# WITHDRAWN:
# Machine Learning Algorithms and Whole Exome Sequencing Data from Breast Cancer Patients in the UK Biobank Predict Survival

**Bum-Sup Jang**
Seoul National University Bundang Hospital
**In Ah Kim**
inah228@snu.ac.kr

College of Medicine, Seoul National University https://orcid.org/0000-0001-9838-5399

**Research article**

**Keywords:** UK Biobank, Whole Exome Sequencing, Machine Learning, Breast cancer

**Posted Date:** December 2nd, 2020

**DOI:** https://doi.org/10.21203/rs.3.rs-115867/v1



**Version of Record:** A version of this preprint was published at Biomarkers in Medicine on November 1st, 2021. See the published version at https://doi.org/10.2217/bmm-2021-0280.

## EDITORIAL NOTE:

The full text of this preprint has been withdrawn by the authors while they make corrections to the work. Therefore, the authors do not wish this work to be cited as a reference. Questions should be directed to the corresponding author.

# Abstract

**Background:** Using by machine learning algorithms, we aimed to identify the mutated gene set from the whole exome sequencing (WES) data of blood in the cancer, which is associated with overall survival in breast cancer patients.

**Methods:** WES data from 1,181 female breast cancer patients within the UK Biobank cohort was collected. The number of mutations for each gene was summed and defined as the blood-based mutation burden per patient. Using by Long short-term memory (LSTM) machine learning algorithm and a XGBoost—a gradient-boosted tree algorithm, we developed the model to predict patient overall survival.

**Results:** From the UK biobank-breast cancer cohort, most altered genes in blood samples were related with the TP53 pathway. In the LSTM model, the minimum 50 genes were found to predict high vs. low mutation burden. In the XGBoost survival model, the gene-set could predict overall survival showing the concordance index of 0.75 and the scaled Brier-score of 0.146 from the held-out testing set (20%, N=236). In older patients (≥ 56 years), the high mutation group based on this gene-set showed inferior overall survival compared to the low mutation group (log-rank test, P=0.042)

**Conclusion:** The machine learning algorithms revealed the gene-signature in the UK biobank breast cancer cohort. Mutational burden observed in blood was associated with overall survival in relatively old patients. This gene-signature should be verified in prospective setting.


# Background

Compared to tissue-based biopsies, blood-based assays are less invasive and more convenient for cancer patients. Due to advancements in sequencing, cell-free DNA (cfDNA) and circulating-tumor DNA (ctDNA) can be extracted from blood and sequenced for prognostic and predictive biomarkers, respectively. The US Food and Drug Administration recently approved the first blood test for the epidermal growth factor receptor (EGFR) gene mutation in cfDNA in patients with non-small cell lung cancer (NSCLC). It has been shown that blood-based tumor mutation burdens derived from cfDNA [1] and ctDNA [2] are promising predictive biomarkers for objective response rates of immunotherapy in patients with advanced-stage NSCLC. However, these studies used deep sequencing technologies with median exome coverage > 800 X and 3417 X, respectively. It has been suggested that shallow whole genome sequencing (sWGS) with low coverage (~ 0.5 X) can be used to profile cfDNA and ctDNA. Several studies of various cancers [3–7] demonstrated the feasibility of sWGS to identify biomarkers. Based on these studies, we hypothesized that whole exome sequencing (WES) of blood from cancer patients could reveal overall survival prognosis.

Currently, the UK Biobank database collects blood samples from volunteer participants. WES data of 50,000 blood samples in this large-scale national database was recently published [8]. This population was a prospective cohort, and some participants developed various diseases, including cancer. Of the 50,000 participants, ~ 5,700 patients (11.1%) were diagnosed with cancer. Baseline epidemiological

factors, results of biological samples, and online follow-up data are available. Because national death and cancer registry data are linked to participants, researchers can identify the date of diagnosis and the date of last follow-up or death. Of the various types of cancer, we focused on female breast cancer patients. Due to the sparsity of the dataset, a machine learning approach was applied.

Recently, machine learning algorithms have been used in medical research to resolve clinical challenges. The deep learning method is used when handling sparse and large high-dimensional datasets, such as genomic data. For example, the long short-term memory (LSTM) network was used to identify binding sequences and structural motifs of RNA-binding proteins in RNA sequencing data [9]. Because the LSTM network can be trained with time series data, mutation time series data were used to predict tumor progression in colon and lung cancer [10]. Extreme gradient boosting (XGBoost) [11] is another popular classifier due to its high performance and model interpretability. We hypothesized that the XGBoost machine learning algorithm would extract important mutational features associated with overall survival.

In this study, the primary endpoint was to identify the optimal gene set to predict the blood-based mutation burden. The secondary endpoint was to determine the association between the gene signature and overall survival.

## Methods

# UK Biobank Study Approval and Study Population

The UK Biobank approved our study proposal and we signed a Material Transfer Agreement contract. The UK Biobank obtained approval from the North West Multi-centre Research Ethics Committee and the Community Health Index Advisory Group. All participants provided written informed consent to the UK Biobank. Details on data collection and protocols are described on the UK Biobank home page (https://www.ukbiobank.ac.uk). We also obtained approval waivers from our institutional review board for this study.

Primary female breast cancer patients without any other malignancies were included in this study. From a total population of 502,507, we identified 1,426 breast cancer patients based on the ICD-10 code. We excluded patients who had other malignancies (N = 172) and were male breast cancer patients (N = 3). We also excluded patients whose follow-up period was < 30 days (N = 70). Finally, we included 1,181 female breast cancer patients within the UK Biobank cohort in this study.

# Assessment Of The Mutation Burden In Blood

For the UK Biobank exome sequencing and variant calling, 39 Mbp of the human genome (19,396 genes) were targeted and captured by the IDT xGen Exome Research Panel v1.0. The depth exceeded 20X at 94.6% of target sites, on average. Samples were sequenced with dual-indexed 75 × 75 bp paired-end reads on the Illumina NovaSeq 6000 platform. All raw sequencing data was converted to FASTQ files according

to Illumina NovaSeq best practices and was aligned to the GRCh38 reference. After read-duplicate marking, single nucleotide variants (SNVs) and indels were called with WeCall (GenomicsPLC) to generate a genomic variant call format (gVCF) per sample. Then, additional "Functionally Equivalent" (FE) VCF files were generated according to the published protocol [12]. We downloaded processed gVCF files after study approval.

ANNOVAR (version 2019Oct24) software was used to annotate mutations according to the Catalogue Of Somatic Mutations In Cancer (COSMIC) release v88 (https://cancer.sanger.ac.uk/cosmic). Annotated mutation files were imported as MAF objects using the maftools package [13] in R software version 3.6.1. To investigate mutations related to cancer in blood, gene mutations not registered in the COSMIC database v88 were excluded. Then, the number of mutations for each gene was summed and defined as the blood-based mutation burden per patient.

# Machine Learning Algorithm

We first used the preprocessing step and the LSTM algorithm according to Auslander et al. [10] which demonstrated feasibility of the machine learning algorithm to analyze the dynamics of a mutational time series for tumor progression. The LSTM algorithm was used to train time series somatic mutation events and to predict tumor mutation burden. Before training the LSTM algorithm, the order score was calculated to sort the mutational data. The order score is a score for every mutation as a ratio of its occurrence in the presence and absence of other mutations; detailed methods and rationale are described in the study by Auslander et al. [3]. Using publicly available code (https://github.com/noamaus/LSTM-Mutational-series), we calculated the order score for every mutation and sorted gene mutations based on the assigned score. The LSTM structure suggested by Auslander et al. [10] was used in this study to identify a subset of genes to predict the blood-based mutation burden. Model performance was evaluated by 5-fold validation. The average area under the receiver operating characteristic curve (AUC) was calculated for each fold.

To investigate the predictive value of a gene set in terms of survival, we trained XGBoost [11], a machine learning gradient-boosted tree algorithm. For censored survival outcomes, we used the modified version of XGBoost, the xgboost.surv R program package. The study population was divided into a testing set (80%) and a training set (20%). The survival model was trained with several combinations of parameters, including maximum depth of a tree, step size shrinkage, minimum loss reduction, and minimum sum of instance weight (Hessian) needed in a child. Hyperparameters were optimized by a grid search. The trained survival model was validated in the held-out testing set using scaled Brier scores. A perfect model demonstrates a scaled Brier score of 1. Negative value scores indicate that the model is not useful. To interpret the model output, the SHAP (SHapley Additive exPlanation) value was estimated and visualized using the SHAPforxgboost R program package. Then, we sorted the genes based on the importance of each gene in terms of overall survival. The concordance index (C-index) was calculated to evaluate how well the model predicted patient survival. A perfect model had a C-index of 1.

# Gene Set Enrichment And Network Analysis

To explore the biological functions of the gene set selected by the machine learning algorithm, we queried the selected gene set in the Genotype-Tissue Expression (GTEx) expression portal (https://www.gtexportal.org/home/multiGeneQueryPage). A heat map showed the gene expression distribution pattern across the human organs. To perform human interaction network analysis of the selected gene set, we used the netboxr [14] R program package. This tool maps mutated gene sets onto the Human Interaction Network, which was derived from four curated data sources: The Human Protein Reference Database, Reactome, NCI-Nature Pathway Interaction Database, and The MSKCC Cancer Cell Map. After generating the network, we performed enrichment analysis for genes within the same module using the Gene Ontology database.

# Statistical Analysis

Overall survival time was estimated from the date of diagnosis to the date of the last follow-up or death. A Kaplan-Meier curve was plotted and the log-rank test was used to compare overall survival between groups. Statistical analyses were performed using STATA version 16.0. In the enrichment analysis, the adjusted P-value was estimated using the Benjamini-Hochberg method. The false discovery rate-adjusted p-value (q-value) was also calculated to filter for significant enrichment pathways.

## Results

# Gene Set Predicting Blood-based Mutation Burden

All mutations found in the COSMIC database were investigated. The median blood-based mutation burden was 73 (47–103) in this study population. Based on the median value, patients were separated into high mutation (N = 640, 54.2%) and low mutation (N = 541, 42.8%) groups. Each blood-based mutation was mapped to known oncogenic signaling pathways derived from TCGA cohorts. In this study, 20.4% of patients showed an altered RTK-RAS pathway in their blood sample and most genes in the TP53 pathway were affected (5/6 genes, 83.3%, Fig. 1a). The blood-based mutation burden correlated positively, but weakly, with age (Spearman's rho = 0.063, P = 0.041, Fig. 1b). Of the mutated genes, we sought the optimal and minimal gene sets that can predict the total blood-based mutation burden. As suggested by Auslander et al. [3], we calculated the order score for each gene and sorted the genes in ascending order. As a time series sequence, sorted genes were used to train the LSTM model to classify high vs. low mutation burden groups. The mean AUC values were plotted against the number of genes involved along with the 5-fold internal validation (Fig. 1c). We found that mean AUC values increased as the number of genes trained increased. The mean AUC value was saturated at 0.948 with 50 genes, thus we used a 50 gene mutational profile in following analysis.

# Gene Set Enrichment Analysis And Network Analysis

For the 50 selected genes, we generated a heat map showing their expression profiles in human organs (Fig. 2). Most genes were expressed across the human organs, and strong expression of *ATN1* gene was observed in most organs. The NetBoxr algorithm [15] was used to identify pathway modules in the pre-defined Human Interactions Network for the altered gene set. Each network module was a cluster of connected nodes/circles (genes) and lines (interactions). In this study, five network modules were revealed (Fig. 3a). Module 1 contained genes *ZYX*, *VIM*, *YWHAE*, and *ZNF384.* Enrichment analysis revealed that Module 1 is primarily involved in the response to interferon gamma (adjusted P-value = 0.056, Q-value = 0.021, Fig. 3b). Module 2 was composed of the *MAPK14*, *MAFA*, *TRAF2*, and *MAP3K1* genes, which associate with MAP kinase activity (adjusted P-value = 0.002, Q-value < 0.001, Fig. 3c). Module 3 was the largest module containing genes *ATN1*, *TRIP6*, *BAT2*, *EFEMP2*, *ATXN1*, *RAD54L2*, *MBP*, *RBM9*, *DMPK*, *GAPDH*, *RBPMS*, *MYST3*, *USP54*, *KIAA0913*, and *RNF31* (N = 15), however enrichment analysis showed no association with any significant biological process. The results of the enrichment analyses for Modules 4 and 5 can be found in Additional file 1: Supplementary Table S1.pdf

# Development Of The Survival Model

We used a modified version of the XGBoost algorithm that can be trained for survival data. Eighty percent of the samples (N = 945) were used to train the XGBoost model, and 20% of the samples (N = 236) were allocated to the held-out testing set. The final model had a scaled Brier score of 0.146 with optimal hyperparameters and a C-index of 75%. To reveal the importance of each gene, the distribution and mean SHAP values were plotted (Fig. 4a); a high SHAP value indicates high importance. Genes with mean SHAP values > 0.1 include *SYT15*, *MUC12*, *CTAGE4*, *NOP9*, *PRAMEF10*, *FRG2C*, *ANKRD20A4*, *LNP1*, *AGAP5*, and *ANKRD35* (N = 10). Mutations in these 10 genes were counted for each patient and divided by the median mutation value to classify patients into the high or low mutation group, based on these genes. The median age of the patient was 56 years (range 28–76). Among older patients ≥ 56 years, overall survival was lower in the high mutation group than in the low mutation group (log-rank test, P = 0.042, Fig. 4b). Among younger patients < 56 years, there was no difference in overall survival between the high and low mutation groups (log-rank test, P = 0.181).

## Discussion

In the UK Biobank breast cancer cohort, most of the altered genes in the blood samples were involved in the TP53 pathway. Using the LSTM machine learning algorithm, we identified a gene set (N = 50) that can predict the total blood-based mutation burden. The number of mutations in the gene set had a very weak positive correlation with age. Expression of the gene set was enriched in biological processes involving MAP kinase activity and the response to interferon gamma. Using the XGBoost machine learning algorithm, we determined the top 10 most important genes which mutations were significantly associated with inferior overall survival in patients ≥ 56 years of age.

Blood samples acquired from cancer patients can conveniently provide considerable information. With deep sequencing-based approaches, ctDNA or cfDNA can be used to detect early cancer, monitor the tumor mutation burden, and evaluate the treatment response [16]. Liu et al. [17] showed that the methylation pattern of cfDNA could differentiate multiple cancer types at all stages. In patients with advanced NSCLC, a prospective study [18] demonstrated that driver mutations in ctDNA could guide targeted therapy. Meanwhile, additional cost-effective strategies for sequencing coverage (0.1X-0.5X) are emerging. In patients with high-stage neuroblastoma, sWGS of cfDNA showed that the copy number alteration profiles in cfDNA and primary tumors were similar [4]. Copy number profiles determined from sWGS could detect histologic differences between non-small and small lung cancers [5]. In metastatic breast cancer patients, a high level of ctDNA, which was determined by sWGS, correlated with a worse prognosis [7].

Thus, we hypothesized that a relatively low coverage (20X) of exome sequencing of the UK Biobank data could reveal somatic mutations in tumors, even in a blood sample. Using the COSMIC database, we selected somatic mutations found in cancer from blood samples. Non-disease tissue, such as blood, is commonly sequenced to filter germline variants, even though the strict definition of germline variants is genomic variations found in germ cells. Rubinstein el al. [19] determined that genomic variations in a blood sample included a large fraction of postzygomatic *de novo* genomic variations rather than germline variations. Indeed, the current challenge is not knowing whether ctDNA variations originated from the germline or somatic mutations [16]. Furthermore, a somatic mutation landscape study [20] also showed that cancer mutations were enriched in non-disease tissue.

To analyze genomic data robustly, we used machine learning algorithms in this study. By using the LSTM algorithm suggested by Auslander et al. [10], we identified 50 genes that could predict the overall somatic mutation count in blood. The LSTM algorithm is intended for sequence classification or sequence prediction. Gene set enrichment analyses revealed that expression of the gene set was related to the interferon gamma response and MAP kinase activity. Consistent with this, García-Nieto et al. [20] demonstrated that genes with a high somatic mutation load were enriched in the immune response. Using the XGBoost algorithm, we identified a more refined gene signature (N = 10), for which a high mutation burden was associated with a lower overall survival in patients > 56 years of age. The XGBoost algorithm is a scalable end-to-end tree-boosting system that can be used for genome variant detection [21]. In high-dimensional datasets, such as the UK Biobank dataset, the XGBoost algorithm was the best at predicting cardiovascular disease risk [22]. A strength of the algorithm is that the importance of each variable in the prediction model is calculated, thus allowing for the identification of a gene signature for survival.

Regarding immunotherapy, a high tumor mutation burden, which was measured using targeted next-generation sequencing (NGS) panels, was associated with better survival in most cancers [23]. This trend was not observed in breast cancer, perhaps due to the relatively low tumor mutation burden [24]. In this study, we found that the total mutation burden of the gene signature in blood samples was associated with age-dependent survival. Age contributes to the increase of somatic mutations. WGS analysis of

hematopoietic progenitors [25] revealed that 14 somatic mutations accumulate per cell per year. García-Nieto et al. [20] also demonstrated that the mutation load in blood samples was associated with age. In this study, we found that the blood-based mutation burden showed a very weak positive correlation with age. Nevertheless, a significant survival difference between a high or low mutation load in relatively older patients was observed. Because mutations found in the COSMIC database were considered in this study, high mutational loads may reflect the cancerous state of older patients.

There are several limitations in this current study. The gene signature should be validated in a prospective study with a long-term follow-up period. Because there was no sequencing data derived from the primary tumor, we could not filter strictly for germline mutations. Although detailed descriptions of baseline patient characteristics should have been provided, no clinicopathologic information, including molecular subtype, treatment regimen/sequence, and progression-free survival, was available. Without information on copy number aberrations, it is unknown whether the functions of the mutated genes were amplified or lost.

# Conclusions

We identified a gene signature among the UK Biobank breast cancer cohort using machine learning algorithms. The mutational burden in blood samples was associated with overall survival in relatively old patients. This gene signature should be verified in a prospective setting.

# Abbreviations

WES, Whole exome sequencing; LSTM, Long short-term memory; AUC, area under a receiver operating characteristic curve; cfDNA, cell-free DNA; ctDNA, circulating-tumor DNA; sWGS, shallow whole genome sequencing; NSCLC, non-small cell lung cancer; TCGA, The cancer genome atlas; SHAP, SHapley Additive exPlanation.

# Declarations

### Ethics approval and consent to participate:

This study was waived by the institutional review board of our institution because dataset was achieved from UK Biobank. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Because that the UK biobank obtained the informed consent from participants, requirement for obtaining informed consent of participants included in the study was exempted

### Consent for publication:

Not applicable.

**Availability of data and material:**

Whole exome-seq data and clinical data of cohort are available in the UK Biobank under approval. Processing code and machine learning model will be uploaded in public repository after publication.

**Competing interests:**

The authors declare that they have no competing interests.

**Funding:**

This work was supported by the grands from Korean Ministry of Science and Information & Communication Technology National Research Foundation (#2020R1A2C2005141), Seoul National University Big Data Institute via The Data Science Research Project 2019, and the SNUBH Research Fund (#14-2019-031) to In Ah Kim.

**Authors' contributions:**

IAK designed and supervised the study. BSJ contributed the sample collection and preprocessing. Statistical analysis was performed by BSJ. IAK and BSJ equally drafted the manuscript. All authors have read and approved the manuscript as submitted.

**Acknowledgements:**

None.

# References

1. Gandara DR, Paul SM, Kowanetz M, Schleifman E, Zou W, Li Y, et al. Blood-based tumor mutational burden as a predictor of clinical benefit in non-small-cell lung cancer patients treated with atezolizumab. Nat Med. 2018;24. doi:10.1038/s41591-018-0134-3.
2. Wang Z, Duan J, Cai S, Han M, Dong H, Zhao J, et al. Assessment of Blood Tumor Mutational Burden as a Potential Biomarker for Immunotherapy in Patients with Non-Small Cell Lung Cancer with Use of a Next-Generation Sequencing Cancer Gene Panel. JAMA Oncol. 2019;5:696–702.
3. Mouliere F, Mair R, Chandrananda D, Marass F, Smith CG, Su J, et al. Detection of cell-free DNA fragmentation and copy number alterations in cerebrospinal fluid from glioma patients. EMBO Mol Med. 2018;10:1–6.
4. Van Roy N, Van Der Linden M, Menten B, Dheedene A, Vandeputte C, Van Dorpe J, et al. Shallow whole genome sequencing on circulating cell-free DNA allows reliable noninvasive copy-number profiling in neuroblastoma patients. Clin Cancer Res. 2017;23:6305–15.
5. Raman L, Van Der Linden M, Van Der Eecken K, Vermaelen K, Demedts I, Surmont V, et al. Shallow whole-genome sequencing of plasma cell-free DNA accurately differentiates small from non-small

cell lung carcinoma. Genome Med. 2020;12:1–12.

6. Mohan S, Foy V, Ayub M, Leong HS, Schofield P, Sahoo S, et al. Profiling of Circulating Free DNA Using Targeted and Genome-wide Sequencing in Patients with SCLC. J Thorac Oncol. 2020;15:216–30.
7. Bourrier C, Pierga J, Xuereb L, Lockhart BP. Guigal-stephan N. Shallow Whole-Genome Sequencing from Plasma Identifies FGFR1 Amplified Breast Cancers and Predicts Overall Survival. 2020.
8. Van Hout CV, Tachmazidou I, Backman JD, Hoffman JX, Ye B, Pandey AK, et al. Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank. 2019. doi:10.1101/572347.
9. Pan X, Rijnbeek P, Yan J, Shen H, Bin. Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. BMC Genom. 2018;19:1–11.
10. Auslander N, Wolf YI, Koonin EV. In silico learning of tumor evolution through mutational time series. Proc Natl Acad Sci. 2019;116:201901695. doi:10.1073/pnas.1901695116.
11. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.
12. Regier AA, Farjoun Y, Larson DE, Krasheninina O, Kang HM, Howrigan DP, et al. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. Nat Commun. 2018;9:1–8. doi:10.1038/s41467-018-06159-4.
13. Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. Maftools: Efficient and comprehensive analysis of somatic variants in cancer. Genome Res. 2018.
14. Liu EM, Luna A, Dong G, Sander C. NetBoxR. Automated Discovery of Biological Process Modules by Network Analysis in R. 2020;:1–11.
15. Cerami E, Demir E, Schultz N, Taylor BS, Sander C. Automated network analysis identifies core pathways in glioblastoma. PLoS One. 2010;5.
16. Yi X, Ma J, Guan Y, Chen R, Yang L, Xia X. The feasibility of using mutation detection in ctDNA to assess tumor dynamics. Int J Cancer. 2017;140:2642–7.
17. Liu MC, Oxnard GR, Klein EA, Swanton C, Seiden MV, Liu MC, et al. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. Ann Oncol. 2020;31:745–59.
18. Sabari JK, Offin M, Stephens D, Ni A, Lee A, Pavlakis N, et al. A Prospective Study of Circulating Tumor DNA to Guide Matched Targeted Therapy in Lung Cancers. J Natl Cancer Inst. 2019;111:575–83.
19. Rubinstein JC, Nicolson NG, Rottmann D, Morotti R, Korah R, Carling T, et al. Choice of control tissue impacts designation of germline variants in a cohort of papillary thyroid carcinoma patients. Ann Oncol. 2020.
20. García-Nieto P, Morrison A, Fraser H. The somatic mutation landscape of the human body. Somat Mutat Landsc Hum body. 2019;:668624.

21. Zhang Z, Yin L, Hao L, Cao L, Liu C, Chen M. GVC: An ultra-fast and all-round genome variant caller. 2017.
22. Alaa AM, Bolton T, Angelantonio E, Di, Rudd JHF, van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. PLoS One. 2019;14:1–17.
23. Samstein RM, Lee C-H, Shoushtari AN, Hellmann MD, Shen R, Janjigian YY, et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. Nat Genet. 2019;51:202–6. doi:10.1038/s41588-018-0312-8.
24. Vonderheide RH, Domchek SM, Clark AS. Immunotherapy for breast cancer: What are we missing? Clin Cancer Res. 2017;23:2640–6.
25. Osorio FG, Rosendahl Huber A, Oka R, Verheul M, Patel SH, Hasaart K, et al. Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. Cell Rep. 2018;25:2308–16.e4. doi:10.1016/j.celrep.2018.11.014.

# Figures

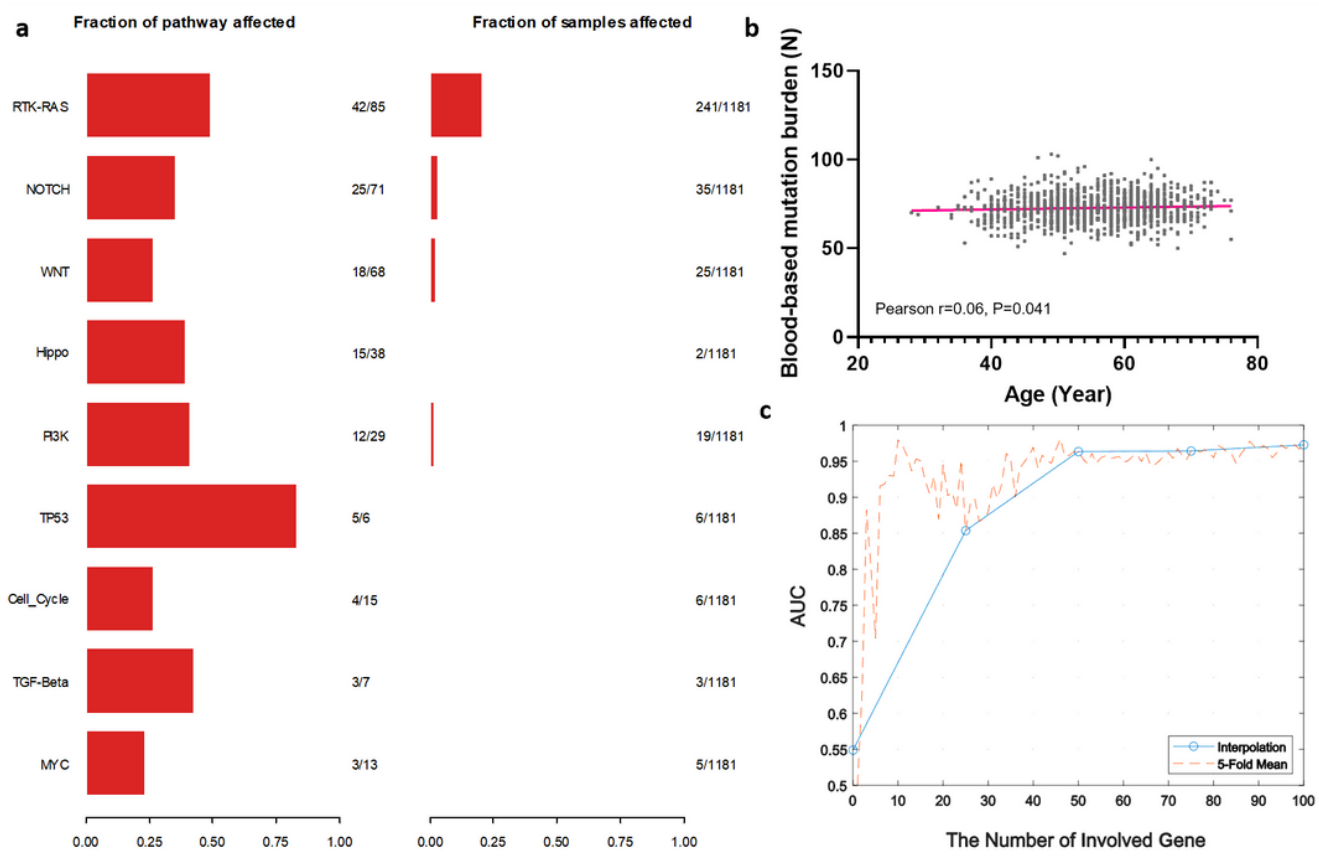
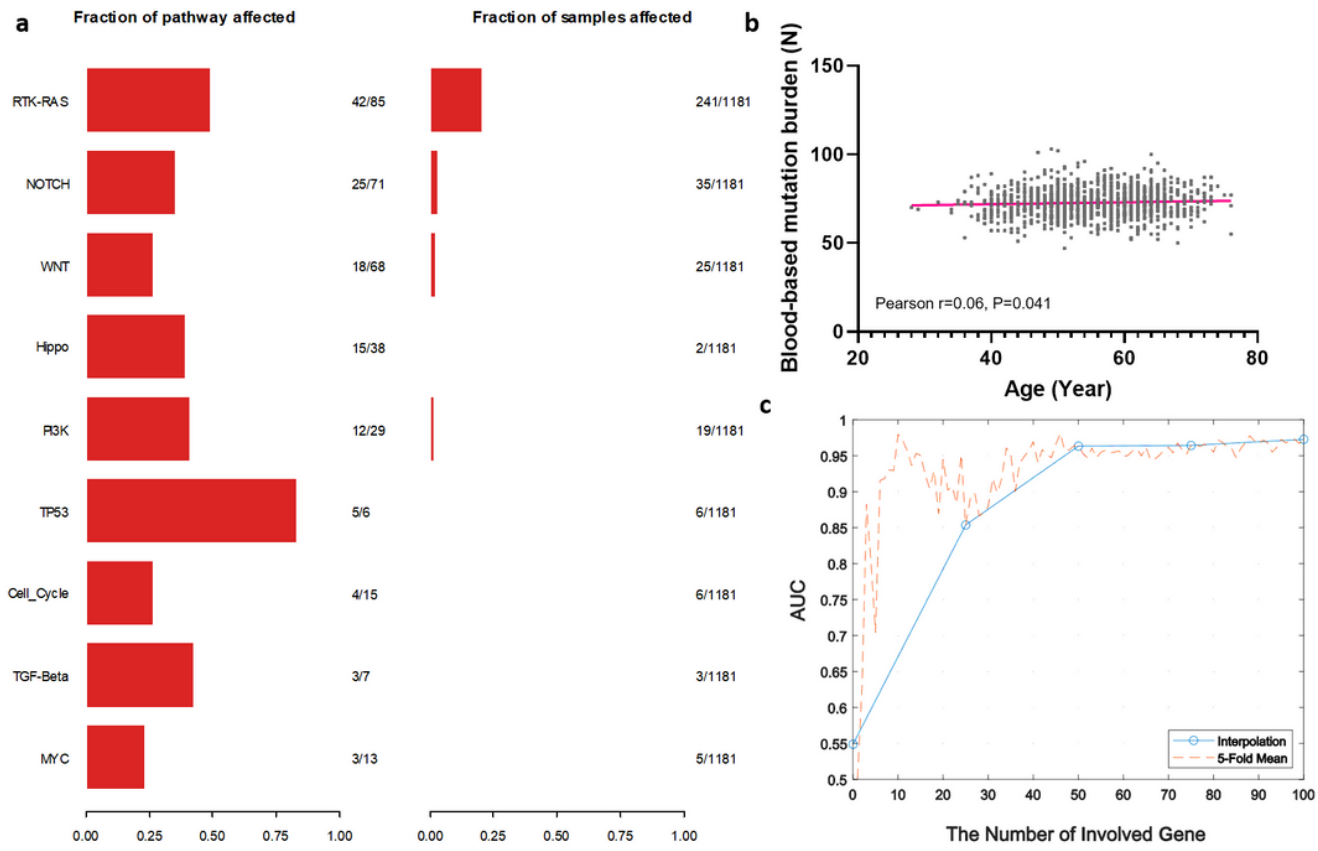



**Figure 1**

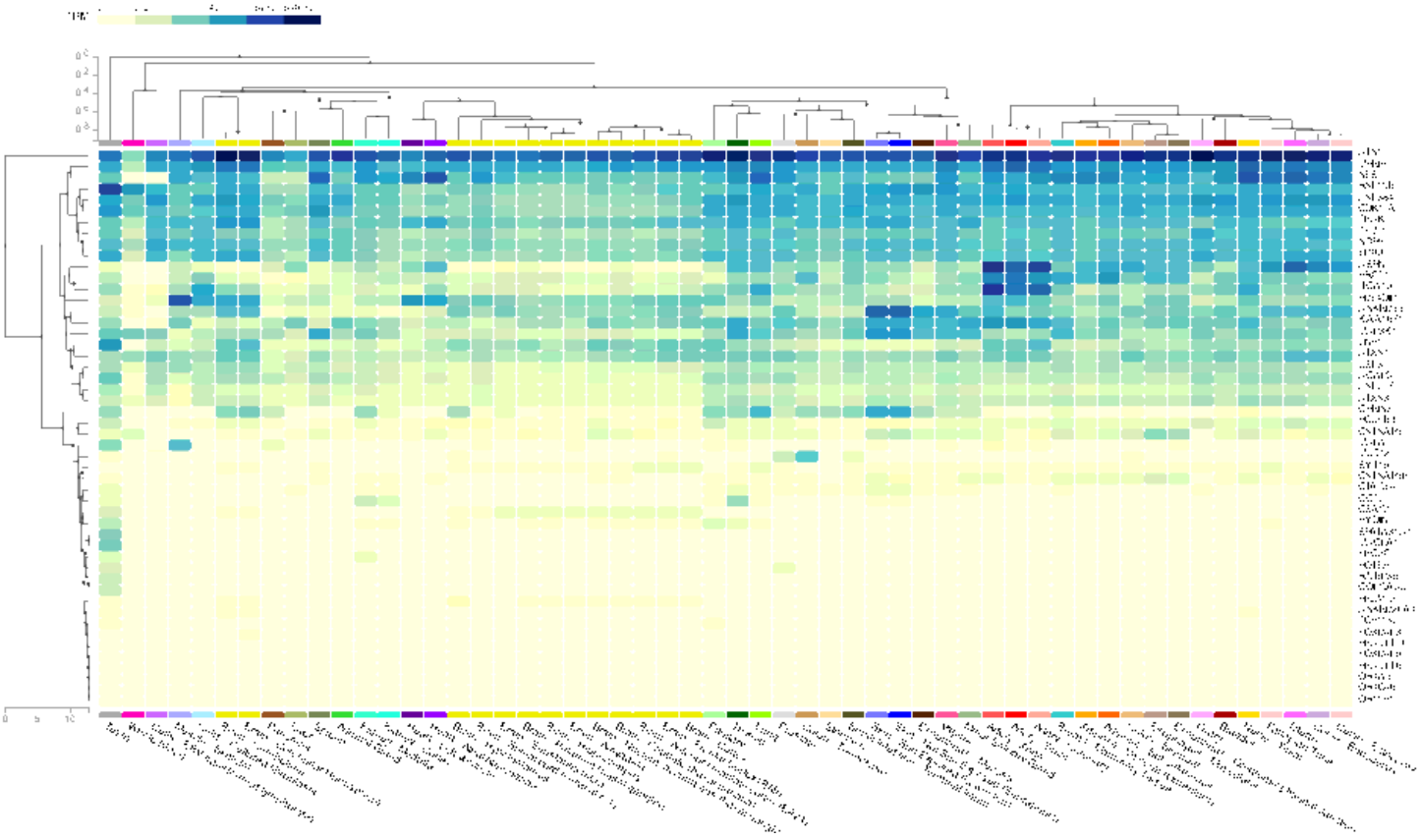
(A) The proportion of the indicated pathways that had mutations and the proportion of blood samples carrying mutations in the indicated pathways. For example, 5/6 of the genes in the TP53 pathway had

mutations and 241/1181 (20.4%) of blood samples had mutations in the RTK-RAS pathway. (B) Correlation plot and interpolation (magenta line) of the blood-based mutation burden and age. The Pearson coefficient and the P-value are shown. (C) Mean AUC value of the 5-fold validation (orange) and the interpolation (blue line) according to the number of trained genes in the LSTM model. AUC, area under a receiver operating characteristic curve.



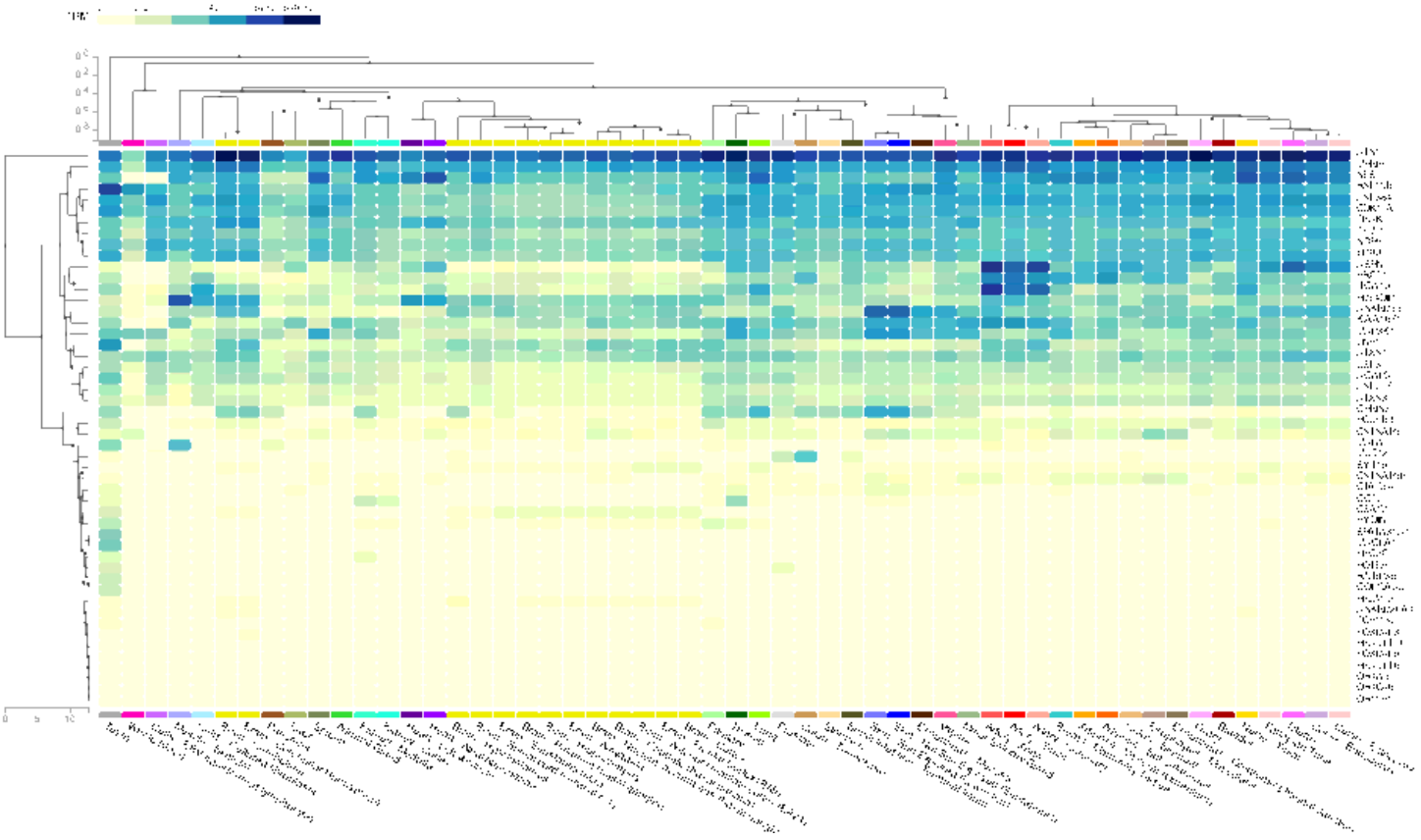

## Figure 1

(A) The proportion of the indicated pathways that had mutations and the proportion of blood samples carrying mutations in the indicated pathways. For example, 5/6 of the genes in the TP53 pathway had mutations and 241/1181 (20.4%) of blood samples had mutations in the RTK-RAS pathway. (B) Correlation plot and interpolation (magenta line) of the blood-based mutation burden and age. The Pearson coefficient and the P-value are shown. (C) Mean AUC value of the 5-fold validation (orange) and the interpolation (blue line) according to the number of trained genes in the LSTM model. AUC, area under a receiver operating characteristic curve.



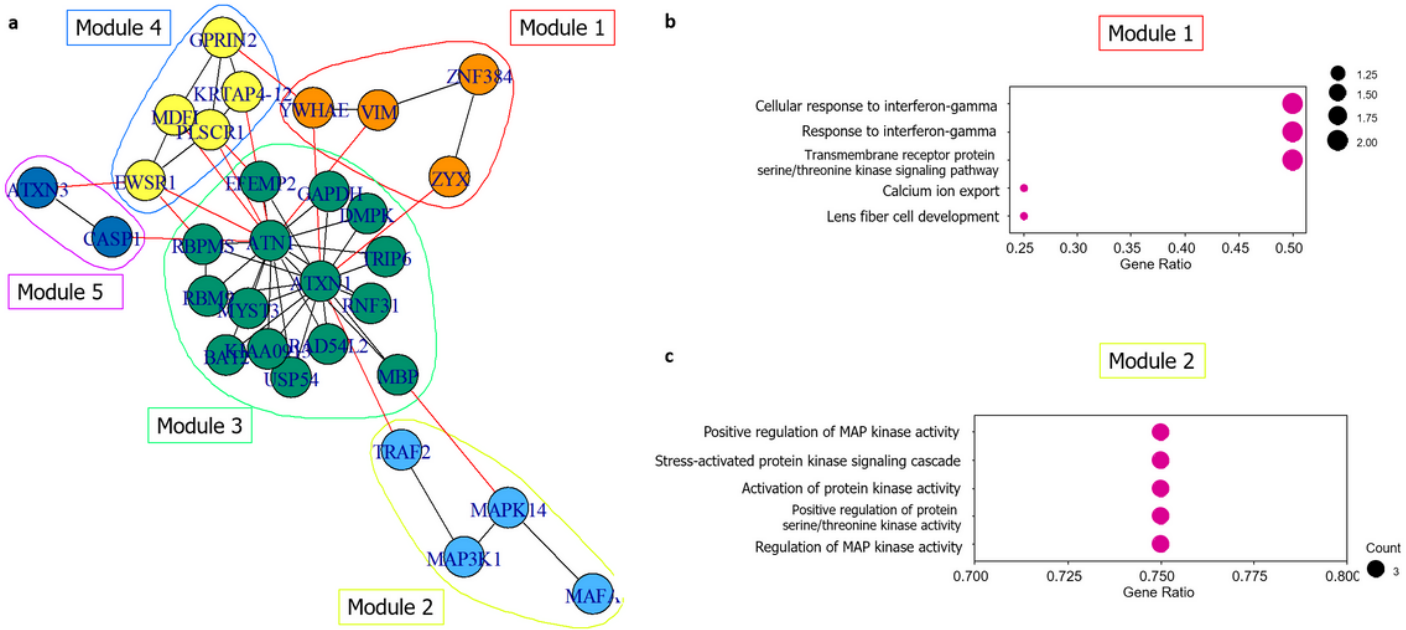
## Figure 2

A heat map representing expression of genes in the gene set (rows) in various human organs (columns) based on the Genotype-Tissue Expression database. TPM, transcripts per million.



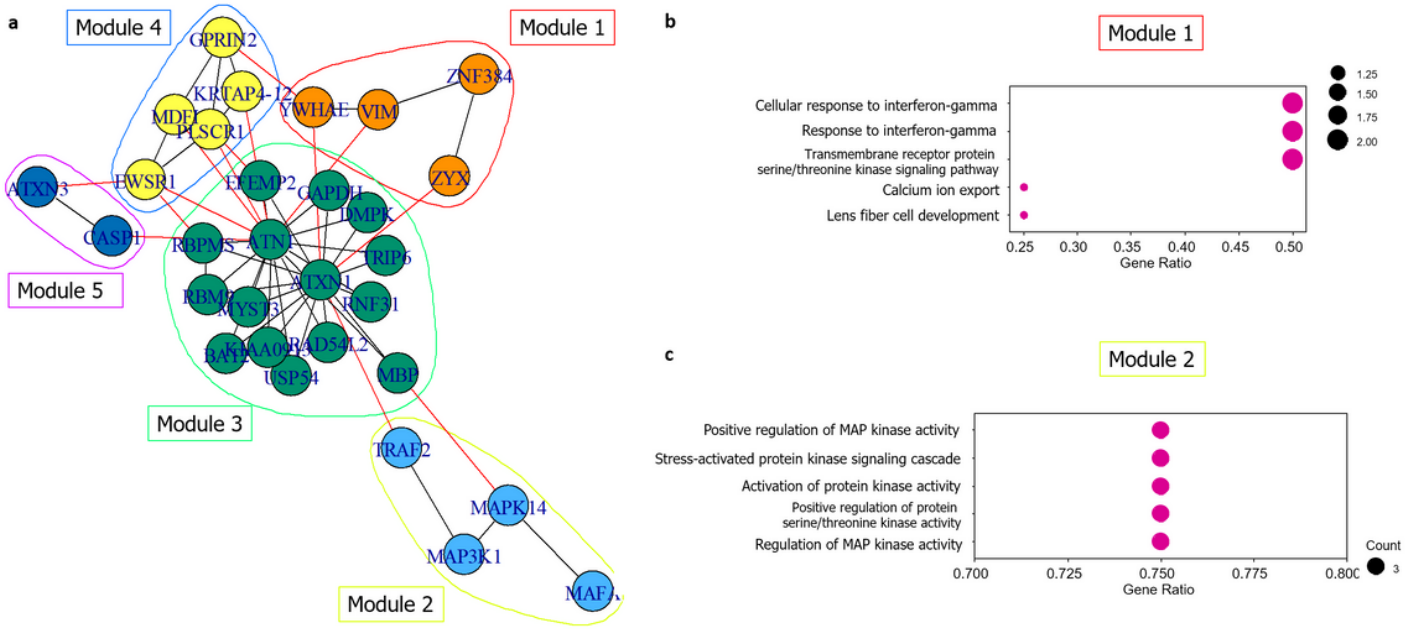
## Figure 2

A heat map representing expression of genes in the gene set (rows) in various human organs (columns) based on the Genotype-Tissue Expression database. TPM, transcripts per million.



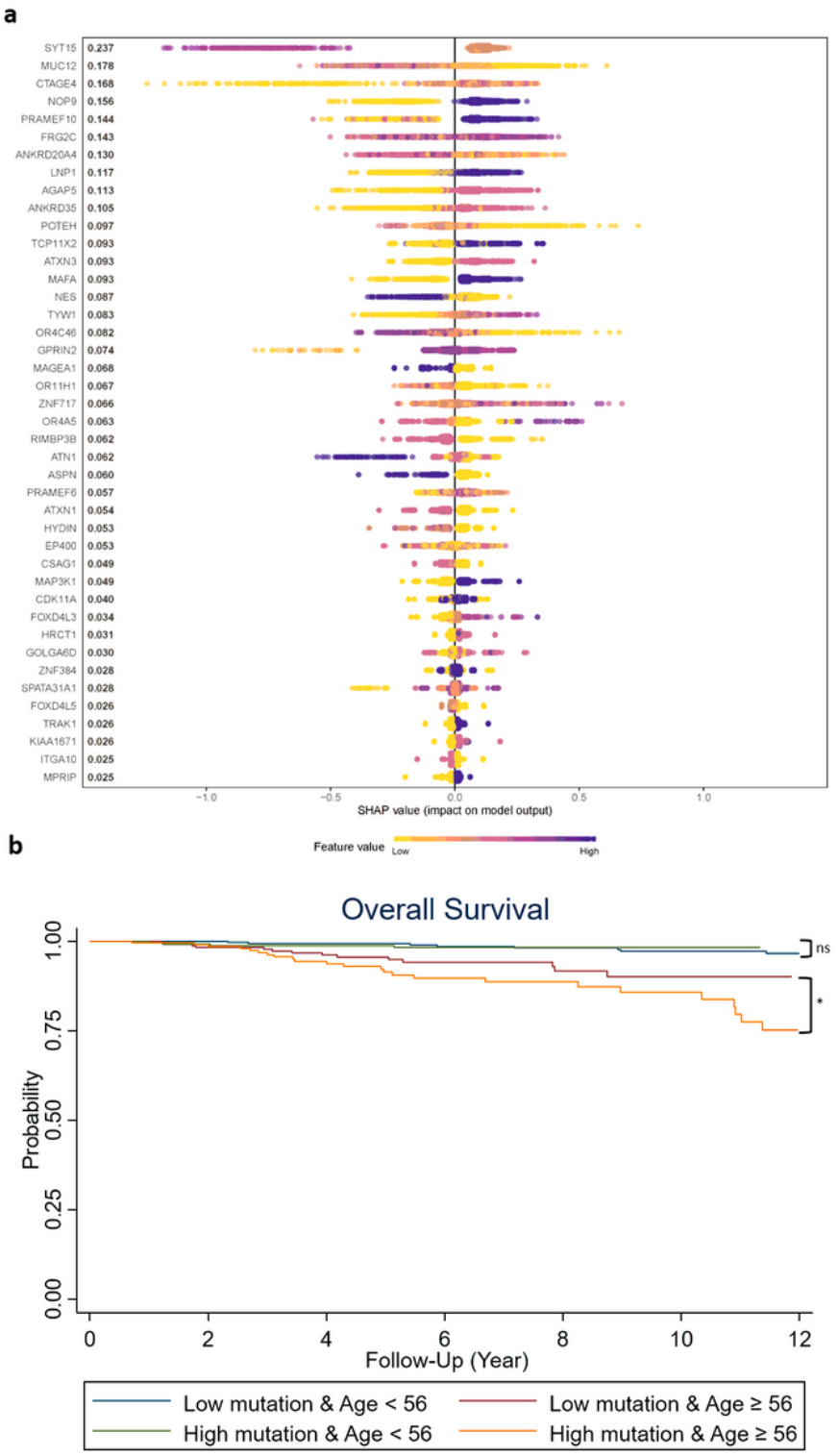

## Figure 3

(A) Four human interaction network modules for the gene set based on The Human Protein Reference Database, Reactome, NCI-Nature Pathway Interaction Database, and The MSKCC Cancer Cell Map database. The colored circle nodes represent genes, the black lines represent intra-module interactions, and the red lines represent inter-module interactions. Module enrichment analyses were performed for module 1 (B) and module 2 (C).



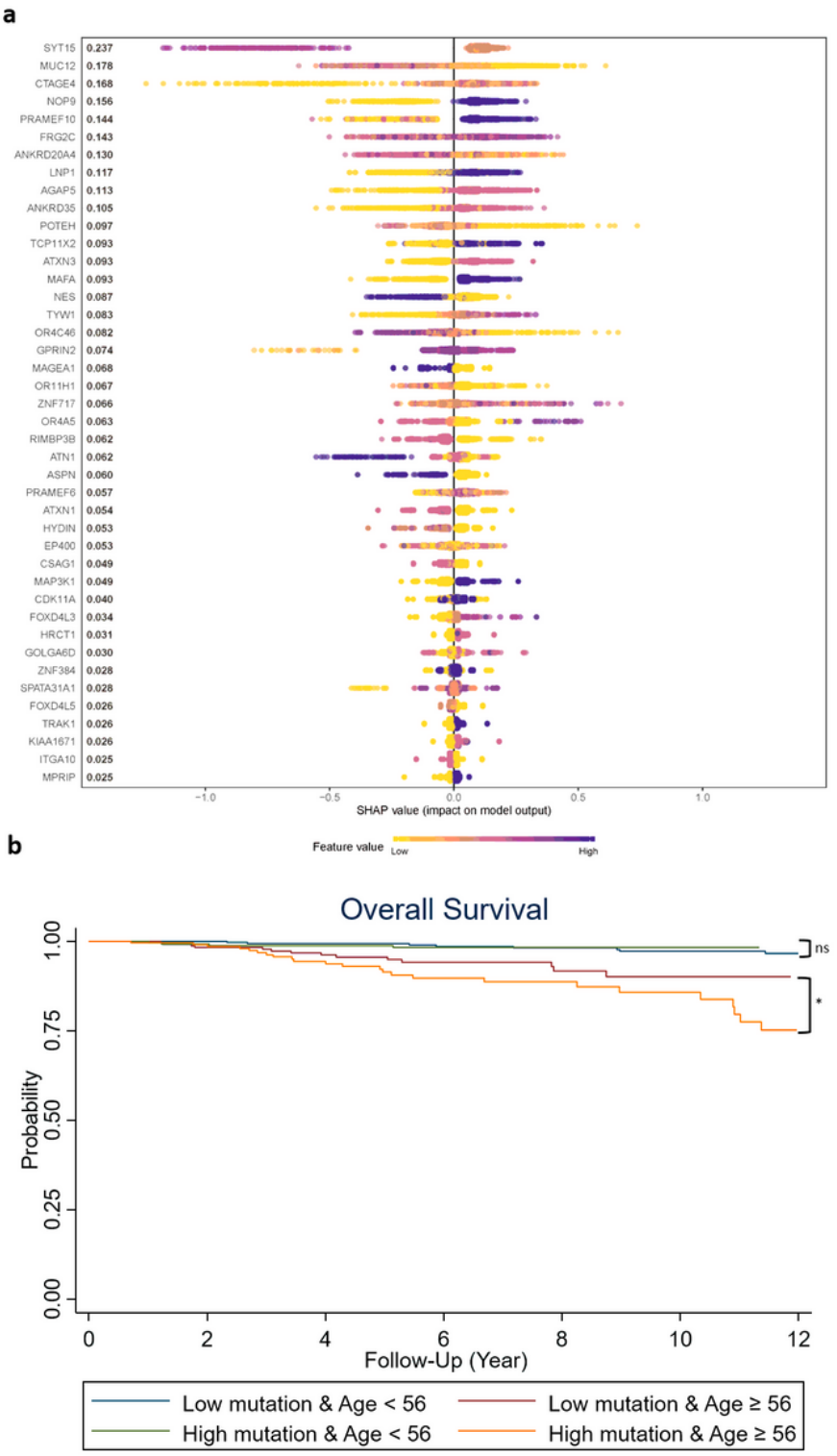

## Figure 3

(A) Four human interaction network modules for the gene set based on The Human Protein Reference Database, Reactome, NCI-Nature Pathway Interaction Database, and The MSKCC Cancer Cell Map database. The colored circle nodes represent genes, the black lines represent intra-module interactions, and the red lines represent inter-module interactions. Module enrichment analyses were performed for module 1 (B) and module 2 (C).




## Figure 4

(A) SHAP value-sorted genes (row) in terms of overall survival. Higher values indicate higher importance in the prediction model. (B) Kaplan-Meier survival curve showing the difference in overall survival according to mutation burden and age groups. P-values were estimated by log-rank test. ns, non-significant; *, P=0.042.




## Figure 4

(A) SHAP value-sorted genes (row) in terms of overall survival. Higher values indicate higher importance in the prediction model. (B) Kaplan-Meier survival curve showing the difference in overall survival according to mutation burden and age groups. P-values were estimated by log-rank test. ns, non-significant; *, P=0.042.

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTableS1.docx
- SupplementaryTableS1.docx