HGGA, Volume *2*

Supplemental information

Common deletion variants causing protocadherin-α deficiency contribute to the complex genetics of BAV and left-sided congenital heart disease

Polakit Teekakirikul, Wenjuan Zhu, George C. Gabriel, Cullen B. Young, Kylia Williams, Lisa J. Martin, Jennifer C. Hill, Tara Richards, Marie Billaud, Julie A. Phillippi, Jianbin Wang, Yijen Wu, Tuantuan Tan, William Devine, Jiuann-huey Lin, Abha S. Bais, Jonathan Klonowski, Anne Moreau de Bellaing, Ankur Saini, Michael X. Wang, Leonid Emerel, Nathan Salamacha, Samuel K. Wyman, Carrie Lee, Hung Sing Li, Anastasia Miron, Jingyu Zhang, Jianhua Xing, Dennis M. McNamara, Erik Fung, Paul Kirshbom, William Mahle, Lazaros K. Kochilas, Yihua He, Vidu Garg, Peter White, Kim L. McBride, D. Woodrow Benson, Thomas G. Gleason, Seema Mital, and Cecilia W. Lo

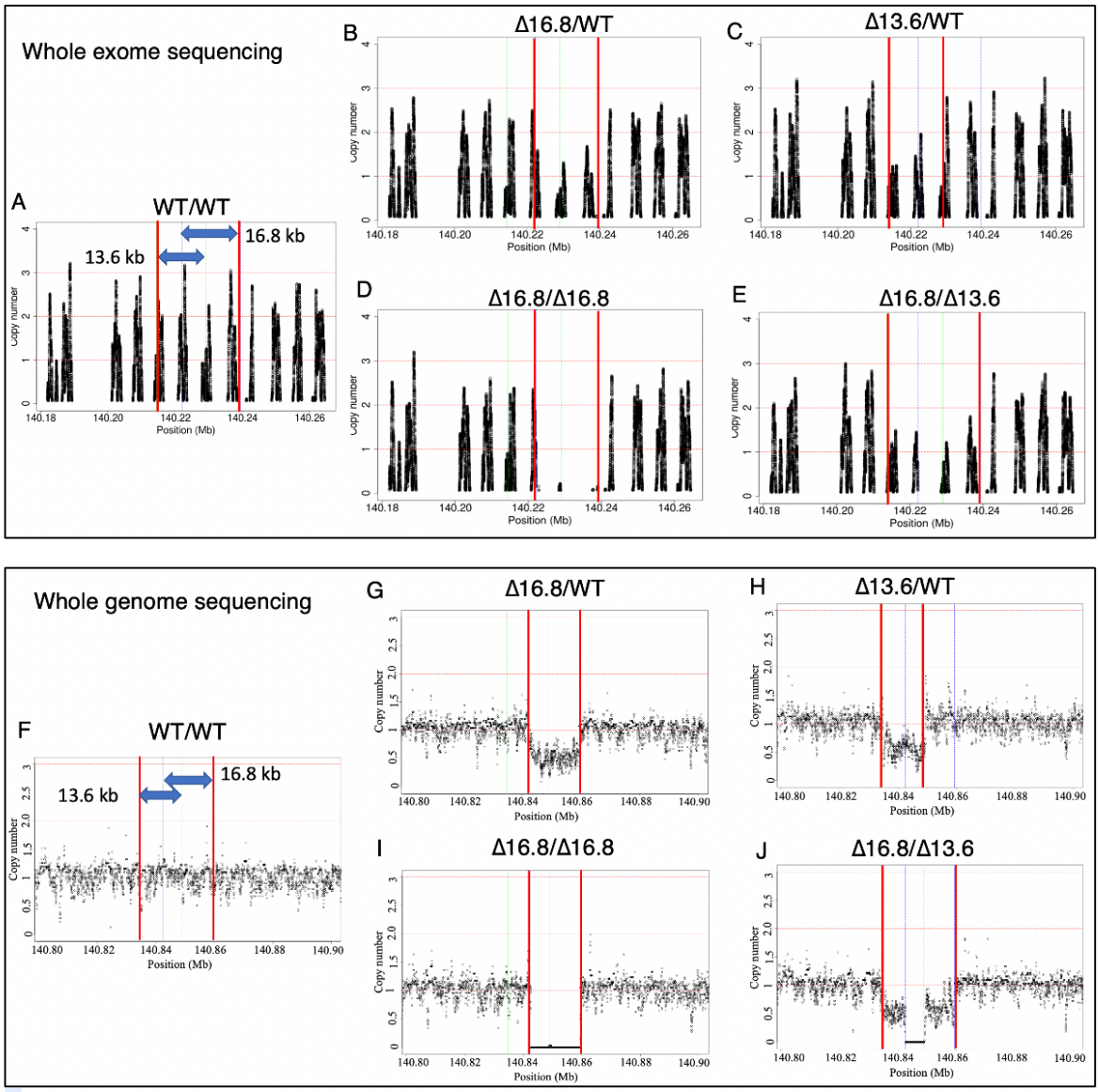

Figure S1. Read depths in the *PCHDA* gene cluster allow determination of *PCDHA* delCNV genotypes from WES and WGS data

Shown are WES (A) and WGS (F) read depths over the region spanning the 13.6 kb and 16.8 kb *PCHDA8-10* deletion interval in an individual with wildtype *PCDHA* genotype. Blue double arrows denote region comprising the *PCHDA7-9* interval. The red line delineate the boundary of the 13.6/16.8 kb delCNV. In (B-E) are examples of read depth in the WES data and (G-J) WGS data representing the different *PCHDA* delCNV genotypes as indicated. The red lines denote the boundary of the deletions associated with the 16.8 or 13.6 kb *PCHDA* delCNVs. Note the complete absence of reads in individuals homozygous for the 16.8 kb delCNV. No individuals are found homozygous for the 13.6 kb delCNV, as the 13.6 kb delCNV are much more rare. Note coordinates of 13.6kb (nsv4684081) and 16.8kb (nsv4655880) deletion are chr5:140214374-140228422 and chr5:140222140-140238927 in human genome GRCh37 reference or chr5:140834789-140848837 and chr5:140842555-140859342 in human genme GRCh38/hg38 reference.

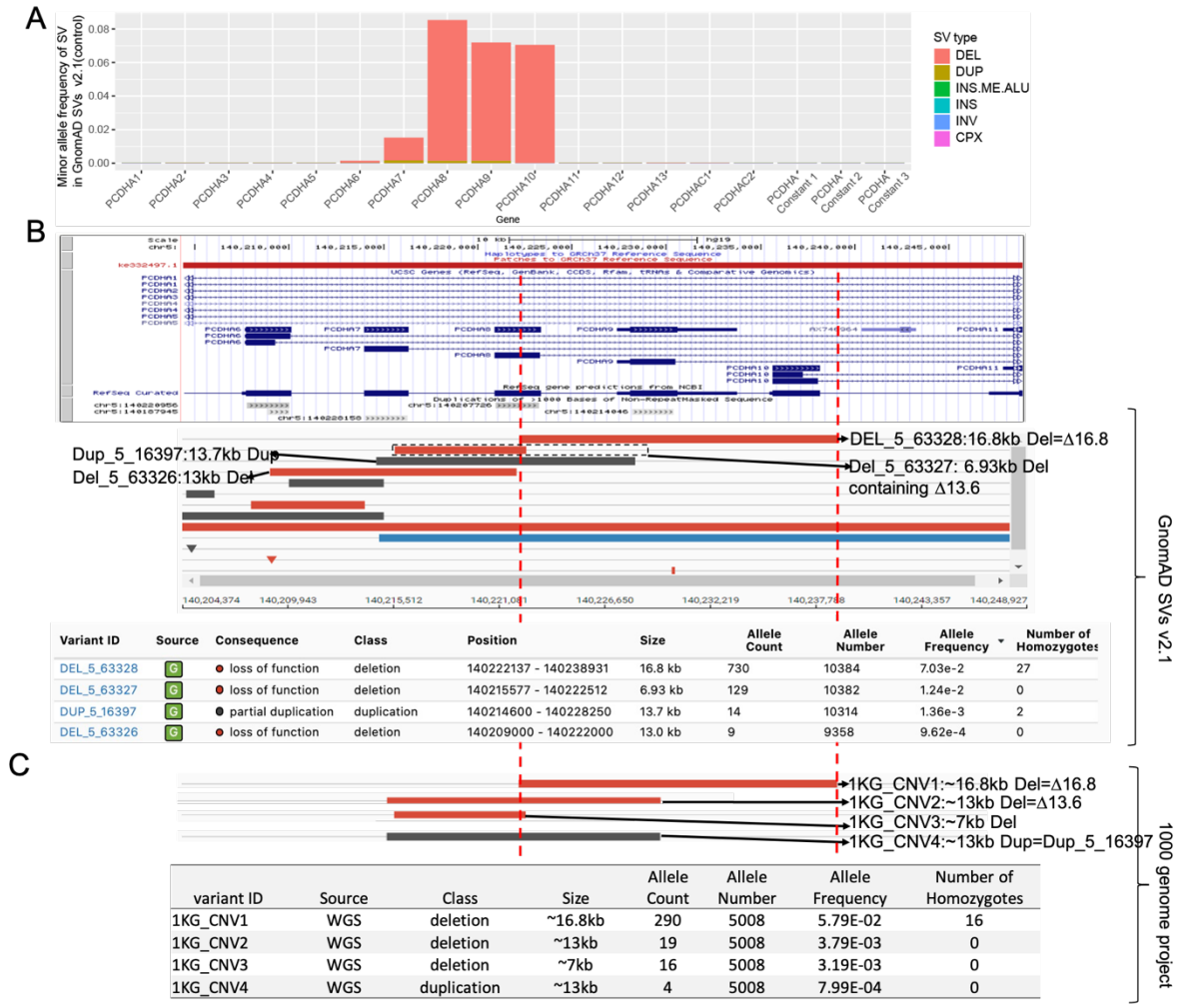

Figure S2. CNV in *PCDHA* gene cluster detected in GnomAD and 1KG database.

The two common delCNVs in the *PCDHA* gene cluster analyzed in our study comprise: Δ16.8kb deletion (nsv4655880) and Δ13.6kb deletion (nsv4684081). The coordinates of the Δ16.8kb and Δ13.6kb (hg19) delCNV correspond to chr5:140222140-140238927 and chr5:140214374-140228422 (or chr5:140215019-140229049), respectively. For Δ13.6kb, precise breakpoints cannot be identified due to segmental duplications around 5' and 3' breakpoints of the 13.6kb delCNV.

(A) Minor allele frequency of structural variants (SVs) from GnomAD SVs v2.1 control database (5,192 samples) for *PCDHA* gene family. This shows *A7, A8, A9* and *A10* genes in the *PCDHA* cluster are prone to deletion. DEL: Deletion; DUP: Duplication; INS: Insertion; INV: Inversion; CPX: Complex SV; INS:ME:ALU: Alu element insertion.

(B) Allele-frequency for top 4 SVs from GnomAD SV v2.1 control database (Collins et al. [1]) covering *PCDHA7-PCDHA10*. SV image and variant information were downloaded from

GnomAD Browser: https://gnomad.broadinstitute.org/gene/ENSG00000204962?dataset=gnomad_sv_r2_1_controls.

(C) Allele-frequency for top 4 SVs in *PCDHA* cluster from 1000 genome project (1KG). 1KG 2504 high coverage whole genome sequencing (WGS) data (~30X) from the New York Genome Center was used to manually identify the del CNVs. The 16.8kb deletion CNV (DEL_5_63328 in GnomAD, 1KG_CNV1 in 1KG and nsv4655880 in dbVAR) was common in both GnomAD and 1KG database. A ~7kb deletion (DEL_5_63327) common in each population in GnomAD database likely encompasses multiple SVs, including the ~13.6 kb deletion (1KG_CNV2 in 1KG=Δ13.6kb, nsv4684081 in dbVAR, covering PCDHA7, PCDHA8 and PCDHA9), as this 7kb deletion (MAF=0.0121) is only found in African population in 1000 genome project, and ~13.6kb deletion is a variant (MAF=0.0119) found mostly only in European population in 1000 genome project. The issue noted here is likely related to the limitation of the three software used for SV detection comprising cn.MOPS (Klambauer et al. [2]), Manta (Chen et al. [3]) and DELLY (Rausch et al. [4]).

1. Collins, R.L., Brand, H., Karczewski, K.J., Zhao, X., Alfoldi, J., Francioli, L.C., Khera, A.V., Lowther, C., Gauthier, L.D., Wang, H., et al. (2020). A structural variation reference for medical and population genetics. Nature 581, 444-451.
2. Klambauer, G., Schwarzbauer, K., Mayr, A., Clevert, D.A., Mitterecker, A., Bodenhofer, U., and Hochreiter, S. (2012). cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. Nucleic Acids Res 40, e69.
3. Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Kallberg, M., Cox, A.J., Kruglyak, S., and Saunders, C.T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics 32, 1220-1222.
4. Rausch, T., Zichner, T., Schlattl, A., Stutz, A.M., Benes, V., and Korbel, J.O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28, i333-i339.

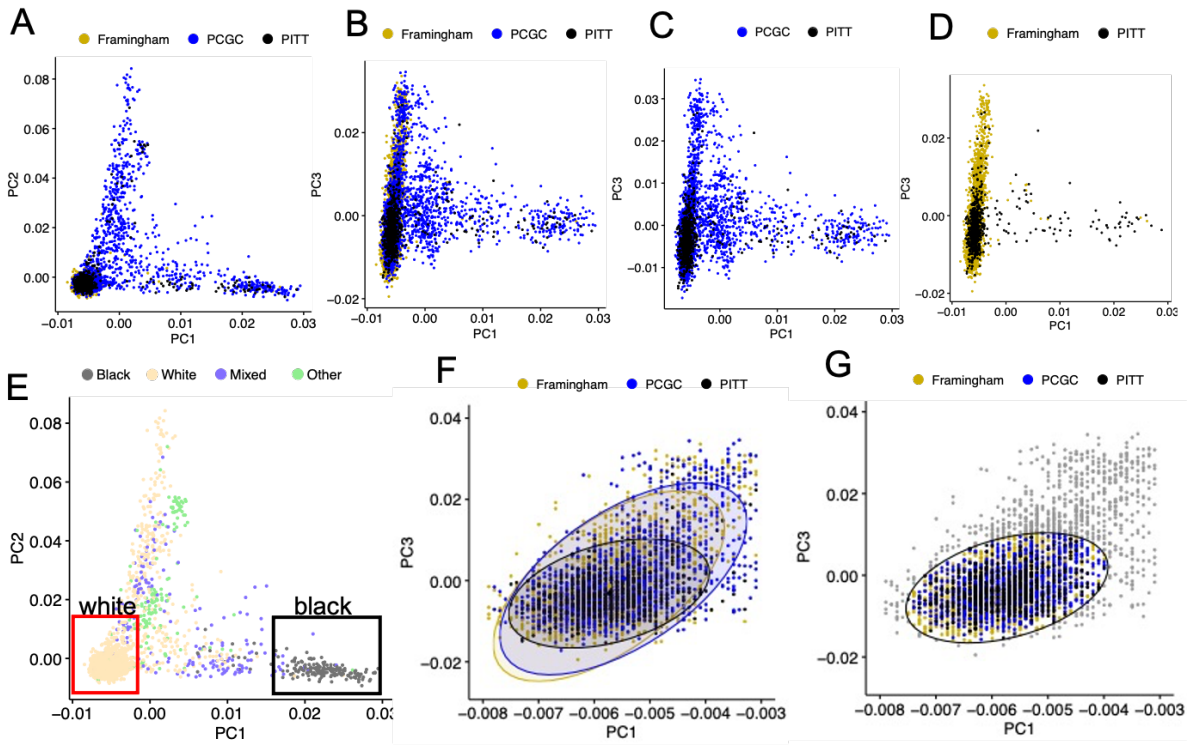

Figure S3. Ancestry matching with principal component analysis

(A-D) PCA plots of WES data from PCGC and Pittsburgh patients and Framingham control subjects. PC1 and PC2 are shown in (A), while PC3 and PC1 are shown in (B-D). Analysis along PC3 show tight clustering of Pittsburgh patients (D) as compared to the PCGC (C) and Framingham cohorts (B).

(E) Same PCA plot as in (A) except self declared ancestry is indicated (black, white, mixed, other). Other" includes Hispanics, Asians, American Indian, etc. and "Mixed" refers to mixed racial background. The patients included in the red box comprise the subpopulation of white subjects from Pittsburgh, Framingham and PCGC that were further stratified along PC3 in the PC3/PC1 plot in (F,G). The black box comprise black subjects from PCGC and Pittsburgh.

(F,G) The three ellipses shown in (F) delineate the centroid for white subjects in each of the three cohorts - Pittsburgh (black), PCGC (blue), and Framingham (gold). The black ellipse comprising most of the white subjects from Pittsburgh (F) were used for ancestry matching the white subjects from the PCGC and Framingham cohorts. Shown in Panel G are the ancestry matched white patients from all three cohorts used for the *PCDHA* delCNV analyses.