SUPPLEMENTARY MATERIAL

TITLE: Cancer genes mutations in congenital pulmonary airway malformation patients

AUTHORS:

Jacob Shujui HSU[1,4], PhD; Ruizhong ZHANG[2], PhD; Fanny YEUNG[3], MD; Clara SM TANG[3], PhD; John KL WONG[1], PhD; MD; Man-Ting SO[3], MSc; Huimin XIA[2], MD; Pak SHAM[1,4], PhD; Paul K TAM[3], MD; Miaoxin LI[1,4], PhD; Kenneth WONG[3], PhD; Maria-Mercè GARCIA-BARCELO[3], PhD.

METHODS

Generation of Whole Exome Sequencing (WES) data

Illumina's TruSeq® DNA Sample Prep v.2, TruSeq® Exome Enrichment Kits (Illumina, San Diego, CA, USA) and SeqCap EZ+ UTR Human Exome capture kits (Roche NimbleGen, Madison, WI, USA) were used. The exome kit captures 20,940 genes, non-coding DNA in flanking regions and regulatory elements. The captured DNA was sequenced as paired-end 100 base reads (PE100) on an Illumina HiSeq 2000, aiming to achieve 50 reads per base (50X) in average. The overall sequence quality metrics (QC) and detailed methods are described in Table S2 and supplementary material respectively. Sequencing reads were aligned to the human genome reference hg19 by the Burrows-Wheeler Aligner (BWA v0.7.12) to produce the sequence alignment file. Briefly, calling and filtering of single nucleotide variants (SNVs) and indels (small insertions/deletions) were done by the Genome Analysis Toolkit (GATK 3.6) haplotype-caller and Variant Quality Score Recalibration (VQSR) module respectively (details in supplementary material) and as recently described[1]. To enhance the calling accuracy and quality, 699 samples from ethnic Chinese individuals participating in an adult-onset disorder (Degenerative Disc Disease) exome sequencing project were used to check population stratification. PLINK 1.9 was used for the detection of contamination, relatedness and identity by descent (IBD) estimation

Only variants meeting the following criteria were considered for downstream analysis: i) minimal genotype Phred score (quality score) of 20; ii) minimal coverage of 8X; iii) if alternative alleles were detected in less than 5% of the reads then they were considered as reference alleles; iv) if alternative alleles were detected in more than 25% of the reads then they were considered as heterozygous alleles; v) the minimal average quality Phred score of each variant across all samples

had to be 30; vi) minimal mapping quality Phred score for each variant, 20 per read; vii) Phred-scaled *P*-value for overall strand bias, by using Fisher's exact test, 60 per each variant; viii) missing genotyping rate, <20%; ix) for SNVs, variants with Variant Quality Score Recalibration (VQSR) >99.50 were excluded, and for small indels, we excluded those with VQSR > 95. Variant annotation and filtering were performed by KGGSeq.

For *de novo* and homozygous analyses, only those variants predicted deleterious by both SIFT and Polyphen2 that mapped to in genes expressed in fetal lung were considered. *De novo* variants were validated by Sanger sequencing. CH variant sets were only selected when the predictor "possibly damaging" generated by SIFT or PolyPhen appear at least twice (2 variants =4 prediction results) or when one of the variants was a stop-gain/loss, splicing or frameshift. If there were more than one predicting outcome from one tool, we considered 1) only predicted results rather than missing values, 2) which outcome outnumber others, 3) damaging > possible damaging > benign when numbers are equal. Besides, we excluded all CH variants in *TTN* genes as this is the longest gene in the genome and as such has a special genetic architecture due to evolutionary processes. All abovementioned mutated genes were to be expressed in fetal lung. For digenic mutations, only recurrent interactive protein pairs not present in the parents were included

Single nucleotide (SNV) and small insertion/deletion (Indels) variant analysis

Given the sporadic presentation of CPAM, we assumed monogenic and/or digenic recessive inheritance models. Thus, the following variants were considered: i) *de novo*; ii) homozygous; iii) compound heterozygous (CH) and, iv) variants in two different but interacting genes (protein-protein interactions; PPIs). For the "*de novo* hypothesis", only novel and deleterious variants were selected. For inherited models, variants had to be inherited from different

parents and we restricted the minor allele frequency (MAF) to <=1% giving an incidence for any such combination in an individual of <=0.01% which is in accordance with the CPAM incidence. MAF information was obtained from those reported for East and South Asian populations in ExAC and in the latest 1000 Human Genome project data, respectively. Only rare non-synonymous variants were prioritized according to the hypothesis being tested. The deleteriousness of each non-synonymous SNV was estimated by SIFT, Polyphen2(HDIV) and CADD among others.

For *de novo* and homozygous analyses, only those variants predicted deleterious by both SIFT and Polyphen2 that mapped to in genes expressed in fetal lung were considered. *De novo* variants were validated by Sanger sequencing. CH variant sets were only selected when the predictor "possibly damaging" generated by SIFT or PolyPhen appear at least twice (2 variants =4 prediction results) or when one of the variants was a stop-gain/loss, splicing or frameshift (supplementary data).

To inquire the potential pathogenicity of the mutated genes, those were examined in various public databases including ExAC, RefGene, ClinVar, OMIM, mouse genome informatics (MGI), international mouse phenotyping consortium (IMPC), Deciphering Developmental Disorders (DDD) and the Catalogue of Somatic Mutations in Cancer (COSMIC v87) and The Cancer Genome Atlas TCGA v13.. Also, to predict the gene pathogenicity according to the inheritance mode of the variants, the Inheritance Modes Specific Pathogenicity Prioritization (ISPP) program was used[2]. ISPP-PAE scores indicate if the genes tested share pathogenic features with the pediatric disease genes as defined in the clinical genomic database CGD (https://research-nhgri-nih-gov.eproxy2.lib.hku.hk/CGD/).

The expression of mRNA, protein and involvement of the latter in fetal lung development was queried in Roadmap Fetal lung (http://www.roadmapepigenomics.org/data/tables/fetal#), in

the recently launched Lung Gene Expression Analysis (LGEA) database (https://research.cchmc.org/pbge/lunggens/)[3], human protein atlas (https://www.proteinatlas.org/) and through literature search.

For the interacting genes/proteins model, we resorted to the Disease Association Protein-Protein Link Evaluator (DAPPLE) and String (functional protein association networks; https://string-db.org/) to determine if the genes with rare non-synonymous SNVs (MAF<=5%) present in a given patient encoded interacting proteins. Not only the "interactive" gene pair should not be present in the parents, but also those recurrent PPI events in patients were considered.

Sanger sequencing was used for validation of selected non-synonymous *de novo* and inherited qualifying variants.

Copy Number Variation (CNV) data generation, detection and analysis

Patients' DNA was hybridized to Infinium HTS assay (Illumina, San Diego, CA, USA) according to standard protocols and detection and analysis was performed as recently described[1]. In brief, we used The GenomeStudio program (Illumina, San Diego, CA, USA) to normalize and analyse the SNP array data. PennCNV and Plink were used to call and to prioritize CNVs. Loss (deletion) and gain (duplication) were defined as the change of a minimum of a 10kb region, with a maximum standard deviation of Log R radio = 0.35 and Wave Factor threshold = 0.1. Only CPAM patient unique CNVs were included. Bedtools and Plink were used to cross-check CPAM's CNV calls against the DGV; http://dgv.tcag.ca/dgv/app/home) as well as our in-house local reference data that consist of 140 healthy controls plus 30 non-CPAM patients. The identified CNVs were annotated if they overlapped with genes. The overlapping genes were cross-checked with the databases described above.

Gene enrichment and hypergeometric tests

To identify any statistically over-represented group of mutated genes belonging to a gene-set or pathway, we performed hypergeometric and gene-enrichment tests against the pediatric disease causal genes reported in CGD (as per March 2018) as well as in curated gene sets (MSigDB; http://software.broadinstitute.org/gsea/msigdb/collections.jsp#C2). Gene-enrichment tests were performed on the genes carrying rare and damaging variants.

REFERENCES

1. Hsu JSJ, So M, Tang CSM, Karim A, Porsch RM, Wong C, Yu M, Yeung F, Xia H, Zhang R, Cherny SS, Chung PHY, Wong KKY, Sham PC, Ngo ND, Li M, Tam PKH, Lui VCH, Garcia-Barcelo MM. De novo mutations in Caudal Type Homeo Box transcription Factor 2 (CDX2) in patients with persistent cloaca. *Human molecular genetics* 2018: 27(2): 351-358.
2. Hsu JS, Kwan JS, Pan Z, Garcia-Barcelo MM, Sham PC, Li M. Inheritance-mode specific pathogenicity prioritization (ISPP) for human protein coding genes. *Bioinformatics* 2016: 32(20): 3065-3071.
3. Du Y, Kitzmiller JA, Sridharan A, Perl AK, Bridges JP, Misra RS, Pryhuber GS, Mariani TJ, Bhattacharya S, Guo M, Potter SS, Dexheimer P, Aronow B, Jobe AH, Whitsett JA, Xu Y. Lung Gene Expression Analysis (LGEA): an integrative web portal for comprehensive gene expression data analysis in lung development. *Thorax* 2017: 72(5): 481-484.

SUPPLEMENTARY FIGURES

Figure S1: Principal Component Analysis (PCA) indicates no population stratification.

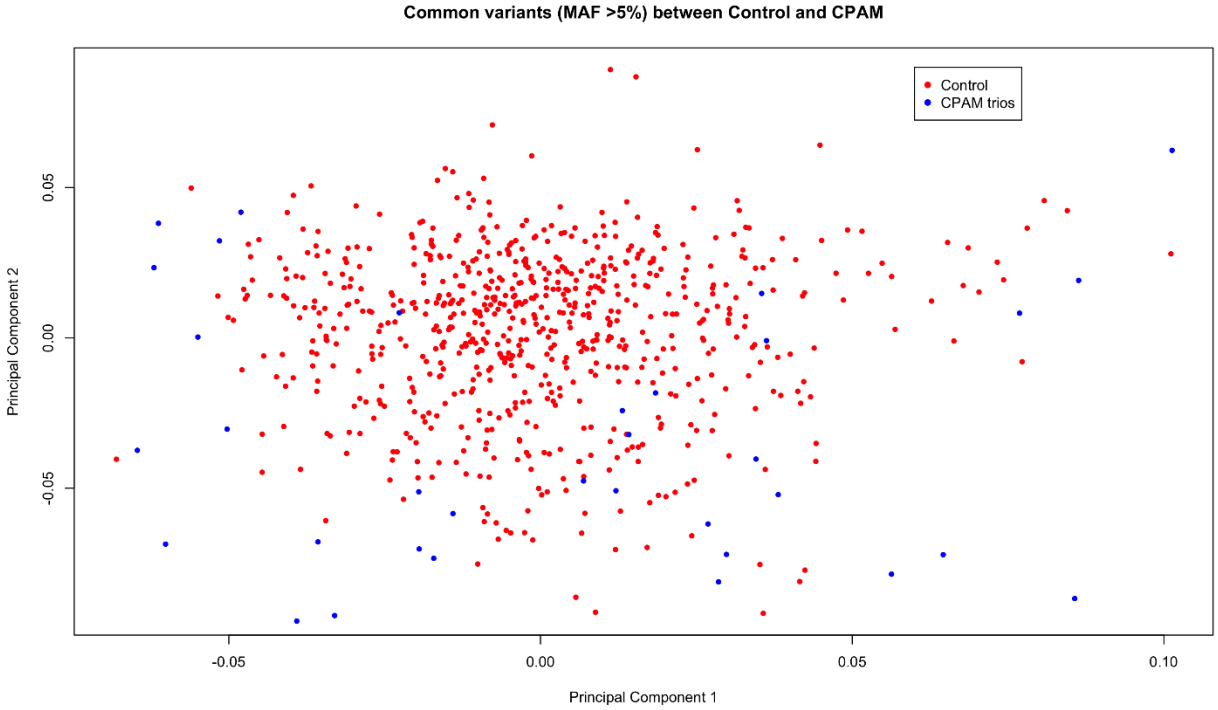
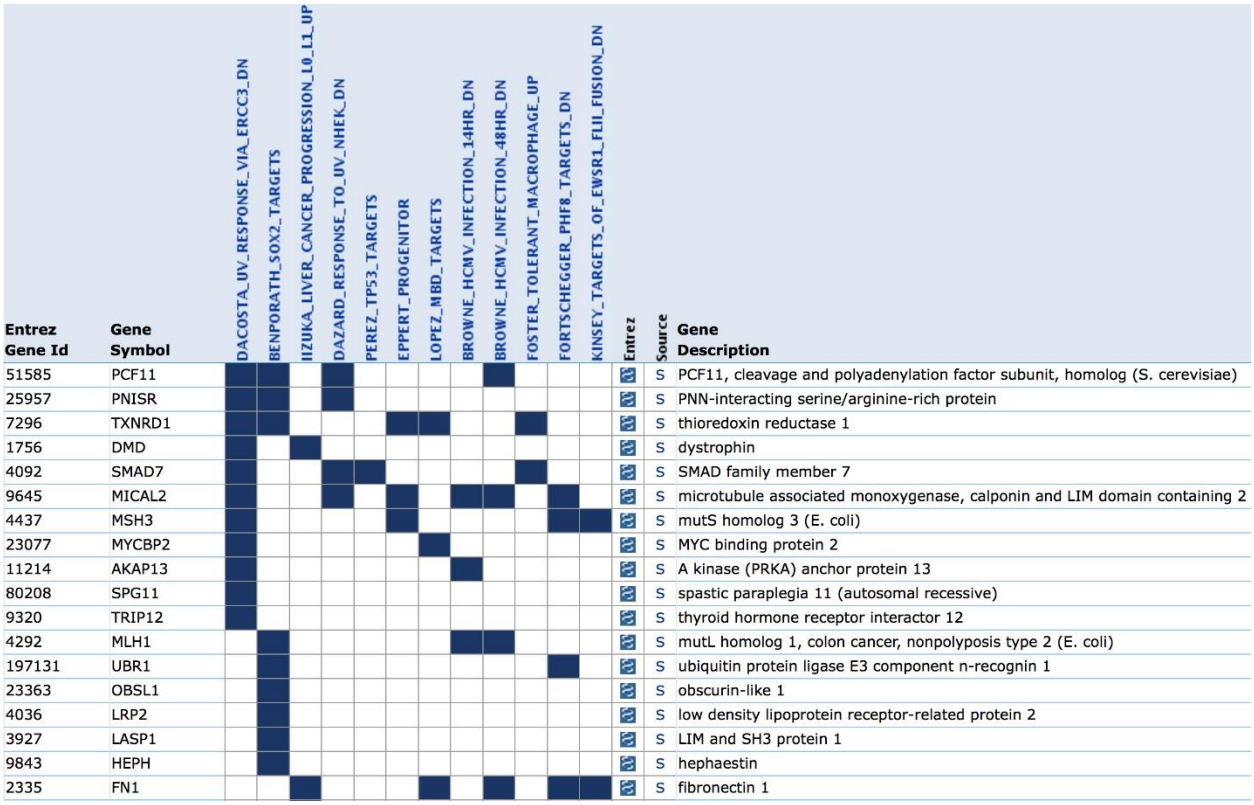


Figure S2. Gene-enrichment tests indicates that the genes with damaging rare variants are over-represented in cancer related pathways.



Entrez Gene Id	Gene Symbol	DACOSTA_UV_RESPONSE_VIA_ERCC3_DN	BENPORATH_SOX2_TARGETS	IIZUKA_LIVER_CANCER_PROGRESSION_L0_L1_UP	DAZARD_RESPONSE_TO_UV_NHEK_DN	PEREZ_TP53_TARGETS	EPPERT_PROGENITOR	LOPEZ_MBD_TARGETS	BROWNE_HCMV_INFECTION_14HR_DN	BROWNE_HCMV_INFECTION_48HR_DN	FOSTER_TOLERANT_MACROPHAGE_UP	FORTSCHEGGER_PHF8_TARGETS_DN	KINSEY_TARGETS_OF_EWSR1_FLI_FUSION_DN	Entrez	Source	Gene Description
51585	PCF11	■	■		■					■					S	PCF11, cleavage and polyadenylation factor subunit, homolog (S. cerevisiae)
25957	PNISR	■	■		■										S	PNN-interacting serine/arginine-rich protein
7296	TXNRD1	■	■				■	■			■				S	thioredoxin reductase 1
1756	DMD	■		■											S	dystrophin
4092	SMAD7	■			■	■					■				S	SMAD family member 7
9645	MICAL2	■			■		■		■	■		■			S	microtubule associated monoxygenase, calponin and LIM domain containing 2
4437	MSH3	■					■					■	■		S	mutS homolog 3 (E. coli)
23077	MYCBP2	■						■							S	MYC binding protein 2
11214	AKAP13	■							■						S	A kinase (PRKA) anchor protein 13
80208	SPG11	■													S	spastic paraplegia 11 (autosomal recessive)
9320	TRIP12	■													S	thyroid hormone receptor interactor 12
4292	MLH1		■						■	■					S	mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)
197131	UBR1		■									■			S	ubiquitin protein ligase E3 component n-recognin 1
23363	OBSL1		■												S	obscurin-like 1
4036	LRP2		■												S	low density lipoprotein receptor-related protein 2
3927	LASP1		■												S	LIM and SH3 protein 1
9843	HEPH		■												S	hephaestin
2335	FN1			■				■		■		■	■		S	fibronectin 1