Oncogenic effects of germline mutations in lysosomal storage disease genes

Junghoon Shin, M.D.[1,2], Daeyoon Kim, M.S.[2], Hyung-Lae Kim, M.D., Ph.D.[3], Murim Choi, Ph.D.[4], Jan O. Korbel, Ph.D.[5], Sung-Soo Yoon, M.D., Ph.D.[1,2], and Youngil Koh, M.D., Ph.D.[1,2,6] on behalf of the PCAWG Germline Cancer Genome Working Group and the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Network

[1]Division of Hematology and Medical Oncology, Department of Internal Medicine, Seoul National University Hospital, Seoul, Republic of Korea; [2]Cancer Research Institute, Seoul National University College of Medicine, Seoul, Republic of Korea; [3]Department of Biochemistry, Ewha Womans University School of Medicine, Seoul, Republic of Korea; [4]Department of Biomedical Sciences, Seoul National University College of Medicine, Seoul, Republic of Korea; [5]European Molecular Biology Laboratory, Genome Biology Unit, 69117, Heidelberg, Germany; [6]Biomedical Research Institute, Seoul National University College of Medicine, Seoul, Republic of Korea

Junghoon Shin and Daeyoon Kim contributed equally to this article.

Correspondence should be addressed to Sung-Soo Yoon and Youngil Koh.

Correspondence

Sung-Soo Yoon, M.D., Ph.D.

Division of Hematology and Medical Oncology, Seoul National University Hospital, 101, Daehak-ro, Jongno-gu, Seoul 03080, Republic of Korea

Tel: +82-2-2072-3079, Fax: +82-2-762-9662

Email: ssysmc@gmail.com

Youngil Koh, M.D., Ph.D.

Division of Hematology and Medical Oncology, Seoul National University Hospital, 101, Daehak-ro, Jongno-gu, Seoul 03080, Republic of Korea

Tel: +82-2-2072-7217, Fax: +82-2-2072-7379

Email: go01@snu.ac.kr

Abstract

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 lysosomal dysfunction may facilitate carcinogenesis.

Methods: Using whole genome sequencing data from the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Network (Pan-Cancer cohort) and the 1000 Genomes project (1000 Genomes cohort), we analyzed the association between potentially pathogenic variants (PPVs) in 42 LSD genes and cancer. We evaluated age at diagnosis of cancer and patterns of somatic mutation and gene expression according to the PPV carrier status.

Results: PPV prevalence in the Pan-Cancer 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 the ExAC cohort as an independent control. Cancer developed earlier in carriers of PPVs in pancreatic adenocarcinoma (*MAN2B1*, *GALNS*, and *GUSB*), skin cancer (*NPC2*), and chronic myeloid disorder (*SGSH*) as well as in the Pan-Cancer cohort. Analysis of RNA-Seq data from the pancreatic cancer cohort revealed 508 genes differentially expressed according to the PPV carrier status, which were highly enriched in the known core signaling pathways of pancreatic cancer.

Conclusion: Carriers of germline PPVs in LSD 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.

Introduction

Lysosomal storage diseases (LSDs) comprise more than 50 disorders caused by inborn errors of metabolism, which involve the impaired function of endosome-lysosome proteins.[1] In LSDs, defects in genes encoding lysosomal hydrolases, transporters, or enzymatic activators result in macromolecule accumulation in the late endocytic system.[2] The disruption of lysosomal homeostasis is linked to increased endoplasmic reticulum and oxidative stress, which not only is a common mediator of apoptosis in LSDs but also can induce oncogenic cellular phenotype and promote development of malignancy.[3,4]

Typical LSD patients have severely impaired organ functions and 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.[5] Clinical observations have also shown that the patients with Gaucher disease or Fabry disease are at increased risk of cancer, indicating the link between dysregulated lysosomal metabolism and carcinogenesis.[6,7] However, the precise relationship between the lysosomal dysfunction and cancer remains unclear; this uncertainty can be attributed in part to the diverse and nonspecific phenotypes of LSDs and the resulting difficulty in recognizing patients with mild symptoms. The extensive allelic heterogeneity and the complex genotype-phenotype relationships make the diagnosis more challenging.[8] Furthermore, growing evidence suggests that the single allelic loss is also functionally significant, even though the impact may not be sufficient to develop overt disease.[9] Considering the above along with the recessive inheritance nature of most LSDs, we hypothesized that there would be a large number of undetected carriers of LSD-causing mutations with mild functional impairment, and these carriers would be at increased risk of cancer.

Here we report the results of a comprehensive association analysis between potentially pathogenic germline mutations in LSD-related genes and cancer using data from global sequencing projects. We aimed 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 focus on the pancreatic adenocarcinoma.

Methods

Study populations

We used matched tumor-normal pair whole genome and tumor whole transcriptome sequences and the clinical data of 2567 cancer patients (Pan-Cancer cohort) from the International Cancer Genome Consortium (ICGC)/The Cancer Genome Atlas (TCGA) Pan-Cancer Analysis of Whole Genomes (PCAWG) project.[10] 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).[11] The second data set contained exomes of 53105 unrelated individuals from a subset of the Exome Aggregation Consortium release 1.0 that did not include TCGA subset (ExAC cohort).[12]

Potentially pathogenic variant selection

Through an extensive literature review, we identified 42 LSD genes (Table 1).[1,8,13-15] The potentially pathogenic variants (PPVs) that showed strong evidence of disease causation or functional defect were selected based on their consequence on transcripts or proteins, clinical and experimental evidence obtained from curated databases and medical literature, and *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 association between PPVs and cancer. In the first step, the Pan-Cancer 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.[16] We adjusted for population structure using principal component analysis on 10494 tag single nucleotide polymorphisms (Figure S2 and

S3 in the Supplementary Appendix). The second step used the ExAC cohort as a control population, confined the analysis to protein-coding regions covered by more than half of the ExAC exomes, and used Fisher's exact tests to validate the preceding results. Age at diagnosis of cancer was compared using Wilcoxon rank sum tests and linear regression. 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) pathways, respectively.[17-19]

Correction for multiple testing was conducted using the false discovery rate (FDR) estimation procedure, and the tail area-based FDR (also referred to as *q*-value) was reported.[20] All tests were two-tailed unless otherwise specified. We considered FDR<0.1 or P<0.05 (when not adjusted for multiple testing) significant. Statistical analysis was performed using R version 3.4.4 (R Foundation for Statistical Computing, Vienna, Austria) with packages of Bioconductor version 3.6. Details on 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 Pan-Cancer 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 at diagnosis of cancer 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 Pan-Cancer cohort.[11] Seven populations were represented in the ExAC cohort, among which the American and Non-Finnish European accounted for more than 60% collectively.[12]

PPV prevalence in the Pan-Cancer and 1000 Genomes cohorts

In the 42 LSD genes, 7187 germline single nucleotide variants and small insertions and deletions in protein-coding regions, essential splice junctions, and 5' and 3' untranslated regions

were identified in the aggregate variant call set of the Pan-Cancer 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 and Table S2 in the Supplementary Appendix). Overall, PPV prevalence was 20.7% in the Pan-Cancer 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. This tendency was broadly consistent across individual tiers, although some differences were not significant for each tier.

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 Pan-Cancer and 1000 Genomes cohort after adjustment for population structure, indicating that the enrichment of the Pan-Cancer cohort for PPVs was not likely due to batch effects (Figure 2B). The gene-specific prevalence of PPVs and RSVs in each cohort is shown in Figure S5B and S5C in the Supplementary Appendix, respectively. The results show that PPVs were relatively more abundant in the Pan-Cancer 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 it was significantly higher in seven subtypes of cancer than in the 1000 Genomes cohort (Figure S6A in the Supplementary Appendix). Results of tier-based analyses were consistent, except that prevalence of tier 2 PPVs in myeloproliferative neoplasm was numerically lower than that of the 1000 Genomes cohort (Figure S6B–D in the Supplementary Appendix). In contrast, RSV prevalence showed much less variation across

cohorts and was higher in the 1000 Genomes cohort than any histological subtype of cancer (Figure S6E in the Supplementary Appendix), reflecting the more heterogeneous nature of ancestry in this cohort. SKAT-O analysis with adjustment for population structure unveiled 37 significantly associated cancer-gene pairs and four genes (*GBA*, *SGSH*, *HEXA*, and *CLN3*) with the pan-cancer association (Figure 2D, Figure S5B, and Table S3 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 subtype. No evidence of systematic inflation of test statistics was observed (Figure 2C).

PPV prevalence in the Pan-Cancer 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 Pan-Cancer cohort or three or more histological subtypes of cancer 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 Pan-Cancer, pancreatic adenocarcinoma, medulloblastoma, pancreatic neuroendocrine carcinoma, and osteosarcoma cohorts. In addition, all tested PPV sets except *GBA* were more prevalent in the Pan-Cancer cohort than in the ExAC cohort, and six were significantly different (Figure S7B).

Age at diagnosis of cancer according to the PPV carrier status

The age at diagnosis of cancer in the major cancer cohorts is shown in Figure S8 in the Supplementary Appendix. We compared the age according to the PPV carrier status, with emphasis on the cancer cohorts and PPV sets that underwent validation analysis. The age at diagnosis of cancer was lower in the PPV carriers in all evaluated cohorts (one-sided $P<0.5$; Figure 3A). This difference was significant in the following cohorts: Pan-Cancer (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 Pan-Cancer cohort, carriers of PPVs in tier 1, tier 3, *HGSNAT*, *CLN3*, and *NPC2* exhibited

significantly earlier diagnosis of cancer (Figure 3B). Moreover, the PPV load (number of PPVs per individual) showed a significant negative linear correlation with age at diagnosis of cancer in the Pan-Cancer and pancreatic adenocarcinoma cohorts (Figure S9 in the Supplementary Appendix). Expanded analysis encompassing all histologies and genes revealed earlier cancer diagnosis in PPV carriers in five additional cancer-gene pairs (Figure 3C), three of which (pancreatic adenocarcinoma-*MAN2B1*, cutaneous melanoma-*NPC2*, and chronic myeloid disorder-*SGSH*) were in concordance with the SKAT-O test results (Figure 2D).

Somatic mutations and gene expression patterns in the 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 the PPV-bearing tumors (n=55) and the PPV-free tumors (n=177). The 50 most frequently mutated genes in each group are shown in Figure S10 in the Supplementary Appendix. 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.[21,22] Non-silent mutation burden was similar between groups (mean 57.1 versus 56.3 mutations per sample for PPV-bearing and PPV-free tumors, respectively; P=0.9). Mutational signature also did not differ according to the PPV carrier status (Figure S11 in the Supplementary Appendix). On the other hand, DEG analysis using the RNA-Seq data revealed 287 significant upregulations and 221 significant downregulations in the PPV-bearing tumors (Figures S12 and S13 and Table S4 in the Supplementary Appendix). Pathway-based analysis with GAGE identified 63 KEGG pathways significantly segregated by the PPV carrier status (Figure 4 and Figure S14 in the Supplementary Appendix). Remarkably, these pathways included at least six among 13 core signaling pathways whose component genes have been shown to be recurrently mutated in pancreatic cancer: Ras signaling, Wnt signaling, axon guidance, cell cycle regulation, focal adhesion, cell adhesion, and ECM-receptor interaction pathways.[22,23] In addition, our data suggested that deleterious mutations in LSD genes can provoke perturbations in neurodegenerative disease pathways involved in

development of Parkinson disease, Alzheimer disease, and Huntington disease, all of which have shown association with LSDs.[1] The glycerophospholipid metabolism pathway was also identified, indicating dysregulation in lysosomal function.

Supportive data

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

Discussion

In this study, we revealed previously unidentified associations between the pathogenic germline mutations in LSD genes and cancer using global genome-wide sequencing data. Analysis by subgrouping of PPVs into three tiers based on different selection criteria, validation with an independent control cohort, and comparison of the results with those obtained by using synonymous variants with matched allele frequency all showed broadly consistent results, substantiating the robustness of findings in our study. The genetic association was further supported by the significant difference in age at diagnosis observed for PPV carriers of at least three histological subtypes as well as the Pan-Cancer cohort compared with the PPV-free cancer patients.

The lysosome is involved in many other cellular functions than biomolecule catabolism, such as intracellular signaling, nutrient sensing, cellular growth regulation, plasma membrane repair, and phagocytosis.[15] These diverse roles of the lysosome are consistent with complex phenotypes of LSDs which can involve virtually any organ.[1] It has long been evident that patients with Gaucher disease are at markedly increased risk of malignancy, especially multiple myeloma with the risk estimated at approximately 50-fold.[24] However, despite the largely shared pathogenesis, the relationship between most other LSDs and cancer has not been studied due to the rarity and phenotypic heterogeneity of each disease, leaving this area largely unexplored. The wide spectrum of tumor histologies and LSD genes covered in this study enabled elucidation of numerous cancer-gene pairwise associations most of which had previously been unknown.

In a gene-based analysis, we identified four genes showing a significant pan-cancer association; among those, *SGSH* and *CLN3* were also strongly linked to five and four histological subtypes of cancer, 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, an adult-onset disease that presents primarily with visceral manifestations without neurological abnormality has also been reported.[25] A recent *in vivo* study showed the importance of oxidative stress in the pathobiology of Sanfilippo syndrome A.[26] Because the oxidative stress is also a key mediator of cancer cell growth, invasiveness, and angiogenesis, the inherited *SGSH* mutations may contribute to an elevated cancer risk via persistent cellular exposure to oxidative stress.[4] CLN3 is a late endosomal and lysosomal transmembrane protein, and its 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.[14] In line with its control over apoptosis, CLN3 also regulates cancer cell growth, suggesting its therapeutic implications.[27] Results from our study warrant further investigation of this protein as a molecular target for cancer therapy.

Almost 5 to 10 percent of pancreatic cancer patients are diagnosed before the age of 50.[28] For these patients, positive family history is the most important risk factor, indicating the importance of inherited genetic variants.[29] Indeed, many predisposition genes have been identified (e.g., *BRCA1/2* and *PALB2*). However, in a majority of the early-onset pancreatic cancer patients, the cause remains unclear.[30] In our histology-specific analysis, pancreatic adenocarcinoma showed a particularly strong association with PPVs and early cancer diagnosis in mutation carriers, motivating us to evaluate signatures of somatic mutations and gene expression in this histological subset. The DEG analysis revealed many genes with a significantly higher or lower level of expression in PPV carriers, and GAGE analysis provided novel insights into the pathways that are involved in pancreatic carcinogenesis in these patients. Remarkably, many of these pathways were previously implicated in pancreatic cancer development in transcriptome and exome

sequencing studies.[22,23] In contrast, the somatic mutation burden and mutational signatures were comparable between the PPV-bearing and PPV-free tumors. Overall, these results suggest that transcriptional misregulation is a key mediator linking defective lysosomal function to pancreatic carcinogenesis. In addition, decades of research on tumor suppressor genes (TSGs) involved in hereditary cancer predisposition syndromes have proposed a continuum model of tumor suppression emphasizing the importance of subtle change in expression of TSGs in oncogenesis.[31] Given the rarity of individual PPVs, almost all PPV carriers observed in our study were heterozygous for the particular variants. Therefore, our findings corroborate the dosage effect model for LSD genes that was already suggested in a previous study.[9]

It is important to note that the Pan-Cancer cohort consisted primarily of adult patients. Thus, highly pathogenic variants that cause fatal outcomes in childhood might be underrepresented in the selected PPVs. This can explain the higher-than-expected population PPV frequency and the relatively low ORs compared to those observed for known high-penetrance CPMs.[32,33] In addition, the 1000 Genomes cohort is not a completely cancer-free population, because donors were selected based on their previous history at the time of enrollment. Therefore, it is not surprising that over one-tenth of the population carried PPVs.

This study has limitations. First, because we did not process the raw sequence data but used variant call sets downloaded online, the possibility of batch effects cannot be excluded, even considering the similarity in pipelines used to generate each data set.[10-12] Second, we could not adjust for the population structure in analysis of the ExAC cohort as a control because the individual-level genotype data were not accessible. However, the ExAC cohort represents a similar mix of populations to the Pan-Cancer cohort, with ~70% of the entire cohort comprised of the Americans and Europeans.[12] Therefore, consistent results from the validation analysis corroborate our findings. Third, an independent cancer cohort with a sufficient size was not available for external validation. Fourth, hematological malignancies such as myeloma, the most widely known LSD-associated cancer, were poorly represented in the Pan-Cancer cohort, and the numbers of patients with individual cancer types were not sufficiently large to draw reliable histology-specific conclusions.

From a therapeutic perspective, LSD genes are attractive targets because of the mechanistically intuitive nature of enzyme replacement and substrate reduction therapies. Enzyme replacement therapy has already been approved for at least seven LSDs.[34] Other promising therapeutic approaches include pharmacological chaperones, gene therapy, and compounds that "read through" the early stop codon introduced by nonsense mutations.[34] Although it is unclear whether preemptive treatment can prevent or delay long-term complications of LSDs such as cancer, our study renders it promising to utilize LSD therapies for individualized cancer prevention and treatment. In conclusion, the present study provides a comprehensive landscape of associations between germline mutations in LSD genes and cancer. Investigating the crosstalk between treatable metabolic diseases and cancer is crucial because it can form a basis for precision cancer prevention. Diverse and increasingly sophisticated therapeutic options to restore lysosomal functions are currently available or being developed. Future clinical trials of these agents guided by individual mutation profiles may pave a new way toward personalized cancer prevention and treatment.

Acknowledgments

This study was supported by the Korean Cancer Foundation (K20170519); Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant numbers HI14C0072 and HI14C2399); National Research Foundation through contract N-16-NM-CR01-S01; and the Program of Construction and Operation for Large-scale Science Data Center (K-16-L01-C06-S01). This manuscript has been edited by native English-speaking experts from BioScience Writers LLC, Houston, TX, USA.

References

1. Parenti, G., Andria, G. & Ballabio, A. Lysosomal storage diseases: from pathophysiology to therapy. Annu. Rev. Med. 66, 471-486 (2015).
2. Platt, F.M. Sphingolipid lysosomal storage disorders. Nature 510, 68 (2014).

3. Wei, H., et al. 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. 17, 469-477 (2008).
4. Halliwell, B. Oxidative stress and cancer: have we moved forward? Biochem. J. 401, 1-11 (2007).
5. Shachar, T., et al. Lysosomal storage disorders and Parkinson's disease: Gaucher disease and beyond. Mov. Disord. 26, 1593-1604 (2011).
6. Arends, M., van Dussen, L., Biegstraaten, M. & Hollak, C.E. Malignancies and monoclonal gammopathy in Gaucher disease; a systematic review of the literature. Br. J. Haematol. 161, 832-842 (2013).
7. Cassiman, D., et al. Bilateral renal cell carcinoma development in long-term Fabry disease. J. Inherit. Metab. Dis. 30, 830-831 (2007).
8. Wang, R.Y., Bodamer, O.A., Watson, M.S. & Wilcox, W.R. Lysosomal storage diseases: Diagnostic confirmation and management of presymptomatic individuals. Genet. Med. 13, 457-484 (2011).
9. Wang, R.Y., Lelis, A., Mirocha, J. & Wilcox, W.R. Heterozygous Fabry women are not just carriers, but have a significant burden of disease and impaired quality of life. Genet. Med. 9, 34-45 (2007).
10. Campbell, P.J., Getz, G., Stuart, J.M., Korbel, J.O. & Stein, L.D. Pan-cancer analysis of whole genomes. bioRxiv (2017).
11. The Genomes Project, C. A global reference for human genetic variation. Nature 526, 68-74 (2015).
12. Lek, M., et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285-291 (2016).
13. Scriver, C.R. The metabolic and molecular bases of inherited disease, (McGraw-Hill, New York, 2001).
14. Boustany, R.-M.N. Lysosomal storage diseases—the horizon expands. Nature reviews Neurology 9, 583-598 (2013).

15. Futerman, A.H. & van Meer, G. The cell biology of lysosomal storage disorders. Nat. Rev. Mol. Cell Biol. 5, 554-565 (2004).

16. Lee, S., Wu, M.C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics 13, 762-775 (2012).

17. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550 (2014).

18. Luo, W., Friedman, M.S., Shedden, K., Hankenson, K.D. & Woolf, P.J. GAGE: generally applicable gene set enrichment for pathway analysis. BMC Bioinformatics 10, 161 (2009).

19. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27-30 (2000).

20. Storey, J.D. A Direct Approach to False Discovery Rates. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 64, 479-498 (2002).

21. Waddell, N., et al. Whole genomes redefine the mutational landscape of pancreatic cancer. Nature 518, 495-501 (2015).

22. Biankin, A.V., et al. Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. Nature 491, 399-405 (2012).

23. Jones, S., et al. Core Signaling Pathways in Human Pancreatic Cancers Revealed by Global Genomic Analyses. Science 321, 1801-1806 (2008).

24. de Fost, M., et al. Increased incidence of cancer in adult Gaucher disease in Western Europe. Blood Cells Mol. Dis. 36, 53-58 (2006).

25. Van Hove, J.L.K., et al. Late-Onset visceral presentation with cardiomyopathy and without neurological symptoms of adult Sanfilippo A syndrome. American Journal of Medical Genetics Part A 118A, 382-387 (2003).

26. Trudel, S., et al. Oxidative stress is independent of inflammation in the neurodegenerative sanfilippo syndrome type B. J. Neurosci. Res. 93, 424-432 (2015).

27. Rylova, S.N., et al. The CLN3 gene is a novel molecular target for cancer drug discovery. Cancer Res. 62, 801-808 (2002).

28. Siegel, R.L., Miller, K.D. & Jemal, A. Cancer statistics, 2016. CA Cancer J. Clin. 66, 7-30

(2016).

29. Vincent, A., Herman, J., Schulick, R., Hruban, R.H. & Goggins, M. Pancreatic cancer. The Lancet 378, 607-620 (2011).

30. Klein, A.P. Identifying people at a high risk of developing pancreatic cancer. Nature reviews. Cancer 13, 66-74 (2013).

31. Berger, A.H., Knudson, A.G. & Pandolfi, P.P. A continuum model for tumour suppression. Nature 476, 163-169 (2011).

32. Zhang, J., et al. Germline Mutations in Predisposition Genes in Pediatric Cancer. N. Engl. J. Med. 373, 2336-2346 (2015).

33. Pritchard, C.C., et al. Inherited DNA-Repair Gene Mutations in Men with Metastatic Prostate Cancer. N. Engl. J. Med. 375, 443-453 (2016).

34. Hollak, C.E.M. & Wijburg, F.A. Treatment of lysosomal storage disorders: successes and challenges. J. Inherit. Metab. Dis. 37, 587-598 (2014).

Figure Legends

Figure 1. Populations of the Pan-Cancer and 1000 Genomes cohorts.

Panels A and B show the populations of the Pan-Cancer and 1000 Genomes cohorts, respectively. Panel C shows the populations of the Pan-Cancer cohort according to the 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 Pan-Cancer 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 single, double, and triple PPV carriers (individuals carrying 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 the double and triple tier 3 PPV carriers and the overall triple PPV carriers are 7.54, infinite, and 7.4, respectively, with the corresponding bars cut off at the top edge of the plot. Panel B shows odds ratios for the prevalence of RSVs (mean allele frequency <0.5%) analyzed in the same manner for comparison. Error bars indicate 95% confidence intervals. Panel C shows the quantile-quantile plot of minus logarithmic P-values determined by SKAT-O analysis. Expected P-values under the null hypothesis (no association between 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 panel D. Panel D shows the SKAT-O associations between 30 major cancer types with more than 15 patients and PPVs in each LSD 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 FDR 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 the histological subtypes are defined in Table S1 in the Supplementary Appendix. RSV denotes rare synonymous variants, FDR false discovery rate, and SKAT-O optimal sequence kernel association test.

Figure 3. Age at diagnosis of cancer.

Panel A shows the age at diagnosis of cancer in seven cancer cohorts significantly associated with PPVs in the SKAT-O analysis (osteosarcoma was excluded because of lack of age at diagnosis information). Panel B shows the age at diagnosis of cancer according to the carrier status of 11 PPV groups significantly associated with the Pan-Cancer cohort or more than two histological cancer subgroups in the SKAT-O analysis. Panel C shows all cancer-gene pairs in which age at diagnosis of cancer differs significantly according to the PPV carrier status. We performed one-sided Wilcoxon rank sum tests, and the corresponding P-values are shown above each violin plot. 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 represents the median. PCAN denotes the Pan-Cancer, PACA pancreatic adenocarcinoma, BRCA breast cancer, CMBT medulloblastoma, SKCM cutaneous melanoma, PAEN pancreatic neuroendocrine carcinoma, and CMDI chronic myeloid disorder.

Figure 4. KEGG pathways significantly segregated by the potentially pathogenic variant carrier status in the pancreatic adenocarcinoma.

Only pathways of particular interest discussed in the manuscript are shown in this figure, and all pathways with FDR <0.1 are shown in Figure S14. ECM denotes extracellular matrix.

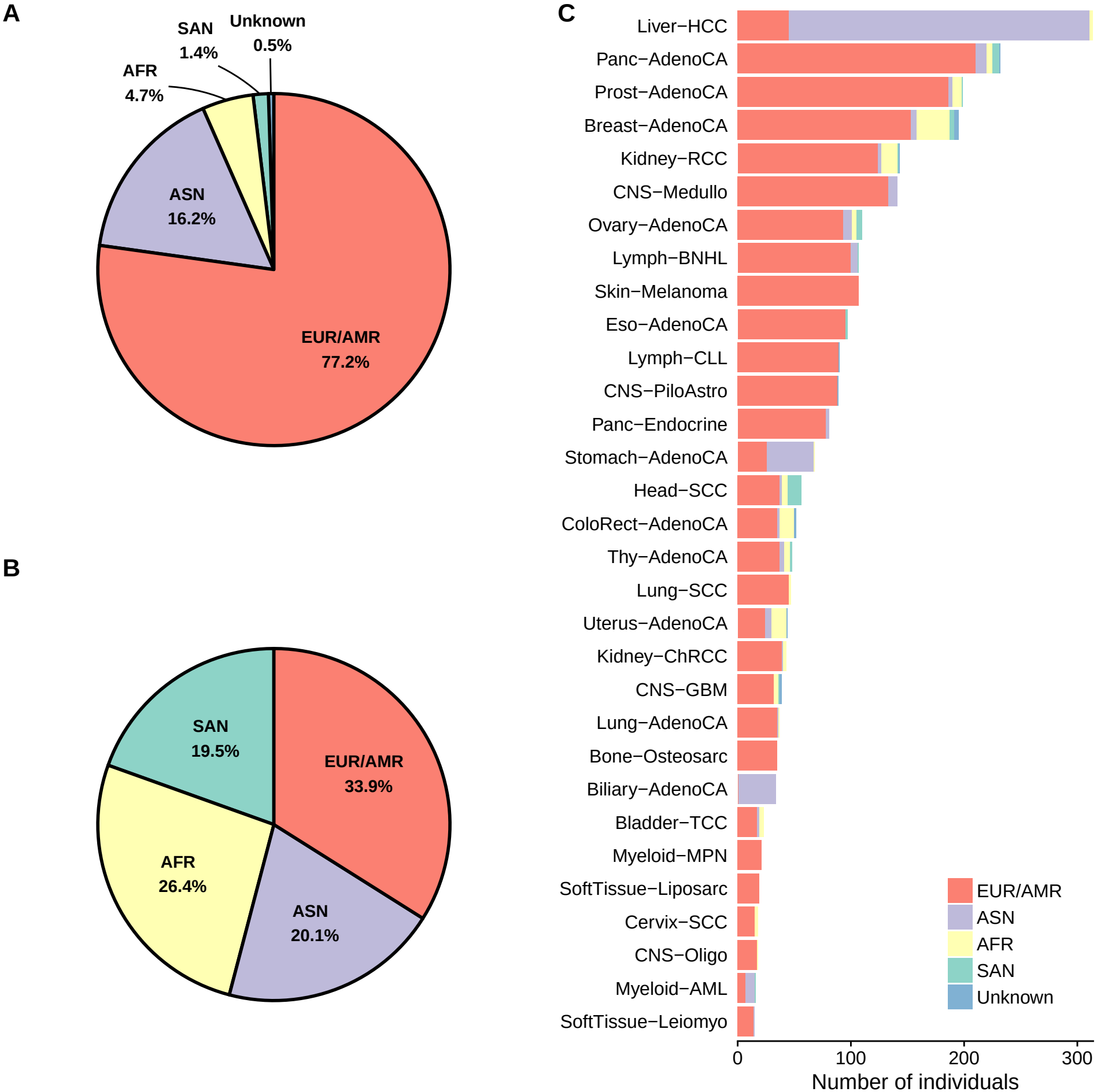
Table 1. Lysosomal storage disease genes included in this study.

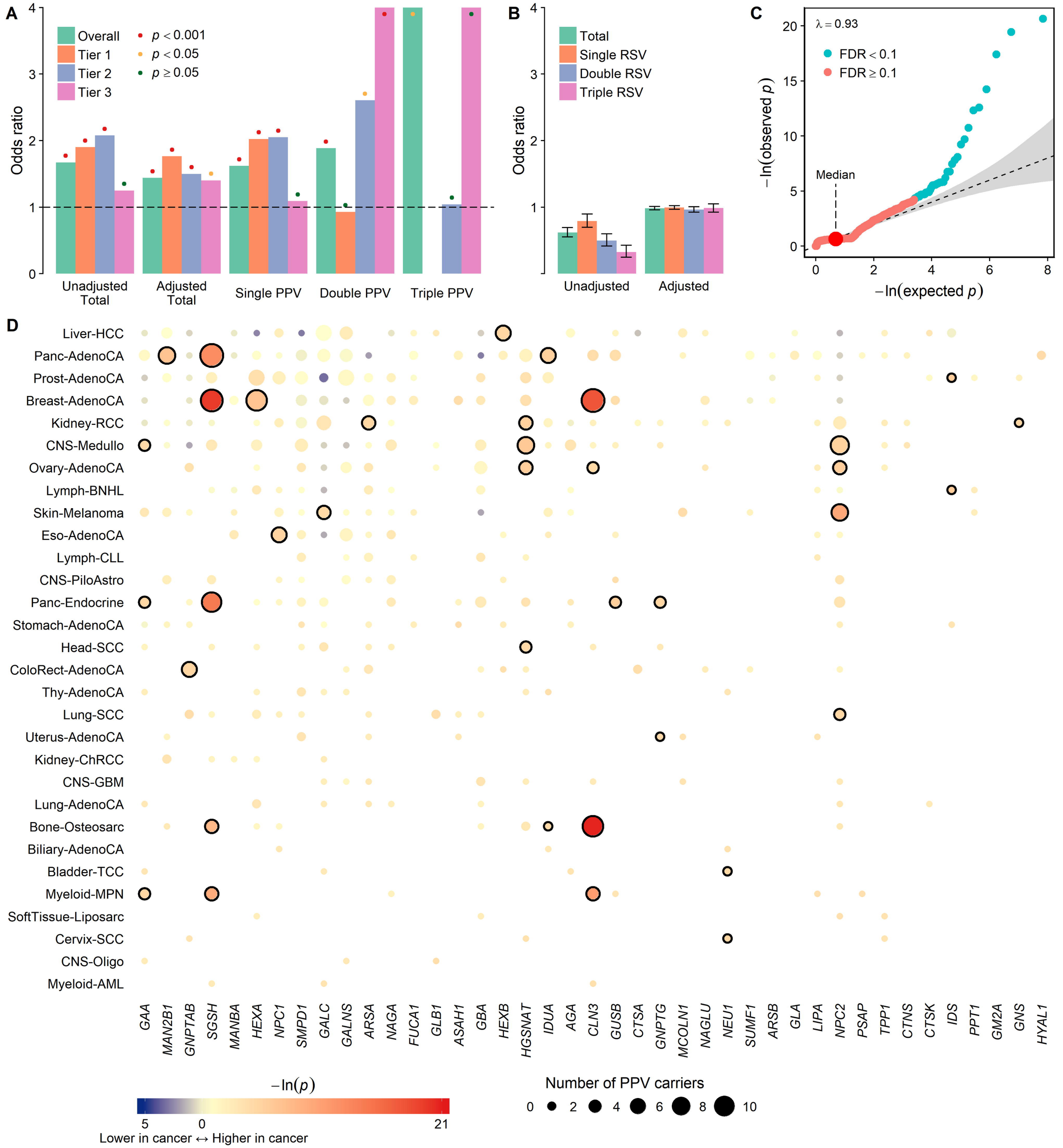
HGNC Symbol	Chromosome	Associated Lysosomal Storage Disease	Inheritance*
AGA	4	Aspartylglycosaminuria	Autosomal recessive
ARSA	22	Metachromatic leukodystrophy	Autosomal recessive
ARSB	5	Mucopolysaccharidosis VI (Maroteaux–Lamy syndrome)	Autosomal recessive
ASAH1	8	Farber lipogranulomatosis	Autosomal 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

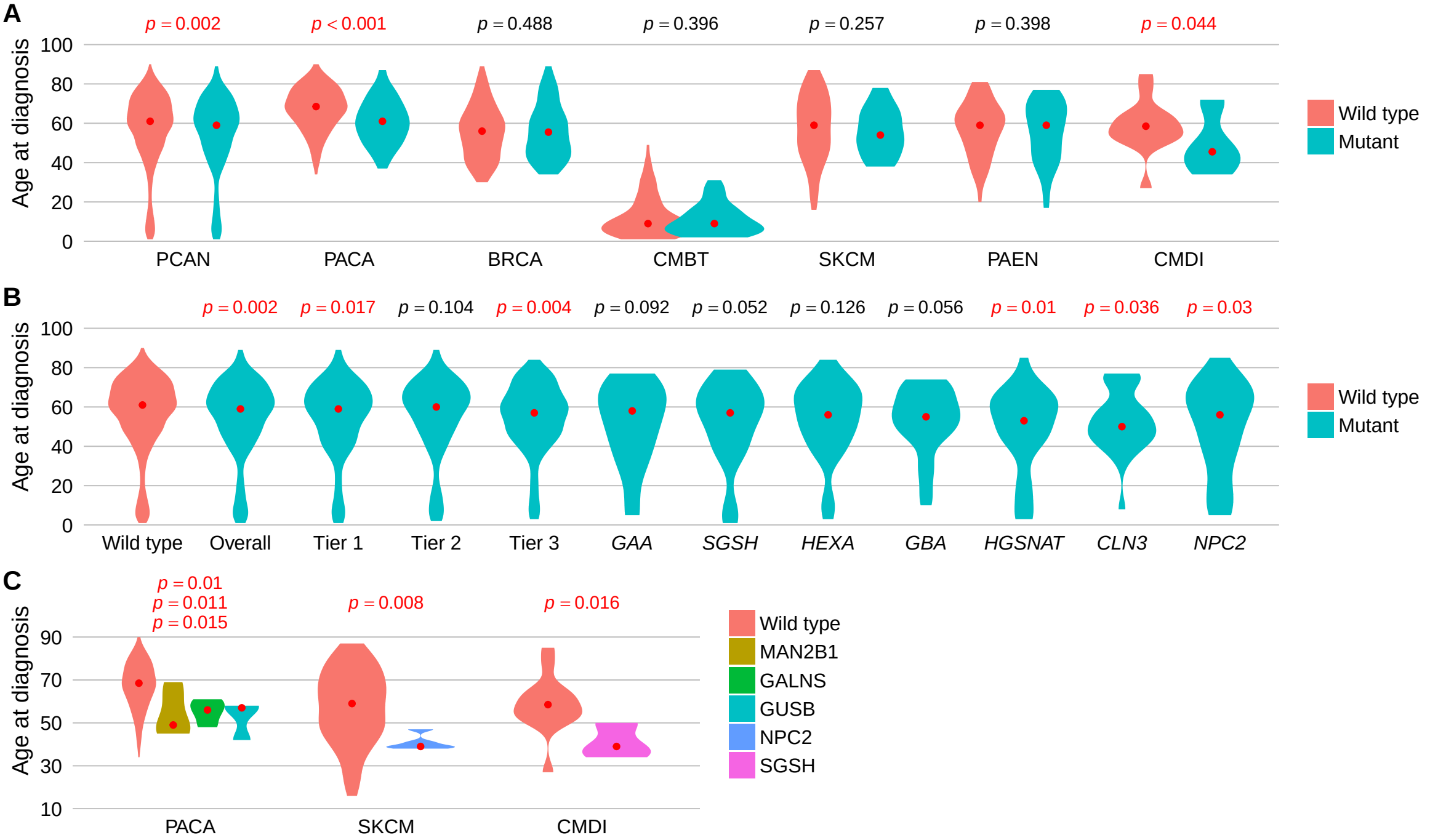
NPC2	14	Niemann–Pick disease type C2	Autosomal recessive
PPT1	1	Neuronal ceroid lipofuscinosis 1 (infantile NCL)	Autosomal recessive
PSAP	10	Gaucher disease Metachromatic leukodystrophy	Autosomal recessive
SGSH	17	Mucopolysaccharidosis IIIA (Sanfilippo syndrome A)	Autosomal recessive
SMPD1	11	Niemann–Pick 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

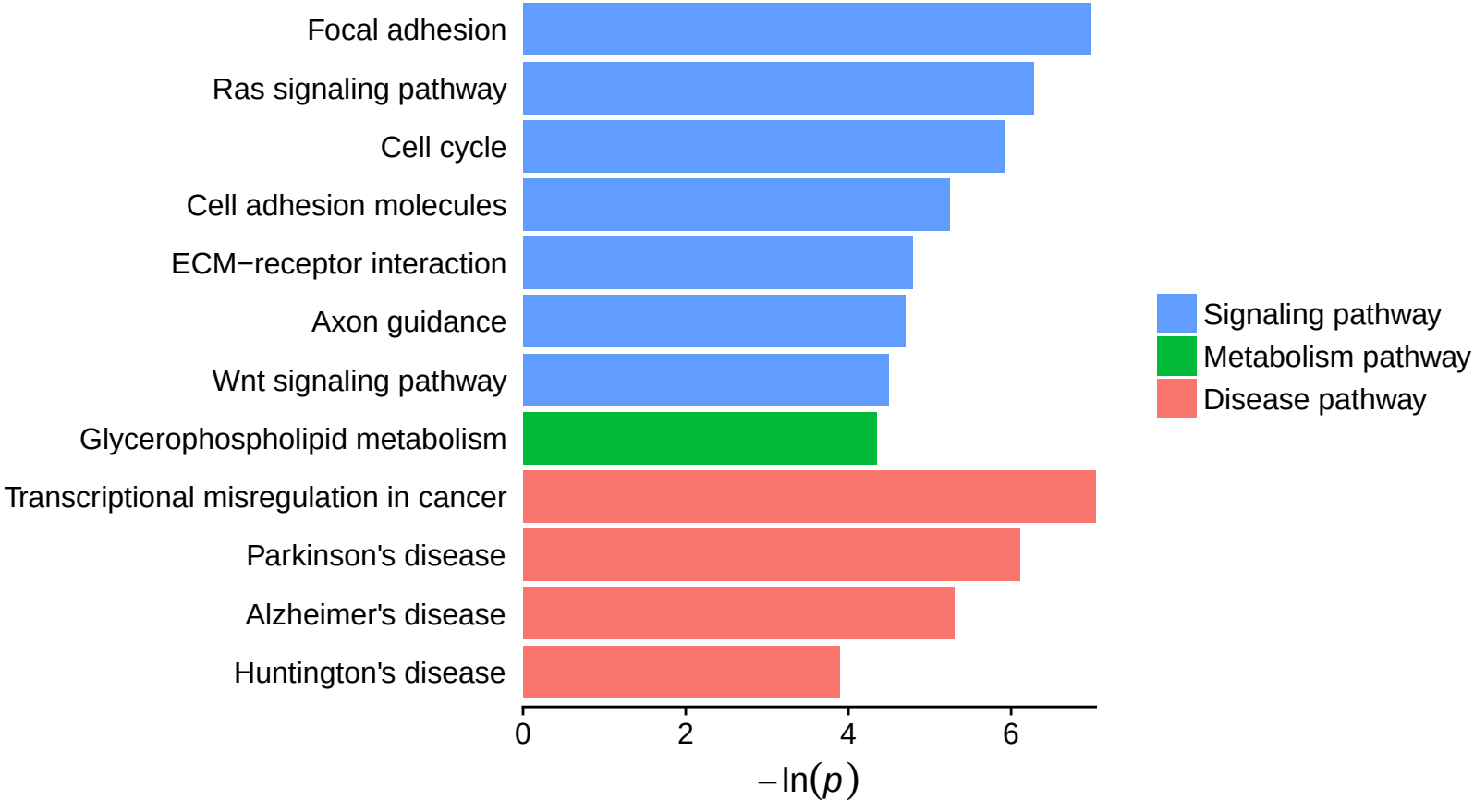
*Inheritance patterns based on the information provided in the Online Mendelian Inheritance in Man (OMIM) database (https://www.omim.org/).

HGNC denotes HUGO Gene Nomenclature Committee.


A
SAN
1.4%
Unknown
0.5%
AFR
4.7%
ASN
16.2%
EUR/AMR
77.2%
B
SAN
19.5%
EUR/AMR
33.9%
AFR
26.4%
ASN
20.1%
C
Liver−HCC
Panc−AdenoCA
Prost−AdenoCA
Breast−AdenoCA
Kidney−RCC
CNS−Medullo
Ovary−AdenoCA
Lymph−BNHL
Skin−Melanoma
Eso−AdenoCA
Lymph−CLL
CNS−PiloAstro
Panc−Endocrine
Stomach−AdenoCA
Head−SCC
ColoRect−AdenoCA
Thy−AdenoCA
Lung−SCC
Uterus−AdenoCA
Kidney−ChRCC
CNS−GBM
Lung−AdenoCA
Bone−Osteosarc
Biliary−AdenoCA
Bladder−TCC
Myeloid−MPN
SoftTissue−Liposarc
Cervix−SCC
CNS−Oligo
Myeloid−AML
SoftTissue−Leiomyo
EUR/AMR
ASN
AFR
SAN
Unknown
0
100
200
300
Number of individuals


A
Odds ratio
Overall
Tier 1
Tier 2
Tier 3
p < 0.001
p < 0.05
p ≥ 0.05
Unadjusted Total
Adjusted Total
Single PPV
Double PPV
Triple PPV
B
Total
Single RSV
Double RSV
Triple RSV
Unadjusted
Adjusted
C
λ = 0.93
FDR < 0.1
FDR ≥ 0.1
Median
−ln(observed p)
−ln(expected p)
D
Liver-HCC
Panc-AdenoCA
Prost-AdenoCA
Breast-AdenoCA
Kidney-RCC
CNS-Medullo
Ovary-AdenoCA
Lymph-BNHL
Skin-Melanoma
Eso-AdenoCA
Lymph-CLL
CNS-PiloAstro
Panc-Endocrine
Stomach-AdenoCA
Head-SCC
ColoRect-AdenoCA
Thy-AdenoCA
Lung-SCC
Uterus-AdenoCA
Kidney-ChRCC
CNS-GBM
Lung-AdenoCA
Bone-Osteosarc
Biliary-AdenoCA
Bladder-TCC
Myeloid-MPN
SoftTissue-Liposarc
Cervix-SCC
CNS-Oligo
Myeloid-AML
GAA MAN2B1 GNPTAB SGSH MANBA HEXA NPC1 SMPD1 GALC GALNS ARSA NAGA FUCA1 GLB1 ASAH1 GBA HEXB HGSNAT IDUA AGA CLN3 GUSB CTSA GNPTG MCOLN1 NAGLU NEU1 SUMF1 ARSB GLA LIPA NPC2 PSAP TPP1 CTNS CTSK IDS PPT1 GM2A GNS HYAL1
−ln(p)
5 0 21
Lower in cancer ↔ Higher in cancer
Number of PPV carriers
0 2 4 6 8 10


A
p = 0.002
p < 0.001
p = 0.488
p = 0.396
p = 0.257
p = 0.398
p = 0.044
Age at diagnosis
100
80
60
40
20
0
PCAN
PACA
BRCA
CMBT
SKCM
PAEN
CMDI
Wild type
Mutant
B
p = 0.002
p = 0.017
p = 0.104
p = 0.004
p = 0.092
p = 0.052
p = 0.126
p = 0.056
p = 0.01
p = 0.036
p = 0.03
Age at diagnosis
100
80
60
40
20
0
Wild type
Overall
Tier 1
Tier 2
Tier 3
GAA
SGSH
HEXA
GBA
HGSNAT
CLN3
NPC2
Wild type
Mutant
C
p = 0.01
p = 0.011
p = 0.015
p = 0.008
p = 0.016
Age at diagnosis
90
70
50
30
10
PACA
SKCM
CMDI
Wild type
MAN2B1
GALNS
GUSB
NPC2
SGSH


Focal adhesion
Ras signaling pathway
Cell cycle
Cell adhesion molecules
ECM−receptor interaction
Axon guidance
Wnt signaling pathway
Glycerophospholipid metabolism
Transcriptional misregulation in cancer
Parkinson's disease
Alzheimer's disease
Huntington's disease
0
2
4
6
−ln(p)
Signaling pathway
Metabolism pathway
Disease pathway