의학석사 학위논문

융해소체축적병 연관 유전자의 생식선 돌연변이와 암 발생의 연관성

Oncogenic potential of germline mutations in lysosomal storage disease-associated genes

2018년 2월

서울대학교 대학원

의과대학 의학과 중개의학전공

신정훈

Oncogenic potential of germline mutations in lysosomal storage disease-associated genes

지도교수 윤성수

이 논문을 의학석사 학위논문으로 제출함

2017년 10월

서울대학교 대학원

의과대학 의학과 중개의학전공

신정훈

신정훈의 석사학위논문을 인준함

2017년 12월

위 원 장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

Abstract

Oncogenic potential of germline mutations in lysosomal storage disease-associated genes

Junghoon Shin

Graduate Program in Translational Research

College of Medicine

Seoul National University Graduate School

Introduction

Lysosomal storage diseases (LSDs) comprise inborn metabolic disorders caused by mutations in genes related to lysosomal function. As already observed in certain LSDs, the accumulation of macromolecules caused by LSDs may facilitate carcinogenesis.

Methods

Using whole genome sequence data from the International Cancer Genome Consortium (ICGC) PanCancer Analysis of Whole Genomes (PCAWG) and the 1000 Genomes projects, we analyzed the relationship between potentially pathogenic variants (PPVs) in 42 LSD genes and cancer. We evaluated age of

cancer onset and patterns of somatic mutation and gene expression according to PPV carrier status (wild type versus mutant).

Results

PPV prevalence in the ICGC-PCAWG cohort was significantly higher than that of the 1000 Genomes cohort (20.7% versus 13.5%, $P=8.7\times10^{-12}$). Cancer risk was increased in individuals with a greater number of PPVs. Population structure-adjusted SKAT-O analysis revealed 37 significantly associated cancer type-LSD gene pairs. These results were validated using ExAC cohort as a control population. Cancer developed earlier in carriers of PPVs in LSD genes in pancreatic adenocarcinoma (*MAN2B1*, *GALNS*, and *GUSB*), skin cancer (*NPC2*), and chronic myeloid disorder (*SGSH*). Analysis of transcriptome data from the pancreatic cancer project revealed 508 genes that were differentially expressed according to PPV carrier status, which were highly relevant for pancreatic cancer-related signaling pathways.

Conclusion

Carriers of germline PPVs in LSD-related genes have an increased incidence of cancer. The available therapeutic options to restore lysosomal function suggest the potential of personalized cancer prevention for these patients.

Keywords: Lysosomal storage disease; germline mutations; rare variant; cancer; association; transcriptome; gene expression

Table of Contents

Abstract	i
Table of Contents	iii
List of Tables	iv
List of Figures	v
Introduction	1
Methods	3
Results	6
Discussion	25
References	31
Supplementary Appendix	36
국문초록	86
감사의 글	89

List of Tables

Table 1. Lysosomal storage disease genes included in this study. 11

List of Figures

Figure 1. Populations of the PanCancer and 1000 Genomes cohorts. 15

Figure 2. Association of cancer with potentially pathogenic variants (PPVs) in lysosomal storage disease (LSD) genes. 16

Figure 3. SKAT-O associations between 30 major cancer types with more than 15 patients and potentially pathogenic variants (PPVs) in each lysosomal storage disease gene. 17

Figure 4. The quantile-quantile plot of minus logarithmic P-values determined by SKAT-O analysis. 18

Figure 5. Age of cancer onset. 19

Figure 6. Somatic mutational landscape of the pancreatic adenocarcinoma according to the potentially pathogenic variant (PPV) carrier status. 20

Figure 7. Differentially expressed genes (DEGs) between the potentially pathogenic variant (PPV)-bearing pancreatic adenocarcinoma and the PPV-free pancreatic adenocarcinoma. 22

Figure 8. KEGG pathways with significant segregation by the potentially pathogenic variant (PPV) carrier status in the pancreatic adenocarcinoma. 23

Figure 9. FPKM-UQ-normalized expression level of genes that were significantly associated with pancreatic adenocarcinoma in the SKAT-O analysis by potentially pathogenic variant (PPV) carrier status. 24

Introduction

Lysosomal storage diseases (LSDs) comprise more than 50 disorders caused by inborn errors of metabolism, which generally involve impaired function of endosome-lysosome proteins.[1] In LSDs, defects in genes encoding lysosomal hydrolases, transporters, and enzymatic activators result in macromolecule accumulation in the late endocytic system.[2] The disruption of lysosomal homeostasis is also linked to increased endoplasmic reticulum and oxidative stress, which is a common mediator of apoptosis in LSDs, suggesting signaling crosstalk between the lysosome and endoplasmic reticulum.[3]

LSD patients are generally thought to have severely impaired function with short life expectancy. However, a considerable number of undiagnosed LSD patients have mildly impaired lysosomal function and survive into adulthood.[1] These patients are often diagnosed after they develop secondary diseases such as Parkinsonism that is attributable to insidious LSDs.[4] At the cellular level, stress caused by lysosomal impairment may cause selective cell death or dysfunction, resulting in secondary manifestations such as progressive neurodegeneration or heart disease.[5]

Cancer development is related to cellular oxidative stress;[6] therefore, ongoing cellular stress for more than several decades would result in mutations or epigenetic alterations, potentially causing cancer. Consistent with this idea, clinical observations have shown that Gaucher and Fabry diseases are associated with certain histological types of cancer, providing evidence for the link between

dysregulated lysosomal metabolism and carcinogenesis.[7,8] However, the precise relationship between lysosomal dysfunction and cancer remains unclear. This may be due to the difficulty in recognizing LSD patients with mild symptoms and the extensive allelic heterogeneity with complex genotype–phenotype relationships.[9] Furthermore, accumulating evidence suggests that single allelic loss is functionally significant, even though the impact may not be sufficient to develop overt disease.[10] Considering the above along with recessive inheritance nature of most LSDs, we hypothesized that there are a large number of undetected heterozygous carriers of LSD-causing mutations under chronic cellular stress, and these carriers could be identified by using genome-wide sequencing analysis.

Here we report the results of a comprehensive pancancer analysis of potentially pathogenic germline mutations in LSD-related genes using data from global sequencing projects. We attempted to elucidate the oncogenic potential of these mutations in a histology-specific manner. Potential carcinogenic mechanisms were also investigated using tumor genomic and transcriptomic data with a special focus on pancreatic adenocarcinoma.

Methods

Study populations

We used matched tumor-normal pair whole genome and tumor whole transcriptome sequences and the clinical data of 2582 cancer patients (PanCancer cohort) constituting the International Cancer Genome Consortium (ICGC) PanCancer Analysis of Whole Genomes (PCAWG) project.[11] As controls, we used publicly available variant call sets from two global sequencing projects of individuals without known cancer histories. The first control data set comprised 2504 genomes from the 1000 Genomes project (1000 Genomes cohort).[12] The second data set contained exomes of 53,105 unrelated individuals from a subset of the Exome Aggregation Consortium release 0.3.1 that did not include The Cancer Genome Atlas (ExAC cohort).[13]

Potentially pathogenic variant selection

We performed an extensive literature review, which identified 42 LSD genes (Table 1).[1,9,14-16] The potentially pathogenic variants (PPVs) showed strong evidence of disease causation or functional defect and were selected based on their consequence on transcripts or proteins, clinical and experimental evidence obtained from curated databases and medical literature, or in silico prediction of mutational effects on protein function. PPV selection was carried out using these three positive selection criteria and an automated algorithm-based approach. The PPVs were grouped into three tiers with partial overlaps, each tier corresponding to one of the

selection criteria (Figure S1 in the Supplementary Appendix).

Statistical analysis

A two-step approach was used to examine the relationship between PPVs and cancer. In the first step, the PanCancer and 1000 Genomes cohorts were analyzed with the optimal sequence kernel association test (SKAT-O) for rare variant association and Fisher's exact test and logistic regression for direct comparison of mutation prevalence.[17] We adjusted for population structure using principal component analysis on 10,494 tag single nucleotide polymorphisms (Figure S2 and S3 in the Supplementary Appendix). The second step used the ExAC cohort as a control population, limited the analysis to coding regions covered in more than half of the ExAC individuals, and used Fisher's exact tests to validate the preceding results. We performed Wilcoxon rank sum tests and linear regression to compare age of cancer onset between groups. Differentially expressed gene (DEG) and gene set analyses were performed using the DESeq2 Bioconductor package and generally applicable gene set enrichment (GAGE) method based on the framework of the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway maps, respectively.[18-20]

Correction for multiple comparisons was conducted using the false discovery rate (FDR) estimation procedure, and the tail area-based FDR (also referred to as *q*-value) was reported.[21] All tests were two-tailed unless otherwise specified. We considered FDR<0.1 or P<0.05 (when not adjusted for multiple comparisons) significant. Statistical analysis was performed using R version 3.4.0 (R Foundation for Statistical Computing, Vienna, Austria) with pipelines of Bioconductor version

3.5.

A more detailed description of data sources, variant interpretation, PPV selection, and analysis are provided in the Supplementary Methods section of the Supplementary Appendix.

Results

Characteristics of study cohorts

The PanCancer cohort consisted of four populations and 38 histological subtypes of pediatric or adult cancer (Figure 1 and Table S1 in the Supplementary Appendix). The median age of cancer onset was 60 years (range, 1 to 90). The most common populations were European and American in most cancer types. The 1000 Genomes cohort comprised five populations; we combined the European and American populations for comparison with the PanCancer cohort.[12] The seven populations defined in the ExAC cohort did not match precisely to the populations defined in the other cohorts.[13]

PPV prevalence in the PanCancer and 1000 Genomes cohorts

In the 42 LSD genes, 7187 germline single nucleotide variants and small insertions and deletions in coding sequences, splice junctions, and 5’ or 3’ untranslated regions were identified in the aggregate variant call set of the PanCancer and 1000 Genomes cohorts (Figure S4 in the Supplementary Appendix). Of those, 4019 (55.9%) were singletons (variants found in only one individual), and 3′ untranslated region variants accounted for the largest proportion (37.7%).

Using a predefined algorithm, a total of 432 PPVs were selected within 41 genes; *LAMP2* contained no PPVs (Figure S5A in the Supplementary Appendix). PPV prevalence was 20.7% in the PanCancer cohort, which was significantly

higher than the 13.5% PPV prevalence of the 1000 Genomes cohort (odds ratio [OR], 1.67; 95% confidence interval [CI], 1.44–1.94; $P=8.7\times10^{-12}$; Figure 2A). This association remained significant after adjustment for population structure (OR, 1.44; 95% CI, 1.22–1.71; $P=2.4\times10^{-5}$). The ORs for cancer risk were higher in individuals with a greater number of PPVs and these were broadly consistent in individual tiers, although the difference was not significant for each tier (Figure 2A).

For comparison, we also examined the prevalence of rare synonymous variants (RSVs) with mean allele frequency <0.5% and found that it did not differ between the PanCancer and 1000 Genomes cohort after adjustment for population structure (Figure 2B). The gene-specific prevalences of PPVs and RSVs in each cohort are displayed in Figure S5B and S5C in the Supplementary Appendix, respectively. The results show that PPVs were relatively more abundant in the PanCancer cohort versus the 1000 Genomes cohort with respect to the abundance of RSVs, for 33 of the 42 genes (78.6%; exact binomial test P<0.001).

Association of PPVs with histological subtypes of cancer

Among the 30 major histological subtypes of cancer (>15 individuals per subtype), the PPV prevalence ranged from 8.8% to 48.6%, and the PPV prevalence in seven subtypes were significantly higher than that of the 1000 Genomes cohort (Figure S6A in the Supplementary Appendix). Results of the tier-based analysis were similar, except that prevalence of PPVs in myeloproliferative neoplasm in tier 2 was lower than that of the 1000 Genomes cohort (Figure S6B–D in the Supplementary Appendix). In contrast, RSV prevalence showed much less

variation and was higher in the 1000 Genomes cohort (Figure S6E in the Supplementary Appendix). SKAT-O analysis adjusted for population structure uncovered 37 significantly associated cancer-gene pairs and four genes (*GBA*, *SGSH*, *HEXA*, and *CLN3*) with pancancer association (Figure 3, Figure S5B, and Table S2 in the Supplementary Appendix). Overall, 19 histological subtypes were significantly associated with at least one gene, and 18 LSD genes were associated with at least one cancer type. No evidence of systematic inflation of test statistics was observed (Figure 4).

PPV prevalence in the PanCancer and ExAC cohorts

Validation was conducted for (1) the eight cancer cohorts with significantly higher PPV prevalence than that of the 1000 Genomes cohort and (2) 10 PPV sets significantly associated with the PanCancer cohort or three or more histological subtypes identified in the preceding SKAT-O analysis. As shown in Figure S7A, PPV prevalence was higher in all eight cancer cohorts than in the ExAC cohort, and this difference was significant for the PanCancer, pancreatic adenocarcinoma, medulloblastoma, pancreatic neuroendocrine carcinoma, and osteosarcoma cohorts. In addition, all tested PPV sets except *GBA* were more prevalent in the PanCancer cohort than in the ExAC cohort, and six were significantly different (Figure S7B).

Age of cancer onset according to PPV carrier status

The age of cancer onset in the major cancer cohorts is shown in Figure S8 in the Supplementary Appendix. We compared the age of onset according to PPV carrier status, with emphasis on the cancer cohorts and PPV sets that underwent validation

analysis. The age of cancer onset was lower in PPV carriers in all evaluated cohorts (one-sided P<0.5; Figure 5A). This difference was significant in the following cohorts: PanCancer (median age, 59 versus 61 years; P=0.002), pancreatic adenocarcinoma (median age, 61 versus 68.5 years; P<0.001), and chronic myeloid disorder (median age, 45.5 versus 58.5 years; P=0.044). In the entire PanCancer cohort, carriers in tier 1 and 3 and those with *HGSNAT*, *CLN3*, and *NPC2* PPVs experienced significantly earlier cancer development (Figure 5B). Moreover, PPV load (number of PPVs per individual) showed significant negative linear correlation with age of cancer onset in the PanCancer cohort and the pancreatic adenocarcinoma group (Figure S9 in the Supplementary Appendix). Extended analysis encompassing all histologies and genes revealed earlier cancer onset in PPV carriers in five additional cancer-gene pairs (Figure 5C), three of which (pancreatic adenocarcinoma-*MAN2B1*, cutaneous melanoma-*NPC2*, and chronic myeloid disorder-*SGSH*) were in concordance with the SKAT-O association.

Somatic mutations and gene expression signatures in PPV-bearing pancreatic adenocarcinoma

Based on the results of genetic and clinical analyses, we focused on pancreatic adenocarcinoma to determine whether differentiating patterns of somatic mutations and gene expression exist between PPV-bearing tumors (n=55) and PPV-free tumors (n=177). The 50 most frequently mutated genes in each group are shown in Figure 6. The five top-ranked genes were common in both groups (*KRAS*, *TP53*, *SMAD4*, *CDKN2A*, and *TTN*), and the first four of these were in agreement with previous studies.[22,23] Non-silent mutation burden was similar between groups

(mean 57.1 versus 56.3 per sample for PPV-bearing and PPV-free tumors, respectively; P=0.9). Mutational signature also did not differ significantly according to PPV carrier status (Figure S10 in the Supplementary Appendix). On the other hand, DEG analysis of the transcriptome sequence data revealed 287 upregulations and 221 downregulations in PPV-bearing tumors (Figures 7 and S11 in the Supplementary Appendix). Furthermore, gene set analysis with GAGE identified 18 KEGG pathways significantly segregated by PPV carrier status (Figure 8). These pathways incorporate well-known cancer-relevant signaling molecules such as Rap1, PPAR, Ras, and MAPK, some of which have been implicated in pancreatic cancer.[24] The expression level of *SGSH* and *IDUA*, PPVs of which showed significant SKAT-O association with pancreatic adenocarcinoma, did not differ significantly between the PPV-bearing and PPV-free tumors (Figure 9). No pancreatic adenocarcinoma specimen with transcriptome sequence data was available which carried PPVs in *MAN2B1*. Collectively, these results suggest that the existence of PPVs does not suppress LSD gene expression but causes functional impairment in the lysosomal components, and pancreatic carcinogenesis may be facilitated by altered gene expression in multiple cancer-related signaling pathways in the PPV carriers, providing hints for attractive diagnostic and therapeutic targets.

Supportive data

Additional supportive data are provided in the Supplementary Results section in the Supplementary Appendix.

Table 1. Lysosomal storage disease genes included in this study.

HGNC Symbol	Chromosome	Associated Lysosomal Storage Disease	Inheritance*
AGA	4	Aspartylglycosaminuria	Autosomal recessive
ARSA	22	Metachromatic leukodystrophy	Autosomal recessive
ARSB	5	Mucopolysaccharidosis VI (Maroteaux–Lamy syndrome)	Autosomal recessive
ASAH1	8	Farber lipogranulomatosis	Autosomal recessive
CLN3	16	Neuronal ceroid lipofuscinosis (NCL) 3 (juvenile NCL or Batten disease)	Autosomal recessive
CTNS	17	Cystinosis	Autosomal recessive
CTSA	20	Galactosialidosis	Autosomal recessive
CTSK	1	Pycnodysostosis	Autosomal recessive
FUCA1	1	Fucosidosis	Autosomal recessive
GAA	17	Glycogen storage disease type II (Pompe disease)	Autosomal recessive
GALC	14	Globoid cell leukodystrophy (Krabbe disease)	Autosomal recessive

GALNS	16	Mucopolysaccharidosis IVA (Morquio A syndrome)	Autosomal recessive
GBA	1	Gaucher disease	Autosomal recessive
GLA	X	Fabry disease	X-linked recessive
GLB1	3	Mucopolysaccharidosis IVB (GM1 gangliosidosis and Morquio B syndrome)	Autosomal recessive
GM2A	5	GM2-gangliosidosis type AB	Autosomal recessive
GNPTAB	12	Mucolipidosis II (I-cell disease) Mucolipidosis IIIA (pseudo-Hurler polydystrophy)	Autosomal recessive
GNPTG	16	Mucolipidosis IIIC (mucolipidosis III gamma)	Autosomal recessive
GNS	12	Mucopolysaccharidosis IIID (Sanfilippo syndrome D)	Autosomal recessive
GUSB	7	Mucopolysaccharidosis VII (Sly syndrome)	Autosomal recessive
HEXA	15	GM2 gangliosidosis type I (Tay-Sachs disease)	Autosomal recessive
HEXB	5	GM2 gangliosidosis type 2 (Sandhoff disease)	Autosomal recessive
HGSNAT	8	Mucopolysaccharidosis IIIC (Sanfilippo syndrome C)	Autosomal recessive
HYAL1	3	Mucopolysaccharidosis IX	Autosomal recessive

IDS	X	Mucopolysaccharidosis II (Hunter syndrome)	X-linked recessive
IDUA	4	Mucopolysaccharidosis I (Hurler, Scheie, and Hurler/Scheie syndromes)	Autosomal recessive
LAMP2	X	Danon disease	X-linked dominant
LIPA	10	Wolman disease Cholesteryl ester storage disease	Autosomal recessive
MAN2B1	19	α-Mannosidosis	Autosomal recessive
MANBA	4	β-Mannosidosis	Autosomal recessive
MCOLN1	19	Mucolipidosis IV	Autosomal recessive
NAGA	22	Schindler disease types I and II (Kanzaki disease)	Autosomal recessive
NAGLU	17	Mucopolysaccharidosis IIIB (Sanfilippo syndrome B)	Autosomal recessive
NEU1	6	Sialidosis	Autosomal recessive
NPC1	18	Niemann–Pick disease type C1	Autosomal recessive
NPC2	14	Niemann–Pick disease type C2	Autosomal recessive
PPT1	1	Neuronal ceroid lipofuscinosis 1 (infantile NCL)	Autosomal recessive

PSAP	10	Gaucher disease Metachromatic leukodystrophy	Autosomal recessive
SGSH	17	Mucopolysaccharidosis IIIA (Sanfilippo syndrome A)	Autosomal recessive
SMPD1	11	Niemann–Pick disease type A and B	Autosomal recessive
SUMF1	3	Multiple sulfatase deficiency	Autosomal recessive
TPP1	11	Neuronal ceroid lipofuscinosis 2 (Classic late-infantile NCL)	Autosomal recessive

*Inheritance pattern based on information provided in the Online Mendelian Inheritance in Man (OMIM) database (https://www.omim.org/).

HGNC denotes HUGO Gene Nomenclature Committee.

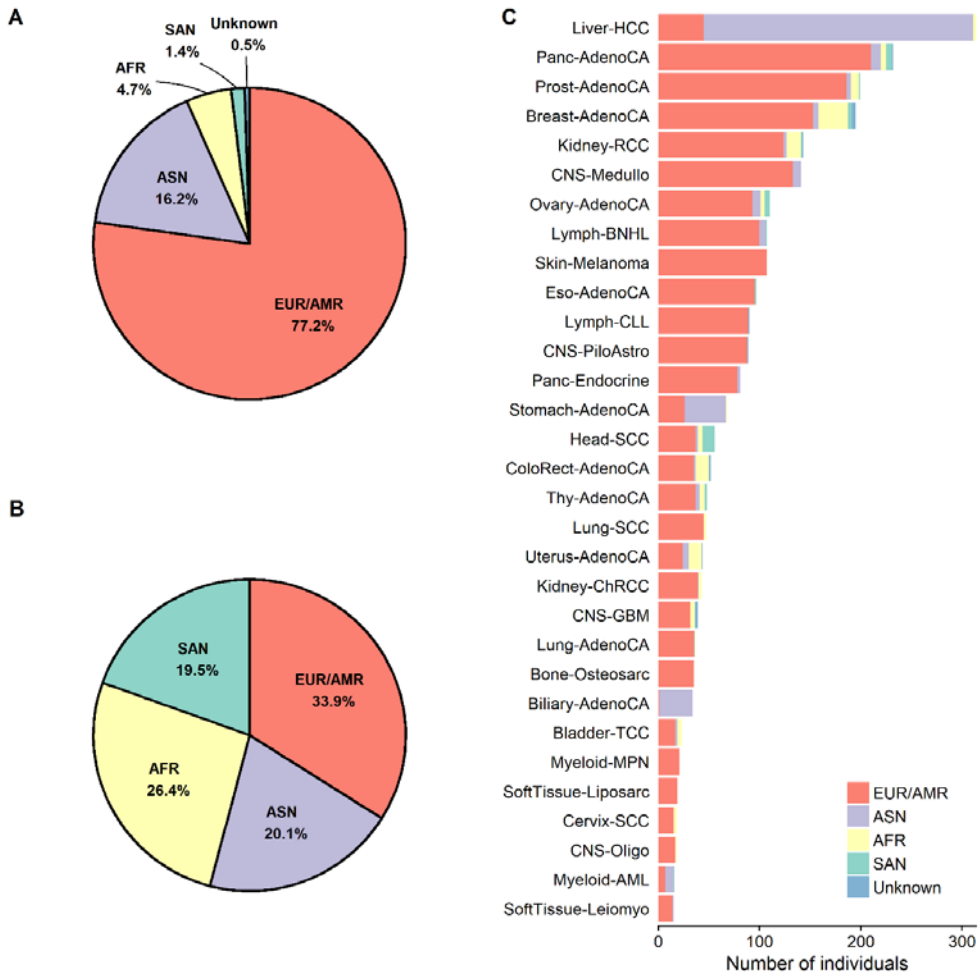

Figure 1. Populations of the PanCancer and 1000 Genomes cohorts.

Panels A and B show the populations of the PanCancer and 1000 Genomes cohorts, respectively. Panel C shows the populations of the PanCancer cohort according to histological subtype. Abbreviations of the histological subtypes are defined in Table S1 in the Supplementary Appendix. EUR denotes European, AMR American, ASN East Asian, AFR African, and SAN South Asian.

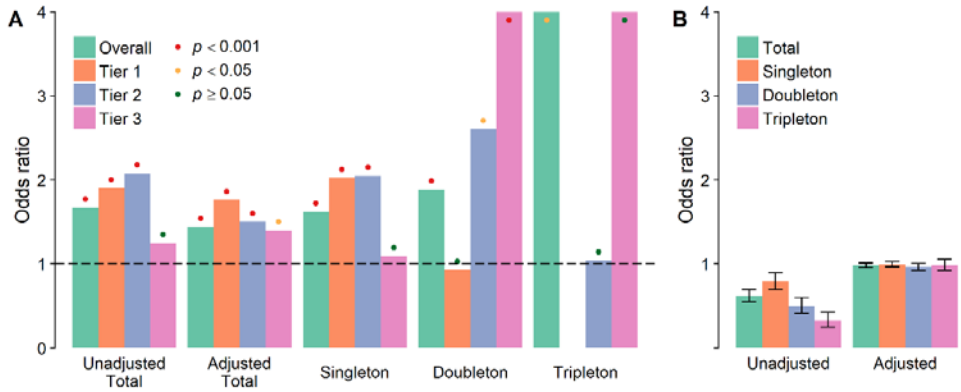

Figure 2. Association of cancer with potentially pathogenic variants (PPVs) in lysosomal storage disease (LSD) genes.

Panel A shows the odds ratios for PPV prevalence in the PanCancer cohort versus those of the 1000 Genomes cohort for the entire PPV set and each of three tiers with or without adjustment for population structure. Data were also analyzed separately for the prevalence of singleton, doubleton, and tripleton carriers (individuals with one, two, or three PPVs, respectively) without adjustment. The color of the dot above each bar denotes P-value obtained using Fisher's exact test (unadjusted analyses) or logistic regression (adjusted analyses). The horizontal dashed line indicates an odds ratio of one (no association between PPVs and cancer). Note that the estimated odds ratios for tier 3 doubletons and tripletons and the overall tripletons are 7.54, infinite, and 7.4, respectively, with corresponding bars cut off at the top edge of the plot. Panel B shows odds ratios for the prevalence of rare synonymous variants (mean allele frequency <0.5%) analyzed in the same manner for comparison. Error bars indicate 95% confidence intervals.

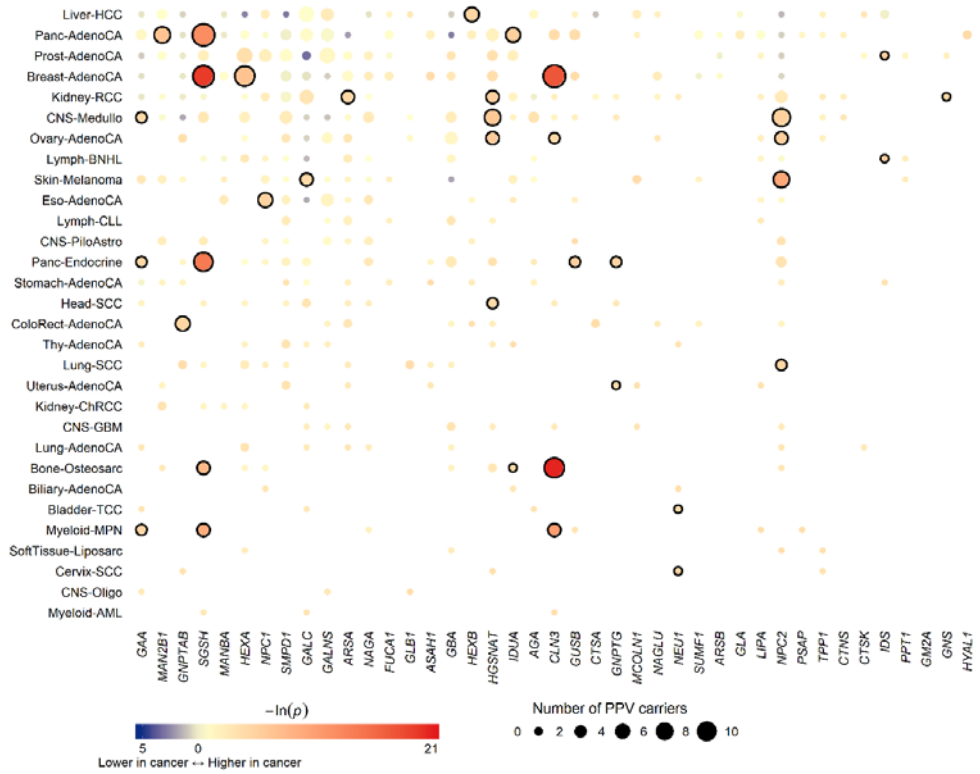

Figure 3. SKAT-O associations between 30 major cancer types with more than 15 patients and potentially pathogenic variants (PPVs) in each lysosomal storage disease gene.

The area of each dot is proportional to the number of PPV carriers in the corresponding cohort-gene pair. The color of each dot corresponds to the logarithm of the P-value determined by the SKAT-O test adjusted for population structure. Significantly associated cohort-gene pairs at the 0.1 false discovery rate threshold are encircled by bold rings. Cohorts are shown in descending order according to the number of individuals they contain (top to bottom), and genes are shown in descending order according to the number of unique PPVs (left to right). Abbreviations for histological subtype are defined in Table S1 in the Supplementary Appendix. FDR denotes false discovery rate.

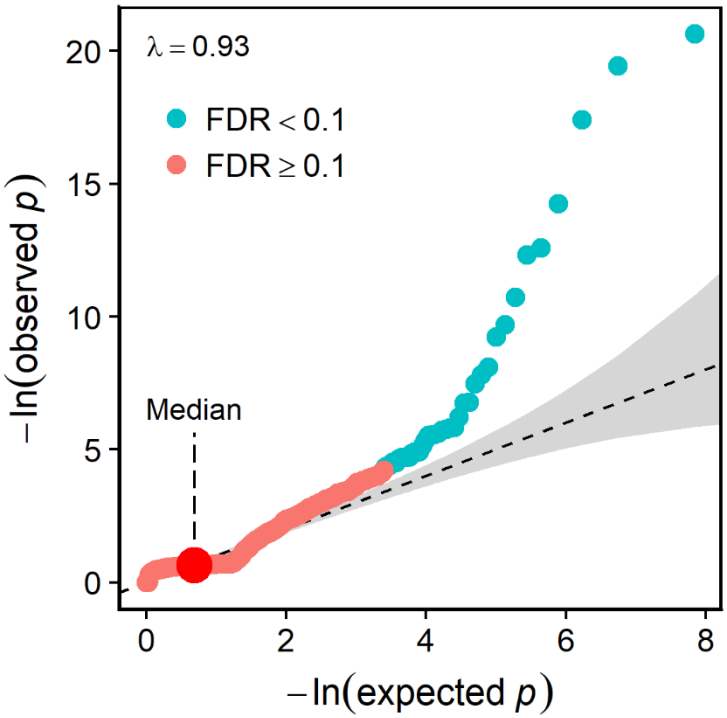

Figure 4. The quantile-quantile plot of minus logarithmic P-values determined by SKAT-O analysis.

Expected P-values under the null hypothesis (no association between potentially pathogenic variants [PPVs] and cancer) and observed P-values determined by the SKAT-O analysis are plotted on the logarithmic scale of the x- and y-axes, respectively. A group-based inflation factor (λ) is displayed. Gray shading indicates the 95% confidence interval. Note that each dot in this panel corresponds to a dot in Figure 3.

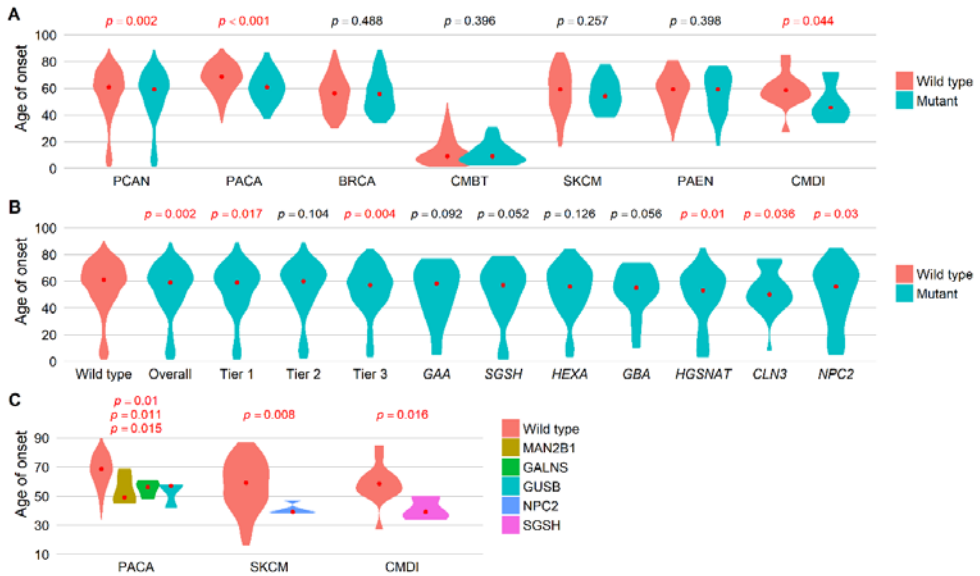

Figure 5. Age of cancer onset.

Panel A shows the age of cancer onset in seven cancer cohorts significantly associated with PPVs in the SKAT-O analysis (osteosarcoma was excluded because lack of age at onset information). Panel B shows the age of cancer onset according to carrier status of 11 PPV groups significantly associated with the PanCancer cohort or more than two histological cancer subgroups in the SKAT-O analysis. Panel C shows all cancer-gene pairs in which age of cancer onset differs significantly according to PPV carrier status. We performed one-sided Wilcoxon rank sum tests and the corresponding P-values are shown above the violin plots. The vertically aligned P-values from top to bottom for PACA in panel C correspond to the three genes displayed from left to right. The red dot in each violin plot represent the median. PCAN denotes the PanCancer, PACA pancreatic adenocarcinoma, BRCA breast cancer, CMBT medulloblastoma, SKCM cutaneous melanoma, PAEN pancreatic neuroendocrine carcinoma, and CMDI chronic myeloid disorder.

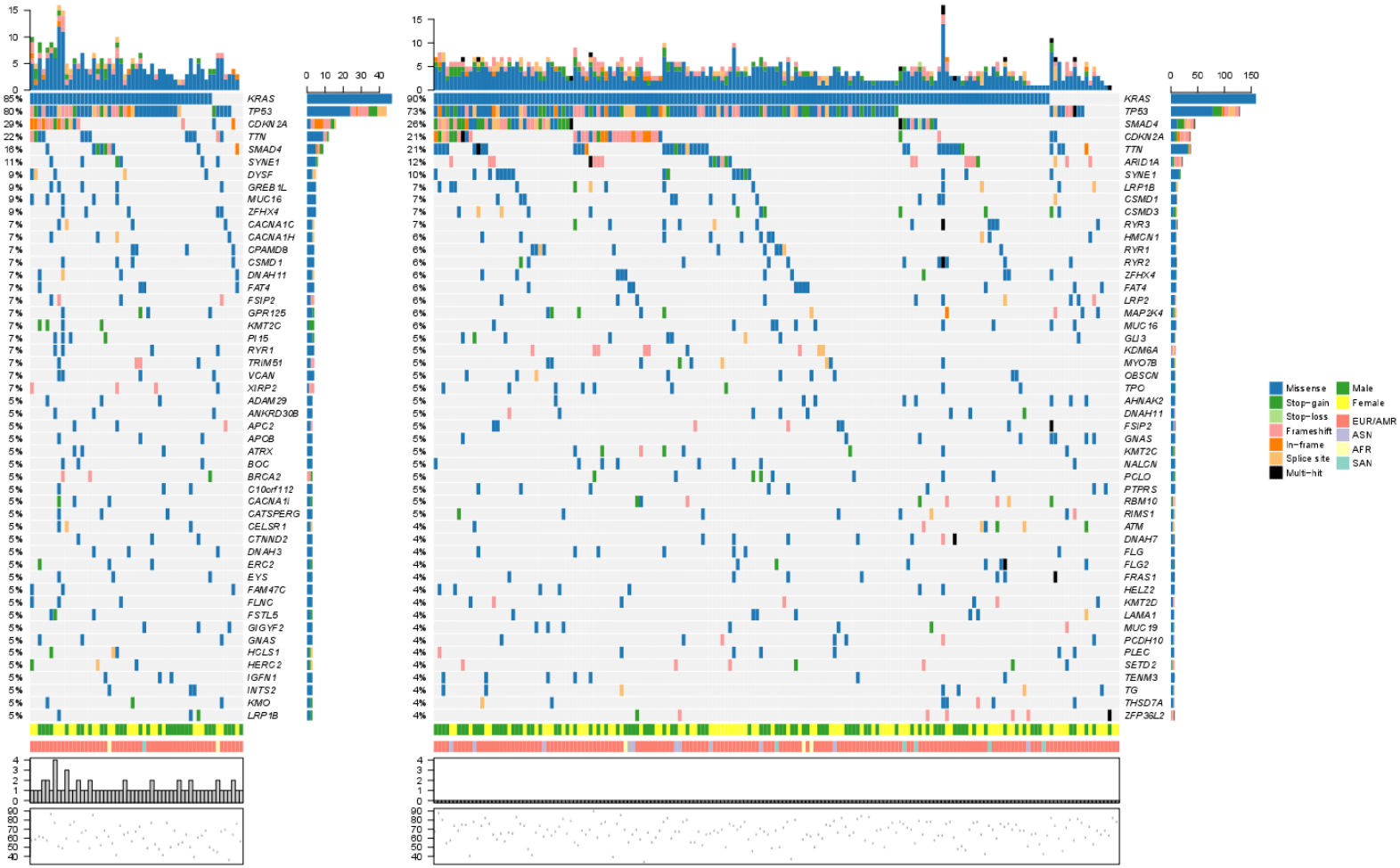
KRAS
TP53
CDKN2A
TTN
SMAD4
SYNE1
DYSF
GREB1L
MUC16
ZFHX4
CACNA1C
CACNA1H
CPAMD8
CSMD1
DNAH11
FAT4
FSIP2
GPR125
KMT2C
PI15
RYR1
TRIM51
VCAN
XIRP2
ADAM29
ANKRD30B
APC2
APOB
ATRX
BOC
BRCA2
C10orf112
CACNA1I
CATSPERG
CELSR1
CTNND2
DNAH3
ERC2
EYS
FAM47C
FLNC
FSTL5
GIGYF2
GNAS
HCLS1
HERC2
IGFN1
INTS2
KMO
LRP1B
SMAD4
ARID1A
LRP1B
CSMD1
CSMD3
RYR3
HMCN1
RYR1
RYR2
ZFHX4
FAT4
LRP2
MAP3K4
MUC16
GLI3
KDM6A
MYO7B
OBSCN
TPO
AHNAK2
DNAH11
FSIP2
GNAS
KMT2C
NALCN
PCLO
PTPRS
RBM10
RIMS1
ATM
DNAH7
FLG
FLG2
FRAS1
HELZ2
KMT2D
LAMA1
MUC19
PCDH10
PLEC
SETD2
TENM3
TG
THSD7A
ZFP36L2
Missense
Stop-gain
Stop-loss
Frameshift
In-frame
Splice site
Multi-hit
Male
Female
EUR/AMR
ASN
AFR
SAN

Figure 6. Somatic mutational landscape of the pancreatic adenocarcinoma according to the potentially pathogenic variant (PPV) carrier status.

Each rainfall plot shows non-silent somatic mutations in the 50 most frequently mutated genes in PPV-bearing (left, n=55) and PPV-free (right, n=177) pancreatic adenocarcinomas. Gender is indicated with color bars below the rainfall plots, below which is shown PPV load (number of PPVs per sample) indicated with bar plots, and age of cancer onset indicated with dot plots. The number of mutations per sample for the displayed genes and affected samples per gene are indicated with bar plots at the top and right side of the rainfall plots, respectively. EUR denotes European, AMR American, ASN East Asian, AFR African, and SAN South Asian.

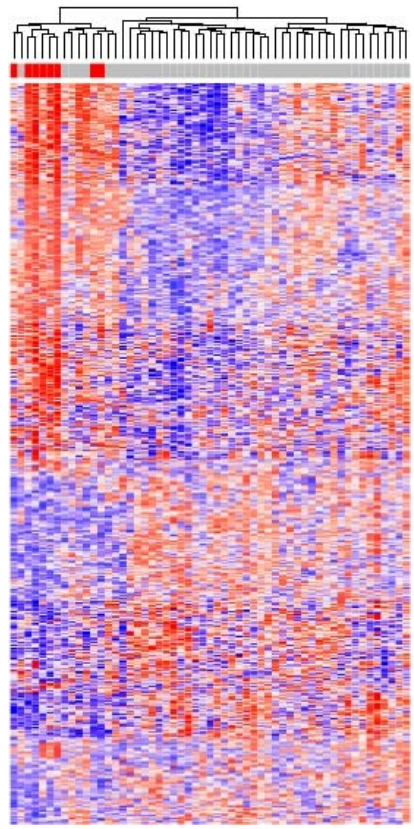

Figure 7. Differentially expressed genes (DEGs) between the potentially pathogenic variant (PPV)-bearing pancreatic adenocarcinoma and the PPV-free pancreatic adenocarcinoma.

The heat map shows relative mRNA expression for 55 pancreatic adenocarcinoma samples with available transcriptome data, including genes significantly up- or downregulated in the PPV-bearing cancers at the 0.1 false discovery rate (FDR) threshold (508 genes in total). Samples are ordered as columns by hierarchical clustering based on Euclidean distance and complete linkage. Genes are ordered as rows in the same manner (dendrogram not shown). High and low relative expression are indicated by progressively more saturated red and blue color, respectively. Below the dendrogram is a color bar indicating PPV-bearing tumors with red.

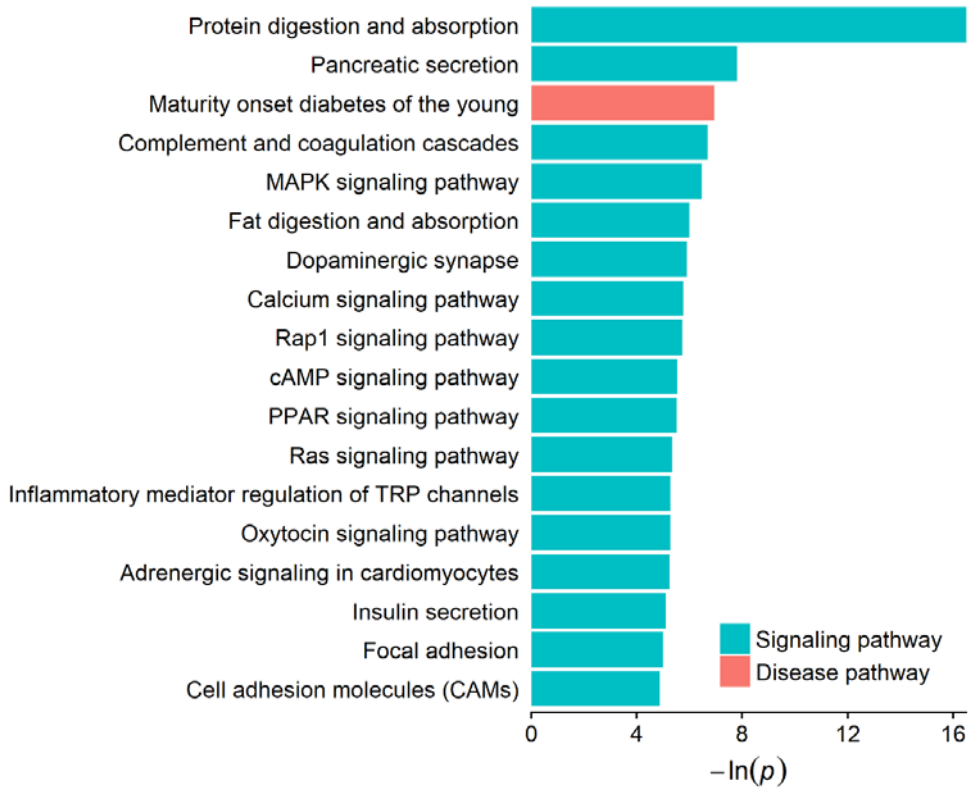

Figure 8. KEGG pathways with significant segregation by the potentially pathogenic variant (PPV) carrier status in the pancreatic adenocarcinoma.

The bar plot shows 18 KEGG pathways with significant segregation between PPV-bearing and PPV-free pancreatic adenocarcinoma at the 0.1 FDR threshold, as determined by generally applicable gene set enrichment analysis.

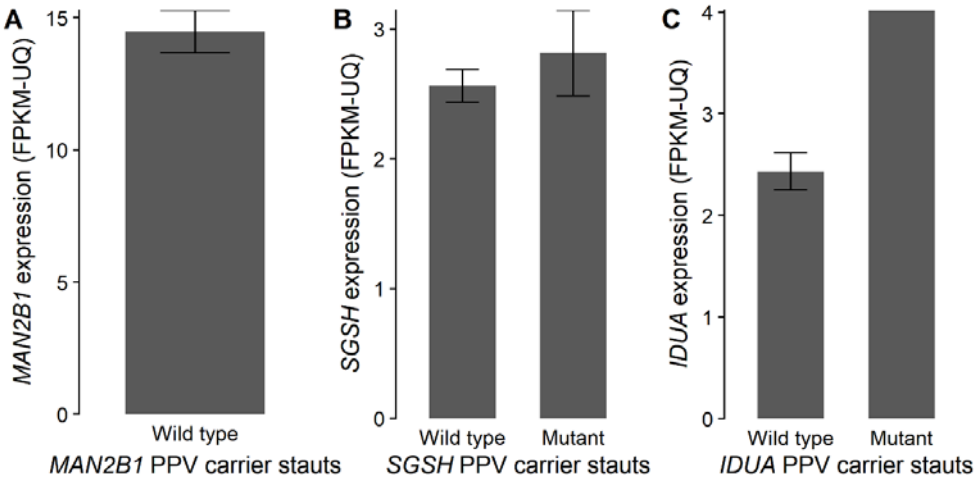

Figure 9. FPKM-UQ-normalized expression level of genes that were significantly associated with pancreatic adenocarcinoma in the SKAT-O analysis by potentially pathogenic variant (PPV) carrier status. Error bars indicate standard errors. SKAT-O denotes optimal sequence kernel association test.

Discussion

In this study, we showed previously unidentified associations between pathogenic germline mutations in LSD genes and cancer using unprecedentedly large-scale sequencing data. Results of subgrouping of PPVs into three tiers based on different selection criteria showed broadly consistent cancer association results, indicating the robustness of the results. The genetic association was further supported by the significant difference in age of cancer onset observed for PPV carriers of at least three histological subtypes and the PanCancer cohort compared with the PPV-free cancer patients.

Decades of research shed light on the wide range of lysosomal functions other than biomolecule catabolism, which include intracellular signaling, nutrient sensing, cellular growth regulation, plasma membrane repair, and phagocytosis.[16] These diverse roles of the lysosome are consistent with the complex phenotypes of LSDs, which can involve virtually any organ and cause neurological disabilities, visceral manifestations, hematological and skeletal abnormalities, and cancer.[1] It has long been evident that patients with Gaucher disease are at markedly increase risk of malignancy, especially multiple myeloma with increased risk estimated at ~50-fold.[25] However, the extreme rarity and heterogeneity of LSDs had hindered our understanding of the relationship between other LSDs and carcinogenesis, leaving this area largely unexplored. The wide spectrum of tumor histologies and LSD genes covered in this study has enabled the elucidation of numerous cancer-gene pairwise associations, most of which were previously unknown.

In a gene-specific analysis, we identified four genes showing significant pancancer association, and *SGSH* and *CLN3* were strongly linked to five and four histological subtypes, respectively. *SGSH* encodes sulfamidase, a lysosomal hydrolase that degrades heparan sulfate. Deficiency of sulfamidase leads to Sanfilippo syndrome A (mucopolysaccharidosis IIIA), which is characterized by progressive mental and behavioral deterioration that typically presents in childhood. However, adult onset disease that presents primarily with visceral manifestations without neurological abnormality has also been reported.[28] A recent *in vivo* study showed the importance of oxidative stress in the pathobiology of Sanfilippo syndrome A independently of inflammation in these patients.[29] The contribution of oxidative stress to cancer cell growth, invasiveness, and angiogenesis[6] suggests that inherited *SGSH* mutations elevate cancer risk via persistent exposure to oxidative stress. On the other land, CLN3 is a late endosomal and lysosomal transmembrane protein, and its genetic defect causes classic juvenile neuronal ceroid lipofuscinosis (CLN3 disease). In CLN3 disease, impaired trafficking of galactosylceramide to the plasma membrane promotes the generation of proapoptotic ceramide and subsequent activation of caspases, which in turn accelerates apoptosis.[15] Because it plays a role in controlling apoptosis, CLN3 also regulates cancer cell growth, suggesting therapeutic implications of this protein.[30] Results from our study confirm the hypothesis and demonstrate the utility of clinical sequencing based on a biological rationale.

The epidemiology of pancreatic cancer shows the existence of an early-onset subset, with ~5% to ~10% of all cases diagnosed before the age of 50. For these patients, a positive family history is the most important risk factor, indicating the

importance of germline genetic factors.[31] Although many predisposition genes have been identified (e.g., *BRCA1/2* and *PALB2*), the genetic cause remains unclear in most cases.[32] In our histology-specific analysis, pancreatic adenocarcinoma showed a particularly strong PPV association and early cancer onset in mutation carriers, motivating us to evaluate signatures of somatic mutations and gene expression in this histological subset. Results from the DEG and GAGE analysis provided novel insights into pathways that appear to be involved in pancreatic carcinogenesis precipitated by impaired lysosomal components. We also note that many of the significantly segregated KEGG pathways were involved in the physiological function of the pancreas (e.g. protein digestion and absorption). This result suggests that impaired lysosomal function may affect the normal organ function which is not directly linked to carcinogenesis as well. Somatic mutation patterns were comparable between PPV-bearing and PPV-free tumors. Collectively, these results expand the current catalogue of pancreatic cancer predisposition genes and suggest the role of epigenetic alterations in provoking cancer-driving signaling cascades. Future research is needed to validate this association and dissect the biological network.

Identification of cancer-predisposing mutations (CPMs) is crucial because they (1) can be exploited for surveillance strategies and risk-reducing interventions in high-risk individuals,[33] (2) may aid prognostic stratification,[34] and (3) can guide tailored therapy with promising outcomes.[35] From a clinical viewpoint, LSD genes are attractive therapeutic targets because of the mechanistically intuitive nature of enzyme replacement and substrate reduction therapy. Enzyme replacement therapy has already been approved for at least seven LSDs.[36] Other therapeutic approaches

that have shown promising results include pharmacological chaperones, gene therapy, and compounds that “read through” the early stop codon introduced by nonsense mutations.[36] Although it is not clear whether preemptive treatment can prevent or delay long-term complications of LSDs such as cancer, our study suggests the potential of harnessing these sophisticated LSD therapies for individualized prevention and management of cancer.

This study has limitations. Because we did not process the raw sequence data but used variant call sets available online, the possibility of batch effects cannot be excluded, even considering the similarity in pipelines used to generate each data set.[11-13] Nevertheless, the clear contrast between prevalence patterns of PPVs and silent variants between PanCancer and 1000 Genomes cohorts and the consistency of results obtained using two separate control cohorts support our conclusions. Another limitation is that the population structure difference could not be adjusted for the ExAC cohort because individual-level genotype data were not accessible. In addition, reliable phenotype information of cancer patients was limited to basic demographic information such as age of cancer onset, hampering in-depth clinical analysis. Although a wide variety of tumor histologies were analyzed, hematological malignancies such as myeloma, the most widely known LSD-associated cancer, were poorly represented in our study. And the number of patients with individual cancer types was not sufficiently large to draw a reliable histology-specific signals.

It is important to note that the PanCancer cohort consisted primarily of adult patients. Thus, highly pathogenic variants that cause fatal outcomes in childhood might be underrepresented in the selected PPVs, which may explain the higher-

than-expected population PPV frequency and the relatively low ORs compared to those observed for known high-penetrance CPMs.[37,38] For this reason we used the SKAT-O method for the association analysis to minimize loss of statistical power compared to the classic burden test.[17] In addition, the 1000 Genomes cohort is not a completely cancer-free population, because donors were selected solely based on their previous clinical history at the time of enrollment; therefore, it is not surprising that over one-tenth of the population carried PPVs. Similarly, a CPM screening study of pediatric cancer reported that 11 (1.1%) pathogenic or probably pathogenic mutations were found in known cancer predisposition genes in the control data set from the 1000 Genomes Project.[37]

Although the explosive growth of genome-wide sequence data has greatly contributed to the discovery of CPMs, expanding the catalogue of cancer predisposition genes,[37,39] there still exists a considerable amount of missing heritability.[40] Because previous studies typically focused on candidate genes that were selected based on their established relevance to cancer,[37-39] little is known outside of the current knowledge framework. This study expands our understanding of mechanisms underlying cancer predisposition, directly addressing the largely unexplored bridge linking monogenic metabolic diseases and cancer.

In conclusion, this study provides the most comprehensive landscape of the associations between germline mutations in LSD genes and cancer to date, substantially advancing our understanding of a new cancer-predisposing biochemical network, as well as cancer-relevant lysosomal components. Investigating the crosstalk between treatable metabolic diseases and cancer predisposition is invaluable, because it has a potential to turn the promise of

precision cancer prevention into reality. Diverse and increasingly sophisticated therapeutic options to restore defective lysosomal functions are currently available or being developed. Future prospective trials of these agents guided by individual mutation profiles may open a new horizon of personalized cancer prevention and treatment.

References

1. Parenti G, Andria G, Ballabio A. Lysosomal storage diseases: from pathophysiology to therapy. Annu Rev Med 2015;66:471-86.

2. Platt FM. Sphingolipid lysosomal storage disorders. Nature 2014;510:68-75.

3. Wei H, Kim S-J, Zhang Z, Tsai P-C, Wisniewski KE, Mukherjee AB. ER and oxidative stresses are common mediators of apoptosis in both neurodegenerative and non-neurodegenerative lysosomal storage disorders and are alleviated by chemical chaperones. Hum Mol Genet 2008;17:469-77.

4. Shachar T, Bianco CL, Recchia A, Wiessner C, Raas-Rothschild A, Futerman AH. Lysosomal storage disorders and Parkinson's disease: Gaucher disease and beyond. Mov Disord 2011;26:1593-604.

5. Linhart A, Elliott PM. The heart in Anderson-Fabry disease and other lysosomal storage disorders. Heart 2007;93:528-35.

6. Benz CC, Yau C. Ageing, oxidative stress and cancer: paradigms in parallax. Nature reviews Cancer 2008;8:875-9.

7. Arends M, van Dussen L, Biegstraaten M, Hollak CE. Malignancies and monoclonal gammopathy in Gaucher disease; a systematic review of the literature. Br J Haematol 2013;161:832-42.

8. Cassiman D, Claes K, Lerut E, et al. Bilateral renal cell carcinoma development in long-term Fabry disease. J Inherit Metab Dis 2007;30:830-1.

9. Wang RY, Bodamer OA, Watson MS, Wilcox WR. Lysosomal storage

diseases: Diagnostic confirmation and management of presymptomatic individuals. Genet Med 2011;13:457-84.

10. Wang RY, Lelis A, Mirocha J, Wilcox WR. Heterozygous Fabry women are not just carriers, but have a significant burden of disease and impaired quality of life. Genet Med 2007;9:34-45.

11. Hudson TJ, Anderson W, Artez A, et al. International network of cancer genome projects. Nature 2010;464:993-8.

12. The Genomes Project C. A global reference for human genetic variation. Nature 2015;526:68-74.

13. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 2016;536:285-91.

14. Scriver CR. The metabolic and molecular bases of inherited disease. 8th ed. New York: McGraw-Hill; 2001.

15. Boustany R-MN. Lysosomal storage diseases—the horizon expands. Nature reviews Neurology 2013;9:583-98.

16. Futerman AH, van Meer G. The cell biology of lysosomal storage disorders. Nat Rev Mol Cell Biol 2004;5:554-65.

17. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics 2012;13:762-75.

18. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550.

19. Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ. GAGE: generally applicable gene set enrichment for pathway analysis. BMC Bioinformatics 2009;10:161.

20. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 2000;28:27-30.

21. Storey JD. A Direct Approach to False Discovery Rates. Journal of the Royal Statistical Society Series B (Statistical Methodology) 2002;64:479-98.

22. Waddell N, Pajic M, Patch A-M, et al. Whole genomes redefine the mutational landscape of pancreatic cancer. Nature 2015;518:495-501.

23. Biankin AV, Waddell N, Kassahn KS, et al. Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. Nature 2012;491:399-405.

24. Jones S, Zhang X, Parsons DW, et al. Core Signaling Pathways in Human Pancreatic Cancers Revealed by Global Genomic Analyses. Science 2008;321:1801-6.

25. de Fost M, vom Dahl S, Weverling GJ, et al. Increased incidence of cancer in adult Gaucher disease in Western Europe. Blood Cells Mol Dis 2006;36:53-8.

26. Mistry PK, Taddei T, Dahl Sv, Rosenbloom BE. Gaucher Disease and Malignancy: A Model for Cancer Pathogenesis in an Inborn Error of Metabolism. Crit Rev Oncog 2013;18:235-46.

27. Pandey MK, Burrow TA, Rani R, et al. Complement drives glucosylceramide accumulation and tissue inflammation in Gaucher disease. Nature 2017;543:108-12.

28. Van Hove JLK, Wevers RA, Van Cleemput J, et al. Late-Onset visceral presentation with cardiomyopathy and without neurological symptoms of adult Sanfilippo A syndrome. American Journal of Medical Genetics Part A 2003;118A:382-7.

29. Trudel S, Trécherel E, Gomila C, et al. Oxidative stress is independent of inflammation in the neurodegenerative sanfilippo syndrome type B. J Neurosci Res 2015;93:424-32.

30. Rylova SN, Amalfitano A, Persaud-Sawin D-A, et al. The CLN3 gene is a novel molecular target for cancer drug discovery. Cancer Res 2002;62:801-8.

31. Vincent A, Herman J, Schulick R, Hruban RH, Goggins M. Pancreatic cancer. The Lancet 2011;378:607-20.

32. Klein AP. Identifying people at a high risk of developing pancreatic cancer. Nature reviews Cancer 2013;13:66-74.

33. Villani A, Shore A, Wasserman JD, et al. Biochemical and imaging surveillance in germline TP53 mutation carriers with Li-Fraumeni syndrome: 11 year follow-up of a prospective observational study. The Lancet Oncology 2016;17:1295-305.

34. Castro E, Goh C, Olmos D, et al. Germline BRCA Mutations Are Associated With Higher Risk of Nodal Involvement, Distant Metastasis, and Poor Survival Outcomes in Prostate Cancer. J Clin Oncol 2013;31:1748-57.

35. Robson M, Im S-A, Senkus E, et al. Olaparib for Metastatic Breast Cancer in Patients with a Germline BRCA Mutation. N Engl J Med 2017 June 4 (Epub ahead of print).

36. Hollak CEM, Wijburg FA. Treatment of lysosomal storage disorders: successes and challenges. J Inherit Metab Dis 2014;37:587-98.

37. Zhang J, Walsh MF, Wu G, et al. Germline Mutations in Predisposition Genes in Pediatric Cancer. N Engl J Med 2015;373:2336-46.

38. Pritchard CC, Mateo J, Walsh MF, et al. Inherited DNA-Repair Gene

Mutations in Men with Metastatic Prostate Cancer. N Engl J Med 2016;375:443-53.

39. Lu C, Xie M, Wendl MC, et al. Patterns and functional implications of rare germline variants across 12 cancer types. Nature Communications 2015;6:10086.

40. Rahman N. Realizing the promise of cancer predisposition genes. Nature 2014;505:302-8.

Supplementary Appendix

Supplementary Methods

Data sources

We downloaded germline and somatic (tumor) variant data sets for single nucleotide variants (SNVs) and small insertions and deletions (indels) of the PCAWG project cohort (PanCancer cohort) as variant call format (VCF) and mutation annotation format (MAF) files, respectively, from the sftp server (sftp://dccsftp.nci.nih.gov/pancan/). The germline variant call sets encompassed all 2834 PCAWG donors and were produced using the DKFZ/EMBL pipeline. The tumor somatic MAF file contained data of 2583 whitelist samples (only one representative tumor from each multi-tumor donor) and were generated by the PCAWG consensus strategy consolidating outputs from the Sanger, Broad, DKFZ/EMBL, and MuSE pipelines for SNVs and from the SMuFin, DKFZ, Sanger, and Snowman pipelines for indels. Pass-only variants were used for the analysis. Tumor whole transcriptome sequencing (RNA-Seq) data for the protein-coding RNAs were downloaded via Synapse (https://www.synapse.org/#!Synapse:syn3104297) as both raw and normalized read counts. Read alignment was conducted by TopHat2, counted using the htseq-count script from the HTSeq framework version 0.6.1p1 against the reference general transfer format of GENCODE release 19, and normalized using the FPKM-UQ normalization technique.[1] Clinical and histological information were obtained from the PCAWG wiki page (https://wiki.oicr.on.ca/pages/) in version 9 (generated on

November 22, 2016 and August 21, 2017, respectively).

As the main control cohort, individual-level genotype data of SNVs and indels for 2504 individuals of the 1000 Genomes project phase 3 (1000 Genomes cohort) were downloaded as VCF files (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3).[2] Likewise, population-level variant allele frequency (AF) data of SNVs and indels for 53,105 unrelated individuals from the Exome Aggregation Consortium release 0.3.1 (ExAC cohort), excluding those from The Cancer Genome Atlas subset, were downloaded for use in validation analysis (ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3.1).[3]

Quality control

Quality assessment of all PCAWG sequence data was carried out to determine whether to include each donor and RNA-Seq aliquot in the study according to three-level criteria (library, sample, and donor level), as guided by the PCAWG project team. This multi-level quality control process was necessary since individual donors may have multiple samples, and individual samples may have multiple libraries. As a rule, a sample was blacklisted if all of its libraries were low-quality, and whitelisted if all of its libraries were high-quality. Similarly, a donor was blacklisted if all associated samples were low-quality, and whitelisted if all associated samples were high-quality. Those that were neither blacklisted nor whitelisted were graylisted. Only whitelisted individuals and samples were included in this study (2583 tumor-normal pair genomes and 1094 RNA-Seq samples). Quality control criteria for each level are provided in the PCAWG flagship paper.

Composition of the PanCancer cohort

The original PCAWG project covered 2834 individuals encompassing 40 major histological subtypes as part of the International Cancer Genome Consortium, which included 48 projects and 20 primary tumor sites. Among those, 2583 whitelisted donors who met multi-level quality control criteria were selected. Sixteen whitelisted donors with histological diagnosis indicating benign neoplasms of the bone including chondroblastoma, chondromyxoid fibroma, osteofibrous dysplasia, or osteoblastoma were later excluded, leaving 2567 cancer patients in the final PanCancer cohort. Tumor specimens from nine patients had multiple histological diagnoses: both myeloproliferative neoplasm and acute myeloid leukemia in eight patients, and both hepatocellular carcinoma and cholangiocarcinoma in the ninth patient. For consistency in the histology-specific analysis, these eight patients were classified as acute myeloid leukemia and the ninth patient as cholangiocarcinoma. To analyze the age of cancer onset, we combined multiple histological cohorts that shared similar clinical and pathological features (e.g., breast invasive ductal, lobular, and micropapillary carcinomas were classified as breast cancer [BRCA], and myeloproliferative neoplasm and myelodysplastic syndrome as chronic myeloid disorder [CMDI]; Table S1). Among the 2567 patients, only 1075 had whitelisted RNA-Seq data from the matched tumor sample. Since 19 donors had more than one tumor sample, RNA-Seq data from 1094 tumors were available.

Gene selection and variant interpretation

Of the genes associated with lysosomal system functions, which include substrate hydrolysis, post-translational modification of hydrolases, intracellular trafficking, and enzymatic activation, 42 were selected based on disease relevance, as determined by a comprehensive literature review.[4-8] The genomic loci of the selected genes based on the GRCh37/hg19 human reference genome assembly were screened for all germline SNVs and indels in each VCF file. Variants were identified based on the GENCODE release 19 gene model (https://www.gencodegenes.org/releases/19.html). Functional annotation of variants was carried out using both ANNOVAR and Variant Effect Predictor version 85, and we cross-checked and manually curated the outputs to achieve the most relevant characterization of each identified variant.[9,10] From here, our analysis focused on variants within the protein-coding sequence, splice donor and acceptor sites on the intron side and within two base pairs of the exon-intron junction (GT-AG conserved sequence), and 5' and 3' untranslated regions (UTRs). Variants were classified into ten categories according to the computationally predicted consequence: missense, start-loss, stop-gain, stop-loss, synonymous, frameshift indel, in-frame indel, essential splice junction, and 5' and 3' UTR variants. When a variant allele had multiple consequences according to the transcripts used for calculation, it was classified into the most functionally disruptive category (e.g., protein-truncating rather than missense, and missense rather than UTR or synonymous). For example, rs373496399 (NC_000017.10:g.78184457G>A) could be either a missense or 3' UTR variant, depending on the transcript isoform, and was classified as missense. In this way, each variant belonged to a unique functional class that was used for subsequent analysis. In silico prediction of the

mutation's effect on protein function was conducted by 19 distinct computational algorithms with the use of dbNSFP version 3.3 (Figure S12).[11-28]

Potentially pathogenic variant selection

The prevalence of individual lysosomal storage diseases (LSDs) ranges from one per tens of thousands to one in millions of live births in Western countries, and considerable allelic heterogeneity exists.[29,30] Therefore, a single variant with a population AF >0.5% is extremely unlikely to be causative, even considering the possibility of underdiagnosis. A recent analysis of the prevalence of known Mendelian disease variants in a large number of exomes suggested that a substantial proportion of variants with AF >1% were in fact benign, highlighting the importance of filtering potentially pathogenic variants (PPVs) based on their frequency in a sufficiently large reference population.[3] It is natural to reason that Mendelian disease variants under high negative selection pressure are rapidly eliminated over generations, with de novo mutations continuously replacing them, maintaining the rarity of each variant. Based on this theoretical background and our data showing the rarity of deleterious variants, mostly with AF <0.5% (Figure S13), we excluded variants with mean AF (arithmetic mean of the AF in the PanCancer and 1000 Genomes cohorts) >0.5% from the PPV selection process.

We accessed the curated databases ClinVar, Human Gene Mutation Database Professional 2016.2 (HGMD), and locus-specific mutation databases and extensively reviewed the medical literature to identify disease-causing mutations. The examined locus-specific mutation databases are listed in Table S3. Variants were initially classified into five non-overlapping categories, as recently proposed

by the American College of Medical Genetics and Genomics (ACMG) and Association for Molecular Pathology (AMP) based on curated clinical significance data in ClinVar.[31] For variants with multiple interpretations of pathogenicity, priority was assigned to the category with the stronger evidence, hence “benign” rather than “likely benign,” and “pathogenic” rather than “likely pathogenic.” When interpretations indicating both pathogenic (“pathogenic’ or “likely pathogenic”) and benign (“benign’ or “likely benign”) directions of effect coexisted for a single variant, or no pathogenicity interpretation was provided in standard terminology, data in the HGMD and the locus-specific mutation databases along with supporting evidence obtained from the direct literature search were reviewed to determine the most relevant interpretation according to the ACMG and AMP guidelines.

As the role of defective regulatory function of microRNA in carcinogenesis has been spotlighted in recent years,[32,33] a number of SNVs in 3′ UTR microRNA-binding sites have been linked to increased or decreased cancer risk through altered expression of the gene products.[34-38] Although much less examined, the 5′ UTR also contains binding motifs for microRNAs, and its sequence affects messenger RNA (mRNA) stability.[39,40] Since UTR variants may create or destroy an microRNA-binding motif affecting microRNA-mediated regulation of gene expression or mRNA degradation, the biological consequence of UTR variants may be reflected in RNA transcript amounts.[41,42] Therefore, we analyzed RNA-Seq data to identify UTR variants associated with significantly depressed mRNA levels of the corresponding genes. Among the 3192 unique UTR variants with mean AF <0.5% (rare UTR variants), 795 and 2397 were present in 5' and 3' UTRs,

respectively. We compared the tissue mRNA level of each LSD gene after variance-stabilizing transformation of the RNA-Seq read counts[43] between UTR variant carriers and donors who did not carry any rare UTR variants in the corresponding gene, using linear regression analysis. Because the expression of each LSD gene varied considerably across histological subtypes (Figure S14), the regression model was adjusted for cancer histology. As a result, only one 3' UTR variant in *IDS*, rs145834006 (ENST00000340855:c.*3950A>G), was statistically significant at the 0.1 false discovery rate (FDR) threshold (Figure S15).

After inspection of all obtained information, PPVs strongly suspected to cause LSDs were selected by using three positive selection criteria (Figure S1). Tier 1 included all frameshift, start-loss, stop-gain, and splice donor and acceptor variants, and a UTR variant associated with significantly depressed mRNA expression of the mutated gene (rs145834006). Thus, most of these variants were considered loss-of-function in principle. Tier 2 included variants classified into the “pathogenic” or “likely pathogenic” category per the ACMG and AMP guidelines, based on information obtained from ClinVar and relevant medical literature, disease-causing mutations (the “DM” category) in HGMD, and pathogenic mutations ascertained in locus-specific mutation databases among those with uncertain clinical significance or conflicting interpretations in ClinVar. Of the variants without curated pathogenicity information in both ClinVar and HGMD (i.e., with unknown clinical significance), those predicted to be functionally deleterious by 19 separate in silico prediction tools (100% concordance) were classified into tier 3. The score threshold for classifying a variant as deleterious or benign was set at the provided default when available, or the median of all evaluated variants otherwise. Because

some variants were not successfully annotated by all 19 prediction tools using dbNSFP version 3.3, the available prediction scores were used in those cases.

PPV-cancer association analysis using the PanCancer and 1000 Genomes cohorts

Because the rarity of PPVs resulted in insufficient statistical power in variant-specific analysis, we used tier- and gene-based aggregate association analysis using the optimal sequence kernel association test (SKAT-O) method with an optimal ρ parameter chosen from a grid of eight points (0, 0.1^2, 0.2^2, 0.3^2, 0.4^2, 0.5^2, 0.5, 1), which could be interpreted as a pairwise correlation among the genetic effect coefficients.[44] As previously described, the SKAT-O method is robust to the presence of highly or moderately pathogenic and benign variants and is thus suitable when no uniform assumption can be made about the genetic effect of each variant, as in this study.[44] To determine if differences in the variant calling pipelines affected our results, we also compared the PPV-to-synonymous variant prevalence ratios between the cancer groups and 1000 Genomes cohort using weighted logistic regression analysis. We also assessed the variant-specific association to specific histological cancer types using logistic regression analysis assuming a multiplicative risk model. All analyses were adjusted for population structure using the method described below.

Adjustment for population structure

Throughout the study, we paid particular attention to avoid bias due to the difference in population structure between cohorts. Previous research on germline

cancer predisposition mutations usually confined the study population to a single ethnic group or a narrow range of ethnic groups or did not address this issue in depth.[45-47] However, disregarding the population structure could lead to a biased conclusion in the association study, and simply narrowing the scope to a superficially homogeneous ethnic group may not be sufficient when studying highly penetrant cancer predisposition mutations, because they may show a completely different pattern of population stratification compared to common genetic polymorphisms.[48] Thus we decided to conduct principal component analysis using the individual-level genotype data of tag single nucleotide polymorphisms (tag-SNPs), which represent the population-distinguishing linkage disequilibrium blocks.

First, 1,555,886 candidate tag-SNPs were downloaded from the phase 3 HapMap ftp server (ftp://ftp.ncbi.nlm.nih.gov/hapmap/phase_3/). We converted the genomic positions of SNPs into the framework of the GRCh37/hg19 human reference genome assembly using the Batch Coordinate Conversion (liftOver) tool created by the UCSC Genome Browser Group (https://genome.ucsc.edu/cgi-bin/hgLiftOver). Meanwhile, VCF files from both PCAWG and 1000 Genomes cohorts were merged using the Genome Analysis Toolkit to calculate broad AF.[49] The VCFtools version 1.13 was then used to extract the candidate tag-SNPs with AF $\geq$5% and $\leq$50% from the merged VCF, leaving 16,304 SNPs in the aggregate genotype data.[50] Among those, population-stratifying tag-SNPs were identified using the PLINK pruning method.[51] During this process, we carried out a recursive sliding-window procedure to exclude SNPs with a variance inflation factor >5 within a sliding window of 50 SNPs, shifting the window forward by 5 SNPs at each step, thereby

pruning linkage disequilibrium panels containing a number of SNPs to a reduced subset of representative tag-SNPs.[52] As a result, 10,494 tag-SNPs were selected for use in the subsequent principal component analysis.

A total of 5086 principal components (PCs) were obtained by performing principal component analysis on the combined genotype data of the 10,494 tag-SNPs from the PanCancer and 1000 Genomes cohorts. Correlations between each PC and the binary phenotype (cancer versus normal) and PPV load (number of PPVs per individual) were analyzed. Predictably, PC1, PC2, and PC3 collectively accounted for more than 12% of the total variance and were significantly correlated with both the binary phenotype and PPV load at 0.1 FDR threshold. The remaining 5083 PCs each accounted for less than 1% of the variance and were correlated with either the phenotype or mutational load or neither. Therefore, the three top-ranked PCs were considered potential confounders of the association between PPVs and cancer. Based on these results, we included PC1, PC2, and PC3 as covariates in the subsequent association analysis. Including more PCs in the model could result in loss of statistical power while providing only a small benefit of adjustment for the confounding effects of population structure.

Although an optimal method for population structure adjustment in a rare variant association study (typically for those with AF <1%) has not yet been developed, a recent study showed that principal component analysis using common variants (AF >5%), a conventional approach widely used in genome-wide association studies, is also appropriate for use in rare variant studies.[53] To examine the possibility of systematic inflation of statistics, a group-based inflation factor (λ) for the SKAT-O association test set was calculated using a previously described method (Figure

4).[54] The value was 0.93 (the greater the value above 1, the greater the possibility of bias), thereby disproving systematic inflation bias. This is in line with a recent analysis suggesting that rare variants are less susceptible to inflated association when the non-genetic risk is complex and widely distributed across the population, as in cancer.[48]

RNA-Seq data analysis

From the downloaded RNA-Seq read count data, we filtered out genes with zero reads across all samples to improve the computational speed in subsequent analysis while minimizing influence on the final results. Since the data were generated on the framework of Ensembl gene classification, we converted the Ensembl ID to Entrez gene ID for compatibility in analyses using Bioconductor pipelines. When multiple Ensembl genes matched to a single Entrez gene ID, those with the largest variance or standard deviation across all samples were retained, and others were removed from the count matrix.

We investigated differential mRNA expression patterns between PPV-bearing and PPV-free tumors using the DESeq2 Bioconductor package, after applying shrinkage estimation for fold changes and dispersions to improve the stability of the estimates.[55] The shrunken fold changes are plotted on the logarithmic scale in Figure S11A. Before estimating FDRs for multiple testing adjustment of the differentially expressed gene analysis results, we performed independent filtering of low-count genes to improve statistical power using the genefilter Bioconductor package.[56] For the heat map visualization in Figure 7, we ranked the samples according to the FPKM-UQ-normalized read counts for each gene and used the

rank numbers for heat map drawing to uniformalize the visual contrast.

Before pathway analysis using the GAGE method, variance-stabilizing transformation of the raw read counts was performed to achieve homoscedasticity of the count matrix and decrease the influence of genes with excessively large variation across samples. The GAGE analysis was based on group-on-group comparison, which could be controlled by the “compare” argument provided in the “gage” function of the gage Bioconductor package. We simultaneously tested for upregulation and downregulation of the expression of gene constituents comprising each KEGG pathway. Only pathways with FDR <0.1 are shown in Figure 8.

Validation analysis using the ExAC cohort as a control

Because the ExAC data set covered only the exome region consisting of GENCODE release 19 coding regions and their flanking 50 base pairs,[3] we focused our analysis on the coding regions covered in more than half of the ExAC individuals (median coverage depth ≥1) in the second-step validation analysis. Coverage depth information of the ExAC sequence data were downloaded from the ftp site (ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3.1/coverage). This strategy minimized bias due to the difference in sequencing coverage between the PanCancer and ExAC cohorts in the association analysis. We then selected PPVs from the aggregate variant data set of the PanCancer and ExAC cohorts with the algorithm used in the preceding analysis with the 1000 Genomes cohort. Finally, 1267 PPVs were selected, including 942 and 475 in tier 1 and 2, respectively, with 150 overlaps. No tier 3 variants were identified because the thresholds for classifying a variant as deleterious were stricter than in the first-step analysis for

some of the 19 scoring systems, including PhastCons.[26] This was due to the algorithmic decision to set the threshold at the median value of all evaluated variants (from both the PanCancer and ExAC cohorts), which differed from the median value of the merged variant set from the PanCancer and 1000 Genomes cohorts.

Although we excluded The Cancer Genome Atlas subset from the ExAC cohort to avoid misleading results from contamination of the control population with cancer patients, a large number of individuals with diseases that may be associated with LSD-causing mutations (e.g., schizophrenia and bipolar disorder) were present in the ExAC cohort.[3] Furthermore, population structure adjustment was unfeasible here because the individual-level genotype data of the ExAC cohort were not accessible when this study was conducted. As shown in Figure S16, the mean PPV frequency varied considerably across populations, and between-group pairwise correlations of the PPV frequency were relatively low for the East Asian and African populations in the ExAC cohort. Therefore, results from the association analysis using this cohort as a control should be interpreted with caution and considered as supportive data only.

Supplementary Results

Detection capability of the tier 3 PPV selection criterion for known pathogenic variants

Over the last decade, computational variant interpretation tools were developed for functional prediction of novel variants identified in sequencing studies, and

each was promoted as superior to previous tools. However, because of the intrinsic complexity and implicitness of machine learning algorithms, it is difficult to compare the performance of the tools. For that reason, we used all 19 available tools for the stringent selection of tier 3 PPVs. Our results showed that 50 (32.5%) of the 154 tier 2 PPVs, most of which are known pathogenic variants, passed the defined pathogenicity thresholds of all 19 prediction tools. This proportion was far higher than that of tier 3 PPVs in the entire pool of 7187 candidate variants (2.7%), indicating the high discriminating power of the tier 3 PPV selection criterion to identify truly pathogenic variants.

PPV-to-synonymous variant ratios

We hypothesized that the relative prevalence of PPVs to synonymous variants would be higher in cancer cohorts than in the control population if the observed enrichment of cancer patients with PPVs was not due to the batch effect. Therefore, we assessed the relative PPV enrichment in cancer patients, offsetting the frequency of synonymous variants by using weighted logistic regression analysis on the PPV-to-synonymous variant prevalence ratios (Figure S17A-D). Briefly, the results were very concordant with those obtained from a direct comparison of absolute PPV prevalence between cancer and control cohorts (Figure S6) and in striking contrast to the rare synonymous variant (mean AF <0.5% in the PanCancer and 1000 Genomes cohorts)-to-synonymous variant prevalence ratios (Figure S17E). This result supports the association between cancer and PPVs identified in the preceding analysis.

SGSH PPV prevalence in European and American populations

To examine the impact of the ethnicity imbalance between the cancer and control cohorts in the SKAT-O results, we directly compared the PPV prevalence within the *SGSH* gene. The results showed the broadest cancer association across histologies, with five cancer subtypes and the PanCancer cohort significantly associated with *SGSH* PPVs (Figure S18). Confining this analysis to the European and American populations had little effect on the prevalence pattern across cohorts. This result supports the major findings of the SKAT-O analysis despite the population structure difference.

Variant-specific association with cancer

Although variant-specific analysis has low statistical power because of the rarity of each PPV, some results deserve attention. A splice donor variant within the exon 4-intron 4 junction of the *NPC2* gene, rs140130028 (ENST00000434013:c.441+1G>A), was the PPV most strongly associated with various cancer types including medulloblastoma (P=0.008), ovarian adenocarcinoma (P=0.022), cutaneous melanoma (P=0.003), and lung squamous cell carcinoma (P=0.019). In addition, this variant was associated with the PanCancer cohort (PC-adjusted OR, 2.64; 95% CI, 0.96–7.25; P=0.059), but this association was not significant.

Inactivating mutations of *NPC2* can cause Niemann–Pick disease type C, which typically presents with progressive neurological abnormalities including developmental delay, cognitive impairment, cerebellar ataxia, and a wide spectrum of psychological conditions. The relationship between Niemann–Pick disease type

C and medulloblastoma was implied by the structurally similar domains of NPC1 and Patched transmembrane protein (Ptch) and loss-of-function mutations that were linked to medulloblastoma.[57] Ptch is regulated by Hedgehog signaling, and its inhibition by vismodegib showed promise in mouse models, leading to clinical trials of this hedgehog signaling inhibitor in medulloblastoma patients.[58-61] However, no previous studies published direct evidence for a biological connection between Niemann–Pick disease type C and medulloblastoma. Our study provides genetic evidence for the tumorigenic potential of inactivating mutations of *NPC2* in medulloblastoma.

Another example, rs145834006, which is the only 3′ UTR variant in the *IDS* gene significantly associated with downregulated mRNA expression ($P=0.1\times10^{-4}$; FDR = 0.07; Figure S15) and is therefore classified in tier 1, showed a strong association with non-Hodgkin B-cell lymphoma (PC-adjusted OR, 156; 95% CI, 11–2228; $P=2.2\times10^{-4}$). This finding is in accordance with the significant SKAT-O association between *IDS* and non-Hodgkin B-cell lymphoma (P=0.005; FDR=0.068; Figure 3). Considering the relatively high *IDS* mRNA levels in lymphoid tissue, which implies an essential role in these organs (Figure S14), downregulation of *IDS* caused by rs145834006 might be a key factor in B-cell lymphomagenesis. Nonetheless, *in vitro* and *in vivo* studies are required to confirm its functional significance.

Table S1. Patient characteristics of the PanCancer cohort.

Organ System	Histological Diagnosis	Histological Cohort	Clinical Cohort*	Number	Age at Diagnosis	Gender		Ethnicity				
						Male	Female	EUR/AMR	ASN	AFR	SAN	Unknown
Liver	Hepatocellular carcinoma	Liver-HCC	LIHC	314	67 (23-89)	226 (72%)	88 (28%)	45 (14.3%)	266 (84.7%)	3 (1%)	0 (0%)	0 (0%)
Pancreas	Adenocarcinoma	Panc-AdenoCA	PACA	232	67 (34-90)	118 (50.9%)	114 (49.1%)	210 (90.5%)	10 (4.3%)	5 (2.2%)	6 (2.6%)	1 (0.4%)
Prostate gland	Adenocarcinoma	Prost-AdenoCA	PRAD	199	59 (38-80)	199 (100%)	0 (0%)	186 (93.5%)	4 (2%)	8 (4%)	1 (0.5%)	0 (0%)
Breast	Invasive ductal carcinoma	Breast-AdenoCA	BRCA	195	56 (30-89)	1 (0.5%)	194 (99.5%)	153 (78.5%)	5 (2.6%)	29 (14.9%)	4 (2.1%)	4 (2.1%)
Kidney	Adenocarcinoma, clear cell type	Kidney-RCC	KIRC	143	60 (38-84)	90 (62.9%)	53 (37.1%)	124 (86.7%)	3 (2.1%)	14 (9.8%)	1 (0.7%)	1 (0.7%)
Central nervous system	Medulloblastoma	CNS-Medullo	CMBT	141	9 (1-49)	76 (53.9%)	65 (46.1%)	133 (94.3%)	8 (5.7%)	0 (0%)	0 (0%)	0 (0%)
Ovary	Adenocarcinoma	Ovary-AdenoCA	OV	110	60 (39-81)	0 (0%)	110 (100%)	93 (84.5%)	8 (7.3%)	4 (3.6%)	5 (4.5%)	0 (0%)
Lymph nodes	Non-Hodgkin B-cell lymphoma	Lymph-BNHL	BNHL	107	57 (4-85)	56 (52.3%)	51 (47.7%)	100 (93.5%)	6 (5.6%)	0 (0%)	1 (0.9%)	0 (0%)
Skin	Malignant melanoma	Skin-Melanoma	SKCM	107	57 (16-87)	69 (64.5%)	38 (35.5%)	107 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Esophagus	Adenocarcinoma	Eso-AdenoCA	ESAD	97	70 (43-87)	83 (85.6%)	14 (14.4%)	95 (97.9%)	0 (0%)	0 (0%)	2 (2.1%)	0 (0%)
Blood, bone marrow, and hematopoietic system	Chronic lymphocytic leukemia	Lymph-CLL	CLLE	90	61 (40-86)	60 (66.7%)	30 (33.3%)	89 (98.9%)	0 (0%)	0 (0%)	0 (0%)	1 (1.1%)
Central nervous system	Pilocytic astrocytoma	CNS-PiloAstro	LGG	89	8 (1-50)	42 (47.2%)	47 (52.8%)	88 (98.9%)	0 (0%)	0 (0%)	0 (0%)	1 (1.1%)
Pancreas	Neoroendocrine carcinoma	Panc-Endocrine	PAEN	81	59 (17-81)	53 (65.4%)	28 (34.6%)	78 (96.3%)	3 (3.7%)	0 (0%)	0 (0%)	0 (0%)
Stomach	Adenocarcinoma	Stomach-AdenoCA	STAD	68	65 (36-90)	50 (73.5%)	18 (26.5%)	26 (38.2%)	41 (60.3%)	1 (1.5%)	0 (0%)	0 (0%)
Head and neck	Squamous cell carcinoma	Head-SCC	HNSC	56	53 (19-76)	46 (82.1%)	10 (17.9%)	37 (66.1%)	2 (3.6%)	5 (8.9%)	12 (21.4%)	0 (0%)
Colon and rectum	Adenocarcinoma	ColoRect-AdenoCA	COAD	52	67.5 (31-89)	24 (46.2%)	28 (53.8%)	35 (67.3%)	2 (3.8%)	13 (25%)	0 (0%)	2 (3.8%)
Thyroid gland	Adenocarcinoma	Thy-AdenoCA	THCA	48	50.5 (17-85)	11 (22.9%)	37 (77.1%)	37 (77.1%)	4 (8.3%)	5 (10.4%)	2 (4.2%)	0 (0%)
Lung and bronchus	Squamous cell carcinoma	Lung-SCC	LUSC	47	68 (47-83)	37 (78.7%)	10 (21.3%)	45 (95.7%)	0 (0%)	2 (4.3%)	0 (0%)	0 (0%)

Uterus	Adenocarcinoma	Uterus-AdenoCA	UCEC	44	69 (35-90)	0 (0%)	44 (100%)	24 (54.5%)	6 (13.6%)	13 (29.5%)	0 (0%)	1 (2.3%)
Kidney	Adenocarcinoma, chromophobe type	Kidney-ChRCC	KICH	43	47 (17-86)	24 (55.8%)	19 (44.2%)	39 (90.7%)	1 (2.3%)	3 (7%)	0 (0%)	0 (0%)
Central nervous system	Glioblastoma	CNS-GBM	GBM	39	59 (21-76)	27 (69.2%)	12 (30.8%)	32 (82.1%)	0 (0%)	4 (10.3%)	1 (2.6%)	2 (5.1%)
Lung and bronchus	Adenocarcinoma	Lung-AdenoCA	LUAD	37	65.5 (41-81)	17 (45.9%)	20 (54.1%)	35 (94.6%)	1 (2.7%)	1 (2.7%)	0 (0%)	0 (0%)
Bones and joints	Osteosarcoma	Bone-Osteosarc	BOCA	35	Not provided	16 (45.7%)	19 (54.3%)	35 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Gallbladder and extrahepatic bile ducts	Cholangiocarcinoma	Biliary-AdenoCA	BTCA	34	64 (37-84)	19 (55.9%)	15 (44.1%)	1 (2.9%)	33 (97.1%)	0 (0%)	0 (0%)	0 (0%)
Urinary bladder	Transitional cell carcinoma	Bladder-TCC	BLCA	23	65 (34-84)	15 (65.2%)	8 (34.8%)	17 (73.9%)	2 (8.7%)	4 (17.4%)	0 (0%)	0 (0%)
Blood, bone marrow, and hematopoietic system	Myeloproliferative neoplasm	Myeloid-MPN	CMDI	21	53 (27-85)	9 (42.9%)	12 (57.1%)	21 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Bone and soft tissue	Liposarcoma	SoftTissue-Liposarc	SARC	19	61 (43-82)	14 (73.7%)	5 (26.3%)	19 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Uterine cervix	Squamous cell carcinoma	Cervix-SCC	CESC	18	39 (21-58)	0 (0%)	18 (100%)	15 (83.3%)	0 (0%)	3 (16.7%)	0 (0%)	0 (0%)
Central nervous system	Oligodendroglioma	CNS-Oligo	LGG	18	40.5 (17-62)	9 (50%)	9 (50%)	17 (94.4%)	0 (0%)	1 (5.6%)	0 (0%)	0 (0%)
Blood, bone marrow, and hematopoietic system	Acute myeloid leukemia	Myeloid-AML	LAML	16	54 (35-78)	10 (62.5%)	6 (37.5%)	7 (43.8%)	8 (50%)	0 (0%)	1 (6.2%)	0 (0%)
Bone and soft tissue	Leiomyosarcoma	SoftTissue-Leiomyo	SARC	15	63 (43-80)	5 (33.3%)	10 (66.7%)	14 (93.3%)	1 (6.7%)	0 (0%)	0 (0%)	0 (0%)
Breast	Invasive lobular carcinoma	Breast-LobularCA	BRCA	13	52.5 (40-76)	0 (0%)	13 (100%)	10 (76.9%)	0 (0%)	2 (15.4%)	0 (0%)	1 (7.7%)
Bone and soft tissue	Chordoma	Bone-Epith	BOCA	10	Not provided	6 (60%)	4 (40%)	9 (90%)	1 (10%)	0 (0%)	0 (0%)	0 (0%)
Breast	Duct micropapillary carcinoma	Breast-DCIS	BRCA	3	55 (40-61)	0 (0%)	3 (100%)	3 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Uterine cervix	Adenocarcinoma	Cervix-AdenoCA	CESC	2	39 (32-46)	0 (0%)	2 (100%)	1 (50%)	0 (0%)	1 (50%)	0 (0%)	0 (0%)
Blood, bone marrow, and hematopoietic system	Myelodysplastic syndrome	Myeloid-MDS	CMDI	1	77 (77-77)	1 (100%)	0 (0%)	1 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)

Table S2. Significantly associated cancer-gene pairs revealed by the population structure-adjusted SKAT-O analysis.

HGNC Symbol	Histological Cohort	Prevalence	*p*-value	*q-value*
CLN3	Bone-Osteosarc	22.9%	1.1E-09	3.5E-07
SGSH	Breast-AdenoCA	4.6%	3.7E-09	5.8E-07
CLN3	Breast-AdenoCA	5.1%	2.8E-08	2.9E-06
SGSH	Panc-Endocrine	8.6%	6.6E-07	5.2E-05
GBA	PanCancer	1.2%	3.4E-06	2.2E-04
SGSH	Panc-AdenoCA	4.3%	4.5E-06	2.4E-04
CLN3	Myeloid-MPN	14.3%	2.2E-05	9.9E-04
NPC2	Skin-Melanoma	4.7%	6.2E-05	2.4E-03
SGSH	Myeloid-MPN	14.3%	9.9E-05	3.5E-03
SGSH	PanCancer	1.8%	3.0E-04	9.6E-03
SGSH	Bone-Osteosarc	8.6%	4.1E-04	1.2E-02
CLN3	PanCancer	1.4%	5.7E-04	1.5E-02
HEXA	Breast-AdenoCA	4.1%	1.1E-03	2.7E-02
MAN2B1	Panc-AdenoCA	2.2%	1.2E-03	2.7E-02
HGSNAT	CNS-Medullo	3.5%	2.0E-03	4.2E-02
HEXA	PanCancer	1.3%	3.0E-03	5.5E-02
HGSNAT	Ovary-AdenoCA	2.7%	3.0E-03	5.5E-02
IDS	Prost-AdenoCA	0.5%	3.1E-03	5.5E-02
IDS	Lymph-BNHL	0.9%	3.3E-03	5.5E-02
NPC2	Ovary-AdenoCA	2.7%	3.6E-03	5.5E-02

GUSB	Panc-Endocrine	2.5%	3.8E-03	5.5E-02
IDUA	Panc-AdenoCA	1.7%	3.9E-03	5.5E-02
HGSNAT	Kidney-RCC	2.1%	4.0E-03	5.5E-02
NPC2	CNS-Medullo	4.3%	4.8E-03	6.4E-02
NPC1	Eso-AdenoCA	4.1%	6.0E-03	7.6E-02
HEXB	Liver-HCC	1.3%	7.4E-03	8.5E-02
GAA	Myeloid-MPN	9.5%	7.5E-03	8.5E-02
GNPTG	Panc-Endocrine	2.5%	7.6E-03	8.5E-02
GNPTAB	ColoRect-AdenoCA	7.7%	7.8E-03	8.5E-02
GNS	Kidney-RCC	0.7%	9.0E-03	8.8E-02
GAA	Panc-Endocrine	2.5%	9.1E-03	8.8E-02
NPC2	Lung-SCC	4.3%	9.2E-03	8.8E-02
NEU1	Cervix-SCC	5.6%	9.2E-03	8.8E-02
ARSA	Kidney-RCC	2.1%	9.7E-03	8.8E-02
NEU1	Bladder-TCC	4.3%	1.1E-02	9.3E-02
GALC	Skin-Melanoma	2.8%	1.1E-02	9.3E-02
CLN3	Ovary-AdenoCA	1.8%	1.1E-02	9.3E-02
GAA	CNS-Medullo	1.4%	1.2E-02	9.7E-02
IDUA	Bone-Osteosarc	2.9%	1.3E-02	9.8E-02
GNPTG	Uterus-AdenoCA	2.3%	1.3E-02	9.8E-02
HGSNAT	Head-SCC	3.6%	1.3E-02	9.8E-02

Table S3. Locus-specific mutation databases examined for tier 2 potentially pathogenic variant selection.

HGNC Symbol	Database	URL
GBA	Leiden Open Variation Database	https://research.cchmc.org/LOVD2/home.php?select_db=GBA
HEXA	HEXdb	http://www.hexdb.mcgill.ca/?Topic=HEXAdb
GAA	Leiden Open Variation Database	http://databases.lovd.nl/shared/genes/GAA https://research.cchmc.org/LOVD2/home.php?select_db=GAA
IDUA	Leiden Open Variation Database	https://grenada.lumc.nl/LOVD2/mendelian_genes/home.php?select_db=IDUA
HGSNAT	Leiden Open Variation Database	http://chromium.lovd.nl/LOVD2/home.php?select_db=HGSNAT
GLA	Leiden Open Variation Database	http://grenada.lumc.nl/LOVD2/MR/home.php?select_db=GLA https://research.cchmc.org/LOVD2/home.php?select_db=GLA
IDS	Leiden Open Variation Database	http://grenada.lumc.nl/LOVD2/MR/home.php?select_db=IDS
PPT1	NCL Mutation and Patient Database	http://www.ucl.ac.uk/ncl/index.shtml
TPP1	NCL Mutation and Patient Database	http://www.ucl.ac.uk/ncl/index.shtml
CLN3	NCL Mutation and Patient Database	http://www.ucl.ac.uk/ncl/index.shtml
	Retina International's Scientific Newsletter	http://www.retina-international.org/files/sci-news/cln3mut.htm

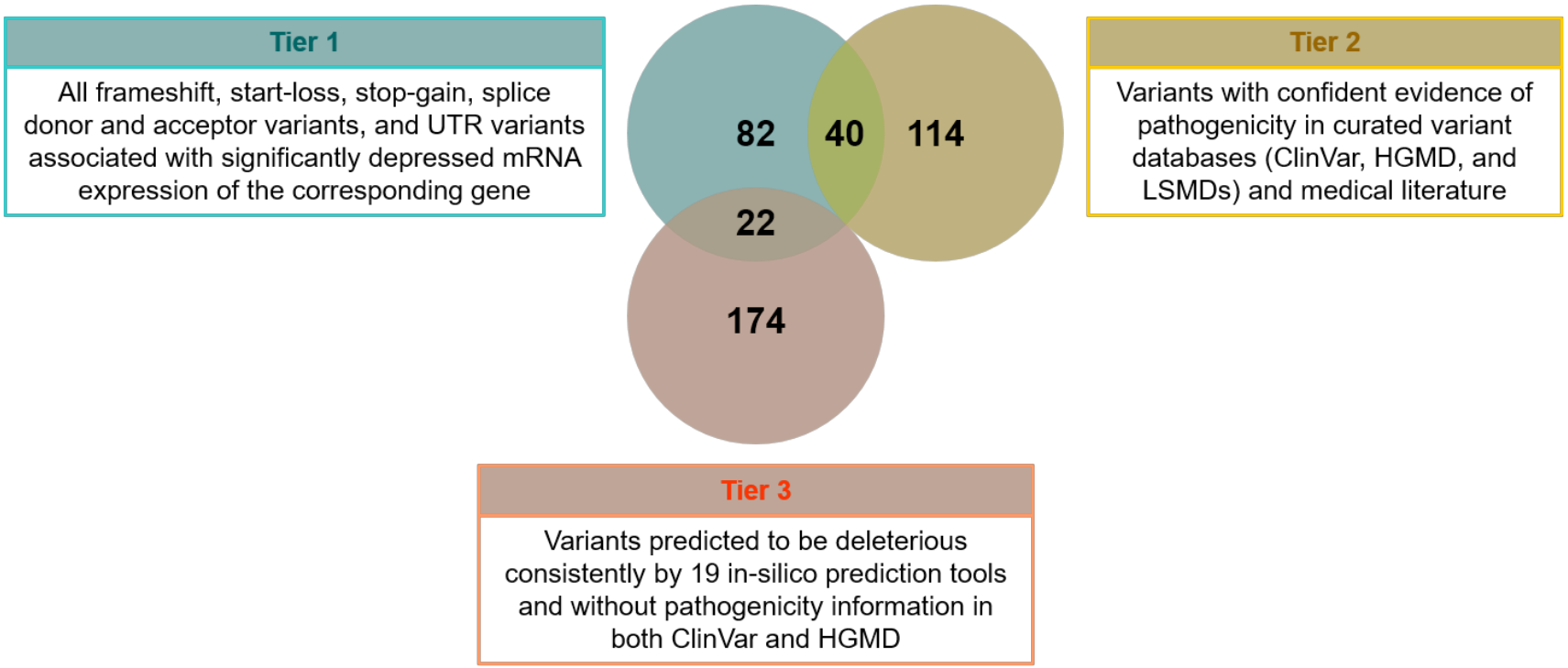

Figure S1. Venn diagram of potentially pathogenic variants grouped into three tiers.

UTR denotes untranslated region, mRNA messenger RNA, HGMD Human Gene Mutation Database, and LSMD locus-specific mutation database.

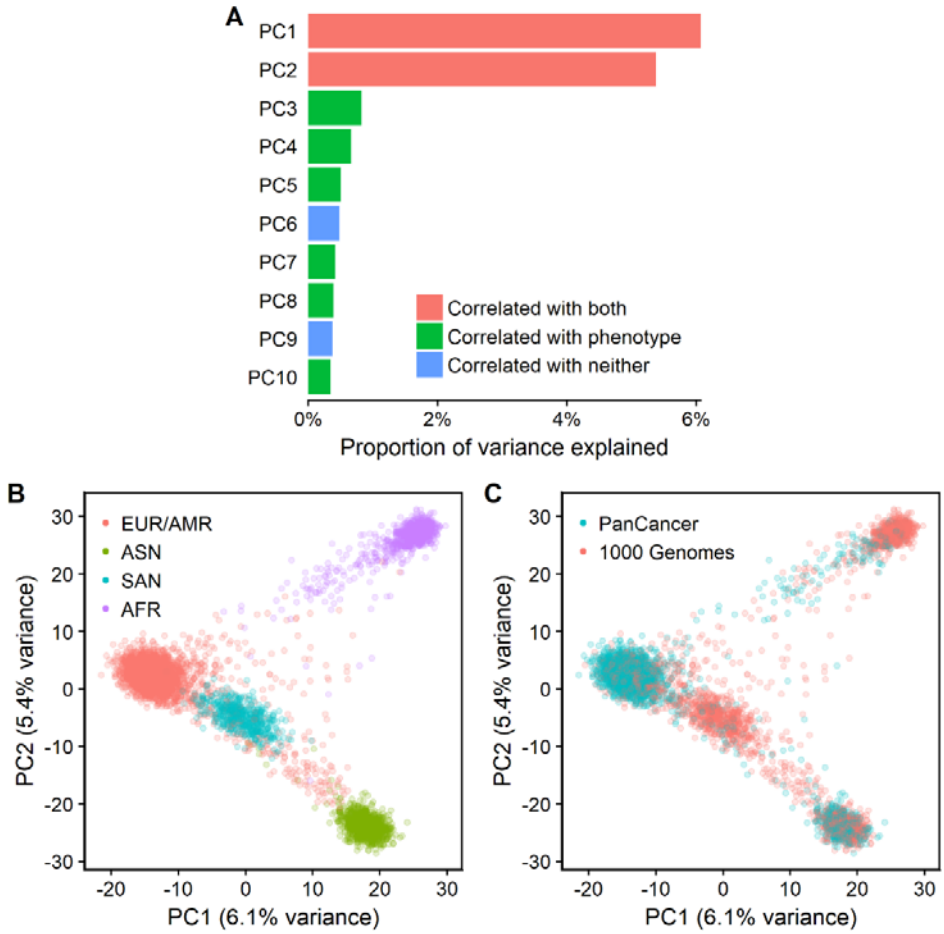

Figure S2. Most important principal components (PCs) obtained from principal component analysis (PCA) of genotype data from 10,494 tag single nucleotide polymorphisms (tag-SNPs) in the PanCancer and 1000 Genomes cohorts.

Panel A shows the proportion of total variance explained by the 10 most important PCs. Significant correlations between the binary phenotype (cancer versus normal) and potentially pathogenic variant (PPV) load (number of PPVs per individual) are indicated. Panels B and C show the ability of PC1 and PC2 to differentiate population and phenotype, respectively. EUR denotes European, AMR American, ASN East Asian, AFR African, and SAN South Asian.

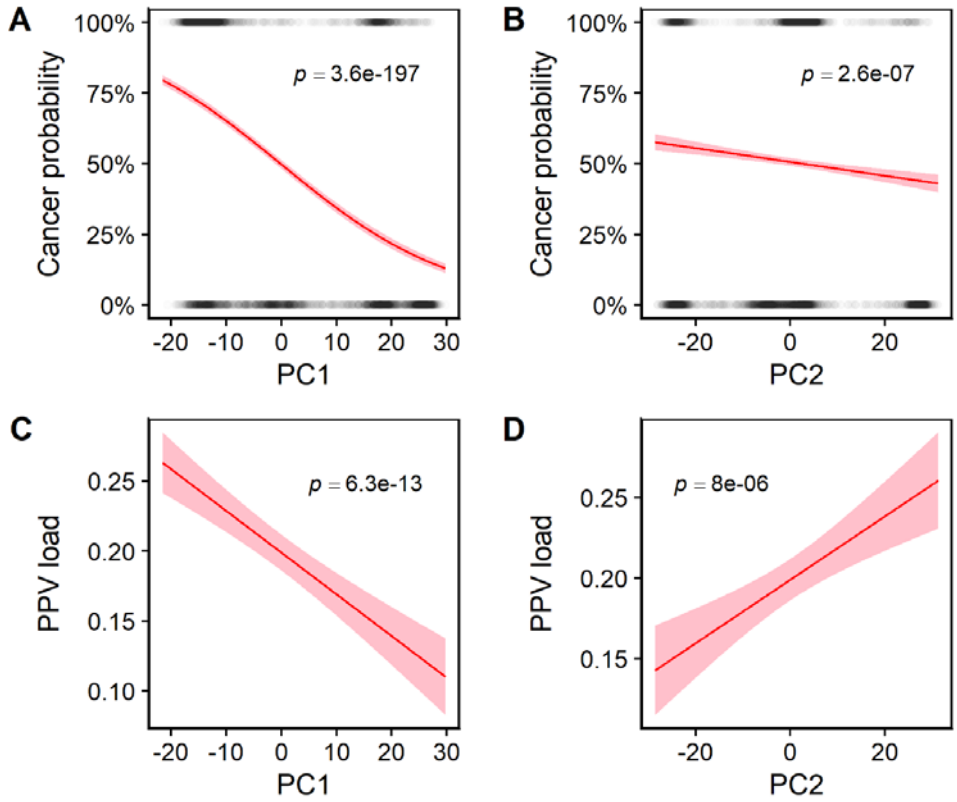

Figure S3. Correlations between the three most important principal components (PCs) and estimated cancer probability and potentially pathogenic variant (PPV) load.

Panels A and B show the correlation between cancer probability and PC1 and PC2, respectively, using logistic regression analysis. Each faint dot represents an individual in the PanCancer cohort (cancer probability = 100%) or 1000 Genomes cohort (cancer probability = 0%). Panels C and D show the linear relationships between PPV load and PC1 and PC2 using simple linear regression analysis. Pink shading indicates 95% confidence interval. Corresponding P-values are shown in each panel.

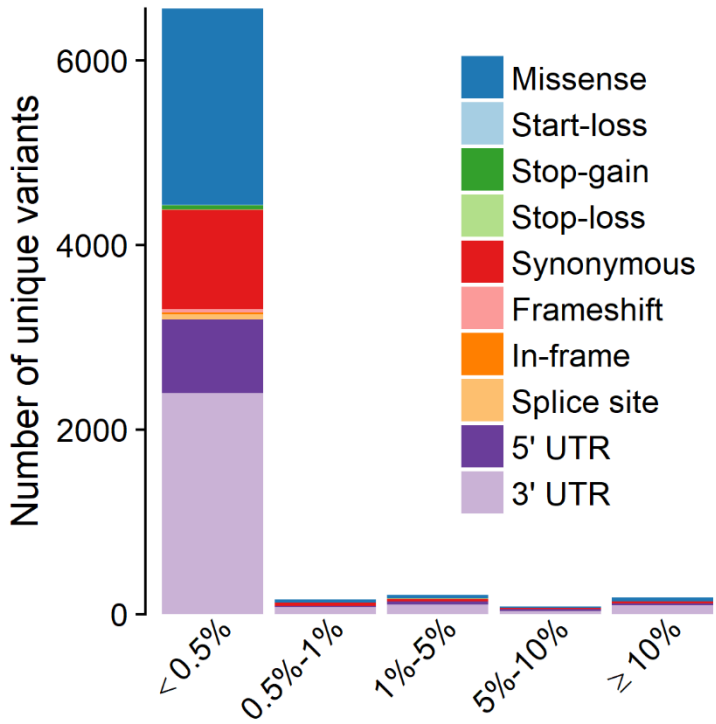

Figure S4. All germline single nucleotide variants and small insertions and deletions identified in the 42 lysosomal storage disease genes from the aggregate variant data set of the PanCancer and 1000 Genomes cohorts.

Numbers of unique variants belonging to each category are shown on the y-axis. UTR denotes untranslated region.

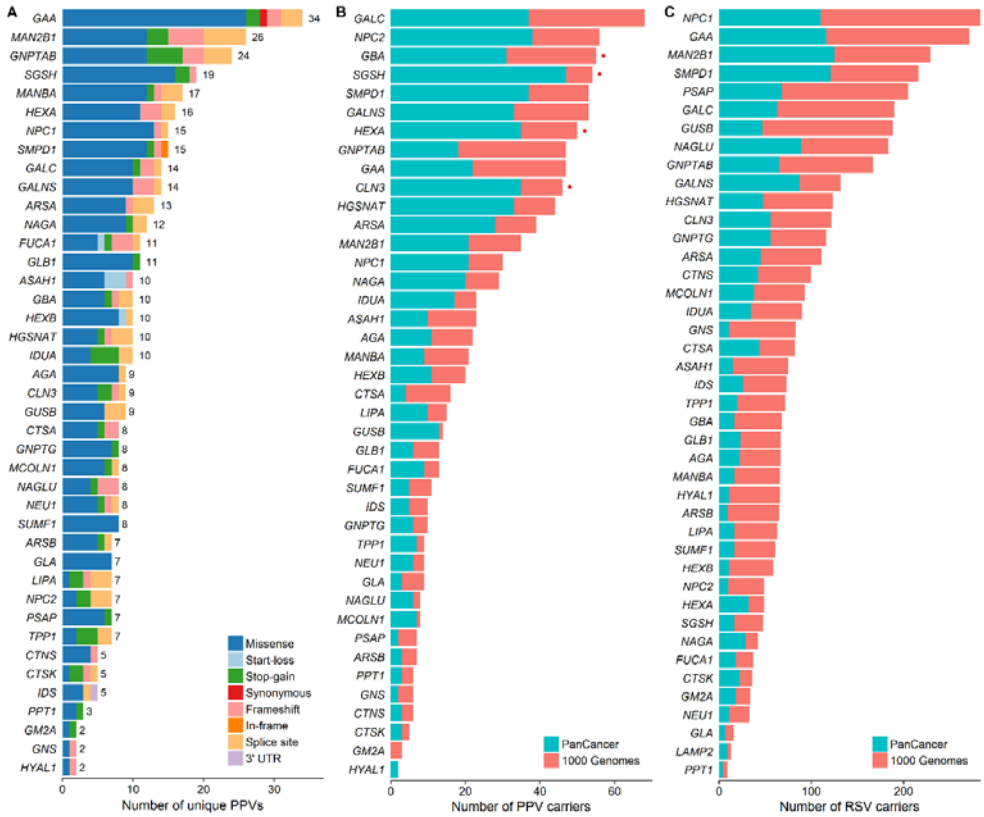

Figure S5. Number of unique potentially pathogenic variants (PPVs) in each lysosomal storage disease gene and gene-wise prevalence of PPVs and rare synonymous variants (RSVs) in the PanCancer and 1000 Genomes cohorts. Panel A shows the numbers of unique PPVs in 41 lysosomal storage disease genes (we identified no PPVs in *LAMP2* in our data set). Panels B and C show the prevalence of PPVs and RSVs, respectively, in each gene for the PanCancer and 1000 Genomes cohorts, which are indicated by color as shown by the legend in each panel. In panel B, genes significantly associated with cancer, as determined by SKAT-O analysis adjusted for population structure, are indicated with red dots on the right side of each bar. UTR denotes untranslated region.

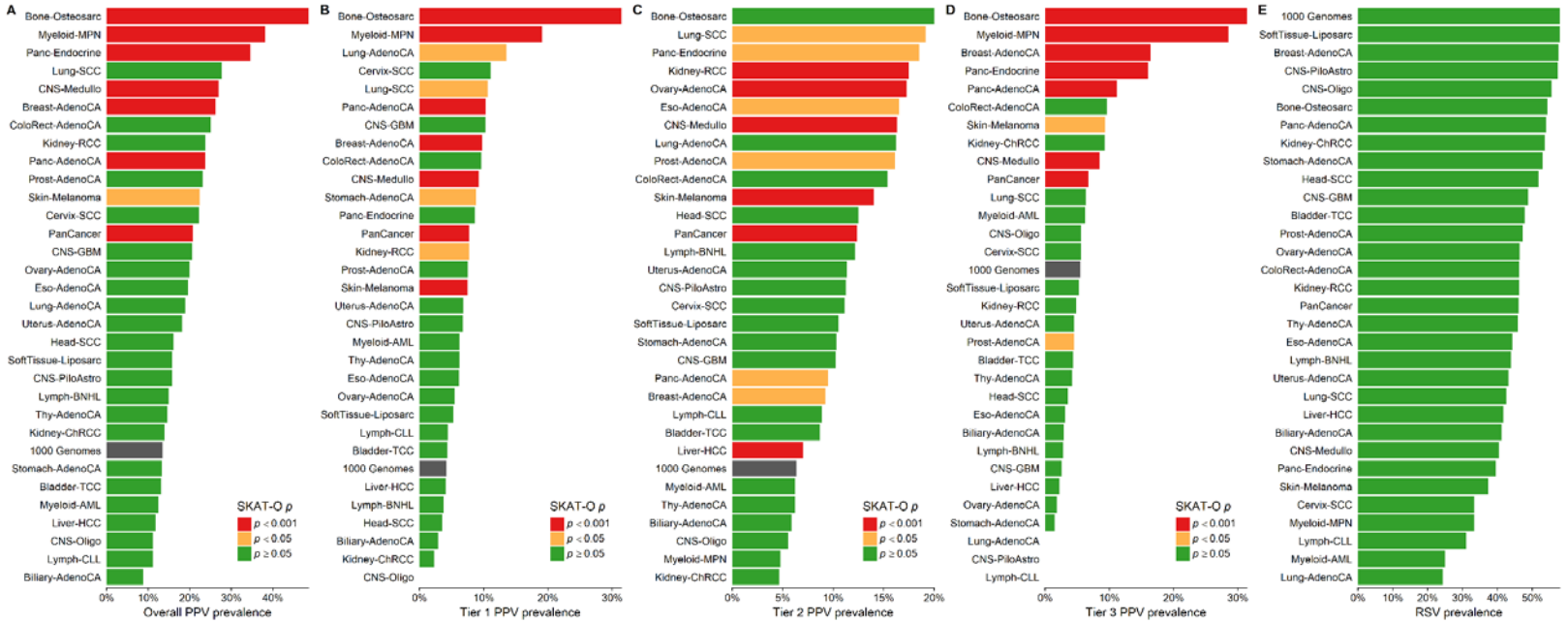

Figure S6. Prevalence of potentially pathogenic variants (PPVs) and rare synonymous variants (RSVs) in the 30 major histological subtypes of cancer (>15 individuals per subtype), PanCancer cohort, and 1000 Genomes cohort.

Panels A, B, C, and D show the prevalence of the entire PPV set and tier 1, 2, and 3 subsets, respectively. SKAT-O analysis adjusted for population structure was performed, and the corresponding P-values are indicated by color, as shown by the legend in each panel. Panel E shows the prevalence of RSVs. No statistical comparison was performed for RSV prevalence.

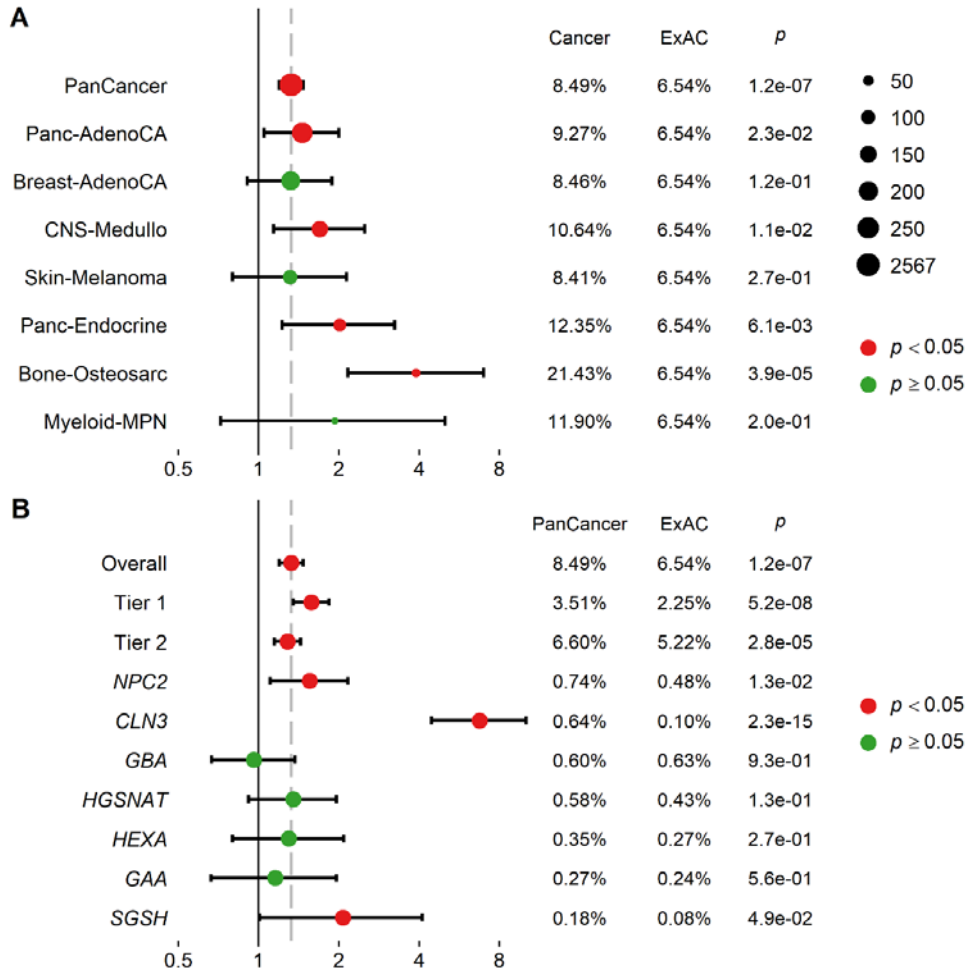

Figure S7. Validation analysis using the ExAC cohort as the control group.

Panel A shows the odds ratios and corresponding 95% confidence intervals for the prevalence of potentially pathogenic variants (PPVs) in eight significant cancer cohorts identified in the SKAT-O analysis versus the ExAC control cohort. The right side of the forest plot shows the PPV allele frequency (AF) in each cancer cohort and the ExAC cohort. The size of each dot corresponds to the number of patients in the cohort, as indicated in the legend at the right side of the panel. Panel B shows the odds ratios and corresponding 95% confidence intervals for the prevalence of 10 significant PPV sets identified in the SKAT-O analysis of the

PanCancer versus ExAC cohort. The right side of the forest plot shows the AF of PPVs belonging to each subset of the PanCancer and ExAC cohorts. The gray dashed lines in panels A and B represent the odds ratios for the PanCancer cohort and overall PPV set, respectively. In both panels, P-values are based on Fisher's exact tests. Because there were no tier 3 variants in the validation analysis (see Supplementary Methods section), the PPV frequency in the PanCancer cohort is lower here (8.5%) than in the main analysis (12.2%), which used the 1000 Genomes cohort as a control.

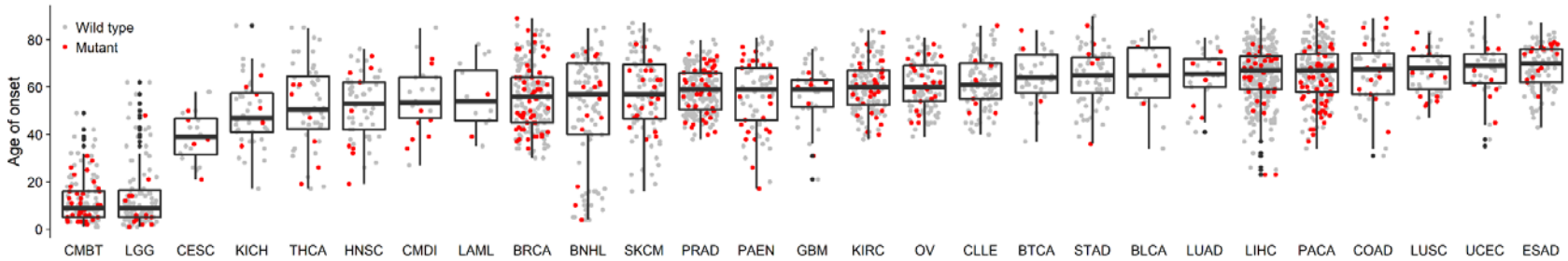

Figure S8. Age of cancer onset in major cancer cohorts.

Each patient is represented by a dot in a color that indicates the potentially pathogenic variant carrier status (red, mutant; gray, wild type). Only clinical cancer cohorts with more than 15 patients are included in this plot. Cohorts are shown in ascending order according to the median age of cancer onset (left to right). Abbreviations of cohort names are shown in Table S1.

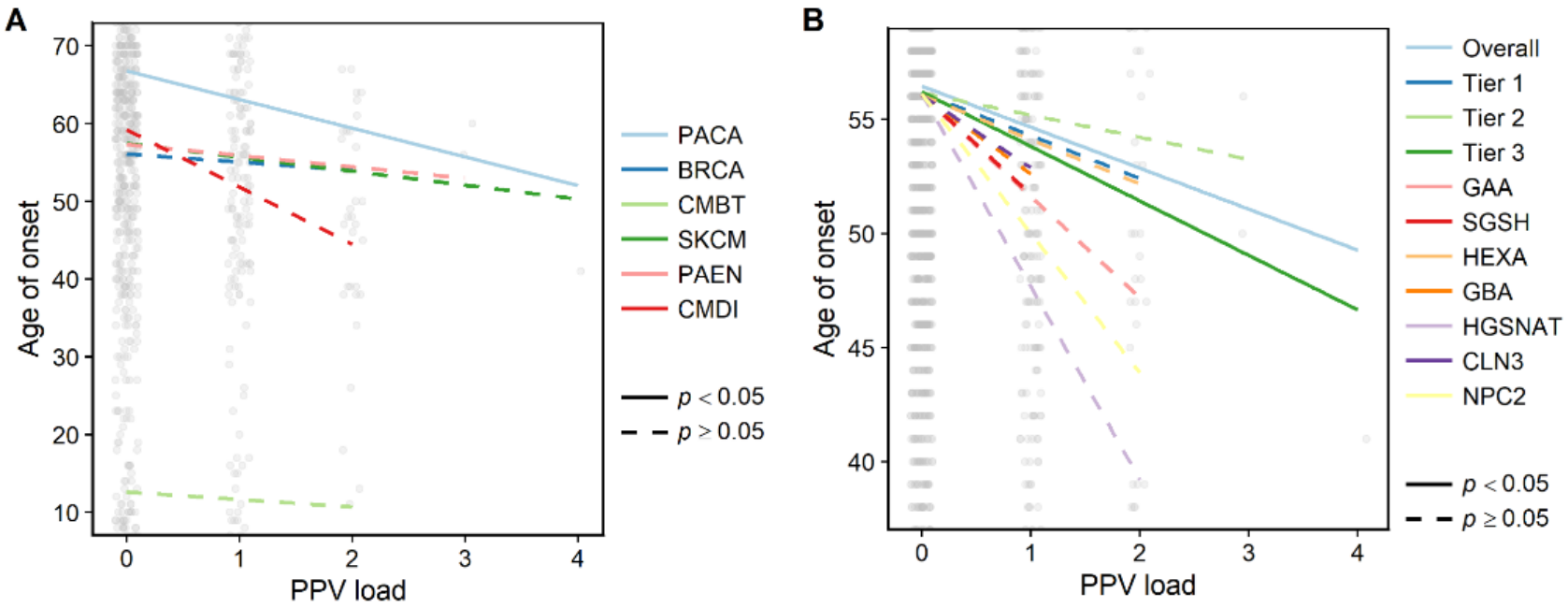

Figure S9. Linear correlation between potentially pathogenic variant (PPV) load and age of cancer onset.

Panel A shows linear correlations between PPV load (number of PPVs per individual) and age of cancer onset for six clinical cancer cohorts significantly associated with PPVs. Simple linear regression was performed for each cohort to draw and test the regression line for statistical significance. Panel B shows the linear correlation between PPV load and age of cancer onset in the PanCancer cohort for each set of PPVs

significantly associated with cancer. Linear regression adjusted for cancer histology was performed for each set of PPVs to draw and test the regression line for statistical significance. In both panels, each faint gray dot corresponds to a single patient. Note that not all patient dots are shown because the plots were magnified to clearly distinguish between regression lines. Abbreviations of the cohort names are defined in Table S1.

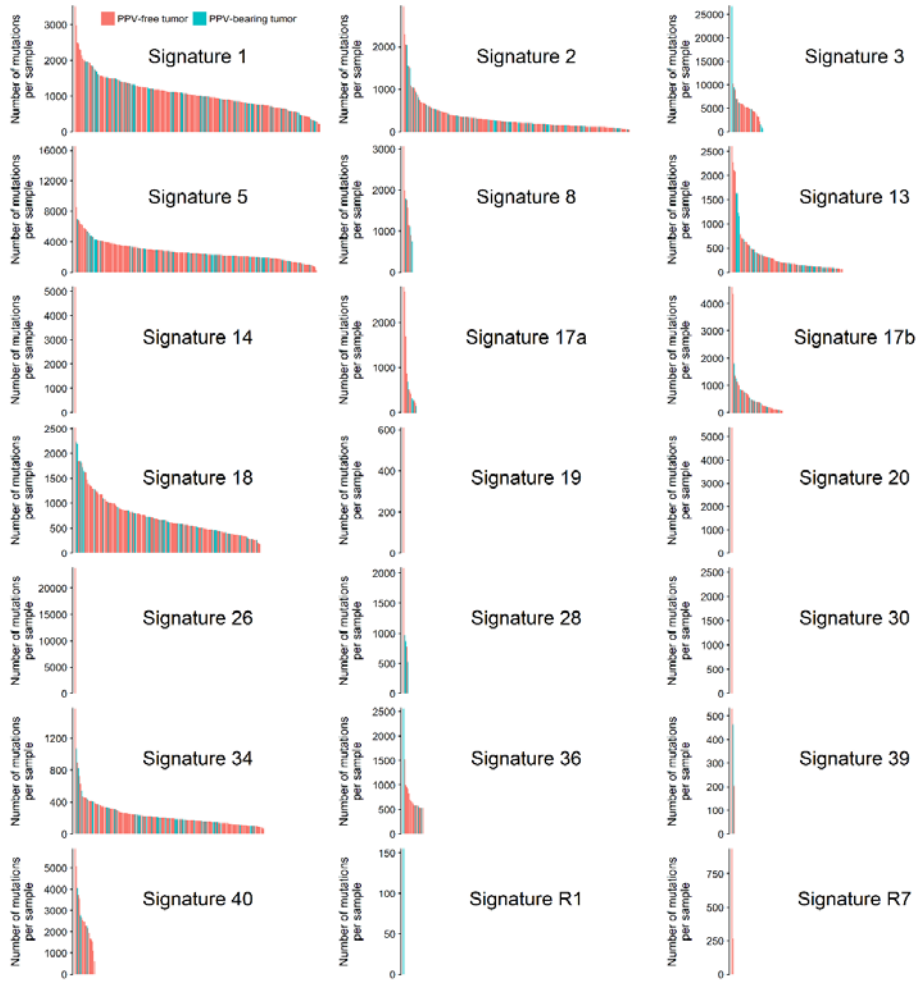

Figure S10. Contribution of each somatic mutation signature in pancreatic adenocarcinoma (n=232) ordered by mutational burden.

Tumors are shown in rows, and potentially pathogenic variant (PPV) carrier status is indicated by color. The height of each bar indicates the prevalence of the somatic mutation belonging to the specific mutation signature. A comparison between the PPV-bearing and PPV-free tumors of the proportion positive (Fisher's exact test) and mutational load (Student's *t*-test) for each mutation signature showed no significant differences (P>0.05). The data used here were the official beta version of mutation signatures for single nucleotide substitutions generated and uploaded to

the sftp server (sftp://dccsftp.nci.nih.gov/pancan/) by the PanCancer Analysis of Whole Genomes project team in March 2017.

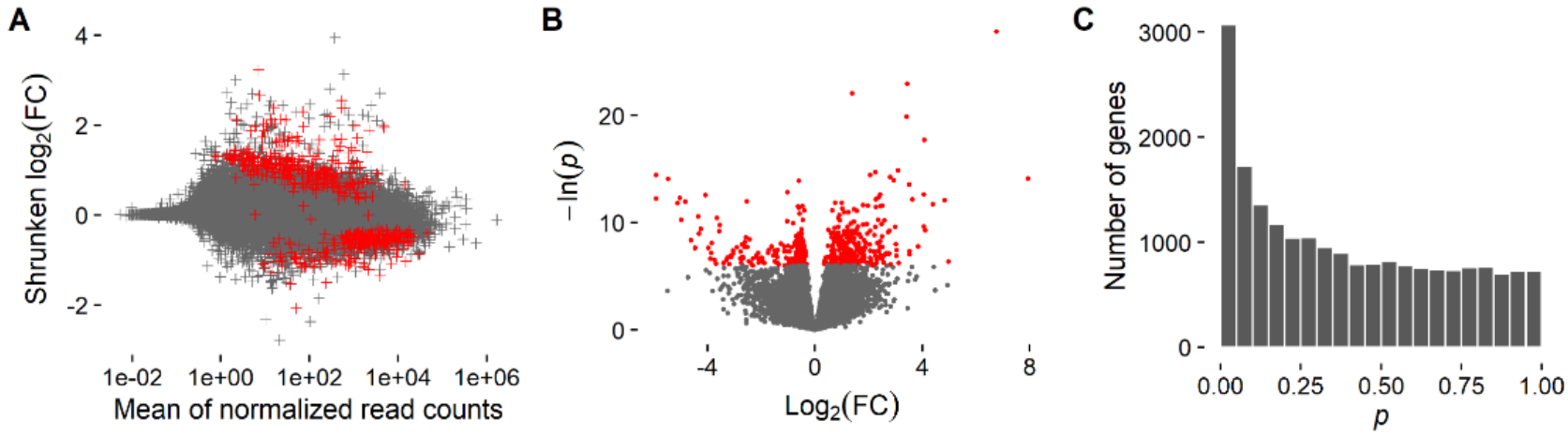

Figure S11. Differentially expressed gene (DEG) analysis reveals 287 gene upregulations and 221 downregulations in potentially pathogenic variant (PPV)-bearing pancreatic adenocarcinoma.

In Panel A the MA plot shows the average of the read counts normalized by size factor on the x-axis and shrunken fold change on the logarithmic scale of the y-axis. In Panel B the volcano plot shows fold changes on the logarithmic scale of the x-axis and P-values as minus logarithm on the y-axis. In both panels, genes with a false discovery rate (adjusted P-value) below 0.1 are shown in red. In Panel C the histogram of P-values shows a peak frequency below 0.05, strongly suggesting the existence of significantly up- or downregulated genes. FC denotes fold change.

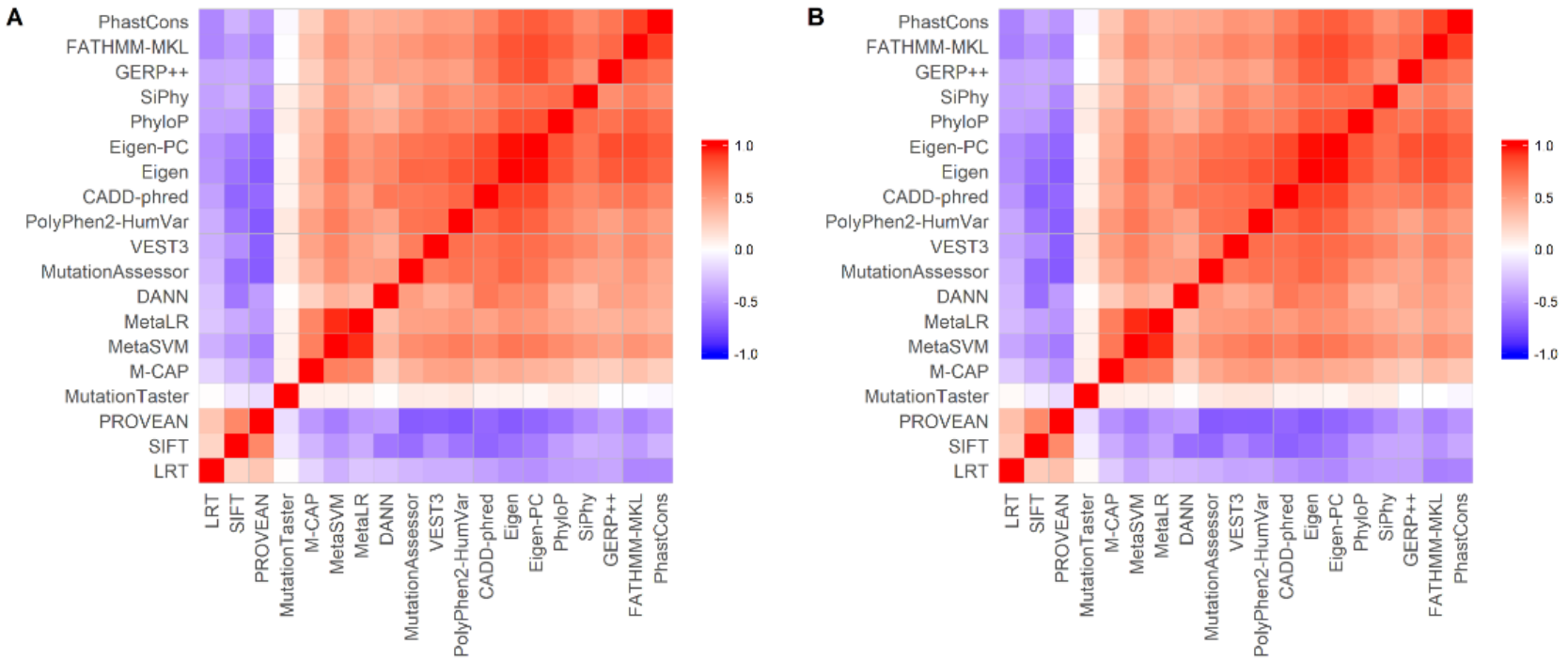

Figure S12. Pairwise correlations of the computationally predicted scores assigned to nonsynonymous single nucleotide variants (SNVs) by 19 separate in silico prediction tools.

Panel A and B show pairwise correlations of scores for nonsynonymous SNVs in the aggregate data set of the PanCancer and 1000 Genomes cohorts and the aggregate data set of the PanCancer and ExAC cohorts, respectively. The scores were provided by the tools themselves.

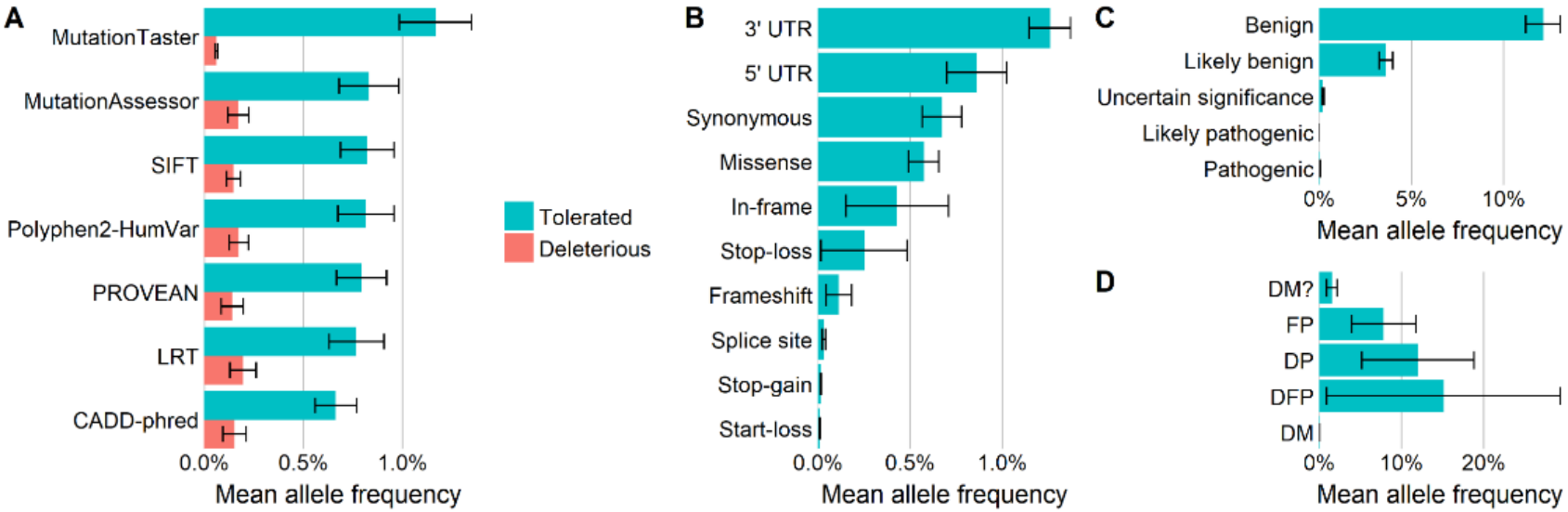

Figure S13. Deleterious variants are rare.

Panel A shows mean allele frequency of variants according to classification by seven representative in silico prediction algorithms. Panels B, C, and D show mean allele frequency of variants according to consequence type, ClinVar clinical significance, and HGMD classification, respectively. Error bars indicate standard errors. UTR denotes untranslated region, DM disease-causing mutation, DFP disease-associated polymorphism with supporting functional evidence, DP disease-associated polymorphism, FP functional polymorphism, and DM? possible pathological mutation but with some degree of uncertainty or conflicting evidence of pathogenicity.[62]

Figure S14. FPKM-UQ-normalized read counts of 42 lysosomal storage disease genes across histological subtypes of cancers with available transcriptomes.

Genes are arranged in descending order according to the number of unique variants

in the PanCancer and 1000 Genomes cohorts from left to right in each row and from top row to bottom row. Cancer histologies are arranged in descending order of the median FPKM-UQ read count from top to bottom in each plot. Red dots inside the violin plots indicate median values. Abbreviations of histological diagnosis are defined in Table S1.

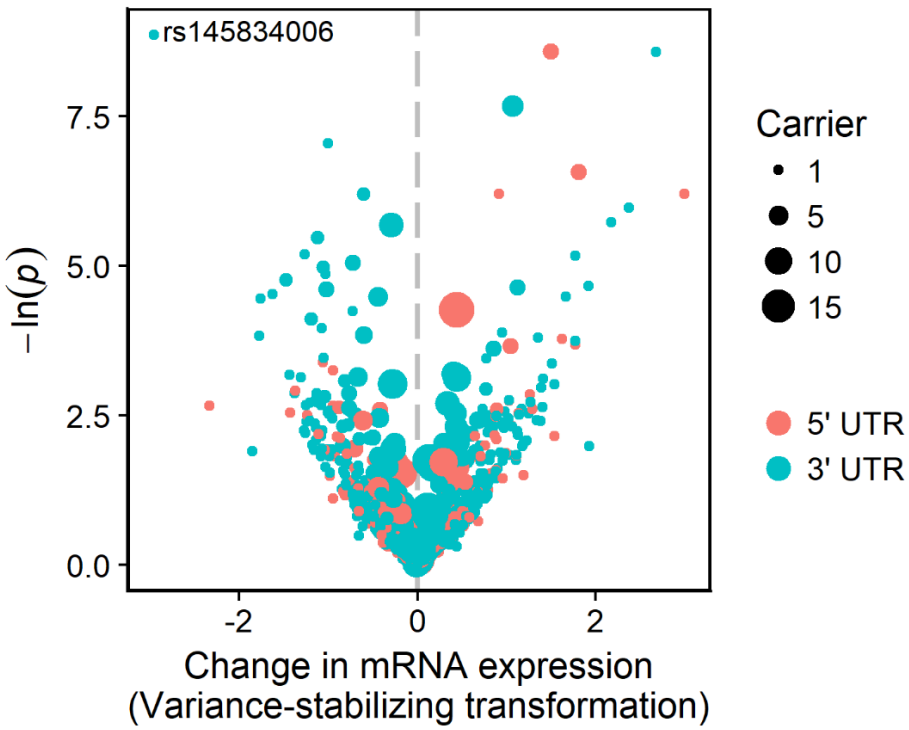

Figure S15. Relationship between 5' or 3' untranslated region (UTR) variants and change in mRNA expression.

All 3192 unique UTR variants with arithmetic mean of the PanCancer and 1000 Genomes allele frequency less than 0.5% are plotted as single dots in the volcano plot. The estimated change in the variance-stabilizing-transformed read count is plotted on the x-axis, and minus logarithm of the P-values are plotted on the y-axis. The size of each dot indicates the total number of carriers in the PanCancer and 1000 Genomes cohorts. Only one 3' UTR variant was associated with significantly lower mRNA expression of the corresponding gene (*IDS*) at the threshold of false discovery rate 0.1 and was annotated with the reference SNP ID (rs145834006).

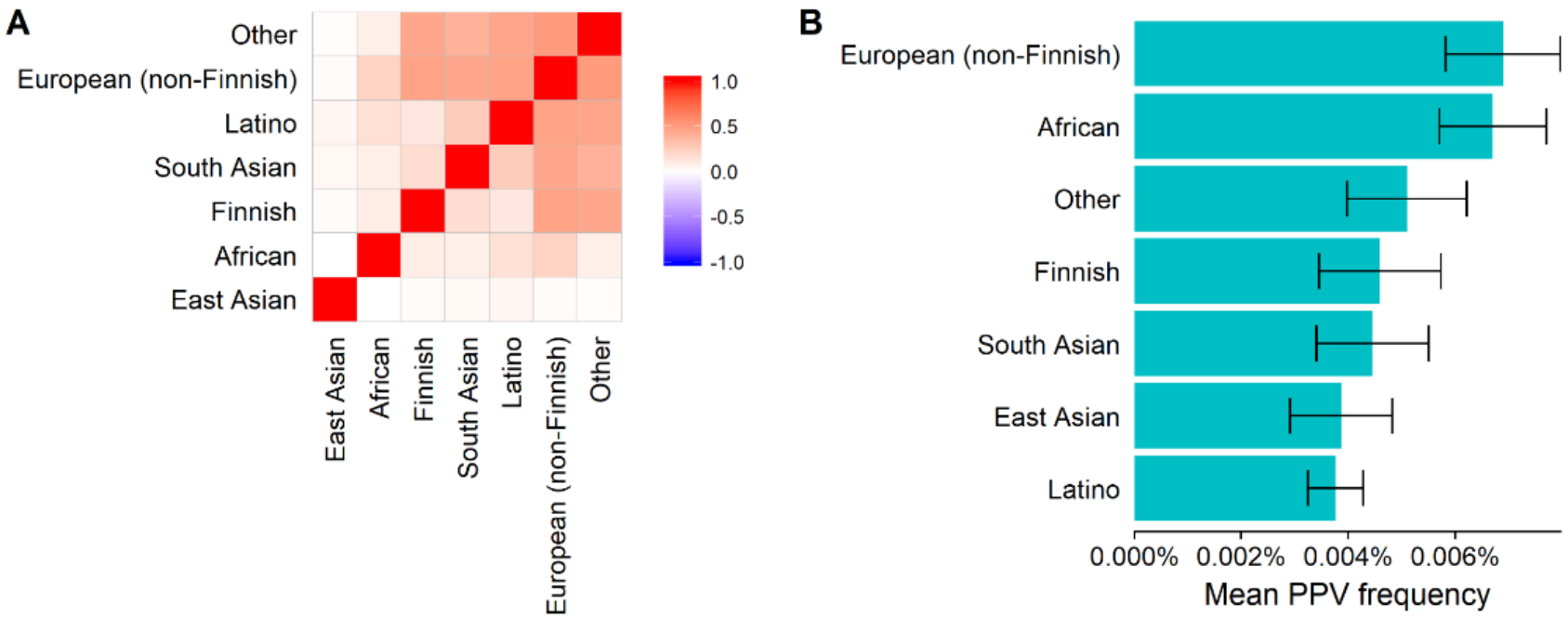

Figure S16. Ethnicity-specific potentially pathogenic variant (PPV) prevalence and its pairwise correlations among populations in the ExAC cohort.

Panel A shows the pairwise correlations of the PPV frequency between populations comprising the ExAC cohort. Panel B shows the average PPV frequency in each population of the ExAC cohort.

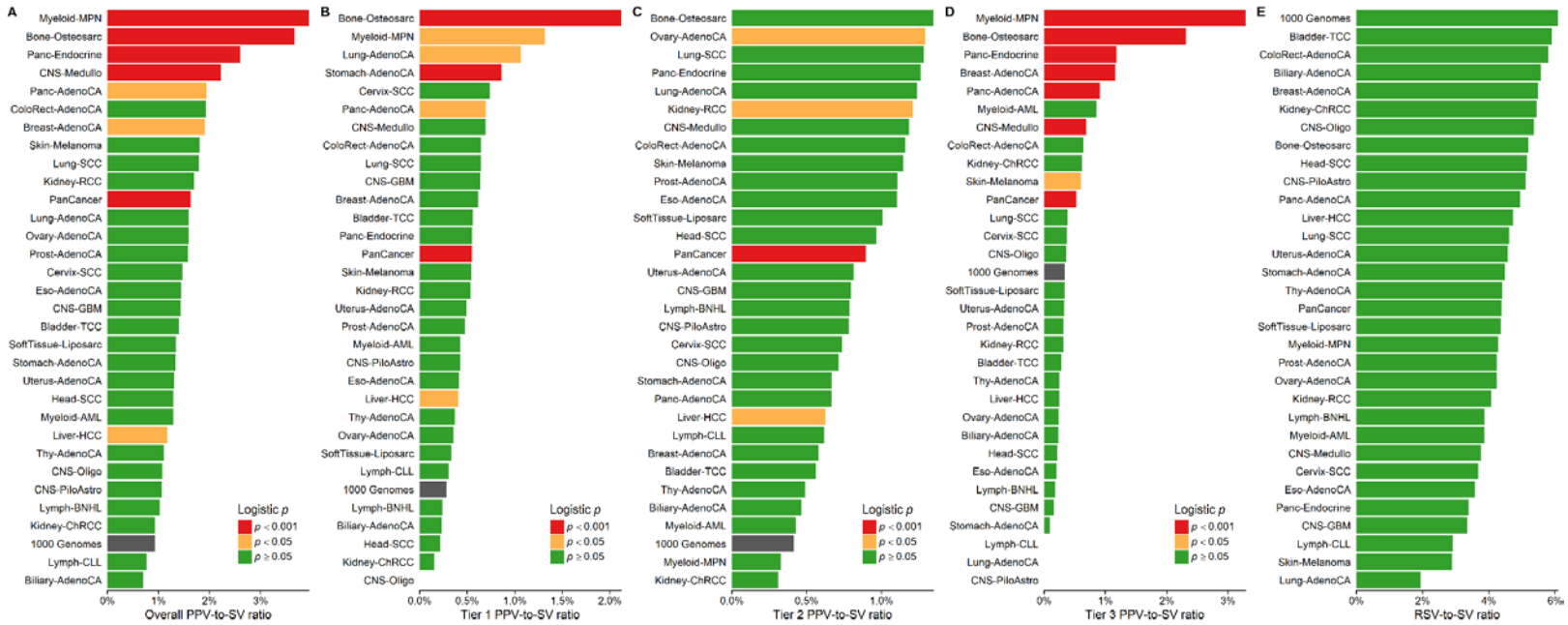

Figure S17. The ratio of the prevalence of potentially pathogenic variants (PPVs) and rare synonymous variants (RSVs) to synonymous variants in the 30 major histological subtypes of cancer (>15 individuals per subtype), PanCancer cohort, and 1000 Genomes cohorts.

Panel A, B, C, and D show the PPV-to-synonymous variant prevalence ratios for the entire PPV set and tier 1, 2, and 3 subsets, respectively. Weighted logistic regression analysis adjusted for population structure was performed, and the corresponding P-values are indicated by color. Panel E shows the RSV-to-synonymous variant prevalence ratios. No statistical comparison was performed for this variable.

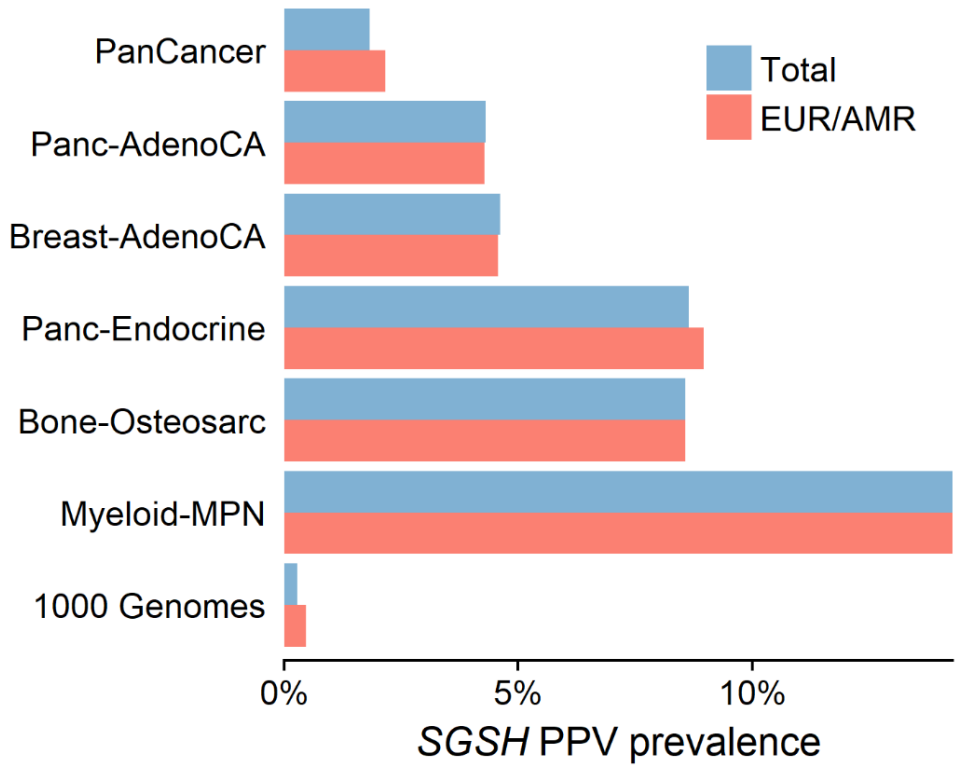

Figure S18. Prevalence of potentially pathogenic variants (PPVs) within the *SGSH* gene across the PanCancer cohort, the 1000 Genomes cohort, and five cancer subtypes that show significant SKAT-O association with *SGSH*.

The prevalence of PPVs in the *SGSH* gene in the entire cohort (blue) and in the subset of European and American individuals (red). Abbreviations for histological diagnoses are defined in Table S1. EUR and AMR denote European and American, respectively.

Supplementary References

1. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics 2010;11:94.

2. The Genomes Project C. A global reference for human genetic variation. Nature 2015;526:68-74.

3. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 2016;536:285-91.

4. Scriver CR. The metabolic and molecular bases of inherited disease. 8th ed. New York: McGraw-Hill; 2001.

5. Wang RY, Bodamer OA, Watson MS, Wilcox WR. Lysosomal storage diseases: Diagnostic confirmation and management of presymptomatic individuals. Genet Med 2011;13:457-84.

6. Boustany R-MN. Lysosomal storage diseases—the horizon expands. Nature reviews Neurology 2013;9:583-98.

7. Parenti G, Andria G, Ballabio A. Lysosomal storage diseases: from pathophysiology to therapy. Annu Rev Med 2015;66:471-86.

8. Futerman AH, van Meer G. The cell biology of lysosomal storage disorders. Nat Rev Mol Cell Biol 2004;5:554-65.

9. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38:e164-e.

10. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. Genome Biol 2016;17:122.

11. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Hum Mutat 2016;37:235-41.

12. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. Nat Meth 2010;7:248-9.

13. Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 2003;31:3812-4.

14. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 2014;46:310-5.

15. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res 2009;19:1553-61.

16. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Meth 2014;11:361-2.

17. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res 2011;39:e118-e.

18. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics 2015;31:2745-7.

19. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics 2015;31:761-3.

20. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the Variant Effect Scoring Tool. BMC Genomics 2013;14:S3.

21. Jagadeesh KA, Wenger AM, Berger MJ, et al. M-CAP eliminates a

majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet 2016;48:1581-6.

22. Dong C, Wei P, Jian X, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet 2015;24:2125-37.

23. Shihab HA, Rogers MF, Gough J, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics 2015;31:1536-43.

24. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet 2016;48:214-20.

25. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 2010;20:110-21.

26. Siepel A, Bejerano G, Pedersen JS, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 2005;15:1034-50.

27. Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics 2009;25:i54-62.

28. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 2010;6:e1001025.

29. Wenger DA, Coppola S, Liu S. Insights into the diagnosis and treatment

of lysosomal storage diseases. Arch Neurol 2003;60:322-8.

30. Pinto R, Caseiro C, Lemos M, et al. Prevalence of lysosomal storage diseases in Portugal. Eur J Hum Genet 2003;12:87-92.

31. Richards S, Aziz N, Bale S, et al. Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in medicine : official journal of the American College of Medical Genetics 2015;17:405-24.

32. Farazi TA, Hoell JI, Morozov P, Tuschl T. MicroRNAs in human cancer. MicroRNA Cancer Regulation: Springer; 2013:1-20.

33. Ryan BM, Robles AI, Harris CC. Genetic variation in microRNA networks: the implications for cancer research. Nat Rev Cancer 2010;10:389-402.

34. Yu Z, Li Z, Jolicoeur N, et al. Aberrant allele frequencies of the SNPs located in microRNA target sites are potentially associated with human cancers. Nucleic Acids Res 2007;35:4535-41.

35. Chin LJ, Ratner E, Leng S, et al. A SNP in a let-7 microRNA Complementary Site in the KRAS 3′ Untranslated Region Increases Non–Small Cell Lung Cancer Risk. Cancer Res 2008;68:8535-40.

36. Zhang L, Liu Y, Song F, et al. Functional SNP in the microRNA-367 binding site in the 3'UTR of the calcium channel ryanodine receptor gene 3 (RYR3) affects breast cancer risk and calcification. Proc Natl Acad Sci U S A 2011;108:13653-8.

37. Wang X, Ren H, Zhao T, et al. Single nucleotide polymorphism in the microRNA-199a binding site of HIF1A gene is associated with pancreatic ductal

adenocarcinoma risk and worse clinical outcomes. Oncotarget 2016;7:13717-29.

38. Tchatchou S, Jung A, Hemminki K, et al. A variant affecting a putative miRNA target site in estrogen receptor (ESR) 1 is associated with breast cancer risk in premenopausal women. Carcinogenesis 2009;30:59-64.

39. Lee I, Ajay SS, Yook JI, et al. New class of microRNA targets containing simultaneous 5'-UTR and 3'-UTR interaction sites. Genome Res 2009;19:1175-83.

40. Wang G, Guo X, Floros J. Differences in the translation efficiency and mRNA stability mediated by 5′-UTR splice variants of human SP-A1 and SP-A2 genes. American Journal of Physiology - Lung Cellular and Molecular Physiology 2005;289:L497-L508.

41. Valencia-Sanchez MA, Liu J, Hannon GJ, Parker R. Control of translation and mRNA degradation by miRNAs and siRNAs. Genes Dev 2006;20:515-24.

42. Fabian MR, Sonenberg N, Filipowicz W. Regulation of mRNA translation and stability by microRNAs. Annu Rev Biochem 2010;79:351-79.

43. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol 2010;11:R106.

44. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics 2012;13:762-75.

45. Zhang J, Walsh MF, Wu G, et al. Germline Mutations in Predisposition Genes in Pediatric Cancer. N Engl J Med 2015;373:2336-46.

46. Lu C, Xie M, Wendl MC, et al. Patterns and functional implications of rare germline variants across 12 cancer types. Nature Communications 2015;6:10086.

47. Pritchard CC, Mateo J, Walsh MF, et al. Inherited DNA-Repair Gene

Mutations in Men with Metastatic Prostate Cancer. N Engl J Med 2016;375:443-53.

48. Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. Nat Genet 2012;44:243-6.

49. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20:1297-303.

50. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. Bioinformatics 2011;27:2156-8.

51. Purcell S, Neale B, Todd-Brown K, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics 2007;81:559-75.

52. Novembre J, Johnson T, Bryc K, et al. Genes mirror geography within Europe. Nature 2008;456:98-101.

53. Zhang Y, Guan W, Pan W. Adjustment for population stratification via principal components in association analysis of rare variants. Genet Epidemiol 2013;37:99-109.

54. Liu Q, Nicolae DL, Chen LS. Marbled Inflation From Population Structure in Gene-Based Association Studies With Rare Variants. Genet Epidemiol 2013;37:10.1002/gepi.21714.

55. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15:550.

56. Bourgon R, Gentleman R, Huber W. Independent filtering increases detection power for high-throughput experiments. Proceedings of the National

Academy of Sciences 2010;107:9546-51.

57. Chen JK, Taipale J, Cooper MK, Beachy PA. Inhibition of Hedgehog signaling by direct binding of cyclopamine to Smoothened. Genes Dev 2002;16:2743-8.

58. Lee MJ, Hatton BA, Villavicencio EH, et al. Hedgehog pathway inhibitor saridegib (IPI-926) increases lifespan in a mouse medulloblastoma model. Proceedings of the National Academy of Sciences 2012;109:7859-64.

59. Coon V, Laukert T, Pedone CA, Laterra J, Kim KJ, Fults DW. Molecular Therapy Targeting Sonic Hedgehog and Hepatocyte Growth Factor Signaling in a Mouse Model of Medulloblastoma. Mol Cancer Ther 2010;9:2627-36.

60. Gajjar A, Stewart CF, Ellison DW, et al. Phase I Study of Vismodegib in Children with Recurrent or Refractory Medulloblastoma: A Pediatric Brain Tumor Consortium Study. Clin Cancer Res 2013;19:6305-12.

61. Robinson GW, Orr BA, Wu G, et al. Vismodegib Exerts Targeted Efficacy Against Recurrent Sonic Hedgehog–Subgroup Medulloblastoma: Results From Phase II Pediatric Brain Tumor Consortium Studies PBTC-025B and PBTC-032. J Clin Oncol 2015;33:2646-54.

62. Stenson PD, Mort M, Ball EV, Shaw K, Phillips AD, Cooper DN. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet 2014;133:1-9.

초록

융해소체축적병 연관 유전자의 생식선 돌연변이와 암 발생의 연관성

신정훈

의과대학 의학과 중개의학전공

서울대학교 대학원

서론

융해소체축적병은 융해소체의 기능에 관여하는 유전자의 돌연변이에 의해 발생하는 선천성 대사이상 질환이다. 일부 융해소체축적병에서 세포 내의 거대분자 축적이 장기적으로 암 발생을 조장할 수 있다는 사실이 기 알려져 있다.

방법

국제 암 유전체 컨소시움 (International Cancer Genome Consortium; ICGC) 및 1000 유전체 프로젝트에서 생산한 암 환자 (ICGC 코호트) 및

정상 성인 (1000 유전체 코호트)의 전장 유전체 염기서열 자료를 이용하여, 42개의 융해소체축적병 연관 유전자 내에 존재하는 생식선 돌연변이와 암 발생의 연관성을 분석하였다. 또한 이러한 생식선 돌연변이를 가진 암 환자와 그렇지 않은 암 환자 간의 암 발병 연령의 차이 및 암 조직의 체세포 돌연변이와 유전자 발현 양상의 차이를 규명하였다.

결과

융해소체축적병-연관 생식선 돌연변이의 수는 1000 유전체 코호트에 비해 ICGC 코호트에 유의하게 높은 빈도로 관찰되었다 (20.7% 대 13.5%, 유의확률 = 8.7×10^{-12}). 암 발생의 위험도는 이러한 생식선 돌연변이를 많이 가지고 있는 사람일수록 더 높은 경향성을 보였다. 인구집단 구조의 차이를 보정한 SKAT-O 분석에서 36개의 종양 아형-유전자 쌍이 통계적으로 유의한 연관성을 보이는 것으로 확인되었다. 이러한 결과는 ExAC 코호트를 대조군으로 이용한 분석에서도 재현되었다. 췌장암, 피부암, 만성 골수질환에서 융해소체축적병-연관 생식선 돌연변이를 동반한 환자의 암 발병 연령이 그렇지 않은 환자의 암 발병 연령에 비해 유의하게 낮은 것으로 확인되었다. 232개의 췌장암 조직의 전사체 데이터를 분석한 결과 508개의 유전자가 생식선 돌연변이 유무에 따라 발현 강도가 유의하게 달라지는 것으로 확인되었고, 이러한 유전자는 특히 췌장암 발병에 관여하는 것으로 기 알려진 신호전달경로에

집중적으로 분포되어 있었다.

결론

융해소체축적병 연관 유전자의 생식선 돌연변이를 가진 사람은 암 발생 위험도가 정상 성인에 비해 높다. 이러한 유전자의 기능 이상을 회복시킬 수 있는 다양한 치료 방법들이 존재하므로 본 연구 결과는 향후 맞춤형 암 예방에 활용될 수 있을 것으로 기대된다.

주요어: 융해소체축적병; 생식선 돌연변이; 희귀 변이; 암; 연관성; 전사체; 유전자 발현