
869   Wnt5a is required for cardiac outflow tract septation in mice. Pediatr Res. 2007;61(4):386–
870   91.

871 34. Sinha T, Li D, Théveniau-Ruissy M, Hutson MR, Kelly RG, Wang J. Loss of Wnt5a
872   disrupts second heart field cell deployment and may contribute to OFT malformations in
873   DiGeorge syndrome. Hum Mol Genet. 2015;24(6):1704–16.

874 35. Reichman DE, Park L, Man L, Redmond D, Chao K, Harvey RP, et al. Wnt inhibition
875   promotes vascular specification of embryonic cardiac progenitors. Development.
876   2018;145(1):dev159905.

877   36.   Wang Y, Sano S, Oshima K, Sano M, Watanabe Y, Katanasaka Y, et al. Wnt5a-Mediated
878   Neutrophil Recruitment Has an Obligatory Role in Pressure Overload-Induced Cardiac
879   Dysfunction. Circulation. 2019;140(6):487–99.

880   37.   Birgmeier J, Esplin ED, Jagadeesh KA, Guturu H, Wenger AM, Chaib H, et al. Biallelic
881   loss-of-function WNT5A mutations in an infant with severe and atypical manifestations of
882   Robinow syndrome. Am J Med Genet Part A. 2018;176(4):1030–6.

883   38.   Person AD, Beiraghi S, Sieben CM, Hermanson S, Neumann AN, Robu ME, et al.
884   WNT5A mutations in patients with autosomal dominant Robinow syndrome. Dev Dyn.
885   2010;239(1):327–37.

886   39.   S Atalay, B Ege, A Imamoğlu, E Suskan, B Ocal HG. Congenital heart disease and
887   Robinow syndrome [Internet]. Clinical Dysmorphology; 1993. p. 3. Available from:
888   https://europepmc.org/abstract/med/8287182

889   40.   Benz PM, Ding Y, Stingl H, Loot AE, Zink J, Wittig I, et al. AKAP12 deficiency impairs
890   VEGF-induced endothelial cell migration and sprouting. Acta Physiol. 2019;(May
891   2019):1–15.

892   41.   Wang Y, Hoeppner LH, Angom RS, Wang E, Dutta S, Doeppler HR, et al. Protein kinase
893   D up-regulates transcription of VEGF receptor-2 in endothelial cells by suppressing
894   nuclear localization of the transcription factor AP2β. J Biol Chem. 2019;294(43):15759–
895   67.

896   42.   Villalobos E, Criollo A, Schiattarella GG, Altamirano F, French KM, May HI, et al.
897   Fibroblast Primary Cilia Are Required for Cardiac Fibrosis. Circulation.
898   2019;139(20):2342–57.

899   43.   Fearnley GW, Odell AF, Latham AM, Mughal NA, Bruns AF, Burgoyne NJ, et al. VEGF-
900   A isoforms differentially regulate ATF-2-dependent VCAM-1 gene expression and
901   endothelial-leukocyte interactions. Mol Biol Cell. 2014;25(16):2509–21.

902   44.   Tessneer KL, Pasula S, Cai X, Dong Y, Mcmanus J, Liu X, et al. Genetic reduction of
903   vascular endothelial growth factor receptor 2 rescues aberrant angiogenesis caused by
904   epsin deficiency. Arterioscler Thromb Vasc Biol. 2014;34(2):331–7.

905   45.   Tomczak A, Mortensen JM, Winnenburg R, Liu C, Alessi DT, Swamy V, et al.
906   Interpretation of biological experiments changes with evolution of the Gene Ontology and
907   its annotations. Sci Rep. 2018;8(1):1–10.

908   46.   Marshall CR, Howrigan DP, Merico D, Thiruvahindrapuram B, Wu W, Greer DS, et al.
909   Contribution of copy number variants to schizophrenia from a genome-wide study of
910   41,321 subjects. Nat Genet [Internet]. 2017 Jan 21;49(1):27–35. Available from:
911   http://www.nature.com/articles/ng.3725

912   47.   Reuter MS, Walker S, Thiruvahindrapuram B, Whitney J, Cohn I, Sondheimer N, et al.
913   The personal genome project Canada: Findings from whole genome sequences of the
914   inaugural 56 participants. Cmaj. 2018;190(5):E126–36.

915   48.   Depristo MA, Banks E, Poplin R, Garimella K V., Maguire JR, Hartl C, et al. A
916   framework for variation discovery and genotyping using next-generation DNA sequencing
917   data. Nat Genet. 2011;43(5):491–501.

918 49. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A,
919   et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit
920   Best Practices Pipeline. In: Current Protocols in Bioinformatics [Internet]. Hoboken, NJ,
921   USA: John Wiley & Sons, Inc.; 2013. p. 11.10.1-11.10.33. Available from:
922   http://doi.wiley.com/10.1002/0471250953.bi1110s43

923 50. Wang K, Li M, Hakonarson H. ANNOVAR: Functional annotation of genetic variants
924   from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):1–7.

925 51. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An
926   integrated map of structural variation in 2,504 human genomes. Vol. 526, Nature. 2015. p.
927   75–81.

928 52. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. Variation
929   across 141,456 human exomes and genomes reveals the spectrum of loss-of-function
930   intolerance across human protein-coding genes. bioRxiv [Internet]. 2019;531210.
931   Available from: https://www.biorxiv.org/content/10.1101/531210v1

932 53. Rodriguez JM, Rodriguez-Rivas J, Di Domenico T, Vázquez J, Valencia A, Tress ML.
933   APPRIS 2017: Principal isoforms for multiple gene sets. Nucleic Acids Res.
934   2018;46(D1):D213–7.

935 54. Ware JS, Samocha KE, Homsy J, Daly MJ. Interpreting de novo Variation in Human
936   Disease Using denovolyzeR. In: Current Protocols in Human Genetics [Internet].
937   Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2015. p. 7.25.1-7.25.15. Available from:
938   http://doi.wiley.com/10.1002/0471142905.hg0725s87

939 55. Gibbs RA, McGrath LM, Stevens C, Boerwinkle E, Rehnström K, Palotie A, et al. A
940   framework for the interpretation of de novo mutation in human disease. Nat Genet
941   [Internet]. 2014;46(9):944–50. Available from: http://dx.doi.org/10.1038/ng.3050

942